Spring 5-10-2019

The Influence of Question Sequencing Using Formative Assessment in Introductory Statistics

Bryan Nelson
Duquesne University

THE INFLUENCE OF QUESTION SEQUENCING USING FORMATIVE ASSESSMENT IN INTRODUCTORY STATISTICS

A Dissertation

Submitted to the School of Education

Duquesne University

In partial fulfillment of the requirements for

the degree of Doctor of Education

By

Bryan T. Nelson

May 2019

THE INFLUCENCE OF QUESTION SEQUENCING USING FORMATIVE ASSESSMENT IN INTRODUCTORY STATISTICS

By

Bryan T. Nelson

Approved March 8, 2019

Dr. Rachel Ayieko
Assistant Professor of Mathematics Education
(Committee Co-Chair)

Dr. Misook Heo
Professor of Instructional Technology
(Committee Co-Chair)

Dr. Gibbs Kanyongo
Associate Professor of Educational Statistics
(Committee Member)

Dr. Cindy M. Walker
Dean, School of Education
Professor

Dr. Jason Ritter
Chair, Department of Instruction and Leadership in Education

ABSTRACT

THE INFLUENCE OF QUESTION SEQUENCING USING FORMATIVE ASSESSMENT IN INTRODUCTORY STATISTICS

By

Bryan T. Nelson

May 2019

Dissertation supervised by Dr. Rachel Ayieko and Dr. Misook Heo

Formative assessment has long been used to gauge students' understanding of course material prior to taking an exam. With the advent of more advanced technology, only recently have instructors been able to combine formative assessment with a student response system to allow students to respond to questions in real time during class. Previous studies show mixed findings on the relationship between the use of student response systems and student learning. For example, some studies show that those students who used a student response system performed better on exams or in the course when compared to those who did not, while others found no significant difference. In addition, the influence of testing students formatively multiple times before a summative assessment has been of little focus.

A quasi-experimental design was used in this study to test students on 112 concepts in introductory statistics at three time points: during class using a student response system, during an online quiz about one week later, and on an exam at the end of the unit. Each concept was associated with one of four course units and was assigned a level of cognitive demand. The primary goal of the study was to determine if the sequences of correct and incorrect responses that students provided on two formative assessments influenced their ability to answer a corresponding summative assessment question correctly using a logistic regression model and Monte Carlo simulation. Also, a series of loglinear models was used to determine if the sequence of responses, unit of the course, and level of cognitive demand were independent.

The results of this study indicate that students who answered both formative assessments correctly performed the best on the exam, followed by those who answered only the quiz question correctly. However, students who answered only the student response system question correctly fared no better on the exam than those who missed both formative assessments. Students who completed more sequences were more likely to overachieve their predicted exam results. Moreover, the results showed that students' sequences of responses, the course unit, and the level of cognitive demand were not independent. Students tended to overachieve on less cognitively demanding sequences requiring descriptive statistics and on strategic thinking exam questions requiring inference, but underachieved in the probability unit and on more challenging descriptive statistics questions.

This study provided insight on the influence of that repeated practice on exam performance, suggesting that working independently after learning a concept is more beneficial to student learning than using a student response system in class. This study also demonstrated how statistics education can effectively use formative assessment in the classroom and test higher-order thinking using multiple-choice questions. University instructors may find the results useful in reevaluating the use of active learning in their classroom.

DEDICATION

To my parents, Scott and Kathy, whose unwavering support and patience over the past four years allowed me to simultaneously pursue my doctorate, achieve my dream of becoming a university lecturer, and establish a startup company. I would not be as successful as I am today without the invaluable advice from my most steadfast supporters, especially during those difficult times when it felt like there was no light at the end of the tunnel. My ability to complete this dissertation began many years before beginning this program because of their willingness to help me study and proofread my papers.

To my younger brother and fellow statistician, Kevin, who at times assisted me in thinking through the design and results of this study when I needed an unbiased opinion, a clear mind, and a fresh set of eyes.

To my grandmother, Kate Nelson, who I always looked forward to visiting after finishing my classes at Duquesne for the day, dating all the way back to my years as an undergraduate.

To my grandfather, Jim Nelson, for instilling in me the desire to pursue as much education as I possibly could.

To my grandparents, Andy and LaVerne Sickle, who wholeheartedly believed that I could accomplish whatever I set my mind to and become whatever I wanted to be in life. Upon reaching the pinnacle of education, I can only hope that I have made them proud.

ACKNOWLEDGEMENT

Nearly thirteen years ago, I embarked on a journey to earn my Bachelor's degree in mathematics at Duquesne University. Never in my wildest dreams did I imagine that journey culminating in the School of Education over a decade later by completing my doctorate in instructional technology. This dissertation would not have been possible without the support and guidance of several mentors.

I would like to extend my deepest gratitude to Dr. Rachel Ayieko for guiding me through the process of writing this dissertation from start to finish and for her unwavering support of this study over the past two years. Despite the frustrations of writing and rewriting sections multiple times, I have become a much stronger academic writer due to her supervision that will serve me well throughout my career.

The completion of my dissertation would not have been possible without the support of Dr. Misook Heo, whose keen eye provided invaluable feedback and insightful suggestions that greatly enhanced the quality of this dissertation.

I would also like to thank Dr. Gibbs Kanyongo for serving on my committee. His methodological suggestions ensured that the design of this study was solid while still allowing me to conduct it primarily on my own terms.

Finally, I would like to acknowledge Dr. Nancy Pfenning, my colleague at the University of Pittsburgh. The idea for this study emerged as a result of sitting in on her class three years ago while I was an adjunct lecturer and learning how to effectively use a student response system to enhance student learning.

Table of Contents

LIST OF TABLES

Page

Chapter 1

Introduction

Background

The use of lectures to convey information dates back over 2,500 years ago to ancient Samaria and continued through ancient Roman times into the early middle ages when Pope Gregory VII used monks to educate the clergy (Beichner, 2014). Passive learning environments, where instructors convey information directly to their students, are often structured so that students simply listen and accept information from the instructor (Tregonning, Doherty, Hornbuckle, & Dickinson, 2012). While a great deal of content can be covered, lectures are passive with the drawback that student attention drops significantly after the first 15 to 20 minutes of class, hindering the amount of information they retain (Premkumar & Coupal, 2008).

Conversely, active learning involves students being engaged in their own learning through activities (Bonwell & Eison, 1991). The first recorded instance of active learning occurred in the 1800s when European scientists Friedrich Stromeyer and Johann von Fuchs began including laboratory sessions with their lectures for students to gain first-hand experience in chemistry (Beichner, 2014). Although the notion of active learning emerged over 200 years ago, it did not become common in the college classroom until the early 1990s (Mitchell, Petter, & Harris, 2017).

Active learning is associated with constructivism, an educational learning theory in which learners construct personal foundations of knowledge by creating their own interpretation of learning experiences and challenging previous conceptions of the material (Carr, Palmer, & Hagel, 2015; Hartle, Baviskar, & Smith, 2012; Hedden, Worthy, Akins, Slinger-Friedman, & Paul, 2017). By engaging in this constructive process, active learning environments not only help engage students and encourage higher-order thinking (Falconer, 2016), but they narrow the achievement gap between advantaged and disadvantaged students and decrease failure rates (Freeman et al., 2014).

Active learning activities take on many forms, allowing the instructor to change the atmosphere of a lecture-intensive course to one where students think independently (Mitchell et al., 2017). Students are encouraged to engage in higher-order thinking and share their ideas with peers through organized class discussions and games as opposed to listening to the instructor lecture (Mitchell et al., 2017; Slavich & Zimbardo, 2012). Alternatively, writing activities, such as minute papers, instant feedback quizzes, and student-generated questions, allow students to reflect on recent material while also providing the instructor with a quick method of individual assessment (Bidgood, Hunt, & Joliffe, 2010; Mitchell et al. 2017).

Instructional technology has further facilitated active learning by encouraging students to create digital material through multimedia software and coding, watch simulations, and interact with their peers and experts in the field via the Internet (USED, 2017). Many courses now allow online instruction through a course management system, further exposing students to alternative methods of learning via instructional technology (Gikandi, Morrow, & David, 2011; Tishkovskaya & Lancaster, 2012). Student response systems, which are handheld devices that allow students to respond to formative assessment questions, have become popular methods of invoking active learning into the college classroom via technology (Bojinova & Oigara, 2011). They allow all students to answer questions posed by the instructor and receive instant feedback (Bojinova & Oigara, 2011; Klein & Kientz, 2013).

While a wide variety of disciplines use student response systems, they are more commonly used in science, engineering, and medical fields (Carr et al., 2015). Student response systems were more effective primarily in STEM education disciplines because questions tend to have objective responses rather than subjective (Chachashvili-Bolotin, Milner-Bolotin, & Lissitsa, 2016; Hunsu, Adesope, & Bayly, 2016). Although some instructional technologies are universal, the field of statistics uses more specific technologies such as software packages to eliminate basic calculations and applets to display graphs and probability distributions (Chance, Ben-Zvi, Garfield, & Medina, 2007; GAISE, 2016).

In education, assessment can be used for measuring student learning, evaluating teaching quality, and analyzing the stability of a department or university (Fletcher, Meyer, Anderson, Johnston, & Rees, 2012). Assessment can be either formative or summative. Formative assessment is low-risk assessment that aims to improve student learning, enhance teaching, and identify areas of difficulty (Dixson & Worrell, 2016). Summative assessments are higher-risk, cumulative assessments where the goal is to measure student retention and understanding of course material (Dixson & Worrell, 2016). Whereas formative assessment is frequently used to provide students with feedback (Dixson & Worrell, 2016), higher education tends to focus on students' performance on summative assessments such as exams and papers (Gikandi et al., 2011).

Feedback received on summative assessments informing students how they are progressing is often received too late to be meaningful (Gikandi et al., 2011; Hernández, 2012). Instead, immediate feedback from the instructor on formative assessment at the time of questioning can correct misconceptions and narrow gaps in student knowledge (Gikandi, et al., 2011; López-Pastor & Sicilia-Camacho, 2017). Student retention of material also increases if concepts are tested formatively multiple times prior to a summative assessment (Glass, Brill, & Ingate, 2008; King & Joshi, 2008; Yeo, Ke, & Chatterjee, 2015). Formative assessment is also useful to the instructor because the instructor can adjust their method of instruction in the future to better meet the needs of the learners (Trumbull & Lash, 2013).

Assessment items can be rated according to various pedagogical taxonomies such as Bloom’s taxonomy (Bush, Daddysman, & Charnigo, 2014), the structure of observed learning outcomes (SOLO) taxonomy (Stalne, Kjellström, & Utriainen, 2016), and Webb’s depth of knowledge (Wyse & Viger, 2011). One way to measure the difficulty of an item is through the level of cognitive demand, which measures the mental effort required to solve a problem (Wise & Viger, 2011). Webb’s depth of knowledge measures a combination of the cognitive demand and amount of knowledge required to answer the item, which differs from Bloom’s taxonomy (Webb, 2002).

In sum, engaging students through active learning encourages students to think critically and make personal connections to previous material. Using a student response system gives both students and instructors valuable information on students’ progress in the course. Specifically, offering opportunities to experience a second formative assessment is correlated with higher achievement on summative assessments (Glass et al., 2008). In particular, statistics education requires the application of concepts whose difficulties may vary depending on the level of cognitive demand. This study builds on this body of research through the evaluation of repeated formative assessment that includes the use of a student response system, on students’ knowledge gain on summative assessments for concepts in introductory statistics.

Statement of the Problem

Though formative assessment has long been believed to assist in student learning, the lack of a consistent methodology to test this notion has produced mixed results (e.g. Lee, Sbeglia, Ha, Finch, & Nehm, 2015; Yeo et al., 2015). For example, Lee et al. (2015) found that student response system responses had predictive power in determining a student's final grade whereas Yeo et al. (2015) reported that online formative assessments were not associated with students' final grades. These mixed results could be a consequence of inconsistency between the formative and summative assessments. Therefore, writing assessment items at a consistent level of difficulty to test the same knowledge is integral in determining if formative assessment enhances student learning (Glass et al., 2008; Lee et al. 2015). Studies comparing formative and summative assessment scores have found that dissimilar question types have led to lower correlations between the scores, making it difficult to attribute where student learning occurs (King & Joshi, 2008; Yeo et al., 2015). Missing from these studies is a validation that the assessment items are testing the same concepts with the same level of difficulty. Moreover, these studies also lack an intermediate assessment between the initial formative assessment and summative assessment. These omissions make it difficult to ascertain if increases in student knowledge are a result of the formative assessment.

Student response systems have allowed instructors to create an active learning environment that utilizes formative assessment, even in large lectures where interaction with all students was previously impossible (Mateo, 2010). While student response systems appear popular among most students (Bojinova & Oigara, 2011; Haeusler & Lozanovski, 2010; Mateo, 2010; Shaffer & Collura, 2009; Tregonning et al., 2012), questions remain regarding their effectiveness in helping students learn. Some studies on student response systems have found significant improvement in exam scores compared to control groups where they are not implemented (Mayer et al., 2009; Yourstone, Kraye, & Albaum, 2008), while others found no difference in exam scores between an experimental student response system group and control group (Bojinova & Oigara, 2011; Richardson, 2011; Roth, 2012; Sutherlin, Sutherlin, &

Akpanudo, 2013). These particular studies use the average exam scores to compare interventions using the student response systems. However, this method ignores potential confounding variables such as a teaching effect or summative assessment improvements not being uniformly distributed across all course material but instead on a few concentrated topics.

There is limited research on student response systems in statistics education. A few studies such as that done by Richardson (2011) found no significant difference in the grade distribution of an introductory statistics course that used student response systems compared to one that did not. Other studies in introductory statistics (e.g. Büyükkurt, Li, & Cassidy, 2012; Dunham, 2009; Titman & Lancaster, 2011) reported on the student perception of the use of a student response system. Further insight on the influence of formative assessment and student response systems on student learning is possible through a study that includes multiple levels of formative assessment prior to the summative assessment. Unlike previous studies on statistics education that focus on final course grades or perception of student response systems, this study reports on individual item analysis.

Purpose of the Study

The purpose of this study was to investigate if the sequences of students' responses on the formative assessment items were related to their response patterns. Specifically, this study examined whether the sequence of correct and incorrect responses provided by students on two formative assessments (student response system and quiz) on concepts in an introductory statistics course predicted the probability that the student answered the corresponding summative assessment item correctly on the unit exam. The two levels of formative assessment controlled for any learning that may have occurred between the initial formative assessment and summative assessment which prior studies did not account for. In addition to analyzing the effectiveness of student responses systems, this study further examined if student responses differed according to the unit of the course to determine if student tended to require more time to understand more complex concepts in statistics.

This study also built upon the limited research in statistics education by incorporating formative assessment into a field that requires computational, inferential, and analytical skills. This study fulfilled a need to understand how students apply their newly gained knowledge to real-life scenarios in STEM-based courses as opposed to memorizing facts. In addition to analyzing the effectiveness of student response systems, this study further examined if student responses differed according to the unit of the course to determine if students tended to require more time to understand more complex concepts in statistics.

Research Questions

This study analyzed the sequence of correct and incorrect responses provided by individual students on two formative assessments (student response system and quiz) on material in an introductory statistics course to determine if they predicted the probability that the student answered the corresponding summative assessment correctly. The relationship between sequences of responses, unit of the course, and level of cognitive demand was also analyzed.

The questions guiding this study were:

1. How do the sequences of responses provided by students on formative assessment items affect the probability that the student will answer the related exam item correctly while controlling for:
    1) The unit of the course in which the concept is presented
    2) The level of cognitive demand required to answer the concepts' assessment items

3) Students' previous statistics experience

4) Students' interest in the course

5) Students' interest in using a student response system

6) Time of day

2. When responses are aggregated over all students, what is the relationship between:

1) Students' sequences of responses to all three items testing a concept

2) The unit of the course in which the concept is presented

3) The level of cognitive demand required to answer the concepts' assessment items

Significance of the Study

This study provided further insights into assessment in higher education. The analysis of the sequences of responses on similar items helped to determine if a second level of formative assessment was effective in assisting students' understanding introductory statistical concepts. The findings related to the active engagement in class were beneficial to students' overall conceptual understanding of the subject matter on summative assessments, which provided more evidence to support decisions on redefining attendance policies and participation in higher education courses. Similarly, this study showed support for strategies that assist instructors in identifying students who are at risk and allow students the opportunity for self-evaluation.

This study also demonstrated that objective format questions are effective as the sole means of assessment in a computational field such as statistics. Identifying common mistakes and including these incorrect answers as distractors allowed the instructor to address these misconceptions universally. This feedback, which did not need to be personalized, was effective in improving student understanding of statistical concepts between evaluations.

Finally, this study provided insight into possible methods for modifying instruction in statistics education. Having access to students' responses on formative assessments allows instructors to identify concepts where students repeatedly chose incorrect answers. This study recognized statistical concepts where the feedback was beneficial as a result of high success on the summative assessments while also detecting concepts where feedback could be modified due to lower than expected success on the exam.

Definitions

Formative assessments are questions that students answer with the goal of improving understanding and knowledge. In this study, formative assessment items are those that students answered using the student response system and on the online quizzes.

Summative assessments are questions that students answer in order to gauge knowledge gain. In this study, summative assessment items are those from the unit exams.

The level of cognitive demand refers to the mental effort required to solve a problem. In this study, the level of cognitive demand was determined according to Webb's depth of knowledge. From lowest to highest, the levels of cognitive demand in this study are recall, basic application of skill, and strategic thinking (Wyse & Viger, 2011).

Chapter 2

Literature Review

Introduction

Many prior studies have attempted to quantify the impact that formative assessment has on student learning. Some have examined the impact of multiple formative assessments (e.g. Glass et al., 2008) while others have solely analyzed the use of a student response system as a way of implementing formative assessment (e.g. Mayer et al., 2009; Richardson, 2011). However, the outcome of interest has not always been students' individual responses to summative assessments.

By employing a student response system, instructors simultaneously get students actively involved and test their knowledge using formative assessment (Blood & Gulchak, 2012). While results are mixed as to whether formative assessment improves grades, students tend to have positive feelings towards using student response systems in higher education (Blood & Gulchak, 2012; Bojinova & Oigara, 2011; Glass et al., 2008). Moreover, when students are actively engaged during class, they are more likely to understand the course material and less likely to fail (Falconer, 2016). Constructivists claim one benefit of active learning results from social interaction, while cognitivists assert that learning is a result of being engaged (Mayer et al., 2009; Park & Choi, 2014).

The review of the literature, which describes the topics guiding this study, is structured as follows. It begins with a discussion of constructivism and active learning. Different means of assessing students in higher education are presented next. A brief review of educational taxonomies follows. This is succeeded by reviewing the literature on formative assessment, student response systems, and other studies that implemented question sequences. Statistics education with a focus on educational technology is reviewed next before concluding with a critique of the literature.

Constructivism

Constructivism is a student-centered learning theory whereby students construct their own individual foundations of knowledge through active engagement in a topic (Hartle et al., 2012). The notion of constructivism emerged from Jean Piaget's research on cognitive development where he discovered that learners acquire new knowledge by combining new information with preexisting schema (Harow, Cummings, & Aberasturi, 2006). According to Piaget, learners' success in solving problems is dependent upon their current skill set and level of understanding (Keengwe, Onchwari, & Agamba, 2014). Constructivism has branched off into several different versions, including radical constructivism and social constructivism (Van Bergen & Parsell, 2019). Whereas radical constructivism is distinguished by an individual's realizations with the material, social constructivism believes that knowledge is constructed through interaction with peers (Van Bergen & Parsell, 2019). Social constructivism emerged from Piaget's research and emphasizes learning through observation, knowledge sharing, and active learning (Barak, 2016). Due to its emphasis on peer interaction, critical thinking, and problem solving, teachers in STEM education have been encouraged to implement constructivist learning techniques in their classes (Barak, 2016). Despite their differences, both radical and social constructivism agree that learning is an active process whereby learners construct their own knowledge through personal experiences (Van Bergen & Parsell, 2019).

Constructivism encourages teachers to create challenging in-class activities that require critical thinking and problem solving from students (Poelmans & Wessa, 2015). One method of accomplishing this is by using formative assessment, which allows the instructor to immediately correct misconceptions by offering feedback to the entire class that students can compare with their prior knowledge (Hartle et al., 2012). To construct new knowledge, students enter the classroom with prior knowledge from previous learning experiences, which they then combine with the new learning experience (Mvududu, 2005).

Constructivism requires that students be active learners who interact with teachers, classroom materials, and other students (Keengwe et al., 2014; Mvududu, 2005). These interactions encourage students to update their base of knowledge by resolving how it differs from the new information they receive during active learning (Slavich & Zimbardo, 2012). Students begin to understand how their collective knowledge can be practically applied through cognitive constructivism, an educational learning theory where students perform relevant real-world tasks that interact with prior knowledge to actively construct new knowledge (Tishkovskaya & Lancaster, 2012). To construct this new knowledge, learners must have cognitive presence, which is a multistage process by which learners identify the problem, explore new ideas, integrate old knowledge with new ideas, and resolve the problem (Garrison, Anderson, & Archer, 1999).

Students may also engage in social constructivism, which considers learning to be primarily a cooperative social activity where learners explore new venues, experience increased engagement, co-construct content, and provide and receive feedback (Poelmans & Wessa, 2015; Barak, 2017). Activities in the classroom should present a challenge to students by forcing them to realize that their current body of knowledge is insufficient to solve the current problem creating a cognitive imbalance and defining the point at which students develop knowledge (Hartle et al., 2012; Barak, 2017). This cognitive dissonance allows students to learn best through modifying prior knowledge by observing gaps and inconsistencies, reinforcing new knowledge through repetition to overwrite previous biases, and reflecting on the differences between prior and new knowledge (Hartle et al., 2012; Barak, 2017; Slavich & Zimbardo, 2012). Teachers must also interact with students by providing feedback on formative assessments to correct any misconceptions that will assist them in learning (Keengwe et al., 2014). Due to its interactive nature, formative assessment plays a major role in constructivism (Slavich & Zimbardo, 2012).

While instructors control the content of the class and present the material to all students simultaneously, each learner has different baseline knowledge that assists them in constructing their own individual meaning to the information presented (Brooks & Brooks, 1999). When preparing lessons, constructivist teachers choose material that challenges students' prior knowledge, assesses student learning, and is relevant to the real world (Brooks & Brooks, 1999). Formative assessment plays a major role in constructivism due to its interactive nature. This includes giving quizzes to students, encouraging students to work in groups, or using a student response system for in-class participation (Slavich & Zimbardo, 2012; Hartle et al., 2012). Instructors must understand how to integrate technology into learning activities that encourage inquiry-based learning so that students become proficient in their own learning (Keengwe et al., 2014). Bringing interactive learning activities into the classroom that allow students to construct their own knowledge is critical to the success of active learning (Hedden et al., 2017).

Active learning.

The primary paradigm in modern education is constructivism, which claims that knowledge must be constructed by the learner (Kumar, McLean, Nash, & Trigwell, 2017). One way of doing so is through active learning, where students challenge their current base of knowledge with new information they discover while actively exploring new topics (Carr et al., 2015). European scientists Friedrich Stromeyer and Johann von Fuchs were among the first to implement active learning by giving students hands-on experience in chemistry laboratory sessions (Beichner, 2014). Instructors did not begin moving away from lectures and towards active learning until about thirty years ago; nonetheless, it has rapidly developed into a useful method of engaging students through methods such as case studies, formative assessment, and technology simulations (Mitchell et al., 2017).

Active learning occurs when students become engaged in the course material by completing activities and thinking about the actions they are performing (Falconer, 2016). When students actively participate, the instructional approach shifts from lecture to classroom discussion focused on higher order thinking (Ghilay & Ghilay, 2015). Active learning helps increase retention, particularly in STEM fields, which reduces the frequency with which students believe they understand a problem only to find they cannot apply the concept on their own after the lesson (Falconer, 2016). Students learn how to apply their knowledge to real-world situations rather than memorizing facts (Mitchell et al., 2017). When implemented in the classroom, active learning doubled students' knowledge gain, decreased the failure rate, and narrowed the gap between socioeconomically advantaged and disadvantaged students (Falconer, 2016).

Class discussions are the most straightforward means of implementing active learning as they encourage students to converse about how they could use the material in new situations, which assists in long-term retention (Ghilay & Ghilay, 2015). Peer learning also employs active learning because it helps students comprehend the course material while understanding how to work as a team (Kroning, 2014). While working in small groups, students can learn from their peers, collaborate to find solutions to problems, and engage in role-playing (Coorey, 2016;

Hedden et al., 2017). In larger classes, instructors may use case studies to ask questions that require students to analyze and synthesize information (Carloye, 2017). Active learning can continue outside of the classroom by immersing students in semester-long projects that may necessitate collaboration in small groups (Hedden et al., 2017; Mitchell et al., 2017). Students collect their own data on a personally selected topic and present their findings in class, allowing them to both offer and receive peer feedback (Strangfeld, 2013).

Due to the increased accessibility to technology, instructors have more options for integrating active learning into their pedagogy (Mitchell et al., 2017). To promote active learning within the classroom, instructors can use interactive whiteboards, student response systems, simulations, team games, and videos (Mitchell et al., 2017; Tomei, 2013). To keep students learning while not in class, instructors have numerous options that were unavailable until recently such as wikis, blogs, social media, and podcasts (Mitchell et al., 2017; Tomei, 2013). Previously viewed as distractions, cell phones and laptops have found a place in the active learning classroom. Both offer the ability to allow all students to respond to instructor-posed questions during the class (Ghilay & Ghilay, 2015; Klein & Kientz, 2013). Instructors can pose an open-ended question and ask students to perform active research during class to find an answer on the Internet (Kroning, 2014). Not only can instructors test students' knowledge of basic information, but they can ask more difficult questions that require students to think analytically before providing them with feedback, which is a key facet of formative assessment (Klein & Kientz, 2013). Thus, student response systems are an effective way of utilizing formative assessment because all students participate simultaneously and the instructor can provide universal feedback to all students (Lee et al., 2015).

Assessment in Education

Assessment in higher education is a means by which instructors can determine gains in student knowledge, provide feedback, and evaluate teaching effectiveness (Fletcher et al., 2012). Assessment should be continuous to support students in their learning by informing them of their progress in the course and offering many opportunities to demonstrate knowledge before assigning a final grade (Fletcher et al., 2012; Hernández, 2012). Student progression in higher education depends on students being assessed according to the standards in the curriculum to ensure they are prepared to continue to more advanced courses (Bearman et al., 2017).

There are several ways of classifying assessment items. Instructors may use earlier assessments for teaching and later assessments for testing retention of material (Gikandi et al., 2011) while allowing students to either choose one of several possible answers or compose their own response (Ozuru, Briner, Kurby, & McNamara, 2013). The level of cognitive demand required to answer a question is another way of classifying an assessment item (Bush et al., 2014; Stalne et al., 2016; Webb, 2002).

Types of assessment items.

Two primary question types exist that allow instructors to assess student knowledge: multiple-choice and open-ended (Ozuru et al., 2013). Multiple-choice items are questions or incomplete statements where the respondent chooses one of several possible options to answer the question or complete the statement (Bush, 2015). Time is best spent creating plausible incorrect answers as they are more discriminating than having many farfetched distractors that are easily eliminated (Haladyna, Downing, & Rodriguez, 2002). Multiple-choice exams tend to have high reliability as long as each question is valid and there are a sufficient number of items on the test (Palmer & Devitt, 2007). The respondent's level of prior knowledge, which has a positive correlation with performance on the item, is one confounding variable on multiple-choice items (Ozuru et al., 2013). Although test banks are somewhat limited in some content areas, computational fields such as mathematics and statistics can easily generate new items by changing the numbers in the problem (Bush, 2015).

Open-ended questions in a computational field, such as statistics, are items that emphasize students' conceptual understanding of a problem (Sanchez, 2013). Open-ended questions all students to display partial knowledge of a topic (Attali, Laitusis, & Stone, 2016) but may also expose students' lack of comprehension (Sanchez, 2013). Both essay questions and modified essay questions have high reliability; moreover, they are uncorrelated with a student's level of prior knowledge because students must possess pertinent and correct facts to answer the question (Ozuru et al., 2013; Palmer & Devitt, 2007).

Studies on multiple-choice and open-ended questions.

Though students respond in different ways, studies have shown that multiple-choice and open-ended questions can parallel one another (Attali et al., 2016; Palmer & Devitt, 2007; Wainer & Thissen, 1993). If a multiple-choice and open-ended question require the same cognitive demand and test the same content, then the ability of the multiple-choice question to predict performance on an open-ended question is high (Attali et al., 2016; Palmer & Devitt, 2007). The correlations between multiple-choice and open-ended sections of the mathematics, computer science, and chemistry Advanced Placement exams exceeded .80, largely due to the high level of cognitive demand required to answer the questions (Wainer & Thissen, 1993). The level of cognitive demand is correlated with performance on both types of assessment items, but subjects in an eighth-grade mathematics class reported having increased worry on open-ended questions (O'Neill & Brown, 1998).

Types of assessment in education.

Assessment can be either formative, with the goal of improving both teaching and learning, or summative, which is a means to test student knowledge at the end of a unit or course (Gikandi et al., 2011). However, instructors and students often view assessment differently. Instructors tend to use assessment to understand what students have learned (Fletcher et al., 2012). Conversely, students have varying views of assessment; some believe assessment is important to obtain their degree while others find assessment and irrelevant and unfair (Fletcher et al., 2012). The following sections will detail the features of each type of assessment.

Formative assessment.

Formative assessment is a type of assessment characterized by its goal of improving both teaching and learning in a low-stakes environment, often with feedback provided at the conclusion of the exercise (Dixson & Worrell, 2016). Formative assessment has several strengths, including increasing student motivation, correcting gaps in knowledge, and improving academic performance (López-Pastor & Sicilia-Camacho, 2017). Providing students with effective formative feedback accomplishes these strengths.

Formative feedback should provide pertinent insight about student learning, incite a dialogue between the students and instructor, allow students the opportunity to take a step closer to their desired level of understanding, and encourage reflection about learning (Gikandi et al., 2011). Formative feedback is most effective when it explains to students how they could have completed the assessment to a higher degree or improve upon their skills rather than offering only praise or punishment (López-Pastor & Sicilia-Camacho, 2017). Formative assessment is not limited to the physical classroom. Online formative assessment allows the instructor to monitor student learning outside of the classroom while also providing timely and constructive feedback (Gikandi et al., 2011). Regardless of the type of feedback, it must be timely and directly related to the knowledge that students already possess; otherwise, students tend to ignore the comments (Gikandi et al., 2011).

Summative assessment.

Summative assessments, on the other hand, gauge the material student has learned against developed course standards (Dixson & Worrell, 2016). They are typically administered periodically throughout the course and designed to test student knowledge over a set of topics covered over the previous several weeks or months (USED, 2017). Summative assessments can be analyzed using many different metrics such as reliability and validity for individual assessments and student contributions for group projects (Mulder, Pearce, & Baik, 2014). Receiving a score that largely impacts a students' course grade is a key identifying factor of summative assessment (Hernández, 2012). Students may also receive feedback from summative assessments through self-correcting multiple-choice exams (Grüng & Cheng, 2014).

Educational assessment taxonomies.

Assessment items in education differ according to the level of cognitive demand, or mental effort, required to solve a problem (Wyse & Viger, 2011). Asking questions at higher levels of cognitive demand tends to increase student engagement, which may lead to higher academic success (Paige, Sizemore, & Neace, 2013; Rush, Rankin, & White, 2016). Many educational taxonomies have been developed that classify questions according to their difficulty or level of cognitive demand. Most renowned is Bloom's taxonomy, which classifies learning objectives and assessments into six ascending levels (Bush et al., 2014). Other taxonomies have conceptualized following Bloom's taxonomy such as the SOLO taxonomy (Stalne et al., 2016) and Webb's depth of knowledge (Webb, 2002).

Bloom's taxonomy.

Bloom's taxonomy, developed by Benjamin Bloom in 1956, assigns one of six levels to an assessment item according to the educational goal: remembering, understanding, applying, analyzing, evaluating, and creating (Bush et al., 2014). Bloom's taxonomy categorizes tasks according to their complexity rather than their difficulty (Dunham, Yapa, & Yu, 2015). Items classified as remembering or understanding require a lower level of cognitive demand, while the levels from applying through creating entail higher levels of cognitive demand (Dunham et al., 2015). In science education, solving problems requires higher levels of Bloom's taxonomy while replicating processes involves lower levels (Bush et al., 2014).

SOLO taxonomy.

The SOLO taxonomy, developed by John Biggs and Kevin Collis in the 1970s, assigns a category to students' responses according to the degree of understanding students display (Hattie & Brown, 2004). Prestructural responses are those that demonstrate a failure to comprehend the problem (Stalne, 2016). Unistructural or multistructural answers show a basic understanding of one or two concepts on the surface, while responses displaying deeper knowledge and interconnected ideas are classified as relational or extended abstract (Hattie & Brown, 2004). Unlike Bloom's taxonomy, the SOLO taxonomy allows for the classification of students' responses rather than questions (Hattie & Brown, 2004).

Webb's depth of knowledge.

Webb's depth of knowledge is a set of criteria developed in 1997 that describes the level of reasoning that students must possess to respond correctly to an assessment item (Wyse & Viger, 2011). The four depth of knowledge levels identified by Webb from lowest to highest are recall, basic application of skill, strategic thinking, and extended thinking (Holmes, 2011).

Recall requires that students remember a fact or definition; basic application of skill involves using some mental processing and demonstrating basic understanding; strategic thinking requires reasoning and using evidence; extended thinking necessitates complex reasoning, possibly over a period of time (Holmes, 2011). The level of cognitive demand required is low for recall, moderate for basic application of skill, and high for both strategic thinking and extended thinking (Son, 2012). Item difficulty is independent of the level of cognitive demand in Webb's depth of knowledge as a challenging item may require little cognitive demand (Wyse & Viger, 2011).

Webb (2002) defined characteristics in several content areas, including mathematics, that were indicative of each of level of cognitive demand. Applying a basic algorithm or definition is at the recall level for mathematics (Webb, 2002). The basic application of skill level incorporates making comparisons and decisions on how to approach a mathematics problem (Webb, 2002). Tasks get more complex at both the strategic thinking level, characterized by coming to conclusions and justifying responses, and the extended thinking level, which may require the synthesizing of ideas or having students conduct their own study over a period of time (Webb, 2002).

Webb's depth of knowledge has found a niche in test design when the level of cognitive demand is of interest (Hess, Carlock, Jones, & Walkup, 2009). For instance, it was used to evaluate an information technology curriculum (Harris & Patten 2015) and assess creative learning principles in high school fine arts classes (Ellis, 2016). Heller, Daehler, Wong, Shinohara, and Miratrix (2012) used Webb's depth of knowledge to classify content knowledge questions on tests given to teachers. Webb (2002) described how depth of knowledge can be directly applied to mathematics and science education, of which statistics education is a subset.

Figure 1 displays the structures of Bloom's taxonomy, the SOLO taxonomy, and Webb's depth of knowledge.

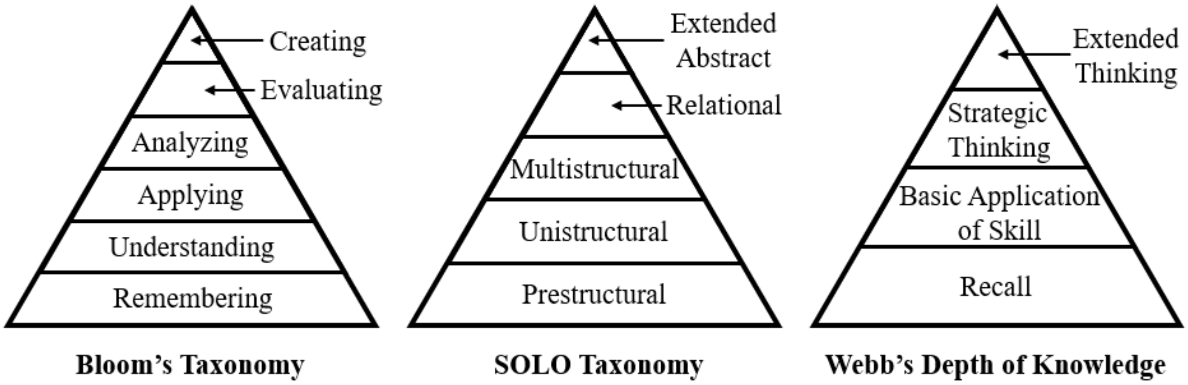


Figure 1. Webb's depth of knowledge levels as compared to Bloom's taxonomy and the SOLO taxonomy

Student response systems as a form of formative assessment

Student response systems combine active learning and formative assessment in higher education by breaking up the monotony of a lecture into smaller sections interrupted by a question (Egelandsdal & Krumsvik, 2017). Students benefit from using a student response system by receiving feedback from the instructor, seeing a graph of the responses from the entire class, and interacting with their peers (Katz, Hallam, Duvall, & Polsky, 2017; Lantz, 2010). Student response systems make students feel more engaged in their own learning, which leads to better performance on summative assessments (Bojinova & Oigara, 2011). Moreover, student response systems can be used as an incentive to increase participation or earn points for correctness (White, Syncox, & Alters, 2011). Student response systems motivate students by promoting attendance and increasing interest in course material while simultaneously providing feedback to the instructor about students' current base of knowledge through class-wide voting (Blood & Gulchak, 2012).

Despite some instructors believing that electronic devices detract from learning, cell phones and laptops have found a place in the college classroom because they can be used to respond to formative assessment items and be connected to a learning management system (Ghilay & Ghilay, 2015; Kroning, 2014). Using a personal device eliminates the need for students to buy their own student response system at an additional cost and prevents wasting valuable class time distributing and collecting student response systems that are provided by the department (Katz et al., 2017). Cell phones and laptops have some disadvantages. Both require a consistent wireless connection and cell phones often require unlocking with each use (Katz et al., 2017). For these alternate student response systems to be successful, the instructor must possess knowledge in both pedagogy and technology (Ghilay & Ghilay, 2015). Many recent studies have attempted to quantify the impact that student response systems have on perception and achievement, which the following section will summarize.

Results are mixed as to whether a student response system aids in student learning. Some studies have found that students who used a student response system scored significantly higher on midterm and final exams than those who did not (e.g., Mayer et al., 2009; Shaffer & Collura, 2009; Yourstone et al., 2008). This was particularly true when students had seen concepts that were previously tested using formative assessment (Yourstone et al., 2008). Conversely, other studies uncovered no significant differences in grade distributions (Richardson, 2011), individual final exam questions (Roth, 2012), or average midterm exam scores (Symister, VanOra, Griffin, & Troy, 2014).

Feedback also plays a major role in student success when combined with a student response system. Differences in midterm and final exam scores in two operations management classes were attributed to the feedback students who used the student response system received compared to the delay caused by traditional assignments (Yourstone et al., 2008). Students who used a student response system in a psychology course and received feedback scored higher than the control groups that either did not use a student response system or did not receive feedback (Lantz & Stawiski, 2014).

The effectiveness of a student response system may be dependent upon the strength of the student within the discipline. High performing students tend to score well regardless of the method of instruction while lower-performing students who used a student response system scored significantly higher than their peers who did not use one (Roth, 2012). Students who are interested in the course material tend to have mastery goal orientation, which is also associated with higher exam scores (Harlow, Harrison, & Meyertholen, 2014; Zingaro, 2015). Student response systems often encourage attendance, which could help lower-performing students as they tend to be the most negatively affected by missing class (Westerman, Perez-Batres, Coffey, & Pouder, 2011).

Student perception of formative assessment with student response systems.

A review of the literature shows that students generally have positive opinions when their college courses use a student response system, regardless of the discipline. Multiple studies have reported results where student enjoyment is in the majority (e.g., Bojinova & Oigara, 2011; Haeusler & Lozanovski, 2010; Tregonning et al., 2012). Ninety-eight percent of students in an obstetrics and gynecology course reported enjoying using a student response system for a summative assessment at the end of eight lectures (Tregonning et al., 2012). Bojinova and Oigara (2011) found that 77 percent of economics students and 95 percent of geography students believed that the material was more interesting due to using a student response system. Students in pre-service science and mathematics courses reported that the student response system was more interesting if implemented throughout class as an instructional method rather than as a tool to test conceptual knowledge all at once (Haeusler & Lozanovski, 2010).

Anonymity is a major factor that strongly influences students' enjoyment of student response systems, particularly for those who are introverted or easily embarrassed (Blood & Gulchak, 2012). Students were more willing to participate in an introductory business course because the student response system provided anonymity, leaving no threat of embarrassment for answering incorrectly (Heaslip, Donovan, & Cullen, 2014). Over 80 percent of medical students in a study by Tregonning et al. (2012) reported that using a student response system was an appropriate means of assessment because mistakes remained anonymous. Bojinova and Oigara (2011) found that 80 percent of economics students and 70 percent of geography students appreciated the anonymity of responses.

Students reported mixed feelings about the benefits of using a student response system to learn the material in the course. When students found a student response system to be enjoyable and valuable to their learning, instructors observed increases in intrinsic and extrinsic motivation (Buil, Catalán, & Martínez, 2016). Over three-quarters of students in introductory biology and chemistry courses felt that using a student response system assisted in learning the course material more effectively, although it may distract motivated students in higher level courses (Sutherlin et al., 2013). About two-thirds of students in the study by Bojinova and Oigara (2011) economics and geography classes felt they learned the material better using a student response system compared to a traditional lecture while the other third was indifferent. More mixed results have been found from studies in statistics courses. Half of the students in a general introductory statistics course felt the student response system helped them understand the content better and one-sixth believed it was beneficial to the course (Mateo, 2010), while a third of business statistics students felt the student response system was either useless or not helpful (Büyükkurt et al., 2012).

Some researchers observed increases in both attendance and participation when using student response systems in their classes. Students are more likely to attend class if they are engaged with the course material, which is a primary goal of student response systems (Katz et al. 2017). Students have reported that lectures utilizing a student response system are more intellectually stimulating than those without because they encourage student participation, create discussion, and promote interactivity (Shaffer & Collura, 2009). Student response systems create a dialogue between students and the instructor that is beneficial to the learning process (Haeusler & Lozanovski, 2010; Yourstone et al., 2008). However, results within the statistics discipline are mixed. Nearly three-quarter of students in an introductory statistics course reported that the use of a student response system increased their likelihood of attending class (Mateo, 2010). Conversely, Büyükkurt et al. (2012) found that a majority of students in a business statistics course were not more motivated to attend, prepare, or study for the class.

Researchers have nearly universally found that the feedback received from using a student response system is beneficial to students. Feedback is important because it assists students in understanding their level of performance, which is related to higher levels of self-efficacy and gives students more confidence in controlling their future learning (Buil, Catalán, & Martínez, 2016). Higher levels of self-efficacy are correlated with higher exam scores (Galyon, Blondin, Yaw, Nalls, & Williams, 2011). Over three-quarters of students in an introductory statistics course reported that the feedback they received from using a student response system was beneficial in determining which concepts they understood compared to 8 percent who deemed the feedback unhelpful (Mateo, 2010). Students benefit from peer interaction after receiving feedback from the results of a student response system question (Katz et al., 2017).

One characteristic that appears to make student response systems quite popular among students is their ease of use. Students in an Irish business statistics course gave the student response system an average rating of 4.3 on a five-point scale for ease of use, implying that the student response system was effortless and simplified interactivity (Heaslip et al., 2014). When used in a digital marketing course, students responded similarly about ease of use, giving the student response system an average rating of 5.81 on a seven-point scale with satisfaction levels tending to increase as the student response system becomes easier to use (Rana & Dwivedi, 2016). Students are more likely to report increased concentration during class when they find the student response system easy to use (Bojinova & Oigara, 2011).

Student response systems in statistics courses.

Despite the increase in the popularity of student response systems, studies analyzing their effectiveness in assessing student learning in introductory statistics courses at the university level are limited. One study found a significant difference on only one of the four exams and no significant difference in the final grade distributions when comparing a class that used a student response system and one that did not (Richardson, 2011). Several other studies investigated how students perceived using student response systems in class, but they did not assess student performance or student learning. Statistics students viewed a student response system as both useful and motivating in an introductory business statistics course (Büyükkurt et al., 2012). Students in a small statistics class reported that the student response system helped identify personal strengths and weaknesses, but student performance was not analyzed (Titman & Lancaster, 2011). Dunham (2009) used a student response system in two different introductory classes, noticing the benefits of feedback, discussion, and the ability to revisit difficult concepts, but never addressed if they improved student knowledge.

Question sequences in formative assessment.

While the focus of student response system use has typically been comparing their use with students' grades on exams and in the course, a few studies matched students' responses on formative assessments with answers on corresponding exam questions. In a general chemistry class, 82 percent of students who attempted a student response system question answered the similar exam questions correctly, regardless of their answer on the formative assessment was correct or incorrect (King & Joshi, 2008). Students who did not attempt the formative assessment answered only 72 percent exam questions correctly, which was significantly worse than their peers who attempted the formative assessment (King & Joshi, 2008).

Glass et al. (2008) compared four formative assessment conditions (online formative assessment, in-class formative assessment, both, or neither) with the probability of answering a multiple-choice exam question correctly in a general psychology course. Students who answered both types of formative assessment were significantly more likely to answer the exam question correctly than any of the other three groups, leading to the conclusion that repeated testing produces better retention (Glass et al., 2008). Moreover, students who answered one type were significantly more likely to respond to the exam question correctly than students with no exposure to formative assessment (Glass et al., 2008).

Yeo et al. (2015) offered optional online formative assessments in an accounting course. While students who completed the quizzes scored significantly higher on each individual midterm topic, the online quizzes were not found to significantly predict final exam scores (Yeo et al., 2015). This may have occurred because the midterm consisted of multiple-choice questions that were similar to the quizzes while the final exam was comprised of open-ended questions (Yeo et al., 2015). This premise is consistent with Glass et al. (2008) who suggested that both the additional study hypothesis, where students learn little more than the material presented on formative assessments, and the repeated testing hypothesis, where students tend to focus on content presented on formative assessments, explain why pretest scores correlate highly with performance on corresponding exam questions. Hubbard and Couch (2018) observed that having multiple formative assessments often benefits higher-performing students as they tend to retain content knowledge better than their lower-performing peers.

Statistics Education

Content in an introductory statistics course at the university level is divided into four distinct processes: data collection, descriptive statistics, probability, and inference (Pfenning, 2011). The Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report recommends less focus on probability beyond the basic rules with the possible exception of the binomial and normal distributions (GAISE, 2016). Woodard and McGowan (2012) redesigned an introductory statistics course according to the GAISE guidelines, completely eliminating probability topics except for the normal distribution, and including topics in sampling, summary statistics, confidence intervals, hypothesis testing, linear regression, and experiments. An analysis of 978 introductory statistics exam questions from ten different teachers found that 66.1 percent of questions inquired about topics in descriptive statistics, compared to only 15.2 percent and 18.7 percent for probability and inference respectively, indicating a tendency to focus on less cognitive demanding topics in general (Salcedo, 2014).

Constructivism is the main learning theory that has historically guided research in mathematics education (Garfield, 1995) while cognitive constructivism has recently become a focus in mathematics and science education (Tishkovskaya & Lancaster, 2012). Correcting students' errors during class helps students make the closest possible connections between their thoughts and the immediate feedback while they are still interested in learning and have the time to reflect (Garfield, 1995; Lovett & Greenhouse, 2000).

A lack of statistical literacy stems from focusing on the mechanical aspects of statistical knowledge rather than assessing students through their ability to interpret data (Tishkovskaya & Lancaster, 2012). According to the SOLO taxonomy, nearly two-thirds of 978 analyzed exam questions were classified as unistructural, the lowest level of cognitive thinking, while only four were at the highest level, extended abstraction (Salcedo, 2014). Furthermore, 83.2 percent of the exam questions tested statistical literacy, which requires only basic knowledge of terminology; the remaining 16.8 percent required higher order cognitive thinking in the form of statistical reasoning and statistical thinking that allowed students to demonstrate a deeper understanding of statistical processes, results, and investigation (Salcedo, 2014). While technology simplified the required calculations, there is little evidence that the technology itself improves student learning. In a study by Meletiou-Mavrotheris, Lee, and Fouladi (2007), technology provided students with additional practice in collecting and working with data, but student learning did not appear to improve with regards to concepts in statistical inference.

The GAISE College Report (2016) recommends that college instructors use active learning, allowing students to discover statistical concepts and engage in statistical thinking on their own by not listening to a lecture for the duration of class time. Active learning can take on several forms in the statistics classroom. Students may also work in small groups to solve problems while the instructor circulates the room to provide feedback (GAISE, 2016; Weltman & Whiteside, 2010; Woodward & McGowan, 2012). Despite the recommendation that students engage in active learning, group work may not the best means. Traditional statistics lectures were correlated with higher exam scores for students with higher grade point averages (GPAs); however, high performing students' exam scores dropped closer to the class mean when they worked with other students during class (Weltman & Whiteside, 2010). Conversely, students with lower GPAs scored better on exams when working in a group compared to listening to a lecture (Weltman & Whiteside, 2010). Instructors may have students explore statistical concepts through software packages, applets, and simulations, which allow for visualization of complex topics and direct interaction in addition to using a student response system for formative assessment (Chance et al., 2007; Garfield, 1995; Tishkovskaya & Lancaster, 2012).

Educational technology in statistics education.

The GAISE College Report recommends that instructors "use technology to explore concepts and analyze data" and to "perform most computations using technology to allow greater emphasis on understanding concepts and interpreting results" (GAISE, 2016, p. 16, 19). Technology has allowed students to focus on learning how to interpret graphs and computer outputs, which reduces their cognitive load (Chance et al., 2007; Salcedo, 2014). Course management systems have also enhanced statistics education in recent years with a wide range of tools such as discussion boards, online tutorials, and online quizzes and homework that can provide immediate feedback (Chance et al., 2007). Using a course management system can engage students using formative assessment by completing weekly multiple-choice quizzes and receiving feedback after the deadline (Hodgson & Pang, 2012). Over 90 percent of students believed that they could achieve greater understanding through the quizzes, and 82 percent said that they could judge how well they understood the material because they could see the results of each attempt (Hodgson & Pang, 2012). One drawback to this format was the inability to evaluate higher-order cognitive skills through only multiple-choice questions (Hodgson & Pang, 2012).

Critique of Previous Research

The theory of constructivism encourages active learning so that students can make their own connections between prior and new knowledge. The fact that Weltman and Whiteside (2010) found that students with lower GPAs tended to score lower when engaged in a workshop with other students suggests that group work harms higher achieving students by not challenging them enough. While active learning should be used to teach introductory statistics, several methods of active learning could be simultaneously integrated into the classroom that provide students with "the maximum amount of instructor expertise and direction" and assist the highest performing students in achieving more profound understanding (Weltman & Whiteside, 2010, p. 9). Combining a traditional lecture with formative assessments using a student response system may invoke the best of both worlds by engaging students at all levels while still providing a great deal of expertise from the instructor.

Attali et al. (2016) showed that the correlations between multiple-choice and open-ended questions are strong when structured with the same stem and requiring the same level of cognitive demand, supporting the notion that multiple-choice questions can be used as the sole means of assessment. Because the level of prior knowledge is more highly correlated with multiple-choice questions as shown by Ozuru et al. (2013), students who have taken a statistics class could act as a possible confounding variable.

Using a course management system similar to the method utilized by Hodgsen and Pang (2012) by providing students with feedback only after the deadline for the module has passed creates a way to both prevent cheating and offer students additional advice on a topic. Asking multiple-choice questions in the strategic thinking level of Webb's depth of knowledge in both formative and summative assessments could challenge the notion that multiple-choice questions cannot be used to test higher-order cognitive thinking.

Glass et al. (2008) conjectured that the additional study hypothesis could have led to the effectiveness of the formative assessments. Using different contexts in the repeated questions could have allowed the authors to test repeated concepts and analyze if having answered formative assessment questions led to a higher success rate on summative exams. This would have eliminated the possibility that students memorized factual information and were instead applying their knowledge of the concepts they learned via the formative assessments.

Chapter 3

Methodology

Overview

Assessment in education informs the instructor about how students are advancing in their learning. An analysis of students' individual responses can provide insight into how quickly they self-correct after receiving appropriate feedback. When formative assessment items test the same concept and require the same level of cognitive demand, the instructor can analyze the sequences of responses to identify when student knowledge has improved rather than examining each assessment independently. In total, 112 concepts in introductory statistics were identified and tested at three different time points throughout the semester.

Chapter 3 is organized as follows. First, the research questions are presented again along with the null hypotheses for each. A description of the setting, participants, and variables follows. Next, the instruments and responses to common threats to validity are presented. The experimental procedure, which describes the design and how the instruments were used, is outlined next. Threats to validity are discussed next. The chapter concludes with a description of the data collection process and methods by which the research questions were analyzed.

Research Questions and Hypotheses

The goal of this study was to analyze the sequences of correct and incorrect responses provided by students on two formative assessments that test students' knowledge of a statistical concept to ascertain if they predict the probability that the student will answer the corresponding summative assessment correctly. This study also examined the relationship between the sequences of responses provided on all three assessment items, the unit of the course, and the level of cognitive demand when the data was aggregated over all students partaking in the study. The two research questions that guided this study are:

1. How do the sequences of responses provided by students on formative assessment items affect the probability that the student will answer the related exam item correctly while controlling for:
    1) The unit of the course in which the concept is presented
    2) The level of cognitive demand required to answer the concepts' assessment items
    3) Students' previous statistics experience
    4) Students' interest in the course
    5) Students' interest in using a student response system
    6) Time of day
2. When responses are aggregated over all students, what is the relationship between:
    1) Students' sequences of responses to all three items testing a concept
    2) The unit of the course in which the concept is presented
    3) The level of cognitive demand required to answer the concepts' assessment items

The hypotheses for the first research question were:

H_0: The sequence of responses provided on formative assessment items will have no impact on the probability that a student will respond to the corresponding exam item correctly.

H_A: The sequence of responses provided on formative assessment items will impact the probability of answering the corresponding exam item correctly with the probabilities from lowest to highest in the following order:

1) Both formative assessment items incorrect
2) Student response system question correct, but quiz question incorrect
3) Quiz question correct, but student response system question incorrect
4) Both formative assessment items correct

The ordering presented in the alternative hypothesis stemmed from the expectation that students should demonstrate constant improvement. Failure to answer either formative assessment question correctly would demonstrate a lack of understanding while answering both correctly displayed a grasp of the concept. While the other two combinations each have one correct answer, students would show improvement when answering the quiz question correctly but not the student response system question. Conversely, they would regress when answering the student response system question correctly, but not the quiz question. Potential reasons for this regression could include randomly guessing correctly on the student response system question or obtaining the answer from their peers in class, neither of which would be conducive to assisting them on the quiz question.

The hypotheses for the second research question are:

H_0: The sequences of responses to the three assessment items on each concept, the unit of the course, and the level of cognitive demand are independent.

H_A: The sequences of responses to the three assessment items on each concept, the unit of the course, and the level of cognitive demand are not independent.

Research Setting

This study was conducted in three business statistics courses at a large, research-intensive university located in the northeast United States. During the 2017-2018 academic year, undergraduate enrollment was approximately 18,000 and full-time graduate enrollment was roughly 7,500 at the university. A strong majority (71%) of students identified as Caucasian. Moreover, 10% of students identified their ethnicity as Asian, 5% as African-American, 4% as Hispanic, and 4% as international.

The course was taught using a lecture format with student response system questions interspersed throughout the PowerPoint lectures. Poll Everywhere, a stable web-based audience response system, was used to gather student responses. This service not only allowed the instructor to generate student response system questions, but it also kept track of all responses provided by each student given that they created a free account.

Students were tested on each concept for the first time during class using a student response system question, which served as the first formative assessment. Each question was created by the instructor on the Poll Everywhere website by inputting the question, the correct answer, several distractors, and any necessary images. The questions were embedded in the lecture slides using the PowerPoint add-in. When each question appeared on the instructor's PowerPoint lecture slide during class, polling opened and students were typically allotted between one and two minutes to answer the question. Additional time was provided if a large number of students were still thinking. Students had the option of submitting their answer using the Poll Everywhere application on their cell phone, through their account on the Poll Everywhere website, or by texting the letter of their response to a code associated with the instructor's account.

After all students submitted their responses, the instructor displayed a bar graph of the results and provided feedback by explaining the logic behind the correct answer as well as why commonly chosen distractors were incorrect. The course consisted of 33 lectures, each lasting for 50 minutes. Each section contained approximately 80 students, the maximum enrollment capacity of the course. The lectures covered in each unit of the course are presented in Table 1. A list of the topics and dates on which they were covered is contained in Appendix A.

Table 1

Lectures to Be Covered in Each Unit

Unit	Lectures	Number of lectures
Data collection and descriptive statistics	1-8	8
Probability	9-16	8
One-sample inference	17-23	7
Two-sample inference	24-33	10

Approximately one week later, students took an online quiz that comprised the second level of formative assessment. Students completed each of the twelve quizzes on their own time outside of class using CourseWeb, the university's course management system. Each quiz opened at 10:00 AM on Tuesday and closed the following Monday at 11:59 PM regardless of the section time, giving students slightly less than one week to complete each.

Students completed a summative assessment at the end of each unit that covered material from the previous three to four weeks for the third response in the sequence. The first three unit exams were administered on the same day but at times that differed by a few hours. The final exam was administered according to the university's schedule, forcing the three sections to take the exam on three different days. Students had 50 minutes each to finish the first three unit exams and one hour and 50 minutes to complete the comprehensive final exam, which included the concepts from the two-sample inference unit of the course. Each student could use a calculator and create one sheet of handwritten notes to reference during the exam.

Research Method

The following sections will detail the sampling procedure for obtaining participants, the variables and instruments used in the study, and the potential threats and protections for internal and external validity.

Participants and Sampling

All participants in this study were enrolled in one of the researcher's three introductory business statistics courses. They were all undergraduate college students, most between the ages of 18 and 22. The instructor taught all three sections using the same set of lecture slides and assessments. Randomization was not possible because students self-selected into the available business statistics courses. Instead, students were recruited through voluntary sampling. All 240 enrolled students had the opportunity to participate in the study. To recruit subjects, the researcher's teaching assistants provided a description of the study during the final recitation of the semester, at which point students had completed all assessments for the final exam. The teaching assistants then explained that willing participants could volunteer to take part in the study. Students gave consent by signing and agreeing to the terms in the consent form, which allowed the instructor to analyze their responses to the student response system questions, quiz questions, and exam questions.

The sample was chosen from 240 potential subjects. According to the rosters of students who enrolled for one of the three sections, potential subjects were primarily freshmen (66%) and sophomores (27%), but a small percentage were juniors (5%) or seniors (2%). Of the 240 potential subjects, 64% were male; approximately 80% identified as Caucasian, 5% as Asian, and 3% as African-American. This was similar to the general makeup of the 2,100 students in the business school in which 60% of students were male and 78% were Caucasian. Slightly more than half of the potential participants had aspirations of enrolling in the business school and were required to take this business statistics course as their fourth prerequisite. However, the course also satisfied a requirement for many other programs so participants included students from several other majors in the arts and sciences.

Variables

The dependent variable in this study was a categorical variable denoting whether the student's response to each exam item is correct or incorrect.

Each student had the opportunity to respond to two formative assessment items for each of the 112 concepts: one student response system question and one quiz question. Each response was classified as being either correct or incorrect. The independent variable in this study was the sequence of responses students provided on the formative assessments for each concept. This was a categorical variable with four levels: both formative assessment items correct, the student response system question correct with the quiz question incorrect, the quiz question correct with the student response system question incorrect, and both formative assessment items incorrect. The four levels of the sequence of formative responses are defined in Table 2 below.

Table 2

Possible Sequences of Formative Assessment Responses

Combination	SRS question result	Quiz question result
1	Incorrect	Incorrect
2	Correct	Incorrect
3	Incorrect	Correct
4	Correct	Correct

In an attempt to account for potential influences on the dependent variable, this study controlled for the following factors:

- Unit of the course: Categorical variable describing the skills required by the concept to answer the question; assumed four levels: data collection and descriptive statistics, probability, one-sample inference, and two-sample inference
- Level of cognitive demand: Categorical variable describing how challenging the concept is according to Webb's depth of knowledge; assumed three levels: recall, basic application of skill, and strategic thinking
- Course section: Categorical variable describing which of the three sections the student was enrolled in; assumed three levels: 10:00 AM, 11:00 AM, or 2:00 PM
- Previous statistics course: Categorical variable describing if a student had ever taken a prior statistics course either in high school or college; assumed two levels: yes and no
- Course excitement level: Ordinal variable depicting a student's self-described excitement level for taking a statistics course at the beginning of the semester; assumed integer values from 1 through 6 with 1 being the lowest level of excitement and 6 being the highest level of excitement
- Student response system excitement level: Ordinal variable describing a student's self-described excitement level for using a student response system during class at the beginning of the semester; assumed integer values from 1 through 6 with 1 being the lowest level of excitement and 6 being the highest level of excitement

Instruments

There were four different types of instruments used in this study, each designed by the researcher and reviewed by a content expert in the field of statistics. Three of these instruments were assessments: student response system questions, quiz questions, and exam questions. The fourth instrument was a survey designed by the researcher that students filled out at the beginning of the semester to collect information about students' previous statistics experience, excitement for the course, and excitement for using a student response system.

All student response system questions, quiz questions, and exam questions were multiple-choice, which ensured that all students were graded identically in all aspects of the course. In total, the instructor presented 112 questions of each type throughout the semester for a total of 336 questions. Each of the 33 lectures contained between one and six student response system questions. There were 12 quizzes given throughout the semester, each covering between five and 13 concepts. The first exam on data collection and descriptive statistics and the third exam testing one-sample inference both consisted of 30 questions. Because the concepts tended to be more computationally intensive and time-consuming, there were only 20 questions on the probability exam. The final exam was required to be cumulative due to departmental rules, but only the 32 questions inquiring about two-sample inference topics were included in the analysis of this study. The other 28 questions on the final exam tested previously covered concepts previously but did not have corresponding formative assessment items to create additional sequences that could be analyzed in the scope of this study. All questions had at least three answer choices. At the suggestion of Haladyna et al. (2002), each question included as many plausible distractors as possible to reduce the probability of students blindly guessing the correct answer.

Validity

The following sections will discuss potential threats to internal validity and how they will be mitigated as well as how external validity will be maintained.

Internal validity.

This study did not have a control group because all subjects were exposed to the same treatments. As a result, diffusion of treatment, compensatory demoralization, and compensatory rivalry were not threats to internal validity in this study. Subjects' history of statistical knowledge could have posed a threat to internal validity. However, the survey that provided information regarding if subjects had taken a prior statistics course controlled for the fact that some students may have had additional knowledge of statistics coming into the course. Maturation did not pose a threat to internal validity from a physical standpoint because the study was conducted over a period of only four months.

As this course was an introductory course that serves as a prerequisite for additional statistics courses, subjects were not enrolled in another statistics course that could have advanced their knowledge of the course material. Regression and selection did not pose an internal threat to validity because most subjects were business school candidates and took the course because they did not place out of it through an earlier placement exam or other exemption. The risks involved in being in the study were minimal so recruitment of subjects was not hindered by potential risks. None of the 240 potential subjects withdrew from the course so mortality was not a threat to internal validity.

Although the same concepts were tested over the three time points, students' knowledge of the concepts was tested using different contexts. Students performed similar tasks on all three items in the sequence, but each item had a different scenario that used different data and parameters. As a result, students recognized how to approach each item using prior knowledge learned in the course; however, they had not seen the exact context of the quiz or exam question prior to taking it. Thus, the memorization of previous assessment items was impossible.

As this course was a requirement and taught using a lecture style, the potential for students to mindlessly choose an answer to the student response system question out of boredom or inattentiveness was a possibility. To control for mindless responding, students' self-described excitement levels for the course and for using a student response system were included as control variables. Thus, the threat to internal validity from testing was mitigated.

External validity.

The student response system and quiz questions in this study were tested during a pilot study conducted by the researcher in three business statistics courses taught during the fall 2017 semester. Students used a student response system in class to answer formative assessment questions. The class averages on the student response system questions were compared with the average quiz scores and homework scores for the corresponding lectures. The student response system question averages were significantly and positively correlated with the average quiz scores, but unrelated to students' average homework scores. Between semesters, the instructor adjusted the wording and modified the distractors on a few questions, but a large majority of the questions remained unaltered. The quiz questions used in this study appeared as exam questions in the courses taught by the researcher during the pilot study. Students who took the course in the previous semester were not permitted to keep their exams so there was limited risk in the questions or solutions being passed on to students taking the course while the study was being performed in the spring 2018 semester. Difficulty indices, presented in Table 3, demonstrate that the student response system and quiz questions were an appropriate level of difficulty.

Table 3

Difficulty Indices for Student Response System and Quiz Assessment Items

Unit	Student response system	Quiz
Data collection and descriptive statistics	.7791	.8415
Probability	.7240	.6475
One-sample inference	.6720	.7969
Two-sample inference	.6974	.7685

All exam questions used in this study were new assessment items that mimicked the formatting of the student response system and quiz questions. A content expert in the field of statistics deemed them to be a similar level of difficulty as the corresponding formative assessment items. Grading was consistent across all students because each assessment item is of the objective format, allowing all questions to be classified as either correct or incorrect.

The Kuder-Richardson 20 (KR20) score was reported to test for reliability on exams. The KR20 score was provided on the report generated by the testing center at the university and was an appropriate measure of reliability because of the dichotomous nature of the response. Responses to the student response system questions from the fall 2017 semester were aggregated by the unit of the course. The KR20 scores were calculated to ensure reliability and are reported in Table 4. Because the quiz questions in this study were asked as exam questions in the previous semester, the KR20 score for each of the unit exams is also reported in Table 4 below as confirmation of reliability for the quiz questions. A content expert in statistics verified that the sequences were testing the same concept and were assigned the correct assigned level of cognitive demand (Secolsky & Denison, 2012). The sequences and their corresponding course units and levels of cognitive demand and defined in Appendix B.

Table 4

KR20 Scores for Student Response System and Quiz Assessment Items

Unit	Student response system	Quiz
Data collection and descriptive statistics	.7884	.5657
Probability	.7891	.6569
One-sample inference	.7967	.8057
Two-sample inference	.7941	.8384

The instruments did not alter the typical classroom setting as the researcher integrated the student response system questions into the PowerPoint lecture slides. These questions would have been used as formative assessment regardless of if the study had been performed. The quizzes were taken outside of regular classroom hours on students' own time and exams were given as scheduled on the course calendar. The students' teaching assistant administered the survey during the weekly recitation and took less than five minutes to complete.

Experiment Structure

This study employed both a predictive and correlational quasi-experimental design where each subject had the opportunity to answer all three questions on each of the 112 concepts in introductory statistics at three different time points. Each subject was exposed to the same set of experimental conditions for several reasons. The researcher taught three sections of the same course, each of which met on the same three days of the week for 50 minutes each. Students occasionally attended the class at a different time if another conflict arose in their schedule. Teaching the sections differently was not an option as it would have either created another confounding variable that could not have been easily controlled for due to exposure of two dissimilar learning environments. Moreover, students would not have been able to attend a different section, which would have been detrimental to their learning. Similarly, subjects enrolled in one section were familiar with subjects enrolled in other sections so discussion of the course material outside of class could have created another confounding variable. The section of the course was included as a variable to control for any differences in the characteristics of students or for a possible inadvertent teaching effect. Because a majority of students were competing for admission to the business school, it would have been unethical to teach the sections of the course differently.

Assessment items.

Students answered the student response questions, the first level of formative assessment, using the web-based service Poll Everywhere. The instructor asked between five and eight questions throughout each lecture that tested students' knowledge of concepts that were recently covered; however, not all questions were part of one of the 112 sequences. Students were not aware of the concepts included on the second level of formative assessment or the summative assessment. Students earned one point of extra credit for each question answered to incentivize responding and a second point for answering correctly to prevent students from randomly choosing an answer simply to earn points. The 112 student response system questions used in the sequences are contained in Appendix C.

Twelve quizzes, which comprised the second level of formative assessment, were offered online within the university's course management system. Each quiz consisted of between five and 13 questions, each corresponding directly to a concept tested in a student response system question from the previous week. Table 5 summarizes the unit, lectures covered, and the number of questions on each quiz. All quizzes opened on a Tuesday morning at 10:00 AM and were due the following Monday night at 11:59 PM, giving students slightly less than one week to complete them. Students could use the entire week to complete each quiz, even doing so in multiple sessions, and were permitted to use their notes.

Table 5

Summary of Quizzes

Quiz	Unit	Lectures covered	Number of questions
1	Data collection/descriptive statistics	1-3	6
2	Data collection/descriptive statistics	4-5	11
3	Data collection/descriptive statistics	6-8	13
4	Probability	9-11	9
5	Probability	12-14	6
6	Probability	15-16	5
7	One-sample inference	17-19	10
8	One-sample inference	20-22	12
9	One-sample inference	22-23	8
10	Two-sample inference	24-25	9
11	Two-sample inference	26-29	10
12	Two-sample inference	30-33	13

To reduce the potential for cheating, only after the deadline for the quiz has passed were students able to review their answers, the correct answers, and the feedback provided for each question by the instructor. The course management system graded the quizzes automatically. Because all questions were multiple-choice, personalized feedback was not necessary like it would have been for open-ended questions. All students saw the same feedback for all questions regardless of if they answered correctly or incorrectly. The twelve quizzes given throughout the semester are presented in Appendix D.

The summative assessments were timed, in-class unit exams where each question corresponded to an earlier student response system question and quiz question. Students took each of the first three unit exams on the same day but at different times according to their class schedule. The university determined dates and times for the final exam according to the days of the week and time the class met. Students in the 10:00 AM, 11:00 AM, and 2:00 PM sections took their final exam on Thursday, Saturday, and Friday morning respectively. All three assessment items tested the same statistical concept. Students created one handwritten sheet of notes to reference during each exam. Students recorded their exam responses on Scantron answer sheets that were graded by the university's testing center. There were four summative exams given throughout the semester, each covering one course unit. A summary of the lectures covered and the number of questions on each exam is contained in Table 6.

Table 6

Summary of Exams

Exam	Unit	Lectures covered	Number of questions
1	Data collection/descriptive statistics	1-8	30
2	Probability	9-16	20
3	One-sample inference	17-23	30
4	Two-sample inference	24-33	60 (32 used in study)

Each of the first three exams, covering data collection and descriptive statistics, probability, and one-sample inference, lasted for 50 minutes. Students had one hour and 50 minutes to complete the cumulative final exam that covered the concepts from two-sample inference in addition to questions testing concepts taught earlier in the semester. However, only questions pertaining to the two-sample inference unit were included in the analysis of this study because the remaining items on the final exam did not have corresponding formative assessment items. The summative exams are contained in Appendix E.

Classification of concepts.

Webb's depth of knowledge was used to classify questions in each sequence according to the level of cognitive demand required by students to answer the question. Because each sequence tested the same concept, all three questions in the sequence were developed to test the same level of cognitive demand. Questions that required students to apply a definition or remember a fact were classified at the recall level. Problems that required direct calculations or application of a rule were classified at the basic application of skill level. Strategic thinking questions were those that required either complex thinking and reasoning or a multistep process to arrive at a computational answer. None of the questions required a cognitive demand of extended thinking because questions at this level are often time-consuming to answer and require research, making them unsuitable for objective-format type questions or as formative assessment during a 50-minute class. A numbered list of the 112 sequences with the three corresponding items for each as well as their level of cognitive demand according to Webb's depth of knowledge is contained in Appendix B.

Control variables.

During the first week of the semester, students filled out a three-question survey. On this survey, students disclosed if they had taken a prior statistics course, rated their level of excitement in taking a statistics course on a Likert scale from 1 to 6, and rated their level of excitement in using a student response system in class on a Likert scale from 1 to 6. The assigned teaching assistant for the section was in charge of distributing and collecting the surveys as well as keeping them in their possession until grades were finalized. The survey is presented in Appendix F.

Data Collection

Data was collected in three separate phases. Results from the student response system questions were downloaded from the Poll Everywhere website while responses to the quiz questions were downloaded from the university's course management system. The university's testing center graded exams electronically and provided a spreadsheet of all responses to the researcher. For each student and concept, the responses to the student response system, quiz, and exam questions were recorded along with an identification number for the student, the unit of the course, and the level of cognitive demand. All identifying factors were removed from the data so the researcher could not associate student names with observations. Any observation where at least of the three assessment items went unanswered was removed from the dataset. These observations were used in a post-hoc analysis comparing complete and incomplete sequences.

Data Analysis

Several different statistical models were created to assess the research questions. The following sections detail the analyses that were performed to assess each research question.

Modeling participant performance on formative and summative assessments.

To analyze the first research question, the following logistic regression model was generated that predicted the probability that each summative assessment item would be answered correctly:

$$\text{logit}(p_i) = \beta_0 + \sum_{j=1}^{3} \gamma_1^{(j)} Z_{1i}^{(j)} + \sum_{k=1}^{3} \gamma_2^{(k)} Z_{2i}^{(k)} + \sum_{l=1}^{2} \gamma_3^{(l)} Z_{3i}^{(l)} + \sum_{m=1}^{2} \gamma_4^{(m)} Z_{4i}^{(m)} + \gamma_5 Z_{5i} + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_{ijklm} \qquad (1)$$

The dependent and independent variables in equation (1) are defined as follows:

- p_i: Predicted probability that the exam question for the ith observation was answered correctly
- $Z_{1i}^{(1)}$: 1 if the ith observation resulted in a sequence where both formative assessment items were answered correctly; 0 otherwise
- $Z_{1i}^{(2)}$: 1 if the ith observation resulted in a sequence where the quiz question was answered correctly, but the student response system question was answered incorrectly; 0 otherwise

- $Z_{1i}^{(3)}$: 1 if the ith observation resulted in a sequence where the student response system question was answered correctly, but the quiz question was answered incorrectly; 0 otherwise

The control variables used in this study that are expressed in equation (1) are defined as follows:

- $Z_{2i}^{(1)}$: 1 if the ith observation was from the probability unit of the course; 0 otherwise
- $Z_{2i}^{(2)}$: 1 if the ith observation was from the one-sample inference unit of the course; 0 otherwise
- $Z_{2i}^{(3)}$: 1 if the ith observation was from the two-sample inference unit of the course; 0 otherwise
- $Z_{3i}^{(1)}$: 1 if the ith observation required a level of cognitive demand of basic application of skills; 0 otherwise
- $Z_{3i}^{(2)}$: 1 if the ith observation required a level of cognitive demand of strategic thinking; 0 otherwise
- $Z_{4i}^{(1)}$: 1 if the student for the ith observation was in the 11:00 AM section taught by the researcher; 0 otherwise
- $Z_{4i}^{(2)}$: 1 if the student for the ith observation was in the 2:00 PM section taught by the researcher; 0 otherwise
- Z_{5i}: 1 if the student was previously enrolled in any type of statistics course; 0 otherwise
- X_{1i}: Student's self-described excitement level to be taking a statistics course rated on a scale from 1 to 6

- X_{2i}: Student's self-described excitement level to be using a student response system in class rated on a scale from 1 to 6

After the model in equation (1) was generated and analyzed using the statistical software package Minitab, the researcher used a script written in Python to analyze individual student performance through a Monte Carlo simulation. Equation (1) returned the predicted probability that each exam question would be answered correctly. For each observation in the dataset, a random number from a continuous uniform distribution on the interval $[0,1]$ was generated. These numbers were compared against the predicted probability of answering the corresponding exam question correctly and classified according to the following rule:

$$q_i = \begin{cases} 1 \text{ if } u_i \le \hat{p}_i \\ 0 \text{ if } u_i > \hat{p}_i \end{cases} \tag{2}$$

In the above expression, q_i is the predicted result of observation i with 1 representing a correct answer and 0 representing an incorrect answer, u_i is the uniform random number generated for observation i, and $\hat{p}_i$ is the predicted probability of answering the exam question in observation i correctly according to the probabilities obtained from equation (1). For each student, the simulated exam results were summed over all 112 questions, providing a simulated total out of 112. This process was repeated 10,000 times, a Monte Carlo simulation sample size deemed more than sufficient (Mundform, Schaffer, Kim, Shaw, & Thongteeraparp, 2011).

Based on the results of the simulation, the researcher calculated a 95% confidence interval for each subject's total number of exam questions correct by eliminating the largest 2.5% and smallest 2.5% of the 10,000 simulated results. For each subject in the study, the confidence intervals were compared with the actual number of questions answered correctly. Students were then classified according to their actual performance against expectation: better than expected, as well as expected, or worse than expected on the exam questions. The sequences of formative assessment responses were then aggregated in a cross-classification table against performance against expectation. A chi-square test for independence was then performed to identify any relationship between students' overall performance relative to expectation on the exams and the sequences of correct and incorrect answers on the formative assessments.

Modeling relationship between cognitive demand, course unit, and sequence of responses.

The second research question was analyzed by building a series of loglinear models, all performed using the statistical software package Minitab. The sequence of assessment responses is a categorical variable with eight possible combinations, displayed in Table 7 below. The unit of the course is a categorical variable with four categories: data collection and descriptive statistics, probability, one-sample inference, and two-sample inference. The level of cognitive demand is a categorical variable with three categories: recall, basic application of skill, and strategic thinking. Each complete sequence was assigned to the corresponding unit of the course and level of cognitive demand. All observations were then aggregated in a $4 \times 3 \times 8$ cross-classification table that compared the unit of the course, the level of cognitive demand, and the sequences of formative and summative assessment responses defined in Table 7.

Table 7

Possible Sequences of Responses to Assessment Items

Sequence	SRS question result	Quiz question result	Exam question result
1	Incorrect	Incorrect	Incorrect
2	Correct	Incorrect	Incorrect
3	Incorrect	Correct	Incorrect
4	Correct	Correct	Incorrect
5	Incorrect	Incorrect	Correct
6	Correct	Incorrect	Correct
7	Incorrect	Correct	Correct
8	Correct	Correct	Correct

The deviance statistic was used to determine lack of fit in each of the loglinear models to determine the model with the best fit. In the following loglinear models, U denotes the unit of the course, C denotes the level of cognitive demand, and S denotes the sequence of assessment responses. The following model was first to test for mutual independence of all three variables:

$$\log(\mu_{ijk}) = \lambda + \lambda_i^U + \lambda_j^C + \lambda_k^S \tag{3}$$

If the model in equation (3) exhibited lack of fit, then the model in equation (4) that includes all second order interaction was built:

$$\log(\mu_{ijk}) = \lambda + \lambda_i^U + \lambda_j^C + \lambda_k^S + \lambda_{ij}^{UC} + \lambda_{ik}^{US} + \lambda_{jk}^{CS} \tag{4}$$

If the model in equation (4) did not exhibit lack of fit, then combinations of second-order terms were removed from the model systematically to determine which variables were conditionally independent before settling on a final model with all necessary second-order terms that fits the data.

If the model in equation (4) also exhibited lack of fit, then the saturated model in equation (5) that included the third-order interaction term was built and considered for the final model:

$$\log(\mu_{ijk}) = \lambda + \lambda_i^U + \lambda_j^C + \lambda_k^S + \lambda_{ij}^{UC} + \lambda_{ik}^{US} + \lambda_{jk}^{CS} + \lambda_{ijk}^{UCS} \tag{5}$$

Due to the large sample size, each possible model was checked for practical significance by comparing the odds ratios of the first order interaction terms at all levels of the third variable. If these odds ratios were close, then statistical significance may have resulted from the large sample size rather than practical significance. This indicated that a simpler model was adequate for describing the relationship between the sequence of responses, the unit of the course, and the level of cognitive demand. Pearson residuals were also analyzed to determine if a simpler model was adequate for describing the relationship between these variables.

Chapter 4

Results

Overview

The results of the semester-long study into the sequences of formative assessment responses provided by students and their relationship with the corresponding summative assessment responses are presented in this chapter. The chapter is organized beginning with a description of the sample and a summary of the descriptive statistics for the formative assessment responses, broken down by sequence, unit of the course, and level of cognitive demand. The results of the logistic regression model and Monte Carlo simulation used to analyze individual participant performance are presented next. The chapter ends with a conclusion containing the results of the loglinear model used to analyze the aggregated responses of all participants.

Description of the Sample Participating in the Study

This section contains a description of the participants whose responses to the assessment questions were analyzed in the two research questions in this study. Participants in this study were students at a large, research-intensive university during the spring 2018 semester and were enrolled in one of the researcher's three business statistics courses. In total, 240 students had the opportunity to participate in the study, of which 145 (60.41%) consented. Participation rates were nearly identical across all three sections of the course. However, the sequence completion rate was slightly lower for the 10:00 AM section than the 11:00 AM and 2:00 PM sections. A summary of the number of students from each section who agreed to participate, the total number enrolled, and information regarding the total number of complete sequences from each section is contained in Table 8.

Table 8

Breakdown of Study Participation by Section

Section	Participating	Enrolled	Participation rate	Complete sequences	Sequence completion rate
10:00 AM	47	79	59.49%	4,438	50.16%
11:00 AM	49	81	60.49%	4,920	54.23%
2:00 PM	49	80	61.25%	4,764	53.17%
Total	**145**	**240**	**60.42%**	**14,122**	**52.54%**

At the beginning of the semester, participants filled out a survey regarding their previous statistics experience, their excitement level for the course, and their excitement level for using a student response system in the course. The survey presented participants with a six-point Likert scale to gauge both excitement levels where 1 represented being extremely unexcited and 6 represented being extremely excited. These acted as control variables for the first research question, which explored whether the sequences of responses provided on formative assessment items affected the probability of answering the related exam item correctly while controlling for six different factors: the unit of the course, the level of cognitive demand, students' previous statistics experience, students' interest in the course, students' interest in using a student response system, and the time of day.

Based on the survey results, participants in the study were fairly diverse in their statistics backgrounds and excitement levels for the course and for the student response system. Participants were slightly more likely to have not taken a previous statistics course (51.72%) than to have taken one (48.28%). However, the proportion of students who had taken a statistics course did not differ significantly from 0.50 according to the one-sample proportion test (Z = 0.42, p = 0.678), suggesting a negligible difference in the composition of these groups. Participants also exhibited a moderate excitement level in both the course (Mean: 4.03; SD: 0.89) and for using a student response system (Mean: 4.20; SD: 0.85) with the student response system being of slightly more interest. Few students expressed excitement levels at the extreme ends of the spectrum for either question. Breakdowns of the responses to the Likert scale questions are provided in Appendix G.

Summary of Participant Performance on Formative and Summative Assessments

This section contains a summary of the sequences of responses provided by students on the formative and summative assessments in compared to the unit of the course and level of cognitive demand as well as information regarding the reliability of the study. Had all 145 participants answered all three questions for all 112 concepts, there would have been 16,240 complete sequences. After removing incomplete sequences where at least one of the three responses was missing, 14,122 complete sequences remained for the final analysis.

The percentage of completed sequences declined slightly as the semester progressed. In the descriptive statistics unit covered during the first three weeks, 91.06% of sequences were completed. This percentage dropped to 87.24% in the probability unit, 82.48% in the one-sample inference unit, and 87.13% in the two-sample inference unit. The four concepts presented during lecture 19 in the one-sample inference unit was taught on the last day before spring break began. Student response system questions on this day received approximately half as many responses as usual, resulting in the loss of 259 complete sequences.

Participants were slightly less likely to complete a sequence with a higher level of cognitive demand (i.e. strategic thinking). Among all incomplete sequences, 87.77% resulted from a missing response to only the student response system question, whereas 9.60% of sequences were incomplete because of a missing quiz question only and 2.63% of sequences where both formative assessment responses were missing. Appendix H contains summaries of the sequence completion percentage by unit and cognitive demand and a summary of the missing formative assessment that caused each incomplete sequence.

Overall, participants answered 84.92% of exam questions correctly. Participants responded to both formative assessment questions correctly in a majority (58.72%) of sequences. On 22.65% of sequences, participants corrected a mistake on the student response system question and answered the quiz question correctly. Conversely, participants regressed by missing the quiz question after answering the student response system question correctly on 10.85% of sequences. They missed both formative assessment items on 7.78% of sequences. A breakdown by the sequence of formative assessment responses and the corresponding conditional percentages for each group is contained in Table 9.

Table 9

Breakdown of Exam Response by Formative Assessment Sequence

	Exam response result		
Formative sequence[1]	Correct (%)	Incorrect (%)	Total (%)
Incorrect-incorrect	815 (74.16)	284 (25.84)	1,099 (7.78)
Correct-incorrect	1,183 (77.22)	349 (22.78)	1,532 (10.85)
Incorrect-correct	2,617 (81.81)	582 (18.19)	3,199 (22.65)
Correct-correct	7,378 (88.98)	914 (11.02)	8,292 (58.72)
Total	**11,993 (84.92)**	**2,129 (15.08)**	**14,122 (100.00)**

Note. [1] First response: student response system question; second response: quiz question

In general, the percentage of correct exam responses increased as performance on the formative assessment improved. When both formative assessment items were missed, participants answered less than three-quarters of exam questions correctly; conversely, they answered nearly 89% of summative assessment items correctly when both formative assessment items were answered correctly. In situations where exactly one formative assessment item was correct, the quiz question was more indicative of success on the summative assessment (81.81%) compared to the student response system question (77.22%).

Aggregating the summative assessment responses across the units of the course, participants performed the best on concepts related to inference, answering nearly 88% of exam questions correctly in each unit. Participants struggled the most in the probability unit, answering slightly more than three-quarters of exam questions correctly. Exam success in the descriptive statistics unit was better than the probability unit but worse than both inference units. If the cognitive demand was lower, participants tended to respond correctly to the exam question more frequently. Despite the increase in difficulty, participants had only slightly less success on exam questions rated as basic application of skill (87.34%) compared to those classified as recall (90.35%); however, they were twice as likely to miss the exam question on questions requiring strategic thinking (75.39%). Summaries of exam responses by course unit and level of cognitive demand are presented in Appendix I.

To ensure the questions were reliable, the KR20 score was used as each question could be classified dichotomously as being either correct or incorrect. The KR20 score was calculated in two different ways: once for the questions in each unit of the course across for student response system, quiz, and exam questions, and once across the three questions in each sequence. For each of the three types of assessments, the KR20 scores on the questions in each unit were above 0.74, exceeding the widely accepted cutoff score of 0.70 (Fraenkel, Wallen, & Hyun, 2012). Thus, there was high reliability of assessment items within the same unit of the course.

Reliability scores for the 112 three-question sequences were moderately high (Mean: 0.58, *SD*: 0.18) across the board, indicating that the questions asked within each sequence produced relatively consistent results. When broken down by unit, KR20 scores were slightly lower in descriptive statistics (Mean: 0.5821) and probability (Mean: 0.4911) than in inference. These are consistent with the KR20 scores observed when the reliability was calculated based on the concepts presented in each unit rather than the sequence. Summaries of KR20 scores by assessment type and the by unit of the course are contained in Appendix J.

Influence of Formative Assessment Response Sequences

The first research question involved analyzing the effect of the sequences of formative assessment responses on the probability that the corresponding summative assessment item was answered correctly. It was assessed using two different methods. First, the logistic regression model defined in equation (1) was built to analyze the following hypotheses:

H_0: The sequence of responses provided on formative assessment items will have no impact on the probability that a student will respond to the corresponding exam item correctly.

H_A: The sequence of responses provided on formative assessment items will impact the probability of answering the corresponding exam item correctly with the probabilities from lowest to highest in the following order:

1) Both formative assessment items incorrect
2) Student response system question correct, but quiz question incorrect
3) Quiz question correct, but student response system question incorrect
4) Both formative assessment items correct

Second, a 95% confidence interval for the expected number of summative assessment questions each student would have been expected to answer correctly was calculated using a Monte Carlo simulation based on the predicted probabilities generated from the logistic regression model. These confidence intervals were compared against the actual number of questions answered correctly to determine students who underachieved, overachieved, or performed as expected.

The logistic regression model defined in equation (1) was statistically significant based on the deviance statistic (χ^2 = 746.59, p < 0.001, r^2 = 0.062), indicating that at least one of the independent or control variables was statistically significant. The assumptions for the logistic regression model were satisfied as the response was binary, the observations were independent, and no issues arose with collinearity between the continuous predictors. To further assess the effect of each variable, the individual coefficients, presented in Table 10, were analyzed.

Table 10

Coefficients and Statistical Tests for Formative Assessment Sequences and Control Variables Used to Predict Success on Summative Assessment

Variable	*Coefficient*	*SE*	*Z*	*p*	95% CI
Constant	1.533	0.175	8.74	0.000	(1.189, 1.876)
Sequence[1]					
Correct-incorrect	0.1072	0.0948	1.13	0.258	(-0.079, 0.293)
Incorrect-correct	0.3205	0.0851	3.77	0.000	(0.159, 0.487)
Correct-correct	0.8134	0.0807	10.09	0.000	(0.655, 0.972)
Unit					
Probability	-0.2438	0.0672	-3.63	0.000	(-0.375, -0.112)
One-sample inference	0.5296	0.0697	7.60	0.000	(0.393, 0.666)
Two-sample inference	0.4160	0.0671	6.19	0.000	(0.284, 0.548)
Cognitive demand					
Basic application of skill	-0.3577	0.0674	-5.31	0.000	(-0.490, -0.226)
Strategic thinking	-1.0611	0.0674	-15.75	0.000	(-1.193, -0.929)
Previous stat course					
Yes	0.1112	0.0491	2.26	0.024	(0.015, 0.208)
Course excitement	0.1120	0.0293	3.83	0.000	(0.055, 0.169)
Clicker excitement	-0.0942	0.0298	-3.16	0.002	(-0.152, -0.036)
Time					
11:00 AM	-0.0967	0.0605	-1.60	0.110	(-0.215, 0.021)
2:00 PM	-0.1829	0.0608	-3.01	0.003	(-0.302, -0.064)

Note. [1] First response: student response system question; second response: quiz question. Incorrect-incorrect used as baseline

The baseline category for the sequence of formative assessment responses was when both formative assessment items were answered incorrectly. As participants answered more formative assessment questions correctly, the probability that they answered the corresponding summative assessment item in the sequence correctly increased in the order suggested by the alternative hypothesis. Answering only the quiz question correctly ($Z = 3.77$, $p < 0.001$) and answering both formative assessment items correctly ($Z = 10.09$, $p < 0.001$) both increased the probability of responding to the exam question correctly over the baseline. Participants who answered only the student response system question correctly were more likely to answer the summative assessment question correctly compared to the baseline ($Z = 1.13$, $p = 0.258$), but the effect was not statistically significant.

Participants who answered both formative assessment questions correctly were more than twice as likely to answer the summative assessment question correctly than students who missed the quiz question, regardless of if they answered the student response system question correctly. Although participants who answered only the quiz question correctly performed better than those who did not, they were still significantly more likely to miss the summative assessment item than those who answered both formative assessment questions correctly. Participants who answered only the quiz question correctly were significantly more likely to answer the summative assessment item correctly than those who only answered the student response system question correctly, which was suggested in the alternative hypothesis. Answering only the student response system item correctly did not significantly improve students' abilities to answer the summative assessment item correctly, which was the only combination of sequences that were not significantly different. Odds ratios, which measure how much more likely a student is to answer the summative assessment item given the one sequence of formative assessment responses compared to a different sequence, are presented in Table 11.

Table 11

Odds Ratios Comparing Each Sequence of Formative Assessment Responses

First level[1]	Second level[1]	*Odds ratio*	*95% CI*
2	1	1.1132	(0.9244, 1.3405)
3	1	1.3778	(1.1662, 1.6278)
4	1	2.2556	(1.9258, 2.6419)
3	2	1.2377	(1.0614, 1.4434)
4	2	2.0263	(1.7588, 2.3345)
4	3	1.6371	(1.4562, 1.8404)

Note. [1] Formative assessment response sequence: 1 – Both formative assessment items incorrect; 2 – Student response system question correct, but quiz question incorrect; 3 – Student response system question incorrect, but quiz question correct; 4 – Both formative assessment items correct

Influence of unit of the course on summative assessment responses.

Participants were significantly less likely to answer the summative assessment question in the probability unit of the course than in any of the other three units (descriptive statistics, one-sample inference, or two-sample inference). Similarly, they were less likely to answer the summative assessment item correctly in the descriptive statistics unit than in either inference unit. Participants were slightly more likely to answer the summative assessment item correctly in the one-sample inference unit than the two-sample inference unit, but the difference was not statistically significant. The effect of the unit of the course was statistically significant for five of the six combinations as displayed in Table 12.

Table 12

Odds Ratios Comparing Effect of Course Units

First level	Second level	Odds ratio	95% CI
Probability	Descriptive statistics	0.7837	(0.6870, 0.8939)
One-sample inference	Descriptive statistics	1.6983	(1.4813, 1.9469)
Two-sample inference	Descriptive statistics	1.5158	(1.3289, 1.7290)
One-sample inference	Probability	2.1671	(1.8822, 2.4950)
Two-sample inference	Probability	1.9343	(1.6883, 2.2162)
Two-sample inference	One-sample inference	0.8926	(0.7753, 1.0276)

Influence of level of cognitive demand on summative assessment responses.

The level of cognitive demand was highly significant in predicting if a participant would answer the exam question correctly. In general, as the level of cognitive demand for the concept increased (i.e. recall to basic application of skill to strategic thinking), participants were less likely to answer the summative assessment item correctly in the sequence. There was a smaller effect between the recall and basic application of skill levels (Odds ratio: 1.4301, 95% CI: [1.2530, 1.6321]) than between the basic application of skill and strategic thinking levels (Odds ratio: 2.0206, 95% CI: [1.8126, 2.2523]). Students were nearly three times more likely to answer an exam question correctly at the recall level than at the strategic thinking level (Odds ratio: 2.8893, 95% CI: [2.5323, 3.2971]).

Analysis of control variables from pre-semester survey questions.

Participants who had taken a previous statistics course were slightly more likely to answer the summative assessment item correctly (Odds ratio: 1.1176, 95% CI: [1.0150, 1.2306]). However, this effect is likely primarily due to the large sample size and is not practically significant considering the lower bound of the interval lies just above 1. The effects of participants' excitement for the course ($Z = 3.83$, $p < 0.001$) and for using a student response system ($Z = -3.16$, $p = 0.002$) were both statistically significant. Participants who were more excited to take the course were significantly more likely to answer the exam question correctly. While the effect of participants' excitement for using the student response system was significant, those who were more excited to use a student response system were significantly less likely to answer the summative assessment item correctly.

Influence of class time on summative assessment responses.

Evidence of a slight teaching effect due to the class time did surface. Participants enrolled in the 11:00 AM class did not demonstrate any significant difference in the probability of answering the exam item correctly compared to either the 10:00 AM (Odds ratio: 1.1014, 95% CI: [0.97854, 1.2401]) or 2:00 PM (Odds ratio: 1.0900, 95% CI: [0.9713, 1.2232]) classes. However, participants in the 10:00 AM class were significantly more likely to answer the summative assessment item correctly than those in the 2:00 PM class (Odds ratio: 1.2006, 95% CI: [1.0656, 1.3526]), although this may be a matter of statistical significance rather than practical significance given the large sample sizes in each class and how close the lower bound of the confidence interval is to 1.

Participant performance relative to expectations.

Using the predicted probabilities from the model in equation (1), 10,000 sets of exam responses were simulated using a Monte Carlo simulation to generate confidence intervals for the number of summative assessment questions each participant would have been expected to answer correctly. Because incomplete sequences were removed from the dataset, each participant's confidence interval estimated the number of correct exam responses out of the total number of sequences they completed personally. The table summarizing the number of sequences each participant completed, the number of summative assessment questions they answered correctly, the predicted number of correct exam answers, and a 95% confidence interval for the predicted number of correct summative assessment items for each participant is contained in Appendix K.

Of the 145 participants who participated in the study, 109 (75.17%) performed as expected with their actual number of correct answers on the summative assessments falling inside their corresponding 95% confidence interval. Fifteen participants (10.35%) performed significantly worse than expected with their actual number of correct answers falling below the lower bound of the confidence interval while the remaining 21 (14.48%) participants overachieved by answering more questions correctly than expected. On average, participants who answered more questions correctly than expected completed more sequences than students who answered as many questions as expected. Similarly, those participants who underachieved completed fewer sequences than those who fell inside of their confidence intervals. A summary of the three groups is provided in Table 13.

Table 13

Summary Statistics of Number of Complete Sequences According to Performance Relative to Simulated Confidence Interval

Group	*Mean*	*SD*	Sample size
Above	106.29	9.08	21
In	96.86	15.87	109
Below	88.80	17.19	15

A post-hoc analysis was performed to study the association between the number of completed sequences and participant performance relative to expectations. A one-way analysis of variance (ANOVA) tested if the number of completed sequences by students who overachieved, underachieved, and performed as expected were equal. According to the ANOVA test, the three population means were not all equal ($F = 6.023$, $p = 0.003$).

Fisher's Least Significant Difference (LSD) method of multiple comparisons was used in conjunction with a one-way ANOVA analysis to determine specific pairwise differences in population means while maintaining a family error rate for situations when three groups are compared (Curran-Everett, 2000). A further analysis of the number of complete sequences using Fisher's LSD method of multiple comparisons found that participants who performed better than expected completed significantly more sequences than both participants who performed as expected ($t = 2.59$, $p = 0.010$, 95% CI: [2.24, 16.60]) and participants who underachieved ($t = 3.39$, $p = 0.001$, 95% CI: [7.30, 27.67]). There was a moderate significant difference in the number of completed sequences between participants who performed as expected and underperformed ($t = 1.92$, $p = 0.057$, 95% CI: [-0.24, 16.36]).

Aggregating the sequences of formative assessment responses over all participants according to their performance relative to their confidence interval returns the results in Appendix L. A chi-square test for independence found a highly significant relationship (χ^2(6, n = 14,112) = 185.27, $p < 0.001$) between participants' performance relative to their confidence interval and the sequences of formative assessment responses. Participants who answered both formative assessment questions correctly were much more likely to perform better than expected. Alternatively, participants who missed the quiz question, regardless of their success on the student response system question, were much more likely to underperform expectations. For participants who performed as well as expected, the counts for all four sequences of formative responses were all close to their expected counts. The actual counts, expected counts, and chi-square components relating the sequence of formative assessment responses and performance relative to expectations are presented in Table 14.

Table 14

Cross-Classification Table of Formative Assessment Sequences by Performance Relative to Confidence Interval

		Sequence[1]				
Performance	Count	1	2	3	4	**Total**
Above	Actual	91.00	195.00	379.00	1,567.00	**2,232**
	Expected	173.70	242.10	505.60	1,310.60	
	χ^2	39.37	9.16	31.70	50.16	
In	Actual	872.00	1,179.00	2,446.00	6,061.00	**10,558**
	Expected	821.60	1,145.40	2,391.70	6,199.30	
	χ^2	3.09	0.99	1.23	3.09	
Below	Actual	136.00	158.00	374.00	664.00	**1,332**
	Expected	103.70	144.50	301.70	782.10	
	χ^2	10.06	1.26	17.33	17.83	
Total		**1,099.00**	**1,532.00**	**3,199.00**	**8,292.00**	**14,122**

Note. [1] Formative assessment response sequence: 1 – Both formative assessment items incorrect; 2 – Student response system question correct, but quiz question incorrect; 3 – Student response system question incorrect, but quiz question correct; 4 – Both formative assessment items correct

A further post-hoc analysis aimed to identify if excitement levels for either the course or the student response system were related to performance relative to expectations. Participants who performed as expected reported the highest level of excitement for the course; conversely, excitement for using a student response system decreased as students performed worse relative to expectations. Summaries of the excitement levels by relative performance are contained in Table 15 below.

Table 15

Mean Course Excitement Levels and Student Response System Excitement Levels by Performance Relative to Expectations

	Course excitement		Student response system excitement	
Group	*Mean*	*SD*	*Mean*	*SD*
Above	3.76	1.09	4.05	0.87
In	4.11	0.80	4.22	0.86
Below	3.80	1.15	4.27	0.80

One-way ANOVA tests comparing the excitement levels across the three levels of performance found no significant difference in the mean course excitement levels ($F(2, 143) = 1.92, p = 0.15$) or in the mean student response system excitement levels ($F(2, 143) = 0.41, p = 0.67$). A weak but significant positive correlation ($r = 0.26, t = 3.19, p = 0.002$) existed between participants' excitement for the course and their excitement for using a student response system.

Comparison of Exam Success for Complete and Incomplete Sequences

Because the focus of this study was on the influence of formative assessment responses on the probability of answering a summative assessment question correctly, a post-hoc analysis was run to better understand the impact of failing to respond to at least one of the formative assessment questions in a sequence. The 145 participants involved in this study combined to leave 2,118 (13.04%) sequences incomplete by failing to answer at least one of the formative assessment questions. Participants answered 77.67% of exam questions correctly when they failed to respond to at least one formative assessment question. This is lower than the 84.92% of correct exam answers that occurred when participants answered both formative assessments.

When participants answered the student response system question correctly but did not complete the quiz question, they were slightly more likely to answer the exam question correctly than those who regressed by missing the quiz question after getting the student response system question correct. Failure to respond to the student response system question did not have any impact on participants' success when they answered the quiz question correctly. Given that the participant answered the quiz question correctly, 81.81% of summative assessment questions were answered correctly when participants missed the student response system question compared to 81.80% for participants who forwent the student response system question. Participants performed significantly worse when they missed their only attempt on a formative assessment, answering less than 70% of questions correctly in these situations; however, they answered about 6% more exam questions correctly when the quiz question was attempted rather than the student response system question. Participants answered about two-thirds of exam questions correctly when they failed to respond to both formative assessments. Success rates on exams questions for the corresponding sequence where at least one formative assessment question was unanswered are contained in Table 16.

Table 16

Exam Success by Success on Incomplete Formative Assessment Sequences

	Exam response result		
Formative assessments	Correct (%)	Incorrect (%)	Total
SRS incorrect, quiz unanswered	38 (62.30)	23 (37.70)	61
SRS correct, quiz unanswered	113 (79.58)	29 (20.42)	142
SRS unanswered, quiz incorrect	332 (68.45)	153 (31.55)	485
SRS unanswered, quiz correct	1,124 (81.80)	250 (18.20)	1,374
Both unanswered	38 (67.86)	18 (32.14)	56
Total	**1,645 (77.67)**	**473 (22.33)**	**2,118**

To determine if the correctness on the exam question was related to the formative assessment questions that were answered or unanswered, a chi-square test for independence was performed on the count data presented in Table 16 as well as the counts in Table 9 that display counts for when both formative assessment questions were answered. The results of the chi-square test showed a significant relationship (χ^2 (df = 8, n = 16240) = 421.876, $p < 0.001$) between the responses provided on the formative assessments and whether the exam question was answered correctly, indicating the variables were not independent. The Pearson residuals are presented in Table 17.

Table 17

Pearson Residuals Comparing Formative Assessment Responses with Exam Correctness

	Exam response result		
Formative assessments	Correct	Incorrect	**Total**
Both answered	1.227	-2.810	**8,292**
SRS incorrect, quiz unanswered	-1.848	4.231	**61**
SRS correct, quiz unanswered	-0.572	1.310	**142**
SRS unanswered, quiz incorrect	-3.731	8.541	**485**
SRS unanswered, quiz correct	-0.879	2.012	**1,374**
Both unanswered	-1.316	3.014	**56**
Totals	**13,638**	**2,602**	**16,240**

Only the group for which both formative assessment questions were answered performed significantly better than if the variables had been independent. Participants who did not answer any formative assessment questions correctly, regardless of if the question was missed or unanswered, performed the worst and answered significantly fewer exam questions correctly than would have been expected under independence. Those who answered one formative assessment question correctly but did not answer the other performed worse than expected under independence but fared better than those who did not answer either question correctly.

Relationship Between Course Unit, Level of Cognitive Demand, and Sequence of Responses

The second research question involves identifying relationships between the unit of the course, the level of cognitive demand, and the sequence of all three responses on a concept when aggregated over all participants. A $4 \times 3 \times 8$ cross-classification table aggregating the responses from all participants and comparing the unit of the course, level of cognitive demand, and the eight combinations of sequences defined in Table 6 can be found in Appendix L.

To identify the model that fits the data the best, nine different loglinear models were generated with varying interaction terms for the unit of the course, the level of cognitive demand, and sequence according to the models specified in equations (3), (4), and (5). One model included only the factors with no interaction terms. Three models included a single two-way interaction term. Three additional models included two two-way interaction terms. One model included all three two-way interaction terms. The final model was the saturated model including the three-way interaction term that guarantees a perfect fit. The deviance statistics for each model are contained in Table 18.

Table 18

Deviance Statistics for Considered Loglinear Models

Model[1]	*df*	Deviance	*p*
(U,C,S)	83	2,174.75	<0.001
(UC,S)	77	1,702.56	<0.001
(US,C)	62	1,729.84	<0.001
(CS,U)	69	1,312.84	<0.001
(UC,US)	56	1,257.66	<0.001
(UC,CS)	63	840.66	<0.001
(US,CS)	48	867.94	<0.001
(UC,US,CS)	42	379.89	<0.001
(UCS)	-	-	-

Note.[1] *U* denotes the unit of the course, *C* denotes the level of cognitive demand, and *S* denotes the sequences of responses from the three questions in the sequence. *UC*, *US*, and *CS* denote the two-way interaction terms between the individual variables. *UCS* denotes the three-way interaction term.

In determining an appropriate loglinear model, the following hypotheses were used for each of the above nine models:

H_0: The model is a good fit for the data.

H_A: The model is a poor fit for the data.

The null hypothesis was rejected (χ^2 = 2,174.55, p < 0.001) for the mutual independence model (U, C, S) defined in equation (3), indicating it was not an adequate fit for the data. There was some dependency between the unit of the course, level of cognitive demand, and sequence of responses.

Adding interaction terms to the mutual independence model assisted in fitting the data slightly better, but none of the models that included two-way interaction terms were statistically significant. Each remained a poor fit for the data. While the deviance statistic decreased as more interaction terms were added, none of the models with second-order terms fit well enough to describe the relationship between the three variables.

The saturated model in equation (5) fit the data perfectly. Due to the large sample size, it was compared against the model with all second-order interaction terms to check for practical significance. This was accomplished by checking if the odds ratios for combinations of two variables changed at different levels of the third variable across the two models. Because the actual count in the cell for participants who missed all three questions in one-sample inference unit on recall questions was zero, a count of 0.5 was imputed to allow the loglinear model to converge. The odds ratios for the model with all two-factor interaction terms compared against the saturated model are contained in Appendix M.

The odds ratios across the different levels of the third variable change significantly whenever the third-order interaction term was added to the loglinear model to generate the saturated model. Because of these large differences, the difference in the deviance statistics in the models (UC, US, CS) and (UCS) was not due to the large sample size but rather a strong deviation from independence. Thus, the saturated model (UCS) was the best fit for the data, indicating that the unit of the course, level of cognitive demand, and sequence of responses were all dependent. There was conditional independence between all three pairs of variables because the saturated model was the best fit for the data. Appendix N contains tables with the expected counts under the assumption that all three variables are mutually independent as well as the Pearson residuals.

Analysis of Pearson residuals.

Pearson residuals larger than 3 in magnitude are an indication that observed count in the group was highly unusual compared to the count that would have been expected if the variables had been independent (Agresti, 2007) although some residuals this large may occur by chance with large samples sizes. To make it more likely the residuals being discussed were due to the relationship between the variables, this analysis will focus on residuals with magnitudes greater than 5, indicating an extremely unusual count under independence. There were 14 cells with Pearson residuals greater than 5 and eight with Pearson residuals less than -5. Large negative Pearson residuals resulted in smaller than expected counts in the following situations:

- Strategic thinking sequences where all three questions were answered correctly in all four units of the course
- Basic application of skill sequences where both formative assessment questions were answered incorrectly but the exam question was answered correctly in the descriptive statistics unit

- Recall sequences where all three questions were answered correctly in the probability unit
- Recall sequences where only the quiz question was answered correctly in the one-sample inference unit

Large positive Pearson residuals resulted in larger than expected counts in the following situations:

- Recall and basic application of skill sequences where all three questions were answered correctly in the descriptive statistics unit
- Strategic thinking sequences in the descriptive statistic, probability, and two-sample inference units where the exam question was answered incorrectly
- Recall sequences in probability where both formative assessment questions were answered incorrectly
- Strategic thinking sequences in one-sample inference and two-sample inference where the student response system question was answered incorrectly, but the exam question was answered correctly

Conclusion

The results of this study demonstrated that the responses provided by a participant on a sequence of formative assessments were associated with the probability of correctly answering the corresponding exam question. The unit of the course and level of cognitive demand also played roles in participants' success on the summative assessment question. Participants who responded to more formative assessment questions correctly in the sequence were more likely to answer the exam question correctly. In situations where participants answered only one formative assessment question correctly, those who answered the quiz question correctly were more likely to respond to the summative assessment question correctly.

The result of the quiz question was more highly correlated with success on the exam question compared to the result of the student response system question. Participants who did not improve upon their success in the formative assessments during a sequence by either missing both questions or regressing by missing the quiz question showed no significant difference in the probability of answering the exam question correctly, indicating that the student response system may not have been effective in these situations. Alternatively, participants who answered the quiz question correctly were significantly more like to answer the exam question correctly regardless of success on the student response system question.

Concepts that required higher levels of cognitive demand were associated with significantly lower probabilities of success on the exam question. Participants struggled the most on sequences in the probability unit but performed the best on sequences in the inference units. Those participants who had previously taken a statistics course were more likely to answer the corresponding exam question correctly than those had not taken one. Pre-course excitement levels were also significant in predicting the probability the exam question was answered correctly. Participants who were more excited to take the course were significantly more likely to answer the exam question correctly while those who more excited to use the student response system were actually less likely to have success on the summative assessment.

In the post-hoc analysis of completed sequences, participants who completed more sequences tended to overperform expectations on the summative assessments while those who attended less frequently tended to underperform. Participants who answered both formative assessment questions correctly were more likely to overperform expectations; alternatively, those who missed both formative assessment questions were more likely to underperform expectations.

The data also revealed that the unit of the course, level of cognitive demand required on the sequences, and the responses on the sequences were not independent. In all four units, there were many fewer sequences where all three questions were answered correctly when the level of cognitive demand was at the strategic thinking level than would have been expected under independence. Students tended to struggle significantly more than expected on exam questions at the strategic thinking level. Moreover, students missed both formative assessments at the recall level in probability more frequently than would have been expected but overachieved at lower levels of cognitive demand in the descriptive statistics unit.

Chapter 5

Discussion

Introduction

The purposes of this study were twofold: to determine if the sequences of responses that participants provided on two formative assessments were predictive of their success on a corresponding summative exam question and to analyze the relationship between the sequences of responses, the unit of the course, and the level of cognitive demand. In this chapter, the conclusions from each research question will be analyzed along with their relationship to the current body of literature. Implications for teaching introductory statistics and the contribution to the literature are discussed next. The chapter concludes with the limitations of the study and recommendations for future research.

Influence of Formative Assessment Response Sequences

The goal of the first research question was to determine how the sequences of responses that participants provided on two formative assessments were related to the probability that they will answer the corresponding exam question correctly while controlling for six other factors. These six factors include the unit of the course in which the concept was presented, the level of cognitive demand required to answer the concept's assessment items, participants' previous statistics experience, their interest in the course, their interest in using a student response system, and the time of day the class was taught.

The findings from the study indicate that participants who answered both formative assessment questions correctly were significantly more likely to answer the summative assessment item correctly when compared to those who missed at least one formative assessment question. Thus, the probability of answering the summative exam item correctly increased when more formative assessments were answered correctly. The quiz question had a larger increase in the predicted probability than the student response system question.

Formative assessment using both a student response system for specific concepts during class and a quiz outside of class aided participants' performance on the corresponding summative assessment question. This finding is consistent with the results from Glass et al. (2008), who found that students who were exposed to two types of formative assessment performed better on the summative assessment compared to those exposed to either zero or one formative assessments. However, the number of formative assessments may not be the only factor that determines how well students perform on summative assessments. Both the order of the formative assessments and the learners' cognitive ability may play a role in students' success on an exam. Glass et al. (2008), who had students answer a quiz question outside of class first and then followed up with a student response system question in-class, witnessed an improvement in the overall success rate between the two formative assessments, but decline thereafter on the exam. While the order in the study by Glass et al. (2008) was reversed from this study where students could discuss the question in class with peers first, both found that student performance on exams was higher if the instructor offered two formative assessments rather than one. Hubbard and Couch (2018), who used two in-class formative assessments at different time points, suggested that higher-performing students may benefit from repeated formative assessment. They may have been able to retain content knowledge better than lower-performing students, whose scores on the summative assessments tended to regress to levels similar to those on the first formative assessment. A future study that retains the same two question formative assessment format and order but separates students according to their performance on exams could identify significant differences in performance across the groups at each of the three assessments.

By using different questions on the summative assessment than on the formative assessments, participants were more likely to learn how to do statistics. This variation in questions was to ensure that participants were applying their knowledge of statistics to new scenarios rather than memorizing a process. Glass et al. (2008) applied a similar technique when they tested the same factual knowledge during each assessment but worded the question slightly differently. In a later study, Glass (2009) found that by varying the wording of questions on subsequent assessments, students' scores initially decreased on the formative assessment outside of class compared to in class but rose again on the exam because they had the ability to generalize concepts rather than memorize facts.

The student response system had some impact on the probability of the participants answering the corresponding summative assessment correctly. Answering only the student response system question correctly did not lead to a significantly higher probability of answering the exam question correctly compared to answering both formative assessments incorrectly. However, answering both formative assessment questions correctly yielded the highest probability of answering the summative assessment question correctly. This is an indication that better exam performance was more correlated with quiz performance than student response system performance. As all student response system questions were multiple-choice, students who did not understand the concept well may have chosen the correct answer through random guessing, which was a tactic that Kulikovshikh, Prokhorov, and Suchkova (2017) found to be employed more frequently by lower-performing students. Guessing could have been their choice of action because the participants may not have understood the concept with which to begin.

Participants were permitted to discuss the student response system questions so the results in this study could have been confounded by lower-performing participants receiving the correct answer from a peer as suggested by Hubbard and Couch (2018). Similarly, this collaboration may have also inadvertently assisted stronger participants by reinforcing the concepts so they were more likely to answer the quiz and exam questions correctly. When working in small groups, lower-performing students tend to learn less through collaborative efforts than higher-performing students (Kulikovshikh et al., 2017). Conversely, higher-performing students may have benefited from explaining concepts to their lower-performing peers (Hubbard & Couch, 2018). The offering of an extra credit point for answering the student response system question correctly may have hindered weaker students. These participants may have opted for a small immediate reward by obtaining the correct answer from a peer rather than attempting the question on their own, which would have given them the opportunity to retain new knowledge through constructivism. This is consistent with the study by White et al. (2011), who found that student response systems help to create a learning experience through constructivism when used interactively rather than for attendance or participation.

Participants who answered the quiz question correctly fared significantly better than those who missed the quiz question, regardless of their success on the student response system question. These participants either understood the concept from the beginning by answering both formative assessments correctly or learned from their mistakes in class when attempting the second and third assessments in the sequence. They may have benefited from the instructor's quick and informative verbal feedback that arose from receiving the responses within seconds, an advantage to using a student response system recognized by Gikandi et al. (2011). Not only does feedback tend to clarify concepts for students, but it also may assist them in identifying weaknesses (Garfield et al. 2011). The instructor also adjusted his instruction at times to better convey concepts on which participants struggled during class. This change in pedagogy assisted in better meeting students' needs (Terrion & Aceti, 2011). Allowing students the opportunity to correct initial mistakes at a later time, such as on a self-correcting midterm as in a study by Grühn & Cheng (2014), promotes the reinforcement of the concept and improves scores on later forms of assessment.

Participants who were more excited to take a statistics course were more likely to answer the summative assessment items correctly. Their personal goals may have contributed to their success on the summative assessments. Students tend to adopt one of two types of goal orientations: mastery goal orientation, where their focus is on acquiring as much knowledge about the subject matter as possible, and performance goal orientation, where their attention is on succeeding on assessments rather than learning (Zingaro, 2015). Zingaro (2015) found that students who reported a higher level of interest were more likely to have mastery goals and thus, score higher on the final exam; conversely, students with performance goals tended to score worse on the final exam. Harlow et al. (2014) identified a similar relationship in that students who enrolled in a physics class for their own interest scored significantly higher than those who took the course solely because it was a requirement. Participants who were interested in taking a statistics course may have had higher levels of self-efficacy, which is associated with better performance on exams and higher participation during class (Galyon et al., 2011). Similarly, participants with a high interest level may also have taken a previous statistics course, which is significant in predicting exam performance the second time through (Sutherlin et al., 2013).

Participants who were more excited to use a student response system were significantly less likely to answer the exam questions correctly. Their excitement levels for the course did not appear to confound the relationship between use of the student response system and success on exam questions the correlation between their excitement for the course and their excitement for using a student response system was positive and significant. Instead, the negative effect of the excitement for the student response system may have resulted from participants who more motivated to come to class to earn the extra credit points (i.e. performance goal orientation), which would help them reach their ultimate goal of admission into the business school. Participants may have seen the extra credit offered by responding to the student response system as an easier way to achieve the requisite grade of a B. Conversely, those who focused on learning and improvement (i.e. mastery goal orientation) may have seen the student response system as a distraction to deeper learning. Considering the negative relationship between excitement for the student response system and exam performance and the positive relationship between excitement for the course and exam performance, this study is in line with Zingaro's finding (2015) that students who had performance goals rather than mastery goals tended to perform more poorly on the final exam. This negative association also could have occurred because students' opinions of the student response system improved throughout the semester, as found in a study by Sutherlin et al. (2013). High performing students may have initially rated their interest in the student response system as low, but later found them enjoyable as the semester progressed. Participants' initial impression of the student response system may not have been reflective of their opinion throughout the majority of the semester. The effectiveness of a student response system is dependent upon the users believing they are useful (Rana & Dwivedi, 2016). Thus, then the negative effect may have resulted from higher-performing participants not realizing that their true interest level in using the student response system was actually much higher than they initially reported.

Participants who completed more sequences were significantly more likely to overachieve on the exam questions; conversely, those who completed fewer sequences were significantly more likely to underachieve on exam questions. All participants involved in this study answered every question on all four exams and the completion rates on the online quizzes exceeded 90%. Thus, the most frequent occurrence for the unfinished sequence was the failure to answer the student response system question, which largely resulted from not attending class. Higher-performing participants' grades were not impacted as much by missing class as their success rate on the summative assessments nearly matched those participants who answered the quiz question correctly but missed the student response system question. This observation matches that of Westerman et al. (2011). Conversely, missing class affected lower-performing participants' grades more than higher-performing participants. Their success rate on the exam questions was lower than the participants who answered but missed both formative assessments. Simply answering the formative assessments, even incorrectly, likely benefits lower-performing students as they could receive some beneficial feedback from a wrong answer (King & Joshi, 2008; Lantz & Stawiski, 2014). Lower-performing students tended to decline throughout the semester, which is consistent with the findings of Lee et al. (2015), likely because later material in a statistics course is dependent upon concepts learned earlier in the course. This decline could have resulted from a lack of consistently attending class, low interest in the course, or low interest in using the student response system.

Influence of Relationship Between Course Unit, Level of Cognitive Demand, and Sequence of Responses

The goal of the second research question was to determine if the unit of the course, the level of cognitive demand, and the sequences of responses on all three assessment items were independent. A level of cognitive demand (recall, basic application of skill, or strategic thinking) was assigned to each sequence according to the definitions set by Webb's depth of knowledge. Each sequence was also presented during one of the four units of the course: data collection and descriptive statistics, probability, one-sample inference, or two-sample inference. The results of the study indicated that the unit of the course, the level of cognitive demand, and the sequence of responses were not independent and that every pair of variables was conditionally independent.

The lack of independence between the unit of the course, level of cognitive demand, and sequence of responses indicates that participants learn at different rates during separate units of the course and also at different rates for higher levels of cognitive demand. Sequences requiring a higher level of cognitive demand or that were presented in more difficult units of the course were associated with more incorrect answers. Garfield et al. (2011) endorsed using questions at different levels of difficulty because they assist the instructor in identifying a consistent level of difficulty in the future. The inclusion of questions at the strategic thinking level of cognitive demand in this study counters the argument by Hodgson and Pang (2012) that higher-order thinking cannot be assessed through multiple-choice questions. This study also challenged the narrative presented by Salcedo (2014) that too many statistics assessment questions either require a low level of cognitive demand or only test concepts in descriptive statistics; nearly one-third of sequences were asked at the strategic thinking level of cognitive demand and over half were regarding inference.

Participants performed significantly better than expected in the descriptive statistics unit of the course on questions involving recall and basic application of skill, likely because they had prior exposure to introductory concepts in statistics. Approximately half of the participants involved in this study had taken a previous statistics course. These participants certainly would have had exposure to data collection and descriptive statistics as they are typically the first units covered in a statistics course (Pfenning, 2011). Having previous exposure to material from a prior course improves performance the second time taking a similar course (Champagne & Klopfer, 1982). Even participants who had not taken a previous statistics course likely had experience working with the collection, display, and summarizing of data as Franklin et al. (2007) emphasize the teaching of these concepts in mathematics courses from pre-kindergarten through the twelfth grade. Participants may have used the lecture to recall this prior knowledge, improving their chances of answering the student response system question correctly, as suggested by Hartle et al. (2012). Because all questions were multiple-choice, randomly guessing correctly on the student response system question without actually knowing the correct answer could have led to a higher than expected count, as suggested by Kulikovshikh et al. (2017). Students were not under a time constraint on the quizzes so there was little pressure to answer questions quickly. The unusually high number of exam questions answered correctly could be attributed to the fact that the descriptive statistics unit had the most sequences where participants answered two formative assessments. This is supported in the study by Glass et al. (2008) who found that the success rate on the exam was lower when students answered only one type of formative assessment as opposed to two.

Participants tended to struggle with probability unit the most. There were many more sequences than expected involving incorrect responses to both the student response system and exam questions. This is likely due to the content and requirement of problem-solving rather than the sequencing of the questions. Utts (2003) stated that students have a poor intuition for probabilistic concepts, particularly those involving conditional probabilities. Poor intuition could have led participants to repeat mistakes made with the student response question on the quiz and exam questions. One common area of confusion is the inability to distinguish between independent and disjoint events (Keeler & Steinhorst, 2001). This difference, if misunderstood and never corrected, can negatively affect students' ability to decide on how to approach a problem. Similarly, students may use the standard deviation instead of the standard error when working with sampling distributions, which will lead to an incorrect standardized score (Motulsky, 2014). Students tend to understand when to use the standard error instead of the standard deviation once the basics of hypothesis testing have been covered in inference; however, when first introduced in probability, the transition from sampling one observation to several can lead to confusion about when to incorporate the sample size in the calculation. There was an unusually high number of sequences in the recall and strategic thinking levels of cognitive demand in the probability unit where participants answered both formative assessments incorrectly but answered the exam question correctly. The verbal and written feedback students received after the formative assessments may also have finally started to resonate by the exam, as proposed by Gikandi et al. (2011), King and Joshi (2008), and the U.S. Department of Education (2017).

In general, the concepts in inference tended to have significantly more sequences with correct exam answers than expected. Participants may have performed better on inferential topics for two reasons. First, descriptive statistics questions require more explanation of findings, whereas inference questions are more algorithmic. Once students identified the correct inferential technique, performing a hypothesis test or calculating a confidence interval was straightforward. The difficulty arose when students had to identify confounding variables or describe the relationship between variables, possibly for the first time. Horton (2015) emphasized that such multivariable methods are critical to teach students in their first statistics course. Second, concepts in inferential statistics often refer to concepts in descriptive statistics that students have since improved upon. Inferential statistics builds on the topics from descriptive statistics, but creates an investigative and algorithmic process when testing hypotheses or estimating using a confidence interval (GAISE, 2016). Participants may also have scored better on inferential topics because they were more familiar with the terminology and the techniques required to examine data, two potentially underdeveloped skills at the beginning of the course. This aligns with Garfield et al. (2011) who found that new statistics students take time to learn statistical literacy and understand the process of investigating data.

The significant differences in achievement depending on the level of cognitive demand support the use of Webb's depth of knowledge (Webb, 2002) as a means of classifying questions and assessing students in a computational field such as statistics. When participants answered the exam question correctly, sequences with higher counts than expected tended to be classified at the recall level; conversely, sequences with lower counts than expected tended to require strategic thinking. This relationship is likely a result of the questions becoming more difficult, which is reinforced in the study by Rush et al. (2016), which found that the percentage of correct answers on exams decreased as the complexity of the question increased. However, this is contrary to the findings of Kibble and Johnson (2011) who found no relationship between the level of cognitive demand and student performance. The lack of a relationship can be attributed to the fact that items requiring higher-order cognition are more difficult, more discriminating, and correlate with student learning (Rush et al., 2016). Each level of cognitive demand in this study was associated with a different type of knowledge: questions at the recall level required factual knowledge, questions requiring basic application of skill required procedures, and strategic thinking questions required conceptual knowledge. As expressed by Banerjee, Rao, and Ramanathan (2015), questions requiring conceptual knowledge tend to be more difficult than those requiring either factual or procedural knowledge, which explains the increase in difficulty across the three levels of cognitive demand.

Performance between the recall and basic application of skill levels was more similar than performance between the basic application of skill and strategic thinking levels. One possible explanation is that students do not benefit from a student response system if it is used for factual types of questions, such as many of the concepts classified as recall, as previously reported by White et al. (2011). A second possibility is that students were more engaged with formative assessment questions that required a higher level of cognitive demand. Paige et al. (2013) found that engagement tended to be positively associated with the level of cognitive demand required of the task, but declined more quickly during class for questions requiring lower levels of cognitive demand. If engagement declined towards the end of class on less cognitively demanding questions but remained high on those requiring strategic thinking, the interaction between when the question was posed and the level of cognitive demand could have further confounded the differences in effect sizes.

Limitations

By analyzing how the sequences of responses on two types of formative assessment were related to student success on the exam on an individual level, this study investigated a use of active learning and feedback that had been otherwise unexplored. However, this study did not examine the impact of having only one level of formative assessment because the sequences of responses were the variable of interest. While answering the student response system question correctly by itself was not significantly associated with a higher probability of getting the exam question correct, eliminating the student response system and only using the quiz for formative assessment would not necessarily lead to similar results. The active engagement and verbal feedback provided during class by the instructor may have been useful in helping participants make connections for the quizzes although the type of feedback and how it was delivered were not a focus of this study.

The possible collaboration of participants on the formative assessments was not controlled for in this study. The instructor did not want to discourage collaboration between participants during class because of the potential for them to learn from one another while answering the student response system question. However, permitting communication allowed grade driven participants to receive correct answers from their peers without fully understanding the concept. This could explain why answering only the student response system question correctly did not have a significantly different impact from answering both formative assessments incorrectly. In cases where participants blindly copied their peers, some of the sequences with only a correct student response system question should have been sequences with two incorrect formative assessment responses because they did not actually know the correct answer. As participants completed the quizzes outside of class without instructor and teaching assistant supervision, it is possible that participants collaborated. However, few quizzes were completed in under five minutes so the likelihood that participants directly copied answers from a peer is low.

This study could not consider the impact that random guessing had on the results. Many sequences where participants answered only the student response system questions correctly likely arose from randomly guessing the correct answer because respondents were not penalized for incorrect responses. Participants would need to have regressed a substantial amount to grasp the concept initially, but miss the second formative assessment and summative assessment questions. However, the true effect of random guessing cannot be inferred from this study.

Time constraints limited participants on the first three unit exams by the 50-minute class length, which did not occur during the online quizzes. They were permitted to complete the quizzes at their own leisure, even in multiple sessions if they chose. Participants who worked slower and more methodically may have been negatively impacted by the time constraint, leading to more incorrect answers than would have otherwise occurred had the setting been similar to the quizzes. Thus, additional pressure to finish the exam or careless mistakes due to rushing could have confounded the results.

The sample in this study may not have been representative of the general population of participants who would be taking an introductory statistics course. This sample was obtained from students whose primary interest was enrolling in the business school. These students tended to have above average analytical skills with which to begin. Moreover, recruitment was performed during the final recitation of the semester so only participants who were motivated to attend could have elected to participate. Participants who were more comfortable with the material may have forwent attending recitation, leading to fewer complete sequences that had many correct answers because they were more likely to attend class regularly. Conversely, disinterested potential participants also likely did not attend recitation and thus, could not consent. These participants were probably more likely to have several incomplete sequences, but their lack of participation likely led to fewer sequences with incorrect answers and fewer incomplete sequences in general, balancing out the impact of the two types of non-attendees.

Another potential limitation is with regards to the sample sizes in the one-way ANOVA that compared the number of completed sequences for participants who overachieved, underachieved, and performed as expected. Because participants' performance relative to expectation was determined by 95% confidence intervals, it is reasonable to expect many more participants to perform in line with expectations. The assumption of homogeneity of variances in one-way ANOVA for which no widely accepted rule of thumb exists was not severely violated by the lack of equal sample sizes. Participants who overachieved were more likely to complete more sequences but were also bounded above by 112, the total number of sequences, likely causing the smaller variability in this highest performing group.

Implications for Teaching

The results of this study provide support for several potential modifications for teaching an introductory statistics course. As answering the student response system question correctly was marginally associated with a higher probability of success on the summative assessment, instructors may consider implementing a student response system in the classroom, as recommended by the GAISE College Report (2016). However, the student response system question by itself was not sufficient to assist students on the summative assessments. A second formative assessment that checks for understanding of concepts before issuing a summative assessment may be one way of improving student success.

Instructors may consider reevaluating their course policies regarding attendance to require that students attend. This policy change is supported by the fact that participants who completed more sequences consistently overperformed on the summative assessments. Any student response questions that were answered resulted from voluntary attendance because this study did not require participants to attend class. While putting a limit on the number of classes that students can miss may frustrate some, having consistent exposure to material that constantly builds on itself and actively engaging students during class will help them in the long term.

The findings from this study indicate that using formative assessment and feedback builds on students' understanding of statistical concepts. Statistics instructors should note that repeated practice can lead to greater mastery of a concept, as demonstrated in this study. While students who previously took an introductory statistics course had a slight advantage, the effect was small enough that the changes in the course structure do not appear necessary to accommodate for any gaps in student knowledge.

Recommendations for Future Research

Future research related to the findings of this study could extend in a number of directions. While this study only analyzed the effect of the sequences of responses on formative assessment, future research could perform the study over multiple semesters, eliminating one level of formative assessment each time. This would allow for a comparison of the impact of only the student response system or quizzes on exam responses with the results from this study because the student response system was not as helpful to students' overall success as the quizzes were.

Future research could include other variables, such as adding a third level of formative assessment by using homework problems. A corresponding question for each sequence could be implemented between the student response system question and quiz question to create eight different sequences on the formative assessments. Because of the potential for students to collaborate on homework, an online system where different numbers are uniquely generated for each student would be necessary to stop prevent direct copying.

While the multiple-choice nature of all the questions allowed the instructor to easily determine if the answer was correct, this also allowed for the possibility of blindly guessing the correct answer. Future research could alleviate this problem in two ways. One method would be to ask students to rate the confidence of their response or eliminate answer choices they know are incorrect. This added variable would allow the instructor to clearly identify the concepts on which students are guessing compared to those where students believe they understand the concept but are actually mistaken. A second modification to this study would be to restructure the assessments so the questions require short free response answers. These questions could require students to demonstrate some mastery of the topic by writing a sentence or doing a short calculation but can be graded easily to accommodate for large class sizes.

Conclusions

This study found that offering participants the opportunity for more formative assessment was associated with a higher probability of answering an exam question correctly on a similar concept. While the student response system question did not appear to significantly increase the probability that a student will get the summative assessment correct, the second level of formative assessment, the quiz, did increase this probability. Moreover, this study discovered that participation on the formative assessments was correlated with higher than expected scored on the summative assessment. Despite statistics being a computational field, this study was successful in finding that multiple-choice questions combined with verbal or written feedback can test participants' knowledge of statistical concepts given the improvement displayed through the different levels of formative assessment. In general, this study found that participants can learn statistical concepts, review and correct their mistakes, and apply their new knowledge to different situations encountered in the statistical world.

References

Agresti, A. (2007). *An Introduction to Categorical Data Analysis* (2nd ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Attali, Y., Laitusis, C., & Stone, E. (2016). Differences in reaction to immediate feedback and opportunity to revise answer for multiple-choice and open-ended questions. *Educational and Psychological Measurement*, *76*, 787-802. doi:10.1177/0013164415612548

Banerjee, S., Rao, N. J., & Ramanathan, C. (2015). Rubrics for assessment item difficulty in engineering courses. *2015 IEEE Frontiers in Education Conference*. El Paso, TX. 1-8

Barak, M. (2017). Science teacher education in the twenty-first century: A pedagogical framework for technology-integrated social constructivism. *Research in Science Education*, *47,* 283-303. doi:10.1007/s11165-015-9501-y

Bearman, M., Dawson, P., Bennett, S., Hall, M., Molloy, E., Boud, D., & Joughlin G. (2017). How university teachers design assessments: a cross-disciplinary study. *Higher Education*, *74,* 49-64. doi:10.1007/s10734-016-0027-7

Beichner, R. J. (2014). History and evolution of active learning. *New Directions for Teaching and Learning*, *2014*, 9-16. doi:10.1002/tl.20081

Bidgood, P, Hunt N., & Jolliffe, F. (2010). *Assessment Methods in Statistical Education: An International Perspective*. Chichester, United Kingdom: John Wiley & Sons Ltd.

Blood, E., & Gulchak, D. (2012). Embedding “clickers” into classroom instruction: Benefits and strategies. *Intervention in School and Clinic*, *48*, 246-253. doi:10.1177/1053451212462878

Bojinova, E. D., & Oigara, J. N. (2011). Teaching and learning with clickers in higher education. *International Journal of Teaching and Learning in Higher Education*, *25*, 154-165. Retrieved from http://www.isetl.org/ijtlhe/

Bonwell, C. C., & Eison, J. A. (1991). *Active learning: Creating excitement in the classroom*. ASHE-ERIC Higher Education Report No. 1. Washington, DC: The George Washington University, School of Education and Human Development.

Buil, I., Catalán, S., & Martínez, E. (2016). Do clickers enhance learning? A control-value theory approach. *Computers and Education*, *103*, 170-182. doi:10.1016/j.compedu.2016.10.009

Bush, M. (2015). Reducing the need for guesswork in multiple-choice tests. *Assessment & Evaluation in Higher Education*, *40*, 218-231. doi:10.1080/02602938.2014.902192

Bush, H. M., Daddysman, J., & Charnigo, R. (2014). Improving outcomes with Bloom's taxonomy: From statistics education to research partnerships. *5*, 1-3, doi:10.4172/2155-6180.1000e130

Büyükkurt, M., Li, Y., & Cassidy, R. (2012). Using a classroom response system in an introductory business statistics course: Reflections and lessons learned. *Informing Science & IT Education Conference*, Montreal, 379-388.

Carloye, L. (2017). Mini-case studies: Small infusions of active learning for large-lecture courses. *Journal of College Science Teaching*, *46*, 63-67. doi:10.2505/4/jcst17_046_06_63

Carr, R., Palmer, S., & Hagel, P. (2015). Active learning: The importance of developing a comprehensive measure. *Active Learning in Higher Education*, *16*, 173-186. doi:10.1177/1469787415589529

Chachashvili-Bolotin, S., Milner-Bolotin, M., & Lissitsa, S. (2016). Examination of factors predicting secondary students' interest in tertiary STEM education. *International Journal of Science Education, 38*, 366-390. doi:10.1080/09500693.2016.1143137

Champagne, A. B., & Klopfer, L. E. (1982). A causal model of students' achievement in a college physics course. *Journal of Research in Science Teaching*, *19*, 299-309. doi:10.1002/tea.3660190404

Chance, B., Ben-Zvi, D., Garfield, J., & Medina, E. (2007). The role of technology in improving student learning of statistics. *Technology Innovations in Statistics Education*, *1*(1), 1-26. Retrieved from https://escholarship.org/uc/uclastat_cts_tise

Cobb, G. W. (1998). *The objective-format question in statistics: Dead horse, old bath water, or overlooked baby?*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Curran-Everett, D. (2000). Multiple comparisons: philosophies and illustrations. *American Journal of Physiology – Regulatory, Integrative, and Comparative Physiology*, *279*, 1-8. doi:10.1152/ajpregu.2000.279.1.R1

Dixson, D. D., & Worrell, F. C. (2016). Formative and summative assessment in the classroom. *Theory Into Practice*, *55*, 153-159. doi:10.1080/00405841.2016.1148989

Dunham, B. (2009). *Statistics clicks: Using clickers in introductory statistics courses*. Paper presented at the International Conference on Improving University Teaching, Vancouver, BC. Retrieved from http://www.cwsei.ubc.ca/SEI_research/files/Stat/BDunham_ClickersInStat.pdf

Dunham, B., Yapa, G., & Yu, E. (2015). Calibrating the difficulty of an assessment tool: The Blooming of a statistics examination. *Journal of Statistics Education*, *23*(3), 1-33. doi:10.1080/10691898.2015.11889745

Egelandsdal, K. & Krumsvik, R. J. (2017). Clickers and formative feedback at university lectures. *Education and Information Technologies*, *22*, 55-74. doi:10.1007/s10639-015-9337-x

Ellis, V. A. (2016). Introducing the creative learning principles: Instructional tasks used to promote rhizomatic learning through creativity. *The Clearing House: A Journal of Educational Strategies, Issues, and Ideas*, *89*, 125-134. doi:10.1080/00098655.2016.1170448

Falconer, J. L. (2016). Why not try active learning?. *AlChE Journal, 62*, 4174-4181. doi:10.1002/aic.15387

Fletcher, R. B., Meyer, L. H., Anderson, H., Johnston, P., & Rees M. (2012). Faculty and students conceptions of assessment in higher education. *Higher Education*, *64*, 119-133. doi:10.1007/s10734-011-9484-1

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to Design and Evaluate Research in Education* (8th ed.). New York, NY: McGraw Hill.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R. (2007). *Guidelines for assessment and instruction in statistics education (GAISE) report: A pre-k-12 curriculum framework.* Retrieved from http://www.amstat.org/education/gaise

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 8410-8415. doi:10.1073/pnas.1319030111

GAISE College Report ASA Revision Committee, (2016). *Guidelines for Assessment and Instruction in Statistics Education College Report 2016*. Retrieved from http://www.amstat.org/education/gaise

Galyon, C. E., Blondin, C. A., Yaw, J. S., Nalls, M. L., & Williams, R. L. (2011). The relationship of academic self-efficacy to class participation and exam performance. *Social Psychology of Education*, *15*, 233-239. doi:10.1007/s11218-011-9175-x

Garfield, J. (1995). How students learn statistics. *International Statistical Review*, *63*, 25-34. doi:10.2307/1403775

Garfield, J., Zieffler, A., Kaplan, D., Cobb, G. W., Chance, B. L., & Holcomb, J. P. (2011). Rethinking assessment of student learning in statistics courses. *The American Statistician*, *65*, 1-10. doi:10.1198/tast/2011.08241

Garrison, D. R., Anderson, T., & Archer, W. (1999). Critical inquiry in a text-based environment: Computer conferencing in higher education. *The Internet and Higher Education*, *2*, 87-105. doi:10.1016/S1096-7516(00)00016-6

Ghilay, Y., & Ghilay, R. (2015). TBAL: Technology-based active learning in higher education. *Journal of Education and Learning*, *4*, 10-18. doi:10.5539/jel.v4n4p10

Gikandi, J. W., Morrow, D., & David, N. E. (2011). Online formative assessment in higher education: A review of the literature. *Computers & Education*, *57*, 2333-2351. doi:10.1016/j.compedu.2011.06.004

Glass, A. L. (2009). The effect of distributed questioning with varied examples on exam performance on inference questions. *Educational Psychology*, *29*, 831-848. doi:10.1080/01443410903310674

Glass, A. L., Brill, G., & Ingate, M. (2008). Combined online and in-class pretesting improves exam performance in general psychology. *Educational Psychology*, *28*, 482-503. doi:10.1080/01443410701777280

Grühn, D. & Cheng, Y. (2014). A self-correcting approach to multiple-choice exams improves students' learning. *Teaching of Psychology*, *41*, 335-339. doi:10.1177/0098628314549706

Haeusler, C. E., & Lozanovski, C. (2010). *Student perception of 'clicker' technology in science and mathematics education*. Paper presented at the 2010 Enhancing Learning Experiences in Higher Education Conference, Hong Kong. Retrieved from https://www.cetl.hku.hk/conference2010/pdf/Haeusler_2.pdf

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*, 309-333. doi:10.1207/S15324818AME1503_5

Harlow, J. J. B., Harrison, D. M., & Meyertholen, A. (2014). Correlating student interest and high school preparation with learning and performance in an introductory physics course. *Physics Review Special Topics – Physics Education Research*, *10*. doi:10.1103/PhysRevSTPER.10.010112

Harow, S., Cummings, R., & Aberasturi, S. M. (2006). Karl Popper and Jean Piaget: A rationale for constructivism. *The Educational Forum*, *71*, 41-48. doi:10.1080/00131720608984566

Harris, M. A., & Patten, K. P. (2015). Using Bloom's and Webb's taxonomies to integrate emerging cybersecurity topics into a computing curriculum. *Journal of Information Systems Education*, *26*, 219-234. Retrieved from http://jise.org/

Hartle, R. T., Baviskar, S., & Smith, R. (2012). A field guide to constructivism in the college science classroom. *Bioscene*, *38*(2), 31-35. Retrieved from https://www.acube.org/bioscene/

Hattie, J., & Brown, G. T. L. (2004). *Cognitive processes in asTTle: The SOLO taxonomy*. (Assessment Tools for Teaching and Learning Technical Report #43). Retrieved from https://auckland.rl.talis.com

Heaslip, G., Donovan, P., & Cullen, J. G. (2014). Student response systems and learner engagement in large classes. *Active Learning in Higher Education*, *15*, 11-24. doi:10.1177/1469787413514648

Hedden, M. K., Worthy, R., Akins, E., Slinger-Friedman, V., & Paul R. C. (2017). Teaching sustainability using an active learning constructivist approach: Discipline-specific case studies in higher education. *Sustainability*, *9*, 1320-1337. doi:10.3390/su9081320

Heller, J. I., Daehler, K. R., Wong, N., Shinohara, M., & Miratrix, L. W. (2012). Differential effects of three professional development models on teacher knowledge and student achievement in elementary science. *Journal of Research in Science Teaching*, *49*, 333-362. doi:10.1002/tea.21004

Hess, K. K., Carlock, D., Jones, B., & Walkup, J. R. (2009). *What exactly do "fewer, clearer, and higher standards" really look like in the classroom? Using a cognitive rigor matrix to analyze curriculum, plan lessons, and implement assessments*. Retrieved from https://www.nciea.org/sites/default/files/publications/cognitiverigorpaper_KH12.pdf

Hernández, R. (2012). Does continuous assessment in higher education support student learning?. *Higher Education*, *64*, 489-502. doi:10.1007/s10734-012-9506-7

Hodgson, P., & Pang, M. Y. C. (2012). Effective formative e-assessment of student learning: A study on a statistics course. *Assessment & Evaluation in Higher Education*, *37*, 215-225. doi:10.1080/02602938.2010.523818

Holmes, V. (2011). Depth of teachers' knowledge: Frameworks for teachers' knowledge of mathematics. *Journal of STEM Education*, *13*(1), 55-71. Retrieved from https://www.jstem.org/index.php/JSTEM

Horton, N. J. (2015). Challenges and opportunities for statistics and statistical education: Looking back, looking forward. *The American Statistician*, *69*, 138-145. doi:10.1080/00031305.2015.1032435

Hubbard, J. K., & Couch, B. A. (2018). The positive effect of in-class clicker questions on later exams depends on initial student performance level but not question format. *Computers & Education*, *120*, 1-12. doi:10.1016/j.compedu.2018.01.008

Hunsu, N. J., Adesope, O., & Bayly, D. J. (2016). A meta-analysis of the effects of audience response systems (clicker-based technologies) on cognition and affect. *Computers & Education*, *94*, 102-119. doi:10.1016/j.compedu.2015.11.013

Katz, L., Hallam M. C., Duvall, M. M., & Polsky, Z. (2017). Considerations for using personal Wi-Fi enabled devices as "clickers" in a large university class. *Active Learning in Higher Education*, *18*, 25-35. doi:10.1177/1469787417693495

Keeler, C., & Steinhorst, K. (2001). A new approach to learning probability in the first statistics course. *Journal of Statistics Education*, *9*, 1-23. doi:10.1080/10691898.2001.11910539

Keengwe, J., Onchwari, G., & Agamba, J. (2014). Promoting effective e-learning practices through the constructivist pedagogy. *Education and Information Technologies*, *19*, 887-898. doi:10.1007/s10639-013-9260-1.

Kibble, J. D. & Johnson, T. (2011). Are faculty predictions or item taxonomies useful for estimating the outcome of multiple-choice examinations?. *Advanced Physiological Education, 35*, 396-401. doi:10.1152/advan.00062.2011

King, D. B. & Joshi, S. (2008). Gender differences in the use and effectiveness of personal response devices. *Journal of Science Education and Technology*, *17*, 544-552. doi:10.1007/s10956-008-9121-7

Klein, K. & Kientz, M. (2013). A model for successful use of student response systems. *Nursing Education Perspectives*, *34*, 334-338. doi:10.5480/1536-5026-34.5.334

Kroning, M. (2014). The importance of integrating active learning in education. *Nurse Education in Practice*, *14*, 447-448. doi:10.1016/j.nepr.2014.06.001

Kulikovshikh, I. M., Prokhorov, S. A., & Suchkova, S. A. (2017). Promoting collaborative learning through regulation of guessing in clickers. *Computers in Human Behavior*, *75*, 81-91. doi:10.1016/j.chb.2017.05.001

Kumar, S., McLean, L., Nash, L., & Trigwell, K. (2017). Incorporating active learning in psychiatry education. *Australasian Psychiatry*, *25*, 304-309. doi:10.1177/1039856217689912

Lantz, M. E. (2010). The use of 'Clickers' in the classroom: Teaching innovation or merely an amusing novelty?. *Computers in Human Behavior*, *26*, 556-561. doi:10.1016/j.chb.2010.02.014

Lantz, M. E., & Stawiski, A. (2014). Effectiveness of clickers: Effect of feedback and the timing of questions on learning. *Computers in Human Behavior, 31*, 280-286. doi:10.1016/j.chb.2013.10.009

Lee, U. J., Sbeglia, G. C., Ha, M., Finch, S. J., & Nehm, R. H. (2015). Clickers score trajectories and concept inventory scores as predictors for early warning systems for large STEM classes. *Journal of Science Education and Technology*, *24*, 848-860. doi:10.1007/s10956-015-9568-2

López-Pastor, V., & Sicilia-Camacho, A. (2017). Formative and shared assessment in higher education: Lessons learned and challenges for the future. *Assessment & Evaluation in Higher Education*, *42*, 77-97. doi:10.1080/02602938.2015.1083535

Lovett, M. C., & Greenhouse, J. B. (2000). Applying cognitive theory to statistics instruction. *The American Statistician*, *54*, 196-206. doi:10.1080/00031305.2000.10474545

Mateo, Z. F. (2010). Creating active learning in a large introductory statistics class using clicker technology. *Proceedings of the 8th International Conference on Teaching Statistics*, Ljubljana, Slovenia. Retrieved from http://icots.info/8/cd/pdfs/invited/ICOTS8_9E4_MATEO.pdf

Mayer, R. E., Stull, A., DeLeeuw, K., Almeroth, K., Bimber, B., Chun, D.,…Knight, A. (2009). Clickers in college classrooms: Fostering learning with questioning methods in large lecture classes. *Contemporary Educational Psychology*, *34*, 51-57. doi:10.1016/j.cedpsych.2008.04.002

Meletiou-Mavrotheris, M., Lee, C., & Fouladi, R. T. (2007). Introductory statistics, college student attitudes and knowledge – A qualitative analysis of the impact of technology-based instruction. *International Journal of Mathematical Education in Science and Technology, 38*, 65-83. doi:10.1080/00207390601002765

Mitchell, A., Petter, S., & Harris, A. L. (2017). Learning by doing: Twenty successful active learning exercises for information systems courses. *Journal of Information Technology Education: Innovations in Practice*, *16*, 21-46. doi:10.28945/3643

Motulsky, H. J. (2014). Common misconceptions about data analysis and statistics. *British Journal of Pharmacology*, 172, 2126-2132. doi:10.1111/bph.12856

Mundform, D. J., Schaffer, J., Kim, M., Shaw, D., & Thongteeraparp, A. (2011). Number of replications required in Monte Carlo simulation studies: A synthesis of four studies. *Journal of Applied Statistical Methods*, *10*, 19-28. doi:10.22237/jmasm/1304222580

Mulder, R. A., Pearce, J. M., & Baik, C. (2014). Peer review in higher education: Student perceptions before and after participation. *Active Learning in Higher Education*, *15*, 157-17. doi:10.1177/1469787414527391

Mvududu, N. (2005). Constructivism in the statistics classroom: From theory to practice. *Teaching Statistics: An International Journal for Teachers, 27*, 49-54. doi:10.1111/j.1467-9639.2005.00208.x

O'Neil Jr., H. F., & Brown, R. S. (1998). Differential effects of question formats in math assessment on metacognition and affect. *Applied Measurement in Education*, *11*, 331-351. doi:10.1207/s15324818ame1104_3

Ozuru, Y., Briner, S., Kurby, C. A., & McNamara, D. S. (2013). Comparing comprehension measured by multiple-choice and open-ended questions. *Canadian Journal of Experimental Psychology*, *67*, 215-227. doi:10.1037/a0032918

Paige D. D., Sizemore, J. M., & Neace, W. P. (2013). Working inside the box: Exploring the relationship between student engagement and cognitive rigor. *NASSP Bulletin*. *97*, 105-123. doi:10.1177/0192636512473505

Palmer, E. J., & Devitt, P. G. (2007). Assessment of higher order cognitive skills in undergraduate education: Modified essay or multiple choice questions? Research paper. *BMC Medical Education*, *7*, 49-56. doi:10.1186/1472-6920-7-49

Park, E. L., & Choi, B. K. (2014). Transformation of college classroom spaces: Traditional versus active learning classrooms in college. *Higher Education*, *68*, 749-771. doi:10.1007/s10734-014-9742-0

Pfenning, N. (2011). *Elementary Statistics: Looking at the Big Picture* (1st ed.). Boston, MA: Cengage Learning.

Poelmans, S., & Wessa, P. (2015). A constructivist approach in a blended e-learning environment for statistics. *Interactive Learning Environments*, *23*, 385-401. doi:10.1080/10494820.2013.766890

Premkumar, K., & Coupal, C. (2008). Rules of engagement–12 tips for successful use of “clickers” in the classroom. *Medical Teacher*, *30*, 146-149. doi:10.1080/01421590801965111

Rana, P. R. & Dwivedi, Y. K. (2016). Using clickers in a large business class: Examining use behavior and satisfaction. *Journal of Marketing Education*, *38*, 47-64. doi:10.1177/0273475315590660

Richardson, A. (2011). Experiences with clickers in an introductory statistics course, *Proceedings of the Fourth Annual ASEARC Conference* (pp. 31-38). Paramatta, Australia: University of Western Sydney.

Roth, K. (2012). Assessing clicker examples versus board examples in calculus. *PRIMUS*, *22*, 353-364. doi:10.1080/10511970.2011.623503

Rush, B. R., Rankin, D. C., & White, B. J. (2016). The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC Medical Education*, *16*, 250-265. doi:10.1186/s12909-016-0773-3

Salcedo, A. (2014). Statistics test questions: Content and trends. *Statistics Education Research Journal*, *13*(2), 202-217. Retrieved from https://iase-web.org/Publications.php?p=SERJ

Sanchez, W. B. (2013). Open-ended questions and the process standards. *The Mathematics Teacher*, *107*, 206-211. doi:10.5951/mathteacher.107.3.0206

Secolsky, C., & Denison, D. B. (2012). *Handbook on measurement, assessment, and evaluation in higher education*. New York, NY: Routledge

Shaffer, D. M., & Collura, M. J. (2009). Evaluating the effectiveness of a personal response system in the classroom. *Teaching of Psychology*, *36*, 273-277. doi:10.1080/00986280903175749

Slavich, G. M., & Zimbardo P. G. (2012). Transformational teaching: Theoretical underpinnings, basic principles, and core methods. *Educational Psychology Review*, *24*, 569-608. doi:10.1007/s10648-012-9199-6.

Son, J. (2012). A cross-national comparison of reform curricula in Korea and the US in terms of cognitive complexity: The case of fraction addition and subtraction. *ZDM – The International Journal on Mathematics Education*, *44*, 161-174. doi:10.1007/s11858-012-0386-1

Stalne, K., Kjellström, S., & Utriainen, J. (2016). Assessing complexity in learning outcomes – A comparison between the SOLO taxonomy and the model of hierarchical complexity.

Assessment & Evaluation in Higher Education, *41*, 1033-1048. doi:10.1080/02602938.2015.1047319

Strangfeld, J. A. (2013). Promoting active learning: Student-led gathering in undergraduate statistics. *Teaching Sociology*, *41*, 199-206. doi:10.1177/0092055X12472492

Sutherlin, A. L., Sutherlin, G. R., & Akpanudo, U. M. (2013). The effect of clickers in university science courses. *Journal of Science Education and Technology*, *22*, 651-666. doi:10.1007/s10956-012-9420-x

Symister, P., VanOra, J., Griffin, K. W., & Troy, D. (2014). Clicking in the community college classroom: Assessing the effectiveness of clickers on student learning in a general psychology course. *The Community College Enterprise*, *20*(2), 10-24. Retrieved from https://www.schoolcraft.edu/cce/community-college-enterprise

Terrion, J. L., & Aceti, V. (2011). Perceptions of the effects of clicker technology on student learning and engagement: A study of freshmen chemistry students. *Research in Learning Technology*, *20*, 16150-16161. doi:10.3402/rlt.v20i0.16150

Tishkovskaya, S., & Lancaster, G. A. (2012). Statistical education in the 21st century: A review of challenges, teaching innovations and strategies for reform. *Journal of Statistics Education*, *20*(2), 1-56. doi:10.1080/10691898.2012.11889641

Titman, A. C. & Lancaster G. A. (2011). Personal response systems for teaching postgraduate statistics to small groups. *Journal of Statistics Education*, *19*(2). doi:10.1080/10691898.2011.11889614

Tomei, L. A. (2013). Top 10 technologies for designing 21st century instruction. *International Journal of Information and Communication Technology Education*, *9*, 80-93. doi:10.4018/jicte.2013070106

Tregonning, A. M., Doherty, D. A., Hornbuckle, J., & Dickinson, J. E. (2012). The audience response system and knowledge gain: A prospective study. *Medical Teacher*, *34*, doi:10.3109/0142159X.2012.660218

Trumbull, E., & Lash, A. (2013). Understanding formative assessment: Insights from learning theory and measurement theory. San Francisco, CA: WestEd. Retrieved from https://www.wested.org/online_pubs/resource1307.pdf

U.S. Department of Education, Office of Educational Technology. (2017). *Reimagining the Role of Technology in Education: 2017 National Education Technology Plan Update*. Retrieved from https://tech.ed.gov/files/2017/01/NETP17.pdf

Utts, J. (2003). What educated citizens should know about statistics and probability. *The American Statistician*, *57*, 74-79. doi:10.1198/0003130031630

Van Bergen, P., & Parsell, M. (2019). Comparing radical, social, and psychological constructivism in Australian higher education: A psycho-philosophical perspective. *The Australian Educational Researcher*, *46*, 41-58. doi:10.1007/s13384-018-0285-8

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education*, *6*, 103-118. doi:10.1207/s15324818ame0602_1

Webb, N. L. (2002). *Depth-of-knowledge levels for four content areas*. Retrieved from http://facstaff.wcer.wisc.edu/normw/All%20content%20areas%20%20DOK%20levels%2032802.pdf

Weltman, D., & Whiteside, M. (2010). Comparing the effectiveness of traditional and active learning methods in business statistics: Convergence to the mean. *Journal of Statistics Education*, *18*, 1-13. doi:10.1080/10691898.2010.11889480

Westerman, J. W., Perez-Batres, L. A., Coffey, B. S., & Pouder, R. W. (2011). The relationship between undergraduate attendance and performance revisited: Alignment of student and instructor goals. *Decision Sciences Journal of Innovative Education*, *9*, 49-62. doi:10.1111/j.1540-4609.2010.00294.x

White, P., Syncox, D., & Alters, B. (2011). Clicking for grades? Really? Investigating the use of clickers for awarding grade-points in post-secondary education. *Interactive Learning Environments*, *19*, doi:10.1080/10494821003612638

Wyse, A. E., & Viger, S. G. (2011). How item writers understand depth of knowledge. *Educational Assessment*, *16*, 185-206. doi:10.1080/10627197.2011.634286

Woodard, R., & McGowan, H. (2012). Redesigning a large introductory course to incorporate the GAISE guidelines. *Journal of Statistics Education*, *20*(3), 1-24. doi:10.1080/10691898.2012.1889650

Yeo, C. H., Ke, K., Chatterjee, B. (2015). An investigation into the relationship between on-line formative assessments and performance of students. *e-Journal of Business Education & Scholarship of Teaching*, *9*(1), 1-42. Retrieved from http://www.ejbest.org/

Yourstone, S. A., Kraye, H. S., & Albaum, G. (2008). Classroom questioning with immediate electronic response: Do clickers improve learning?. *Decision Sciences Journal of Innovative Education*, *6*, 75-88. doi:10.1111/j.1540-4609.2007.00166.x

Zingaro, D. (2015). Examining interest and grades in computer science 1: A study of pedagogy and achievement goals. *ACM Transactions on Computing Education*, *15*, 1-18. doi:10.1145/2802752

Appendix A

List of Topics and Dates Covered

Table A.1

List of Topics and Dates Covered

Lecture	Date	Topic	Due date (Lectures)
1	January 10	Variable Types and Sampling Methods	
2	January 12	Surveys and Observational Studies	
3	January 17	Experiments	
4	January 19	Displaying and Summarizing Categorical Data	
5	January 22	Comparing Categorical Variables	Quiz #1 Due (1-3)
6	January 24	Displaying and Summarizing Quantitative Data I	
7	January 26	Displaying and Summarizing Quantitative Data II	
8	January 29	Comparing Quantitative Variables	Quiz #2 Due (4-5)
9	January 31	Basics of Probability	
10	February 2	More Probability Rules	
11	February 5	Random Variables	Quiz #3 Due (6-8)
12	February 9	Binomial Distribution	
13	February 12	Calculating Normal Probabilities	Quiz #4 Due (9-11)
14	February 14	Finding Normal Distribution Percentiles	
15	February 16	Sampling Distribution of the Sample Mean	
16	February 19	Sampling Distribution of the Sample Proportion	Quiz #5 Due (12-14)
17	February 21	Confidence Intervals	
18	February 23	Hypothesis Testing I	Quiz #6 Due (15-16)
19	March 2	Hypothesis Testing II	
20	March 12	Hypothesis Testing III	Quiz #7 Due (17-19)
21	March 14	Inference for a Population Mean	
22	March 16	Inference for a Population Proportion	
23	March 19	Goodness of Fit Test	Quiz #8 Due (20-22)
24	March 21	Matched Pairs Test	
25	March 23	Inference for the Difference of Two Means	Quiz #9 Due (22-23)
26	March 30	Analysis of Variance (ANOVA)	
27	April 2	Multiple Comparisons	Quiz #10 Due (24-25)
28	April 4	Inference for the Difference of Two Proportions	
29	April 6	Chi-Squared Test for Independence	
30	April 9	Introduction to Simple Linear Regression	Quiz #11 Due (26-29)
31	April 11	Assessing the Linear Regression Model	
32	April 13	Prediction Intervals and Confidence Intervals	
33	April 16	Multiple Linear Regression	
34	April 18	Lying with Statistics	Quiz #12 Due (30-33)

Appendix B

Table of Sequences and Concepts

Table B.1

Table of Sequences and Concepts

Question sequence	Concept	Webb's Depth of Knowledge	SRS question location	Quiz question location	Exam question location
1	Identify variable situation	Recall	Lecture 1 Question 1	Quiz 1 Question 1	Exam 1 Question 1
2	Identify population	Recall	Lecture 1 Question 2	Quiz 1 Question 2	Exam 1 Question 2
3	Identify statistic and parameter	Recall	Lecture 4 Question 1	Quiz 1 Question 4	Exam 1 Question 3
4	Identify sampling method	Recall	Lecture 1 Question 3	Quiz 1 Question 5	Exam 1 Question 4
5	Identify error in data collection	Recall	Lecture 1 Question 4	Quiz 1 Question 3	Exam 1 Question 5
6	Identify issue in survey question	Recall	Lecture 2 Question 1	Quiz 1 Question 6	Exam 1 Question 6
7	Identify type of experimental design	Recall	Lecture 3 Question 3	Quiz 2 Question 2	Exam 1 Question 7
8	Identify ethical standard in experiment	Recall	Lecture 3 Question 4	Quiz 2 Question 3	Exam 1 Question 8
9	Determine appropriate assignment of subjects in experiment	Basic Application of Skill	Lecture 3 Question 2	Quiz 2 Question 5	Exam 1 Question 9
10	Identify control group in experiment	Recall	Lecture 3 Question 1	Quiz 2 Question 4	Exam 1 Question 10
11	Determine appropriate sampling method	Strategic Thinking	Lecture 4 Question 2	Quiz 2 Question 6	Exam 1 Question 11
12	Identify data collection method with explanatory and response variables	Basic Application of Skill	Lecture 2 Question 2	Quiz 2 Question 1	Exam 1 Question 12
13	Identify correct type of graph to use to display data	Recall	Lecture 8 Question 5	Quiz 3 Question 1	Exam 1 Question 13
14	Use pie chart to calculate number of observations in group	Basic Application of Skill	Lecture 4 Question 3	Quiz 2 Question 7	Exam 1 Question 14
15	Calculate conditional proportion from cross-classification table	Basic Application of Skill	Lecture 5 Question 1	Quiz 2 Question 8	Exam 1 Question 15

16	Calculate medians of ordinal data and determine which is larger	Strategic Thinking	Lecture 4 Question 4	Quiz 2 Question 9	Exam 1 Question 16
17	Determine largest conditional proportion of four groups based on grouped bar graph	Basic Application of Skill	Lecture 5 Question 2	Quiz 2 Question 10	Exam 1 Question 17
18	Evaluate the truth of three statements about stacked bar graph/mosaic plot	Strategic Thinking	Lecture 5 Question 3	Quiz 2 Question 11	Exam 1 Question 18
19	Calculate proportion of observations in a range using a histogram	Basic Application of Skill	Lecture 6 Question 1	Quiz 3 Question 2	Exam 1 Question 19
20	Calculate mean and standard deviation using Empirical Rule	Basic Application of Skill	Lecture 6 Question 2	Quiz 3 Question 3	Exam 1 Question 20
21	Use Empirical Rule to determine truth of four statements	Basic Application of Skill	Lecture 6 Question 3	Quiz 3 Question 4	Exam 1 Question 21
22	Determine outliers using 1.5*IQR rule	Basic Application of Skill	Lecture 7 Question 1	Quiz 3 Question 5	Exam 1 Question 22
23	Evaluate truth of three statements about percentiles and boxplots	Strategic Thinking	Lecture 7 Question 3	Quiz 3 Question 6	Exam 1 Question 23
24	Determine statistics that would change if a single observation in a dataset were altered	Strategic Thinking	Lecture 7 Question 2	Quiz 3 Question 7	Exam 1 Question 24
25	Determine skewness of a histogram based on a boxplot	Recall	Lecture 7 Question 4	Quiz 3 Question 8	Exam 1 Question 25
26	Evaluate truth of three statements about side-by-side boxplots	Strategic Thinking	Lecture 7 Question 5	Quiz 3 Question 9	Exam 1 Question 26
27	Calculate correlation and classify strength of relationship	Recall	Lecture 8 Question 1	Quiz 3 Question 10	Exam 1 Question 27
28	Order scatterplots from weakest to strongest correlation	Basic Application of Skill	Lecture 8 Question 2	Quiz 3 Question 11	Exam 1 Question 28
29	Evaluate truth of three statements in simple linear regression	Strategic Thinking	Lecture 8 Question 4	Quiz 3 Question 12	Exam 1 Question 29
30	Calculate predicted value using linear regression line	Basic Application of Skill	Lecture 8 Question 3	Quiz 3 Question 13	Exam 1 Question 30
31	Identify if events are intersections	Recall	Lecture 9 Question 1	Quiz 4 Question 1	Exam 2 Question 1

32	Calculate conditional probability	Recall	Lecture 10 Question 1	Quiz 4 Question 2	Exam 2 Question 2
33	Calculate joint probability using a table of probabilities	Basic Application of Skill	Lecture 10 Question 2	Quiz 4 Question 3	Exam 2 Question 3
34	Calculate probability of a union using a table of probabilities	Basic Application of Skill	Lecture 9 Question 2	Quiz 4 Question 4	Exam 2 Question 4
35	Determine if events are independent and/or mutually exclusive	Recall	Lecture 10 Question 4	Quiz 4 Question 5	Exam 2 Question 5
36	Calculate probability of an intersection using General Multiplication Rule	Basic Application of Skill	Lecture 10 Question 3	Quiz 4 Question 6	Exam 2 Question 6
37	Calculate joint probability using probability distribution of two random variables	Basic Application of Skill	Lecture 11 Question 1	Quiz 4 Question 7	Exam 2 Question 7
38	Compare mean and standard deviation of two probability distributions	Strategic Thinking	Lecture 11 Question 3	Quiz 4 Question 8	Exam 2 Question 8
39	Calculate mean and standard deviation of a transformation	Basic Application of Skill	Lecture 11 Question 2	Quiz 4 Question 9	Exam 2 Question 9
40	Calculate binomial distribution probability	Strategic Thinking	Lecture 12 Question 1	Quiz 5 Question 1	Exam 2 Question 10
41	Use binomial distribution to determine how unusual an event is	Strategic Thinking	Lecture 12 Question 2	Quiz 5 Question 2	Exam 2 Question 11
42	Order observations using standardized scores	Recall	Lecture 13 Question 1	Quiz 5 Question 3	Exam 2 Question 12
43	Calculate normal distribution probability	Basic Application of Skill	Lecture 14 Question 1	Quiz 5 Question 4	Exam 2 Question 13
44	Unstandardize a normal distribution percentile	Basic Application of Skill	Lecture 14 Question 2	Quiz 5 Question 5	Exam 2 Question 14
45	Convert an observation between two different normal distributions	Basic Application of Skill	Lecture 14 Question 3	Quiz 5 Question 6	Exam 2 Question 15
46	Determine sampling distribution of a sample mean	Strategic Thinking	Lecture 15 Question 1	Quiz 6 Question 1	Exam 2 Question 16
47	Calculate probability of a sample mean	Basic Application of Skill	Lecture 15 Question 3	Quiz 6 Question 2	Exam 2 Question 17
48	Determine if changes to the sampling distribution of a mean	Strategic Thinking	Lecture 15 Question 2	Quiz 6 Question 3	Exam 2 Question 18

	increase or decrease a probability				
49	Determine sampling distribution of a sample proportion	Strategic Thinking	Lecture 16 Question 1	Quiz 6 Question 4	Exam 2 Question 19
50	Calculate probability of a sample proportion	Basic Application of Skill	Lecture 16 Question 2	Quiz 6 Question 5	Exam 2 Question 20
51	Identify correct form for a confidence interval for a population mean	Recall	Lecture 17 Question 1	Quiz 7 Question 1	Exam 3 Question 1
52	Use a confidence interval for a population mean to evaluate a hypothesis	Basic Application of Skill	Lecture 19 Question 3	Quiz 7 Question 2	Exam 3 Question 2
53	Determine if changes to a confidence interval will change the width	Strategic Thinking	Lecture 17 Question 2	Quiz 7 Question 3	Exam 3 Question 3
54	Determine how many confidence intervals will contain the true population mean	Basic Application of Skill	Lecture 17 Question 3	Quiz 7 Question 4	Exam 3 Question 4
55	Determine necessary sample size to attain confidence interval width	Basic Application of Skill	Lecture 17 Question 4	Quiz 7 Question 5	Exam 3 Question 5
56	Identify critical value(s) of a Z-test	Recall	Lecture 19 Question 2	Quiz 7 Question 6	Exam 3 Question 6
57	Calculate p-value of a Z-test	Basic Application of Skill	Lecture 19 Question 1	Quiz 7 Question 7	Exam 3 Question 7
58	Determine conclusion of Z-test given partial output	Strategic Thinking	Lecture 18 Question 3	Quiz 7 Question 8	Exam 3 Question 8
59	Determine result of making a Type II error	Strategic Thinking	Lecture 20 Question 2	Quiz 7 Question 9	Exam 3 Question 9
60	Determine if a confidence interval would contain two specified values based on result of Z-test	Strategic Thinking	Lecture 19 Question 4	Quiz 7 Question 10	Exam 3 Question 10
61	Determine hypotheses in a t-test	Basic Application of Skill	Lecture 21 Question 1	Quiz 8 Question 1	Exam 3 Question 11
62	Determine if conditions for a t-test are satisfied	Basic Application of Skill	Lecture 21 Question 2	Quiz 8 Question 2	Exam 3 Question 12
63	Identify correct test statistic for a t-test	Recall	Lecture 21 Question 3	Quiz 8 Question 3	Exam 3 Question 13

64	Determine if changes to a t-test would lead to a smaller p-value	Strategic Thinking	Lecture 21 Question 5	Quiz 8 Question 4	Exam 3 Question 14
65	Determine levels of significance for which the null hypothesis would be rejected	Basic Application of Skill	Lecture 18 Question 1	Quiz 8 Question 5	Exam 3 Question 15
66	Determine differences between Z-test and t-test	Strategic Thinking	Lecture 21 Question 4	Quiz 8 Question 6	Exam 3 Question 16
67	Interpret confidence interval for a population proportion	Basic Application of Skill	Lecture 22 Question 3	Quiz 9 Question 1	Exam 3 Question 17
68	Determine if width of a confidence interval for a population proportion would change	Basic Application of Skill	Lecture 22 Question 6	Quiz 9 Question 2	Exam 3 Question 18
69	Evaluate truth of three statements about a confidence interval for a population proportion	Strategic Thinking	Lecture 22 Question 5	Quiz 9 Question 3	Exam 3 Question 19
70	Compare confidence intervals for a population proportion from two different samples	Strategic Thinking	Lecture 22 Question 4	Quiz 9 Question 4	Exam 3 Question 20
71	Identify hypothesis for a one-sample proportion test	Recall	Lecture 22 Question 1	Quiz 8 Question 7	Exam 3 Question 21
72	Interpret p-value of a hypothesis test	Strategic Thinking	Lecture 18 Question 2	Quiz 8 Question 8	Exam 3 Question 22
73	Determine decision and conclusion of a one-sample proportion test	Basic Application of Skill	Lecture 22 Question 2	Quiz 8 Question 9	Exam 3 Question 23
74	Determine type of decision made in a hypothesis test	Basic Application of Skill	Lecture 20 Question 2	Quiz 8 Question 10	Exam 3 Question 24
75	Determine p-value of two-sided test given p-value of one-sided test	Recall	Lecture 20 Question 3	Quiz 8 Question 11	Exam 3 Question 25
76	Identify probability of making a Type I error	Recall	Lecture 23 Question 4	Quiz 8 Question 12	Exam 3 Question 26
77	Identify hypotheses in goodness of fit test	Basic Application of Skill	Lecture 23 Question 1	Quiz 9 Question 5	Exam 3 Question 27
78	Calculate chi-squared contribution to test statistic for a category in a goodness of fit test	Basic Application of Skill	Lecture 23 Question 5	Quiz 9 Question 6	Exam 3 Question 28
79	Calculate degrees of freedom in a goodness of fit test	Recall	Lecture 23 Question 2	Quiz 9 Question 7	Exam 3 Question 29

80	Determine conclusion of goodness of fit test	Strategic Thinking	Lecture 23 Question 3	Quiz 9 Question 8	Exam 3 Question 30
81	Identify hypotheses in difference of two means test	Recall	Lecture 25 Question 1	Quiz 10 Question 4	Exam 4 Question 13
82	Identify test statistic in difference of two means test	Recall	Lecture 25 Question 2	Quiz 10 Question 5	Exam 4 Question 14
83	Determine conclusion of differences of two means test	Basic Application of Skill	Lecture 25 Question 4	Quiz 10 Question 6	Exam 4 Question 15
84	Determine if a confidence interval would contain two specified values based on result of difference of two means test	Strategic Thinking	Lecture 25 Question 6	Quiz 10 Question 7	Exam 4 Question 16
85	Determine if changes to a difference of two means test would lead to a smaller p-value	Strategic Thinking	Lecture 25 Question 5	Quiz 10 Question 8	Exam 4 Question 17
86	Determine changes to hypothesis test if populations are reversed in difference of two means test	Strategic Thinking	Lecture 25 Question 3	Quiz 10 Question 9	Exam 4 Question 18
87	Calculate degrees of freedom in ANOVA	Recall	Lecture 26 Question 1	Quiz 11 Question 1	Exam 4 Question 21
88	Determine conclusion of ANOVA test given output	Basic Application of Skill	Lecture 26 Question 2	Quiz 11 Question 2	Exam 4 Question 22
89	Determine if changes to an ANOVA test would lead to a smaller p-value	Strategic Thinking	Lecture 27 Question 2	Quiz 11 Question 3	Exam 4 Question 23
90	Determine group means that are significantly different using multiple comparisons confidence intervals	Basic Application of Skill	Lecture 27 Question 1	Quiz 11 Question 4	Exam 4 Question 24
91	Identify variables and dependency in a matched pairs test	Basic Application of Skill	Lecture 24 Question 1	Quiz 10 Question 1	Exam 4 Question 26
92	Identify correct ways to write hypotheses in matched pairs test	Recall	Lecture 24 Question 2	Quiz 10 Question 2	Exam 4 Question 27
93	Determine conclusion in matched pairs test from an output	Basic Application of Skill	Lecture 24 Question 3	Quiz 10 Question 3	Exam 4 Question 28
94	Determine conclusion in difference of two proportions test from an output	Basic Application of Skill	Lecture 28 Question 1	Quiz 11 Question 5	Exam 4 Question 30
95	Evaluate the truth of three statements about a confidence	Strategic Thinking	Lecture 28 Question 2	Quiz 11 Question 6	Exam 4 Question 32

	interval for the difference of two proportions				
96	Determine hypotheses in test for independence	Basic Application of Skill	Lecture 29 Question 3	Quiz 11 Question 7	Exam 4 Question 34
97	Calculate degrees of freedom in test for independence	Recall	Lecture 29 Question 2	Quiz 11 Question 8	Exam 4 Question 35
98	Calculate chi-squared contribution to test statistic for a cell in a test for independence	Basic Application of Skill	Lecture 29 Question 1	Quiz 11 Question 9	Exam 4 Question 36
99	Determine conclusion of test for independence given output	Strategic Thinking	Lecture 29 Question 4	Quiz 11 Question 10	Exam 4 Question 37
100	Classify a point as being an outlier and/or influential point on scatterplot	Basic Application of Skill	Lecture 30 Question 3	Quiz 12 Question 1	Exam 4 Question 38
101	Describe strength of linear relationship given equation of the regression line and coefficient of determination	Strategic Thinking	Lecture 30 Question 1	Quiz 12 Question 2	Exam 4 Question 39
102	Identify form of linear model	Recall	Lecture 31 Question 1	Quiz 12 Question 3	Exam 4 Question 40
103	Calculate residual in simple linear regression	Basic Application of Skill	Lecture 30 Question 2	Quiz 12 Question 4	Exam 4 Question 41
104	Identify hypotheses for testing slope in simple linear regression	Recall	Lecture 31 Question 2	Quiz 12 Question 5	Exam 4 Question 42
105	Determine conclusion about relationship in simple linear regression	Basic Application of Skill	Lecture 31 Question 3	Quiz 12 Question 6	Exam 4 Question 43
106	Determine if conditions are satisfied in simple linear regression	Basic Application of Skill	Lecture 31 Question 4	Quiz 12 Question 7	Exam 4 Question 44
107	Identify correct interpretation of prediction interval in simple linear regression	Basic Application of Skill	Lecture 32 Question 1	Quiz 12 Question 8	Exam 4 Question 45
108	Identify ways in which a confidence or prediction interval in regression can become wider	Strategic Thinking	Lecture 32 Question 2	Quiz 12 Question 9	Exam 4 Question 46
109	Identify hypotheses to test multiple linear regression model	Recall	Lecture 33 Question 2	Quiz 12 Question 10	Exam 4 Question 47
110	Determine if any predictors are significant in multiple linear regression model	Basic Application of Skill	Lecture 33 Question 3	Quiz 12 Question 11	Exam 4 Question 48

111	Determine specific predictors in multiple linear regression that are significant and have positive or negative relationships with the response	Basic Application of Skill	Lecture 33 Question 4	Quiz 12 Question 12	Exam 4 Question 49
112	Interpret slope coefficient in multiple linear regression model	Basic Application of Skill	Lecture 33 Question 1	Quiz 12 Question 13	Exam 4 Question 50

Appendix C

Student Response System Questions

Lecture 1

1. “10% of the world’s population is left-handed.” Which variable situation is reflected in this statement?

   a. One categorical
   b. One quantitative
   c. Two categorical
   d. One categorical and one quantitative
   e. Two quantitative

2. 500 people were selected from a list of registered voters in Allegheny County. 40% of those sampled were registered Independents. What is the population in this study?

   a. All adults in Allegheny County
   b. The 500 people selected from the list of registered voters
   c. The 200 Independents selected from the list of registered voters
   d. All registered voters in Allegheny County
   e. All registered Independents in Allegheny County

3. Gallup wants to gauge Trump’s approval rating. They randomly sample 1000 registered voters from a list of voters in the 2016 election. What type of sampling method was used?

   a. Simple random sample
   b. Stratified random sample
   c. Systematic sample
   d. Convenience sample
   e. Voluntary sample

4. Residents of Oakland are left off of the list of potential candidates for jury duty. What type of error is this?

   a. Sampling error
   b. Data acquisition error
   c. Nonresponse bias
   d. Selection bias

Lecture 2

1. A closed survey question asked respondents, “Are you in favor of repealing Obamacare and replacing it with a new health care bill?” with the options of “Yes” and “No”. What type of issue exists with this survey question?

a. Complicated question
b. Vague question
c. Leading question
d. Central tendency bias
e. Error prone response options

2. Suppose we are looking to identify if a relationship exists between if a person has lung cancer and if a person smokes. We do the study by recruiting a group of smokers and a group of nonsmokers and following them over the next 10 years. Which of the following best describes this study?

   a. Retrospective observational study where lung cancer status is the explanatory variable and smoking status is the response variable
   b. Retrospective observational study where smoking status is the explanatory variable and lung cancer status is the response variable
   c. Prospective observational study where lung cancer status is the explanatory variable and smoking status is the response variable
   d. Prospective observational study where smoking status is the explanatory variable and lung cancer status is the response variable

Lecture 3

1. We want to study the effects of a drug that cures depression. A sugar pill will be given to the subjects not assigned to the drug. 30 subjects with no major characteristic differences are recruited. They rate their depression on a scale from 1-10 at the beginning and end of the study. Which group is acting as the control?

   a. Subjects assigned the drug
   b. Subjects assigned the sugar pill
   c. Subjects who see an improvement in depression
   d. Subjects who see no change in their depression

2. A history professor wants to see if students do better on the final exam if they do a project or write a paper. She believes that underclassmen and upperclassmen may learn differently. Of the 100 students in the class, 70 are underclassmen and 30 are upperclassmen. How should subjects be assigned to treatments?

   a. Everyone does both a paper and a project
   b. Underclassmen do a project; upperclassmen do a paper
   c. 35 random underclassmen and 15 random upperclassmen do a project; other 35 underclassmen and 15 upperclassmen write a paper
   d. 50 randomly selected students do a project; remaining 50 students write a paper

3. We want to study the reaction times of subjects after they have been awake for 6, 12, or 18 hours. Each subject is assigned to stay awake for one of the assigned periods of time and

then take a reaction time test. 60 people are recruited with no supposed differences in any major characteristics. What type of experimental design should be used?

a. Matched pairs
b. Completely randomized design
c. Block design

4. Facebook manipulated the news feeds of 700,000 randomly selected users for one week to change how many positive and negative posts they saw. The goal was to determine if users' emotions changed during the study by comparing the content they saw and the posts they wrote. Facebook published a journal article before without notifying any users that they had been studied. What ethical standard was violated in the Facebook experiment?

   a. Autonomy
   b. Beneficence
   c. Confidentiality
   d. Deception
   e. Informed consent

Lecture 4

1. Survey of 796 college students found that 288 of them reported binge drinking at some point in the past month. How should the proportion .362 be denoted? How should the overall proportion of all college students who binge drink be denoted?

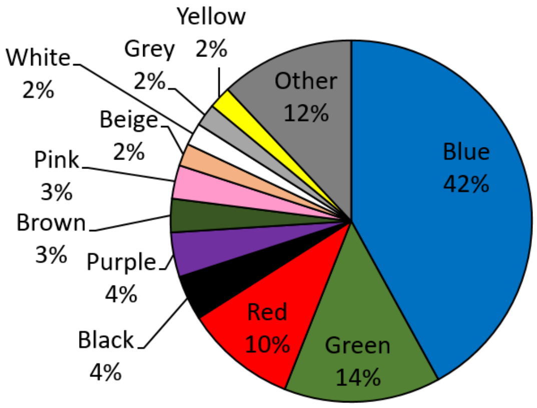
   a. Sample Proportion: $\hat{p}$; Population Proportion: $p = .362$
   b. Sample Proportion: p; Population Proportion: $\hat{p} = .362$
   c. Sample Proportion: $\hat{p} = .362$; Population Proportion: p
   d. Sample Proportion: $p = .362$; Population Proportion: $\hat{p}$

2. Three students perform studies on the proportion of people who wear glasses.
   - A: Observes 21 of 30 people (70%) on the morning bus ride downtown wearing glasses
   - B: Surveys 300 people downtown during the day and finds 200 (66.7%) who respond they wear glasses
   - C: Surveys 30 people downtown during the day and finds 21 (70%) who respond they wear glasses

   Is one student's data more convincing that a majority of people wear glasses?

   a. Yes: A's is most convincing
   b. Yes: B's is most convincing
   c. Yes: C's is most convincing
   d. Yes: Both A and C are equally convincing
   e. No: All three samples are equally convincing

3. A survey of adults asks for their favorite color with the responses contained in the pie chart below. If 630 people responded that blue was their favorite color, how many chose red?



a. 26
b. 150
c. 264
d. 1,500
e. 2,646
f. 15,000

4. Patients are asked to rate the amount of pain they are in. The tables below compare patients with sprains with those with broken bones. What can we say about the medians?

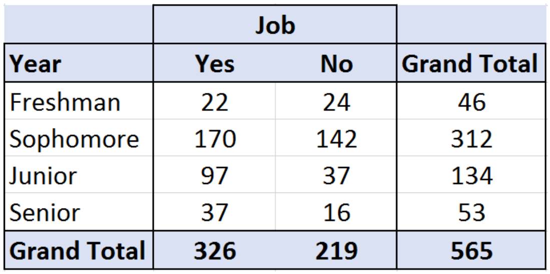
Broken Bones	Percentage
Little/no pain	10%
Moderate pain	45%
Severe pain	30%
Acute pain	15%

Sprains	Percentage
Little/no pain	20%
Moderate pain	35%
Severe pain	40%
Acute pain	5%

a. Median for broken bones is greater
b. Median for sprains is greater
c. Medians are the same
d. Medians cannot be computed for either

Lecture 5

1. The table below compares year in school against if a student currently has a job. How should we calculate the proportion of freshmen who have jobs?



	Job		
Year	Yes	No	Grand Total
Freshman	22	24	46
Sophomore	170	142	312
Junior	97	37	134
Senior	37	16	53
Grand Total	**326**	**219**	**565**

a. $\frac{22}{326} = .0675$

b. $\frac{22}{565} = .0389$

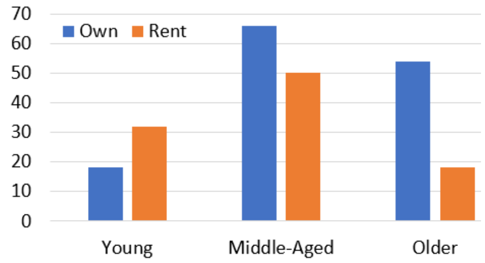
c. $\frac{46}{565} = .0814$

d. $\frac{46}{326} = .1411$

e. $\frac{22}{46} = .4783$

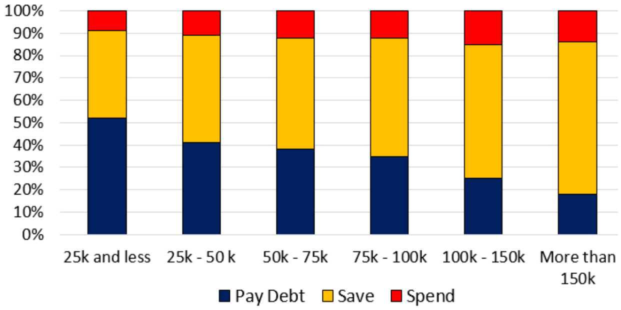
f. $\frac{326}{565} = .5770$

2. The graph below compares the age bracket of a sample of adults with if they own or rent their home. Which conditional proportion is the largest?



a. Young adults who own their home
b. Young adults who rent
c. Middle-aged adults who own their home
d. Middle-aged adults who rent
e. Older adults who own their home
f. Older adults who rent

3. The graph below compared income level and intention on how to use tax returns based on a *Google Consumer* survey. Which of the following are patterns that exist?

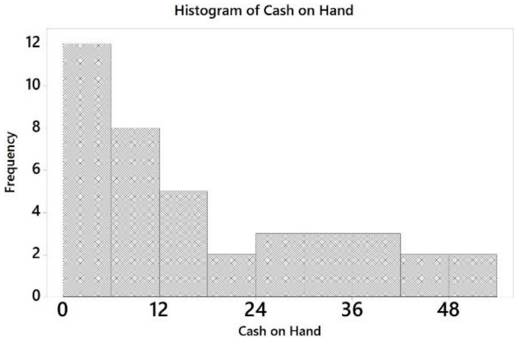


I. People of all incomes are about equally likely to spend
II. Higher incomes are more likely to pay debt
III. Lower incomes are more likely to save

a. None are true
b. I only
c. II only
d. III only
e. I and II
f. I and III
g. II and III
h. I, II, and III

Lecture 6

1. The histogram below shows the amount of money that a random sample of 40 people were carrying on them. What percentage of people were carrying between $18 and $36?



a. $\frac{3}{40} = .075$
b. $\frac{8}{40} = .20$
c. $\frac{13}{40} = .325$
d. $\frac{16}{40} = .40$

2. Suppose ACT scores follow a normal distribution where about 95% of students score between 11 and 31. What are the mean and standard deviation of ACT scores?

a. Mean is 21 and standard deviation is 5
b. Mean is 21 and standard deviation is 10
c. Mean is 21 and standard deviation is 15
d. Mean is 42 and standard deviation is 5
e. Mean is 42 and standard deviation is 10
f. Mean is 42 and standard deviation is 15

3. IQ scores follow a normal distribution with a mean of 100 and standard deviation of 15. Which of the following is not true?

   a. About 32% of people have IQs above 115.
   b. A person with an IQ of 147 is considered highly unusual.
   c. A person with an IQ of 90 has a negative Z-score
   d. A person whose IQ is 80 is more standard deviations from the mean than someone whose IQ is 105.

Lecture 7

1. A random sample of 10 exam scores from a large class returned a five-number summary of 32, 68, 79, 88, 97. Are there any outliers?

   a. No
   b. Yes: 32 is the only outlier
   c. Yes; 97 is the only outlier
   d. Yes: 32 and 97 are both outliers

2. If the student who scored 97 had their score inputted as 87 instead, what would have happened to the mean, median, and standard deviation?

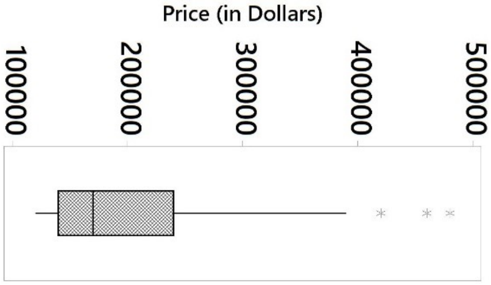
   a. Mean increases; Median decreases; No change to standard deviation
   b. Mean increases; No change to median; Standard deviation decreases
   c. Mean and median decrease; No change to standard deviation
   d. Mean and median decrease; Standard deviation decreases
   e. Mean and median decrease; Standard deviation increases
   f. Mean and standard deviation decrease; No change to median
   g. Mean decreases; No change to median; Standard deviation increases
   h. Mean, median, and standard deviation all decrease

3. The number of home runs hit by qualified baseball players in MLB last year had five-number summary: 3, 14, 22, 30, 47. If a player's home runs were in the 80th percentile, what can we conclude?

   I. He hit between 22 and 30 home runs.
   II. He hit fewer home runs than 20% of all other players.
   III. The value is in the lower whisker.

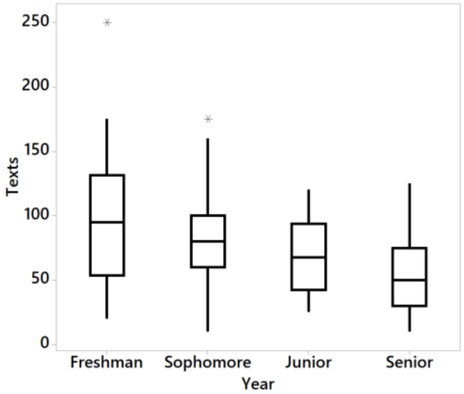
a. None are true
b. I only
c. II only
d. III only
e. I and II
f. I and III
g. II and III
h. I, II, and III

4. A random sample of 90 houses in a city is taken. The assessed values are plotted on a boxplot. What can we deduce about the shape of the histogram and the relationship between the mean and median?



a. Right-skewed histogram with mean > median
b. Right-skewed histogram with median < mean
c. Left-skewed histogram with mean > median
d. Left-skewed histogram with mean < median
e. Symmetric histogram with mean ≈ median

5. A random sample of college students compared the number of texts that students send and receive each day with the year in school they are in. The side-by-side boxplots of the results are below. Which of the following statements are true?



I. Freshmen tend to send and receive the most texts.
II. Sophomores have the smallest spread for the number of texts sent and received.
III. Students tend to send and receive fewer texts as they get older.

a. None are true
b. I only
c. II only
d. III only
e. I and II
f. I and III
g. II and III
h. I, II, and III

Lecture 8

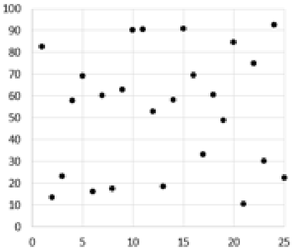
1. Suppose we want to compare students' midterm and final exam scores. The covariance and sample standard deviations are presented in the table below. What type of linear relationship exists between the variables?

Covariance	$s_{MIDTERM}$	s_{FINAL}
53.36	7.34	7.90

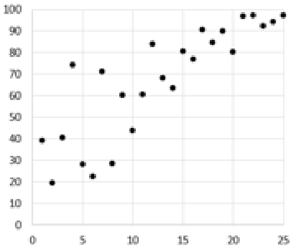
a. Strong negative linear relationship
b. Moderate negative linear relationship
c. Weak negative linear relationship
d. No linear relationship
e. Weak positive linear relationship
f. Moderate positive linear relationship
g. Strong positive linear relationship

2. The scatterplots below show observations with predictors from 1 to 25 and responses from 0 to 100. Order the scatterplots from weakest correlation to strongest.

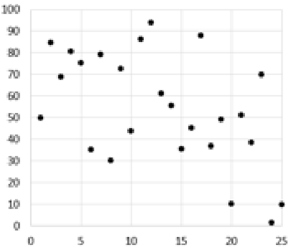
A.



B.



C.



a. A, B, C
b. A, C, B
c. B, A, C
d. B, C, A
e. C, A, B
f. C, B, A

3. The regression line for predicting the sales price of a used car based on its age (in years) is $\hat{y} = 11{,}725 - 351.74x$. How much would we expect to pay for a used car that is 7 years old?

a. $2462,18
b. $9262.82
c. $11,373.26
d. $12,076.74
e. $14,187.18

4. The regression line for predicting the sales price of a used car based on its age (in years) is $\hat{y} = 11{,}725 - 351.74x$. Which of the following statements are true?

I. There is a strong, negative linear relationship.
II. A car that was 2 years old a sold for $4,000 is close to what we would expect.
III. Based on the regression line, a car that is 8 years old would sell for more than a car that is 5 years old.

a. None are true
b. I only
c. II only
d. III only
e. I and II
f. I and III
g. II and III
h. I, II, and III

5. We take a random sample of high school students and ask survey each to see if they play an instrument, play a sport, or participate in a club. What type of graph is best to display the results of the survey?

a. Bar graph
b. Pie chart
c. Histogram
d. Side-by-side boxplots
e. Scatterplot

Lecture 9

1. Randomly sample two people (X and Y) and ask if they use a PC or a Mac. Which of the following events are intersections?

I. Both people use PCs
II. At least one of the two people uses a PC
III. Neither person uses a PC

a. None are true
b. I only
c. II only

d. III only
e. I and II
f. I and III
g. II and III
h. I, II, and III

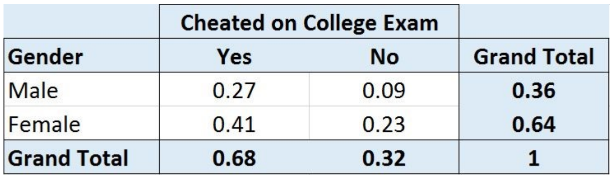
2. Probabilities of freshmen taking math or English classes in their first semester are displayed in the table below. What is the probability a freshman has to take either a math or an English class?

	English	No English	**Total**
Math	.30	.15	**.45**
No Math	.45	.10	**.55**
Total	**.75**	**.25**	**1.00**

a. .30
b. .45
c. .75
d. .90
e. 1.20

Lecture 10

1. An anonymous survey was given to college students about if they had cheated on an exam. The probabilities are displayed in the table below. Given that a person has not cheated, what is the probability she is female?



	Cheated on College Exam		
Gender	**Yes**	**No**	**Grand Total**
Male	0.27	0.09	**0.36**
Female	0.41	0.23	**0.64**
Grand Total	**0.68**	**0.32**	**1**

a. $\frac{23}{1.00} = .23$
b. $\frac{.23}{.64}$ =.359
c. $\frac{.32}{.64} = .50$
d. $\frac{.23}{.41} = .561$
e. $\frac{.23}{.32} = .712$

2. We want to compare a person's highest level of education with their self-described political ideology. The table of probabilities is presented below. What is the probability that a person does not describe themselves as a moderate and has a Bachelor's degree as their highest level of education?

	Liberal	Moderate	Conservative	Total
High School	.17	.10	.08	**.35**
Bachelor's	.20	.15	.11	**.46**
Master's/PhD	.08	.07	.04	**.19**
Total	**.45**	**.32**	**.23**	**1.00**

a. .15
b. .31
c. .32
d. .46
e. .54
f. .68

3. An urn contains 12 black balls and 8 white balls. Select one ball, look at the color, do not replace it, and select another ball. What is the probability that both balls selected were black?

a. .33
b. .3474
c. .36
d. .3789

4. 30% of athletes have used steroids. Given that an athlete has used steroids, the probability of testing positive is .80. Given that he has not used steroids, the probability of a positive test is .10. Let A be the event an athlete has used steroids and let B be the event an athlete tested positive. Which of the following best describe the events A and B?

a. Independent and mutually exclusive
b. Independent, but not mutually exclusive
c. Mutually exclusive, but not independent
d. Neither independent nor mutually exclusive

Lecture 11

1. The number of living great grandparents that a child has at birth is displayed in the table below. Suppose we independently sample two children. What is the probability that the first has at least 6 living great grandparents and the second has no more than 1?

GGP (X)	0	1	2	3	4	5	6	7	8
Probability	.07	.13	.18	.20	.16	.14	.07	.04	.01

a. $(.07)(.12) = .0084$
b. $(.05)(.07) = .0035$
c. $(.12)(.20) = .024$
d. $.07 + .13 = .20$
e. $.07 + .13 + .07 + .04 + .01 = .32$

2. The mean score on a final exam is 76 with a standard deviation of 6. Suppose the professor decides to curve the scores by adding 5 points. What are the new mean and standard deviation of the curved exam scores?

   a. $\mu = 76$ and $= 6$
   b. $\mu = 76$ and $\sigma = 11$
   c. $\mu = 76$ and $\sigma = 30$
   d. $\mu = 81$ and $\sigma = 6$
   e. $\mu = 81$ and $\sigma = 11$
   f. $\mu = 81$ and $\sigma = 30$

3. Suppose we have two weighted fair dice with the probability distributions shown below. Without doing any calculations, what can we say about the means and standard deviations?

Roll (Die A)	1	2	3	4
Probability	.10	.30	.40	.20

Roll (Die B)	1	2	3	4
Probability	.40	.20	.10	.30

   a. Mean of Die A is larger and standard deviation of Die A is larger
   b. Mean of Die B is larger and standard deviation of Die A is larger
   c. Mean of Die A is larger and standard deviation of Die B is larger
   d. Mean of Die B is larger and standard deviation of Die B is larger

Lecture 12

1. 57% of college students are female. Select 8 students independently of one another. What is the probability that more than 6 students are female?

   a. $(.57)^6(.43)^2 = .0063$
   b. $(.57)^7(.43)^1 + (.57)^8(.43)^0 = .0195$
   c. $(.57)^6(.43)^2 + (.57)^7(.43)^1 + (.57)^8(.43)^0 = .0259$
   d. $\frac{8!}{6!2!}(.57)^6(.43)^2 = .1776$
   e. $\frac{8!}{7!1!}(.57)^7(.43)^1 + \frac{8!}{8!0!}(.57)^8(.43)^0 = .0783$
   f. $\frac{8!}{6!2!}(.57)^6(.43)^2 + \frac{8!}{7!1!}(.57)^7(.43)^1 + \frac{8!}{8!0!}(.57)^8(.43)^0 = .2559$

2. Suppose you flip a coin 100 times and count 32 heads. If the coin is actually fair, how unusual is this result?

   a. Highly unusual: Assumption of a fair coin is almost certainly false
   b. Unusual: Assumption of a fair coin is probably false
   c. Slightly unusual: Assumption of a fair coin may or may not be false
   d. Not unusual: Assumption of a fair coin is probably true

Lecture 13

1. Scores for three students in three different classes are displayed below. Order the students from best to worst relative to the rest of their class.

   A: Scored 80 in a class with $\mu = 71$ and $\sigma = 6$
   B: Scored 83 in a class with $\mu = 74$ and $\sigma = 9$
   C: Scored 87 in a class with $\mu = 85$ and $\sigma = 4$

   a. A, B, C
   b. A, C, B
   c. B, A, C
   d. B, C, A
   e. C, A, B
   f. C, B, A

Lecture 14

1. The average speed that cars travel on I-79 is normal with a mean of 75 mph and standard deviation of 6 mph. What proportion of cars travel between 65 and 72 mph?

   a. .0475
   b. .1210
   c. .2610
   d. .3085
   e. .8790

2. Airplane decibel levels follow normal distribution with mean 100 and standard deviation 8. What decibel level corresponds to the loudest 20% of takeoffs?

   a. 93.28
   b. 93.6
   c. 106.4
   d. 106.72

3. The cost for a 40" flat screen TV in 2005 was normal with a mean of $2460 and standard deviation of $272. In 2017, the cost was also normal with a mean of $379 and a standard deviation of $56. If you paid $1850 for a TV in 2005, what would we expect the cost for the same TV to be in 2017?

   a. 284.36
   b. 376.45
   c. 381.55
   d. 473.64

Lecture 15

1. Die roll follows a uniform distribution with a mean of 3.5 and a standard deviation of 1.72. Roll 36 fair dice and average the rolls. What is the sampling distribution of the sample mean of 36 die rolls?

   a. Normal with $\mu_{\bar{X}} = 3.5$ and $\sigma_{\bar{X}} = \frac{1.72}{\sqrt{36}} = .287$
   b. Normal with $\mu_{\bar{X}} = 3.5$ and $\sigma_{\bar{X}} = \frac{1.72}{36} = .048$
   c. Normal with $\mu_{\bar{X}} = \frac{3.5}{36} = .097$ and $\sigma_{\bar{X}} = \frac{1.72}{\sqrt{36}} = .287$
   d. Normal with $\mu_{\bar{X}} = \frac{3.5}{36} = .097$ and $\sigma_{\bar{X}} = \frac{1.72}{36} = .048$
   e. Undetermined with $\mu_{\bar{X}} = 3.5$ and $\sigma_{\bar{X}} = \frac{1.72}{\sqrt{36}} = .287$
   f. Undetermined with $\mu_{\bar{X}} = 3.5$ and $\sigma_{\bar{X}} = \frac{1.72}{36} = .048$
   g. Undetermined with $\mu_{\bar{X}} = \frac{3.5}{36} = .097$ and $\sigma_{\bar{X}} = \frac{1.72}{\sqrt{36}} = .287$
   h. Undetermined with $\mu_{\bar{X}} = \frac{3.5}{36} = .097$ and $\sigma_{\bar{X}} = \frac{1.72}{36} = .048$

2. SAT scores follow a normal distribution with a population mean of 1060 and a population standard deviation of 140. Which of the following would decrease this probability further?

   I. Increase population mean SAT score to 1200
   II. Increase population standard deviation to 200
   III. Increase sample size to 25

   a. None would decrease the probability
   b. I only
   c. II only
   d. III only
   e. I and II
   f. I and III
   g. II and III
   h. I, II, and III

3. Die rolls follow a uniform distribution with a mean of 3.5 and a standard deviation of 1.72. Roll 36 fair dice and average the rolls. What is the probability the sample mean is less than 3?

   a. $P(\bar{X} < 3) = P\left(Z < \frac{3-3.5}{1.72}\right) = P(Z < -0.29)$
   b. $P(\bar{X} < 3) = P\left(Z < \frac{3.5-3}{1.72}\right) = P(Z < 0.29)$
   c. $P(\bar{X} < 3) = P\left(Z < \frac{3-3.5}{1.72/\sqrt{36}}\right) = P(Z < -1.74)$

d. $P(\bar{X} < 3) = P\left(Z < \frac{3-3.5}{1.72/\sqrt{36}}\right) = P(Z < 1.74)$

e. $P(\bar{X} < 3) = P\left(Z < \frac{3-3.5}{1.72/36}\right) = P(Z < -10.46)$

f. $P(\bar{X} < 3) = P\left(Z < \frac{3-3.5}{1.72/36}\right) = P(Z < 10.46)$

Lecture 16

1. A basketball player makes 85% of his free throws. During one week, he took 30 shots. What is the sampling distribution of the sample proportion of shots he could have made in that week?

a. Normal with mean $\mu_{\hat{p}} = .85$ and standard error $\sigma_{\hat{p}} = \sqrt{\frac{.85(1-.85)}{30}} = .065$

b. Normal with mean $\mu_{\hat{p}} = .85$ and standard error $\sigma_{\hat{p}} = \frac{.85}{\sqrt{30}} = .155$

c. Normal with mean $\mu_{\hat{p}} = .25.5$ and standard error $\sigma_{\hat{p}} = \sqrt{30(.85)(1-.85)} = 1.96$

d. Undetermined with mean $\mu_{\hat{p}} = .85$ and standard error $\sigma_{\hat{p}} = \sqrt{\frac{.85(1-.85)}{30}} = .065$

e. Undetermined with mean $\mu_{\hat{p}} = .85$ and standard error $\sigma_{\hat{p}} = \frac{.85}{\sqrt{30}} = .155$

f. Undetermined with mean $\mu_{\hat{p}} = .25.5$ and standard error $\sigma_{\hat{p}} = \sqrt{30(.85)(1-.85)} = 1.96$

2. Flip a fair coin 50 times. Let ‘Heads’ be a success. How should we set up the probability that the sample proportion of heads is less than .40?

a. $P\left(Z < \frac{.40-.50}{\sqrt{\frac{.40(1-.40)}{50}}}\right) = P(Z < -1.44) = .0749$

b. $P\left(Z < \frac{.40-.50}{\sqrt{\frac{.50(1-.50)}{50}}}\right) = P(Z < -1.41) = .0793$

c. $P\left(Z < \frac{.50-.40}{\sqrt{\frac{.50(1-.50)}{50}}}\right) = P(Z < 1.41) = .9207$

d. $P\left(Z < \frac{.50-.40}{\sqrt{\frac{.40(1-.40)}{50}}}\right) = P(Z < 1.44) = .9251$

Lecture 17

1. A random sample of heights of 21 men (in inches) found a sample mean of 70.2 inches. Assume heights are normal and $\sigma = 4$ inches. What is a 99% confidence interval for the mean height of males?

a. $70.2 \pm 1.645\left(\frac{4}{\sqrt{21}}\right) = (68.76, 71.64)$
b. $70.2 \pm 1.645(4) = (63.62,76.78)$
c. $70.2 \pm 1.96\left(\frac{4}{\sqrt{21}}\right) = (68.49, 71.91)$
d. $70.2 \pm 1.96(4) = (62.36, 78.04)$
e. $70.2 \pm 2.576\left(\frac{4}{\sqrt{21}}\right) = (67.95, 72.45)$
f. $70.2 \pm 2.576(4) = (59.90, 80.50)$

2. A 95% confidence interval for the mean SAT score is   Keeping all other statistics the same, which of the following will lead to a narrower interval?

I. Using 90% confidence
II. Using a population standard deviation of 200
III. Using a sample size of 50

a. None will lead to a narrower interval
b. I only
c. II only
d. III only
e. I and II
f. I and III
g. II and III
h. I, II, and III

3. IQ scores are normal with $\mu = 100$ and $\sigma = 15$. Take 10 random samples of size 25 and calculate a 90% confidence interval for each. How many of the intervals should we expect to contain 100?

a. 0
b. 1
c. 5
d. 9
e. 10

4. A random sample of heights of 21 men (in inches) found a sample mean of 70.2 inches. Assume heights are normal and $\sigma = 4$ inches. How large a sample would be needed to reduce the width of a 99% confidence interval to 1.5 inches?

a. 7
b. 14
c. 48
d. 189

Lecture 18

1. In testing the IQ of Pitt students to see if the mean is significantly greater than 115, we found a test statistic of 2.00, which yielded a p-value of $.0228$. For what levels of significance would the null hypothesis be rejected?

   a. Reject for $\alpha = .01$, $\alpha = .05$, and $\alpha = .10$
   b. Reject for $\alpha = .05$ and $\alpha = .10$
   c. Reject for $\alpha = .10$
   d. Reject for $\alpha = .01$
   e. Reject for $\alpha = .01$ and $\alpha = .05$

2. We want to determine if the average AP Statistics exam score in the state is significantly less than the national average of 3 at the 5% level of significance. The sample mean was 2.8, which yielded a p-value of .1469. What is the correct interpretation of the p-value?

   a. If the true mean AP Statistics exam score is actually 3, then the probability of getting a sample mean of 2.8 or less is .1469.
   b. The probability that the true mean AP Statistics exam score is equal to 3 is .1469.
   c. The probability that the true mean AP Statistics exam score is less than 3 is .1469.
   d. If the true mean AP Statistics exam score is actually 3, then the probability of getting a sample mean of exactly 2.8 is .1469.

3. We want to determine if the average AP Statistics exam score in the state is significantly less than the national average of 3 at the 5% level of significance. The test statistic is -1.05, which yields a p-value of .1469. What conclusion should be made regarding the null hypothesis?

   a. Average AP Statistics exam score is less than 3
   b. Average AP Statistics exam score is not less than 3
   c. Average AP Statistics exam score is greater than 3
   d. Average AP Statistics exam score is not greater than 3
   e. Average AP Statistics exam score is equal to 3
   f. Average AP Statistics exam score is not equal to 3

Lecture 19

1. In testing $H_0: \mu = 50$ vs. $H_1: \mu \neq 50$, we found a test statistic of 1.40. What is the p-value of this test?

   a. .0404
   b. .0808
   c. .1616
   d. .8384
   e. .9192
   f. .9596

2. We want to determine if the average body temperature of athletes is different from 98.6 degrees at the 1% level of significance. What is/are the critical value(s)?

   a. Only $Z = 1.282$
   b. Both $Z = \pm 1.645$
   c. Only $Z = 1.645$
   d. Both $Z = \pm 1.96$
   e. Only $Z = 2.326$
   f. Both $Z = \pm 2.576$

3. A confidence interval for the average spending habits of college students is $(1004, 1102)$. Is \$1000 a plausible amount for the average amount that college students spend each semester?

   a. Yes: 95% confidence interval contains 1000
   b. Yes: 95% confidence interval is entirely above 1000
   c. Yes: 95% confidence interval is entirely below 1000
   d. No: 95% confidence interval contains 1000
   e. No: 95% confidence interval is entirely above 1000
   f. No: 95% confidence interval is entirely below 1000

4. We tested to see if a machine was making screws that were 50 mm long using a 10% level of significance. After finding a sample mean of 50.13, we failed to reject the null hypothesis and concluded that the average length did not differ from 50. What can we conclude about a 90% confidence interval for the mean screw length?

   a. It would contain 50.13, but not 50
   b. It would contain 50, but not 50.13
   c. It would contain both 50 and 50.13
   d. It would contain neither 50 nor 50.13

Lecture 20

1. A researcher wants to test the effects of a cancer curing drug with the following hypotheses:

   H_0: Drug has no effect on cancer
   H_1: Drug is effective in curing cancer

   She concludes that the drug is effective in curing cancer, but it actually has no effect. What type of decision was made?

   a. Type I error
   b. Type II error
   c. Correct decision

2. We are testing to see if the mean temperature of chicken dishes at a restaurant is equal to 165 degrees or is some temperature less than 165. What would be the result of making a Type II error?

   a. Concluding the mean temperature is equal to 165 when it is actually less than 165
   b. Concluding the mean temperature is equal to 165 when it is actually greater than 165
   c. Concluding the mean temperature is equal to 165 when it is actually equal to 165
   d. Concluding the mean temperature is less than 165 when it is actually equal to165
   e. Concluding the mean temperature is less than 165 when it is actually less than 165

3. A fair die has mean 3.5 and standard deviation 1.71. Suspect a die might be weighted so that all 6 sides do not come up equally. Roll the die 40 times and average the results. If a two-sided test had been used, the p-value would have been .0644. What would be the p-value of the corresponding upper one-sided test?

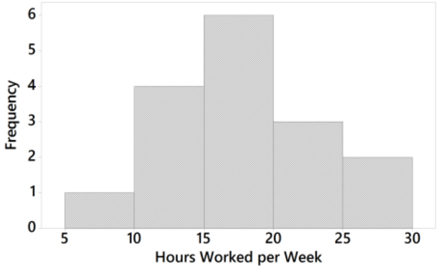
   a. .0322
   b. .0644
   c. .1288
   d. .9356

Lecture 21

1. A study found that high school students who work more than 15 hours per week were more likely to drop out. Take a random sample of high school students with part time jobs. What are the hypotheses for determining if high school students are working more than 15 hours per week on average?

   a. $H_0: \mu = 15; H_1: \mu \neq 15$
   b. $H_0: \mu = 15; H_1: \mu < 15$
   c. $H_0: \mu = 15; H_1: \mu > 15$
   d. $H_0: \mu \neq 15; H_1: \mu = 15$
   e. $H_0: \mu < 15; H_1: = 15$

2. A study found that high school students who work more than 15 hours per week were more likely to drop out. The random sample of 16 high school students with part time jobs yielded the following histogram. Are the conditions for a t-test satisfied?



a. Yes: The sample size is large enough to guarantee the shape of the sample mean is normal
b. Yes: The histogram of the data is approximately normal
c. No: The sample size is too small and the histogram is severely skewed

3. The sample of 16 students yielded a sample mean of 17.5 hours with a sample standard deviation of 5.50 hours. What is the correct test statistic to test if the population mean is greater than 15 hours?

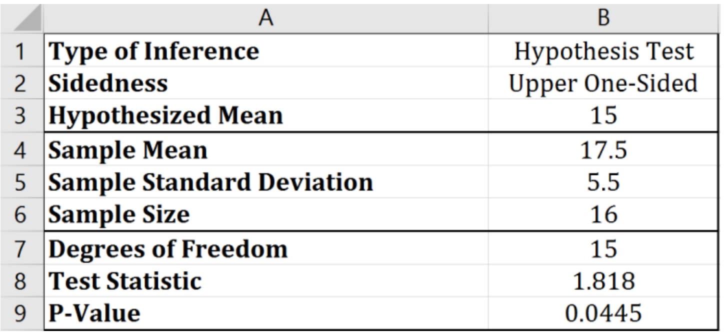
a. $Z = \frac{15-17.5}{5.50/\sqrt{16}} = -1.818$
b. $Z = \frac{17.5-15}{5.50/\sqrt{16}} = 1.818$
c. $t = \frac{15-17.5}{5.50/\sqrt{16}} = -1.818$
d. $t = \frac{17.5-15}{5.50/\sqrt{16}} = 1.818$

4. Suppose we knew that the population standard deviation for the number of hours high school students work each week is 5.50 hours. What would have been different in this hypothesis test?

I. Hypotheses
II. Value of test statistic
III. P-value

a. None would be different
b. I only
c. II only
d. III only
e. I and II
f. I and III
g. II and III
h. I, II, and III

5. We collect data on the number of minutes students spend on their smartphones daily returned the results in the output below. Keeping all other statistics the same, which of the following would reduce the p-value?



	A	B
1	**Type of Inference**	Hypothesis Test
2	**Sidedness**	Upper One-Sided
3	**Hypothesized Mean**	15
4	**Sample Mean**	17.5
5	**Sample Standard Deviation**	5.5
6	**Sample Size**	16
7	**Degrees of Freedom**	15
8	**Test Statistic**	1.818
9	**P-Value**	0.0445

I. Finding $\bar{x} = 165$
II. Finding $s = 20$
III. Using $n = 25$

a. None would reduce the p-value
b. I only
c. II only
d. III only
e. I and II
f. I and III
g. II and III
h. I, II, and III

Lecture 22

1. 10% of the world's population is left-handed. Left-handed tennis players may have an advantage because they put a slightly different spin on the ball so there may be a higher proportion of lefties than usual. What are the hypotheses for determining if more than 10% of tennis players are left-handed?

   a. $H_0: p = .10$ vs. $H_1: p > .10$
   b. $H_0: \hat{p} = .10$ vs. $H_1: \hat{p} > .10$
   c. $H_0: \bar{x} = .10$ vs. $H_1: \bar{x} > .10$
   d. $H_0: \mu = .10$ vs. $H_1: \mu > .10$

2. 10% of the world's population is left-handed. Left-handed tennis players may have an advantage because they put a slightly different spin on the ball so there may be a higher proportion of lefties than usual. The test statistic is 1.33 and the p-value is .0912. What decision and conclusion can we make regarding the proportion of left-handed tennis players using a 5% level of significance?

   a. Reject the null hypothesis and conclude that the proportion of left-handed tennis players is greater than .10
   b. Reject the null hypothesis and conclude that the proportion of left-handed tennis players is not greater than .10

c. Fail to reject the null hypothesis and conclude that the proportion of left-handed tennis players is greater than .10
d. Fail to reject the null hypothesis and conclude that the proportion of left-handed tennis players is not greater than .10

3. When estimating the proportion of students at his university who have used marijuana, Student B obtained a sample proportion of .52 and calculated a 95% confidence interval of $(.458, .582)$ from a sample of 250 people. Which of the following is the correct interpretation of this confidence interval?

   a. 95% of all confidence intervals calculated regarding the use of marijuana by college students will contain the sample proportion of .52.
   b. We are 95% confident that the proportion of college students who have used marijuana is between .458 and .582.
   c. We are 95% confident that between 45.8% and 58.2% of the next 250 college students surveyed will admit to having used marijuana.
   d. The probability that 95% of college students have used marijuana is between .458 and .582.

4. Student A directly asked 25 students a survey question regarding marijuana usage and obtained a sample proportion of .44 with a 95% confidence interval of $(.246, .634)$. Student B used an anonymous survey and obtained a sample proportion of .52 with a 95% confidence interval of $(.458, .582)$. Which of the following statements is not true?

   a. Student A's proportion likely underestimates the proportion of students who have used marijuana.
   b. Student B's interval is narrower because the sample size is larger.
   c. Student B's interval is likely to have less bias from students lying.
   d. We cannot make a decision about which interval provides a better estimate because different sampling methods were used.

5. A 95% confidence interval for the proportion of college students who have used marijuana is $(.458, .582)$ according to Student B. Which of the following statements are true?

   I. We can conclude that more than half of college students have used marijuana.
   II. .56 is a plausible estimate for the proportion of college students who have used marijuana.
   III. A 99% confidence interval would also contain .46.

   a. None are true
   b. I only
   c. II only
   d. III only
   e. I and II
   f. I and III
   g. II and III

h. I, II, and III

6. Student B obtained a sample proportion of .52 and calculated a 95% confidence interval of $(.458, .582)$ from a sample of 250 people. Maintaining a sample size of 250, what would have happened to the width of the interval if a sample proportion of .60 had been obtained instead of .52?

   a. Width increases
   b. Width decreases
   c. No change in width

Lecture 23

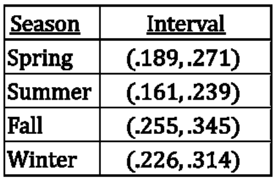
1. 400 high school football players were surveyed and asked what season their birthday is in. We want to know if the birthdays of high school football players are evenly distributed across all 4 seasons. What are the hypotheses?

   a. $H_0\colon p_{Sp} = .25, p_{Su} = .25, p_F = .25, p_W = .25$; H_1: At least one proportion differs from .25
   b. $H_0\colon p_{Sp} = .25, p_{Su} = .25, p_F = .25, p_W = .25$; H_1: All four proportions are different from .25
   c. $H_0\colon \hat{p}_{Sp} = .25, \hat{p}_{Su} = .25, \hat{p}_F = .25, \hat{p}_W = .25$; H_1: At least one proportion differs from .25
   d. $H_0\colon \hat{p}_{Sp} = .25, \hat{p}_{Su} = .25, \hat{p}_F = .25, \hat{p}_W = .25$; H_1: All four proportions are different from .25
   e. $H_0\colon \mu_{Sp} = 100, \mu_{Su} = 100, \mu_F = 100, \mu_W = 100$; H_1: At least one mean differs from 100
   f. $H_0\colon \mu_{Sp} = 100, \mu_{Su} = 100, \mu_F = 100, \mu_W = 100$; H_1: All four means differ from 100

2. 400 high school football players were surveyed and asked what season their birthday is in. We want to know if the birthdays of high school football players are evenly distributed across all 4 seasons. How many degrees of freedom does this test have?

   a. 3
   b. 4
   c. 399
   d. 400

3. 400 high school football players were surveyed and asked what season their birthday is in. The test statistic is 9.28 and the critical value is 7.815. Using this information and the confidence intervals below, which categories' proportions differ significantly from .25?



Season	**Interval**
Spring	**(.189,.271)**
Summer	**(.161,.239)**
Fall	**(.255,.345)**
Winter	**(.226,.314)**

a. None of the proportions of high school football players born in any season differ significantly from .25
b. Proportion of high school football players born in spring and winter differ significantly from .25
c. Proportion of high school football players both in summer and fall differ significantly from .25
d. Proportion of high school football players both in all four seasons differ significantly from .25

4. We found a test statistic of 9.28, which corresponds to a p-value of .0258. This led us to reject the null hypothesis at the 5% level of significance. What is the probability of making a Type I error?

   a. .0258
   b. .05
   c. .95
   d. .9742

5. In a random sample of M&M's, we observed 49 red M&M's, but would have expected 60.13 to be red. What is the chi-squared contribution from the red M&M's?

   a. $\frac{49-60.97}{49} = -.2271$
   b. $\frac{49-60.97}{60.97} = -.1851$
   c. $\frac{60.97-49}{60.97} = .1851$
   d. $\frac{60.97-49}{49} = .2271$
   e. $\frac{(49-60.97)^2}{60.97} = 2.350$
   f. $\frac{(49-60.13)^2}{49} = 2.528$

Lecture 24

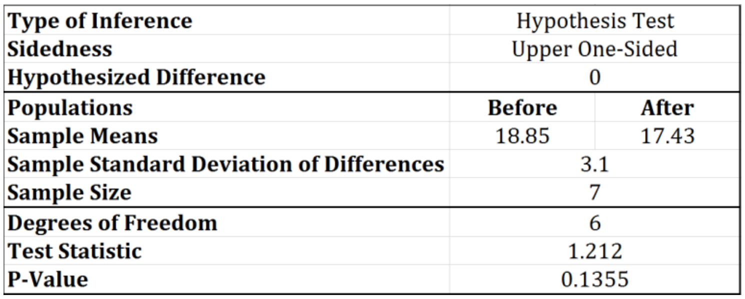
1. A shoe company is testing new running shoe. 12 sprinters run the 100-meter dash in old shoes first, take a break, and then run the same race in the new shoe the next day. We want to know if the times posted in the new shoe are faster than those in the old show. Which of the following describes the situation?

   a. Samples are independent. Running times are quantitative and type of show is categorical
   b. Samples are independent. Running times are categorical and type of shoe is quantitative
   c. Samples are dependent. Running times are quantitative and type of show is categorical
   d. Samples are dependent. Running times are categorical and type of shoe is quantitative

2. One way to write the null hypothesis by setting the before and after means equal to one another. Which of the following is not an appropriate way to write the null hypothesis?

a. $_0: \mu_B = \mu_A = 0$
b. $H_0: \mu_B - \mu_A = 0$
c. $H_0: \mu_D = 0$

3. Cameras are installed at 7 lights. A police department studies how many cars run the red light daily before and after the cameras are installed. The results of the test are presented in the output below. What conclusion can we come to using a 5% level of significance?



Type of Inference	Hypothesis Test	
Sidedness	Upper One-Sided	
Hypothesized Difference	0	
Populations	**Before**	**After**
Sample Means	18.85	17.43
Sample Standard Deviation of Differences	3.1	
Sample Size	7	
Degrees of Freedom	6	
Test Statistic	1.212	
P-Value	0.1355	

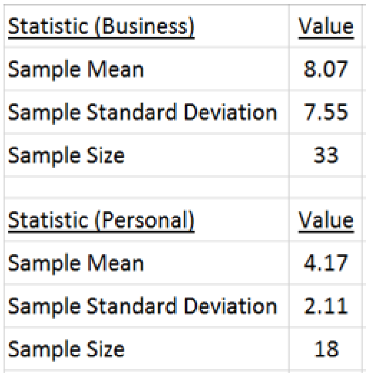
a. Fewer cars ran the red light after installing the cameras, but the difference is not significant
b. Significantly fewer cars ran the red light after installing the cameras.
c. Installation of the cameras caused fewer cars to run the red lights.
d. Installation of the cameras did not cause fewer cars to run the red light.

Lecture 25

1. An Uber driver collects data on his passengers regarding distance driven and if the trip was personal or for business. We want to determine if there is a significant difference in the average ride length for personal and business trips. What are the hypotheses?

a. $H_0: \mu_B = \mu_P; H_1: \mu_B \neq \mu_P$
b. $H_0: \mu_B = \mu_P; H_1: \mu_B > \mu_P$
c. $H_0: \mu_B = \mu_P; H_1: \mu_B < \mu_P$
d. $H_0: \mu_B \neq \mu_P; H_1: \mu_B = \mu_P$

2. The statistics below summarize the lengths of the rides for business and personal Uber rides. Which of the following is the correct test statistic?



Statistic (Business)	Value
Sample Mean	8.07
Sample Standard Deviation	7.55
Sample Size	33
Statistic (Personal)	Value
Sample Mean	4.17
Sample Standard Deviation	2.11
Sample Size	18

a. $t = \frac{8.07-4.17}{7.55^2+2.11^2/\sqrt{51}} = .453$

b. $t = \frac{8.07-4.17}{\sqrt{\frac{7.55^2}{33}+\frac{2.11^2}{18}}} = 2.775$

c. $t = \frac{8.07-4.17}{7.55+2.11/\sqrt{51}} = 2.883$

d. $t = \frac{8.07-4.17}{\sqrt{\frac{7.55}{33}+\frac{2.11}{18}}} = 6.630$

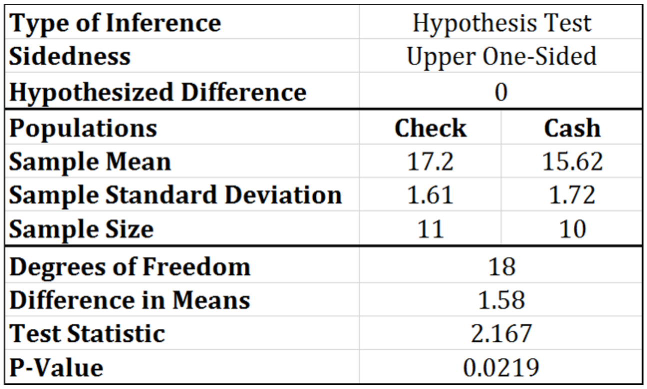
3. Which of the following would change in the hypothesis test if we instead used personal trips as population 1 and business trips as population 2?

Type of Inference	Hypothesis Test	
Sidedness	Two-Sided	
Hypothesized Difference	0	
Populations	**Business**	**Personal**
Sample Mean	8.07	4.17
Sample Standard Deviation	7.55	2.11
Sample Size	33	18
Degrees of Freedom	40	
Difference in Means	3.9	
Test Statistic	2.775	
P-Value	0.0083	

I. Hypotheses
II. Test Statistic
III. P-Value

a. None would change
b. I only
c. II only
d. III only
e. I and II
f. I and III
g. II and III
h. I, II, and III

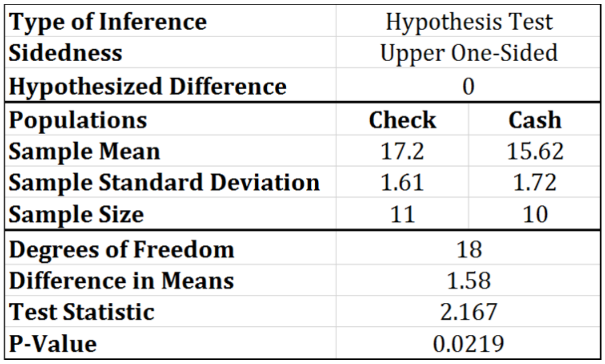
4. We want to compare the mean percentage customers leave as a tip at a restaurant in cash vs. on a check. Based on the output below, what conclusion can we come to using a 5% level of significance?



Type of Inference	Hypothesis Test	
Sidedness	Upper One-Sided	
Hypothesized Difference	0	
Populations	**Check**	**Cash**
Sample Mean	17.2	15.62
Sample Standard Deviation	1.61	1.72
Sample Size	11	10
Degrees of Freedom	18	
Difference in Means	1.58	
Test Statistic	2.167	
P-Value	0.0219	

a. Mean tip percentage left on check is greater than the mean tip percentage left in cash, but the difference is not significant.
b. Mean tip percentage left on check is less than the mean tip percentage left in ash, but the difference is not significant.
c. Mean tip percentage left on check is significantly greater than the mean tip percentage left in cash.
d. Mean tip percentage left on check is significantly less than the mean tip percentage left in cash.

5. We want to compare the mean percentage customers leave as a tip at a restaurant in cash vs. on a check. Keeping all other statistics the same, which of the following would reduce the p-value?



Type of Inference	Hypothesis Test	
Sidedness	Upper One-Sided	
Hypothesized Difference	0	
Populations	**Check**	**Cash**
Sample Mean	17.2	15.62
Sample Standard Deviation	1.61	1.72
Sample Size	11	10
Degrees of Freedom	18	
Difference in Means	1.58	
Test Statistic	2.167	
P-Value	0.0219	

I. Double the difference between the sample means
II. Double the standard deviations
III. Double the sample sizes

a. None would reduce the p-value
b. I only
c. II only

d. III only
e. I and II
f. I and III
g. II and III
h. I, II, and III

6. In comparing the lengths of Uber rides for business and personal trips, we rejected $H_0: \mu_B = \mu_P$ and concluded $H_1: \mu_B \neq \mu_P$. What can we conclude about a 95% confidence interval for the difference between the means?

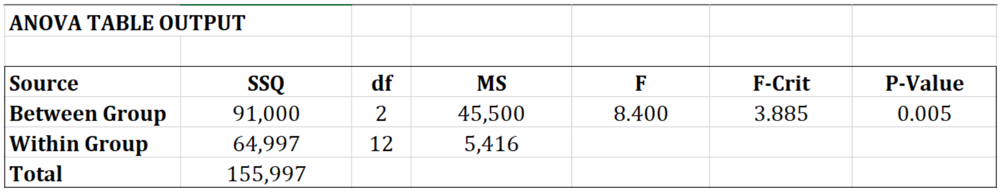
   a. If would contain both 3.9 and 0
   b. It would contain 0, but not 3.9
   c. It would contain 3.9, but not 0
   d. It would contain neither 0 nor 3.9

Lecture 26

1. We want to compare the average math SAT scores for computer science, economics, and history majors. Sample 5 students from each of the three majors. What are the degrees of freedom?

   a. Numerator: 2; Denominator: 2
   b. Numerator: 2; Denominator; 4
   c. Numerator: 2; Denominator: 12
   d. Numerator: 3; Denominator: 2
   e. Numerator: 3; Denominator; 4
   f. Numerator: 3; Denominator: 12

2. We want to compare the average math SAT scores for computer science, economics, and history majors. The results of the test are contained in the output below. What conclusion can we come to using the just output below?



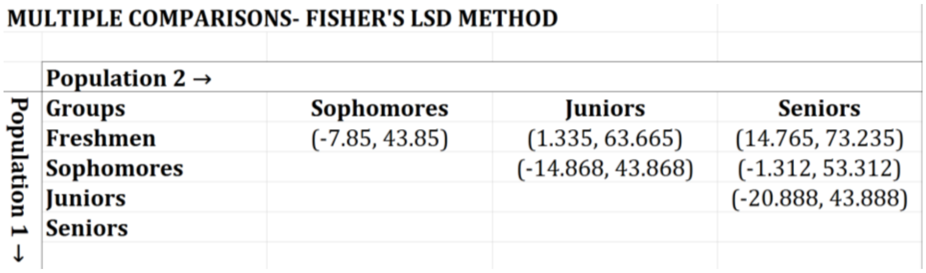
ANOVA TABLE OUTPUT

Source	SSQ	df	MS	F	F-Crit	P-Value
Between Group	91,000	2	45,500	8.400	3.885	0.005
Within Group	64,997	12	5,416			
Total	155,997					

   a. None of the mean SAT scores are significantly different.
   b. Exactly two of the mean SAT scores are significantly different
   c. At least two of the mean SAT scores are significantly different
   d. All three mean SAT scores are significantly different from one another

Lecture 27

1. The output below gives 95% Fisher's LSD confidence intervals for the difference between each pair of means. How many and which groups have significantly different means?



MULTIPLE COMPARISONS- FISHER'S LSD METHOD

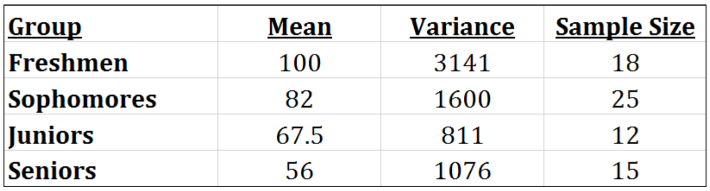
Population 2 →

Population 1 ↓

Groups	Sophomores	Juniors	Seniors
Freshmen	(-7.85, 43.85)	(1.335, 63.665)	(14.765, 73.235)
Sophomores		(-14.868, 43.868)	(-1.312, 53.312)
Juniors			(-20.888, 43.888)
Seniors			

a. None of the groups have significantly different means
b. 2 groups: Freshmen and juniors, freshmen and seniors
c. 4 groups: Freshmen and sophomores, sophomores and juniors, sophomores and seniors, junior and seniors
d. All 6 pairs of means are significantly different

2. The output below shows the sample statistics for the number of texts send each day by each of the four groups. Keeping all other statistics the same, which of the following would reduce the p-value?



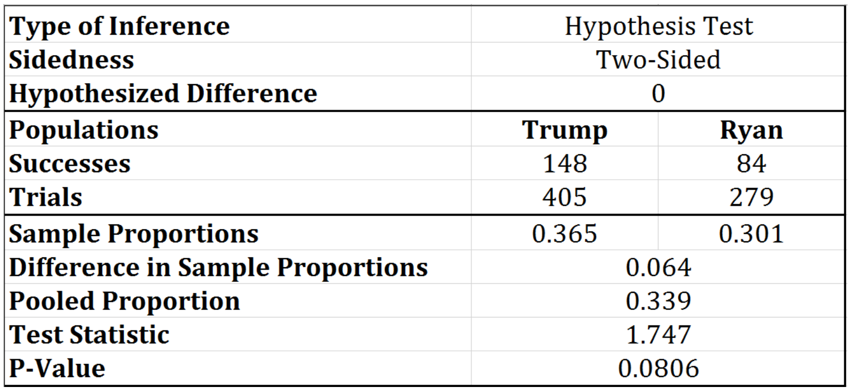
Group	Mean	Variance	Sample Size
Freshmen	100	3141	18
Sophomores	82	1600	25
Juniors	67.5	811	12
Seniors	56	1076	15

I. Sample mean for freshmen equal to 110
II. Sample variance for sophomores equal to 900
III. Sample size for juniors equal to 24

a. None would reduce the p-value
b. I only
c. II only
d. III only
e. I and II
f. I and III
g. II and III
h. I, II, and III

Lecture 28

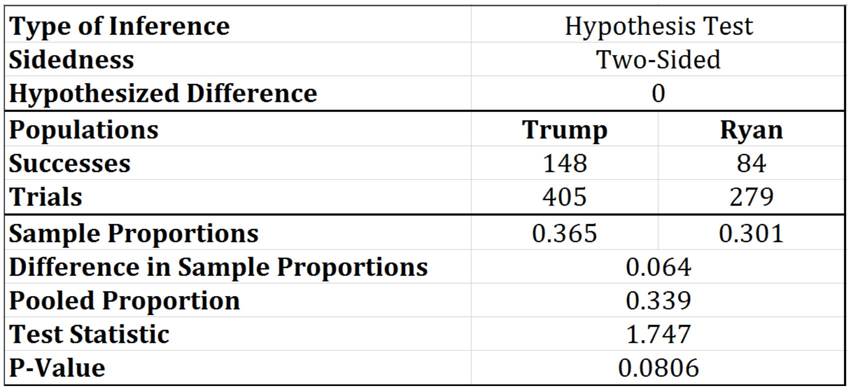
1. We want to compare the approval ratings of Donald Trump and Paul Ryan based on a June 2017 survey. From the output below, what conclusion can we come to using a 5% level of significance?



Type of Inference	Hypothesis Test	
Sidedness	Two-Sided	
Hypothesized Difference	0	
Populations	**Trump**	**Ryan**
Successes	148	84
Trials	405	279
Sample Proportions	0.365	0.301
Difference in Sample Proportions	0.064	
Pooled Proportion	0.339	
Test Statistic	1.747	
P-Value	0.0806	

a. Trump's approval rating is significantly different from Ryan's with indications it is higher.
b. Trump's approval rating is significantly different from Tyan's with indications it is lower.
c. Trump's approval rating is lower than Ryan's, but the difference is not significant.
d. Trump's approval rating is higher than Ryan's, but the difference is not significant.

2. We want to compare the approval ratings of Donald Trump and Paul Ryan based on a June 2017 survey. The output for the test is below. Which of the following statements are true?



Type of Inference	Hypothesis Test	
Sidedness	Two-Sided	
Hypothesized Difference	0	
Populations	**Trump**	**Ryan**
Successes	148	84
Trials	405	279
Sample Proportions	0.365	0.301
Difference in Sample Proportions	0.064	
Pooled Proportion	0.339	
Test Statistic	1.747	
P-Value	0.0806	

I. A 95% confidence interval would contain .064.
II. A 95% confidence interval would contain only positive values
III. Zero is a plausible value for the difference between proportions

a. None would reduce the p-value
b. I only
c. II only
d. III only
e. I and II
f. I and III
g. II and III
h. I, II, and III

Lecture 29

1. The tables below show the observed and expected counts of handedness by gender for a random sample of 500 people. What is the contribution to the chi-squared test statistic for left-handed females?

Observed	Left	Right
Male	30	240
Female	20	210

Expected	Left	Right
Male	27	243
Female	23	207

   a. $\frac{20-23}{20} = -0.15$
   b. $\frac{20-23}{23} = -0.13$
   c. $\frac{23-20}{23} = 0.13$
   d. $\frac{23-20}{20} = 0.15$
   e. $\frac{(20-23)^2}{23} = 0.39$
   f. $\frac{(20-23)^2}{20} = 0.45$

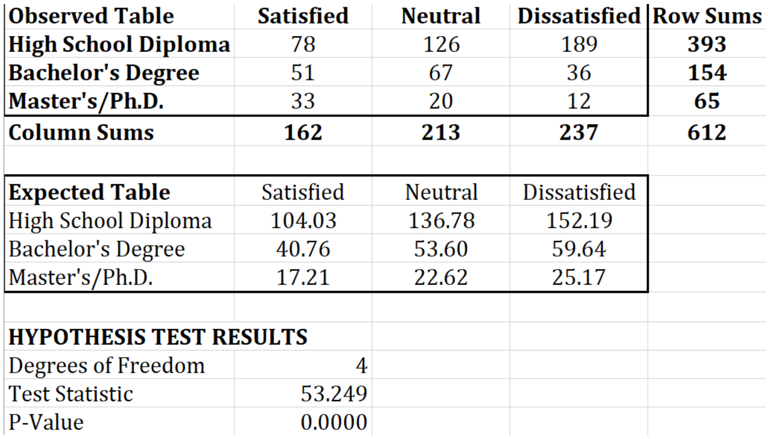
2. We want to determine if there is a relationship between gender and handedness by taking a random sample of 500 people. How many degrees of freedom does this test have?

   a. 1
   b. 2
   c. 3
   d. 4
   e. 499
   f. 500

3. A random sample of people were asked what their highest level of education was and how satisfied they were with their job. We want to know if highest level of education and job satisfaction are independent. What are the hypotheses?

   a. H_0: Highest level of education and job satisfaction are independent.
      H_1: Highest level of education and job satisfaction are not independent.
   b. H_0: Highest level of education and job satisfaction are not independent.
      H_1: Highest level of education and job satisfaction are independent.
   c. H_0: People of all education levels are satisfied with their jobs.
      H_1: People of all education levels are dissatisfied with their jobs.
   d. H_0: People of all education levels are dissatisfied with their jobs.
      H_1: People of all education levels are satisfied with their jobs.

4. The output below shows the results of comparing highest level of education with job satisfaction. What conclusion can we come to using a 5% level of significance?



Observed Table	Satisfied	Neutral	Dissatisfied	Row Sums
High School Diploma	78	126	189	**393**
Bachelor's Degree	51	67	36	**154**
Master's/Ph.D.	33	20	12	**65**
Column Sums	**162**	**213**	**237**	**612**
Expected Table	Satisfied	Neutral	Dissatisfied	
High School Diploma	104.03	136.78	152.19	
Bachelor's Degree	40.76	53.60	59.64	
Master's/Ph.D.	17.21	22.62	25.17	
HYPOTHESIS TEST RESULTS				
Degrees of Freedom	4			
Test Statistic	53.249			
P-Value	0.0000			

a. Highest level of education and job satisfaction are independent
b. Highest level of education and jobs satisfaction are not independent: More education is correlated with higher job satisfaction
c. Highest level of education and job satisfaction are not independent: Less education is correlated with higher job satisfaction

Lecture 30

1. We want to use the number of attractions at an amusement park to predict admission price. The equation of the regression line is $\hat{y}$=19.07+.89x and the coefficient of determination is .5693. How should we describe the strength of the linear relationship?

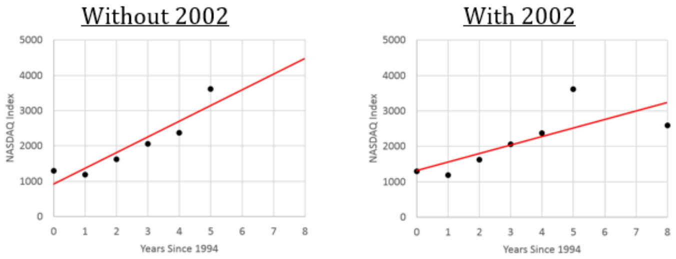
   a. Strong negative linear relationship
   b. Moderate negative linear relationship
   c. Weak negative linear relationship
   d. No linear relationship
   e. Weak positive linear relationship
   f. Moderate positive linear relationship
   g. Strong positive linear relationship

2. The predicted cost to enter a park with 20 attractions is \$36.87. You pay \$50 to enter a park with 20 attractions. What is the residual?

   a. $36.87 - 50 = -13.13$
   b. $\frac{50-36.87}{36.87} = .356$
   c. $\frac{(50-36.87)^2}{36.87} = 4.68$
   d. $50 - 36.87 = 13.13$
   e. $50 + 36.87 = 86.87$

3. We want to predict the NASDAQ index at beginning of the year between 1994 and 1999 using years since 1994 as the predictor. The scatterplots below show the regression line for

data between 1994 and 1999 with and without 2002 included. Which of the following describe the observation from 2002?



a. Only an outlier
b. Only an influential point
c. Both an outlier and an influential point
d. Neither an outlier nor an influential point

Lecture 31

1. We want to use the number of attractions at an amusement park to predict admission price. The output for the linear regression is presented below. What is the linear model?

	Coefficient	Std. Error	t-Stat	P-Value	95% LB	95% UB
Intercept	19.015	5.380	3.535	0.001	8.070	29.960
Attractions	0.891	0.135	6.604	0.000	0.617	1.166

a. $y = \beta_0 + \beta_1 x$
b. $\hat{y} = \beta_0 + \beta_1 x$
c. $y = \beta_0 + \beta_1 x + \varepsilon$
d. $\hat{y} = \beta_0 + \beta_1 x + \varepsilon$
e. $y = 19.015 + .891x$
f. $\hat{y} = 19.015 + .891x$
g. $y = 19.015 + .891x + \varepsilon$
h. $\hat{y} = 19.015 + .891x + \varepsilon$

2. We want to use the number of attractions at an amusement park to predict admission price. What are the hypotheses?

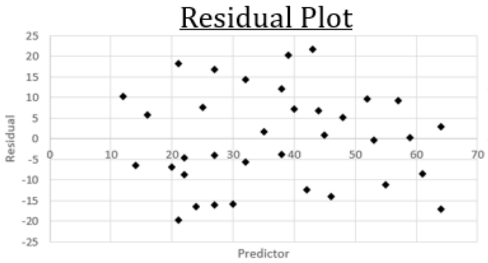
a. $H_0: b_0 = 0; H_1: b_0 \neq 0$
b. $H_0: b_1 = 0; H_1: b_1 \neq 0$
c. $H_0: \beta_0 = 0; H_1: \beta_0 \neq 0$
d. $H_0: \beta_1 = 0; H_1: \beta_1 \neq 0$

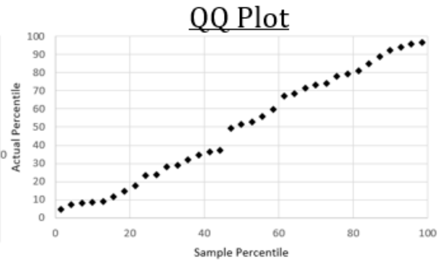
3. We want to use the number of attractions at an amusement park to predict admission price. The output for the linear regression is presented below. What conclusion can we make at the 5% level of significance?

	Coefficient	Std. Error	t-Stat	P-Value	95% LB	95% UB
Intercept	19.015	5.380	3.535	0.001	8.070	29.960
Attractions	0.891	0.135	6.604	0.000	0.617	1.166

a. Number of attractions is a significant predictor of admission price with an indication the relationship is positive
b. Number of attractions is a significant predictor of admission price with an indication the relationship is negative
c. A positive linear relationship exists between the number of attractions and admission price, but it is not significant
d. A negative linear relationship exists between the number of attractions and admission price, but it is not significant.

4. We want to use the number of attractions at an amusement park to predict admission price. The residual plot and QQ-plot are below. Are any error conditions violated?





a. No
b. Yes: Normality
c. Yes: Homoscedasticity
d. Yes: Both normality and homoscedasticity

Lecture 32

1. A 95% prediction interval for the admission price for parks with 37 rides is (27.57, 76.39). What is the correct interpretation of this interval?

a. We are 95% confident that the average admission price for all amusement parks is between $27.57 and $76.39.
b. We are 95% confident that the average admission price for all amusement parks with 37 rides is between $27.57 and $76.39.
c. We are 95% confident that the admission price for a single amusement park is between $27.57 and $76.39.
d. We are 95% confident that the admission price for a single amusement park with 37 rides is between $27.57 and $76.39.

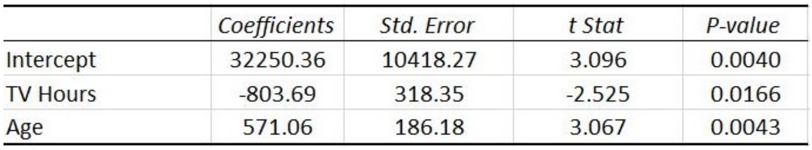
2. A 95% confidence interval for the admission price for parks with 37 rides is (47.92, 56.04). Which of the following intervals would be wider?

II. A 95% prediction interval for admission price for parks with 37 rides
III. A 95% confidence interval for admission price for parks with 50 rides
IV. A 99% confidence interval for admission price for parks with 37 rides

a. None would be wider
b. I only
c. II only
d. III only
e. I and II
f. I and III
g. II and III
h. I, II, and III

Lecture 33

1. We want to use hours of television watched weekly and age to predict a person's annual income. The output of the multiple linear regression is below. What is the correct interpretation of the slope coefficient for number of hours of television watched weekly?



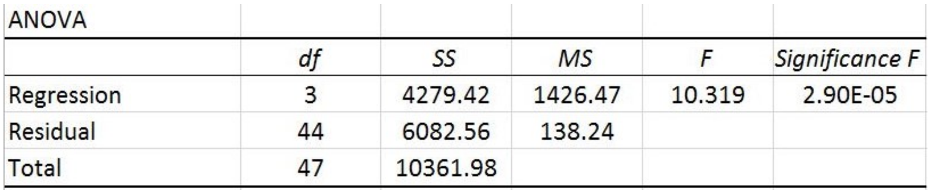
	Coefficients	*Std. Error*	*t Stat*	*P-value*
Intercept	32250.36	10418.27	3.096	0.0040
TV Hours	-803.69	318.35	-2.525	0.0166
Age	571.06	186.18	3.067	0.0043

a. For every additional hour of television watched each week, there is an $803.69 decrease in the predicted annual income.
b. For every additional hour of television watched each week, there is an $803.69 decrease in the annual income.
c. For every additional hour of television watched each week, holding age constant, there is an $803.69 decrease in the predicted annual income.
d. For every additional hour of television watched each week, holding age constant, there is an $803.69 decrease in the annual income.

2. We want to use lectures skipped, number of late assignments, and midterm grade to predict a student's final grade. What are the hypotheses?

a. $H_0: \beta_1 = \beta_2 = \beta_3 = 0$; $H_1: \beta_1 \neq 0, \beta_2 \neq 0, \beta_3 \neq 0$
b. $H_0: \beta_1 = \beta_2 = \beta_3 = 0$; H_1: At least one $\beta_i \neq 0$
c. $H_0: b_1 = b_2 = b_3 = 0$; $H_1: b_1 \neq 0, b_2 \neq 0, b_3 \neq 0$
d. $H_0: b_1 = b_2 = b_3 = 0$; H_1: At least one $b_i \neq 0$

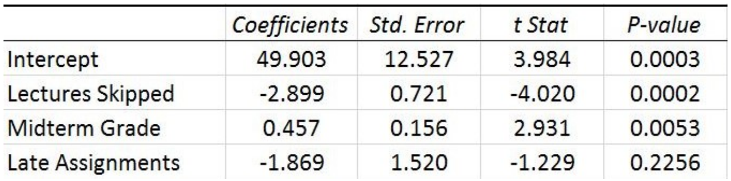
3. We want to use lectures skipped, number of late assignments, and midterm grade to predict a student's final grade. Using the output below, what conclusion can we come to regarding the predictors?



ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	4279.42	1426.47	10.319	2.90E-05
Residual	44	6082.56	138.24		
Total	47	10361.98			

a. None of the predictors are significant
b. Exactly one predictor is significant
c. Exactly two predictors are significant
d. All three predictors are significant
e. At least one predictor is significant

4. We want to use lectures skipped, number of late assignments, and midterm grade to predict a student's final grade. Using the output below, what conclusion can we come to regarding the predictors?



	Coefficients	*Std. Error*	*t Stat*	*P-value*
Intercept	49.903	12.527	3.984	0.0003
Lectures Skipped	-2.899	0.721	-4.020	0.0002
Midterm Grade	0.457	0.156	2.931	0.0053
Late Assignments	-1.869	1.520	-1.229	0.2256

a. None are significant
b. Only lectures skipped
c. Only midterm grade
d. Both lectures skipped and midterm grade
e. Both lectures skipped and late assignments
f. Both midterm grade and late assignments
g. All three predictors are significant

Appendix D

Quizzes

Quiz 1

1. Identify the variable situation reflected in this statement: “Female teenagers are more likely to have summer jobs than male teenagers.”

    a. One categorical
    b. One quantitative
    c. Two categorical
    d. Two quantitative
    e. One categorical and one quantitative

Use the following scenario for Questions 2, 3, and 4.

Verizon is interested in how much data their average customer uses in a typical month so they can determine if their prices are reasonable. They send out a survey to 1500 randomly selected customers, but receive only 300 responses. From the 300 responses, Verizon found that the respondents used a **total** of 1200 gigabytes of data.

2. What was the population of interest in the poll?

    a. All people in the world
    b. All people in the world with cell phones
    c. All Verizon customers
    d. All Verizon customers who were mailed a survey
    e. All respondents to the survey

3. What type of error occurred when collecting the data?

    a. Sampling error
    b. Nonresponse bias
    c. Data acquisition error
    d. Selection bias

4. Which of the following best describes how we should denote the statistic and parameter from the poll?

    a. Statistic: $\bar{x}$ unknown; Parameter: $\mu = 4$
    b. Statistic: μ unknown; Parameter: $\overline{\ } = 4$
    c. Statistic: p unknown; Parameter: $\hat{p} = .25$
    d. Statistic: $\hat{p}$ unknown; Parameter $p = .25$
    e. Statistic: $\bar{x} = 4$; Parameter: μ unknown
    f. Statistic: $\mu = 4$; Parameter: $\bar{x}$ unknown

g. Statistic: $p = .25$; Parameter: $\hat{p}$ unknown
h. Statistic: $\hat{p} = .25$; Parameter p unknown

5. The Human Resources department at a major university wants to investigate faculty members' opinions of working conditions through a survey. They want to obtain opinions from faculty members at all levels (adjunct, associate professors, and tenured professors) so they send out a survey to 50 randomly selected faculty members at each of the three levels. What type of sampling method was used to generate the sample?

   a. Simple random sample
   b. Stratified random sample
   c. Systematic sample
   d. Convenience sample
   e. Voluntary sample

6. A single question on a survey about impressions on Saturday Night Live asked respondents, "On a scale from 1 to 6 with 1 being 'not funny at all' and 6 being 'extremely funny', rate Alec Baldwin's impression of Donald Trump and Melissa McCarthy's impression of Sean Spicer." What is the problem with this survey question?

   a. Complicated question
   b. Vague concept
   c. Leading question
   d. Central tendency bias
   e. Error prone response option

Quiz 2

1. A pediatrician wants to determine if adult males who played high school football have higher rates of CTE, a disease caused by repeated blows to the head, than other high school athletes. They recruit a group of 1000 adult males who played sports in high school. Each subject reported the sports he played and was then analyzed for CTE. Which of the following is true about this study?

   a. This is a retrospective observational study where the occurrence of CTE is the explanatory variable and the sports played in high school is the response.
   b. This is a retrospective observational study where the sports played in high school is the explanatory variable and the occurrence of CTE is the response.
   c. This is an experiment where the occurrence of CTE is the explanatory variable and the sports played in high school is the response.
   d. This is an experiment where the sports played in high school is the explanatory variable and the occurrence of CTE is the response.
   e. This is a prospective observational study where the occurrence of CTE is the explanatory variable and the sports played in high school is the response.
   f. This is a prospective observational study where the sports played in high school is the explanatory variable and the occurrence of CTE is the response.

Use the following scenario for Questions 2 and 3.

Plant therapists believe that plants can promote productivity in the workplace, but that the impact of plants may differ between males and females. A company volunteers to have its employees take part in a study. This company has 12 employees: 6 men and 6 women. Before the work day began, the owner placed a plant on the desks of 3 randomly selected men and 3 randomly selected women because they believed men and women may be distracted by different things. Each employee was watched throughout the day to see how many times they did something that was considered “off-task”. No employees were aware that they were being studied.

2. What type of experimental design was used in this study?

   a. Matched pairs design
   b. Completely randomized design
   c. Block design

3. What ethics violation occurred in this study?

   a. Autonomy
   b. Beneficence
   c. Confidentiality
   d. Deception
   e. Informed consent

4. A group of psychologists wants to test two different treatments that could potentially reduce anxiety. Recruited subjects are asked to rate their anxiety levels on a scale from 1 to 10 at the beginning of the study. They are then divided into four groups and assigned to one of four treatment groups. One group receives both group therapy and a medication; a second group receives only group therapy; the third group receives only the medication. The final group receives a placebo instead of a medication and does not attend group therapy. At the end of six weeks, each subject is asked to report their anxiety levels. What group of subjects is acting as the control?

   a. Subjects who receive the medication and also attend group therapy
   b. Subjects who are only receiving the medication
   c. Subjects who are only attending group therapy
   d. Subjects who receive a placebo and do not attend group therapy
   e. Subjects who noticed a decrease in their anxiety levels
   f. Subjects who noticed no change in their anxiety levels

5. A doctor is curious as to if the flu shot or a nasal spray called the FluMist is more effective in preventing the flu. In total, 100 subjects (50 men and 50 women) agree to take part. However, the doctor has no reason to believe that incidences of the flu differ by gender. To prevent the treatments from interacting with one another and causing a side effect, subjects

can only be assigned one treatment. At the end of winter, subjects report if they developed the flu. What is the most appropriate way to assign subjects to the two treatments?

a. Assign all 100 subjects to be given both the flu shot and FluMist.
b. Assign the 50 men to the flu shot and the 50 women to the FluMist.
c. Assign 50 subjects to the flu shot and the other 50 to the FluMist, ignoring subjects' gender.
d. Give 25 men and 25 women the flu shot and give the other 25 men and 25 women the FluMist.
e. Give 50 subjects the flu shot and the other 50 the FluMist at the beginning of winter. Halfway through winter, give each subject the other treatment.
f. Give the 50 men the flu shot and the 50 women the FluMist at the beginning of winter. Halfway through winter, give the 50 men the FluMist and the 50 women the FluMist.

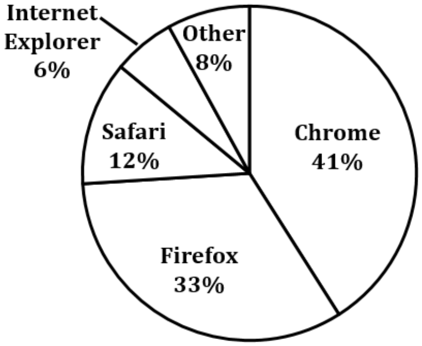
6. Three representatives from a company are asking for opinions about a new product. Each representative gets responses regarding the quality of the product on a scale from 1 to 5, with 1 being terrible and 6 being fantastic, but each representative uses a different method of surveying people.

- A: Select the 100 people who purchased the product most recently
- B: Select 100 people randomly from a list of all owners of the product
- C: Randomly select 25 reviews of the product from Amazon and use the number of stars as the quality

Which of the following statements is not true?

a. B will resemble the population the closest
b. A and B will likely produce very similar results because the sample sizes are the same.
c. C will likely provide a biased estimate of the average product rating.
d. A and B are both better sampling methods than C
e. C is likely going to have the largest sampling error because the sample size is the smallest

7. The following pie chart shows the primary Internet browser used by respondents of a survey. There were 300 people who primarily use Safari. How many people responded that they primarily use Firefox?



a. 36

b. 109
c. 825
d. 909
e. 2,500

8. The following cross-classification table compares a person's gender with the type of car that they drive. What proportion of men drive trucks?

Gender	Car	Truck	SUV	No Vehicle	Total
Male	1,000	600	700	100	2,400
Female	1,400	150	1,100	50	2,600
Total	2,400	750	1,800	150	5,000

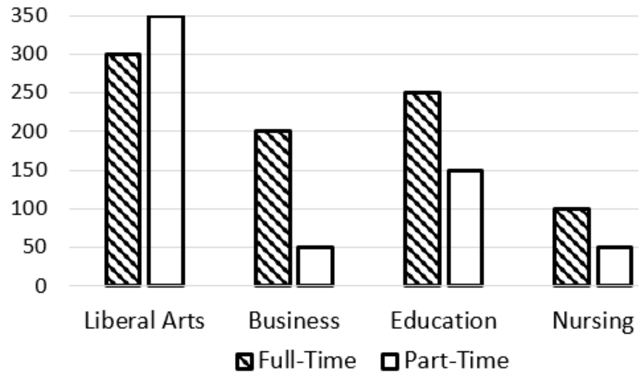
a. .12
b. .15
c. .25
d. .48
e. .80

9. Random samples of males and females were asked to choose their favorite sport out of the four major sports in the United States. The results are presented in the relative frequency tables below. Which of the following statements describes the medians of these two groups?

Males	Percentage	Females	Percentage
Baseball	18%	Baseball	32%
Basketball	33%	Basketball	19%
Football	38%	Football	28%
Hockey	11%	Hockey	21%

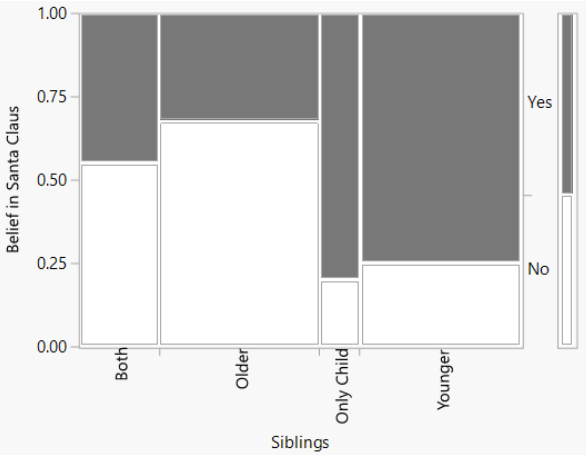
a. Males' median response is basketball, while females' median response is football
b. Males' median response is basketball, while females' median response is baseball.
c. The median response for both males and females is basketball.
d. The medians cannot be computed, and thus no comparison can be made.

10. The following double bar graph summarizes the employment status of a random sample of college graduates. Responses are separated based on the type of degree the person earned. For which group is the conditional proportion of respondents who work full-time the highest?



a. Liberal arts
b. Business
c. Education
d. Nursing

11. The following mosaic plot contains the results of a random sample of 7-year-olds regarding their belief in Santa Claus. Results are further broken down based on if the child has only older siblings, only younger siblings, both, or is an only child. Based on the information provided in the graph, which of the following statements are true?



I. The category with the highest frequency of 'Yes' responses came from those who are only children.
II. 7-year-olds with an older sibling were less likely to believe in Santa Claus than those without an older sibling.
III. About the same number of children with only younger siblings were surveyed as those with only older siblings.

a. None are true
b. I only
c. II only
d. III only

e. I and II
f. I and III
g. II and III
h. I, II, and III

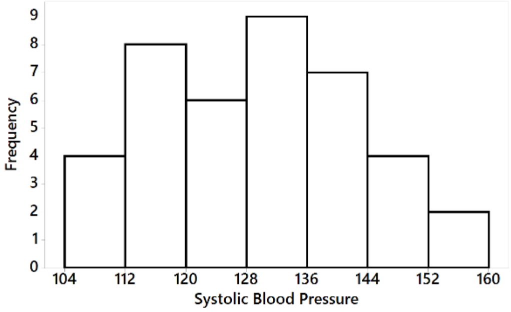
Quiz 3

1. The following table shows the results of a survey given to 20 students asking what their GPA is and what year in school they are in. What is the best type of graphical display to use to compare this data?

Year	GPAs
Freshman	3.91, 2.97, 3.61, 2.28, 2.84
Sophomore	3.83, 3.58, 2.92, 3.31, 2.07
Junior	3.70, 2.86, 3.36, 3.41, 2.90
Senior	3.23, 2.89, 2.12, 3.74, 3.09

a. Bar graph
b. Mosaic plot
c. Histogram
d. Side-by-side boxplots
e. Scatterplot

2. A nurse took the blood pressure of 40 patients who came to a clinic for a physical exam. The histogram below shows the systolic blood pressure of these patients. What proportion of patients had systolic blood pressures between 112 and 136?



a. .15
b. .20
c. .35
d. .575
e. .675

3. Suppose that about 99.7% of pairs of tennis shoes at a shoe store cost between \$32 and \$128 and that a histogram of these prices has a normal shape. What are the mean and standard deviation of the cost of a pair of tennis shoes at this shoe store?

a. Mean is $80 and standard deviation is $48
b. Mean is $80 and standard deviation is $24
c. Mean is $80 and standard deviation is $16
d. Mean is $160 and standard deviation is $48
e. Mean is $160 and standard deviation is $24
f. Mean is $160 and standard deviation is $16

4. Data for the amount of time that Americans spend commuting to work each day creates a histogram that has a normal shape. The mean is 35 minutes and the standard deviation is 10 minutes. According to the Empirical Rule, which of the following is not a conclusion that we can make?

   a. It would be common to find a commuter who takes 43 minutes to commute to work.
   b. About 5% of commuters take longer than 55 minutes to commute to work.
   c. A commuter who takes 30 minutes to commute to work would have a negative Z-score.
   d. A commuter who takes 25 minutes to commute to work is more standard deviations from the mean than a commuter who takes 40 minutes.

Use the following scenario for Questions 5, 6, and 7.

The following table shows the weights of the five slimmest and five heaviest U.S. Presidents. The mean weight is 190 pounds with a standard deviation of 20 pounds. The five-number summary is 122, 175, 210, 220, 332.

Slimmest	Weight
James Madison	122
Andrew Jackson	154
John Tyler	160
Franklin Pierce	162
William Henry Harrison	162

Heaviest	Weight
William Howard Taft	332
Grover Cleveland	275
Chester Arthur	246
Theodore Roosevelt	234
Bill Clinton	231

5. Which of the presidents are considered outliers according to the IQR rule?

   a. James Madison, William Howard Taft, Grover Cleveland, and Chester Arthur
   b. James Madison, Andrew Jackson, William Howard Taft, and Grover Cleveland
   c. James Madison, William Howard Taft, and Grover Cleveland
   d. James Madison and William Howard Taft
   e. James Madison only
   f. William Howard Taft and Grover Cleveland
   g. William Howard Taft only
   h. No presidents are outliers

6. When Barack Obama left office, his weight was in the 33rd percentile. Which of the following statements are true?

   I. His weight was between 175 and 210 pounds.

II. He was heavier than 33% of U.S. presidents.
III. The observation is contained in the lower half of the box in a boxplot.

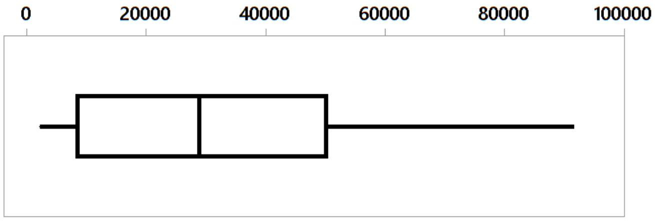
a. None are true
b. I only
c. II only
d. III only
e. I and II
f. I and III
g. II and III
h. I, II, and III

7. Suppose William Howard Taft's weight was incorrectly recorded as 432 pounds instead of 332 pounds. Which of the following statistics would **increase** as a result of this error?

I. Mean
II. Median
III. Standard Deviation

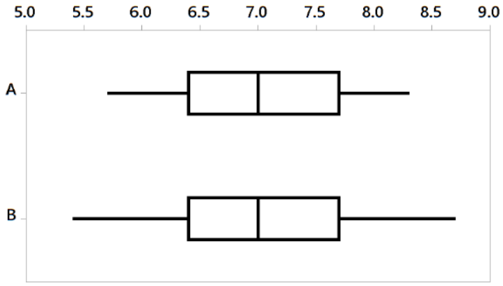
a. None would increase
b. I only
c. II only
d. III only
e. I and II
f. I and III
g. II and III
h. I, II, and III

8. The following boxplot displays the number of people incarcerated in each of the 50 states. What would a histogram of the data look like based on the boxplot?



a. Symmetric with Mean ≈ Median
b. Left-skewed with Mean > Median
c. Left-skewed with Mean < Median
d. Right-skewed with Mean > Median
e. Right-skewed with Mean < Median

9. The following boxplots compare the weights of newborn babies (in pounds) at two different hospitals on a single day. Which of the following are true statements about the boxplots?



I. There are more babies in the first quartile at Hospital B than there are in the first quartile of at Hospital A.
II. The median weights and interquartile ranges at both hospitals are both approximately the same.
III. The range of weights at Hospital B is larger than the range of weights at Hospital A.

a. None are true
b. I only
c. II only
d. III only
e. I and II
f. I and III
g. II and III
h. I, II, and III

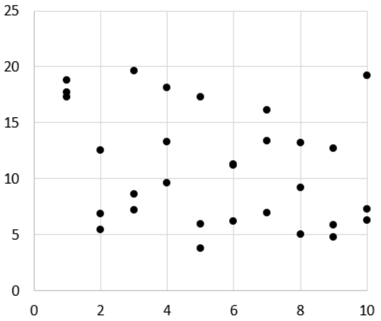
10. We want to use the number of runs scored by MLB teams in 2017 to predict the number of runs they allowed. An analysis of the data reveals the following summary statistics. What type of linear relationship exists between runs scored and runs allowed?

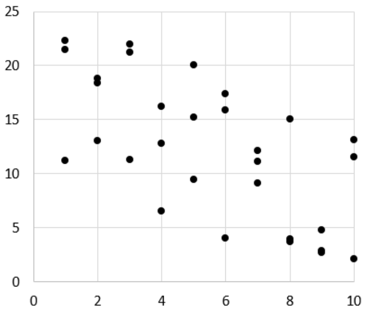
Covariance	$s_{RUNS\ SCORED}$	$s_{RUNS\ ALLOWED}$
-2219.04	65.19	82.85

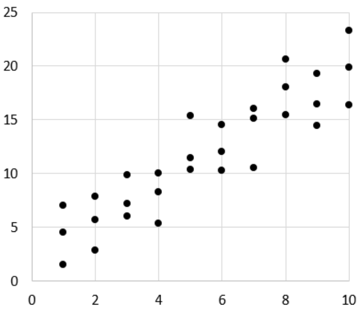
a. Strong negative linear relationship
b. Moderate negative linear relationship
c. Weak negative linear relationship
d. No linear relationship
e. Weak positive linear relationship
f. Moderate positive linear relationship
g. Strong positive linear relationship

11. Three scatterplots have predictor values between 0 and 10 and responses between 0 and 25. Order the scatterplots from weakest correlation to strongest correlation.

A B C







a. A, B, C
b. A, C, B
c. B, A, C
d. B, C, A
e. C, A, B
f. C, B, A

Use the following scenario for Questions 12 and 13.

Climatologists believe that Los Angeles has been getting significantly less rain each year since 2000. To test this theory, they use the number of inches of rain as the response and the number of years since 2000 as the predictor. (i.e. The value of the predictor for 2000 would be 0, the value of the predictor for 2001 would be 1, etc.) After fitting a linear regression line to the data, they find an intercept of 13.34, a slope of -0.26, and a correlation of $-.60$.

12. Which of the following statements are true?

I. There is a moderate, positive linear relationship between years since 2000 and rainfall
II. Because the year 2010 had 16.36 inches of rain, the point on the scatterplot falls above the regression line.
III. As the number of years since 2000 increases, the predicted rainfall each year also increases.

a. None are true
b. I only
c. II only
d. III only
e. I and II
f. I and III
g. II and III
h. I, II, and III

13. Assuming this trend holds, what would be the predicted rainfall for Los Angeles in 2017?

a. 3.14
b. 8.92
c. 13.08

d. 13.6
e. 17.76

Quiz 4

1. Suppose three cards are picked at random from a full deck. Which of the following events are intersections?

   I. Selecting at least one face card
   II. Selecting three clubs
   III. Having the individual cards be the 7 of hearts, 8 of diamonds, and 9 of spades

   a. None
   b. I only
   c. II only
   d. III only
   e. I and II
   f. I and III
   g. II and III
   h. I, II, and II

Use the following scenario for Questions 2, 3, and 4.

The following table of probabilities compares the location of a person's job (urban, suburban, or rural) with the mode of transportation they use to get to their job.

	Car	Bus	Walk	Total
Urban	.37	.25	.03	.65
Suburban	.13	.12	.02	.27
Rural	.04	.03	.01	.08
Total	.54	.40	.06	1.00

2. Given that a person's job is in a suburban area, what is the probability that they take the bus to work?

   a. .120
   b. .270
   c. .300
   d. .400
   e. .444
   f. .550
   g. .670

3. What is the probability that a person does not work in an urban area and also does not take the bus to work?

a. .20
b. .21
c. .26
d. .35
e. .40
f. .60
g. .95
h. 1.05

4. What is the probability that a person works in a rural area or drives a car to work?

a. .0400
b. .0741
c. .1481
d. .4320
e. .5000
f. .5768
g. .5800
h. .6200

5. An urn contains 5 white marbles and 10 black marbles. You select one marble at random from the urn, put it back in, and select a second marble. Let A be the event that the first marble is white and B be the event the second marble is white. Which of the following best describes events A and B?

a. Independent, but not mutually exclusive
b. Mutually exclusive, but not independent
c. Both independent and mutually exclusive
d. Neither independent nor mutually exclusive

6. A completely indifferent voter is going to vote for two members of city council by randomly selecting two people on the ballot. Of the 12 people up for election, 4 are Democrats and 8 are Republicans. If the voter randomly selects two candidates without replacement, what is the probability both are Democrats?

a. .0833
b. .0909
c. .1111
d. .1212
e. .2500
f. .5000
g. .6060

7. The following probability distributions show the number of interceptions thrown by the Pittsburgh Steelers and Cleveland Browns during a football game. Assume that the number of interceptions thrown by each team is independent of the number thrown by the other team.

Pittsburgh Steelers

Interceptions	0	1	2	3	4	5
Probability	.20	.35	.25	.10	.05	.05

Cleveland Browns

Interceptions	0	1	2	3	4	5
Probability	.35	.20	.20	.15	.10	.05

What is the probability that both teams throw more than 2 interceptions in a game?

a. .0500
b. .0600
c. .2250
d. .3000
e. .3025
f. .4500
g. .5000
h. .6000
i. .9500

8. The probability distributions of two spinners used in a board game are displayed below:

Spin (A)	1	2	3	4
Probability	.10	.50	.30	.10

Spin (B)	1	3	5	7
Probability	.10	.50	.30	.10

Which of the following statements accurately describes the mean and standard deviations of the distributions?

a. The mean of spinner A is larger and the standard deviation of spinner A is larger
b. The mean of spinner A is larger and the standard deviation of spinner B is larger
c. The mean of spinner A is larger and the standard deviations of both spinners are the same
d. The mean of spinner B is larger and the standard deviation of spinner A is larger
e. The mean of spinner B is larger and the standard deviation of spinner B is larger
f. The mean of spinner B is larger and the standard deviations of both spinners are the same
g. The means of both spinners are the same and the standard deviation of spinner A is larger
h. The means of both spinners are the same and the standard deviation of spinner B is larger
i. The means and standard deviations of both spinners are the same

9. The cost of a kitchen appliance at a hardware store has a population mean of $800 with a population standard deviation of $200. The owner of the store decides to run a sale where everything is 25% off. What are the new mean and standard deviation of the sales price of kitchen appliances?

a. $\mu = 150$ and $\sigma = 50$
b. $\mu = 150$ and $\sigma = 150$
c. $\mu = 150$ and $\sigma = 200$
d. $\mu = 600$ and $\sigma = 50$
e. $\mu = 600$ and $\sigma = 150$
f. $\mu = 600$ and $\sigma = 200$

g. $\mu = 800$ and $\sigma = 50$
h. $\mu = 800$ and $\sigma = 150$
i. $\mu = 800$ and $\sigma = 200$

Quiz 5

1. 86% of school aged children have said that summer is their favorite season. Suppose 5 children are independently sampled. What is the probability that at least one says summer is **not** their favorite season?

   a. .0766
   b. .3829
   c. .4704
   d. .5296
   e. .6171
   f. .9234

2. In the United States, 34% of the population has a college degree. Suppose that out of 400 Americans who were surveyed, 100 of them had a college degree. Which of the following describes how unusual it would be to see 100 people out of a sample of 400 with a college degree?

   a. 100 is extremely low compared to the number of Americans we would expect to have a college degree
   b. 100 is somewhat low compared to the number of Americans we would expect to have a college degree
   c. 100 is very close to the number of Americans we would expect to have a college degree
   d. 100 is somewhat high compared to the number of Americans we would expect to have a college degree
   e. 100 is extremely high compared to the number of Americans we would expect to have a college degree

3. Three bowlers bowl a single game against one another. Each bowler's results are presented below along with the population mean and population standard deviation of their typical bowling scores. Assume all three distributions are normal. (Note that higher bowling scores are better.)

   - Bowler A: Bowled 170; $\mu = 150$ and $\sigma = 20$
   - Bowler B: Bowled 220; $\mu = 210$ and $\sigma = 25$
   - Bowler C: Bowled 130; $\mu = 110$ and $\sigma = 10$

   Order the bowlers from best to worst score relative to their typical bowling scores.

   a. A, B, C
   b. A, C, B
   c. B, A, C

d. B, C, A
e. C, A, B
f. C, B, A

Use the following scenario for Questions 4, 5, and 6.

Gas prices have increased throughout the years, but have tended to remain normally distributed. The mean and standard deviation for a gallon of gas (in dollars) during four different years are presented below. Round any Z-scores to the nearest hundredth.

- 1950: $\mu = 0.27$ and $\sigma = 0.03$
- 1970: $\mu = 0.39$ and $\sigma = 0.05$
- 1990: $\mu = 1.12$ and $\sigma = 0.12$
- 2010: $\mu = 2.74$ and $\sigma = 0.24$

4. What is the probability of paying between \$0.24 and \$0.28 for a gallon of gas in 1950?

   a. .0672
   b. .1587
   c. .2514
   d. .4706
   e. .6293
   f. .7486
   g. .7880
   h. .9082

5. Suppose that a randomly selected person spent more than 21.19% of other people for a gallon of gas in 1970. How much did this person pay for a gallon of gas, rounded to the nearest cent?

   a. .08
   b. .35
   c. .38
   d. .40
   e. .43
   f. .47

6. Suppose a person spent \$0.91 for a gallon of gas in 1990. What would the equivalent price have been in 2010?

   a. 2.32
   b. 2.53
   c. 2.75
   d. 2.95
   e. 2.97
   f. 3.16

Quiz 6

1. The average amount of weight that a person gains while on vacation follows a normal distribution with a mean of 1.1 pounds and standard deviation of 2.5 pounds. Suppose we take a random sample of 25 people who just returned from vacation and find out the amount of weight gained. What is the sampling distribution of the sample mean amount of weight gained for this sample of 25 people?

    a. Normal with mean $\mu_{\bar{X}} = 1.1$ pounds and standard error $\sigma_{\bar{X}} = 0.5$ pounds
    b. Normal with mean $\mu_{\bar{X}} = 1.1$ pounds and standard error $\sigma_{\bar{X}} = 0.1$ pounds
    c. Normal with mean $\mu_{\bar{X}} = .044$ pounds and standard error $\sigma_{\bar{X}} = 0.5$ pounds
    d. Normal with mean $\mu_{\bar{X}} = .044$ pounds and standard error $\sigma_{\bar{X}} = 0.1$ pounds
    e. Undetermined shape with mean $\mu_{\bar{X}} = 1.1$ pounds and standard error $\sigma_{\bar{X}} = 0.5$ pounds
    f. Undetermined shape with mean $\mu_{\bar{X}} = 1.1$ pounds and standard error $\sigma_{\bar{X}} = 0.1$ pounds
    g. Undetermined shape with mean $\mu_{\bar{X}} = .044$ pounds and standard error $\sigma_{\bar{X}} = 0.5$ pounds
    h. Undetermined shape with mean $\mu_{\bar{X}} = .044$ pounds and standard error $\sigma_{\bar{X}} = 0.1$ pounds

2. Suppose the weights of adults are normally distributed with a mean of 185 pounds and a standard deviation of 30 pounds. An elevator can hold 2000 pounds. If 10 people get on the elevator, what is the probability that the average weight of the passengers exceeds the amount it can hold? (Hint: Think about what the average person would have to weigh to equal the weight limit for the elevator.)

    a. .0571
    b. .1984
    c. .3085
    d. .6915
    e. .7688
    f. .9429

3. Let X be the random variable denoting the amount of money a customer spends on concessions at a movie theater. Suppose X follows a normal distribution with a mean of \$7.50 and a standard deviation of \$2.40. Suppose we randomly sample 9 customers and average the amount they spent on concessions. The probability that this sample mean exceeds \$8 is .2659; that is, $P(\bar{X} > 8) = .2659$. Which of the following changes would result in this probability decreasing?

    I. Decrease the population mean amount spent to \$7
    II. Decrease the population standard deviation to \$1.80
    III. Decrease the number of customers sampled to 4

    a. None would decrease the probability
    b. I only
    c. II only
    d. III only

e. I and II
f. I and III
g. II and III
h. I, II, and III

4. The proportion of adults in this country who watch the show *This Is Us* is .04. Suppose you take a random sample of 200 adults and ask if they watch *This Is Us*. What is the sampling distribution of the sample proportion of adults who watch the show in this sample?

a. Normally distributed with mean $\mu_{\hat{p}} = 8$ and standard error $\sigma_{\hat{p}} = 2.771$
b. Normally distributed with mean $\mu_{\hat{p}} = .04$ and standard error $\sigma_{\hat{p}} = .196$
c. Normally distributed with mean $\mu_{\hat{p}} = .04$ and standard error $\sigma_{\hat{p}} = .0138$
d. Undetermined shape with mean $\mu_{\hat{p}} = 8$ and standard error $\sigma_{\hat{p}} = 2.771$
e. Undetermined shape with mean $\mu_{\hat{p}} = .04$ and standard error $\sigma_{\hat{p}} = .196$
f. Undetermined shape with mean $\mu_{\hat{p}} = .04$ and standard error $\sigma_{\hat{p}} = .0138$

5. In a large city, 19% of people are in favor of building a new housing development in a park. Suppose a random sample of 100 people is selected. What is the probability that less than one-quarter of the people sampled support building the housing development? (Use four decimal places in the calculation of the standard error.)

a. .0630
b. .0823
c. .3612
d. .5182
e. .9177
f. .9370

Quiz 7

Use the following scenario for Questions 1-5.

In a laboratory, a standard weight that is known to weigh 1000 grams is repeatedly weighed 9 times on the same scale to test if the scale is calibrated correctly. The resulting measurements yielded a sample mean weight of 998 grams. Assume that the weightings by the scale are normally distributed and that the population standard deviation of the measurements is 4 grams.

1. Which of the following represents a 95% confidence interval for the mean weights reported by the scale?

a. $998 \pm 1.645\left(\frac{4}{\sqrt{9}}\right) = (995.81, 1000.19)$
b. $1000 \pm 1.645\left(\frac{4}{\sqrt{9}}\right) = (997.81, 1002.19)$
c. $998 \pm 1.96\left(\frac{4}{\sqrt{9}}\right) = (995.39, 1000.61)$

d. $1000 \pm 1.96\left(\frac{4}{\sqrt{9}}\right) = (997.39, 1002.61)$
e. $998 \pm 2.576\left(\frac{4}{\sqrt{9}}\right) = (994.57, 1001.43)$
f. $1000 \pm 2.576\left(\frac{4}{\sqrt{9}}\right) = (996.57, 1003.43)$

2. Based on the confidence interval, does it appear as if the scale needs to be recalibrated? Why?

a. Yes: 95% confidence interval contains 1000
b. Yes: 95% confidence interval is entirely above 1000
c. Yes: 95% confidence interval is entirely below 1000
d. No: 95% confidence interval contains 1000
e. No: 95% confidence interval is entirely above 1000
f. No: 95% confidence interval is entirely below 1000

3. Which of the following would have led to a narrower confidence interval?

I. Sample mean of 1001 grams
II. Population standard deviation of 5 grams
III. Using a 99% level of confidence

a. None would have led to a narrower interval
b. I only
c. II only
d. III only
e. I and II
f. I and III
g. II and III
h. I, II, and III

4. Suppose the true population mean reading provided by the scale is actually 1000. If this process of weighing a 1000 gram weight was repeated 60 times and a 95% confidence interval was calculated for each, how many of these confidence intervals would we expect to contain 1000?

a. 0
b. 3
c. 30
d. 57
e. 60

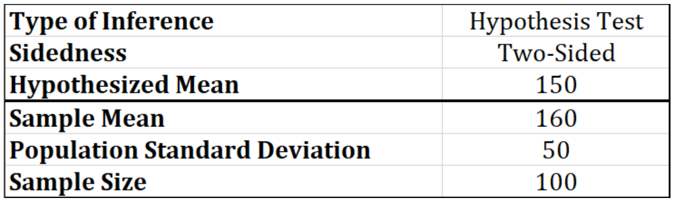
5. Suppose we wanted to obtain an interval that was 1.5 grams wide, but to be more certain we capture the population mean, we want to use 99% confidence instead. How large a sample would need to be taken to attain this width?

a. 7

b. 14
c. 48
d. 189

Use the following scenario for Questions 6-10.

According to a 2010 demographic report, a typical American household spends $150 per day. Suppose a random sample of Pittsburgh households was surveyed regarding how much money they spend in a typical day. The results of the survey are presented in the output below. Use a 10% level of significance.



Type of Inference	Hypothesis Test
Sidedness	Two-Sided
Hypothesized Mean	150
Sample Mean	160
Population Standard Deviation	50
Sample Size	100

6. What is/are the critical value(s) of this test?

   a. $Z = 1.28$
   b. $Z = \pm 1.645$
   c. $Z = 1.645$
   d. $Z = \pm 1.96$
   e. $Z = 2.326$
   f. $Z = \pm 2.576$
   g. $Z = 2.00$
   h. $Z = -2.00$
   i. $Z = \pm 2.00$

7. What is the p-value of this test?

   a. .0114
   b. .0228
   c. .0456
   d. .9544
   e. .9772
   f. .9886

8. Based on the results of this test, what conclusion can we come to regarding the average amount that Americans spend each day?

   a. Average amount Pittsburghers spend each day is significantly less than $150
   b. Average amount Pittsburghers spend each day is significantly different from $150 with an indication it may be higher
   c. Average amount Pittsburghers spend each day is significantly different from $150 with an indication it may be lower

d. Average amount Pittsburghers spend each day is equal to \$150
e. Average amount Pittsburghers spend each day is not significantly different from \$150

9. The result of a Type II error would be concluding that the true average amount Pittsburghers spend each day is:

    a. Equal to \$150 when it is really lower than \$150.
    b. Equal to \$150 when it is really some value other than \$150.
    c. Lower than \$150 when it is really equal to \$150.
    d. Equal to some value other than \$150 when it is really equal to \$150.
    e. Equal to \$150 when it is really equal to \$160.

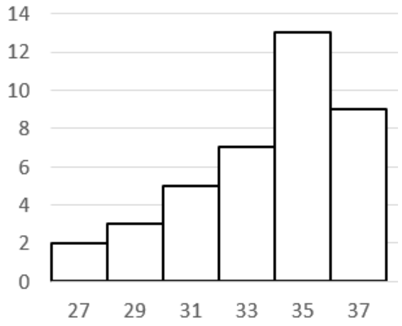
10. Which of the following describes what we could conclude about a 90% confidence interval for the mean amount that Pittsburghers spend daily?

    a. It would contain both 150 and 160
    b. It would contain 150, but not 160
    c. It would contain 160, but not 150
    d. It would contain neither 150 nor 160
    e. It would contain 150, but we cannot determine if it would contain 160
    f. It would not contain 150, but we cannot determine if it would contain 160
    g. It would contain 160, but we cannot determine if it would contain 150
    h. It would not contain 160, but we cannot determine if it would contain 150

Quiz 8

Use the following scenario for Questions 1-6.

The following output contains the ages at which a random sample of professional baseball players retired. Major League Baseball is interested in determining if the average retirement age is younger than 35.



Type of Inference	Hypothesis Test
Sidedness	
Hypothesized Mean	35
Sample Mean	34.33
Sample Standard Deviation	2.94
Sample Size	39
Degrees of Freedom	38
Test Statistic	
P-Value	0.0814

1. What are the null and alternative hypotheses?

    a. $H_0: \mu = 35$ vs. $H_1: \mu > 35$
    b. $H_0: \mu = 35$ vs. $H_1: \mu < 35$

c. $H_0: \mu = 35$ vs. $H_1: \mu \neq 35$
d. $H_0: \mu > 35$ vs. $H_1: \mu = 35$
e. $H_0: \mu < 35$ vs. $H_1: \mu = 35$
f. $H_0: \mu \neq 35$ vs. $H_1: \mu = 35$

2. What can we determine about the conditions needed to perform this test?

a. Satisfied because the histogram is approximately normal
b. Satisfied because the sample size is large enough
c. Not satisfied because the original population is not known to be normal
d. Not satisfied because the histogram is severely skewed
e. Not satisfied because the histogram is severely skewed and the original population is not known to be normal

3. What are the distribution and value of the test statistic?

a. $Z = \frac{34.33-35}{2.94/\sqrt{39}} = -1.423$
b. $Z = \frac{35-34.33}{2.94/\sqrt{39}} = 1.423$
c. $t = \frac{34.33-35}{2.94/\sqrt{39}} = -1.423$
d. $t = \frac{35-34.33}{2.94/\sqrt{39}} = 1.423$
e. $t = \frac{34.33-35}{2.94/\sqrt{38}} = -1.405$
f. $t = \frac{35-34.33}{2.94/\sqrt{38}} = 1.405$

4. Which of the following would have led to a smaller p-value?

I. Hypothesized mean of 34
II. Sample size of 50
III. Sample standard deviation of 4

a. None would have led to a smaller p-value
b. I only
c. II only
d. III only
e. I and II
f. I and III
g. II and III
h. I, II, and III

5. For which levels of significance would the null hypothesis be rejected?

a. Do not reject the null hypothesis for $\alpha = .01$, $\alpha = .05$, or $\alpha = .10$

b. Reject for $\alpha = .01$ and $\alpha = .05$, but not for $\alpha = .10$
c. Reject for $\alpha = .10$, but not for $\alpha = .01$ or $\alpha = .05$
d. Reject for $\alpha = .01$, $\alpha = .05$, and $\alpha = .10$

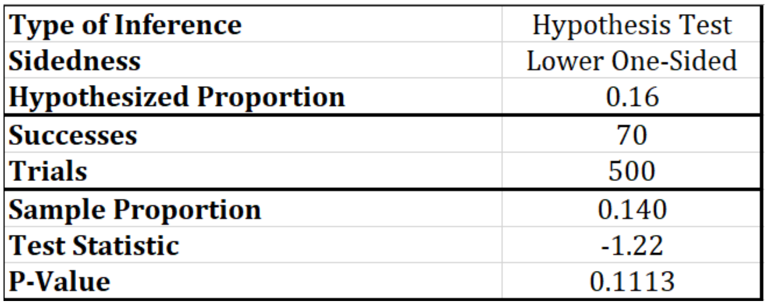
6. Suppose instead that we had known the population standard deviation was 2.94. Which of the following would have been different?

I. Hypotheses
II. Value of the test statistic
III. P-value

a. None would have been different
b. I only
c. II only
d. III only
e. I and II
f. I and III
g. II and III
h. I, II, and III

Use the following scenario for Questions 7-12.

Nationwide, 16% of adults smoke cigarettes. A survey was conducted to estimate the proportion of students at a large university who are smokers. Consider sampling a person who is a smoker a success. Use the output below to determine if the proportion of students who smoke at this university is less than the national proportion at the 5% level of significance.



Type of Inference	Hypothesis Test
Sidedness	Lower One-Sided
Hypothesized Proportion	0.16
Successes	70
Trials	500
Sample Proportion	0.140
Test Statistic	-1.22
P-Value	0.1113

7. What are the null and alternative hypotheses?

a. $H_0: \mu = .16$ vs. $H_1: \mu < .16$
b. $H_0: \bar{x} = .16$ vs. $H_1: \bar{x} < .16$
c. $H_0: p = .16$ vs. $H_1: p < .16$
d. $H_0: \hat{p} = .16$ vs. $H_1: \hat{p} < .16$

8. What is the correct interpretation of the p-value?

a. The probability that less than 16% of students who smoke is .1113

b. If the true proportion of students who smoke is actually .16, then the probability of obtaining a sample proportion of exactly .140 is .1113.
c. The probability that 16% is the true percentage of students who smoke is .1113.
d. If the true proportion of students who smoke is actually .16, then the probability of obtaining a sample proportion less than or equal to .16 is .1113.

9. What decision and conclusion can be made based on the results of this test?

a. Reject H_0 and conclude that less than 16% of students are smokers
b. Fail to reject H_0 and conclude that less than 16% of students are smokers
c. Reject H_0 and conclude that the percentage of students who smoke is not less than 16%
d. Fail to reject H_0 and conclude that the percentage of students who smoke is not less than 16%

10. Suppose the true proportion of students who smoke is actually .10. What type of decision was made based on the results of the hypothesis test?

a. Type I error
b. Type II error
c. Correct decision

11. If a two-sided test had been used instead, what would the p-value have been?

a. .0557
b. .025
c. .1113
d. .10
e. .2226

12. What is the probability of making a Type I error?

a. .025
b. .05
c. .1113
d. .8887
e. .95

Quiz 9

Use the following scenario for Questions 1-4.

A simple random sample of 200 married people were asked the question, “Would you remarry your spouse a second time if you were given the chance?”. Based on the 200 responses, a 95% confidence interval for the proportion of people who reported they would remarry their spouse is $(.812, .908)$.

1. Which of the following is the correct interpretation of the confidence interval?

   a. 95% of all confidence intervals calculated regarding the proportion of people who would remarry their spouse a second time will contain the sample proportion of .86.
   b. The probability that 95% of people would remarry their spouse is between .812 and .908.
   c. We are 95% confident that the true proportion of people who would remarry their spouse is between 81.2% and 90.8%.
   d. We are 95% confident that between 81.2% and 90.8% of the next 200 married people surveyed would be willing to remarry their spouse.

2. If the sample proportion had been .90 instead of .86 as observed in the original poll while still using a sample size of 200 and a 95% confidence interval, then the confidence interval would have been:

   a. Wider
   b. Narrower
   c. The same width

3. Which of the following statements are true based on the confidence interval?

   I. The proportion of people who would remarry their spouse appears to be significantly greater than .80.
   II. A 90% confidence interval would contain .80
   III. .90 is a plausible estimate for the true proportion of people who would remarry their spouse.

   a. None are true
   b. I only
   c. II only
   d. III only
   e. I and II
   f. I and III
   g. II and III
   h. I, II, and III

4. Suppose a simple random sample of 800 recently divorced men are surveyed in a separate poll and asked the same question regarding if they would remarry their spouse. A 95% confidence interval is calculated for these responses as well. Which of the following statements is **not** true?

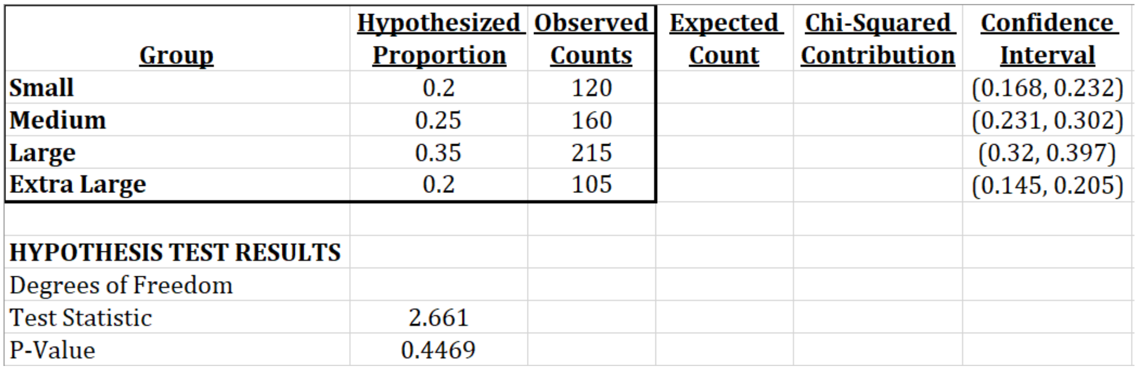
   a. The interval calculated from the sample of divorced men will be narrower than the interval calculated by taking the sample from the population of people who are currently married.
   b. The interval calculated from the sample of divorced men is a closer approximation of the proportion of people who would remarry their spouse because the sample size is larger.

c. The sample proportion of divorced men who reported they would remarry their spouse should not be used to approximate the proportion of all people who would remarry their spouse because the estimate is likely biased downward.

d. The intervals are estimating different parameters so no conclusions can be made about the proportion of all married people who would remarry their spouse based on the interval that consisted of only divorced men.

Use the following scenario for Questions 5-8.

A retail store wants to test its hypothesized proportions of T-shirts sold by size against sales in the past month. The hypothesized proportions and observed counts for each of the four sizes (small, medium, large, and extra-large) are displayed below. Note that 600 shirts were sold in total. Use the output to answer the following questions using a 5% level of significance.



	Hypothesized	**Observed**	**Expected**	**Chi-Squared**	**Confidence**
Group	**Proportion**	**Counts**	**Count**	**Contribution**	**Interval**
Small	0.2	120			(0.168, 0.232)
Medium	0.25	160			(0.231, 0.302)
Large	0.35	215			(0.32, 0.397)
Extra Large	0.2	105			(0.145, 0.205)
HYPOTHESIS TEST RESULTS					
Degrees of Freedom					
Test Statistic	2.661				
P-Value	0.4469				

5. What are the null and alternative hypotheses?

a. $H_0: \hat{p}_S = .20, \hat{p}_M = .25, \hat{p}_L = .35, \hat{p}_{XL} = .20$
H_1: At least one proportion differs significantly from its hypothesized value

b. $H_0: p_S = .20, p_M = .25, p_L = .35, p_{XL} = .20$
H_1: At least one proportion differs significantly from its hypothesized value

c. $H_0: \mu_S = .20, \mu_M = .25, \mu_L = .35, \mu_{XL} = .20$
H_1: At least one mean differs significantly from its hypothesized value

d. $H_0: \hat{p}_S = .20, \hat{p}_M = .25, \hat{p}_L = .35, \hat{p}_{XL} = .20$
H_1: All proportions differ significantly from their hypothesized values

e. $H_0: p_S = .20, p_M = .25, p_L = .35, p_{XL} = .20$
H_1: All proportions differ significantly from their hypothesized values

f. $H_0: \mu_S = .20, \mu_M = .25, \mu_L = .35, \mu_{XL} = .20$
H_1: All means differ significantly from their hypothesized values

6. What is the chi-squared contribution from the category ‘Extra Large’?

a. -0.143
b. -0.125
c. 0.125
d. 0.143

e. 1.875
f. 2.143

7. How many degrees of freedom does this test have?

a. 3
b. 4
c. 599
d. 600

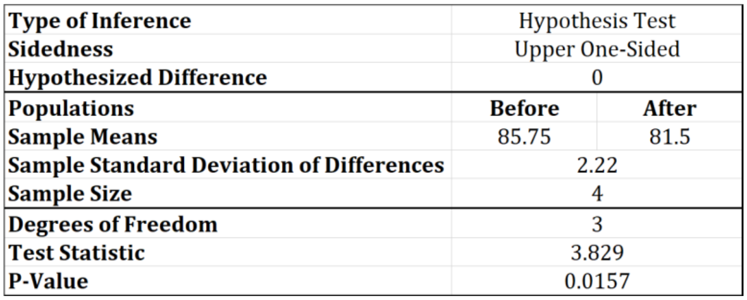
8. What conclusion can we come to based on the results of the test and the confidence intervals?

a. None of the proportions of shirt sizes sold differ significantly from their hypothesized proportions.
b. The proportion of small, medium, and extra-large shirts differ significantly from their hypothesized proportions.
c. The proportion of small, large, and extra-large shirts differ significantly from their hypothesized proportions.
d. Only the proportion of large shirts sold differs significantly from its hypothesized proportion.
e. Only the proportion of extra-large shirts sold differs significantly from its hypothesized proportion.
f. The proportion of all four shirt sizes differ significantly from their hypothesized proportions.

Quiz 10

Use the following scenario for Problems 1-3.

A golf instructor wants to determine how effective his golf lessons are. He has four new students record their scores in a round of golf before beginning the lessons. After each student received six lessons, they play another round of golf and record their scores. Assume that golf score differences are normal. Use the output to determine if the students' golf scores were reduced using $\alpha = .05$. (Note: It is better to have a lower score in golf than a higher score.)



Type of Inference	Hypothesis Test	
Sidedness	Upper One-Sided	
Hypothesized Difference	0	
Populations	**Before**	**After**
Sample Means	85.75	81.5
Sample Standard Deviation of Differences	2.22	
Sample Size	4	
Degrees of Freedom	3	
Test Statistic	3.829	
P-Value	0.0157	

1. Which of the following is true?

a. Samples are independent; Golf score is categorical and golfing before or after lessons is quantitative
b. Samples are dependent: Golf score is categorical and golfing before or after lessons is quantitative
c. Samples are independent: Golf score is quantitative and golfing before or after lessons is categorical
d. Samples are dependent: Golf score is quantitative and golfing before or after lessons is categorical

2. Which of the following is **not** a correct way to write the hypotheses?

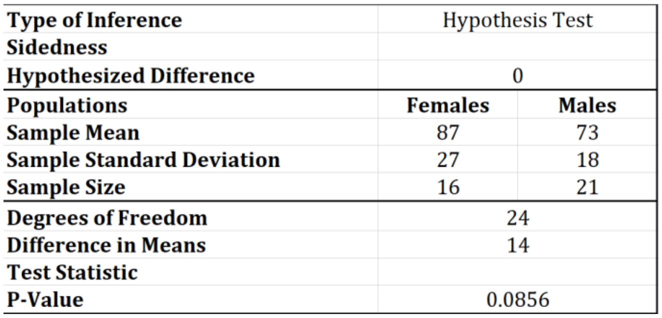
a. $H_0: \mu_B - \mu_A = 0$ vs. $H_1: \mu_B - \mu_A > 0$
b. $H_0: \mu_B = \mu_A$ vs. $H_1: \mu_B > \mu_A$
c. $H_0: \mu_D = 0$ vs. $H_1: \mu_D > 0$
d. $H_0: \mu_B = \mu_A = 0$ vs. $H_1: \mu_B \neq 0$ or $\mu_A \neq 0$

3. What conclusion can you draw based on the result of the hypothesis test?

a. Average golf scores were significantly greater after the lessons.
b. Average golf scores were significantly lower after the lessons.
c. Average golf scores were higher before the lessons, but the difference is not significant
d. Average golf scores were lower before the lessons, but the difference is not significant

Use the following scenario for Problems 4-9.

Suppose we are interested in testing if the average number of texts that females send during a day differs from the average number that males send. Assume both populations are normally distributed. Use the Excel output to answer the following questions at the 5% level of significance.



Type of Inference	Hypothesis Test	
Sidedness		
Hypothesized Difference	0	
Populations	**Females**	**Males**
Sample Mean	87	73
Sample Standard Deviation	27	18
Sample Size	16	21
Degrees of Freedom	24	
Difference in Means	14	
Test Statistic		
P-Value	0.0856	

4. What are the hypotheses?

a. $H_0: \mu_F - \mu_M = 0$ vs. $H_1: \mu_F - \mu_M > 0$
b. $H_0: \mu_F - \mu_M = 0$ vs. $H_1: \mu_F - \mu_M \neq 0$
c. $H_0: \mu_F - \mu_M = 0$ vs. $H_1: \mu_F - \mu_M < 0$
d. $H_0: \mu_F - \mu_M \neq 0$ vs. $H_1: \mu_F - \mu_M = 0$

5. What is the test statistic?

   a. $t = \frac{87-73}{16^2+21^2/\sqrt{37}} = 0.122$
   b. $t = \frac{87-73}{45/\sqrt{37}} = 1.892$
   c. $t = \frac{87-73}{\sqrt{\frac{27}{16}+\frac{18}{21}}} = 8.776$
   d. $t = \frac{87-73}{\sqrt{\frac{27^2}{16}+\frac{18^2}{21}}} = 1.792$

6. What conclusion can you come to about the number of texts sent by females compared to males based on the result of the above test?

   a. Mean number of texts sent by females is greater than the mean for males, but the difference is not significant.
   b. Mean number of texts sent by females is less than the mean for males, but the difference is not significant.
   c. Mean number of texts sent by females is significantly different from the mean number sent by males with an indication it is greater.
   d. Mean number of texts sent by females is significantly different from the mean number sent by males with an indication it is lower.
   e. Mean number of texts sent by females is significantly greater than the mean number set by males.
   f. Mean number of texts sent by females is significantly less than the mean number set by males.

7. Which of the following describes what we could conclude about a 95% confidence interval for the difference between the means?

   a. It would contain both 0 and 14
   b. It would contain 0, but not 14
   c. It would contain 14, but not 0
   d. It would contain neither 0 nor 14

8. Which of the following would cause the p-value to decrease?

   I. Increasing the sample means to 100 texts for the females and 80 for the males
   II. Decreasing the sample standard deviations to 20 for the females and 10 for the males
   III. Increasing the sample sizes to 32 for the females and 42 for the males

   a. None would decrease the p-value
   b. I only
   c. II only
   d. III only
   e. I and II

f. I and III
g. II and III
h. I, II, and III

9. Which of the following would have changed if we had decided to use the males as population 1 and the females as population 2?

I. Value of the test statistic
II. P-value
III. Decision

a. None would change
b. I only
c. II only
d. III only
e. I and II
f. I and III
g. II and III
h. I, II, and III

Quiz 11

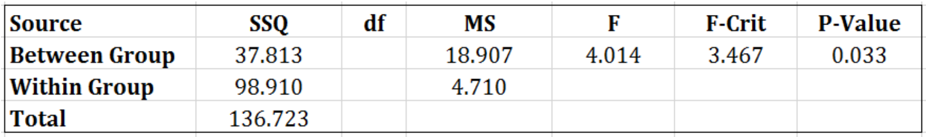
Use the following scenario for Problems 1-4.

Archaeologists recently analyzed the skeletal remains of 8 humans from each of three different time periods (Prehistoric Era, Middle Ages, and the Modern Era) to determine if there was a significant difference in their average height. Assume that the heights in all three populations are normally distributed.

1. What are the numerator and denominator degrees of freedom?

a. Numerator: 2, Denominator: 5
b. Numerator: 2, Denominator: 21
c. Numerator: 3, Denominator: 5
d. Numerator: 3, Denominator: 21

2. Using only the ANOVA table, what conclusion can we come to regarding the mean heights across the groups?



Source	SSQ	df	MS	F	F-Crit	P-Value
Between Group	37.813		18.907	4.014	3.467	0.033
Within Group	98.910		4.710			
Total	136.723					

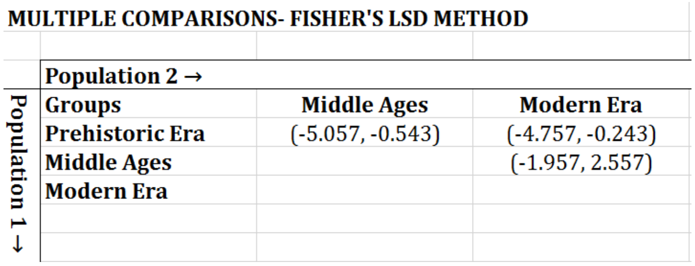
a. None of the mean heights are significantly different
b. Exactly two of the mean heights are significantly different
c. At least two of the mean heights are significantly different

d. All three mean heights are significantly different from one another

3. Which of the following would cause the p-value to decrease?

I. Increasing the sample mean of each group by 1 inch
II. Increasing the sample standard deviation of each group by 1 inch
III. Doubling the sample size of each group

a. None would cause the p-value to decrease
b. Only I
c. Only II
d. Only III
e. I and II
f. I and III
g. II and III
h. I, II, and III



MULTIPLE COMPARISONS- FISHER'S LSD METHOD

Population 2 →

Population 1 →

Groups	Middle Ages	Modern Era
Prehistoric Era	(-5.057, -0.543)	(-4.757, -0.243)
Middle Ages		(-1.957, 2.557)
Modern Era		

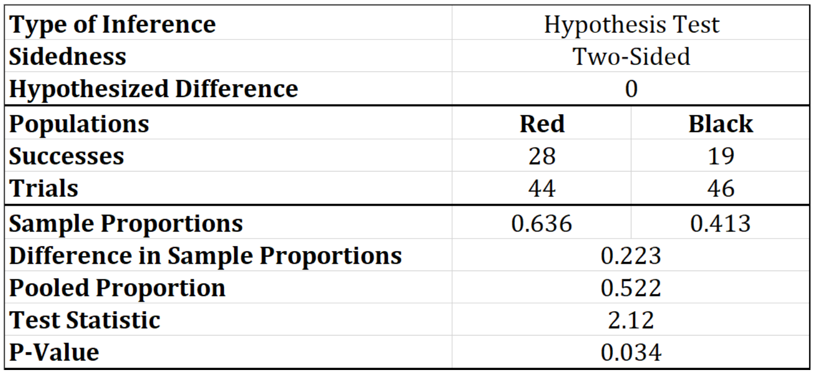
4. The above output contains 95% confidence intervals for the difference between each pair of group means using Fisher's LSD method. Use the confidence intervals to determine which pairs of means are significantly different.

I. Prehistoric Era vs. Middle Ages
II. Prehistoric Era vs. Modern Era
III. Middle Ages vs. Modern Era

a. No significant difference between any pairs
b. Only I
c. Only II
d. Only III
e. I and II
f. I and III
g. II and III
h. I, II, and III

Use the following scenario for Problems 5-6.

There are two common 4-letter English words that begin with the string "FAI": fail and fair. A college professor wants to compare the proportion of students using a red pen who fill in the blank with an “L” with the proportion of students using a black pen that finish the string with an “L”. In a large lecture hall, he randomly gives 100 students a slip of paper reading “F A I __”, as well as either a red pen or a black pen. Students are instructed to write the first letter that comes to mind that forms an English word before turning in the slips at the end of class. In the end, 90 students returned the slips. The results of the experiment are presented in the following output.



Type of Inference	Hypothesis Test	
Sidedness	Two-Sided	
Hypothesized Difference	0	
Populations	**Red**	**Black**
Successes	28	19
Trials	44	46
Sample Proportions	0.636	0.413
Difference in Sample Proportions	0.223	
Pooled Proportion	0.522	
Test Statistic	2.12	
P-Value	0.034	

5. What conclusion can we come to using a 5% level of significance?

   a. Proportion of students using a red pen who complete the word with an ‘L’ is significantly different from the proportion of students using a black pen who complete the word with an “L” with an indication that the proportion using a red pen is higher.
   b. Proportion of students using a red pen who complete the word with an ‘L’ is significantly different from the proportion of students using a black pen who complete the word with an “L” with an indication that the proportion using a black pen is higher.
   c. Proportion of students using a red pen who complete the word with an ‘L’ is greater than the proportion of students using a black pen who complete the word with an “L”, but the difference is not significant.
   d. Proportion of students using a black pen who complete the word with an ‘L’ is greater than the proportion of students using a red pen who complete the word with an “L”, but the difference is not significant.

6. Suppose we calculated a 95% confidence interval for the difference of two proportions by taking the proportion of students who finished the word with an ‘L’ using a red pen and subtracting the proportion who finished the word with an ‘L’ using a black pen. Which of the following are true about the confidence interval?

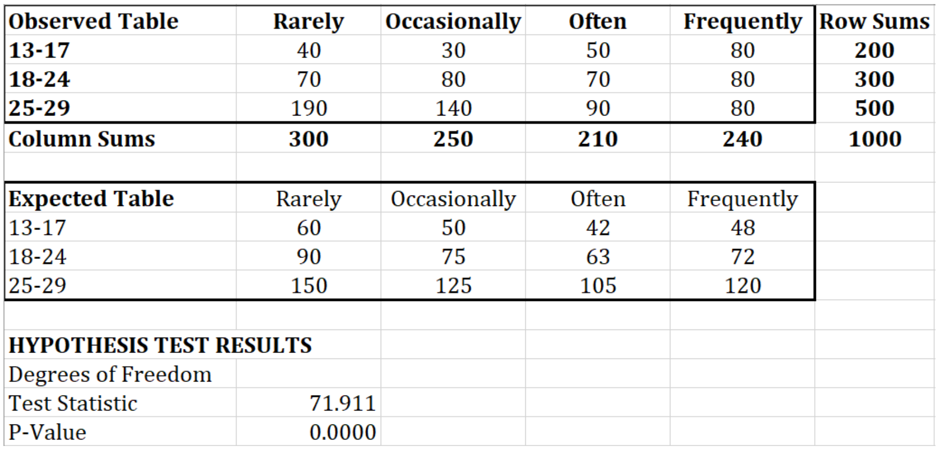
   I. The interval would contain .223
   II. Zero is a plausible value for the difference between the two proportions
   III. The entire interval would be positive

   a. None are true
   b. Only I
   c. Only II
   d. Only III
   e. I and II
   f. I and III

g. II and III
h. I, II, and III

Use the following scenario for Problems 7-10.

A statistician is interested in determining if there is a relationship between age and video game playing in teenage and young adult males. A random sample of 1000 males were classified based on their age (13-17, 18-24, or 25-29) and how frequently they played video games (rarely, occasionally, often, or frequently).



Observed Table	**Rarely**	**Occasionally**	**Often**	**Frequently**	**Row Sums**
13-17	40	30	50	80	**200**
18-24	70	80	70	80	**300**
25-29	190	140	90	80	**500**
Column Sums	**300**	**250**	**210**	**240**	**1000**
Expected Table	Rarely	Occasionally	Often	Frequently	
13-17	60	50	42	48	
18-24	90	75	63	72	
25-29	150	125	105	120	
HYPOTHESIS TEST RESULTS					
Degrees of Freedom					
Test Statistic	71.911				
P-Value	0.0000				

7. What are the null and alternative hypotheses?

a. H_0: Older gamers and younger gamers play video games at the same rate
H_1: Older gamers play video games more frequently than younger gamers
b. H_0: Older gamers and younger gamers play video games at the same rate
H_1: Younger gamers play video games more frequently than younger gamers
c. H_0: Age and video game playing frequency are independent
H_1: Age and video game playing frequency are not independent
d. H_0: Age and video game playing frequency are not independent
H_1: Age and video game playing frequency are independent

8. How many degrees of freedom does this test have?

a. 5
b. 6
c. 11
d. 12
e. 994
f. 999

9. What is the chi-squared contribution to the test statistic from the group of respondents in the 13-17 age bracket who responded that they play video games ‘occasionally’?

a. -0.40
b. -0.667
c. 8
d. 13.33
e. 30
f. 50

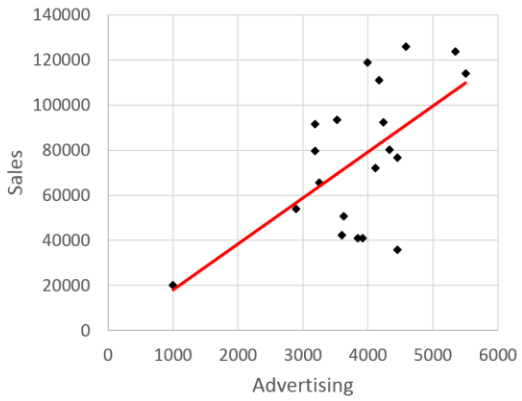
10. What conclusion can we come to based on the results of the test?

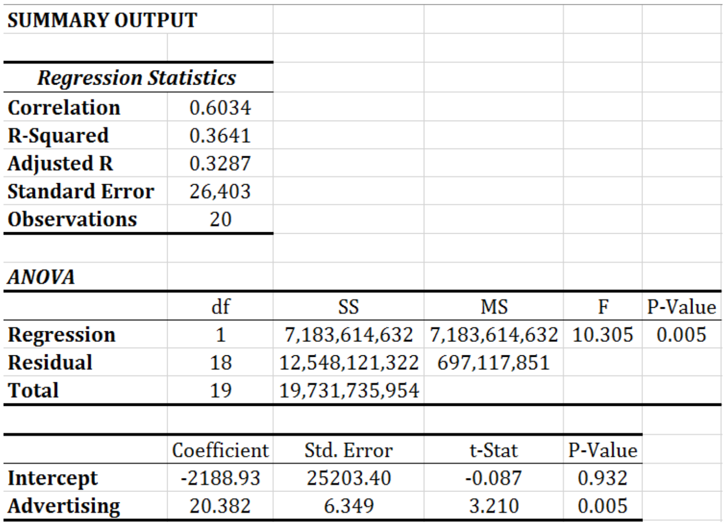
a. Age and video game playing frequency are independent.
b. Age and video game playing frequency are not independent: Older gamers are more likely to play more often than younger gamers.
c. Age and video game playing frequency are not independent: Younger gamers are more likely to play more often than older gamers.

Quiz 12

Use the following scenario for Problems 1-9.

A random sample of independently owned small businesses in Allegheny County was taken. Each was asked to anonymously report the amount of money (in dollars) spent on advertising during 2016 as well as the total amount in sales. The line on the plot is the least squares regression line. Use a 5% level of significance to make your decision on any hypothesis tests.





SUMMARY OUTPUT					
Regression Statistics					
Correlation	0.6034				
R-Squared	0.3641				
Adjusted R	0.3287				
Standard Error	26,403				
Observations	20				
ANOVA					
	df	SS	MS	F	P-Value
Regression	1	7,183,614,632	7,183,614,632	10.305	0.005
Residual	18	12,548,121,322	697,117,851		
Total	19	19,731,735,954			
	Coefficient	Std. Error	t-Stat	P-Value	
Intercept	-2188.93	25203.40	-0.087	0.932	
Advertising	20.382	6.349	3.210	0.005	

1. One business reported advertising costs of \$1000 and total sales of \$20,000. This point is located in the bottom left corner of the scatterplot. Which of the following best describes the characteristics of this point?

a. Only an outlier
b. Only an influential point
c. Both an outlier and an influential point

d. Neither an outlier nor an influential point

2. Calculate the correlation and use it to describe the strength of the linear relationship.

   a. Strong positive linear relationship
   b. Moderate positive linear relationship
   c. Weak positive linear relationship
   d. No linear relationship
   e. Weak negative linear relationship
   f. Moderate negative linear relationship
   g. Strong negative linear relationship

3. What is the linear model?

   a. $y = \beta_0 + \beta_1 x$
   b. $y = \beta_0 + \beta_1 x + \varepsilon$
   c. $\hat{y} = \beta_0 + \beta_1 x$
   d. $\hat{y} = \beta_0 + \beta_1 x + \varepsilon$
   e. $y = -2188.93 + 20.382x$
   f. $y = -2188.93 + 20.382x + \varepsilon$
   g. $\hat{y} = -2188.93 + 20.382x$
   h. $\hat{y} = -2188.93 + 20.382x + \varepsilon$

4. What is the residual for a small business that spent \$3,500 on advertising and actually made \$65,000 during the quarter? (Round your answer to the nearest dollar.)

   a. $-\$69{,}148$
   b. \$69,148
   c. \$4,148
   d. $-\$4{,}148$
   e. \$61,500
   f. $-\$61{,}500$

5. What are the null and alternative hypotheses for determining if advertising cost is a significant predictor of total sales?

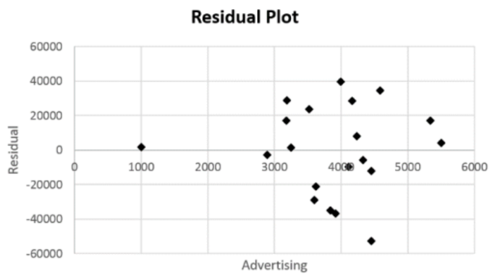
   a. $H_0{:}\,\beta_0 = 0$ vs. $H_1{:}\,\beta_0 \neq 0$
   b. $H_0{:}\,b_0 = 0$ vs. $H_1{:}\,b_0 \neq 0$
   c. $H_0{:}\,\beta_1 = 0$ vs. $H_1{:}\,\beta_1 \neq 0$
   d. $H_0{:}\,b_1 = 0$ vs. $H_1{:}\,b_1 \neq 0$

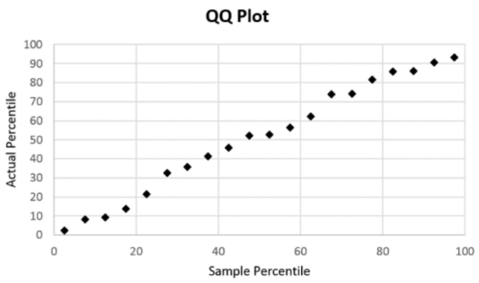
6. Based on the output what can we deduce about the relationship between advertising cost and total sales?

   a. Advertising cost is a significant predictor of total sales; relationship is positive.
   b. Advertising cost is a significant predictor of total sales; relationship is negative.

c. A positive linear relationship exists between advertising cost and total sales, but is not statistically significant
d. A negative linear relationship exists between advertising cost and total sales, but is not statistically significant

7. The residual plot (left) and QQ-plot (right) are displayed below. Which of the following best describes any violations of requirements of the error term?





a. Homoscedasticity is violated
b. Normality is violated
c. Both homoscedasticity and normality are violated
d. Neither homoscedasticity nor normality is violated

8. A 95% prediction interval for total sales at an advertising cost of $3,600 is (14,241, 128,131). Which of the following is the correct interpretation of this interval?

a. We are 95% confident that the average total sales for all stores in 2016 is between $14,241 and $128,131.
b. We are 95% confident that the average total sales for all stores in 2016 that spent $3,600 on advertising is between $14,241 and $128,131.
c. We are 95% confident that the total sales in 2016 for a single store sampled is between $14,241 and $128,131.
d. We are 95% confident that the total sales in 2016 for a single store sampled that spent $3,600 on advertising is between $14,241 and $128,131.

9. On average, these businesses spent $4,000 on advertising. A 95% prediction interval for total sales at an advertising cost of $3,600 is (14,241, 128,131). Which of the following intervals would be wider than the one provided?

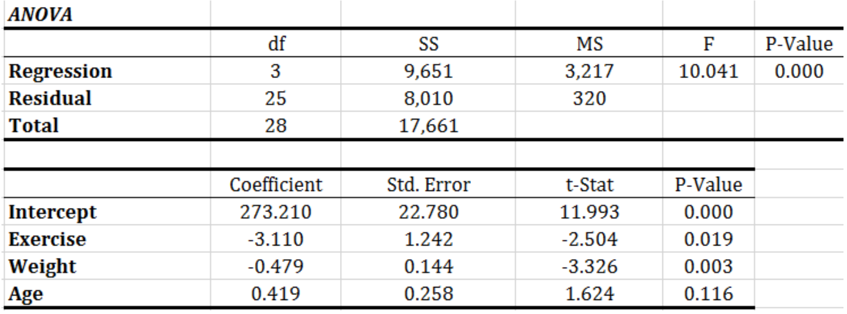
I. A 99% prediction interval for total sales with an advertising cost of $3,600
II. A 95% confidence interval for total sales with an advertising cost of $3,600
III. A 95% prediction interval for total sales with an advertising cost of $2,500

a. None would be wider
b. I only
c. II only
d. III only
e. I and II

f. I and III
g. II and III
h. I, II, and III

Use the following scenario for Problems 10-13.

Researchers looking for a way to predict a person’s cholesterol level recruited 29 people and recorded three predictor variables. The three variables used in the prediction of cholesterol level (Y) are: hours of exercise per week (X_1), weight (X_2), and age (X_3). Use the multiple linear regression output below to answer the following questions. Use a 5% level of significance to make a decision on any hypothesis tests.



ANOVA

	df	SS	MS	F	P-Value
Regression	3	9,651	3,217	10.041	0.000
Residual	25	8,010	320		
Total	28	17,661			

	Coefficient	Std. Error	t-Stat	P-Value
Intercept	273.210	22.780	11.993	0.000
Exercise	-3.110	1.242	-2.504	0.019
Weight	-0.479	0.144	-3.326	0.003
Age	0.419	0.258	1.624	0.116

10. What are the null and alternative hypotheses for determining if at least one of the three predictor variables are significant in predicting cholesterol level?

a. $H_0: b_1 = b_2 = b_3 = 0$ vs. H_1: At least one $b_i \neq 0$
b. $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ vs. H_1: At least one $\beta_i \neq 0$
c. $H_0: b_1 = b_2 = b_3 = 0$ vs. $H_1: b_1 \neq 0, b_2 \neq 0, \text{and } b_3 \neq 0$
d. $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ vs. $H_1: \beta_1 \neq 0, \beta_2 \neq 0, \text{and } \beta_3 \neq 0$

11. Using the appropriate output and the hypotheses from Problem 10, what conclusion can you come to regarding the three predictor variables?

a. None of the predictors are statistically significant in predicting cholesterol level.
b. Exactly one of the predictors is statistically significant in predicting cholesterol level.
c. At least one of the predictors is statistically significant in predicting cholesterol level.
d. All three predictors are statistically significant in predicting cholesterol level.

12. Which of the following describes the relationship of the predictor variables with cholesterol level?

a. Exercise and weight have a nonsignificant negative relationship; age has a significant positive relationship
b. Exercise and weight have significant negative relationships; age has a significant positive relationship
c. Exercise and weight have significant negative relationships; age has a nonsignificant positive relationship

d. Exercise and weight have a nonsignificant negative relationship; age has a nonsignificant positive relationship

13. Which of the following is the correct interpretation for the coefficient for the predictor variable “Exercise”?

    a. For every additional hour of exercise each week, there is a 3.11 point decrease in cholesterol level.
    b. For every additional hour of exercise each week, holding weight and age constant, there is a 3.11 point decrease in cholesterol level.
    c. For every additional hour of exercise each week, there is a 3.11 point decrease in predicted cholesterol level.
    d. For every additional hour of exercise each week, holding weight and age constant, there is a 3.11 point decrease in predicted cholesterol level.

Appendix E

Exams

Exam 1

1. Identify the variable situation reflected in this newspaper headline: "Study finds that as low temperatures increase, people tend to get fewer hours of sleep."

   a. One categorical
   b. One quantitative
   c. Two categorical
   d. Two quantitative
   e. One categorical and one quantitative

Use the following scenario for Questions 2 and 3.

In June 2017, Yale University conducted a poll of 1800 people who voted for Trump in the 2016 election and found that 846 of them were in favor of the United States staying in the Paris Agreement.

2. What was the population of interest in the poll?

   a. All registered voters in the United States
   b. All voters in the 2016 election
   c. All respondents in the survey who wanted the United States to stay in the Paris Agreement
   d. All respondents to the survey
   e. All voters in the 2016 election who voted for Trump

3. Which of the following best describes how we should denote the statistic and parameter from the poll?

   a. Statistic: $\bar{x}$ unknown; Parameter: $\mu = 846$
   b. Statistic: μ unknown; Parameter: $\bar{x} = 846$
   c. Statistic: p unknown; Parameter: $\hat{p} = .47$
   d. Statistic: $\hat{p}$ unknown; Parameter $p = .47$
   e. Statistic: $\bar{x} = 846$; Parameter: μ unknown
   f. Statistic: $\mu = 846$; Parameter: $\bar{x}$ unknown
   g. Statistic: $p = .47$; Parameter: $\hat{p}$ unknown
   h. Statistic: $\hat{p} = .47$; Parameter p unknown

4. Suppose that Pitt wanted to investigate students' opinions on the new two-step authentication required to log in to My Pitt. They randomly select 400 students to take part in a survey, making sure to select 100 each from the freshman, sophomore, junior, and senior classes. What type of sampling method was used to generate the sample?

a. Simple random sample
b. Stratified random sample
c. Systematic sample
d. Convenience sample
e. Voluntary sample

5. A high school is interested in identifying if their students' math skills improve between standardized tests taken in eighth grade and eleventh grade. They collect the percentile that each student's score fell into nationally after each of the two tests. Suppose the district originally had 240 eighth graders, but 20 of them moved away before taking the test in eleventh grade. What type of error occurred in this scenario?

   a. Sampling error
   b. Nonresponse bias
   c. Data acquisition error
   d. Selection bias

6. A survey question about driving habits asked respondents, "When was the last time you were pulled over for speeding by a police officer while you were driving?" Respondents could choose from the following answers:

   - Within the last month
   - More than one month ago, but less than one year ago
   - More than one year ago

   Why is this survey question biased?

   a. Complicated question
   b. Vague concept
   c. Leading question
   d. Central tendency bias
   e. Error prone response options

Use the following scenario for Questions 7 and 8.

A researcher believes that consuming caffeine before taking a test will decrease the amount of time it takes the subject to complete the test. Subjects enter the testing room and randomly sit down at a desk. Each desk contains a number, a pill (either the caffeine or the placebo), and a 100-question test involving basic math skills. The researcher sits at a desk in the back of the room where the subjects cannot see him, instructs the subjects to ingest the pill, and then begin the test. He remains there throughout the entire duration of the test. Subjects turn their test in to the researcher when they have completed the test and immediately leave the room. The researcher then records the amount of time the subject took to complete the test.

7. What type of experimental design was used in this scenario?

a. Matched pairs design
b. Completely randomized design
c. Block design

8. Suppose one subject initially consents to participate in the study, shows up for the test, but at the last minute decides not to participate because the pill is too large to swallow. The researcher tells the subject that because they agreed to participate, they must follow through with the experiment. What ethical standard was violated?

   a. Autonomy
   b. Beneficence
   c. Confidentiality
   d. Deception
   e. Informed consent

Use the following scenario for Questions 9 and 10.

A sleep psychologist is interested in understanding if different types of background noise help patients with insomnia falls asleep faster. She recruits 30 patients with insomnia: 15 men and 15 women. Each is assigned to one of three treatments: white noise, classical music, or no noise. The amount of time between laying down and falling asleep is the response. The sleep psychologist also believes that different sounds may help men and women differently.

9. Which of the following describes the most appropriate way to assign subjects to a treatment?

   a. Assign all 30 subjects to listen to all three sounds.
   b. Assign 10 people to listen to white noise, 10 to classical music, and 10 to no noise, ignoring gender.
   c. Assign 5 men to listen to white noise, 5 to listen to classical music, and 5 to no noise. Also assign 5 women to listen to white noise, 5 to listen to classical music, and 5 to no noise.
   d. Assign the men to listen to white noise and assign the women to listen to classical music. Then have all 30 subjects fall asleep to no noise.

10. What is the control group in this study?

   a. Subjects who listen to white noise
   b. Subjects who listen to classical music
   c. Subjects who listen to no noise
   d. Subjects who fell asleep faster than normal
   e. Subjects who took longer to fall asleep than normal
   f. There is no control

11. Three professors who teach the same course are interested in their students' opinions of the textbook that the department has chosen for the class. They create a survey, asking students

to rate the textbook on a scale from 1 (very poor textbook) to 10 (excellent textbook), but each uses a different method to collect data.

- A: Give a survey to every other person entering the room; 40 observations were obtained
- B: Put the survey online and offer the opportunity for all students to respond; 40 observations were obtained
- C: Choose every fourth person from an alphabetical list of students; 20 observations were obtained

Which of the following statements is **not** true?

a. A will likely resemble the population the closest.
b. A and B will likely produce very similar results because the sample sizes are the same.
c. B will likely provide a biased estimate of the average textbook rating.
d. Regardless of the sample size, A and C are both better sampling methods than B.
e. C provides the most variability in its estimate because the sample size is the smallest.

12. A group of doctors believes that exercise during the winter months may help prevent people from getting a cold. They ask a random sample of patients to record each time they exercise as well as if and when they caught a cold between December 1 and February 28. Upon returning to the clinic in March, doctors plan to ask patients if they developed a cold in the past three months. Which of the following is true about this study?

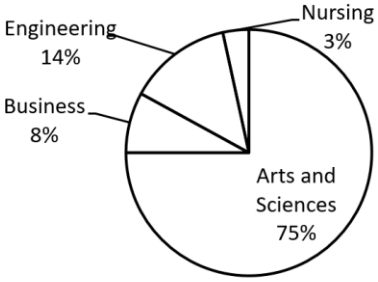
    a. This is a retrospective observational study where the number of days of exercise is the explanatory variable and whether or not the patient developed a cold is the response.
    b. This is a prospective observational study where the number of days of exercise is the explanatory variable and whether or not the patient developed a cold is the response.
    c. This is an experiment where the number of days of exercise is the explanatory variable and whether or not the patient developed a cold is the response.
    d. This is a retrospective observational study where the number of days of exercise is the response and whether or not the patient developed a cold is the explanatory variable.
    e. This is a prospective observational study where the number of days of exercise is the response and whether or not the patient developed a cold is the explanatory variable.
    f. This is an experiment where the number of days of exercise is the response and whether or not the patient developed a cold is the explanatory variable.

13. The following frequency table displays the percentage of college students who have used various illicit drugs at some point during their college careers. What is the best type of graph to use to display this data?

Drug	Percentage
Marijuana	56%
Cocaine	12%
LSD	8%
Ecstasy	13%
Heroin	2%

a. Pie chart
b. Bar graph
c. Mosaic plot
d. Histogram
e. Scatterplot

14. The following pie chart shows the percentage of freshmen who enrolled in each of the four entry level schools at Pitt last year. There were 3,300 students who enrolled in the School of Arts and Sciences. How many students enrolled in the Business School?



a. 264
b. 352
c. 550
d. 2,475
e. 4,400

15. The following cross-classification table compares the age of a random sample of American women in 2016 with their marital status. What proportion of women between the ages of 35 and 49 were married?

Age	Single	Married	Divorced	Widowed	Total
18-34	1,273	1,305	602	20	3,200
35-49	813	1,140	946	101	3,000
50-64	532	965	709	194	2,400
65+	147	590	334	329	1,400
Total	2,765	4,000	2,591	644	10,000

a. .114
b. .285
c. .300
d. .380
e. .400

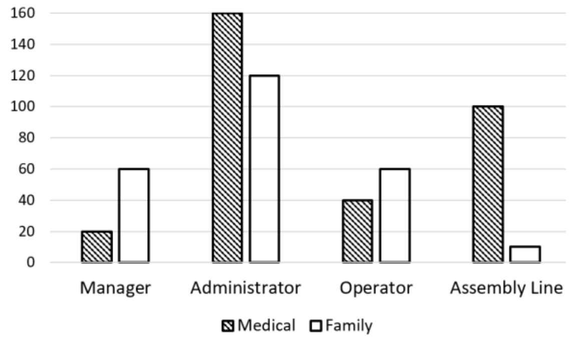
16. Random samples of Democrats and Republicans were asked to rate how often they trust Congress to do what is right. The results are presented in the relative frequency tables below. Which of the following statements describes the medians of these two groups?

Democrats	Percentage	Republicans	Percentage

Nearly all the time	4%	Nearly all the time	16%
Most of the time	12%	Most of the time	30%
Some of the time	44%	Some of the time	36%
Almost never	40%	Almost never	18%

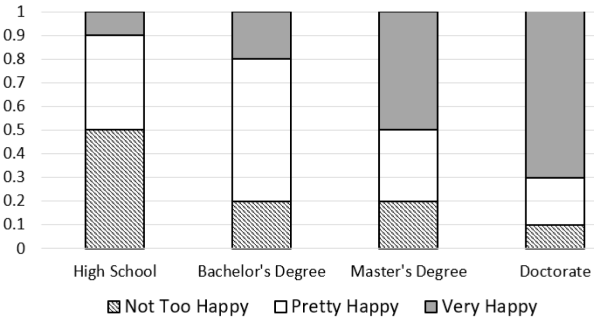
a. Democrats’ median response gives more confidence to Congress than Republicans’ median response.
b. Republicans’ median response gives more confidence to Congress than Democrats’ median response.
c. The median response from Democrats and Republicans gives the same amount of confidence to Congress.
d. The medians cannot be computed, and thus no comparison can be made.

17. The following double bar graph summarizes the reasons employees at a manufacturing plant have given to miss work. Responses are separated based on if the employee is a manager, administrator, operator, or assembly line worker. For which group is the conditional proportion of absences due to family emergencies the highest?



a. Manager
b. Administrator
c. Operator
d. Assembly Line

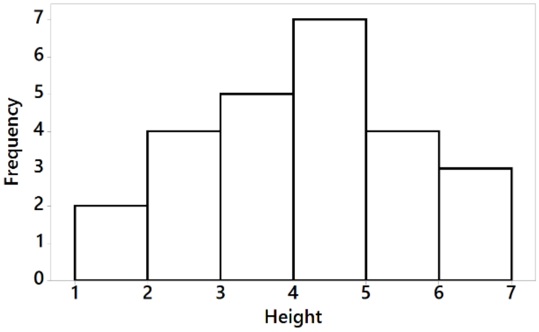
18. The following stacked bar graph compares a person’s highest level of education with their reported happiness level. Values on the vertical axis report the proportion of respondents in each happiness category. Based on the information provided in the graph, which of the following statements are true?



I. More people with Master's degrees than with Bachelor's degrees describe themselves as very happy.
II. Lower education levels tend to be associated with higher levels of happiness.
III. The proportion of people who are pretty happy does not appear to change depending on the highest level of education attained.

a. None are true
b. I only
c. II only
d. III only
e. I and II
f. I and III
g. II and III
h. I, II, and III

19. A 7th grade science class is growing plants using hydroponics (a method of growing plants without soil). Each of the 25 students has their own plant. At the end of the year, each plant is measured in inches. Based on the histogram below, what proportion of plants are shorter than 4 inches tall?



a. .11
b. .20
c. .36
d. .44
e. .64

20. Suppose that 95% of adult males consume between 1900 and 3100 calories per day and that these values follow a normal distribution. What are the mean and standard deviation of the number of calories consumed by adult males each day?

a. Mean is 2500 and standard deviation is 600
b. Mean is 2500 and standard deviation is 300
c. Mean is 2500 and standard deviation is 200
d. Mean is 5000 and standard deviation is 600
e. Mean is 5000 and standard deviation is 300
f. Mean is 5000 and standard deviation is 200

21. Heights of American males follow a normal distribution with a mean of 69 inches and a standard deviation of 3 inches. According to the Empirical Rule, which of the following is not a conclusion that we can make?

    a. An adult male who is 67 inches tall would be considered common.
    b. About 2.5% of adult males are shorter than 63 inches tall.
    c. An adult male who is 66 inches tall has a negative Z-score.
    d. An adult male who is 75 inches tall is more standard deviations from the mean than an adult male who is 64 inches tall.

Use the following scenario for Questions 22, 23, and 24.

The following table shows the ages of the five oldest and five youngest British Prime Ministers when they left office. The mean age is 60.73 with a standard deviation of 10.79 years. The five-number summary is 34, 54, 61, 69, 84.

Youngest	Age	Oldest	Age
Duke of Grafton	34	Marquess of Salisbury	72
Duke of Devonshire	37	John Russell	73
Earl of Shelburne	45	Benjamin Disraeli	75
Viscount Goderich	45	Winston Churchill	80
Henry Addington	46	William Gladstone	84

22. Which Prime Ministers are considered outliers?

    a. None are outliers
    b. Duke of Grafton only
    c. Duke of Grafton and Duke of Devonshire
    d. William Gladstone only
    e. Winston Churchill and William Gladstone
    f. Duke of Grafton and William Gladstone
    g. Duke of Grafton, Duke of Devonshire, Winston Churchill, and William Gladstone

23. When Margaret Thatcher left office, her age was in the 63rd percentile. Which of the following statements are true?

    I. She was between 61 and 69 when she left office.
    II. She was older than only 37% of all other Prime Ministers when she left office.
    III. The observation is contained in the upper half of the box in a boxplot.

    a. None are true
    b. I only
    c. II only
    d. III only
    e. I and II
    f. I and III

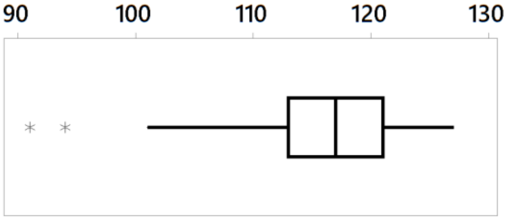
g. II and III
h. I, II, and III

24. Suppose the Duke of Grafton's age was incorrectly recorded as '23' instead of '34'. Which of the following statistics would **decrease** as a result of this error?

I. Mean
II. Median
III. Standard Deviation

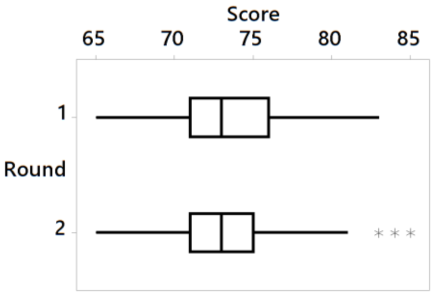
a. None
b. I only
c. II only
d. III only
e. I and II
f. I and III
g. II and III
h. I, II, and III

25. The following boxplot displays the high temperatures in Death Valley during the summer of 2016. What would a histogram of the data look like based on the boxplot?



a. Symmetric with Mean ≈ Median
b. Left-skewed with Mean > Median
c. Left-skewed with Mean < Median
d. Right-skewed with Mean > Median
e. Right-skewed with Mean < Median

26. The following boxplots compare the scores of golfers in the first and second rounds of the 2017 U.S. Open. Which of the following are true statements about the boxplots?



I. Neither round of golf produced any outliers.
II. The range of scores is larger for Round 1 than Round 2.
III. The IQR of scores is larger for Round 1 than Round 2.

a. None are true
b. I only
c. II only
d. III only
e. I and II
f. I and III
g. II and III
h. I, II, and III

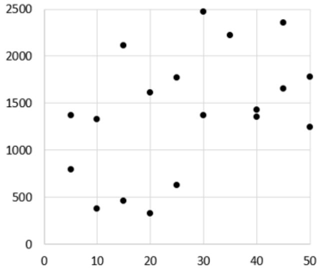
27. It is commonly quoted that defense wins championships in professional sports. We want to use the number of goals allowed for each team in the Premier League during the 2016-17 season to predict the number of points they finished the season with. An analysis of the data reveals the following summary statistics. What type of linear relationship exists between goal differential and points earned?

Covariance	$s_{GOALS\ ALLOWD}$	$s_{POINTS\ EARNED}$
-248.48	14.94	19.85

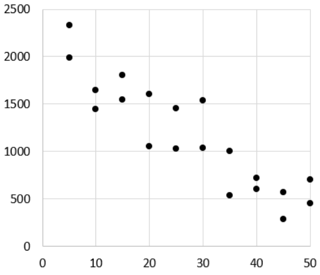
a. Strong negative linear relationship
b. Moderate negative linear relationship
c. Weak negative linear relationship
d. No linear relationship
e. Weak positive linear relationship
f. Moderate positive linear relationship
g. Strong positive linear relationship

28. The following scatterplots have predictor values between 0 and 50 and responses between 0 and 2500. Rank the following scatterplots in order from **weakest** correlation to **strongest** correlation.

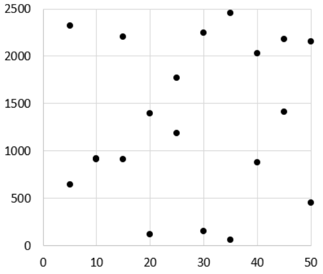
A



B



C



a. A, B, C
b. A, C, B
c. B, A, C

d. B, C, A
e. C, A, B
f. C, B, A

Use the following scenario for Questions 29 and 30.

A doctor wants to compare the heights and weights of his patients. In a random sample of 40 patients, he uses his patients' height as the predictor and their weight as the response. The correlation was found to be $r = 0.25$. The slope of the regression line was found to be 5.6 with a y-intercept of -202.

29. Which of the following statements are true?

I. There is a weak, positive linear relationship between the height and weight.
II. A person who is 72 inches tall and weighs 200 pounds would be close to the regression line if plotted on a scatterplot.
III. If the doctor drew a regression line on a scatterplot to fit the data, then the slope would be positive.

a. None are true
b. I only
c. II only
d. III only
e. I and II
f. I and III
g. II and III
h. I, II, and III

30. What would be the predicted weight of a person who is 68 inches tall?

a. 95.2
b. 134
c. 178.8
d. 270
e. 380.8

<u>Exam 2</u>

1. Suppose three fair dice are rolled. Which of the following events are intersections?

I. Rolling a sum of either 7 or 11
II. Rolling a sum of less than 18
III. Having the individual rolls consist of one 4, one 5, and one 6

a. None
b. I only

c. II only
d. III only
e. I and II
f. I and III
g. II and III
h. I, II, and III

Use the following scenario for Questions 2, 3, and 4.

The following table of probabilities compares what part of the United States a person lives in (East, Central, West) with their preferred vacation location.

	Beach	Camping	Visiting a City	Total
East	.25	.02	.17	.44
Central	.13	.08	.10	.31
West	.12	.10	.03	.25
Total	.50	.20	.30	1.00

2. Given that a person lives in the West, what is the probability that they prefer to camp for vacation?

   a. .10
   b. .20
   c. .25
   d. .35
   e. .40
   f. .45
   g. .50

3. What is the probability that a person does not live in the East and also does not prefer visiting a city for vacation?

   a. .132
   b. .170
   c. .200
   d. .392
   e. .400
   f. .430
   g. .830
   h. 1.260

4. What is the probability that a person lives in the Central part of the country or prefers to go to the beach for vacation?

   a. .1300
   b. .2600

c. .4194
d. .5000
e. .6800
f. .8100

5. An urn contains 5 white marbles and 10 black marbles. You select two marbles at random from the urn consecutively without replacing the first marble. Let A be the event that the first marble is white and B be the event the second marble is white. Which of the following best describes events A and B?

   a. Independent, but not mutually exclusive
   b. Mutually exclusive, but not independent
   c. Both independent and mutually exclusive
   d. Neither independent nor mutually exclusive

6. A committee made up of two people is to be formed. Members will be randomly chosen without replacement from a group of 8 people that includes 5 women and 3 men. What is the probability that both members of the committee are women?

   a. .3125
   b. .3571
   c. .3906
   d. .4464
   e. .2500
   f. .5000
   g. .9143

7. The number of phone lines that are in use by the customer service center at a software manufacturer depends on if it is a weekday or weekend. The following probability distributions display the probability that a certain number of phone lines are busy at noon on weekends and on weekdays.

Weekdays

Busy Phone Lines	0	1	2	3	4	5
Probability	.30	.25	.15	.10	.15	.05

Weekends

Busy Phone Lines	0	1	2	3	4	5
Probability	.45	.20	.15	.10	.05	.05

What is the probability that more than 3 phone lines are in use on both a Sunday and Monday at noon?

   a. .01
   b. .02
   c. .06
   d. .20
   e. .30
   f. .50

g. .56
h. .60
i. .72

8. The probability distributions of two spinners used in a board game are displayed below:

Spin (A)	1	2	3	4
Probability	.50	.25	.15	.10

Spin (B)	3	4	5	6
Probability	.50	.25	.15	.10

Which of the following statements accurately describes the mean and standard deviations of the distributions?

a. The mean of spinner A is larger and the standard deviation of spinner A is larger
b. The mean of spinner A is larger and the standard deviation of spinner B is larger
c. The mean of spinner A is larger and the standard deviations of both spinners are the same
d. The mean of spinner B is larger and the standard deviation of spinner A is larger
e. The mean of spinner B is larger and the standard deviation of spinner B is larger
f. The mean of spinner B is larger and the standard deviations of both spinners are the same
g. The means of both spinners are the same and the standard deviation of spinner A is larger
h. The means of both spinners are the same and the standard deviation of spinner B is larger
i. The means and standard deviations of both spinners are the same

9. After a natural disaster, prices for a case of water tend to increase by approximately 25%. If the mean price for a case of water is typically \$5.00 with a population standard deviation of \$0.40, what would be the distribution after adjusting the prices due to a natural disaster?

a. $\mu = 5.25$ and $\sigma = 0.40$
b. $\mu = 6.25$ and $\sigma = 0.40$
c. $\mu = 5.25$ and $\sigma = 0.50$
d. $\mu = 6.25$ and $\sigma = 0.50$
e. $\mu = 5.25$ and $\sigma = 0.65$
f. $\mu = 6.25$ and $\sigma = 0.65$

10. In Allegheny County, 34% of adults have at least a Bachelor's degree. Suppose 6 people are independently sampled from Allegheny County. What is the probability that at least one person has a Bachelor's degree?

a. .0799
b. .0827
c. .2555
d. .3382
e. .3600
f. .7446
g. .9173

11. In Liechtenstein, 56% of all babies born are male. During a single year, 196 of 375 babies born were male. Which of the following describes how unusual it would be to see 196 male births out of 375 in Liechtenstein?

   a. 196 is extremely low compared to the number of male births we would expect
   b. 196 is somewhat low compared to the number of male births we would expect
   c. 196 is very close to the number of male births we would expect
   d. 196 is somewhat high compared to the number of male births we would expect
   e. 196 is extremely high compared to the number of male births we would expect

12. Three runners participate in races of three different lengths. Runner A competed in a 5K, runner B completed a 10K, and runner C ran a marathon. Each runner's results are presented below along with the population mean and population standard deviation of completion times for all other runners in the race. Assume all three distributions are normal and that times are given in minutes.

   - Runner A: Completed in 22 minutes; $\mu = 25$ and $\sigma = 3$
   - Runner B: Completed in 46 minutes; $\mu = 52$ and $\sigma = 7$
   - Runner C: Completed in 270 minutes; $\mu = 300$ and $\sigma = 20$

   Order the runners from best to worst finishes relative to the rest of the runners they competed against in their respective races. (Reminder: Lower times are better when running in a race.)

   a. A, B, C
   b. A, C, B
   c. B, A, C
   d. B, C, A
   e. C, A, B
   f. C, B, A

Use the following scenario for Questions 13, 14, and 15.

The amount of time that high school students spend on homework nightly differs by grade level, but the distributions tend to be normally distributed. Assume all times given are in minutes.

- Freshmen: $\mu = 75$ and $\sigma = 16$
- Sophomores: $\mu = 100$ and $\sigma = 20$
- Juniors: $\mu = 150$ and $\sigma = 40$
- Seniors: $\mu = 120$ and $\sigma = 30$

13. What is the probability that a randomly selected freshman will spend less between 63 and 83 minutes on homework one night?

   a. .2266
   b. .4013
   c. .4649

d. .6914
e. .8944
f. .9181

14. Suppose that for a randomly selected sophomore, 34.46% of sophomores spend more time on homework than the senior selected. How much time does this student spend on homework each night?

a. 65.54
b. 92.00
c. 93.11
d. 106.89
e. 108.00
f. 134.46

15. Suppose a junior spends 142 minutes on homework each night. Assuming they continue with the same study habits, how much time would we expect them to spend on homework as a senior?

a. 107.38
b. 112.00
c. 114.00
d. 126.00
e. 128.00
f. 132.62

16. The number of miles driven when a person rents a car follows a normal distribution with a mean of 512 miles and a standard deviation of 128 miles. Suppose we take a random sample of 16 people who recently returned a rental car. What is the sampling distribution of the sample mean number of miles driven for these 16 people?

a. Normally distributed with mean $\mu_{\bar{X}} = 512$ miles and standard error $\sigma_{\bar{X}} = 8$ miles
b. Normally distributed with mean $\mu_{\bar{X}} = 512$ miles and standard error $\sigma_{\bar{X}} = 32$ miles
c. Normally distributed with mean $\mu_{\bar{X}} = 32$ miles and standard error $\sigma_{\bar{X}} = 8$ miles
d. Normally distributed with mean $\mu_{\bar{X}} = 32$ miles and standard error $\sigma_{\bar{X}} = 32$ miles
e. Undetermined shape with mean $\mu_{\bar{X}} = 512$ miles and standard error $\sigma_{\bar{X}} = 8$ miles
f. Undetermined shape with mean $\mu_{\bar{X}} = 512$ miles and standard error $\sigma_{\bar{X}} = 32$ miles
g. Undetermined shape with mean $\mu_{\bar{X}} = 32$ miles and standard error $\sigma_{\bar{X}} = 8$ miles
h. Undetermined shape with mean $\mu_{\bar{X}} = 32$ miles and standard error $\sigma_{\bar{X}} = 32$ miles

17. The average credit an American has on their credit cards is $4000 with a population standard deviation of $1500. Suppose a random sample of 36 Americans is taken and the amount of credit they have is calculated. What is the probability that the sample mean credit of these 36 Americans exceeds $4375?

a. .1057

b. .2563
c. .4012
d. .5988
e. .7338
f. .8944

18. Let X be the random variable denoting the amount of money a student in Pennsylvania pays for tuition each year. Suppose X follows a normal distribution with a mean of $12,000 and a standard deviation of $3,000. Suppose we randomly sample 4 students and average their yearly tuition. The probability that this sample mean exceeds $13,500 is .1587; that is, $P(\bar{X} > 13{,}500) = .1587$. Which of the following changes would result in this probability decreasing?

I. Decrease the population mean tuition to $10,000
II. Increase the population standard deviation to $4,000
III. Increase the number of students sampled to 16

a. None would decrease the probability
b. I only
c. II only
d. III only
e. I and II
f. I and III
g. II and III
h. I, II, and III

19. The proportion of adults in this country who are eligible to vote is .92. Suppose you take a random sample of 50 adults and ask if they are eligible to vote. What is the sampling distribution of the sample proportion of adults who are eligible to vote in this sample?

a. Normally distributed with mean $\mu_{\hat{p}} = 46$ and standard error $\sigma_{\hat{p}} = 1.92$
b. Normally distributed with mean $\mu_{\hat{p}} = .92$ and standard error $\sigma_{\hat{p}} = .038$
c. Normally distributed with mean $\mu_{\hat{p}} = .92$ and standard error $\sigma_{\hat{p}} = .130$
d. Undetermined shape with mean $\mu_{\hat{p}} = 46$ and standard error $\sigma_{\hat{p}} = 1.92$
e. Undetermined shape with mean $\mu_{\hat{p}} = .92$ and standard error $\sigma_{\hat{p}} = .038$
f. Undetermined shape with mean $\mu_{\hat{p}} = .92$ and standard error $\sigma_{\hat{p}} = .130$

20. In a large city, 38% of people are in favor of raising taxes to pay for a new stadium. Suppose a random sample of 40 people is selected. What is the probability that less than half of the people sampled support raising taxes to fund the new stadium? (Use four decimal places in the calculation of the standard error.

a. .0572
b. .0655
c. .1934
d. .8117

e. .9345
f. .9418

Exam 3

Use the following scenario for Questions 1-5.

In 2012, President Obama's administration issued a requirement that all new cars average greater than 36 miles per gallon (MPG) by the end of 2016. To determine how close automobile manufacturers are coming to the new regulation, the EPA takes a random sample of 16 cars and measures the number of miles per gallon each gets. The results found a sample mean of 38.05 MPG. Assume the population standard deviation is equal to 5 MPG and that the number of miles per gallon is normally distributed.

1. Which of the following represents a 95% confidence interval for the mean?

   a. $38.05 \pm 1.645\left(\frac{5}{\sqrt{16}}\right) = (35.994, 40.106)$
   b. $36 \pm 1.645\left(\frac{5}{\sqrt{16}}\right) = (33.944, 38.056)$
   c. $38.05 \pm 1.96\left(\frac{5}{\sqrt{16}}\right) = (35.67, 40.57)$
   d. $36 \pm 1.96\left(\frac{5}{\sqrt{16}}\right) = (33.55, 38.45)$
   e. $38.05 \pm 2.576\left(\frac{5}{\sqrt{16}}\right) = (34.83, 41.27)$
   f. $36 \pm 2.576\left(\frac{5}{\sqrt{16}}\right) = (32.78, 39.22)$

2. Based on the confidence interval, does it appear as if President Obama's requirement has been met? Why?

   a. Yes: 95% confidence interval contains 36
   b. Yes: 95% confidence interval is entirely above 36
   c. Yes: 95% confidence interval is entirely below 36
   d. No: 95% confidence interval contains 36
   e. No: 95% confidence interval is entirely above 36
   f. No: 95% confidence interval is entirely below 36

3. Which of the following would have led to a narrower confidence interval?

   I. Sample mean of 35 MPG
   II. Population standard deviation of 6 MPG
   III. Using a 99% level of confidence

   a. None would have led to a narrower interval
   b. I only
   c. II only
   d. III only

e. I and II
f. I and III
g. II and III
h. I, II, and III

4. Suppose the true population mean miles per gallon for new cars is actually 37. If 80 random samples of cars had been taken and a 95% confidence interval calculated for each, how many of these confidence intervals would we expect to contain 37?

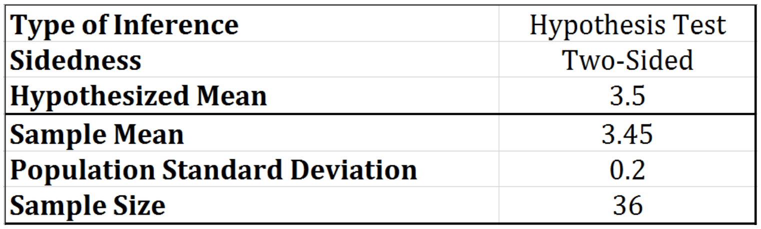
a. 0
b. 4
c. 40
d. 76
e. 80

5. Suppose we wanted to obtain an interval that was 2 MPG wide, but to be more certain we capture the population mean, we want to use 99% confidence instead. How large a sample would need to be taken to attain this width?

a. 7
b. 13
c. 42
d. 166

Use the following scenario for Questions 6-10.

Ivy League schools claim that the average GPA of its graduates is 3.50. A simple random sample of recent graduates yielded the following results. Use a 1% level of significance.



Type of Inference	Hypothesis Test
Sidedness	Two-Sided
Hypothesized Mean	3.5
Sample Mean	3.45
Population Standard Deviation	0.2
Sample Size	36

6. What is/are the critical value(s) of this test?

a. $Z = 1.28$
b. $Z = \pm 1.645$
c. $Z = 1.645$
d. $Z = \pm 1.96$
e. $Z = 2.326$
f. $Z = \pm 2.576$
g. $Z = 1.50$
h. $Z = -1.50$

i. $Z = \pm 1.50$

7. What is the p-value of this test?

   a. .0334
   b. .0668
   c. .1336
   d. .8664
   e. .9332
   f. .9666

8. Based on the results of this test, what conclusion can we come to regarding the average GPA of Ivy League graduates?

   a. Average GPA of Ivy League graduates is significantly less than 3.50
   b. Average GPA of Ivy League graduates is significantly different from 3.50 with an indication it may be lower
   c. Average GPA of Ivy League graduates is significantly different from 3.50 with an indication it may be higher
   d. Average GPA of Ivy League graduates is equal to 3.50
   e. Average GPA of Ivy League graduates is not significantly different from 3.50

9. The result of a Type II error in this situation would be concluding that the true population mean GPA of Ivy League graduates is:

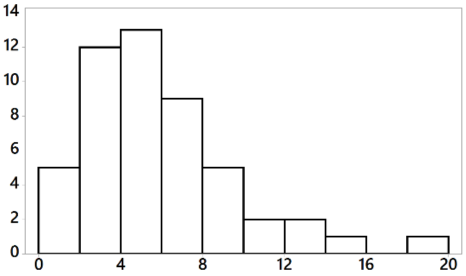
   a. Equal to 3.50 when it is really lower than 3.50.
   b. Equal to 3.50 when it is really some value other than 3.50.
   c. Lower than 3.50 when it is really equal to 3.50.
   d. Equal to some value other than 3.50 when it is really equal to 3.50.
   e. Equal to 3.50 when it is really equal to 3.45.

10. Which of the following describes what we could conclude about a 99% confidence interval for the mean GPA of Ivy League graduates?

   a. It would contain both 3.45 and 3.50
   b. It would contain 3.45, but not 3.50
   c. It would contain 3.50, but not 3.45
   d. It would contain neither 3.45 nor 3.50
   e. It would contain 3.45, but we cannot determine if it would contain 3.50
   f. It would not contain 3.45, but we cannot determine if it would contain 3.50
   g. It would contain 3.50, but we cannot determine if it would contain 3.45
   h. It would not contain 3.50, but we cannot determine if it would contain 3.45

Use the following scenario for Questions 11-16.

The following output contains the results of a survey given to 50 skiers regarding the number of times they visit a ski resort throughout the year. The resort wants to make sure it is profitable and believes that the break-even point is to have its customers visit an average of 5 times per year.



Type of Inference	Hypothesis Test
Sidedness	
Hypothesized Mean	5
Sample Mean	5.75
Sample Standard Deviation	3.74
Sample Size	50
Degrees of Freedom	49
Test Statistic	
P-Value	0.0813

11. What are the null and alternative hypotheses?

a. $H_0: \mu = 5$ vs. $H_1: \mu > 5$
b. $H_0: \mu = 5$ vs. $H_1: \mu < 5$
c. $H_0: \mu = 5$ vs. $H_1: \mu \neq 5$
d. $H_0: \mu > 5$ vs. $H_1: \mu = 5$
e. $H_0: \mu < 5$ vs. $H_1: \mu = 5$
f. $H_0: \mu \neq 5$ vs. $H_1: \mu = 5$

12. What can we determine about the conditions needed to perform this test?

a. Satisfied because the histogram is approximately normal
b. Satisfied because the sample size is large enough
c. Not satisfied because the original population is not known to be normal
d. Not satisfied because the histogram is severely skewed
e. Not satisfied because the histogram is severely skewed and the original population is not known to be normal

13. What are the distribution and value of the test statistic?

a. $Z = \frac{5.75-5}{3.74/\sqrt{50}} = 1.418$
b. $Z = \frac{5-5.75}{3.74/\sqrt{50}} = -1.418$
c. $t = \frac{5.75-5}{3.74/\sqrt{50}} = 1.418$
d. $t = \frac{5-5.75}{3.74/\sqrt{50}} = -1.418$
e. $t = \frac{5.75-5}{3.74/\sqrt{49}} = 1.404$
f. $t = \frac{5-5.75}{3.74/\sqrt{49}} = -1.404$

14. Which of the following would have led to a smaller p-value?

    I. Hypothesized mean of 5.50
    II. Sample size of 100
    III. Sample standard deviation of 4

    a. None would have led to a smaller p-value
    b. I only
    c. II only
    d. III only
    e. I and II
    f. I and III
    g. II and III
    h. I, II, and III

15. For which levels of significance would the null hypothesis be rejected?

    a. Do not reject the null hypothesis for $\alpha = .01$, $\alpha = .05$, or $\alpha = .10$
    b. Reject for $\alpha = .01$ and $\alpha = .05$, but not for $\alpha = .10$
    c. Reject for $\alpha = .10$, but not for $\alpha = .01$ or $\alpha = .05$
    d. Reject for $\alpha = .01$, $\alpha = .05$, and $\alpha = .10$

16. Suppose instead that we had known the population standard deviation was 3.74. Which of the following would have been different?

    I. Hypotheses
    II. Value of the test statistic
    III. P-value

    a. None would have been different
    b. I only
    c. II only
    d. III only
    e. I and II
    f. I and III
    g. II and III
    h. I, II, and III

Use the following scenario for Questions 17-20.

A simple random sample of 800 voting age Americans were contacted regarding their concern about climate change. Suppose 600 responded that they were worried about climate change. Based on the 800 responses, a 90% confidence interval for the proportion of people who reported they are worried about climate change is $(.725, .775)$.

17. Which of the following is the correct interpretation of the confidence interval?

    a. 90% of all confidence intervals calculated regarding the proportion of people who are worried about climate change will contain the sample proportion of .75.
    b. The probability that 90% of people are worried about climate change is between .725 and .775
    c. We are 90% confident that the true proportion of people who are worried about climate change is between 72.5% and 77.5%.
    d. We are 90% confident that between 72.5% and 77.5% of the next 800 people surveyed will be worried about climate change.

18. If the sample proportion had been .65 instead of .75 as observed in the original poll while still using a sample size of 800 and a 90% confidence interval, then the confidence interval would have been:

    a. Wider
    b. Narrower
    c. The same width

19. Which of the following statements are true based on the confidence interval?

    I. The proportion of people who are worried about climate change appears to be significantly less than .80.
    II. A 99% confidence interval would also contain .73.
    III. .77 is a plausible estimate for the true proportion of people who are worried about climate change.

    a. None are true
    b. I only
    c. II only
    d. III only
    e. I and II
    f. I and III
    g. II and III
    h. I, II, and III

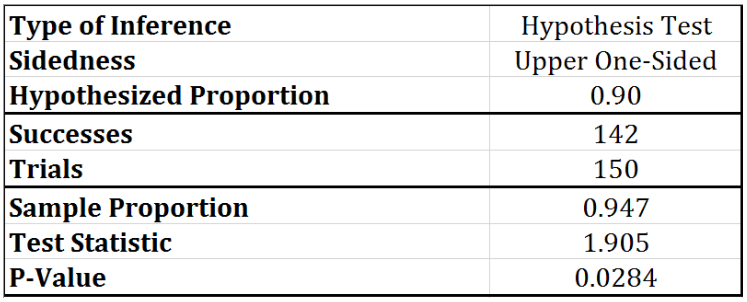
20. Suppose a simple random sample of 1600 Democrats, who have made protecting the environment a primary concern in their platform, are surveyed in a separate poll and asked the same question regarding climate change. A 90% confidence interval is calculated for these responses as well. Which of the following statements is **not** true?

    a. The interval calculated from the sample of Democrats will be narrower than the interval calculated by taking the sample from the population of all voting age Americans.
    b. The interval calculated from the sample of Democrats is a closer approximation of the proportion of Americans who are worried about climate change because the sample size is larger.

c. The sample proportion of Democrats who reported they are concerned about climate change is likely an overestimate of the proportion of all Americans who are worried about climate change.
d. The intervals are estimating different parameters so no conclusions can be made about the proportion of all Americans who are worried about climate change based on the interval that consisted of only Democrats.

Use the following scenario for Questions 21-26.

The Department of Transportation is interested in increasing the proportion of flights that arrive at their destinations on time. A random sample of 150 flights were followed one day and classified as being "on-time" or "late". Consider a flight that was on-time as a success. Use the output below to perform the test at the 5% level of significance.



Type of Inference	Hypothesis Test
Sidedness	Upper One-Sided
Hypothesized Proportion	0.90
Successes	142
Trials	150
Sample Proportion	0.947
Test Statistic	1.905
P-Value	0.0284

21. What are the null and alternative hypotheses?

a. $H_0: \mu = .90$ vs. $H_1: \mu > .90$
b. $H_0: \bar{x} = .90$ vs. $H_1: \bar{x} > .90$
c. $H_0: p = .90$ vs. $H_1: p > .90$
d. $H_0: \hat{p} = .90$ vs. $H_1: \hat{p} > .90$

22. What is the correct interpretation of the p-value?

a. The probability that more than 90% of flights are on-time is .0284
b. If the true proportion of flights that are on-time is actually .90, then the probability of obtaining a sample proportion of exactly .947 is .0284.
c. The probability that 90% is the true percentage of flights that are on-time is .0284.
d. If the true proportion of flights that are on-time is actually .90, then the probability of obtaining a sample proportion greater than or equal to .947 is .0284.

23. What decision and conclusion can be made based on the results of this test?

a. Reject H_0 and conclude that more than 90% of flights are on-time
b. Fail to reject H_0 and conclude that more than 90% of flights are on-time
c. Reject H_0 and conclude that the percentage of on-time flights is not greater than 90%
d. Fail to reject H_0 and conclude that the percentage of on-time flights is not greater than 90%

24. Suppose the true proportion of flights that are on time is actually .90. What type of decision was made based on the results of the hypothesis test?

a. Type I error
b. Type II error
c. Correct decision

25. If a two-sided test had been used instead, what would the p-value have been?

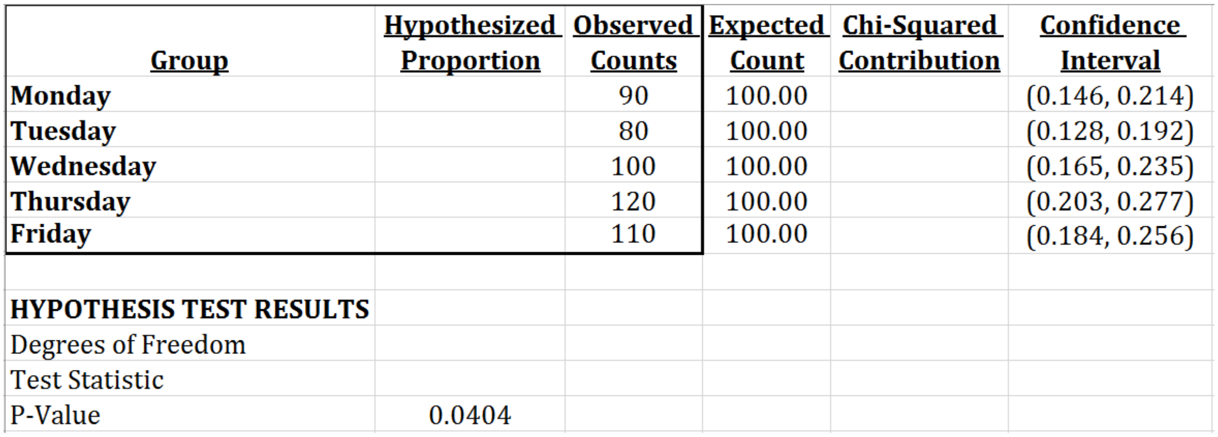
a. .0142
b. .025
c. .0284
d. .0568
e. .10

26. What is the probability of making a Type I error?

a. .0284
b. .05
c. .0568
d. .95
e. .9716

Use the following scenario for Questions 27-30.

A small business that is open Monday through Friday keeps track of the number of customers it has each day for a week. In total, the business had 500 customers during the week. Use the results in the output below to determine if the proportion of customers is the same on all five days that the business is open using a 10% level of significance.



Group	Hypothesized Proportion	Observed Counts	Expected Count	Chi-Squared Contribution	Confidence Interval
Monday		90	100.00		(0.146, 0.214)
Tuesday		80	100.00		(0.128, 0.192)
Wednesday		100	100.00		(0.165, 0.235)
Thursday		120	100.00		(0.203, 0.277)
Friday		110	100.00		(0.184, 0.256)
HYPOTHESIS TEST RESULTS					
Degrees of Freedom					
Test Statistic					
P-Value	0.0404				

27. What are the null and alternative hypotheses?

a. $H_0: p_M = p_{Tu} = p_W = p_{Th} = p_F = .20$

H_1: All proportions differ significantly from $.20$

b. $H_0: \mu_M = \mu_{Tu} = \mu_W = \mu_{Th} = \mu_F = 100$
   H_1: All population means differ significantly from 100
c. $H_0: \hat{p}_M = \hat{p}_{Tu} = \hat{p}_W = \hat{p}_{Th} = \hat{p}_F = .20$
   H_1: All proportions differ significantly from .20
d. $H_0: p_M = p_{Tu} = p_W = p_{Th} = p_F = .20$
   H_1: At least one proportion differs significantly from $.20$
e. $H_0: \mu_M = \mu_{Tu} = \mu_W = \mu_{Th} = \mu_F = 100$
   H_1: At least one population mean differs significantly from 100
f. $H_0: \hat{p}_M = \hat{p}_{Tu} = \hat{p}_W = \hat{p}_{Th} = \hat{p}_F = .20$
   H_1: At least one proportion differs from .20

28. What is the chi-squared contribution from the category ‘Tuesday’?

a. $-.25$
b. $-.20$
c. $.20$
d. $.25$
e. 4
f. 5

29. How many degrees of freedom does this test have?

a. 4
b. 5
c. 499
d. 500

30. What conclusion can we come to based on the results of the test and the confidence intervals?

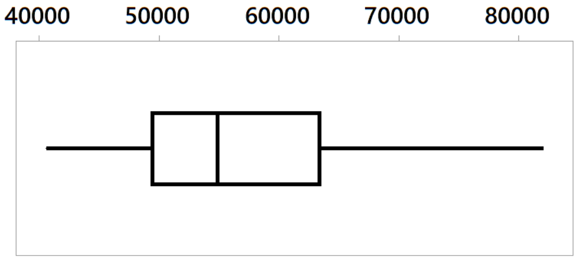
a. The proportion of customers is the same on all five days of the week.
b. The proportion of customers on Tuesday and Thursday differ from their hypothesized proportions.
c. The proportion of customers on Monday, Wednesday, and Friday differ from their hypothesized proportions.
d. The proportion of customers differs from their hypothesized proportions on all five weekdays.

Exam 4

Use the following scenario for Problems 1-4.

The median household income was calculated for each of the 50 states and plotted in the boxplot below.

1. Based on the boxplot, what would the shape of the histogram look like?



   a. Left skewed
   b. Right skewed
   c. Symmetric

2. What can we say about the relationship between the mean and median?

   a. Mean is much greater than the median
   b. Mean is much less than the median
   c. Mean and median are approximately equal

3. Which of the following statements is **not** true about the boxplot?

   a. About 75% of states have median annual household incomes above $50,000.
   b. No states appear to have median annual household incomes that are outliers.
   c. The fourth quartile contains the most observations.
   d. A state with a median annual household income of $60,000 is in the third quartile.

4. Suppose Puerto Rico's median annual household income of $18,626 was included as well. If this observation is added to the data, what would happen to the mean and standard deviation?
   a. Both the mean and standard deviation would decrease
   b. Mean would increase; standard deviation would decrease
   c. Mean would decrease, standard deviation would increase
   d. Both the mean and standard deviation would increase

5. Consider the following terms in descriptive statistics. Which of the following **could** take on negative values?

   I. Mean
   II. Interquartile Range
   III. Correlation

   a. None can be negative
   b. Only I
   c. Only II
   d. Only III
   e. I and II

f. I and III
g. II and III
h. I, II, and III

6. Consider the following terms in probability. Which of the following **could** take on negative values?

I. Z-score
II. Probability
III. Standard Deviation

a. None can be negative
b. Only I
c. Only II
d. Only III
e. I and II
f. I and III
g. II and III
h. I, II, and III

7. Consider the following terms in inference. Which of the following **could** take on negative values?

I. P-value
II. Standard Error
III. Slope Coefficient

a. None can be negative
b. Only I
c. Only II
d. Only III
e. I and II
f. I and III
g. II and III
h. I, II, and III

8. Weights for 25-year-old males follow a normal distribution with a mean of 170 pounds and standard deviation $\sigma = 32$ pounds. Suppose that a doctor tells her patient that the Z-score for his weight is 0.75. Which **one** of the following is **not** a correct description of the patient's weight?

a. His weight is 194 pounds.
b. 75% of 25-year-old males weigh less than he does.
c. His weight is 0.75 standard deviations above the average weight for 25-year-old males.
d. According to the Empirical Rule, his weight is not considered to be unusual for someone his age.

9. On the final day of the NBA season, the 76ers played the Knicks and the Celtics played the Bucks. Which **one** of the following best describes the relationship between the events “76ers win” and “Celtics win”?

   a. Mutually exclusive
   b. Independent
   c. Both mutually exclusive and independent
   d. Neither mutually exclusive nor independent

10. Two random samples are taken independently of one another and 90% confidence intervals are calculated for each. If 10 is actually the true population mean, what is the probability that neither interval contains 10?

   a. .01
   b. .10
   c. .81
   d. .90
   e. .99

11. Suppose you are interested in discovering the true proportion of children under the age of 10 who go trick-or-treating on Halloween. You want to test the hypotheses $H_0\colon p = .80$ vs. $H_1\colon p \neq .80$. You take 50 independent samples from the population and perform a hypothesis test on each sample. Assuming that the true population proportion is actually $p = .80$, how many tests would you expect to reject using $\alpha = .10$?

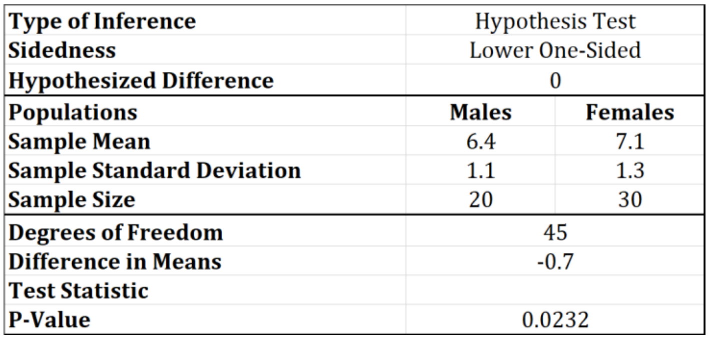
   a. 0
   b. 5
   c. 10
   d. 25
   e. 40
   f. 45
   g. 50

12. In the Pennsylvania Lottery’s Pick 3 game, 3 numbers between 0 and 9 are chosen. One lottery player notices that the number 7 tends to be selected quite often. He takes a random sample of lottery drawings and tests the hypotheses $H_0\colon p = 10$ vs. $H_1\colon p > .10$. He concludes that the true proportion of the time that 7 is selected is significantly greater than .10. However, in reality, the true proportion of the time that 7 is chosen is actually .10. What type of decision did he make?

   a. Type I error
   b. Type II error
   c. Correct decision

Use the following scenario for Problems 13-18.

Suppose we are interested in testing if the average amount of sleep for college students who commute is less than the average amount of sleep that students who live on-campus get. Assume both populations are normally distributed and values are presented in hours. Use the output to answer the following questions at the 5% level of significance.



Type of Inference	Hypothesis Test	
Sidedness	Lower One-Sided	
Hypothesized Difference	0	
Populations	**Males**	**Females**
Sample Mean	6.4	7.1
Sample Standard Deviation	1.1	1.3
Sample Size	20	30
Degrees of Freedom	45	
Difference in Means	-0.7	
Test Statistic		
P-Value	0.0232	

13. What are the hypotheses?

    a. $H_0: \mu_C - \mu_{OC} = 0$ vs. $H_1: \mu_C - \mu_{OC} < 0$
    b. $H_0: \mu_C - \mu_{OC} \neq 0$ vs. $H_1: \mu_C - \mu_{OC} = 0$
    c. $H_0: \mu_C - \mu_{OC} = 0$ vs. $H_1: \mu_C - \mu_{OC} > 0$
    d. $H_0: \mu_C - \mu_{OC} = 0$ vs. $H_1: \mu_C - \mu_{OC} \neq 0$

14. What is the test statistic?

    a. $t = \frac{6.4-7.1}{{}^{1.1^2+1.3^2}/_{\sqrt{50}}} = -1.707$
    b. $t = \frac{6.4-7.1}{{}^{2.4}/_{\sqrt{50}}} = -2.062$
    c. $t = \frac{6.4-7.1}{\sqrt{\frac{1.1^2}{20}+\frac{1.3^2}{30}}} = -2.048$
    d. $t = \frac{6.4-7.1}{\sqrt{\frac{1.1}{20}+\frac{1.3}{30}}} = -2.23$

15. What conclusion can you come to about the mean amount of sleep that commuters get compared to students who lie on campus based on the result of the above test?

    a. Mean amount of sleep commuters get is significantly less than the mean amount of sleep that on-campus students get.
    b. Mean amount of sleep commuters get is significantly greater than the mean amount of sleep that on-campus students get.
    c. Mean amount of sleep commuters get is less than the mean amount of sleep that on-campus students get, but the difference is not significant.
    d. Mean amount of sleep commuters get is greater than the mean amount of sleep that on-campus students get, but the difference is not significant.

16. Which of the following describes what we could conclude about a 90% confidence interval for the difference between the means?

    a. It would contain both 0 and -0.7
    b. It would contain 0, but not -0.7
    c. It would contain -0.7, but not 0
    d. It would contain neither 0 nor -0.7

17. Which of the following would cause the p-value to decrease?

    I. Obtaining sample means of 6 hours from the commuters and 8 from the on-campus students
    II. Obtaining sample standard deviations of 1 hour for both commuters and on-campus students
    III. Obtaining the same means and standard deviations from samples of size 50 and 60 for the commuters and on-campus students respectively

    a. None would decrease the p-value
    b. I only
    c. II only
    d. III only
    e. I and II
    f. I and III
    g. II and III
    h. I, II, and III

18. Which of the following would have changed if we had decided to use the on-campus students as population 1 and the commuters as population 2?

    I. Test statistic
    II. P-value
    III. Decision

    a. None would change
    b. I only
    c. II only
    d. III only
    e. I and II
    f. I and III
    g. II and III
    h. I, II, and III

Use the following scenario for Problems 19-21.

A movie fanatic is interested in studying if the average lengths of movies differ depending on their rating (PG, PG-13, and R). She visits IMDB.com and selects a random sample of 10 movies from each of the three ratings. Use the following output to determine if there is a significant difference in the average length of PG, PG-13, and R rated movies.

19. What type of sampling method was used to collect the data?

a. Simple random sample
b. Stratified random sample
c. Systematic sample
d. Voluntary sample

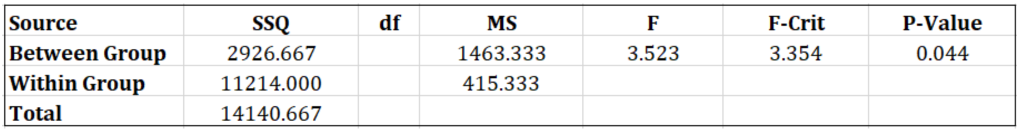
20. Which graphical display would best display the relationship between movie rating and length of movie?

a. Bar graph
b. Histogram
c. Side-by-side boxplots
d. Scatterplot

21. What are the numerator and denominator degrees of freedom?

a. Numerator: 2, Denominator: 8
b. Numerator: 2, Denominator: 27
c. Numerator: 3, Denominator: 8
d. Numerator: 3, Denominator: 27

22. Using only the ANOVA table, what conclusion can we make regarding the mean movie lengths across the groups?



Source	SSQ	df	MS	F	F-Crit	P-Value
Between Group	2926.667		1463.333	3.523	3.354	0.044
Within Group	11214.000		415.333			
Total	14140.667					

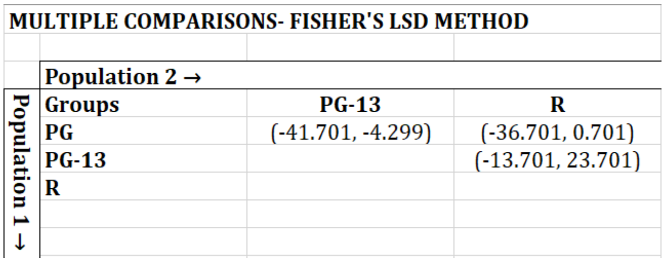
a. None of the mean movie lengths are significantly different
b. Exactly two of the mean movie lengths are significantly different
c. At least two of the mean movie lengths are significantly different
d. All three mean movie lengths are significantly different from one another

23. Which of the following would cause the p-value to decrease?

I. Decreasing the sample mean of each group by 5 minutes
II. Decreasing the sample standard deviation of each group by 5 minutes
III. Doubling the sample size of each group

a. None would cause the p-value to decrease
b. Only I
c. Only II

d. Only III
e. I and II
f. I and III
g. II and III
h. I, II, and III



MULTIPLE COMPARISONS- FISHER'S LSD METHOD

Population 2 →

Population 1 →

Groups	**PG-13**	**R**
PG	(-41.701, -4.299)	(-36.701, 0.701)
PG-13		(-13.701, 23.701)
R		

24. The above output contains 95% confidence intervals for the difference between each pair of group means using Fisher's LSD method. Use the confidence intervals to determine which pairs of means are significantly different.

I. PG vs. PG-13
II. PG vs. R
III. PG-13 vs. R

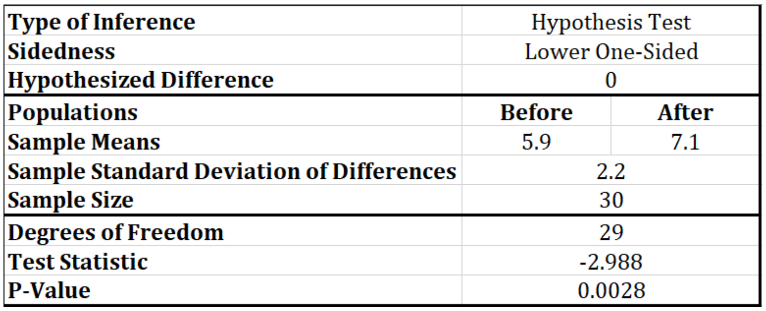
a. No significant difference between any pairs
b. Only I
c. Only II
d. Only III
e. I and II
f. I and III
g. II and III
h. I, II, and III

25. Why are the widths of all three multiple comparisons confidence intervals the same?

a. Number of intervals calculated equals the number of different groups being compared
b. Test statistic is greater than the critical value
c. Sample size taken from each group is the same
d. Original populations are all normally distributed

Use the following scenario for Problems 26-28.

Police trainees were being trained on how to correctly identify license plate numbers. They were shown 10 license plates for five seconds each. At the end of the session, the trainees were instructed to write down as many of the license plates as they could remember. They then took a week-long memory training course. At the end of the course, the process of being shown the same 10 license plates and writing down as many as they could remember was repeated. Use the output below to determine if there is a significant difference in the number of license plates trainees could recall before and after the memory training course using a 5% level of significance.



Type of Inference	Hypothesis Test	
Sidedness	Lower One-Sided	
Hypothesized Difference	0	
Populations	**Before**	**After**
Sample Means	5.9	7.1
Sample Standard Deviation of Differences	2.2	
Sample Size	30	
Degrees of Freedom	29	
Test Statistic	-2.988	
P-Value	0.0028	

26. Which of the following is true?

    a. Samples are independent: Plates correct is quantitative and taking test before or after training is categorical
    b. Samples are independent: Plates correct is categorical and taking test before or after training is quantitative
    c. Samples are dependent: Plates correct is quantitative and taking test before or after training is categorical
    d. Samples are dependent: Plates correct is categorical and taking test before or after training is quantitative

27. Which of the following is **not** a correct way to write the hypotheses?

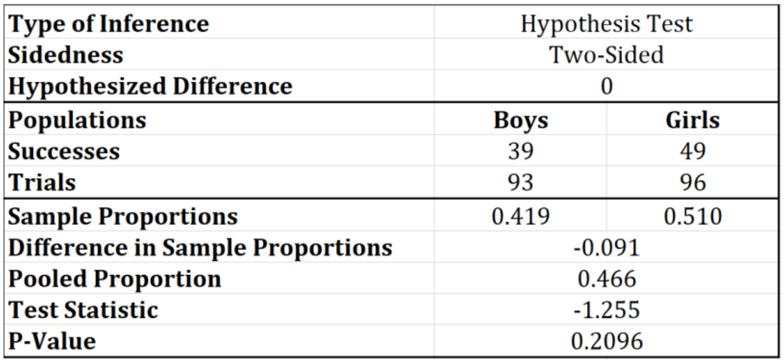
    a. $H_0: \mu_B - \mu_A = 0$ vs. $H_1: \mu_B - \mu_A < 0$
    b. $H_0: \mu_B = \mu_A = 0$ vs. $H_1: \mu_B \neq 0$ or $\mu_A \neq 0$
    c. $H_0: \mu_D = 0$ vs. $H_1: \mu_D < 0$
    d. $H_0: \mu_B = \mu_A$ vs. $H_1: \mu_B < \mu_A$

28. What conclusion can you draw based on the result of the hypothesis test?

    a. Average number of license plates identified correctly was significantly lower after the memory training.
    b. Average number of license plates identified correctly was lower after the memory training, but the difference is not significant.
    c. Average number of license plates identified correctly was significantly higher after the memory training.
    d. Average number of license plates identified correctly was higher after the memory training, but the difference is not significant.

Use the following scenario for Problems 29-33.

A pediatrician is interested in if the proportion of children who are able to walk by their first birthday differs by gender. Random samples of patients who come in for their one-year checkup are taken at several area doctor's offices. The results of the study are presented in the output below. A success in this study should be considered a child who was able to walk independently on by their first birthday. Use a 5% level of significance.



Type of Inference	Hypothesis Test	
Sidedness	Two-Sided	
Hypothesized Difference	0	
Populations	**Boys**	**Girls**
Successes	39	49
Trials	93	96
Sample Proportions	0.419	0.510
Difference in Sample Proportions	-0.091	
Pooled Proportion	0.466	
Test Statistic	-1.255	
P-Value	0.2096	

29. Which of the following would **not** be an appropriate way to visualize or display the data for a difference of two proportions test?

    a. Mosaic plot
    b. Histogram
    c. Double bar graph
    d. Cross classification table

30. What conclusion can we come to using a 5% level of significance?

    a. Proportion of boys who can walk on their first birthday is significantly different than the proportion of girls who can walk on their first birthday with an indication the proportion of girls is higher.
    b. Proportion of boys who can walk on their first birthday is significantly different than the proportion of girls who can walk on their first birthday with an indication the proportion of boys is higher.
    c. Proportion of boys who can walk on their first birthday is higher than the proportion of girls, but the difference is not significant.
    d. Proportion of girls who can walk on their first birthday is higher than the proportion of boys, but the difference is not significant.

31. What would the p-value have been if a one-sided alternative had been used instead of a two-sided alternative?

    a. .1048
    b. .2096
    c. .4192

32. Suppose we calculated a 95% confidence interval for the difference of two proportions by taking the proportion of boys who can walk on their first birthday and subtracting the proportion of girls who can walk on their first birthday. Which of the following are true about the confidence interval?

I. The interval would contain −.091
II. Zero is a plausible value for the difference between the two proportions
III. The entire interval would be negative.

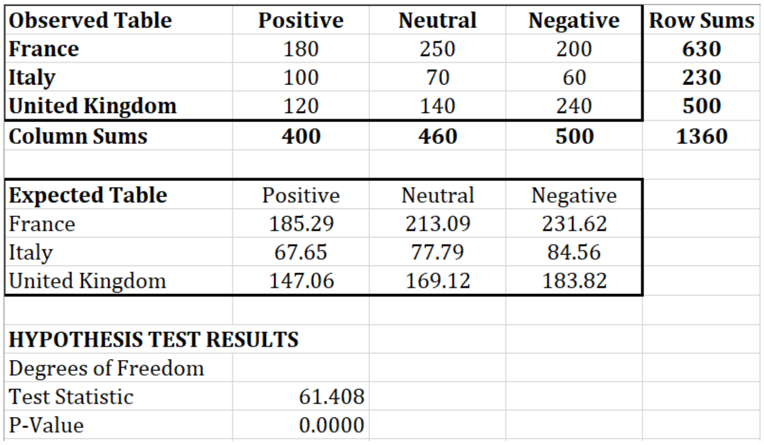
a. None are true
b. Only I
c. Only II
d. Only III
e. I and II
f. I and III
g. II and III
h. I, II, and III

33. The above output shows the results for a difference of two proportions test. What other equivalent inferential technique could we have used to analyze this data?

a. Goodness of fit test
b. Matched pairs test
c. Difference of two means test
d. $2x2$ test for independence

Use the following scenario for Problems 34-37.

Residents of three European countries (France, Italy, and the United Kingdom) were asked if they viewed the European Union positively, negatively, or if they were neutral. The results are presented in the output below. Use a 5% level of significance to make any conclusions.



Observed Table	**Positive**	**Neutral**	**Negative**	**Row Sums**
France	180	250	200	**630**
Italy	100	70	60	**230**
United Kingdom	120	140	240	**500**
Column Sums	**400**	**460**	**500**	**1360**

Expected Table	Positive	Neutral	Negative
France	185.29	213.09	231.62
Italy	67.65	77.79	84.56
United Kingdom	147.06	169.12	183.82

HYPOTHESIS TEST RESULTS	
Degrees of Freedom	
Test Statistic	61.408
P-Value	0.0000

34. What are the null and alternative hypotheses?

a. H_0: Residents of all three countries have a positive view of the European Union
H_1: Residents of at least one country do not have a positive view of the European Union
b. H_0: Country and opinion of the European Union are not independent
H_1: Country and opinion of the European Union are independent

c. H_0: Residents of all three countries have a positive view of the European Union
   H_1: Residents of all three countries do not have a positive view of the European Union
d. H_0: Country and opinion of the European Union are independent
   H_1: Country and opinion of the European Union are not independent

35. How many degrees of freedom does this test have?

a. 4
b. 5
c. 8
d. 9
e. 1351
f. 1359

36. What is the chi-squared contribution to the test statistic from the group of respondents from France who responded that they have a neutral opinion of the European Union?

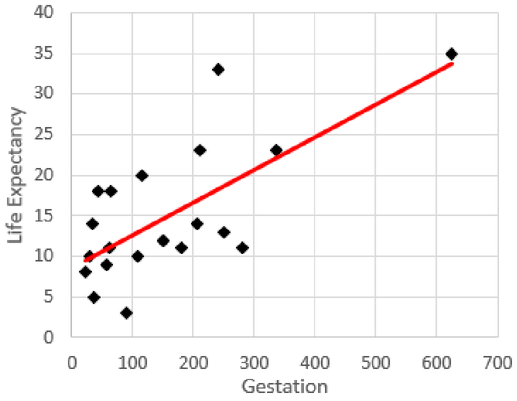
a. $-.173$
b. $-.148$
c. .148
d. .173
e. 5.449
f. 6.393

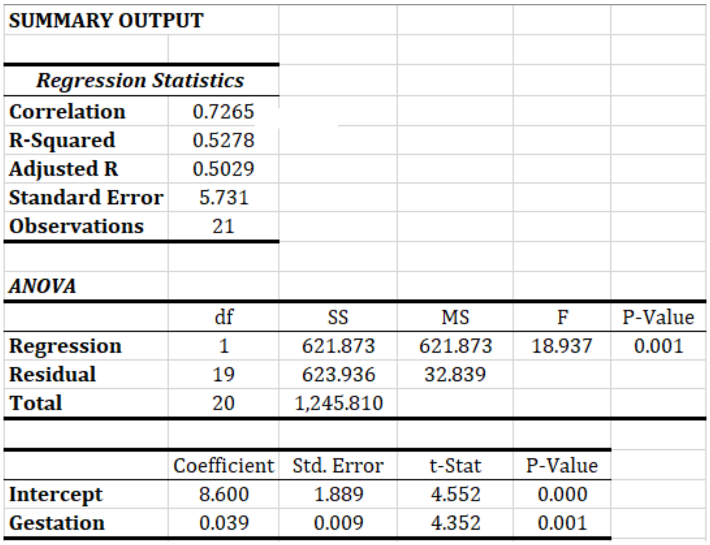
37. What conclusion can we come to based on the results of the test?

a. Country and opinion of the European Union are independent.
b. Country and opinion of the European Union are not independent: Residents of Italy have the most favorable opinions while residents of the United Kingdom have the least favorable opinions
c. Country and opinion of the European Union are not independent: Residents of France have the most favorable opinions while residents of Italy have the least favorable opinions
d. Country and opinion of the European Union are not independent: Residents of France have the most favorable opinions while residents of the United Kingdom have the least favorable opinions
e. Country and opinion of the European Union are not independent: Residents of Italy have the most favorable opinions while residents of France have the least favorable opinions

Use the following scenario for Problems 38-46.

The following output and plots display the results of using the average gestation period (in days) to predict the average life expectancy (in years) of 21 randomly selected animals. The line of the scatterplot is the least squares regression line. Use a 5% level of significance to make your decision on any hypothesis tests.





SUMMARY OUTPUT

Regression Statistics	
Correlation	0.7265
R-Squared	0.5278
Adjusted R	0.5029
Standard Error	5.731
Observations	21

ANOVA

	df	SS	MS	F	P-Value
Regression	1	621.873	621.873	18.937	0.001
Residual	19	623.936	32.839		
Total	20	1,245.810			

	Coefficient	Std. Error	t-Stat	P-Value
Intercept	8.600	1.889	4.552	0.000
Gestation	0.039	0.009	4.352	0.001

38. One animal has a gestation period of 240 days and a life expectancy of 33 years. Which of the following best describes the characteristics of this point?

a. Only an outlier
b. Only an influential point
c. Both an outlier and an influential point
d. Neither an outlier nor an influential point

39. Calculate the correlation and use it to describe the strength of the linear relationship.

a. Strong positive linear relationship
b. Moderate positive linear relationship
c. Weak positive linear relationship
d. No linear relationship
e. Weak negative linear relationship
f. Moderate negative linear relationship
g. Strong negative linear relationship

40. What is the linear model?

a. $y = \beta_0 + \beta_1 x$
b. $y = \beta_0 + \beta_1 x + \varepsilon$
c. $\hat{y} = \beta_0 + \beta_1 x$
d. $\hat{y} = \beta_0 + \beta_1 x + \varepsilon$
e. $y = 8.6 + .039x$
f. $y = 8.6 + .039x + \varepsilon$
g. $\hat{y} = 8.6 + .039x$
h. $\hat{y} = 8.6 + .039x + \varepsilon$

41. What is the residual for an animal that has a gestation period of 400 days and actually has a life expectancy of 20 years?

a. -380
b. 380
c. -4.2
d. 4.2
e. 24.2
f. -24.2

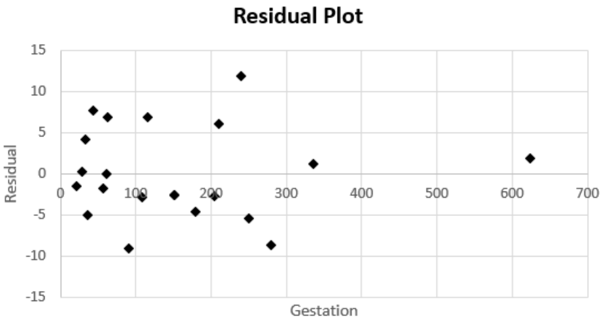
42. What are the null and alternative hypotheses for determining if gestation period is a significant predictor of life expectancy?

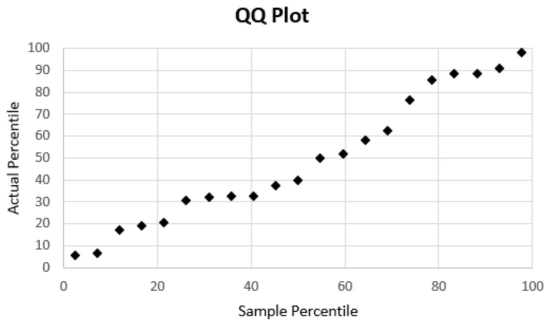
a. $H_0: \beta_0 = 0$ vs. $H_1: \beta_0 \neq 0$
b. $H_0: b_0 = 0$ vs. $H_1: b_0 \neq 0$
c. $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$
d. $H_0: b_1 = 0$ vs. $H_1: b_1 \neq 0$

43. Based on the output what can we deduce about the relationship between gestation period and life expectancy?

a. Gestation period is a significant predictor of life expectancy; relationship is positive.
b. Gestation period is a significant predictor of life expectancy; relationship is negative.
c. A positive linear relationship exists between gestation period and life expectancy, but is not statistically significant
d. A negative linear relationship exists between gestation period and life expectancy, but is not statistically significant

44. The residual plot (left) and QQ-plot (right) are displayed below. Which of the following best describes any violations of requirements of the error term?





a. Homoscedasticity is violated
b. Normality is violated
c. Both homoscedasticity and normality are violated
d. Neither homoscedasticity nor normality is violated

45. A 95% prediction interval for life expectancy at a gestation period of 200 days is (4,17, 16.48). Which of the following is the correct interpretation of this interval?

a. We are 95% confident that the average life expectancy for all animals is between 4.17 and 16.48.
b. We are 95% confident that the average life expectancy for a single animal is between 4.17 and 16.48.
c. We are 95% confident that the average life expectancy for a single animal with an average gestation period of 200 days is between 4.17 and 16.48.
d. We are 95% confident that the average life expectancy for all animals with an average gestation period of 200 days is between 4.17 and 16.48.

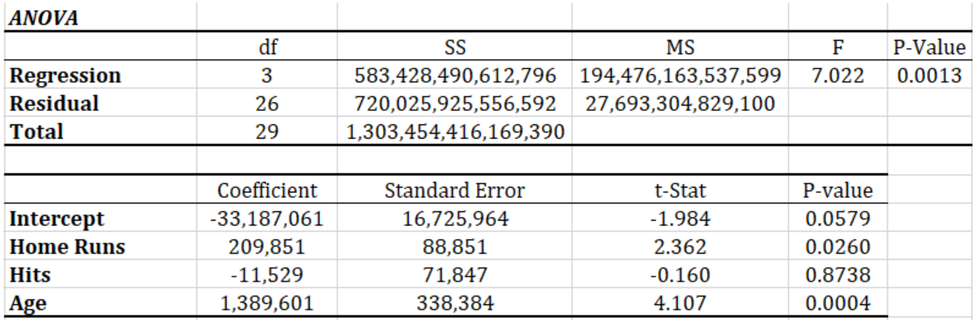
46. On average, the gestation period of these animals was 140 days. A 95% prediction interval for life expectancy at a gestation period of 200 days is (4,17, 16.48). Which of the following intervals would be wider than the one provided?

I. A 90% prediction interval for life expectancy with a gestation period of 200 days
II. A 95% prediction interval for life expectancy with a gestation period of 150 days
III. A 95% confidence interval for life expectancy with a gestation period of 200 days

a. None would be wider
b. I only
c. II only
d. III only
e. I and II
f. I and III
g. II and III
h. I, II, and III

Use the following scenario for Problems 47-50.

A general manager in baseball is interested in the relationship between a player's salary and a number of his statistics from the previous season. He uses the number of home runs (X_1), the number of hits (X_2), and the player's age (X_3) to predict his salary. Use the multiple linear regression output below to answer the following questions. Use a 5% level of significance to make a decision on any hypothesis tests.



ANOVA					
	df	SS	MS	F	P-Value
Regression	3	583,428,490,612,796	194,476,163,537,599	7.022	0.0013
Residual	26	720,025,925,556,592	27,693,304,829,100		
Total	29	1,303,454,416,169,390			

	Coefficient	Standard Error	t-Stat	P-value
Intercept	-33,187,061	16,725,964	-1.984	0.0579
Home Runs	209,851	88,851	2.362	0.0260
Hits	-11,529	71,847	-0.160	0.8738
Age	1,389,601	338,384	4.107	0.0004

47. What are the null and alternative hypotheses for determining if at least one of the three predictor variables are significant in predicting salary?

a. $H_0: b_1 = b_2 = b_3 = 0$ vs. $H_1: b_1 \neq 0, b_2 \neq 0,$ and $b_3 \neq 0$
b. $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ vs. $H_1: \beta_1 \neq 0, \beta_2 \neq 0,$ and $\beta_3 \neq 0$
c. $H_0: b_1 = b_2 = b_3 = 0$ vs. H_1: At least one $b_i \neq 0$
d. $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ vs. H_1: At least one $\beta_i \neq 0$

48. Using the appropriate output and the hypotheses from Problem 47, what conclusion can you come to regarding the three predictor variables?

a. Exactly one of the predictors is statistically significant in predicting salary.
b. All three predictors are statistically significant in predicting salary.
c. None of the predictors are statistically significant in predicting salary.
d. At least one of the predictors is statistically significant in predicting salary.

49. Which of the following describes the relationship of the predictor variables with salary?

a. Home runs and age have a nonsignificant positive relationship; hits have a nonsignificant negative relationship
b. Home runs and age have a nonsignificant positive relationship; hits have a significant negative relationship
c. Home runs and age have significant positive relationships; hits have a nonsignificant negative relationship
d. Home runs and age have significant positive relationships; hits have a significant negative relationship

50. Which of the following is the correct interpretation for the coefficient for the predictor variable "Home Runs"?

a. For every additional home run hit, there is a $209,851 increase in predicted salary.
b. For every additional home run hit, there is a $209,851 increase in salary.
c. For every additional home run hit, holding hits and age constant, there is a $209,851 increase in predicted salary.
d. For every additional home run hit, holding hits and age constant, there is a $209,851 increase in salary.

Match each scenario with the most appropriate hypothesis test. Write the letter corresponding to the hypothesis test on the line next to the scenario. Each hypothesis test will be used exactly once.

a. Z-test
b. t-test
c. One Sample Proportion Test
d. Goodness of Fit Test
e. Matched Pairs
f. Difference of Two Means Test
g. ANOVA

h. Difference of Two Proportions Test
i. Test for Independence
j. Linear Regression

51. An airline wants to test the claim that the no-show rate for its passengers is less than 4%.
52. A hospital wants to test if the average weight of newborn males differs significantly from the average weight of newborn females.
53. A marketing company conducts a survey asking if people are in favor of raising the tax on cigarettes and wants to determine if smokers and non-smokers differ with respect to their opinions on raising the tax.
54. UPS wants to determine if the average weight of a package delivered by their drivers is greater than 10 pounds. The population standard deviation of the weights of the packages is unknown.
55. A teachers union claims that the average annual salary of teachers in their first five years of teaching is less than $35,000. The population standard deviation of teachers' salaries is known to be $10,000.
56. A police department wants to know if the average speed driven by teenagers with their learner's permit is greater than their average speed after taking a driver's safety class.
57. A software company wants to determine if gender and preferred type of computer are related.
58. A doctor wants to determine if there is a significant relationship between the heights and weights of his patients.
59. An economist wants to compare the average incomes for people who did not complete high school, only have high school diplomas, or have college degrees to see if they are all equal.
60. We want to test if the proportion of M&M's colors (red, blue, green, yellow, orange, brown) are all equal.

Appendix F

Survey

Name: ____

1. Have you ever taken a previous statistics class (either in high school or college)?

   Yes No

2. On a scale from 1 to 6 with 1 being completely unexcited and 6 being extremely excited, how excited are you to be taking a statistics course this semester?

   1 2 3 4 5 6

3. On a scale from 1 to 6 with 1 being completely unexcited and 6 being extremely excited, how exited are you to be using a student response system in class this semester?

   1 2 3 4 5 6

Appendix G

Summaries of Survey Question Responses

Table G.1

Summary of Course and Student Response System Excitement

			Distribution of responses (Percentage)					
Variable	Mean	*SD*	1	2	3	4	5	6
Course excitement	4.03	0.89	1.38	3.45	17.93	46.90	28.97	1.38
Clicker excitement	4.20	0.85	0.69	2.07	11.72	53.79	25.52	6.21

Appendix H

Summary of Complete and Incomplete Sequences by Unit and Level of Cognitive Demand

Table H.1

Number of Complete and Possible Sequences by Unit

Unit	Complete sequences	Possible sequences	Percent complete
Descriptive statistics	3,961	4,350	91.06
Probability	2,530	2,900	87.24
One-sample inference	3,588	4,350	82.48
Two-sample inference	4,043	4,640	87.13

Table H.2

Number of Complete and Possible Sequences by Level of Cognitive Demand

Cognitive demand	Complete sequences	Possible sequences	Percent complete
Recall	3,988	4,495	88.72
Basic application of skill	6,273	7,250	86.52
Strategic thinking	3,861	4,495	85.90

Table H.3

Number of Complete and Possible Sequences by Unit and Level of Cognitive Demand

Unit	Cognitive demand	Complete sequences	Possible sequences	Percent complete
Descriptive statistics	Recall	1,597	1,740	91.78
	Basic application of skill	1,447	1,595	90.72
	Strategic thinking	917	1,015	90.34
Probability	Recall	527	580	90.86
	Basic application of skill	1,267	1,450	87.38
	Strategic thinking	736	870	84.60
One-sample inference	Recall	843	1,015	83.05
	Basic application of skill	1,538	1,885	81.59
	Strategic thinking	1,207	1,450	83.24
Two-sample inference	Recall	1,021	1,160	88.02
	Basic application of skill	2,021	2,330	87.11
	Strategic thinking	1,001	1,160	86.29

Table H.4

Reason for Incomplete Sequences by Unit

Unit	SRS missing (%)	Quiz missing (%)	Both missing (%)	Total
Descriptive statistics	318 (81.75)	67 (17.22)	4 (1.03)	389
Probability	306 (82.70)	45 (12.16)	19 (5.14)	370
One-sample inference	688 (90.29)	55 (7.22)	19 (2.49)	762
Two-sample inference	535 (91.61)	35 (5.99)	14 (2.40)	584
Total	**1,847 (89.17)**	**202 (9.60)**	**56 (2.66)**	**2,118**

Appendix I

Summary of Exam Reponses by Unit and Level of Cognitive Demand

Table I.1

Breakdown of Exam Response by Course Unit

	Exam Response Result		
Unit	Correct (%)	Incorrect (%)	Total
Descriptive statistics	3,328 (84.02)	633 (15.98)	3,961
Probability	1,960 (77.47)	570 (22.53)	2,530
One-Sample inference	3,154 (87.90)	434 (12.10)	3,588
Two-Sample inference	3,551 (87.83)	492 (12.17)	4,043
Total	**11,993 (84.92)**	**2,129 (15.08)**	**14,122**

Table I.2

Breakdown of Exam Response by Level of Cognitive Demand

	Exam response result		
Unit	Correct (%)	Incorrect (%)	Total
Recall	3,603 (90.35)	385 (9.65)	3,988
Basic application of skill	5,479 (87.34)	794 (12.66)	6,273
Strategic thinking	2,911 (75.39)	950 (24.61)	3,861
Total	**11,993 (84.92)**	**2,129 (15.08)**	**14,122**

Appendix J

Summary of KR20 Scores by Unit

Table J.1

KR20 Scores for Assessment Items by Unit

Unit	SRS	Quiz	Exam
Data collection and descriptive statistics	0.7961	0.8654	0.8640
Probability	0.7405	0.7904	0.7980
One-sample inference	0.8498	0.9186	0.9357
Two-sample inference	0.8752	0.9080	0.9283

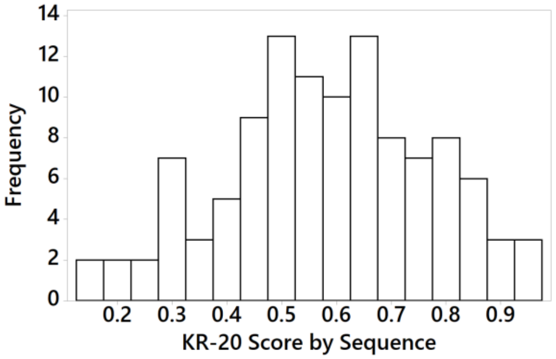


Figure J.1. Histogram of the KR20 scores for the 112 concepts, calculated to test for reliability across the three questions in each sequence

Table J.2

KR20 Score Statistics for Sequences Broken Down by Unit

Unit	Mean	*SD*	Sample Size
Data collection and descriptive statistics	0.5821	0.1826	30
Probability	0.4911	0.1739	20
One-sample inference	0.6183	0.1727	30
Two-sample inference	0.6163	0.1817	32

Appendix K

Summary of Monte Carlo Simulation Results by Student

Table K.1

Summary of Monte Carlo Simulation Results by Student

ID	Complete Sequences	Correct Exam Answers	Predicted Correct Exam Answers	95% Confidence Interval	Result
1	98	76	82.29	(75, 89)	In
2	72	51	57.24	(50, 63)	In
3	54	40	42.96	(37, 48)	In
4	107	77	85.68	(78, 93)	Below
5	66	53	51.95	(45, 58)	In
6	97	79	82.95	(76, 89)	In
7	82	66	70.75	(64, 76)	In
8	108	96	92.02	(85, 99)	In
9	112	107	98.16	(91, 105)	Above
10	106	82	88.10	(81, 95)	In
11	110	100	95.13	(88, 102)	In
12	110	90	93.35	(86, 100)	In
13	94	70	79.54	(73, 86)	Below
14	64	58	56.03	(51, 61)	In
15	104	91	87.64	(80, 94)	In
16	112	95	93.64	(86, 101)	In
17	104	92	88.66	(81, 95)	In
18	98	81	82.46	(75, 89)	In
19	87	82	76.14	(70, 82)	In
20	108	103	91.89	(85, 99)	Above
21	92	85	79.87	(73, 86)	In
22	102	82	88.10	(81, 94)	In
23	106	89	90.70	(84, 97)	In
24	100	93	85.01	(78, 92)	Above
25	64	47	53.21	(47, 59)	In
26	93	71	79.84	(73, 86)	Below
27	89	69	72.90	(66, 80)	In
28	106	103	91.78	(85, 98)	Above
29	106	84	90.27	(83, 97)	In
30	101	87	84.09	(77, 91)	In
31	98	76	81.94	(75, 89)	In
32	112	88	92.76	(85, 100)	In
33	110	91	89.39	(81, 97)	In
34	108	97	94.91	(88, 101)	In
35	99	86	83.48	(76, 90)	In

36	112	100	100.20	(94, 106)	In
37	92	86	79.75	(73, 86)	In
38	112	100	95.88	(88, 102)	In
39	106	104	91.43	(84, 98)	Above
40	100	69	83.75	(77, 91)	Below
41	112	93	94.58	(87, 102)	In
42	93	71	77.53	(70, 84)	In
43	111	91	97.24	(90, 104)	In
44	110	96	91.85	(84, 99)	In
45	112	99	99.15	(92, 105)	In
46	106	89	90.24	(83, 97)	In
47	111	103	94.47	(87, 101)	Above
48	63	46	51.61	(46, 57)	In
49	110	78	94.46	(87, 101)	Below
50	65	57	56.69	(51, 61)	In
51	112	106	97.64	(91, 104)	Above
52	103	75	85.71	(78, 93)	Below
53	92	66	75.03	(68, 82)	Below
54	75	60	65.74	(60, 71)	In
55	90	57	74.16	(67, 81)	Below
56	109	101	93.36	(86, 100)	Above
57	94	79	81.05	(74, 87)	In
58	112	105	97.23	(90, 104)	Above
59	101	97	84.36	(77, 91)	Above
60	111	97	93.09	(85, 100)	In
61	73	55	63.84	(58, 69)	Below
62	109	97	91.55	(84, 99)	In
63	67	61	57.76	(52, 63)	In
64	112	102	96.26	(89, 103)	In
65	112	90	92.61	(85, 100)	In
66	108	96	90.90	(83, 98)	In
67	89	80	76.01	(69, 82)	In
68	111	105	97.21	(90, 104)	Above
69	107	94	91.10	(84, 98)	In
70	81	70	70.21	(64, 76)	In
71	109	95	93.95	(87, 100)	In
72	103	95	87.25	(80, 94)	Above
73	112	101	99.20	(93, 105)	In
74	59	52	49.45	(44, 54)	In
75	111	94	91.81	(84, 99)	In
76	112	95	97.16	(90, 104)	In
77	99	87	84.83	(78, 91)	In
78	112	87	94.78	(87, 102)	In
79	108	91	92.92	(86, 100)	In
80	110	104	97.37	(90, 103)	Above

81	100	76	83.29	(76, 90)	In
82	111	95	97.98	(91, 104)	In
83	106	84	88.25	(81, 95)	In
84	90	73	73.35	(66, 80)	In
85	112	101	95.44	(88, 102)	In
86	107	101	94.21	(87, 100)	Above
87	88	75	74.12	(67, 80)	In
88	84	73	72.88	(67, 78)	In
89	111	94	96.46	(89, 103)	In
90	91	67	77.43	(71, 84)	Below
91	78	62	67.09	(61, 73)	In
92	103	87	88.27	(81, 95)	In
93	112	109	98.57	(92, 105)	Above
94	101	78	83.75	(76.025, 91)	In
95	75	56	63.31	(57, 69)	Below
96	111	96	92.27	(84, 99)	In
97	103	95	88.01	(81, 95)	In
98	107	86	91.03	(84, 98)	In
99	104	92	89.13	(82, 96)	In
100	111	106	95.43	(88, 102)	Above
101	90	78	77.58	(71, 83)	In
102	106	78	86.14	(78, 94)	In
103	108	98	90.18	(83, 97)	Above
104	102	84	86.46	(79, 93)	In
105	44	36	38.89	(35, 43)	In
106	102	90	87.63	(81, 94)	In
107	97	78	83.07	(76, 89)	In
108	101	85	85.01	(78, 92)	In
109	106	94	92.01	(85, 98)	In
110	92	81	76.67	(70, 83)	In
111	82	61	65.41	(58, 72)	In
112	112	95	90.78	(83, 98)	In
113	111	104	93.29	(86, 100)	Above
114	65	57	54.95	(49, 60)	In
115	78	73	69.24	(64, 74)	In
116	73	56	62.59	(57, 68)	Below
117	108	97	91.35	(84, 98)	In
118	92	76	79.12	(72, 85)	In
119	72	56	60.71	(54, 66)	In
120	108	98	88.92	(81, 96)	Above
121	91	77	79.17	(73, 85)	In
122	52	35	41.47	(36, 47)	Below
123	111	92	95.97	(89, 102)	In
124	97	87	82.96	(76, 89)	In
125	68	45	58.42	(53, 63.975)	Below

126	109	96	94.94	(88, 101)	In
127	109	99	94.64	(87.025, 101)	In
128	111	76	91.02	(83, 98)	Below
129	108	98	92.05	(85, 99)	In
130	112	100	92.38	(85, 100)	In
131	101	82	84.87	(78, 92)	In
132	104	97	90.07	(83, 96)	Above
133	102	85	85.29	(78, 92)	In
134	103	78	84.14	(76, 91)	In
135	68	57	57.02	(51, 63)	In
136	93	78	80.31	(74, 86)	In
137	112	93	89.76	(81, 97)	In
138	96	80	81.80	(75, 88)	In
139	107	98	90.82	(84, 98)	In
140	71	57	59.40	(53, 65)	In
141	85	70	73.61	(67, 79)	In
142	110	99	94.40	(87, 101)	In
143	70	64	56.71	(50, 63)	Above
144	91	75	77.03	(70, 83)	In
145	90	81	76.52	(70, 83)	In

Appendix L

Aggregated Response Sequences by Unit and Level of Cognitive Demand

Table L.1

Aggregated Response Sequences by Unit and Level of Cognitive Demand

		Sequence[1]								
Unit	Cognitive demand	1	2	3	4	5	6	7	8	**Total**
Descriptive statistics	Recall	15	31	34	104	57	123	243	990	**1,597**
	Basic application of skill	8	26	27	67	21	115	155	1,028	**1,447**
	Strategic thinking	32	80	91	118	76	96	135	289	**917**
Probability	Recall	28	16	27	24	73	49	134	176	**527**
	Basic application of skill	38	47	52	127	64	130	186	623	**1,267**
	Strategic thinking	43	31	78	59	64	76	132	253	**736**
One-sample inference	Recall	0	5	4	24	21	63	206	520	**843**
	Basic application of skill	19	15	58	117	64	109	354	802	**1,538**
	Strategic thinking	36	32	62	62	129	137	268	481	**1,207**
Two-sample inference	Recall	6	5	23	39	60	51	215	622	**1,021**
	Basic application of skill	21	44	40	88	84	170	329	1245	**2,021**
	Strategic thinking	38	17	86	85	102	64	260	349	**1,001**
Total		**284**	**349**	**582**	**914**	**815**	**1,183**	**2,617**	**7,378**	**14,122**

Note. [1] Sequence of all three responses: student response system, quiz, and exam: 1 – All three questions incorrect; 2 – Only student response system question correct; 3 – Only quiz question correct; 4 – Student response system and quiz question correct, exam question incorrect; 5 – Only exam question correct, quiz question incorrect; 6 – Student response system question and exam question correct, quiz question incorrect; 7 – Quiz question and exam question correct, student response system question incorrect; 8 – All three questions correct

Appendix M

Comparison of Odds Ratios for Two-Factor Interaction Loglinear Model and Saturated Loglinear Model

Table M.1

Comparison of Odds Ratios for Two-Factor Interaction Loglinear Model and Saturated Loglinear Model

Interaction Levels[1, 2, 3]	Third Variable	(UC, US, CS) Odds Ratio	(UCS) Odds Ratios at Every Level of Third Variable

U = P
C = BAS
	Sequence	2.74	(6.36, 10.49, 5.87, 49.4, 5.70, 7.77, 4.71, 11.25)

U = P
C = ST
	Sequence	2.06	(0.44, 0.47, 0.98, 4.26, 0.36, 4.01, 3.56, 45.6)

U = 1-SI
C = BAS
	Sequence	2.12	(17.29, 2.69, 7.39, 3.97, 4.26, 1.55, 1.95, 1.31)

U = 1-SI
C = ST
	Sequence	2.56	(11.13, 2.08, 3.60, 1.80, 2.92, 2.05, 1.80, 2.20)

U = 2-SI
C = BAS
	Sequence	2.28	(19.11, 33.45, 4.31, 7.61, 8.85, 7.69, 4.06, 2.83)

U = 2-SI
C = ST
	Sequence	1.74	(7.54, 2.64, 2.25, 3.60, 1.80, 2.66, 4.31, 3.42)

U = P
S = 1
	Cognitive Demand	0.33	(0.06, 0.09, 0.06)

U = P
S = 2
	Cognitive Demand	0.51	(0.12, 0.12, 0.28)

U = P
S = 3
	Cognitive Demand	0.36	(0.00, 0.04, 0.05)

U = P
S = 4
	Cognitive Demand	0.71	(0.36, 0.32, 0.3)

U = P
S = 5
	Cognitive Demand	0.38	(0.01, 0.02, 0.13)

U = P
S = 6
	Cognitive Demand	0.46	(0.03, 0.02, 0.22)

U = P
S = 7
	Cognitive Demand	0.24	(0.00, 0.00, 0.08)

U = 1-SI
S = 1
	Cognitive Demand	0.34	(2.83, 0.44, 0.53)

U = 1-SI
S = 2
	Cognitive Demand	0.80	(2.27, 0.97, 0.73)

U = 1-SI
S = 3
	Cognitive Demand	0.84	(3.78, 0.87, 0.61)

U = 1-SI
S = 4
	Cognitive Demand	1.77	(5.21, 1.28, 1.36)

U = 1-SI
S = 5
	Cognitive Demand	1.05	(6.62, 0.59, 1.22)

U = 1-SI
S = 6
	Cognitive Demand	2.03	(9.49, 1.07, 1.54)

U = 1-SI
S = 7
	Cognitive Demand	1.05	(6.82, 0.52, 1.35)

U = 2-SI
S = 1
	Cognitive Demand	0.44	(0.26, 0.46, 0.09)

U = 2-SI
S = 2
	Cognitive Demand	0.83	(2.59, 0.58, 0.77)

U = 2-SI
S = 3
	Cognitive Demand	0.45	(0.89, 2.23, 0.42)

U = 2-SI
S = 4
	Cognitive Demand	1.12	(6.49, 3.00, 1.55)

U = 2-SI
S = 5
	Cognitive Demand	0.61	(1.07, 0.43, 0.38)

U = 2-SI
S = 6
	Cognitive Demand	1.00	(5.00, 1.06, 2.86)

U = 2-SI
S = 7
	Cognitive Demand	0.66	(2.53, 0.38, 1.15)

C = BAS
S = 1
	Unit	1.62	(2.41, 3.97, 0.38, 4.22)

C = BAS
S = 2
	Unit	1.25	(2.18, 2.01, 0.93, 0.49)

C = BAS
S = 3
	Unit	1.37	(1.51, 11.70, 0.35, 0.60)

C = BAS
S = 4
	Unit	0.64	(0.48, 0.43, 0.12, 0.22)

C = BAS
S = 5
	Unit	1.17	(3.42, 4.18, 0.31, 0.73)

C = BAS
S = 6
	Unit	0.76	(1.49, 1.11, 0.17, 0.32)

C = BAS
S = 7
	Unit	1.05	(4.53, 8.00, 0.34, 0.67)

C = ST
S = 1
	Unit	1.09	(1.65, 1.79, 0.31, 0.58)

C = ST
S = 2
	Unit	1.26	(1.84, 4.14, 0.59, 0.55)

C = ST
S = 3
	Unit	0.61	(0.16, 1.54, 0.03, 0.08)

C = ST
S = 4
	Unit	0.57	(0.27, 0.22, 0.07, 0.06)

C = ST
S = 5
	Unit	0.45	(0.05, 0.48, 0.01, 0.02)

C = ST
S = 6
	Unit	0.32	(0.02, 0.14, 0.00, 0.01)

C = ST
S = 7
	Unit	0.21	(0.00, 0.21, 0.00, 0.00)

Note. [1] Unit: P – Probability, 1-SI – One-sample inference; 2-SI – Two-sample inference; Descriptive statistics used as baseline

Note. [2] Level of cognitive demand: BAS – Basic application of skill; ST – Strategic thinking; Recall used as baseline

Note. [3] Sequence of all three responses: student response system, quiz, and exam: 1 – All three questions incorrect; 2 – Only student response system question correct; 3 – Only quiz question correct; 4 – Student response system and quiz question correct, exam question incorrect; 5 – Only exam question correct, quiz question incorrect; 6 – Student response system question and exam question correct, quiz question incorrect; 7 – Quiz question and exam question correct, student response system question incorrect; 8 – All three questions correct

Appendix N

Expected Counts and Pearson Residuals for Saturated Loglinear Model

Table N.1

Expected Counts for Saturated Loglinear Model

		Sequence[1]								
Unit	Cognitive demand	1	2	3	4	5	6	7	8	**Total**
Descriptive statistics	Recall	32.1	39.5	65.8	103.4	92.2	133.8	295.9	834.3	**1,597**
	Basic application of skill	29.1	35.8	59.6	93.7	83.5	121.2	268.1	756.0	**1,447**
	Strategic thinking	18.4	22.7	37.8	59.3	52.9	76.8	169.9	479.1	**917**
Probability	Recall	10.6	13.0	21.7	34.1	30.4	44.1	97.7	275.3	**527**
	Basic application of skill	25.5	31.3	52.2	82.0	73.1	106.1	234.8	661.9	**1,267**
	Strategic thinking	14.8	18.2	30.3	47.6	42.5	61.7	136.4	384.5	**736**
One-sample inference	Recall	17.0	20.8	34.7	54.6	48.7	70.6	156.2	440.4	**843**
	Basic application of skill	30.9	38.0	63.4	99.5	88.8	128.8	285.0	803.5	**1,538**
	Strategic thinking	24.3	29.8	49.7	78.1	69.7	101.1	223.7	630.6	**1,207**
Two-sample inference	Recall	20.5	25.2	42.1	66.1	58.9	85.5	189.2	533.4	**1,021**
	Basic application of skill	40.6	49.9	83.3	130.8	116.6	169.3	374.5	1,056	**2,021**
	Strategic thinking	20.1	24.7	41.3	64.8	57.8	83.9	185.5	523.0	**1,001**
Total		**284.0**	**349.0**	**582.0**	**914.0**	**815.0**	**1,183.0**	**2,617.0**	**7,378.0**	**14,122**

Note. [1] Sequence of all three responses: student response system, quiz, and exam: 1 – All three questions incorrect; 2 – Only student response system question correct; 3 – Only quiz question correct; 4 – Student response system and quiz question correct, exam question incorrect; 5 – Only exam question correct, quiz question incorrect; 6 – Student response system question and exam question correct, quiz question incorrect; 7 – Quiz question and exam question correct, student response system question incorrect; 8 – All three questions correct

Table N.2

Pearson Residuals for Saturated Loglinear Model

		Sequence[1, 2]								
Unit	Cognitive demand	1	2	3	4	5	6	7	8	**Total**
Descriptive statistics	Recall	-3.02	-1.35	-3.92	0.06	-3.67	-0.93	-3.08	**5.39**	**1,597**
	Basic application of skill	-3.91	-1.64	-4.22	-2.76	***-6.84***	-0.56	-6.91	**9.89**	**1,447**
	Strategic thinking	3.17	**12.03**	**8.65**	**7.62**	3.18	2.19	-2.68	***-8.68***	**917**
Probability	Recall	**5.34**	0.83	1.14	-1.73	**7.73**	0.74	3.67	***-5.98***	**527**
	Basic application of skill	2.48	2.81	-0.03	4.97	-1.06	2.32	-3.18	-1.51	**1,267**
	Strategic thinking	**7.33**	3.00	**8.67**	1.65	3.30	1.82	-0.38	***-6.71***	**736**
One-sample inference	Recall	-4.12	-3.46	***-5.21***	-4.14	-3.97	-0.90	3.98	3.79	**843**
	Basic application of skill	-2.14	-3.73	-0.68	1.75	-2.63	-1.74	4.09	-0.05	**1,538**
	Strategic thinking	2.37	0.40	1.74	-1.82	**7.10**	3.57	2.96	***-5.96***	**1,207**
Two-sample Inference	Recall	-3.20	-4.02	-2.94	-3.33	0.14	-3.73	1.88	3.84	**1,021**
	Basic application of skill	-3.08	-0.84	-4.74	-3.74	-3.02	0.05	-2.35	**5.82**	**2,021**
	Strategic thinking	3.99	-1.55	**6.96**	2.51	**5.81**	-2.17	**5.47**	***-7.61***	**1,001**
Total		**284.00**	**349.00**	**582.00**	**914.00**	**815.00**	**1,183.00**	**2,617.00**	**7,378.00**	**14,122**

Note.[1] Bold cells indicate a positive Pearson residual greater than 5. Bold and italicized cells indicate a negative Pearson residual less than -5.

Note.[2] Sequence of all three responses: student response system, quiz, and exam: 1 – All three questions incorrect; 2 – Only student response system question correct; 3 – Only quiz question correct; 4 – Student response system and quiz question correct, exam question incorrect; 5 – Only exam question correct, quiz question incorrect; 6 – Student response system question and exam question correct, quiz question incorrect; 7 – Quiz question and exam question correct, student response system question incorrect; 8 – All three questions correct