APPENDIX FOR DAY 4

Suggestions for dealing with the File Drawer Problem: *(p. 6)*

Suggestions for researchers:

- Carefully review the literature *and* any relevant research registries before you embark on new research.
- Take the file drawer problem into account when writing a literature review.
- These considerations are especially important when conducting a meta-analysis.
- Make every effort to publish good research, even if results are not statistically significant, are not practically significant, or do not meet hopes or expectations.

Suggestion for reviewers, editors, etc:

- Accept papers on the quality of the research and writing, *not* on the basis of whether or not the results are statistically or practically significant or whether or not they are as expected.
- If necessary, work to implement this as the policy of the journals you are affiliated with.

Suggestions for consumers of research:

- Do not let a single research result convince you of anything.
- If you are reading a meta-analysis, check whether and how well the authors have taken the file-drawer problem into account.

Suggestions for data snooping professionally and ethically *(p. 18)*

1. Educate yourself on the limitations of statistical inference: Model assumptions, the problems of Types I and II errors, power, and multiple inference, including the "hidden comparisons" that may be involved in data snooping (as in the example on p. 18 of the notes).

2. Plan your study to take into account the problems involving model assumptions, Types I and II errors, power, and multiple inference. Some specifics to consider:

a. ***If you will be gathering data***, decide *before gathering the data*:

- The questions you are trying to answer.
- How you will gather the data and the inference procedures you intend to use to help answer your questions.
    - *These need to be planned together, to maximize the chances that the data will fit the model assumptions of the inference procedures.*
- Whether or not you will engage in data snooping.
- The type I error rate (or false discovery rate) and power that would be appropriate (considering the consequences of these types of errors in the situation you are studying).
    - Be sure to allow some portion of Type I error for any data snooping you think you might do.

Then do a power analysis to see what sample size is needed to meet these criteria.

- Take into account any relevant considerations such as intent-to-treat analysis (see below), or how you will deal with missing data.
- *If the sample size needed is too large for your resources, you will need to either obtain additional resources or scale back the aims of your study.*

b. ***If you plan to use existing data***, you will need to go through a process similar to that in (a) *before looking at the data*:

- Decide on the questions you are trying to answer.
- *Find out how the data were gathered.*
- Decide on inference procedures that i) will address your questions of interest and ii) have model assumptions *compatible with how the data were collected.*
  - *If this turns out to be impossible, the data are not suitable.*
- Decide whether or not you will engage in data snooping.
- Decide the type I error rate (or false discovery rate) and power that would be appropriate (considering the consequences of these types of errors in the situation you are studying).

Then do a power analysis to see what sample size is needed to meet these criteria.

- Take into account any relevant considerations such as those listed above.
- *If the sample size needed is larger than the available data set, you will need to either scale back the aims of your study, or find or create another larger data set.*

c. ***If data snooping is intended to be the purpose or an important part of your study***, then *before you look at the data*, divide it randomly into two parts: One to be for used for discovery purposes (generating hypotheses), the other to be used for confirmatory purposes (testing hypotheses).

- Be careful to do the randomization in a manner that preserves the structure of the data.
  - For example, if you have students nested in schools nested in school districts, you need to preserve the nesting.
  - e.g., if a particular student is assigned to one group (discovery or confirmatory), then the student's school and school district need to be assigned to the same group.
- Using a type I error rate or false discovery rate may not be obligatory in the discovery phase, but may be practical to help you keep the number of hypotheses you generate down to a level that you will be able to test (with a reasonable bound on Type I error rate or false discovery rate, and a reasonable power) in the confirmatory phase

- A preliminary consideration of Type I errors and power should be done to help you make sure that your confirmatory data set is large enough.

    - Be sure to then give further thought to consequences of Type I and II errors for the hypotheses you generate with the discovery data set, and set an overall Type I error rate (or false discovery rate) for the confirmatory stage.

3. Report your results carefully, aiming for honesty and transparency

- State clearly the questions you set out to study.
- State your methods, *and your reasons for choosing those methods*. For example:
    - Why you chose the inference procedures you used;
    - Why you chose the Type I error rate and power that you used.
- Give details of how your data were collected.
- State clearly what (if anything) was data snooping, and how you accounted for it in overall Type I error rate or False Discovery Rate.
- Include a "limitations" section, pointing out any limitations and uncertainties in the analysis. Examples:
    - If power was not large enough to detect a practically significant difference;
    - Any uncertainty in whether model assumptions were satisfied;
    - If there was possible confounding;
    - If missing data created additional uncertainty, etc.
- Be careful not to inflate or over-interpret conclusions, either in the abstract or in the results or conclusions sections.

Suggestions for Checking Model Assumptions of Equal Variance or Normality (p. 29)

Checking for Equal Variance

- Plot residuals against fitted values (in most cases, these are the estimated conditional means, according to the model), since it is not uncommon for conditional variances to depend on conditional means, especially to increase as conditional means increase.

    - This would show up as a funnel or megaphone shape to the residual plot.

- Especially with complex models, plotting against factors or regressors might also pick up unequal variance.

- *Caution*: Hypothesis tests for equality of variance are often *not* reliable, since they also have model assumptions and are typically not robust to departures from those assumptions.

Checking for Normality or Other Distribution

Caution: A histogram (whether of outcome values or of residuals) is *not* a good way to check for normality, since histograms of the same data but using different bin sizes (class-widths) and/or different cut-points between the bins may look quite different.

Instead, use a *probability plot* (also know as a *quantile plot* or *Q-Q plot*).

- Most statistical software has a function for producing these.
- *Caution*: Probability plots for small data sets are often misleading; it is very hard to tell whether or not a small data set comes from a particular distribution.

Example where a Linear Model Fits with Two Predictors, but Dropping One Results in a Model Requiring a Non-linear Functionl: (p. 32)

If the model with two predictors X_1 and X_2, and response variable Y, has conditional linear mean function

$$E(Y|X_1, X_2) = 1 + 2X_1 + 3X_2$$

but also X_1 and X_2 are related by

$$E(X_1|X_2) = \log(X_1),$$

then it can be calculated that

$$E(Y|X_1) = 1 + 2X_1 + 3\log(X_1),$$

which says that a linear model does *not* fit when Y is regressed on X_1 alone.

More on Fixed vs Random effects Terminology: (p. 39)

Usage of "random" in this and similar contexts is not uniform.

- For example, some authors, in discussing hierarchical (multilevel) analysis, may refer to an intercept as "random" when interest is restricted to a finite population with all members present in the data (e.g., the various states of the U.S.A.), but the intercept is allowed to be different for different members of the population.
- Using the term "variable intercept" can help emphasize that, although the intercept is allowed to vary, interest is only in the finite population, with no implication of inference beyond that population.

Additional Comments about Fixed and Random Factors (p. 41)

- The standard methods for analyzing random effects models assume that the random factor has infinitely many levels, but usually still work well if the total number of levels of the random factor is at least 100 times the number of levels observed in the data.

    - Situations where the total number of levels of the random factor is less than 100 times the number of levels observed in the data require special "finite population" methods.
- An interaction term involving both a fixed and a random factor should be considered a random factor.
- A factor that is nested in a random factor should be considered random.

Suggestions for Dealing with Pseudoreplication (p. 47)

1. *Avoid it if at all possible.*

Key in doing this is to

- Carefully determine what the experimental/observational units are;
- Then be sure that each treatment is randomly applied to more than one experimental/observational unit.

For example, in comparing curricula (Example 3 above), if ten schools participated in the experiment and five were randomly assigned to each treatment (i.e., curriculum), then each treatment would have five replications; this would give some information about the variability of the effect of the different curricula.

2. *If it is not possible to avoid pseudoreplication*, then:

a. Do whatever is possible to minimize lack of independence in the pseudo-replicates.
- For example, in the study of effect of CO_2 on plant growth, the researcher rearranged the plants in each growth chamber each day to mitigate effects of location in the chamber.

b. Be careful in analyzing and reporting results.
- Be open about the limitations of the study.
- Be careful not to over-interpret results.
- For example, in Example 2, the researcher could calculate what might be called "pseudo-confidence intervals" that would not be "true" confidence intervals, but which could be interpreted as giving a lower bound on the margin of error in the estimate of the quantity being estimated.

c. Consider the study as preliminary (for example, for giving insight into how to plan a better study), or as one study that needs to be combined with many others to give more informative results.

How can over-fitting be avoided? (p. 52)

As with most things in statistics, there are no hard and fast rules that guarantee success.

- However, here some guidelines.

- They apply to many other types of statistical models (e.g., multilinear, mixed models, general linear models, hierarchical models) as well as least squares regression.

1. ***Validate*** your model (for the mean function, or whatever else you are modeling) if at all possible. Good and Hardin (2006, p. 188) list three general types of validation methods:

i. Independent validation (e.g., wait till the future and see if predictions are accurate)
- This of course is not always possible.

ii. Split the sample.
- Use one part for model building, the other for validation.
- See item II(c) of Data Snooping for more discussion.

iii. Resampling methods.

- See Chapter 13 of Good and Hardin (2006), and the further references provided there, for more information.

2. Gather plenty of (ideally, well-sampled) data.

- If you are gathering data (especially through an experiment), be sure to consult the literature on *optimal design* to plan the data collection to get the tightest possible estimates from the least amount of data.

- For regression, the values of the explanatory variable (x values, in the above example) do not usually need to be randomly sampled; choosing them carefully can minimize variances and thus give tighter estimates.

- Unfortunately, there is not much known about sample sizes needed for good modeling.

    - Ryan (2009, p. 20) quotes Draper and Smith (1998) as suggesting that the number of observations should be at least ten times the number of terms; this may be overly optimistic.

    - Good and Hardin (2006, p. 183) offer the following conjecturally:

    "If $\boldsymbol{m}$ points are required to determine a univariate regression line with sufficient precision, then it will take at least $\boldsymbol{m^n}$ observations and perhaps

n!m" observations to appropriately characterize and evaluate a regression model with ***n*** variables."

3. Pay particular attention to transparency and avoiding over-interpretation in reporting your results.

- For example, be sure to state carefully what assumptions you made, what decisions you made, your basis for making these decisions, and what validation procedures you used.
- Provide (in supplementary online material if necessary) enough detail so that another researcher could replicate your methods.

REFERENCES FOR DAY 4:

American Statistical Association (1997), *Ethical Guidelines for Statistical Practice*, http://www.amstat.org/committees/ethics/index.html

Y. Benjamini and Y. Hochberg (1995), Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society, Series B (Methodological)*, Vol. 57 No. 1, 289 – 300

Y. Benjamini and D. Yekutieli (2001), The Control of the False Discovery Rate in Multiple Testing under Dependency, *The Annals of Statistics*, vol. 29 N. 4, 1165 - 1186.

Y. Benjamini and D. Yekutieli (2005), False Discovery Rate–Adjusted Multiple Confidence Intervals for Selected Parameters, *Journal of the American Statistical Association,* March 1, 2005, 100(469): 71-81

Buja, A. et al (2009), Statistical inference for exploratory data analysis and model diagnostics, *Phil. Trans. R. Soc. A*, vol 367, 4361 - 4383

Cook, R.D. and S. Weisberg (1999) *Applied Regression Including Computing and Graphics*, Wiley

Dallal, Jerry, 100 Independent 0.05 Level Tests For An Effect Where None Is Present, http://www.jerrydallal.com/LHSP/multtest.htm

> This simulates the results of 100 independent hypothesis tests, each at 0.05 significance level. Click the "test/clear" button to see the results of one set of 100 tests (that is, for one sample of data). Click the button two more times (first to clear and then to do another simulation) to see the results of another set of 100 tests (i.e., for another sample of data). Notice as you continue to do this that i) which tests give type I errors (i.e., are statistically significant at the 0.05 level) varies from sample to sample, and ii) which samples give type I errors for a given test varies from test to test. (To see the latter point, it may help to focus just on the first column.)

Freedman, D. A. (2005) *Statistical Models: Theory and Practice*, Cambridge

Freedman, D.A. (2006) "Statistical models for causation: What inferential leverage do they provide?" *Evaluation Review* vol. 30 pp. 691–713. Preprint at http://www.stat.berkeley.edu/%7Ecensus/oxcauser.pdf

Harris, A. H. S., R. Reeder and J. K. Hyun (2009), Common statistical and research design problems in manuscripts submitted to high-impact psychiatry journals: What editors and reviewers want authors to know, *Journal of Psychiatric Research*, vol 43 no15, 1231 -1234

Heffner, Butler, and Reilly (1996) Pseudoreplication Revisited, *Ecology* 77(8), pp. 2558 - 2562

Hochberg, Y. and Tamhane, A. (1987) *Multiple Comparison Procedures*, Wiley

Hopewell, S. et al (2009). Publication Bias in Clinical Trials due to Statistical Significance or Direction of Trial Result, *Cochrane Review* 2009, Issue 1; abstract available at www.thecochrane library.com

The authors conclude that "Trials with positive findings are published more often, and more quickly, than trials with negative findings."

S. H. Hurlbert (1984) Pseudoreplication and the design of ecological field experiments, *Ecological monographs* 54(2), pp. 187 – 211

Kass, Robert (2011) Statistical inference: The big picture, *Statistical Science*, to appear. Preprint available at http://www.imstat.org/sts/future_papers.html.

See also the discussion papers by Stephen Goodman, Hal Stern, Andrew Gelman, and Robert McCulloch, as well as Kass' rejoinder (all available at the same website.)

Lau, J. et al (2006) The case of the misleading funnel plot, *BMJ* 333 (7568), 597-600. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1570006/?tool=pmcentrez

Maxwell, S. E. and K Kelley (2011), Ethics and Sample Size Planning, Chapter 6 (pp. 159 - 183) in Panter, A. T. and S. K. Sterba, *Handbook of Ethics in Quantitative Methodology*, Routledge

Maxwell, S.E. (2004), The persistence of underpowered studies in psychological research: Causes, consequences, and remedies, *Psychological Methods* 9 (2), 147 - 163.

Miller, R.G. (1981) *Simultaneous Statistical Inference* 2nd Ed., Springer

Moher, D. et al (2010), CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials, BMJ 2010; 340:c869, open access at http://www.bmj.com/content/340/bmj.c869.full

Potcner and Kowalski (2004), How to Analyze a Split-Plot Experiment, Quality Progress, December 2004, pp. 67 – 74, http://www.minitab.com/uploadedFiles/Shared_Resources/Documents/Articles/analyze_split_plot_experiment.pdf

Rice Virtual Lab in Statistics, Robustness Simulation, http://onlinestatbook.com/stat_sim/robustness/index.html

R. Rosenthal (1979) The "file drawer problem" and tolerance for null results, *Psychological Bulletin,* Vol. 86, No. 3, 838-641.

J. Scargle (2000) Publication bias: The "file-drawer" problem in scientific inference, *Journal of Scientific Exploration*, Vol. 14, No. 1, pp. 91-106.

F. Song et al (2009), Extent of publication bias in different categories of research cohorts: a meta-analysis of empirical studies, *BMC Medical Research Methodology* 2009, 9:79, http://www.biomedcentral.com/1471-2288/9/79.

Reports on a meta-analysis of studies that examine a cohort of research studies for publication bias. In the studies examined, publication bias tended to occur in the form of not presenting results at conferences and not submitting them for publication. The paper also discusses different types of evidence for publication bias.

T. D. Sterling, W. L. Rosenbaum and J. J. Weinkam (1995), Publication Decisions Revisited: The Effect of the Outcome of Statistical Tests on the Decision to Publish and Vice Versa, *The American Statistician*, vol 49 No. 1, pp. 108 – 112.

Reviews the literature through 1995, and reports on an additional study indicating the existence of publication bias, with results reported in the literature showing statistical significance being over-represented compared to what would be expected (although the rate depended on the field). They also provide anecdotal evidence that papers may be rejected for publication on the basis of having a result that is not statistically significant.

A. M. Strasak et al (2007), The Use of Statistics in Medical Research, *The American Statistician*, February 1, 2007, 61(1): 47-55

G. van Belle (2008) *Statistical Rules of Thumb*, Wiley

Wainer, Howard (2011) Value-added models to evaluate teachers: A cry for help, *Chance* vol.24, No. 2. Available at http://chance.amstat.org/2011/02/value-added-models/

A nice discussion of the difficulties of statistical modeling in a topic of current wide interest.