UNIVERSITY OF WASHINGTON

Practical Statistics for Human-Computer Interaction

Independent Study using SAS JMP and IBM SPSS.

Jacob O. Wobbrock, Ph.D.

The Information School | DUB Group

University of Washington

Seattle, WA 98195 USA

Acknowledgement

This work was supported in part by the National Science Foundation under CAREER grant IIS-0952786. Any opinions, findings, conclusions or recommendations expressed in this work are those of the author and do not necessarily reflect those of the National Science Foundation.

Table of Contents

Introduction

I have often found that students in my field of human-computer interaction (HCI), particularly from the computer science tradition, lack, as I did, the proper statistical skills to evaluate results of studies, whether experimental, quasi-experimental, observational, or response-based. These fundamentals can be hard to come by. On the one hand, formal coursework in statistics departments is, especially at the graduate level, overly theoretical for easy application to real-world HCI studies. On the other hand, students who simply read help documentation for a particular statistics package are likely to make numerous errors due to a lack of statistics fundamentals. What has been missing is the ground between these two extremes, a place for students to learn enough principled material to avoid blunders, while also learning the practical handling of real-world data sets in a statistics package.

To provide this middle ground is my purpose in creating this independent study. It is designed to be completed in a 10-week period, which fits nicely into the quarter system at many universities. While the modules are meant to be doable without aid of an instructor, periodic feedback from an instructor is presumed. I ask my own students to email me their modules each week, and after annotating them, I discuss any problems with them the following week. Students can also come to me when they are having difficulties along the way, although I encourage them first to reach the boiling point of frustration as they *independently* try to resolve their confusion. That is the nature of an independent study, after all.

Many, but not all, of the sample data sets that accompany this study are situated in the context of human-computer interaction. Others are from experiments in other fields. Most of the data sets are fictitious, designed by me to exhibit certain distributional properties and so forth.

The pedagogical goal for this study is that by its end, a student will be proficient in many common statistical analyses, and know how to recognize the proper analysis from simply viewing a data table, even without knowing how the experiment was run. This level of proficiency is sufficient for both comprehending statistical results in HCI papers, and for producing them in the first place.

A trial version of the SAS JMP statistics package is available from http://www.jmp.com/. A trial version of the IBM SPSS statistics package is available from http://www.spss.com/ . I readily admit the downside of having two proprietary statistics packages as the focus of this independent study, but these are the tools with which I am familiar. If someone wants to reproduce this study and its data sets for use with **R**, I would be most grateful.

I welcome comments, suggestions, and errata. You can reach me by email at wobbrock@uw.edu.

Jacob O. Wobbrock, Ph.D.
June 2010
Seattle, WA

01. Concepts and definitions

For the following statistics terms, discover the formal definition (*e.g.*, on-line or in a textbook), and then write a "common sense" definition reflecting the *intuition* of the concept. With each common sense definition, supply an example that shows you really know what the term means. You may have to read a considerable amount before gaining enough awareness to define each of these terms correctly. You might consider first writing the description for a single experiment, and then referring to various aspects of that experiment in your answers.

1. Factor
2. Level
3. Independent variable
4. Dependent variable
5. Measure
6. Trial
7. Covariate
8. Within-subjects factor
9. Between-subjects factor
10. Factorial design
11. Main effect
12. Interaction
13. Within-subjects design
14. Between-subjects design
15. Mixed factorial design
16. Confound
17. Control
18. Carryover effect
19. Counterbalancing
20. Nominal variable
21. Categorical variable
22. Ordinal variable
23. Continuous variable
24. Scalar variable
25. Fixed effect
26. Random effect
27. Mixed-effects model

02. Understanding a data table

dinner.jmp

Open the JMP data table *dinner.jmp*. Examine the table and then answer the following questions.

1. Does it represent a within-subjects, between-subjects, or mixed design? How do you know?
2. What columns are the independent variables? What columns are the dependent variables?
3. What are the factors? What are the levels?
4. Which factors are nominal? Ordinal? Continuous?
5. How would you represent this with the multiplicative pattern of describing experiments (*e.g.*, "a 3×2×3 within-subjects design")?
6. Given the variable modeling types, what type of analysis should be performed: ANOVA, Chi-Square, nominal logistic regression, ordinal logistic regression, or numeric regression? For each of these, say what the independent variable types and the dependent variable type must be.
7. What is a "balanced design?" Is this a balanced design? Why or why not?
8. Explain the structural differences between long-format and a wide-format data tables. Is *dinner.jmp* a long-format table or a wide-format table? How do you know?
9. Give a plausible story about why and how this experiment was conducted.
10. Try doing an analysis and reporting your findings.

03. Introduction to ANOVA

dinner.jmp

For this module, we're still going to use *dinner.jmp*. You may have to look up certain things in the JMP help documentation. You may also have to refer to module 2 to remember how you interpreted the experiment design.

After opening *dinner.jmp*, please do the following:

Part 1, Descriptive statistics. Go to Analyze > Distribution. In Y, Columns put the dependent variable. In By, put only one independent variable. (Do this again in a separate analysis for the other independent variable.)

1. Eyeballing the descriptive statistics and histograms, what seems to be the case about the effect of each independent variable on the dependent variable?
2. What might you expect the ANOVA to show? How do you know? Write your answer before going further.

Part 2, ANOVA. Go to Analyze > Fit Model. This is where you provide a model that, when fitted to your data, will attempt to predict the outcome variable from the independent variables and their interactions. As before, put your outcome variable in Y. Put your two independent variables and their interaction (also called their "cross") in the Construct Model Effects box. Then click Run Model. Study the Summary of Fit, Analysis of Variance, Effect Tests, and the Residual by Predicted Plot.

3. Using published papers for reference, craft a paragraph that precisely reports the conclusions of this ANOVA. It should probably be about 2-4 sentences with statistics cited in-line in the appropriate places.

Part 3, REML. In part 2, you ran a traditional between-subjects ANOVA, but the state-of-the-art is an analysis of variance based on regression that uses REstricted (or REsidual) Maximum Likelihood, or REML. While the mathematics of this approach is beyond the current scope, you should know that REML analyses of variance can do everything ANOVA can do, and more. Unlike in traditional ANOVA, with REML you can have dropped/missing data. Unbalanced designs are easily handled. Also, a mixture of categorical and continuous independent variables can be used. And for within-subjects or repeated measures designs, REML allows for the modeling of each subject with covariance structures, which can take into account learning over time. These issues will be covered in more detail in later modules.

For simple analyses, ANOVA and REML will agree. Prove this to yourself: go to Analyze > Fit Model. Add the same model terms you did before but this time, also add *Subject* to the model effects. Then highlight *Subject* and click on the Attributes menu and select Random Effect. Be sure the Method being used is REML and the Emphasis is Effect Leverage. Then run the model.

4. Verify that the results are the same as before.
5. Does the REML model fit any better or worse than the ANOVA model did? How do you know?

Part 4, Random Effects. REML analyses use what are called mixed-effect models, where you have some model effects that are called "fixed," and some, in this case *Subject*, that are called "random." Fixed effects are those whose levels are chosen intentionally and beyond which generalizations are not meant to be drawn. Random effects, on the other hand, are those whose levels are pulled randomly from a population over which the results are meant to generalize. The specific levels of a random effect, therefore, are not of any interest, only that they are randomly sampled. *Subject* is a classic random effect.

6. In light of the description here, why is *Subject* a random effect?
7. Provide an example of another random effect from a hypothetical or real experiment. (Hint: think of soil plots in agricultural studies.)

04. Statistics tools

posts., postsctrl.*, postsbtwn.*, postswthn.*, postsmix.*, postsmfx.**

Now that you have some initial exposure to statistical concepts, data sets, and analyses of variance, we are going to spend one module familiarizing ourselves with two industrial-strength statistics packages, SAS JMP and IBM SPSS. Many researchers use two (or more) tools when analyzing data. Doing so helps to compare results to ensure correctness. (Another popular tool is **R**, which is free but uses a command-line interface. You are welcome to use **R** but it is currently not supported as part of this independent study.)

Unlike the rest of the modules in this independent study, this module provides low-level details on carrying out analyses using these tools, and requires less independent discovery on your part. The goal here is to increase your comfort and familiarity with both JMP and SPSS, and to see how different statistics packages "name and frame" their features. The rest of the modules in this independent study will be based mainly on JMP, but with occasional forays into SPSS. After this module, cross-package exploration and comparison should be easy and is heretofore encouraged.

Free trial versions are available for both JMP and SPSS. JMP can be downloaded at http://www.jmp.com/. SPSS can be downloaded at http://www.spss.com/. (This module is based on JMP 7.0 and SPSS 19.0, but later versions of these products should suffice.) The file extensions for data sets in JMP are **.jmp*, and for SPSS are **.sav*.

Follow the instructions for each type of analysis *carefully*, and write up your answers to the questions that appear after each analysis. You are meant to perform each analysis in JMP and SPSS concurrently. This module can be completed with only one tool, but the comparative benefits are lost.

Part 1, *t*-test. Open the files *posts.** of a hypothetical study of 40 college students' Facebook posting behavior using one of two platforms: Apple's iOS or Google's Android OS. The data show the number of Facebook posts subjects made during a particular week using their mobile platform.

SPSS. Be sure View > Value Labels is checked for the data labels to show correctly. Perform a *t*-test by going to Analyze > Compare Means > Independent-Samples T Test. Transfer *Posts* into the Test Variable(s) box and *Platform* into the Grouping Variable. Click Define Groups. Enter "1" for Group 1 and "2" for Group 2. (This works because the *Platform* variable is defined as having one of two integer values, with 1 corresponding to "iOS," and 2 corresponding to "Android." To see this, go to the Variable View pane, and examine the Values column entry for *Platform*.) Click Continue. Then click OK.

JMP. Perform a *t*-test by going to Analyze > Fit Y by X. Put *Posts* into the Y, Response box and *Platform* into the X, Factor box. Click OK. Then access the menu from the little red arrow on the Oneway Analysis of Posts by Platform heading. (You can also always access the same menu by right-clicking anywhere on headings with red arrows.) From the menu, choose Means and Std Dev. Also choose Means/Anova/Pooled t. (For the *t*-test, you'll use only the results in the section labeled t Test. You can collapse unwanted sections using the little blue arrows.)

1. Is this a between-subjects or within-subjects experiment? Why?

2. What is the independent variable named?
3. How many levels (*i.e.*, values) does the independent variable have? What are they?
4. How many subjects were in this experiment?
5. How many subjects were exposed to each level of the independent variable? Is the design balanced (*i.e.*, are the numbers equal)?
6. What are the mean and standard deviation number of posts for each level of the independent variable?
7. Assuming equal variances, what is the *t* statistic for this *t*-test? (Hint: this is also called the *t* Ratio.)
8. How many degrees of freedom are there (*df*s)?
9. What is the two-tailed *p*-value resulting from this *t*-test? Is it significant at the $\alpha = .05$ level?
10. The formulation for expressing a significant *t*-test result is: $t(df\text{s}) = t\text{-statistic}$, $p < .05$. For a nonsignificant result, it is: $t(df\text{s}) = t\text{-statistic}$, *n.s.* Write your result in the proper formulation.
11. What can we conclude about the results of this study? Which platform seems to result in more Facebook posts, if any? Write a sentence summarizing the finding from this analysis.

Part 2, One-way ANOVA (F-test). Still using the files *posts.**, we will perform an F-test, which can do everything a *t*-test can do, and more. An F-test, which is the most common analysis of variance, can handle multiple independent variables, or factors, and these factors can have more than two levels. By comparison, a *t*-test can only have one factor with two levels, which is not very useful for many experiment designs.

SPSS. To perform an F-test on the data in *posts.sav*, go to Analyze > Compare Means > One-Way ANOVA. Into the Dependent List, transfer *Posts* as your measure. Into the Factor box, transfer *Platform*. Click Options and check the Descriptive box. Click Continue and click OK.

JMP. To perform an F-test on the data in *posts.jmp*, go to Analyze > Fit Y by X. Put *Posts* into the Y, Response box and *Platform* into the X, Factor box. Click OK. From the red-arrow menu, choose Means/Anova/Pooled t. Now you will use the Analysis of Variance section and the Means for Oneway Anova section.

12. Do the number of observations (*N*'s) and means agree with those produced by the *t*-test? What are they? (If they do not agree, there is an error somewhere!)
13. In the ANOVA table, what is the F-statistic? What is the *p*-value? Is it significant at the $\alpha = .05$ level?
14. How does this *p*-value compare to that produced by the *t*-test? Did you expect it to be otherwise?
15. *SPSS only***.** If this experiment had more than one factor (see parts 5-8, below), we would use Analyze > General Linear Model > Univariate. Use this function on the same data and see if you achieve the same results. (Hints: you don't use *Subject*, and you don't have to select any options within a subdialog beneath the Univariate dialog.) Report the line on which your F-test result appears.
16. *JMP only***.** If this experiment had more than one factor (see parts 5-8, below), we would use Analyze > Fit Model. Use this function on the same data and see if you achieve the same results. (Hint: you don't use *Subject*, but put *Posts* as your response and *Platform* as your model effect.) Report the section(s) in which your results appear.

17. The general formulation for expressing an F-test result is: F(df_{num},df_{den}) = F-ratio, $p < .05$ (or *n.s.*). Report the F-test result, filling in df_{num} and df_{den} with the numerator and denominator (also called "error") degrees-of-freedom, respectively.

Part 3, Mann-Whitney *U* test, Wilcoxon rank sums test. The use of a *t*-test or an F-test requires the upholding of some assumptions, which will be covered in more detail in module 9. One of these assumptions is that the variance in the measures of subjects in each group is similar. Another is that the measures form a roughly normal distribution, or would if enough measures were taken from the population. When these assumptions are violated, a nonparametric test should be used. (Nonparametric tests are not a specific test but a class of test.) Such a test does not presume an underlying normal distribution. They are therefore sometimes called "distribution free." Such tests are useful when, for example, analyzing Likert scale data, which is often non-normal. The equivalent of a between-subjects (independent samples) *t*-test using nonparametric statistics is the Mann-Whitney *U* test. The Mann-Whitney *U* test is a test for an experiment containing one between-subjects factor with two-levels. More detail on nonparametric tests will be given in module 9; this exercise is just to give you some initial exposure.

SPSS. Perform a *U* test on *posts.sav* by going to Analyze > Nonparametric Tests > Legacy Dialogs > 2 Independent Samples. Transfer *Posts* into the Test Variable List. Transfer *Platform* into the Grouping Variable box. Click Define Groups and enter "1" for Group 1 and "2" for Group 2. Click Continue. Click Options and check the Descriptive box. Then check the Mann-Whitney U box. Click OK.

JMP. JMP does not perform a Mann-Whitney *U* test, but a similar test, which will give equivalent results, is the Wilcoxon rank sums test. (This is not to be confused with the popular matched-pairs Wilcoxon signed-rank test.) Perform a Wilcoxon rank sums test on *posts.jmp* by going to Analyze > Fit Y by X. Put *Posts* into the Y, Response box and *Platform* into the X, Factor box. Click OK. From the red-arrow menu, choose Nonparametric > Wilcoxon Test. The results are in the section labeled 2-Sample Test, Normal Approximation.

18. Do the Counts (*N*'s) agree with those produced by the *t*-test and F-test? What are they? (If they do not agree, there is an error somewhere!)
19. *SPSS only***.** What is the Mann-Whitney *U* value produced?
20. *JMP only***.** *JMP only***.** What is the Wilcoxon *W* value produced? (Hint: this is the *S* value displayed.) Does it agree with the SPSS value for Wilcoxon W?
21. This test produces a *Z*-value as its statistic. What is the *Z*-value?
22. What is the two-tailed *p*-value? Is it significant at the $\alpha = .05$ level? How does the result compare to the *t*-test and F-test results for the same data from parts 1 and 2, above?
23. The formulation for a Mann-Whitney *U* test result is *U* = U-value, *Z* = Z-value, *N* = *n*, $p < .05$ (or *n.s.*). The formulation for a Wilcoxon rank sums test is *W* = W-value, *Z* = Z-value, *N* = *n*, $p < .05$ (or *n.s.*). Report both tests, filling in the appropriate values. Note that there are no degrees of freedom in these nonparametric tests.
24. What can we conclude about the results of this study? Which platform seems to result in more Facebook posts, if any? Write a sentence summarizing the finding from this nonparametric analysis.

Part 4, One-way ANOVA with 3+ levels. As noted above, the F-test can handle more than one factor, and also, more than two levels per factor. A one-way ANOVA refers to a single factor design. Similarly, a two-way ANOVA refers to a two-factor design, *i.e.*, two independent variables. In this part, we will still conduct a one-way ANOVA, but this time, our factor will have three levels. Thus, it cannot be analyzed with a *t*-test, which can only handle two levels of a single factor.

SPSS. Open *postsctrl.sav*. This data set is the same for the iOS and Android levels, but now has added 20 new college students as a control group who did not use a mobile device for posting on Facebook but were told to use their desktop computer instead. Thus, the *Platform* factor now has three levels: iOS, Android, and desktop. To perform an ANOVA, repeat the steps you did in part 2 using Analyze > Compare Means > One-Way ANOVA. However, this time, before executing the analysis, click Post Hoc and select LSD and Bonferroni. Click Continue and click OK.

JMP. Open *postsctrl.jmp*. To perform an ANOVA, repeat the steps you did in part 2 using Analyze > Fit Y by X. (On the red-arrow menu, you will choose Means/Anova, as Means/Anova/Pooled t has been automatically renamed to drop the "Pooled t" part because a *t*-test is no longer possible.)

25. Was this a one-way, two-way, or three-way analysis of variance? What is/are the factor(s)? What are each factor's levels?
26. How many data points are there for each level of *Platform*? What are the means and standard deviations for each level? (*JMP* hint: use both Means/Anova and Means and Std Dev from the red-arrow menu.)
27. In this case, the overall, or *omnibus*, F-test is testing for whether any differences exist among the levels of the independent variable. Is the F-test significant at the $\alpha = .05$ level? What is the F-ratio? What is the *p*-value? Construct a proper reporting formulation for this result.
28. *SPSS only.* The omnibus F-test does not tell us whether all three levels of *Platform* are different from one another, or whether just two levels (and which two?) are different. For this, we need *post hoc* comparisons, which are justified only when the omnibus F-test is significant. Examine the Post Hoc Tests output. The LSD output does *not* correct for making multiple *post hoc* comparisons (remember that with $\alpha = .05$, we can expect 1 in 20 tests to be significant purely by chance). Which pairwise comparisons are significant without *post hoc* correction?
29. *JMP only.* Use the red-arrow menu to choose Compare Means > Each Pair, Student's t. This performs *post hoc* pairwise *t*-tests for each pair of levels within our factor. These tests are *not* corrected for multiple comparisons (remember that with $\alpha = .05$, we can expect 1 in 20 tests to be significant purely by chance). Which pairwise comparisons are significant without *post hoc* correction? (Hint: use the Level, -Level bottommost table produced to make this judgment, or use the red-arrow menu on the Comparisons header to select Detailed Comparisons Report.)
30. A Bonferroni correction divides α by the number of *post hoc* comparisons. In this case, with three *post hoc* comparisons, we would use $\alpha = .05 / 3 = .0166$. (Equivalently, we can multiply our *p*-values by 3 and then continue using $\alpha = .05$, which is what SPSS's Bonferroni output does.) Under the Bonferroni correction, which pairwise comparisons are significant? Did any comparisons move from significant before correction to nonsignificant after? If so, which?
31. Assuming a Bonferroni correction, what can we conclude from this study? Write a sentence summarizing the findings from this analysis.

32. This study added one control group. However, one other obvious control group is missing. If we were to add a second control group of 20 college students, what should that control group be?

Part 5, Two-way between-subjects ANOVA. It is often the case that we wish to examine the effects of more than one factor, and we also care about the interaction among factors. Because multiple factors are involved, this is called a factorial design, expressed as $N_1 \times N_2 \times \ldots \times N_n$ for an arbitrary number n of factors, and where each N_i is an integer indicating the number of levels of that factor. In practice, it is difficult to interpret experiments with more than three factors, especially if those factors each have more than two levels.

For this part, we will examine an augmented version of our current study that adds another factor. Open *postsbtwn.**. You will see another column labeled *Day* with values "weekday" and "weekend." These values correspond to the days of the week the subject was allowed to post to Facebook.

SPSS. To perform a two-way between-subjects ANOVA, go to Analyze > General Linear Model > Univariate. You should be able to discover which variables go where. (Hint: only the Dependent Variable and Fixed Factor(s) boxes need to be filled in. *Subject* is not used. It is included in the data tables for clarity, and will come into play later when we use mixed-effects models in part 8, below.) Click Options and transfer all the contents of the Factor(s) and Factor Interactions box to the Display Means for box. Check Compare main effects. Check Descriptive statistics and Estimates of effect size. Click Continue and click OK.

JMP. To perform a two-way between-subjects ANOVA, go to Analyze > Fit Model. You should be able to discover which variables go where. Be sure that your model effects contain not only the independent variables, but also their interaction. (Hint: You should have one Y value and three items as model effects. *Subject* remains unused for now. It is included in the data tables for clarity, and will come into play later when we use mixed-effects models in part 8, below.) Click Run Model.

33. Was this a one-way, two-way, or three-way analysis of variance? What is/are the factor(s)? What are each factor's levels? Express the design using $N_1 \times N_2 \times \ldots \times N_n$ notation.
34. For each identified factor, was it between-subjects or within-subjects? How do you know?
35. What were the means and standard deviations for the number of posts on weekdays? weekends?
36. Write the F-test result for the *Platform* factor. Is it significant at the $\alpha = .05$ level? What is its *p*-value? Does this differ from the finding prior to the inclusion of the *Day* factor (part 2, above)? If so, how?
37. Write the F-test result for the *Day* factor. Is it significant at the $\alpha = .05$ level? What is its *p*-value?
38. Write the F-test result for the *Platform*Day* interaction. Is it significant at the $\alpha = .05$ level? What is its *p*-value?
39. Within each factor, why don't we need to perform any *post hoc* pairwise comparison tests?
40. Interpret these results and craft three sentences describing the results of this experiment, one for each factor and one for the interaction. What can we say about the findings from this study? (Hint: *p*-values between .05 and .10 are often called "trends" or "marginal results," and are often reported, although they cannot be considered strong evidence. Be wary of ever calling such results "marginally significant." A result is either significant or it is not; there is no "marginal significance.")

Part 6, Two-way within-subjects ANOVA. Thus far, we have only considered experiments where one subject was measured once on only one level of each factor. But often we wish to measure a subject more than once, perhaps for different levels of our factor(s), or over time, in which case *time* itself becomes a factor. Such designs are called "repeated measures" designs, and the factors on which we obtain repeated measures are called within-subjects factors (as opposed to between-subjects factors). For repeated measures studies, we can still use an ANOVA, but now we use a "repeated measures ANOVA," and our data table inevitably looks different: for a wide-format table, there are now multiple measures per row (each row still corresponds to just one subject, as it has thus far).

Our current hypothetical study on Facebook posts has been modified to be a purely within-subjects study. Imagine that each college student was issued either an iOS or Android device for one week, and then the other device for the next week. Also, each college student's posts were counted separately on weekdays and weekends. Instead of needing 40 college students as before, we now only need 10 students for the same data, which is shown in *postswthn.**. Open those files and see the wide-format data tables.

SPSS. To perform a traditional repeated measures ANOVA, go to Analyze > General Linear Model > Repeated Measures. For repeated measures analyses, the data table must be in wide-format, which has all measures for a given subject in a single row. (This contrasts with long-format, which has only one measure per subject in a given row, and for repeated measures designs, it must therefore have multiple rows for the same subject, which will occur in part 8, below.) Thus, we must tell SPSS which factors are represented by each column. For the first factor, enter the factor that changes the *slowest* across columns, *e.g.*, *Platform*. It should be clear by now that it has two levels—iOS and Android—so enter "2" in Number of Levels. For the second factor, enter the factor that changes *fastest* across columns, *e.g.*, *Day*. It also has two levels: weekday and weekend. Once these factors are added, click Define. Next, highlight the four result columns and transfer them to the Within-Subjects Variables box. Next click Options. Transfer everything in the Factor(s) and Factor Interactions box into the Displays Means for box. Check the Compare main effects box. Also check Descriptive statistics and Estimates of effect size. Click Continue and click OK.

JMP. To perform a traditional repeated measures ANOVA, go to Analyze > Fit Model. On the Fit Model dialog, put all four results columns into the Y box. For Personality, choose Manova. No model effects are used in the Construct Model Effects box, as everything is within-subjects here (that box would be used if we had any between-subjects factors). Click Run Model. Now click Choose Response > Compound, which is used for two layers of repeated measures. Because our factors are encoded in the columns themselves, we must tell JMP which factors are represented by each column. In the top textbox, change "factor1" to the factor that moves *fastest* across columns, namely *Day*. In the left textbox, change "factor2" to the factor that moves *slowest* across columns, namely *Platform*. Check the boxes for Create Interaction Effect also and Univariate Tests Also. Click OK.

41. Was this a one-way, two-way, or three-way analysis of variance? What is/are the factor(s)? What are each factor's levels? Express the design using $N_1 \times N_2 \times \ldots \times N_n$ notation.
42. For each identified factor, was it between-subjects or within-subjects? How do you know?
43. Verify that the means and standard deviations for the number of posts in each condition match the values produced in the previous version of the data (part 5, above).

44. *SPSS only.* Write the statistical result for the *Platform* factor. (Hint: the table you care about is called Tests of Within-Subjects Effects. Use the Sphericity Assumed line, and remember that the word "error" is synonymous with "denominator" when reporting F-test results.)
45. *JMP only.* Write the statistical result for the *Platform* factor. (Hint: you can collapse all of the Whole Model sections and ignore them. The section you care about is called Intercept within the Platform report. Use the Univar unadj Epsilon line, ignoring the Value column.)
46. Write the statistical result for the *Day* factor.
47. Write the statistical result for the *Platform*Day* interaction.
48. What can we conclude about the results from this study?
49. In general, was the significance level as judged by the *p*-values for each factor stronger or weaker (*i.e.*, numerically lower or higher) than for the same data from 40 subjects in part 5, above? Why do you think that is?

Part 7, Two-way mixed factorial design. After seeing purely between-subjects and purely within-subjects designs, you may wonder whether mixed factorial designs involving factors of each type are possible. Indeed they are, and *postsmix.** contains our same data in a hypothetical mixed design. Imagine that each college student is again assigned only one mobile device for a week, an Apple iOS device or Google Android device, but that the student is allowed to use it both on weekdays and on weekends. This is still considered a repeated measures design because at least one factor is within-subjects, but now it is a mixed factorial design, meaning at least one factor is also between-subjects.

SPSS. To perform a traditional repeated measures ANOVA, go to Analyze > General Linear Model > Repeated Measures and fill in the dialogs based on what you learned in the previous parts. Don't forget to add your between-subjects factor.

JMP. To perform a traditional repeated measures ANOVA, go to Analyze > Fit Model and fill in the dialog based on what you learned in the previous parts. (Hint: you'll have two Y responses.) Don't forget to add your between-subjects factor in the model effects box. Choose Manova as the Personality and click Run Model. Select Choose Response > Repeated Measures and enter for "Y Name" the factor embedded in your Y response columns, namely *Day*. Check the Univariate Tests Also box and click OK.

50. Was this a one-way, two-way, or three-way analysis of variance? What is/are the factor(s)? What are each factor's levels? Express the design using $N_1 \times N_2 \times \ldots \times N_n$ notation.
51. For each identified factor, was it between-subjects or within-subjects? How do you know?
52. Verify that the means for the number of posts in each condition match the values produced in the previous version of the data (parts 5 and 6, above).
53. *SPSS only.* Write the statistical result for the *Platform* factor. (Hint: you care about the Platform line in the Tests of Between-Subjects Effects.)
54. *JMP only.* Write the statistical result for the *Platform* factor. (Hint: your result for Platform is in the Between Subjects > Platform report.)
55. *SPSS only.* Write the statistical result for the *Day* factor. (Hint: you care about the Tests of Within-Subjects Effects.)
56. *JMP only.* Write the statistical result for the *Day* factor. (Hint: use the Univar unadj Epsilon line in the Day report and ignore the Value column.)

57. Write the statistical result for the *Platform*Day* interaction. (*SPSS* hint: you care about the Tests of Within-Subjects Effects, and Error(Day) provides your denominator degrees of freedom.)
58. What can we conclude about the results from this study?
59. In general, was the significance level as judged by the *p*-values for each factor stronger or weaker (*i.e.*, numerically lower or higher) than for the same data from 40 subjects in part 5, above? than from the 10 subjects in part 6, above? In each case, why do you think that is? Remember that there were 20 subjects in this mixed design.

Part 8, Mixed-effects models. Although similar in terminology to a "mixed factorial design," a mixed-effects model analysis of variance means something else entirely. "Mixed effects" refer to the mixture of fixed and random effects, where "effects" are synonymous with factors. A fixed effect is everything we have considered thus far: it is a factor whose specific levels we care to draw conclusions about, and whose values were therefore determined *a priori* during experiment design. Parts 1-7 of this module have dealt only with fixed effects. Indeed, fixed effects ANOVAs represent the traditional method for analyzing parametric data. However, with the advent of personal computing, more computationally intensive analyses have become available. One is to perform analyses of variance (no longer written as "ANOVA" to avoid confusion) using an underlying regression-based model, covariance matrices, and a method known as REstricted (or REsidual) Maximum Likelihood, or REML. You briefly encountered REML in module 3.

A REML analysis has many advantages over a traditional fixed-effects ANOVA. Unlike in traditional ANOVA, a REML analysis can have dropped or missing data. Unbalanced designs are also handled. Also, a mixture of nominal and continuous factors can be used, instead of only nominal factors. And for repeated measures designs, REML allows for separate modeling of each subject over trials using covariance matrices, which takes into account learning. REML analyses are therefore preferred to traditional repeated measures ANOVAs for analyzing repeated measures.

So what is a *random effect* that makes such models "mixed?" Unlike a fixed effect whose levels you specifically care about, a random effect is a factor whose levels are sampled randomly from a larger population about which we care to draw inferences. Accordingly, we do not care about the specific levels of a random effect. Classic examples of random effects are the actual college students used in our Facebook posting study, or the plots of soil into which different types of seeds are planted. We wish to generalize to all college students and all soil plots, respectively, and do not care specifically about which humans were used or in which plots we planted.

Unlike the wide-format data tables used in traditional repeated measures ANOVAs, the table for mixed-effects model analyses must be in long-format. For a given dependent variable, long-format tables have only one measure per row, and with repeated measures, subjects therefore must be listed on multiple rows. We represent the within-subjects data from part 6 in long-format as *postsmfx.**. Remember that there are only 10 unique subjects for this hypothetical within-subjects experiment. So why are there 40 rows in our long-format table? Because each subject is repeated four times. Now our *Subject* column finally matters. Thus, our columns are the same as for our fully between-subjects data in *postsbtwn.**, but now with only 10 unique subjects, not 40.

SPSS. To perform a mixed-effects model analysis of variance on *postsmfx.sav*, go to Analyze > Mixed Models > Linear. Transfer *Subject* into the Subjects box and *Platform* and *Day* into the Repeated

box, as these are repeated factors. For Repeated Covariance Type, choose Compound Symmetry. (This choice will be revisited in module 8.) Click Continue. The Dependent Variable is *Posts*. The Factor(s) are *Platform* and *Day*. Then click Fixed and highlight *Platform* and *Day* in the Factors and Covariates box. Click Add to show these factors plus their interaction in the Model box. Click Continue. Click EM Means, highlight everything in the Factor(s) and Factor Interactions box and transfer it to the Display Means for box. Check the Compare main effects box. Click Continue and click OK.

JMP. To perform the mixed-effects model analysis of variance on *postsmfx.jmp*, go to Analyze > Fit Model. Put *Posts* in the Y box. Highlight both *Platform* and *Day* and choose Macros > Full Factorial to construct your model effects. Next, for the first time, also put *Subject* in the model effects box. Then highlight *Subject* and use the Attributes red-arrow menu to select Random Effect. Subject should become "Subject& Random." Be sure Personality is Standard Least Squares, Emphasis is Effect Leverage, and Method is REML. Click Run Model. You will mostly care about the results in the Fixed Effects Tests section.

60. What are the fixed effects? What are the random effects? How many levels does each of these effects have? What are those levels?
61. Verify that the means and standard deviations for the number of posts in each condition agree exactly with those from the wide-format within-subjects data from part 6, above.
62. Write the statistical result for the *Platform* factor. How does it compare to the result from the traditional repeated measures ANOVA from part 6, above? (Do not just compare the *p*-values, but also the F-ratios and the numerator and denominator degrees-of-freedom.)
63. Write the statistical result for the *Day* factor. How does it compare to the result from the traditional repeated measures ANOVA?
64. Write the statistical result for the *Platform*Day* interaction? How does it compare to the result from the traditional repeated measures ANOVA?
65. What can we conclude about the results from this study?
66. Share any thoughts you have about how mixed-effects models compare, either statistically or conceptually, to traditional ANOVAs.

Part 9, Contrast tests. As we briefly discussed in part 4, above, a significant omnibus test for a given factor simply indicates that there exists *some* significant difference among levels of that factor. If the factor has only two levels, then we know where that difference lies. But if the factor has more than two levels, a *post hoc* comparison, called a "contrast test" or "pairwise comparison," is probably in order using the Bonferroni correction.

Another circumstance in which contrast testing is useful is *across* factors in a factorial design. For example, in *postsmfx.**, suppose we wish to know whether there was a significant difference between numbers of posts in on weekdays using Android, and numbers of posts on weekends using iOS. How would we test for such a difference? Contrast testing provides the answer.

SPSS. First, it is illustrative to graph the interaction. Using the *postsmfx.sav* data table from part 8, go to Graphs > Chart Builder. (Click OK on the dialog that appears.) Highlight Line in the Choose from box, and drag the Multiple Line graph into the chart preview area. From your Variables box, drag *Posts* onto the Y-Axis? box. Drag *Day* onto the X-Axis? box. Drag *Platform* into the Set color area. Then click OK.

Visit your SPSS output to see the graph. You can then visually compare (Android, weekday) to (iOS, weekend).

By default, SPSS will not perform contrast tests *across* factors. To do this, we need to use the SPSS Syntax Editor. After running your mixed-effects model analysis of variance from part 8, go to the SPSS Statistics Viewer and click on the code snippet that SPSS generated for you to perform that analysis. Use CTRL+C to copy it. Then use File > New > Syntax to open the syntax editor. Use CTRL+V to paste the code snippet there. Erase any initial lines before the `MIXED` command. You should see the last four lines of code using the `EMMEANS` command. For *Platform* and *Day*, these lines are accompanied by the `COMPARE` statement, which will compare levels separately within *Platform* and *Day*. No such statement accompanies *Platform*Day*. We can add the `COMPARE` statement to this line, specifying the factor whose levels we wish to compare *within* the levels of the other factor. Thus, `COMPARE(Platform)` will compare platforms within each level of *Day*. However, this cannot be used to compare *across* factors of an interaction. For that, we need the `TEST` subcommand. With this command, we can specify exactly the contrast we want by giving weights.[1] In the end, your script should look like this. Only the last two lines have been modified, with the modifications shown in bold:

```
   MIXED Posts BY Platform Day
     /CRITERIA=CIN(95)  MXITER(100)  MXSTEP(10)  SCORING(1)  SINGULAR(0.000000000001)
HCONVERGE(0, ABSOLUTE) LCONVERGE(0, ABSOLUTE) PCONVERGE(0.000001, ABSOLUTE)
     /FIXED=Platform Day Platform*Day | SSTYPE(3)
     /METHOD=REML
     /REPEATED=Platform*Day | SUBJECT(Subject) COVTYPE(CS)
     /EMMEANS=TABLES(OVERALL)
     /EMMEANS=TABLES(Platform) COMPARE ADJ(LSD)
     /EMMEANS=TABLES(Day) COMPARE ADJ(LSD)
     /EMMEANS=TABLES(Platform*Day) COMPARE(Platform) ADJ(LSD)
     /TEST = 'Contrast of Android,weekday to iOS,weekend' Platform*Day 0 -1 +1 0 .
```

 Unfortunately, for reasons I have not been able to discern, the `TEST` subcommand for the *postsmfx.sav* data set will not execute. SPSS complains that the "L matrix is not estimable." Searching the Web and SPSS documentation did not provide any answers. The above detail is included here for future reference, as the `COMPARE` option is sufficient when we want to test *within* a factor, and apparently the `TEST` command usually works just fine. (If you decode this mystery, please let me know!)

Note: If upon running this script you receive an error concerning the `CRITERIA` subcommand, it is probably because you have a broken line due to line-wrapping. Closely examine your syntax for the `CRITERIA` subcommand and ensure that the script is not broken across lines.

JMP. First, it is illustrative to graph the interaction. On your statistical output from part 8, find the report labeled Platform*Day. Use its red-arrow menu and choose LSMeans Plot. You will see that the *x*-axis is assigned the *Day* factor, and the lines themselves are assigned the *Platform* factor. Use the red-arrow menu again to uncheck LSMeans Plot. Now, hold down SHIFT *before* clicking, and use the red-arrow menu again to choose LSMeans Plot. This time, you can see the graph has been "sliced" the other

[1] See http://www.ats.ucla.edu/stat/spss/faq/test_subcommand_mixed_mod.htm for more detail.

way, with *Platform* on the *x*-axis and *Day* assigned to the lines. Now re-slice the graph the original way (without SHIFT) and then proceed to perform a contrast test.

Use the red-arrow menu to choose LSMeans Contrast. Here, you choose opposing positive and negative weights to indicate which levels you wish to combine and compare. To test (Android, weekday) *vs.* (iOS, weekend), toggle weights +1 and -1 next to each of these, respectively. Then click Done.

67. Describe the graph(s) you created. What main effect does the slant in the two lines describe? What main effect does the vertical gap between the two lines describe? What effect do the parallel lines describe?
68. *JMP only.* What is the result for the contrast test you performed for (Android, weekday) *vs.* (iOS, weekend)? Report and interpret this result as you would in a paper.
69. When performing only one *post hoc* contrast test, why is a Bonferroni correction unnecessary?

05. Repeated measures

cockpits.jmp

In this module, we are going to look again at within-subjects designs due to their prevalence in human-computer interaction studies. As you know, these are also called repeated measures designs because each subject is measured more than once. This is in contrast to between-subjects designs, where each subject is measured only once. Remember that mixed factorial designs have both between- and within-subjects factors. Note that "repeated measures" is a term sometimes used to refer to longitudinal studies, where subjects are measured more than once *over time*, perhaps even with *time* itself being the only independent variable to vary from one measure to the next (*i.e.*, all other factors' levels are kept the same). We will be considering longitudinal studies in module 8.

This module uses data table *cockpits.jmp*, which is the same data as one of the JMP sample files, *cholesterol.jmp*, but recast to represent a fictitious human factors study of combat pilots. Open *cockpits.jmp* table before answering the following questions.

Part 1, Table examination. Carefully examine the data table. Also, right-click on Notes > Edit at the top-left of the data table to read study notes for this fictitious study. (Such notes can be created in JMP by using the red-arrow menu on "cockpits" and choosing New Table Variable. Also, notes can be added to individual columns by right-clicking on the column name, choosing Column Info > Column Properties > Notes.)

1. Is this table in wide-format or long-format? What would the other look like (*i.e.*, how many rows and columns, and what might the column names and variable types be)?
2. Characterize the design of this experiment in the appropriate experimental terms: what are the factors and their levels? For each factor, is it within-subjects or between-subjects?
3. Describe the experiment in everyday terms. What is a plausible account of what was done here? Also, explain the importance of the Placebo level of *CockpitUI*.

Part 2, Descriptive statistics. Using JMP, go to Analyze > Distribution, and place the six result columns in Y, Columns. Then on the Distributions menu, choose Stack and Uniform Scaling. Visually compare the plots.

4. Does the data seem normal? (Non-normal data is not amenable to ANOVA and must be transformed before an ANOVA is performed. Transformations are the topic of module 6.)
5. Can you eyeball the results of the within-subjects significance tests from these distributions? Make your predictions *before* proceeding. (Hint: you should be making predictions about two factors and their interaction here, but there is another factor not visible in these distributions. Which is that factor? Why isn't it visible here?)

Part 3, Repeated measures ANOVA. Now we will perform a repeated measures ANOVA. Note that in papers and other write-ups, calling this a "repeated measures ANOVA" is important, not just an "ANOVA." Reviewers want to ensure this data was analyzed by taking repeated measures into account *i.e.*, that measures on the same subject were not regarded as independent. To perform it, go to Analyze >

Fit Model. Put your six result columns in Y. Put your one between-subjects factor in Construct Model Effects. (Remember that your subject term, *Pilot*, is not used as a model effect unless we are using mixed-effects models with long-format tables.) For Personality, choose Manova. Then click Run Model.

6. Look at the graphs in the Least Squares Means section. (You can drag a graph's borders to enlarge it.) Eyeballing the graphs, can you now make better predictions about which effects are significant? Remake your predictions here. (Hint: you should be making predictions about three factors and all possible interactions among them.)

First, we have to test the between-subjects factor, *CockpitUI*. Using the Choose Response menu, select Sum. Click Run. The report you care about is under the CockpitUI heading. It shows the main effect of *CockpitUI*.

7. Report and interpret the *CockpitUI* effect. Do not just report the statistical result, but also interpret its meaning using means and standard deviations. (Hint: you may find it convenient to use Microsoft Excel to help compute these values.)

Now we want to examine within-subjects factors and their relation to each other and to the between-subjects factor. From the Choose Response menu, choose Compound. This menu item is used when *two* within-subjects factors are represented across columns. (If we had only one within-subjects factor, we could choose Repeated Measures instead.) The factor that changes fastest across columns should be named at the top, and its levels specified. In this case, put "TimeOfDay" in the top textbox and indicate that it has 2 levels. Put "Month" in the left textbox and then study the grid: it should look properly laid out and labeled. Also, check the boxes for Create Interaction Effect also and Univariate Tests Also. Then click OK.

Find the TimeOfDay results section. (It is in small font under the first section called Compound.) The *TimeOfDay* main effect is under the Intercept section, and is on the Univar unadj Epsilon line. We ignore the Value column, but we do care about the Exact F (the F-test value), NumDF (numerator degrees of freedom), DenDF (denominator degrees of freedom), and Prob>F (the *p*-value) columns.

8. Report and interpret the *TimeOfDay* effect. Do not just report the statistical result, but interpret its meaning, also. (Hint: you need to look at means for levels of *TimeOfDay* to do so.)

In the CockpitUI section in the TimeOfDay results, we find the *CockpitUI*TimeOfDay* interaction. Again we care about the result on the Univar unadj Epsilon line.

9. Report and interpret the *CockpitUI*TimeOfDay* interaction. What, conceptually, is this interaction testing?

Now find the *Month* results section. Notice here that we have a Sphericity Test, which is a test that takes place in repeated measures ANOVA but is unnecessary in mixed-effects model analyses of variance. If the Sphericity Test is significant, we cannot use an unadjusted ANOVA result for that particular factor, and instead must use a corrected result, such as the Greenhouse-Geisser (G-G) results. The G-G results can be found on the Univar G-G Epsilon line. For the *Month* factor, the Sphericity Test is nonsignificant ($p = .27$), so we do not have a violation of sphericity and can report the unadjusted result as we have been doing.

10. In the Intercept section for Month, find, report, and interpret the *Month* main effect. What is this testing? What does the result indicate?
11. Now find the *CockpitUI*Month* interaction, and report and interpret that result.

Find the *TimeOfDay*Month* interaction results.

12. Is a Greenhouse-Geisser correction necessary? How do you know?
13. In this report, the Intercept section is the result for the *TimeOfDay*Month* interaction. Is it significant? Report and interpret the result.
14. What is this interaction testing? What does it mean?

Finally, find the results for the *CockpitUI*TimeOfDay*Month* three-way interaction.

15. Report and interpret the *CockpitUI*TimeOfDay*Month* interaction.
16. What is this three-way interaction testing? (You will quickly realize that gaining an intuition for a three-way interaction is difficult, but not impossible. But four-way interactions just might be impossible to intuitively comprehend!)

Part 4, Avoiding confounds from carryover effects. In experiments with repeated measures, carryover effects from one level of a factor to another can confound results if not properly neutralized through counterbalancing. Full counterbalancing requires that all possible orderings of a factor's levels are administered to subjects an equal number of times. Partial counterbalancing is also possible if many levels of repeated factors are present. A popular scheme for partial counterbalancing is called a Latin Square. Carryover effects are unavoidable, but counterbalancing distributes these effects evenly across levels—that is, provided there is no *asymmetric skill transfer*, where the carryover effect differs depending on which levels came first. In cases where asymmetric skill transfer is present, a between-subjects factor should be used instead of a within-subjects factor.

17. For each factor in this study, discuss whether it requires counterbalancing. For those factors that require it, how could counterbalancing be achieved? What are the implications if counterbalancing is not employed? What confounding issues could explain the results if not counterbalanced out? What would be an example of asymmetric skill transfer? How can it be handled?

Part 5, Change table format. Recall from modules 3 and 4 that mixed-effects models are useful, often even preferable, for repeated measures designs. But to perform a mixed-effects model analysis of variance, you need a long-format table. Recall your answers from question 1, above.

18. What columns should your long-format table have?
19. How many rows should it have?
20. For how many rows should each subject be repeated? Be sure you understand exactly what the long-format version of this data should look like. You might consider first sketching it out on paper before proceeding.

Although Tables > Stack provides support for transforming a wide-format table into a long-format one (and Tables > Split does the reverse), Tables > Stack is not well designed and causes more headaches than it relieves. (Tables > Split, on the other hand, is very useful.) For a table this small, it is feasible to

manually create a long-format table using copy and paste. You can do this in Microsoft Excel and paste it back into an empty data table in JMP, or do it directly in JMP. Be extremely careful not to corrupt your data!

21. Once your long-format table is completed, state the number of columns and rows it has, and what its column names are. (These names should correspond exactly to the factor names from your analysis in part 3, above.)
22. Sanity-check your transformed data by computing means and standard deviations (use either Tables > Summary, or easier, Analyze > Distribution). Ensure that these values match those from your wide-format table. If all the means, standard deviations, quartiles, *etc.* agree, you'll know you transformed your data table correctly from wide-format to long-format.

Part 6, Mixed-effects model. Now perform a mixed-effects model analysis of variance using the REML method from modules 3 and 4. Recall that we avoid writing "ANOVA" because that connotes a traditional ANOVA involving only fixed effects. (Hint: go to Analyze > Fit Model, enter your one Y measure. You now have three columns as independent variables. To enter them swiftly, highlight *CockpitUI*, *Month*, and *TimeOfDay* in Select Columns, and then choose Macros > Full Factorial. Then add *Pilot* by itself. Now highlight *Pilot* in Construct Model Effects and *CockpitUI* in Select Columns and click Nest[2] to produce *Pilot[CockpitUI]*. Finally, highlight this term in Construct Model Effects and use the Attributes red-arrow menu to choose Random Effect. The final built term should be *Pilot[CockpitUI]& Random*. Be sure the REML method is used with Standard Least Squares personality and Effect Leverage emphasis.)

Compare the results of the mixed-effects model analysis of variance with the traditional repeated measures ANOVA.

23. Are the same factors significant? With similar F- and *p*-values?
24. How do the degrees-of-freedom compare?
25. Judging by the F- and *p*-values, with which analysis does it seem easier to achieve significance?
26. Write a few sentences describing the statistical analysis you just performed as you would for a publication.

Part 7, More than two within-subjects factors. If you have more than two repeated factors, JMP cannot perform a traditional repeated measures ANOVA because the Analyze > Fit Model > Manova > Choose Response > Compound procedure we used in part 3 cannot handle more than two factors. With 3+ repeated factors for a traditional repeated measures ANOVA, we can use SPSS, which can handle arbitrary numbers of within-subjects factors in a wide-format table.

Run SPSS and paste your original wide-format JMP data from *cockpits.jmp* into a spreadsheet. Name and set the column values properly in the Variable View, including using integers and value labels for

[2] Nesting and nested effects are covered in module 7. The reason for nesting *Pilot* in the between-subjects factor *CockpitUI* is to avoid pooling pilots who used different cockpits. Nesting causes only pilots who used the same cockpit interfaces to be pooled. It turns out that because *Pilot* encodings are all unique for each subject, JMP produces the correct results in this case even if nesting is not used, *i.e.*, using a *Pilot& Random* term. However, if each pilot identifier had been assigned *within* each cockpit interface (*e.g.*, the first pilot in cockpits A, B, Control, and Placebo all get pilot identifier "1"), then nesting would be essential.

CockpitUI, and then use Analyze > General Linear Model > Repeated Measures. (Hint: when you execute this command, you'll provide names for your within-subjects factors and their number of levels, and then map your columns to these. Don't forget to also add the between-subjects factor to your model.) Although the data in *cockpits.jmp* has only two repeated factors and not three, you will see easily how SPSS could be used to support any number of repeated factors.

Run the ANOVA in SPSS and explore the output report. In particular, you want to pay attention to the following sections in the SPSS output:

- *Within-Subjects Factors.* Verify that your within-subjects variables and levels are properly built.
- *Between-Subjects Factors.* Verify that your between-subjects variable(s) appear there with the proper number of levels and data points.
- *Mauchly's Test of Sphericity*. Compare to the Sphericity Test outputs from JMP.
- *Tests of Within-Subjects Effects*. This is where your main results will be—note that the Error(*x*) rows give your denominator degrees-of-freedom.
- *Tests of Between-Subjects Effects*. The Error line gives your denominator *df*s, and the line with the name of your effect gives your numerator *df*s.

Compare your SPSS repeated measures ANOVA output to that from JMP for the repeated measures ANOVA in part 3.

27. How do your results compare among these two reports?
28. Why might exact values not always match?
29. Do the study conclusions change for any of these analyses?

06. Transformations

debugging.jmp, *errors.jmp*, *errorsrare.jmp*

In this module, we will explore the use of common transformations on data for the purposes of making them amenable to analyses of variance. We will focus on the common $\log_e$ transform, but a reading included with these modules discusses a wider range of possible transformations than you'll be exposed to here.

Part 1, Examine the data table. Open *debugging.jmp* and examine it. Before going further, answer the following orientation questions:

1. Is this table in wide-format or long-format? How do you know?
2. How many subjects participated in this experiment?
3. What are the factor(s) in this experiment? For each factor, say what variable type it is and whether it is between-subjects or within-subjects.
4. What is the response in this experiment? What variable type does it have?
5. What type of analysis is warranted based on the factor and response variable types?
6. Write up a plausible explanation of what this study was examining.

Part 2, Explore the distributions. Use the Analyze > Distribution menu to explore this data. First explore it as a whole, and then explore it using the By box to visually compare the two factors to each other. (Hint: use Uniform Scaling and Stack to make visual comparisons easier.)

7. Describe what you see on the overall data. Does it look normally distributed to you?
8. When looking at two plots by *IDE*, what do you see? Do you think there is an effect of *IDE* on *Seconds*?
9. When looking at two plots by *Language*, what do you see? Do you think there is an effect of *Language* on *Seconds*?

Part 3, Test for normality. Again use Analyze > Distribution and examine the whole data set. Use the red-arrow menu on Seconds and choose Fit Distribution > Normal.

10. A normal distribution, based on a Gaussian function, is fully parameterized by two parameters. In general, what are they? And what values do they show for this fitted curve?

Next, use the red-arrow menu on Fitted Normal and select Goodness of Fit. This causes JMP to perform a Shapiro-Wilk *W* test of normality.

11. Conceptually, how do you think the Shapiro-Wilk *W* test works? What is it doing? Report the results of the test, and say whether your *Seconds* data can be considered normal or not. Does it appear to violate the normality assumption of ANOVA?

Next, use the red-arrow menu on Seconds to fit a LogNormal distribution. If a distribution is normal under a logarithmic transform, it is said to be lognormal. Examine the Goodness of Fit.

12. What kind of test is performed? Report the results of the test.
13. According to the test, can your *Seconds* data be considered lognormal?

Part 4, Apply a log transform. Return to your data table and create a new column. Name it log(Seconds) and enter the Formula editor. Under Functions (grouped), choose Transcendental > Log. Under Table Columns, choose Seconds. (You should now see Log(Seconds) as the formula for this column. Note that Log by itself uses base *e*, which is typical by convention. If you wanted to supply a different base, you would press the comma (,) key after choosing Log and a subscript would appear.) Click OK. You should now see your new column populated.

14. Use Analyze > Distribution on your newly created column and perform the same two tests for normality that we did in part 3. Report the test results and conclusions. Be sure to name each test appropriately.
15. When time is a dependent variable, especially times of many seconds or minutes, it often exhibits a lognormal distribution. Think about the nature of time and human performance, and give a brief explanation of why time is often lognormal.

Part 5, Perform the analysis. Perform two analyses of variance, one on your original response variable, and one on your newly transformed response. The analysis you perform is up to you. (Your choices are either a traditional fixed-effects ANOVA or a mixed-effects model analysis of variance. Be sure to describe the analysis you performed and the model you used.)

16. What are the conclusions of these analyses?
17. How do their findings compare?
18. Write up the description and the results for this study as you would for a publication. One long paragraph is sufficient.

Part 6, Using a constant. Next, open the file *errors.jmp* and examine the data table.

19. What does this experiment seem to have been about?
20. What does the dependent variable seem to measure? Does it seem reasonable that it might be lognormally distributed? Why or why not?
21. Examine the distribution of *Total Errors* and fit normal and lognormal distributions to it. What happens? Why? Write up what you find from analyzing the distribution.
22. How could you remedy your inability to fit a lognormal distribution to the data?
23. Attempt to create a log-transformed column for your data. What happens? Why? Again, what could you do to make this work? (Write your response before proceeding.)

One option is to change from doing $y' = \log(y)$ to instead doing $y' = \log(y+1)$. This can be generalized to $y = \log(y+c)$, for some constant c. It is often the case that $c = 1$ suffices, although statisticians have more complex approaches to determining c based on particular data (*see* Berry 1987).

24. Using $c = 1$, apply the log-transformation. Then re-examine the distribution using Analyze > Distribution. Does it work now? Perform a normal fitness test and report the results.
25. Perform a mixed-effects model analysis of variance on both the original response and your newly transformed response. As you did in part 5, write up the experiment, procedure, analysis, and findings as you would for a publication (*see* question 18).

Part 7, Challenges. Now open *errorsrare.jmp* and examine the data table. Pay special attention to the frequency (or lack thereof) of errors in the distributions you examine.

26. What problems can you anticipate with applying a log-transform (or, frankly, most any other transform)?
27. Apply a workable log-transform as you did previously. Re-examine the distribution of data using Analyze > Distribution. Did the log-transform succeed in normalizing the data? Why or why not?

In reality, these error data are not amenable to ANVOA because of the high frequency of zeroes. Even using a log-transform, one cannot change the frequency of outcomes, just their distance from one another. We will see in module 9 how to use nonparametric techniques to analyze data such as these.

28. Despite its inappropriateness, use a mixed-effects model analysis of variance on both your raw outcome and your log-transformed outcome with Analyze > Fit Model and report your findings.

Another challenge of using a transformation is *interpreting* your results. In research write-ups, it is essential to be clear when reporting results based on a transformed response variable. The more intricate the transform, the harder it is to straightforwardly interpret the results.

29. Look back at your answers in previous parts of this module. Were you careful to indicate that you were interpreting findings from a log-transformed response? Or did you write your answers just as you would have had you not transformed your response?

Part 8, Other transforms. Acquire the Hoyle (1973) reading and familiarize yourself with the 19 transformations it reviews in section 4. You'll see that some of them are quite involved.

30. Name the three transformations you think are most relevant to human-computer interaction data. Why those three? What kind of data would warrant their use?

07. Nesting

designers.jmp, *higgins.jmp*

Nesting is an issue that arises in many types of experiments in which a hierarchical or containment structure is used to organize factors. It also arises in mixed factorial designs when long-format tables are used, as experimental units (Subjects) must be nested in between-subjects factors. With nesting, we prevent the introduction of unwanted variation by pooling nested factors only within levels of nesting factors, rather than across levels. Before beginning the exercises, it is important to give you an intuition regarding nested effects so that you can recognize them when they should be built. (A rule of thumb is to be on the lookout whenever you have between-subjects factors. As stated, nested effects occur in mixed factorial designs that have both between- and within-subjects factors, although they can occur for other reasons, *e.g.*, with covariates.)

Part 1, Text entry example. As a fictitious example, consider a study to examine the effects of bodily posture on mobile text entry performance. A researcher examined the text entry performance of subjects while they were sitting, standing, walking, and riding on a bus. Each subject assumed each of the postures in the study in a random order, but because of the learning necessary, each subject used only one of the following text entry methods: *mini-Qwerty*, *soft-Qwerty*, *Graffiti*, *EdgeWrite*, or *ShapeWriter*. As the difference in performance of these methods is already well known, the research question concerns the posture effect, not the relative performance of the methods, *per se*.

1. What are the between-subjects and/or within-subjects factors for the above study?
2. How many levels does each factor have? What are those levels?
3. What would be a reasonable response variable? *Record your answers before proceeding.*

Think in terms of the variance of the response variable. The use of five separate text entry methods introduces a lot of variance into the data. ShapeWriter, for example, can reach 40 wpm or higher while Graffiti may only reach 20 wpm. What to do about this? On one hand, the researcher could decide to simply use *one* text entry method for the whole experiment. But this reduces the generalizability of the results: the effects of posture, then, could be said to only apply to the method tested, and not to mobile text entry in general. So it would be nice to use a selection of methods to increase generalizability.

To account for this unwanted variance, we must nest *Subject* within *Method*, written (counter-intuitively for computer programmers) as *Subject[Method]*. (In general, we write factor *A* is nested in factor *B* as *A[B]*. *A* is said to be the nested factor, and *B* is said to be the nesting factor.)

Mathematically, nesting causes the levels of the nested factor (A) to be modeled *separately* for each level of the nesting factor (B), rather than *across* levels of the nesting factor. In the example above, each subject will be grouped based on whatever *Method* they used. So the subjects that use ShapeWriter will be modeled together, but separately from the subjects that use Graffiti, and so on.

The explanation of nesting from the SAS JMP 7.0 help documentation reads as follows, with letters *A* and *B* from our *A[B]* notation added for clarity:

> JMP uses the terms inside the parentheses (B) as grouping terms for each group (A). For each combination of levels of the nesting terms (B), JMP constructs the effect on the outside of the parentheses (A).

The intuition for when to nest can be thought of in terms of the meaning of the nested effect. If the levels of a factor (*e.g.*, *Subject*) differ in their experimental "meaning" or "interpretation" according to the levels of another factor (*e.g.*, *Method*), then the first should be nested in the second (*e.g.*, *Subject[Method]*). In the example above, although each subject is, in some sense, equivalent to each other subject, the activity that a subject does in the experiment is entirely different based on the text entry method they are assigned. Nesting *Subject[Method]* ensures that each subject is modeled according to the text entry method that they used, allowing us to better isolate the effects of posture. (Note that we must still make *Subject* a random effect; in JMP it will look like "Subject[Method]& Random" in the model effects box.)

Part 2, Agriculture example. Mixed designs are sometimes called *split-plot designs* due to their common use in agriculture. Consider an experiment from Milliken & Johnson (1984) reported in Higgins *et al.* (1990) to test the effects of fertilizer and moisture levels on the dry matter resulting from peat moss. See if you can spot the experimental unit (the "subject"), the between-subjects factor, and the within-subjects factor from Higgins' description:

> In the experiment, four peat pots, with a different level of fertilizer randomly assigned to each, were placed in a tray. Each tray was treated with one of four different moisture levels, the levels being randomly assigned to trays. The response variable was amount of dry matter (Higgins *et al.* 1990, p. 189).

I'll clarify that in this experiment, there were 12 trays, 4 moisture levels, and 4 types of fertilizer.

4. What is the experimental unit in this study? How many of these are there?
5. What is the between-subjects factor? What are its levels?
6. What is the within-subjects factor? What are its levels?
7. What constitutes a single "trial?" Indicate how a trial can be defined in any experiment.

Part 3, Education example. Other examples of nesting occur in educational research, where different sites are used. Consider a fictitious study of high school students' performance on a test. If five different high schools are used, then we will probably want to nest *Student[School]* to reflect that each student came from a different high school. Nesting prevents the pooling of students across schools.

Part 4, Manufacturing example. Wikipedia's entry on *Restricted randomization* has an informative section called *Example of nested data* (http://en.wikipedia.org/wiki/Split-plot_design). Here is its description of a fictitious manufacturing experiment that has *two* nested effects. Note at the end where it also mentions the use of *random effects*, with which you should be already familiar.

> Consider a batch process that uses 7 monitor wafers in each run. The plan further calls for measuring a response variable on each wafer at each of 9 sites. The organization of the sampling plan has a hierarchical or nested structure: the batch run is the

> topmost level, the second level is an individual wafer, and the third level is the site on the wafer.
>
> The total amount of data generated per batch run will be 7 * 9 = 63 observations. One approach to analyzing these data would be to compute the mean of all these points as well as their standard deviation and use those results as responses for each run.
>
> Analyzing the data as suggested above is not absolutely incorrect, but doing so loses information that one might otherwise obtain. For example, site 1 on wafer 1 is physically different from site 1 on wafer 2 or on any other wafer. The same is true for any of the sites on any of the wafers. Similarly, wafer 1 in run 1 is physically different from wafer 1 in run 2, and so on. To describe this situation one says that sites are nested within wafers while wafers are nested within runs.
>
> As a consequence of this nesting, there are restrictions on the randomization that can occur in the experiment. This kind of restricted randomization always produces nested sources of variation.
>
> The objective of an experiment with this type of sampling plan is generally to reduce the variability due to sites on the wafers and wafers within runs (or batches) in the process. The sites on the wafers and the wafers within a batch become sources of unwanted variation and an investigator seeks to make the system robust to those sources.
>
> Because the wafers and the sites represent unwanted sources of variation and because one of the objectives is to reduce the process sensitivity to these sources of variation, treating wafers and sites as random effects in the analysis of the data is a reasonable approach. In other words, nested variation is often another way of saying nested random effects or nested sources of noise. (Wikipedia, accessed June 1, 2010)

Part 5, Study of designers. Open the JMP sample data file *designers.jmp*. This data table is equivalent to JMP sample file *animals.jmp*, recast to serve as an example relevant to human-computer interaction. Examine the table notes and the columns carefully. Then answer:

8. How many designers took part in this study?
9. What are the between- and within-subjects factors in this study? What and how many are their levels?
10. What is the response variable in this study?
11. Give a plausible explanation for what this study may have been about.
12. Now perform the proper analysis on this data. You should use Analyze > Fit Model and use a mixed-effects model analysis of variance. (Hint: nesting is required.) Reproduce your model effects, your response variable, and your results table. (You can either retype it or clip out the

relevant window regions from JMP. If you are on a Windows machine, you can use Start Menu > Accessories > Snipping Tool.)

13. Which factor had to be nested? In which other factor? Why?
14. Look at the levels of *Designer* in the data table. Why does it make intuitive sense that nesting was necessary?
15. Based on the premise for this study and your findings, what can you conclude?

Part 6, Peat moss exercise. Now open *higgins.jmp*, which corresponds to the data from the peat moss experiment in part 2, above. Based on the description in part 2, and now in examining the data table, answer the following:

16. Is this a wide-format or long-format table? How do you know?
17. What are the factors and levels? (Do they match your answers given in questions 5 and 6, above? If not, why not?)
18. Recall from module 4, part 7 that you analyzed a two-way mixed factorial design with a traditional repeated measures ANOVA. Apply that same procedure here to analyze *higgins.jmp*. (Remember to consider Sphericity; if there is a violation, report the test and use an adjustment.) Write up the findings from this experiment as you would for a publication. First introduce the experiment in your own words.
19. Note that as in module 4, part 7, explicit nesting was not an issue; it is also not (yet) relevant to this analysis. Speculate as to why not. (Hint: the data itself warrants nesting, but a traditional repeated measures ANOVA makes explicit nesting unnecessary because of the inherent structure of the wide-format data table. Why?)

After you have analyzed *higgins.jmp* using a traditional repeated measures ANOVA, transform the *higgins.jmp* data table into long-format. You can use Tables > Stack or Microsoft Excel's Paste Special > Transpose command. When you are done, be sure to compute means and standard deviations over both tables to verify that your wide-format to long-format transformation was without error.

20. How many columns does your long-format table have? What are their names?
21. How many rows?
22. Do means and standard deviations agree between your wide-format and long-format tables?

Next, conduct a mixed-effects model analysis of variance on the long-format table using Analyze > Fit Model.

23. What are your model effects?
24. What is your model's response?
25. What is your random effect?
26. You should be nesting one factor within another: which are you nesting and why? Give the rationale. (Hint: remember that *Subject* is indeed a factor, just not usually one of interest.)

After you have run your model, inspect the results in light of what you obtained using your traditional repeated measures ANOVA on *higgins.jmp*.

27. Compare the results between your repeated measures ANOVA and your mixed-effects model analysis of variance. Do the results agree? Compare and contrast your outcomes.

Next, we are no longer going to use a mixed-effects model to analyze our long-format table, but instead use a traditional ANOVA but still on the long-format table. To do this, simply remove the "&Random" attribute from your random effect that you had in the previous model. (Personality should still be Standard Least Squares and Emphasis should still be Effect Leverage.)

28. Reproduce your model effects here.
29. Report the results of this analysis.
30. How do the results for each factor and interaction compare to the mixed-effects model results? Do any of the results from question 29 seem incorrect? Which? How do you know?

Finally, remove the nesting such that your formerly nested effect is no longer nested. (Symbolically, if you had *A[B]* then just use *A* in the model for that term. You will still have *B* separately.)

31. Reproduce your model effects here.
32. What happens when you run the model? (Hint: look at the top of the report for Singularity Details.) Why do you think this happens?
33. In general, how can a singularity problem be remedied?

Part 7, Create your own experiment. Create a mock human-computer interaction experiment where nesting is warranted. Try to make the experiment somewhat plausible and interesting.

34. Write your experiment description.
35. Identify any between-subjects factors, within-subjects factors, random effects, and their levels.
36. Which factor(s) should be nested within which other factor(s)? Why?

Part 8, Nesting in mixed factorial designs. JMP provides an online sample to illustrate the equivalence of wide-format and long-format repeated measures analyses. The sample is #30584: Analyzing Repeated Measures in JMP Software, and is available at http://www.jmp.com/support/notes/30/584.html. It is also provided in PDF form as part of this module as `Sample 30584.pdf`. Open the sample and read it, paying special attention to the wide-format and long-format tables that are figures in the sample.

37. Which figure represents the wide-format data? Which figure represents the long-format data? For each, how do you know?
38. What are the fixed effects and their levels? Which are between-subjects effects and which are within-subjects effects?
39. What is the response measure? How is it encoded in the wide-format and long-format tables?
40. In either table, which column identifies the experimental unit?
41. In the analysis for the long-format table, what is the random effect?
42. In the analysis for the long-format table, which effect is nested? In which effect? Why?

08. Longitudinal data

macbtwn.jmp, macwthn.jmp

Whether it is throwing a javelin or saying the alphabet backwards, most types of human performance improve with practice. The ability for people to enter text by typing on a QWERTY keyboard, for example, follows a well-known "learning curve," where performance, measured as words per minute (WPM), improves steadily over time. Studies where practice and learning play a crucial role are often best conducted *longitudinally*, meaning "over time." A typical practice is to have subjects attend sessions that are no closer than 2 hours apart, no farther than 48 hours apart, and without more than 2 on the same day. Such studies may have within- or between-subjects factors of interest, but all will have a factor named *Session* or *Day* or *Time* that codifies the *x*-axis in Figure 1, reproduced from MacKenzie & Zhang (1999).

Finding "the elusive crossover point" is important when practice and learning play a central role. This point tells us the time it takes for one technique's performance to match that of another, often more familiar, technique. Also, longitudinal data of the kind shown in Figure 1 allows us to fit learning curves modeled by the Power Law of Learning:

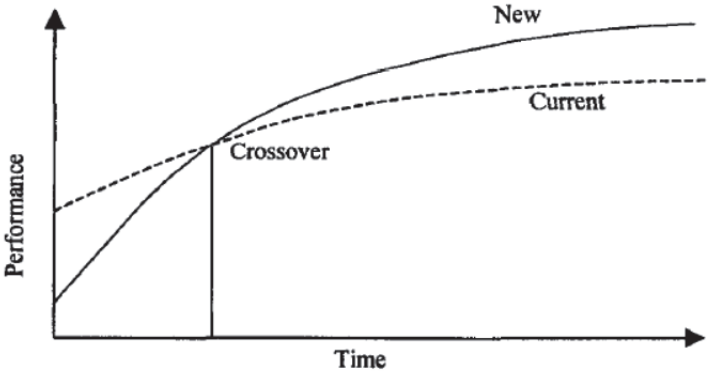

Figure 1. The elusive crossover point

$$\text{Performance} = a \times \text{Session}^{b} ,$$

or, more generally,

$$Y = aX^{b} .$$

Here, the regression coefficient a is the y-intercept and represents initial performance. The other regression coefficient, b, is usually a value between 0.0 and 0.5, and reflects how steep the learning curve is. The response Y is the y-axis measure (*e.g.*, WPM), and the variable X is the temporal variable on the x-axis (*e.g.*, *Session*). As part of this module, we will see how to fit a learning curve to longitudinal data using the Solver in Microsoft Excel.

From a modeling standpoint, the temporal variable is always a within-subjects variable, as every subject is measured at every time point along the x-axis. Although the measures do not happen at the exact same times, relative to each subject, a given *Session* number is the same from one person to the next.

Questions of central interest are:

- How is longitudinal data analyzed statistically?
- Is it possible to statistically operationalize a notion of "when learning stops?" If so, how?
- How is a learning curve of the form $Y = aX^{b}$ fitted?
- How can future performance be estimated based on a learning curve?

Part 1, Repeated measures ANOVA. Consider an example inspired by an experiment conducted by MacKenzie & Zhang (1999). The authors invented an optimized stylus keyboard layout called OPTI, and compared it to a stylus keyboard based on the familiar QWERTY layout. A longitudinal study was used. The main performance measure was words per minute (WPM), although error rates were also recorded (but will be ignored for our purposes). Open *macbtwn.jmp* and answer the following questions.

1. Is this data table in wide-format or long-format? How do you know?
2. How many subjects took part in this study?
3. What are any factors and their levels? For each factor, say whether it is between- or within-subjects.
4. What is the formal multiplicative notation for describing this experiment?
5. In any text entry experiment, we would assume subjects each entered multiple text phrases. Why, then, do you think the data table just shows one measure per subject per session? How was this achieved? What are the implications of making this choice in terms of both *power* and *variance*?

Now we will analyze the data. The procedure is identical to that for module 4, part 7. Follow that procedure for this longitudinal data and then answer the following questions.

6. When using JMP's Manova platform, what name did you give to the repeated factor? What was the default name that JMP placed in the text box where you could enter a name of your own?
7. Using the graph of means over responses created by the JMP Manova platform, eyeball the crossover point (*see* Figure 1, above). In what session does the crossover point occur?
8. On a Windows machine, use Start Menu > Accessories > Snipping Tool (or PrtSc and Paint) to produce the graph here, including its legend. Feel free first to adjust the graph dimensions by dragging its borders.
9. What is each method's performance in the first session?
10. What is each method's average performance across all sessions?
11. What was each method's peak performance averaged over subjects per session? Did each peak occur in the last session?
12. State and interpret the finding for the *Method* factor. What does this result mean?
13. State and interpret the finding for the *Session* factor. What does this result mean?
14. State and interpret the finding for the *Method*Session* interaction. What does this result mean?

Part 2, Mixed-effects model. Now transform this data into a long-format table for the purposes of performing a mixed-effects model analysis of variance. To do this, go to Tables > Stack, place the 20 response columns in Stack Columns, enter "WPM" for Stacked Data Column, and "Session" for Source Label Column. Click OK. Note that the Session column will be a character data type (text), but you want it to be numeric, so right-click on the column, choose Column Info, and for Data Type, set Numeric. (For now, leave its modeling type as nominal.) You may wish to save this data table as *macbtwn.long.jmp*. Then answer the following questions.

15. How many rows *should* there be in the long-format table? How many rows do you have?
16. Prepare for a mixed-effects model analysis of variance by going to Analyze > Fit Model. What is your Y response? What are your model effects? Is nesting appropriate? If so, what is nested in what? If not, why not?

17. What is the fit of this mixed-effects model? Does it fit well? (Hint: look for the R^2 of the model.) You can also judge the model's fit by the -2 LogLikelihood value in the REML Variance Component Estimates section. What is this value? Smaller numbers for -2 LogLikelihood are better model fits.
18. What are the results of this analysis of variance? Do the results agree with those found in your repeated measures ANOVA (questions 12-14)?
19. Now change the modeling type of *Session* to ordinal. What are the results of this analysis? How do they compare to the results from when it was nominal? Which model fits better according to R^2? According to -2 LogLikelihood? How do the degrees-of-freedom compare?
20. Now change the modeling type of *Session* to be continuous. What are the results of this analysis? How do they compare to the results from when it was nominal? ordinal? Which model fits better according to R^2? According to -2 LogLikelihood? How do the degrees-of-freedom compare?
21. Examine the plot of residuals. This plot should ideally seem as randomly scattered as possible; otherwise, a pattern is present in the data that is not being accounted for by the model. How do the residuals compare among the models with nominal, ordinal, and continuous modeling types for *Session*?
22. Which of the three modeling types seems most appropriate based on the above? Why? Which of the three modeling types *conceptually* seems most appropriate? Why?
23. Based on your exploration of *Session* modeling types, what *must* have been the implicit modeling type for *Session* in the traditional repeated measures ANOVA (questions 12-14)? Does your answer impact your final decision for the modeling type of *Session*? If so, how? If not, why not?

Part 3, Learning curves. We now will fit learning curves to this data. It is typical to fit curves to the average performance across subjects per method for each session. Therefore, we need to create an "average table" over subjects per session by method. Use Tables > Summary to create this from your long-format data table. After you have done so, be sure the table is sorted first by Method, then by Session number, since each method was used separately in each session. (If it is not sorted properly by session number, you may have neglected to set the data type of the Session column to numeric before answering question 15. Remedy this before proceeding.) Then answer:

24. What are the columns in your summary table?
25. How many rows does your summary table have?
26. What is the response column named?

Now copy your summary data table into Microsoft Excel and save it as *macbtwn.xlsx*. (You can delete the N Rows column that JMP created.) Before proceeding, create a graph of WPM over sessions for both OPTI and QWERTY. Your graph should look similar to that created by JMP's Manova platform for question 8. However, do not connect the data points in each series.

27. Reproduce your Excel graph here, including its legend and marked axes.

Now we are going to use Microsoft Excel's Solver feature to fit a regression power curve for the Power Law of Learning. After the WPM column, add columns with the names *a*, *b*, Model WPM, RMSD, and Sum. As seeds, put the value 1.0 in row 2 under *a* and 0.1 in row 2 under *b*. These two columns, your regression coefficients, will only have one value in each, not a column full of numbers. Of course, you will need separate *a* and *b* values for the OPTI and QWERTY models, so halfway down, where

QWERTY begins, put another 1.0 and 0.1 for QWERTY's *a* and *b*. (This should be in row 22 of the Microsoft Excel table.)

The Model WPM column contains model predictions based on the Power Law of Learning. At the start of this module you learned the equation for the Power Law, so you should be able to create the model predictions in this column. Fill up the column with these temporary seed predictions.

28. Do that now, and describe what you did here. Include the text of your Microsoft Excel formula.

The RMSD column stands for "root mean square deviation."[3] Each cell in this column contains a value for the difference between the observed response and the predicted response. Compute this column as the square root of the squared difference between observed WPM and model WPM. In symbols:

$$RMSD = \sqrt{(Y_{\text{observed}} - Y_{\text{model}})^2}.$$

We need a way of knowing *overall* how far off our model predictions are from the observed data points. Do this by summing up RMSD in the Sum column, one sum for OPTI in row 2, and one sum for QWERTY in row 22.

29. What are the initial Sum values for OPTI and QWERTY?

Our goal now will be to *minimize Sum separately for each of the models*. Once we've done this, we'll have the best-fitting model possible. This requires an iterative improvement algorithm. Microsoft Excel provides this in its Solver feature, which for its first use must be installed. Go to the Office Button > Excel Options > Add-Ins and click Go. Check the Solver Add-In box, and click OK. Install the add-in when prompted.

30. The Solver can change the values of some cell(s) while minimizing or maximizing the value of another cell. In these terms, what should the solver do for fitting our power law curve for each of OPTI and QWERTY? *Record your answer before proceeding*.

On the Data tab, in the Analysis section, you should now see the Solver. Click it. For fitting the OPTI curve, we want to set the target cell to the OPTI Sum in row 2. We want to minimize this by changing the values of OPTI's regression coefficients *a* and *b* in row 2. Then click Solve. Do this also for QWERTY's curve using the values of Sum, *a*, and *b* in row 22.

31. What are the final Sum, *a*, and *b* values for OPTI? What power law equation does this result in?
32. What are the final Sum, *a*, and *b* values for QWERTY? What power law equation does this result in?
33. Using your model predictions, plot the power curves on the graph you created previously. Do not plot points, just the power curves, so that we can easily see the original data points and the fitted curve over them. Display your graph here, including its legend.

[3] See http://en.wikipedia.org/wiki/Root_mean_square_deviation for more information.

We also want to know the goodness-of-fit of our power curves from R^2 values. Create a new column called R2 after Sum. For OPTI, in row 2, use the square of the CORREL or PEARSON Excel functions over the observed WPM and model WPM. For QWERTY, in row 22, do the same. (You can instead use the RSQ Excel function directly, but then do not square its results!)

34. What are the R^2 values for the OPTI and QWERTY models?
35. Speculate as to why one of the models fits better than the other.
36. This data was estimated from the graph in MacKenzie & Zhang's (1999) original paper. In other words, it is "real human data." Does it surprise you that the Power Law of Learning models this data so well? Why or why not?

Models for data are useful not just for description, but also for *prediction*. Extend your predicted WPM values, and the curve you plotted, from session 20 to session 40 for each method.

37. What does the graph show as the predicted speed for OPTI and for QWERTY in the 40th session? Display your extended graph here, including its legend.

Part 4, Covariance structures. In part 2, we analyzed the *macbtwn.jmp* data using a mixed-effects model analysis of variance. An important but hidden aspect of using mixed-effects models in JMP is that of specifying a *covariance structure*, which JMP does not expose but SPSS does. The covariance structure for repeated measures data indicates how measures "vary and vary together." For example, measures taken across sessions may vary differently with respect to each other than measures using different input methods in the same session. And yet, because the same subject was measured more than once, both are forms of repeated measures data.

As mentioned, SPSS allows us to pick the covariance structure when using mixed-effects models. Although there are numerous possible structures, the most common ones are Autoregressive, Compound Symmetry, Diagonal, Scaled Identity, and Unstructured. Although the underlying mathematical details are beyond the current scope, the intuition behind each one is explained below.

- **Autoregressive Order 1 (AR1).** Successive measures of a subject are more strongly correlated than distant measures. This will often be the case with measures of practice over time, *i.e.*, measurements at times 2 and 3 are more strongly correlated than measurements at times 1 and 3.
- **Compound Symmetry (CS).** This is the unchangeable default in JMP. The relationship between levels of a repeated factor all have constant variance and constant covariance; that is, measures n and m covary the same as measures n and k for $n \neq m \neq k$.
- **Diagonal (DIAG).** This is the changeable default in SPSS. Each level of the repeated factor has its own variance, but has zero covariance with every other level.
- **Scaled Identity (ID).** Each level of a repeated factor has unity (1) variance and zero (0) covariance with every other level.
- **Unstructured (UN).** The relationship between all $n \times m$ levels of the repeated factor is unstructured; that is, all variances and covariances are independent.

As stated, the unchangeable default in JMP is Compound Symmetry, while the settable default in SPSS is Diagonal. To find out which is most appropriate for your data using SPSS, run each model and look at the

Information Criteria output for each. Select the model that yields the *lowest* value for either the Akaike's Information Criterion (AIC) or Schwarz's Bayesian Criterion (BIC) values.

Paste your data from your long-format version of *macbtwn.jmp* into SPSS. After setting variable names, values, and types, perform a mixed-effects model analysis of variance. (Referring to questions 17-23, set the modeling type of *Session* to whatever you decided was most appropriate.) Go to Analyze > Mixed Models > Linear. On the first dialog, put *Subject* in Subjects and *Session* in Repeated. For Repeated Covariance Type, you will try the five covariance structures listed above. For now, start with Compound Symmetry to agree with JMP. Click Continue. Place *WPM* in Dependent Variable and *Method* and *Session* in Factor(s). Click Fixed and add *Method*, *Session*, and *Method*Session* in the Model. Click Continue. Click EM Means and transfer all factors to Display Means for. Click Continue. Click OK.

38. What modeling type did you choose for *Session*? Why?
39. What are the AIC and BIC for Compound Symmetry? What are they for the other four covariance structures listed above? Make a table. Include -2 LogLikelihood for comparisons. Which covariance structure seems best suited to these data?
40. How do your results using the best covariance structure compare to the results from using JMP (*see* questions 18-20)? Do the conclusions change from what you concluded previously based on using the best covariance structure?

Part 5, Within-subjects. The original experiment from MacKenzie & Zhang (1999) had within-subjects data for *Method*, as each subject used either OPTI or QWERTY first in each session, followed by the other. Although MacKenzie & Zhang deterministically alternated the order of *Method* for each subject, let us pretend they chose the order randomly (*e.g.*, with a coin toss).

Open *macwthn.jmp*, which shows fictitious data similar to, but not the same as, that from the between-subjects data. Examine the data table and answer the following questions.

41. What are the columns in this data table?
42. How many rows are in the table? What does each row represent?
43. How many distinct subjects took part in this study? How many rows are there per subject? Explain how this number arises.
44. Is this data table in wide-format or long-format? How can you tell?
45. What are the factors and their levels? For each factor, say whether it is between-subjects or within-subjects.
46. Are there any covariates? If so, what are they? Are they between- or within-subjects?
47. Explain the meaning of the *Order* column. It indicates the order of what, exactly?
48. Explain the meaning of the *Trial* column. What does this tell us about how the experiment was run?
49. What is the formal multiplicative notation for describing this experiment?

Now, prepare to conduct an analysis of variance using a mixed-effects model. *Before* you do so, answer the following questions.

50. What should be the Y response for the model?

51. Ignoring *Handedness* and *Order* for now, what should be the model effects? (Hint: think carefully about whether any effects should be nested, or even doubly nested.)
52. Did you encounter the following issue? If a model contains *A[B]*, then it is not allowed to contain the interaction *B*A[B]*. Why does such an interaction make no sense?

Only after recording your answers to the above questions should you proceed to run your analysis. (Be patient; complex models take a while to execute!)

53. What was the final model you ran? What were any nested and/or random effects?
54. Now open the saved analysis called "WPM." Examine the model effects. How do they differ from the model you used?
55. Why is *Trial[Session, Method]* in the model? What does this effect mean? (Hint: you may want to look in the JMP help documentation for nested effects.)
56. Why are there no interactions for *Trial* in the model?
57. Judging by R^2, how well does this model fit the data?
58. Report and interpret the results of this analysis.
59. How do these findings compare to those for *macbtwn.jmp* in questions 12-14?

Part 6, Covariates. Now let us consider the covariate of handedness. It may be that handedness makes no difference, but since we bothered to record it for each participant, we might as well check. There is a reasonable hypothesis that handedness matters for stylus keyboard use, so it is a justifiable exploration. Conceptually, we are not interested in *Handedness*Session* or *Handedness*Trial[Session, Method]*. It is unclear what such findings would really mean. We are, however, interested in a main effect of handedness or in an interaction between handedness and methods. To examine these, run the saved model "Handedness."

60. How many subjects were right-handed? How many were left-handed?
61. What is a rationale for any nested and/or random effects?
62. What would it mean to have a *Handedness* main effect? Give an example that would give rise to such an effect.
63. What would it mean to have a *Handedness*Method* interaction? Give an example that would give rise to such an interaction.
64. Write up the conclusions regarding the effects of handedness on each method's performance.

Part 7, Order effects. We also care about whether or not the coin toss for counterbalancing method presentation in a given session "worked." The *Order* column allows us to check for this. Conceptually, we are interested in a main effect of *Order* and an *Order*Method* interaction. Run the saved model "Order" before answering the following questions.

65. Over the entire experiment, how many times did OPTI go first? How many times did QWERTY go first? Describe in step-by-step detail how you arrived at these answers.
66. Conceptually, what would an *Order* main effect mean? Give an example that would give rise to such an effect.
67. Conceptually, what would an *Order*Method* interaction mean? Give an example that would give rise to such an interaction.

68. What are your conclusions regarding order effects? Did the coin-toss counterbalancing work? How do you know?
69. What could you argue to save the integrity of your experiment if you ever found an *Order* main effect but not an *Order*Method* interaction? How can this be okay?

Part 8, Repeated measures ANOVA. For questions 53-59, we used a mixed-effects model analysis of variance on *macwthn.jmp*. For comparisons, let's also do this analysis (once again, ignoring *Handedness* and *Order*, having examined those already) using a traditional repeated measures ANOVA. Recall that JMP can only handle up to two repeated factors, but for *macwthn.jmp*, we have three: *Session*, *Method*, and *Trial*. One approach would be to "average down" over trials to create one measure per subject per session per method, as we had in *macbtwn.jmp*. But this costs us statistical power. If we wish to retain *Trial*, we'll need to use SPSS, which as you will recall, allows for an arbitrary number of repeated factors. (See module 5, part 7.) And we'll need a (very!) wide table.

Using Tables > Split, transform *macwthn.jmp* into a wide-format table. (Hint: you will need to split three times, once for each within-subjects factor.) Run SPSS and paste this data into it, naming your data table *macwthn.wide.sav*. You don't want to manually name the 400 data columns you'll have in SPSS; instead, first paste the column names from JMP into Microsoft Excel. Then use the Excel command SUBSTITUTE to replace any spaces with underscores, as SPSS does not allow spaces in its column names. Also, SPSS won't allow numbers to be the lead characters in column names, so if you split such that the session number is the leading character, use CONCATENATE to prepend an "s" to the front of each column name, *e.g.*, s1_OPTI_5 would be session 1, OPTI keyboard, trial 5. Once your column names are prepared, copy them with CTRL+C in Microsoft Excel; in SPSS, drag your mouse down the Name column in the Variable View, and paste the names all at once using CTRL+V.

Once the SPSS data table is ready, conduct a repeated measures ANOVA by going to Analyze > General Linear Model > Repeated Measures. Fill out the dialog box and execute the analysis. Remember that unlike with our mixed-effects models, explicit nesting is not applicable here, so *Trial[Session, Method]* is not modeled, but instead *Trial* is an unnested model term just like *Session* and *Method*.

70. What are the results? Write up a publishable paragraph of findings similar to the one you crafted for question 58.
71. How do these results compare to those from your mixed-effects model analysis of variance (*see* question 58)?

Part 9, Learning curves. Referring to the work you did in part 3 graphing performance over sessions and fitting learning curves, return to JMP to prepare your long-format *macwthn.jmp* data table for the same treatment.

72. Describe what procedure you followed to prepare your data for fitting learning curves.
73. Paste your summary data into Microsoft Excel and follow the same procedure as in part 3 to plot the *macwthn.jmp* data, fit a Power Law of Learning curve, compute R^2 values, and predict performance to 40 sessions. Save this data table as *macwthn.xlsx*. Here, give the power law equations for OPTI and QWERTY, their R^2 fits, and their predicted performance in the 40th session.
74. Place your finished graph with performance extended to 40 sessions below.

Part 10, Learning stabilization. It is often important to know when learning has stabilized during a longitudinal study. In casual terms, we might say that the learning curve has "flattened out." Of course, humans will keep improving a skill for a very long time, even if that improvement becomes almost imperceptible. Therefore, we cannot expect a learning curve to truly *flatten*. But we might like, for example, to compare sessions only *after* learning has stabilized by some quantifiable definition. By doing so, we can find a principled way to remove sessions where learning was still prominent, retaining sessions for analysis in which learning was stable. To avoid bias, the method of determining learning stabilization should be determined *before* analyzing the data.

Multiple comparisons. One approach to determining when learning has stabilized is to use the Tukey-Kramer HSD ("honestly significant difference") test for multiple comparisons across sessions comparing each session to the one before it. At least three criteria for determining "stabilized" are possible from such a test. From least to most conservative (modulo chosen thresholds), the point of stabilization would be:

- *Criterion 1.* The first point at which a predetermined number N of consecutive sessions cease to be significantly different. N may be just two, or more, or a percentage (*e.g.*, 20% of conducted sessions).
- *Criterion 2.* The first point at which the number of consecutive sessions that are not significantly different never decreases (ignoring the end boundary), perhaps with a minimum threshold (*e.g.*, of $N = 3$).
- *Criterion 3.* The first point at which all remaining consecutive sessions up through the last session are not significantly different.

We will use your long-format version of *macbtwn.jmp*, although our procedure is not specific to it. (You may have saved it as *macbtwn.long.jmp*.) After you have this data in long-format, go to Analyze > Fit Y by X and place *WPM* as your Y, Response, *Session* as your X, Factor, and *Method* in the By box. Click OK. You now have plotted the data points separately for each method over sessions. Use the red-arrow menu on the OPTI report and choose Compare Means > All Pairs, Tukey HSD. (Unlike the Student's *t*-test, the Tukey test will adjust alpha (α) for the number of comparisons made.) After executing Tukey, use the newly created red-arrow menu to uncheck everything except the Connecting Letters Report. Then right-click anywhere inside the report and choose Make into Data Table. On the data table, right-click the Level column and choose Column Info. Set the Data Type to Numeric. Then sort the table using Tables > Sort by descending Level order. Repeat this procedure for the QWERTY results. Then answer the following questions.

75. Carefully study the connecting letters report from the Tukey-Kramer HSD test. Describe what this report shows. What is in the leftmost column? What do you notice about it? What is the rightmost column? How is the table sorted? Why did you have to make your own tables sorted by Level?
76. Using the first of the three criterion to determining stabilization above, and using $N = 4$ (or 20% of conducted sessions), at which session would we say learning stabilized for OPTI? for QWERTY?
77. Using the second criterion and a threshold of $N = 3$, at which session would we say learning stabilized for OPTI? for QWERTY? (Hint: make an extra column in your data tables and in it, place the number of consecutive sessions that are not significantly different; a value of "1"

indicates that a session is only not different from itself. Then find the first number ≥*N* after which the numbers never decrease, ignoring the inevitable decrease at the end of sessions.)

78. Using the third criterion, at which session would we say learning stabilized for OPTI? for QWERTY?
79. Return to your long-format data table of 200 rows and, using your answer for each criterion from questions 76-78, exclude all sessions for which learning was not yet stabilized. (You may wish to use Rows > Row Selection > Select Where (CTRL+SHIFT+W) and then Rows > Exclude (CTRL+E).) For each criterion, how many sessions are considered "stable?" How many total rows are thereby excluded?
80. Using each criterion to exclude the proper rows, run a mixed-effects model analysis of variance on *macbtwn.jmp* to compare the methods' stable performance. (Hint: because you are now unbalanced, you must make *Session* continuous and you will see fractional degrees-of-freedom.) Report the model fit (R^2 and -2 LogLikelihood) and your F-test results for each of the three criterion. You may wish to make one or more tables to convey your results. Then reflect on any trends you see in your results based on the conservativeness of the stabilization criterion.

Successive linear regression. Another approach to determining learning stabilization is to first create a summary table of session means (per method), and then fit regression lines through the session means, removing points from the early sessions successively until the regression is not significantly different from horizontal. When a regression line is fitted in JMP, an ANOVA result is produced, which indicates whether the regression fit is significantly better than a horizontal line set at the mean value of the data.

Compute a summary of *macbtwn.jmp* containing session means for each method. After doing so, turn *Session* into a continuous variable for regression purposes. Use Analyze > Fit Y by X with *WPM* in Y, Response, *Session* in X, Factor, and *Method* in By. On each graph, use the red-arrow menu to select Fit Line. Examine the Analysis of Variance table and the *p*-value there. As long as it remains $p < .05$, exclude points, each in turn, from the starting sessions using Rows > Exclude (CTRL+E). After each exclusion, repeat this process until the linear fit is not significantly different from horizontal. This can be used as an indication for where learning stabilized.

81. Having followed this process above for OPTI and QWERTY, indicate at which session each method stabilizes under this scheme. Give the nonsignificant regression ANOVA results for each method.
82. Repeat your mixed-model analysis of variance on the main long-format data table of 200 rows after excluding sessions for each method according to your answer to question 81. How many rows are excluded for each method? What are the results of your analysis?
83. How does the successive linear regression approach compare to the answers obtained for each of the three criteria based on multiple comparisons? Is it more or less conservative?
84. Of the four approaches, which do you think is conceptually most sound? least sound? Why?
85. Of the four approaches, which do you think is easiest to employ? hardest to employ? Why?

09. Nonparametric tests

nonparam.xlsx

Until now, our modules have mostly used parametric tests. Although deeply technical definitions exist for "parametric" and "nonparametric," intuitively, they can be thought of as, in the former, assuming an underlying normal distribution, and in the latter, being free of assumptions about distribution. By "underlying normal distribution," we mean that the measure in question taken over the whole population would exhibit a Gaussian curve. While this is true for many measures (*e.g.*, height, weight, age, certain abilities), it is not true for all (*e.g.*, household income, prison sentence durations). Some measures follow non-normal distributions; common alternatives are *lognormal*, *exponential*, *binomial* and *Poisson*. Still other data exhibits no discernable distribution at all. You encountered these concepts briefly in module 6.

You ran a nonparametric test in module 4, part 3, when you learned how to use the JMP and SPSS statistical tools. Some common analyses are available in only one or the other of these tools. Still others are not widely available at all.

An important feature of many nonparametric tests is that they use *ranks* on the data being analyzed, rather than the raw data points themselves. By ranking data, and using average ranks (or "midranks") for ties, distributions that were present in the data are destroyed.

In module 6, you learned how to examine the distributions of data and test for normality. Because we cannot know the distribution of the whole population, we must use our sample as an approximation. Thus, if our sample's measure is normally distributed, we can assume that the population's measure (also known as the "population parameter") is also normally distributed. This is but one of the assumptions required for parametric *t*- and *F*-tests.

Part 1, ANOVA assumptions. There are classically *three* assumptions required for one-way ANOVA. Discover these and list them below. Beside each item, give *your own* definition that indicates you grasp the *intuition* behind the assumption. Then say how you would check each assumption using a tool like Microsoft Excel, JMP, or SPSS.

1.

2.

3.

Part 2, Parallel tests. Study the table on the next page, which lists the parametric tests based on number of factors and levels, and whether or not those factors are between- or within-subjects. By exploring the features and help available in JMP or SPSS or on the Web, find a way to enter the proper tests for the semiparametric/nonparametric column on the right. The tests to properly place, mixed with tests you should *not* place, are:

Student's *t*, paired-samples *t*, ANOVA, repeated measures ANOVA, MANOVA, Tukey-Kramer HSD, Chi-Square (χ^2), Fisher's exact, Mann-Whitney *U*, Kruskal-Wallis, Wilcoxon signed-rank, Friedman, Pearson correlation, Spearman correlation, R^2, Kolmogorov-Smirnov *D*, Shapiro-Wilk *W*, Aligned Rank Transform (ART), General Linear Models (GLMs), Generalized Linear Models (GZLMs), Generalized Linear Mixed Models (GLMMs), Generalized Estimating Equations (GEEs).

No. Factors	**No. Levels**	**between-subjects or within-subjects**	**Parametric Test**	**Semiparametric or Nonparametric Equivalent**
1	2	between	independent-samples *t*	4)
1	2	within	paired-samples *t*	5)
1	3+	between	one-way ANOVA	6)
1	3+	within	repeated measures ANOVA	7)
2+	2+ ea.	between *only* (cannot do within)	*n*-way ANOVA	8)
2+	2+ ea.	within (can also do between)	repeated measures ANOVA	9) or, 10) or, 11)

STOP! Complete this table before proceeding to the next page.

No. Factors	No. Levels	between-subjects or within-subjects	Parametric Test	Semiparametric or Nonparametric Equivalent
1	2	between	independent-samples *t*	4) Mann-Whitney *U*
1	2	within	paired-samples *t*	5) Wilcoxon signed-rank
1	3+	between	one-way ANOVA	6) Kruskal-Wallis
1	3+	within	repeated measures ANOVA	7) Friedman
2+	2+ ea.	between *only* (cannot do within)	*n*-way ANOVA	8) GZLMs
2+	2+ ea.	within (can also do between)	repeated measures ANOVA	9) ART or, 10) GLMMs or, 11) GEEs.

Part 3, Discrepancies. The above table contains the answers to part 2. Examine it closely.

12. Describe and explain any discrepancies from what you had in your table. Where did the confusions arise and why?

Part 4, Touring the tests. In this part, the mechanics for performing each of the above nonparametric tests are given for JMP and SPSS. Afterwards, a collection of generic unlabeled data tables are given to you in *nonparam.xlsx*. Your job will be to examine the tables' structures, perform tests for normality (and report the results), and determine what nonparametric tests, and their parametric equivalents, should be conducted on the data, writing up both results and comparing them. All data tables are given as separate worksheets in *nonparam.xlsx*. It is up to you to create JMP or SPSS data tables as necessary.

• Mann-Whitney *U* test • The Mann-Whitney *U* test is the nonparametric equivalent of the independent-samples *t*-test. It analyzes one between-subjects factor with two levels.

JMP can analyze two independent nonparametric samples, but a Mann-Whitney *U* test is not performed; rather, a Wilcoxon rank sums test is used. (This is not to be confused with a Wilcoxon signed-rank test.) SPSS is therefore preferred for the Mann-Whitney *U* test. However, the results from JMP will agree. In JMP, go to Analyze > Fit Y by X, and fill out the Y, Response and X, Factor fields. On the Oneway Analysis report, use the red-arrow menu to select Nonparametric > Wilcoxon Test. The result of the Wilcoxon rank sums test appears in the 2-Sample Test report, where the *S*, *Z*, and *p*-values are of interest.

In SPSS, go to Analyze > Nonparametric Tests > Legacy Dialogs > 2 Independent Samples. Put the response variable in Test Variable List. Put the factor with two levels in Grouping Variable and click Define Groups to indicate each level. Ensure only Mann-Whitney U is checked and click OK. The results are in the Test Statistics report, where the *U*, *Z*, and Asymp. Sig. (2-tailed) values are of interest.

• Wilcoxon signed-rank test • The Wilcoxon signed-rank test is the nonparametric equivalent of the paired-samples *t*-test. It analyzes one within-subjects factor with two levels.

In JMP, go to Analyze > Matched Pairs. Place the two matched response columns in Y, Paired Responses. Click OK. (A paired-samples *t*-test is shown by default.) On the Matched Pairs report, use the

red-arrow menu to choose Wilcoxon Signed Rank. The Test Statistic is the *Z* value and the two-sided *p*-value is Prob > |z|.

In SPSS, go to Analyze > Nonparametric Tests > Legacy Dialogs > 2 Related Samples. Put a response in each of the variable slots for Pair 1. Be sure only Wilcoxon is checked and click OK. The *Z* value and *p*-value are given.

• Kruskal-Wallis test • The Kruskal-Wallis test is the nonparametric equivalent of a one-way ANOVA. It analyzes one between-subjects factor with two or more levels.

In JMP, go to Analyze > Fit Y by X. Fill out the dialog and click OK. On the Oneway Analysis report, use the red-arrow menu to select Nonparametric > Wilcoxon Test. A Kruskal-Wallis test is given in the 1-way Test report. It is reported as a Chi-Square (χ^2).

In SPSS, go to Analyze > Nonparametric Tests > Legacy Dialogs > K Independent Samples. Place the response in the Test Variable List, and the factor in the Grouping Variable box. Click Define Range to define the minimum and maximum levels of the factor. Ensure that only Kruskal-Wallis H is checked and click OK. The Test Statistics report shows the results, reported as a Chi-Square (χ^2).

• Friedman test • The Friedman test is the nonparametric equivalent of a repeated measures ANOVA on one factor. It analyzes one within-subjects factor with two or more levels.

Unfortunately, JMP does not provide the Friedman test.

In SPSS, go to Analyze > Nonparametric Tests > Legacy Dialogs > K Related Samples. Transfer the response columns into Test Variables. Ensure that Friedman is checked and click OK. In the Test Statistics report, all rows are of interest. The test is reported as a Chi-Square (χ^2).

• Generalized Linear Models (GZLMs) • Before discussing Generalized Linear Models, let's briefly discuss General Linear Models (GLMs). GLMs are what you have used many times already to perform ANOVAs. In fact, you have used SPSS's Analyze > General Linear Model menu to do so. A GLM is a parametric model that assumes a normal distribution in a continuous response variable. In essence, a GLM is the machinery you've come to regard as simply "an ANOVA," whether repeated measures or not.

Generalized Linear Models (GZLMs) advance GLMs by no longer assuming that a continuous response variable must be normally distributed. Instead, the response may be Poisson, exponential, binomial, and so on. GZLMs are useful for Likert data, event-count data, binary data, interval-censored survival data, and many other types of data.

GZLMs require their user to specify a distribution and a link function. Certain distributions and link functions commonly go together. JMP sets the most common link function as the default whenever you pick a distribution. SPSS enables you to simply choose a model type to set both the distribution and link function at once. For example, a GZLM using the Normal distribution and Identity link function is equivalent to a GLM with only between-subjects factors.

A major limitation of GZLMs is that they assume each observation is independent. In other words, GZLMs can only be used for between-subjects factors, as within-subjects factors are correlated within experimental units (*i.e.*, subjects). Also, GZLMs are not "distribution free" like rank-based methods. Rather, GZLMs simply have the power and flexibility to not require a normal distribution. If a rank-based method is desired instead, use the Aligned Rank Transform (ART) procedure for factorial between-subjects data, covered next.

GZLMs can be referred to using certain shorthands depending on their configuration. "Poisson regression" is used on event-count responses and refers to GZLMs with Poisson distribution and Log link

functions. "Binomial logistic regression" is used on binary responses and refers to GZLMs with Binomial distributions and Logit link functions. "Nominal logistic regression" is used on nominal responses and refers to GZLMs with Multinomial distributions and Logit link functions.[4] "Ordinal logistic regression" is used on ordinal responses such as Likert scale data and refers to GZLMs with Multinomial distributions and Cumulative Logit link functions. The SPSS 19 help documentation for Generalized Linear Models has a "Model Types" section that explains common distributions and their link functions.

In JMP, Analyze > Fit Model provides access to GZLMs. Nominal logistic regression and ordinal logistic regression are set as the Personality automatically when a nominal or ordinal response is placed into Y. Personality can also be manually set to Generalized Linear Model, which allows access to distributions and link functions other than for nominal and ordinal logistic regression. As usual, you place your model effects into the Construct Model Effects box. Remember: GZLMs cannot handle non-independent observations, so you cannot execute an analysis containing random effects. After running the model, if the Lack of Fit test is significant, it is likely that more model terms are necessary. (You want the highest possible RSquare (U) model fit, which is the R^2 value of the model.) The significance results are shown in the Effect Likelihood Ratio Tests report as a Chi-Square (χ^2).

In SPSS, Analyze > Generalized Linear Models > Generalized Linear Models provides access to GZLMs. The Type of Model tab provides convenient shortcuts for selecting both a distribution and a link function at once. Custom combinations can also be set at the bottom of the tab. For example, to analyze Likert scale data, Choose Ordinal logistic in the Ordinal Response section. On the Response tab, place the response in the Dependent Variable box. On the Predictors tab, place all factors in the Factors box. On the Model tab, create the full factorial model in the Model box. On the Estimation tab, set the Method to Newton-Raphson. On the Statistics tab, choose Likelihood ratio in the Chi-square Statistics box and Profile likelihood in the Confidence Interval Type. Click OK. The Tests of Model Effects box will be of greatest interest.

• Aligned Rank Transform • The Aligned Rank Transform (ART) is a procedure for preparing data in such a way that a parametric ANOVA can be used but the procedure can still be considered nonparametric in nature. The ART is therefore the nonparametric equivalent of either an *n*-way ANOVA or a repeated measures ANOVA on any number of factors. It can analyze between-subjects or within-subjects factors with two or more levels.

Although the regular rank transform (RT) has been studied since the early 1980s, it is known to produce unreliable interaction effects, making it unsuitable for factorial designs. The ART procedure produces accurate interaction effects by first *aligning* the data for each effect, main or interaction, and then ranking it. Data aligning is a procedure by which all potential effects other than the one of interest are subtracted from a continuous response by using marginal means, leaving only the effect of interest present, if there is one.

Rather than produce a lengthy explanation of the ART procedure here, you are required to read the following Note from CHI 2011:

---

[4] In SPSS, the Type of Model tab on the dialog available from Analyze > Generalized Linear Models > Generalized Linear Models does not permit using a Multinomial distribution with a Logit link function. Such a combination *is* available in SPSS 19 for Generalized Linear Mixed Models (GLMMs) from Analyze > Mixed Models > Generalized Linear. Owing to its simplicity, JMP is preferred when analyzing nominal responses.

Wobbrock, J.O., Findlater, L., Gergle, D. and Higgins, J.J. (2011). The Aligned Rank Transform for nonparametric factorial analyses using only ANOVA procedures. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '11)*. Vancouver, British Columbia (May 7-12, 2011). New York: ACM Press, pp. 143-146.

Data aligning is tedious, so the above publication provided a tool to do it for you. Please read the write-up for the ARTOOL program posted at this URL: http://depts.washington.edu/aimgroup/proj/art/. Next, download ARTOOL and save your data to a comma separated file (*.csv). (You can do this in Microsoft Excel.) Run your data through ARTOOL and ensure you understand the output file and its columns. Once the output is created, you can paste it into a statistics package like JMP or SPSS for performing your ANOVA.

As the write-up for ARTOOL says, it is good practice to first perform an appropriate analysis of variance on the "aligned" columns to ensure that all effects except the one for which the column was aligned are indeed "stripped out." Then perform an analysis of variance as appropriate on the "ART" columns, interpreting the result only for the effect matching each column.

• Generalized Linear Mixed Models (GLMMs) • Generalized Linear Mixed Models (GLMMs) are like the linear mixed-effects models you have used in previous modules to carry out analyses of variance. Like those models, GLMMs can contain both fixed and random effects. However, as *generalized* models, they are like GZLMs in not requiring the response variable to be normally distributed. GLMMs can use a variety of distributions and link functions, just like GZLMs, discussed above.

Because GLMMs can have random effects, they remedy the major limitation of GZLMs requiring independent observations and only between-subjects factors. GLMMs can handle correlated measurements and can therefore accommodate within-subjects factors, just as linear mixed models have done for us all along. GLMMs are therefore extremely powerful, and only recently has the computing power necessary to implement them become widely available.

Unfortunately, JMP does not provide access to GLMMs.

With version 19, GLMMs became available for the first time in SPSS. To access GLMMs, go to Analyze > Mixed Models > Generalized Linear.[5] The Data Structure tab enables you to use drag-and-drop to specify how your data table is laid out. Drop the name of your experimental unit (*e.g.*, "Subject") on the Subjects rectangle on the Canvas. Drop any within-subjects factors on the Repeated Measures rectangle on the Canvas. Do this in such a way that the picture developed on the Canvas resembles your tabulated data.

On the Fields & Effects tab, with Target selected at left, choose the Target, which is your response measure. Depending on its modeling type, different radio buttons will be enabled in the box labeled Target Distribution and Relationship (Link) with the Linear Model. Choosing "Linear model" here is equivalent to running a mixed-effects model analysis of variance as you have done in previous modules. SPSS provides a description of the currently selected option in the Description box, which can be useful.

[5] If you are using a GLMM to analyze repeated measures data, the response variable must have its Measure set to Scale or Nominal on the Variable View pane, even if it is ordinal in nature. This limits the utility of GLMMs for ordinal response data, but GLMMs remain useful for analyzing other types of nonparametric data (*e.g.*, event-count data using Poisson regression). For repeated measures on ordinal responses, use Generalized Estimating Equations (GEEs) or the Aligned Rank Transform (ART) procedure.

If you are uncertain as to which option to choose, view the distribution of your response measure to gain an intuition. An easy way to do this is in JMP using the Analyze > Distribution command.

On the Fields & Effects tab, with Fixed Effects selected at left, highlight and drag each fixed effect separately onto the rectangle labeled Main in the Effect builder. Similarly, highlight combinations of factors simultaneously and drop them on the 2-way, 3-way, or * boxes for fixed interaction effects.

On the Fields & Effects tab, with Random Effects selected at left, use the Add Block button to add random effects to the model. In general, intercepts should be used with the Include intercept checkbox. If the experimental unit is the only random effect, choose it from the Subject combination drop-down, include an intercept, and leave the Effect builder otherwise empty.

With these settings in place, click Run. Double-click the Model Summary graphic in the output viewer and maximize the Model Viewer window that opens. Find the fixed effects thumbnail on the left and click it. Set the Style drop-down to Table at the bottom of the window to see the analysis of variance results.

• Generalized Estimating Equations (GEEs) • Like GLMMs, Generalized Estimating Equations (GEEs) extend GZLMs to accommodate correlation among measurements, and therefore can handle within-subjects factors from studies with repeated measures. GEEs allow for correlation but do not require the structure of that correlation to be identified, and instead use estimators to derive the structure from the data. Also, unlike GLMMs, GEEs can accommodate repeated measures on ordinal response variables.

Unfortunately, JMP does not provide access to GEEs.

In SPSS, go to Analyze > Generalized Linear Models > Generalized Estimating Equations. On the Repeated tab, place all columns necessary to uniquely identify the experimental units in the Subject variables box. Place any within-subjects variables in their eponymous box. Set the Structure in the Working Correlation Matrix to reflect the covariance structure across measurements. (The default is "Independent." "Exchangeable" is synonymous with Compound Symmetry; *see* module 8, part 4.) On the Type of Model tab, choose the model to use as you did for GZLMs. On the Response tab, fill in Dependent Variable. On the Predictors tab, add all Factors, which must be nominal, and any Covariates, which must be scalar. On the Model tab, construct the full factorial model in the Model box. On the Estimation tab, change the Method to Newton-Raphson. Click OK to run the GEE analysis. In the output, the Tests of Model Effects table is of most interest. Results are reported as a Wald Chi-Square (χ^2).

For *each* of the six data sets A-F in *nonparam.xlsx*, answer the following questions. You should first duplicate these questions for each data set A-F. Also, since each method named in questions 4-11 will be performed, you may wish to first label *all* of the data tables with the proper test(s), and then begin performing the tests themselves. For the data set requiring the analyses in the last row of the table (ART, GLMM, GEE), perform all three tests on it.

13. What type of experimental design appears to have been used? Factors? Levels? Within- or between-subjects? Responses? Number of subjects? Is the table in wide- or long-format?
14. Using Analyze > Distribution in JMP, perform a test for normality and report the result. Do the data seem to violate the normality assumption?
15. Using Fit Y by X > UnEqual Variances in JMP, perform Levene's test for homogeneity of variances. (Hint: if you have multiple factors, you will have to test for each factor separately.) Report the result of each test. Do the data seem to violate the homoscedasticity assumption?

16. What *nonparametric* test should be used? Perform and report the test. Indicate whether JMP or SPSS was used.
17. What *parametric* test should be used? Perform and report the test. Indicate whether JMP or SPSS was used.
18. How do the results and conclusions compare between your nonparametric and parametric tests for this data table?

10. Categories, counts and proportions

prefs2., prefs3.*, liked.jmp, liked.freq.jmp, likedwthn.sav, comm.jmp, ptgbtwn.jmp, ptgmixed.jmp*

For some studies in human-computer interaction, especially field studies where behavior is logged, *count* measures are common. Take, for example, the deployment of a novel text entry method to a selection of participants "in the wild" for a period of 10 weeks. From raw field data, it would be very difficult to measure words per minute or error rates, because a researcher cannot know the intention of the participant. What were they *trying* to write? Did someone interrupt the participant? Did the phone ring mid-sentence? Did the participant change his mind mid-sentence? However, log files could show *counts* of events, such as letters, numbers, spaces and backspaces, use of cursor keys, and so on. If, say, a word prediction system was installed, we could count how many characters were entered via this feature, and thus approximate the character savings from word prediction.

Counts also appear in data where *events* are being coded from videos or interviews. If participants are videotaped for later coding, that coding often entails counting the number of times a certain behavior occurred.

Counts become *rates* when measured per unit time. In our example of fielding a text entry method, we could compute certain rates, like the number of characters entered per day. Words per minute is also a rate, but it has a formal definition which requires a laboratory setting.

A count can sometimes be turned into a *proportion*. If we can answer, "how many letters were entered?" we can sometimes also answer, "of all characters entered, what was the proportion of letters?" By definition, to have a proportion, we must be able to define a "denominator," *i.e.*, the maximum possible value in which our count occurs.

Counts, events, rates, and proportions are often not suitable to ANOVA. You learned in module 6 about normality tests and module 9 about nonparametric tests. This module continues your nonparametric training but with methods specifically designed to evaluate counts that fall into various response categories. Some of these methods are new, and some are those from these prior modules.

Part 1, One-sample Chi-Square (χ^2) test of proportions. You have seen the Chi-Square (χ^2) statistic before, *e.g.*, when reporting the results of a sphericity test, a Friedman test, or logistic regression. However, there are also Chi-Square tests of proportions. A *one-sample* test compares counts occurring among C unordered categories for whether those counts differ from hypothesized values, often the values set by *chance* (*i.e.*, equal probabilities for each category).

Open *prefs2.jmp*, which contains subjective preferences from 30 subjects who participated in a within-subjects study of the EdgeWrite versus Graffiti text entry methods. Each subject performed with both text entry methods, and at the end of the study, chose which method they preferred. Note that although the study itself had repeated measures, there is only one response in *prefs2.jmp* from each subject, making the data suitable to analysis with a Chi-Square test.

Although the *Subject* column is often omitted in such tables, it is a good idea to keep it for clarity. The question of interest can be framed as whether the proportion of subjects who liked one method or the other was significantly different than chance. To answer this question, go to Analyze > Distribution. Put *Preference* in Y, Columns and click OK. Using the Preference red-arrow menu, choose Test Probabilities. To test with equal weighting, one only needs to enter equal values into the two Hypoth Prob boxes. (If the entered values do not sum to 1.00, JMP will automatically rescale them according to their relative proportion.) For ease, simply enter "1" into each box. Be sure the first radio button option is selected to perform a two-sided test, and click Done.

1. What happened to the Hypoth Prob values after you clicked Done? Why did this happen?
2. Report the result from the Pearson one-sample Chi-Square test of proportions. What can you conclude?

The *prefs2.jmp* file is in long-format, with one row for each subject's response. Another common table format for Chi-Square tests is a "count table," "frequency table," or "weighted cases table." To make this table, go to Tables > Summary, highlight *Preference* on the left, and click Group. Then click OK. The resulting table will show, for each preference, how many rows matched it, giving you the count for that response.

Now go to Analyze > Distribution again, putting *Preference* in Y, Columns and *N Rows* in Freq. Click OK and answer the following questions.

3. How does the resulting report compare to the one you viewed for questions 1-2?
4. Conduct the one-sample Chi-Square test of proportions, report the result, and compare it to that from question 2.

Open *prefs2.sav* in SPSS to conduct a one-sample Chi-Square analysis on the long-format table. Go to Analyze > Nonparametric Tests > Chi-square. Transfer *Preference* into Test Variable List. Ensure All categories equal is selected in Expected Values. (This setting causes SPSS to do what you achieved by typing "1" into all Hypoth Prob cells in JMP.) Click OK.

5. Report the result for this one-sample Chi-Square test of proportions.
6. Comparing your results to those from question 2, determine whether SPSS is using a Likelihood Ratio Chi-Square result or a Pearson Chi-Square result. How do you know?

Open *prefs3.jmp* in JMP and see fictitious results for a within-subjects study of three text entry methods, EdgeWrite and Graffiti, as before, and now also ShapeWriter. The counts of responses can now fall into one of three categories instead of just two. Following the same procedures as above, answer the following:

7. Using the long-format table, determine whether the proportions of preferences among the three text entry methods are significantly different from chance.
8. Transform the long-format table into a weighted cases table. Describe what this table looks like.
9. Using the weighted cases table, perform the Chi-Square test again and report your results. How do they compare to those from question 7?

Open *prefs3.sav* in SPSS.

10. Perform a one-sample Chi-Square test of proportions on the long-format table. Report your results. How do they compare to those from question 7?

Part 2, Chi-Square (χ^2) test of proportions with independent groups. A *two-sample* Chi-Square test of proportions compares counts, but now coming from two independent groups. A *three-sample* test compares counts coming from three independent groups. And so on. Note that Chi-Square tests always presume the independence of responses, so they are not appropriate for multiple responses obtained from one subject.

In general, a Chi-Square test using independent groups compares whether *C* counts spread among *R* unordered categories are significantly different. The variables *R* and *C* are chosen because such data forms a *R*ow × *C*olumn table called a "contingency table." And again, an assumption of the test is that responses are independent, *i.e.*, each subject (or other experimental unit) can be counted only once.[6]

Open *liked.jmp*, which is a long-format table containing data from a 7-point Likert scale from *disliked* (1) to *liked* (7). (Note that these are ordered categories, but to begin with, we will simply regard them as seven separate unordered categories.) Two text entry methods have been judged after use, each by 32 different subjects who this time only used one text entry method or the other, constituting independent groups.

Now open *liked.freq.jmp*, a weighted cases table. Note the *Count* column indicating how many subjects marked the corresponding Likert value (1-7) for each text entry method.

11. Describe how both *liked.jmp* and *liked.freq.jmp* show the same data. (Do they? Are you sure? How can you verify this?)
12. Verify that the mean Likert response in both tables is the same for EdgeWrite and Graffiti. What are these means? How did you calculate the means for each method in *liked.jmp*? How did you do so in *liked.freq.jmp*? (Hint: the long-format table method means can be calculated in one step; it will require multiple steps, perhaps involving the creation of formula columns, to do so for the weighted cases table.)

Now perform a Chi-Square test of proportions on *liked.freq.jmp*. We have only one factor, *Method*, so we can first use Analyze > Fit Y by X. Put *Disliked-Liked* in Y, Response; *Method* in X, Factor; and *Count* in Freq. The Tests report shows the *N*, *df*, χ^2, and *p*-value necessary to report your Chi-Square result. (Here we will use the Likelihood Ratio result, but the Pearson result is fine in general, too.)

13. What is the result of this test? What does the result mean? Report the result and interpret the findings.

We can also perform the same test using Analyze > Fit Model, which would allow for more than one factor if we required it. However, whereas the Fit Y by X platform regards a nominal or ordinal response variable as equivalently categorical (unordered), the Fit Model platform changes its analysis based on this

[6] If correlated (non-independent) measures are taken, McNemar's test may be used, but it is only applicable to 2×2 tables, and thus has limited utility. McNemar's test is not covered in these modules. See http://en.wikipedia.org/wiki/McNemar%27s_test.

modeling type. Therefore, for now, flip the modeling type of *Disliked-Liked* to nominal. Then use Analyze > Fit Model with Nominal Logistic for Personality.

14. Examine the output and report the results. Do they match the results from question 13?

Now turn to *liked.jmp*. As before, flip *Disliked-Liked* to nominal for the sake of performing a Chi-Square test. Use both Analyze > Fit Y by X and Analyze > Fit Model on the long-format table. Note that there is no Freq parameter required when using the long-format table.

15. Examine the output and report the results from Fit Y by X. Do they match the results from questions 13-14?
16. Examine the output and report the results from Fit Model. Do they match the results from questions 13-15?

Note when using Analyze > Fit Y by X, two warnings are produced due to low cell counts. A rule of thumb for using the Likelihood Ratio Chi-Square is that the average cell count is 5 or more. A rule of thumb for using the Pearson Chi-Square is that 80% of cells have a count of 5 or more. When this is not the case *and when 2×2 designs are analyzed*, Fisher's Exact test can be used instead. If we had only two Likert values (say, "disliked" and "liked") along with two methods (EdgeWrite and Graffiti), then JMP would automatically perform Fisher's Exact Test for us. We'll see this in part 3, below.

Because semantically, we know that these data represent ordered Likert scores, we can do better than a mere Chi-Square test of proportions of unordered categories. Return to *liked.freq.jmp* and flip *Disliked-Liked* back to ordinal. Use Analyze > Fit Model to run ordinal logistic regression on this ordinal response.

17. Examine the output and report the results. How do they compare to those for questions 13-16?

Similarly, on *liked.jmp*, flip *Disliked-Liked* back to ordinal, and use Analyze > Fit Model to analyze this long-format data using ordinal logistic regression.

18. Examine the output and report the results. How do they compare to those for questions 13-17?

Now we will perform a Chi-Square test of proportions in SPSS using a weighted cases table. Open a blank data table in SPSS and replicate the data from *liked.freq.jmp*. Provide appropriate column names and values, with "EdgeWrite" as 1 and "Graffiti" as 2. (Be sure View > Value Labels is turned on.) For your variable modeling types, *Method* should be nominal, *Disliked2Liked* should be ordinal, and *Count* should be scalar. You may wish to save this table as *liked.freq.sav*.

Now perform a Chi-Square test of proportions. First go to Data > Weight Cases, choose Weight cases by, and set *Count* as the Frequency Variable. Then use Analyze > Descriptive Statistics > Crosstabs and for Row(s) put *Method*, for Column(s) put *Disliked2Liked*. Click Statistics and check Chi-square. Click Continue. Click OK.

19. Examine the output from the Chi-Square Tests report. Report the result and conclusion of this Chi-Square test.
20. In light of the results you saw from JMP when *Disliked-Liked* was nominal (questions 13-16) and when it was ordinal (questions 17-18), what can you conclude about how SPSS treated *Disliked2Liked*, which you modeled as ordinal, in this procedure?

Part 3, Fisher's Exact test. As noted above, when cell counts are low, the Chi-Square test is suspect. In the case of 2×2 designs, Fisher's Exact test can be used. JMP produces this test automatically in its Analyze > Fit Y by X platform when it is warranted.

First, we need to turn our *Disliked-Liked* ordinal outcome into a dichotomous one in *liked.jmp*. Define a new column called *Liked?*[7] and set it equal to the following formula:

21. Semantically, why is the value "4" the correct value in the above formula for defining *Liked?*.

Before we perform our tests, we need to define the value ordering for the *Liked?* column. The first value will determine how the odds ratio is constructed. Think of this as determining whether we ask about the relative odds of having 0/1 *vs*. 1/0. Semantically, we'd like to define the question as 1/0 because we are naming the column *Liked?*. Therefore, for the *Liked?* column, go to Column Info > Column Properties > Value Ordering and move the "1" above the "0". Then click OK.

Having defined the value order, use Analyze > Fit Y by X on *Liked?* and examine the results.

22. What is the Likelihood Ratio Chi-Square result for this test? Report and interpret the finding.
23. How does this result compare to the findings from using the *Disliked-Liked* response as a nominal outcome (questions 13-16, 19-20)? as an ordinal outcome (questions 17-18)?
24. What is the Fisher's Exact test result? Report and interpret the finding.
25. Is a Fisher's Exact test warranted in this case? Why or why not?

Also, analyze *Liked?* using nominal logistic regression with Analyze > Fit Model. (Recall that you would have no option of using Fit Y by X if you had a factorial design.)

26. Report the result of the Chi-Square test. Does it agree with what you found in question 22?

Part 4, Binomial logistic regression. Recall from module 9, part 4 that binomial logistic regression is a Generalized Linear Model (GZLM) with a binomial distribution and logit link function designed to handle dichotomous (0/1) outcomes. Recall that GZLMs cannot handle within-subjects factors, but Generalized Linear Mixed Models (GLMMs) and Generalized Estimating Equations (GEEs) can, effectively extending and remedying GZLMs in this way.

For practice, we'll use two new tables to perform binomial logistic regression on the data in *liked.jmp*. (In reality, you already performed binomial logistic regression for question 26 on the long-format table because *Liked?* is dichotomous. These are two new table formats.)

[7] In JMP, I prefer to use a question mark (?) at the end of dichotomous variables, whether they are independent factors or dependent measures. This serves as a visual reminder that the variable is dichotomous. SPSS does not allow this symbol in column names.

The first table we need is a weighted cases table. Create this table from *liked.jmp* such that you have the following columns: *Method*, *Liked?*, and *Count*. Define the value ordering on *Liked?* such that 1 is first.

27. How did you create the new table? How many rows are in the table?
28. What are the modeling types of each variable?
29. What are the counts for the four combinations of *Method* = {EdgeWrite, Graffiti} and *Liked?* = {1, 0}?

You may wish to save this table as *likedornot.jmp*. Now perform your analysis using Analyze > Fit Model. You should by now understand how to fill out the model dialog. Don't forget to place your *Count* column in Freq.

30. What is the outcome of this test? Report and analyze your finding.
31. How does it compare to the results from questions 22 and 26?

The second type of table, which we'll call *likedornot.gzlm.jmp*, allows us to explicitly choose a Generalized Linear Model for binomial logistic regression. This table will be similar to *likedornot.jmp* but has just two rows and an additional column named *Trials*, indicating the total trials. So its columns should be *Method*, *Liked?*, *Count*, and *Trials*. The *Liked?* column is superfluous at this point, and should just have "1" in each row. (We'll keep it for clarity.) *Count* should now have the total number of 1's for each method. *Trials* has the total number of trials. In essence, in moving from *likedornot.jmp* to *likedornot.gzlm.jmp*, we've removed the two rows that contained *Liked?* = 0, and replaced them with the *Trials* column that identifies the total number of responses.

32. How did you create the new table? What are the *Count* and *Trials* values for the EdgeWrite row? for the Graffiti row?

When your table is ready, go to Analyze > Fit Model. For Personality, select Generalized Linear Model. For Distribution, select Binomial. (Leave the link function as Logit.) Add *Method* to your model effects box. Place both *Count* and *Trials* in the Y box. The first value here identifies the number of 1's, while the second value identifies the total number of trials. Click Run Model.

33. Examine the Effect Test box. Report the result, being careful to report the correct *N* in the Chi-Square result.
34. How does the result compare to those from questions 22, 26, and 30?

It turns out we can also use our *liked.jmp* data to perform explicit binomial logistic regression. Return to that table and choose Analyze > Fit Model, setting Personality to Generalized Linear Model, Distribution as Binomial, Y as *Liked?*, and *Method* as the model effect. Click Run Model.

35. Report the result of this analysis.
36. How does the result compare to questions 22, 26, 30, and 33?

As one last exploration, we'll also analyze *liked.jmp* in SPSS. Open a blank spreadsheet in SPSS and paste the data from *liked.jmp* into it. On the Variable View, set variable names to *Subject* (Nominal),

Method (Numeric, Nominal), *Disliked2Liked* (Ordinal), and *LikedP*[8] (Nominal). For *Method*, define Values as 1 = "EdgeWrite" and 2 = "Graffiti". Then enter these values in your spreadsheet (*Method* rows 1-32 get "1", 33-64 get "2"). (Be sure View > Value Labels is turned on.)

When your table is ready, go to Analyze > Generalized Linear Models > Generalized Linear Models. On Type of Model, select Binary logistic. On Response, place *LikedP* in the Dependent Variable box. On Predictors, place *Method* in the Factors box. On Model, place *Method* in the Model box. On Estimation, set Method to be Newton-Raphson. On Statistics, choose Likelihood ratio for Chi-Square Statistics and Profile likelihood for Confidence Interval Type. Then click OK.

37. Report the results. How do the findings compare to those from questions 22, 26, 30, 33, and 35?

Part 5, Logistic regression with repeated measures. In parts 2 and 4, above, you analyzed Likert scale data and dichotomous data with Generalized Linear Models (GZLMs). Recall that GZLMs cannot handle repeated measures, *i.e.*, all factors must be between-subjects, not within-subjects. Generalized Linear Mixed Models (GLMMs) and Generalized Estimating Equations (GEEs) remedy these limitations. Both of these analyses are unavailable in JMP but are available in SPSS as of version 19.

In SPSS, open *likedwthn.sav*, which contains the same data as *liked.jmp* except that now there are only 32 subjects, not 64.

38. Is the data table *likedwthn.sav* in long format or wide format? How do you know?
39. For each factor in *likedwthn.sav*, identify whether it is between-subjects or within-subjects and say how you know.
40. How many dependent measures are shown in *likedwthn.sav*? What are their names and types?

Recall that liked linear mixed-effects models, Generalized Linear Mixed Models (GLMMs) must have a scalar outcome for repeated measures analyses. While we could turn *Disliked2Liked* from "Ordinal" to "Scale" in the Variable View, we can leave it as an ordinal measure if we use Generalized Estimating Equations (GEEs) to carry out our analysis. You may wish to refer back to GEEs from module 9, part 4 before continuing.

41. Run a GEE on *Disliked2Liked* as an ordinal measure in *likedwthn.sav*. What are the results, and how do they compare to the between-subjects analyses of the ordinal response using GZLMs from questions 17-18?

For question 37, above, you analyzed *LikedP* as a dichotomous response using binomial logistic regression, *i.e.*, using Generalized Linear Models (GZLMs) and the Binary logistic model type. Now use Generalized Linear Mixed Models (GLMMs) to analyze *LikedP* in *likedwthn.sav*. You may wish to refer back to GLMMs from module 9, part 4 before continuing.

---

[8] SPSS does not allow question marks (?) in variable names. The use of suffix "P" harkens back to Lisp programming, where Boolean variables, by convention, are given this suffix. Nick Parlante (http://www-cs-faculty.stanford.edu/~nick/), computer science lecturer at Stanford University, tells a story about querying his roommates for their pizza orders by asking, "Pizza-P?"

42. Run a GLMM on *LikedP* as a nominal dichotomous measure in *likedwthn.sav*. What are the results, and how do they compare to the between-subjects analyses using GZLMs from questions 22, 26, 30, 33, 35, and 37?

Part 6, Nominal logistic regression. You've already seen nominal logistic regression at work; it is the name for the analysis JMP's Fit Model platform uses when the response is nominal. In SPSS, nominal logistic regression is indicated by a Multinomial distribution and Logit link function. It can be considered the generalization of binomial regression with *N* categorical outcomes, not just two.

We will use this opportunity to examine a between-subjects factorial design. The file *comm.jmp* has fictitious data about the communication preferences of teenagers and young adults. Thirty-six people, 18 of each gender, were surveyed as to whether they prefer the new communication media of Facebook or texting to email. (Phrased this way, it requires email to be our last category; examine *Pref*'s Column Info > Column Properties > Value Ordering, and see how email has been placed last. The default would have been alphabetical order.)

43. What are the factors and levels in this data table? (Hint: if we want to be picky, there are no factors here, but other things.) Are these between-subjects or within-subjects variables?
44. What is the response variable?
45. Form three hypotheses about these data, one for each variable and one for their interaction.
46. Use Analyze > Distribution to examine all three hypotheses. Report what you see and revise your hypotheses if necessary.

Use Analyze > Fit Model and run the factorial nominal logistic regression.

47. What is the effect of gender on communication preference? Report and interpret the finding.
48. What is the effect of age on communication preference? Report and interpret the finding.
49. Is there an interaction between gender and age on communication preferences? Report and interpret the finding. (Hint: remember that *p*-values between .05 and .10 are "trends" and deserve some extra interpretive attention.)

Ignore gender by using Analyze > Fit Y by X to explore the relationship between *Age* and *Pref*. You should produce a graph called "Logistic Fit of Pref By Age." This graph is very useful once you know how to read it. Consult the JMP help documentation's section on "Nominal Logistic Regression," particularly the subsection called "The Cumulative Logistic Probability Plot." Google Books has the same text online if you do not have access to the JMP help documentation. (Direct your browser here: http://books.google.com/books?q="The Cumulative Logistic Probability Plot".)

50. Explain how to read the odds of a preference being Facebook for a given age from the plot. Then do the same for texting and email.
51. Judging from the graph only, what age is most likely to prefer Facebook? texting? email? For each of these preferences, say how many subjects out of all subjects this age made that choice.

Part 7, Poisson regression. As you read in module 9, part 4, Poisson regression is a Generalized Linear Model (GZLM) that is useful when analyzing event-count data. Poisson regression is especially suited to rare event data, meaning small counts are common and large counts are exceedingly rare. Consider a study of the number of suicides per school district for a state in a given decade. We should expect, and

hope, that most districts in the state have counts of zero; but some will have one, fewer will have two, and still fewer, three. In human-computer interaction, recognition errors for an accurate gesture recognizer might be considered rare events, or other situations when errors are infrequent. Importantly, Poisson regression requires all response data to be non-negative integers. If you have error rates expressed as decimals, you will need to multiply them by 100 and regard the rates as *counts per hundred*.

Open *ptgbtwn.jmp* and examine this data table. Then answer the following questions.

52. What is an educated guess as to what this experiment shows? (After you answer, read the Column Info > Notes for the *Errors* column to discover what this fictitious experiment was about.)
53. Express this study design in multiplicative notation identifying the factors and levels and whether they are within- or between-subjects.
54. Use Analyze > Distribution to examine how *Errors* are distributed for each level of *Device*. Do this again for each level of *Cursor*. Show six test results for examining whether the response is Poisson-distributed within each factor. You may wish to make a table.
55. Now analyze the distribution of *Errors* as a whole. Is it Poisson-distributed? How can this be in light of what you found in question 54?
56. Now to go to Analyze > Fit Model and run a Generalized Linear Model with Poisson distribution, Log link, and overdispersion checked. Report and interpret the conclusions from the Effect Tests table.
57. Also run an analysis of variance on these data for comparisons. Report and interpret the conclusions. How do the results compare to those from the Poisson regression in question 56?

Find the Device*Cursor report and use the red-arrow menu to select LSMeans Plot. The default places the *Cursor* factor on the *x*-axis and the *Device* factor as the line types. Now remove this plot and once again use the red-arrow menu to select LSMeans Plot, *but before you click to open the menu, hold down SHIFT*. This will "slice" the graph the other way, placing *Device* on the *x*-axis and *Cursor* as the line types.

58. In looking at this graph, does the reason for the significant *Device*Cursor* interaction become clear? What is it? Explain your reasoning.

Part 8, Contrast tests. Continuing with your analysis of variance output from questions 57-58, run a contrast test by using the red-arrow menu on the Device*Cursor report to select LSMeans Contrast. Use the + and – buttons to select the levels you want to compare. From the graph, it appears possible that the point cursor and area cursor each do not significantly differ between the mouse and trackball.

59. Perform a contrast to see if the point cursor and area cursor differ within the mouse. Report and interpret your finding.
60. Perform a contrast to see if the point cursor and area cursor differ within the trackball. Report and interpret your finding.
61. In light of the findings from questions 59 and 60, it may be that the point and area cursors do not differ at all. If they do, what level of *Device* must this be due to? Perform the relevant contrast test to confirm or refute your hypothesis.
62. Investigate whether the point and area cursors are significantly different *overall* by performing a contrast on the Cursor report. Report and interpret your finding.

63. Now investigate the same thing *but using a contrast on the Device*Cursor report*. (Hint: you will be setting three positive weights against three negative weights.) Do your results agree with question 62? Explain why or why not.
64. Finally, return to your Poisson regression output and use the red-arrow menu on Generalized Linear Model Fit to select Contrast. Determine whether the point and area cursors significantly differ. What contrast did you perform? Report and interpret your results. Are these cursors closer or further from being significantly different using Poisson regression than ANOVA?

Part 9, Counts from a mixed factorial design. The event-count data has thus far been between-subjects, *i.e.*, each subject has only contributed one count. What if we have factorial event-count data? Recall from module 9 that the Wilcoxon signed-rank and Friedman tests handle non-independent samples for only one factor. If a factorial design is used, the Aligned Rank Transform (ART) procedure, Generalized Linear Mixed Models (GLMMs), and Generalized Estimating Equations (GEEs) can be used. Each of these analyses can be used for event-count data, which is what we'll do here.

Open *ptgmixed.jmp*, which shows the same pointing error data as *ptgbtwn.jmp*, but this time, with only 12 subjects instead of 36. Each subject used one cursor type with three input devices, resulting in a mixed factorial design. The between-subjects factor is *Cursor*, and the within-subjects factor is *Device*. Note the column order has been changed to reflect this relationship clearly in the table.

65. Before conducting the nonparametric procedures, first analyze this data set using a (linear) mixed-effects model analysis of variance. Report and interpret your findings.
66. How do your results for the mixed factorial design compare to those from the between-subjects design (question 57)? Make a table comparing your degrees-of-freedom, F-ratios, and *p*-values.

Now copy your data from JMP to Microsoft Excel, including column names, and then save it as an MS-DOS *.csv file. Run the ARTOOL program available at http://depts.washington.edu/aimgroup/proj/art/ to align and rank the data. Then open the aligned *.csv file in Microsoft Excel, and paste the data into a new JMP table. Paste in the column names as well. Save the new JMP table as *ptgmixed.art.jmp* to reflect the *.csv filename created for you by ARTOOL.

67. Use Tables > Summary to sum the aligned columns. Aligned columns in a long-format table should always sum to zero. Do they?
68. Run a (linear) mixed-effects model analysis of variance on the "aligned" columns. Are all effects for which the response was *not* aligned properly "stripped out?"
69. Now run the model on the "ART" columns. Report and interpret your findings for *Device*, *Cursor*, and their interaction.
70. How do each of these findings compare to those from the parametric analysis (question 65)? Reuse the mixed factorial design results from your table in question 66, and fill in the other half with the results from question 69 to facilitate comparisons.

Now we will employ Generalized Linear Mixed Models (GLMMs) and Generalized Estimating Equations (GEEs) to analyze *Errors* for the *ptgmixed.jmp* data. As these two analyses are unavailable in JMP, recreate the *ptgmixed.jmp* data in SPSS 19 and save the new table as *ptgmixed.sav*. Then answer the following questions.

71. Run a GLMM on *Errors* as a scalar response, *Cursor* as a between-subjects nominal factor, *Device* as a within-subjects nominal factor, and *Subject* as a random effect. Use Poisson regression, accessed via choosing Loglinear as the model type. On the Random Effects pane, be sure to add *Subject* as a random effect by adding a block, choosing *Subject* in the Subject combination dropdown and clicking Include intercept. What are the results of the statistical analyses?
72. Run a GEE on *Errors* as a scalar response, *Cursor* as a between-subjects nominal factor, *Device* as a within-subjects nominal factor, and *Subject* as a Subject variable. Use a Poisson loglinear model. What are the results of the statistical analyses?
73. Create a table to compare the results from your ART, GLMM, and GEE analyses of the *ptgmixed.** data. Which analysis seems most powerful for this data set?

Bibliography

Aitchison, J. and Brown, J. A. C. (1957). *The Lognormal Distribution*. Cambridge, England: Cambridge University Press.

Akritas, M. G. and Brunner, E. (1997). A unified approach to rank tests for mixed models. *Journal of Statistical Planning and Inference 61* (2), pp. 249-277.

Akritas, M. G. and Osgood, D. W. (2002). Guest editors' introduction to the special issue on nonparametric models. *Sociological Methods and Research 30* (3), pp. 303-308.

Anderson, T.W. and Darling, D.A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Annals of Mathematical Statistics 23* (2), pp. 193-212.

Anderson, T.W. and Darling, D.A. (1954). A test of goodness of fit. *Journal of the American Statistical Association 49* (268), pp. 765-769.

Bartlett, M. S. (1947). The use of transformations. *Biometrics 3* (1), pp. 39-52.

Bennington, C. C. and Thayne, W. V. (1994). Use and misuse of mixed model analysis of variance in ecological studies. *Ecology 75* (3), pp. 717-722.

Berry, D. A. (1987). Logarithmic transformations in ANOVA. *Biometrics 43* (2), pp. 439-456.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B 26* (2), pp. 211-252.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association 88* (421), pp. 9-25.

Card, S. K., Moran, T. P. and Newell, A. (1983). *The Psychology of Human-Computer Interaction*. Hillsdale, New Jersey: Lawrence Erlbaum.

Conover, W.J. (1972). A Kolmogorov goodness-of-fit test for discontinuous distributions. *Journal of the American Statistical Association 67* (339), pp. 591-596.

Conover, W. J. and Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics *The American Statistician 35* (3), pp. 124-129.

D'Agostino, R.B. (1986). Tests for the normal distribution. In *Goodness-of-Fit Techniques*, R. B. D'Agostino and M. A. Stephens (eds.). New York: Marcel Dekker, pp. 367-420.

Fawcett, R. F. and Salter, K. C. (1984). A Monte Carlo study of the F test and three tests based on ranks of treatment effects in randomized block designs. *Communications in Statistics: Simulation and Computation 13* (2), pp. 213-225.

Findlater, L., Moffatt, K., McGrenere, J. and Dawson, J. (2009). Ephemeral adaptation: The use of gradual onset to improve menu selection performance. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '09)*. Boston, Massachusetts (April 4-9, 2009). New York: ACM Press, pp. 1655-1664.

Frederick, B. N. (1999). Fixed-, random-, and mixed-effects ANOVA models: A user-friendly guide for increasing the generalizability of ANOVA results. In *Advances in Social Science Methodology*, B. Thompson (ed). Stamford, Connecticut: JAI Press, pp. 111-122.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association 32* (200), pp. 675-701.

Geisser, S. and Greenhouse, S. W. (1958). An extension of Box's results on the use of the F distribution in multivariate analysis. *Annals of Mathematical Statistics 29* (3), pp. 885-891.

Greenhouse, S. W. and Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika 24* (2), pp. 95-112.

Hanley, J. A., Negassa, A., Edwardes, M. D. deB. and Forrester, J.E. (2003). Statistical analysis of correlated data using generalized estimating equations: An orientation. *American Journal of Epidemiology 157* (4), pp. 364-375.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

Hayter, A. J. (1984). A proof of the conjecture that the Tukey-Kramer multiple comparisons procedure is conservative. *Annals of Mathematical Statistics 12*, pp. 61-75.

Higgins, J. J., Blair, R. C. and Tashtoush, S. (1990). The aligned rank transform procedure. *Proceedings of the Conference on Applied Statistics in Agriculture*. Manhattan, Kansas: Kansas State University, pp. 185-195.

Higgins, J. J. and Tashtoush, S. (1994). An aligned rank transform test for interaction. *Nonlinear World 1* (2), pp. 201-211.

Higgins, J. J. (2004). *Introduction to Modern Nonparametric Statistics*. Pacific Grove, California: Duxbury Press.

Hodges, J.L. and Lehmann, E.L. (1962). Rank methods for combination of independent experiments in the analysis of variance. *Annals of Mathematical Statistics 33* (2), pp. 482-497.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics 6* (2), pp. 65-70.

Hoyle, M. H. (1973). Transformations: An introduction and a bibliography. *International Statistical Review 41* (2), pp. 203-223.

Kackar, R. N. and Harville, D. A. (1984). Approximations for standard errors of estimators of fixed and random effect in mixed linear models. *Journal of the American Statistical Association 79* (388), pp. 853-862.

Kaiser, J. F. and Reed, W. A. (1977). Data smoothing using low-pass digital filters. *Review of Scientific Instruments 48* (11), pp. 1447-1457.

Kaptein, M., Nass, C. and Markopoulos, P. (2010). Powerful and consistent analysis of Likert-type rating scales. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '10)*. Atlanta, Georgia (April 10-15, 2010). New York: ACM Press, pp. 2391-2394.

Kenward, M. G. and Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics 53* (3), pp. 983-997.

Kramer, C. Y. (1956). Extension of multiple range tests to group means with unequal numbers of replications. *Biometrics 12*, pp. 309-310.

Kristensson, P.-O. and Zhai, S. (2004). SHARK^2: A large vocabulary shorthand writing system for pen-based computers. *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST '04)*. Santa Fe, New Mexico (October 24-27, 2004). New York: ACM Press, pp. 43-52.

Krueger, C. and Tian, L. (2004). A comparison of the general linear mixed model and repeated measures ANOVA using a dataset with multiple missing data points. *Biological Research for Nursing 6* (2), pp. 151-157.

Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association 47* (260), pp. 583-621.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics 38* (4), pp. 963-974.

Lehmann, E. L. (2006). *Nonparametrics: Statistical Methods Based on Ranks*. New York: Springer.

Levene, H. (1960), Robust tests for the equality of variances. In I. Olkin (ed.), *Contributions to Probability and Statistics*. Stanford, California: Stanford University Press.

Limpert, E., Stahel, W. A. and Abbt, M. (2001). Log-normal distributions across the sciences: Keys and clues. *BioScience 51* (5), pp. 341-352.

Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics 46* (3), pp. 673-687.

Littell, R. C., Milliken, G. A., Stroup, W. W. and Wolfinger, R. D. (1996). Analysis of repeated measures data. In *SAS System for Mixed Models*. Cary, North Carolina: SAS Institute, Inc., pp. 87-134.

Littell, R. C., Henry, P. R. and Ammerman, C. B. (1998). Statistical analysis of repeated measures data using SAS procedures. *Journal of Animal Science 76* (4), pp. 1216-1231.

Ludbrook, J. (1998). Multiple comparison procedures updated. *Clinical and Experimental Pharmacology and Physiology 25* (12), pp. 1032-1037.

MacKenzie, I. S. and Zhang, S. X. (1999). The design and evaluation of a high-performance soft keyboard. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '99)*. Pittsburgh, Pennsylvania (May 15-20, 1999). New York: ACM Press, pp. 25-31.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics 18* (1), pp. 50-60.

Mansouri, H. (1999). Aligned rank transform tests in linear models. *Journal of Statistical Planning and Inference 79* (1), pp. 141-155.

Mansouri, H. (1999). Multifactor analysis of variance based on the aligned rank transform technique. *Computational Statistics and Data Analysis 29* (2), pp. 177-189.

Mansouri, H., Paige, R. L. and Surles, J. G. (2004). Aligned rank transform techniques for analysis of variance and multiple comparisons. *Communications in Statistics: Theory and Methods 33* (9), pp. 2217-2232.

Massey, F.J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association 46* (253), pp. 68-78.

Milliken, G. A. and Johnson, D. E. (1984). Analysis of Messy Data vol. I: Designed Experiments. New York: Van Nostrand Reinhold Company.

Petocz, P. (2005). An upper bound on standard deviation as a function of range. *Teaching Statistics 27* (2), pp. 42-44.

Richter, S. J. (1999). Nearly exact tests in factorial experiments using the aligned rank transform. *Journal of Applied Statistics 26* (2), pp. 203-217.

Salter, K. C. and Fawcett, R. F. (1985). A robust and powerful rank test of treatment effects in balanced incomplete block designs. *Communications in Statistics: Simulation and Computation 14* (4), pp. 807-828.

Salter, K. C. and Fawcett, R. F. (1993). The art test of interaction: A robust and powerful rank test of interaction in factorial models. *Communications in Statistics: Simulation and Computation 22* (1), pp. 137-153.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin 2* (6), pp. 110-114.

Sawilowsky, S. S. (1990). Nonparametric tests of interaction in experimental design. *Review of Educational Research 60* (1), pp. 91-126.

Schuster, C. and von Eye, A. (2001). The relationship of ANOVA models with random effects and repeated measurement designs. *Journal of Adolescent Research 16* (2), pp. 205-220.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology 46* (1), pp. 561-584.

Shah, A. K. (1985). A simpler approximation for areas under the standard normal curve. *The American Statistician 39* (1), p. 80.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika 52* (3 & 4), pp. 591-611.

Stephens, M.A. (1972). EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association 69* (347), pp. 730-737.

Tukey, J. (1953). A problem of multiple comparisons. Dittoed manuscript of 396 pages, Princeton University.

Vermunt, J. K. (1997). *Log-linear Models for Event Histories*. Thousand Oaks, California: Sage Publications.

Ware, J.H. (1985). Linear models for the analysis of serial measurements in longitudinal studies. *American Statistician 39* (2), pp. 95-101.

Welch, B.L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika 38* (3 & 4), pp. 330-336.

West, B.T. (2009). Analyzing longitudinal data with the linear mixed models procedure in SPSS. *Evaluation and the Health Professions 32* (3), pp. 207-228.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin 1* (6), pp. 80-83.

Winship, C. and Mare, R. D. (1984). Regression models with ordinal variables. *American Sociological Review 49* (4), pp. 512-525.

Wobbrock, J. O., Chau, D. H. and Myers, B. A. (2007). An alternative to push, press, and tap-tap-tap: Gesturing on an isometric joystick for mobile phone text entry. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '07)*. San Jose, California (April 28-May 3, 2007). New York: ACM Press, pp. 667-676.

Wobbrock, J.O., Findlater, L., Gergle, D. and Higgins, J.J. (2011). The Aligned Rank Transform for nonparametric factorial analyses using only ANOVA procedures. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '11)*. Vancouver, British Columbia (May 7-12, 2011). New York: ACM Press, pp. 143-146.

Wobbrock, J. O., Myers, B. A. and Kembel, J. A. (2003). EdgeWrite: A stylus-based text entry method designed for high accuracy and stability of motion. *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST '03)*. Vancouver, British Columbia (November 2-5, 2003). New York: ACM Press, pp. 61-70.

Wobbrock, J. O. and Myers, B. A. (2006). Trackball text entry for people with motor impairments. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '06)*. Montréal, Québec (April 22-27, 2006). New York: ACM Press, pp. 479-488.

Wobbrock, J. O., Rubinstein, J., Sawyer, M. W. and Duchowski, A. T. (2008). Longitudinal evaluation of discrete consecutive gaze gestures for text entry. *Proceedings of the ACM Symposium on Eye Tracking Research and Applications (ETRA '08)*. Savannah, Georgia (March 26-28, 2008). New York: ACM Press, pp. 11-18.

Wobbrock, J. O., Wilson, A. D. and Li, Y. (2007). Gestures without libraries, toolkits or training: A $1 recognizer for user interface prototypes. *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST '07)*. Newport, Rhode Island (October 7-10, 2007). New York: ACM Press, pp. 159-168.

Zhai, S. and Kristensson, P. (2003). Shorthand writing on stylus keyboard. *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '03)*. Ft. Lauderdale, Florida (April 5-10, 2003). New York: ACM Press, pp. 97-104.

Zeger, S.L., Liang, K.-Y. and Albert, P.S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics 44* (4), pp. 1049-1060.