

The t-test

- Original motivation: yields of barley. Let's illustrate it with a problem of **quality control** for the Guinness stout.
- Say you are measuring barley concentration in small beer samples. Your sample is assumed to follow some unknown i.i.d. Gaussian:

$$X^{(i)} \sim N(\mu, \sigma^2), i = 1, 2, \dots, n$$



The t-test

- We would like to know if we are correctly manufacturing it with target mean μ_0 . We can formulate it as a hypothesis test

$$H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_0$$

- Notice that in most cases the **alternative hypothesis** is just the negation of H_0 .

The t-test

- Gosset (a.k.a. “Student”) derived the distribution of the following statistic

$$T = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{S_n} \sim \mathcal{T}(n - 1)$$

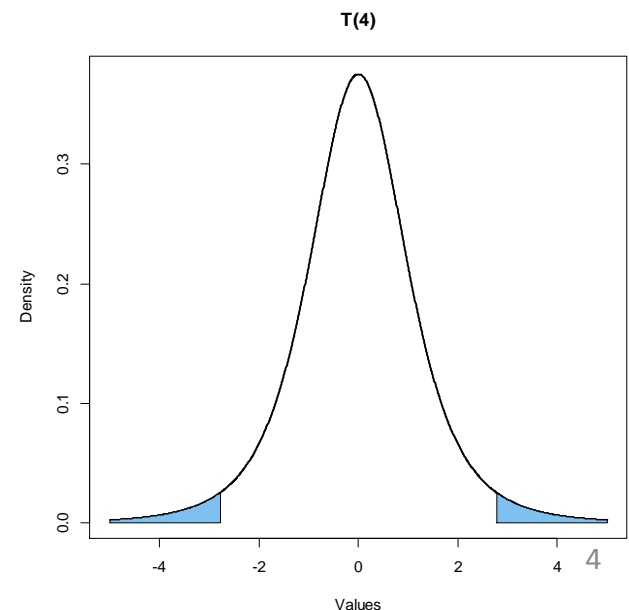
\bar{X}_n : sample mean

S_n : sample standard deviation

- This is called a t-distribution with $n - 1$ degrees of freedom. The formula is ugly, but for large n it is essentially a Normal (0, 1).

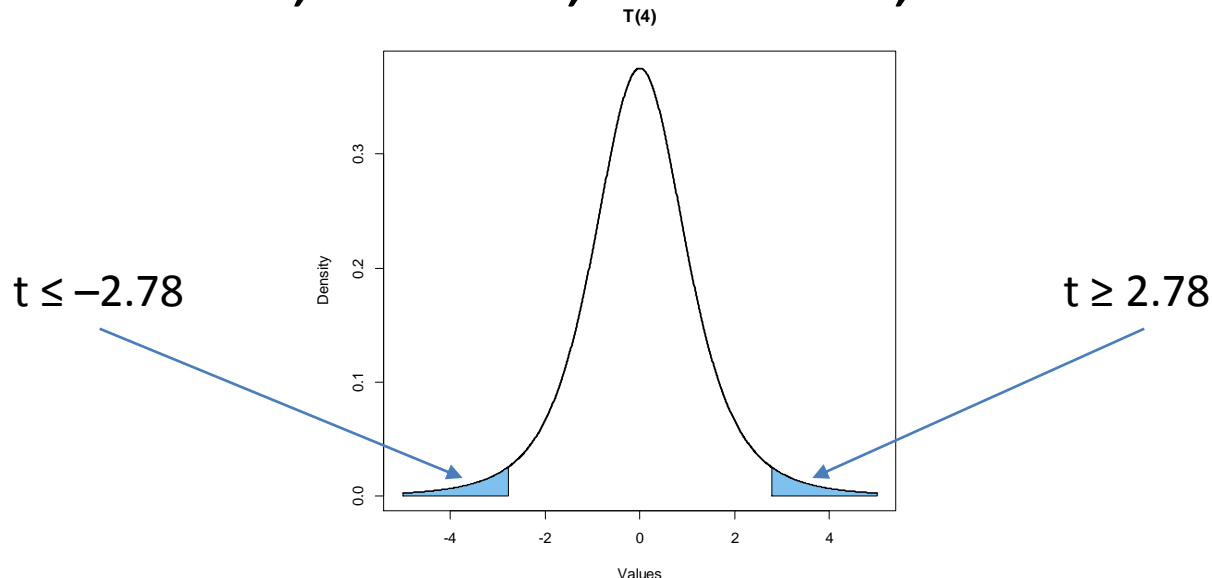
The t-test

- Now, we can look at the probability of “extreme values”.
- The true mean can differ from μ_0 by being smaller or larger. A two-tailed test is used in this example.
- The blue in the figure is the critical region.



The t-test

- We reject H_0 with level α only if t is smaller than the respective $\alpha / 2$ quantile, or larger than the $1 - \alpha / 2$ quantile.
- For instance, if $n = 5$, $\alpha = 0.05$, we have



The Wald Test

- Recall our friend, the Central Limit Theorem.
 - In a simplified way: averages of big samples look like Gaussian random variables.
- The t-test is motivated by small, Gaussian samples.
- The Wald test uses the same statistic (!) as the t-test. The interpretation is different.
 - Samples $X^{(i)}$ can be of “any” distribution.
 - Sample sizes are assumed to be “big enough” to that the CLT kicks in.
 - Hence the distribution of the statistic (let’s call it W , but it’s the same formula as T) is $N(0, 1)$ now.

Goodness-of-fit Tests

- We can think of testing more general assumptions. The t-test goes for a particular mean. What about other **constraints** in the distribution?

Goodness-of-fit Tests

- Like Michelangelo's "statue trapped in the stone", your model is not what you add, but what you remove. *Hypothesis testing doesn't answer whether what you added to the model was valid, but whether what you subtracted didn't hurt you.*
- In modelling terms: subtraction = constraints.



Wikimedia Commons

Goodness-of-fit Tests

- Example: testing whether two discrete variables are independent.
- In probability terms, this means

$$P(X, Y) = P(X)P(Y)$$

– notice the implication: $P(Y | X) = P(Y)$

Goodness-of-fit Tests

- **Contingency table** of twin data
 - D_j = depression, sibling j
 - A_j = dependence on alcohol, sibling j

		$D_1 = 0$		$D_1 = 1$	
		$D_2 = 0$	$D_2 = 1$	$D_2 = 0$	$D_2 = 1$
$A_1 = 0$	$A_2 = 0$	288	80	92	51
	$A_2 = 1$	15	9	7	10
$A_1 = 1$	$A_2 = 0$	8	4	8	9
	$A_2 = 1$	3	2	4	7

- Is there an association between depression and alcohol dependency across different subjects?

Goodness-of-fit Tests

- The **chi-squared test** compares “expected” versus “observed” outcomes.
- In a nutshell: in this case, for every combination of values of the two variables, compare its frequency of co-occurrences against the product of its marginal frequencies. We can derive a test statistic using a particular way of aggregating these numbers.
- This statistic, Pearson’s χ^2 , has a so-called chi-squared distribution. In general, this test can be used to check whether a particular pmf explains some observed **multinomial data**.

Paired Tests

- As a final example, consider comparing measurements that are tied to a single unit (a person, a beer vat, and so on).
- This is typical when we apply two treatments to a same individual and contrast the results.
 - Further details? Yes, *STATG002*.



BEFORE



AFTER

Paired Tests

- A **signed rank test** (Wilcoxon test) compares the differences of two $X^{(i)}$ and $Y^{(i)}$ measurements within a single individual i .
- The idea is to build data as if it came from the null. For that, build differences $X^{(i)} - Y^{(i)}$ for all possible pairs, look at the sign. Under the null, it is possible to find the distribution of a test statistic based on these rearrangements.
- The null here is whether $P(X^{(i)} > Y^{(i)}) = P(Y^{(i)} > X^{(i)})$.
 - Question: What would you do if we assume they are Gaussian distributed with the same variance?

Hypothesis testing

WORDS OF CAUTION AND PRAGMATIC ADVICE

Before We Conclude

- Let's remind ourselves why we are doing this!

Objections to Hypothesis Testing

- The null is “always false”, especially depending on the amount of precision used.
- Dichotomization of decisions may lead to inconsistencies. We can “accept” some null H_0^i , and “reject” some H_0^j , even if $H_0^i \Rightarrow H_0^j$!
- It is confusing and people often misuse it.
 - Well, don’t you disappoint me!

Why Do Hypothesis Testing

- A typical industry practice: A/B testing
 - Do two treatments. Say, offer product with two different variations (price, colour, user interface, etc.)
 - Does the distribution of outcomes (sales, consumer satisfaction, etc.) change in some way (mean, variance, maximum value, etc.)?
 - Set H_0 as the “no change” hypothesis.

Kohavi et al. (2009) "Controlled experiments on the web: survey and practical guide"

<http://dl.acm.org/citation.cfm?id=1485091>

Why Do Hypothesis Testing

- Granted, you may be/should be interested on the *size* of the effect (e.g. difference in sales means).
- However, would you have a large enough sample to distinguish it from zero?
- The machinery of hypothesis testing helps to tell you whether you are asking a ridiculous question to begin with.
 - that is, estimating effect size when the sample size you have can't really distinguish it from zero

Why Do Hypothesis Testing

- Models rely on assumptions that sometimes are “good enough”. For example, Gaussianity and independence.
- There is a more convincing story, instead of meaningless handwaving, if you actually show that your data cannot falsify these assumptions!

Why Do Hypothesis Testing

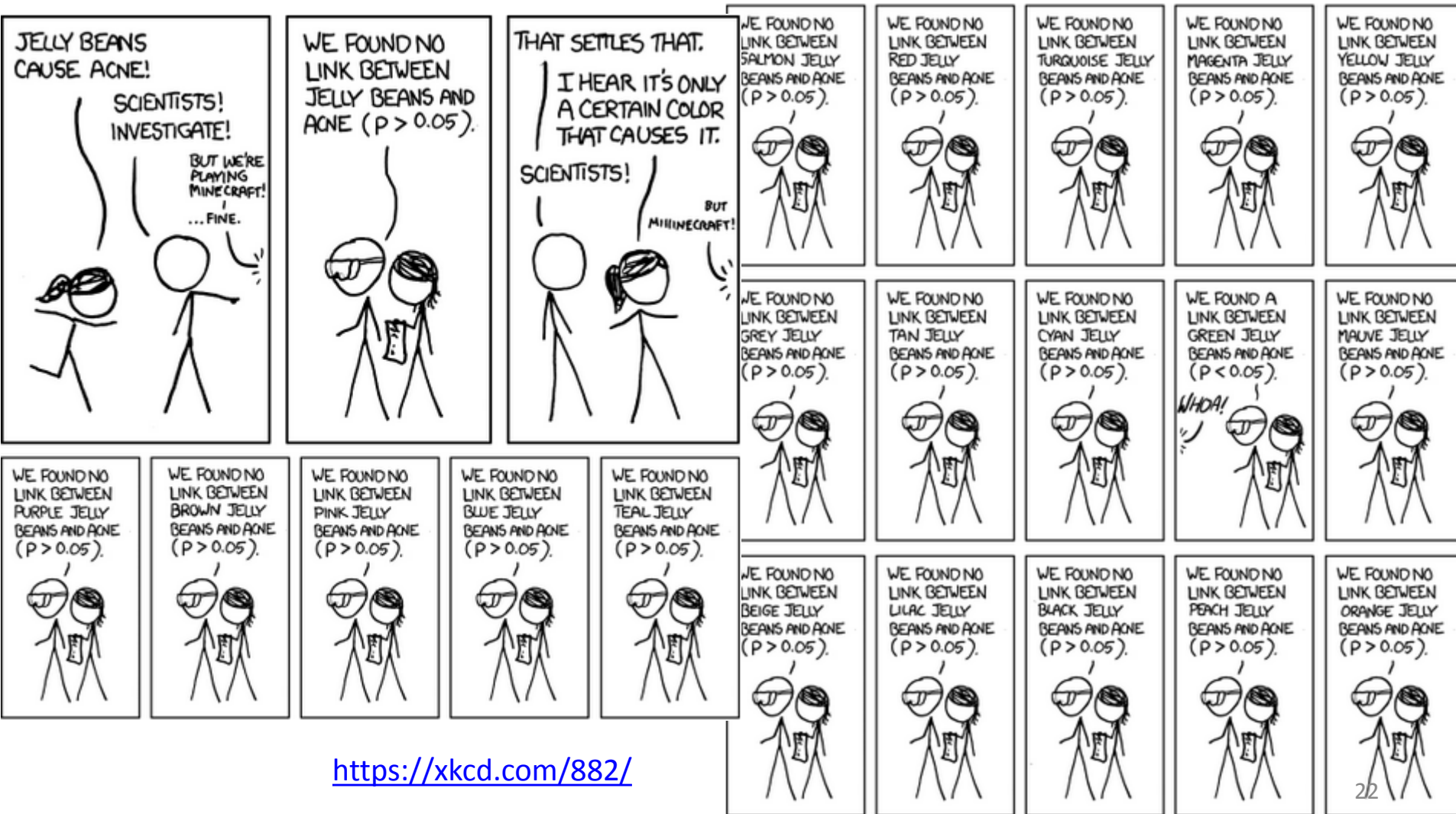
- Physical/social/psychological measurements have practical limits.
- A null hypothesis can be highly precise, and yet not falsifiable with the given technology.
- Moreover, background knowledge might tell us that the precision of the null is good enough. See also: Boson, Higgs.

Why *Not* Do It

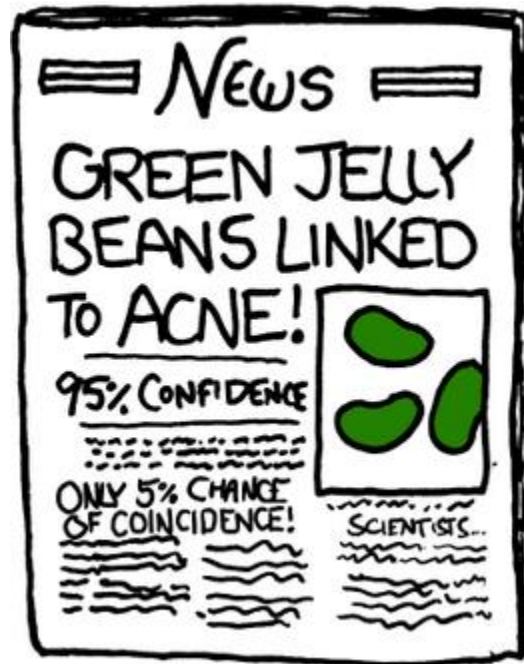
- Essentially, if you think you can get away without doing anything else.
- Hypothesis testing is hardly (or should be) convincing by itself. Effect size matters. Validation of assumptions/sample size may be just the starting point of a solid analysis.

Statistical significance \neq practical significance

The Sociology of Hypothesis Testing and p-hacking



The Sociology of Hypothesis Testing and p-hacking



The Sociology of Hypothesis Testing and p-hacking

- There are perverse incentives for “p-hacking”: making a selective reporting of p-values.
 - Multiple tests on a single data set are not independent. **And the minimum of a set is not going to be uniformly distributed.**
- Be responsible.

The Sociology of Hypothesis Testing and p-hacking

- Two different types of bad incentives:
 - “the null is bad”. If I’m proposing a new treatment and the null is a zero difference with respect to the old treatment. Down with the p-value!
 - “the null is good.” If I’m proposing a model and the null is “the model generated the data”. Up with the p-value!

Multiple Testing

- There is a considerable literature on multiple testing, which I will not cover.
- I will just mention one of the simplest techniques, the **Bonferroni correction**. It is motivated by $P(A \cup B) \leq P(A) + P(B)$.
- So if you have k hypotheses to test, think of A s and B s as the Type I error events.

Multiple Testing

- Without knowing the (complicated) joint distribution of these events, it suffices to control the level of the joint test by changing the level α of each individual test to α / k .
- It will control the level, but the probability of Type I error maybe much smaller than α . Bad power is likely to follow...

Take-Home Message

- Hypothesis testing: putting assumptions to test by deriving their consequences to the observed data.
- Despite its shortcomings, it is a common type of diagnostics. As long as we do not take them as the ultimate goal of an analysis (only in rare cases), they can be valuable.

Confidence Intervals

GENERAL CONCEPTS

Recall the NHANES Data

- Height data, where we found the estimated height expectation of 168 cm:

$$\hat{\mu} = 168$$

- We asked: what if **different** 19,219 individuals had been sampled?

Using Simulation to Understand Confidence Intervals

- Imagine the following experiment: let's generate **simulated data**, “God playing dice”, like we have been doing in our R examples.
- This is called a **(Monte Carlo) simulation**, for which there are very old computer algorithms based on **pseudo-randomness**.

Using Simulation to Understand Confidence Intervals

- Let's say we generate datasets of size 50, based on a $N(\mu = 168, \sigma^2 = 103)$ distribution.
- Let's do it over and over again, say 20 times, calculate the sample average each time.
- R demo

Using Simulation to Understand Confidence Intervals

- The sample average

$$\bar{X} \equiv \frac{1}{n} \sum_{i=1}^n X_i$$

is of course a random variable, since it is a function of the data.

- Importantly, it is a function only of the data, not of the unknown parameters of the model.
 - Recall the name: statistic!
- But its *distribution* should be a function the data distribution.

Recall (from Chapter 1)

- If

$$X \sim N(\mu, \sigma^2)$$

then

$$E[X] = \mu$$

- (You might want to check that yourself. Remember your calculus)

$$E[X] = \int x \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2} \right\} dx = \mu$$

In Particular

$$E[\bar{X}] = E \left[\frac{1}{n} \sum_{i=1}^n X_i \right] = \mu$$

- (check this, if you are interested. It follows from the definition of $E[\]$)
- That's why \bar{X} is a plausible estimator of μ . We typically denote estimators with hats, so in our case we chose $\hat{\mu} = \bar{X}$.

More Than This

- We can characterize the whole distribution of the sample average if each X_i is $N(\mu, \sigma^2)$:

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

- We can now use this to our advantage. Ask yourself: what is the distribution of

$$\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} ?$$

(Use the fact that the $\text{Var}(Y / c) = \text{Var}(Y) / c^2$ for any random variable Y and constant c , and linear manipulations of a Gaussian are also Gaussian-distributed)

Bounding μ

- It is a $N(0, 1)$. So for instance we can make claims such as (the cutoff below is arbitrary),

$$P\left(\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \leq 1.5\right) = P(Z \leq 1.5) \approx 0.93$$

where $Z \sim N(0, 1)$.

- So we can make the analogous claim

$$P\left(\mu \geq \bar{X} - 1.5\sigma/\sqrt{n}\right) \approx 0.93$$

Two IMPORTANT Observations

- In general, $\bar{X} - 1.5\sigma/\sqrt{n}$ is not a statistic!!
- Assume for now σ^2 is known, to simplify the argument (so the above IS a statistic in this case).
- Also: how to interpret this statement?

$$P(\mu \geq \bar{X} - 1.5\sigma/\sqrt{n}) \approx 0.93$$

The Key Point Concerning Confidence Intervals

- The randomness is not on μ !
- The randomness is in the data, which here is summarized by \bar{X} !
- What on Earth does it mean? After all, “I got my data already, it’s right here in front of me”.



Coverage

- Regardless of what μ is, if my sample size is n and my data follows a $N(\mu, \sigma^2)$, then in the limit of infinite repetitions of my dataset, the interval

$$[\bar{X} - 1.5\sigma/\sqrt{n}, +\infty)$$

will contain μ (approximately) 93% of the time.

- Another way of saying this: the **coverage** of this interval is 93%.

Lessons

- As I've mentioned before, nobody expects you to collect “infinitely many datasets” for a same problem.
- What this means is the following: if you provide a (say) 93% confidence interval for your quantity of interest in each problem you work on through your career, then in the long run the intervals you provided will contain the quantity of interest 93% of the time. You will earn the title “Mr/Ms/Mrs 93%”.
- **You cannot know (without further data) for which intervals you got it right, just the long-run performance!**

In Practice

- There will be several assumptions in your model that will be violated, so coverage will not be exact.
- Despite being an idealization, reporting confidence intervals is of major importance in many applications.
 - **At the very least as a way of being more humble about which conclusions you can draw.**
- Just because some modern models have difficult-to-interpret parameters (e.g., neural nets), it doesn't mean confidence intervals will not be used at some point in your application (e.g., in the estimation the empirical performance of a neural net).

In Practice

- When we get to other topics (such as regression), we will talk again about confidence intervals in those contexts.

Another Interval

- It is much more common to report lower and upper bounds.
- Going back to our example. Say you want a 95% interval (the default in much software). This is typically done by finding the 2.5% and 97.5% quantiles of the distribution of your statistic. For instance,

$$[\bar{X} + z_{0.025}\sigma/\sqrt{n}, \bar{X} + z_{0.975}\sigma/\sqrt{n}]$$

2.5% quantile of $N(0, 1)$

97.5% quantile of $N(0, 1)$

A “Real” Statistic for the Gaussian Mean CI

- I owe you that. Start from:

$$\frac{\bar{X} - \mu}{\sqrt{S^2/n}} \sim \mathcal{T}(n - 1)$$

where S^2 is the sample variance:

$$S^2 \equiv \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

- Exercise: can you derive a 95% confidence interval using this?

A “Real” Statistic for the Gaussian Mean CI

- Notice: for large n , which we are assuming anyway, the following is used in practice
 - (the funny double-tilde is just an informal notation for “approximately distributed”)

$$\frac{\bar{X} - \mu}{\sqrt{S^2/n}} \approx N(0, 1)$$

Confidence Intervals

APPROXIMATIONS: CLT AND THE BOOTSTRAP

Those Bounds Again

- Say you want to trap a parameter of interest θ with coverage probability c . This is the general template of a confidence interval:

$$P(lower_c(\mathbf{X}) \leq \theta \leq upper_c(\mathbf{X})) = c$$

so depending on your choice of c , you will get (random) lower and upper bounds that depend on data \mathbf{X} .

- In general, it is not at all easy to find the actual distribution of $lower_c(\mathbf{X})$ and $upper_c(\mathbf{X})$!
 - It would be bonkers to assume Gaussianity in general. However, the good old Gaussian is useful in a different way.

Help me, Central Limit Theorem!

- [You may have seen this coming.]
- Many lower/upper functions depend on averages.
- We have seen what happens to averages for large n , correct?

Practical Advice

- Many confidence intervals in software packages rely on the Central Limit Theorem under the hood.
- It is not always obvious what a “large sample” is. The theory is about **asymptotics**. Sometimes checks can be done.

A Missing Ingredient

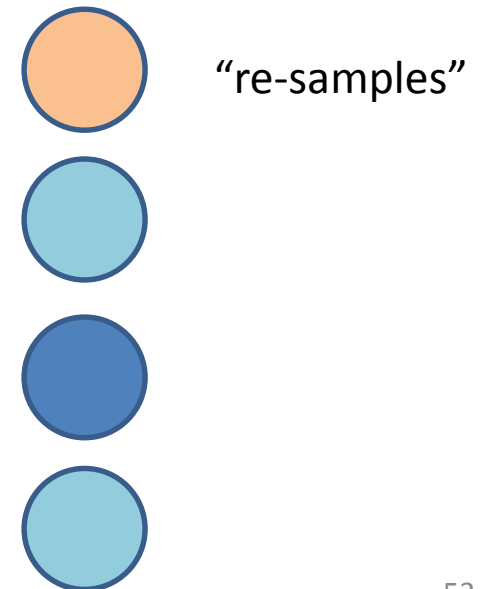
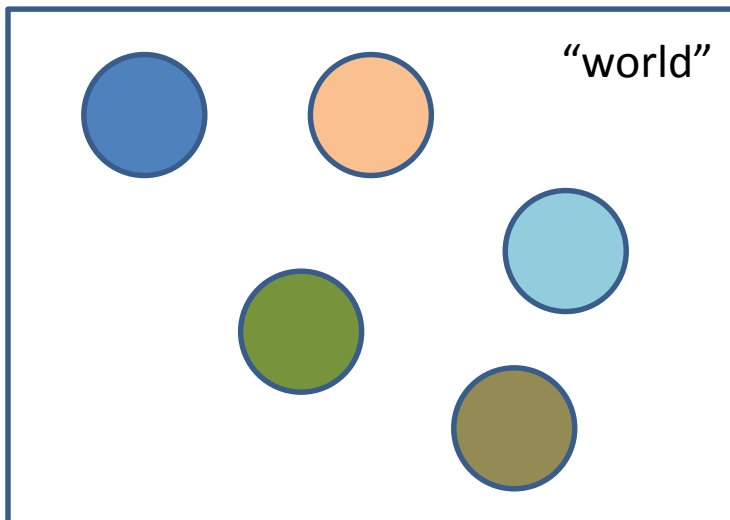
- Normal approximations for averages are fine in many cases. But what is the variance of your statistic??
- Sometimes this can be written down in a simple way. In many cases, not at all.
- When good old fashioned algebra fails, we should resort to computer-intensive alternatives. **Enter the bootstrap.**

The Bootstrap

- “The world” generated your data.
- **What if *your data* was the world?** What does it mean by “your data generating data”?

The Basic Idea

- Think of your sample as if it was the “world”, the **population**.
- A box which we can, now, *play with by sampling from it*.



Sampling with Replacement, and the Size of the Sample

- This is **sampling with replacement**: choose a data point, add to the “re-sample”, but “put it back in the box”.
- We do this to generate a re-sample of the same size as the original sample. The idea is to mimic the same process that generated your data, down to the sample size.

The Bootstrap

- So, we are “using the data we have to generate data.” This will inform us about sampling variability.



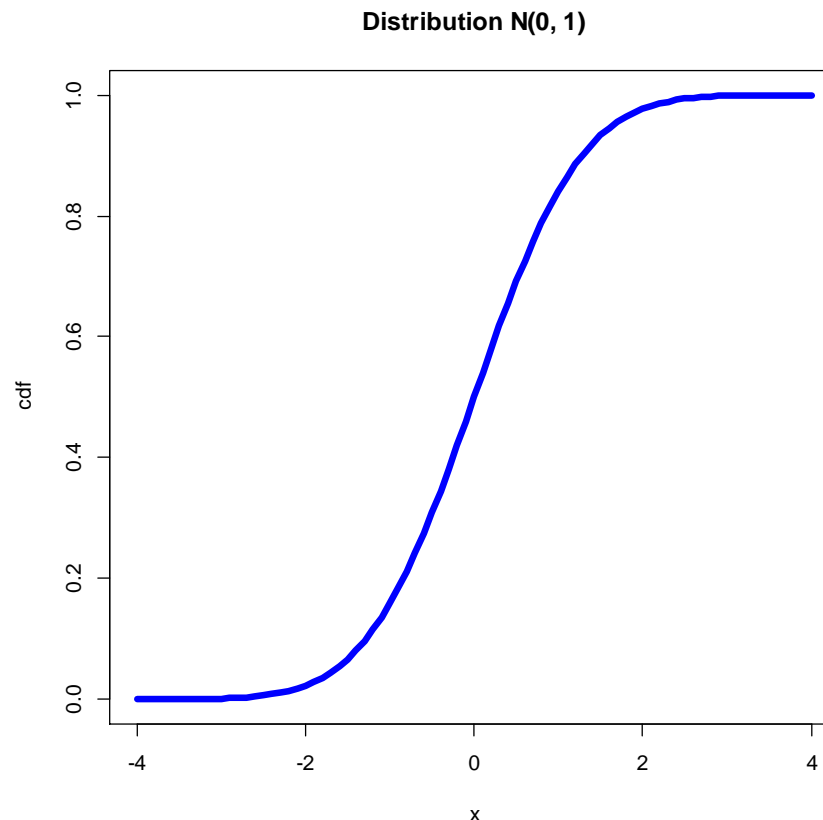
Baron Munchausen (Wikimedia Commons)

Why?

- Because we can now do it many many times to simulate in a computer the idea of “infinite replicates of a dataset”.
 - Ideally, we would do infinitely many replicates. In practice, we choose a large number and accept that we will get some approximation error.
- We can then see how our statistic varies across these synthetic replicates.

An Intuition of Why This Works

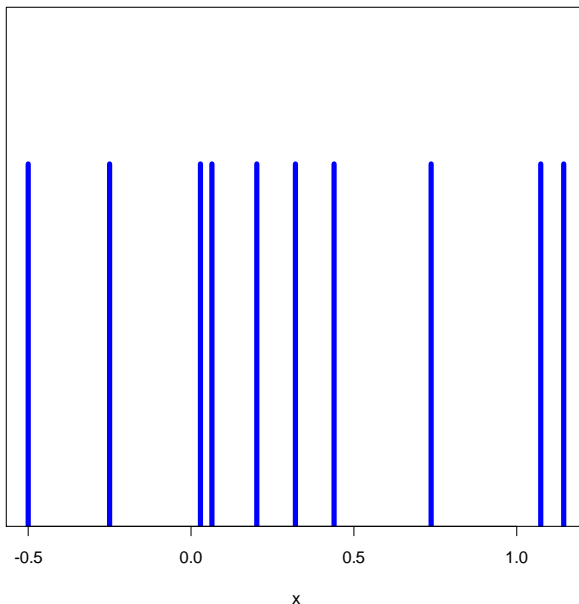
- Recall our friend, the cumulative distribution function $F(x) \equiv P(X \leq x)$. For instance,



The Empirical CDF

- Thought experiment: the data as our population. What would be the cdf when new data should be only at particular locations, with equal probability?

Samples from a $N(0, 1)$



Each stick is placed at a particular draw from a $N(0, 1)$.

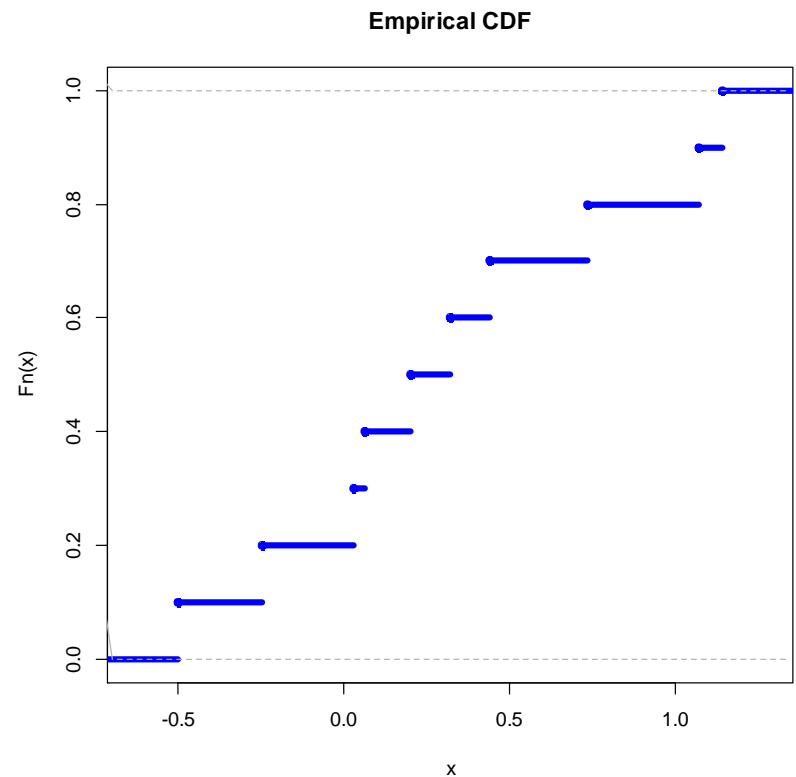
Now imagine you are generating x values only at the locations previously chosen by our draws.

The Empirical CDF

- For any particular level x , just count the frequency of points no larger than x .

$$\hat{F}_n(x) \equiv \frac{\text{\#data points no greater than } x}{\text{\#data points (i.e., } n\text{)}}$$

$$\hat{F}_n(x) \equiv \frac{1}{n} \sum_{i=1}^n I(x^{(i)} \leq x)$$




In the Limit

- Analogous to the **law of large numbers** (averages converge to means as sample size increases), empirical CDFs converge to the population cdfs (R demo).
- So, informally, we can say that the empirical distribution (what we get by re-sampling) carry some information about sampling according to the true cdf.

Using the Bootstrap

- The simplest way is to use it to calculate the variance of the statistic of interest along with the CLT approximation.
- Going back to our UK Gas consumption, let's create a 95% confidence interval for the mean.

standard error (“deviation”)
obtained by bootstrap



$$[\bar{X} + z_{0.025}\hat{se}_{boot}, \bar{X} + z_{0.975}\hat{se}_{boot}]$$

The Algorithm

1. Draw $X^{(1)\star}, \dots, X^{(n)\star} \sim \hat{F}_n$
2. Compute \bar{X}_n^\star by averaging $X_1^\star, \dots, X_n^\star$
3. Repeat steps 1 and 2, B times, to get $\bar{X}_{n,1}^\star, \dots, \bar{X}_{n,B}^\star$
4. Let

$$s.e.\text{boot} \equiv \sqrt{\frac{1}{B} \sum_{b=1}^B \left(\bar{X}_{n,b}^\star - \frac{1}{B} \sum_{r=1}^B \bar{X}_{n,r}^\star \right)^2}$$

[Recall the empirical variance definition: $\frac{1}{n} \sum_{i=1}^n (X^{(i)} - \bar{X})^2$]

$B?$

- Ideally, we would average over *every possible re-sample*, but this is not in general doable.
- So this is an approximation, which itself is justifiable by the law of the large numbers! Some packages might choose B for you.
- R demo

Wait

- The empirical mean is simple enough that we can find a decent approximation for its variance in the literature. There wasn't really a need for the bootstrap here.
- The bootstrap shines when this is not the case. In the UK energy example, we might be interested in a confidence interval for the *median* (can you guess why?).

Rephrasing It

- The idea is exactly the same. Instead of the sample average, our statistic is the sample median. Just think in terms of abstract T s.

1. Draw $X^{(1)\star}, \dots, X^{(n)\star} \sim \hat{F}_n$
2. Compute T_n^\star from $X_1^\star, \dots, X_n^\star$ according to its definition
3. Repeat steps 1 and 2, B times, to get $T_{n,1}^\star, \dots, T_{n,B}^\star$
4. Let

$$s.e.boot \equiv \sqrt{\frac{1}{B} \sum_{b=1}^B \left(T_{n,b}^\star - \frac{1}{B} \sum_{r=1}^B T_{n,r}^\star \right)^2}$$

Rephrasing It

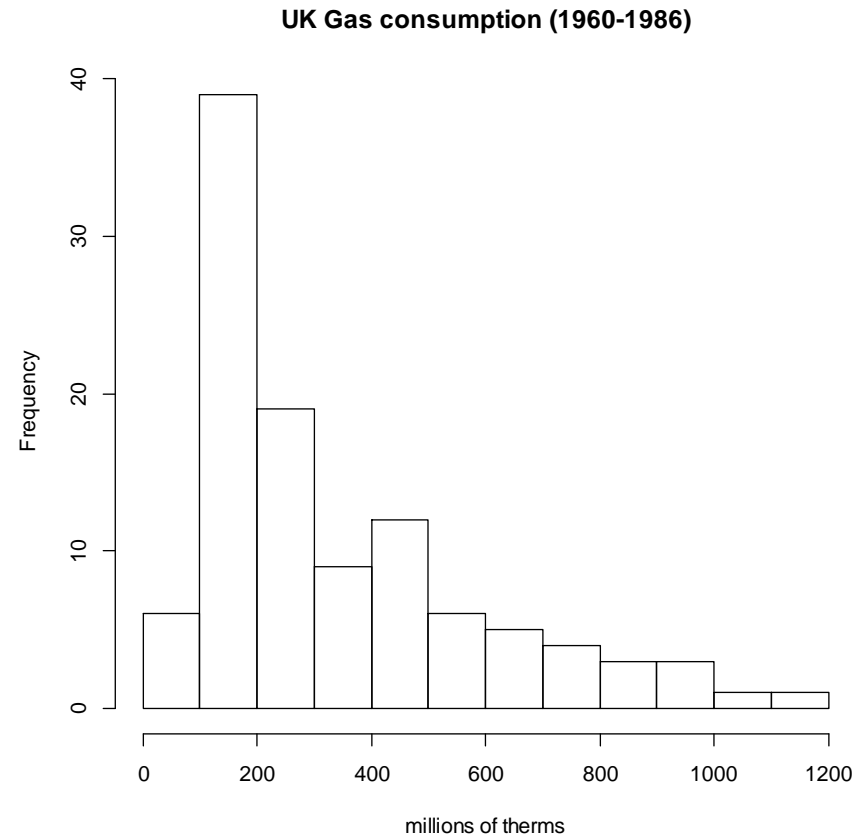
- But this only gives variances. What if we want a more precise way of building confidence intervals without using the Normal approximation?

Confidence Intervals

THE STORY SO FAR

Recap

- From data like this, we would like **to learn a property of the population**, like the mean or median.
- Preferably, **not only a single point**, but a **set** (an **interval**, more precisely) that will **contain the true value with high probability**.



Recap

- For instance, to learn about the mean μ , we can make use of the following claim

$$\bar{X} \approx N(\mu, S^2/n)$$

where n is the sample size and S^2 is the sample variance $S^2 \equiv \frac{1}{n} \sum_{i=1}^n (X^{(i)} - \bar{X})^2$.

- Why is this useful?

Recap

1. We can rewrite it like this $\frac{\bar{X} - \mu}{\sqrt{S^2/n}} \approx N(0, 1)$
2. Then find the (say) 0.025 and 0.975 quantiles of this known distribution to claim that
$$P\left(-1.96 \leq \frac{\bar{X} - \mu}{\sqrt{S^2/n}} \leq 1.96\right) = 0.95,$$
3. And then re-express it in a way to emphasize μ :
$$P\left(\bar{X} - 1.96\sqrt{S^2/n} \leq \mu \leq \bar{X} + 1.96\sqrt{S^2/n}\right) = 0.95,$$

Note

- We actually have a name for quantities like this:

$$\frac{\bar{X} - \mu}{\sqrt{S^2/n}}$$

- This is called a **pivot**, a function of the parameter of interest which has a known distribution.
 - Notice a pivot is not a statistic by construction!
 - **In general, hard to find!**

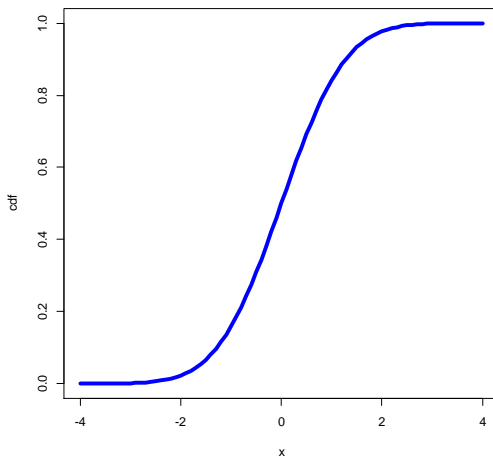
Recap

- For instance: how to get a confidence interval for the median? Which pivot to use?

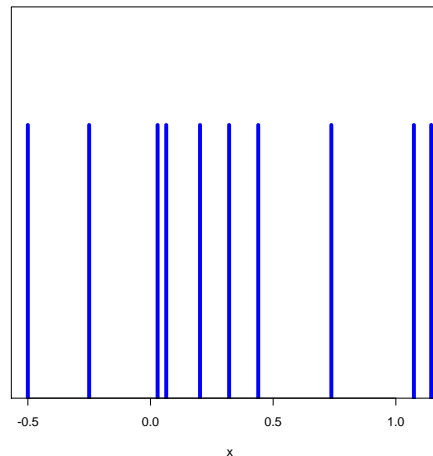
Recap

- Enter the bootstrap: a general trick that uses the **empirical distribution**.
 - We saw it to calculate variances. Now let's see how to build pivots with it.

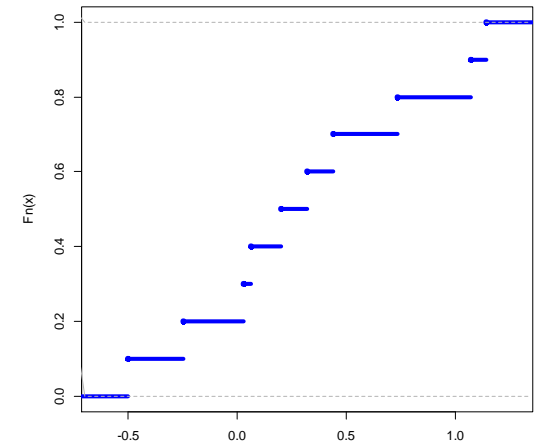
Distribution $N(0, 1)$



Samples from a $N(0, 1)$



Empirical CDF

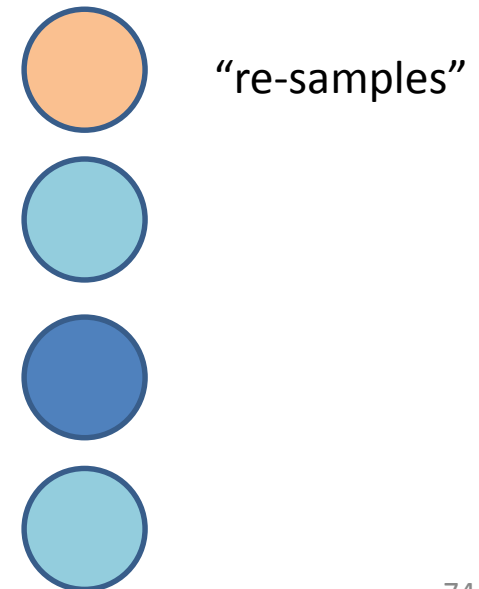
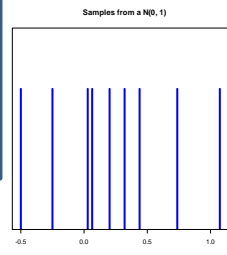
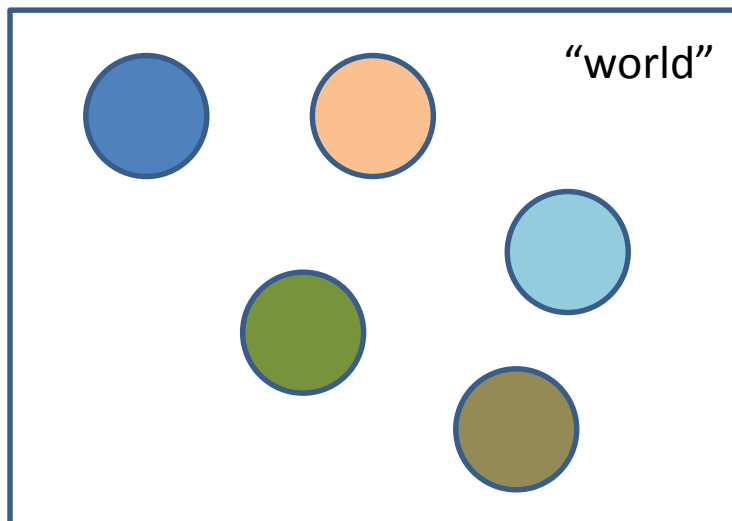


$$\hat{F}_n(x) \equiv \frac{\text{\#data points no greater than } x}{\text{\#data points (i.e., } n\text{)}}$$

$$\hat{F}_n(x) \equiv \frac{1}{n} \sum_{i=1}^n I(x^{(i)} \leq x)$$

Recap

- Basic idea: think of your sample as if it was the “world”, the **population**.
- A box which we can, now, *play with by sampling from it*.



The Bootstrap Pivotal Interval

- The Normal approximation may not be good, as in estimating the median.
- One of the most used bootstrap variants of confidence intervals is the **pivotal interval**.
- The problem: find confidence interval for parameter θ based on an estimator $\hat{\theta}_n$ built from a sample of size n .
 - Notice the explicit subscript n .

Alternative Use

- Essentially, use bootstrap to estimate the distribution of $\hat{\theta}_n - \theta$, **which we will be our pivot**. Then find its quantiles of interest.

Idea

- Let $H(r)$ be the cdf of the pivot, that is

$$H(r) \equiv P(\hat{\theta}_n - \theta \leq r)$$

- Define quantiles such that we get coverage $1 - \alpha$:

$$P(a(\hat{\theta}_n) \leq \theta \leq b(\hat{\theta}_n)) = 1 - \alpha$$

- I won't show to you, but the following satisfies the above:

$$a(\hat{\theta}_n) = \hat{\theta}_n - H^{-1}(1 - \alpha/2)$$

$$b(\hat{\theta}_n) = \hat{\theta}_n - H^{-1}(\alpha/2)$$

The Problem

- What are the distributions of these little monstrosities? Hard to find!

$$\begin{aligned}a(\hat{\theta}_n) &= \hat{\theta}_n - H^{-1}(1 - \alpha/2) \\ b(\hat{\theta}_n) &= \hat{\theta}_n - H^{-1}(\alpha/2)\end{aligned}$$

- Solution: bootstrap 'em. First, generate B resampled datasets, with the respective **bootstrapped pivots** $R_{n,b}^*$, replacing $\hat{\theta}_n - \theta$!

This replaces $\hat{\theta}_n$ And this replaces θ !

$$R_{n,b}^* \equiv \theta_{n,b}^* - \hat{\theta}_n$$

Final Trick

- Use the distribution of the bootstrap samples to get an estimated $\hat{H}(r)$.

$$\begin{aligned}a(\hat{\theta}_n) &= \hat{\theta}_n - \hat{H}^{-1}(1 - \alpha/2) \\ b(\hat{\theta}_n) &= \hat{\theta}_n - \hat{H}^{-1}(\alpha/2)\end{aligned}$$

- What is $\hat{H}^{-1}(\alpha)$? It's just the empirical quantile of the bootstrap distribution. Sort all of your B bootstrap pivots, pick the one in position $B\alpha$, or the closest integer.

That Is

- To get e.g. $b(\hat{\theta}_n) = \hat{\theta}_n - \hat{H}^{-1}(\alpha/2)$
 - Find $\hat{H}^{-1}(\alpha/2) = \theta_{n, B\alpha/2}^* - \hat{\theta}_n$
 - Set $b(\hat{\theta}_n) = 2\hat{\theta}_n - \theta_{n, B\alpha/2}^*$
 - Note: $B\alpha / 2$ is rounded
- R demo

Take-Home Messages

- Estimation is important, but so is the assessment of your uncertainty.
 - In your career as Data Scientist, you will be pressed for easy answers. Don't fall for that.
- Confidence intervals provide coverage: an interval which traps the parameter of interest, regardless of its true value, with the advertised probability.

Take-Home Messages

- However, everything is predicted on given assumptions. Including the bootstrap. Don't ever forget that.
 - You should verify them as best as you can.
 - Again, **the game is about being “less wrong”, it is not about being infallible.**
- If the intervals look wide and uninformative, even with large sample sizes: tough luck. **Information doesn't come for free.** If you want more certainty, you need more assumptions (or more data).

Take-Home Messages

Hypothesis testing and confidence intervals are two core building blocks of statistical inference.

They will show up as again as needed, when we study regression and modelling during the rest of this course.