Hypothesis Testing

CCBB Introduction to Biostats

September 18, 2015

Why test? Why be critical of tests?

Separate fiction from fact before it ends up in a psychology journal?

Journal of Personality and Social Psychology

0022-3514/11/$12.00 DOI: 10.1037/a0021524

Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect

Daryl J. Bem
Cornell University

The term *psi* denotes anomalous processes of information or energy transfer that are currently unexplained in terms of known physical or biological mechanisms. Two variants of psi are *precognition* (conscious cognitive awareness) and *premonition* (affective apprehension) of a future event that could not otherwise be anticipated through any known inferential process. Precognition and premonition are themselves special cases of a more general phenomenon: the anomalous retroactive influence of some future event on an individual's current responses, whether those responses are conscious or nonconscious, cognitive or affective. This article reports 9 experiments, involving more than 1,000 participants, that test for retroactive influence by "time-reversing" well-established psychological effects so that the individual's responses are obtained before the putatively causal stimulus events occur. Data are presented for 4 time-reversed effects: precognitive approach to erotic stimuli and precognitive avoidance of negative stimuli; retroactive priming; retroactive habituation; and retroactive facilitation of recall. The mean effect size (*d*) in psi performance across all 9 experiments was 0.22, and all but one of the experiments yielded statistically significant results. The individual-difference variable of stimulus seeking, a component of extraversion, was significantly correlated with psi performance in 5 of the experiments, with participants who scored above the midpoint on a scale of stimulus seeking achieving a mean effect size of 0.43. Skepticism about psi, issues of replication, and theories of psi are also discussed.

Keywords: psi, parapsychology, ESP, precognition, retrocausation

Why test? Why be critical of tests?

Separate fiction from fact before it ends up in a psychology journal?

Journal of Personality and Social Psychology
© 2011 American Psychological Association
0022-3514/11/$12.00 DOI: 10.1037/a0021524

Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect

Research Article

ASSOCIATION FOR PSYCHOLOGICAL SCIENCE
Psychological Science
XX(X) 1–10
© The Author(s) 2013
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797612466416
pss.sagepub.com
SAGE

The Fluctuating Female Vote: Politics, Religion, and the Ovulatory Cycle

Kristina M. Durante[1], Ashley Rae[1], and Vladas Griskevicius[2]
[1]College of Business, University of Texas, San Antonio, and [2]Carlson School of Management, University of Minnesota

Abstract
Each month, many women experience an ovulatory cycle that regulates fertility. Although research has found that this cycle influences women's mating preferences, we proposed that it might also change women's political and religious

Null hypothesis testing: an outline

The basic idea:

'Validate' hypothesis by rejecting a contrary (often simpler) hypothesis.

Null hypothesis testing: an outline

The basic idea:

'Validate' hypothesis by rejecting a contrary (often simpler) hypothesis.

The four steps to null hypothesis testing:

1. Determine statistical hypothesis (model)

Null hypothesis testing: an outline

The basic idea:

'Validate' hypothesis by rejecting a contrary (often simpler) hypothesis.

The four steps to null hypothesis testing:

1. Determine statistical hypothesis (model)
2. Determine statistical null hypothesis (null model)

Null hypothesis testing: an outline

The basic idea:

'Validate' hypothesis by rejecting a contrary (often simpler) hypothesis.

The four steps to null hypothesis testing:

1. Determine statistical hypothesis (model)
2. Determine statistical null hypothesis (null model)
3. Evaluate evidence for null hypothesis

Null hypothesis testing: an outline

The basic idea:

'Validate' hypothesis by rejecting a contrary (often simpler) hypothesis.

The four steps to null hypothesis testing:

1. Determine statistical hypothesis (model)
2. Determine statistical null hypothesis (null model)
3. Evaluate evidence for null hypothesis
4. Evidence against null hypothesis?

   Yes: reject null
   No: *fail to reject* null

Definitions

Example of hypothesis:
the application of neonicotinoid pesticide influences bumblebee survival.

A **model** is a mathematical expression of the hypothesized mechanism generating the data.

Example of model:

- (Pesticide) treatment groups have differ in mean survival time.
- Survival time is distributed as a normal random variable.

$$\mathbb{E}[Y_{\text{pesticide}}] \neq \mathbb{E}[Y_{\text{control}}]$$

Definitions

Example of hypothesis:
the application of neonicotinoid pesticide influences bumblebee survival.

A **null hypothesis** is a (simpler) model lacking the hypothesized mechanism.

Example of null hypothesis:

There is no effect of pesticide application on bumblebee survival.

Definitions

Example of hypothesis:
the application of neonicotinoid pesticide influences bumblebee survival.

A **null model** is the mathematical expression of the null hypothesis.

Example of null model:

- Treatment groups have the same mean survival time.
- Survival time is distributed as a normal random variable.

$$\mathbb{E}[Y_{\text{pesticide}}] = \mathbb{E}[Y_{\text{control}}]$$

Null hypotheses can be stupid (*a priori* false)

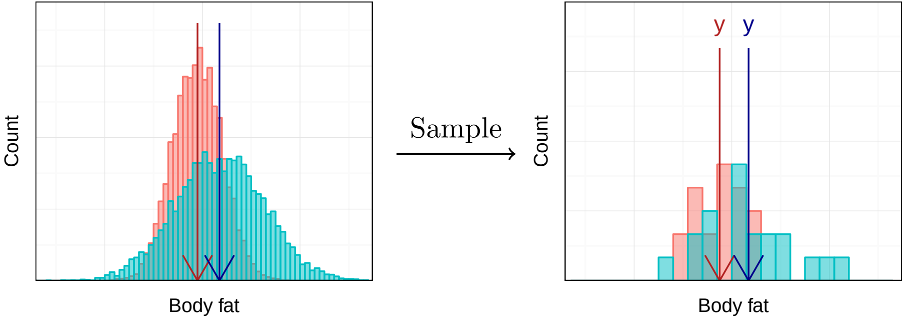
Silly hypothesis

Hyp.: The average body fat of ducks differs between years.
Null: The average body fat of ducks does not differ between years.

Null hypotheses can be stupid (*a priori* false)

Silly hypothesis

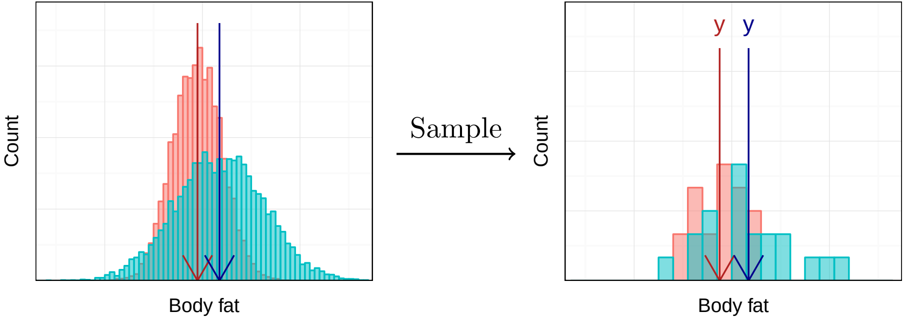
Hyp.: The average body fat of ducks differs between years.
Null: The average body fat of ducks does not differ between years.



Null hypotheses can be stupid (*a priori* false)

Silly hypothesis

Hyp.: The average body fat of ducks differs between years.
Null: The average body fat of ducks does not differ between years.



We already know that the years differ!
There is *no way* they could not, even if by a small amount.

Null hypotheses can be stupid (*a priori* false)

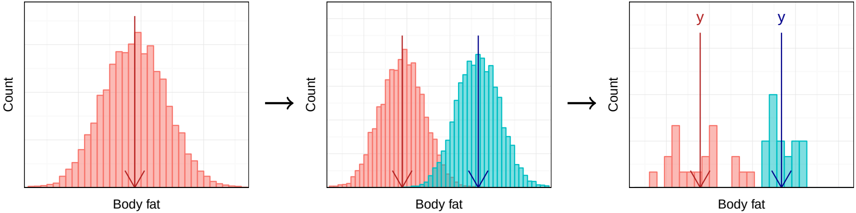
Reasonable hypothesis

Hyp.: The average body fat of ducks differs after feeding treatment.
Null: The average body fat of ducks is the same after treatment.

Null hypotheses can be stupid (*a priori* false)

Reasonable hypothesis

Hyp.: The average body fat of ducks differs after feeding treatment.
Null: The average body fat of ducks is the same after treatment.

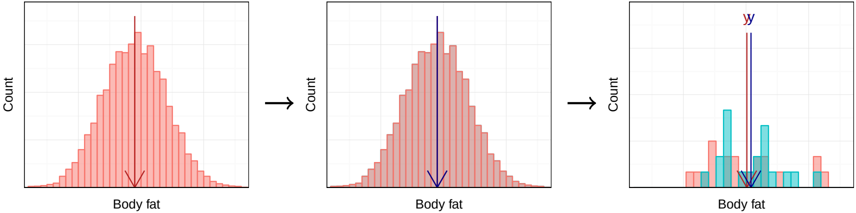


Case 1: the treatment alters the population mean.

Null hypotheses can be stupid (*a priori* false)

Reasonable hypothesis

Hyp.: The average body fat of ducks differs after feeding treatment.
Null: The average body fat of ducks is the same after treatment.

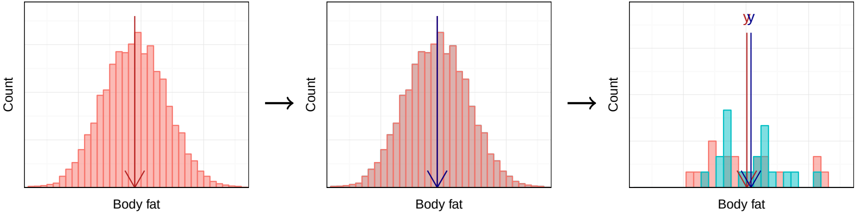


Case 2: the treatment does not alter the population mean.

Null hypotheses can be stupid (*a priori* false)

Reasonable hypothesis

Hyp.: The average body fat of ducks differs after feeding treatment.
Null: The average body fat of ducks is the same after treatment.



Case 2: the treatment does not alter the population mean.

Obvious–but the point is we do not know *a priori* which case is true.

How to evaluate *evidence* for null hypothesis?

The typical approach:

1. Pick a test statistic
    - (i.e. the parameter of interest)

Test statistics

The **test statistic** is a standardized measure of **effect size**, associated with a parameter of interest in the model. It is *calculated from data.*

Two main types of effect size:

Test statistics

The **test statistic** is a standardized measure of **effect size**, associated with a parameter of interest in the model. It is *calculated from data.*

Two main types of effect size:

1. The raw parameter estimate
    - (i.e. the difference in means between two treatment groups).

i.e. **T-statistic**

$$T = \frac{\text{difference between group means}}{\text{standard error of difference}}$$

Test statistics

The **test statistic** is a standardized measure of **effect size**, associated with a parameter of interest in the model. It is *calculated from data.*

Two main types of effect size:

1. The raw parameter estimate
   - (i.e. the difference in means between two treatment groups).
2. A measure of improvement in explanatory power associated with adding a parameter
   - (i.e. the increase in the likelihood, associated with adding a parameter)

i.e. χ^2**-statistic**

$$\chi^2 = \frac{\text{likelihood of model with effect}}{\text{likelihood of null model}}$$

How to evaluate *evidence* for null hypothesis?

The typical approach:

1. Pick a test statistic
   - (i.e. the parameter of interest)
2. Calculate test statistic with observed data
   - (i.e. fit the data to a model)

How to evaluate *evidence* for null hypothesis?

The typical approach:

1. Pick a test statistic
   - (i.e. the parameter of interest)
2. Calculate test statistic with observed data
   - (i.e. fit the data to a model)
3. Calculate (or approximate) the distribution of the test statistic **if the null model were true**
   - (this is called the *null distribution* of the test statistic.)

Null distributions

The **null distribution** is the sampling distribution of the test statistic, if the null model is true.

Example by Shiny!

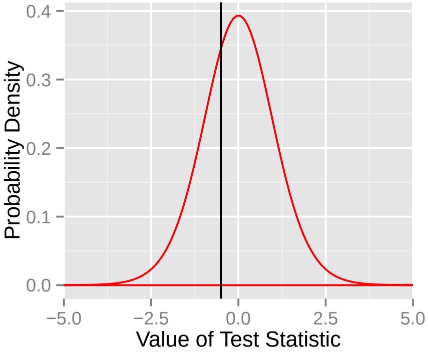
(code at class GitHub repo)

How to evaluate *evidence* for null hypothesis?

The typical approach:

1. Pick a test statistic
   - (i.e. the parameter of interest)
2. Calculate test statistic with observed data
   - (i.e. fit the data to a model)
3. Calculate (or approximate) the distribution of the test statistic **if the null model were true**
   - (this is called the *null distribution* of the test statistic.)
4. Calculate the probability of finding a test statistic of greater value than the observed test statistic, under the null distribution.
   - this is called the *p-value* of the test statistic.

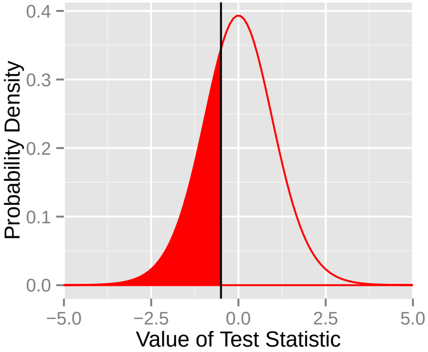
Cumulative Distribution Function



The **probability distribution function** measures the height of the curve at specified point.

$$\Pr(X = -0.5) = 0.34 = \mathrm{PDF}(-0.5)$$

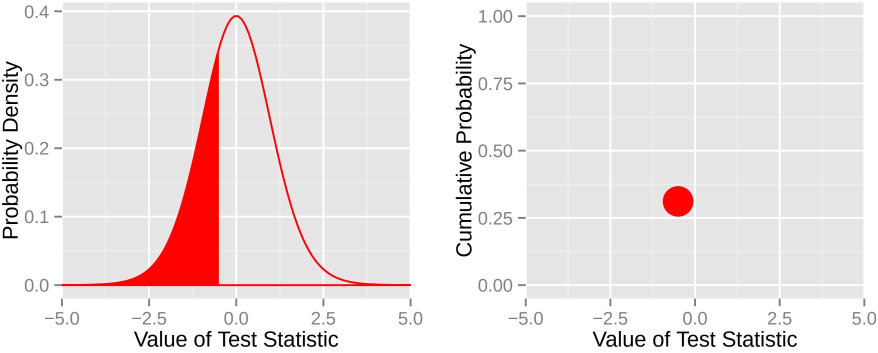
Cumulative Distribution Function



The **cumulative distribution function** measures the area under the curve *up to* specified point.

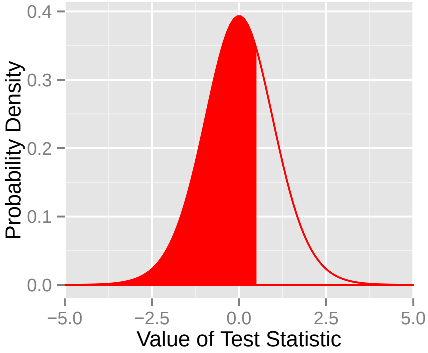
$$\Pr(X < -0.5) = 0.31 = \text{CDF}(-0.5)$$

Cumulative Distribution Function



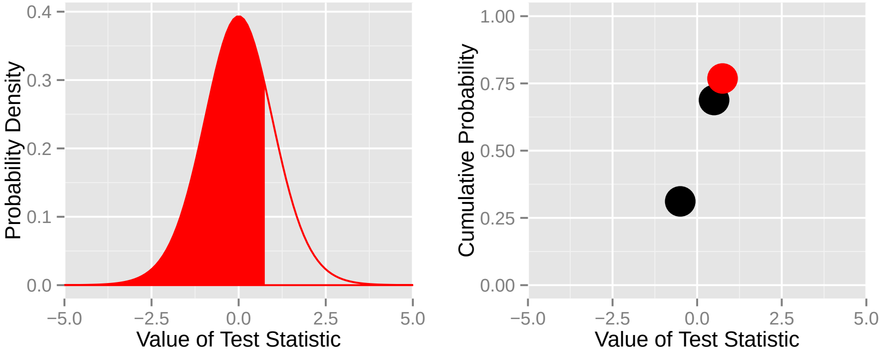
The CDF gives the probability that a random variable is *less than* a value.

Cumulative Distribution Function



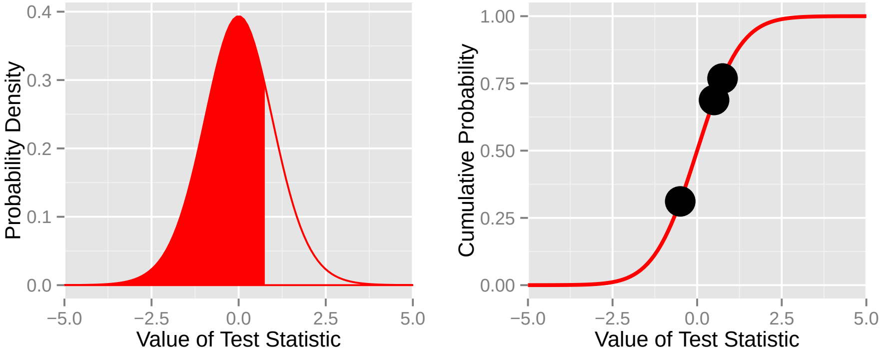
The CDF gives the probability that a random variable is *less than* a value.

Cumulative Distribution Function



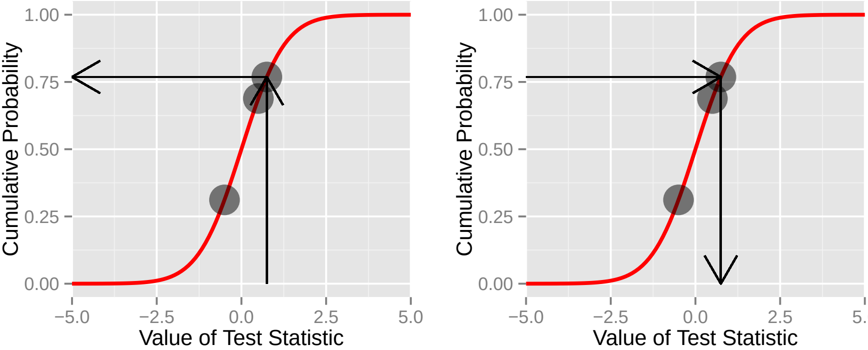
The CDF gives the probability that a random variable is *less than* a value.

Cumulative Distribution Function



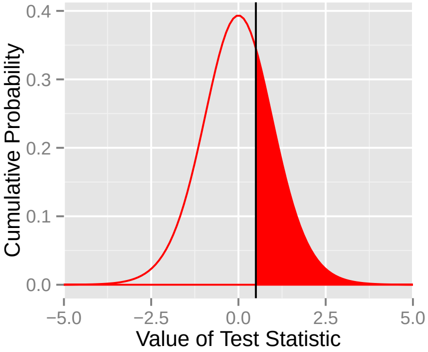
The CDF gives the probability that a random variable is *less than* a value.

Quantile Function



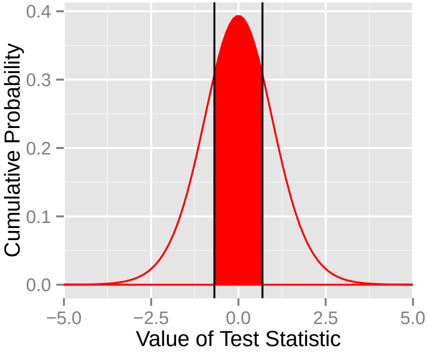
The Quantile function (right) is the inverse of the CDF (left).

Transformations of the CDF



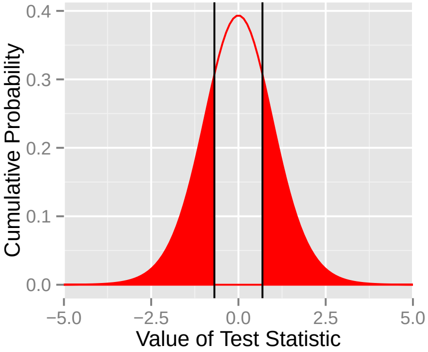
$1 - \mathrm{CDF}(x)$ gives the area of the curve beyond the value x
(The probability of a greater value than x.)

Transformations of the CDF



$\mathrm{CDF}(x+z) - \mathrm{CDF}(x)$ gives the area of the curve between x and $x+z$ (The probability of a value between x and $x+z$.)

Transformations of the CDF



$1 - [\text{CDF}(x+z) - \text{CDF}(x)]$ ...
(The probability of a value lower than x or greater than $x+z$.)

The p-value and the CDF

The p-value

A **p-value** gives the probability of getting a *more extreme value* then our observed test statistic, if the null model were true.

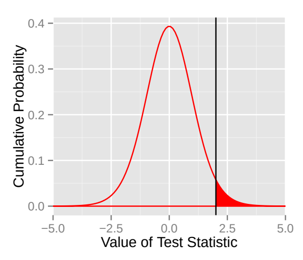
The p-value and the CDF

The p-value

A **p-value** gives the probability of getting a *more extreme value* then our observed test statistic, if the null model were true.

Case 1: one-tailed

More extreme means greater than the actual value of the test statistic. Use $1 - \mathrm{CDF}(T)$.



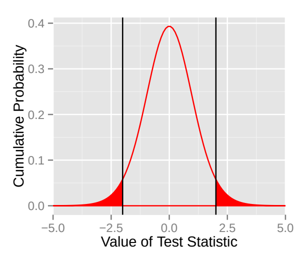
The p-value and the CDF

The p-value

A **p-value** gives the probability of getting a *more extreme value* then our observed test statistic, if the null model were true.

Case 2: two-tailed

More extreme means greater than the absolute value of the test statistic. Use $1 - [\text{CDF}(|T|) - \text{CDF}(-|T|)]$.



The p-value and the CDF

The p-value

A **p-value** gives the probability of getting a *more extreme value* then our observed test statistic, if the null model were true.

You don't need an equation for the CDF

You can simulate the sampling distribution.

The p-value

What the p-value is

The probability of observing an effect size of greater magnitude than the effect size in our data, **if the null hypothesis is true**.

The p-value

What the p-value is

The probability of observing an effect size of greater magnitude than the effect size in our data, **if the null hypothesis is true**.

What the p-value isn't

The p-value

What the p-value is

The probability of observing an effect size of greater magnitude than the effect size in our data, **if the null hypothesis is true**.

What the p-value isn't

- The probability that the null hypothesis is true.

The p-value

What the p-value is

The probability of observing an effect size of greater magnitude than the effect size in our data, **if the null hypothesis is true**.

What the p-value isn't

- The probability that the null hypothesis is true.
- The probability of our data, if the the null hypothesis is true.

The p-value

What the p-value is

The probability of observing an effect size of greater magnitude than the effect size in our data, **if the null hypothesis is true**.

What the p-value isn't

- The probability that the null hypothesis is true.
- The probability of our data, if the the null hypothesis is true.
- One minus the probability of our data, if the hypothesis is true.

The p-value

What the p-value is

The probability of observing an effect size of greater magnitude than the effect size in our data, **if the null hypothesis is true**.

What the p-value isn't

- The probability that the null hypothesis is true.
- The probability of our data, if the the null hypothesis is true.
- One minus the probability of our data, if the hypothesis is true.
- One minus the probability that the hypothesis is true.

The p-value

What the p-value is

The probability of observing an effect size of greater magnitude than the effect size in our data, **if the null hypothesis is true**.

What the p-value isn't

- The probability that the null hypothesis is true.
- The probability of our data, if the the null hypothesis is true.
- One minus the probability of our data, if the hypothesis is true.
- One minus the probability that the hypothesis is true.

The p-value is a noisy measure of support for the null hypothesis.

History

R.A. Fisher

- Invented p-value in experimental setting
- ‘Continuous measure of evidence’

History

R.A. Fisher

- Invented p-value in experimental setting
- 'Continuous measure of evidence'

- p-value as decision making tool
- Confidence threshold (ie. $P < 0.05$)

J. Neyman

The confidence threshold and type errors

		Null Hypothesis Rejected?	
		Yes	No
Null Hypothesis True?	Yes	Type I error	True negative
	No	True rejection	Type II error

The confidence threshold and type errors

		Null Hypothesis Rejected? Yes	No
Null Hypothesis True?	Yes	Type I error	True negative
	No	True rejection	Type II error

Multiple comparisons inflate Type I error:
these are multiple hypothesis tests on the same sample.

The confidence interval

Takehome message:

The **confidence interval** provides a plausible range for the true value of the parameter.

The confidence interval

Takehome message:

The **confidence interval** provides a plausible range for the true value of the parameter.

The confidence interval is a function of the sample data, the sample is random, so the CI is random too.

The confidence interval

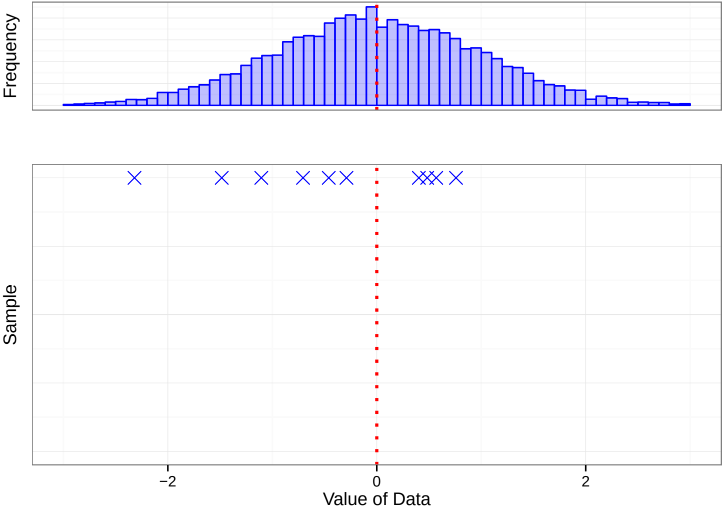
Takehome message:

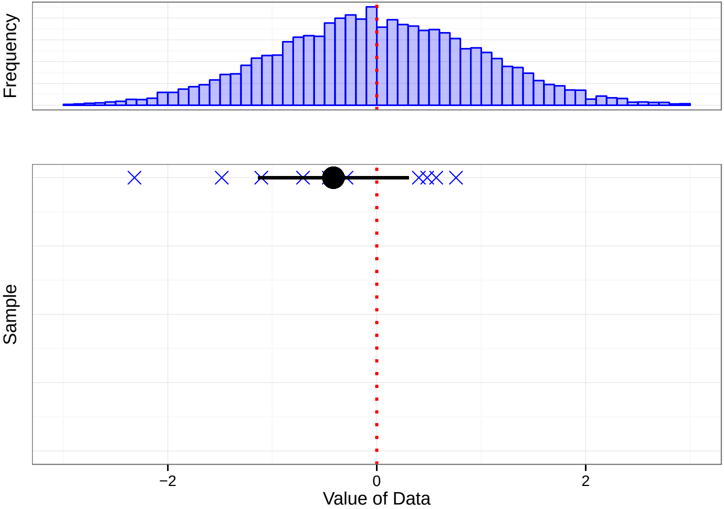
The **confidence interval** provides a plausible range for the true value of the parameter.

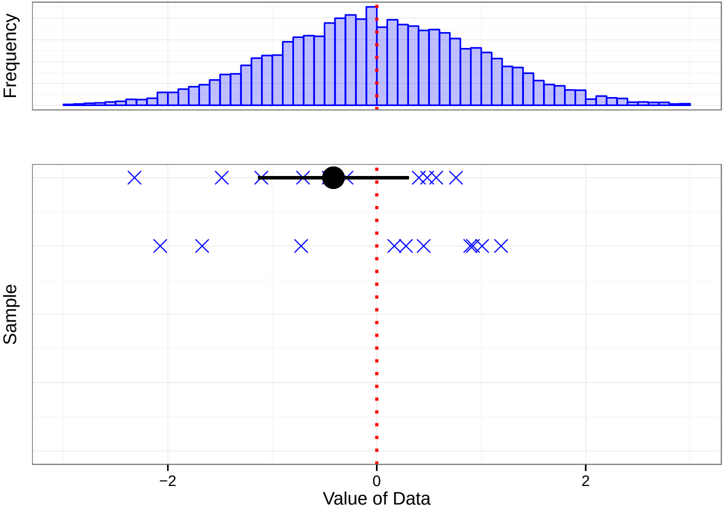
The confidence interval is a function of the sample data, the sample is random, so the CI is random too.

The definition of a $P\%$ confidence interval:

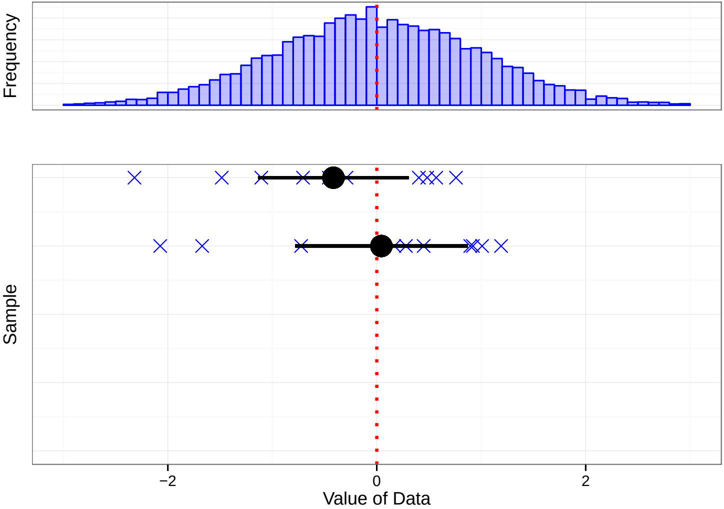
$P\%$ of these confidence intervals calculated from random samples, will contain the true value of the parameter.

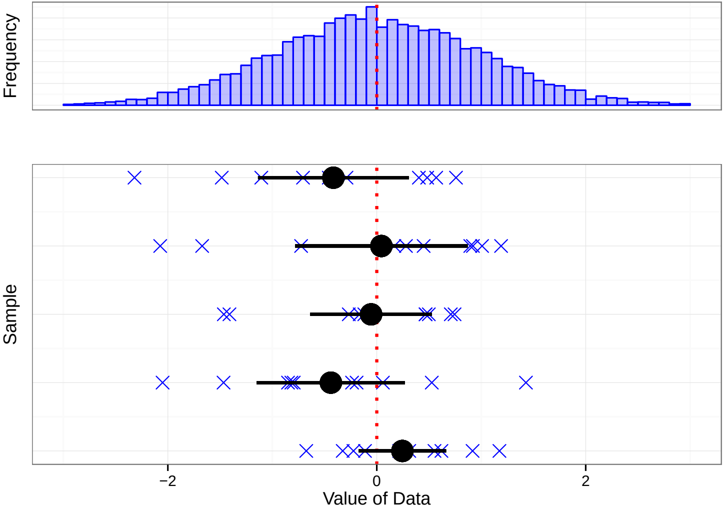

Frequency
Sample
−2
0
2
Value of Data

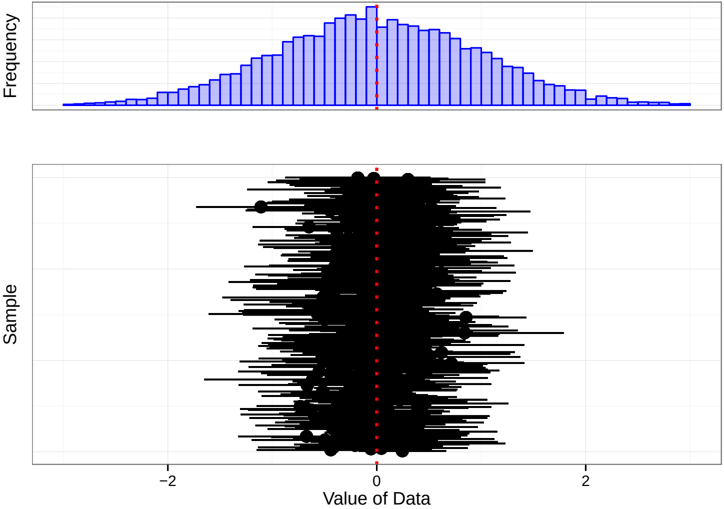

Frequency
Sample
−2
0
2
Value of Data

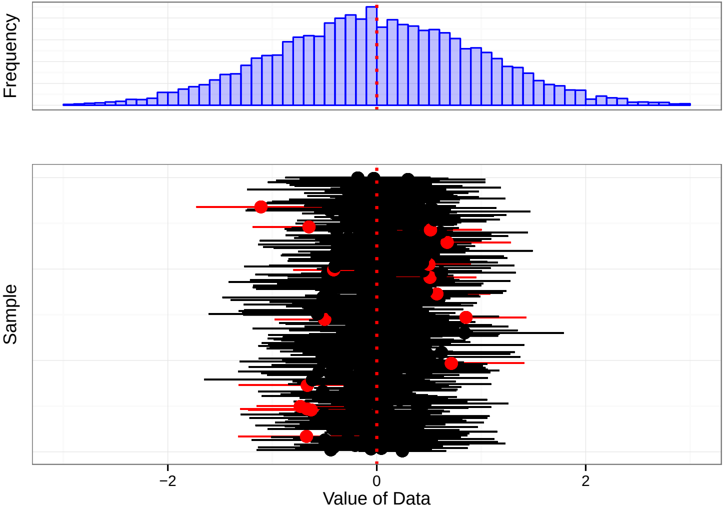


Frequency
Sample
−2
0
2
Value of Data


Frequency
Sample
−2
0
2
Value of Data


Frequency
Sample
−2
0
2
Value of Data


Frequency
Sample
−2
0
2
Value of Data



Out of 300 confidence intervals, 16 (5.3%) do not cover the true value

Criticisms in summation

Criticisms:

- P-values, confidence intervals are not intuitive

Criticisms in summation

Criticisms:

- P-values, confidence intervals are not intuitive
- Null hypotheses are often false a priori

Criticisms in summation

Criticisms:

- P-values, confidence intervals are not intuitive
- Null hypotheses are often false a priori
- Rejecting a null hypotheses does not confirm the hypothesis

Criticisms in summation

Criticisms:

- P-values, confidence intervals are not intuitive
- Null hypotheses are often false a priori
- Rejecting a null hypotheses does not confirm the hypothesis
- The P-value is a very indirect measure of support for the hypothesis of interest

Criticisms in summation

Criticisms:

- P-values, confidence intervals are not intuitive
- Null hypotheses are often false a priori
- Rejecting a null hypotheses does not confirm the hypothesis
- The P-value is a very indirect measure of support for the hypothesis of interest
- The 0.05 cutoff is completely arbitrary

Criticisms in summation

Criticisms:

- P-values, confidence intervals are not intuitive
- Null hypotheses are often false a priori
- Rejecting a null hypotheses does not confirm the hypothesis
- The P-value is a very indirect measure of support for the hypothesis of interest
- The 0.05 cutoff is completely arbitrary

On the other hand:

Criticisms in summation

Criticisms:

- P-values, confidence intervals are not intuitive
- Null hypotheses are often false a priori
- Rejecting a null hypotheses does not confirm the hypothesis
- The P-value is a very indirect measure of support for the hypothesis of interest
- The 0.05 cutoff is completely arbitrary

On the other hand:

- Null hypotheses *can* be meaningful (especially in experimental settings)

Criticisms in summation

Criticisms:

- P-values, confidence intervals are not intuitive
- Null hypotheses are often false a priori
- Rejecting a null hypotheses does not confirm the hypothesis
- The P-value is a very indirect measure of support for the hypothesis of interest
- The 0.05 cutoff is completely arbitrary

On the other hand:

- Null hypotheses *can* be meaningful (especially in experimental settings)
- P-values *are* a measure of evidence, just not very accurate

Criticisms in summation

Criticisms:

- P-values, confidence intervals are not intuitive
- Null hypotheses are often false a priori
- Rejecting a null hypotheses does not confirm the hypothesis
- The P-value is a very indirect measure of support for the hypothesis of interest
- The 0.05 cutoff is completely arbitrary

On the other hand:

- Null hypotheses *can* be meaningful (especially in experimental settings)
- P-values *are* a measure of evidence, just not very accurate
- P-values are very easy to calculate

Alternatives

1. Bayesian (-like) approaches

- Directly calculate the probability of the data, the probability of the hypothesis, etc.
- Do so using the **Bayes evidence** aka marginal likelihood
- Leads to **Bayes factors**, **model-averaging**, etc.
- Non-Bayesian attempts at something similar, like AIC-based model selection

The takehome:

Principled, meaningful, but potentially hard to calculate

Alternatives

2. Minimize prediction error

- Really, what is a meaningful test?
- How about: *does the model predict out-of-sample data better than other models?*
- Leads to **cross-validation**, **AUROCH**, etc.
- But needs a substantial amount of data

The takehome:

Simple to calculate, meaningful w.r.t. predictive accuracy, can be hard to interpret

Some references pertaining to NHT

- Cohen. 1994. The world is round ($p < 0.05$). American Psychologist.
- Johnson. 1999. The insignificance of statistical significance. Journal of Wildlife Management.
- Ellison et al. 2014. P-values, hypothesis testing, and model selection: it's deja vu all over again. Ecology.
- Murtaugh. 2014. In defense of p-values. Ecology.