## APPENDIX FOR DAY 4

### *Suggestions for data snooping professionally and ethically* *(p. 17)*

**→ Be sure to see the *Suggestions for Researchers* posted at the course website for additional information and references.**

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 overall 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 the overall Type I error rate 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 what overall Type I error rate (or false discovery rate) and power would be appropriate (considering the consequences of these types of errors in the situation you are studying).
  - Be sure to allow some portion of the overall Type I error rate 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 or how you will handle missing data.
- *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
  - Depending on the aims of the study and the size of the sample, this might be done by random assigning students within each school to the discovery or confirmatory set, or by randomly assigning school districts (plus all their students in the data set) to the discovery or confirmatory set.
- 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. Before doing any research, register your research plan, to help keep yourself honest as well as to insure transparency in reporting your research.

4. Keep notes on any decisions you made while collecting or analyzing data.

5. 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.

## V. METHODS FOR CHECKING MODEL ASSUMPTIONS

***Examples of Checking Model Assumptions Using Well-established Facts or Theorems***

*Recall*:

- This is not possible very often.
- Here, "well established" means *well established by empirical evidence and/or sound mathematical reasoning*.
- This is *not* the same as "well accepted," since sometimes things may be well accepted without sound evidence or reasoning.

1. *Using laws of physics*

Hooke's Law says that when a weight that is not too large (below what is called the "elastic limit") is placed on the end of a spring, the length of the (stretched) spring is approximately a linear function of the weight.

- This tells us that if we do an experiment with a spring by putting various weights (below the elastic limit) on it and measuring the length of the spring, we are justified in using a linear model,

  $$\text{Length} = A \times \text{Weight} + B$$

*2. Using the Central Limit Theorem*

One form of The Central Limit Theorem says that for most distributions, a linear combination (e.g., the sum or the mean) of a large enough number of independent random variables is approximately normal.

- Thus, if a random variable in question is the sum of independent random variables, then it's usually safe to assume that the variable is approximately normal.
- For example, adult human heights (at least if we restrict to one sex) are the sum of many heights: the heights of the ankles, lower legs, upper legs, pelvis, many vertebrae, and head.
  - Empirical evidence suggests that these heights vary roughly independently (e.g., the ratio of height of lower leg to that of upper leg varies considerably).
  - Thus it's plausible by the Central Limit Theorem that human heights are approximately normal.
  - This in fact is supported by empirical evidence.
- *Caution*: "Most" is not "all." There are some distributions for which the central limit theorem is not valid. One notable exception is distributions that are "heavy-tailed" (also called *leptokurtic*). Such distributions occur in certain situations, such as seed dispersal in biology.
  - Try it on the Sampling Distribution demo.
- The Central Limit Theorem can also be used to reason that some distributions are approximately *lognormal* -- that is, that the logarithm of the random variable is normal.
  - For example, the distribution of a pollutant might be determined by successive independent dilutions of an original emission.
  - This translates into mathematical terminology by saying that the amount of pollution (call this random variable Y) in a given small region is the *product* of independent random variables.
  - Thus logY is the *sum* of independent random variables.
  - If the number of successive dilutions is large enough, the reasoning above shows that logY is approximately normal, and hence that Y is approximately lognormal.

***Using Plots to Check Model Assumptions***

*Overall Cautions*:

1. Unfortunately, these methods are typically better at telling you when the model assumption does *not* fit than when it does.

2. There's inherently an element of subjectivity in using model-checking plots.

- Some people are more likely than others to "see things that aren't really there."
- Buja et al (2009) have recently proposed some protocols for taking this into account.
- The smaller the sample size, the more of a problem this will be.

3. Different techniques have different model assumptions, so will need different model checking plots.

- Be sure to consult a good reference *for the particular technique* you are considering using.

*General Rule of Thumb*:

1. First check any independence assumptions;
2. then any equal variance assumption;
3. then any assumption on distribution (e.g., normal) of variables.

*Rationale*: Techniques are usually least robust to departures from independence, and most robust to departures from normality.

- See van Belle (2008), pp. 173 - 177 and the references given there for more detail.

***Suggestions and Guidelines for Checking Independence Assumptions***

Independence assumptions are usually formulated in terms of error terms rather than in terms of the outcome variables.

- For example, in simple linear regression, the model equation is
$$Y = \alpha + \beta x + \varepsilon,$$
where Y is the outcome (response) variable and $\varepsilon$ denotes the error term (also a random variable).
- It's the error terms that are assumed to be independent, not the values of the response variable.
- In more detail: The model assumptions are
  - $E(Y|x) = \alpha + \beta x$
  - For each x, $\varepsilon$ is normal with mean 0 and standard deviation $\sigma$.
  - The values of $\varepsilon$ for different x's are independent.

We *do <u>not</u> know* the values of the error terms $\varepsilon$, so we can only plot the residuals $e_i$ (defined as the observed value $y_i$ minus the fitted value, according to the model), which approximate the error terms.

*Rule of Thumb*: To check independence, plot residuals against:

- Any time variables present (e.g., order of observation)
- Any spatial variables present,
- Any variables used in the technique (e.g., factors, regressors)

A pattern that's *not* random suggests *lack* of independence.

*Rationale*: Dependence on time or on spatial variables is a common source of lack of independence, but the other plots might also detect lack of independence.

*Comments*:

1. Since time or spatial correlations are so frequent, it is important when making observations to *record any time or spatial variables that could conceivably influence results.*

- This not only allows you to make the residual plots to detect possible lack of independence, but also allows you to change to a technique incorporating additional time or spatial variables if lack of independence is detected in these plots.

2. Since it's known that the residuals sum to zero (in least squares regression), they're *not* independent, so the plot is really a very rough *approximation*.

3. Some models only require that errors are uncorrelated, not independent; model checks are the same as for independence.

## *Suggestions for Checking Model Assumptions of Equal Variance or Normality*

*Caution:* These suggestions are things you should do, but they are <u>not</u> guaranteed to find all departures from equal variance or normality.

*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.
- See R. Wayne Oldford (2016), Self-Calibrating Quantile-Quantile Plots, *The American Statistician 70*, 74 – 90 for more details. (You might be able to get a copy via https://www.researchgate.net/profile/R_Oldford)

*Checking for Linearity*

When considering a *simple linear regression model*, it's important to check the linearity assumption -- i.e., that the *conditional means* of the response variable are a linear function of the predictor variable.

Graphing the response variable vs. the predictor can often give a good idea of whether or not this is true.

However, one or both of the following refinements may be needed:

1. Plot residuals (instead of response) vs. predictor.
    - A non-random pattern suggests that a simple linear model is not appropriate; you may need to transform the response or predictor, or add a quadratic or higher term to the mode.

2. Use a scatterplot smoother such as lowess (also known as loess) to give a visual estimation of the conditional mean.
    - Such smoothers are available in many regression software packages.
    - *Caution*: You may need to choose a value of a smoothness parameter. Making it too large will over smooth; making it too small will not smooth enough.

When considering a *linear regression with just two terms*, plotting response (or residuals) against the two terms (making a three-dimensional graph) can help gauge suitability of a linear model, especially if your software allows you to rotate the graph.

***Caution*: It's *not* possible to gauge from scatterplots whether a linear model in *more than two predictors* is suitable.

- One way to address this problem is to try to transform the predictors to approximate multivariate normality.
    - See, e.g., Cook and Weisberg (1999), pp. 324 – 329.
- Multivariate normality will ensure not only that a linear model *is* appropriate for all (transformed) predictors together, but *also* that a linear model is appropriate even when some transformed predictors are dropped from the model.

*Note*: It's a **common mistake** to assume that if a linear model fits with all predictors included, then a linear model will still fit when some predictors are dropped. (p. 46)

*Example:* 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.

### *Additional Comments about Fixed and Random Factors in ANOVA* *(p. 48)*

- The standard methods for analyzing random effects ANOVA 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.
- Sometimes (especially in mixed models) ANOVA software will give an "estimates" of a random effect for particular levels, but these should be considered "predictions" rather than estimates,_.

### *Suggestions for Dealing with Pseudoreplication* (p. 55)

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.

### *Trying to avoid over-fitting (p. 60)*

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 $\boldsymbol{n!m^n}$ observations to appropriately characterize and evaluate a regression model with $\boldsymbol{n}$ 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:**
**See also the *Suggestions for Researchers* on the course website for additional suggestions and references.**


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

Christoph Bernau, Markus Riester, Anne-Laure Boulesteix, Giovanni Parmigiani, Curtis Huttenhower, Levi Waldron, and Lorenzo Trippa (2014). Cross-study validation for the assessment of prediction algorithms. Bioinformatics, 30(12):i105-i112 [See also Donoho (2015) p. 30 for a summary.]

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

Couzin-Frankel, Jennifer (2013), The Power of Negative Thinking, *Science* 342, 4 October, 2014, pp. 68 – 69, *http://www.sciencemag.org/content/342/6154/68.full*

Donoho, David (2015) 50 Years of Data Science (Based on a presentation at the Tukey Centennial workshop, Princeton NJ Sept 18 2015), http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf

Donoho, David and J. Jin (2015), Higher Criticism for Large-Scale Inference, Especially for Rare and Weak Effects, *Statistical Science 30*, 1-25, preprint at http://arxiv.org/abs/1410.4743

A review article on Higher Criticism, a method developed (based on an idea of Tukey) by the authors in 2004 for dealing with multiple inference in large-scale data studies.

B. Efron (2010), Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction, Cambridge. (Table of contents and prologue at statweb.stanford.edu/~ckirby/brad/other/2010LSIexcerpt.pdf)

Flom, Peter (2012), The Perils of Categorizing Continuous Variables, http://www.statisticalanalysisconsulting.com/the-perils-of-categorizing-continuous-variables/

Flom, Peter and David Cassell (2007) "Stopping Stepwise: Why stepwise and similar selection methods are bad and what you should use", NorthEast SAS Users Group, http://www.nesug.org/proceedings/nesug07/sa/sa07.pdf

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

Gelman, Andrew (2013), Preregistration of studies and mock reports, *Political Analysis 21*, 40-41. http://www.stat.columbia.edu/~gelman/research/published/mock.pdf

Gelman, Andrew (2013) Reregistration of Studies and Mock Reports, *Political Analysis 21*:40–41, doi:10.1093/pan/mps032

Gelman, A., J. Hill and M. Yajima (2012). Why We (Usually) Don't Have to Worry About Multiple Comparisons, *J. Res. on Educational Effectiveness*, 5: 189 – 211,

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*. Technical report, Department of Statistics, Columbia University, New York, NY. , http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf. [For a shorter published version for a more general audience, see Gelman and Loken (2014b)]

Gelman, Andrew and Eric Loken (2014a) "The AA tranche of subprime science, Ethics and Statistics column", *Chance vol 27.1*, pp. 51 – 56, http://www.stat.columbia.edu/~gelman/research/published/ChanceEthics10.pdf; see also discussion at http://andrewgelman.com/2014/01/20/aaa-tranche-subprime-science/

Gelman, Andrew and Eric Loken (2014b), The Statistical Crisis in Science, *American Scientist*, November-December 2014, Vol 102, Number 6, p. 460, available at http://www.americanscientist.org/issues/feature/2014/6/the-statistical-crisis-in-science or http://www.stat.columbia.edu/~gelman/research/published/ForkingPaths.pdf]

Goeman, J. and A. Solari (2011), Multiple Testing for Exploratory Research (with discussion and rejoinder), *Statistical Science v. 26 no.*4, pp. 584 – 612, available from Project Euclid at https://projecteuclid.org/euclid.ss/1330437927

Humphreys et al (2013), Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration, *Political Analysis 21*, 1-20, http://pan.oxfordjournals.org/content/21/1/1.full [See also Gelman (2013).]

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

Humphreys, Macartan, Raul Sanchez de la Sierra, and Peter van der Windt. 2013. Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration. Political Analysis 21:1–20

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

Ioannidis JPA (2005) Why Most Published Research Findings Are False. PLoS Med 2(8): e124. doi:10.1371/journal.pmed.0020124, available at http://www.plosmedicine.org/article/info:doi/10.1371/journal.pmed.0020124

Ioannidis JPA (2008) Why most discovered true associations are inflated, *Epidemiology* 19, 640 – 648.

Ioannidis JPA (2012) Scientific inbreeding and same-team replication: Type D personality as an example, *Journal of Psychosomatic Research* 73, 408 – 410

S. Jauhar, P. J. McKenna, J. Radua, E. Fung, R. Salvador, K. R. Laws (2014) The British Journal of Psychiatry 204 (1) 20-29, DOI: 10.1192/bjp.bp.112.116285, http://bjp.rcpsych.org/content/204/1/20.full.

See also http://keithsneuroblog.blogspot.com/2015/05/science-politics-of-cbt-for-psychosis.html for brief discussion of the paper and a link to an invited hour address before the British Psychological Society about the background and results of the paper.

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.)

Klaus, B. and K. Strimmer (2013) Signal identification for rare and weak features: higher criticism or false discovery? *Biotatistics 14* (1), 129 – 143, available at http://biostatistics.oxfordjournals.org/content/14/1/129

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

Lee and Rubin (2016) Evaluating the Validity of Post-Hoc Subgroup Inferences : A Case Study, *The American Statistician* 70, pp. 39 - 46

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

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

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

Muller et al (2006), FDR and Bayesian Multiple Comparisons Rules, Proc. Valencia/ISBA 8th World Meeting on Bayesian Statistics, http://biostats.bepress.com/jhubiostat/paper115/

Pashler, Harold and Christine Harris, Is the Replicability Crisis Overblown?, *Psychological Science 7*, 531-536, free download at http://pps.sagepub.com/content/7/6/531

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

Silberzahnn, R. and E. L. Uhlmann (2015), Crowdsourced research: Many hands make tight work, *Nature 526*, 189-191.

U. Simonsohn et al (2013) P-curve: A Key to the File Drawer, forthcoming, *Journal of Experimental Psychology: General*.

Proposes a method to help detect selective reporting (whether publication bias or p-hacking). See also the authors' website, http://www.p-curve.com/, which has a link to the paper, a related web app, and supplemental materials.

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.

Zimmerman, D.W. (2004) A note on preliminary tests of equality of variances, *Britsh Journal of Mathematical and Statistical Psychology 57*, 173 - 181