Statistical inference (not only) for dummies

Peter Pütz & Thomas Kneib
Centre for Statistics, University of Göttingen

October 11 – October 12, 2017

Organisational matters

- (Interactive) lectures and tutorials
- In general: If you don't know what's meant, just ask!
- If you need a break, just tell me!
- Further issues?

A current debate

Andrew Gelman (2016): The Problems With P-Values are not Just With P-Values, Online discussion of the ASA Statement on Statistical Significance and P-Values, *The American Statistician*, 70.

'I put much of the blame on statistical education: [...]' '[...] it seems to me that statistics is often sold as a sort of alchemy that transmutes randomness into certainty, an "uncertainty laundering" that begins with data and concludes with success as measured by statistical significance.'

'If researchers have been trained with the expectation that they will get statistical significance if they work hard and play by the rules, if granting agencies demand power analyses in which researchers must claim 80% certainty that they will attain statistical significance, and if that threshold is required for publication, it is no surprise that researchers will routinely satisfy this criterion, and publish, and publish, and publish, even in the absence of any real effects, or in the context of effects that are so variable as to be undetectable in the studies that are being conducted.'

Aims of this workshop

- Learning something from observed data – plus potential prior evidence – about the underlying data generating process
- Understanding basic (but often misunderstood) statistical concepts
  - Unbiasedness, p-values, power, confidence intervals, ...
  - Frequentist and Bayesian reasoning
- Critical assessment of empirical research
  - Publication bias, p-hacking
  - How to do better empirical research

Adopted course material & references

Courses:

- Thomas Kneib: *Statistical Inference: Likelihood & Bayes* (given in this winter term)
- Daniel Lakens: *Improving your statistical inference* (you may take part on `https://www.coursera.org/`)
- Holger Sennhenn-Reulen: *Basic Mathematics & Statistics* (was given at the Leibniz Science Campus on Primate Cognition)

References:

- Aitken (2010): *Statistical Inference*
- Held & Sabanés Bové (2012): *Applied Statistical Inference*
- Migon & Gamerman (1999): *Statistical Inference*
- Royall (1997): *Statistical Evidence*

Random variables

Statistics is about extracting information from data that contain an inherently unpredictable component. Random variables are the mathematical construct used to build models of such variability:

- A random variable takes on a (numerical) value, at random, each time it is observed.
- In advance: Only probability statements possible about the values / events which could occur.

Two types of random variables:

- **discrete**, and
- **continuous**.

Discrete random variables

A discrete random variable X is one which may take on only a countable number of distinct values.

Example: Tossing a coin

- Random variable

$$X(\omega) = \begin{cases} 1, & \text{if } \omega = heads \\ 0, & \text{if } \omega = tails \end{cases}$$

- Probabilities

$$P(X = 1) = P(\text{``}heads\ up\text{''}) = 0.5$$

$$P(X = 0) = P(\text{``}tails\ up\text{''}) = 0.5$$

Continuous random variables

A random variable X is called continuous if the number of possible values is not countable.

Example: Waiting time for spotting a bird in an oil palm plantation.

- For a continous random variable the probability of a single outcome x is zero, i.e. $P(X = x) = 0$.

- However, we can assign probabilities to intervals, i.e. $P(a \leq X \leq b)$ may be bigger than 0.

Probability distributions

Important probability distributions for **discrete random variables**:

- **Bernoulli distribution** (0 or 1, e.g. coin flip), with the **Binomial distribution** for the sum of a sequence of Bernoulli trials ("number of successes"), and
- **Poisson distribution** (for counts, e.g. number of birds observed in a certain time interval: 0, 1, 2, $\ldots$).

The most important probability distribution for **continuous random variables** is the **normal distribution** (one main reason is the *central limit theorem*, see the references).

Note that distributions are characterized by **parameters**, e.g. the Bernoulli distribution by the "success probability" π, $X \sim \mathrm{Ber}\,(\pi)$.
Similarly: $X \sim N(\mu, \sigma^2)$ (Normal distribution).

Statistical Inference

Statistical inference is the task to learn about an empirical phenomenon based on observed data.

Definition (Inference): The process of drawing conclusions about a population on the basis of measurements or observations made on a sample of individuals from the population (from Everitt - The Cambridge Dictionary of Statistics).

→ We are not only interested in the observed data (as in descriptive statistics) but want to draw conclusions on underlying general principles.

Examples

- Estimate the expected (average) tree height in all oil palm plantations in Indonesia (this is the population) based on a sample of observations.
  - Assume that the sample is (in some sense) representative for the population.
  - Use the sample mean

    $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

    as an estimator for the expected tree height in the population.

Statistical models

- Empirical phenomena are described via statistical models:
  - The result of such a phenomenon is represented as a random variable X with a certain probability distribution.
  - More random or non-random variables may be incorporated (e.g. regression model).
- We will assume parametric distributions for X such that statistical inference relies on the parameters of the distribution of X.
- Usual assumption: We observe independent and identically distributed (i.i.d.) replications $X_1, \ldots, X_n$ where n denotes the sample size.

What are the random variable(s) and the statistical model in your paper?

Three General Tasks in Statistical Inference:

Notation: Let θ be the only unknown parameter of interest of the distribution of X.

- **Point estimation**: Determine a suitable “guess” $\hat{\theta}$ for the true θ based on the observations $X_1, \ldots, X_n$. A good point estimate should
  - be as close to θ as possible.
  - vary not too much, i.e. should be reliable.
  - converge to the true value as more and more data get available.
- Interval estimation
- Statistical tests

Note: Model choice and prediction are not covered in this workshop.

Statistics & (point) estimators

- A function of random variables $g(X_1, ..., X_n)$ is called a **statistic**.
- The sample mean is an example of a statistic:

$$g(X_1, ..., X_n) = \bar{X} = \frac{1}{n}(X_1 + ... + X_n)$$

- Statistics used to estimate **unknown** quantities from a population are called **estimators**.
- The sample mean is an estimator for the expected value / location of a distribution.
- Statistics and estimators are functions of random variables and therefore themselves random variables. So they have a distribution determined by parameters, for instance expected value and variance.

Illustration (difference emerging by chance)

- Using R, we can draw samples from random variables for which we know the underlying true distribution function (in contrast to data from an empirical study).
- This gives us the ability, e.g., to investigate the variability in the statistic between repeated samples from the same distribution.
- For example by using the R function `rnorm`, we are able to generate two samples, `x` and `y` that come from the same normal distribution with equal expected value and equal standard deviation:

```
set.seed(123)
(x <- rnorm(7))
[1] -0.56047565 -0.23017749  1.55870831  0.07050839  0.12928774
[6]  1.71506499  0.46091621
(y <- rnorm(7))
[1] -1.26506123 -0.68685285 -0.44566197 1.2240818  0.3598138
[6] 0.4007715  0.1106827
```

Illustration (difference emerging by chance, continued)

If we calculate the means of these two samples, we see that they are different:

```
mean(x)
[1] 0.07462564
mean(y)
[1] 0.208622
```

- Since we know that both samples and both sample means were generated from the same distribution, we know that this difference in the sample means is solely a matter of chance.
- But in any real study, we don't know the underlying truth, and consequently cannot draw this conclusion.
- So we need tools to derive probabilities for obtaining certain differences, assuming an underlying truth.

Exercise (sample mean differences): The following R code draws random samples from the normal distributions and plots the differences in sample means.

- Vary the values for `mean_x`, `mean_y`, `sd_x`, `sd_y`.
- What conclusions can you draw?

```
smd <- function(mean_x, mean_y, sd_x, sd_y, ylim){
  samplesizes <- 5:1000
  result <- rep(0, length(samplesizes))
  for(i in 1:length(samplesizes)){
    x <- rnorm(samplesizes[i], mean = mean_x, sd = sd_x)
    y <- rnorm(samplesizes[i], mean = mean_y, sd = sd_y)
    result[i] <- mean(x) - mean(y)}
  plot(samplesizes, result, type = "n", ylim = ylim,
       main = paste("Means: (", mean_x, ", ", mean_y, "),
       SDs: (", sd_x, ", ", sd_y, ")", sep = ""))
  abline(h = 0, col = rgb(0, 92, 169, maxColorValue = 255), lwd = 2)
  lines(samplesizes, result)}
set.seed(123)
par(mfrow = c(1, 3))
smd(mean_x = 0, mean_y = 0, sd_x = 3, sd_y = 3, ylim = c(-3, 3))
smd(mean_x = 0, mean_y = 0, sd_x = 1, sd_y = 1, ylim = c(-3, 3))
smd(mean_x = 0, mean_y = 0, sd_x = 0.1, sd_y = 0.1, ylim = c(-3, 3))
```

Screw example

- A company produces a special sort of screws.
- The machine that makes 1000 screws a day is old and sometimes defective.
- How many screws produced can't be sold because of the failure?
- The random variable

$$X = \begin{cases} 1, & \text{if the screw has a failure} \\ 0, & \text{if the screw can be sold} \end{cases}$$

is repeatedly measured for one day (i.e. sample size $n = 1000$).

- Assumption: $X_i \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}\left(\pi\right), \; i = 1, \ldots, n$, i.e. the random variables $X_1, \ldots, X_{1000}$ are assumed to be independent from each other and identically distributed (i.i.d.).

Screw example, continued

- For the resulting random variables $X_1, ..., X_{1000}$ we define the statistics

$$g_1(X_1, ..., X_{1000}) = \bar{X} = \frac{1}{1000}(X_1 + ... + X_{1000})$$

and

$$g_2(X_1, ..., X_{1000}) = \sqrt{\overline{X^2}} = \sqrt{\frac{1}{1000}(X_1^2 + ... + X_{1000}^2)}$$

- Target: a good point estimator for the unknown failure probability π
- We use both statistics to estimate the probability π :

$$\hat{\pi}_1 = \bar{X}$$

$$\hat{\pi}_2 = \sqrt{\overline{X^2}}$$

Screw example, continued

- We observe that 144 screws have a failure $(X_1 = 0, X_2 = 0, X_3 = 1, ...)$.
- Therefore the parameter π is estimated

$$\hat{\pi}_1 = \bar{X} = \frac{144}{1000} = 0.144$$

$$\hat{\pi}_2 = \sqrt{\overline{X^2}} = \frac{12}{\sqrt{1000}} \approx 0.38$$

- Which estimator should we take?
- For the expected value $\mathbb{E}$ of the first estimator ("the relative frequency") it holds

$$\mathbb{E}(\hat{\pi}) = \mathbb{E}(\bar{X}) = \mathbb{E}(\frac{1}{1000}(X_1 + ... + X_{1000})) = \frac{1}{1000} \cdot 1000 \cdot \pi = \pi.$$

Unbiasedness

- An estimator $\hat{\theta} = g(X_1, ..., X_n)$ ($\hat{\theta} = \hat{\pi}$ in the previous example) is called unbiased if
$$\mathbb{E}(\hat{\theta}) = \theta.$$
- $\hat{\theta}$ is called biased if
$$\mathbb{E}(\hat{\theta}) \neq \theta.$$
- The quantity
$$Bias(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$$
is called the bias.
- Interpretation: If you repeatedly drew a sample from the population and computed the statistic, then the average over all of the sample statistics would (nearly) equal the population parameter.

Variance

- Unbiasedness ensures that the estimate is close to the true value "on average". In individual experiments, we may still observe large deviations from the true θ.
- We therefore consider a measure for the variation of $\hat{\theta}$ across different trials, i.e. a measure for the reliability of $\hat{\theta}$.
- The most well known example is the variance of an estimator:

$$\mathrm{Var}(\hat{\theta}) = E\left[(\hat{\theta} - E(\hat{\theta}))^2\right].$$

- The variance as a measure of reliability only makes sense for unbiased estimates. For example, the estimator $\hat{\theta} = 1$ has minimal variance $\mathrm{Var}(\hat{\theta}) = 0$ but does not even depend on the data.
- If an estimator has smaller variance than any other unbiased estimator, it is said to be efficient.

Mean squared error

- A combined measure for the quality of estimators is the mean squared error (MSE)

$$MSE(\hat{\theta}) = \mathbb{E}(\hat{\theta} - \theta)^2,$$

  which can be split up in the estimator's variance and squared bias:

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + Bias(\hat{\theta})^2.$$

- An estimator with

$$\lim_{n \to \infty} MSE(\hat{\theta}) = 0$$

  is called consistent, which implies that, for large samples, the sample estimate will be "very close" to the true value.

So far, we just chose some point estimators such as the sample mean without further justification or explaining their origin.

But is there a general strategy for obtaining “good” point estimators, even for complicated statistical models?

Maximum likelihood (ML)

- Again we think of an i.i.d. sample of random variables $X_1, ..., X_n$ from a distribution with probability function or density $f(x|\theta)$ with parameter θ.

- For a given data set $x_1, \ldots, x_n$ the likelihood function is

$$L(\theta) = f(x_1, ..., x_n|\theta)$$

  for continuous data and

$$L(\theta) = P(X_1 = x_1, ..., X_n = x_n|\theta)$$

  for discrete data, respectively.

- The likelihood function $L(\theta)$ gives the probability (density) of the observed sample for all different values of the unknown parameter θ.

Maximum likelihood point estimator

- Idea: Parameter values which make the observed data appear to be probable according to the given model are more likely to be correct than parameter values which make the data appear relatively improbable.
- Thus: For a fixed dataset $X_1 = x_1, ..., X_n = x_n$ with joint density $f(x_1, \ldots, x_n|\theta)$, the method of maximum likelihood selects the value of the model parameter θ that produces a distribution that gives the observed data the greatest density.
- As formula:

$$L(\theta) = f(x_1, ..., x_n|\theta) = f(x_1|\theta) \cdot ... \cdot f(x_n|\theta) \rightarrow \max_{\theta}$$

- Equivalent: Discrete data with joint probability $P(X_1 = x_1, ..., X_n = x_n|\theta)$.
- How do we maximize a function?

Deriving the ML point estimator

- Maximizing the likelihood by setting the first derivative to zero.

- Problem: Maximizing
$$L(\theta) = f(x_1|\theta) \cdot ... \cdot f(x_n|\theta).$$
would need the chain rule.

- Idea: Instead of maximizing $L(\theta)$ we maximize the log-likelihood (monotonous transformation of the likelihood).

$$\begin{aligned} l(\theta) = \log(L(\theta)) &= \log(f(x_1|\theta)) + ... + \log(f(x_n|\theta)) \\ &= \sum_{i=1}^{n} \log(f(x_i|\theta)) \end{aligned}$$

- Equivalent for discrete $x_1, \ldots, x_n$.

Likelihood example

- Is a tossed coin fair, i.e. $\pi = P(\text{“}tails\,up\text{”}) = 0.5$?

- Therefore you toss a coin ten times. Note how often $tails$ $(X_i = 1)$ is flipped.

$\Rightarrow$ $X_1, ..., X_{10}$ from an i.i.d. $\mathrm{Ber}(\pi)$ distribution.

- Some data $X_1 = 1, X_2 = 0, X_3 = 1, \ldots, X_{10} = 0$ is observed, for instance 3 times tails and 7 times heads.

Likelihood example, continued

- Using the probability function of the Bernoulli distribution:

$$f_\pi\left(x\right) = \pi^x \cdot \left(1-\pi\right)^{1-x},$$

we can calculate a likelihood by the product of all $n = 10$ likelihood contributions.

- Probability of that sample for a given π:

$$\begin{aligned} L(\pi) = P(X_1 = 1, \ldots, X_{10} = 0|\pi) &= P(X_1 = 1|\pi) \cdot \ldots \cdot P(X_{10} = 0|\pi) \\ &= \pi^1(1-\pi)^0 \cdot \ldots \cdot \pi^0(1-\pi)^1 \\ &= \pi^3\left(1-\pi\right)^7. \end{aligned}$$

- Exercise: Compute the ML estimator $\hat{\pi}_{ML}$ for this (or your) dataset!

Likelihood example, continued

We can illustrate this likelihood in R by plotting the complete path of $\mathrm{L}\left(\pi|x_1,\ldots,x_n\right)$ across $\pi \in [0,1]$:

```
tails <- 3
heads <- 7
pi <- seq(0, 1, length = 200)
par(mar = c(4.1, 6, 1, 1))
plot(pi, pi^(tails)*(1-pi)^(heads), type = "l", bty = "n", yaxt = "n",
     xlab=expression(pi), ylab = "",ylim=c(0,0.0025))
axis(2, las = 2)
mtext(side= 2, line = 4.5, "Likelihood")
# For which value of pi the occurence of the data set at hand is most likely?
abline(v=tails/(tails+heads), col=2,lwd=2)
# For comparison: a fair coin
abline(v=0.5, col=1,lwd=2)
```

Likelihood example, continued

We can use the likelihoods under two hypothesis H_0 and H_1 to calculate the **likelihood ratio**, for instance $H_0 : \pi = 0.5$ and some value under H_1 (choose one):

```
# Likelihood of Maximum-likelihood estimate
max(pi^(tails)*(1-pi)^(heads))
# Likelihood of fair coin (black)
.5^(tails)*(1-.5)^(heads)
# Likelihood ratio
max(pi^(tails)*(1-pi)^(heads))/(.5^(tails)*(1-.5)^(heads))
abline(h=max(pi^(tails)*(1-pi)^(heads)),col=2,lwd=2,lty=2)
abline(h=(.5^(tails)*(1-.5)^(heads)),col=1,lwd=2,lty=2)
```

A likelihood ratio gives **relative** evidence for H_1 vs. H_0 (and may be used for hypotheses testing). Both hypotheses might be unlikely.
Exercise: Choose an example where the likelihood ratio is large but both hypotheses are unlikely!

Properties of the ML estimator (without derivation)

The Maximum likelihood estimator is:

- not generally unbiased;
- but generally asymptotically (for big sample size) unbiased;
- consistent;
- asymptotically efficient;
- asymptotically normally distributed.

Question: How many observations do we need to live in $asymptotia$?

Calculating point estimates without illustrating the uncertainty surrounding that estimate is somewhat insufficient.

Three general tasks in statistical inference:

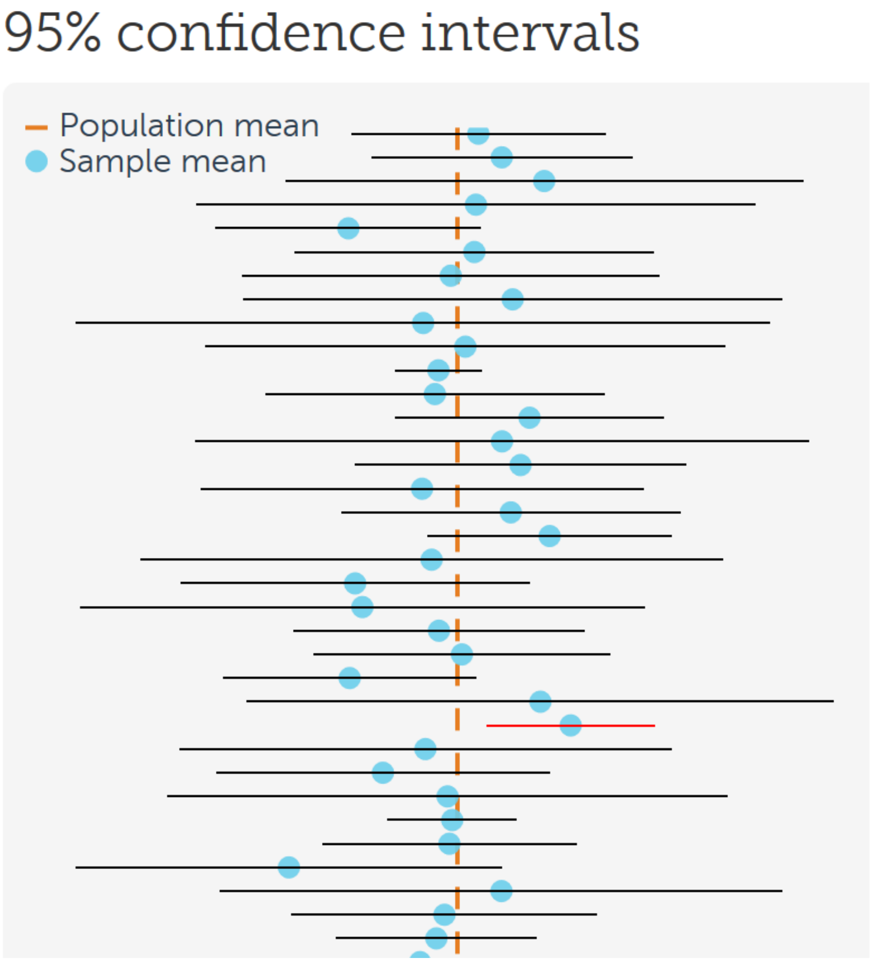
- Point estimation
- **Interval estimation**: Determine an interval $[\hat{\theta}_{\mathrm{low}}, \hat{\theta}_{\mathrm{up}}]$ based on the data such that the interval covers the true value θ with a high probability. A good interval estimate should
    - maintain the desired coverage probability.
    - be as narrow as possible given the coverage probability.
    - get smaller and smaller as the sample size increases.
- Statistical tests

Confidence intervals

- Confidence intervals are random variables. Why?
- After collecting the data a confidence interval either contains the population parameter θ or not.
- Frequency interpretation of a 95% confidence interval: If the data generating experiment is repeated again and again, a fraction of 95% of the resulting confidence intervals will cover the true parameter.
- Alternative frequency interpretation: There is a 95% probability that when I compute a confidence interval from data of this kind (by performing the same experiment again and again or taking samples from the population again and again), the true value θ will be covered by the confidence interval.
- Confidence intervals are a statement about the percentage of (future) confidence intervals that contain the true parameter value.

Confidence intervals: Visualization

`http://rpsychologist.com/d3/CI/`



- In practice, one usually uses confidence intervals of the form

$$\hat{\theta} \pm q_\alpha \sqrt{\mathrm{Var}(\hat{\theta})}$$

  where q_α is a constant chosen such that the coverage probability (e.g. 95%) holds.

  - The confidence interval is symmetric around the estimate $\hat{\theta}$.
  - The width of the confidence interval is proportional to the standard deviation of the estimate.

- Given the coverage probability, a confidence interval (or region) should be as small as possible.

Example: Confidence interval for the mean of a normal distribution.

- Let $X_1, \ldots, X_n$ be i.i.d. $\mathrm{N}(\mu, \sigma^2)$.
- If σ^2 is known, a $(1 - \alpha)$ confidence interval for μ is given by

$$\bar{x} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

where $z_{1-\frac{\alpha}{2}}$ is the $z_{1-\frac{\alpha}{2}}$-quantile of the standard normal distribution.

Maximum likelihood confidence intervals

- Maximum likelihood confidence intervals around the point estimator typically rely on the asymptotic normal distribution of the ML estimator.

- For small or moderate sample sizes, this can be problematic.

- Alternative 1: Bootstrap (highly generic and powerful tool).

- Alternative 2: Construct interval solely based on observed likelihood function.

Alternative 1: Bootstrap percentile confidence intervals

- Assume that $X_1, \ldots, X_n$ are i.i.d. replications from a distribution with expectation μ.
- An asymptotic maximum likelihood confidence interval for μ is then given by

  $$\bar{x} \pm z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$$

  i.e. we treat the estimator $\bar{x}$ as if it was approximately normally distributed.
- Bootstrap procedures avoid this assumption and are based on resampling (with replacement) from the given sample.

Example: Algorithm for a (nonparametric) bootstrap

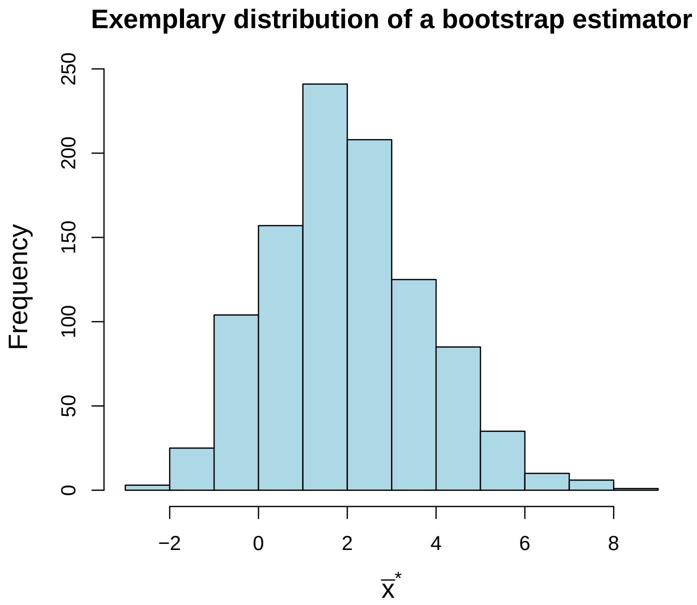
- Sample n observations $x_1^*, \ldots, x_n^*$ with replacement from the observed values $x_1, \ldots, x_n$ (this is called a bootstrap sample).
- Compute the average in the bootstrap sample, i.e.

$$\bar{x}^* = \frac{1}{n} \sum_{i=1}^{n} x_i^*.$$

- Repeat these steps many times (say B times) to obtain the averages $\bar{x}_1^*, \ldots, \bar{x}_B^*$.
- Use the $\alpha/2$ and $1 - \alpha/2$ quantiles from the sample of averages as lower and upper bound of the confidence interval.

Idea of a (nonparametric) bootstrap

- The underlying idea is as follows: If we could repeatedly draw random samples from the true population model, the above procedure would yield correct confidence intervals.
- Since the true population model is not available, we approximate it by the empirical distribution (by drawing with replacement from the observed sample and calculating the estimator each time).



Alternative 2: Confidence intervals based on likelihood function

- Recall: Likelihood function gives for all possible values of the unknown θ (or π) the probability / density to generate the dataset at hand.
- Compute a 95% confidence interval such that the range of possible values for θ (or π) in the resulting interval has the cumulative probality of 95% to generate the dataset at hand.

Likelihood example, interval based on likelihood function

```
tails <- 3
heads <- 7
pi <- seq(0, 1, length = 200)
par(mar = c(4.1, 6, 1, 1))
plot(pi, pi^(tails)*(1-pi)^(heads), type = "l", bty = "n", yaxt = "n",
     xlab=expression(pi), ylab = "",ylim=c(0,0.0025))
axis(2, las = 2)
mtext(side= 2, line = 4.5, "Likelihood")
# For which value of pi the occurence of the data set at hand is most likely?
abline(v=tails/(tails+heads), col=2,lwd=2)
# For comparison: a fair coin
abline(v=0.5, col=1,lwd=2)
pi <- seq(0, 1, length = 500)
L <- function(pi){
  pi^(tails)*(1-pi)^(heads)
}
L_integral <- integrate(L, lower = 0, upper = 1)[[1]]
cdf_p <- diff(pi)[1]*cumsum(L(pi))/L_integral
p_lo <- pi[which.min(abs(cdf_p - 0.025))]
p_up <- pi[which.min(abs(cdf_p - 0.975))]
## Approximate 95% central interval for pi is then:
abline(v = print(c(p_lo, p_up)), lty = 2)
```

The resulting interval has a direct bayesian interpretation, as we will see later.

Three general tasks in statistical inference:

- Point estimation
- Interval estimation
- **Statistical tests**: Decide whether a hypothesis about the parameter θ is true or not. Typical examples include hypotheses of the form

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq \theta_0$$

$$H_0 : \theta \leq \theta_0 \quad \text{vs.} \quad H_1 : \theta > \theta_0$$

$$H_0 : \theta \geq \theta_0 \quad \text{vs.} \quad H_1 : \theta < \theta_0$$

with a fixed, prespecified value θ_0 (often: $\theta_0 = 0$). A good test should

- often make the right decision.
- tend to get better when the sample size increases.

Example: Tests for the mean of a normal distribution

- Let $X_1, \ldots, X_n$ be i.i.d. $\mathrm{N}(\mu, \sigma^2)$ with σ^2 known.
- We are interested in tests of the form

$$
\begin{aligned}
H_0 &: \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu \neq \mu_0 \\
H_0 &: \mu \leq \mu_0 \quad \text{vs.} \quad H_1 : \mu > \mu_0 \\
H_0 &: \mu \geq \mu_0 \quad \text{vs.} \quad H_1 : \mu < \mu_0
\end{aligned}
$$

- A suitable test statistic for all three problems is

$$T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}.$$

- Example: For the one-sided test $H_0 : \mu \geq \mu_0$ vs. $H_1 : \mu < \mu_0$, large negative values of T (corresponding to $\bar{X} < \mu_0$) are an indication against H_0 and in favour of H_1.

Error types in statistical tests

- When performing a statistical test, we can make the following decisions:

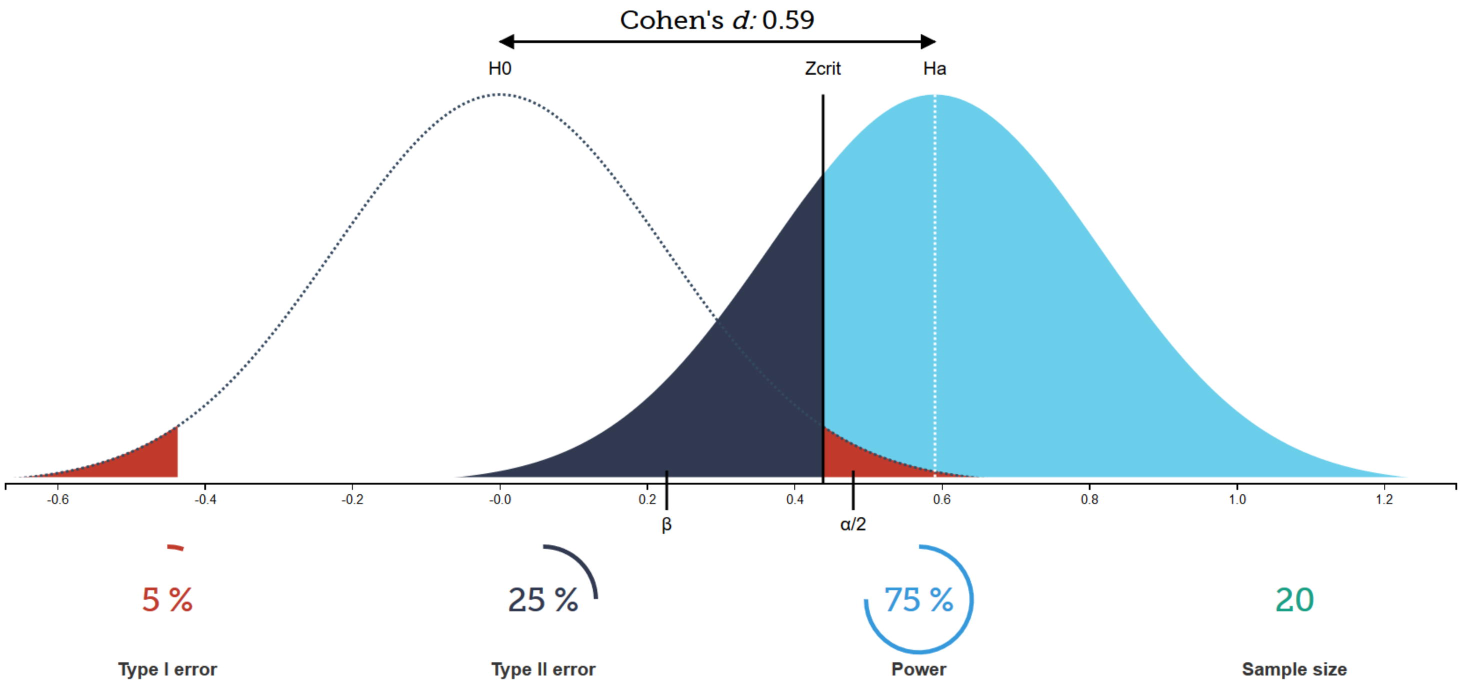
	H_0 true	H_1 true
retain H_0	$\surd$ $(1-\alpha)$	type II error (β)
reject H_0	type I error (α)	$\surd$ $(1-\beta)$

  - type I error: reject H_0 although H_0 is in fact true.
  - type II error: H_0 is not rejected although H_1 is true.
- In practice, it is usually not possible to make the probabilities for both types of errors (i.e. α and β) small simultaneously.

  $\Rightarrow$ Specify an upper bound α for the type I error and minimize the probability β for the type II error given this constraint.
- The upper bound for the type I error is called the significance level (often $\alpha = 0.05$).
- $1 - \beta$ is the probability of a significant result when H_1 is true (statistical power).

Error types in a statistical test: Visualization

`http://rpsychologist.com/d3/NHST/`



- Cohen's d is the effect size measuring the standardized difference between two means.
- Task: Play around with the sample size, error probabilites and effect size.

Error types in statistical tests

- It is desirable that asymptotically type 1 and type 2 probabilities tend to zero (consistency of a test).
- On the other hand, it also indicates that, for large sample sizes, we will be able to find arbitrarily small deviations from H_0 statistically significant.
- In addition to statistical significance, estimated effects should also be relevant from a subject matter perspective.
- In practice: Which error rate should you minimize?
- Note: Type 1 and type 2 errors hold again only if you repeat the **same** experiment/study many times. Does this happen in practice?

A meta-view on error rates

- Another (meta-) view on error rates: Scanning some journals in your field including many **different** experiments.
  Before the next study, what result can you expect if the probability in your field of research that H_0 is true is 70%? For all of the studies in your field, let $\alpha = 5\%$ if H_0 is true and $1 - \beta = 80\%$ for some true effect under H_1:

		truth	
		H_0 true	H_1 true
decision	retain H_0	$95\% * 70\% = 66.5\%$	$20\% * 30\% = 6\%$
	reject H_0	$5\% * 70\% = 3.5\%$	$80\% * 30\% = 24\%$

- The most likely finding of your next study is a correctly non-rejected H_0.
- But: If you find a significant effect, how likely is it that this finding is wrong? $\alpha = 5\%$? Calculate!

- In this example, the probability that a significant finding is false in your field of research is $\frac{3.5\%}{3.5\%+24\%} \approx 12.7\%$.
- Often, studies have low power such that many more than 5% of the significant findings are false.
- Significant findings in your field of research are more likely to be indeed true if there are
  - fewer studies with no real effects, i.e. H_0 is not true very often.
  - for given α: many studies with high power (large samples, big real effects).
  - fewer incentives and possibilites for p-hacking (discussed later).
- Do not overrate the evidence of a single significant effect.

More to read on this topic

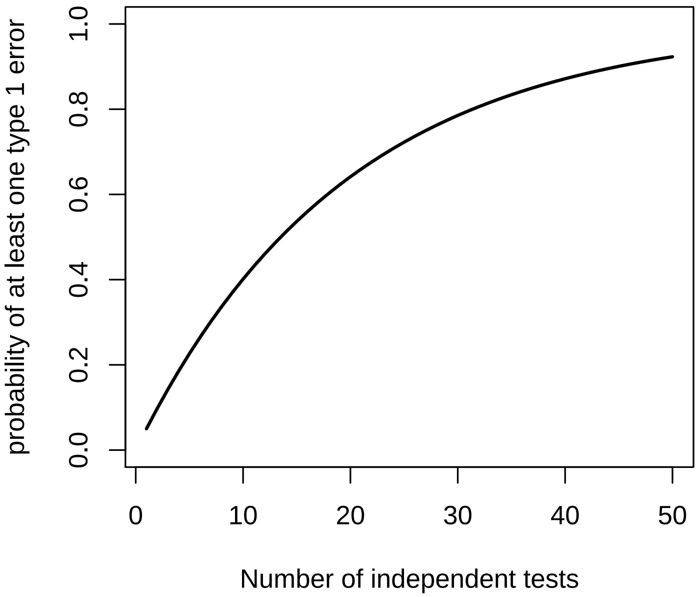
- Colquhoun (2014): *An investigation of the false discovery rate and the misinterpretation of p-values*
- Ioannidis (2005): *Why Most Published Research Findings Are False*
- Sterne & Smith (2001): *Sifting the evidence-what's wrong with significance tests?*

Exercise positive predictive value

Note: P-hacking will be discussed later.

Type 1 error control

- Set an upper limit α for the type 1 error (before the study), e.g. $\alpha = 0.05$.
- Multiple tests increase the type 1 error rate.
- For four (independent) tests, the type 1 error rate is: $1 - (0.95)^4 \approx 0.19 > 0.05$



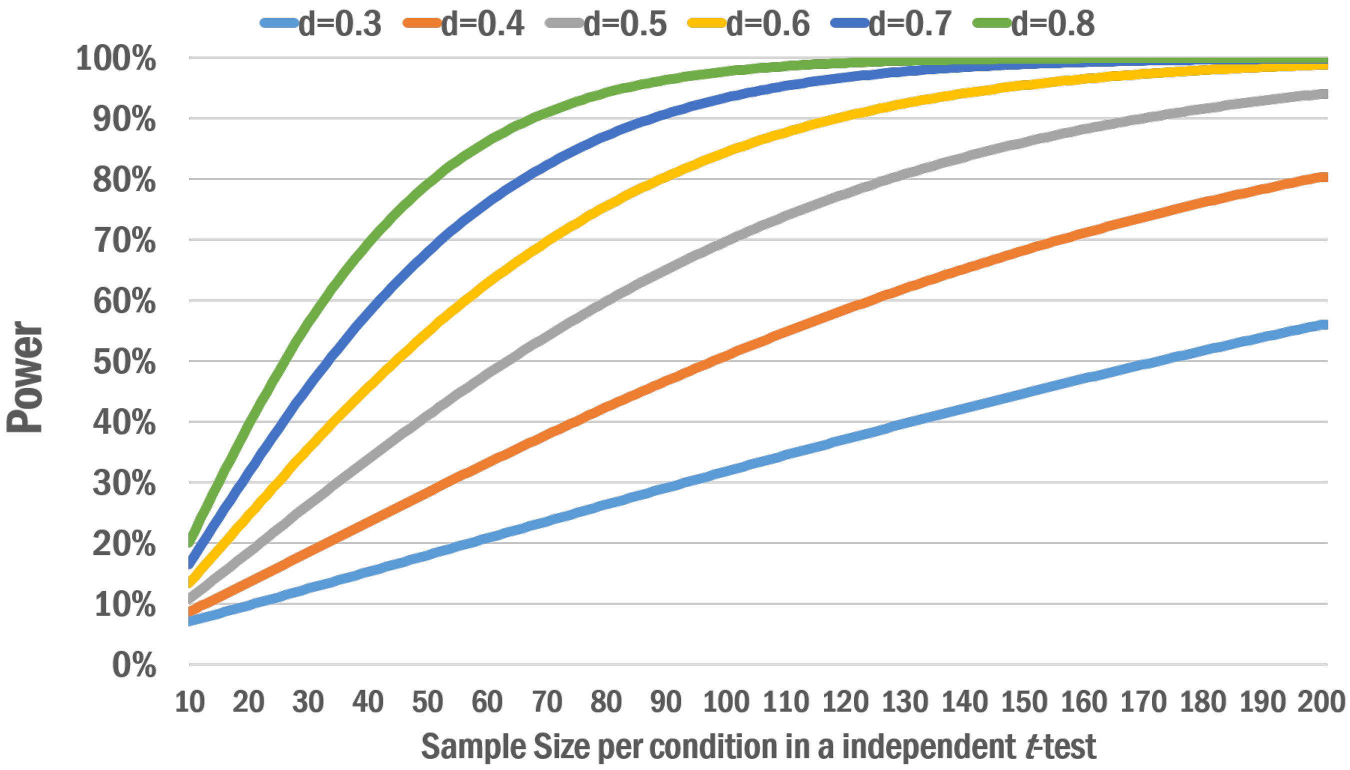
- Multiple testings corrections (Bonferroni, Holm, etc.) may help.

Type 2 error control

- Type 2 errors are more difficult to control than type 1 errors, but may be just as important.
- Studies with low power do not yield much information and can be harmful from a meta perspective (see previous slides and `http://www.nature.com/nrn/journal/v14/n5/full/nrn3475.html?foxtrotcallback=true`).
- Power can be increased by
  - larger samples
  - decreasing measurement error
  - using one-sided tests
  - using within-designs (measure same observations over time) when high within correlation is likely

Type 2 error control: Sample size planning

- Only possible in experimental research.
- How do you determine the sample size for a new study?
- You need to justify the sample size of a study. What goal do you want to achieve?



from Daniel Lakens

- Which **unknown** true effect should be assumed? Take it from the literature?
- You should decide on a minimal true effect which you are interested in.
- Assuming this effect, a statistically siginificant result should then be detected in a certain fraction of many repeated studies with the same sample size (power).
- Of course, feasibility matters.
- For difficult models: Ask a statistician for a power analysis before collecting data.

P-values

- The p-value is the probability of getting the observed or more extreme data (if repeating the experiment again and again), assuming the null hypothesis is true.
- The p-value is small, if the observed data or more extreme data is very unlikely under the null hypothesis.
- The p-value can also be interpreted as the smallest significance level, for which we would be able to reject H_0.
- We therefore have the equivalences
  - p-value $\leq \alpha \Leftrightarrow$ reject H_0.
  - p-value $> \alpha \Leftrightarrow$ retain H_0.
- $p > \alpha$ does not mean there is no true effect. You need large samples to detect small effects.

Example: Two-sided test for the mean of a normal distribution

- Assume that we have observed

$$t_{\text{obs}} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}.$$

- Which values of T provide at least as much evidence against H_0 as t_{obs}?

$\Rightarrow$ Those with $|T| \geq |t_{\text{obs}}|$.

- The p-value is then given by

$$P_{H_0}(|T| \geq |t_{\text{obs}}|) = P_{H_0}(T \geq |t_{\text{obs}}|) + P_{H_0}(T \leq -|t_{\text{obs}}|)$$

where $T \sim \mathrm{N}(0,1)$ (under H_0).

Two ways to perform a statistical test:

- P-value: Reject if the p-value is smaller than the significance level α.
- Confidence interval: Reject if θ_0 is not contained in the confidence interval.
- Ergo: Confidence intervals and p-value lead to the same test decisions. Differences?

Exercise Confints

P-values

- P-values tell you how surprising the data is, assuming H_0 is true.
- Why can‘t we just say: The p-value is the probability of H_0 being true?
- You can‘t get the probability the null hypothesis is true, given the data, from a p-value (Bayesian Statistics needed for this):

$$P(\text{observed data or more extreme data}|H_0 \text{ is true}) \neq P(H_0 \text{ is true}|\text{data})$$

Further p-value misconceptions

From Goodmann, 2008: *A Dirty Dozen*

- A statistically significant finding is important from a subject matter perspective.
- Studies with p-values on opposite sides of 0.05 are conflicting.
- P-values are properly written as inequalities (e.g., $p < .02$ when $p = 0.015$).
- A scientific conclusion or treatment policy should be based on whether or not the p-value is significant.

Discuss!

More about p-values

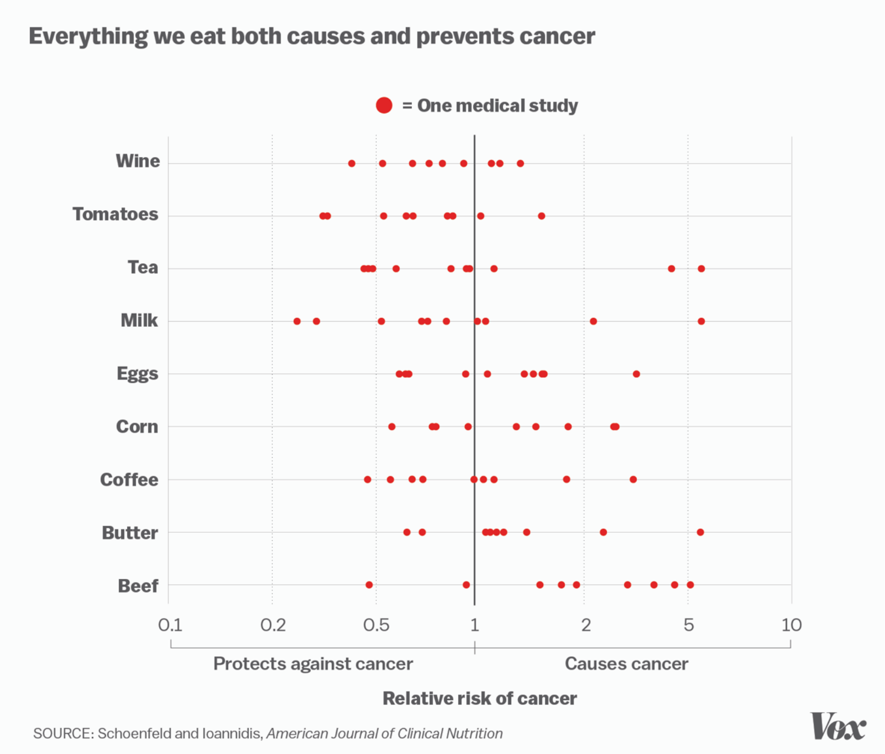
- A p-value tells us how compatible the observed (and more extreme) data are with a specified hypothesis H_0.
- But is H_0 (often: “no effect”) bravely and reasonably chosen?
- If H_0 is rarely true, rejecting it does not tell us much.
- So shouldn’t we ban p-values like the journal *Basic and Applied Social Psychology*? (`http://www.nature.com/news/psychology-journal-bans-p-values-1.17001`)

Recommendations for the practice

- Refrain from making binary decisions (yes / no).
- Do not report only p-values (and report actual p-values, not $p < 0.05$ or the like).
- Report additionally point estimators, effect sizes, confidence / bootstrap intervals, full likelihoods, likelihood ratios, Bayesian tools (later), ...
- Read Greenland et. al (2016): *Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations*.
- Interpret your results critically and correctly.
- Report the uncertainty instead of claiming certainty.
- Do not care too much about one single study (except for the case when the sample is large).

Meta-analytic thinking

- Often, several studies related to one topic yield mixed results, but are more meaningful than single studies.



SOURCE: Schoenfeld and Ioannidis, *American Journal of Clinical Nutrition*

- Open LikelihoodResultsApp.R

Exercise: Meta-analytic thinking

Think of a journal in your field of research. Assume an average power of 50% to detect some true effect opposing the null hypothesis. Then, only about 3.125% of all articles that report 5 experiments / significance tests should contain exclusively significant results (if the alternative hypothesis is true). If you pick up a random issue and a study from the journal, and see an article reporting 5 significance tests which all yield statistically significant results, would you trust the reported findings more, or less, than when all these articles had reported mixed results? Why?

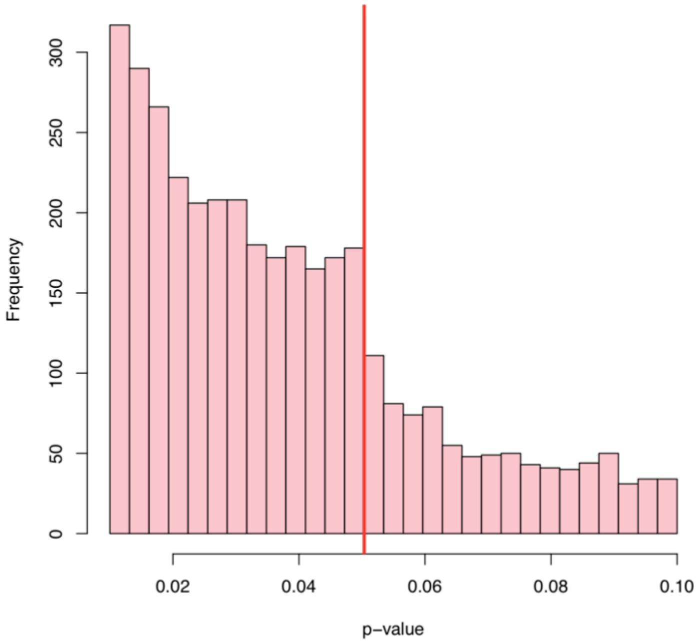
Meta-Analytic thinking

- If possible, report all data and perform a small scale meta-analysis when publishing.
- But: Do I find all relevant studies related to one topic in the literature?

Publication bias

- Due to the strong focus on results with $p < 0.05$, tests with p-values below 0.05 are much more likely to be published than those above 0.05 (publication bias).

- Papers with “non-significant results” less likely to be accepted by journals.



- There are tools for detecting and to some extent correcting publication bias.

P-hacking

- P-hacking ("flexibility" in data analysis) as a response to the publication system.
- Opportunities to hack p-values:
  - selecting the model that gives you the results you want
  - selecting a dependent variable that yields a "significant effect"
  - data manipulation (e.g. deleting "adverse observations")
- Note: Not all sloppy research necessarily originates from deliberate p-hacking.
- Note: The hacking strategies also apply to confidence intervals etc.!
- Some literature:
  - Brodeur et al. (2016): *Star Wars: The Empirics Strike Back*
  - Simmons et al. (2011): *False-Positive Psychology*

Exercises p-hacking

- Run p-hacking.R.
- Take a look at your study. How could the author could have hacked the p-value? Did she / he hack the p-value? What do you think?

Detecting and preventing p-hacking and sloppy research

- **Before** data collection: Study pre-registration / pre-analysis plan
  - Describe experimental design, primary outcome variable, multiple testing procedures, statistical model(s), planned number of observations, power calculations, ...
  - Difficulties?
  - Read Olken (2015): *Promises and Perils of Pre-Analysis Plans*.
  - Site for pre-registration: `https://www.socialscienceregistry.org/`.
  - What about your field? Do you know journals which require pre-registration?
  - Are pre-analysis plans useful in observational research where you rely on secondary data? Consequences?

Detecting and preventing p-hacking and sloppy research

- **After** data collection: Data analysis
  - Is the pre-analysis plan followed? Are there justified deviations?
  - Without pre-analysis plan: are there many possible models but only one is reported (which might even not be the model you would choose)?
  - Are data and software code available (threat of replication)?
  - Is the study replicable?
  - When replicating (highly recommended): Are the findings robust (when applying different models, removing unreasonable outliers, etc.).
- Homework: Replicate a study!

Are reported p-values reasonable?
Do the exercise on p-values under H_0 and H_1.

In a nutshell: How should you do empirical research?

- If possible: Publish pre-analysis plans, think about your study beforehand.
- If possible: Use large sample sizes or include data from related studies to conduct a meta-analysis.
- Be critical with yourself.
- Report all your models and findings.
- Report more than p-values.
- Avoid dichotomous thinking (effect or no effect).
- Explain the pitfalls of your study, be honest and interpret correctly.
- Provide the data and well commented software code.
- Be critical with other research, conduct replications if possible.

Is the statistical philosophy we are using wrong?

So far: Frequentist view on empirical research.

- What can we expect if we repeat an experiment many times? (p-value, power, confidence intervals, unbiasedness, ...)
- P-value: The p-value is the probability of getting the observed or more extreme data, assuming the null hypothesis is true:
  $P(\text{observed data or more extreme data}|H_0 \text{ is true})$.
- Similar: The likelihood function gives probabilites to observe the data set at hand for all possible values of the unknown parameter(s) and not vice versa:

$$L(\theta) = P(\text{data}|\theta) \neq P(\theta|\text{data})$$

- Actually, we would like to have the probability that a hypothesis is true (that a parameter has a certain value) - given the data set at hand...

Bayes' theorem

- Bayes' Theorem gives us exactly what we want to have:

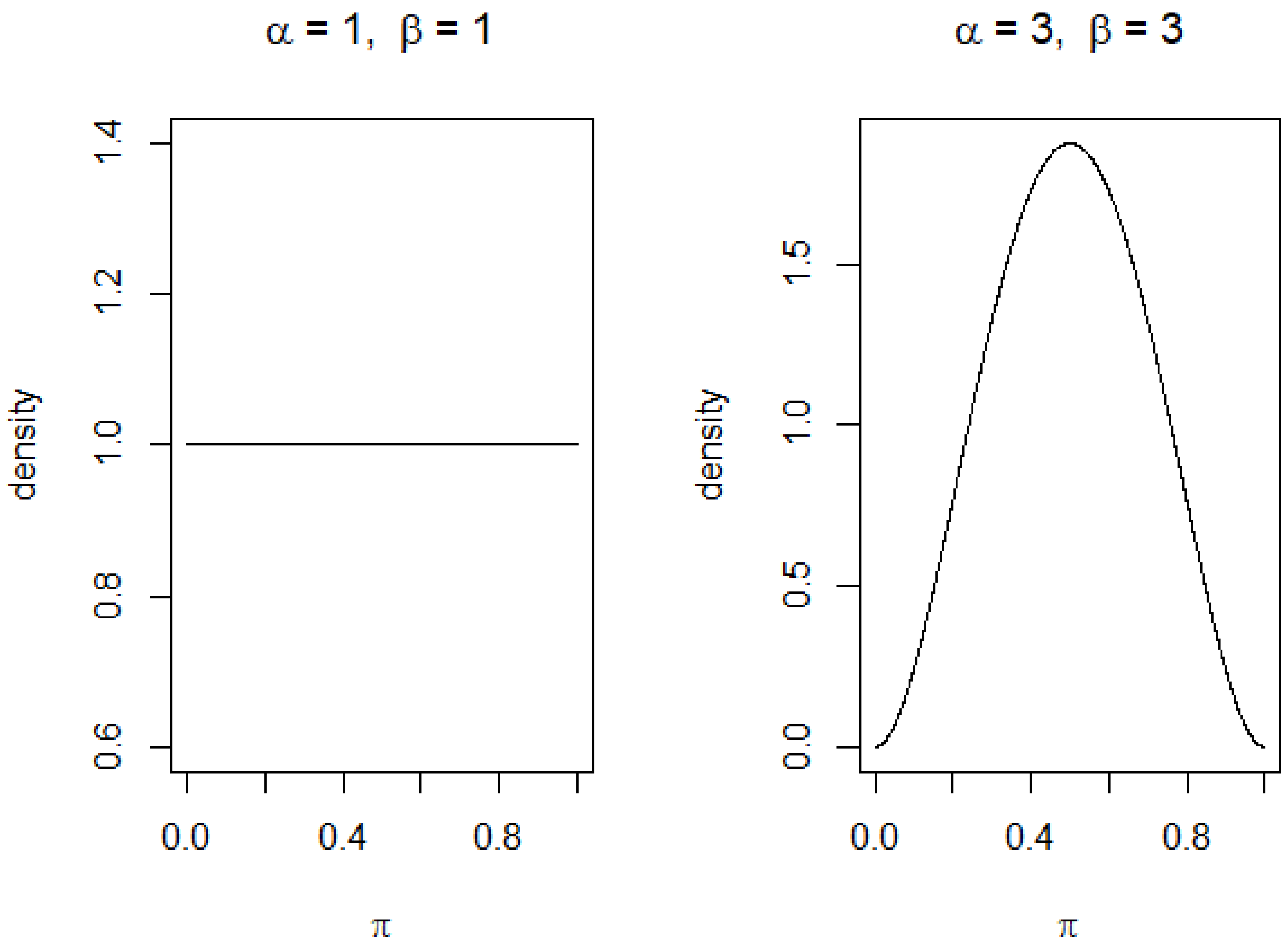
$$P(\theta|\mathrm{data}) = \frac{P(\mathrm{data}|\theta) * P(\theta)}{P(\mathrm{data})}$$

- The denominator $P(\mathrm{data})$ is just a constant for normalization, all we need to specify are the **likelihood** $L(\theta) = P(\mathrm{data}|\theta)$ and a belief about θ before data collection: the **prior** $P(\theta)$.
- The distribution of interest $P(\theta|\mathrm{data})$ is called **posterior** distribution.

Bayesian statistics: Coin toss example

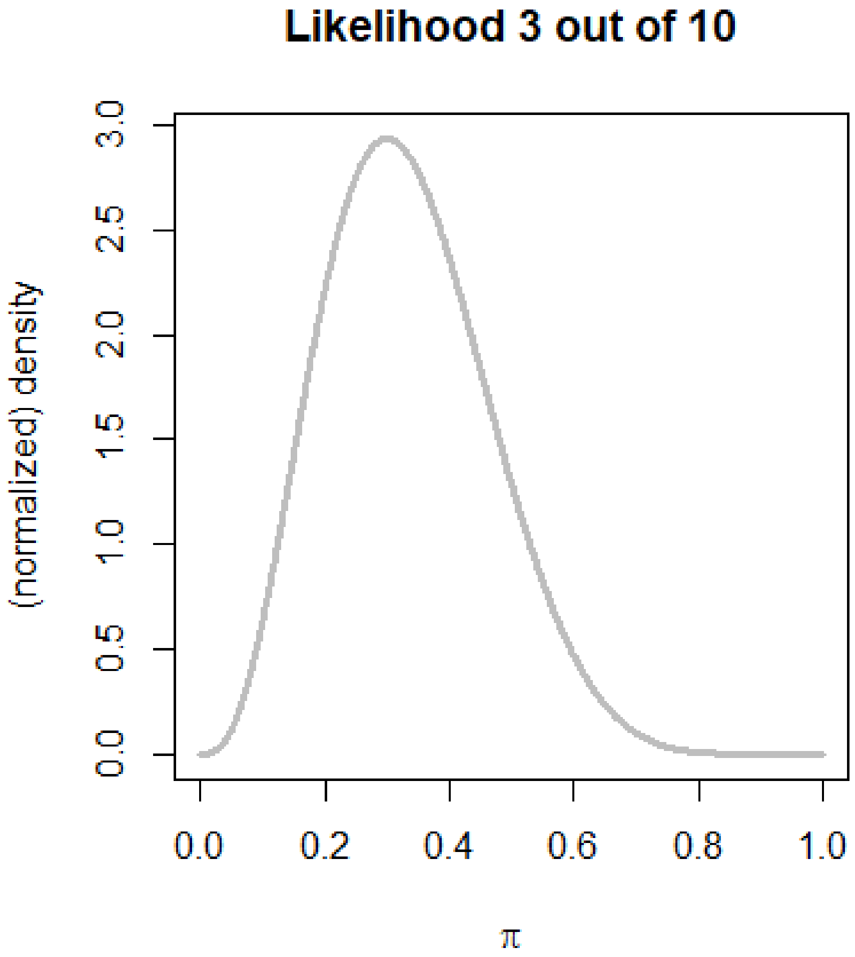
- What is the probability π for “tails up”? Here: $\theta = \pi$.
- For the prior, a beta distribution is used (ensures values for π between 0 and 1). The beta prior is determined by two parameters α and β: $B(\alpha, \beta)$.
- Note: α and β are not error types here!
- When we expect intermediate values for π to be more likely, e.g. choose $B(3, 3)$.
- Flat prior (also misleadingly called noninformative prior): $B(1, 1)$.

Coin toss example: Beta priors

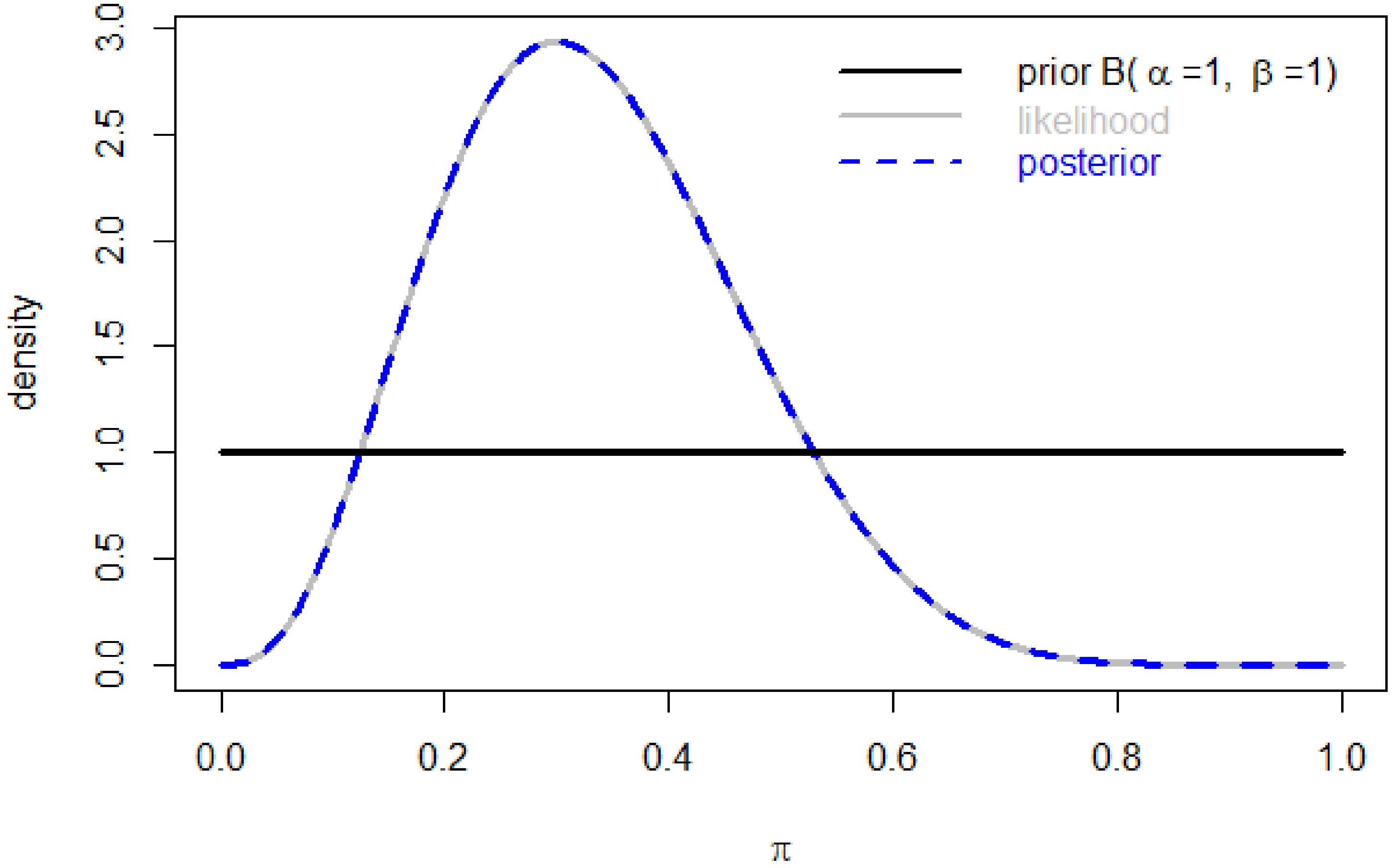


Coin toss example: Likelihood

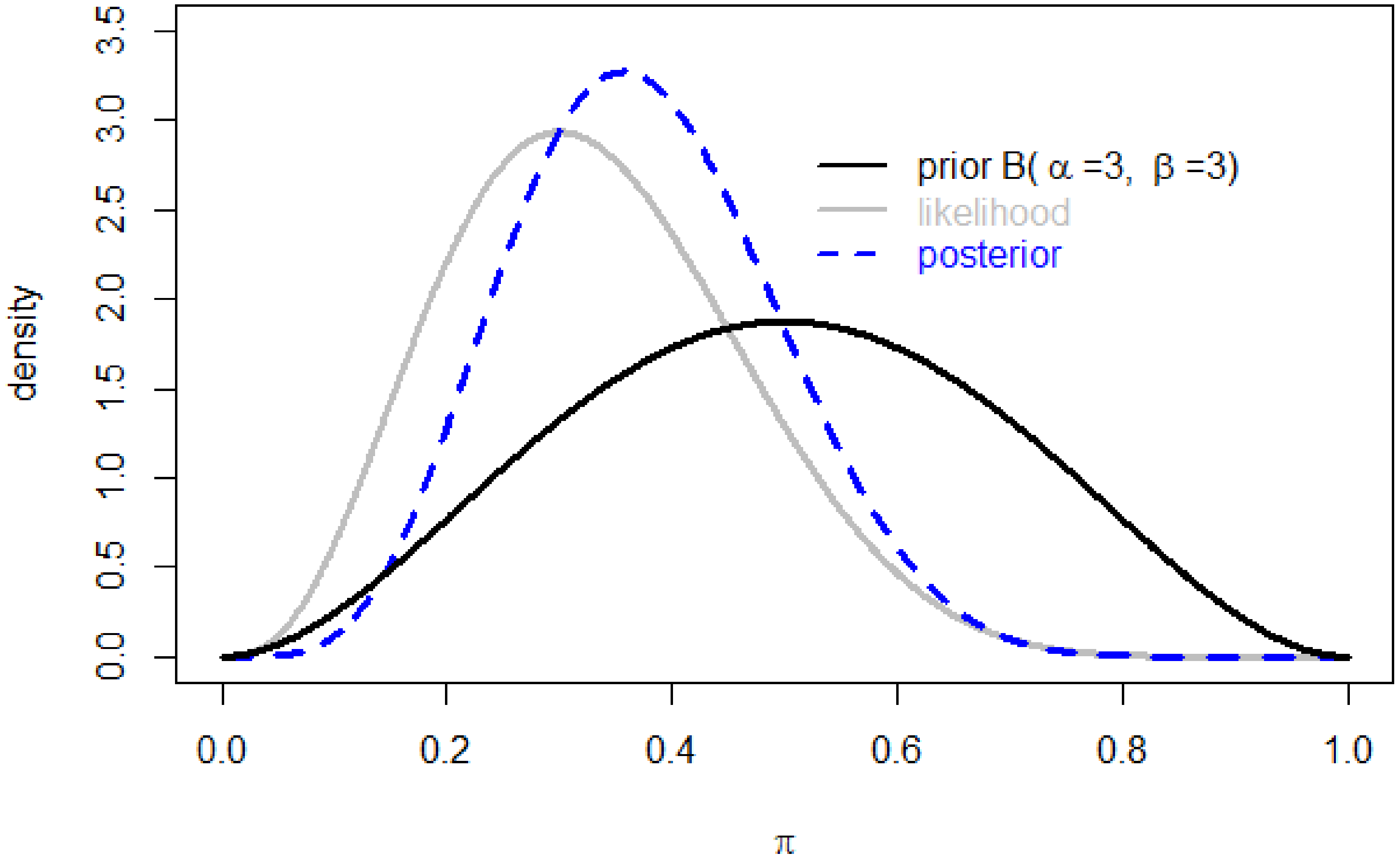
Assume: 10 toin cosses, 3 times tails.



Coin toss example: Posterior distributions



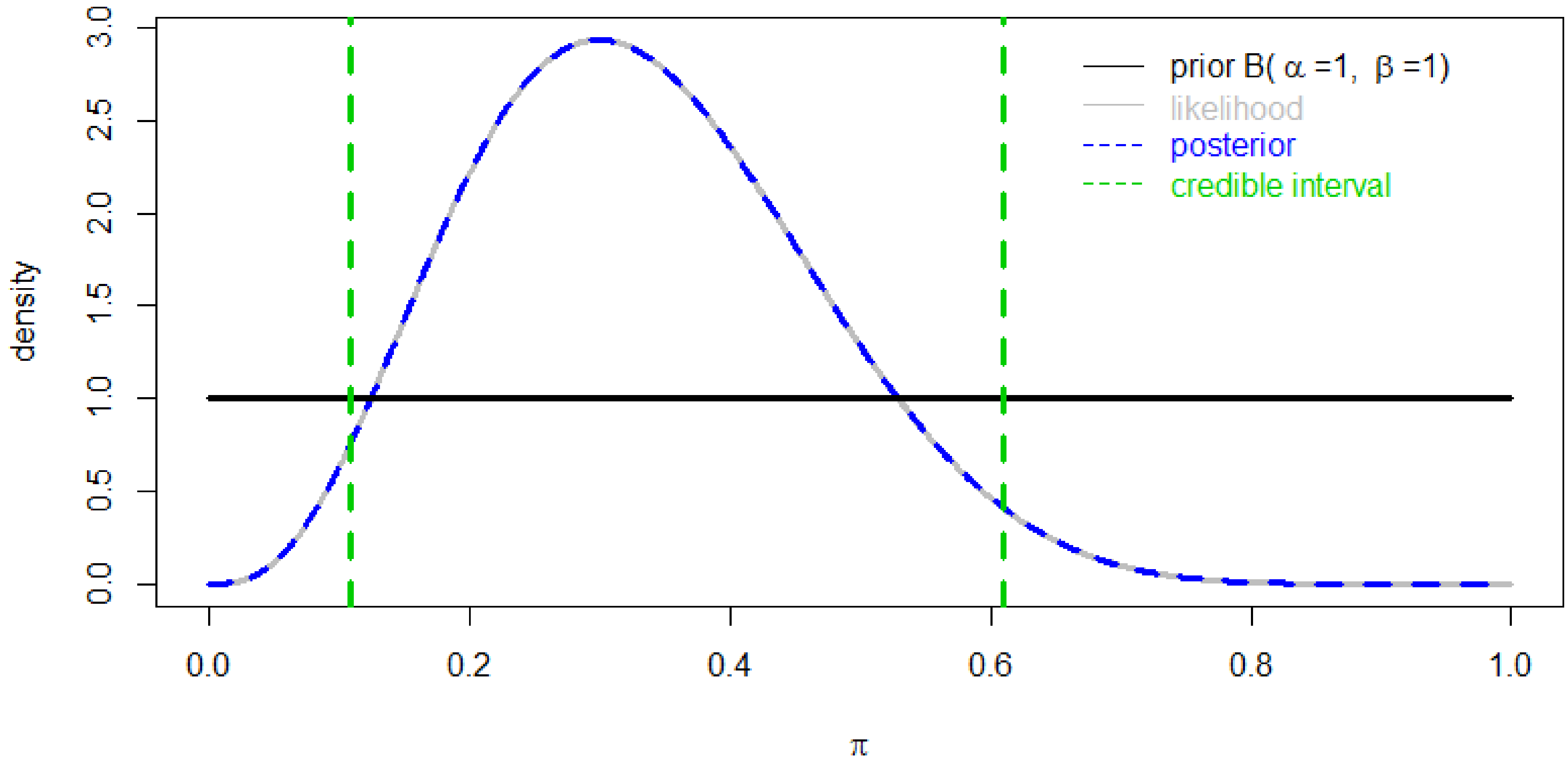
Coin toss example: Posterior distributions



Bayesian statistics

- Bayesian statistics allow you to update prior beliefs by seeing the data.
- We know how to determine a likelihood, but how should we decide for a prior for θ?
  - Often we have some idea about a reasonable prior.
  - Subject matter knowledge may help.
  - Check sensitivity of posterior distribution to different priors.
  - The choice of the prior loses its importance for growing sample size.
- You obtain probabilities / densities for possible values of the unknown parameter π after seeing the data.
- You can also use the posterior to calculate so-called credible intervals which cover 95% of the most plausible values for π.

Credible intervals



Credible intervals vs. confidence intervals

The **credible interval** says that some percentage (e.g. 95%) of the posterior distribution for a parameter π lies within a particular region:

'Given our observed data and the prior, there is a probability of 95%
that the true value of π falls within the credible region.'

We can directly compare this to the interpretation of a 95% **confidence interval**:

'If the data generating experiment is repeated again and again, a fraction of 95% of
the resulting confidence intervals will cover the true value of π. '

Interestingly, the credible interval on the previous slide is the same as the confidence interval based on the likelihood function (as computed earlier). The reason for that is the flat prior.

Direct bayesian interpretation of a likelihood

From Aitkin (2010): Statistical inference: an integrated Bayesian/likelihood approach, *CRC Press*, section 1.4.2.:

> 'The theory [to Bayesian inference approaches] requires that we express prior information as a probability distribution.
>
> In many cases, we may not have well-developed information or views which are easily expressed as a probability distribution, and much use is made, by many Bayesians, of weak or non-informative priors, which are 'uninformative' relative to the information in the data: the data were presumably collected to obtain information about parameters for which we had little prior information, and so the prior should reflect this lack of information.
>
> A non-informative prior for the case of two parameter values would be one with equal prior probabilities, leading to the posterior odds being equal to the likelihood ratio.'

So with using a flat prior ($f(\pi) \propto 1$), we can directly move from a likelihood for π to a posterior density for π.

Bayesian vs. frequentist thinking

- Bayesian reasoning and interpretation nicer than frequentist one.
- When prior information available
  - Bayesian statistics are preferable.
  - Importance of prior does not matter much for large sample sizes.
- When no prior information available
  - Analogy between Bayesian approach and approach based on likelihood function can be shown (but Bayesian computation advantageous for complex models).
  - Bootstrap is a valuable (frequentist) alternative even when sample sizes are small and models are complex.
  - For large sample sizes, asymptotic maximum likelihood theory sufficient (this is what is usually implemented for common functions in statistical software).

There are other opinions... Read more if you are interested:

- Goodman: *A Dirty Dozen: Twelve P-Value Misconceptions*
- Kruschke (2017): *The Bayesian New Statistics*
- Morey et al. (2016): *The fallacy of placing confidence in confidence intervals*
- Wagenmakers et. al: *Bayesian Versus Frequentist Inference* (in Hoijtink et al. (2008): *Bayesian Evaluation of Informative Hypotheses*)

Some concluding remarks

- So far, we talked a lot about statistical inference.
- We did not say that the choice of the used statistical model (e.g. the likelihood, the explanatory variables, ...) is often difficult.
- In practice, there are many possible models, potentially relevant variables, ...
- It is of utmost importance to think about the appropriateness of your model (checking model assumptions might sometimes help).
- Even more important: Think about your study beforehand
  - What is the population behind my sample?
  - What is the question I want to answer and is this possible with my data?
  - Is it possible to identify a causal effect?
  - ...

Some concluding remarks

- It is long way to improve quantitative research.
- Empirical research is far more uncertain than usually claimed.
- Support and take part in reproducible science (share well commented data, software code etc.).
- Use open source software if possible (accessible to everyone).
- Declare conflicts of interests.
- Conduct and support replications.

Some concluding remarks

- The absence of publication bias would be a marvelous step ahead in quantitative science.
- Times are changing – you will (hopefully) be rewarded for conducting accurate research.

Some references about how to make progress in quantitative research:

- Ioannidis et al. (2016): *The power of bias in economics research*
- Maniadis et al. (2016): *How to Make Experimental Economics Research More Reproducible: Lessons from Other Disciplines and a New Proposal*
- Munafo et al. (2017): *A manifesto for reproducible science*

Great blogs to follow the current debate (and beyond):

- `http://daniellakens.blogspot.de/`
- `http://andrewgelman.com/`