Inference Review

Confidence Intervals

Prof. Maria Tackett

01.13.20

Click for PDF of slides

Announcements

- Complete surveys and consent form by Wed at 11:59p
- Reading for today & Wednesday
- Office hours start today. See course homepage for schedule
- **New to R or need a refresher?**
  - Attend an Intro to R workshop
  - Offered Jan 13 - 16, 6p - 7:30p, Gross Hall 270 (choose 1 night to attend)
  - Click here to sign up
- Find more info about statistics related events on Sakai

Any questions from last class?

Today's Agenda

- Sampling distributions & the Central Limit Theorem
- Calculating confidence intervals

Sesame Street

- *Sesame Street* is a television series designed to teach children ages 3-5 basic education skills such as reading (e.g. the alphabet) and math (e.g. counting)
- Today we are going to analyze data from an study conducted by the Educational Testing Service in the early 1970s to test the effectiveness of the program.

STA 210

Sesame Street study

- Children from 6 locations around the United States (including Durham!) participated in the study. The children were split into two groups (`treatment`):
    - **Group 1**: Those who were encouraged to watch the show (assume watched regularly)
    - **Group 2**: Those who didn't get encouragement to watch the show (assume didn't watch regularly)
- Each child was given a test before and after the study to measure their knowledge of basic math, reading, etc.
- We will focus on the change in reading (identifying letters) scores (`change`)

Sesame Street Data - Full Description Original Study: *Ann Bogatz, Gerry & Ball, Samuel. (1971). The Second Year of Sesame Street: A Continuing Evaluation. Volume 1. vols. 1 & 2.*

Let's look at the data

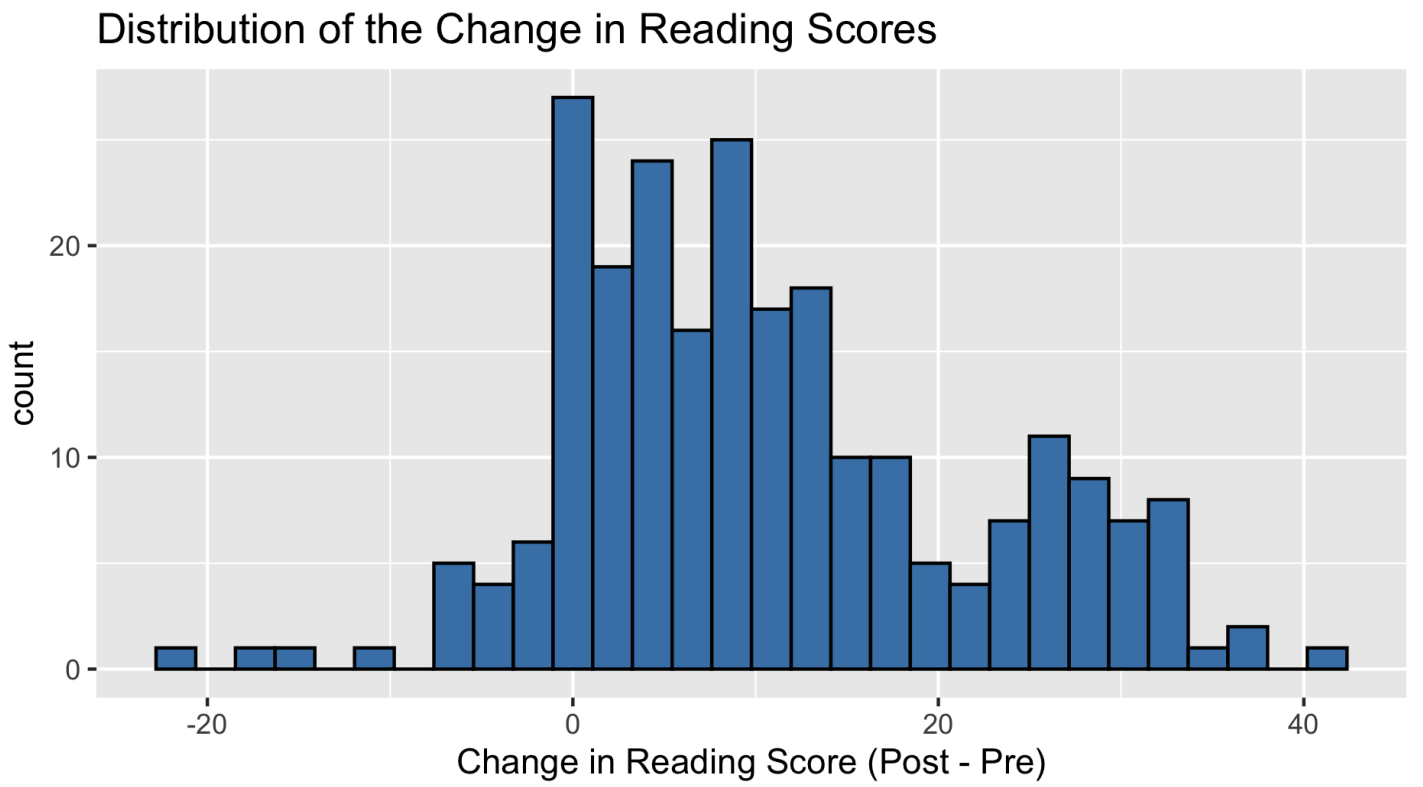
`sesame_street.csv` is available in the datasets repo on GitHub.

```
sesame_street %>%
  slice(1:10)
```

```
## # A tibble: 10 x 4
##    treatment      prelet postlet change
##    <chr>           <dbl>   <dbl>  <dbl>
##  1 Encouraged         23      30      7
##  2 Encouraged         26      37     11
##  3 Not Encouraged     14      46     32
##  4 Not Encouraged     11      14      3
##  5 Not Encouraged     47      63     16
##  6 Not Encouraged     26      36     10
##  7 Not Encouraged     12      45     33
##  8 Encouraged         48      47     -1
##  9 Encouraged         44      50      6
## 10 Encouraged         38      52     14
```

Exploratory Data Analysis - Univariate

```
ggplot(data = sesame_street, mapping = aes(x = change)) +
  geom_histogram(fill = "steelblue", color = "black") +
  labs(x = "Change in Reading Score (Post - Pre)" ,
       title = "Distribution of the Change in Reading Scores")
```



Exploratory Data Analysis - Univariate

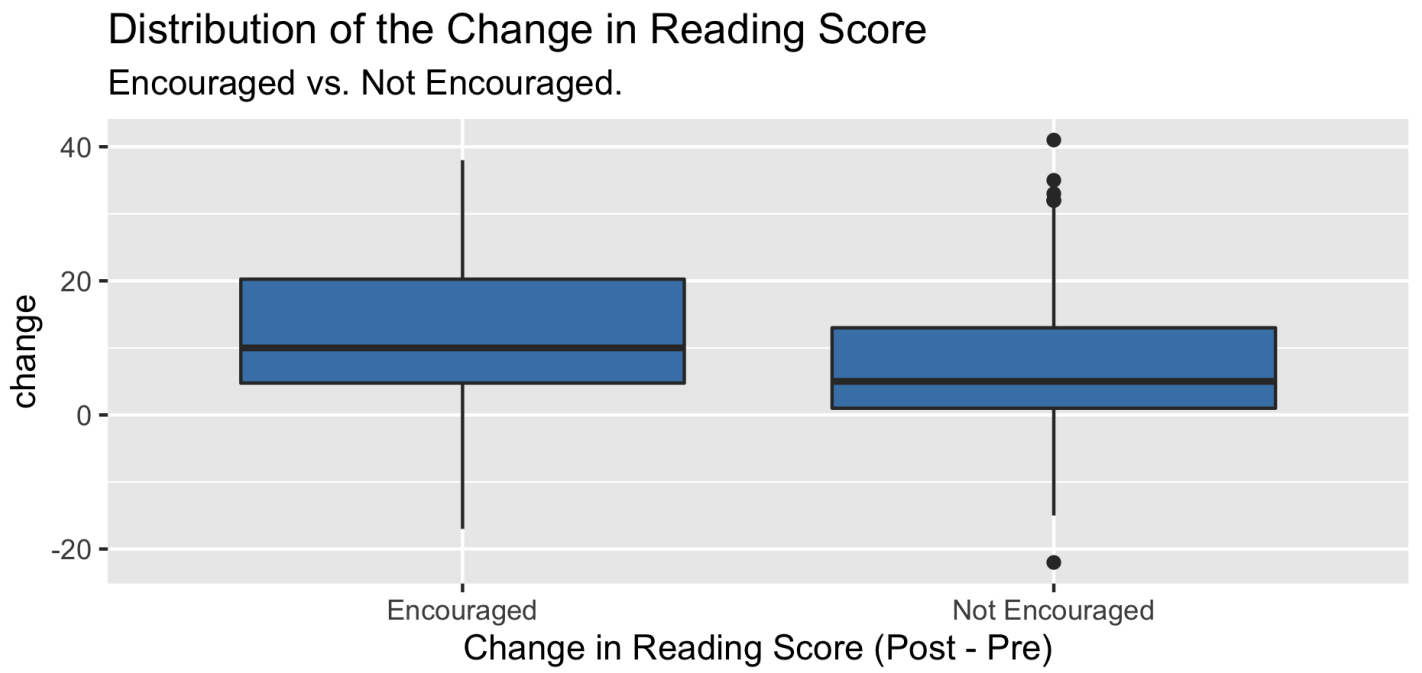
- Calculate summary statistics for change

```
sesame_street %>%
  summarise(n = n(), min = min(change), median = median(change), m
            IQR = IQR(change),
            mean = mean(change), std_dev = sd(change))
```

```
## # A tibble: 1 x 7
##       n   min median   max   IQR  mean std_dev
##   <int> <dbl>  <dbl> <dbl> <dbl> <dbl>   <dbl>
## 1   240   -22      9    41    15  10.8    11.2
```

Exploratory Data Analysis - Bivariate

```
ggplot(data = sesame_street,
       mapping = aes(y = change, x = treatment)) +
  geom_boxplot(fill = "steelblue") +
  labs(x = "Change in Reading Score (Post - Pre)",
         title = "Distribution of the Change in Reading Score",
       subtitle = "Encouraged vs. Not Encouraged.")
```



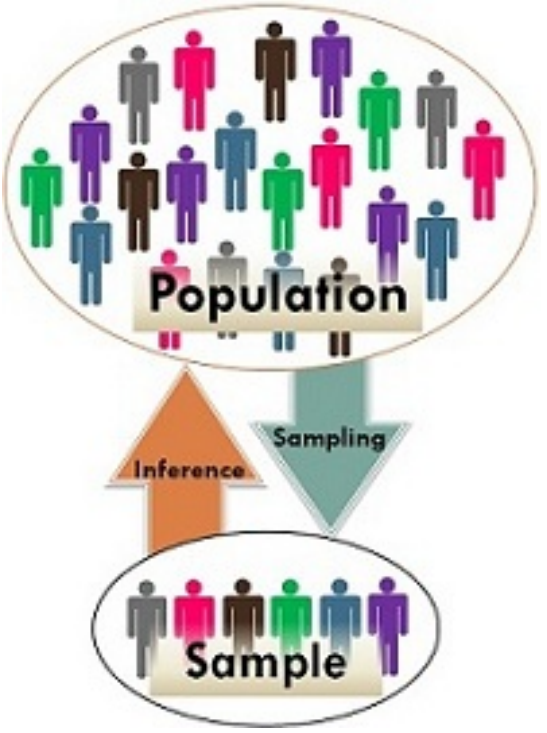
Exploratory Data Analysis - Bivariate

Calculate summary statistics for `change` for each group of `treatment`

```
sesame_street %>%
  group_by(treatment) %>%
  summarise(n = n(), min = min(change), median = median(change),
            max = max(change), IQR = IQR(change),
            mean = mean(change), std_dev = sd(change))
```

```
## # A tibble: 2 x 8
##   treatment           n   min median   max   IQR  mean std_dev
##   <chr>           <int> <dbl>  <dbl> <dbl> <dbl> <dbl>   <dbl>
## 1 Encouraged        152   -17     10    38  15.5 12.5     10.7
## 2 Not Encouraged     88   -22      5    41  12    7.88    11.4
```

What is statistical inference?



- **Statistical inference** is the process of using sample data to make conclusions about the underlying population from which the sample was taken
- Types of inference:
  - **Confidence Intervals:** Estimate the parameter of interest
  - **Hypothesis Tests:** Test a specified claim or hypothesis

Image source: https://keydifferences.com/difference-between-population-and-sample.html

Confidence Interval for a Population Mean

Confidence Intervals

- Developed by Jerzy Neyman (in the 1930s)
- **What**: Plausible range of values for a population parameter
  - Assuming sample data is a random sample from the population
- **Why**: Because the statistic is a random variable, its value is subject to chance error, i.e. random variability
  - We want to take that variability into account by reporting a range of plausible values the parameter can take rather than solely relying on a single statistic

Let's think about the *Sesame Street* data

- We want to know the true mean change in reading scores for all children in the U.S. ages 3 - 5 after 26 weeks (the length of the study). This is the **population parameter**.

- We aren't able to collect data on all children in the U.S. ages 3 - 5, but we do have data on 240 children who participated in the study. This is the **sample statistic**.

Let's think about the *Sesame Street* data

- Our best guess for the true mean change in reading scores is the mean change in reading scores from our sample: 10.8

- If we redid the study using 240 different children, we'd expect the mean change in reading score in that sample to be differ from 10.8. This is sampling variability.

- Our goal, then is to account for that sampling variability and calculate a plausible range of values the true mean can take.

Sampling distribution

- A **sampling distribution** is the distribution of sample statistics from random samples of of the same size taken with replacement from a population
- In practice it is impossible to construction sampling distributions, since it would require having access to the entire population. However, we have theorems that tell us what the sampling distribution will look like (more on this shortly.)
- For now, let's do a couple of demonstrations to get an idea about some basic properties of sampling distributions.
    - For the demonstration, we will make the unrealistic assumption that we have access ot the population and will construct the sample distribution.

The population

```
set.seed(011320)
norm_pop <- tibble(x = rnorm(n = 100000, mean = 20, sd = 3))
ggplot(data = norm_pop, aes(x = x)) +
  geom_histogram(binwidth = 1) +
  labs(title = "Population distribution")
```

Sampling from the population - 1

```
samp_1 <- norm_pop %>%
  sample_n(size = 50, replace = TRUE)
```

```
samp_1 %>%
  summarise(x_bar = mean(x))
```

```
## # A tibble: 1 x 1
##   x_bar
##   <dbl>
## 1  20.1
```

Sampling from the population - 2

```
samp_2 <- norm_pop %>%
  sample_n(size = 50, replace = TRUE)
```

```
samp_2 %>%
  summarise(x_bar = mean(x))
```

```
## # A tibble: 1 x 1
##   x_bar
##   <dbl>
## 1  19.7
```

Sampling from the population - 3

```
samp_3 <- norm_pop %>%
  sample_n(size = 50, replace = TRUE)
```

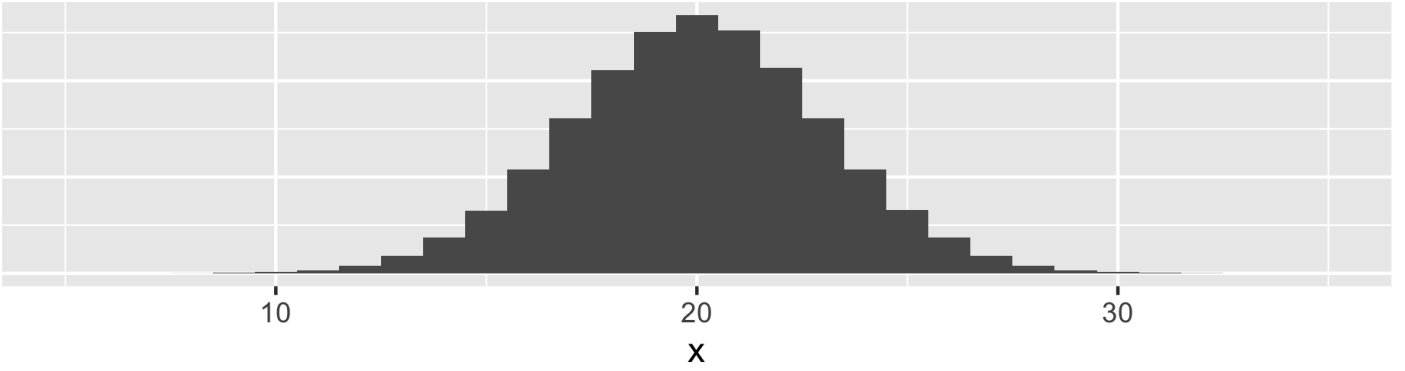
```
samp_3 %>%
  summarise(x_bar = mean(x))
```

```
## # A tibble: 1 x 1
##   x_bar
##   <dbl>
## 1  19.5
```

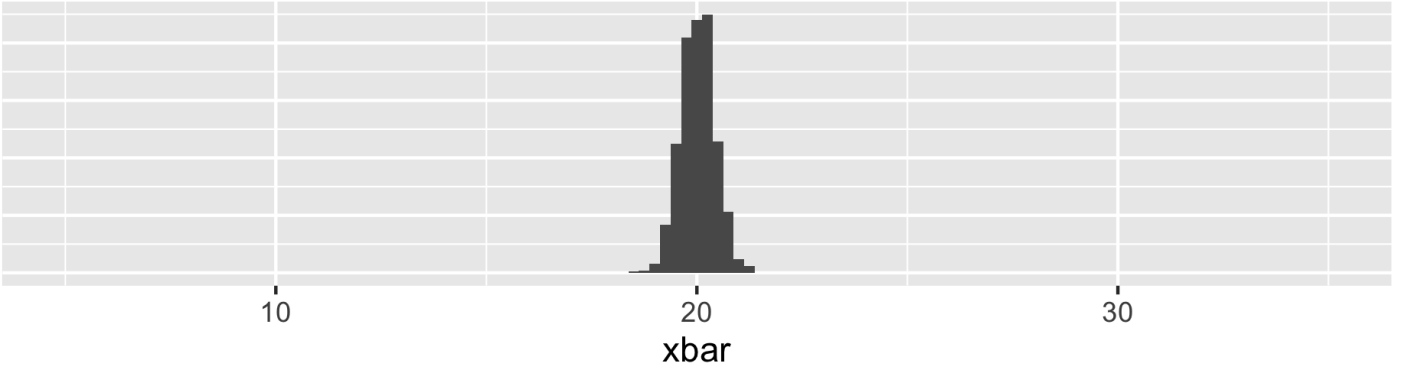
keep repeating...

Population vs. sampling distributions

Population distribution



Sampling distribution of sample means



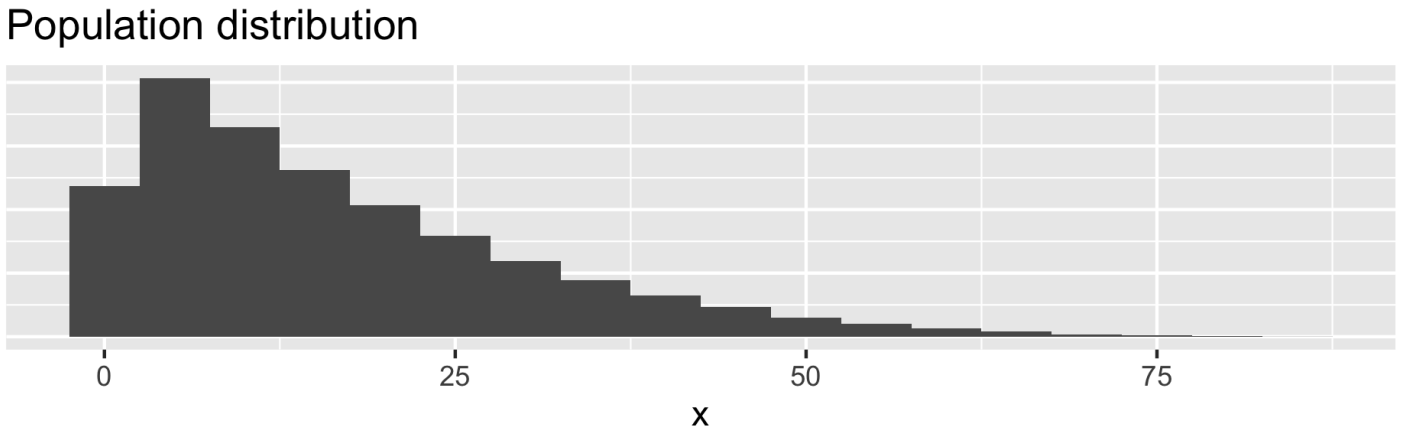
Discussion

Take a minute to discuss the following with 1 - 2 people around you:

How do the shapes, centers, and spreads of these distributions compare?

Let's simulate another distribution

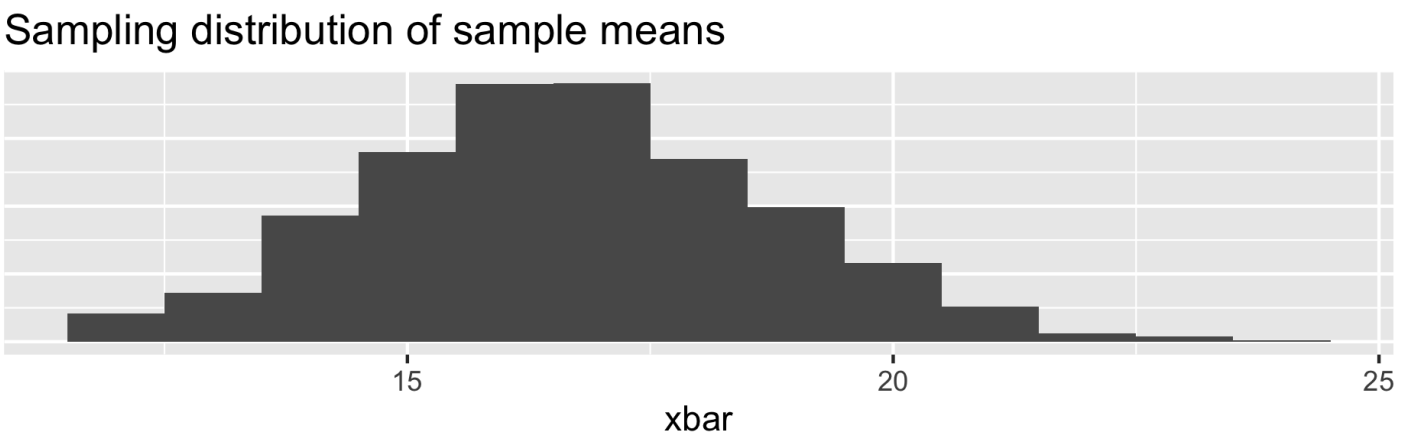
```
rs_pop <- tibble(x = rbeta(100000, 1, 5) * 100)
```




```
## # A tibble: 1 x 2
##      mu sigma
##   <dbl> <dbl>
## 1  16.6  14.1
```

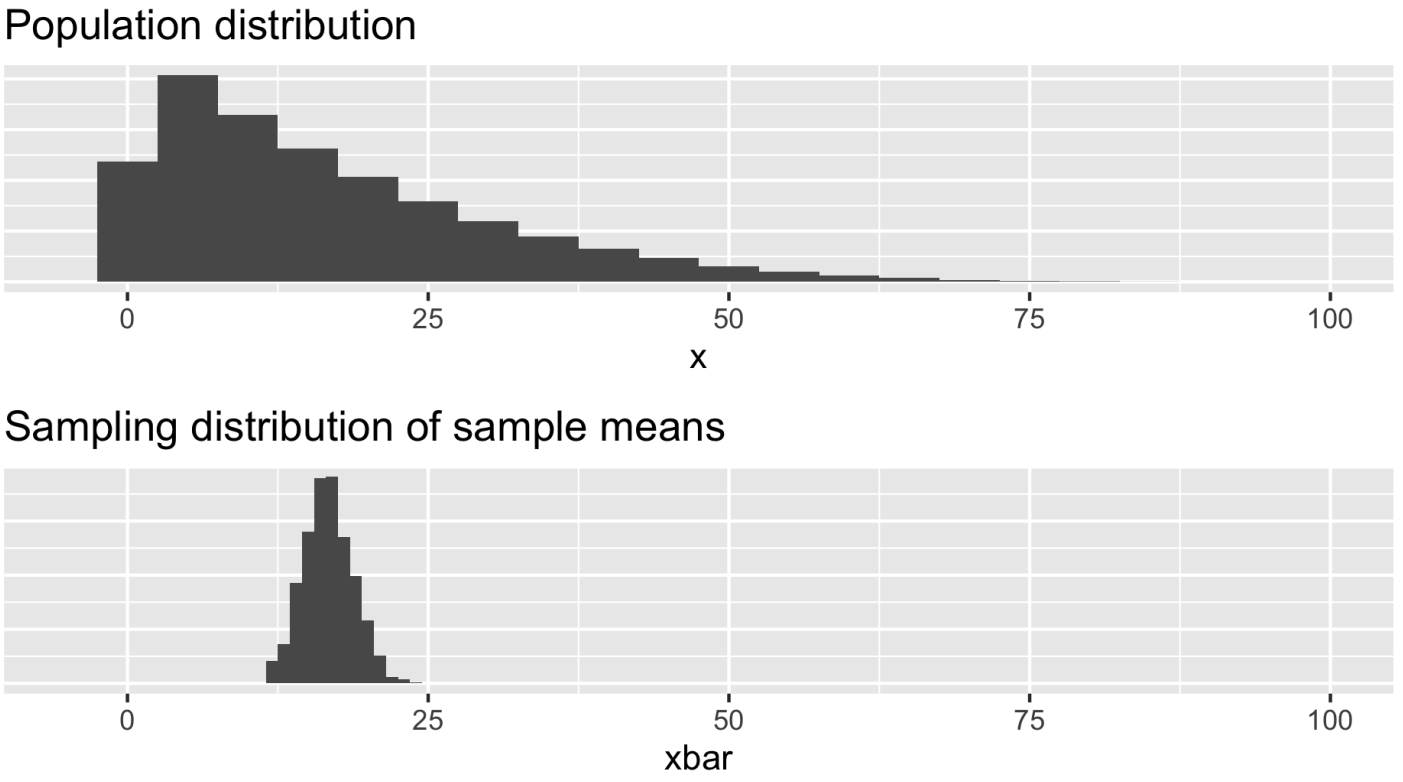
Sampling distribution

```
sampling <- rs_pop %>%
  rep_sample_n(size = 50, replace = TRUE, reps = 1000) %>%
  group_by(replicate) %>%
  summarise(xbar = mean(x))
```




```
## # A tibble: 1 x 2
##    mean    se
##   <dbl> <dbl>
## 1  16.7  2.08
```

Population vs. sampling distribution



In-class exercise

- Use the two examples we just discussed to answer the questions: http://bit.ly/sta210-sp20-samp
- Use **NetId@duke.edu** for your email address.
- You are welcome (and encouraged!) to discuss these questions with 1 - 2 people around you, but **each person** must submit a response.

Central Limit Theorem

- Using the Central Limit Theorem (CLT) we know the form of the sampling distribution for certain statistics such as the mean, proportion, difference in means, etc.
    - CLT does not apply to all statistics (e.g. the median)
- By the Central Limit Theorem, when the conditions are met, we know the sampling distribution of the sample statistic will..
    - be approximately Normal
    - have a mean equal to the unknown population parameter
    - have a standard deviation proportional to the inverse of the square root of the sample size.
- Get more details on the derivation of the CLT in STA 240 & STA 250

CLT for a population mean

Suppose we have a population with mean μ and standard deviation σ. By the CLT, when conditions are met, the sampling distribution of the sample mean is

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Conditions for CLT

- **Independence**: The sampled observations must be independent. This is difficult to check, but the following are useful guidelines:
  - the sample must be random
  - if sampling without replacement, sample size must be less than 10% of the population size
- **Sample Size**: Sample size is large. Usually $n > 30$ is considered large enough sample. Need larger sample size if population distribution is extremely skewed.
- **Independent Groups**: If comparing two populations, the groups must be independent of each other, and all conditions should be checked for both groups.

Standard Error

- By the CLT, the standard deviation of the sampling distribution of $\bar{x}$ is $\sigma/\sqrt{n}$.
- In practice, we don't know the population standard deviation σ, but we can estimate it using the sample standard deviation s.
- The **standard error (SE)** is the *standard deviation* of the *sampling distribution*, calculated using sample statistics

$$SE = \frac{s}{\sqrt{n}}$$

- We will use the standard error for calculations of confidence intervals and hypothesis tests

Confidence interval for the mean

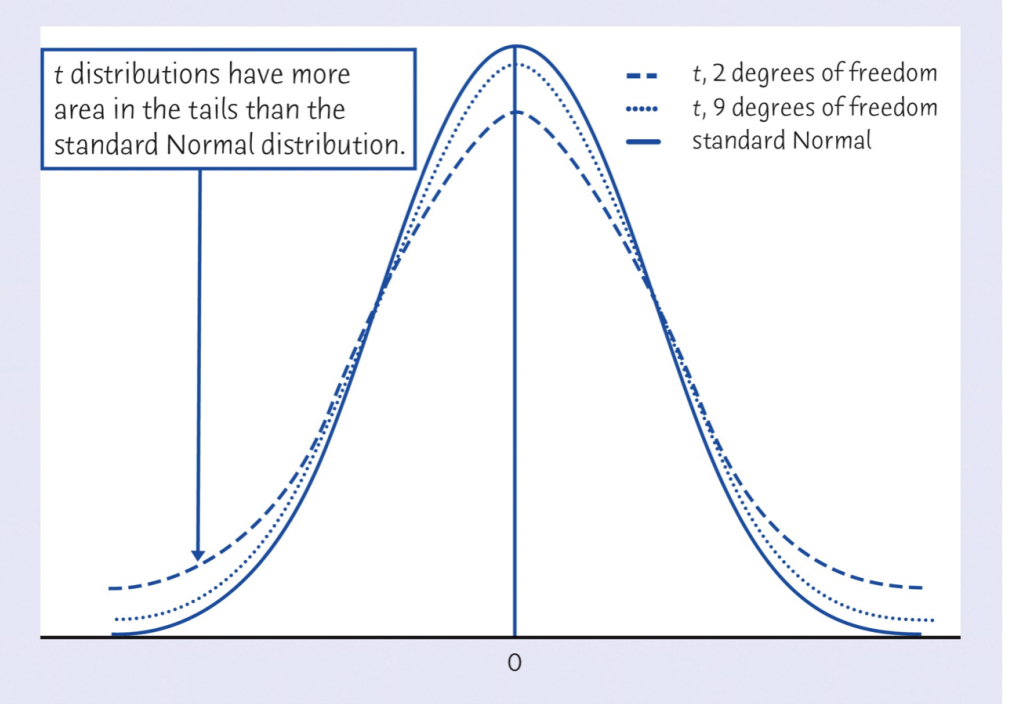
The C confidence interval to estimate μ is

$$\bar{x} \pm t^*_{df} \times \frac{s}{\sqrt{n}}$$

where t^*_{df} is the critical value calculated from the t distribution with $n - 1$ degrees of freedom.

t-distribution vs. Normal

- We need to account for the extra variability that comes from using $s/\sqrt{n}$ (instead $\sigma/\sqrt{n}$) Therefore, we will use the t distribution for the shape of the sampling distribution of $\bar{x}$ in our calculations.



Picture from *The Basic Practice of Statistics (7th edition)*

95\% CI for mean change in reading scores

Let's write the equation for the 95\% confidence interval for the mean change in reading scores in 26 weeks.

```
## # A tibble: 1 x 3
##       n  mean std_dev
##   <int> <dbl>   <dbl>
## 1   240  10.8    11.2
```

```
(t_star <- qt(0.975, 239))
```

```
## [1] 1.969939
```

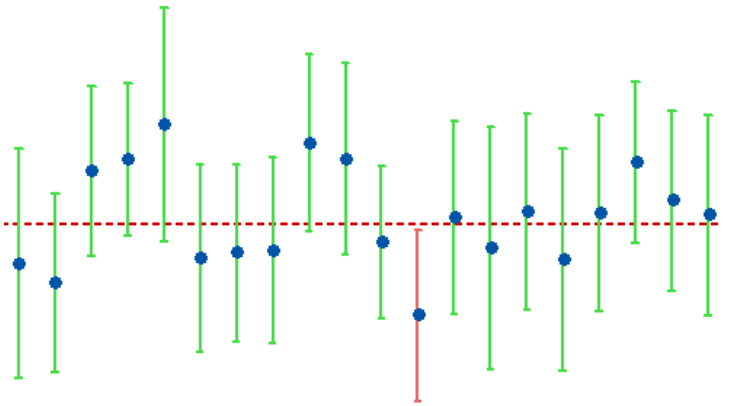
95\% CI for mean change in reading scores

We can also calculate the 95\% confidence interval using the `t.test` function in R

```
t.test(sesame_street$change, conf.level = 0.95)
```

```
##
##     One Sample t-test
##
## data:  sesame_street$change
## t = 14.987, df = 239, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   9.384038 12.224296
## sample estimates:
## mean of x
##  10.80417
```

What does "95\% confidence" mean?



- Suppose we take a lot of samples and calculate a 95\% confidence interval from each
- We would expect about 95% of these intervals to contain the true population mean, i.e. the parameter of interest
- Some sample means will be far away from the parameter and that's OK. The interval is only a plausible range of values. We may conclude that other values are not plausible based on our data, but that doesn't mean other values are impossible.

In-class exercise

- Answer the questions at http://bit.ly/sta210-sp20-CI
- Use **NetId@duke.edu** for your email address.
- You are welcome (and encouraged!) to discuss these questions with 1 - 2 people around you, but **each person** must submit a response.

Confidence Interval for the Difference in Two Means

Difference in mean reading score change

Let Group 1 be the those encouraged to watch *Sesame Street* and Group 2 who got no encouragement to watch the show

```
sesame_street %>%
  group_by(treatment) %>%
  summarise(n = n(), mean = mean(change), sd = sd(change))
```

```
## # A tibble: 2 x 4
##   treatment           n  mean    sd
##   <chr>           <int> <dbl> <dbl>
## 1 Encouraged        152 12.5   10.7
## 2 Not Encouraged     88  7.88  11.4
```

- **Parameter**: $\mu_1 - \mu_2$
- **Statistic**: $\bar{x}_1 - \bar{x}_2$
- We want to estimate the difference in the mean change in reading scores between the two groups, i.e. estimate $\mu_1 - \mu_2$.]

Sample distribution of $\bar{x}_1 - \bar{x}2$

- In the *Sesame Street* example, the parameter of interest is the difference in means, $\mu_1 - \mu_2$. Let's look at the confidence interval for $\mu_1 - \mu_2$ based on the CLT
- The statistic is the difference in sample means $\bar{x}_1 - \bar{x}_2$
- Assuming the conditions for the CLT are met (independent observations, large n, independent groups), the sampling distribution for $\bar{x}_1 - \bar{x}_2$ is

$$\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Confidence interval for the difference in means

The C confidence interval to estimate $\bar{\mu}_1 - \bar{\mu}_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm t^*_{df} \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where t^*_{df} is the critical value calculated from the *t* distribution with *df* degrees of freedom

Standard Error of $\bar{x}_1 - \bar{x}_2$

- In practice, we don't know the population standard deviations σ_1 and σ_2
- We will use the sample standard deviations s_1 and s_2 to estimate σ_1 and σ_2
- Thus, the **standard error of $\bar{\mathbf{x}}_\mathbf{1} - \bar{\mathbf{x}}_\mathbf{2}$** is

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Calculating the critical value

The critical value, t^*, follows a t distribution with degrees of freedom given by the formula:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2} \approx min\{n_1 - 1, n_2 - 1\}$$

In practice, we can use R to calculate the degrees of freedom.

95% confidence interval for the difference in means

```
sesame_street %>%
  group_by(treatment) %>%
  summarise(n = n(), mean = mean(change), std_dev = sd(change))
```

```
## # A tibble: 2 x 4
##   treatment           n  mean std_dev
##   <chr>           <int> <dbl>   <dbl>
## 1 Encouraged        152 12.5     10.7
## 2 Not Encouraged     88  7.88    11.4
```

```
(df <-  (var1/n1 + var2/n2)^2/((var1/n1)^2*(n1-1)^(-1) + (var2/n2)
```

```
## [1] 173.5923
```

```
(t_star <- qt(0.975, df))
```

```
## [1] 1.973724
```

95% confidence interval for the difference in means

We can also calculate the 95% confidence interval for the difference in means using the `t.test` function

```
t.test(change ~ treatment, data = sesame_street, conf.level = 0.95
```

```
##
##     Welch Two Sample t-test
##
## data:  change by treatment
## t = 3.102, df = 173.59, p-value = 0.002244
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.682256 7.567744
## sample estimates:
##     mean in group Encouraged mean in group Not Encouraged
##                       12.500                       7.875
```

95% confidence interval for the difference in means

The 95% confidence interval for the difference in mean reading score change is [1.682, 1.757].

Interpret this interval in the context of the data.

Using this interval, is there evidence of a statistically significant difference in the mean change in reading scores between those encouraged to watch *Sesame Street* and those who got no encouragement?

Accessing RStudio & GitHub

Access RStudio & GitHub

RStudio

- Go to https://vm-manage.oit.duke.edu/containers and login using your NetId credentials
- Click to log in to the Docker container called **STA 210 - Regression Analysis**

GitHub

- Go to https://github.com/sta210-sp20
- Click to accept the invite at the top of the page. This will make you a member of the sta210-sp20 organization on GitHub