Introduction to SPSS

Stephanie Dickinson

Biostatistics Consulting Center
School of Public Health - Bloomington

Section 1

Introductions

Introductions

- **Biostatistics Consulting Center**, School of Public Health --- biostats.indiana.edu

   Professional statisticians for health-related research. Free consulting Tu/Th 10-12 @ SSRC.

- **Social Science Research Commons** --- ssrc.indiana.edu

- **Research Analytics**, UITS RT --- https://rt.iu.edu/

- **Indiana Statistical Consulting Center**, --- iscc.indiana.edu

- **Center for Survey Research** --- csr.indiana.edu

Software

- **SPSS** – easy "point & click", good for most "off the shelf" analyses

- **SAS** – syntax code, industry standard, public health,…

- **R** – syntax code, free & flexible

- **STATA** – syntax w/ "point & click", political science, sociology,…

- **JMP** – "point & click", good mix of stats and graphs – good for exploring data

- **MATLAB** – powerful numerical computing, matrix manipulations

Software Acquisition

- **Purchase** SPSS ($100) or JMP ($50) from Research Analytics --- https://kb.iu.edu/d/bfhv
  - Download from **Iuware.iu.edu**

- **IUanyWare.iu.edu**
  - Free software, streaming online
  - Download the Citrix Workspace (or use the "light" version in the browser)

- **cloudstorage.iu.edu**
  Link your preferred storage and have access to your files from campus computers and apps.
  - Box.iu.edu
  - Dropbox
  - File server

Practice Project: Local Food in Indiana

Comparing consumers (n=302) who purchase food 1 of 3 places in their motivations towards local food.

- Farmer's Markets,
- CSA's (Community Supported Agriculture), or
- Neither.

Shoppers were surveyed at the Farmer's Market, CSA members, and public.

Thank you to James Farmer (SPH) for sharing his data.

Getting Started in SPSS

Open SPSS

- Double-click an SPSS data file (.sav)
- Or, Open SPSS first
  - and open an SPSS data file
  - Or Import an Excel or .csv file.

✓ Open 'CSA Farmer Market.sav'

Data View

- Columns are 'Variables'

- Rows are subjects, or 'Observations'

  o (Usually one row per subject…)

BIOSTATISTICS CONSULTING CENTER

Variable View

- Type (Numeric, String)

- Label

- Values

- Missing

- Measure (Nominal, Scale, Ordinal)

Example Data: Local Food in Indiana

1.   Please indicate your level of agreement for the following statements on a scale from **S**trongly **D**isagree (SD), **D**isagree (D), Neutral (N), **A**gree (A), to **S**trongly **A**gree (SA).

		1 SD	2 D	3 N	4 A	5 SA
1	Purchasing organically grown food is very important to me. ← **Q1MOTORGANIC**					
2	I give preference to foods that are grown with few chemical applications. ← **Q1MOTFEWCHEM**					
3	I give preference to foods that were picked just a few days before my purchase. ← **Q1MOTFRESH**					
	~~Over half of the foods/groceries I purchase are fresh produce.~~					
4	The nutritional value of a food is an important part of my purchasing decisions.					
5	I give preference to animal products that have been derived in a humane manner.					
6	I give preference to animal products that are free from growth hormones.					
7	The expense of fresh local produce deters me from purchasing it as often as I would like.					
8	I generally purchase whole foods, rather than processed foods. ← **Q1MOTWHOLE**					
9	I give preference to purchasing foods that come from within 100 miles of my location.					
10	I give preference to eating foods that are in season, for example, tomatoes in July-October.					
11	I give preference to food purchase decisions that support the local economy.					
12	I give preference to food purchase decisions that support local farmers.					
13	I believe consuming food produced locally is better for the environment.					

Data Types (Measures)

Data Types		Example
Continuous/ Interval/ Scale		Test scores Height, weight, age Response Time <Percent, proportions >, <Counts> <Likert-type items>
Ordinal		Education: Bachelor, Masters, PhD. Likert-type items
Categorical: Nominal (≥2) Binary (2 levels)		Treatment Group (A,B,C) Sex: Male/female, Yes/no, right/wrong, 0/1.
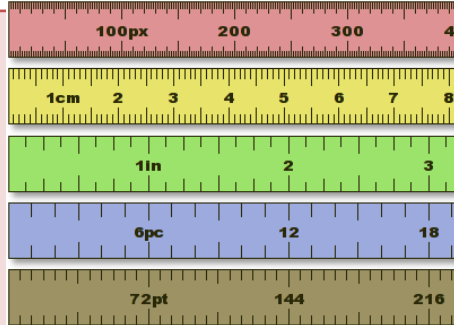
Likert-type scales

Likert Scales

Please fill in the number that represents how you feel about the computer software you have been using

I am satisfied with it

① ② ③ ④ ⑤
Strongly Agree   Agree   Neither   Disagree   Strongly Disagree

It is simple to use

① ② ③ ④ ⑤
Strongly Agree   Agree   Neither   Disagree   Strongly Disagree

It is fun to use

① ② ③ ④ ⑤
Strongly Agree   Agree   Neither   Disagree   Strongly Disagree

It does everything I would expect it to do

① ② ③ ④ ⑤
Strongly Agree   Agree   Neither   Disagree   Strongly Disagree

I don't notice any inconsistencies as I use it

① ② ③ ④ ⑤
Strongly Agree   Agree   Neither   Disagree   Strongly Disagree

DEBATE over whether it's "okay" to treat these as Continuous scales… (Argument is that the items are not equal distance apart)

Yes, it's truly *Ordinal* but usually needs to be **treated** as *Categorical* or *Continuous* for standard analyses.

Likert items are ambidextrous…could go either way…

(Summary scales, average across 5 items, would be Continuous)

Data Cleaning and Preparation

- Missing data?
  - How is it coded?
- Invalid data?
- Outliers?

- **Prepare data for analysis**
  - Calculate new variables
  - Recodes
  - Transformations
- **Codebook**
  - Formats/Values
- **Syntax**

The MAJORITY of your time will be cleaning, coding, and planning! Don't short-change it!

Research Questions

What are the differences in Food Preferences & Motivations for Local Food between people who purchase produce through a CSA, at the Farmer's Market, or neither?

- And what are the differences by Gender?

- **Outcome Measures (DV):** Food Preferences
  - Individual items? (Purchasing Organically Grown Food (Q1MOTORGANIC), etc)
  - Composite summary scores
- **Predictor Variables (IV):**
    - Venue Type (CSA, Farmer's Market, neither)
    - Sex (Male, Female)

Create 3 summary scores

- "Organic, Whole, Humane" = Mean of items 1, 2, 4, 5, 6, 8
- "Fresh, Local, In Season" = Mean of items 3, 9-13

Purchasing organically grown food is very important to me.	A	← Q1MOTORGANIC
I give preference to foods that are grown with few chemical applications.	A	← Q1MOTFEWCHEM
I give preference to foods that were picked just a few days before my purchase.	B	
~~Over half of the foods/groceries I purchase are fresh produce.~~		
The nutritional value of a food is an important part of my purchasing decisions.	A	← Q1MOTNUTR
I give preference to animal products that have been derived in a humane manner.	A	← Q1MOTANIMHUMA
I give preference to animal products that are free from growth hormones.	A	← Q1MOTANIHORMONE
The expense of fresh local produce deters me from purchasing it as often as I would like.	C	
I generally purchase whole foods, rather than processed foods.	A	← Q1MOTWHOLE
I give preference to purchasing foods that come from within 100 miles of my location.	B	
I give preference to eating foods that are in season, for example, tomatoes in July-October.	B	
I give preference to food purchase decisions that support the local economy.	B	
I give preference to food purchase decisions that support local farmers.	B	
I believe consuming food produced locally is better for the environment.	B	

Organic, Whole, Humane

Fresh, Local, In season

Expensive

Do any items need to be "reverse coded" ?
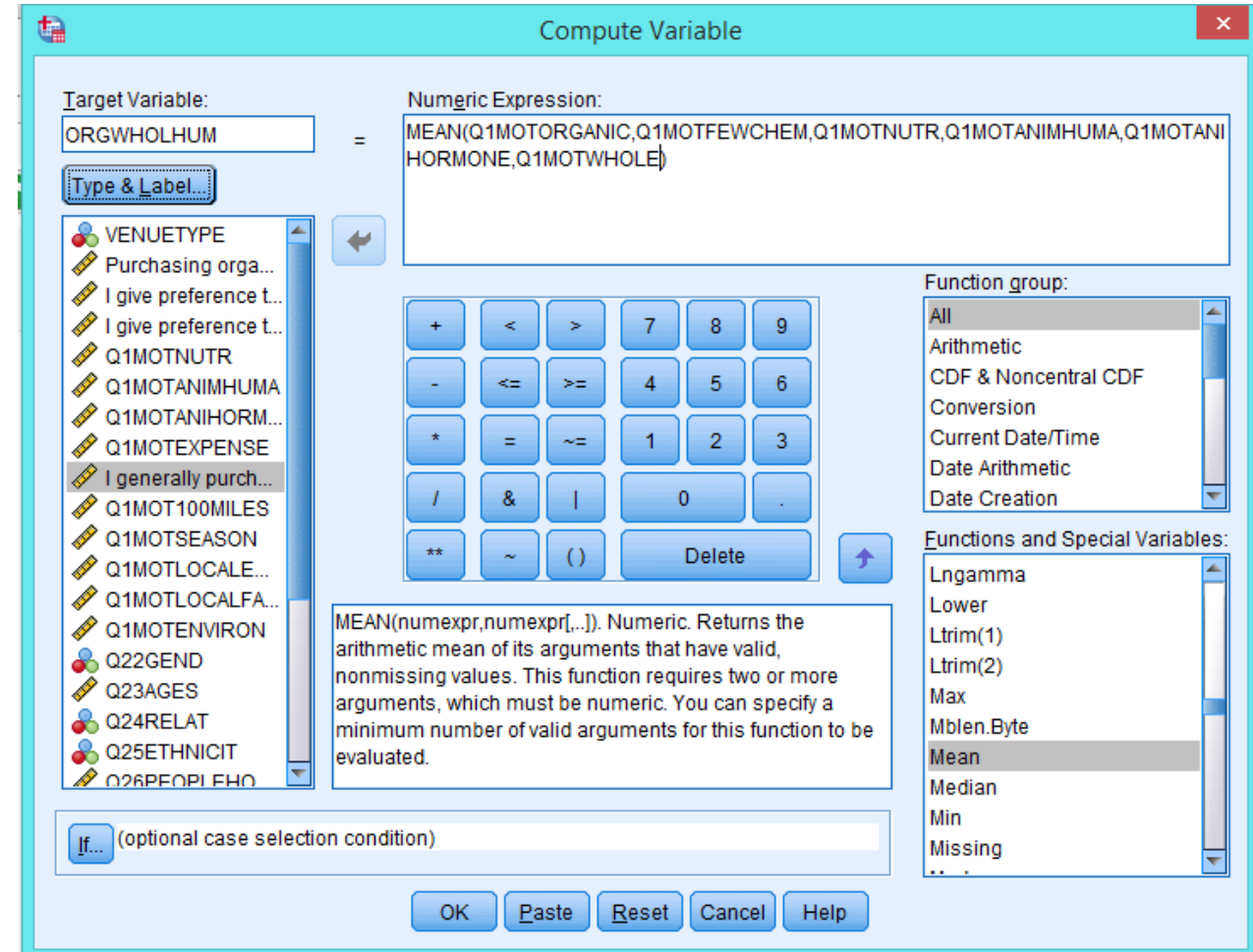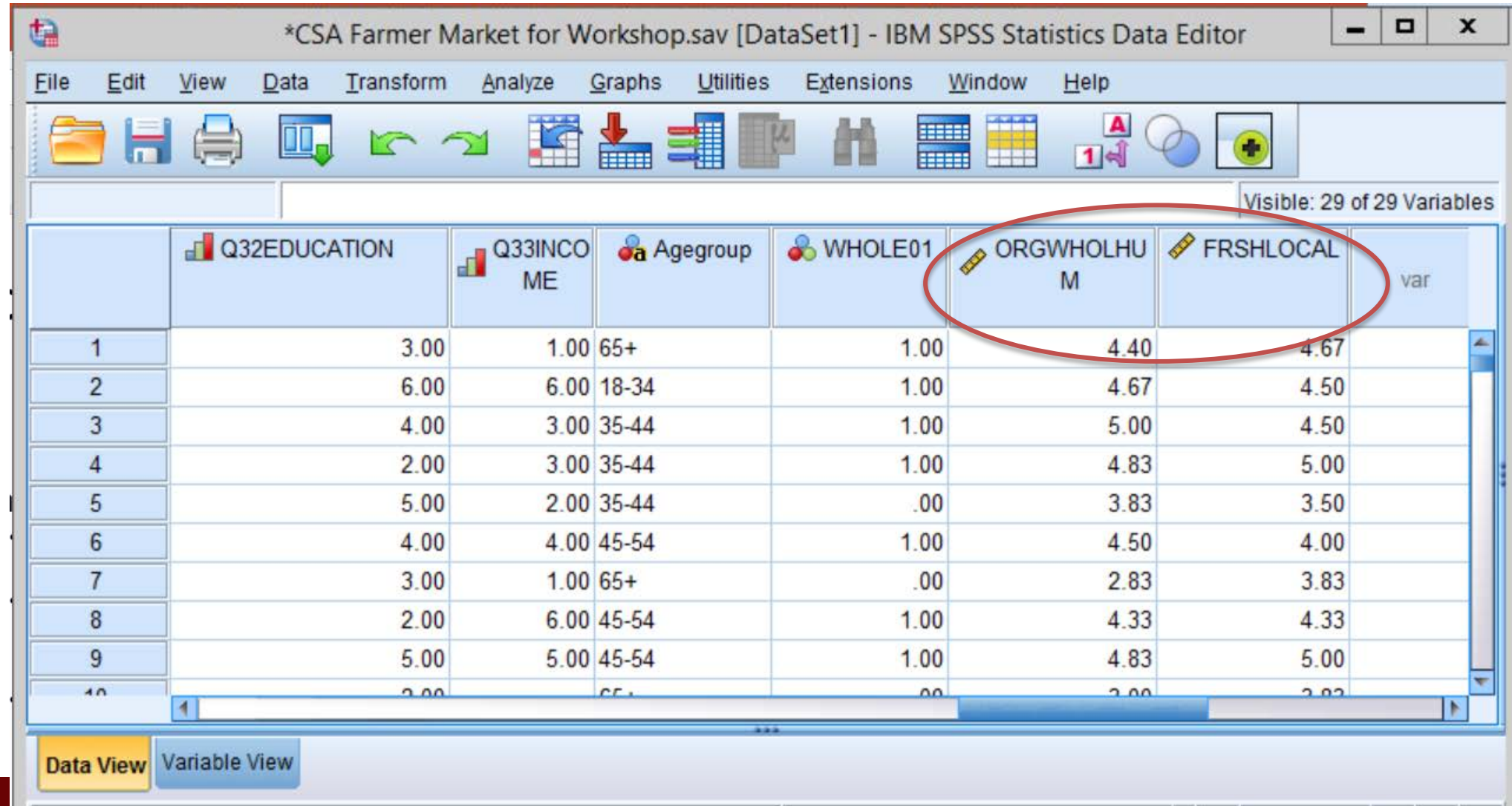
Creating New Variables

- Calculate new variables which are **Summary scores**

❑ Transform > Compute Variable
  - "Organic, Whole, Humane"
    ORGWHOLHUM = Mean of items 1,2,4,5,6,8
  - "Fresh, Local, In Season"
    FRSHLOCAL = Mean of items 3,9-13

  - ORGWHOLHUM & FRSHLOCAL are then added at the end of the dataset

Created variables appear at the end of the data set

Data Recodes

Maybe we need to recode some of the 5-point scales into binary (high/low) splits.

Preference for Whole Foods? (Yes/No)

- **Recodes**
- ❑ Transform > Recode into Different Variables. Select 'Q1MOTWHOLE' as Numeric Variable, and Name Output variable 'WHOLE01'. Select 'Old and New Values'.
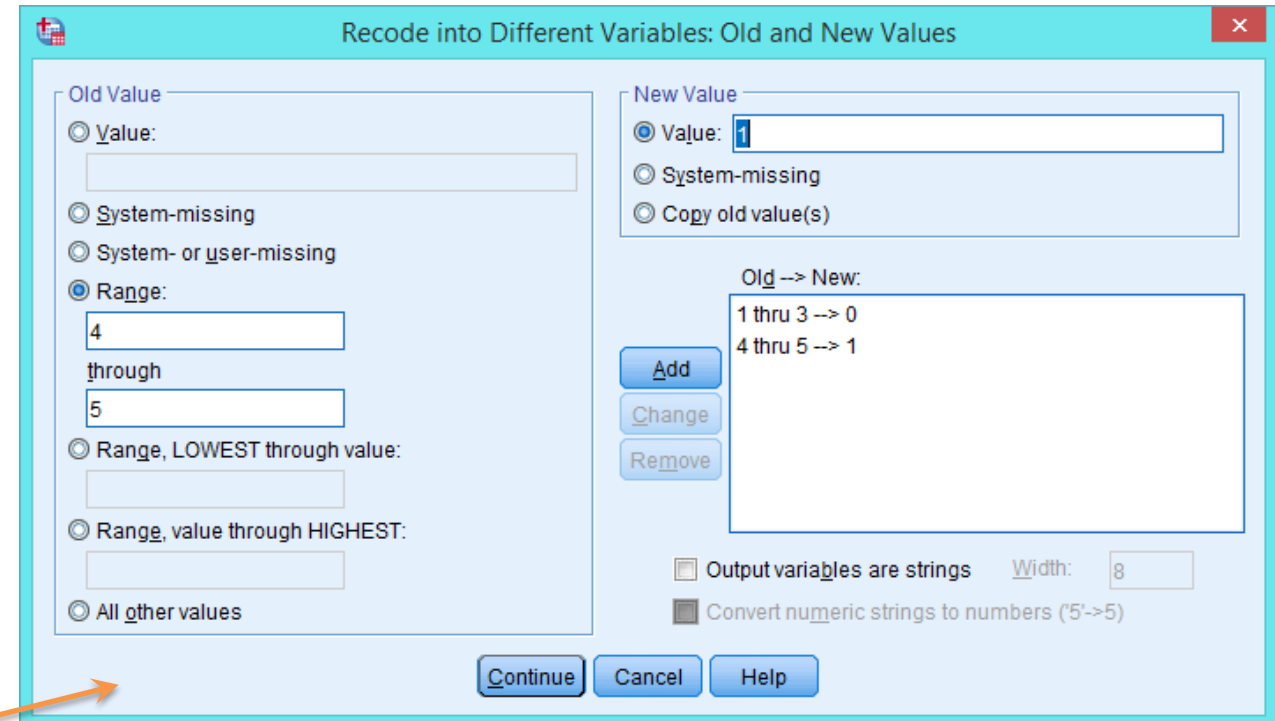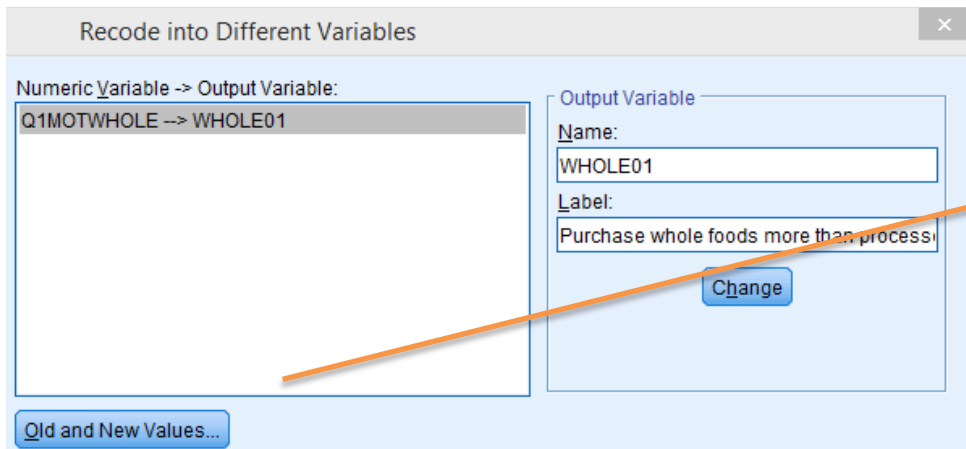
WHOLE01=1 for Yes (Q1MOTWHOLE 4-5),
WHOLE01=0 for No (Q1MOTWHOLE 1-3)

Syntax

Save syntax from the
Output window to
Paste into a Syntax
window to re-run later.
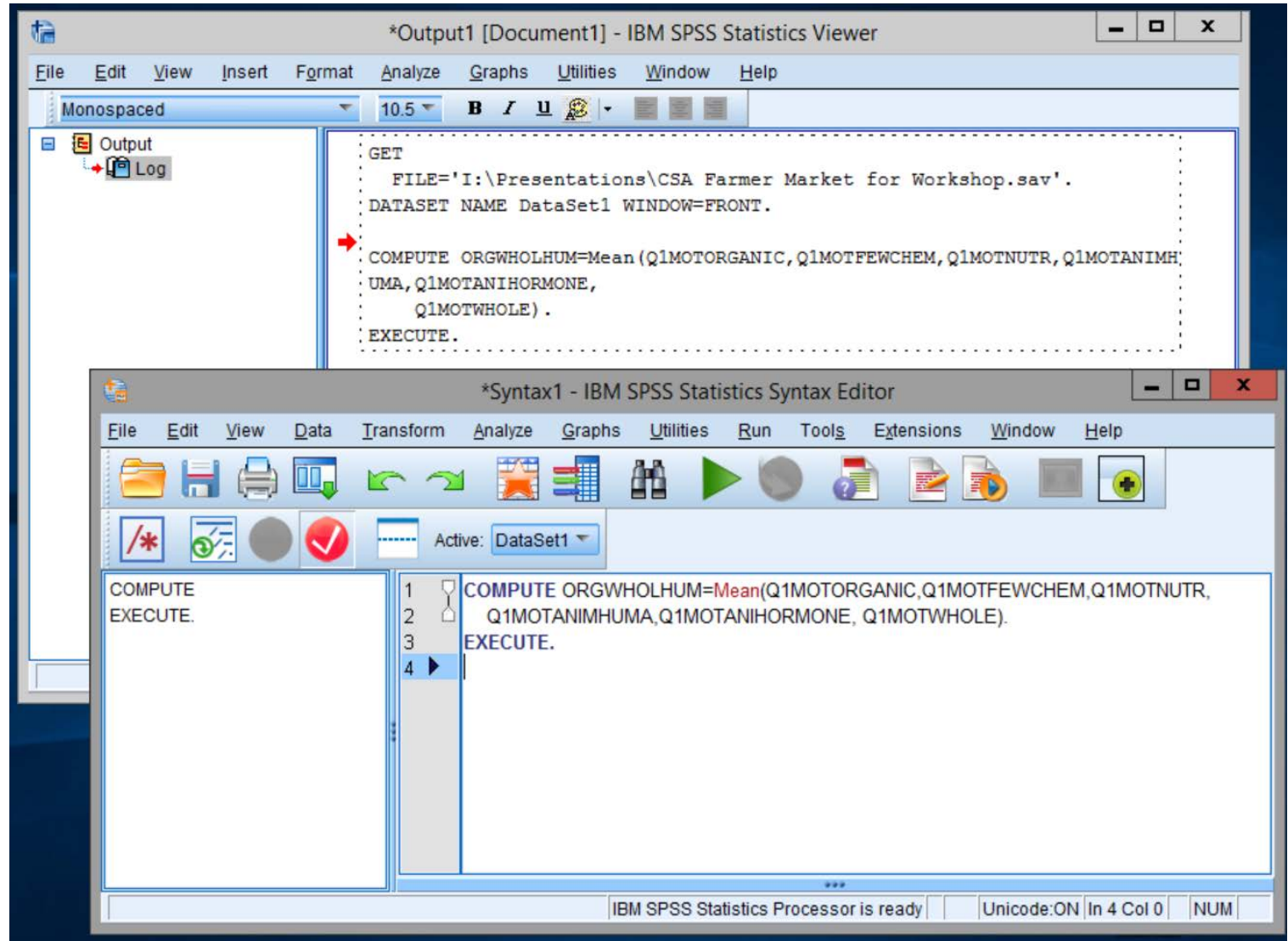
Data Analysis SPSS

Overview of Basic Analyses

		Descriptive Stats & Plots	Inferential Stats (Significance tests)	
Univariate (1 at a time)	1 Scale Variable	Mean, Med, Min, Max, Std. Dev., Histogram		1
	1 Nominal Variable	Frequency, Percent, Bar chart		2
Bivariate Association (1 on 1)	2 Scale Variables	Scatterplot	Correlation	3
	2 Nominal Variables	Cross-tab	Chi-square Test	4
	1 Scale Variable w/ 1 Nominal Variable (Comparing Groups)	Means by Subgroup, Box-plots	T-test (Compare 2 groups), ANOVA (Compare 3+ groups)	5 / 6
Multiple variables	Scale variable as response with multiple predictors		Linear Regression, GLM	7
	Nominal variable as response with multiple predictors		Logistic Regression	8

Data Types (Measures)

- **Q1MOTORGANIC** (etc) → *Scale*
  - 1 to 5 discreet (Strongly Disagree to Strongly Agree)
  - (Note that this is technically *Ordinal* but we will treat as linear *Scale*.)
- **ORGWHOLHUM & FRSHLOCAL** → *Scale*
  - 1 to 5 scale (Mean of discreet items)
- **VENUETYPE** → *Nominal*
  - 1=CSA
  - 2=Farmer's Market
  - 3=Neither
- **SEX** → *Nominal*
  - 1=Female
  - 2=Male

① Descriptive Stats

Summarizing Scale variables

Describe how much people like to purchase organically grown food.

- **Histograms**
- ❑ Graphs > Legacy > Histogram > Select 'Q1MOTORGANIC' as Variable

- **Descriptive Stats** (Mean, SD, Med, Min, Max)
- ❑ Analyze > Descriptive Stats > Descriptives > Select 'Q1MOTORGANIC' as Variable

Mean = 3.92
Std. Dev. = 1.00
N = 299

Purchasing organically grown food is very important to me
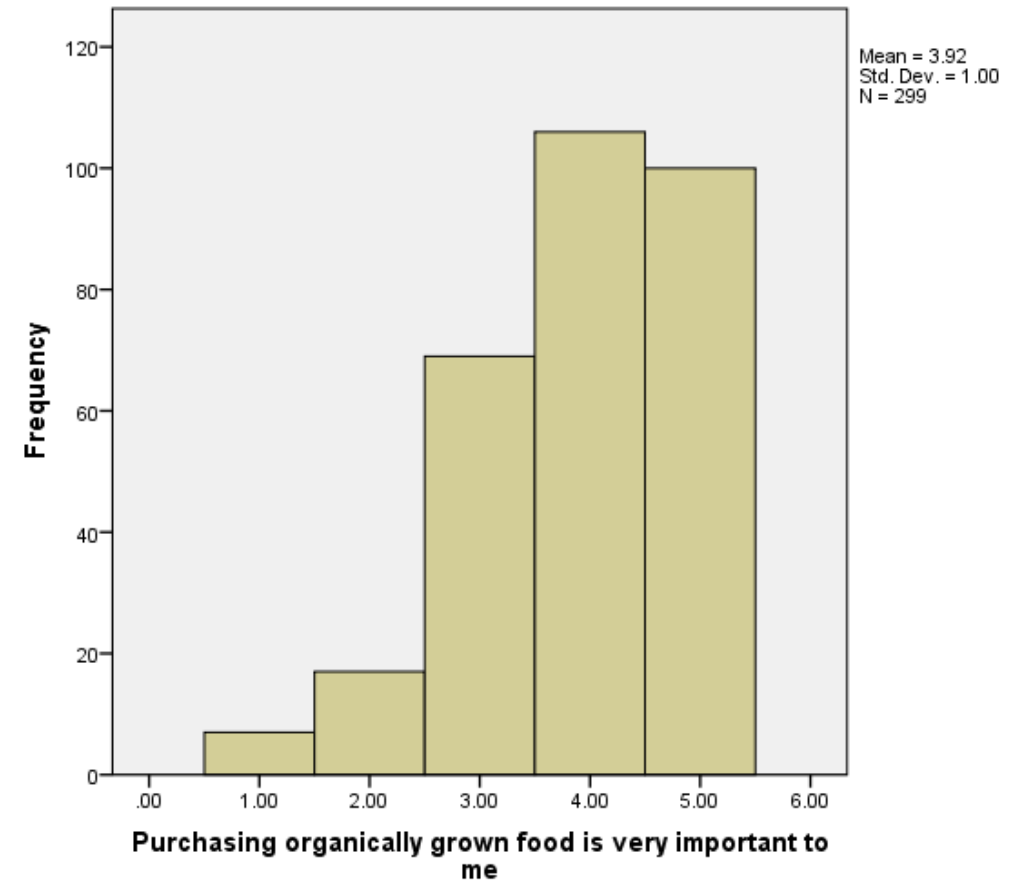
	N	Minimum	Maximum	Mean	Std. Deviation
Purchasing organically grown food is very important to me	299	1.00	5.00	3.9197	1.00012
I give preference to foods that are grown with few chemical applications.	299	1.00	5.00	4.1538	.92490

Frequencies

Summarizing Nominal variables

How many people responded to each survey?

- **Bar Graph**
- ❑ Graphs > Legacy > Bar > Simple > Select 'VENUE' (or 'SEX') as Category Axis

- **Frequency table** (n, %)
- ❑ Analyze > Descriptive Stats > Frequencies > Select 'VENUE' as Variable
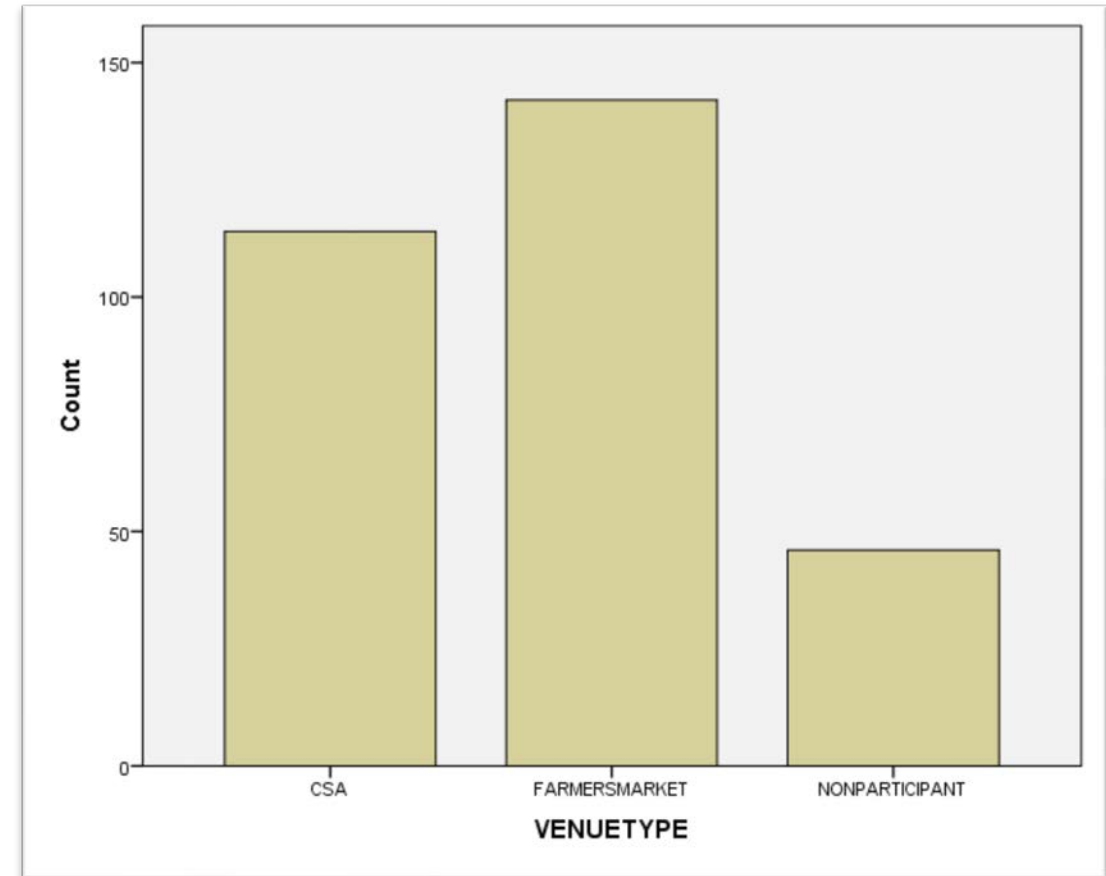
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	CSA	114	37.7	37.7	37.7
	FARMERSMARKET	142	47.0	47.0	84.8
	NONPARTICIPANT	46	15.2	15.2	100.0
	Total	302	100.0	100.0	

Correlation

Relationship between 2 Scale variables

Is motivations for organic food correlated with motivations to have fewer chemicals?

- **Scatterplot**
- ❑ Graphs > Legacy > Scatter > Simple > Select 'Q1MOTORGANIC' as Y Axis and 'Q1MOTFEWCHEM' as X Axis.
  (Note: Use syntax to "jitter" the points)

- **Correlation**
- ❑ Analyze > Correlate > Bivariate > Select 'Q1MOTORGANIC' and 'Q1MOTFEWCHEM' as Variables
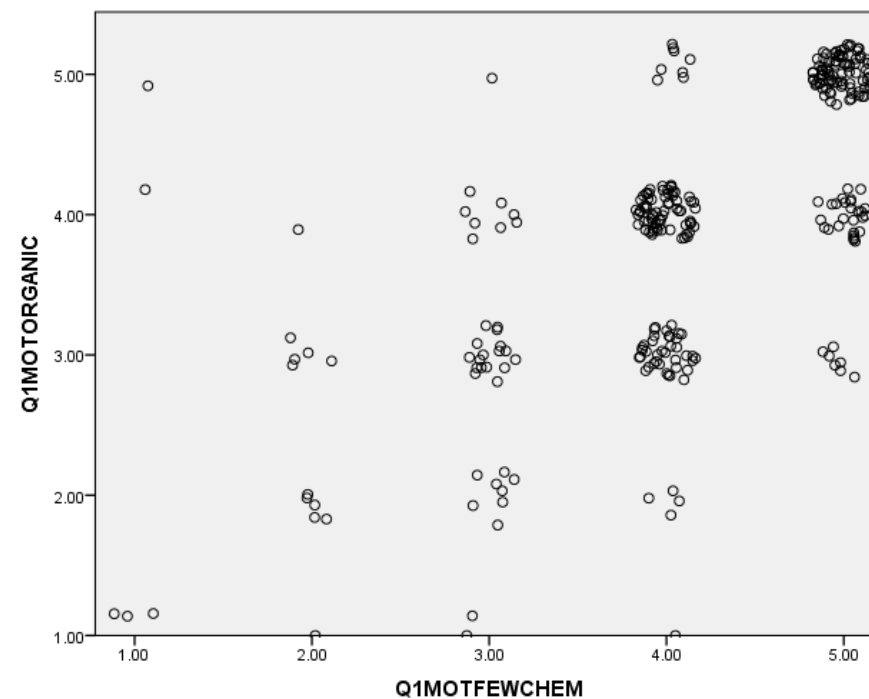
Correlations

		Q1MOTORGANIC	Q1MOTFEWCHEM
Q1MOTORGANIC	Pearson Correlation	1	.685**
	Sig. (2-tailed)		.000
	N	299	296
Q1MOTFEWCHEM	Pearson Correlation	.685**	1
	Sig. (2-tailed)	.000	
	N	296	299

**. Correlation is significant at the 0.01 level (2-tailed).

Chi-square

(4)

Relationship between 2 Nominal variables

Is gender associated with Venue type?

- **Stacked Bar Chart**
- ❑ Graphs > Legacy > Bar > Stacked > Select 'VENUE' as Category Axis and 'Q22Gend' to Define Clusters

- **Crosstab w/ Chi-square test**
- ❑ Analyze > Descriptive > Crosstab > Select 'VENUETYPE' as Rows and 'Q22Gend' Columns Statistics button > Select Chi-square. Cells button > Select Row %
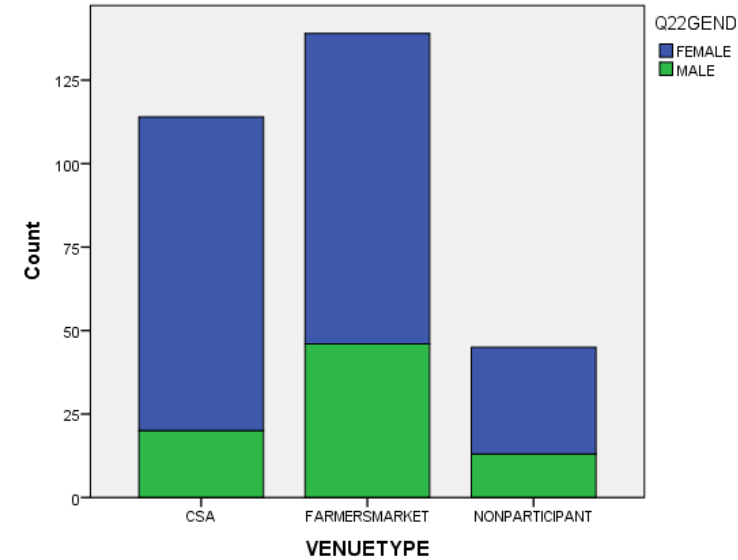
			Q22GEND FEMALE	MALE	Total
VENUETYPE	CSA	Count	94	20	114
		% within VENUETYPE	82.5%	17.5%	100.0%
	FARMERSMARKET	Count	93	46	139
		% within VENUETYPE	66.9%	33.1%	100.0%
	NONPARTICIPANT	Count	32	13	45
		% within VENUETYPE	71.1%	28.9%	100.0%
Total		Count	219	79	298
		% within VENUETYPE	73.5%	26.5%	100.0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	7.927[a]	2	.019

T-Test : Comparing 2 Groups

1 Scale Var. w/ 1 Nominal Var. (2 levels)

Is there a difference in preferences for organic food between males and females?

- **Box-plot**
- ☐ Graphs > Legacy > Boxplot > Simple > Select 'Q1MOTWHOLE' as Variable and 'Q22Gend' as Category Axis

- **T-test**
- ☐ Analyze > Compare Means > Independent Samples T-test > Select 'Q1MOTWHOLE' as Test Variable and 'Q22Gend' as Grouping Variable. Define Groups – levels 1 & 2
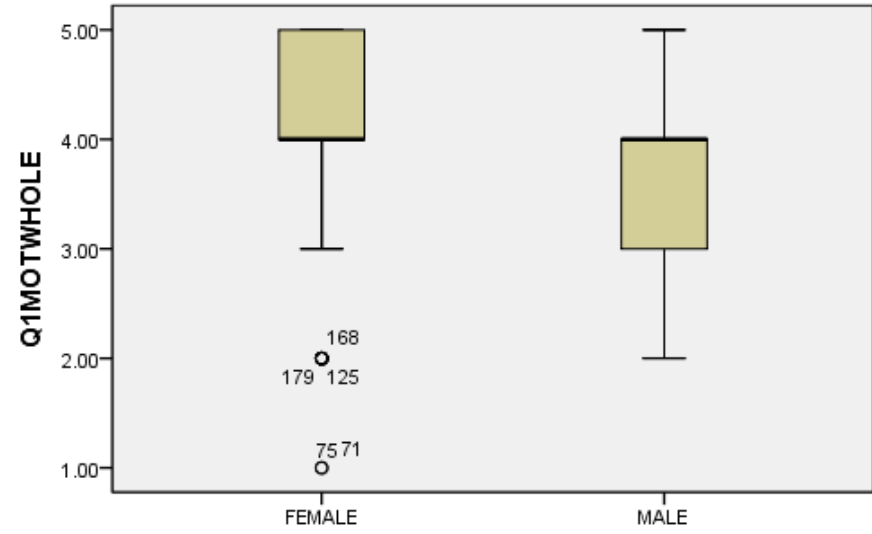
Group Statistics

	Q22GEND	N	Mean	Std. Deviation	Std. Error Mean
Q1MOTWHOLE	FEMALE	218	4.0321	.87140	.05902
	MALE	79	3.6962	.80630	.09072

Independent Samples Test

		Levene's Test for Equality of Variances				
		F	Sig.	t	df	Sig. (2-tailed)
Q1MOTWHOLE	Equal variances assumed	.035	.852	2.993	295	.003
	Equal variances not assumed			3.104	148.444	.002

BIOSTATISTICS CONSULTING CENTER

ANOVA : Comparing 3+ Groups

**1 Scale Var. w/ 1 Nominal Var. (3+ levels)**

Is there a difference in preferences for organic food between venue types?

- **Box-plot**
- ❑ Graphs > Legacy > Boxplot > Simple > Select 'Q1MOTWHOLE' as Variable and 'VENUETYPE' as Category Axis

- **ANOVA**
- ❑ Analyze > Compare Means > One-way ANOVA , or…
- ❑ Analyze > General Linear Model > Univariate > Select 'Q1MOTWHOLE' as Dependent and 'VENUETYPE' as Fixed Factor. Also button for 'Post Hoc…' > 'VENUETYPE'> 'Tukey'.
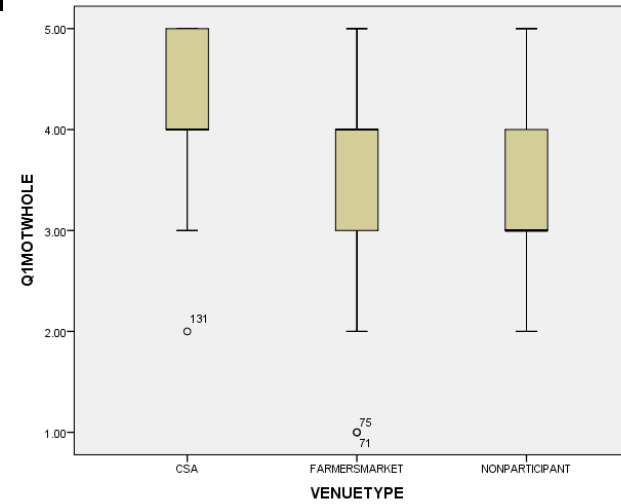
Tests of Between-Subjects Effects

Dependent Variable: Q1MOTWHOLE

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	28.012ª	2	14.006	21.172	.000
Intercept	3535.455	1	3535.455	5344.327	.000
VENUETYPE	28.012	2	14.006	21.172	.000
Error	197.137	298	.662		
Total	4914.000	301			
Corrected Total	225.150	300			

Linear Regression/ GLM

1 Scale Var. w/ Multiple predictors

How do gender and venue type together predict organic motivations?

- **GLM, Factorial ANOVA, ANCOVA**
- ❑ Analyze > General Linear Model > Univariate. Select 'Q1MOTWHOLE' as DV.
  - ▪ Nominal vars are "Fixed Factors" (VenueType, Q22 Gend)
  - ▪ Scale vars are "Covariates"

- **Linear Regression**
- ❑ Analyze > Regression > Linear. Select 'Q1MOTWHOLE' as DV.
- ❑ All IV's must be Scale or Binary "dummies" (0/1)

Tests of Between-Subjects Effects

Dependent Variable: Q1MOTWHOLE

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	32.290[a]	5	6.458	9.905	.000
Intercept	2555.977	1	2555.977	3920.113	.000
VENUETYPE	13.872	2	6.936	10.638	.000
Q22GEND	2.247	1	2.247	3.446	.064
VENUETYPE * Q22GEND	.856	2	.428	.656	.520
Error	189.737	291	.652		
Total	4839.000	297			
Corrected Total	222.027	296			

Coefficients[a]

Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	4.202	.180		23.352	.000
	Q23AGES	-.005	.003	-.088	-1.505	.133

a. Dependent Variable: Q1MOTWHOLE

Logistic Regression

1 Nominal Var. w/ Multiple predictors

How do gender, venue type and age predict preferences for whole food (high/low)?

- **Logistic Regression**
- ❑ Analyze > Regression > **Binary Logistic**. Select 'WHOLE01' as DV. Put IV's as Covariates, and use 'Categorical' button to identify Nominal vars.

- ➤ 'Purchase Whole foods vs Processed?' WHOLE01=1 for Yes (Q1MOTWHOLE 4-5), WHOLE01=0 for No (Q1MOTWHOLE 1-3)
- ➤ Someone who uses a CSA has 7.1 times the odds of saying they purchase Whole Foods compared to Nonparticipant.

Categorical Variables Codings

		Frequency	Parameter coding (1)	(2)
VENUETYPE	CSA	114	1.000	.000
	FARMERSMARKET	135	.000	1.000
	NONPARTICIPANT	44	.000	.000
Q22GEND	FEMALE	214	1.000	
	MALE	79	.000	

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	VENUETYPE			21.531	2	.000	
	VENUETYPE(1)	1.965	.424	21.467	1	.000	7.134
	VENUETYPE(2)	.992	.363	7.476	1	.006	2.698
	Q23AGES	.002	.009	.049	1	.825	1.002
	Q22GEND(1)	.361	.303	1.421	1	.233	1.435
	Constant	-.466	.642	.526	1	.468	.628

a. Variable(s) entered on step 1: VENUETYPE, Q23AGES, Q22GEND.

THANK YOU!

Stephanie L. Dickinson

Senior Biostatistician

Biostatistics Consulting Center

Hire a pro! http://biostats.Indiana.edu

Free DIY consultations Tu/Th 10-12 in the SSRC (ssrc.Indiana.edu)

Department of Epidemiology & Biostatistics

School of Public Health-Bloomington

Indiana University

License Terms

- Please cite as: *Dickinson, SL (2019). Introduction to SPSS [PowerPoint slides]. Retrieved from* [IUScholarWorks handle link].

- Except where otherwise noted, contents of this presentation are copyright 2013 by the Trustees of Indiana University.

- This document is released under the Creative Commons Attribution 3.0 Unported license (http://creativecommons.org/licenses/by/3.0/). This license includes the following terms: You are free to share – to copy, distribute and transmit the work and to remix – to adapt the work under the following conditions: attribution – you must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work). For any reuse or distribution, you must make clear to others the license terms of this work.