# Commented Literature Review

Fabian Dablander

March 21, 2015

## 1 General Linear Model

- Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American psychologist*, *12*(11), 671

- Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin*, *70*(6p1), 426

- Kéry, M. (2010). *Introduction to WinBUGS for ecologists: Bayesian approach to regression, ANOVA, mixed models and related analyses*. Academic Press

  **The first four chapters are arguably the best introduction to the General Linear Model that there is. Featuring R code - most prominently the *model.matrix* function - the author introduces the t-test, ANOVA, multiple regression etc. as being basically the same thing. Highly recommended.**

- Poline, J.-B. & Brett, M. (2012). The General Linear Model and fMRI: does love last forever? *Neuroimage*, *62*(2), 871–880

  **This is a hilarious and well written paper. It gives an easy to understand introduction to the GLM using simple notation, and offers some psychological investigations on why neuroimaging people love it (one can easily incorporate a lot of different things, parameters for low frequency drift, motion correction etc.) The appendix derives $\hat{\beta}$ as the ordinary least square solution, which in the case of the GLM is the best linear unbiased estimate.**

- Monti, M. M. (2011). Statistical analysis of fMRI time-series: a critical review of the GLM approach. *Frontiers in human neuroscience*, *5*(article 28)

  **This paper discusses the problems of the "Massive Univariate Approach" in-depth, ranging from heteroscedasticity to violations of additivity in modeling the HRF-response. Single subject analysis as well as multiple**

**subject analysis (using the summary statistics approach) are discussed. In general, this paper shows how a research community's creativity is boosted when a beloved tool - the General Linear Model - gets into (deep) trouble.**

## 2 Deficits of Frequentist Statistics


Bakan, D. (1966). The test of significance in psychological research. *Psychological bulletin*, *66*(6), 423

Royall, R. M. (1986). The effect of sample size on the meaning of significance tests. *The American Statistician*, *40*(4), 313–315

Cohen, J. (1994a). The earth is round (p $<$ .05). *American Psychologist*, *49*(12), 997–1003

Cohen, J. (1994b). The earth is round (p $<$ .05): rejoinder. *American Psychologist*, *49*(12), 1103

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current directions in psychological science*, 161–171

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological methods*, *5*(2), 241

Sellke, T., Bayarri, M., & Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *The American Statistician*, *55*(1), 62–71

Haller, H. & Krauss, S. (2002). Misinterpretations of significance: a problem students share with their teachers. *Methods of Psychological Research*, *7*(1), 1–20

Lee, M. D. & Wagenmakers, E.-J. (2005). Bayesian statistical inference in psychology: comment on Trafimow (2003).

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic bulletin & review*, *14*(5), 779–804


**In my opinion the best exposition of the deficits of null hypothesis significance testing. Uses extremely illustrative examples to show that p values depend on data never observed (sampling distribution), on the subjective intentions of researchers (binomial vs. negative binomial), can lead to highly inflated Type I error rates due to optional stopping (see also Simmons, Nelson, and Simonsohn, 2011), do not quantify statistical evidence, and differ in their interpretation in small versus large sample sizes (the p-postulate is false).**

**Also introduces Bayesian estimation and hypothesis testing on a coin toss example. Goes on to suggest the Bayesian Information Criterion (BIC) as simple alternative to significance testing.**

$$BIC(H_i) = -2 \log L_i + k_i \log n$$

**where $L_i$ is the maximum likelihood for model $H_i$, $k_i$ is the number of free parameters and $n$ is the sample size. Approximating the probability of the data given model $M_i$ we can write $P_{BIC}(D|H_i) = \exp \frac{-BIC(H_i)}{2}$ such that**

$$BF_{01} \sim \frac{P_{BIC}(D|H_0)}{P_{BIC}(D|H_1)} = \exp \frac{BIC(H_0) - BIC(H_1)}{2}$$

Wagenmakers, E.-J., Lee, M., Lodewyckx, T., & Iverson, G. J. (2008). Bayesian versus frequentist inference. In *Bayesian evaluation of informative hypotheses* (pp. 181–207). Springer

**Is similar to Wagenmakers, 2007 in that it discusses the disadvantages of null hypothesis significance testing. Additionally notes that the p-value cannot be used to compare non-nested models, and that it relies on ad-hoc principles since it does not have as solid a foundation as Bayesian inference (probability theory) and thus does not specify a unique solution to every statistical problem. Discusses the advantages of Bayesian inference, i.e. coherence, automatic parsimony, extension to non-nested models, no optional stopping, possibility to quantify evidence in favour of the null etc. Features a hilarious anecdote about the subjective nature of the p-value.**

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic bulletin & review*, *21*(5), 1157–1164

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2014). The fallacy of placing confidence in confidence intervals. *submitted*

Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., & Wagenmakers, E.-J. (2014). The p <. 05 rule and the hidden costs of the free lunch in inference. *Manuscript under review*

Hoenig, J. M. & Heisey, D. M. (2001). The abuse of power. *The American Statistician*, *55*(1)

Wagenmakers, E.-J., Verhagen, J., Ly, A., Bakker, M., Lee, M. D., Matzke, D., ... & Morey, R. D. (2014). A power fallacy. *Behavior research methods*, 1–5

**Power is a pre-experimental concept that averages over all possible outcomes of an experiment, only one of which is actually observed. Thus low powered experiment can lead highly valuable information, while high powered experiments might be uninformative. As Ioannidis Ioannidis, 2005 and others Button et al., 2013 have shown, low power effects the likelihood that a significant finding actually reflects a true effect; because using Bayes' theorem**

$$\frac{P(H_1|p<\alpha)}{P(H_0|p<\alpha)} = \frac{P(p<\alpha|H_1)}{P(p<\alpha|H_0)} \times \frac{P(H_1)}{P(H_0)}$$

**where $P(p<\alpha|H_1)$ is $1-\beta$ and $P(p<\alpha|H_0)$ is $\alpha$. Thus $\frac{1-\beta}{\alpha}$ is the extent to which the observation $p<\alpha$ changes the prior odds that $H_1$ rather than $H_0$ is true.**

**However, when the actual data is observed, we can go beyond measures of diagnosticity such as power. Assume two urns, $H_0$ with nine green balls and one blue ball, $H_1$ with nine green balls and one orange ball. Your task is decide which urn you draw from (the true urn is $H_1$). In one experiment, you only draw on ball - yielding statistical power of 0.10. However, you were lucky and drew an orange ball, which means that the urn is $H_1$. Although low powered, your experiment yielded decisive support. Conversely, you could have drawn 22 balls, yielding statistical power of 0.90, all of which where green. Despite high power, your experiment would have been uninformative. The paper features another example using a t-test; go read it!**

# 3 Bayesian Statistics

## 3.1 Motivation

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological review*, *70*(3), 193

**This is the earlierst exposition of Bayesian ideas specifically written for psychologists. It is quite extensive, in-depth and gives a broad overview about Bayesian merits. It is considered a classic.**

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & Van Der Maas, H. L. (2011). Why psychologists must change the way they analyze their data: the case of psi: comment on Bem (2011).

Dienes, Z. (2011). Bayesian versus orthodox statistics: which side are you on? *Perspectives on Psychological Science*, *6*(3), 274–290

**Clarifies the differences between frequentist and Bayesian statistics with respect to optional stopping, planned versus post hoc comparisons and multiple testing. Discusses the likelihood principle and how classical statistics violates it. Argues that if you want to be rational (which scientist would not?) you should be using Bayesian inference.**

Rouder, J. N., Morey, R. D., & Pratte, M. S. (2013). Hierarchical Bayesian models. *Practice*, *1*(5), 10

Wagenmakers, E.-J., Verhagen, A. J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., & Morey, R. D. (in press). The need for Bayesian hypothesis testing in psychological science. In *Psychological science under scrutiny: recent challenges and proposed solutions*. John Wiley and Sons

Dienes, Z. (under revision). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*

## 3.2 Priors

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, 187–192

**The "Jeffreys-Lindley" paradox is an intriguing finding. Assume a normal model $N(\theta, \sigma^2)$ with known variance $\sigma^2$, $H_0 : \theta = \theta_0$ and $H_1 : \theta \neq \theta_1$. For many prior distributions over $\theta$ and any $\alpha \in [0, 1]$ we can find a sample size $N$ such that**

- the sample mean $\hat{x}$ is significantly different from $\theta_0$ at level $\alpha$
- the posterior probability that $\theta = \theta_0$ is at least as big as $1 - \alpha$

**see also Sprenger, 2013 and Robert, 2014. Stated differently, while the p-value would lead to the rejection of $H_0$, the Bayes factor- under certain diffuse prior distributions - would lead to acceptance of $H_0$. It also shows that as the prior on $\theta$ gets more diffuse, the Bayes factor favours $H_0$ without bound. It is important to note that the p-value does not look at the data under $H_1$, while the Bayes factor quantifies evidence under $H_0$ and $H_1$ - the marginal likelihoods (this is one reason why in frequentist statistics one cannot accept $H_0$).**

Liu, C. C. & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*(6), 362–375

**Shows on a simple example how the Bayes factor is influenced by the prior. The marginal likelihood, i.e. the probability of the data given some Model $M$ is $p(D|M) = \int p(D|\theta, M)p(\theta|M)\mathrm{d}\theta$ where $p(D|\theta, M)$ is the likelihood and $p(\theta|M)$ is the prior over the parameter vector $\theta$. Stated differently, the marginal likelihood is the sum of the likelihood evaluated at each value of the model parameters weighted by the prior over the**

**parameters. Parameter estimation is not heavily influenced by prior assumptions. Other methods of model comparison such as the *posterior predictive loss*, the *Deviance information criterion* and the *posterior likelihood ratio* are discussed, all of which capitalize on the robustness of the posterior relative to the prior. When using the Bayes factor, make sure to run a sensitivity analysis; that is, see if the conclusions change when using a variety of different priors.**

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*(6), 491–498

**Argues that the marginal likelihood - and thus the Bayes factor - is an appropriate measure for evaluating psychological models precisely because of its sensitivity to the prior. Priors instantiate psychological theory and restrict the parameter values in the formal model; they are thus the antidote to the *Greek letter syndrome*, which leads sufferers to not consider the meaning of the parameters and ignore the underlying theory they instantiate. Concludes in arguing that one does not need sensitivity analysis, since the prior instantiates theory (it is not an arbitrary assumption). Contrary to Liu and Aitkin, 2008, users of posterior measures should use sensitivity analysis. Note that the paper talks about evaluation of formal mathematical models that instantiate theories, not models used for data analysis like regression.**

## 3.3 Bayes Factor

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic bulletin & review*, *16*(2), 225–237

Wetzels, R., Raaijmakers, J. G., Jakab, E., & Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: a flexible WinBUGS implementation of a default Bayesian t test. *Psychonomic bulletin & review*, *16*(4), 752–760

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive psychology*, *60*(3), 158–189

**In general, the Bayes factor - or more precisely the marginal likelihood - is hard to compute. An incredible mathematical result, the Savage-Dickey density ratio, allows one to calculate the Bayes factor in one go:**

$$BF_{01} = \frac{p(D|H_0)}{p(D|H_1)} = \frac{p(\Phi = \Phi_0|D, H_1)}{p(\Phi = \Phi_0|H_1)}$$

**This only works for nested models (a constraint also for the p-value), where $H_0$ posits that $\Phi = \Phi_0$, i.e. some substantive parameter is zero. Graphically, Savage-Dickey is the ratio between the height of the posterior and the height of the prior at the parameter value of interest. One clearly sees how the prior affects the Bayes factor, see also Liu and Aitkin, 2008, and that *nuisance parameters* irrelevant - they are present in both models and cancel out. The paper reviews much of Bayesian inference on a binomial example and features two real world examples applying Savage-Dickey. A must read; but see R. D. Morey, Rouder, Pratte, and Speckman, 2011 for a better computational approach.**

Rouder, J. N. & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, *47*(6), 877–903

**Enlightening paper. Frequentist statistics cannot support $H_0$ - absence of evidence is not evidence of absence - since significance testing is asymmetric; alternative models are not specified; free lunch problem, see also Rouder et al., 2014. This leads to the p-value being biased against $H_0$, see also Wagenmakers, 2007; Rouder et al., 2009. The Bayes factor is introduced via example of linear regression.**

**Instead of only allowing for point alternative hypotheses - like likelihood-ratio tests do -, Bayes factor allows for interval alternatives; the alternative becomes the average likelihood of the point alternatives weighted by the prior. By reparameterizing the model in terms of standardized effect size with $\beta \sim Cauchy(s)$ the Bayes factor becomes invariant to transformations. Additional beneficial properties include *Consistency*, i.e. if $M_1$ then $B_{10} \sim \inf$; conversely if $M_0$ then $B_{10} \sim 0$. Note that p-values are not consistent, since they are randomly distributed under $H_0$; and *Consistency in information*, i.e. data should affect the Bayes factor only through $R^2$.**

**Generalizes the default prior to multiple regression. Cauchy is computationally hard, so a multivariate normal prior is used for the slope parameters, with an inverse-gamma specified over the variance of the standardizes slope. After a short example, the authors discuss three concerns: *the null model is never true*, *all models are wrong* and the *subjective nature* of Bayes.**

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374

Wetzels, R. & Wagenmakers, E.-J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. *Psychonomic bulletin & review*, *19*(6), 1057–1064

Ly, A., Verhagen, A. J., & Wagenmakers, E.-J. (2014). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *submitted*

**Introduces the "default" or "objective" priors developed by Harrold Jeffreys to have certain desirable properties. Introduces them by means of the t-test and the Pearson correlation. Very in-depth, not light on mathematics.**

## 3.4 Parameter Estimation

Kruschke, J. K. (2013). Bayesian estimation supersedes the t-test. *Journal of Experimental Psychology: General*, *142*(2), 573

Wagenmakers, E.-J., Lee, M., Rouder, J. N., & Morey, R. D. (2014). Another statistical paradox. *submitted*

## 3.5 Model Comparison

Vandekerckhove, J., Matzke, D., & Wagenmakers, E.-J. (2014). Model comparison and the principle of parsimony. *Oxford Handbook of Computational and Mathematical Psychology. Oxford University Press, Oxford*

## 3.6 Applications

Rouder, J. N. & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*, *18*(4), 682–689

Morey, R. D. & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological methods*, *16*(4), 406

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: an empirical comparison using 855 t-tests. *Perspectives on Psychological Science*, *6*(3), 291–298

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in psychology*, *5*

Wagenmakers, E.-J., Verhagen, A., & Ly, A. (2015). How to quantify evidence for the absence of a correlation. *submitted*

**Builds on work by Donnellan, Lucas, and Cesario, 2014 who in nine experiments with over 3.000 participants in total failed to replicate Bargh and Shalev, 2012. Since one cannot within the framework of null hypothesis significance testing support the null hypothesis, Wagenmakers**

**and colleagues use Bayesian inference to do just that. Also excellent exposition on how high powered experiments do not necessarily yield high evidential value, see also Wagenmakers, Verhagen, et al., 2014. Three studies consisting of a sample size of 210, 494 and 553 participants were uninformative!**

Nieuwenstein, M. R., Wierenga, T., Morey, R. D., Wicherts, J. M., Blom, T. N., Wagenmakers, E.-J., van Rijn, H., et al. (2015). On making the right choice: a meta-analysis and large-scale replication attempt of the unconscious thought advantage. *Judgment and Decision Making*, *10*(1), 1–17

Verhagen, J. & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*(4), 1457

### 3.7 Problems?

Sanborn, A. N. & Hills, T. T. (2014). The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychonomic bulletin & review*, *21*(2), 283–300

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic bulletin & review*, *21*(2), 301–308

## 4 Software

Morey, R. (2015). Bayesfactor: 0.9.11 CRAN. doi:10.5281/zenodo.16085

**This is an awesome R package written by Richard Morey. It does all the GLM things and contingency tables. Generalized Linear Models such as logistic regressions have yet to be implemented. Checkout the excellent documentation which features many examples.**

Love, J., Selker, R., Marsman, M., Jamil, T., Verhagen, A. J., Ly, A., ... & Wagenmakers, E.-J. (n.d.). JASP (Version 0.6.5)[Computer Software]

**JASP is a low fat alternative to SPSS, and a Bayesian trojan horse. It is easy and intuitive to use. Checkout the homepage which also includes materials from workshops.**

## 5 Statistical Power

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of abnormal and social psychology*, *65*(3), 145–153

Sedlmeier, P. & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*(2), 309

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

Cumming, G. & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*(4), 532–574

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376

Fraley, R. C. & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PloS one*, *9*(10), e109019

Yarkoni, T. (2009). Big correlations in little studies: Inflated fMRI correlations reflect low statistical power—Commentary on Vul et al.(2009). *Perspectives on Psychological Science*, *4*(3), 294–298

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124

**Very famous paper arguing how publication bias, false incentives and low power lead to most published research being false. Introduces a simple formula for the posterior predictive value of studies $PPV = \frac{(1-\beta)R}{(1-\beta)R+\alpha}$, i.e. the probability of a research finding being true, where $R$ are the pre-study odds of the hypothesis being true, $\beta$ is the Type II error rate, and $\alpha$ is the Type I error rate. For further reading on how low power undermines evidential values of studies, see also Button et al., 2013**

# 6 Miscellaneous

## 6.1 p-hacking

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 0956797611417632

**Introduces the term "p-hacking", that is exploiting one's freedom in data collection, statistical analysis and reporting in order to increase the chance of a significant finding (which then will most likely turn out to be a false positive). Demonstrates that basically anything can be presented as significant, using an experimental study and simulations.**

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, 0956797611430953

**Asking over 2.000 psychologists, report the prevalence of "questionable research practices" such as "failing to report all dependent measures in a study" ($\sim 64\%$), "optional stopping" ($\sim 56\%$), "harking" ($\sim 31\%$) etc.**

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. *Downloaded January, 30*, 2014

**Beautiful paper arguing that "p-hacking" can be an entirely unconscious process where researchers have only the best intentions in mind! Draws on examples from the literature to discuss a hidden multiple comparison problem - the garden of forking paths. Psychological studies, although presented as confirmatory, are frequently exploratory in nature.**

## 6.2 Historical Notes on Modern Statistics

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. *A handbook for data analysis in the behavioral sciences: Methodological issues*, 311–339

**Excellent, must read paper about the "inference revolution". Using a Freudian analogy, describes R.A. Fisher as the Ego, specifying no alternative hypothesis, having no fixed $\alpha$-level and interpreting the p-value as strength of evidence. The Neyman-Pearson approach to statistical testing is described as the Superego, including an alternative hypothesis, specifying a pre-fixed $\alpha$ and $\beta$-level and interpreting the p-value as either being above the treshold or below - not as strength of evidence, but as decision to act: either reject the null hypothesis, or stay in limbo. Finally, the Freudian Id is Bayes, giving us what we really want: the probability of the hypothesis given the data $P(H|D)$. The p-value *cannot* achieve this; it merely is the probability of the data given data as extreme or more extreme as the once observed, $P(D|H)$. Incredibly fun read.**

Berger, J. O. et al. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, *18*(1), 1–32

Gigerenzer, G. & Marewski, J. N. (2014). Surrogate science the idol of a universal method for scientific inference. *Journal of Management*, 0149206314547522

**Recent exposition of statistical inference in general, and the problems with dogma. Also discusses Bayesian statistics.**