Week 8

Definition (Population and Sample): Population is a set containing all related subjects under investigation. A sample is a subset of the population. While a Population is denoted by N, a sample is denoted by n. The numerical values related to Population are called parameters, the numerical values pertinent to the sample is called statistics. Any data can be examined either using descriptive statistics or running inferential statistics. The former is related to taking the snapshot of the data under investigation, which means that summarizing the data, the latter deals with a forecast, estimation, and prediction, and modeling the data.

Any data set could have errors that could be classified as follows:

1.The errors stemmed from the persons who measures

2.The errors that are stemmed from a subject or object

3.The error that is related to the device used

4.the errors that are related to the environment

5.The errors relate to a factor called chance, which means that the factor that we do not have control over

Definition (Measure-Scale): Objects could have either numerical or string values. Based on these values, the four measurement scales are defined. 1.Nominal, 2.Ordinal, 3.Interval, 4.Ratio.

Types of Scales

1.Nominal Scale: The data based on nominal scale can only be grouped. No mathematical calculation can be conducted using this scale. For example, sex, marital status, religion, working status, ID, vocation is measured using nominal scale. Only frequencies and mode statistics are calculated for this scale.

Student ID	Department	Where was born	Sex	Marital Status
1123	Comp. Eng.	İstanbul	Male	Single
1876	Ind.Eng	Diyarbakır	Female	Single
1911	Food.Eng	İzmir	Female	Single
1001	Civil Eng.	Trabzon	Male	Married

2. Ordinal Scale: Ordinal scale is the extended version of nominal scale by taking into account the order when classifying the values of the variables. For example, when service evaluations are conducted for hotels, restaurants and hospitals. Services are assessed by some remarks such as “very good, “good”, “medium”, “poor, “very poor”. By using this scale, some statistics such as frequency, mean, mode and decimals and percentiles could be computed.

Employee	Education	Income	Age	The Level of Experience
Ali Temiz	Middle school	Low	Young	Low
Mustafa Uykulu	High school	High	Middle	Medium
Mehmet Üşümez	High School	Medium	Old	High
Ayşe Koşar	University	High	Old	High

3.Interval Scale: This scale uses non real starting point and comparisons cannot be made. All statistical procedures are conducted for the data measured by interval scale. For example, when measuring temperature either using Celsius or Fahrenheit, or grading exam pares.

4.Ration Scale: The highest level of measurement. The stating point is real and comparisons can be made. For example, Length, height, and speed

Descriptive Statistics

1.Central Tendency Measure

a. Mean(Average)

b. Median

c. Mode

2.Central Dispersion Measures

a. Interval value

b. Absolute deviation from mean

c. Variance – standard deviation

d. Decimals

e. Percentiles

3.Distribution measures

a. Kurtosis

b. Skewness

Software that can be used to run statistical analysis

a. requiring licenses (SPSS-SAS-STATA-E-views-Minitab-Mat Lab, JMP)

b.no licenses (R)

Inferential Statistics

1.Modeling (regression, binary regression, discriminant analysis, factor analysis, ANOVA, ANCOVA, MANCOVA and so on)

2.Forecast (Time series analysis, regression)

3.Classification (k-means clustering, hierarchical clustering, regression trees, and so on)

1.Central Tendency Measures

a. Definition (Mean/Average): The arithmetic mean or mean or average is defined as the summation of all numerical values divided by the number of observations.

Data: 34500/30700/32900/36000/34100/33800/32500

Find mean.

$$\overline{x} = \frac{34500 + 30700 + 32900 + 36000 + 34100 + 33800 + 32500}{7}$$

$$\overline{x} = 33500$$

Definition (Median): The median of the numerical values is computed based on the algorithm provided as follows:

Algorithm 1:

a. The number of observations (n) is odd:

a1. Numerical values are organized in ascending order

a2. Find $\frac{(n+1)}{2}$ th observation

a3. The numerical value of this observation is assigned to the median

Algorithm 2:

b: The number of observations (n) is even:

b1. Numerical values are organized in ascending order

b2. Find $\frac{n}{2} th\ and\ \frac{n}{2} + 1$ th observations

b3. The numerical values of these observations are added and divided by 2, which means that the average is computed for these numerical values. Average is called the median.

Data:34500/30700/32900/36000/34100/33800/32500

Find median

n=7, the number of observations is odd.

a1.30700/32500/32900/33800/34100/34500/36000

a2. $\frac{(n+1)}{2} = 4th\ observation$

a3.median=33800

Definition (Mode): The observation with the highest frequency is called mode.

Data:34500/30700/32900/36000/34100/33800/32500

Find mode

The mode does not exist.

Data:34500/30700/32900/36000/34100/33800/32500/32900

Find mode

Mode=32900

There exists one Mode.

Data:34500/30700/32900/36000/34100/33800/32500

Find mode

Remark: Any data set could have no mode, or one mode, or more than one mode.

2.Central Dispersion Measures

a.Interval

Data:30700/32900/36000/34100/33800/32500/32900

Find interval.

Interval=Max-Min=36000-30700=5300

b. **Definition (Deviation from Absolute Mean): the formula given below is used to calculate Deviation from Absolute Mean**

$$DAM = \frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n}$$

x_i	$\lvert x_i - \overline{x} \rvert$
34500	|34500-33500|
30700	|30700-33500|
32900	|32900-33500|
36000	|36000-33500|
34100	|34100-33500|
33800	|33800-33500|
32500	|32500-33500|

OMS=1257.14

c. **Variance-Standard Deviation**

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

x_i	$(x_i - \overline{x})^2$
34500	$(34500-33500)^2$
30700	$(30700-33500)^2$
32900	$(32900-33500)^2$
36000	$(36000-33500)^2$
34100	$(34100-33500)^2$
33800	$(33800-33500)^2$
32500	$(32500-33500)^2$

$$s^2 = 1678.92$$

S=986.33

d. Decimals

Definition (Deciamals):Decimals of any data set can be computed by an algorithm given below

D1. Numerical values are organized in ascending order

D2. $\frac{k(n+1)}{10}$ **th term is obtained.**

D3. $D_k = Lower\ Bound$**+fraction value * (Upper Bound-Lower Bound)**

Data:34500/30700/32900/36000/34100/33800/32500

Find the third decimal or 30 percent of the data.

1. 30700/32500/32900/33800/34100/34500/36000

2.k=3, $\frac{3(7+1)}{10}$**=2.4 th term. However, there exists no such a term in the data set. On the other hand, this statistic could be estimated since it is a value between the second term and the third term, namely it is between 30700 and 32900.**

Using the formula given above

3.32500+0.4(32900-32500)=32660

While 30 percent of the data is located on the left side of 32600, 70 percent of the data is located on the right side of 32600.

e.Quantiles (First and third ones)

Definition (Quantiles): Quantile are statistics that split data into 25 percent and 75 percent portions respectively. Algorithm given below helps find those statistics.

First quantile:

E1. Numerical values are organized in ascending order

E2. $\frac{(n+1)}{4}$ **th decimal value is obtained**

E3. $Q_1 =$**Lower Bound+fraction value *(Upper Bound-Lower Bound)**

Third quantile

E1. Numerical values are organized in ascending order

E2. $\frac{3(n+1)}{14}$ **th decimal value is obtained**

E3. $Q_3 =$ **Lower Bound+fraction value *(Upper Bound-Lower Bound)**

Data:34500/30700/32900/36000/34100/33800/32500, find Q_1 and Q_3.

1. 30700/32500/32900/33800/34100/34500/36000

2. $\frac{(7+1)}{4}$=2 nd term. In this case, we do not use the formula to predict the value. The second term exists in the data set.

3. Ç$_1$=32500

While 25 percent of the data is located on the left side of 32500, 75 percent of the data is located on the right side of 32500.

1. 30700/32500/32900/33800/34100/34500/36000

2. $\frac{3(7+1)}{4}$=6 ci terim. Böyle tek bir sayısal değer Ç3 değeri formül kullanmadan elde edilir.

3. Ç$_1$=34500

While 75 percent of the data is located on the left side of 35400, 25 percent of the data is located on the right side of 35400.

f.Percentiles

Definition (Percentiles): With the help of percentiles, when data is organized in ascending order, any percentile value could be directly obtained or estimated by a formula. An algorithm below is used to compute it.

F1. Numerical values are organized in ascending order

F2. $\frac{m(n+1)}{100}$ th percentile value is obtined where m denotes the percentile value

F3. $P_m =$ Lower Bound+fraction value *(Upper Bound-Lower Bound)

Data:34500/30700/32900/36000/34100/33800/32500, Find percentile of 33 and 82.

F1. Numerical values are organized in ascending order

30700/32500/32900/33800/34100/34500/36000

F2. $\frac{33(n+1)}{100}$ =2.64 th percentile value is obtained. However, there is no such term in the data set. On the other hand, we know the numerical values of the second and third terms in the data set. Then, the 2.64th term could be estimated using the numerical values of the second and third terms.

F3.Y$_{33}$= Lower Bound+fraction value *(Upper Bound-Lower Bound)

=32500+0.64(32900-32500)=32756

While 33 percent of the data is located on the left side of 32756, 67 percent of the data is located on the right side of 32756.

F1. Numerical values are organized in ascending order

---30700/32500/32900/33800/34100/34500/36000

F2. $\frac{82(n+1)}{100}$ **=6.56 the percentile term is obtained. However, there is no such term in the data set. On the other hand, we know the numerical values of the sixth and seventh terms in the data set. Then, the 6.56th term could be estimated using the numerical values of the sixth and seventh terms.**

F3.P_{82}= **Lower Bound+fraction value *(Upper Bound-Lower Bound)**

=34500+0.56(3600-34500)=35340

While 82 percent of the data is located on the left side of 35340, 18 percent of the data is located on the right side of 35340.