L4: Measures of variation

Measures of variation used mostly for quantitative variables.

1-Range

The sample range of the variable is the difference between its maximum and minimum values in a data set:

$$\text{Range} = \text{Max} - \text{Min}.$$

There are several characteristics of range:

1- The sample range of the variable is quite easy to compute.

2- In using the range, a great deal of information ignored, that is, only the largest and smallest values of the variable are considered; the other observed values are disregarded.

3- The range cannot ever decrease, but can increase, when additional observations are included in the data set and that in sense the range is overly sensitive to the sample size.

Example 5.3. Prices of hotdogs ($/oz.):

0.11, 0.17, 0.11, 0.15, 0.10, 0.11, 0.21, 0.20, 0.14, 0.14, 0.23, 0.25, 0.07, 0.09, 0.10, 0.10, 0.19, 0.11, 0.19, 0.17, 0.12, 0.12, 0.12, 0.10, 0.11, 0.13, 0.10, 0.09, 0.11, 0.15, 0.13, 0.10, 0.18, 0.09, 0.07, 0.08, 0.06, 0.08, 0.05, 0.07, 0.08, 0.08, 0.07, 0.09, 0.06, 0.07, 0.08, 0.07, 0.07, 0.07, 0.08, 0.06, 0.07, 0.06

The range in SPSS:

Analyze -> Descriptive Statistics -> Frequencies,

Analyze -> Descriptive Statistics -> Descriptives

Range of the prices of hotdogs

	N	Range	Minimum	Maximum
Price ($/oz)	54	.20	.05	.25
Valid N (listwise)	54			

Example 5.2. 7 participants in bike race had the following finishing times in minutes: 28, 22, 26, 29, 21, 23, 24, 2, 44, 12, 23.

What is the range?

Example 5.3. 8 participants in bike race had the following finishing times in minutes: 28, 22, 26, 29, 21, 23, 24, 50,30,44, 60

What is the range?

2- Interquartile range

Before we can define the sample interquartile range, we have to first define the percentiles, the deciles and the quartiles of the variable in a data set.

1- **The percentiles** of the variable divide observed values into **hundredths**, or 100 equal parts.

   The first percentile, P1, is the number that divides the bottom 1% of the observed values from the top 99%; **second percentile, P2**, is the number that divides the bottom 2% of the observed values from the top 98%; and so forth. **The median is the 50th percentile**.

2- **The deciles** of the variable divide the observed values into **tenths**, or 10 equal parts. The variable has nine deciles, denoted by D1,D2, . . . ,D9. **The first decile D1 is 10th percentile**, the second decile D2 is the 20th percentile, and so forth.

3- The most commonly used percentiles are **quartiles**. The quartiles of the variable divide the observed values **into quarters**, or **4 equal parts**.

The variable has three quartiles, denoted by Q1,Q2 and Q3: Arrange the observed values of variable in a data in **increasing order**.

1. The first quartile Q1 is at position n+1/4
2. The second quartile Q2 (the median) is at position n+1/ 2

3. The third quartile Q3 is at position 3(n+1)/4 in the ordered list.

4- The sample interquartile range of the variable(IQR) :is the difference between the first and third quartiles of the variable, that is:

$$IQR = Q3 - Q1$$

IQR gives the range of the middle 50% of the observed values.

Example 5.4. 7 participants in bike race had the following finishing times in minutes: 28,22,26,29,21,23,24.

What is the interquartile range?

Example 5.5. 8 participants in bike race had the following finishing times in minutes: 28,22,26,29,21,23,24,50.

What is the interquartile range?

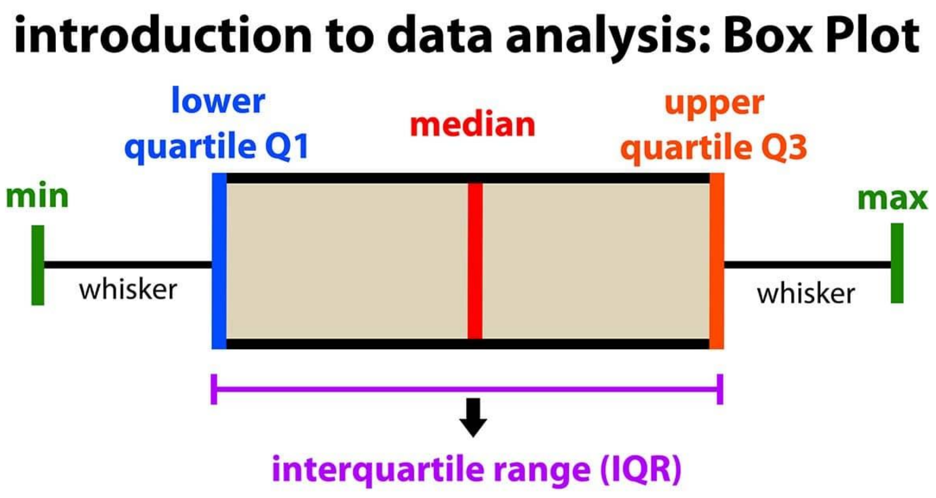
Analyze -> Descriptive Statistics -> Explore

Noticeable: Five-number summary and boxplots

The five-number summary of the variable consists of minimum, maximum, and quartiles written in increasing order:

Min,Q1,Q2,Q3,Max.

A boxplot is based on the five-number summary and can be used to provide a graphical display of the center and variation of the observed values of variable in a data set.



Standard deviation –SD

The sample standard deviation is the most frequently used measure of variability,

although it is not as easily understood as ranges. It can be considered as a kind of average of the absolute deviations of observed values from the mean of the variable in question.

Standard deviation is a statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance.

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}.$$

$$\mathrm{SD} = \sqrt{\frac{\sum |x - \mu|^2}{N}}$$

where:

$\sum$ means "sum of"

x is a value in the data set

μ is the mean of the data set

N is the number of data points in the population.

Since the standard deviation is defined using the sample mean μ of the variable x.

<u>Note</u> :

1- the standard deviation is always positive number, i.e., SD≥ 0.

In a formula of the standard deviation, the sum of the squared deviations from the mean.

2- The formula above is for finding the standard deviation of a population. If you're dealing with a sample, you'll want to use a slightly different formula (below), which uses ***n-1*** instead of N.

$$SD_{sample} = \sqrt{\frac{\sum |x - \overline{x}|^2}{n - 1}}$$

Step 1: Find the **mean**. **Step 1: Finding** μ **in** $\sqrt{\frac{\sum |x - \mu|^2}{N}}$

Step 2: For each data point, find the **square of its distance to the mean**.

Step 2: Finding $|x - \mu|^2$

In this step, we find the distance from each data point to the mean (i.e., the deviations) and square each of those distances.

Step 3: Sum the values from Step 2. **Step 3: Finding** $\sum |x - \mu|^2$

Step 4: Divide by the number of data points. **Step 4: Finding** $\frac{\sum |x - \mu|^2}{N}$

Step 5: Take the **square root**.

Step 5: Finding the standard deviation

$$\sqrt{\frac{\sum |x-\mu|^2}{N}}$$

EX: find the SD of these values (6, 3, 2, 1)

$$\mu = \frac{6+2+3+1}{4} = \frac{12}{4} = 3$$

Step 1: Find the mean μ.

Step 2: Find the square of the distance from each data point to the mean $|x-\mu|^2$.

x	$\lvert x-\mu\rvert^2$
6	$\lvert 6-3\rvert^2 = 3^2 = 9$
2	$\lvert 2-3\rvert^2 = 1^2 = 1$
3	$\lvert 3-3\rvert^2 = 0^2 = 0$
1	$\lvert 1-3\rvert^2 = 2^2 = 4$

Steps 3, 4, and 5:

$$\text{SD} = \sqrt{\frac{\sum |x-\mu|^2}{N}}$$

$$= \sqrt{\frac{9+1+0+4}{4}}$$

$$= \sqrt{\frac{14}{4}}$$ Sum the squares of the distances (Step 3).

$$= \sqrt{3.5}$$ Divide by the number of data points (Step 4).

$$\approx 1.87$$ Take the square root (Step 5).

Q- Find the standard deviation of the data set.

1, 4, 7, 2, 6