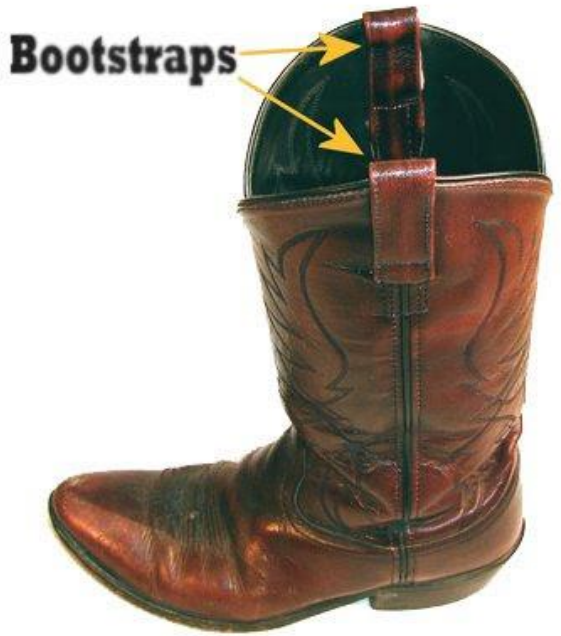
Sampling distributions, confidence intervals, and the bootstrap



Overview

Very quick review of probability functions

Sampling distributions

Confidence intervals

Computing confidence intervals using the bootstrap

Announcements

Information about learning group has been sent out

- There is still time to sign up: email stephan.billingslea@yale.edu

UConn Sports Analytics Symposium online on Saturday Oct 9th

- https://statds.org/events/ucsas2021/
- $5 registration fee

I posted my Introductory Statistics slides under Resources on Canvas

- Note: I wrote special functions for that class so some of the code won't work in base R (e.g., the "do_it" function instead of for loops).

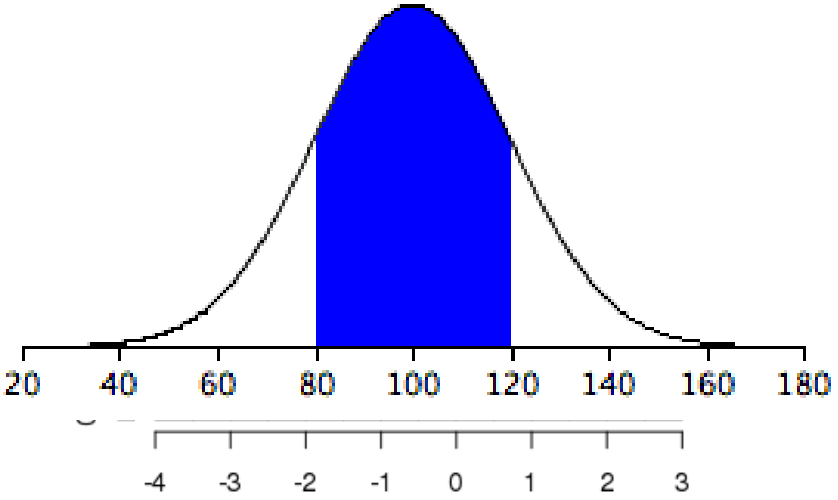
Very quick review of probability functions

To **generate random data** we use functions that start with the letter ***r***

```
> rand_data <- rnorm(100)
> hist(rand_data)
```

To **plot probability density functions** we use functions that start with the letter ***d***

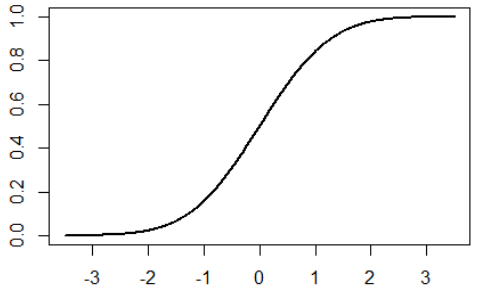
```
> x <- seq(-3, 3, by = .001)
> y <- dnorm(x)
> plot(x, y, type = "l")
```



$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

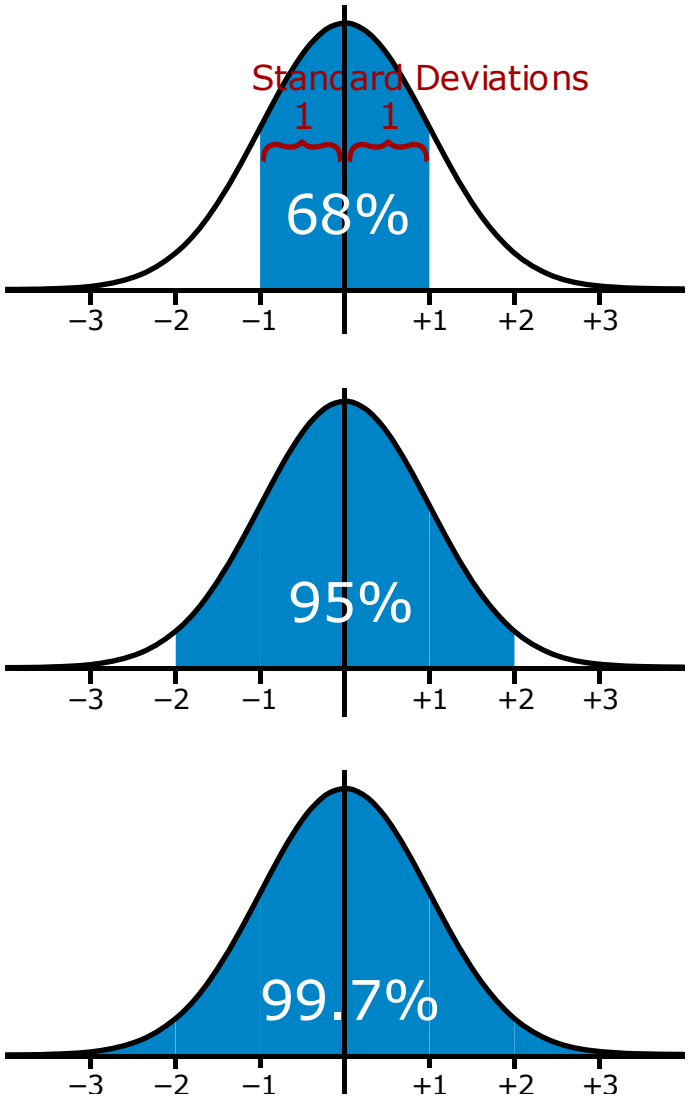
To **get the probability** that that a random number X is less than x, $P(X \leq x)$, we use functions that start with ***p***

```
> pnorm(1)
```



$$P(X \leq x) = \int_{-\infty}^{x} f(x)dx$$

Normal density function



Sampling distributions

A ***sampling distribution*** is a distribution of ***statistics***

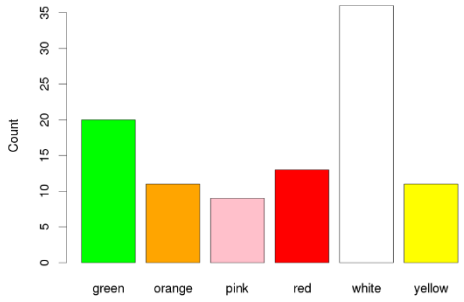
- (a ***statistic*** is a number computed from a sample of data)

Reminder: For a single ***categorical variable***, the main statistic of interest is the ***proportion*** ($\hat{p}$) in each category

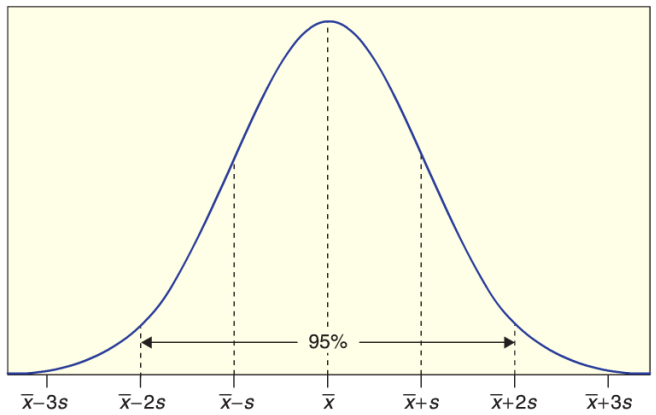
- (shadow of the parameter π)

$$\hat{p} = \text{Proportion in a category} = \frac{\text{number in that category}}{\text{total number}}$$

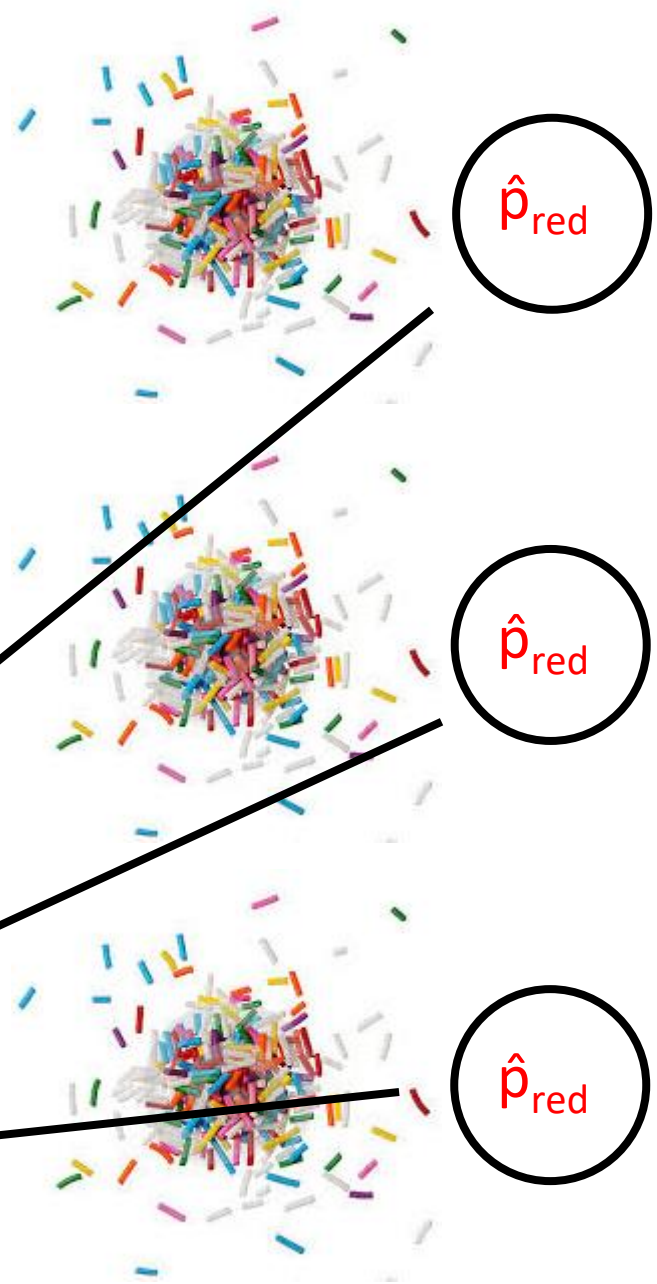
n = 100



π_{red}



Sampling distribution!

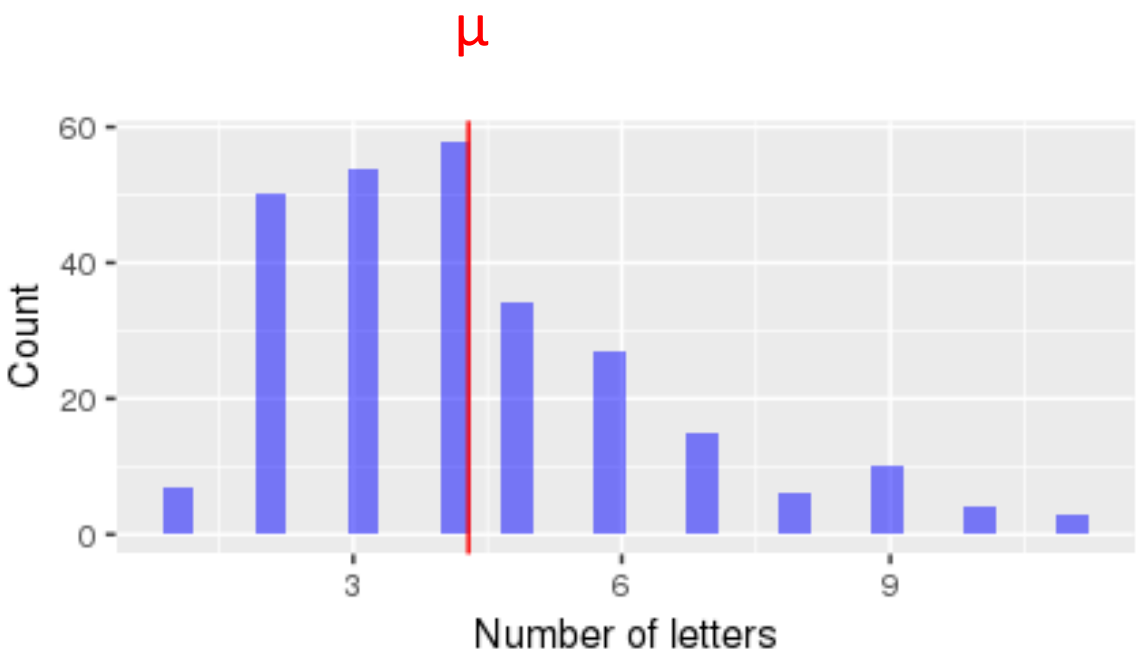


$\hat{p}_{red}$

$\hat{p}_{red}$

$\hat{p}_{red}$

Another sampling distribution illustration



10, 3, 3, 3, 4,
3, 2, 6, 10, 5

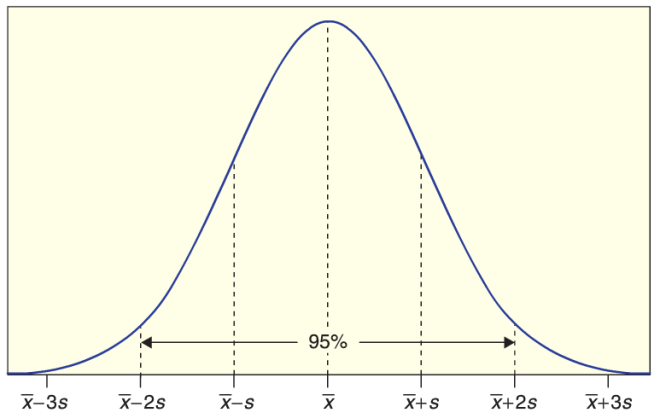
$\bar{x} = 5$

2, 6, 2, 6, 6,
2, 5, 3, 2, 9

$\bar{x} = 4.3$

3, 9, 3, 4, 4,
3, 6, 6, 2, 2

$\bar{x} = 4.2$

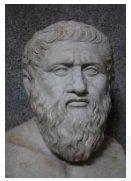


Sampling distribution!

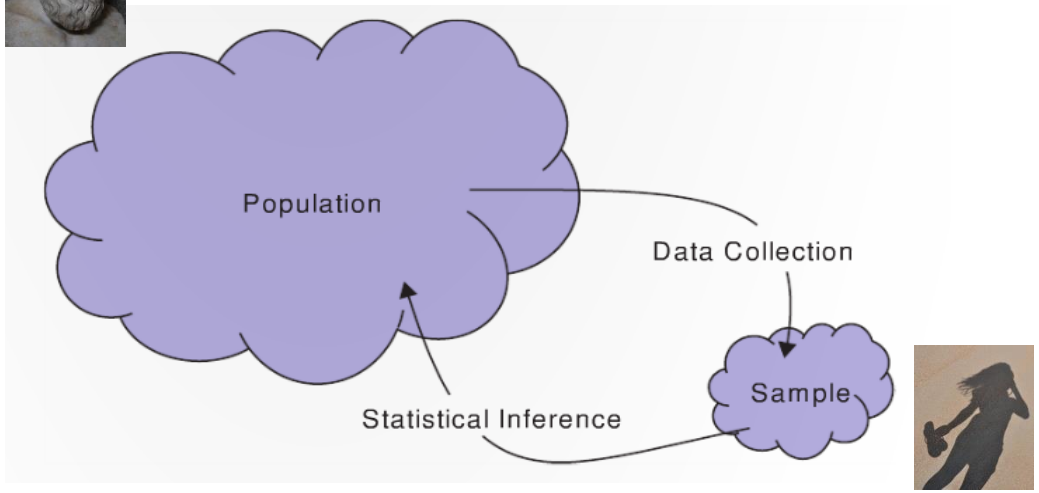
Sampling distribution

Why are we interested in the sampling distribution?

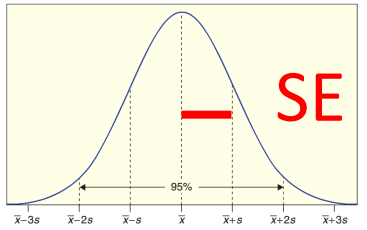
- If we knew what the sampling distribution was, then we could evaluate how much we should trust individual statistics



Parameters: π, μ, σ, ρ, β



Sampling distribution

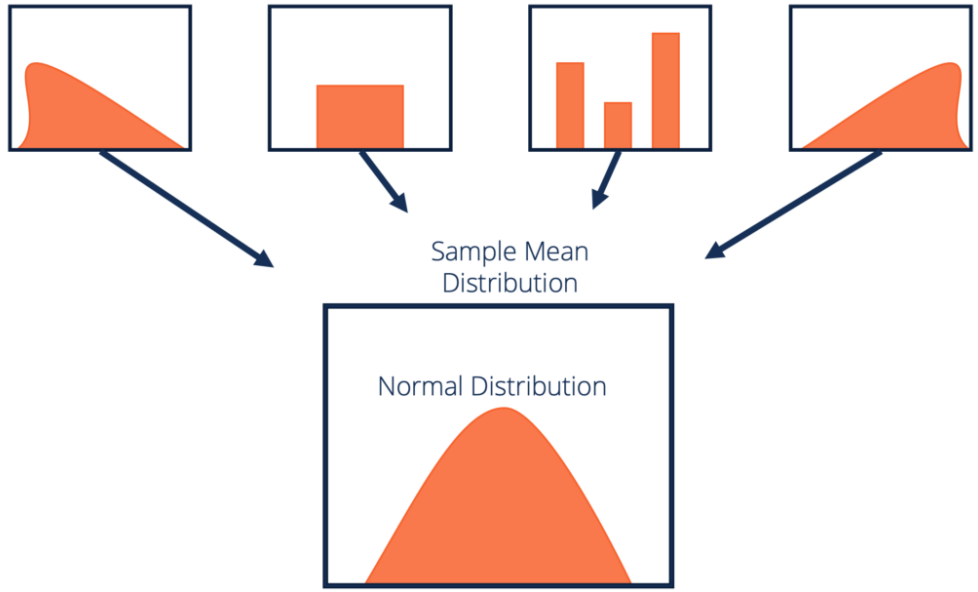


Statistics: $\hat{p}$, $\bar{x}$, s, r, b

The standard deviation of a sampling distribution is called the standard error (SE)

The central limit theorem

The **central limit theorem** establishes that, in many situations, when independent random variables are summed up, their properly normalized sum tends toward a normal distribution.



Since many statistics we use are the sum of randomly data, many of our sampling distributions will be approximately normal

- You will explore this more on homework 2

Statistics: $\hat{p}$, $\bar{x}$, s, r, b

Sampling distributions in R…

Simulating a sampling distribution

```
sampling_dist <- NULL
for (i in 1:1000) {
        rand_data <- runif(100)    # generate n = 100 points from U(0, 1)
        sampling_dist[i] <- mean(rand_data)    # save the mean
}

hist(sampling_dist)
```

Confidence intervals

Point Estimate

We use the statistics from a sample as a **point estimate** for a population parameter

- $\bar{x}$ is a point estimate for…? μ

A NPR/PBS NewHour/Marist poll listed Biden's approval rating at 43%

Symbols:

π: Biden's approval for all voters

$\hat{p}$: Biden's approval for those voters in our sample

Interval estimate based on a margin of error

An **interval estimate** give a range of plausible values for a population parameter

One common form of an interval estimate is:

Point estimate ± margin of error

Where the **margin of error** is a number that reflects the precision of the sample statistic as a point estimate for this parameter

Example: NPR/PBS NewHour/Marist poll

43% of American approve of Biden's job performance, plus or minus 3%

How do we interpret this?

Says that the population parameter (π) lies somewhere between 40% to 46%

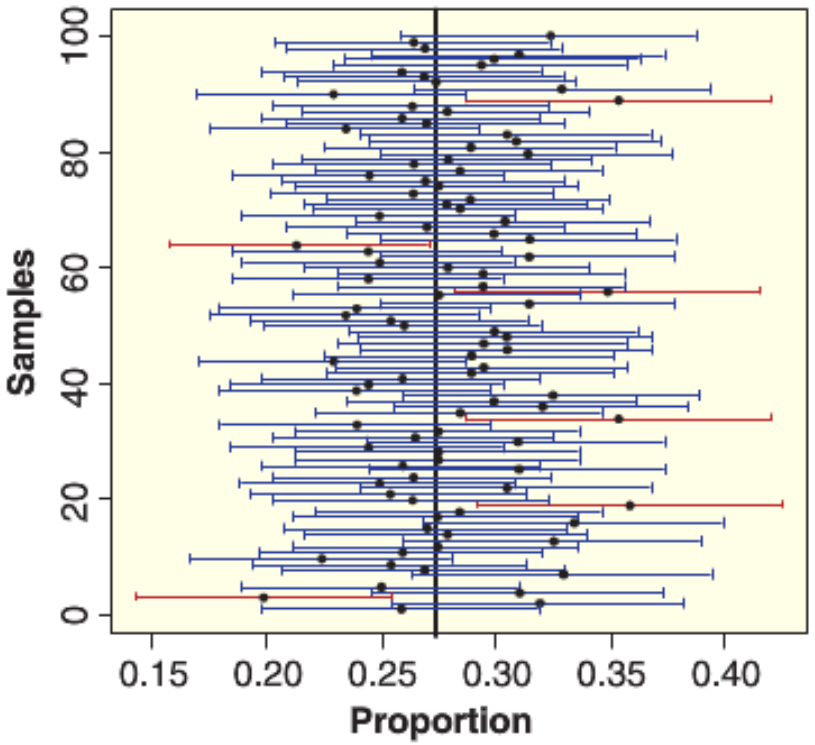
i.e., if they sampled all voters the true population proportion (π) would be likely be in this range

Confidence Intervals

A **confidence interval** is an interval computed by a method that will contain the ***parameter*** a specified percent of times

- i.e., if the estimation were repeated many times, the interval will have the parameter x% of the time

The **confidence level** is the percent of all intervals that contain the parameter



Think ring toss…

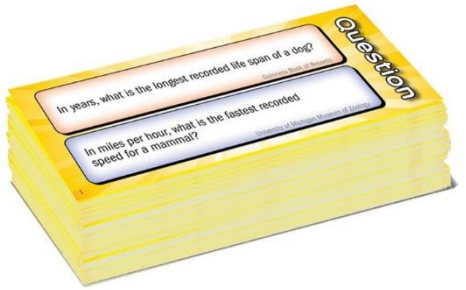
Parameter exists in the ideal world

- E.g., it's a single number

We toss intervals at it

95% of those intervals capture the parameter

Wits and Wagers: 90% confidence intervals estimators



I am going to ask you 10 questions

You need to produce an **interval range** that contains the true answer for 9 out of the 10 questions I ask

Please write down your answers on a piece of paper

Wits and Wagers…

Question 1: What year was Yale University founded?

Question 2: In what year did Benjamin Franklin prove that lightning was electricity, after flying his kite in a thunderstorm?

Question 3: How many floors does the leaning tower of Pisa have?

Wits and Wagers…

Question 4: In feet, how tall was the tallest giraffe ever recorded?

Question 5: In years, what is the longest recorded life span of a dog?

Question 6: How many pounds does one gallon of whole milk weigh?

Question 7: In pounds, what was the weight of the heaviest domesticated cat ever recorded?

Wits and Wagers…

Question 8: If a person weighs 100 pounds on Earth, how many pounds would they weigh on the surface of the moon?

Question 9: What percentage of American adults say that reading is their favorite leisure-time activity?

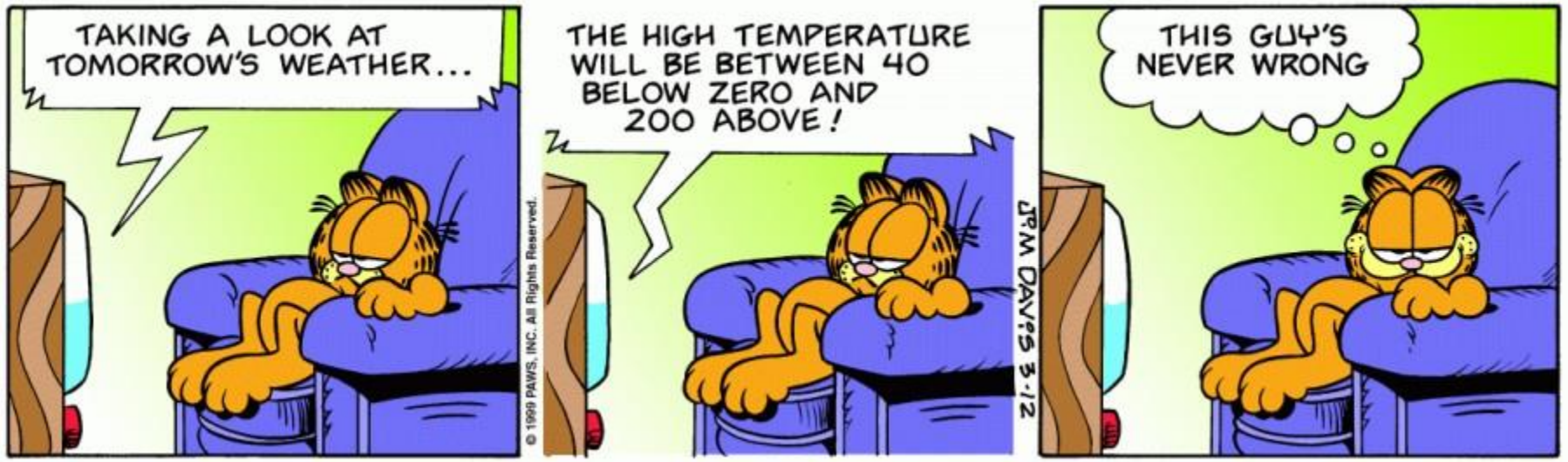
Question 10: How many cups of coffee does the average American drink per year?

Can I take a picture of the moon?

Pisa Tower: yea sorry


How did
we do?

100% confidence intervals



There is a <u>tradeoff</u> between the **confidence level** (percent of times we capture the parameter) and the **confidence interval size**

Note

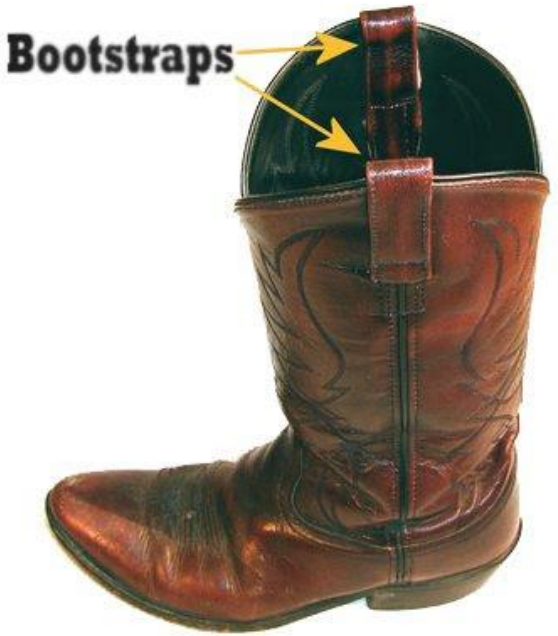
For any given confidence interval we compute, we don't know whether it has really captured the parameter

But we do know that if we do this 100 times, 90 of these intervals will have the parameter in it

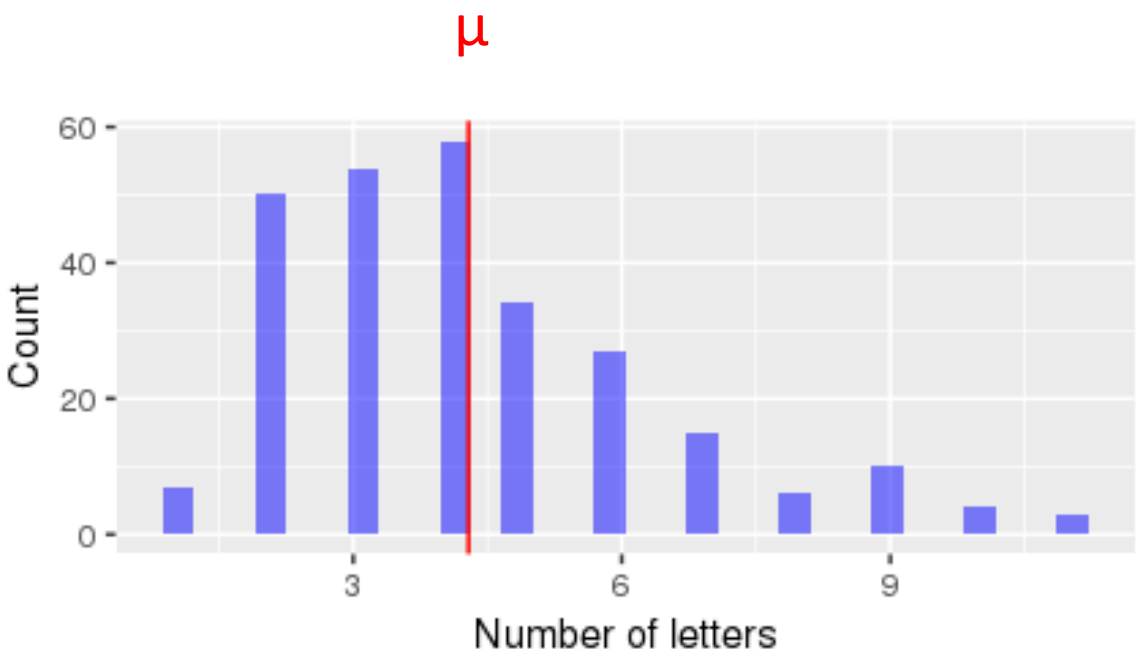
(for a 90% confidence interval)

Computing confidence intervals

Let’s now discuss how we can compute confidence intervals…



Recall: sampling distribution illustration



10, 3, 3, 3, 4,
3, 2, 6, 10, 5

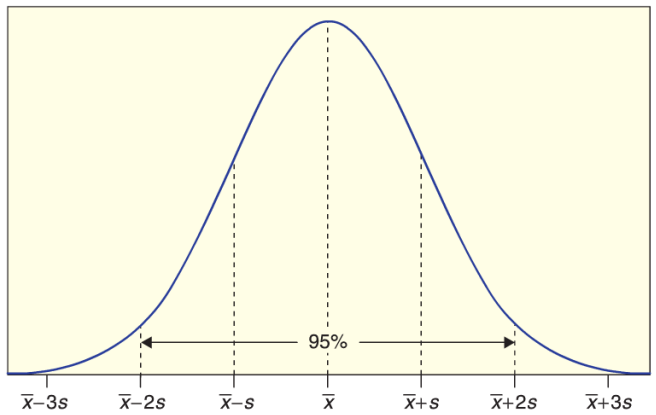
$\bar{x} = 5$

2, 6, 2, 6, 6,
2, 5, 3, 2, 9

$\bar{x} = 4.3$

3, 9, 3, 4, 4,
3, 6, 6, 2, 2

$\bar{x} = 4.2$

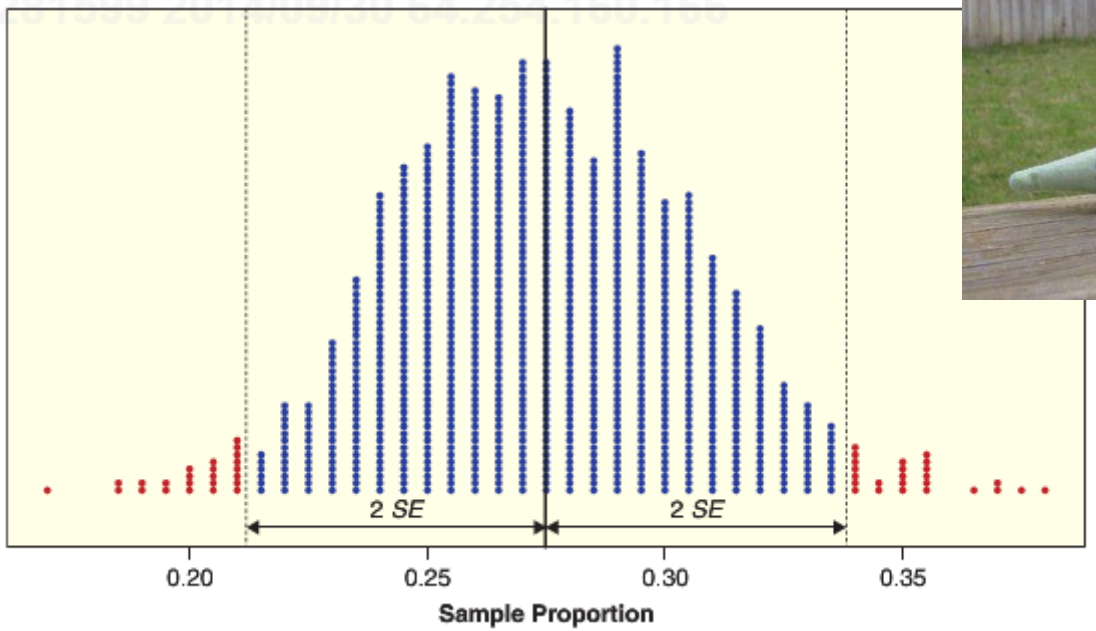


Sampling distribution!

Sampling distributions

Q: For a sampling distribution that is a normal distribution, what percentage of ***statistics*** lie within 2 standard deviations (SE) for the population mean?

A: 95%



Sampling distributions

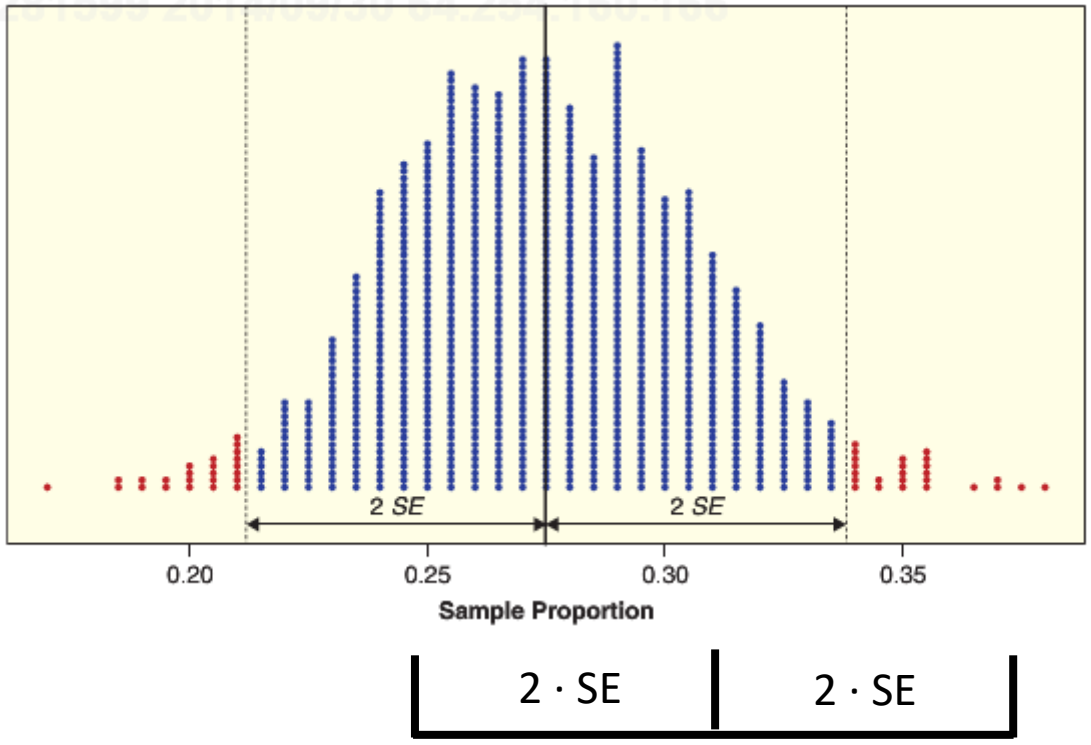
Q: Suppose we had:

- A statistics value
- The SE
- The sampling distribution was normal

Could we compute a 95% confidence interval?

A: Yes!

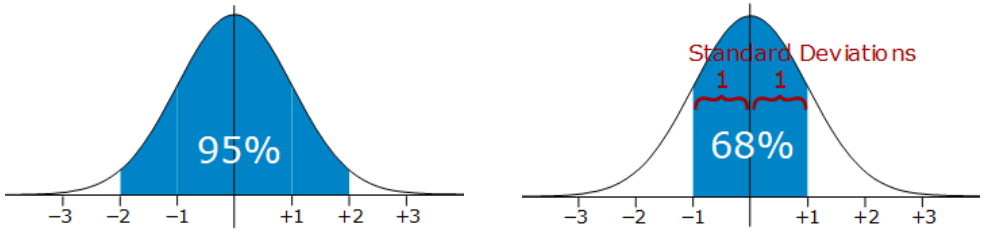
CI = statistic value ± 2 · SE



95% confidence interval: stat ± 2 · SE

Confidence interval

Sampling distributions



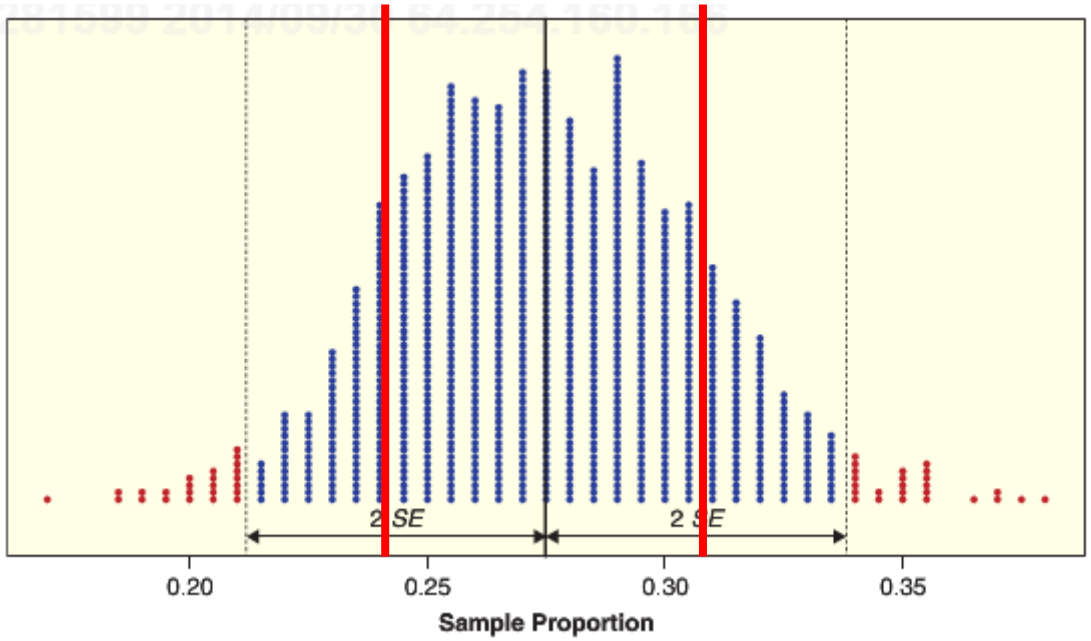
Q: Suppose we had:

- A statistics value
- The SE
- The sampling distribution was normal

Could we compute a 95% confidence interval?

A: Yes!

CI = statistic value ± 2 · SE



What would happen if we made the margin of error smaller?

- E.g., ME = 1 · SE

95% confidence interval: stat ± 2 · SE

Confidence interval

Sampling distributions

Q: Could we repeat the sampling process many times to create a sampling distribution and then calculate the SE?

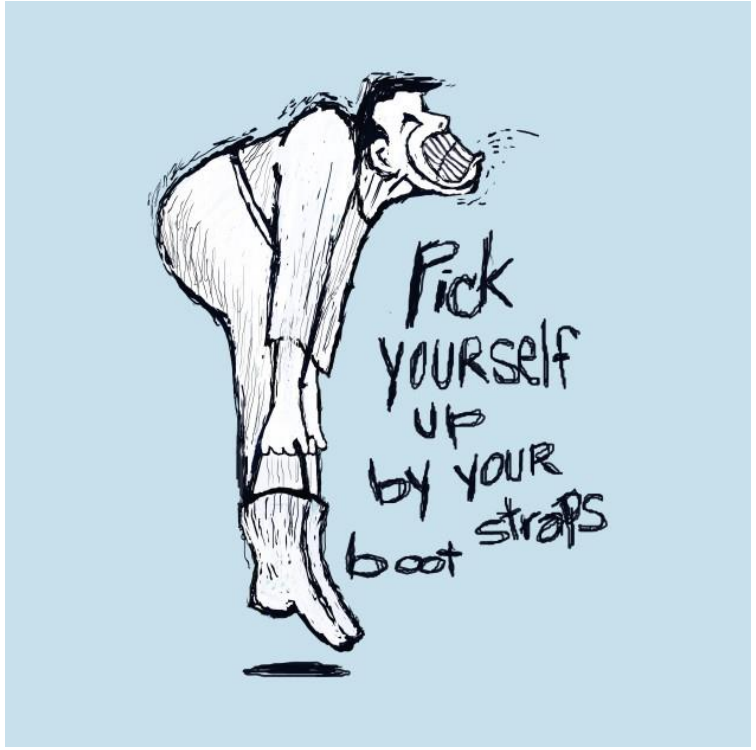
- A: Not in the real world because it would require running our experiment over and over again…

Sampling distributions

Q: If we can't calculate the sampling distribution, what's else could we do?

- A: We could pick ourselves up from the bootstraps

1. Estimate SE with $\hat{SE}$
2. Then use $\bar{x} \pm 2 \cdot \hat{SE}$ to get the 95% CI



Plug-in principle

Suppose we get a sample of size *n* from a population

We pretend that *the sample is the population* (plug-in principle)

1. We then sample *n* points *with replacement* from our sample, and compute our statistic of interest

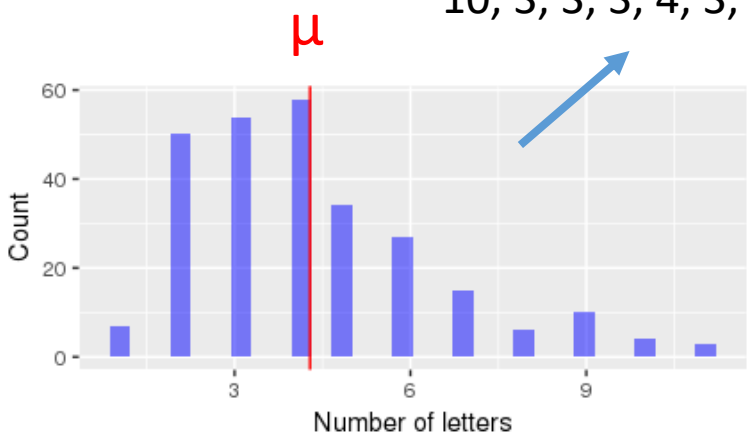
2. We repeat this process 1000's of times and get a ***bootstrap sample distribution***

3. The standard deviation of this bootstrap distribution (SE* bootstrap) is a good approximate for standard error SE from the real sampling distribution

Bootstrap distribution illustration

The sample (n = 10)

10, 3, 3, 3, 4, 3, 2, 6, 4, 5



3, 3, 3, 5, 3,
4, 5, 2, 2, 10

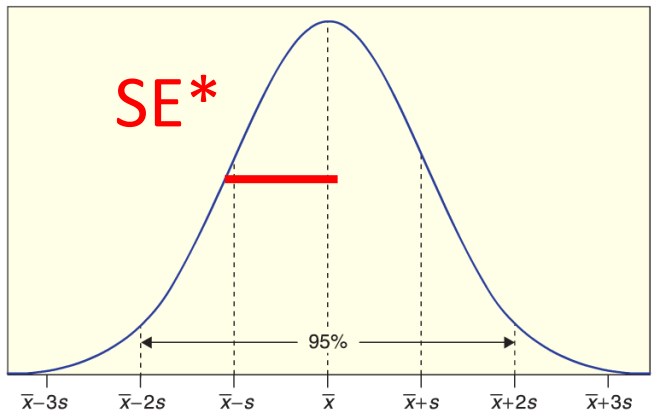
$\bar{x}^* = 4$

3, 3, 2, 3, 6,
4, 6, 5, 3, 6

$\bar{x}^* = 4.1$

5, 3, 2, 3, 3,
3, 10, 3, 4, 3

$\bar{x}^* = 3.9$



Bootstrap distribution!

Notice there is no 9's in the bootstrap samples

95% Confidence Intervals

When a bootstrap distribution for a sample statistic is approximately normal, we can estimate a 95% confidence interval using:

$$Statistic \pm 2 \cdot SE^*$$

Where SE* is the standard error estimated using the bootstrap

Let’s try it in R…

Formulas for the standard error of the mean

As you likely learned in intro statistics class, there is formula the **standard error of the mean (SE mean)** which is:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad\qquad s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Where:

- σ is population standard deviation parameter
- n is the sample size
- s is the sample standard deviation

Formula for the standard error of a proportion

Likewise, there is a formula for **standard error of a proportion (SE proportions)** which is:

$$\sigma_{\hat{p}} = \sqrt{\frac{\pi \cdot (1-\pi)}{n}} \qquad s_{\hat{p}} = \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}}$$

Where:

- π is the population proportion parameter
- n is the sample size
- p̂ is the sample proportion statistic