UNIVERSITÄT BIELEFELD
Fakultät für Wirtschaftswissenschaften

Moving to a World Beyond $p < 0.05$

Reading Course, Winter Term 2020/2021

Peter Pütz
peter.puetz@uni-bielefeld.de

Agenda

1. Organisational issues

2. Introductory presentation

Organisational issues

- Course language: English
- Meetings: Monday, 2 pm (sharp) - 3:30 pm,
  Placeholder
- Those who present should be in the zoom room 10 minutes earlier
- Syllabus:
  - Literature about a current topic in statistics
  - Students study and prepare material independently
  - Presentation and discussion of materials in class

Introduction round

- Are you familiar with the programming language `R` and if not, with any other statistics software (if so, which one)?
- Do you have any (very!) basic knowledge of Bayesian statistics?
- Have you ever heard about the following topics?
  - p-hacking
  - Publication bias
  - Replication crisis
  - Positive predictive value

Topic

Moving to a World Beyond $p < 0.05$

see also https://www.tandfonline.com/toc/utas20/73/sup1

About this course

- Requirements
  - Before each meeting, read and understand the selected paper
  - Each of you has to give a presentation on one of the papers and to be the discussant of another paper (in a group of two persons)
  - As a discussant team, you should ask at least one question per person regarding the paper to start the discussion
  - Participate in discussions
- Grading
  - Mainly based on presentation (75%), but also on contributions to discussions including the role as discussant (25%)
- Feedback
  - Always possible and desired

Presentation

- 30-45 minutes (about 15 minutes per person)
- Summarize the paper in your own words (try to find a balance between theoretical background, practical relevance and applications)
- Highlight the **main** ideas of the paper
- If appropriate, include illustrating examples and graphics by using `R` and showing some `R` code (other software is also fine)
  - You might use the data and code used in the paper
  - Feel free to write your own code and use your own data (it would be very cool to see an interesting application!)
- Send me your presentation as well as your software code and data by Saturday evening preceding your presentation (the earlier, the better)
- Grading of presentation (including responses to questions if appropriate):
  - 10% Structure
  - 10% Formalities / Readability (Formulae, Grammar, slide style etc.)
  - 40% Content accuracy
  - 40% Understandability (Didactics, applications etc.)

Aims of this course

- At the end of the course, you should...
  - ... be critical regarding common practices in empirical research (e.g., using holy *p*-value thresholds to summarize results of a study)
  - ... know about alternatives in statistical inference
  - ... have an insight into a currently discussed hot topic in empirical research
  - ... have some experience how to present research papers
  - ... still love statistics ;)

About the papers

- Some papers are more technical, some are more “philosophical”, applications are sometimes included
- Familiarity with basic statistics including statistical inference is expected
- If you understand everything of today’s presentation, you are well prepared
- If not, learning the basics does not require advanced mathematical skills
- In general: If you are lost, just ask!

Schedule

Please move to

Placeholder

and put your name into two cells, once as a presenter and once as a discussant.

Schedule

Date	Paper
02.11.20	Organisation & Introduction
	Interpreting and using p
09.11.20	Betensky – The p-Value Requires Context, Not a Threshold
16.11.20	Greenland – Valid P-Values Behave Exactly as They Should: Some Misleading Criticisms of P-Values and Their Resolution With S-Values
	Supplementing or replacing p
23.11.20	Goodman et al. – A Proposed Hybrid Effect Size Plus p-Value Criterion: Empirical Evidence Supporting its Use
30.11.20	Benjamin et al. – Three Recommendations for Improving the Use of p-Values
07.12.20	Colquhoun – The False Positive Risk: A Proposal Concerning What to Do About p-Values
14.12.20	Matthews – Moving Towards the Post $p < 0.05$ Era via the Analysis of Credibility
	Adopting more holistic approaches
21.12.20	Billheimer – Predictive Inference and Scientific Reproducibility
04.01.21	Amrhein et al. – Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don't Expect Replication
11.01.21	McShane et al. – Abandon Statistical Significance
18.01.21	Ziliak – How Large Are Your G-Values? Try Gosset's Guinnessometrics When a Little "p" Is Not Enough
25.01.21	van Dongen et. al – Multiple Perspectives on Inference for Two Simple Statistical Scenarios
	Optional
01.02.21	Wasserstein et al. – Moving to a World Beyond "$p < 0.05$"
08.02.21	to be decided

Introduction:

Statistical significance and inference – Basics, pitfalls and the current debate

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).

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

Statistical tests

- Based on your data, decide whether a hypothesis about an unknown population parameter θ is true or not. Typical examples include hypotheses of the form

$$\begin{aligned} H_0 &: \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq \theta_0 \\ H_0 &: \theta \leq \theta_0 \quad \text{vs.} \quad H_1 : \theta > \theta_0 \\ H_0 &: \theta \geq \theta_0 \quad \text{vs.} \quad H_1 : \theta < \theta_0 \end{aligned}$$

  with a fixed, prespecified value θ_0 (often: $\theta_0 = 0$, e.g. a drug has no effect).
- A good test should
  - often make the right decision.
  - tend to get better when the sample size increases.

Error types in statistical tests

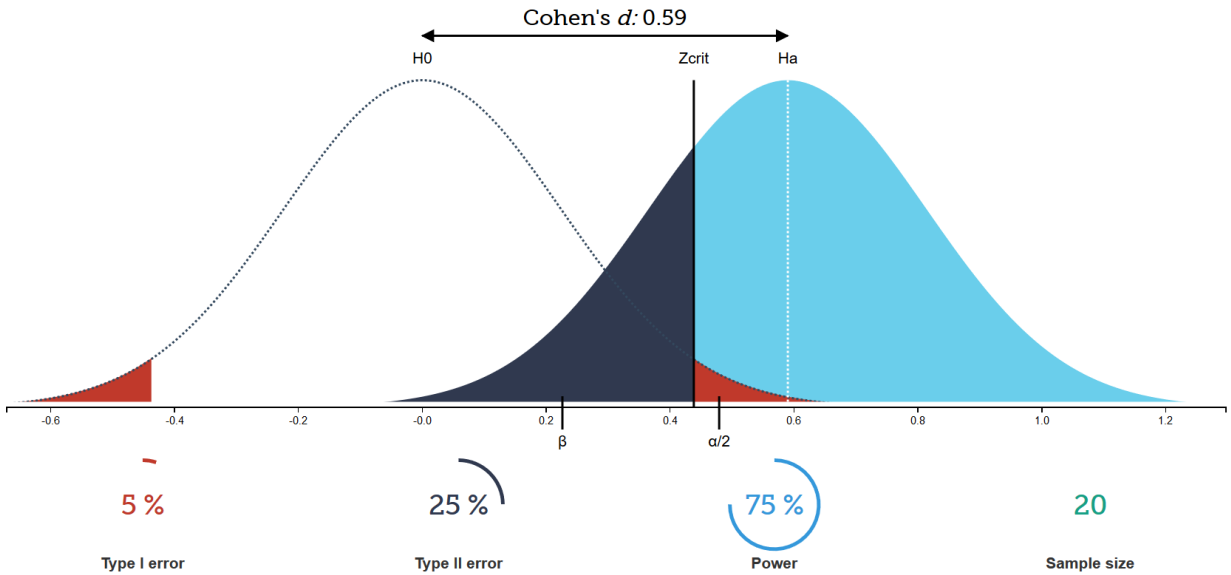
- When performing a statistical test, we can make the following decisions:

	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 often called the $\alpha = 0.05$ 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 that the power of a test moves towards one (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 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 **different** experiments that investigate the effectiveness of drugs against Covid-19.
  Before the next drug study, what result can you expect if the probability that H_0 is true (the drug is ineffective) is 70%? For all of the studies in this field, let $\alpha = 5\%$ if H_0 is true and $1 - \beta = 80\%$ for some true effect under H_1 (drug is effective):

		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 drug study is a correctly non-rejected H_0.
- But: If you find a significant effect, how likely is it that this finding is wrong (that the drug is in fact ineffective)? 5%? Calculate!

- In this example, the probability that a significant finding is false is $\frac{3.5\%}{3.5\%+24\%} \approx 12.7\%$.
- $1 - 12.7\% = 87.3\%$ is called the **Positive Predictive Value**.
- Often, studies have low power such that many more than 5% of the significant findings are false.
- Significant findings in a 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 on this topic

- Exercise positive predictive value (available in the ”Lernraum”)
- 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?*

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 α 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.

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):

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

Two ways to perform a statistical test:

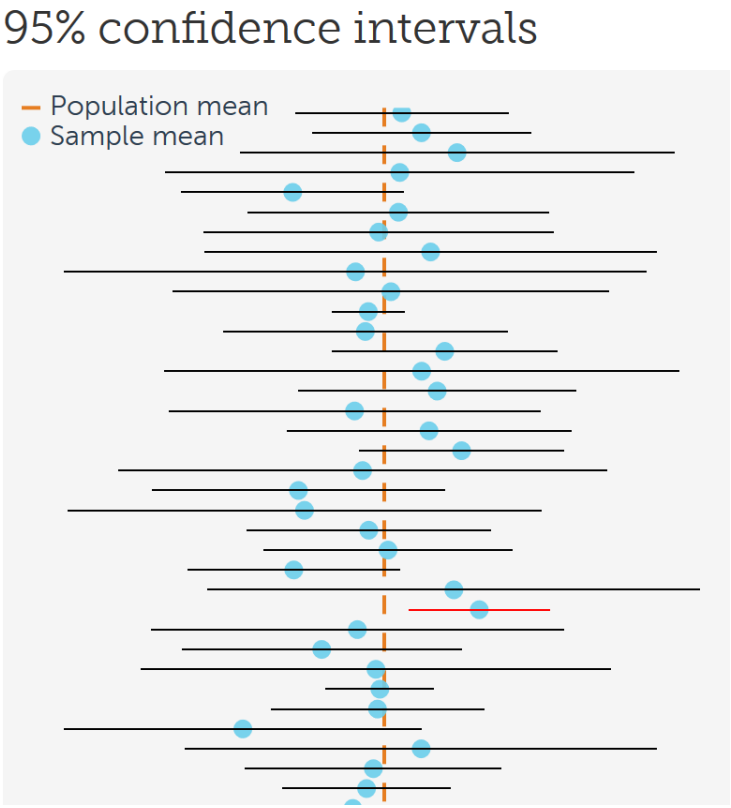
- P-value: Reject if the p-value is smaller than α.
- Confidence interval: Reject if θ_0 is not contained in the confidence interval.
- Ergo: Confidence intervals and p-value lead to the same test decisions. Differences?

Confidence intervals

- 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.

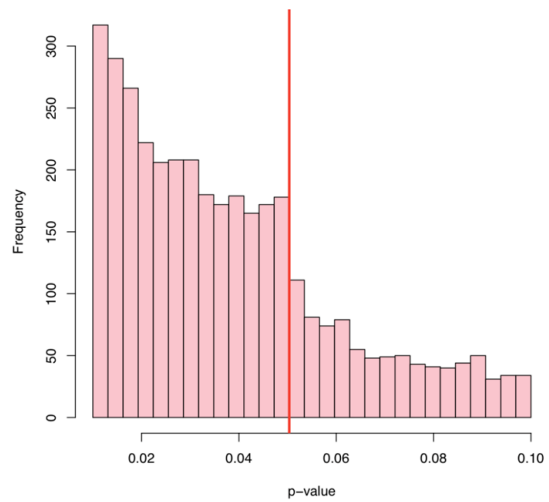
Confidence intervals: Visualization

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



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” are less likely to be accepted by journals.



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:
  - Pütz and Bruns (2020): *The (Non-)Significance of Reporting Errors in Economics: Evidence from Three Top Journals*
  - Brodeur et al. (2016): *Star Wars: The Empirics Strike Back*
  - Simmons et al. (2011): *False-Positive Psychology*

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, ...)
- 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 (frequentist statistics) 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:

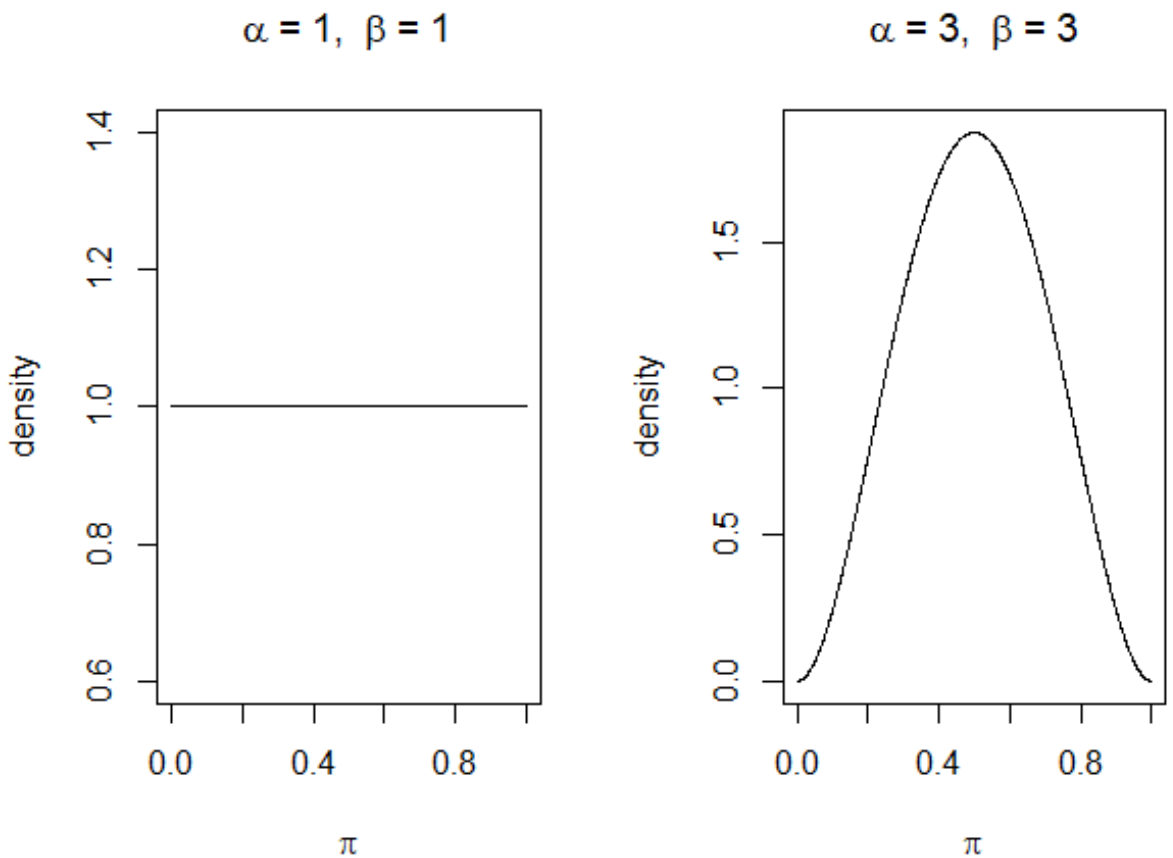
$$P(\theta|\text{data}) = \frac{P(\text{data}|\theta) * P(\theta)}{P(\text{data})}$$

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

Bayesian statistics: Coin toss example

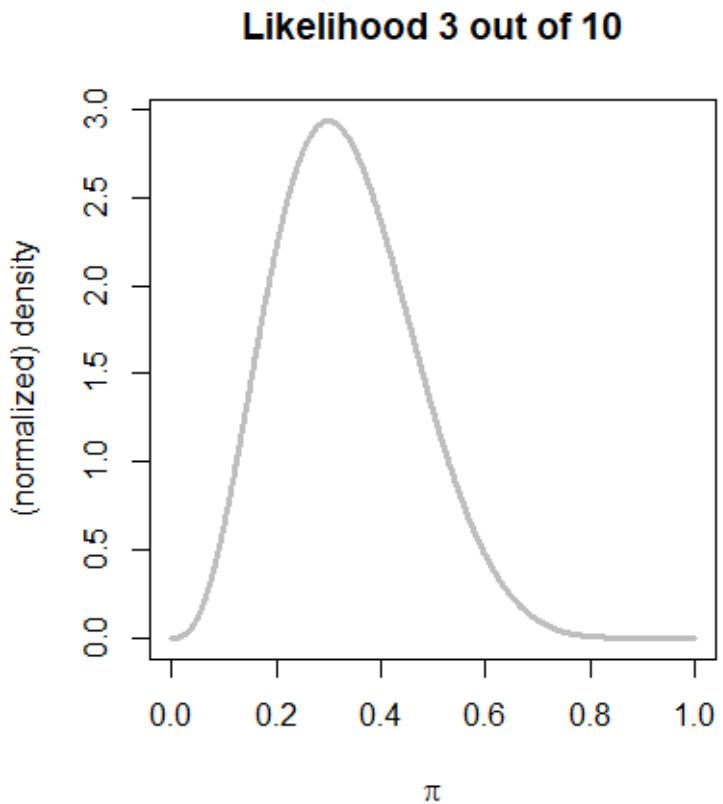
- What is the probability π for "tails up"? (Using previous notation: $\pi = \theta$)
- For the prior, a beta distribution is used (ensures values for π between 0 and 1). The beta prior is determined by two parameters α and β: $B(\alpha, \beta)$.
- Note: α and β are not error types here!
- When we expect 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

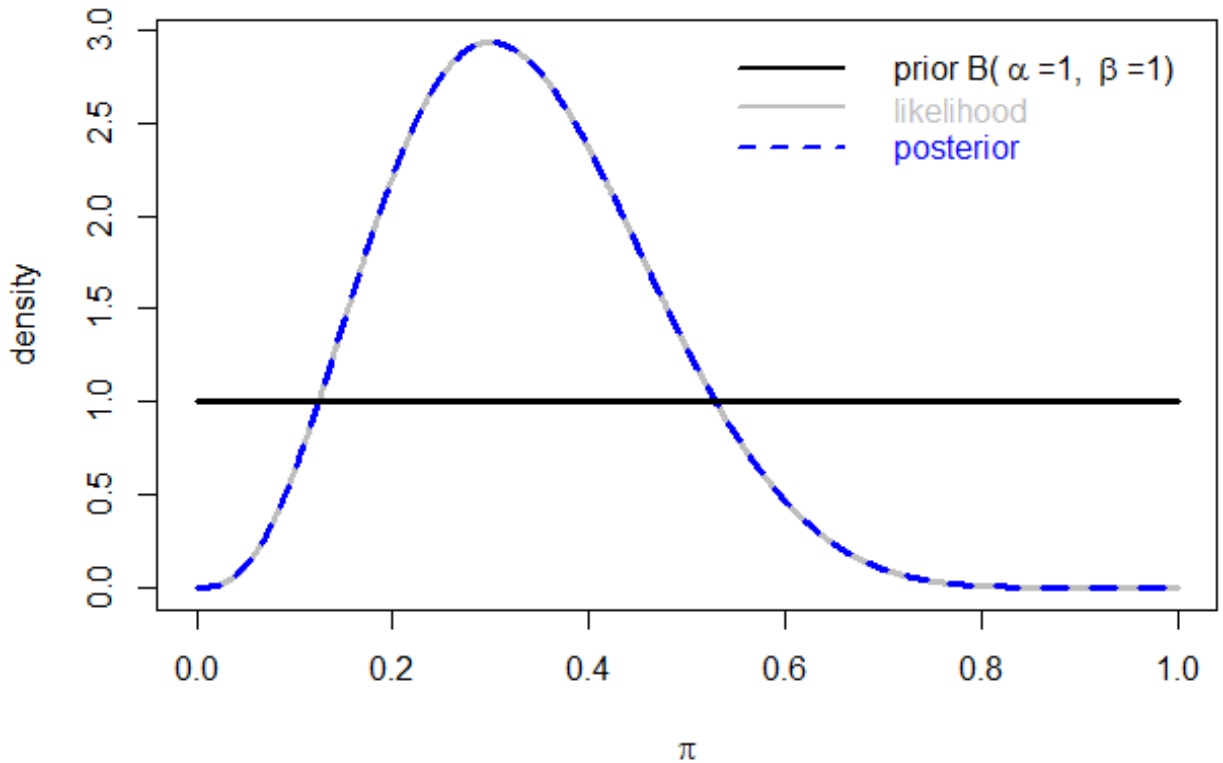


Coin toss example: Likelihood

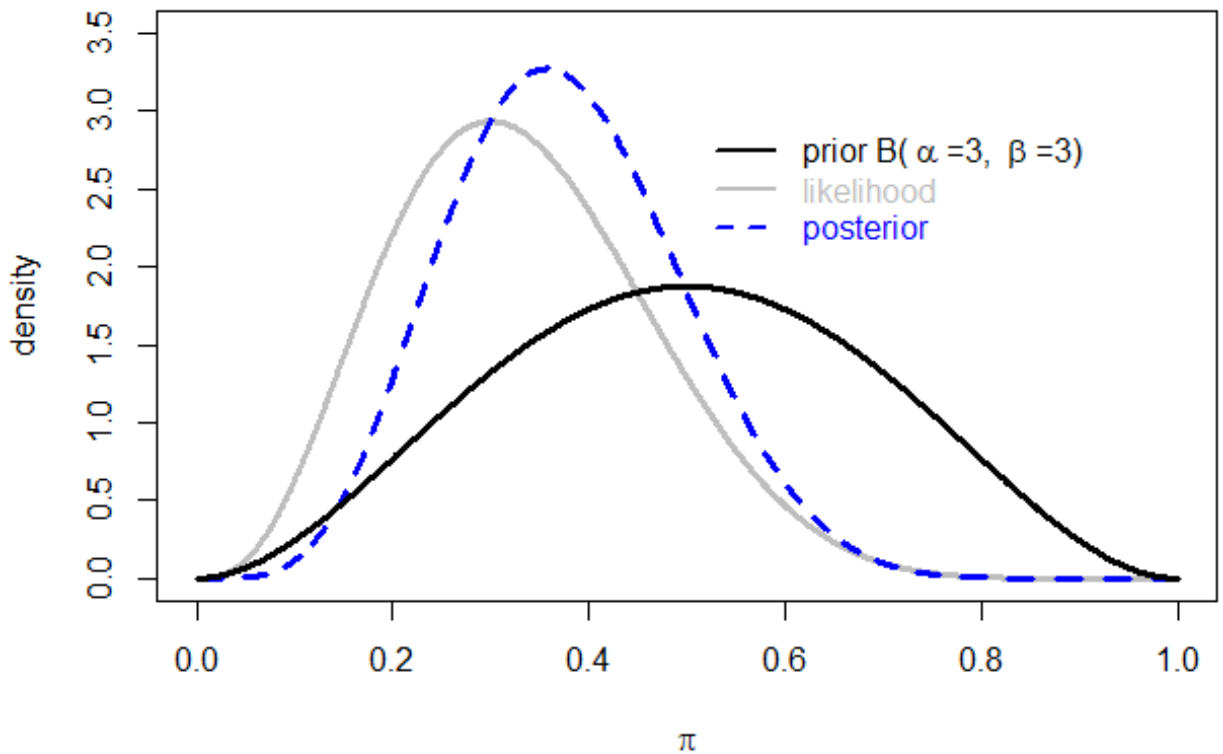
Assume: 10 toin cosses, 3 times tails.



Coin toss example: Posterior distributions



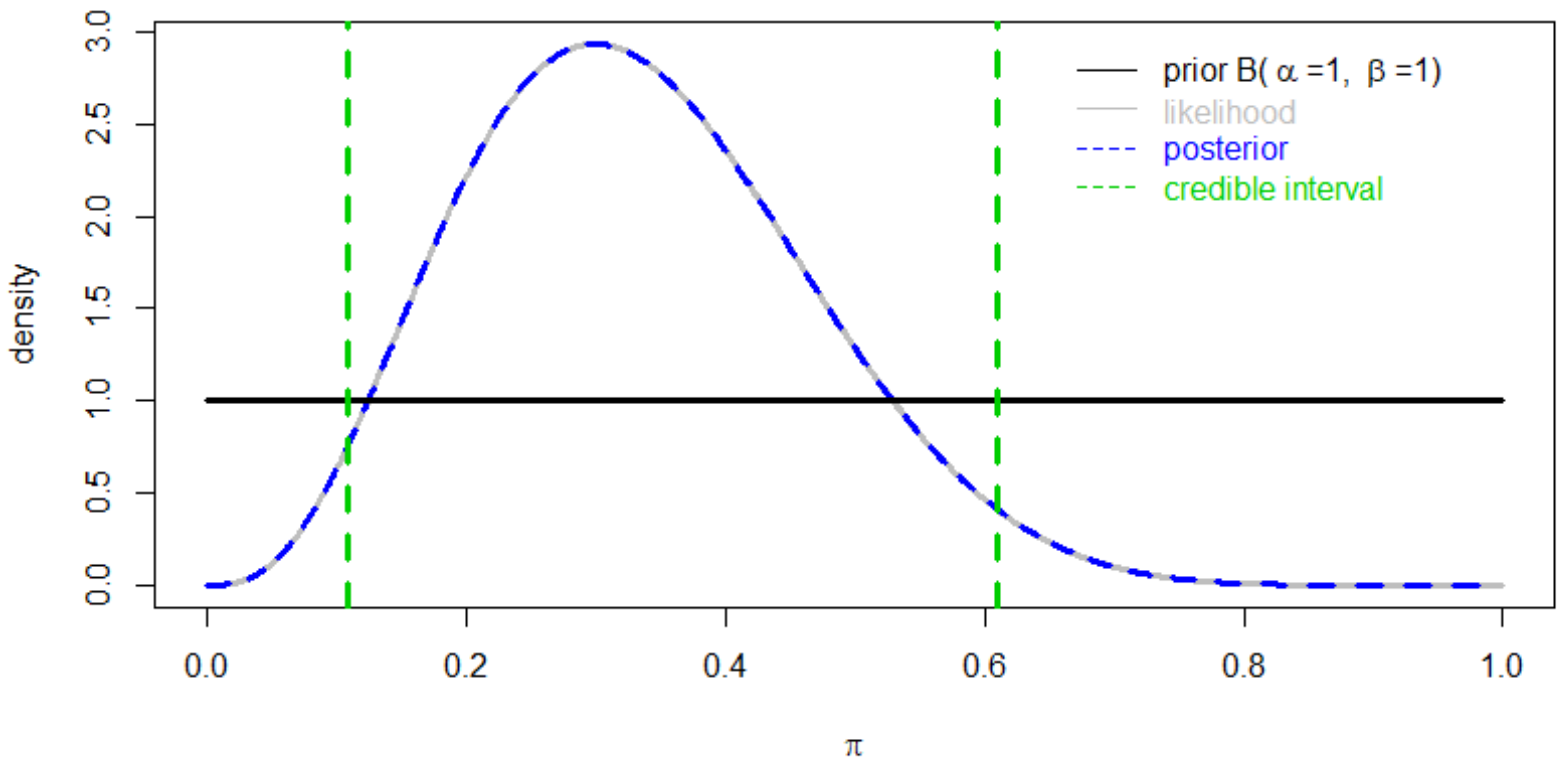
Coin toss example: Posterior distributions



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. '

A current debate on *p*-values and beyond

- The debate is actually quite old and pitfalls of significance testing have been discussed for decades.
- However, common practices have changed very little, significance tests based on *p*-values are omnipresent in empirical sciences.
- The topic received much attention by the critical ASA (American Statistical Association) statement on *p*-values in 2016 (https://amstat.tandfonline.com/doi/full/10.1080/00031305.2016.1154108)
- Plenty of papers and comments followed the ASA statement...

A current debate on *p*-values and beyond

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.'*

A current debate on *p*-values and beyond

Redefine statistical significance

We propose to change the default *P*-value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries.

Daniel J. Benjamin, James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony Greenwald, Jarrod D. Hadfield, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, Daniel J. Hruschka, Kosuke Imai, Guido Imbens, John P. A. Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael McCarthy, Don Moore, Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D. Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman and Valen E. Johnson

The lack of reproducibility of scientific studies has caused growing concern over the credibility of claims of new discoveries based on 'statistically significant' findings. There has been much progress

not address the appropriate threshold for confirmatory or contradictory replications of existing claims. We also do not advocate changes to discovery thresholds in fields that have already adopted more stringent

probabilities. By Bayes' rule, this ratio may be written as:

$$\frac{\Pr(H_1 \mid x_{\text{obs}})}{\Pr(H_{\cdot} \mid x_{\cdot})}$$

A current debate on *p*-values and beyond

A current debate on *p*-values and beyond

Justify Your Alpha: A Response to "Redefine Statistical Significance"

Daniel Lakens[abc1], Federico G. Adolfi[bc2], Casper J. Albers[ab3], Farid Anvari[d4], Matthew A. J. Apps[a5], Shlomo E. Argamon[ab6], Thom Baguley[ab7], Raymond B. Becker[ac8], Stephen D. Benning[a9], Daniel E. Bradford[a10], Erin M. Buchanan[ab11], Aaron R. Caldwell[d12], Ben van Calster[ab13], Rickard Carlsson[d14], Sau-Chin Chen[a15], Bryan Chung[a16], Lincoln J Colling[a17], Gary S. Collins[b18], Zander Crook[ab19], Emily S. Cross[d20], Sameera Daniels[ab21], Henrik Danielsson[a22], Lisa DeBruine[a23], Daniel J. Dunleavy[ab24], Brian D. Earp[ab25], Michele I. Feist[bc26], Jason D. Ferrell[ab27], James G. Field[ab28], Nicholas W. Fox[abc29], Amanda Friesen[d30], Caio Gomes[d31], Monica Gonzalez-Marquez[abc32], James A. Grange[abc33], Andrew P. Grieve[d34], Robert Guggenberger[d35], James Grist[d36], Anne-Laura van Harmelen[ab37], Fred Hasselman[bc38], Kevin D. Hochard[d39], Mark R. Hoffarth[a40], Nicholas P. Holmes[abc41], Michael Ingre[ab42], Peder M. Isager[b43], Hanna K. Isotalus[ab44], Christer Johansson[d45], Konrad Juszczyk[d46], David A. Kenny[d47], Ahmed A. Khalil[abc48], Barbara Konat[d49], Junpeng Lao[ab50], Erik Gahner Larsen[a51], Gerine M. A. Lodder[ab52], Jiří Lukavský,[d53], Christopher R. Madan[d54], David Manheim[ab55], Stephen R. Martin[abc56], Andrea E. Martin[ab57], Deborah G. Mayo[d58], Randy J. McCarthy[a59], Kevin McConway[ab60], Colin McFarland[61], Amanda Q. X. Nio[ab62], Gustav Nilsonne[ab63], Cilene Lino de Oliveira[b64], Jean-Jacques Orban de Xivry[ab65],Sam Parsons[bc66], Gerit Pfuhl[ab67], Kimberly A. Quinn[b68], John J. Sakon[a69], S. Adil Saribay[a70], Iris K. Schneider[ab71], Manojkumar Selvaraju[d72], Zsuzsika Sjoerds[b73], Samuel G. Smith[b74], Tim Smits[a75], Jeffrey R. Spies[b76], Vishnu Sreekumar[abc77], Crystal N. Steltenpohl[abc78], Neil Stenhouse[a79], Wojciech Świątkowski[a80], Miguel A. Vadillo[a81], Marcel A. L. M. Van Assen[ab82], Matt N. Williams[ab83], Samantha E. Williams[d84], Donald R. Williams[ab85], Tal Yarkoni[b86], Ignazio Ziano[d87], Rolf A. Zwaan[ab88]

A current debate on *p*-values and beyond

A current debate on *p*-values and beyond

- In 2017, the ASA held a symposium on statistical inference.
- Afterwards, the ASA encouraged scientists to contribute to a special issue on this topic...
- ...the special issue is the foundation of this certainly intriguing reading course!