

University of Bahrain
Business Administration College
Marketing and Management Department

QM250

Introduction to Statistics

Dr. Amin Al-Agha, Notes

Written by: Shawqi Radhi

21/9/01

QM250

Chapter 1: What is Statistics?

* Definitions

Statistics is a science that is concerned with collections, organizing, analysing and interpreting (explain) data to make better decision.

* 4 Important Collections

Figures & Organizing

* Analysing

* Interpreting [explain]

* ~~Populations~~ Collections:

Data can be collected from different resources depending on the nature and purpose of this data.

Data can be collected from:

[1] Population: the number of all items & objects which are "of interest" to our study.

[2] Sample: is a small part of the whole population.

* Reasons for collecting data from a Sample:

[1] Because the population could be too large and needs (request) time, money & effort.

[2] The population could be difficult to define.

[3] Because the study of the population could be destroyed [terminate].

* Random Samples: Choose a random sample depending on luck or chance.

* Not every sample is a good one, the good one should be a random sample.

21/9/2010

For a sample to be random,

- 1) Every member should have equal chances of being selected 'chosen' in this sample.

* Variables: could be Age, country, GPA.
It's to vary "change".

* constant: It's fixed, like birthday, names

* Types of Variables:

1) Qualitative: non-numeric [country, states, etc]

2) Quantitative: numeric [Age, birthday, etc] Speed, Time

3) Discrete: whole number without fraction
4) Continuous: can take whole or fraction

23/9/2010

* Variables can be classified according to the level of measurement: 4 types:

1) Nominal: non-numeric qualitative [colour, names, type]

2) Ordinal: ranking or ordering [performance...]

3) Ratio: must be a zero, the zero is the start of the scale

4) Interval: could be zero or no zero, if there is zero in the scale, this zero is not start of the scale, and zero does not mean there is nothing from the variable

salary
prices

Scale
Zero means
there is
nothing of value

"one good example for interval is the temperature"
"other example is shoes size & clothes size."

23/9/20 Chapter 2 : Describing Data; Frequency Distribution
+ ch: 3, 4 \Rightarrow Organizing data

* After data are collected; they need to be organized or presented in a ~~series~~ form. This form could be either:
* Numerical measure, or Graph measure.

* Data can be raw or ungrouped or they can be grouped into classes.

I. Organizing ungrouped or raw data:
example II

26/9/2010 a) The mean: The sum of numbers divided by their number. [like Average].

• Assume that "scores" is the X variable.

The mean = $\frac{\text{Sum of data values}}{\text{number of values}} \Rightarrow \bar{X}$ [mean]

If the variable y then mean = \bar{y}

Note:

Σ [sigma]

\hookrightarrow sum.

$$\bar{X} = \frac{\Sigma X}{n}$$

where, \bar{X} is the mean of the data values

ΣX is the sum of the data values

n : is the values in the data (sample size)

$n \Rightarrow$ number of values

$$\text{Ex: } \bar{X} = \frac{85 + 50 + 78 + 85 + 92}{5} = \frac{390}{5} = 78$$

* Features of means

1. For every set of data there is only one mean.
2. All data values are used on the calculation of the mean.
3. The mean is also known as: the Average, the arithmetic mean, and the expected value.

4* The mean is affected or influenced by extreme values.

[illegible]

X	\bar{X}	$X - \bar{X}$
85	78	$85 - 78 = +7$
50		$50 - 78 = -28$
78		$78 - 78 = 0$
85		$85 - 78 = +7$
92		$92 - 78 = +14$

$$\Rightarrow \sum (x - \bar{x}) = 0$$

2/10/20

26/9/2010

The median:

Is the middle or central value of an ordered set of data. by ordered set of data we mean, the data values should be rearranged from smallest to largest, and then pick up the central data

* Rearrange our data values: 85 50 78 85 92

In order \Rightarrow 50 78 85 85 92

median \Rightarrow 85

* ~~Other~~ other examples 12 10 15 8

In order \Rightarrow 8 10 12 15

median \Rightarrow $\frac{10+12}{2} = \frac{22}{2} = \underline{11}$

* Location (Position): to find it, we must use this equation of the median $\Rightarrow = 0.5(n+1)$

* $\Rightarrow 0.5(4+1) = 2.5$ is the position.

* $\Rightarrow 0.5(5+1) = 3$ is the position.

* Features of the median:

[1] 50% of the data values fall below the median,

[2] For every set of data there is only one median.

[3] The median is not affected, or influenced by extreme values

28/9/2010

* The Mode:

is the value (or values) that occurs more frequently than any other data value.

Example: 85 50 78 85 92

* In our example, the mode is 85 but it's occurs more than other values.

* Features of the Mode:

① Any set of data could have:

no mode, one mode, 2 mode, or more than 2 mode.

Examples

* 10 8 15 17 22 [No mode]

10 8 10 8 15 15 [No mode]

20 10 8 10 8 15 15 [3 modes]

20 30 25 17 30 [one mode, 30]

30 20 25 20 25 [2 mode, 20 & 25]

~~40 45 40 40 40 45 48~~ [one mode, 40]

40 45 40 40 40 45 48 45 [2 mode, 40, 45]

② The mode is not affected or ~~influenced~~ or influenced by extreme values:

* 10 15 10 300 \Rightarrow mode is 10

* 10 15 10 \Rightarrow mode is 10

③ The mode can be used to describe quantitative as well as qualitative data [nominal & non-nominal]

* Black Blue Black \Rightarrow Mode is Black

* Tall Short Short tall Short \Rightarrow mode is short

Notes the above 3 ~~measures~~ mean, median, mode are known as the measuring by "location" of of

28/9/2019 The
(ii) Range:

Is the difference between the highest data value and the smallest data value.

* Our examples 85 50 78 85 92

$$\text{Range} = \text{highest data} - \text{smallest data} \\ = 92 - 50 \Rightarrow 42$$

Features:

- ① It's easy to calculate.
- ② For every set of data there is only one range.
- ③ The range is affected or influenced by extreme value.
- ④ It is a weak measure of variation, because it depends the highest & lowest only & ignores the others.

Deviation
↓
Difference

* The Standard deviation [Most Important]

* Shows how far the data values are from their mean.

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

Where,

S = Standard Deviation.

\sum = Sum, total

x = the values

\bar{x} = the mean

n = number of data values, or sample size,

28/9/2010

X	$(X - \bar{X})$ $\bar{X} = 78$	$(X - \bar{X})^2$
50	$50 - 78 = -28$	$(-28)^2 = 784$
78	$78 - 78 = 0$	$(0)^2 = 0$
85	$85 - 78 = +7$	$(7)^2 = 49$
85	$85 - 78 = +7$	$(7)^2 = 49$
92	$92 - 78 = +14$	$(14)^2 = 196$
$\Sigma 390$	$\Sigma(X - \bar{X}) = \text{Zero}$	$\Sigma(X - \bar{X})^2 = 1078$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{390}{5} = 78$$

$$S = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n - 1}}$$

$$S = \sqrt{\frac{1078}{5 - 1}} = \sqrt{\frac{1078}{4}} = \sqrt{269.5}$$

$S = 16.42$ The Standard Deviation

30/9/2010, Note: $\Sigma(X - \bar{X})^2$ is the sum of the squared deviation (difference) between the data values and their mean.

30/6/2010

* The Variance:

Is the squared of Standard deviation.

The variance is referred to " S^2 "

$$S = 16.42 \Rightarrow S^2 = (16.42)^2 = 269.5$$

* The Standard deviation is the squared root of the variance.

$$S \Rightarrow S = \sqrt{S^2}$$

$$S = \sqrt{269.5} \Rightarrow S = 16.42$$

$$S^2 = \frac{\sum (x - \bar{x})^2}{n-1}$$

* The features of Standard Deviation: (S)

(1) It cannot be negative.

(2) It can be positive or zero.

(3) If $S = 0$, this means that there is no deviation or variation between the data values & their mean.

All data values are equal.

Examples

x	$x - \bar{x}$	$(x - \bar{x})^2$
30	0	0
30	0	0
30	0	0
30	0	0

$$\left. \begin{array}{l} \sum x = 120 \\ \bar{x} = \frac{120}{4} = 30 \end{array} \right\} S = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} = 0$$

(L) It is [Standard Deviation] is influenced or affected by extreme values.

30/9/20(5) If the value of " S " is small, this means that the data values are more consistent.

→ Consistent, numbers that same or order or close to each other

like: 30 32 28 30 31

(the smaller, the value of " S ", the better)

& It is better to compare.

(6) It can be used to compare, 2 or more sets of data which has same units of measurement.

* The features of variance:

- (1) It is influenced, affected by extreme values
- (2) If the variance is zero, this means that there is no deviation between the data values & their mean, & the data values are all equal.

~~133~~

Note 3 [1] the range, Standard deviation, Variance are known as:

"The measures of variation"

Note 3 [2] "Variation" is also known as:

- Variability.
- Deviation,
- Difference.
- Dispersion
- Spread
- Change.

30/9/2010

iii. First & third quartiles

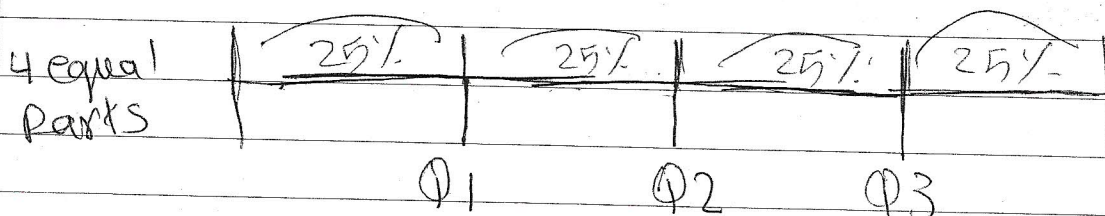
Quartile $\rightarrow 1/4$

* They are the values that ~~data~~ divide an ordered ~~the~~ set of data into 4 equal parts.

* an ordered: to arrange the data values should from smallest to largest.

* There are 3 quartiles:

- 1) First Quartile, Q_1
- median 2) Second Quartile, Q_2
- 3) Third Quartile, Q_3



3/10/2010

* to calculate quartiles we need to arrange data values from smallest to largest.

* 50¹ 78² 85³ 85⁴ 92⁵

Q_1
64

* First Quartile: Q_1

location of $Q_1 \Rightarrow 0.25(n+1)$

$$= 0.25(5+1) = 1.5$$

$$\text{Value of } Q_1 = \frac{50+78}{2} = 64$$

* This value means that ~~how~~ 25% of the data value fall below it

30/10/20

* Second quartiles: Q_2

$$\begin{aligned} \text{location of } Q_2 &= 0.5(n+1) \\ &= 0.5(5+1) = 3 \end{aligned}$$

$$\text{value of } Q_2 = 85$$

* Third Quartiles:

$$\begin{aligned} \text{location of } Q_3 &= 0.75(n+1) \\ &= 0.75(5+1) = 4.5 \end{aligned}$$

$$\text{Value of } Q_3 = \frac{85 + 92}{2} = 88.5$$

Note/To find the quartiles:

[1] Arrange values from smallest to largest

[2] find the location

[3] find the value of location

* The Interquartile Range:

$$\begin{aligned} IQR &= Q_3 - Q_1 \quad [\text{like the range}] \\ &= 88.5 - 64 = 24.5 \end{aligned}$$

* The quartile deviation:

$$\begin{aligned} QD &= \frac{Q_3 - Q_1}{2} = \frac{IQR}{2} \\ &= \frac{24.5}{2} = 12.25 \end{aligned}$$

30/10/2010

Q: i.v : 60th percentile:

* Are the values that divided an ordered set of data into 100 equal parts.

Same thing: - Rearrange values [from smallest to large]
- Find location.
- Find the value.

* 60th percentile 50 78 85 ^{3.6} 85 92
P₆₀

$$\begin{aligned} \text{* location of } P_{60} &= 0.6(n+1) \\ &= 0.6(5+1) = 3.6 \end{aligned}$$

$$\begin{aligned} \text{* Value of } P_{60} &= 3\text{rd value} + 0.6(4\text{th value} - 3\text{rd value}) \\ &= 85 + 0.6(85 - 85) = \boxed{85} \end{aligned}$$

Important: Note, other examples

$$\text{* location} = 2.75$$

$$\text{Value} = 2\text{nd value} + 0.75(3\text{rd} - 2\text{nd})$$

$$\text{* location} = 110.32$$

$$\text{Value} = 110\text{th value} + 0.32(111\text{th} - 110\text{th})$$

$$\text{* location} = 7.63$$

$$\text{Value} = 7\text{th value} + 0.63(8\text{th} - 7\text{th})$$

~~we~~ we can applied this method for median, quartiles

~~Notes~~

$$\text{location of } Q_1 = 0.25(n+1)$$

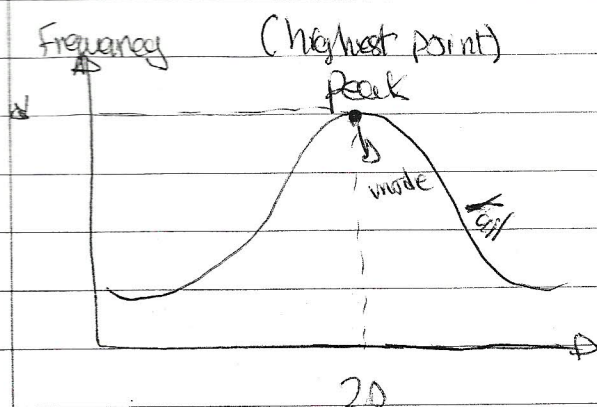
$$= 0.25(6) = 1.5$$

$$\text{Value of } Q_1 = 1\text{st} + 0.5(2\text{nd} - 1\text{st})$$

$$= 50 + 0.5(78 - 50)$$

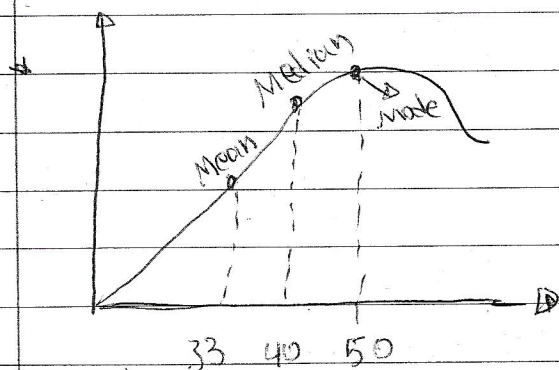
Interpret it: This means that 60% of data value fall [explain] below 85

5/10/2010 V. Coefficient of skewness



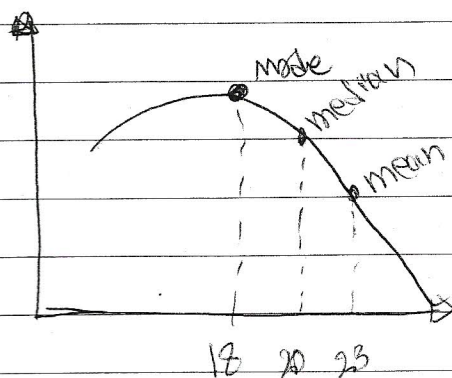
[1] Normal or symmetric curve

Always:
mean = median = mode
Is the highest point (peak)



[2] The curve is skewed to the left (It has negative skewness).

Always:
Mode \gg median $>$ mean



[3] The curve is skewed to the right (It has ~~negative~~ positive skewness)

Always:
Mean $>$ median $>$ mode

* Any set of data can take one of three types of curves on their distribution. If the two tails are equal like number [1] above [Normal], we say that the curve is normal or symmetric, In this case there is no skewness.

5/10/2010

* Skewness is happen, when one of the tails of the curve is longer than other, like on curve two or three.

* To measure the amount or degree of Skewness on the data, we use the coefficient of SK (Skewness) we use this equation:

$$SK = \frac{3(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

Data values:

85 50 78 85 92

Mean = 78 } SD = 16.42

Median = 85 }

$$\Rightarrow \frac{3(78 - 85)}{16.42} = -1.28 \quad \boxed{-ve}$$

* This means, that there is a negative skewness on the data

-ve
negative

+ve
positive

* Features of SK :

1. SK ranges between -3 and +3

$$\Rightarrow -3 \leq SK \leq +3$$

i.e: SK cannot be 4 or more.

7/10/2010

2. If SK = 0, It means there is no skewness on the data, and the data curve is normal or symmetric.

3. If SK is negative [-ve], This means, the data curve is skewed to the left.

4. If SK is positive [+ve], This means, the data curve is skewed to the right.

7/10/2010

VI. Coefficient of Variation [C.V]

It measure, how large the standard deviation is, compared to the mean.

$$CV = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100 \Rightarrow \frac{s}{\bar{x}} \times 100$$

$$CV = \frac{16.42}{78} \times 100 = 21.1 \% \text{ of the mean}$$

• Other names for [C.V]:

- Relative Dispersion.
- Variability percentage.

* CV can be used to compare, the variation of two or more sets (groups) of data, which have different units of measurement (like money on any currency + age on years).

30/12/2010
29/12/2010
10/12/2010
* Because the two samples (the scores sample & GPA Sample) have different units of measurement, we use the CV, to compare their variation (or variability).

* For scores Sample:

$$CV = \frac{16.42}{78} \times 100 = 21.1 \%$$

* For GPA Sample:

$$CV = \frac{\sqrt{\text{Variance}}}{\bar{x}} \times 100$$

$$= \frac{\sqrt{0.436}}{2.75} \times 100 = 24 \%$$

This means that the variation on the GPA Sample is higher than the variation on Scores Sample.

7/10/2010

B. Population mean:

$$\text{Population, } \mu = \frac{\sum X}{N}$$

Where,

μ is the population mean, $\sum X$ is the sum of the values, N is the population sizes.

$$* \mu = \frac{390}{5} = 78$$

Population Standard Deviation:

σ (sample sigma), Σ (Capital sigma)

$$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}}$$

Where,

σ is the population standard deviation, μ is the population mean, & $\sum (X - \mu)^2$ is the sum value of the square deviating between the data value of the means,

N = Population.

10/10/2010

10/10/2010 * The population Variance, σ^2

$$\sigma^2 = \frac{\sum (x - M)^2}{N}$$

* $\sum (x - M)^2$ = The sum of the squared deviations between data values and their mean.

In our examples

x	$x - M$	$(x - M)^2$
50	$50 - 78 = -28$	784
78	0	0
85	+7	49
85	+7	49
92	+14	196

$$\begin{aligned}\sum x &= 390 \\ M &= 78\end{aligned}$$

$$\sum (x - M) = 0$$

$$\sum (x - M)^2 = 1078$$

$$\sigma = \sqrt{\frac{\sum (x - M)^2}{N}} = \sqrt{\frac{1078}{5}} = \sqrt{215.6} = 14.68$$

$$\sigma^2 = 215.6$$

10/10/2010

* Organising Grouped Sample Data :

* The data values which we used in the last example are called "Ungrouped or raw data".

* If the data values are too many, we need to group them into groups or classes.

* Each class contains a ~~se~~ certain number of data values, that has a beginning and ending.

* Examples class 10 - 20
Begin. [low limit] \rightarrow ending [upper limit]

* Frequency Distribution Table :

a) Group data into classes. \Rightarrow ~~Important steps~~

~~1) Determine the number of classes.~~

~~2) $2^k > N$, where "k" is the number of data.~~

$N = 10 \Rightarrow 2^1 > 10$ X less

$2^2 = 4 > 10$ X less

$2^3 = 8 > 10$ X less

$2^4 = 16 > 10$ \checkmark stop here

$K = 4$, we need Four classes.

~~3) Determine the class interval or width (i)~~

$$i \geq \frac{H - L}{K} \rightarrow \text{range}$$

where, H : Highest data value

L : Lowest data value

K : Number of classes.

12/10/2010
14/10/2010

3/4 Make a frequency Distribution table

Sl. No.	Class limits	Tally	f	rf	Cf	m	$\sum fm$	$(m - \bar{x})$	$(m - \bar{x})^2$	$(m - \bar{x})^2 f$
1	20 up to < 25	///	3	.3	3	22.5	3 x 22.5 = 67.5	22.5 - 28 = -5.5	(-5.5) ² = 30.25	90.75
2	25 upto < 30	////	4	.4	7	27.5	110	-5	0.25	1.0
3	30 upto < 35	///	2	.2	9	32.5	65	-4.5	20.25	40.5
4	35 upto < 40	/	1	.1	10	37.5	37.5	-9.5	90.25	90.25
Total			10	1.0			280			222.5

K=4

i=5

37 22 34 20 28 30 27 20 25 29

Notes [1] The lower limit of the first class is the smallest ~~lower~~ data value.

[2] The upper limit of the Class, lower limit plus i.

[3] For class number one, the limit can be written also as follows: ~~range~~
20 up to under 25 OR 20 up to 24.999 and so on.

[4] f mean frequency, which means, the number of data values, that belongs to this class.

[5] $\sum f = n$, sum of frequency ^{should} always equal 'n'

[6] Relative frequency is the $\frac{f}{\sum f}$ or $\frac{f}{n}$
 $\sum rf = 1.0$ [should]

^{slw} [7] Cf = Cumulative frequency

[8] m = Midpoint of a class. [Short for middle]
= $\frac{\text{lower limit} + \text{upper limit}}{2}$

14/10/2016

[a] The difference between any mid-point & the following one, is the same as the class interval (i)

[10] The mean of grouped sample data

$$\bar{X} = \frac{\sum fm}{\sum f} = \frac{\sum fm}{n}$$

$$\bar{X} = \frac{280}{10} = 28$$

[11] In the calculation of the mean of grouped data we use the mid point (fm) to represent all the values in the class.

* The Standard deviation of grouped data

$$S = \sqrt{\frac{\sum (m - \bar{x})^2 f}{\sum f - 1}} = \sqrt{\frac{\sum (m - \bar{x})^2 f}{n - 1}}$$

$$\text{Our example} = \sqrt{\frac{222.5}{10 - 1}} = 4.97$$

* The Variance, S^2

$$S^2 = (4.97)^2 = 24.72$$

* The Mode:

$$\text{mode} = L + i \left(\frac{d_1}{d_1 + d_2} \right)$$

Model
class

* The modal class: Is the class that contains the mode on it and that has the highest frequency in the table.

* on our example, the second class has the highest frequency of four. therefore it is the modal class.

14/10/2010

* i = Class Interval.

* $d1$ = highest frequency - previous frequency
(d) is short for difference.

on our example = $4 - 3 \Rightarrow \boxed{1} \Rightarrow d1$

* $d2$ = highest frequency - following frequency

on our example = $4 - 2 \Rightarrow \boxed{2} \Rightarrow d2$

So, the Mode 15^{th} LL_{modal class} + $i \left(\frac{d1}{d1+d2} \right)$

$$= 25 + 5 \left(\frac{1}{1+2} \right) = \boxed{26.67}$$

is the mode

* The Medians of grouped data

first \Rightarrow location of median:

$$= 0.5(n+1)$$

$$= 0.5(10+1) = 5.5$$

Second \Rightarrow value of median:

17/10/2010 * from the "cf", column: location "5.5" is among the 7, of the second class. Therefore, the second class is the median class.

$$* \text{Median} = LL_{\text{median class}} + i \left(\frac{0.5n - CF_{\text{class before}}}{CF_{\text{median class}} - CF_{\text{class before}}} \right)$$

$$= 25 + 5 \left(\frac{0.5(10) - 3}{7 - 3} \right)$$

$$= 25 + 5(1.5) = 27.5$$

17/10/2010

* First & the third quartiles:

$$\bullet \text{ location of } Q_1 = 0.25(n+1) = 0.25(10+1) \\ = 2.75$$

~~location of Q_1~~

"from the cf column, location of 2.75 ^{is in} the first class"
therefore, the first class is the Q_1 class.

$$\begin{aligned} \text{value of } Q_1 &= LL_{Q_1 \text{ class}} + i \left(\frac{0.25n - CF_{\text{before}}}{CF_{Q_1 \text{ class}} - CF_{\text{before}}} \right) \\ &= 20 + \cancel{10}^5 \left(\frac{0.25(10) - 0}{3 - 0} \right) \\ &= 20 + 4.17 = 24.17 \end{aligned}$$

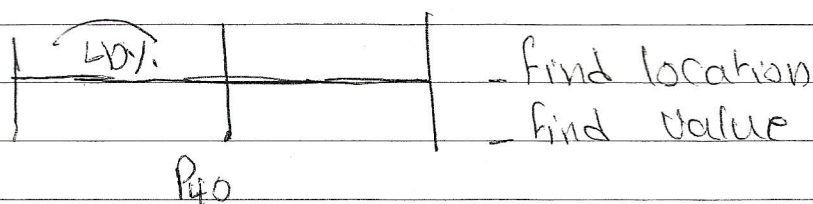
$$\bullet \text{ location of } Q_3 = 0.75(n+1) \\ = 0.75(10+1) \\ = 8.25$$

* From the cf column, location of 8.25 is in the third class, therefore the third class is the Q_3 class.

$$\begin{aligned} Q_3 &= LL_{Q_3 \text{ class}} + i \left(\frac{0.75n - CF_{\text{before}}}{CF_{Q_3 \text{ class}} - CF_{\text{before}}} \right) \\ &= 30 + \cancel{10}^5 \left(\frac{0.75(10) - 7}{9 - 7} \right) \\ &= 30 + 1.25 = 31.25 \end{aligned}$$

17/10/200

* The Age for the first 40% of the sample



$$\begin{aligned} \text{* location of } P_{40} &= 0.4(n+1) \\ &= 4.4 \end{aligned}$$

* From the CF column, location of the 4.4 is in the ~~2nd~~ second class, therefore ~~the~~ ^{second} class is P_{40} class.

$$\begin{aligned} \text{Value of } P_{40} &= LL_{P_{40} \text{ class}} + i \left(\frac{0.4n - CF_{\text{before}}}{CF_{P_{40} \text{ class}} - CF_{\text{before}}} \right) \\ &= 25 + 5 \left(\frac{0.4(10) - 3}{7 - 3} \right) \\ &= 25 + 1.25 = 26.25 \end{aligned}$$

* This means that 40% of the data values, full ~~below~~ below 26.25.

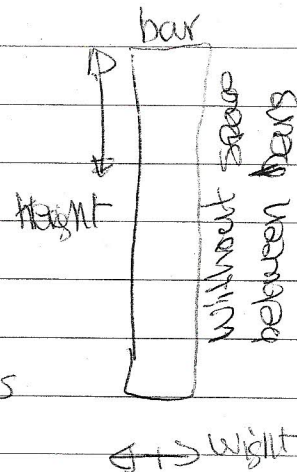
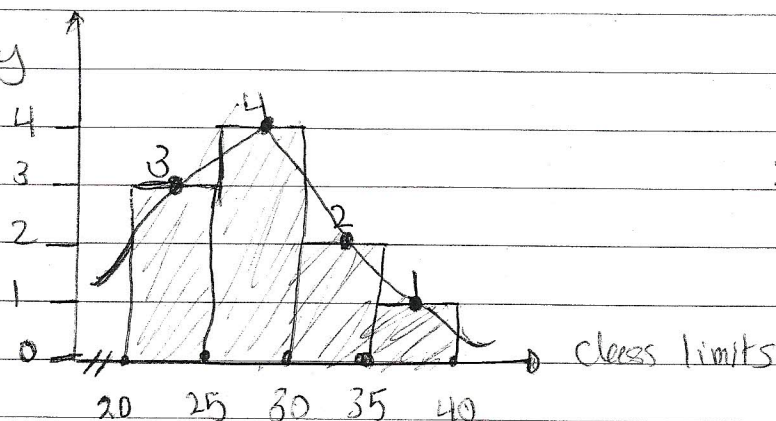
17/10/2010

C. Draw a histogram & a cumulative frequency polygon (an ogive):

- * Histogram is a ~~bar~~ ~~graph~~ drawn between two ~~axes~~ axes, one for the frequency (f) & the second for the class limits.
- * There are no gaps (spaces) between the bars.

19/10/2010

frequency
(f)



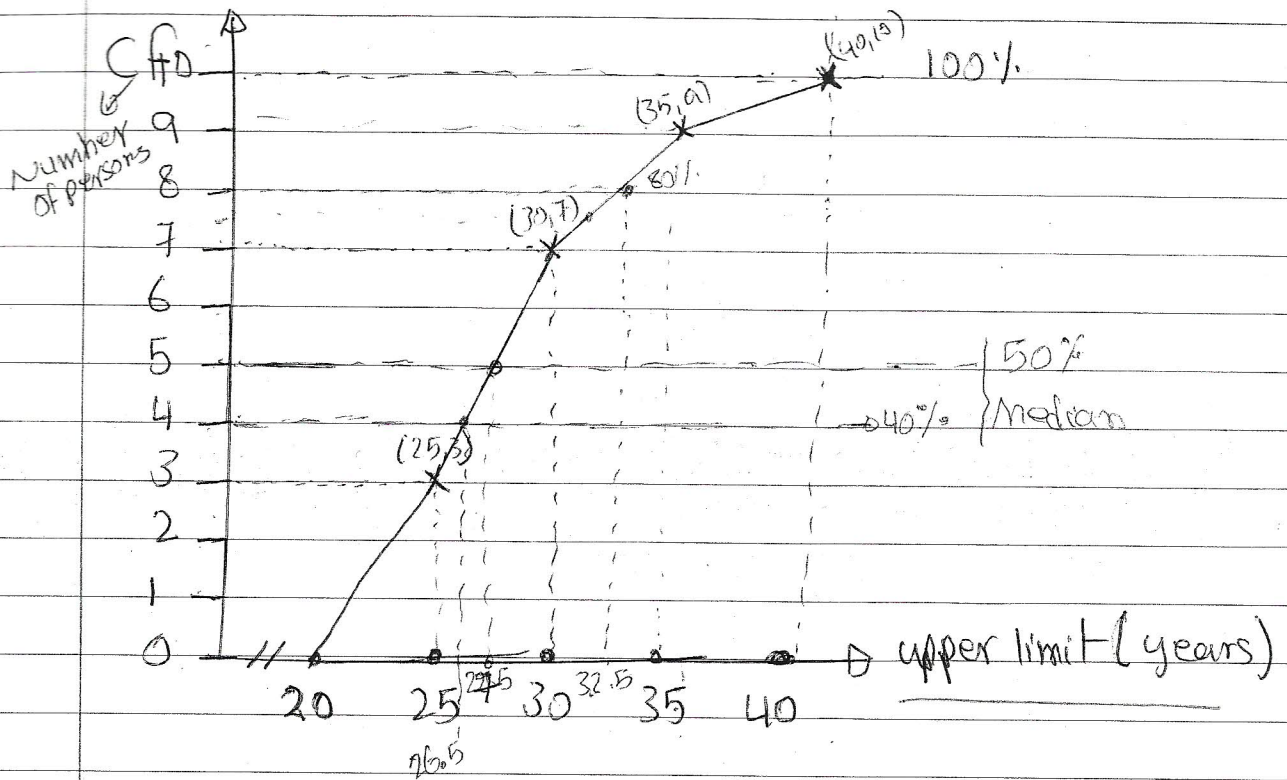
Majority
→ highest

* From the histogram, we can see that the Age of the Majority of Persons on Sample is between "25-30" years, while the Age of minority of persons on Sample is 35-40 year

Minority
→ smallest

note * The histogram is the basis of every curve for of the data.

10/10/20 * Cumulative frequency polygon (an ogive):



d. from the ogive find the: i, ii, iii, iv [Example sheet]:

i. Number of persons who are:

* (1) younger than 30 years = [go left from ogive]

The number of person who are younger than 30 (less than 30) are "7"

* (2) Older than 35 years =

The number of person who are older than 35 is equal to $10 - 9 = 1$

where 10 is total number and 9 is 35 & less.

* (3) between 30 and 35 years

The number of persons who are between 30 & 35 years old is equal to $9 - 7 = 2$

21/10/2010

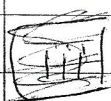
21/10/2019



Percent of persons who are 32.5 years or less;

* The percent of persons who are 32.5 years or less is 80%. " $\frac{8}{10} \times 100 = 80\%$ "

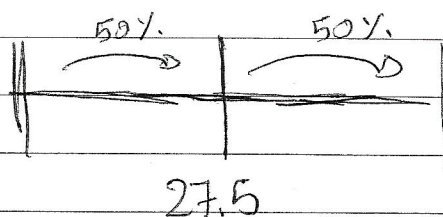
Note: Draw a line on an ogive curve for the number 32.5 and find the number of person and find the percent by using the following: the number of person occur on the ogive divided by total person and then multiply by 100.



Age below which 50% of the persons;

* The Age below which 50% of the persons is 27.5 years.

This is the median of the data.



40th Percentile: ~~the 40th~~

* The Age below which 40% of the persons is about 26.5 years.

* A 40% is occur on # 4 [see ogive curve].

Other examples: $n=72$, $P=65$

$$\Rightarrow 65\% \times 72 = 46.8$$

$n=120$, $P=70$

$$\Rightarrow 70\% \times 120 = 84$$

- & then draw a curve on ogive to get the number.

21/10/2010

(IV) $IQR = Q_3 - Q_1$ [like the range]
 $= P_{75} - P_{25}$

$n=10$

Notes: $75\% \times 10 = 7.5 \Rightarrow Q_3 = 31.25$

$25\% \times 10 = 2.5 \Rightarrow Q_1 = 24.25$

$$S_o = Q_3 - Q_1$$

$$P_{75} - P_{25}$$

$= 31.25 - 24.25$

$IQR = 7 \text{ years}$

Finish

On 1, 2, 3, 4

General Notes:

- Not because the sample is smaller than the population, that the sample mean (\bar{x}) should be smaller than the population mean (μ). (\bar{x}) could be equal to, smaller, or greater than (μ).

24/10/2010

- You could be given part of frequency table to answer question. In this case, just complete ~~the~~ the frequency table and answer the question

i.e.

limits	f
70 up to 80	15
80 up to 90	22
90 up to 100	13

Q3

- For grouped population data $\mu = \frac{\sum fm}{N}$

$$\sigma = \sqrt{\frac{\sum (m - \mu)^2 f}{N}}$$

Sample
Statistic

Population
Parameter

- Any numerical measure calculated for sample data (\bar{x} , s , s^2 ...etc) is called a Statistic, while any measure calculated from population data is called Parameter (μ , σ , σ^2 ...etc).

24/10/2010

5. The Science of Statistics, into 2 types; / branches

1 - Descriptive Statistics (to describe)

2 - Inferential Statistics (to infer: to conclude).

① Descriptive Statistics:

This a branche concerns all the methods to organize, summarize, and present data on a useful form.

② Inferential Statistics:

This a branch concerns all the methods used to make ~~descriptions~~ decision about population, based on sample data.

Exercises:

① The following Samples 15 30 -25 B has a mean of 10, find the value of B.

Solution:

$$\bar{X} = \frac{\sum X}{n} \Rightarrow \frac{10}{1} = \frac{15 + 30 + (-25) + B}{4}$$

$$40 = 20 + B$$

$$B = 40 - 20 = 20$$

② If: CV = 20%, Variance = 16, Find \bar{X}

Solution:

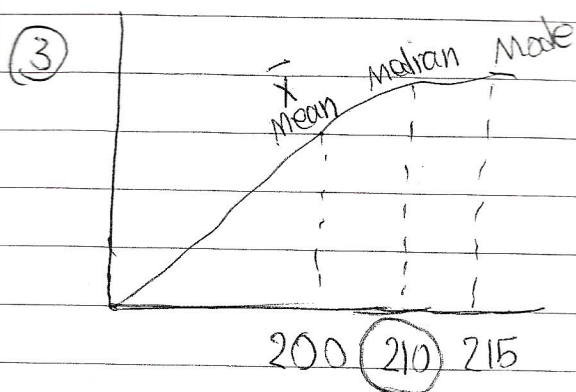
$$CV = \frac{\text{Standard Deviation}}{\bar{X}} \times 100$$

$$20 = \frac{\sqrt{16}}{\bar{X}} \times 100$$

$$20 = \frac{400}{\bar{X}}$$

$$20\bar{X} = 400$$

$$\bar{X} = \frac{400}{20} = 20$$

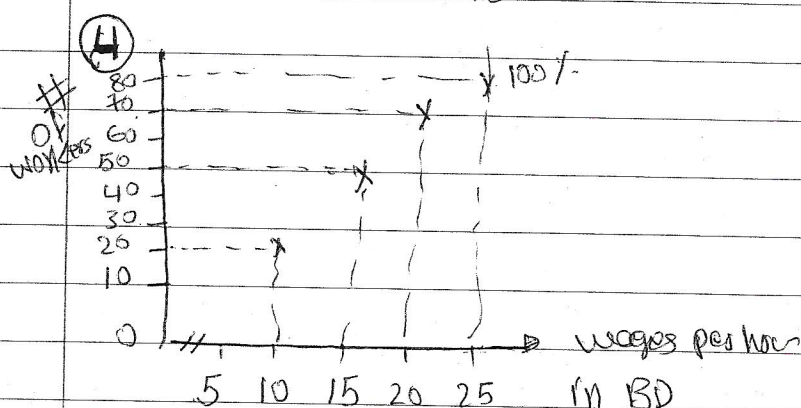


(a) $Q_2 = 210$

(b) If $S^2 = 225$ Find SK ?

$SK = \frac{3(\bar{x} - \text{Median})}{S}$

$$= \frac{3(200 - 210)}{15} = -2$$



(a) what is the name of this diagram?

⇒ an ogive, or cumulative frequency polygon

(b) How much the sample size?

⇒ 80 [Number of workers]

(c) How much is the class interval i ?

⇒ 5 [10-5]

(d) How many workers earn more than 15 BD per hour?

⇒ $80 - 50 = 30$ [total number - # of workers on 15]

(e) How many workers earn between 10 & 20 BD?

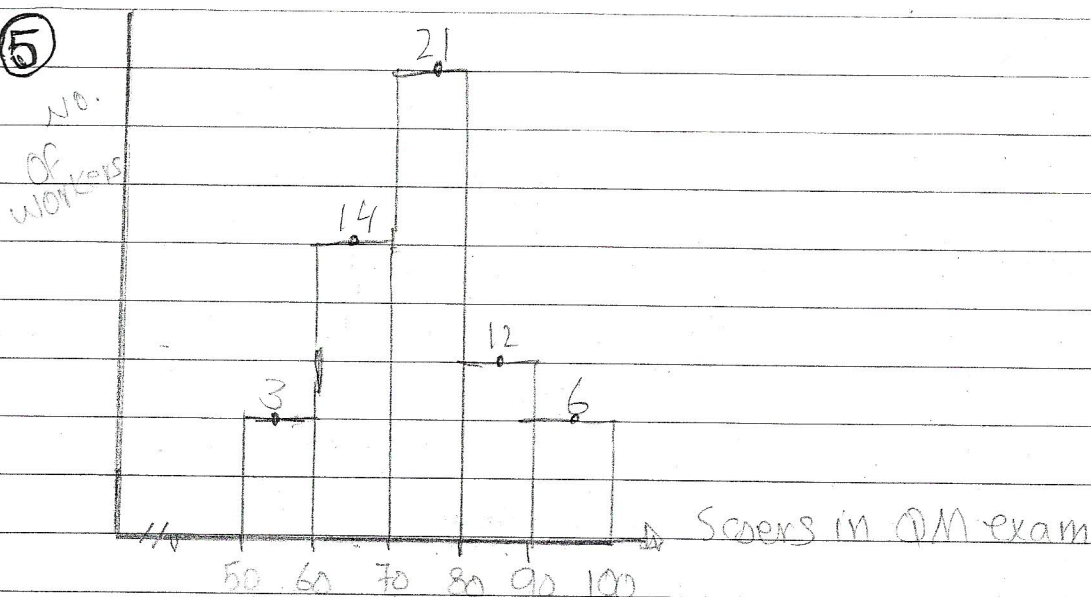
⇒ $70 - 20 = 50$

26/10/2012

(f) 25% of workers earn, how many or less?

⇒ 25% of workers is 25% of 80 = 20 workers they earn 10 80 or less.

(5)



① How many student do PM exam?

$$3 + 14 + 21 + 12 + 6 = 56 \text{ Students}$$

② How much is the class interval?

$$60 - 50 = 10$$

③ How many classes?

$$= 5$$

④ What is the class mid-point? For 14

$$m = \frac{\text{upper} + \text{lower}}{2} = \frac{60 + 70}{2} = 65$$

⑤ How many² student is full in exam (get ~~at~~ f)

$$= 3$$

⑥ How many student gets A?

$$= 6$$

⑦ How many student gets less than 70?

$$14 + 3 = 17$$

⑧ What is the lowest score in the class?

$$= 50$$

⑨ What is the r f For second class = $\frac{14}{56} = 0.25$

26/10/2010

⑩ what is the cumulative frequency of fourth class?
 $\Rightarrow 3 + 14 + 21 + 12 = 50$

* QUIZ 3/11/2010

11-12

1-19

out of 10

got $\Rightarrow 9$

26/10/2020 Chapter 5: Probability Concepts (Prot.)

4* Can called:
- Probabilition
- Possibility
- likelihood

* probability: Is a number between 0 and 1 that is used to describe the possibility of occurrence of a certain events.

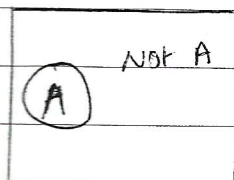
* Shortcut: $P(A)$, where P is the probability & (A) is the event.

* So, $0 \leq P(A) \leq 1$
- Note that, the "P" can not be negative (less than 0) or more than 1.

* Could used the percentage or decimal or fraction
i.e: 80%, 0.8, $\frac{8}{10}$
⇒ Can expressed either on above forms.

28/10/2020 * The probability of occurrence of a certain event plus the probability of non-occurrence of the same event should equal one (100%).

* *



* Venn Diagram:
 $P(A) + P(\text{not } A) = 1.0$

* The event "A" and the event "Not A" are called Complementary event.

28/10/2010

Two events A & B are complementary

IF $P(A) + P(B) = 1.0$

example: - $P(A) = .7$, $P(B) = .4 \Rightarrow$ not comple.

- $P(A) = .65$, $P(B) = .35 \Rightarrow$ complementary

- $P(A) + P(B) = 1.0$, $P(A) = .65$ & $P(B) = ?$

$P(B) = 1 - .65 = .35$ } IF unknown

$P(A) = 1 - .35 = .65$ } we can find it.

die
dice

Sample Space:

- Sample space for flipping a coin is [head, tail] or { }

- Sample space for throwing a die is [1, 2, 3, 4, 5, 6]

- Sample space for playing a football game [win, lose, draw]



So, Sample Space is the set of all possible outcomes of an experiment.



The Sum of prob. of the outcomes that belong to the sample space equals one.

* Probable
* likely
* possible

Methods for estimating Prob. or assigning probabilities:



- The classical method of equally likely outcomes.

by equally likely outcomes, we mean the outcomes

have equal probability of occurrence.

Example: coins / game / die.

$$P(A) = \frac{\text{Number of ways by which "A" occurs}}{\text{number of all possible outcomes}}$$

$$P(\text{head}) = \frac{1}{2} = 0.5$$

$$P(\text{1 on die}) = \frac{1}{6} = 0.167$$

22/10/2010

- [2]** The Empirical method of relative frequencies
- Empirical means: more common, natural.

$$P(A) = \frac{\text{number of times it occurs}}{\text{number of observations (sample size)}}$$

$$\text{So, } P(A) = \frac{f}{n}$$

- This method is used when events are not equally likely.

- [3]** The Subjective method:
- Depended on ~~the~~ feeling.
- This method estimates prob.t based on personal feelings, past experience, and judgment.

31/10/2010 * Example 5.1 :

a) P of getting number 4 :

$$= \frac{\text{Number of ways by which number 4 occurs}}{\text{number of all possible outcomes}} \\ = \frac{1}{6} = 0.167 = 16.67\%$$

b) an even number : 2, 4, 6

$$= \frac{3}{6} = 0.5 = 50\%$$

c) a non 4 number: 1, 2, 3, 5, 6 without "4"

$$= \frac{5}{6} = 0.833 = 83.33\%$$

$$\text{OR } P(\text{non 4}) = 1 - P(4) \\ = 1 - \frac{1}{6} = 0.833$$

31/10/2010

Example 5.2:

* 40 pens : 16 Red

10 blue

& Rest is black $= (16 + 10) - 40 = 14$

$$a) \text{Red Pen} = \frac{\text{Frequency of red}}{\text{Sample size}} = \frac{16}{40} = 0.4$$

$$b) \text{non black} = \frac{\text{frequency of red \& blue}}{\text{Sample size}} = \frac{16 + 10}{40} = 0.65$$

OR

$$P(\text{non black}) = 1 - P(\text{black}) \\ = 1 - \frac{14}{40} = \frac{26}{40} = 0.65$$

* Probability Rules:

* These rules are used to find the probability of compound events (more than one event occurring).

* There are two rules:

1. The addition rule.
2. The multiplication rule.

* The addition rule:

• This rule is used to find the probability of one event occurring or the other.

• There are two equation for this rule:

$$① P(A \text{ or } B) = P(A) + P(B)$$

$$② P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

3/10/2013

The Multiplication rule:

• This rule is used to find the probability of two or ~~over~~ more events occurring together.

• There are also two equations for this rule:

③ $P(A \text{ and } B) = P(A) \cdot P(B)$ "Independent"

④ $P(A \text{ and } B) = P(A) \cdot P(B|A)$ "dependent"

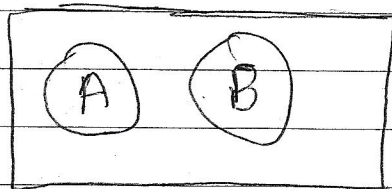
See, Example 5.3 :

Q/i. either a Nokia or an LG. [Addition rule]

$$\begin{aligned} \Rightarrow P(\text{Nokia or LG}) &= P(\text{Nokia}) + P(\text{LG}) \\ &= 7/20 + 3/20 = \frac{10}{20} = \frac{1}{2} \\ &= 0.5 \Rightarrow 50\% \end{aligned}$$

Note: The two events, Nokia & LG cannot occur together in same mobile.

• (Mutually exclusive events) \Rightarrow M. E. events.
That means: If one occur the other is out/not occur.



\Rightarrow M. E. events

$$P(A \text{ or } B) = P(A) + P(B)$$

$$P(A \text{ and } B) = 0$$

Q/ii. either Nokia or a mobile with a camera

$$\begin{aligned} \Rightarrow P(\text{Nokia or camera}) &= P(\text{Nokia}) + P(\text{mobile with a camera}) \\ &= 7/20 + 9/20 - 3/20 \\ &= 0.65 = 65\% \end{aligned}$$

$P(\text{Nokia with camera})$

$$\Rightarrow P(\text{Nokia or with camera})$$

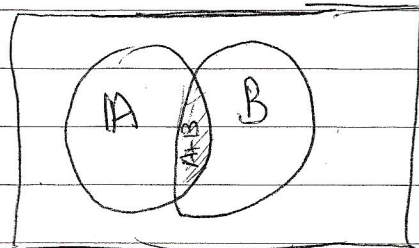
$$= P(\text{Nokia}) + P(\text{mobile with a camera}) - P(\text{Nokia with cam})$$

Note: The two events "Nokia mobile" and "Mobile with a camera" can occur together in the same mobile.

If two events occur together they are called: not mutually exclusive.

24

2/11/2017



* Events A & B are not M.E. events, can occur together.

* $P(A \cup B)$

$$= P(A) + P(B) - P(A \cap B)$$

* $P(A \cap B)$ is called a joint probability.

\Rightarrow we minus to avoid the double counting.

Example 5.3:

b/i Replaced in the box before drawing the 2nd.

* Drawing a mobile with replacement.

~~P(1st mobile is Samsung and the 2nd is LG)~~

$P(1^{\text{st}} \text{ mobile is Samsung and the } 2^{\text{nd}} \text{ is LG})$

$$= P(1^{\text{st}} \text{ Samsung}) \cdot P(2^{\text{nd}} \text{ LG})$$

$$= 10/20 \cdot 3/20$$

$$= 0.075 \text{ or } 3/40$$

Notes When we draw the first mobile & then we replaced ~~it~~ in the box, the drawing of second mobile is not affected by drawing the first mobile. In this case, the two events are independent.

b/ii not Replaced in the box before drawing the 2nd

* Drawing a mobile without replacement.

$P(1^{\text{st}} \text{ mobile is Samsung and the } 2^{\text{nd}} \text{ is LG}) =$

$$= P(1^{\text{st}} \text{ Samsung}) \cdot P(2^{\text{nd}} \text{ LG} | \text{the } 1^{\text{st}} \text{ is Samsung})$$

$$= 10/20 \cdot 3/19 = 0.079 \text{ or } 3/38$$

Notes When there is no replacement of the first mobile, the probability of the second event is changed ~~or~~ or affected, because now, we draw the second mobile out of 19 mobile only. In this case, the two events are dependent.

2/11/2010 In order to show that events are not independent we use "Slash" in the second probability.

* The slash means : — IF — assuming
— given — provided

Note / In Multiplication Rule :

$$* P(A \text{ and } B) = P(A) \cdot P(B|A)$$

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

Note : The multiplication rule for dependent event is as above.

& $P(B|A)$ is called conditional probability.

example * $P(K|L) = \frac{P(K \text{ and } L)}{P(L)}$ = both event
the event after slash.

$$* P(H|M) = \frac{P(H \text{ and } M)}{P(M)}$$

H — The event that we want to find the prot. for
M — The event that has already occur and affect the first.

2/11/2010 * Contingency Table 8

* This table summarizes the outcomes of an experiment according to two or more features in the form of rows and columns.

* Example 5.4 8

4/11/2010

	Single	Married	Divorced	total
Male	60	80	10	150
Female	32	160	8	200
Total	92	240	18	350

* Sample size = 350

* to avoid
Double
counting

(a) i: $P(\text{female}) = \frac{200}{350} = 0.571 \Rightarrow 57.14\%$

ii: $P(\text{female or a married person})$

$$= P(\text{female}) + P(\text{married person}) - P(\text{female married person})$$

$$= \frac{200}{350} + \frac{240}{350} - \frac{160}{350}$$

$$= 0.8 \Rightarrow 80\%$$

iii: $P(\text{single or a married person})$

$$= P(\text{single}) + P(\text{married person})$$

$$= \frac{92}{350} + \frac{240}{350} = 0.94 \Rightarrow 94.85\%$$

* after given
is divorced
given = slash
is before given
is male
→ condition

iv: $P(\text{male given that he is a divorced person})$

$$\Rightarrow P(\text{male} | \text{married}) = \frac{P(\text{Male and Divorce})}{P(\text{divorce})}$$

$$= \frac{(10 \div 350)}{(18 \div 350)}$$

$$= 0.56 = 55.56\%$$

4/11/2010

$$\begin{aligned}
 \text{vi: } P(\text{Married person if the employee is a female}) \\
 &= P(\text{Married} | \text{Female}) = \frac{P(\text{Married} \& \text{Female})}{P(\text{Female})} \\
 &= \frac{(160 \div 350)}{(200 \div 350)} = 0.8 \Rightarrow 80\%
 \end{aligned}$$

if is
a condition

Male
Not Divorced
Single
+ Married
= 60 + 80
= 140

$$\begin{aligned}
 \text{vi: } P(\text{Male assuming that the employee is not a divorced person}) \\
 &= P(\text{Male} | \text{Not divorced}) = \frac{P(\text{Male} \& \text{not divorced})}{P(\text{Not divorced})} \\
 &= \frac{(140 \div 350)}{(332 \div 350)} = 0.42 \Rightarrow 42.16\%
 \end{aligned}$$

Not Divorced
92 + 240
= 332
male + female

$$\begin{aligned}
 \text{B) } P(\text{Male} | \text{Not married}) &= \frac{P(\text{Male and not married})}{P(\text{Not married})} \\
 &= \frac{(70 \div 350)}{(110 \div 350)} = 0.63 \\
 &\Rightarrow 63.63\%
 \end{aligned}$$

Male & not
married
= 60 + 10
= 70
(single + Divorced)

Notes

- The number of employees = 0.636×350
= $222.6 \approx 223$
- Male given not married is 223 person;

Not married
= male + female
= 92 + 18
= 110

4/11/2010

[C]

i: The rule says: Event A & B are M.E events if $P(A+B) = 0$; otherwise they are not M.E events.

* Male and a divorced person are M.E events if $P(\text{male} \& \text{divorced person}) = 0$; otherwise they are not.

& From the table: $P(\text{Male} \& \text{Divorced}) = 10/350 = 0.0285$
 $\Rightarrow 0.029 \neq 0$ So they are not M.E events.

M.E
Mutually
Exclusive

ii: They are M.E events if $P(\text{male} \& \text{female}) = 0$
 otherwise they are not.

& From the table: there is no intersection between male & female, that means $P(\text{male} \& \text{female}) = 0$
 So, they are M.E events

7/11/2010

iii: They are Independent if:

The rule: events A and B are Independent if $P(A \text{ and } B) = P(A) \cdot P(B)$ or if the second rule: $P(A) = P(A|B)$, otherwise, they are not Independent.

* Female and married person are independent if $P(\text{Female and married}) = P(\text{Female}) \cdot P(\text{married})$

Table $\Rightarrow 160/350 = (200/350) (240/350)$
 $0.457 \neq 0.392$

The two sides of the equation are not equal; the event are not Independent.

* Or alternatively:

$$P(\text{Female}) = \frac{P(\text{Female} | \text{Married})}{P(\text{Female} \& \text{married})}$$

$$200/350 = \frac{160 \div 350}{240 \div 350}$$

$$0.571 \neq 0.667$$

- This means that, the two events are not Independent
- we can use one of the equation

7/11/2010

iv: The rule is A & B are complementary if $P(A) + P(B) = 1.0$, otherwise, they are not.

- Single and married are complementary events IF:
 $P(\text{single}) + P(\text{married}) = 1$, otherwise they are not.
 $92/350 + 240/350 = 1$
 $0.949 = 1$

- They are not complementary because they are not equals to 1.

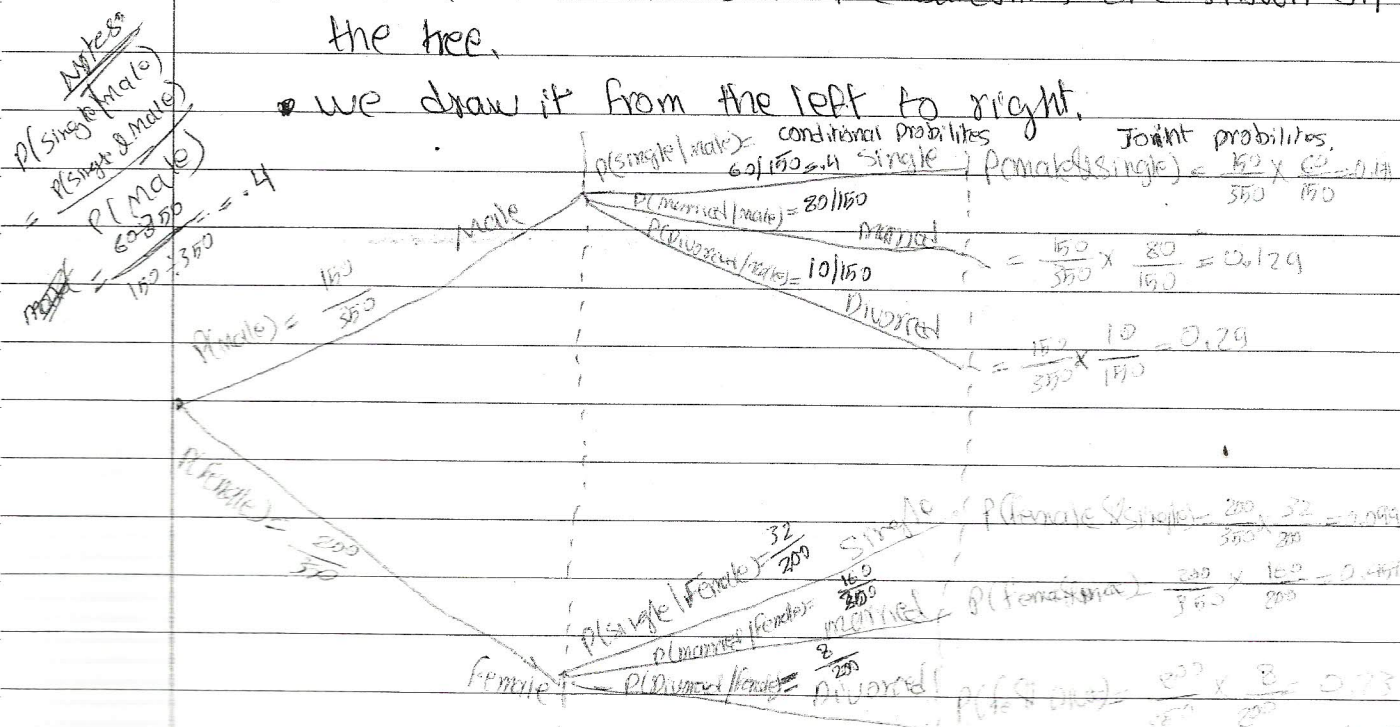
D $P(\text{Single} / \text{Male}) = \frac{P(\text{Single and Male})}{P(\text{Male})}$
 $= \frac{60 \div 350}{150 \div 350} = 0.4$

$P(\text{single} / \text{married}) = 0 \Rightarrow \text{M-E events}$

E Tree Diagrams

- Is a diagram that shows all the possible outcomes of an experiment on the form of tree branches.
- All the probabilities of the outcomes are shown on the tree.

- we draw it from the left to right.



9/11/2010

$$\rightarrow P(\text{single}|\text{male}) = \frac{P(\text{single} \& \text{male})}{P(\text{male})} = \frac{60 \div 350}{240 \div 350} =$$

$$\rightarrow P(\text{married}|\text{male}) = \frac{P(\text{married} \& \text{male})}{P(\text{male})} = \frac{10 \div 350}{150 \div 350} =$$

$$\rightarrow P(\text{divorced}|\text{male}) = \frac{P(\text{divorced} \& \text{male})}{P(\text{male})} = \frac{10 \div 350}{150 \div 350} =$$

$$\rightarrow P(\text{single}|\text{female}) = \frac{P(\text{single} \& \text{female})}{P(\text{female})} = \frac{32 \div 350}{200 \div 350} =$$

$$\rightarrow P(\text{married}|\text{female}) = \frac{P(\text{married} \& \text{female})}{P(\text{female})} = \frac{140 \div 350}{200 \div 350} =$$

$$\rightarrow P(\text{divorced}|\text{female}) = \frac{P(\text{divorced} \& \text{female})}{P(\text{female})} = \frac{8 \div 350}{200 \div 350} =$$

9/11/200

* Final notes:

* Events "a" and "b" are:

1) Mutually exclusive IF:

$P(a \text{ and } b) = 0$, otherwise they are not.

2) Independent IF:

$$P(a \text{ and } b) = P(a) \cdot P(b)$$

or $P(a) = P(A|B)$

3) Complementing IF:

$$P(a) + P(b) = 1, \text{ otherwise they are not}$$

$$* P(x \text{ and } y) = P(x) \cdot P(y)$$

If events are independent

9/11/2010 * Revision:

* Ex 1: for events "x" and "y", $P(x) = 4/7$ and for $P(x \text{ or } y) = 5/8$ and $P(x \text{ and } y) = 3/8$. Find $P(y)$

$$P(x \text{ or } y) = P(x) + P(y) - P(x \text{ and } y)$$
$$5/8 = 4/7 + P(y) - 3/8$$

$$0.625 = 0.571 + P(y) - 0.375$$

$$P(y) = 0.429 \Rightarrow \text{Three number after decimal value}$$

$$\hookrightarrow 0.625 - 0.571 + 0.375$$

$$0.625 - 0.146 = 0.429$$

* Ex 2: If K and L are independent events with $P(K) = 0.65$, $P(L) = 0.73$. Find $P(K|L)$.

$$P(K|L) = \frac{P(K \& L)}{P(L)}$$

$$= \frac{P(K) \cdot P(L)}{P(L)}$$

$$= \frac{(0.65)(0.73)}{0.73} = 0.65$$

$$P(K) = P(K|L) \Rightarrow 0.65$$

9/11/2010 Chapter 6 : Probability Distributions

- * For any experiment, we always study a certain variable.
- * This variable could be "discrete" or "continuous".
Whether, this variable is discrete or continuous it has a probability distributions.
- * Probability Distributions; means assigning or giving a probability to all the outcomes of the variable in the experiment according to a certain procedure.
- * Therefore, Probability Distributions can be either discrete or continuous depending on the variable we study.
- * The sum of the probabilities in a distributions should equals one.

11/11/2010

chapter 6 : Probability Distributions:

• Could be:

① Discrete

② Continuous.

• Discrete:

- ch. 6 {
1. The discrete population probability ~~the~~ distributions (DPPD)
 2. The binomial Distributions.
 3. The ~~Pois~~ Poisson //

• Continuous:

ch. 7 { 1. Normal Distributions.

ch. 8 { 2. The Sampling distribution of the sample mean (\bar{X} distribution)

#11: The discrete Population Probability Distribution (DPPD)
How to find: mean, Standard Deviation

Equations:

① The mean, or the expected value

$$\mu = \sum [x \cdot P(x)]$$

② The standard deviation of the distribution

$$\sigma = \sqrt{\sum (x - \mu)^2 \cdot P(x)}$$

③ $\sum P(x) = 1$

11/11/2010

Example 6.1 / Worksheet:

Ⓐ Expected Number (mean) = $\sum [x \cdot P(x)] \Rightarrow 1.25$

x	P(x)	x · P(x)	x - M	(x - M) ²	(x - M) ² · P(x)
0	0.15	0	-1.25	1.563	0.234
1	0.45*	0.45	-0.25	0.063	0.028
2	0.40	0.80	0.75	0.563	0.225
Total	1.0	1.25			0.487

* $1 - 0.15 + 0.40 = 0.45$

Ⓑ The Standard Deviation: $\sigma = \sqrt{\sum (x - M)^2 \cdot P(x)}$
 $\Rightarrow \sigma = \sqrt{0.487} = 0.698$

Note: Variance is $\sigma^2 \Rightarrow 0.487$

Ⓒ Probability:

i: Exactly 2 times a week:

$\Rightarrow P(x=2) = 0.40$ [From the table]

ii: At least one time a week: means that one is minimum, ≥ 1

$\Rightarrow P(\text{at least one}) = P(x=1) + P(x=2)$

From the table $= 0.45 + 0.4 = 0.85$

iii: Four times a week:

$\Rightarrow P(x=4) = 0$

because, the outcome "4" is not part of our experiment

Notes: * Probabilities could have the following probabilities:

① $P(\text{maximum of 1}) = P(x=0) + P(x=1)$

② $P(\text{no more than 1}) = P(x=0) + P(x=1)$

③ $P(\text{one or fewer}) = P(x=0) + P(x=1)$

④ $P(\text{at most one}) = P(x=0) + P(x=1)$

⑤ $P(\text{one or more}) = P(x=1) + P(x=2)$

0
1 } x
2 }
From the table

11/11/2010

⑥ $P(\text{more than 1}) = P(x=2)$

⑦ $P(\text{at least one}) = P(x=1) + P(x=2)$

or $= 1 - P(x=0)$

21/11/2010

[2] The ^{mean two} Binomial Distribution:

- This is another "discrete" distribution.
- Not any experiment is a binomial experiment
- In order for the experiment to be binomial, 4 conditions must be satisfied:

like

① There must be a fixed number of ^{trials} trials, "n" sample size

Cons:

② There are two outcomes for the experiment.

③ The outcomes are independent.

④ The probability of occurrence "success" and probability of non-occurrence "failure", remain constant. "fixed"

• Equations of Distribution: mean, Standard Deviation, probability

① Mean:

$$\mu = n \cdot p$$

② Standard Deviation:

$$\sigma = \sqrt{n \cdot p \cdot q}$$

③ Probability:

$$P(x) = {}^n C_x \cdot p^x \cdot q^{n-x}$$

Note that:

Probability of occurrence +
probability of non-occurrence
= one

$$\Rightarrow p + q = 1$$

! : Factorial

$$3! = 3 \times 2 \times 1$$

$$5! = 5 \times 4 \times 3 \times 2 \times 1$$

• Where, μ is the mean or expected value, & "n" is the sample size or trials & "p" is the probability of success.

& "q" is the probability of failure & σ is the standard deviation

& "x" is the variable we studied & different value of "x" means number of success obtained out of a trials.

• ${}^n C_x$, is the combination of "x" success out of "n" trials

$${}^n C_x = \frac{n!}{x!(n-x)!} \quad \text{So } n! = n(n-1)(n-2)(n-3) \dots 1$$

= "Factorial"

21/11/2018

* or by using the calculator. ~~find~~

* Or by using the Calculator.

$P(\text{smoking}) = 5\% \Rightarrow 0.05$

8. sample size = 10 "n"

23/11/2010

• Situation:

$$\textcircled{a} \mu = n \cdot p \quad \text{N sinderey}^n$$

$$= 10 \cdot 0.05 = .5$$

- Smoking here is "Success"

- The outcome which is given in the equation is "success"

$$\text{Standard Deviation } \sigma = \sqrt{npq} = \sqrt{10 \times 0.05 \times 0.95} = \sqrt{0.475} = 0.69$$

o Variance $\Rightarrow \sigma^2 = npq$
 $= 0.475$

23/11/2010

(b) $\mu = n \cdot p$

$= 10 \times 0.95 = 9.5$

"No smoking"

we use the probability of 0.95 as the probability of success; because, Success, here is not smoking

(i) at least 2 smoker: $[0 \ 1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 7 \ 8 \ 9 \ 10]$

that means, 2 & more... the success is "smoke" = 5%

So $\Rightarrow P(\text{at least 2 smoke}) = 1 - [P(x=0) + P(x=1)]$
 $= 1 - [{}^{10}C_0 (0.05)^0 (0.95)^{10} + {}^{10}C_1 (0.05)^1 (0.95)^9]$
 $= 1 - [1 \times 1 \times 0.6 + (10 \times 0.05 \times 0.63)]$
 $= 1 - [0.6 + 0.315]$
 $= 1 - [0.915] = 0.085 \Rightarrow 8.5\%$

(ii) Exactly 8 do not smoke: $[0 \ 1 \ 2 \ 3 \ 4 \ 5 \ 6 \ 7 \ 8 \ 9 \ 10]$

that means, $X=8$ exactly, the success "not smoking" = 95%

So $\Rightarrow P(x=8) = {}^{10}C_8 (0.95)^8 (0.05)^2$
 $= 45 \times 0.663 \times 0.0025$
 $= 0.075 \Rightarrow 7.45\%$

iii. None of them is smoke

$P(\text{none-smoke}) = {}^nC_x \cdot p^x \cdot q^{n-x}$
 $= {}^{10}C_0 (0.95)^0 (0.05)^{10}$
 $= 1 \times 0.6 \Rightarrow 0.6 \Rightarrow 60\%$

Exercise: the expected value of the binomial variable "x" is 16 and the variance is 12. Find the probability of success, p

$\mu = np = 16$

$\sigma^2 = np \cdot q$

$12 = 16 \cdot q$

$q = \frac{12}{16} = 0.75$

$p = 1 - q$

$= 1 - 0.75 = 0.25 \Rightarrow 25\%$

Notes

always equals sample size "n"

$P(x) = n \cdot C_x \cdot p^x \cdot q^{n-x}$

the power always equals the sample size "n"

sample size is $n=10$

sample space is 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

is all the possible outcomes & the zero is excluded

there is always a possibility of zero.

Poisson / Fish

23/11/2010

31 The Poisson Distribution

• This is another discrete distribution.

• It has the following features:

25/11/2010

- ① ~~It is~~ It is a special case of binomial distribution, where "n" is too large & "p" is too small.
- ② The sample size "n" can be given or not.
- ③ "X" represents the number of times & events happens within a certain interval "always given the interval".
- ④ The expected value "μ" is proportional "related directly" to the interval.

example:

<u>Interval</u>	<u>μ</u>
1 week	2
2 week	4
3 week	6

- ⑤ It is always positively skewed, "to the right".
- ⑥ The expected value "μ" equals the variance.
 $\Rightarrow \mu = \sigma^2$ OR $\sigma^2 = \mu$

Equations: mean, Standard Deviation, Probability

① Mean $\Rightarrow \mu = n.p$

② Standard Deviation $\Rightarrow \sigma = \sqrt{\mu} = \sqrt{n.p}$

③ Probability $\Rightarrow P(x) = \frac{\mu^x \cdot e^{-\mu}}{x!}$

Note: "e" is constant "fixed" which is equals to 2.71828

Note, variance = $\sigma^2 = \mu$

25/11/2010

Example / work sheet:

6.3

a) $\mu = n \cdot p$ & $\sigma^2 = \text{variance}$

$$\mu = 3.2$$

b) Standard Deviation $\Rightarrow \sigma = \sqrt{\mu} = \sqrt{3.2}$
 $= 1.79$

b) (i) Interval is week, $X=0$, what is "P"?

If the number of accidents is "X" then no accidents means $X=0$

Note: $P(x) = \frac{\mu^x \cdot e^{-\mu}}{x!}$

$$P(x=0) = \frac{(3.2)^0 \cdot e^{-3.2}}{0!} = \frac{(1)(0.0408)}{1} = 0.0408 = 4.07\%$$

two weeks
So, the " μ "
should be
doubled
etc

(ii) at most one accident occurs in a period of two weeks,

So,

$$P(\text{at most one in 2 weeks}) = P(x=0) + P(x=1)$$

$$P(x=0) = \frac{(6.4)^0 \cdot e^{-6.4}}{0!} + \frac{(6.4)^1 \cdot e^{-6.4}}{1!}$$

$$= \frac{(1)(0.0017)}{1} + \frac{(6.4)(0.00167)}{1}$$

$$= 0.0017 + 0.0106$$

$$= 0.012$$

Interval weeks	μ
1	3.2
2	6.4
4	12.8

25/11/2019

6.4

Variable is busy signal. $n = 750$ (given)at least one $\Rightarrow 0 \ 1 \ 2 \ 3 \dots 750$

$$\Rightarrow P(\text{at least one busy signal}) = 1 - P(X=0)$$

$$\Rightarrow 1 - \frac{\mu^0 \cdot e^{-\mu}}{0!}$$

Note that we don't have " μ " we should find it

$$\mu \Rightarrow \textcircled{1} \mu = nP \quad \checkmark$$

$$\text{or } \textcircled{2} \mu = 6^2$$

$$\mu = n \cdot P \quad \text{note: } P = 750 \text{ \& } P = \frac{2}{250}$$

$$= (750) \left(\frac{2}{250} \right)$$

$$= 1750 \cdot \frac{2}{250} = 6 = 750 \cdot \frac{2}{250} = 6$$

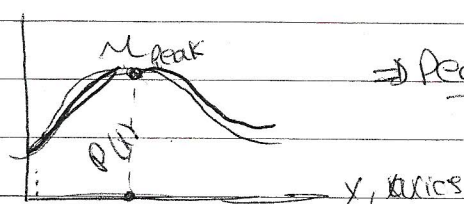
$$\text{So } \Rightarrow 1 - \frac{(6)^0 \cdot e^{-6}}{0!} = 1 - \frac{(1)(0.00247)}{1} = 1 - 0.00247$$

$$= 1 - 0.00247 = 0.998$$

28/11/2010 Chapter 7 : Normal Distribution

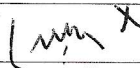

- This is a continuous Distribution. ~~It~~ It is the most important distribution, because many variables around us would follow normal Distribution.

- This distribution is known by its ~~smooth curve~~ "Normal curve"

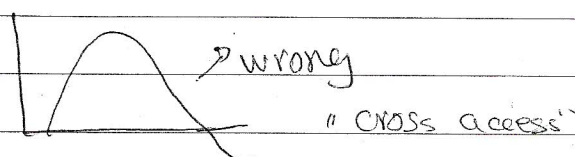
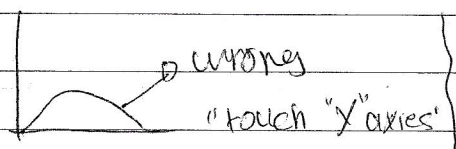


⇒ Peak μ mean = mode = median

- The normal curve have the following features:

- It is symmetric. (No skewness).
- It is smooth. (there are no up or down on curve) 
- It is bell-shaped with one peak (highest point). 
- The peak falls exactly at the mean, which is the same as median ~~th~~ and the mode.
- The tails of the curve come very close to the "x" axis but ~~we never~~ they never touch or cross the axes (Parallel)

(16)



- The area under the curve, represent the probabilities of the values taken by the variable.
- The total area under the curve equal 1 or 100%.
- The shape of the curve is determine by standard deviation, " σ " the ~~smaller~~ "σ" is the slimmer.
- The expected value or the mean can be either negative, zero, or positive.

28/11/2010

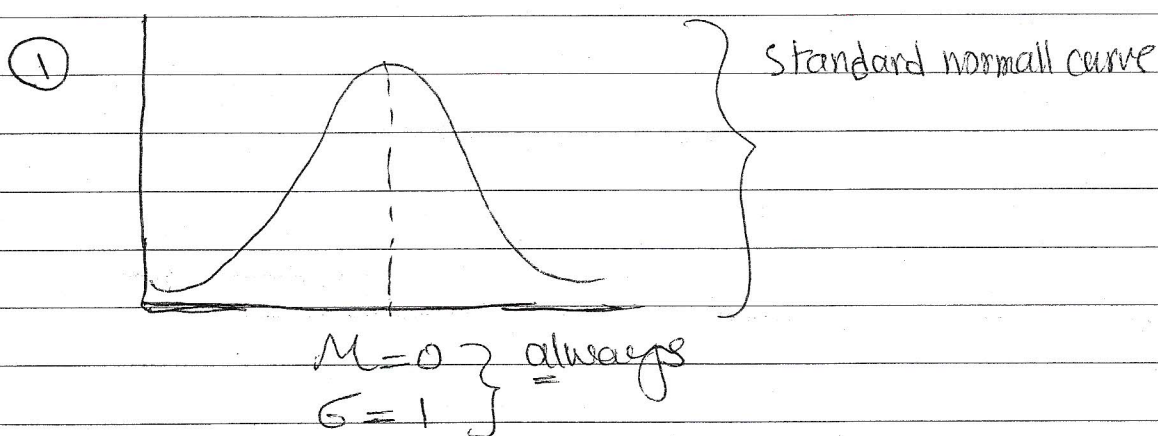
• Find the probability of a normal variable:

$$P(a \leq x \leq b) = \int_a^b \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Notes

• The above is difficult to use; there is an alternative way of finding the probability, using two items:

- ① Standard normal curve (Z-curve), always $\mu=0$, $\sigma=1$
- ② Z-table



• Any normal distribution value "x", can be transfer "chang to" to a standard normal curve; using the following equation

$$Z = \frac{x - \mu}{\sigma}$$

where "Z" is the standard normal value that will be used to find the probability.

& x, is the original normal variable

& μ , is the mean of normal distribution

& σ , is standard deviation of normal distribution

2/12/2010

* The value of Z , we find from the above equation is used on the Z -table to find the probability.

* The Z -table consists only of Z -value and the probability.

* Probability means under the curve.

Example 7.18

• variable: monthly sales

• mean: 1200

• Standard Deviation: 220

i) Between 1200 and 1400 units.

$$P(1200 \leq X \leq 1400)$$

① * First, transfer " X " value to " Z " by using $\left[Z = \frac{x - \mu}{\sigma} \right]$

So,

$$\Rightarrow P\left(\frac{1200 - 1200}{220} \leq Z \leq \frac{1400 - 1200}{220}\right)$$

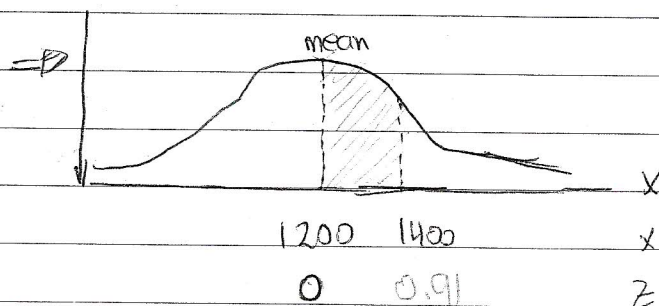
$$= P(0 \leq Z \leq 0.91)$$

Con. Notes

In other words you can not use the Z table to find the probability if Z does not start from 0 example.

② * Second, Sketch the curve

So,



③ * Third, See the Z -table

So,

$$\Rightarrow Z \begin{array}{|c|c|} \hline 0.0 & 0.1 \\ \hline 0.9 & 0.3186 \\ \hline \end{array} \text{ the probability is } = 0.3186$$

Notes: we always check the value of Z

Note inside the bracket to make sure that " Z " starts from zero, In this case you can use the Z -table to find the prob. or Area.

2/12/2010 ii) between 980 and 1100

• First, transfer "x" to "z"

So,

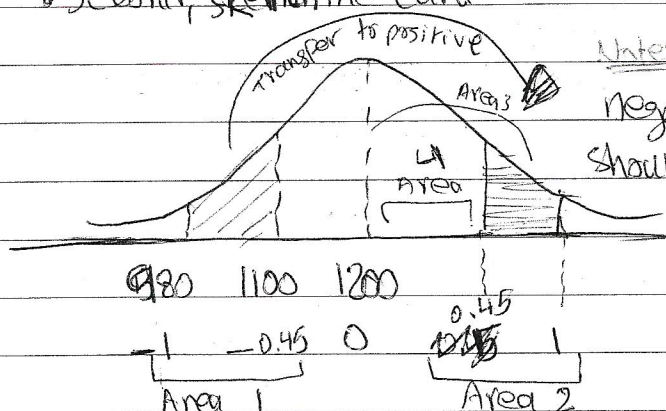
$$\Rightarrow P(980 \leq X \leq 1100)$$

$$= P\left(\frac{980 - 1200}{220} \leq Z \leq \frac{1100 - 1200}{220}\right)$$

$$= P(-1 \leq Z \leq -0.45)$$

$$\Rightarrow \text{After transfer: } P(0.45 \leq Z \leq 1)$$

• Second, sketch the curve



Notes: If the required area is in negative side of Z curve, it should be transferred to positive side of the curve.

5/12/2010

• Third, see the Z-table

So,

$$\Rightarrow \text{After transfer: } = P(0.45 \leq Z \leq 1) \Rightarrow \text{Area 2}$$

$$\text{Area 3} \Rightarrow P(0 \leq Z \leq 1)$$

$$\text{Area 4} \Rightarrow P(0 \leq Z \leq 0.45)$$

$$\Rightarrow \text{to find "Area 2"} \Rightarrow \text{Area 3} - \text{Area 4}$$

$$= P(0 \leq Z \leq 1) - P(0 \leq Z \leq 0.45)$$

$$\text{From the Z-table} = 0.3413 - 0.1736$$

$$\text{Table} = 0.1677$$

Notes 3 steps for finding the probability from the Z-table:-

① Transfer "x" into "z" using the Z-equation $Z = \frac{x - \mu}{\sigma}$

② Sketch the curve to show the required Area.

③ Check the value of "z" inside the brackets, If "z" starts from "zero" then use the 'z' table directly to find probability of "P"

5/12/2010 But, If "Z" does not start with zero, we look for other Areas that start from zero & can help us to find our area.

④ If the required area falls in the negative side of the "Z" curve, we transferred to positive side of the curve & then, go to step "3".

iii) Between 1000 and 1250

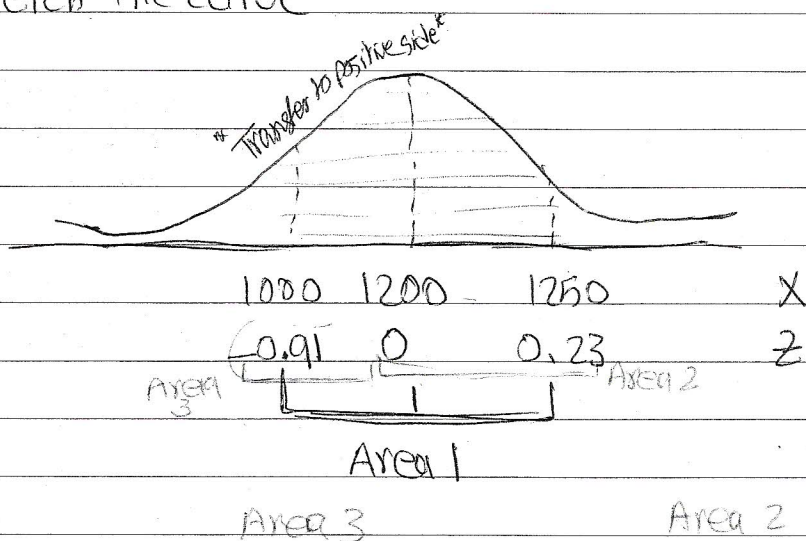
$z = \frac{x - \mu}{\sigma}$
 $S.D = 220, \mu = 1200$
 $= P(1000 \leq X \leq 1250)$

④ transfer it to "Z"

$$= P\left(\frac{1000 - 1200}{220} \leq Z \leq \frac{1250 - 1200}{220}\right)$$

$$= P(-0.91 \leq Z \leq 0.23)$$

② Sketch the curve



③

$$= P(-0.91 \leq Z \leq 0) + P(0 \leq Z \leq 0.23)$$

$$= P(0 \leq Z \leq 0.91) + P(0 \leq Z \leq 0.23)$$

$$= 0.3186 + 0.0910$$

$$= 0.4096$$

9/12/2010

9/12/2010 iv. less than 850

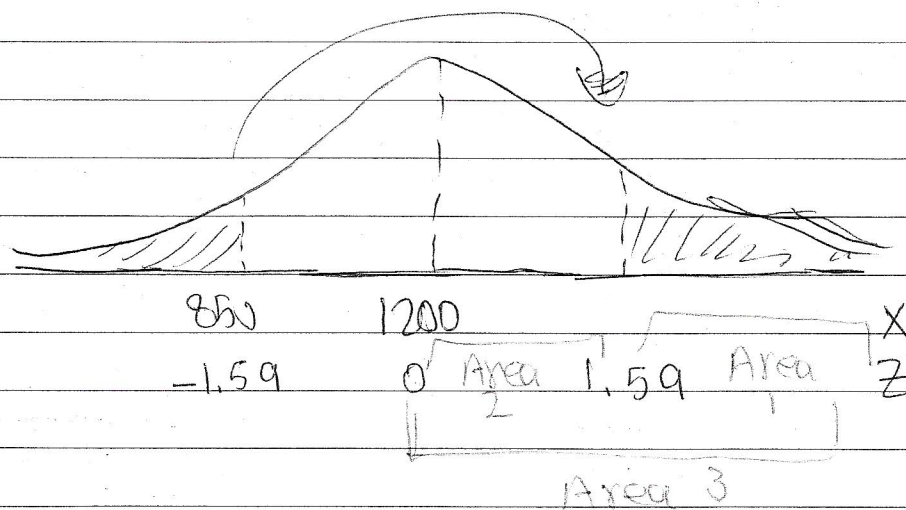
① transfer "X" to "Z":

$$P(X < 850) = P\left(Z < \frac{850 - 1200}{220}\right)$$

$$= P(Z < -1.59)$$

$$= P(Z > 1.59)$$

② Sketch the curve;



③ $= P(Z < -1.59)$

$$= P(Z > 1.59)$$

$$= 0.5 - P(0 \leq Z \leq 1.59)$$

$$= 0.5 - 0.4441 = 0.0559$$

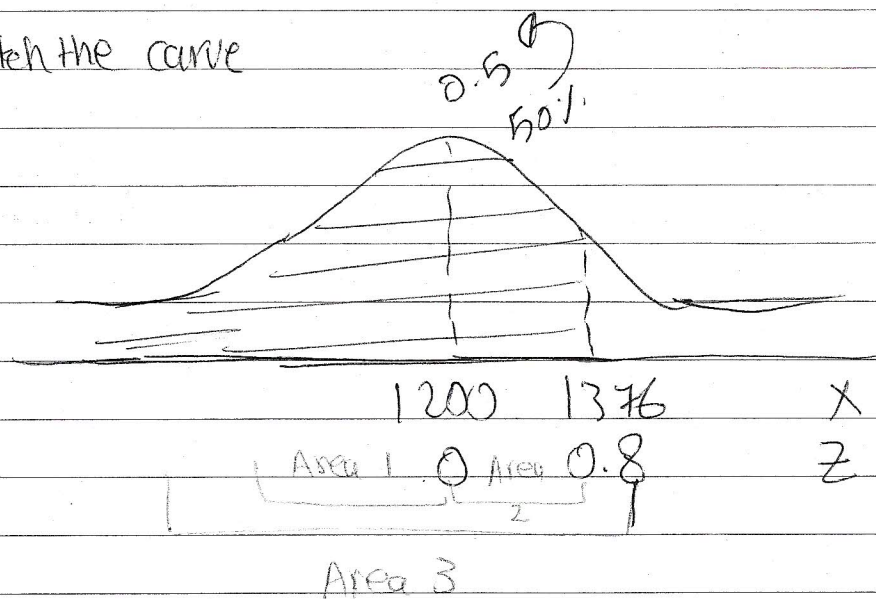
12
9/10/2010

U. less than ≤ 1376

① transfer "x" to "z"

$$\begin{aligned} P(X < 1376) &= P\left(Z < \frac{1376 - 1200}{220}\right) \\ &= P(Z < 0.8) \end{aligned}$$

② Sketch the curve



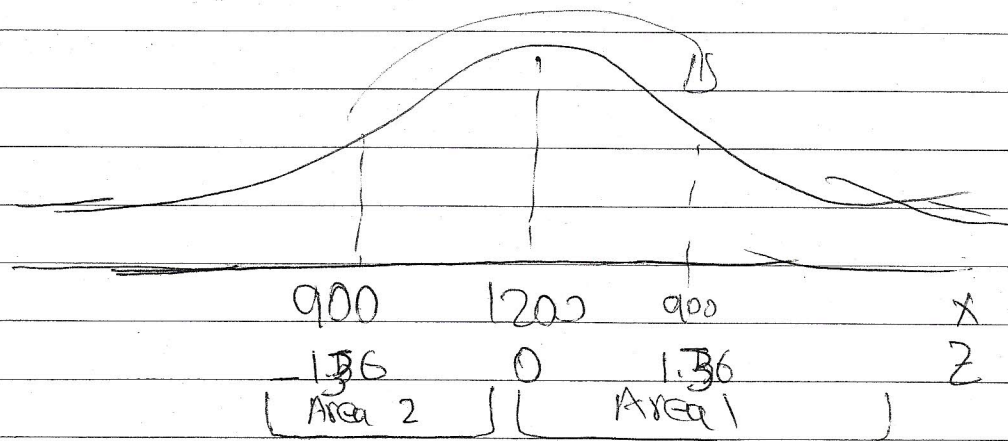
$$\begin{aligned} \textcircled{3} &= P(Z < 0.8) \\ &= 0.5 + P(0 \leq Z \leq 0.8) \\ &= 0.5 + 0.2881 \\ &= 0.7881 \end{aligned}$$

9/10/2010 VI more than 900

① transfer "x" to "z"

$$\begin{aligned} P(X > 900) &= P\left(Z > \frac{900 - 1200}{220}\right) \\ &= P(Z > -1.36) \end{aligned}$$

② Sketch the curve



$$\begin{aligned} \textcircled{3} \quad &= P(Z > -1.36) \\ &= P(-1.36 \leq Z \leq 0) + 0.5 \\ &= P(0 \leq Z \leq 1.36) + 0.5 \\ &= 0.4131 + 0.5 \\ &= 0.9131 \end{aligned}$$

9/12/2010 Part (B) \Rightarrow 7.1

Seq ②

①

① sketch the curve

$\bullet X_1$: max. no. of units.

X

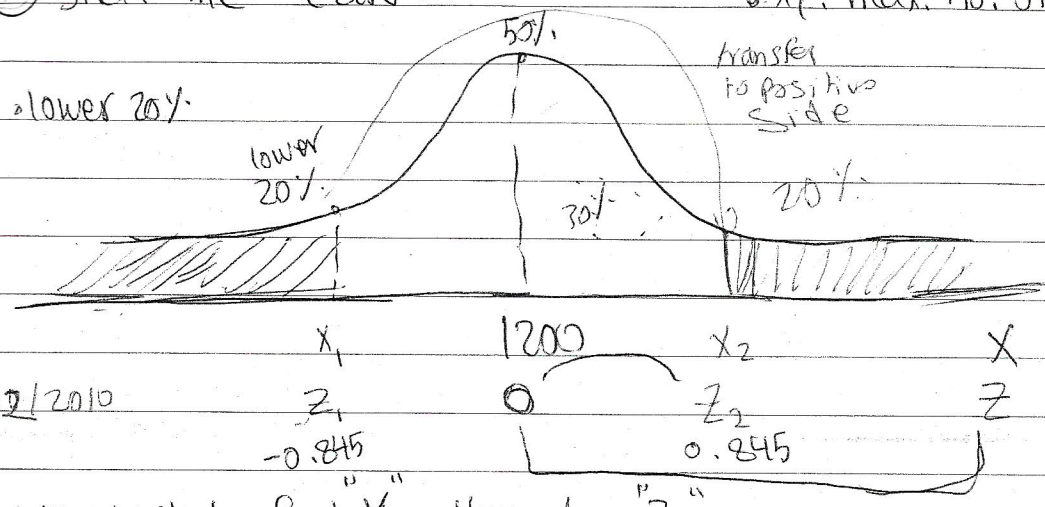
Equation

Z

table

Probab
(Area)

12/12/2010



\bullet we want to find " X_1 " through " Z_1 "

$$\Rightarrow P(0 \leq Z \leq Z_2) = 0.5 - 0.2 = 0.3000$$

\bullet IF $P = 0.3$, from the " Z " table, check the body not Z

$$P = 0.2995 \quad 0.3023$$

$$Z = 0.84 \quad 0.85$$

$$Z_2 = \frac{0.84 + 0.85}{2} = 0.845$$

$$Z_1 = -0.845$$

\bullet equation $Z = \frac{X - \mu}{\sigma}$

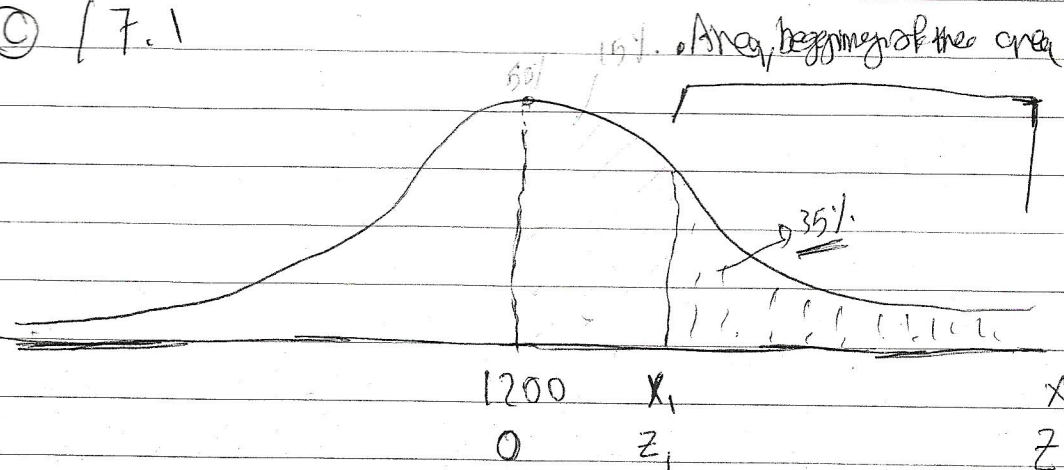
$$\Rightarrow -0.845 = \frac{X_1 - 1200}{220}$$

$$X_1 = -(0.845 \times 220) + 1200 = 1014$$

$$= \textcircled{1014} \text{ units } 1386.9 \Rightarrow 1386$$

max. no. of units for lower 20%.

12/12/010 Part © / 7.1



• we want to find "x," through "z."

$$\Rightarrow P(0 \leq z \leq z_1) = 0.5 - 0.35 \\ = 0.15$$

• $P = 0.1480, 0.1517$

$z = 0.38, 0.39$

$$z_1 = \frac{0.38 + 0.39}{2} = 0.385 \quad \text{"Average z"}$$

• using the equation

$$z = \frac{x - \mu}{\sigma}$$

$$0.385 = \frac{x_1 - 1200}{220}$$

$$x_1 = (0.385 * 220) + 1200 \\ = 1284.7 \text{ units "No need to round"}$$

12/12/2010

Chapter 8 :

The central limit theorem & the sampling distribution
& the sample mean (\bar{x}).

study
 \bar{x} as
a value
not x

the mean
of set of
Data (\bar{x})

... Brief ideas

* The sampling Distribution of the sample mean is:

* A continuous Distribution; for the means of all samples of equal sizes taken from a population.

* $\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4, \dots, \bar{x}_n$

The sampling Distribution.

* The mean of distribution: $M_{\bar{x}}$

$M_{\bar{x}}$: the mean of Distribution

$$\Rightarrow M_{\bar{x}} = \mu$$

where; ($M_{\bar{x}}$) is the mean of sampling distribution of sample mean \bar{x} . This Distribution is also known as:

" \bar{x} Distribution" \Rightarrow Short name.

(μ) is population mean.

14/12/2010

* The standard Deviation of the Distribution:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

where; $\sigma_{\bar{x}}$ is the standard Deviation of the \bar{x} & known as the standard error of their distribution.

• Probability of \bar{x} being between 2 values:

It has been found that if the sample size is large enough ($n \geq 30$), then the sampling distribution of the sample mean would follow the normal distribution.

This is known as "the central limit theorem"

theorem;
small theory

1 sample
normal
Distribution

- Once the sample size is large enough, we follow the steps of finding the probability of normal distribution to this distribution.

① \Rightarrow For this Distribution, the 'Z' equation is:

The equation is:

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}}$$

$$\Rightarrow \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Example sheet: 8.1

8.1/A: Prob. f; between 20 & 22 / Sample mean

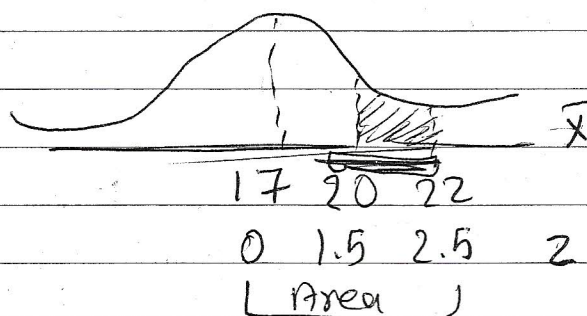
- Since, $n > 30$ the \bar{X} distribution we follow (apply, similar) the normal Distribution.

$$\Rightarrow P(20 \leq \bar{X} \leq 22)$$

$$= P\left(\frac{20 - 17}{12 / \sqrt{36}} \leq Z \leq \frac{22 - 17}{12 / \sqrt{36}}\right)$$

$$= P(1.5 \leq Z \leq 2.5)$$

- Sketch the curve



$$\begin{aligned}
 &= P(1.5 \leq Z \leq 2.5) \\
 &= P(0 \leq Z \leq 2.5) - P(0 \leq Z \leq 1.5) \\
 &= 0.4938 - 0.4332 \\
 &= 0.0606 \Rightarrow \text{Prob. f}
 \end{aligned}$$

14/12/2018 B) Standard error / Standard Deviation

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

$$= \frac{12}{\sqrt{36}} = \frac{12}{6} = 2$$

14/12/2010

Inferential
Stats

Chapter 9: Estimation and confidence intervals

- This chapter is about "Inferential Statistics" where we use the sample mean (\bar{X}) to estimate the population (μ).

- There are two methods of estimation:

① Point estimation:

This method depends on using a single value of " \bar{X} " as the value of " μ " (Samplely: $\bar{X} = \mu$).

② Confidence Interval estimation:

In this case, " μ " is estimated to fall between two values, i.e. μ is estimated using an interval of values; the interval can be written as follows:

$$\text{* lower limit} \leq \mu \leq \text{upper limit}$$

$$(\bar{X} - E) \leq \mu \leq (\bar{X} + E)$$

Note: "E" is the maximum allowable error (margin error).

i.e. "give or take two".

E \Rightarrow to calcn of "E" depends on two cases:

21/12/2010

Case ①: large sample

IF $n \geq 30$: then
$$E = \frac{Z \sigma}{\sqrt{n}}$$

• If " σ " is unknown then $E = \frac{Z S}{\sqrt{n}}$
use "S" instead of " σ "

• the value of "Z" depends on:

1 - Confidence level (CL):

2 - Probability, (P), ~~PA~~ ~~CL~~ ~~P~~ $P = \frac{CL}{2}$

3 - Z-table

21/12/2010

Case # ②: small sample

If $n < 30$, then $E = \frac{t \cdot s}{\sqrt{n}}$

• where "s" is assumed to be \sqrt{n} unknown (not given)

• the value of "t" depends on:

① the confidence level (CL)

② the number of degrees of freedom (df).

$$df = n - 1$$

③ t-table.

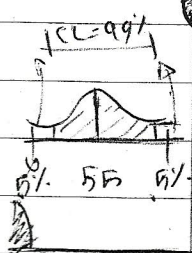
Example 9.1 / work sheet:

• $n = 49$, $\bar{x} = 55$, $s^2 = 100$, confidence level = 99%.

① the point estimate of the population mean is the sample mean \bar{x} . then $\mu = \bar{x} = 55$

② Since the sample size is large ($n = 49$), then we use the Z-distribution to estimate "E".

then: $E = \frac{z \cdot s}{\sqrt{n}}$, since "s" is unknown $\Rightarrow E = \frac{z \cdot \sigma}{\sqrt{n}}$



• To find Z:

① $CL = 99\% = 0.99$

② $P = CL/2 = \frac{0.99}{2} = 0.495$

③ Z-table. If $P = 0.495$

$Z = \frac{2.57 + 2.58}{2} = 2.575$

$$\Rightarrow 55 - \frac{2.575 \cdot \sqrt{100}}{\sqrt{49}} \leq \mu \leq 55 + \frac{2.575 \cdot \sqrt{100}}{\sqrt{49}}$$

$$\Rightarrow \bar{x} - E \leq \mu \leq \bar{x} + E$$

Interval $\Rightarrow 51.32 \leq \mu \leq 58.68$

• we are 99% confident, that the population mean is within the above interval.

21/12/2010 Example 9.2 / Worksheet

$n = 10$, $\bar{x} = 20$, $S = 5$, $CL = 95\%$ → use t-ratio

$$\bar{x} = \frac{\sum x}{n}$$

S.D. standard deviation

• Since the sample size small, we use the "t" distribution to find the error (E).

$$E = \frac{t \cdot s}{\sqrt{n}}$$

t depends on: t: from t distribution
s: Standard Deviation

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

un group data

① $CL = 0.95$

② $df = 10 - 1 = 9$

23/12/2010

③ t-table, the t-distribution. = 2.262

note that:

the question

might not

give you \bar{x}

and S

directly but

gives the single

data values from

them, you can

find \bar{x} and

S.D

10-1

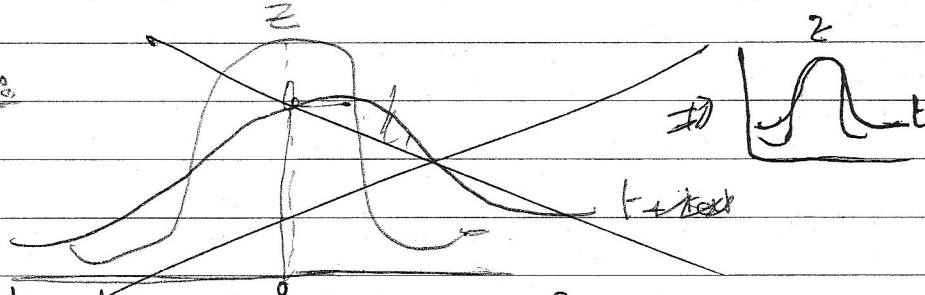
df = 9

CL = 0.95%

t: 2.262

$\bar{x} = 20$

Note:



• the t-distribution has these features:

① Symmetric and smooth.

② It has a mean of zero "like z distribution"

③ It is flatter "bigger"

④ As the sample size increased and ^{decreases} ~~resumes~~ to 30 & more the t-curve will become like the z-curve.

$$\bar{x} - \frac{ts}{\sqrt{n}} \leq \mu \leq \bar{x} + \frac{ts}{\sqrt{n}}$$

$$\bar{x} - E \leq \mu \leq \bar{x} + E$$

$$\Rightarrow 20 - \frac{2.262 \cdot 5}{\sqrt{10}} \leq \mu \leq 20 + \frac{2.262 \cdot 5}{\sqrt{10}}$$

$$20 - 3.576 \leq \mu \leq 20 + 3.576$$

$$16.42 \leq \mu \leq 23.58$$

important

always

write this

• we are 95% confident that the population mean μ is within the above interval.

23/12/2010 • It is not reasonable "acceptable" to conclude that the population mean of 25, because it is outside the interval: $16.42 \leq \mu \leq 23.58$

• Heading: Choosing the appropriate "right" Sample size:

Case one: If ~~large~~ sample size: $n \geq 30$

USE:

$$E = \frac{Z \sigma}{\sqrt{n}}$$

to find "n" sample size:

$$E \sqrt{n} = Z \sigma$$

$$\sqrt{n} = \frac{Z \sigma}{E}$$

$$\textcircled{1} n = \left(\frac{Z \sigma}{E} \right)^2$$

• If " σ " is not known, we will use " s " instead.

$$\text{or } \textcircled{2} n = \left(\frac{Z \cdot s}{E} \right)^2 \}$$

Notes: If you want to reduce "n" we could either:

① reduce confidence level to have lower value of "Z"

② Increase the allowable error "E".

& If you want to increase "n" we could either:

① Increase confidence level to have higher value of "Z"

② reduce the max allowable error "E".

E: max
allowable
error

23/12/2010

Example 9.3

$$\textcircled{1} CL = 90\% = 0.9$$

$$\textcircled{2} \text{Prop. } r = \frac{CL}{2} = \frac{0.9}{2} = 0.45$$

$$\textcircled{3} Z_{\text{table}}$$

$$\text{Prop. } r = 0.4495, 0.4505$$

$$Z = 1.64, 1.65$$

$$= \frac{1.64 + 1.65}{2} = \cancel{1.645} 1.645$$

$$n = \left(\frac{Z_{\alpha}}{E} \right)^2$$

$$= \left(\frac{1.645 \cdot \sqrt{0.45}}{3} \right)^2 = 43.3 \approx 44$$

Notes we rounded the number to least correct number like the "i" \Rightarrow Interval.

26/12/2010

Chapter 10: Hypothesis Testing

level of significance
if $CL = 80\%$
 $\alpha = 20\%$

• This chapter is about using " \bar{X} distribution" to test and assumption "Hypothesis" about the value of the population mean " μ ".

• The testing will be to decide whether " μ " is different, greater than; or less than a certain value.

• To do the testing we will use a large sample & a small sample; according to certain steps:

Example: 10.1 / work sheet:

3 types of testing

1. different

$n \geq 30$
 $n < 30$

2. greater than

$n \geq 30$
 $n < 30$

3. smaller than

$n \geq 30$
 $n < 30$

① Testing whether " μ " is different from 60,000 km using large sample of "50"

Steps: ① Set the Hypothesis.

$H_0: \mu = 60,000$: null

$H_1: \mu \neq 60,000$

اختلاف افتراف
اختلاف بزرگتر

Note: H_0 is the Hypothesis that want to test, whether the test true or not. It is also known as "the null hypothesis", the same as zero " H_0 ".

• H_1 is the alternative Hypothesis, that we accept if we reject H_0 .

• The question determines the sign of " H_1 " Hypothesis, where it is " $\neq, >, <, \geq, \leq$ ".

• Based on H_1 we can write H_0

~~② Find the test~~

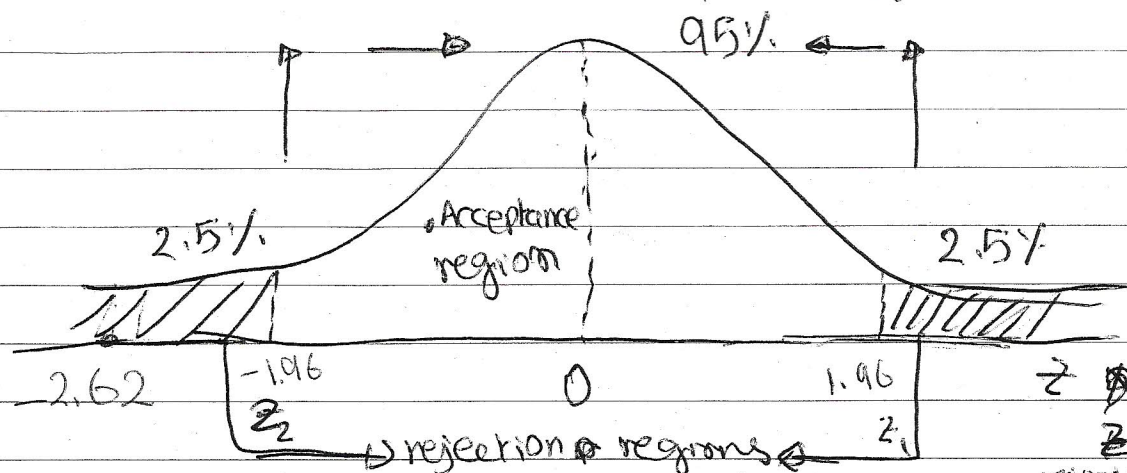
Null Hypothesis

226/12/2010

Step

(2) Find the critical values and the acceptance region.

- The critical values depends on sample size.
- Since we have large sample "n 730" than we use the Z - distribution to find the critical values.



- notes
- the level of significance α represent rejection regions
 - the confidence level (CL) represent acceptable region

$$P(0 \leq Z \leq Z_1)$$

$$= 0.5 - 0.025 = 0.4750 \text{ from Z-table}$$

$$Z_1 = 1.96$$

$$Z_2 = -1.96$$

& the acceptance region is from $[-1.96 \text{ to } 1.96]$
 ∴ these are the critical values.

Step (3) Make the decision rule

- If the Calculated "Z" is in the acceptance region, we accept the null hypothesis (H_0) otherwise, we accept (H_1)

Step (4) Find the Calculated "Z"

$$Z = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{58,000 - 60,000}{5400 \div \sqrt{50}} = -2.62$$

• This "Calculated Z" is different from "Z", because we will calculate it, but "Z" is ready from the Z-table, Also known as: "test statistic"

28/12/2010

Step ⑤ Make your decision/last step.

- Since the "calculated Z " $\Rightarrow (-2.62)$ is in left of the rejection region, we reject H_0 and take H_1 .
- Which means that $\mu \neq 60,000$, then the hre company claim is not true.

Notes By rejection " H_0 " we could make an error.
IF (H_0) was true.

• This error is called: Type I error

- The probability of making "Type I error" equals to level of significance (5%). is the " α "

Section
B/Work
sheet

(b) Testing wither the " μ " is different from 60,000 using a small sample of 27

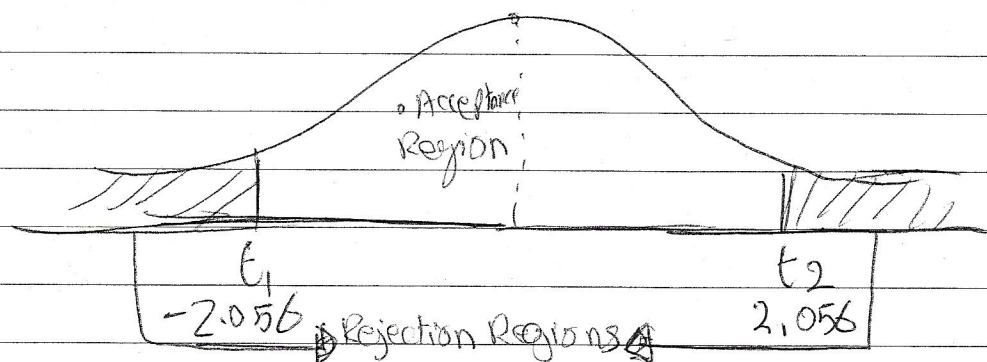
Small
sample

\downarrow
t-table

- Since the sample size " n " is small we use the "t curve & t-table" to find the critical values & acceptance region.

large
sample

\downarrow
Z-table



• To find the "table; t " we need/use:

- ① level of significance " α " = 5% = 0.05
- ② Degrees of freedom $\Rightarrow DF = n - 1 \Rightarrow 27 - 1 = 26$
- ③ t-table.

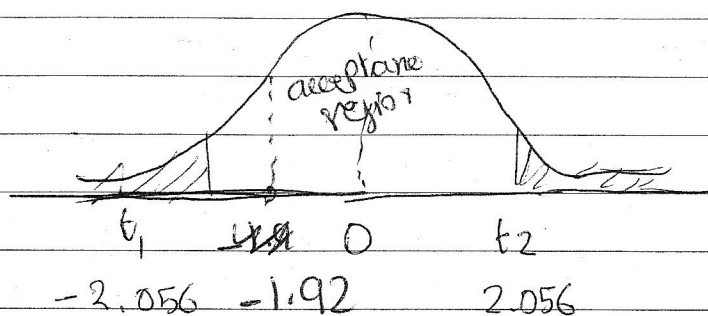
Notes ~~we~~ when we ~~use~~ testing wither " μ " is different from certain value, the rejection region " α " is divided equally between the two tails of the curve. Both of them are using on testing on

28/12/2010 \Rightarrow this case; this is called "the two tails" test; could be on one tail, right or left.

& This example of "the two tails":

\Rightarrow the critical values are $(-2.056 \text{ \& } 2.056)$
the acceptance region between: $(-2.056 \text{ \& } 2.056)$

30/12/2010



$$\Rightarrow t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{58,000 - 60,000}{5400 \div \sqrt{27}} = -1.92$$

• Since the Calculated 't' is in the acceptance region we accept the null hypothesis (H_0). ~~Acceptance~~

Note • by accepting " H_0 " we could make a "type I error"; IF H_0 is false, "that means IF we accept " H_0 " that ~~is~~ may be a mistake;" therefore there are probability of making error

• the probability of "type II error", is (Beta) β :

• type I occurs when we reject a true " H_0 ," the probability of this error equals level of significance " α "

• type II occurs when we accept a false " H_0 " the probability of this error equals Beta " β "

New
type
of error

α
" H_0 " is
null hypothesis

30/12/2010 ii) Greater than 60,000

$$n = 50, n = 27$$

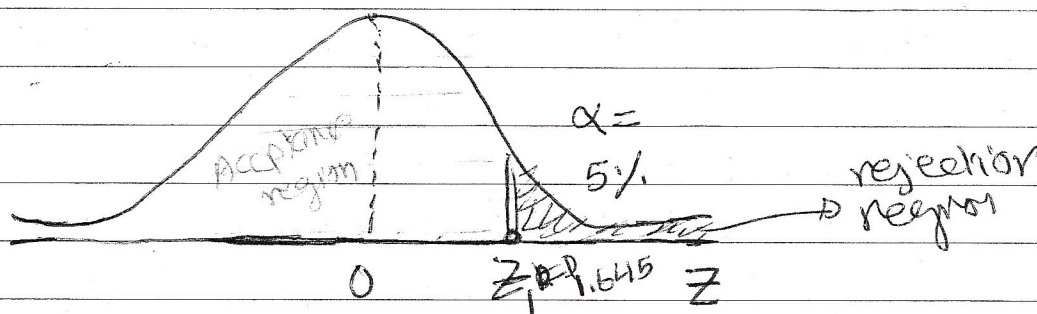
testing whether, μ is greater than 60,000 KM
using a large sample of 50; Z, table

Step ①: Set the Hypothesis:

- $H_0: \mu \leq 60,000$

- $H_1: \mu > 60,000$

Step ②: Find the critical values & acceptance region
"there is always a curve"



Notes: • the sign on " H_1 " points to the direction "Direction", the level of Significance " α ", rejection region

- $P(0 \leq Z \leq z_1)$

$$= 0.5 - 0.05 = 0.45 \text{ from "Z table"}$$

$$z_1 = \frac{1.64 + 1.65}{2} = 1.645 \text{ critical values}$$

Separate rejection from acceptance

- The acceptance region is all the area less than 1.645

Step ③ Make the decision rule:

- If the Calculated "Z" is in the acceptance region we accept the null hypothesis " H_0 " otherwise we accept " H_1 "

Step ④ Find the Calculated Z

I reject
 $G \alpha$

$$Z = \frac{\bar{X} - \mu}{S / \sqrt{n}} = \frac{58,000 - 60,000}{5400 \div \sqrt{50}} = -2.62 \quad \text{on acceptance region}$$

II accept
 $G \beta$

Step ⑤ Make your decision

- Since the Calculated Z is in acceptance region, we accept the H_0 this means $\mu \leq 60,000$.

Notes: • By accepting " H_0 " we could make type II error if H_0 is false

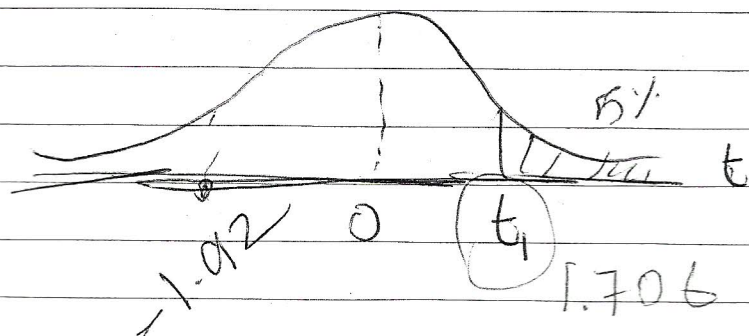
- the probability of making error is β "Beta",

• Testing whether μ is greater than 60,000 KM using a small sample of 27

- Since the sample small we use the t-curve & t-table to find the critical values & the acceptance region

• to find the "t" we use;

- ① level of significance " α " = 5%.
- ② Degree of Freedom $DF = n - 1 \Rightarrow 27 - 1 = 26$
- ③ t table for one tail test



3/11/2011

* $t_c = 1.706$

• this is the critical values

• the acceptance region is all the area below 1.706

$$\text{Calculated } t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

$$= \frac{58,000 - 60,000}{5400/\sqrt{27}} = -1.92$$

• since the Calculated "t" is in the acceptance region we accept H_0

• this means that " μ " is equal or less (\leq) 60,000

• By accepting " H_0 " we could make "type II" error if " H_0 " is false

• the probability of making error is β (Beta).

iii/ less than 60,000 KM

* Testing whether " μ " is less than 60,000 KM using a large sample of 50

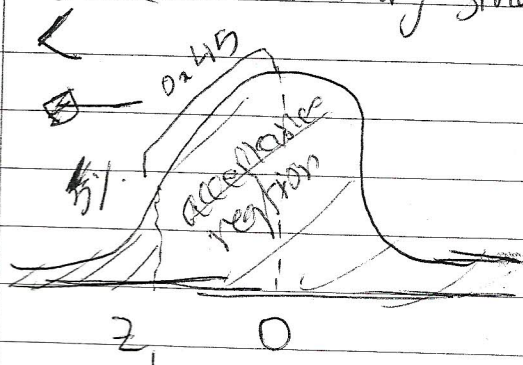
① Step 1; Set the Hypothesis

• $H_0: \mu \geq 60,000$

• $H_1: \mu < 60,000$

② Step 2: Find critical values & acceptance region

• Since this is a large sample we use the "Z" curve & table.



~~Step 2~~

• $M.P. = 0.45$, from "Z" table

$$Z = \frac{1.64 + 1.65}{2} = 1.645$$

$$Z_1 = -1.645$$

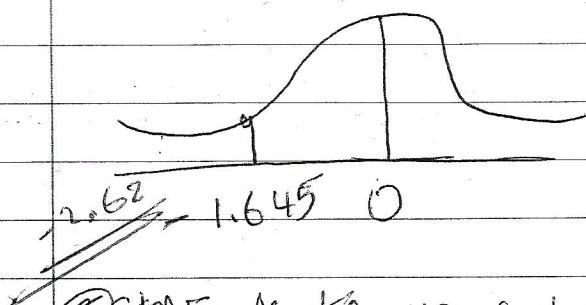
• the acceptance region is all the area greater than -1.645

3/1/2011 (3) Step 3: Make decision rule

• If the Calculated Z is in the acceptance region we accept H_0 , otherwise we accept H_1 .

(4) Step 4: Calculated Z

$$Z = \frac{\bar{X} - \mu}{s / \sqrt{n}} = \frac{58,002 - 60,000}{5400 / \sqrt{50}} = -2.62$$



(5) Step 5: Make your decision

- Since the Calculated Z is in the rejection region, we reject H_0 and accept H_1 .
 - That means μ is less than 60,000.
 - By accepting H_1 & rejecting H_0 we could make an error of type I.
 - & the probability = $\alpha \Rightarrow 5\%$.
- testing whether μ is less than 60,000 RM using a sample size of 27
- Since the sample is small we use the "t" curve & table to find the critical value & acceptance & rejection.
 - to find "t" we use:
 - (1) level of significance = $\alpha \Rightarrow 5\%$.
 - (2) Degree of freedom = $DF = n - 1 = 27 - 1 = 26$
 - (3) t-table of one tail.

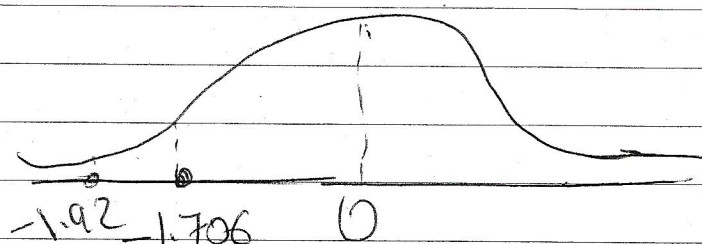
3/1/2011

$$t = 1.706, \Rightarrow t_1 = -1.706$$

- the acceptance region all the area more than -1.706

- Calculated t :

$$t = \frac{\bar{X} - \mu}{S / \sqrt{n}} = \frac{58000 - 60000}{5400 / \sqrt{27}} = -1.92$$



- since the Calculated " t " in the rejection area, we ~~do~~ reject H_0 & accept H_1 .
- this means that μ less than 60,000
- by rejection H_0 we could make type I & have probab of $\alpha = 5\%$.

3/11/2011

Chapter 13: Regression and Correlation.

- This chapter is about describing the relationship between two variables X and y in the form of an equation.

- This equation is known as regression equation which has the following form

$$\Rightarrow \hat{y} = a + bx \quad \text{If the relation is direct}$$

$$\text{OR } \Rightarrow \hat{y} = a - bx \quad \text{If the relation is indirect}$$

Where, \hat{y} is the estimated or predicted value of y
 x is the independent variable that makes change in y .

- b is the slope of the line; that shows the relationship between x and y
- a is the y -intercept; this is the amount of y -axis which is cut by the line

- All the work on this chapter is about how to find the values of a & b

- Sometimes there is ~~relationship~~ relationship between x & y \Rightarrow ① Direct; ① S & P;
② Indirect; ② D & P;

- Sometimes there is ^{no} relationship between x & y
Like: age & Price of Oil
GPA & Inflation of India.

\hat{y}
 y_{hat}

\checkmark
regress
go
back

3/1/2011

Example 13

Direct (plus)

	* 2	1	3	4	Adver. exp.
* 7	3	8	10	Sales Rev.	

Section (a): Develop a regression equation and Interpret the regression coefficients

 Σ (total)

X, Adver. exp.	y, Sales rev.	X^2	y^2	$X \cdot y$
2	7	4	49	14
1	3	1	9	3
3	8	9	64	24
4	10	16	100	40
$\Sigma X = 10$	$\Sigma Y = 28$	$\Sigma X^2 = 30$	$\Sigma Y^2 = 222$	$\Sigma XY = 81$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{10}{4} = 2.5$$

$$\bar{Y} = \frac{\Sigma Y}{n} = \frac{28}{4} = 7$$

$$SS_{XX} = \Sigma X^2 - \frac{(\Sigma X)^2}{n} = 30 - \frac{10 \times 10}{4} = 5$$

$$SS_{YY} = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n} = 222 - \frac{28 \times 28}{4} = 26$$

$$SS_{XY} = \Sigma XY - \frac{\Sigma X \Sigma Y}{n} = 81 - \frac{10 \times 28}{4} = 11$$

$$b = \frac{SS_{XY}}{SS_{XX}} = \frac{11}{5} = 2.2 \quad \text{"Slop"}$$

$$a = \bar{Y} - b\bar{X} = 7 - (2.2 \times 2.5) = 7 - 5.5 = 1.5 \quad \text{"y-intercept"}$$

$$\text{The regression equation} \Rightarrow \hat{Y} = a + bX$$

$$\Rightarrow \hat{Y} = 1.5 + 2.2X$$

This equation is used to estimate the value of "y" for any given value of "x"

3/10/2021 • the regression coefficients are a & b

* Interpretation of a & b

• $a = 1.5$

• this means that the sales revenues equals 1.5 m BD, even if there is no advertising expenses.

• In the absence of advertising expenses, the company can still make 1.5 million BD, as sales revenues,

• $b = 2.2$

this means an increase of one unit in "x" can make (cause) an increase of 2.2 units in "y"

i.e. If you spend \$1 as (x) the (y) increases by (2.2×1) in y

Section (B) Find the estimated sales revenue if adv. exp = 5m

* $\hat{y} = 1.5 + 2.2x$

\hat{y} is the estimated value of y

$\hat{y} = 1.5 + (2.2 \times 5) = 12.5$

Vol 8, Pt 1 Section (C) Calculate the correlation ^{coef} & comment on it

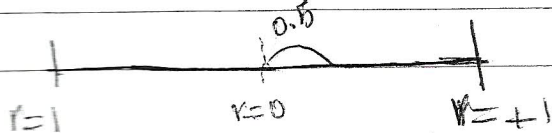
• Correlation coefficient r are to measure the weak and strong of x & y

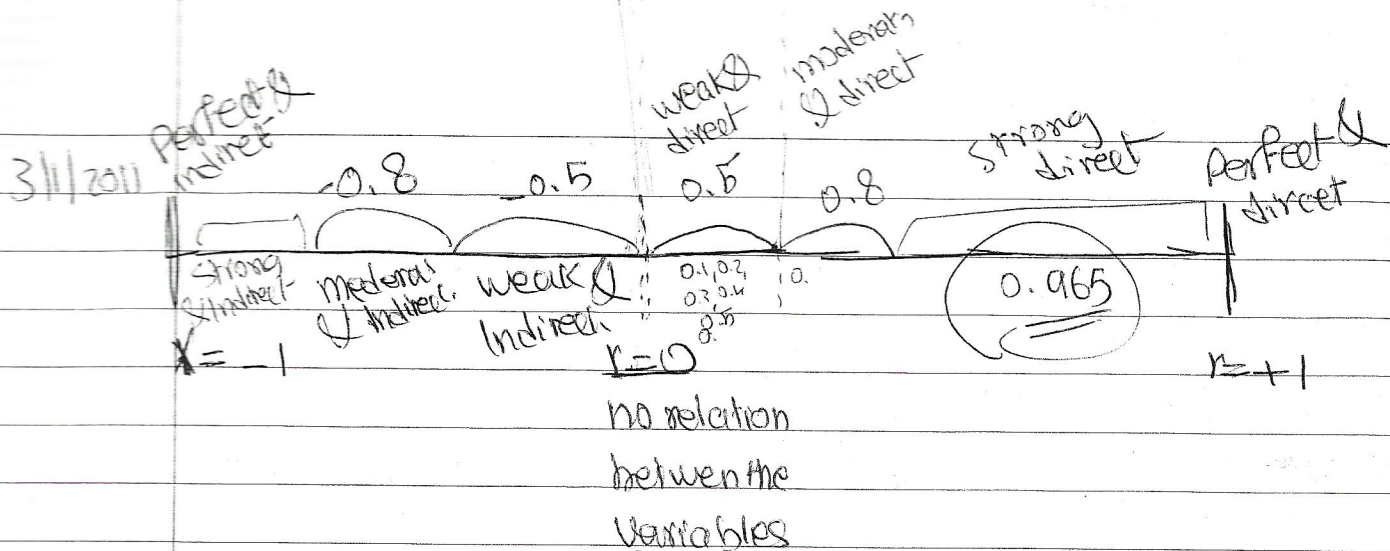
•
$$r = \frac{SS_{x,y}}{\sqrt{SS_{xx} \cdot SS_{yy}}} = \frac{11}{\sqrt{5 \times 26}} = \frac{11}{\sqrt{130}} = 0.965$$

+1 that x only the factor to efficiency

• $-1 \leq r \leq +1 \Rightarrow$ only -1 or $+1$ or between them.

• $r = 0$, when no relation between variables
• $r = +1$, perfect direct
• $r = -1$, perfect indirect





From the value of r , we understand that, the relationship between x & y is direct & Strong. not (x)

Section (D) Find the determination coefficient & explain

- Determination coefficient = $r^2 \Rightarrow 0 \leq r^2 \leq 1$
 $= (0.965)^2 = 0.931 \times 100 = 93.1\%$

Section (C) How much is the unexplained variation in sales revenue.

- 93.1% of the variation, in y , can be explain by the variation in x .

- So, 93.1% change on x will effect the y by 93.1%

another words

- the variation of " x ", is responsible for 93.1% variation on y .

Note: that, the total variation on $y = SS_y = 26$

93.1% can be explained by " x "

So: 26 total $\Rightarrow 0.931 \times 26 = 24.21$

6.9% unexplained by " x "

$\Rightarrow 0.069 \times 26 = 1.79$

Note: Total variation in y = explained variation + unexplained variation
 $= 93.1\% + 6.9\% \Rightarrow 100\%$
 as a percentage

3/11/2011 → Section (f): Draw the regression line on a scatter diagram

4/11/2011 → Section (e): Unexplained Variation in Sales revenue

Unexplained variation

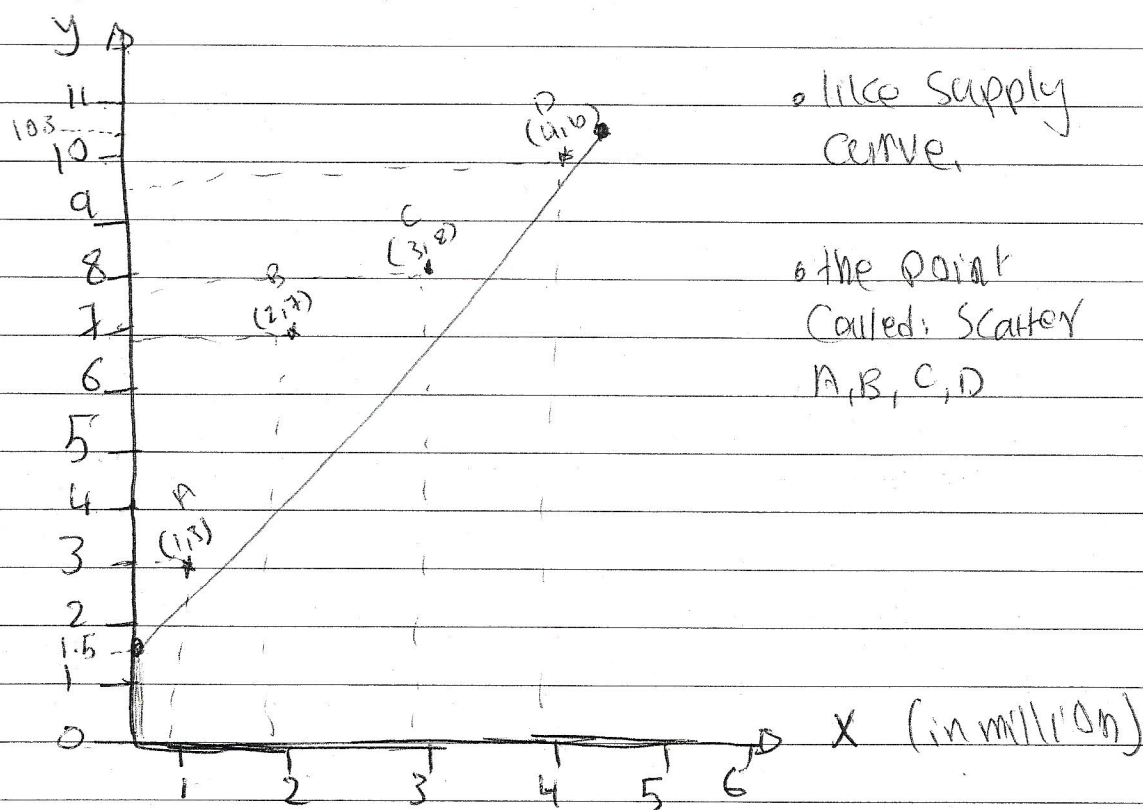
$$= \text{Total variation} - \text{explained variation}$$

$$= 26 - 24.21$$

$$= 1.79$$

Section (f): Draw the regression line on a scatter diagram

It is a relationship between X & Y



always starts from zero & than the highest one

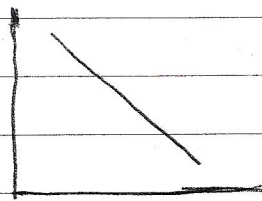
Y		X	
1.5		0	
10.3		(4)	
Y = 1.5 + 2.2X		X	
regression line equation		The purpose of the regression line; is to use it in estimating or predicting the value of "Y" if we know the value of "X". This is the same purpose as the regression equation.	
1.5 + 2.2(0)		0	
1.5 + 2.2(4)		(4)	
used to estimate Y through X, draw a line of curve			

4/11/2011

Notes

- If the relationship between x & y is indirect, the regression equation will be on the following form:
 $\Rightarrow y = a - bx$.

- the correlation coefficient (r) in this case will be a negative side; the regression line will look like.



- like the demand curve.

4/11/2010

Ex. chapter 9 & 10:

① Which of the following can be used as a point estimate of the population μ ?

- a) $\frac{x}{n}$ ☒ b) $\frac{\sum x}{\sqrt{n}}$ ☒ c) $\frac{\sum x}{n}$ ☒ d) $\frac{\sum x}{N}$ ☒ e) s

• Answer: C

② Which value does ^{the} null hypothesis ~~make~~ (H_0) make a ~~claim~~ claim about?

- ☒ a) Population Parameter " μ "
☐ b) Sample statistic
☐ c) type type I
☐ d) type I I

• Answer: A

③ Which of following is true about null hypothesis (H_0)?

- ☐ a) $\mu > 60,000$
☐ b) $\mu < 60,000$
☐ c) $\mu \neq 60,000$
☒ d) $\mu = 60,000$

• Answer: D

④ Which of following represent the critical value in hypothesis testing?

- ☐ a) the calculated "Z" or "t"
☐ b) the rejection region
☐ c) the acceptance region
☒ d) tabulated "Z" or "t" "from the table"

• Answer: D

$H_0: \mu = 60,000$
 $H_1: \mu \neq 60,000$

different

$H_0: \mu = 60,000$
 $H_1: \mu \neq 60,000$

greater than

$H_0: \mu \leq 60,000$
 $H_1: \mu > 60,000$

less than

$H_0: \mu \geq 60,000$
 $H_1: \mu < 60,000$

different

greater

less

4/11/2011 : Lind's
 Finish : 2/2/2011
 89