ECON - STAT I Lecture Notes | CHAPTERS: 2 & 3

DATA TYPES

I Qualitative (Categorical) Data (Kategorik)
- (i) Nominal
- (ii) Ordinal

II Quantitative (Numerical) Data (Sayısal)
- (i) Ratio
- (ii) Interval
- Discrete
- Continuous

I Qualitative Data ∋ Sayılar kodlamak için kullanılır ya da karşılaştırmalı anlam ifade eder. Sayıların "ölçüm" ya da "sayı" değeri olmaz.

(i) Nominal; Kategori belirtmek için kullanıyoruz. Sayıları farklı kategorilere eşleşik de olurdu.

Ex: 0 - Male
1 - Female

1 - Engineering
2 - Education
3 - Administration
4 - Science
5 - Other

(ii) Ordinal; Sayılar birbirlerine göreceli derece ifade eder ama sayı değeri yoktur.

Ex: How is your teacher performing?

1 - Very Low
2 - Low
3 - Moderate
4 - Good
5 - Very Good

* Sorudan bağımsız, "Hocanın performansı 4" desek nesela, neye göre 4?

II Quantitative Data ⇒ Sayıların "ölçüm" ya da "Sayı", "Çokluk" değeri olur.

(i) Ratio; "Natural Zero": Sıfır sayısı "Yok" anlamını taşır.

Ex: Weight, Height, Money ... etc.

(ii) Interval; "No natural Zero": Sıfır sayısı bir ölçüm ifade eder, "yok" anlamı taşımaz.

Ex: Temperature: 0°C bir değerdir, hava sıcaklığı yok anlamına gelmez.

Location: ——+—— 0: Centre location.

Discrete vs. Continuous Data;

Discrete ⇒ Count: How many?

Continuous ⇒ Measurement: Weight, Distance ... etc.

Examples;

2.1 State whether each of the following variables is categorical or numerical. If categorical, give the level of measurement. If numerical, is it discrete or continuous?

a. Number of e-mail messages sent daily by a financial planner
b. Actual cost of a student's textbooks for a given semester
c. Your monthly electricity bill
d. Faculty ranks (professor, associate professor, assistant professor, instructor)

a. Ratio, Discrete
b. Ratio, Continuous
c. Ratio, Continuous
d. Ordinal.

2.4 Faculty at one university were asked a series of questions in a recent survey. State the type of data for each question.

a. Indicate your level of satisfaction with your teaching load (very satisfied; moderately satisfied; neutral; moderately dissatisfied; very dissatisfied).
b. How many articles did you have published in refereed journals during the last year?
c. Did you attend the last university faculty meeting?
d. Do you think that the teaching evaluation process needs to be revised?

a. Ordinal

b. Ratio, Discrete

c. Yes/No ⇒ Nominal

d. Yes/No ⇒ Nominal

GRAPHICAL DISPLAY of CATEGORICAL VARIABLES

Frequency Distribution;

Classes OR Groups ⇒ All possible responses

Frequencies ⇒ Number of observations.

Classes	Freq.
Class 1	n_1
Class 2	n_2
⋮	⋮
Class k	$+\ n_k$
	n: # of observations.

Bar Charts and Pie Charts

Bar Charts; interest is in frequency

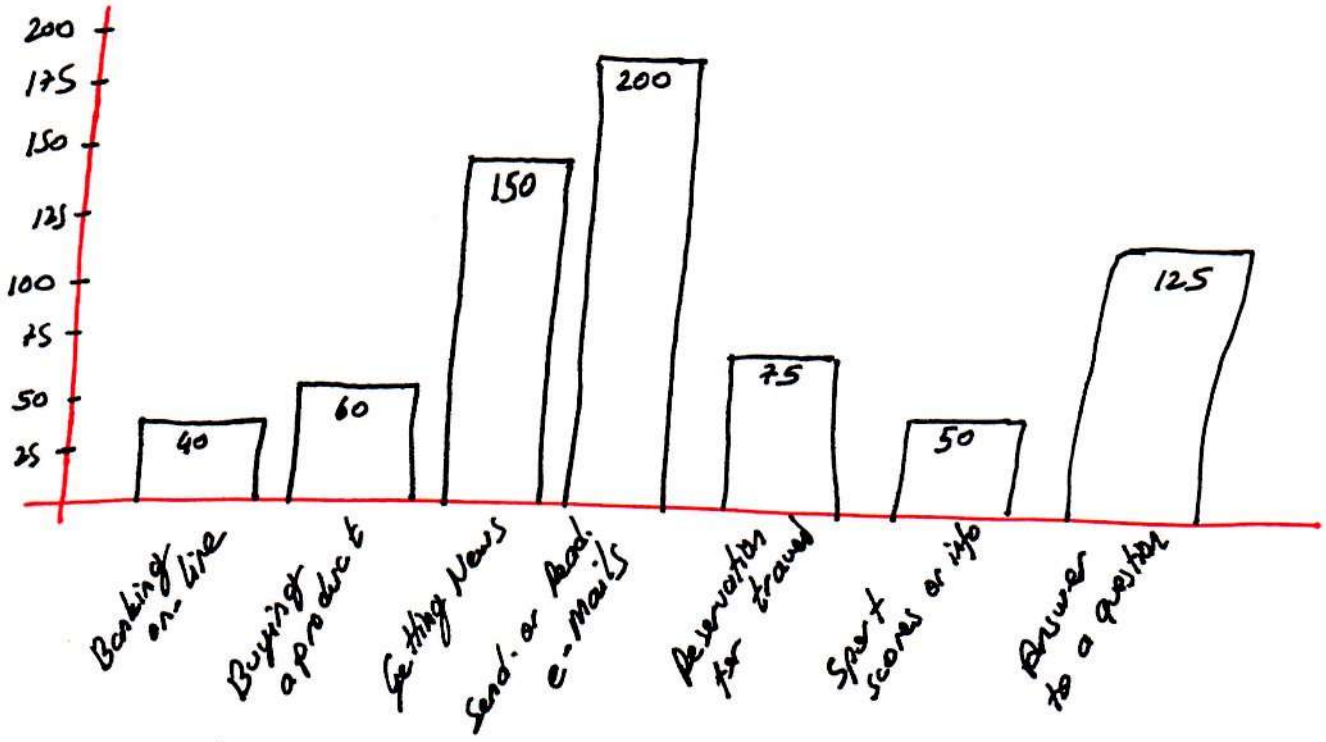
Pie Charts; interest is in percentage.

Ex:

2.16 On what type of Internet activity do you spend the most time? The responses from a random sample of 700 Internet users were banking on-line, 40; buying a product, 60; getting news, 150; sending or reading e-mail, 200; buying or making a reservation for travel, 75; checking sports scores or information, 50; and searching for an answer to a question, 125. Describe the data graphically.

Classes	Frequency
Banking on-line	40
Buying a product	60
Getting News	150
Send. or Read. e-Mails	200
Reservation for travel	75
Sport scores or info.	50
Answer to a question	$+\ 125$
	700

(i) Bar Chart

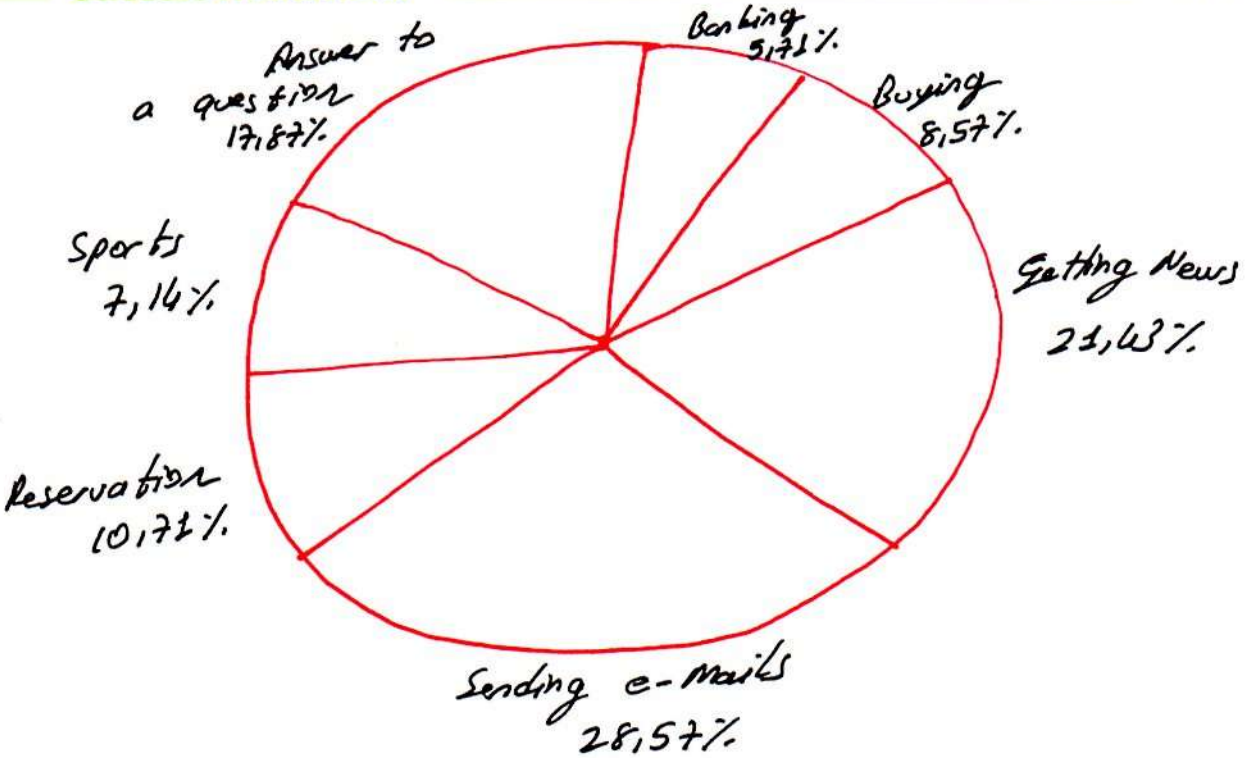


(ii) Pie-Chart

Pie → 360°, $n = 700$ ⇒ Herbir class'ın kaç derecelik açıya karşılık geldiğini ve yüzde kaçlık dilime karşılık geldiğini bulalım. Tabii ki 25% → 90° olacak, 4'te bir olduğu için.

$$Açı = \frac{n_i}{n} \times 360 \qquad Yüzde = \frac{n_i}{n} \times 100$$

Classes	Frequency	Degree	Percentage
Banking	40	20,6°	5,71 %
Buying	60	30,9°	8,57 %
Getting News	150	77,2°	21,43 %
Sending e-Mails	200	102,9°	28,57 %
Reservation	75	38,6°	10,71 %
Sports	50	25,7°	7,14 %
Answer to Question	125	64,1°	17,87 %



Pareto Diagram; is a bar chart that displays the frequency of defect causes. (Burada büyükten küçüğe sıralama dalgası var.) Cumulative frequency line is also shown in the diagram.

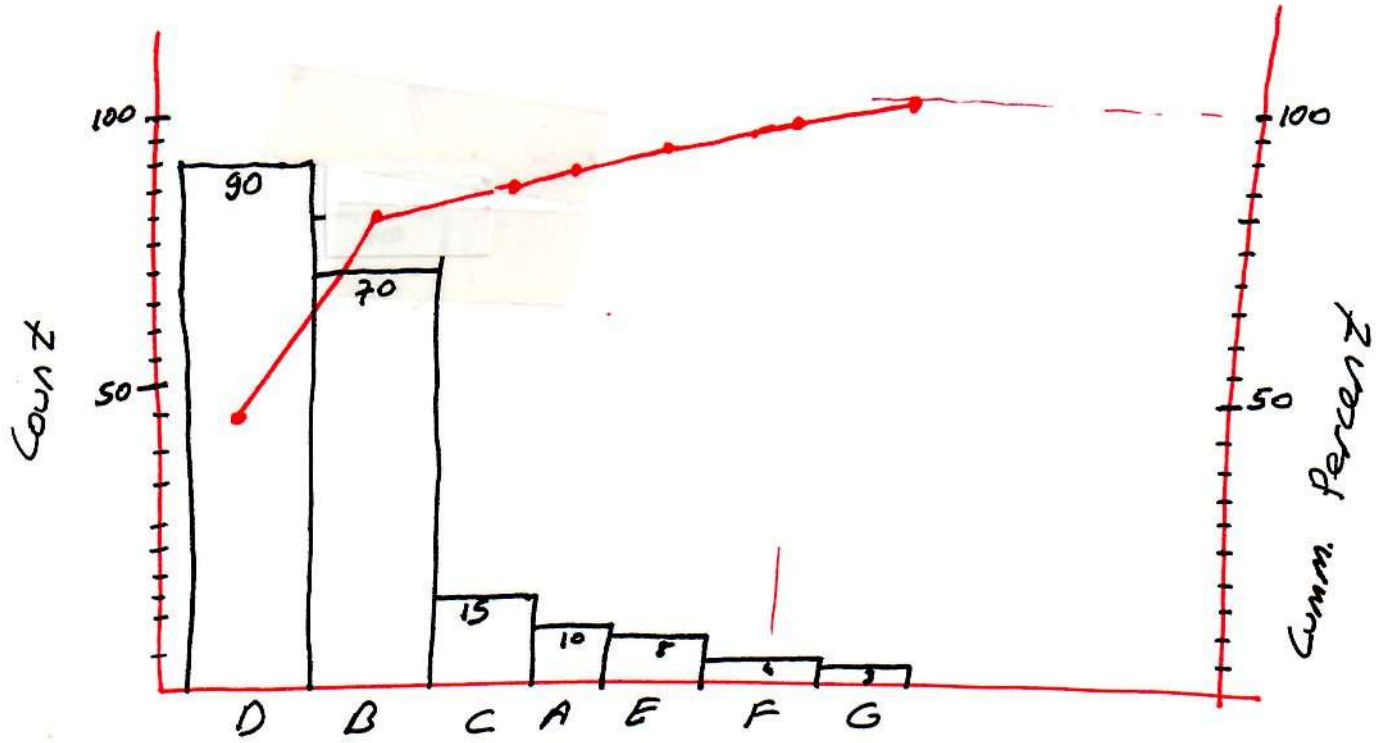
Ex16

2.10 A company has determined that there are seven possible defects for one of its product lines. Construct a Pareto diagram for the following defect frequencies:

Defect Code	Frequency
A	10
B	70
C	15
D	90
E	8
F	4
G	3

(RANKED)

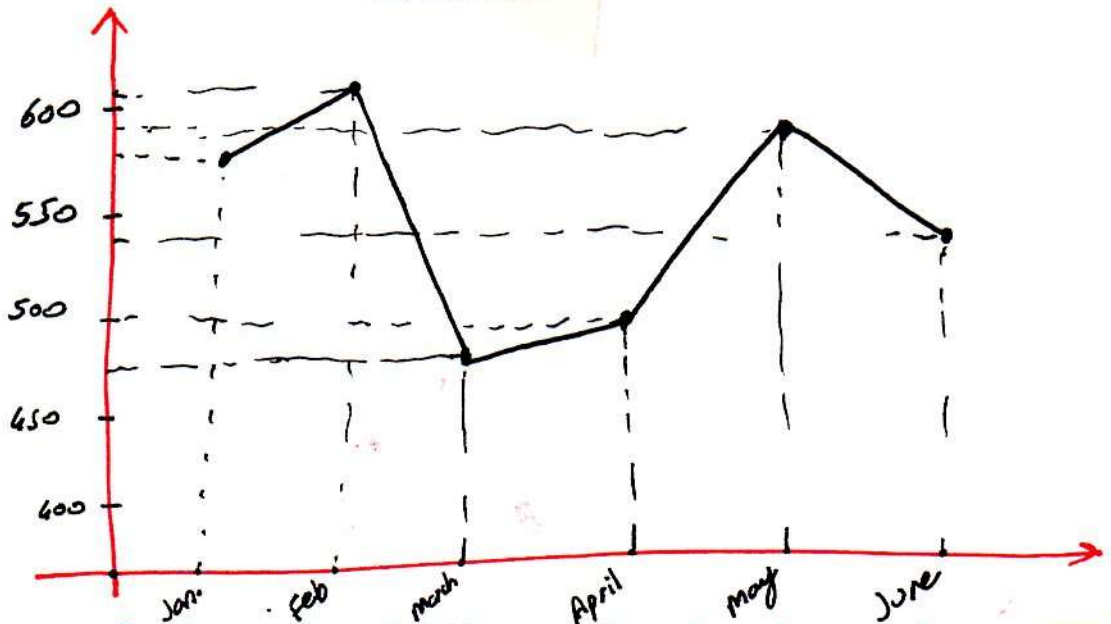
Defect	Count	Percent	Cum.%
D	90	45%	45%
B	70	35%	80%
C	15	7,5%	87,5%
A	10	5%	92,5%
E	8	4%	96,5%
F	4	2%	98,5%
G	3	1,5%	100%
	n = 200	100%	



Ethe Time Series Data

2.20 Construct a time-series plot for the following data on weekend or night mobile phone usage (in minutes):

Month	Weekend or Night
January	575
February	603
March	469
April	500
May	586
June	540



GRAPHS TO DESCRIBE NUMERICAL VARIABLES

Frequency Distributions

(i) Number of classes;

Sample Size	# of Classes
$n < 50$	5-7
$50 \leq n < 100$	7-8
$101 \leq n < 500$	8-10

(ii) Class with;

$$w = \frac{\text{Max}(x_i) - \text{Min}(x_i)}{\#\text{ of classes.}}$$

→ Round up this Number

(iii) $\text{Min}(x_i)$'in biraz altından başlayarak tüm data'yı cover et, herbir class with = w olacak.

(iv) Tally: Herbir data'yı uygun aralığa çiz, 5. Tally'yi /// yap ki tally bitince saymak kolay olsun.

(v) Herbir class'a düşen Frekans'ı yaz.

Histogram; Bar-Chart gibi ama aralar aralık değil.

O-Give; Cumm. Percentage çiziyoruz. Pareto'da da çizmiştik, ama noktalar "Upper Bound"dan geçecek.

Alp Giray Özen | 0533 549 91 08 | alp@lecturemania.com | www.lecturemania.com

2.32 Consider the following data:

Etiy

17	62	15	65
28	51	24	65
39	41	35	15
39	32	36	37
40	21	44	37
59	13	44	56
12	54	64	59

a. Construct a frequency distribution.
b. Draw a histogram.
c. Draw an ogive.
d. Draw a stem-and-leaf display.

(i) $n = 28$

Use $k = 5$ classes

Bunu rastgele seçiyoruz, sorunun tek bir doğru cevabı yok.

(ii) $\max(x_i) = 65$

$\min(x_i) = 12$

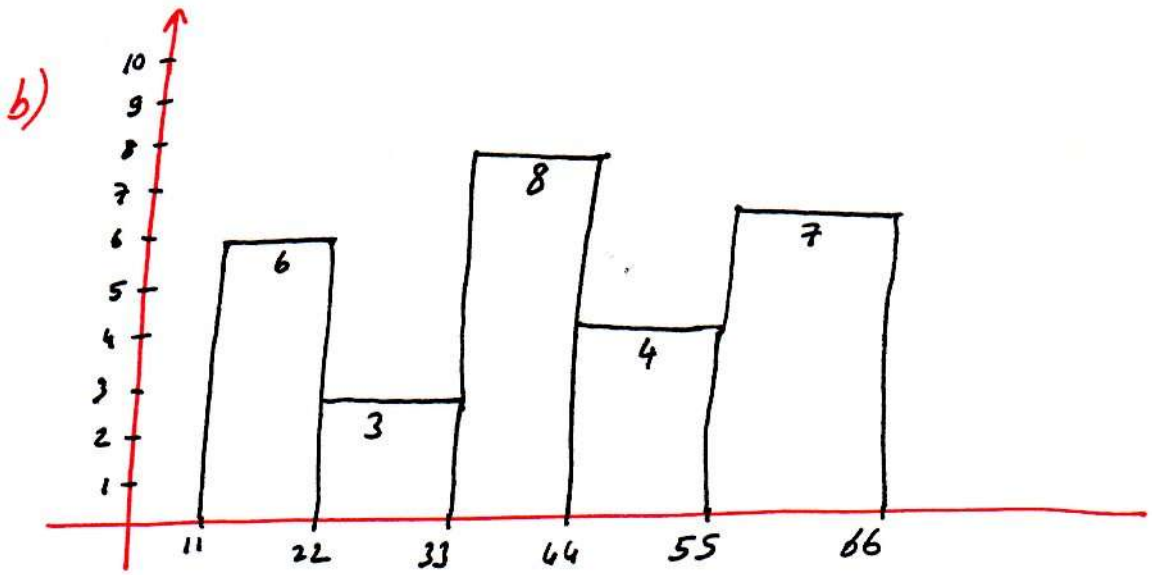
$$w = \frac{65 - 12}{5} = 10{,}6 \Rightarrow w = 11$$

(10.3 olsa da 11 dedik)

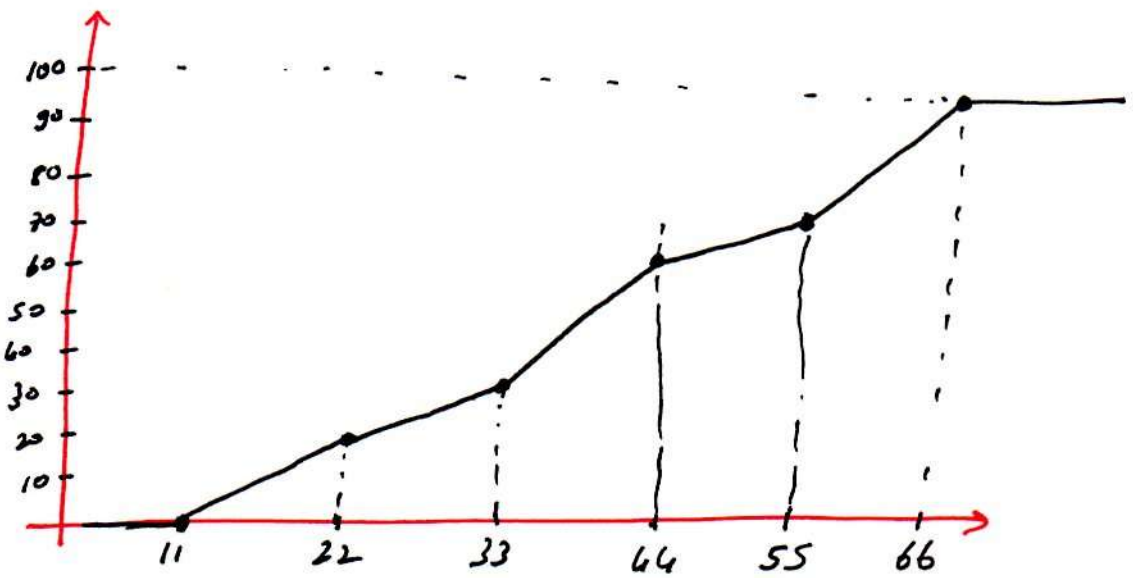
a)

Classes	Tally	Frequency	Percentage	Cumm.%
$11 \leq x_i < 22$	//// /	6	21,4%	21,4%
$22 \leq x_i < 33$	///	3	10,7%	32,1%
$33 \leq x_i < 44$	//// ///	8	28,6%	60,7%
$44 \leq x_i < 55$	////	4	14,3%	75,0%
$55 \leq x_i < 66$	//// //	7	25,0%	100,0%
		$n = 28$	100%	

b)



c)



d) Stem & Leaf Display: Sayıları onluk ve birlik basamaklarına ayırıyoruz.

1	7	5	5	3	2			
2	8	4	1					
3	9	5	9	2	6	7	7	
4	1	0	4	4				
5	1	9	6	4	9			
6	2	5	5	4				

onluk

⇒ RANK!

1	2	3	5	5	7
2	1	4	8		
3	2	5	6	7	7	9	9
4	0	1	4	4	
5	1	4	6	9	9
6	2	4	5	5	

Mesela; 40, 41, 44, 44 ⇒ 40'lı sayılar.

Hisaya dikkat et, dağılımla ilgili fikir versin.

Scatter Plot & Cross Tables; Relationship Between Variables.

2.44 Acme Delivery offers three different shipping rates for packages under 5 pounds delivered from Maine to the West Coast: regular, $3; fast, $5; and lightning, $10. To test the quality of these services, a major mail-order retailer shipped 15 packages at randomly selected times from Maine to Tacoma, Washington. The packages were shipped in groups of three by the three services at the same time to reduce variation resulting from the shipping day. The following data show the shipping cost, x, and the number of days, y, in (x, y) pairs:

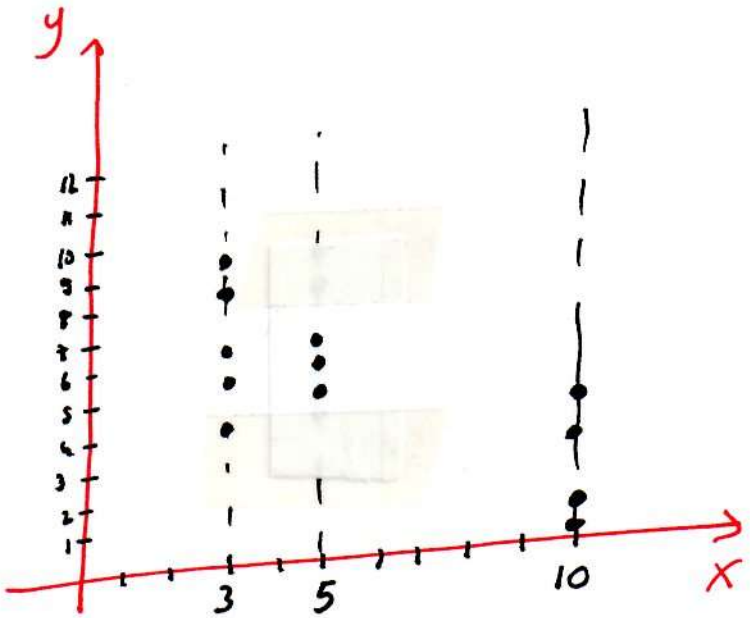
(3, 7) (5, 5) (10, 2) (3, 9) (5, 6) (10, 5) (3, 6) (5, 6) (10, 1)
(3, 10) (5, 7) (10, 4) (3, 5) (5, 6) (10, 4)

Prepare a scatter plot of the points and comment on the relationship between shipping cost and observed delivery time.

X: Shipping Cost
Y: Delivery Time

X	Y
3	7
5	5
10	2
3	9
5	6
10	5
3	6
5	6
10	1
3	10
5	7
10	4
3	5
5	6
10	4

Scatter-Plot

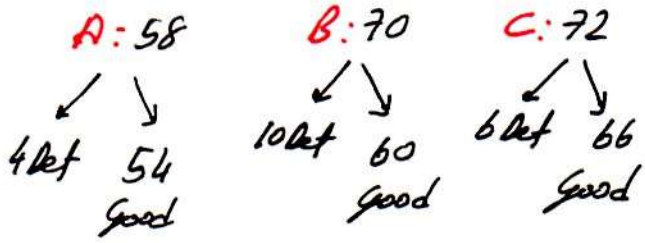


As might be expected, as shipping gets more costly, we have less delivery time. The variables are conversely related.

2.42 Three subcontractors, A, B, and C, supplied 58, 70, and 72 parts, respectively, to a plant during the last week. Of the parts supplied by subcontractor A, only 4 were defective. From the parts supplied by subcontractor B, 60 were good parts; from those supplied by subcontractor C, only 6 were defective.

Set up a cross table for the data.

Ex



Supplier / Quality	A	B	C	TOTAL
Defective	4	10	6	20
Non-Defective	54	60	66	180
TOTAL	58	70	72	200

MEASURES of CENTRAL TENDENCY

(Arithmetic) Mean, Average, Expected Value: Ortalama için bu üç terim kullanılır.

Median: Ortanca → Sayıları küçükten büyüğe sıralayınca tam ortadaki sayı

Mode: En çok tekrar eden sayı (Moda)

Mode, mean veya median'dan farklı olarak olmayabilir, ya da birden fazla olabilir.

Population versus Sample

→ n = 50 (Mesela)

Random Sample: Rasgele 50 kişi seçtik. Bu değerleri kullanarak "Population" hakkında çıkarım yapacağız.

Population: Araştırmanın yapıldığı tüm birimler. Mesela, Bilkent University students. N: Population size, N=12000

X: Random Variable: Üzerine çalışılan değişken
Mesela, X: weekly food expenditure of a student at Bilkent.

	Population Parameters (Unknown Constants)	Sample statistics (Known Variables)
Mean	μ	$\bar{X}$
Variance	σ^2	s^2
Standard Deviation	σ	s

* Tüm öğrencilere bakabilseydik, sabit bir sayı bulurduk. Ama pratikte bunun için zaman ve paramız olmaz. Onun yerine, rasgele seçilen n kişiden yola çıkarak bir değer elde ediyoruz. Bu değer, "başka n kişi seçsek başka sayı olurdu" anlamında değişkendir.

Sample Mean: $\bar{X} = \frac{\sum X_i}{n} = \frac{\text{Sayıların toplamı}}{\text{Eleman sayısı}}$

Median: Ortanca; (i) Rank The Data

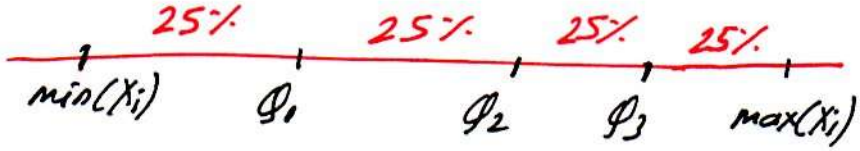
(ii) Find Position of Median = $\frac{n+1}{2}$

(iii) Find Median → Tam ortadaki sayı

Mode: En çok tekrar eden sayı.

Quartiles & 5-Number Summary

Quartiles: Data'yı 4'e bölen sayılar. Median datayı 2'ye böldüğü için "2nd Quartile" olur.

Data: 

(ii) Position of Quartiles;

$Q_1 \rightarrow (0,25) \cdot (n+1)$

$med = Q_2 \rightarrow (0,50) \cdot (n+1)$

$Q_3 \rightarrow (0,75) \cdot (n+1)$

5 Number Summary: $Min(X_i) < Q_1 < Q_2 < Q_3 < Max(X_i)$

In General; Position of p^{th} percentile is found as; $\left(\frac{p}{100}\right) \cdot (n+1)$

* Mean, Median, Mode, Quartiles and Percentiles are location statistics because they measure "where" the data is.

Alp Giray Özen | 0533 549 91 08 | alp@lecturemania.com | www.lecturemania.com

3.3 Ten economists were asked to predict the percentage growth in the Consumer Price Index over the next year. Their forecasts were

3.6	3.1	3.9	3.7	3.5
3.7	3.4	3.0	3.7	3.4

a. Compute the sample mean.
b. Compute the sample median.
c. What is the mode?

Ex

a) $\bar{X} = \frac{\sum X_i}{n} = \frac{3,6+3,1+\cdots+3,4}{10}$

$\bar{X} = \frac{35}{10} = 3,5$

b) (i) 3.0 3.1 3.4 3.4 3.5 3.6 3.7 3.7 3.7 3.9

(ii) $\frac{n+1}{2} = \frac{11}{2} = 5,5 \rightarrow$ 5. sayı ile 6. sayının ortalaması

(iii) Median $= \frac{3.5+3.6}{2} = 3.55$

c) Mode = En çok tekrar eden sayı = 3.7

3.1 A random sample of 5 weeks showed that a cruise agency received the following number of weekly specials to the Caribbean:

20	73	75	80	82

a. Compute the mean, median, and mode.
b. Which measure of central tendency best describes the data?

Ex

a) Mean $= \bar{X} = \frac{20+73+\cdots+82}{5}$

$\bar{X} = \frac{330}{5} = 66$

b) Data is already ranked. Median $= 3^{rd}$ # $= 75$

c) "Mean" uç değerlerden etkilenir. Data'daki 20 sayısı, diğer sayılardan çok farklı, dolayısıyla mean, sayıların çoğundan küçük. Uç değer olan data'da "median", "Central Tendency"yi daha iyi açıklar.

3.4 A department store chain randomly sampled 10 stores in a state. After a review of sales records, it was found that, compared with the same period last year, the following percentage increases in dollar sales had been achieved over the Christmas period this year:

10.2	3.1	5.9	7.0	3.7
2.9	6.8	7.3	8.2	4.3

a. Calculate the mean percentage increase in dollar sales.
b. Calculate the median.
c. Comment on symmetry.

Ex

a) $\bar{X} = \frac{10.2+3.1+\cdots+4.3}{10}$

$\bar{X} = \frac{59.4}{10} = 5.94$

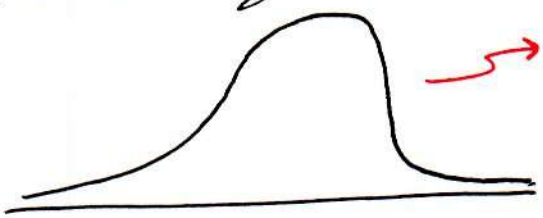
b) (i) 2.9 3.1 ~~3.7~~ 4.3 5.9 6.8 7.0 7.3 8.2 10.2

(ii) Pos. of median $= \frac{n+1}{2} = \frac{11}{2} = 5,5$

(iii) Median $= \frac{5.9+6.8}{2} = 6.35$

c) SYMMETRY:

* If the data is symmetric, then median ≃ Mean.



→ Can Eğrisi → Değerlerin çoğu ortada toplanmış, çok yüksek veya düşük değerler az.
(Etki Grades)

* Left (Negatively) Skewed Data.

mean < median

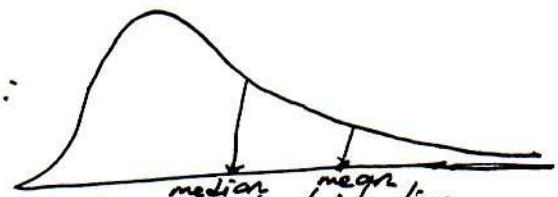
↳ our case

→ Değerlerin çoğu yüksek sayılardan oluşuyor.
Etki Hemoglobin level.

(Slightly Skewed to the left, çünkü mean ile median yakın)

* Right (Positively) Skewed Data:

mean > median



Etki income distribution.

Measures of Dispersion

* Range = $Max(x_i) - Min(x_i)$
* Interquartile Range (IQR) = $Q_3 - Q_1$
* Variance & Standard Deviation.

→ Population;

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$ ↳ Variance

$$\sigma = \sqrt{\sigma^2}$$ ↳ Std. Dev.

→ Sample;

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$ → Definition Formula

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

$$s^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}$$ → Calculation Formula

* Range, IQR, Variance & Standard Deviation are Dispersion statistics because they measure how wide the data is, independent of their location.

* Coefficient of Variation: When we compare two data whose means are not equal, standard deviation does not make sense. To compare the dispersion of such two data, we use $CoV = \frac{s}{\bar{x}} \times 100\%$.

3.15 The time (in seconds) that a random sample of employees took to complete a task is

Ex ⇒

23	35	14	37	28	45
12	40	27	13	26	25
37	20	29	49	40	13
27	16	40	20	13	66

a. Find the mean time.
b. Find the standard deviation.
c. Find the five-number summary.
d. Find the coefficient of variation.

X_i	X_i^2
23	23^2
35	35^2
14	14^2
⋮	⋮
+ 66	66
$\Sigma X_i = 695$	$\Sigma X_i^2 = 24361$

$n = 24$

a) $\bar{X} = \frac{\Sigma X_i}{n} = \frac{695}{24} = 28,96$

b) $s^2 = \frac{\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n}}{n-1} = \frac{24361 - \frac{695^2}{24}}{23} = 184,13$

$s = \sqrt{184,13} = 13,6$

c) To rank the data, stem & leaf display is useful;

$\min(X_i) = 12 \quad \max(X_i) = 66$

1	4 2 3 3 6 3
2	3 8 7 6 5 0 9 7 0
3	5 7 7
4	5 0 9 0 0
5	
6	6

⇒

1	2 3 3 3 4 6
2	0 0 3 5 6 7 7 8 9
3	5 7 7
4	5 0 9 0 0
5	
6	6

Pos. of $Q_1 = 0,25 \cdot (24+1) \cong 6$ → $Q_1 = 16$

Pos. of $Q_2 = 0,5 \cdot 25 = 12,5$ → $Q_2 = \frac{26+27}{2} = 26,5$

Pos. of $Q_3 = 0,75 \cdot 25 \cong 19$ → $Q_3 = 45$

5 Number Summary : $12 < 16 < 26.5 < 45 < 66$

d) $CoV = \frac{s}{\bar{X}} \times 100 = \frac{13,6}{28,96} \times 100 = 47\%$

3.16 The following stem-and-leaf display contains sample data:

Ex

Stem unit:

3	01
4	588
5	0345789
6	1479
7	369
8	037

a. Calculate *IQR*.
b. Find the 8th decile.
c. Find the 92nd percentile.

a) $n = 22$

Pos. of $Q_1 = 0{,}25 \cdot 23 \cong 6$

$Q_1 = 50$

Pos of $Q_3 = 0{,}75 \cdot 23 \cong 17$

$Q_3 = 73$

$IQR = 73 - 50 = 23$

b) Pos. of $D_8 = 0{,}80 \cdot 23 \cong 18{,}5$

$D_8 = \frac{76+79}{2} = 77.5$

c) Pos. of $P_{92} = 0{,}92 \cdot 23 \cong 21$

$P_{92} = 83$

3.20 The annual percentage returns on common stocks over a 7-year period were as follows:

Ex

4.0% 14.3% 19.0% –14.7% –26.5% 37.2% 23.8%

Over the same period the annual percentage returns on U.S. Treasury bills were as follows:

6.5% 4.4% 3.8% 6.9% 8.0% 5.8% 5.1%

a. Compare the means of these two population distributions.
b. Compare the standard deviations of these two population distributions.

Let: X_i : Common stocks; $n_x = 7$

y_i : Treasury bills; $n_y = 7$

X_i	X_i^2	y_i	y_i^2
4.0	4.0^2	6,5	$6{,}5^2$
14,3	$14{,}3^2$	4,4	$4{,}4^2$
⋮	⋮	⋮	⋮
+ 23,8	$23{,}8^2$	5,1	$5{,}1^2$
$\Sigma X_i = 57{,}1$	$\Sigma X_i^2 = 3458{,}3$	$\Sigma y_i = 40{,}5$	$\Sigma y_i^2 = 247{,}3$

a) $\mu_x = \frac{57{,}1}{7} = 8{,}16$ $\mu_y = \frac{40{,}5}{7} = 5{,}79$

Common stocks are more profitable on the average

b) $\sigma_x^2 = \frac{3458{,}3 - 57{,}1^2/7}{7} = 427{,}5$ $\sigma_y^2 = \frac{247{,}3 - 40{,}5^2/7}{7} = 1{,}85$

$\sigma_x = \sqrt{427{,}5} = 20{,}7$ $\sigma_y = \sqrt{1{,}85} = 1{,}36$ → Treasury bills are less risky.

Chebychev's Theorem;

For any population; AT LEAST $\left(1-\frac{1}{k^2}\right)\times 100\%$ of the data lies within the interval $\mu \pm k.\sigma$

Empirical Rule;

If the data are mounded (bell-shaped) and symmetric, APPROXIMATELY;

$\mu \pm 1.\sigma$ interval contains 68%,

$\mu \pm 2\sigma$ interval contains 95%,

$\mu \pm 3\sigma$ interval contains 99,7% of the data

(we will see that these approximation follows Normal Distribution.)

3.17 A random sample of data has a mean of 75 and a variance of 25.

Ex

a. Use Chebychev's theorem to determine the percent of observations between 65 and 85.
b. If the data are mounded, use the empirical rule to find the approximate percent of observations between 65 and 85.

$\mu = 75 \quad \sigma^2 = 25$

$\sigma = 5$

interval (65 ; 85) is;

(75 - 2.5 ; 75 + 2.5)

$\Rightarrow \mu \pm 2\sigma$

a) At least $\left(1-\frac{1}{2^2}\right).100 = 75\%$ of the data is in the interval

b) Approximately 95% of the data is in the interval

c) Find an interval that at least 60% of the data lies.

$1-\frac{1}{k^2} = 0,60 \Rightarrow \frac{1}{k^2} = 0,60 \Rightarrow k^2 = 2,5 \Rightarrow k = 1,58$

Interval is; $75 \pm 1,58.5 \Rightarrow (67,1 ; 82,9)$

Weighted mean & Variance

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i} = \frac{w_1 x_1 + w_2 x_2 + \dots + w_n x_n}{n} \quad \hookrightarrow n = \sum w_i$$

→ Herbir x_i'ye ağırlığına göre önem veriyoruz, ya da w_1 tane x_1, w_2 tane x_2 ... var gibi...

$$s^2 = \frac{\sum w_i x_i^2 - \frac{(\sum w_i x_i)^2}{n}}{n-1}$$

Approximate Mean and Variance for Grouped Data (or Frequency Dist.)

$$\bar{x} = \frac{\sum f_i m_i}{n} \qquad s^2 = \frac{\sum f_i m_i^2 - \frac{(\sum f_i m_i)^2}{n}}{n-1}$$

where $n = \sum f_i$ and m_i is midpoint of class i.

(herbir classın tüm elemanlarını midpointine eşit varsayıyoruz. Bu tahmin (ya da yaklaşım) $\bar{x}$ için yerindedir ama s^2 olduğundan yüksek tahmin edilebilir.)

3.30 A random sample of 50 personal property insurance policies showed the following number of claims over the past two years.

Number of claims	0	1	2	3	4	5	6
Number of policies	21	13	5	4	2	3	2

a. Find the mean number of claims per day.
b. Find the sample variance and standard deviation.

Ex

x_i	w_i	$w_i \cdot x_i$	x_i^2	$w_i \cdot x_i^2$
0	21	0	0	0
1	13	13	1	13
2	5	10	4	20
3	4	12	9	36
4	2	8	16	32
5	3	15	25	75
+ 6	2	12	36	72

$n = \sum w_i = 50 \quad \sum w_i x_i = 70 \quad \sum w_i x_i^2 = 248$

$$\bar{x} = \frac{70}{50} = 1{,}4$$

$$s^2 = \frac{248 - \frac{70^2}{50}}{49} = 3{,}06 \; ; \; s = \sqrt{3{,}06} = 1{,}75$$

3.32 A sample of 20 financial analysts was asked to provide forecasts of earnings per share of a corporation for next year. The results are summarized in the following table:

Forecast ($ per share)	9.95 < 10.45	10.45 < 10.95	10.95 < 11.45	11.45 < 11.95	11.95 < 12.45
Number of analysts	2	8	6	3	1

a. Estimate the sample mean forecast.
b. Estimate the sample standard deviation.

Class	f_i	m_i	$f_i \cdot m_i$	m_i^2	$f_i \cdot m_i^2$
9.95 ⊢ 10.45	2	10.2	20,4	104,4	208,8
10.45 ⊢ 10.95	8	10,7	85,6	114,5	916,0
10.95 ⊢ 11.45	6	11,2	67,2	125,4	752,4
11.45 ⊢ 11.95	3	11,7	35,1	136,9	410,7
11.95 ⊢ 12.45	1	12,2	12,2	168,9	168,9
+					
	$n = \Sigma f_i = 20$		$\Sigma f_i m_i = 220,5$		$\Sigma f_i m_i^2 = 2436,8$

$$\bar{X} = \frac{220,5}{20} = 11,03 \qquad s^2 = \frac{2436,8 - 220,5^2/20}{19} = 0,305$$

$$s = \sqrt{0,305} = 0,55$$

Linear Relationship between two VARIABLES

Covariance & Correlation Coefficient

	Population	Sample
Covariance	σ_{xy}	s_{xy}
Correlation coefficient	ρ	r

İki değişken arasındaki linear ilişkiyi ölçüyoruz. Değişkenlerden birisi artıyorken diğeri de "sabit bir eğim ile" artıyorsa pozitif linear ilişki, azalıyorsa negatif linear ilişki olur.

Standard sapma her zaman pozitifdi: Buna göre, negatif linear ilişki için "Covariance" negatif olur.

* $Cov(x,y) = \sigma_{xy} = \dfrac{\sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y)}{N}$ → definition formula

$\widehat{Cov(x,y)} = s_{xy} = \dfrac{\sum_{i=1}^{n} xy - \frac{\Sigma x \cdot \Sigma y}{n}}{n-1}$ → calculation formula

$-\infty < s_{xy} < \infty$

* $\rho = \dfrac{Cov(x,y)}{\sigma_x \sigma_y}$; $r = \dfrac{\widehat{Cov(x,y)}}{s_x \cdot s_y}$

$-1 \leq r \leq 1$

* Procedure to calculate r

(i) Σx, Σy, Σx^2, Σy^2, $\Sigma x.y$

(ii) $s_x^2 = \dfrac{\Sigma x^2 - (\Sigma x)^2/n}{n-1}$; $s_y^2 = \dfrac{\Sigma y^2 - (\Sigma y)^2/n}{n-1}$

$s_{xy} = \dfrac{\Sigma xy - \frac{\Sigma x \cdot \Sigma y}{n}}{n-1}$

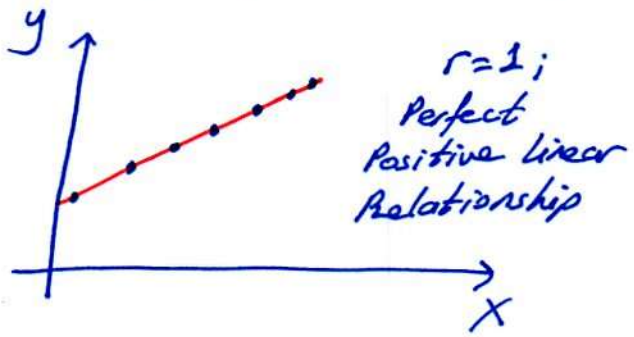
(iii) $r = \dfrac{s_{xy}}{\sqrt{s_x^2 \cdot s_y^2}} = \dfrac{s_{xy}}{s_x \cdot s_y}$

OR; Note that; (Çünkü n-ler sadeleşir.)

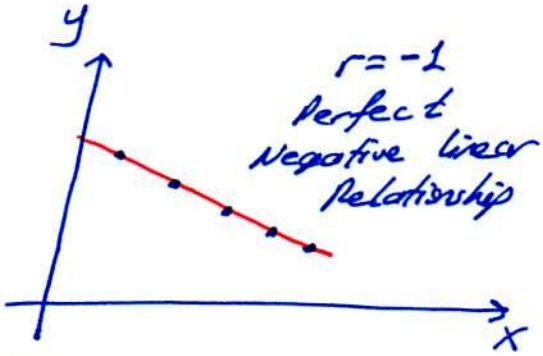
$$r = \frac{\Sigma xy - \frac{\Sigma x \Sigma y}{n}}{\sqrt{\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right) \cdot \left(\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right)}}$$

Alp Giray Özen | 0533 549 91 08 | alp@lecturemania.com | www.lecturemania.com

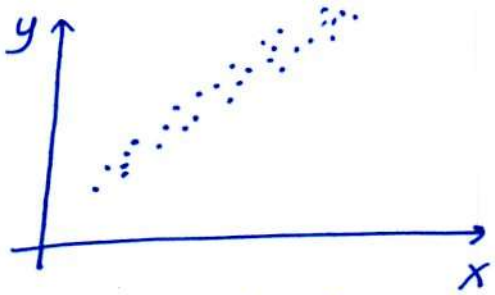
Interpretation of r



r = 1; Perfect Positive Linear Relationship

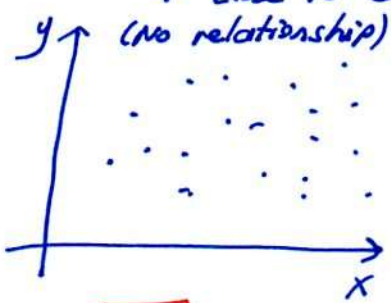


r = -1 Perfect Negative Linear Relationship

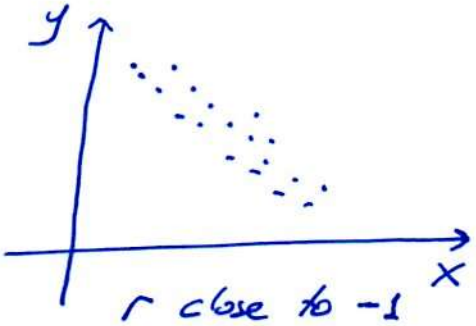


r close to 1 (Ex: r = 0,893)

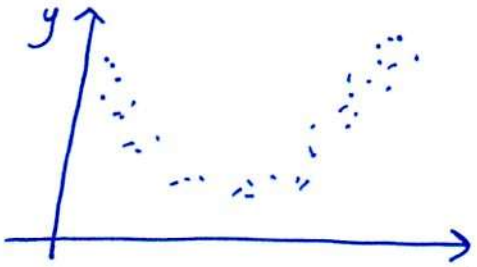
r close to 0 (No relationship)



OR



r close to -1 (Ex: r = -0,769)



r close to 0 (There exist relationship but NOT linear)

Ex 7

3.37 Following is a random sample of price per piece of plywood, X, and quantity sold, Y (in thousands):

Price per Piece (X)	Thousands of Pieces Sold (Y)
$6	80
7	60
8	70
9	40
10	0

a. Compute the covariance.
b. Compute the correlation coefficient.

X	y	X^2	y^2	X.y
6	80	36	6400	480
7	60	49	3600	420
8	70	64	4900	560
9	40	81	1600	360
+ 10	0	100	0	0

$n = 5$

$\Sigma x_i = 40$ $\Sigma y_i = 250$ $\Sigma y_i^2 = 16500$ $\Sigma x_i y_i = 1820$

$\Sigma x_i^2 = 330$

$$s_x^2 = \frac{330 - \frac{40^2}{5}}{4} = 2{,}5 \qquad s_y^2 = \frac{16500 - \frac{250^2}{5}}{4} = 1000$$

a) $$s_{xy} = \frac{1820 - \frac{40 \cdot 250}{5}}{4} = -45$$

b) $$r = \frac{-45}{\sqrt{2{,}5 \cdot 1000}} = -0{,}9$$

Negative linear relationship between Price & Sales (as might be expected)

Alp Giray Özen | 0533 549 91 08 | alp@lecturemania.com | www.lecturemania.com