You might be asking yourself “Why are we studying probability and statistics?” Well, one important reason is that probability and statistics are covered on the FE exam. That alone should be enough to get your attention. Also, the purpose of probability and statistics is to deal with uncertainty.

What’s uncertain in the field of civil engineering? Everything!

- The future. (How many vehicles will use this highway over the next 20 years?)
- Human behavior. (What is the water demand in this neighborhood between 6 am and 8 am?)
- Material behavior. (How strong is this batch of steel compared to the last batch?)
- Nature. (What kind of wind loads will this billboard have to withstand?)

Statistics is concerned with (a) collecting data (i.e., observing the phenomenon of interest), (b) organizing and summarizing this data, and (c) analyzing the data and (d) drawing conclusions from it.

The branch of statistics that deals with organizing and summarizing data is called descriptive statistics and the branch of that deals with drawing conclusions from the data is called inferential statistics.

Let’s look at the speed of vehicles driving on Central Avenue as an example. The population might consist of the speeds of all vehicles driving down Central Avenue this year (about seven million cars). One way to summarize all of those speeds is to compute the average. That’s the population parameter of interest..

It would be prohibitively expensive to sit on the sidewalk with a radar gun and measure the speed of all seven million cars just to compute their average! Instead, we collect a random sample of vehicle speeds (perhaps by going out at random times on random days and measuring the speeds of the first 10 cars we see). Then we compute the average value of those observations. That’s the sample statistic we’re going to use in order to infer the population average.
FREQUENCY DISTRIBUTIONS

A frequency distribution is a tabular summary of sample data organized into categories or classes. Suppose we have collected a sample of 50 speed measurements on Central Avenue. (For the sake of brevity, I won’t list them). One way to summarize this data is to organize it into 5-mph speed ranges (or class intervals) and determine the number of observations within each class (which we call the class frequencies):

<table>
<thead>
<tr>
<th>Class Intervals</th>
<th>Class Frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>40–45</td>
<td>1</td>
</tr>
<tr>
<td>45–50</td>
<td>9</td>
</tr>
<tr>
<td>50–55</td>
<td>15</td>
</tr>
<tr>
<td>55–60</td>
<td>10</td>
</tr>
<tr>
<td>60–65</td>
<td>7</td>
</tr>
<tr>
<td>65–70</td>
<td>5</td>
</tr>
<tr>
<td>70–75</td>
<td>3</td>
</tr>
<tr>
<td>TOTAL</td>
<td>50</td>
</tr>
</tbody>
</table>

This might be easier to grasp if we plot these class frequencies as a bar chart, with each bar representing one of the classes. This graphical representation of a frequency distribution is called a histogram:
If an observation happens to fall exactly on a class boundary (e.g., a speed of 55.0 mph), then it is usually placed in the lower of the two classes. In other words, each class includes those observations who value is greater than the lower bound and less than or equal to the upper bound of the class.

Sometimes, the horizontal axis of the histogram is labeled to show the range of values in each class:

![Histogram with class boundaries](image)

and sometimes it is labeled to show the mid-point of each class (which is sometimes called the class mark):

![Histogram with class marks](image)

Regardless, they are all histograms. The important thing is that the area of each column is proportional to the frequency of the corresponding class.
A relative frequency distribution is a more useful way of summarizing data. Here, each class frequency is divided by the total number of observations (i.e., the sample size) and expressed as a percentage of the total sample. That way, it doesn’t matter what the sample size is. We can directly compare the distributions from samples of various sizes.

### Vehicle Speeds on Poplar Avenue

<table>
<thead>
<tr>
<th>Speeds (kph)</th>
<th>Vehicles Counted</th>
<th>Percentage of Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>40–45</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>45–50</td>
<td>9</td>
<td>18</td>
</tr>
<tr>
<td>50–55</td>
<td>15</td>
<td>30</td>
</tr>
<tr>
<td>55–60</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>60–65</td>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td>65–70</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>70–75</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>TOTAL</td>
<td>50</td>
<td>100</td>
</tr>
</tbody>
</table>

You can also plot a bar chart of the relative frequencies. This is called a relative frequency histogram:
CUMULATIVE FREQUENCY DISTRIBUTIONS

A cumulative frequency distribution is an even more useful way of summarizing data in which the class frequencies in the relative frequency distribution are replaced by running totals of the class frequencies:

<table>
<thead>
<tr>
<th>Speeds (mph)</th>
<th>Vehicles Counted</th>
<th>Percentage of Sample</th>
<th>Cumulative Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>40–45</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>45–50</td>
<td>9</td>
<td>18</td>
<td>20</td>
</tr>
<tr>
<td>50–55</td>
<td>15</td>
<td>30</td>
<td>50</td>
</tr>
<tr>
<td>55–60</td>
<td>10</td>
<td>20</td>
<td>70</td>
</tr>
<tr>
<td>60–65</td>
<td>7</td>
<td>14</td>
<td>84</td>
</tr>
<tr>
<td>65–70</td>
<td>5</td>
<td>10</td>
<td>94</td>
</tr>
<tr>
<td>70–75</td>
<td>3</td>
<td>6</td>
<td>100</td>
</tr>
<tr>
<td>TOTAL</td>
<td>50</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

You can plot a bar chart of the cumulative frequencies, too. This is called a cumulative frequency diagram:

Note that the height of each column is proportional to the percentage of the sample having values less than or equal to the upper bound of the class. The class containing the sample maximum—and all classes to its right—have a height of 100% (because the entire sample is less than or equal to the maximum value).
Cumulative frequency distributions can also be plotted as a line plot rather than a bar chart. Here, each point on the plot represents the percentage of the sample with values less than or equal to the abscissa value.

**HOW MANY CLASSES IS ENOUGH?**

If you have too few classes, you can’t tell anything about the distribution of the data because everything gets lumped into a couple of classes. If you have too many classes, you also can’t tell anything about the data because each class has only one or two members (or perhaps no members at all). A good rule of thumb is that the number of classes should be approximately equal to the square root of the number of observations.

**SUMMARY STATISTICS**

Frequency distributions and histograms are invaluable tools for summarizing a set of observations, but it is sometimes necessary to use numerical values to describe the data. These numerical descriptions are called statistics. The most basic descriptive statistics tell you what a “typical” value is for the sample (these are called measures of central tendency) and how widely the individual observations vary from that typical value (these are called measures of dispersion).

**MEASURES OF CENTRAL TENDENCY**

Most data tend to cluster about a central point. One very simple way of summarizing a data sample is to state that central value. Such a value is called a measure of central tendency. The most common measures are as follows:

- **Arithmetic Mean** – the average of all the observations
- **Median** – the middle value in the sample (half the sample greater, half the sample less)
- **Mode** – the data value that occurs most frequently
The various measures of central tendency are best explained with an example. Consider the following set of nine observations (of who-knows-what):

\[8 \ 6 \ 4 \ 10 \ 3 \ 8 \ 4 \ 8 \ 5\]

The **sample mean** is given by:

\[
\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{8 + 6 + 4 + 10 + 3 + 8 + 4 + 8 + 5}{9} = 6.22
\]

The symbol \( \bar{x} \) is commonly used for the sample mean and the symbol \( \mu \) is used for the population mean. Greek letters always represent *population parameters* while Latin letters represent *sample statistics*.

The **median** of the sample is the value that divides the sample exactly in half. Rearranging the data in ascending order, we get:

\[3 \ 4 \ 4 \ 5 \ 6 \ 8 \ 8 \ 8 \ 10\]

The median value is obviously six because there are four observations with values less than six and four with values greater than six.

What happens when there are an *even* number of points? Then you would use the average of the two middle values, instead.

The **mode** of the sample is the value that occurs most frequently. For this sample, the mode is eight because there are more occurrences of that value than any others:

\[3 \ 4 \ 4 \ 5 \ 6 \ 8 \ 8 \ 8 \ 10\]

It is, of course, possible for a sample to have more than one mode. For example, if there were three 4’s as well as three 8’s, then the sample would be described as *bimodal*:

\[3 \ 4 \ 4 \ 4 \ 6 \ 8 \ 8 \ 8 \ 10\]

A sample could even have no mode (which happens when each value occurs the same number of times).

If the frequency distribution is symmetrical, the mean, median, and mode will be nearly identical, in which case the mean is the most commonly used measure of central tendency. In cases where the distribution is highly asymmetrical, then the median can be more useful than the mean. For example, economists often talk about **median income**. Half the population makes less than this amount and the other half makes more than this amount. Most of the population has modest levels of income. In 1999, for example, the middle 60% of households made between $16,000 and $82,000. But a select few make such an extraordinary amount of money (millions or tens of millions of dollars) that the average income is unrepresentative. The median value (about $38,000 in 1999) is much more representative of the population as a whole.
MEASURES OF DISPERSION

Even though most data tend to cluster about a central point, there is almost always some variation from one observation to the next. From the standpoint of describing the data set, it would be useful to know if that variability is small or large. A value that describes the amount of variation or dispersion in the data is called a measure of dispersion. Measures of dispersion are based on the deviation of each observation from the sample mean. The most common measure of dispersion is the sample variance:

\[ s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1} \]

The sample variance is really just an average of all the squared deviations. Note, however, that the sum is divided by \( n - 1 \) rather than just \( n \). The denominator is called the degrees of freedom. The sample mean has \( n \) degrees of freedom, which means you need to know all \( n \) data values \( (x_i) \) before you can calculate \( \bar{x} \). On the other hand, you only need to know \( n - 1 \) of the deviations to calculate \( s^2 \). Why? The sum of the \( n \) deviations about the mean is always zero, so you can calculate a missing deviation by subtracting the \( n - 1 \) known deviations from zero!

For a quick “proof” we can start with the definition of the sample mean:

\[ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \Rightarrow n\bar{x} = \sum_{i=1}^{n} x_i \Rightarrow 0 = \sum_{i=1}^{n} x_i - n\bar{x} \]

On the right-hand side, there are \( n \) values of \( x_i \) inside the summation and \( n \) values of \( \bar{x} \) outside. If we bring the \( n \) values of \( \bar{x} \) inside the summation, we get the following:

\[ 0 = \sum_{i=1}^{n} (x_i - \bar{x}) \]

which simply says that the sum of the \( n \) deviations from the mean is equal to zero!

The variance is hard to conceptualize because it has squared units (such as \( \text{ft}^2 \) or \( \text{sec}^2 \) or \( \text{lb}^2 \)). The amount of dispersion would be easier to understand if it had the same units as the data. The sample standard deviation is the square root of sample variance:

\[ s = \sqrt{s^2} \]

Because they pertain to the sample, \( s \) and \( s^2 \) are statistics. The corresponding population parameters are the population variance \((\sigma^2)\) and the population standard deviation \((\sigma)\). We use the sample statistics as estimates of the (usually unknown) population parameters.
ILLUSTRATION

Suppose we are given the following 10 speed measurements (in kph):

\[
\begin{array}{ccccccc}
61 & 75 & 43 & 61 & 54 \\
60 & 60 & 58 & 61 & 65 \\
\end{array}
\]

To calculate the sample mean, we start by summing the speeds:

\[\sum x_i = 598 \text{ kph}\]

then divide by the number of observations:

\[\bar{x} = \frac{598}{10} = 59.8 \text{ kph}\]

To calculate the sample variance, we could calculate the individual deviations from this mean, then sum the squared deviations, but there is an easier way to calculate the sample variance:

\[s^2 = \frac{\sum (x_i)^2 - n(\bar{x})^2}{n-1}\]

We already know the mean speed \(\bar{x}\), so all we have left to do is calculate the sum of the squared speeds:

\[\sum (x_i)^2 = 61^2 + 75^2 + 43^2 + 61^2 + 54^2 + 60^2 + 60^2 + 58^2 + 61^2 + 65^2 = 36,342 \text{ kph}^2\]

and substitute into the equation for the sample variance:

\[s^2 = \frac{36,342 - 10(59.8)^2}{10-1} = 64.62 \text{ kph}^2\]

The square root of this number is the standard deviation:

\[s = \sqrt{64.62} = 8.04 \text{ kph}\]

One last measure of dispersion is the coefficient of variation (COV), which is just the standard deviation expressed as a percentage of the mean:

\[COV = \frac{s}{\bar{x}} \times 100\% = \frac{8.04 \text{ kph}}{59.8 \text{ kph}} \times 100\% = 13.44\%\]

This is a good way to compare measures of dispersion between different samples whose values don’t necessarily have the same magnitude (or, for that matter, the same units!).

11
DESCRIPTIVE STATISTICS IN EXCEL

EXCEL provides multiple ways of calculating the descriptive statistics covered in this chapter:

1. You can use the Function Wizard (the \( f_x \) button) to select from any of a number of statistical functions, including MEDIAN, MODE, AVERAGE (mean), VAR (variance), and STDEV (standard deviation):

   ![Function Wizard](image)

   Note that there are also functions available called VARP and STDEVP. These give you the population variance and the population standard deviation and should only be used when your sample is actually the entire population, rather than a subset of the population. Mathematically, the difference is that the population variance has \( n \) degrees of freedom while the sample variance has \( (n – 1) \) degrees of freedom. As a rule, you want to stay away from VARP and STDEVP.

2. Alternatively, you can get all of your descriptive statistics (including some we haven’t even talked about) in a single operation. To do this, go to the Tools menu, select Data Analysis (which should be at the bottom of the menu), then select the Descriptive Statistics tool from the list and click OK.

   **NOTE:** If you can’t find Data Analysis on the Tools menu, it means EXCEL’s Analysis ToolPak hasn’t been loaded yet. Normal EXCEL installations don’t include the ToolPak because very few people have a clue what it is or how to use it. (See what elite company you’re already keeping?) It’s actually on your hard drive, but it has to be added to the menu. To do that, go back to the Tools menu, select Add-ins, put a check in the box marked Analysis ToolPak, the click OK.

When the Descriptive Statistics dialog box appears, you’ll find it divided into two sections. The upper half is where you tell EXCEL where to find the data. The lower half is where you tell EXCEL what type of statistics you want (in this case, Summary Statistics) and where you want to put them:
Excel will then produce a report containing the values of the Mean, Median, Mode, Standard Deviation, Variance, Minimum, Maximum, Range, Sum, and Count of the data (plus the Kurtosis and Skewness, which we won’t go into here). Just remember that this tool produces a report, which means it’s just a bunch of text on a page. If, for some reason, the original data in the Input Range changes, the report does not get updated.

3. If you want to calculate a frequency distribution, you can use the **FREQUENCY** function, which has two arguments:

\[
\text{FREQUENCY(data\_array, bins\_array)}
\]

The **data\_array** argument identifies the data whose frequency distribution you seek (e.g., the 50 vehicle speeds referred to earlier in the chapter). These data are specified as a horizontal, vertical, or rectangular array of cells in the spreadsheet.

The **bins\_array** argument identifies the “bins” or “classes” into which you want to sort the data (e.g., the 7 speed ranges referred to earlier in the chapter). These bins are specified as a **vertical array** of cells containing the **upper bound** of each bin. (Don’t try to use a horizontal array of cells, it won’t work).

The **FREQUENCY** function is an **array function**, which means it returns an array of values (the counts associated with each of the specified bins) rather than a single value. That means you have to simultaneously enter the function into an array of cells rather than a single cell. To do this, select the appropriate **vertical array** of cells (usually adjacent to the **bins\_array** cells so you know which count goes with which bin), type the function into the formula bar at the top of the spreadsheet, then simultaneously hit [Ctrl], [Shift], and [Enter] to enter the formula into all the cells at the same time.
4. If you want to create a frequency distribution and plot a histogram, go to the Tools menu, select Data Analysis, then select the Histogram tool from the list and click OK.

As with the Descriptive Statistics tool, the upper half of the dialog box is where you tell EXCEL the location of the cells containing the data (referred to as the “Input Range”) and the location of the cells containing the upper bounds of the classes (referred to as the “Bin Range”).

The bottom half is where you tell EXCEL what information to include in the output and where you want the output to go. If you don’t check anything, you’ll just get a table containing the frequency distribution. If you check the Cumulative Percentage box, EXCEL will also include cumulative frequencies in the table. If you check the Chart Output box, EXCEL will also output a histogram and, if Cumulative Percentage was also checked, a cumulative frequency plot.