Business Statistics - Course Summary

Business Statistics (University of Technology Sydney)

Business Statistics

Course Summary

26134

Table of Contents

Lecture 1 – Introduction to Stats, Data and Graphing Data

Chapter 1: Introduction to Statistics

Sources of Data:

Primary data: data collected to address a specific need.

Secondary data: data that was collected for some other purpose, and is thus already available.

If secondary data is available, we should use it – as collecting primary data will be expensive, and wasteful if the data (or very similar data) is already available.

Nature of Data:

Qualitative: dealing with characteristics of objects – such as gender, grade, colour, country.

Quantitative: dealing with numbers, numerical data. These numbers are usually arising from some sort of scale.

Definition of Statistics:

'the science that deals with the collection, classification and use of numerical facts or data, bearing on a subject matter'.

Statistics has two branches – **descriptive statistics** and **inferential statistics**. To understand these, we must understand the following definitions:

Population: a collection of persons, objects or items of interest. E.g 'cars'

Census: a process of gathering data from a whole population for a given measurement of interest.

Sample: a portion of the whole, which, if taken properly, is representative of the whole.

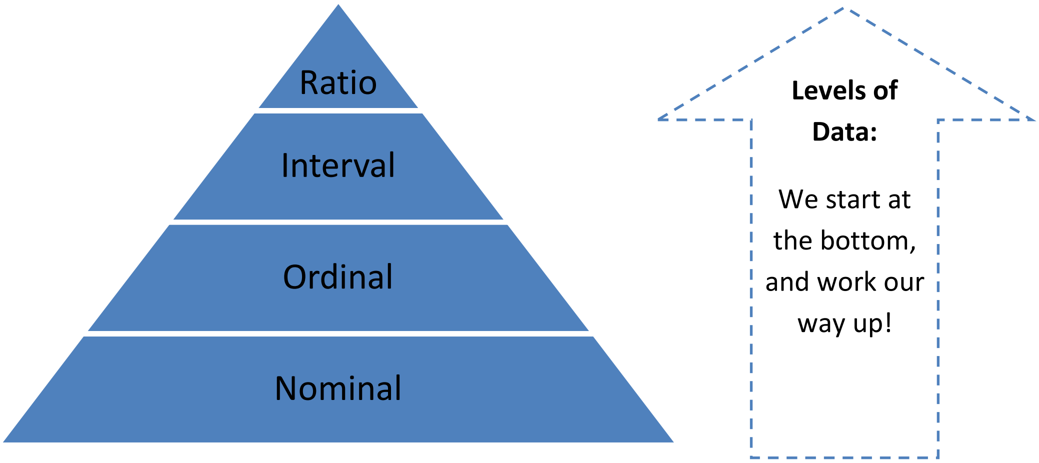
Therefore we can now define descriptive and inferential statistics.

Descriptive statistics: using data that has been gathered on a group to describe or reach conclusions **about that same group.**

Inferential statistics: using data that have been gathered from a sample to reach conclusions **about the population from which the sample was taken.**

Data Measurement:

depending on the data we're using, we must alter our analysis. For example, sometimes the mean is irrelevant, or impossible to calculate... for example 'the mean of Australia and Canada' – these countries names are just labels. Likewise, averaging the numbers on the back of football jerseys is possible, but pointless. Therefore we must classify data into different types. There is a set hierarchy; so we work our way up until the data fails to meet a criteria of that category – thus rendering it an object of the previous category.



Levels of Data:

We start at the bottom, and work our way up!

Nominal Data: used only to **classify** or **categorise.** **Non-metric / Qualitative**

Common examples:

- Staff ID
- Gender
- Religion
- Ethnicity
- Geographic location
- Place of birth
- Tax file numbers
- Telephone numbers
- Post codes

This sort of data should never be used to calculate the mean, for example. Even when numbers are assigned to each category, this is usually for data entry purposes. E.g. 1=Erina, 2=Holgate etc… whilst they have numbers in them, taking the average is not going to provide a meaningful number. Nominal-level data generally has limited uses – however the Chi Square is an example of a statistic that can be used on nominal data.

Ordinal Data: can be used to order or rank items, objects or people. **Non-metric / Qualitative**

For example, using numbers 1, 2 and 3 to rank performance of employees. Thus it's no longer a label, it has meaning, 1 comes before 2 etc. Thus we can see that 1 person was the most productive, 1 person was the least productive etc.

However, **the distance between the numbers in ordinal data varies.** E.g. employee 1 might be almost as good as employee 2, however employee 3 might be terrible. Thus the gap between numbers follows so net scale, merely an order or ranking.

In essence, the distances or spacing between consecutive numbers is not always equal.

Ordinal data is often qualitative data.

Interval Data: the intervals between consecutive numbers have meaning, and the zero-point has no meaning beyond convention or convenience. I.e. it is possible to go below zero – such as with temperature.

Metric / Quantitative

Examples:

- Temperature – degrees Celsius

Ratio Data: the same as interval level data – but the zero point carries a meaning, and the ratio of two numbers is meaningful.

Metric / Quantitative

Examples:

- Height
- Weight
- Time volume
- Kelvin temperature
- Production cycle time
- Work measurements time
- Passengers kilometres
- Sales
- Complaints
- Number of employees

Categorical vs. Numerical:

Categorical data is non-numerical data that are frequency counts of categories from one or more variables.

Numerical data can be either **discrete** or **continuous.**

Discrete data has a finite number of values possible, and these values cannot be subdivided meaningfully.

Continuous data can take on values at every point over a given interval.

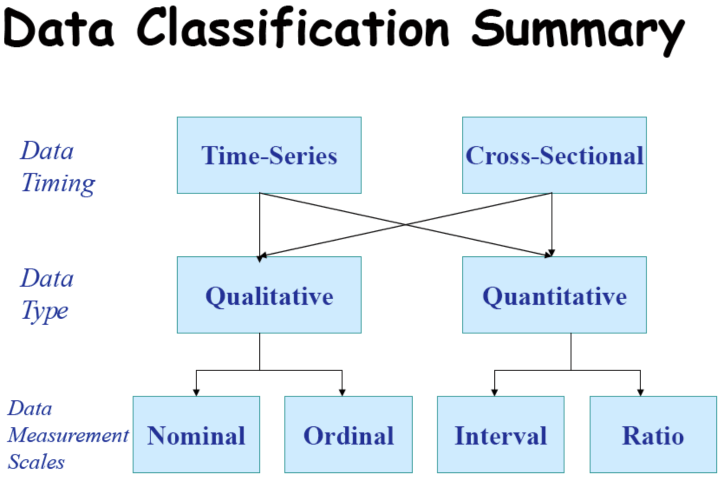
Inferential Methods:

- Parametric (mostly we use this sort of data).
    - Requires:
        - Interval or ratio data
        - Certain assumptions about the distribution of the data
- Nonparametric
    - Requires:
        - Nominal or ordinal data

Data Timing:

Cross Sectional: looking at a sample at one point in time.

Time Series: looking at a sample over time, so we can see changes / patterns etc.



Chapter 2: Charts and Graphs

Grouped Data: data that have been organised into a frequency distribution

Ungrouped Data: raw data, or data that have not been summarised in any way

Frequency Distribution: a summary of data presented in the form of **class intervals and frequencies**.

Range: the difference between the largest and the smallest values in a set of data.

Class midpoint: the value halfway across a class interval

Class mark: another name for class midpoint.

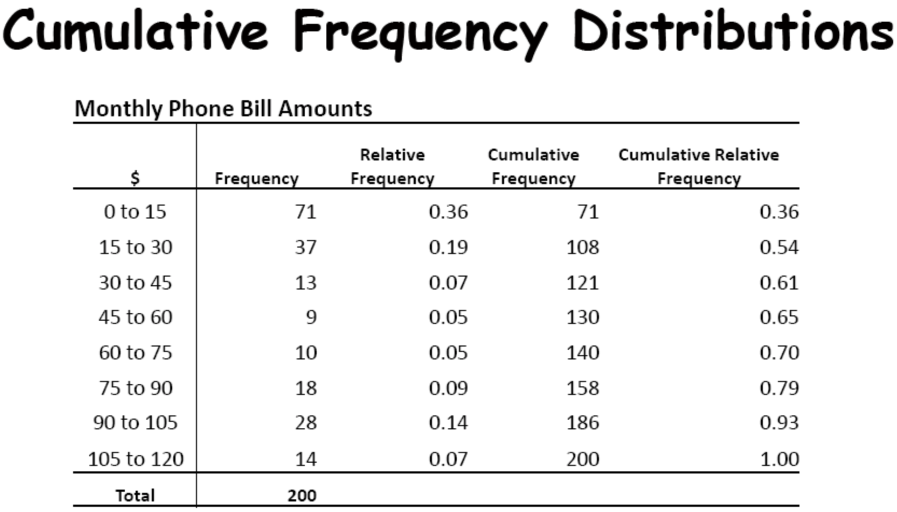
Relative Frequency: is the proportion of the total frequency that is in any given class interval in a frequency distribution

$$\boldsymbol{Relative\ Frequency} = \frac{\text{Individual frequency}}{\text{Total frequency}}$$

Cumulative Frequency: a running total of frequencies through the classes of a frequency distribution.

Ogive: a cumulative frequency polygon – useful when we want to make decisions about our running totals… e.g. running cost of a car over a year. Steep slopes show big increases in frequencies.

Example:



Cumulative Frequency Distributions

Monthly Phone Bill Amounts

$	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
0 to 15	71	0.36	71	0.36
15 to 30	37	0.19	108	0.54
30 to 45	13	0.07	121	0.61
45 to 60	9	0.05	130	0.65
60 to 75	10	0.05	140	0.70
75 to 90	18	0.09	158	0.79
90 to 105	28	0.14	186	0.93
105 to 120	14	0.07	200	1.00
Total	200			

Summary from Example:

Relative Frequency = Frequency / Total

Cumulative Frequency = Frequency Above + This Record’s Frequency

Cumulative Relative F /e Frequency

Lecture 2: Descriptive Statistics

Chapter 3: Descriptive Statistics

Describing Measures of Location:

- Mean: average of data values
- Median: middle observation of the ordered data
- Mode: most frequent

Measures of Central Tendency

- Percentiles: data broken into groupings of 10% of ordered data in each group
- Quartiles: data broken into 25% of ordered scores per group. Used for box & whisker plots.

Notation:

i = observation number

n = sample size

∑ = summation

x = variable

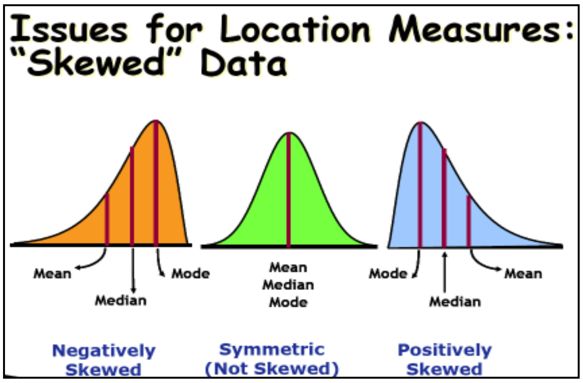
x_1 = value of variable x at observation i

Outliers will ALWAYS change the mean, causing data to be skewed… never change median / mode.

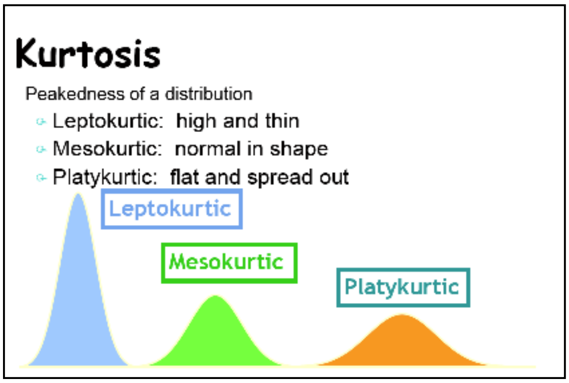
Type	Measurement Scale	Measures of Central Location
Qualitative	Nominal	Mode
	Ordinal	Mode Median
Quantitative	Interval	Mean Mode Median
	Ratio	Mean Mode Median

Skewness:

Kurtosis:



Issues for Location Measures: "Skewed" Data



Kurtosis

Peakedness of a distribution

- Leptokurtic: high and thin
- Mesokurtic: normal in shape
- Platykurtic: flat and spread out

Location and Spread – what do they tell us?

- Measures of **location** indicate the **typical centrality** of data values.
- **Spread** indicates **how close** typical data values are to **central location.**

Measures of Variability: Describe the spread or dispersion of a set of data.

- Range: largest - smallest
- Inter-quartile range
- Variance
- Standard Deviation
- Coefficient of Variation

Notation:
$\bar{x}$= sample mean
m (mew) = Population Mean
µ or M = population mean

s = sample standard deviation
σ = population standard deviation

s^2 = sample variance
σ^2 = population variance

Sample and Population <u>Variance</u>

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

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

i.e. use ALL of our scores for the top bit, then sum them, then divide it by the denominator.

Standard Deviation:

$$Sample\ Standard\ Deviation = s = \sqrt{s^2}$$

$$Population\ Standard\ Deviation = \sigma = \sqrt{\sigma^2}$$

s^2 and σ^2 come from the variance in above formula.

The bigger the standard deviation, the bigger our bell curve.

Sample and Population <u>Coefficient of Variation:</u>

$$Sample\ Coefficient\ Variation = \frac{standard\ deviation}{mean} \times 100$$

$$Population\ Coefficient\ Variation = \frac{standard\ deviation}{mean} \times 100$$

Coefficient of Variation shows how large the standard deviation is in relation to the mean. (Best suited to non-negative data sets).

Methods of Detecting Outliers:

- Box Plots → don't use box plots… they're difficult!
- Z-Scores → use Z-Scores, they're much quicker and easier!

Note – "outliers can be legitimate scores; however they usually are not representative of the population as a whole". E.g. CEO's salaries.

Z-Scores:

Sample Z-Score: $Z_i = \frac{x_i - \bar{x}}{s}$

Population Z-Score: $Z_i = \frac{x_i - \bar{x}}{\sigma}$

RULE – FOR DETECTING OUTLIERS:

- Formulate a "minimum" and a "maximum" value. If a score is found to have a value below the minimum, or above the maximum, then it is an outlier.

Minimum Value = $\bar{x}$ - 3 x standard deviations

Maximum Value = $\bar{x}$ + 3 x standard deviations

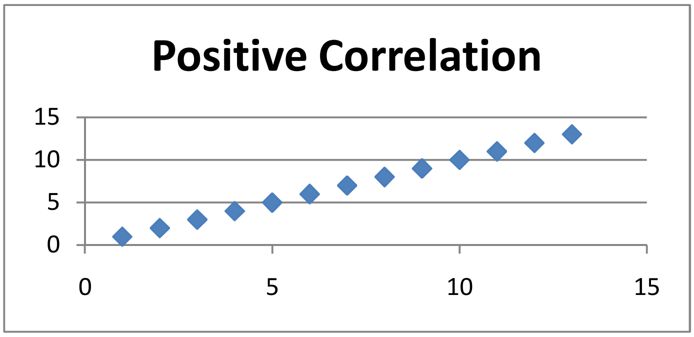
Rule with Z-Scores: $\sum(x - \bar{x}) = 0$ This is because all the + and – should balance out.

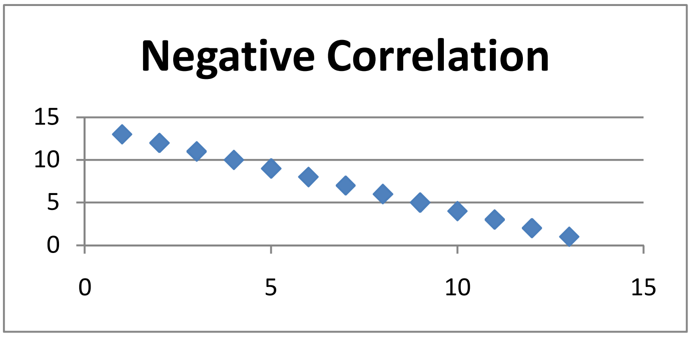
Empirical Rule: shows us a guideline of the approximate percentage of values within that range of standard deviations from the mean.

$\mu \pm 1\,\sigma = 68\%$

$\mu \pm 2\,\sigma = 95\%$

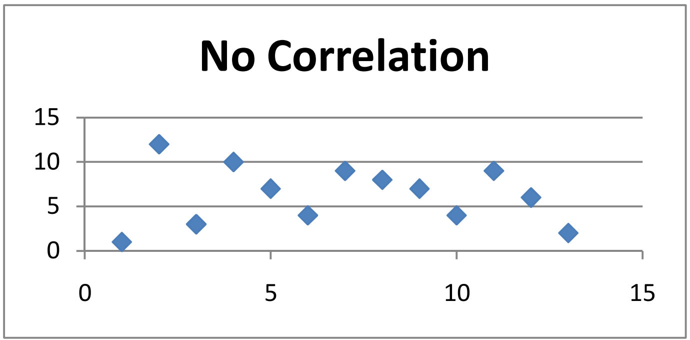
$\mu \pm 3\,\sigma = 99.7\%$





Correlation: the degree of relatedness between variables

1	Positive	Strong: +1 to +0.7
0	Negative	Weak: +0.7 to -0.7
-1		Strong: -0.7 to -1



Lecture 3: Introduction to Probability

Chapter 4: Probability

Three general methods of assigning probabilities:

- Classical Method
  - Assigning probability based on assumptions of equally likely outcomes
    - E.g. P(Heads) = .5
      - We assume there is 50/50 chance either way.
  - $P(E) = \frac{n_e}{N}$
  - n_e = number of outcomes in which the event occurs
  - N=total possible number of outcomes in an experiment

- Relative Frequency of Occurrence Method
  - Assigning probability based on experimentation or historical data
    - E.g. P(Heads)= T, H, H, H, T, T, T, H = .50
  - $P(E) = \frac{x}{N}$
  - x = number of times event occurred
  - N = total number of opportunities for the event to occur.

- Subjective Probabilities
  - Assigning probability based on the intuition or reasoning of the person determining the probability.

Rules of Probability:

∩ = AND = Multiply	
OR = Add	Use this for Sub Events e.g. probability of coke = P(1L coke) + P(500mL coke)
$0 \leq P(E_i) \leq 1$	i.e. all probabilities lie between 0 and 1
$\sum P(E_i)=1$	i.e. the sum of all probabilities = 1
' = "not"	i.e. P(Coke') = the probability of NOT coke… can also be A* A' A^C A~
Compliment Rule:	P(E) = 1 – P(E')

Mutually Exclusive Events:

If event A can only occur if event B does not occur, then the events are mutually exclusive: P(A∩B)=0

A special rule of mutually exclusive events is: P(X ∪ Y) = P(X) + P(Y)

The Contingency Table – Joint Probability Table:

P(A∩B)	P(A∩B')	**P(A)**
P(A'∩B)	P(A'∩B')	**P(A')**
P(B)	**P(B')**	**1.00**

Marginal Probability: Independent Events:

P(A) = P(A∩B) + P(A∩B') e.g. Rain and Going to Class independent, thus no dependence.

Independence Test:

Events are Independent if: P(A|B) = P(A|B’) = P(A)

Example:

		Obese		
		Yes	No	
Pregnant	Yes	0.5192	0.3382	**0.8573**
	No	0.1044	0.0383	**0.1427**
		0.6235	**0.3765**	**1.0000**

Independence Test: P(Obese) = P(Obese | Pregnant)

P(Obese) = 0.6235

P(Pregnant) = 0.8573

P(Obese | Pregnant) = $\frac{0.5192}{0.8573}$= 0.6056 (or 0.61)

Therefore, 0.62 = 0.61, so the events are independent.(close enough..ish)

If independent then the following is true: P(A∩B) = P(A) x P(B)

Conditional Factors:

- We don’t determine things like our age or gender
- We can thus assume that it’s “given”
- Therefore, we’re often needing to know “Given you’re 21 years old, what is the probability that you’ve used drugs?”)

P(A|B) = the probability of A occurring, given that event B occurs.

$$P(A|B) \; = \frac{P(A \cap B)}{P(B)}$$

P(A∩B)	P(A∩B’)	**P(A)**
P(A’∩B)	P(A’∩B’)	**P(A’)**
P(B)	**P(B’)**	**1.00**

Combinations enumerate the number of ways to arrange **x** items from **n** possible ways.
NOTE: only works where order DOES NOT matter.

$$\boldsymbol{{}^{n}C_{x} = \binom{n}{x} = \frac{n!}{x!\,(n-x)!}}$$

n = number to choose from

x= number to select

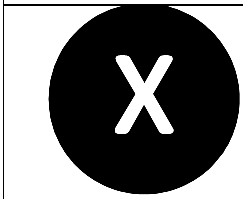
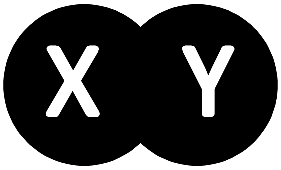
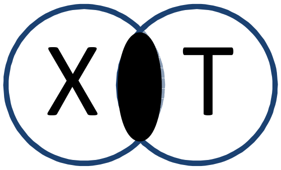
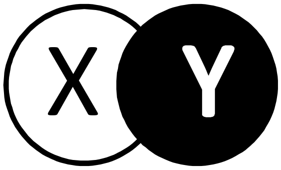
Or just use the nCr button on the calculator ☺

Where do we use combinations?

Example:

I have 5 workers that are equally qualified, but I can only promote 2 of them. How many possible ways are there to choose these 2 workers?

10 nCr 2 = 10 workers

Marginal	Union	Joint	Conditional
P(X)	P(X ∪ Y)	P(X ∩ Y)	P(X|Y)
The probability of X occurring Uses total possible outcomes denominator	The probability of X or Y occurring Uses total possible outcomes in denominator	The probability of X and Y occurring Uses total possible outcomes in denominator	The probability of X occurring, given that Y has occurred Uses subtotal of the possible outcomes in denominator
		 i.e. just the overlap	 i.e. just Y circle incl. overlap
	P(X ∪ Y) = P(N)+(S) – P(N ∩ S) This allows for the overlap... because just adding X and Y would see some bits ‘double counted’. Text pg191		

Lecture 4: Discrete Probability Distributions

Chapter 5: Discrete Distributions

Discrete vs. Continuous Distributions:

- **Random Variable:** denotes the outcomes of a chance experiment; it is a numerical description of the outcome of an experiment. The outcome is not known until the event occurs.
  - **Discrete Random Variables:** random variables in which the set of all possible values is at most a finite or a countable infinite number. i.e. all possible outcomes can be listed; even if it will take a long time
  - **Continuous Random Variables:** random variables that take on values at very point over a given interval. This is an exact number – e.g. time, distance, weight.

Discrete Distributions describe discrete random variables.

- Binomial distribution
- Poisson distribution
- Hypergeometric distribution (ignore it – we don't use it)

Continuous Distributions describe continuous random variables.

- Normal distribution
- Uniform distribution
- Exponential distribution
- t distribution
- chi-square distribution
- F Distribution.

These are in Chapter 6, covered in next lecture (5)

Examples of each from slides:

Example	Distribution Type
The **distance** a car travels on one tank of petrol	Continuous Normal
The **number of accidents** that occur annually on a busy stretch of highway	Discrete Poisson
The **number of satisfied customers out of the total customers** served in a day	Discrete Binomial
Time taken by a respondent to complete an online survey	Continuous Exponential

Expected Value and Variance for **discrete distributions:**

The **expected value**, or mean, of a random variable is a measure of its central location:

$$E(x) = \mu = \sum x \times f(x)$$

i.e. we work it out for each score / frequency… just do x times by the frequency of x.. then sum the result of all scores.

This expected value is our mean / average value. It can be compared with the variance, however we will never need to calculate the variance.

see L4: slide 12 on how that is done.

Binomial Experiments:

Use binomial when there is 2 outcomes only – general things like 'success or failure'

Example from lecture 4 – slide 15→

Supto is a telephone company, looking at their drop out rates.
Their trials indicate that the average number of 'drop outs' is 10 for every 1000 calls made.
Supto is **targeting people who make 3 calls per day, and are worried about drop out rates.**
What is the probability that callers will experience 2 drop outs.

Sample size = n = 3 calls
Event of interest = x = two drop outs.
Note that it's a **binary event** – it either happens or it doesn't happen. (i.e. mutually exclusive)

$P(x) = {}^{n}\boldsymbol{C}_{\boldsymbol{x}} \times \mathrm{p}^{\mathrm{x}} \times \mathrm{q}^{(\mathrm{n}-\mathrm{x})}$

Where:
P(x) = the probability of x successes in n trials
n = number of trials
p = the probability of success (of the event in interest) i.e. in the example, a success IS a drop out.
q = 1–p   (thus this is the probability of failure).

Thus from the example:
n=3
x=2
p=0.01 (10/1000)
q=1-p = .99

Substitute in:

$P(2) = {}^{3}\boldsymbol{C}_{\boldsymbol{2}} \times 0.01^{2} \times 0.99^{(3-2)}$
P(2) = 3 x 0.0001 x 0.99
P(2) = 0.000297

Now for when P isn't just =to… like if its P(x≥2) etc…
Can do:
P(x≥2)= 1 – P(x=0) + P(x=1)

Poisson Probability Distribution:
A Poisson distributed random variable is often useful in estimating the number of occurrences over a **specified interval of time or space.**
e.g. number accidents per highway.
It is a discrete random variable – that may assume an **infinite sequence of values**.

The mean rate of occurrence always stays constant.

Use Poisson when there is more than 2 outcomes / events. Or if you want the probability between two intervals – e.g. between time x and y.

Poisson Probability Function:

$$P(x) = \frac{\lambda^x \times e^{-\lambda}}{x!}$$

Where:
P(x) = the probability of x occurrences in an interval
λ = mean number of events in a continuous interval (doesn't have to be discrete).
e = is constant… in calculator.
x= the number…0, 1, 2, 3, … etc – you can **never** scale the x. must scale the λ instead.

Steps:

1. Express the Poisson parameter (λ) as **occurrences per interval.**
2. Determine the value of x to put into equation
3. Use formula

Example:
The number of leaking joints is estimated to be 1 per 10km of pipe.
What is the probability that there are 3 leaking joints in a 10km stretch of pipe?

$$P(x) = \frac{1^3 \times e^{-1}}{3!} \qquad = P(x) = \frac{1 \times 0.367879}{6} \qquad = 0.0613$$

NB: λ and x were in the same units – thus we did not have to convert λ into x's units.

Example 2:
The number of missing baggage in a small city for well known airline averages twelve per day (16 operating hours).
What is the probability that, in any given hour, there will be less than 2 claims?
Must convert λ into the units of 1 hour… thus 12 per 16 hours = 0.75 per 1 hour.
Now need to use Poisson formula 2 times – once for 0 bags, once for 1 bag. Then add the two.

$$P(x=0) = \frac{.75^0 \times e^{-.75}}{0!} = .472 \qquad P(x=1) = \frac{.75^1 \times e^{-.75}}{1!} = .354$$

Thus **P(X<2) = .472+.354 = .826**

Lecture 5: Continuous Probability Distributions

Chapter 6: Continuous Distributions

- Uniform Probability Distribution
- Normal Probability Distribution
- Exponential Probability Distribution

Intervals matter – not points:

A continuous random variable can assume any value in an interval on the real line or in a collection of intervals.

It is not possible to talk about the probability of a random variable assuming a particular value. Instead, we talk about the probability of the random variable assuming a value within a given interval.

Continuous probability uses a **probability density function** since the (Px=x_0)≈0

The Uniform Distribution:

- Key word in question will be 'follows no pattern' or 'told nothing else'.

You know that a friend will meet you between 11am and 12pm. However, you need to leave between 11:12am and 11:36am. What is the probability that they arrive, when you're gone.

Uniform Probability Density Function:

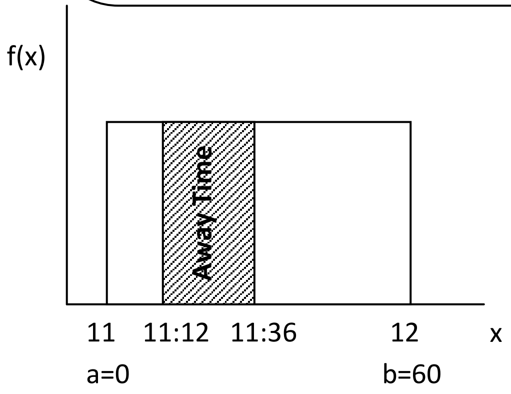
$$f(x) = \frac{1}{b-a}$$

Where a = smallest value the variable can assume, b=largest value the variable can assume.

Uniform Formula:

$$P(x_1 < x < x_2) = \left(\frac{x_2 - x_1}{b-a}\right)$$

Where the x_2-x_1 is the area we want, and the b-a is the total area.



$$P(12 < x < 36) = \left(\frac{36-12}{60-0}\right) = \ 0.40$$

The Normal Distribution: (The MOST IMPORTANT!)

Normal Probability Density Function:

$$f(x) = \frac{1}{\theta\sqrt{2\pi}} e^{-\left(\frac{1}{2}\right)\left[\frac{x-\mu}{\sigma}\right]^2}$$

This is not in the exam ☺

μ determines the location (left or right) of the standard normal bell curve

σ determines the shape / size… taller and skinny or shorter and fatter.

The Standard Normal Distribution:

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

This is the z-score formula.

This standardises scores – making the mean 0.

The scores are in the z-space, thus we can use z-tables to do probabilities.

The Exponential Probability Distribution:

Cumulative Probabilities:

$$P(x \geq x_0) = e^{-\lambda \times x_0}, x_0 \geq 0$$

Where

x$_0$ = some specific value of x.

λ= mean number of events per unit of time = $\frac{1}{\mu}$

****NB: this calculates the area to the RIGHT. Must remember to do 1 – at the end if we want $\leq$.**

Example:

An accountant notes that, on average, it takes 30 minutes to talk to two clients… with the time per visit following an exponential distribution. What is the probability that visiting a single client will take less than 10 minutes?

$$P(x \geq x_0) = e^{-\lambda \times x_0}, x_0 \geq 0$$

$$P(x \geq 10) = e^{-\frac{1}{15} \times 10}$$

$P(x \geq 10)$=0.51

Thus 1-0.51

=0.49

Lecture 6: Sampling and Sampling Distributions

Chapter 7: Sampling and Sampling Distributions

Reasons for not using the entire population:

- Cost
- Time
- Revelation
- Statistical interference
- Infinity
- Killing the population
- Disruption – e.g. survey about what you spend on fast food might cause people to change their habits, after realising their answers to the survey.

Parameters & Population vs. Statistics and Samples:

- **Population**: all elements of interest
- **Census**: information from all members of the population
- **Sample**: a subset of the population
- **Parameter**: a measure computed from the entire population
- **Sampling error**: is the difference between a value computed from a sample (a statistic), and the corresponding value computed from the population (parameter).
- **Statistical inference**: to develop estimates and test hypotheses about the population, using a sample.
- **Point estimate**: a single value statistic.

Gathering a Sample:

- Sample frame: master list of the population of interest
- Probability sample: simple, systemic, cluster, stratified
- Non-probability sample: convenience, judgement, quota
- Errors (every sample has errors):
  - Sampling error
  - Non-sampling error
  - Error = Parameter – Estimate

Random Sampling Techniques:

- Simple random sampling
  - A sampling method in which every (possible) sample has the same probability of being sampled; consequently, every unit of the population has an equal probability of being selected for the sample.
- Stratified random sampling
  - A type of random sampling in which the population is divided into various non-overlapping strata and then items are randomly selected for the sample from each stratum.
    - Proportionate stratified random sampling
      - The proportions of the items selected for the sample from the strata reflect the proportions of the strata in the population.
    - Disproportionate stratified random sampling
      - The proportions of items selected from the strata for the sample do not reflect the proportions of the strata in the population.
- Systematic random sampling
  - A random sampling technique in which every k^{th} item or person is selected from the population.
  - $k = \frac{N}{n}$ Where k= size of interval, N = size of population, n=size of sample.
- Cluster (area) random sampling
  - A type of random sampling in which the population is divided into non-overlapping areas or clusters, and elements are randomly sampled from the areas or clusters. E.g. states, towns, etc.

Statistical Inference Process:

1. Population with mean? No →
2. A simple random sample of *n* elements is selected from the population
3. A sample mean is calculated
4. The value of the sample mean is used to make inferences about the population mean.

Finite vs. Infinite Populations

- **Finite:** population size N is known
- **Infinite:** the elements cannot be listed and numbered (i.e. goes on forever)
- Some finite populations are so large that for practical purposes, it must be treated as infinite.

Standard ERROR of the Sample Mean (equivalent of Standard Deviation – for sample means)

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

Use where:

- Infinite population (N is unknown)
- Finite population where N is sufficiently large

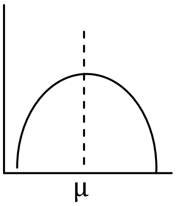
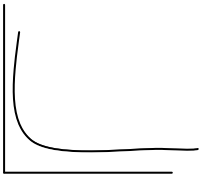
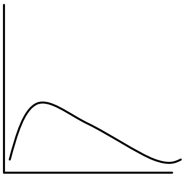
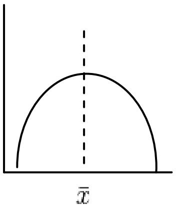
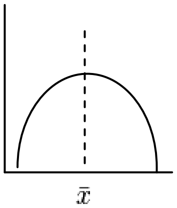
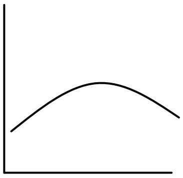
Notes:

- when σ increases, $\sigma_{\bar{x}}$ increases
- when n increases, $\sigma_{\bar{x}}$ decreases

Central Limit Theorem:

- in selecting simple random samples of size **n** from **a population**, the sampling distribution of the sample mean (x - bar) can be approximated by a normal probability distribution as the sample size becomes large (when n ≥ 30), **regardless of the shape of the population distribution.**
- **Whenever the population has a normal probability distribution**, the sampling distribution of x-bar has a **normal probability distribution** for **any** sample size.

Central Limit Theorem – diagrams:

	Normal	Exponential	Chi-Square
Population			
Sample	*When x is normal, and n is any size.*	*When x is not normal, and n ≥ 30*	*When x is not normal and n < 30*
			

i.e.

- Regardless of the distribution, if n ≥ 30, it is **normally distributed**
- When x is stated as normally distributed, regardless of n, it is normally distributed.
- When n < 30, and x is not stated as normal, it is **not normally distributed.**

Standard Error – Finite Correction:

If $\frac{n}{N} < 0.05$ we do NOT need FCF

(Also use this if N is unknown – as we assume it is infinite).

$$\sigma \quad = \frac{\sigma}{\sqrt{n}}$$

If $\frac{n}{N} > 0.05$ we NEED FCF

$$\sigma \quad = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$$

Z Scores – for the sample means

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

Where:

$\bar{x}$ = sample mean

$\mu_{\bar{x}}$ = mean of sample means

$\sigma_{\bar{x}}$ = standard error of sample mean (basically the standard deviation). ($\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$)

Example:

Suppose the average grade of 225 applicants is calculated to be 106.3 with a standard deviation of 12. Assuming the population of applicants is infinite, what is the probability that:

- The sample mean is computed to be within +/- 1 mark of the population grade.

Assume the average grade follows a normal distribution.

n=225

$\bar{x}$=106.3

s = 12

N = infinite (thus no FCF is needed)

P(105.3< Expected Value $\bar{x}$ < 107.3)

Need to: calculate the z-score of 105.3 and 107.3

Standard Error:

$$\frac{12}{\sqrt{225}} = 0.8$$

Z Upper:

Z = $\frac{107.3 - 106.3}{0.8}$ = 1.25

Z Lower:

Z = $\frac{105.3 - 106.3}{0.8}$ = - 1.25

Look up in Z table:

Z=1.25 = 0.8944

Z=-1.25 = 0.1056

Therefore, 0.8944 – 0.1056 = 0.7888…. or 78.88%

Sampling Distributions: PROPORTIONS:

$$Proportion = \frac{number\ in\ group}{number\ in\ population}$$

E($\hat{p}$)=p

Where p = the population proportion

Standard Deviation of $\hat{p}$ = $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$

Note: q = 1-p

$\sigma_{\hat{p}}$ is used as the standard error of the proportion

Can we use A Normal Approximation?

The sampling distribution of $\bar{x}$ can be approximated by a normal probability distribution whenever the sample size is *large*.

Something is *large* if:

n ≥ 30

The sampling distribution of $\hat{p}$ can be approximated by a normal probability distribution whenever the sample size is *large*.

But in this case, large is determined by:

$n \geq \frac{5}{p}$ AND $n \geq \frac{5}{1-p}$

- Always round UP the answers.

Finite Population Correction:

- If a population Is neither infinite nor extremely large, the **standard error must be adjusted** to account for this

Finite population correction is necessary when: $\frac{n}{N} > 0.5$

For **proportions**:

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} \times \sqrt{\frac{N-n}{N-1}}$$

For **Quantitative Data:**

$$\sigma_x = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$$

Lecture 7: Interval Estimation

Chapter 8: Statistical Inference: Interval Estimation for a Single Population

USING THE NORMAL DISTRIBUTION: FIND THE PROBABILITY GIVEN Xi

Rule: $P(x \leq x_i) = P(Z \leq \frac{x_i - \mu}{\sigma})$ (basically, this is the z-score approach)

Example:

Companies that are part of the All Ordinaries Index have an average market capitalisation of $4.1 billion.

The average changes as companies' share prices change. The standard deviation is $0.55 billion.

What is the probability that a company will have a market capitalisation of less than $4.0 billion?

Answer:

$$P(x < 4.0) = Z = \left(\frac{4.0 - 4.1}{0.55}\right)$$

=P (z < -0.18)

look up -0.18 in Z-Table

P(x<4.0) = 0.4286

Normal Distribution – The Graph – What makes it move around?

When…

- Mean + = moves →
- Mean - = moves ←
- SD + = flatter
- SD - = pointier / taller

USING THE NORMAL DISTRIBUTION: FINDING Xi GIVEN THE PROBABILITY

Rule: $P(z \leq Z_i) = P(x \leq \mu + Z_i \sigma)$

Example:

The sales manager entry program requires his team to score in the top 1% of the company's nationwide sales skills program.

The manager knows that scores are normally distributed, with a mean of 490 and standard deviation of 61.

What should be the minimum score for accepting team members into the sales team?

Answer:

$\bar{x} = 490$

$\sigma = 61$

We want to know P(x>0.01) or P(x<0.99) (i.e. top 1%, or bottom 99%... same thing)

0.99 = P(z<2.33) ***from z-table… where is the probability 0.99? it is closest at 2.33***

$$2.33 = P\left(\frac{x - 490}{61}\right)$$

= 61 x 2.33 = x - 490

=623 (rounded) is the score that is necessary to be in the top 1%, and thus accepted into team.

Confidence Intervals – What are they all about?

- We want to make statements about some **population parameter**
    - E.g. "Our customers, on average, are taking 25 minutes to eat their meals"
    - E.g. "The mean proportion of people who would be happy to drink their own crap (recycled water) is close to 80%"

Remember, we **cannot** observe our population parameter (the whole population)
But, we would like to make comments about it.
So, we use **sample statistics** to make comments about it.
Our inference statements depend on the accuracy we want to make and the nature of the data itself.

With how much consider can we make the statements written above? We need to express each statement in terms of an ***upper*** *and* ***lower*** *limit –* ***this is the 'interval'.***

Confidence Interval: an interval developed from sample values such that if all possible values of a given width were constructed, a percentage of these intervals, known as the confidence level, would include the true population parameter.

Confidence Intervals: Format:

Point Estimate ± Margin of Error

Point Estimate ± (Critical Value) (Standard Error)

*point estimate = the mean

Commonly Used Confidence Levels

1-α (α= 'not confident')	**α**	$\frac{\alpha}{2}$	$Z_{\frac{\alpha}{2}}$
.90	.10	.05	1.645
.95	.05	.025	1.96
.99	.01	.005	2.575

Critical Value

Critical Level

Example: n>30, σ known
A random selection of 250 teenagers reported the number of CDs purchased in the previous 12 months
We find that 4.26 CDs are purchased on average.
Estimate with 99% confidence, the mean number of CDs purchased annually by teenagers. (assume the population standard deviation to be σ=3.0 CDs)
Answer:

Standard Error: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3.0}{\sqrt{250}} = 0.19$

99% confidence = 2.575 Critical Value

Upper Value	**Lower Value**
4.26 + 2.575 x 0.19 = 4.75	4.26 - 2.575 x 0.19 = 3.77

Thus, we are 99% *CDs per year.*

Confidence Intervals – worked out steps:

1. Calculate the average… or get it from the question

2. Do you know the population standard deviation? (σ)

Yes – σ is known	**No – σ is not known**
Apply Formula: $\bar{x} \pm Z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$ $Z_{\frac{\alpha}{2}}$ = Critical Value (use golden table – below) σ = Population Standard Deviation n = sample size	Use 's' (sample standard deviation) instead HAVE TO USE T TABLE FOR CRITICAL VALUE $\bar{x} \pm T_{\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}$ $T_{\frac{\alpha}{2}}$ = Critical T from T Table... (see below for how to look up) s = sample standard deviation n = sample size
90% CI = Z = 1.645 95% CI = Z =1.96 99% CI = Z = 2.575	T – TABLE d.f. = n-1, α is same – remember to divide by 2.

Margin of Error and Required Sample Size:

You want to use the smallest possible sample size, but still be accurate.
So what sample size do we need to get our desired level of confidence?

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

Where:

σ^2 = population standard deviation, squared (aka the Variance) if not given, use ¼ of range.
E^2 = Margin of Error = $\bar{x} - Mue\ (population\ mean)$ this is always given. Sometimes it's referred to as 'precision'.

ALWAYS **<u>ROUND UP</u> the answer you get **

Example:

A medical researcher wishes to investigate the time taken to relieve headaches using a new painkiller. She believes the population is normally distributed with a standard deviation of 20 minutes. How large a sample size should be taken to estimate the mean time to within 1 minute at a 95% confidence level?

n = ? $Z_{\frac{\alpha}{2}}$ = 1.96 σ = 20 mins E = 1

$$n = \frac{[(1.96)^2 \times 20^2]}{1^2} = 1536.64 \quad round\ up \quad = 1537\ people$$

Proportions and Confidence Interval:

$$\hat{p} \pm Z_{\frac{\alpha}{2}} \times \sqrt{\frac{\hat{p} \times \hat{q}}{n}}$$

***ALWAYS USE A Z-SCORE ***

Must test to ensure we can use this formula – it can only be used when

n x $\hat{p} \geq 5$

n x $\hat{q} \geq 5$

Example:

An electronics company has found that 10% of all circuit boards fail.

The company based this percentage on a test sample of 50 circuit boards.

Construct a 95% confidence interval estimate for the circuit board failure population proportion.

Answer:

$\hat{p} = .10$

$\hat{q} = .90$

n = 50

test:

n x $\hat{p} \geq 5$	50 x .10 = 5	☑
n x $\hat{q} \geq 5$	50 x .90 = 45	☑

$$0.10 \pm 1.96 \times \sqrt{\frac{0.10 \times 0.90}{50}}$$

= 0.18 (upper)

= 0.02 (lower)

Lecture 8: Hypothesis Testing

Chapter 9: Statistical Inference: Hypothesis Testing for Single Populations

Hypothesis Testing involves a specific value and two outcomes… these are the two competing hypotheses.
Examples:

Ho: Parameter relationship value
Ha: Parameter opposing relationship value (and the > points to the tale end!)

Example	***Hypotheses***	***In Statistics***
Jenny wants to know if waiting at KFCs drive through, on average, is less than five minutes	Ho: M ≥ 5 Ha: M < 5	M = 5 M < 5
Jeff wants to know if the average distance travelled by his sales team significantly exceeds the budgeted distance of 100km	Ho: M ≤ 100 Ha: M > 100	Ho: M = 100 Ha: M > 100
Jim wants to know if spending money on one of the new software systems has meant average call times is exactly within the specified 3 minute milestone.	Ho: M = 3 Ha: M ≠ 3	M = 3 M ≠ 3

HTAB System to Test Hypotheses:

1. Hypothesise
2. Test
3. Take Statistical Action
4. Determine the business implications.

Two Hypotheses:

- **Null Hypothesis (Ho) –** the 'status quo'
  - a maintained hypothesis that is held to be true unless sufficient evidence to the contrary is obtained
- **Alternative Hypothesis (Ha)**
  - A hypothesis against which the null hypothesis is testing and which will be held true if the null is held false.

Things to remember:

- The null hypothesis represents the situation assumed to be true unless the evidence is strong enough to convince the decision maker it is not true
- Whatever we're investigating is the Alternative Hypothesis
- The alternative hypothesis is what the test is attempting to establish
- The alternative hypothesis is really what answers the question given.

- ***You must always form a hypothesis such that you hope to find evidence to reject the null, and favour the competing (alternative) hypothesis – since the alternative hypothesis is the research hypothesis!!!***

Tails – what do they mean?

Ho: $\mu \geq \mu_o$ Ha: $\mu < \mu_o$	Ho: $\mu \leq \mu_o$ Ha: $\mu > \mu_o$	Ho: $\mu = \mu_o$ Ha: $\mu \neq \mu_o$
One tailed – lower	One tailed – upper	Two-tailed

*Ha will NEVER have an '=' in it.

Example:

Gizmo Importers have stipulated that the sales of any new product line must achieve an average sales return of more than 15%.

The sales executive team have asked you to check this stipulation by looking at only 100 products out of their 10,000 products that they import.

- *What is the null and alternative hypotheses?*

Answer:

Ho: M ≤ .15

Ha: M > .15

*check for finite correction issues…

$\frac{100}{10,00} = 0.01$ no need for FCF

Errors:

We only look at 'false positive' – type I errors… The probability is α

Alpha represents the probability that we will make a Type I error.

Generally assumed to be α 5% (i.e. 5% chance we have an error, 95% chance we don't)

Hypothesis Testing Probabilities:

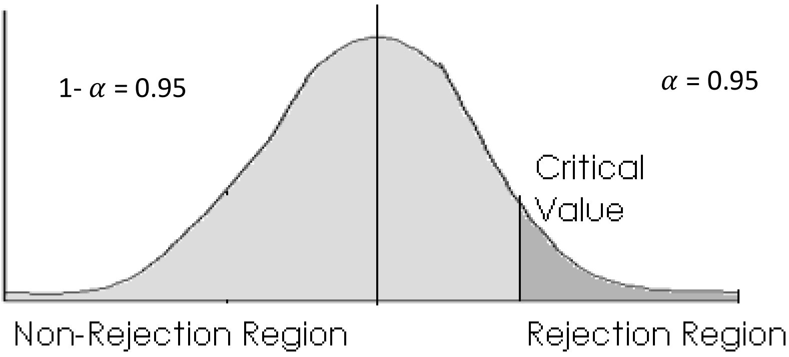
Common Critical Values:

- Standardised values allow us to get probabilities
- Get critical values from appropriate t-distribution, OR
- Z-distribution uses the table below.

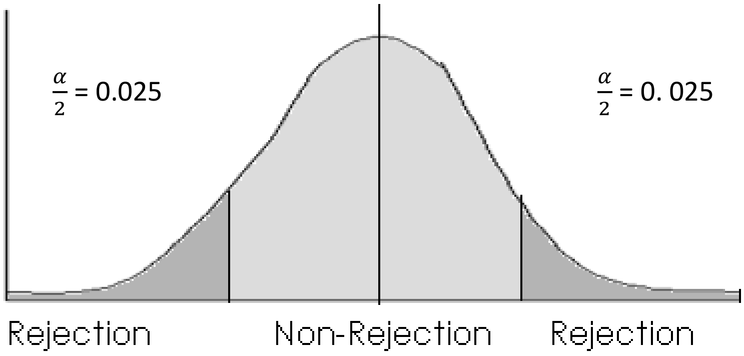
% Sign	***Alpha***	***One-Tailed***	***Two Tailed***
.99	0.01	2.326	2.576
.95	0.05	1.645	1.96
.90	0.10	1.282	1.645
.80	0.20	0.842	1.282

<u>What Does It Look Like?</u>

<u>One Tailed Test:</u>



<u>Two Tailed Test:</u>



<u>Steps:</u>

1. Determine: One or Two Tail Test?

> or < indicates ONE TAILED TEST

= / ≠ indicates TWO TAILED TEST

2. T or Z Distribution?

Sample (e.g. given an ‘s’ – sample standard deviation) = T

Population (e.g. given an ‘σ’ – population standard deviation) = Z

3. Calculate test scores / observed scores:

T	Z
$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$ d.f. = n-1	$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

(bottom part is just the standard error for samples (t) and standard error for populations (z).

4. Look up the Critical Values – they come from the tables

Remember to use α from the question (or assume 95% if not given!)

Example:

A claim made by Normo Electronics is that the average life of a battery they provide to various devices exceeds 4,000 hours. To test this claim, a random sample of 12 components is examined, tracing the life between installation and failure. The data revealed an average life of 4366 hours with a sample standard deviation of 1000.849 hours.

(i) What conclusion should be reached based on the sample data given, assuming a 0.05 level of significance.

~~(ii) Suppose that the random sample consisted of 500 components. What is your conclusion now?~~

(iii) What assumptions regarding normality need to be made in both (i) and (ii)?

Answer:

(i)

$\bar{x}$ = 4366   s = 1000.849   n = 12   d.f. = 11   t-Value (it's a sample)

Hypotheses:

Ho: M ≤ 4000

Ha: M > 4000

Observed T:

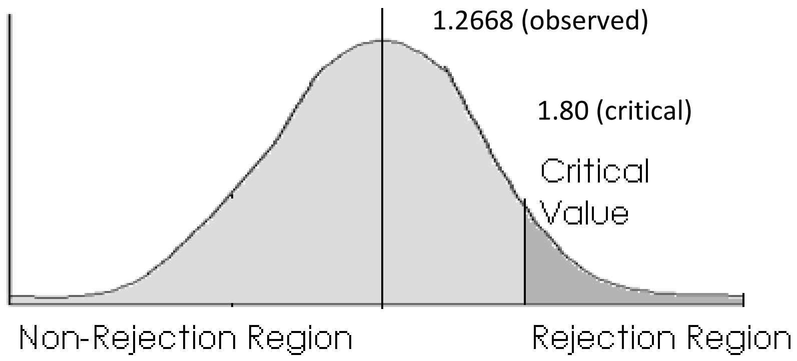
$$t = \frac{4366 - 4000}{\frac{1000.849}{\sqrt{12}}}$$

= 1.2668

Critical T:

d.f. = 11, α 0.05   (1 tailed, upper)

Critical T = 1.80



<u>Thus, as the test score lies within the non-rejection region, we cannot reject the null. As such, the company is claiming the Alternative, when the evidence suggests that the Null cannot be rejected.</u>

(iii) The assumption that the outcomes are normally distributed must be made to be able to test hypotheses and r where n>30.

Proportions and Hypothesis Testing:

Proportions ALWAYS use the z-distribution, as long as:

n x $\hat{p} \geq 5$

and

n x $\hat{q} \geq 5$

Same as before:

Ho: $\mu \geq \mu_o$ Ha: $\mu < \mu_o$	Ho: $\mu \leq \mu_o$ Ha: $\mu > \mu_o$	Ho: $\mu = \mu_o$ Ha: $\mu \neq \mu_o$
One tailed – lower	One tailed – upper	Two-tailed

Still use significance level given by α.

The Formula:

$$\sigma_{\hat{p}} = \sqrt{\frac{p_0(1-p_0)}{n}} \neq \sqrt{\frac{\hat{p}\hat{q}}{n}} \qquad Z = \frac{\hat{p}-p_0}{\sqrt{\frac{p \times q}{n}}}$$

Example:

The sate energy commission determines that too much energy is being wasted due to lack of insulation in houses. The commission is contemplating a law that specifies at least 80% of new houses be insulted. This means that no more than 20% can be non-insulated.

However, the builders association believes that the proportion of houses that are non-insulated is already less than 20%, so the law in unnecessary.

The energy commission surveys 224 new houses and finds that 52 houses are non-insulated.

Using α = 0.10 level of significance, test the hypothesis that the proportion of houses that are non-insulated is less than 20%.

Answer:

Ho: M ≤ 20

Ha: M > 20

$$\hat{p} = \frac{x}{n} = \frac{52}{224} = .23$$

$$Z = \frac{\hat{p}-p_0}{\sqrt{\frac{p \times q}{n}}} \qquad = \frac{.23-.20}{\sqrt{\frac{.2 \times .8}{224}}} \qquad = \mathbf{1.12}$$

Critical Z (all proportions use z, even if it’s a sample)

α = .10, 1 tailed (upper)

Critical Z = 1.282

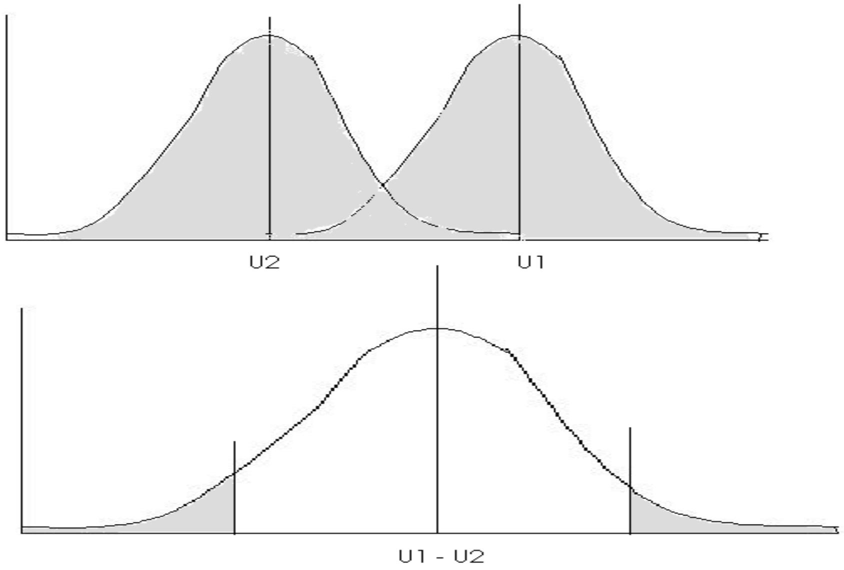
Thus 1.12 < 1.282… so it lies within the non-rejection zone (draw it graphically, rejection zone is above 1.282)

therefore, we cannot reject the null

Lecture 9: Comparisons Involving Means and Proportions

Chapters 10, 11 & 12: Statistical Inference: For Two Populations

"Comparison of TWO groups against each other"
e.g. average male salary vs. average female salary.

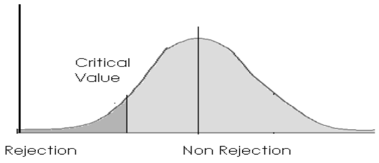
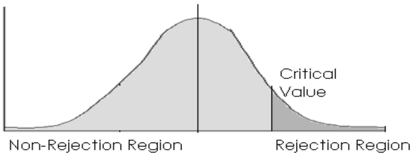
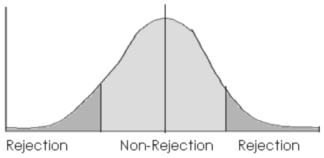


Ho: Mm = Mf
Ha: Mm ≠ Mf

Thus:

Ho: Mm – Mf = 0
Ha: Mm – Mf ≠ 0

The same rules of rejection apply.

	 Critical Value Non-Rejection Region Rejection Region	
Ho: $\mu \geq \mu_o$ Ha: $\mu < \mu_o$	Ho: $\mu \leq \mu_o$ Ha: $\mu > \mu_o$	Ho: $\mu = \mu_o$ Ha: $\mu \neq \mu_o$
One tailed – lower	One tailed – upper	Two-tailed

The hypothesis test is looking at whether or not there is a difference between means of two populations: Independent Samples.

Common Critical Values (SAME AS USUAL)

- Standardised values allow us to get probabilities
- Get critical values from appropriate t-distribution, OR
- Z-distribution uses the table below.

% Sign	***Alpha***	***One-Tailed***	***Two Tailed***
.99	0.01	2.326	2.576
.95	0.05	1.645	1.96
.90	0.10	1.282	1.645
.80	0.20	0.842	1.282

STANDARD ERROR:

Four Formulas (things with formulas are in bold / underlined): Slide 8, Lecture 9… FINAL EXAM

- **Large Sample *(n_1 AND $n_2 \geq 30$)***
    - **Standard Deviation Known**
    - **Standard Deviation Unknown**
- **Small Sample Case *(n_1 OR $n_2 < 30$)***
    - Equal Variance (assumed)
        - **Standard Deviation Known**
        - **Standard Deviation Unknown**
    - Unequal Variance (cannot do this – but we always assume equal variances anyway)

STANDARD ERROR (I.e. the denominator for test statistic...explained below)

LARGE SAMPLE ***(n_1 AND $n_2 \geq 30$)***	**Small Sample** ***(n_1 OR $n_2 < 30$)***
Standard Deviation **KNOWN** $\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$	Standard Deviation **KNOWN** $\sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$
Standard Deviation **UNKNOWN** $\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$	Standard Deviation **UNKNOWN** (EXAM!!) $\sqrt{\left(\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{((n_1+n_2)-2)}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$

Test Statistic – comparing two groups against each other:

$$Test\ Statistic = \frac{Evidence - Hypothesised\ Value}{\boldsymbol{Standad\ Error}\ of\ the\ Value\ in\ Question}$$

***μ_1 - μ_2 = hypothesised value.*

Simply put:

Z (where σ is known)	***T (where σ is NOT known)***
$z = \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{Standard\ Error\ (above)}$	$t = \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{Standard\ Error\ (above)}$
Critical comes from COMMON TABLE of CRIT Z's	Critical comes from T table; d.f = n-1.

Example:

A shopping chain is thinking about building a new centre at one of two locations:

- *Mt Gravatt where a sample of 100 households reveals an average annual household income of $29,980 with standard deviation of $4,470.*
- *Logan where a sample of 100 houses reveals an average annual household income of $28,650 with standard deviation of $5,365.*

The company knows building costs are lower in Logan City and will build there only if average household income at Mt Gravatt exceeds Logan considerably (significance level of 5%).

Answer:

	Mt Gravatt	**Logan**
i	1	2
n	100	100
$\bar{x}$	29,980	28,650
s	4,740	5,365

M = Average Annual Household Income

Ho: M1 ≤ M2

Ha: M1 > M2

i.e. we want to try and prove that Mt Gravatt's annual household income is significantly more than Logan's: meaning that despite the higher building costs, the average annual income is higher… thus better chance of customers spending more money in centre.

Ho: M2 – M2 ≤ 0

Ha: M1 – M2 > 0 (Determines Rejection End)

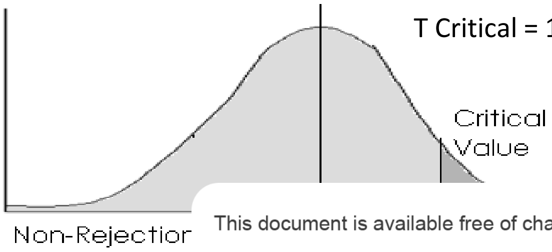
It's large, with a sample standard deviation!

Test Stat:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \qquad = \frac{(29{,}980 - 28{,}650) - (0 - 0)}{\sqrt{\frac{4{,}740^2}{100} + \frac{5{,}365^2}{100}}} \qquad = \frac{1{,}330}{\sqrt{512{,}508.25}} = 1.86$$

T critical = 198 d.f. α=0.05 T Critical = 1.65

We are in the REJECTION AREA. Thus, we can **reject the null hypothesis**.



T Critical = 1.65

Test Stat = 1.86

Test Statistic – When events are DEPENDENT (i.e. not told to assume the events are independent):
This allows us to compare **two DEPENDENT groups.**

$$Test\ Statistic = \frac{Evidence - Hypothesised\ Value}{\boldsymbol{Standad\ Error}\ of\ the\ Value\ in\ Question}$$

$$\text{Test statistic } = \frac{\bar{d} - D}{\frac{Sd}{\sqrt{n}}}$$

Where:
$\bar{d}$ = sample difference
D = population difference
Sd = sample standard deviation
n = size of sample
Example:
People completed a survey using pen & paper, and then the same survey online. The difference in time between the two tasks is calculated for 100 individuals, and reported as ***di = online time i – offline time i.*** *Suppose the sample mean and standard deviation of the variable* ***di*** *is reported to be ½ a minute and 1.3 minutes respectively. Is there a significant difference in the time taken for each task at a 95% confidence level? Pretend that you have no a priori convictions about which is quicker (i.e. a two tailed test).*

(i) **Specify the difference of interest (a difference of zero)**
i.e. are they different?

(ii) **Formulate the null and alternative hypothesis**
The hypothesis is that there is no difference; the alternative is that there is a difference.
Ho: U Online – U pen and paper = 0 (status quo; they're the same)
Ha: U Online – U pen and paper ≠ 0 (the research hypothesis)

Significance level = 95% thus α=0.05 This test is two tailed, so α=0.025

(iii) **Compute the test-statistic (paired or independent test?)**
Paired test; because the SAME people are being measured twice.
d bar is the average of the **di** differences, and the **standard deviation** is the standard deviation of the **di** differences.
The standard deviation was estimated – so we must use a t test statistic.

$$Test\ Stat = \frac{\bar{d} - \mu_d}{\frac{Sample\ Std\ Dev}{\sqrt{n}}} = \frac{0.5 - 0}{\frac{1.3}{\sqrt{100}}} = \frac{0.50}{3.846} = 3.846$$

(iv) **Reach a decision (do we reject?)**
T critical = d.f. 100-1 = 99, α=0.025, T Critical = 1.98
Thus, we reject the null. There is a significant difference between online and offline time.

Comparisons Involving Proportions: **(lecture 9, slide 25)**

Same rule applies, ALWAYS USE A Z SCORE METHOD as long as:

n x $\hat{p} \geq 5$

n x $\hat{q} \geq 5$

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} \qquad \bar{q} = 1 - \bar{p}$$

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{(\bar{p} \times \bar{q}) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

NB: $(we\ dont\ know\ p1\ and\ p2..they're\ for\ population\ proportions\ \dots thus = 0)$

Example:

In a study, the effect of smoking on the birth weight was studied for 189 women. Of the respondents, 115 were non-smokers and 74 were smokers.

Of the non-smokers, there were 29 low birth weight babies and of the smokers there were 30 low weight babies.

Is the proportion of low birth weight babies higher for smokers or non-smokers?

		Low		
		No	**Yes**	**Total**
Smoker	**No**	86	29	115
	Yes	44	30	74
	Total	130	59	189

Answer:

Ho: Psmokers ≤ Pnon smokers

Ha: Psmokers > Pnon smokers

Thus,

Ho: Psmokers – Pnon smokers ≤ 0

Ha: Psmokers – Pnon smokers > 0

$$\hat{p}_{smoker} = \frac{30}{74} \qquad \hat{p}_{non\ smoker} = \frac{29}{115}$$

$$\bar{p} = \frac{30 + 29}{74 + 115} = \frac{59}{189} = 0.3122$$

$$\bar{q} = 1 - \frac{59}{189} = 0.6878$$

$$z = \frac{\left(\frac{30}{74} - \frac{29}{115}\right) - 0}{(0.3122 \times 0.6878) \times \left(\frac{1}{74} + \frac{1}{115}\right)} = \frac{0.1532}{0.069056} = 2.219$$

Z critical = 95% c

ECT the null.

ANOVA – Analysis of Variance: See lecture 9, slides 28 and 29.

Ho: u_1=u_2=u_3... etc

Ha: at least one u_i is different from the others.

*always a one tailed test, using 95% confidence.

ANOVA can be done using:

- F distribution
- P value

ANOVA						
Source of Variance	**SS**	**dF**	**MS**	**F**	**P-Value**	**F Crit**
Between Groups (**NUMERATOR**)	1.051439	2	0.52572	14.15436	6.25E-05	3.354131
Within Groups (**DENOMINATOR**)	1.002831	27	0.037142			
Total	2.054271	29				

F Distribution Method:

F Stat = 14.15436

F Critical = 3.35 (using numerator and denominator given)

14.15436 > 3.35, thus we REJECT THE NULL... thus one of them must be different.

P-Value Method:

Rule: if α > P Value, reject the Ho (null).... can use α=0.05 as our default.

Thus, 0.0000625 < 0.05, so we reject the null (same as above).

RULE:

If P > 0.05 – we reject Ho

If P ≤ 0.05 – we reject Ha

Chi Square Test:

- Uses categorical values; where an average would be useless. (e.g. can't average males and females)
- Instead, you count... e.g. 10 males, 15 females.
- Used where you want to test the relationship between two categorical variables – e.g. Gender.

Ho: No relationship

Ha: There IS a relationship

NB: Ingo uses terms:

Relationship = dependent,

No relationship = independent

Steps for Chi Square Test:

1. Examine observed counts
    a. Calculate row totals
    b. Calculate column totals
    c. Calculate the grand total
    d. (i.e. we're calculating all the marginal's)
2. Calculate **expected** counts for each cell

    $Expected\ Count = \frac{row\ total \times column\ total}{Grand\ Total}$

    Do this for EVERY cell.
3. Apply following formula to **all** cells

    $\left(\frac{Observed - Expected}{Expected}\right)^2$
4. Add up **every cell** from step 3... and **this is our chi-square statistic (observed.. aka x^2)**
5. Work out the 'chi critical'

    d.f. = (number of rows – 1) x (number of columns – 1)

    Assume α = 0.05, one tail, the upper tail.
6. So if observed > critical, we reject Ho (null). **Thus we can conclude that the 2 events are not independent.**

    Likewise, if **test statistic < critical value,** events are **independent.**

Lecture 10: Introduction to Regression

Chapter 13: Simple Regression Analysis

Independent Variable – are not influenced by anything

Dependent Variable is influenced, by the independent variable.

e.g.

Independent Variable = hours of study

Dependent Variable = success

We use regression to see if these independent variables really do influence the dependent variables.

Correlation only helps us to determine relations, not causation. i.e. just because something is correlated, doesn't mean that one thing causes the other.

Simple linear regression lets us say "yes if you increase study, your marks will go up"
It shows **the relationship between 1 dependent and 1 independent variable**.

Equation of a line in stats is: $y = b_0 + b_1x$
Where:
y = dependent variable (y axis on all graphs)
b_1 = gradient
b_0 = independent = y intercept
x = independent variable (x axis on all graphs)

Regression tells us if the variable is having a significant impact and quantifies this impact.

- Every time x goes up by 1, y goes up by the value of the gradient.

Spurious Correlation is where there is a correlation, but no theory underlying it. This can be used to mislead – correlation ≠ causation.

Regression assumes underlying theory when choosing which variable is the dependent and the independent.

Least-Squares Slope and Intercept (Simple Linear Regression) – Lecture 10, Slide 12.

$y = b_0 + b_1 x$

b_0 = y intercept (where x=0)

$$b_1 = \frac{\sum(x_1 - \bar{x})(y_1 - \bar{y})}{\sum(x_1 - \bar{x})^2}$$

- find b_1 then b_0 – that is easier!

Calculate Coefficients, using above formula:

Example

x	y	(x-$\bar{x}$)	(y-$\bar{y}$)	(x-$\bar{x}$)^2	(x-$\bar{x}$) x (y-$\bar{y}$)
5	4	-1.6	-1	2.56	1.6
7	5	0.4	0	0.16	0
4	3	-2.6	-2	6.76	5.2
8	6	1.4	1	1.96	1.4
9	7	2.4	2	5.72	4.8
$\bar{x}$ = 6.6	**$\bar{y}$ = 5**			**Sum = 17.2**	**Sum = 13**

Thus, $b_1 = \frac{13}{17.2} = 0.756$

b_0 was given as 0.011628 in the question.

Thus –

y= 0.011628 + 0.756 X

Further; if x=6, what is y? just sub in!

y= 0.011628 + (0.756 x 6)

y= 4.55….. so if x = 6% unemployment rate, there would be 4.55 crimes per 100 people.

(blank table for exam)

x	y	(x-$\bar{x}$)	(y-$\bar{y}$)	(x-$\bar{x}$)^2	(x-$\bar{x}$) x (y-$\bar{y}$)
$\bar{x}$ =					**Sum =**

When does the intercept (b_0) have a meaning?

- When the data is ratio data; i.e. when 0 has a meaning.

E.g. from previous example (tute 10, q 3), when the unemployment rate is 0,

y= 0.11628 + (0.756 x 0)
= 0.011628
Therefore, there will still be SOME crimes occurring, despite full employment.

Prediction Error (aka Residuals)
Each observation has a residual. The residual is "what is left over".
Residual = Actual Y – Estimate Y
The regression line / line of best fit attempts to minimise errors

NB: on the regression line, a dot above the line = positive error… a dot below the line = negative error.
The sum of all errors must = 0.

Coefficients – Excel Output:
Degrees of freedom = n – k – 1
Where:
n = number of data
k = number of x variables (always one in this topic, but in next topic, can be more)

Always a t distribution.
Always 2 tailed test (alpha is always 0.025)
Null hypothesis is 0

Outcomes – Coefficient of Determination – EXAM THING!!!
We want to know, what percent variation in y is explained by x?
To do this, we use the R^2 value, which is given.
This is the 'coefficient of determination' – a percentage of variability in y explained by the x variables as a group.
We use R^2 to decide how good our model is… e.g. if the R^2 is 0.92, it accounts for 92% - which is good… where as if our R^2 is 0.20, it means only 20% of the variability in Y is explained by our x's… thus, our model isn't as good.

Excel Output: Testing Estimates:

Example:

	Coefficient	Std Error	T Stat	P Value	Lower 95%	Upper 95%
Intercept	44.12	135.82	0.32	0.76	-305.00	393.25
Temp	47.58	6.11	7.79	0.00	31.88	63.27

Check the hypothesis that the coefficients are significantly different from 0.

Critical t = 2.57 (7-1-1 = 5 d.f. and alpha = 0.025)

$$t\ stat = \frac{coefficient}{standard\ error}$$

Often it is GIVEN IN THE TABLE!! So look!!

Ho: $b_o = 0$ Ha: $b_0 \neq 0$	Ho: $b_1 = 0$ Ha: $b_1 \neq 0$
$t\ stat = \frac{44.12 - 0}{135.82} = 0.32\ (given\ in\ table!)$	$t\ stat = \frac{47.58 - 0}{6.11} = 7.79\ (given\ in\ table!)$
Do not reject	Reject

We can again use P Values:

When α > P value, reject the null!

Excel Outputs: Confidence Intervals:

Example:

	Coefficient	Std Error	T Stat	P Value	Lower 95%	Upper 95%
Intercept	44.12	135.82	0.32	0.76	-305.00	393.25
Temp	47.58	6.11	7.79	0.00	31.88	63.27

Confidence intervals: **Coefficient ± (standard error) x (critical t value)**

NB: Ingo possibly screwed up – the last two columns of the table "lower 95%" and "upper 95%" show the confidence intervals… so check it out ☺

Again note:

Always a t distribution.

Always 2 tailed test (alpha is always 0.025)

Other things to think / remember about:

Predicting beyond the data we have is ‘dangerous’.
See Ingo’s example in lecture 10, slide 23

Outliers cannot be removed unless we are told to – because they are possibly legitimate.

REMEMBER THE EQUATION OF THE REGRESSION LINE IN STATS IS:

$$y = \ b_0 + b_1 X$$

If letters are in Greek, it’s population data
If letters are in Latin, it’s sample data

If it has a hat, it’s estimated… e.g. $\hat{y}$

Assumptions with Regression:

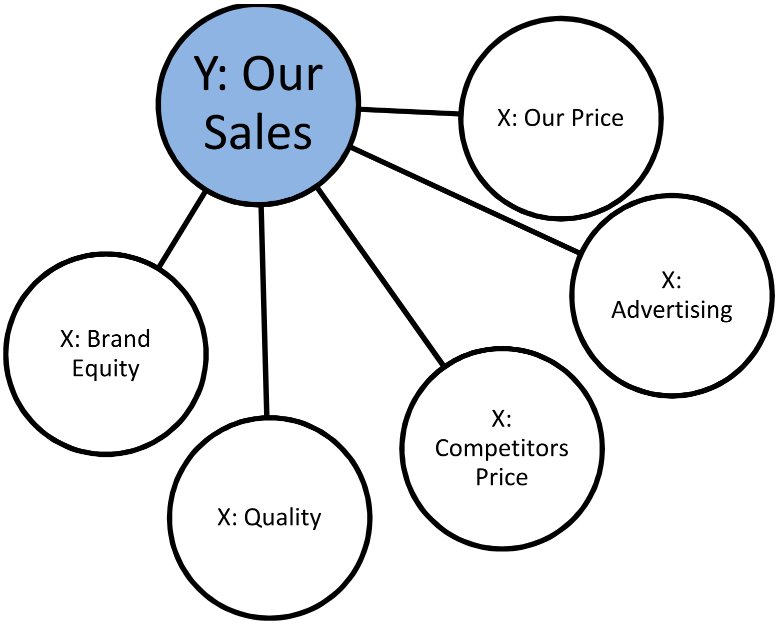
- Random
- Constant variance
- Sum of errors = 0

Lecture 11: Additional Topics in Regression

Chapter 14: Multiple Regression Analysis

Regression is a causal model.

Dependent and Independent Variables: Multiple Regression:



Simple Linear Regression:

$y = b_0 + b_1 X$

Multiple Regression:

$y = b_0 + b_1 X_1 + b_2 X_2 \ldots \ldots.. + b_k X_k$

A=0.025 (always 2 tailed, so alpha is always 0.025)

k=number of independent (x) variables. Use for degrees of freedom = n-k-1

(i.e. just as many $+b_1 X_1$ until all the x variables are put into it)

Example:

$y = b_0 + b_1(brand\ equity) + b_2(competitors\ price) + b_3(our\ price)\quad etc$

Example:

$y = 5000 - 200x_1 + 10x_2 + 25x_3$

Where:

Y=sales in dollars

X1 = our price

X2 = our average quality rating

X3 = our advertising spending.

Predict y when our price is $10, our average quality rating is 5 and our advertising spending is $200.

$y = 5000 - 200(10) + 10(5) + 25(200)$

=$8050

We can also just look at the **marginal effect** – this is just the coefficient of which ever x variable is asked… e.g. Marg

Qualitative Variables using Dummy (indicator) coding:

Just where you use a 0 or a 1 to replace words…
e.g. 0 = not in city, 1 = in city
Example:
What would sales be predicted if price = 20, quality = 50, advertising =100 and the store is in the city?
(formula was given)

$$y = 5000 - 200(10) + 10(50) + 25(100) + \mathbf{30(1)}$$

In city, y = 4030

Multiple Regression – Example Excel Output – Table of Coefficients (L:11 – S:9)

	Coefficients	Standard Error	T Stat	P Value
Intercept	-48.3548382	28.86499897	-1.675206649	0.100830166
Radiation	0.0262412	0.031892314	0.822811036	0.414954509
Temperature	1.488901257	0.344534142	4.321491187	8.4562E-05
Wind	-2.931705667	0.835403985	-3.50932689	0.001032728

Data was collected over 49 days.

Is the impact of temperature on ozone significant?
Equation would be:

$$y = 48.35 - 0.03x_1 + 1.489x_2 - 2.93x_3$$

d.f. = n-k-1 = 49-3-1 = 45
α=0.05/2 = 0.025
thus T CRITICAL = 2.015

because Ingo already included the t stat (4.321491187) in the table, we can easily just compare…
4.321491187>2.015
Therefore, it's in the reject region, so we reject the null hypothesis. (Remember the null is always that there is no significant impact (Ho = 0)… therefore by rejecting it, in this case, temperature may have a significant impact.)

P – Values

As with simple linear, we can just use P-values… they're heaps easier!

RULE:
If P > 0.05 – we reject Ho
If P ≤ 0.05 – we reject Ha

Thus in this instance… p= 8.4562E-05 which is 0.000084562

0.025 > 0.000084562… so again, we **reject**.

Coefficients – check the sign!

Put simply,

$$b_1 \text{ is the impact of that factor on } X_1, \text{holding all else constant}$$

- Regression should only be done if we believe variation in Y can be explained by X.
- We should have **a priori** ideas about what to expect.

Output – R^2 again

- Same as simple linear regression, explains that “this % of variation on Y is explained by X”
- Goes up as u ad more x variables (k’s)
- The higher, the better our equation is.

Adjusted R^2

- Do not use it.
- Need to know the theory behind it – which is that:
    - It adjusts your R^2 so as to penalise you for adding too many x variables (k’s)
    - Does same thing as normal R^2 in that it tells you a percentage.
    - If an extra x variable actually DOES NOT contribute to the overall change in Y, the Adjusted R^2 will actually decrease!! (wow… how exciting!)

Multiple Regression – Overall Significant

- Tests how well our GROUP of X variables are at explaining Y
- Uses **F Distribution**
- A = 0.05, 1 tailed.

H_0: the set of variables **DO NOT** predict Y very well
H_a: at least one X variable from the set predict Y **better than chance.**

Example: (See L:11, S:17)
H_0: neither radiation, temperature or wind predict ozone levels, better than chance
H_a: either radiation, temperature and / or wind can predict ozone levels better than chance.

How is F Calculated?

	df	**SS**	**MS**	**F**	**Significance F (like P Value)**
Regression (numerator) (=k)	3	26751.34268	8917.11426	22.0056589	6.38267E-09
Residual (denominator) (=n-k-1)	45	18234.8614	405.2191423		
Total (n-1)	48	44986.20408			

F Test Stat is Given… 22.0056589

F Critical is from the F Distribution table… 2.81

Thus, we REJECT the null.

As such, at least one X variable from the set predicts Y **better than chance.**

Significance F
Alike the p value, same rules:
RULE:
If F > 0.05 – we reject Ho
If F ≤ 0.05 – we reject Ha

When F is NOT rejected → you get left with the constant (i.e. all the x's = 0)
When F is rejected → the variables work well as a group (i.e. this is a good thing!)

read through lecture 11, slides 21 – 22 for some notes about testing individual coefficients
see slides 24 – 25 for 'issues' such as multi-collinearity