Statistics for Machine Learning and Big Data

An Introduction

Part 2: distributions and inference

Jeff Abrahamson

7 April 2015

Distributions

Words

Given a distribution X, the

probability distribution function (pdf) (continuous) or **probability mass function (pmf)** is the probability that the variate has value x:

$$\Pr(a \leq X \leq b)$$

or

$$\Pr(X = a)$$

Words

Given a distribution X, the

cumulative probability function (cdf) is the probability that the variate is less than x:

$$\Pr(X \leq x) = \int_{-\infty}^{x} \mathrm{pdf}(x)\,\mathrm{d}x \text{ or } \sum_{i \leq x} \Pr(X = x)$$

Words

Given a distribution X, the

percent point function (ppf) is the inverse of the cdf. Given a probability, what's x? Also called the **inverse distribution**.

Words

Given a distribution X, the

survival function (sf) is the probability that the variate takes a value greater than x:

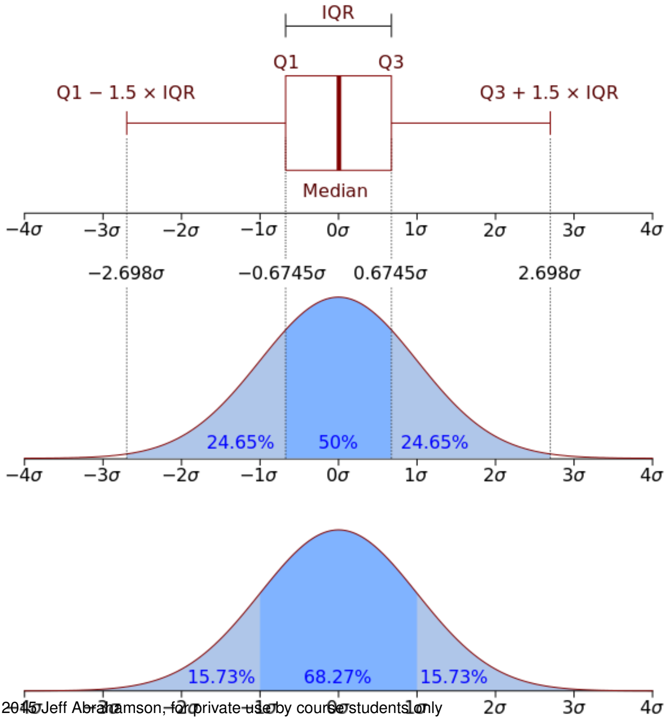
$$\mathrm{ss}(x) = \Pr(X > x) = 1 - \mathrm{cdf}(x)$$

Words

Given a distribution X, the

inverse survival function (isf) is the inverse of the survival function:

$$\text{isf}(\alpha) = \text{ppf}(1 - \alpha)$$


IQR
Q1
Q3
Q1 − 1.5 × IQR
Q3 + 1.5 × IQR
Median
−4σ −3σ −2σ −1σ 0σ 1σ 2σ 3σ 4σ
−2.698σ
−0.6745σ
0.6745σ
2.698σ
24.65%
50%
24.65%
−4σ −3σ −2σ −1σ 0σ 1σ 2σ 3σ 4σ
15.73%
68.27%
15.73%
−4σ −3σ −2σ −1σ 0σ 1σ 2σ 3σ 4σ

Code lab

ipython notebook

Uniform distribution

Takes value 1 with probability 1.

Model	identical events
Parameters	none
Mean	$\frac{1}{2}$
Variance	$\frac{1}{12}$

Bernoulli distribution

Takes value 1 with probability p and 0 with probability $1 - p$.

Model	turning coins
Parameters	$p \in [0, 1]$
Mean	p
Variance	$p(1 - p)$

The pmf(k) (mass function) is $1 - p$ if $k = 0$ and p if $k = 1$.

Geometric distribution

Two definitions:

- The probability distribution of the number X of Bernoulli trials needed to get one success, supported on the set $\{1, 2, 3, \dots\}$
- The probability distribution of the number $Y = X - 1$ of failures before the first success, supported on the set $\{0, 1, 2, 3, \dots\}$

Wikipedia, geometric distribution

Geometric distribution

Two definitions:

- The probability distribution of the number X of Bernoulli trials needed to get one success, supported on the set $\{1, 2, 3, \ldots\}$
- The probability distribution of the number $Y = X - 1$ of failures before the first success, supported on the set $\{0, 1, 2, 3, \ldots\}$

Model	$\ldots$
Parameters	$p \in (0, 1]$
Mean	$\frac{1}{p}$ or $\frac{1-p}{p}$
Variance	$\frac{1-p}{p^2}$

Poisson distribution

Probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event.

Model	radioactive decay, network packets
Parameters	$\lambda \in \mathbb{R}^+$
Mean	λ
Variance	λ

$$\text{pmf}(k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

(Sometimes we say μ instead of λ.)

Binomial distribution

$\mathbf{B}(n, p) =$ Number of successes in a sequence of n independent bernoulli trials (yes/no experiments), each of which yields success with probability p.

Model	sequences of coin tosses
Parameters	n, p
Mean	np
Variance	$np(1-p)$

Binomial distribution

$\mathbf{B}(n, p) =$ Number of successes in a sequence of n independent bernoulli trials (yes/no experiments), each of which yields success with probability p.

Model	sequences of coin tosses
Parameters	n, p
Mean	np
Variance	$np(1-p)$

$$\text{pmf}(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for $k \in \{0, 1, \ldots, n\}$.

Binomial distribution: Example

19% of the British public smokes. A study of 500 people reports 90 smokers. What is the probability of finding 90 or fewer smokers in a sample of 500 UK residents chosen at random?

http://en.wikipedia.org/wiki/Smoking_in_the_United_Kingdom

Binomial distribution: Example

19% of the British public smokes. A study of 500 people reports 90 smokers. What is the probability of finding 90 or fewer smokers in a sample of 500 UK residents chosen at random?

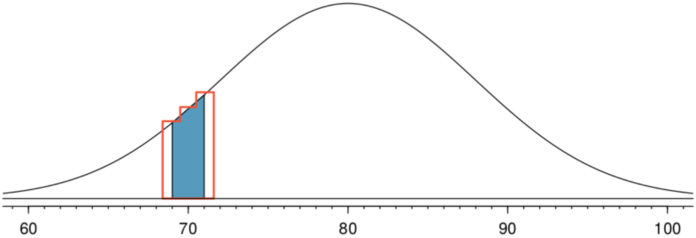
Exercise

Binomial distribution: Example

19% of the British public smokes. A study of 500 people reports 90 smokers. What is the probability of finding 90 or fewer smokers in a sample of 500 UK residents chosen at random?

Binomial distribution: Example

19% of the British public smokes. A study of 500 people reports 90 smokers. What is the probability of finding 90 or fewer smokers in a sample of 500 UK residents chosen at random?



Negative binomial distribution

How many Bernoulli trials with parameter p until we have n successes?

Model	
Parameters	p, n
Mean	$\frac{pn}{1-p}$
Variance	$\frac{pn}{(1-p)^2}$

$$\mathrm{pmf}(k) = \binom{k+n-1}{n-1} p^n (1-p)^k$$

for $k \geq 0$.

Comparing discrete distributions

Binomial distribution.
Fixed number of trials, measures probability of success.

Negative binomial distribution.
Fixed number of successes, measures probable number of trials.

Poisson distribution.
Fixed number of trials, measures probable number of successes.

Normal distribution

$\mathcal{N}(\mu, \sigma^2)$, about which we will say a great deal over the next hour and the rest of the day.

Model	cf. CLT
Parameters	μ, σ^2
Mean	μ
Variance	σ^2

Normal distribution

$\mathcal{N}(\mu, \sigma^2)$, about which we will say a great deal over the next hour and the rest of the day.

Model	cf. CLT
Parameters	μ, σ^2
Mean	μ
Variance	σ^2

$$\text{pdf}(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

Z score

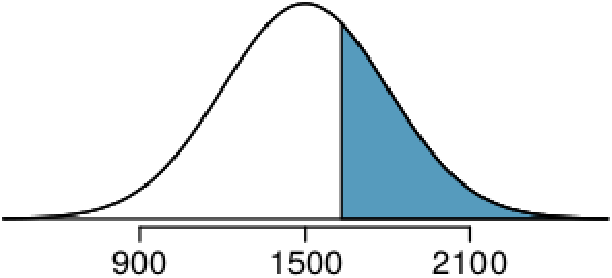
$$Z = \frac{x - \mu}{\sigma}$$

Example

The scores on an exam are approximately $\mathcal{N}(1500, 300)$. What is the probability a random exam taker (one about whom we know nothing a priori) scores above 1630?

Example

The scores on an exam are approximately $\mathcal{N}(1500, 300)$. What is the probability a random exam taker (one about whom we know nothing a priori) scores above 1630?



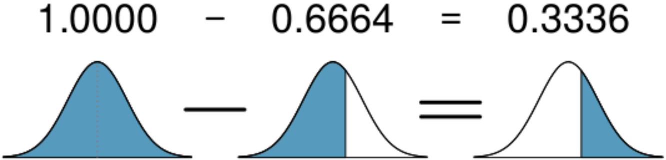
Example

The scores on an exam are approximately $\mathcal{N}(1500, 300)$. What is the probability a random exam taker (one about whom we know nothing a priori) scores above 1630?

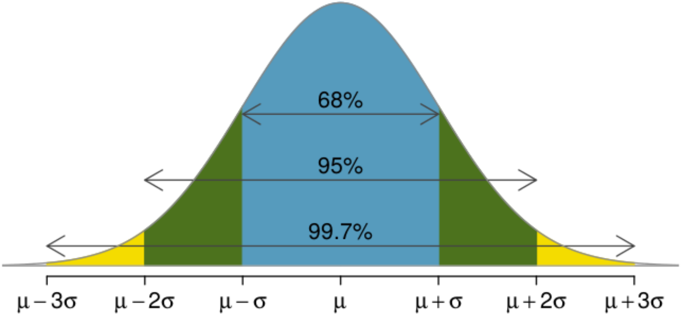
$$Z = \frac{1630 - 1500}{300} = .43$$

Example

The scores on an exam are approximately $\mathcal{N}(1500, 300)$. What is the probability a random exam taker (one about whom we know nothing a priori) scores above 1630?

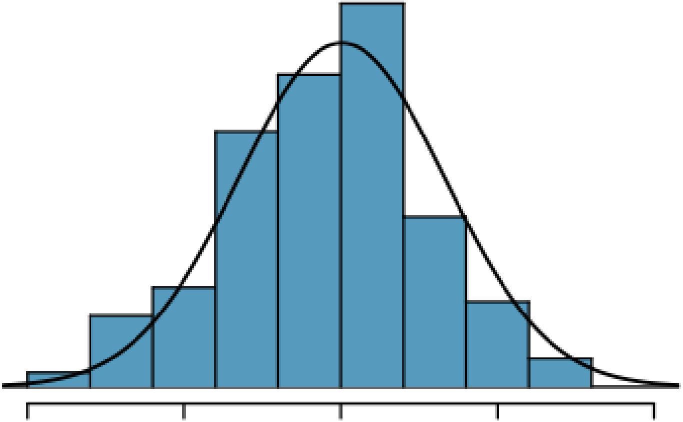


65-95-99.7 Rule



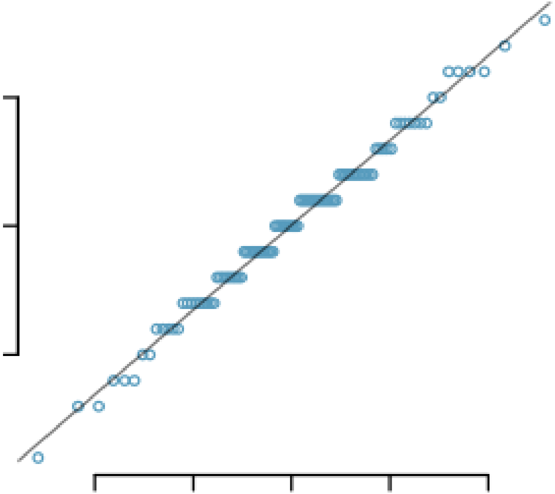
Evaluating Normal Approximations

Easy technique 1: visually compare to normal plot.



Evaluating Normal Approximations

Easy technique 2: normal probability plot.



Also known as a quantile-quantile plot.

Inference

Inference

This is the important part.

Inference

Goal: Understand the quality of parameter estimates.

Inference

Goal: Understand the quality of parameter estimates.

Examples:

- How close is $\overline{x}$ to μ?

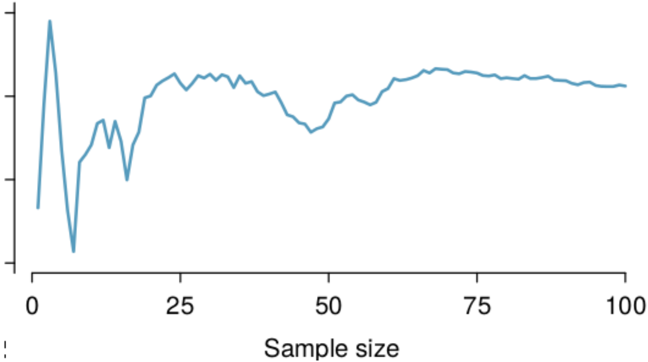
Point estimates

Population mean (μ) $\neq$ sample mean ($\overline{x}$).

Inference Concepts

Running mean. Sequence of partial sums (divided by number in sum).

Running mean. Sequence of partial sums (divided by number in sum).



Inference Concepts

Running mean. Sequence of partial sums (divided by number in sum).

Sampling variation. Change of $\overline{x}$ from one sample to the next.

Inference Concepts

Running mean. Sequence of partial sums (divided by number in sum).

Sampling variation. Change of $\overline{x}$ from one sample to the next.

Sampling distribution. The distribution of possible point samples of a fixed size from a given population.

Inference Concepts

Running mean. Sequence of partial sums (divided by number in sum).

Sampling variation. Change of $\overline{x}$ from one sample to the next.

Sampling distribution. The distribution of possible point samples of a fixed size from a given population.

"All problems in computer science can be solved by another level of indirection" *—David Wheeler*

Inference Concepts

Running mean. Sequence of partial sums (divided by number in sum).

Sampling variation. Change of $\overline{x}$ from one sample to the next.

Sampling distribution. The distribution of possible point samples of a fixed size from a given population.

"All problems in computer science can be solved by another level of indirection, except of course for the problem of too many indirections." *—David Wheeler*

Inference Concepts

Running mean. Sequence of partial sums (divided by number in sum).

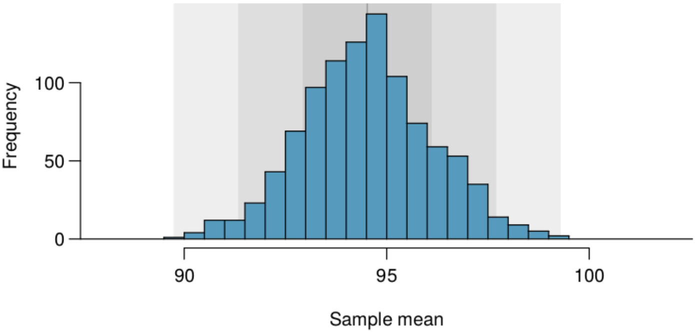
Sampling variation. Change of $\overline{x}$ from one sample to the next.

Sampling distribution. The distribution of possible point samples of a fixed size from a given population.

"All problems in computer science can be solved by another level of indirection, except of course for the problem of too many indirections." *—David Wheeler*

http://en.wikipedia.org/wiki/David_Wheeler_%28British_computer_scientist%29

Sampling distribution



Sampling distribution

Exercise: Generate uniform population, sample, and plot sampling distribution.

Sampling distribution

Exercise: Generate highly skewed population, sample, and plot sampling distribution.

Sampling distribution

Given n observations, the standard error of the sample means is

$$\text{S.E.} = \frac{\sigma}{\sqrt{n}}$$

Rules of Thumb

When is the preceding a good approximation?

- $n > 30$
- The population is not too skewed.

When is the sample likely to be independent?

- Use simple random sampling.
- $n < 10\%$ of the population.

Summary

- Point estimates estimate population parameters.
- Point estimates have error and vary among samples.
- The standard error quantifies the variation among point estimates.

Confidence intervals

Sample n points, choose an interval around the sample mean.

A 95% confidence interval means if we sample repeatedly, about 95% of the samples will contain the population mean.

Confidence intervals

Sample n points, choose an interval around the sample mean.

A 95% confidence interval means if we sample repeatedly, about 95% of the samples will contain the population mean.

95% confidence means ± 2 S.E.

Confidence intervals

Sample *n* points, choose an interval around the sample mean.

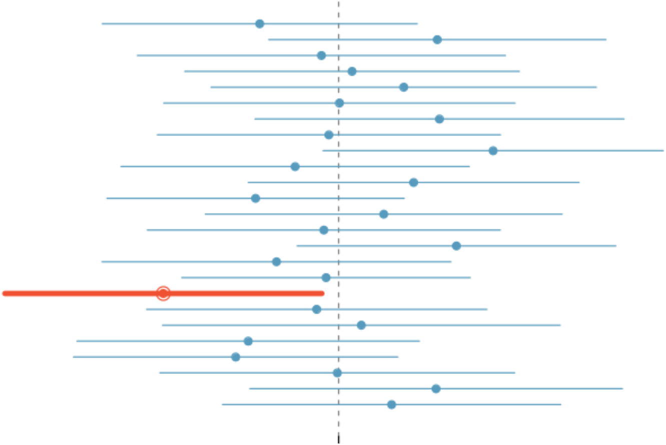
A 95% confidence interval means if we sample repeatedly, about 95% of the samples will contain the population mean.

95% confidence means ± 1.96 S.E.

Confidence intervals

Sample n points, choose an interval around the sample mean.

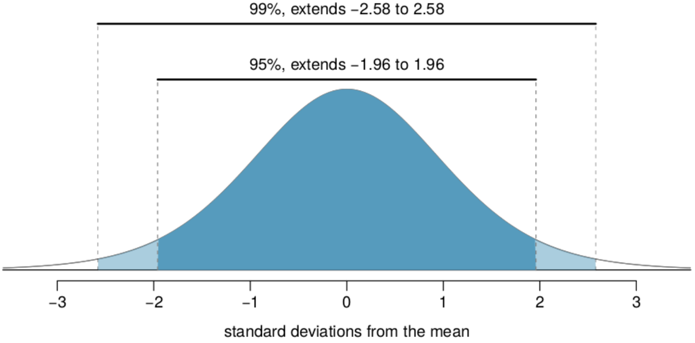
A 95% confidence interval means if we sample repeatedly, about 95% of the samples will contain the population mean.



Confidence intervals

Sample *n* points, choose an interval around the sample mean.

A 95% confidence interval means if we sample repeatedly, about 95% of the samples will contain the population mean.



p-value

Do people die later now than 10 years ago?

Compute the mean death age from 10 years ago.
Compute a sample mean now and confidence interval.

> H_0: *no*
> H_A: *yes*

If the mean is within the confidence interval, we reject H_A.
If the mean is outside the confidence interval, we accept H_A.

p-value

The *p*-value is a way of quantifying the strength of the evidence against the null hypothesis and in favor of the alternative.

p-value

The *p*-value is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis is true.

Example: p-value

Googlers gain on average 7 kgs within one year of working at the company.

Facebook thinks their food is better, they say their employees gain more weight in the first year.

H_0: $\mu = 7$
H_A: $\mu > 7$

Sample $n = 110$ facebookers, $\overline{x} = 7.42$ and $s = 1.75$.

Example: p-value

Googlers gain on average 7 kgs within one year of working at the company. *(This may not be true, but it's a common joke.)*

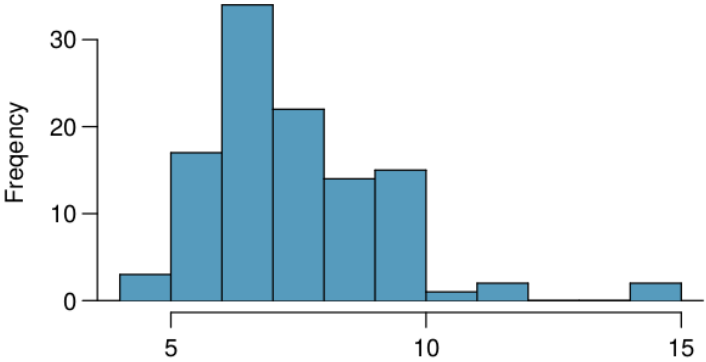
Facebook thinks their food is better, they say their employees gain more weight in the first year. *This would be a perverse argument for joining facebook.*

H_0: $\mu = 7$
H_A: $\mu > 7$

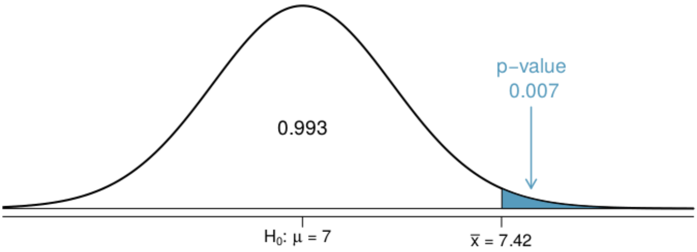
Sample $n = 110$ facebookers, $\overline{x} = 7.42$ and $s = 1.75$.

S.E. $= \frac{s}{\sqrt{n}} = \frac{1.75}{\sqrt{110}} \approx 0.17$

Example: *p*-value



Example: *p*-value

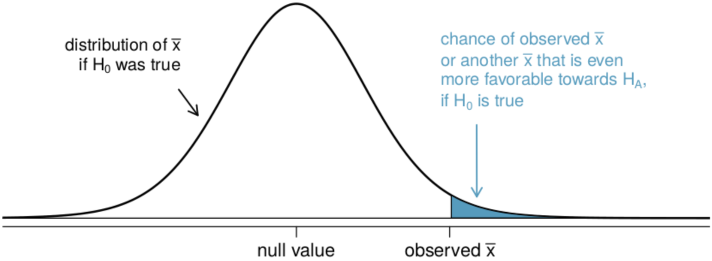


Example: *p*-value

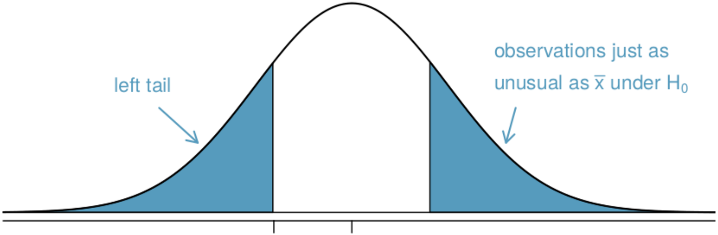
$$Z = \frac{\overline{x} - \mu}{\text{S.E.}} = \frac{7.42 - 7}{0.17} \approx 2.47$$

$\Pr(\overline{x} < 7.42) = .993$, so $p = 1 - .993 = .007$.

Example: *p*-value



Example: *p*-value



Central Limit Theorem (CLT)

The distribution of $\overline{x}$ is approximately normal. The approximation can be poor if the sample size is small, but it improves with larger sample sizes.

Open Statistics

Central Limit Theorem (CLT)

Given certain conditions, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined expected value and well-defined variance, will be approximately normally distributed, regardless of the underlying distribution.

http://en.wikipedia.org/wiki/Central_limit_theorem

Central Limit Theorem (CLT)

Suppose $\{X_1, X_2, \dots\}$ is a sequence of i.i.d. random variables with $\mathbf{E}[X_i] = \mu$ and $\mathbf{Var}[X_i] = \sigma^2 < \infty$. Then as n approaches infinity, the random variables $\sqrt{n}(S_n - \mu)$ converge in distribution to a normal $\mathcal{N}(0, \sigma^2)$:

$$\sqrt{n}\bigg(\bigg(\frac{1}{n}\sum_{i=1}^{n} X_i\bigg) - \mu\bigg) \ \xrightarrow{d}\ N(0,\ \sigma^2).$$

http://en.wikipedia.org/wiki/Central_limit_theorem

Vocabulary

A point estimate is **unbiased** if the sampling distribution of the estimate is centred at the parameter it estimates.

Significance (revisited)

- Flip a coin *N* times. Get 49% heads. Is the coin fair?
- Congress/parliament has *N* members of whom *m* are male. Do we discriminate against women?

- Careful about what we conclude. Here we can't conclude anything about cause or source.

Code lab: A/B testing

Exercises:

For each,

- What are H_0 and H_A?
- Should we adopt method B based on the evidence?

A/B Jeopardy

a game in pairs

Questions?

`purple.com/talk-feedback`

Break

