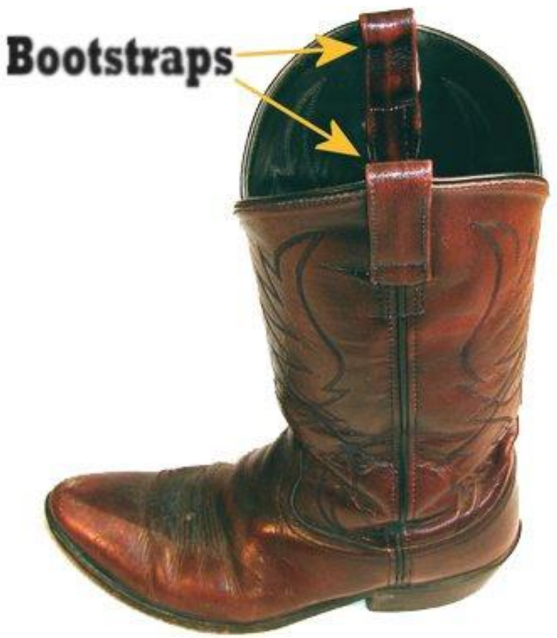
Sampling distributions, confidence intervals, and the bootstrap



Overview

Quick review of sampling distributions

Confidence intervals

Computing confidence intervals using the bootstrap

Announcements

Homework 2 has been posted

- Due Sunday (9/18) at 11pm

Notes:

- There are some useful links for learning R under the "Resources" page on Canvas
- If you get + symbol in the R console in mean you entered an incomplete line of code
  - E.g., `sqrt(`
  - To get back to the regular console (i.e., > symbol) press the escape key

Quick review: for loops

For loops are useful when you want to repeat a piece of code many times under similar conditions

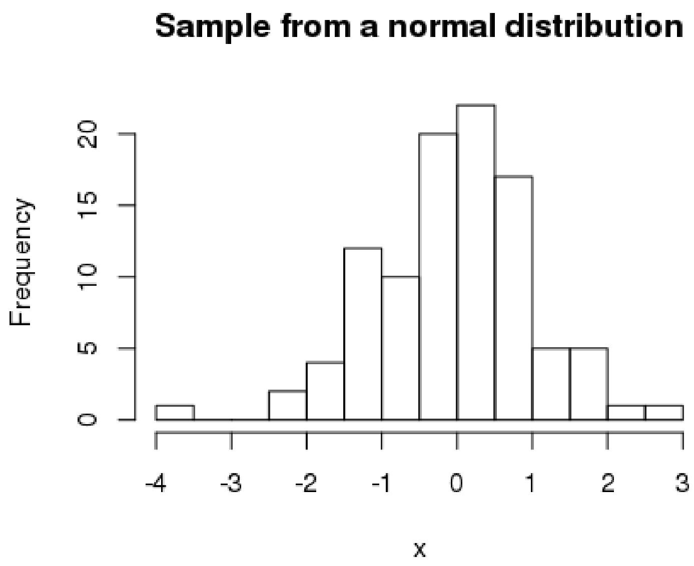
For loops are particularly useful in conjunction with vectors…

```
my_results <- NULL      # create an empty vector to store the results
for (i in 1:100) {
        my_results[i] <- i^2
}
```

Quick review: generating random data and sampling

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

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



We can sample from a vector using the sample function:

```
> my_vec <- 1:100
> my_sample <- sample(my_vec, 30)
> my_sample2 <- sample(my_vec, 30, replace = TRUE)
```

Review: Sampling distributions

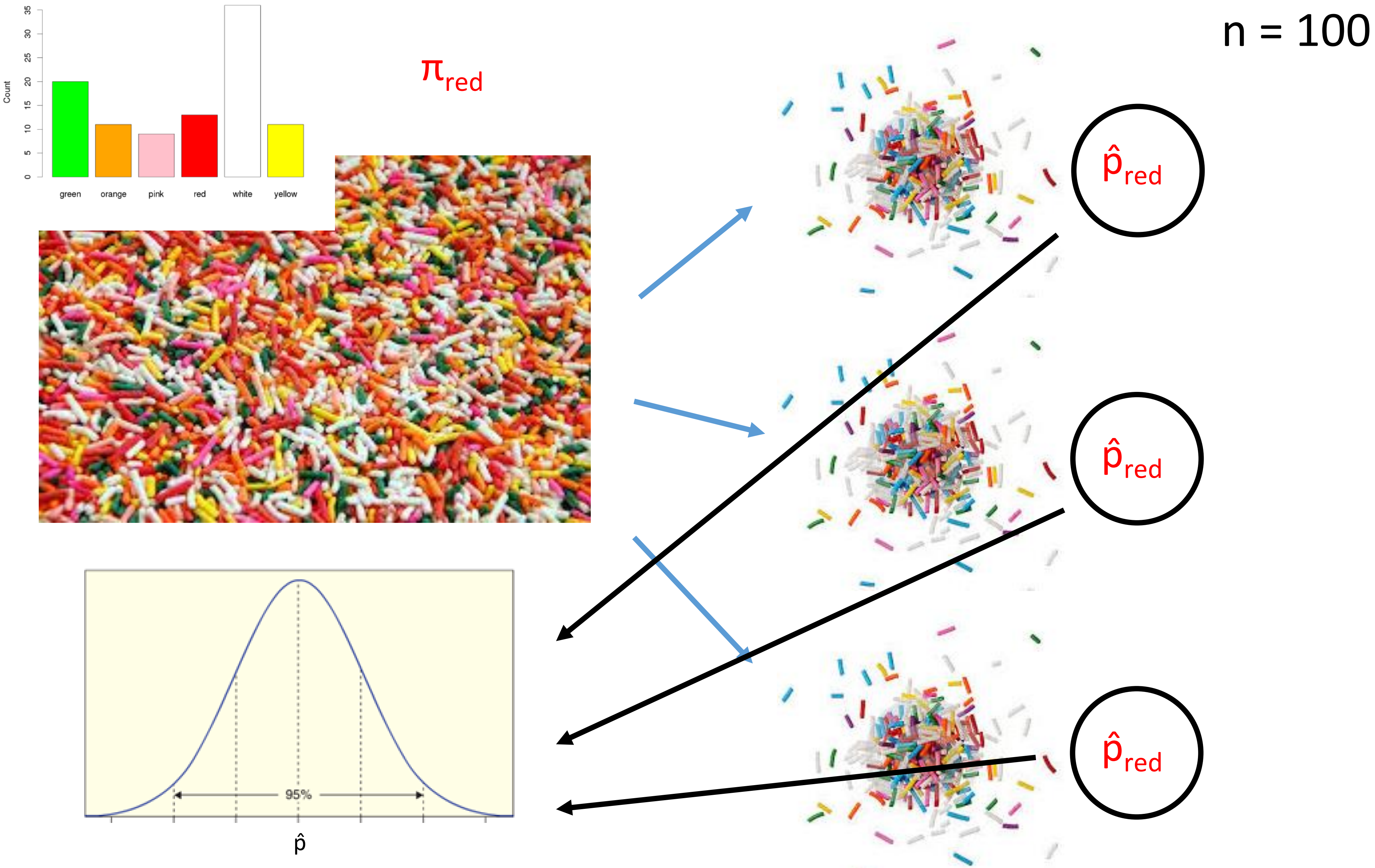
A ***sampling distribution*** is a distribution of ***statistics***

- (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

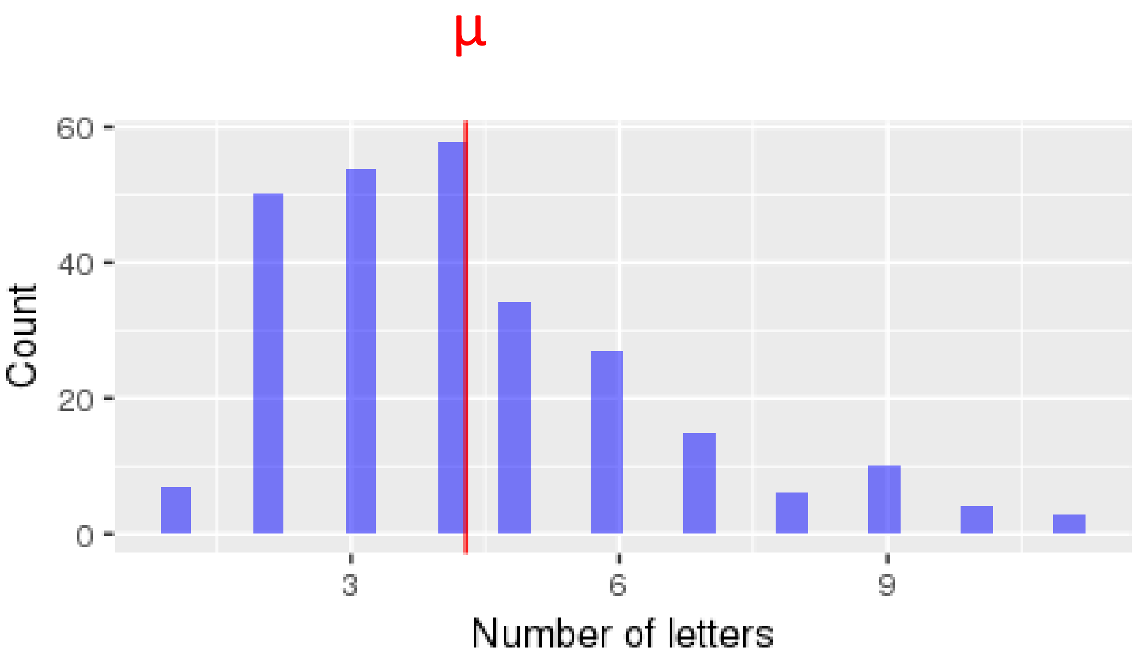
- (shadow of the parameter π)

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



Sampling distribution!

Another sampling distribution illustration



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

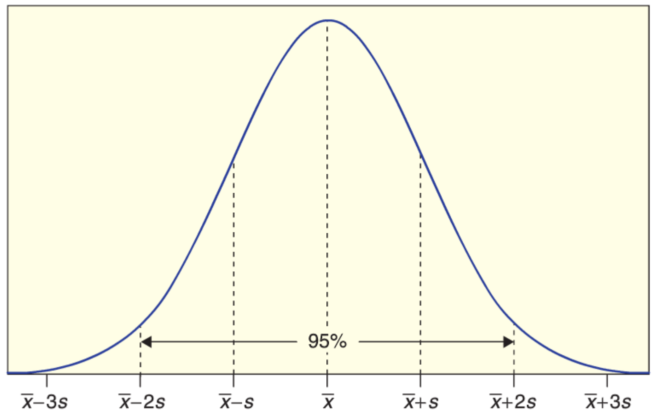
$\overline{x} = 5$

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

$\overline{x} = 4.3$

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

$\overline{x} = 4.2$

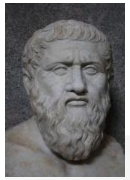


Sampling distribution!

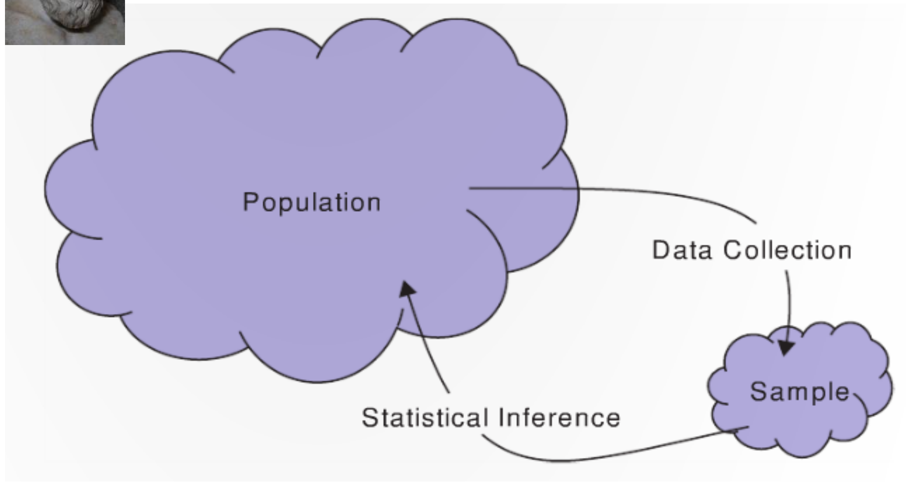
Review: Sampling distribution

Why are we interested in the sampling distribution?

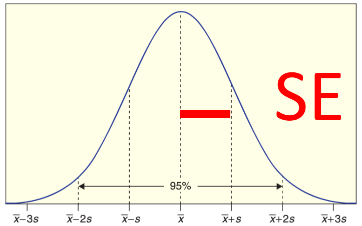
- If we knew what the sampling distribution was, then we could evaluate how much we should trust individual statistics



Parameters: π, μ, σ, ρ, β



Sampling distribution



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

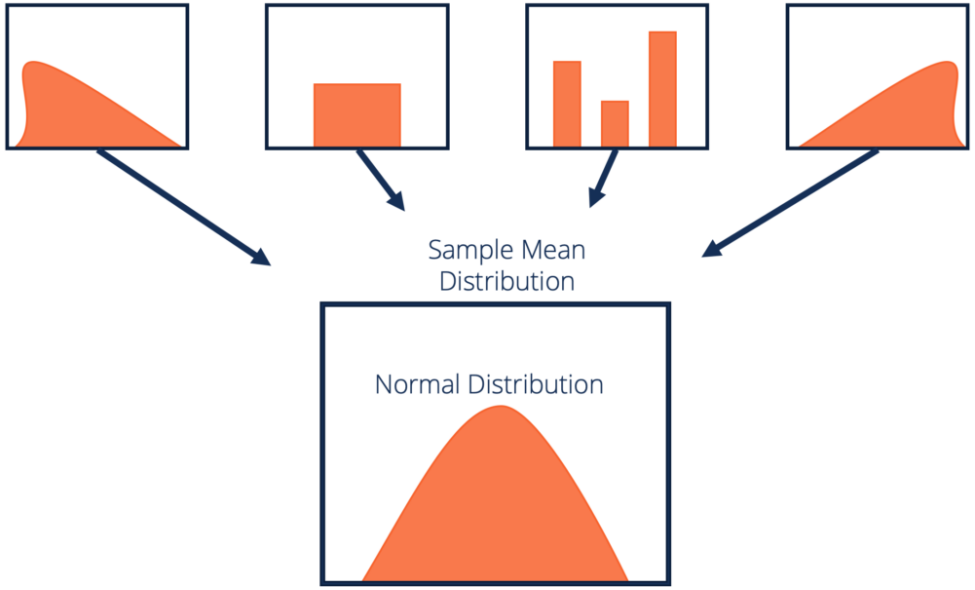
Statistics: $\hat{p}$, $\bar{x}$, s, r, b

Review: 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.

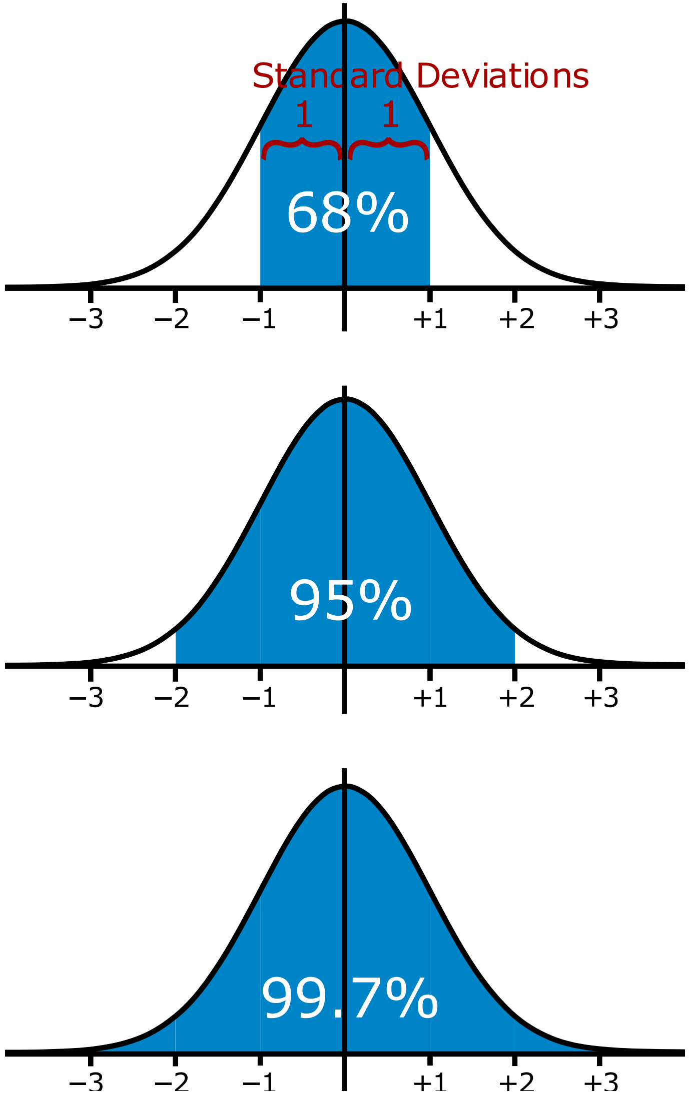
Since many statistics we use are the sum of randomly generated data, many of our sampling distributions will be approximately normal

- You will explore this more on homework 2



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

Normal density function



Review: Simulating a sampling distributions in R

```
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)         # visualize the sampling distribution
SE <- sd(sampling_dist)    # get the standard error
```

Review: Simulating sampling distributions from data

Distribution of OkCupid user's heights n = 100

```
heights <- profiles$height

# get one random sample of heights from 100 people
height_sample <- sample(heights, 100)

# get the mean of this sample
mean(height_sample)
```

Review: Simulating sampling distributions from data

Distribution of OkCupid user's heights n = 100

```
sampling_dist <- NULL
for (i in 1:1000) {
        height_sample <- sample(heights, 100)    # sample 100 random heights
        sampling_dist[i] <- mean(height_sample)    # save the mean
}

hist(sampling_dist)
```


QUESTIONS

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 recent YouGov poll of 2,335 adults showed Biden's approval rating at 40.2%

Symbols:

π: Biden's approval for all voters

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

CBS News Poll – September 5-8, 2023
Adults in the U.S.

YouGov

Sample	2,335 Adults in the U.S.
Margin of Error	±2.7%

1. Generally speaking, do you feel things in America today are going...

Very well	5%
Somewhat well	23%
Somewhat badly	34%
Very badly	38%

2. How would you rate the condition of the national economy today?

Very good	8%
Fairly good	21%
Fairly bad	31%
Very bad	35%
Not sure	5%

3. Do you approve or disapprove of the way Joe Biden is handling his job as president?

Approve	40%
Disapprove	60%

Interval estimate based on a margin of error

An **interval estimate** give a range of plausible values for a <u>population parameter</u>

One common form of an interval estimate is:

Point estimate ± margin of error

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

Example: YouGov poll

40.2% of American approve of Biden's job performance, with a margin of error of 2.7%

- i.e., plus or minus 2.7%

How do we interpret this?

Says that the <u>population parameter</u> (π) lies somewhere between:

40.2 - 2.7 to 40.3 + 2.7 = 37.5 to 42.9

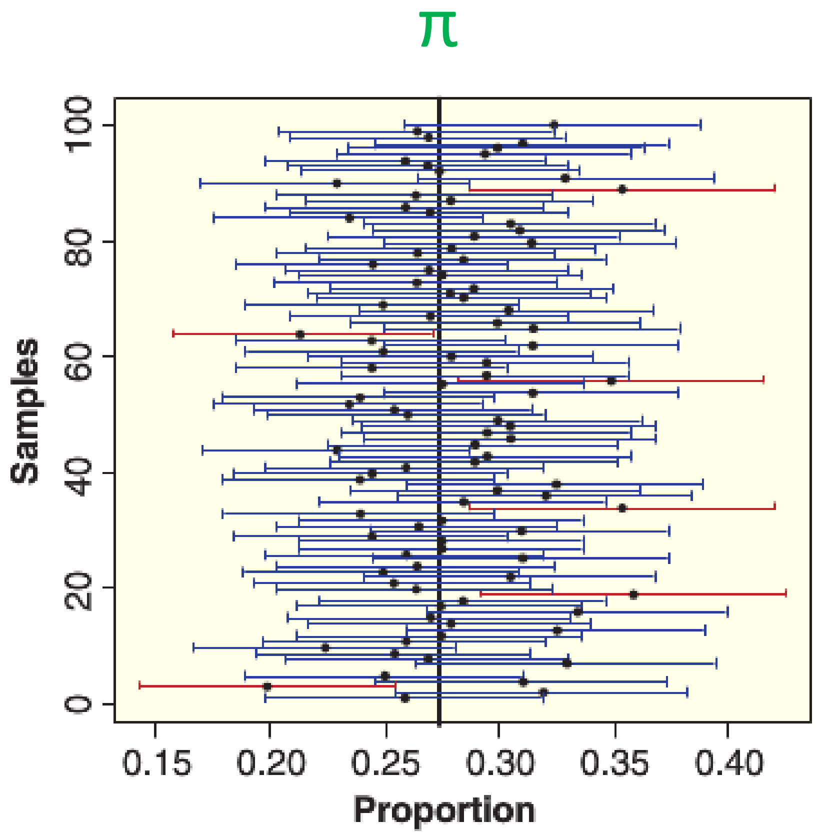
i.e., if they sampled all voters the true population proportion (π) would be likely be in this range

Confidence Intervals

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

- i.e., if the interval was calculated repeatedly from many different random samples, the parameter will be in p% of these intervals

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



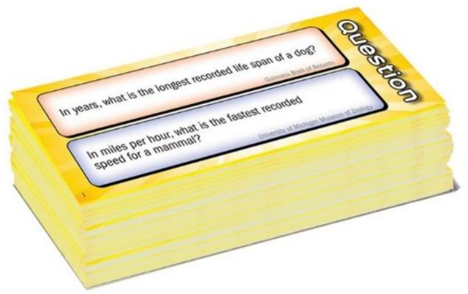
Think ring toss…

Parameter exists in the ideal world

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

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

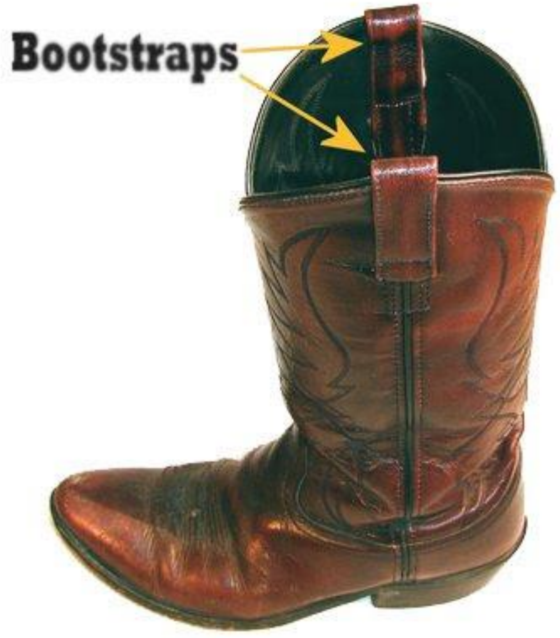
For any given confidence interval we compute, we don't know whether it has really captured the parameter

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

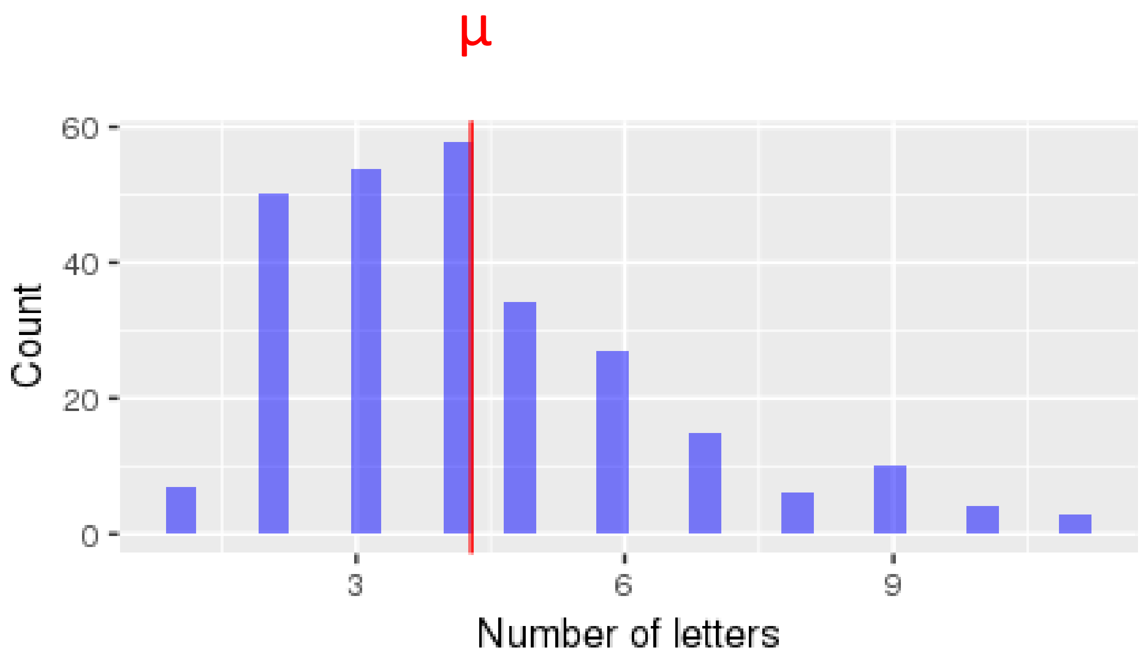
(for a 90% confidence interval)

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

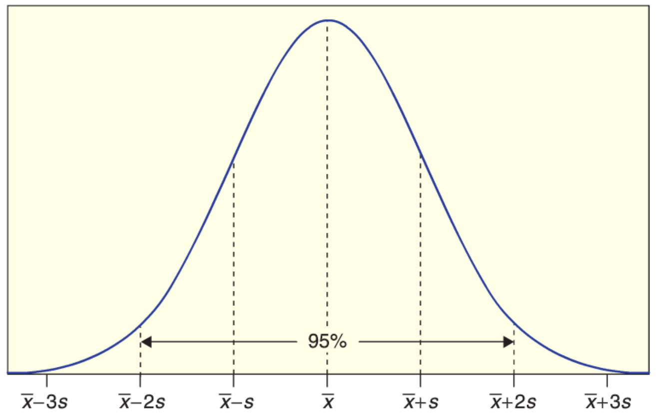
$\bar{x} = 5$

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

$\bar{x} = 4.3$

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

$\bar{x} = 4.2$

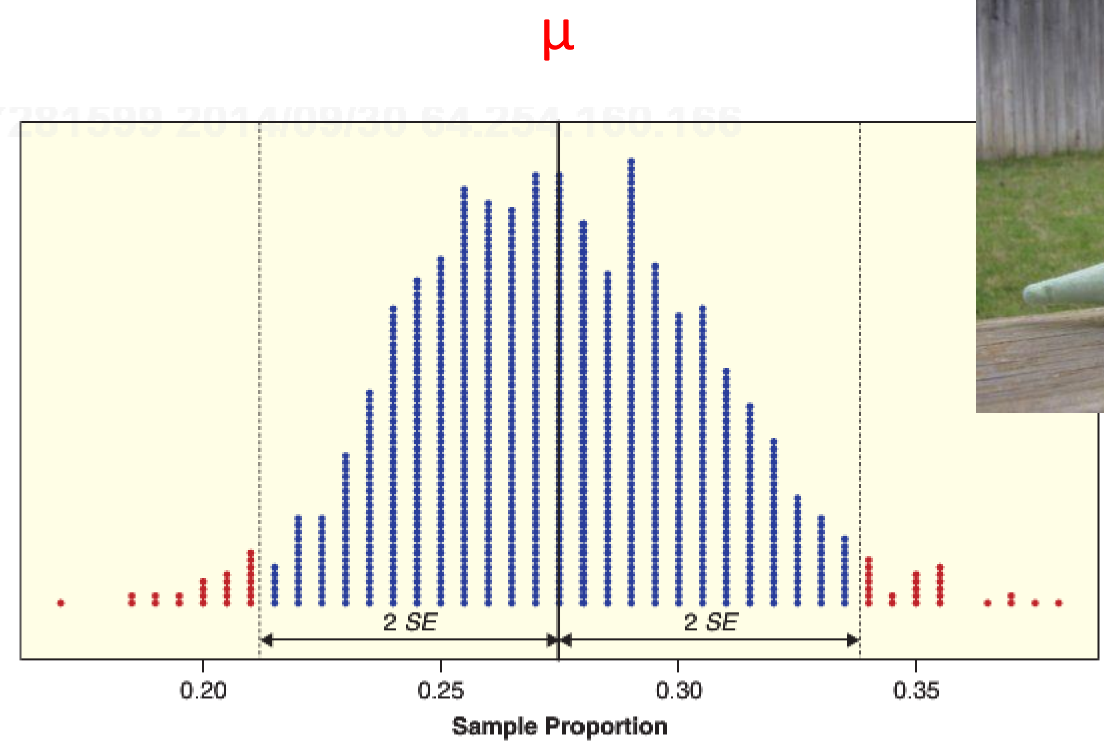


Sampling distribution!

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

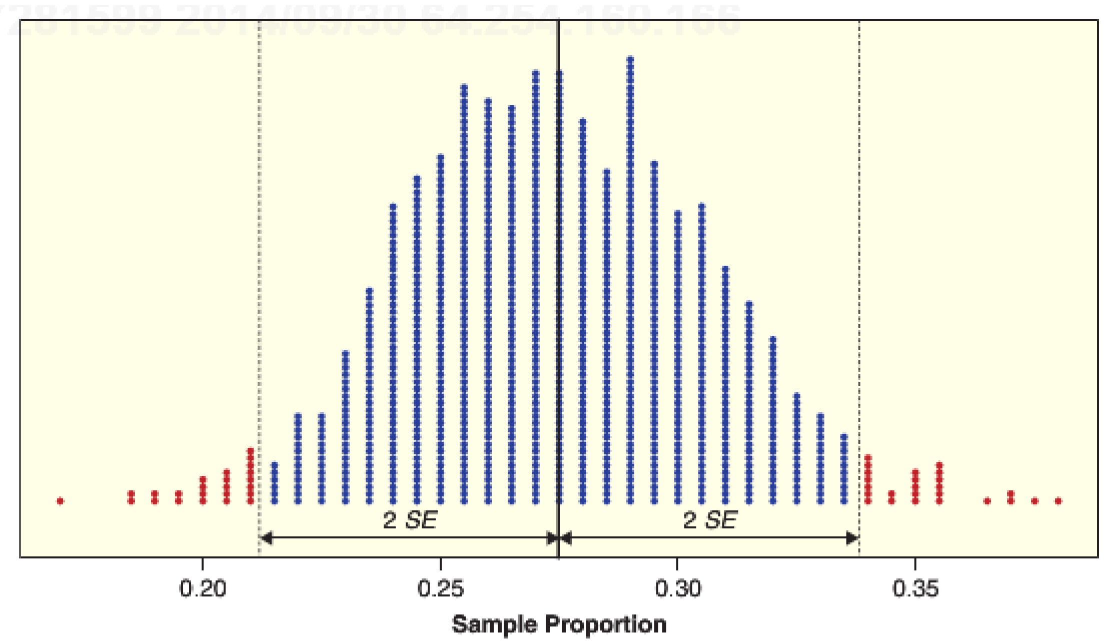
Q: Suppose we had:

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

Could we compute a 95% confidence interval?

A: Yes!

CI = statistic value $\pm$ $2 \cdot$ SE

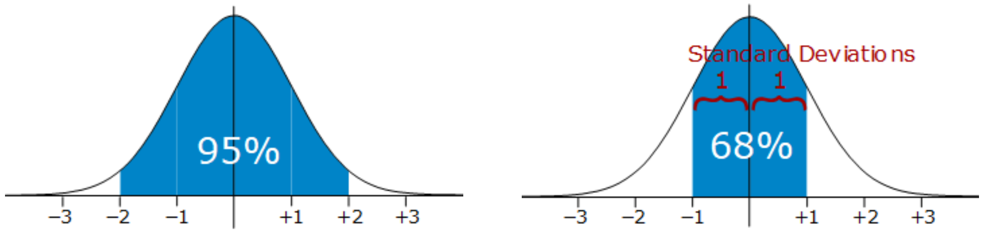


$\underbrace{2 \cdot SE \quad\quad 2 \cdot SE}_{}$

95% confidence interval: stat $\pm$ $2 \cdot$ SE

Confidence interval

Sampling distributions



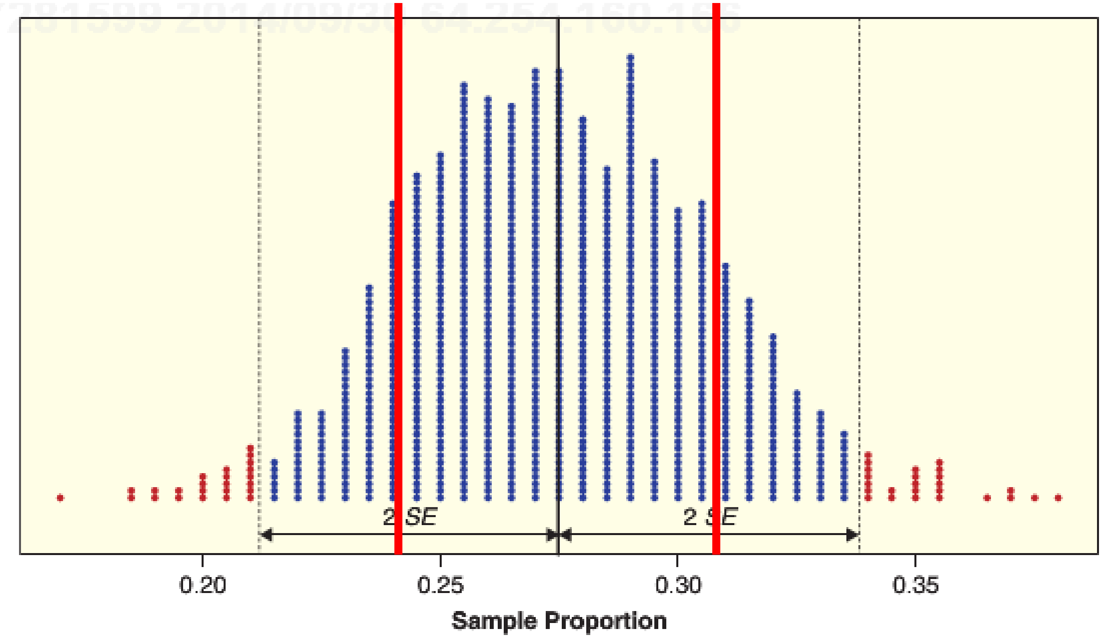
Q: Suppose we had:

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

Could we compute a 95% confidence interval?

A: Yes!

CI = statistic value ± 2 · SE



2 · SE 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

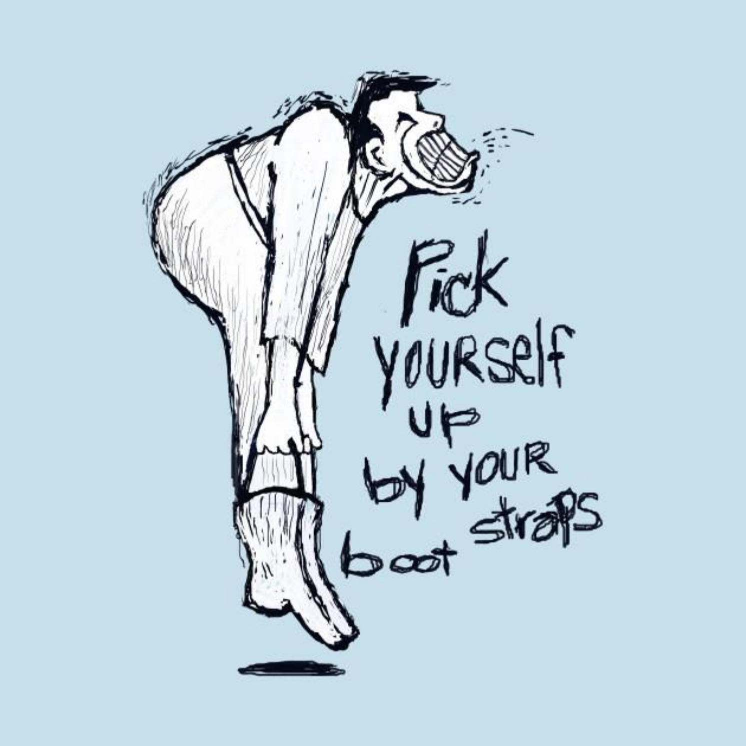
Q: Could we repeat the sampling process many times to create a sampling distribution and then calculate the SE?

- 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 $\overline{x} \pm 2 \cdot \hat{SE}$ to get the 95% CI

Plug-in principle

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

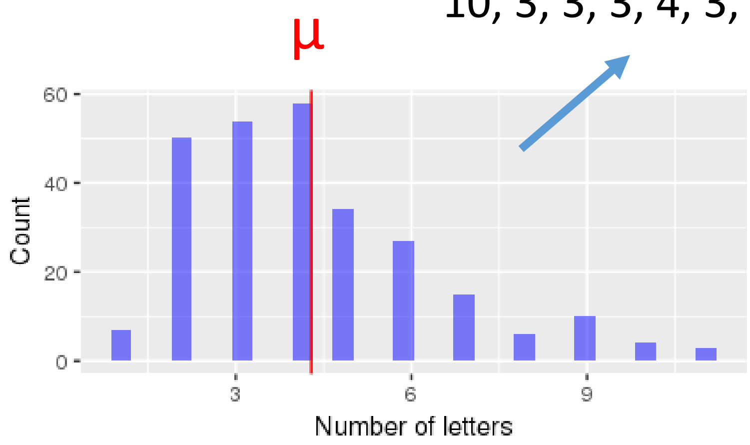
We pretend that *the sample is the population* (plug-in principle)

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



<u>The sample (n = 10)</u>
10, 3, 3, 3, 4, 3, 2, 6, 4, 5

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

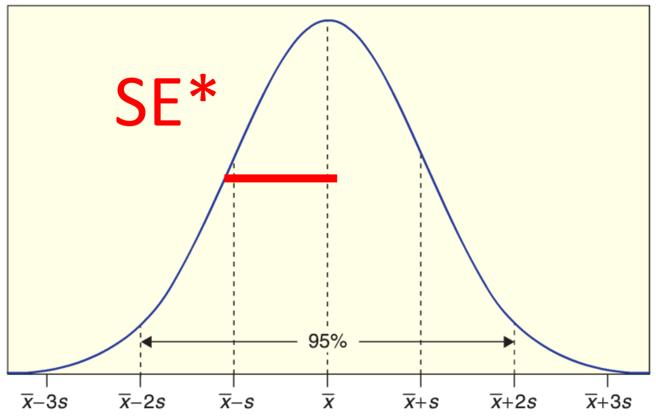
$\overline{x}^* = 4$

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

$\overline{x}^* = 4.1$

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

$\overline{x}^* = 3.9$



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:

$$\textit{Statistic} \;\pm\; 2 \cdot SE^*$$

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:

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad \hat{SE}_{\bar{x}} = \frac{s}{\sqrt{n}}$$

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:

$$SE_{\hat{p}} = \sqrt{\frac{\pi \;\cdot\; (1-\pi)}{n}} \qquad \hat{SE}_{\hat{p}} = \sqrt{\frac{\hat{p} \;\cdot\; (1-\hat{p})}{n}}$$

Where:

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