On the logic and history of statistical tests

Prof. Dr. Uwe Saint-Mont

Nordhausen University of Applied Sciences

Munich, 23 March 2016

Theory and data

Every experiment may be said to exist only to give the **facts** *a chance of disproving the null* **hypothesis**. *(Fisher 1935)*

Basic setup:

1. General tier: Population $X \sim P_H$
2. Concrete tier: (Independent) sample $x_1, \ldots, x_n$

Deterministic conclusion

Suppose X is a discrete random variable, and there is the following situation:

1. $P_H(X = x) = 0$, i.e., given H, the data x cannot be observed
2. Observation: x

Conclusion: H is wrong, H cannot be the case

Almost deterministic conclusion

Suppose X is a discrete random variable, and there is the following situation:

1. $P_H(X = x) = \varepsilon$ where $\varepsilon > 0$ small
2. Observation: x

Conclusion: **Either** an exceptionally rare chance has occurred, **or** the [hypothesis] is not true.

(Fisher 1956, his emphasis)

What is small?

Information theory confirms this intuition: $I(p) = -\log p$

Thus $I(1) = 0$, if $q < p$ then $I(p) < I(q)$, and $I(0) = \infty$

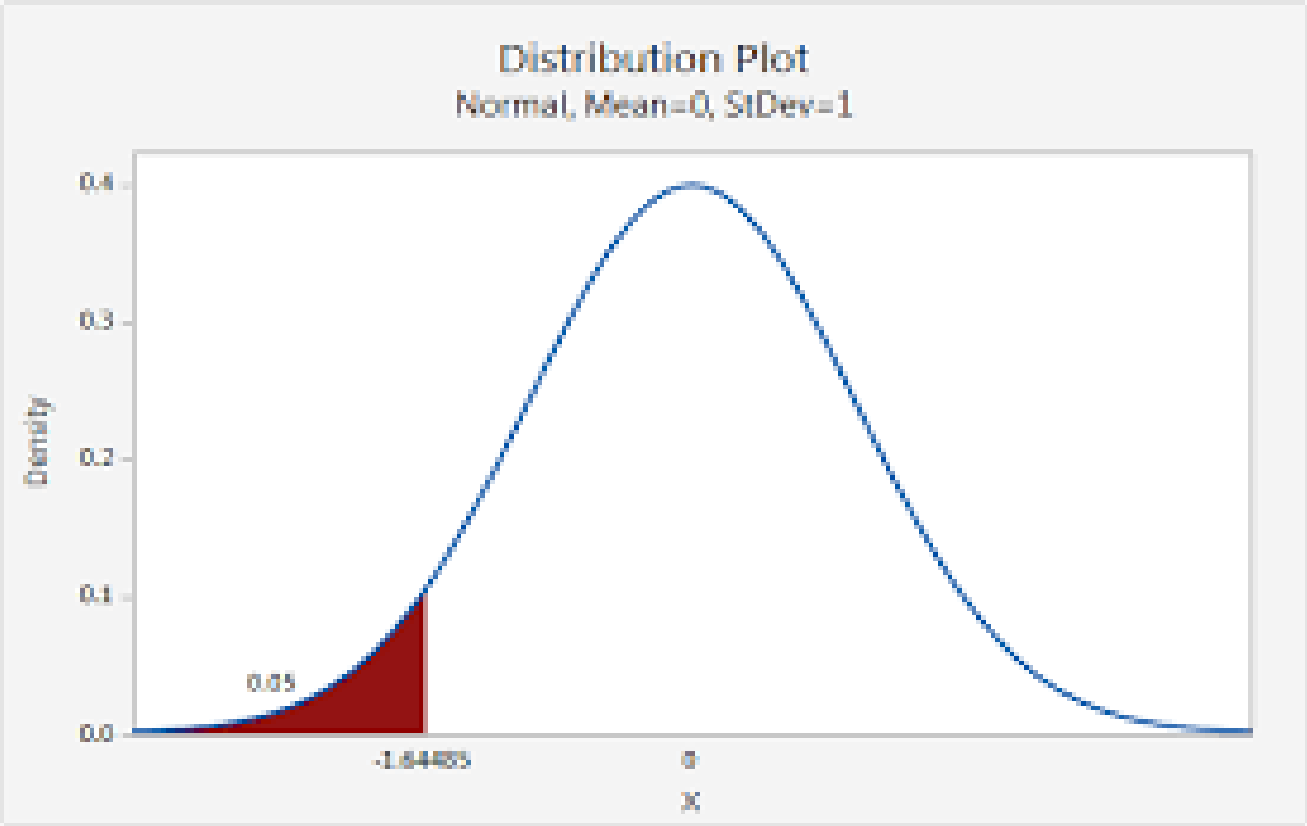
However:

1. "Small" depends on the distribution
2. What if all values have (about) the same probability?

Thus no general theory has evolved from this approach

Refined idea: "Extreme" values

Given a "typical" distribution, small (and / or large) observations are often suspicious:



Thus one gets the P-value (P-integral)

Interpreting *P*-Values

is notoriously difficult. For example:

- If my findings are not significant, then I know that they probably just occurred by chance
- If the result is significant, then I know I have a reliable finding
- The *P*-values from the significance test tell me whether the relationship in my data are large enough to be important or not
- I can also determine from the *P*-value what the chances are that these findings would replicate if I conducted a new study

However: Every one of these thoughts about the benefits of significance testing is false. (Schmidt 1996)

A **small** *P*-value can even be evidence **in favour** of the hypothesis. (Lindley 1957)

The basic flaw

Objection has sometimes been made that the method of calculating confidence limits by setting an assigned value such as 1% on the frequency of observing 3 or less ... is unrealistic treating values less than 3, which have not been observed, in exactly the same manner as 3, which is the one that has been observed. This feature is indeed not very defensible save as an approximation. (Fisher 1956)

An hypothesis that may be true is rejected because it has failed to predict observable results that have not occurred. This seems a remarkable procedure. (Jeffreys 1939)

The straightforward remedy: Several hypotheses

. . . if there is any alternative hypothesis . . . you will be much more inclined to consider that the original hypothesis is not true . . .” (Gosset / Student 1926)

. . . the only valid reason for rejecting a statistical hypothesis is that some alternative hypothesis explains the observed events with a greater degree of probability. (E.S. Pearson 1938)

Crucial decision (bifurcation)

Given two hypotheses H and K, make your choice:

1. Either stick to P-integrals
2. Or compare $P_H(x)$ and $P_K(x)$

Notice, however, that in order to remedy the constructional flaw mentioned above, there is a fundamental difference:

1. The integrals $\alpha = P_H(X \leq x_\alpha)$ and $\beta = P_K(X \geq x_\beta)$, say, have to be defined a priori
2. The ratio $P_H(x)/P_K(x)$ is a posteriori, conditional on the observation x

Neyman & Pearson: The first choice

A statistical test is a fixed decision procedure with four ingredients:

- The error of the first kind (decision in favour of K, although H is correct)
- The error of the second kind (decision in favour of H, although K is correct)
- The effect size, e.g. $EX_K - EX_H$
- Sample size n

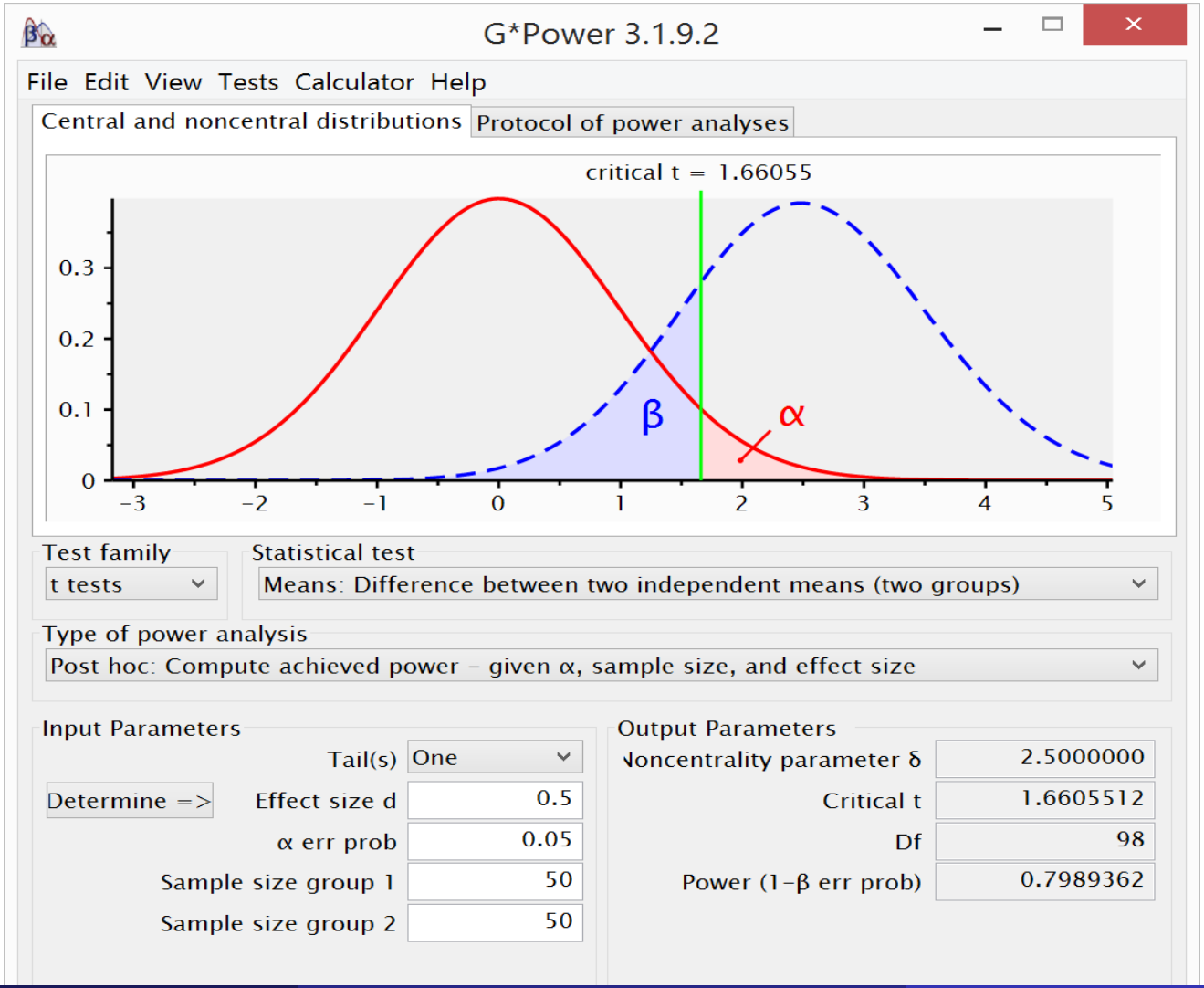
The hypotheses are treated asymmetrically: H represents the nil hypothesis, K is a substantial alternative

Thus α and β are also treated differently

At the same time, the parameters are on an equal footing, given three of them, the fourth can be computed

Neyman & Pearson: Standard situation

Given two normal distributions, the acceptance / rejection regions are intervals:



Consequence 1: "Economy" of experimentation

Given α and β, and the effect size, there is an optimum sample size:

> *The appropriate test is one which, while involving (through the choice of its significance level* $[\alpha]$*) only a very small risk of discarding my working hypothesis* $[H]$ *prematurely will enable me to demonstrate with assurance* $[1-\beta]$ *(but without any unnecessary amount of experimentation) the reality of the influences which I suspect may be present [K]. (E.S. Pearson 1955)*

Consequence 2: "Error statistics"

The crucial concept becomes error and how to control it (Mayo 1996)

In particular, α is a limited, non-renewable resource:

> *Once we have spent this error rate, it is gone.*
>
> *[Thus] a very few prespecified comparisons will be allowed to eat up the available error rate, and the remaining comparisons have the logical status of hints, no matter what statistical techniques may be used to study them. (Tukey 1991)*

Consequence 3: Advance Planning and Restrictions

1. Adjustment: Spent the available error rate with great care
2. Emphasis on prespecified analyses:

   There is a popular solution to this problem, a simple way to prevent experimental trials from evolving into demonstration trials: do not allow those who are conducting the trial to look at the results as they accumulate. That is, . . . conceal the evidence from the physician until the trial is completed. (Royall 1991)

3. Penalize multiple looks at the data (fishing for signifcant results)
4. In a nutshell: Encourage hypothesis-driven research (Popper), discourage data-driven explorations (induction)

Criticism 1: Lack of symmetry

The asymmetrical treatment of the hypotheses is not mandatory, rather it is Neyman's and Pearson's choice.

(Too) much rests on this decision:

> *It is clear that the entire basis for sequential analysis depends upon nothing more profound than a preference for minimizing β for given α rather than minimizing their linear combination. Rarely has so mighty a structure and one so surprising to scientific common sense, rested on so frail a distinction and so delicate a preference. (Cornfield 1966)*

Criticism 1a: Lack of symmetry $\longrightarrow$ conservativism

The overall attitude becomes conservative:

> *For a time it was fashionable to detect 'hidden periodicities' for sunspots, wheat prices, poetic creativity, etc. Such hidden periodicities used to be discovered as easily as witches in medieval times, but even strong faith must be fortified by a statistical test. (Feller 1971)*

> *Apparently, Feller did not believe in the sunspot periodicity, which no responsible scientist has doubted for over a century. The evidence for it is so overwhelming that nobody needs a 'statistical test' to see it ... the eyeball is a more reliable indicator of an effect than an orthodox [test]. (Jaynes 2003)*

Criticism 1b: Lack of symmetry$\longrightarrow$ conservativism

Adjusting for α straightforwardly leads to an ultraconservative attitude that effectively prevents the detection of real effects:

> *Instead of dealing with the very credible threat of Type II errors, researchers have been imposing increasingly stringent controls to deal with the relatively unlikely threat of Type I errors...*
>
> *In view of these trade-offs, adjusting alpha may be a bit like spending $1,000 to buy insurance for a $500 watch. (Ellis 2010)*

Criticism 2: Effect Size

Effect size, n, and (the probabilites of) error should **not** be on a par.

In particular, the effect size summarizes the difference between two groups, e.g. the impact of a certain treatment:

> *. . . the emphasis on significance levels tends to obscure a fundamental distinction between the size of an effect and its statistical significance.*
>
> *Regardless of sample size, the size of an effect in one study is a reasonable estimate of the size of an effect in replication. (Tversky and Kahneman 1971)*

Criticism 3: Sample size

In general, information is crucial.

The larger n, the more information there is:

> *There are no inferential grounds whatsoever for preferring a small sample ...*
>
> *The larger the sample size the more stable the estimate of effect size; the better the information, the sounder the basis from which to make a decision ...*
>
> *the larger the sample the better ... (Oakes 1986)*

Criticism 4: Experimentation in practice

An experiment involving an image-producing apparatus often ends appropriately with a 'golden event', that is, a picture or image of something whose existence has been conjectured, but possibly questioned.

An experiment involving a counting apparatus often ends appropriately when a decision based on some probability model suggests that enough counts have been taken for some purpose. (Ackermann 1989)

Criticism 4a: Logic of experimentation

An experimenter, having made n observations in the expectation that they would permit the rejection of a particular hypothesis, at some predesignated significance level, say .05, *finds that he has not quite attained his critical level.*

He still believes that the hypothesis is false and asks how many more observations would be required to have reasonable certainty of rejecting the hypothesis.

Criticism 4a: Logic of experimentation

Under these circumstances it is evident that there is no amount of additional information, no matter how large, which would permit rejection at the .05 *level.*

It the hypothesis being tested is true, there is a .05 *of its having been rejected after the first round of observations. To this chance must be added the probability of rejecting after the second round, given failure to reject after the first, and this increases the total chance of erroneous rejection to above* .05

Thus no amount of additional evidence can be collected which would provide evidence against the hypothesis equivalent to rejection at the $P = 0.05$ *level . . . (Cornfield 1966)*

Criticism 5: Scientific common sense

. . . it is indeed unsatisfactory to have to defend, perhaps in the face of senior, highly qualified substantive scientists, our **mainstream statistical thinking** *which assumes that you are* **not supposed to look at the data** *when searching for methods of optimal analysis with the purpose of gaining new knowledge. (Keiding 1995)*

Basic technical question: Why integrals?

It is not difficult to see how 'Student' and Fisher found themselves defending the use of the P integral.

For if one accepts that it is possible to test a null hypothesis **without specifying an alternative**, *and that the test must be based on the value of a test statistic in conjunction with its known sampling distribution on the null hypothesis, then the integral of the distribution between specified limits is the only measure which is invariant to transformation of the statistic.*

If follows that one is virtually forced to consider the area between the realized value of the statistic and a boundary as the rejection area - the P integral, in fact. (Edwards 1972)

Likelihood et al.: The second choice?

Basic insight (Royall 1997):

> *. . . a proper measure of strength of evidence should not depend on probabilities of unobserved values.*

Given two hypotheses H, K, and the observation x, just consider the ratio $P_K(x)/P_H(x)$. More generally:

$$r_n = r_n(x_1, \ldots, x_n) = \frac{P_K(x_1, \ldots, x_n)}{P_H(x_1, \ldots, x_n)} = \prod_{i=1}^{n} \frac{P_K(x_i)}{P_H(x_i)}$$

Decide in favour of K if r_n exceeds some preassigned number t.

Levels of evidence

Jeffreys's rule of thumb:

$r_n > 1$ supports K

$1 > r_n > 0.3$ supports H, "but not worth more than a bare comment."

The evidence in favour of H (and thus, equivalently, against K) is

substantial	if	$0.3 > r_n > 0.1$
strong	if	$0.1 > r_n > 0.03$
very strong	if	$0.03 > r_n > 0.01$
decisive	if	$0.01 > r_n$

Likelihood-based inference

It is well known that

$$P\left(\prod_{i=1}^{n}\frac{P_K(X_i)}{P_H(X_i)} \geq t \text{ for some } n = 1, 2, \ldots\right) \leq \frac{1}{t}$$

if H is correct (Robbins 1970).

Notice, that "if an unscrupulous researcher sets out deliberately to find evidence supporting his favourite hypothesis [K] over his rival's [H], which happens to be correct, by a factor of at least [t], then the chances are good that he will be eternally frustrated." (Royall 1997)

In general, evidence accrues. You just have to be patient...

Bayesian extension

Suppose $\pi(H)$ and $\pi(K)$ are the prior probabilities of H and K.

Then Bayes theorem straightforwardly implies:

$$\frac{\pi(K|x_1,\ldots,x_n)}{\pi(H|x_1,\ldots,x_n)} = \frac{\pi(K)}{\pi(H)} \cdot \prod_{i=1}^{n} \frac{p_K(x_i)}{p_H(x_i)} = r_n(x_1,\ldots,x_n) \cdot \frac{\pi(K)}{\pi(H)}$$

General results (Walker 2004) guarantee fast convergence if $0 < \pi(H) < 1$ and $0 < \pi(K) < 1$.

Modern treatment (information theory)

Define Kullback-Leibler divergence:

$$D(P_H||P_K) = \sum_x P_H(x) \log \frac{P_H(x)}{P_K(x)}$$

If $\hat{P}_n$ is the empirical distribution, the crucial insight is

$$\log r_n = \log \frac{P_K(x_1, \ldots, x_n)}{P_H(x_1, \ldots, x_n)} = n \cdot D(\hat{P}_n||H) - n \cdot D(\hat{P}_n||K)$$

Thus, if H is true, owing to the law of large numbers,

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{P_H(X_1, \ldots, X_n)}{P_K(X_1, \ldots, X_n)} \to D(H||K)$$

in probability.

Modern treatment (Likelihood test)

A decision in favour of H if

$$\frac{P_K(x_1, \ldots, x_n)}{P_H(x_1, \ldots, x_n)} < \varepsilon$$

is equivalent to a decision in favour of H if

$$D(\hat{P}_n||H) - D(\hat{P}_n||K) < \frac{1}{n} \log \varepsilon$$

Thus, without prior preference ($\varepsilon = 1$), one should decide in favour of H if

$$D(\hat{P}_n||H) < D(\hat{P}_n||K),$$

i.e., if $\hat{P}_n$ is "closer" to H.

Modern treatment (Bayesian test)

Assume $D(P_H||P_K) < \infty$, and let A_n be the acceptance region for H, depending on n. Thus $\alpha_n = P_H(A_n^c)$, $\beta_n = P_K(A_n)$.

Symmetry assumption: The goal ist to minimize the total probability of error $P_n^e = \pi(H)\alpha_n + \pi(K)\beta_n$

Then, given a sample $x_1, \ldots, x_n$ from either H or K, the optimum decision rule is to decide in favour of H if

$$\frac{1}{n} \log \frac{\pi(K)}{\pi(H)} + \frac{1}{n} \sum_{i=1}^{n} \log \frac{P_K(x_i)}{P_H(x_i)} < 0$$

and in favour of K if the left-hand side is > 0.

Modern treatment (Single Hypothesis)

For a single hypothesis H, consider the "ball" of all distributions P that are close to H. That is, given constant $c > 0$, we have $D(P||H) \leq c$.

Given a sample of size n, Hoeffding's "universal test" (1965) decides in favour of H, if $D(\hat{P}_n||H) \leq c_n$.

Owing to the law of large numbers, the sequence c_n converges to 0 rapidly.

General remarks

Evaluation of the prior perspective:

- Rather formal, mathematical point of view
- Preferred statistical techniques (e.g. statistical tests, randomization, regression) supersede problem orientation
- Context information is not formalized, and therefore cannot be dealt with in a precise way
- Cult of the single study considered in isolation (Nelder 1999) "Let the data speak for themselves"
- From a historic perspective, no comprehensive theory has evolved Rather: much ad hocery, toolkit statisticians serving "cookbook statistics"

Focus on the information in the data

Many major developments in the last decades:

- EDA, Data Mining, Machine Learning, Neuronal networks
- Modelling (find the information in the data at hand)
- Meta Analysis (systematic summary of all the available evidence)
- Evidence from various sources, non-classical collection of information: quasi-experiments, control functions, information markets, non-random "information rich" sampling
- Likelihood-based inference: MLE, ME, AIC, BIC, Bayes
- "Extended Bayes": Full Probability Modelling, Nonparametric Bayes, Imprecise Probability, Dempster-Shafer theory
- Combining probability with other mathematical theories, in particular structural equations (SEM) and causal graphs (DAG)
- Strong link to information & communication theory, and the computer sciences in general

Conclusions

- There is still a chasm between mathematical statistics and scientific data analysis (deduction vs. induction)
- Tests of hypotheses can serve as a particularly simple model to study these matters (they may serve as a "test bed")
- Prior view: Measures to be taken in order to obtain informative data (optimum collection methods, best decisions)
- Posterior view: Extract all the information in the available data
- Information is the most important idea in statistics and other sciences
- A unified theory could be possible

Thank you!

- Ackermann, R. (1989). The New Experimentalism. *Brit. J. Phil. Sci.* **40**, 185-190.
- Cornfield, J. (1966). Sequential Trials, Seq. Analysis and the Likelihood Principle. *Am Stat* **20(2)**, 18-23.
- Edwards, A.W.F. (1972). Likelihood. Johns Hopkins Univ. Press, Baltimore, MD
- Ellis, P.D. (2010). The essential guide to effect sizes. Cambridge University Press, New York
- Feller, W. (1971). An Introduction to Probability Theory and its Applications. Vol. 2 (2nd ed.), Wiley
- Fisher, R.A. (1935). The Logic of Inductive Inference. *J. Royal Stat. Soc.* **98**, 39-54.
- Fisher, R.A. (1956). Statistical Methods and Scientific Inference. Oliver and Boyd, Edinburgh
- Hoeffding, W. (1965). Asymptotically optimal tests for multinomial distrib. *Ann Math Stat* **36**, 369-408.
- Jaynes, E.T. (2003). Probability Theory. The Logic of Science. Cambridge Univ. Press, Cambridge
- Jeffreys, H. (1939). Theory of Probability. Clarendon Press, Oxford
- Keiding, N. (1994). Comment on Spielhalter, D.J.; Freedman, L.S.; and M.K.B. Parmar: Bayesian Approaches to Randomized Trials. *J. Royal Stat. Soc. A* **157**, 357-416.
- Lindley, D.V. (1956). On a Measure of the Information Provided by an Experiment. *Ann Math Stat* **27**, 986-1005.
- Mayo, D.G. (1996). Error and the Growth of Experimental Knowledge. Univ. Chicago Press, Chicago
- Nelder, J.A. (1999). Statistics for the Millenium. *The Statistician* **48**, 257-269.
- Oakes, M. (1986). Statistical Inference: A Commentary for the Social and Behavioral Sciences, Wiley
- Pearson, E.S. (1938). Student as Statistician. *Biometrika* **30**, 210-250.
- Pearson, E.S. (1955). Statistical Concepts and their Relation to Reality. *J. Royal Stat. Soc. B* **17**, 204-207.
- Robbins, H. (1970). Statistical methods related to the LIL. *Ann Math Stat* **41**, 1397-1409.
- Royall, R.M. (1991). Ethics and Statistics in Randomized Clinical Trials. *Statistical Science* **6**(1), 52-88.
- Royall, R.M. (1997). Statistical Evidence. A Likelihood Paradigm. Chapman & Hall, London
- Saint-Mont (2011). Statistik im Forschungsprozess. Physica-Verlag (Springer), Heidelberg
- Schmidt, F. (1996). Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for Training of Researchers. *Psychological Methods* **1(2)**, 115-129.
- Tukey, J.W. (1991). The Philosophy of Multiple Comparisons. *Stat. Science* **6(1)**, 100-116.
- Tversky, A.; and D. Kahneman (1971). Belief in the Law of Small Numbers. *Psych. Bulletin* **76**, 105-110.
- Walker, S. (2004). New approaches to Bayesian consistency. *Ann Stat* **32(5)**, 2028-2043.