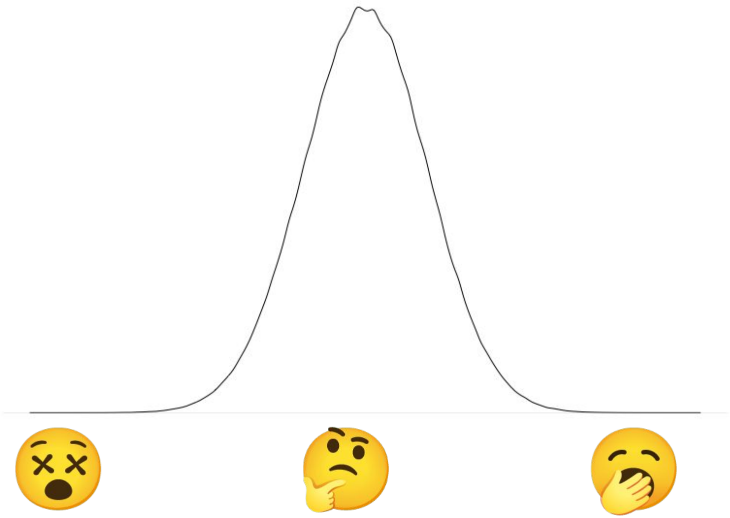
Philosophy of Science and Statistical Reasoning

Frequentist Inference

But first, …

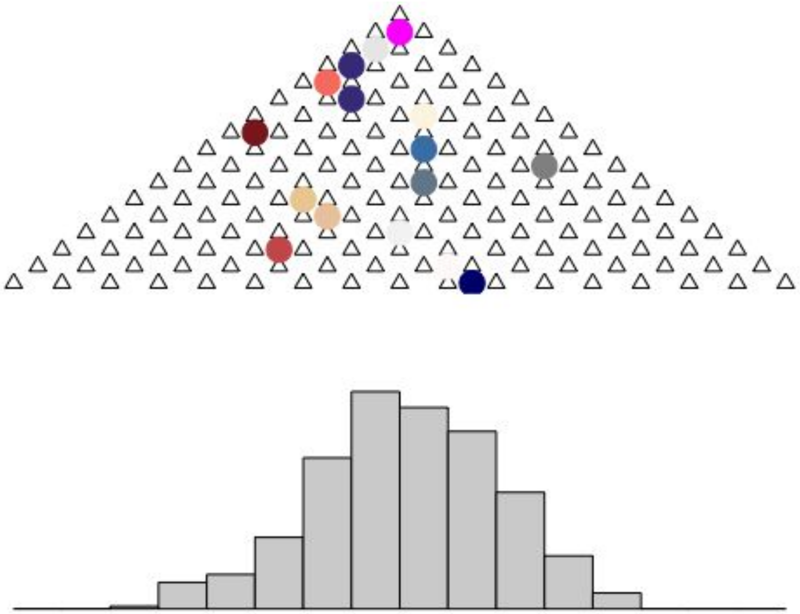
📢

- Access SOWISO
- Recidivists
- Canvas Discussions
- Open book exam
- Exam material



Student Puzzlement Scale

Previously, on statistical reasoning



You take an exam with 15 two-choice items. Where on the Galton board are the answers to the following questions captured?

- The sum score of the test is 8. How many possible ways can get you to that sum score? (1)
- What is the probability of sum score 8? (2)
- What is the probability of precisely this series (with sum score 8): 000010110011111?
- What is the probability of sum score 8 or less? (3)
- Which lowest sum scores have a probability of 10% or less? (4)
- What are the two factors that determine these probabilities? (binomial theorem!)

💭

🧑‍💻

```
choose(15, 8)  # (1) number of ways to score an 8
dbinom(8, 15, .5)  # (2) probability of sum score 8
pbinom(8, 15, .5)  # (3) probability of sum score 8 or less
qbinom(.1, 15, .5)  # (4) lowest sum scores with probability 10% or less
rbinom(n = 500, size = 15, prob = .5)  # sample 500 sum scores
```

🚁 _binom(); _norm(); _t(); _f()

You take an exam with 15 two-choice items. Where on the Galton board are the answers to the following questions captured?

- The sum score of the test is 8. How many possible ways can get you to that sum score? (1)
- What is the probability of sum score 8? (2)
- What is the probability of precisely this series (with sum score 8): 000010110011111?
- What is the probability of sum score 8 or less? (3)
- Which lowest sum scores have a probability of 10% or less? (4)
- What are the two factors that determine these probabilities? (binomial theorem!)

Pub quiz

What will we learn today?

Topics

Statistical reasoning
Empirical cycle
Probability distributions
Frequentist inference
Sample / sampling distribution
Central limit theorem
Normal distribution
P-value
Type I/II errors
Effect size
Confidence interval
Power
Test statistics
Linear regression
t-Test
Moderation
ANOVA
Nonparametric inference
Bayesian inference

Questions

How can we use probability distributions to test a null hypothesis?

How can we determine how likely our observed data is, given a null hypothesis?

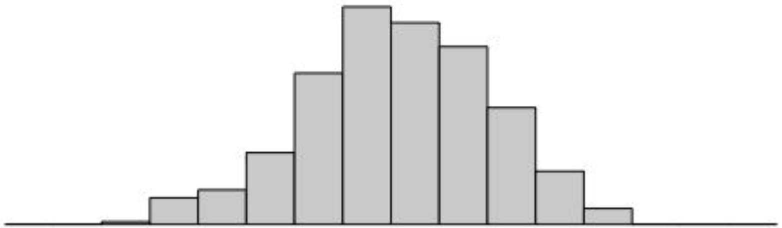
How can we make a decision about our null hypothesis? What affects this decision?

How can we quantify the uncertainty in our test result?

Frequentist inference

> Statistical inference makes propositions about a population, using data drawn from the population with some form of sampling. — [Wikipedia]

- 🪙 Frequentist inference
- 🪙 Bayesian inference
- 🪙 Nonparametric inference



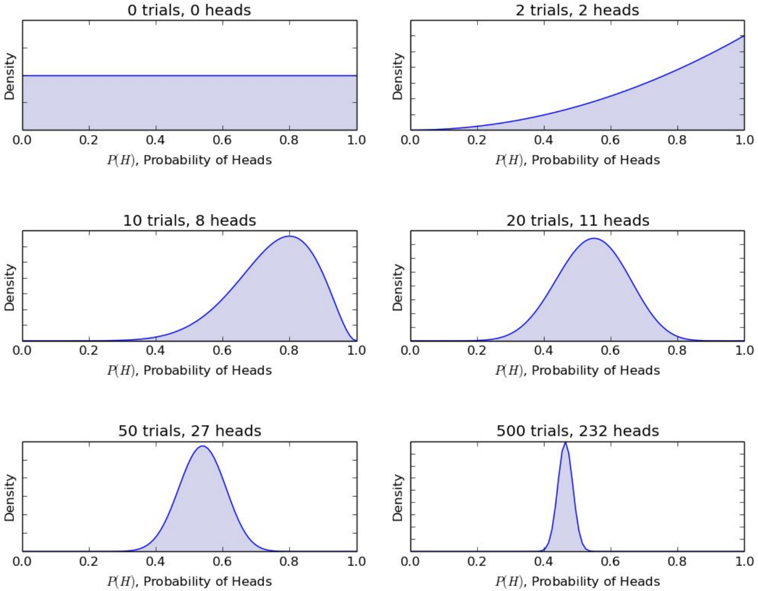
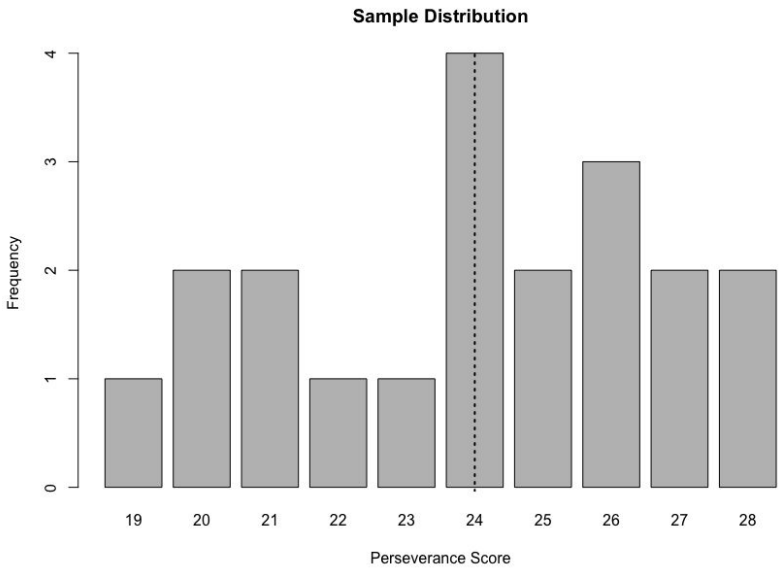


Illustration by QuantStart

Sample distribution



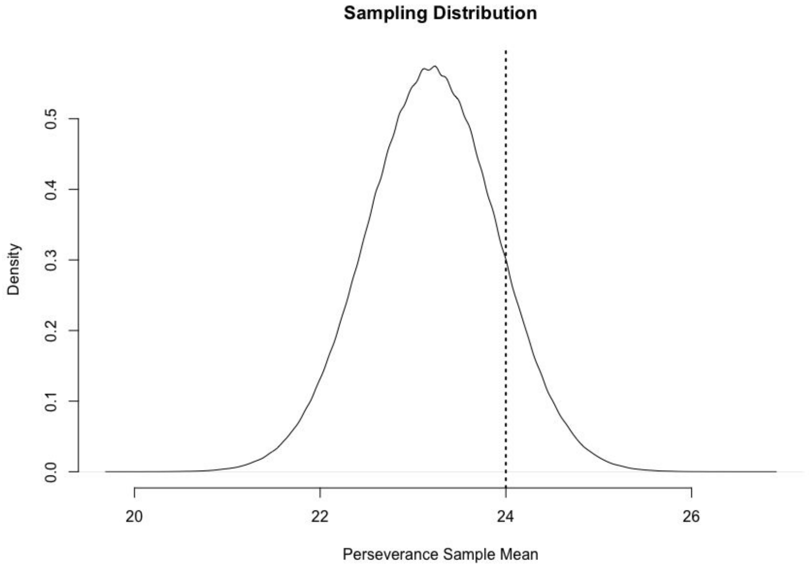
Student perseverance scale

- 40 two-choice items
- ‘0’ = (very) low perseverance
- ‘40’ = (very) high perseverance
- Psychobiology sample (n = 20)

We know that on average students score ‘23.2’.

How do we know how (un)likely the mean score of our sample is in relation to the average student?

Sampling distribution (of the sample means)



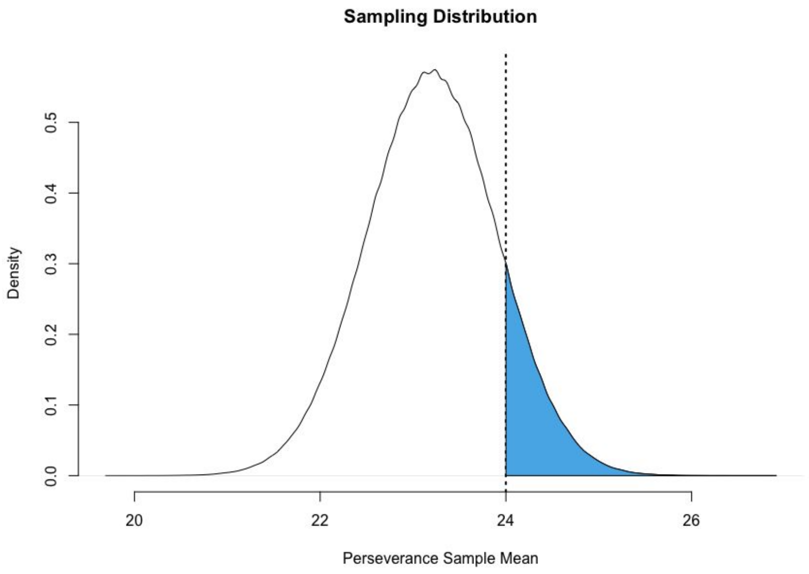
What to expect under the null hypothesis

- H_0: psychobiology student perseverance = average student perseverance

💭 How likely is our observation? Is it possible to visualize the probability of our sample mean?

💭… or a more extreme mean? (Why an area?)

How likely is our observation?



How do we know how (un)likely the mean score of our sample is in relation to the average student?

💭 Can we calculate the area?

You can't just simply take that many samples (expensive, time-consuming).

Your sample mean must come from the same distribution as the null hypothesis. Which one is that?

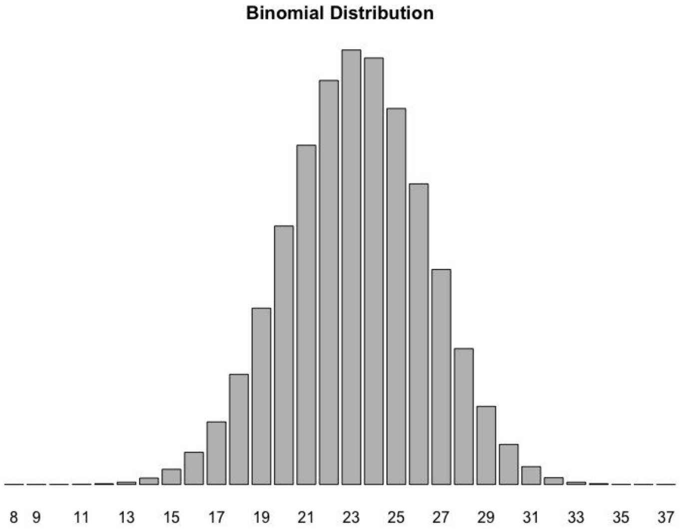
Now what? How can we learn how the null hypothesis is distributed?

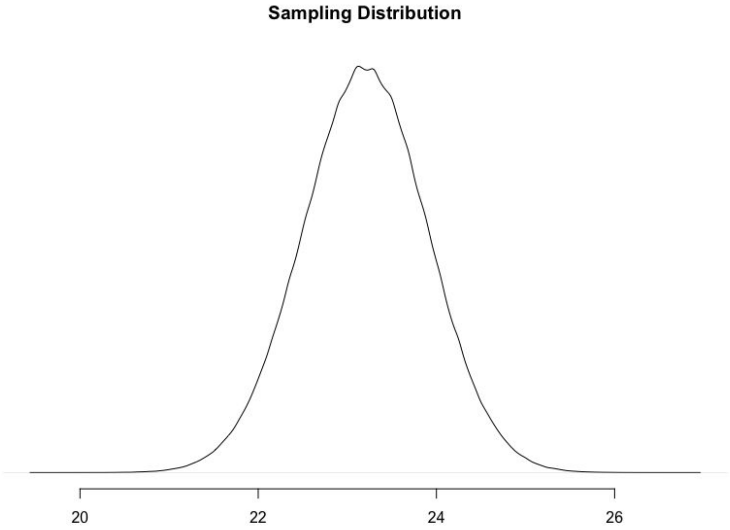
Central limit theorem

🤯 Sample means are approximately normally distributed (if the sample size is large enough). Even if the population is not normally distributed itself.

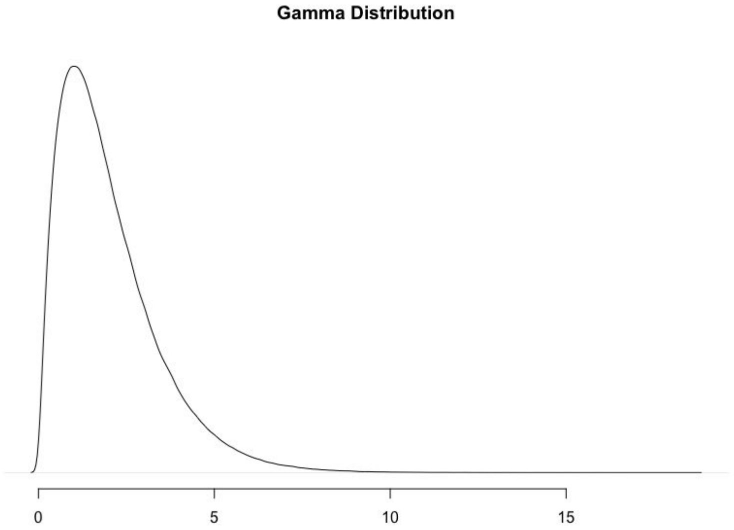
💭 Can you dissect and explain this definition?

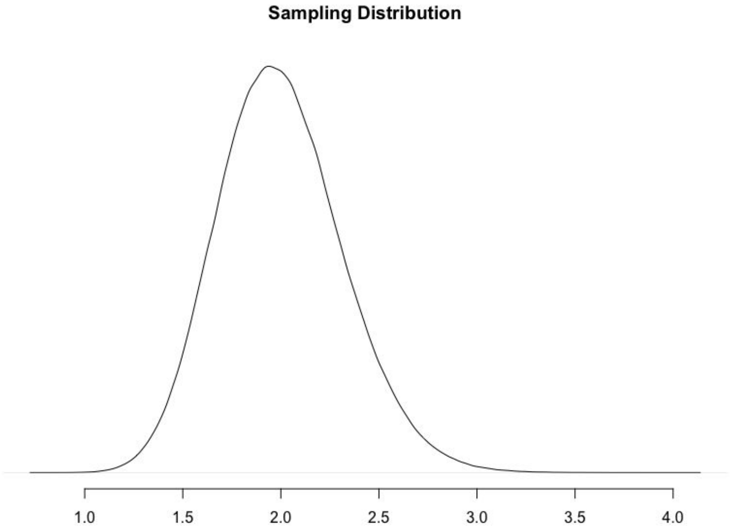
Binomial distribution





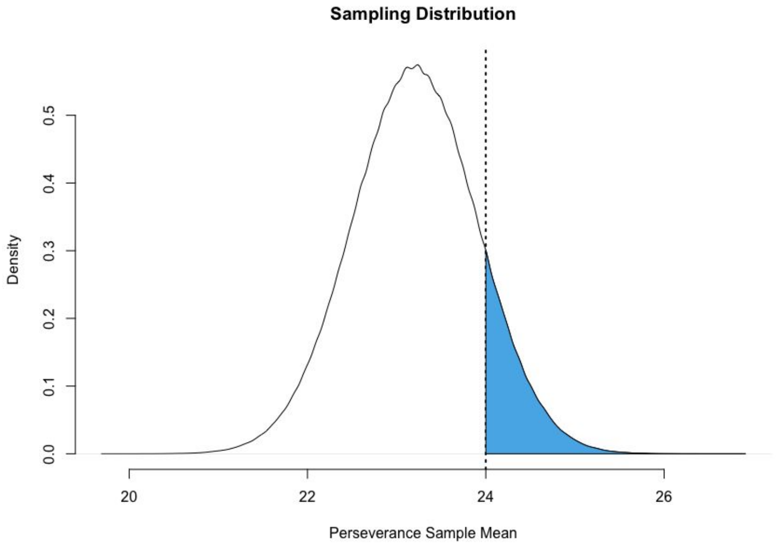
Gamma distribution



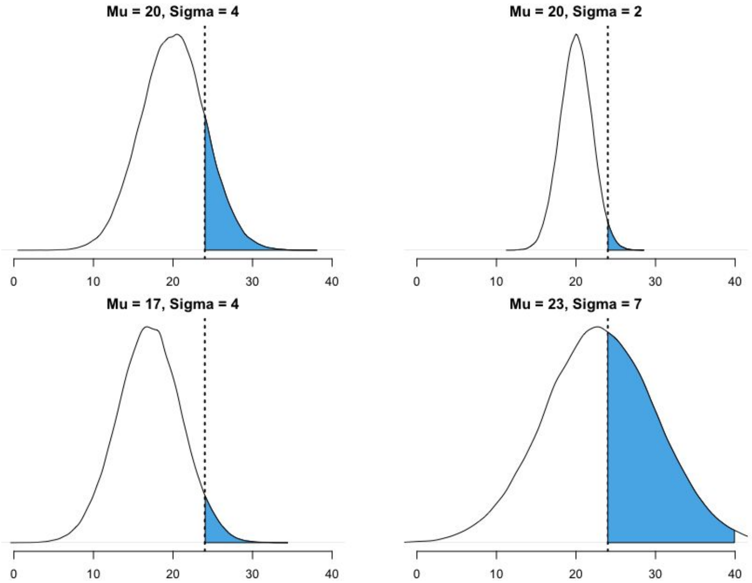


Normal distribution

💭 Can we now calculate how (un)likely our sample mean is?



No! We don't know the mean and the standard deviation of this distribution.



Central limit theorem

How do we determine the mean?

- The average student perseverance was '23.2'.

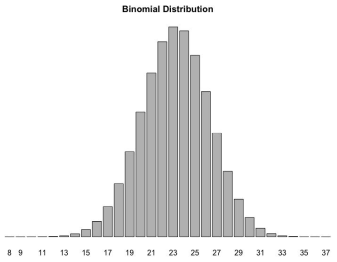
How do we determine the standard deviation?

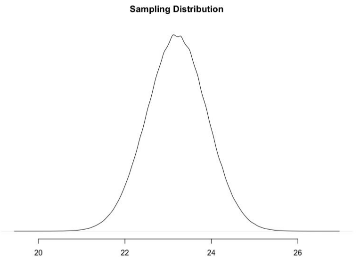
- 🤯 The central limit theorem also holds for the standard deviation!
- If our sample is large enough, we know that its standard deviation originates from an (approximately) normal distribution.

We can therefore use this **standard deviation** (s) for the sampling distribution.

The standard deviation of the sampling distribution is called the **standard error** (SE):
SE = s / √(n)

💭 Why do we need to divide s by √(n)?

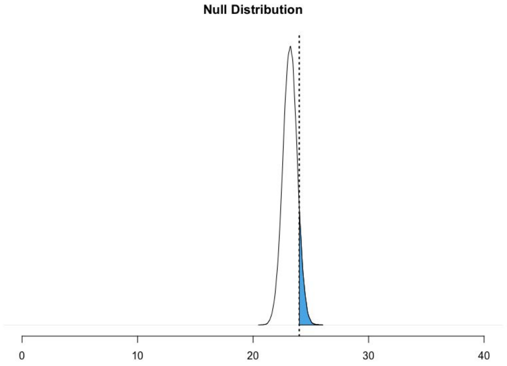




Frequentist inference

> Statistical inference makes propositions about a population, using data drawn from the population with some form of sampling. — Wikipedia

💭 What's another name for this area under the curve?



💡
Bunnies, Dragons and the 'Normal' World (CreatureCast, NYTimes)

🛠️
R simulation using the animation package.

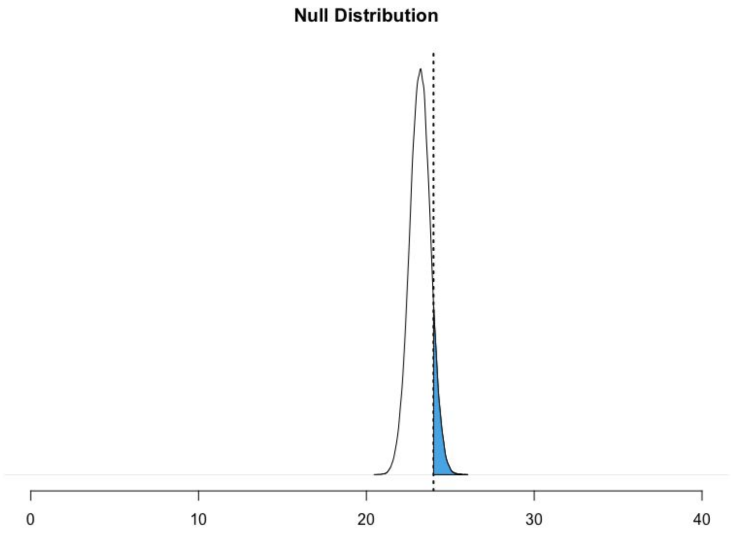
🧑‍💻
```
library(animation)
ani.options(interval = 1)
par(mar = c(3, 3, 1, 0.5), mgp = c(1.5, 0.5, 0), tcl = -0.3)
lambda = 4
f = function(n) rpois(n, lambda)
clt.ani(FUN = f, mean = lambda, sd = lambda)
```

☕

Illustration by Randall Munroe (wtf)

15:00

Decisions in Frequentist Inference



Previously

- Null hypothesis, alternative hypothesis
- Sample distribution, sample mean, sample standard deviation
- Sampling distribution, standard error
- Population mean

Next

- Test statistic
- *P*-value
- Effect size
- Power
- Type I/II error
- Confidence interval

Test statistic

> “ A test statistic is a statistic (a quantity derived from the sample) used in statistical hypothesis testing. A hypothesis test is typically specified in terms of a test statistic, considered as a numerical summary of a data-set that reduces the data to one value that can be used to perform the hypothesis test. — Wikipedia

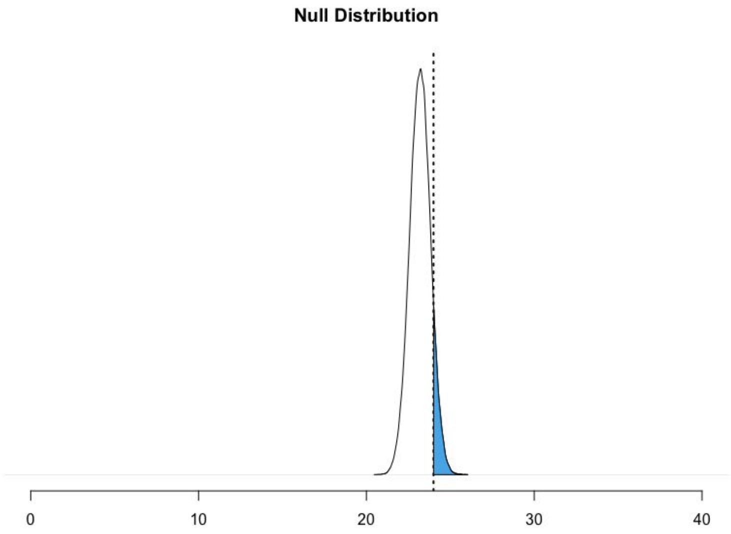
Mean, *t*-statistic, *F*-test, *Z*-test, etc.

💡 Anscombe’s quartet

🧑‍💻

```
library("datasauRus")
datasaurus_dozen %>%
  ggplot2::ggplot(aes(x = x, y = y, color = dataset)) +
  ggplot2::geom_point() +
  ggplot2::theme_void() +
  ggplot2::theme(legend.position = "none", text = element_text(size = 30)) +
  ggplot2::facet_wrap(~dataset, ncol = 3)
```

P-value



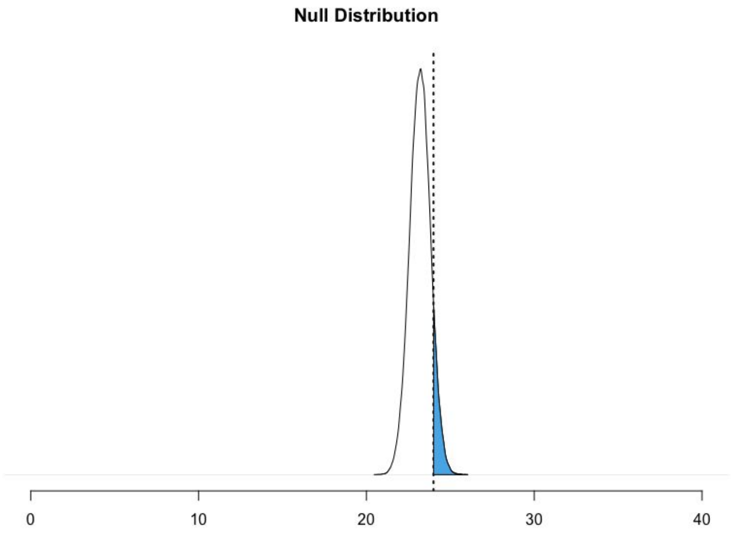
> "The probability of the observed data (or of more extreme data points), given that the null hypothesis is true: p(D|H0).
> — Gigerenzer, 2004 🔒

💭 Say, we find a *p*-value of .049. Is it statistically significant?

💭 How should you pick the alpha level?

💭 How should you pick between a one-sided and two-sided test?

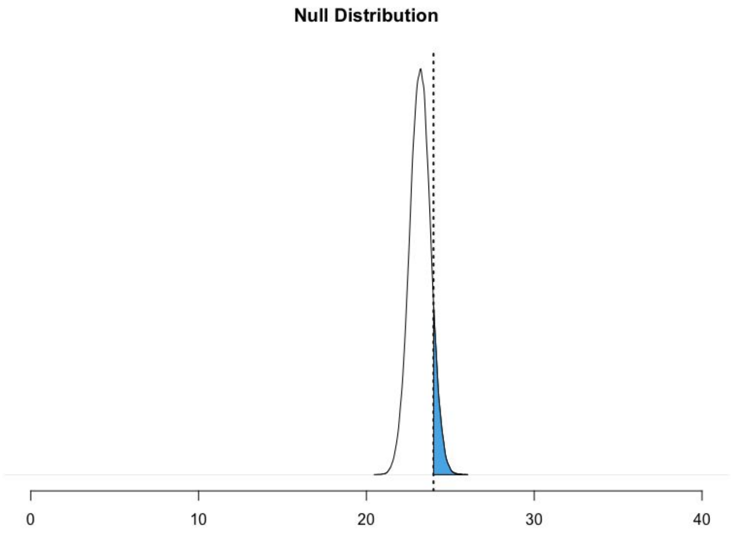
P-value



💡 How reliable are *p*-values? Should we redefine or abandon statistical significance (interesting discussion)?

🛠️ An interactive visualization of the *p*-distribution by Kristoffer Magnusson.

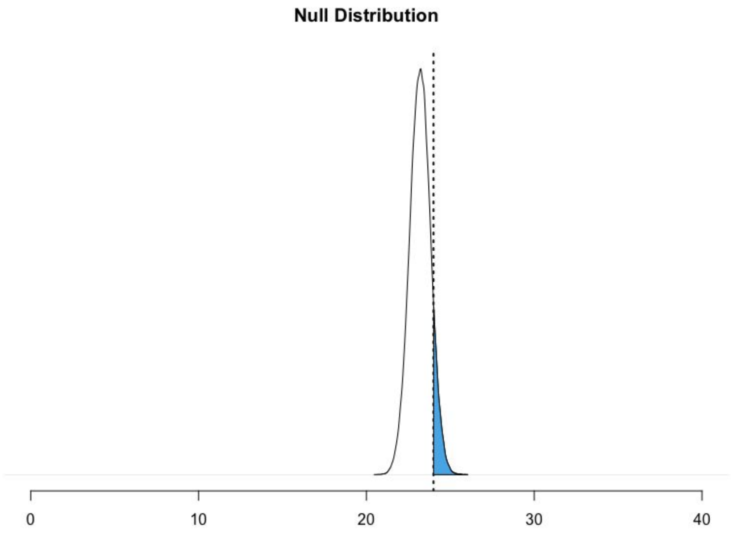
Effect size



> In statistics, an effect size is a value measuring the strength of the relationship between two variables in a population, or a sample-based estimate of that quantity.
> — Wikipedia

💭 How to express the effect size in our example?

Effect size



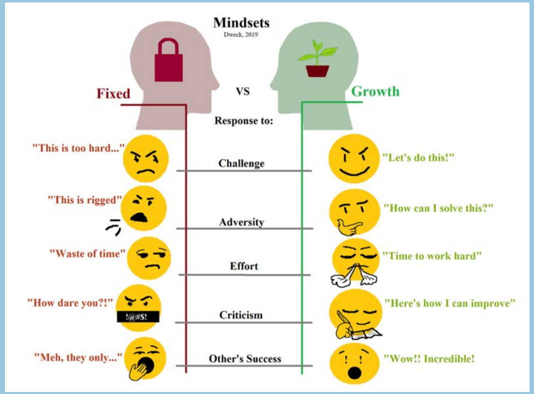
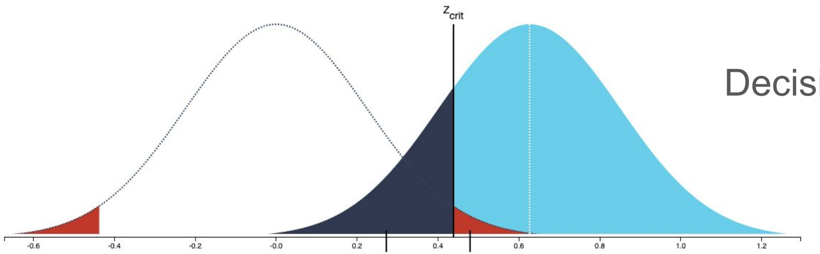


Illustration by [Wikipedia]

💡 Interesting discussion: [Brooke Macnamara], [Carol Dweck], [Andrew Gelman]
🚁 Sense and nonsense: [Funder & Ozer, 2019]

Decisions

Decision	Null Hypothesis: True	Null Hypothesis: False
Don't Reject 🗄️	**…** *(true negative)* $P(\neg Reject \mid H_0) =$ *1 − Alpha*	**Type II Error** *(false negative)* $P(\neg Reject \mid \neg H_0) =$ *Beta* (β)
Reject 🥇	**Type I Error** *(false positive)* $P(Reject \mid H_0) =$ ***Alpha*** (α)	**Power** *(true positive)* $P(Reject \mid \neg H_0) =$ *1 − Beta*



Interactive visualization by [Kristoffer Magnusson]

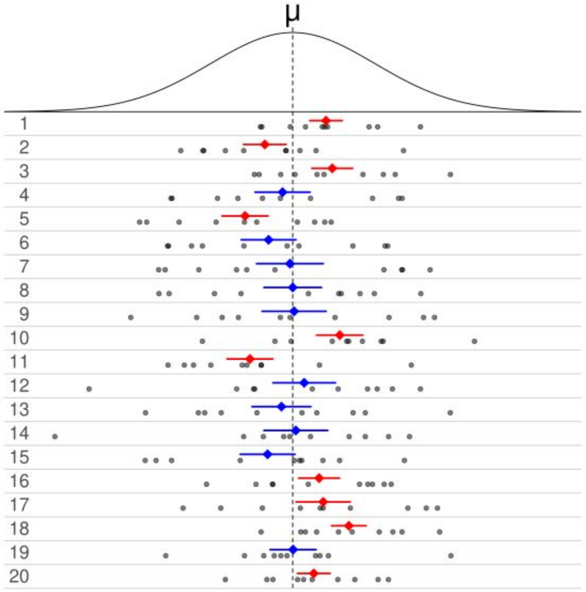
Decision	Null Hypothesis: True	Null Hypothesis: False
Don't Reject 🗄	**…** *(true negative)* $P(\neg Reject \mid H_0) = 1 - Alpha$	**Type II Error** *(false negative)* $P(\neg Reject \mid \neg H_0) =$ Beta (β)
Reject 🥇	**Type I Error** *(false positive)* $P(Reject \mid H_0) =$ ***Alpha*** (α)	**Power** *(true positive)* $P(Reject \mid \neg H_0) = 1 - Beta$

🚁

		Null Hypothesis	
		True	False
Decision	Don't Reject 🗄️	**…** *(true negative)* $P(\neg Reject \mid H_0) = 1 - Alpha$	**Type II Error** *(false negative)* $P(\neg Reject \mid \neg H_0) = Beta\ (\beta)$
	Reject 🥇	**Type I Error** *(false positive)* $P(Reject \mid H_0) =$ **Alpha** (α)	**Power** *(true positive)* $P(Reject \mid \neg H_0) = 1 - Beta$

- 💡 Type III/IV error
- 💡 Type S/M error (underpowered studies)
- 💡 File drawer problem
- 🚁 Confusion matrix

Confidence interval



Illustrations by Wikipedia (top, right)

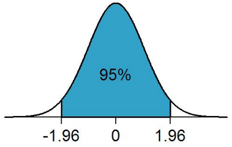
> “ If we would repeat the experiment, in *..%* of the time, the true population mean will fall within the constructed interval.

sample mean ± SE × …

90% CI = 1.64

95% CI = 1.96

99% CI = 2.56



🧑‍💻 qnorm(p = .05 / 2) # two-sided

🛠️ Web simulation by Seeing Theory, adapted from Kristoffer Magnusson. R simulation using the animation package.

Cooling down

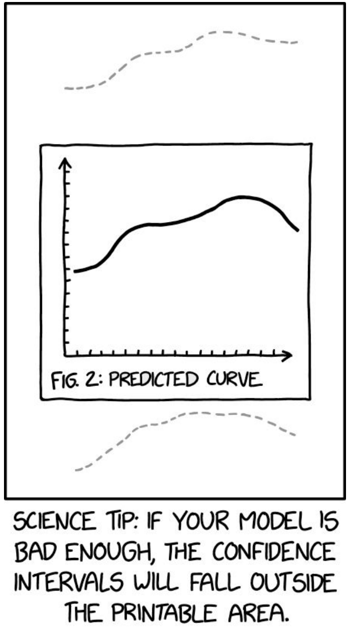
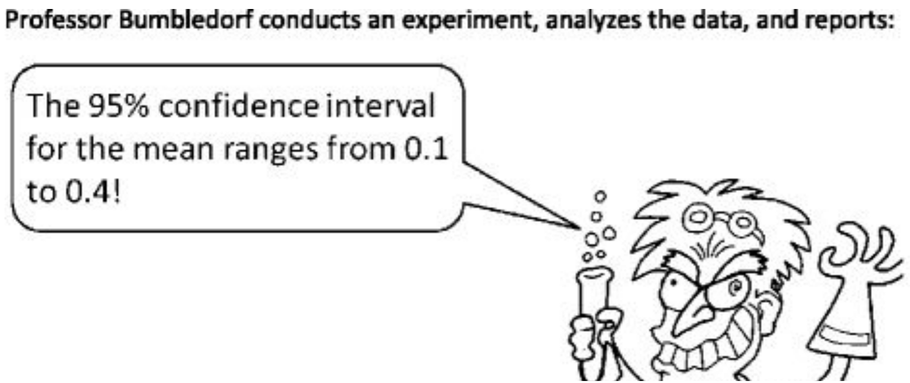


Illustration by Randall Munroe (wtf)

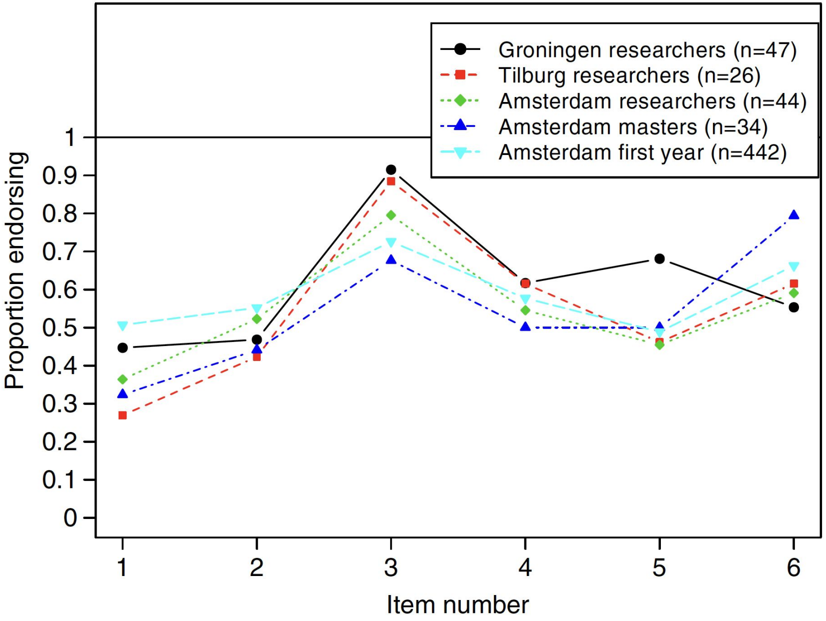
What did we learn?



1. The probability that the true mean is greater than 0 is at least 95%.
2. The probability that the true mean equals 0 is smaller than 5%.
3. The “null hypothesis” that the true mean equals 0 is likely to be incorrect.
4. There is a 95% probability that the true mean lies between 0.1 and 0.4.
5. We can be 95% confident that the true mean lies between 0.1 and 0.4.
6. If we were to repeat the experiment over and over, then 95% of the time the true mean falls between 0.1 and 0.4.

Hoekstra et al., 2014

What did we learn?



[Hoekstra et al., 2014]

Topics

Statistical reasoning
Empirical cycle
Probability distributions
Frequentist inference
Sample / sampling distribution
Central limit theorem
Normal distribution
P-value
Type I/II errors
Effect size
Confidence interval
Power
Test statistics
Linear regression
t-Test
Moderation
ANOVA
Nonparametric inference
Bayesian inference

Illustration by Jennifer Cheuk

Take-home assignments

📆 Weekly assignment

Q10 will not be graded (skip it)
Q14 read ‘post-hoc power’ simply as ‘power’

🍻 Pub quiz

Create an *informative* four-choice question about the content of today’s lecture.

An informative question has a large spread in responses across answer options.

Clarify answer options (which are (in)correct and why).

Illustration adapted from Snippets.com

Colophon

Slides

alexandersavi.nl/teaching/

License