MHPE 494: Data Analysis

Alan Schwartz, PhD
Matt Lineberry, PhD
Department of Medical Education

College of Medicine
University of Illinois at Chicago

Welcome!

- Your name, specialty, institution, position
- Experience in data analysis
- Why this class?
  What are your expectations and goals?

The Analytic Process

- Formulate research questions
- Design study
- Collect data
- Record data
- Check data for problems
- Explore data for patterns
- Test hypotheses with the data
- Interpret and report results

Covered in Research Design/Grant Writing

Covered in Writing for Scientific Publication

Monday AM

- Introduction
- Syllabus
- Data Entry
- Data Checking
- Exploratory Data Analysis

Data entry

or,

"Garbage in, garbage out"

Data Entry

- Data entry is the process of recording the behavior of research subjects (or other data) in a format that is efficient for:
  - Understanding the coded responses
  - Exploring patterns in the data
  - Conducting statistical analyses
  - Distributing your data set to others
- Data entry is often given low regard, but a little time spent now can save a lot of time later!

Methods of data entry

- Direct entry by participants
- Direct entry from observations
- Entry via coding sheets

- Entry to statistical software
- Entry to spreadsheet software
- Entry to database software

Data file layout

- Most data files in most statistical software use “standard data layout”:
  - Each row represents one subject
  - Each column represents one variable measurement
- Special formats are sometimes used for particular analyses/software
  - Doubly multivariate data (each row is a subject at a given time)
  - Matrix data

“Standard data layout”

Id	Female	YrsOld	GPA
1	1	19	3.5
2	0	21	3.4
3	1	20	3.4

Missing data

- Data can be missing for many reasons:
  - Random missing responses
  - Drop-out in longitudinal studies (censoring)
  - Systematic failure to respond
  - Structure of research design
- Knowing why data is missing is often the key to deciding how to handle missing data

Missing data

- Approaches to dealing with missing data:
  - Leave data missing, and exclude that cell or subject from analyses
  - Impute values for missing data (requires a model of how data is missing)
  - Use an analytic technique that incorporates missing data as part of data structure

Naming Variables

- Variables should have both a short name (for the software) and a descriptive name (for reporting)
- Name for what is measured, not inferred
- Short names should capture something useful about the variable (its scale, its coding)
- Better names:
  - Q1-Q20, IQ, MALE, IN_TALL, IN_TALLZ
- Worse names:
  - INTEL, SEX, SIZE

Coding Variables

- Depends on *measurement scale*
  - Nominal, two categories: Name variable for one category and code 1 or 0
  - Nominal, many categories: Use a string coding or meaningful numbers
  - Ordinal: Code ranks as numbers, decide if lower or higher ranks are better
  - Interval/Ratio: Code exact value

Labeling Variable Values

- For nominal and ordinal variables, *values* should also be labeled unless using string coding.
- Value labels should precise indicate the response to which the value refers.
  - Example: Educational level ordinal variable:

    1 = grade school not completed
    2 = grade school completed
    3 = middle school completed
    4 = high school completed
    5 = some college
    6 = college degree

Error Checking

- Goal: Identify errors made due to:
  - Faulty data entry
  - Faulty measurement
  - Faulty responses
- Prior to analyses. Not hypothesis-based

Range checking

- The first basic check that should be performed on all variables
- Print out the range (lowest and highest value) of every variable
- Quickly catches common typos involving extra keystrokes

Distribution checking

- Examining the distribution of variables to insure that they'll be amenable to analysis.
- Problems to detect include:
  - Floor and ceiling effects
  - Lack of variance
  - Non-normality (including skew and kurtosis)
  - Heteroscedascity (in joint distributions)

Eccentric subjects

- Patterns of data can suggest that particular subjects are eccentric
  - Subjects may have misunderstood instructions
  - Subjects may understand instructions but use response scale incorrectly
  - Subjects may intentionally misreport (to protect themselves or to subvert the study as they see it)
  - Subjects may actually have different, but coherent views!

Verbal protocols

- Verbal protocols (written or otherwise recorded) can help to distinguish subjects who don't understand from subjects who understand, but feel differently than most others.
  - "What was going through your head while you were doing this?"
  - "How did you decide to response that way?"
  - "Do you have any comments about this study?"
- Debriefing interviews can be used similarly

Holding subjects out

- If a subject is indeed eccentric, you must decide whether or not to hold the subject out of the analysis. Document these choices.
  - Pros: Data will be cleaner (sample will be more homogenous, less noisy)
  - Cons: Ability to generalize is reduced, bias may be introduced
- If a group of subjects are eccentric in the same way, it's probably better to analyze them as a subgroup, or use individual-level techniques.

Cleaning data

- When only a few data points are eccentric, a case can sometimes be made for *cleaning* the data.
  - Example: Subjects were asked to respond on a computer keyboard to money won or lost in a game on a scale from -50 (very unhappy) to 50 (very happy). One subject's ratings were:
  - +$5 = "10", -$5 = "-3", -$20 = "-40", -$10 = "20"
  - Should the "20" response be changed to "-20"?
- Document these choices.

Four great SPSS commands

- *Transform…Compute*: create a new variable computed from other variables.
- *Transform…Recode*: create a new variable by recoding the values of an existing variable.
- *Data…Select cases*: choose cases on which to perform analyses, setting others aside.
- *Data…Split file*: choose variables that define groups of cases, and run following analyses individually for each group.

Exploratory Data Analysis

- The goal of EDA is to apprehend patterns in data
- The better you understand your data set, the easier later analyses will be.
- EDA is *not:*
  - "data-mining" (an atheoretical look for any significant findings in the data, capitalizing on chance)
  - hypothesis testing (though it may help with this)
  - data presentation (though it **does** help with this)

EDA Tools: Stem-and-leaf plots

```
Starting Salary Stem-and-Leaf Plot
 Frequency    Stem &  Leaf
     1.00        0 .  &
     9.00        0 .  889
     9.00        1 .  001
    22.00        1 .  2222333
    20.00        1 .  4555555
    39.00        1 .  6666777777777
    57.00        1 .  8888888888999999999
   139.00        2 .  00000000000000000000000000000011111111111111111
   118.00        2 .  2222222222222222233333333333333333333333
   126.00        2 .  444444444444444444455555555555555555555555
   132.00        2 .  66666666666666666666666677777777777777777777
    98.00        2 .  88888888888888889999999999999999
   113.00        3 .  0000000000000000000000011111111111111
    94.00        3 .  2222222222222222233333333333333
    55.00        3 .  444444444555555555
    21.00        3 .  6666677
    15.00        3 .  88889
    11.00        4 .  0001
     6.00        4 .  23
     2.00        4 .  4
    13.00 Extremes    (>=45000)
 Stem width:  10000 Each leaf: 3 case(s) & denotes fractional leaves.
```

EDA Tools: Central Tendency

- Measures of central tendency: what one number best summarizes this distribution?
- Most common are mean, median, and mode
- Others include trimmed means, etc.
- Example:

Starting salary (N=1100)	
Mean	26064.20
Median	26000.00
Mode	20000

EDA Tools: Variability

- Measures of variability: how much and in what way do the data vary around their center?
- Most common: standard deviation, variance (sd squared), skew, kurtosis

Starting salary (N=1100)	
Mean	26064.20
Std. Deviation	6967.98
Variance	48552771.77
Skewness	.488
Std. Error of Skewness	.074
Kurtosis	1.778
Std. Error of Kurtosis	.147

EDA Tools: Norms and percentiles

- Percentiles are pieces of the frequency distribution: for what score are x% of the scores below that score. They can be used to set norms.

Starting salary (N=1100)		
Percentiles		
	5	15000.00
	25	21000.00
	50	26000.00
	75	30375.00
	95	36595.00

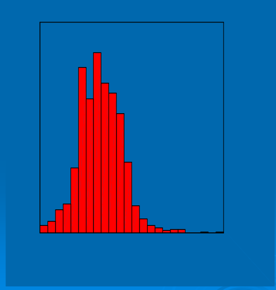
EDA Tools: Graphing

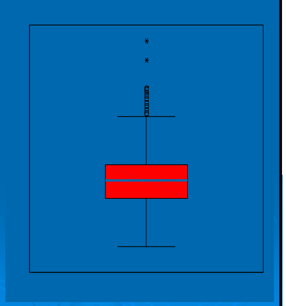
- Graphing puts the inherent power of visual perception to work in finding patterns in data
- Choice of graph depends on:
  - Number of dependent and independent variables
  - Measurement scale of variables
  - Goal of visualization (compare groups? seek relationships? identify outliers?)

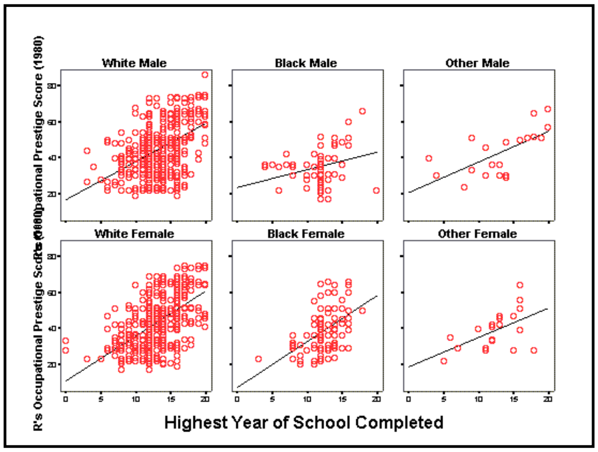
Types of graphs

- One variable: Frequency histogram, stem-and-leaf
- Two variables (independent x dependent):
  - nominal x interval: bar chart
  - interval x nominal: histogram
  - interval x interval: scatter plot
- Three variables (ind x ind x dep):
  - nominal x nominal x interval: 3d or clustered bar chart
  - nominal x interval x interval: line chart
  - interval x interval x interval: 3d scatter plot
- Four variables (ind x ind x ind x dep): matrix

Examples

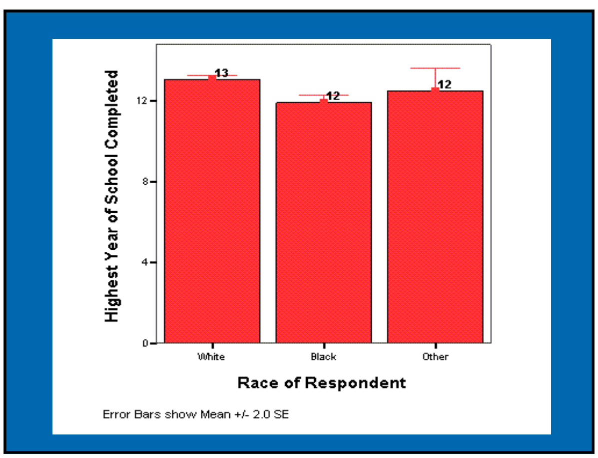






Error bars

- Most graphs provide measures of central tendency or aggregate response
- Error bars are a natural way to indicate variability as well. Some common choices to show:
  - 1 standard deviation (when describing populations)
  - 1 standard error of the mean
  - 95% confidence interval
  - 2 standard errors of the mean



SD & SE: commonly confused

- Standard deviation: How do individual scores cluster around the mean score?
  - There's some true variation in the world, and we're getting an estimate of it.
- Standard error of the mean: If we repeatedly estimate the mean, how will those estimates cluster around the true mean?
  - There's one true mean in the world, but each estimate we make has some noise
  - Larger sample size → less noise, smaller SE

SE & CI: Inherently related

- Standard error of the ____: If we repeatedly estimate the ____, how will those estimates cluster around the true ___?
- X% confidence interval around the ____: If we repeatedly estimate the ____, what interval around the estimate would be wide enough to ensure that X% of those interval estimates would contain the ____?
- (For normal distributions, the 95% CI around the mean is ± 1.96 SE)

Assignment

- Explore the hyp data, and describe the distribution of each of the variables.

Monday PM

- Presentation of AM results
- Hypothesis testing review
- Testing means
  - One sample t-test
  - Two sample t-test
  - One way ANOVA
  - Paired sample t-test

Hypothesis testing: a review

- In hypothesis testing statistics, we set up a null hypothesis about the data, and then proceed to try to reject this hypothesis.
- The null hypothesis usually represents “no effect”, “no difference”, or “no relationship”, though it may represent other possibilities as well.

Errors in hypothesis testing

	H_o is false (effect)	H_0 is true (no effect)
Reject H_0	Sig. effect (TP) 1-β (power)	Type I error (FP) α
Fail to reject H_0	Type II error (FN) β	No effect (TN) 1-α (confidence)

One- and two-tailed tests

- Tests can be one-tailed or two-tailed
  - A two tailed test looks for any difference from the null hypothesis, no matter what direction.
  - A one tailed test looks for a specified *directional* difference from the null hypothesis, and does not test for differences in the other direction.
- One-tailed tests are more powerful for a given α, but two-tailed tests can find effects in either direction.

Degrees of freedom

- Inferential (sample) statistics essentially involve fitting a model of the null hypothesis to the data, and finding that the model is a poor fit.
- The more data you have and the simpler the model, the less constraint there is upon how the data can be distributed, and the more ways the data might not be fitted.
- Formally, the number of unconstrained data points that the model is free to fit or not are the statistic's *degrees of freedom*, or *df*.

Degrees of freedom, example

- A line is a model defined by two parameters: a slope and an intercept.
- If I give you two points, you can always fit a perfect line. There are no degrees of freedom left to determine if the line fits well or not.
- If I give you three points, you need two of them to fit a line, and one is left to test whether a line is a good fit to the data. You have 1 df. This is a weak test.
- If I give you 100 points, you need 2 to fit the line, and you have 98 df. This is a powerful test.

One-sample t-test

- Goal: Given a sample, test to see if it comes from a population with a given mean value of a variable
- Example: Is the mean GPA of medical students different from 3.0?
- H_0: $\mu = k$
- H_1 $\mu \neq k$ (2-tailed) or $\mu > k$ (1-tailed)

One-sample t-test in SPSS

- Analyze…Compare Means…One-sample t-test
- Enter variable and test value

```
                                           N  Mean    SD      SE mean
Highest Year of School Completed      1510    12.88   2.98
         7.68E-02

One-Sample Test
Test Value = 8
         t      df      Sig.   Mean Diff    95% CI of diff

                                         Lower      Upper
HYSC    63.602  1509    .000     4.88       4.73      5.03
```

- On average, Americans sampled had more than 8 years of education ($t(1509) = 63.6$, $p < .05$).

Two-sample t-test

- Goal: Given 2 samples, test to see if they come from populations with different mean values of a variable.
- Example: Is the mean GPA of male medical students greater than that of female?
- H_0: $\mu_m = \mu_f$
- H_1 $\mu_m \neq \mu_f$ (2-tailed) or $\mu_m < \mu_f$ (1-tailed)

Two-sample t-test in SPSS

- Analyze…Compare means…Independent samples t-test
- Enter test (dependent) variable, and grouping (independent) variable, and define the two groups by their value on the grouping variable.

Group Statistics

	Sex	N	Mean	SD	SE Mean
HYSC	Male	633	13.23	3.14	.12
	Female	877	12.63	2.84	9.59E-02

Independent Samples Test

Levene's Test for Equality of Variances: F=11.226, p < .001

t-test, equal variances not assumed:

t	df	Sig.	Mean Diff	SE Diff	95% CI Diff
3.824	1276.454	.000	.60	.16	[.29,.91]

Two-sample t-test, reporting

- Men had a significant higher mean number of years of education than women (unequal-variance t(1276)=3.83, p<.05).
- On average, men had 0.6 more years of school than women (95% CI: [.29,.91]).

One-way ANOVA (analysis of variance)

- Goal: Given many samples, test to see if they come from populations with different mean values of a variable.
- Example: Do the mean GPAs of medical students from Sâo Paulo, Marilia, and Botucatu differ?
- H_0: $\mu_{sp} = \mu_m = \mu_b$
- H_1: at least one mean differs

One-way ANOVA in SPSS

- Analyze…Compare means…One-way ANOVA
- Enter test and group vars, optionally set contrasts, post-hoc tests, and options.

Descriptives - Highest Year of School Completed

	N	Mean	SD	Std. Error
White	1262	13.06	2.95	8.32E-02
Black	199	11.89	2.68	.19
Other	49	12.47	4.00	.57

ANOVA - Highest Year of School Completed

	Sum Squares	df	Mean Square	F	Sig.
Betw Grps	240.725	2	120.362	13.746	.000
W/in Grps	13195.994	1507	8.756		
Total	13436.719	1509			

One-way ANOVA, reporting

- An ANOVA was conducted to examine the effect of race (white, black, other) on highest year of education completed.
- There was a significant difference in education between races ($F(2,1507) = 13.7$, $p < .05$).

Contrasts in a one-way ANOVA

- If you reject the overall null hypothesis, you still haven't show which particular means are different from each other.
- You can analyze specific contrasts or comparisons of means to do this.
- You can either do *t* tests on pairs, or (better) program the contrasts into the ANOVA analysis itself. This is more powerful, because it will use the ANOVA's error estimate, which is based on the full sample and usually more accurate

Contrasts in SPSS

- To specify a contrast in a one way ANOVA, use the Contrasts button.
- Contrasts are specified as weights for each group in the grouping variable. For example, if the groups are white, black, other, some contrasts are:

1,-1,0	Compare white and black
-1,1,0	Compare black and white
1,0,-1	Compare white and other
1,-0.5, -0.5	Compare white and mean nonwhite

Contrast output and reporting

Contrast Coefficients

Race of Respondent

White	Black	Other
1	-1	0

Contrast Tests - Highest Year of School Completed

	Value of Contrast	SE	t	df	Sig. (2-tailed)
Equal var	1.16	.23	5.147	1507	.000
Unequal var	1.16	.21	5.607	279.767	.000

- A planned contrast between white and black respondents found that white respondents had significantly more schooling than black ($t(1507)=5.15$, $p<.05$).

Post-hoc tests

- Post-hoc tests are *unplanned* comparisons, performed after looking at the data pattern.
- As such, they capitalize on chance in the data, and should not be accorded as much weight as planned comparisons.
- Moreover, if you perform a large number of statistical tests (common in post-hoc analyses), you must consider the *familywise* (Type I) error rate, which is much larger than the per-test rate.

Example: Bonferroni test

Bonferroni: Highest Year of School Completed

(I) Race	(J) Race	Mean Diff	SE	Sig.
White	Black	1.16*	.23	.000
	Other	.59	.43	.520
Black	White	-1.16*	.23	.000
	Other	-.57	.47	.670
Other	White	-.59	.43	.520
	Black	.57	.47	.670

* The mean difference is significant at the .05 level.

- Corrects for familywise error by setting each test's α to 0.05/(number of tests)

Paired-sample t-test

- Goal: Given 2 measurements of a variable from the same sample, test to see if their means differ between measurements.
- Example: For graduating medical students, are mean GPAs higher at the end of year 2 or the end of year 4?
- H_0: $\mu_{GPA,2} = \mu_{GPA,4}$
- H_1: $\mu_{GPA,2} \neq \mu_{GPA,4}$ (2 tailed)
  or $\mu_{GPA,2} < \mu_{GPA,4}$ (1 tailed)

Paired-sample t-test in SPSS

- Analyze…Compare means…Paired-sample t-test
- Enter pairs of variables

Paired Samples Statistics: LE in 109 countries

	Mean	N	SD	SE Mean
Female LE	70.16	109	10.57	1.01
Male LE	64.92	109	9.27	.89

Paired Samples Test

Paired Differences							
Mean	SD	SE	95% CI		t	df	sig (2-tail)
5.24	2.27	.22	4.81	5.67	24.11	108	.000

- Average life expectancy for women is higher than for men in the same country ($t(108)=24.11$, $p<.05$)

Assumptions of t-tests

- t tests and ANOVA are *parametric* tests: they make assumptions about distribution of scores in the populations from which the means are taken:
  - Distributions are assumed to be normal
  - If two or more population means are being compared, populations are assumed to have equal variances
- These are fairly strong assumptions, but the tests are often ok even if they're violated moderately.
- We'll see *nonparametric* tests later that don't make these assumptions

Monday PM assignment

- Using the hyp data set, test these hypotheses:
  1. The mean spatial perception score is 50
  2. The mean midterm score is different for case-based and lecture formats
  3. The mean final score is higher for case-based than lecture formats
  4. Mean final scores are higher than mean midterm scores
  5. Create a new variable with 3 categories: new (<5 years post-MD), medium (5-15 years post-MD) and old (>15 years post-MD). Do mean satisfaction scores differ by this category?

Tuesday AM

- Presentation of yesterday's results
- Factorial design concepts
- Factorial analyses
  - Two-way between-subjects ANOVA
  - Two-way mixed-model ANOVA
  - Multi-way ANOVA

Factorial designs

- A factorial design measures a variable at different levels of two or more "factors" (categorical independent variables).
- For example, one might measure the efficacy of a drug given in two different forms and at three different dosages.

Factorial designs

- Factors: drug form, drug dosage
- Levels of drug form: oral, inhaled
- Levels of drug dosage: low, medium, high
- Dependent variable: time to pain relief

	low	medium	high
oral	$\mu_{t,l-o}$	$\mu_{t,m-o}$	$\mu_{t,h-o}$
inhaled	$\mu_{t,l-i}$	$\mu_{t,m-i}$	$\mu_{t,h-i}$

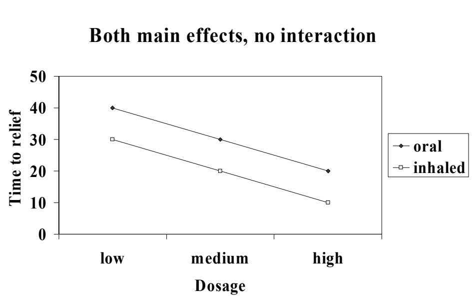
Factorial analyses

- Overall analyses of factorial designs are broken down into main effects and interactions
  - Main effect of dosage
  - Main effect of form
  - Interaction between dosage and form
- When there is no interaction, the main effects are easily interpreted as the independent effects of each factor, as if you'd done t-tests or one-way ANOVAs on the factors.

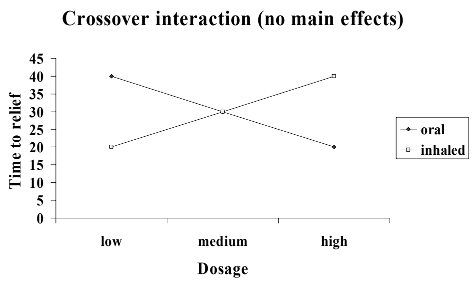
Interactions

- When an interaction is present, the effect of one variable depends on the level of another (for example, inhaled drugs might only be effective at high doses).
- Main effects may or may not be meaningful.
- Graphing the means can show the nature of the interaction.

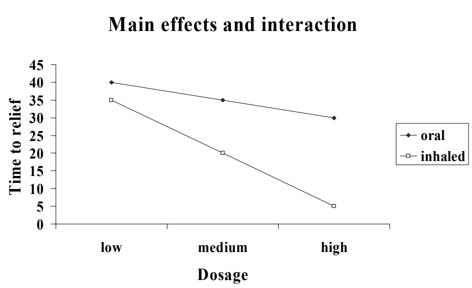
Interaction graphs



Interaction graphs



Interaction graphs



Simple effects and contrasts

- Simple effects are the effects of one variable at a fixed level of another (like doing a one-way ANOVA on dosage for only the oral form).
- Just as you might use contrasts in a one-way ANOVA to identify specific significant differences, you can do the same in factorial analyses

Two-way between-subject ANOVA

- Goal: Determine effects of two different between-subject factors on the mean value of a variable.
- Each cell of the table of means is a different group of subjects.
- Example: Do mean exam scores of students taking PBL or nonPBL versions of physiology taught in Spring, Fall, or Summer differ?
- Each main effect (instruction method, semester) and the interaction has its own null hypothesis

Two-way ANOVA in SPSS

- Analyze…General Linear Model…Univariate
- Enter dependent variable, and fixed factors, and optionally ask for contrasts, plots, tables of means, post-hoc tests, etc.

Tests of Between-Subjects Effects: Occupational Prestige

Source	**SS**	**df**	**Mean Square**	**F**	**Sig.**
SEX	54.460	1	54.460	.330	.566
RACE	7632.679	2	3816.340	23.119	.000
SEX * RACE	1255.778	2	627.889	3.804	.023
Error	233079.627	1412	165.071		

- There was a significant interaction between race and sex ($F(2,1412) = 3.8$, $p < .05$) and a main effect of race ($F(2,1412) = 23.1$, $p < .05$)…. Explain the effects...

Two-way mixed-model ANOVA

- Goal: Determine effects of a b/s and a w/s factor on the mean value of a variable.
- Each row of the table of means is a different group of subjects; each column are the same subjects

	Traditional test	Computer test
Spring	$\mu_{test,traditional-spring}$	$\mu_{test,computer-spring}$
Summer	$\mu_{test,traditional-summer}$	$\mu_{test,traditional-summer}$
Fall	$\mu_{test,traditional-fall}$	$\mu_{test,computer-fall}$

Two-way mixed-model ANOVA

- In standard data format, each of the levels of the within-subject factor is a separate variable (column).
- Analyze…General Linear Model…Repeated Measures
- Name the within subject factor, and give the number of levels, then click Define
- Assign a variable to each level of the within-subject factor
- Assign a variable to code the between-subject factor
- Optionally select contrasts, post-hoc tests, plots, etc.

Two-way mixed-model ANOVA

- Effects of sex (within-country) and predominant religion (between-country) on country's life expectancy

Tests of Within-Subjects Effects

Source	SS	df	Mean Square	F	Sig.
SEX	263.354	1	263.354	143.32	.000
SEX*RELIGION	97.529	9	10.837	5.897	.000
Error(SEX)	180.077	98	1.838		

Tests of Between-Subjects Effects

Source	SS	df	Mean Square	F	Sig.
Intercept	215459.270	1	215459.270	1260.5	.000
RELIGION	4313.969	9	479.330	2.804	.006
Error	16751.749	98	170.936		

Multi-way ANOVA

- Of course, you are not limited to two factors. You can do an ANOVA with any number of factors, between- or within-subjects, and any number of levels per factor, if you have enough data.
- In larger and more complex ANOVAs, however, planned contrasts are often more important than overall interaction effects, etc.

Multivariate ANOVA

- Sometimes you have measurements of multiple different variables (not repeats of the same variable) for the same subjects. You could do a set of ANOVAs on each, or a single multivariate ANOVA (aka MANOVA).
- Sometimes you have repeated measurements of multiple variables for the same subjects. This is called *doubly multivariate* data.
- SPSS can do either with the GLM procedure.

Tuesday AM assignment

- Using the osce data set, test for effects of rater and of patient on the ratings of each of these:
  1. Reasoning
  2. Knowledge
  3. Communication
- If you find any significant effects, plot or table the cell means to illustrate the effects.
- What kind of analyses are these?

Tuesday PM

- Presentation of AM results
- What are nonparametric tests?
- Nonparametric tests for central tendency
  - Mann-Whitney U test (aka Wilcoxon rank-sum test)
  - Sign test, Wilcoxon signed-ranks test
  - Nonparametric ANOVA
- Chi-squared

Nonparametric tests

- As mentioned on Monday, t-tests and ANOVAs are *parametric*: they make assumptions about the distribution of populations (typically, normal distributions)
- *Nonparametric* tests don't require normality, but…
  - They are less powerful (require more subjects)
  - They test slightly different null hypotheses

Mann-Whitney U Test

- Goal: Determine whether two groups differ on a variable. "Nonparametric indepedent t-test"
- Equivalent to the *Wilcoxon rank-sum test*
- Works by ranking all scores across groups, and computing the sum of the ranks within each group. Those rank-sums should be similar if the distributions are similar in each group.
- U or W is reported, with significance.

Mann-Whitney U in SPSS

- Analyze…Nonparametric tests…2 independent samples
- Enter test (dependent) variable and grouping variable
- Do Asian Pacific countries have significant larger populations than Eastern European countries? (t-test might be too sensitive to skew in distribution):

	Test Statistics
Mann-Whitney U	51.000
Wilcoxon W	156.000
Z	-2.699
Asymp. Sig. (2-tailed)	.007
Exact Sig. [2*(1-tailed Sig.)]	.006

- AP countries have significantly larger populations than EE (Mann-Whitney U=51, p<.06)

Sign test

- Goal: Determine whether a variable, measured twice, differs between measurements. “Nonparametric paired t-test”
- Works by examining the difference between each pair of scores, and categorizing it as positive, negative, or zero.
- If the measurements differ, there should be significantly more positive or negative differences.

Sign test

- Analyze…Nonparametric tests…2 related samples
- Enter pairs of variables

Avg male LE - Avg female LE in 109 countries:	
Negative Differences	107
Positive Differences	1
Ties	1
Total	109
Test Statistics	
Z	-10.104
Asymp. Sig. (2-tailed) .000	

- Female life expectancy exceeds male life expectancy in nearly all countries (sign test, Z=-10.1, p < .05).

Wilcoxon signed-ranks test

- Goal: Determine whether a variable, measured twice, differs between measurements. "Nonparametric paired t-test"
- Works by ranking absolute differences between measurements, summing them up for positive and negative differences, and comparing the sums.
- Unlike sign test, gives more weight to pairs that show large differences than to pairs that show small differences.

Wilcoxon signed-ranks test in SPSS

- Analyze…Nonparametric tests…2 related samples
- Enter pairs of variables

Ranks: Avg male LE - Avg female LE

	N	Mean Rank	Sum of Ranks
Negative Ranks	107	54.98	5883.00
Positive Ranks	1	3.00	3.00
Ties	1		

Test Statistics

Z	-9.039
Asymp. Sig. (2-tailed)	.000

- Female LE exceeds male LE across countries (Wilcoxon signed-ranks test, $Z=-9.0$, $p < .05$).

Nonparametric ANOVA

- SPSS also offers nonparametric tests for:
  - 3+ independent groups (Kruskal-Wallis H) "Nonparametric one-way between-subject ANOVA"
  - 3+ repeated measures of same variable (Friedman's test) "Nonparametric one-way within-subject ANOVA"
  - 3+ measures by different raters (Kendall's W)

Chi-squared

- χ^2 is one of the most useful nonparametric statistics. It can be applied to many problems:
  - Is an observed distribution of responses different from an expected on?
  - Are there independent or interactive effects of two categorical variables on a distribution of responses?
  - Are there differences in two related proportions (e.g. proportion of students scoring >90% before and after an educational intervention)?

One-way χ^2

- Given:
  - a set of observed responses divided into categories
  - a set of expected responses divided into categories
    (often a null hypothesis of 'equal distribution')
- Goal: Determine if the observed distribution is significantly different than the expected distribution.

One-way χ^2: example

- Students are asked to choose if they prefer exams in the morning or afternoon. Is there a significant preference?

	Prefer AM	Prefer PM	Total
Observed	39	21	60
Expected	30	30	60

- $\chi^2 = \Sigma(O-E)^2/E = (39-30)^2/30 + (21-30)^2/30 = 5.4$
- Significantly more students prefer morning to afternoon exams ($\chi^2(1)=5.4$, $p<.05$)

One-way χ^2 in SPSS

- Nonparametric tests…Chi-square
- Enter test variable and set expected values if not equally distribute across categories
- Example: We are designing an evaluation in which residents are given a case and asked to make a yes or no decision about performing an LP. We don't expect the residents, on average, to know the right answer, so we expect equal numbers to say yes and no. Did that happen?

One-way χ^2 output

LP Decision

	Observed N	Expected N	Residual
No	28	20.0	8.0
Yes	12	20.0	-8.0
Total	40		

Test Statistics

Chi-Square	6.400
df	1
Asymp. Sig.	.011

- Significantly more residents believed they should not do the LP ($\chi^2(1)=6.4$, $p<.05$)

Two-way χ^2

- Given data in a contingency table (relating responses to two categorical variables)

	Prefer AM	Prefer PM	Total
M1	25	25	50
M2	15	35	50
Total	40	60	100

- Are the effects of the two categorical variables independent or related?
- Same algorithm as one-way (compute expected frequencies based on marginal totals)

Two-way χ^2 in SPSS

- A second case is developed about use of CT (and tested on different residents). Are the distribution of responses to the CT and LP cases the same?
- Analyze…Descriptive statistics…Crosstabs
- Enter a row and column variable to define the contingency table.
- Hit "Options" and check the box for chi-square

Two-way χ^2 output

Form * Prior Decision Crosstabulation

	Prior Decision No	Yes	Total
CT	25	20	45
LP	28	12	40
Total	53	32	85

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1.882	1	.174
Continuity Correction	1.317	1	.251
Likelihood Ratio	1.897	1	.168

- The distributions of responses to the two items were not significantly different.

McNemar's test of correlated proportions

- Given two related proportions, is one significantly higher than the other?
- Example: 85 residents answered the LP case, and were then given a journal abstract that did not support doing LP in the case, and were asked to answer the case again. Did significant fewer do the LP after the evidence?

McNemar's test in SPSS

- Analyze…Nonparametric tests…2 related samples
- Enter variable pair and select McNemar checkbox

	Prior Decision 0	Prior Decision 1
Post Decision 0	50	12
Post Decision 1	3	20

Test Statistics	
N	85
Exact Sig. (2-tailed)	.035

- Residents were significantly less likely to order the LP after reading the evidence (McNemar's test, $p < 0.05$)

χ^2 data considerations

- Observations are assumed to be independent
  (except in McNemar's test)
- χ^2 is not reliable if the expected cell frequencies are smaller than about 5.
- A "correction for continuity" may be applied when expected frequencies are small, but there is argument about appropriateness (see Howell, p 146).

Tuesday PM assignment

- Using the clerksp data set, examine the i1/i1post items (self-rated differential diagnosis skills):
  - Are post-test scores higher than pre-test? Test this question using a paired t-test, a sign test, and the Wilcoxon signed-ranks test. How do the results differ?
  - Create a new variable, nastydoc, coded "1" for clerks whose pre-test i1 rating is higher than their pre-test i15 (expresses caring) rating, and "0" for others. Test whether more than half the clerks are nastydocs using one-way χ^2
  - Create a new variable, IM, coded "1" for clerks whose 1st choice residency before the clerkship was internal medicine, and "0" for all others. Is there a relationship between IM and nastydoc? Test using two-way χ^2 and interpret.

Wednesday AM

- Presentation of yesterday’s results
- Associations
- Correlation
- Linear regression
- Applications: reliability

Associations

- We’re often interested in the association between two variables, especially two interval scales.
- Associations are measured by their:
  - direction (positive, negative, u-shaped, etc.)
  - magnitude (how well can you predict one variable by knowing the score on the other?)

Correlation

- The (Pearson) correlation (r) between two variables is the most common measure of association
  - Varies from -1 to 1
  - Sign represents direction
  - r^2 is the proportion of variance in common between the two variables (how much one can account for in the other)
  - Relationship is assumed to be *linear*

Correlation in SPSS

- Analyze…Correlate…Bivariate
- Enter variables to be correlated with one other.

		Q1	Q2	Q3
Q1	Pearson Correlation	**1.000**	**.105**	**.109**
	Sig. (2-tailed)	.	.111	.099
	N	233	233	231
Q2	Pearson Correlation	**.105**	**1.000**	**.616**
	Sig. (2-tailed)	.111	.	.000
	N	233	234	232
Q3	Pearson Correlation	**.109**	**.616**	**1.000**
	Sig. (2-tailed)	.099	.000	.
	N	231	232	232

- There was a significant positive correlation between Q2 and Q3 ($r = 0.62$, $p < .05$).

Linear regression

- Correlation is a measure of association based on a linear fit.
- Linear regression provides the equation for the line itself (e.g. $Y = b_1X + b_0$)
- That is, in addition to providing a correlation, it tells how much change in the independent variable is produced by a given change in the dependent variable...
- ... in both natural units and standardized units.

Linear regression in SPSS

- Analyze…Regression…Linear
- Enter dependent and independent variables
- Three parts to output:
  - Model summary: how well did the line fit?
  - ANOVA table: did the line fit better than a null model?
  - Regression equation: what is the line? How much change in the dependent variable do you get from a 1 unit (or 1 standard deviation) change in the independent variable

Linear regression output

- Predicting Q2 from Q3:

Model Summary

R	R Square	Adjusted R Square
.616	.380	.377

- R is the correlation
- R^2, the squared correlation, is proportion of variance in Q2 accounted for by variance in Q3
- Adjusted R^2 is a less optimistic estimate

Linear regression output

ANOVA

	Sum of Sq	df	Mean Square	F	Sig.
Regression	153.924	1	153.924	140.8	.000
Residual	251.455	230	1.093		
Total	405.379	231			

- Shows that the regression equation accounts for a significant amount of the variance in the dependent variable compared to a null model.
- (A null model is a model that says that the mean of Q2 is the predicted Q2 for all subjects).

Linear regression ouput

Coefficients

	Unstandardized		Standardized		
	B	Std. Error	Beta	t	Sig
(Constant)	.804	.315		2.554	.011
Q3	.693	.058	.616	11.866	.000

- Unstandardized coefficients (B) give the actual equation:

  Q2 = 0.693 * Q3 + 0.804

  - These are raw units. An increase of 1 point in Q3 increases Q2 by 0.693 points on average. People who have Q3 = 0 have Q2 = 0.804 on average, etc.
  - Because SE of B is estimated, we can perform t-tests to see if a B is significantly different than 0 (has a significant effect).
- Standardized coefficients (β) give the amount of change in Q2 caused by a change in Q3, measured in standard deviation units. They are useful in multiple regression (later)...

Measuring reliability of a scale

- Test-retest reliability is usually measured as the correlation between tests (ranks of subjects stay the same at each testing)
- Cronbach's α is another common internal reliability measure based on the average inter-item correlation of items in a scale.

Cronbach's α in SPSS

- Analyze…Scale...Reliability analysis
- Enter item variables that make up the scale
- Go to Statistics dialog box and ask for *scale* and *scale if item deleted* descriptives.

Cronbach's α in SPSS

Item-total Statistics

	Scale Mean if Item Deleted	**Scale Variance if Item Deleted**	**Corrected Item-Total Correlation**	**Alpha if Item Deleted**
Q1	21.2913	9.2466	.3133	.6071
Q2	23.4000	6.0576	.4507	.5325
Q3	22.5826	6.4975	.4798	.5096
Q4	21.9043	8.5148	.3565	.5840
Q5	22.2130	7.4173	.3448	.5870

Reliability Coefficients

N of Cases = 230.0 **N of Items** = 5

Alpha = .6229 **Standardized item alpha** = .6367

Wednesday AM assignment

- Using the clerksp data set:
  - Examine the correlations between items 1-17 (self-ratings of different clerkship skills). What do you notice about the correlation matrix?
  - Select any one of those 17 items. Run a linear regression to determine if the pre-clerkship rating on that item predicts the post-clerkship rating.
  - Assume that we want to combine post-clerkship items 1-17 into a single scale of self-related clerk skill. What would the reliability of this scale be?

Wednesday PM

- Presentation of AM results
- Multiple linear regression
  - Simultaneous
  - Stepwise
  - Hierarchical
- Logistic regression

Multiple regression

- Multiple regression extends simple linear regression to consider the effects of multiple independent variables (controlling for each other) on the dependent variable.
- The line fit is: $Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots$
- The coefficients (b_i) tell you the independent effect of a change in one dependent variable on the independent variable, in natural units.

Multiple regression in SPSS

- Same as simple linear regression, but put more than one variable into the independent box.
- Equation output has a line for each variable:

Coefficients: Predicting Q2 from Q3, Q4, Q5

	Unstandardized		Standardized		
	B	**SE**	**Beta**	**t**	**Sig.**
(Constant)	.407	.582		.700	.485
Q3	.679	.060	.604	11.345	.000
Q4	-.028	.095	-.017	-.295	.768
Q5	.112	.066	.095	1.695	.091

- Unstandardized coefficients are the average effect of each independent variable, controlling for all other variables, on the dependent variable.

Standardized coefficients

- Standardized coefficients can be used to compare effect sizes of the independent variables *within the regression analysis.*
- In the preceding analysis, a change of 1 standard deviation in Q3 has over 6 times the effect of a change of 1 sd in Q5 and over 30 times the effect of a change of 1 sd in Q4.
- However, βs are not stable across analyses and can't be compared.

Stepwise regression

- In simultaneous regression, all independent variables are entered in the regression equation.
- In stepwise regression, an algorithm decides which variables to include.
- The goal of stepwise regression is to develop the model that does the best prediction with the fewest variables.
- Ideal for creating scoring rules, but atheoretical and can capitalize on chance (post-hoc modeling)

Stepwise algorithms

- In *forward* stepwise regression, the equation starts with no variables, and the variable that accounts for the most variance is added first. Then the next variable that can add new variance is added, if it adds a significant amount of variance, etc.
- In *backward* stepwise regression, the equation starts with all variables; variables that don't add significant variance are removed.
- There are also hybrid algorithms that both add and remove.

Stepwise regression in SPSS

- Analyze…Regression…Linear
- Enter dependent variable and independent variables in the independents box, as before
- Change "Method" in the independents box from "Enter" to:
  - Forward
  - Backward
  - Stepwise

Hierarchical regression

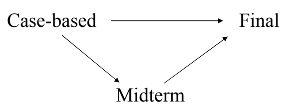
- In hierarchical regression, we fit a hierarchy of regression models, adding variables according to theory and checking to see if they contribute additional variance.
- You control the order in which variables are added
- Used for analyzing the effect of dependent variables on independent variables in the presence of moderating variables.
- Also called *path analysis*, and equivalent to *analysis of covariance (ANCOVA)*.

Hierarchical regression in SPSS

- Analyze…Regression…Linear
- Enter dependent variable, and the independent variables you want added for the smallest model
- Click "Next" in the independents box
- Enter additional independent variables
- …repeat as required…

Hierarchical regression example

- In the hyp data, there is a correlation of -0.7 between case-based course and final exam.
- Is the relationship between final exam score and course format moderated by midterm exam score?



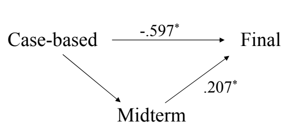
Hierarchical regression example

- To answer the question, we:
  - Predict final exam from midterm and format (gives us the effect of format, controlling for midterm,
    and the effect of midterm, controlling for format)
  - Predict midterm from format
    (gives us the effect of format on midterm)
- After running each regression, write the βs on the path diagram:

Predict final from midterm, format

Coefficients

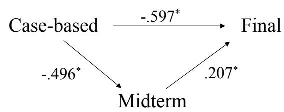
	B	SE	Beta	t	Sig.
(Constant)	50.68	4.415		11.479	.000
Case-based course	-26.3	3.563	**-.597**	-7.380	.000
midterm exam score	.156	.061	**.207**	2.566	.012



Predict midterm from format

Coefficients

	B	SE	Beta	t	Sig.
(Constant)	63.43	3.606		17.59	.000
Case-based course	-29.2	5.152	**-.496**	-5.662	.000



- Conclusions: The course format affects the final exam both directly and through an effect on the midterm exam. In both cases, lecture courses yielded higher scores.

Logistic regression

- Linear regression fits a line.
- Logistic regression fits a cumulative logistic function
  - S-shaped
  - Bounded by [0,1]
- This function provides a better fit to binomial dependent variables (e.g. pass/fail)
- Predicted dependent variable represents the probability of one category (e.g. pass) based on the values of the independent variables.

Logistic regression in SPSS

- Analyze…Regression…Binary logistic (or multinomial logistic)
- Enter dependent variable and independent variables
- Output will include:
  - Goodness of model fit (tests of misfit)
  - Classification table
  - Estimates for effects of independent variables
- Example: Voting for Clinton vs. Bush in 1992 US election, based on sex, age, college graduate

Logistic regression output

- Goodness of fit measures:

```
-2 Log Likelihood     2116.474          (lower is better)
Goodness of Fit       1568.282          (lower is better)
Cox & Snell - R^2         .012          (higher is better)
Nagelkerke - R^2          .016          (higher is better)

                    Chi-Square    df   Significance
Model                  18.482     3        .0003
```

(A significant chi-square indicates poor fit (significant difference between predicted and observed data), but most models on large data sets will have significant chi-square)

Logistic regression output

```
Classification Table
The Cut Value is .50
                        Predicted
                      Bush   Clinton    Percent Correct
                        B   |    C
Observed           -------------------
   Bush        B   |     0  |   661  |      .00%
                   -------------------
   Clinton     C   |     0  |   907  |   100.00%
                   -------------------
                                Overall  57.84%
```

Logistic regression output

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
FEMALE	.4312	.1041	17.2	1	.0000	.0843	1.5391
OVER65	.1227	.1329	.85	1	.3557	.0000	1.1306
COLLGRAD	.0818	.1115	.53	1	.4631	.0000	1.0852
Constant	-.4153	.1791	5.4	1	.0204		

- B is the coefficient in log-odds; Exp(B) = e^B gives the effect size as an odds ratio.
- Your odds of voting for Clinton are 1.54 times greater if you’re a woman than a man.

Wednesday PM assignment

- Using the semantic data set:
  - Perform a regression to predict total score from semantic classification. Interpret the results.
  - Perform a one-way ANOVA to predict total score from semantic classification. Are the results different?
  - Perform a stepwise regression to predict total score. Include semantic classification, number of distinct semantic qualifiers, reasoning, and knowledge.
  - Perform a logistic regression to predict correct diagnosis from total score and number of distinct semantic qualifiers. Interpret the results.

Thursday AM

- Presentation of yesterday's results
- Factor analysis
- A conceptual introduction to:
  - Structural equation models
  - Mixed models

Factor analysis

- Given responses to a set of items (e.g. 36 likert-scaled questions on a survey)…
- Try to extract a smaller number of *common latent factors* that can be combined additively to predict the responses to the items.
- Variance in response to an item is made up of:
  - Variance in common factors that contribute to the item
  - Variance specific to the item
  - Error

Factor analysis: survey design

- Typically, a large set of likert-scaled items
- Design points:
  - 5 (or better, 7) response categories per item
  - 3-5 items per expected factor
  - 3-5 subjects per item
- Example: residency training survey data set
  - Likert scale with 7 categories per item
  - 41 items in 5 expected factors (3-16 per factor)
  - 234 subjects (nearly 6 subjects per item)

Factor analysis: decisions

- *Exploratory* or *confirmatory* analysis?
- How will factors be *extracted*? (initial solution)
  - Principal components analysis
  - Maximum likelihood methods
- How will I choose *how many* factors to extract?
  - Based on theory
  - By scree plot
  - By eigenvalue

Factor analysis: decisions

- How will factors be *rotated*? (rotated solution)
  - Orthogonal rotation (Varimax, etc.)
  - Oblique rotation (Promax, Oblimin, Quartimin)
- How should factors be *interpreted*?
  - Pattern matrix
  - High and low items

Factor analysis in SPSS

- Analyze…Data reduction…Factor
- Enter items in Variables box
- Click "Extraction" and choose extraction method and how number of factors will be determined.
- Click "Rotation" and choose rotation method.
- Click "Scores" if you want to save factor scores
- Click "Options" and ask to have coefficients ("loadings") sorted by size and to have small coefficients suppressed.

(c) Alan Schwartz, UIC DME, 1999

Use of factor scores

- Once factors are derived, *factor scores* can be computed for each subject on each factor
- Factor scores indicate how the subject perceives each of the factors.
- Factor scores can be used as variables in regression analyses (including path analyses).

Factor analysis assignment

- Conduct factor analyses on the residency training data set and see what you can learn:
  - Vary some of the “decisions” and see how the results change.
  - If you find an interpretable solution, save the factor scores and see if they are related to any of the residency program demographics.

Structural equation models

- Structural equation modeling is a technique that combines confirmatory factor analysis (the *measurement model*) and path analysis (the *structural model*) and does both at the same time.
- Requires specialized statistical software
  - Lisrel
  - EQS
  - Amos for SPSS

Mixed models

- Aka:
  - General (or generalized) linear models with fixed and random effects
  - Random-effects models
  - Random-intercept models
  - Hierarchical linear models
  - Multilevel models

Why mixed models? Clustering

- Participants clustered in groups
  - Example: test the association between MCAT scores and a new rating instrument administered in a medicine clerkship.
  - There may be differences between each clerkship rotation that would cause the ratings of clerks in a given clerkship to be not wholly independent of one another.
  - Because the usual correlation coefficient (or linear regression, or t-tests, etc.) assumes independent observations, you would not be able to use it.
- Observations clustered in participants
  - Example: clerks are rated on communication skills five times during the year
  - Compare the rate of improvement (or decay) for clerks who get a special training course at the start of the year vs those who don't.
  - Scores are clustered within the clerks and not truly independent observations.
  - Scores taken from consecutive months may be more closely correlated
  - Some clerks may be missing a rating (at random)
- Multiple cases, multiple raters, etc. problems

Random effects: The key concept

- Instead of assuming that a regression coefficient is fixed value we want to estimate,
- Assuming that the coefficient is a random variable, and we want to estimate its mean and variance
- That can mean something like: "Each group in the regression gets its own intercept, drawn from a normal distribution around the overall effect"
- It can also mean that we can model a variety of nonindependant relationships between variables

How to do this stuff

- Think about whether clustering is present in your research design
- Discuss the research design and plan the analysis with a statistician or data analyst in advance. Bring up the issue of clustering with the statistician in that discussion, and determine an appropriate way to control for it
- Get the assistance of the statistician in interpreting the results of the analysis. You might want to ask whether the analysis suggests that observations did have substantial independence or not (sometimes this is part of the research question, but often it's just reassuring to hear that you had dependence and to pat yourself on the back for employing a mixed model and controlling for it!)

Resources

- Applied Mixed Models in Medicine (Brown and Prescott) – Introductory chapters are particularly good.
- SAS for Mixed Models (Littell, et al.)
- Hierarchical Linear Models (Raudenbush & Bryk) – for some people, a more intuitive way to think about the problem that reduces to the same math
- Linear Mixed Models (West, Welch, & Galecki) – covers SAS, SPSS, Stata, and others
- Mixed Models for Repeated (Longitudinal) Data (Howell) – very well written: http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Mixed%20Models%20for%20Repeated%20Measures.pdf
- Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models (Singer) – also very well written, a classic: http://gseweb.harvard.edu/~faculty/singer/Papers/sasprocmixed.pdf
- The University of Bristol also offers an excellent online course in multilevel modeling called LEMMA, with very good self-assessment quizzes. It's at: http://www.cmm.bristol.ac.uk/learning-training/course.shtml

(c) Alan Schwartz, UIC DME, 1999