Review: Sampling Distributions and Confidence Intervals

Review: Sampling Distributions and Confidence Intervals

-

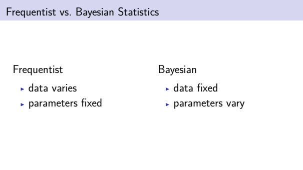
Frequentist vs. Bayesian Statistics

Frequentist

- data varies
- parameters fixed

Bayesian

- data fixed
- parameters vary



Frequentist vs. Bayesian Statistics

Frequentist

- data varies
- parameters fixed

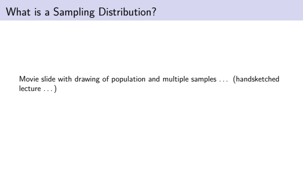
Bayesian

- data fixed
- parameters vary

- two different world views in stats
- Frequentist world view is traditional and we'll use it throughout this class
  - sampling is a repeatable process in which the data varies but the pop. parameters are fixed
  - observed sample is one of infinitely many possible samples . . . this is where sampling distributions come in.
- Bayesian world view
  - the sample data is fixed
  - the values of the parameters vary and are described probabilistically
- the merits of each approach are the subject of ongoing debate
- while this class is about frequentist statistics, the bayesian view is gaining traction in data science

What is a Sampling Distribution?

Movie slide with drawing of population and multiple samples . . . (handsketched lecture . . .)



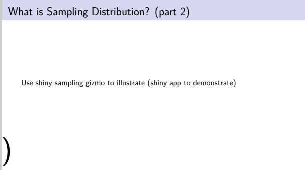
What is a Sampling Distribution?

Movie slide with drawing of population and multiple samples . . . (handsketched lecture . . .)

- In traditional statistics, the observed sample is one of infinitely many possible samples
- Different samples will, of course, yield different values for any computed statistics
- To understand the variation and uncertainty of a statistic we study the sampling distribution of that statistic
- What is a sampling distribution? It works like this . . .

What is Sampling Distribution? (part 2)

Use shiny sampling gizmo to illustrate (shiny app to demonstrate)



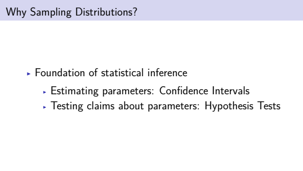
What is Sampling Distribution? (part 2)

Use shiny sampling gizmo to illustrate (shiny app to demonstrate)

- start with means, move to median, etc.
- when a statistician asks "how variable is your statistic?", they want to know how spread out the sampling distribution is
- in particular, they want to know the standard error of the statistic, the standard error is the standard deviation of the sampling distribution of a statistic
- visit `https://github.com/DataScienceUWL/samplingShiny` to see how to run the app yourself.

Why Sampling Distributions?

- Foundation of statistical inference
  - Estimating parameters: Confidence Intervals
  - Testing claims about parameters: Hypothesis Tests



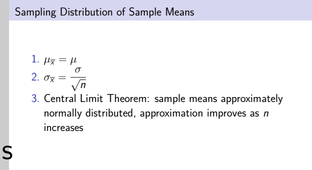
Why Sampling Distributions?

- Foundation of statistical inference
  - Estimating parameters: Confidence Intervals
  - Testing claims about parameters: Hypothesis Tests

- to estimate or test claims about population parameters we account for uncertainty by using a sampling distribution
- often we can predict properties of the sampling distribution without actually choosing or simulating a large number of samples
- if we don't have a model for the sampling distribution we can simulate the sampling distribution through a process called bootstrapping

Sampling Distribution of Sample Means

1. $\mu_{\overline{x}} = \mu$
2. $\sigma_{\overline{x}} = \dfrac{\sigma}{\sqrt{n}}$
3. Central Limit Theorem: sample means approximately normally distributed, approximation improves as *n* increases



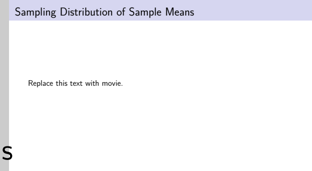
Sampling Distribution of Sample Means

1. $\mu_{\overline{x}} = \mu$
2. $\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$
3. Central Limit Theorem: sample means approximately normally distributed, approximation improves as *n* increases

- Every statistic has a sampling distribution, but for the next few slides we'll focus on sample means since they come up so often.
- We know a lot about the sampling distribution of sample means without actually having to sample data
- One. The mean of all the sample means is the same as the original population mean.
- Two. The standard deviation of the sample means, also known as the standard error of the mean is given by $\sigma/\sqrt{n}$ which shows the variability of sample means decreases as n increases
- Three. The central limit theorem guarantees that sample means will always be approximately normally distributed for large enough samples. How large is large enough depends on the distribution of the random variable being sampled.

Sampling Distribution of Sample Means

Replace this text with movie.

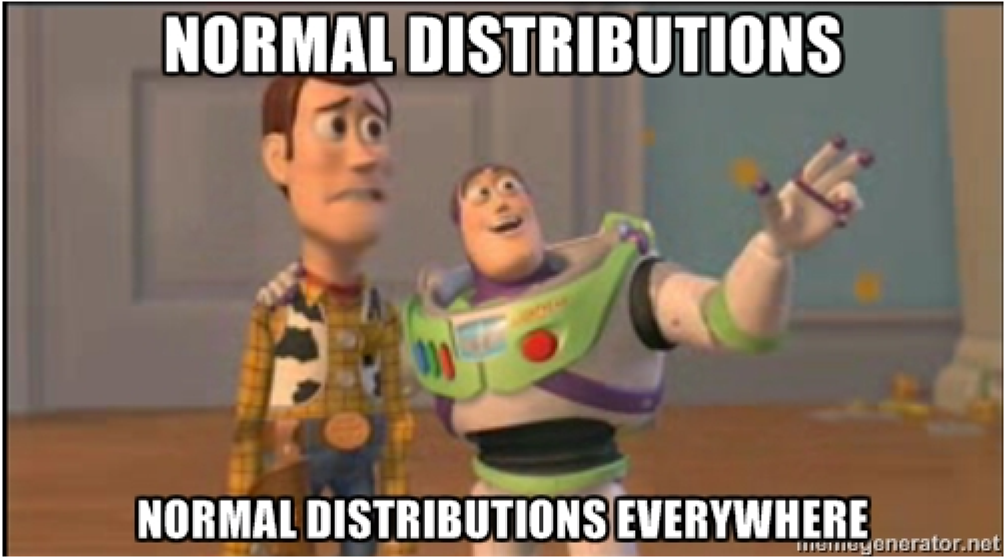


Sampling Distribution of Sample Means

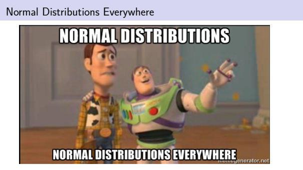
Replace this text with movie.

- start with normal parent, show center, spread and shape
- go to log normal, show center, spread and shape, note that n needs to be fairly large for good normal approx
- HEAVY tails repeat, show center, spread, and shape, note that n needs to be really large for good normal approx

Normal Distributions Everywhere



└Normal Distributions Everywhere


Normal Distributions Everywhere

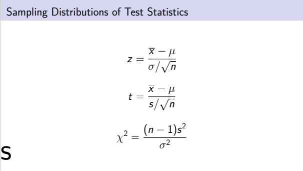
- as an aside, the central limit theorem is part of the reason normal distributions are everywhere
- Loosely, the central limit theorem tells us that any random variable which is the sum of many independent random variables that have similar distribution will be approximately normally distributed
- heights, test scores, blood pressures, IQ scores all have approximately normal distributions
- a test score for instance is the sum of scores on multiple questions . . .

Sampling Distributions of Test Statistics

$$z = \frac{\overline{x} - \mu}{\sigma / \sqrt{n}}$$

$$t = \frac{\overline{x} - \mu}{s / \sqrt{n}}$$

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$



Sampling Distributions of Test Statistics

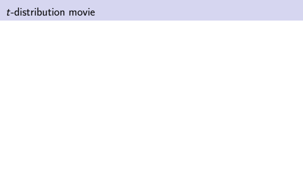
$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$$

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

- for statistical inference the sampling distribution of test statistics are especially important
- a test statistic is a special statistic that is used on estimating and testing population parameters
- for now, we'll focus on the t test statistic

t-distribution movie

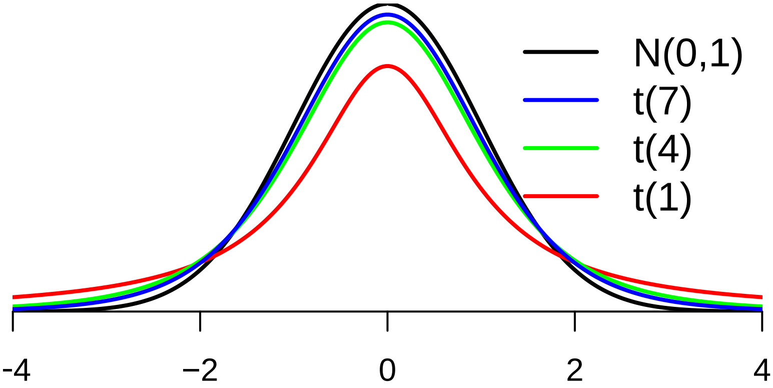


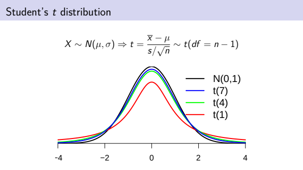
t-distribution movie

- use shiny app to demonstrate t-distribution
- show that normal leads to t
- the more not normal the population is, the further from t

Student's t distribution

$$X \sim N(\mu, \sigma) \Rightarrow t = \frac{\overline{x} - \mu}{s/\sqrt{n}} \sim t(df = n - 1)$$





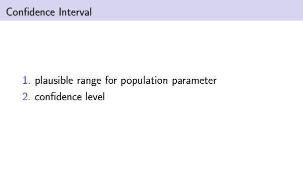
Student's *t* distribution

$$X \sim N(\mu, \sigma) \Rightarrow t = \frac{\overline{x} - \mu}{s/\sqrt{n}} \sim t(df = n - 1)$$

- in 1908 William Gosset published a paper deriving the t-distribution under the pseudonym "Student", hence the name
- also called simply a t-distribution
- each t-distribution is characterized by its degrees of freedom
- as the degrees of freedom increases the distribution approaches the standard normal
- we will see, shortly, how to use the t-distribution to get a confidence interval for the population mean

Confidence Interval

1. plausible range for population parameter
2. confidence level



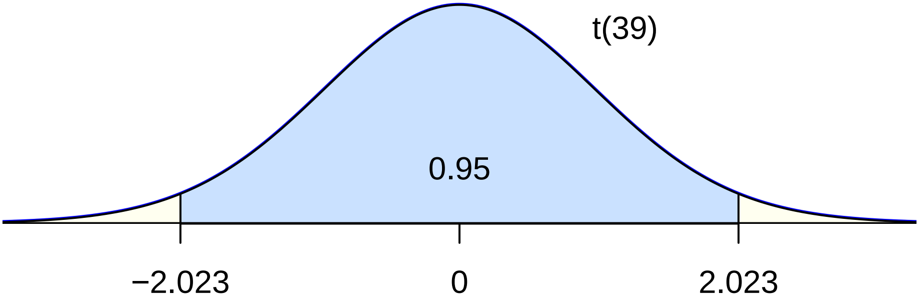
Confidence Interval

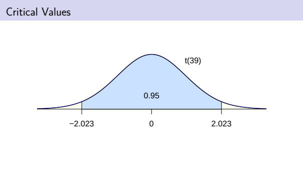
1. plausible range for population parameter
2. confidence level

- a confidence interval estimates a population parameter, such as the mean, while accounting for uncertainty
- it has two elements

1. a plausible range of values for the population parameter, think of it as a fuzzy estimate of the paramter value
2. a confidence level that expresses our belief in the estimate, the larger the level, the higher our belief that the range captures the true value (but also the wider the range)

Critical Values





Critical Values

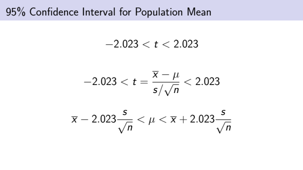
- this is a picture of the *t* distribution with 39 degrees of freedom
- the middle 95% is shaded and the t values, -2.023 and 2.023, that carve out that 95% are called critical values
- this distribution and the critical values are the basis for constructing a 95% CI for the population mean

95% Confidence Interval for Population Mean

$$-2.023 < t < 2.023$$

$$-2.023 < t = \frac{\overline{x} - \mu}{s/\sqrt{n}} < 2.023$$

$$\overline{x} - 2.023\frac{s}{\sqrt{n}} < \mu < \overline{x} + 2.023\frac{s}{\sqrt{n}}$$



95% Confidence Interval for Population Mean

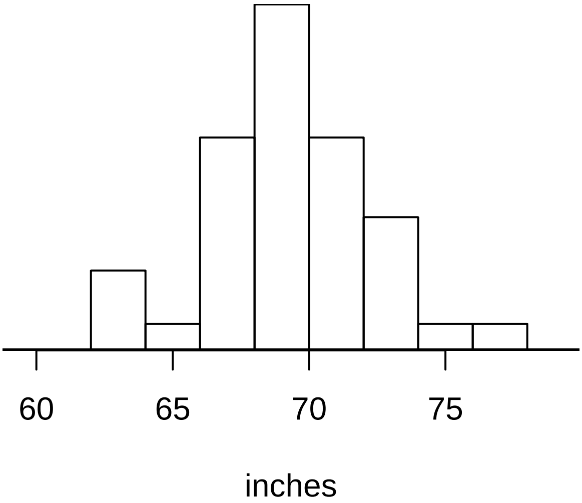
$$-2.023 < t < 2.023$$

$$-2.023 < t = \frac{\overline{x} - \mu}{s/\sqrt{n}} < 2.023$$

$$\overline{x} - 2.023\frac{s}{\sqrt{n}} < \mu < \overline{x} + 2.023\frac{s}{\sqrt{n}}$$

- (Line 1) we just saw on the last slide that for df=39, there is a 95% chance t will be between -2.023 and 2.023
- (Line 2) if the sample data comes from a normally distributed random variable, then the t-test statistic follows a t-distribution with df = n - 1 . . .
- (Line 2 cont) so 95% of samples of size 40 from a normally distributed random variable will give t - test statistics in that range
- (Line 3) solving the inequality for μ gives the last row . . . 95% of samples yield intervals that capture μ
- this is a 95% CI. A larger confidence level like 99% results in a wider interval.

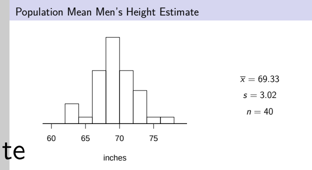
Population Mean Men's Height Estimate



$\overline{x} = 69.33$

$s = 3.02$

$n = 40$



Population Mean Men's Height Estimate

$\overline{x} = 69.33$

$s = 3.02$

$n = 40$

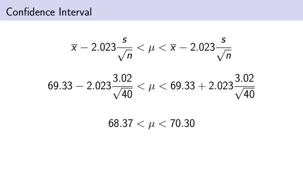
- based on our sample

Confidence Interval

$$\overline{x} - 2.023\frac{s}{\sqrt{n}} < \mu < \overline{x} - 2.023\frac{s}{\sqrt{n}}$$

$$69.33 - 2.023\frac{3.02}{\sqrt{40}} < \mu < 69.33 + 2.023\frac{3.02}{\sqrt{40}}$$

$$68.37 < \mu < 70.30$$



Confidence Interval

$$\overline{x} - 2.023\frac{s}{\sqrt{n}} < \mu < \overline{x} - 2.023\frac{s}{\sqrt{n}}$$

$$69.33 - 2.023\frac{3.02}{\sqrt{40}} < \mu < 69.33 + 2.023\frac{3.02}{\sqrt{40}}$$

$$68.37 < \mu < 70.30$$

- once we have a sample and we believe that it came from an approximately normal distribution, then we can build the interval

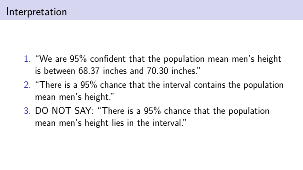
Confidence Interval in R

```r
t.test( h, conf.level = 0.95)$conf.int
```

```
## [1] 68.3693 70.3007
## attr(,"conf.level")
## [1] 0.95
```

Interpretation

1. “We are 95% confident that the population mean men’s height is between 68.37 inches and 70.30 inches.”
2. “There is a 95% chance that the interval contains the population mean men’s height.”
3. DO NOT SAY: “There is a 95% chance that the population mean men’s height lies in the interval.”



Interpretation

1. "We are 95% confident that the population mean men's height is between 68.37 inches and 70.30 inches."
2. "There is a 95% chance that the interval contains the population mean men's height."
3. DO NOT SAY: "There is a 95% chance that the population mean men's height lies in the interval."

- The second and third statements seem identical, but the second one implies that the interval changes and 95% of samples give intervals that work.
- However, the third one implies that the population mean changes and there is a 95% chance it lands in the range. - In the frequentist world view, parameters do not change and do not have associated probabilites
- `http://www.graphpad.com/guides/prism/6/statistics/index.htm?stat_more_about_confidence_interval.htm` is a nice site discussing the subtleties of interpreting confidence intervals.
- BTW, the population mean height of men in the US is about 70 inches. This is one of the 95% of intervals that contain the population mean.

CI's movie

Use rossman chance applet to record a movie illustrating what confidence means, similar information in OTT but this is nicer