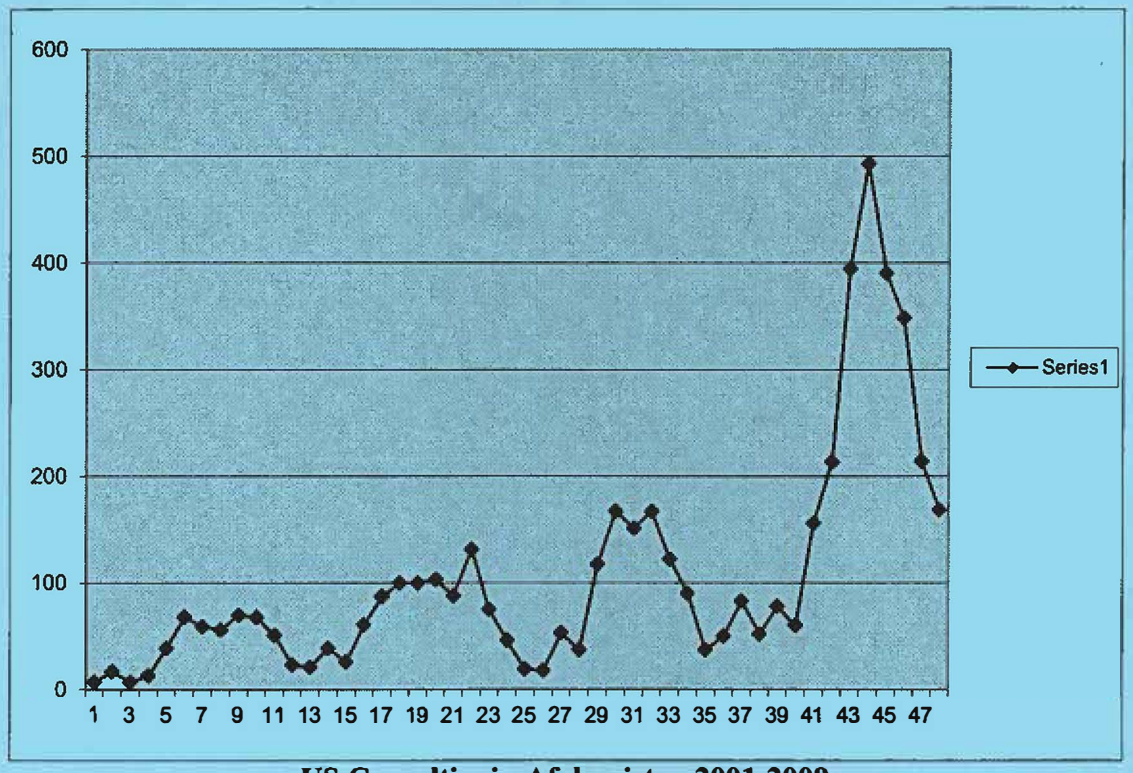
2012

Study Guide Mathematical Modeling for Decision Making II DA 3410

Fox, William Dr.

Monterey, California. Naval Postgraduate School

STUDY GUIDE
Mathematical Modeling for Decision Making II
DA 3410



US Casualties in Afghanistan 2001-2009

Dr. William P. Fox

MODELING FOR MILITARY DECISION MAKING II

DA 3410

Introduction and background

The mission of the U.S. Army Special Operations Command is to organize, train, educate, man, equip, fund, administer, mobilize, deploy and sustain Army special operations forces to successfully conduct worldwide special operations, across the range of military operations, in support of regional combatant commanders, American ambassadors and other agencies as directed.

This mission requires military leaders to manage, obtain, and utilize massive resources in order to successfully conduct worldwide operations as directed. To adequately do the jobs and missions required, someone needs to use mathematical tools to "expect the unexpected, analyze the requirements, forecast the needs, maintain, and win." This someone is you.

Every day we are bombarded by data. Whether you get your news from CNN, ESPN, the New York Times, or the internet, raw data will not help you if you cannot make sense of it. Military leaders at all levels need to understand how to interpret the great variety of data they see in order to be effective. This course will give you the ability to turn raw data into information that you can use. Crudely put, *statistics* is the mathematical study of data, while *probability* is the mathematical study of chance.

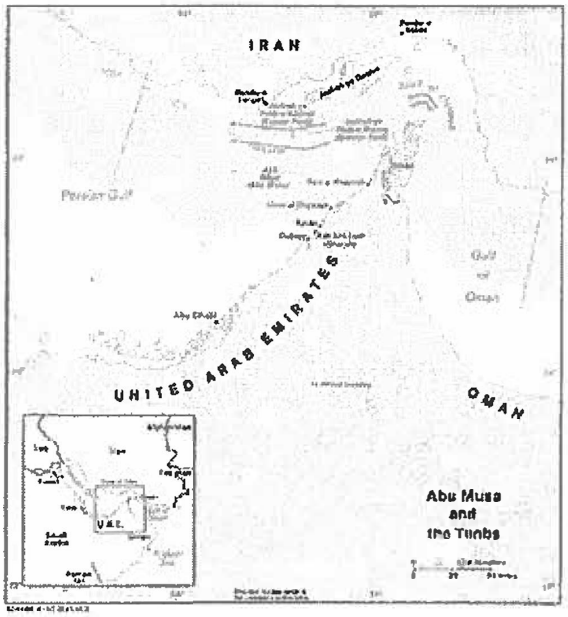
Critical to this is the study and use to statistics and probability to analyze, predict, and forecast. The collection of data that can be organized summarized and analyzed has been pursued for hundreds of years. The part of statistics that deals with methods for performing those operations is called **descriptive statistics**. We will spend a few lessons on this topic. When the data is a sample and the objective is to draw conclusions about the population based on the sample information, methods from **inferential statistics** are used. We will spend lesson discusses these techniques also.

Think of the S-1 who collects information of SIDPERS daily. These reports provide data on the readiness of the unit. Think of the motor officer who is trying to improve the readiness rates of vehicles dead-lined for repairs or parts. In some units that are monthly (or more frequent) meetings with the generals to discuss these readiness indicators and discuss how to improve the numbers. Consider the fact the military is growing with the addition of new units. How are we going to attract and keep enough soldiers and officer to meet the future needs with an all-volunteer force?

When we plan a mission, we do want to know the answer to the question "what is our chance of success?". When the chances are small we probably will not execute that mission but when the chances are good to great we have more confidence in the success of the mission.

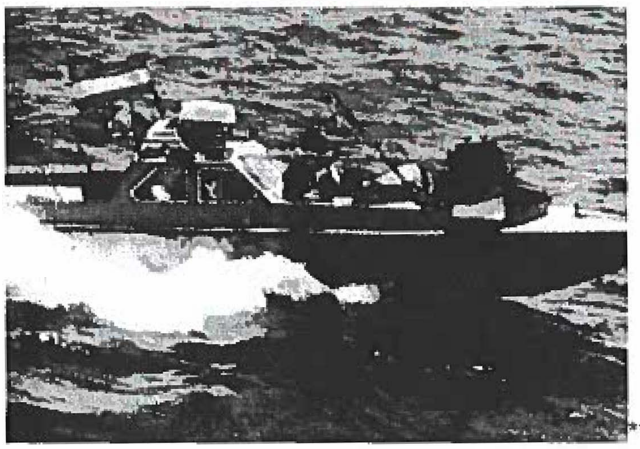
One of the biggest mission failures could be considered the rescue mission of the hostages in Iran. A little more mathematically based analysis should have told the planners to send more helicopters to insure that six (the minimum number required for mission success) helicopters were operational when the mission was to begin. The mission was aborted because not enough operational helicopters were on the ground in the desert when the mission was to take off.

Consider this extracted from a recent article:
The U.S. Navy has determined that Iran has amassed a fleet of fast patrol boats in the 43-kilometer straits. Iran's Islamic Revolutionary Guard Corps, responsible for strategic programs, leads the effort.



At this point, officials said, IRGC has deployed more than 1,000 FPBs in and around the straits. The vessels, armed with cruise missiles, mines, torpedoes and rocket-propelled grenades, are up to 23 meters in long and can reach a speed of 100 kilometers per hour. *** "This marks the implementation of Iran's swarm program, where dozens of armed speed boats attack much larger naval vessels from all sides," an official said.
***

IRGC swarming tactics envision a group of more than 100 speedboats attacking a target, such as a Western naval vessel or a commercial oil tanker. They said 20 or more speedboats would strike from each direction, making defense extremely difficult.

 ***

"We have devised various tactics and other ways of coping," U.S. commander Vice Adm. Kevin Cosgriff said. "You just don't get 1,000 or 500 or even 20 of anything under way and tightly orchestrated over a large body of water to create a specific effect at a specific time and specific place. They have their own challenges."
Wait, they won't just magically appear all around a carrier battle group all at once? Even with their incredible Iranian stealth attack ground effect boat/planes?

More on "swarm" tactics and on the Iranian stealth effort here under the title of "Iran's Doctrine of Asymmetric Naval Warfare" -

Swarming tactics are not new; they have been practiced by land armies for thousands of years. Such tactics require light, mobile forces with substantial striking power, capable of rapidly concentrating to attack an enemy from multiple directions and then rapidly dispersing.

Iranian naval swarming tactics focus on surprising and isolating the enemy's forces and preventing their reinforcement or resupply, thereby shattering the enemy's morale and will to fight. Iran has practiced both mass and dispersed swarming tactics. The former employs mass formations of hundreds of lightly armed and agile small boats that set off from different bases, then converge from different directions to attack a target or group of targets. The latter uses a small number of highly agile missile or torpedo attack craft that set off on their own, from geographically dispersed and concealed locations, and then converge to attack a single target or set of targets (such as a tanker convoy). The dispersed swarming tactic is much more difficult to detect and repel because the attacker never operates in mass formations.

During the Iran-Iraq War, the Pasdaran navy used mass swarming tactics; as a result, its forces proved vulnerable to attack by U.S. naval and air power. Because of this, it is unlikely that such tactics would be used for anything but diversionary attacks in the future. In today's Iranian naval forces, mass swarming tactics have largely given way to dispersed swarming.

Dispersed swarming tactics are most successful when attackers can elude detection through concealment and mobility, employ stand-off firepower, and use superior

situational awareness (intelligence), enabling them to find and engage the enemy first. This accounts for a number of trends in Iranian naval force development in the past two decades. The first is the acquisition and development of small, fast weapons platforms—particularly lightly armed small boats and missile-armed fast-attack craft; extended- and long-range shore- and sea-based antiship missiles; midget and diesel attack submarines (for intelligence gathering, covert mine laying, naval special warfare, and conventional combat operations); low-signature reconnaissance and combat unmanned aerial vehicles (UAVs); and the adaptation of the Shahab-3 medium-range surface-to-surface missile armed with a cluster warhead reportedly carrying 1,400 bomblets, for use against enemy naval bases and carrier battle groups.

Iran has also sought to improve its ability to achieve surprise by employing low-observable technologies (such as radar-absorbent paints), strict communications discipline, stringent emissions control measures, passively or autonomously guided weapons systems (such as the Kowsar series of television-guided antiship missiles), and sophisticated command-and-control arrangements. To support its naval swarm tactics, Iran has encouraged decentralized decision making and initiative, as well as autonomy and self-sufficiency among naval combat elements.

Dispersed swarming? Adm Cosgrove has it right - a coordinated attack is difficult to conceal and an uncoordinated attack can lead to forces being defeated *seriatim*.

So are swarm tactic effective? How and when should they be used?

How does one answer these questions? Perhaps by understanding some statistical analysis one can begin to get a handle on the issue.

DA 3410 Statistical Symbols & Notation

Probability and statistics symbols table

Symbol	Symbol Name	Meaning / definition	Example
$P(A)$	probability function	probability of event A	$P(A) = 0.5$
$P(A \cap B)$	probability of events intersection	probability that of events A and B	$P(A \cap B) = 0.5$
$P(A \cup B)$	probability of events union	probability that of events A or B $P(A \cup B)$=P(A)+P(B)-$P(A \cap B)$	$P(A \cup B) = 0.5$
$P(A \mid B)$	conditional probability function	probability of event A given event B occurred	$P(A \mid B) = 0.3$
$f(x)$	probability density function (pdf)	$P(a \leq x \leq b)$	
$F(x)$	cumulative distribution function (cdf)	$F(x) = P(X \leq x)$	
μ	population mean	mean of population values	$\mu = 10$
$E[X]$	expected value	expected value of random variable X	$E(X) = 10$
$\bar{x}$	Sample mean	Mena of sample values	$\bar{x} = 10$
$var(X)$	variance	variance of random variable X	$var(X) = 4$
σ^2	Population variance	variance of population values	$\sigma^2 = 4$
S^2	Sample variance	Variance of sample values	S^2=4.2
$std(X)$	standard deviation	standard deviation of random variable X	$std(X) = 2$
σ_X	standard deviation	standard deviation value of random variable X from a population	$\sigma_X = 2$

s	Sample standard deviation	standard deviation value of random variable X from a sample	s=2
$\tilde{x}$	median	middle value of random variable x	$\tilde{x} = 5$
$cov(X,Y)$	covariance	covariance of random variables X and Y	$cov(X,Y) = 4$
$corr(X,Y)$	correlation	correlation of random variables X and Y	$corr(X,Y) = 0.6$
r_{xy} or $\rho_{X,Y}$	correlation	correlation of random variables X and Y	$\rho_{X,Y} = 0.6$=r
$\sum$	summation	summation - sum of all values in range of series	$\sum_{i=1}^{4} x_i = x_1 + x_2 + x_3 + x_4$
$Mode$	mode	value that occurs most frequently in population	
MR	mid-range	$MR = (x_{max}+x_{min})/2$	
Md	sample median	half the population is below this value	
Q_1	lower / first quartile	25% of population are below this value	
Q_2	median / second quartile	50% of population are below this value = median of samples	
Q_3	upper / third quartile	75% of population are below this value	
$\bar{x}$	sample mean	average / arithmetic mean	$\bar{x} = (2+5+9) / 3 = 5.333$
s^2	sample variance	population samples variance estimator	$s^2 = 4$
s	sample standard deviation	population samples standard deviation estimator	$s = 2$
z_x	standard score	$z_x = (x-x) / s_x$	
$X \sim$	distribution of X	distribution of random variable X	$X \sim N(0,3)$
$N(\mu,\sigma^2)$	normal distribution	gaussian distribution	$X \sim N(0,3)$
$U(a,b)$	uniform distribution	equal probability in range a,b	$X \sim U(0,3)$
$exp(\lambda)$	exponential distribution	$f(x) = \lambda e^{-\lambda x}$, $x \geq 0$	

$\chi^2(k)$	chi-square distribution	$f(x) = x^{k/2-1} e^{-x/2} / (2^{k/2} \Gamma(k/2))$	
$F(k_1, k_2)$	F distribution		
$Bin(n,p)$	binomial distribution	$f(k) = {}_nC_k \, p^k (1-p)^{n-k}$	
$Poisson(\lambda)$	Poisson distribution	$f(k) = \lambda^k e^{-\lambda} / k!$	
$Geom(p)$	geometric distribution	$f(k) = p(1-p)^k$	
t-dist	Student T-distribution	Normal with both μ and σ unknown. Estimate with $\bar{x}$ and S.	
$Bern(p)$	Bernoulli distribution	a 0 or 1 variable.	

Combinatorics Symbols

Symbol	Symbol Name	Meaning / definition	Example
$n!$	factorial	$n! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot n$	$5! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 = 120$
${}_nP_k$	permutation	${}_nP_k = \frac{n!}{(n-k)!}$	${}_5P_3 = 5! / (5-3)! = 60$
${}_nC_k$ $\binom{n}{k}$	combination	${}_nC_k = \binom{n}{k} = \frac{n!}{k!(n-k)!}$	${}_5C_3 = 5!/[3!(5-3)!] = 10$

Block I: Introduction to Statistics, Data, measures, and Displays

Lesson 1-3 Objectives:

1. Know the different types of data (categorical (qualitative) and quantitative) and which displays are used for the type of data that you are using & displaying.
2. Know what constitutes a bad display. Avoid them.
3. Know how to use Excel to create "good" and useful displays of data.
4. Know how to obtain and use descriptive statistics.
5. Know how to interpret key measures mean, median, mode, range, standard deviation, variance & skewness.
6. Distinguish measures of location from measure of spread.
7. Learn how to use Excel to obtain descriptive statistics
8. Learn how to use Excel to obtain displays

Lesson 1 Introduction to Statistics

In probability problems properties of the population are assumed to be known and questions regarding a sample are posed and answered. In a statistics problem, characteristics of the sample are available to the experimenter, and this information enables the experimenter to draw conclusions about the population. The relationship between the two disciplines is illustrated in figure 1.

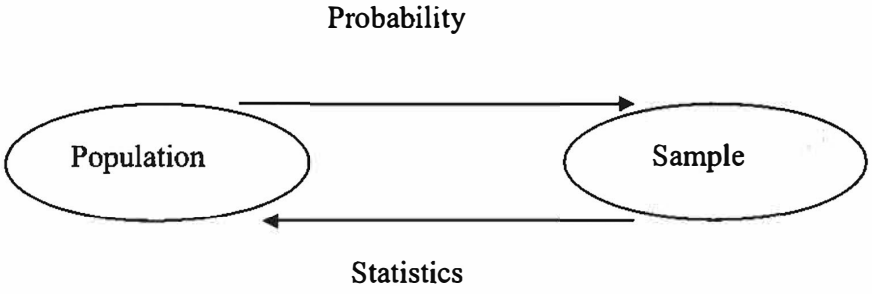


Figure 1. The relationship between Probability and inferential statistics

In this course we will learn the following topics so that we can readily use these topics to analyze data or make decisions about mission and/or operations:

Course Coverage

Statistics ←→ Data
Populations & Samples
Displays of Statistical Data
- Qualitative data displays: Pie and Bar Charts
- Quantitative data displays: Stem and Leaf, Histogram, Boxplot

Descriptive Statistics
- Data Types
  - Quantitative and Qualitative (categorical)
- Measures of Location & Dispersion
  - Location
    - Mean
    - Median
    - Mode
  - Dispersion
    - Range
    - Standard Deviation
    - Variance
    - Coefficient of Skewness

Classical Probability
- Probability Axioms & Rules
- Tree Diagrams
- Conditional Probability
- Independence
- Bayes' Theorem & Bayes' Rule

Random Variables
- PMF and CDF for Discrete Distributions
  - Poisson and Binomial Distributions
- PDF and CDF Continuous Distributions
  - Exponential and Normal Distributions
- Expected Value
- Hypothesis tests

Regression Analysis
- Simple linear regression (review)
- Multiple Linear Regression
- Nonlinear Regression
- Logistic Regression
- Poisson Regression

Multi-attribute Decisions Making
- Data Envelopment Analysis
- Analytical Hierarchy Process
- TOPSIS

ACTIVITY : Class Data Survey –

_____ 1. Age (in years)

_____ 2. Sex
(female =2, male =1)

_____ 3. Your height in inches.

_____ 4. Your weight in pounds.

_____ 5. Estimate your average travel time to school in minutes.

_____ 6. What is your most frequent mode of transportation to school?
(Foot = 1, Car =2, Bike = 3, Other = 4)

_____ 7. Number of people in your household.

_____ 8. Employment status.
(working = 1, not working = 2, looking for a job = 3)

_____ 9. Your attitude towards mathematics.
(hate it = 1, tolerate it = 2, like it = 3)

_____10. Branch of service(1-army, 2- navy, 3 -air force, 4- marine, 5-international, 6 –civilian, 7-other)

_____11. Department of Study Code

Did we cover all the possible answers that someone could put?

Should we modify our survey (yes or no)? Why?

In a recent report issued in GEN(Ret) Barry McCaffrey stated that the situation in Afghanistan is going to get much worse.

NEWSER) – America should prepare for thousands of casualties when the beefed-up US force in Afghanistan begins its spring offensive against the Taliban, warns a retired general and military expert. Gen. Barry McCaffrey, an analyst who has visited the war zone several times, predicts US casualties will rise to as many as 500 killed and wounded per month, matching or exceeding the deadliest months of last year's fighting.

"What I want to do is signal that this thing is going to be up to $10 billion a month and 300 to 500 killed and wounded a month by next summer," McCaffrey tells the *Army Times*. "That's what we probably should expect. And that's light casualties." McCaffrey predicts that building a viable Afghan state able to take care of its own security will take as long as 10 years[1]

[1] http://www.newser.com/story/77563/general-brace-for-thousands-of-gi-casualties.html#ixzz0v5jSGM3W

We will use US casualty data from Afghanistan for most of this course. The plot of the data, figure 2, over time is our cover plot,

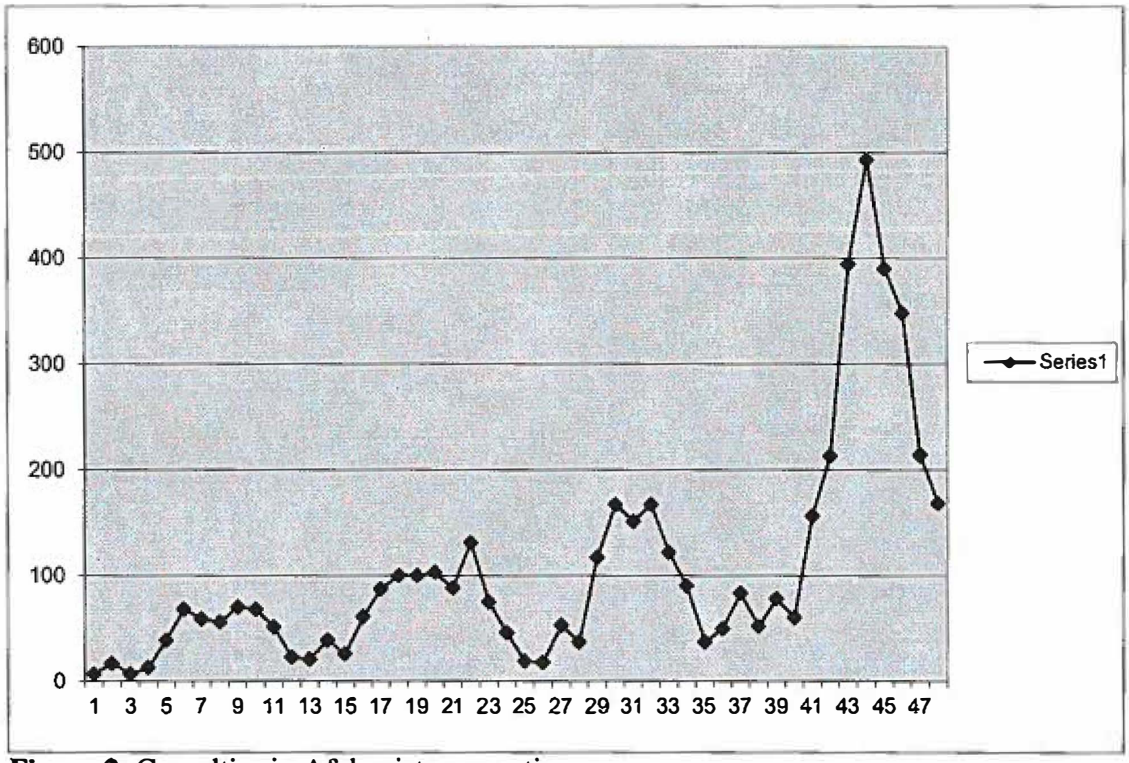


Figure 2. Casualties in Afghanistan over time

We will want to know everything there is to know about this data. We can apply our methods to any data that we encounter in our careers that require statistical analysis.

How to Prepare for a Modeling Class

This is a suggestion on how to prepare for each mathematical modeling class. Review the previous class material, prepare questions for class, and ask them. Next, read (scan) the lesson material for that day's class. If you come in cold, it is very difficult to stay up with the instructor. There are many areas of statistics that other DA professors have indicated the need for coverage. To cover that material some prep material is usually required as well. That makes for quite a lot of material in a short quarter. Thus, we go fast and expect work done by the student outside of class. Additionally, this is a modeling class and building the model and interpreting the model is key. We can't do so without adequate preparation. Thus, read over the class material to see what is coming. Look over the posted PowerPoint slides that narrow down the focus of emphasis for that class. Bring them to class and take notes on them. Try the homework problems after class. Ask questions in class or come visit me during office hours. Do not get behind. Repeat the process for the next class.

Lesson 1

Objectives

1. **Know the different types of data (categorical (qualitative) and quantitative) and which displays are used for the type of data that you are using & displaying.**
2. **Know the difference between a *population* and a *sample.***
3. **Know what constitutes a bad display. Avoid them.**
4. **Know how to use Excel to create "good" and useful displays of data.**
5. **Know how to obtain and use descriptive statistics.**
6. **Know how to interpret key measures mean, median, mode, range, standard deviation, variance & skewness.**
7. **Distinguish measures of location from measure of dispersion.**

Basic Statistics-UNIVARIATE DATA

1.1 Discovering Basic Statistics

Statistics is the ***science of reasoning from data***, so a natural place to begin your study is by examining what is meant by the term "data." The most fundamental principle in statistics is that of variability. If the world were perfectly predictable and showed no variability, there would be no need to study statistics. You will need to discover the notion of a **variable** and then first learn how to classify variables.

Any characteristic of a person or thing that can be expressed as a number is called a ***variable***. A *value* of that variable is the actual number that describes that person or thing. Think of the variables that might be used to describe you: height, weight, income, rank, branch of service, and gender.

Data can be ***quantitative*** or ***categorical (qualitative)***.

Quantitative means that the data are numerical where the number has relative meaning. Examples could be a list of heights of students in your class, weights of soldiers in your unit, or batting averages of the starting line-up for the 1998 New York Yankees.

Heights:

5'10"	6'2"	5'5"	5'2"	6'	5'9"

Weights in a squad

135	155	215	192	173	170	165	142

Yankee Batting Averages for their lineup:

.276	.320	.345	.354	.269	.275	.300	.254	.309

IED deaths over the from 2001-2014 (http://icasualties.org/oef/)

IED Fatalities

Period	IED	Total	Pct
2001	0	4	0.00
2002	4	25	16.00
2003	3	26	11.54
2004	12	27	44.44
2005	20	73	27.40
2006	41	130	31.54
2007	78	184	42.39
2008	152	263	57.79
2009	275	451	60.98
2010	368	630	58.41
2011	252	492	51.22
2012	132	312	42.31
2013	52	117	44.44
2014	3	13	23.08

These data elements provide numerical information. We can determine from the data which height is the tallest or smallest or which batting average is the greatest or the smallest. We can also compare and contrast "mathematically" these values.

Quantitative data can be either discrete (counting data) or continuous. These types become important as we analyze them and use them in models later in the course. Quantitative data allows us to "do meaningful mathematics." (+,-,*,/)

Categorical (qualitative) data can describe the objects, such as recording the people with a particular hair color as: blonde = 1 or brunette = 0. If we had four colors of hair: blonde, brunette, black and red; we could use as codes: brunette = 0, blonde = 1, black = 2, and red = 3. We certainly cannot have an average hair color from these numbers. It would not make sense. Another example is categories by gender: male = 0 and female =1. In general, it may not make sense to do any arithmetic using categorical variables. Ranks: LT, CPT, MAJOR, LTC, etc are categories as well as services: Army, Navy, Air Force, Marine, Coast Guard, or International.

Once you have learned to distinguish between quantitative and categorical data, we need to move on to a fundamental principle of data analysis: "begin by looking at a visual display of the data set", see figure 2.

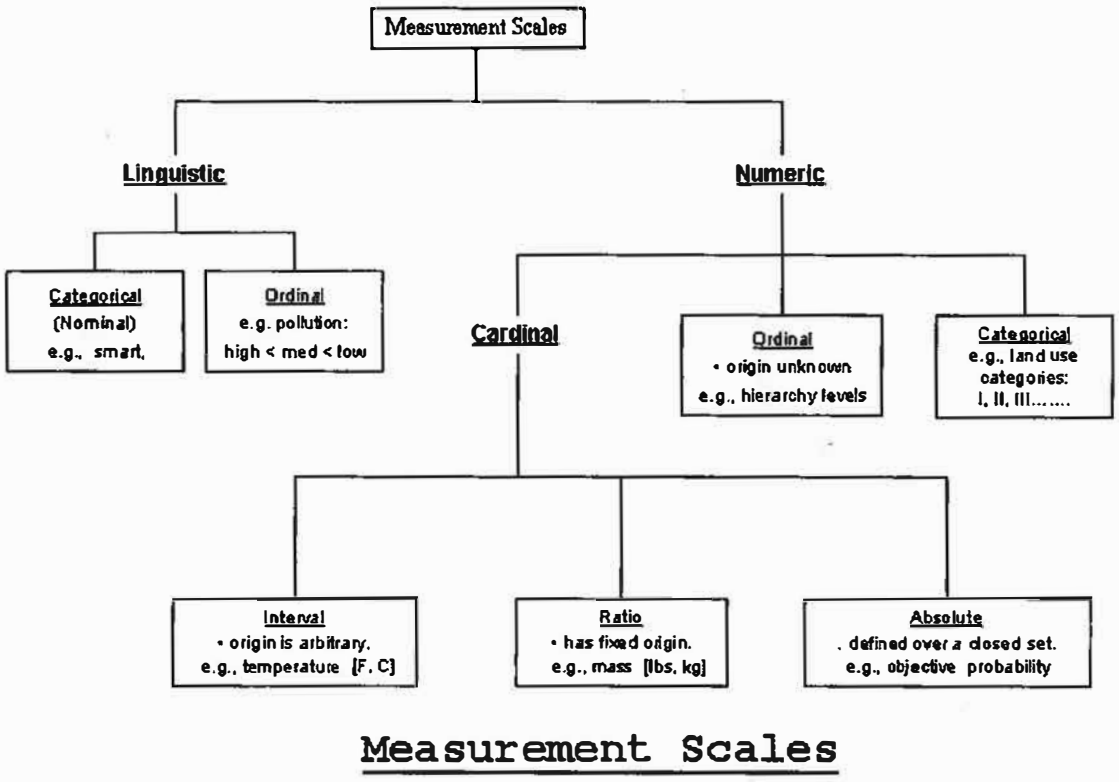


Figure 2. Measurement Scales

1.1 Exercises:

Determine whether the following variables would be **quantitative** or **categorical.** Provide an example of the value of such a variable and include the units, if any units exist.

(1) Flip a penny that lands as a "head" or "tail"

(2) the color of M&M's

(3) the number of calories in the local fast food selections

(4) the life expectancy for males in the United States

(5) the life expectancy females for in the United States

(6) the number of babies born on New Year's eve

(7) the dollars spent each month out of the allocated supply budget

(8) the number of hours that a soldier works per week

(9) amount of car insurance paid per year

(10) whether the bride is older, younger, or the same age as the groom

(11) the difference in ages of a couple at a wedding

(12) average low temperature in Monterey, CA, in January

(13) the eye color of a soldier

(14) the gender of a soldier

(15) the number of intramural sports a person plays per year

(16) the distance a bullet travel from a specific weapon

(17) the number of deployments in 3 years

(18) the location a bullet hits on a target

(19) DATA INTERPRETATION: The following table represents the numbers of sports-related injuries treated in U.S. hospital emergency rooms in 2001, along with an estimate of the number of participants in that sport.

Sport	Injuries	Participants	Sport	Injuries	Participants
Basketball	646,678	26,200,000	Fishing	84,115	47,000,000
Bicycling	600,649	54,000,000	Skateboard	56,435	8,000,000
Baseball	459,542	36,100,000	Hockey	54,601	1,800,000
Football	453,684	13,300,000	Golf	38,626	24,700,000
Soccer	150,449	10,000,000	Tennis	29,936	16,700,000
Swimming	130,362	66,200,000	Water skiing	26,663	9,000,000
Weightlifting	86,398	39,200,000	Bowling	25,417	40,400,000

(a) If we want to use the number of injuries as a measure of the hazardousness of a sport, which sport is more hazardous between bicycling and football? Between soccer and hockey?

(b) use either a calculator or a computer to calculate the *rate* of injuries per thousand participants. *Rate* is defined as the average number of injuries out of the total participants.

(c) rank order this new measure for the sports.

(d) how do your answers in part (a) compare if we do the hazardous analysis using the *rates* in (b). If different, why are the results different?

1.2 Displaying the Data

Why do we want to display data? Visual displays such as bar graphs, pie charts, and histograms are very useful because they provide a quick and efficient way to present the data revealing characteristics of the data. These displays allow our eyes to take in the overall pattern and see if there are unusual observations of data elements. Graphs and numbers that we will introduce to describe the data are not ends in themselves, but merely aids to our overall understanding of the data.

Displaying data badly is a problem. What is meant by displaying data badly? We illustrate and explain.

The three fundamental elements of bad graphical display are these: **Data Ambiguity, Data Distortion, and Data Distraction.**

Data Ambiguity:

Data ambiguity arises from the failure to precisely define just what the data represent. Every dot on a scatterplot, every point on a time series line, every bar on a bar chart represents a number (actually, in the case of a scatterplot, two numbers). It is the job of the legend and labels (text) on the chart to tell us just what each of those numbers represents. If a number represented in a chart is, say, 33½, the text in the graph -- in the title, the axis labels, the data labels, the legend, and sometimes the footnote -- must answer question: "Thirty-three and a half what?".

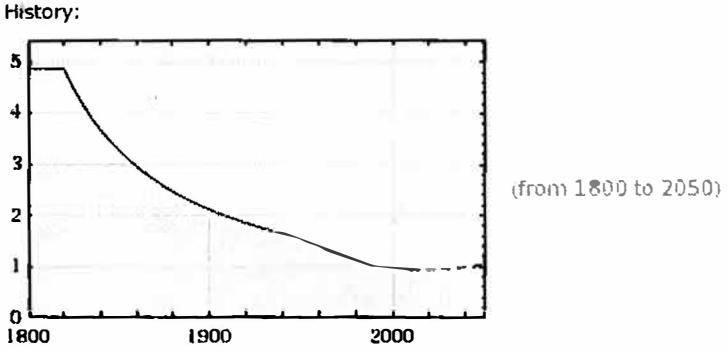


Figure 3. Data ambiguity in that this nice plot tells us nothing other than heights from 0-5 and years from 1800-2000+.

Data Distortion.

Before the development of spreadsheet graphing, the most common graphical mistake was the use of artist-drawn 3-D images with the height of 3-D objects representing the magnitude of the data points. In these charts, both the height and the width of the drawn object increase proportionate to the magnitude of the data points. The effect is to exaggerate the differences in magnitude as the viewer tends to perceive the area of the

figures rather than just the height as representing the magnitude. The incredible shrinking family doctor (shown in Tufte, p. 69) is a classic example. In this chart the 1990 doctor is a bit less than half the height of the 1964 doctor. Each doctor has the same relative shape. Imagine two doctors with the same average physical shape, one less than 4 feet tall, the other 8 feet tall. If the 4 ft. doctor weighed 100 lbs., how much would the 8 ft. doctor weigh? Certainly much more than 200 lbs.

The Shrinking Doctor

source: LA Times, August 5, 1979
from: Tufte, p. 69

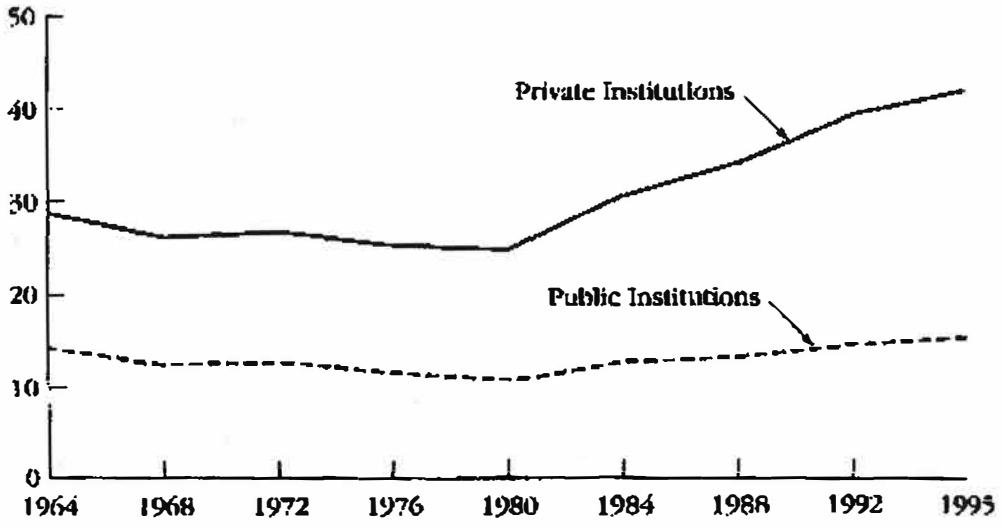
The Shrinking Dollar

source: Washington Post October 25, 1978
from: Tufte, p. 70

Figure 4. Data distortion were the imagines over or under measure the actual change.

With the development of spreadsheet graphics, such visual distortions are no longer common, and the Art of Lying with graphics has become a technology rather than an art. Today, altogether new forms of bad graphical design predominate.

Most of the bad charting described thus far has the redeeming feature that it does not for the most part distort the data being represented by exaggerating or understating the values of some of the data points. We will consider now some of the more complicated ways of using graphical display to mislead. Figure 5 is a time series chart originally printed in a public policy textbook authored by four professors of political science employed by three public universities.



SOURCE: U.S. Dept. of Education.

Figure 5. Average Tuition, Room and Board as a percentage of Median Family Income, 1964-1995

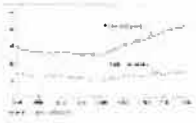


chart source: Cochran, 347

The interpretation of the graph is as follows:

There is some evidence that the cost of higher education may not have escalated so much... Figure 5 reflect the average cost for tuition, room, and board as a percentage of median family income from 1964 to 1995. While private institutions have increased costs substantially, public university costs have remained constant. This indicates that the increased costs associated with higher education may be quite reasonable when compared to family income levels. (Cochran 346-7)

Note the ways in which the authors have understated the rising costs of public university education. First, the costs are deflated not by adjusting for the consumer price index but by median family income -- especially for the years after 1982, median family income rose much faster than the consumer price index. Second, graphing both the private and public data on the same graph enlarges the scale on which the public data is displayed. It's hard to tell from the graph, but between 1980 and 1995 it appears that public university costs increased from around 11% of family income to near 15% -- in effect the share of family income going to public university costs has increased by a third. The third way of minimizing the cost increases that have occurred since 1980 is to extend the time series back to 1965.

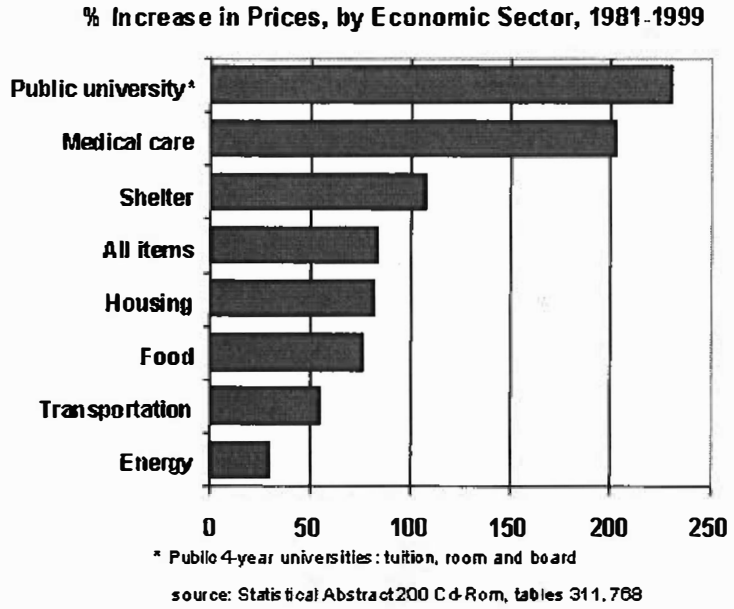


Figure 6. Bar chart

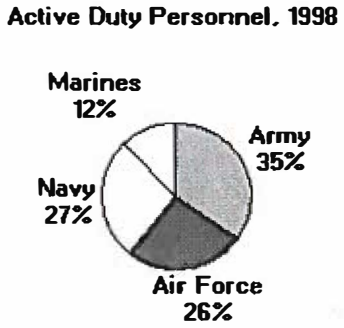
A completely different picture emerges if one were to compare the rate of increase in public university costs to the rate of increases in other sectors of the economy. On the left, we see that from 1981 to 1999 -- over the lifetime of today's college student -- public university costs have risen faster than any other sector of the economy. Faster even than rising medical care costs. In addressing the topic of health care inflation, the same authors note that: "Cost escalation in the medical field has been constant," and spend four pages of text addressing the reasons for the increases.

Data Distraction:

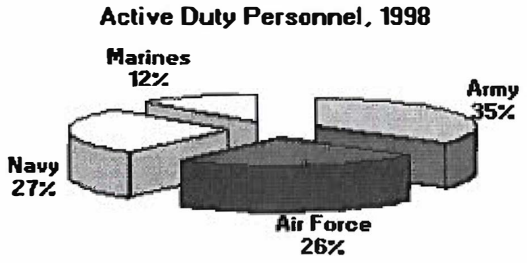
Edward Tufte's fundamental rule of efficient graphical design is to **minimize the ratio of ink-to-data**. This is essentially the same advice offered by Strunk and White to would be writers:

"A sentence should contain no unnecessary words, a paragraph no unnecessary sentences for the same reason that a drawing should contain no unnecessary lines and a machine no unnecessary parts."

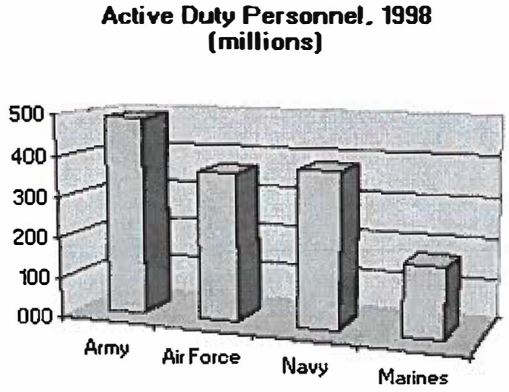
The primary source of extraneous lines in charting graphics today are the 3-D options offered by conventional spreadsheet graphics. These 3-D options serve no useful purpose; they add only ink to the chart, and more often than not make it more difficult to estimate the values represented. Even worse are the spreadsheet options that allow one to rotate the perspective. For those who would take bad graphical display to even higher levels, the Excel spreadsheet program offers the option of doughnut, radar, cylinder, cone, bubble charts.



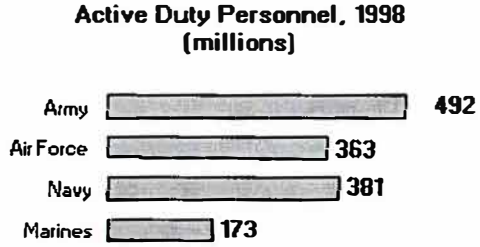
2-D Pie Chart



3-D Pie, Exploded



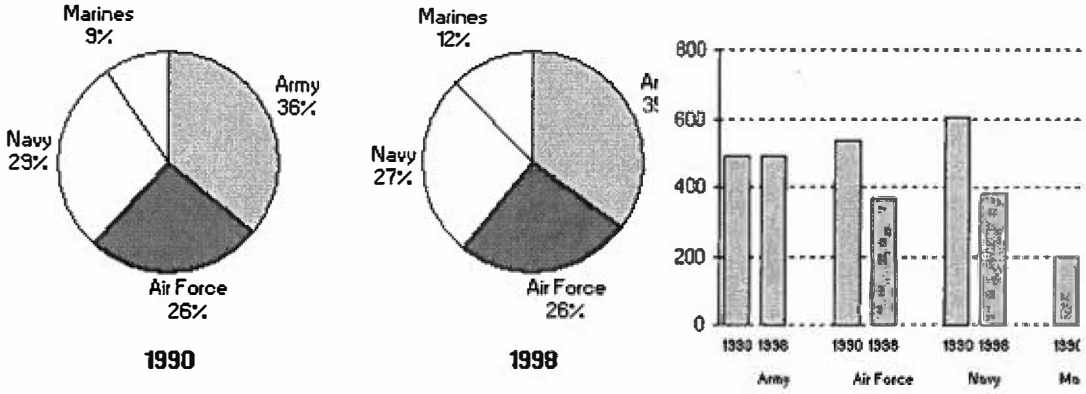
3-D Column Bar



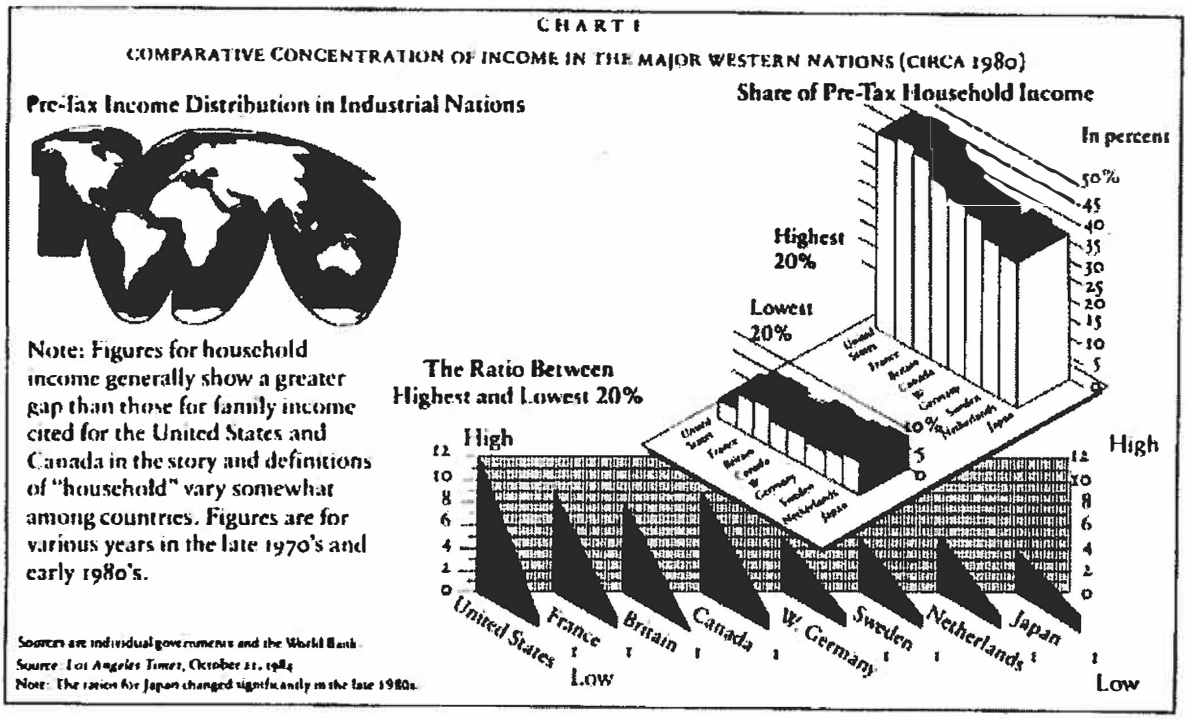
Simple 2-D Bar

Figure 7. Pie and bar charts

Pie charts should rarely be used. It is more difficult for the eye to discern the relative size of pie slices than it is to assess relative bar length. With the pie chart, without looking at the numbers, it is difficult to figure out whether the Navy or Air Force is larger; from the bar charts it is obvious. 3-D pie charts are even worse, as they also add a visual distortion (in this case, making the Air Force appear much larger). Note how much less ink the 2-D bar charts uses compared to the 3-D bar. Using data labels rather than a y-axis scale in this case reduces the number of numbers displayed from 6 to 4, and adds precision as well. Normally, I would have sorted the data here, so that the Navy would be between the Army and Air Force, but since the Marines are a part of the Navy (and the Air Force, originally, a part of the Army), this order made more sense. A strict application of the ink-to-data in this case, however, would eliminate the bars altogether and simply present the data as a table.



Pies are even less effective when an additional variable is added and comparisons between pies are required (sometimes by adjusting the relative size of the pies).



Not content with the distractions and distortions made possible by the use of 3-D effects, charters sometimes feel the need to add all sorts of other Chartjunk to a graph. In the graphics on the left, Kevin Phillips (1991, 9) is trying to make the point that income is more inequitably distributed in the United States than in other countries.

Note the extraneous features of this in this graphic.

- A completely irrelevant map of the world.
- Two entirely different kinds of 3-D charts displayed at two different perspectives.
- Country names are repeated three times.
- To display 24 numeric data points, 28 numbers are used to define the scales.
- The countries are sorted in no apparent order (not even alphabetically).
- Note the use of the letter " I " to separate the countries on the bottom chart.

While it might be possible to display these data better graphically, a table does the job quite nicely:

Pre-Tax income Distribution in Industrial Nations

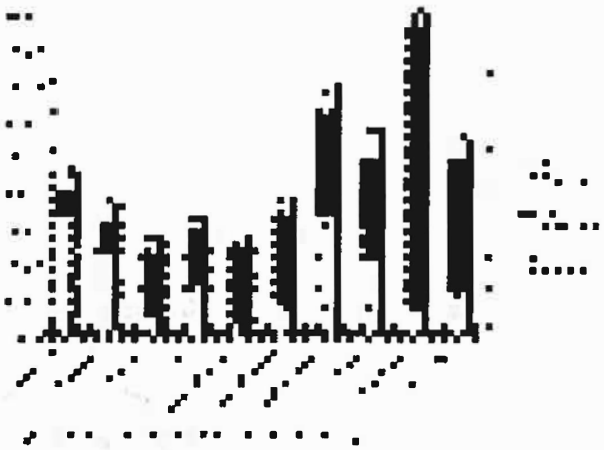
	Share of Pre-tax Household Income		Ratio:
	Top income quintile	Bottom income quintile	Top to bottom shares
United States	45	4	12
Canada	42	4	9
France	47	5	9
Britain	45	6	8
W. Germany	39	8	5
Sweden	38	8	5
Netherlands	37	7	5
Japan	36	9	4

*data estimated from chart.

More Chart-Junk.

Two chart types that should always be avoided.

Two common charts easily produced by spreadsheet programs that should almost always be avoided are the stacked bar chart and the pie chart. The stacked bar chart, made even worse by the use of 3-D effects in figure 3, makes it very difficult to estimate the values of the variables represented on the top of the bars. Similar "stacking" can also been done with time series area charts and should be avoided as well.

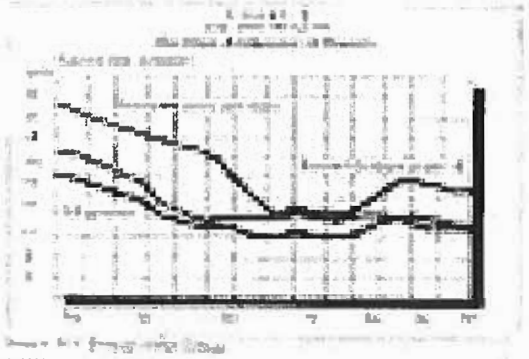


source: Putnam, p. 227

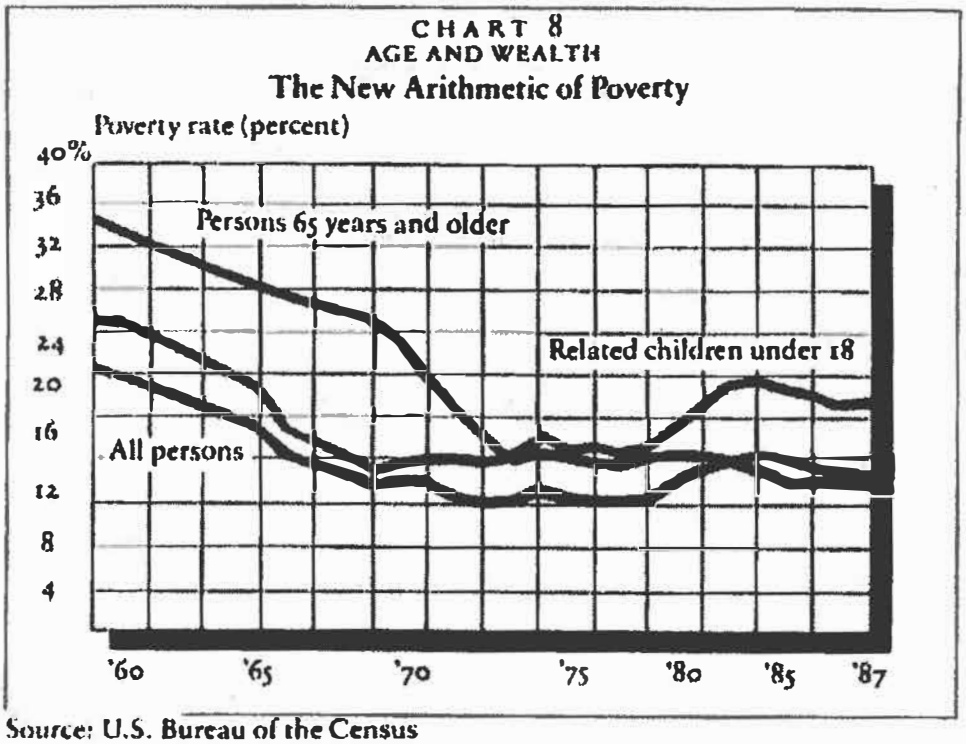
Figure 8: Stacked 3-D bar chart

Pie charts are fun to look at, but generally involve using a great deal of ink to display very little data. In addition, the charts often make it difficult to discern the exact magnitude of the size of the pie slices. Using multiple pie charts to display more than one variable is also a bad idea. All this is made even worse by exploiting the power of the spreadsheet technology to produce 3-D pie charts and "exploding" 3-D pie charts. If you think that you really must use a pie chart, make sure it is for data that does indeed at up to a total (i.e., the percentages for the slices add up to 100) and stay away from the fancy stuff.

Bad Chart 2: Where do the lines cross?



Phillips, p. 206



Source: U.S. Bureau of the Census

References:

Clarke Cochran et. al. *American Public Policy: An Introduction* (1999: St. Martin's Press)

Kevin Phillips, *The Politics of Rich and Poor* (1991: Harper Perennial)

Putnam, Robert D., *Bowling Alone* (Simon and Schuster, 2000)

Strunk, William Jr., and E. B. White, *The Elements of Style 3d edition* (MacMillan publishing, 1976).

1.3 Good Displays of the Data

We will discuss five methods of displaying data: pie chart, bar chart, stem and leaf (by hand only), histogram, and boxplot (by hand only). The displays should supply visual information to the viewer without a struggle.

Data Display	**Categorical**	**Quantitative: Continuous or Discrete**	**Comment**
	Pie Chart	*Stem & leaf*	
	Bar Chart	*Dot plot*	
		Histogram	
Concern	*Comparisons*	*Shape & Skewness*	Often overlay the distribution of interest over the histogram.

Displaying Categorical Data
PIE Chart

The *pie chart* is useful to show the division of a total quantity into component parts. A pie chart, if done correctly, is usually safe from misinterpretation. The total quantity, or 100%, is shown as the entire circle. Each wedge of the circle represents a component part of the total. These parts are usually labeled with *percentages* of the total. Thus, a pie chart helps us see what part of the whole each group forms.

Let's review percentages. Let a represent the partial amount and b represent the total amount. Then P represents a percentage calculated by $P=a/b\ (100)$.

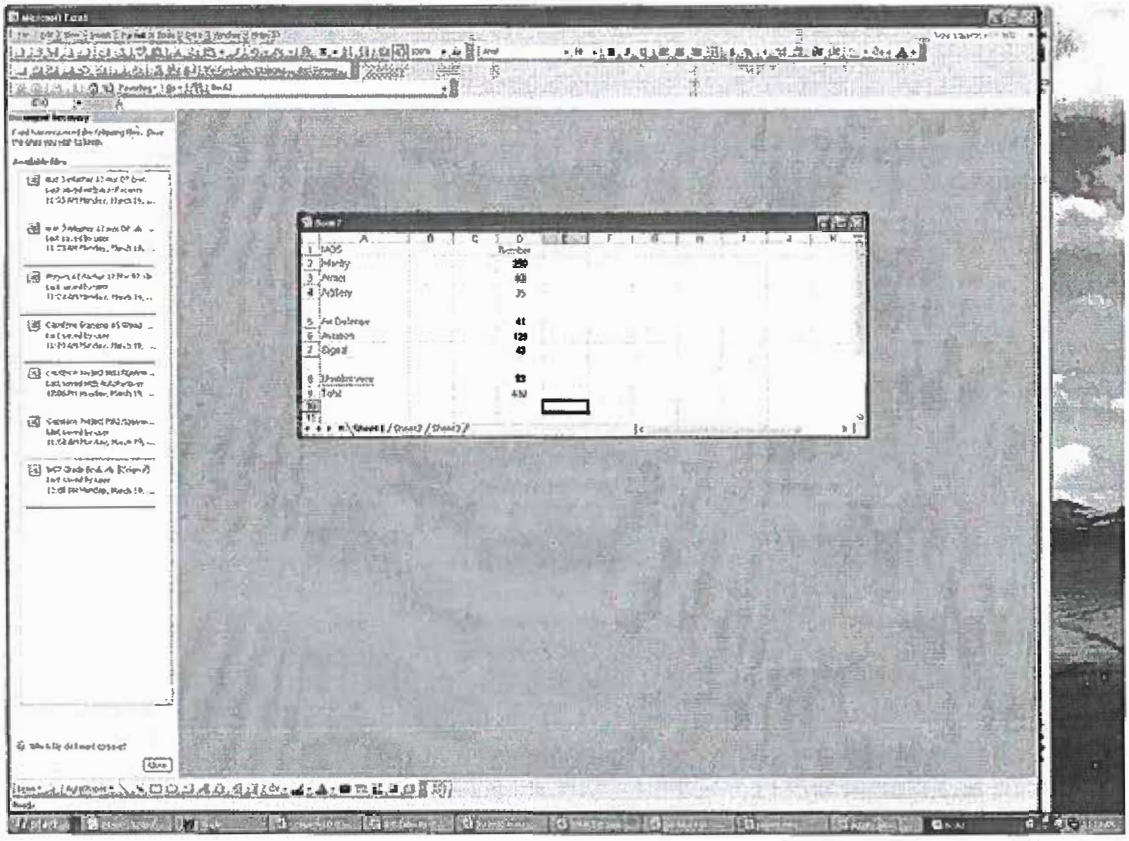
A percentage is thus a part of a whole. For example, \$0.25 is what part of \$1.00? We let $a=25$ and $b=100$. Then, $P= 25/100\ (100)= 25\%$.

Now, let's see how Excel would create a pie chart for us in the following scenario.

Consider soldiers choosing their MOS. Out of the 632 new soldiers recruited in SC that actually choose a MOS, the breakdown of selection is as follows.

Infantry	250
Armor	53
Artillery	35
Air Defense	41
Aviation	125
Signal	45
Maintenance	83
Total	632

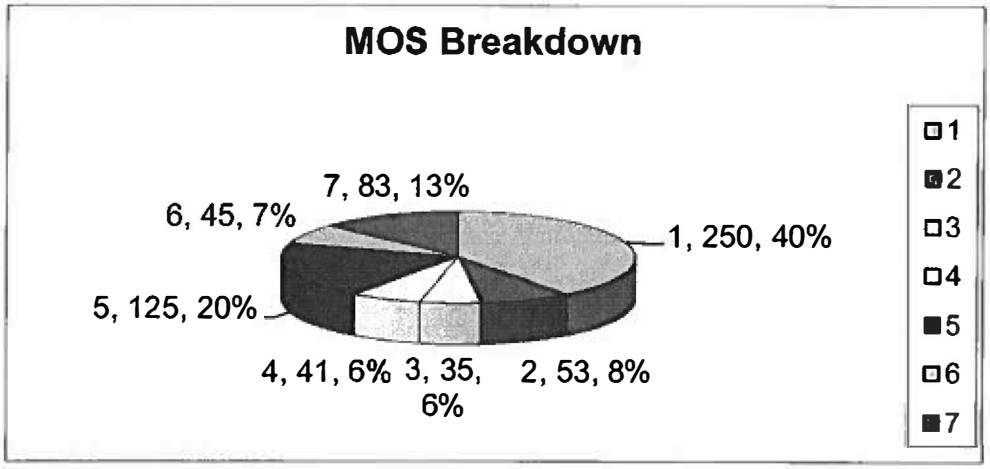
We begin by entering the data into labeled columns.



To select a pie chart, we highlight the labels and the numbers for the seven majors. We then click on Chart Wizard with the mouse, click on Pie, and follow editing directions.

In the Pie chart section, we select the type of chart to display. Select *No Legend*, show label and percent, and choose to display as a new sheet.

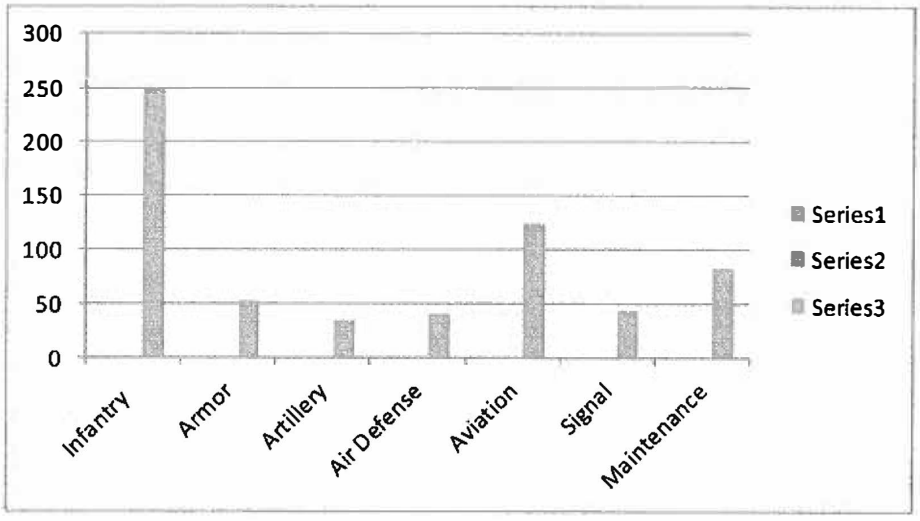
The output for the Pie Chart looks as follows:



Each of the shaded regions displays the percentage (%) of soldiers out of 632 that chose that MOS. Clearly Infantry has the largest percent of recruits. Which MOS appears to have the least?

What advantages and disadvantages can you see with using Pie Charts?

Let's view a bar chart:



Is this clearer to make your point that the pie chart?

ACTIVITY :

Given the following 1990 data from the U.S. Bureau of Census, Current Population Reports, p 385, in *The World Almanac 1993,* construct a pie chart. From the pie chart, rank order the age distribution of the U.S. population in 1990.

Category	Category #	Population
under 5	1	18,354,443
5 to 17	2	45,249,989
18 to 20	3	11,726,668
21 to 24	4	15,010,898
25 to 44	5	80,754,835
45 to 54	6	25,223,066
55 to 59	7	10,531,756
60 to 64	8	10,616,167
65 and up	9	31,241,831

Is this all the information that you would need to discuss population trends?

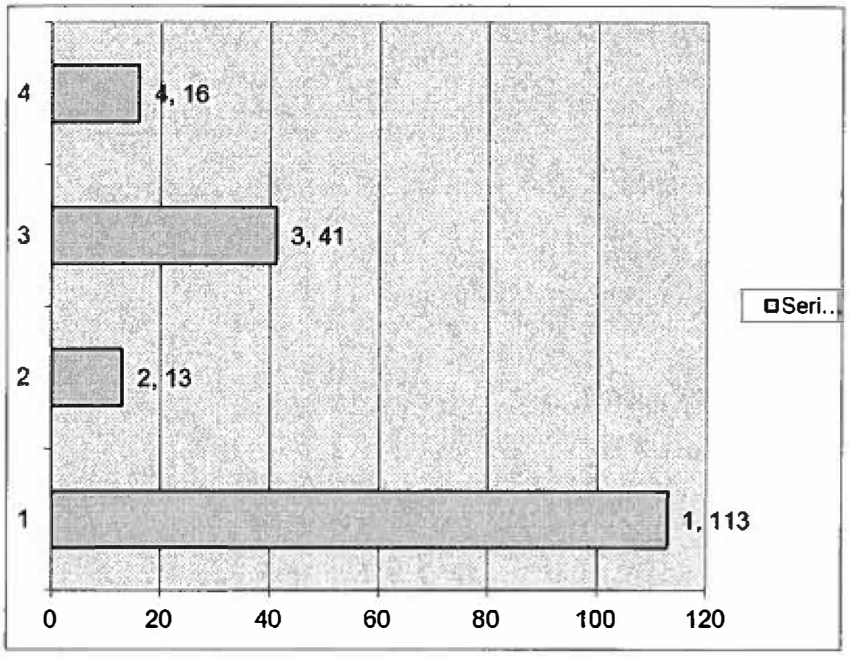
What other information would you like to have been given?

Bar Chart

Bar charts are useful when comparing relative sizes of data groups especially when they come from **categorical** variables. For example, consider the eye color from patients visiting the local eye clinic last year.

Eye Color	*Count*	*Percent*
Blue	*113*	*61.7486*
Green	*13*	*7.10383*
brown	*41*	*22.4044*
mixed	*16*	*8.74317*
Total	*183*	*100*

Click on Chart Wizard and obtain a Bar Chart.



You can quickly and clearly compare the relative sizes of the color groups. From the bar chart, which eye color occur the most frequently? Blue eyes occur the most frequently. It appears twice as large as the next most frequent - - brown eyes. Again, bar charts are most useful to display categorical data.

Displaying Quantitative Data

In quantitative data we are concerned with the shape of the data. Shape refers to symmetry of data." Is it symmetric?"" Is it skewed?" are questions we ask and answer.

Stem and Leaf

A stem-and Leaf plot uses the real data points in making a plot. The plot will appear strange because your plot is sideways. The rules are as follows:

Step 1:	Order the data
Step 2:	Separate according to the one or more leading digits. List stems in a vertical column.
Step 3:	Leading digit is the stem and trailing digit is the leaf. For example 32, 3 is the stem and 2 is the leaf. Separate the stem from the leafs by a vertical line.
Step 4.	Indicate the units for stems and leafs in the display.

You will probably create these plots by hand (***Excel*** will not produce a stem and leaf plot).

Example: Grades for 20 students in a course

53, 55, 66, 69, 71, 78, 75, 79, 77, 75, 76, 73, 82, 83, 85, 74, 90, 92, 95, 99

Stems are the leading digit:

5
6
7
8
9

Standing for 50-s, 60's, 70's, 80's, and 90's.

If there had been a score of 100, then the leading digit is in 100's. So we would need:

05
06
07
08
09
10

for 50's, 60's, 70's, 80's, 90's, and 100's

Draw a vertical line after each stem.

5|
6|
7|
8|
9|

Now add the leafs, which are the trailing digits,

53, 55, 66, 69, 71, 73, 74, 75, 75, 76, 77, 78,79, 82, 83, 85, 90, 92, 95, 99

5| 3, 5
6| 6, 9
7| 1, 3, 4, 5, 5, 6, 7, 8, 9
8| 2, 3, 5
9| 0, 2, 5, 9

We can characterize this shape as almost *symmetric*. Note how we read the values from the stem-and-leaf.

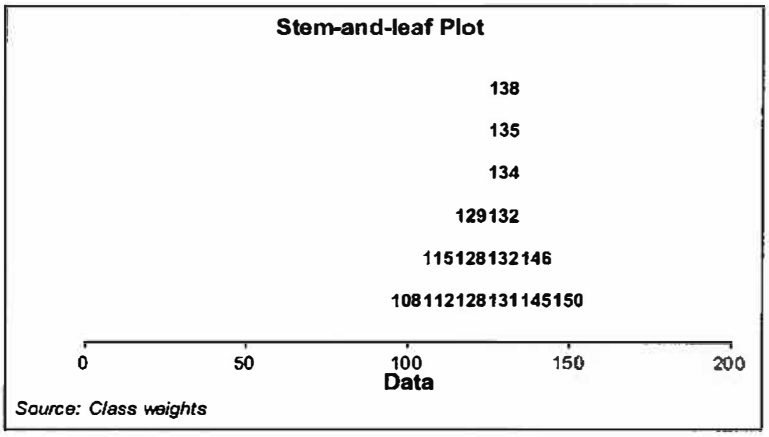
For example, we read
5| 3, 5
as data elements 53 and 55.

We have a program in Excel that gives us a stem & leaf plot. It is still up to us to determine the shape. Here is an example with class weights.

Back

Increment	10
Data	
135	
132	
112	
115	
128	
132	
145	
129	
134	
138	
131	
128	
150	
146	
108	

Title	Stem-and-leaf Plot
Source	*Class weights*

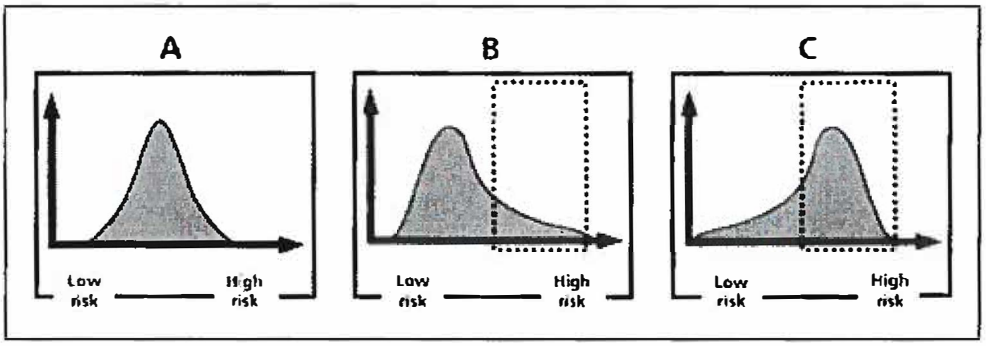


We might accept this as symmetric.

SYMMETRY ISSUES

We look at these shapes as symmetric or skewed. Symmetric looks like a bell shaped curve while skewed means that the plot appears lopsided.

Three generic risk distribution shapes (symmetric, skewed right and skewed left)



Note: The shape of the distribution has important implications from a risk management standpoint. In the above Figure (A), the risk distribution is symmetric, and as a result, there are an equal number of people experiencing a high risk as there are a low risk. In Figure (B), the risk distribution is skewed to the right, with most people experiencing a low risk and a few experiencing a high risk, compared to Figure (C), where the distribution is skewed to the left translating to many people experiencing a higher risk and only a few people experiencing a lower risk. From a risk management or policy perspective, each of these situations would need to be assessed differently in light of the following considerations: the population (children, elderly, etc.) experiencing the higher risk; what the actual magnitude of higher risk is (high risk as defined in this context may not be very high when compared to other competing risks), if the higher risk is being borne as a result of voluntary or involuntary actions, and whether the people bearing the higher risk are in control of the risk situation, etc.

A key aspect of the risk characterization stage is providing insight not only into the risk estimates, but also our confidence in the generated assessment. Such insights include:

- The steps that could be taken to reduce the risk;
- Points in the process about which we have uncertainty and could benefit from more information;
- Points that have a significant impact on the risk, and as such would be ideal areas to focus more attention on so as to ensure they are under control.

In general, quantitative risk assessment models can be considered as contributing towards risk management decision-making by providing input along four avenues:

- focusing attention on risk reducing areas;
- focusing attention on research areas;
- helping in the formulation of risk reduction strategies;
- providing a tool to test out formulated risk reduction strategies prior to implementation.

1.3 Exercises.

Make a stem and leaf of the following data sets and comment about the shape.

1. 100, 105, 111, 115, 121, 129, 131, 131, 133, 135, 137, 145, 146, 150, 160, 180
2. .10 , .15, .22, .23, .50, .62 , .62, .65, .66, .69, .72
3. 63, 65, 72, 81, 83, 85, 92, 93, 94, 105, 106, 121, 135

1.4 Displaying quantitative data with Histograms

We begin by stating that there is a difference between bar charts and histograms. Bar charts have discrete values as their horizontal axis. Thus, bars are centered at discrete values. A histogram has continuous values as its horizontal axis. Thus, there are no spaces between the bars unless no data in in that range. Since most of the data that you will use are large, we will go quickly to displays with technology.

Steps:

(1) Obtain descriptive statistics for the data or order the data smaller to larger
(2) Determine the Interval [smallest, largest]
(3) Calculate the class intervals (largest-smallest)/n where n is the number of intervals desired. The value of n must be between 5 and 20. Start with 5 and go up until a good view of the histogram is obtained.
(4) List the endpoints as Bin values
(5) Go to Data Analysis, Histogram and bring up dialog box. Put data in data input and endpoints in bins.
(6) The output is a table.
(7) Highlight the frequencies of the table and go to insert Bar chart
(8) Right click in bar char (on a bar) and close GAP size to 0.
(9) Comment on shape in regard to symmetry and skewness.

Histograms of data series can be created using the Analysis ToolPak's Histogram tool. Data is grouped into intervals (known as bins) and the number of observations that fall into each are displayed both in a table and, also graphically, as a bar chart. We must edit the bar chart so that the gap width is 0 to be a histogram.

Example

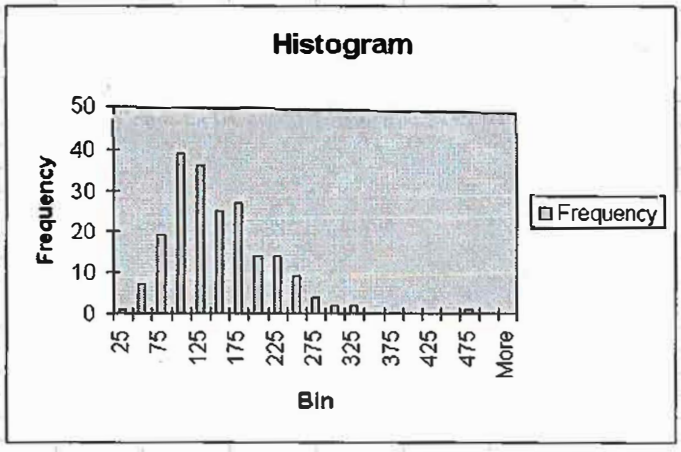
Using the potato data select **Tools > Data Analysis... > Histogram...** from the menu bar and a dialog box will appear. Insert the input range by either entering the reference by keyboard or highlighting the input range on the worksheet. Tick the **label** box if labels are included in the input range.

Set up the ranges for the histogram divisions (e.g. a column of numbers starting at 50 and going up in steps of 50) and enter the reference for the bin range. If cells E2=50,E3=100,...E11=500, then the reference given is $E$2:$E$11. If this box is left empty Excel will generate a default range. Define where you want the output to appear by entering the cell reference of the top left corner of where you want it to go.

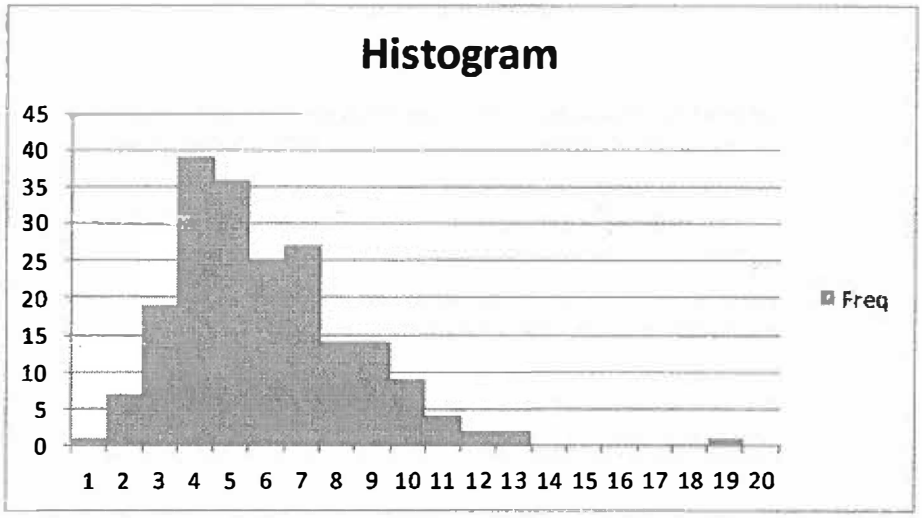
Tick the chart output box to obtain the histogram and click on the OK button when you have finished. The histogram for the variable **weight** could look like the one on the right, after you have stretched it vertically.

Below is the output generated by the Histogram tool for the **weight** data using a step-size of 25 instead of 50 as in the previous graph.

Bin	Frequency
25	1
50	7
75	19
100	39
125	36
150	25
175	27
200	14
225	14
250	9
275	4
300	2
325	2
350	0
375	0
400	0
425	0
450	0
475	1
500	0
More	0



We must close the gap width to 0 for a histogram:



Our examination shows the data is skewed right.

Histogram

Using the Histogram tool will allow us to make a histogram and create a frequency distribution chart at the same time.

1. Click Tools > Data Analysis.... If you don't have a Data Analysis option under Tools, see step #3 of "How do I get started with Excel?".

2. In the Data Analysis window, select Histogram and click OK.

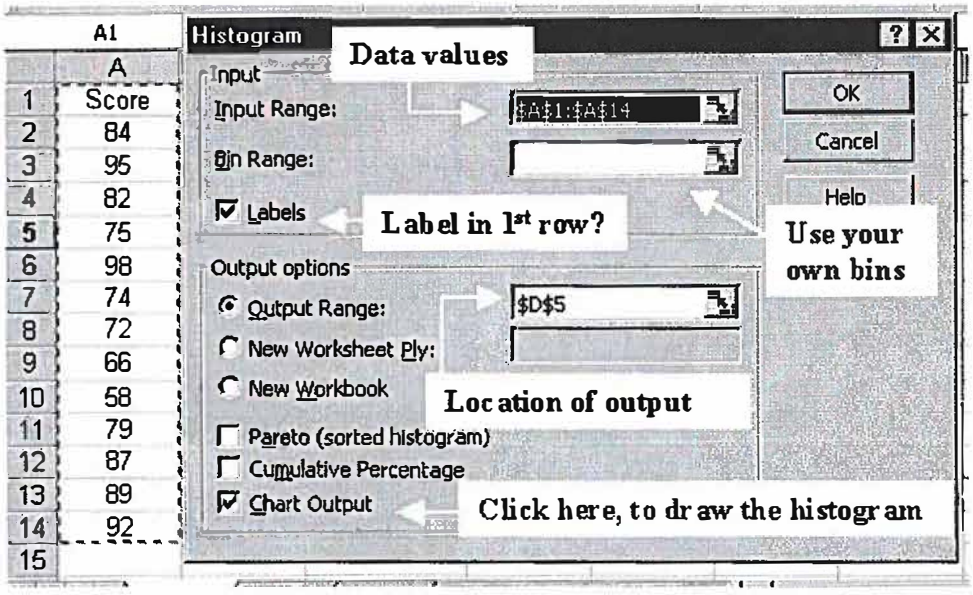
3. A new window titled Histogram should appear. This window has many options. Below is a brief explanation of each:

    - Input Range is where the data being used to create the histogram goes. Simply put your cursor back into the spreadsheet and highlight the variable name and all the data in that column.
    - More information about Bin Range.
    - Click Labels. If a variable name was highlighted in the Input Range, then this needs to be checked.
    - You must select one of the following Output options:
        - Click Output Range if you want the histogram to be placed on the current sheet. Next, simply input the cell where you want the output to be placed.
        - Click New Worksheet Ply if you want the histogram to be placed on a new sheet. Next, type the name of the new sheet where you want the output to be placed.
    - Clicking Cumulative Percentage will list the cumulative percentage for each class and include a cumulative percentage line on your histogram.

- Click Chart Output under Output options. This step is ***necessary*** to obtain the histogram. If this is not highlighted you will only receive a frequency distribution chart.

4. Click OK. The histogram and frequency distribution chart should be placed onto your spreadsheet.

An example:



We will present the information on how to construct a histogram using EXCEL.

Histogram:

Step 1.	Determine and select the classes, 5-15 classes. Find the range (lowest to highest value). Classes should be evenly spaced if possible.
Step 2.	Tally the data in the classes.
Step 3.	Find the numerical (relative) frequencies from the tallies.
Step 4.	Find the cumulative frequencies.

Histogram: connects class interval as a base and tallies (or relative frequencies) as the height of a rectangle. Rectangle is centered at the mid-point of class interval.

53, 55, 66, 69, 71, 73, 74, 75, 75, 76, 77, 78,79, 82, 83, 85, 90, 92, 95, 99

Possible class intervals:

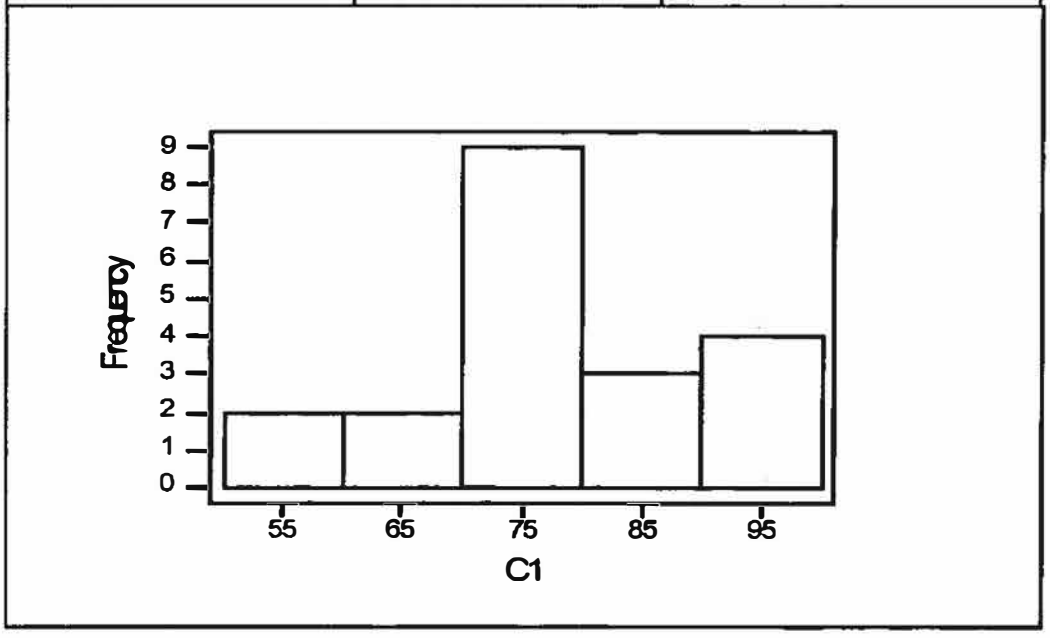
(a) Classes 51-60,61-70,71-80,81-90, 91- 100
(5 classes intervals)

(b) Classes 50-59, 60-69,70-79, 80-89, 90-99, 100-109
(6 classes intervals)

(c) Classes 51-55, 56-60, 61-65, 66-70, 71-75, 76-80, 81-85, 86-90, 91-95, 96-100
(10 class intervals)

Let's use selection (a)

Interval	Tally	Decimal
51 - 60	2	2/20 = .10
61 - 70	2	2/20 = .10
71 - 80	9	9/20 = .45
81 - 90	4	4/20 = .2
91 - 100	3	3/20 = .15
Total	20	20/20 = 1.00



We note the data is somewhat symmetric.

1.4 Exercises.

Make a histogram using at least 5 class intervals of the following data sets and comment about the shape. Use a graphing calculator as well as doing these by hand.

1. 100, 105, 111, 115, 121, 129, 131, 131, 133, 135, 137, 145, 146, 150, 160, 180

2. 0.10 ,0 .15, 0.22, 0.23, 0.50, 0.62 , 0.62, 0.65, 0.66, 0.69, 0.72

3. 63, 65, 72, 81, 83, 85, 92, 93, 94, 105, 106, 121, 135

1. 5 Boxplots for comparisons

We will present the information on how to construct and use a boxplot. I believe a have an Excel program to do boxplots. Boxplots are a good way to compare data sets from multiple sources. For example, let's look at violence in a 10 regions in Afghanistan. Putting the 10 boxplots together allows us to compare many aspects such as medians, ranges, and dispersions.

Boxplot

Step 1.	Draw a horizontal measurement scale that includes all data within the range of data.
Step 2.	Construct a rectangle (the box) whose left edge is the lower quartile value and whose right edge is the upper quartile value.
Step 3.	Draw a vertical line segment in the box for the median value.
Step 4.	Extend line segments from rectangle to the smallest and largest data values (these are called whiskers).

53, 55, 66, 69, 71, 73, 74, 75, 75, 76, 77, 78, 79, 82, 83, 85, 90, 92, 95, 99

The values are in numerical order. What is needed are the range, the quartiles, and the median.

Range is the smallest and largest values from the data: 53 and 99.

The median is the middle value. It is the average of the 10th and 11th values as we will see later: *(76 + 77)/ 2 = 76.5*

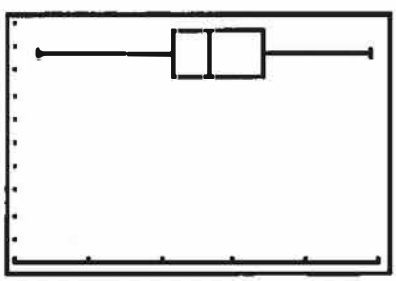
The quartiles values are the median of the lower and upper half of the data.

Lower quartile values: *53, 55, 66, 69, 71, 73, 74, 75, 75, 76.* Its median is 72.

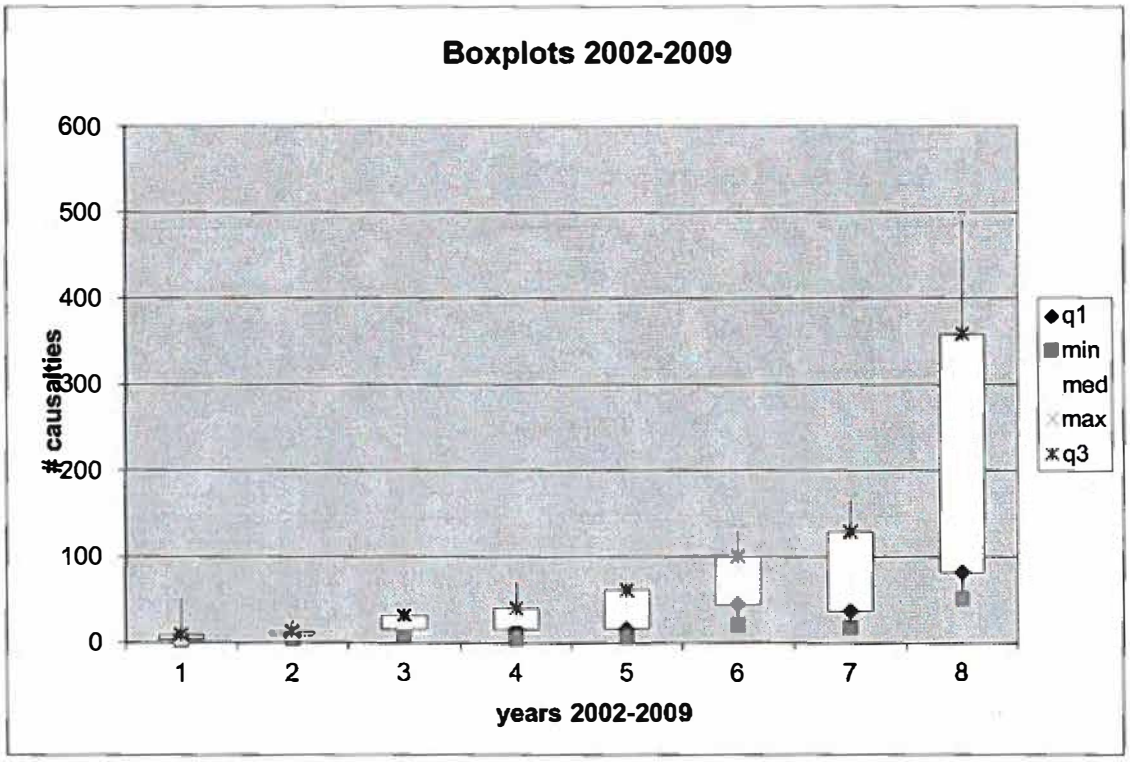
Upper quartile values: *77, 78,79, 82, 83, 85, 90, 92, 95, 99.* Its median is 84.

You draw a rectangle from 72 to 84 with a vertical line at 76.5

Then draw a whisker to the left to 53 and to the right to 99.
It would look something like this:



Comparisons;
Consider our data for casualties in Afghanistan through the years 2002-2009. This is presented to you as a commander. What information can you interpret from this boxplot?



1.5 Exercises.

Make a boxplot of each of the following data sets and comment about the shape.

1. 100, 105, 111, 115, 121, 129, 131, 131, 133, 135, 137, 145, 146, 150, 160, 180

2. .10, .15, .22, .23, .50, .62 , .62, .65, .66, .69, .72

3. 63, 65, 72, 81, 83, 85, 92, 93, 94, 105, 106, 121, 135

SUMMARY: Displays of Data

Categorical Displays

Pie Chart:

The circle (pie) represents the whole or 100%. The wedges or pieces of the pie represent the proportion or part of the total for that category.

Bar Chart:

Bars can be horizontal or vertical, but they should be uniformly spaced and of the same width. The lengths of the bars represent the values of the categories we wish to compare.

Quantitative Displays

Stem and Leaf:

Step 1. Order the data

Step 2. Separate according to the one or more leading digits. List stems in a vertical column.

Step 3. Leading digit is the stem and trailing digit is the leaf. For example 32, 3 is the stem and 2 is the leaf. Separate the stem from the leafs by a vertical line.

Step 4. Indicate the units for stems and leafs in the display.

Histogram:

Step 1. Determine and select the classes, 5-15 classes. Find the range (lowest to highest value). Classes should be evenly spaced if possible.

Step 2. Tally the data in the classes.

Step 3. Find the numerical (relative) frequencies from the tallies.

Step 4. Find the cumulative frequencies.

Histogram: connects class intervals as a base and tallies (or relative frequencies) as the height of a rectangle. The rectangles are centered at the mid-point of each class interval.

Boxplot:

Step 1. Draw a horizontal measurement scale that includes all data within the range of data.

Step 2. Construct a rectangle whose left edge is the lower quartile value and whose right edge is the upper quartile value.

Step 3. Draw a vertical line segment in the box for the median value.

Step 4. Extend line segments from rectangle to the smallest and largest data values (these are called whiskers).

Summary Exercises:

1. Make a pie chart for the following data:

Never Married	43.9 million
Married	116.7 million
Widowed	13.4 million
Divorced	17.6 million

What information is best displayed by a pie chart?

2. Make a bar chart for this data: female doctorates as a percent of graduates in that field that were females:

Computer Science	15.4%
Education	60.8%
Engineering	11.1%
Life Sciences	40.7%
Physical Sciences	21.7%
Psychology	62.2%
Mathematics	10%

Can you make a pie chart? What do you have to do first?

3. Display the following data: In 1995 there were 90,402 deaths from accidents in the US. Among these there were 43,363 from motor vehicles, 10,483 from falls, 9072 from poisoning, 4350 from drowning, and 4235 from fires. How many deaths were due to other unknown causes?

4. In Math I class last semester the final averages were: 88, 63, 82, 98, 89, 72, 86, and 70. Display the data as a stem and leaf. Are they symmetric? Are they skewed?

5. In Math II class last semester, the final averages were: 66, 61, 78, 54, 75, 40, 78, 91, 84, 82, 76, and 65. Display the data as a histogram. Are they symmetric? Are they skewed?

6. Using the grades in Math I (#4) and Math II (#5), display the data as two boxplots side by side. Is each display symmetric? Is each display skewed? Can you compare the two data sets? Which class had the higher grades? Which class has grades that are more spread out? Does the symmetry and skewness of the data sets tell us anything about the grades?

Lesson 2-3

Statistical Measures

Lesson Objectives

1. Learn how to use Excel to obtain and interpret descriptive statistics
2. Learn how to use Excel to obtain displays and be able to interpret them.
3. Know what the measures are and which ones are used for quantitative data and which one are better for qualitative data.

Measures of Central Tendency or Location

Describing the Data

In addition to plots and tables, numerical descriptors are often used to summarize data. Three numerical descriptors, the *mean*, the *median*, and the *mode* offer different ways to describe and compare data sets. These are generally known as the *measures of location.*

The *MEAN.*

The mean is the arithmetic average, with which you are probably very familiar. For example, your academic average in a course is generally the arithmetic average of your graded work. The mean of a data set is found by summing all the data and dividing this sum by the number of data elements.

The following data represent ten scores earned by a student in a college algebra course: *55, 75, 92, 83, 99, 62, 77, 89, 91, 72.*

Compute the student's average.

The mean can be found by summing the 10 scores

$$55 + 75 + 92 + 83 + 99 + 62 + 77 + 89 + 91 + 72 = 795$$

and then dividing by the number of data elements (10), $795/10 = 79.5$

To describe this process in general we can represent each data element by a letter with a numerical subscript. Thus, for a class of n tests, the scores can be represented by $a_1, a_2, \ldots, a_n$. The mean of these n values of $a_1, a_2, \ldots, a_n$ is found by adding these values and then dividing this sum by n, the number of values. The Greek letter Σ (called sigma) is used to represent the sum of all the terms in a certain group. Thus, we may see this

written as $\sum_{i=1}^{n} a_i = a_1 + a_2 + \ldots + a_n$ mean $= \dfrac{\sum_{i=1}^{n} a_i}{n}$

Think of the mean as the average. Notice that the mean does not have to equal any specific value of the original data set. The mean value of 79.5 was not a score ever earned by our student.

Batting average is the total number of hits divided by the total number of official at bats. Is batting average a mean? Explain.

THE *MEDIAN*.

The median locates the true middle of a numerically ordered list. The hint here is that You need to make sure that your data is in numerical order listed from smallest to largest along the *x*- number line. There are two ways to find the median (or middle value of an ordered list) depending on n (the number of data elements):

(1) if there is an odd number of data elements then the middle (median) is the exact data element that is the middle value. For example, here are 5 ordered math grades earned by a student: *55, 63, 76, 84, 88.*

The middle value is 76 since there are exactly two scores on each side (lower and higher) of 76. Notice that with an odd number of values that the median is a real data element.

(2) if there is a even number of data elements then there is no true middle value within the data itself. In this case, we need to find the mean of the two middle numbers in the ordered list. This value, probably not a value of the data set, is reported as the median. Let's illustrate with several examples.

(a) Here are 6 math scores for student one: *56, 62, 75, 77, 82, 85*

The middle two scores are 75 and 77, because there are exactly two scores below 75 and exactly 2 scores above 77. We average 75 and 77. *(75+77) / 2 = 152/2 = 76*

76 is the median. Note that 76 is not one of the original data values.

(b) Here are 8 scores for student two: *72, 80, 81, 84, 84, 87, 88, 89*

The middle two scores are 84 and 84, because there are exactly 3 scores lower than 84 and 3 scores higher than 84. The average of these two scores is 84. Note that this median is one of our data elements.

It is also very possible for the mean to be equal to the median.

THE *MODE*.

The value that occurs the most often is called the mode. It is one of the numbers in our original data. The mode does not have to be unique. It is possible for there to be more than one mode in a data set. As a matter of fact, if every data element is different from the other data elements then every element is a mode.

For example, consider the following data scores for a mathematics class.
75, 80, 80, 80, 80, 85, 85, 90, 90, 100

The number of occurrences for each value is:

Value	Number of Occurrences
75	1
80	4
85	2
90	2
100	1

Since 80 occurred 4 times and that is the largest value among the number of occurrences, then 80 is the mode.

Measures of Dispersion

Variance and Standard Deviation

Measures of variation or *measures of the spread* of the data include the variance and standard deviation. They measure the spread in the data, how far the data are from the mean. The sample variance has notation S^2 and the sample deviation has notation S.

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$ where n is the number of data elements .

$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$ where n is the number of data elements.

EXAMPLE 1: Consider the following 10 data elements:

50, 54, 59, 63, 65, 68, 69, 72, 90, 90.

The mean, $\bar{x}$, is 68. The variance is found by subtracting the mean, 68, from each point, squaring them, add them up, and divide by *n-1*.

$$S^2 = [(50-68)^2 + (54-68)^2 + (59-68)^2 + (63-68)^2 + (65-68)^2 + (68-68)^2 + (69-68)^2 + (72-68)^2 + (90-68)^2 + (90-68)^2]/9 = 180$$

$$S = \sqrt{S^2} = 13.42.$$

EXAMPLE 2: Consider a person's metabolic rate at which the body consumes energy. Here are 7 metabolic rates for men who took part in a study of dieting. The units are calories in a 24-hour period.

1792 1666 1362 1614 1460 1867 1439

The researchers reported both $\bar{x}$ and S for these men.

The mean:

$$\bar{x} = \frac{1792+1666+1362+1614+1460+1867+1439}{7} = \frac{11,200}{7} = 1600$$

To see clearly the nature of the variance, start with a table of the deviations of the observations from the mean.

Observations	Deviations	Squared Deviations
X_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1792	1792-1600=192	36,864
1666	1666-1600=66	4,356
1362	1362-1600=-238	56,644
1614	1614-1600=14	196
1460	1460-1600=-140	19,600
1867	1867-1600=267	71,289
1439	1439-1600=-161	25,921
	Sum = 0	Sum=214,870

The variance, $S^2 = 214{,}870 / 6 = 35,811.67$

The standard deviation, $S = \sqrt{35{,}811.67} = 189.24$

Some properties of the standard deviation are:

- S measures spread about the mean.
- S = *0* only when there is no spread.
- S is strongly influenced by extreme outliers.

Measures of Symmetry or Skewness

We define a measure, the coefficient of skewness, S_k. Mathematically, we determine this value from formula:

$$S_k = \frac{3 \cdot (\bar{X} - \tilde{X})}{S}$$

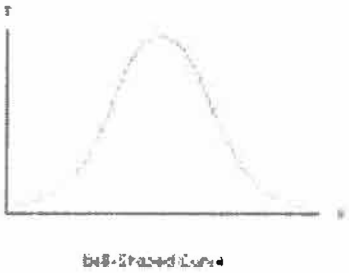
We use the following rules for skewness and symmetry.

If Sk ≈ 0, the data is symmetric.

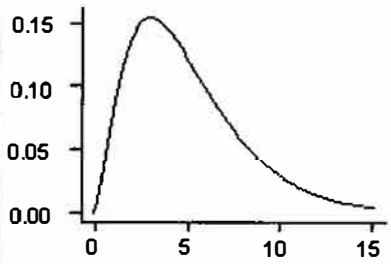
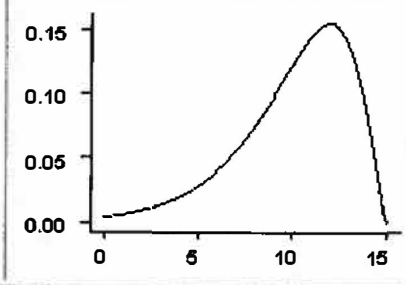
If Sk > 0 the data is positively skewed (skewed right).

If Sk < 0, the data is negatively skewed (skewed left).

We use the bell-shaped curve to denote symmetry. The following figures illustrate:



(a) Symmetric

An example of positive skewness (skew to the right)	An example of negative skewness (skew to the left)
	

(b) Skewness

Range is a measure that takes the maximum and minimum values of the data. Often, this is provided a single number. Assume we have the following data:

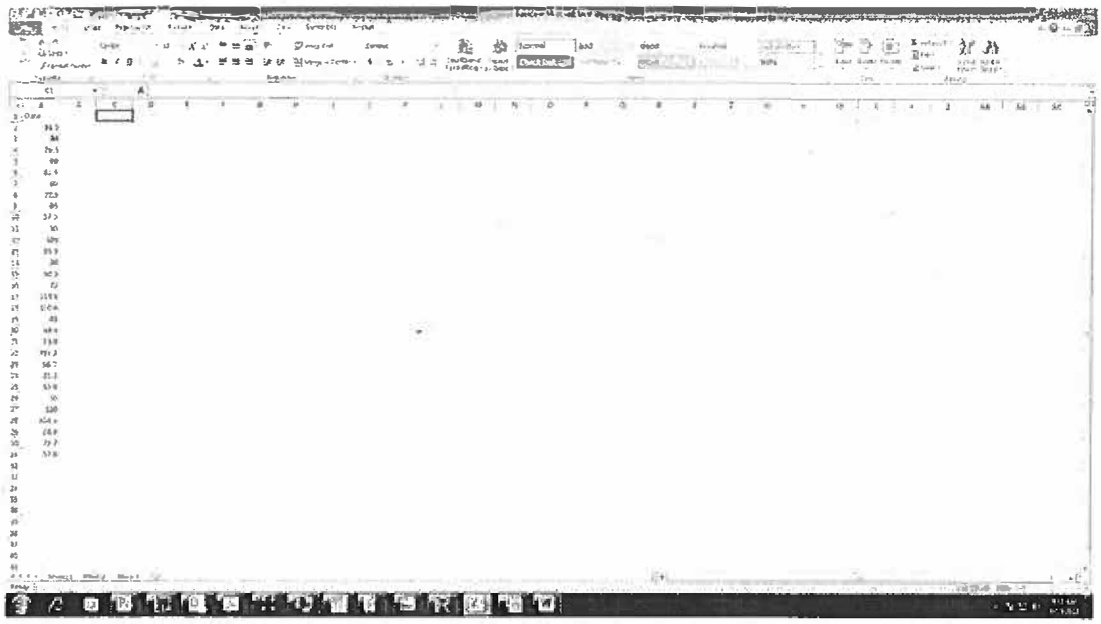
1792
1666
1362
1614
1460
1867
1439

The maximum value is 1867 and the minimum value is 1362. If you take the difference, 1867-1362=505. What does 505 represent? I suggest you give the range as an interval [1362, 1867].

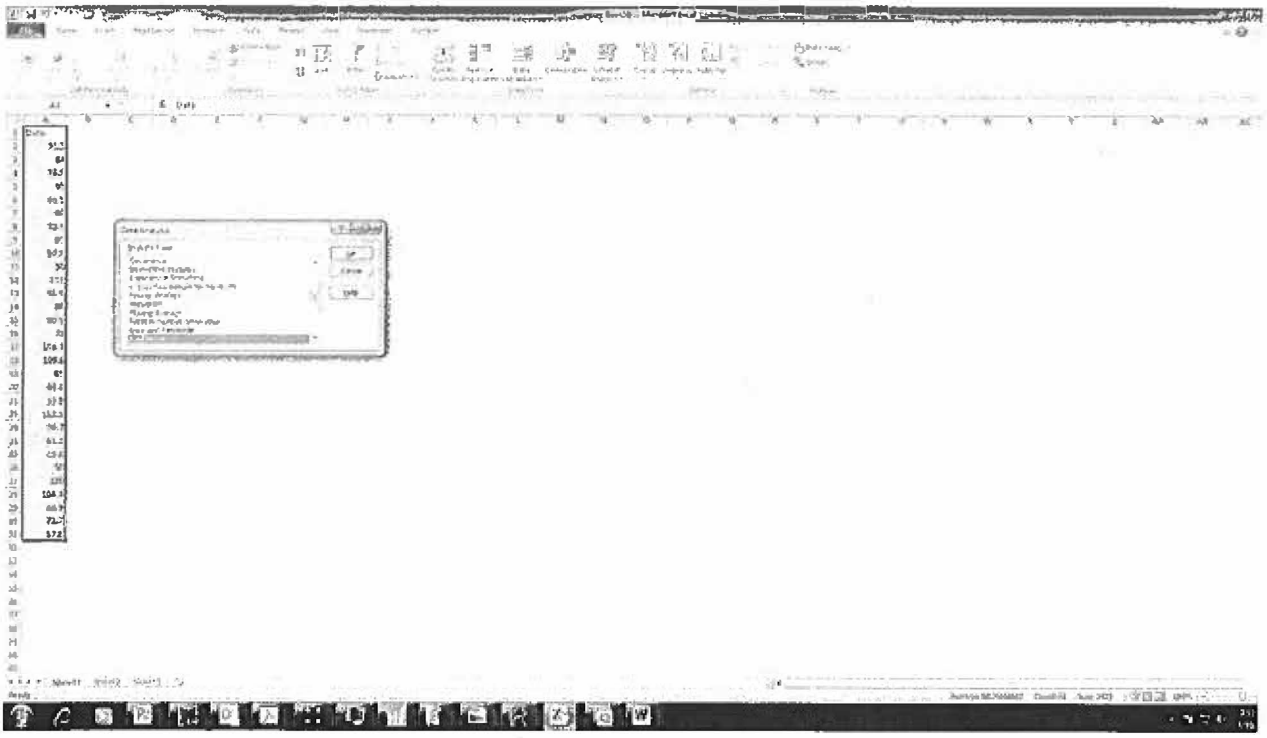
Summary with Excel

We can obtain all this information quickly in Excel. Given a list of data we can use the Data Analysis package, Descriptive Statistics, Summary Statistics option.
We first enter the data in a column in Excel.

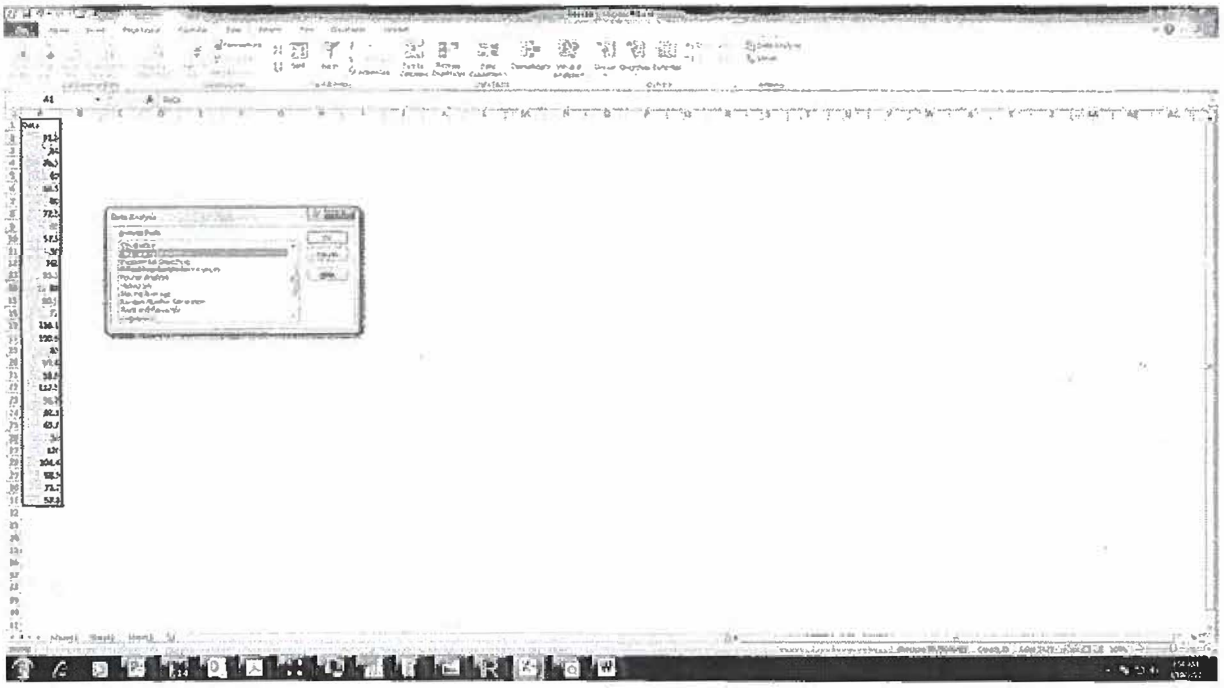
Data
91.5
84
76.5
69
61.5
80
72.5
65
57.5
50
103
95.5
88
80.5
73
116.1
100.6
85
69.4
53.9
112.3
96.7
81.1
65.6
50
120
104.4
88.9
73.7
57.8



Then go to Data→Data Analysis

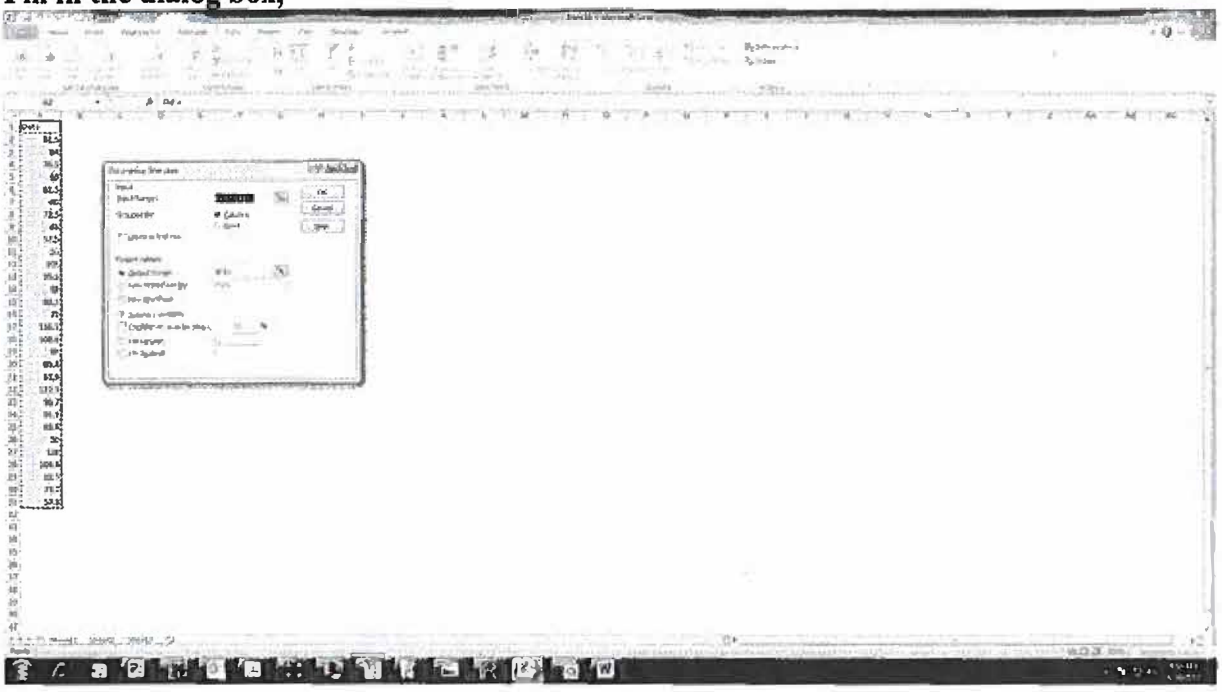


Highlight Descriptive Statistics



Press OK

Fill in the dialog box,



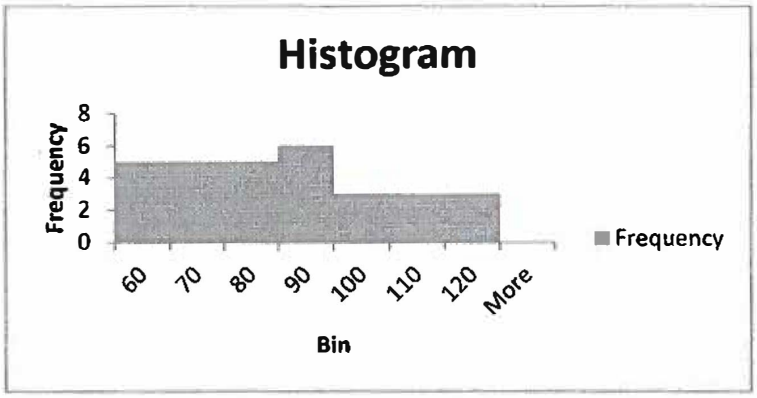
Press OK.

View the output:

Extracted the descriptive statistics are:

	Column1
Mean	80.76666667
Standard Error	3.549341313
Median	80.25
Mode	50
Standard Deviation	19.44054301
Sample Variance	377.9347126
Kurtosis	-0.681118616
Skewness	0.281013454
Range	70
Minimum	50
Maximum	120
Sum	2423
Count	30

Now, we are tempted to discuss this but I suggest that we need one more piece of information---the display. I will create a histogram first. Note the Range is from 50 to 120 or 70 units. I will create the following Bins: 60, 70,80,90,100,110,120.



The histogram appears somewhat symmetric. The values in the descriptive statsitcs support this.

Now let's sumamrize the interpretations of the statsitics.

Definition of 'Skewness'
Describe asymmetry from the normal distribution in a set of statistical data. Skewness can come in the form of "negative skewness" or "positive skewness", depending on whether data points are skewed to the left (negative skew) or to the right (positive skew) of the data average.

Definition of 'Standard Deviation'
A measure of the dispersion of a set of data from its mean. The more spread apart the data, the higher the deviation. Standard deviation is calculated as the square root of variance.

Definition of 'Variance'
A measure of the dispersion of a set of data points around their mean value. Variance is a mathematical expectation of the average squared deviations from the mean.

Definition of 'Mean'
The simple mathematical average of a set of two or more numbers. The mean for a given set of numbers can be computed in more than one way, including the arithmetic mean method, which uses the sum of the numbers in the series, and the geometric mean method. However, all of the primary methods for computing a simple average of a normal number series produce the same approximate result most of the time

Definition of 'Median'
The middle number in a sorted list of numbers. To determine the median value in a sequence of numbers, the numbers must first be arranged in value order from lowest to highest. If there is an odd amount of numbers, the median value is the number that is in the middle, with the same amount of numbers below and above. If there is an even amount of numbers in the list, the middle pair must be determined, added together and

divided by two to find the median value. The median can be used to determine an approximate average.

Definition of 'Mode'

A statistical term that refers to the most frequently occurring number found in a set of numbers. The mode is found by collecting and organizing the data in order to count the frequency of each result. The result with the highest occurrences is the mode of the set.

Statistics and Measure Summary

Type of Data	Best measures of Location	Best measures of Dispersion
Quantitative	Mean, median, mode	Range, Variance, standard deviation, skewness
Qualitative	Median, Mode	Range

Exercises:

1. The 1994 live birth rates per thousand population in the mountain states of Idaho, Montana, Wyoming, Colorado, New Mexico, Arizona, Utah, and Nevada were 12.9, 15.5, 13.5, 14.8, 16.7, 17.4, 20.1 and 16.4, respectively. What is the mean, variance, and standard deviation?

2. In five attempts, it took a person 11, 15, 12, 8, and 14 minutes to change a tire on a car. What is the mean, variance, and standard deviation?

3. A soldier is sent to the range to test a new bullet that the manufacturer says is very accurate. You send your best shooter with his weapon. He fires 10 shots with each using the standard ammunition and then the new ammunition. We measure the distance from the bull's eye to each shot. location. Which appears to the better ammunition? Explain.
   Standard Ammunition: -3,-3,-1,0,0,0,1,1,1,2

   New Ammunition -2,-1,0,0,0,0,1,1,1,2

4. AGCT Scores: **AGCT-score**

AGCT stands for Army General Classification Test. These scores have a mean of 100, with a standard deviation of 20.0. Here are the AGCT scores for a unit:

79, 100, 99, 83, 92, 110, 149, 109, 95, 126, 101, 101, 91, 71, 93, 103, 134, 141, 76, 108, 122, 111, 97, 94, 90, 112, 106, 113, 114, 117

Find the mean, median, mode, standard deviation, variance, and coefficient of skewness for the data. Provide a brief summary to your S-1 about this data.

Lesson 4-8 Mathematical Modeling of Decision Making using Multi-Attribute Decision Making Tools

CHAPTER 2 MADM

Objectives:

(1) Students understand the power and limitations to MADM techniques.

(2) Students understand the power and limitations of decision maker weights.

(3) Students understand the concept of pairwise comparison in analysis.

(4) Students can use TOPSIS template with AHP weights.

(5) Students understand and interpret output.

(6) Students understand the importance of "sensitivity analysis"

There are three "good" tools for MADM. Generally they are used when you have more than 2 alternatives (course of action) each of which has several attributes or characteristics that distinguish them from one another. The three tools are Data Envelopment Analysis (DEA), Analytical Hierarchy Process (AHP), and Technique of Order preference by Similarity to Ideal Solution (TOPSIS). Although DEA is a tool, we will not address it in this course.

2.1 Analytical Hierarchy Process (AHP)

2.1.1 Description and Uses

AHP is a multi-objective decision analysis tool first proposed by Satty (1980). It is designed when either subjective and objective measures or just subjective measures are being evaluated in terms of a set of alternatives based upon multiple criteria, organized in a hierarchical structure. At the top level, the criteria are evaluated or weighted, and at the bottom level the alternatives are measured against each criterion. The decision maker assesses their evaluation by making pairwise comparisons in which every pair is subjectively or objectively compared. The subjective method involves a 9 point scale that we present later in Table 3.

We only desire to briefly discuss the elements in the framework of AHP. AHP can be described as a method to decompose a problem into sub-problems. In most decisions, the decision maker has a choice among several to many alternatives. Each alternative has a

set of attributes or characteristics that can be measured, either subjectively or objectively. We will call these attributes, criteria. The attribute elements of the hierarchal process can relate to any aspect of the decision problem—tangible or intangible, carefully measured or roughly estimated, well- or poorly-understood—anything at all that applies to the decision at hand.
In its simplest sense we can state that in order to perform AHP we need an objective, a set of alternatives, each with criteria (attributes) to compare. Once the hierarchy is built, the decision maker(s) systematically evaluate its various elements pairwise (by comparing them to one another two at a time), with respect to their impact on an element above them in the hierarchy. In making the comparisons, the decision makers can use concrete data about the elements, but they typically use their judgments about the elements' relative meaning and importance. It is the essence of the AHP that human judgments, and not just the underlying information, both can be used in performing the evaluations.

The AHP converts these evaluations to numerical values that can be processed and compared over the entire range of the problem. A numerical weight or priority is derived for each element of the hierarchy, allowing diverse and often incommensurable elements to be compared to one another in a rational and consistent way. This capability distinguishes the AHP from other decision making techniques.

In the final step of the process, numerical priorities are calculated for each of the decision alternatives. These numbers represent the alternatives' relative ability to achieve the decision goal, so they allow a straightforward consideration of the various courses of action.

While it can be used by individuals working on straightforward decisions, the Analytic Hierarchy Process (AHP) is most useful where teams of people are working on complex problems, especially those with high stakes, involving human perceptions and judgments, whose resolutions have long-term repercussions. It has unique advantages when important elements of the decision are difficult to quantify or compare, or where communication among team members is impeded by their different specializations, terminologies, or perspectives.

Decision situations to which the AHP can be applied include the following where we desire ranking:

- Choice - The selection of one alternative from a given set of alternatives, usually where there are multiple decision criteria involved.
- Ranking - Putting a set of alternatives in order from most to least desirable
- Prioritization - Determining the relative merit of members of a set of alternatives, as opposed to selecting a single one or merely ranking them
- Resource allocation - Apportioning resources among a set of alternatives
- Benchmarking - Comparing the processes in one's own organization with those of other best-of-breed organizations

- Quality management - Dealing with the multidimensional aspects of quality and quality improvement
- Conflict resolution - Settling disputes between parties with apparently incompatible goals or positions

2.1.2 Methodology of the Analytic Hierarchy Process

The procedure for using the AHP can be summarized as:

Step 1. Build the hierarchy for the decision

Goal	*Select the best alternative*
Criteria	$c_1, c_2, c_3, \ldots, c_m$
Alternatives:	$a_1, a_2, a_3, \ldots, a_n$

Step 2. Judgments and Comparison

Build a numerical representation using a 1-9 point scale in a pairwise comparison for the attributes criterion and the alternatives. The goal, in AHP, is to obtain a set of eigenvectors of the system that measures the importance with respect to the criterion. We can put these values into a matrix or table based on the values from Table 2.1.

Table 2.1. Saaty's 9-Point Scale

Intensity of Importance in Pair-wise Comparisons	Definition
1	Equal Importance
3	Moderate Importance
5	Strong Importance
7	Very Strong Importance

9	Extreme Importance
2,4,6,8	For comparing between the above
Reciprocals of above	In comparison of elements *i* and *j* if *i* is 3 compared to *j*, then j is 1/3 compared to *i*.
Rational	Force consistency; measure values available

We must insure that this pairwise matrix is consistent according to Saaty's scheme to compute *CR*. The value of *CR* must be less than or equal to 0.1 to be considered consistent. Saaty's computed the random index, *RI*, for random matrices for up to 10 criteria, see Table 2.2.

Table 2.2 *RI* from Saaty

n	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
RI	*0*	*0*	*0.52*	*0.89*	*1.1*	*1.24*	*1.35*	*1.4*	*1.45*	*1.49*

Next, we approximate the largest eigenvalue, λ, using the power method (Burden et al., 2012). We compute the consistency index, *CI*, with the formula:

$$CI=(\lambda-n)/(n-1)$$

We compute *CR* using:

$$CR=CI/RI$$

If $CR \leq 0.1$, then our pairwise comparison matrix is consistent and we may continue. If not, we must go back to our pairwise comparison and make fix the inconsistencies until the $CR \leq 0.1$. In general the consistency insures that if $A > B$, $B > C$, that $A > C$ for all A,B, and C.

Step 3. Finding all the eigenvectors combined in order to obtain a comparative ranking.

Step 4. After the $m \times 1$ criterion weights are found and the $n \times m$ matrix for n alternatives by m criterion, we use matrix multiplication to obtain the $n \times 1$ final rankings.

Step 5. We order the final ranking and interpret.

2.1.3 Strengths and Limitations of AHP

Like all modelling methods, the AHP has strengths and limitations.
The main advantage of the AHP is its ability to rank choices in the order of their effectiveness in meeting conflicting objectives. If the judgments made about the relative importance of criteria and those about the alternatives' ability to satisfy those objectives, have been made in good faith and effort, then the AHP calculations lead to the logical consequence of those judgments. It is quite hard – but not impossible – to 'fiddle' with the pair-wise judgments to get some predetermined result. A further strength of the AHP is its ability to detect inconsistent judgments in the pair-wise comparisons using the *CR* value.

The limitations of the AHP are that it only works because the matrices are all of the same mathematical form – known as a positive reciprocal matrix. The reasons for this are explained in Saaty's book, which is not for the mathematically daunted, so we will simply state that point. To create such a matrix requires that, if we use the number 9 to represent 'A is absolutely more important than B', then we have to use 1/9 to define the relative importance of B with respect to A. Some people regard that as reasonable; others do not.

Some suggest a drawback is in the possible scaling. However, understanding that the values obtained simply say that something is relatively better than another at meeting some objective. For example, if the AHP ranks found were (0.392, 0.406, 0.204) then they only implies that alternatives A and B are about equally good at 0.4, while C is worse at 0.2. It does not mean that A and B are twice as good as C.

In less clear-cut cases, it would not be a bad thing to change the rating scale and see what difference it makes. If one option consistently scores well with different scales, it is likely to be a very robust choice.

In short, the AHP is a useful technique for discriminating between competing options in the light of a range of objectives to be met. The calculations are not complex and, while the AHP relies on what might be seen as a mathematical trick, you don't need to understand the mathematics to use the technique. Be aware that it only shows relative values.

Although AHP has been used in many applications of the public and private sectors, Hartwich (1999) noted several limitations. First and foremost, AHP was criticized for not providing sufficient guidance about structuring the problem to be solved, forming the levels of the hierarchy for criteria and alternatives, and aggregating group opinions when team members are geographically dispersed or are subject to time constraints. Team members may carry out rating items individually or as a group. As the levels of hierarchy increase, so does the difficulty and time it takes to synthesize weights. One remedy in preventing these problems is by conducting "AHP Walk-throughs" (i.e., meetings of decision-making participants who review the basics of the AHP methodology and work through examples so that concepts are thoroughly and easily understood).

Another critique of AHP is the "rank reversal" problem, i.e., changes in the importance ratings whenever criteria or alternatives are added-to or deleted-from the initial set of alternatives compared. Several modifications to AHP have been proposed to cope with this and other related issues. Many of the enhancements involved ways of computing, synthesizing pair-wise comparisons, and/or normalizing the priority and weighting vectors. We mention now that TOPSIS corrects this rank reversal issue.

2.1.4 Sensitivity Analysis

Since AHP, at least in the pair-wise comparisons, is based upon subjective inputs using the 9 point scale that sensitivity analysis is extremely important. How often do we change our minds about the relative importance of an object, place, or thing? Often enough that we should alter the pair-wise comparison values to determine how robust our rankings are in the AHP process. We might even want to find the "break-point" values of the decision maker weights that change the rankings of our alternatives. Since the pair-wise comparisons are subjective matrices compiled using the Saaty method, we suggest as a minimum a "trial and error" sensitivity analysis.

A simple method of sensitivity analysis can be obtained from changing one weight at a time using a scheme:

Step 1. Choose the largest weighted criteria, call this weight w_p.

Step 2. Choose a Δ, for example, Δ=-0.1

Step 3. The new weight for w_p is called w_p'.

Step 4, Find the other new weights using the formula in equation (2.1)

$$w_j' = w_j\left(\frac{(1-w_p')}{(1-w_p)}\right) \quad (2.1)$$

Step 5. Use these new weights to calculate and find the new rankings

Step 6. Plot the alternative values and use to analyze changes.

We will illustrate later in our discussion.

2.1.5 Illustrative Examples

Example 1. Car Selection (data from Consumer's Report and US News and World Report on-line data) using AHP

We are considering six cars: Ford Fusion, Toyota Prius, Toyota Camry, Nissan Leaf, Chevy Volt, and Hyundai Sonata. For each car we have data on seven criteria that were extracted from Consumer's Report and US News and World Report data sources. They are *cost, mpg city, mpg highway, performance, interior & style, safety,* and *reliability*. We provide the extracted information in the Table 2.3:

Table 2.3 Raw data for cars

Cars	*Cost ($000)*	*MPG City*	*MPG HW*	*Performance*	*Interior & Style*	*Safety*	*Reliability*
Prius	27.8	44	40	7.5	8.7	9.4	3
Fusion	28.5	47	47	8.4	8.1	9.6	4
Volt	38.668	35	40	8.2	6.3	9.6	3
Camry	25.5	43	39	7.8	7.5	9.4	5
Sonata	27.5	36	40	7.6	8.3	9.6	5
Leaf	36.2	40	40	8.1	8.0	9.4	3

Step 1. Build the hierarchy and prioritize the criterion from your highest to lower priority.

Goal — *Select the best car*
Criteria — $c_1, c_2, c_3, \ldots, c_m$
Alternatives: — $a_1, a_2, a_3, \ldots, a_n$

For our cars example we choose the priority as follows: Cost, MPG City, Safety, Reliability, MPG Highway, Performance, and Interior & style. Putting these in an order allows for an easier assessment of the pairwise comparisons.

Step 2. Perform the pairwise comparisons using Saaty's 9-point scale with the EXCEL template.

	Element A		Element B	More Important	Intensity (1-9)
1	cost	compared with	mpg city	A	2
2			safety	A	2
3			reliability	A	3
4			mpg hw	A	4
5			performance	A	5
6			interior & style	A	6
7					
1	mpg city	compared with	safety	A	2
2			reliability	A	3
3			mpg hw	A	4
4			performance	A	5
5			interior & style	A	5
6					
1	safety	comp. with	reliability	A	2
2			mpg hw	A	2
3			performance	A	3
4			interior & style	A	3
5					
1	reliability	comp. with	mpg hw	A	1
2			performance	A	2
3			interior & style	A	3
4					
1	mpg hw	vs	performance	A	2
2			interior & style	A	3
3					
1	performance	vs	interior & style	A	2
2					
1		vs			

This yields the following decision criterion matrix,

Matrix 0

		1 cost	2 mpg city	3 safety	4 reliability	5 mpg hw	6 performance	7 interior & style
1	cost	1	2	2	3	4	5	6
2	mpg city	1/2	1	2	3	4	5	5
3	safety	1/2	1/2	1	2	2	3	3
4	reliability	1/3	1/3	1/2	1	1	2	3
5	mpg hw	1/4	1/4	1/2	1	1	2	3
6	performance	1/5	1/5	1/3	1/2	1/2	1	2
7	interior & style	1/6	1/5	1/3	1/3	1/3	1/2	1

We check the *CR*, the consistency ration, to insure it is less than 0.1. For our pairwise decision matrix the *CR=0.00695*. Since the *CR < 0.1*, we continue.

We use the power method or discrete dynamical system, within the template, and find the eigenvector for the decision weights as:

6	Eigenvector Criterion Weights	
7		
8	cost	0.342407554
9	mpg city	0.230887543
10	safety	0.151297361
11	reliability	0.094091851
12	mpg hw	0.080127732
13	performance	0.055515667
14	interior & style	0.045672293
15	n/a	0
16		

All you want is to obtain these weights from the template.

Step 3. For the alternatives, we either have the data as we obtained it from consumer reports for each car under each decision criterion or we must use pairwise comparisons, by criteria for how car fares versus its competitors. In this example, we take the data from before and ***normalize*** the columns except cost. We cannot use cost directly since a lower cost is better while for all variables a higher value is better. Thus, we have three courses of action (1) use 1/cost to replace cost, (2) use a pairwise comparison using the nine point scale, or (3) remove cost and then completely and then do a *benefit /cost* ratio to rank the results.

Normalize→ this means that you take each value in a column and divide by the sum total of values in that column..

Step 4. We multiply the matrix of the *normalized raw data* from Consumer Reports and the *matrix* of weights to obtain the rankings. Using option (1) discussed in Step 3, we obtain the following results:

Car	AHP Values
Prius	0.170053
Fusion	0.178746
Volt	0.145801
Camry	0.18166
Sonata	0.17162
Leaf	0.15212

We rank order these and get:

5	Rank Order these Values	
6		
7	Car	AHP Values
8	Camry	0.18166
9	Fusion	0.178746
10	Sonata	0.17162
11	Prius	0.170053
12	Leaf	0.15212
13	Volt	0.145801
14		

Camry is our first choice, followed by Fusion, Sonata,, Prius, Leaf, and Volt.

If we use option (2) then in the matrix, we replace the actual costs with the following pairwise results (CR=0.031):

Pairwise weights for alternatives under criterion, *cost:*

Prius	0.107059
Fusion	0.073259
Volt	0.046756
Camry	0.465277
Sonata	0.256847
Leaf	0.050802

We place the column of raw data for cost with the pairwise comparisons values, normalize the data, and perform the matrix multiplication to obtain the results:

Car	AHP Values
Camry	0.274347
Sonata	0.197785
Prius	0.145595
Fusion	0.144216
Leaf	0.122581
Volt	0.115476

If we do option (3) and obtain a benefit/cost ratio then our results are

Camry	1.205076
Fusion	1.174729
Prius	1.137356
Sonata	1.089212
Leaf	0.822593
Volt	0.725676

Sensitivity Analysis

Trial and Error

We altered our decision pairwise values to obtain a new set of decision weights to use in (1) to obtain the results Camry, Fusion, Sonata, Prius, leaf, and Volt. The new weights and model's results are:

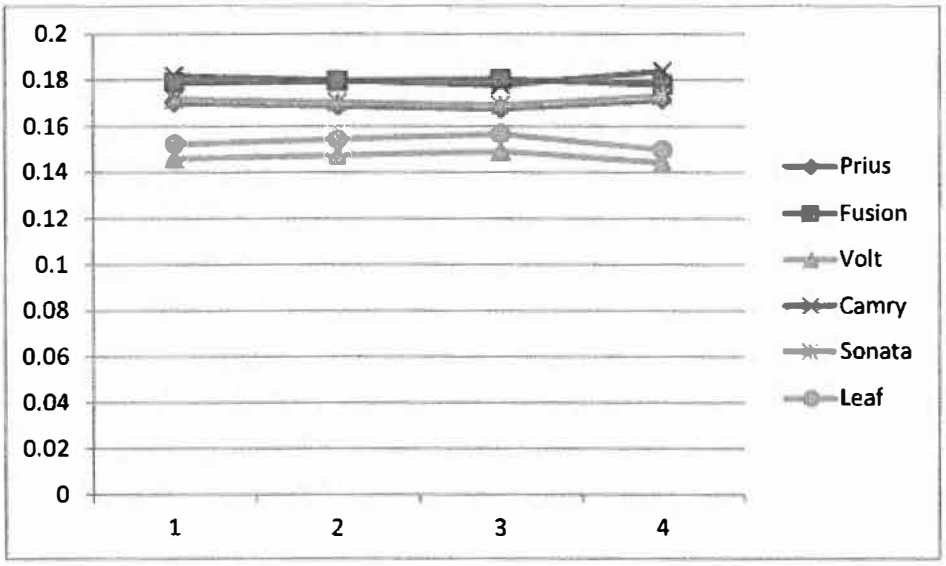
Cost	**0.311155922**
MPG City	**0.133614062**
MPG Hw	**0.095786226**
Performance	**0.055068606**
Interior	**0.049997069**
Safety	**0.129371535**
Reliability	**0.225006578**

Alternatives	Values	
Prius	**0.10882648**	**4**
Fusion	**0.11927995**	**2**
Volt	**0.04816882**	**5**
Camry	**0.18399172**	**1**
Sonata	**0.11816156**	**3**
Leaf	**0.04357927**	**6**

The resulting values have changed but not the relative rankings of the cars.

Using the equation (2.1), varying the largest weighted criteria, cost, in increments of ±0.1

We recomputed the AHP values for the new weights each time and then perform a line-plot of the columns.



We see from the plot that Camry is replaced as the top alternative by Fusion when Δ is between -0.1 and -0.2. This tells us that depending on how we weight the criteria we obtain a different ranking.

Example 2. Kite Network

Assume all we have are the outputs from ORA which we do not show here due to the volume of output produced. We take the metrics from ORA and normalize each column. The columns for each criterion are placed in a matrix X with entries, x_{ij}. We define *wj* as the weights for each criterion. We set up the linear program using equation *(1)* with the output from the Kite Network.

Next, we assume we can obtain pairwise comparison matrix from the decision maker concerning the criterion. We use the output from ORA and normalize the results for AHP to rate the alternatives within each criterion. We provide a sample pairwise comparison matrix for weighting the criterion from the Kite example using Saaty's 9-point scale.

Pairwise Comparison Matrix

	Central	*Eigenvector*	*In-degree*	*Out-degree*	*Information centrality*	*Betweenness*

Central	*1*	*3*	*2*	*2*	*½*	*1/3*
Eigenvector	*1/3*	*1*	*1/3*	*1*	*2*	*½*
In-degree	*½*	*3*	*1*	*½*	*½*	*¼*
Out-degree	*½*	*½*	*1*	*1*	*¼*	*¼*
Information Centrality	*2*	*2*	*4*	*4*	*1*	*1/3*
Betweenness	*3*	*2*	*4*	*4*	*3*	*1*

The *CR* is *0.0828,* which is less than *0.1*, so our pairwise matrix is consistent and we continue.

We obtain the steady state values that will be our weights, where the sum of the weights equals 1.0. There exist many methods to obtain these weights. We obtain the following weights from our template.

0.1532
0.1450
0.1194
0.0672
0.1577
0.3575

These values provide the weights for each criterion: *centrality =0.1532, eigenvectors =0.1450, in-centrality =0.1194, out-centrality = 0.0672, information centrality=0.1577,* and *betweenness = 0.3575.*

We multiply the matrix of the weights and the normalized matrix of the metrics from ORA to obtain our output and ranking are:

Susan	**0.159421**	**2**
Steven	**0.132728**	**3**
Sarah	**0.113323**	**4**
Tom	**0.075717**	**6**
Claire	**0.075717**	**6**
Fred	**0.061358**	**8**
David	**0.061358**	**8**
Claudia	**0.175856**	**1**
Ben	**0.109015**	**5**
Jennifer	**0.035507**	**10**

For this example with AHP Claudia, *cl*, is the key node. However, the bias of the decision maker is important in the analysis of the criterion weights. The criterion,"*Betweenness*", is 2 to 3 times more important than the other criterion.

Sensitivity Analysis

Changes in the pairwise decision criterion cause fluctuations in the key nodes. We change our pairwise comparison so that "*Betweenness*" is not so dominant a criterion.

	Centrality	IN	OUT	Eigen	EIGENC	Close	IN-Close	Betw	INFO Cen.
t	0.111111	0.111111	0.111111	0.114399	0.114507	0.100734	0.099804	0.019408	0.110889
c	0.111111	0.111111	0.111111	0.114399	0.114507	0.100734	0.099804	0.019408	0.108891
f	0.083333	0.083333	0.083333	0.093758	0.094004	0.097348	0.09645	0	0.097902
s	0.125	0.138889	0.111111	0.137528	0.137331	0.100734	0.111826	0.104188	0.112887
su	0.180556	0.166667	0.194444	0.175081	0.174855	0.122743	0.107632	0.202247	0.132867
st	0.138889	0.138889	0.138889	0.137528	0.137331	0.112867	0.111826	0.15526	0.123876
d	0.083333	0.083333	0.083333	0.093758	0.094004	0.097348	0.107632	0	0.100899
cl	0.083333	0.083333	0.083333	0.104203	0.104062	0.108634	0.107632	0.317671	0.110889
b	0.055556	0.055556	0.055556	0.024123	0.023985	0.088318	0.087503	0.181818	0.061938
j	0.027778	0.027778	0.027778	0.005223	0.005416	0.070542	0.069891	0	0.038961

10 alternatives and 9 attirbutes or criterion

Criterion weights

w1	0.034486
w2	0.037178
w3	0.045778
w4	0.398079
w5	0.055033
w6	0.086323
w7	0.135133
w8	0.207991

With these slight pairwise changes, we now find Susan is now ranked first, followed by Steven and then Claudia. The ***AHP process is sensitive*** to changes in the criterion weights.

Tom	0.098628	Susan	0.161609
Claire	0.098212	Steven	0.133528
Fred	0.081731	Claudia	0.133428
Sarah	0.12264	Sarah	0.12264
Susan	0.161609	Tom	0.098628
Steven	0.133528	Claire	0.098212
David	0.083319	David	0.083319
Claudia	0.133428	Fred	0.081731
Ben	0.0645	Ben	0.0645
Jennifer	0.022405	Jennifer	0.022405

2.2 Technique of Order Preference by Similarity to the Ideal Solution (TOPSIS)

2.2.1 Description and Uses

The Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) is a multi-criteria decision analysis method, which was originally developed in a dissertation from Kansas State University (Hwang and Yoon, 1981). It has been further development by other (Yoon, 1987; Hwang et al, 1993). TOPSIS is based on the concept that the chosen alternative should have the shortest geometric distance from the positive ideal solution and the longest geometric distance from the negative ideal solution. It is a method of compensatory aggregation that compares a set of alternatives by identifying weights for each criterion, normalizing the scores for each criterion and calculating the geometric distance between each alternative and the ideal alternative, which is the best score in each criterion. An assumption of TOPSIS is that the criteria are monotonically increasing or decreasing. Normalization is usually required as the parameters or criteria are often of incompatible dimensions in multi-criteria problems. Compensatory methods such as TOPSIS allow trade-offs between criteria, where a poor result in one criterion can be negated by a good result in another criterion. This provides a more realistic form of modeling than non-compensatory methods, which include or exclude alternative solutions based on hard cut-offs.

We only desire to briefly discuss the elements in the framework of TOPSIS. TOPSIS can be described as a method to decompose a problem into sub-problems. In most decision, the decision maker has a choice among several to many alternatives. Each alternative has a set of attributes or characteristics that can be measured, either subjectively or objectively. The attribute elements of the hierarchal process can relate to any aspect of the decision problem—tangible or intangible, carefully measured or roughly estimated, well- or poorly-understood—anything at all that applies to the decision at hand.

2.2.2 Methodology

The TOPSIS process is carried out as follows:

Step 1 Create an evaluation matrix consisting of m alternatives and n criteria, with the intersection of each alternative and criteria given as x_{ij}, giving us a matrix $(X_{ij})_{m \times n}$.

$$D = \begin{array}{c} \\ A_1 \\ A_2 \\ A_3 \\ \cdot \\ \cdot \\ \cdot \\ A_m \end{array} \begin{array}{c} \begin{array}{cccccccc} x_1 & x_2 & x_3 & \cdot & \cdot & \cdot & x_n \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdot & \cdot & \cdot & x_{1n} \\ x_{21} & x_{22} & x_{23} & \cdot & \cdot & \cdot & x_{2n} \\ x_{31} & x_{32} & x_{33} & \cdot & \cdot & \cdot & x_{3n} \\ \cdot & \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & \cdot & & & & \cdot \\ x_{m1} & x_{m2} & x_{m3} & \cdot & \cdot & \cdot & x_{mn} \end{bmatrix} \end{array}$$

Step 2 The matrix shown as D above then normalized to form the matrix $R=(R_{ij})_{m\,x\,n}$, using the normalization method

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum x_{ij}^2}}$$

for $i=1,2...,m;\ j=1,2,...n$

Step 3 Calculate the weighted normalized decision matrix. First we need the weights. Weights can come from either the decision maker or by computation.

Step 3 a. Use either the decision maker's weights for the attributes $x_1,x_2,...x_n$ or compute the weights through the use Saaty's [10] AHP's decision maker weights method to obtain the weights as the eigenvector to the attributes versus attribute pair-wise comparison matrix.

$$\sum_{j=1}^{n} w_j = 1$$

The sum of the weights over all attributes must equal 1 regardless of the method used.

Step 3b. Multiply the weights to each of the column entries in the matrix from *Step 2* to obtain the matrix, T.

$$T = (t_{ij})_{m\,x\,n} = (w_j r_{ij})_{m\,x\,n}, i = 1,2,\dots,m$$

Step 4 Determine the worst alternative (A_w) and the best alternative (A_b) : Examine each attribute's column and select the largest and smallest values appropriately. If the values imply larger is better (profit) then the best alternatives are the largest values and if the values imply smaller is better (such as cost) then the best alternative is the smallest value.

$$A_w = \{\langle \max(t_{ij}|i = 1,2,\dots,m|j \in J_-\rangle, \langle \min(t_{ij}|i = 1,2,\dots,m)|j \in J_+\rangle\}$$
$$\equiv \{t_{wj}|j = 1,2,\dots,n\},$$
$$A_{wb} = \{\langle \min(t_{ij}|i = 1,2,\dots,m|j \in J_-\rangle, \langle \max(t_{ij}|i = 1,2,\dots,m)|j \in J_+\rangle\}$$
$$\equiv \{t_{bj}|j = 1,2,\dots,n\},$$

where,

$J_+ = \{j = 1,2,\dots n|j)$ associated with the criteria having a positive impact, and

$J_- = \{j = 1,2,\dots n|j)$associated with the criteria having a negative impact.

We suggest that if possible make all entry values in terms of positive impacts.

Step 5 Calculate the L2-distance between the target alternative i and the worst condition A_w

$$d_{iw} = \sqrt{\sum_{j=1}^{n}(t_{ij} - t_{wj})^2}, i=1,2,...m$$

and the distance between the alternative i and the best condition A_b

$$d_{ib} = \sqrt{\sum_{j=1}^{n}(t_{ij} - t_{bj})^2}, i=1,2,...m$$

where d_{iw} and d_{ib} are L2-norm distances from the target alternative i to the worst and best conditions, respectively.

Step 6 Calculate the similarity to the worst condition:

$$s_{iw} = \frac{d_{iw}}{(d_{iw} + d_{ib})}, 0 \leq s_{iw} \leq 1, i = 1,2, \ldots, m$$

$S_{iw}=1$ if and only if the alternative solution has the worst condition; and

$S_{iw}=0$ if and only if the alternative solution has the best condition.

Step 7 Rank the alternatives according to their value from S_{iw} $(i=1,2,...,m)$.

Normalization

Two methods of normalization that have been used to deal with incongruous criteria dimensions are linear normalization and vector normalization.

Linear normalization can be calculated as in *Step 2* of the TOPSIS process above. Vector normalization was incorporated with the original development of the TOPSIS method (Yoon, 1987), and is calculated using the following formula:

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum x_{ij}^2}} \quad \text{for } i=1,2...,m; j=1,2,...n$$

In using vector normalization, the non-linear distances between single dimension scores and ratios should produce smoother trade-offs (Hwang and Yoon, 1981).

Let's suggest two options for the weights in Step 3. First, the decision maker might actually have a weighting scheme that they want the analyst to use. In not, we suggest using Saaty's 9-Point pair-wise method developed for the Analytical Hierarchy Process (AHP) (Saaty, 1980). We refer the reader to our discussion in the AHP section for the decision weights using the Saaty's 1-9 point scale and pairwise comparisons. In TOPSIS, we have the following scheme.

Objective Statement ← This is the decision desired
Alternatives: *1, 2, 3, ..., n*
For each of the alternatives there are criteria (attributes) to compare:
Criteria (or Attributes): $c_1, c_2, ..., c_m$

Once the hierarchy is built, the decision maker(s) systematically evaluate its various elements pairwise (by comparing them to one another two at a time), with respect to their impact on an element above them in the hierarchy. In making the comparisons, the decision makers can use concrete data about the elements, but they typically use their judgments about the elements' relative meaning and importance. It is the essence of the TOPSIS that human judgments, and not just the underlying information, can be used in performing the evaluations.

TOPSIS converts these evaluations to numerical values that can be processed and compared over the entire range of the problem. A numerical weight or priority is derived for each element of the hierarchy, allowing diverse and often incommensurable elements to be compared to one another in a rational and consistent way. This capability distinguishes the TOPSIS from other decision making techniques.

In the final step of the process, numerical priorities or ranking are calculated for each of the decision alternatives. These numbers represent the alternatives' relative ability to achieve the decision goal, so they allow a straightforward consideration of the various courses of action.

While it can be used by individuals working on straightforward decisions, TOPSIS is most useful where teams of people are working on complex problems, especially those with high stakes, involving human perceptions and judgments, whose resolutions have long-term repercussions. It has unique advantages when important elements of the decision are difficult to quantify or compare, or where communication among team members is impeded by their different specializations, terminologies, or perspectives.

Decision situations to which the TOPSIS might be applied and are identical to what we presented earlier for AHP:

- Choice - The selection of one alternative from a given set of alternatives, usually where there are multiple decision criteria involved.
- Ranking - Putting a set of alternatives in order from most to least desirable
- Prioritization - Determining the relative merit of members of a set of alternatives, as opposed to selecting a single one or merely ranking them
- Resource allocation - Apportioning resources among a set of alternatives
- Benchmarking - Comparing the processes in one's own organization with those of other best-of-breed organizations

- Quality management - Dealing with the multidimensional aspects of quality and quality improvement
- Conflict resolution - Settling disputes between parties with apparently incompatible goals or positions

2.2.3 Strengths and Limitations

TOPSIS is based on the concept that the chosen alternative should have the shortest geometric distance from the positive ideal solution and the longest geometric distance from the negative ideal solution. It is a method of compensatory aggregation that compares a set of alternatives by identifying weights for each criterion, normalizing scores for each criterion and calculating the geometric distance between each alternative and the ideal alternative, which is the best score in each criterion. An assumption of TOPSIS is that the criteria are monotonically increasing or decreasing. Normalization is usually required as the parameters or criteria are often of incongruous dimensions in multi-criteria problems. Compensatory methods such as TOPSIS allow trade-offs between criteria, where a poor result in one criterion can be negated by a good result in another criterion. This provides a more realistic form of modelling than non-compensatory methods, which include or exclude alternative solutions based on hard cut-offs. TOPSIS corrects the rank reversal that was a limitation in strictly using the AHP method. TOPSIS also allows the user to state which criterion are maximized and which are minimized for better results. In the late 1980's TOPSIS was a department of defense standard for performing selection of systems across all branches in tight budget years.

2.2.4 Sensitivity Analysis

The decision weights are subject to sensitivity analysis to determine how the affect the final ranking. The same procedures discussed in section 4.4 are valid here. Sensitivity analysis is essential to good analysis.

2.2.5 Illustrate examples

Example 1. Revisit the Car Selection from example 2.1.5

We might assume that our decision maker weights from the AHP section are still valid for our use.
Weights from before:

Cost	**0.38960838**
MPG City	**0.11759671**
MPGHW	**0.04836533**
Performance	**0.0698967**
Interior	**0.05785692**
Safety	**0.10540328**
Reliability	**0.21127268**

	cost	MPG_city	MPG_HW	Perf.	Interior	safety	reliability	N/A
Cost	1	4	6	5	6	4	2	0
MPG_city	0.25	1	6	3	5	1	0.333333333	0
MPG_HW	0.166667	0.166667	1	0.5	0.5	0.333333	0.25	0
Perf.	0.2	0.333333	2	1	2	0.5	0.333333333	0
Interior	0.166667	0.2	2	0.5	1	0.5	0.333333333	0
safety	0.25	1	3	2	2	1	0.5	0
reliability	0.5	3	4	3	3	2	1	0
N/A	0	0	0	0	0	0	0	1

We use the identical data from the car example from AHP but we apply step 3-7 from TOPSIS to our data. We are able to keep the cost data and just inform TOPSIS that a smaller cost is better. We obtained the following rank ordering of the cars: Camry, Fusion, Prius, Sonata, Volt, and Leaf.

4	**0.82154128**	**Camry**
2	**0.74622988**	**Fusion**
1	**0.72890117**	**Prius**
5	**0.70182382**	**Sonata**
6	**0.15580913**	**Leaf**
3	**0.11771999**	**Volt**

It is critical to perform sensitivity analysis on the weights to see how they affect the final ranking. This time we work toward finding the break point where the order of cars actually changes.

Example 2. *Kite Network Analysis* with TOPSIS to find influences on a network
We present the output from ORA that we used in Table 2.2.1.
Table 2.2.1. Summary of ORA's output for Kite Network.

	IN	OUT	Eigen	EigenL	Close	IN-Close	Betweenr	INF Centr
Tom	0.4	0.4	0.46	0.296	0.357	0.357	0.019	0.111
Claire	0.4	0.4	0.46	0.296	0.357	0.357	0.019	0.109
Fred	0.3	0.3	0.377	0.243	0.345	0.345	0	0.098
Sarah	0.5	0.4	0.553	0.355	0.357	0.4	0.102	0.113
Susan	0.6	0.7	0.704	0.452	0.435	0.385	0.198	0.133
Steven	0.5	0.5	0.553	0.355	0.4	0.4	0.152	0.124
David	0.3	0.3	0.377	0.243	0.345	0.385	0	0.101
Claudia	0.3	0.3	0.419	0.269	0.385	0.385	0.311	0.111
Ben	0.2	0.2	0.097	0.062	0.313	0.313	0.178	0.062
Jennifer	0.1	0.1	0.021	0.014	0.25	0.25	0	0.039

We use the decision weights from AHP (unless a decision maker gives us their own weights) and find the eigenvectors for our eight metrics as:

w1	0.034486
w2	0.037178
w3	0.045778
w4	0.398079
w5	0.055033
w6	0.086323
w7	0.135133
w8	0.207991

We take the metrics from ORA and perform steps 2-7 of TOPSIS.

S+	S-	C		
0.113035	0.133492	0.541489	**Susan**	**0.861966**
0.113199	0.133063	0.540332	**Steven**	**0.721125**
0.134113	0.108322	0.446809	**Sarah**	**0.675151**
0.077748	0.161588	0.675151	**Claudia**	**0.649058**
0.034008	0.212364	0.861966	**Tom**	**0.541489**
0.064772	0.167489	0.721125	**Claire**	**0.540332**
0.133751	0.109236	0.449555	**David**	**0.449555**
0.083224	0.153919	0.649058	**Fred**	**0.446809**
0.183362	0.059804	0.245938	**Bon**	**0.245938**
0.224337	0	0	**Jennfier**	**0**

We rank order the final output from TOPSIS as shown in the last column above. We interpret the results as follows: The key node is *Susan* followed by *Steven, Sarah,* and *Claire.*

Sensitivity Analysis

Again we suggest the use of equation (2.1) to systemically change the weights, recomputed the TOPSIS ranks, and plot the changes by criteria changed. The purpose is to attempt to find the break point values where the top ranking change.

References

W. Cooper, L. Seiford, and K. Tone. Data envelopment analysis. Kluwer Academic Press. London, UK. 2000.

A. Charnes , W. Cooper, and E. Rhodes. Measuring the efficiency of decision making units. European Journal of Operations Research, 2 (1978),429-444.

W. Cooper, , Li, S., Seiford, L.M., Thrall, R.M., and Zhu, Joe, Sensitivity and stability analysis in DEA: some recent developments, *Journal of Productivity Analysis*, Vol. 15, No. 3 (2001), 217-246.

J. Callen (1991). Data envelopment analysis: practical survey and managerial accounting applications, Journal of Management Accounting Research, 3(1991), 35-57/

W. Winston, W (1995) Introduction to mathematical programming. Duxbury Press, Belmont, CA., 322-325, 1995.

M. A. Trick (1996). Multiple Criteria Decision Making for Consultants, **November 1996,** http://mat.gsia.cmu.edu/classes/mstc/multiple/multiple.html accessed April 2014.

M.A. Trick, Data Envelopment Analysis, Chapter 12, http://mat.gsia.cmu.edu/classes/QUANT/NOTES/chap12.pdf

L. Neralic (1998). Sensitivity analysis in models of data envelopment analysis. Mathematical Communications 3(1998), 41-59.

D. Krackhardt. Assessing the political landscape: Structure, cognition, and power in organizations. Admin. Science Quarterly, 35 (1990), 342-369.

Carley, K. M. 2001-2011. *Organizational risk analyzer (ORA).* Pittsburgh, PA: Center for Computational Analysis of Social and Organizational Systems (CASOS): Carnegie Mellon University.

Fox, W. and S. Everton, (2013), Mathematical Modeling in Social Network Analysis: Using TOPSIS to Find Node Influences in a Social Network, Journal of Mathematics and System Science, 3(10), 531-541.

Fox, W. and S. Everton, (2014), Mathematical Modeling in Social Network Analysis: Using Data Envelopment Analysis and Analytical Hierarchy Process to Find Node Influences in a Social Network,

Journal of Defense Modeling and Simulation accepted December 2013 for 2014 publication. Vol XX, 1-9.

P.C. Fishburn, (1967). Additive Utilities with Incomplete Product Set: Applications to Priorities and Assignments, Operations Research.

Consumer's Report Car Guide, 2008-2012.

T. Satty (1980). The analytical hierarchy process. McGraw Hill, United States, 1980.

Burden, R. & D. Faires (2013), *Numerical analysis*, Cengage Publishers, Boston, Ma.

Hartwich F, 1999. *Weighting of Agricultural Research Results: Strength and Limitations of the Analytic Hierarchy Process (AHP)*, Universitat Hohenheim. Retrieved from https://entwicklungspolitik.uni-hohenheim.de/uploads/media/DP_09_1999_Hartwich_02.pdf

C. L. Hwan and K. Yoon. Multiple attribute decision making: Methods and applications. New York: Springer-Verlag. 1981.

K. Yoon, K. A reconciliation among discrete compromise situations. Journal of Operational Research Society 38, (1987) 277–286.

C. L. Hwang, Y, Lai, and T.Y. Liu. A new approach for multiple objective decision making. Computers and Operational Research 20 (1993) 889–899.

W. P. Fox. Mathematical modeling of the analytical hierarchy process using discrete dynamical systems in decision analysis, Computers in Education Journal, July-Sept. (2012) 27-34.

F. Giordano, W. Fox, and S. Horton. A first course in mathematical modeling, Brooks-Cole Publishers, Boston, MA. 2008

G. Zhenhua. The application of DEA/AHP method to supplier selection. 2009 International Conference on Information Management, Innovation Management and Industrial Engineering, (2009) 449-451.

E. Thanassoulis. Introduction to the theory and application of data envelopment analysis-A foundation text with integrated software. Kluwer Academic Press. London, UK., 2011.

Additional Reading

B. Huang, J. Keisler, and I. Linkov. Multi-criteria decision analysis in environmental science: Ten years of applications and trends. Science of the Total Environment 409 (2011) 3578–3594.

Carley, K. M. 2001-2011. *Organizational risk analyzer (ORA)*. Pittsburgh, PA: Center for Computational Analysis of Social and Organizational Systems (CASOS): Carnegie Mellon University.

Everton, S. F. 2012. *Disrupting dark networks*. Cambridge and New York: Cambridge University Press.
Fox, W. P. and Everton (2103)
Roberts, N. & S. Everton. (2011). Strategies for combating dark networks, *Journal of Social Structure*, **12**, 1-32.

Sterne, J. (2010). *Social media metrics: How to measure and optimize your market investments*. John Wiley & Sons, New York: NY.

Trick, M. http://mat.gsia.cmu.edu/classes/mstc/dea/node3.html (accessed October 25, 2012).

Wellman, B. (1996) For a social network analysis of computer networks: A sociological perspective on collaborative work and virtual community. *Proceedings of SIGCPR/SIGMIS*. Denver, CO. ACM Press.

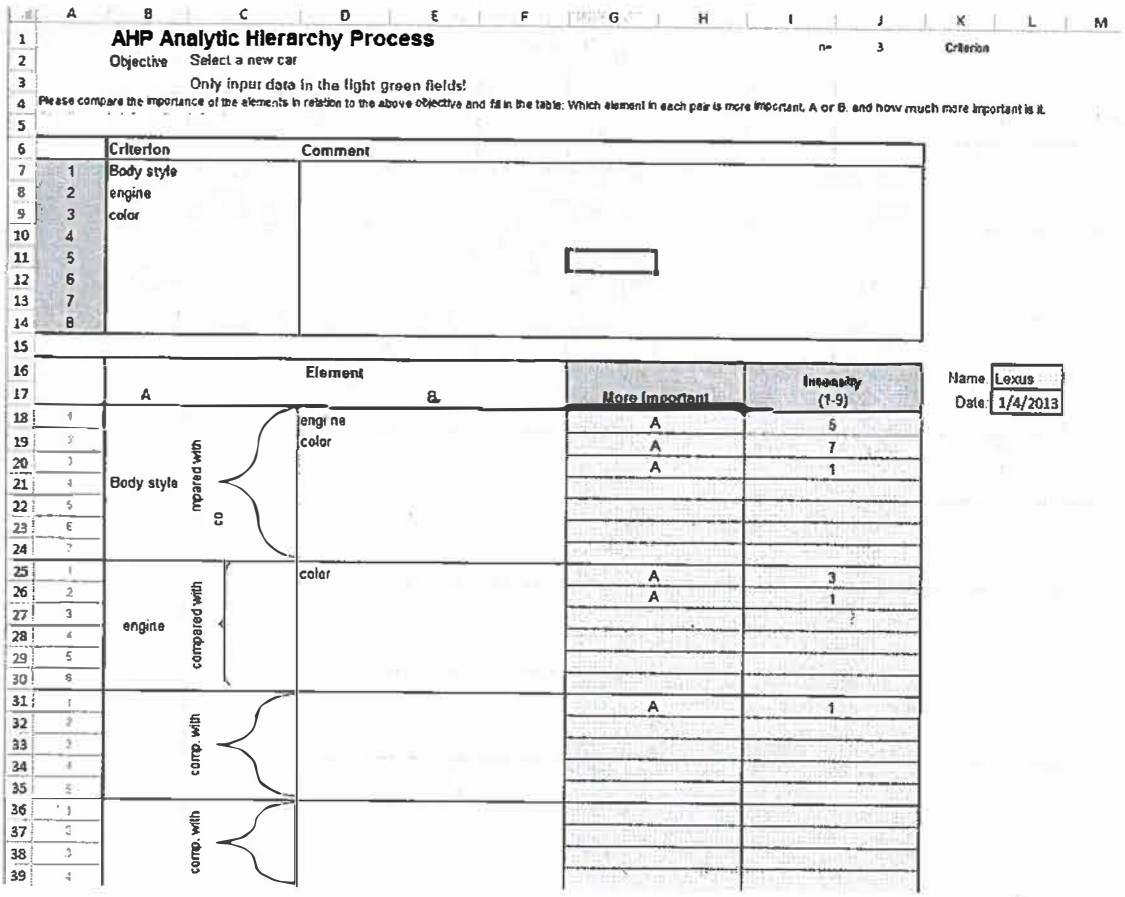
Zhenhua, G. (2009). The application of DEA/AHP method to supplier selection. 2009 International Conference on Information Management, Innovation Management and Industrial Engineering, 449-451.

http://en.wikipedia.org/wiki/Social_network_analysis Accessed October 2, 2012.
http://www.orgnet.com/sna.html How to do Social Network Analysis?"-Accessed October 2, 2012.

Using the EXCEL TEMPLATE

AHP-use template for up to *8 x 8* of alternatives and criterion.

Fill out template information and interpret results.



AHP Analytic Hierarchy Process

Objective: Select a new car

Only input data in the light green fields!

Please compare the importance of the elements in relation to the above objective and fill in the table: Which element in each pair is more important, A or B, and how much more important is it.

n= 3 Criterion

	Criterion	Comment
1	Body style	
2	engine	
3	color	
4		
5		
6		
7		
8		

Name: Lexus
Date: 1/4/2013

	Element A	compared with	Element B	More important	Intensity (1-9)
1	Body style	compared with	engine	A	5
2			color	A	7
3				A	1
4					
5					
6					
7					
8					
9	engine	compared with	color	A	3
10				A	1
11					
12					
13					
14					
15	comp. with			A	1
16					
17					
18					
19					
20					
21	comp. with				
22					
23					
24					

Insure your matrix inputs above provide a CR ratio < 0.1.

λ	**3.02784016**
CI	**0.01392008**
RI	**0.52**
CR=	**0.02676939**
	consistent

In this case the CR is 0.0267 < 0.1 so we are OK.

Repeat for all alternatives versus criterion

Go to Summary for results

25		
26	**Results**	
27		
28	**Alternatives**	**Values**
29	**Acura**	**0.07139135**
30	**Buick**	**0.18076036**
31	**C-Max**	**0.26791739**
32	**Escape**	**0.4799309**
33	**Car 5**	**0**
34	**Car 6**	**0**
35	**Car 7**	**0**
36	**Car 8**	**0**
37		

Put these in numerical value order: Escape #1, C-max #2, Buick #3, Acura #4.

MADM procedure that ranks alternatives based upon real and subjective criterion weights. You fill in the pairwise comparison. YOU MUST Have a CR less than or equal to 0.1 or GO back and change your pair-wise values.

The result is a ranking of alternatives based upon 100% total.

TOPSIS-template

If you need Decision weights, then use the AHP template to obtain the criterion weights. Transfer the values to TOPSIS.

Again we will allow up to 8 alternatives and up to 8 criterion for our template. TOPSIS ranks order alternatives based upon distances from an ideal solution. You may use either inputs weights or use the criterion weight methods from AHP. The values in TOPSIS are either real values (where bigger is better) or subjectively entered values where bigger is better. I think the template will be self-explanatory.

	A	B	C	D	E	F	G	H	I	J	K
1	**TOPSIS Example**										
2	Step 1. Enter the number of alternatives and number of criterion below:										
3					Valid						
4	Enter the number of alternatives				4 yes						
5	Enter the number of criterion				4 yes						
6	Step 2.										
7	Do you have your own criterion weights?				0 Enter a 0 or 1.						
8	1="yes", 0 = "no"										
9	Step 3. Either fill in row 12 with weights or go to Decision_criterion sheet and fill in yellow information.										
10	If you answered 1 enter your weights here										
11	c1	c2	c3	c4	c5	c6	c7	c8	Sum	Valid	
12	1	0	0	0	0	0	0	0	1	VALID	
13	0.054931	0.414559405	0.408603909	0.121906151	0	0	0	0	1	VALID	
14	If you answered No to criterion weights then go to Decision_Criterion worksheet and fill it ou										
15	Step 4. Fill in values for alternatives comapred to the other alternatives per criterion.										
16	Recommend using a 9 point scale										
17	1	2	3	4	5	6	7	8	9		
18	equal		moderate		strong		very strong		extreme		
19											
20	Alternati	Criterion	C1	C2	C3	C4	C5	C6	C7	C8	
21	A1		7	9	9	8	0	0	0	0	
22	A2		8	7	8	7	0	0	0	0	
23	A3		9	6	8	9	0	0	0	0	
24	A4		6	7	8	6	0	0	0	0	
25	A5		0	0	0	0	0	0	0	0	
26	A6		0	0	0	0	0	0	0	0	
27	A7		0	0	0	0	0	0	0	0	
28	A8		0	0	0	0	0	0	0	0	
29											

Enter into yellow spaces only.

Read your output in green.

Step 11. Rankings

Alternative	Larger better
1	0.83598049
2	0.32686258
4	0.31089873
3	0.22409112
5	0
6	0
7	0
8	0

Alternative #1 is best followed by alt #2, alt #4, and alt #3.

Lessons 9-13

CHAPTER 3
CLASSICAL PROBABILITY, DISTRIBUTIONS, AND HYPOTHESIS TESTING

A terrorist bomber explodes a bomb on a cruise ship in the Mediterranean Sea. The table below gives the results.

	Men	Women	Boys	Girls	Total
Survived	332	318	29	27	706
Died	1360	104	35	18	1517
Total	1692	422	64	45	2223

One rule of a disaster at sea is rescue woman and children first. Was this rule followed?

Some basic calculations, that we will show the formulas later, reveal that only 19.6% of the men (332 out of 1692 survived) and 70.4% of the woman and children survived (374 out of 531). Such simple calculations are powerful tools in analyzing information and providing insights into results.

3.1 Introduction to Classical Probability

Probability is a measure of the likelihood of a random phenomenon or chance behavior. Probability describes the long-term proportion with which a certain **outcome** will occur in situations with short-term uncertainty. Probability deals with experiments that yield random short-term results or outcomes yet reveal long-term predictability.
The long-term proportion with which a certain outcome is observed is the probability of that outcome.

The Law of Large Numbers

As the number of repetitions of a probability experiment increases, the proportion with which a certain outcome is observed gets closer to the probability of the outcome.
In probability, an **experiment** is any process that can be repeated in which the results are uncertain. A **simple event** is any single outcome from a probability experiment. Each simple event is denoted e_i. The **sample space, S,** of a probability experiment is the collection of all possible simple events. In other words, the sample space is a list of all possible outcomes of a probability experiment. An **event** is any collection of outcomes from a probability experiment. An event may consist of one or more simple events. Events are denoted using capital letters such as E.

Example 1: Consider the probability experiment of flipping a fair coin twice

(a) Identify the simple events of the probability experiment.
(b) Determine the sample space.
(c) Define the event E = "have only one head".

Solution:

(a) Events for two flips
H=head
T=tail
(b) Sample space {HH,HT,TH, TT }
(c) Having one head {HT,TH}

The **probability of an event**, denoted $P(E)$, is the likelihood of that event occurring.

Properties of Probabilities

1. The probability of any event E, $P(E)$, must be between 0 and 1 inclusive. That is,
$$0 \leq P(E) \leq 1.$$
2. If an event is **impossible**, the probability of the event is 0.
3. If an event is a **certainty,** the probability of the event is 1.
4. If $S = \{e_1, e_2, \ldots, e_n\}$, then
$$P(e_1) + P(e_2) + \ldots + P(e_n) = 1.$$

where S is the sample space and e_i are the events.

P(only One head in two flips)= Number of outcomes with only one head/ total number

of outcomes= 2/4=1/2

The classical method of computing probabilities requires *equally likely outcomes.*
An experiment is said to have **equally likely outcomes** when each simple event has the same probability of occurring. An example of this is a flip of a fair coin where the chance of flipping a head is ½ and the chance of flipping a tail is ½..

If an experiment has n equally likely simple events and if the number of ways that an event E can occur is m, then the probability of E, $P(E)$, is

$$P(E) = \frac{\text{Number of way that E can occur}}{\text{Number of Possible Outcomes}} = \frac{m}{n}$$

So, if S is the sample space of this experiment, then

$$P(E) = \frac{N(E)}{N(S)}$$

Example 2: Suppose a "fun size" bag of M&Ms contains 9 brown candies, 6 yellow candies, 7 red candies, 4 orange candies, 2 blue candies, and 2 green candies. Suppose that a candy is randomly selected.
(a) What is the probability that it is brown?
(b) What is the probability that it is blue?
(c) Comment on the likelihood of the candy being brown versus blue.
Solution:
(a) P(brown) = 9/30 = .3
(b) P(blue) = 2/30 = .066666
(c) Since there are more brown candies than blue candies, it is more likely to draw a brown candy than a blue candy.

Probability from Data

The probability of an event E is approximately the number of times event E is observed divided by the number of repetitions of the experiment.

$$P(E) \approx \text{ relative frequency of E}$$

$$= \frac{\text{frequency of E}}{\text{number of trials of experiment}}$$

Now, lets' return to our terrorist attack on the cruise ship. We can use this method to compute the probabilities.

	Men	Women	Boys	Girls	Total
Survived	332	318	29	27	706
Died	1360	104	35	18	1517
Total	1692	422	64	45	2223

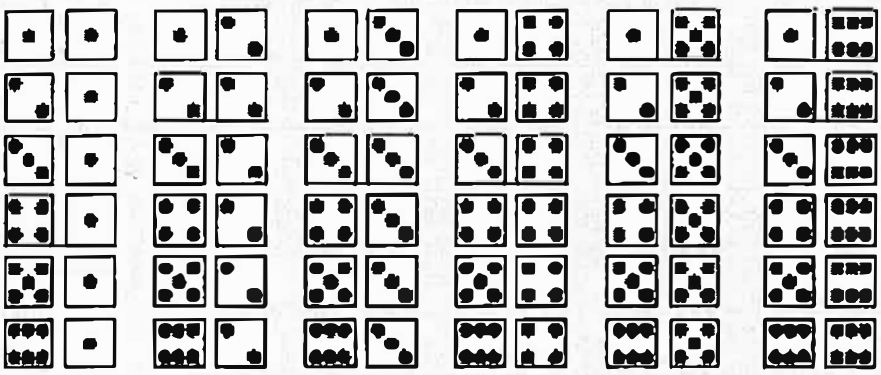
P(Survived the attack) = 706/2223 = 0.3176
P(Died)= 1517/2223 = 0.6824

P(Woman and children survived) = (318+29+27)/(422+64+45) = 374/531
P(Men survived) = 332/1692

Intersections and Unions

Now, let E and F be two events.
E* and *F is the event consisting of simple events that belong to *both* E and F. The notation is $\cap$ (intersection) , $\mathbf{E \cap F}$
E* or *F is the event consisting of simple events that belong to *either* E *or* F or both. The notation is $\cup$(union), $\mathbf{E \cup F}$.

Suppose that a pair of dice are thrown. Let E = "the first die is a two" and let F = "the sum of the dice is less than or equal to 5". Find $P(E \cap F)$ and P(E $\cup$ F) directly by counting the number of ways E or F could occur and dividing this result by the number of possible outcomes.



Event F = {1-1,1-2,1-3,1-4,2-1,2-2,2-3,3-1,3-2,4-1}

There are 36 outcomes above.
P(E) = 6/36 = 1/6
P(F) = 10/36=5/18

$(E \cap F)$ = {2-1, 2-2, 2-3}
(E $\cup$ F) = **{1-1,1-2,1-3,1-4,2-1,2-2,2-3,3-1,3-2,4-1,2-4,2-5, 2-6}**

$P(E \cap F)$ = 3/36 = 1/12
P(E $\cup$ F)= 13/36

The Addition Rule
For any two events E and F,

$$P(E \text{ or } F) = P(E) + P(F) - P(E \text{ and } F)$$

$$P(E \cup F) = P(E) + P(F) - P(E \cap F)$$

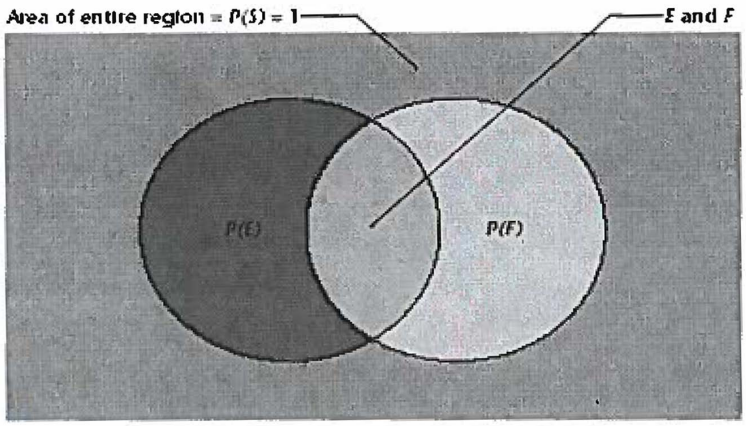
Let's consider the following example. Let event A be the event the a soldier on post takes the local newspaper and let event B be the event that a soldier on post take the USA today. There are 1,000 soldiers living on post and we know 750 take the local paper, and 500 take USA today. We are told 450 take both papers.
P(A $\cap$ B) = 450/1000 = .45
P(A) = .75
P(B) = .50

We can find the union, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$P(A \cup B) = .75+.50-.45 = .8$

Thus, 80% of the soldiers take at least one of the two newspapers.

Venn diagrams represent events as circles enclosed in a rectangle. The rectangle represents the sample space and each circle represents an event.

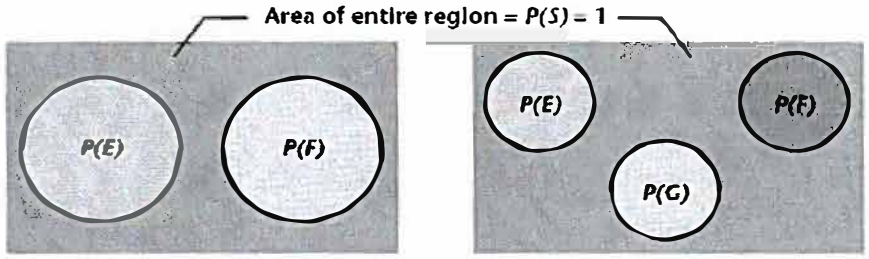


If events E and F have no simple events in common or cannot occur simultaneously, they are said to be **disjoint** or **mutually exclusive**, $E \cap F = \varnothing$ (the null set)

Addition Rule for Mutually Exclusive Events

If E and F are mutually exclusive events, then $P(E \text{ or } F) = P(E) + P(F)$. In general, if E, F, G, ... are mutually exclusive events, then

$$P(E \text{ or } F \text{ or } G \text{ or } \ldots) = P(E) + P(F) + P(G) + \ldots$$



Complement Rule

If E represents any event and $\bar{E}$ represents the complement of E then

$P(\bar{E}) = 1 - P(E)$

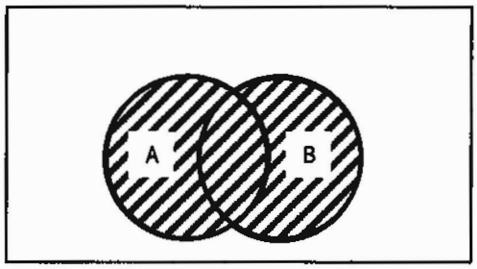
Complement of an Event

Let S denote the sample space of a probability experiment and let E denote an event. The **complement of E**, denoted $\bar{E}$, is all simple events in the sample space S that are not simple events in the event E.

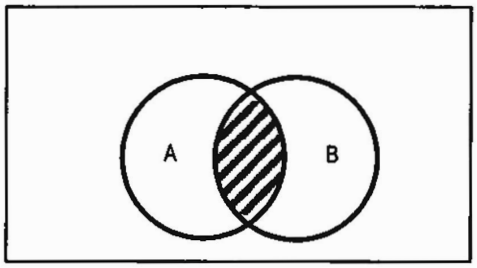
Example: Consider a roll of a single die

Let S = {1,2,3,4,5,6}
Let event A be the roll in an even number. A = { 2,4,6}

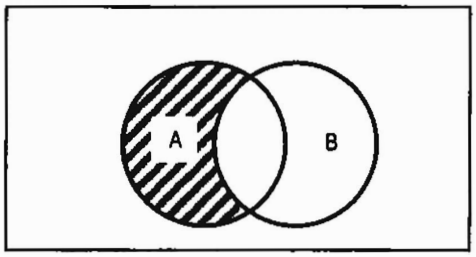
The event $\overline{A}$ would be {1,3,5}.
The following diagrams, called **VENN DIAGRAMS**, illustrate all the above set operations: complement, intersection, and union. Venn diagram are also very useful to find probabilities. Here sets are represented by simple plane areas and U, the universal set, by the area in the entire rectangle.



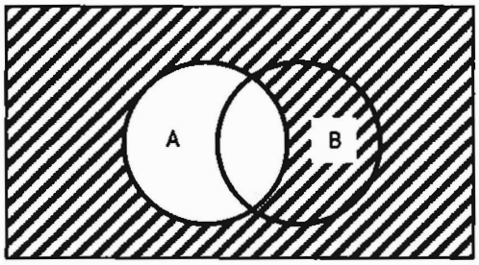
$A \cup B$ is shaded.



$A \cap B$ is shaded.

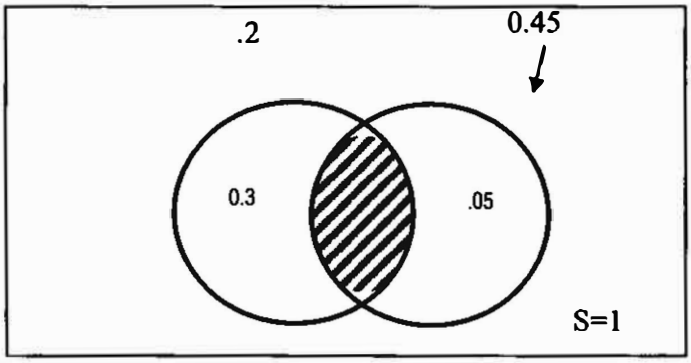


$A - B$ is shaded. This is only A



$\overline{A}$ is shaded. Everything that is NOT A is shaded.

Consider the newspaper example, the Venn Diagram would look like



The following probabilities can be used or found from the Venn Diagram. We always start filling in probabilities from inside the intersection of the events and move our way out. The sum total of all probabilities within the Venn Diagram rectangle, S, the sample set is 1.0.

P(A) = 0.75
P(B) = 0.5
P(A ∩ B) =0.45
P(A ∪ B) = P(A) + P(B)- P(A ∩ B)=0.8
P (only A) =0. 3
P(only B) =0.05

P(only take 1 paper) = P(only A) + P(only B) = .3 + .05 = 0.35
P(a soldier does not take a paper) = 0.2

P($\overline{A}$ **)** = 1-.75= 0.25

Conditional Probability

The notation $P(F \mid E)$ is read "the probability of event F given event E". It is the probability of an event F given the occurrence of the event E. The idea in a Venn Diagram here is if an event has happened then we only consider that circle of the Venn Diagram and we look for the portion of that circle that is intersected by another Event circle.

The Multiplication Rule

The probability that two events E and F both occur is

$$P(E \text{ and } F) = P(E) \cdot P(F \mid E)$$

In words, the probability of E and F is the probability of event E occurring times the probability of event F occurring given the occurrence of event E.

Think of this formula as

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

In most cases these conditional probabilities led to different probabilities as answers. Let's return to our newspaper example. Find the P(A|B) and P(B|A).

$P(A \cap B) = 0.45$

$P(A) = 0.75$

$P(B) = .5$

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{.45}{.50} = .9$$

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{.45}{.75} = .60$$

Notice that the probabilities increased as we obtained more information about the events occurring. The probabilities do not always increase, they could decrease, or remain the same. They do not have to be affected the same way.

Independence

Two events E and F are **independent** if the occurrence of event E in a probability experiment does not affect the probability of event F. Two events are **dependent** if the occurrence of event E in a probability experiment affects the probability of event F.

Definition of Independent Events

Two events E and F are independent if and only if

$$P(F \mid E) = P(F) \text{ or } P(E \mid F) = P(E)$$

Another way to see this is if

$P(A \cap B) = P(A) \cdot P(B)$ then the events A & B are independent.
If $P(A \cap B) \neq P(A) \cdot P(B)$ then the events are dependent.

Multiplication Rule for Independent Events

If E and F are independent events, the probability E and F both occur is

$$P(E \text{ and } F) = P(E) \cdot P(F)$$

In words, the probability of E and F is the probability of event E times the probability of event F.

Example: Are the events of getting the local newspaper and USA Today independent events?
Solution:
P(A) = 75 P(B) = .5
P(A) * P(B) = (.75)*(.5) = .375
P(A ∩ B) = .45
Since $P(A \cap B) \neq P(A) \cdot P(B)$ then these events are not independent.

Example: Given the following information:

P(E) = .2 P (F) = .6 P(E∪F) =
Are E and F independent events?
Solution:
P(E) * P(F) = .12

P(E∩F) is not given and must be found first. We do not assume independence and use the product rule. We use the addition rule where
P(E∪F)=P(A)+P(B)- P(E∩F) and solve for P(E∩F).
.68 = .2+.6-. P(E∩F)
P(E∩F) = .12
Since P(E∩F) = 12 and P(A)*P(B) = 12 then events E & F are independent.

Example: Suppose we have a box full of 500 golf balls. In the box, there are 50 Titlist golf balls.

Suppose a golf ball is selected at random and then replaced. A second golf ball is then selected. What is the probability they are both Titlists? NOTE: When sampling with replacement, the events are independent
Solution:
When selecting two golf balls, the following can occur: both are Titlists, both are other, one of each.
We assume independence so **P(both Titlists)= P(T ∩ T)=P(T)*P(T) = 0.1*0.1 =.01**

Conditional Probability Rule

If *E* and *F* are any two events, then

$$P(F \mid E) = \frac{P(E \text{ and } F)}{P(E)} = \frac{N(E \text{ and } F)}{N(E)}$$

The probability of event *F* occurring given the occurrence of event *E* is found by dividing the probability of *E* and *F* by the probability of *E*. Or, the probability of event *F* occurring given the occurrence of event *E* is found by dividing the number of simple events in *E* and *F* by the number of simple events in *E*.

Mutually exclusive and independent are not synonymous.

Review/Summary

Elementary probability theory is required for understanding this chapter in discrete *Stochastic Models*. We will provide a quick review of some important concepts in probability.
An **event** is any collection of results or outcomes of a procedure or experiment.
A **simple event** is an outcome that cannot be broken down into simpler components.
The **sample space** for a procedure or an experiment consists of all possible outcomes (as simple events).

Examples

Experiment	Example of an event	Sample Space
Flip of a coin	Head	{Head, Tail}
Roll of a die	5 (simple event)	{1,2,3,4,5,6}

Roll of two die 7 (not a simple event) {1-1,1-2,…,6-5,6-6}
Explanation: if we have only one die, we can roll a die to get a 5 only one way. If we have two die and roll a 7, it could be done as 1-6, 2-5, 3-4, 4-3, 5-2, 6-1 so it is not a simple event.

We define the **probability** that **event** A occurs, P(A), as the number of times A occurs out of the total number of possible outcomes in the **sample space:**

$$\mathbf{P(A)} = \frac{number\ of\ times\ A\ occurred}{number\ of\ events\ in\ the\ sample\ space}.$$

Sampling and Experiments

An event is any collection of results or outcomes of an experiment. A sample space is a listing of all possible outcomes from an **experiment**. An experiment is any process that allows researchers to obtain observations (data). In a flip of a fair coin 2 times the sample space are all the possible outcomes of 2 flips of a fair coin. If we call a head, H, and a tail, T, then the possible outcomes are:

HH HT TH TT

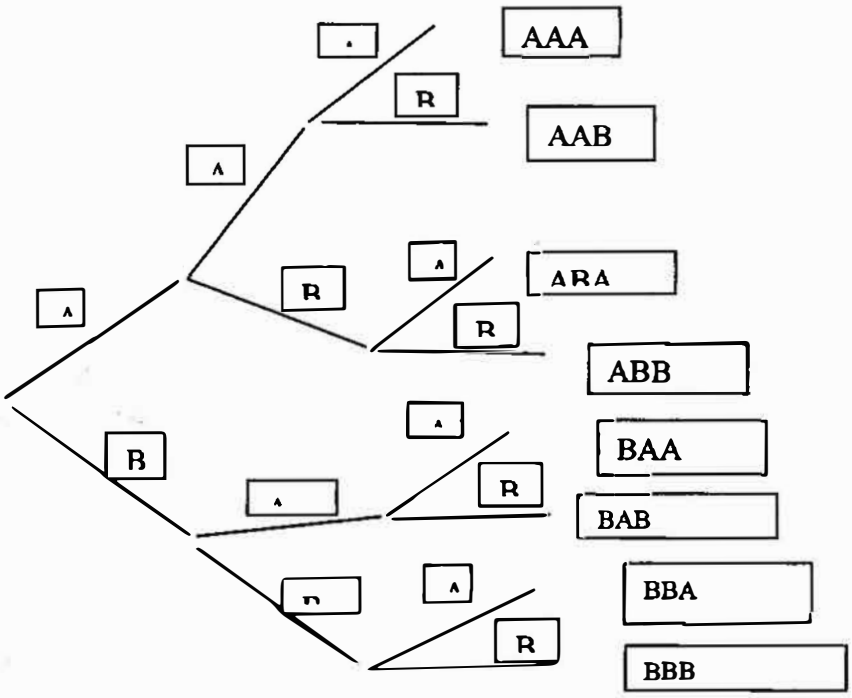
This set constitutes the entire sample space. Let's call event A-- the event that exactly one head appeared in the two flips. That occurred in flip HT and flip TH, or 2 times. Since there were four possible outcomes, the probability of event A, P(A), is 2/4 = 0.5.

Let's consider a tennis match between player A and player B where the winner is the first to win three sets. The sample space for the winner is:
{AAA, ABAA, ABBAA, AABA, AABBA, ABABA, BBB, BBAB, BAABB, BABAB, BAABA, BBAAB, BABB, BAAA, ABBB, ABABB, BABAA, BBAAA, AABBB, ABBAB}. If A and B were equally likely to win a set, then we can compute the probability of each event in the sample space. Thus, the probability that A wins is 0.5 or 10/20.

Tree Diagrams

Tree diagrams are a useful way to delineate the outcomes of a sample space. For example, consider wanting to find out what happened after only three sets of the match. Each branch of the tree signifies the winner of that set. For a three sets, we could have the following tree (note there are only two outcomes that show a winner at that time).

A



There are 8 simple events in 3 sets. Of the 3 sets, only two results in a identifying a winner. The events {*AAA*} and {*BBB*} are winners. All other outcomes must continue to a fourth or more sets. Thus, The probability that *A* wins in 3 sets is (1/8).
Show that the probability that *A* wins in 4 sets is 3(1/16) and in 5 sets 6(1/32).

Now let's assume that *A* is a higher ranked player with odds to win a set is 3:1. We can re-compute the probabilities that *A* wins from the given sample space in 3, 4 or 5 sets. Odds are a way of weighting the outcomes so that they are no longer equally likely.
P(A wins in 3 sets) = (0.75^3)= *0.421875*
P(A wins in 4 sets) = $3(.75^3)(.25)$= *0.3164065*
P(A wins in 5 sets) = $6(0.75^3)(.25^2)$= *0.158203125*
The probability that *A* wins this match, *P(A wins the match)*, is the sum of the probability that *A* wins in 3 sets + probability that *A* wins in 4 sets + probability that *A* wins in 5 sets = *0.421875+0.3164065+0.158203125 = 0.89648625.*

Review of Probability Laws

The are some important rules for probability. These rules that we will use, include the following:
1. The Law of Large Numbers states that if an experiment is repeated again and again, the relative frequency probability of an event approaches its probability. We will see this in our chapter on simulations.
2. Addition Rule: *P(A or B) = P(A) + P(B) - P(A and B)*
3. For any event *A* in the sample space, *S*,
   a. $P(A) \geq 0$
   b. $P(A) \leq 1$
   c. $P(S) = 1$

d. *P(not A) = 1 - P(A)*

4. For any events *A* & *B* in the sample space S
   a. If mutually exclusive, *P(A and B) = 0*
   b. If independent, *P(A and B) = P(A) * P(B)*
   c. Otherwise, *P(A and B) = P(A) + P(B) - P(A or B)*
5. Conditional probability: *P(A given B has occurred already) p(A|B) = P(A and B) / P(B)*

Bayes' Theorem

OBJECTIVES

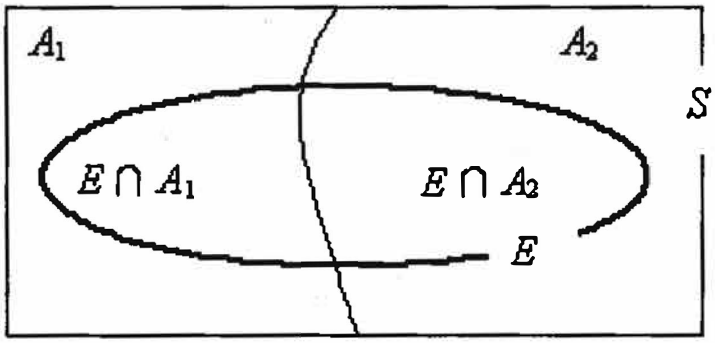
❶ Use the Theorem of Total Probability
❷ Use Bayes' Theorem to Compute Probabilities

❶ The Theorem of Total Probability

Theorem of Total Probability
Let E be an event that is a subset of a sample space S. Let $A_1, A_2, \ldots, A_n$ be a partition of the sample space, S. Then,

$$P(E) = P(A_1) \cdot P(E \mid A_1) + P(A_2) \cdot P(E \mid A_2) + \ldots + P(A_n) \cdot P(E \mid A_n)$$

FIGURE

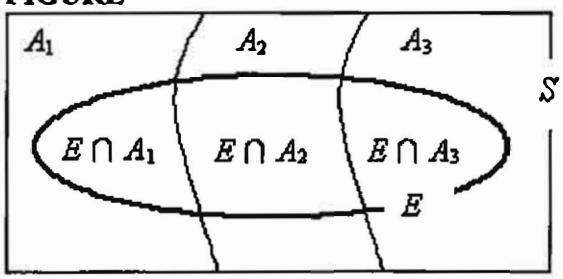


If we define E to be any event in the sample space S, then we can write event E as the union of the intersections of event E with A_1 and event E with A_2.

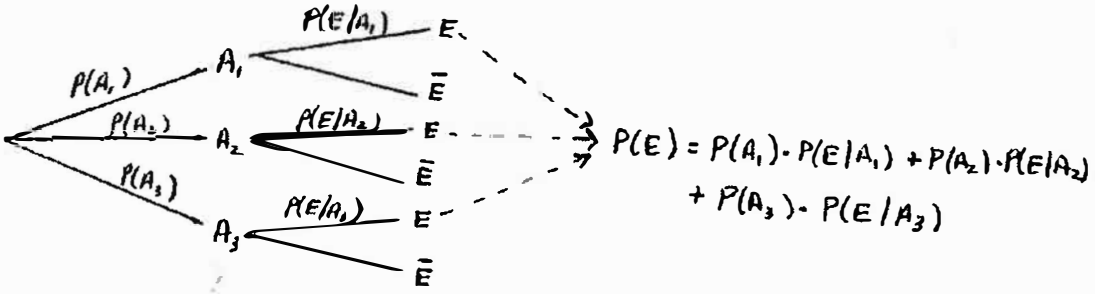
$$E = (E \cap A_1) \cup (E \cap A_2)$$

If we have more events, we just expand the union of the number of events that E intersects with as in the next Figure.

FIGURE



$$\begin{aligned} P(E) &= P(A_1 \cap E) + P(A_2 \cap E) + P(A_3 \cap E) \\ &= P(E \cap A_1) + P(E \cap A_2) + P(E \cap A_3) \\ &= P(A_1) \cdot P(E \mid A_1) + P(A_2) \cdot P(E \mid A_2) + P(A_3) \cdot P(E \mid A_3) \end{aligned}$$



❷ Bayes' Theorem

Bayes' Theorem
Let $A_1, A_2, \ldots, A_n$ be a partition of a sample space S. Then for any event E that is a subset of S for which $P(E) > 0$, the probability of event A_i for $i = 1, 2, \ldots, n$ given the event E, is

$$\begin{aligned} P(A_i \mid E) &= \frac{P(A_i) \cdot P(E \mid A_i)}{P(E)} \\ &= \frac{P(A_i) \cdot P(E \mid A_i)}{P(A_1) \cdot P(E \mid A_1) + P(A_2) \cdot P(E \mid A_2) + \ldots + P(A_n) \cdot P(E \mid A_n)} \end{aligned}$$

EXAMPLE ***Unemployed Women***

Problem: According to the United States Census Bureau 21.1% of American adult women are single, 57.6% of American adult women are married, and 21.3% of American adult women are widowed or divorced (other). Of the single women, 7.1% are unemployed; of the married women, 2.7% are unemployed; of the "other" women, 4.2% are unemployed. Suppose that a randomly selected American adult woman is determined to be unemployed. What is the probability that she is single?

Approach: Define the following events: U: unemployed
S: single

M: married
O: other

We have the following probabilities: $P(S) = 0.211$; $P(M) = 0.576$; $P(O) = 0.213$
$P(U \mid S) = 0.071$; $P(U \mid M) = 0.027$; $P(U \mid O) = 0.042$

and from the Theorem of Total probability , we know $P(U) = 0.039$

We wish to determine the probability that a woman is single given the knowledge that she is unemployed. That is, we wish to determine $P(S \mid U)$. We will use Bayes' Theorem as follows:

$$P(S \mid U) = \frac{P(S \cap U)}{P(U)} = \frac{P(S) \cdot P(U \mid S)}{P(U)}$$

Solution:
$$P(S \mid U) = \frac{0.211(0.071)}{0.039} = 0.384$$

There is a 38.4% probability that a randomly selected unemployed woman is single.

■

In Words
A priori probabilities are probabilities computed prior to any knowledge. *A posteriori* probabilities are computed after gaining some knowledge.

We say that all the probabilities $P(A_i)$ are *a priori* probabilities. These are probabilities of events prior to any knowledge regarding the event. However, the probabilities $P(A_i \mid E)$ are *a posteriori* probabilities because they are probabilities computed after some knowledge regarding the event. In our example, the *a priori* probability of a randomly selected woman being single is 0.211. The *a posteriori* probability of a woman being single knowing that she is unemployed is 0.384. Notice the information that Bayes' Theorem gives us. Without any knowledge of the employment status of the woman, there is a 21.1% probability that she is single. But, with the knowledge that the woman is unemployed, the likelihood of her being single increases to 38.4%.

Let's do one more example.

EXAMPLE ***Work Disability***

Problem: A person is classified as work disabled if they have a health problem that prevents them from working in the type of work they can do. Table 1 contains the proportion of
Americans that are 16 years of age or older that are work disabled by age.

TABLE 1

Age	Event	Proportion Work Disabled
16 – 24	A_1	0.078
25 – 34	A_2	0.123
35 – 44	A_3	0.209
45 – 54	A_4	0.284
55 and older	A_5	0.306

Source: United States Census Bureau

If we let M represent the event that a randomly selected American who is 16 years of age or older is male, then we can also obtain the following probabilities:

$P(\text{male} \mid 16 - 24) = P(M \mid A_1) = 0.471$ $P(\text{male} \mid 25 - 34) = P(M \mid A_2) = 0.496$

$P(\text{male} \mid 35 - 44) = P(M \mid A_3) = 0.485$ $P(\text{male} \mid 45 - 54) = P(M \mid A_4) = 0.497$

$P(\text{male} \mid 55 \text{ and older}) = P(M \mid A_5) = 0.460$

(a) If a work disabled American aged 16 years of age or older is randomly selected, what is the probability that the American is male?

(b) If the work disabled American that is randomly selected is male, what is the probability that he is 25 – 34 years of age?

Approach:

(a) We will use the Theorem of Total Probability to compute $P(M)$ as follows:

$$P(M) = P(A_1) \cdot P(M \mid A_1) + P(A_2) \cdot P(M \mid A_2) + P(A_3) \cdot P(M \mid A_3) + P(A_4) \cdot P(M \mid A_4) + P(A_5) \cdot P(M \mid A_5)$$

(b) We use Bayes' Theorem to compute $P(25 - 34 \mid \text{male})$ as follows:

$P(A_2 \mid M) = \dfrac{P(A_2) \cdot P(M \mid A_2)}{P(M)}$ where $P(M)$ is found from part (a).

Solution:

(a)

$$P(M) = P(A_1) \cdot P(M \mid A_1) + P(A_2) \cdot P(M \mid A_2) + P(A_3) \cdot P(M \mid A_3) + P(A_4) \cdot P(M \mid A_4) + P(A_5) \cdot P(M \mid A_5)$$

$= (0.078)(0.471) + (0.123)(0.496) + (0.209)(0.485) + (0.284)(0.497) + (0.306)(0.460)$

$= 0.481$

There is a 48.1% probability that a randomly selected work disabled American is male.

(b) $P(A_2 \mid E) = \dfrac{P(A_2) \cdot P(E \mid A_2)}{P(E)} = \dfrac{0.123(0.496)}{0.481} = 0.127$

There is a 12.7% probability that a randomly selected work disabled American who is male is 25 – 34 years of age.

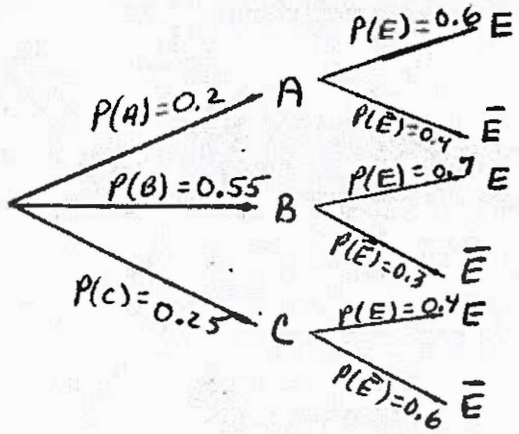
Notice that the *a priori* probability (0.123) and the *a posteriori* probability (0.127) do not differ much. This means that the knowledge that the individual is male does not yield much information regarding the age of the work disabled individual.

■

Bayes' Theorem Exercises

- **Basic Skills**

In Problems 1 – 14, find the indicated probabilities by referring to the tree diagram given below and by using Bayes' Theorem.



1. $P(E \mid A)$ 2. $P(E \mid B)$ 3. $P(\overline{E} \mid A)$ 4. $P(\overline{E} \mid B)$ 5. $P(E \mid C)$
6. $P(\overline{E} \mid C)$ 7. $P(E)$ 8. $P(\overline{E})$ 9. $P(A \mid E)$ 10. $P(A \mid \overline{E})$
11. $P(C \mid E)$ 12. $P(B \mid \overline{E})$ 13. $P(B \mid E)$ 14. $P(C \mid \overline{E})$

15. Suppose that events A_1 and A_2 form a partition of the sample space S with $P(A_1) = 0.55$ and $P(A_2) = 0.45$. If E is an event that is a subset of S and $P(E \mid A_1) = 0.06$ and $P(E \mid A_2) = 0.08$, find $P(E)$.

16. Suppose that events A_1 and A_2 form a partition of the sample space S with $P(A_1) = 0.35$ and $P(A_2) = 0.65$. If E is an event that is a subset of S and $P(E \mid A_1) = 0.12$ and $P(E \mid A_2) = 0.09$, find $P(E)$.

17. Suppose that events A_1, A_2, and A_3 form a partition of the sample space S with $P(A_1) = 0.35$, $P(A_2) = 0.45$, and $P(A_3) = 0.2$. If E is an event that is a subset of S and $P(E \mid A_1) = 0.25$, $P(E \mid A_2) = 0.18$, and $P(E \mid A_3) = 0.14$, find $P(E)$.

18. Suppose that events A_1, A_2, and A_3 form a partition of the sample space S with $P(A_1) = 0.3$, $P(A_2) = 0.65$, and $P(A_3) = 0.05$. If E is an event that is a subset of S and $P(E \mid A_1) = 0.05$, $P(E \mid A_2) = 0.25$, and $P(E \mid A_3) = 0.5$, find $P(E)$.

- **Applying the Concepts**

19. **Color Blindness** The most common form of colorblindness is so-called "red-green" colorblindness. People with this type of colorblindness cannot distinguish between green and red. Approximately 8% of all males have red-green colorblindness, while only about 0.64% of women have red-green colorblindness. In 2000, 49.1% of all Americans were male and 50.9% were female according to the United States Census Bureau.
(a) What is the probability that a randomly selected American is colorblind?
(b) What is the probability that a randomly selected American who is colorblind is female?

20. **The Elias Test** The standard test for the HIV virus is the Elias test that tests for the presence of HIV antibodies. If an individual does not have the HIV virus, the test will come back negative for the presence of HIV antibodies 99.8% of the time and will come back positive for the presence of HIV antibodies 0.2% of the time (a false positive). If an individual has the HIV virus, the test will come back positive 99.8% of the time and will come back negative 0.2% of the time (a false negative). The latest reports available indicate that approximately 0.7% of the world population has the HIV virus.
(a) What is the probability that a randomly selected individual has a test that comes back positive?
(b) What is the probability that a randomly selected individual has the HIV virus if the test comes back positive?

21. **Educational Attainment** The data in the following table represent the proportion of Americans 25 years of age or older at the various levels of educational attainment in 2000.

Level	Event	Proportion
Not a High School Graduate	A_1	0.158
High School Graduate	A_2	0.331
Some College, No Degree	A_3	0.176
Associate's Degree	A_4	0.078
Bachelor's	A_5	0.171

Degree Advanced Degree	A_6	0.086

Source: United States Census Bureau

If we let M represent the event that a randomly selected American who is 25 years of age or older is male, then we can also obtain the following probabilities:

$P(M \mid A_1) = 0.477$ $P(M \mid A_2) = 0.460$ $P(M \mid A_3) = 0.472$
$P(M \mid A_4) = 0.434$ $P(M \mid A_5) = 0.500$ $P(M \mid A_6) = 0.555$

(a) What is the probability that a randomly selected American 25 years of age or older is male?
(b) What is the probability that an American male 25 years of age or older has an advanced degree?
(c) What is the probability that an American male 25 years of age or older is not a high school graduate?

22. **Educational Attainment** Refer to Problem 29. If we let E represent the event that a randomly selected American who is 25 years of age or older is employed, then we can also obtain the following probabilities:

$P(E \mid A_1) = 0.402$ $P(E \mid A_2) = 0..628$ $P(E \mid A_3) = 0.699$
$P(E \mid A_4) = 0.755$ $P(E \mid A_5) = 0.785$ $P(E \mid A_6) = 0.791$

(a) What is the probability that a randomly selected American 25 years of age or older is employed?
(b) What is the probability that an employed American 25 years of age or older has a Bachelor's degree?
(c) What is the probability that an employed American 25 years of age or older is not a high school graduate?

23. **Voting Pattern** The following data represent the proportion of Americans who are voting age at the various levels of educational attainment in 2000.

Level	Event	Proportion
Grade School	A_1	0.163
High School Graduate	A_2	0.600
College Graduate	A_3	0.237

Source: Statistical Abstract, 2002

If we let D represent the event that a randomly selected American who is voting age voted Democratic in the 2000 Presidential election, then we can also obtain the following probabilities:

$P(D \mid A_1) = 0.74$ $P(D \mid A_2) = 0.540$ $P(D \mid A_3) = 0.500$

(a) What is the probability that a randomly selected American who is voting age voted Democratic in the 2000 Presidential election?
(b) What is the probability that an American who is voting age and voted Democratic has graduated from college?
(c) What is the probability that an American who is voting age and voted Democratic has a grade school education?

24. **Murder Victims** The following data represent the proportion of murder victims at the various age levels in 2000.

Level	Event	Proportion
Less than 17 years	A_1	0.082
17 – 29	A_2	0.424
30 – 44	A_3	0.305
45 – 59	A_4	0.125
At least 60 years	A_5	0.064

Source: Federal Bureau of Investigation

If we let M represent the event that a randomly selected murder victim was male, then we can also obtain the following probabilities:

$P(M \mid A_1) = 0.622$ $P(M \mid A_2) = 0.843$ $P(M \mid A_3) = 0.733$
$P(M \mid A_4) = 0.730$ $P(M \mid A_5) = 0.577$

(a) What is the probability that a randomly selected murder victim was male?
(b) What is the probability that a randomly selected male murder victim was 17 – 29 years of age?
(c) What is the probability that a randomly selected male murder victim was less than 17 years of age?

25. **Espionage** Suppose that the CIA suspects that one of its operatives is a double agent. Past experience indicates that 95% of all operatives suspected of espionage are, in fact, guilty. The CIA decides to administer a polygraph to the suspected spy. It is known that the polygraph returns results that indicate a person is guilty 90% of the time if they are guilty. The polygraph returns results that indicate a person is innocent 99% of the time if they are innocent. What is the probability that this particular suspect is innocent given that the polygraph indicates that he is guilty?

6.2 Probability Distributions

6.2.1 Discrete DISTRIBUTIONS in Modeling

We will also use several probability distributions for discrete random variables. A random variable is a rule that assigns a number to every outcome of a sample space. A discrete random variable takes on counting numbers 0,1,2,3,...etc. These are either finite or countable. Then, a probability distribution gives the probability for each value of the random variable.

Let's return to our coin flipping example earlier. Let the random variable F be the number of heads of the two flips of the coin. The possible values of the random variable F are 0, 1, and 2. We can count the number of outcomes that fall into each category of F as shown in the *probability mass function* table below.

Random Variable	0	1	2
Occurrences	1	2	1
Corresponding to events	TT	TH,HT	HH
P(F)	¼	2/4	1/4

Note that the $\Sigma P(F) = 1/4 + 2/4 + 1/4 = 1$. This is a rule for any probability distribution. Let's summarize these rules:

1. P(each event) ≥ 0
2. Σ P(events) = 1

Thus, the coin flip experiment is a probability distribution.

All probability distributions have means, μ, and variances, σ^2. We can find the mean and the variance for a random variable X using the following formulas:

$$\mu = E[X] = \Sigma \, x \, P(X=x)$$

$$\sigma^2 = E[X^2]-(E[X])^2$$

For our example, we compute the mean and variance as follows:

$\mu = E[X] = \Sigma \, x \, P(X=x) = 0(1/4)+1(2/4)+2(1/4) = 1$
$\sigma^2 = E[X^2]-(E[X])^2 = 0(1/4) + 1(2/4) + 4(1/4) - 1^2 = .5$
We can also find the standard deviation, σ.

$$\sigma = \sqrt{\sigma^2}$$

Thus, we find the variance first and then take its square root.

$\sigma = \sqrt{.5}$

There will be several discrete distributions that will arise in our modeling: Bernoulli, Binomial and Poisson.

Consider an experiment made up of a repeated number of independent and identical trials having only two outcomes, like tossing a fair coin {Head, Tail}, or a {red, green} stoplight. These experiments with only two possible outcomes are called *Bernoulli trials*. Often they are found by assigning either a S (success) or F (failure) or a 0 or 1 to an outcome. Something either happened (1) or did not happen (0).

A binomial experiment is found counting the number of successes in N trials.

Binomial experiment:
(a) Consists of n trials where n is fixed in advance.
(b) Trials are identical and can result in either a success or a failure.
(c) Trials are independent.
(d) Probability of success is constant from trial to trail.

Formula: $b(x; n, p) = p(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$ 0for $x = 0,1,2,\ldots n$

Cumulative Binomial: $p(X \leq x) = B(x; n, ,p) = \sum_{y=0}^{x} \binom{n}{y} p^y (1-p)^{n-y}$ 0 for $x = 01,2,\ldots n$

Mean: $\mu = n\,p$

Variance: $\sigma^2 = n\,p\,(1\text{-}p)$

Example, our coin flip experiment follows these rules and is a binomial experiment. The probability that we got 1 heads in 2 flips is:

$P(X=1) = \binom{2}{1} .5^1 (1-.5)^{2-1} = .50$

If we wanted 5 heads in 10 flips of a fair coin then we can compute:

$$P(X=5) = \binom{10}{5}.5^5(1-.5)^{10-5} = 0.2461$$

LIGHT BULBS:

Light bulbs are manufactured in a small local plant. In testing the light bulbs, prior to packaging and shipping, they either work, S, or fail to work, F. The company cannot test all the light bulbs but does test a random batch of 100 light bulbs per hour. In this batch, they found 2 % that did not work but all batches were shipped to distributors.
As a distributor, you are worried about past performance of these lights bulbs that you sell individually off the shelf. If a customer buys 20 lights bulbs, what is the probability that all work?
Problem ID: Predict the probability that x lights bulbs out of N work.

Assumptions: The lights bulbs follow the binomial distribution rules stated earlier.

Model: Formula: $b(x; n, p) = p(X = x) = \binom{n}{x}p^x(1-p)^{n-x}$ for $x = 0,1,2,\ldots n$

We can use EXCEL. The following is from the help page in Excel.

We show the use of Excel below:

Binomial Distribution In EXCEL

BINOMDIST(20,20,.998,false)
0.960751

To determine the probability that $x \leq 10$ given n=20 and P(s) = .50 we would use the cumulative distribution, BINOMDIST(10,20,..50, True).

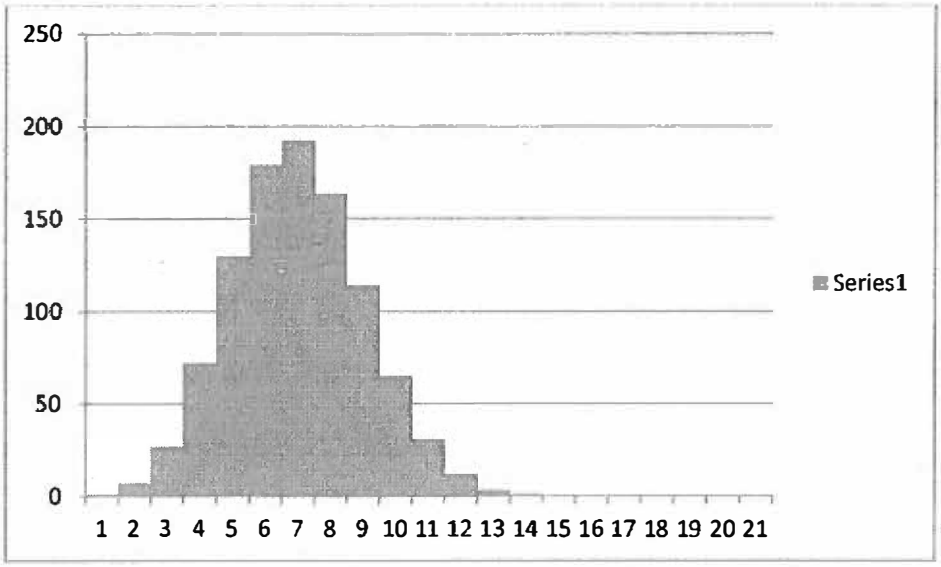
0.588098526

To find P(X>16), we need to know P(X≤15), so

=1-
BINOMDIST(15,20,.5),

would give us
0.005908966

If we have discrete data that follows a binomial distribution then its histogram might look as follows:



It is symmetric. The keys are the assumptions for the binomial as well as it being discrete.

Example: A weapon has a 93% accuracy on average. If we fire 10 shots at a target, what is the probability that we hit the target 5 times, at most 5 times, at least 5 times?

Solution: This is a binomial distribution because shots are fired independently, the probability of success is known (93%), and we know in advance the number of shots fired, n (10 shots fired).

First, we use Excel to generate the PDF and the CDF given below.

n	PDF	CDF
0	0.0000	0.0000
1	0.0000	0.0000
2	0.0000	0.0000
3	0.0000	0.0000
4	0.0000	0.0000
5	0.0003	0.0003
6	0.0033	0.0036
7	0.0248	0.0283
8	0.1234	0.1517
9	0.3643	0.5160
10	0.4840	1.0000

a) *P(X=5).* This is a pdf value that we extract from n=5, under pdf. The value is 0.0003. *P(X=5)=0.0003.* Interpretation: If we fired 10 shots at a target the probability the exactly 5 of the 10 hit the target is 0.0003.

b) *At most 5 hit the target → P(X≤5). This is a CDF value that we extract from n=5 since we include 5 under the CDF, P(X≤5)=0.0003. Interpretation. If we fire 10 shots at a target the probability the 5 or fewer hit the target is P(X≤5)=0.0003.*

c) *At least 5 hit the target → P(X≥5). This is NOT one of our known forms. We must convert the probability to its complement. P(X≥5) = 1 – P(X<5) = 1 –P(X≤4). We obtain P(X≤4) from the cdf table and obtain 0.000 (to four decimal places). 1-0.0000=1. Interpretation, we expect 5 or more rounds to hit the target with probability of 1.*

Poisson Distribution:

A random variable is said to have a Poisson Distribution if the probability distribution function of X is:

$$p(x;\lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$ 0, for x =0,1,2,3… for some $\lambda > 0$.

We consider λ as a *rate per unit time or per unit area*. A key assumption is that with a Poisson distribution the mean and the variance are the same.

For example, let X represent the number of minor flaws on the surface of a randomly selected F-16. It has been found that on average, 5 flaws are found per F-16 surface. Find the probability that a randomly selected-16 has exactly 2 flaws.

POISSON

Returns the Poisson distribution. A common application of the Poisson distribution is predicting the number of events over a specific time, such as the number of cars arriving at a toll plaza in 1 minute.

Syntax

POISSON(x,mean,cumulative)

X is the number of events.

Mean is the expected numeric value.

Cumulative is a logical value that determines the form of the probability distribution returned. If cumulative is TRUE, POISSON returns the cumulative Poisson probability that the number of random events occurring will be between zero and x inclusive; if FALSE, it returns the Poisson probability mass function that the number of events occurring will be exactly x.

Remarks

If x is not an integer, it is truncated.
If x or mean is nonnumeric, POISSON returns the #VALUE! error value.
If x < 0, POISSON returns the #NUM! error value.
If mean < 0, POISSON returns the #NUM! error value.
POISSON is calculated as follows.

For cumulative = FALSE:

$$POISSON = \frac{e^{-\lambda}\,\lambda^{x}}{x!}$$

For cumulative = TRUE:

$$CUMPOISSON = \sum_{k=0}^{x} \frac{e^{-\lambda} \lambda^{k}}{k!}$$

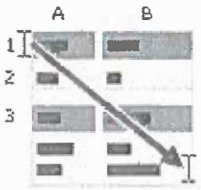
Example

The example may be easier to understand if you copy it to a blank worksheet.

How to copy an example

1. Create a blank workbook or worksheet.
2. Select the example in the Help topic.

   NOTE Do not select the row or column headers.



   Selecting an example from Help

3. Press CTRL+C.
4. In the worksheet, select cell A1, and press CTRL+V.
5. To switch between viewing the results and viewing the formulas that return the results, press CTRL+` (grave accent), or on the **Formulas** tab, in the **Formula Auditing** group, click the **Show Formulas** button.

	A	B
1	**Data**	**Description**
2	2	Number of events
3	5	Expected mean
	Formula	**Description (Result)**
	=POISSON(A2,A3,TRUE)	Cumulative Poisson probability with the terms above (0.124652)
	=POISSON(A2,A3,FALSE)	Poisson probability mass function with the terms above (0.084224

POISSON(2,5,false)
0.084224337

$$p(X=2)=\frac{e^{-5}5^2}{2!}=.084$$

A ***Poisson distribution*** has a

mean, μ, of λ and

variance σ^2 of λ.

A ***Poisson process*** is a Poisson distribution that varies over time (generally its time). There exists a rate, called α for a short time period. Over a longer period of time λ becomes αt.
Here is an example:
Suppose, your pulse is read by an electronic machine at a rate of five times per minute. Find the probability that your pulse is read 15 times in a 4 minute interval.

$\lambda = \alpha t =$ 5 times 4 minutes = 20 pulses in a 4 minute period

$$p(X=15)=\frac{e^{-20}20^{15}}{15!}=.052$$

Poisson(15,20,false)
= 0.051648854

Poisson data usually at least is slightly positively skewed.

EXERCISES

1. If 75% of all purchases at Wal-Mart are made with a credit card and X is the number among ten randomly selected purchases made with a credit card, then find the following:

a. p(X=5)

b. $p(X \leq 5)$

c. μ, and σ^2

2. Russell Stover's produces fine chocolates and its known from experience that 10% of its chocolate boxes have flaws and must be classified as "seconds."

a. Among six randomly selected chocolate boxes, how likely is it that one is a second?

b. Among the six randomly selected boxes, what is the probability that at least two are seconds?

c. What is the mean and variance for "seconds?"

3. Consider the following TV ad for an exercise program: 17% of the participants lose 3 pounds, 34% lose 5 pounds, 28% lose 6 pounds, 12% lose 8 pounds, and 9 % lose 10 pounds. Let X = the number of pounds lost on the program.

   a. Give the probability mass function of X in a table.

   b. What is the probability that the number of pounds lost is at most 6? At least 6?

   c. What is the probability that the number of pounds lost is between 6 and 10?

   d. What are the values of μ and σ^2?

4. A machine fails on average 0.4 times a month (30 consecutive days). Determine the probability that there are 10 failures in the next year.

Projects

1. Iran Hostage Rescue Attempt. In 1979, President Carter authorized an attempt to rescue American hostages held in Iran. DoD estimated that at least 6 helicopters would have to complete the mission successfully, but that the total number of helicopters needed to be kept as small as possible for security reasons. Each helicopter was believed to have a 95% chance of completing the mission (based on historical maintenance records). DoD used 8 helicopters. Three helicopters failed so the mission was aborted. Defend the use of the Poisson distribution over the Binomial distribution. Determine the minimum number of helicopters necessary to have successfully completed the mission.

2. Military Aircraft Accidents. In a 7-day period in September 1997, 6 military aircraft crashed, prompting the Secretary of Defense to suspend all training flights. There were 277 crashes in the previous 4 years. Show that this is a rare event. Was there anything special about this week (7-day period) other than the 6 crashes? How many 7-day period could have occurred in a 4 year period? What should the Secretary of Defense done in this matter? Make some recommendations based upon sound probability analysis.

6.2.3 Continuous Probability Models

Introduction

Some random variables do not have a discrete range of values. In the previous chapter, we saw examples of discrete random variables and discrete distributions. What if we were looking at time, as a random event? Time has a continuous range of values and thus, as a continuous random variable can be continuous probability distribution. We define a continuous random variable as any random variable measured on continuous scale. Other examples include altitude of a plane, the percent of alcohol in a person's blood, net weight of a package of frozen chicken wings, the distance a round misses a designated target, or the time to failure of an electric light bulb. We cannot list the sample space because the sample space is infinite. We need to be able to define a distribution as well as its domain and range.

For any continuous random variable, we can define the cumulative distribution function (CDF) as

$$F(b) = P(X \le b).$$

For those that have seen calculus, the probability density function (PDF) of $f(x)$ is defined to be

$$P(a \le x \le b) = \int_a^b f(x)dx.$$

To be a valid probability density function (PDF):

a) $f(x)$ must be greater than or equal to zero for all x in its domain, and

b) the integral $\int_{-\infty}^{\infty} x \cdot f(x)dx = 1$ = the area under the entire graph of $f(x)$.

Expected value or average value of a random variable x, with PDF defined as above, is defined as $E[X] = \int_{-\infty}^{\infty} x \cdot f(x)dx$.

In this chapter, we will see some modeling applications using many continuous distributions such as the exponential distribution and the normal distribution. For each of these two distributions, we will not have to use calculus to get our answers to probability questions.

Since we do not require calculus, we will discuss only a few of these distributions that we obtain results with Excel.

The Normal Distribution

A continuous random variable X is said to have a normal distribution with parameters μ and σ (or μ and σ^2), where $-\infty < \mu < \infty$ and $\sigma > 0$, if the pdf of X is

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{(2\sigma^2)}}, -\infty \le x \le \infty$$

The plot of the normal distribution is our bell-shaped curve, see figure 3.3 below.

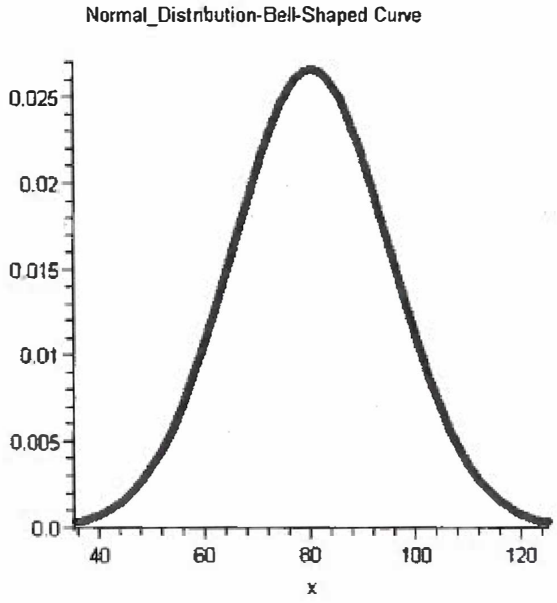


Figure 3.3 Bell-Shaped Curve of the Normal Distribution

To compute P(a < x < b) when X is a normal random variable, with parameters μ and σ, we must evaluate $\int \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{(2\sigma^2)}} dx$.

Since none of the standard integration techniques can be used to evaluate this integral, the standard normal random variable Z with parameters $\mu = 0$ and $\sigma = 1$ has been numerically evaluated and tabulated for certain values. Since most applied problems do not have parameters of $\mu = 0$ and $\sigma = 1$, "standardizing" transformation can be used $Z = \frac{x-\mu}{\sigma}$.

For example, the amount of fluid dispensed into a can of diet coke is approximately a normal random variable with mean 11.5 fluid ounces and a standard deviation of 0.5 fluid ounces. We want to determine the probability that between 11 and 12 fluid ounces, P(11 < x < 12) are dispensed.

$$Z_1 = (11 - 11.5)/.5 = -1$$

$$Z_2 = (12 - 11.5)/.5 = 1$$

This probability statement P(11 < x < 12) is equivalent to P(-1 < Z < 1). If we used the tables, we can compute this to be 0.8413 – 0.1587 = 0.6826. However, we can use Excel to compute the area between 11 and 12. This is displayed in the figure below.

Excel Command is **Normdist(value, mean, standard deviation, TRUE)**

To find the P(a < x < b) or in our case P(11< x < 12)

=Normdist(12,11.5,.5,True)-Normdist(11,11.5,.5,True)

Find Normal Probability	
P(a < X < b)	
a	11
b	12
mean	11.5
standard dev	0.5
`True`	Yes

Probability is 0.682689

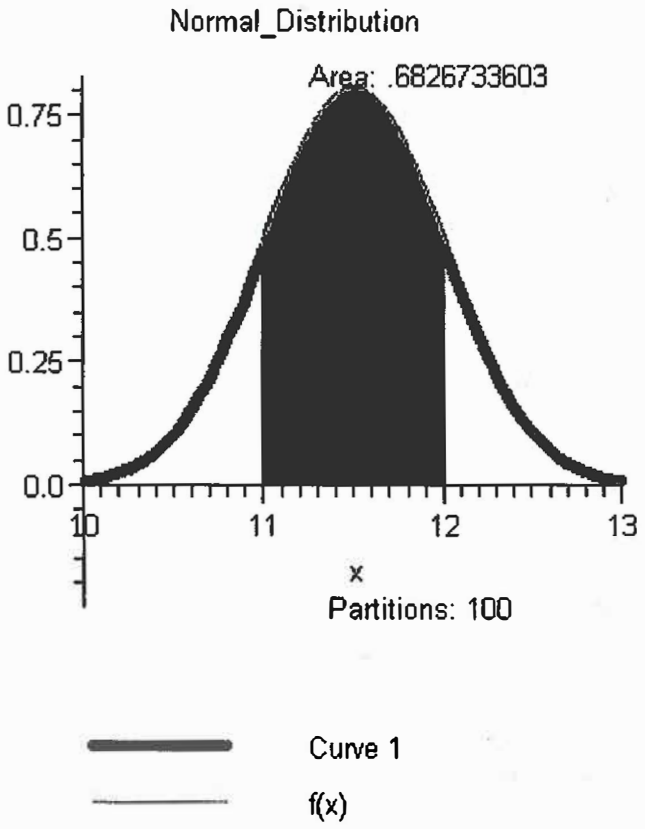


Figure 3.4 Normal Distribution area from 11 to 12.

Therefore, 68.26% of the time the cans are filled between 11 and 12 fluid ounces.

CENTRAL LIMIT THEOREM: looking at average or total results not individual random variables.

The Central Limit Theorem is one of the most important theorems in probability. It states that if $X_1, X_2, \ldots X_n$ are a random sample from a distribution with a mean μ and a standard deviation σ and n is sufficiently large $(\mathbf{n > 30})$ then the distribution of the average $\overline{X}$ or TT(The total) has a normal distributions with parameters:

$$\mu_{\overline{X}} = \mu \qquad \sigma^2_{\overline{X}} = \frac{\sigma^2}{n} \qquad \mu_T = n \cdot \mu \qquad \sigma^2_T = n \cdot \sigma$$

For example, when a batch of a certain pharmaceutical is prepared the amount of natural substance aloe is a random variable with mean value 4.0 grams and standard deviation 0.35 grams. If 50 batches are prepared, what is the probability that the sample average of the aloe is between 3.5 and 3.8 grams?

Since n = 50(n > 30), then the sample average aloe random variable, $\overline{X}$, follows a normal distribution with mean 4.0 and standard deviation, $\frac{0.35}{\sqrt{50}} = 0.4950$.

$$P(3.5 < \overline{X} < 3.8) = P\left(\frac{(3.5-4.0)}{0.4950} < Z < \frac{(3.8-4)}{0.4950}\right) = 0.1869$$

Using our Excel commands, we must take the difference between the two calculated values in order to get the area between the 2 points. Draw it.
=Normdist(3.8,4,.4950, True)-Normdist(3.5,4,.4950, True)

Find Normal Probability	
P(a < X < b)	
a	3.5
b	3.8
mean	4
standard dev	0.495
`True`	Yes
Probility is	0.186868

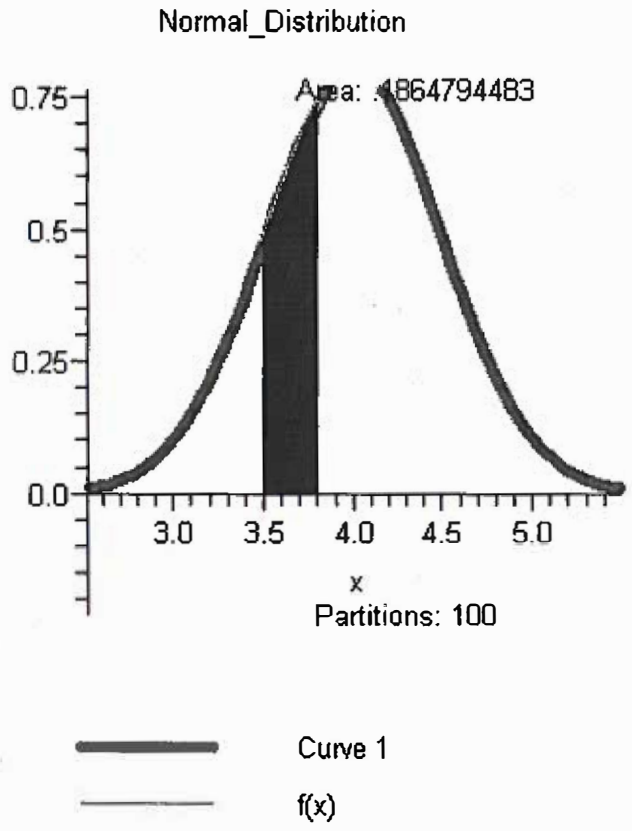


Figure 3.5 Normal Curve with mean = 4, standard deviation = 0.4950 from 3.5 to 3.8.

The normal distribution and the central limit (when applicable) is used in many applications of confidence intervals and hypothesis testing.

EXERCISES.

Find the following probabilities:

1. $X \sim N(\mu = 10, \sigma = 2)$, $P(X > 6)$
2. $X \sim N(\mu = 10, \sigma = 2)$, $P(6 < x < 14)$

3. Determine the probability that lies within one standard deviation of the mean, two standard deviations of the mean, and three standard deviations of the mean. Draw a sketch of each region.

4. A tire manufacturer thinks that the amount of wear per normal driving year of the rubber used in their tire follows a normal distribution with mean = 0.05 inches and standard deviation 0.05 inches. If 0.10 inches is considered dangerous, then determine the probability that P(X > 0.10)

6.4 Confidence Intervals (optional) and Hypothesis Testing

The basic concepts and properties of confidence intervals involve initially understanding and using two assumptions:

1) the population distribution is normal and
2) the standard deviation σ is known or can be easily estimated.

In its simplest form, we are trying to find a region for μ (and thus a confidence interval) that will contain the value of the true parameter of interest. The formula for finding the confidence interval for an unknown population mean from a sample is $\overline{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

The value of $Z_{\frac{\alpha}{2}}$ is computed from the normality assumption and the level of confidence, $1 - \alpha$, desired.

Let's consider a variation of the diet coke example in the previous section. For example, the amount of fluid dispensed into a can of diet coke is approximately a normal random variable with unknown mean fluid ounces and a standard deviation of 0.5 fluid ounces. We want to determine a 95% confidence interval for the true mean. A sample of 36 diet cokes was taken and a sample mean of $\bar{x} = 11.35$ was found.

Now, $1 - \alpha = 0.95$. Therefore, $\alpha = 0.5$ and since there are two regions then we need $\frac{\alpha}{2} = 0.25$ and $Z_{\frac{\alpha}{2}} = 1.96$. This is seen in the figure below.

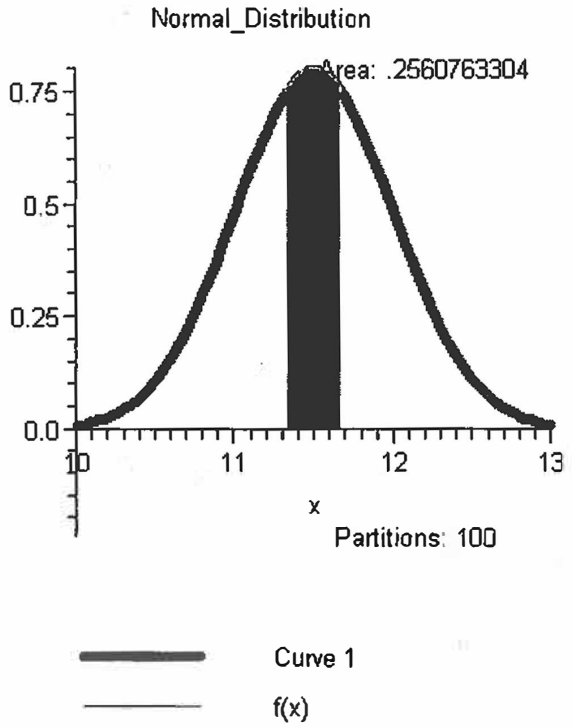


Figure 3.6 Confidence Interval $11.35 \pm 1.96 \cdot \frac{0.5}{\sqrt{36}}$

Our confidence interval for the parameter, μ, is $11.35 \pm 1.96 \cdot \frac{0.5}{\sqrt{36}}$.

Let's interpret this or any confidence interval. If we took 100 experiments of 36 random samples each, and calculated the 100 confidence intervals in the same manner, $\bar{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$.

Thus, 95 of the 100 confidence intervals would contain the true mean, μ. We do not know which of the 95 confidence intervals contain the true mean. Thus, to a modeler, each confidence interval built will either contain the true mean or it will not contain the true mean.
In EXCEL the command is CONFIDENCE(alpha,st_dev,size) and it only proved the value of $Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$, we must still combine to get the interval $\bar{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$.
In Excel we find our confidence interval for our example:

xbar	11.35
s	0.5
n	36
alpha	0.05

$Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

0.16333
Upper 11.51333
Lower 11.18667

Simple Hypothesis Testing

A more powerful technique for interring information about a parameter is a hypothesis test. A statistical hypothesis test is a claim about a single population characteristic or about values of several population characteristics. There is a null hypothesis (which is the claim initially favored or believe to be true) and is denoted by H_0. The other hypothesis, the alternate hypothesis, is denoted as H_a. We will always keep equality with the null hypothesis. The objective is to decide, based upon sample information, which of the two claims is correct. Typical hypothesis tests can be categorized by three cases:

CASE 1:	H_0:	$\mu = \mu_0$	versus	H_a:	$\mu \neq \mu_0$
CASE 2:	H_0:	$\mu \leq \mu_0$	versus	H_a:	$\mu > \mu_0$
CASE 3:	H_0:	$\mu \geq \mu_0$	versus	H_a:	$\mu < \mu_0$

There are two types of errors that can be made in hypothesis testing, Type 1 errors called α error and Type II errors called β errors. It is important to understand these. Consider the information provided in the table below.

State of Nature

		H_0 True	H_a True
Test Conclusion	Fail to Reject H_0	$1 - \alpha$	β
	Reject H_0	α	$1 - \beta$

Some important facts about both α and β:

(1) $\alpha = P(\text{reject } H_0 \mid H_0 \text{ is true}) = P(\text{Type I error})$

(2) $\beta = P(\text{fail to reject } H_0 \mid H_0 \text{ is false}) = P(\text{Type II error})$

(3) α is the level of significance of the test

(4) $1 - \beta$ is the power of the test

Thus, referring to the table we would like α to be small, since it is the probability that we reject H_0 when H_0 is true. We would also want $1 - \beta$ to be large since it represents the probability that we reject H_0 when H_0 is false. Part of the modeling process is to determine which of these errors is the most costly, and work to control that error as your primary error of interest.

The following template if provided for hypothesis testing:

STEP 1: **Identify the parameter of interest**
STEP 2: **Determine the null hypothesis, H_0**
STEP 3: **State the alternative hypothesis, H_a**
STEP 4: **Give the formula for the test statistic based upon the assumptions that are satisfied**
STEP 5: **State the rejection criteria based upon the value of α**
STEP 6: **Obtain your sample data and substitute into your test statistic**
STEP 7: **Determine the region in which your test statistics lies (rejection region or fail to reject region)**
STEP 8: **Make your statistical conclusion. Your choices are to either reject the null hypothesis or fail to reject the null hypothesis. Insure the conclusion is scenario oriented.**

You run a small aviation transport company for a major corporation. You are tired of hearing management complain that your crews rest too much during the day. Aviation rules require a crew to get around 9 hours of rest each day. You collect a sample of 37 crew members and determine that their sample average, $\bar{x}$, is 8.94 hours with a sample deviation of 0.2 hours.

The parameter of interest is the true population mean, μ.

H_0: $\mu \geq 9$
H_a: $\mu < 9$

The test statistic is $Z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$. This is a one-tailed test.

We select α to be 0.05.
We reject H_0 at $\alpha = 0.05$, if $Z < -1.645$.

From our sample of 36 aviators, we find $Z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{8.94 - 9}{.2/\sqrt{36}} = -0.06(6)/.2 = -1.8$. Z = -1.8.

Since, -1.8 < -1.645 then we reject null hypothesis that aviators rest 9 or more hours per day and conclude the alternate hypothesis is true, that your aviators rest less than 9 hours per day. Rejecting the null hypothesis is the better strategy because it is now concluded that we reject the null hypothesis that the aviator crews rest 9 or more hours a day.

P-Value. The P-Value is the appropriate probability related to the test statistic. It is written so that the result is the smallest alpha level in which we may reject the null hypothesis. It is normal probability. From above our test statistic is -1.8, We are doing a lower tail test. P-Value is P(Z<-1.8) = 0.0359. Thus, we reject the null hypothesis for all values of alpha > 0.0359. Thus, we reject of alpha is 0.05 but fail to reject if alpha is 0.01.

In statistical significance testing, the **p-value** is the probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. In this context, value *a* is considered more "extreme" than *b* if *a* is less likely to occur under the null. One often "rejects the null hypothesis" when the p-value is less than the significance level α (Greek alpha), which is often 0.05 or 0.01. When the null hypothesis is rejected, the result is said to be statistically significant.

The P value is a probability, with a value ranging from zero to one. It is the answer to this question: If the populations really have the same mean overall, what is the probability that random sampling would lead to a difference between sample means as large (or larger) than you observed?

We usually use either a normal distribution directly or evoke the central limit theorem

(for n>30) for testing means. Let's say we think our mean of our distribution is ½. We want to test if our sample comes from this distribution.

H_0: μ=1.2

H_a: $\mu \neq$ ½

The test statistic is key. From our data with sample size n=49, we find that the mean is 0.41 and the standard deviation is 0.2.

The test statistic for a one sample test of a proportion is

$$z = \frac{p - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

So we substitute $p=1/2$, $p_o=.41$, $(1-p_o)=.59$, $n=49$

$$z=\frac{0.5-0.41}{\sqrt{.41(1-.41)/49}}$$

We find $z=1.28$

We next need to find the probability that corresponds to the statement $P(Z\geq 1.28)$.

If z were negative we would want $P(Z\leq -1.28)$.

We need to find that probability. We find $P(Z\leq 1.28) = 0.8997$. So $P(Z>1.28) = 1-p(Z<1.28)= 1-0.8997=0.1003$

We compare that value, p, to our level of significance.

If $p < a$ then we have significance (a is usually either *0.05* or 0.01).

Statistical calculations can answer this question: If the populations really have the same mean, what is the probability of observing such a large difference (or larger) between sample means in an experiment of this size? The answer to this question is called the *P value.*

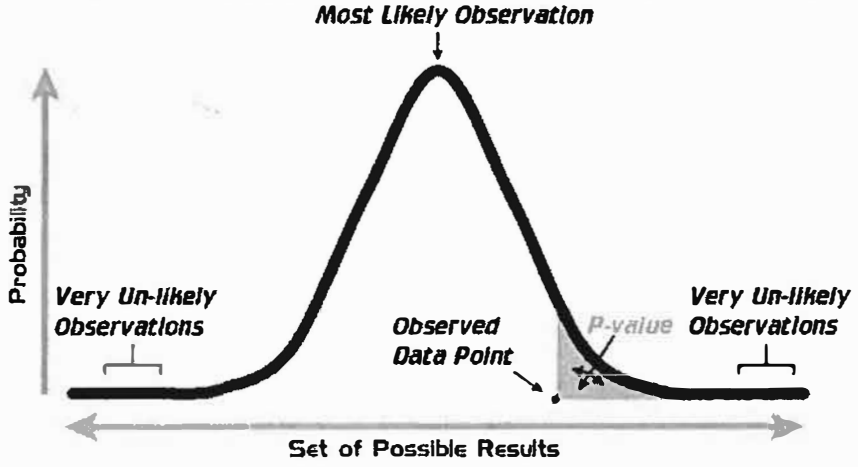
The P value is a probability, with a value ranging from zero to one. If the P value is small, you'll conclude that the difference between sample means is unlikely to be a coincidence. Instead, you'll conclude that the populations have different means.

IMPORTANT:

Pr (Observation | Hypothesis) ≠ Pr (Hypothesis | Observation)

The probability of observing a result given that some hypothesis is true is not equivalent to the probability that a hypothesis is true given that some result has been observed.

Using the p-value as a "score" is commiting an egregious logical error:
The Transposed Conditional Fallacy



A p-value (shaded green area) is the probability of an observed (or more extreme) result arising by chance

EXCEL TEMPLATEs

Hypothesis Test Template

2 Tail Test		Test Stat	Value
Mean	8.94	$Z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$	-1.8
Population Mean, hyp	9		
Standard Deviation, S	0.2		
N, sample size	36		
Alpha Level	0.05	-1.64485	Results
Enter tail information	2		Reject
Upper tail as 0			
Lower tail as 1			
Both tails as 2			

User inputs are in yellow

Given our hypothesis test above, the probability of a Type I error, α, is the area under the normal bell-shaped curve centered at μ_0 corresponding to the rejection region. This value is 0.05.

3.4 Exercises

Discuss how to set up each of the following as a hypothesis test.

1. Does drinking coffee increase the risk of getting cancer?
2. Does taking aspirin every day reduce the chance of a heart attack?
3. Which of two gauges is more accurate?
4. Why is a person "innocent until proven guilty"?
5. Is the drinking water safe to drink?
6. Set up a fake trial for a suspected felon. Build a matrix for their innocence or guilt with an appropriate null hypothesis. Which error, Type I or Type II, is the worst error?

7. Numerous complaints have been made that a certain hot coffee machine is not dispensing enough hot coffee into the cup. The vendor claims that on average the machine dispenses at least 8 ounces of coffee per cup. You take a random sample of 36 hot drinks and calculate the mean to be 7.65 ounces with a standard deviation of 1.05 ounces. Find a 95% confidence interval for the true mean.

8. Numerous complaints have been made that a certain hot coffee machine is not dispensing enough hot coffee into the cup. The vendor claims that on average the machine dispenses at least

8 ounces of coffee per cup. You take a random sample of 36 hot drinks and calculate the mean to be 7.65 ounces with a standard deviation of 1.05 ounces. Set up and conduct a hypothesis test to determine if the vendors claim is correct. Use an $\alpha = .05$ level of significance. Determine the Type II error if the true mean were 7.65 ounces.

Further hypothesis needed. Simple means, simple proportions, two means, and two proportions.

Hypothesis Testing

Questions: to test an hypothesis—is the sample Normally distributed? or is the sample large (n>30) since the test concern means?

Notation and Definitions

H_0 the null hypothesis & is what we assume to be true.
Ha the alternative hypothesis and generally what is the worst case or what we want to prove.
α= P(Type I error) known as level of significance (usually 0.05 or 0.01).
β = P(Type II error)
Type I error rejects the null hypothesis when it is true.
Type II error fail to reject the null hypothesis when it is false.
Power of test=1-β. We want this to be large. This is the probability that someone guilty is found guilty.
Conclusions: reject H_0 or fail to reject H_0.
One tail test from H_a.
Two tail test from H_a.
Test statistic, T_s, comes from our data and is found by $z = \frac{\bar{x}-\mu}{\left(\frac{s}{\sqrt{n}}\right)}$
Rejected region: that area under the normal curve where we reject the null hypothesis.
P-values is the smallest level of significance at which H_0 would be rejected when a specified test procedure is used on a given data set. We compare P-value to our given α. If P-value $\leq \alpha \rightarrow$ we reject H_0 at level α. If P-value > a $\rightarrow$ we fail to reject H_0 at level α. It is usually thought of as the probability associated with the test statistic , P(X-bar>T_S)

HYPOTHESIS TESTING

	H_0 is true	H_0 is false
Reject H_0	Type I error P(Type I error) = α	Correct decision
Fail to Reject H_0	Correct decision	Type II error P(Type II error) = β

Example: H_0: The defendant is innocent
H_A: The defendant is guilty

What is a Type I error: Someone who is innocent is convicted—we want that to me small.
What is a Type II error: Someone who is guilty is cleared, we want that small also.

Example: H_0: The drug is not safe and effective
H_A: The drug is safe and effective

What is a Type I error: Unsafe/ineffective drug is approved

What is a Type II error: Safe/Effective drug is rejected
The reason we do it this way is we want to prove that the drug is safe and effective.

Mathematically when we examine hypothesis tests we always put the = with H_0!!!!!