



Stephanie Dickinson
Senior Statistical Consultant
Department of Epidemiology & Biostatistics/
Social Science Research Commons
sd3@indiana.edu

YOUR STATISTICAL TOOL BELT

OVERVIEW

“Let’s build something together”

Get the right tools for the right project.

Workshop outline:

Part I:

- ❑ Your materials (data)
- ❑ Your project (research questions)
- ❑ Your tools (analysis methods)
- ❑ Which tools to use for which projects

Part II: - *On your own!*

Practice in SPSS

"Comparing motivations for shopping at Farmer’s markets, CSA’s, or neither."

STATISTICAL CONSULTING

- *Dept. of Statistics, Indiana Statistical Consulting Center (ISCC) – all fields & topics*
www.indiana.edu/~iscc
- *Dept. of Epidemiology & Biostatistics Consulting Center – health-related research*
go.iu.edu/epi_bio_consulting

Free Consultation & Funded Collaboration

- ✘ What kind of analysis should I use to answer my research questions?
- ✘ What is the statistical output telling me about me data?
- ✘ How do I address the reviewer's comments about the stats in the manuscript I submitted?

SSRC in Woodburn 200: M-F 9-12

Appointments recommended



RESOURCES

At IU:

- ✘ Research Analytics (UITS) - software support
- ✘ Center for Survey Research (CSR)
- ✘ WIM/ISCC workshops: <http://ssrc.indiana.edu/seminars/wim.shtml>
- ✘ UITS training <http://ittraining.iu.edu/training> (SPSS, SAS,...)
- ✘ Stats courses <http://statclasses.indiana.edu>

On the web:

- ✘ UCLA Stats Consulting <http://www.ats.ucla.edu/stat/>

Books:

- ✘ *Discovering Statistics Using SPSS*, by Andy Field

SOFTWARE

- × **SPSS** – easy “point & click”, good for most “off the shelf” analyses
- × **STATA** – syntax w “point & click”- political science, sociology,...
- × **SAS** – syntax - industry standard, public health,...
- × **R** – free & flexible (but less documented and maintained)
- × **MATLAB** – powerful numerical computing, matrix manipulations
- × **JMP** – “point & click”, good mix of stats and graphs

- × **IUanyWare.iu.edu**
Free software, streaming online

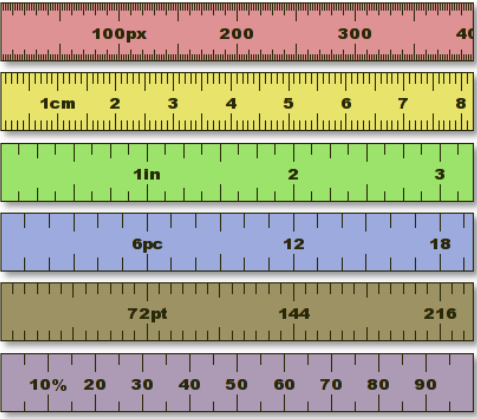


- × **cloudstorage.iu.edu**
 - × **Box.iu.edu**
 - × **File server**





MATERIALS

DATA TYPES

Data Type		Examples
Continuous/ Interval/ Scale		Test score Height, weight, age Response Time <Percent, proportions > <Likert-type items> <Counts>
Ordinal		Educ: 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 ITEMS

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)

INDEPENDENT OBSERVATIONS?

...if each observation is a random “independent” draw from the larger population.

Only one measure for each person in the analysis

Important to know the structure because:

- ✘ Most standard analyses (T-test, ANOVA, Regression, Chi-square,...) assume independent observations.
- ✘ This assumption is built in to the calculation of the p-value for “significant” inferences.

CORRELATED DATA

- ✘ Multiple measurements within subject across time, condition, or item (Repeated Measures)
 - + Panel data, ex: countries across years
- ✘ Observations are clustered in groups (Random effects/HLM)
 - + Students within class, class within school
 - + Mice within litters
 - + Plants within plots

...REPEATED MEASURES

“wide” format

ID	Group	Week1	Week2	Week3
1	Trt	142	139	120
2	Control	155	156	135
3	Trt	151	149	150
Etc...				

“long” format

ID	Group	Week	Sys BP
1	Trt	1	142
1	Trt	2	139
1	Trt	3	120
2	Control	1	155
2	Control	2	156
2	Control	2	135
Etc			

...RANDOM EFFECTS (CLUSTERED)

ID	School	Treatment	Math	Reading
1	A	Trt	256	189
2	A	Trt	213	178
3	B	Cntrl	354	210
4	C	Trt	187	190
5	B	Cntrl	210	221
6	D	Cntrl	185	196
Etc...				

- ✘ Subjects are clustered within school
- ✘ May need random effect for School...



TOOLS

EXAMPLE

Local Food in Indiana

Comparing consumers
(n=302) who purchase food
at Farmer's Markets, CSA's,
or neither, in their
motivations towards local
food.



SURVEY

1. Please indicate your level of agreement for the following statements on a scale from **Strongly Disagree (SD)**, **Disagree (D)**, **Neutral (N)**, **Agree (A)**, to **Strongly Agree (SA)**.

	SD	D	N	A	SA
Purchasing organically grown food is very important to me.					
I give preference to foods that are grown with few chemical applications.					
I give preference to foods that were picked just a few days before my purchase.					
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.					
I give preference to animal products that have been derived in a humane manner.					
I give preference to animal products that are free from growth hormones.					
The expense of fresh local produce deters me from purchasing it as often as I would like.					
I generally purchase whole foods, rather than processed foods.					
I give preference to purchasing foods that come from within 100 miles of my location.					
I give preference to eating foods that are in season, for example, tomatoes in July-October.					
I give preference to food purchase decisions that support the local economy.					
I give preference to food purchase decisions that support local farmers.					
I believe consuming food produced locally is better for the environment.					

DATA

SPSS Data View

- ✗ Columns are 'Variables'
- ✗ Rows are subjects, or 'Observations'

CSA Farmer Market reduced SD.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

Visible: 27 of 27 Variables

	VENUETYPE	Q1MOTORG ANIC	Q1MOTFEW CHEM	Q1MOTFRES H	Q1MOTNU...	Q1MOTANIM HUMA	Q1MOTANIH ORMONE	Q1MOTEXPE NSE	Q1MOTWHO LE	Q1MOT100M LES	Q1MOT SOM
1	2.00	4.00	.	4.00	4.00	5.00	5.00	4.00	4.00	4.00	
2	2.00	4.00	5.00	4.00	5.00	5.00	5.00	4.00	4.00	4.00	
3	1.00	5.00	5.00	4.00	5.00	5.00	5.00	2.00	5.00	4.00	
4	1.00	5.00	5.00	5.00	5.00	5.00	5.00	4.00	4.00	5.00	
5	2.00	4.00	4.00	5.00	4.00	5.00	4.00	3.00	2.00	2.00	
6	1.00	4.00	5.00	4.00	4.00	5.00	5.00	1.00	4.00	4.00	
7	3.00	3.00	3.00	4.00	2.00	3.00	4.00	2.00	2.00	3.00	
8	1.00	4.00	4.00	4.00	5.00	4.00	4.00	1.00	5.00	5.00	
9	2.00	5.00	5.00	5.00	5.00	5.00	5.00	1.00	4.00	5.00	
10	2.00	4.00	2.00	4.00	4.00	.00	2.00	4.00	3.00	3.00	
11	2.00	4.00	5.00	5.00	4.00	5.00	5.00	4.00	4.00	5.00	
12	1.00	4.00	5.00	4.00	4.00	5.00	4.00	2.00	5.00	3.00	
13	1.00	4.00	3.00	3.00	4.00	3.00	4.00	3.00	4.00	4.00	

Data View Variable View

IBM SPSS Statistics Processor is ready

VARIABLES

SPSS Variable View

*CSA Farmer Market reduced SD.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	VENUETYPE	Numeric	8	2		{1.00, CSA}...	None	9	Right	Nominal
2	Q1MOTORGANIC	Numeric	8	2	Purchasing organi...	{1.00, stron...	.00	8	Right	Scale
3	Q1MOTFEWCHEM	Numeric	8	2	I give preference t...	{1.00, stron...	.00	8	Right	Scale
4	Q1MOTFRESH	Numeric	8	2	I give preference t...	{1.00, stron...	.00	8	Right	Scale
5	Q1MOTNUTR	Numeric	8	2	The nutritional val...	{1.00, stron...	.00	8	Right	Scale
6	Q1MOTANIMHUMA	Numeric	8	2	I give preference t...	{1.00, stron...	.00	8	Right	Scale
7	Q1MOTANIHORMONE	Numeric	8	2	I give preference t...	{1.00, stron...	.00	8	Right	Scale
8	Q1MOTEXPENSE	Numeric	8	2	The expense of fre...	{1.00, stron...	.00	8	Right	Scale
9	Q1MOTWHOLE	Numeric	8	2	I generally purcha...	{1.00, stron...	.00	8	Right	Scale
10	Q1MOT100MILES	Numeric	8	2	I give preference t...	{1.00, stron...	.00	8	Right	Scale
11	Q1MOTSEASON	Numeric	8	2	I give preference t...	{1.00, stron...	.00	8	Right	Scale
12	Q1MOTLOCALECON	Numeric	8	2	I give preference t...	{1.00, stron...	.00	8	Right	Scale
13	Q1MOTLOCALFARME...	Numeric	8	2	I give preference t...	{1.00, stron...	.00	8	Right	Scale
14	Q1MOTENVIRON	Numeric	8	2	I believe consumin...	{1.00, stron...	.00	8	Right	Scale
15	Q22GEND	Numeric	8	2		{1.00, FEM...	.00	8	Right	Nominal
16	Q23AGES	Numeric	8	2		None	.00	8	Right	Scale
17	Q24RELAT	Numeric	8	2		{1.00, SING...	.00	8	Right	Nominal
18	Q25ETHNICIT	Numeric	8	2		{1.00, AFRI...	.00	18	Right	Nominal
19	Q26PEOPLEHOUSE	Numeric	8	2		None	None	8	Right	Scale

Data View Variable View



DESCRIBING AND EXPLORING

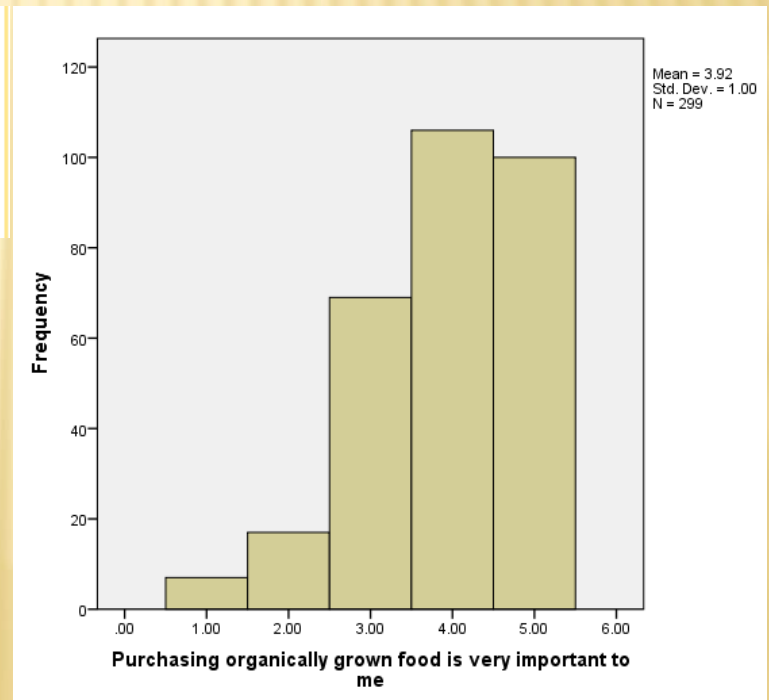
DESCRIBING

Continuous (Scale) Variables

- ✘ Histograms, QQ-plot
- ✘ Descriptive Stats (Mean, SD, Median, Min, Max)

	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

Most analyses (T-test, ANOVA, Regression, etc) prefer a Normal (bell-shaped), symmetric distribution...



Categorical (Discrete) Variables

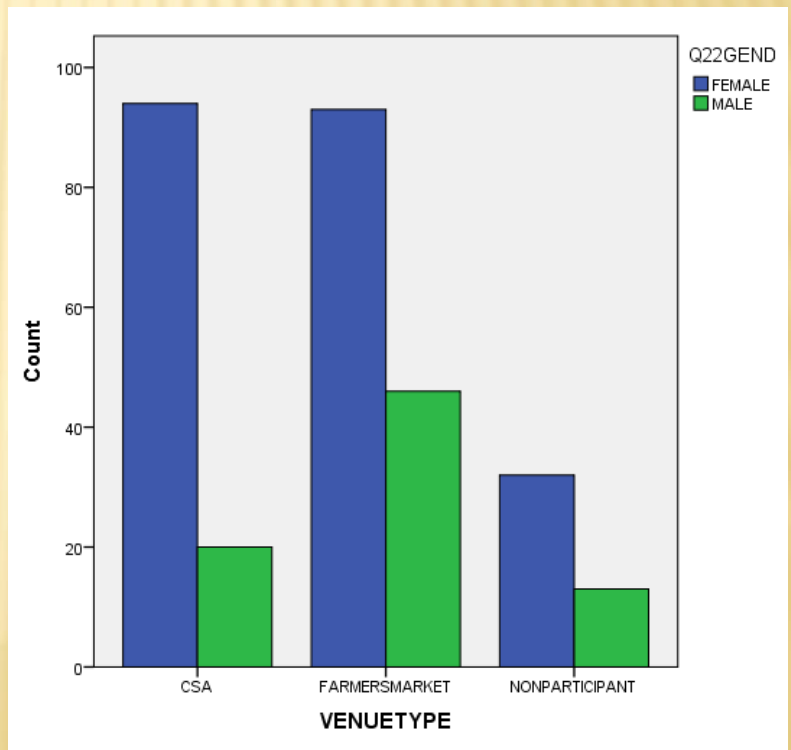
✘ Frequency tables: Counts & Percentages

VENUETYPE

		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	

Q22GEND

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	FEMALE	219	72.5	73.5	73.5
	MALE	79	26.2	26.5	100.0
	Total	298	98.7	100.0	
Missing	.00	3	1.0		
	System	1	.3		
	Total	4	1.3		
Total		302	100.0		



DATA CLEANING

- ✘ Re-codes, grouping
- ✘ Transformations (Log, square-root, etc)
- ✘ Summary scores
- ✘ Re-structure

Created 3 summary scores by averaging across responses in grouped items.

Purchasing organically grown food is very important to me.	A
I give preference to foods that are grown with few chemical applications.	A
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
I give preference to animal products that have been derived in a humane manner.	A
I give preference to animal products that are free from growth hormones.	A
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
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

EXPLORING RELATIONSHIPS

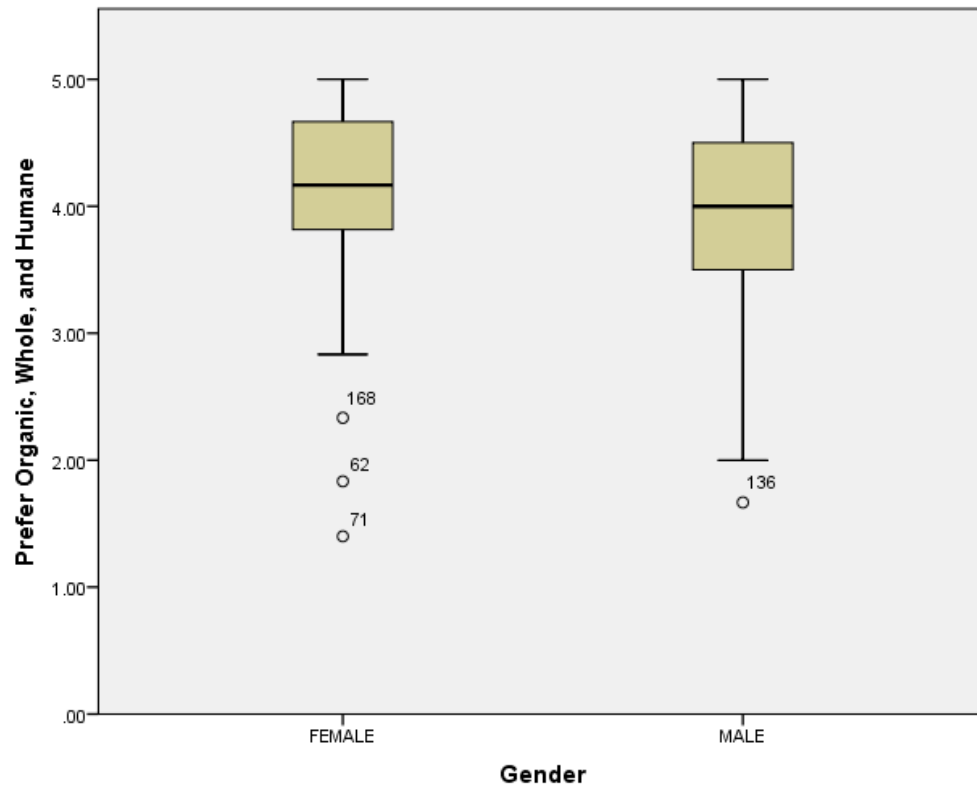
1 Continuous w/ 1 Categorical variable

- ✗ Comparison of Means
- ✗ Boxplots

Prefer Organic, Whole, and Humane

Gender	N	Mean	Std. Deviation
FEMALE	219	4.1548	.66674
MALE	79	3.8570	.75965
Total	298	4.0758	.70370

Summary Score (Average)

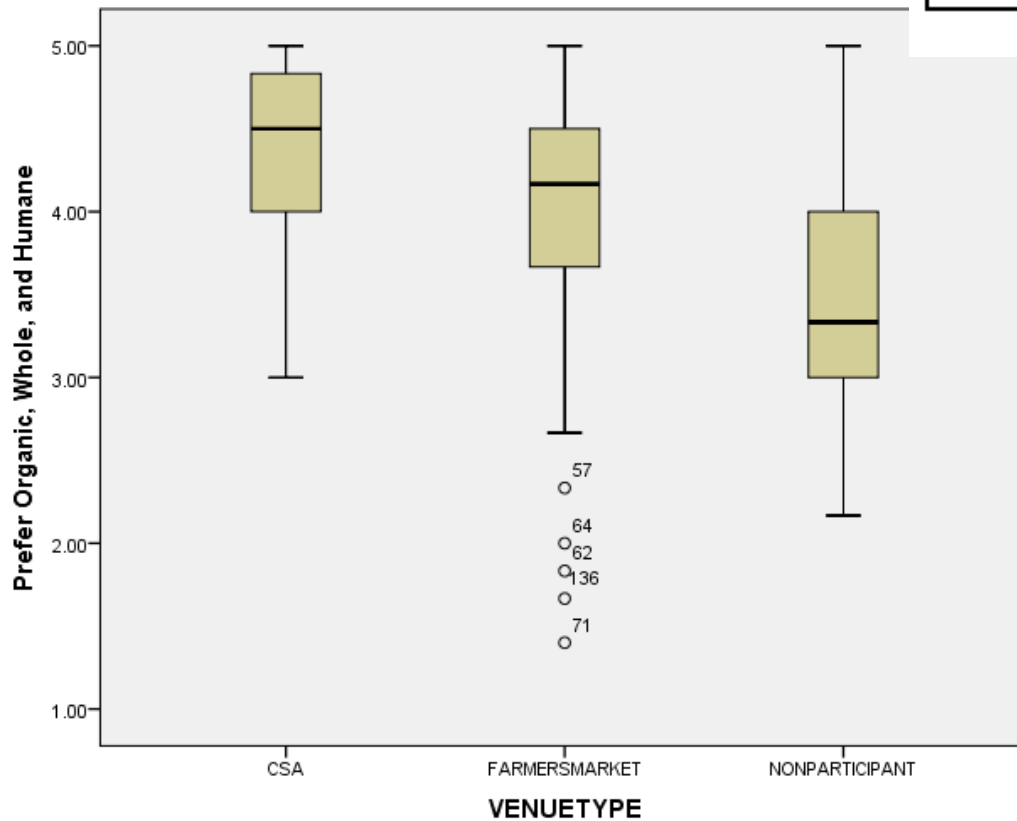


A box-plot is kind-of a histogram on its side.

*Comparing 2 groups...
Think about T-test...*

Prefer Organic, Whole, and Humane

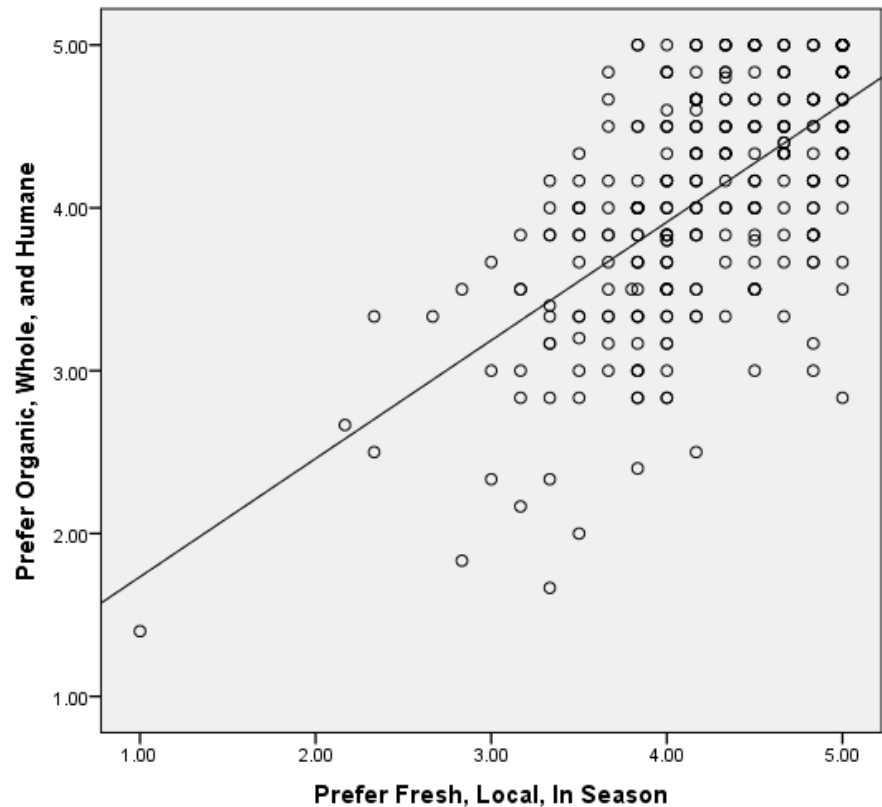
VENUETYPE	N	Mean	Std. Deviation
CSA	114	4.3547	.52182
FARMERSMARKET	142	4.0408	.72734
NONPARTICIPANT	46	3.4826	.67870
Total	302	4.0743	.70902



*3 or more groups...
Think about ANOVA...*

2 Continuous variables

- ✗ Correlation
- ✗ Scatterplot



Correlations

		Prefer Organic, Whole, and Humane	Prefer Fresh, Local, In Season
Prefer Organic, Whole, and Humane	Pearson Correlation	1	.616**
	Sig. (2-tailed)		.000
	N	302	302
Prefer Fresh, Local, In Season	Pearson Correlation	.616**	1
	Sig. (2-tailed)	.000	
	N	302	302

Think about
Pearson correlation...

2 Categorical variables

- ✗ Crosstabs
- ✗ Comparison of Proportions

Income ^ VENUETYPE Crosstabulation

			VENUETYPE			Total
			CSA	FARMERSMARKET	NONPARTICIPANT	
Income	0-29K	Count	8	24	8	40
		% within VENUETYPE	7.5%	19.0%	19.0%	14.6%
	30-45K	Count	13	24	13	50
		% within VENUETYPE	12.3%	19.0%	31.0%	18.2%
	45-59K	Count	11	16	7	34
		% within VENUETYPE	10.4%	12.7%	16.7%	12.4%
	60-74K	Count	10	16	6	32
		% within VENUETYPE	9.4%	12.7%	14.3%	11.7%
	75-89K	Count	11	11	1	23
		% within VENUETYPE	10.4%	8.7%	2.4%	8.4%
	90+K	Count	53	35	7	95
		% within VENUETYPE	50.0%	27.8%	16.7%	34.7%
Total		Count	106	126	42	274
		% within VENUETYPE	100.0%	100.0%	100.0%	100.0%

*Think about Pearson
Chi-square test...*



MAKING PLANS

OUTCOMES & PREDICTORS

General Linear Models

...The set of tools for modeling one (or more) outcome(s) (Y) as a function of one or more predictors (X).

Dependent Variable (DV) = the *outcome* measure (Y)

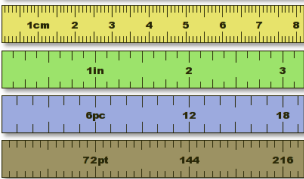
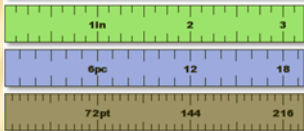


Independent Variable (IV) = the *predictor* variable(s) (X's)

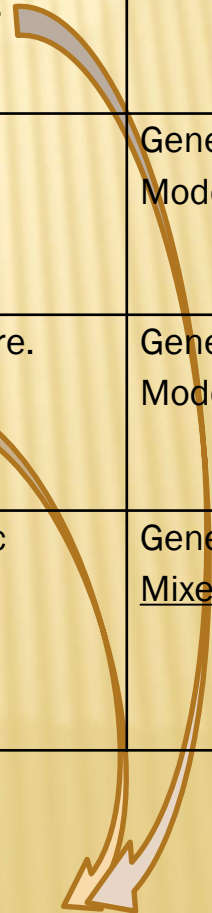
$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \varepsilon$$



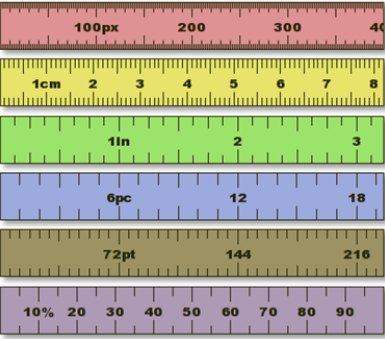

- ✘ GLM is a framework for ANOVA, Regression, etc
- ✘ Can be used hypothesis tests for research questions:
 - + Is there a difference between groups (sex (X1)) in some variable (height (Y))?
 - + Is there an association between one variable (tree density (X1)) on some outcome (seedling density (Y)), controlling for other covariates (X2, etc)?

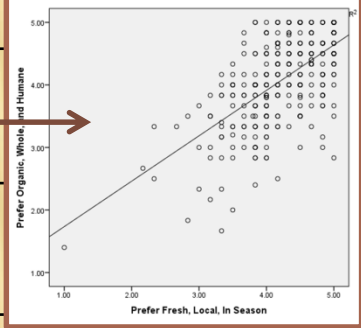
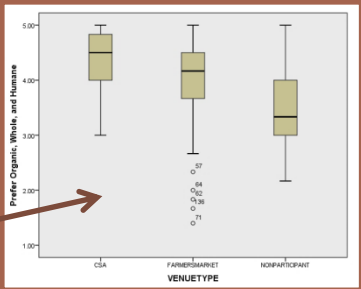
Which bucket of tools do you use with given materials?

DV	Data structure	Analyses	Model type
DV is Continuous 	Independent Observations	T-test, Correlation, ANOVA, ANCOVA, Repeated Measures ANOVA, Linear Regression.	General Linear Model
DV is Continuous 	Correlated Data	“Mixed” Models. Repeated Measures. Random Effects (HLM).	General Linear <u>Mixed</u> Model
DV is Categorical 	Independent Observations	Crosstab, Pearson Chi-square. Logistic Regression. Poisson, Neg. Binomial.	Generalized <u>Linear</u> Model
DV is Categorical 	Correlated Data	Repeated Measures Logistic Regression. GEE, GLIMMIX	Generalized <u>Linear</u> <u>Mixed</u> Model



...zooming in on Independent Observations

DV	IV	Analyses
DV is Continuous 	IV is Categorical IV is Continuous Any IV's	T-test (1 IV: 2 groups (Binary)), One way ANOVA (1 IV: >2 groups), Two-way ANOVA (2 IV's) Factorial ANOVA (>2 IV's) Pearson Correlation (1 IV) Simple Linear Regression (1 IV) Multiple Linear Regression (>1 IV) ANCOVA Multiple Linear Regression
Multiple DV's (Continuous)		Paired T-test (1 IV, 2 levels) Repeated Measures ANOVA (≥ 2 levels) MANOVA (≥ 2 DV's)
DV is Counts	Any IV's	Poisson Regression Neg. Binomial Regression.
DV is Categorical 	IV is Categorical Any IV's Any IV's	Pearson Chi-square (1 IV). Logistic Regression (>1 IV). Binary Logistic Regression Multinomial Logistic Regression

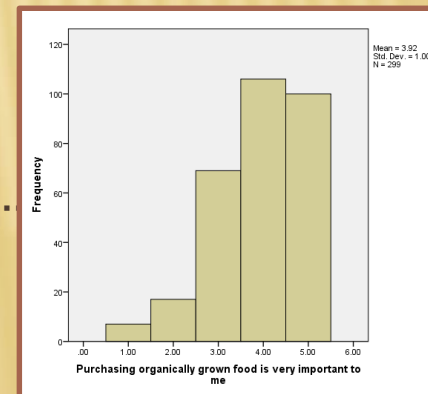
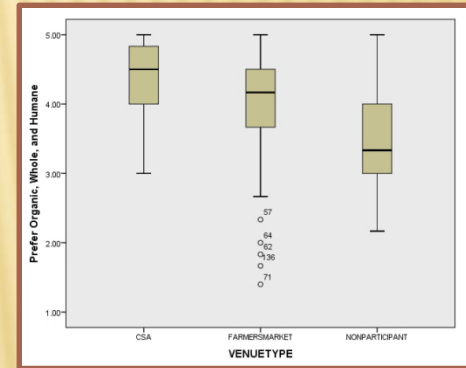


		VENUETYPE		
		CSA	FARMERSMARKET	NONPARTICIPANT
0-29K	Count	8	24	8
	% within VENUETYPE	7.5%	19.0%	19.0%
30-45K	Count	13	24	13
	% within VENUETYPE	12.3%	19.0%	31.0%
45-59K	Count	11	16	7
	% within VENUETYPE	10.4%	12.7%	16.7%
60-74K	Count	10	16	6
	% within VENUETYPE	9.4%	12.7%	14.3%
75-89K	Count	11	11	1
	% within VENUETYPE	10.4%	8.7%	2.4%
90+K	Count	53	35	7
	% within VENUETYPE	50.0%	27.8%	16.7%
	Count	106	126	42

ASSUMPTIONS

Assumptions of GLM (T-test, ANOVA, Regression)

- ✘ Observations are **independent**
(or else modeled appropriately in a Repeated Measures or “Mixed” model)
- ✘ There are **equal variances** between the groups (or across values of continuous predictor variables).
 - + Evaluate standard deviation in each group
 - ✘ boxplots or scatterplot of DV vs IV
 - + *Maybe* use Levene’s test for homogeneity of variance
 - + Residuals have equal variance across levels of IV’s
- ✘ Residuals are **Normally distributed**.
 - + Normally distributed DV is a ‘proxy’ for this...
 - + Inspect histogram, qq-plot, skewness, and kurtosis; boxplot
 - + Shapiro-Wilks tests normality, but p-value not always helpful..



But...

When assumptions are (or seem to be) violated

✘ Not independent observations?

- + Maybe aggregate data to the individual level? (esp. binary data!)
- + Model the correlation structure in Repeated Measures ANOVA or Mixed Models

✘ Not equal variances?

- + Levene's test is only one diagnostic measure... (careful with p-values)
- + What is Std. Dev. in each group? How different are they? Is one SD more than twice as big as the other SD?
- + If sample sizes between groups are equal, ANOVA is robust to this
- + Log transformations of skewed data often help with variances

✘ Not Normally distributed DV/residuals?

- + Be skeptical of tests of normality (Shapiro-Wilks)...p-value is more significant with larger sample size, but...
- + larger sample sizes are more robust (Central Limit Theorem means ANOVA is Robust)
- + Skewness and Kurtosis are helpful (skewness <1 or 2?)
- + Try transformations, like taking the log, square root (or try Box-Cox)

When assumptions are (or seem to be) violated

How bad is too bad?

- ✘ “Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance”, Glass, Peckham and Sanders. 1972 42: 237 *REVIEW OF EDUCATIONAL RESEARCH*

- ✘ Non-parametric tests where possible:
 - + Wilcoxon Rank-sum (comparing 2 groups; T-test)
 - + Kruskal-Wallis (comparing 3 groups; One way ANOVA)
 - + A *little* less powerful.
 - + Still assume independent observations.
 - + More robust

- ✘ Or bootstrap your own p-values...

PUTTING IT IN PRACTICE

SPSS

Note that I am not particularly promoting SPSS over other stats software except that it's the easiest to pick up and use quickly.

If you don't have a copy of SPSS locally, use **IUanyware.IU.edu**

+ Install Citrix client first

Open the CSA Farmer's Market data...

<https://iu.box.com/ISCCWorkshops>



DESCRIPTIVES

For descriptive stats and exploratory plots

- ✘ Analyze > Descriptive Stats > **Descriptives** (Select 'Organic' as Variable)
- ✘ Graphs > Legacy > **Histogram** (Select 'Organic' as Variable)
- ✘ Graphs > Legacy > **Boxplot** > Simple (Select 'Organic' as Variable, and Gender as 'Category Axis')
- ✘ Analyze > Compare Means > **Means** (Select 'Organic' as 'Dependent', and Gender as 'Independent')
- ✘ Graphs > Legacy > **Scatter/Dot** > Simple Scatter (Select 'Organic' as Y-Axis and 'Local' as X-Axis)
- ✘ Analyze > **Correlate** > Bivariate (Select 'Organic' and 'Local')

T-TEST

(Independent Samples)

- ✘ Compare Continuous DV between 2 groups (1 Categorical IV w/ 2 levels)

Is there a difference between men and women in how highly they rate the importance of buying Organic/Whole food?

IV: Gender (M/F)

DV: Organic

Q22GEND	Q23AGES	Q24RELAT	Q25ETHNI...	Q26PEOPLE HOUSE	Q26CHILDR HOUSE	Q27TYPEGR EWUP	Q28TYPELIV ETODAY	Q29RELIGIOSITY	Q32EDUCATION	Q33INCOME	Agegroup	Organic
FEMALE	73.00	SINGLE	WHITE	1.00	.00	URBAN	URBAN	1+ TIMES A WEEK	HS OR GED	0-29K	65+	4.40
MALE	33.00	MARRIED	WHITE	4.00	2.00	RURAL	URBAN	dont attend	PROFESSIONAL...	90+K	18-34	4.67
MALE	42.00	SINGLE	WHITE	1.00	.00	SUBURBAN	URBAN	dont attend	GRADUATE DE...	45-59K	35-44	5.00
FEMALE	36.00	MARRIED	WHITE	4.00	2.00	RURAL	URBAN	dont attend	BA OR BS	45-59K	35-44	4.83
FEMALE	43.00	SINGLE	WHITE	2.00	1.00	RURAL	RURAL	dont attend	SOME COLLEG...	30-45K	35-44	3.83
FEMALE	48.00	SINGLE	WHITE	2.00	.00	URBAN	URBAN	dont attend	GRADUATE DE...	60-74K	45-54	4.50
FEMALE	72.00	.	WHITE	1.00	.00	SUBURBAN	RURAL	1+ TIMES A WEEK	HS OR GED	0-29K	65+	2.83
FEMALE	46.00	MARRIED	WHITE	5.00	3.00	RURAL	RURAL	1+ TIMES A WEEK	BA OR BS	90+K	45-54	4.33
FEMALE	49.00	MARRIED	WHITE	5.00	.00	URBAN	URBAN	1+ TIMES A WEEK	SOME COLLEG...	75-89K	45-54	4.83
FEMALE	67.00	MARRIED	WHITE	2.00	.00	RURAL	URBAN	1+ TIMES A WEEK	BA OR BS	.	65+	3.00
FEMALE	70.00	MARRIED	WHITE	4.00	.00	RURAL	RURAL	several times a year	.	0-29K	65+	4.50
FEMALE	54.00	MARRIED	WHITE	4.00	2.00	SUBURBAN	SUBURBAN	1+ TIMES A WEEK	BA OR BS	90+K	45-54	4.50
FEMALE	41.00	MARRIED	HISPANIC	3.00	1.00	SUBURBAN	RURAL	several times a year	SOME COLLEG...	90+K	35-44	3.67
MALE	32.00	MARRIED	WHITE	3.00	1.00	SUBURBAN	SUBURBAN	1+ TIMES A WEEK	SOME COLLEG...	60-74K	18-34	3.33

T-TEST

...T-test in SPSS

Analyze > Compare Means > Independent Samples T-Test

- ✘ Put DV (Organic) in the 'Test Variable(s)'.
- ✘ Put IV (Gender) in the Grouping Variable. Define Groups 1 and 2.

Output:

- ✘ Inspect Descriptive Stats.
- ✘ Check Levene's test.
- ✘ Use corresponding "Sig." value (= p-value)

There is not a significant difference between males and females in how important organic food is to them.

ANOVA

- ✦ Compare Continuous DV between 3 or more groups (1 Categorical IV w/ 3+ levels)

Is there a difference between the three venues in how highly respondents rate the importance of buying Organic/Whole food?

IV: Venue (CSA, Farmer's Market, neither)

DV: Organic

(note box-plot above)



ANOVA

...One-way ANOVA in SPSS

Analyze > Compare Means > One-way ANOVA

- ✘ Put DV (Organic) as 'Dependent', and Venue as 'Factor'
- ✘ 'Post Hoc' > Tukey
- ✘ Options > Descriptives and Homogeneity of variance.

Output:

- ✘ Inspect Descriptive Stats.
- ✘ Check Levene's test.
- ✘ Use "Sig." value (= p-value) from ANOVA table

There is a significant difference between the three venues in how respondents rate Organic food ($F(2,301)=29.9$, $p<.001$). CSA members give the highest ratings for the Organic/Whole/Animal items ($M=4.35$, $SE=.05$), followed by the Farmer's Market participants ($M=4.04$, $SE=.06$), and lastly those who use neither ($M=3.5$, $SE=.10$). All pairwise differences are significant (Tukey, p 's $<.001$).

Note!

Post-hoc tests (comparing trt 1 vs 2, 1 vs 3, and 2 vs 3) are begging for a p-value "correction" so that you don't over-test your data.

Bonferroni is easy (takes p-value times # of comparisons, or $\alpha/\text{comparisons}$) but too conservative/stingy.

Tukey is more accurate.

ANOVA

For more than one Categorical IV...

Is there a difference between the three venues AND by income in how highly respondents rate the importance of buying Organic/Whole food?

IV: Venue (CSA, Farmer's Market, neither); Income levels

DV: Organic

ANOVA

...GLM in SPSS

Analyze > General Linear Model > Univariate

- ✘ Put DV (Organic) as 'Dependent'.
- ✘ Put Gender and Income as 'Fixed Factors'.
- ✘ Click 'Model' to specify interactions.
(“ANOVA” usually thinks about interactions...)
- ✘ Post Hoc > Tukey
- ✘ (Optional) Save > Standardized Residuals
- ✘ Options > Display Means for: Gender Income Gender*Income
(Note: 'Compare main effects' can do Bonferroni or Sidak, but Tukey not an option here)
(Optional) Get 'Descriptive Stats', 'Estimates of effect size', 'Homogeneity tests'


Output:

- ✘ Tests of Between-Subjects Effects (F-tests & “sig” p-values)
- ✘ Estimated Marginal Means
- ✘ Post-hoc tests

SPSS note!

'Fixed Factors' are for
Categorical variables.

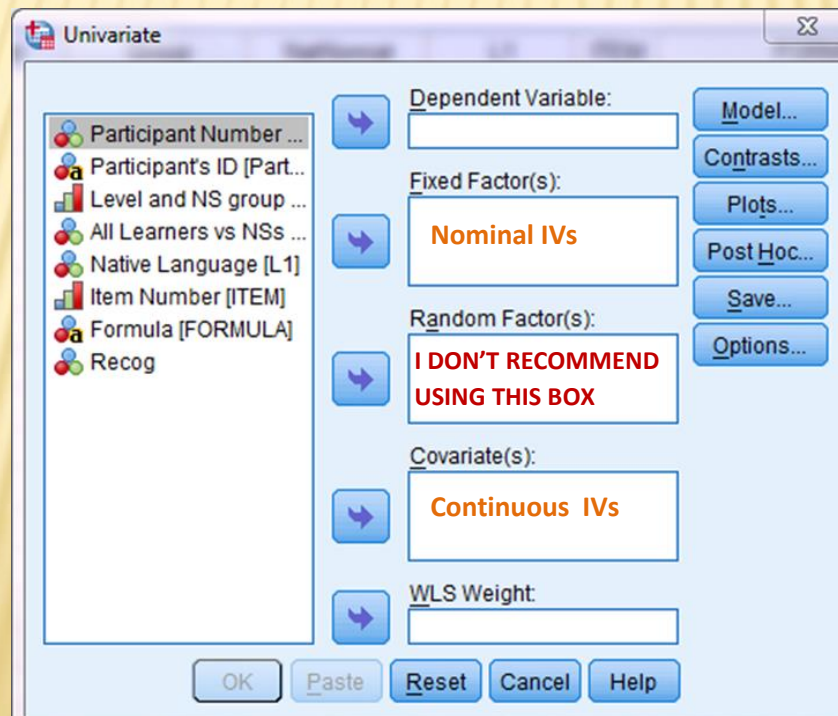
'Covariates' are for
Continuous variables.



ANOVA

“Factors” vs “Covariates” in GLM

- ✘ In SPSS, “Factors” are any categorical IV. “Covariates” are any continuous IV.
- ✘ Regression procedure only permits continuous variables or dummy (0/1)
- ✘ In SAS, the “Class” statement is for any categorical IV. Others are continuous.
- ✘ Your model will be “wonky” to say the least if you mix them up...
- ✘ Do Random effects in Linear Mixed Model rather than ANOVA



ANCOVA

- ✘ Compare Continuous DV between groups (Categorical IV), adjusting for Continuous “**covariate**” IV

What’s the difference in ‘Organic’ ratings between venues and gender, controlling for age?

IV: Venue (CSA, FM, neither); Gender (M/F); Age

DV: Organic

ANCOVA

... in SPSS

Analyze > General Linear Model > Univariate

- ✘ Put DV (Organic) as 'Dependent'.
- ✘ Put Venue and Gender as 'Fixed Factors'.
- ✘ Put Age as 'Covariate'
- ✘ (other options same as above)
- ✘ Also, Options > Parameter Estimates (to get 'slope' for continuous variables: age)

Output:

- ✘ Tests of Between-Subjects Effects (F-tests & "sig" p-values)
- ✘ Parameter estimates for continuous variables
- ✘ Estimated Marginal Means for categorical variables
- ✘ Post-hoc tests for categorical variables

MANOVA (OR MANCOVA)

- ✘ Compare more than 1 Continuous (related) DV between groups (Categorical IV), adjusting for Continuous “covariate” IV

What’s the difference in all 13 of the food motivations by Venue and Gender, and adjusting for age?

IV: Venue (CSA, FM, neither); Gender (M/F); Age

DV: Item #1, 2, 3....13 (Q1MOTORGANIC, Q1MOTFEQCHEM, etc)

MANOVA (OR MANCOVA)

... in SPSS

Analyze > General Linear Model > Multivariate

- ✘ Put Q1MOTORGANIC, Q1MOTFEQCHEM, etc as 'Dependent'.
- ✘ Put Gender and Venue as 'Fixed Factors'.
- ✘ Put Age as 'Covariate'
- ✘ (other options same as above)

Output:

- ✘ Multivariate Tests (the gatekeeper to individual ANOVA's, $p < .05$?)
- ✘ Tests of Between-Subjects Effects (F-tests & "sig" p-values)
- ✘ Parameter estimates for continuous variables
- ✘ Estimated Marginal Means for categorical variables

Note (!) that the only difference between MANOVA and separate ANOVA's is the omnibus "gatekeeper" tests first. The following "tests of between-subjects effects" are the same as if you had run separate ANOVA's.

PAIRED T-TEST

- ✘ Compare 2 Continuous DV's "paired" within subject

Do people rate the importance of buying organic food higher than the expense which might deter them?

DV: Q1MOTORGANIC, Q1MOTEXPENSE

IV: (NA)

Or could call the DV the "rating" while the IV is the "motivation (A or B)"

PAIRED T-TEST

... in SPSS

Analyze > Compare Means > Paired samples t-test

- ✘ Put Q1MOTORGANIC and Q1MOTEXPENSE into 'paired variables'

Output:

- ✘ Inspect Descriptive Stats and Mean difference
- ✘ Find "Sig" level

Note: You could run this same analysis as a "Repeated Measures" (under General Linear Model) by leaving the factors and covariate blank...see below.

There is a significant difference ($p < .001$) where the respondents overall rated the organic motivation higher than the deterrent of the expense.

However, what if you did the analysis separately by VENUE?...

- ✘ Data > Split File > Compare groups by Venue

REPEATED MEASURES ANOVA

- ✘ Compare multiple Continuous DV's within-subject, and also IV's between-subject

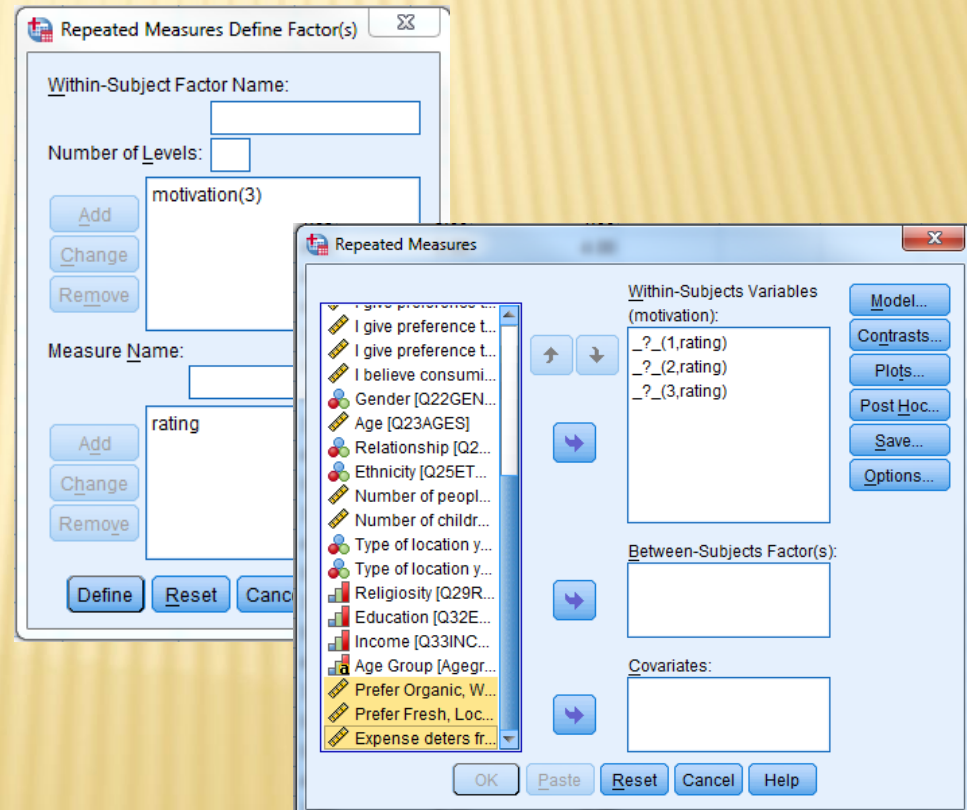
Which motivation do people rate the highest: 'Organic/Whole Food & Animals', 'Local & Fresh', or 'Too expensive'? How does it depend on survey venue?

DV: (Ratings of) Organic, Local, Expensive

IV: Venue; 'Motivation'

Note! 'Motivation' is not a variable in your dataset, but you will have to label the 'within-subject' variable defined by the three motivations.

(See example in 2nd half)



REPEATED MEASURES ANOVA

... in SPSS

Analyze > General Linear Model > Repeated Measures

- ✘ Put 'motivation' as the Within-subject Factor, with 3 levels
- ✘ (optional) Name the measure Ratings
- ✘ Enter Organic, Local, Expensive as the 'within-subject' variables
- ✘ Enter Venue as a 'between-subject' factor
- ✘ Consider 'Model' or 'Plots'
- ✘ 'Post-Hoc' for Trt, with Tukey
- ✘ 'Options' > Display Means for everything

Output:

- ✘ Multivariate tests > Wilk's lambda
- ✘ Or, Tests of Within-subject effects > Sphericity assumed*
- ✘ Tests of Between-subject effects
- ✘ Estimated Marginal Means & Plots

Can always use Wilks-Lambda, but others (Pillai's trace, etc) might be more powerful in some cases.

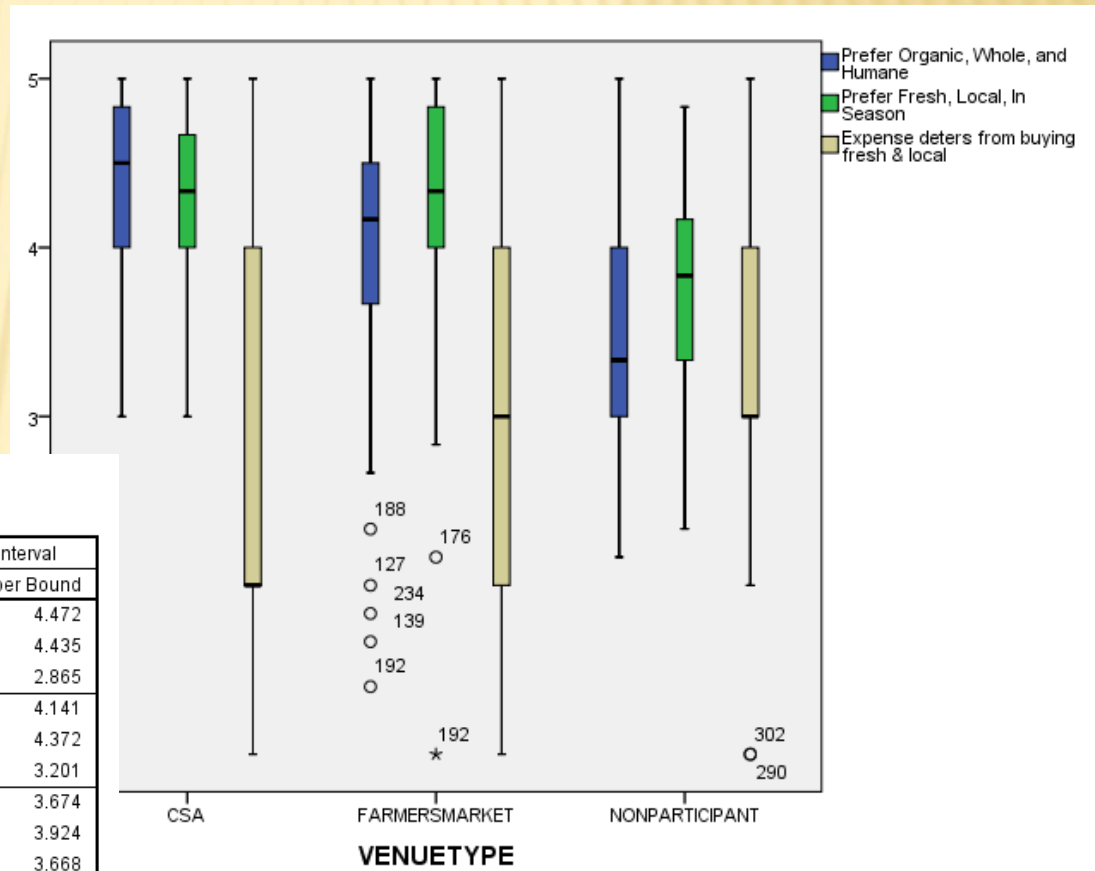
* If Mauchy's test is significant $p < .05$, we can NOT assume Sphericity (simple correlation structure). But sometimes more powerful if sphericity satisfied.

REPEATED MEASURES ANOVA

Results

Both main effects and interaction are significant.

- ✘ There is an overall difference between the 3 motivations.
- ✘ There is an overall difference between the 3 venues.
- ✘ The difference between the motivations is different across the venues.



4. VENUETYPE * motivation

Measure: Rating

VENUETYPE	motivation	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
CSA	1	4.351	.062	4.229	4.472
	2	4.329	.054	4.222	4.435
	3	2.625	.122	2.385	2.865
FARMERSMARKET	1	4.033	.055	3.924	4.141
	2	4.276	.048	4.181	4.372
	3	2.986	.109	2.771	3.201
NONPARTICIPANT	1	3.482	.097	3.291	3.674
	2	3.756	.085	3.587	3.924
	3	3.289	.193	2.910	3.668

...More Practice

Repeated Measures ANOVA for Diet Study

This is a hypothetical data file containing the results of a study of a hypothetical diet (loosely based on the "Stillman diet" (Rickman et al., 1974)). Each case corresponds to a separate subject, and records their weights in pounds and triglyceride levels in mg/100 ml at five stages of the diet.

We want to know if the patients' weight (DV) decreases over time (IV, factor), and is weight-loss related to age (IV, covariate) and gender (IV, factor).

✘ **Data & Demo** : <https://iu.box.com/ISCCWorkshops>

2014-09-01 Statistical Toolbelt folder

- + Diet Study Demos:
- + 'dietstudy.sav' – dataset
- + 'Diet Study RM ANOVA_demo.pdf' – slideshow for SPSS commands



Diet Study RM ANOVA_demo.pdf

Created Today by Stephanie Dickinson · 13.9 MB



dietstudy.sav

Created Today by Stephanie Dickinson · 2.2 KB

CORRELATION

- ✘ Test for relationship between 2 Continuous variables (~IV & 1 DV)

What's the correlation (or association/ relationship) between age and each of the three motivations?

- + Organic, Local, Expensive

CORRELATION

...in SPSS

Analyze > Correlate > Bivariate

- ✘ Enter Age and the three motivations (Organic, Local, Expensive).
- ✘ Check Pearson and/or Spearman (Non-parametric test)

Output:

- ✘ Pearson r correlation value (or Spearman rho)
- ✘ Corresponding p-value
- ✘ Sample size (N)

Note that the Pearson correlation (r) is the square root of the R-squared from a *simple linear regression* ...

REGRESSION

Multiple Linear Regression

- ✘ Test for effect of one (or more) predictor variables (IV: any type) on one Continuous outcome (DV)

DV: Expensive

IV's: Venue, Age, Gender, Income, Education

REGRESSION

Linear Regression in SPSS

- ✘ For continuous IV's (or dummy variables 0/1):
 - + Analyze > Regression > Linear
 - + Note: 'Method': 'Enter' to enter all IV's simultaneously, or 'Stepwise' selection

- ✘ For continuous and categorical IV's:
 - + Analyze > General Linear Model > Univariate
 - + Enter Age as a 'covariate'
 - + Enter Gender as a 'factor'
 - + (same options and output as above, but make sure you get **Parameter Estimates**)
 - + Parameter estimates are what make it more 'regression-like'.
 - + Means and 'forced' interactions make it more 'anova-like'.
 - + Note: If we would enter 'trt' as a factor, this analysis would be *identical* to the ANCOVA above!

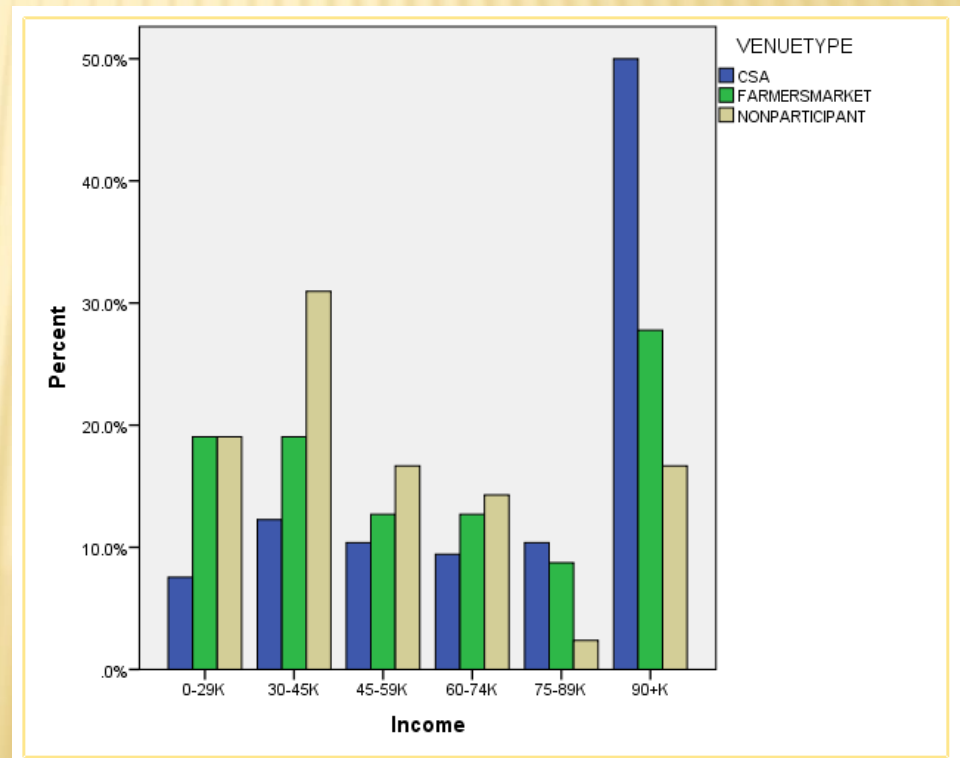
CHI-SQUARE TEST

Pearson Chi-square Test

- ✘ Test for relationship between 2 Categorical vars, also a comparison of proportions

What's the difference in age group distribution between respondents in the three venues?

Venue (CSA, FM, neither)
Age group



CHI-SQUARE TEST

...in SPSS

Analyze > Descriptive Stats > Crosstabs

- ✘ Enter Income as 'Rows' and Venue as 'Columns' (or vice-versa)
- ✘ Statistics > Chi-square
- ✘ Cells > Percentages > Columns (or Rows)
- ✘ Since we sampled by Venue, this will give the % in each age group by Venue.

Output:

- ✘ Frequencies & Percentages
- ✘ 'sig' p-value from Pearson Chi-square

Income ^ VENUETYPE Crosstabulation

			VENUETYPE			Total
			CSA	FARMERSMA RKET	NONPARTICI PANT	
Income	0-29K	Count	8	24	8	40
		% within VENUETYPE	7.5%	19.0%	19.0%	14.6%
	30-45K	Count	13	24	13	50
		% within VENUETYPE	12.3%	19.0%	31.0%	18.2%
	45-59K	Count	11	16	7	34
		% within VENUETYPE	10.4%	12.7%	16.7%	12.4%
	60-74K	Count	10	16	6	32
		% within VENUETYPE	9.4%	12.7%	14.3%	11.7%
	75-89K	Count	11	11	1	23
		% within VENUETYPE	10.4%	8.7%	2.4%	8.4%
	90+K	Count	53	35	7	95
		% within VENUETYPE	50.0%	27.8%	16.7%	34.7%
Total		Count	106	126	42	274
		% within VENUETYPE	100.0%	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	28.676 ^a	10	.001
Likelihood Ratio	29.739	10	.001
Linear-by-Linear Association	23.964	1	.000
N of Valid Cases	274		

a. 2 cells (11.1%) have expected count less than 5. The minimum expected count is 3.53.

LOGISTIC REGRESSION

- ✘ Test for effect of any IV's on 1 Categorical DV with 2 or more levels.
 - + 2 levels for DV (Yes/No) is Binary Logistic
 - + 3 or more levels for DV is Multinomial

What variables are most “predictive” of your food shopping group (CSA, FM, neither) ?

DV: Group (CSA, FM, neither)

IV: Age, Gender, Race, Education, Income



Beware!
Multinomial (3+
groups) can be a
BEAR to interpret!

LOGISTIC REGRESSION

...in SPSS

Analyze > Regression > Multinomial [usually Binary]

- ✘ Enter Venue as 'Dependent'.
- ✘ Enter Age, as 'Covariates'; Gender, and Race as Factors
- ✘ For Binary: if you have numeric categorical variables, enter as covariates and click 'Categorical' to specify.
- ✘ Note: Method > 'Enter' or 'Forward: LR'

Output:

- ✘ DV Encoding (Predicting '1' vs '0')
- ✘ Categorical variable coding (reference levels '0')
- ✘ Block 0 - Variables not in the Equation
- ✘ Block 1 - Variables in the Equation

LINEAR MIXED MODEL

Correlated data...

- ✘ Longitudinal data, Panel data, Hierarchical Linear Models
- ✘ Data in “long” format
- ✘ Better than RM ANOVA if missing data across repeated measures
- ✘ Necessary if IV’s are also changing across repeated measures (“time varying covariates”)

Repeated Measures

- ✘ ...if you can enumerate/items measurements across time or task
- ✘ ex: each person is measured once a year for 5 years, or each person does 5 different tasks, or you measure response time for 32 different trials

Random Effects

- ✘ ...if you cannot enumerate specific items but there is just a “bucket” of observations for each subject/group, then subject (or group) is the Random effect.
- ✘ ex: students within class or school (HLM), words spoken by person

LINEAR MIXED MODEL

...For Practice

Example: Diet Study

- ✘ Diet data is restructured into “long” form with multiple rows for each subject. ('Dietstudy_long.sav')
- ✘ Note that under some circumstances the LMM on “long” data can be identical to the RM ANOVA on “wide” data. ('time' is categorical, no missing data, 'compound symmetry' correlation structure)
- ✘ **Data & Demo** : <https://iu.box.com/ISCCWorkshops>
 - + '2014-09-01 Statistical Toolbelt' folder
 - + 'dietstudy_long.sav' – dataset
 - + 'Diet Study LMM_demo.pdf' – slideshow for SPSS commands
- ✘ More info:
 - + '2011-10-04 GLM Workshop' folder
 - + 'GLM workshops slides Part 2 2011-10-03.pdf'
 - + UCLA stat computing <http://www.ats.ucla.edu/stat/spss/library/spssmixed/mixed.htm>

THE END

- ✘ Please fill out the WIM feedback survey.
- ✘ Let me know any questions:
 - + Stephanie Dickinson (sd3@indiana.edu)