

Decision Making – As a manager, strategic, tactical and operational decisions are made on a frequent basis and the information that is used to formulate these decisions is derived from statistics.

Business Analytics – the scientific process of transforming data into insight for making better decisions. The tools of business analytics can aid decision making by creating insights from data, aids for more accurately forecast for planning and yielding better alternatives through analysis and optimization.

Analytical Methods and Models

Data collected from the past

Descriptive Analytics – encompasses the set of techniques that describes what has happened in the past (describing characteristics of a set of data) **eg.** data queries, reports, descriptive statistics

Diagnostic – discovering patterns or correlations in data

Data for the predictions of the future

Predictive Analytics – consists of techniques that use models constructed from past data to predict future outcomes **eg.** past data on product scales

Prescriptive Analytics – using models to design actions that will influence outcomes, the output of a prescriptive model is a best decision **eg.** price markdown models in retailing

Big Data – a set of data that cannot be managed, processed, or analysed with commonly available software in a reasonable amount of time **eg.** Facebook processes more than 250 million picture uploads per day

Sample: In many cases it is not feasible to collect data from the population of all elements of interest, therefore data is collected from a subset of the population known as a sample

Types of data:

Numerical (discrete, continuous) and Categorical (qualitative)

Data Sources

- Data comes from many sources; company databases, government reporting, market research etc
- The challenge lies with linking in the data about one subject, thus loyalty programs are so valuable
- Surveys are used to extract target information and they allow us to relate information together

example: characteristics (income, age) and preferences (about facilities that would need to be provided)

Challenge with surveys: questions must extract the right information

- they do not have an ambiguous meaning
- impartial (can't be biased or give hints)
- wording that does not invite dishonesty (domestic violence example, choosing the right time)

✚ Samples need to be representative, a non-representative sample is biased

Sources of Bias:

Selection Bias – the sample is chosen in a way that makes it not representative (ie. choosing first 50 people who watched the video to give feedback) **solution???** take a random sample that represents the characteristics of the student body

Non-Response Bias – not all who are selected for a sample will respond, if there is a pattern in the non-responses, this creates bias in the sample (ie. people who could not be bothered responding)

Descriptive Analytics

- Always start with a question of interest and ensure that the data you are dealing with is relevant for the question
- Data might have to be scaled or adjusted before meaningful descriptions can be made

Scaling Data: data must be scaled appropriately to suit the question of interest (refer to CO2 example in booklet)

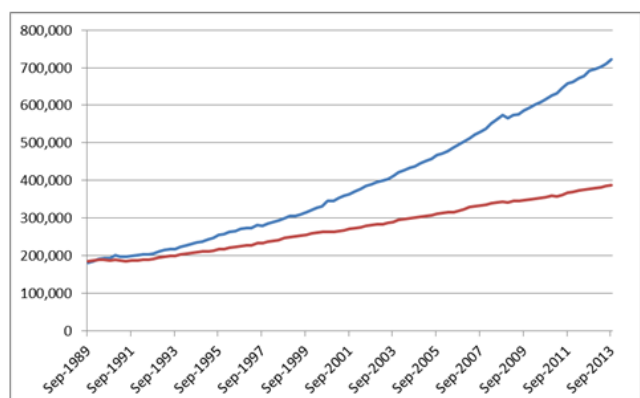
- when looking at CO2 emissions (kilo tonnes pa) it is evident that China has the most, however this may be because of their large population. Use this measurement when looking at environmental policies
- when looking at CO2 emissions (tonnes per capita (person) pa) China drops off the top list, and therefore this scaling should be used to look at individual countries

Types of scaling:

- obtaining relative measures (ie. percentages, proportions, per-person, per-hectare etc)
- adjusting for price differences/inflation **example:** Gross Domestic Product in Australia in booklet

Adjusting for Inflation

- Nominal/Actual/Current Prices – the \$\$ value of a series is actually measured at each point in time
- Real/Constant Prices – the value of a series would have taken if prices remained fixed at the same point in history (the “Base” period)
- Consumer Price Index (CPI) – A weighted average of prices of consumer items, indexed to 100 in the “Base” period
- $REAL = \frac{NOMINAL}{(CPI/100)}$
- The CPI tells us how much prices have changed over time and when we adjust by the CPI, we are essentially taking out the part of the increase that is due to inflation
- Where Nominal and Real GDP are similar – the year the CPI has its base of 100
- The Nominal series have the seasonal pattern removed so the trend is more easily seen



Nominal – increases because of price rise

Real – “true” increases

Descriptive Analytics – Categorical Data

Describing Categorical Data – Frequency Distribution/Pivot Table, Visualising with a Bar Chart or a Pie Chart

Describing Bivariate Categorical Data – Pivot Table/Frequency Distribution/Contingency Table

- different ways to present the pivot table (counts, % of total, % of column, % of row)
- **example:** Employee training scenario in booklet
- rows = took part in training
- columns = became employed

Count of Individual		Column Labels	
Row Labels	no	yes	Grand Total
no	69.23%	30.77%	100.00%
yes	4.55%	95.45%	100.00%
Grand Total	55.00%	45.00%	100.00%

- when deciding which way to represent the data; if we think one factor might cause the other, we take the **% of causal factor** (ie. training causes employment, so use % of row)
- Using a pivot table allows us to summarise the overall pattern in the data and see a bigger picture

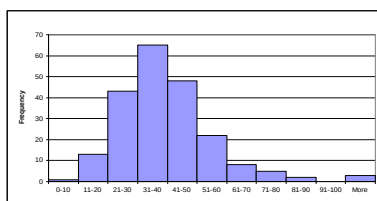
Descriptive Analytics – Numerical Data

Describing Numerical Data in Tables

- ❖ Frequency Distribution Table (upper limit of category with frequency, best stated in percentage terms)
- ❖ Cumulative Distribution Table (can be constructed from the frequency distribution table)

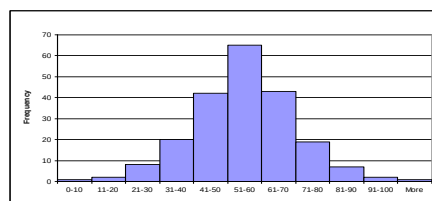
Histograms – used to describe the shape of a distribution of data (no spaces between bars)

Positively skewed (tails off to the right)



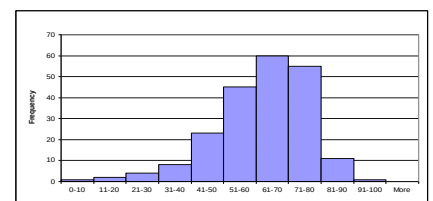
Mean > Median

Symmetric



Mean = Median

Negatively Skewed (tails off to the left)



Mean < Median

- ✚ For numerical data we can describe data by looking at the average and spread
- ✚ **Average** = Mean, formula (ie. =Average(C1:C1222) **example:** The average number of doctor's visits per quarter is 2.28.
- ✚ **Median** = the middle point, 50% of values are equal to or higher than the median, and 50% are equal to or lower, formula (ie. =Median(C1:C1222) **example:** The median number of doctors' visits per quarter is 1
- ✚ Variances with the mean and median relate to the skewness of the distribution
- ✚ Mean = median (symmetric distribution)
- ✚ Mean > Median (positive skewness) – higher values pull up the mean

✚ Mean < Median (negative skewness) – lower values pull the mean down

Variation in Data:

- **Range** (or spread of the data), formula (ie. =max(C1:C1222)-min(C1:C1222))
- The **range** can be **influenced** by one unusually **extreme value** **example:** *The range from the fewest to biggest number of doctors' visits per year is 60 visits. However, only one person had 60 visits, next most common was 30. The data is not as spread as the range of 60 suggests*
- **Standard deviation**, formula (ie. =stdev.s(C1:C1222))
- **Simple Explanation:** roughly the standard deviation measures the average amount the X values vary from the mean, above or below **example:** *The standard deviation of doctors' visits per quarter was 3.8019, which measures the average amount the X values varied from the mean, above or below*
- **Technical Explanation:** Explaining the formula

The term inside the bracket gives the difference between the actual value (x_i) and the mean value ($\bar{x}$). (the deviations from the mean)

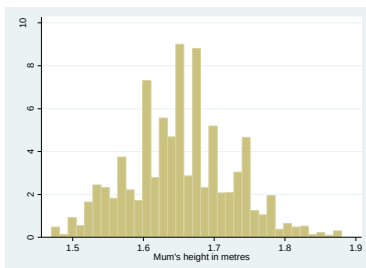
This will produce both positive and negative results. This is then squared to remove the negative differences as we are only concerned with the size and not direction of the deviation.

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

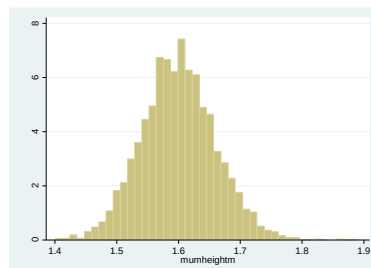
We then average these over (n-1) observations and take the square root to remove the effect of squaring the deviations.

Example:

Heights of Australian mothers



Height of Zimbabwe mothers



Australia has a higher mean, and Australia is more spread

The Normal Distribution (Bell Curve)

- symmetric around the mean
- values become less likely the further they are from the mean
- very unlikely to get odd/extreme values
- 67% of the data is within 1 SD of the mean
- 95% of the data is within 2 SDs of the mean
- 99.7% of the data is within 3 SDs of the mean

Standard Normal Probability Distribution

- suppose X has a normal distribution with a mean (μ) and standard deviation (σ)
- different means shift the distribution sideways along the axis, but don't change the shape
- changing the SD affects the spread of the normal, smaller SDs make it narrower, bigger SDs spread it out more

- $Z = \frac{X - \mu}{\sigma}$ Z scores represent how many standard deviations below or above the mean the X value is
- In lecture example: Looking at z scores of child's heights from Zimbabwe, even though they tend to follow a normal distribution for the most part, there are often a small number of extremes/outlier values that can distort the mean. Therefore in the z score calculations the **median** was used instead of the mean
- Using the median protects us from this distortion, being a more robust measure of central tendency

Diagnostic Analytics – Categorical Data

Diagnostic Analytics is an exploratory process of discovering patterns in data, usually used to discover relationships between different variables/factors.

Diagnostics with Bivariate Categorical Data

Count of Individual Column Labels			
Row Labels	no	yes	Grand Total
no	69%	31%	100%
yes	5%	95%	100%
Grand Total	55%	45%	100%

Pivot Table as % of Row

- rows = took part in training
- columns = became employed

These are conditional probabilities:

- Prob (employed, given took part in training) = 0.95
- Prob (employed, given no training) = 0.31

Independent Events

$$\Pr(A | B) = \Pr(A | B') = \Pr(A)$$

If these do not equal each other, then the characteristic of B is not independent of the characteristic of A

Example: Check whether demand for beer is independent of gender

$$\Pr(\text{spend } \$0 | F) = \Pr(\text{spend } \$0 | M) = \Pr(\text{spend } \$0)$$

$$\frac{61}{94} \neq \frac{45}{106} \neq \frac{106}{200}$$

Therefore they are not independent; gender does make a difference

Beer Expenditure (\$ per annum)	Females	Males	Total
0	61	45	106
1-100	5	4	9
101-500	17	29	46
501-1000	9	22	31
More than 1000	2	6	8
Total	94	106	200

Same Contingency Table as % of Total:

- Intersection/joint probabilities

eg. $\Pr(X=\text{yes} \cap Y=\text{yes}) = 0.21$

- Marginal probabilities

eg. $\Pr(X=\text{no}) = 0.78$

eg. $\Pr(Y=\text{yes}) = 0.45$

Count of Individual Column Labels			
Row Labels	no	yes	Grand Total
no	54%	24%	78%
yes	1%	21%	22%
Grand Total	55%	45%	100%

Example: % of Total

Count of C22a. E-mail to communicate with clients or suppliers			
		Column Labels	
Row Labels	No	Yes	Grand Total
Large >=100	0.67%	1.33%	2.00%
Medium >= 20 and <=99	12.67%	26.67%	39.33%
Small >=5 and <=19	44.00%	14.67%	58.67%
Grand Total	57.33%	42.67%	100.00%

Joint Probabilities

Pr (medium size $\cap$ not use email) = 0.1267

Pr (small size $\cap$ not use email) = 0.44

Marginal Probabilities

Pr (use email) = 0.4267

Example: % of Row

Conditional Probabilities

25% of small firms use email

67% of large firms use email

Count of C22a. E-mail to communicate with clients or suppliers			
		Column Labels	
Row Labels	No	Yes	Grand Total
Large >=100		33%	67%
Medium >= 20 and <=99		32%	68%
Small >=5 and <=19		75%	25%
Grand Total	57%	43%	100%

Diagnostic Analytics – Comparing Across Groups

- To discover patterns in data, we can break up the people into categories, and compare doctors visits across the categories
- example:** do men visit the doctor more or less than women? Do people with higher incomes visit the doctor more or less than others?

C					E				
AGE	DOCVISIT	EDUC	MALE	Income	SUMME	SPRING	AUTUM	WINTER	
48	0	14.5	1	3467.818	1	0	0	0	
52	2	18	1	3152.017	1	0	0	0	
49	0	12	1	1500.96	0	0	0	0	1
59	4	11.5	1	2969.174	1	0	0	0	0
39	6	10.5	1	6416.904	1	0	0	0	0
40	12	9	1	780.4994	0	1	0	0	0
58	1	16	1	2619.857	0	0	0	0	1
45	0	11.5	1	1746.571	0	0	0	0	1
28	0	11.5	1	2825.337	1	0	0	0	0
60	0	12	1	2486.298	0	1	0	0	0
51	3	11.5	1	2134.699	0	1	0	0	0
27	0	13	1	1951.249	0	1	0	0	0
24	0	10.5	1	3274.823	0	0	0	0	1
23	4	10	1	1582.19	0	1	0	0	0
33	0	18	1	1245.242	1	0	0	0	0

Comparing Gender: using sumif and countif functions

- =sumif(E2:1222," = 1", C2:C1222)/countif (E2:E1222," = 1") = $\bar{X}$ male = 1.871
- for females, same formula is used except E2:1222 = 0, = $\bar{X}$ female = 2.682

Comparing Income: divide groups: low and high (median = \$1,779, so use \$1,800 as the dividing point)

$$\bar{X}_{\text{low}} = 2.272 \quad \bar{X}_{\text{high}} = 2.289$$

- these are different but compared to differences across gender, this difference is very small

✚ There may be reasons to compare other characteristics of data across groups, such as:

Spread

Example: compare standard deviation for number of doctor's visits in winter compared to other seasons:

$$s_{\text{winter}} = 3.26, s_{\text{other seasons}} = 4.06$$

Values of particular interest

Example: what proportion of people have no doctor's visits? Does this vary across seasons?

Full sample: 38.0%

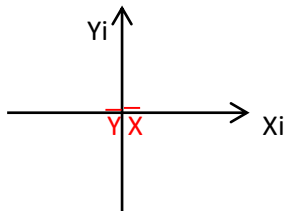
Winter: 37.9% Other Seasons: 38.1%

Diagnostic Analytics – Correlation

Correlation – a measure of relationship between numerical variables

Scatter Diagram: Scatter Diagrams help address 3 questions. Is there a relationship between X and Y? (any pattern?) What direction is the relationship? And is the relationship linear?

❖ Correlation measures the strength and direction of the linear relationship between X and Y



$$\rho = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{X}}{s_x} \right) \left(\frac{y_i - \bar{Y}}{s_y} \right)$$

- ❖ Product of deviations of x_i from $\bar{X}$ and y_i from $\bar{Y}$
- ❖ Deviations can be positive or negative, so their product and average can be positive or negative
- ❖ Correlations range between -1 and 1 with 0 having no correlation

Correlation is not Causality (could be a number of other factors leading to an outcome)

Outliers can have a big impact on the correlation value in the data set

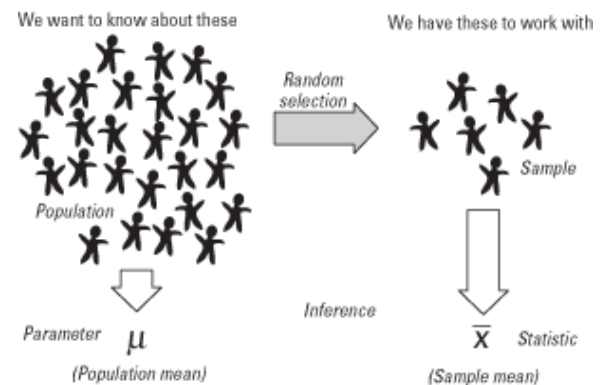
Predictive Analytics – Estimation

Real Difference vs. Chance Difference

- Some differences in means and proportions are small, and some correlations are small
- Small differences are probably just zero and the non-zero value was due to random variation
- How do we decide what is small?
- Need to identify real/ substantive differences vs. small/coincidental differences
- Real, substantive differences are likely to persist so relevant to decisions and predicting behaviour
- Predictive Analytics requires ideas of **Statistical Interference**

Statistical Interference – use a sample of data to estimate characteristics of a population

A sample statistic is used as an estimator of a Population Parameter



Population Parameter	Sample Statistic
Mean μ	Sample Mean $\bar{X}$
Proportion π	Sample Proportion p
Difference between means $\mu_2 - \mu_1$	Difference between sample means $\bar{X}_2 - \bar{X}_1$
Difference in proportions $\pi_2 - \pi_1$	Difference in sample proportions $p_2 - p_1$
Correlation ρ	Sample Correlation r
etc	etc

- Whenever we use a sample statistic to estimate a population parameter, there is uncertainty; the estimate is only an approximation to the true value
- If we take a new sample, we will get a different estimate. Both cannot be correct!
- The uncertainty is captured with the ideas of confidence intervals, hypothesis testing, standard errors

Students T Distribution

- continuous distribution - it is similar in shape to the normal distribution as (mean = 0 and sd = 1) but with more area in the extremes, it is bell shaped but more spread out
- Requires input of degrees of freedom, ie. if $n = 10$ then degrees of freedom = 9

Predictive Analytics – Confidence Intervals

Confidence intervals are a way of showing the uncertainty in estimating population parameters with sample statistics

- To calculate confidence intervals in excel, create a new column titled mean and input 1's in every row (eg. Y = no. of doctors' visits and X = columns with all values equal to 1)
- ✚ Short cut methods to calculating confidence intervals, using the binomial distribution
- ✚ Sample proportion p has binomial distribution: mean = π and standard deviation = $\sqrt{\frac{\pi(1-\pi)}{n}}$

✚ 95% confidence interval is: $p - 1.96 \sqrt{\frac{p(1-p)}{n}}$ to $p + 1.96 \sqrt{\frac{p(1-p)}{n}}$

✚ If you know n and sample size you can construct the confidence interval

Confidence intervals for correlation

- What is the correlation between number of doctors' visits (Y) and years of education (X)?
- Scale Y and X by their standard deviations $Y^* = y/S_y$ and $X^* = x/S_x$ (these give you the 95% confidence interval for the correlation)

Predictive Analytics – Sampling Distribution Theory

Confidence intervals are constructed using some statistical theory for sampling distributions

Sampling Distribution for Sample Mean

Sample Mean – take a random sample of n values and calculate a sample mean ($\bar{X}$)

Important Concept: sampling distribution of $\bar{X}$ is the theoretical probability distribution of all the possible $\bar{X}$ we would get if we took many, many samples and calculated $\bar{X}$ for each of them. (this is difficult as we don't take many samples)

- ❖ $\bar{X}$ has a mean μ
- ❖ $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$
- ❖ Since the data is from a normal distribution, the distribution of $\bar{X}$ will also be normal
- ❖ The sampling distribution of $\bar{X}$ is thus $\bar{X} \sim (\mu, \sigma^2/n)$
- ❖ $\Pr(z > 1.96) = 0.025$, and $\Pr(z < -1.96) = 0.025$.
- ❖ So there is a 95% chance that z is between -1.96 and +1.96.

Lecture 7 Sampling Distribution Mean Example

$\bar{X} \sim (\mu, \sigma^2/n)$

- Call centre designing a good monitoring system for average phone call ratings
- Decide an acceptable minim average rating (eg. 8 out of 10)
- Each day take a random sample of calls to be monitored, listen to these and give rating. Calculate the mean of that sample
- If the true mean equals 8, 50% of the time we will get a sample mean less than 8
- We are looking for a sample mean much less than 8, suggesting the population mean is unlikely to be 8 and if that is the case we have a problem
- We use the sampling distribution to decide a rule when to flag a problem (C).
- We will say there is a problem when the sample mean falls below C, where C is defined by;

$$\Pr(\bar{X} < C | \mu = 8) = 0.05$$

- For a variable Z that follows the standard normal distribution, we can use Excel to workout

$$\Pr(Z < -1.645) = 0.05$$

- This can be used to work out the appropriate value of c;

$$Z = \frac{C-8}{\sigma/\sqrt{n}} \quad Z = -1.645 \quad \text{Therefore, } C = 8 - 1.645\sigma/\sqrt{n} \quad (\text{Cut off})$$

- The population SD is unknown so we use the sample SD to estimate it, and use Student's T distribution instead of the standard normal to calculate our value of C
- If $n = 20$, then $t_{.05,19} = 1.729$
- So

$$C = 8 - 1.729 S / \sqrt{20}$$

Theory of Confidence Intervals

- To obtain confidence intervals, need an estimate of standard deviation σ . Use sample standard deviation s .
- This change introduces a little extra uncertainty so now use Student's t Distribution
- So 1.96 changes to something a bit bigger. Value depends on n , and will be very similar to 1.96 for larger n
- Confidence Interval Formula:

$$\bar{X} - t_{.025, n-1} S / \sqrt{n} \quad \text{and} \quad \bar{X} + t_{.025, n-1} S / \sqrt{n}$$

SUMMARY OUTPUT						
Regression Statistics						
Multiple	0.5144802					
R Square	0.2646899					
Adjusted	0.2638702					
Standard	3.8018768					
Observat	1221					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	1	6347.793612	6347.794	439.164	1.59E-83	
Residual	1220	17634.20639	14.45427			
Total	1221	23982				
Coefficients						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A
Mean	2.2800983	0.108802835	20.95624	1.56E-83	2.066637	2.49356

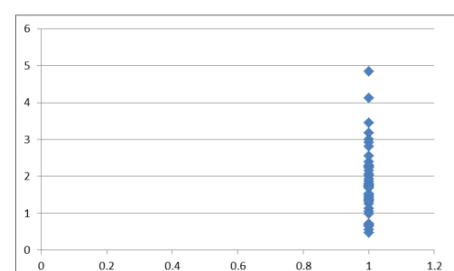
$$t_{.025,1220} = 1.96$$

Lecture 6 – Adult Income and Lifestyle Choices

- Estimating the mean household income
- Sample mean of \$1,677
- The sample mean is only an estimate, to capture the uncertainty we construct a confidence interval for the possible values

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
mean	1676.6	16.55413	101.2798	0	1644.144	1709.056	1644.144	1709.056

- We are 95% confident that the mean income for Australian households is between \$1644 and \$1709.
- Why does this regression estimate the mean?
- In this case all of our x values are equal to 1 so our scatter diagram would look like this
- Fitting a line through this data, there is no information about what would happen to y if we change x . The “best” horizontal line through this data will go through the mean and this is why we get the coefficient of the sample mean



Income and Age

Do we expect to see a correlation between household income and the parent's age?

- we can estimate the correlation and confidence interval by; transforming the data for income and age dividing each by their SD, regression of income on age, including intercept

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-0.06427	0.098953	-0.64952	0.516035	-0.25827	0.129724	-0.25827	0.129724
x/sd	0.22912	0.015816	14.48688	1.53E-46	0.198114	0.260126	0.198114	0.260126

- Correlation = 0.229, a reasonably strong positive correlation (point estimate)
- Confidence interval for the correlation: between 0.198 and 0.260 (interval estimate)

Differences in Means

Do males go to the doctor less often than females?

$$\bar{X}_{\text{male}} = 1.871, \bar{X}_{\text{female}} = 2.682$$

$$\text{So } \bar{X}_{\text{difference}} = 0.811$$

- Whilst this difference appears to be big, is it systematic or just due to a chance outcome?
- Construct a confidence interval for the difference
- Hypothesis testing allows significant effects to be identified

Regression	1	200.6262229	200.6262	14.02829	0.000188
Residual	1219	17433.58017	14.30154		
Total	1220	17634.20639			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	2.6818182	0.152370567	17.60063	5.82E-62	2.382881	2.980756
MALE	-0.8107438	0.216461781	-3.74544	0.000188	-1.23542	-0.38606

$$\bar{X}_{\text{male}} - \bar{X}_{\text{female}}$$

Confidence Interval

- We are 95% confident that on average, males pay

between 0.386 and 1.235 less visits to the doctor than females

Hypothesis Testing – is a formal approach in deciding if there is a difference between two groups

Null Hypothesis (H_0) – the conservative view we take unless there is clear evidence to reject it

Alternative Hypothesis (H_1) – the alternative view against which we are testing the null

$H_0: \mu_{\text{male}} = \mu_{\text{female}}$ (that is the mean number of visits is the same for males and females)

$H_1: \mu_{\text{male}} \neq \mu_{\text{female}}$ (that is there is a gender difference in mean number of visits)

Is the difference “statistically significant”?

- this is determined by looking at the p value of the output
- **reject H_0 if p value < significance level** (usually 0.05 or 5%)
- the p value is the probability of a value this far or further from H_0 assuming H_0 is true

Difference in Proportions

- For Doctor's visits, construct a variable = 1 if number of Doctors visits = 0, and = 0 otherwise
- Regress on intercept and Male dummy variable

Proportion of females
with no visits
Diff in proportion
(Male – Female)

	df	SS	MS	F	Significance F
Regression	1	9.234354	9.234354	40.42794	2.88E-10
Residual	1219	278.438	0.228415		
Total	1220	287.6724			

	Coefficient	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	0.293831	0.019256	15.25899	3.05E-48	0.256052	0.33161
MALE	0.173937	0.027356	6.358297	2.88E-10	0.120267	0.227607

- 29.4% of females had no doctor's visits
- 46.8% of males had no doctor's visits
- 95% confident that the difference in proportions is between 12% and 22.8%

Hypothesis

$H_0: \pi_{\text{male}} - \pi_{\text{female}} = 0$

$H_1: \pi_{\text{male}} - \pi_{\text{female}} \neq 0$

P value = 0.000000000288, therefore reject H_0 as there is a clear difference

How Regression Works

- ✖ A regression model defines a linear relationship between Y and X, with an error e

$$y = \beta_0 + \beta_1 x + e$$

- ✖ For this example, when $x = 0$, $y = \beta_0 + e$ (β_0 is the mean of y when a person is female, the proportion of females who never visited the doctor)
- ✖ β_1 is the difference in proportion for males compared to females

Hypothesis Testing in more Depth

Consider estimating the population mean with the sample mean $\bar{X}$ and using the estimator to test the following null hypothesis.

$H_0: \mu = 0$

This hypothesis can be tested using the **p value approach**. A p value less than the significance level suggests we should reject the null hypothesis.

$H_1: \mu \neq 0$

$\bar{X}$

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A
Mean	2.2800983	0.108802835	20.95624	1.56E-83	2.066637	2.49356

- P-value for testing $H_0: \mu = 0$

output produced by using the column of 1's on the X variable

Reject the null hypothesis as p value is < 0.05

Errors in Hypothesis Testing

Type I Error: where you reject H_0 when H_0 is true (occasionally when we pick up a problem but there isn't one)

- ❖ significance level = the probability of a type 1 error (most commonly 5% of the time)

Type II Error: where you don't reject H_0 when H_0 is false (sometimes it takes a few days to pick up a problem)

Significance Level

- Choosing 10% means we reject H_0 when we have some evidence that it is false
- Choosing 5% means we only reject H_0 when we are reasonably confident that it is false (most common)
- 1% means we want convincing evidence before rejecting H_0

A smaller significance level reduces Pr(type 1 error) BUT increase Pr (type II error)

T Statistic Approach to Hypothesis Testing

ANOVA						
	df	SS	MS	F	Significance F	
Regression	1	6347.793612	6347.794	439.164	1.59E-83	
Residual	1220	17634.20639	14.45427			
Total	1221	23982				

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A
Mean	2.2800983	0.108802835	20.95624	1.56E-83	2.066637	2.49356

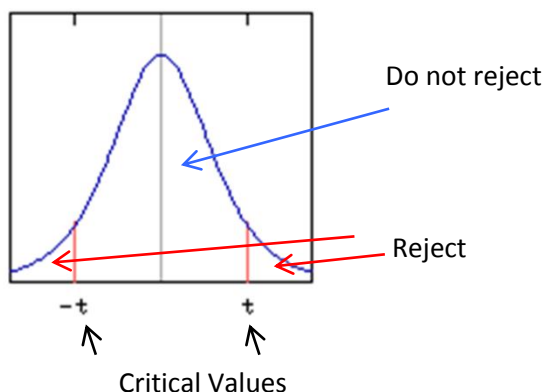
Degrees of Freedom

$\bar{X}$
 $s/\sqrt{n}$
 t-statistic for testing $H_0: \mu = 0$

- T-stat: guide of how close we are to the mean, if t-stat is close to zero don't reject the null
- $t - stat = \frac{(\bar{X} - \mu_0)}{s/\sqrt{n}}$
- μ_0 is the value of μ when H_0 is true
- $s/\sqrt{n}$ = the standard error
- The t-stat produced by excel assumes $\mu_0 = 0$
- The further $\bar{X}$ is from μ_0 , the further the t stat is from zero
- Use a **critical value** to determine how far the t stat needs to be away from zero to reject the null
- **If t stat > critical value, reject H_0**

The critical value is found from the student's T distribution

$$t - stat = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim \text{Student's } t \text{ with } df = n - 1$$



A big negative or a big positive value both indicate evidence against H_0

One Sided Hypothesis Test – T stat Approach

- Sometimes we know the direction of the alternative hypothesis, so we want to perform a one-sided test
- **example:** $H_0: \mu = 0$ against $H_1: \mu > 0$
- the only difference is the critical value, as all the probability is in one side of the t-distribution
- calculate the t statistics, see which direction the alternative view follows
- critical value will be closer to zero than for two sided test

One Sided Hypothesis Test – P-value Approach

Excel assumes a two sided test so we need to;

1. check which side of μ_0 , $\bar{X}$ is to see if there is any evidence against H_0 in the direction of H_1

example: if alternative is $H_1: \mu > 0$, then $\bar{X}$ values that are less than zero provide no evidence against H_0 in the direction of H_1 (looking for positive $\bar{X}$ values)

2. if $\bar{X}$ does vary from μ_0 in the direction of H_1 , **halve the p-value** in Excel output, and perform the test with the smaller p value

example: p value in excel gives 1.56×10^{-83}

change to 1 sided value (divide by 2) = 0.78×10^{-83} or 7.8×10^{-82}

Data Analytics Models – Linear Regression Model

Output (Y) – an outcome of interest

Inputs (X) – one or more variables that we believe cause Y or may be used to predict Y

Process – linear equation

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k + e$$

The model is: $\hat{y} = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k$

- Therefore, the error e is the difference between the actual data value (y) and the model's prediction ($\hat{y}$)
- β_0 is the model's prediction of y when all x values = 0
- β_1 is the change in the model's prediction for y for a one unit change in x_1 leaving other x 's unchanged (similar interpretations for the rest of the parameters)

Lecture 8 Worked Example

- 📊 A simple Linear Regression Model for Child Reading Score
- 📊 Y = child's reading score (for ages) out of 100
- 📊 X = hours of TV watched in an average week
- 📊 Output from Excel's Data analysis, regression;

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	43.92091983	1.012263583	43.38882	2.5E-290	41.935717	45.906123
Hours TV	-0.331215291	0.045417012	-7.29276	4.36E-13	-0.420285	-0.2421456

Model: $\hat{y} = 43.92 - 0.331x$

Intercept: $\beta_0 = 43.92$

The average reading score amongst children who watch no TV during a typical week is estimated to be 43.92 out of 100.

Hours TV: $\beta_1 = -0.331$

Take two children, the first of whom watches 1 more hour of TV per week than the second. The model predicts the first child to have a reading score 0.331 (out of 100) lower than the second child who watched less TV.

Is this effect statistically significant? Perform a Hypothesis Test

1. Formulate Null and Alternative

$H_0: \beta_1 = 0$ Watching TV has no impact on a child's reading score

$H_1: \beta_1 \neq 0$ Watching TV does have an impact on a child's reading score

2. Decide a "significance" level

Test at 5% significance ie. $\alpha = 0.05$

3. Calculate the p-value

P value from output = 4.36×10^{-13}

4. Make a decision

The decision rule is to reject $H_0: \beta_1 = 0$ if the p value $< \alpha$

Since $4.36 \times 10^{-13} < 0.05$, we reject H_0 and conclude that watching TV does have an impact on a child's reading score

🔧 How can the effect of TV not be practically important but statistically very important? May need to add in other explanatory variables and perform a multiple linear regression

Multiple Linear Regression for Income in South Africa – Lecture 9

- ❖ We are interested in the factors that contribute to variations in income for all employed South Africans
- ❖ What sort of independent variables should we use?

Years of Education (Continuous variable - change in x by one unit and how it affects y)

Age and Age squared (Quadratic – reasonable to expect age to have a non-linear effect on income)

Gender (2 category dummy – male = 1 if the person is male, 0 otherwise)

Rural Area (2 category dummy – rural = 1 if the person lives in a rural area, 0 otherwise)

Formal Sector Work (2 category dummy variable – formal = 1 if the persons occupation is considered formal, 0 otherwise)

Racial Group (multiple category dummy)

- Asian = 1 if the person is officially classified as Asian, 0 otherwise
- Black = 1 if the person is officially classified as Black, 0 otherwise
- Coloured = 1 if the person is officially classified as Coloured, 0 otherwise
- White = 1 if the person is officially classified as White, 0 otherwise

We need to leave one of these categories out as the “base” person

Summary Output

Some things to note from this output

Regression Statistics	
Multiple R	0.45832025
R Square	0.21005745
Adjusted R Square	0.20978353
Standard Error	2953.10043
Observations	25965

ANOVA					
	df	SS	MS	F	Significance F
Regression	9	60189403947	6.69E+09	766.8689	0
Residual	25955	2.26348E+11	8720802		
Total	25964	2.86538E+11			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-4024.90727	230.2115478	-17.4835	4.7E-68	-4476.13464	-3573.6799
Years education	155.298602	5.575820719	27.85215	3E-168	144.3696845	166.22752
Age	160.25091	11.78286097	13.60034	5.55E-42	137.1558508	183.34597
Age2	-1.5572072	0.150502523	-10.3467	4.84E-25	-1.85220047	-1.2622139
Male	702.109899	37.85555529	18.54708	2.68E-76	627.910916	776.30888
Rural	-167.603349	43.26951229	-3.87347	0.000108	-252.413987	-82.792711
Formal sector	464.462196	42.04836341	11.0459	2.65E-28	382.0450768	546.87931
Asian	861.094517	98.35219265	8.755214	2.16E-18	668.3187776	1053.8703
Coloured	169.063625	48.93105339	3.45514	0.000551	73.15605294	264.9712
White	2817.52716	60.92993739	46.24208	0	2698.101109	2936.9532

- All our explanatory variables are highly significant, meaning they all help to explain income
- R^2 is poor (only 21% of the variation in y can be explained by the variation in x)
- A standard error of 2953 Rand is extremely high, considering the overall mean income is 1992 Rand

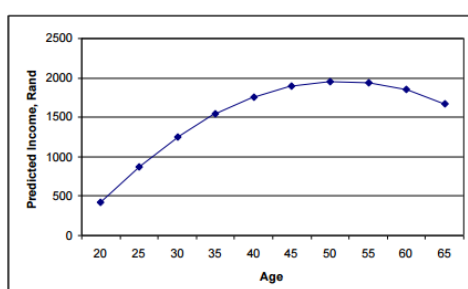
Intercept: $b_0 = -4024.9$ (this intercept makes no sense to interpret)

Years of Education: $b_1 = 155.3$

Take **two people**, who are the same age, same gender, live in the same area, are employed under the same employment sector and are the same race, except **one** has **1 more year of education** than the other. The **model predicts** the person with the extra years of education to **earn 155.3 Rand per month more** than the person with the lesser education.

Age: $b_2 = 160.3$ and **Age Squared:** $b_3 = -1.6$

Need to interpret these both together and physically calculate how $\hat{y}$ changes with different values of age. Plot a graph and interpret (ie. 20-65 years of age, 12 years of education and all other variables = 0)



The inverted U shape makes sense, as b_2 is a large positive and b_3 is a small negative. For small ages values b_2 component dominates whilst b_3 dominates that larger age component. The turning point makes sense too, once someone reaches the age of 50 they may reduce working hours, may not get anymore promotions etc.

Male: $b_4 = 702.109899$

Take 2 people the same in every other respect (age, education, rural/urban, employment sector, and race), except that one is male and the other is female. The model predicts the male to earn 702.1 Rand per month more than the female

Rural: $b_5 = -167.603349$

Take 2 people the same in every other respect, except that one lives in a rural area and the other lives in an urban area. The model predicts the rural resident to earn 167.6 Rand less per month than the urban resident

Formal Sector: $b_6 = 464.462196$

Take 2 people the same in every other respect, except that one has a formal sector job and the other has an informal sector job. The model predicts the person working in the formal sector to earn 464.5 Rand more per month than the person working in the informal sector

Race Dummies

Asian: $b_7 = 861.094517$

Coloured: $b_8 = 169.063625$

White: $b_9 = 2817.52716$

Take two people the same in every other respect, except that one is Asian South African and the other a Black South African. The model predicts the Asian South African to earn 861.1 Rand per month more than the Black South African

- repeat for coloured and white

Estimating Parameters in Linear Regression Models

- the parameters of the model are population values and are unknown
- use the sample to estimate these parameters (with hats on)
- to reflect the fact that sample estimates are only estimates, one can use sampling distribution theory to;
 - Construct confidence intervals for the true slope parameters
 - Test hypotheses about particular values for parameters

Estimates of Parameters

Residual	1214	17141.04	14.11948			
Total	1220	17634.21				
	Coefficient	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	3.977565	0.545068	7.297379	5.29E-13	2.908186	5.046944
EDUC	-0.14278	0.046674	-3.059	0.002269	-0.23435	-0.05121
MALE	-0.75453	0.215834	-3.49587	0.00049	-1.17798	-0.33108
Income	8.4E-05	0.000123	0.68079	0.496134	-0.00016	0.000326
SUMMER	0.222004	0.349067	0.635991	0.524902	-0.46284	0.906846
AUTUMN	0.065177	0.2398	0.271799	0.785823	-0.40529	0.535647
WINTER	1.778408	0.547687	3.247122	0.001198	0.703888	2.852924

Confidence Intervals for parameters

Example: Estimated slope parameter (and its confidence interval) for male suggest:

Take two people of same income, education and at the same time of year, but one male and one female. We are 95% confident the male will make between 0.33 and 1.18 fewer doctor's visits per quarter, on average.

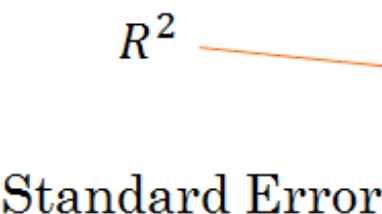
Example: Does lack of income limit people's ability to visit the doctor? Is there an income gradient for access to health services?

- If yes slope parameter for income will exceed zero
- $H_0: \beta_3 = 0$ vs $H_1: \beta_3 > 0$
- $\beta_3 = 0.000084$ which is away from zero in the direction of H_1 , so we need to calculate the p value (dividing by 2 because it is a one-sided test)
- p value = $0.496/2 = 0.248$
- p value is much bigger than 0.05 so do NOT reject H_0
- There is no evidence of an income gradient for health

Evaluating Regression Models

1. Careful thought on the selection of X variables (are the main important factors taken into account?)
 - Hypothesis tests for relevance of variables
2. Two guides to "goodness of fit":
 - R^2
 - Standard Error

Example: Number of Doctor's visits per quarter



Regression Statistics	
Multiple F	0.167231
R Square	0.027966
Adjusted R Square	0.023162
Standard Error	3.757589
Observations	1221

R^2 – measures the proportion of variation in y that is explained by the model

The closer R^2 is to 1.0, the better the model. An R^2 of zero indicates a model explains nothing of the variation in y

$R^2 = 0.028$

This value suggests that education, gender, income and season (all the x variables) can only explain 2.8% of the variation in number of doctors' visits people make

- ✖ Adding more x variables always increases R^2 so it is an unreliable way to determine which x variables to include. Use hypothesis test for that purpose

Standard Error – measures the standard deviation of the estimated error term from the model

- Since the error term has a mean of zero, the standard error is the "square root of the average squared error"; a good guide to how big the typical error would be if we used the model to predict y

Standard error = 3.76 doctors' visits per quarter

Is this error big or small? (compare to the mean)

- ❖ Since the average number of doctors' visits is 2.28 per quarter, this is actually a big error! It is hard to predict people's visits to the doctor just based on a few demographic characteristics

- ❖ This model is not much use for predictive analytics (using it to predict people's behaviour, shouldn't stereotype based on gender, education and income)
- ❖ But this model is useful for Prescriptive Analytics; we learn a lot about variations across income, education and gender that inform public health campaigns, health funding models etc.

Regression Models for Prediction

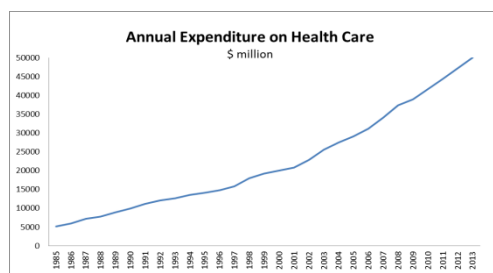
Prediction with models is mostly used to predict future values – “Forecasting”

- Time Series Models include specific set of X variables that capture the main features of Y
- Most common features to capture – trend and seasonality

Trend – long term steady pattern in how a series evolves over time. Usually need 15-20 years of data to identify a trend

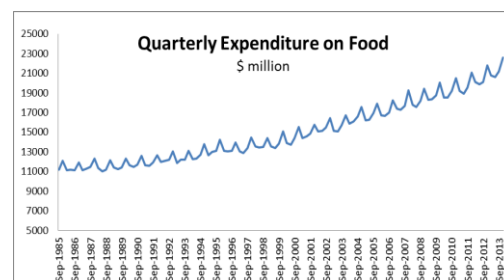
Trend Model – linear trend

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 t$$



Seasonality – a pattern that recurs in regular intervals across time (monthly, quarterly)

- in this graph there are three low quarters with a peak quarter



Cyclical – a pattern that occurs with peaks and troughs that takes over a year

Weird – any outliers or irregular components of the plot?

Linear Trend Model with quarterly seasonality

$$\hat{y}_t = \underbrace{\hat{\beta}_0 + \hat{\beta}_1 t}_{T} + \underbrace{\hat{\beta}_2 Q1_t + \hat{\beta}_3 Q2_t + \hat{\beta}_4 Q3_t}_{S}$$

Interpretation:

	A	B	C	D	E	F
	Quarter	Food Expenditure \$m	t	Q1	Q2	Q3
1	Sep-1985	11156	1	0	0	1
2	Dec-1985	12047	2	0	0	0
3	Mar-1986	11118	3	1	0	0
4	Jun-1986	11190	4	0	1	0
5	Sep-1986	11143	5	0	0	1
6	Dec-1986	11922	6	0	0	0
112	Mar-2013	20789	111	1	0	0
113	Jun-2013	20593	112	0	1	0
114	Sep-2013	21175	113	0	0	1
115	Dec-2013	22552	114	0	0	0

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	10672.5973	165.5211688	64.47875	1.23E-88	10344.53972	11000.6549
t	89.8392857	1.896164706	47.37947	1.48E-74	86.08114891	93.5974225
Q1	-1168.86515	176.5063532	-6.62223	1.38E-09	-1518.695	-819.03529
Q2	-1299.133	176.4961679	-7.36069	3.66E-11	-1648.94267	-949.32334
Q3	-989.643473	174.9513811	-5.65668	1.26E-07	-1336.39142	-642.89552

Intercept: the estimated trend value of the series at time $t=0$ (quarter 2, 1985) is \$10,673m

t: the estimated average quarterly growth in the series, after adjusting for seasonality is \$89.8m

Q1: the estimated food expenditure in quarter 1 is \$1169m less than quarter 4, after adjusting for trend

Q2: the estimated food expenditure in quarter 2 is \$1299m less than quarter 4, after adjusting for trend

Q3: the estimated food expenditure in quarter 3 is \$990m less than quarter 4, after adjusting for trend

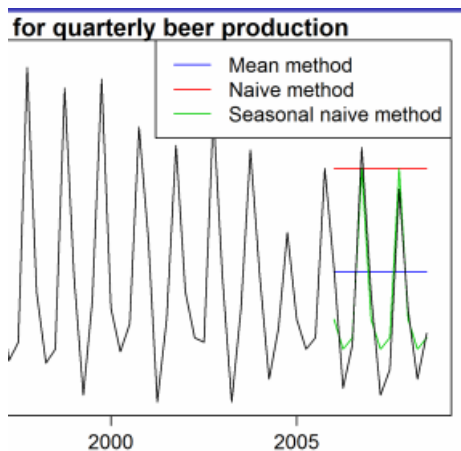
Forecast the Australian food expenditure for Q1 2014:

✓ $t=114$ corresponds to Q4 2013, therefore $t=115$ corresponds to Q1 2014

✓ $\hat{y}_{115} = 10673 + 90 \times 115 - 1169 \times 1 - 1299 \times 0 - 990 \times 0 = \$19835m$

Forecast Accuracy:

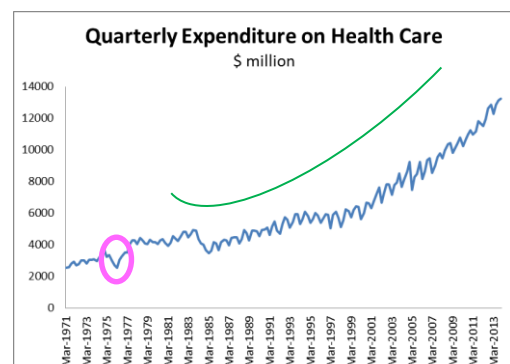
- Mean Method – just calculate the average between peaks and troughs and roughly estimates
- Naïve Method – doesn't take into account seasonality, it predicts the future based on a recent past ie. from Sunday to Monday
- Seasonal Naïve Method – Takes into account seasonality and predicts the future based on previous results for that period, ie. Monday to Monday



Extensions:

With this series we may need to deal with:

- non-linear trends
- outliers
- “structural changes”



Tricky Questions throughout Revision

Expected Value – “average”

Denote:

r the effective interest rate per annum
 p^t the (known) probability of being alive t years into the scheme

The probability distribution for payments in a particular year, Year t , in present value terms, is:

X	$\text{Pr}(X)$
\$0	$1-p^t$
$\$20,000 / (1+r)^t$	p^t

What situation do the two x values represent?

\$0 – represents adults who are not retired yet as they won't be receiving the payments until they invest their superannuation into the company or it could be they passed away

$\$20,000 / (1+r)^t$ – represents adults who are retired and have invested a lump sum with the finance company, and who will receive payments of \$20000 in the present value at t years with the effective interest rate per annum.

Calculate the Expected Value of X in Year t

$$\begin{aligned} E(X) &= (0 \times 1 - p^t) + (20000/(1+r)^t \times p^t) \\ &= 20000 p^t / (1+r)^t \end{aligned}$$

For most values of t and p , the Expected Value of X in Year t is not actually a value that X can take. Explain carefully what is meant by the Expected Value of X

The expected value is a weighted average of \$0 vs \$20000. The average could be \$16000

A simple random sample size of n that is $X_1, X_2, X_3, \dots, X_n$ is taken and the sample mean $\bar{X}$, is calculated. Let μ denote the population mean of X .

Show that $E(\bar{X}) = \mu$

$$E(X) = \mu$$

What does this mean in practice?

$$E(\bar{X}) = E\left(\frac{1}{n} \sum x_i\right)$$

$$= \frac{1}{n} E\left(\sum x_i\right)$$

$$= \frac{1}{n} \sum_{i=1}^n E(x_i)$$

$$= \frac{1}{n} \sum_{i=1}^n \mu$$

$$= \frac{1}{n} \times n\mu$$

$$= \mu$$

If you calculate the mean of many sample sizes of n and then you calculated the expected value of these means (average of the means) you should obtain the true population mean

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	158.73135	24.55982767	6.463048	8.34E-10	110.2880282	207.17468
Time	0.2395433	0.108879846	2.20007	0.029001	0.024781977	0.4543047
February	-54.797191	31.35758453	-1.74749	0.082159	-116.648834	7.0544529
March	15.492093	31.35815161	0.494037	0.621848	-46.3606693	77.344855

Time variable – On average rainfall increases by 0.2395mm per month, after adjusting for seasonality

The p value for time is 0.029001 which is below a significance level of 0.05. In this case as the p value < significance level we would reject the null hypothesis and conclude that the average rainfall increase has an effect on child health in Aceh. The higher the rainfall each month the more crops will flourish improving agriculture and thus improving children's health.

October	1.9070521	31.372703	0.060787	0.951592	-59.9744121	63.788516
November	54.015156	31.37629256	1.721528	0.086774	-7.87338802	115.9037
December	21.671495	31.3802595	0.690609	0.49065	-40.2248739	83.567864

Estimated Coefficient of November: In the month of November each year, the average rainfall is 54.015mm higher than the rainfall in January, after adjusting for the trend

The p value for November is relatively large. Explain carefully what this means about rainfall in November.

The p value of 0.086774 is greater than a significance level of 0.05. Therefore, we would not reject the null hypothesis that the coefficient equals zero and conclude that there is insufficient evidence that there is a difference between November rainfall and January Rainfall

Why is the median used instead of the mean in some cases for the measure of central tendency?

Extreme values (outliers) on either side of the distribution can pull the mean up or down (ie. if there is skewness). Therefore, the mean can be distorted by these values and thus the median will be a better measure of central tendency. Data like house prices and income will carry high and low extreme values based on the locations the houses are in and the types of jobs people have, thus the median will be a good indication of central tendency in these cases

Population Parameter: is a measure calculated from all the possible observations in a population of interest

Sample Statistic: is a measure calculated from a subset of observations in that population and used to estimate an unknown population parameter

Non-technical explanation of confidence interval: We are 95% confident that the true mean lies between \$12.40 and \$17.80

Technical Explanation: If we were to take all possible samples in the population of the same sample size (i.e. repeated sampling) and calculated 95% confidence intervals for each sample, 95% of those intervals would contain the true/population mean

TP for quadratic variables: $x = -b / 2a$

Age = 160

Age squared = 1.56

$x = -160 / 2 \times -1.56 \approx 53$ Therefore the income is at its maximum at around 53 years of age

Tricky Question

2	3302.0.55.001 Life Tables, Australia, 2004–2006								
3	Table 1 Life Tables, Australia								
4		Males					Females		
5	Age	lx	qx	Lx	ex	lx	qx	Lx	ex
6	0	100000	0.00544	99520	70.7	100000	0.00451	99603	63.5
7	1	99456	0.00043	99433	78.1	99549	0.00034	99531	82.9
8	2	99414	0.00024	99401	77.2	99515	0.00022	99504	81.9
9	3	99390	0.00019	99380	76.2	99494	0.00017	99485	80.9
10	4	99371	0.00016	99363	75.2	99477	0.00014	99470	79.9
11	5	99355	0.00014	99348	74.2	99463	0.00011	99458	78.9
12	6	99341	0.00012	99335	73.2	99452	0.00009	99448	77.9
13	7	99329	0.00011	99323	72.2	99443	0.00008	99439	76.9
14	8	99318	0.00010	99312	71.2	99435	0.00007	99431	76.0
15	9	99307	0.00010	99302	70.3	99428	0.00007	99424	75.0
16	10	99297	0.00010	99292	69.3	99421	0.00007	99417	74.0
17	11	99287	0.00011	99282	68.3	99414	0.00007	99411	73.0
18	12	99276	0.00012	99270	67.3	99407	0.00008	99403	72.0
19	13	99264	0.00014	99258	66.3	99399	0.00010	99394	71.0
20	14	99251	0.00016	99243	65.3	99389	0.00013	99383	70.0
21	15	99234	0.00022	99224	64.3	99376	0.00018	99368	69.0
22	16	99213	0.00032	99198	63.3	99358	0.00022	99348	68.0
23	17	99181	0.00050	99158	62.3	99336	0.00026	99324	67.0
24	18	99131	0.00065	99100	61.4	99310	0.00029	99296	66.0
25	19	99066	0.00075	99030	60.4	99282	0.00030	99267	65.1
26	20	98992	0.00080	98953	59.5	99252	0.00030	99237	64.1
27	21	98913	0.00081	98873	58.5	99222	0.00030	99207	63.1
28	22	98832	0.00081	98792	57.5	99192	0.00030	99177	62.1
29	23	98752	0.00081	98712	56.6	99162	0.00030	99148	61.1
30	24	98672	0.00082	98632	55.6	99133	0.00030	99118	60.2
31	25	98591	0.00084	98550	54.7	99103	0.00031	99088	59.2

Recall that q_x can be defined as $\Pr(\text{Dying in next year} \mid \text{Age}=x)$

- What is the probability of a male dying in the first year of their life? (1 mark)
 $\Pr(\text{Dying in first year of life} \mid \text{age} = 0) = 0.0054$
- What is the probability of a male dying in the second year of their life? (1 mark)
 $\Pr(\text{Dying in second year of life} \mid \text{age} = 1) = 0.0043$
- Calculate the probability of a male dying in the first two years of their life? Hint: It is not correct to just add the answers to parts i. and ii. (3 marks)

$$\Pr(\text{die in year 1} \mid \text{age} = 0) + \Pr(\text{die in year 2} \cap \text{didn't die in year 1})$$

$$\Pr(\text{die in year 1} \mid \text{age} = 0) + \Pr(\text{didn't die in year 1}) \times \Pr(\text{die in year 2} \mid \text{didn't die in year 1})$$

$$= 0.00544 + (1 - 0.00544) \times 0.00043$$

Recall that e_x is the expected value of the number of years a person will live, for a person whose current age is x

- Based on these values, what is the life expectancy of a 25 year-old male, and why is it bigger than that of a male aged zero? (3 marks)

$$\text{Life expectancy for a 25 year old male} = 25 + 54.7 = 94.7$$

$$\text{Life expectancy for a 0 year old} = 0 + 78.7 = 78.7$$

The life expectancy for a 25 year old is bigger than a male aged zero because they have a stronger immune system built over the early stages of life

Suppose you were offering a simple life insurance policy, which has a one-off premium P , and if a person dies in the next year, there is a fixed payout at the end of the year. This payout has a present value of \$10,000. The premium is allowed to vary with gender and age

- v. Use the Principle of Equivalence to calculate the premium P for a policy offered to a 20 year-old female. Show your working. (5 marks)

$$\Pr(\text{died in year 1} \mid \text{age} = 20) = 0.00030$$

$$\Pr(\text{didn't die in year 1} \mid \text{age} = 20) = 1 - 0.0003 = 0.9997$$

Principle of Equivalence: Expected payout = Expected Income

$$E(\text{payout}) = 0.00030 \times \$10000 + 0.9997 \times \$0$$

$$P = E(\text{payout}) = \$3$$

- v. Given the information in the Life Table, which people (i.e. what gender and age range) would have the lowest premium? Who would have the highest premium?

Females aged between 8 and 11 years old will have the lowest premium

Males aged zero years old will have the highest premium

Diversification – Risk and Return – Correlation

Stock example – Week 5

Stock A and C that were performing in the opposite weather conditions would have a correlation of -1

Stock A and B that were performing the same in each weather condition would have a correlation of 1

- If \$1 is invested in A and C you will get the same returns regardless (no risk)

Portfolio P which has half the stocks in A and half in B will have an average return of:

$$\mu_p = 0.5\mu_a + 0.5\mu_b$$

Quality Cycle

1. Plan, with use of Baseline Data eg. ask for student evaluations of units, and detect problem with specific units
2. Act – this might involve some treatment where particular subjects are chosen for a different type of input eg. we send some of our lecturers to a training workshop
3. Evaluate – collecting some data and comparing results – tests of means and differences of means (those who did receive the treatment and those who didn't) eg. compare improvements for the staff who did the training with those who didn't do the training
4. Improve – what factors are within our control actually contributed to the outcomes – regression analysis eg. student performance as it relates to evaluation scores