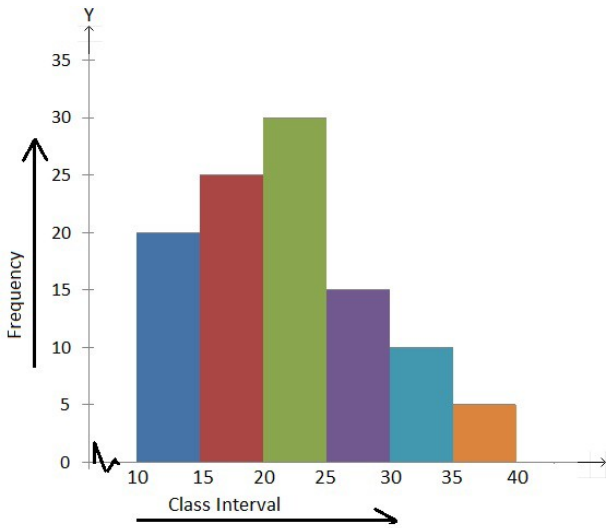
Displaying distribution with graphs

1 Visualizing Data

Four most common questions statisticians may ask themselves about their data are: **What is the shape of my data-set? Do I notice anything unusual? What is the center of my data-set? What is the spread of my data-set?** These questions can be answered when we study the distribution of our sample of data, and one of the most common ways to do it is to create a histogram.

2 Histograms

Histograms display your data by putting observations into intervals called bins of equal width. The height of the bin is the frequency. Frequency is the number of times a value was recorded when data was collected. The x axis be your independent numerical variable. And the y axis will have the various frequencies at which those variables occur.



When examining a distribution, we pay attention to its features so that we can properly describe it. We do so by analyzing the **shape, center** and **spread** of the distribution.

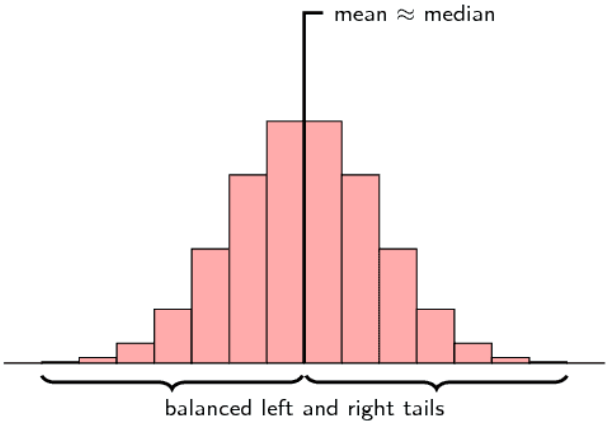
3 Shape

Let's first begin with the shape. When you want to analyze the **shape** of a distribution, you want to ask yourself the following questions:

1. Is the distribution symmetric or skewed?

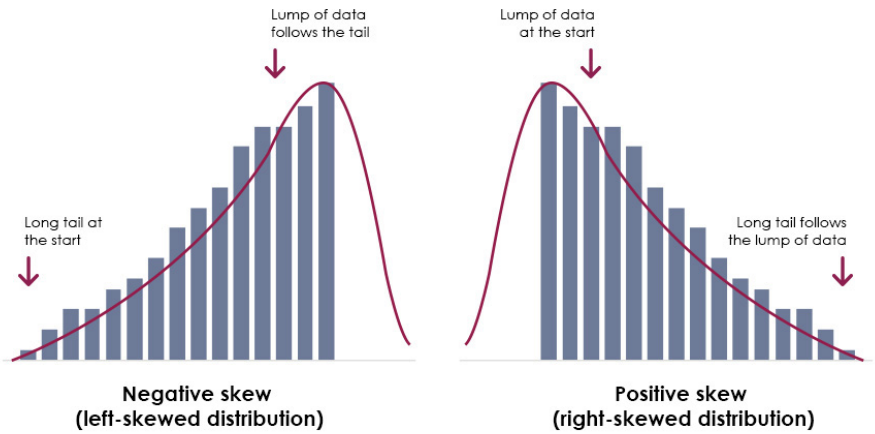
2. How many major mounds appear? None? One? Two? Several?

To answer these questions we are going to see a lot of examples. Let's first take a look at a **symmetrical histogram**. An important feature about the symmetrical histogram below is that the typical value falls around the center and the data is almost symmetrically distributed on both sides, hence the name.

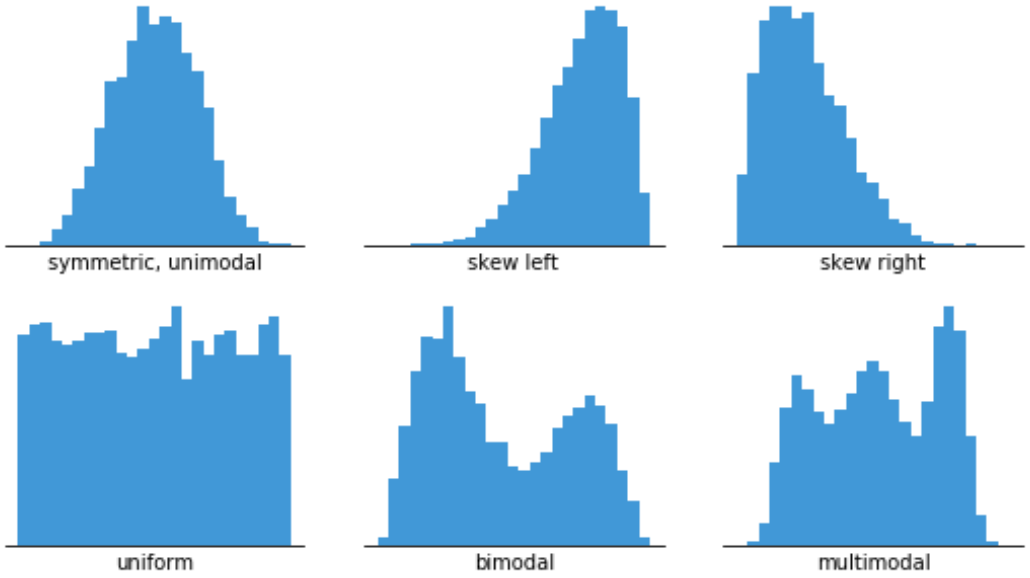


Now let's observed a skewed histogram. There are two types of skewed histograms, **right skewed** and **left skewed**.

- You will see in a **left skewed** distribution, lots of data is towards the right side of the histogram, which leaves the tail end to the left.
- In a **right skewed** distribution, lots of data is towards the left side of the histogram, which leaves the tail end to the right.



Now to answer question 2, you can also describe a histogram by how many peaks it has. A distribution may be uniform, have one major peak, two, or several. Those are shown below as well as the other shapes summarized.



Examples: What shape would you expect to see in a histogram of the following data sets?

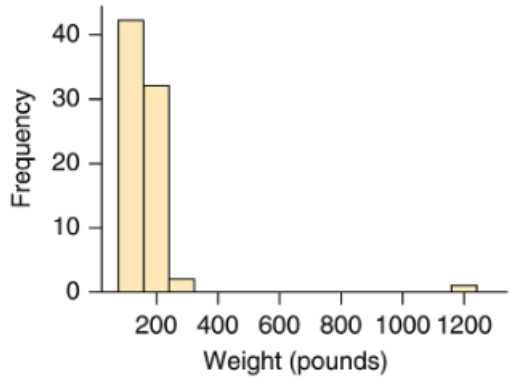
- GPA of college students - *Left Skewed*
- SAT scores - *Symmetric*
- Last digit of Social Security numbers for a random sample of students - - *Uniform*
- Income of USA residents - *Right Skewed*

What about modality in these data-sets?

- Morning and evening sales at a restaurant - *Bimodal*
- Men and women's heights - *Bimodal*

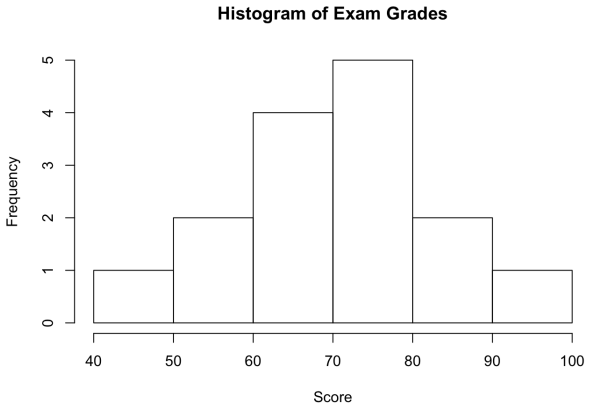
4 Do I notice anything unusual?

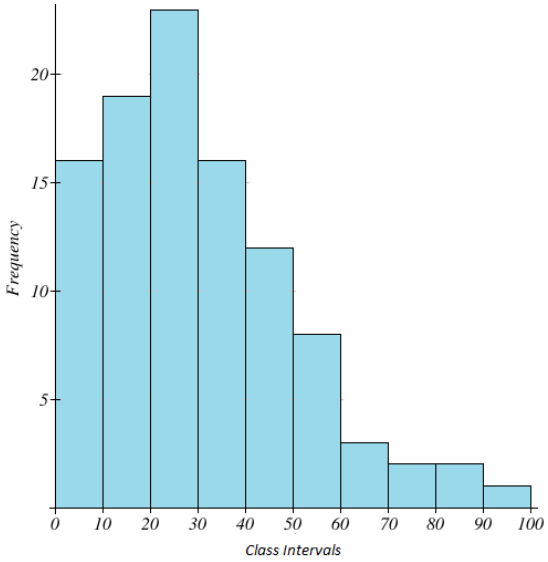
Some extremely large or small data values that don't fit the pattern of the rest of the data are called outliers. Outliers can be errors (such as typos) or be genuine. Genuine outliers are unusually interesting data values!
The histogram below shows one outlier on the right.



5 Center

The **center** of a distribution is like the typical amount. It is usually where you find most of your data. Based on the histograms below, what is the **shape** and what is the **center** or typical value?

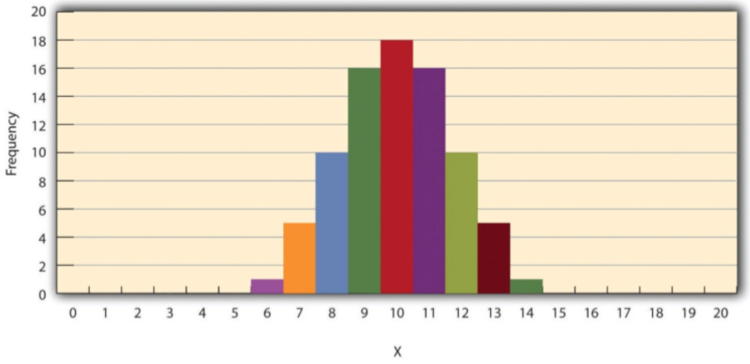


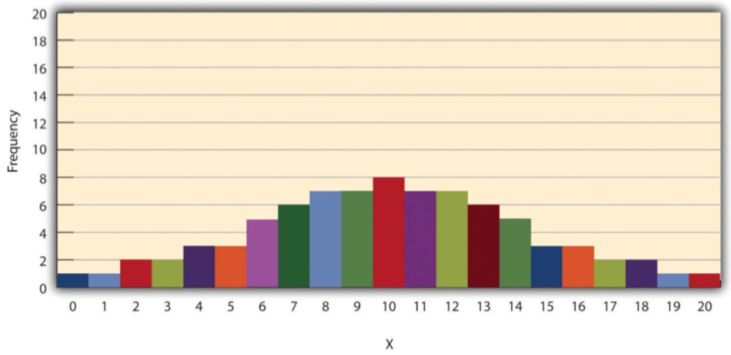


The histogram on the left has it's center between 70-75 and the histogram on the right has center somewhere between 28-33.

6 Variability/Spread

We describe the amount of variation in our data by looking at how much **spread** there is along the x axis. Based on the histograms below, which has more variability? Why?



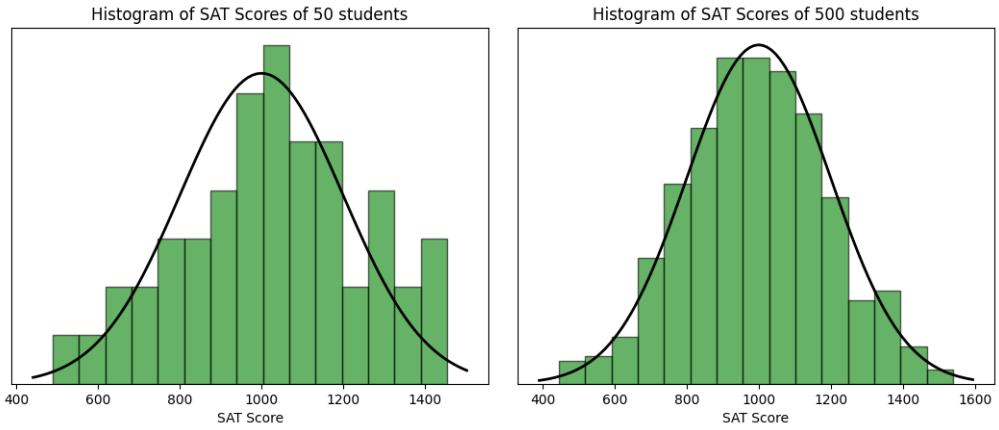


Since the histogram on the bottom has more horizontal spread, it has more spread/variability in the data.

7 The Normal Distribution

There are many distributions that we observe in real world datasets; however, the Normal distribution is the most frequently seen distribution for continuous numerical variables. Hence it is of particular interest to statisticians as many numerical variables of interest to researchers tend to have distributions that closely match the Normal model.

We always start with unimodal and symmetric dataset to begin with. The following histogram displays the SAT scores for 50 students and 500 students. We have a computer-drawn curve to fit the histogram data. The overlaying of smooth curve on the histograms is to illustrate the unimodal and symmetric shape to you. It's easy to imagine that as more data is collected, the histogram would fill in the curve more precisely, eventually matching the shape almost perfectly. It is also evident that the histogram of SAT score with 500 students fits more closely to the curve than the histogram with 50 students.



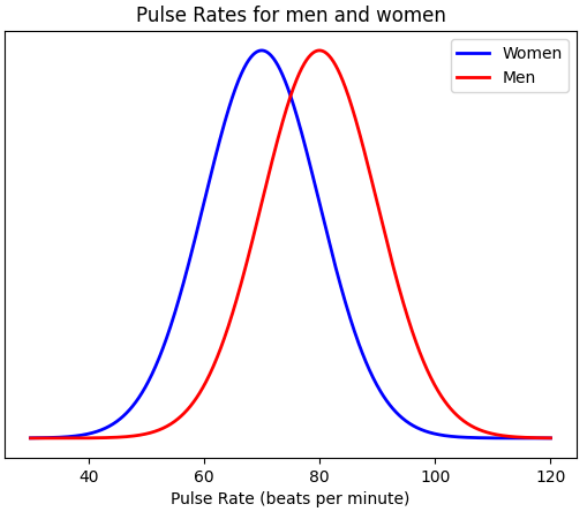
Examples: Height and weight of people, reaction of time of people, length of natural sleep, IQ scores, SAT scores, pulse rate are some examples of normally distributed datasets.

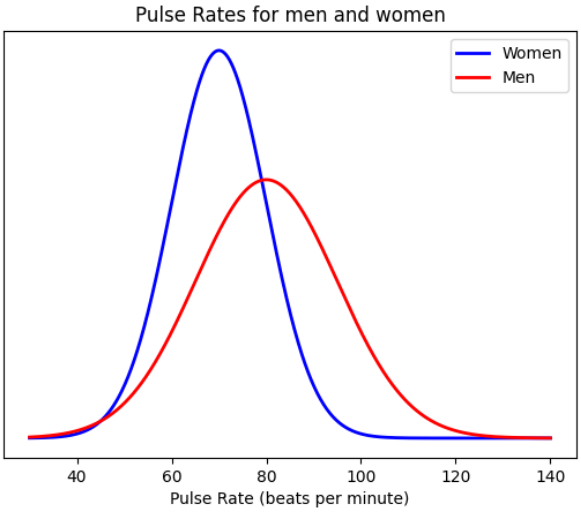
8 Understanding Normal Distribution

Colloquially, Normal distribution is also called as normal curve or bell curve. The main characteristic of a normal distribution is that it describes a symmetrical plot of data around its typical value (center), where the width of the curve indicates how spread out the data set is.

Let's visualize these with the help of examples. We have the dataset for resting pulse rate for men and women in beats per minute. The blue curve is the normally distributed pulse rate for women and it's center is at 70. Hence, the typical value of resting pulse rate in women is 70 beats per minute. Similarly, the red curve is the normally distributed pulse rate for men and it's center is at 80. Hence, the typical value of resting pulse rate in women is 80 beats per minute.

Similarly in the following figure, the spread for women is 10 beats per minute. Hence, the blue curve is narrow in width. However, the spread for men is 25 beats per minute. Thus, we have a wider curve.

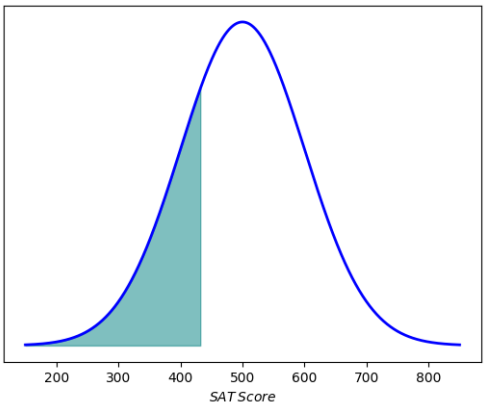




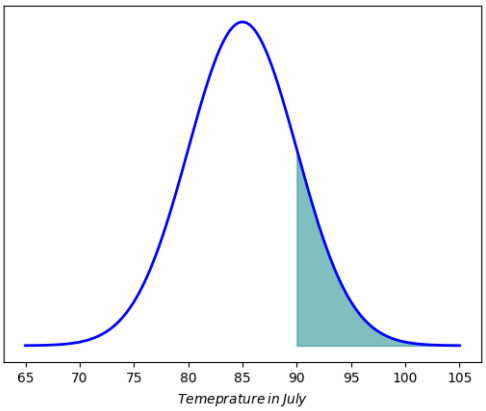
9 Percentiles

You frequently hear the term "percentiles" used in everyday life. In most of the cases, percentiles come from the Normal distribution. Percentile boils down to the areas under a normal curve and in between given intervals. Let's see some examples.

Example 1: Assume that the SAT math exam scores are normally distributed. If Harry obtains a score of 432 on the SAT math exam, and say 24.83% of students scored less than or equal to Harry, we say hat Harry's score is at 25th percentile. 25th percentile is highlighted in the normal curve below.

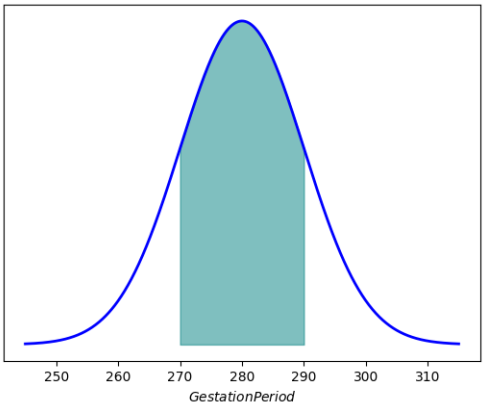


Example 2: In a certain city, the average daily temperature in July follows a normal distribution with a center of 85°F and a spread of 5°F. Let's say that a certain day in July had a temperature of 90°F and the temperature was in the 84th percentile. Let's see what this means.



This mean that 84% of temperature in July are below the 90°F and 16% of temperature in July are above the 90°F which is shown in the normal curve above.

Example 3: Suppose that gestation period for human pregnancies are found to follow a Normal distribution, with a center of 280 days and a spread of 10 days. 16th percentile gestation period is of 270 days and 84th percentile gestation period is of 290 days. Let's visualize what percentile of pregnancies are between 270 and 290 days.



The shaded region shows the middle 68% of pregnancies, corresponding to the 16th to 84th percentile range.