SESSION-10: STATA DESCRIPTIVE STATISTICS-CROSS TABLE, MEASURES OF CENTRAL TENDENCY

Course detail: http://julhas.com/jsedutech/stata-level-one.html
Mentor: Julhas Sujan

Recap-Session 9

- Descriptive statistics – One way table

Session-10

- **Cross table**

Gender	Age group	Height (m)	Weight (kg)
Female	Adult	1.4	60
Male	Child	1.2	15
Male	Adult	1.5	85
Female	Adult	1.5	74
Male	Adult	1.3	77
Female	Elderly	1.6	65

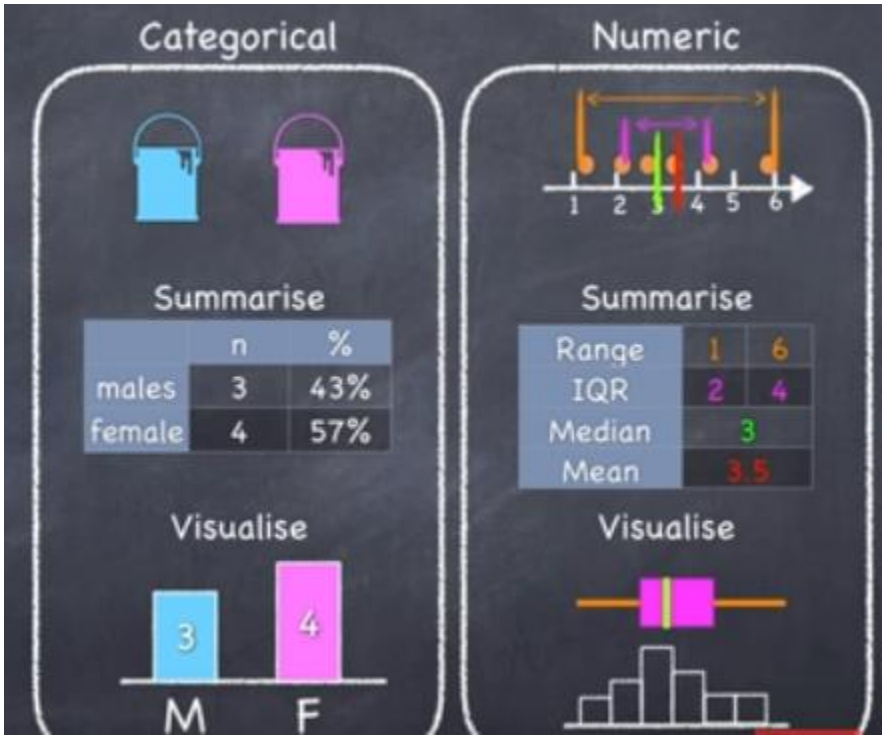
Variables/ Objectes: Gender, Age group, Height, Weight

Categorical variables: Gender, Age group

Numeric variabales: Height, Weight

Presentation of numeric variables:

- Summarize: Range, IQR, Median, Mean (measures of central tendency)
- Visualize: Box plot, Histogram



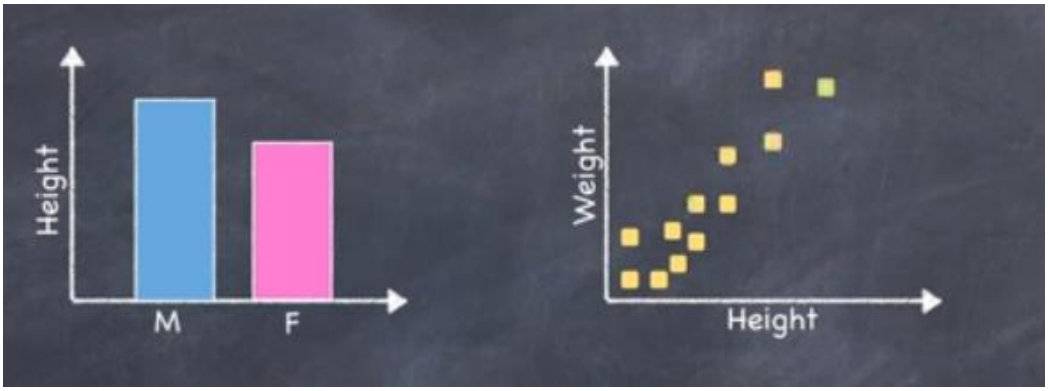
Presentation of Categorical Variables:

- Summarize as table
- Visualize as graph

Example: Combination of categoricala nd numerical variable

Interpretation-1: Average age of Male (1.58) is greater than female (1.54).

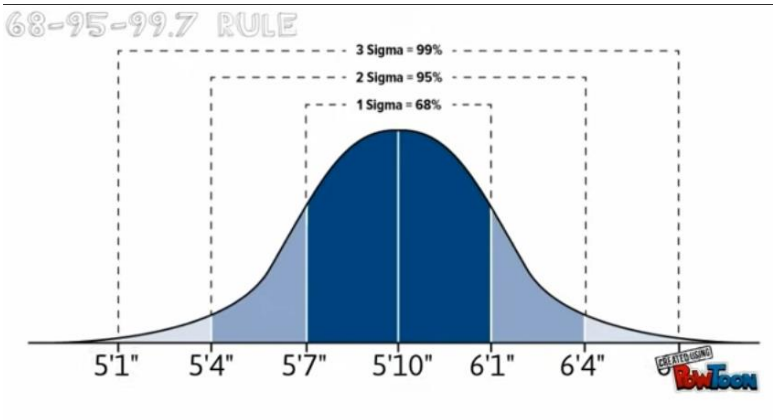
Example-2: Relationship among categorical and numerical variables:



Meaures of central tendency (Descriptive statistics):

Why do we need Mean, Median, Mode, Range, IQR, Variance, Standrd deviation?

- **Mean:** the mathematical average of all the terms.
- **Median: Median** is a **statistical** measure that determines the middle value of a dataset listed in ascending order (i.e., from smallest to largest value). The measure divides the lower half from the higher half of the dataset.
- **Mode:** The **mode** is useful when there are a lot of repeated values in a dataset.
- **Percentile: Percentiles** are **used** to understand and interpret **data**. The nth **percentile** of a set of **data** is the value at which n percent of the **data** is below it. In everyday life, **percentiles** are **used** to understand values such as test scores, health indicators, and other measurements.
- **Quartile: Quartiles tell us** about the spread of a **data** set by breaking the **data**set into quarters, just like the median breaks it in half.
- **Range:** The **range** is the spread of your data from the lowest to the highest value in the distribution.
- **IQR:** The **interquartile range** (**IQR**) is the distance between the first and third quartile marks. The **IQR** is a measurement of the variability about the median. More specifically, the **IQR tells us** the range of the middle half of the **data**.
- **Variance: Variance** is a statistical figure that determines the average distance of a set of variables from the average value in that set. It is used to provide insight into the spread of a set of data, mainly through its **role** in calculating standard deviation.
- **Standard deviation:** It is a measure of the average distance between the values of the **data** in the set and the mean. One, two and three standard deviations. Example: Average height of American people is 5'.10'' when SD is .3



Excel:

Index	Gender	Gender2	Age group	Height (m)	Weight (kg)	Height - Ordered	xbar	x-xbar	(x-xbar)^2	Measures of Center	Manually	Automated	Measures of Variability	Result
1	Female	1	Adult	1.4	60	1.1	1.39	0.01	0.0001	Mean	1.39	1.39	Range	0.5
2	Male	0	Child	1.2	15	1.2	1.39	-0.19	0.0361	Median (Asc order> average of mid 2)	1.45	1.45	IQR (Q3-Q1)	0.2
3	Male	0	Adult	1.5	85	1.3	1.39	0.11	0.0121	Mode (Most frequent number)	1.5	1.5	Variance-S^2 = ∑(Xi-X-bar)^2/n-1	0.025444
4	Female	1	Adult	1.5	74	1.3	1.39	0.11	0.0121	Percentile (Find Index of number)	75 Percentile =10*0.75, 7.5 index = 1.5	1.5	Standard Deviation (SQRT)	0.159513
5	Male	0	Adult	1.3	77	1.4	1.39	-0.09	0.0081	Quartile-Q1 (1/4(n+1)th)	Formula =1/4*(10+1), Result: 2.75th = 3 == 1.3	1.3	Coef. Of Variation	11.00091
6	Female	1	Elderly	1.6	65	1.5	1.39	0.21	0.0441	Quartile-Q3 (3/4(n+1)th)	Formula =3/4*(10+1), Result: 8.25 th = 8 = 1.5	1.5		
7	Female	1	Adult	1.3	44	1.5	1.39	-0.09	0.0081	Quartile-Q2 (Q3-Q1)	8-3 = 5 th == 1.4	1.45		
8	Male	0	Child	1.1	18	1.5	1.39	-0.29	0.0841				1.5*IQR	0.3
9	Male	0	Adult	1.5	35	1.5	1.39	0.11	0.0121				Lower Fence	1
10	Male	0	Elderly	1.5	82	1.6	1.39	0.11	0.0121				Upper Fence	1.2
			Mean:	1.39					0.229					

Stata commands:

Summarize: ***summarize heightm***

```
    Variable[a] |       Obs[b]        Mean[c]    Std. Dev.[d]       Min[e]        Max[f]
-------------+--------------------------------------------------------
       heightm |        10         1.39    .1595131           1.1         1.6
```

```
summarize heightm weightkg
```

Variable	Obs	Mean	Std. Dev.	Min	Max
heightm	10	1.39	.1595131	1.1	1.6
weightkg	10	55.5	26.00534	15	85

a. **Variable** – This column indicates which variable is being described. You can list more than one variable after the summarize command; when you do, you will see each variable on its own line of the output.

b. **Obs** – This column tells you the number of observations (or cases) that were valid (i.e., not missing) for that variable. If you had 10 observations in your data set, but you had 2 missing values for the variable **heightm,** then the number in this column would be 8.

c. **Mean** – This is the mean of the variable. In this case, our variable **heightm** ranges from 0 to 1 (the min and max values), so the mean is actually the proportion of observations coded as 1.

d. **Std. Dev.** – This is the standard deviation of the variable. This gives information regarding the spread of the distribution of the variable.

summarize heightm, detail

```
                        writing score
-------------------------------------------------------------
      Percentiles      Smallest[i]
 1%[e]       1.1            1.1
 5%          1.1            1.2
10%          1.15           1.3       Obs[b]                  10
25%[f]       1.3            1.3       Sum of Wgt.[k]          10
50%[g]       1.45                     Mean[c]                 1.39
                        Largest[j]    Std. Dev.[d]            .1595131
75%[h]       1.5            1.5
90%          1.55           1.5       Variance[l]             .0254444
95%          1.6            1.5       Skewness[m]            -.5228832
99%          1.6            1.6       Kurtosis[n]             2.104784
```

e. **1%** – This is the first percentile. Percentiles are calculated by ordering the values of a variable from lowest to highest, and then finding the value that corresponds to whatever percent you are interested in, in this case, 1%. Hence, 1% of the values of the variable **heightm** are equal to or less than 1.1.

f. **25%** – This is the 25th percentile, also known as the first quartile.

g. **50%** – This is the 50th percentile, also known as the median. If you order the values of the variable from lowest to highest, the median would be the value exactly in the middle. In other words, half of the values would be below the median, and half would be above. This is a good measure of central tendency if the variable has outliers.

h. **75%** – This is the 75th percentile, also known as the third quartile.

i. **Smallest** – This is a list of the four smallest values of the variable. In this example, the four smallest values are all 1.1.

j. **Largest** – This is a list of the four largest values of the variable. In this example, the four largest values are all 1.6.

b. **Obs** – This column tells you the number of observations (or cases) that were valid (i.e., not missing) for that variable. If you had 10 observations in your data set, but you had 2 missing values for the variable **heightm,** then the number in this column would be 8.

k. **Sum of Wgt**. – This is the sum of the weights. In Stata, you can use different kinds of weights on your data. By default, each case (i.e., subject) is given a weight of 1. When this default is used, the sum of the weights will equal the number of observations.

c. **Mean** – This is the arithmetic mean across the observations. It is the most widely used measure of central tendency. It is commonly called the average. The mean is sensitive to extremely large or small values.

d. **Std. Dev.** – This is the standard deviation of the variable. This gives information regarding the spread of the distribution of the variable.

l. **Variance** – This is the standard deviation squared (i.e., raised to the second power). It is also a measure of spread of the distribution.

m. **Skewness** – Skewness measures the degree and direction of asymmetry. A symmetric distribution such as a normal distribution has a skewness of 0, and a distribution that is skewed to the left, e.g., when the mean is less than the median, has a negative skewness.

n. **Kurtosis** – Kurtosis is a measure of the heaviness of the tails of a distribution. A normal distribution has a kurtosis of 3. Heavy tailed distributions will have kurtosis greater than 3 and light tailed distributions will have kurtosis less than 3.

Command table produces frequencies and descriptive statistics per category.

```
table gender, contents(freq mean heightm mean weightkg)
```

Source: https://stats.idre.ucla.edu/stata/output/descriptive-statistics-using-the-summarize-command/