Descriptive Statistics in Excel

Descriptive statistics is a way to describe data using measures of central tendencies and variation (how spread out the data is). These allow us to make evidence-based decisions.

Basic Descriptors

Count	The total number of data points in the set		=COUNT()
Minimum	Smallest value in the data set		=MIN()
Maximum	Largest value in the data set		=MAX()
Range	Different between the maximum and minimum values (range = max – min) *helpful in identifying outliers		=MAX()-MIN()

Measures of Central Tendency

Mean (average)	Sum of all the data points divided by the total number of data points (sensitive to outliers*)	$\overline{X} = \frac{\sum X}{N}$	=AVERAGE()
Median	Middle number of the data set (when data is listed from smallest to largest)		=MEDIAN()
Mode	Most frequently occurring number	If you expect one mode: If you expect multiple modes:	=MODE() =MODE.SNGL() =MODE.MULT()

Data Spread

Standard Deviation	The measure of the amount of variation in a data set (square root of the variance)**	$\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{N-1}}$	=STDEV.S()
Variance	Average of the squared difference of each data point from the mean (standard deviation squared)	$\frac{(X-\mu)^2}{N}$	=VAR.S()

*An outlier is a data point that is far away from the rest of the data

**A low standard deviation suggests that the data points are close to the mean. A high standard deviation suggests that the data points are more spread out

Outliers vs Errors

Outliers and errors can both impact the descriptive statistics of a data set.

Outlier – a data point far away from the rest of the data. The reason for outliers depends on the context of the data. For example, did a machine run slower for some reason?

Error – a mistake that occurred when data is collected. For example, did a number get entered wrong? Did a decimal get displaced? 1.23 can easily be entered as 12.3.

Take a look at the data sets below. There are two values that could be outliers or errors.

Student	Grade
1	59
2	40
3	67
4	89
5	94
6	99
7	12
8	97
9	88
10	84
11	86
12	79
13	58
14	92
15	880
16	64
17	74
18	41
19	63
20	49
21	69
22	90
23	76
24	87

Before doing any calculations or plots, a decision must be made about what to do with the errors or outliers.

Decision #1: delete the values from the data set, then perform the calculations.

You should never just delete data from a data set without a thorough investigation.

Decision #2: Leave the values in the data set and just proceed with the calculations.

Because the values are so different from the others and could have an impact on the descriptive statistics, this is not the best option.

Decision #3: Find the source of the data set to get more information on the possible reasons for the values.

This is the best option to be able to make an informed decision.

In this case, student 7's grade of 12% was correctly entered into the data set. It would be considered an outlier.

However, student 15 did not earn 880%. This data point doesn't make sense in the context of a grade as there's usually a maximum of 100%. This would be an error. The grade was supposed to be entered as 80%