APPENDIX TO DAY 3

Considerations on determining sample size to give desired power: *(pp. 23, 29)*

- The difference used in calculating sample size (i.e., the specific alternative used in calculating sample size) should be decided on the base of practical significance and/or "worst case scenario," depending on the consequences of decisions.

- Even when the goal is a hypothesis test, it may be wise to base the sample size on the width of a confidence interval rather than just ability to detect the desired difference: Even when power is large enough to detect a difference, the uncertainty, as displayed by the confidence interval, may still be too large to make the conclusions very credible to a knowledgeable reader.

- Determining sample size to give desired power and significance level will usually require some estimate of parameters such as variance, so will only be as good as these estimates.

    - These estimates usually need to be based on previous research, experience of experts in the field, or a pilot study.

    - In many cases, it may be wise to use a conservative estimate of variance (e.g., the upper bound of a confidence interval from a pilot study), or to do a sensitivity analysis to see how the sample size estimate depends on the parameter estimate. See Lenth (2001) for more details.

- Even when there is a good formula for power in terms of sample size, "inverting" the formula to get sample size from power is often not straightforward

    - This may require some clever approximation procedures.

    - Such procedures have been encoded into computer routines for many (not all) common tests.

    - See Russell Lenth's website or John C. Pezzullo's Interactive Statistics Pages for links to a number of online power and sample size calculators.

    - *Caution*: If you use software routines to calculate power, be sure it calculates *a priori* power, *not* *retrospective (or observed) power.* (See below)

- Good and Hardin (2006, p. 34) report that using the default settings for power and sample size calculations is a **common mistake** made by researchers.

- For *discrete* distributions, the "power function" (giving power as a function of sample size) is often saw-toothed in shape.

- A consequence is that software may not necessarily give the optimal sample size for the conditions specified.

- Good software for such power calculations will also output a graph of the power function, allowing the researcher to consider other sample sizes that might give be better than the default given by the software.

References for tests of equivalence (p. 27):

- Hoenig, John M. and Heisey, Dennis M. (2001
- Graphpad.com, Statistical Tests for Equivalence, http://www.graphpad.com/library/biostatsspecial/article_182.htm
- Lauchenbruch, P. A. (2001)

Note regarding Cohen's d (p. 29):

Figure 1 of Browne (2010) shows that, for the two-sample t-test, Cohen's classification of "large" d as 0.8 still gives substantial overlap between the two distributions being compared; d needs to be close to 4 to result in minimal overlap of the distributions.

Suggestions for dealing with the File Drawer Problem: *(p. 36)*

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 and professional societies that 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.

REFERENCES FOR DAY THREE:

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

Browne, Richard H. (2010). The t-Test p Value and Its Relationship to the Effect Size and P(X>Y), *The American Statistician*, February 1, 2010, 64(1), p. 31

Claremont Graduate University WISE Project Statistical Power Demo, http://wise.cgu.edu/powermod/power_applet.asp

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.)

Doshi P., M. Jones and T. Jefferson (2012). Rethinking credible evidence synthesis, *British Medical Journal* 344, Article Number: d7898 DOI: 10.1136/bmj.d7898

B. Efron (2010), Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction, Cambridge.

Festing, Michael, Statistics and animals in biomedical research, *Significance* Volume 7 Issue 4 (December 2010), http://www.significancemagazine.org/details/magazine/879779/Statistics-and-animals-in-biomedical-research-.html

Good and Hardin (2006 or 2010), *Common Errors in Statistics*, Wiley

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

Hoenig, John M. and Heisey, Dennis M. (2001), "The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis," *The American Statistician*, 55, 19-24

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."

Kilkenny et al, (2010) Improving bioscience research reporting: The ARRIVE guidelines for reporting animal research. PLoS Biology, 8, http://www.plosbiology.org/article/info%3Adoi%2F10.1371%2Fjournal.pbio.1000412

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

Lauchenbruch, P. A. (2001), Equivalence Testing, http://www.fda.gov/ohrms/dockets/ac/01/slides/3735s1_02_lachenbruch/index.htm

Lenth, Russell V. (2000), Two Sample-Size Practices that I Don't Recommend, comments from panel discussion at the 2000 Joint Statistical Meetings in Indianapolis, http://www.stat.uiowa.edu/%7Erlenth/Power/2badHabits.pdf

Lenth, Russell V. (2001) Some Practical Guidelines for Effective Sample Size Determination, *American Statistician*, 55(3), 187 – 193.
A discussion of many considerations in deciding on sample size. An early version and some related papers can be downloaded from his website (below)

Lenth, Russell, Power website, http://www.stat.uiowa.edu/~rlenth/Power/ Has several online applets for calculating power, some advice on using the applets, and links to some papers on power.

Lipsey, MW (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage.

Marshall E. (2011). Unseen world of clinical trials emerges from US database. *Science* 333:145.

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.

McClelland, Gary H. (2000) Increasing statistical power without increasing sample size, *American Psychologist* 55(8), 963 – 964

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

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.

Pezzullo, John C., Power, Sample Size and Experimental Design Calculations, *Interactive Statistics Pages*, http://statpages.org/#Power

> Has links to online power calculations; I'd suggest trying Russell Lenth's page (above) first.

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

Wuensch, K. L. (1994). Evaluating the Relative Seriousness of Type I versus Type II Errors in Classical Hypothesis Testing, http://core.ecu.edu/psyc/wuenschk/StatHelp/Type-I-II-Errors.htm

Wuensch, K.L. et al (2003), "Retrospective (Observed) Power Analysis, Stat Help website, http://core.ecu.edu/psyc/wuenschk/stathelp/Power-Retrospective.htm