STAT 216 Coursepack

Spring 2025
Montana State University

Melinda Yager
Jade Schmidt
Stacey Hancock

This resource was developed by Melinda Yager, Jade Schmidt, and Stacey Hancock in 2021 to accompany the online textbook: Hancock, S., Carnegie, N., Meyer, E., Schmidt, J., and Yager, M. (2021). *Montana State Introductory Statistics with R*. Montana State University. https://mtstateintrostats.github.io/IntroStatTextbook/.

Contents

Preface

This coursepack accompanies the textbook for STAT 216: Montana State Introductory Statistics with R, which can be found at https://mtstateintrostats.github.io/IntroStatTextbook/. The syllabus for the course (including the course calendar), data sets, and links to D2L Brightspace, Gradescope, and the MSU RStudio server can be found on the course webpage: https://math.montana.edu/courses/s216/. Other notes and review materials are linked in D2L.

Each of the activities in this workbook is designed to target specific learning outcomes of the course, giving you practice with important statistical concepts in a group setting with instructor guidance. In addition to the in-class activities for the course, video notes are provided to aid in taking notes while you complete the required videos. Bring this workbook with you to class each class period, and take notes in the workbook as you would your own notes. A well-written completed workbook will provide an optimal study guide for exams!

All activities and labs in this coursepack will be completed during class time. Parts of each lab will be turned in on Gradescope. To aid in your understanding, read through the introduction for each activity before attending class each day.

STAT 216 is a 3-credit in-person course. In our experience, it takes six to nine hours per week outside of class to achieve a good grade in this class. By "good" we mean at least a C because a grade of D or below does not count toward fulfilling degree requirements. Many of you set your goals higher than just getting a C, and we fully support that. You need roughly nine hours per week to review past activities, read feedback on previous assignments, complete current assignments, and prepare for the next day's class. A typical week in the life of a STAT 216 student looks like:

- *Prior to class meeting*:
  - Read assigned sections of the textbook, using the provided reading guides to take notes on the material.
  - Watch the provided videos, taking notes in the coursepack.
  - Read through the introduction to the day's in-class activity.
  - Read through the week's homework assignment and note any questions you may have on the content.
- *During class meeting*:
  - Work through the guided activity, in-class activity or weekly lab with your classmates and instructor, taking detailed notes on your answers to each question in the activity.
- *After class meeting*:
  - Complete any parts of the activity you did not complete in class.
  - Review the activity solutions in the Math and Stat Center, and take notes on key points.
  - Complete any remaining assigned readings for the week.
  - Complete the week's homework assignment.

MODULE 1

Exploring Quantitative Data: Exploratory Data Analysis and Hypothesis Testing for a Single Quantitative Variable

1.1 Vocabulary Review and Key Topics

Review the Golden Ticket posted in the resources at the end of the coursepack for a summary of a single quantitative variable.

1.1.1 Key topics

Module 6 will introduce exploratory data analysis and hypothesis testing using both simulation-based and theory-based methods for a single quantitative variable. The **summary measure** for one quantitative variable is the **mean**. Additionally, we can find the five number summary (min, Q1, median, Q3, max) as well as the sample standard deviation.

- Notation for a sample mean: $\bar{x}$
- Notation for a sample standard deviation: s
- Notation for a population mean: μ
- Types of plots for a single categorical variable:
  - Histogram
  - Boxplot
  - Dotplot

1.1.2 Vocabulary

Sample statistics for a single quantitative variable

- **Mean**, $\bar{x}$: the average

  $$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

  where $x_1, x_2, \ldots, x_n$ are the data values and n is the sample size.
- **Median**: value at the 50th percentile; approximately 50% of data values are at or below the value of the median.
- **Quartile 1** (lower quartile), Q_1: value at the 25th percentile; approximately 25% of data values are at or below the value of Q_1.
- **Quartile 3** (upper quartile), Q_3: value at the 75th percentile; approximately 75% of data values are at or below the value of Q_3.

- **Sample standard deviation**, s: on average, each value in the data set is s units from the mean of the data set ($\bar{x}$). We will always calculate s using R, but it is calculated using the following formula:

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}},$$

where $x_1, x_2, \ldots, x_n$ are the data values, $\bar{x}$ is the sample mean, and n is the sample size.

- **Interquartile range**: the range of the data between the two quartiles: $IQR = Q_3 - Q_1$.
- R code to find the summary statistics for a quantitative variable:

```
object %>% # Data set piped into...
    summarise(favstats(variable))
```

Plotting one quantitative variable

- **Histogram**: sorts a quantitative variable into bins of a certain width. R code to create a histogram:

```
object %>% # Data set piped into...
    ggplot(aes(x = variable)) +   # Name variable to plot
    geom_histogram(binwidth = 10) +  # Create histogram with specified binwidth
    labs(title = "Don't forget to title the plot!", # Title for plot
        x = "x-axis label", # Label for x axis
        y = "y-axis label") # Label for y axis
```

- **Boxplot**: plots the values of the five-number summary and shows any outliers in the data set. R code to create a boxplot:

```
object %>% # Data set piped into...
    ggplot(aes(x = variable)) + # Name variable to plot
    geom_boxplot() + # Create boxplot
    labs(title = "Don't forget to title the plot!", # Title for plot
        x = "x-axis label", # Label for x axis
        y = "y-axis label") # Label for y axis
```

- **Dotplot**: plots each value as a dot along the x-axis. R code to create a dotplot:

```
object %>% # Data set piped into...
    ggplot(aes(x = variable)) + # Name variable to plot
    geom_dotplot() + # Create dotplot
    labs(title = "Don't forget to title the plot!", # Title for plot
        x = "x-axis label", # Label for x axis
        y = "y-axis label") # Label for y axis
```

- Four characteristics of a distribution of a single quantitative variable:
  - Shape (symmetric, skewed left, or skewed right)
  - Center
  - Spread
  - Outliers?

Hypothesis testing

Hypotheses in notation for a single mean: In the hypotheses below, μ_0 is the **null value**.

$$H_0 : \mu = \mu_0$$

$$H_A : \mu \left\{ \begin{array}{c} < \\ \neq \\ < \end{array} \right\} \mu_0$$

Simulation-based Hypothesis Testing

- R code to use for **simulation-based methods** for one quantitative variable to find the p-value, `one_mean_test`, is shown below. Review the comments (instructions after the #) to see what each should be entered for each line of code.

```
one_mean_test(object$variable,#Enter the object name and variable
          null_value = xx, #Enter the null value for the study
          summary_measure = "mean",  #Can choose between mean or median
          shift = xx, #Difference between the null value and the sample mean
          as_extreme_as = xx, #Value of the summary statistic
          direction = "xx", #Specify direction of alternative hypothesis
          number_repetitions = 10000)
```

Theory-based Hypothesis Testing

- Theory-based methods should give the same results as simulation-based methods if conditions are met. For a single quantitative variable, conditions are met if either the data themselves follow a normal distribution or if the sample size is large enough. We call this the "normality condition."
- **Conditions for the sampling distribution of $\bar{x}$ to follow an approximate normal distribution**:
  - **Independence**: The sample's observations are independent, e.g., are from a simple random sample. (*Remember*: This also must be true to use simulation methods!)
  - **Normality Condition**: Either the sample observations come from a normally distributed population or we have a large enough sample size. To check this condition, use the following rules of thumb:
    - $n < 30$: The distribution of the sample must be approximately normal with no outliers.
    - $30 \geq n < 100$: We can relax the condition a little; the distribution of the sample must have no extreme outliers or skewness.
    - $n > 100$: Can assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.
- **t-distribution**: a theoretical distribution that is bell-shaped with mean zero. Its degrees of freedom determine the variability of the distribution. For very large degrees of freedom, the t-distribution is close to a standard normal distribution. For a single quantitative variable, the degrees of freedom are calculated by subtracting one from the sample size: $n - 1$. A t-distribution with $n - 1$ degrees of freedom is denoted by: t_{n-1}.
- **Standard error of the sample mean**: measures the how far each possible sample mean is from the true mean, on average, and is calculated using the formula below:

  $$SE(\bar{x}) = \frac{s}{\sqrt{n}}$$

  where s is the sample standard deviation.
- **Standardized sample mean**: standardized statistic for a single quantitative variable calculated using:

  $$T = \frac{\bar{x} - \mu_0}{SE(\bar{x})},$$

  If the conditions for the sampling distribution of $\bar{x}$ to follow an approximate normal distribution are met, and if the true value of μ is equal to the null value of μ_0, the standardized sample mean, T, will have an approximate t-distribution with $n - 1$ degrees of freedom.

- The following R code is used to find the p-value using theory based methods for a single quantitative variable.
  - `pt` will give you a p-value using the t-distribution with $n-1$ df (enter for `yy`)
  - Enter the value of the standardized statistic for `xx`
  - If a greater than alternative, change `lower.tail = TRUE` to FALSE.
  - If a two-sided test, multiply by 2.

```
pt(xx, df = yy, lower.tail=TRUE)
```

1.2 Video Notes: Exploratory Data Analysis and Hypothesis Testing of Quantitative Variables

Read Chapters 5 and 17 in the course textbook. Use the following videos to complete the video notes for Module 6.

1.2.1 Course Videos

- QuantitativeData
- 5.5to5.6
- 5.7
- 17.2
- 17.3TheoryTests

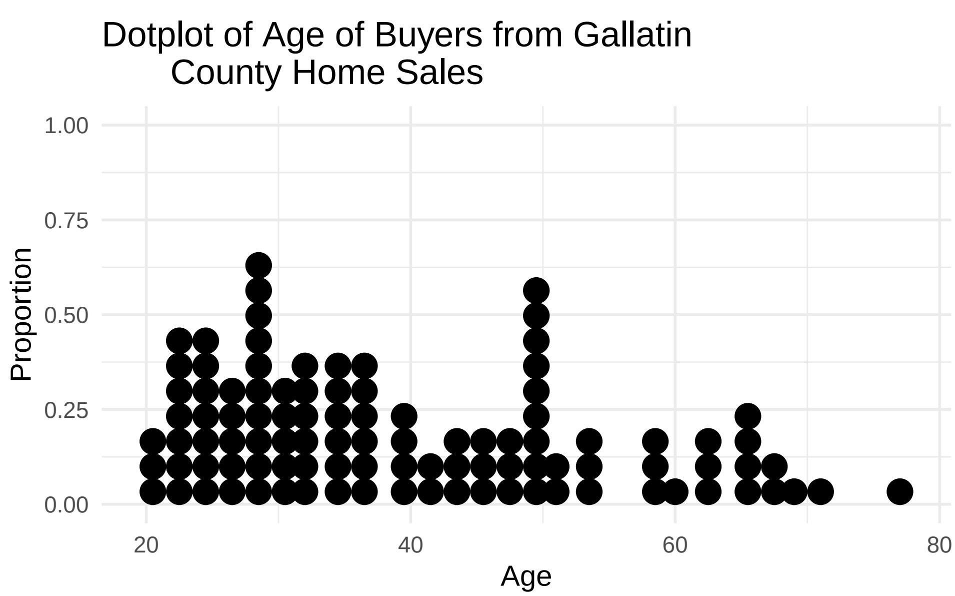
Summarizing quantitative data - Videos 5.2to5.4 and 5.5to5.6

Types of plots

We will revisit the moving to Montana data set and plot the age of the buyers.

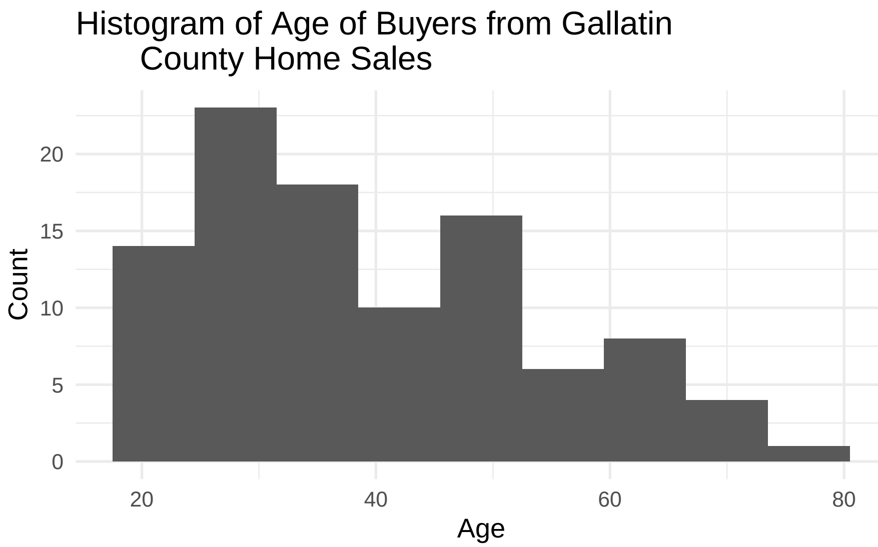
Dotplot:

```
moving %>%
  ggplot(aes(x = Age))+ #Enter variable to plot
  geom_dotplot() +
  labs(title = "Dotplot of Age of Buyers from Gallatin
       County Home Sales", #Title your plot
       x = "Age", #x-axis label
       y = "Proportion") #y-axis label
```



Histogram:

```
moving %>%
  ggplot(aes(x = Age))+
  geom_histogram(binwidth = 7) +
  labs(title = "Histogram of Age of Buyers from Gallatin
       County Home Sales",
       #Title your plot
       x = "Age",
       y = "Count")
```



Quantitative data can be numerically summarized by finding:

Two measures of center:

- Mean: __________________ of all the ____________________ in the data set.
    - Sum the values in the data set and divide the sum by the sample size
- Notation used for the population mean:
    - Single quantitative variable:

      - One categorical and one quantitative variable:

        - Subscripts represent the _________________ variable groups
- Notation used for the sample mean:

      - Single quantitative variable:

      - One categorical and one quantitative variable:

- Median: Value at the ____________________ percentile
  - _______________ % of values are at and _________________ and at and _________________ the value of the _____________________.
  - Middle value in a list of ordered values

Two measures of spread:

- Standard deviation: Average _____________________________ each data point is from the _____________________ of the data set.

  - Notation used for the population standard deviation

  - Notation used for the sample standard deviation

- Interquartile range: middle 50% of data values

  Formula:

  Quartile 3 (Q3) - value at the 75th percentile

  - __________________ % of values are at and ____________________ the value of Q3

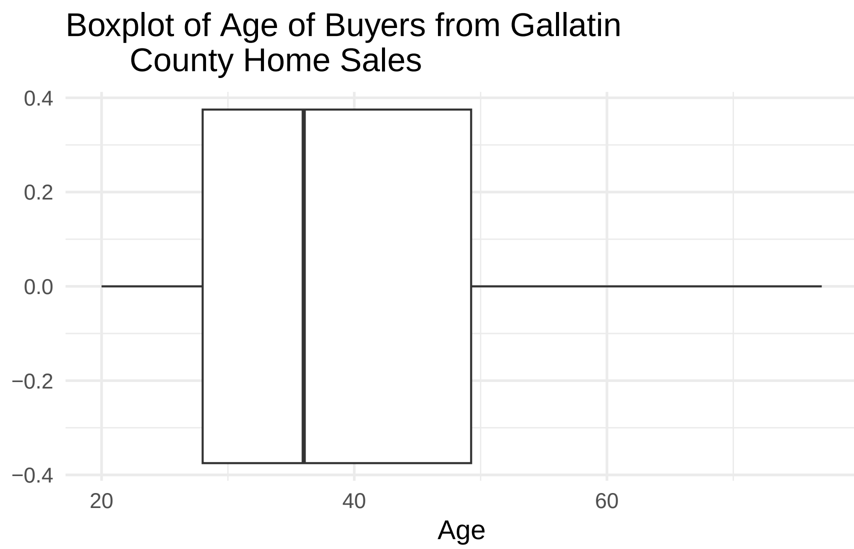
  Quartile 1 (Q1) - value at the 25th percentile

  - ____________________ % of values are at and ____________________ the value of Q1

Boxplot (3rd type of plot for quantitative variables)

```
- Five number summary: minimum, Q1, median, Q3, maximum
```

```
moving %>%
  ggplot(aes(x = Age))+ #Enter variable to plot
  geom_boxplot() +
  labs(title = "Boxplot of Age of Buyers from Gallatin
       County Home Sales", #Title your plot
       x = "Age", #x-axis label
       y = "") #y-axis label
```




```
favstats(moving$Age)
```

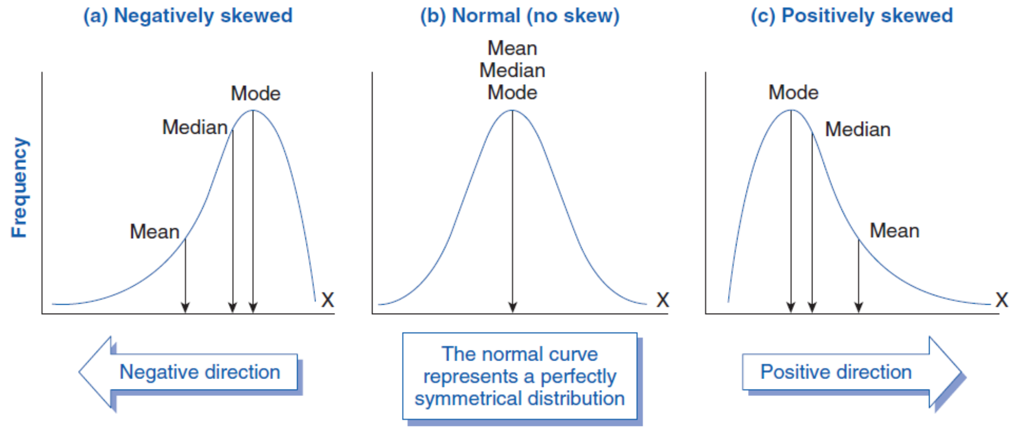
```
#>  min Q1 median    Q3 max  mean       sd   n missing
#>   20 28     36 49.25  77 39.77 14.35471 100       0
```

Interpret the value of Q_3 for the age of buyers.

Interpret the value of s for the age of buyers.

Four characteristics of plots for quantitative variables

- Shape: overall pattern of the data



- What is the shape of the distribution of age of buyers for Gallatin County home sales?

- Center:

Mean or Median

- Report the measure of center for the boxplot of age of buyers for Gallatin County home sales.

- Spread (or variability):

Standard deviation or IQR

- Report the IQR for the distribution of age of buyers from Gallatin County home sales.

- Outliers?

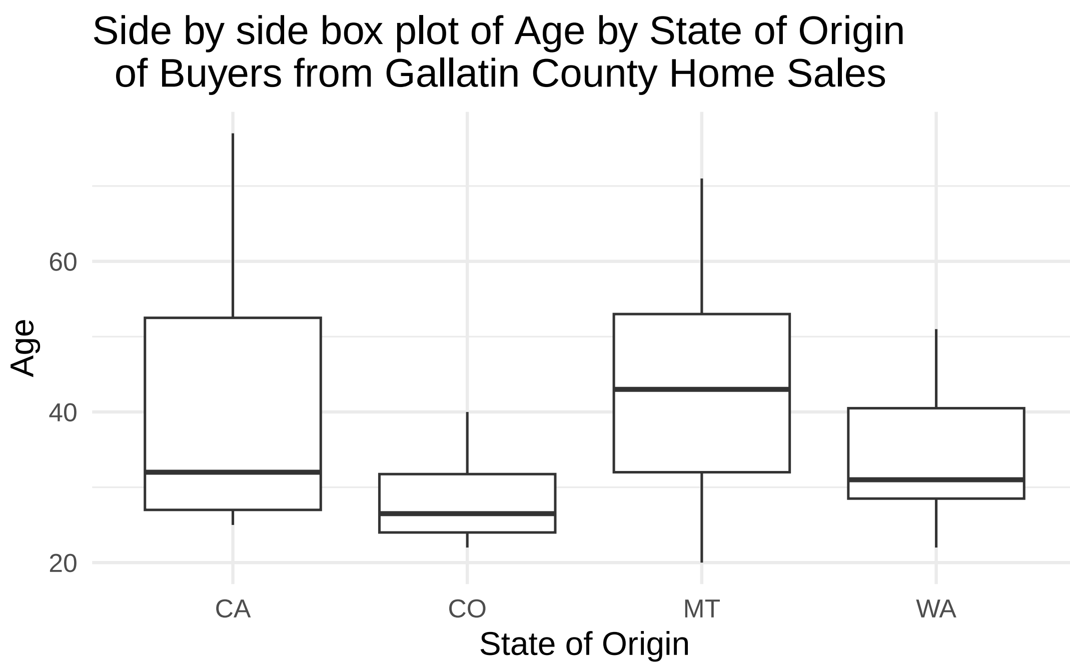
values $< Q_1 - 1.5 \times IQR$

values $> Q_3 + 1.5 \times IQR$

- Use these formulas to show that there are no outliers in the distribution of age of buyers from Gallatin County home sales.

Let's look at side-by-side boxplot of the variable age by state of origin moved from.

```
moving %>%  # Data set piped into...
  ggplot(aes(y = Age, x = From))+  # Identify variables
  geom_boxplot()+  # Tell it to make a box plot
  labs(title = "Side by side box plot of Age by State of Origin
  of Buyers from Gallatin County Home Sales",  # Title
       x = "State of Origin",    # x-axis label
       y = "Age")  # y-axis label
```



- Which state of origin had the oldest median age of buyers from sampled home sales?

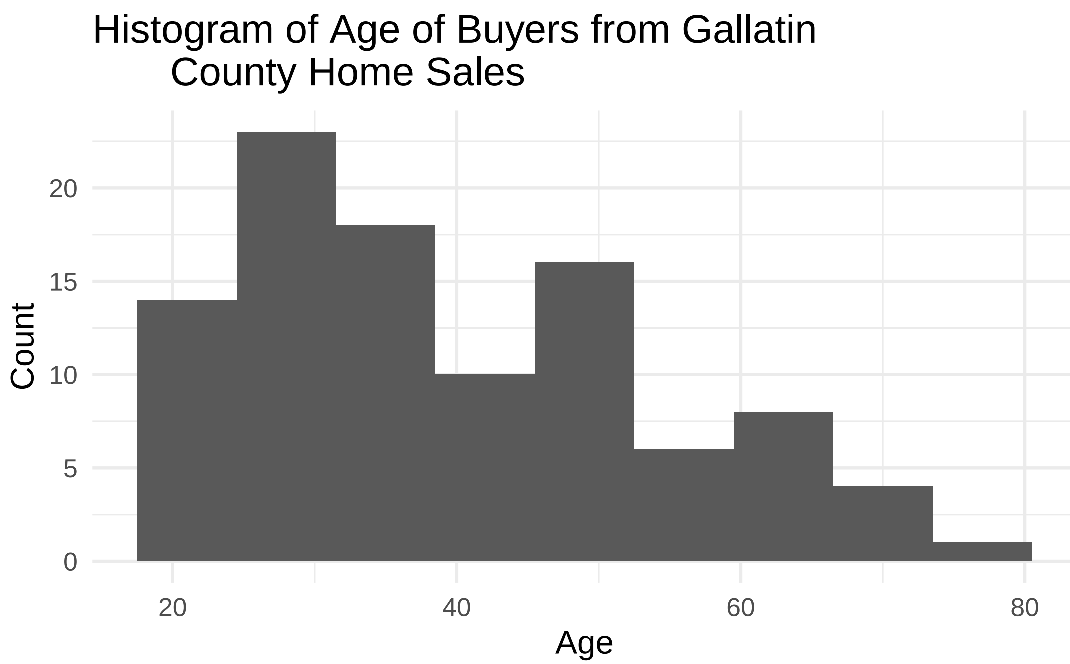
- Which state of origin had the most variability in age of buyers from sampled home sales?

- Which state of origin had the most symmetric distribution of ages of buyers from sampled home sales?

- Which state of origin had outliers for the age of buyers from sampled home sales?

Robust statistics - Video 5.7

Let's review the summary statistics and histogram of age of buyers from sampled home sales.




```
#>  min Q1 median    Q3 max  mean       sd   n missing
#>   20 28     36 49.25  77 39.77 14.35471 100       0
```

Notice that the ___________________ has been pulled in the direction of the ______________________.

- The ________________ is a robust measure of center.
- The ________________ is a robust measure of spread.
- Robust means not _________________________ by outliers.

When the distribution is symmetric use the _________________ as the measure of center and the ________________ as the measure of spread.

When the distribution is skewed with outliers use the ___________________ as the measure of center and the _________________ as the measure of spread.

1.2.2 Video notes single quantitative variable inference

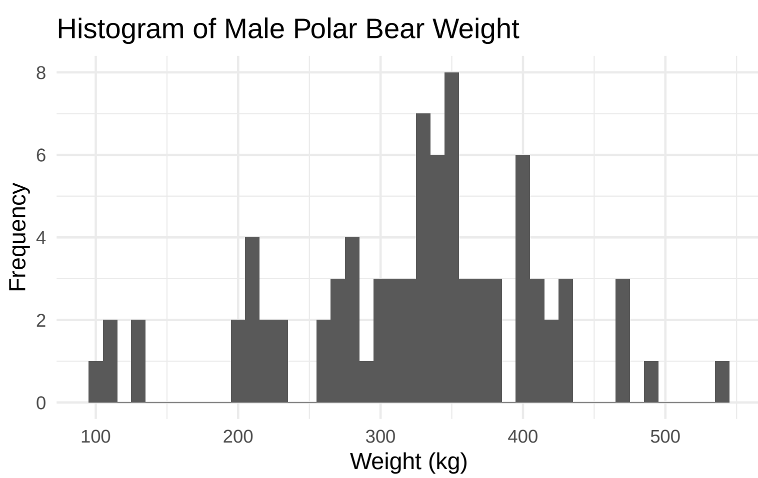
Example: What is the average weight of adult male polar bears? The weight was measured on a representative sample of 83 male polar bears from the Southern Beaufort Sea.

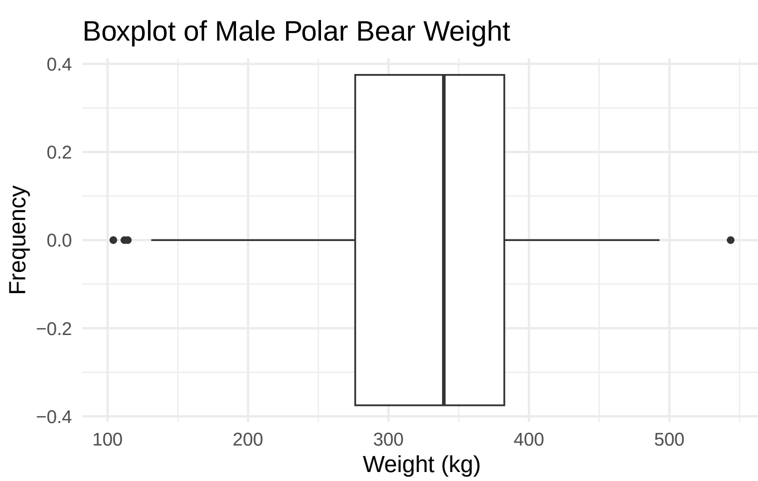
```
pb <- read.csv("https://math.montana.edu/courses/s216/data/polarbear.csv")
```

Plots of the data:

```
pb %>%
    ggplot(aes(x = Weight)) +    # Name variable to plot
    geom_histogram(binwidth = 10) +  # Create histogram with specified binwidth
    labs(title = "Histogram of Male Polar Bear Weight", # Title for plot
       x = "Weight (kg)", # Label for x axis
       y = "Frequency") # Label for y axis

pb %>% # Data set piped into...
ggplot(aes(x = Weight)) +   # Name variable to plot
  geom_boxplot() +  # Create boxplot
  labs(title = "Boxplot of Male Polar Bear Weight", # Title for plot
       x = "Weight (kg)", # Label for x axis
       y = "Frequency") # Label for y axis
```





Summary Statistics:

```
pb %>%
  summarise(favstats(Weight)) #Gives the summary statistics
#>     min    Q1 median     Q3   max     mean       sd  n missing
#> 1 104.1 276.3  339.4 382.45 543.6 324.5988 88.32615 83       0
```

Hypothesis testing

- Hypotheses are always written about the ____________________________________. For a single mean we will use the notation ________________.

Null Hypothesis:

H_0 :

Alternative Hypothesis:

H_A :

- Direction of the alternative depends on the __________________________ ____________________________.

Simulation-based method

- Simulate many samples assuming $H_0 : \mu = \mu_0$
  - Shift the data by the difference between μ_0 and $\bar{x}$
  - Sample with replacement n times from the shifted data
  - Plot the simulated shifted sample mean from each simulation
  - Repeat 1000 times (simulations) to create the null distribution
  - Find the proportion of simulations at least as extreme as $\bar{x}$

Example: Is there evidence that male polar bears weigh less than 370kg (previously recorded measure), on average? The weight was measured on a representative sample of 83 male polar bears from the Southern Beaufort Sea.

Hypotheses:

In notation:

H_0 :

H_A :

In words:

H_0 :

H_A :

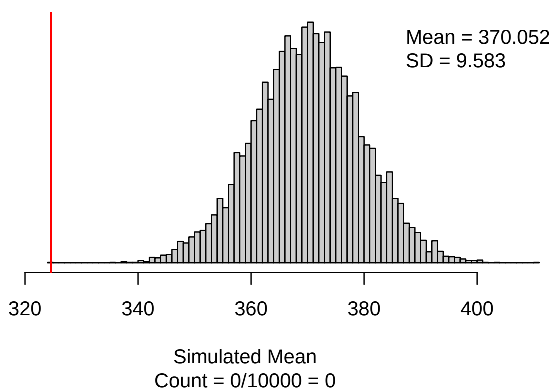
Reminder of summary statistics:

```
pb %>%
  summarise(favstats(Weight)) #Gives the summary statistics
#>     min    Q1 median     Q3   max     mean       sd  n missing
#> 1 104.1 276.3  339.4 382.45 543.6 324.5988 88.32615 83       0
```

Find the difference:

$\mu_0 - \bar{x} =$

```
set.seed(216)
one_mean_test(pb$Weight,   #Enter the object name and variable
              null_value = 370, #Enter null value for the study
              summary_measure = "mean",  #Can choose between mean or median
              shift = 45.4,   # Shift needed for bootstrap hypothesis test
              as_extreme_as = 324.6,  # Observed statistic
              direction = "less",  # Direction of alternative
              number_repetitions = 10000)  # Number of simulated samples for null distribution
```



Interpretation of the p-value:

- Statement about probability or proportion of samples
- Statistic (summary measure and value)
- Direction of the alternative
- Null hypothesis (in context)

Conclusion:

- Amount of evidence
- Parameter of interest
- Direction of the alternative hypothesis

Theory-based method

Conditions for inference using theory-based methods:

- Independence:
- Large enough sample size:

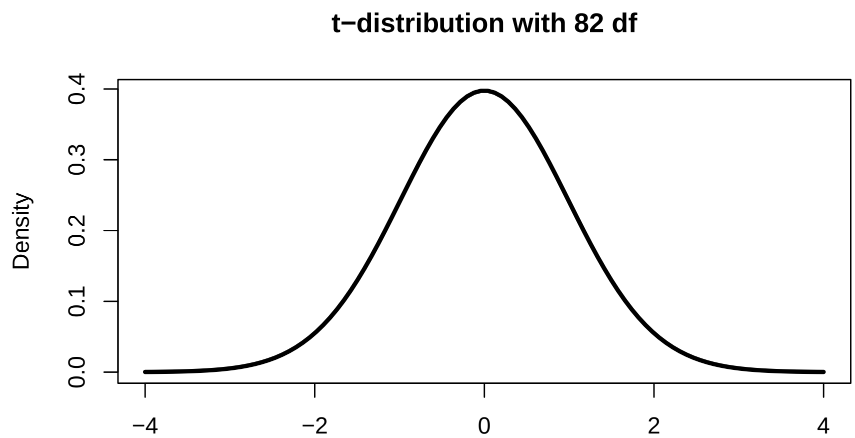
t-distribution

In the theoretical approach, we use the Central Limit Theorem (CLT) to tell us that—under certain conditions—the distribution of sample means will be approximately normal, centered at the assumed true mean under H_0, and with standard deviation $\frac{\sigma}{\sqrt{n}}$.

$$\bar{x} \sim N\left(\mu_0, \frac{\sigma}{\sqrt{n}}\right)$$

- Estimate the population standard deviation, σ, with the ____ standard deviation, ____________.
- For a single quantitative variable we use the ______ - distribution with ______________________ degrees of freedom to approximate the sampling distribution.

The t^* multiplier is the value at the given percentile of the t-distribution with $n-1$ degrees of freedom.

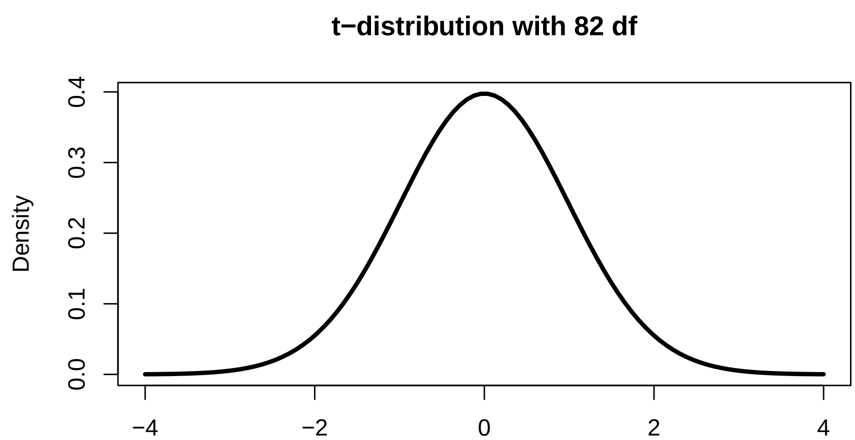


- Calculate the standardized statistic
- Find the area under the t-distribution with $n - 1$ df at least as extreme as the standardized statistic

Equation for the standard error of the sample mean:

Equation for the standardized sample mean:

Calculate the standardized sample mean weight of adult male polar bears:



Interpret the standardized sample mean weight:

To find the theory-based p-value:

```
pt(-4.683, df=82, lower.tail=TRUE)
#> [1] 5.531605e-06
```

1.2.3 Concept Check

Be prepared for group discussion in the next class. One member from the table should write the answers to the following on the whiteboard.

1. What plots can be used to summarize quantitative data?

2. Which measure of center is robust to outliers?

3. How do we determine the direction of the alternative hypothesis?

1.3 Activity 11: Summarizing Quantitative Variables

1.3.1 Learning outcomes

- Identify and create appropriate summary statistics and plots given a data set or research question for quantitative data.
- Interpret the following summary statistics in context: median, lower quartile, upper quartile, standard deviation, interquartile range.

1.3.2 Terminology review

In today's activity, we will review summary measures and plots for quantitative variables. Some terms covered in this activity are:

- Two measures of center: mean, median
- Two measures of spread (variability): standard deviation, interquartile range (IQR)
- Plots of quantitative variables: dotplots, boxplots, histograms
- Given a plot or set of plots, describe and compare the distribution(s) of quantitative variables (center, variability, shape, outliers).

To review these concepts, see Chapter 5 in the textbook.

1.3.3 The Integrated Postsecondary Education Data System (IPEDS)

These data were collected on a subset of higher education institutions that met the following selection criteria (Education Statistics 2018):

- Degree granting
- United States only
- Title IV participating
- Not for profit
- 2-year or 4-year or above
- Has full-time first-time undergraduates

Some of the variables collected and their descriptions are below. Note that several variables have missing values for some institutions (denoted by "NA").

Variable	**Description**
`UnitID`	Unique institution identifier
`Name`	Institution name
`State`	State abbreviation
`Sector`	whether public or private
`LandGrant`	Is this a land-grant institution (Yes/No)
`Size`	Institution size category based on total student enrolled for credit, Fall 2018: Under 1,000, 1,000$-4, 999, 5, 000-9, 999, 10, 000-$19,999, 20,000 and above
`Cost_OutofState`	Cost of attendance for full-time out-of-state undergraduate students
`Cost_InState`	Cost of attendance for full-time in-state undergraduate students
`Retention`	Retention rate is the percent of the undergraduate students that re-enroll in the next year
`Graduation_Rate`	6-year graduation rate for undergraduate students

Variable	Description
SATMath_75	75th percentile Math SAT score
ACT_75	75th percentile ACT score

Identifying variables in a data set

Look through the provided table of variable descriptions. The UnitID and Name are identifiers for each observational unit, *US degree-granting higher education institutions in 2018.*

1. Identify in the table which variables collected on the US institutions are categorical (C) and which variables are quantitative (Q).

Summarizing quantitative variables

The favstats() function from the mosaic package gives the summary statistics for a quantitative variable. The R output below provides the summary statistics for the variable Graduation_Rate. The summary statistics provided are the two measures of center (mean and median) and two measures of spread (standard deviation and the quartile values to calculate the IQR) for undergraduate 6-year graduation rate.

- Highlight and run lines 1–12 in the provided R script file to load the data set. Check that the summary statistics match the output given in the coursepack.
- Notice that the 2-year institutions were removed so the observational units for this study are **4-year US degree-granting higher education institutions in 2018.**

```
IPEDS <- read.csv("https://www.math.montana.edu/courses/s216/data/IPEDS_2018.csv")
IPEDS <- IPEDS %>%
  filter(Sector != "Public 2-year") # Filters the data set to remove Public 2-year
IPEDS <- IPEDS %>%
  filter(Sector != "Private 2-year") # Filters the data set to remove Private 2-year
IPEDS %>%
    summarize(favstats(Graduation_Rate))
```

```
#>   min Q1 median Q3 max     mean       sd    n missing
#> 1   0 38     53 67 100 52.48749 20.63192 1918      49
```

2. Report the values for the two measures of center (mean and median).

3. Calculate the interquartile range ($IQR = Q_3 - Q_1$) of graduation rates.

4. Report the value of the standard deviation and interpret this value in context of the problem.

5. Interpret the value of Q_3 in context of the study.

Displaying a single quantitative variable

There are three type of plots used to plot a single quantitative variable: a dotplot, a histogram or a boxplot. A dotplot of graduation rates would plot a dot for the graduation rate for each 4-year US higher education institution.

First, let's create a histogram of the variable `Graduation_Rate`.

- Enter the name of the variable in line 19 for `variable` in the R script file.
- Replace the word title for the plot in line 21 between the quotations with a descriptive title. **A title should include: type of plot, variable or variables plotted, and observational units.**
- Highlight and run lines 18–23 to create the histogram.

```
IPEDS %>% # Data set piped into...
ggplot(aes(x = xx)) +   # Name variable to plot
  geom_histogram(binwidth = 10) +  # Create histogram with specified binwidth
  labs(title = "Don't forget to title the plot!", # Title for plot
       x = "Graduation Rate", # Label for x axis
       y = "Frequency") # Label for y axis
```

Notice that the **bin width** for the histogram is 10. For example the first bin consists of the number of institutions in the data set with a graduation rate of 0 to 10%. It is important to note that a graduation rate on the boundary of a bin will fall into the bin above it; for example, 20 would be counted in the bin 20–30.

6. Which range of Graduation Rates have the highest frequency?

Next we will create a boxplot of the variable `Graduation_Rate`.

- Enter the name of the variable in line 28 for `variable` in the R script file.
- Highlight and run lines 28–36 to create the boxplot.

```
IPEDS %>% # Data set piped into...
ggplot(aes(x = variable)) +   # Name variable to plot
  geom_boxplot() +  # Create boxplot with specified binwidth
  labs(title = "Boxplot of Graduation Rates for \n 4-year Higher Education Institutions",
          # Title for plot
          # Note the \n starts a new line
       x = "Graduation_Rate", # Label for x axis
       y = "") + # Remove y axis label
    theme(axis.text.y = element_blank(),
          axis.ticks.y = element_blank()) # Removes y-axis ticks
```

7. Sketch the boxplot created and identify the values of the 5-number summary (minimum value, Q_1, median, Q_3, maximum value) on the plot. Use the following formulas to find the invisible fence on both ends of the distribution. Draw a dotted line at the invisible fence to show how the outliers were detected (any values less than the lower fence or greater than the upper fence were flagged as outliers).

Lower Fence: $Q_1 - 1.5 \times IQR$ Upper Fence: $Q_3 + 1.5 \times IQR$

When describing distributions of quantitative variables we discuss the **shape** (symmetric or skewed), the **center** (mean or median), **spread** (standard deviation or IQR), and if there are **outliers** present.

8. What is the shape of the distribution of graduation rates?

9. From which plot (histogram or boxplot) is it easier to determine the shape of the distribution?

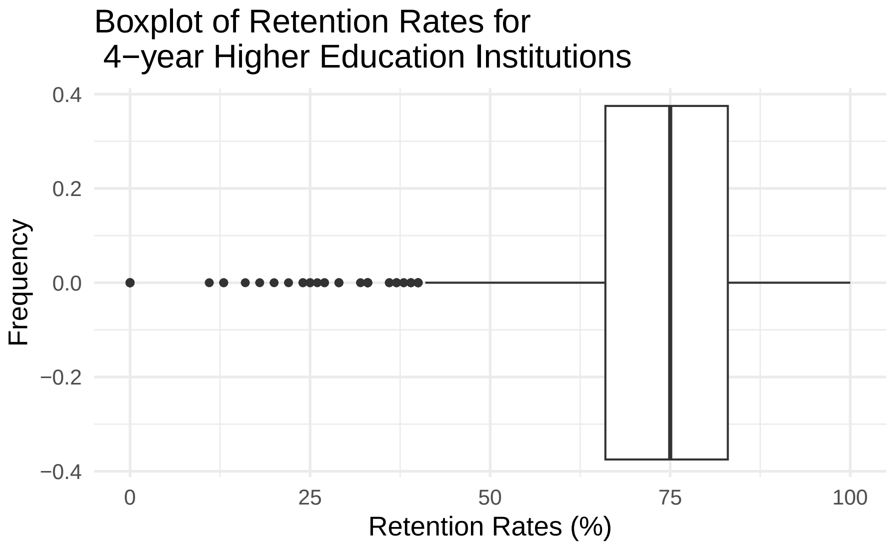
10. From which plot is it easier to determine if there are outliers?

Robust statistics

Let's examine how the presence of outliers affects the different summary measures for center and spread. For this part of the activity, we will look at the retention rate variable (`Retention`) in the IPEDS data set.

```
IPEDS %>% # Data set piped into...
    summarise(favstats(Retention))
#>   min Q1 median Q3 max    mean       sd    n missing
#> 1   0 66     75 83 100 73.8525 15.14323 1817     150

IPEDS %>% # Data set piped into...
    ggplot(aes(x = Retention)) + # Name variable to plot
    geom_boxplot() + # Create boxplot
    labs(title = "Boxplot of Retention Rates for \n 4-year Higher Education Institutions",
           # Title for plot
         x = "Retention Rates (%)", # Label for x axis
         y = "Frequency") # Label for y axis
```



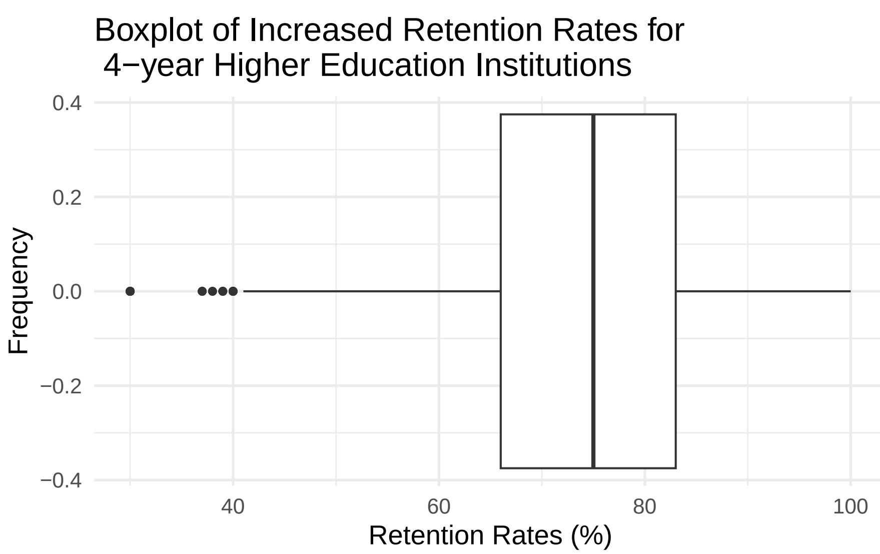
11. Report the values for the two measures of center for these data.

12. Report the values for the two measures of spread for these data.

To show the effect of outliers on the measures of center and spread, the smallest values of retention rate in the data set were increased by 30%. This variable is called `Retention_Inc`.

```
IPEDS %>% # Data set piped into...
    summarise(favstats(Retention_Inc))
#>   min Q1 median Q3 max     mean       sd    n missing
#> 1  30 66     75 83 100 74.49642 13.41255 1817     150

IPEDS %>% # Data set piped into...
    ggplot(aes(x = Retention_Inc)) + # Name variable to plot
    geom_boxplot() + # Create histogram
    labs(title = "Boxplot of Increased Retention Rates for \n 4-year Higher Education Institutions",
        # Title for plot
        x = "Retention Rates (%)", # Label for x axis
        y = "Frequency") # Label for y axis
```



13. Report the values for the two measures of center for this new data set.

14. Report the values for the two measures of spread for this new data set.

15. Which measure of center is robust to outliers? Explain your answer.

16. Which measure of spread is robust to outliers? Explain your answer.

1.3.4 Take-home messages

1. Histograms, box plots, and dot plots can all be used to graphically display a single quantitative variable.
2. The box plot is created using the five number summary: minimum value, quartile 1, median, quartile 3, and maximum value. Whiskers extend to the lowest value and highest value that are *not* considered outliers. Values in the data set that are less than $Q_1 - 1.5 \times IQR$ or greater than $Q_3 + 1.5 \times IQR$ are considered outliers and are graphically represented by a dot outside of the whiskers on the box plot.
3. Data should be summarized numerically and displayed graphically to give us information about the study.
4. When comparing distributions of quantitative variables we look at the shape, center, spread, and for outliers. In this course, we only consider two measures of center (mean and the median), and two measures of spread (standard deviation and the interquartile range, $IQR = Q_3 - Q_1$).

1.3.5 Additional notes

Use this space to summarize your thoughts and take additional notes on today's activity and material covered.

1.4 Activity 12: Hypothesis Testing of a Single Quantitative Variable

1.4.1 Learning outcomes

- Given a research question involving one quantitative variable, construct the null and alternative hypotheses in words and using appropriate statistical symbols.
- Investigate the process of creating a null distribution for one quantitative variable.
- Find, evaluate, and interpret a p-value from the null distribution.

1.4.2 Terminology review

In today's activity, we will use simulation-based and theory-based methods to analyze a single quantitative variable. Some terms covered in this activity are:

- Null hypothesis
- Alternative hypothesis
- Null distribution
- t-distribution
- p-value

To review these concepts, see Chapters 9 and 17 in the textbook.

1.4.3 College student sleep habits

According to an article in *Sleep* (Watson 2015), experts recommend adults (>18 years old) get at least 7 hours of sleep per night. A professor at MSU is interested in the sleep habits of MSU students. The professor obtained a representative sample of MSU students and asked each student to report the amount of sleep they get on a typically night. Is there evidence that MSU students get less than the recommended 7 hours of sleep per night, on average?

Summarizing quantitative variables

- Download the R script file and data file for this activity
- Upload both files to the RStudio server and open the R script file
- Enter the name of the dataset for datasetname.csv
- Highlight and run lines 1–8 to load the data

```
sleep <- read.csv("datasetname.csv")
```

Ask a research question

1. Write the parameter of interest in context of the study.

2. Write the null hypothesis in words in context of the study.

3. Write the alternative hypothesis in notation.

Summarize and visualize the data

The `favstats()` function from the `mosaic` package gives the summary statistics for a quantitative variable.

- Enter the variable name, `SleepHours`, for `variable` in line 13
- Highlight and run lines 12–13

```
sleep %>%
    summarize(favstats(variable))
```

4. About how far is each number of hours of sleep for a Stat 216 student from the mean number of hours of sleep, on average?

Create a boxplot of the variable `SleepHours`.

- Enter the name of the variable in line 19 for `variable` in the R script file.
- Enter a title in line 21 for the plot between the quotations.
- Highlight and run lines 18–25.

```
sleep %>% # Data set piped into...
    ggplot(aes(x = variable)) +   # Name variable to plot
    geom_boxplot() +  # Create boxplot with specified binwidth
    labs(title = "Don't forget to title your plot!", # Title for plot
       x = "Amount of sleep (hrs)", # Label for x axis
       y = "") + # Remove y axis label
    theme(axis.text.y = element_blank(),
          axis.ticks.y = element_blank()) # Removes y-axis ticks
```

5. Describe the distribution of number of hours of sleep using the four characteristics of boxplots.

Simulation methods

To simulate the null distribution of sample means we will use a bootstrapping method. Recall that the null distribution must be created under the assumption that the null hypothesis is true. Therefore, before bootstrapping, we will need to *shift* each data point by the difference $\mu_0 - \bar{x}$. This will ensure that the mean of the shifted data is μ_0 (rather than the mean of the original data, $\bar{x}$), and that the simulated null distribution will be centered at the null value.

6. Calculate the difference $\mu_0 - \bar{x}$. Based on the sign of this difference, will we need to shift the data up or down?

- Open the data set (`sleep_college`) in Excel.
- Create a new column labeled Shift.
- In the column, Shift, add the shifted value (answer to Q6) to each value in the `SleepHours` column.
- Save the file and upload again to the RStudio server.
- Find the `favstats` of the variable, `Shift`.
- Highlight and run lines 30–32

```
sleep <- read.csv("sleep_college.csv")
sleep %>%
    summarize(favstats(Shift))
```

7. Report the mean of the `Shift` variable. Why does it make sense that this value is the same as the null value?

8. Report the standard deviation of the `Shift` variable. How does this compare to the standard deviation for the variable `SleepHours`? Explain why these values are the same.

9. What inputs should be entered for each of the following to create the simulated null distribution?

- Null value (What is the null value for the study?):
- Summary measure ("mean" or "median"):
- Shift (difference between $\mu_0 - \bar{x}$):
- As extreme as (enter the value for the observed sample mean):
- Direction (`"greater"`, `"less"`, or `"two-sided"`):
- Number of repetitions:

The `one_mean_test` will be used to find the p-value for the simulation test. Following the instructions below to complete the code.

- Enter your answers for question 9 in place of the `xx`'s to produce the null distribution with 10000 simulations.
- Highlight and run lines 36–42.

```
one_mean_test(sleep$SleepHours,#Enter the object name and variable
              null_value = xx,
              summary_measure = "xx",  #Can choose between mean or median
              shift = xx, #Difference between the null value and the sample mean
              as_extreme_as = xx, #Value of the summary statistic
              direction = "xx", #Specify direction of alternative hypothesis
              number_repetitions = 10000)
```

10. Interpret the p-value of the test in context of the problem.

11. Write a conclusion to the test in context of the problem. Include the appropriate population to which the study design allows us to generalize.

1.4.4 Take-home messages

1. We use bootstrapping—sampling with replacement—from the shifted data to generate a null distribution of simulated sample means. In order to ensure that the null distribution is centered at the null value, μ_0, we shift the data by adding $\mu_0 - \bar{x}$ to each value in the original data set. Note that if this value of the shift is negative, we are shifting the data down; if it is positive, we shift the data up.
2. The mean of the shifted data will equal the null value, μ_0, but the standard deviation of the shifted data will be the same as the standard deviation of the original data.
3. As in the one proportion scenario, we calculate the p-value for a simulation-based hypothesis test for a single mean by finding the proportion of simulated sample means that are as or more extreme (in the direction of H_A) as the observed sample mean, $\bar{x}$.

1.4.5 Additional notes

Use this space to summarize your thoughts and take additional notes on today's activity and material covered.

1.5 Activity 13: Body Temperature

1.5.1 Learning outcomes

- Given a research question involving a quantitative variable, construct the null and alternative hypotheses in words and using appropriate statistical symbols.
- Describe and perform a theory-based hypothesis test for a single mean.
- Interpret and evaluate a p-value for a theory-based hypothesis test for a single mean.

1.5.2 Terminology review

In today's activity, we will analyze quantitative data using theory-based methods. Some terms covered in this activity are:

- Normality
- t-distribution
- Degrees of freedom
- T-score

To review these concepts, see Chapters 11 and 17 in the textbook.

1.5.3 Body Temperature

It has long been reported that the mean body temperature of adults is 98.6°F. There have been a few articles that challenge this assertion. (LUETKEMEIER 2017) In 2018, a sample of 52 Stat 216 undergraduates were asked to report their body temperature. Is there evidence that body temperatures of adults differ from the known temperature of 98.6°F?

Ask a research question

1. Write out the null hypothesis in proper notation for this study.

2. Write out the alternative hypothesis in words for this study.

In general, the sampling distribution for a sample mean, $\bar{x}$, based on a sample of size n from a population with a true mean μ and true standard deviation σ can be modeled using a Normal distribution when certain conditions are met.

Conditions for the sampling distribution of $\bar{x}$ to follow an approximate Normal distribution:

- **Independence**: The sample's observations are independent.
- **Normality**: The data should be approximately normal or the sample size should be large.
  - $n < 30$: If the sample size n is less than 30 and the distribution of the data is approximately normal with no clear outliers in the data, then we typically assume the data come from a nearly normal distribution to satisfy the condition.

- $30 \le n < 100$: If the sample size n is between 30 and 100 and there are no particularly extreme outliers in the data, then we typically assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.
- $n \ge 100$: If the sample size n is at least 100 (regardless of the presence of skew or outliers), we typically assume the sampling distribution of $\bar{x}$ is nearly normal, even if the underlying distribution of individual observations is not.

Like we saw in Chapter **5**, we will not know the values of the parameters and must use the sample data to estimate them. Unlike with proportions, in which we only needed to estimate the population proportion, π, quantitative sample data must be used to estimate both a population mean μ and a population standard deviation σ. This additional uncertainty will require us to use a theoretical distribution that is just a bit wider than the standard Normal distribution. Enter the ***t*-distribution**!

As you can seen from Figure 1.1, the t-distributions (dashed and dotted lines) are centered at 0 just like a standard Normal distribution (solid line), but are slightly wider. The variability of a t-distribution depends on its degrees of freedom, which is calculated from the sample size of a study. (For a single sample of n observations or paired differences, the degrees of freedom is equal to $n-1$.) Recall from previous classes that larger sample sizes tend to result in narrower sampling distributions. We see that here as well. The larger the sample size, the larger the degrees of freedom, the narrower the t-distribution. (In fact, a t-distribution with infinite degrees of freedom actually IS the standard Normal distribution!)

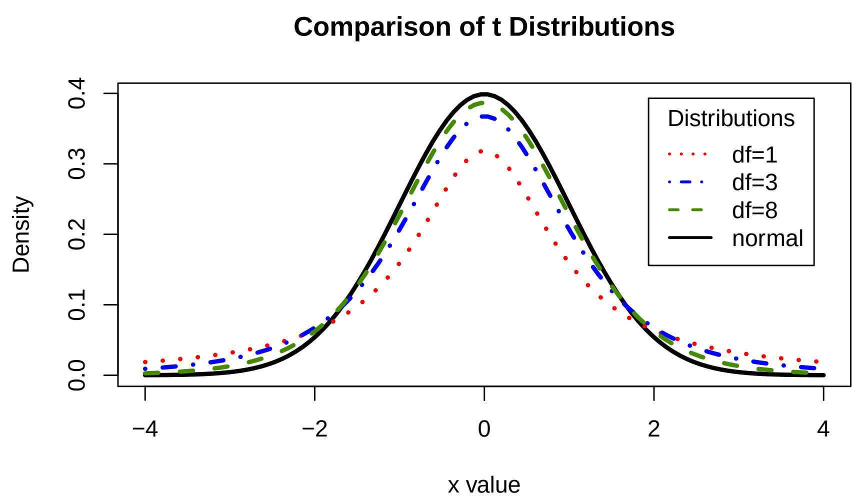


Figure 1.1: Comparison of the standard Normal vs t-distribution with various degrees of freedom

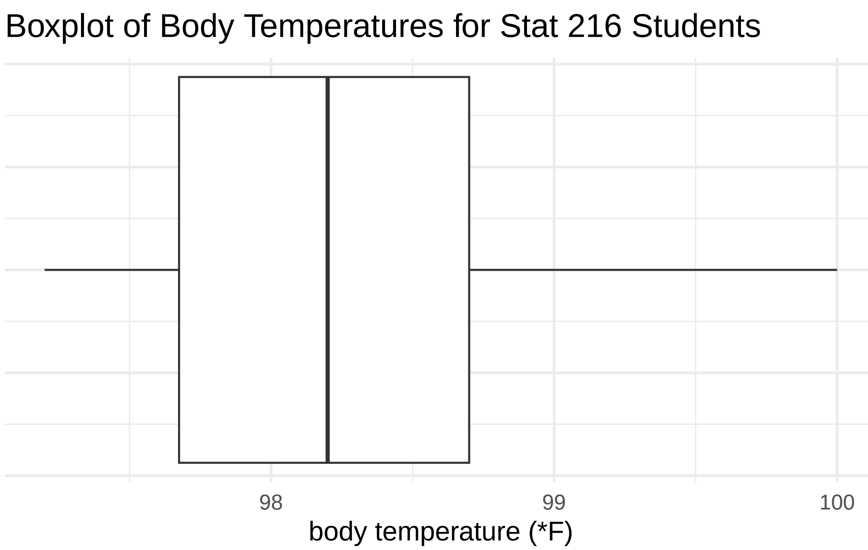
Summarize and visualize the data

The following code is used to create a boxplot of the data.

- Download the R script file upload to the R studio server.
- Open the R script file and highlight and run lines 1–14.

```
bodytemp <- read.csv("https://math.montana.edu/courses/s216/data/normal_temperature.csv")
bodytemp %>%
  ggplot(aes(x = Temp))+
  geom_boxplot()+
  labs(title="Boxplot of Body Temperatures for Stat 216 Students",
       x = "body temperature (*F)") +
```

```r
      theme(axis.text.y = element_blank(),
        axis.ticks.y = element_blank()) # Removes y-axis ticks
```



- Highlight and run lines 17 - 18 to get the summary statistics for the variable `Temp`.

```r
bodytemp %>%
  summarise(favstats(Temp))
```

```
#>    min     Q1 median   Q3 max     mean        sd  n missing
#> 1 97.2 97.675   98.2 98.7 100 98.28462 0.6823789 52       0
```

Check theoretical conditions

3. Report the sample size of the study. Give appropriate notation.

4. Report the sample mean of the study. Give appropriate notation.

5. How do you know the independence condition is met for these data?

6. Is the normality condition met to use the theory-based methods for analysis? Explain your answer.

Use statistical inferential methods to draw inferences from the data

To find the standardized statistic for the mean we will use the following formula:

$$T = \frac{\bar{x} - \mu_0}{SE(\bar{x})},$$

where the standard error of the sample mean is:

$$SE(\bar{x}) = \frac{s}{\sqrt{n}}.$$

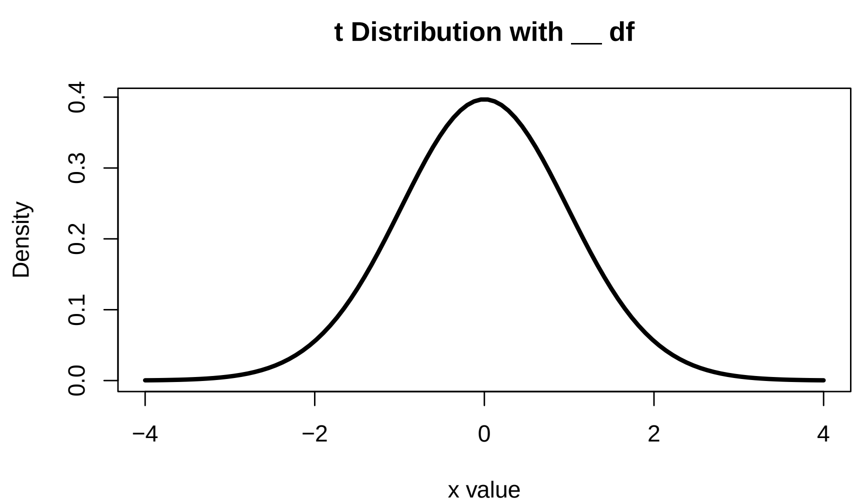
7. Calculate the standard error of the sample mean.

8. Interpret the standard error in context of the study.

9. Calculate the standardized mean.

10. We model a single mean with a t-distribution with $n - 1$ degrees of freedom. Calculate the degrees of freedom for this study and use it to fill in the blank in the title of the t-distribution displayed below.

11. Mark the value of the standardized statistic on the t-distribution and illustrate how the p-value is found.



To find the p-value for the theory-based test in R:

- Enter the value for the standardized statistic for `xx` in the `pt` function.
- Enter the degrees of freedom for `yy` in the `pt` function.
- Highlight and run line 24.

```
pt(xx, df=yy, lower.tail=FALSE)
```

12. What does this p-value mean, in the context of the study? Hint: it is the probability of what…assuming what?

13. Write a conclusion to the test in context of the study.

14. Can we generalize the results of the study to all adults? Explain your answer.

1.5.4 Take-home messages

1. In order to use theory-based methods for a quantitative variable, the independent observational units and normality conditions must be met.
2. In order to find a theory-based p-value, we use R to calculate the area under a t-distribution with $n-1$ degrees of freedom (df) that is at or more extreme than the observed T-score. To find a two-sided p-value using theory-based methods we need to multiply the one-sided p-value by 2.
3. A t^* multiplier is found by obtaining the bounds of the middle X% (X being the desired confidence level) of a t-distribution with $n-1$ df.

1.5.5 Additional notes

Use this space to summarize your thoughts and take additional notes on today's activity and material covered

"Average Driving Distance and Fairway Accuracy." 2008. https://www.pga.com/ and https://www.lpga.com/.

Banton, et al, S. 2022. "Jog with Your Dog: Dog Owner Exercise Routines Predict Dog Exercise Routines and Perception of Ideal Body Weight." *PLoS ONE* 17(8).

Bhavsar, et al, A. 2022. "Increased Risk of Herpes Zoster in Adults ≥50 Years Old Diagnosed with COVID-19 in the United States." *Open Forum Infectious Diseases* 9(5).

Bulmer, M. n.d. "Islands in Schools Project." https://sites.google.com/site/islandsinschoolsprojectwebsite/home.

"Bureau of Transportation Statistics." 2019. https://www.bts.gov/.

"Child Health and Development Studies." n.d. https://www.chdstudies.org/.

Darley, J. M., and C. D. Batson. 1973. ""From Jerusalem to Jericho": A Study of Situational and Dispositional Variables in Helping Behavior." *Journal of Personality and Social Psychology* 27: 100–108.

Davis, Smith, A. K. 2020. "A Poor Substitute for the Real Thing: Captive-Reared Monarch Butterflies Are Weaker, Paler and Have Less Elongated Wings Than Wild Migrants." *Biology Letters* 16.

Du Toit, et al, G. 2015. "Randomized Trial of Peanut Consumption in Infants at Risk for Peanut Allergy." *New England Journal of Medicine* 372.

Edmunds, et al, D. 2016. "Chronic Wasting Disease Drives Population Decline of White-Tailed Deer." *PLoS ONE* 11(8).

Education Statistics, National Center for. 2018. "IPEDS." https://nces.ed.gov/ipeds/.

"Great Britain Married Couples: Great Britain Office of Population Census and Surveys." n.d. https://discovery.nationalarchives.gov.uk/details/r/C13351.

Group, TODAY Study. 2012. "A Clinical Trial to Maintain Glycemic Control in Youth with Type 2 Diabetes." *New England Journal of Medicine* 366: 2247–56.

Hamblin, J. K., K. Wynn, and P. Bloom. 2007. "Social Evaluation by Preverbal Infants." *Nature* 450 (6288): 557–59.

Hirschfelder, A., and P. F. Molin. 2018. "I Is for Ignoble: Stereotyping Native Americans." Retrieved from https://www.ferris.edu/HTMLS/news/jimcrow/native/homepage.htm.

Hutchison, R. L., and M. A. Hirthler. 2013. "Upper Extremity Injuies in Homer's Iliad." *Journal of Hand Surgery (American Volume)* 38: 1790–93.

"IMDb Movies Extensive Dataset." 2016. https://kaggle.com/stefanoleone992/imdb-extensive-dataset.

Kalra, et al., Dl. 2022. "Trustworthiness of Indian Youtubers." Kaggle. https://doi.org/10.34740/KAGGLE/DSV/4426566.

Keating, D., N. Ahmed, F. Nirappil, Stanley-Becker I., and L. Bernstein. 2021. "Coronavirus Infections Dropping Where People Are Vaccinated, Rising Where They Are Not, Post Analysis Finds." *Washington Post*. https://www.washingtonpost.com/health/2021/06/14/covid-cases-vaccination-rates/.

Laeng, Mathisen, B. 2007. "Why Do Blue-Eyed Men Prefer Women with the Same Eye Color?" *Behavioral Ecology and Sociobiology* 61(3).

Levin, D. T. 2000. "Race as a Visual Feature: Using Visual Search and Perceptual Discrimination Tasks to Understand Face Categories and the Cross-Race Recognition Deficit." *Journal of Experimental Psychology* 129(4).

LUETKEMEIER, et al., M. 2017. "Skin Tattoos Alter Sweat Rate and Na+ Concentration." *Medicine and Science in Sports and Exercise* 49(7).

Madden, et al, J. 2020. "Ready Student One: Exploring the Predictors of Student Learning in Virtual Reality." *PLoS ONE* 15(3).

Miller, G. A. 1956. "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information." *Psychological Review* 63(2).

Moquin, W., and C. Van Doren. 1973. "Great Documents in American Indian History." Praeger.

"More Americans Are Joining the 'Cashless' Economy." 2022. https://www.pewresearch.org/short-reads/2022/10/05/more-americans-are-joining-the-cashless-economy/.

National Weather Service Corporate Image Web Team. n.d. "National Weather Service – NWS Billings." https://w2.weather.gov/climate/xmacis.php?wfo=byz.

O'Brien, Lynch, H. D. 2019. "Crocodylian Head Width Allometry and Phylogenetic Prediction of Body Size in Extinct Crocodyliforms." *Integrative Organismal Biology* 1.

"Ocean Temperature and Salinity Study." n.d. https://calcofi.org/.

"Older People Who Get Covid Are at Increased Risk of Getting Shingles." 2022. https://www.washingtonpost.com/health/2022/04/19/shingles-and-covid-over-50/.

"Physician's Health Study." n.d. https://phs.bwh.harvard.edu/.

Porath, Erez, C. 2017. "Does Rudeness Really Matter? The Effects of Rudeness on Task Performance and Helpfulness." *Academy of Management Journal* 50.

Quinn, G. E., C. H. Shin, M. G. Maguire, and R. A. Stone. 1999. "Myopia and Ambient Lighting at Night." *Nature* 399 (6732): 113–14. https://doi.org/10.1038/20094.

Ramachandran, V. 2007. "3 Clues to Understanding Your Brain." https://www.ted.com/talks/vs_ramachandran_3_clues_to_understanding_your_brain.

"Rates of Laboratory-Confimed COVID-19 Hospitalizations by Vaccination Status." 2021. CDC. https://covid.cdc.gov/covid-data-tracker/#covidnet-hospitalizations-vaccination.

Richardson, T., and R. T. Gilman. 2019. "Left-Handedness Is Associated with Greater Fighting Success in Humans." *Scientific Reports* 9 (1): 15402. https://doi.org/10.1038/s41598-019-51975-3.

Stephens, R., and O. Robertson. 2020. "Swearing as a Response to Pain: Assessing Hypoalgesic Effects of Novel "Swear" Words." *Frontiers in Psychology* 11: 643–62.

Stewart, E. H., B. Davis, B. L. Clemans-Taylor, B. Littenberg, C. A. Estrada, and R. M. Centor. 2014. "Rapid Antigen Group a Streptococcus Test to Diagnose Pharyngitis: A Systematic Review and Meta-Analysis" 9 (11). https://doi.org/10.1371/journal.pone.0111727.

Stroop, J. R. 1935. "Studies of Interference in Serial Verbal Reactions." *Journal of Experimental Psychology* 18: 643–62.

Subach, et al, A. 2022. "Foraging Behaviour, Habitat Use and Population Size of the Desert Horned Viper in the Negev Desert." *Soc.Open Sci* 9.

Sulheim, S., A. Ekeland, I. Holme, and R. Bahr. 2017. "Helmet Use and Risk of Head Injuries in Alpine Skiers and Snowboarders: Changes After an Interval of One Decade" 51 (1): 44–50. https://doi.org/10.1136/bjsports-2015-095798.

"Titanic." n.d. http://www.encyclopedia-titanica.org.

"US COVID-19 Vaccine Tracker: See Your State's Progress." 2021. Mayo Clinic. https://www.mayoclinic.org/coronavirus-covid-19/vaccine-tracker.

US Environmental Protection Agency. n.d. "Air Data – Daily Air Quality Tracker." https://www.epa.gov/outdoor-air-quality-data/air-data-daily-air-quality-tracker.

Wahlstrom, et al, K. 2014. "Examining the Impact of Later School Start Times on the Health and Academic Performance of High School Students: A Multi-Site Study." *Center for Applied Research and Educational Improvement*.

Watson, et al., N. 2015. "Recommended Amount of Sleep for a Heathy Adult: A Joint Consensus Statement of the American Academy of Sleep Medicine and Sleep Research Society." *Sleep* 38(6).

Weiss, R. D. 1988. "Relapse to Cocaine Abuse After Initiating Desipramine Treatment." *JAMA* 260(17).

"Welcome to the Navajo Nation Government: Official Site of the Navajo Nation." 2011.Retrieved from https://www.navajo-nsn.gov/.

Wilson, Woodruff, J. P. 2016. "Vertebral Adaptations to Large Body Size in Theropod Dinosaurs." *PLoS ONE* 11(7).