

Chapter Four: Comparing Distributions

Review

- **Chapter One**
 - Datasets and fundamental vocabulary
 - Quantitative versus Categorical variables
- **Chapter Two: Categorical Variables**
 - Visualizing *one* categorical variable
 - Frequency tables and relative frequency tables
 - Visualizing *two* categorical variables
 - Contingency tables
 - Conditional and Marginal Distributions
- **Chapter Three: Quantitative Variables**
 - Data visualization: Histograms and stem-and-leaf plots
 - Summary Statistics: Mean, Median, Range, IQR

So we have looked at the interaction of two categorical variables and how to analyze one quantitative variable. We will now learn how to analyze data when we have a categorical variable and a quantitative variable.

Just as a review:

- **Quantitative Variable:**

- **Categorical Variable:**

So when we come across a relationship between a categorical and a quantitative variable, we:

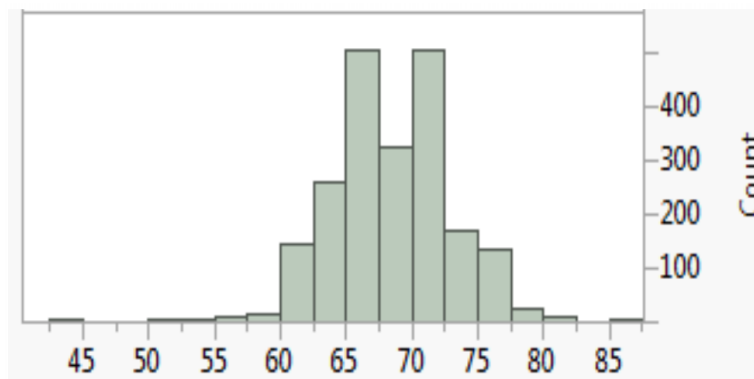
1.

2.

When we are trying to establish an association we give each variable roles. Often:

- **Response Variable:**
- **Explanatory Variable:**

Let's illustrate how we compare categorical and quantitative variables with an example. Suppose we have data about the height of STAT 101 students shown below. Describe the shape of the distribution.

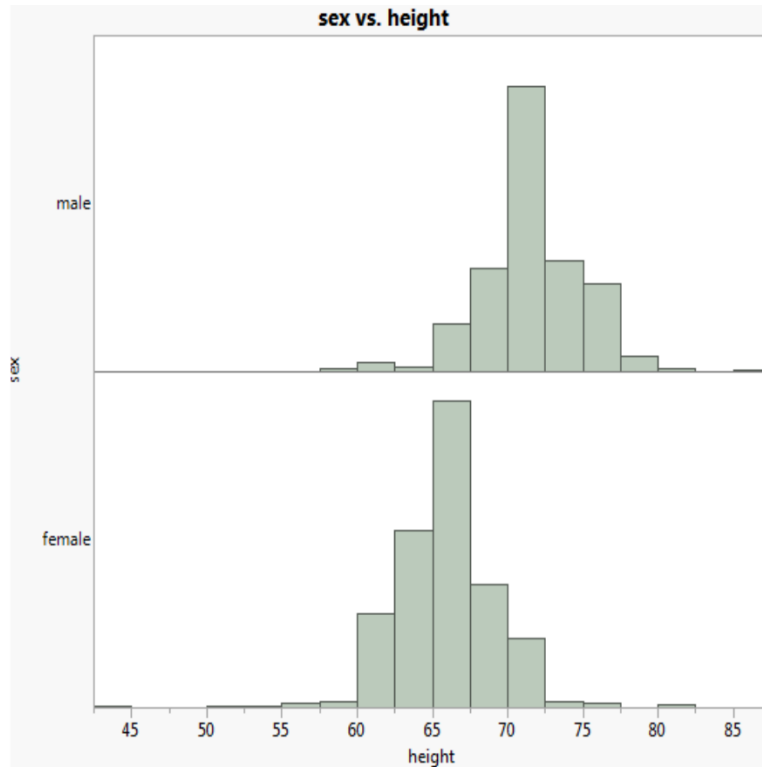


Other variables were also collected in the study:

- Sex
- Year in College
- Number of Brothers and Sisters
- Exercise Status

Do you think there could be an association between height and any of these variables?

So we have two variables in this case, height in inches which is the _____, and we have sex which is the _____. One way we can compare these two variables is by comparing them graphically. We can create two histograms, one for each _____ of our explanatory variable.



Let's describe these two histograms and compare them.

We can also compare the two distribution numerically. We can present a table containing a five number summary for each category.

| Heights | | |
|---------|--------|------|
| | Female | Male |
| Min | 44 | 58 |
| Q1 | 64 | 70 |
| Median | 66 | 72 |
| Q3 | 68 | 73 |
| Max | 82 | 85 |

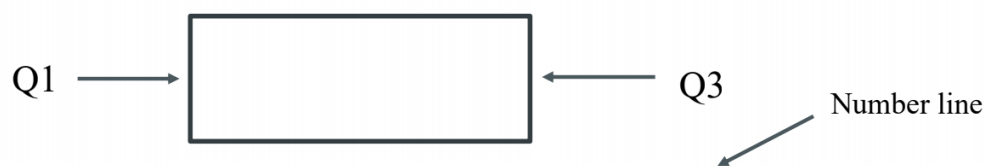
While this table can illustrate a lot about the distribution, a better way might be to make a box plot.

Box Plots

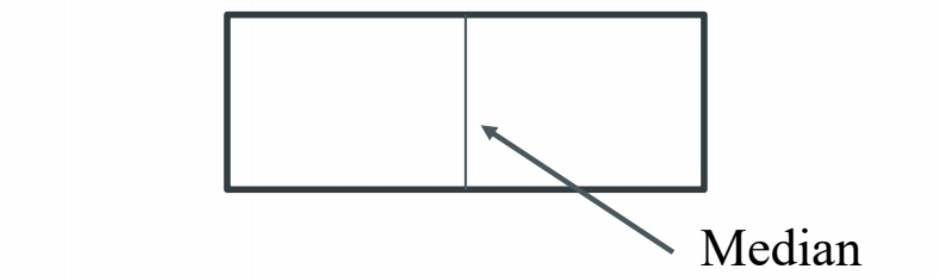
A boxplot, essentially communicates the same information as a five number summary but is visually more appealing and can help make comparisons between categories of the explanatory variable.

1.

2.



3.

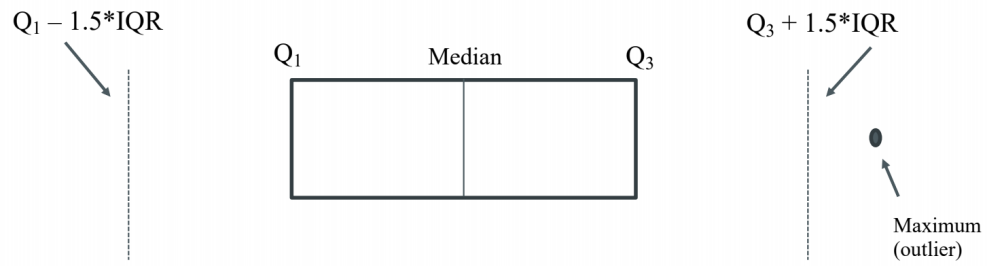


4.

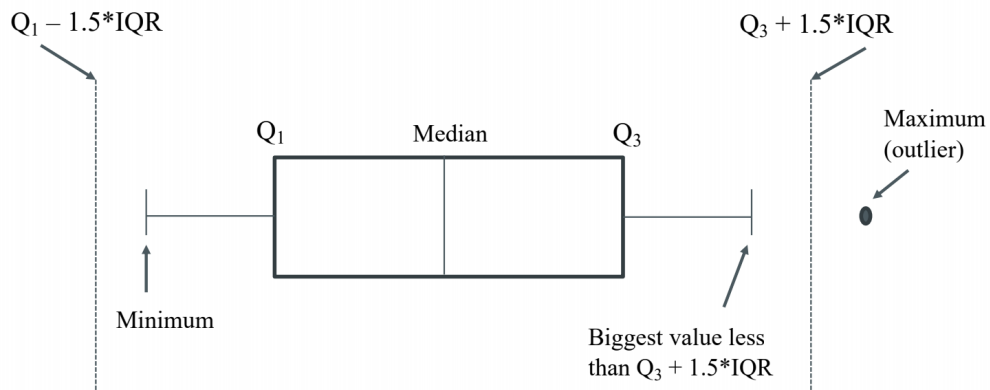
•

•

5.



6.



Now let's practice creating a box plot by hand.

Example Below are 20 sampled ACT scores from a small group of students. Create a boxplot from the following data.

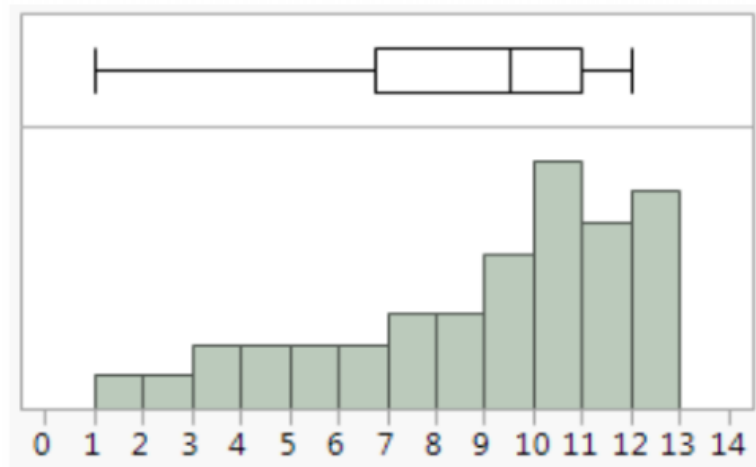
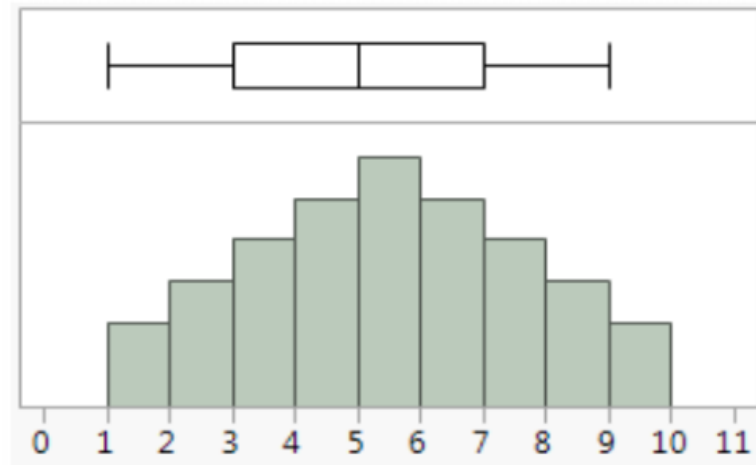
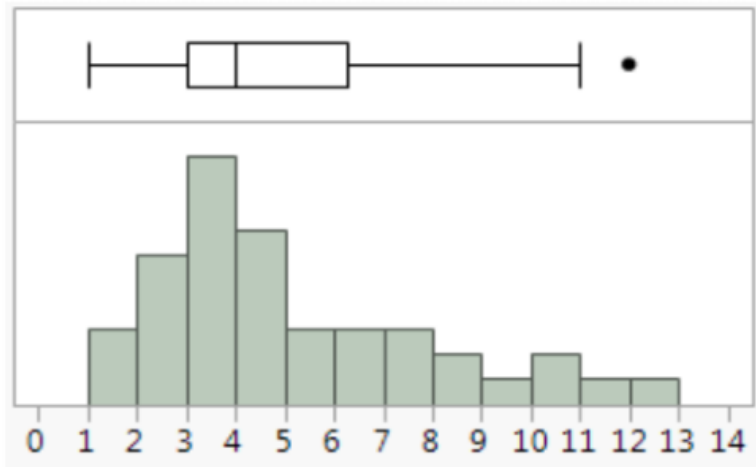
25, 26, 26, 27, 28, 28, 28, 28, 28, 28,
26, 29, 29, 29, 29, 30, 30, 31, 31, 32

Interpretation

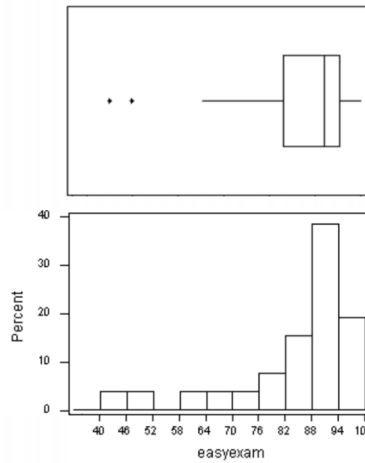
From a boxplot, we can learn about:

-
-

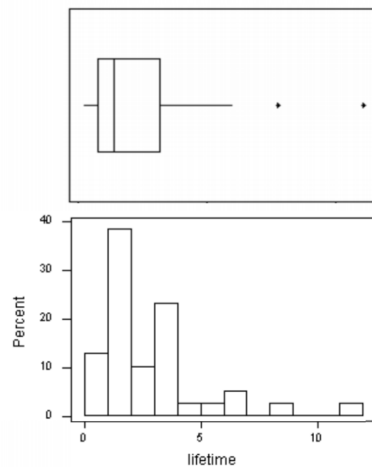
We can also identify _____ and _____ but not _____ from a boxplot.



Example Suppose a chemistry exam was given to an 8:00am lecture section of 26 students. Describe the shape of the distribution in terms of Modes, Symmetry, and Outliers based on the picture given, using the box plot.

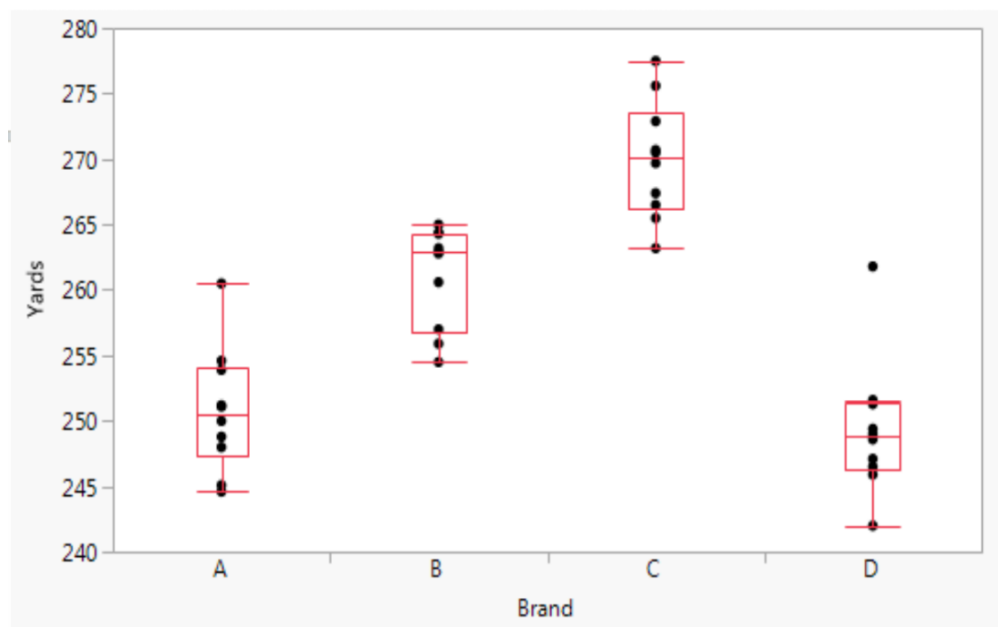


Example A random sample of 39 batteries were tested to find their lifetime in years. Describe the shape of the distribution in terms of Modes, Symmetry, and Outliers based on the picture given, using the box plot.



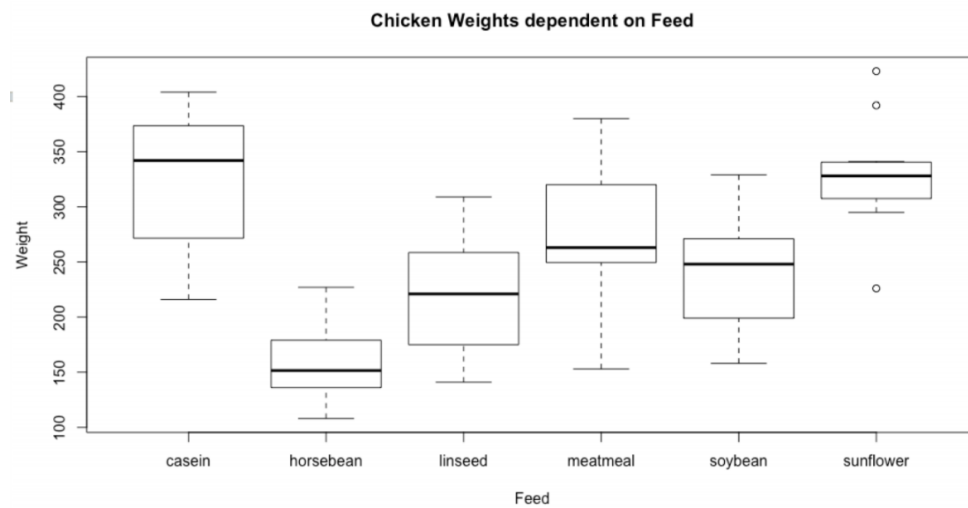
So when do we choose to use a boxplot or a histogram? If our goal is to describe the distribution of a single quantitative variable, then we use a _____. However, if our goal is to compare the distribution of a quantitative variable between groups or categories then we should use a _____.

Example Based on the plot below of golfball brands versus distance (in yards) to answer the following questions.



1. Which brand has gone the largest distance? The smallest distance?
2. Which brand has a median around 264 yards?
3. Which brand has the smallest IQR?
4. True or False, Q_1 for brand B is bigger than Q_3 for brand A.

Example Based on the plot below of food type versus chicken weight (in grams) to answer the following questions.



1. Which feed makes chickens the lightest?
2. Which feed has outliers?
3. Which feed has the smallest IQR?
4. What is the shape of the distribution of each feed?