Note: this week's material is on Chapter 3: Displaying and Summarizing Quantitative Data

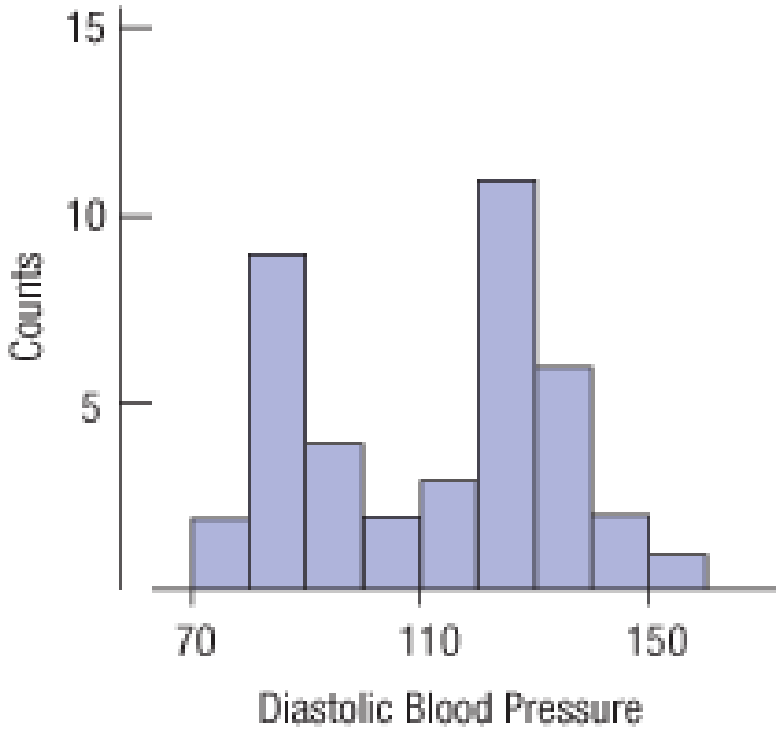
Histogram

- A histogram is used to display one quantitative variable.
- The bins, together with these counts, gives the distraction of this variable.
- The differences between a histogram and a bar chart
    - A bar chart is used to display a category variable.
    - The gaps in a bar chart is used to separate categories; in a histogram, the gap means there is a bin with no cases.

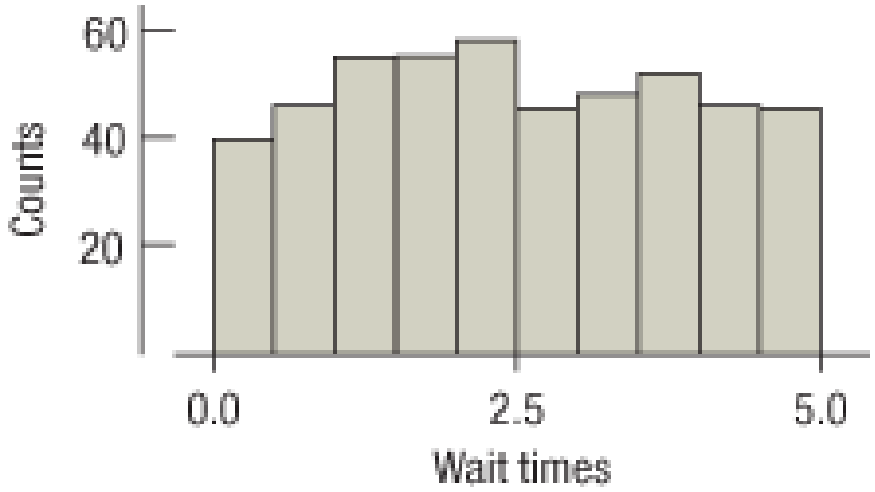
Describing Distribution

- When you describe a distribution, you should always discuss its shape, center, and spread
- Shape(mode, symmetry, and outlier)
    - Does the histogram have a single, central hump or several separated humps?
    - These humps are called **modes**.
    - A histogram with one peak is called **unimodal**; histograms with two peaks are **bimodal**; and those with three or more are called **multimodal.**

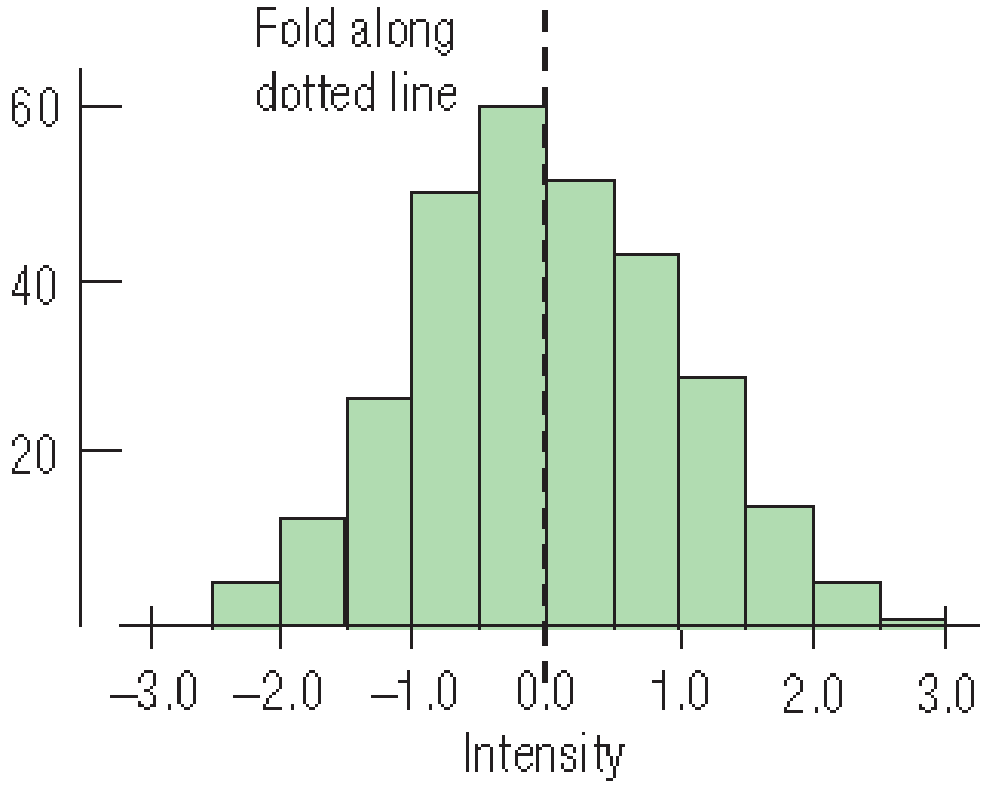
      E.g. A bimodal histogram with two peaks



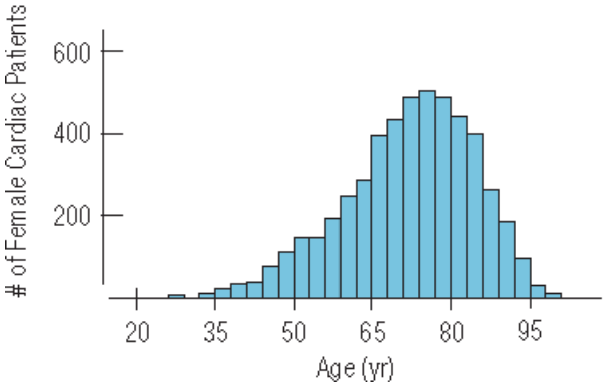
E.g. In this histogram, the bars are all about the same height. The histogram doesn't appear to have a mode and is called uniform

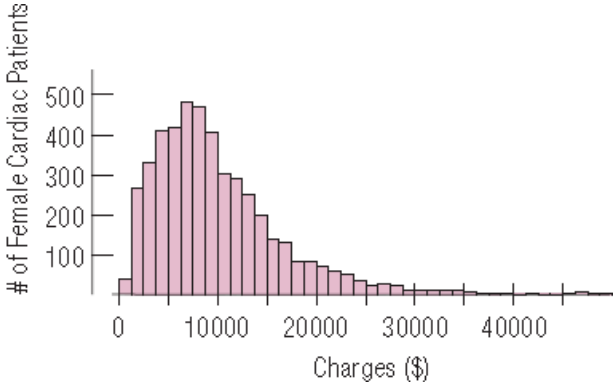


E.g. a symmetric histogram can fold in the middle so that the two sides almost match

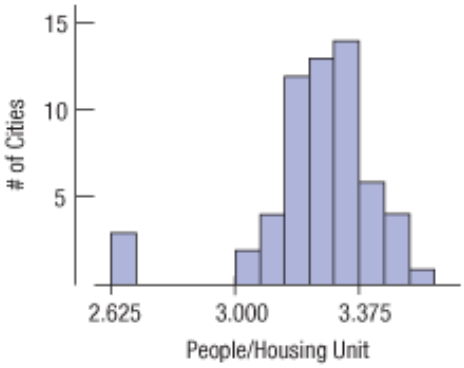


E.g. Left-skewed (longer tail to the left) and right-skewed (longer tail to the right)





E.g. Outliers



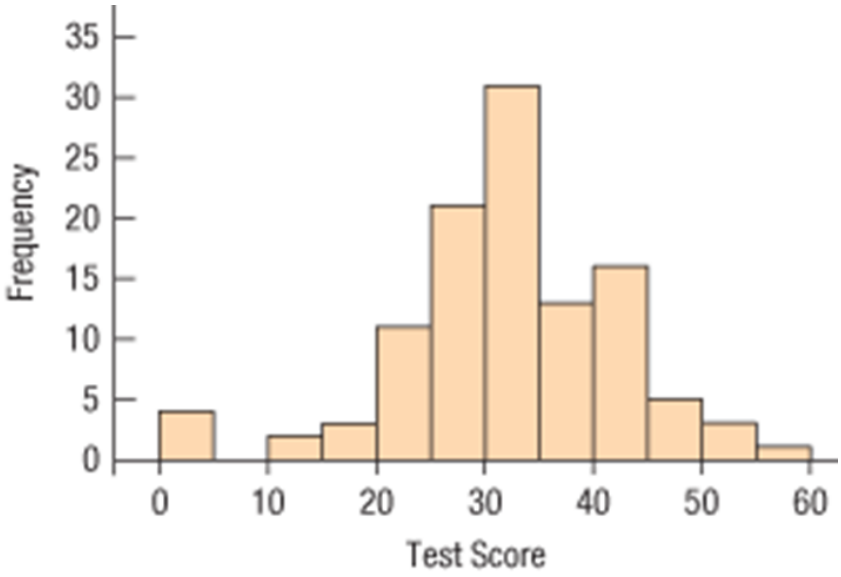
- Center(one number that represents the distribution)
  - Medium, mode, and mean
    Median: the middle value in the dataset that divides the dataset into two equal parts
    Mean: average
  - In a unimodal symmetric distribution with no outliers, generally mode = medium = mean
  - How to calculate median?
    Odd case: $\frac{n+1}{2}$th position
    Even case: the average of the numbers at the $\frac{n}{2}$th and the $\frac{n}{2}+1$ th position
- Spread(how the data varies around the center)
  - Range: difference between the max and min number
  - IQR: upper quartile – lower quartile
    The lower and upper quartiles are also known as the 25th and 75th percentiles of the data, respectively, since 25% of the data falls below the lower quartile and 75% of the data falls below the upper quartile.

Boxplot and 5-number Summary

- The 5-number summary of a distribution reports its median, quartiles, and extremes (maximum and minimum)
- The range of the data is defined as the difference between the maximum and minimum values
- Once we have a 5-number summary of a (quantitative) variable, we can display that information in a boxplot

Question 1: Chapter 3 Question 23

TEST SCORES Below is a histogram of 110 students' test scores on a one-hour test in STA220 at the University of Toronto during summer 2007. The test was out of 50, not 100.



a) Approximately what percentage of students got A grades? An A was 40 or higher, since the test was out of 50 (some scored above 50 due to bonus marks).

b) What percent got C or D grades? (i.e., between 50% and 69%, or 25 and 35 in the graph since any scores of 35 went into the higher bin)

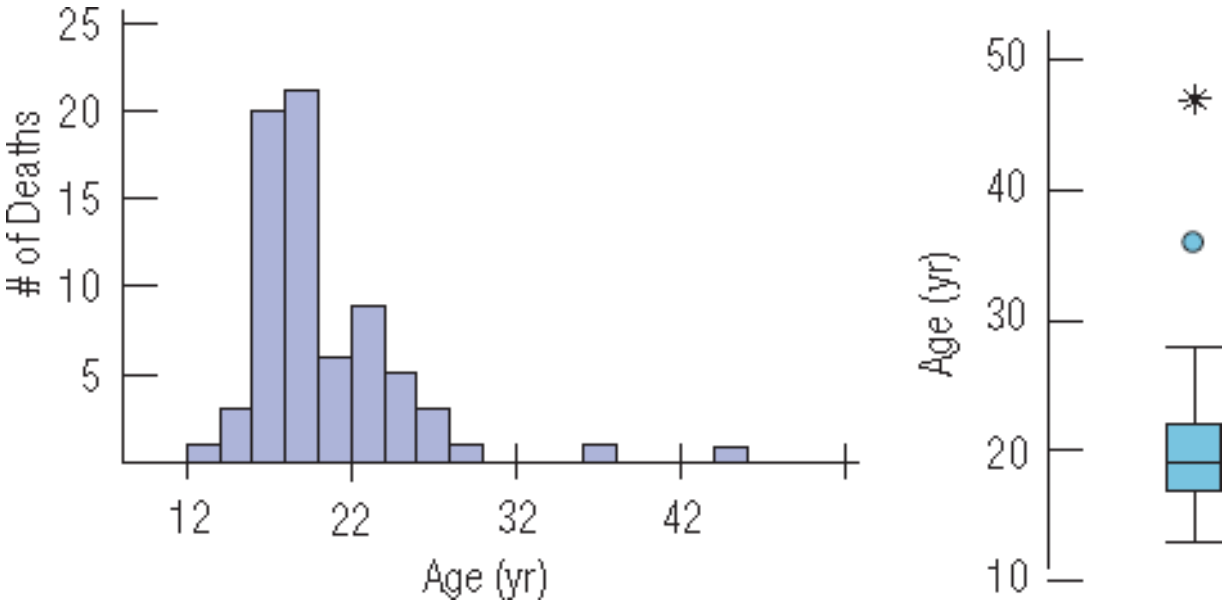
c) Write a brief description of this distribution (shape, centre, spread, and unusual features). Can you account for any of the features you see here?

Solution:

a)22.7% b) 47.3% c) The centre (the median) is around 32. The scores range from 0–60, but the few scores close to zero may be outliers (with a gap of just one bin we might not be able to conclude they are clear outliers, but they are somewhat unusual compared to the rest of the scores).

Question 2: Chapter 3 Question 71

ROCK CONCERT ACCIDENTS Crowd Management Strategies (www.crowdsafe.com) monitors accidents at rock concerts. In their database, they list the names and other variables of victims whose deaths were attributed to “crowd crush” at rock concerts. Here are the histogram and boxplot of the victims’ ages for data from a recent one-year period.



a) What features of the distribution can you see in both the histogram and the boxplot?
b) What features of the distribution can you see in the histogram that you could not see in the boxplot?
c) What summary statistic would you choose to summarize the centre of this distribution? Why?
d) What summary statistic would you choose to summarize the spread of this distribution? Why?

Solution:

a) The histogram and boxplot of the distribution of ages both show that a typical crowd crush victim was approximately 18–20 years of age, that the range of ages is 36 years, and that there are two outliers, one at age 36–38 and another at age 46–48.

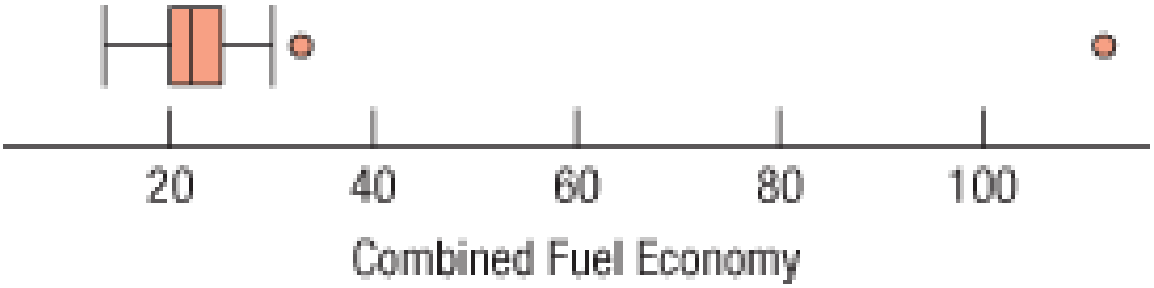
b) This histogram shows that there may have been two modes in the distribution of ages, one at 18–20 years of age and another at 22–24 years of age, but this is not shown in the boxplot.

c) Median is the better measure of centre, since the distribution of ages has outliers.

d) IQR is a better measure of spread, since the distribution of ages has outliers.

Question 3:

FUEL ECONOMY The boxplot shows the fuel economy (miles per gallon) ratings for 67 model year 2012 subcompact cars. Some summary statistics are also provided. The extreme outlier is the Mitsubishi i-MiEV, an electric car whose electricity usage is *equivalent* to 112 mpg.



Mean	sd	Min	Q1	Med	Q3	Max	n
23.76	11.87	14	20	22	25	112	67

a) If that electric car is removed from the data set, how will the standard deviation be affected? The IQR? (which is better to describe the sampel? Mean or Median)