Your Statistical Tool Belt

Stephanie Dickinson
Senior Consultant
Department of Statistics
Indiana Statistical Consulting Center (ISCC)
sd3@indiana.edu

Indiana Statistical Consulting Center

ISCC

"You can build it. We can help."

http://www.indiana.edu/~iscc
iscc@indiana.edu

free consultation, funded collaboration, & educational outreach

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?

Contact us! http://www.indiana.edu/~iscc/

Social Science Research Commons (SSRC) in Woodburn 200

Overview

"Let's build something together"

Get the right tools for the right project.

Workshop outline:

Part I:

1. Your materials (data)
2. Your project (research questions)
3. Your tools (analysis methods)
4. Which tools to use for which projects

Part II:

Practice in SPSS

DIY Resources

At IU:

Stats courses

Research Analytics (UITS) - software support

WIM/ISCC workshops: R coming up…

UITS training http://ittraining.iu.edu/training (SPSS, SAS)

On the web:

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

Books:

Discovering Statistics Using SPSS, by Andy Field

A Guide to Doing Statistics in Second Language Research Using SPSS ,
by Jenifer Larson-Hall

Software

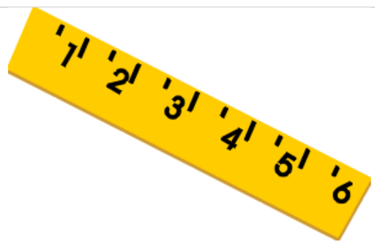
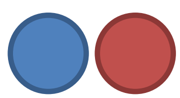
- SPSS – easy “point & click”, should do what you need
- R – free & flexible (but less documented and maintained)
- STATA – very powerful - political science, sociology,…
- SAS – very powerful - industry standard, public health,…

IUanyWare.iu.edu

- Free software, running online
- cloudstorage.iu.edu
    - Box.iu.edu
    - Dropbox
    - File server

Materials

Before you can build, consider the nature of your materials…

Data types		Examples
Continuous/ Interval/ Scale		Test score Height, weight, age Response Time <Percent, proportions > <Likert scales> <Counts>
Ordinal		Educ: Bachelor, Masters, PhD. <Likert scales>
Categorical/ Nominal (≥2) ~ Binary (2 levels)		Treatment Group (A,B,C) Sex: Male/female, Yes/no, right/wrong, 0/1.

NOTE: It's one thing to know the TRUE nature of the data (Continuous, Categorical, Ordinal), but it may be a separate decision how you will TREAT the data in a particular analysis.

Ex:

- Likert Scale (1 to 5) may be truly *Ordinal* (i.e. not truly continuous) but usually needs to be treated as *Categorical* or *Continuous* for standard analyses

There is not a clear "recipe" or a "one-size fits all" for these decisions…

What is meaningful for your data analysis? ...

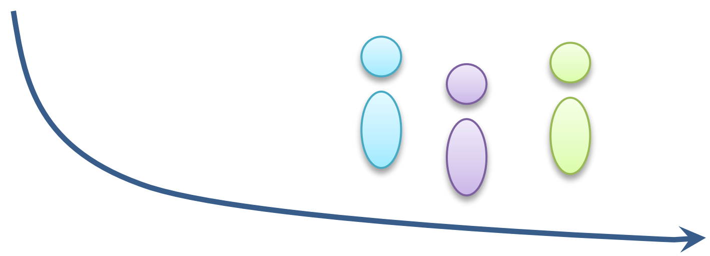
- Are you willing to assume a straight linear relationship across the scale (continuous… as age goes up, blood pressure goes up),
- or do you want to fit separate means for each discrete level (categorical …average blood pressure for separate age groups)?

Independent Observations vs Correlated Data

Independent Observations

…if each observation is a random "independent" draw from the larger population.

Only one measure for each person in the analysis



ID	Group	Test Score
1	Trt	98
2	Trt	83
3	Control	67
4	Control	79
etc…		

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 inferences.

…> *p-value* is the probability of seeing a "significant" effect by chance

Correlated Data

- Multiple measurements within subject across time, condition, or item (Repeated Measures)

ID	Group	Pre	Post
1	Trt	82	98
2	Trt	78	93
3	Control	67	83
etc…			

"wide"

- Observations are clustered in groups (Random effects/HLM)
  - Students within class, class within school

ID	Group	Item	Response time
1	Trt	1	3.2
1	Trt	2	2.5
1	Trt	3	3.7
2	Trt	1	4.2
etc…			

"long"

ID	School	Group	Test Score
1	1	Trt	98
2	1	Trt	93
3	1	Control	67
4	2	Control	79
etc…			

Example data – HSB2 "High School and Beyond"

This data file contains 200 observations from a sample of high school students with demographic information about the students, such as their gender (female), socio-economic status (ses) and ethnic background (race). It also contains a number of scores on standardized tests, including tests of reading (read), writing (write), mathematics (math) and social studies (socst).

SPSS data at http://www.ats.ucla.edu/stat/spss/webbooks/reg/hsb2.sav

*hsb2.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	id	Numeric	9	2		None	None	8	Right	Scale
2	female	Numeric	9	2		{.00, male}...	None	8	Right	Nominal
3	race	Numeric	12	2		{1.00, hispanic}...	None	8	Right	Nominal
4	ses	Numeric	9	2		{1.00, low}...	None	8	Right	Ordinal
5	schtyp	Numeric	9	2	type of school	{1.00, public}...	None	8	Right	Nominal
6	prog	Numeric	9	2	type of program	{1.00, general}...	None	8	Right	Nominal
7	read	Numeric	9	2	reading score	None	None	8	Right	Scale
8	write	Numeric	9	2	writing score	None	None	8	Right	Scale
9	math	Numeric	9	2	math score	None	None	8	Right	Scale
10	science	Numeric	9	2	science score	None	None	8	Right	Scale
11	socst	Numeric	9	2	social studies score	None	None	8	Right	Scale
12										
13										

Variable View (previous slide)

Data View (here)

Columns are “variables”.

	id	female	race	ses	schtyp	prog	read	write	math	science	socst
1	70.00	.00	4.00	1.00	1.00	1.00	57.00	52.00	41.00	47.00	57.00
2	121.00	1.00	4.00	2.00	1.00	3.00	68.00	59.00	53.00	63.00	61.00
3	86.00	.00	4.00	3.00	1.00	1.00	44.00	33.00	54.00	58.00	31.00
4	141.00	.00	4.00	3.00	1.00	3.00	63.00	44.00	47.00	53.00	56.00
5	172.00	.00	4.00	2.00	1.00	2.00	47.00	52.00	57.00	53.00	61.00
6	113.00	.00	4.00	2.00	1.00	2.00	44.00	52.00	51.00	63.00	61.00
7	50.00	.00	3.00	2.00	1.00	1.00	50.00	59.00	42.00	53.00	61.00
8	11.00	.00	1.00	2.00	1.00	2.00	34.00	46.00	45.00	39.00	36.00
9	84.00	.00	4.00	2.00	1.00	1.00	63.00	57.00	54.00	58.00	51.00
10	48.00	.00	3.00	2.00	1.00	2.00	57.00	55.00	52.00	50.00	51.00
11	75.00	.00	4.00	2.00	1.00	3.00	60.00	46.00	51.00	53.00	61.00
12	60.00	.00	4.00	2.00	1.00	2.00	57.00	65.00	51.00	63.00	61.00
13	[illegible]	[illegible]	4.00	3.00	1.00	2.00	73.00	60.00	71.00	61.00	71.00
14	[illegible]	[illegible]	4.00	3.00	1.00	2.00	54.00	63.00	57.00	55.00	46.00
15	38.00	.00	3.00	1.00	1.00	2.00	45.00	57.00	50.00	31.00	56.00
16	115.00	.00	4.00	1.00	1.00	1.00	42.00	49.00	43.00	50.00	56.00
17	76.00	.00	4.00	3.00	1.00	2.00	47.00	52.00	51.00	50.00	56.00
18	195.00	.00	4.00	2.00	2.00	1.00	57.00	57.00	60.00	58.00	56.00
19	114.00	.00	4.00	3.00	1.00	2.00	68.00	65.00	62.00	55.00	61.00
20	85.00	.00	4.00	2.00	1.00	1.00	55.00	39.00	57.00	53.00	46.00

Rows are subjects (observations).

Describing and Exploring Data

Distributions

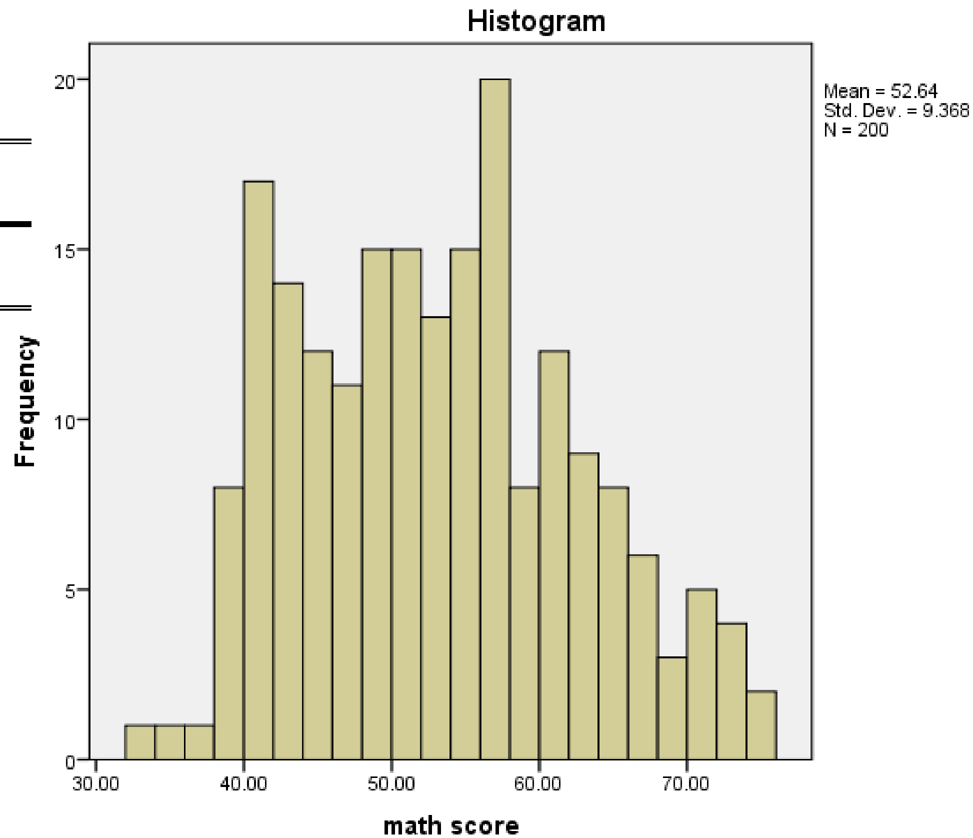
Continuous variables

- Descriptive Stats: Mean, Std. Dev., Median, Min, Max
- Histograms

Math

N	Mean	Std. Dev	Median	Min	Max
200	52.6	9.37	52.0	33.0	75.0

Note: Most analyses (ANOVA, Regression, etc) prefer a Normal (bell-shaped), symmetric distribution…



Categorical variables

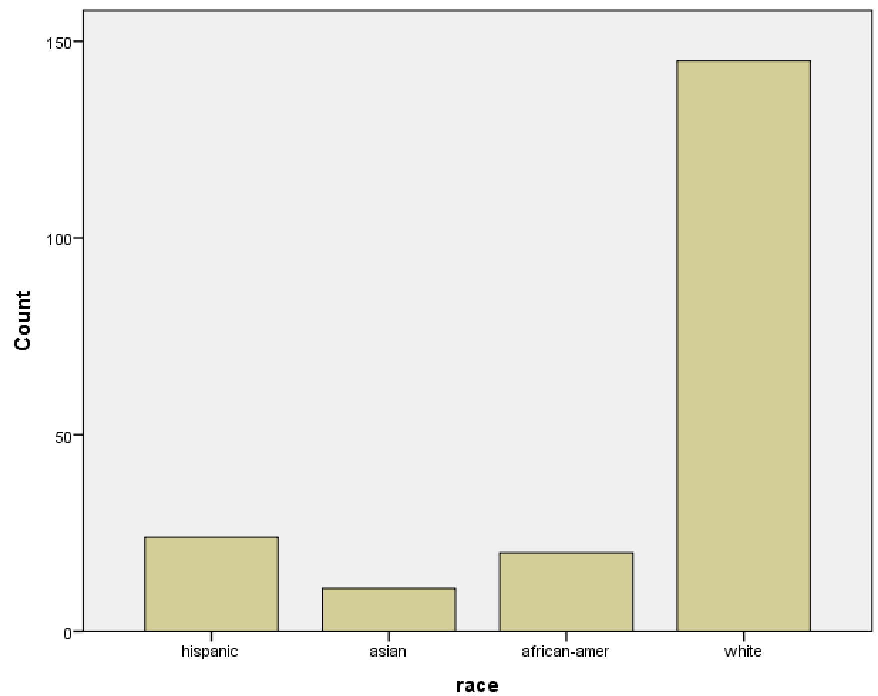
- Frequency tables: Counts and Percentages

female

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	male	91	45.5	45.5	45.5
	female	109	54.5	54.5	100.0
	Total	200	100.0	100.0	

race

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	hispanic	24	12.0	12.0	12.0
	asian	11	5.5	5.5	17.5
	african-amer	20	10.0	10.0	27.5
	white	145	72.5	72.5	100.0
	Total	200	100.0	100.0	



ses

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	low	47	23.5	23.5	23.5
	middle	95	47.5	47.5	71.0
	high	58	29.0	29.0	100.0
	Total	200	100.0	100.0	

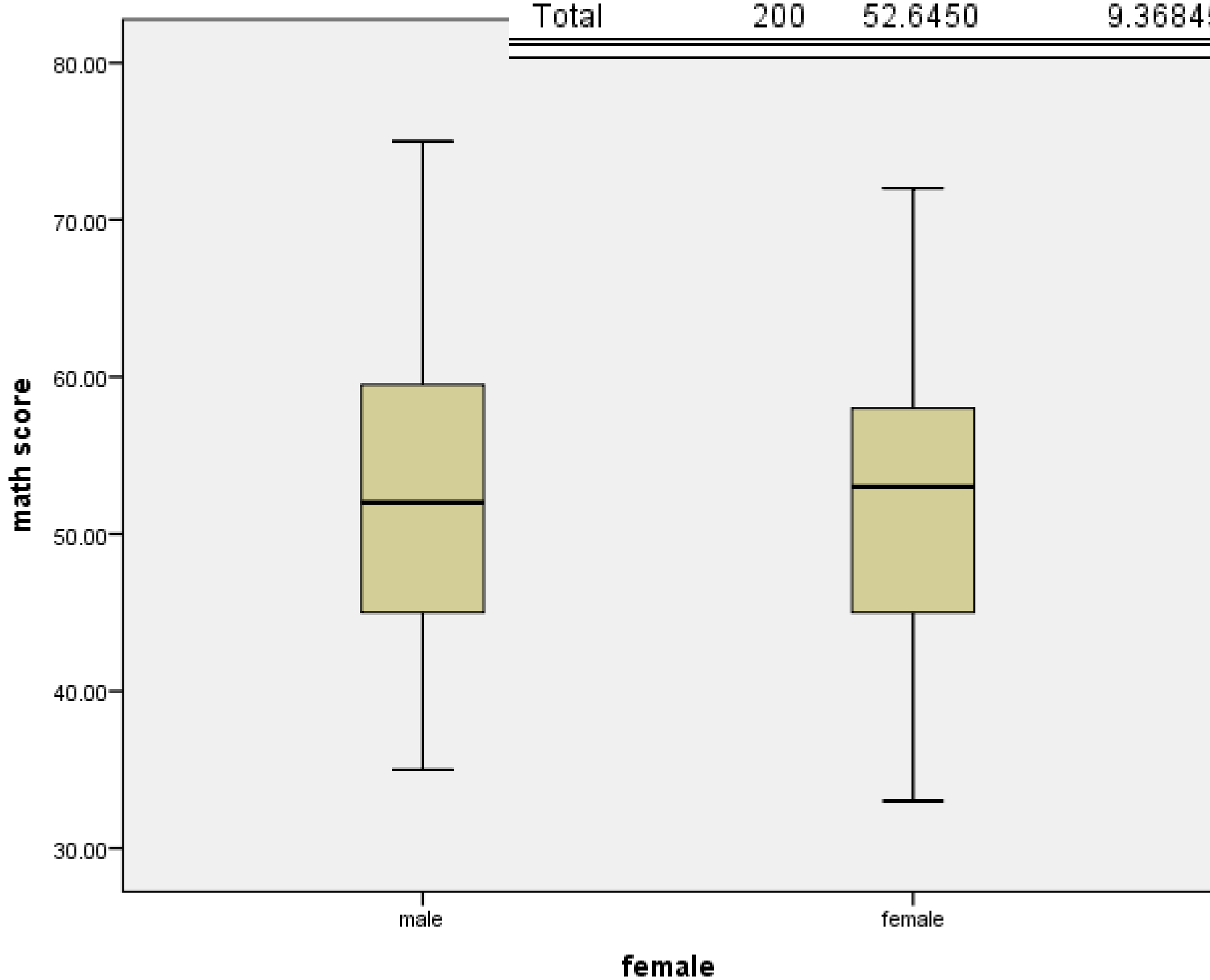
1 Continuous var w/ 1 Categorical var

Comparison of Means

Boxplots

math

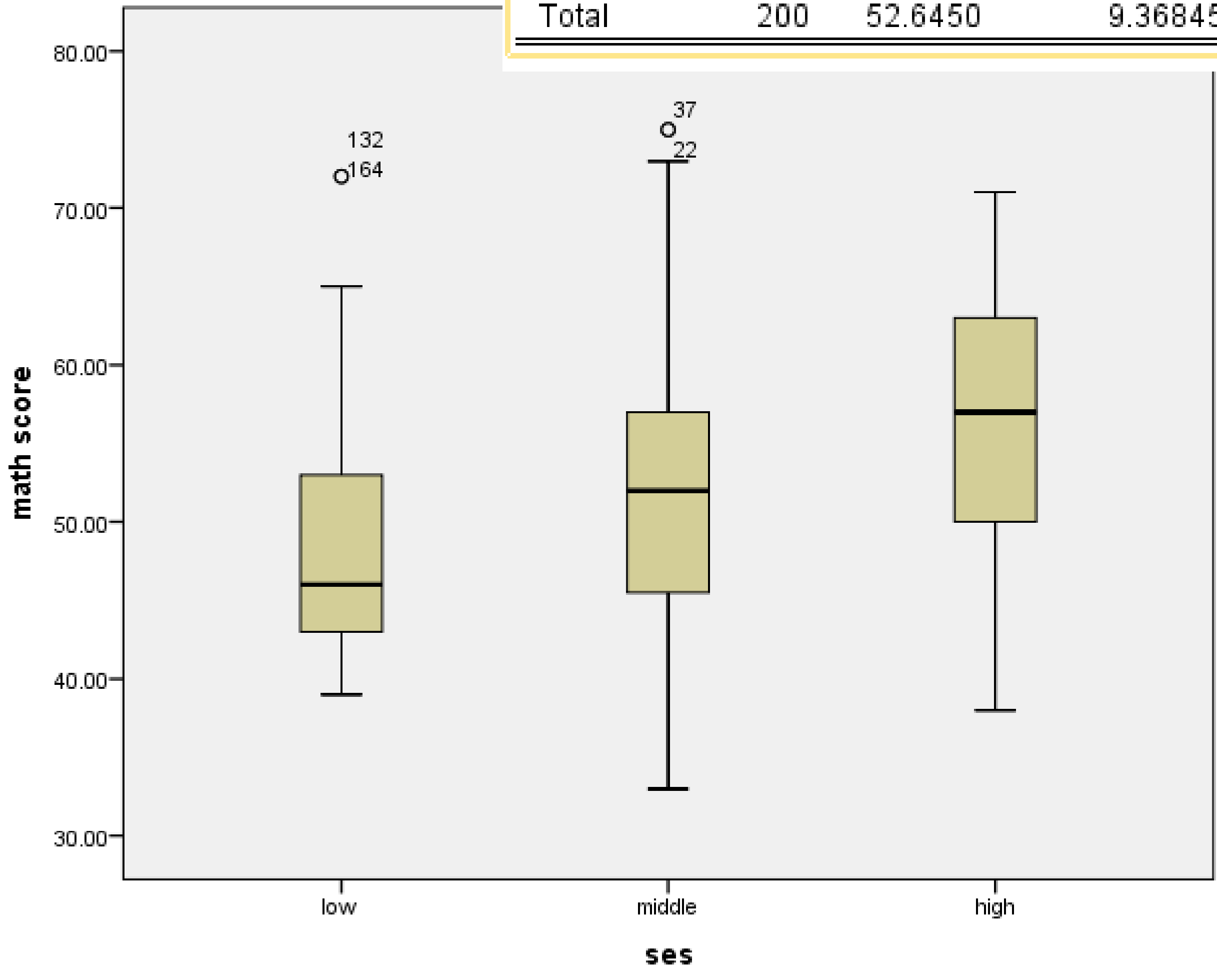
female	N	Mean	Std. Deviation	Median	Minimum	Maximum
male	91	52.9451	9.66478	52.0000	35.00	75.00
female	109	52.3945	9.15102	53.0000	33.00	72.00
Total	200	52.6450	9.36845	52.0000	33.00	75.00



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

Think about T-test...

math						
ses	N	Mean	Std. Deviation	Median	Minimum	Maximum
low	47	49.1702	8.87674	46.0000	39.00	72.00
middle	95	52.2105	9.35573	52.0000	33.00	75.00
high	58	56.1724	8.69188	57.0000	38.00	71.00
Total	200	52.6450	9.36845	52.0000	33.00	75.00



Think about ANOVA...

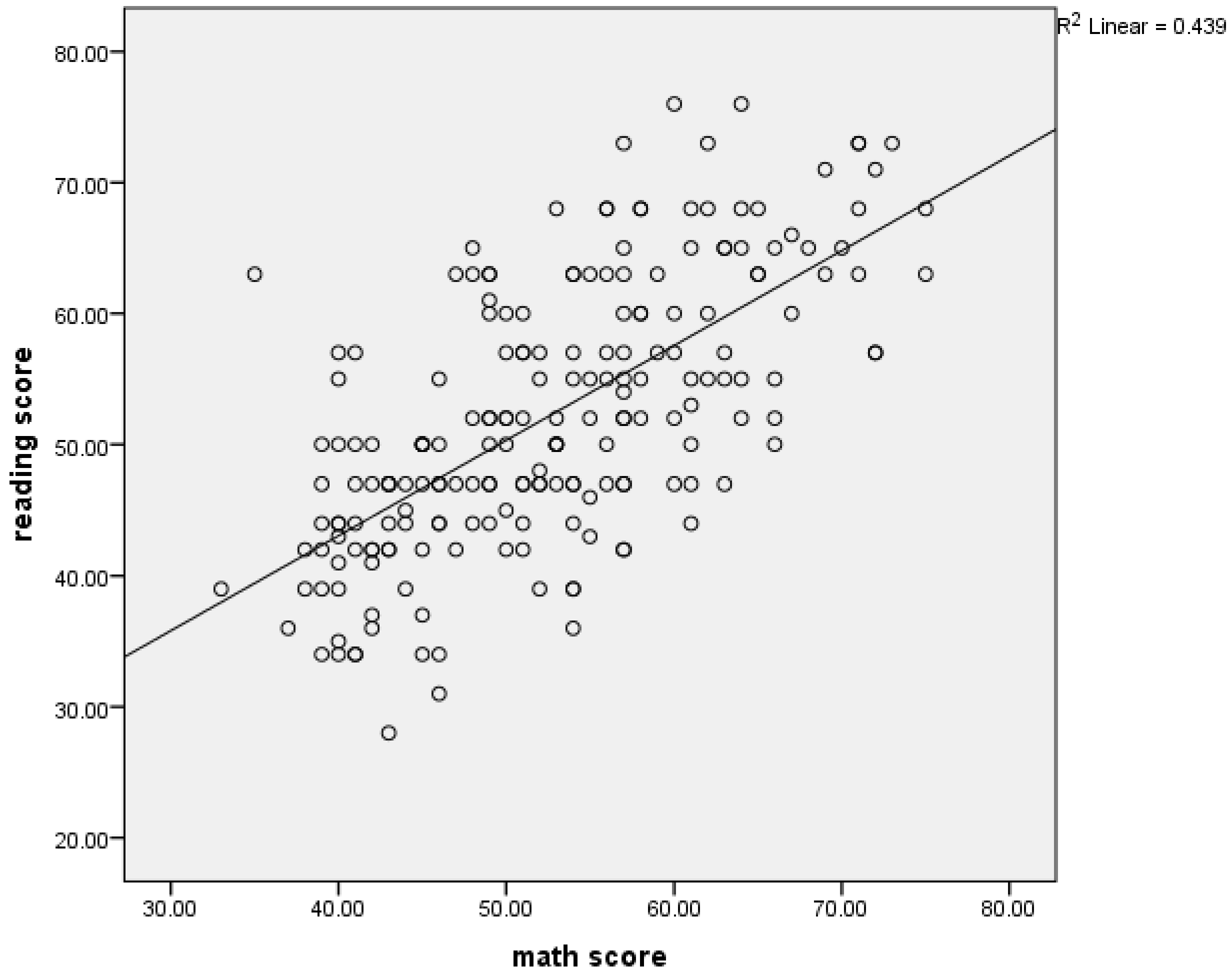
2 Continuous vars

Correlation

Scatterplot

Correlations

		read	math
read	Pearson Correlation	1	.662**
	Sig. (2-tailed)		.000
	N	200	200
math	Pearson Correlation	.662**	1
	Sig. (2-tailed)	.000	
	N	200	200



Think about Pearson Correlation…

2 Categorical vars

Comparison of Proportions

Crosstabs

ses * type of program Crosstabulation

			prog			
			general	academic	vocation	Total
ses	low	Count	16	19	12	47
		% within ses	34.0%	40.4%	25.5%	100.0%
	middle	Count	20	44	31	95
		% within ses	21.1%	46.3%	32.6%	100.0%
	high	Count	9	42	7	58
		% within ses	15.5%	72.4%	12.1%	100.0%
Total		Count	45	105	50	200
		% within ses	22.5%	52.5%	25.0%	100.0%

Think about Pearson Chi-square test…

Tools for Making Inferences

(about the population)

General Linear Models

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

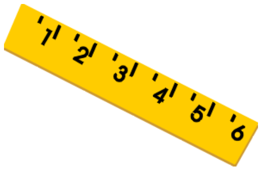
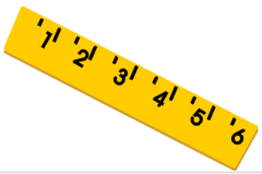
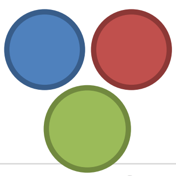
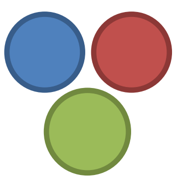
Dependent Variable (DV) = the *outcome* measure (Y)

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

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

- A framework for ANOVA, Regression, Logistic 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)?

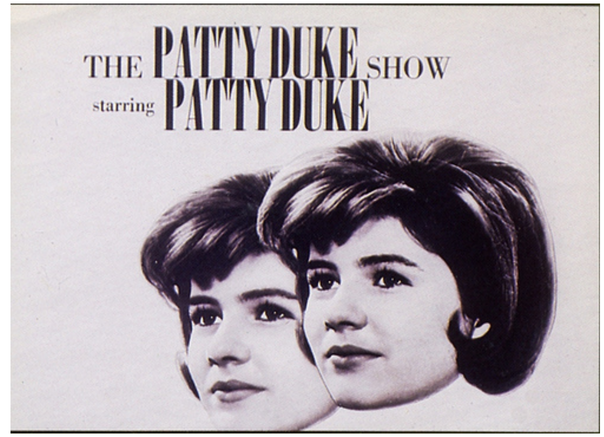
Which bucket of tools do you use for given materials?

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 Mixed Model
DV is Categorical 	Independent Observations	Crosstab, Pearson Chi-square. Logistic Regression. Poisson, Neg. Binomial.	Generalized Linear Model
DV is Categorical 	Correlated Data	Repeated Measures Logistic Regression. GEE, GLIMMIX	Generalized Linear Mixed Model

Independent Observations

DV is Continuous	IV is Categorical	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)	Comparison of Means
	IV is Continuous	Pearson Correlation (1 IV) Simple Linear Regression (1 IV) Multiple Linear Regression (>1 IV)	Correlations
	Any IV's	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	Pearson Chi-square (1 IV). Logistic Regression (>1 IV).	Comparison of Proportions
2 levels	Any IV's	Binary Logistic Regression	
>2 levels		Multinomial Logistic Regression	

ANOVA and Regression are 'identical cousins'



What's the difference?

- More about a matter of perspective - experimental design vs observational data
- Mathematically, the model for an ANCOVA (1 categorical IV with 1 continuous "covariate") is identical to a Regression with 1 categorical IV and 1 continuous IV.
- A "covariate" is just another Independent Variable.
- ANOVA often includes factorial interactions between IV "factors", but it's really up to you…
- Software for "ANOVA" and "Regression" usually have different defaults, but you can always get one from the other. And "GLM" does both!
- ANOVA often includes interactions, shows sums of squares with F tests
- Regression will not include interactions by default, shows beta parameter coefficients

Project Checklist

1. Formulate research question
2. Study design
   - Identify measures: IV's and DV's for hypothesis tests (and data types)
   - Consider data structure (independent or correlated data)
3. Analysis plan
4. Data collection (experiment, observation)
5. Data management
   - cleaning, possibly restructure (aggregate?) , eligibility
6. Perform descriptive statistics and univariate plots
   - Frequency tables for categorical variables
   - Means, SD, Min, Med, Max for continuous variables
   - Histograms, Box-plots for continuous variables, especially DV's
7. Consider re-codes (categorical) or transformations (continuous), or mix

8. Plot relationship of DV on IV's
   - Box plots for 1 continuous variable on 1 categorical variable
   - Scatterplots for 1 continuous variable on 1 continuous variable
9. Perform hypothesis tests, modeling
   - T-test, ANOVA, Regression, etc
10. Check assumptions & diagnostics
    - Normal residuals
    - Equal variance
    - Outliers/influential points
11. Interpret Results
    - Predicted values/means
    - Estimated Marginal Means (for categorical IV's)
      - Post Hoc tests
    - Beta "slope" coefficients (for continuous IV's)

Assumptions of T-test/ANOVA/Regression

~ Data are a representative sample of some larger population which I'm making inferences about.

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

~ Note that multicollinearity is not an assumption of the model in the same way …just something important to consider.

When Assumptions ~~are~~ *seem to be* violated

1. Not independent observations?
   - Maybe aggregate data to the individual level? (esp. binary data!)
   - Model the correlation structure in Repeated Measures ANOVA or Mixed Models
2. 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 more than twice as big as the other?
   - If sample sizes between groups are equal, ANOVA is robust
   - Log transformations of skewed data often help with variances
3. Not Normally distributed DV/residuals?
   - Be skeptical of 'tests' of normality (Shapiro-Wilks)…p-value is more significant with larger sample size, but…but larger sample sizes are more robust (Central Limit Theorem means ANOVA is Robust)
   - Skewness and Kurtosis are helpful (skewness <2?)
   - Try transformations, like taking the log, square root (or try Box-Cox)

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*

Also 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.

Start IBM SPSS

Open data by browsing to exercise.xls

(note 'Files of type' defaults to SPSS but can switch to Excel)

Select Sheet1 since our data is on the first Sheet in 'exercise.xls'

Descriptives

For descriptive stats and exploratory plots of our *continuous* DV (Systolic BP) (as seen above…)

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

Analyses

T-test (Independent Samples)

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

What's the difference in Systolic blood pressure between male/females?

IV: Gender (M/F)

DV: Systolic BP

Age	Gender	Max HR	SysBP	DiastBP	Race	Stress	Trt
13	Female	192	135	71	White	4	1
16	Female	194	146	78	Afr. Amer.	6	1
19	Female	199	165	78	Afr. Amer.	10	1
15	Female	177	165	60	White	10	1
11	Male	184	122	74	Afr. Amer.	1	1
13	Female	195	143	68	White	5	1
14	Female	197	163	62	Afr. Amer.	10	1
8	Female	150	144	70	White	6	1
16	Male	163	160	85	Afr. Amer.	9	1
12	Female	206	142	81	White	7	1
15	Female	192	143	68	White	7	1
11	Female	194	163	62	Afr. Amer.	10	1
13	Male	190	144	70	White	4	2
14	Female	177	160	85	Afr. Amer.	9	2
8	Female	172	142	81	White	3	2

T-test in SPSS.

Analyze > Compare Means > Independent Samples T-Test

Put DV (SysBP) in the 'Test Variable(s)'.

Put IV (Gender) in the Grouping Variable. Define Groups 'Male' and 'Female'.

Output:

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

There is not a significant difference between males (M=140.9, SD=19.2) and females (M=148.9, SD=14.7) in Systolic blood pressure (t(36)=-1.28, p=.208)

ANOVA

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

What's the difference in Systolic blood pressure between treatment groups?

IV: Treatment group (1, 2, 3)

DV: Systolic BP

Age	Gender	Max HR	SysBP	DiastBP	Race	Stress	Trt
13	Female	192	135	71	White	4	1
16	Female	194	146	78	Afr. Amer.	6	1
19	Female	199	165	78	Afr. Amer.	10	1
15	Female	177	165	60	White	10	1
11	Male	184	122	74	Afr. Amer.	1	1
13	Female	195	143	68	White	5	1
14	Female	197	163	62	Afr. Amer.	10	1
8	Female	150	144	70	White	6	1
16	Male	163	160	85	Afr. Amer.	9	1
12	Female	206	142	81	White	7	1
15	Female	192	143	68	White	7	1
11	Female	194	163	62	Afr. Amer.	10	1
13	Male	190	144	70	White	4	2
14	Female	177	160	85	Afr. Amer.	9	2
8	Female	172	142	81	White	3	2

One-way ANOVA in SPSS.

Analyze > Compare Means > One-way ANOVA

Put DV (SysBP) as 'Dependent', and Trt as 'Factor'

'Post Hoc' > Tukey

Options > Descriptives and Homogeneity of variance.

Output:

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

There is not a significant difference between the three treatment groups in Systolic blood pressure (F(2,35)=.260, p=.773)

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) but too conservative/stingy. Tukey is more generous.

ANOVA, continued

for more than 1 Categorical IV's

What's the difference in Systolic blood pressure between treatment groups, and does it depend on gender?

IV: Treatment group (1, 2, 3); Gender (M/F)

DV: Systolic BP

Age	Gender	Max HR	SysBP	DiastBP	Race	Stress	Trt
13	Female	192	135	71	White	4	1
16	Female	194	146	78	Afr. Amer.	6	1
19	Female	199	165	78	Afr. Amer.	10	1
15	Female	177	165	60	White	10	1
11	Male	184	122	74	Afr. Amer.	1	1
13	Female	195	143	68	White	5	1
14	Female	197	163	62	Afr. Amer.	10	1
8	Female	150	144	70	White	6	1
16	Male	163	160	85	Afr. Amer.	9	1
12	Female	206	142	81	White	7	1
15	Female	192	143	68	White	7	1
11	Female	194	163	62	Afr. Amer.	10	1
13	Male	190	144	70	White	4	2
14	Female	177	160	85	Afr. Amer.	9	2
8	Female	172	142	81	White	3	2

ANOVA in SPSS, continued.

Analyze > General Linear Model > Univariate

Put DV (SysBP) as ‘Dependent’.

Put Gender and Trt as ‘Fixed Factors’.

Click ‘Model’ to specify interactions.
(“ANOVA” usually thinks about interactions…)

Post Hoc > Trt > Tukey

(Optional) Save > Standardized Residuals

Options > Display Means for: <Gender Trt Gender*Trt>
(Note: Compare main effects can do Bonferroni, 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

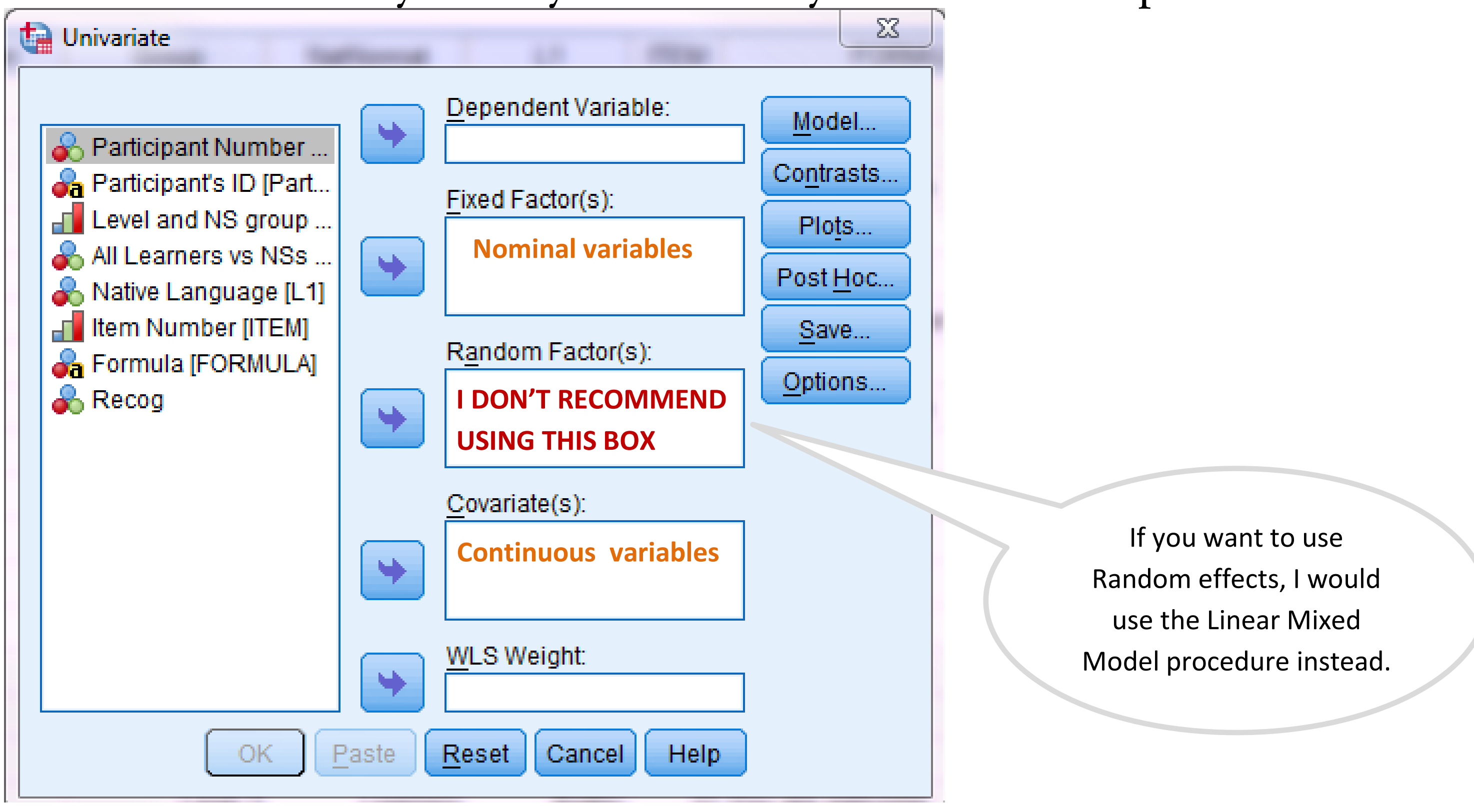
SPSS note!

‘Fixed Factors’ are for Categorical variables.

‘Covariates’ are for Continuous variables.

“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…



ANCOVA

Compare Continuous DV between groups (Categorical IV), adjusting for Continuous "covariate" IV

What's the difference in Systolic blood pressure between treatment groups and gender, controlling for age?

IV: Treatment group (1, 2, 3); Gender (M/F), Age

DV: Systolic BP

Age	Gender	Max HR	SysBP	DiastBP	Race	Stress	Trt
13	Female	192	135	71	White	4	1
16	Female	194	146	78	Afr. Amer.	6	1
19	Female	199	165	78	Afr. Amer.	10	1
15	Female	177	165	60	White	10	1
11	Male	184	122	74	Afr. Amer.	1	1
13	Female	195	143	68	White	5	1
14	Female	197	163	62	Afr. Amer.	10	1
8	Female	150	144	70	White	6	1
16	Male	163	160	85	Afr. Amer.	9	1
12	Female	206	142	81	White	7	1
15	Female	192	143	68	White	7	1
11	Female	194	163	62	Afr. Amer.	10	1
13	Male	190	144	70	White	4	2
14	Female	177	160	85	Afr. Amer.	9	2
8	Female	172	142	81	White	3	2

ANCOVA in SPSS.

Analyze > General Linear Model > Univariate

Put DV (SysBP) as 'Dependent'.

Put Gender and Trt 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

MANCOVA (or MANOVA)

Compare more than 1 Continuous (related) DV between groups (Categorical IV), adjusting for Continuous "covariate" IV

What's the difference in Systolic blood pressure and Diastolic blood pressure between treatment groups and genders, controlling for age?

IV: Treatment group (1, 2, 3); Gender (M/F), Age

DV: Systolic BP and Diastolic BP

Age	Gender	MaxHR	SysBP	DiastBP	Race	Stress	Trt
13	Female	192	135	71	White	4	1
16	Female	194	146	78	Afr. Amer.	6	1
19	Female	199	165	78	Afr. Amer.	10	1
15	Female	177	165	60	White	10	1
11	Male	184	122	74	Afr. Amer.	1	1
13	Female	195	143	68	White	5	1
14	Female	197	163	62	Afr. Amer.	10	1
8	Female	150	144	70	White	6	1
16	Male	163	160	85	Afr. Amer.	9	1
12	Female	206	142	81	White	7	1
15	Female	192	143	68	White	7	1
11	Female	194	163	62	Afr. Amer.	10	1
13	Male	190	144	70	White	4	2
14	Female	177	160	85	Afr. Amer.	9	2

MANOVA/MANCOVA in SPSS.

Analyze > General Linear Model > Multivariate

Put SysBP and DiastBP as 'Dependent'.

Put Gender and Trt 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

What's the difference in Systolic blood pressure between time point 1 and 2?

DV: Systolic BP at timepoint 1 (SBP1) and timepoint 2 (SBP2)

IV: Time (1, 2)

Age	Gender	Max HR	SysBP	DiastBP	Race	Stress	Trt	SBP1	SBP2
13	Female	192	135	71	White	4	1	133	134
16	Female	194	146	78	Afr. Amer.	6	1	143	140
19	Female	199	165	78	Afr. Amer.	10	1	166	164
15	Female	177	165	60	White	10	1	166	162
11	Male	184	122	74	Afr. Amer.	1	1	123	120
13	Female	195	143	68	White	5	1	140	136
14	Female	197	163	62	Afr. Amer.	10	1	164	163
8	Female	150	144	70	White	6	1	143	142
16	Male	163	160	85	Afr. Amer.	9	1	160	158
12	Female	206	142	81	White	7	1	141	139
15	Female	192	143	68	White	7	1	145	145
11	Female	194	163	62	Afr. Amer.	10	1	163	161
13	Male	190	144	70	White	4	2	143	144
14	Female	177	160	85	Afr. Amer.	9	2	160	157
8	Female	172	142	81	White	3	2	144	146

Paired t-test in SPSS.

Analyze > Compare Means > Paired samples t-test

Put SBP1 and SBP2 into 'paired variables' (systolic blood pressure at time 1 and time 2)

Output:

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

There is a significant decrease in systolic blood pressure from time 1 (M=147) to time 2 (M=146) (t(37)=5.66, p<.001)

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

Repeated Measures ANOVA

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

What's the difference in Systolic blood pressure between time point 1 and 2, for each treatment? Did the 3 trt groups change similarly or differently?

DV: Systolic BP at timepoint 1 (SBP1) and timepoint 2 (SBP2)

IV: *Time (1, 2)*; Trt (1,2,3)

Note that 'Time' is not actually a variable in the dataset. We have to specify that below.

Age	Gender	Max HR	SysBP	DiastBP	Race	Stress	Trt	SBP1	SBP2
13	Female	192	135	71	White	4	1	133	134
16	Female	194	146	78	Afr. Amer.	6	1	143	140
19	Female	199	165	78	Afr. Amer.	10	1	166	164
15	Female	177	165	60	White	10	1	166	162
11	Male	184	122	74	Afr. Amer.	1	1	123	120
13	Female	195	143	68	White	5	1	140	136
14	Female	197	163	62	Afr. Amer.	10	1	164	163
8	Female	150	144	70	White	6	1	143	142
16	Male	163	160	85	Afr. Amer.	9	1	160	158
12	Female	206	142	81	White	7	1	141	139
15	Female	192	143	68	White	7	1	145	145
11	Female	194	163	62	Afr. Amer.	10	1	163	161
13	Male	190	144	70	White	4	2	143	144
14	Female	177	160	85	Afr. Amer.	9	2	160	157
8	Female	172	142	81	White	3	2	144	146

Repeated Measures in SPSS.

Analyze > General Linear Model > Repeated Measures

Put 'time' as the 'within-subject' factor, with 2 levels

(optional) Name the 'measure' SBP

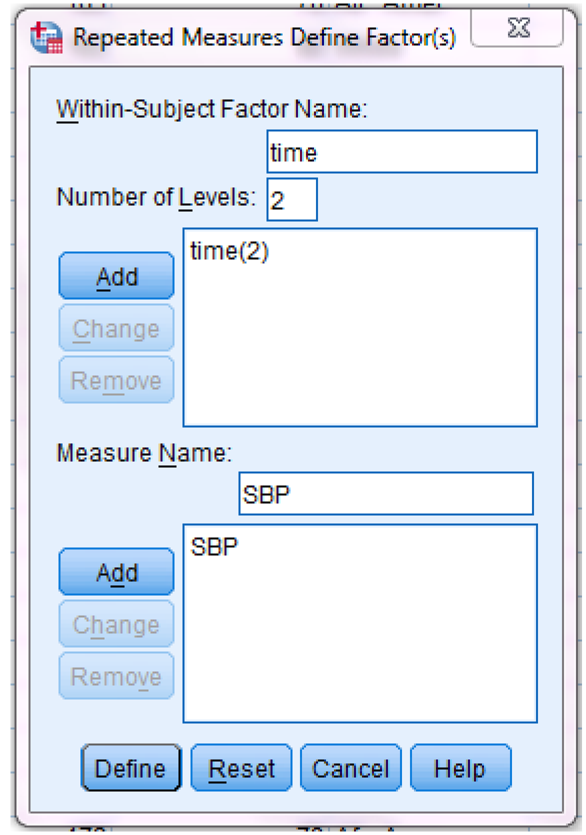
Enter SBP1 and SBP2 as the 'within-subject' variables

Enter Trt as a 'between-subject' factor

Consider 'Model' or 'Plots'

'Post-Hoc' for Trt, with Tukey

'Options' > Display Means for <time, trt, time*trt>



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.

Also see…

Repeated Measures ANOVA for Diet Study.

Using ‘Dietstudy.sav’

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).

Tutorial “demo” slideshow for SPSS commands in ‘Diet Study RM ANOVA_demo.pdf’

Demo:
http://www.indiana.edu/~iscc/files/WIM_Fall_2012/Diet%20Study%20RM%20ANOVA_demo.pdf

Data: http://www.indiana.edu/~iscc/files/WIM_Fall_2012/dietstudy.sav

Correlation

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

What's the correlation (or association or relationship) between age and SBP?

DV: Systolic Blood Pressure

IV: Age

Age	Gender	Max HR	SysBP	DiastBP	Race	Stress	Trt	SBP1	SBP2
13	Female	192	135	71	White	4	1	133	134
16	Female	194	146	78	Afr. Amer.	6	1	143	140
19	Female	199	165	78	Afr. Amer.	10	1	166	164
15	Female	177	165	60	White	10	1	166	162
11	Male	184	122	74	Afr. Amer.	1	1	123	120
13	Female	195	143	68	White	5	1	140	136
14	Female	197	163	62	Afr. Amer.	10	1	164	163
8	Female	150	144	70	White	6	1	143	142
16	Male	163	160	85	Afr. Amer.	9	1	160	158
12	Female	206	142	81	White	7	1	141	139
15	Female	192	143	68	White	7	1	145	145
11	Female	194	163	62	Afr. Amer.	10	1	163	161
13	Male	190	144	70	White	4	2	143	144
14	Female	177	160	85	Afr. Amer.	9	2	160	157
8	Female	172	142	81	White	3	2	144	146

Correlation in SPSS.

Analyze > Correlate > Bivariate

Enter Age and SysBP.

(Note Pearson and/or Spearman (Non-parametric test))

Output:

- Pearson r correlation value
- Corresponding p-value
- Sample size (N)

Correlations

		Age	SysBP
Age	Pearson Correlation	1	.347*
	Sig. (2-tailed)		.033
	N	38	38
SysBP	Pearson Correlation	.347*	1
	Sig. (2-tailed)	.033	
	N	38	38

*. Correlation is significant at the 0.05 level (2-tailed).

Note that this correlation 'r' is the square root of the R-squared from a ***simple linear regression*** of SysBP on age…

(Multiple) Linear Regression

Test for effect of 1 (or more) Continuous (or Categorical) IV on 1 Continuous DV

What's the effect of age and gender on Systolic blood pressure?

IV: Age; Gender (M/F)

DV: Systolic BP

Age	Gender	Max HR	SysBP	DiastBP	Race	Stress	Trt
13	Female	192	135	71	White	4	1
16	Female	194	146	78	Afr. Amer.	6	1
19	Female	199	165	78	Afr. Amer.	10	1
15	Female	177	165	60	White	10	1
11	Male	184	122	74	Afr. Amer.	1	1
13	Female	195	143	68	White	5	1
14	Female	197	163	62	Afr. Amer.	10	1
8	Female	150	144	70	White	6	1
16	Male	163	160	85	Afr. Amer.	9	1
12	Female	206	142	81	White	7	1
15	Female	192	143	68	White	7	1
11	Female	194	163	62	Afr. Amer.	10	1
13	Male	190	144	70	White	4	2
14	Female	177	160	85	Afr. Amer.	9	2
8	Female	172	142	81	White	3	2

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's)

Test for relationship between 1 Categorical IV and 1 Categorical DV, also a comparison of proportions

What's the difference between males and females in the proportion (or percent) with hypertension?

DV: Hypertension (0=No, 1=Yes)

IV: Gender (Male, Female)

Age	Gender	Max HR	SysBP	DiastBP	Race	Stress	Trt	SBP1	SBP2	Hypertension
13	Female	192	135	71	White	4	1	133	134	0
16	Female	194	146	78	Afr. Amer.	6	1	143	140	1
19	Female	199	165	78	Afr. Amer.	10	1	166	164	1
15	Female	177	165	60	White	10	1	166	162	1
11	Male	184	122	74	Afr. Amer.	1	1	123	120	0
13	Female	195	143	68	White	5	1	140	136	1
14	Female	197	163	62	Afr. Amer.	10	1	164	163	1
8	Female	150	144	70	White	6	1	143	142	1
16	Male	163	160	85	Afr. Amer.	9	1	160	158	1
12	Female	206	142	81	White	7	1	141	139	1
15	Female	192	143	68	White	7	1	145	145	1
11	Female	194	163	62	Afr. Amer.	10	1	163	161	1
13	Male	190	144	70	White	4	2	143	144	1
14	Female	177	160	85	Afr. Amer.	9	2	160	157	1
8	Female	172	142	81	White	3	2	144	146	1

Pearson Chi-square test in SPSS.

Analyze > Descriptive Stats > Crosstabs

Enter Gender as 'Rows' and Hypertension as 'Columns' (or vice-versa)

Statistics > Chi-square

Cells > Percentages > Rows (or Columns)

If Hypertension is DV (in Columns), and Gender is IV (in Rows), then we want 'Row' percentages to get the % with hypertension within each Gender

Output:

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

Gender * Hypertension Crosstabulation

			Hypertension		Total
			0	1	
Gender	Female	Count	5	25	30
		% within Gender	16.7%	83.3%	100.0%
	Male	Count	4	4	8
		% within Gender	50.0%	50.0%	100.0%
Total		Count	9	29	38
		% within Gender	23.7%	76.3%	100.0%

There is a significant difference between genders in the percent of hypertension (Chi-square(1)=3.883, p=.049).

(Multiple) Logistic Regression

Test for effect of any IV's on 1 Categorical DV with 2 or more levels.

What are the predictors of hypertension? Age, Gender, Race?

DV: Hypertension (0=No, 1=Yes)

IV: Gender (M, F); Race (White, Af.Am); Age

Age	Gender	Max HR	SysBP	DiastBP	Race	Stress	Trt	SBP1	SBP2	Hypertension
13	Female	192	135	71	White	4	1	133	134	0
16	Female	194	146	78	Afr. Amer.	6	1	143	140	1
19	Female	199	165	78	Afr. Amer.	10	1	166	164	1
15	Female	177	165	60	White	10	1	166	162	1
11	Male	184	122	74	Afr. Amer.	1	1	123	120	0
13	Female	195	143	68	White	5	1	140	136	1
14	Female	197	163	62	Afr. Amer.	10	1	164	163	1
8	Female	150	144	70	White	6	1	143	142	1
16	Male	163	160	85	Afr. Amer.	9	1	160	158	1
12	Female	206	142	81	White	7	1	141	139	1
15	Female	192	143	68	White	7	1	145	145	1
11	Female	194	163	62	Afr. Amer.	10	1	163	161	1
13	Male	190	144	70	White	4	2	143	144	1
14	Female	177	160	85	Afr. Amer.	9	2	160	157	1
8	Female	172	142	81	White	3	2	144	146	1

Logistic Regression in SPSS.

Analyze > Regression > Binary Logistic

Enter Hypertension as 'Dependent'.

Enter Age, Gender, and Race as 'Covariates'

If you have numeric categorical variables, click 'Categorical' to specify.

Note: Method > 'Enter' or 'Forward: LR'

For model selection, my favorite is 'Forward: LR'. LR=Likelihood Ratio tests are a bit better than Wald.

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 Models

- 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

Example…

Using 'Dietstudy_long.sav'

Diet data is restructured into "long" form with multiple rows for each subject.

Tutorial "demo" slideshow for SPSS commands in 'Diet Study LMM_demo.pdf'

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)

Demo: http://www.indiana.edu/~iscc/files/WIM_Fall_2012/Diet%20Study%20LMM_demo.pdf

Data: http://www.indiana.edu/~iscc/files/WIM_Fall_2012/dietstudy_long.sav

More info from a previous workshop:

http://www.indiana.edu/~iscc/files/WIM_Fall_2012/GLM workshop slides Part 2_ 2011-10-03.pdf

Ucla stat computing:

http://www.ats.ucla.edu/stat/spss/library/spssmixed/mixed.htm