Bus 216: Business Statistics II

Introduction

Business statistics II is purely inferential or applied statistics.

Study Session 1

1. Random Variable

A random variable is a variable that assumes numerical value associated with the random outcome of experiment where one and only one numerical value is assigned on each sample point, in other words, a variable whose level is determined by chance is called a random variable.

When a specific outcome is uncertain is treated as a variable. The random variable assumes actual numeric value after all relevant outcomes are known. Many different terms of random variable can be generated by a single situation.

Mathematically, a random variable is a function that we use to map the elementary events unto their corresponding point of the numerical scale.

The random variable is a fundamental concept in applying probability theory to decision making. The relationship between the value of a random variable and their probabilities is summarized by probability distribution.

1.1 Discrete Random Variable

Random that can assume a countable number of values are called discrete random variable.

1.1.1 Binomial Theorem

Is an experiment conducted by a man called Binomial and named after him as Binomial theory. A Binomial Situation to be recognized by the following characteristics.

The experiment consists of "n" repeated trials each trial results in an outcome that may be described as a success or a failure.

The probability of a success denoted by "p" remains constant from trial to trial. The repeated trials are independent.

Binomial Probability Formular:

$^{n}C_{2}\ p^{x}\ q^{n-x}$

p = probability of success

x = what we are looking for or the sample point

q = probability of failure

n = Number of repeated trial

c = Combination sign

$\binom{n}{x}\ p^{x}\ q^{n-x}$

$\binom{n}{x} = \frac{n!}{(n-x)!x!}$

X = 0, 1, 2, 3----n

Illustration

Suppose a patient has 40% chance of surviving a surgical operation, granted that they are five repeated trials what is the probability that:

(a) Three trials will be successful?

(b) Three trials will not be successful?

Solution

n = 5, x = 3, p = 4%, p = 40%q = 60%

$^{n}C_{x}\ \ p^{x}\ \ q^{n-x}\ \ =\ \ C^{5}{}_{3}\ \ \ (0.40)^{3}\ \ (0.60)^{5-3}$

$$\left(\begin{matrix}(0.40)\ 0.60) \\ \frac{5!}{(5-3)!3!}=\frac{5\times4\times3\times2\times1}{2\times1\times3\times2\times4}=10\end{matrix}\right)$$

(a) $10\times0.40^{3}\ \ \times0.60^{2} = 10\times0.064\ \ \times0.36 = 0.2304$. probability = 0.2304×100 = $\underline{\underline{23\%l}}$

(b) Will not be successful

1-0.2304 = 0.7696 ×100 = $\underline{\underline{77\%}}$

(c) More than 3 will be successful

n=5 p = 40% q= 60% × =>3

Assume × = 0,1,2,3 - - -n

$^{5}C_{0}\ \ (0.40)^{0}\ \ (0.6)^{5-1} = 0.0778$

$^5C_1$ $(0.40)^1$ $(0.6)^{5-1}$ = 0.2592

$^5C_2$ $(0.40)^2$ $(0.6)^{5-2}$ = 0.3456

$^5C_3$ $(0.40)^3$ $(0.6)^{5-3}$ = 0.2304

0.0778+ 0.2592 + 0.3456+ 0.2304

= $\underline{\underline{0.913}}$

X >3 = 1-0.913 = $\underline{\underline{0.087}}$

1.1.2 Poisson Distribution

To use the Binomial distribution we must be able to count the number of successes and the number of failures; where as in many applications, you may be able to count the number of successes, you often cannot count the number of failures example:

The divisional police officer in charge of Gwagwalada could easily count the number of district calls he responded to in one day. How can he determine how many district calls he did not receive, obviously in this case. The number of possible outcomes (success plus failures) is difficult if not impossible to determine. If the trial outcome of possible outcome cannot be determined then the Binomial distribution cannot be applied as a decision making aid fortunately the Poisson probability distribution can be applied in this situation. To apply Poisson distribution we need to only know the average number of successes for a given segment

Characteristics of Poisson distribution

A physical situation must possess certain characteristics before it can be described by the Poisson distribution.

1. The physical event must be considered rare event
2. The physical events must be random and independent from each other an event occurrence must not be predictable nor can it influence the chances of another even no occurring.

The Poisson distribution is described by a single parameter- lamder which is the average occurrence per segment. The value of lamder depend on the situation being described, for example ,lamder could be the average number of machines breaks down a month or the average number of customers arriving at a check out sound a ten minutes period

Once lamder has been determined; we can calculate the average occurrence for multiple segments = lamder T

lamder and T must be in a compatible unit. If we have lamder equals to 20 arrivals per hour, we cannot multiply this by a time period measured minute's i.e.

ƛ = 20/hr t= 30 min –must be in compatible, 10hrs or minutes.
ƛ t= (20) ½ = 10
ƛ t= N – (the average)

Poisson Dis – formular ; $p(x) = ƛ\frac{t^{x} –e-}{X!} ƛ^{t}$

or $\frac{e-m-mx}{x!}$

Where ƛt = expected number of successes in segment “T”

x= t = Nπnumber of successes in segment “T”

e= 2.71828 –Based on a natural number.

1.2 Continuos Random Variable

Normal Distribution

Normal distribution is the name given to a types or distribution of continuous date that occur. Frequently it is a distribution of natural phenomenon such as weight, distance, time.

Characteristics of Normal Distribution

1. It has a symmetry curve about the mean of the distribution. One half of the curve is a mirror image of the other.

2. The diagrams of the values take to cluster about the mean with the greatest frequency at the mean itself.

3. The frequency of the values tapper away (symmetrically) either side of the mean giving the curve a characteristic bear bell shape. The 2 score for any value of normal distribution having the mean and standard deviation:

$Ƶ = \frac{x-N}{Y}$ Ƶ is the standardization, x = what we are looking for

$Ƶ = \frac{x-N}{Y} = \frac{8-6.4}{1.2} = \underline{\underline{1.33}}$

Ƶ = 1.33 = 0.908 (probability: 908)

Illustration

A time taken to complete a particular job is distributed appropriately normally with mean 1.8 hrs and standard deviation 0.1hrs.

If normal time work finishes at 6:00pm and a job is started at 4:00pm what is the probability that the job will need over time payment.

What estimated completion time is the nearest minute should be set so that there is a 90 chance that the job is completed on time.

The time taken for a particular job is approximately normal distribution in time or on time –find this out

Solution

We are looking for two hrs ie 6:00-4:00

$Ƶ = \frac{x-N}{Y} = \frac{2-1.8}{0.1} \quad \frac{0.2}{0.1}$

ƻ = 2

Prob 0.9772 = 97.7% - the prob that the job will not need over time payment Prob that the job will need an overtime payment = 1-09772= 0.02 =2% 90% supposedly ƻ = convert to probability nor above 0.90 but could be assumed as 1.2 =ƻ $\frac{\substack{(x)-findx\\ N}}{\text{ɤ}}$

Normal Approximation Binomial

In a binomial situation when “N” is large the “P” is not too small or large. The normal distribution can be used to approximate to Binomial. The mean of the normal approximation to binomial

It small but not too small it’s < than 50 i.e. 40.

μ=np n> 30

ɤ2 = npq

ɤ = $\sqrt{npq}$

Illustration

From past records 40% of a firm order is for export. The record for export is 48% in a particular financial quarter. If they expect to satisfy about 80 orders in the next financial quarter, what is the probability that they will break their previous export record.

If we put an order for export as success; this is a Binomial situation with trial (n) = trial = order.

n= trial = order. Trial success = an order for export. Number of trials i.e n= 80, p the probability of a success = 40%

Since n is large and p is not too small we can use normal approximation to binomial our μ = np = 80 (0.4) =32

$\sqrt{npq} = \sqrt{80x\,0.4\,x\,0.6 = 4.4}$

Low, in order that the previous record is brought to the firm needs more than 48% of 80 orders.

48% of 80 = 0.48 (80) = 38.4

ƶ = $\frac{x-N-mean}{T}$ = $\frac{38.4-32}{4.4}$

ƶ = 1.45

ƶ = $\underline{\underline{0.9265}}$

Study Session 2

2. Confidence Limit

Confident limit specify a range of value within which some unknown population value (mean or value) lies with a stated degree of confidence, they are always based on the result of a same.

2.1 Confidence Interval for a Mean

Given a random sample from some population, a confidence internal from an unknown population mean is

s

$\bar{x} \pm$ ƻ $\sqrt{n}$

Where $\bar{x}$ = is the sample mean

S= is the sample. Sample standard deviation

n= sample size

ƻ = confidence factor

64 for 90%

76 for 95%

58 for 99%

s

Where $\bar{x} \pm$ ƻ $\sqrt{n \qquad}$ is known as the standard error of the mean

Suppose a sample of 100 invoices a mean gross have of ₦3.24k

= 100 = 45.50

= 3.24 1.96 – is assumed 95% confidence

$\bar{x} \pm$ ƻ $\sqrt{n}^{\,s}$ = 45.50 ± 1096 $(\frac{3.24}{100})$

45.50$\bar{x}$ 1.96 (0.324)

45.50 $\bar{x}$ 0.635

45.50 – 0.635 = 44. 865

45.50 + 0.635 = 46-135

2.2 Confidence Limit for a Proportion

When a random sample from some population confidence interval from the unknown population:

P $\bar{x}$ ƻ $\frac{\sqrt{p(1-p)}}{n}$

Where p is the sample proportion

₦ n is the sample size

ƻ = confidence factor

Assumed 1.64 for 90%, 1.96 for 95%, 2.5 for 99%

Note that the conjugation $\frac{\sqrt{p(1-p)}}{n}$ is known as the standard error

Suppose 4,40 component are discovered in a random sample at 20 finished component taken from a production line what statement can we make about the defective rate of all finished component, assuming a 95% confidence interval.

First find the proportion: - in a sample of 20, 4 are 40 = 4/20 = p = 0.2 i.e out of 20, 4 components

n

P $\bar{x}$ ƻ $\frac{\sqrt{p(1-p)}}{n}$ 0.2 ± 1.96 $\frac{\sqrt{0.2(1-0.2)}}{20}$ = 0.2 ± 1.96 (0.085)

= 0. 2 ±0.175

2 ± 0.175 = 0. 2+ 0. 175 = 0.375

0.2 – 0.175 = 0.025

Same five was tested on 21 similar cars identical conditions fined consumption was found to have a mean of 41.6m by with a D of 3.2mbg. only 14 of cars were found completely satisfy a currents exhaust emission confidence .

Study Session 3

3. Test of Significance

Test of significance are directly connected to confidence limit and are based don normal distribution concepts of tests weather a sample of size “n” with mean and standard deviation “s” could be considered as having being drawn a population with a mean “N”. The test statistics = ƻ = $\frac{\bar{x}\text{-u}}{s}$

must lie within the range _1.96 to +1.96 $\frac{s}{n}$ in the test, we are looking for evidence of a different between population mean and the sample mean.

The evidence is found if ƻ lies outsides the above stated limit (i.e -1.96 to + 1.96) if ƻ lies within the limit, we say; `there’s no evidence that the sample mean is different to the population mean`.

Illustration

A manager is convinced that a new type of machine does not affect production at the company’s major shop floor. In order to test this, twelve sample of the weekly hourly output is taken and the average production per hour is measured as 11,58 with a S.D 71 given that the output per hour averaged 1196 before the machine was introduced, test the manager’s conviction.

Note that the sample is measuring now the population is measuring what was before thus, evidence of a difference between the sample population will show evidence of a change

$$= \frac{\frac{1158-1196}{71}}{12}$$

$$\therefore \quad Ƶ = \frac{\bar{X} - N}{\left(\frac{S}{\sqrt{n}}\right)}$$

Ƶ = -1.86 - it is within -1.9600 + 1.96

Thus there is no evidence of any difference between sample and population mean. The manager's diction is therefore supported

Illustration

Is life of electric life bulbs from a particular manufacturer have an average life of 800 hours. A sample of 25 bulbs was taken and found to have mean life of 850, life with S.D 80hours. Is there evidence that the sample bulbs have perform differently to the population norm.

$$Ƶ = \frac{\bar{X} - N}{\left(\frac{S}{\sqrt{n}}\right)}$$

$$Ƶ = \frac{850-800}{\left(\frac{80}{\sqrt{25}}\right)} = Ƶ = 3.\ 125$$

ƻ = ?

$\overline{X}$ = 850 hours

ນ = 800 hours

S = 80 hours

n = 25 bulbs

ƻ = $\frac{850-800}{\left(\frac{80}{\sqrt{25}}\right)}$ = $\frac{50}{\left(\frac{80}{5}\right)}$ = $\frac{50}{16}$

ƻ = $\frac{50}{16}$, ƻ = 3.13

ƻ = -1.96 + 1.96 = 3.13 is outside this limit, which means, there's evidence that there's a difference that the sample bulb perform differently from the population norm.

(b) what is the large whole number value the sample mean would have taken to reverse this decision:

-1.96 - +1.96

ɤ = 1.96

$\overline{X}$ = ?

ນ = 800 hour

n = 25

s = 80 hours

$1.96 = \frac{\bar{X} - 800}{\left(\frac{80}{\sqrt{25}}\right)}$

$1\text{-}96 = \frac{\bar{X} - 800}{\left(\frac{80}{5}\right)} \qquad = \frac{\bar{X} - 800}{(16)}$

$1.96 = \frac{\bar{X} - 800}{(16)}$

Cross multiply

$1.96 \times 16 = \bar{X} - 800 = 31.36 = \bar{X} -800$

Inter change

$\bar{X} - 800 = 31.36$

Collect the like term

$\bar{X} = 31.36 + 800$

$\bar{X} = 831.36$ hours

That means this is the value the sample mean would have taken to reverse the earlier decision.

Study Session 4

4. Statistical Decision Theory

Very often in practice we are called upon to make decision about the decision on the basis of sample information. Such decisions are called statistical decision. For example we may wish to decide on the basis of sample, whether a new serum is really effective in curing a disease whether one educational procedure is better than another or whether a given point is coin is loaded

4.1 Statistical Hypothesis

In attempting to reach decision, it is used to make assumptions (or guesses) about the population involved such assumptions which may or may not be true are called Statistical Hypothesis. They are generally statements about the probability distribution of the population.

Types of Statistical Hypothesis

There are basically two types of statistical hypothesis. These are:
Null hypothesis
Alternate hypothesis

Null Hypothesis: in many instance we formulate a statistical hypothesis for the sole purpose of rejecting or nullifying it. For example if we want to decide whether a given coin is head we formulate the hypothesis that the coin is fair i.e. P = 0.5 where P is the probability of getting x. similarly, if we want to decide whether one procedure is better than another, we formulate the hypothesis that there is no difference between the procedure. (i.e. any observed differences were due mainly to fluctuations in sampling from the same population) such hypothesis are often call Null hypothesis and are denoted by “H_o”. in order words a null hypothesis states that there is no difference between population parameters and that

any difference found between sample statistics are due to chance variation and are of no importance.
In other words - $H_o : \bar{x}_1 = \bar{x}_2$

This means there's no difference between the two population means. The null hypothesis simply says that the two phenomena are the same.

Alternative Hypothesis

Any hypothesis that differs from a given hypothesis is called an alternative hypothesis.

The hypothesis alternative to null hypothesis is denoted as H_1: $\bar{x}_1 \neq \bar{x}_2$

Alternative hypothesis takes that there is a significant difference between two population parameters
i.e. $H_1 : \bar{x}_1 \neq \bar{x}_2$

Test of Hypothesis

If we suppose that a particular hypothesis is true but found that the result observed in a random sample differ markedly from the results expected under the hypothesis i.e. (expected on the hypothesis of pure chance, using sampling theory). Then we would say that the observed differences are significant and we thus will be inclined to expect the hypothesis (or at least not accepting it on the basis of the evidence obtained).

Procedure that enable us to determine whether to accept or reject hypothesis are called

Type I And Type Ii Errors

If we reject the hypothesis when it should be accepted, we say that a type I error has been committed if on the others hand we accept the hypothesis when it should be rejected we say that a type II error has been made either case, a wrong decision or error judgment has occurred

Level of Significance

Testing a given hypothesis, the maximum probability with which we should be willing to risk a type one error is called LEVEL OF SIGNIFICANCE or SIGNIFICANCE LEVEL. This probability often denoted by "α" is generally specified before any sample or samples are drawn, so that the result obtained will not influence our choice

Study Session 5

5. Regression Analysis

Regression is a technique used to describe a relationship between two variables in mathematical term. Regression is concerned with obtaining a mathematical equation which describes the relationship between two variables. The equation can be used for comparison purposes or estimation purposes. Regression analysis or curve fitting is a statistical technique which can be used for medium term for casting, which seeks to establish the line of best fit to the observed data.

The process of obtaining a linear regression or relationship for a given set of bivariate data is often refers to as fitting the regression line. There are three methods commonly used to find a regression line for a given set of bivariate data, these are;

Inspection: This method is simple and consist of fitting a scattered diagram for the data and then drawing in the line that most suitable for the data. The main disadvantage of this method is that different people will probably draw different line using the same data. It sometimes helps to plot the main point of the data 91.e. the mean of x and y respectively) and ensure the regression line passes through this.

Semi-Averages: The techniques consist of splitting the data into two equal groups, plotting the mean joint for each group and joining these two points with straight line.

Least Square: This is considered to be standard method of obtaining a regression line; the derivation of the technique is purely mathematical.

Illustration:

The table below shows the age x and the symbolic Blood pressure of y of twelve (12) women.

Age (x)	**BP (y)**	**Xy**	**X^2**	**Y^2**
56	147	8,232	3,136	21,609
42	125	5,250	1,764	15,625
72	160	11,520	5,184	25,600
36	118	4,248	1,296	13,924
63	149	9,387	3,969	22,201
47	128	6,016	2,209	16,384
55	150	8,250	3,025	22,500
49	145	7,105	2,401	21,025
38	115	4,370	1,444	13,225
42	140	5,850	1,764	19,600
68	152	10,336	4,624	23,104
60	155	9,300	3,600	24,025
628	1,684	89,416	34,416	238,822

(a) Determine the least-square regression equation of y and x

(b) Estimate the blood pressure of a woman whose age is 45 years

(c) Estimate the blood pressure of a woman whose age is 75 years

(d) Find the correlation coefficient between x and y square

Regression equation is given by: $y = a + bx$

Where:

y is a dependent variable.

a is a constant and an intercept of y axis

b is a slope of the regression line

x is an independent variable

Where $b = \frac{n\sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$

And $a = \frac{\sum y}{n} - \frac{b\sum x}{n}$

Solution

$b = \frac{(12)(89894) - (628)(1684)}{12\,(34416) - (628)^2}$

$b = -1.14$

$a = \frac{1684}{12} - 1.14\,\frac{(628)}{12}$

$a = 80.67$

(a) To determine the least square regression of y on x

Solution

$Y = a + bx$

$Y = 80.67 + 1.14x$

or

$a_n + b\sum x = \sum y$

$a_{12} + 628 = 1684 \text{ x } 628$ --- equation (i) a 628 + b34416 x 12 ....equation (ii)

Using elimination method:

a7536 + b394384 = 1,057,552

a7536 + b412992 = 1078728

Subtract equation I from equation II

b18608 = -21176

Making b the subject of the formula

divide both sides by 18608

$$= \frac{b18608}{18608} = \frac{-21176}{18608}$$

b = 1.14

Study Session 6

6. Correlation Technique

Correlation is a technique used to measure the strength of the relationship between two variables. Correlation is concerned with describing the strength of the relationship between two variables by measuring the degree of `scatter` of the data values.

6.1 Product Moment Correlation

Formula:

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}}$$

$$r = \frac{1{,}078{,}728 - 1{,}057}{\sqrt{412{,}992 - 394{,}384} \times \sqrt{2{,}865{,}864 - 2{,}835{,}856}}$$

$$r = \frac{21{,}176}{\sqrt{18{,}608} \times \sqrt{30{,}008}}$$

$$= \frac{21{,}176}{136.4 \times 173.228} = \frac{21{,}176}{23{,}628.32}$$

r = 0.896 - which means that there's a strong elation between the range of x and y is called product moment correlation

6.2 Spear Man Rank Correlation

An alternative method of measuring correlation based on the ranks of the sizes of the item values is available and known as rank correlation most commonly used is known as spearman's rank correlation coefficient:

1. Rank the values of x
2. Rank the values of y

Note that ranking of the x values are performed quite independent of the ranking of the y values and the ranking is normally performed in an ascending order.

3. for each spear of rank calculate d^2

$d^2 = (rx - ry)^2$

4. calculate summation d^2

$$r^2 = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

Study Session 7

7. Time Series Analysis

A time series is a name given to the value of some statistical variables measured over a uniform set of term point. Any business large or small sales, purchase, the value of sock heed and VAT, this could be recorded daily, weekly, monthly, quarterly, or yearly.

7.1 Component of Term Series

1. **Seasonal variation:** This is a periodic rice and fall in sales that leads or repeat itself annually.
2. **Trend:** The long term tendency of the whole sales to rise and fall
3. **Cyclical factors:** Is the periodic rise and all of the whole sales over a number of years a long term.
4. **Random or Residual Variation:** Is the remaining variation in the data, which cannot be attributed to the components of time series (i.e. 1,2,3).

7.2 Coefficient of Determination

This measure calculate word proportion of variation in actual values may be provided by ranges in values of x