Week 6: Bootstrapping Confidence Intervals

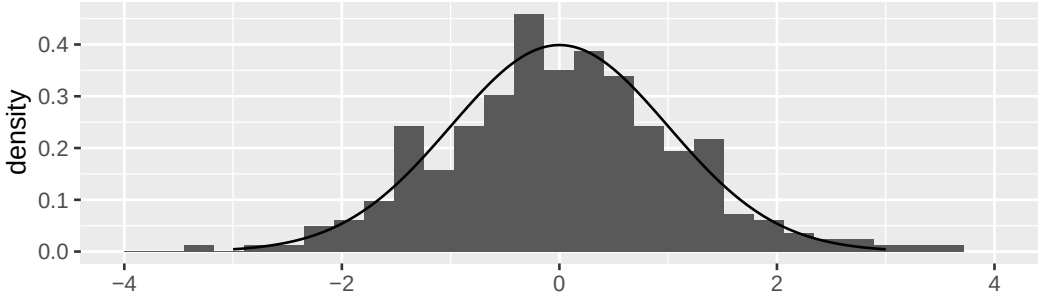
AKA Estimation, a new kind of Statistical Inference Tool

Scott Schwartz

October 17, 2021

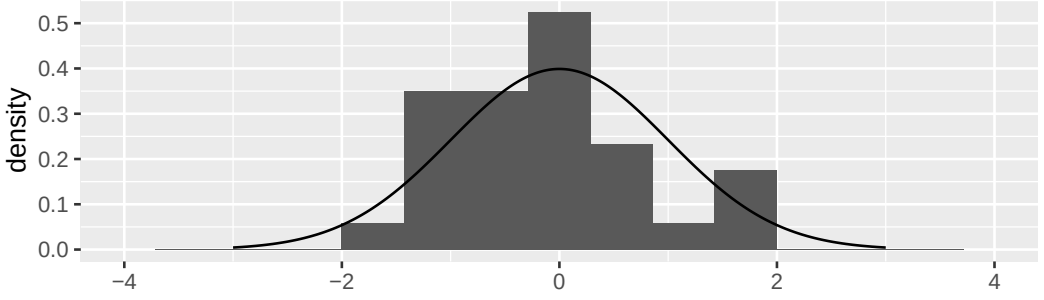
Samples Approximate Populations

```
library(tidyverse); set.seed(130); n <- 300; support <- seq(-3,3,.01)
normal_pdf <- geom_line(data = tibble(`normal pdf`=dnorm(support), x=support),
                        aes(x=x, y=`normal pdf`))
normal_sample <- geom_histogram(data = tibble(x=rnorm(n=n)),
                                aes(x=x, y=..density..), bins=30)
# https://r-charts.com/distribution/histogram-density-ggplot2/
# https://stackoverflow.com/questions/16712800/overlay-lines-and-hist-with-ggplot2
ggplot() + normal_sample + normal_pdf + xlim(-4,4) # {r,fig.width=6, fig.height=2}
```



Samples Approximate Populations

```
set.seed(130); n <- 30
normal_pdf <- geom_line(data = tibble(`normal pdf`=dnorm(support), x=support),
                        aes(x=x, y=`normal pdf`))
normal_sample <- geom_histogram(data = tibble(x=rnorm(n=n)),
                                aes(x=x, y=..density..), bins=15)
# https://r-charts.com/distribution/histogram-density-ggplot2/
# https://stackoverflow.com/questions/16712800/overlay-lines-and-hist-with-ggplot2
ggplot() + normal_sample + normal_pdf + xlim(-4,4) # {r,fig.width=6, fig.height=2}
```



How Well Do Samples Approximate Populations?

Clearly samples can't totally approximate populations

but if we're using them to learn population **parameters** it may be "sufficient"...

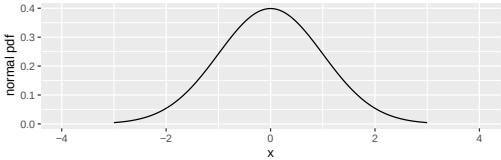
$$\text{(statistic) } \bar{x} \text{ approximates (parameter) } p$$

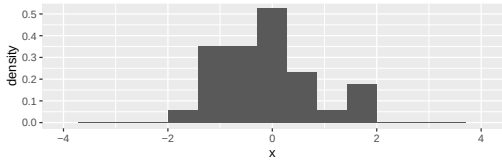
if x_i is only either 0 or 1 (and $p = Pr(x_i = 1)$)

$$\text{(statistic) } \bar{x} \text{ approximates (parameter) } \mu$$

otherwise (and $\mu = E[x_i]$)

How Well Do Samples Approximate Populations?






```r
set.seed(130); n <- 30; x <- rnorm(n=n); N <- 1000#0000000?
population_sample_means <- 1:N; bootstrap_sample_means <- 1:N
set.seed(130); for(i in 1:N){
  population_sample_means[i] <- mean(rnorm(n=n))
  bootstrap_sample_means[i] <- mean(sample(x, prob=rep(1/n,n), size=n, replace=TRUE))
} # Does Bootstrap Approximation Work?
```

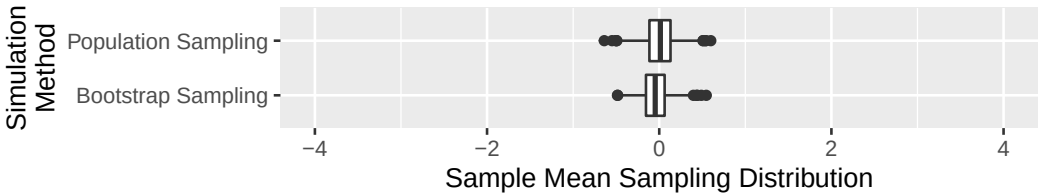
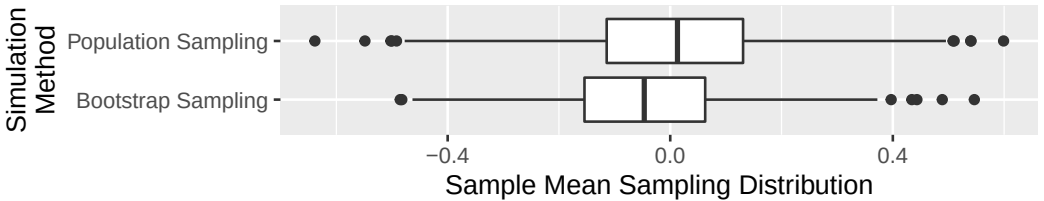
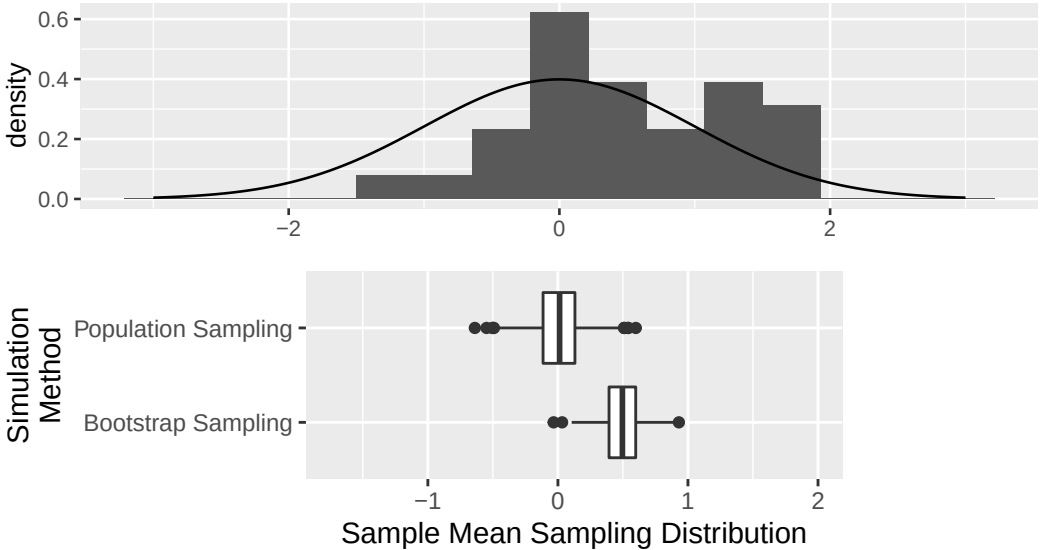


Figure Code

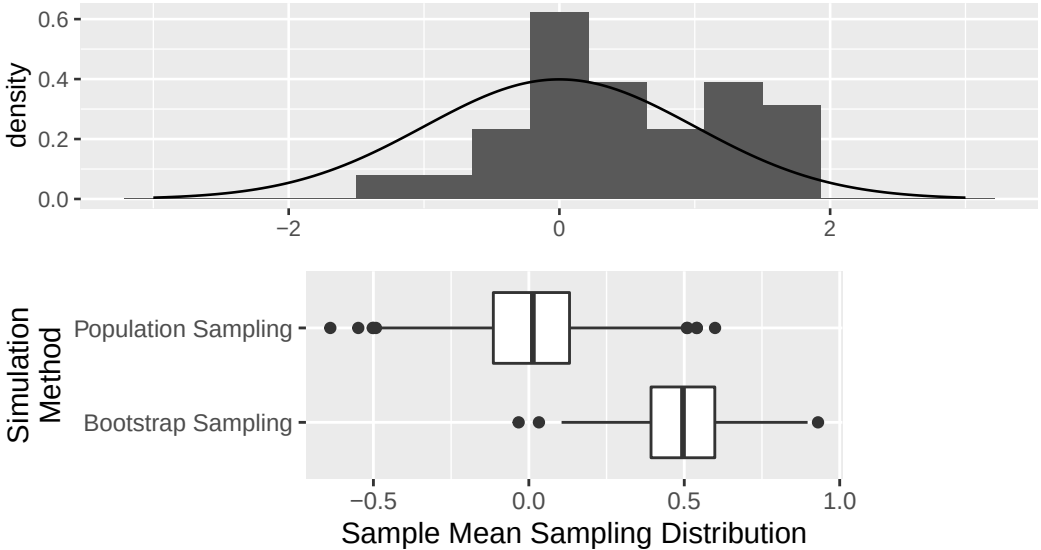
```
# {r, include=TRUE, echo=TRUE, fig.width=6, fig.height=1.25}
sampling_distribution <- c(population_sample_means,
                           bootstrap_sample_means)
simulation_method <- c(rep("Population Sampling",N),
                       rep("Bootstrap Sampling",N))
tibble(`Sampling Distribution` = sampling_distribution,
       `Simulation\nMethod` = simulation_method) %>%
  ggplot(aes(x=`Sampling Distribution`, y=`Simulation\nMethod`)) +
  geom_boxplot() + xlab("Sample Mean Sampling Distribution")
# Does Bootstrap Approximation Work?
```



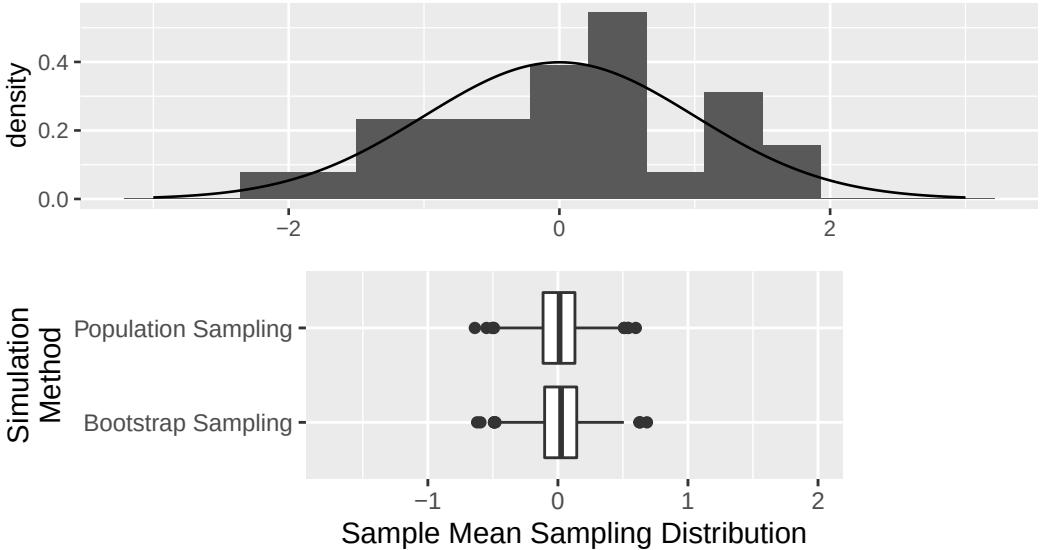
How Well Do Samples Approximate Populations?



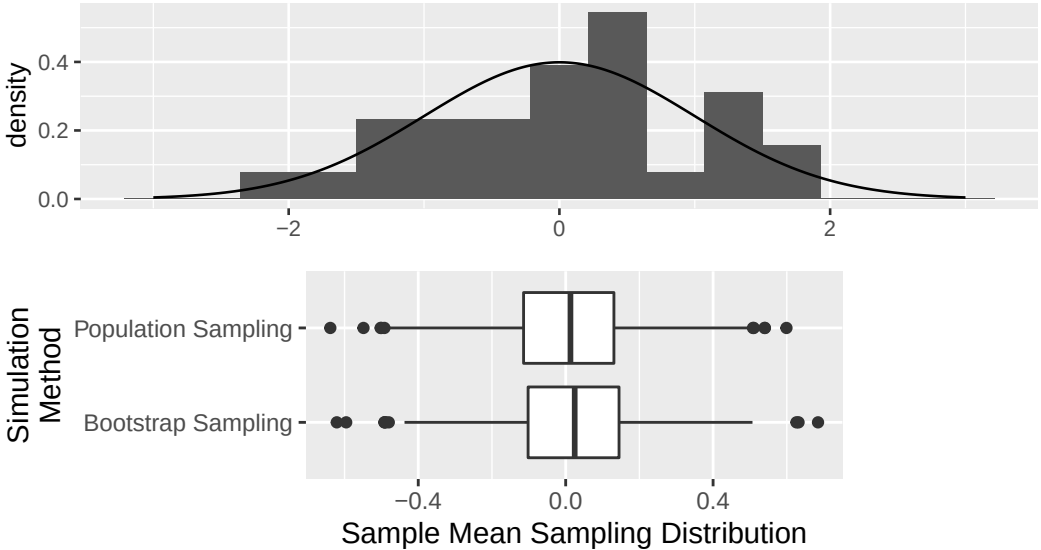
How Well Do Samples Approximate Populations?



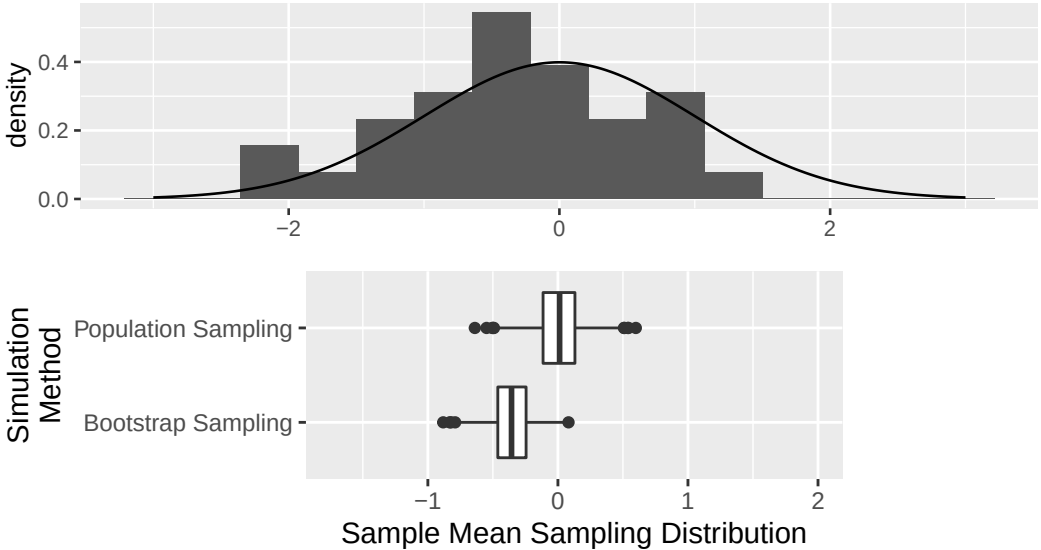
How Well Do Samples Approximate Populations?



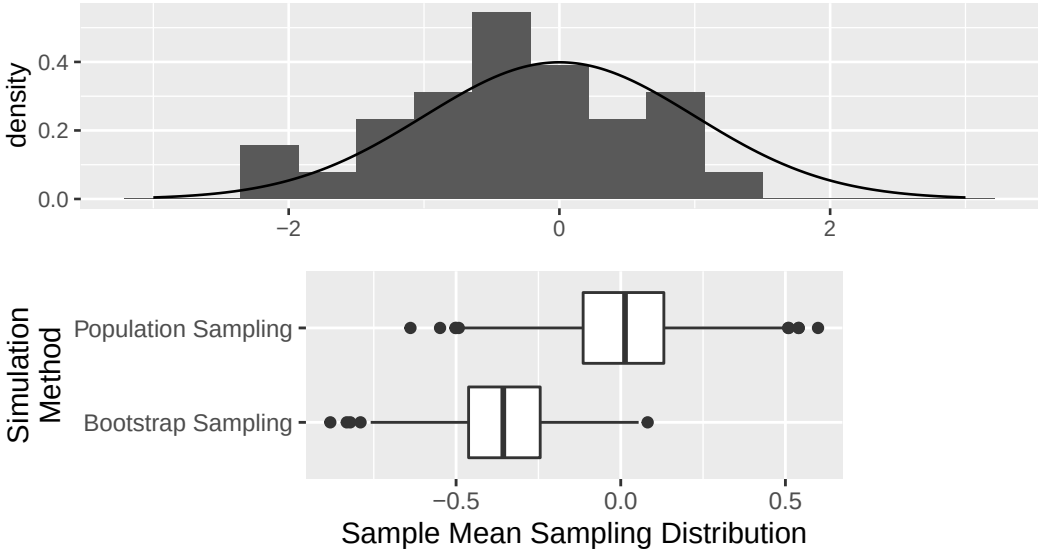
How Well Do Samples Approximate Populations?



How Well Do Samples Approximate Populations?



How Well Do Samples Approximate Populations?



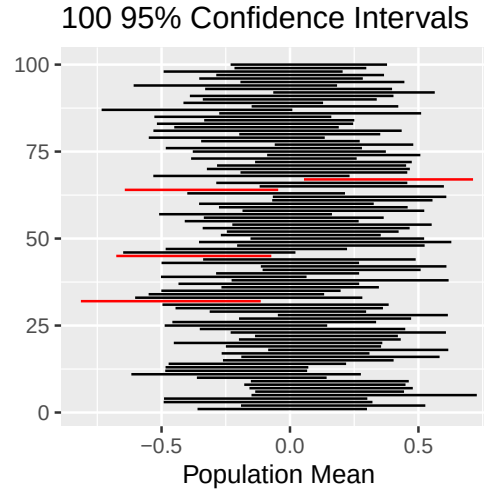
Sanity Check

1. What is x_i?
2. What is the **population** of x_i?
3. What is a parameter?
4. What is the sample `x = c(x_1,x_2,...x_n)`?
5. What makes the sample a better approximation of the population?
6. What happens if you get a "bad" sample? Can anything be done about this?
7. What is a statistic?
8. What is $\bar{x}$?
9. What is the **sampling distribution** of $\bar{x}$?
10. How is the **sampling distribution** of $\bar{x}$ created?

How Well Do Samples Approximate Populations?

```
plot<-ggplot();mu<-0; n<-30;set.seed(130)
for(i in 1:R){
  x <- rnorm(mean=mu, n=n)
  bootstrap_xbar <- 1:N
  for(j in 1:N){
    tmp <- sample(x, replace=TRUE)
    bootstrap_xbars[j] <- mean(tmp)
  }
  ConfidenceInterval <-
    quantile(bootstrap_xbars, percentiles)
  if( all(ConfidenceInterval < mu) |
      all(ConfidenceInterval > mu) ){
      col="red"}else{col="black"}
  plot <- plot + geom_line(color = col,
    data = tibble(x=ConfidenceInterval,
      y=c(i,i)), aes(x=x, y=y))
}; plot+xlab("Population Mean")+labs(title=
  paste(R, " ", (1-2*half_alpha)*100,
        "% Confidence Intervals", sep=""))+
  theme(axis.title.y=element_blank())
```

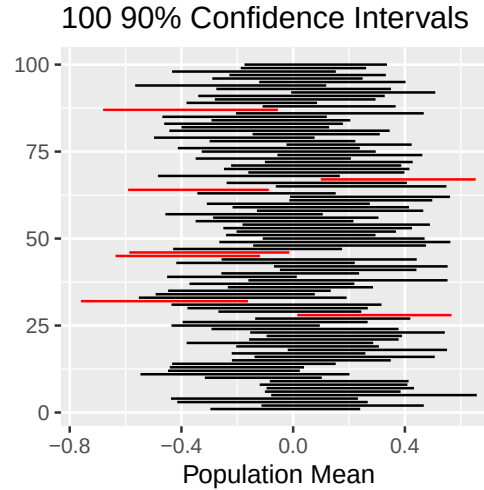
```
half_alpha <- 0.025; R <- 100; N <- 1000#00?
percentiles <- c(half_alpha, 1-half_alpha)
```



How Well Do Samples Approximate Populations?

```
plot<-ggplot();mu<-0; n<-30;set.seed(130)
for(i in 1:R){
  x <- rnorm(mean=mu, n=n)
  bootstrap_xbar <- 1:N
  for(j in 1:N){
    tmp <- sample(x, replace=TRUE)
    bootstrap_xbars[j] <- mean(tmp)
  }
  ConfidenceInterval <-
    quantile(bootstrap_xbars, percentiles)
  if( all(ConfidenceInterval < mu) |
      all(ConfidenceInterval > mu) ){
      col="red"}else{col="black"}
  plot <- plot + geom_line(color = col,
    data = tibble(x=ConfidenceInterval,
      y=c(i,i)), aes(x=x, y=y))
}; plot+xlab("Population Mean")+labs(title=
  paste(R, " ", (1-2*half_alpha)*100,
        "% Confidence Intervals", sep=""))+
  theme(axis.title.y=element_blank())
```

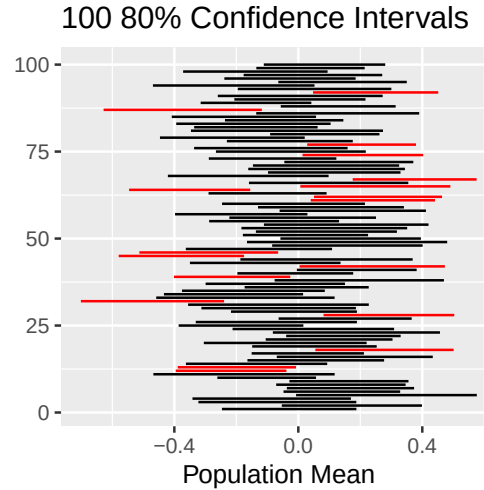
```
half_alpha <- 0.05; R <- 100; N <- 1000#00?
percentiles <- c(half_alpha, 1-half_alpha)
```



How Well Do Samples Approximate Populations?

```
plot<-ggplot();mu<-0; n<-30;set.seed(130)
for(i in 1:R){
  x <- rnorm(mean=mu, n=n)
  bootstrap_xbar <- 1:N
  for(j in 1:N){
    tmp <- sample(x, replace=TRUE)
    bootstrap_xbars[j] <- mean(tmp)
  }
  ConfidenceInterval <-
    quantile(bootstrap_xbars, percentiles)
  if( all(ConfidenceInterval < mu) |
      all(ConfidenceInterval > mu) ){
      col="red"}else{col="black"}
  plot <- plot + geom_line(color = col,
    data = tibble(x=ConfidenceInterval,
      y=c(i,i)), aes(x=x, y=y))
}; plot+xlab("Population Mean")+labs(title=
  paste(R, " ", (1-2*half_alpha)*100,
        "% Confidence Intervals", sep=""))+
  theme(axis.title.y=element_blank())
```

```
half_alpha <- 0.1; R <- 100; N <- 1000#00?
percentiles <- c(half_alpha, 1-half_alpha)
```



Sanity Check

1. What is the "population" when bootstrapping a sampling distribution?
2. Should `replace=TRUE` or `replace=FALSE` when bootstrapping with `sample()`?
3. Why would we use the `quantile()` function in the bootstrapping context?
4. What value of the `probs` parameter gives a 90% confidence intervals?
5. How does the confidence level relate to the width of the Confidence Intervals?
6. Are there one or two `for` loops when we're bootstrapping a confidence interval?
7. Why?

Statistical Grammar Police

For a 95% Confidence Interval we say

we have 95% **Confidence** the true parameter value is contained in the Interval

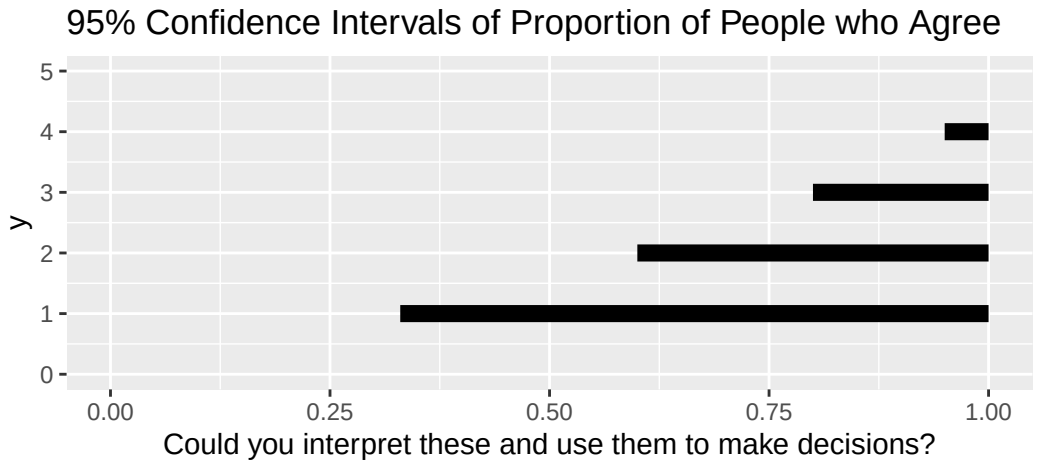
- We use the term **Confidence** (as opposed to *probability* or *chance*) to intentionally signal that this is a **Confidence Interval** formulation.
- We **DO NOT** want to say there's a 95% probability (or chance) that the true parameter value will be contained in some some interval.
  - This **might** be misinterpreted as saying that the true parameter is usually in the interval but sometimes not; but parameters are only just in the interval or not*

Are we just splitting hairs here?

The chance the constructed confidence interval bounds the true parameter value is 95%

Using Confidence Intervals

Do we need formal Hypotheses? Can actionable decisions be made with these?



Confidence Intervals VS Hypothesis Testing

STATISTICAL INFERENCE: Parameter *ESTIMATION*

α-significance level **Hypothesis Testing** formally rejects implausible parameter values

- *What if we'd instead like to provide a range of plausible parameter values?*

At a fixed confidence level,
narrower intervals are more meaningful and therefore more likely actionable.

- "We have '95% Confidence' that anywhere between 1% to 99% of people agree!"

How Do We Get Tighter Confidence Intervals: ?

We previously saw that for the same data

- 80% Confidence Intervals are narrower than
  - 90% Confidence Intervals, which are narrower than
    - 95% Confidence Intervals, which are narrower than...

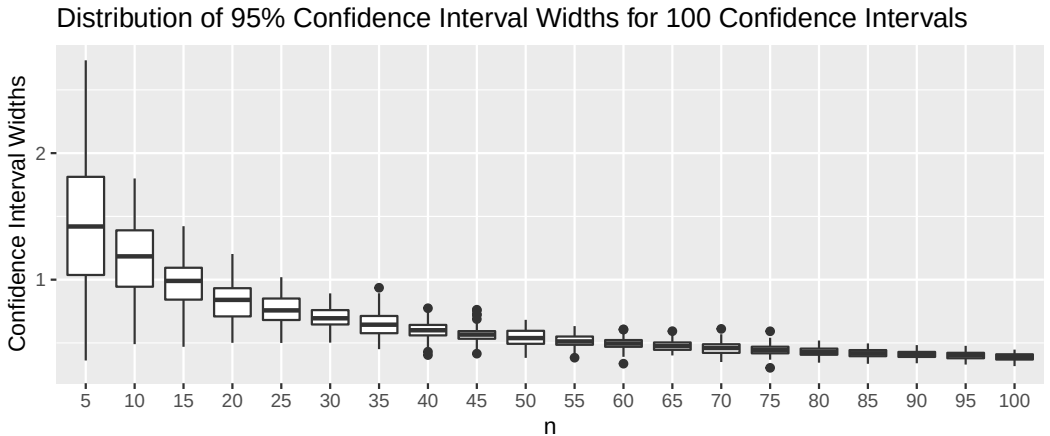
This won't really help us though...

Sanity Check

So what's the difference between Confidence Intervals and Hypothesis Tests?

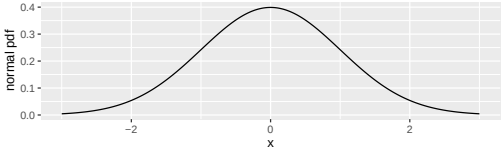
1. What are the different "use cases" for each?
2. Which one gives *Estimation*?
3. Are they both *Inference*?
4. Can both be use to make decisions?
5. How so?
6. When are these decisions more power/useful/trustworthy/actionable?

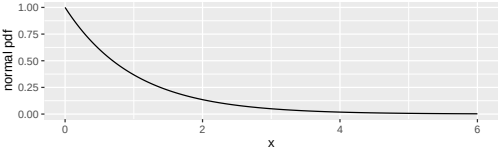
How Do We Get Tighter Confidence Intervals? n

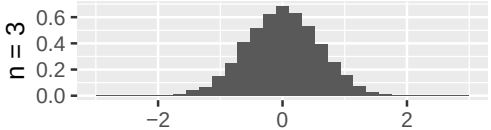


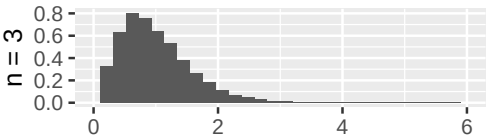
https://www.zoology.ubc.ca/~whitlock/Kingfisher/CLT.htm

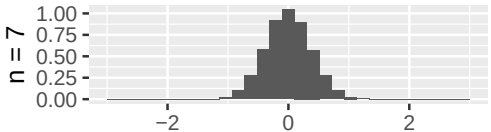
The Sampling Distribution of $\bar{x}$ VS Skewness

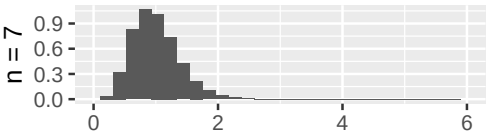


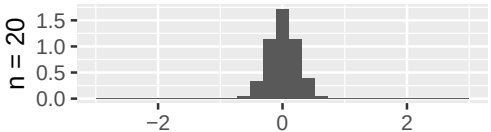


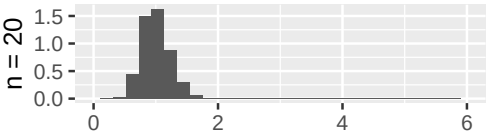












Confidence Intervals VS Hypothesis Testing

Confidence Intervals $[\hat{\mu}_{lower}, \hat{\mu}_{upper}]$

1. **Approximate** population as sample
2. Bootstrap **sampling distribution**
3. Define Confidence Interval with sampling distribution percentiles

$[(1-\alpha)\times 100]$% Confidence Interval

- α chance the confidence interval does not bound the true parameter value
- This is not $Pr(\mu \in [\hat{\mu}_{lower}, \hat{\mu}_{upper}])$, it is $Pr([\hat{\mu}_{lower}, \hat{\mu}_{upper}]$ bounds $\mu)$
- *True μ isn't a random thing, but $[\hat{\mu}_{lower}, \hat{\mu}_{upper}]$ based on the sample is

Hypothesis Testing $H_0 : \mu = \mu_0$

1. **Assume** a population through H_0
2. Get **sampling distribution** under H_0
3. Compute the sample p-value and either Reject or Fail to Reject H_0

α-level significance testing

- α chance of a Type I Error if H_0 rejected for a p-value smaller than α
- p-values are neither $Pr(H_0$ is `TRUE`$)$ nor $Pr(\mu = \mu_0)$
- *True μ and hence H_0 aren't random things, but sample-based p-values are

Sanity Check

For a 95% Confidence Interval and α-level significance test

1. What is the probability $\mu \in [\hat{\mu}_{lower}, \hat{\mu}_{upper}]$
2. What is the probability that $[\hat{\mu}_{lower}, \hat{\mu}_{upper}]$ bounds μ?
3. What is $Pr(H_0$ is `TRUE`$)$?
4. What is $Pr(\mu = \mu_0)$?
5. What is the p-value?
6. What is the role of the test statistic, parameter, population, and sampling distribution relative to the *p-value*? And how about for a *confidence interval*?

The p-value ConTROVersy/ContraVersy

- Why are p-values controversial?
- What a nerdy debate about p-values shows about science and how to fix it
- The reign of the p-value is over: what alternative analyses could we employ to fill the power vacuum?
- Scientists rise up against statistical significance
- Statistics experts urge scientists to rethink the p-value

The main problem is wrongly interpreting p-values as $Pr(H_0 \text{ is } \texttt{TRUE})$ and $Pr(\mu = \mu_0)$

but the deeper problems with p-values are

- introduced here and presented here
- and rely upon understanding the simulation here

The p-value ConTROVersy/ContraVersy

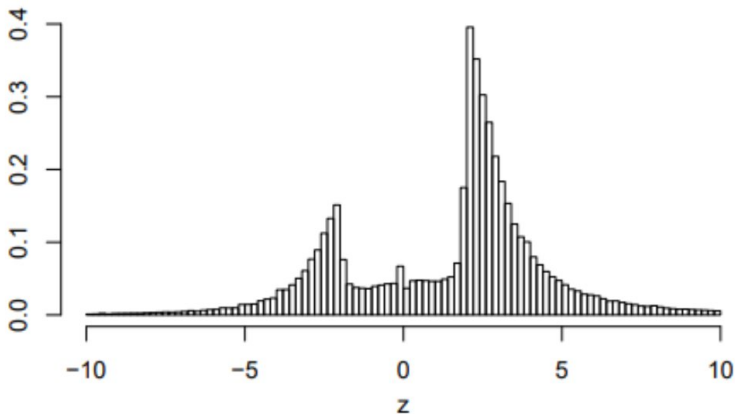


Figure 1: The distribution of more than one million z-values from Medline (1976–2019).

https://twitter.com/fmg_twtr/status/1334884184675012609

The p-value ConTROVersy/ContraVersy

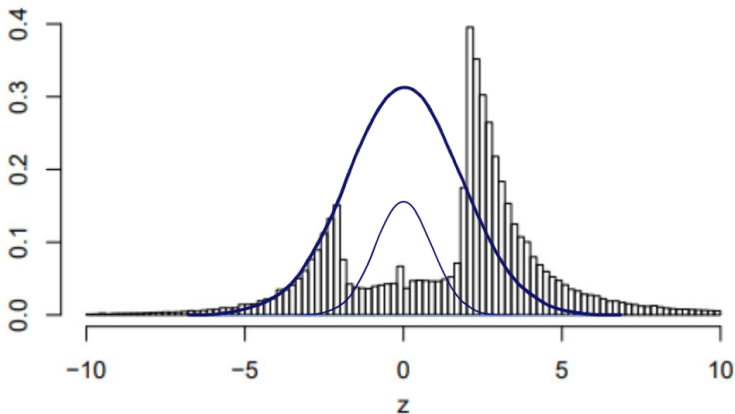


Figure 1: The distribution of more than one million z-values from Medline (1976–2019).

https://twitter.com/fmg_twtr/status/1334884184675012609

The p-value ConTROVersy/ContraVersy

George Cobb, Professor Emeritus, Mount Holyoke College

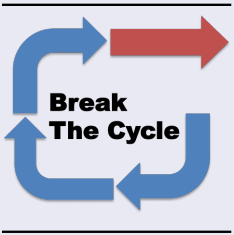
Q: Why do colleges/grad schools use $\alpha = 0.05$ *thresholds for statistical significance?*

A: Because that's still what the scientific community and journal editors use.

Q: Why do so many people use $\alpha = 0.05$ *thresholds for statistical significance?*

A: Because that's what they were taught in college or grad school.

$\alpha = 0.05$ is arbitrary: better to either simply comment on the strength of the evidence against H_0 by reporting the p-value (or at least choose α before calculating the p-value)



p-value	evidence against H_0
above 0.1	None
0.05 to 0.1	Weak
0.01 to 0.05	Moderate
0.0001 to 0.01	Strong
below 0.0001	Very Strong

FIX The p-value ConTROVersy/ContraVersy

0. Don't interpret p-values as $Pr(H_0 \text{ is } \texttt{TRUE})$ or $Pr(\mu = \mu_0)$:

   p-values are the probability of observing a test statistic that is *as or more extreme* than the one we got **if the NULL Hypothesis was actually** `TRUE`

1. Want to control Type I error? Set α and do a α-significance test
2. Want to use a "measure of evidence" perspective without controlling Type I error?

   Don't retroactively interpret p-value in terms of Type I error, and instead, say

p-value	above 0.1	0.05 to 0.1	0.01 to 0.05	0.0001 to 0.01	below 0.0001
evidence against H_0	None	Weak	Moderate	Strong	Very Strong

3. Or, use a **Confidence Interval**: get BOTH the estimate *AND* its strength. How?

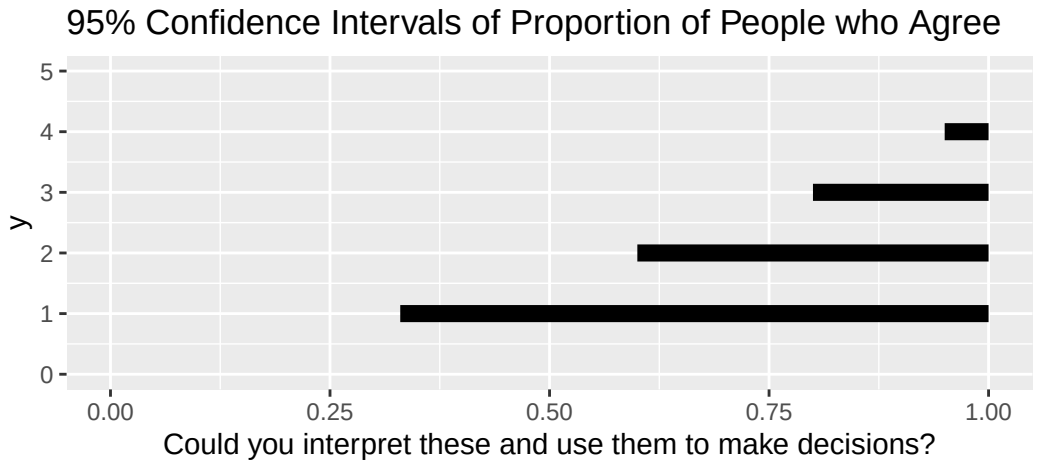
Winning


"THE ONLY THING I'M ADDICTED TO
RIGHT NOW
IS WINNING."


I CAN'T HEAR YOU
I'M TOO BUSY WINNING

FIX The p-value ConTROVersy/ContraVersy

Do we need formal Hypotheses? Can actionable decisions be made with these?



Sanity Check

- True/False: If you have relevant data for each individual in the population, you can calculate the true value of parameters.
- True/False: In general, we know the true value of parameters.
- True/False: We only know the true value of parameters when we're doing Hypothesis Testing, not when we're estimating them with Confidence Intervals.
- True/False: A statistic is calculated from observed data and is an estimate of a true parameter value.
- True/False: Every random sample drawn from the population will yield the same values for statistics.