**Introduction**

In this module, we present an introduction to the science of statistics which will provide a valuable foundation for which the rest of the modules will build.

Statistics is a science that deals with the collection, classification, analysis, and interpretation of information or data. It is not just the tabulation of data, statistics deals with data collection, evaluation and interpretation. Statisticians use data to find patterns, answer important scientific questions, and draw conclusions. Statistics can be used to solve issues across many different applications including business, government, physical, social, and medical sciences.

There are two main areas of statistics:

1. Describing data (including numerical and graphical summaries)
2. Drawing conclusions about data (making estimates, predictions, and decisions) from data collected via sampling

The first part of this module focuses on mostly on the first area: describing data. We will learn how to utilize numerical summaries and graphical methods to look for patterns and to summarize information in a convenient and meaningful way.

The latter half of the module, we begin to introduce the concepts underlying inferential statistics including the normal distribution, sampling, an introduction to inference, and the Central Limit Theorem.

**Fundamental Elements of Statistics**

Statistical methods are used for studying, analyzing, and learning about populations of experimental units.

An experimental unit (or observational unit) is an object (for example, a person, thing, or event) about which we collect data about.

A population is a set of units (for example, people, objects, or events) that we are interested in studying.

For example, we may be interested in the population of all people that have diabetes, or the population of unemployed workers in the United States. Or, we may be interested in the population of red trucks coming off a manufacturing line, or all the orders placed at a particular restaurant.

In studying a population, we focus on one or more characteristics or properties of the units of the population. We call these characteristics variables. For example, for the population of unemployed workers, we may be interested in their age, gender, and/or number of weeks unemployed. The value of the variable differs from one experimental unit to another.

Variables can be classified into one of two general types:

1. Qualitative
2. Quantitative

Qualitative variables are those that place experimental units into categories. Qualitative variables are sometimes referred to as categorical variables. Gender, race, smoking status, and eye color are examples of qualitative variables.

Quantitative variables, on the other hand, are those that contain numeric data. Quantitative variables are also sometimes referred to as continuous variables. For example, age, blood pressure, and heart rate are quantitative variables that may be measured on people. If our population of interest were transactions at a restaurant, the cost of the order would be a quantitative variable that we’d likely be interested in recording. We can take arithmetic operations (such as the sum or the average) on quantitative variables.

Sometimes continuous variables are collected or summarized in such a way that they become categorical. For example, if instead of asking individuals how old they were, we collected information about which age category each person belonged to (<18<18, 1818–2929, 3030–3939, etc.). In general, when collecting information or measuring a particular variable, it is generally best to collect it as precisely as possible. If you have each subject’s age, you can create the age categories using the precise ages in any way that you want. If you only collect information about age category and then later wish to modify the categories slightly (from <18<18, 1818–2929, 3030–3939, etc. to ≤30≤30, 3131–4040, 4141–5050, etc.), then you will be unable to do so.

**Data Summaries**

Regardless of the type of variable, one of the first steps in data analysis is to summarize data. Even when the interest is to look at the relationship between two or more variables (as we will look at in modules later in the course), it is important to first examine and characterize each variable by itself.

The methods for summarizing data depend on the type of variable that you have. Methods for summarizing both qualitative and quantitative variables are described below. Numerical summaries concisely summarize data into meaningful quantities. Graphical summaries help take that information and display it so that it is easier to understand your data. Generally, graphical summaries are to assist you in seeing the overall pattern of the data as well as to look for any striking deviations from that pattern.

**Qualitative Data Summaries**

Numerically, we can summarize qualitative data in two ways: (1) by computing the class frequency or (2) by computing the class relative frequency. The class frequency is the number of observations in the dataset that fall into a particular class. The class relative frequency is the proportion of the number of observations in the dataset that fall into a particular class to the total number of observations in the dataset.

Graphically, we can display qualitative data using pie charts and bar graphs. Bar graphs show the frequency or the relative frequency by the height of each rectangle in the graph. Pie charts visually show the relative frequency of each class as “pie slices” where the size (or angle) of the slice is proportional to the class relative frequency. Graphical displays of categorical data are often limited in utility since the numerical summaries are often sufficient for understanding the distribution of a single variable. Categories may be shown in any order in either of these displays. If there is a natural ordering of categories (as in the example below), this may be preferred. Otherwise, alphabetical or from highest to lowest relative frequency can be used to order the categories.

Data from all possible categories should be shown (and as such, the sum of all relative frequencies should equal 100% and the sum of the frequencies should equal the total number of experimental units measured). Sometimes this means you may have an “Other” (if there are a lot of small categories that you’d prefer not to display individually) or “Unknown” class. Graphical displays can be modified to group similar categories together.

For any graphical summary, it is important to ensure that it is labelled appropriately. Vertical and Horizontal axes should always be labelled. Depending on the data and the “customer” of the output, you may also want to ensure that data labels are on the graph (for example, to show the exact frequency or relative frequency for each class).

**Example 1.1**

Suppose we are interested in understanding the education level of the most well paid chief executive officers (CEOs). We have data about the highest level of education received by the top 40 most well paid CEOs as published in a recent article in *Forbes* magazine. For this study, the variable of interest is the highest level of education which is a qualitative variable. Highest level of education is not numeric in nature and must be classified into categories or classes. The classes collected for highest education level in this case include high school, college, masters, other advanced degree (e.g., Law or PhD).  The data for the 40 most well paid CEOs is shown in the table below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| High School | Other | College | Other | Master’s |
| Master’s | Master’s | High School | Master’s | College |
| College | Master’s | Master’s | High School | College |
| Master’s | College | High School | Other | College |
| Master’s | Other | Other | Master’s | Master’s |
| Master’s | Other | College | College | College |
| High School | College | Master’s | College | High School |
| Master’s | Master’s | College | College | College |

We can summarize this data as follows:

|  |  |  |
| --- | --- | --- |
| Class | Frequency | Proportion |
| High School | 66 | 6/40=0.15=15%6/40=0.15=15% |
| College | 1414 | 14/40=0.15=35%14/40=0.15=35% |
| Master’s | 1414 | 14/40=0.35=35%14/40=0.35=35% |
| Other | 66 | 6/40=0.15=15%6/40=0.15=15% |
| **Total** | 4040 |  |

The class frequencies (shown in the Frequency column) came from counting up the number of CEOs with each particular class. For example, the frequency for the class High School was determined by counting up the number of CEOs that reported High School as their highest level of education. The class relative frequencies (shown in the Proportion column) came from dividing the respective class frequency by the total number of CEOs (in this case, 40). From the table above, it appears that 85%85% of the highest paid CEOs have more than a high school degree. About half have received formal education after attaining their bachelor’s degree. It may be surprising that only 15%15% of the highest paid CEOs have a higher degree than a master’s degree.

Graphically, these data can be represented in a pie chart or a bar graph (either showing the class frequency or the class relative proportion):

Pie chart showing the highest education level of the top 40 best paid CEOs

Bar graph showing the highest education level of the top 40 best paid CEOs

Bar graph showing the highest education level of the top 40 best paid CEOs

**Test Yourself (Multipart) 1.1**

The average number of babies born on each day of the week in the last year for the US is shown below:

|  |  |
| --- | --- |
| Day | Frequency |
| Sunday | 77317731 |
| Monday | 1101811018 |
| Tuesday | 1242412424 |
| Wednesday | 1218312183 |
| Thursday | 1189311893 |
| Friday | 1201212012 |
| Saturday | 86548654 |

Calculate the class relative frequencies and the best way to display this information.

What are the class relative frequencies?

Since the total number of births per week is 75,91575,915, we can find the relative frequency by taking each relative frequency and dividing by 75,91575,915:

|  |  |  |
| --- | --- | --- |
| Day | Frequency | Relative Frequency |
| Sunday | 77317731 | 10.2%10.2% |
| Monday | 1101811018 | 14.5%14.5% |
| Tuesday | 1242412424 | 16.4%16.4% |
| Wednesday | 1218312183 | 16.0%16.0% |
| Thursday | 1189311893 | 15.7%15.7% |
| Friday | 1201212012 | 15.8%15.8% |
| Saturday | 86548654 | 11.4%11.4% |

For example, the relative frequency for Sunday is 7731/75,915=0.1027731/75,915=0.102 or 10.2%10.2%.

What is the best way to display this information?

Given the large number of classes for this variable and the similarity between relative frequencies on the majority of days, this data is best represented as a bar chart as below. As shown in the graph, week days (Monday–Friday) seem to have a similar number of births. The weekend days, however, have a much lower percentage of births.

**Quantitative Data Summaries**

Since quantitative data by definition are data that are recorded on a numerical scale, there are more methods available for both numerical and graphical summaries of this type of data.

**Graphical Summaries of Quantitative Data**

Graphical summaries for quantitative data can be used to describe the shape, center, and spread of the data. Though numerical summaries may be the most convenient way to summarize the center of the distribution (as will be discussed in the next section), we can still get a fairly good idea of the center (by looking at the approximate midpoint where roughly half of the observations are above and about half of the observations are below) and the spread (by looking at the smallest and largest values) using graphical summaries. Outliers (data points that do not fit with the general pattern of the data) can also be detected using graphical summaries to look for outliers.

The most popularly used graphical summaries for quantitative data include

1. Histograms
2. Box Plots

Other summaries (not detailed in this course) include dot plots and stem and leaf plots.

Histograms are perhaps the most popular graphical summary of quantitative variables. Data are first categorized into classes (categories) of equal width and then frequencies and relative frequencies are calculated. Some judgment is needed to select how many classes should be displayed (which ultimately influences the width of each class). Choosing too many classes will produce a “flat” graph. Choosing too few classes may result in a “skyscraper” graph. Generally, histograms look similar to bar graphs, but the bars are positioned side-by-side so that there are no spaces between. The classes of a histogram should cover the entire range of values of a variable.

The shape of a distribution can be described by looking at the histogram. A distribution is described as symmetric if the right and left sides of the histogram are approximately mirror images of each other. A distribution is skewed to the right if the right side of the histogram extends much further out than the left side. It is skewed to the left if the left side of the histogram extends much farther to the left than to the right side. See an example of a symmetric distribution as well as examples of distributions skewed to the right or to the left below.

**Example 1.2**

We are interested in summarizing the age of employees at company X. Below are the ages of each of the employees at the company. Create a histogram to show the distribution of ages and interpret it.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 4242 | 2020 | 3232 | 4747 | 3131 |
| 6666 | 2525 | 6464 | 2525 | 4646 |
| 7676 | 5656 | 3232 | 2020 | 5050 |
| 6060 | 5858 | 3131 | 8383 | 5151 |
| 2222 | 3232 | 6464 | 4949 | 7575 |
| 4040 | 4343 | 5454 | 4444 | 6262 |
| 4646 | 2727 | 3232 | 4949 | 3737 |
| 3838 | 5959 | 3333 | 5959 | 7373 |
| 2626 | 2626 | 8383 | 7171 | 3939 |
| 3535 | 3333 | 3535 | 2828 | 3535 |

First, we select the classes. For this example, we will pick the following classes: 2020–2929, 3030–3939, 4040–4949, 5050–5959, 6060–6969, 7070–7979, and 8080–8989. When choosing how many and the width of each class, you want to ensure that you have a reasonable amount of bars and that you define the classes precisely so that each experimental unit only falls into a single class (that is, each class definition needs to be non-overlapping). You also need to ensure that each class is the same width.

Next, we count the individuals in each class (the frequency). We can also calculate the relative frequency by dividing each frequency by 5050 (the number of employees). The counts and relative frequencies are summarized in the table below:

|  |  |  |
| --- | --- | --- |
| Class | Frequency | Relative Frequency |
| 2020–2929 | 99 | 18%18% |
| 3030–3939 | 1414 | 28%28% |
| 4040–4949 | 99 | 18%18% |
| 5050–5959 | 77 | 14%14% |
| 6060–6969 | 55 | 10%10% |
| 7070–7979 | 44 | 8%8% |
| 8080–8989 | 22 | 4%4% |

Finally, the histogram can be generated. The histogram for this data is shown in the figure below.

Note that the range of the data is represented here (2020 to 9090). Each axis is labeled so it is clear what is being represented. Each class is represented by a separate bar which has a height that is proportional to relative frequency.

We can use the histogram directly to describe the data. This histogram shows a single peak (representing ages between 3030 and 3939). The distribution is slightly skewed to the right. The midpoint is likely between 4040 and 4949 years of age. Ages generally range between 2020 and 8383, but less than 5%5% of employees are 8080 or older.

We will return to box plots after we review the numerical summaries of qualitative data.

**Numerical Summaries of Quantitative Data**

Numerical summaries of quantitative data generally focus on measures that describe the center and the spread.

The most common measures of the center are the mean and the median.

The mean is the ordinary arithmetic average. To find the mean of a set of observations, you simply add up their values and divide by the number of observations. If we have nn observations, x1,x2,…,xnx1,x2,…,xn, then the mean (commonly represented as x¯x¯), is calculated as x¯=x1+x2+⋯+xnn=1n∑ni=1xix¯=x1+x2+⋯+xnn=1n∑i=1nxi. (Note, the Greek letter sigma (∑∑) in the formula means to sum or to add up all of the values. The subscripts on the xx’s are just a way of keeping track of each of the nn observations. For example, x1x1 just means the value from the first observation (experimental unit), x2x2 just means the value from the second observation, ⋯⋯ and xnxn just means the value from the nnth (the last) observation. x¯x¯ is pronounced “xx-bar” and it designates the mean over all the observations.

**Example 1.3**

We are interested in summarizing the age of employees at company X. Below are the ages of each of the 50 employees at the company. What is the mean age?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 4242 | 2020 | 3232 | 4747 | 3131 |
| 6666 | 2525 | 6464 | 2525 | 4646 |
| 7676 | 5656 | 3232 | 2020 | 5050 |
| 6060 | 5858 | 3131 | 8383 | 5151 |
| 2222 | 3232 | 6464 | 4949 | 7575 |
| 4040 | 4343 | 5454 | 4444 | 6262 |
| 4646 | 2727 | 3232 | 4949 | 3737 |
| 3838 | 5959 | 3333 | 5959 | 7373 |
| 2626 | 2626 | 8383 | 7171 | 3939 |
| 3535 | 3333 | 3535 | 2828 | 3535 |

To calculate the mean, we must add up all of the ages and then divide by the total number of observations (5050 in this case).

x¯=x1+x2+⋯+xnn=42+66+76+60+22+⋯+73+39+3550=226450=45.28x¯=x1+x2+⋯+xnn=42+66+76+60+22+⋯+73+39+3550=226450=45.28

**Test Yourself 1.2**

The earnings (in thousands of dollars) for 99 college graduates are shown below. Find the mean.

|  |  |  |
| --- | --- | --- |
| 3535 | 4040 | 145145 |
| 3333 | 3030 | 4242 |
| 3232 | 3232 | 2525 |

Show Answer

The median is the midpoint of the distribution such that half of the data points are below it and the other half are larger than it. To find the median, you must order the values from smallest to largest. If the number of observations is odd, the median is the center observation in the ordered list. It is the n+12n+12th smallest (or largest) observation. If the number of observations is even, then the median is the mean of the two center observations. That is, it is the average of the n2n2th and the n2+1n2+1th smallest observations.

The formulas above do not give the median, but give the position of the median in the ordered list. Note that finding the median in a small dataset is relatively easy. However, finding the median when the dataset is large is tedious (as it involves ordering the data from smallest to largest which is more difficult to do when the number of observations gets large).

**Example 1.4**

We are interested in summarizing the age of employees at company X. Below are the ages of each of the 5050 employees at the company. What is the median age?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 4242 | 2020 | 3232 | 4747 | 3131 |
| 6666 | 2525 | 6464 | 2525 | 4646 |
| 7676 | 5656 | 3232 | 2020 | 5050 |
| 6060 | 5858 | 3131 | 8383 | 5151 |
| 2222 | 3232 | 6464 | 4949 | 7575 |
| 4040 | 4343 | 5454 | 4444 | 6262 |
| 4646 | 2727 | 3232 | 4949 | 3737 |
| 3838 | 5959 | 3333 | 5959 | 7373 |
| 2626 | 2626 | 8383 | 7171 | 3939 |
| 3535 | 3333 | 3535 | 2828 | 3535 |

To calculate the median, we must first order the values from smallest to largest:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | 11 | 2020 | | 22 | 2020 | | 33 | 2222 | | 44 | 2525 | | 55 | 2525 | | 66 | 2626 | | 77 | 2626 | | 88 | 2727 | | 99 | 2828 | | 1010 | 3131 | |  | |  |  | | --- | --- | | 1111 | 3131 | | 1212 | 3232 | | 1313 | 3232 | | 1414 | 3232 | | 1515 | 3232 | | 1616 | 3333 | | 1717 | 3333 | | 1818 | 3535 | | 1919 | 3535 | | 2020 | 3535 | |  | |  |  | | --- | --- | | 2121 | 3737 | | 2222 | 3838 | | 2323 | 3939 | | 2424 | 4040 | | 2525 | 4242 | | 2626 | 4343 | | 2727 | 4444 | | 2828 | 4646 | | 2929 | 4646 | | 3030 | 4747 | |  | |  |  | | --- | --- | | 3131 | 4949 | | 3232 | 4949 | | 3333 | 5050 | | 3434 | 5151 | | 3535 | 5454 | | 3636 | 5656 | | 3737 | 5858 | | 3838 | 5959 | | 3939 | 5959 | | 4040 | 6060 | |  | |  |  | | --- | --- | | 4141 | 6262 | | 4242 | 6464 | | 4343 | 6464 | | 4444 | 6666 | | 4545 | 7171 | | 4646 | 7373 | | 4747 | 7575 | | 4848 | 7676 | | 4949 | 8383 | | 5050 | 8383 | |

In the above, the left hand column is the order (11–5050) and the right column is the data.

Since we have an even number of observations (5050), the median is the average of the n2n2th and the n2+1n2+1th smallest observations. That is, the median is the average of the 2525th and the 2626th smallest observations. The 2525th smallest observation is 4242 and the 2626th smallest observation is 4343, so the median in this case is 42.542.5.

**Example 1.5**

Find the median of the following 55 numbers:

1116592211165922

First, we order the numbers:

5911162259111622

Since we have an odd number of observations, the median is the n+12n+12th smallest observation. Here, n=5n=5, so the median is the 33rd smallest observation. The third number in the list is 1111, so the median is 1111.

Another method that is sometimes used to find the median (which works the same as the method above) is to cross out values from the left and then the right continually until there is only one value left:

/591116/22⧸591116⧸22

Then, we repeat that process:

/5/911/16/22⧸5⧸911⧸16⧸22

If there were more than one value left, we’d continue to cross out values on the left and right. Since the only value left is 1111, this is the median.

If there were an even number of values as in the example below, we can still use the cross out method. We just will be left with two values in the middle (as we need to make sure that we cross out the same number of values on the left and right sides):

/5/131719/20/22⧸5⧸131719⧸20⧸22

To find the median in this case, we take the average of the two remaining values. Here the median is 1818. We would have gotten this same value using the ordered approach. Since there are 66 observations, the median is the average of the 33rd and 44th smallest values (1717 and 1919).

**Test Yourself 1.3**

The earnings (in thousands of dollars) for 99 college graduates are shown below. Find the median.

|  |  |  |
| --- | --- | --- |
| 3535 | 4040 | 145145 |
| 3333 | 3030 | 4242 |
| 3232 | 3232 | 2525 |

Show Answer

In the test yourself question above, the mean and median were not similar. The mean was 4646 and the median was 3333. You may have noticed that the value of 145145 in the data didn’t seem to fit with the other values. It was more than three times the next highest observation. If we calculated the mean and median excluding that value from the dataset, we’d get 33.62533.625 and 32.532.5 for the mean and median, respectively. By removing this one data point, the mean changes by more than 1010 units (thousands of dollars in this case), while the median remains virtually unchanged.

This example illustrates an important point about the differences between the mean and the median. A single high value (or a skewed distribution) will pull the mean towards the outlier (or towards the skew). The median on the other hand isn’t influenced by outliers. The value of 145145 is just an observation that is above the center in the calculation of the median. How far 145145 is from the center doesn’t matter. If the value had been 5050 or 200200, the median would have been the same (which cannot be said for the calculation of the mean).

If the distribution is symmetric, then the mean and the median will be relatively similar. If the distribution is skewed, the mean will be farther out toward the tail of the skew than the median. In cases where the distribution is skewed or there are outliers present, the median is usually the preferred measure of center (given the mean’s sensitivity to skewness). However, the best choice often depends on the context.

The mean and median are numerical measures of the center of a distribution. The quartiles and the variance (and standard deviation) are measures of the spread of a distribution. Both measures of the center and the spread are important pieces of information in trying to understand a set of data.

For example, consider summaries of the census data. A recent report indicated that the median annual income of American households was approximately $42,000$42,000. This means that half of the households in the US had annual incomes that were lower than $42,000$42,000 and half of the households had annual incomes that were higher. The mean annual income was nearly $60,000$60,000. The difference between the mean and the median here indicates that the distribution is skewed to the right and/or has outliers that are pulling the mean higher than the median. But, this may not tell the whole story. The bottom 20%20% of households had annual incomes that were less than $18,000$18,000 while the top 5%5% of households took in more than $150,000$150,000. This type of information comes from understanding the spread of the distribution along with the center. Generally, numerical summaries of quantitative data include both a measure of the center and the spread.

The simplest way to measure the spread is to give the minimum and the maximum data point. However, the minimum and the maximum are highly sensitive to outliers and may not give the best picture of the spread of the data.

The spread of the middle half of the data is often used instead of the minimum and maximum to describe the spread of the distribution. The middle half of the data is referred to as the quartiles of the data. The first quartile is the value in which 25%25% of the data lie below it and 75%75% of the data are above it. The third quartile is the value in which 75%75% of the data lie below it and 25%25% of the data are above it. The second quartile is the median which is the value in which 50%50% of the data lie above and below it.

**Example 1.6**

We are interested in summarizing the age of employees at company X. Below are the ages of each of the 5050 employees at the company. Calculate the first and third quartiles.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 4242 | 2020 | 3232 | 4747 | 3131 |
| 6666 | 2525 | 6464 | 2525 | 4646 |
| 7676 | 5656 | 3232 | 2020 | 5050 |
| 6060 | 5858 | 3131 | 8383 | 5151 |
| 2222 | 3232 | 6464 | 4949 | 7575 |
| 4040 | 4343 | 5454 | 4444 | 6262 |
| 4646 | 2727 | 3232 | 4949 | 3737 |
| 3838 | 5959 | 3333 | 5959 | 7373 |
| 2626 | 2626 | 8383 | 7171 | 3939 |
| 3535 | 3333 | 3535 | 2828 | 3535 |

To calculate the quartiles, we must use the ordered values as we did when we calculated the median:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | 11 | 2020 | | 22 | 2020 | | 33 | 2222 | | 44 | 2525 | | 55 | 2525 | | 66 | 2626 | | 77 | 2626 | | 88 | 2727 | | 99 | 2828 | | 1010 | 3131 | |  | |  |  | | --- | --- | | 1111 | 3131 | | 1212 | 3232 | | 1313 | 3232 | | 1414 | 3232 | | 1515 | 3232 | | 1616 | 3333 | | 1717 | 3333 | | 1818 | 3535 | | 1919 | 3535 | | 2020 | 3535 | |  | |  |  | | --- | --- | | 2121 | 3737 | | 2222 | 3838 | | 2323 | 3939 | | 2424 | 4040 | | 2525 | 4242 | | 2626 | 4343 | | 2727 | 4444 | | 2828 | 4646 | | 2929 | 4646 | | 3030 | 4747 | |  | |  |  | | --- | --- | | 3131 | 4949 | | 3232 | 4949 | | 3333 | 5050 | | 3434 | 5151 | | 3535 | 5454 | | 3636 | 5656 | | 3737 | 5858 | | 3838 | 5959 | | 3939 | 5959 | | 4040 | 6060 | |  | |  |  | | --- | --- | | 4141 | 6262 | | 4242 | 6464 | | 4343 | 6464 | | 4444 | 6666 | | 4545 | 7171 | | 4646 | 7373 | | 4747 | 7575 | | 4848 | 7676 | | 4949 | 8383 | | 5050 | 8383 | |

From above, we know that the median is 42.542.5. To find the first quartile, we look to find the “median” of the set of the data that is below the median. That is, we calculate the median for the first 2525 observations. The third quartile will be the median of the second half of the observations.

In this case, the first quartile is the 1313th value, 3232. The third quartile is the 3838th value (or the 1313th value in the latter 2525 observations), 5959.

**Example 1.7**

The earnings (in thousands of dollars) for 99 college graduates are shown below. Find the first and third quartiles.

|  |  |  |
| --- | --- | --- |
| 3535 | 4040 | 145145 |
| 3333 | 3030 | 4242 |
| 3232 | 3232 | 2525 |

As above, we first order the values:

25303232333540421452530323233354042145

Since we know from above that 3333 is the median, we use this as a reference point. We find the “median” of the values that are below 3333 to find the first quartile and we find the “median” of the values above 3333 to find the third quartile.

That is, the first quartile is the midpoint of these numbers:

2530323225303232

which is 3131 (the average of 3030 and 3232).

And, the third quartile is the midpoint of these numbers:

354042145354042145

which is 4141 (the average of 4040 and 4242).

The variance and the standard deviation are another measure of the spread. Both of these quantities measure how far each observation is from the mean.

The variance, denoted s2s2, is the average of the squared deviations of each observation from the mean. This is represented in the following equation:

s2=(x1−x¯)2+(x2−x¯)2+⋯+(xn−x¯)2n−1=1n−1∑i=1n(xi−x¯)2s2=(x1−x¯)2+(x2−x¯)2+⋯+(xn−x¯)2n−1=1n−1∑i=1n(xi−x¯)2

The standard deviation, ss, is the square root of the variance.

s=s2−−√=(x1−x¯)2+(x2−x¯)2+⋯+(xn−x¯)2n−1−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−√=1n−1∑i=1n(xi−x¯)2−−−−−−−−−−−−−−−√s=s2=(x1−x¯)2+(x2−x¯)2+⋯+(xn−x¯)2n−1=1n−1∑i=1n(xi−x¯)2

As before, the notation ∑ni=1∑i=1n means to sum up the values for each data point.

Values for standard deviation are always positive (s≥0s≥0). A standard deviation of 00 means that there is no spread (all values for all observations are the same). The standard deviation (and variance) increases as the values become more spread out about their mean. The units for the standard deviation are the same units of measure as the original observations. For example, if you are looking at age in years, the units of measure of the standard deviation are years (while the units of the variance are years squared). For this reason, standard deviation is more “interpretable” and is more commonly used to describe the spread than the variance. Although at first it may seem as though the standard deviation is difficult to interpret even though it is in the “right” units, we will see in the next section the importance of this quantity and its interpretation in symmetric distributions will become more clear.

Like the mean, the standard deviation and variance are not resistant to skewness and outliers. In fact, skewness and outliers have the potential to greatly inflate these measures of the spread.

**Example 1.8**

We are interested in summarizing the age of employees at company X. Below are the ages of each of the 5050 employees at the company. The mean is 45.2845.28. Calculate the variance and standard deviation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 4242 | 2020 | 3232 | 4747 | 3131 |
| 6666 | 2525 | 6464 | 2525 | 4646 |
| 7676 | 5656 | 3232 | 2020 | 5050 |
| 6060 | 5858 | 3131 | 8383 | 5151 |
| 2222 | 3232 | 6464 | 4949 | 7575 |
| 4040 | 4343 | 5454 | 4444 | 6262 |
| 4646 | 2727 | 3232 | 4949 | 3737 |
| 3838 | 5959 | 3333 | 5959 | 7373 |
| 2626 | 2626 | 8383 | 7171 | 3939 |
| 3535 | 3333 | 3535 | 2828 | 3535 |

We use the formula above to calculate the variance.

s2=(x1−x¯)2+(x2−x¯)2+⋯+(xn−x¯)2n−1=(42−45.28)2+(66−45.28)2+⋯+(35−45.28)250−1=(−3.28)2+(20.72)2+⋯+(−10.28)249=10.7584+429.3184+⋯+105.678449=14,502.2849≈295.96s2=(x1−x¯)2+(x2−x¯)2+⋯+(xn−x¯)2n−1=(42−45.28)2+(66−45.28)2+⋯+(35−45.28)250−1=(−3.28)2+(20.72)2+⋯+(−10.28)249=10.7584+429.3184+⋯+105.678449=14,502.2849≈295.96

The standard deviation is the square root of the variance:

s=s2−−√=295.96−−−−−√≈17.20s=s2=295.96≈17.20

**Test Yourself 1.4**

The earnings (in thousands of dollars) for 99 college graduates are shown below. The mean is 4646. Find variance and the standard deviation.

|  |  |  |
| --- | --- | --- |
| 3535 | 4040 | 145145 |
| 3333 | 3030 | 4242 |
| 3232 | 3232 | 2525 |

Show Answer

Due to the outlier of 145145 in the test yourself question above, the variance and standard deviation are inflated. If we would remove this observation, the variance drops to 29.429.4 (from 14041404) and the standard deviation drops to 5.45.4 (from 37.537.5). As discussed above, this example illustrates the fact that a single high value (or a skewed distribution) will drive the value of the variance and standard deviation up. The quartiles (similar to the median) are more resistant to outliers and skewness. Given this, the interquartile range (3rd quartile - 1st quartile) is the preferred summary of spread when the distribution is skewed.

Now that we have reviewed the key numerical summaries for quantitative data, let’s return to the second type of graphical summary for quantitative data, the box plot or the box and whisker plot (now that we have the knowledge of the numerical pieces that are needed for it). The median, minimum, maximum, 11st and 33rd quartiles are to create box plots. In a box plot, a central box spans the 11st and 33rd quartiles. The median is shown as a line in the box. Lines extend out from the box to show the minimum and maximum.

These are generally used to compare different distributions side-by-side (we will return to this concept in later modules). From the box plot, the spread of the middle 50%50% of the data is the focus. It also gives an indication of the symmetry. In a symmetric distribution, the median will be roughly half way between the first and third quartiles. If the distribution is skewed to the right, then the median will be closer to the first quartile than it is to the third quartile. Similarly, if the distribution is skewed to the left, then the median will be closer to the third quartile than it is to the first quartile.

**Example 1.9**

We are interested in summarizing the age of employees at company X. The table below summarizes the various numerical summaries of the data. Use this information to create a box plot.

|  |  |
| --- | --- |
| Mean | 45.2845.28 |
| Variance | 295.96295.96 |
| Standard Deviation | 17.2017.20 |
| Median | 42.542.5 |
| 1st Quartile | 3232 |
| 3rd Quartile | 5959 |
| Minimum | 2020 |
| Maximum | 8383 |

The box plot is displayed below. Each piece is labeled so it is clear what is being presented.

As we saw from the histogram of this data, the distribution is slightly skewed to the right (which is indicated by the fact that the median is closer to the 11st quartile than to the 33rd quartile).

**Example 1.10**

The earnings (in thousands of dollars) for 99 college graduates are shown below. The table below summarizes the various numerical summaries of the data. Use this information to create a box plot.

|  |  |
| --- | --- |
| Mean | 4646 |
| Variance | 14041404 |
| Standard Deviation | 37.537.5 |
| Median | 3333 |
| 1st Quartile | 3232 |
| 3rd Quartile | 4040 |
| Minimum | 2525 |
| Maximum | 145145 |

The box plot generated using software is below:

You may have noticed that this box plot’s upper line does not extend to the maximum (145145). This is because many software programs evaluate each point to see if it is an outlier.  Generally, this is defined as any points that are less than the first quartile −1.5⋅−1.5⋅ IQR or any points that are greater than the 3rd quartile +1.5⋅+1.5⋅ IQR where IQR is the inter-quartile range and is defined as the difference between the quartiles. In this case, IQR =40−32=8=40−32=8. So, any observation in this dataset that was less than 2020 or greater than 5252 would be considered an outlier and thus ignored for the purposes of developing the plot.

Generally, we use software to help calculate numerical summaries of the spread (both quartiles as well as the variance and standard deviation). The most commonly reported numerical summaries of the center and spread of a distribution are the mean and the standard deviation, especially when the distribution is symmetric. When the distribution is skewed or when there are outliers, the more resistant numerical summaries (the median and the quartiles) are preferred along with a graphical summary (either a histogram or a box and whisker plot). In these cases, graphical summaries are really the best way to describe and summarize the data. Software is also often used to help us create the graphical summaries discussed above.

**Normal Distribution**

As detailed in the previous sections, histograms are useful tools to graphically display quantitative data. They are also useful in helping us determine if a particular set of data appears to adhere to any standard idealized patterns. One of the most common and useful such patterns is the Normal distribution.

The Normal distribution is an important distribution in statistics. Though it is termed the “Normal” distribution, it is not “normal” in the sense of being average or natural. Instead, “normal” here refers to the regularity of the distribution. The density curve (a mathematical model that represents the pattern of data) of a normal distribution has a very familiar, bell curve shape with a single peak. It is perfectly symmetrical.

Many statistical inference procedures are based on the normal distribution or similar symmetric distributions like the tt-distribution which we will learn about in the next module.

For now, we will focus on the fact that real life data often are normally distributed and the fact for such distributions, we can use the well-known and understood properties of the normal distribution to describe and make predictions about the probability of events. For example, standardized test scores taken by large groups of people (such as the SAT and many psychological tests), repeated careful measurements of the same quantity, and characteristics of biological populations (such as weights or heights of individuals or animals or yields of crops) are generally normally distributed.

**Example 1.11**

In a 2006 article (Low Birth Weight, a Risk Factor for Cardiovascular Diseases in Later Life, Is Already Associated with Elevated Fetal Glycosylated Hemoglobin at Birth) in Circulation, a histogram is shown of the birth weights of newborns in their sample. The figure from the paper is shown below.

As evidenced by the histogram, the distribution of birth weights is symmetric with a single peak. The tails of the distribution fall off smoothly. There are no gaps or obvious outliers. The normal density curve is shown on top of the histogram (the bell shaped curve shown in the solid black line). The normal density curve seems to fit the data well (that is, the curve mirrors the shape of the histogram). It is an “idealized” description of the data. It gives the general picture of the data, ignoring minor irregularities. In situations like this, we can use properties of the normal distribution to make statements about the quantitative variable.

Remember that the area of the bars in a histogram represent the proportion of the observations. Even when the y-axis in a histogram are counts instead of relative frequencies, the area is proportional to the proportion of observations that the bar or bars correspond to. As such, if you were interested in knowing the proportion of newborns who weighed more than 40004000 grams, you could use the histogram above to find this. Or, alternatively you could use the normal distribution.

Using the histogram, we could determine the number of infants of the 12951295 that weighed more than 40004000 grams. This would be equivalent to the proportion of the area of the bars of the histogram to the right of 40004000 grams to the area for all of the bars in the histogram (as pictured in the graph below where the area of the bars in the histogram to the right of 40004000 are shaded in green). Alternatively, if we scale the curve such that the total area under the curve is 100%100%, then the area to the right of 40004000 grams under the curve (shown shaded in blue in the figure below) would be an approximation of the proportion of infants weighing more than 40004000 grams.

As in the example above, no real life dataset will perfectly follow a normal distribution. However, the normal distribution density curve provides an approximation that is easy to use and accurate enough for most practical applications.

The exact density curve for a particular Normal distribution is specified by giving its mean, μμ (the Greeek letter “mu”), and its standard deviation, σσ (the Greek letter “sigma”). This is often denoted N(μ,σμ,σ). The mean of a Normal distribution is the center of the distribution and is the point that splits the area under the bell shaped curve in half.

Two Normal distributions with the same standard deviation but with different means have exactly the same spread but are centered on different places along the x-axis (horizontal axis). This concept is shown in the figure below where the Normal distribution shown by the green density curve has a larger mean than the Normal density curve shown in blue.

Two Normal distributions with the same mean but with different standard deviations are both centered on the same place along the x-axis, but will have different amounts of spread. In the figure below, the normal distribution shown in green has a larger standard deviation than the one shown in blue.

**The 68-95-99.7 Rule**

Although there are many different Normal density curves, they all share common properties.

One of the most useful properties is called the 68-95-99.7 rule. This property allows you to quickly determine probabilities without making detailed calculations. It states that for a normal distribution with mean, μμ, and standard deviation, σσ,

* 68%68% of the observations fall within one standard deviation of the mean
* 95%95% of the observations fall within two standard deviations of the mean
* 99.7%99.7% of the observations fall within three standard deviations of the mean

This property is shown in the figure below.

**Example 1.12**

Let’s assume that the distribution of birth weights is normally distributed with a mean of 35003500 grams with a standard deviation of 500500 grams. What proportion of infants weigh between 30003000 and 40004000 grams at birth? What proportion weigh between 25002500 and 45004500 grams? What percentage weigh more than 40004000 grams?

All of these questions can be answered using the 68-95-99.7 rule. The rule tells us that 68%68% of the observations fall within one standard deviation of the mean. This means that 68%68% of the data are between μ−σμ−σ and μ+σμ+σ. Since here, μ=3500μ=3500 grams and σ=500σ=500 grams, this means that 68%68% of the data are between μ−σ=3500−500=3000μ−σ=3500−500=3000 grams and μ+σ=3500+500=4000μ+σ=3500+500=4000 grams.

The rule also tells us that 95%95% of the observations fall within two standard deviations of the mean. This means that 95%95% of the data are between μ−2⋅σμ−2⋅σ and μ+2⋅σμ+2⋅σ. Since here, μ=3500μ=3500 grams and σ=500σ=500 grams, this means that 95%95% of the data are between μ−2⋅σ=3500−2⋅500=2500μ−2⋅σ=3500−2⋅500=2500 grams and μ+2⋅σ=3500+2⋅500=4500μ+2⋅σ=3500+2⋅500=4500 grams.

Since 68%68% of the data are between 30003000 grams and 40004000 grams, this also means that 100%−68%=32%100%−68%=32% of the data are outside of this range (as the total area under the curve must equal 100%100%). Given the symmetry of the Normal distribution, this means that half of this (16%16%) is below 30003000 grams and half (16%16%) is above 40004000 grams. This is shown in the figure below.

**Test Yourself 1.5**

The distribution of SAT scores for the verbal section for high school seniors is approximately a normal distribution with a mean of 504504 and a standard deviation of 111111. What proportion of seniors score between 282282 and 726726?

Show Answer

**The Standardized Normal Distribution**

As the 68-95-99.7 rule suggests, if we measure in standard deviation units, all normal distribution density curves are the same. That is, if we define an area under a normal distribution in terms of standard deviation units, then regardless of the actual values of the mean and standard deviation, that area corresponds with the same proportion of observations.

We can convert a value into standard deviation units by calculating the value’s zz-score. If xx is an observation from a distribution with mean, μμ, and standard deviation, σσ, then its zz-score is calculated as

z=x−μσz=x−μσ

The zz-score is considered the standardized value of xx as it tells us how many standard deviations xx is from the mean. Positive zz-scores indicate that the original value was greater than the mean while negative zz-scores indicate that the original value was less than the mean.

If the original variable was normally distributed with mean, μμ, and standard deviation, σσ, the distribution of standardized zz values follows a normal distribution with a mean of 00 and a standard deviation of 11 (the standard Normal distribution).

This is an exceptional result, since this means that regardless of what the mean and standard deviation are for a particular normal distribution, we can use this standardization to “transform” any Normal distribution into a standard Normal distribution, and we can use a single distribution (the standard Normal distribution) to compute areas under the curve (which translate into proportions of observations). That is, if we have a normally distributed variable and we are interested in answering questions about what proportion of the observations lie in some range of values, then we can express the range in terms of the standard normal distribution (by using the zz-score above). Since any normal distribution can be expressed as a standard normal distribution by using this transformation, we can focus our understanding on the standard normal distribution and any information and properties about this particular distribution will apply any time there is a variable that is normally distributed.

Areas under the standard normal curve require special calculations or software to compute. That is, there are not simple formulas or rules (besides the one mentioned in the previous section) to help compute areas. The most common way of computing areas “by hand” involve using tables (which give areas under the standard normal curve). Check out [an example of such a table](https://onlinecampus.bu.edu/bbcswebdav/pid-11024813-dt-content-rid-76146031_1/courses/22fallmetcs555_a2/Course_Content/documents/Table_A_Standard_Normal_Probabilities.pdf). We will use these tables for any “by hand” calculations. In these tables, entries give the area under the standard Normal density curve to the left of zz. The interpretation of these values is the proportion of observations under a normal curve to the left of zz. Or, it can also be interpreted as the probability that a single observation from a standard normal distribution is less than zz.

Some important properties about the standard normal density curve are summarized below:

1. The total area under the standard normal curve is 1.001.00.
2. The curve is perfectly symmetrical. If zz is greater than 00, then the proportion of observations that are less than −z−z is equal to the proportion of observations that are greater than zz. That is, if zz is greater than 00, then the probability that a given observation is less than −z−z is equal to the probability that a given observation is greater than zz. This is shown in the figure below. The two areas to the left and right are equal.
3. The standard normal curve is centered on 00. Then, the probability that zz is greater than 00 or less than 00 is 50%50%. That is, the area to the left of z=0z=0 is 0.50000.5000. The area to the right of z=0z=0 is 0.50000.5000. The areas in green (to the left of z=0z=0) and white (to the right of z=0z=0) in the figure below are both equal to 0.500000.50000.
4. The probability that z≥az≥a is equal to the probability that z>az>a. There is no area under the curve for a single point. Similarly, the probability that z≤az≤a is equal to the probability that z<az<a. There is no area under the curve for a single point.

**Example 1.13**

Using the tables, find the area under the standard normal curve to the left of z=1.53z=1.53.

Here, we are looking to find the area under the standard normal curve to the left of z=1.53z=1.53. This area is shown in the figure below shaded in green.

Since z>0z>0, we use the second page of the table (as shown below). To find the area under the curve to the left of z=1.53z=1.53, we first locate the row associated with 1.51.5 (in the column on the far left). Then, we locate the remaining digit 33 by looking for the .03.03 column (in the top row). The value in the table where this row and column intersect is the proportion of observations from the standard normal distribution that is less than z=1.53z=1.53. Per the table, the area under the standard Normal curve to the left of zz is 0.93700.9370.

**Example 1.14**

Using the tables, find the proportion of observations greater than z=−0.58z=−0.58 under the standard normal.

The proportion of observations greater than z=−0.58z=−0.58 is the area under the standard normal curve to the right of z=−0.58z=−0.58. This area is shown in the figure below shaded in green. To find this, we first find the area to the LEFT of z=−0.58z=−0.58 (since this is the area that we can find using the tables) and then we use this to find the area to the RIGHT of z=−0.58z=−0.58.

Since z<0z<0, we use the first page of the table (as shown below). To find the area under the curve to the LEFT of z=−0.58z=−0.58, we first locate the row associated with −0.50−0.50 (in the column on the far left). Then, we locate the remaining digit 88 by looking for the .08.08 column (in the top row). The value in the table where this row and column intersect is the proportion of observations from the standard normal distribution that is less than z=−0.58z=−0.58. Per the table, the area under the standard Normal curve to the LEFT of zz is 0.28100.2810. Since the total area under the curve is equal to 11 (or 100%100%), the area to the RIGHT of z=−0.58z=−0.58 is 1–0.2810=0.71901–0.2810=0.7190. So, the proportion of observations under the standard normal curve that are less than z=−0.58z=−0.58 is 71.90%71.90%.

**Example 1.15**

Using the tables, find the area under the standard normal curve between z=−1.25z=−1.25 and z=0.77z=0.77.  
To find the area between z=−1.25z=−1.25 and z=0.77z=0.77, we first need to find the area to the left of z=−1.25z=−1.25 as well as the area to the left of z=0.77z=0.77. Using the tables, the area to the left of z=−1.25z=−1.25 is 0.10560.1056. The area to the left of z=0.77z=0.77 is 0.77940.7794.

It helps to see what this looks like. See the figure below.

Then to find the area between, we can just subtract the areas from each other. That is, the area between z=−1.25z=−1.25 and z=0.77z=0.77 is equal to 0.7794−0.1056=0.67380.7794−0.1056=0.6738.

**Test Yourself 1.6**

Using the tables, find the probability that zz is greater than 1.961.96.

Show Answer

We can answer any question about the proportion of the observations in a Normal distribution by standardizing and then using the standard Normal table. Here are the steps we generally need to take:

1. State the problem in terms of xx
2. Standardize xx to restate the problem in terms of the standard normal variable zz. Draw a picture to show the area under the standard normal curve.
3. Find the required area under the standard normal curve using Table A and the fact that the total area under the curve is 11.

**Example 1.16**

Let’s assume that the distribution of birth weights is normally distributed with a mean of 35003500 grams with a standard deviation of 500500 grams. What proportion of infants weigh less than 28002800 grams?

In order to answer this, we will follow the steps laid out above:

1. State the problem in terms of xx

Let xx be the birth weight of infants. The variable xx has the N(3500,5003500,500) distribution. We are interested in the proportion of infants with x<2800x<2800 grams.

1. Standardize xx to restate the problem in terms of the standard normal variable zz. Draw a picture to show the area under the standard normal curve.

x<2800x<2800

x−3500500<2800−3500500x−3500500<2800−3500500

z<−1.40z<−1.40

1. Find the required area under the standard normal curve using Table A and the fact that the total area under the curve is 11.

To find the area to the left of z=−1.40z=−1.40, we can use the value in Table A directly. The value in Table A for z=−1.40z=−1.40 is 0.08080.0808. As such, 8.08%8.08% of infants weigh less than 28002800 grams.

**Example 1.17**

Let’s assume that the distribution of birth weights is normally distributed with a mean of 35003500 grams with a standard deviation of 500500 grams. What proportion of infants weigh greater than or equal to 32003200 grams?

In order to answer this, we will follow the steps laid out above:

1. State the problem in terms of xx

Let xx be the birth weight of infants. The variable xx has the N(3500,5003500,500) distribution. We are interested in the proportion of infants with x≥3200x≥3200 grams.

1. Standardize xx to restate the problem in terms of the standard normal variable zz. Draw a picture to show the area under the standard normal curve.

x≥3200x≥3200

x−3500500≥3200−3500500x−3500500≥3200−3500500

z≥−0.60z≥−0.60

1. Find the required area under the standard normal curve using Table A and the fact that the total area under the curve is 11.

To find the area to the right of z=−0.60z=−0.60, we can first use Table A to find the area to the left of z=−0.60z=−0.60 and then subtract the result from 11. The value in Table A for z=−0.60z=−0.60 is 0.27430.2743 which corresponds to the area to the left. Then the area to the right is equal to 1–0.2743=0.72571–0.2743=0.7257. As such, 72.57%72.57% of infants weigh greater than or equal to 32003200 grams.

Note that here the question asked about infants weighing *greater than or equal to* 32003200 grams. If the question was asking about infants weighing *greater than* 32003200 grams, the result would have been the same. That’s because as noted in the properties discussed above, the probability that z≥az≥a is equal to the probability that z>az>a since there is no area under the curve for a single point.

**Example 1.18**

Let’s assume that the distribution of birth weights is normally distributed with a mean of 35003500 grams with a standard deviation of 500500 grams. What proportion of infants between 32503250 and 37503750 grams?

In order to answer this, we will follow the steps laid out above:

1. State the problem in terms of xx

Let xx be the birth weight of infants. The variable xx has the N(3500,5003500,500) distribution. We are interested in the proportion of infants with 3250<x<37503250<x<3750 grams.

1. Standardize xx to restate the problem in terms of the standard normal variable zz. Draw a picture to show the area under the standard normal curve.

3250<x<37503250<x<3750

3250−3500500<x−3500500<3750−35005003250−3500500<x−3500500<3750−3500500

−0.50<z<0.50−0.50<z<0.50

1. Find the required area under the standard normal curve using Table A and the fact that the total area under the curve is 11.

To find the area between z=−0.50z=−0.50 and z=0.50z=0.50, we can subtract the area to the left of z=−0.50z=−0.50 from the area to the left of z=0.50z=0.50. The values in Table A associated with z=−0.50z=−0.50 and z=0.50z=0.50 are 0.30850.3085 and 0.69150.6915, respectively. Then, the area between z=−0.50z=−0.50 and z=0.50z=0.50 is 0.6915–0.3085=0.38300.6915–0.3085=0.3830. As such, 38.30%38.30% of infants weigh between 32503250 and 37503750 grams.

**Test Yourself (Multipart) 1.7**

The distribution of SAT scores for the verbal section for high school seniors is approximately a normal distribution with a mean of 504504 and a standard deviation of 111111. What proportion of seniors score greater than 600600?

(Click each step to reveal its solution.)

State the problem in terms of xx.

Standardize xx to restate the problem in terms of the standard normal variable zz. Draw a picture to show the area under the standard normal curve.

Find the required area under the standard normal curve using Table A and the fact that the total area under the curve is 11.

To find the area to the right of z=0.86z=0.86, we can first use Table A to find the area to the left of z=0.86z=0.86 and then subtract the result from 11. The value in Table A for z=0.86z=0.86 is 0.80510.8051 which corresponds to the area to the left. Then the area to the right is equal to 1–0.8051=0.19491–0.8051=0.1949. As such, approximately 20%20% of seniors score higher than 600600 on the verbal section of the SAT.

**Test Yourself (Multipart) 1.8**

The distribution of SAT scores for the verbal section for high school seniors is approximately a normal distribution with a mean of 504504 and a standard deviation of 111111. What proportion of seniors score between 504504 and 760760?

(Click each step to reveal its solution.)

State the problem in terms of xx.

Let xx be the SAT verbal scores of high school seniors. The variable xx has the N(504,111504,111) distribution. We are interested in the proportion of seniors with 504<x<760504<x<760.

Standardize xx to restate the problem in terms of the standard normal variable zz. Draw a picture to show the area under the standard normal curve.

504<x<760504<x<760

504−504111<x−504111<760−504111504−504111<x−504111<760−504111

0<z<2.310<z<2.31

Find the required area under the standard normal curve using Table A and the fact that the total area under the curve is 11.

To find the area between z=0z=0 and z=2.31z=2.31, we can subtract the area to the left of z=0z=0 from the area to the left of z=2.31z=2.31. The values in Table A associated with z=0z=0 and z=2.31z=2.31 are 0.50000.5000 and 0.98960.9896, respectively. Then, the area between z=−0.50z=−0.50 and z=0.50z=0.50 is 0.9896–0.5000=0.48960.9896–0.5000=0.4896. As such, 48.96%48.96% of seniors score between 504504 and 760760 on the verbal section of the SAT.

In the examples above, we used Table A to find the proportion of the observations that satisfies some condition that is defined based on a value or values of xx. Sometimes, however, we are interested in finding the observed value of xx that’s associated with a given proportion of the observations above or below it. To do this, we have to do the calculations that we just did, only backwards!

**Example 1.19**

Let’s assume that the distribution of birth weights is normally distributed with a mean of 35003500 grams with a standard deviation of 500500 grams. How much must an infant weigh to be in the top 10%10% of infant birth weights?

We want to find the birth weight xx with an area of 0.10000.1000 to the right under the Normal curve with mean μ=3500μ=3500 and standard deviation of σ=500σ=500. This is the same as finding the birth weight xx with an area of 0.90000.9000 to the left. This question is displayed in the figure below.

Since Table A has areas to the left of zz-values, we want to look to see which zz-value is associated with an area of 0.90000.9000. So, we look for a table entry as close to 0.90000.9000 as possible. The closest value in the table is 0.89970.8997 which corresponds to a zz-score of 1.281.28. So z=1.28z=1.28 is the standardized value with area 0.900.90 to its left.

To transform this value of zz back into the original xx scale, we need to “unstandardized.” We know that

x−μσ=zx−μσ=z

x−3500500=1.28x−3500500=1.28

Solving for xx we get,

x=z⋅σ+μ=1.28⋅500+3500=4140x=z⋅σ+μ=1.28⋅500+3500=4140

That is, 41404140 is 1.281.28 standard deviations above the mean on this particular Normal curve. An infant must weigh more than 41404140 grams to be in the top 10%10% of infant weights.

Now that we have a better understanding of the normal distribution, we switch gears now and discuss sampling. We will return to concepts involving the Normal distribution at the end of the module.

**Sampling**

In the previous sections, we have looked at various methods for summarizing data using both graphical visualizations as well as numerical summaries. These methods allow us to understand and describe the data we are examining.

However, most times we are not specifically interested in just the data from our sample, but we are interested in answering questions about people, animals, or things outside of the data we have available to us. Many times, constraints on time, money, and resources prevent us from being able to gather information on the entire group of people or things we are interested in. In order to answer such questions, we instead gather information about only part of the group and use this information to make educated guesses about qualities of the larger group. However, we need to be sure that the part of the group (our sample) that we collect information about is selected in such a way that allows for inference to the (larger) group of interest (the population).

For example, suppose we were interested in knowing what percentage of Americans adults have an iPhone. In order to attain this information, we could ask adults in the US. It would be impossible and/or impractical to ask every adult in America whether or not they have an iPhone, but we could ask the question to a small sample of American adults chosen in such a way that we feel they represent the entire population of adults in America and use this information to make a guess at the proportion of all American adults that have an iPhone.

The entire group of individuals or things that we want information about or to answer questions about is called the population. A sample is a subset of the population that we actually examine and gather information about.

In the example above, the population of interest would be all American adults. In this case, we’d be interested in finding a smaller sample of adults that are representative of the general US adult population.

**Example 1.20**

Suppose we were interested in knowing what percentage of college students approve of the current president. In order to answer this question, suppose we questioned a group of 100 undergraduate students at Brigham and Young University.

In this case, our population of interest is all college students. Our sample, on the other hand, would be the group of undergraduate students at Brigham and Young University to whom we asked about their approval of the president.

The data we collect from a sample is only as good as the methods used to collect it. A sample that is not representative of the population of interest or one that favors a specific outcome will lead to biased inference about the population.

**Example 1.21**

In the example above where we were interested in the approval rating of the president among American college students, our sample was from Brigham and Young University. Since Brigham and Young University has traditionally been a conservative school, the percentage of Brigham and Young undergraduates that are republican, for example, is likely higher than the percentage of all American college students that are republican. This is likely to influence their opinion and approval of the president. As such, sampling from this particular university alone may not be the best way to attain information about the president’s approval rating among all American college students.

Generally, we want to try to select a sample that is as similar to the population as possible. The characteristics of the sample should be similar to what we’d expect from the larger population. The distribution of characteristics of the study sample is often one way we have of assessing how representative a sample is of the population. For example, if we took a sample of college students and all of those in our sample were male, then we’d have evidence of bias in our sampling methodology since we know that not all college students are male.

In determining how to select a sample, we must try to ensure that all groups and individuals have an equal chance of being selected. That is, we’d like to ensure that the way we selected those in our sample didn’t systematically (even if inadvertent) exclude those of a particular socioeconomic status, race, education level, gender, or other characteristic that is part of the population of interest.  For example, opinion polls conducted by telephone will miss Americans without telephones (or who are on the “Do Not Call” list). Door to door surveys miss homeless people, those in prison, and students living in college dormitories.

Many polls use shopping malls to find samples. Though such places are convenient places to find people to interview, they may not be representative of the targeted sample (may be of higher socioeconomic status, may be younger (teenagers) or older (retirees)). Also, poll interviewers may inadvertently only approach particular types of people (those that seem less in a rush, those that are more “put together,” etc).

**Example 1.22**

In trying to pick out a new restaurant to try out, many of us turn to Yelp reviews to help select where to go. However, how many of us have actually written reviews of the places we have visited?  Generally, voluntary response samples (samples that are comprised of responses from individuals who elected to participate on their own free will to a general call for an opinion on a topic) are often biased since those with strong opinions, especially those that are negative, are more likely to respond. This is also true of polls from news organizations (which suggest that viewers call in or respond to an online survey after reading an article on the topic).

In many ways, all samples will have some type of responder bias since those who elect to participate in surveys or designed experiments tend to be different than those that who chose not to participate.

We must also be sensitive to biases that could be imposed by the questions that we ask (and the way that we ask them).

**Example 1.23**

Top of Form

In Africa, the “big five” game animals are the lion, the elephant, the buffalo, the leopard, and the rhinoceros. These are termed the “big five” as they are known as the five most dangerous and difficult animals in Africa to hunt on foot. Which is your favorite?

|  |  |
| --- | --- |
|  | Lion |
|  | Elephant |
|  | Buffalo |
|  | Leopard |
|  | Rhinoceros |

Show Answer

Bottom of Form

In order to avoid biases introduced by sampling (often termed selection bias), we attempt to try to choose samples by chance (instead of letting a poll interviewer or responder choose who is selected for the sample).  If we have a list of all members of a population (if this is even possible), then we can randomly choose our sample. The simplest way of thinking about this is to think of names in a hat (the population) and selecting out a handful (sample). In doing this, we ensure that each individual or element of the population has an equal chance of being selected.  Ensuring that each individual or element from the population has an equal chance of being selected is key to protecting against selection bias.

Even in the most carefully selected samples, bias may be introduced if we have a large amount of non-response. That is, if we select 50 households from a community, many of those that we select may either be unreachable or may elect not to participate. If those that chose not to participate would have answered differently than those that chose to participate (something nearly impossible to measure in practice), then we have introduced bias into our data collection.

In the following sections and modules, we will focus on the analysis of data from samples. Though we won’t focus in most cases on where the data came from or how it was collected, we should be mindful of this in our own research as the methods used for the collection of the data that we analyze is critically important to our ability to answer questions and to make inferences and conclusions. We should be mindful of sampling methods and potential for bias as we critically analyze our results.

**Inference about a Population**

In real life applications, information about a population is unknown. As described above, we take a random sample from a population and use this information to make inferences about the population. We use specific vocabulary and notation to differentiate information about a sample from information about a population:

A **p**arameter is a number that describes a **p**opulation. In practice, the value of the parameter is unknown since we often are not able to examine the entire population.

A **s**tatistic is a number that is computed from the **s**ample data. In practice, we use information contained in sample statistics to estimate the unknown population parameter.

The table below provides a list of typical notation for population parameters and their corresponding sample statistic.

|  |  |  |
| --- | --- | --- |
|  | Population Parameter | Sample Statistic |
| Mean | μμ (“mu”) | x¯x¯ (“x bar”) |
| Variance | σ2σ2 (“sigma squared”) | s2s2 |
| Standard Deviation | σσ (“sigma”) | ss |
| Proportion | pp | p^p^ (“p hat”) |

The Greek letter mu (μμ) is used to denote the mean of a population. This is a fixed quantity that is generally unknown and which we wish to estimate using information from a sample. The mean of the sample, x¯x¯, is as calculated in the previous sections as the arithmetic average of the observations in the sample. The sample value, x¯x¯, is dependent on the sample taken. If we took two separate random samples, we’d like end up with two different values of the sample mean. The sample mean, x¯x¯, is an estimate of the population mean, μμ.

**Example 1.24**

A researcher was interested in estimating the mean income level for college graduates aged 2525–3030. In order to do so, a random sample of 20002000 college graduates was taken. The mean income of the sample was x¯=$45,455x¯=$45,455. The number $45,455$45,455 is a statistic because it describes the sample. The unknown population parameter, μμ, would be the mean of all college graduates aged 2525–3030. This latter quantity is unknown and is estimated by the sample mean. The actual value could only be obtained if you took the arithmetic average of all approximately 1212 million college graduates aged 2525–3030.

**Test Yourself 1.9**

Voter registration for the state of Massachusetts showed that 76%76% of party affiliated voters were registered as Democrats in October of 2012. A survey of 200200 randomly selected voters exiting the polls in the last election in late 2014 showed that 74%74% of voters were registered as democrats. Is 76%76% a population parameter or a sample statistic? Is 74%74% a population parameter or a sample statistic?

Show Answer

**Distributions of Sample Statistics**

In order to know which sample statistic to use to estimate a particular parameter, we must first evaluate the properties of the sample statistic. For example, if we want to estimate the population parameter μμ, then we could use a number of sample statistics for this purpose. We might consider the sample mean x¯x¯ or the sample median which we will denote mm.

**Example 1.25**

Suppose we are interested in estimating the mean age of employees at company X. For convenience, let’s assume that the population is made up of only 5050 individuals with the following ages:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 42 | 20 | 32 | 47 | 31 |
| 66 | 25 | 64 | 25 | 46 |
| 76 | 56 | 32 | 20 | 50 |
| 60 | 58 | 31 | 83 | 51 |
| 22 | 32 | 64 | 49 | 75 |
| 40 | 43 | 54 | 44 | 62 |
| 46 | 27 | 32 | 49 | 37 |
| 38 | 59 | 33 | 59 | 73 |
| 26 | 26 | 83 | 71 | 39 |
| 35 | 33 | 35 | 28 | 35 |

A histogram of the ages of workers at company X would look like this:

The population mean of the ages of employees at company X is 45.2845.28 years. However, let’s assume that due to limited resources and limited access to information, this value is unknown and we are looking to estimate it.

Let’s assume we take two random samples of size 55. Here are the resulting ages in each of the two samples.

Sample 1:  20,44,46,20,4420,44,46,20,44  
Sample 2:  83,32,31,50,3283,32,31,50,32

The first sample has a sample mean of 34.634.6 and a sample median of 4444. The second sample has a sample mean of 45.645.6 and a sample median of 3232.

In the first sample, the sample median was closer to the population value of 45.2845.28. In the second sample, the sample mean was closer to the population mean.

As you can see from the example above, neither the sample mean nor the sample median will *always* fall closer to the population mean in a given sample. As such, to evaluate each of these sample statistics and their ability to estimate the true population value, we must not rely on just one example (one sample). Instead, it would be best to see how each of these statistics does if we were to perform this experiment many, many times. That is, we’d like to compare the distribution of the sample mean and sample median if we take hundreds of random samples of the same size. In practice, we only have one sample and the true value of the underlying population parameter is unknown. As such, we must rely on the properties of an estimator.

**Example 1.26**

Let’s return to the example above where we were attempting to estimate the population mean of the ages of employees at company X by taking a random sample of size 55. Since we have the entire population for this particular example, we can use it to evaluate how well the sample mean and sample median perform as estimators of the population mean by using a computer to randomly select 10,00010,000 samples of size 55 (just as in the example above). For each sample, we calculated the sample mean and the sample median. Here’s what our results look like:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Sample | x1x1 | x2x2 | x3x3 | x4x4 | x5x5 | x¯x¯ | mm |
| 1 | 20 | 44 | 46 | 20 | 44 | 34.6 | 43 |
| 2 | 83 | 32 | 31 | 50 | 32 | 45.6 | 32 |
| 3 | 58 | 49 | 31 | 32 | 50 | 44 | 49 |
| 4 | 49 | 38 | 31 | 71 | 32 | 44.2 | 38 |
| 5 | 40 | 27 | 76 | 47 | 62 | 50.4 | 47 |
| 6 | 59 | 27 | 32 | 66 | 46 | 46 | 46 |
| 7 | 31 | 64 | 38 | 42 | 62 | 47.4 | 42 |
| 8 | 83 | 49 | 50 | 39 | 22 | 48.6 | 49 |
| 9 | 40 | 26 | 22 | 49 | 54 | 38.2 | 40 |
| 10 | 27 | 47 | 64 | 39 | 54 | 46.2 | 47 |
| … | | | | | | | |
| 10,000 | 25 | 26 | 33 | 60 | 71 | 43 | 33 |

In order to “see” this data better, we can look at the distribution of the sample mean (x¯x¯) by using a histogram:

The distribution of the sample median (mm) for the 10,00010,000 samples of size 55 is shown below:

As above, the population mean of the ages of employees at company X is 45.2845.28 years. Which estimator generally did a better job of estimating the population mean? From comparing the distributions of each of the sample statistics above as displayed by the histograms (which show the frequency of each value taken on by each sample statistic across the 10,00010,000 random samples), the sample mean appears to be a better estimate of the population mean. The distribution for the sample mean is centered over the true value μμ (since the mean of the sample means or the middle of the distribution is close to the value of μμ) and has less spread or variability than the distribution of the sample median (as evidenced by the width of the distribution and/or the standard deviation of the sample means across the 10,00010,000 random samples). These characteristics are preferred since in a real life example, we only get one sample. Given the distributions above, the probability is higher that if we take a single sample and calculate the sample mean from this sample that this value will be closer to the true value of the population mean than the sample median will be.

Often, many different point estimates (sample statistics that are “candidates” for estimating a population parameter) are possible for the same population parameter. Each estimator will have its own distribution. The exact distribution of a point estimate can be derived mathematically. However, the theory and mathematics for doing so are beyond the scope of this course. In general, though, these distributions (whether derived mathematically or estimated by using computer simulation as we had done in the example above) are used to compare estimators and to select the best one for a particular parameter of interest. As in the above example, we tend to prefer estimators that are centered over the true value and those that have less spread around the true population parameter. Sometimes, it is a bit of a balancing act (if one estimator is centered over the true value and has a lot of variability while the other is slightly off centered but has less variability).

In real life applications, we are often interested in estimating the population mean. For example, we may be interested in estimating the mean birth weight of infants, the mean number of adults per household, the mean yield of products per factory worker per hour, etc. In each of these cases, the best estimator of this particular parameter is of interest since it is rare that we would be able to look at the entire population. As shown in the above example, the sample mean provides a good estimate of the population mean when we take a random sample. The properties that we saw above were not “magic” or coincidental; they were just a demonstration of the inherent qualities of this estimator. The sample mean has some very interesting and important mathematical properties which we will explore. These properties make much of the theory underlying the material presented throughout the remainder of this course possible.

**The Central Limit Theorem**

The Central Limit Theorem tells us that when the number of samples taken from a population is sufficiently large, the sampling distribution of the sample mean, x¯x¯, will be approximately normally distributed with an expected value of μμ and a standard deviation of σn√σn where μμ and σσ are the mean and the standard deviation from the population.

That is, if you take a random sample of size nn from a population (with *any* underlying distribution, this doesn’t even have to be a normal distribution) with mean μμ and standard deviation σσ, then when nn is sufficiently large, the distribution of sample means will be approximately normally distributed with a mean of μμ and a standard deviation σn√σn. The larger the sample size, the closer the sampling distribution of the sample means will be to the normal distribution (and the smaller the variance will be of the sample mean).

This result means that the sampling distribution of the sample mean is centered on the population mean (this can also be conveyed by calling the sample mean an unbiased estimator). It also means that the sample mean is less variable than the individual observations from the population. Finally, this result also tells us that the sample mean from a larger sample is less variable than the sample mean taken from a smaller sample.

This concept is perhaps best illustrated through pictures and examples. Let’s return to the example above.

**Example 1.27**

Let’s return to the example above where we were attempting to estimate the population mean of the ages of employees at company X by taking a random sample. We had originally looked at the distribution of the population which showed that the distribution of ages was slightly skewed:

The central limit theorem tells us that if we take a random sample of size nn, then the distribution of sample means will be approximately normally distributed if nn is sufficiently large. It also tells us that as nn increases, the distribution will look more and more like the normal distribution and that the variability will reduce. We can “see” this by taking a look at the distribution of the sample means with various sizes for nn. (Click the numbered boxes in the object below to see n=3,5n=3,5, etc.)

*  3
*  5
*  10
*  15
*  20
*  25
*  30

As shown in the graphs above, as nn increases, the distribution of the sample means looks more and more like a normal distribution. Also, the variability of the sample means (and the width of the distribution) decreases as nn  increases. In the graphs above, the axes are the same so as to facilitate easier comparisons across graphs.

The reduction in variability can also be observed numerically. The population had a mean of 45.2845.28 and a standard deviation of 17.2017.20. The same statistics on the simulated data are shown below.

|  |  |  |
| --- | --- | --- |
| nn | x¯x¯x¯x¯ | σx¯σx¯ |
| 33 | 45.2545.25 | 9.579.57 |
| 55 | 45.2745.27 | 7.197.19 |
| 1010 | 45.2045.20 | 4.804.80 |
| 1515 | 45.2845.28 | 3.723.72 |
| 2020 | 45.2545.25 | 2.972.97 |
| 2525 | 45.2845.28 | 2.432.43 |
| 3030 | 45.2845.28 | 1.971.97 |

In each case, the center of the distribution (the sample mean of the sample means) is approximately equal to μμ and the standard deviation of the sample means decreases as nn increases.

How large must nn be for the theory underlying the Central Limit Theorem to work? That is, how large is “sufficiently large”? This really depends on the shape of the underlying population distribution. If the underlying population is approximately normal to begin with, then even small values of nn will give a sampling distribution of sample means that are normally distributed (as shown in the figure below). For more skewed population distributions, nn must be larger before the sampling distribution is sufficiently normally distributed. Generally, the rule of thumb is that nn should be ≥30≥30 for the distribution of the sample means to be reasonably normally distributed.

Sampling Distribution of x¯x¯ for…

We can use the normality of the sampling distribution of the sample means to find probabilities relating to the sample mean. This is similar to the way that we calculated probabilities using the normal distribution in previous sections. In these cases (where we are interested in making statements about the sample mean), we will use the normality and the distribution of the sampling distribution of the sample mean (this means that we will use the parameters associated with the distribution of the sample mean as opposed to using the population parameters directly).

**Example 1.28**

Suppose we have selected a random sample of n=36n=36 from a population with a mean of 8080 and a standard deviation of 66. Find the probability that the sample mean will be between 7979 and 8181.

The Central Limit Theorem tells us that regardless of the shape of the underlying population, the sampling distribution of x¯x¯ is approximately normal when n≥30n≥30 and that the sampling distribution will have a mean of μx¯=80μx¯=80 and σx¯=σn√=636√=66=1σx¯=σn=636=66=1.

Using the normal distribution, we can calculate the probability that the sample mean will be between 7979 and 8181.We first standardize (from the N(μx¯=80,σx¯=1μx¯=80,σx¯=1) distribution) and then we use Table A to calculate the probabilities:

79<x<8179<x<81

79−801<x−801<81−80179−801<x−801<81−801

−1.00<z<1.00−1.00<z<1.00

The area under the standard normal curve between z=−1z=−1 and z=+1z=+1 corresponds with a probability of 0.68260.6826 (calculated from areas found in Table A: 0.8413–0.1587=0.68260.8413–0.1587=0.6826). As such, the probability that the sample mean will be between 7979 and 8181 is 68.26%68.26%.

**Test Yourself 1.10**

Suppose we have selected a random sample of n=100n=100 from a population with a mean of 8080 and a standard deviation of 66. Find the probability that the sample mean will be between 7979 and 8181.

Show Answer

As demonstrated from these examples, as nn increases, the precision of the sample mean increases. The Central Limit Theorem tells us that variability of the sample mean decreases as the sample size increases. As a result, the sample mean becomes more and more accurate in estimating the sample mean as nn gets larger.

Based on these properties, we can estimate the population mean from the sample mean of a single sample.

**Summary/Lead in for next Section**

This module focused on introducing the science of statistics. The remaining modules in the course will build upon the concepts described in this section. Graphical and numerical summaries help us to describe, understand, and visualize data in a convenient and meaningful way. Concepts of inferential statistics were also introduced. Using data collected in a sample to make estimates and conclusion about a larger population is key to the science of statistics. The Central Limit Theorem is a powerful result which allows us to use the normal distribution to construct confidence intervals and perform statistical testing relating to the population mean using the sample mean.

In the next section, we use the foundation built upon in this section as we learn about statistical inference where we will using data collected in a sample to make estimates, decisions, predictions, and other generalizations about a larger set of data. Important concepts include confidence intervals and hypothesis testing which will be critical throughout the remaining modules. We will also review in the next module the most commonly used statistical procedures for quantitative variables. That is, we will explore procedures about a population mean from a single group as well as procedures that compare means between two groups. The remaining modules will focus on inference for relationships between variables. We will return to qualitative variables in the last module where we will look at procedures about a population proportion from a single group as well as procedures that compare proportions between two groups before exploring procedures that focus on inference for relationships between a qualitative variable and other variables (regardless of type).

**Module 1 R Code – Introduction to R Programing**

**Installing R**

* Install R Software on Your Laptop
* Go to R main website <https://cran.r-project.org/> and download R based on your operating system.
* Install of R on Windows Operating System.
  + **Installation on Windows OS** Step by Step installation Video <https://www.youtube.com/watch?v=mfGFv-iB724>
  + **Installation on MacOS** Step by Step installation Video <https://www.youtube.com/watch?v=uxuuWXU-7UQ>

To be able to work much easier, we recommand to use a graphical user interface — Integrated Development Environment (IDE) for R Programming.

**RStudio** (IDE Recommended )

RStudio <https://www.rstudio.com/> is a free and open-source Integrated Development Environment (IDE) for R Programming.

How to install RStudio. Step by step video <https://www.youtube.com/watch?v=cX532N_XLIs>

**Introduction to R**

In this mini tutorial we will review some basic concepts of R programing.

Run R and check the following commands:

# R includes numeric types: interger, double

> 348

# Characters

> "my string"

# logical

> TRUE

> FALSE

# Arithmetic operators as you'd expect

> 42 + 1 \* 2^4

# We have logical operators/comparison

TRUE | FALSE

> 1 + 7 != 7

# Other logical operators:

# &, |, !

# <,>,<=,>=, ==, !=

**Variable Assignment**

# Variables assignment is done with the <- operator

> mynumber <- 483

# typeof() tells use type

> typeof(mynumber)

[1] "double"

# we can convert between types

myint <- as.integer(mynumber)

> typeof(myint)

[1] "integer"

**Vector of Data**

# The vector is the most important data structure

# create it with c() - named combiner function

my.vec <- c(1, 2, 67, -8)

# get some properties

str(my.vec)

##

num [1:4] 1 2 67 -8

> length(my.vec)

[1] 4

# access elements with []

> my.vec[3]

[1] 67

# We can use another vector to index , like following command.

> my.vec[c(3,4)]

# can do assignment too

my.vec[5] <- 41.2

# Working directory - Setting/Getting}

# It is the default location of all input and output files

# List all the objects in the current workspace

> getwd()

# Set working directory

> setwd("/YOUR-HOME-FOLDER/YOURFOLDER")

Remember to use double backslashes or use a single forward slash "/"

# List all the objects in the current workspace

# For example like following settings in Windows.

> setwd("C:/Users/xyz/Documents/work/R")

\greenb{You can use the RStudio menus to set your working directory.}

Reading Data into R

Read a Comma-Separated Values (CSV) data file from a text file

> read.csv("filename")

First Line is the header, default value for header is True

> read.csv("filename", header=True)

It reads a \blueb{Dataframe} into R. \red{Datatrame is an important data type in R.}

**Dataframes in R**

> read.csv("filename", header=True)

# It data type similar to an Excel Sheet or a Database Table like:

# Like following data

"age","job","marital","education","balance","housing","loan","contact"

30,"unemployed","married","primary",1787,"no","no","cellular"

33,"services","married","secondary",4789,"yes","yes","cellular"

35,"management","single","tertiary",1350,"yes","no","cellular"

30,"management","married","tertiary",1476,"yes","yes","unknown"

**Installing Library Packages in R**

# Install a package (only need to do it once)

> install.packages("package name")

# It will recognize dependencies between packages and install required sub packages

# Access the package and loading it into memory

> library("package name")

# View a list of installed packages

> library()

**R Session Commands**

Session Commands

> q() # end R session

Save workshpace image? [y/n/c]:

# y - yes

# n – no

# c – cancel

# Save content of the current workspace into .Rdata file

> save.image()

> save.image(file = "abc.Rdata")

# Save some objects of the current workspace into the file

> save.image(a, b, file = "abc.Rdata")

**Load Stored Objects**

# You can load a set of objects in R from a Rdata binary file.

> load("abc.Rdata")

# List all the objects in the current workspace

> ls()

OR

> objects()

# Remove objects from the current workspace

> rm(a, b)

# delete a file

> unlink("myFile.Rdata")

**Learning R in R: swirl**

You can learn R in R. Step by Step Tutorial: the package swirl <http://swirlstats.com/students.html> can be used to learn R.

> install.packages("swirl")

> library("swirl")

> swirl()

**Module 1 R Code – R Functions for Normal Distribution**

**pnorm()** – computes the probability} that a normally distributed random number will be less than the given number.

This function is also called the "Cumulative Distribution Function" (CDF).

Calculate the area to the left of z: P(Z <= z)

pnorm(z)

Non-standardized normal distribution

pnorm(x, mean=a, sd=b) # calculate the area to the left of x

Find the area under the standard curve  
z = 1.53

pnorm(1.53)

0.9369916

**Standardized Normal Distribution – An Example**

Let’s assume that the birth weights of newborns are normally distributed with a mean of 3500g with a standard deviation of 500g.

What proportion of infants weigh less than 2800g?

The variable x, the birth weight, has the N(3500,500) distribution.

> pnorm(2800, mean=3500, sd=500)

0.08075666

What proportion of infants weigh between 3250g and 3750g?

> pnorm(3750, mean=3500, sd=500)-pnorm(3250, mean=3500, sd=500)

0.3829249

**R Functions on Normal Distribution – Normal Distribution Probability Density Function (PDF)**

**dnorm()** – Given a set of values it returns the height of the probability distribution at each point. If you only give the points it assumes you want to use a mean of 0 and standard deviation of 1.

density\_standard\_norm <- function(x){1/sqrt(2\*pi)\*exp(-0.5\*x^2)}

> dnorm(0)

0.3989423

> dnorm(0, mean=4, sd=10)

0.03682701

> pnorm(0)

0.5

In statistics, quantiles are cut points dividing the range of a probability distribution into contiguous intervals with equal probabilities.

Given a probability p and a distribution, we want to calculate the corresponding quantile for p: the value x such that P(X <= x) = p

# For non-standardized normal distribution

> qnorm(x)

# For any normal distribution with mean and sd

> qnorm(x, mean=a, sd=b)

**R Function on Normal Distribution**

Density, distribution function, quantile function and random generation for the normal distribution with mean equal to mean and standard deviation equal to sd.

Usage:

dnorm(x, mean = 0, sd = 1, log = FALSE)

pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)

qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)

rnorm(n, mean = 0, sd = 1)

**Summary**

**pnorm** – calculates the probabilty value given a critical value

**qnorm** – is invese computation and calculate a critical value given a probabilty value

**dnorm** – calculates density value for a given critical value (you can understand this as y value for a given z or x critical point)

**rnorm** – generates random samples from a normal distribution

For more information read the manuals.

To read the manual, use one of the following commands:

?pnorm

or

help(pnorm)

**Data Summarization**

In R you can simply pass your data object to summary function.

* summary() function is a very general function and we use it in R a lot to pass data and model object to it.
* summary() is a intelligent function and can return well-formatted prints depending on object types.

> earning <- c(35, 40, 145, 33, 30, 42, 32, 32, 25)

> summary(earning)

Min. 1st Qu. Median Mean 3rd Qu. Max.

25 32 33 46 40 145

You can calculate the results from the above summary() using the following separate functions:

> mean(earning)

> median(earning)

> min(earning)

> max(earning)

> quantile(earning)

You can calculate the Variance and standard deviation using following functions:

> var(earning)

> sd(earning)

**Data Summarization and Visualiztion – R Functions – Graphical data summaries**

**Histograms**

> hist(data$variable)

> hist(data$variable, bins) # specify the number of bins

> hist(data$variable, breaks=c(x,y,z..)) # specify cutpoints

> hist(data$variable, breaks=seq(a,b,by=c)) # specify cutpoints

data$variable is a vector of data that we read for example from a dataframe "data" and access specific variable of it using a dollar sign.

**Boxplot**

We can create a boxplot using the **boxplot()** function.

boxplot(earning)

boxplot() function in R can detect outliers and visualize it.

Read the documentations of boxplot function to find out how you can turn this automated process ON or OFF.

**Formatting**

You can make your graphs look much better and exactly what you want to have:

* **Labeling:** Each graph in R can have a title a X and Y-Axis label. You can use the following lables, main, xlab and ylab.
* Title: main="Histogram of xyz"
* X-axis label: xlab="Nile flow"

Y-axis lable: ylab = "Frequency"

* **Colors:** You can fund [a list of collor numbers](http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf).
* **Controlling the window:** you can use the following attributes xlim and ylim to have a limitation for x and y axis.
* X-axis: xlim=c(min, max)

Y-axis: ylim=c(min, max)

**Combine multiple plots into one overall graph**

Creating graphs side-by-side. Before you create your plot you need to set an enviroment variable that controls the positioning of plots.

You can use the **par()** function and use the **mfrow** attribute

par(mfrow=c(2,2)) # 2 by 2 panels

par(mfrow=c(1,1)) # Go back to single graph mode

**R Functions – Qualitative Data Summaries**

* Numerical summary:
  + Class Frequencies

> table(data$variable)

Or

> summary(data$variable)

* + Relative Class Frequencies: Divide class frequencies by number of rows in the dataset using nrow(data).
* Graphical summary: You can create a Pie chart or a barplot.
* > Pie(table(data$variable))

> Barplot(table(data$variable))

**Module 1 R Code – Examples**