Review of Module 6

Vocabulary Review

Sample statistics for a single quantitative variable

- **Mean**, $\bar{x}$, the average value in the data set

Interpreting the sample mean:

Include:

- Summary measure (mean)
  - What numerical value are we calculating?
  - This is dependent on the type of variable(s) in our study
  - Give the value of the statistic
- Context
  - Observational units/cases – what or whom are we collecting data on
  - Variable of interest

Example Activity 9:

- $\bar{x} = 52.487$
  - The mean 6-year graduation rate for undergraduate students at 4-year US degree-granting higher education institutions in 2018 is 52.487 percentage points.

units

- **Median**: value at the 50th percentile; approximately 50% of data values are at or below the value of the median.
- **Quartile 1** (lower quartile), $Q1$: value at the 25th percentile; approximately 25% of data values are at or below the value of $Q1$.
  - Could also be interpretated as approximately 75% of data values are at and above the value of $Q1$.
- **Quartile 3** (upper quartile), $Q3$: value at the 75th percentile; approximately 75% of data values are at or below the value of $Q3$.

Interpreting the Quartiles:

Include:

- *Reference associated percentile for given quartile*
- *Data values*
- *At or above/below given quartile value*
- *Context:*
  - *Observational units/cases*
  - *Response variable*

Example Activity 9:

- $Q_3 = 67$
  - Approximately 75% of 4-year US degree-granting higher education institutions in 2018 have a 6-year graduation rate for undergraduate students of 67% and below.

OUs

variable

value

units

  - OR

- Approximately 25% of 4-year US degree-granting higher education institutions in 2018 have a 6-year graduation rate for undergraduate students of 67% and above.

- **Sample standard deviation**, s: on average, each value in the data set is s units from the mean of the data set.

Interpreting the sample standard deviation:
Include:

- Data values
- Distance from sample mean, on average
- Context:
  - Observational units/cases
  - Response variable

Example Activity 9:

- s = 20.632
  - In 2018, each 4-year US degree-granting higher education institution's graduation rate is 20.632% from the mean graduation rate of 52.487%, on average.
- **Interquartile range**: the range of the data between the two quartiles: $IQR = Q3 - Q1$.

- Notation for the mean
  - *Parameter: μ (mu)*
  - *Statistic: $\bar{x}$ (xbar)*
- Visualize with:
  - *Histogram*
  - *Dotplot*
  - *Boxplot*
- Examples of the statistic written in words:
  - *Mean GPA of Stat 216 students*
    - Variable: GPA of Stat 216 students
  - *Mean temperature for Bozeman, MT in December*
    - Variable: *temperature for Bozeman, MT in December*
- Four features used for comparing quantitative distributions
  - Shape
    - Symmetric
      - The mean will be approximately the same value as the median
    - Left-skewed (negatively skewed)
      - The long tail of the distribution is in the direction of the skew
      - The mean will be less than the value of the median
    - Right-skewed (positively skewed)
      - The mean will be greater than the value of the median

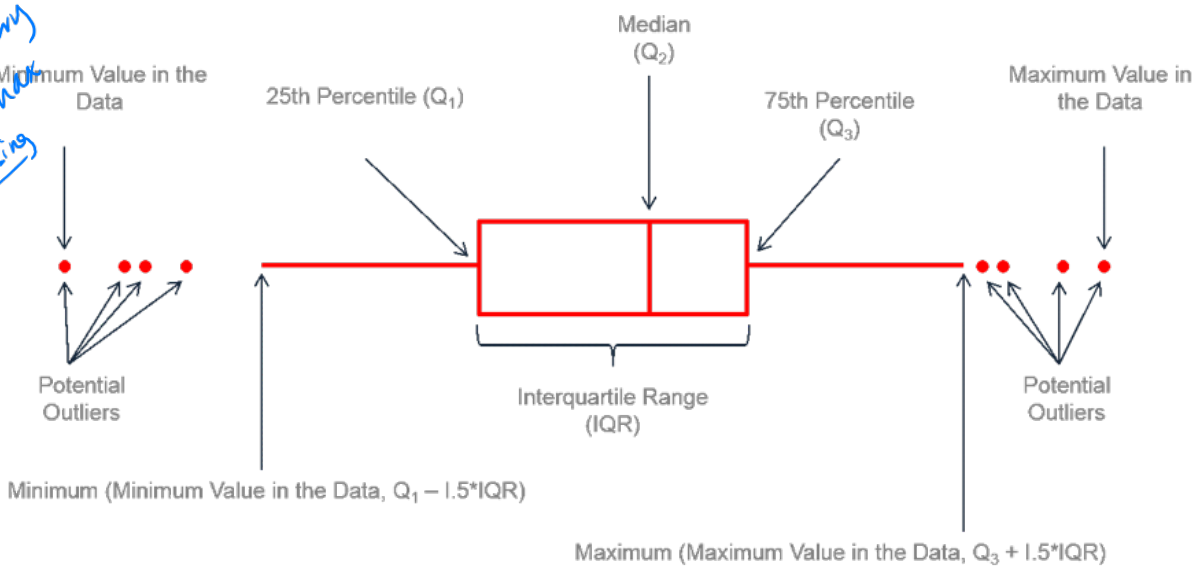
  - Center
    - Mean or Median
  - Spread
    - Standard deviation or IQR

visually on a boxplot

- Outliers?
  - Values in the data set $> Q_3 + 1.5xIQR$
  - Values in the data set $< Q_1 - 1.5xIQR$

- Components of the boxplot (box and whisker plot)
  - Minimum (smallest value)
  - Q_1 = Lower Quartile = 25th Percentile = median of lower half of the ordered data values (not including median)
  - Median = middle value = 50th Percentile
  - Q_3 = Upper Quartile = 75th Percentile = median of upper half of the ordered data values (not including median)
  - Maximum (largest value)

fivestats & 5 number summary min, Q_1, median, Q_3, max mean, s, n, missing



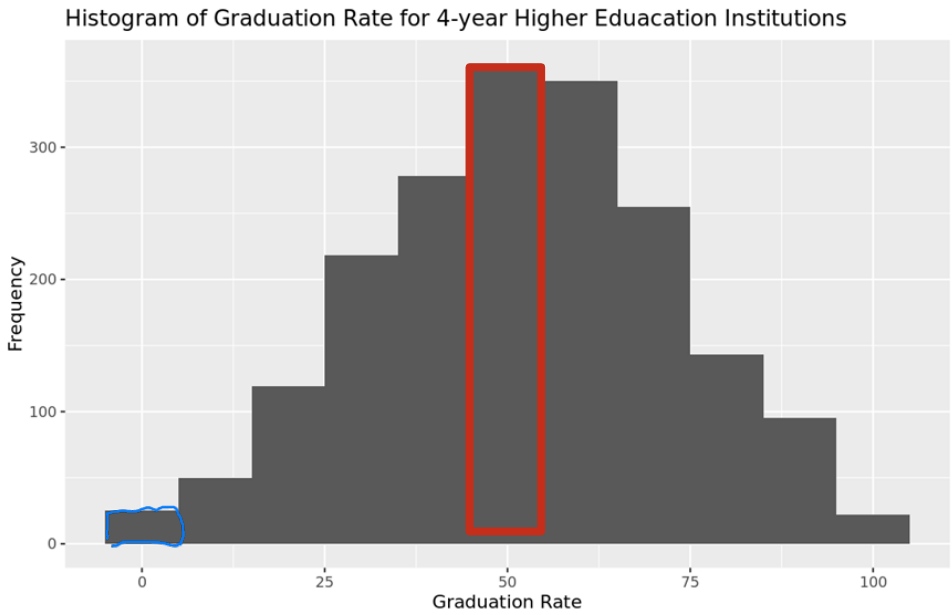
- In the first part of the Activity 9, we looked at the distribution of graduation rates for 4-year higher education institutions. The summary statistics are shown below.

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

- Notice there are 49 missing values – this means that 49 institutions in the data set did not report their graduation rate. This indicates that there is non-response bias in the study.

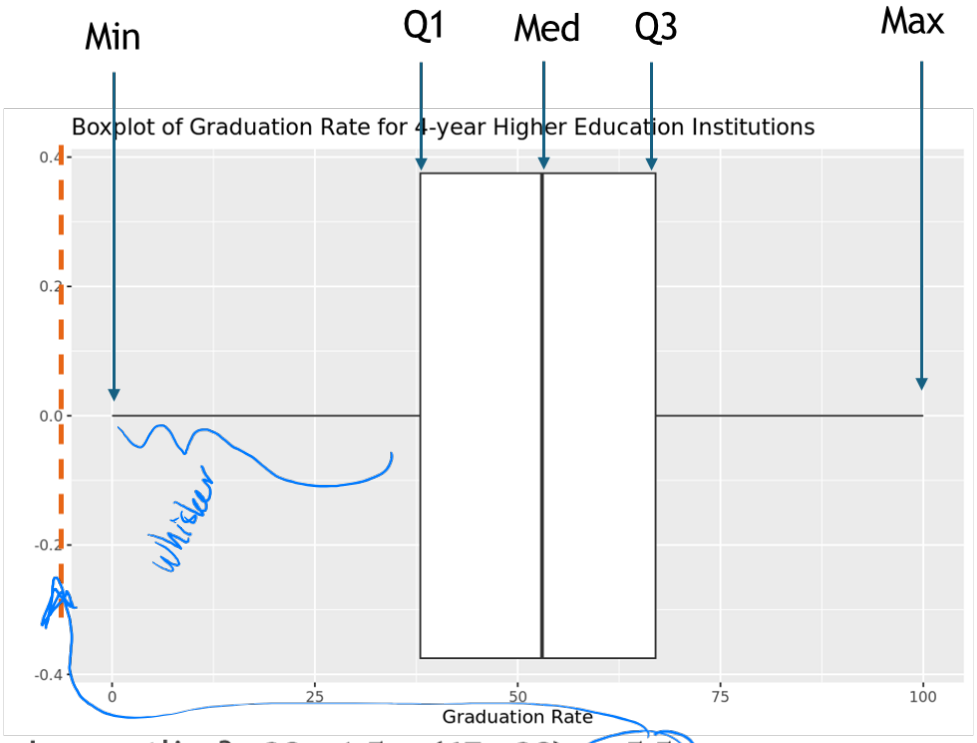
Review of Module 6

From the histogram and dot plot we can see that the distribution of graduation rates is symmetric.



- Shape = symmetric
- Highest frequency bin = 50 – 60 % points

The boxplot shows no outliers in the distribution.



- Low outlier? $38 - 1.5 \times (67 - 38) = -5.5$
  - Minimum = 0 is inside the lower fence → no low outliers
- High outlier? $67 + 1.5 \times (67 - 38) = 110.5$
  - Maximum = 100 is inside the upper fence → no high outliers

Review of Module 6

Parameter of Interest: What information do we want to know about the population?:

- The parameter of interest is used in the hypotheses statements, in conclusions, and in many interpretations!

Include:

- Reference of the population (true, long-run, population, all)
    - Clearly refer to the population
- Summary measure (mean)
    - What numerical value are we calculating?
    - This is dependent on the type of variable(s) in our study
- Context
    - Observational units/cases – what or whom are we collecting data on
    - Variable of interest

Example Activity 10:

- ***μ represents the true mean number of hours of sleep per night for MSU students***

Hypothesis Test (test of significance/inference): test to show evidence based on the sample statistic against the null hypothesis

Hypotheses:

- Null Hypothesis: This is the known claim that we are trying to disprove; may be based on random chance
    - $H_0: \mu = 0$
- Alternative: this is the claim we are testing that is based on the research question
    - $H_A: \mu \begin{Bmatrix} > \\ \neq \\ < \end{Bmatrix} 0$
        - The direction of the alternative (the sign) is determined by the research question
- **Example Activity 10:** Is there evidence that MSU students get less than the recommended 7 hours of sleep per night, on average?
    - *H_0: **The true mean number of hours of sleep per night for MSU students is 7 hrs***
        - *Note: we are assuming that the the average amount of sleep per night for MSU students is the same as the recommend amount of sleep of 7 hrs*
        - $H_o: \mu = 7$
    - *H_A: **The true mean number of hours of sleep per night for MSU students is less than 7 hrs.***
        - *The direction of the alternative is less than because the research question asks for evidence that MSU students get less than the recommended 7 hours of sleep*
        - $H_A: \mu < 7$

Review of Module 6

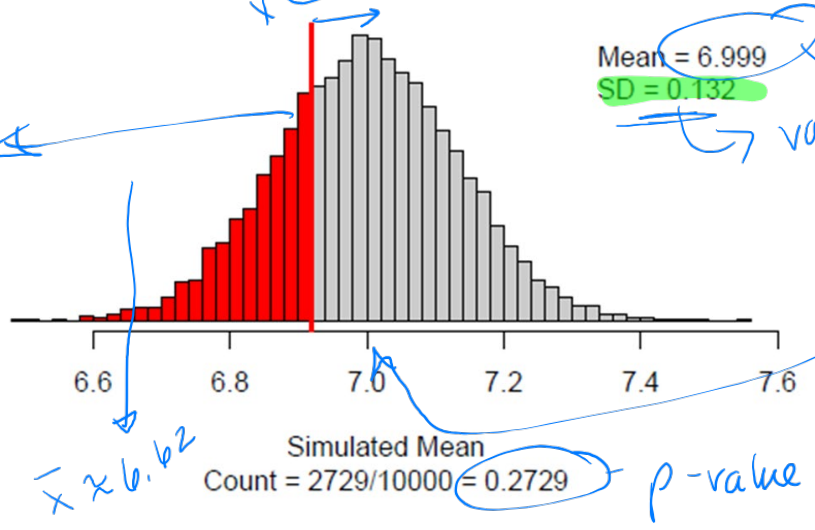
Simulation methods for a single mean:

Null Distribution: simulation distribution created based on the assumption that the null hypothesis is true; centered at the null value

- **How can we use cards to simulate one sample for the null distribution?**
  - Label cards with the values for each observational units
  - Shift the data by $\mu_0 - \bar{x}$
  - Sample with replacement n times
  - Plot the mean from each simulated sample

- **Example Activity 10**: Label 49 cards with the number of sleep per night for MSU students. Add the value (7 – (6.919) = 0.081) to each value. Sample with replacement 49 times. Plot the mean number of hours of sleep for one simulated sample.
- We use the one mean test function in catstats.

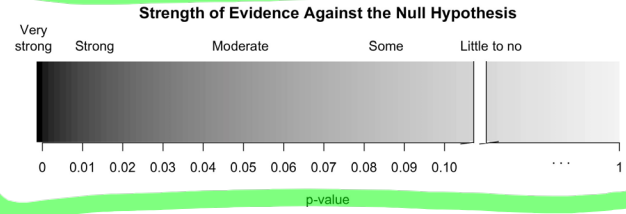
```
one_mean_test(sleep$SleepHours,   #Enter the object name and variable
              null_value = 7,
              summary_measure = "mean",  #can choose between mean or median
              shift =0.081,  #Difference between the null value and the sample mean
              as_extreme_as = 6.919,  #Value of the summary statistic
              direction = "less",  #Specify direction of alternative hypothesis
              number_repetitions = 10000)
```



Review of Module 6

Strength of Evidence: How much evidence does the p-value provide **against** the null?

- Use the guidelines for the strength of evidence



- The smaller the p-value the MORE evidence there is against the null hypothesis

There are FOUR things we ask about the p-value (we will learn the 4th in Module 7)

Evaluation of a p-value:

- How much evidence does the p-value provide AGAINST the null hypothesis?
- **Example Activity 10:** The simulation p-value for this study was found to be 0.2729.
  - ***There is little to no evidence against the null hypothesis that the true mean number of hours of sleep per night for MSU students is 7 hours.***

Interpretation of a p-value:

- What the p-value measures: the probability of observing the sample statistic or more extreme if the null hypothesis is true (Don't forget the context!)
- Include in the interpretation:
  - Statement about probability (in x% of simulated samples, in x out of 1000 simulated samples, with a probability of x%)
  - Statistic in context (give the value and in words what the statistic represents)
  - more extreme (direction of the alternative)
  - If the null hypothesis is true in context (give the null value and in words what the null represents)
    - Note: context only needs to be included in either the statistic OR the null

$P(\bar{x} \text{ or more extreme} \mid H_0 \text{ is true}) = p\text{-value}$

- **Example Activity 10:** The simulation p-value for this study was found to be 0.2729.
  - ***We would observe a sample mean of 6.919 hours or less with a probability of 0.2729, if we assume the true mean number of hours of sleep per night for MSU students is 7 hours.***

$\bar{x}$ $<$ p-value $H_0: \mu = ?$

OR

- ***If the true mean number of hours of sleep per night for MSU students is 7 hours, we would observe a sample mean of 6.919 hours or less with a probability of 27.29%.***

OR

p-value $\bar{x}$ $<$ $H_0: \mu = ?$

- ***There is 27.29% chance we would observe a sample mean of 6.919 hours or less, if we assume the true mean number of hours of sleep per night for MSU students is 7 hours.***

Conclusion: Answers the research question. Write a conclusion as the amount of evidence in support of the alternative.

- **Example Activity 10**: The simulation p-value for this study was found to be 0.2729.
  - ***There is little to no evidence*** *that the* ***true mean number of hours of sleep per night for MSU students is less than 7 hours.***

Simulation methods to estimate the parameter of interest (Confidence Interval)

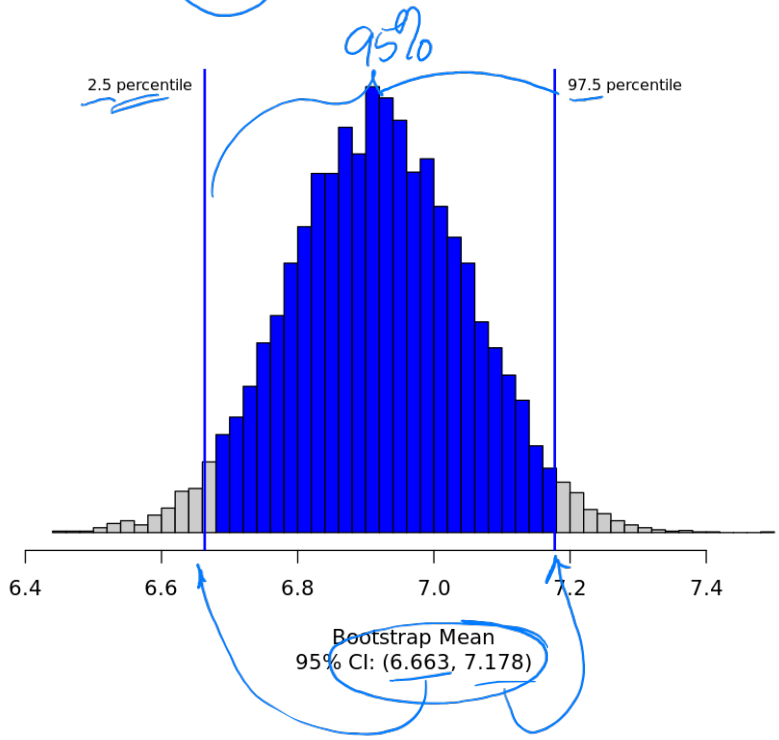
Bootstrap Distribution: simulation distribution created based on sampling with replacement from the original statistic; centered at the sample statistic, ($\bar{x}$)

- **How can we use cards to simulate one sample for the bootstrap distribution?**
  - Label cards with the value for each observational unit
  - Sample with replacement n times
  - Plot the mean from each simulated sample

- We use the one mean CI function in catstats.

Example: Activity 10

```
one_mean_CI(sleep$SleepHours, #Enter the name of the variable
            summary_measure = "mean", #choose the mean or median
            number_repetitions = 10000, #Number of simulations
            confidence_level = 0.95) #Confidence level in decimal form
```



Review of Module 6

Interpretation of a confidence interval:

- Include in the interpretation:
  - How confident you are (90%, 95%, 99%)
  - Parameter of interest *in context*
    - *Population word (true, long-run, population)*
    - *Summary measure (difference in proportion)*
    - *Observational units*
    - *Variable of interest*
  - Calculated Interval

Example Activity 10: ***We are 95% confident, the true mean number of hours of sleep per night for MSU students is between 6.663 and 7.178 hours.***

units