Null Hypothesis Significance Testing

Disclaimer up top

- This lecture is not a stand-in for *actual* statistics training
- Strongly recommend taking a course that covers:
  - Overview and applications of the generalized linear model (GLM)
  - Multilevel (hierarchical) modeling and regression
  - Thorough discussion of power analysis and/or precision analysis
- Additional recommendations:
  - Improving your statistical inferences (Coursera)
  - Statistical Rethinking (book)

p-values: what's the fuss?

comment

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

Benjamin et al. (2017)

comment

Justify your alpha

In response to recommendations to redefine statistical significance to $P \leq 0.005$, we propose that researchers should transparently report and justify all choices they make when designing a study, including the alpha level.

Daniel Lakens, Federico G. Adolfi, Casper J. Albers, Farid Anvari, Matthew A. J. Apps, Shlomo E. Argamon, Thom Baguley, Raymond B. Becker, Stephen D. Benning, Daniel E. Bradford, Erin M. Buchanan, Aaron R. Caldwell, Ben Van Calster, Rickard Carlsson, Sau-Chin Chen, Bryan Chung, Lincoln J. Colling, Gary S. Collins, Zander Crook, Emily S. Cross, Sameera Daniels, Henrik Danielsson, Lisa DeBruine, Daniel J. Dunleavy, Brian D. Earp, Michele I. Feist, Jason D. Ferrell, James G. Field, Nicholas W. Fox, Amanda Friesen, Caio Gomes, Monica Gonzalez-Marquez, James A. Grange, Andrew P. Grieve, Robert Guggenberger, James Grist, Anne-Laura van Harmelen, Fred Hasselman, Kevin D. Hochard, Mark R. Hoffarth, Nicholas P. Holmes, Michael Ingre, Peder M. Isager, Hanna K. Isotalus, Christer Johansson, Konrad Juszczyk, David A. Kenny, Ahmed A. Khalil, Barbara Konat, Junpeng Lao, Erik Gahner Larsen, Gerine M. A. Lodder, Jiří Lukavský, Christopher R. Madan, David Manheim, Stephen R. Martin, Andrea E. Martin, Deborah G. Mayo, Randy J. McCarthy, Kevin McConway, Colin McFarland, Amanda Q. X. Nio, Gustav Nilsonne, Cilene Lino de Oliveira, Jean-Jacques Orban de Xivry, Sam Parsons, Gerit Pfuhl, Kimberly A. Quinn, John J. Sakon, S. Adil Saribay, Iris K. Schneider, Manojkumar Selvaraju, Zsuzsika Sjoerds, Samuel G. Smith, Tim Smits, Jeffrey R. Spies, Vishnu Sreekumar, Crystal N. Steltenpohl, Neil Stenhouse, Wojciech Świątkowski, Miguel A. Vadillo, Marcel A. L. M. Van Assen, Matt N. Williams, Samantha E. Williams, Donald R. Williams, Tal Yarkoni, Ignazio Ziano and Rolf A. Zwaan

Lakens et al. (2018)

THE AMERICAN STATISTICIAN
2019, VOL. 73, NO. S1, 235–245: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1527253

Abandon Statistical Significance

Blakeley B. McShane[a], David Gal[b], Andrew Gelman[c], Christian Robert[d], and Jennifer L. Tackett[e]

[a]Department of Marketing, Kellogg School of Management, Northwestern University, Evanston, IL; [b]Department of Managerial Studies, College of Business Administration, University of Illinois at Chicago, Chicago, IL; [c]Department of Statistics and Department of Political Science, Columbia University, New York, NY; [d]Centre de Recherche en Mathématiques de la Décision (CEREMADE), Université Paris-Dauphine, Paris, France; [e]Department of Psychology, Northwestern University, Evanston, IL

ABSTRACT

We discuss problems the null hypothesis significance testing (NHST) paradigm poses for replication and more broadly in the biomedical and social sciences as well as how these problems remain unresolved by proposals involving modified *p*-value thresholds, confidence intervals, and Bayes factors. We then discuss our own proposal, which is to abandon statistical significance. We recommend dropping the NHST paradigm—and the *p*-value thresholds intrinsic to it—as the default statistical paradigm for research, publication, and discovery in the biomedical and social sciences. Specifically, we propose that the *p*-value be demoted from its threshold screening role and instead, treated continuously, be considered along with currently subordinate factors (e.g., related prior evidence, plausibility of mechanism, study design and data quality, real world costs and benefits, novelty of finding, and other factors that vary by research domain) as just one among many pieces of evidence. We have no desire to "ban" *p*-values or other purely statistical measures. Rather, we believe that such measures should not be thresholded and that, thresholded or not, they should not take priority over the currently subordinate factors. We also argue that it seldom makes sense to calibrate evidence as a function of *p*-values or other purely statistical measures. We offer recommendations for how our proposal can be implemented in the scientific publication process as well as in statistical decision making more broadly.

ARTICLE HISTORY
Received October 2017
Revised September 2018

KEYWORDS
Null hypothesis significance testing; *p*-Value; Replication; Sociology of science; Statistical significance

McShane et al. (2019)

COMMENT · 20 MARCH 2019

Scientists rise up against statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

Valentin Amrhein ✉, Sander Greenland & Blake McShane

Amrhein et al. (2019)

BASIC AND APPLIED SOCIAL PSYCHOLOGY, 37:1–2, 2015

ISSN: 0197-3533 print/1532-4834 online
DOI: 10.1080/01973533.2015.1012991

Routledge
Taylor & Francis Group

Editorial

David Trafimow and Michael Marks
New Mexico State University

The *Basic and Applied Social Psychology* (BASP) 2014 Editorial emphasized that the null hypothesis significance testing procedure (NHSTP) is invalid, and thus authors would be not required to perform it (Trafimow, 2014). However, to allow authors a grace period, the Editorial stopped short of actually banning the NHSTP. The purpose of the present Editorial is to announce that the grace period is over. From now on, BASP is banning the NHSTP.

With the banning of the NHSTP from BASP, what

a strong case for rejecting it, confidence intervals do not provide a strong case for concluding that the population parameter of interest is likely to be within the stated interval. Therefore, confidence intervals also are banned from BASP.

Bayesian procedures are more interesting. The usual problem with Bayesian procedures is that they depend on some sort of Laplacian assumption to generate numbers where none exist. The Laplacian assumption is that when in a state of ignorance, the researcher should

Viewpoint

August 7, 2019

More

Is It Time to Ban the *P* Value?

Helena Chmura Kraemer, PhD[1]

» Author Affiliations | Article Information

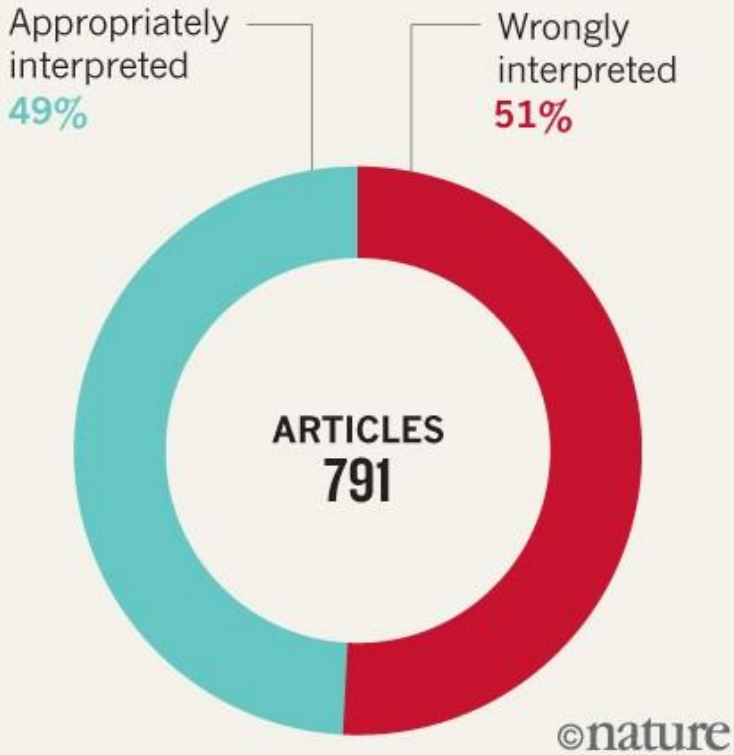
JAMA Psychiatry. 2019;76(12):1219-1220. doi:10.1001/jamapsychiatry.2019.1965

For more than 20 years, there have been rumbles about banning the *P* value,[1-3] because it is so often misused, miscomputed, and, even when used and computed correctly, misinterpreted. Consequently, findings that affect medical decision-making, policy, and research are often misled by the very research that is supposed to provide their evidence base.[4] Recently, such rumbles have increased.[5-7] Is now the time to ban the *P* value in all medical research?

Tramifow & Marks (2015); Kraemer (2019)

WRONG INTERPRETATIONS

An analysis of 791 articles across 5 journals* found that around half mistakenly assume non-significance means no effect.



*Data taken from: P. Schatz *et al. Arch. Clin. Neuropsychol.* **20**, 1053–1059 (2005); F. Fidler *et al. Conserv. Biol.* **20**, 1539–1544 (2006); R. Hoekstra *et al. Psychon. Bull. Rev.* **13**, 1033–1037 (2006); F. Bernardi *et al. Eur. Sociol. Rev.* **33**, 1–15 (2017).

Nuzzo (2014)

Misuse of *p*-values

From Wikipedia, the free encyclopedia

Misuse of *p*-values is common in scientific research and scientific education. *p*-values are often used or interpreted incorrectly; the American Statistical Association states that *p*-values can indicate how incompatible the data are with a specified statistical model.[1] From a Neyman–Pearson hypothesis testing approach to statistical inferences, the data obtained by comparing the *p*-value to a significance level will yield one of two results: either the null hypothesis is rejected (which however does not prove that the null hypothesis is *false*), or the null hypothesis *cannot* be rejected at that significance level (which however does not prove that the null hypothesis is *true*). From a Fisherian statistical testing approach to statistical inferences, a low *p*-value means *either* that the null hypothesis is true and a highly improbable event has occurred *or* that the null hypothesis is false.

Contents [hide]

“The problem is not that people use p-values poorly, it is that the vast majority of data analysis is not performed by people properly trained to perform data analysis.”

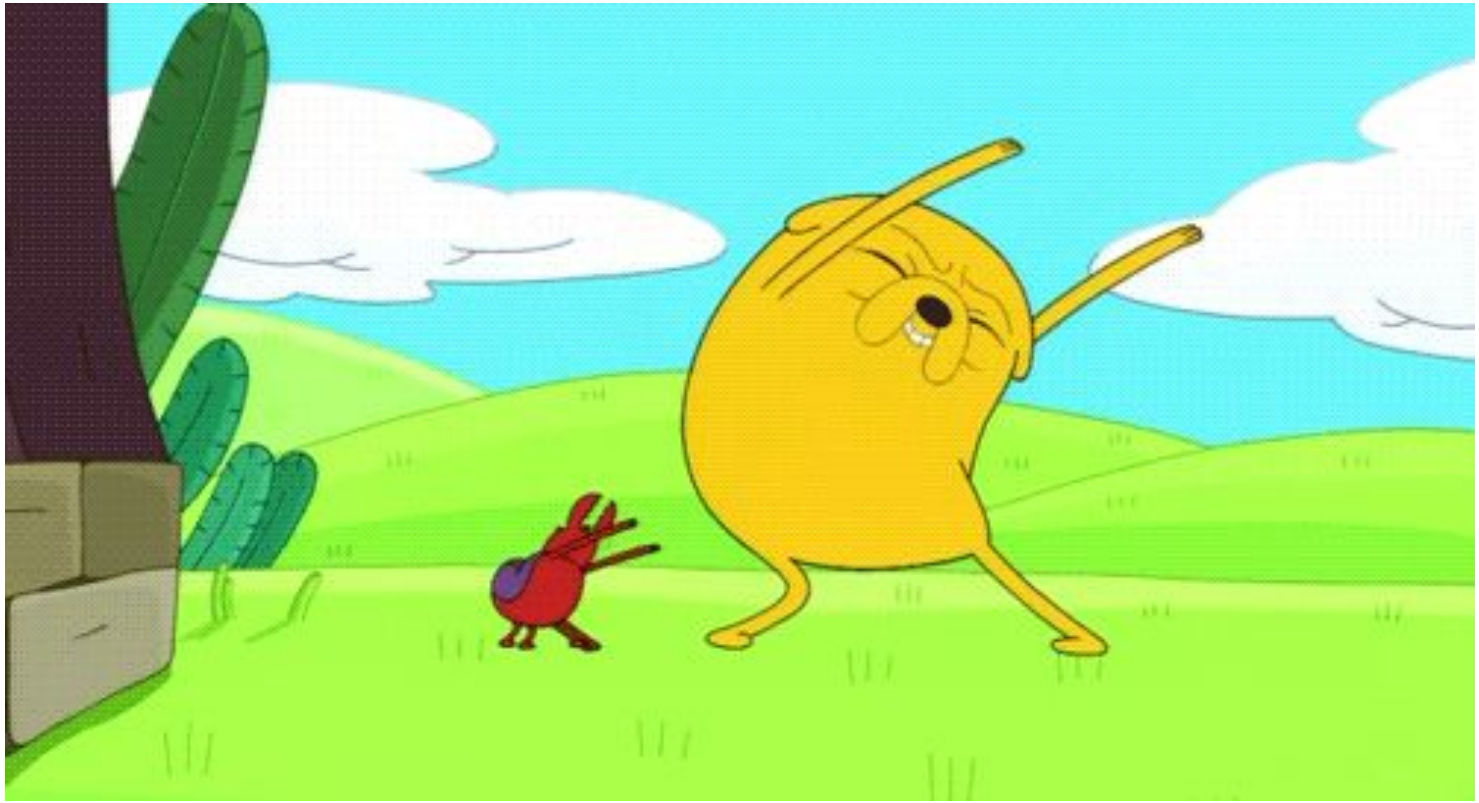
Leek (2014)

Goals for today

- Review and unpack the definition of *p*-values
- Explore some properties of *p*-values
- Build intuitions for interpreting *p*-values

The usual scientific workflow

1. Conceive a (brilliant) hypothesis
2. Design an experiment
3. Collect some data
4. Run a statistical test
5. **Interpret the p-value**



$p < 0.05$

What really is a p-value though?

Assuming the null hypothesis is true, the p-value is the probability of observing data (or a test statistic) as or more extreme than what was actually observed.

Assuming the null hypothesis is true, the p-value is the probability of observing data (or a test statistic) **as or more extreme** than what was actually observed.

Frequentist statistics: a primer

- We assume there are true states of nature (objective probabilities)
  - Events are generated by *latent* data generating processes
  - Data generating processes are inherently stochastic

Frequentist statistics: a primer

- We assume there are true states of nature (objective probabilities)
  - Events are generated by *latent* data generating processes
  - Data generating processes are inherently stochastic
- We devise and test discrete hypotheses about the state of nature
  - Hypotheses of interest are tested against an idealized *null* hypothesis
  - Each hypothesis has a corresponding model
  - Typically we compare only two (H_0, H_1) of an infinite number of hypotheses

Frequentist statistics: a primer

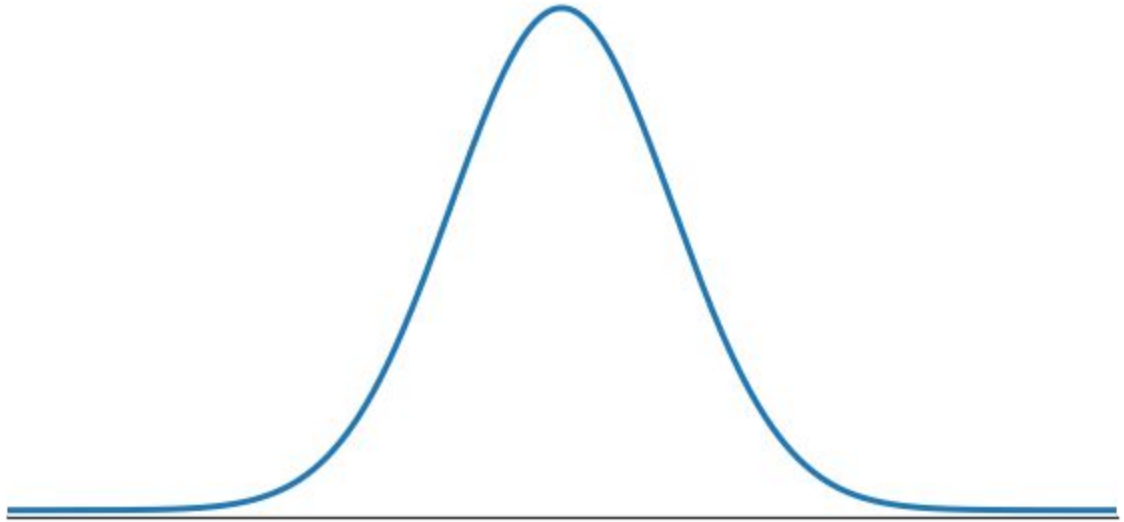
- We assume there are true states of nature (objective probabilities)
  - Events are generated by *latent* data generating processes
  - Data generating processes are inherently stochastic
- We devise and test discrete hypotheses about the state of nature
  - Hypotheses of interest are tested against an idealized *null* hypothesis
  - Each hypothesis has a corresponding model
  - Typically we compare only two (H_0, H_1) of an infinite number of hypotheses
- With near-infinite experiments, we could simply count the number of events consistent with each hypothesis

Frequentist statistics: a primer

- We assume there are true states of nature (objective probabilities)
  - Events are generated by *latent* data generating processes
  - Data generating processes are inherently stochastic
- We devise and test discrete hypotheses about the state of nature
  - Hypotheses of interest are tested against an idealized *null* hypothesis
  - Each hypothesis has a corresponding model
  - Typically we compare only two (H_0, H_1) of an infinite number of hypotheses
- With near-infinite experiments, we could simply count the number of events consistent with each hypothesis
- With a single experiment, we can use our models to approximate how likely an observation is under the *null* model

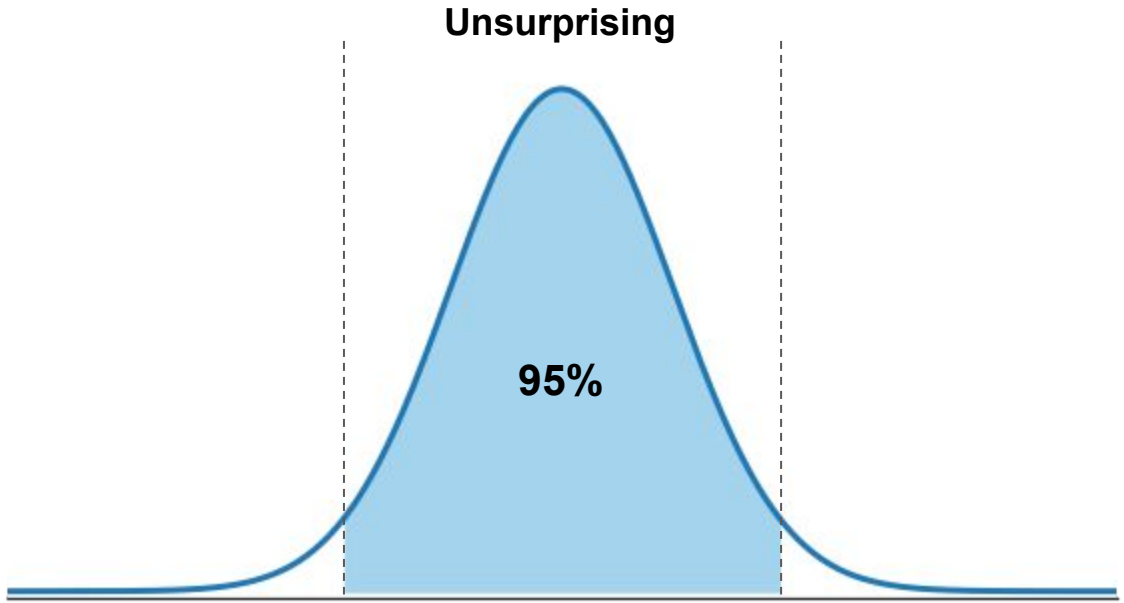
Assuming the null hypothesis is true, the p-value is the probability of observing data (or a test statistic) **as or more extreme** than what was actually observed.

The shape of a null distribution

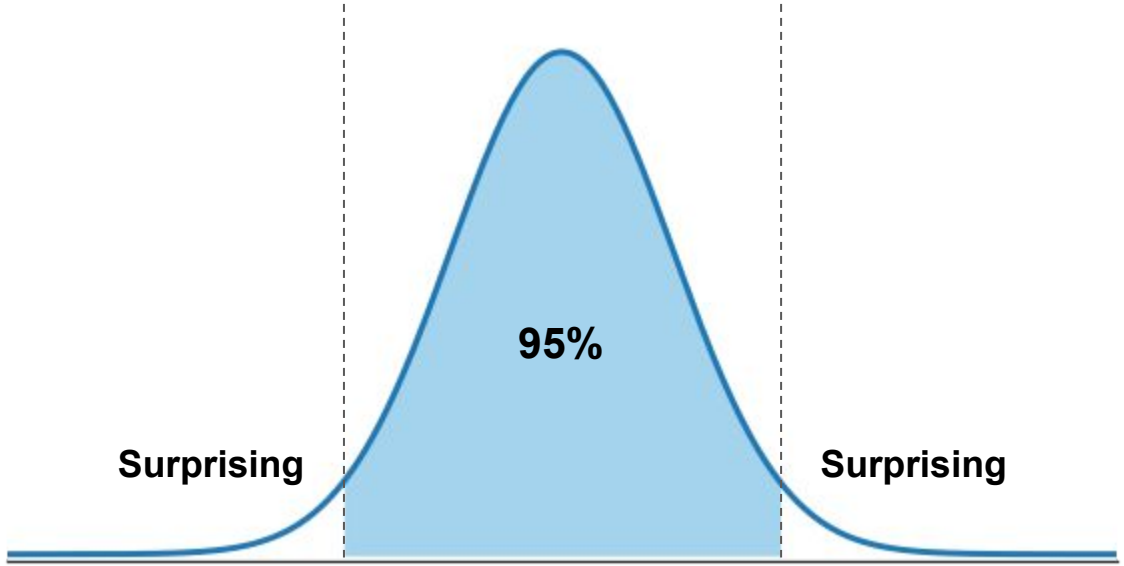


Magnitude of effect under null hypothesis

The shape of a null distribution



The shape of a null distribution

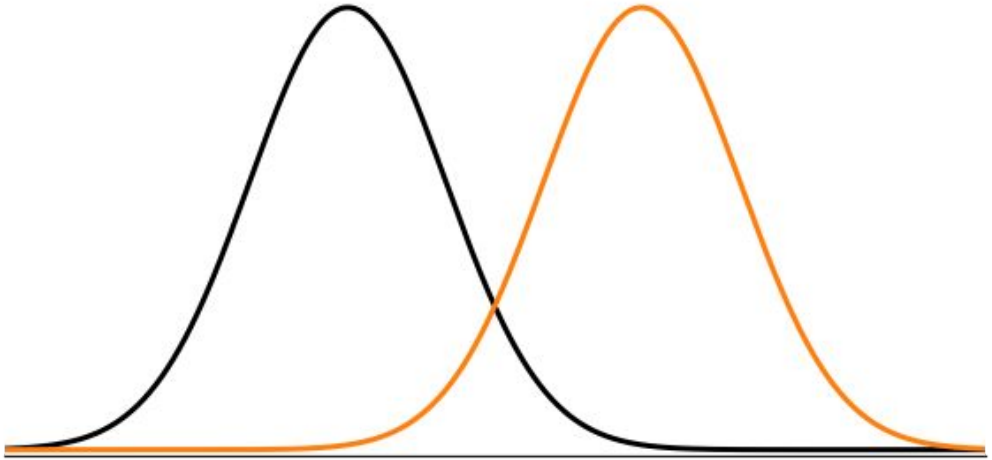


Magnitude of effect under null hypothesis

Common tests and their assumed shapes

Statistical Test	Type of Data	Assumed Distribution
T-test (one sample) T-test (paired difference) T-test (two sample)	Continuous	Normal (or T) distribution
ANOVA	Continuous	F distribution
Proportion test (one sample) Proportion test (two sample)	Proportion (ratio)	Binomial distribution
Chi-square	Count	Chi-square distribution

An example: comparing two (normal) samples



H_0: $\mu_1 = \mu_2$ H_1: $\mu_1 =/= \mu_2$

Two sample *t*-test

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{N_1 - 1} + \frac{s_2^2}{N_2 - 1}}}$$

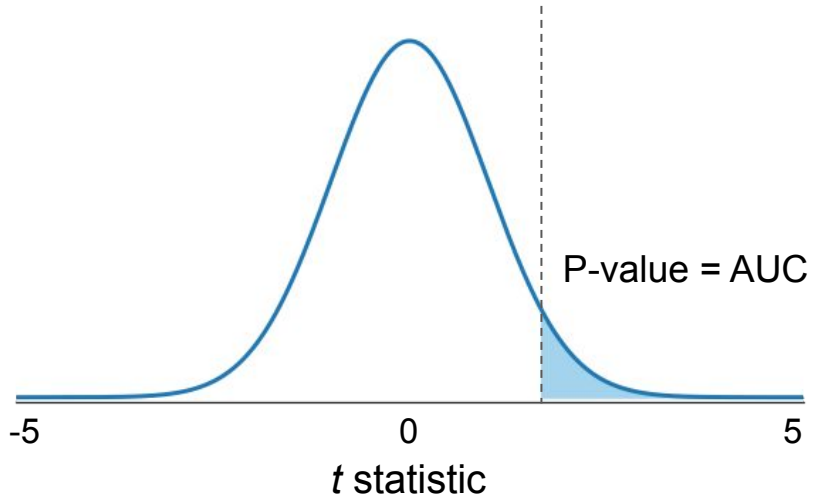
X = sample mean

s = sample standard deviation

N = sample size

Two sample *t*-test

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{N_1 - 1} + \frac{s_2^2}{N_2 - 1}}}$$



Empirical p-values

Permutation test: test of statistical significance in which the *distribution* of the test statistic under the null hypothesis is obtained by calculating all possible values of the test statistic under many possible rearrangements of the observed data points.

In our example, under the null hypothesis, group membership (1 or 2) is arbitrary. To compute distribution of test statistics under null, shuffle group membership (while maintaining sample sizes) and re-compute *t*-statistic.

Exercise 1

p-values are random variables

- A p-value reflects the degree of surprisingness under the null of a *single* experiment
- Because latent data generating processes are stochastic, we should never expect the p-values for successive experiments to be identical (although they may be similar in magnitude)

Exercise 2

Discussion Questions

What can you conclude from a single test of statistical significance? Do p-values reflect theories or observations?

What can you say about a hypothesis when a statistical test is significant?

What can you say about a hypothesis when a statistical test is not significant?

How p-values are distributed

- For a data generating process, we can form expectations for how p-values should be distributed
- These expectations motivate Frequentist statistics as a framework for **error control**

	Fail to reject null	**Reject null**
H_0 true	Correct conclusion	Type 1 error
H_1 true	Type 2 error	Correct conclusion

p-values under the null hypothesis

- The default threshold for statistical significance ($\alpha = 0.05$) is motivated by the distribution of p-values under the null distribution
- The usual justification is that setting $\alpha = 0.05$ ensures that we make a Type I error only 5% of the time.
- Where does this come from?

Exercise 3

p-values under the alternative hypothesis

- Statistical power is how often we expect to reject the null hypothesis, given that the alternative hypothesis is true (i.e. avoid Type II error)
- Statistical power is influenced by many features

p-values under the alternative hypothesis

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{N_1 - 1} + \frac{s_2^2}{N_2 - 1}}}$$

Factors influencing statistical power:

- **Effect strength:** larger experimental effects is proportional to test statistic size
- **Measurement noise:** smaller measurement noise is inversely proportional to test statistic size
- **Sample size:** larger sample size is proportional to test statistic size

Exercise 4

Discussion Questions

Is the p-value a measure of effect size?

When statistical power is very high and you observe $p = 0.04$, is the result more consistent with the null or alternative hypothesis?