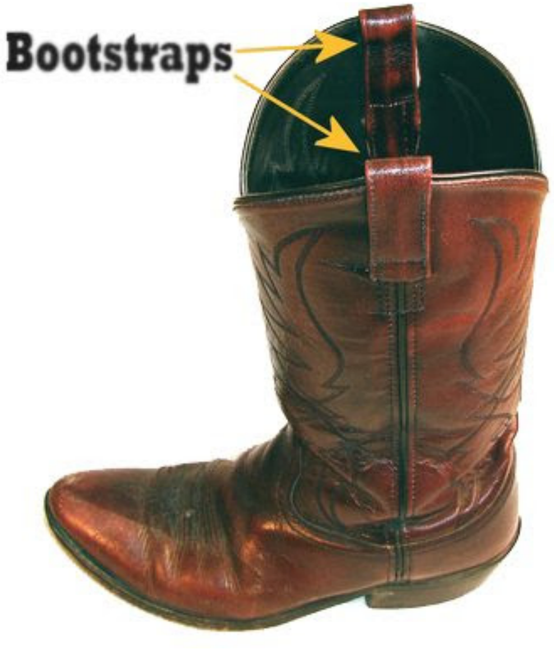
Sampling distributions, confidence intervals, and the bootstrap



Overview

Quick review of probability density functions

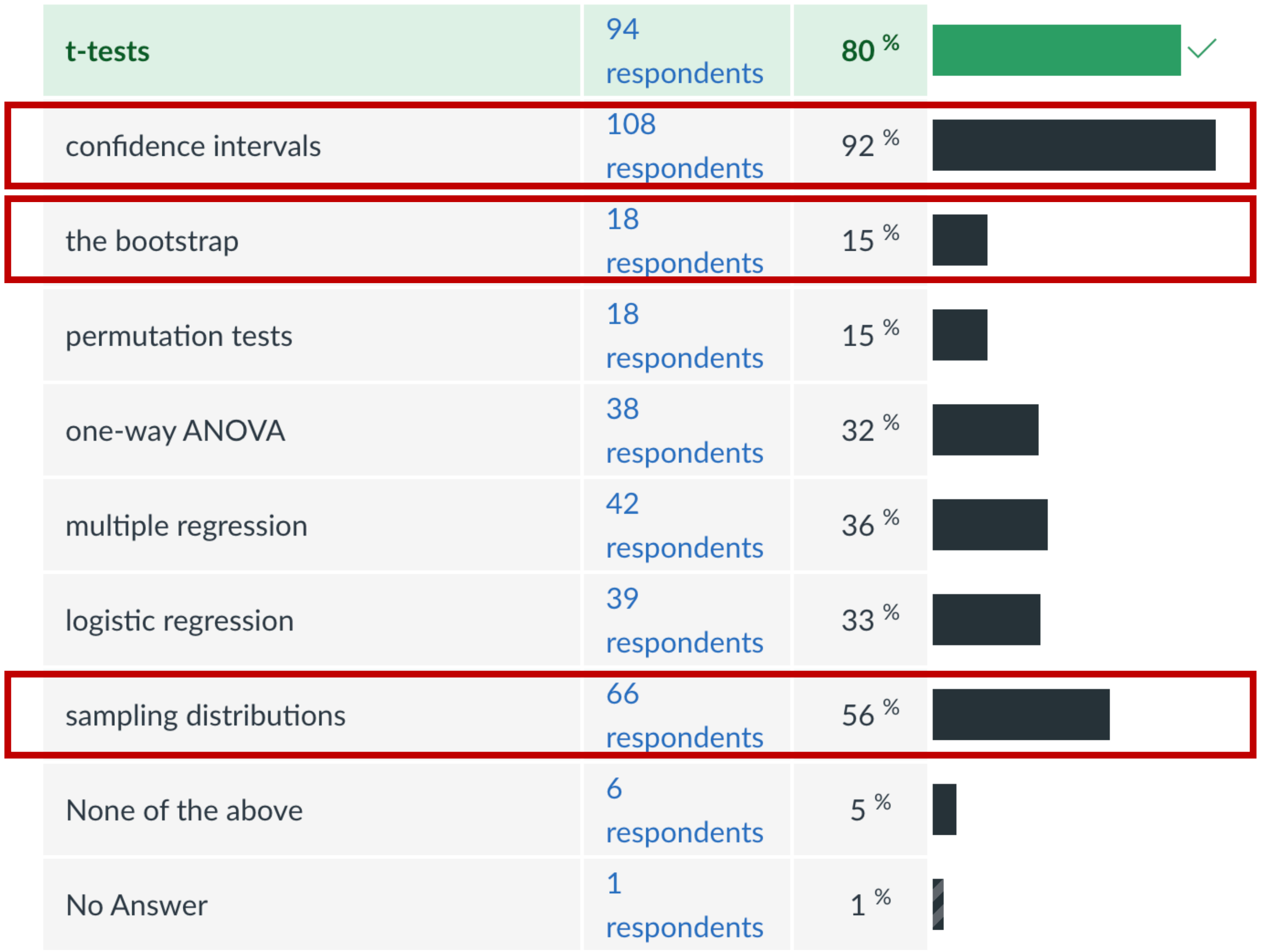
Sampling distributions

Confidence intervals

Computing confidence intervals using the bootstrap

Survey results

Which Statistics methods/concepts are you comfortable with?



t-tests	94 respondents	80 %	✓
confidence intervals	108 respondents	92 %	
the bootstrap	18 respondents	15 %	
permutation tests	18 respondents	15 %	
one-way ANOVA	38 respondents	32 %	
multiple regression	42 respondents	36 %	
logistic regression	39 respondents	33 %	
sampling distributions	66 respondents	56 %	
None of the above	6 respondents	5 %	
No Answer	1 respondents	1 %	

Quick review of density functions

Density Curves

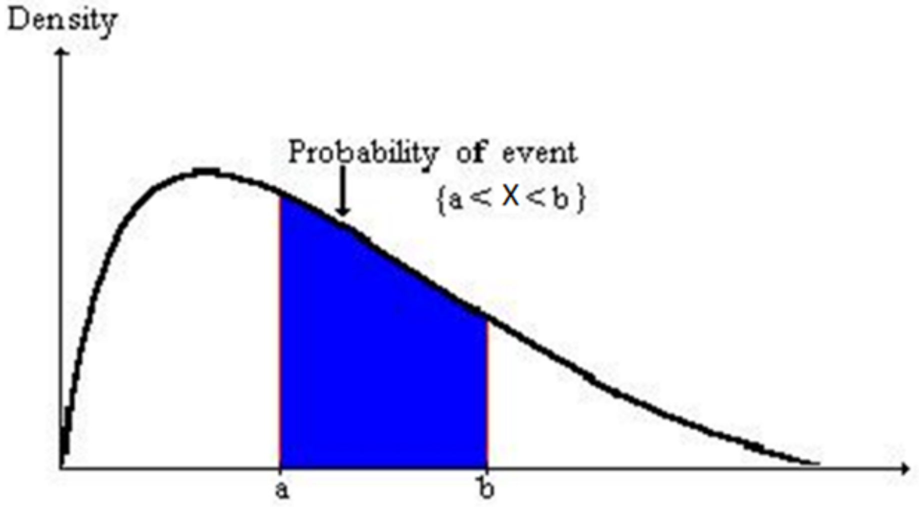
A **density curve** is a mathematical function f(x) that has two important properties:

1. The total area under the curve f(x) is equal to 1

2. The curve is always ≥ 0

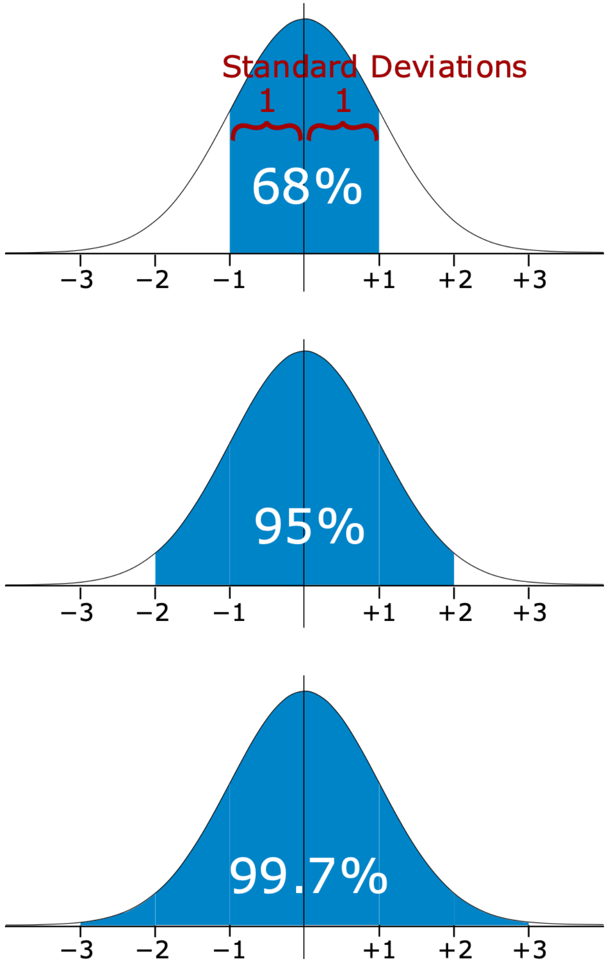
$$P(a < X < b) = \int_a^b f(x)dx$$

The area under the curve in an interval [a, b] models the probability that a random number X will be in the interval

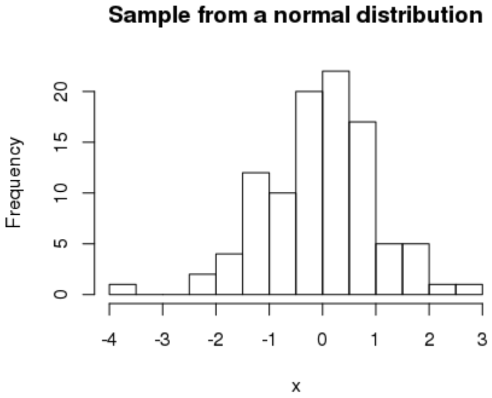


Normal density function

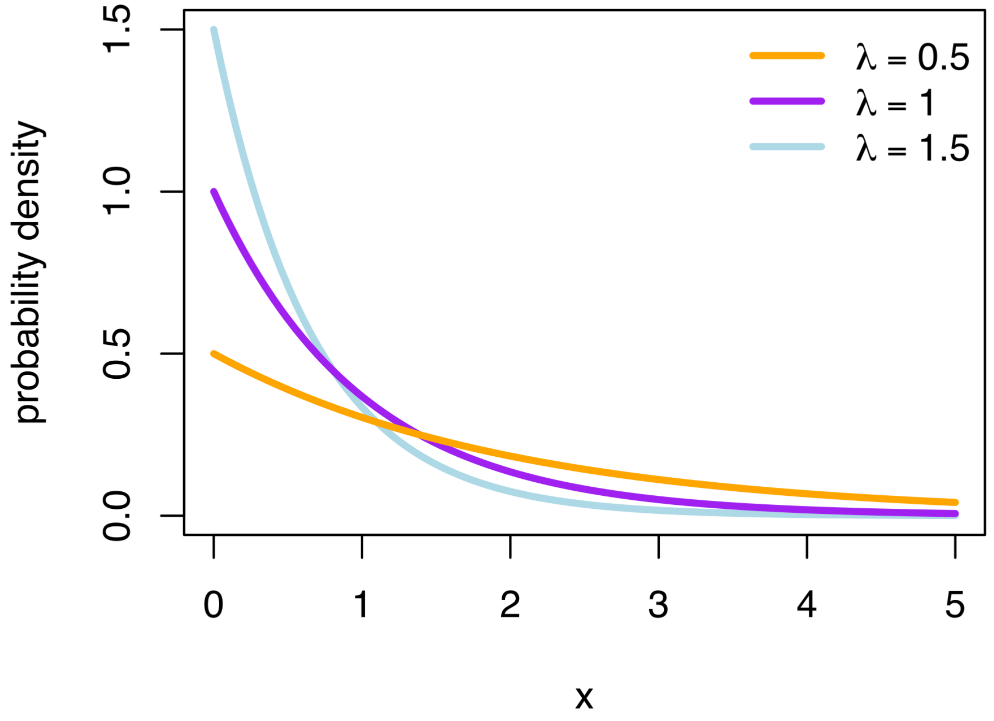
$$f(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



```
dnorm(x, 0, 1)
rand_data <- rnorm(100, 0, 1)
hist(rand_data)
```

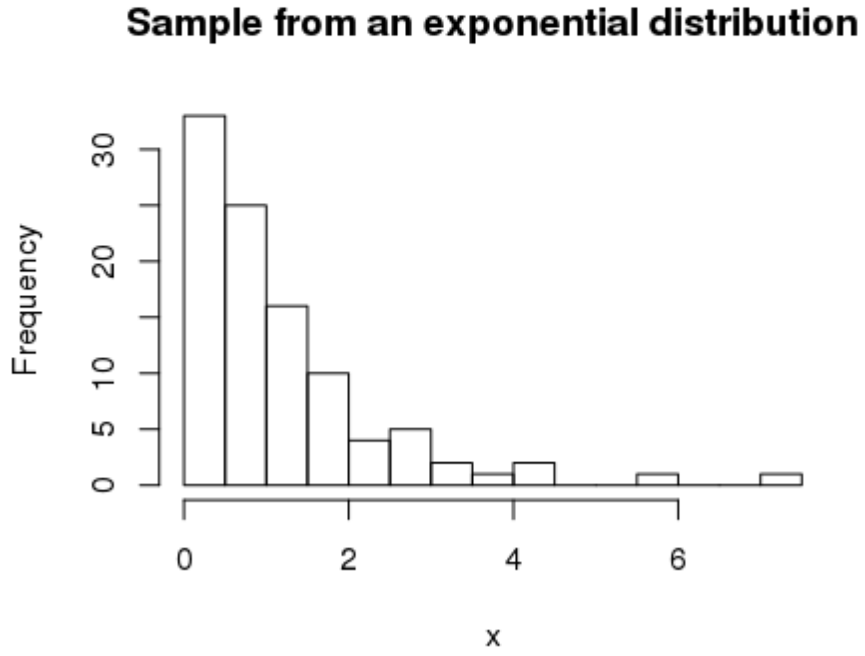


Exponential distribution



$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \ge 0, \\ 0 & x < 0. \end{cases}$$

```
dexp(x, 1)
rand_data <- rexp(100, 1)
hist(rand_data)
```



More review/practice with density functions

There is also quick review of density functions in the class 5 R Markdown document which can be downloaded using:

```
SDS230::download_class_code(5)
```

Sampling distributions

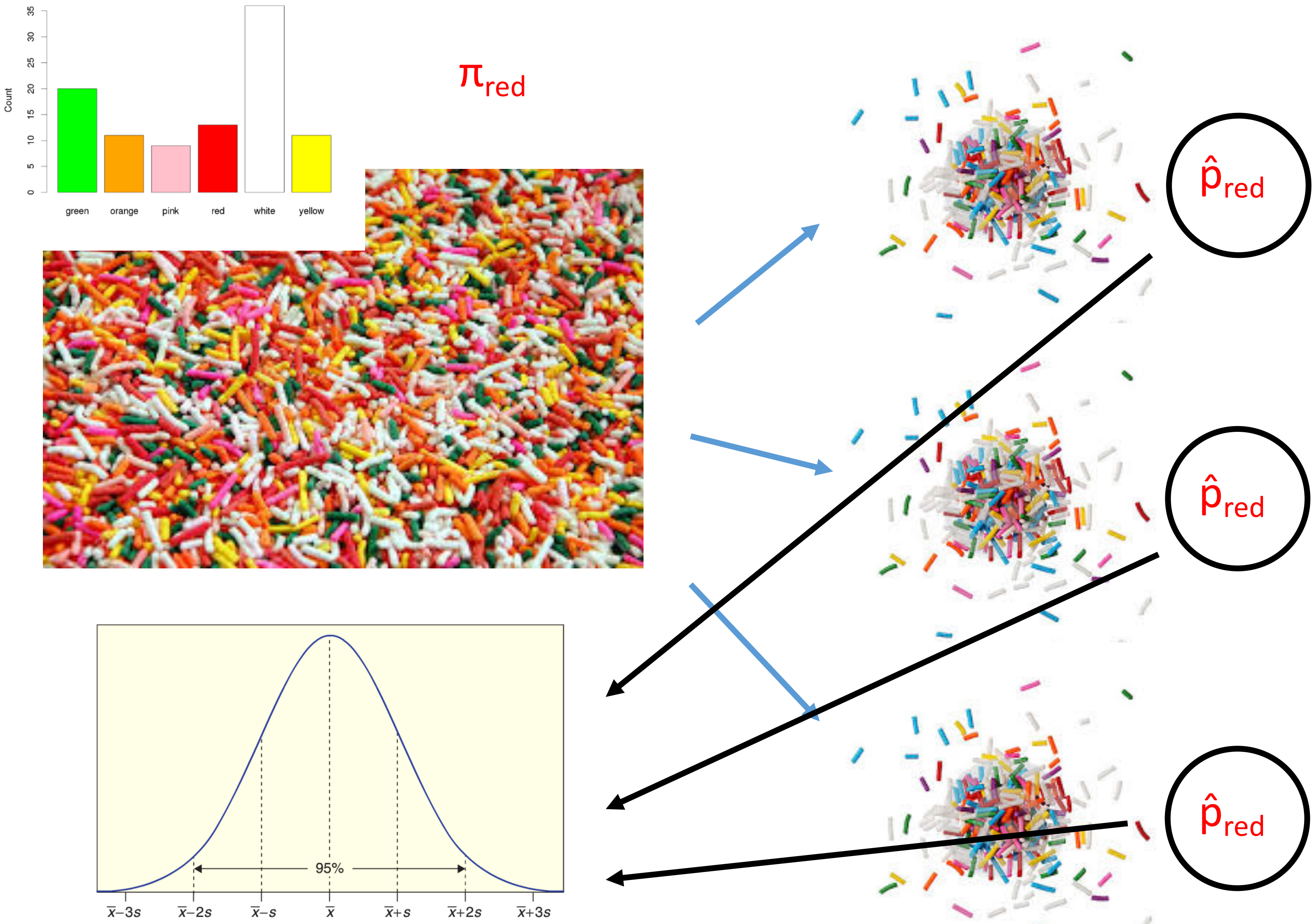
A ***sampling distribution*** is *a* distribution of ***statistics***

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}

$\hat{p}_{red}$

$\hat{p}_{red}$

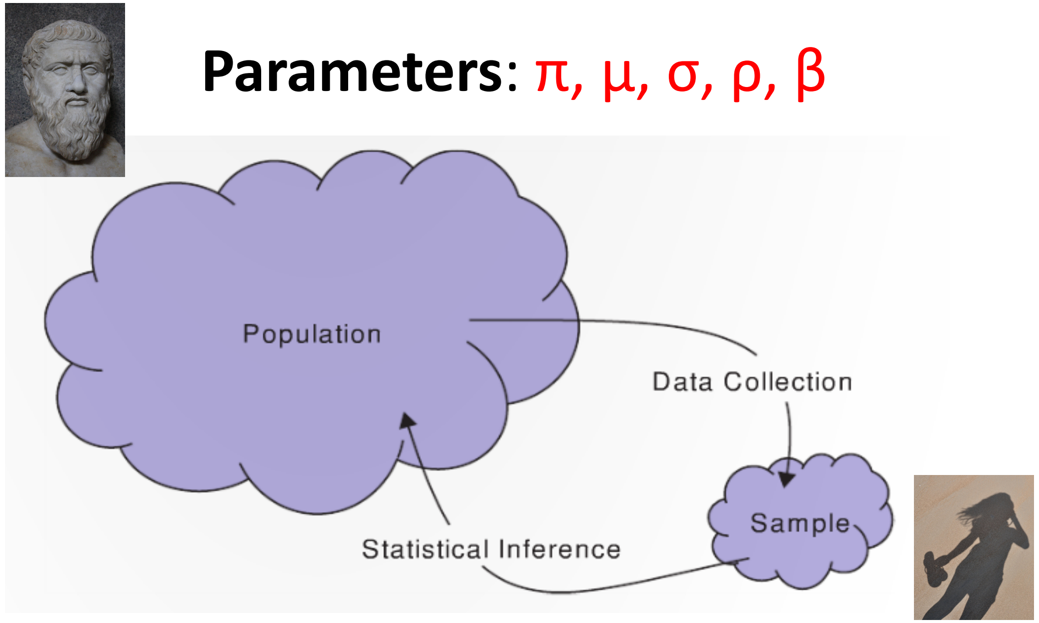
$\hat{p}_{red}$

Sampling distribution!

Sampling distribution

Why would we be interested in the sampling distribution?

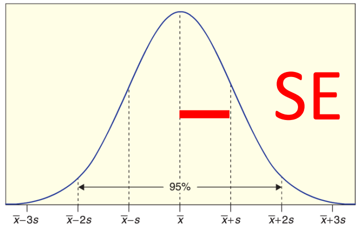
- If we knew what the sampling distribution was, then we could evaluate how much we should trust individual statistics



Parameters: π, μ, σ, ρ, β

Statistics: p̂, $\bar{x}$, s, r, b

Sampling distribution



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

Sampling distributions in R

Sampling distributions

```
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)
```

Sampling distributions

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)
```

Sampling distributions

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)
```

Confidence intervals

Survey question 0: Which of the following are true statements about 90% confidence intervals?

1. ~90% of such intervals contain the statistic of interest

2. ~90% of such intervals contain the parameter of interest

3. For a given confidence interval, there is a 90% probability that the interval contains the parameter

Point Estimate

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

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

A Gallup poll in August listed Trump's approval rating at 42% for likely voters

Symbols:

π: Trump's approval for all voters

$\hat{p}$: Trump'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: Fox news poll

44% of American approve of Trump's job performance, plus or minus 3%

How do we interpret this?

Says that the population parameter (π) lies somewhere between 41% to 47%

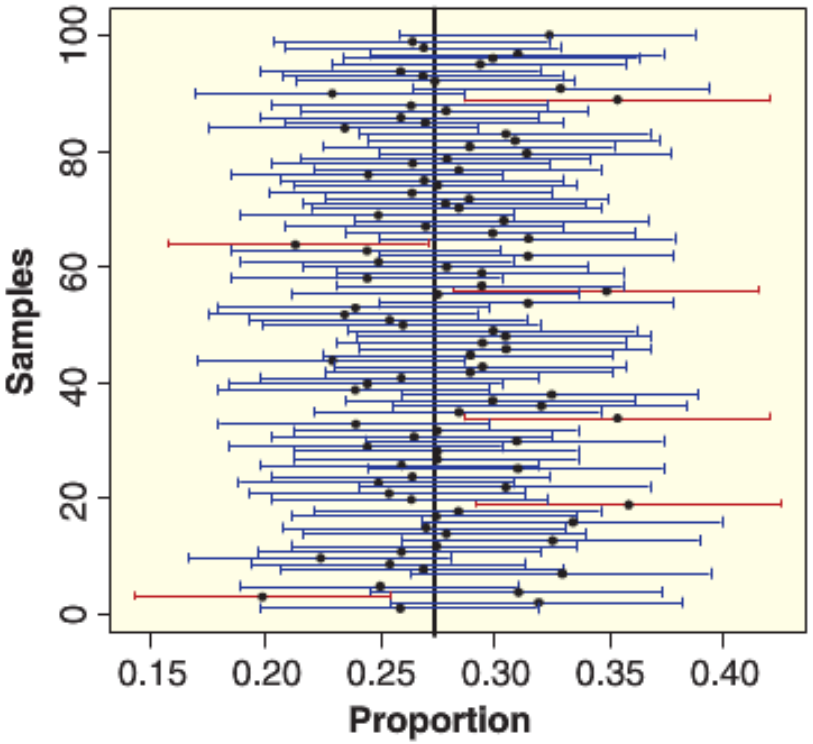
i.e., if they sampled all voters the true population proportion (π) would be likely be in this range

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



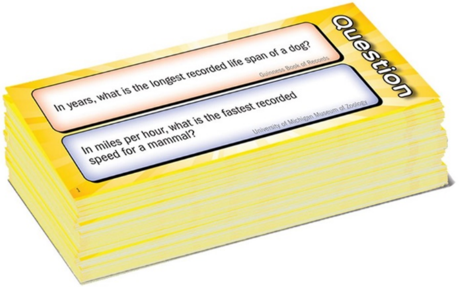
Think ring toss…

Parameter exists in the ideal world

We toss intervals at it

95% of those intervals capture the parameter

Wits and Wagers: 90% confidence intervals estimators



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

Guinness Book of Records

In miles per hour, what is the fastest recorded speed for a mammal?

University of Michigan Museum of Zoology

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 do the following:

- 1. Write down your answers on a piece of paper
- 2. Enter your answers on the class survey
    - Enter as two numbers separated by a comma: 2, 90
- 3. Indicate if you are willing to share your answers anonymously

Bring your answers to class on Thursday to see which questions got right

- And whether you are indeed a 90% confidence interval estimator

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: In feet, how tall is The Statue of Liberty including the pedestal?

Wits and Wagers…

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

Question 5: In pounds, what was the weight of the heaviest domesticated cat ever recorded?

Question 6: In years, what is the longest recorded life span of a dog?

Question 7: How many pounds does one gallon of whole milk weigh?

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?

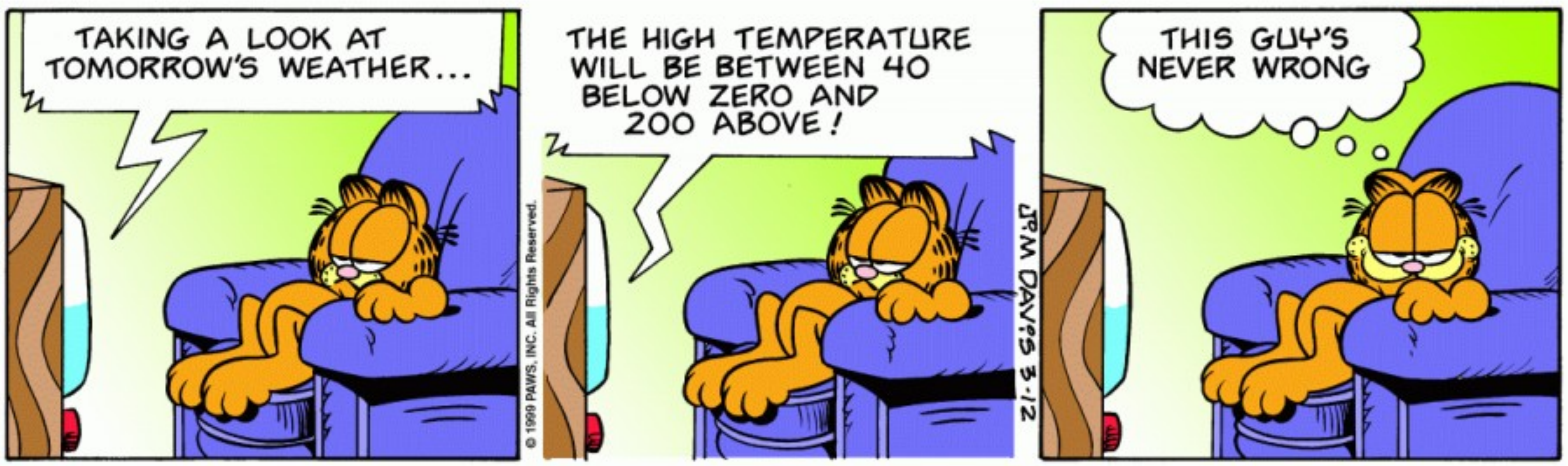
Question 9: What percentage of American adults say that reading is their favorite leisure-time activity?

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

Answers

Come to class on Thursday to see the answers!

100% confidence intervals


TAKING A LOOK AT TOMORROW'S WEATHER...
THE HIGH TEMPERATURE WILL BE BETWEEN 40 BELOW ZERO AND 200 ABOVE!
THIS GUY'S NEVER WRONG
© 1999 PAWS, INC. All Rights Reserved.
JIM DAVIS 3-12

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

Note

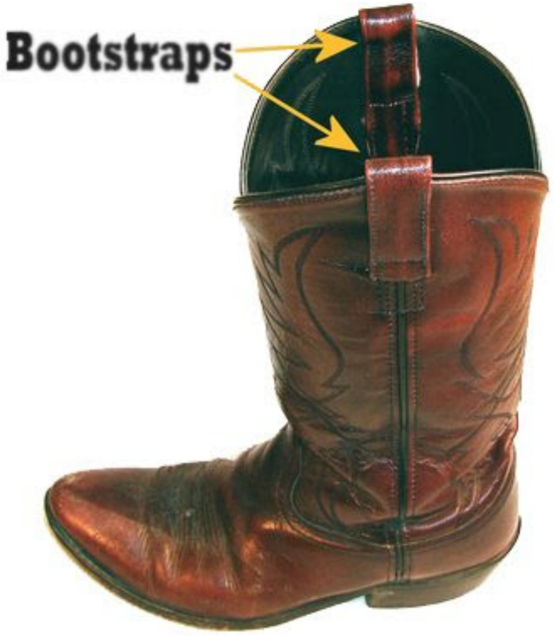
For any given confidence interval we compute, we don’t know whether it has really captured the parameter

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

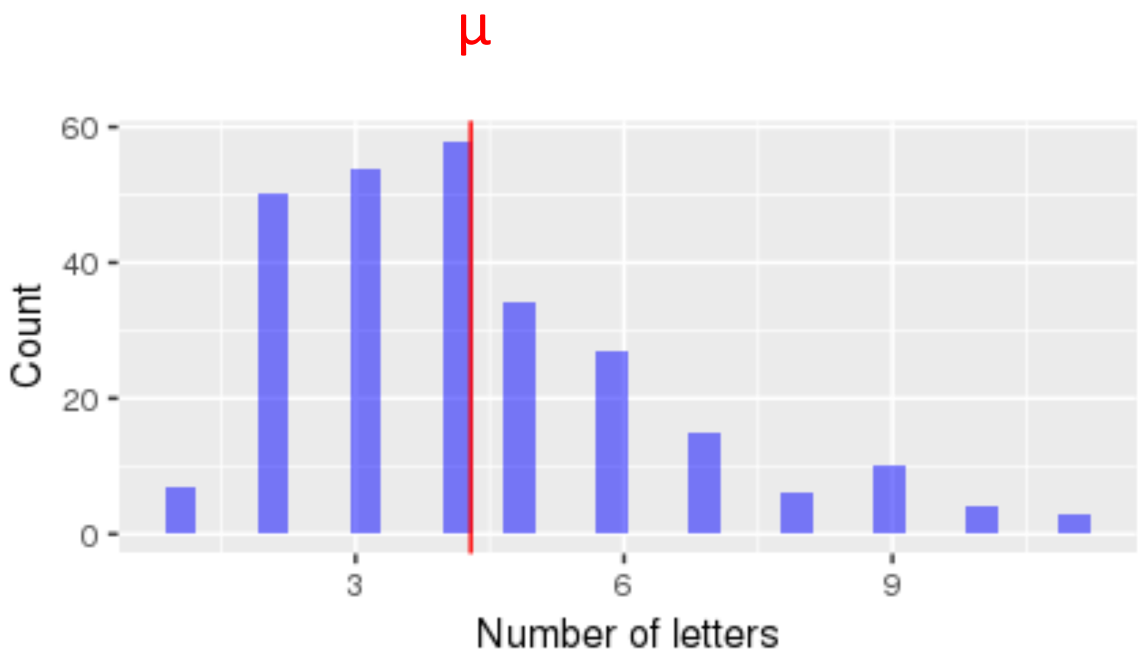
(for a 90% confidence interval)

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

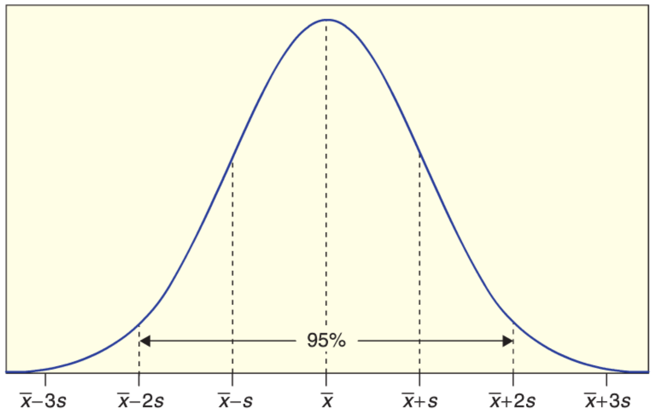
$\bar{x} = 5$

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

$\bar{x} = 4.3$

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

$\bar{x} = 4.2$

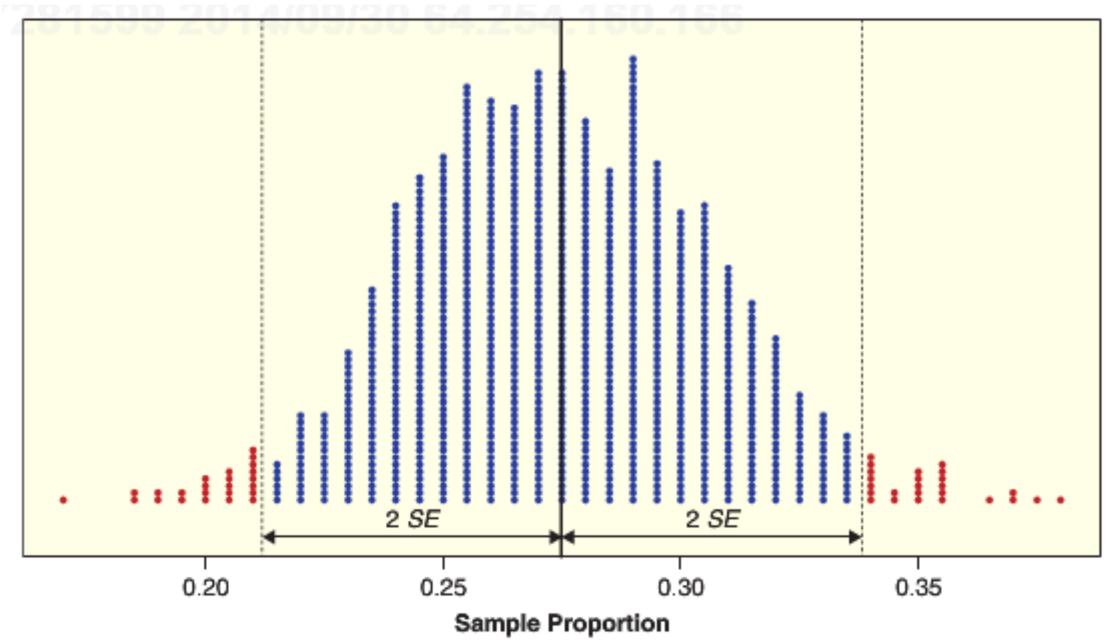


Sampling distribution!

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

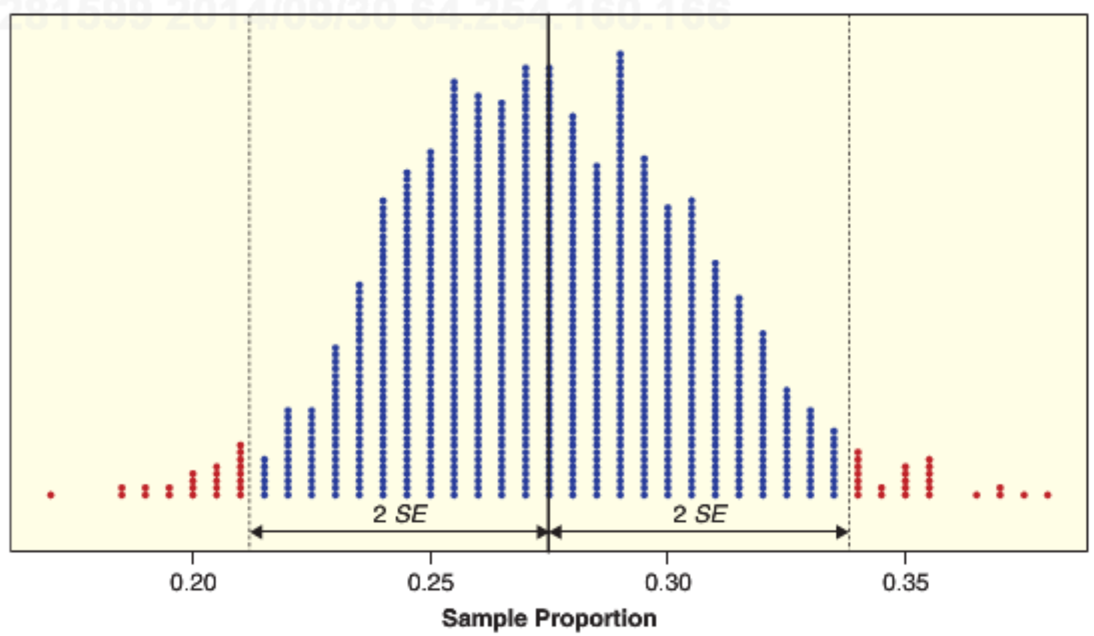
Q: If we had:

- A statistics value
- The SE

Could we compute a 95% confidence interval?

A: Yes!

CI = statistic value ± 2 · SE



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 from a population of size *n*

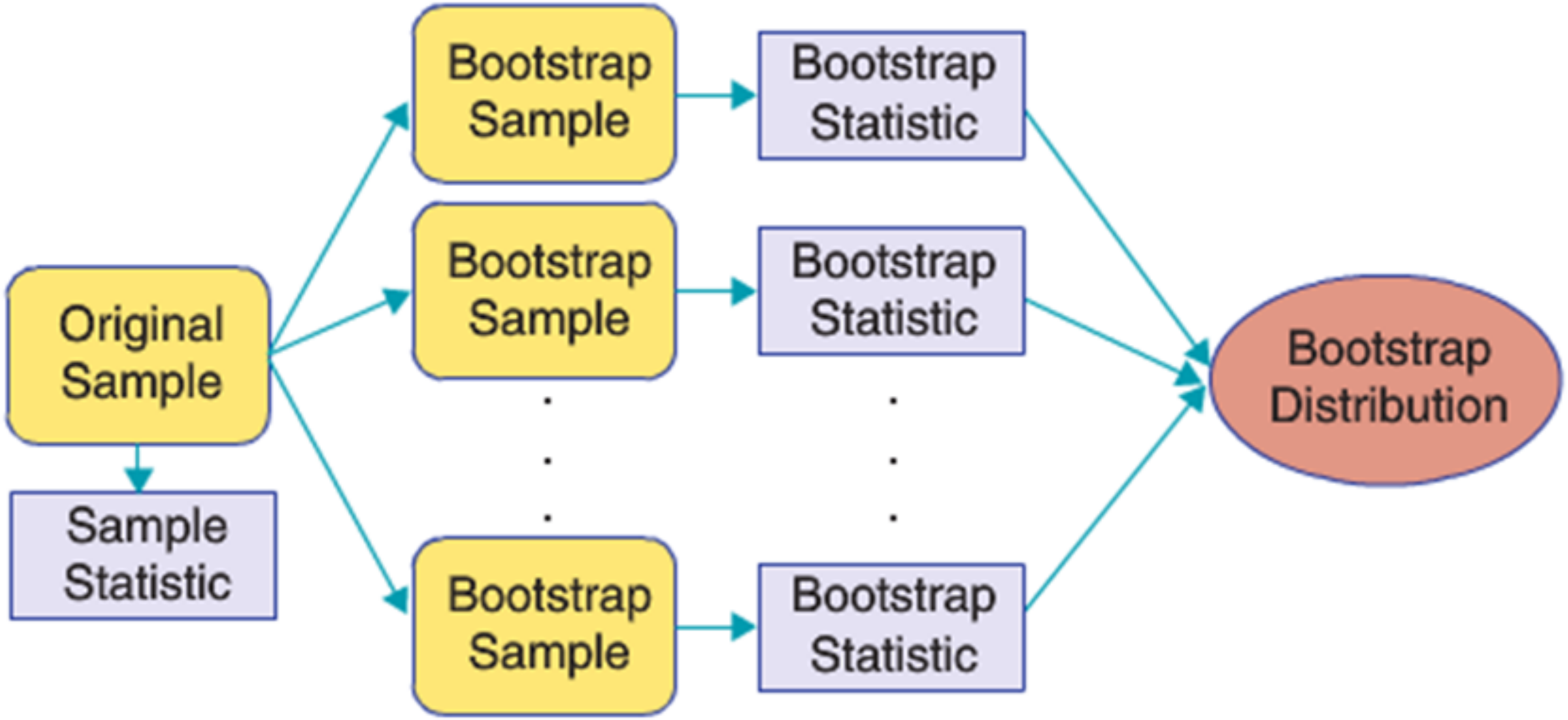
We pretend that *the sample is the population* (plug-in principle)

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 process


Original Sample

Sample Statistic

Bootstrap Sample

Bootstrap Sample

Bootstrap Sample

Bootstrap Statistic

Bootstrap Statistic

Bootstrap Statistic

Bootstrap Distribution

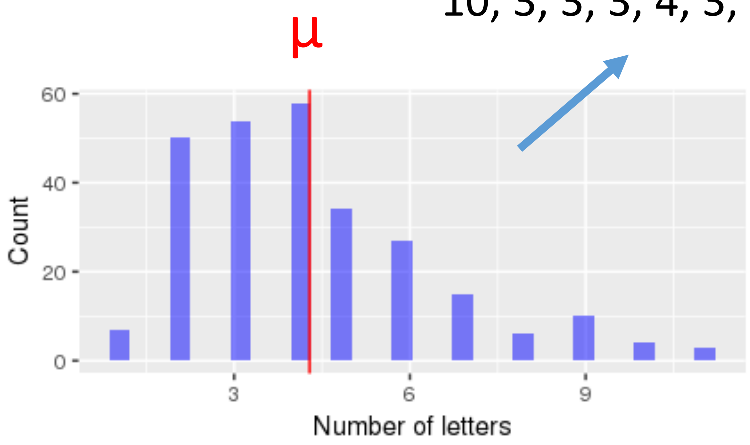
95% Confidence Intervals

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

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

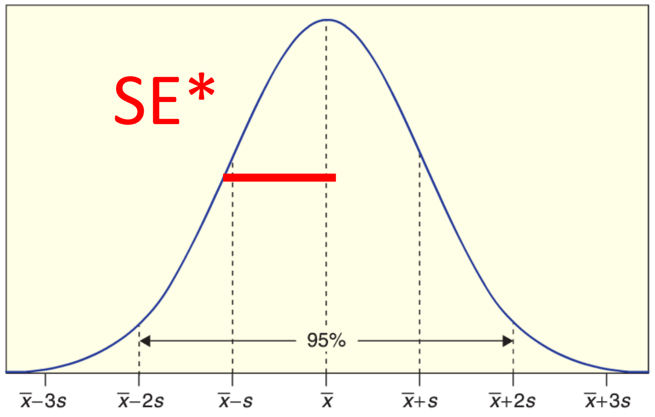
$\overline{x}^* = 4$

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

Let’s try it in R…