İST292 STATISTICS LESSON 1

1. INTRODUCTION TO STATISTICS AND SOME IMPORTANT DEFINITIONS

Statistics is the science of data. This involves collecting, classifying, summarizing, organizing, analyzing, presenting and interpreting numerical information.

Statistical methods are particularly useful for studying, analyzing and learning about ***populations*** of ***experimental units***.

Experimental (or Observational) Unit: An ***experimental (or observational) unit*** is an object (e.g., person, thing, transaction (işlem), or event) about which we collect data.

Population: A ***population*** is a set of units (usually people, objects, transactions, o events) that we are interested in studying. It is the data which we have not completely examined but to which our conclusions refer. The population size is usually indicated by a capital ***N***.

Examples:
- all unemployed workers in Turkey
- all registered voters in Ankara
- all the cars produced last year by a particular assembly line (montaj hattı/montaj üretim hattı)
-the set of all accidents occurring on a particular stretch (alan/kısım/saha/bölüm) of interstate highway (eyaleetlerarası otoban) during a holiday period.
In studying a population, we focus on one or more characteristics or properties of the units in the population. We call such characteristics ***variables***. For Example, we may be interested in the variables *age, gender and number of years of education* of the people currently unemployed in Turkey.

Variable: A ***variable*** is a characteristic or property of an individual experimental (or observational) unit in the population. A variable is a characteristic (e.g., temperature, age, race, growth, education level, etc.) that we would like to measure on individuals. The actual measurements recorded on individuals in the sample are called **data.**

Two Types: ***Quantitative variables*** have measurements (data) on a numerical scale. ***Categorical (Qualitative) variables*** have measurements (data) where the values simply indicate group membership.

Example: Which of the following variables are quantitative in nature? Which are categorical?

- age - ***Quantitative variables***
- advertising medium (reklam aracı) (radio/TV/internet)- ***Categorical variables***
- number of cigarettes smoked per day- ***Quantitative variables***
- smoking status (yes/no)- ***Categorical (Qualitative) variables***

Measurement: ***Measurement*** is the process we use to assign numbers to variables of individual population units. We might, for example, measure the performance of the president by asking a registered voter to rate it on a scale from 1 to 10. Or we might measure the age of the Turkey workforce simply by asking each worker, "How old are you?" In other cases, measurement involves the use of instruments such as stopwaches (kronometre), scales (ölçek) and calipers (kalınlık/çap ölçer). It can be used a scale from 0 to 10 or 0 to 100 for evaluating exam papers in a course.

Census (sayım, populasyon sayımı): If the population you wish to study is small, it is possible to measure a variable for every unit in the population. When we measure a variable for every unit of a population, it is called a ***census*** of the population. In Turkey, national cencus is made for every 5 years. The number of the units in the population is shown as capital N

Sample: A ***sample*** is a subset of the units of a population. The sample size is shown as lower case ***n***.

Often data are selected from some larger set of data whose characteristics we wish to estimate. This selection process is called ***sampling***.

Example:
If your company manufactures one million laptops, they might take a sample of say, 500, of them to test quality. The population N = 1000000 and the sample n= 500.

Parameter: A characteristic of a population. The population mean, μ and the population standard deviation, σ, are two examples of population parameters. If you want to determine the population parameters, you have to take a census of the entire population. Taking a census is very costly.

Statistic: A ***statistic*** is a measure that is derived from the sample data. For example, the sample mean $(\overline{X})$ and the sample standard deviation (S) are statistics. They are used to estimate the population parameters.

Statistics involves two different processes:
(1) describing sets of data and **(2)** drawing conclusions (making estimates, decisions, predictions, etc.) about the sets of data on the basis of sampling. So, the applications of statistics could be divided into two broad areas: ***descriptive statistics*** and ***statistical inference***.

Descriptive statistics utilizes numerical and graphical methods to look for patterns in a data set, to summarize the information revealed in a data set, and to present that information in a convenient form. Those statistics that summarize a sample of numerical data in terms of averages and other measures for the purpose of description, such as the mean and standard deviation. This includes the presentation of data in the form of graphs, charts, and tables.

[For example, after grading an exam, a teacher may calculate the average grade to summarize the overall performance of the class. No inferences being made here.]

Statistical Inference is an estimate, prediction, or some other generalization about a population based on information contained in a sample.

The process of using sample statistics to draw conclusions about population parameters is known as ***statistical inference***. For example, using $(\overline{X})$ based on a sample of, say, n=1000) to draw conclusions about μ (population of, say, 300 million). This is a measure of performance in which the sample measurement is used to estimate the population parameter.

Example 1: ***Nielsen television ratings***
The Nielsen ratings are based on a sample, not the population.
The sample consists of about 5,000 TV households
Population of more than 115,000,000 TV households
For example, if a show has a 10.0 rating, this means that 10% of the entire sample were watching that show. [Note: "Share of audience" is the percentage of those who have the TV on, i.e. of those actually watching TV.]

2. TYPES OF SAMPLES

A) Nonprobability Samples – based on convenience (uygunluk, elverişlik) or judgment (yargı, hüküm). In these types of sampling methods sample of size n is selected from population of size N according to personal opinion. Hence, the problem with a nonprobability sample is that we do not know how representative our sample is of the population.

B) Probability Samples

Probability Sample: A sample collected in such a way that every element in the population has a known chance of being selected.

a) Simple Random Sample: A sample collected in such a way that every element in the population has an equal chance of being selected.

b) Systematic Random Sample: Choose the first element randomly, then every kth observation, where k = N/n

c) Stratified (Tabakalı) Random Sample: The population is sub-divided based on a characteristic and a simple random sample is conducted within each stratum (tabaka).

d) Cluster Random Sample: First take a random sample of clusters from the population of cluster. Then, a Stratified Random Sample (SRS) within each cluster. Example, election district (seçim bölgesi).

3. DATA

3.1. Types of Data

Statistics is the science of data and that data are obtained by measuring the values of one or more variables on the units in the sample (or population). All data (and hence the variables we measure) can be classified as one of two general types: ***quantitative (nicel/sayısal) data and qualitative (nitel) data***.

A) Qualitative (or Categorical) Data: Qualitative data are measurements that cannot be measured on a natural numerical scale; they can only be classified into one of a group of categories.

Qualitative Data Examples:
1. Sex □Male □Female (Nominal). Nanionality: German, French, British, Turkish, Arabian, etc.
2. The political party affiliation (üyeliği) (Democrat, Rebuplican, or Independent) in a sample of 50 voters. (Nominal)
3. The defective status (defective or not) of each of 100 computer chips manufactured by Intel
4. A taste tester's ranking (best, worst, etc.) of four brands of barbecue sauce for a panel of 10 testers (Ordinal)

Often, we assign arbitrary (keyfi) numerical values to qualitative data for ease of computer entry and analysis. But these assigned numerical values are simply codes: They can't be meaningfully added, subtracted, multipilied or divided. For example, we might code Democrat=1, Rebuplican=2, and Independent=3. Smilarly, a taste tester might rank the barbecue sauces from 1 (best) to 4 (worst). These are simply arbitrarily (keyfi olarak) selected numerical codes for the categories and have no utility beyond that.

B) Quantitative Data: ***Quantitative data*** are measurements that are recorded on a naturally occurring numerical scale.

Quantitative Data Examples:
1. The temperature (in degrees Celsius) at which each piece in a sample of 20 pieces of heat-resistant plastic begins to melt (Interval)
2. The current unemployment rate (measured as a percentage) in each of the 81 cities (Ratio)
3. The number of convicted (mahkum edilmiş) murderers (katil) who receive the death penalty each year over a 10-year period (Ratio)

Quantitative data result in numerical responses, and may be ***discrete*** or ***continuous***.

a) Discrete data arise from a counting process.
Example:
How many courses have you taken at this College? _________

b) Continuous data arise from a measuring process.
Example:
How much do you weigh? ________

One way to determine whether data is continuous, is to ask yourself whether you can add several decimal places to the answer. You may weigh 150 pounds but in actuality may weigh 150.23568924567 pounds. On the other hand, if you have 2 children, you do not have 2.3217638 children.

3.2. Levels of Data

Qualitative (Categorical) data can be subclassified as either ***nominal data*** or ***ordinal data.*** The categories of an ***ordinal data*** set can be ranked or meaningfully ordered, but the categories of a ***nominal data*** set can't be ordered.

A) Nominal Data

Classification – categories

When objects are measured on a nominal scale, all we can say is that one is different from the other
Examples: gender (sex) (cinsiyet), male, female
occupation (iş/meslek), doctor, teacher, researcher, astronuat, engineer
ethnicity (etnik yapı),
marital status (medeni hal), single, maried, divorced
etc.

B) Ordinal Data

Ranking, but the intervals between the points are not equal. We can say that one object has more or less of the characteristic than another object when we rate them on an ordinal scale. Thus, a category 5 hurricane is worse than a category 4 hurricane which is worse than a category 3 hurricane, etc.

Examples: hardness of minerals scale, income as categories, class standing (kaçıncı sınıf), rankings of football teams, military rank (general, colonel (albay), major (binbaşı), lieutenant (yüzbaşı/üsteğmen), sergeant (astsubay/çavuş), etc.), hurricane rankings (category 1, 2, …, category 5)

Quantitative data can be subclassified as either ***interval data*** or ***ratio data***. For ***ratio data***, the origin (i.e. the value 0) is a meaningful number. But the origin has no meaning with interval data. Consequently, we can add and subtract interval data, but we can't multiply and divide them.

C) Interval Data

Equal intervals, but no "true" zero.
Examples: IQ, temperature

Since there is no true zero – the complete absence (eksiklik) of the characteristic you are measuring – you can't speak about ratios.

Example:
Suppose
New York temperature = 40 degrees
Buffalo temperature = 20 degrees
Does that mean it is twice as cold in Buffalo as in NY?

İST292 STATISTICS LESSON 2
DESCRIPTIVE STATISTICS

1. DESCRIBING DATA SETS

The numerical findings of a study should be presented clearly, concisely, and in such a manner that an observer can quickly obtain a feel for the essential characteristics of the data. Over the years it has been found that tables and graphs are particularly useful ways of presenting data, often revealing important features such as the range, the degree of concentration (yığma), and the symmetry of the data. In this section we present some common graphical and tabular ways for presenting data.

1.1. Frequency Tables and Graphs

How we display data distributions depends on the type of variable(s) or data that we are dealing with.

Categorical: pie charts, bar charts

Quantitative: line graphs, frequency polygon, stemplots, histograms, boxplots

A data set having a relatively small number of distinct (farklı) values can be conveniently presented in a ***frequency table***. For instance, ***Table1 is a frequency table*** for a data set by asking 500 students how many cigarettes they smoked.

Table 1. The number of cigarettes smoked

Number of cigarettes smoked per day	frequency
0	12
5	15
8	33
10	17
12	23
15	215
20	100
25	30
30	50
40	5
Total	**500**

Table 1 tells us, among other things, that the ***lowest number cigarettes of 0*** was smoked by ***12 of the students*** (nonsmokers were 12), whereas the ***highest number cigarettes of 40*** were smoked by ***5 students.*** The most ***common number of cigarettes were 15***, and smoked by ***215 of the students.***

A) Line Graph

Data from a frequency table can be graphically represented by a **line graph** that plots the ***distinct data values on the horizontal axis*** and indicates ***their frequencies by the heights of vertical lines.*** A line graph of the data presented in Table 1 is shown in Figure 1.

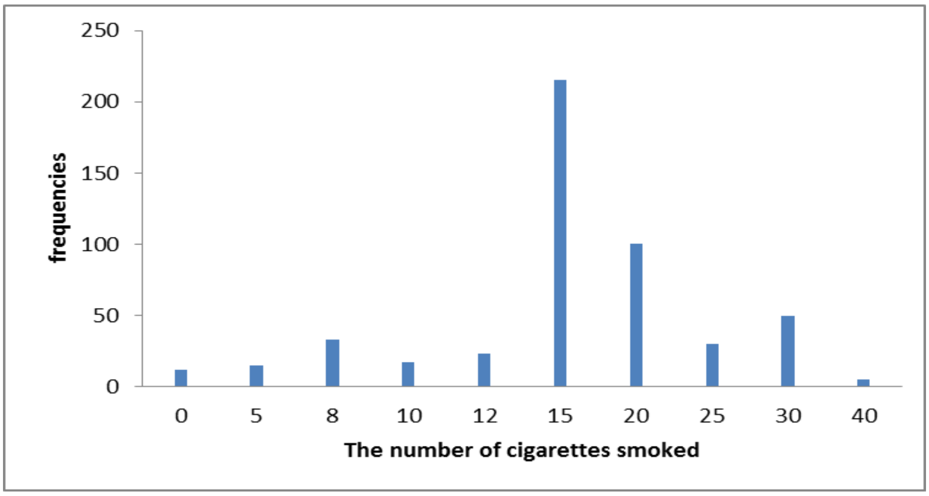


Figure 1. The number of cigarettes smoked per day.

B) Bar Graph

When the lines in a line graph are given added thickness, the graph is called a **bar graph**. Figure 2 presents a bar graph.

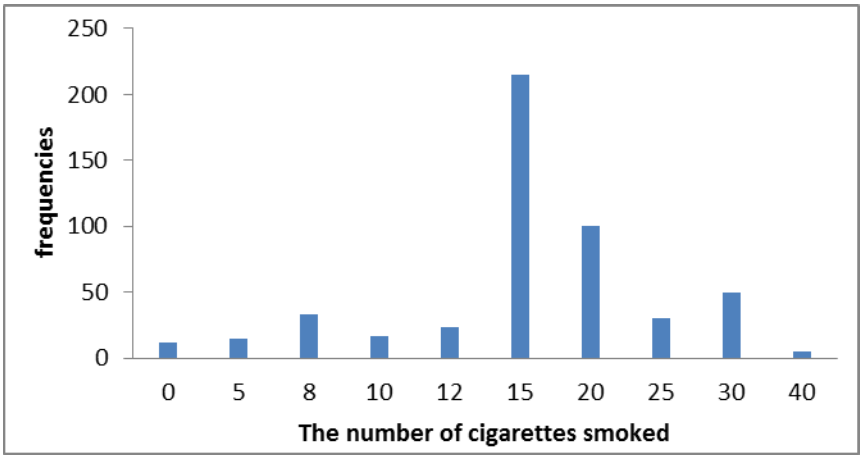
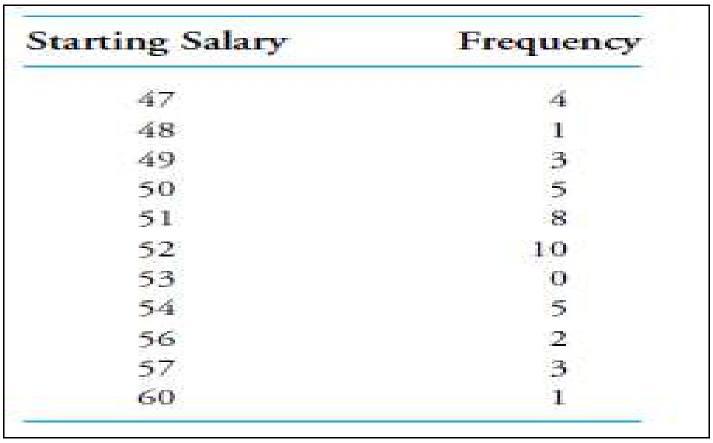


Figure 2. Bar graph for the number of cigarettes smoked per day.

C) Frequency Polygon

A data set having a relatively small number of distinct (farklı) values can be conveniently presented in a ***frequency table***. For instance, ***Table 2 is a frequency table*** for a data set consisting of the starting yearly salaries (to the nearest thousand dollars) of 42 recently graduated students with B.S. degrees (dört yıllık üniversite diploması) in electrical engineering.

Table 2. Starting Yearly Salaries



Starting Salary	Frequency
47	4
48	1
49	3
50	5
51	8
52	10
53	0
54	5
56	2
57	3
60	1

Another type of graph used to represent a frequency table is the **frequency polygon**, which plots the ***frequencies of the different data values on the vertical axis***, and then connects the plotted points with straight lines. Figure 3 presents a ***frequency polygon*** for the data of Table 2.

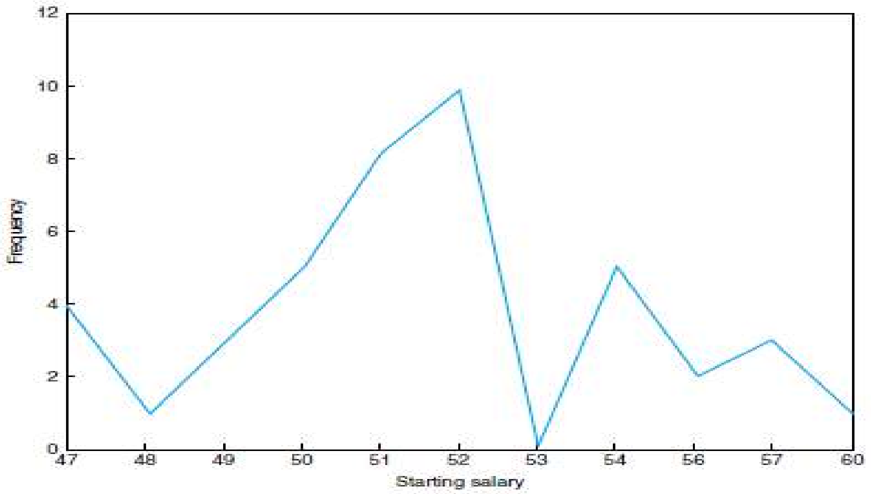


Figure 3. Frequency polygon for starting salary data.

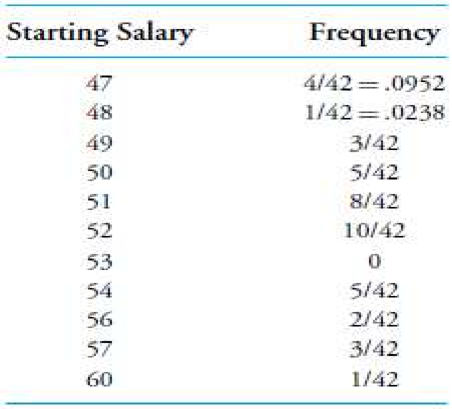
1.2. Relative Frequency Tables and Graphs

Consider a data set consisting of ***n*** values. If ***f*** is the frequency of a particular value, then the ratio ***f* /*n*** is called its **relative frequency**. That is, the relative frequency of a data value is the proportion of the data that have that value.

The ***relative frequencies*** can be represented graphically by ***a relative frequency line*** or ***bar graph*** or by ***a relative frequency polygon***. Indeed, these relative frequency graphs will look like the corresponding graphs of the absolute frequencies except that the labels on the vertical axis are now the old labels (that gave the frequencies) divided by the total number of data points.

Example: Table 3 is a ***relative frequency table*** for the data of Table 2. The relative frequencies are obtained by dividing the corresponding frequencies of Table 2 by ***42***, the size of the data set.

Table 3 Relative Frequency Table of the Salary Data



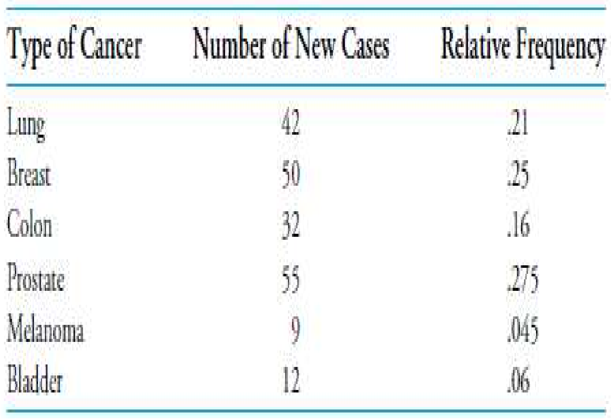
Starting Salary	Frequency
47	4/42 = .0952
48	1/42 = .0238
49	3/42
50	5/42
51	8/42
52	10/42
53	0
54	5/42
56	2/42
57	3/42
60	1/42

D) Pie Chart

The easiest way to portray the distribution of a categorical variable is to use a table of counts and/or percentages.

A pie chart is often used to indicate ***relative frequencies when the data are not numerical in nature. A circle is constructed and then sliced into different sectors; one for each distinct type of data value.*** The relative frequency of a data value is indicated by the area of its sector (daire dilimi, bölüm), this area being equal to the total area of the circle multiplied by the relative frequency of the data value.

Example: The following data relate to the different types of cancers affecting the 200 most recent patients to enroll (kaydetmek, kayda geçirmek) at a clinic specializing in cancer. These data are represented in the ***pie chart*** presented in Figure 4. Melanoma (kara tumor), Bladder (mesane, sidik torbası), Colon (kalın bağırsak), breast (meme), prostate (prostat). Example, 200*0.21=42 patients are Lung cancers.



Type of Cancer	Number of New Cases	Relative Frequency
Lung	42	.21
Breast	50	.25
Colon	32	.16
Prostate	55	.275
Melanoma	9	.045
Bladder	12	.06

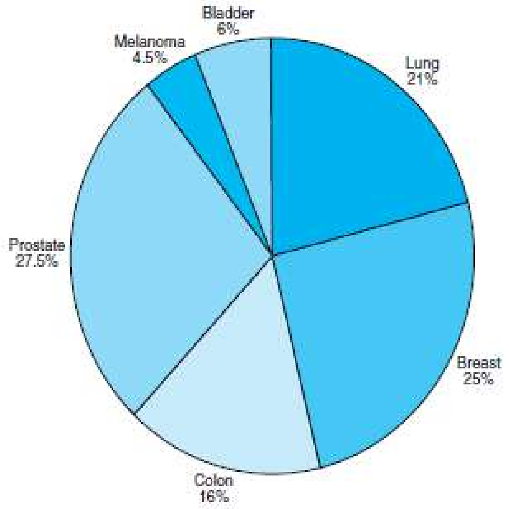


Figure 4

1.3. Grouped Data, Histograms, Cumulative Frequency Plot, and Stem and Leaf Plots

As seen in Subsection 1.2, using a ***line or a bar graph*** to plot the frequencies of data values is often ***an effective way of portraying a data set***. ***However, for some data sets the number of distinct values is too large to utilize this approach.***

Instead, in such cases, it is useful to divide the values into groupings, or **classes**, ***and then plot the number of data values falling in each class.*** The number of classes chosen should be a trade-off between **(1)** choosing too few classes at a cost of losing too much information about the actual data values in a class and **(2)** choosing too many classes, which will result in the frequencies of each class being too small for a pattern to be discernible (farkedilebilir, görülebilir). **Although 5 to 20** number of classes are typical, the appropriate number is a subjective choice, and of course, you can try different numbers of classess to see which of the resulting charts appears to be most revealing about the data. ***It is common, although not essential, to choose class intervals of equal length.***

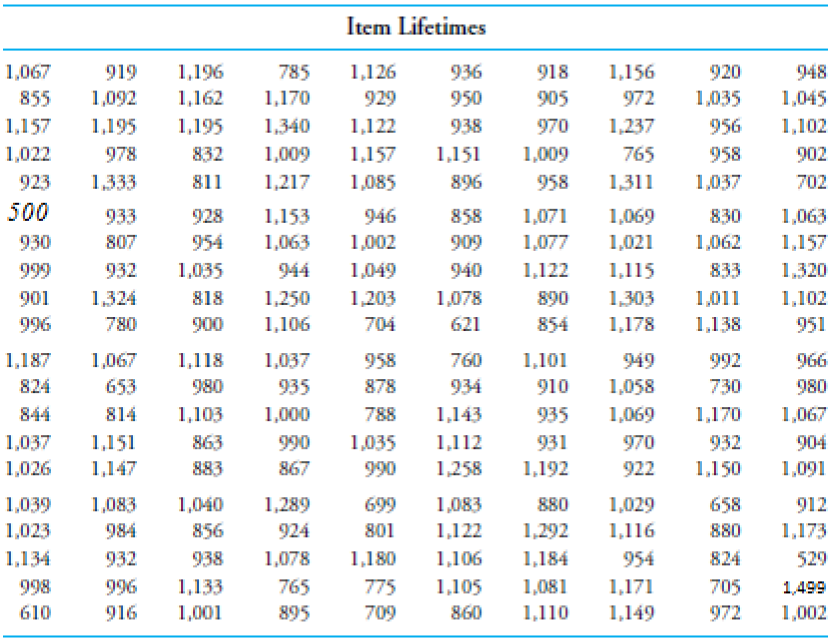
The endpoints of a class are called the **class boundaries.** For example, ***the class boundary 20–30*** contains all values that are both greater than *or equal to* 20 and less than or equal to 30.

The number of classes is shown by ***k***, if it is not given, calculated by the formula $k = 1 + 3.3\log_{10}(n)$.

Class interval is calculated by the formula $c = \dfrac{Range + a}{k}$, for example, ***a*** is equal to 1 if all data values are integer, ***a*** is equal to 0.1, if the data values have only one digit after decimal point.

Table 4 presents the lifetimes of 200 lamps (elektrik lambası). A ***class frequency table*** for the data of Table **4** is presented in Table 5. The class intervals are of length 100, with the first one starting at 500.

Table 4. Life in Hours of 200 Lamps



Item Lifetimes									
1,067	919	1,196	785	1,126	936	918	1,156	920	948
855	1,092	1,162	1,170	929	950	905	972	1,035	1,045
1,157	1,195	1,195	1,340	1,122	938	970	1,237	956	1,102
1,022	978	832	1,009	1,157	1,151	1,009	765	958	902
923	1,333	811	1,217	1,085	896	958	1,311	1,037	702
500	933	928	1,153	946	858	1,071	1,069	830	1,063
930	807	954	1,063	1,002	909	1,077	1,021	1,062	1,157
999	932	1,035	944	1,049	940	1,122	1,115	833	1,320
901	1,324	818	1,250	1,203	1,078	890	1,303	1,011	1,102
996	780	900	1,106	704	621	854	1,178	1,138	951
1,187	1,067	1,118	1,037	958	760	1,101	949	992	966
824	653	980	935	878	934	910	1,058	730	980
844	814	1,103	1,000	788	1,143	935	1,069	1,170	1,067
1,037	1,151	863	990	1,035	1,112	931	970	932	904
1,026	1,147	883	867	990	1,258	1,192	922	1,150	1,091
1,039	1,083	1,040	1,289	699	1,083	880	1,029	658	912
1,023	984	856	924	801	1,122	1,292	1,116	880	1,173
1,134	932	938	1,078	1,180	1,106	1,184	954	824	529
998	996	1,133	765	775	1,105	1,081	1,171	705	1,499
610	916	1,001	895	709	860	1,110	1,149	972	1,002

Table 5. Class Frequency of Table 4.

Group no/Class no	LL	UL	s_i	f_i	p_i	%
1	500	599	549,5	2	0,01	1
2	600	699	649,5	5	0,025	2,5
3	700	799	749,5	12	0,06	6
4	800	899	849,5	25	0,125	12,5
5	900	999	949,5	58	0,29	29
6	1000	1099	1049,5	41	0,205	20,5
7	1100	1199	1149,5	43	0,215	21,5
8	1200	1299	1249,5	7	0,035	3,5
9	1300	1399	1349,5	6	0,03	3
10	1400	1499	1449,5	1	0,005	0,5
				200	1	100

E) Histogram

A bar graph plot of class data, with the bars placed adjacent to each other, is called a **histogram**. The vertical axis of a histogram can represent either the class frequency or the relative class frequency; in the former case the graph is called a **frequency histogram** and in the latter a **relative frequency histogram**. Figure 5 presents a frequency histogram of the data in Table 4.

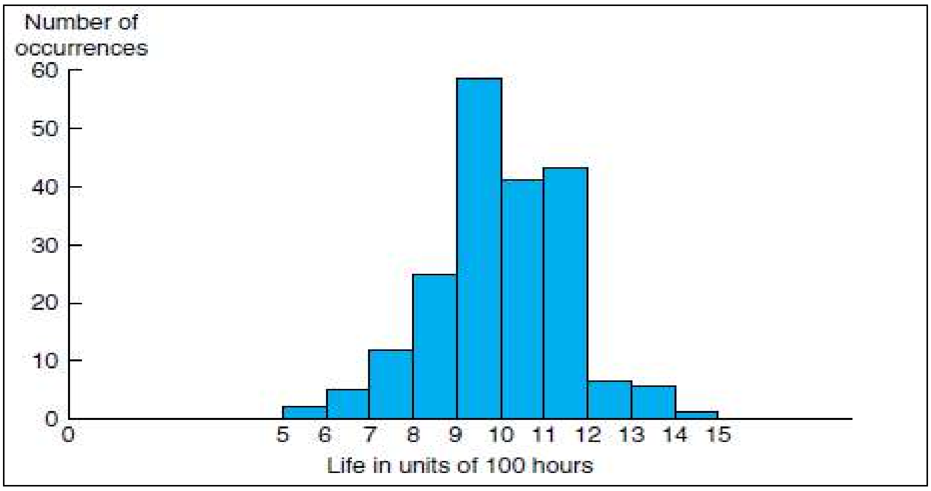


Figure 5. A frequency histogram.

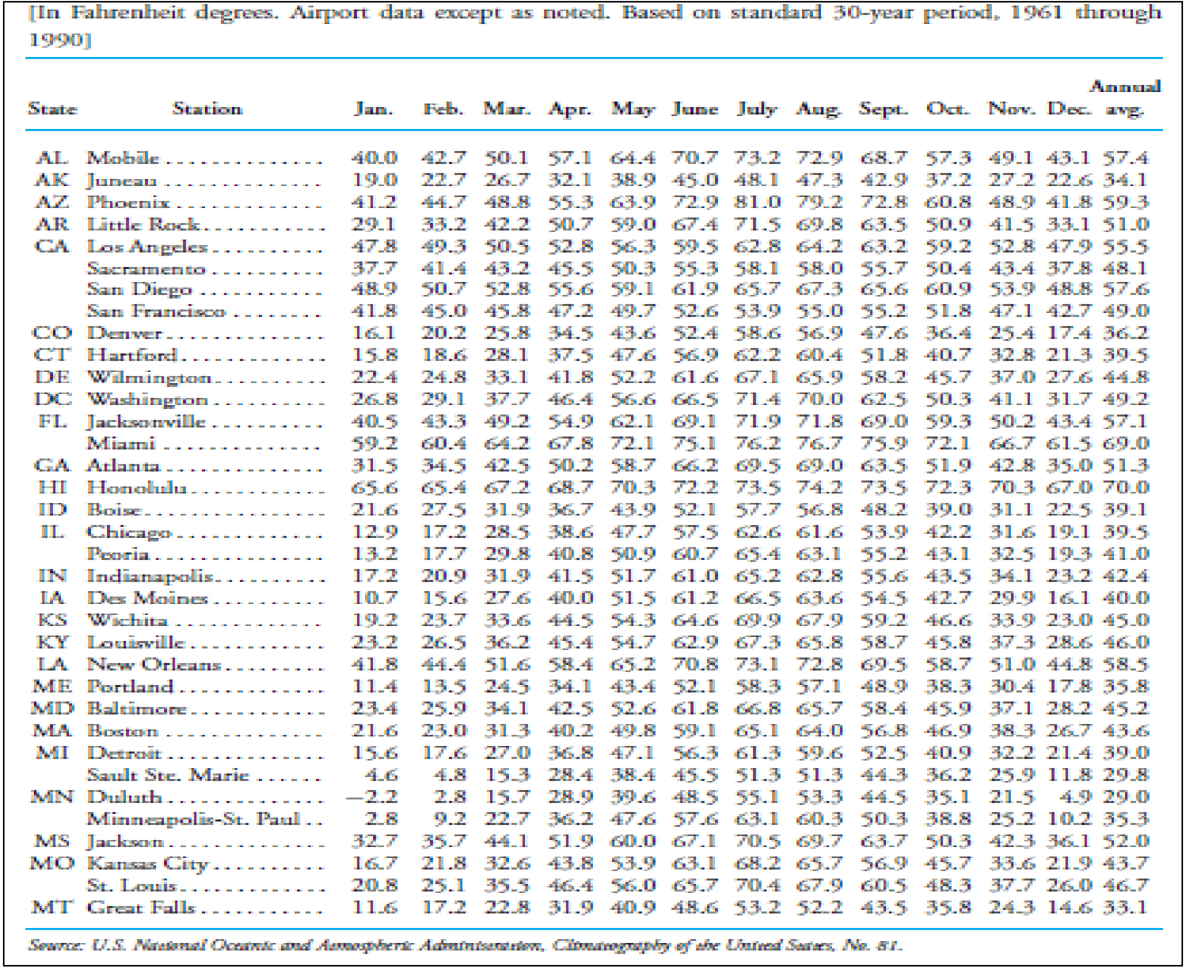
F) Stem and leaf plot (Dal ve yaprak grafiği)

An efficient way of organizing a small- to moderate-sized data set is to utilize a **stem and leaf plot.** Such a plot is obtained by first dividing each data value into two parts —its stem (dal) and its leaf (yaprak). For example, if the data are all two-digit numbers, then we could let the stem part of a data value be its tens digit and let the leaf be its ones digit. Thus, for instance, the value 62 is expressed as

Stem Leaf
6 2
and the two data values 62 and 67 can be represented as
Stem Leaf
6 2,7

Example: Table 6 gives the monthly and yearly average daily minimum temperatures in 35 U.S. cities.

Table 6. Normal Daily Minimum Temperature— Selected Cities

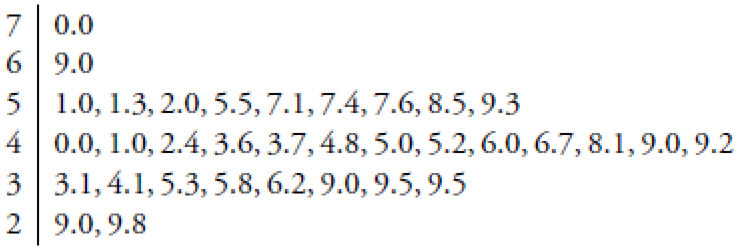


[In Fahrenheit degrees. Airport data except as noted. Based on standard 30-year period, 1961 through 1990]

State	Station	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.	Annual avg.
AL	Mobile	40.0	42.7	50.1	57.1	64.4	70.7	73.2	72.9	68.7	57.3	49.1	43.1	57.4
AK	Juneau	19.0	22.7	26.7	32.1	38.9	45.0	48.1	47.3	42.9	37.2	27.2	22.6	34.1
AZ	Phoenix	41.2	44.7	48.8	55.3	63.9	72.9	81.0	79.2	72.8	60.8	48.9	41.8	59.3
AR	Little Rock	29.1	33.2	42.2	50.7	59.0	67.4	71.5	69.8	63.5	50.9	41.5	33.1	51.0
CA	Los Angeles	47.8	49.3	50.5	52.8	56.3	59.5	62.8	64.2	63.2	59.2	52.8	47.9	55.5
	Sacramento	37.7	41.4	43.2	45.5	50.3	55.3	58.1	58.0	55.7	50.4	43.4	37.8	48.1
	San Diego	48.9	50.7	52.8	55.6	59.1	61.9	65.7	67.3	65.6	60.9	53.9	48.8	57.6
	San Francisco	41.8	45.0	45.8	47.2	49.7	52.6	53.9	55.0	55.2	51.8	47.1	42.7	49.0
CO	Denver	16.1	20.2	25.8	34.5	43.6	52.4	58.6	56.9	47.6	36.4	25.4	17.4	36.2
CT	Hartford	15.8	18.6	28.1	37.5	47.6	56.9	62.2	60.4	51.8	40.7	32.8	21.3	39.5
DE	Wilmington	22.4	24.8	33.1	41.8	52.2	61.6	67.1	65.9	58.2	45.7	37.0	27.6	44.8
DC	Washington	26.8	29.1	37.7	46.4	56.6	66.5	71.4	70.0	62.5	50.3	41.1	31.7	49.2
FL	Jacksonville	40.5	43.3	49.2	54.9	62.1	69.1	71.9	71.8	69.0	59.3	50.2	43.4	57.1
	Miami	59.2	60.4	64.2	67.8	72.1	75.1	76.2	76.7	75.9	72.1	66.7	61.5	69.0
GA	Atlanta	31.5	34.5	42.5	50.2	58.7	66.2	69.5	69.0	63.5	51.9	42.8	35.0	51.3
HI	Honolulu	65.6	65.4	67.2	68.7	70.3	72.2	73.5	74.2	73.5	72.3	70.3	67.0	70.0
ID	Boise	21.6	27.5	31.9	36.7	43.9	52.1	57.7	56.8	48.2	39.0	31.1	22.5	39.1
IL	Chicago	12.9	17.2	28.5	38.6	47.7	57.5	62.6	61.6	53.9	42.2	31.6	19.1	39.5
	Peoria	13.2	17.7	29.8	40.8	50.9	60.7	65.4	63.1	55.2	43.1	32.5	19.3	41.0
IN	Indianapolis	17.2	20.9	31.9	41.5	51.7	61.0	65.2	62.8	55.6	43.5	34.1	23.2	42.4
IA	Des Moines	10.7	15.6	27.6	40.0	51.5	61.2	66.5	63.6	54.5	42.7	29.9	16.1	40.0
KS	Wichita	19.2	23.7	33.6	44.5	54.3	64.6	69.9	67.9	59.2	46.6	33.9	23.0	45.0
KY	Louisville	23.2	26.5	36.2	45.4	54.7	62.9	67.3	65.8	58.7	45.8	37.3	28.6	46.0
LA	New Orleans	41.8	44.4	51.6	58.4	65.2	70.8	73.1	72.8	69.5	58.7	51.0	44.8	58.5
ME	Portland	11.4	13.5	24.5	34.1	43.4	52.1	58.3	57.1	48.9	38.3	30.4	17.8	35.8
MD	Baltimore	23.4	25.9	34.1	42.5	52.6	61.8	66.8	65.7	58.4	45.9	37.1	28.2	45.2
MA	Boston	21.6	23.0	31.3	40.2	49.8	59.1	65.1	64.0	56.8	46.9	38.3	26.7	43.6
MI	Detroit	15.6	17.6	27.0	36.8	47.1	56.3	61.3	59.6	52.5	40.9	32.2	21.4	39.0
	Sault Ste. Marie	4.6	4.8	15.3	28.4	38.4	45.5	51.3	51.3	44.3	36.2	25.9	11.8	29.8
MN	Duluth	−2.2	2.8	15.7	28.9	39.6	48.5	55.1	53.3	44.5	35.1	21.5	4.9	29.0
	Minneapolis-St. Paul	2.8	9.2	22.7	36.2	47.6	57.6	63.1	60.3	50.3	38.8	25.2	10.2	35.3
MS	Jackson	32.7	35.7	44.1	51.9	60.0	67.1	70.5	69.7	63.7	50.3	42.3	36.1	52.0
MO	Kansas City	16.7	21.8	32.6	43.8	53.9	63.1	68.2	65.7	56.9	45.7	33.6	21.9	43.7
	St. Louis	20.8	25.1	35.5	46.4	56.0	65.7	70.4	67.9	60.5	48.3	37.7	26.0	46.7
MT	Great Falls	11.6	17.2	22.8	31.9	40.9	48.6	53.2	52.2	43.5	35.8	24.3	14.6	33.1

Source: U.S. National Oceanic and Atmospheric Administration, Climatography of the United States, No. 81.

The annual average daily minimum temperatures from Table 6 are represented in the following stem and leaf plot.



7	0.0
6	9.0
5	1.0, 1.3, 2.0, 5.5, 7.1, 7.4, 7.6, 8.5, 9.3
4	0.0, 1.0, 2.4, 3.6, 3.7, 4.8, 5.0, 5.2, 6.0, 6.7, 8.1, 9.0, 9.2
3	3.1, 4.1, 5.3, 5.8, 6.2, 9.0, 9.5, 9.5
2	9.0, 9.8

2. SUMMARIZING DATA SETS

In this section we present some ***summarizing statistics***, where a statistic is a numerical quantity whose value is determined by the data.

2.1. Measures of Location/Center (Sample Mean, Sample Median, and Sample Mode)

2.1.1. Sample Mean

If the n observations in a sample are denoted by $x_1, x_2, \cdots, x_n$, the ***sample mean*** is

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

<u>**Example 1:**</u> Let's consider the eight observations collected from the prototype (ön model, ilk örnek) engine (motor, makine) connectors. The eight prototype units are produced and their pull-of forces measured, resulting in the following data (in pounds): $x_1 = 12.6,\ x_2 = 12.9,\ x_3 = 13.4,\ x_4 = 12.3,\ x_5 = 13.6,\ x_6 = 13.5,\ x_7 = 12.6,\ x_8 = 13.1.$ The sample mean is $\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{8} x_i}{8} = \frac{12.6 + 12.9 + \cdots + 13.1}{8} = \frac{104}{8} = 13.0$ pounds.

<u>**Example 2:**</u> The winning scores in the U.S. Masters golf tournament in the years from 1999 to 2008 were as follows: 280, 278, 272, 276, 281, 279, 276, 281, 289, 280. Find the sample mean of these scores.

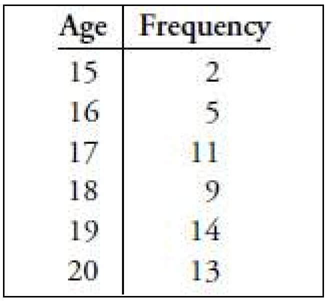
$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{2792}{10} = 279.2$$

Sometimes we want to determine the sample mean of a data set that is presented in a frequency table listing the ***k*** distinct values (class/group's value: average of the limits of group) ***s₁,…,sₖ*** having corresponding frequencies ***f₁,…,fₖ***. Since such a data set consists of $n = \sum_{i=1}^{k} f_i$ observations, with the value *sᵢ* appearing *fᵢ* times, for each *i =1,...,k,* it follows that the sample mean of these ***n*** data values is

$$\bar{x} = \sum_{i=1}^{k} s_i f_i / n$$

By writing the preceding as $\bar{x} = \frac{f_1}{n} s_1 + \frac{f_2}{n} s_2 + \cdots + \frac{f_k}{n} s_k$ we see that the sample mean is a ***weighted average*** of the distinct values, where the weight given to the value s_i is equal to the proportion of the *n* data values that are equal to s_i , i =1, . . . , *k*.

Example: The following is a frequency table giving the ages of members of a symphony orchestra for young adults. Find the sample mean of the ages of the 54 members of the symphony.



Age	Frequency
15	2
16	5
17	11
18	9
19	14
20	13

$$\bar{x} = (15\cdot 2 + 16\cdot 5 + 17\cdot 11 + 18\cdot 9 + 19\cdot 14 + 20\cdot 13)/54 \approx 18.24$$

2.1.2. Sample Median

Another statistic used to indicate the center of a data set is the **sample median**; loosely speaking, it is the middle value when the data set is arranged in increasing order.

The ***median*** is the middle of the data (after data is arranged in ascending or descending order); half the observations are less than the median and half are more than the median. To get the median, we must first rearrange the data into an ***ordered array***. *Generally, we order the data from the lowest value to the highest value.* The median is the data value such that half of the observations are larger than it and half are smaller. *It is also the 50th percentile (we will be learning about percentiles).*

If n is odd, the median is the middle observation of the ordered array. If n is even, it is midway between the *two* central observations. Order the values of a data set of size *n* from smallest to largest. If *n* is odd, the **sample median** is the value in position *(n*+1*)*/2; if *n* is even, it is the average of the values in positions *n*/2 and *n*/2+1.

Example 1: The data set is 10, 20, 30, 40, 50, 60 and n=6 and **median=(30+40)/2=35**

Example 2: Find the sample median for the ages of members of a symphony orchestra for young adults data. Since there are 54 data values, it follows that when the data are put in increasing order, the sample median is the average of the values in positions 27 and 28. Thus, the sample median is 18.5.

Note that the mean and median are UNIQUE for a given set of data.

Advantage: ***The Median is not affected by extreme values.*** In the ***previous*** example, if you change the 60 to 6,000, the median will still be 35. The mean, on the other hand will change by a great deal. The sample mean and sample median are both useful statistics for describing the central tendency of a data set. The sample mean makes use of all the data values and is affected by extreme values that are much larger or smaller than the others; the sample median makes use of only one or two of the middle values and is thus not affected by extreme values.

2.1.3. Sample Mode

Another statistic that has been used to indicate the central tendency of a data set is the **sample mode,** defined to be the value that occurs with the greatest frequency. If no single value occurs most frequently, then all the values that occur at the highest frequency are called *modal values*. Mode is corresponded to the value which is the most observed in the data.

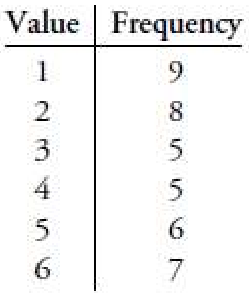
Example1: 1, 1, 1, 2, 3, 4, 5

The mode is 1 since it occurs three times. The other values only appear once in the data set.

Example 2: 5, 5, 5, 6, 8, 10, 10, 10

The modes for this data set are 5 and 10. This is a ***bi-modal (iki tepeli)*** dataset.

Example 3: The following frequency table gives the values obtained in 40 rolls of a die.



Value	Frequency
1	9
2	8
3	5
4	5
5	6
6	7

Find **(a)** the sample mean, **(b)** the sample median, and **(c)** the sample mode.

(a) The sample mean is $\overline{x}$=(9+16+15+20+30+42)/40=3.05

(b) The sample median is the average of the 20th and 21st smallest values, and is thus equal to 3.

(c) The sample mode is 1, the value that occurred most frequently.

Problems: The mode may not exist. The mode may not be unique.

2.2. Sample Quartiles and Box Plots

Quantiles are also measures of locations.

Quartiles divide the ordered set of data ***into four equal parts,*** the division points are called ***quartiles.***

Q1 – First Quartile – 25% of the observations are smaller than Q1 and 75% of the observations are larger than Q1.

Q2 – Second Quartile – 50% of the observations are smaller than Q2 and 50% of the observations are larger than Q2. Same as the Median. It is also the 50th percentile.

Q3 – Third Quartile – 75% of the observations are smaller than Q3 and 25% of the observations are larger than Q3.

As in the case of median, the quartiles may not be unique.

The quartiles, like the median, either take the value of one of the observations, or the value halfway between two observations.

First Quartile:

$$Q_1 = \begin{cases} x_i & i = \dfrac{n+1}{4}, \text{ n is odd} \\ \dfrac{x_i + x_{i+1}}{2} & i = \dfrac{n}{4}, \text{ n is even} \end{cases}$$

Second Quartile **(Median)**:

$$Q_2 = \begin{cases} x_i & i = \dfrac{n+1}{2}, \text{ n is odd} \\ \dfrac{x_i + x_{i+1}}{2} & i = \dfrac{n}{2}, \text{ n is even} \end{cases}$$

***Third Quartile*:**

$$Q_3 = \begin{cases} x_i & i = \dfrac{3(n+1)}{4}, \text{ n is odd} \\ \dfrac{x_i + x_{i+1}}{2} & i = \dfrac{3n}{4}, \text{ n is even} \end{cases}$$

Example 1:

Original data: 3, 10, 2, 5, 9, 8, 7, 12, 10, 0, 4, 6
Ordered data: 0, 2, 3, 4, 5, 6, 7, 8, 9, 10, 10, 12

Mean: $\bar{x} = \frac{\sum_{i=1}^{12} x_i}{12} = \frac{76}{12} = 6.33$, Mode=10

n=12 (even) for Q1; $i = \frac{12}{4} = 3$ and $Q1 = \frac{x_3 + x_4}{2} = \frac{3+4}{2} = 3.5$ (25% smaller; 75% larger)

For Q2 (Median); $i = \frac{12}{2} = 6$ and $Q2 = \frac{x_6 + x_7}{2} = \frac{6+7}{2} = 6.5$ (50% smaller; 50% larger)

For Q3; $i = \frac{3 \times 12}{4} = 9$ and $Q3 = \frac{x_9 + x_{10}}{2} = \frac{9+10}{2} = 9.5$ (75% smaller; 25% larger)

Example 2: Ordered data: 210, 220, 225, 225, 225, 235, 240, 250, 270, 280

Mean: $\bar{x} = \frac{\sum_{i=1}^{12} x_i}{10} = \frac{2380}{10} = 238$, Mode=225

n=10 (even) for Q1; $i = \frac{10}{4} = 2.5$ and $Q1 = \frac{x_2 + x_3}{2} = \frac{220 + 225}{2} = 222.5$ (25% smaller; 75% larger)

For Q2 (Median); $i = \frac{10}{2} = 5$ and $Q2 = \frac{x_5 + x_6}{2} = \frac{225 + 235}{2} = 230$ (50% smaller; 50% larger)

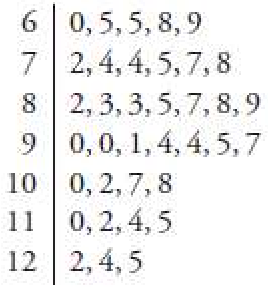
For Q3; $i = \frac{3 \times 10}{4} = 7.5$ and $Q1 = \frac{x_7 + x_8}{2} = \frac{240 + 250}{2} = 245$ (75% smaller; 25% larger)

Example 3: Noise is measured in decibels, denoted as dB. One decibel is about the level of the weakest sound that can be heard in a quiet surrounding by someone with good hearing; a whisper measures about 30 dB; a human voice in normal conversation is about 70 dB; a loud radio is about 100 dB. Ear discomfort usually occurs at a noise level of about 120 dB.

The following data give noise levels measured at 36 different times directly outside of Grand Central Station in Manhattan. Determine the quartiles.

82, 89, 94, 110, 74, 122, 112, 95, 100, 78, 65, 60, 90, 83, 87, 75, 114, 85
69, 94, 124, 115, 107, 88, 97, 74, 72, 68, 83, 91, 90, 102, 77, 125, 108, 65

A stem and leaf plot of the data is as follows:



6	0, 5, 5, 8, 9
7	2, 4, 4, 5, 7, 8
8	2, 3, 3, 5, 7, 8, 9
9	0, 0, 1, 4, 4, 5, 7
10	0, 2, 7, 8
11	0, 2, 4, 5
12	2, 4, 5

The **first quartile** is 76, the average of the 9th and 10th smallest data values; the **second quartile** is 89.5, the average of the 18th and 19th smallest values; the **third quartile** is 104.5, the average of the 27th and 28th smallest values.

n=36 (even) for Q1; $i=\frac{36}{4}=9$ and $Q1=\frac{x_9+x_{10}}{2}=\frac{75+77}{2}=76$ (25% smaller; 75% larger)

For Q2 (Median); $i=\frac{36}{2}=18$ and $Q2=\frac{x_{18}+x_{19}}{2}=\frac{89+90}{2}=89.5$ (50% smaller; 50% larger)

For Q3; $i=\frac{3\times 36}{4}=27$ and $Q1=\frac{x_{27}+x_{28}}{2}=\frac{102+107}{2}=104.5$ (75% smaller; 25% larger)

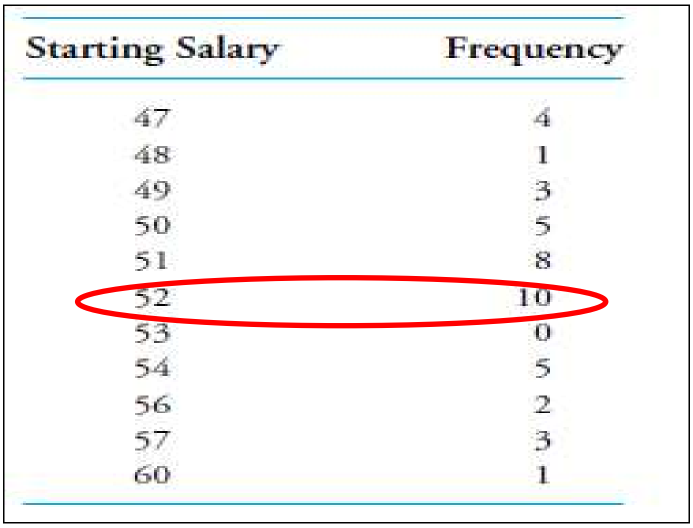
BOX PLOT

A **box plot** ***is often used to plot some of the summarizing statistics of a data set. A straight line segment stretching from the smallest to the largest data value is drawn on a horizontal axis; imposed on the line is a "box," which starts at the first and continues to the third quartile, with the value of the second quartile indicated by a vertical line***

Example:

A data set having a relatively small number of distinct (farklı) values can be conveniently presented in a ***frequency table***. For instance, ***Table 7 is a frequency table*** for a data set consisting of the starting yearly salaries (to the nearest thousand dollars) of 42 recently graduated students with B.S. degrees (dört yıllık üniversite diploması) in electrical engineering.

Table 7. Starting Yearly Salaries



Starting Salary	Frequency
47	4
48	1
49	3
50	5
51	8
52	10
53	0
54	5
56	2
57	3
60	1

The 42 data values presented in Table 7 go from a low value of 47 to a high value of 60. The value of the first quartile ((equal to the average of the 10th and 11th smallest on the list) is 50; the value of the second quartile (equal to the average of the 21st and 22nd smallest values) is 51.5; and the value of the third quartile (equal to the average of the 31st and 32nd smallest on the list) is 53. The box plot for this data set is shown in Figure 6.

n=42 (even) for Q1; $i = \frac{42}{4} = 10.5$ and $Q1 = \frac{x_{10} + x_{11}}{2} = \frac{50+50}{2} = 50$ (25% smaller; 75% larger)

For Q2 (Median); $i = \frac{42}{2} = 21$ and $Q2 = \frac{x_{21} + x_{22}}{2} = \frac{51+52}{2} = 51.5$ (50% smaller; 50% larger)

For Q3; $i = \frac{3 \times 42}{4} = 31.5$ and $Q1 = \frac{x_{31} + x_{32}}{2} = \frac{52+54}{2} = 53$ (75% smaller; 25% larger)

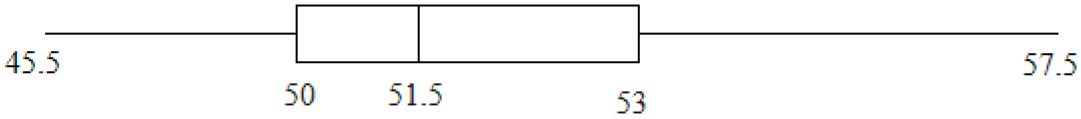


Figure 6. A box plot.

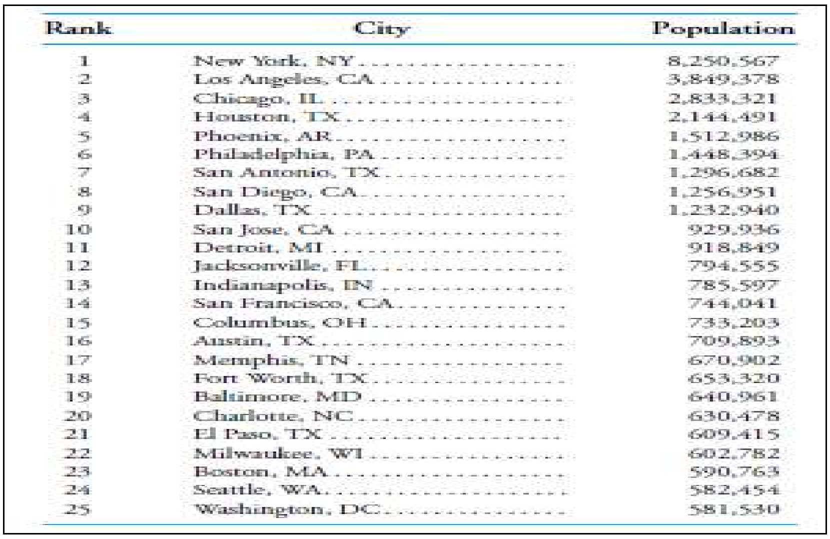
L Fence=Q1-1.5(Q3-Q1)=50-1.5×(53-50)=45.5
U Fence=Q3+1.5(Q3-Q1)=53+1.5×(53-50)=57.5

P.C. Interquartile Range =Q3-Q1

The length of the line segment on the box plot, equal to the Upper Fence minus the Lower Fence value, is called the ***range*** of the data. Also, the length of the box itself, equal to the third quartile (Q_3) minus the first quartile (Q_1), is called the ***interquartile range.***

Example: Table 8 lists the populations of the 25(odd) most populous U.S. cities for the year 1994. For this data set, find **(a)** the sample 25 percentile and **(b)** the sample 75 percentile.

Table 8. *Population of 25 Largest U.S. Cities, July 2006*



Rank	City	Population
1	New York, NY	8,250,567
2	Los Angeles, CA	3,849,378
3	Chicago, IL	2,833,321
4	Houston, TX	2,144,491
5	Phoenix, AR	1,512,986
6	Philadelphia, PA	1,448,394
7	San Antonio, TX	1,296,682
8	San Diego, CA	1,256,951
9	Dallas, TX	1,232,940
10	San Jose, CA	929,936
11	Detroit, MI	918,849
12	Jacksonville, FL	794,555
13	Indianapolis, IN	785,597
14	San Francisco, CA	744,041
15	Columbus, OH	733,203
16	Austin, TX	709,893
17	Memphis, TN	670,902
18	Fort Worth, TX	653,320
19	Baltimore, MD	640,961
20	Charlotte, NC	630,478
21	El Paso, TX	609,415
22	Milwaukee, WI	602,782
23	Boston, MA	590,763
24	Seattle, WA	582,454
25	Washington, DC	581,530

Example: Table 8 lists the populations of the 25(odd) most populous U.S. cities for the year 1994. For this data set, find **(a)** the sample 25 percentile and **(b)** the sample 75 percentile.

(a) Because the sample size is 25(odd) and 26(0.25) =6.5, the sample 25 percentile is average of 6th and 7th values, equal to (1 448 394 +1 296 682)/2=1 372 538

(b) Because 26(0.75)=19.5, the sample 75 percentile is the average of the nineteenth and the twentieth values. Hence, the sample 75 percentile is (640 961 +630 478)/2=635 719.5.

2.3. Measures of Dispersion (Saçılım/Dağılım) / Variability(Değişkenlik) (Range, Interquartile range, Sample Variance, Sample Standard Deviation, Coefficient of Variation)

Whereas we have presented statistics that describe the central tendencies of a data set, we are also interested in ones that describe ***the spread or variability of the data values.***

Dispersion is the amount of spread, or variability, in a set of data. **There are mainly 5 major measures of dispersion:**

2.3.1. Range

Range = Largest Value – Smallest Value

Example: ordered data: 1 2 3 4 8 then Range = 8 – 1 = 7
Problem: The range is influenced by extreme values at either end.

2.3.2. Interquartile Range

IQR = Q3 – Q1

It is basically the range encompassed (kuşatılmış, içine alınmış) by the central 50% of the observations in the distribution. It is less sensitive to extreme values in the sample than is the ordinary sample range.

Problem: The interquartile range does not take into account the variability of the *total* data (only the central 50%). We are “throwing out” half of the data.

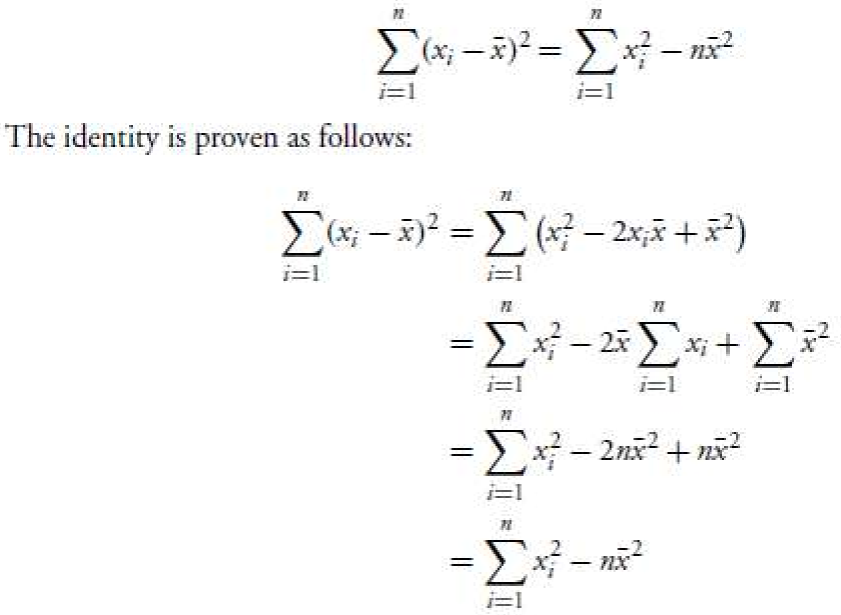
2.3.3. Sample Variance

A statistic that could be used for this purpose would be one that measures the average value of the squares of the distances between the data values and the sample mean. This is accomplished by the **sample variance**, which for technical reasons divides the sum of the squares of the differences by *n*−1 rather than *n*, where *n* is the size of the data set.

The population variance is $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$ but it is very rare that we ever take a census (tamsayım) of the population and deal with N. Normally, we work with a sample and calculate the sample measures, like the sample mean and the sample variance s^2.

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

EXTRA INFORMATION: The reason we divide by n-1 instead of n is to assure that s is an unbiased estimator of σ. We have taken a shortcut: in the second formula, we are using $\bar{x}$, a statistic, instead of μ, a parameter. To correct for this – which has a tendency to understate the true standard deviation – we divide by n-1 which will increase s somewhat and make it an unbiased estimator of σ. Later on in the course we will refer to this as “losing one degree of freedom.”



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

The identity is proven as follows:

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\bar{x})^2&=\sum_{i=1}^{n}\left(x_i^2-2x_i\bar{x}+\bar{x}^2\right)\\&=\sum_{i=1}^{n}x_i^2-2\bar{x}\sum_{i=1}^{n}x_i+\sum_{i=1}^{n}\bar{x}^2\\&=\sum_{i=1}^{n}x_i^2-2n\bar{x}^2+n\bar{x}^2\\&=\sum_{i=1}^{n}x_i^2-n\bar{x}^2\end{aligned}$$

Example 1: Find the sample variances of the data sets **A** and **B** given below.

A: 3, 4, 6, 7, 10   **B:**−20, 5, 15, 24

As the sample mean for data set **A** is $\bar{x}=(3+4+6+7+10)/5=6$

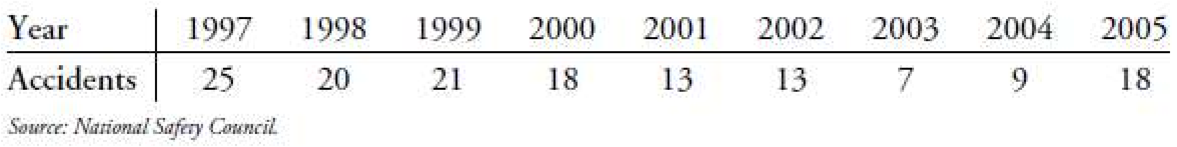
$$s^2=\frac{\sum_{i=1}^{n}x_i^2-n\bar{x}^2}{n-1}=\frac{210-\left(5\times6^2\right)}{5-1}=7.5$$

The sample mean for data set **B** is also 6; its sample variance is

$$s^2=\frac{\sum_{i=1}^{n}x_i^2-n\bar{x}^2}{n-1}=\frac{1226-\left(4\times6^2\right)}{4-1}=360.67$$

Thus, although both data sets have the same sample mean, there is a much greater variability in the values of the **B** set than in the **A** set.

Example 2: The following data give the worldwide number of fatal (ölümcül) airline accidents of commercially scheduled air transports in the years from 1997 to 2005.



Year	1997	1998	1999	2000	2001	2002	2003	2004	2005
Accidents	25	20	21	18	13	13	7	9	18

Source: National Safety Council.

$$s^2=\frac{\sum_{i=1}^{n}x_i^2-n\bar{x}^2}{n-1}=\frac{2582-\left(9\times16^2\right)}{9-1}=34.75$$

2.3.4. Sample Standard Deviation

The positive square root of the sample variance is called the sample standard deviation. The quantity *s*, defined by

$$s=\sqrt{s^2}=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}=\sqrt{\frac{\sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}}{n-1}}=\sqrt{\frac{\sum_{i=1}^{n}x_i^2-n\bar{x}^2}{n-1}}$$

is called the ***sample standard deviation*.** The sample standard deviation is measured in the same units as the data.

If we want to determine the sample mean of a data set that is presented in a frequency table listing the ***k*** distinct values (class/group's value: average of the limits of group) $s_1,\ldots,s_k$ having corresponding frequencies $f_1,\ldots,f_k$. Since such a data set consists of $n=\sum_{i=1}^{k}f_i$ observations, with the value s_i appearing f_i times, for each $i=1,\ldots,k$, it follows that the sample variance and sample standard deviation of these ***n*** data values, respectively:

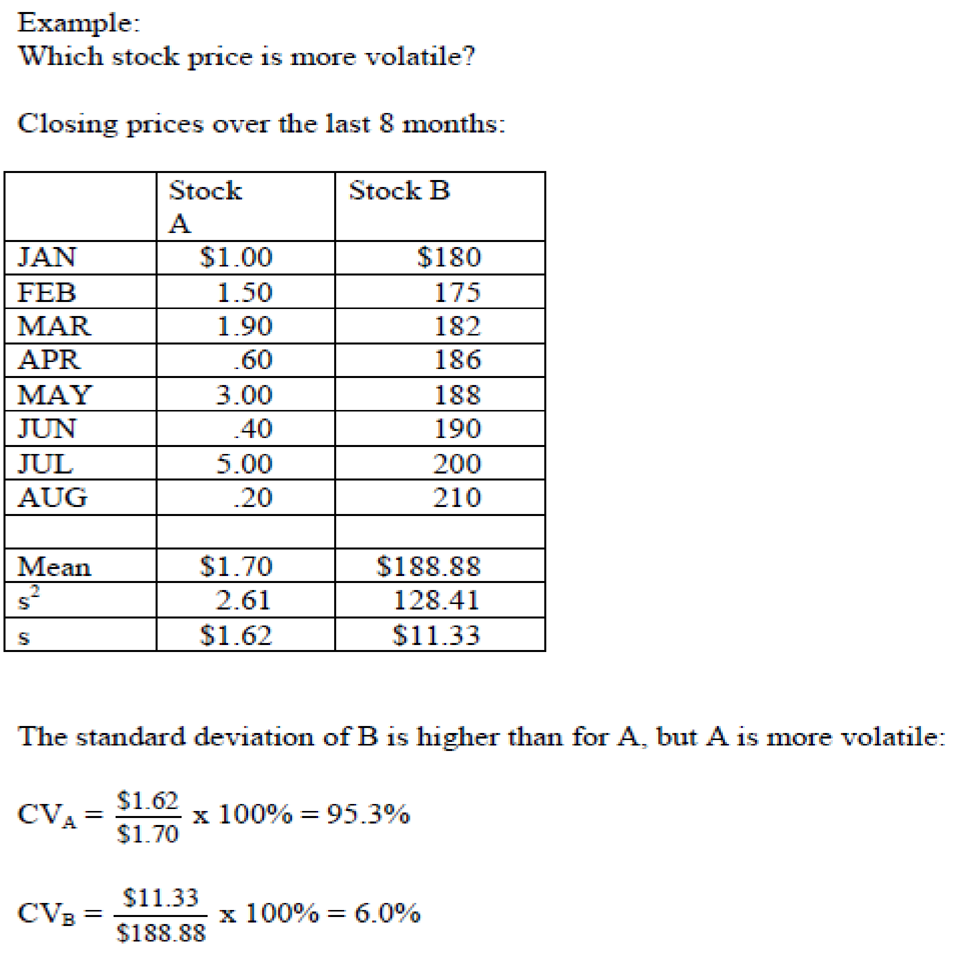
$$s^2=\frac{\sum_{i=1}^{k}f_is_i^2-\frac{\left(\sum_{i=1}^{k}f_is_i\right)^2}{n}}{n-1}$$ and thus sample standart devaiation $s=\sqrt{s^2}$

2.3.5. Coefficient of Variation

The problem with s^2 and s is that they are in the "original" units. This makes it difficult to compare the variability of two data sets, if they are in different units or if the magnitude of the numbers is very different. Suppose you wish to compare two stocks and one is in dollars and the other is in yen; if you want to know which one is more volatile (oynak, değişken), you should use the **coefficient of variation**. It is also not appropriate to compare two stocks of vastly different prices even if both are in the same units. The standard deviation for a stock that sells for around \$300 is going to be very different than one where the price is around \$0.25. The coefficient of variation will be a better measure of dispersion when comparing the two stocks than the standard deviation (see example below).

$$CV=\frac{s}{\bar{x}}\times 100\%$$

Example:



Example:
Which stock price is more volatile?

Closing prices over the last 8 months:

	Stock A	Stock B
JAN	$1.00	$180
FEB	1.50	175
MAR	1.90	182
APR	.60	186
MAY	3.00	188
JUN	.40	190
JUL	5.00	200
AUG	.20	210
Mean	$1.70	$188.88
s^2	2.61	128.41
s	$1.62	$11.33

The standard deviation of B is higher than for A, but A is more volatile:

$$CV_A = \frac{\$1.62}{\$1.70} \text{ x } 100\% = 95.3\%$$

$$CV_B = \frac{\$11.33}{\$188.88} \text{ x } 100\% = 6.0\%$$

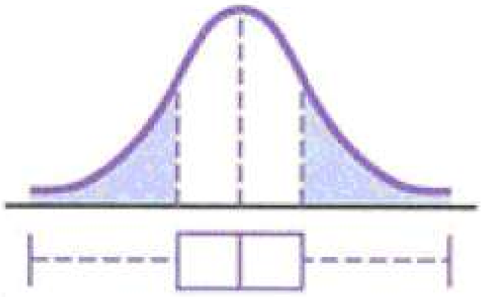
2.4. Measure of Shape

A third important property of data is its shape. **Shape** is the distribution symmetric or skewed? Symmetric distributions are those whose left and right-hand sides look like mirror images of one another (perfect symmetry is a rarity in real life).

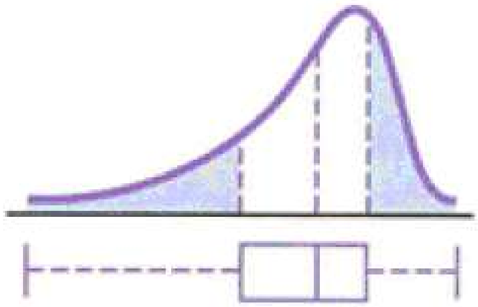
Shape can be described by degree of asymmetry (i.e., skewness).

if a data distribution is skewed right, the mean will be greater than the median.
(mean > median positive or right-skewness)
if a data distribution is perfectly symmetric, the median and mean will be equal.
(mean = median symmetry or zero-skewness)
if a data distribution is skewed left, the mean will be less than the median.
(mean < median negative or left-skewness)

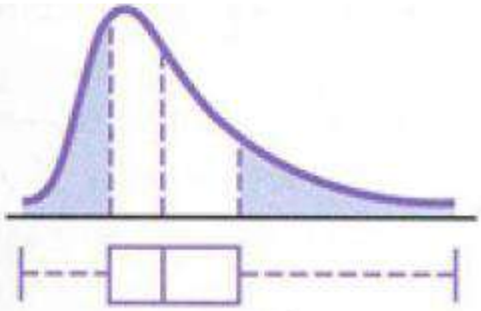
Positive skewness arises when the mean is increased by some unusually high values. Negative skewness occurs when the mean is decreased by some unusually low values.



Panel A
Symmetric distribution



Panel B
Left-skewed distribution



Panel C
Right-skewed distribution

Example:

Hours to complete a task:

2	3	8	8	9	10	10	12	15	18	22	63

$\bar{x} = \frac{180}{12} = 15$ n=12 (even) $i = \frac{n}{2} = \frac{12}{2} = 6$ median $= \frac{x_6 + x_7}{2} = \frac{10+10}{2} = 10$

mean > median positive or right-skewness

$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} = \frac{5568 - \left(12 \times 15^2\right)}{12-1} = 260.72$ $s = \sqrt{s^2} = \sqrt{260.72} = 16.15$

$CV = \frac{s}{\bar{x}} \times 100\% = \frac{16.15}{15} \times 100\% = 107.7$

Example 1: A data shows the scores of the exam which was taken by a group of students.

- **a)** Specify the type of data (scale)
- **b)** Build up frequency table for the data.
- **c)** Explain the highest frequency and the lowest percentile in the table.
- **d)** Draw a suitable plot of data.
- **e)** Find central measurements (mean, median, mode) using from raw data.
- **f)** Find sample mean using frequency table.
- **g)** Find and explain Q1, Q2, Q3 quartiles.
- **h)** Calculate dispersion measurements (variance, standard deviation)
- **i)** Calculate skewness and kurtosis measurements for the data

Table 1. Students' exam scores

14.5	46.6	59.5	70.5	75.5
18.5	48.4	62.4	70.5	77.5
20.6	50.5	63.4	71.0	83.5
25.3	51.5	65.4	71.5	84.0
28.8	54.8	65.5	71.6	84.4
40.6	54.8	66.5	71.8	87.4
42.5	55.0	69.0	72.0	88.5
43.0	56.8	69.9	75.0	92.0
43.5	57.8	70.0	75.3	98.4

$$\sum_{i=1}^{n} x_i = 2765.6 \qquad \sum_{i=1}^{n} x_i^2 = 187277.6 \qquad \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^3 = \text{-198789,9} \qquad \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^4 = 19294316{,}2$$

ANSWERS OF EXAMPLE 1:

- **a)** Quantitative data-(Continuous data) the scale of it is "interval"

- **b)**

Table 2. Frequency table for the students' exam scores data **(answer of b)**

Group No /Class No	Lower Limit (LL))	Upper Limit (UL)	Class Value (s_i)	Frequency (f_i)	Relative Frequency (p_i=f_i/n)	$f_i \times s_i$	$f_i \times s_i^2$
1	14.5	28.4	(14.5+28.4)/2=21.45	4	4/45=0.09	85.80	1840.41
2	28.5	42.4	35.45	2	2/45=0.04	70.90	2513.405
3	42.5	56.4	49.45	10	0.22	494.50	24453.03
4	56.5	70.4	63.45	11	0.24	697.95	44284.93
5	70.5	84.4	77.45	14	0.31	1.084.30	83979.04
6	84.5	98.4	91.45	4	0.09	365.80	33452.41
			TOTAL	**n=45**	**1**	**2799.25**	**190523.2**

- **c)** The highest frequency is 14 that means in this exam the most of the students (or 14 of the students) got the scores between 70.5 and 84.4. The lowest percentile is 0.04 that means in this exam few of the students (or 4% of the students) got the scores between 28.5 and 42.4.

d)

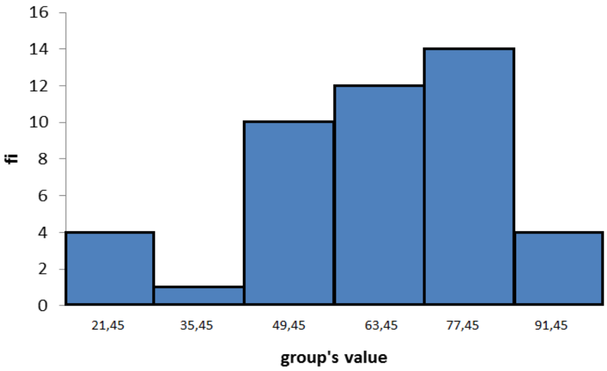


Figure 1. Histogram chart for the data given in Table 2. **(answer of d)**

e) Find central measurements (mean. median. mode) using from raw data.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{2765.6}{45} = 61.46$$

$\tilde{x}' = x_{23} = 65.5$ since for n=45. $i = \frac{n+1}{2} = \frac{45+1}{2} = 23$ if n is odd. That means the half of the students' exam scores are smaller/lower than 65.5 and the other half of the students' exam scores are larger/higher than 65.5. In other words. 50% of the students' exam scores are smaller/lower than 65.5 and 50% of the students' exam scores are larger/higher than 65.5. In other words 50%percent of students got score less than 65.5 and %50 percent of students got score greater than 65.5.

$\hat{x} =$ 54.8 and 70.5 (two times repeated), that is there are two modes.

f) Find sample mean using frequency table:

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i s_i}{n} = \frac{2799.25}{45} = 62.2056$$

g) Find and explain Q1. Q2. Q3 quartiles.

$Q1 = \frac{x_{11} + x_{12}}{2} = \frac{48.4 + 50.5}{2} = 49.45$ for n=45. $i = \frac{n+1}{4} = \frac{45+1}{4} = 11.5$ if n is odd.

Q1 – First Quartile – 25% of the students' exam scores are smaller/lower than 49.45 and 75% of the students' exam scores are larger/higher than 49.45.

$Q2 = \tilde{x}' = x_{23} = 65.5$ for n=45. $i = \frac{n+1}{2} = \frac{45+1}{2} = 23$ if n is odd.

Q2 – Second Quartile- 50% of the students' exam scores are smaller/lower than 65.5 and 50% of the students' exam scores are larger/higher than 65.5. Same as the median.

$Q3 = \frac{x_{34} + x_{35}}{2} = \frac{72 + 75}{2} = 73.5$ for n=45. $i = \frac{3\times(n+1)}{4} = \frac{3\times(45+1)}{4} = 34.5$ if n is odd.

Q3 – Third Quartile-75% of the students’ exam scores are smaller/lower than 73.5 and 25% of the students’ exam scores are larger/higher than 73.5.

h) Calculate dispersion measurements (variance. standard deviation).

For raw data:

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} = \frac{187277.6 - 45\times(61.46)^2}{44} = 393.13$$

$$s = \sqrt{s^2} = \sqrt{393.13} = 19.82750$$

Another formula for s^2 is: $s^2 = \dfrac{\sum_{i=1}^{n} x_i^2 - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}$**, it could also be used and a close value will be obtained.**

For grouped data (from frequency table):

$$s^2 = \frac{\sum_{i=1}^{k} f_i s_i^2 - \dfrac{\left(\sum_{i=1}^{k} f_i s_i\right)^2}{n}}{n-1} = \frac{190523.2 - \dfrac{(2799.25)^2}{45}}{44} = 372.598$$

$$s = \sqrt{s^2} = \sqrt{372.598} = 19.3028$$

i) Calculate skewness and kurtosis measurements for the data.

$skewness = \dfrac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^3}{nS^3} = \dfrac{-198789.9}{45\times\left(19.82750\right)^3} \cong -0.57$ since skewness<0 the distribution of the data is **skewed left.**

$kurtosis = \dfrac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^4}{nS^4} - 3 = \dfrac{19294316.2}{45\times\left(19.82750\right)^4} \cong 2.8 - 3 = -0.2$ since kurtosis<0 the distribution of the data is platykurtic (basık) according to a standard normal distribution.

Kurtosis Value Range

- Normal distribution kurtosis = 0
- A distribution that is more peaked and has fatter tails than normal distribution has kurtosis value greater than 0 (the higher kurtosis, the more peaked and fatter tails). Such distribution is called ***leptokurtic*** or ***leptokurtotic***.
- A distribution that is less peaked and has thinner tails than normal distribution has kurtosis value less than 0. Such distribution is called ***platykurtic*** or ***platykurtotic***.

Example 2: A data set shows the books sales (daily) of a publishing house during a year. The book sales are given for randomly selected 22 days in a year.

Table3. Books sales data.

19	39	58	75	135	195	196	200	235	254	255
286	312	314	356	370	371	373	373	430	433	490

$$\sum_{i=1}^{n} x_i = 5769 \qquad \sum_{i=1}^{n} x_i^2 = 1900063$$

- **a)** Specify the type of data (scale)
- **b)** Build up frequency table for the data.
- **c)** Explain the highest frequency and the lowest percentile in the table.
- **d)** Draw a suitable plot of the data.
- **e)** Find central measurements (mean. median. mode) using from raw data.
- **f)** Find sample mean using the frequency table.
- **g)** Find and explain Q1. Q2. Q3 quartiles.
- **h)** Calculate dispersion measurements (variance. standard deviation)

ANSWERS OF EXAMPLE 2:

- **a)** Quantitative data-(Discrete data) the scale of it is “ratio”.
- **b)**

Table 4. Frequency table for the books sales data **(answer of b).**

Group No /Class No	Upper Limit (UL)	Lower Limit (LL)	Class Value (s_i)	Frequency (f_i)	Relative Frequency ($p_i=f_i/n$)	$f_i \times s_i$	$f_i \times s_i^2$
1	19	77	(19+77)/2=96/2=48	4	0.18	192	9216
2	78	136	(78+136)/2=214/2=107	1	0.05	107	11449
3	137	195	166	1	0.05	166	27556
4	196	254	225	4	0.18	900	202500
5	255	313	284	3	0.14	852	241968
6	314	372	343	4	0.18	1372	470596
7	373	431	402	3	0.14	1206	484812
8	432	490	461	2	0.09	922	425042
			Total	**n=22**	**1**	**5717**	**1873139**

c) The highest frequency is 4 but it has seen three times. That means 4 of the days randomly selected in a year number of books sales are between 19 and 77. Another 4 of the days randomly selected in a year the number of books sales are between 196 and 254. Another 4 of the days randomly selected in a year the number of books sales are between 314 and 372. The lowest percentile is 0.05 but it has seen two times. That means 5% of the days for randomly selected 22 days in a year the number of books sales are between 78 and 136. Another 5% of the days for randomly selected 22 days in a year the number of books sales are between 137 and 195.

d)

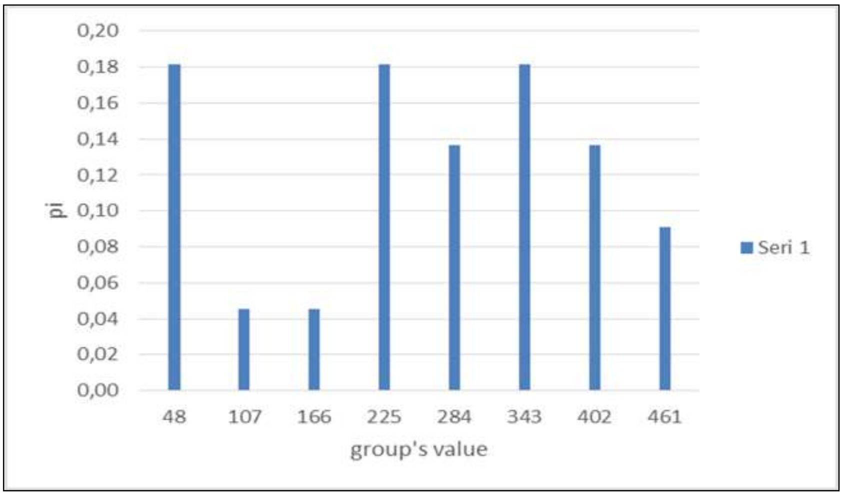


Figure 2. Line chart for the data given in Table 4 **(answer of d).**

e) Find central measurements (mean. median. mode) using from raw data.

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{5769}{22} = 262.2272$$

$\overline{x}' = \frac{x_{11} + x_{12}}{2} = \frac{255+286}{2} = 270.5 \cong 271$ for n=22. $i = \frac{n}{2} = \frac{22}{2} = 11$ if n is even. That means 11 of the days randomly selected in a year. the number of books sales are lower than 270.5 and the other 11 of the days randomly selected in a year. number of books sales are higher than 270.5. In other words. 50% of the days number of books sales are lower than 270.5 and 50% of the days number of books sales are higher than 270.5.

$\hat{x} =$ 373 (observed 2 times)

f) Find sample mean using frequency table.

$$\overline{x} = \frac{\sum_{i=1}^{k} f_i s_i}{n} = \frac{5717}{22} = 259.8636$$

g) Find and explain Q1. Q2. Q3 quartiles.

$Q1 = \frac{x_5 + x_6}{2} = \frac{135+195}{2} = 165$ for n=22. $i = \frac{n}{4} = \frac{22}{4} = 5.5$ if n is even.

<u>Q1 – First Quartile</u> – 25% of the days number of books sales are lower than 165 and 75% of the days number of books sales are higher than 165.

$Q2 = \overline{x}' = \frac{x_{11} + x_{12}}{2} = \frac{255+286}{2} = 270.5$ for n=22. $i = \frac{n}{2} = \frac{22}{2} = 11$ if n is even

<u>Q2 –Second Quartile</u> – 50% of the days number of books sales are lower than 270.5 and 50% of the days number of books sales are higher than 270.5.

$Q3 = \frac{x_{16} + x_{17}}{2} = \frac{370 + 371}{2} = 370.5$ for n=22. $i = \frac{3 \times n}{4} = \frac{3 \times 22}{4} = 16.5$ if n is even.

Q3 – Third Quartile - 75% of the days number of books sales are lower than 370.5 and 25% of the days number of books sales are higher than 370.5.

h) Calculate dispersion measurements (variance. standard deviation).

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} = \frac{1900063 - 22\times(262.2273)^2}{21} = 18441.6155$$

$$s = \sqrt{s^2} = \sqrt{18411.6155} = 135.79989$$

$$s^2 = \frac{\sum_{i=1}^{k} f_i s_i^2 - \frac{\left(\sum_{i=1}^{k} f_i s_i\right)^2}{n}}{n-1} = \frac{1873139 - \frac{(5717)^2}{22}}{21} = 18452.31385 \qquad s=\sqrt{s^2}=\sqrt{18452.31385}=135.8393$$

Example 3: The distribution of computers in lab 1 is given in below. Create cumulative frequency and cumulative relative frequency columns and draw pie chart for data.

Table 5. Frequency table for the computers in lab 1.

speed	f_i	p_i	Cumulative Frequency (F_i)	Cumulative Relative Frequency (P_i)	Angles
low	5	0.25	5	0.25	360×0.25=90º
medium	6	0.30	11	0.55	360×0.30=108º
high	9	0.45	20	1.00	360×0.45=162º
Total	**20**	**1.00**			**360**

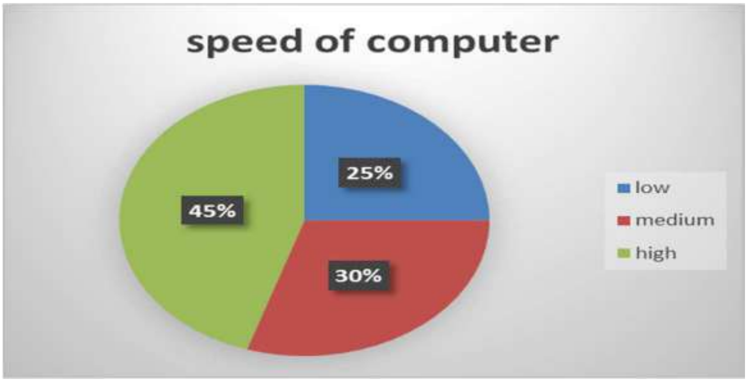


Figure 3. Pie graph for the data given in Table 5.

PERSONAL STUDY QUESTIONS

Example 4: A data set shows the amount of boron reserves (gr/1000) in computer processors (bilgisayar işlemcileri) produced by a producer.

a) Specify the type of data (scale)
b) Build up frequency table for the data.
c) Explain the highest frequency and the lowest percentile in the table.
d) Draw a suitable plot of data.
e) Find central measurements (mean. median. mode) using from raw data.
f) Find sample mean using frequency table.
g) Find and explain Q1. Q2. Q3 quartiles.
h) Calculate dispersion measurements (variance. standard deviation)
i) Calculate skewness and kurtosis measurements for the data

Table 6. Boron reserves data.

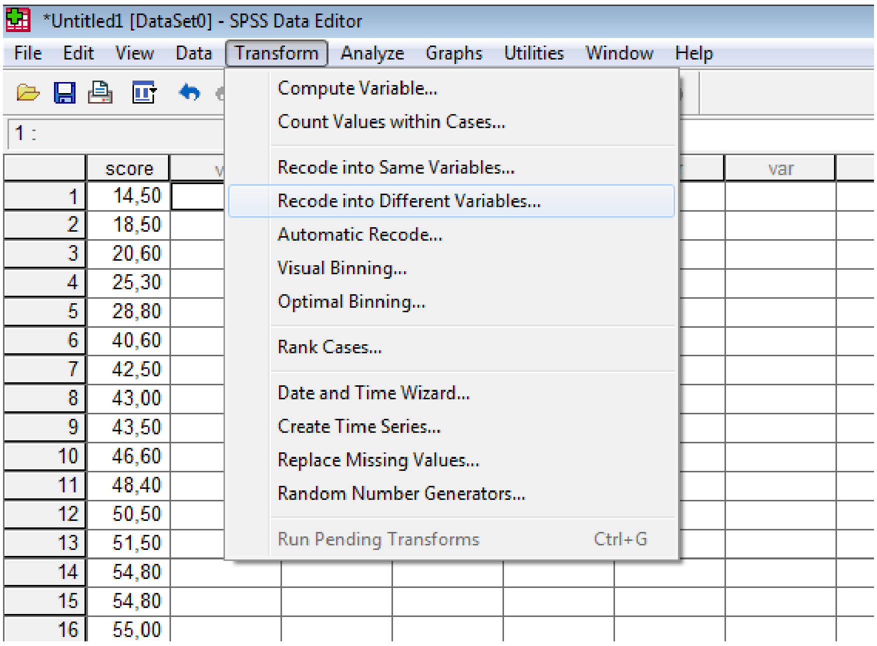
83.4	88.5	90.4	92.6
84.3	88.6	90.6	92.7
87.5	89.0	90.7	93.0
87.8	89.2	90.9	93.7
87.9	89.6	91.0	94.4
88.2	89.7	91.2	94.7
88.3	89.9	91.6	96.5
88.3	90.1	91.8	98.8
88.4	90.4	92.2	

Example 5: A research is done for describing the computer users (for which purpose/purposes they use computers). 263 computer users attended to this research and 112 of them stated that they use computer for playing games. 57 of them using computer for utilizing internet. 82 of them using computer for their work. 12 of them using computer for mixed purpose (playing games. work. internet etc.). Build up frequency table for the data. Create cumulative frequency and cumulative relative frequency columns and draw pie chart for data.

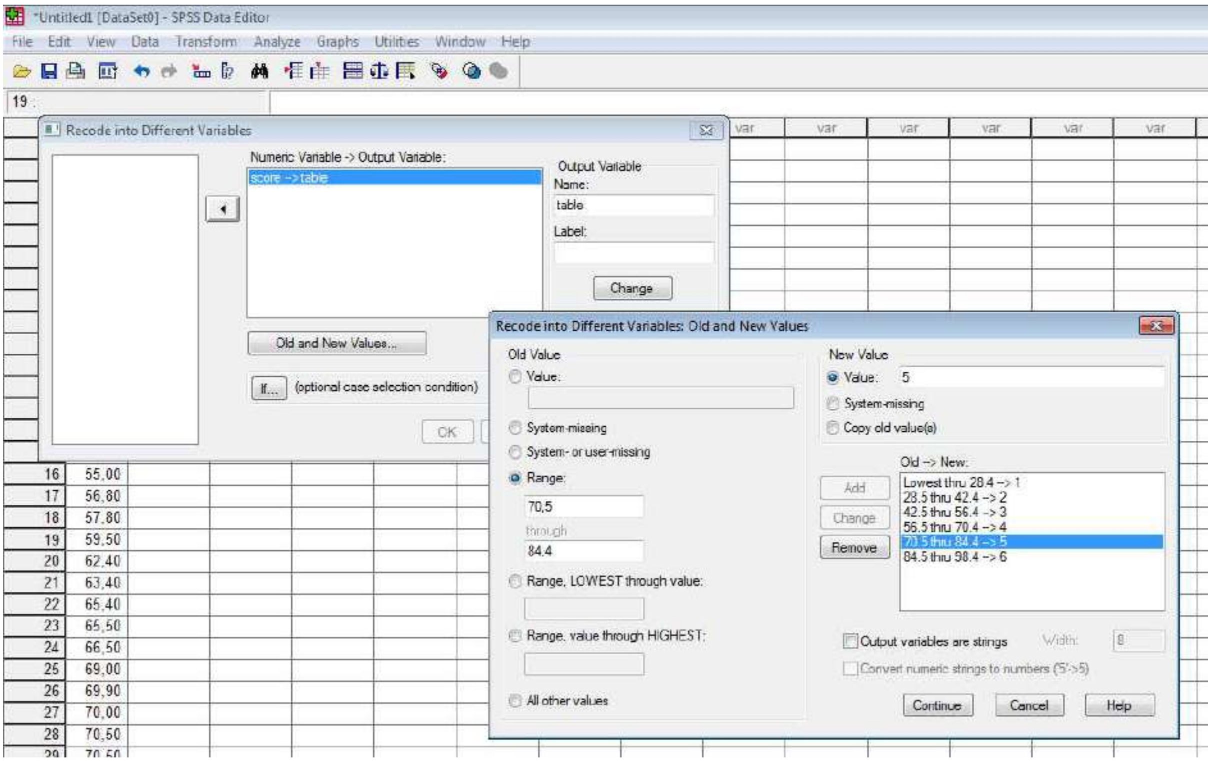
Table 7. Frequency table for describing computer users.

Purpose of Computer Usage	f_i	p_i	**Cumulative Frequency (F_i)**	**Cumulative Relative Frequency (P_i)**	**Angles**
Game					
Internet					
Work					
Mixed Purpose					
Total					

APPLICATION WITH SPSS


*Untitled1 [DataSet0] - SPSS Data Editor
File
Edit
View
Data
Transform
Analyze
Graphs
Utilities
Window
Help
1 :
Compute Variable...
Count Values within Cases...
Recode into Same Variables...
Recode into Different Variables...
Automatic Recode...
Visual Binning...
Optimal Binning...
Rank Cases...
Date and Time Wizard...
Create Time Series...
Replace Missing Values...
Random Number Generators...
Run Pending Transforms Ctrl+G
var

	score
1	14,50
2	18,50
3	20,60
4	25,30
5	28,80
6	40,60
7	42,50
8	43,00
9	43,50
10	46,60
11	48,40
12	50,50
13	51,50
14	54,80
15	54,80
16	55,00


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Old and New Values...
If...
OK
Change
Add
Change
Remove
Continue
Cancel
Help

19 :

Recode into Different Variables

Numeric Variable -> Output Variable:
score --> table

Output Variable
Name: table
Label:

(optional case selection condition)

Recode into Different Variables: Old and New Values

Old Value
○ Value:
○ System-missing
○ System- or user-missing
● Range:
70,5
through
84.4
○ Range, LOWEST through value:
○ Range, value through HIGHEST:
○ All other values

New Value
● Value: 5
○ System-missing
○ Copy old value(s)

Old --> New:
Lowest thru 28.4 --> 1
28.5 thru 42.4 --> 2
42.5 thru 56.4 --> 3
56.5 thru 70.4 --> 4
70.5 thru 84.4 --> 5
84.5 thru 98.4 --> 6

□ Output variables are strings Width: 8
□ Convert numeric strings to numbers ('5'->5)

16	55,00
17	56,80
18	57,80
19	59,50
20	62,40
21	63,40
22	65,40
23	65,50
24	66,50
25	69,00
26	69,90
27	70,00
28	70,50
29	70,50

Analyze→Descritive statistics→Frequencies

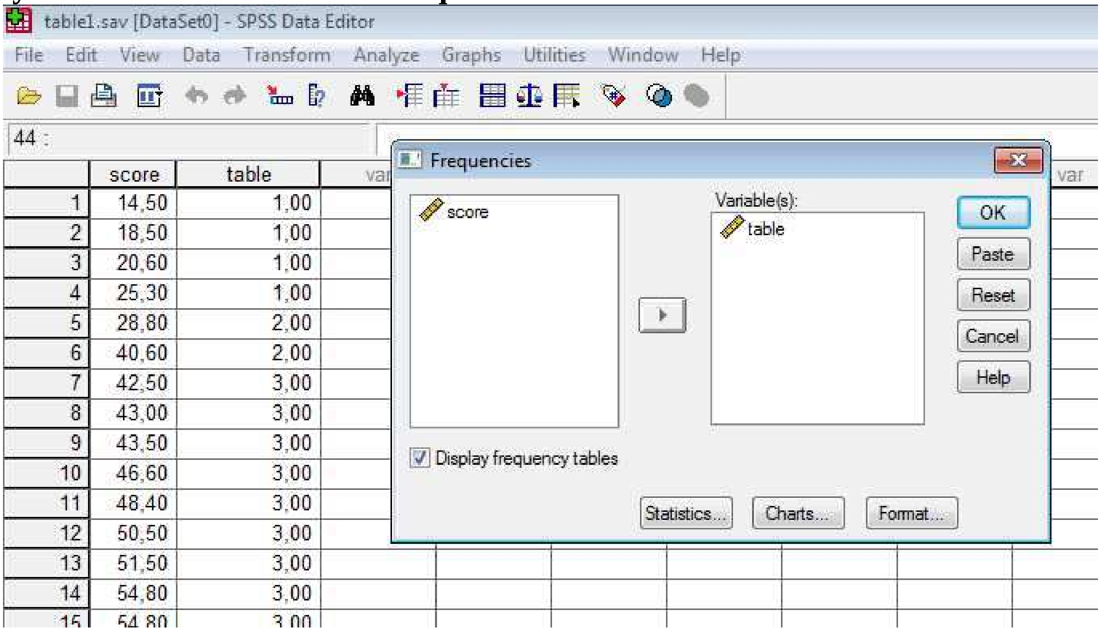


table1.sav [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
OK
Paste
Reset
Cancel
Help
Statistics...
Charts...
Format...

44 :

	score	table
1	14,50	1,00
2	18,50	1,00
3	20,60	1,00
4	25,30	1,00
5	28,80	2,00
6	40,60	2,00
7	42,50	3,00
8	43,00	3,00
9	43,50	3,00
10	46,60	3,00
11	48,40	3,00
12	50,50	3,00
13	51,50	3,00
14	54,80	3,00
15	54,80	3,00

Frequencies

score

Variable(s):
table

☑ Display frequency tables

Statistics

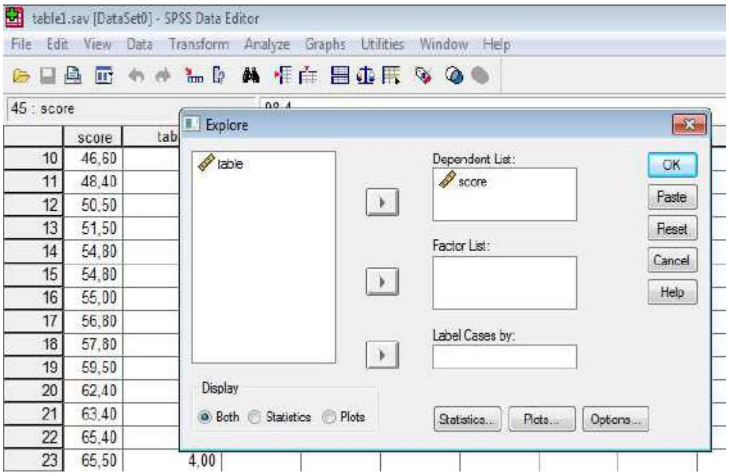
table

N	Valid	45
	Missing	0

table

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	4	8.9	8.9	8.9
	2.00	2	4.4	4.4	13.3
	3.00	10	22.2	22.2	35.6
	4.00	11	24.4	24.4	60.0
	5.00	14	31.1	31.1	91.1
	6.00	4	8.9	8.9	100.0
	Total	45	100.0	100.0	

Analyze→Descriptive statistics→Explore…



Dependent List: score

Display: ● Both ○ Statistics ○ Plots

Descriptives

			Statistic	Std. Error
score	Mean		61,4556	2,95781
	95% Confidence Interval for Mean	Lower Bound	55,4945	
		Upper Bound	67,4166	
	5% Trimmed Mean		62,0944	
	Median		65,5000	
	Variance		393,688	
	Std. Deviation		19,84156	
	Minimum		14,50	
	Maximum		98,40	
	Range		83,90	
	Interquartile Range		24,05	
	Skewness		-,605	,354
	Kurtosis		,027	,695

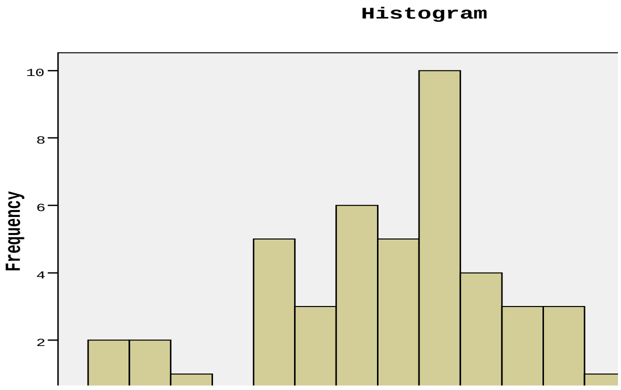
Note: SPSS is using modified the kurtosis formula, so $kurtosis = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^4}{nS^4} - 3$ **gives a different result from SPSS output.**

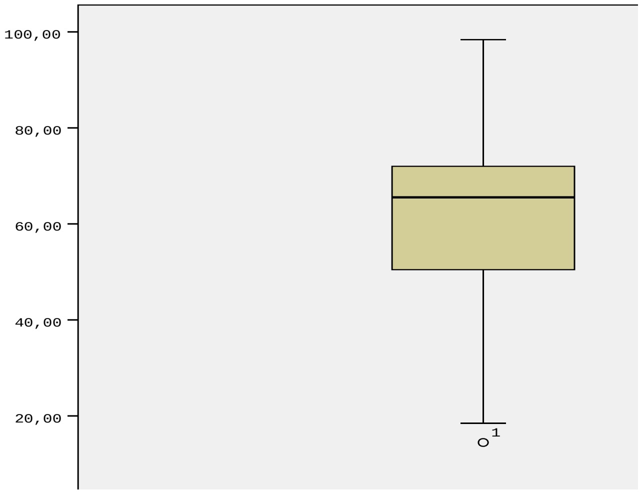
```
score Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     1,00 Extremes    (=<15)
     1,00        1 .  8
     3,00        2 .  058
      ,00        3 .
     6,00        4 .  023368
     8,00        5 .  01445679
     7,00        6 .  2355699
    12,00        7 .  000111125557
     5,00        8 .  34478
     2,00        9 .  28

 Stem width:     10,00
 Each leaf:       1 case(s)
```





İST292 STATISTICS LESSON 3
NORMAL DISTRIBUTION AND CENTRAL LIMIT THEOREM

3. NORMAL DISTRIBUTION

Many statistical phenomen are modelled by normal distribution. For example, human charecteristics such as height, weight, strength; the speed of any particule in gas, errors in measument of quantities. It has bell-shaped symetric curve and the probability is interpreted as "area under the curve".

Characteristics of the Normal Distribution

- Symmetric, bell shaped
- Continuous for all values of X between -∞ and ∞ so that each conceivable (possible) interval of real numbers has a probability other than zero.
- X random variable is defined as $-\infty \leq x \leq \infty$
- Two parameters, μ and σ. Note that the normal distribution is actually a family of distributions, since μ and σ determine the shape of the distribution.
- The rule for a normal density function is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty$$

- The notation $N(\mu, \sigma^2)$ means normally distributed with mean μ and variance σ^2. If we say $X \sim N(\mu, \sigma^2)$ we mean that X is distributed $N(\mu, \sigma^2)$.
- About 2/3 of all cases fall within one standard deviation of the mean, that is

  $P(\mu - \sigma \leq X \leq \mu + \sigma) = .6827$.
- About 95% of cases lie within 2 standard deviations of the mean, that is

  $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = .9545$

Why is the normal distribution useful?

- Many things actually are normally distributed, or very close to it. For example, height and intelligence are approximately normally distributed; measurement errors also often have a normal distribution
- The normal distribution is easy to work with mathematically. In many practical cases, the methods developed using normal theory work quite well even when the distribution is not normal.

- There is a very strong connection between the size of a sample n and the extent to which a sampling distribution approaches the normal form. Many sampling distributions based on large N can be approximated by the normal distribution even though the population distribution itself is definitely not normal.

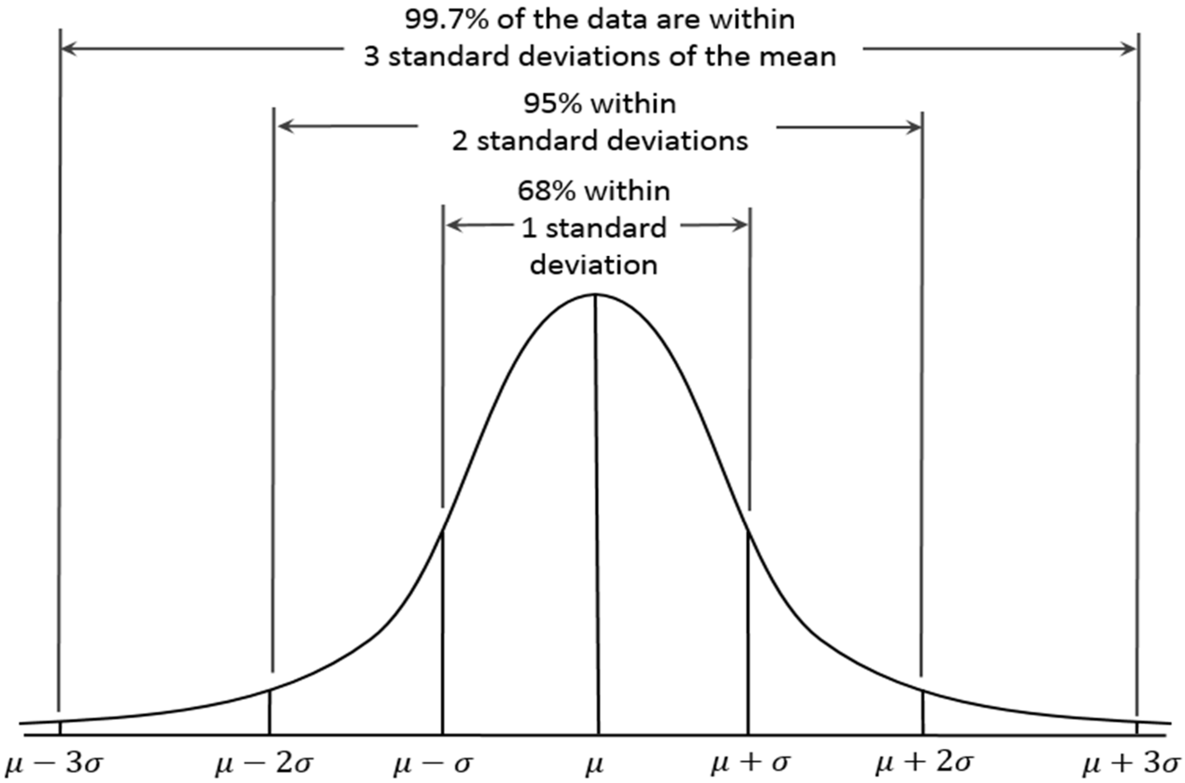


Figure 3.1. Empirical rule of normal distribution.

For the normal distribution, the values less than one standard deviation away from the mean account for 68.27% of the set; while two standard deviations from the mean account for 95.45%; and three standard deviations account for 99.73%. (https://en.wikipedia.org/wiki/Normal_distribution)

Standart Normal Distribution

Standart normal distribution is a spacial case of the normal distribution. If X has a normal disribution (X~N(μ, σ^2)), $Z = \frac{X - \mu}{\sigma}$ will have a standard normal distribution (Z~N(0, 1)). Every normal random variable X can be transformed into a z score via the following equation:

$Z = \frac{X - \mu}{\sigma}$ and we call that standardized random variable.

Example: Find P(Z ≤ a) for a = 1.65, -1.65, 1.0, -1.0

To solve: for positive values of a, look up and report the value for Φ (a) given in SND Table. For negative values of a, look up the value for Φ (-a) (i.e. Φ (absolute value of a)) and report 1 - Φ (-a).

P(Z ≤ 1.65) = Φ (1.65) = .95

P(Z ≤ -1.65) = Φ (-1.65) = 1 - Φ (1.65) = .05

Example: Find a for P(Z ≤ a) = .6026, .9750, .3446

To solve: for p ≥ .5, find the probability value in Table I, and report the corresponding value for Z. For p < .5, compute 1 - p, find the corresponding Z value, and report the negative of that value, i.e. -Z.

P(Z ≤ .26) = .6026

P(Z ≤ 1.96) = .9750

P(Z ≤ -.40) = .3446 (since 1 - .3446 = .6554 = Φ (.40))

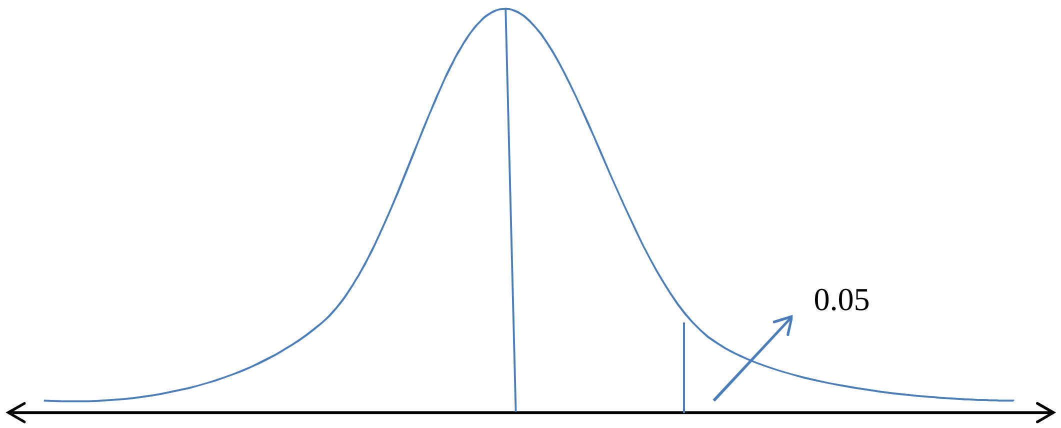
EXAMPLES

1. The top 5% of applicants (as measured by GRE scores) will receive scholarships. If GRE ~ N(500, 100^2), how high does your GRE score have to be to qualify for a scholarship?

Solution: X ~ N(500, 100^2) and $P(X > a) = 0.05$

$$P\left(\frac{X-\mu}{\sigma} > \frac{a-\mu}{\sigma}\right) = P\left(\frac{X-500}{100} > \frac{a-500}{100}\right) = 0.05$$

$$P\left(Z > \frac{a-500}{100}\right) = 1 - \Phi\left(\frac{a-500}{100}\right) = 0.05$$



$$\frac{a-500}{100}=1.65 \Rightarrow a=665$$

2. Family income ~ N(\$25000, \$10000^2). If the poverty (fakirlik) level is \$10,000, what percentage of the population lives in poverty?

Solution: Let X = Family income. We want to find P(X ≤ \$10,000). This is too hard to compute directly, so let

Z = (X - \$25,000)/\$10,000.

If x = \$10,000, then z = (\$10,000 - \$25,000)/\$10,000 = -1.5. So,

P(X ≤ \$10,000) = P(Z ≤ -1.5) = Φ (-1.5) = 1 - Φ (1.5) = 1 - .9332 = .0668. Hence, a little under 7% of the population lives in poverty.

3. A new tax law is expected to benefit “middle income” families, those with incomes between \$20,000 and \$30,000. If Family income ~ N(\$25000, \$10000^2), what percentage of the population will benefit from the law?

Solution: Let X = Family income. We want to find P(\$20,000 X ≤ \$30,000). To solve, let

Z = (X - \$25,000)/\$10,000.

Note that when x = \$20,000, z = (\$20,000 - \$25,000)/\$10,000 = -0.5, and when x = \$30,000, z = +0.5. Hence,

P(\$20,000 ≤ X ≤ \$30,000) = P(-.5 ≤ Z ≤.5) = 2Φ (.5) - 1 = 1.383 - 1 = .383. Thus, about 38% of the taxpayers will benefit from the new law.

Normal Data Set

Many of the large data sets observed in practice have histograms that are similar in shape. These histograms often reach their peaks at the sample median and then decrease on both sides of this point in a bell-shaped symmetric fashion. Such data sets are said to be normal and their histograms are called normal histograms. Figure 3.2 is the histogram of a normal data set.

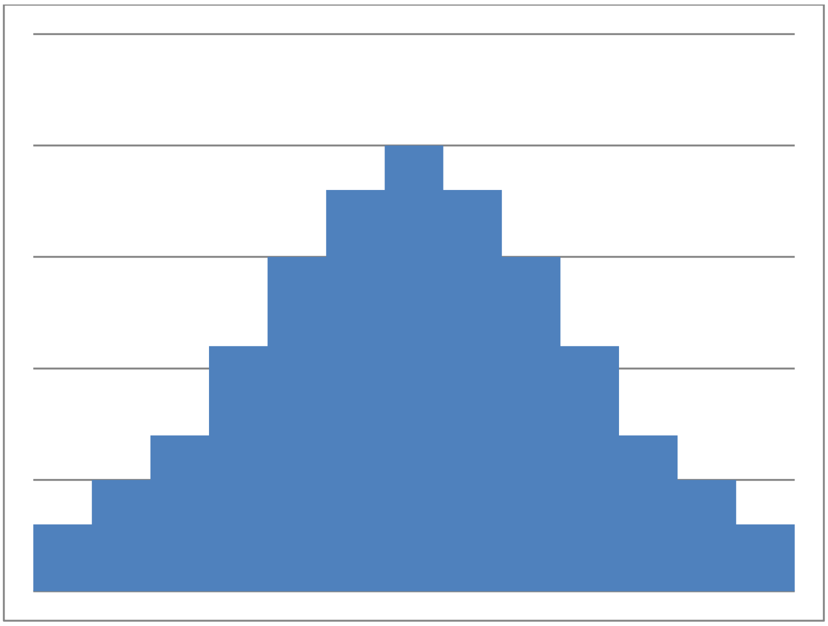
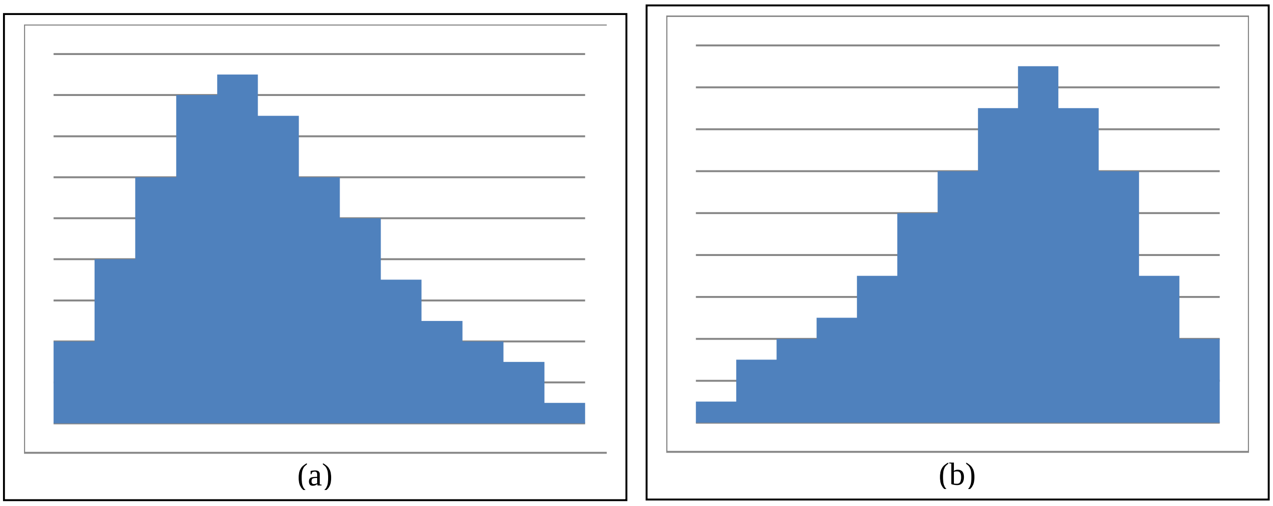


Figure 3.2. Histogram of a normal data set.



(a)

(b)

Figure 3.3. (a) Histogram of a data set skewed to the right. (b) Histogram of a data set skewed to the left.

The histogram in the Figure 3.3 (a) is called as right skewed distribution and (b) is called as left skewed distribution.

Distribution of Sample Mean $\bar{X}$

Suppose a population has normal distrbution with μ and σ^2 parameters. Suppose a sample of n independent measurements is selected from this population:

$X_1, X_2, ..., X_n \sim N(\mu, \sigma^2)$.

Sample mean (sample statistic) $\frac{\sum_{i=1}^{n} X_i}{n}$ is a random variable hence it has probability distribution and so $E(\bar{X})$ and $V(\bar{X})$ are found as follow:

$E(\bar{X}) = \frac{1}{n} E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\left[\sum_{i=1}^{n} E(X_i)\right] = \mu$, (it is not required that $X_1, X_2, ..., X_n$ are independent)

$V(\bar{X}) = \frac{1}{n^2} V\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\left[\sum_{i=1}^{n} V(X_i)\right] = \frac{\sigma^2}{n}$, if $X_1, X_2, ..., X_n$ are independent.

CENTRAL LIMIT THEOREM

Suppose a population has the uniform distribution given in Figure 3.4. The mean and standard deviation of this probability distribution are μ=.5, and σ=.29 (σ^2=0.08333333). Now suppose a sample of 11 measurements is selected from this population. Describe the sampling distribution of the sample mean $\bar{X}$ based on the 1000 sampling experiments.

Central Limit Theorem: For large sample size, the mean $\bar{X}$ of the a sample from a population with mean μ and standard deviation σ possesses a sampling distribution that is approximately normal-regardless of the probability distribution of the sampled population. For the large the sample size, the approximation will be better.

From the theorem, the distibution of $\bar{X}$,

$$n \to \infty \quad \bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

where the term of ∞ is coresponded to $n \geq 30$. Using the result of theorem, the standardized $\bar{X}$ has a standard normal distibution:

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim \mathrm{N}(0,1).$$

This theoren can be enlarged for sample statistic $\sum_{i=1}^{n} X_i$, when $n \to \infty$

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}} \sim \mathrm{N}(0,1).$$

Example: The life of electronic devices has a distribution with mean 500 hours and standard deviation 80. Suppose a system used 100 devises. Find probability if the expected life of system is at least 490 and at most 510.

$$P(490 < X < 510) = \mathrm{P}\left(\tfrac{490-500}{80/10} < Z < \tfrac{510-500}{80/10}\right) = P(-1.25 < Z < 1.25) = 2\times 0.3944 = 0.7888$$

Example: From past experience, it is known that the number of tickets purchased by a student standing in line at the ticket window for the football match of UCLA against USC follows a distribution that has mean μ= 2:4 and standard deviation σ= 2:0. Suppose that few hours before the start of one of these matches there are 100 eager students standing in line to purchase tickets. If only 250 tickets remain, what is the probability that all 100 students will be able to purchase the tickets they desire?

We are given that μ = 2.4; σ = 2; n = 100. There are 250 tickets available, so the 100 students will be able to purchase the tickets they want if all together ask for less than 250 tickets.

The probability for that is $P(T \le 250) = P(\mathrm{Z} \le \tfrac{250-100\times 2.4}{2\sqrt{100}}) = P(Z \le 0.5) = 0.6915$

EXAMPLES

1. A large freight elevator can transport a maximum of 9800 pounds. Suppose a load of cargo containing 49 boxes must be transported via the elevator. Experience has shown that the weight of boxes of this type of cargo follows a distribution with mean μ=205 pounds and standard deviation σ=15 pounds. Based on this information, what is the probability that all 49 boxes can be safely loaded onto the freight elevator and transported?

2. The amount of mineral water consumed by a person per day on the job is normally distributed with mean 19 ounces and standard deviation 5 ounces. A company supplies its employees with 2000 ounces of mineral water daily. The company has 100 employees.

a) Find the probability that the mineral water supplied by the company will not satisfy the water demanded by its employees.

b) Find the probability that in the next 4 days the company will not satisfy the water demanded by its employees on at least 1 of these 4 days. Assume that the amount of mineral water consumed by the employees of the company is independent from day to day.

c) Find the probability that during the next year (365 days) the company will not satisfy the water demanded by its employees on more than 15 days.

Useful Notes

It was investigated first in the 18th century when the scientists observed an astonishing (hayret verici) degree of regulatity in error of measurement. They found the patterns that they observed could be closely approximated by countinous curves, which they referred to as “normal curves of errors” and attributed (atfedilmiş) to the laws of chance (Miller and Miller, 2004 “John E. Freund’s Math. Stat. with Applications). It is also called the "Gaussian curve" after the mathematician Karl Friedrich Gauss.

“The importance of the normal curve stems primarily from the fact that the distributions of many natural phenomena are at least approximately normally distributed. One of the first applications of the normal distribution was to the analysis of errors of measurement made in astronomical observations, errors that occurred because of imperfect instruments and imperfect observers. Galileo in the 17th century noted that these errors were symmetric and that small errors occurred more frequently than large errors. This led to several hypothesized distributions of errors, but it was not until the early 19th century that it was discovered that these errors followed a normal distribution. Independently, the mathematicians Adrain in 1808 and Gauss in 1809 developed the formula for the normal distribution and showed that errors were fit well by this distribution.

This same distribution had been discovered by Laplace in 1778 when he derived the extremely important central limit theorem, the topic of a later section of this chapter. Laplace showed that even if a distribution is not normally distributed, the means of repeated samples from the distribution would be very nearly normally distributed, and that the larger the sample size, the closer the distribution of means would be to a normal distribution.

Most statistical procedures for testing differences between means assume normal distributions. Because the distribution of means is very close to normal, these tests work well even if the original distribution is only roughly normal.

Quételet was the first to apply the normal distribution to human characteristics. He noted that characteristics such as height, weight, and strength were normally distributed. “ (http://onlinestatbook.com/2/normal_distribution/history_normal.html)

Yararlı linkler

http://stattrek.com/probability-distributions/standard-normal.aspx?tutorial=stat

CENTRAL LIMIT THEOREM

Central Limit Theorem: For large sample size, the mean $\bar{X}$ of the a sample from a population with mean μ and standard deviation σ possesses a sampling distribution that is approximately normal-regardless of the probability distribution of the sampled population. For the large the sample size, the approximation will be better.

From the theorem, the distibution of $\bar{X}$,

$$n \to \infty \quad \bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

where the term of ∞ is corresponded to $n \geq 30$. Using the result of theorem, the standardized $\bar{X}$ has a standard normal distibution:

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim \mathrm{N}(0,1).$$

In addition, the distribution of $\sum_{i=1}^{n} X_i$,

$$n \to \infty \quad \sum_{i=1}^{n} X \sim \mathrm{N}\left(n\mu,\, n\sigma^2\right).$$

And then the standardized $\sum_{i=1}^{n} X_i$ has a standard normal distibution:

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}} \sim \mathrm{N}(0,1).$$

Example: The life of electronic devices has a distribution with mean 500 hours and standard deviation 80. Suppose a system used 100 devices. Find probability if the expected life of system is at least 490 and at most 510.

$$P(490 < \bar{X} < 510) = \mathrm{P}\left(\tfrac{490-500}{80/10} < Z < \tfrac{510-500}{80/10}\right) = P(-1.25 < Z < 1.25) = 2 \times 0.3944 = 0.7888$$

Example: From past experience, it is known that the number of tickets purchased by a student standing in line at the ticket window for the football match of UCLA against USC follows a distribution that has mean μ= 2:4 and standard deviation σ= 2:0. Suppose that few hours before the start of one of these matches there are 100 eager students standing in line to purchase tickets. If only 250 tickets remain, what is the probability that all 100 students will be able to purchase the tickets they desire?

We are given that $\mu = 2.4$; $\sigma = 2$; $n = 100$. There are 250 tickets available, so the 100 students will be able to purchase the tickets they want if all together ask for less than 250 tickets.

The probability for that is $P(T \leq 250) = P(\mathrm{Z} \leq \frac{250-100\times2.4}{2\sqrt{100}}) = P(Z \leq 0.5) = 0.6915$

EXAMPLES

1. A large freight elevator (nakliye/yük asansörü) can transport a maximum of 9800 pounds. Suppose a load of cargo containing 49 boxes must be transported via the elevator. Experience has shown that the weight of boxes of this type of cargo follows a distribution with mean μ=205 pounds and standard deviation σ=15 pounds. Based on this information, what is the probability that all 49 boxes can be safely loaded onto the freight elevator and transported?

Solution:

X: Weight of each box

$\mathrm{X} \sim \text{with } \mu = 205, \sigma^2 = 15^2 = 225$

When we load all boxes (49) onto the elevator:

$$\sum \mathrm{X_i} \sim \mathrm{N}(49\times205,\ 49\times225) \qquad \mathrm{E}(\sum \mathrm{X_i}) = \mathrm{n}\mu$$

$$\mathrm{n} \to \infty \ \ \mathrm{N}(10045, 11025) \qquad \mathrm{V}(\sum \mathrm{X_i}) = \mathrm{n}\sigma^2$$

$$(\mathrm{n} > 30)$$

$$\underset{\substack{\downarrow\\ \text{total weight}}}{P\left(\sum_{i=1}^{49} X_i < 9800\right)} = P\left(\frac{\sum X_i - n\mu}{\sqrt{n\sigma^2}} < \frac{9800 - 10045}{\sqrt{11025}}\right)$$

$$= P(Z < -2.33) = 0.5 - 0.4901 = 0.099$$

2. The amount of mineral water consumed by a person per day on the job is normally distributed with mean 19 ounces and standard deviation 5 ounces. A company supplies its employees with 2000 ounces of mineral water daily. The company has 100 employees.

a) Find the probability that the mineral water supplied by the company will not satisfy the water demanded by its employees.

b) Find the probability that in the next 4 days the company will not satisfy the water demanded by its employees on at least 1 of these 4 days. Assume that the amount of mineral water consumed by the employees of the company is independent from day to day.

c) Find the probability that during the next year (365 days) the company will not satisfy the water demanded by its employees on more than 15 days.

Solution:

$$\underset{\substack{\downarrow\\ \text{amount of water}\\ \text{for person per day}}}{X \sim N(19, 25)} \quad \underset{(n=100)}{\Rightarrow n \to \infty} \quad \sum X_i \sim N(1900, 2500)$$

a)

$$P\left(\sum X_i < 2000\right) = P\left(Z < \frac{1900 - 2000}{\sqrt{2500}}\right) = P(Z < -2) = 0.5 - 0.4772 = \underset{\substack{\downarrow\\ \text{corresponds to the lack of water sup lement for all employees.}}}{0.0228}$$

b) X: The number of day of 4 days in which the company would face with the problem lack of water for their employees.

$$P(X \geq 1) = \sum_{x=1}^{4} \binom{4}{x} 0.0228^{x}\, 0.9772^{4-x}$$

c) $X \sim \text{Binomial}(365,\ p = 0.0228)$

$P(X > 15) = ?$

Exact solution: $P(X > 15) = \sum_{x=16}^{365} \binom{365}{x} 0.0228^{x}\, 0.9772^{365-x}$

Approximate solution from Central Limit Theorem

$$P\left(\frac{X - np}{\sqrt{npq}} > \frac{15 - 365 \times 0.0228}{\sqrt{365 \times 0.0228 \times 0.9772}} \right) = P\left(Z > \frac{6.678}{\sqrt{8.1322}} \right)$$

$$P(Z > 2.34) = 0.5 - 0.4904 = 0.0096$$

İST292 STATISTICS LESSON 4
SAMPLING DISTRIBUTIONS

4. SAMPLING DISTRIBUTIONS

Since sample statistics are random variables, they therefore have (possess) probability distributions that are either discrete or continuous. These probability distributions, called ***sampling distributions*** because they characterize the distribution of values of the various statistics over a very large number of samples, are the topic of this lesson.

A ***parameter*** is a numerical descriptive measure of a population. It is calculated from the observations in the population.

Since it is almost impossible to get all observations of a population because it is costly and time consuming, a sample which would be desribed well to the population is taken from a population and then *sample statistics* are used to make inference about the parameters of a population.

A ***sample statistic*** is a numerical descriptive measure of a sample. It is calculated from the observations in the sample.

Not that the term *statistic* refers to a sample quantity and the term *parameter* refers to a population quantity.

If we want to estimate a parameter of a population- say, the population mean μ- there are a number of sample statistics that could be used for the estimation. Two possibilities are the sample mean $\bar{x}$ and the sample median $\bar{x}'$. Which of these do you think will provide a better estimate of μ?

Neither the sample mean nor the sample median will always fall closer to the population mean. Consequently, we can not compare these two sample statistics, or, in general, any two sample statistics, on the basis of their performance for a single sample. Instead, since the sample statistics are themselves random variables, they must be judged and compared on the basis of their *probability distributions* i.e (id est: yani) the collection of values and associated probabilities of each statistic that would be obtained if the sampling experiment were repeated a very large number of times.

The ***sampling distribution*** of a sample statistic calculated from a sample of n measurments is the probability disribution of the statistic.

A ***point estimator*** of a population parameter is rule of formula that tells us how to use the sample data to calculate a single number that can be used as an estimate of the population parameter.

population parameter	***point estimate of the parameter***
μ	$\overline{x}$ (sample mean)
σ^2	S^2 (sample variance)

By examining the sampling distribution, we can determine how large the difference between an estimate and the true value of the parameter (called the error of estimation) is likely to be.

If the sampling distribution of a sample statistic has a mean equal to the population parameter which the statistic is intended to estimate, the statistic is said to be an **unbiased estimate** of the parameter.

If the mean of the sampling distribution is not equal to the parameter, the statistic is said to be a **biased estimate** of the parameter.

The *standard deviation of a sampling distribution* measures another important property of statistics- the spread of these estimates generated by repeated sampling. Suppose two statistics, A and B, are both unbiased estimators of the population parameter. Since the means of the two sampling distributions are the same, we turn to their standard deviations (standard error of the estimate) in order to decide which will provide estimates that fall closer to the unknown population parameter we are estimating. Naturally, we will choose the sample statistic that has the smaller standart deviation.

The Distribution of the Mean

Since statistics (sample quantities: sample mean, variance, etc.) are random variables, their values will vary from a sample to another sample. Sample mean has a probability distribution and also its expected value and variance can be found.

Let $X_1, X_2, ..., X_n$ be a random sample from a an infinite population with the mean μ and the variance σ^2, then

$$E(\bar{X}) = \mu, \quad V(\bar{X}) = \frac{\sigma^2}{n} = \sigma_{\bar{X}}^2$$

$\bar{X}$ is an ***unbised estimator of the population mean*** μ, and $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$ is called ***standard error the mean***. It refers to a measure how close to the mean of a sample is the population mean μ. When the sample size n gets close to ∞, $\frac{\sigma^2}{n} \to 0$, and so we say that $\bar{X} \to \mu$.

The Distribution of the Mean: Finite Populations

If an experiment consists of selecting one or more values from a finite set of numbers $\{c_1, c_2, ...c_N\}$, this set is referred to as ***a finite population size N***. Assume that we are *sampling without replacement* from a finite population size *N*.

If $\bar{X}$ is the mean of a random sample of size n from a finite population size N with the mean μ and the variance σ^2, then

$$E(\bar{X}) = \mu, \quad V(\bar{X}) = \frac{\sigma^2}{n} . \frac{N-n}{N-1}$$

where $\frac{N-n}{N-1}$ is called the ***finite population correction factor***. This term is usually negligible (ihmal edilebilir).

The Distribution of the Mean: A Sample from Normal Distribution (Population)

Let $X_1, X_2, ..., X_n$ be a random sample from normal distribution with the mean μ and the variance σ^2. It is shown as $X_1, X_2, ..., X_n \sim N(\mu, \sigma^2)$, then

$$E(\bar{X}) = \mu, \quad V(\bar{X}) = \frac{\sigma^2}{n}$$

also the distribution of a sample mean $\bar{X}$ is a normal distribution with the mean μ and the variance σ^2 / n and shown $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$. In addition, standardized random variable has the standard normal distribution:

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1) .$$

The Distribution of the Ratio: A Sample from Bernoulli Distribution (Population)

Suppose that we are interested in whether each value of a population has a particular status or not- ***0*** or ***1*** value is assigned, ***0*** refers to ***"failure"*** and ***1*** refers to ***"success"***- say that the population has Bernoulli distribution with probability *p* (each success occurrs with the probability *p*). It is shown as $X \sim Bernoulli(p)$.

Suppose a random sample of n measurements is drawn from this Bernoulli distribution, the total number of successes and the ratio of success are calculated:

If $X_1, X_2, ..., X_n \sim Bernoulli(p)$, $\sum_{i=1}^{n} X_i$ is the ***total number of successes*** and $\frac{\sum_{i=1}^{n} X_i}{n}$ is the ***ratio of success*** in the sample. These are statistics calucated from the sample, then

$$E\left(\sum_{i=1}^{n} X_i\right) = np, \quad V\left(\sum_{i=1}^{n} X_i\right) = npq \text{ and}$$

$$E\left(\sum_{i=1}^{n} X_i \Big/ n\right) = p, \quad V\left(\sum_{i=1}^{n} X_i \Big/ n\right) = \frac{pq}{n} .$$

For example, the Bernoulli random variable might be the status of single computer microchip (good or defective), $\sum_{i=1}^{n} X_i$ is the number of n such chips that are good, and $\frac{\sum_{i=1}^{n} X_i}{n}$ the fraction of good chips in a set of n.

Note that $X_1, X_2, ..., X_n \sim Bernoulli(p)$, $\sum_{i=1}^{n} X_i$ has a ***Binomial distribution*** with parameters (*n,p*).

The Student's t Distribution

As it is known, if a random sample is drawn from normal distribution with the mean μ and the variance σ^2, the random variable –sample mean- has a normal distribution with the mean μ and the variance σ^2 / n; in other words,

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

has the ***standard normal distribution***. S^2 is the variance of random sample of size n from normal distribution with the mean μ and the variance σ^2, let our interest be find the exact disribution of the random variable $\frac{\bar{X} - \mu}{S / \sqrt{n}}$, its distribution is known as ***t distribution with n-1 degrees of freedom***. The *t* distribution was introduced by W. S. Gosset, who published his scientific writings under pen name (takma ad) "Student" since the company which he worked, a brewery (bira fabrikası), did not permit publication by employees. Thus the t distribution is also known as the ***Student t distribution*** or ***Student's t distribution***. (John E. Freund's Math. Stat. with Applications, 2004).

If *T* is a Student *t* distribution with ν degrees of freedom, *T* has zero mean and $\frac{\nu}{\nu - 2}$ (ν>*2*) variance. The density is symmetrical about *t=0* and hence $t_{1-\alpha,\nu} = -t_{\alpha,\nu}$ where $P(T \geq t_{\alpha,\nu}) = \alpha$.

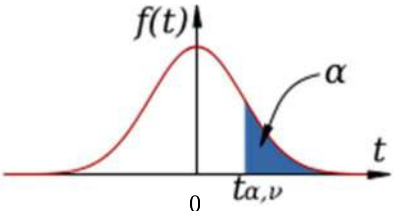


Figure 1. *t* distribution

(https://en.wikibooks.org/wiki/Engineering_Tables/Student%27s_T-Distribution#/media/File:Student_T-Distribution_Table_Diagram.svg)

The Chi-square Distribution

Let $X_1, X_2, ..., X_n$ be a random sample from normal distribution with the mean μ and the variance σ^2. For the sample mean and variance, $\overline{X}$ and S^2, then,

- $\overline{X}$ and S^2 are independent random variables (statistics)
- The random variable $\overline{X}$ has a normal distribution with the mean μ and the variance σ^2 / n, and shown $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$.
- The random variable $\frac{(n-1)S^2}{\sigma^2}$ has a chi-square distribution with *n-1* degrees of freedom, and shown $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$.

If *X* is a chi-square distribution with ν degrees of freedom, *X* has ν mean and 2ν (*ν>0*) variance. The density is positively (right) skewed. The figure shows its density as follow:

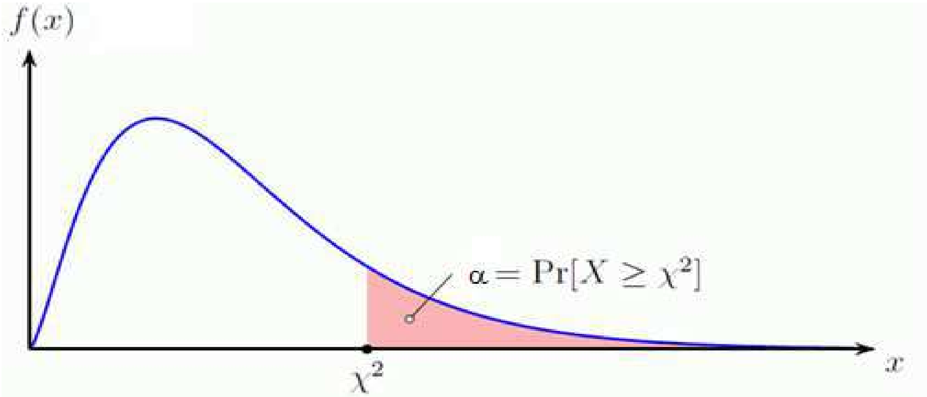


Figure 2. *Chi-square* distribution. (http://www.di-mgt.com.au/chisquare-calculator.html)

The probabilities in *chi-square* table show areas to its right under the chi-square curve with degrees of freedom ν: $P(X \geq \chi^2_{\alpha,\nu}) = \alpha$

When ν is greater than 30, chi-square distributions are usually appoximated with normal distributions.

The F Distribution

Another sampling distribution with related to normal populations is the F distribution, named after Sir Ronald A. Fisher, one of the most prominent statisticians of the last century. Originally, it was studied as the sampling distribution of the ratio of two independent variables with chi-square distributions, each divided by its respective degrees of freedom.

Asume that $X_{11}, X_{12}, ..., X_{1n_1}$ is a random sample from normal distribution with the mean μ_1 and the variance σ_1^2 and independently $X_{21}, X_{22}, ..., X_{2n_2}$ another random sample from normal distribution with the mean μ_2 and the variance σ_2^2.

For a sample, the random variable $\frac{(n_1 - 1)\,\mathrm{S}_1^2}{\sigma_1^2}$ has a chi-square distribution with *n*n_1*-1* degrees of freedom, and shown $\frac{(n_1 - 1)\,\mathrm{S}_1^2}{\sigma_1^2} \sim \chi^2_{(n_1 - 1)}$. For another sample, the random variable $\frac{(n_2 - 1)\,\mathrm{S}_2^2}{\sigma_2^2}$ has a chi-square distribution with n_2*-1* degrees of freedom, and shown $\frac{(n_2 - 1)\,\mathrm{S}_2^2}{\sigma_2^2} \sim \chi^2_{(n_2 - 1)}$.

The ratio of two independent variables with chi-square distributions, each divided by its respective degrees of freedom is defined as:

$$\frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} = \frac{\sigma_2^2}{\sigma_1^2}\frac{S_1^2}{S_2^2} \sim f_{\alpha,(n_1 - 1)(\mathrm{n}_2 - 1)}\,.$$

Like chi-square distribution, *F* distribution is positively (right) skewed, but it has two degrees of freedoms (ν_1, ν_2). $P(F \geq f_{\alpha,\nu_1,\nu_2}) = \alpha$

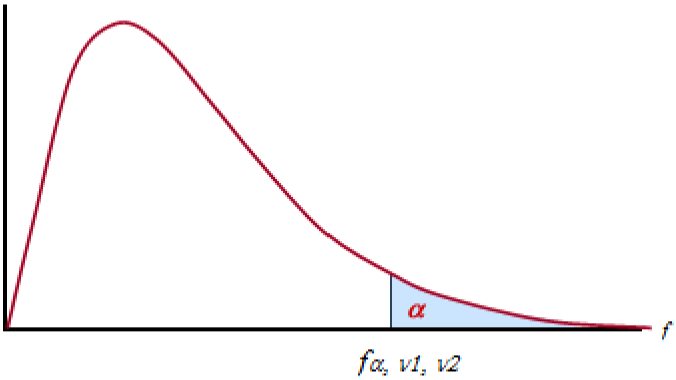


Figure 3. *F* distribution. (http://gmein.uib.es/bioinformatica/statistics/)

Examples 1. If $T \sim t_{(9)}$, find a, b and c values given in $P(T < a) = 0.10$, $P(T > b) = 0.01$, $P(-c < T < c) = 0.96$.

Solution: a=-1.383, b=2.821, c=2.398

Examples 2. If $X \sim \chi^2_{(6)}$, find the probabilities $P(X < 10.64)$, $P(10.65 < X < 14.45)$, $P(X \geq 2.2)$

Solution:

$P(X < 10.64) = 0.90$, $P(10.65 < X < 14.45) = 0.10 - 0.025 = 0.075$ $P(X \geq 2.2) = 0.90$

Examples 3. If $X \sim \mathrm{f}_{(2,9)}$, find a and b values satisfing to $P(X > a) = 0.05$, $P(X > b) = 0.95$.

Solution: if $P(X > a) = 0.05$, then a=4.26.

Using properties of f distribution $f_{(1-\alpha),(\nu_1,\nu_2)} = \dfrac{1}{f_{\alpha,(\nu_2,\nu_1)}}$

$$f_{0.95,(2,9)} = \frac{1}{f_{(0.05),(9,2)}} = \frac{1}{19.38} \cong 0.05$$

Examples 4. If $X \sim \chi^2_{(18)}$, find a and b values given in $P(2X < a) = 0.10$, $P(X - 1 < b) = 0.25$

Solution: $\dfrac{a}{2} = 10.86494 \Rightarrow a = 21.72988$ and $b + 1 = 13.67529 \Rightarrow b = 12.67529$

Examples 5. If $T \sim t_{(30)}$, find a and b values satisfing to $P(-\mathrm{a} < T < b) = 0.88$, $P(T > -a) = 0.98$,

Solution: If $P(T > -a) = 0.98 \Rightarrow a = -2.147$ and so $P(T > b) = 0.10 \Rightarrow b = 1.310$

Examples 6. A random sample of size n=21 was drawn a normal population with variance 10. Find probability of the sample variance being less than 17.085 and greater than 6.22.

Solution: $\frac{(n-1)\,\mathrm{S}^2}{\sigma^2} \sim \chi^2_{(n-1)}$ and the asked probability is

$$P(6.22 < S^2 < 17.085) = P(\frac{20\times 6.22}{10} < \frac{(n-1)S^2}{\sigma^2} < \frac{20\times 17.085}{10})$$
$$= P(12.44 < \chi^2_{(n-1)} < 34.17) = 0.90 - 0.025 = 0.875$$

SAÜ

7. BÖLÜM

ASİMETRİ (ÇARPIKLIK) VE BASIKLIK ÖLÇÜLERİ

PROF. DR. MUSTAFA AKAL

İÇİNDEKİLER

HEDEFLER

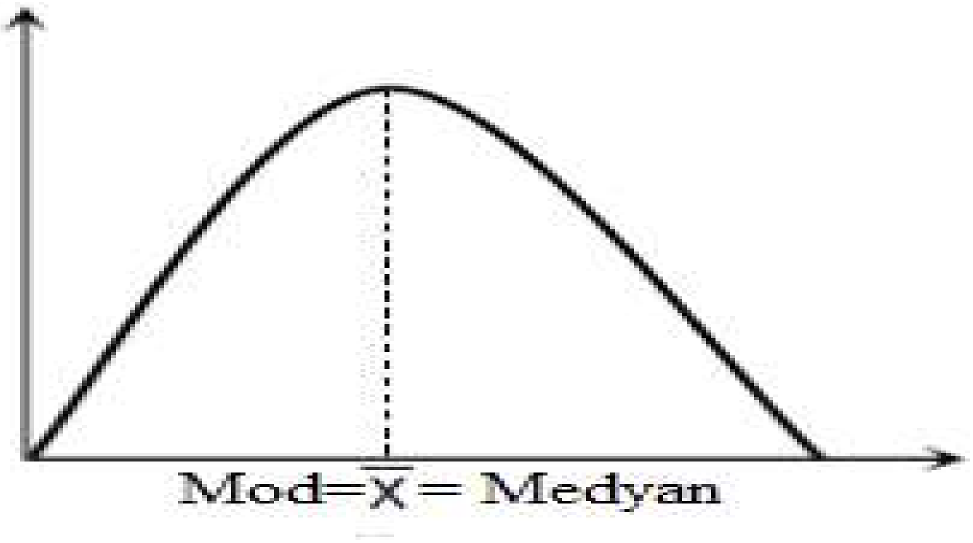
Çarpıklık ve basıklık tanımlarının tanıtılması ve çarpıklık hesaplamalarının gösterilmesi.

1. ÇARPIKLIK VE BASIKLIK

Serilerin dağılımı hakkında ortalamalar ve değişim ölçüleri yardımıyla belli ölçüde bilgi edinebiliriz. Bu iki ölçünün yanında, serilerin simetriden ne kadar uzaklaştığını gösteren **“Çarpıklık Katsayısı”** ve serinin yüksekliğinin normal serinin yüksekliğinden ne kadar uzaklaştığını gösteren **“Basıklık Katsayısı”** hesaplanabilir.

1.1. Ortalamalar Yardımıyla Çarpıklığın (asimetri, skewness) Hesaplanması

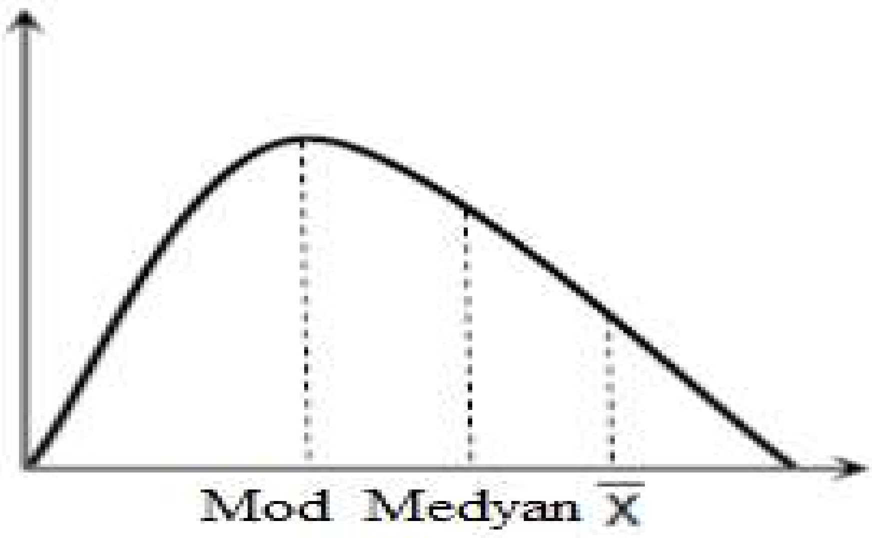
Araştırmacılar, çalışmalarında genellikle ortalamalar ve değişim ölçülerini hesaplayarak seri hakkında ulaşabilecekleri maksimum bilgiye ulaştıklarını ve diğer ölçülerin hesaplanmasının fazla bir bilgi sağlamayacağını savunmaktadırlar. Çoğu zaman bunda haklılık payı olsa da, serinin dağılımının şekli hakkında bilgiler edinmenin araştırmacıya ilave bilgiler sağlayacağında göz ardı edilmemesi gerekir. Serilerin frekans dağılımlarını gösteren aşağıdaki üç şekil incelendiğinde bu daha iyi anlaşılacaktır.



Serinin frekans dağılımını gösteren yukarıdaki şekilden verilerin merkezi eğilim ölçüleri etrafında **simetrik** dağıldığını söyleyebiliriz. Bu serinin mod, medyan ve aritmetik ortalaması birbirine eşittir. Aşırı büyük ve küçük değerlerin frekansları eşit ya da birbirine çok yakındır.

Simetrik serilerde → Mod = Medyan = $\bar{X}$

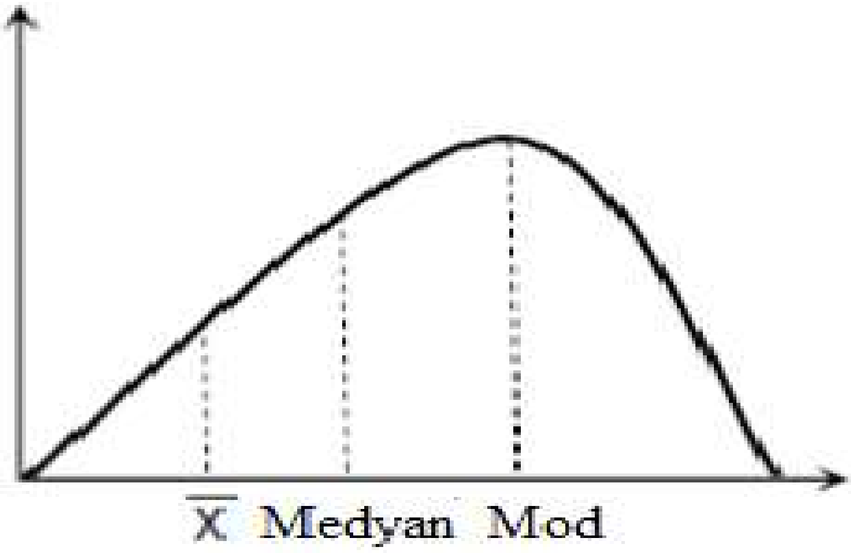
Öğrencilerin istatistik dersinden aldıkları notlar ortalama etrafında simetrik olarak dağılıyorsa yukardaki durum söz konusu olacaktır.



Serinin frekans dağılımını gösteren yukarıdaki şekil sağa doğru uzun kuyrukludur. **Sağa çarpık** frekans dağılımına sahip olan bu seride merkezi eğilim ölçüleri arasındaki ilişki aşağıdaki gibidir.

Sağa çarpık serilerde ⟶ **Mod < Medyan <** $\overline{\mathbf{X}}$

Aşırı küçük değerlerin frekansı büyük değerlerin frekansından daha fazladır. Bundan dolayı aritmetik ortalama medyandan ve medyanda moddan **daha büyüktür.** ***Öğrencilerin istatistik dersinden aldıkları notların çoğunluğu aritmetik ortalamadan küçükse yukarıdaki şekilde olduğu gibi sağa çarpık frekans dağılımı söz konusu olacaktır.*** Yüksek değerli gözlemler geniş bir aralıkta yer alırken düşük değerli gözlemler nispeten bir arada toplanmıştır.



Serinin frekans dağılımını gösteren yukarıdaki şekil sola doğru uzun kuyrukludur. **Sola çarpık** frekans dağılımına sahip olan bu seride merkezi eğilim ölçüleri arasındaki ilişki aşağıdaki gibidir.

Sola çarpık serilerde ⟶ Mod > Medyan > $\bar{X}$

Aşırı büyük değerlerin frekansı küçük değerlerin frekansından daha fazladır. Bundan dolayı aritmetik ortalama medyandan ve medyan da moddan **daha küçüktür.** ***Öğrencilerin istatistik dersinden aldıkları notların çoğunluğu aritmetik ortalamadan büyükse yukarıdaki şekilde olduğu gibi sola çarpık frekans dağılımı söz konusu olacaktır.*** Düşük değerli gözlemler geniş bir aralıkta yer alırken yüksek değerli gözlemler nispeten bir arada toplanmıştır.

Görüldüğü gibi gözlemlerin frekans dağılımları farklılık gösterebilir ve serilerin çarpıklığının ölçülmesi önemli bilgiler içermektedir. Merkezi eğilim ve değişim ölçüleri serilerin çarpıklığı hakkında bilgi içermezler ve bunların değişik yöntemlerle hesaplanması gerekir. *Serilerin asimetrisi (çarpıklığı) merkezi eğilim ölçüleri, kartiller ya da momentler yardımıyla hesaplanırken serilerin basıklığı momentler yardımıyla hesaplanabilir.*

Herhangi bir seride bu 3 ilişkiden 1 tanesi vardır. Serinin sağa veya sola yakınlığı (artıkça) asimetrik ortalama ile mod arasındaki fark belirli şekilde büyür.

1.2. Merkezi Eğilim Ölçüleri Yardımıyla Serilerin Çarpıklığının Hesaplanması

Serilerin mod ya da medyanının aritmetik ortalamadan farkının o serinin standart sapmasına bölünmesi ile serinin asimetrisi yani çarpıklığı hesaplanabilir. Bulucusunun adından dolayı **Pearson Asimetri Ölçüsü** denilen bu çarpıklık katsayıları (ÇK) aşağıdaki formüller yardımıyla hesaplanır.

$$ÇK = \frac{\bar{X}-Mod}{\sigma} \quad \text{veya} \quad ÇK = \frac{3(\bar{X}-Medyan)}{\sigma}$$

Çarpıklık katsayısı -1 ve +1 sınırları arasında olacaktır. Çarpıklık katsayısı +1'e yaklaştıkça serinin sağa çarpıklığı ve -1'e yaklaştıkça serinin sola çarpıklığı artacaktır. Çarpıklık katsayısı sıfıra yaklaştıkça serinin simetrisi (çarpıklığı) artacaktır (azalacaktır).

ÇK = 0 ⟶ Seri simetriktir

ÇK > 0 ⟶ Seri sağa çarpıktır

ÇK < 0 ⟶ Seri sola çarpıktır

ÖRNEK: Bir mahallede yaşayan aileler çocuk sayılarına göre tasnif edilmiş seri olarak aşağıda verilmiştir. Bu dağılımın çarpıklık katsayısını hesaplayınız ve yorumlayınız?

Çocuk Sayısı (X_i)	Aile Sayısı f_i	f_iX_i	$(X_i - \mu)$	$(X_i - \mu)^2$	$f_i(X_i - \mu)^2$
0	1	0	-2.3	5.3	**5.3**
1	3	3	-1.3	1.7	**5.1**
2	8	16	-0.3	0.1	**0.8**
3	5	15	0.7	0.5	**2.5**

4	3	12	1.7	2.9	**8.7**
Toplam	$\sum_{i=1}^{n} f_i = 20$	$\sum_{i=1}^{n} f_i X_i = 46$			**22.4**

İlk aşamada serinin aritmetik ortalama, mod, medyan ve standart sapmasını hesaplamalıyız. Serinin aritmetik ortalama, mod, medyan ve standart sapması aşağıdaki gibi hesaplanmıştır.

$$\mu = \frac{\sum_{i=1}^{k} N_i X_i}{\sum_{i=1}^{k} N_i} = \frac{46}{20} = 2.3, \text{ Mod} = 2, \text{ Medyan} = 2,$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k} N_i \left(X_i - \mu\right)^2}{\sum_{i=1}^{k} N_i}} = \sqrt{\frac{22.4}{20}} = 1.06.$$

Bu seride aritmetik ortalama medyan ve moddan büyük olduğu için sağa çarpıktır. Çarpıklık derecesi ya da katsayısı bulunan değerlerin formülde yerine konulmasıyla aşağıdaki gibi bulunur.

$$ÇK = \frac{\mu - Mod}{\sigma} = \frac{2.3 - 2}{1.06} = 0.28$$

Serinin asimetrisi yani çarpıklığı pozitiftir (sağa çarpık) ve katsayısı 0.28'dir. 0< ÇK < 0.5 olduğundan asimetrisi hafif sağa çarpık seri olarak yorumlanır.

Asimetrisi hafif ya da simetriğe yakın dağılıma sahip serilerde yaklaşık olarak

$Mo \cong 3Me - 2\bar{X}$ **ilişkisi görülür.**

Her iki taraftan $-\bar{X}$ çıkartılıp, Her iki taraf (-1) ile çarpılırsa, serinin asimetrisi hafif ise 3 ortalama arasında şu ilişki vardır:

$$\boxed{\bar{X} - Mo \cong 3\left(\bar{X} - Me\right)}$$

Her iki taraf serinin standart sapması ile bölünürse Perason asimetri ölçüsüne ulaşılır.

$$ASp_I = \frac{\bar{X} - Mo}{\sigma} \quad \text{veya} \quad ASp_{II} = 3\frac{\left(\bar{X} - Me\right)}{\sigma}$$

Pearson asimetri ölçüsü teorik olarak ± 3 sınırları arasında bulunması gereken bu ölçü çoğu zaman ± 1 sınırları arasında gerçekleşir **($-1 \leq AS_p \leq 1$).**

Hesaplanan asimetri ölçüsü ± 1'e yaklaştıkça asimetri derecesi yükselir.

a) simetrik serilerde As_p=0

b) sağa eğik serilerde $As_p>0$

c) sola eğik serilerde $As_p<0$

$As_p=0$ sonucu rastlantısal ise serinin simetrik olduğu kesin değildir. Kesinlik kazandırmak için alternatif asimetri ölçülerine bakılır.

ÖRNEK: Bir serinin $\bar{X}=19$, $M_e=19.5$, $M_o=20$, $\sigma=2$ olsun.

$AS_{p1}=\dfrac{19-20}{2}=0.50; AS_{p2}=\dfrac{3.(19-19.5)}{2}=-0.75$ <0 olduğundan kuvvetli sola eğik seridir.

ÖRNEK: A ve B sınıflanmış serileri için aşağıdaki değerler verilsin.

$\bar{X}_A=3.625$	$\sigma_A=1.74$	$M_0^A=3.14$	$M_e^A=3.43$
$\bar{X}_B=5$	$\sigma_B=1.78$	$M_0^B=5.4$	$M_e^B=5.25$

Buna göre A ve B serilerinin asimetri durumlarını kıyaslayınız.

$$AS_p^{AI}=\frac{\bar{X}-M_0}{\sigma_A}=\frac{3.625-3.14}{1.74}=0.2787>0 \text{ sağa eğik seri}$$

$$AS_p^{AII}=\frac{3(\bar{X}-M_e)}{\sigma_A}=\frac{3(3.625-3.14)}{1.74}=0.336>0$$

$$AS_p^{BI}=\frac{\bar{X}-M_0}{\sigma_A}=\frac{5-5.4}{1.78}=-0.2247<0 \text{ Sola eğik}$$

$$AS_p^{BII}=\frac{3(\bar{X}-M_e)}{\sigma_A}=\frac{3(5-5.4)}{1.78}=-0.421<0 \text{ sola eğik}$$

Her iki serinin asimetrisi hafif çünkü $\boxed{\begin{array}{l}0\le AS_p^A\le 0.5\\ -0.5\le AS_p^B\le 0\end{array}}$ dir.

AS_{PI} ve AS_{PII} çelişkili ise asimetrinin yönünü belirlemede AS_{PI} ölçümü dikkate alınır. Buna göre A serisinin asimetrisi B serisinden fazladır. Çünkü mutlak değerce AS_p^{AI} =0.2789 > AS_p^{BI} =0.2247 dir.

1.3. Kartillere Dayanan Asimetri Ölçüsü

Kartiller arasında 3 türlü ilişki mevcuttur.

a) $Q_3-Q_2\rangle Q_2-Q_1$ **b)** $Q_3-Q_2=Q_2-Q_1$ **c)** $Q_3-Q_2\langle Q_2-Q_1$

sağa eğik seri simetrik seri sola eğik seri

Serinin simetrisi bozuldukça, yani sağa veya sola eğiklik arttıkça söz konusu 2 fark; $(Q_3 - Q_2) - (Q_2 - Q_1)$ farkı sıfırdan uzaklaştıkça sola veya sağa eğiklik artar, değişkenlik artar.

Farkının pozitif olması serinin sağa eğik, negatif olması sola eğik olduğunu gösterir.

ÖRNEK: Q_1 =2.29, Q_2 =3.43, Q_3 =5 olarak verilsin.

$Q_3 - Q_2$ = 5-3.43=1.57 > $Q_2 - Q_1$ = 3.43-2.29=1.14

1.57 >1.14 asimetrisi sağa eğik seridir.

Bowley Asimetri Ölçüsü: kartiller arası farkların, kartiller arası fark toplamlarına oranına dayanır.

$$\boxed{AS_B = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)}}$$

Pearson asimetri ölçüsünde olduğu gibi simetrik serilerde $AS_B = 0$, sağa eğik serilerde $AS_B > 0$ ve sola eğik serilerde $AS_B < 0$’dır. Bowley asimetri ölçüsü $\boxed{-1 \le AS_B \le +1}$ sınırlıdır.

ÖRNEK: Q_1 =2.29, Q_2 =3.43, Q_3 =5 olarak verilsin. Bowley Asimetri ölçüsünü bulunuz?

$$\boxed{AS_B = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)}} = \frac{(5\text{-}3.43) - (3.43\text{-}2.29)}{1.57 \;\; +1.14} = \frac{0.43}{2.71} = 0.16$$

0≤0.16≤0.5 aralığında olduğundan asimetrisi hafif sağa eğik seridir.

ÖRNEK: Bir serinin Q_1 =1.5, Q_2 =3.5, Q_3 = 6.5 olsun.

$$\boxed{AS_B = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)}} \quad \boxed{AS_B = \frac{(6.5 - 3.5) - (3.5 - 1.5)}{(6.5 - 3.5) + (3.5 - 1.5)}} = \frac{3-2}{3+2} = 0.2$$

0< AS_p=0.2 <0.5 asimetrisi hafif sağa eğik bir seridir.

ÖRNEK: P&R şirketlerinde çalışan işçilerin haftalık ücret dağılımlarının a) Asp' ve Asp'' ölçülerini bulunuz? b**)** Bowley asimetri ölçüsünü bulunuz? c) Sadece aritmetik ortalama ve medyan bilindiğinde 65 işçinin model ücretini bulunuz**?**

Ücretler	f_i	m_i	$f_i m_i$
250-259.99	8	255	2040
260-269.99	10	265	2650

270-279.99	16	275	4400
280-289.99	14	285	3990
290-299.99	10	295	2950
300-309.99	5	305	1525
310-319.99	2	315	630
	$\sum f_i$=65		$\sum f_i m_i$=18185

a. $\bar{X} = \frac{18185}{65} = 279.76$

$$M_o = \ell + s.\frac{\Delta_1}{\Delta_1 + \Delta_2} = \ell + s.\frac{\left(N_m - N_{m-1}\right)}{\left(N_m - N_{m-1}\right) + \left(N_m - N_{m-1}\right)}$$

$$= 270 + 10\frac{(16-10)}{6+(16-14)} = 277.58$$

$$M_e = \ell + s.\frac{\left(\frac{N}{2} - N_a\right)}{N_m} = 270 + (108)\frac{(32.5-18)}{16} = 279.06$$

$\sigma = \sqrt{280.2^2 - 279.76^2} = 15.76$, gruplanmış seri olduğundan düzeltmiş standart sapma uygulanır;

$$\sigma' = \sqrt{\sigma^2 - \frac{s^2}{12}} = \sqrt{15.76^2 - \frac{10^2}{12}} = 15.5$$

$\bar{X}$ =279.76, Me=279.06, Mo=277.58, σ' =15.5

Ancak burada gruplanmış seri olduğundan doğru asimetri ölçüsü düzeltilmiş (σ') kullanılmasıyla; σ' =15.5 elde edilir.

$$\boxed{ASp' = \frac{279.76-277.5}{15.5} = 0.1448} \qquad \boxed{ASp'' = \frac{3(279.76-279.06)}{15.5} = 0.1346}$$

Ve $\boxed{0 \leq ASp \leq 0.5}$ asimetrisi hafif sağa eğik seridir. İşçilerin çoğu ortalama ücretin altında ücret almaktadır.

b) Q_1 =259.995+ $\frac{16.25-8}{10}x10 = 268.25$ $\qquad \frac{1N}{4} = \frac{65}{4} = 16.25$. terim

$Q_2 = 269.995 + \frac{32.5-18}{10}x10 = 279.06$ $\qquad \frac{2N}{4} = 32.5$. terim

Q_3 =289.995 + $\frac{48.75-48}{10}x10 = 290.758$ $\qquad \frac{3N}{4} = 48.75$. terim

Buna göre işçilerin %25'i 268.258 veya daha az kazanır.

%50'si 279.06 veya daha az kazanır.

% 75'i 290.75 veya daha az kazanır.

$Q_3 = 290.75$, $Q_2 = 279.06$, $Q_1 = 268.25$ ise R&P serisinde

$\boxed{AS_B = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)}} = \boxed{\frac{(290.75 - 279.06) - (279.06 - 268.25)}{(290.75 - 279.06) + (279.06 - 268.25)} = 0.039}$ olup asimetrisi hafif sağa eğik seridir.

. $V_Q{}^A = \frac{290.75 - 268.25}{290.75 + 268.25} = \frac{22.5}{559} = 0.0402$ =%4

c) $M_0 \cong \bar{X} - 3(\bar{X} - M_e)$ Asimetrisi hafif seri özelliğini kullanırız.

$\cong$ 279.76 – 3.(279.76–279.06) $\cong$ 277.66

$Mo \cong 3Me - 2\bar{X}$ =3(279.06)–2(279.76) $\cong$ 277.66

Ve bizim hesaplanan M_o ücretimiz=277.58 olduğu için yakın bir ilişki vardır.

2. MOMENTLER YARDIMIYLA ÇARPIKLIĞIN (ASİMETRİ, SKEWNESS) VE BASIKLIĞININ (KURTOSİS) HESAPLANMASI

Serilerin frekans dağılımları hakkında momentler yardımıyla da bilgi edinilebilir. Momentler, gözlem değerlerinin aritmetik ortalamadan farklarının kuvvetini alarak gözlem sayısına bölünmesi ile elde edilir. Bu şekilde hesaplanan momentlere aritmetik ortalama etrafındaki momentler denir ve en yaygın kullanılanıdırlar. Momentleri hesaplama formülleri serilerin türüne göre değişecektir. Formüllerdeki **r** sembolü momentin derecesini gösterir.

2.1. Basit Serilerde Ortalamadan Sapmaya Göre Momentler

Birinci moment: $\mu_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})}{N} = 0$

İkinci moment: $\mu_2 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N} = \sigma^2 = varyans$

Üçüncü moment: $\mu_3 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^3}{N}$

Dördüncü moment: $\mu_4 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^4}{N}$

r. moment: $\mu_r = \frac{\sum_{i=1}^{N}(X_i \text{-} \mu)^r}{N}$

2.2. Tasnif Edilmiş Serilerde Ortalamadan Sapmaya Göre Momentler

Birinci moment: $\mu_1 = \frac{\sum_{i=1}^{k} N_i \left(X_i - \mu \right)}{\sum_{i=1}^{k} N_i} = 0.$

İkinci moment: $\mu_2 = \frac{\sum_{i=1}^{k} N_i \left(X_i - \mu \right)^2}{\sum_{i=1}^{k} N_i} = \sigma^2$

Üçüncü moment: $\mu_3 = \frac{\sum_{i=1}^{k} N_i \left(X_i - \mu \right)^3}{\sum_{i=1}^{k} N_i}$

Dördüncü moment: $\mu_4 = \frac{\sum_{i=1}^{k} N_i \left(X_i - \mu \right)^4}{\sum_{i=1}^{k} N_i}$

r. moment: $\mu_r = \frac{\sum_{i=1}^{k} N_i \left(X_i - \mu \right)^r}{\sum_{i=1}^{k} N_i}$

2.3. Gruplanmış Serilerde Ortalamadan Sapmaya Göre Momentler

Birinci moment: $\mu_1 = \frac{\sum_{i=1}^{k} N_i \left(m_i - \mu \right)}{\sum_{i=1}^{k} N_i} = 0.$

İkinci moment: $\mu_2 = \frac{\sum_{i=1}^{k} N_i \left(m_i - \mu \right)^2}{\sum_{i=1}^{k} N_i} = \sigma^2,$ $\boxed{\mu_2^1 = \mu_2 - \frac{S^2}{12}}.$

Üçüncü moment: $\mu_3 = \frac{\sum_{i=1}^{k} N_i \left(m_i - \mu \right)^3}{\sum_{i=1}^{k} N_i}$

Dördüncü moment: $\mu_4 = \frac{\sum_{i=1}^{k} N_i \left(m_i - \mu \right)^4}{\sum_{i=1}^{k} N_i},$ $\boxed{\mu_4^1 = \mu_4 - \frac{S^2}{2}\mu_2 + \frac{7S^4}{240}}$

r. moment: $$\mu_r = \frac{\sum\limits_{i=1}^{k} N_i \left(m_i - \mu\right)^r}{\sum\limits_{i=1}^{k} N_i}$$

2.4. Momentler Yardımıyla Çarpıklık (Asimetri, Skewness) Katsayısının Hesaplanması

Simetrik serilerde aritmetik ortalamadan sapmaların tek dereceli kuvvetlerinin toplamı sıfır olacağından simetrik serilerde birinci ve üçüncü momentler sıfır olacaktır. Buradan **üçüncü momente** bakarak serinin asimetrisi hakkında aşağıdaki sonuçlara ulaşabiliriz.

$\mu_3 = 0$ ⟶ Simetrik seri

$\mu_3 > 0$ ⟶ Sağa çarpık (Asimetrisi pozitif) seri

$\mu_3 < 0$ ⟶ Sola çarpık (Asimetrisi negatif) seri

Bir serinin üçüncü momentine bakarak serinin çarpıklığı konusunda bir fikir sahibi olsak bile çarpıklığın derecesini ölçmek ve farklı birimlerle ölçülen serilerin asimetrisini karşılaştırmak için göreceli bir çarpıklık ölçüsüne ihtiyaç vardır. Serinin hesaplanan üçüncü dereceden momenti yine aynı serinin standart sapmasının üçüncü kuvvetine bölünürse standart bir çarpıklık ölçü birimi elde edilmiş olur. Momentler yardımıyla çarpıklık katsayısı aşağıdaki şekilde formüle edilebilir.

$$ÇK = \alpha_3 = \frac{\frac{\sum\limits_{i=1}^{k} N_i \left(X_i - \mu\right)^3}{\sum\limits_{i=1}^{k} N_i}}{\sigma^3} = \frac{\mu_3}{\sigma^3}$$

$\mu_3 = 0$ ⟶ ÇK = 0, Simetrik seri,

$\mu_3 > 0$ ⟶ ÇK > 0, Sağa çarpık (Asimetrisi pozitif) seri

$\mu_3 < 0$ ⟶ ÇK < 0, Sola çarpık (Asimetrisi negatif) seri

Momentlere Dayalı Asimetri Ölçüsü:-

$$\boxed{\alpha_3 = \frac{\mu_3}{\mu_2^{\frac{3}{2}}} = \frac{aritmetikortalamayagöre3.derecedenmoment}{aritmetikortalamayagöre2.momentin3/2.kuvveti}}$$

$$\alpha_3 = \frac{\mu_3}{\mu_2^{\frac{3}{2}}} = \frac{\mu_3}{\sqrt{\mu_2^3}} = \frac{\mu_3}{\sigma^3}$$

Gruplanmış serilerde yine μ_2^1 uygulanır. Yani $\left(\mu_2^1\right)^{\frac{3}{2}}$ konur.

Sağa eğik seri	α_3**>0**
Sola eğik seri	α_3**<0**
Simetrik seri	α_3**=0**
Asimetrisi kuvvetli	α_3**>0.5**
Asimetrisi zayıf	**-0.5<**α_3**<0**
Asimetrisi kuvvetli seri	$\lvert\alpha_3\rvert = \lvert -0.75 \rvert$ **>0.5**
$\mu_3 = 0$ **Simetrik seri**	α_3**=0**
$\mu_3 > 0$ **Sağa eğik seri**	α_3**>0**
$\mu_3 < 0$ **Sola eğik seri**	α_3**<0**

ÖRNEK: Bir mahallede yaşayan aileler çocuk sayılarına göre tasnif edilmiş seri olarak aşağıda verilmiştir. Bu dağılımın çarpıklık katsayısını momentler yardımıyla hesaplayınız ve yorumlayınız?

Çocuk Sayısı (X_i)	**Aile Sayısı f_i**	**f_iX_i**	$(X_i-\mu)$	$(X_i-\mu)^2$	$f_i(X_i-\mu)^2$	$(X_i-\mu)^3$	$f_i(X_i-\mu)^3$
0	1	0	-2.3	5.29	5.29	-12.167	**-12.167**
1	3	3	-1.3	1.69	5.07	-2.197	**-6.591**
2	8	16	-0.3	0.09	0.72	-0.027	**-0.216**
3	5	15	0.7	0.49	2.45	0.343	**1.715**
4	3	12	1.7	2.89	8.67	4.913	**14.739**
Toplam	**20**	**46**			**22.2**		**-2.52**

İlk aşamada serinin üçüncü momentini ve standart sapmasını hesaplamalıyız. Serinin üçüncü momenti ve standart sapması yukarıdaki tablo yardımıyla aşağıdaki gibi hesaplanmıştır.

$$\mu = \frac{\sum_{i=1}^{k} f_i X_i}{\sum_{i=1}^{k} f_i} = \frac{46}{20} = 2.3,$$

$$\sigma = \sqrt{\mu_2} = \sqrt{\frac{\sum\limits_{i=1}^{k} N_i \left(X_i - \mu\right)^2}{\sum\limits_{i=1}^{k} N_i}} = \sqrt{\frac{22.2}{20}} = \sqrt{1.11} = 1.05 \Rightarrow \sigma^3 = 1.17.$$

$$\mu_3 = \frac{\sum_{i=1}^{k} N_i (X_i\text{-}\mu)^3}{\sum_{i=1}^{k} N_i} = \frac{-2.52}{20} = -0.126$$

$$\alpha_3 = \frac{\mu_3}{\sigma^3} = \frac{-0.126}{1.17} = -0.11$$

Serinin asimetrisi yani çarpıklığı negatiftir (sola çarpık) ve katsayısı - 0.11'dir.

Genel olarak, $|ÇK| > 0.50$ ise kuvvetli çarpık ve -0.5 ≤ ÇK ≤ 0.5 ise zayıf çarpık denilebilir.

2.5. Momentler Yardımıyla Basıklık (Kurtosis) Katsayısının Hesaplanması

İki veya daha fazla serinin aritmetik ortalaması ve standart sapmaları aynı olsa bile frekans dağılımlarının yüksekliği (Basıklık) farklı olabilir. Bu durumlarda serilerin frekans dağılımının basıklığı o seri hakkında bazı ilave bilgiler içermektedir ve hesaplanmasında fayda vardır.

Bir serinin frekans dağılımının basıklığı **dördüncü moment** yardımıyla hesaplanabilir. Serinin hesaplanan dördüncü dereceden momenti yine aynı serinin standart sapmasının dördüncü kuvvetine bölünürse standart bir basıklık ölçü birimi elde edilmiş olur. Momentler yardımıyla basıklık katsayısı (BK) aşağıdaki şekilde formüle edilebilir ve yorumlanabilir.

$$BK = \alpha_4 = \frac{\frac{\sum_{i=1}^{k} N_i (X_i\text{-}\mu)^4}{\sum_{i=1}^{k} N_i}}{\sigma^4} = \frac{\mu_4}{\sigma^4} = 3 \Rightarrow$$ Serinin dağılımının yüksekliği standart normal dağılıma uygundur.

$$\alpha_4 = \frac{\mu_4}{\sigma^4} > 3 \Rightarrow$$ Serinin dağılımının yüksekliği standart normal dağılımın yüksekliğinden daha sivridir.

$$\alpha_4 = \frac{\mu_4}{\sigma^4} < 3 \Rightarrow$$ Serinin dağılımının yüksekliği standart normal dağılımın yüksekliğinden daha basıktır.

ÖRNEK: Bir mahallede yaşayan aileler çocuk sayılarına göre tasnif edilmiş seri olarak aşağıda verilmiştir. Bu dağılımın çarpıklık katsayısını momentler yardımıyla hesaplayınız ve yorumlayınız?

Çocuk Sayısı	Aile Sayısı

(X_i)	f_i	f_iX_i	$(X_i-\mu)$	$(X_i-\mu)^2$	$f_i(X_i-\mu)^2$	$(X_i-\mu)^4$	$f_i(X_i-\mu)^4$
0	1	0	-2.3	5.29	5.29	27.98	27.98
1	3	3	-1.3	1.69	5.07	2.86	**8.57**
2	8	16	-0.3	0.09	0.72	0.0081	**0.06**
3	5	15	0.7	0.49	2.45	0.24	**1.20**
4	3	12	1.7	2.89	8.67	8.35	**25.06**
Toplam	**20**	**46**			**22.2**		**62.87**

İlk aşamada serinin dördüncü momentini ve standart sapmasını hesaplamalıyız. Serinin dördüncü momenti ve standart sapması yukarıdaki tablo yardımıyla aşağıdaki gibi hesaplanmıştır.

$$\mu = \frac{\sum_{i=1}^{k} f_i X_i}{\sum_{i=1}^{k} f_i} = \frac{46}{20} = 2.3,$$

$$\mu_4 = \frac{\sum_{i=1}^{k} N_i (X_i - \mu)^4}{\sum_{i=1}^{k} N_i} = \frac{62.87}{20} = 3.1435$$

$$\sigma = \sqrt{\mu_2} = \sqrt{\frac{\sum_{i=1}^{k} N_i (X_i - \mu)^2}{\sum_{i=1}^{k} N_i}} = \sqrt{\frac{22.4}{20}} = \sqrt{1.12} = 1.06 \Rightarrow \sigma^4 = 1.26.$$

$$\sigma = \sqrt{\mu_2} = \sqrt{\frac{\sum_{i=1}^{k} N_i (X_i - \mu)^2}{\sum_{i=1}^{k} N_i}} = \sqrt{\frac{22.2}{20}} = \sqrt{1.11} = 1.05 \Rightarrow \sigma^4 = 1.2321.$$

$$\alpha_4 = \frac{\mu_4}{\mu_2^2} = \frac{3.1435}{1.2321} = 2.55<3.$$

Serinin basıklık katsayısı 2.55'dır. Seri standart normal dağılıma göre daha basıktır.

ÖRNEK: Sınıflanmış bir seriye ilişkin $\mu_2 = 8, \mu_3 = 0,\ \mu_4 = 71.476$ olsun. Buna göre momentlere dayalı asimetri ve basıklık ölçüleri nedir.

$\alpha_3 = \frac{\mu_3}{\sqrt{\mu_2^3}} = \frac{0}{8^{\frac{3}{2}}} = 0$ seri simetriktir.

$\alpha_4 = \frac{\mu_4}{\mu_2^2} = \frac{71.476}{(8)^2} = 1.117$, α_4<3 olduğu için seri basıktır.

ÖRNEK: Sınıflanmış Seride

X	N	NX	$(X-\bar{X})$	$(X-\bar{X})^2$	$N(X-\bar{X})^2$	$N(X-\bar{X})^3$	$N(X-\bar{X})^4$
32	4	128	-6	36	144	-864	5184
36	20	720	-2	4	80	-160	320
38	24	912	0	0	0	0	0
40	32	1280	2	4	128	256	512
					∑=352	∑=-768	∑=6016

$$\bar{X} = \frac{\sum NX}{\sum N} = \frac{3040}{80} = 38$$

$\alpha_3 = \frac{\mu_3}{\sqrt{\mu_2^3}} = \frac{-9.6}{\sqrt{(4.4)^3}} = -1.04$ <0 olduğu için sola eğik ve mutlak değerce 0.5'den büyük olduğu için asimetrisi kuvvetlidir.

μ_4=76.4 $\quad \mu_3 = -9.6 \quad \mu_2 = 4.4$

$$\alpha_4 = \frac{\mu_4}{\mu_2^2} = \frac{76.4}{19.36} = 3.8573$$

α_4>3 olduğu için seri standart normal dağılıma göre daha sivridir.

Momentlere Dayanan Basıklık Ölçüsü, Bir serinin normal olup olmadığı bir serinin simetrik($\alpha_3 = 0$) yanında belirli bir yüksekliğe bağlı olmasını ($\alpha_4 = 3$) belirlemek için kullanır.

$$\boxed{\alpha_4 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4}}$$

Normal bir seride $\alpha_4 = 3$ sivri bir seride $\alpha_4 > 3$ ve basık bir seride $\alpha_4 < 3$'dir. Her iki ölçüde de aritmetik ortalamaya göre momentler kullanılır. Seri gruplanmış bir seri ise asimetrik ve basıklık ölçüsünü bulmadan önce 2. ve 4. momentlerin düzeltilmiş değerleri buluruz. μ_2 ve μ_4 yerine düzeltilmiş değerler konulur. **Gruplanmış serilerde formüle** μ_2^1 ve μ_4^1 uygulanır.

Gruplanmış bir seriye ilişkin, $\mu_2^1 = 9.2$, $\mu_3 =$ -3.6, $\mu_4^1 = 122$ olsun

$\alpha_3 = \dfrac{\mu_3}{\sqrt{\left(\mu_2^1\right)^3}} = \dfrac{-3.6}{\sqrt{(9.2)^3}} = -0.129 < 0$ olduğu için sola eğik seridir.

$\alpha_4 = \dfrac{\mu_4^1}{\left(\mu_2^1\right)^2} = \dfrac{122}{(9.2)^2} = 1.44 < 3$ basık seridir.

ÖRNEK: Gruplanmış bir seriye ilişkin $\mu_2 = 8, \mu_3 = 0, \ \mu_4 = 128$, sınıf aralığı=4 olsun. Buna göre momentlere dayalı asimetri ve basıklık ölçüleri nedir. Gruplanmış seri olduğundan önce düzeltilmiş momentler hesaplanır.

$$\mu_2' = \mu_2 - \frac{S^2}{12} = 8 - \frac{4^2}{12} = 6.67$$

$$\alpha_3' = \frac{\mu_3}{\sqrt{\mu_2'^3}} = \frac{0}{6.67^{\frac{3}{2}}} = 0 \text{ simetriktir.}$$

$$\boxed{\mu_4' = \mu_4 - \frac{S^2}{2}\mu_2 + \frac{7S^4}{240}} = 128 - (8).8 + \frac{7.(256)}{240} = 71.460$$

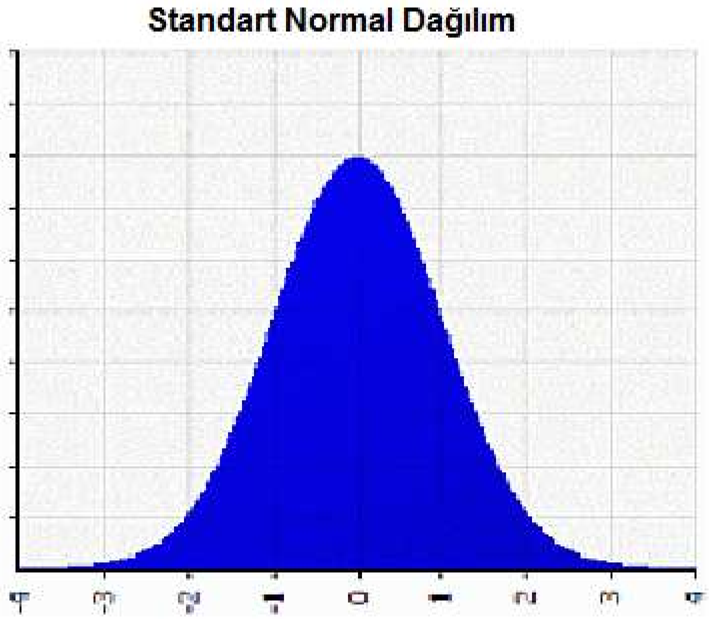
$$\alpha_4' = \frac{\mu_4'}{\mu_2'^2} = \frac{71.46}{(6.67)^2} = 1.606, \quad \alpha_4 < 3 \text{ olduğu için seri basıktır.}$$

3. ORTALAMALAR, ÇARPIKLIK KATSAYISI VE BASIKLIK KATSAYISI YARDIMIYLA BİR SERİNİN ANALİZİ

Seriler frekans dağılımı çarpıklık katsayısına göre simetrik, sağa çarpık ve sola çarpık olarak üç olasılığa sahipken, basıklık katsayısına göre de normal, basık ve sivri olmak üzere üç olasılık mevcuttur. Bir serinin çarpıklığı (Skewness) ve basıklığı (Kurtosis) aşağıdaki tablo yardımıyla özetlenebilir.

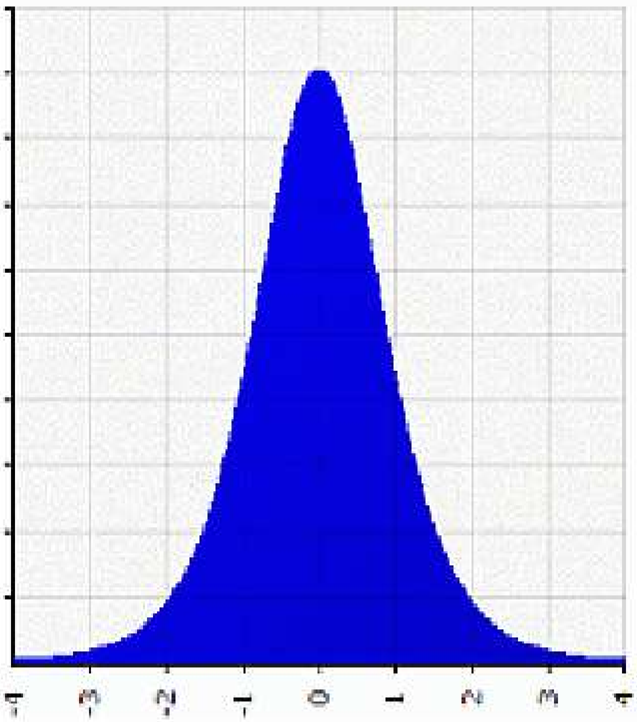
	ÇK = 0	**ÇK > 0**	**ÇK < 0**
BK = 3	Simetrik ve basıklığı normal		
BK > 3		Asimetrisi pozitif ve normale göre sivri	Asimetrisi negatif ve normale göre sivri
BK < 3		Asimetrisi pozitif ve normale göre basık	Asimetrisi negatif ve normale göre basık

Serilerin çarpıklık ve basıklık katsayısının serilerin frekansının dağılımını nasıl belirlediği frekans dağılımının grafiği yardımıyla da görülebilir.



μ = 0, σ = 1, ÇK= 0 ve BK = 3

Yukarıda standart normal bir dağılımın grafiği mevcuttur. Serinin ÇK=0 olması simetrik ve BK = 3 olması basıklığının normal olduğunu göstermektedir.



μ = 0, σ = 0.5, ÇK = 0, BK = 4.2

Yukarıda normal bir dağılımın grafiği mevcuttur. Serinin ÇK=0 olması simetrik ve

BK > 3 olması basıklığının normalden sivri olduğunu göstermektedir. Dikkat edilirse diğer veriler sabitken serinin standart sapmasının azalması BK'nı artıracağından serinin frekans dağılımının sivriliği artacaktır.

5. STANDARTLAŞTIRILMIŞ DEĞİŞKEN

$\boxed{Z_i = \frac{\left(X_i - \bar{X}\right)}{s}}$ veya $\boxed{Z_i = \frac{\left(X_i - \mu_x\right)}{\sigma_x}}$ olup genelde Z puan hesaplarında kullanılır.

ÖRNEK: Bir öğrencinin matematik puanı=84 ve matematik notları serisinin ortalaması $\bar{X}_{mat} = 76$ ve standart sapması $s_m = 10$ dır. Fizik dersi notu 90, sınıf not ortalaması $\bar{X}_{fizik} = 82$, standart sapması $s_f = 32$,dır. Buna göre Z_m, Z_f değerlerini hesaplayınız?

Çözüm:

$$Z_m = \frac{84 - 76}{10} = 0.8$$

$$Z_f = \frac{90 - 82}{32} = 0.25$$

$\Rightarrow$ Z_m= 0.8 olduğundan öğrenci matematikte ortalamanın üzerinde "0.8 standart sapma" kadar bir puan almıştır.

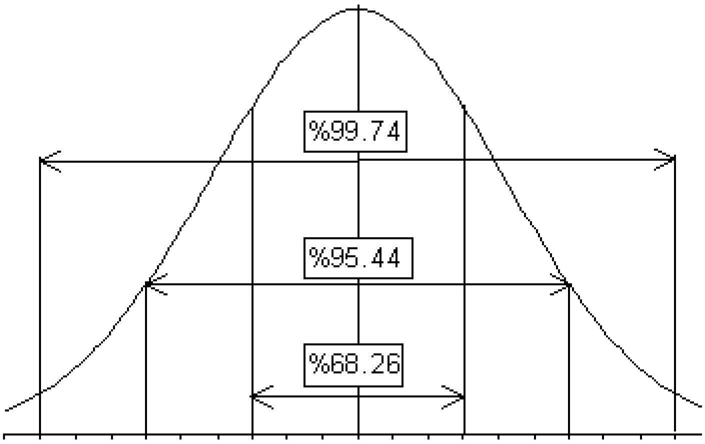
Z_f=0.25 olduğundan öğrenci ortalamanın üzerinde 0.5 standart sapma kadar puan almıştır.

$Z_M = 0.8 > Z_f = 0.25$ olduğundan öğrenci nispi olarak matematikte daha iyi konumdadır.

Oysa sınıf ortalamasına göre öğrenci fizikten daha iyi konumda olduğu kanaati oluşacaktı.

Standart Normal dağılım nedir?

- Varyansı 1, ortalaması sıfır olan,
- Frekans eğrisi çan şeklinde olan simetrik dağılımdır.
- Normal dağılım simetrik olduğu için, normal dağılım gösteren değişkenlerin ortalama, ortanca ve modları eşittir.



Normal Dağılım özelliğinin önemi nedir

- Parametrik testlerin tümünün uygulanabilmesi için gereken varsayımların başında verilerin dağılımının normal olması gelir. *Normal dağılımdan gelmeyen ölçümler kullanıldığında, gerçekte olduğundan daha küçük bir olasılık değeri ya da daha dar bir güven aralığı hesaplanır. Bu durumda, doğru bir hipotezi reddetme olasılığı artar.* Yani, iki grup arasında fark olmadığı halde fark varmış gibi sonuç elde edilebilir

Normal Dağılım Kriterleri

- Dağılımın normal olup olmadığı grafik ve istatistik analiz yöntemleri ile anlaşılır. Histogram, dal ve yaprak grafiği ve normal olasılık grafiği çizilerek dağılımın normal olup olmadığı hakkında fikir edinilebilir.
- Ama bu izlenimin istatistik yöntemlerle de test edilmesi gerekir. Shapiro-Wilks (n<30) ve Lilliefors (n>30) Kolmagorw Simirnov. Yada Shefi testleri bu amaçla sıklıkla kullanılan testlerdir. Bu testlerde p değeri <0.05 ise dağılımın normal olmadığı sonucuna varılır.
- Örneklem büyüklüğü arttıkça, deneklerin dağılımı ve ortalamanın örneklem dağılımı normal dağılıma yaklaşır.
- Genellikle bir örneklemde 30 ya da daha fazla sayıda denek varsa, evren normal dağılım göstermiyorsa bile, ortalamanın örneklem dağılımının normal olduğu varsayılabilir

Verilerin normal dağılmadığı durumlarda iki işlem yapılabilir :

1. Verilere dönüşüm uygulayarak, onların normal dağılıma uymalarını sağlamak.
2. Varolan verilere parametrik olmayan bir test uygulamak

Normal Dağılım sınaması icin hipotezler şöyle ifade edilir:

H_0: Veriler normal dağılım gösterir.

H_1: Veriler normal dağılım göstermez.

Jarque ve Bera sınaması bir Lagranj çarpanı prensipine dayanan bir sınama tipindendir. Sınama istatistiği örneklem basıklık ve çarpıklık ölçülerinin dönüşümlerinden elde edilmiştir. Sıfır hipotezi daha ayrıntılı olarak bir bileşik hipotezdir: beklenen çarpıklığın 0 değerde ve beklenen basıklık fazlalığının 3 değerde olacağı sıfır hipotezdir; çünkü bir normal dağılım için bu değerler gereklidir.

Sınama istatistiği olan JB şöyle elde edilir:

$$JB = \frac{n}{6}\left(S^2 + \frac{(\text{K-3})^2}{4} \right)$$

- Burada *n* gözlem sayısı (veya genellikle serbestlik derecesi); *S* örneklem çarpıklık ölçüsü, *K* örneklem basıklık ölçüsü olur ve bu son iki istatistik şöyle tanımlanır:

$$S=\frac{\mu_3}{\sigma^3}=\frac{\mu_3}{\left(\sigma^2\right)^{3/2}}=\frac{\frac{1}{n}\sum_{i=1}^{n}\left(x-\bar{x}\right)^3}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(x-\bar{x}\right)^2\right)^{3/2}}$$

$$K=\frac{\mu_4}{\sigma^4}=\frac{\mu_4}{\left(\sigma^2\right)^{2}}=\frac{\frac{1}{n}\sum_{i=1}^{n}\left(x-\bar{x}\right)^4}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(x-\bar{x}\right)^2\right)^{2}}$$

Burada $\bar{x}$örneklem ortalaması, σ^2 ikinci moment veya varyans ve sırasıyla μ_3 ve μ_4 üçüncü ve dördüncü merkezsel momentlerdir.

JB sınama istatistiği asimptotik olarak **2** serbestlik derecesi bulunan bir ki-kare dağılımına yaklaşır. Örneklem çarpıklığı '0'dan ve basıklığı '3'den sapma gösterdikçe, *JB* sınama istatistiği büyüme gösterir.

Bu sınama çok kere ekonometriciler tarafından çoklu doğrusal regresyon kestirim sonuçları elde edildikten sonra ele geçen hataların normal dağılım gösterip göstermediğini araştırmak için kullanılır. Bazı ekonometriciler bu sınama istatistiğinin bu hallerde, bağımsız değişken sayısı olan *k* ile düzeltilmesini önermişlerdir.

ÖRNEK: $\alpha_3=-0.129$ ve $\alpha_4=1.44$, n=40 verilerine göre serinin normal dağılımını α=0.05 anlamlılık seviyesinde test ediniz.

H_0: Veriler normal dağılım gösterir

H_1: Veriler normal dağılım göstermez.

Sd=2 , anlamlılık seviyesi 0.05 için kritik değeri; $\chi^2_{2,0.05}=5.99$ dir.

Test istatistiği;

$$\chi^2=\frac{n}{6}\left(S^2+\frac{(\text{K-3})^2}{4}\right)=\frac{n}{6}\left(\alpha_3^2+\frac{(\alpha_4\text{-3})^2}{4}\right)=\frac{40}{6}\left((\text{-0.129})^2+\frac{(1.44\text{-3})^2}{4}\right)=\frac{40}{6}\left(0.016641+\frac{(2.4336)}{4}\right)$$

$\chi^2=\frac{40}{6}(0.625041)=4.16694<\chi^2_{2,0.05}=5.99$ olduğundan H₀ hipotezi kabul edilir. Seri normal dağılıma sahiptir; $X\sim N\left(\mu_x,\sigma_X^2\right)$.

6. KONING TEOREMİ

Aritmetik ortalamaya göre momentlerin sıfıra göre momentler cinsinden hesaplanması ile ilgili teoremdir.

μ’ler ile M’ler arasındaki bağıntı

$$\boxed{\mu_2 = M_2 - M_1^2}$$

$$\boxed{\mu_3 = M_3 - 3M_1M_2 + 2M_1^3}$$

$$\boxed{\mu_4 = M_4 - 4M_1M_3 + 6M_1^2M_2 - 3M_1^4}$$

Aritmetik ortalamadan sapmaya göre hesaplanan çift dereceden moment hesabında gruplaşmış seriler söz konusu olduğunda Shepard düzeltmesi yapılır.

6.1. Basit Serilerde Momentler

X$_i$	X_i^2	X_i^3	X_i^4
2	4	8	16
3	9	27	81
7	49	343	2401
8	64	512	4096
10	100	1000	10000
∑X=30	$\sum X^2$ **=226**	$\sum X^3$ **=1890**	$\sum X^4$ **=16594**

Sıfıra Göre	**Koning Teoremince Aritmetik Ortalamaya Göre Momentler**
$M_1 = \frac{\sum X}{N} = \frac{30}{5} = 6$	$\mu_1 = \frac{(X - \bar{X})}{N} = 0$
$M_2 = \frac{\sum X^2}{N} = \frac{226}{5} = 45.2$	$\mu_2 = M_2 - M_1^2 = 45.2 - 36 = 9.2$
$M_3 = \frac{\sum X^3}{N} = \frac{1890}{5} = 37.8$	$\mu_3 = M_3 - 3M_1M_2 + 2M_1^3 = 378 - 3.(6x45.2) + 2.6^3 = -3.6$
$M_4 = \frac{\sum X^4}{N} = \frac{16594}{5} = 331$	$\mu_4 = M_4 - 4M_1M_3 + 6M_1^2M_2 - 3M_1^4 = 33188 - 4.(6x37.8) + 6.(6^2x45.2) - 3.6^4 = 122$

6.2. Sınıflanmış Serilerde Momentler

Sıfıra Göre Momentler

X_i	**N_i**	X_i^2	X_i^3	X_i^4	**N_i X_i**	**N_i** X_i^2	**N_i** X_i^3	**N_i** X_i^4
2	3	4	8	16	6	12	24	48
3	6	9	27	81	18	54	162	486
4	4	16	64	256	16	64	256	1024
6	7	36	216	1296	42	252	1512	9072
	$\sum N_i=20$				$\sum N_iX_i=82$	$\sum N_i X_i^2=382$	$\sum N_i X_i^3=1954$	$\sum N_i X_i^4=10630$

$$M_1=\frac{\sum X}{N}=\frac{82}{20}=4.1 \qquad M_2=\frac{\sum X^2}{N}=\frac{382}{20}=19.1$$

$$M_3=\frac{\sum X^3}{N}=\frac{1954}{20}=97.7 \qquad M_4=\frac{\sum X^4}{N}=\frac{10630}{20}=531.5\,.$$

Aritmetik Ortalamaya Göre Momentler

X_i	N_i	$\left(X_i-\bar{X}\right)$	$\left(X_i-\bar{X}\right)^2$	$\left(X_i-\bar{X}\right)^3$	$\left(X_i-\bar{X}\right)^4$
2	3	-2.1	4.41	-9.261	19.4481
3	6	-1.1	1.21	-1.331	1.4641
4	4	-0.1	0.01	-0.001	0.0001
6	7	1.9	3.61	6.859	13.0321

$N_i\left(X_i-\bar{X}\right)$	$N_i\left(X_i-\bar{X}\right)^2$	$N_i\left(X_i-\bar{X}\right)^3$	$N_i\left(X_i-\bar{X}\right)^4$
-6.3	13.23	-27.783	58.3443
-6.6	7.26	-7.986	8.7846
-0.4	0.04	-0.004	0.0004

13.3	25.27	48.013	91.2247
$\sum \mathrm{N_i}\left(X_i-\bar{X}\right)$=0	$\sum \mathrm{N_i}\left(X_i-\bar{X}\right)^2$=45.8	$\sum \mathrm{N_i}\left(X_i-\bar{X}\right)^3$=12.24	$\sum \mathrm{N_i}\left(X_i-\bar{X}\right)^4$=158.354

$$\mu_1=\frac{0}{20}=0 \qquad \mu_2=\frac{45.80}{20}=2.29 \qquad \mu_3=\frac{12.24}{20}=0.612 \qquad \mu_4=\frac{158.354}{20}=7.9177$$

6.3. Gruplanmış Serilerde Momentler

Sıfıra Göre Momentler

Sınıflar	**N**	**m**	**Nm**	**N** m^2	**N** m^3	**N** m^4	$(m-\bar{X})$	$(m-\bar{X})^2$	$(m-\bar{X})^3$	$(m-\bar{X})^4$
0-4'den az	2	2	4	8	16	32	-4	16	-64	256
4-8'den az	4	6	24	144	864	5184	0	0	0	0
8-12'den az	2	10	20	200	2000	20000	4	16	64	256
$\sum$	**8**		**48**	**352**	**2880**	**25216**				

$$M_1=\frac{\sum NM}{\sum N}=\bar{X}=\frac{48}{8}=6 \qquad M_3=\frac{\sum NM^3}{\sum N}=\frac{2880}{8}=360$$

$$M_2=\frac{\sum NM^2}{\sum N}=K^2=\frac{352}{8}=44 \qquad M_4=\frac{\sum NM^4}{\sum N}=\frac{25126}{8}=3152$$

Koning Teoremince Aritmetik Ortalamaya Göre Momentler;

$$\boxed{\mu_2=M_2-M_1^2}=44-36=8\,, \qquad \mu_2^1=\mu_2-\frac{S^2}{12}=8-\frac{4^2}{12}=6.67$$

$$\boxed{\mu_3=M_3-3M_1M_2+2M_1^3}=0-3.(44).(0)+2.(0^3)=0$$

$$\boxed{\mu_4=M_4-4M_1M_3+6M_1^2M_2-3M_1^4}=3152-(4).(6).(360)+(6).(6^2).(44)-(3).(6^4)=128$$

$$\boxed{\mu_4^1=\mu_4-\frac{S^2}{2}\mu_2+\frac{7S^4}{240}}=128-(8).8+\frac{7.(256)}{240}=71.46$$

Aritmetik Ortalamaya Göre Momentler

$\mathbf{N}(m-\bar{X})$	$\mathbf{N}(m-\bar{X})^2$	$\mathbf{N}(m-\bar{X})^3$	$\mathbf{N}(m-\bar{X})^4$
-8	32	-128	512
0	0	0	0
8	32	128	512
$\sum\mathbf{N}(m-\bar{X})$ **=0**	$\sum\mathbf{N}(m-\bar{X})^2$ **=64**	$\sum\mathbf{N}(m-\bar{X})^3$**=0**	$\sum\mathbf{N}(m-\bar{X})^4$ **=1024**

$$\mu_1=\frac{\sum \mathrm{N_i}\left(X_i-\bar{X}\right)}{\sum N_i}=\frac{0}{8}=0$$

$$\mu_2=\frac{\sum N_i\left(m_i-\bar{X}\right)^2}{\sum N_i},\ \mu_2=\frac{64}{8}=8 \text{ ve } \mu_2^1=\mu_2-\frac{S^2}{12}=8-\frac{4^2}{12}=6.67 \text{ (DÜZELTİLMİŞ)}$$

$$\mu_3=\frac{\sum N_i\left(m_i-\bar{X}\right)^3}{\sum N_i},\ \mu_3=\frac{0}{8}=0$$

$$\mu_4=\frac{\sum N_i\left(m_i-\bar{X}\right)^4}{\sum N_i}=\frac{1024}{8}=128,\ \boxed{\mu_4^1=\mu_4-\frac{S^2}{2}\mu_2+\frac{7S^4}{240}}=128-(8).8+\frac{7.(256)}{240}=71.460$$

Koning Teoremince, aritmetik ortalamaya göre momentler sıfıra göre momentlerden elde edildikten sonra gruplanmış serilerde çift sayılı momentlerde düzeltme uygulanarak düzeltilmiş momentler elde edilir.

KAYNAKLAR:

1. Yılmaz Özkan, Uygulamalı İstatistik 1, Sakarya Kitapevi, 2008.
2. Özer Serper, Uygulamalı İstatistik 1, Filiz Kitapevi, 1996.
3. Meriç Öztürkcan, İstatistik Ders notları, YTÜ.
4. Andım Oben Balce ve Serdar Demir, İstatistik Ders Notları, Pamukkale Üniversitesi, 2007.
5. Ayşe Canan Yazıcı, Biyoistatistik Ders Notları, Başkent Üniversitesi.
6. Zehra Muluk ve Yavuz Eren Ataman, Biyoistatistik ve Araştırma Teknikleri Ders Notları, Başkent Üniversitesi.

İST292 STATISTICS LESSON 5
INTERVAL ESTIMATION

5. INTERVAL ESTIMATION

Suppose a sample from a population with mean μ and variance σ^2, $X_1, X_2, ..., X_n$, $\bar{x}$ represents to the point estimation of the population mean μ. How can we assess the accuracy of this point estimation/estimator. The ***Central Limit Theorem*** and ***sampling distributions*** help us at this point. For large samples, according to the Central Limit Theorem, the distibution of the sample mean $\bar{X}$, is approximately normal with μ and variance σ^2/n. Using the distribution of $\bar{X}$ we construct ***an interval estimation of the population parameter μ.*** Basically approximately 95% of all values fall between $2\sigma_{\bar{X}}$ away to the mean μ. Hence the interval $\bar{x} \pm 2\sigma_{\bar{x}}$ will contain the mean μ with a probability approximately equal to 0.95. In otherwords, approximately 95% of intervals would contain μ if 100 repeated random samples were drawn from this population. Since there is now way of knowing whether our sample interval is one of the 95% that contain μ or one of the 5% that does not, but the odds (ihtimal, olasılık) certainly favor its containing μ. An importing point here is that the interval estimation is associated with confidence level such as 95%, 99%. That's why we prefer an interval estimation of a parameter to the point estimation of the parameter. The convidence level is refered by 1-α where α is called as ***significance level.***

5.1. Confidence Interval of a Population Mean μ

Suppose a random sample of size n from a **normal population** with mean μ and variance σ^2, $X_1, X_2, ..., X_n$, and thus $\bar{X}$ is a normal distributed random variable with μ and variance σ^2/n.

5.1.1. When the Population Variance σ^2 is Known

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1),$$

the confidence interval of μ is constructed by:

$$P(-z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}) = 1-\alpha$$

$$P(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha .$$

We assert (idda etmek) with (1-α)100% confidence that the interval from $\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ to $\overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ contains the true mean of the population.

If $\overline{X}$, the mean of a random sample of size n from normal population with known variance σ^2, is to be used as an estimator of the mean of the population, the probability is 1-α that the error will be less than $z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$.

NOTE: If we have a sample from nonnormal population with known population variance, by virtue of the limit theorem, this result can be also used for random samples from nonnormal populations provided that n is sufficienly large; that is, n ≥ 30.

Example 1: A team of efficiency experts intends to use the mean of random sample of size n=150 to estimate the average mechanical aptitude (uygunluk, kabiliyet) of assembly-line (montaj hattı) workers in a large industry and suppose that they get $\overline{x} = 69.5$. If, based on experince, the efficiency experts can assume that σ=6.2 for such data, what can they assert (iddia etmek, ileri sürmek) with probability 0.99 about maximum error of their estimate?

Solution: Since n=150 (n≥30), as a result of cental limit theorem, substituting n=150, σ=6.2 and and $z_{0.005} = 2.575$ into the formula of

$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 2.575 \times \frac{6.2}{\sqrt{150}} = 1.30$ as the maximum error of the estimate of μ.

$$P(69.5 - 2.575 \times \frac{6.2}{\sqrt{150}} < \mu < 69.5 + 2.575 \times \frac{6.2}{\sqrt{150}}) = 0.99$$

$$P(68.2 < \mu < 70.8) = 0.99$$

The 99% confidence interval of μ is equal to 69.5 ± 1.30 that is, (68.2, 70.8).

5.1.2. When the Population Variance σ^2 is Unknown

If the sample size **n is enough large (n≥30)**, $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ is approximately $N(0,1)$,

$$P(-z_{\alpha/2} < \frac{\overline{X}-\mu}{S/\sqrt{n}} < z_{\alpha/2}) = 1-\alpha$$

$$P(\overline{x} - z_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \overline{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}) = 1-\alpha.$$

Example 2: Suppose a large hospital wants to estimate the average length of time patients remain in the hospital. It is assumed that the length of time patients staying in the hospital has normal distrubition. To accomplish this objective, the hospital administrators plan to sample 100 of all previous patients records. Find a point estimation and a confidence interval of the mean stay, μ, of all patients' visits using given $\sum_{i=1}^{100} x_i = 465\ days$ $\sum_{i=1}^{100}\left(x_i - \overline{x}\right)^2 = 2387$ for α=%5.

Solution: The sample mean $\overline{x} = 4.65$ represents a point estimation of population mean μ.

$$s^2 = \frac{\sum_{i=1}^{100}\left(x_i-\overline{x}\right)^2}{n-1} = \frac{2387}{99} = 24.11$$ and s=4.9 days.

Then we calculate the interval

$$\overline{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} = 4.65 \pm 1.96 \times \frac{4.9}{10}$$

$$P(4.65 - 1.96\times\frac{4.9}{10} < \mu < 4.65 + 1.96\times\frac{4.9}{10}) = 0.95$$
$$P(3.69 < \mu < 5.61) = 0.95$$

That is, we estimate the mean length stay in the hospital for all patients to fall in the interval 3.69 and 5.61 days with 95 % confidence level.

NOTE: If we have a sample from nonnormal population with unknown population variance, by virtue of the limit theorem, this result can be also used for random samples from nonnormal populations provided that n is sufficienly large; that is, n ≥ 30. In that case, we may also subsitute for σ the value of the sample standard devaition S

But if the sample size n is not enough large (n<30), $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ is not approximately $N(0,1)$, but we know that $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ has a student t distribution with (n-1) degrees of freedom.

$$P(-t_{\alpha/2,(n-1)} < \frac{\bar{X}-\mu}{S/\sqrt{n}} < t_{\alpha/2,(n-1)}) = 1-\alpha$$

$$P(\bar{x} - t_{\alpha/2,(n-1)}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2,(n-1)}\frac{s}{\sqrt{n}}) = 1-\alpha\,.$$

We refer to it as a small-sample confidence interval for μ.

Example 3: A major car manufacture wants to test a new engine to determine whether it meets new air pollution standards. The mean emission μ of all engines of this type must be less than 20 parts per million carbon. It is assumed that the emission measures are normally distributed. Ten engines are manufactured for testing purposes and the mean and the standard deviation of the emission for this sample of engines are determined to be $\bar{x} = 17.1$ parts per million of carbon and s=3.0 parts per million of carbon. Find the 95% confidence interval of the true mean emission. Could you decide whether a new type engine meets the new air pollution standards.

Solution: We calculate the interval by:

$$\bar{x} \pm t_{\alpha/2,(n-1)}\frac{s}{\sqrt{n}} = 17.1 \pm 2.262 \times \frac{3}{\sqrt{10}}$$ ($t_{0.05/2,(10-1)}$=$t_{0.025,9}$=2.262)

$$P(17.1 - 2.262 \times \frac{3}{\sqrt{10}} < \mu < 17.1 + 2.262 \times \frac{3}{\sqrt{10}}) = 0.95$$

$$P(14.95 < \mu < 19.25) = 0.95$$

That is, the interval 14.95 parts per million to 19.25 parts per million contains the true mean emission with 95% confidence. Hence, a new type engine meets the new air pollution standards.

Summary of Which Statistics Used in Each Case of Confidence Interval of a Population Mean μ

	The Distribution of Population is Normal – N(μ, σ^2)		**The Distribution of Population is Nonnormal with mean μ and variance σ^2**	
Sample Size	**with known population variance σ^2**	**with unknown population variance σ^2**	**with known population variance σ^2**	**with unknown population variance σ^2**
n≥30	z statistic (we use σ^2 in formula)	z statistic (we use s^2 in formula)	As a result of central limit theorem, z statistic (we use σ^2 in formula)	As a result of central limit theorem, z statistic (we use s^2 in formula)
n<30	z statistic (we use σ^2 in formula)	t statistic (we use s^2 in formula)	**In that case n must be made larger we do not know a special statistic for this case**	**In that case n must be made larger we do not know a special statistic for this case**

5.2. Confidence Interval of a Population Variance σ^2

Given a random sample of size n from a normal population, we can obtain (1-α)100% confidence interval for σ^2 by making use of the sampling distribution according to which

$$\frac{(n-1)\mathrm{S}^2}{\sigma^2}$$

a random variable having a chi-square distribution with *n-1* degrees of freedom. Thus,

$$P\left(\chi^2_{1-\alpha/2,(n-1)} < \frac{(n-1)\mathrm{S}^2}{\sigma^2} < \chi^2_{\alpha/2,(n-1)}\right) = 1-\alpha$$

$$P\left(\frac{(n-1)s^2}{\chi^2_{\alpha/2,(n-1)}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,(n-1)}}\right) = 1-\alpha$$

Example 4: In 16 tests runs the gasoline consumption of an experimental engine had a standard deviation of 2.2 of gallons. It is assumed that the gasoline consumption of engine has normal distibution. Construct a 99% confidence interval for σ^2, which measures the true variability of the gassoline consumption of the engine.

Solution: We calculate the interval by using the the values $\chi^2_{0.005,15} = 32.801$ and $\chi^2_{0.995,15} = 4.601$ obtained from the chi-square table.

$$P\left(\frac{15(2.2)^2}{32.801} < \sigma^2 < \frac{15(2.2)^2}{4.601}\right) = 0.99$$

$$P\left(2.21 < \sigma^2 < 15.78\right) = 0.99$$

The true variability of the gassoline consumption of the engine, σ, falls in the interval 2.21 to 15.78 gallons with 99% confidence.

5.3. Confidence Interval of a Proportion

In many problems we must estimate proportions, probabilities or rates such as the proportion of defectives in a large shipment (yük, sevkiyat) of transistors, the probability that a car will be written traffic ticket in a particular day, the mortality rate of a disease. For many of these examples, we assume that we are sampling a binomial population and hence that our problem is to estimate the binomial parameter *p*. We can use of the fact that for large n the binomial distribution can be approximated with a normal; that is,

$$Z = \frac{X - np}{\sqrt{np(1-p)}} \sim N(0,1) \text{ (as result of central limit theorem, } n \to \infty \;\; X \sim N(np, np(1-p)) \text{)}$$

can be treated as a random variable having approximately the standard normal distribution. The confidence interval of the ratio *p* corresponding to the proportion of events interested in a population is constructed with using the same way given previously:

$$P(-z_{\alpha/2} < \frac{X - np}{\sqrt{np(1-p)}} < z_{\alpha/2}) = 1 - \alpha$$

$$P(\frac{x}{n}-z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}<p<\frac{x}{n}+z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}})=1-\alpha\,.$$

Substituted $\hat{p}=\frac{x}{n}$ for p in two sides of the equation, we get a following formula

$$P(\hat{p}-z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}<p<\hat{p}+z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}})=1-\alpha$$

as an approximate (*1-α*)100% confidence interval for *p*.

Example 5: A study is made to determine the proportion of the voters in a sizable community who favor the construction of a nuclear power plant (tesis). If 140 of 400 voters selected at random favor the project and we use $\hat{p}=\frac{140}{400}=0.35$ as an estimate of the actual proportion of all voters in the community who favor the project, what can we say with 99% confidence about the maximum error.

Solution: Substituting n=400, $\hat{p}=0.35$ and $z_{0.005}=2.575$ into the formula of

$$z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}=2.575\times\sqrt{\frac{(0.35)(0.65)}{400}}=0.061\cong 0.06$$

$$P(0.35-2.575\times\sqrt{\frac{0.35\times 0.65}{400}}<p<0.35+2.575\times\sqrt{\frac{0.35\times 0.65}{400}})=0.99$$

0.061 or 0.06 rounded to two decimals. From using this maximum error, 99% confidence interval of the actual proportion of voters in the community who favor the project is calculated by 0.35 ± 0.06 and so the actual proportion of voters in the community who favor the project falls in the interval (0.29 , 0.41) with 99% confidence.

INTERVAL ESTIMATION FOR TWO POPULATIONS

The Estimation of Differences between Two Populations Means

Let $X_{11}, X_{12}, ..., X_{1n_1} \sim N(\mu_1, \sigma_1^2)$ and $X_{21}, X_{22}, ..., X_{2n_2} \sim N(\mu_2, \sigma_2^2)$ be two independent random samples from normal populations. $\bar{X}_1$ and $\bar{X}_2$ are two sample means and their distributions are given below:

$$\bar{X}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \text{ and } \bar{X}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right).$$

A random variable $\bar{X}_1 - \bar{X}_2$ has normal distribution as it is shown

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \tfrac{\sigma_1^2}{n_1} + \tfrac{\sigma_2^2}{n_2}\right) \text{ and,}$$

$$Z = \frac{(\bar{\mathrm{X}}_1 - \bar{\mathrm{X}}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}},$$

has the standard normal distribution. If we substitute this expression for Z into

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

the pivotal method yields the following confidence interval formula for $\mu_1 - \mu_2$.

When σ_1^2 and σ_2^2 are known,

$$P\left((\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right) = 1 - \alpha$$

is a (1-α)100% confidence interval for the difference between the two population means.

When σ_1^2 and σ_2^2 are unknown, if n₁≥30 and n₂≥30

The sample variances s_1^2 and s_2^2 being the estimations of σ_1^2 and σ_2^2, are replaced into the formula given below:

$$P\left((\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right) = 1 - \alpha$$

is a (1-α)100% confidence interval for the difference between the two population means.

Example: The management of a restaurant wants to determine whether a new advertising campaign has increased its mean daily income (net). The incomes for each of 50 business days prior to the campaign's beginning are recorded. After conducting the advertising campaign and allowing a 20 day period for advertising to take effect, the restaurant management records the income for 30 business days. These two samples (assumed that they are from independent normal populations) will allow the management to make an inference about the effect of the advertising campaign on the restaurant's daily income. A summary of the two samples is shown in the table. Find a confidence interval for the difference in mean daily incomes before and after the advertising campaign.

Before campaign	After campaign
$\bar{x}_1 = \$1,255$ $s_1 = \$215$	$\bar{x}_2 = \$1,330$ $s_2 = \$238$

Solution: The 95% confidence for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} = (1,255-1,330) \pm 1.96\sqrt{\frac{215^2}{50}+\frac{238^2}{30}}$$
$$= -75 \pm 103.94$$

where we substituted s_1^2 and s_2^2 for σ_1^2 and σ_2^2, because these quantities are provide good approximations to σ_1^2 and σ_2^2 for samples as large as $n_1 \geq 30$ and $n_2 \geq 30$.

Thus, we estimate the difference in mean daily income to fall in the interval -\$178.94 to \$28.94. In other words, we estimate that μ_2, the mean daily income after advertising campaign, could be larger than μ_1, the mean daily income before adverting campaign, by as much as \$178.94 per day or it could be less than μ_1 by \$28.94 per day.

By virtue of the central limit theorem, these two confidence interval formulas above are also used for independent random samples from non-normal populations when n_1 and n_2 are large, that is $n_1 \geq 30$ and $n_2 \geq 30$. If population variances are not known, we may also substitute for σ_1 and σ_2 the values of standard deviations s_1 and s_2 obtained from samples.

Example: Construct a 94% confidence interval for the difference between the mean lifetimes of two kinds of light bulbs (ampul/elektrik lambası), given that a random sample 40 light bulbs of first kind lasted on the average 418 hours of continuous use and 50 light bulbs of the second kind lasted on the average 402 hours of continuous use. The population standard deviations are known to be $\sigma_1=26$, $\sigma_2=22$.

Solution: For α=0.06, we find $z_{0.03}$=1.88 from z distribution table. Therefore, the 94% confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = (418 - 402) \pm 1.88\sqrt{\frac{26^2}{40} + \frac{22^2}{50}}$$

which reduces to $6.3 < \mu_1 - \mu_2 < 25.7$.

Hence, we are 94% confident that the interval from 6.3 to 25.7 hours contains the actual difference between the mean lifetimes of the two kinds light bulbs. The fact that both confidence limits are positive suggests that on the average the first kind of light bulb is superior to the second kind.

When σ_1^2 and σ_2^2 are unknown, but n_1<30 and n_2<30

If $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

is a random variable having the standard normal distribution, and

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

is a unbiased estimator of σ^2 and it is called pooled estimator of σ^2 and it is referred pooled variance. The independent random variables are $\frac{(n_1 - 1)S_1^2}{\sigma^2}$ and $\frac{(n_2 - 1)S_2^2}{\sigma^2}$ having chi-squared distributed with $(n_1 - 1)$ and $(n_2 - 1)$ degrees of freedom respectively. Their sum

$$Y = \frac{(n_1 - 1)S_1^2}{\sigma^2} + \frac{(n_2 - 1)S_2^2}{\sigma^2} = \frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2}$$

has a chi-squared distribution with degrees of freedom $(n_1 + n_2 - 2)$. The random variables Z and Y are independent and then,

$$T=\frac{Z}{\sqrt{\dfrac{Y}{(n_1+n_2-2)}}}=\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{S_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$$

has a t distribution with (n_1+n_2-2) degrees of freedom. Substituting this expression for T into

$$P(-t_{\alpha/2,(n_1+n_2-2)}<T<t_{\alpha/2,(n_1+n_2-2)})=1-\alpha$$

We arrive at the following (1-α)100% confidence interval for $\mu_1-\mu_2$:

$$P\left((\bar{x}_1-\bar{x}_2)-t_{\alpha/2,(n_1-n_2-2)}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}<\mu_1-\mu_2\right.$$
$$\left.<(\bar{x}_1-\bar{x}_2)+t_{\alpha/2,(n_1-n_2-2)}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right)=1-\alpha$$

Example: Suppose a television network want to determine whether major sports events or first-run movies attract more viewers in the prime time hours. It selects 28 prime-time evenings; of these, 13 have programs devoted to major sports events and remaining15 have first-run movies. The number of viewers (estimated by a television viewer rating firm) is recorded for each program. The television network's samples produced the following results:

Sports (n_1=13)	Movies (n_2=15)
$\bar{x}_1=6.8\,million$	$\bar{x}_2=5.3\,million$
$s_1=1.8\,million$	$s_2=1.6\,million$

Construct a 95% confidence interval for the difference between the mean numbers of viewers for major sport events and first-time movies under the samples from the two normal populations with common variances.

Solution: For α=0.05, we find $t_{0.025,\,26}$=2.056 from Student's t distribution table.

$$s_p^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}=\frac{12(1.8)^2+14(1.6)^2}{13+15-2}=2.87$$

$$(\bar{x}_1-\bar{x}_2)\pm t_{\alpha/2,(n_1-n_2-2)}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}=1.5\pm 2.056\sqrt{2.87\left(\frac{1}{13}+\frac{1}{15}\right)}$$

We can be 95% confident the estimated interval form 0.1801 to 2.8198 covers the actual difference between the mean numbers of viewers for major sport events and first-time movies.

In other words, the mean numbers of viewers of sports events is larger than the mean numbers of viewers of first-time movies with 95% confident.

If $\sigma_1^2 \neq \sigma_2^2$, it is called as Behrens Fisher problem in literature. The standard form,

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

is a random variable having t distribution with ν degrees of freedom as defined below:

$$\nu \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2 \left(\frac{1}{n_1 - 1}\right) + \left(\frac{s_2^2}{n_2}\right)^2 \left(\frac{1}{n_2 - 1}\right)}.$$

We arrive at the following (1-α)100% confidence interval for $\mu_1 - \mu_2$:

$$P\left((\bar{x}_1 - \bar{x}_2) - t_{\alpha/2,(\nu)} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2,(\nu)} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right) = 1 - \alpha$$

Example: Construct 95% confidence interval for the differences between the means in the previous example under the samples from the two normal populations with uncommon variances.

Solution: For $\sigma_1^2 \neq \sigma_2^2$, we need to compute the degrees of freedom:

$$\nu \approx \frac{\left(\frac{1.8^2}{13} + \frac{1.6^2}{15}\right)^2}{\left(\frac{1.8^2}{13}\right)^2 \left(\frac{1}{12}\right) + \left(\frac{1.6^2}{15}\right)^2 \left(\frac{1}{14}\right)} \approx 24.$$

Substituted $t_{0.025,(24)}$ =2.064 in the formula we get:

$$(\bar{x}_1-\bar{x}_2)\pm t_{\alpha/2,(v)}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}=(6.8-5.3)\pm 2.064\sqrt{\frac{1.8^2}{13}+\frac{1.6^2}{15}}$$

$$=1.5\pm 1.34$$

We can be 95% confident the estimated interval form 0.16 to 2.84 covers the actual difference between the mean numbers of viewers for major sport events and first-time movies.

The Estimation of Differences between Two Population Means: Paired Difference Experiments

In some studies or researches, paired measurements of a variable for each unit, person, event or observation are obtained such as before/after, first/second, etc. In general these measurements cannot be assumed independent, because these units (persons, events or observations) are same and thus these leads to dependence between the measurements.

Let $D_i = X_{1i} - X_{2i}$ (i=1,2,…,n) be a difference between the pairs of measurements. D$_i$ is normally distributed random variables with mean $\mu_1-\mu_2$ and variance σ_D^2.

$$D_i = X_{1i} - X_{2i} \sim N(\mu_1-\mu_2,\sigma_D^2) \text{ i=1,2,…,n}$$

If σ_D^2 is <u>known</u>, $P\left(\bar{d}-z_{\alpha/2}\frac{\sigma_D}{\sqrt{n}}<\mu_1-\mu_2<\bar{d}+z_{\alpha/2}\frac{\sigma_D}{\sqrt{n}}\right)=1-\alpha$

If σ_D^2 is <u>unknown</u> and also <u>n<30</u>, $P\left(\bar{d}-t_{\alpha/2,(n-1)}\frac{s_D}{\sqrt{n}}<\mu_1-\mu_2<\bar{d}+t_{\alpha/2,(n-1)}\frac{s_D}{\sqrt{n}}\right)=1-\alpha$

where $\bar{d}=\frac{\sum_{i=1}^{n}(x_{1i}-x_{2i})}{n}$ and $s_D^2=\frac{\sum_{i=1}^{n}(d_i-\bar{d})^2}{n-1}$ for the given values of $d_i=x_{1i}-x_{2i}$, (i=1,2,…,n).

Example: Suppose you wish to compare two methods of teaching reading to slow learners by means of a 90% confidence level. Suppose it is possible to measure the slow learners “reading IQ’s” before they are subjected to a teaching method. There are eight slow learners with similar reading IQ’s, they are taught by a new teaching method after they are taught by the standard teaching method and thus their reading test scores were given for the both methods.

pair	New method	Standard method	d_i
1	77	72	5
2	74	68	6
3	82	76	6
4	73	68	5
5	87	84	3
6	69	68	1
7	66	61	5
8	80	76	4
			$\bar{d}=4.375$ $s_D=1.69$

Solution: for α=0.10, we find $t_{0.05,\,7}$=1.895 from Student's t distribution table.

$$\bar{d} \pm t_{\alpha/2,(n-1)} \frac{s_D}{\sqrt{n}} = 4.375 \pm 1.895 \frac{1.69}{\sqrt{8}}$$
$$= 4.375 \pm 1.1323$$

We are %90 confident that the interval estimation (3.2427, 5.5073) covers the actual mean of differences between the scores corresponding to standard method and new method.

The Estimation of Differences between Proportions

Let X_1 be a random variable from binomial distribution with n_1 and p_1 and X_2 be an independent random variable from binomial distribution with n_2 and p_2.

If the respective number of successes are X_1 and X_2 and the corresponding sample proportions are denoted by $\hat{p}_1 = \frac{X_1}{n_1}$ and $\hat{p}_2 = \frac{X_2}{n_2}$, let us investigate the sampling distribution of $\hat{p}_1 - \hat{p}_2$, which is an obvious estimator of $p_1 - p_2$. We have

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$$
$$V(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

and for large samples, X_1, X_2 and hence also their difference, can be approximated with normal distributions, it follows that

$$Z = \frac{\left(\frac{X_1}{n_1} - \frac{X_2}{n_2}\right) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

is a random variable having approximately the standard normal distribution. Substituting this expression for Z into $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha$, we arrive at the following result:

$$p\left(\hat{p}_1 - \hat{p}_2 - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} < p_1 - p_2 \right.$$
$$\left. < \hat{p}_1 - \hat{p}_2 + z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right) = 1-\alpha$$

is an approximate (1-α)100% confidence interval for $p_1 - p_2$.

Example: In a random samples of visitors to a famous tourist attraction, 84 of 250 men and 156 of 250 women bought souvenirs. Construct a 95% confidence interval for the difference between the true proportions of men and women who buy souvenirs at this tourist attraction.

Solution: $\hat{p}_1 = 0.34$ is the estimated proportion of men who buy souvenirs and $\hat{p}_2 = 0.62$ is the estimated proportion of men who buy souvenirs. For α=0.05, we find $z_{0.025}$=1.96 from z distribution table, and then,

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$
$$(0.34 - 0.62) \pm 1.96\sqrt{\frac{0.34(0.66)}{250} + \frac{0.62(0.38)}{250}}$$
$$(-0.28) \pm 0.084$$

We can be 95% confident that the interval from -0.364 to -0.196 cover the true difference between the true proportions of men and women who buy souvenirs at this tourist attraction. In other words, women visitors tend to buy souvenirs more than men visitors in tourist attraction.

The Estimation of the Ratio of Two Variances

If S_1^2 and S_2^2 are the variances of independent random samples of size n₁ and n₂ from normal populations, then, $F = \frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2}$ is a random variable having an f distribution with $(n_1 - 1)$ and $(n_2 - 1)$ degrees of freedom. Thus we can write,

$$P\left(f_{1-\alpha/2, n_1-1, n_2-1} < \frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} < f_{\alpha/2, n_1-1, n_2-1}\right) = 1-\alpha$$

Since $f_{1-\alpha/2,n_1-1,n_2-1} = \dfrac{1}{f_{\alpha/2,n_2-1,n_1-1}}$, it follows that

$$P\left(\frac{s_1^2}{s_2^2}\frac{1}{f_{\alpha/2,n_1-1,n_2-1}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2}\frac{1}{f_{1-\alpha/2,n_1-1,n_2-1}}\right) = 1-\alpha$$

is a (1-α)100% confidence interval for σ_1^2/σ_2^2.

Example: A study has been made to compare the nicotine contents of two brands of cigarettes. Ten cigarettes of Brand A had an average nicotine content of 3.1 milligrams with a standard deviation of 0.5 milligrams, while eight cigarettes of Brand B had an average nicotine content of 2.7 milligrams with standard deviation of 0.7 milligrams. Assuming that two sets of data are independent random samples from normal populations, construct 90% confidence interval for σ_1^2/σ_2^2.

Solution: Substituting n_1=10, n_2=8, s_1=0.5 and s_2=0.7 and $f_{0.05,9,7} = 3.68$, $f_{0.05,7,9} = 3.29$, we get

$$\frac{s_1^2}{s_2^2}\frac{1}{f_{\alpha/2,n_1-1,n_2-1}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2}\frac{1}{f_{1-\alpha/2,n_1-1,n_2-1}}$$

$$\frac{0.25}{0.49}\frac{1}{3.68} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{0.25}{0.49}(3.29)$$

$$0.14 < \frac{\sigma_1^2}{\sigma_2^2} < 1.68$$

The real ratio of two variances is included by the interval from 0.14 to 1.68 with 90% confident interval. Since the interval obtained here includes the possibility that ratio is 1, there is no real evidence against the assumption of equal population variances in this example.

INTERVAL ESTIMATION EXAMPLES

Example 1: Unoccupied seats on flights cause airlines to lose revenue (gelir, hasılat). Suppose a large airplane wants to estimate its average number of unoccupied seats per flight over the past year. To accomplish this, the records of 225 flights are randomly selected, and the number of unoccupied seats is noted for each of the sampled flights. Estimate μ, the mean number of unoccupied seats per flight during the past year, using given $\bar{x} = 11.6$, s=4.1 and a 90 % confidence interval.

Solution: The form of a large-sample 90% confidence interval for a population mean (based on the z-statistic) is:

$$P\left(\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

Substituting n=225, s=4.1 and and $z_{0.05} = 1.96$ into the formula, we get;

$$P\left(11.6 - 1.645\left(\frac{4.1}{\sqrt{225}}\right) < \mu < 11.6 + 1.645\left(\frac{4.1}{\sqrt{225}}\right)\right) = 0.90$$

Then the 90 % confidence interval is approximately $11.6 \pm 1.645\left(\frac{4.1}{\sqrt{225}}\right) = 11.6 \pm 0.45$ or from 11.15 to 12.05. That is, at the 90 % confidence level, we estimate the mean number of unoccupied seats per flight to be between 11.15 and 12.05 dring the sampled year.

Important: We stress that the confidence level for this example, 90%, refersto the procedure used. If we were apply that procedure repeatedly to different samples, qpproximately 90% of the intervals would contain μ. Although we do not know for sure whether this particular interval (11.15, 12.05) is one of the 90% that contain μ or one of the 10% that do not, our knowledge of probability gives us "confidence" that the interval contains μ.

Example 2: Consider the pharmaceutical company that desires an estimate of the mean increase in blood pressure of pateints who take a new drug. The blood pressure increases (measured in points) are measured for the n=6 patients in the human testing phase and the mean and the standard deviation of blood pressure increases for this sample of patients are determined to be $\bar{x} = 2.283$ and s=0.950. Use this information to construct a 95 % confidence interval for μ, the mean increase in blood pressure associated with the new drug for all patients in the population.

Solution: We do not get the normal distribution of $\bar{x}$ "automically" from the Central Limit Theorem when the sample size is small. Instead, we must assume that the measured variable, in this case the increase in blood pressure, is normally distributed in order for the distribution of $\bar{x}$ to be normal.

The confidence interval formula, we get

$$\bar{x} \pm t_{\alpha/2,(n-1)}\frac{s}{\sqrt{n}} = 2.283 \pm 2.571 \times \frac{0.950}{\sqrt{6}} = 2.283 \pm 0.997 \quad (t_{0.05/2,(6-1)} = t_{0.025,5} = 2.571)$$

or 1.286 to 3.280 points. We can be 95 % confident that the mean increase in blood pressure associated with taking this new drug between 1.286 and 3.28 points.

Example 3: Refer to the U.S. Army Corps of Engineers study of contaminated fish in the Tennessee River. The Corps of Engineers has collected data for a random sample of 144 fish contaminated with DDT. (The engineers made sure to capture contaminated fish in several different randomly selected streams and tributaries of the river.). The Army Corps of Engineers wants to estimate the true variation in fish weights in order to determine the true variation in fish weights in order to determine whether the fish are stable enough to allow further testing for DDT contamination. Use the sample data to find a 95 % confidence interval for the true variation in fish weights (σ^2).

Solution: We calculate the interval by using the the values $\chi^2_{0.025,143} \cong 185.800$ and $\chi^2_{0.975,143} \cong 117.985$ obtained from the chi-square table (by looking in the df=150 row of chi-squre table that it is the row with the df closest to 143).

$$P\left(\frac{(144-1)(376.5)^2}{185.500} < \sigma^2 < \frac{(144-1)(376.5)^2}{117.985} \right) = 0.95$$

$$P(109.275 < \sigma^2 < 171.806) = 0.95$$

Thus, the Army Corps of Engineers can be 95 % confident that the variance in weights of the population of contaminated fish ranges between 109.275 and 171.806.

Example 4: Public-opinion polls are conducted regularly to estimate the fraction of U.S. citizens who trust the president. Suppose 1000 people are randomly chosen and 637 answer that they trust the president. How would you estimate the true fraction of all U.S. citizens who trust the president? Construct a 95 % confidence interval for the true percentage of all U.S. citizens who trust the president.

Solution: Substituting n=1000, $\hat{p} = 0.637$ and $z_{0.025} = 1.96$ into the formula of

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.637 \pm 1.96\sqrt{\frac{0.637(0.363)}{1000}} = 0.637 \pm 0.030 = (0.607, 0.667)$$

Then we can be 95 % confident that the interval from 60.7 % to 66.7 % contains the true percentage of all U.S. citizens who trust the president. That is, in repeated constrructions of confidence intervals, approximately 95 % of all samples would produce confidence intervals that enclose p.

İST292 STATISTICS LESSON 6
HYPOTHESES TESTS

6. HYPOTHESES TESTS

If an engineer has to decide on the basis of sample data whether the true average lifetime of a certain kind of tire (tekerlek) is at least 42 000 miles, if an computer engineer has to decide on the basis of samples whether 95 percent of all virus attacks will be detected by a new virus scanner. These problems can all be referred as statistical hypothesis test. In the first case, the engineer has to test the hypothesis that θ, the parameter of an exponential population, is at least 42 000; in the second case we might say that the engineer has to decide whether p, the parameter of binomial population, equals 0.95. In each case it must be assumed that the chosen distribution correctly describes the experimental conditions; that is, the distribution provides the correct stastistical model.

Generally a hypothesis is defined as a proposition whose truth has not proved yet. Basically we can say that a **statistical hypotesis** is defined as a proposition about a population parameter.

A **statistical hypothesis** is an assertion (sav, iddia) of conjecture about the distribution of one or more random variables. If a statistical hypothesis completely specifies the distribution, it is refered to as a **simple hypothesis**; if not, it is refered to as a **composite hypothesis**.

To be able to construct suitable criteria for testing statistical hypotheses, it is necessary that we also formulate **alternative hypotheses**. For instance, in the example dealing with the lifetimes of the tires, we might formulate the alternative hypothesis that the parameter θ of exponential distribution is less than 42 000 miles; in the example of dealing with the ratio of detecting (removing) the virus attacks, we are testing the simple hypothesis p=0.95 against the simple alternative hypothesis p=0.80, where p is the parameter of binomial population for which n is given.

Symbolically, we shall use the symbol H_0, for the null hypothesis that we want to test and H_1, H_A for alternative hypothesis.

The testing of a statistical hypothesis is a decision procedure for determining whether sample data supports to the null hypothesis H_0, or not. The test procedure partitions the possible values of the test statistics into two subsets: an acceptance region for H_0 and a rejection region for H_0. The procedure can lead to two kinds of errors: for instance we want to test a null hypothesis H_0: $\theta=\theta_0$ against an alternative hypothesis H_1: $\theta=\theta_1$. If the true value of θ is θ_0, and the statistician (as a result of the test procedure) incorrectly concludes that $\theta=\theta_1$, he is committing an error referred to as a **type I error**. On the other hand, if the true value of θ is θ_1, and the statistician (as a result of form test procedure) incorrectly concludes that $\theta=\theta_0$, he is committing a second kind of error referred to as a **type II error**.

type I error $\alpha = P\left(H_0 \, is \, rejected \, \middle| H_0 \, is \, true\right)$ **and α is also called significance level of the test.**

type II error $\beta = P\left(H_0 \, is \, accepted \, \middle| H_0 \, is \; false\right)$

Suppose that the pharmaceutical manufacturer of a new medication wants to test the null hypothesis $\theta=0.90$ against the alternative hypothesis $\theta=0.60$. His test statistic is X, the observed number of successes (recoveries) in 20 trials (experimental units or patients), and he will accept the null hypothesis if x>14; otherwise, he will reject it. Find α and β.

In this example, the acceptance region for the null hypothesis is x=15,16,17,18,19 and 20, and, correspondingly, the rejecting region (or critical region) is x=0,1,2,…,14.

$$\alpha = P(X \le 14 | \theta = 0.90) = 0.0114$$

$$\beta = P(X > 14 | \theta = 0.60) = 0.1255$$

A good test procedure is one in which both α and β are small, thereby giving us a good chance of making the correct decision. The probability of the type II error in this example is rather high, but this can be reduced by appropriately changing the critical region. For instance, if we use the acceptance region x>15 in this example so that the critical region x≤15, it can easily be checked that this would make α=0.0433 and β=0.0509. As it is seen, as long as n is held fixed, if the probability of one type error is reduced, that of the other type error is increased. The only way in which we can reduce the probabilities of both types errors is to increase the size of the sample.

We want to test the null hypothesis H_0: $\theta=\theta_0$ against an alternative hypothesis H_1: $\theta\neq\theta_0$. Since it appears reasonable to accept the null hypothesis when our point estimate $\hat{\theta}$ of θ is close to θ_0 and to reject it when $\hat{\theta}$ is much larger or much smaller than θ_0. It would be logical to let the critical region consist of both tails of the sampling distribution of our test statistic. Such a test is referred to as a **two-tailed test**.

On the other hand, if we are testing the null hpothesis H_0: $\theta=\theta_0$ against an alternative hypothesis H_1: $\theta<\theta_0$. It would seem reasonable to reject H_0 only when $\hat{\theta}$ is much smaller than θ_0. Therefore, in this case it would be logical to let the critical region consist only of the left hand tail of the sampling distribution of our test statistic. This test is called as an **one(left)-sided hypothesis test**. Likewise, in testing H_0: $\theta=\theta_0$ against an alternative hypothesis H_1: $\theta>\theta_0$, we reject H_0 only for large values of $\hat{\theta}$, and the critical region consists only of the right tail of the sampling distribution of the test statistic. This test is called as an **one(rigt)-sided hypothesis test.**

For instance, for two sided alternative hypothesis H_1: $\mu\neq\mu_0$, the likelihood ratio technique led to a two-sided test with the critical region

$$|\bar{x}-\mu_0| \geq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

or

$$\bar{x} \leq \mu_0 - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \bar{x} \geq \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

In other words, the test statistic,

$$z=\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$$

where the critical region can be written as $z \leq -z_{\alpha/2}$ and $z \geq z_{\alpha/2}$ for an **α level of significance** corresponding to the amount of the critical region.

6.1 Test Concerning Mean

Suppose a random sample of size n from a normal population with mean μ and variance σ^2, $X_1, X_2, ..., X_n$, and thus $\bar{X}$ is a normal distributed random variable with μ and variance σ^2/n.

Assume that the hypotheses are concerned as follows:

$H_0 : \mu = \mu_0$	$H_0 : \mu = \mu_0 \, (\mu \geq \mu_0)$	$H_0 : \mu = \mu_0 \, (\mu \leq \mu_0)$
$H_1 : \mu \neq \mu_0$	$H_1 : \mu < \mu_0$	$H_1 : \mu > \mu_0$
two-sided test	one(left)-sided test	one(right)-sided test

6.1.1 When the Population Variance σ^2 is Known

For all hypotheses given above, the test statistic is:

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \sim N(0,1)$$

The distribution of the test statistic has standard normal distribution and so the boundaries of critical region for the two-sided test at the size of α :

If $z \leq -z_{\alpha/2}$ and $z \geq z_{\alpha/2}$ or $|z| \geq z_{\alpha/2}$, then the H₀ is rejected.

For the one(left)-sided test with the size of α :

If $z \leq -z_{\alpha}$, then the H₀ is rejected.

For the one(right)-sided test with the size of α :

If $z \geq z_{\alpha}$, then the H₀ is rejected.

Example 1: Suppose that it is known from experience that the standard deviation of the weight of 8- ounce packages of cookies made by a certain bakery is 0.16 ounce. To check whether its productions is under control on given a day, that is, to check whether the true average weight of the packages is 8 ounces, employees select a random sample of 25 packages and find that their mean weight is $\bar{x} = 8.091$ ounces. Since the bakery stands to lose money when μ>8 and the customer loses out when $\mu<8$, test the null hypothesis $\mu=8$ against the alternative hypothesis $\mu\neq8$ at 0.01 level of significance.

Solution:

1. H_0: $\mu=8$

   H_1: $\mu\neq8$

   $\alpha=0.01$
2. The test statistic values is $z=\dfrac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}=\dfrac{8.091-8}{0.16/\sqrt{25}}=2.84$
3. The boundaries of critical region at the size of $\alpha=0.01$ (or the critical values) are found as $-z_{0.005}=-2.575$ and $z_{\alpha/2}=2.575$ by using z (standard normal) distribution table.
4. When we compare the test value with the critical values, we reject the null hypothesis since $|z|=2.84$ exceeds 2.575 and suitable adjustments should be made in the production process at the level $\alpha=0.01$.

6.1.2 When the Population Variance σ^2 is Unknown

For all hypotheses given above, the test statistic is:

$$Z=\frac{\overline{X}-\mu_0}{S/\sqrt{n}}\sim N(0,1)$$

where S is a good estimate of σ for a sufficienly large sample size of n ($n\geq30$) and so the distribution of the test statistic has approximately standard normal distribution.

$H_0:\mu=\mu_0$
$H_1:\mu\neq\mu_0$
If $z\leq-z_{\alpha/2}$ and $z\geq z_{\alpha/2}$ or $|z|\geq z_{\alpha/2}$, then the H_0 is rejected at the size of α.

$H_0:\mu=\mu_0\,(\mu\geq\mu_0)$
$H_1:\mu<\mu_0$
If $z\leq-z_{\alpha}$ or $|z|\geq z_{\alpha}$ then the H_0 is rejected at the size of α.

$H_0:\mu=\mu_0\,(\mu\leq\mu_0)$
$H_1:\mu>\mu_0$
If $z\geq z_{\alpha}$, the H_0 is rejected at the size of α.

Example 2: Suppose that 100 tires made by a certain manufacturer lasted on the average 21819 miles with a standard deviation of 1295 miles. Test the null hypothesis $\mu=22000$ miles against the alternative hypothesis $\mu<22000$ miles at the 0.05 of significance.

Solution:

1. H$_0$: μ=22000

   H$_1$: μ<22000

   α=0.05

2. The test statistic values is $z = \dfrac{\overline{x}-\mu_0}{s/\sqrt{n}} = \dfrac{21819-22000}{1295/\sqrt{100}} = -1.40$

3. The boundary of critical region at the size of α=0.05 (or the critical values) are found as $-z_{0.05} = -1.645$ by using z (standard normal) distribution table.

4. When we compare the test value with the critical value, we say that the null hypothesis can not rejected since |z|=1.40 does not exceed 1.645 and so there is no real evidence that the tires are not as good as assumed under the null hypothesis at the level α=0.05.

NOT: If we have a sample from nonnormal population with unknown population variance, by virtue of the limit theorem, this result can be also used for random samples from nonnormal populations provided that n is sufficienly large; that is, n ≥ 30. In that case, we may also subsitute for σ the value of the sample standard deviation.

<u>But if the sample size n is not enough large (n<30),</u> $\dfrac{\overline{X}-\mu}{S/\sqrt{n}}$ is not approximately $N(0,1)$, but we know that $\dfrac{\overline{X}-\mu}{S/\sqrt{n}}$ has a student t distribution with (n-1) degrees of freedom.

In this case the test statistic has t distribution with (n-1) dergree of freedom:

$$T = \frac{\overline{X}-\mu_0}{S/\sqrt{n}} \sim t_{(n-1)}$$

$H_0: \mu = \mu_0$
$H_1: \mu \neq \mu_0$ If $t \leq -t_{\alpha/2,(n-1)}$ and $t \geq t_{\alpha/2,(n-1)}$ or $|t| \geq t_{\alpha/2,(n-1)}$, then the H$_0$ is rejected at the size of α.

$H_0: \mu = \mu_0\ (\mu \geq \mu_0)$
$H_1: \mu < \mu_0$ If $t \leq -t_{\alpha,(n-1)}$ or $|t| \geq t_{\alpha,(n-1)}$ then the H$_0$ is rejected at the size of α.

$H_0: \mu = \mu_0\ (\mu \leq \mu_0)$
$H_1: \mu > \mu_0$ If $t \geq t_{\alpha,(n-1)}$, the H$_0$ is rejected at the size of α.

Example 3: Suppose that a new drug is testing in a group of patients. The six test patients have blood pressure increase of 1.7, 3.0, 0.78, 3.4, 2.7 and 2.1 points. At level α=0.05 find whether there is evidence that the new drug satisfies the requirement that the resulting increase in blood pressure averages less than 3 points. $\bar{x} = 2.28, \quad s = 0.95$.

Solution:

1. H$_0$: μ=3

   H$_1$: μ<3

   α=0.05

2. The test statistic values is $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{2.28 - 3}{0.95/\sqrt{6}} = -1.86$

3. The boundary of critical region at the size of α=0.05 (or the critical values) are found as $-t_{0.05,5} = -2.015$ by using t (Student's t) distribution table.

4. When we compare the test value with the critical value, we say that the null hypothesis can not rejected since |t|=1.86 does not exceed 2.015 and so there is no real evidence that the mean increase in blood pressure resulting from taking the drug is less than 3 points at the level α=0.05.

6.2 Test Concerning Population Variance

Given a random sample of size n from a normal population, we can obtain (1-α)100% confidence interval for σ^2 by making use of the sampling distribution according to which

$$\frac{(n-1)\mathrm{S}^2}{\sigma^2}$$

a random variable having a chi-square distribution with *n-1* degrees of freedom. Thus, for these hypotheses tests,

$H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 \neq \sigma_0^2$	$H_0: \sigma^2 = \sigma_0^2 (\sigma^2 \geq \sigma_0^2)$ $H_1: \sigma^2 < \sigma_0^2$	$H_0: \sigma^2 = \sigma_0^2 (\sigma^2 \leq \sigma_0^2)$ $H_1: \sigma^2 > \sigma_0^2$
two-sided test	one(left)-sided test	one(right)-sided test

under the null hypothesis is correct, the test statistic is:

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \;.$$

$H_0 : \sigma^2 = \sigma_0^2$
$H_1 : \sigma^2 \neq \sigma_0^2$
If $\chi^2 \leq \chi^2_{1-\alpha/2,(n-1)}$ or $\chi^2 \geq \chi^2_{\alpha/2,(n-1)}$ then the H_0 is rejected at the size of α.

$H_0 : \sigma^2 = \sigma_0^2 (\sigma^2 \geq \sigma_0^2)$
$H_1 : \sigma^2 < \sigma_0^2$
If $\chi^2 \leq \chi^2_{1-\alpha,(n-1)}$ then the H_0 is rejected at the size of α.

$H_0 : \sigma^2 = \sigma_0^2 (\sigma^2 \leq \sigma_0^2)$
$H_1 : \sigma^2 > \sigma_0^2$
If $\chi^2 \geq \chi^2_{\alpha,(n-1)}$, the H_0 is rejected at the size of α.

Example 4: Suppose that the thickness of a part used in semiconductor is its critical dimension and that measurements of the thickness of a random sample of 18 such parts have the variance s^2=0.68, where the measurements are in thousandths (bininci) of an inch. The process is considered to be under control if the variation of thicknesses is given by a variance not greater than 0.36. Check whether the process is under control at the 0.05 level of significance.

Solution:

1. H_0: σ^2=0.36

   H_1: σ^2>0.36

   α=0.05

2. The test statistic values is $\chi^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(18-1)\times 0.68}{0.36} = 32.11$

3. The boundary of critical region at the size of α=0.05 (or the critical values) is found as $\chi^2_{0.05,17} = 27.587$ by using chi-squared distribution table.

4. When we compare the test value with the critical value, we say that the null hypothesis must be rejected since χ^2=32.11 exceeds 27.587 and the process used in manufacture of the parts must be adjusted at the level α=0.05.

6.3 Test Concerning Proportion

If an outcome of an experiment is the number of defect in particular assembly line, the number books in the library which are not checked in, the number of children who are not absent from school on a given day etc., we refer to such data as **count data**. Appropriate models for the analysis of count data are the binomial distribution, the Poisson distribution, the multinomial distribution and some other discrete distributions. The one of the most common tests based on the count data is a test concerning the parameter *p* (or θ) of the binomial distribution. In the test process for large values of n the normal approximation to the binomial distribution is used and is based on:

$$Z = \frac{X - np}{\sqrt{np(1-p)}} \sim N(0,1).$$

Where random variable X shows the number of successes obtained in n trials.

$H_0 : p = p_0$ $H_1 : p \neq p_0$	$H_0 : p = p_0\,(p \geq p_0)$ $H_1 : p < p_0$	$H_0 : p = p_0\,(p \leq p_0)$ $H_1 : p > p_0$
two-sided test	one(left)-sided test	one(right)-sided test

We can test the hypotheses above by using:

$$z = \frac{x - np_0}{\sqrt{np_0(1-p_0)}} \text{ or } z = \frac{\frac{x}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

$H_0 : p = p_0$
$H_1 : p \neq p_0$ If $z \leq -z_{\alpha/2}$ and $z \geq z_{\alpha/2}$ or $|z| \geq z_{\alpha/2}$, then the H_0 is rejected at the size of α.

$H_0 : p = p_0\,(p \geq p_0)$
$H_1 : p < p_0$ If $z \leq -z_{\alpha}$ or $|z| \geq z_{\alpha}$ then the H_0 is rejected at the size of α.

$H_0 : p = p_0\,(p \leq p_0)$
$H_1 : p > p_0$ If $z \geq z_{\alpha}$, the H_0 is rejected at the size of α.

Example 5: An oil company claims that less than 20 percent of all cars owners have not tried its gasoline. Test the claim at the 0.01 level of significance if a random check reveals that 22 of 200 car owners have not tried the oil company's gasoline.

Solution:

1. H_0: p=0.20

   H_1: p<0.20

   α=0.01

2. The test statistic values is $z = \frac{x - np_0}{\sqrt{np_0(1-p_0)}} = \frac{22 - 200(0.20)}{\sqrt{200(0.20)(0.80)}} = -3.18$

3. The boundary of critical region at the size of α=0.01 (or the critical values) are found as $-z_{0.01} = -2.33$ by using z (standard normal) distribution table.
4. When we compare the test value with the critical value, we say that the null hypothesis must be rejected since |z|=3.18 exceeds 2.33 and so we conclude that, as claimed, less than 20 percent of all car owners have not tried the oil company's gasoline at the level α=0.01.

6.4 Test Concering Comparison of Two Variances

If S_1^2 and S_2^2 are the variances of independent random samples of size n_1 and n_2 from normal populations, then, $F = \frac{S_1^2}{S_2^2}\frac{\sigma_2^2}{\sigma_1^2}$ is a random variable having an f distribution with $(n_1 - 1)$ and $(n_2 - 1)$ degrees of freedom. Therefore the hypotheses tests concerning the comparision of two variances are based on f distribution. The hypotheses are:

$H_0: \sigma_1^2 = \sigma_2^2 = \sigma^2$ $H_1: \sigma_1^2 \neq \sigma_2^2$	$H_0: \sigma_1^2 = \sigma_2^2 (\sigma_1^2 \geq \sigma_2^2)$ $H_1: \sigma_1^2 < \sigma_2^2$	$H_0: \sigma_1^2 = \sigma_2^2 (\sigma_1^2 \leq \sigma_2^2)$ $H_1: \sigma_1^2 > \sigma_2^2$
two-sided test	one(left)-sided test	one(right)-sided test

for two-sided the test statistic is:

if $s_1^2 \geq s_2^2$, if $f = \frac{s_1^2}{s_2^2} \geq f_{\alpha/2,(n_1-1),(n_2-1)}$ the null hypothesis is rejected at the size of α.

if $s_2^2 \geq s_1^2$, if $f = \frac{s_2^2}{s_1^2} \geq f_{\alpha/2,(n_2-1),(n_1-1)}$ the null hypothesis is rejected at the size of α.

for one(left)-sided the test statistic is:

if $f = \frac{s_1^2}{s_2^2} \geq f_{\alpha,(n_1-1),(n_2-1)}$ the null hypothesis is rejected at the size of α.

for one(left)-sided the test statistic is:

if $f = \frac{s_2^2}{s_1^2} \geq f_{\alpha,(n_2-1),(n_1-1)}$ the null hypothesis is rejected at the size of α.

6.5 Tests Concerning Differences between Means

In many problems in applied research, we are interested in hypotheses concerning differences between the means of two populations. For instance, we may want to decide upon the basis of suitable samples whether men can perform a certain task as fast as women, or we may want to decide on the basis of an appropriate sample survey whether the average daily time spent on the internet in science faculty's students exceeds those of art faculty's students by at least 2 hours.

$H_0 : \mu_1 - \mu_2 = \delta$ $H_1 : \mu_1 - \mu_2 \neq \delta$	$H_0 : \mu_1 - \mu_2 = \delta \, (\mu_1 - \mu_2 \geq \delta)$ $H_1 : \mu_1 - \mu_2 < \delta$	$H_0 : \mu_1 - \mu_2 = \delta \, (\mu_1 - \mu_2 \leq \delta)$ $H_1 : \mu_1 - \mu_2 > \delta$
two-sided test	one(left)-sided test	one(right)-sided test

When σ_1^2 and σ_2^2 are known, the test statistic

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - \delta}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}},$$

has the standard normal distribution.

$H_0 : \mu_1 - \mu_2 = \delta$
$H_1 : \mu_1 - \mu_2 \neq \delta$
If $z \leq -z_{\alpha/2}$ and $z \geq z_{\alpha/2}$ or $|z| \geq z_{\alpha/2}$, then the H0 is rejected at the size of α.

$H_0 : \mu_1 - \mu_2 = \delta \, (\mu_1 - \mu_2 \geq \delta)$
$H_1 : \mu_1 - \mu_2 < \delta$
If $z \leq -z_\alpha$ or $|z| \geq z_\alpha$ then the H0 is rejected at the size of α.

$H_0 : \mu_1 - \mu_2 = \delta \, (\mu_1 - \mu_2 \leq \delta)$
$H_1 : \mu_1 - \mu_2 > \delta$
If $z \geq z_\alpha$, the H0 is rejected at the size of α.

When σ_1^2 and σ_2^2 are unknown, if n1≥30 and n2≥30,

The test statistic is still used by subsituted the sample variances S_1^2 and S_2^2 for the estimators of σ_1^2 and σ_2^2:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - \delta}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}},$$

has the approximately standard normal distribution. Thus the hypotheses are tested by using the rules:

$H_0: \mu_1 - \mu_2 = \delta$
$H_1: \mu_1 - \mu_2 \neq \delta$ If $z \le -z_{\alpha/2}$ and $z \ge z_{\alpha/2}$ or $|z| \ge z_{\alpha/2}$, then the H$_0$ is rejected at the size of α.

$H_0: \mu_1 - \mu_2 = \delta \, (\mu_1 - \mu_2 \ge \delta)$
$H_1: \mu_1 - \mu_2 < \delta$ If $z \le -z_\alpha$ or $|z| \ge z_\alpha$ then the H$_0$ is rejected at the size of α.

$H_0: \mu_1 - \mu_2 = \delta \, (\mu_1 - \mu_2 \le \delta)$
$H_1: \mu_1 - \mu_2 > \delta$ If $z \ge z_\alpha$, the H$_0$ is rejected at the size of α.

Example 6: Lesley E. Tan investigated the relationship between handedness (tek elini kullanabilme) and motor competence (kabiliyet) in preschool children. Random samples of 41 right-handers and 41 left-handers were administered several tests of motor skills, yielding the means and standard deviations shown in the accompanying table. Is there evidence of a difference between the average motor skill scores of left- and right- handed preschoolers base on this experiment? Use α=0.10 (source: Tan, L. E. "Laterality and motor skills in four –years-olds", Child Development, 56.)

Left-handed	Right-handed
$n_1 = 41$ $\bar{x}_1 = 97.5$ $s_1 = 17.5$	$n_2 = 41$ $\bar{x}_2 = 98.1$ $s_2 = 19.2$

Solution:

1. H$_0$: μ$_1$-μ$_2$=0

   H$_1$: μ$_1$-μ$_2$≠0

   α=0.10

2. The test statistic value's is $z = \frac{(\bar{x}_1 - \bar{x}_2) - \delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(97.5 - 98.1) - 0}{\sqrt{\frac{17.5^2}{41} + \frac{19.2^2}{41}}} = \frac{-0.6}{\sqrt{\frac{674.89}{41}}} = -0.14788$

3. The boundary of critical region at the size of α=0.10 (or the critical value) is found as $-z_{0.05} = -1.645$ by using z (standard normal) distribution table.

   When we compare the test value with the critical value, we say that the null hypothesis cannot be rejected since |z|=0.14788 does not exceed 1.645 and so we conclude that, there is no evidence of a difference between the average motor skill scores of left- and right- handed preschoolers at the level α=0.10.

By virtue of the central limit theorem, these two test statistics above are also used for independent random samples from non-normal populations when n$_1$ and n$_2$ are large, that is n$_1$≥30 and n$_2$≥30. If population variances are not known, we may also substitute for σ$_1$ and σ$_2$ the values of standard deviations s$_1$ and s$_2$ obtained from samples.

When σ_1^2 and σ_2^2 are unknown, but n$_1$<30 and n$_2$<30

Assuming that $\sigma_1^2 = \sigma_2^2 = \sigma^2$ then the test statistic is given by :

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Where the test statistic has t distribution with degrees of freedom $(n_1 + n_2 - 2)$.

In this test first we need test the $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$ against $H_1 : \sigma_1^2 \neq \sigma_2^2$ if we conclude that H$_0$ cannot be rejected, we compare the means by using these rules as follows:

$H_0 : \mu_1 - \mu_2 = \delta$
$H_1 : \mu_1 - \mu_2 \neq \delta$
If $t \leq -t_{\alpha/2,(n_1+n_2-2)}$ and $t \geq t_{\alpha/2,(n_1+n_2-2)}$ or $|t| \geq t_{\alpha/2,(n_1+n_2-2)}$, then the H$_0$ is rejected at the size of α.

$H_0 : \mu_1 - \mu_2 = \delta \, (\mu_1 - \mu_2 \geq \delta)$
$H_1 : \mu_1 - \mu_2 < \delta$
If $t \leq -t_{\alpha,(n_1+n_2-2)}$ or $|t| \geq t_{\alpha,(n_1+n_2-2)}$ then the H$_0$ is rejected at the size of α.

$H_0 : \mu_1 - \mu_2 = \delta \, (\mu_1 - \mu_2 \leq \delta)$
$H_1 : \mu_1 - \mu_2 > \delta$
If $t \geq t_{\alpha,(n_1+n_2-2)}$, then the H$_0$ is rejected at the size of α.

Example 7: Suppose a television network want to determine whether major sports events or first-run movies attract more viewers in the prime time hours. It selects 28 prime-time evenings; of these, 13 have programs devoted to major sports events and remaining 15 have first-run movies. The number of viewers (estimated by a television viewer rating firm) is recorded for each program. The television network's samples produced the following results:

Sports (n_1=13)	Movies (n_2=15)
$\bar{x}_1 = 6.8\,million$ $s_1 = 1.8\,million$	$\bar{x}_2 = 5.3\,million$ $s_2 = 1.6\,million$

Is there evidence the difference between the mean numbers of viewers for major sport events and first-time movies, at level 0.10 under the samples from the two normal populations.

Solution:

1. $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$
   $H_1 : \sigma_1^2 \neq \sigma_2^2$

   α=0.10,

2. The test statistic's value is $f = \frac{s_1^2}{s_2^2} = \frac{1.8^2}{1.6^2} = 1.2656$

3. The boundary of critical region at the size of α=0.10 (or the critical value) is found as $f_{0.05,12,14} = 2.53$ by using f distribution table.

4. When we compare the test value with the critical value, we say that the null hypothesis cannot be rejected since f=1.2656 does not exceed 2.53 and so we conclude that, $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

5. $H_0 : \mu_1 - \mu_2 = 0$
   $H_1 : \mu_1 - \mu_2 \neq 0$

6. The test statistic's value is calculated by

$$s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} = \frac{12(1.8)^2+14(1.6)^2}{13+15-2} = 2.87$$

$$t = \frac{(\bar{x}_1-\bar{x}_2)-0}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{1.5}{\sqrt{2.87\left(\frac{1}{13}+\frac{1}{15}\right)}} = \frac{1.5}{0.6419} = 2.3368$$

7. we find $t_{0.05,26}$=1.706 from Student's t distribution table, then we compare the test value with the critical value, we say that the null hypothesis is rejected since |t|=2.3368 exceeds 1.706 and we conclude that the mean numbers of viewers of sport events and is larger than the mean numbers of viewers of first-time movies at the significance level 0.10.

If we conclude $\sigma_1^2 \neq \sigma_2^2$, The test statistic

$$T = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$$

is a random variable having t distribution with ν degrees of freedom as defined below:

$$\nu \approx \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^2}{n_1}\right)^2\left(\frac{1}{n_1-1}\right)+\left(\frac{s_2^2}{n_2}\right)^2\left(\frac{1}{n_2-1}\right)}.$$

we compare the means by using these rules as follows:

$H_0:\mu_1-\mu_2=\delta$
$H_1:\mu_1-\mu_2\neq\delta$
If $t \le -t_{\alpha/2,\nu}$ and $t \ge t_{\alpha/2,\nu}$ or $|t| \ge t_{\alpha/2,\nu}$, then the H_0 is rejected at the size of α.

$H_0:\mu_1-\mu_2=\delta\,(\mu_1-\mu_2\ge\delta)$
$H_1:\mu_1-\mu_2<\delta$
If $t \le -t_{\alpha,\nu}$ or $|t| \ge t_{\alpha,\nu}$ then the H_0 is rejected at the size of α.

$H_0:\mu_1-\mu_2=\delta\,(\mu_1-\mu_2\le\delta)$
$H_1:\mu_1-\mu_2>\delta$
If $t \ge t_{\alpha,\nu}$, the H_0 is rejected at the size of α.

6.6 Test Concerning Differences between Means: Paired Samples

In some studies or researches, paired measurements of a variable for each unit, person, event or observation are obtained such as before/after, first/second, etc. In general these measurements cannot be assumed independent, because these units (persons, events or observations) are same and thus these leads to a dependence between the measurements.

Let $D_i = X_{1i} - X_{2i}$ (i=1,2,…,n) be a difference between the pairs of measurements. Di is normally distributed random variables with mean $\mu_1 - \mu_2$ and variance σ_D^2.

$D_i = X_{1i} - X_{2i} \sim N(\mu_1 - \mu_2, \sigma_D^2)$ i=1,2,…,n

The hypotheses corresponding to the comparing the means are:

$H_0: \mu_1 - \mu_2 = \delta$ $H_1: \mu_1 - \mu_2 \neq \delta$	$H_0: \mu_1 - \mu_2 = \delta\,(\mu_1 - \mu_2 \geq \delta)$ $H_1: \mu_1 - \mu_2 < \delta$	$H_0: \mu_1 - \mu_2 = \delta\,(\mu_1 - \mu_2 \leq \delta)$ $H_1: \mu_1 - \mu_2 > \delta$
two-sided test	one(left)-sided test	one(rigt)-sided test

If σ_D^2 ***is <u>known</u>***, the test statistic is based on the z distribution: $Z = \dfrac{(\bar{X}_1 - \bar{X}_2) - \delta}{\sigma_D / \sqrt{n}}$

If σ_D^2 **is <u>unknown</u> and also <u>n≥30</u>**, the test statistic is based on the approximately z distribution: $Z = \dfrac{(\bar{X}_1 - \bar{X}_2) - \delta}{S_D / \sqrt{n}}$

If σ_D^2 **is <u>unknown</u> and also <u>n<30</u>**, the test statistic is based on the t distribution with n-1 degrees of freedom: $T = \dfrac{(\bar{X}_1 - \bar{X}_2) - \delta}{S_D / \sqrt{n}}$

where $\bar{D} = \dfrac{\sum_{i=1}^{n}(X_{1i} - X_{2i})}{n}$ and $S_D^2 = \dfrac{\sum_{i=1}^{n}(D_i - \bar{D})^2}{n-1}$ are estimators of μ_1-μ_2 and σ_D^2 respectively

if the values of $d_i = x_{1i} - x_{2i}$ (i=1,2,…,n) are given, $\bar{d} = \dfrac{\sum_{i=1}^{n}(x_{1i} - x_{2i})}{n}$ and $s_D^2 = \dfrac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}$ would be their estimations respectively. Here the test procedures are same as given before.

Example 8: The data shown in the table provide information on the average weekly losses of work hours due to accidents in 10 industrial plants before and after a certain safety program was put into operation. At the 0.05 level significance, test whether the safety program is effective.

before	45	73	46	124	33	57	83	34	26	17
after	36	60	44	119	35	51	77	29	24	11
d_i	9	13	2	5	-2	6	6	5	2	6

$$\sum d_i = 52 \qquad s_D^2 = 16.6222$$

Solution:

1. H_0: μ_1-μ_2=0

   H_1: μ_1-μ_2≠0

   α=0.05

2. The test statistic's value is $t = \dfrac{\bar{d} - \delta}{s_D / \sqrt{n}} = \dfrac{5.2 - 0}{4.077 / \sqrt{10}} = 4.033$

3. The boundary of critical region at the size of α=0.05 (or the critical value) is found as $t_{0.025,9} = 2.262$ by using t distribution table.

4. When we compare the test value with the critical value, we say that the null hypothesis can be rejected since |t|=4.033 exceed 2.262 and so we conclude that, there is evidence of the effectiveness of the safety program on the average weekly losses of work hours at the level α=0.05.

6.7 Test Concerning Differences between Proportions

In many problems in applied research, we must decide whether observed differences of two or more sample proportions, or percentages, are significant or whether they can be attributed to chance. For solving these problems, we use approximation of binomial distribution to normal distribution by virtue of Central Limit Theorem.

Let X_1 be a random variable from binomial distribution with n_1 and p_1 and X_2 be an independent random variable from binomial distribution with n_2 and p_2.

If the respective number of successes are X_1 and X_2 and the corresponding sample proportions are denoted by $\hat{p}_1 = \frac{X_1}{n_1}$ and $\hat{p}_2 = \frac{X_2}{n_2}$, let us investigate the sampling distribution of $\hat{p}_1 - \hat{p}_2$, which is an obvious estimator of $p_1 - p_2$. We have

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$$

$$V(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

and for large samples, X_1, X_2 and hence also their difference, can be approximated with normal distributions, it follows that

$$Z = \frac{\left(\frac{X_1}{n_1} - \frac{X_2}{n_2}\right) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

is a random variable having approximately the standard normal distribution.

The hypotheses corresponding to the difference of proportions are:

$H_0 : p_1 - p_2 = 0$ $H_1 : p_1 - p_2 \neq 0$	$H_0 : p_1 - p_2 = 0 (p_1 - p_2 \geq 0)$ $H_1 : p_1 - p_2 < 0$	$H_0 : p_1 - p_2 = 0 (p_1 - p_2 \leq 0)$ $H_1 : p_1 - p_2 > 0$
two-sided test	one(left)-sided test	one(right)-sided test

When we are interested only in the null hypothesis $p_1 = p_2$, we substitute for p the pooled estimate $\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$ and the formula is used to test:

$$z = \frac{\frac{x_1}{n_1} - \frac{x_2}{n_2}}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$H_0 : p_1 - p_2 = 0$
$H_1 : p_1 - p_2 \neq 0$

If $z \leq -z_{\alpha/2}$ and $z \geq z_{\alpha/2}$ or $|z| \geq z_{\alpha/2}$, then the H₀ is rejected at the size of α.

$$\begin{aligned} &H_0 : p_1 - p_2 = 0 (p_1 - p_2 \geq 0) \\ &H_1 : p_1 - p_2 < 0 \end{aligned}$$ If $z \leq -z_\alpha$ or $|z| \geq z_\alpha$ then the H0 is rejected at the size of α.

$$\begin{aligned} &H_0 : p_1 - p_2 = 0 (p_1 - p_2 \leq 0) \\ &H_1 : p_1 - p_2 > 0 \end{aligned}$$ If $z \geq z_\alpha$, the H0 is rejected at the size of α.

Example 9: In random samples of 250 person with low incomes, 200 person with avergae incomes, and 150 person with high incomes, there were, respectively, 155, 118 and 87 who favor a certain piece of legislation (yasa). Use the 0.05 level of significance to test the null hypothesis $p_{low} = p_{high}$ against the alternative hypothesis $\theta_{low} \neq \theta_{high}$.

Solution:

1. H0: $p_{low} = p_{high}$ or ($p_{low} - p_{high} = 0$)

   H1: $p_{low} \neq p_{high}$ 0 or ($p_{low} - p_{high} \neq 0$)

   α=0.05
2. The test statistic value’s is:

   $$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{155 + 87}{250 + 150} = \frac{242}{400} = 0.605$$

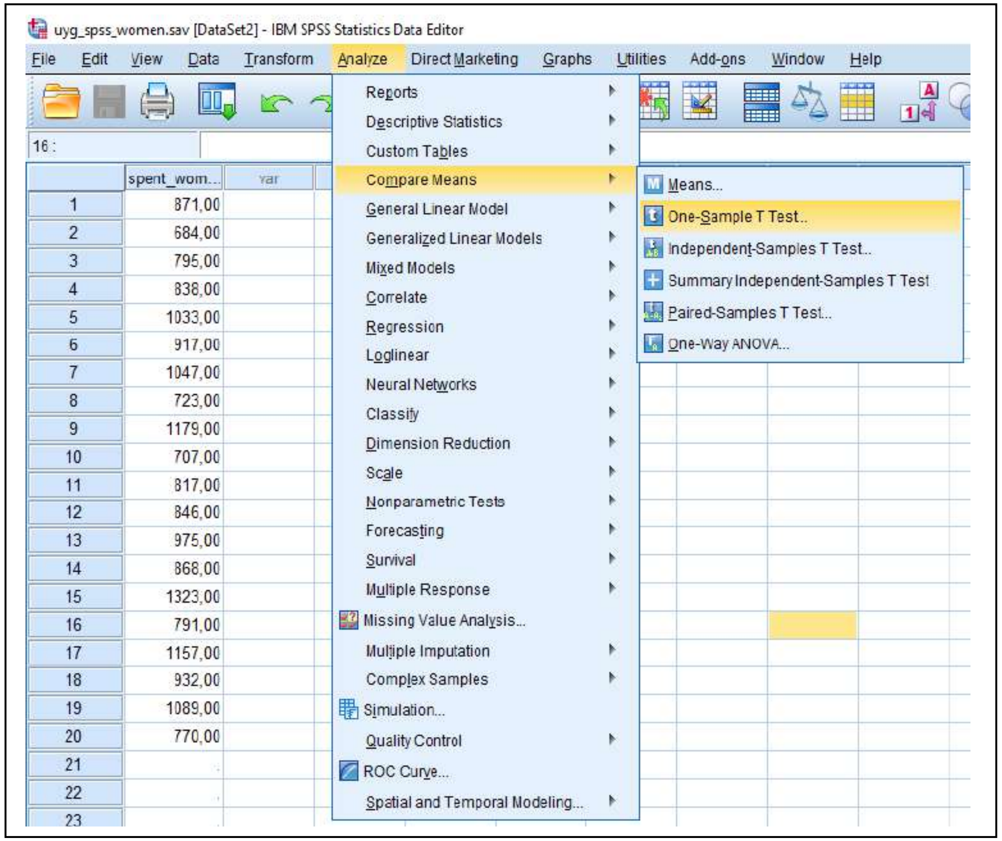
   $$z = \frac{\frac{x_1}{n_1} - \frac{x_2}{n_2}}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(0.62 - 0.58)}{\sqrt{0.605 \times 0.395\left(\frac{1}{250} + \frac{1}{150}\right)}} = \frac{0.04}{0.0504876} = 0.7923$$

3. The boundary of critical region at the size of α=0.05 (or the critical value) is found as $z_{0.025} = 1.96$ by using z (standard normal) distribution table.

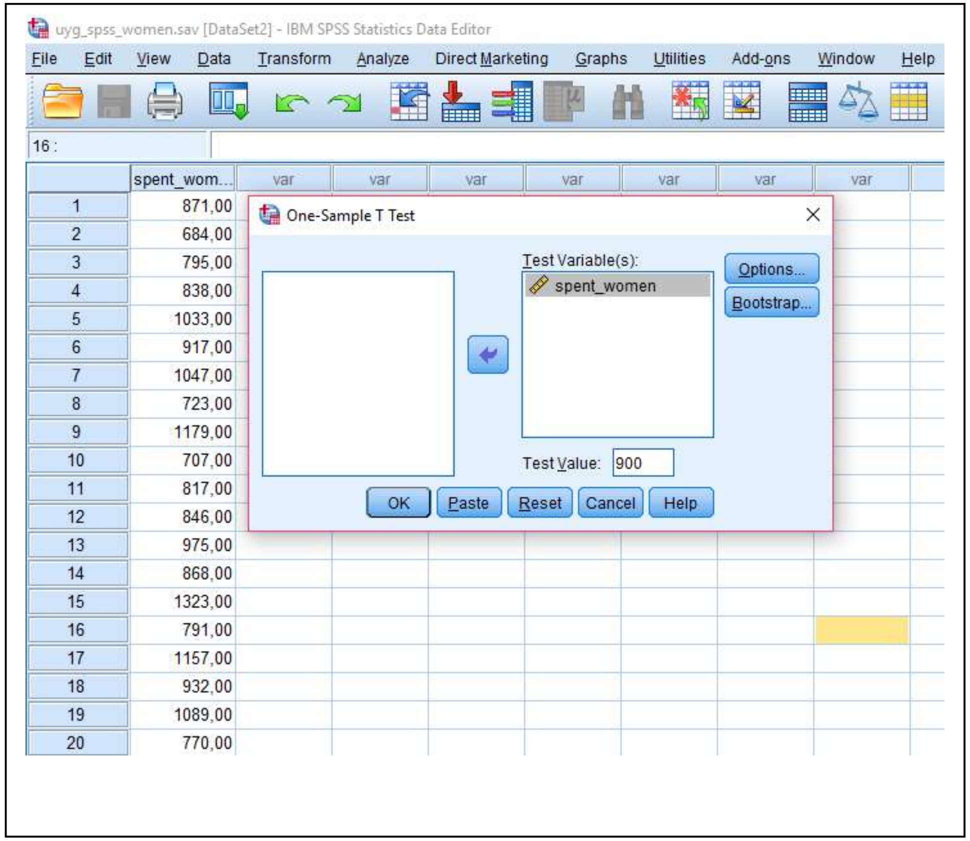
4. When we compare the test value with the critical value, we say that the null hypothesis cannot be rejected since |z|=0.4979 does not exceed 1.96 and so we conclude that, there is no evidence of a difference between the proportion of persons favoring the legislation is the same for two groups with lower and higher incomes at the level α=0.05.

Example 1: Folowing table shows the amount of money paid by women for reparing of their cars. Conduct a test of hypothesis whether if the mean of spent money by women for reparing of their cars equals to 900 or not using α=0.05.

women	871	684	795	838	1033	917	1047	723	1179	707	817	846	975	868	1323	791	1157	932	1089	770



	spent_wom...	var
1	871,00	
2	684,00	
3	795,00	
4	838,00	
5	1033,00	
6	917,00	
7	1047,00	
8	723,00	
9	1179,00	
10	707,00	
11	817,00	
12	846,00	
13	975,00	
14	868,00	
15	1323,00	
16	791,00	
17	1157,00	
18	932,00	
19	1089,00	
20	770,00	
21		
22		
23		


uyg_spss_women.sav [DataSet2] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
16 :
One-Sample T Test
Options...
Bootstrap...
OK
Paste
Reset
Cancel
Help

Test Variable(s): spent_women

Test Value: 900

	spent_wom...	var	var	var	var	var	var	var
1	871,00							
2	684,00							
3	795,00							
4	838,00							
5	1033,00							
6	917,00							
7	1047,00							
8	723,00							
9	1179,00							
10	707,00							
11	817,00							
12	846,00							
13	975,00							
14	868,00							
15	1323,00							
16	791,00							
17	1157,00							
18	932,00							
19	1089,00							
20	770,00							

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
spent_women	20	918,1000	173,01929	38,68829

One-Sample Test

	Test Value = 900					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
spent_women	,468	19	,645	18,10000	-62,8755	99,0755

H_0: $\mu = 900$

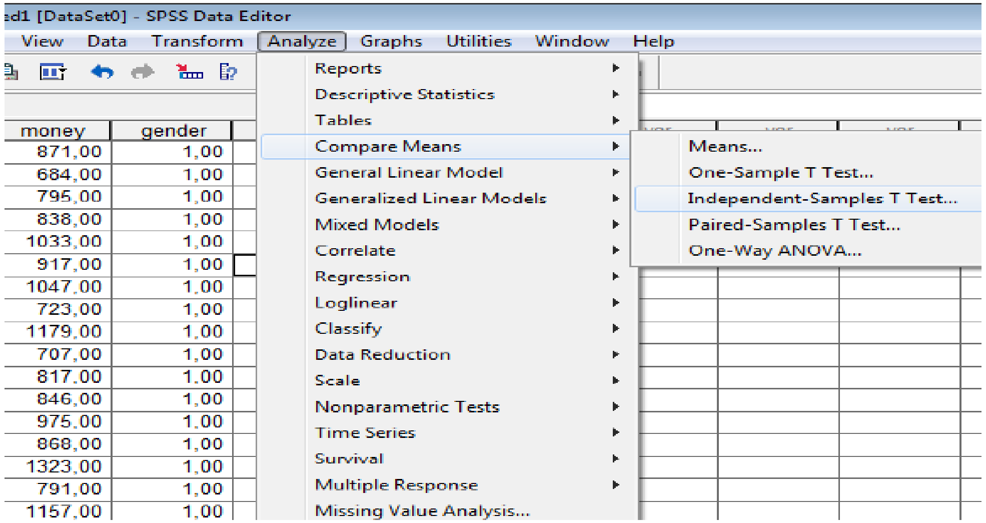
H_1: $\mu \neq 900$

σ^2 is not known and n<30 so for the test, test statistic is $t = \dfrac{\bar{X} - 900}{S/\sqrt{n}} = 0.468$ and critical value using t distribution with 19 degrees of freedom, $t_{0.025,19} = 2.093$. When comparing the test value and critical value at level 0.05, $0.468 < 2.093$ is found and we say that H_0 cannot be rejected. In addition this decision can be taken by using the p value(sig. (2 tailed)) given in the table. Since $p\,value = 0.645 > \alpha = 0.05$, H_0 cannot be rejected at the significance of level α=0.05.

Example 2: Folowing table shows the amount of money paid by women and men for reparing of their cars. Conduct a test of hypothesis comparing the two means of spent money for women and men's car repairs using α=0.05.

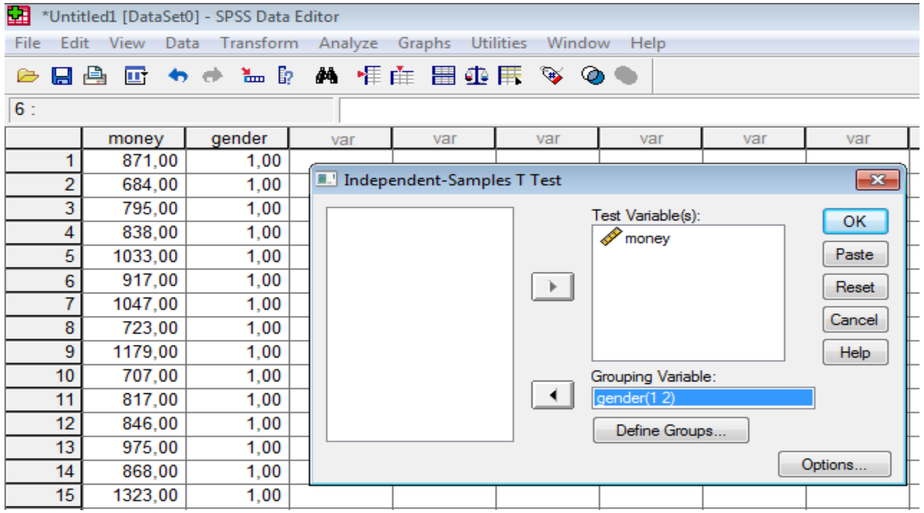
women	871	684	795	838	1033	917	1047	723	1179	707	817	846	975	868	1323	791	1157	932	1089	770
men	792	765	511	520	618	447	548	720	899	788	927	657	851	702	918	528	884	702	839	878

1. Step



money	gender
871,00	1,00
684,00	1,00
795,00	1,00
838,00	1,00
1033,00	1,00
917,00	1,00
1047,00	1,00
723,00	1,00
1179,00	1,00
707,00	1,00
817,00	1,00
846,00	1,00
975,00	1,00
868,00	1,00
1323,00	1,00
791,00	1,00
1157,00	1,00

2. step


*Untitled1 [DataSet0] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
6 :
Independent-Samples T Test
OK
Paste
Reset
Cancel
Help
Define Groups...
Options...

	money	gender	var	var	var	var	var	var
1	871,00	1,00						
2	684,00	1,00						
3	795,00	1,00						
4	838,00	1,00						
5	1033,00	1,00						
6	917,00	1,00						
7	1047,00	1,00						
8	723,00	1,00						
9	1179,00	1,00						
10	707,00	1,00						
11	817,00	1,00						
12	846,00	1,00						
13	975,00	1,00						
14	868,00	1,00						
15	1323,00	1,00						

Test Variable(s): money

Grouping Variable: gender(1 2)

1) H₀: $\mu_1 - \mu_2 = 0$

H₁: $\mu_1 - \mu_2 \neq 0$

For the these hypotheses test we need to decide whether $\sigma_1^2 = \sigma_2^2$ or not.

2) H₀: $\sigma_1^2 = \sigma_2^2$

H₁: $\sigma_1^2 \neq \sigma_2^2$

From the output

$f=\frac{S_1^2}{S_2^2}=1.2686$ critical value by using f distribution $f_{0.025,(df1=19,df2=19)}=2.53$ and comparing the test value and critical value at level 0.05, $1.2686<2.53$ is found and we say that H$_0$ cannot be rejected at the significance of level α=0.05.

NOTE: Moreover, we can also test the hypothesis given in 2) by using "Levene's Test for Equality of Variances" at the table entitled "Independence Samples Test". For the Levene's Test the significance value $p=0.732>\alpha=0.05$ as a result H$_0$: $\sigma_1^2=\sigma_2^2$ hypothesis cannot be rejected at the significance of level α=0.05.

For the hypothesis given 1) we use first line of the table entitled "Independence Samples Test". Here t test statistic's value $t=3.738$ and $p=0.001<0.05$ H$_0$ is rejected at the significance of level α=0.05. We can say that the amount of money paid by women and men for reparing of their cars are not same. Women pay more money then men for reparing the their car.

Group Statistics

	gender	N	Mean	Std. Deviation	Std. Error Mean
money	women	20	918,1000	173,01929	38,68829
	men	20	724,7000	153,61370	34,34907

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
money	Equal variances assumed	,119	,732	3,738	38	,001	193,40000	51,73627	88,66539	298,13461
	Equal variances not assumed			3,738	37,475	,001	193,40000	51,73627	88,61715	298,18285

Example 3: In the research of a psychology department, to compare two methods of solving problem in group, two problems sets 10 groups of each having 4 persons: one was solved by using face to face method; other was solved by using teleconference method. Groups' scores were recorded.

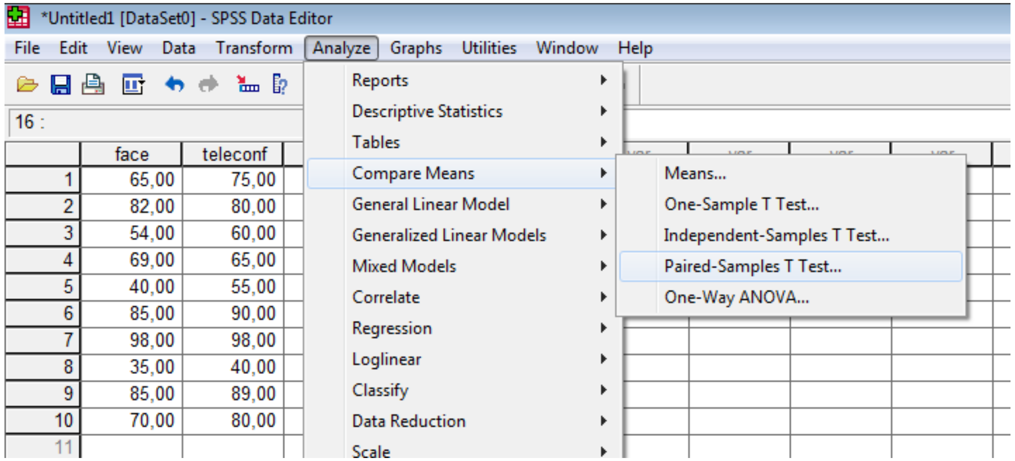
a) Conduct a test of hypothesis comparing the two methods efficiencies using α=0.05.

b) Use a 95 % confidence interval to estimate the difference between the mean of test scores for two methods.

c) Compare parts a) and b).

Face to face	65	82	54	69	40	85	98	35	85	70
teleconference	75	80	60	65	55	90	98	40	89	80

1. Step

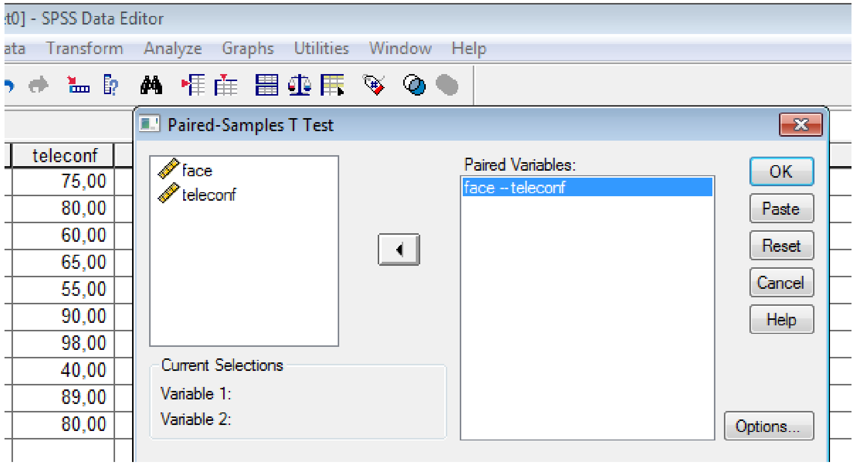


*Untitled1 [DataSet0] - SPSS Data Editor

16 :

	face	teleconf
1	65,00	75,00
2	82,00	80,00
3	54,00	60,00
4	69,00	65,00
5	40,00	55,00
6	85,00	90,00
7	98,00	98,00
8	35,00	40,00
9	85,00	89,00
10	70,00	80,00
11		

2. step



t0] - SPSS Data Editor

Paired-Samples T Test

teleconf
75,00
80,00
60,00
65,00
55,00
90,00
98,00
40,00
89,00
80,00

Paired Variables: face -- teleconf

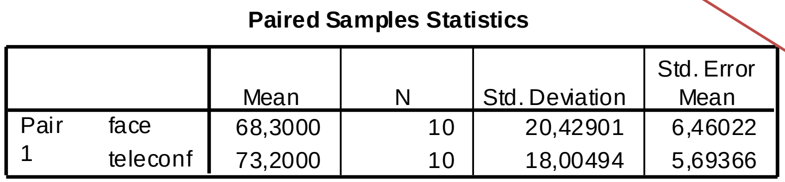
Current Selections
Variable 1:
Variable 2:

H₀: $\mu_1 - \mu_2 = 0$

H₁: $\mu_1 - \mu_2 \neq 0$

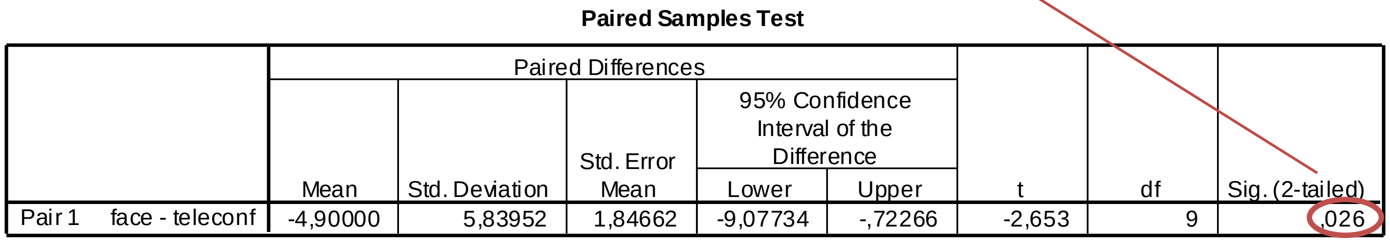
The samples are dependent since two problem sets were asked same 10 groups of each having 4 persons.

σ_D^2 is not known and n<30 so for the test, test statistic is $t=\frac{\bar{d}-0}{S_d/\sqrt{n}}=-2.653$ and critical value using t distribution with 9 degrees of freedom, $t_{0.025,9}=2.262$. When comparing the test value and critical value at level 0.05, $\left|-2.653\right|\geq 2.262$ is found and we say that H₀ is rejected. In addition this decision can be taken by using the p value(sig. (2 tailed)) given in the table. Since $p\,value=0.026<\alpha=0.05$ H₀ is rejected at the significance of level α=0.05.



Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	face	68,3000	10	20,42901	6,46022
	teleconf	73,2000	10	18,00494	5,69366



Paired Samples Test

		Paired Differences							
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	face - teleconf	-4,90000	5,83952	1,84662	-9,07734	-,72266	-2,653	9	,026

NOTE: Since there is no test using z statistic in SPSS, these tests related to z statistic are also done with respect to t statistic. Because the distribution of t statistic approximates to standard normal distribution for large sample sizes.

NOTE: In all of these examples alternative hypotheses are two sided tests since the p-value (Sig.2-tailed) are given in the tables. So that we directly compare the p values with significance level α. However, if we are conducting one-sided test, taking half of the p-value (p-value /2) is compared with significance level α and the decision about the null hypothesis H₀ is taken.

İST292 STATISTICS LESSON 7

REGRESSION ANALYSIS

Many engineering and scientific problems are concerned with determining a relationship between a set of variables. ***Regression analysis***, is a statistical technique that is very useful for these types of problems. *For example,* in a chemical process, suppose that the yield of the product is related to the process-operating temperature. Regression analysis, can be used to build a model to predict yield at a given temperature level. This model can also be used for process optimization, such as finding the level of temperature that maximizes yield, or for process control purposes.

As an illustration, consider data in Table 1. In this table;
y is the purity of oxygen produced in a chemical distillation (damıtma) process,
x is the percentage of hydrocarbons that are present in the main condenser of the distillation unit.

Table 1. Oxygen and Hydrocarbon Levels

Observation Number	**Hydrocarbon Level (x)**	**Purity y(%)**
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

Figure 1 presents a ***scatter diagram***. Inspection of this scatter diagram indicates that, although no simple curve will pass exactly through all the points, there is a strong indication that the points lie scattered randomly around a straight line. Therefore, it is probably reasonable to assume that the mean of the random variable Y is related to x by the following straight-line relationship:

$$E(Y|x) = \mu_{Y|x} = \beta_0 + \beta_1 x$$

,where **the slope (β_0) and intercept (β_1) of the line are called *regression coefficients*.**

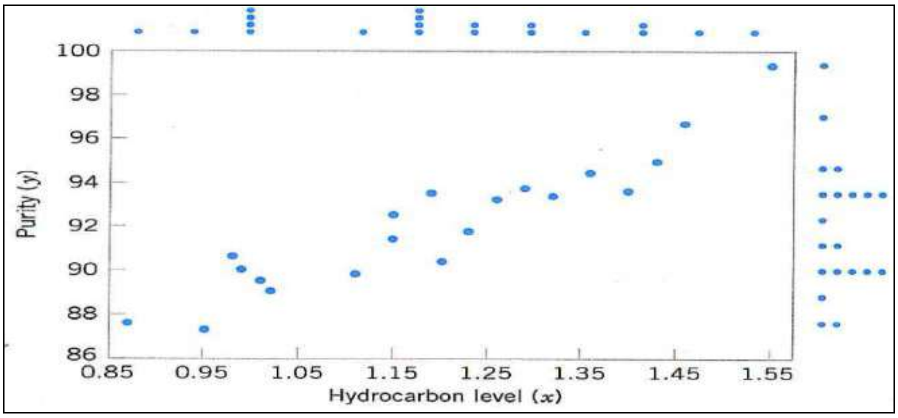


Figure 1. Scatter diagram of oxygen purity versus hydrocarbon level from Table 1.

8.1. Simple Linear Regression

The case of ***simple linear regression*** considers a single ***regressor*** or ***predictor*** x and a dependent or ***response variable*** Y. Suppose that the true relationship between Y and x is a straight line and that the observation Y at each level of x is a random variable. As noted previously, the expected value of Y for each value of *x* is

$$E(Y|x) = \mu_{Y|x} = \beta_0 + \beta_1 x$$

where the **intercept** β_0 **and the slope** β_1 **are unknown regression coefficients.** We assume that each observation Y, can be described by the model

$$Y = \beta_0 + \beta_1 x + \in \qquad (1)$$

Where $\in$ **is a random error with mean zero and (unknown) variance** σ^2**.** The random errors corresponding to different observations are also assumed to be uncorrelated random variables.

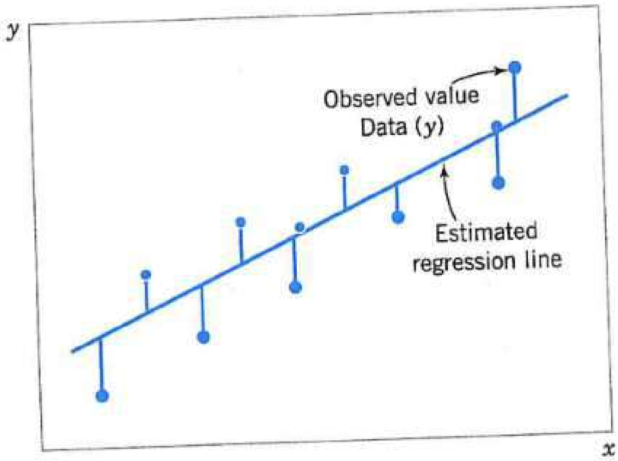


Figure 2. Deviations of the data from the estimated regression model.

Suppose that we have n pairs of observations $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$. Figure 2 shows a typical scatter plot of observed data and a candidate for the estimated regression line. The estimates of β_0 and β_1 should result in a line that is (in some sense) a "best fit" to the data. The German scientist Karl Gauss (1777-1855) proposed estimating the parameters β_0 and β_1 in Equation (1) to minimize the sum of the squares of the vertical deviations in Figure 2.

We call this criterion for estimating the regression coefficients the ***method of least squares***. Using Equation (1), we may express the n observations in the sample as

$$y_i = \beta_0 + \beta_1 x_i + \in_i, \quad i = 1, 2, \cdots, n \tag{2}$$

Definition: The **least squares estimates** of the intercept $\hat{\beta}_0$ and slope $\hat{\beta}_1$ in the simple linear regression model are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{3}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \tag{4}$$

Where $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ and $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

The **fitted** or **estimated regression line** is therefore

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{5}$$

Note that each pair of observations satisfies the relationship

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + e_i, \quad i = 1, 2, \cdots, n$$

where $e_i = y_i - \hat{y}_i$ is called the ***residual***. *The residual describes the error in the fit of the model to the ith observation* y_i.

Notationally, it is occasionally convenient to give special symbols to the numerator and denominator of Equation (4). Given data $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$, let

$$SS_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} \tag{6}$$

$$SS_{xy} = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} \tag{7}$$

Example: We will fit a simple linear regression model to the oxgyen purity data in Table 1. The following quantities may be computed:

$$n=20 \quad \sum_{i=1}^{20} x_i = 23.92 \quad \sum_{i=1}^{20} y_i = 1843.21 \quad \bar{x} = 1.1960 \quad \bar{y} = 92.1605$$

$$\sum_{i=1}^{20} y_i^2 = 170044.5321 \quad \sum_{i=1}^{20} x_i^2 = 29.2892 \quad \sum_{i=1}^{20} x_i y_i = 2214.6566$$

$$SS_{xx} = \sum_{i=1}^{20} x_i^2 - \frac{\left(\sum_{i=1}^{20} x_i\right)^2}{20} = 29.2892 - \frac{(23.92)^2}{20} = 0.68088 \quad \text{and}$$

$$SS_{xy} = \sum_{i=1}^{20} x_i y_i - \frac{\left(\sum_{i=1}^{20} x_i\right)\left(\sum_{i=1}^{20} y_i\right)}{20} = 2214.6566 - \frac{(23.92)(1843.21)}{20} = 10.17744$$

Therefore, the least squares estimates of the slope and intercept are

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{10.17744}{0.68088} = 14.94748 \text{ and } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 92.1605 - (14.94748)1.196 = 74.28331$$

The fitted simple linear regression model (with the coefficients reported to three decimal places) is

$$\hat{y} = 74.283 + 14.947x$$

This model is plotted in Figure 3, along with the sample data.

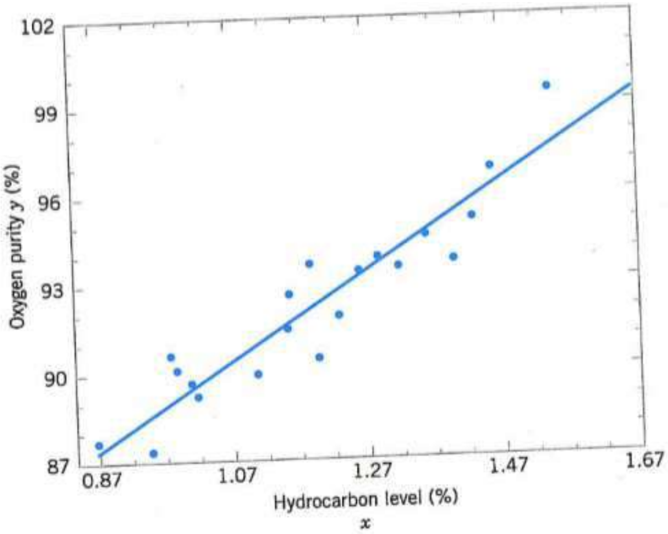


Figure 3. Scatter plot of oxygen purity y versus hydrocarbon level x and regression model $\hat{y} = 74.283 + 14.947x$.

Using the regression model of Example 1, we would predict oxygen purity of $\hat{y} = 89.23\%$ when the hydrocarbon level is $x = 1.00\%$. ***The purity 89.23 % may be interpreted as an estimate of the true population mean purity when x=1.00 %, or as an estimate of a new observation when x=1.00 %.*** These estimates are, of course, subject to error; that is, it is unlikely that a future observation on purity would be exactly 89.23 % when the hydrocarbon level is 1.00 %. In subsequent sections we will see how to use confidence intervals to describe the error in estimation from a regression model.

Estimating σ^2

There is actually another unknown parameter in our regression model, σ^2 (***the variance of the error term*** $\in$). The residuals $e_i = y_i - \hat{y}_i$ are used to obtain an estimate of σ^2. The sum of squares of the residuals, often called the ***error sum of squares***, is

$$SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (8)$$

We can show that the expected value of the error sum of squares is $E(SS_E) = (n-2)\sigma^2$. Therefore an **unbiased estimator** of σ^2 is

$$\boxed{\hat{\sigma}^2 = \frac{SS_E}{n-2} \qquad (9)}$$

Computing SS_E using Equation (8) would be fairly tedious. A more convenient computing formula can be obtained by substituting $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ into Equation (8) and simplifying.

The resulting formula is

$$\boxed{SS_E = SS_T - \hat{\beta}_1 SS_{xy} \qquad (10)}$$

Where $SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$ is the ***total sum of squares of the*** *response variable* y.

The error sum of squares and the estimate of σ^2 for the oxygen purity data;

$$SS_T = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 170044.5321 - 20 \times (92.1605)^2 = 173.3769$$

$$SS_E = SS_T - \hat{\beta}_1 SS_{xy} = 173.3769 - 14.94748 \times 10.17744 = 21.249819$$

$$\hat{\sigma}^2 = \frac{SS_E}{n-2} = \frac{21.249819}{18} = 1.18$$

8.2. Properties of the Least Squares Estimators

The values of $\hat{\beta}_0$ and $\hat{\beta}_1$ depend on the observed y's; thus, the least squares estimators of the regression coefficients may be viewed as random variables. We will investigate the bias and variance properties of the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$.

$$E\left(\hat{\beta}_1\right) = \beta_1 \text{ and } V\left(\hat{\beta}_1\right) = \frac{\sigma^2}{SS_{xx}} \tag{11}$$

Thus, $\hat{\beta}_1$ is an **unbiased estimator** of the true slope $\hat{\beta}_1$.

For the intercept, we show that

$$E\left(\hat{\beta}_0\right) = \beta_0 \text{ and } V\left(\hat{\beta}_0\right) = \sigma^2\left[\frac{1}{n} + \frac{\overline{x}^2}{SS_{xx}}\right] \tag{12}$$

Thus, $\hat{\beta}_0$ is an **unbiased estimator** of the intercept β_0.

The estimate of σ^2 could be used in Equations 11 and 12 to provide estimates of the variance of the slope and the intercept. We call the square roots of the resulting variance estimators the **estimated standard errors** of the slope and intercept, respectively.

Definition:
In simple linear regression the **estimated standard error of the slope** and the **estimated standard error of the intercept** are

$$se\left(\hat{\beta}_1\right) = \sqrt{\frac{\hat{\sigma}^2}{SS_{xx}}} \quad \text{and} \quad se\left(\hat{\beta}_0\right) = \sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\overline{x}^2}{SS_{xx}}\right]}$$

respectively, where $\hat{\sigma}^2$ is computed from Equation (9): $\hat{\sigma}^2 = \frac{SS_E}{n-2}$

8.3. Hypothesis Tests in Simple Linear Regression

An important part of assessing the adequacy of a linear regression model is testing statistical hypotheses about the model parameters and constructing certain confidence intervals. To test hypotheses about the slope and intercept of the regression model, we must make the additional assumption that the error component in the model, ε, is normally distributed. Thus, the complete assumptions are that the errors are normally and independently distributed with mean zero and variance σ^2, abbreviated $N\left(0, \sigma^2\right)$.

8.3.1. Use of t-Tests

Suppose we wish to test the hypothesis that the slope equals a constant, say, $\beta_{1,0}$. The appropriate hypotheses are

$$\begin{aligned} H_0 &: \beta_1 = \beta_{1,0} \\ H_1 &: \beta_1 \neq \beta_{1,0} \end{aligned} \tag{13}$$

where we have assumed a two sided alternative. The statistic

$$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2 / SS_{xx}}} \tag{14}$$

Follows the t distribution with n-2 degrees of freedom under $H_0 : \beta_1 = \beta_{1,0}$. We would reject $H_0 : \beta_1 = \beta_{1,0}$ if

$$|t| > t_{\alpha/2,n-2} \tag{15}$$

where t_0 is computed from Equation (14). The denominator of Equation (14) is the standard error of the slope, so we could write the test statistic as

$$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{se\left(\hat{\beta}_1\right)}$$

A similar procedure can be used to hypotheses about the intercept. To test

$$\begin{aligned} H_0 &: \beta_0 = \beta_{0,0} \\ H_1 &: \beta_0 \neq \beta_{0,0} \end{aligned} \tag{16}$$

we would use the statistic

$$T_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{SS_{xx}} \right]}} = \frac{\hat{\beta}_0 - \beta_{0,0}}{se\left(\hat{\beta}_0\right)} \tag{17}$$

and reject the null hypothesis if the computed value of this test statistics, t_0, is such that $|t_0| > t_{\alpha/2,n-2}$. Note that the dominator of the test statistic in Equation (17) is just the standard error of the intercept.

A very important special case of the hypotheses of Equation (13) is

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 \neq 0 \end{aligned} \tag{18}$$

The hypotheses relate to the **significance of regression.** Failure to reject $H_0 : \beta_1 = 0$ is equivalent to concluding that there is no linear relationship between x and Y. This situation is illustrated in Figure 4. Note that this may imply either that x is of little value in explaining the variation in Y and the best estimator of Y for any x is $\hat{y} = \overline{Y}$ (Figure 4 (a)) or that the true relationship between x and Y is not linear (Figure 4 (b)). Alternatively, if $H_0 : \beta_1 = 0$ is rejected, this implies that x is of value in explaining the variability in Y (see Figure 5). Rejecting $H_0 : \beta_1 = 0$ could mean either that the straight-line model is adequate (Figure 5(a)) or that, although there is a linear effect of x, better results could be obtained with the addition of high order polynomial terms in x (Figure 5b).

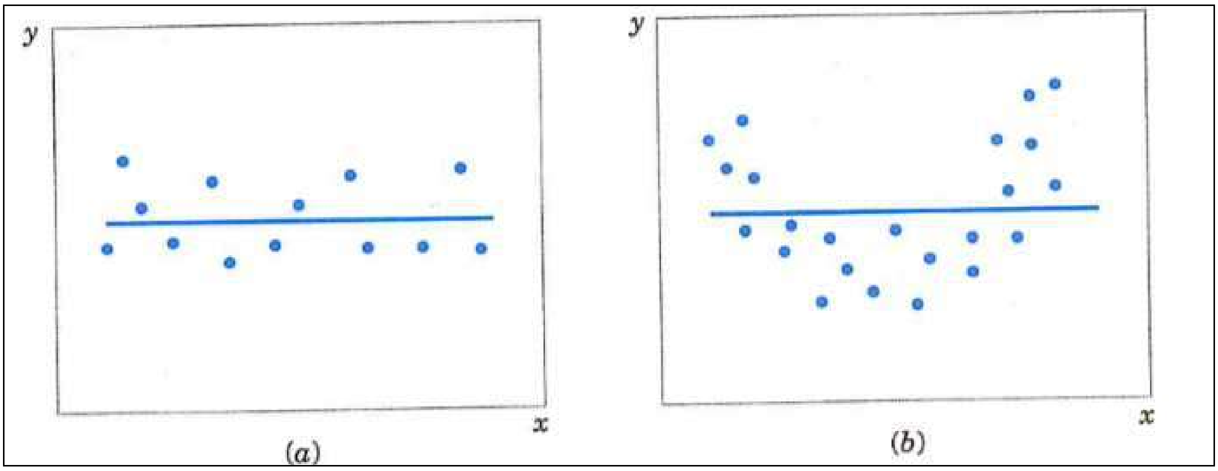


Figure 4. The hypothesis $H_0 : \beta_1 = 0$ is not rejected.

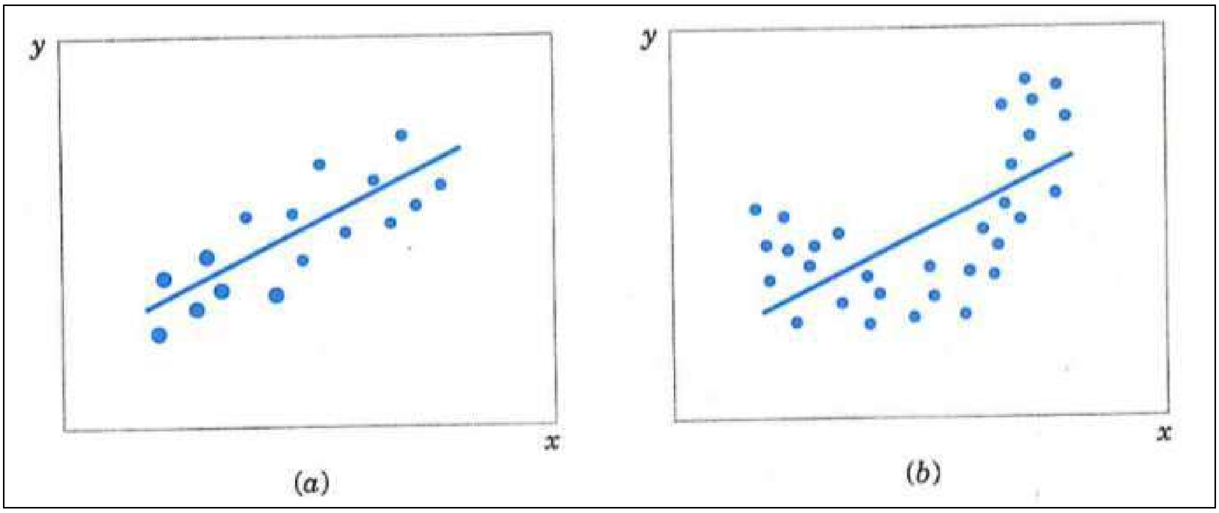


Figure 5. The hypothesis $H_0 : \beta_1 = 0$ is rejected.

Example: We will test for significance of regression using the model for the oxygen purity data. The hypotheses are

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

and we will use $\alpha = 0.05$. We have before obtained $\hat{\beta}_1 = 14.97$ n=20 $S_{xx} = 0.68088$ $\hat{\sigma}^2 = 1.18$ so the t-statistic becomes

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 / SS_{xx}}} = \frac{\hat{\beta}_1}{se\left(\hat{\beta}_1\right)} = \frac{14.947}{\sqrt{1.18/0.68088}} = 11.35$$

Since the reference value of t is $t_{0.025,18} = 2.101$, the value of the test statistic is very far into the critical region, implying that $H_0 : \beta_1 = 0$ should be rejected.

The t-statistic for testing the hypothesis $H_0 : \beta_0 = 0$ is

$$t_0 = \frac{\hat{\beta}_0}{\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}}\right]}} = \frac{\hat{\beta}_0}{se\left(\hat{\beta}_0\right)} = \frac{74.283}{\sqrt{1.18\left[\frac{1}{20} + \frac{1.1960^2}{0.68088}\right]}} = 46.62$$

Since the reference value of t is $t_{0.025,18} = 2.101$, the value of the test statistic is very far into the critical region, clearly, then the hypothesis that the intercept is zero is rejected.

8.3.2. Analysis of Variance Approach to Test Significance of Regression

A method called the **analysis of variance** can be used to test for significance of regression. The procedure partitions the total variability in the response variable into meaningful components as the basis for the test. The **analysis of variance identity** is as follows:

$$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2 + \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \qquad (19)$$

The two components on the right-hand-side of Equation (19) measure, respectively, the amount of variability in y_i accounted for by the regression line and the residual variation left unexplained by the regression line. We usually call $SS_E = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$ the **error sum of squares** and $SS_R = \sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2$ the **regression sum of squares.** Symbolically, Equation (19) may be written as

$$SS_T = SS_R + SS_E \qquad (20)$$

where $SS_T = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = SS_{yy}$ is the **total corrected sum of squares** of y. In Section 8.1 we noted that $SS_E = SS_T - \hat{\beta}_1 SS_{xy}$ (see Equation (10)), so since $SS_T = \hat{\beta}_1 SS_{xy} + SS_E$, we note that the regression sum of squares is $SS_R = \hat{\beta}_1 SS_{xy}$. The total sum of squares SS_T has n-1 degrees of freedom, and SS_R and SS_E have 1 and n-2 degrees of freedom, respectively.

We may show that $E\left[SS_E/(n-2)\right]=\sigma^2$, $E(SS_R)=\sigma^2+\beta_1^2 SS_{xx}$ and that SS_E/σ^2 and SS_R/σ^2 are independent chi-square random variables with n-2 and 1 degrees of freedom, respectively. Thus, if the null hypothesis is true, the statistics

$$F_0=\frac{SS_R/1}{SS_E/(n-2)}=\frac{MS_R}{MS_E} \tag{21}$$

follows the $F_{1,n-2}$ ***distribution, and we would reject*** H_0 ***if*** $f_0>f_{\alpha,1,n-2}$. The quantities $MS_R=SS_R/1$ and $MS_E=SS_E/(n-2)$ are called **mean squares**. In general, a mean square is always computed by dividing a sum of squares by its number of degrees of freedom. The test procedure is usually arranged in an **analysis of variance** table, such as Table 2.

Table 2. Analysis of Variance for Testing Significance of Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F₀
Regression	$SS_R=\hat{\beta}_1 SS_{xy}$	1	MS_R	MS_R/MS_E
Error	$SS_E=SS_T-\hat{\beta}_1 SS_{xy}$	n-2	MS_E	
Total	$SS_T=SS_{yy}$	n-1		

Note that $MS_E=\hat{\sigma}^2$.

Example: We will use the analysis of variance approach to test for significance of regression using the oxygen purity data model. Recall that $SS_T=173.38$, $\hat{\beta}_1=14.947$, $SS_{xy}=10.17744$, and n=20. The regression sum of squares is

$$SS_R=\hat{\beta}_1 SS_{xy}=(14.947)\times 10.17744=152.13$$

and the error sum of squares is

$$SS_E=SS_T-SS_R=173.38-152.13=21.25$$

The analysis of variance for testing $H_0:\beta_1=0$ is as in below:

Table 3. Analysis of Variance for Testing Significance of Regression for Oxygen Purity Data

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F₀
Regression	152.13	1	152.13	128.92
Error	$21.25\leftarrow SS_E$	18	$1.18\leftarrow\hat{\sigma}^2$	
Total	173.38	19		

The test statistic is $f_0=MS_R/MS_E=152.13/1.18=128.92$ and since $f_0=128.92>f_{0.05,1,18}=4.414$ we reject H_0, so we conclude that β_1 is not zero.

** ***Note that the analysis of variance procedure for testing for significance of regression ($\beta_{1,0} = 0$) is equivalent to the t-test. That is, either procedure will lead to the same conclusions.***

8.4. Confidence Intervals on the Slope and Intercept

In addition to point estimates of the slope and intercept, it is possible to obtain **confidence interval** estimates of these parameters. The width of these confidence intervals is a measure of the overall quality of the regression line. If the error term, $\in_i$, in the regression model are normally and independently distributed,

$$\left(\hat{\beta}_1 - \beta_1\right) / \sqrt{\hat{\sigma}^2 / SS_{xx}} \text{ and } \left(\hat{\beta}_0 - \beta_0\right) / \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{SS_{xx}}\right]}$$

are both distributed as t random variables with n-2 degrees of freedom. This leads to the following definition of $100\left(1-\alpha\right)\%$ confidence intervals on the slope and intercept.

Definition: Under the assumption that the observations are normally and independently distributed, a $100\left(1-\alpha\right)\%$ **confidence interval on the slope** β_1 in simple linear regression is

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{SS_{xx}}} \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{SS_{xx}}} \qquad (22)$$

Similarly, a $100\left(1-\alpha\right)\%$ **confidence interval on the intercept** β_0 is

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\overline{x}^2}{SS_{xx}}\right]} \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\overline{x}^2}{SS_{xx}}\right]} \qquad (23)$$

Example: We will find a 95% confidence interval on the slope of the regression line using the oxygen purity data. Recall that $\hat{\beta}_1 = 14.947$, $SS_{xx} = 0.68088$, and $\hat{\sigma}^2 = 1.18$, then,

$$\hat{\beta}_1 - t_{0.025,18}\sqrt{\frac{\hat{\sigma}^2}{SS_{xx}}} \le \beta_1 \le \hat{\beta}_1 + t_{0.025,18}\sqrt{\frac{\hat{\sigma}^2}{SS_{xx}}}$$

or

$$14.947 - 2.101\sqrt{\frac{1.18}{0.68088}} \le \beta_1 \le 14.947 + 2.101\sqrt{\frac{1.18}{0.68088}}$$

This simplifies to

$$12.197 \le \beta_1 \le 17.697$$

8.5. Coefficient of Determination (R^2)

The quantity

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} \qquad 0 \le R^2 \le 1 \tag{24}$$

is called the **coefficient of determination** and *is often used to judge the adequacy of a regression model.* Subsequently, in the case where X and Y are jointly distributed random variables, R^2 is the square of the correlation coefficient between X and Y. We often refer loosely to R^2 as the amount of variability in the data explained or accounted for by the regression model. For the oxygen purity regression model, we have $R^2 = \frac{SS_R}{SS_T} = \frac{152.13}{173.38} = 0.877$; that is, the model accounts for 87.7 % of the variability in the data.

For Example, suppose we wish to develop a regression model relating the shear strength of spot welds to the weld diameter. In this example, weld diameter cannot be controlled. We would randomly select n spot welds and observe a diamater (X_i) and a shear strength (Y_i) for each. Therefore, (X_i, Y_i) are jointly distributed random variables. The **sample correlation coefficient (R)** between X_i and Y_i could be calculated as given in below.

$$R = \frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)\left(X_i - \bar{X}\right)}{\left[\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2\right]^{1/2}} = \frac{SS_{xy}}{\left(SS_{xx}SS_T\right)^{1/2}} \tag{25}$$

$\hat{\beta}_1$ and R are closely related, although they provide somewhat different information. ***The sample correlation coefficient R measures the linear association between Y and X, while*** $\hat{\beta}_1$ ***measures the predicted change in the mean of Y for a unit change in X.*** In the case of a mathematical variable x, R has no meaning because the magnitude of R depends on the choice of spacing of x.

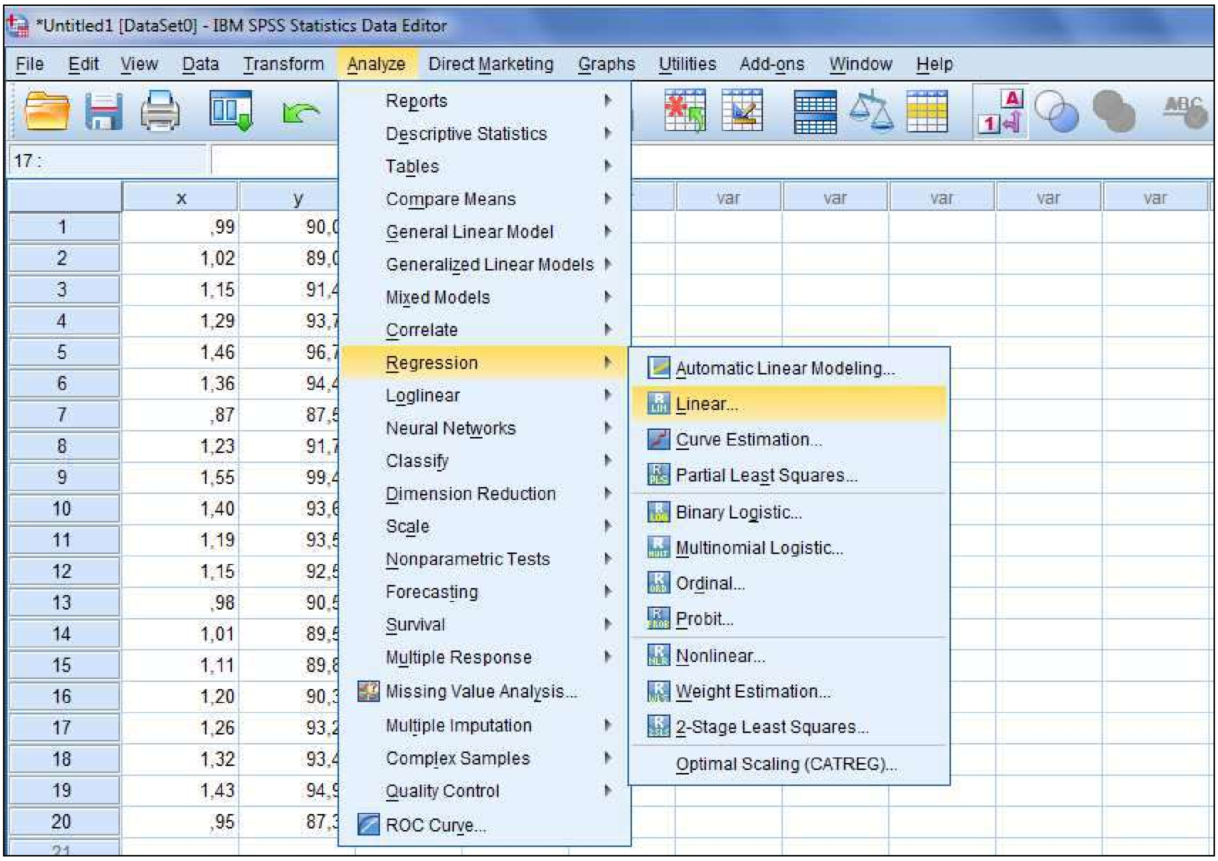
$$R^2 = \hat{\beta}_1^2 \frac{SS_{xx}}{SS_{yy}} = \frac{\hat{\beta}_1 SS_{xy}}{SS_T} = \frac{SS_R}{SS_T} \tag{26}$$

which is just **the coefficient of determination (R^2).** That is, *the coefficient of determination* R^2 *is just the square of the correlation coefficient between Y and X.*

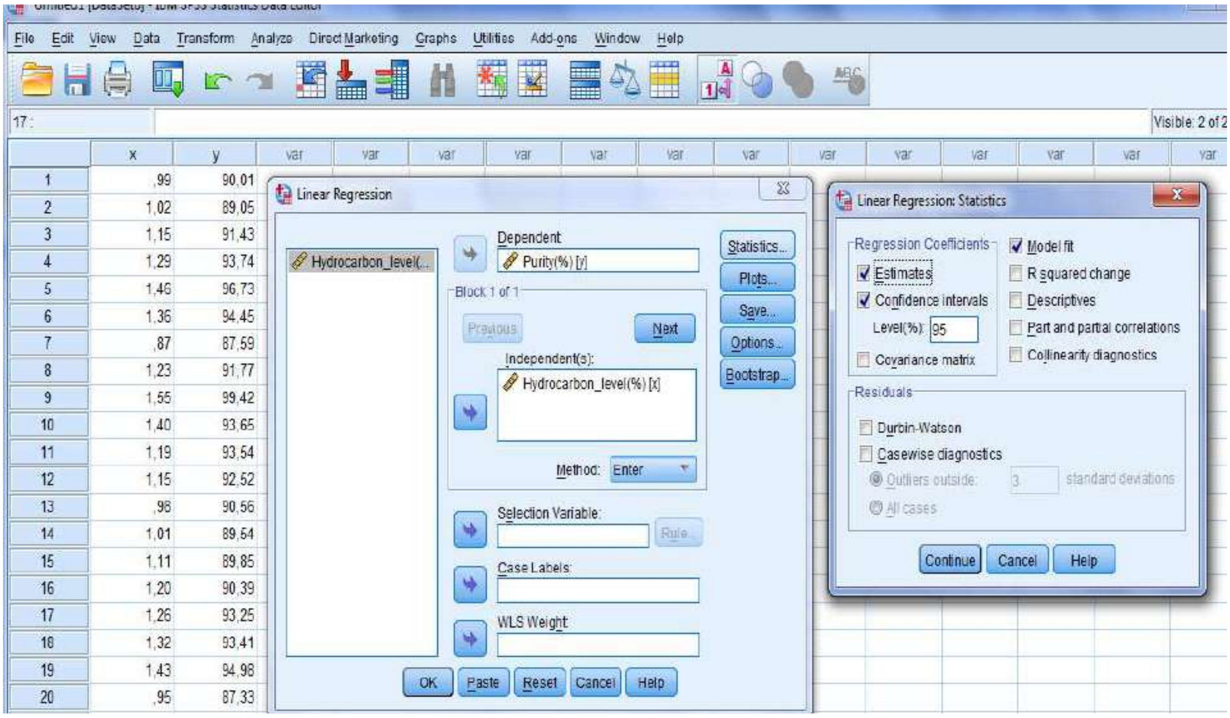
For the oxygen purity regression model, the sample correlation coefficient is $r = \frac{SS_{xy}}{\left[SS_{xx}SS_T\right]^{1/2}} = \frac{10.17744}{\left[0.68088 \times 173.38\right]^{1/2}} = 0.9367$.

İST292 STATISTICS LESSON 7

REGRESSION ANALYSIS IN SPSS



	x	y
1	,99	90,(
2	1,02	89,(
3	1,15	91,4
4	1,29	93,7
5	1,46	96,7
6	1,36	94,4
7	,87	87,5
8	1,23	91,7
9	1,55	99,4
10	1,40	93,6
11	1,19	93,5
12	1,15	92,5
13	,98	90,5
14	1,01	89,5
15	1,11	89,8
16	1,20	90,3
17	1,26	93,2
18	1,32	93,4
19	1,43	94,9
20	,95	87,3



	x	y
1	,99	90,01
2	1,02	89,05
3	1,15	91,43
4	1,29	93,74
5	1,46	96,73
6	1,36	94,45
7	,87	87,59
8	1,23	91,77
9	1,55	99,42
10	1,40	93,65
11	1,19	93,54
12	1,15	92,52
13	,98	90,56
14	1,01	89,54
15	1,11	89,85
16	1,20	90,39
17	1,26	93,25
18	1,32	93,41
19	1,43	94,98
20	,95	87,33

Hydrocarbon_level(...

Dependent: Purity(%) [y]

Independent(s): Hydrocarbon_level(%) [x]

Method: Enter

Selection Variable:

Case Labels:

WLS Weight:

Regression Coefficients

☑ Estimates

☑ Confidence intervals

Level(%): 95

□ Covariance matrix

☑ Model fit

□ R squared change

□ Descriptives

□ Part and partial correlations

□ Collinearity diagnostics

Residuals

□ Durbin-Watson

□ Casewise diagnostics

○ Outliers outside: 3 standard deviations

○ All cases

OUTPUTS

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Hydrocarbon_level(%)[b]	.	Enter

a. Dependent Variable: Purity(%)

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,937[a]	,877	,871	1,08653

a. Predictors: (Constant), Hydrocarbon_level(%)

From **Model Summary Table** it is clear that for the oxygen purity regression model, the sample correlation coefficient is $r = 0.937$. Since, the coefficient of determination R^2 is just the square of the correlation coefficient between y and x, for the oxygen purity regression model's $\mathrm{R}^2 = 0.877$, that is, the model accounts for 87.7 % of the variability in the data.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	152,127	1	152,127	128,862	,000[b]
	Residual	21,250	18	1,181		
	Total	173,377	19			

a. Dependent Variable: Purity(%)

b. Predictors: (Constant), Hydrocarbon_level(%)

From **ANOVA Table (Analysis of Variance for Testing Significance of Regression for Oxygen Purity Data),**

We will use the analysis of variance approach to test for significance of regression using the oxygen purity data model. From ANOVA Table it is clear that ***total sum of squares of the dependent variable*** $\mathrm{SS_T} = 173.377$, ***the regression sum of squares*** is $\mathrm{SS_R} = 152.127$, ***the error sum of squares*** is $\mathrm{SS_E} = \mathrm{SS_T} - \mathrm{SS_R} = 173.377 - 152.127 = 21.250$ and ***the estimate of*** σ^2 ***for the oxygen purity data*** $\hat{\sigma}^2 = \dfrac{\mathrm{SS_E}}{\mathrm{n}-2} = \dfrac{21.250}{18} = 1.181$.

1) We will test for significance of regression using the model for the oxygen purity data. The hypotheses are

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

and we will use $\alpha = 0.05$.

We would reject H_0 if $f_0 > f_{0.05,1,18}$. The quantities $MS_R = SS_R / 1$ and $MS_E = SS_E / 18$ are called **mean squares**. In general, a mean square is always computed by dividing a sum of squares by its number of degrees of freedom. The test statistic is $f_0 = MS_R / MS_E = 152.127 / 1.181 \cong 128.862$ and since $f_0 = 128.862 > f_{0.05,1,18} = 4.414$ we reject H_0, so we conclude that β_1 is not zero. (Remember $MS_E = \hat{\sigma}^2$). We can test this hypothesis also by using p-value (Sig.), since p-value=0.000<0.05, $H_0 : \beta_1 = 0$ is rejected. Comment on $H_0 : \beta_1 = 0$ *is rejected, this implies that x is of value in explaining the variability in y. Rejecting* $H_0 : \beta_1 = 0$ *could mean either that the straight-line model is adequate.*

** ***Note that the analysis of variance procedure for testing for significance of regression is equivalent to the t-test. That is, either procedure will lead to the same conclusions.***

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	74,283	1,593		46,617	,000	70,936	77,631
	Hydrocarbon_level(%)	14,947	1,317	,937	11,352	,000	12,181	17,714

a. Dependent Variable: Purity(%)

From **Coefficients Table** we know that $\hat{\beta}_1 = 14.97$ and the t-statistic is $t_0 = \frac{\hat{\beta}_1}{se\left(\hat{\beta}_1\right)} \cong \frac{14.947}{1.317} = 11.352$. Since the reference value of t is $t_{0.025,18} = 2.101$, the value of the test statistic is very far into the critical region, implying that $H_0 : \beta_1 = 0$ should be rejected. We can test this hypothesis also by using p-value (Sig.), since p-value=0.000<0.05, $H_0 : \beta_1 = 0$ is rejected.

2) From **Coefficients Table** the t-statistic for testing the hypothesis $\begin{matrix} H_0 : \beta_0 = 0 \\ H_1 : \beta_0 \neq 0 \end{matrix}$ is $t_0 = \frac{\hat{\beta}_0}{se\left(\hat{\beta}_0\right)} = \frac{74.283}{1.593} \cong 46.617$. Since the reference value of t is $t_{0.025,18} = 2.101$, the value of the test statistic is very far into the critical region, clearly, then the hypothesis that the intercept is zero is rejected. This hypothesis can also be tested by using p-value (Sig.), since p-value=0.000<0.05, $H_0 : \beta_0 = 0$ is rejected.

3) From **Coefficients Table*, the fitted simple linear regression model (with the coefficients reported to three decimal places) is***

$$\hat{y} = 74.283 + 14.947x$$

Using the regression model, we would predict oxygen purity of $\hat{y} = 89.23\%$ when the hydrocarbon level is $x = 1.00\%$. The purity 89.23 % may be interpreted as an estimate of the true population mean purity when x=1.00 %, or as an estimate of a new observation when x=1.00 %. These estimates are, of course, subject to error; that is, it is unlikely that a future observation on purity would be exactly 89.23 % when the hydrocarbon level is 1.00 %.

4) From **Coefficients Table**, we will find a 95% confidence interval on the slope of the regression line using the oxygen purity data;

$$12.181 \le \beta_1 \le 17.714$$

Similarly, 95 % confidence interval on the intercept β_0 is

$$70.936 \le \beta_0 \le 77.631$$