Abe Mirza **Topics Review** **Statistics**

Part I

To understand God's thoughts,
we must study statistics, for these are the measure of His purpose.

– **Florence Nightingale**

Topics Page

General Introduction

The Purpose of statistics: Statistics has many uses, but perhaps its most important purpose is to help us make decisions about issues that involve uncertainty.

Definition of Statistics:

1. Numerical Facts

1. Average price for one bedroom apartment at the city of Rocklin is $ 895.
2. 80% of Sierra students graduate in 2 years.

2. C O D A **C**ollection, **O**rganization, **D**escription, **A**nalysis and interpretation of data.

Collection: **D**ata **S**ampling

Organization: **F**requency Table (Bar-chart, Pie-chart),
Histogram, **F**requency **P**olygon, **O**give **C**urve

Description: **M**ean, **M**ode, **M**edian,
Range, **V**ariance, **S**tandard **D**eviation **SD**,
Quartiles, **P**ercentiles, **B**ox Plot

Analysis: **C**orrelation and **R**egression, **E**stimation,
Test of **H**ypothesis, Analysis of **V**ariance

Type of Statistics:

Descriptive: **C**ollection, **O**rganization, **D**escription

Inferential: **A**nalysis and interpretation of data

What is the statistics all about?

1. It is about how we test if a new drug is effective in treating cancer.
2. It is about opinion polls, pre-election polls, and exit polls.
3. It is about sports, where we rank players and teams primarily through their statistics.
4. It is about the market research and the effectiveness of advertising
5. It is about how agricultural inspectors ensure the safety of the food supply.

Population versus **Sample**:

Population: Entire elements or subjects under study that share one or more **common characteristic** such as age, gender, major or race. (Keyword all/every), All college students, All Sierra College students, All male Sierra College Students who are taking statistics and majoring in business. Two Elements: **T**ime and **P**lace

Sample: A portion of population.

Census: The collection of data from every element in a population.

Parameter vs. statistic: A numerical measurement describing some characteristic of a
Population vs. a **S**ample (Greek Alphabet vs. lower case English)
μ = avg. σ (sigma) = st. dev χ^2 = Chi-squared…. $\overline{X}$, s, r

HW: *Answer questions* ***A*** *from* ***page 2*** *of practice problem* ***part 1***

Type of Data:

Qualitative (Names, Labels ...) pass / fail, democrat/republican/independent, yes/no, grades (A,B,C,D,F)

Quantitative: **1**. **D**iscrete (Countable): number of accidents in Rocklin each day, number of emergency call to 911 center each day, number of students that will pass Abe stat class

2. **C**ontinuous (Measurable): Speed, weight, time, capacity, length, volume, area

HW: *Answer questions* ***B*** *from* ***page 2*** *of practice problem* ***part 1***

Types of Sampling: R_S_S_C_C

1. **R**andom: Every member of population has equal chance to be selected.
How? Every member will be assigned a different number, and we select random numbers by a computer or a table and match those with the members' numbers.

2. Systematic: We select some starting point and then select every kth (such as every 20^{th}) member in the population.
How? Every 10^{th} customer or client will be selected to be asked questions.

3. Stratified: **Subdivide** the population into at least two different subgroups (strata) sharing the same characteristics (such as gender or age bracket), then we draw a sample from each stratum.
How? a)divide the police officers in Sacramento into male and female group
b)select a random sample of each and collect data regarding the years in service.

4. Cluster: Divide the population into sections (or clusters), and then randomly select some of those clusters, then choose all the members from those selected clusters.
How? To see the customer feedback to a new menu
a) divide Sacramento in different zones,
b) randomly select some of those zones
c) collect data from **all** fast-food branches in those selected zones.

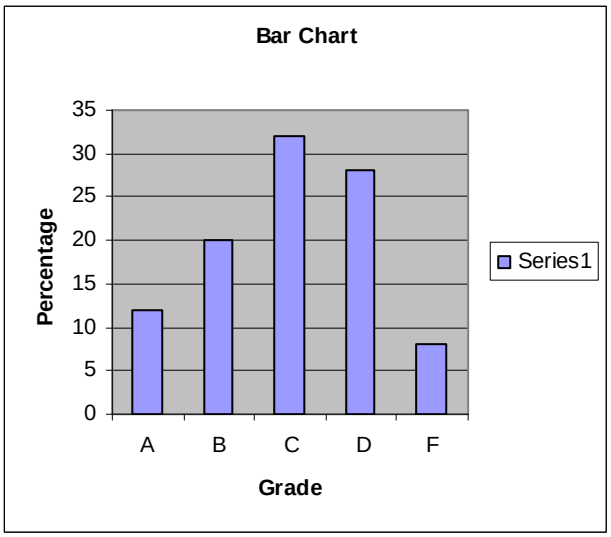
5.**C**onvinence: Use the results are readily available.
How? A math instructor ask some of his students if they use student solution manual to do their homework.

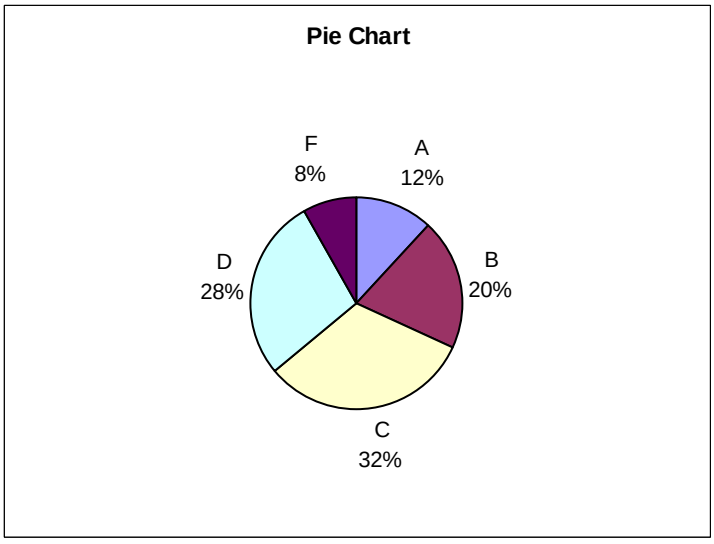
HW: *Answer questions* ***C*** *from* ***page 2*** *of practice problem* ***part 1***

Qualitative Data

Example 1.

Grade	f = **Students**	**Rel. freq %** $\frac{f}{n}\times 100$	**Angles** 360° **(Rel. freq)**
A	**6**	$(6/50)\times 100 = \mathbf{12}$	$.12\times 360 = \mathbf{43.2}^\circ$
B	**10**	$(10/50)\times 100 = \mathbf{20}$	$.20\times 360 = \mathbf{72}^\circ$
C	**16**	$(16/50)\times 100 = \mathbf{32}$	$.32\times 360 = \mathbf{115.2}^\circ$
D	**14**	$(14/50)\times 100 = \mathbf{28}$	$.28\times 360 = \mathbf{100.8}^\circ$
F	**4**	$(4/50)\times 100 = \mathbf{8}$ +	$.8\times 360 = \mathbf{28.8}^\circ$ +
	$n = \sum f = 50$	100?	360° ?





Practice 1:

Grade	f = **Students**	**Rel. freq %** $\frac{f}{n}\times 100$	**Angles** 360° **(Rel. freq)**
A	**22**		
B	**26**		
C	**20**		
D	**8**		
F	**4**	+	+
	$n = \sum f =$	100?	360° **?**

Complete the table and draw the bar chart and the pie chart.

Descriptive Statistics

A) Measure of Central Tendency (**M**ean, **M**edian, **M**ode)

Mean (μ , $\bar{x}$) $x = data$ $\sum = Sum$ N or n = Number of data points

Data: 5, 6, 3 , 9, 7 $\bar{x} = \frac{\sum x}{n} = \frac{5+6+3+9+7}{5} = \frac{30}{5} = 6$

Median: The middle data point in a ranked (largest to smallest or smallest to largest) data, **or**
The median cuts the ranked data in half **one half below** it and **one half above** it.

Example1: Suppose the median score for the first test was 73, it simply means half the class got below 73 and the other half above it.

How to find it?

2, 5, 11, 16, 8, 9, 3, 7, 5 Ranked 2, 3, 5, 5, **7**, 8, 9, 11, 16, **Median** = **7**

2, 3, 5, 5, **7**, 8, 9, 11, 16, 4 Ranked 2, 3, 4, 5, **5**, **7**, 8, 9, 11, 16, **Median** = $\frac{5+7}{2} = \mathbf{6}$

Hint: If there are extreme values in data set (too large or too low with respect of the rest of data) then median is a better than mean to identify the measure of central tendency.

Mode: The data value(s) with the highest occurrence, bimodal, multimodal

2, **8**, 11, 7, **8**, 13 **Mode** = 8

3, **12**, 5, 14, 9, **12**, **7**, 16, **7** **Bimodal** = 7, 12

11, **15**, 7, **2**, 6, 16, **15**, 3, **2**, **11**, 19, 5, 4 **Multimodal** = 2, 11, 15

HW: *Answer questions on columns* ***A-G*** *from* ***page1*** *of practice problem* ***part 1*** ***All the answers are on p.18***

B) Measure of Variation (**R**ange, **S**tandard Deviation, **V**ariance)

Range: It shows how far apart the data points are? **Range** = the highest value - the smallest value

Standard Deviation (σ, s) : It measures the **average dispersion** of data **around the mean.**

Example: Consider the 3 random delivery time (in days) taken by 2 different companies A, and B

	A	**B**
Mean	5	5
Median	5	5
Mode	5	none

At first it seems there are not that much of difference between the delivery times of these two companies but now let's look at their actual data.

	A	B
Delivery time	5	5
Delivery time	5	0
Delivery time	5	10

Dot Plot

Now, it seems that there is **no dispersion** for company A, but an **average dispersion of 5** for company B.

The formula for the Standard Deviation or average dispersion of data around mean $s=\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$

Company A

x	$\bar{x}$	$(x-\bar{x})$	$(x-\bar{x})^2$
5	5	0	0
5	5	0	0
5	5	0	0
			$\sum(x-\bar{x})^2=0$

$$s=\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}=\sqrt{\frac{0}{3-1}}=\sqrt{0}=0$$

Company B

x	$\bar{x}$	$(x-\bar{x})$	$(x-\bar{x})^2$
5	5	0	0
0	5	-5	25
10	5	5	25
			$\sum(x-\bar{x})^2=50$

$$s=\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}=\sqrt{\frac{50}{3-1}}=\sqrt{25}=5$$

Find the mean and standard deviation for 5, 6, 3 , 9, 10, 3, and also draw the **dot-plot**.

x	$\bar{x}=\frac{\sum x}{n}=$	$(x-\bar{x})$	$(x-\bar{x})^2$
5			
6			
3			
9			
10			
3			
$\sum x=$			$\sum(x-\bar{x})^2=$

$s=\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}=\sqrt{\quad}=\sqrt{\quad}=2.97$ **Variance** = $s^2=8.8$

Variance (σ^2 , s^2): Variance is the **square of standard deviation.**

Rule of thumb to **estimate s**: $S=\frac{Range}{4}$

Generally the larger the data set the closer the estimate will be to the exact value.

HW: *Answer questions on columns* ***A-G*** *from* ***page1*** *of practice problem* ***part 1*** ***All the answers are on p.18***

C) Measure of Positions (**Q**uartiles, **P**ercentile, **B**ox-Plot, **Z**-score)

Quartiles: Quartiles breaks the **ranked data** in 3 quartiles (**Q1, Q2, Q3**)

Data: _____25%_____**Q1**_____ 25%_____**Q2**_____25%_____**Q3**_____25%_____

How to find quartiles?

1. Rank the data
2. Find Q2 = Median
3. Find the new medians Q1, Q3 on either side of Q2.

Example 1: Data: 2, 5, 11, 16, 8, 9, 3, 7, 5, 4, 13 Odd number of data

Ranked Data: 2, 3, ,**4**, 5, 5, **7**, 8, 9, **11**, 13, 16,

Q1 Q2 Q3

Example 2: Data: 2, 3, 5, 5, 7, 8, 9, 11, 16, 4 Even number of data

Ranked Data 2, 3, **4**, 5, 5, 7, 8, **9**, 11, 16, **Q2 = Median** = $\frac{5+7}{2} = \mathbf{6}$

Q1 **Q2** = 6 **Q3**

Box-Plot: is mainly used for ungrouped data to show how the data are distributed by showing **center**, **spread**, and **skewness**. **Center** is the **Q2**, **Spread** is how wide the box is, **Skewness** explains the distribution of the data

To construct a box-plot

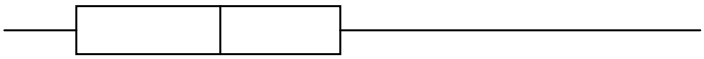
1. Find the **5-number summery** of the data that are **Min, Q1, Q2, Q3, Max**

2. Plot these points on a **scaled** number line.

3. Construct a box by using Q1, Q2, Q3

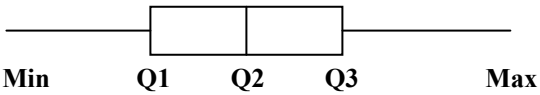
There are many possibilities of where the box in box-plot may be located.

If the box in box-plot is located to the **far Left**, it suggests that distribution of data are **skewed** to the **Right**



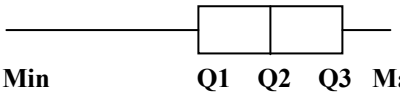
Skewed to the **Right**

If the box in box-plot is located to the **Center**, it suggests that distribution of data are **Centered**.



Centered

If the box in box-plot is located to the **far Right**, it suggests that distribution of data are **skewed** to the **Left**



Skewed to the **Left**

HW: *Answer questions on columns* ***A-G*** *from* ***page1*** *of practice problem* ***part 1*** ***All the answers are on p.18***

Empirical Rules: If and only if the **box-plot or histogram is centered** then we can apply the **three** following empirical rules.

$99.7\% = \bar{x} \pm 3\,s$ **99.7 %** of data are within $3\,s$ of the mean ($\bar{x}$)

$95\% = \bar{x} \pm 2\,s$ **95 %** of data are within $2\,s$ of the mean ($\bar{x}$)

$68\% = \bar{x} \pm s$ **68 %** of data are within $1\,s$ of the mean ($\bar{x}$)

Example: Find all three empirical rules for Abe Stat class if the average was 72 and the standard deviation was 8, assuming that Box-plot was centered.

$99.7\% = 72 \pm 3(8) = 72 \pm 24$ $48 <$ **99.7 %** of class got scores < 96

$95\% = 72 \pm 2(8) = 72 \pm 16$ $56 <$ **95 %** of class got scores < 88

$68\% = 72 \pm 1(8) = 72 \pm 8$ $64 <$ **68 %** of class got scores < 80

HW: *Answer questions* ***C*** *from* ***page1*** *of practice problem* ***part 1*** ***All the answers are on p.18***

Z-score: is used to show the relative position of a data points with respect of the rest of data by measuring how many standard deviation the point is away from the mean. To apply the z-score the box-plot or histogram must be centered.

$$Z = \frac{x - \bar{x}}{s} \quad \textbf{or} \quad Z = \frac{x - \mu}{\sigma}$$

The possible range of Z-values;

---------- -2 ----------0---------- 2 ----------

Unusual Values: $Z < -2$ **Ordinary Values:** $-2 \le Z \le 2$ **Unusual Values:** $Z > 2$

Example 1: Find the z-score of final exam for Tommy Yank in stat class at CSUS, if his score was 87, when the class average was 72 and the standard deviation was 8**.**

$$Z = \frac{x - \mu}{\sigma} = \frac{87 - 72}{8} = \frac{15}{8} = 1.875$$ **Ordinary Or Unusual Value?**

So, he does relatively an ordinary performance relative to the rest of his class.

Example 2: Find the z-score of final exam for Marcy Tank in stat class at UC Davis, if his score was 82, when the class average was 71 and the standard deviation was 4**.**

$$Z = \frac{x - \mu}{\sigma} = \frac{82 - 71}{4} = \frac{11}{4} = 2.75$$ **Ordinary Or Unusual Value?**

So, she does relatively better than the rest of her class.

HW: *Answer questions* ***D*** *from page* ***3*** *of practice problem* ***part 1***

Grouped Data (Freq. Table)

X-axis		**Frequency Polygon**	**Histogram** Y-axis	**Mean**	**St. Dev.**
Quiz Score	**Freq(f)= Students**	m **midpoint**	**Rel. freq %** $\frac{f}{n}\times 100$	$f \times m$	$f \times m^2$
0 – 4	6	2	12%	12	24
4 - 8	10	6	20	60	360
8 – 12	16	10	32	160	1600
12 – 16	14	14	28	196	2744
16 – 20	4	18	8 +	72	1296
	$\sum f = n = 50$			$\sum f \times m = 500$	$\sum f \times m^2 = 6024$

5. Mean: $\bar{X} = \frac{\sum f \times m}{n} = \frac{500}{50} = 10$

6. Standard deviation: $s = \sqrt{\frac{n\sum f \times m^2 - \left(\sum f \times m\right)^2}{n(n-1)}} = \sqrt{\frac{50(6024) - (500)^2}{50(50-1)}} = \sqrt{\frac{51200}{2450}} = 4.57$

7. Variance: $s^2 = 4.57^2 = 20.9 =$

Practice:

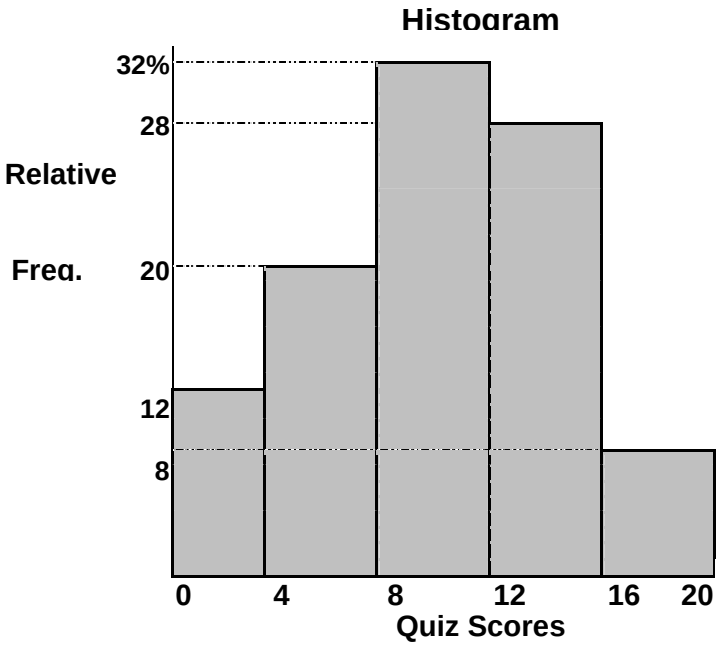
X-axis		**Frequency Polygon**	**Histogram** Y-axis	**Mean**	**St. Dev.**
Quiz Score	**Freq(f)=**	m	**Rel. freq %**	$f \times m$	$f \times m^2$
0 – 10	8		20%		
10 - 20	12			180	
20 – 30	14	25			
30 – 40	6		+		7350
	$\sum f = n =$		**Do they add to 100%?**	$\sum f \times m =$ **780**	$\sum f \times m^2 =$ **19000**

5. Mean: $\bar{X} = \frac{\sum f \times m}{n} =$ _____ = 19.5

6. Standard deviation: $s = \sqrt{\frac{n\sum f \times m^2 - \left(\sum f \times m\right)^2}{n(n-1)}} =$ $\sqrt{\text{_____}}$ = $\sqrt{\text{_____}}$ = 9.86

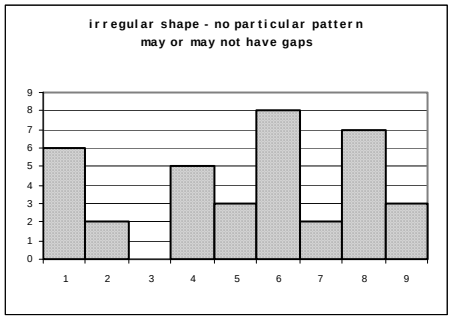
7. Variance: $s^2 = 9.8^2 =$ 97.18

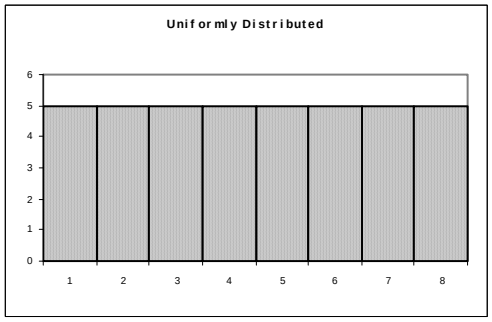
HW: *Answer questions* ***A, B, C, D*** *from* ***pages 4,5*** *of practice problem* ***part 1*** ***All the answers are on p.10-13***

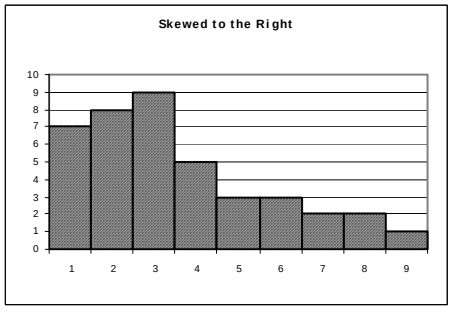


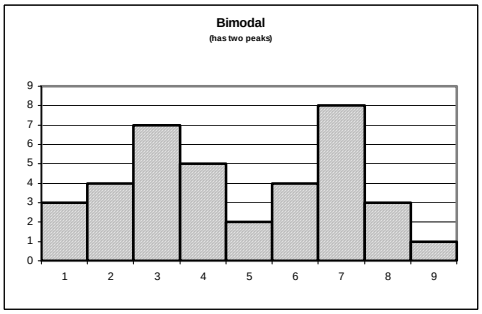
Histogram looks close to a Centered or bell-shaped distribution.

Different possible shapes of Histogram









Regression

Correlation and Regression is the study of the **relationship** between **two variables** (x , y) with the following objectives:

1. To find the **nature of the relationship** (Linear or non-linear, positive or negative relationship) by doing a scattered diagram, y versus x

2. To measure the **strength of this relationship** by computing the correlation coefficient $= r$

3. Finding **slope** and **y-intercept** for equation of the best fitted- line (**regression equation** = $y = a\,x + b$) between x , y variables.

4. Using the regression equation to **estimate or predict** one variable from the other.

Nature of relationship:
Positive: Both variables either increasing or decreasing $x \uparrow\uparrow y$ **or** $x \downarrow\downarrow y$
Negative: When one variable increases the other one decreases or vice versa. $x \uparrow\downarrow y$ **or** $x \downarrow\uparrow y$

Determine the nature of relationship between x and y variables.

	x	y	Nature of relationship Positive, Negative
1	Average number of hours per week to study for stat class	Stat test score	+ , − , *None*
2	Mortgage rate	Number of loans refinanced	+ , − , *None*
3	Average height of the parents	Height of the sons or daughters	+ , − , *None*
4	No. of absences in a semester for stat class	Stat test scores	+ , − , *None*
5	Daily temperature in summer	Water or electric consumption	+ , − , *None*
6	$ amount spent on advertisement	Monthly sales	+ , − , *None*
7	Fat consumption	Cholesterol level	+ , − , *None*
8	Number of years of education	Monthly salary	+ , − , *None*
9	Number of hours watching TV/week	GPA	+ , − , *None*
10	Ice cream sales	Number of drownings	+ , − , *None*
11			
12			

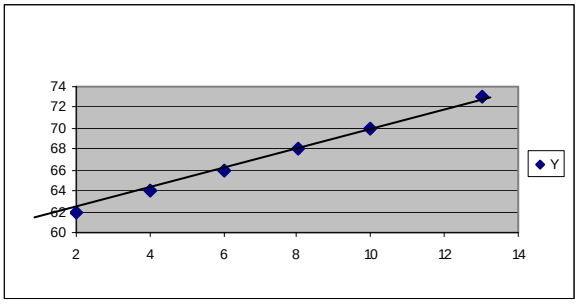
Why do we need to do scattered diagram?

a) To see if data exhibit a **linear pattern** or not b) To see if linear pattern is **positive or negative**

c) To see how closely (**strongly**) the data are **clustered around** the **mean**

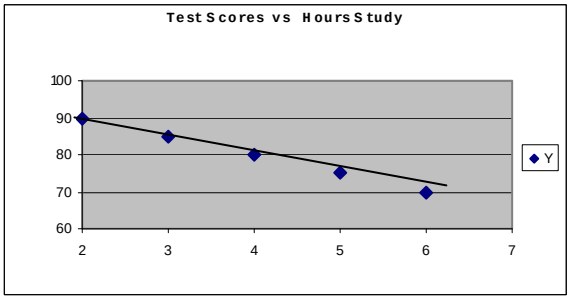
d) To detect any **outlier** (a point that is lying far away from the other data points).

Different Possible shapes of a Scattered Diagram



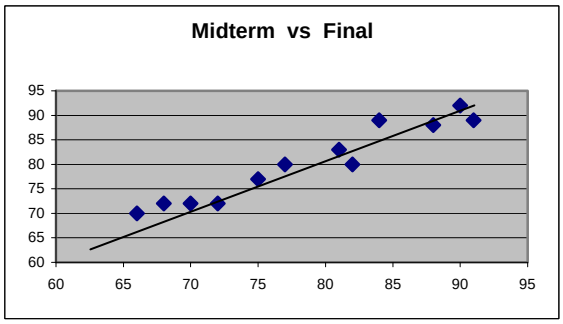
$r = 1$

Perfect Positive Linear Correlation

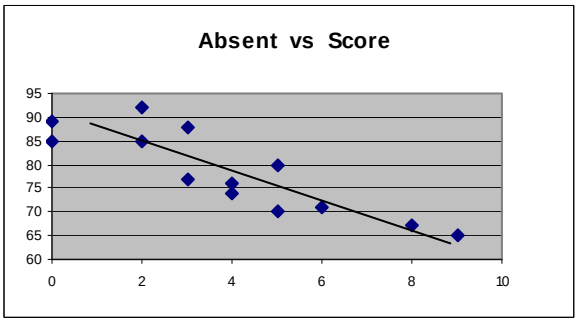


$r = -1$

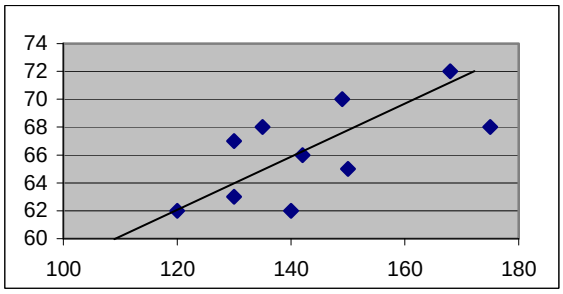
Perfect Negative Linear Correlation



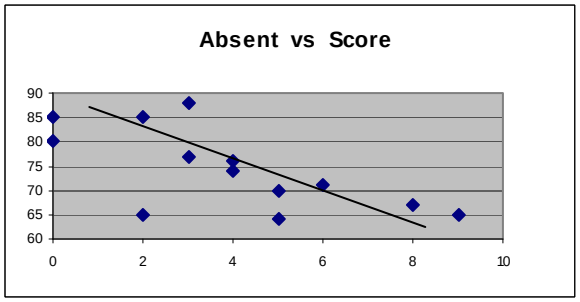
Strong Positive Linear Correlation



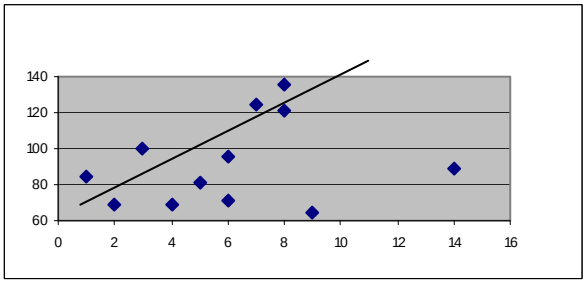
Strong Negative Linear Correlation



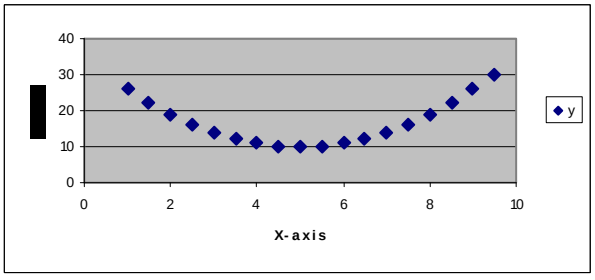
Positive Linear Correlation



Negative Linear Correlation



No Correlation

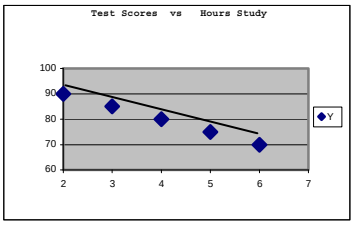


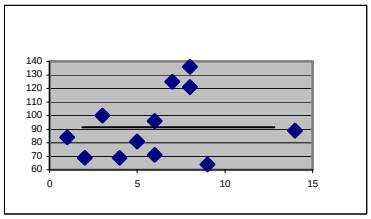
Non linear relationship

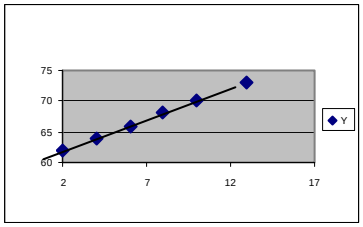
Steps to do a Correlation and Regression problem

1. Constructing a Scattered Diagram and comment on its nature (linear or non-linear, positive or negative, strong or weak relationship)

2. Computing $\boldsymbol{r}$ = Correlation Coefficient and comment on its strength $-1 \le r \le 1$







$-1 \le r \le 1$

-1 ——————————— +1

0

Perfect Negative Linear Correlation | **No Correlation** | **Perfect Positive Linear Correlation**

3. Computing $\overline{x}, \overline{y}, s_x, s_y$

4. Computing Slope ($\boldsymbol{a}$) and y-intercepts ($\boldsymbol{b}$) for the regression equation $\boldsymbol{y = a\,x + b}$

5. Using the regression equation $\boldsymbol{y = a\,x + b}$ to **estimate or predict** one variable from the other.
Estimated values are labeled as y' (y -prime) and x' (x -prime).

Guideline for using the regression line:

1. If there is no significant linear correlation, do not use the regression equation.

2. When using the regression equation for prediction, **stay** within the range of the available sample data.

3. A Regression equation based on old data is not necessarily valid now.

Marginal Change (Slope): in a variable is the amount that it changes in y-variable when the x-variable increases by one unit.

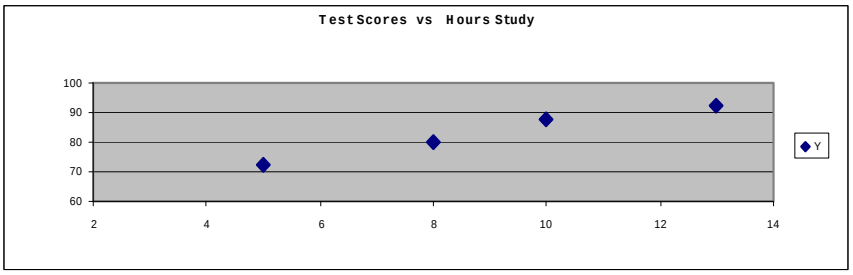
Outlier: is a point that is lying far away from the other data points.

Coefficient of determination $= r^2 \times \% = \dfrac{\text{explained variation}}{\text{total variation}} =$ is the amount of variation in y that is explained by the regression line

Example 1.

	x = Hours Study/week	y = Test Score	x^2	y^2	xy
1	5	72	25	5184	360
2	10	88	100	7764	880
3	13	92	169	8464	1196
4	8	80	64	6400	640
	$\Sigma x = \mathbf{36}$	$\Sigma y = \mathbf{332}$	$\Sigma x^2 = \mathbf{358}$	$\Sigma y^2 = \mathbf{27792}$	$\Sigma xy = \mathbf{3076}$

1. Use the data and plot the data as a scattered diagram and **comment** on the pattern of the points.



Strong Positive

Linear Correlation

2. Compute the correlation coefficient and **comment** on that: ***a very strong positive linear correlation.***

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2}\sqrt{n\sum y^2 - (\sum y)^2}} = \frac{(4)(3076) - (36)(332)}{\sqrt{4(358) - (36)^2}\sqrt{4(27792) - (332)^2}} = \frac{12304 - 11952}{\sqrt{136}\sqrt{944}} = \frac{352}{358.307} = 0.9824$$

3. Compute the slope and y-intercept and write the equation of regression line.

$$\textbf{\textit{Slope}} = \boldsymbol{a} = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} = \frac{4(3076) - (36)(332)}{4(358) - (36)^2} = \frac{12304 - 11952}{1432 - 1296} = \frac{352}{136} = \mathbf{2.588 = 2.59}$$

$$\textbf{\textit{y - itc}} = \boldsymbol{b} = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2} = \frac{(332)(358) - (36)(3076)}{4(358) - (36)^2} = \frac{118856 - 110736}{1432 - 1296} = \frac{8120}{136} = \mathbf{59.71}$$

$$\boldsymbol{y = a\,x + b = 2.59\,x + 59.71}$$

4. Explain the slope based on the regression equation and the in relation of x and y variables.

In general for every additional hour of study per week the score goes up by 2.59 points.

5. Compute average and standard deviation for both x and y variables.

$$\bar{x} = 36/4 = \mathbf{9}\ \ hrs \quad s_x = 3.37 \qquad \bar{y} = 332/4 = \mathbf{83} \quad s_y = 8.87$$

6. If one student studies 10 hours a week, use **Reg. Equ.** to estimate her test score. $x = 10\ hrs, \quad y' = 85.61$

$$x = 10\ hrs, \quad y' = 85.61$$

7. If one student has test score of 90, use **Reg. Equ.** to estimate number of hours he spends studying per week.

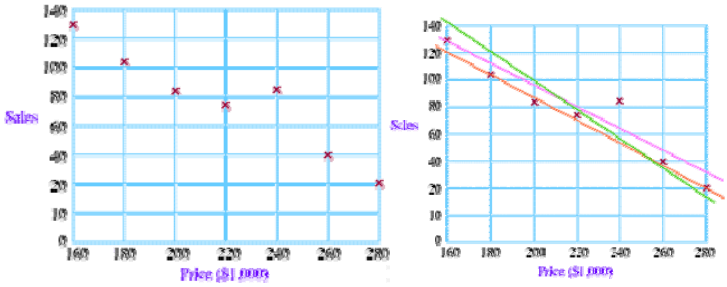
and if $y = 90, \quad x' = 11.69$ hrs

8. Compute the coefficient of determination $(r^2 \times 100)$ and **comment** on that: $(r^2 \times 100) = (.9824^2 \times 100) = 96.5\ \%$, 96.5% of variations in test score are explained by regression equation

Best Fit Line (Regression Line): We start with an attempt to construct a linear demand function. Suppose that your market research of real estate investments reveals the following sales figures for new homes of different prices over the past year.

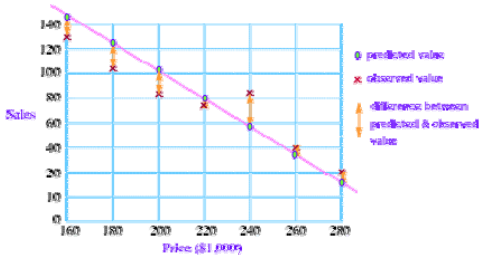
Price (Thousands of $)	$160	$180	$200	$220	$240	$260	$280
Sales of New Homes This Year	126	103	82	75	82	40	20

We would like to use these data to construct a demand function for the real estate market. (Recall that a demand



function gives demand y, measured here by annual sales, as a function of unit price, x.) Here is a plot of y versus x. The data definitely suggest a straight line, more-or-less, and hence a linear relationship between p and q. Here are several possible "straight line fits."

Q Which line best fits the data?
A We would like the sales predicted by the best-fit line (**predicted values**) to be as close to the actual sales (**observed values**)as possible. The differences between the predicted values and the observed values appear as the vertical distances shown in the figure below.



Q Since we want the vertical distances to be as small as possible, why can't we set them all to zero and solve for the slope and intercept of the straight line?
A If this were possible, then there would be a straight line that passes through all the data points. A look at the graph shows that this is not the case.

Q Then why not find the line that minimizes *all* the vertical distances?
A This is not possible either. The line that minimizes the first two distances is the line that passes through the first two data points, since it makes the distances 0. But this line certainly does not minimize the distance to the third point. In other words, there is a trade-off: making some distances smaller makes others larger.

Q So what do we do?
A Since we cannot minimize *all* of the distances, we minimize some reasonable combination of them. Now, one reasonable combination of the distances would be their *sum*, but that turns out the be difficult to work with (because distances are measured in terms of absolute values). Instead, we use the sum of the *squares* of the distances (no absolute values required). The line that minimizes this sum is called the best fit line**,** regression line, **or** least squares line **associated with the given data.**

Principles of Causation

Types of association

An association may be found between two variables for several reasons (show causal modeling figures):

- there may be **direct causation**, e.g. smoking causes lung cancer
- there may be a **common cause**, e.g. ice cream sales and number of drowning both increase with temperature
- there may be a **confounding factor**, e.g. highway fatalities decreased when the speed limits were reduced to 55 mph at the same time that the oil crisis caused supplies to be reduced and people drove fewer miles.
- there may be a **coincidence**, e.g., the population of Canada has increased at the same time as the moon has gotten closer by a few miles.

Establishing cause-and effect:

How do we establish a cause and effect relationship? It is generally agreed that most or all of the following must be considered before causation can be declared.

Strength of the association.
The stronger an observed association appears over a series of different studies, the less likely this association is spurious because of bias.

Dose-response effect.
The value of the response variable changes in a meaningful way with the dose (or level) of the suspected causal agent.

Lack of temporal ambiguity.
The hypothesized cause precedes the occurrence of the effect. The ability to establish this time pattern will depend upon the study design used.

Consistency of the findings.
Most, or all, studies concerned with a given causal hypothesis produce similar findings. Of course, studies dealing with a given question may all have serious bias problems that can diminish the importance of observed associations..

Biological or theoretical plausibility.
The hypothesized causal relationship is consistent with current biological or theoretical knowledge. Note, that the current state of knowledge may be insufficient to explain certain findings.

Coherence of the evidence.
The findings do not seriously conflict with accepted facts about the outcome variable being studied.

Specificity of the association.
The observed effect is associated with only the suspected cause (or few other causes that can be ruled out).

IMPORTANT: NO CAUSATION WITHOUT MANIPULATION!

Examples: Discuss the above in relation to:

- smoking vs. lung cancer.
- amount of studying vs. grades in a course.
- sex education in school vs. having pre-marital intercourse.
- fossil fuel burning and the greenhouse effect.
- free trade vs plant closings.

Basic Probability

$$\text{Probability of an event A } = P(A) = \frac{f}{N} = \frac{\text{The Number of } \textbf{desired outcome} \text{ Can Occur}}{\text{The Total Number Of } \textbf{Possible Outcomes}}$$

$$0 \le P(A) \le 1$$

If the **probability** of occurrence of an event such as event A is between $0 \le P(A) < 5\%$ then its occurrence is called unusual.

Definition	**Example**	
An **experiment** is a situation involving chance or probability that leads to results called outcomes.	Tossing a coin.	Rolling a Die
All possible outcomes of the experiment are called **sample space** Find **N = ?**	**sample space** **N = 2** outcomes (H,T)	**sample space** **N** = 6 outcomes (1,2,3,4,5,6)
What is/are the **desired outcome** or **Outcomes**? Find **f = ?**.	(to be tail) **f = 1**	(an odd number 1,3,5) **f = 3**
Probability is the measure of how likely an event is $= P(A) = \frac{f}{N}$	probability to be tail P(H) = 1/2 =50%	probability to be an odd number l P(odd number) = 3/6 =50%

Example A: Frequency distribution of annual income for U.S. families

Income	Frequency (1000s)
Under $10,000	5,216
$10,000–$14,999	4,507
$15,000–$24,999	10,040
$25,000–$34,999	9,828
$35,000–$49,999	12,841
$50,000–$74,999	14,204
$75,000 & over	12,961
	69,597

Part 1: Find the probability that a randomly selected person from this group makes $75,000 and over

1) Experiment: randomly selecting a person. **2)** Sample space $= N = 69,597$

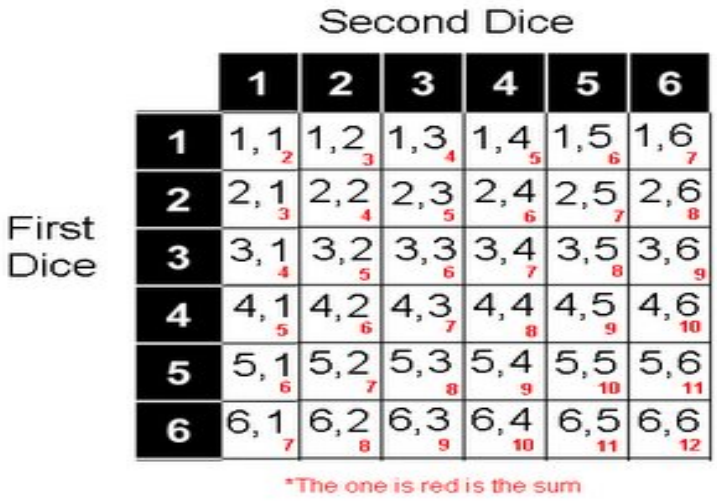
3) His/her income is $75,000 and over: $f = 12,961$ **4)** Prob ($75,000 and over) $= 12,961 / 69,597 = 18.63\,\%$

Part 2: Find the probability that a randomly selected person from this group makes $24,999 or less

1) Experiment: randomly selecting a person. **2)** Sample space $= N = 69,597$

3) His/her income is $75,000 and over: $f = 19,763$ **4)** Prob ($24,999 or less) $= 19,763 / 69,597 = 28.40\,\%$

Example B:
If we roll 2 dice, then there are 36 possible outcomes meaning that the **sample space is 36** or $= N = 36$



Solution:

a) find the probability that their total is 10
Desired outcomes: to get a total of 10 $\Rightarrow \{(4,6),(5,5),(6,4)\} \Rightarrow \ f = 3$
Prob (a sum of 10) $= 3/36 = 1/12 = 8.33\,\%$

b) find the probability that their total 10 or more
Desired outcomes: to get a total 10 or more $\Rightarrow \{(4,6),(5,5),(5,6),(6,4),(6,5),(6,6),\} \Rightarrow \ f = 6$
Prob (a total of 10 or more) $= 6/36 = 1/6 = 16.67\,\%$

c) find the probability that their total is 5
Desired outcomes: to get a total of 5 $\Rightarrow \{(1,4),(2,3),(3,2),(4,1)\} \Rightarrow \ f = 4$
Prob (a total of 5) $= 4/36 = 1/9 = 11.11\,\%$

Example C.

In a deck of 52 cards there are 13 diamonds and 12 faces, and 4 aces. If one card is drawn randomly find the probability that

a) it is a diamond b) it is a face c) It is a diamond and face

Solution:

a) P(diamond) $= 13/52 = 25\,\%$ b) P(face) $= 12/52 = 23.08\,\%$ c) P(diamond and face) $= 3/52 = 5.77\,\%$

Multiplication Rule (Keywords: and, both, all)

$$P(A \text{ and } B \text{ and } C \text{ and} \ldots) = P(A)P(B)P(C)\ldots$$

We use multiplication rule to find the probability that events A, B, C happen together
Hint:
When you make a selection out of a group by using multiplication rule be aware of **with** or **w/o** replacement effect.

In a deck of 52 cards there are 13 diamonds and 12 faces, and 4 aces.

If 2 cards are randomly drawn **w/o replacement,** what is the probability that both are diamonds?

$$\text{P(both diamond)} = \frac{13}{52}\bullet\frac{12}{51} = 5.88\,\%$$

If 2 cards are randomly drawn **with replacement,** what is the probability that both are diamonds?

$$\text{P(both diamond)} = \frac{13}{52}\bullet\frac{13}{52} = 6.25\,\%$$

There are 13 diamonds and 12faces, and 4 aces in a deck of card.
If 4 cards are randomly drawn **w/o replacement** then,

a) What is the probability that all 4 are diamond and how likelihood is this?

$$\boxed{\begin{matrix}13\ \text{Diamo}\\ 39\ \text{Others}\end{matrix}}_{52} \Rightarrow \boxed{\begin{matrix}12\ \text{Diamo}\\ 39\ \text{Others}\end{matrix}}_{51} \Rightarrow \boxed{\begin{matrix}11\ \text{Diamo}\\ 39\ \text{Others}\end{matrix}}_{50} \Rightarrow \boxed{\begin{matrix}10\ \text{Diamo}\\ 39\ \text{Others}\end{matrix}}_{49} \qquad \frac{13}{52}\bullet\frac{12}{51}\bullet\frac{11}{50}\bullet\frac{10}{49} = .26\% \quad \text{very unlikely}$$

b) What is the probability that all 4 are aces and how likelihood is this?

$$\boxed{\begin{matrix}4\ \text{Aces}\\ 48\ \text{Others}\end{matrix}}_{52} \Rightarrow \boxed{\begin{matrix}3\ \text{Aces}\\ 48\ \text{Others}\end{matrix}}_{51} \Rightarrow \boxed{\begin{matrix}2\ \text{Aces}\\ 48\ \text{Others}\end{matrix}}_{50} \Rightarrow \boxed{\begin{matrix}1\ \text{Aces}\\ 48\ \text{Others}\end{matrix}}_{49} \qquad \frac{4}{52}\bullet\frac{3}{51}\bullet\frac{2}{50}\bullet\frac{1}{49} = .000369\ \%\ \text{very much unlikely}$$

c) What is the probability that all 4 are faces and how likelihood is this?

$$\boxed{\begin{matrix}12\ \text{Faces}\\ 40\ \text{Others}\end{matrix}}_{52} \Rightarrow \boxed{\begin{matrix}11\ \text{Faces}\\ 40\ \text{Others}\end{matrix}}_{51} \Rightarrow \boxed{\begin{matrix}10\ \text{Faces}\\ 40\ \text{Others}\end{matrix}}_{50} \Rightarrow \boxed{\begin{matrix}9\ \text{Faces}\\ 40\ \text{Others}\end{matrix}}_{49} \qquad \frac{12}{52}\bullet\frac{11}{51}\bullet\frac{10}{50}\bullet\frac{9}{49} = .001828 = .1828\%\ \text{much unlikely}$$

d) What is the probability that all 4 are non faces and how likelihood is this?

$$\boxed{\begin{matrix}12\ \text{faces}\\ 40\ \text{non-faces}\end{matrix}}_{52} \Rightarrow \boxed{\begin{matrix}12\ \text{faces}\\ 39\ \text{non-faces}\end{matrix}}_{51} \Rightarrow \boxed{\begin{matrix}12\ \text{faces}\\ 38\ \text{non-faces}\end{matrix}}_{50} \Rightarrow \boxed{\begin{matrix}12\ \text{faces}\\ 37\ \text{non-faces}\end{matrix}}_{49} \qquad \frac{40}{52}\bullet\frac{39}{51}\bullet\frac{38}{50}\bullet\frac{37}{49} = .337575 = 33.76\ \%\ \text{It is likely}$$

A. There are 10 men, and 8 women in a group. If two people are selected at random **without replacement**, then,

1.Write all the possibilities **2.** Compute all the probabilities

Possibilities			**Probabilities**
	10 M, 8 W (18) $\Rightarrow$ 9 M, 8 W (17)		
MM	$p(M)=10/18$ · $p(M)=9/17$	$\Rightarrow$	$P(MM)=\frac{10}{18}\cdot\frac{9}{17}=29.4\%$
	10 M, 8 W (18) $\Rightarrow$ 9 M, 8 W (17)		
MW	$p(M)=10/18$ · $p(W)=8/17$	$\Rightarrow$	$P(MW)=\frac{10}{18}\cdot\frac{8}{17}=26.14\%$
	10 M, 8 W (18) $\Rightarrow$ 10 M, 7 W (17)		
WM	$p(W)=8/18$ · $p(M)=10/17$	$\Rightarrow$	$P(WM)=\frac{8}{18}\cdot\frac{10}{17}=26.14\%$
	10 M, 8 W (18) $\Rightarrow$ 10 M, 7 W (17)		
WW	$p(W)=8/18$ · $p(W)=7/17$	$\Rightarrow$	$P(WW)=\frac{8}{18}\cdot\frac{7}{17}=18.3\%$ +
			100%

then, find the following probabilities,

5. Both are men. **29.4 %**

6. Both are women. **18.3 %**

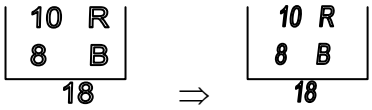
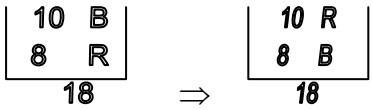
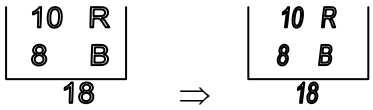
7. At least one woman. $P(MW)+P(WM)+P(WW)=26.14+26.14+18.3=$ **70.6 %**

8. At most one woman $P(MW)+P(WM)+P(MM)=26.14+26.14+29.4=$ **81.7 %**

9. One man and one woman. $P(MW)+P(WM)=26.14+26.14=$ **52.28 %**

B. In a box there are 10 Red and 8 Blue balls. If two balls are drawn at random **with replacement**, then

1.Write all the possibilities **2.** Compute all the probabilities

Possibilities			**Probabilities**
			
RR	$p(R)=10/18$ ·	$p(R)=10/18$ ⇒	$P(RR)=\frac{10}{18}\cdot\frac{10}{18}=30.86\%$
			
RB	$p(R)=10/18$ ·	$p(B)=8/18$ ⇒	$P(RB)=\frac{10}{18}\cdot\frac{8}{18}=24.69\%$
			
BR	$p(B)=8/18$ ·	$p(R)=10/18$ ⇒	$P(BR)=\frac{8}{18}\cdot\frac{10}{18}=24.69\%$
			
BB	$p(B)=8/18$ ·	$p(B)=8/18$ ⇒	$P(BB)=\frac{8}{18}\cdot\frac{8}{18}=19.75\%$ +
			100%

then, find the following probabilities,

5.Both are Red. $P(RR)=$ **30.86 %** **6**.Both are Blue. $P(BB)=$ **19.75 %**

7. At least one Red. $P(RB)+P(BR)+P(RR)=$ 24.69 + 24.69 + 30.86 = **80.25 %**

8. At most one Red. $P(RB)+P(BR)+P(BB)=$ 24.69 + 24.69 + 19.75 = **69.4 %**

10. One Red and one Blue. $P(RB)+P(BR)=$ 24.69 + 24.69 = **49.38 %**