Descriptive Statistics

**Describing your Data**

Generally one of the first things to do with new data is to get to know it by asking some general questions like but not limited to the following:

* What variables are included? What information are we getting?
* What is the format of the variables: string, numeric, etc.?
* What type of variables: categorical, continuous, and discrete?
* Is this sample or population data?

After looking at the data you may want to know

* How many males/females?
* What is the average age?
* How many undergraduate/graduates students?
* What is the average SAT score? It is the same for graduates and undergraduates?
* Who reads the newspaper more frequently: men or women?

You can start answering some of these questions by looking directly at the table, for some other questions you may have to do some calculations by obtaining a set of descriptive statistics. These statistics are a collection of measurements of two things: location and variability.

* *Location* tells you the central value (the mean is the most common measure of this) of your variables.
* *Variability* refers to the spread of the data from the center value (i.e. variance, standard deviation).
* *Statistics* is basically the study of what causes such variability.

|  |  |
| --- | --- |
| **Location** | **Variability** |
| Mean | Variance |
| Mode | Standard deviation |
| Median | Range |

**Use the Descriptive Statistics data file.**

* Click on DATA, then Data Analysis, then Descriptive Statistics.
* For input range, select ALL data.
* You should see the screen below.



* Since we include the labels in first row make sure to check that option.
* For the output option which is the place where excel will enter the results select O1 or you can select a new worksheet or even new workbook.
* Check “Summary statistics” and the press OK.
* You will get the following:



* While the whole descriptive statistics cells are selected go to Format – Cells to change all numbers to have one decimal point. When you get the ‘format cells’ window, select the following:



* Click OK. All numbers should now have one decimal as follows:



Now we know something about our data.

* The average student in this sample is 25.2 years, has a SAT score of 1848.9, got a grade of 80.4, is 66.4 inches tall and reads the newspaper 4.9 times a week.
* We know this by looking at the ‘mean’ value on each variable.

**Mean**

* is the sum of the observations divided by the total number of observations.
* It is the most common indicator of central tendency of a variable.
* If you look at the last two rows: “Sum” and “Count” you can estimate the mean dividing “Sum” by ”Count” (sum/count).
* You can also calculate the mean using the function below (**IMPORTANT**: All functions start with the equal “=” sign):

=AVERAGE(range of cells with the values of interest)

**Sum**

* refers to the sum of all the values in a range of values.
* The excel function for sum is:

=SUM(range of cells with the values of interest)

**Age**

* means the sum of the ages of all students.

=AVERAGE(J2:J31)

**Count**

* refers to the count of cell that contain values (numbers).
* The function is:

=COUNT(range of cells with the values of interest)

**Min**

* is the lowest value in an array of values.
* The function is:

=MIN(range of cells with the values of interest)

**Max**

* is the largest value in an array of values.
* The function is:

=MAX(range of cells with the values of interest)

**Standard Error (SE)**

* indicates how close the sample mean is from the ‘true’ population mean.
* The average age of 25.2 years is just an estimate of this sample of students but it can vary had you used a different set of students.
* The standard error is calculated by *dividing the standard deviation of the population (or the sample) by the square root of the total number of observations.*
* The SE can be used to roughly define a range of certainty for the mean. Using “age”:

|  |  |  |  |
| --- | --- | --- | --- |
| Z | % Certainty | Lower bound | Upper bound |
| 1 (0.99) | 68% | 23.9 | 26.5 |
| 2 (1.96) | 95% | 22.7 | 27.7 |
| 3 (2.58) | 99% | 21.4 | 29.0 |

Lower:  Mean – (SE\*Z) for example 25.2 – (1.3 \* 2) = 22.7

Upper:  Mean + (SE\*Z) for example 25.2 + (1.3 \* 2) = 27.7

* You are 68% certain that the average age is between 23.9 and 26.5 years old
* You are 95% certain that the average age is between 22.7 and 27.7 years old
* You are 99% certain that the average age is between 21.4 and 29.0 years old
* Note that the more certainty wider the gap.

**Median**

* is another measure of central tendency.
* To get the median you have to order the data from lowest to highest. The median is the number in the middle.
* If the number of cases is odd the median is the single value, for an even number of cases the median is the average of the two numbers in the middle.
* The excel function is:

=MEDIAN(range of cells with the values of interest)

**Mode**

* refers to the most frequent, repeated or common number in the data.
* By age there are more students 19 years old in the sample than any other group.
* In the SAT scores the mode is “#N/A” which means that all values are unique.
* The excel function is:

=MODE(range of cells with the values of interest)

**Range**

* is a measure of dispersion.
* It is simple the difference between the largest and smallest value, “max” – “min”.

**Sample Variance**

* measures the dispersion of the data from the mean.
* It is the simple mean of the squared distance from the mean.
* It is calculated by:

SV = sum of (X-mean of X)2 / Number of observation minus 1

* *Higher variance* means more dispersion from the mean.
* The excel function is:

=VAR(range of cells with the values of interest)

**Standard Deviation**

* is the squared root of the variance.
* Indicates how close the data is to the mean.
* Assuming a normal distribution, 68% of the values are within 1 sd from the mean, 95% within 2 sd and 99% within 3 sd.
* The excel formula is:

=STDEV(range of cells with the values of interest)

**Skewness**

* measures the asymmetry of the data, when in an otherwise normal curve one of the tails is longer than the other.
* It is a roughly test for normality in the data (by dividing it by the SE).
* If it is positive there is more data on the left side of the curve (right skewed, the median and the mode are lower than the mean).
* A negative value indicates that the mass of the data is concentrated on the right of the curve (left tail is longer, left skewed, the median and the mode are higher than the mean).
* A normal distribution has a skew of 0.
* Skewness can also be estimated with the following function:

=SKEW(range of cells with the values of interest)

**Kurtosis**

* measures the peak of the distribution.
* It is also an indicator of normality.
* *Positive kurtosis* indicates too few cases in the tails or a tall distribution (leptokurtic), *negative kurtosis* too many cases in the tails or a flat distribution (platykurtic).
* A normal distribution has a kurtosis of 0 (given a correction of –3, otherwise it will have a kurtosis of 3).
* The excel function for kurtosis is:

=KURT(range of cells with the values of interest)

**Quartiles**

During a period of one month, a random sample of approved insurance policies were selected and listed below:

1. Computer the mean, median, first, and third quartile
2. Compute the range, interquartile range, standard deviation.
3. What would you tell a customer who enters the bank to purchase this type of policy and asks how long the approval process is?

