Statistik III

Advanced Statistics

Florian Wickelmaier

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



Summer 2026

Bayesian Statistics: Basics

Parameters

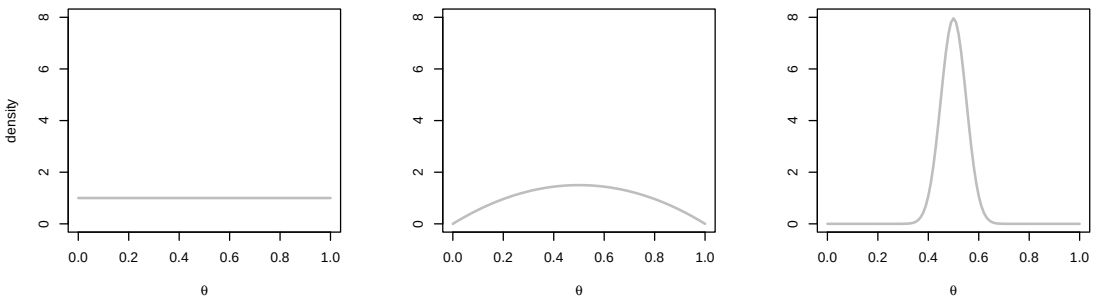
- In contrast to classical statistics, parameters are not considered as fixed values (i. e. numbers), but as random variables
- In the following, Θ will denote a generic parameter
- A specific value θ is considered as a realization of the random variable Θ
- The uncertainty about the value θ of the parameter is represented by a probability distribution of the random variable Θ

Prior Distribution

- Before having available data from a random sample, the uncertainty regarding the parameter value is described by the *prior distribution*
- The prior distribution represents (more or less specific) a priori assumptions, knowledge, or beliefs regarding the parameter value
  - informative, diffuse, uninformative priors
- Let $h(\theta)$ denote the prior density (or mass function in the discrete case) of Θ
- In this context it is said that
  - Bayesian statistics is based on a *subjective notion of probability*
  - classical statistics is based on a *frequentist concept of probability* (probability as a limit of a sequence of relative frequencies)
- Both theories are based on the concept of probability according to Kolmogorov
  - The difference lies in the interpretation of probabilities, where classical statistics relies on repeated random experiments (as in the interpretation of confidence intervals)

Example: Prior Distribution

Let Θ denote the parameter of the Bernoulli distribution for a coin toss, then the following density functions represent different degrees of uncertainty about the parameter value



In each case, h(θ) is a beta distribution

Posterior Distribution

- As soon as data are available, the *posterior distribution* is calculated using an appropriate version of Bayes' formula
- The posterior distribution of the random variable Θ will deviate more or less strongly from the prior distribution when accounting for the data (Bayesian learning, updating, or inference)
- Let $f(\theta \mid x_1, \ldots, x_n)$ denote the posterior density of Θ given the observed data $X_1 = x_1, \ldots, X_n = x_n$

Bayesian Updating

- Notation

$l(x_1, \ldots, x_n \mid \theta)$	probability density of the data given $\Theta = \theta$
$g(x_1, \ldots, x_n)$	marginal density of the data
$h(\theta)$	prior density of Θ (marginal density of Θ)
$f(\theta \mid x_1, \ldots, x_n)$	posterior density of Θ given the data

- Bayes' formula is given by

$$f(\theta \mid x_1, \ldots, x_n) = \frac{l(x_1, \ldots, x_n \mid \theta) \cdot h(\theta)}{g(x_1, \ldots, x_n)}$$

- The denominator does not depend on the parameter, but on the data only
- $l(x_1, \ldots, x_n \mid \theta) = l(x_1 \mid \theta) \cdot \ldots \cdot l(x_n \mid \theta) = L(\theta)$ is the likelihood function

Bayesian Updating

- For a general continuous random variable Θ (the parameter can take on values in a real interval) we obtain the marginal density $g(x_1, \ldots, x_n)$ by integrating over all possible values θ

$$g(x_1, \ldots, x_n) = \int l(x_1, \ldots, x_n \mid \theta) \cdot h(\theta) d\theta$$

- This procedure is also occasionally described as
  - "integrating out" the dependency on the unknown parameter value θ
- Whenever the integral cannot be solved analytically, numerical integration methods, or simulation techniques are to be used

Bayesian Updating

- The above characterizations of the densities $l(x_1, \ldots, x_n \mid \theta)$ and $g(x_1, \ldots, x_n)$ can be substituted into Bayes' formula
- This results in the following equations

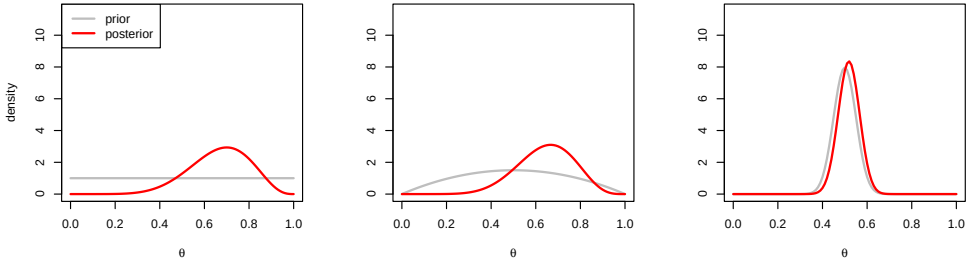
$$f(\theta \mid x_1, \ldots, x_n) = \frac{l(x_1, \ldots, x_n \mid \theta) \cdot h(\theta)}{\int l(x_1, \ldots, x_n \mid \theta) \cdot h(\theta) d\theta}$$

$$= \frac{l(x_1 \mid \theta) \cdot \ldots \cdot l(x_n \mid \theta) \cdot h(\theta)}{\int l(x_1 \mid \theta) \cdot \ldots \cdot l(x_n \mid \theta) \cdot h(\theta) d\theta}$$

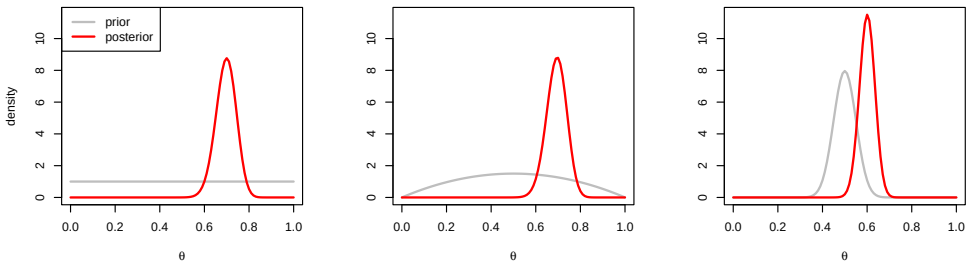
$$= \frac{L(\theta) \cdot h(\theta)}{\int L(\theta) \cdot h(\theta) d\theta}$$

Example: Bayesian Updating

- Data: The coin lands heads up 7 times in 10 tosses



- Data: The coin lands heads up 70 times in 100 tosses



Parameter Estimation

Bayesian estimation methods

- The posterior distribution $f(\theta \mid x_1, \ldots, x_n)$ forms the basis for determining both interval and point estimators of the unknown parameter
- In distinction to the confidence intervals in classical statistics, one speaks of *credible intervals* or *highest posterior density regions/intervals*
- Remember: The confidence level $1 - \alpha$ of a confidence interval characterizes
  - *not* the probability with which the parameter lies in the resulting interval
  - but the probability with which the method used produces an interval containing the parameter (proportion of covering intervals)

Credible Interval

- For credible intervals derived in the framework of Bayesian statistics, the confidence level can actually be interpreted as the probability with which the parameter lies in the resulting interval
- A two-sided credible interval at level $1 - \alpha$ can be specified using the posterior distribution by defining
  - the lower bound g_l as the $(\alpha/2)$-quantile

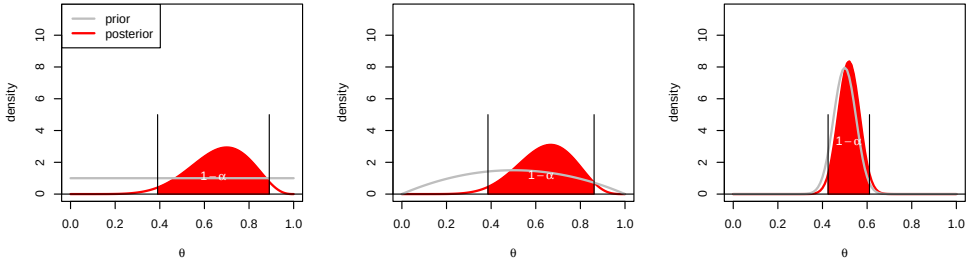
$$\int_{-\infty}^{g_l} f(\theta \mid x_1, \ldots, x_n) d\theta = \frac{\alpha}{2}$$

  - the upper bound g_u as the $(1 - \alpha/2)$-quantile

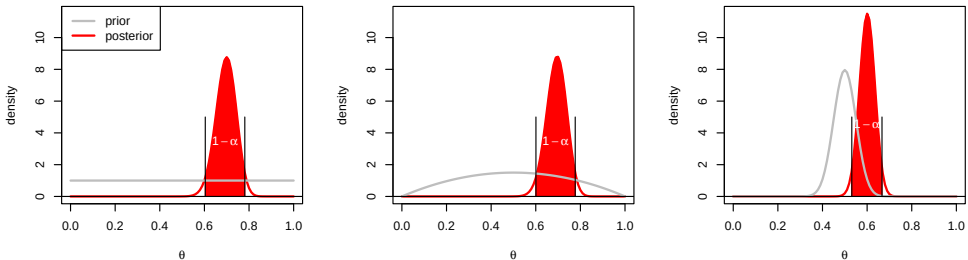
$$\int_{-\infty}^{g_u} f(\theta \mid x_1, \ldots, x_n) d\theta = 1 - \frac{\alpha}{2}$$

Example: Credible Interval

- Data: The coin lands heads up 7 times in 10 tosses



- Data: The coin lands heads up 70 times in 100 tosses



Highest Posterior Density Region/Interval

- The region HPD_θ is determined for given α such that

$$\text{HPD}_\theta = \{\theta \mid f(\theta \mid x_1, \ldots, x_n) > \tau\}$$

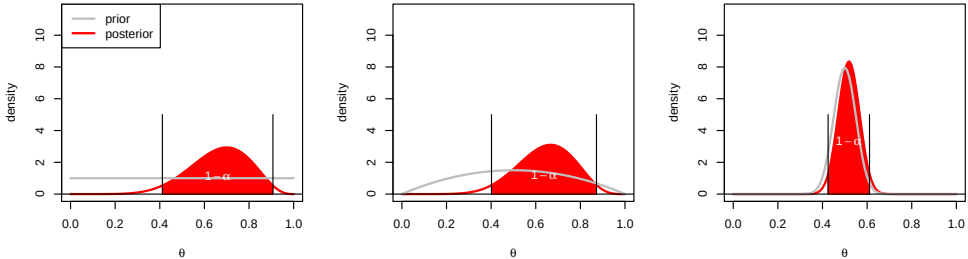
and

$$\int_{\{\theta | f(\theta|x_1,\ldots,x_n) > \tau\}} f(\theta \mid x_1, \ldots, x_n) d\theta = 1 - \alpha$$

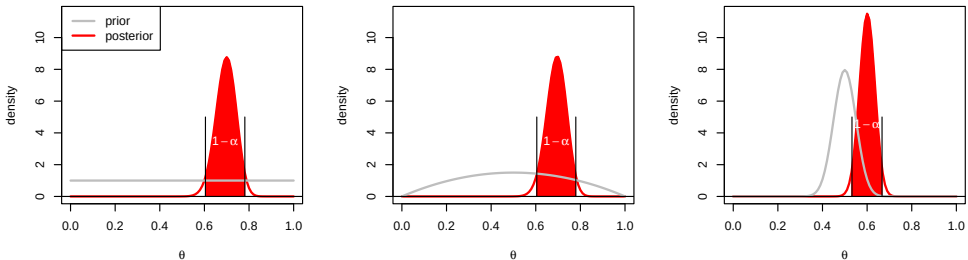
- The result is an interval whenever f is unimodal

Example: HPD Interval

- Data: The coin lands heads up 7 times in 10 tosses



- Data: The coin lands heads up 70 times in 100 tosses



Point Estimation (EAP)

- The *expected a posteriori* (EAP) estimator is calculated as the expected value

$$\hat{\theta}_{EAP} = E(\theta \mid x_1, \ldots, x_n) = \int \theta \cdot f(\theta \mid x_1, \ldots, x_n) d\theta$$

- The integral may have to be determined via numerical integration methods or simulation techniques
- Markov chain Monte Carlo (MCMC) estimation
  - with the posterior samples $\theta_1, \ldots, \theta_m \sim f(\theta \mid x_1, \ldots, x_n)$
  - the MCMC estimator is

$$\hat{E}(\theta \mid x_1, \ldots, x_n) = \frac{1}{m} \sum \theta_i$$

Point Estimation (MAP)

- The calculation of the *maximum a posteriori* (MAP) estimator avoids this computational problem
- Since the integral in the denominator of

$$f(\theta \mid x_1, \ldots, x_n) = \frac{L(\theta) \cdot h(\theta)}{\int L(\theta) \cdot h(\theta) d\theta}$$

  does not depend on θ, the mode can obtained by maximizing the numerator $L(\theta) \cdot h(\theta)$
- This means that the parameter value $\hat{\theta}_{MAP}$ is chosen such that

$$L(\hat{\theta}_{MAP}) \cdot h(\hat{\theta}_{MAP}) = \max_{\theta} L(\theta) \cdot h(\theta)$$

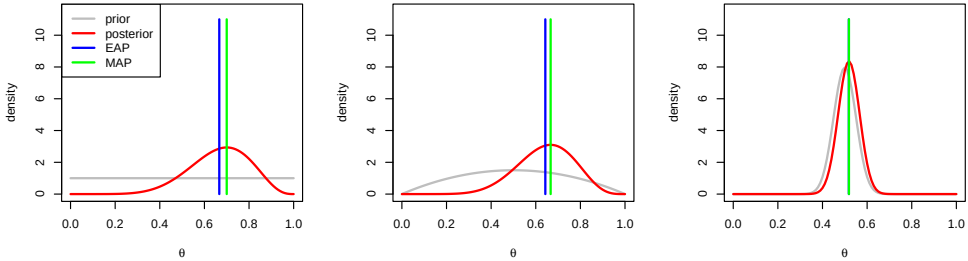
Point Estimation

Relation to maximum likelihood estimation

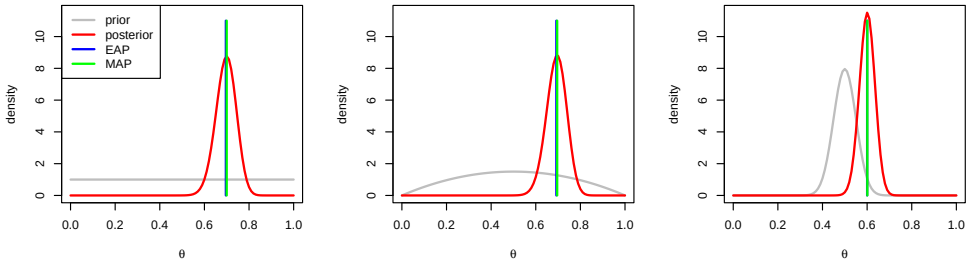
- Instead of the likelihood $L(\theta)$ in ML estimation, MAP maximizes the function $L(\theta) \cdot h(\theta)$ with the prior density $h(\theta)$ as weighting factor
- For an almost flat prior density $h(\theta)$ expressing little prior knowledge regarding the value θ, the ML estimator $\hat{\theta}$ and the MAP estimator $\hat{\theta}_{MAP}$ are barely distinguishable
- In the extreme case, where $h(\theta)$ is a uniform distribution on the range of values taken by θ, the two estimators coincide

Example: Point Estimation

- Data: The coin lands heads up 7 times in 10 tosses



- Data: The coin lands heads up 70 times in 100 tosses



Bayesian Modeling with Stan

https://mc-stan.org/

Stan

- Probabilistic programming language that allows us to flexibly specify $f(\theta \mid x_1, \ldots, x_n) \propto L(\theta) \cdot h(\theta)$
- Algorithm that efficiently draws samples θ_i from the posterior

Stan model

- Three mandatory blocks: `data`, `parameters`, `model`
- Two basic variable types: `int`, `real` (and many not so basic ones)

Interface to R via `rstan::stan("model.stan", data)`

- Transpiles the model code to C++
- Compiles the C++ to an executable
- Calls the executable: draw MCMC samples for all parameters

Example: Bayesian Modeling with Stan

Model specification (`coinflip.stan`)

```
data {
  int<lower=1> n;
  int<lower=0> x;
}
parameters {
  real<lower=0, upper=1> theta;
}
model {
  theta ~ beta(2, 2);      // prior
  x ~ binomial(n, theta);  // likelihood
}
```

Example: Bayesian Modeling with Stan

Calling Stan from R

```
library(rstan)
m <- stan("coinflip.stan", data = list(n = 10, x = 7))
theta <- extract(m)$theta  # MCMC samples
```

Output of the `stan()` function

```
Inference for Stan model: coinflip.
4 chains, each with iter=2000; warmup=1000; thin=1;
post-warmup draws per chain=1000, total post-warmup draws=4000.

       mean se_mean   sd   2.5%   25%   50%   75% 97.5% n_eff Rhat
theta  0.65    0.00 0.12   0.40  0.57  0.65  0.74  0.87  1072    1
lp__  -9.64    0.02 0.75 -11.87 -9.80 -9.35 -9.17 -9.13  1431    1
```

Example: Bayesian Modeling with Stan

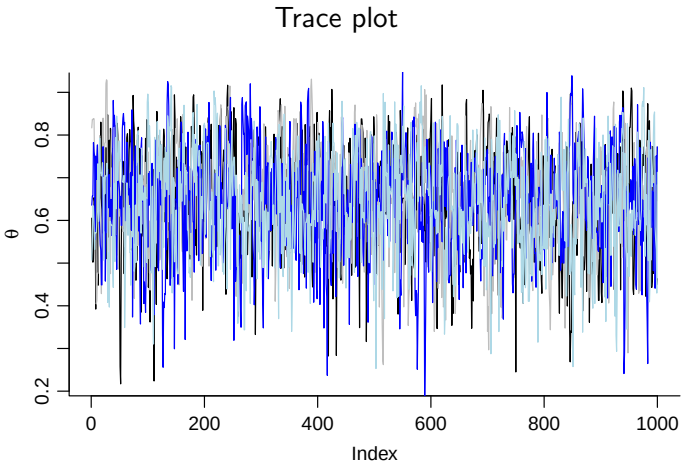
Plotting

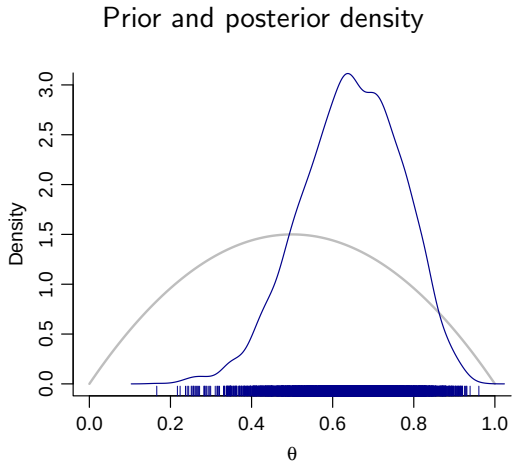
```
## Diagnostics: trace plot
traceplot(m)

## Prior and posterior density
curve(dbeta(x, 2, 2), ylim = c(0, 3), lwd = 2, col = "gray",
      xlab = expression(theta), ylab = "Density")              # prior
density(theta) |> lines(col = "darkblue")                      # posterior
rug(theta, col = "darkblue")                                   # samples
```

Credible interval, highest density interval

```
quantile(theta, c(.025, .975))
# 2.5% 97.5%
# 0.40  0.86
HDInterval::hdi(theta)
# lower upper
#  0.42  0.88
```

Example: Bayesian Modeling with Stan





Hypothesis Testing

Neyman-Pearson hypothesis testing

- The test is formalized as a decision between two hypotheses: the null hypothesis H_0 on the one side, and the alternative hypothesis H_1 on the other side
- Each of the hypotheses is formally and logically the negation of the other one
- Derivation of the test statistic assumes that the null hypothesis H_0 (which in many cases represents no effect) is valid
- The decision situation thus is asymmetric, and H_0 can be discarded in favor of H_1 or retained
- The probabilities of falsely rejecting the null hypothesis (Type I error rate α) and falsely retaining it (Type II error rate β, obtained from a prospective power analysis) are controlled for

Hypothesis Testing

Bayesian hypothesis testing

- Any hypotheses H_1, H_2 can be tested against each other
- Before collecting data, the ratio of the prior probabilities $P(H_1)/P(H_2)$ (prior odds) is considered
- One is interested in its change in the light of data D, i. e., in the ratio of posterior probabilities $P(H_1 \mid D)/P(H_2 \mid D)$ (posterior odds)
- The inference is described by means of Bayes' formula

$$\frac{P(H_1 \mid D)}{P(H_2 \mid D)} = \frac{P(H_1)}{P(H_2)} \cdot \underbrace{\frac{f(D \mid H_1)}{f(D \mid H_2)}}_{\text{Bayes factor}}$$

Bayes Factor

- The transition from prior to posterior odds is determined by the Bayes factor

$$B_{12} = \frac{f(D \mid H_1)}{f(D \mid H_2)}$$

where $f(D \mid H_i)$ denotes the *marginal likelihood* of the data under hypothesis H_i

- This marginal likelihood is calculated via the integral

$$f(D \mid H_i) = \int_{\boldsymbol{\theta}} f(D \mid \boldsymbol{\theta}, H_i) \cdot h(\boldsymbol{\theta} \mid H_i) d\boldsymbol{\theta}$$

over all possible parameter values $\boldsymbol{\theta}$ under hypothesis H_i (which, in general, cannot be solved analytically)

Bayes Factor

- The integration can be viewed as the computation of a "mean likelihood" of the data D under the hypothesis H_i
  - It is a weighted average with the prior density $h(\boldsymbol{\theta} \mid H_i)$ serving as a weighting factor
- To achieve a high marginal likelihood, parameter values $\boldsymbol{\theta}$ that lead to low likelihoods should not have high probability under H_i
- Thus, the Bayes factor (in contrast to maximum likelihood estimation) takes into account the complexity of the theory under consideration
  - Less informative priors, which allow for many (possibly implausible) parameter values with comparable probability, lead to rather low marginal likelihood

Bayes Factor

The Bayes factor B_{12} provides evidence for or against the hypotheses H_1 and H_2

B_{12}	Interpretation
> 100	Decisive evidence for H_1
$30 - 100$	Very strong evidence for H_1
$10 - 30$	Strong evidence for H_1
$3 - 10$	Substantial evidence for H_1
$1 - 3$	Anecdotal evidence for H_1
1	No evidence
$1/3 - 1$	Anecdotal evidence for H_2
$1/10 - 1/3$	Substantial evidence for H_2
$1/30 - 1/10$	Strong evidence for H_2
$1/100 - 1/30$	Very strong evidence for H_2
$< 1/100$	Decisive evidence for H_2

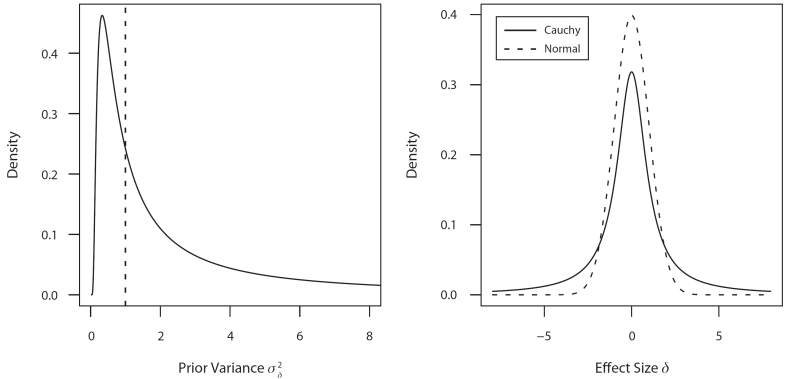
(Jeffreys, 1961)

Bayes Factor

- Based on the Bayes factor, one has either
  - evidence for hypothesis H_1
  - evidence for hypothesis H_2
  - evidence for *none* of the hypothesis (meaning, that data do not provide sufficient information for deciding between the hypotheses)
- The latter argues for collecting more data until there is sufficient evidence to support one of the hypotheses
- Some Bayesian statisticians argue that this can be done
- Classical statistics has to explicitly account for repeatedly analyzing data as more observations are added (through methods of sequential hypothesis testing), and inferences are invalidated otherwise

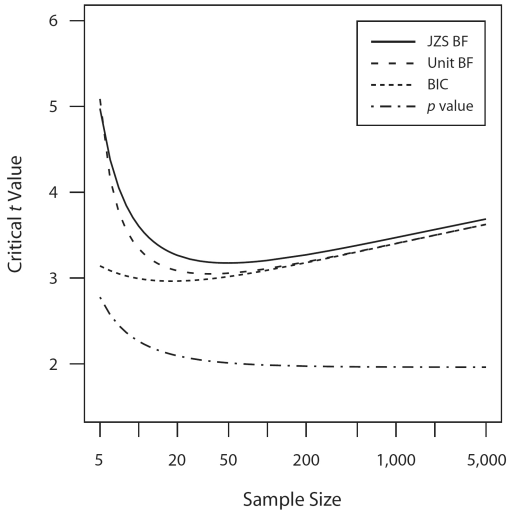
Bayesian t-Test

- Rouder et al. (2009) develop a Bayesian version of the t-test (and common hypothesis tests in R package `BayesFactor`)
- They also show how to calculate the Bayes factor based on a given t-value and the respective degrees of freedom (or alternatively the sample size)
- For variance and effect size in the alternative hypothesis, standard prior distributions are assumed (Jeffreys prior, Cauchy prior)



Bayesian t-Test

- Decisions in favor of the alternative hypothesis according to a t-test ($p < 0.05$) can be compared to those based on the Bayes factor ($BF > 10$)



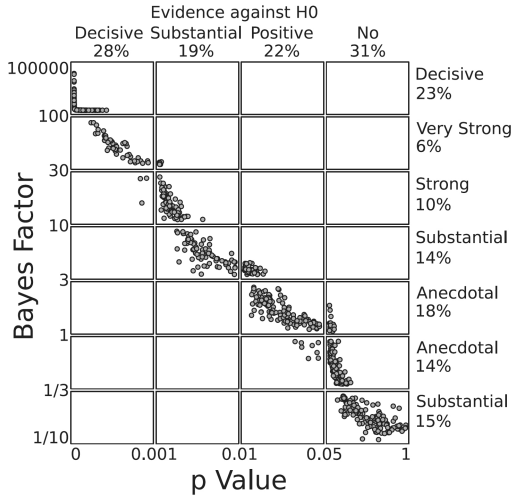
Bayesian t-Test

- The comparison shows that as the sample size increases, the critical t-value required for the particular decision
  - increases for the Bayes factor
  - remains constant for the t-test
- Deciding in favor of the alternative hypothesis requires a noticeably smaller t-value if based on the p-value $p = 0.05$ rather than on the Bayes factor $BF_{10} = 10$
- This means that the Bayes factor does not so easily favor the alternative when there are small effects in large samples
- Wetzels et al. (2011) compare p-values and corresponding Bayes factors for 855 t-tests from articles published in 2007 (PBR and JEP:LMC)

Bayesian t-Test

Results from Wetzels et al. (2011)

- There is quite a number of t-tests with $0.01 \leq p < 0.05$ but only anecdotal evidence for H_1



Cognitive Modeling

- Bayesian statistics provides a framework for formally setting up cognitive theories
- Psychological assumptions may refer to
  - the prior distributions
  - the likelihood
- Assumptions about the prior distributions
  - parameter constraints (e. g., restriction to non-negative values)
  - results of previous research on plausible parameter values (subjective vs. objective priors) supporting a cumulative research strategy
  - may be derived from hierarchical extensions of the theory (hierarchical Bayesian models) that account for individual variation

Cognitive Modeling

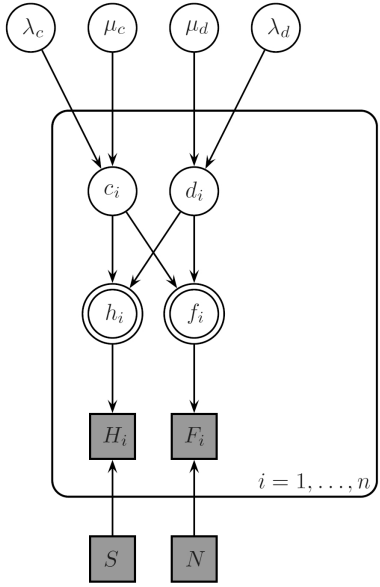
- Assumptions about the likelihood
  - specify the basic mathematical formulation of the theory
  - allow for incorporating individual differences in the context of hierarchical Bayesian models
- These assumptions are often specified in graphical models
  - continuous variables (circular nodes), discrete variables (square nodes)
  - observable variables (shaded nodes)
  - unobservable stochastic (single-bordered nodes) and deterministic (double-bordered nodes) variables
  - independent replications of graphical structure within the whole model are enclosed in bounding rectangles ("plates")

Example: Hierarchical Signal Detection Model

Heit and Rotello (2005) have presented a clever model-based evaluation of the conjecture that both inductive and deductive reasoning involve the same single psychological dimension of *argument strength* (Rips, 2001). Heit and Rotello used SDT (for detailed treatments, see Green & Swets, 1966; Macmillan & Creelman, 2005) to model this conjecture. The basic idea is to assume that the strength of an argument is unidimensional but that different decision criteria control inductive and deductive reasoning. In particular, a relatively lesser criterion of argument strength is assumed to decide between *weak* and *strong* arguments for induction, whereas a relatively greater criterion decides between *invalid* and *valid* arguments for deduction. Under this conception, deduction is simply a more stringent form of induction. (Lee, 2008, p. 9)

Example: Hierarchical Signal Detection Model

Graphical model and prior distributions (Lee & Wagenmakers, 2014, Fig. 9.4)



$$\mu_d \sim \text{Gaussian}(0, .001)$$
$$\lambda_d \sim \text{Gamma}(.001, .001)$$
$$\mu_c \sim \text{Gaussian}(0, .001)$$
$$\lambda_c \sim \text{Gamma}(.001, .001)$$

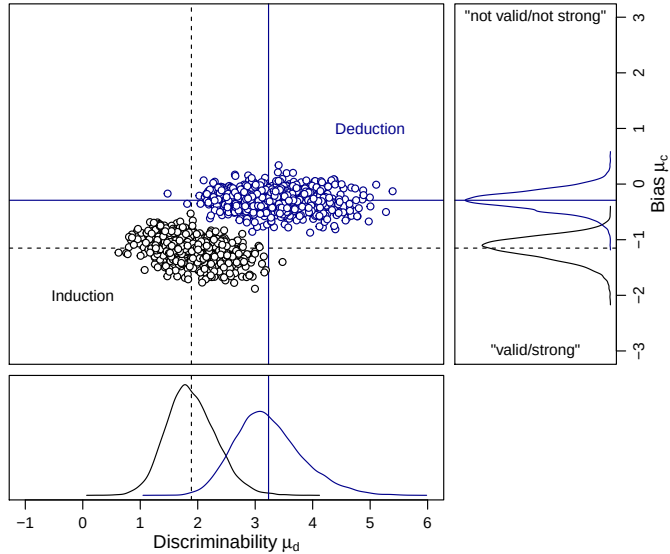
$$d_i \sim \text{Gaussian}(\mu_d, \lambda_d)$$
$$c_i \sim \text{Gaussian}(\mu_c, \lambda_c)$$

$$h_i = \Phi(\tfrac{1}{2}d_i - c_i)$$
$$f_i = \Phi(-\tfrac{1}{2}d_i - c_i)$$

$$H_i \sim \text{Binomial}(h_i, S)$$
$$F_i \sim \text{Binomial}(f_i, N)$$

Example: Hierarchical Signal Detection Model

Posterior distributions (Lee & Wagenmakers, 2014, Fig. 9.6)



Example: Hierarchical Signal Detection Model

The conclusion from the Bayesian analyses is that, in complete agreement with Heit and Rotello (2005), it is important to allow discriminability in the induction condition to be different from that in the deduction condition. The contribution of the Bayesian analysis is that this conclusion has been reached, unlike in the Heit and Rotello analysis, allowing for the possibility of individual differences in discriminability and bias across participants and accommodating the clear limitations in how accurately hit and false alarm rates can be estimated from only four observations per participant. In this way, the application demonstrates the ability of Bayesian methods to implement more realistic theoretical and methodological assumptions in drawing inferences from data. (Lee, 2008, p. 11)

Summary

- Bayesian statistics provides a formal framework for
  - evaluating the evidence of data concerning the plausibility of theories/hypotheses
  - precisely formulating psychological theories
- There is a lively (and sometimes polemic) debate about advantages and disadvantages of Bayesian statistics compared to classical statistics
- Advantages
  - Incorporation of prior knowledge (cumulative research)
  - Inferences in hypothesis testing are finer graded
  - Lower risk of false positive findings
  - Accounting for the complexity of the theories when comparing Bayes factors
  - Invites discussion of the assumptions underlying the psychological theories

Summary

- Disadvantages
  - The complex statistical procedures exhibit a high degree of flexibility
  - Requires transparency and a coherent justification of the underlying assumptions (regarding prior and likelihood)
  - Robustness of the results is to be demonstrated by a sensitivity analysis
- Bayesian statistical methods are gaining more widespread use and impact (also in psychology)
- Various software packages are available (e. g., Stan) that perform the corresponding analyses and offer interfaces to R and Python

References

Etz, A., Gronau, Q.F., Dablander, F., Edelsbrunner, P.A., & Baribault, B. (2018). How to become a Bayesian in eight easy steps: An annotated reading list. Psychonomic Bulletin and Review, 25, 219–234.

Lee, M.D. (2008). Three case studies in the bayesian analysis of cognitive models. Psychonomic Bulletin and Review, 15, 1-15.

Lee, M.D., & Wagenmakers, E.-J. (2014). Bayesian cognitive modeling: A practical course. Cambridge University Press.

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 and Review, 16, 225-237.

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.