**Name: Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Classwork 4**

**Exploring Variation (Continued)**

**Summaries and Graphs**

**Textbook Chapter: 3**

**Advanced Learning Objectives:**

* Describe a distribution in terms of Shape, Center, Spread, and Outliers
* Interpret Summary Statistics and Data Visualizations
* Distinguish between Histograms, Boxplots, and Bar Charts

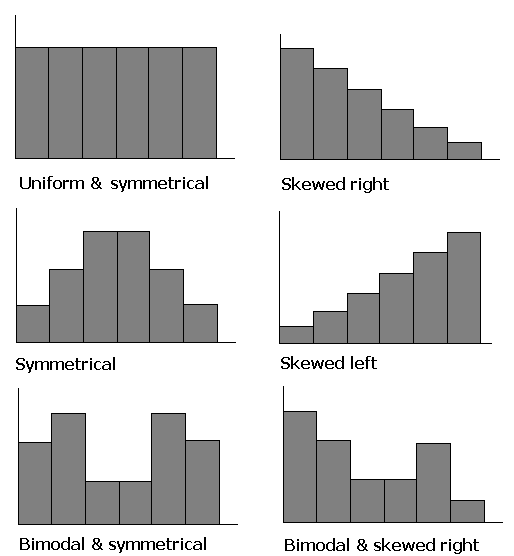
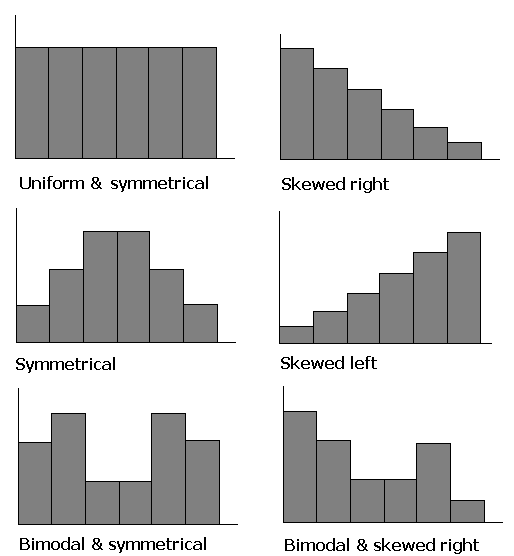
**Exploring Variation: Shape, Center, Spread, and Outliers**

**Quantitative Variables and SHAPE**

Describe each histogram in terms of:

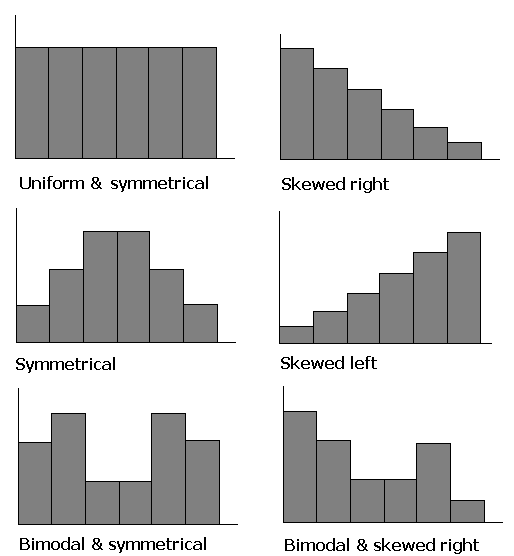
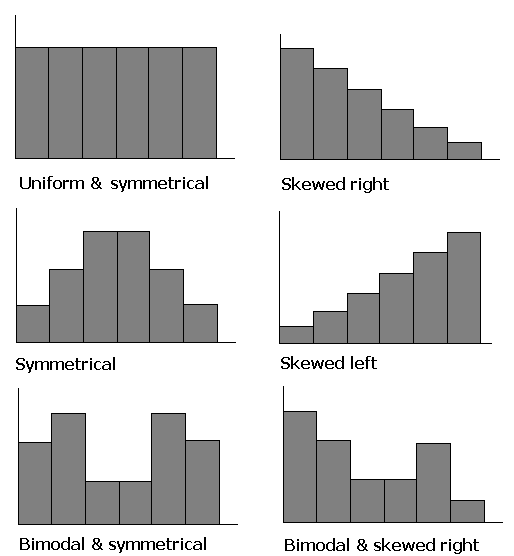
*Symmetry* (symmetrical or skewed right/left)

*Modality* (unimodal, bimodal, or uniform)



\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_



\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

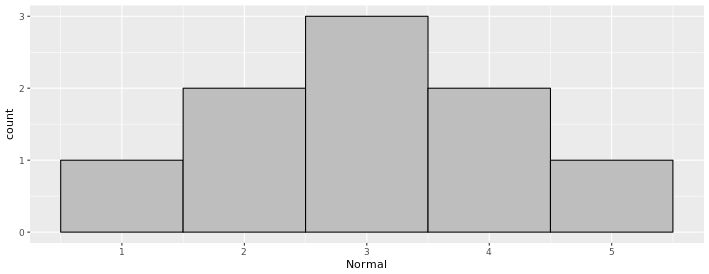
Which of the above histograms represents a “normal” distribution (circle it)?

Does it make sense to describe a bar plot of categorical variables using these descriptions?

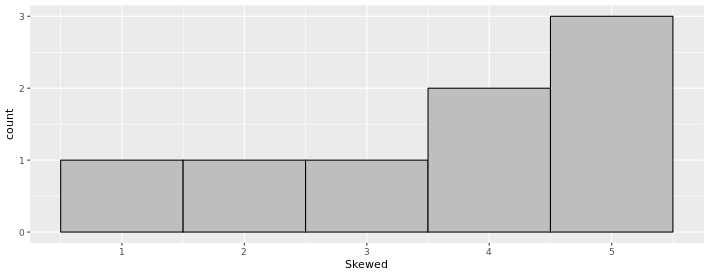
**Quantitative Variables and CENTER**

Describe each distribution in terms of:

*Mean, Median, and/or Mode*

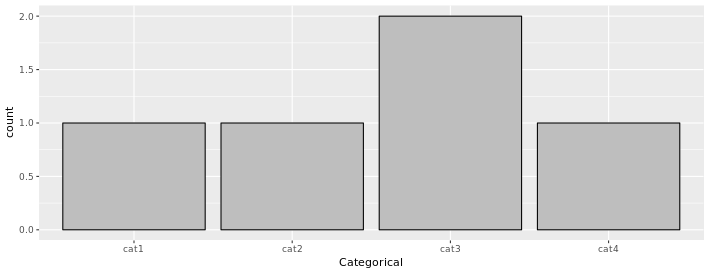


**Mean = \_\_\_\_\_\_\_\_\_\_, Median = \_\_\_\_\_\_\_\_\_\_, Mode = \_\_\_\_\_\_\_\_\_\_**



**Mean = \_\_\_\_\_\_\_\_\_\_, Median = \_\_\_\_\_\_\_\_\_\_, Mode = \_\_\_\_\_\_\_\_\_\_**

*Can we calculate the Mean or Median for categorical variables? How do we describe their central tendency?*

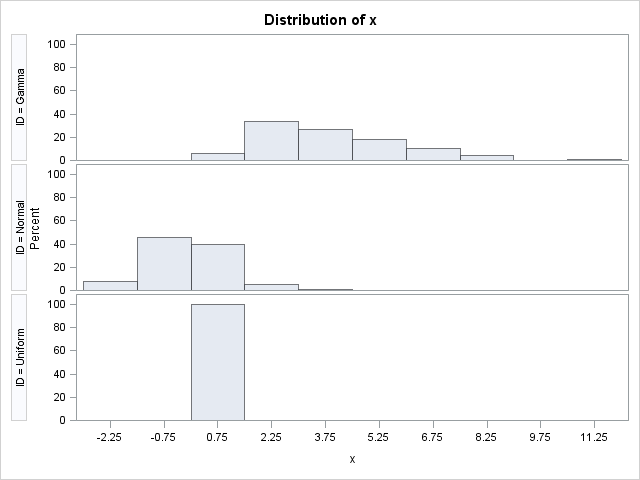
****

**Mean = \_\_\_\_\_\_\_\_\_\_, Median = \_\_\_\_\_\_\_\_\_\_, Mode = \_\_\_\_\_\_\_\_\_\_**

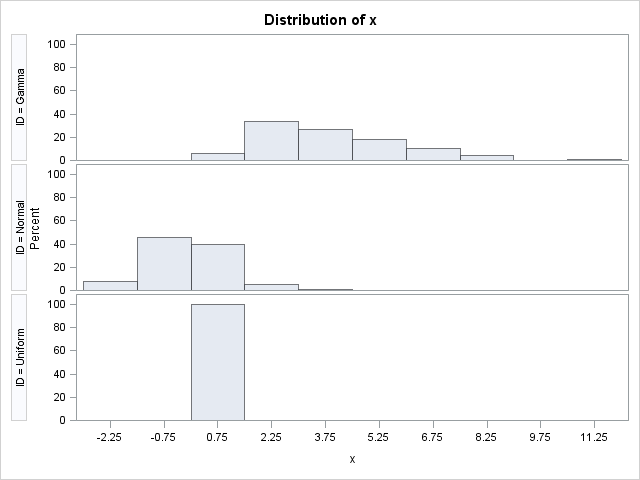
**Quantitative Variables and SPREAD**

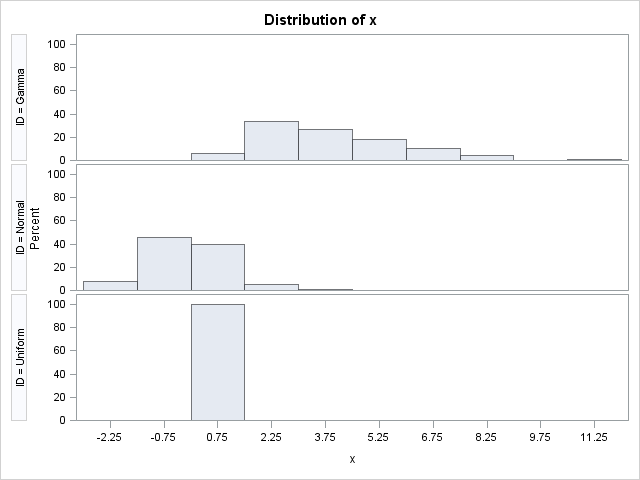
Describe each distribution in terms of:

*High Variation, Low Variation, or No Variation*

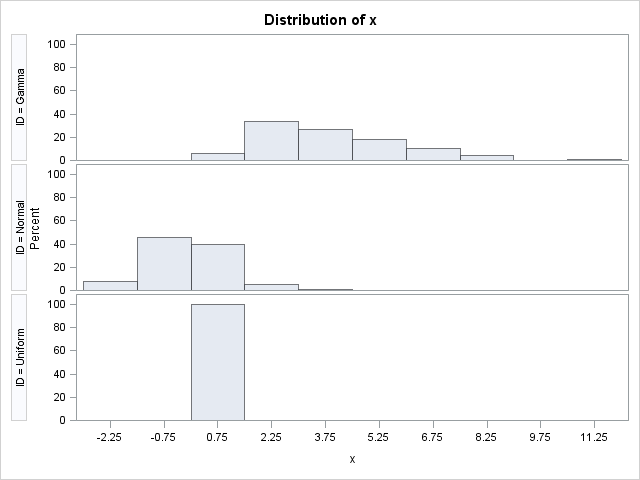
****

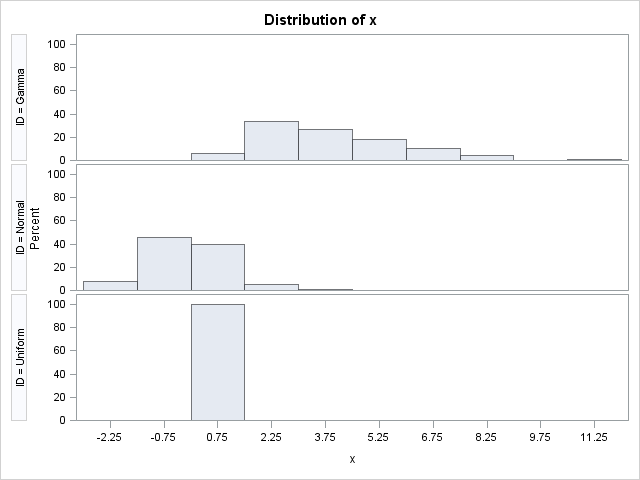
**\_\_\_\_\_\_\_\_\_\_\_\_\_ Variation**

****

****

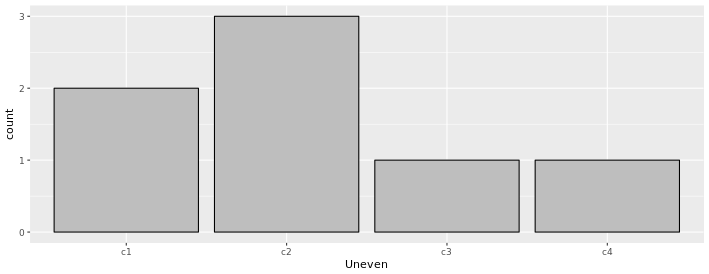
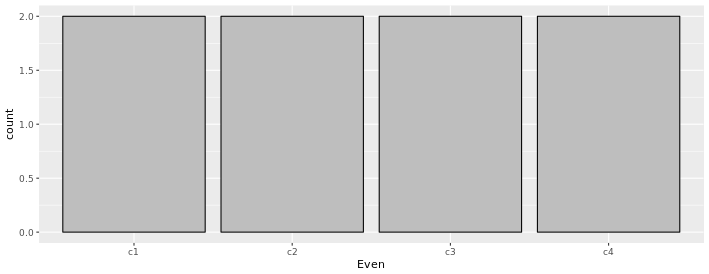
**\_\_\_\_\_\_\_\_\_\_\_\_\_ Variation**

****

****

**\_\_\_\_\_\_\_\_\_\_\_\_\_ Variation**

*Which of these categorical variables displays evenly distributed spread?*

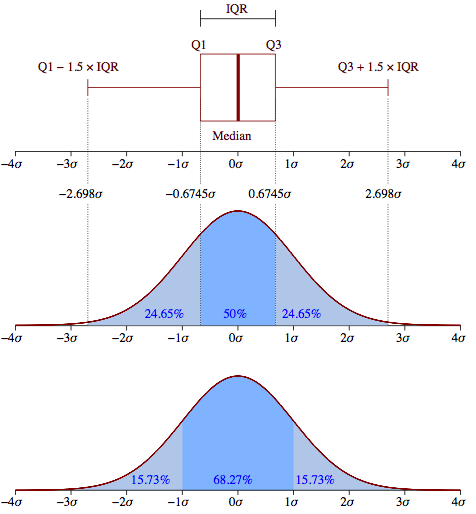
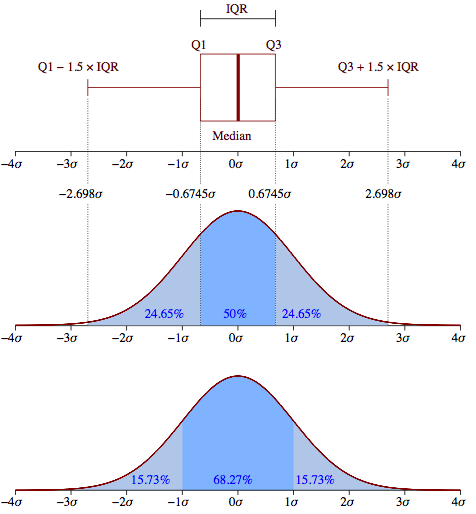
* *

**Quantitative Variables and OUTLIERS**

Describe the distribution in terms of:

*Values below Q1 - (1.5 \* IQR)*

*Values above Q3 + (1.5 \* IQR)*

**

***Examine this sample distribution (hint: first put the values in order):*** *50, 40, 30, 50, 100, 55, 10, 40, 45, 35*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **5 Number Summary** | | | | |
| **Q0 (Minimum)** | **Q1** | **Q2 (Median)** | **Q3** | **Q4 (Maximum)** |
|  |  |  |  |  |

**Range** (max - min) = \_\_\_\_\_\_\_\_\_\_\_\_ **Interquartile Range** (Q3 - Q1) = \_\_\_\_\_\_\_\_\_\_\_\_

**Lower Boundary** =Q1 - (1.5 \* IQR) = \_\_\_\_\_\_\_\_\_\_ **Upper Boundary** =Q3 + (1.5 \* IQR) *=* \_\_\_\_\_\_\_\_\_\_

Which values are outliers? (What function in R could also show us any outliers?)

Can you draw the boxplot?

**Putting it All Together**

**Summary Statistics vs. Data Visualizations**

Data often comprises many, many numbers. Imagine a study with 10,000 participants. It’s hard to keep in mind the results of each participant. Summary measures (such as the mean) are easier to keep in mind. Data visualizations (such as histograms) are easier to explore and think about.

Sometimes people get hung up on the summary measures; these summaries can hide telling patterns. Sometimes people think: all we have to know is the mean and that will be enough information about the *distribution.*

Let’s say you are trying to choose between two cancer drugs. Which of these is a better drug? (Assume the number means additional months lived.) Put them into vectors and generate the **favstats**() for each drug.

**Drug A**: 23, 21, 5, 9, 18, 24, 6, 8, 8, 25, 23, 12 (the mean is 15.17)

**Drug B**: 17, 16, 16, 13, 15, 22, 17, 13, 15, 15, 9, 15 (the mean is 15.25)

If we just looked at the means, we would think that drug B might be slightly better (the mean is .08 of a month more which turns out to be about 2.5 days). Let’s also use the **sort**() function to take a closer look.

1. Why is drug A better than drug B?
2. Why is drug B better than drug A?
3. Imagine someone looked at the means and summarized the two drugs like this: “Drug B basically helps you live 2.5 days longer than drug A.” Is this an accurate statement? Why not?

**Exploring Variation: Histograms vs. Boxplots vs. Bar charts**

1. Which type(s) of data visualization will best help us visually examine the distribution of the number of extra months survived for each drug?
2. Open the csv file of the data. How can we ask R to display a histogram of months survived to us?
   1. How can we modify the function in R to create a relative frequency histogram?
   2. How can we ask R to create separate histograms for each drug?
   3. How can we modify the function to also include a smooth density plot?
   4. Sketch the results and describe each histogram in terms of its shape, center, and spread.
3. How can we ask R to display boxplots of survival months for each drug?
   1. Sketch the results and use the boxplots to indicate the 5 number summary for each drug.
4. Which of the above visualizations (the histograms or the boxplots) would be more helpful to a person who is trying to figure out which drug to take? Why?
5. Now remember that the mean for drug A is 15.17 and 15.25 for drug B. On each of the figures above, put a mark where the means should go. Where do the means go on a boxplot? Where do they go on the histogram? How are they different?
6. Let’s say that the pharmaceutical company tells people, “Well, on average, people live about 15 more months with this drug.”
7. Is that an accurate statement for each drug?
8. Which drug is better described by the mean? Why?
9. Without knowing anything else, which seems like the better drug to take (assuming you want to survive a longer period of time)? Why?
10. How can we explore the variation in occupation for this sample? What type(s) of data visualization would be most appropriate?
    1. How can we ask R to display this visualization?
    2. How can we ask R to display a count for each category?
    3. How can we modify that code to display the count as a proportion?
    4. What can we add to the code to also include the total count/proportion?
    5. Sketch the results and describe the mode and spread of the distribution.
11. Compare the distributions of survival months for each drug to the survival months for each type of occupation (with histograms). Which variable appears to better explain the variation in survival months: treatment or occupation (consider the shape/center/spread of each distribution)?

**Take Home Messages:**

* We can summarize distributions in terms of their shape, center, spread, and outliers.
* Knowing one summary statistic (such as the mean) of a distribution is not always enough to draw conclusions; you should also examine the data visualization (e.g., the histogram or bar chart) in order to determine how spread out the data is.
* Visualizations help us see the big picture, while numeric summaries give us finer detail.
* The mean may be less reliable as a summary value of a distribution if the distribution has a lot of spread (variation) or if the shape is skewed.
* Bar charts represent the frequency for categorical variables, while histograms and boxplots represent the frequency of a single continuous variable.
* Pay attention to the level of measurement (Quantitative or Categorical).