MKC3500 Week 1

Fundamentals Of Data Analysis

Research Designs

Exploratory research

- Objective: to explore and gain insights into the general nature of a problem
- Often the front end of a full research design (the first stage of the research process). Conclusive research may be undertaken subsequently. This is used when there is little or no prior knowledge. It is flexible, versatile, and unstructured (or semi-structured).
- Typical methods:
  - Expert surveys
  - Pilot surveys
  - Interviews
  - Focus groups
- Typically, the qs are open-ended. Interviews and focus groups allow the researcher to alter or develop their qs based on ppants responses.
- Benefits:
  - Discovering new ideas
  - Formulate a problem or define it more precisely
  - Develop hypotheses, identify key variables and relationships
  - Gain insights for developing a solution or establishing priorities for future research
- However, this is very time consuming, and is likely to have a small sample size (if there is a large sample, it will be very expensive).
- The data received is qualitative, so it will be difficult and time consuming to convert this data into quantitative data.

Conclusive research

Descriptive research

- Objective: to describe some aspect of the market enviro e.g. existence of a variable or relationship
  - Answers to the qs of who, what, where, when, and how
  - Needs problem to be clearly stated, hypotheses developed, variables identified
  - To gather info e.g. demographics, psychographics, and attitudes
  - To understand the prevalence of certain behaviour and the degree to which certain variables are associated
  - To make predictions and forecast behaviours
- Typical methods: surveys, secondary data
- Cross-sectional vs longitudinal designs:
  - Cross-sectional design: data is collected across individuals/groups. Can be single or multiple designs. The response rate is higher than longitudinal designs, however there may be confounds e.g. individual differences.
  - Longitudinal design: data involves repeated collection (on same measures) from the same pop at different points in time. This controls for individual differences, increasing reliability. However, it is very expensive, the response rate is low (attrition rate is high), and this may not be appropriate for certain studies.
- Primary vs secondary data:
  - Primary data is collected specifically to address a RQ and tailored to research needs. This is often not feasible.
  - Secondary data is collected by someone other than the researcher and not specifically for the research project. Secondary data can be longitudinal, cross-sectional, or panel.

Causal/experimental research

- Objective: to establish cause and effect in a relationship

- Advantages
  - To understand the nature of the relationship between the causal and outcome variables
  - To test whether your theory/justification for a relationship is correct
- Correlation ≠ causation; third variable problem
- Typical method: experiments allow researchers to control for non-relevant factors, manipulate the focal variables and rigorously test their effect on the outcome variable to establish a cause-effect relationship and eliminate the third variable problem
- A key concern is external validity: the experiment may not reflect what occurs in a natural setting rather than a lab setting

Survey Measurement Scales

Operation	Nominal	Ordinal	Interval	Ratio
Mode	Yes	Yes	Yes	Yes
Median	No	Yes	Yes	Yes
Mean	No	No	Yes	Yes
Ratios (÷, x)	No	No	No	Yes

Nominal scale

- Nominal scale – numbers are assigned to attributes of objects solely for ID
- Objects are assigned to mutually exclusive, labelled categories, but there are no necessary relationships among the categories. No ordering or spacing is implied.
- Useful measures: frequencies or count of each category
- Nominal data is discrete and nominal variables are categorical.
- To include these variables in a regression, we have to convert them to dummy variables.

In my family, I am the ... (pick the option that best applies)

o only child (1)

o eldest child among my siblings (2)

oyoungest child among my siblings (3)

o neither the eldest nor the youngest among my siblings (4)

Ordinal scale

- Ordinal scale – objects are ranked
- This provides order only, but not info about how much difference there is between objects.
- Useful measures: frequencies, median, mode

Please select the option that best describes your employment situation

o I am not currently employed (1)

o I am working part time for no more than 10 hours per week (2)

o I am working part time for over 10 hours but no more than 20 hours per week (3)

o I am working for more than 20 hours per week (4)

Interval scale

- Interval scale – units have the same width throughout the scale
- Numbers used to rank objects represent equal increments of the attribute being measured. Differences between levels can be compared.
- A large range of stat operations can be performed, however the scale has no meaningful 0 or starting point. Only the differences matter, not the absolute values.

	Strongly disagree (1)	Disagree (2)	Somewhat disagree (3)	Neither agree nor disagree (4)	Somewhat agree (5)	Agree (6)	Strongly agree (7)
For most of my previous Monash units I get Ds and HDs (1)	o	o	o	o	o	o	o

Likert scale

- Likert scale – uses statements that a respondent agrees or disagrees with
- This is typically used for psychometric measurements of attitudes and beliefs.

How much do you agree or disagree with each of the following statement about service at Myer department stores

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
1. Myer has friendly staff	1	2	3	4	5
2. Myer's staff are responsive	1	2	3	4	5
3. Myer's staff are knowledgeable 1	2	3	4	5	

Semantic differential scale

- Semantic differential – a rating scale with end points associated with bipolar labels

MYER's staff are

Unfriendly	___	X	___	___	___	Friendly
Ignorant	___	X	___	___	___	Knowledgeable
Unresponsive	___	___	X	___	___	Responsive
	1	2	3	4	5	

How important was performance on these attributes?

	Not Important	Somewhat Important	Important	Very Important
Overall quality	○	○	○	○
Value	○	○	○	○
Purchase experience	○	○	○	○

Ratio scale

- Ratio scale – a special type of interval scale that has a meaningful 0 point
- Ratio data is continuous where both differences and ratios are interpretable. Certain stats e.g. geometric mean, coefficient of variation can only be applied to ratio data.
- It is possible to say how many times greater or smaller one object is than another.

In the last 12 months, how many books did you read outside of school (i.e., exclude textbooks or books that you read because it is assigned for a class)? Write the number down in the space below

What mark do you expect to achieve for this unit (MKF2121) (e.g., 50, 65, 77, 92, etc)? Write down the number in the space below

Preliminary Data Analysis

Exploratory analyses

- Used to describe the data using % and measures of central tendency
- Always start with basic exploratory analyses
- To know about the distribution of variables, use:

- Simple plots
- Histograms
- Frequencies/bar charts
- Box plots
- Descriptive stats

- To know about the relationship between variables, use:
  - Scatter plots
  - Cross-tabs
  - Correlations

Why exploratory analyses?

- Identify if the data is credible/rep (check how much time it has taken to complete the survey, whether all qs have the same answer etc)
- Identify data errors and outliers
- Find basic relationships between variables
- Better understand the data e.g. how variables are distributed
- Provide a basis for evaluating more formal approaches e.g. what model should I use?
- Potentially identify unknown insights e.g. shape of the distribution

Frequencies

- Is the data rep? Are respondents from all categories adequately represented?

X1 - Customer Type

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Less than 1 year	32	32.0	32.0	32.0
	1 to 5 years	35	35.0	35.0	67.0
	Over 5 years	33	33.0	33.0	100.0
	Total	100	100.0	100.0	

Transforming a variable

- You can recode a range of original codes or one value into same/different variables or a value

Recommend

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	83	83.0	83.0	83.0
	1.00	17	17.0	17.0	100.0
	Total	100	100.0	100.0	

a dummy variable where 0 is <8 and 1 is 8 – 10.

If condition

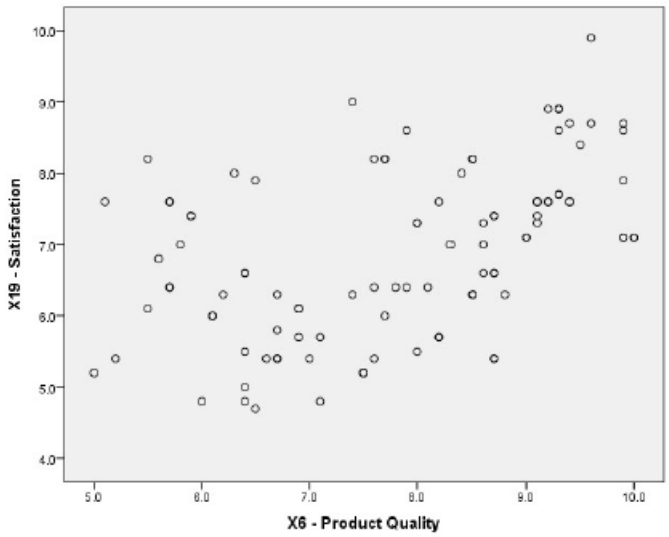
- Focus on a subset of the data

X2 - Industry Type

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Magazine industry	33	54.1	54.1	54.1
	Newsprint industry	28	45.9	45.9	100.0
	Total	61	100.0	100.0	

Scatter plots

- Exploring relationships between continuous variables



Cross-tabulations

- Explore relationships between nominal variables
- If the two variables are independent, the actual and expected cells should be very close.
- Most common tool used on survey data

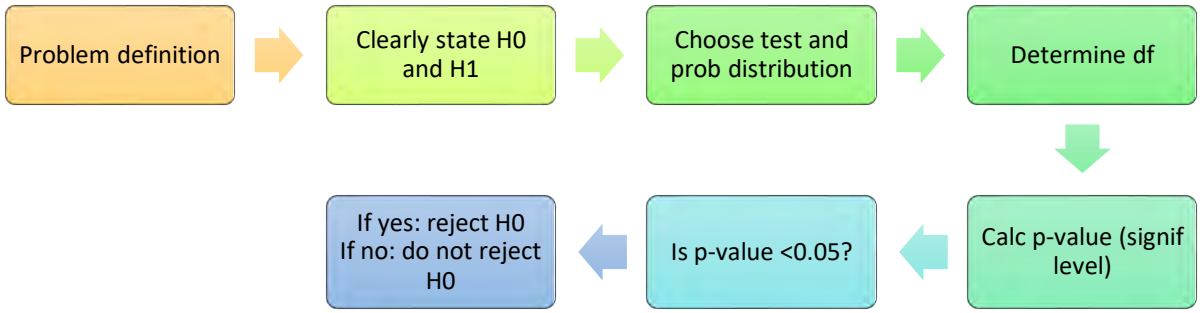
X4 - Region * X2 - Industry Type Crosstabulation

			X2 - Industry Type		
			Magazine industry	Newsprint industry	Total
X4 - Region	USA/North America	Count	19	20	39
		% within X2 - Industry Type	36.5%	41.7%	39.0%
	Outside North America	Count	33	28	61
		% within X2 - Industry Type	63.5%	58.3%	61.0%
Total		Count	52	48	100
		% within X2 - Industry Type	100.0%	100.0%	100.0%

there is some relationship between the variables, but we don't know whether this relationship is signif (use chi-square tests to determine signif)

Hypothesis Testing

- Hypothesis – an idea or explanation for something based on known facts but not yet proven
- Hypothesis testing is typically used to test whether a variable in the pop = some value, or whether there is a relationship between variables.
- When we work with samples, the mean of a variable in any given sample can differ from the true pop mean because of the inherent randomness of the sample. Hence we test how likely we would observe a particular sample if H0 were true.
- If observing that particular sample is rare (<5% chance), we reject H0, meaning that 95% of the time, the sample mean will not be the H0 value.
- H0: there is no relationship/effect between variables
  H1: there is a relationship/effect between variables
- Set the level of signif (5% or 1% if you want to be more stringent) and determine whether it is a one-tailed or two-tailed test



MKC3500 Week 2

Regression Analysis

What Is A Regression Model?

- Regression model – an equation that relates 1 outcome variable (DV) to 1+ causal variables (IVs)
- Explained by the model | The rest

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \cdots + \beta_k X_{ki} + \varepsilon_i$$

  - α or β0: intercept
  - β1: coefficient estimate of X1
  - ε: error term/residual

How is a regression model estimated?

- Using a least squares approach: βs are chosen to minimise the sum of squared residuals (the line is closest to all data points, since it has the smallest residuals). Residuals are squared to remove negative signs.
- H0: β = 0 and H1: β ≠ 0

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	3.676	.598		6.151	.000
	X6 - Product Quality	.415	.075	.486	5.510	.000

a. Dependent Variable: X19 - Satisfaction

β0 = 3.676 has a p-value <0.05, so we reject H0. We expect a satisfaction rating of 3.676 on a 0-10 scale when quality is low/0.

β1 = 0.415 has a p-value <0.05, so we reject H0. Satisfaction increases by 0.415 for every 1 point increase in quality.

Assessing the regression model

- The sum of squared devs provides the basis for assessing the fit of a regression model.
- Sum of squared devs of y data from its mean (SS_Y) = Sum of squares due to the regression (SS_{Reg}) – Residual sum of squares (SS_{Res})
- A good model should have signif higher SS_{Reg} than SS_{Res}.
- Commonly used model diagnostic tests include:
  - F ratio – a measure of how much variability is explained vs unexplained by the model. A good model will have a large F ratio (explain more variability). The F ratio has to be signif for the model to be valid (meaning at least some of the variability in the data is explained by the model).
  - r^2 – the proportion of variation in the DV explained by the variation in IVs. This is used to measure the model's fit. Adjusted r^2 discourages overfitting the model by including a penalty for the number of IVs in the model. When interpreting r^2, refer to all variables included in the model regardless of signif.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	33.260	1	33.260	30.358	.000[b]
	Residual	107.367	98	1.096		
	Total	140.628	99			

a. Dependent Variable: X19 - Satisfaction

b. Predictors: (Constant), X6 - Product Quality

F ratio p-value <0.05 so we can reject H0 that none of the variability in the data is explained by our model. The model fits the data at least to some extent.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.486[a]	.237	.229	1.0467

a. Predictors: (Constant), X6 - Product Quality

23.7% of the variation in the DV can be explained by the variation in the IV (product quality).

Assumptions of the linear regression model

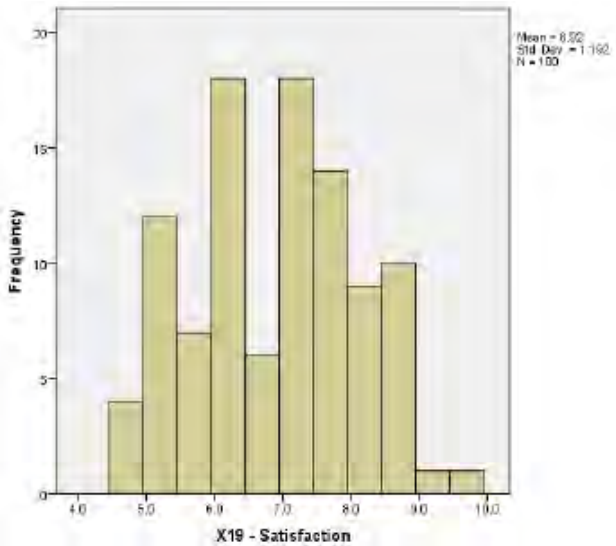
- The relationship between the IV and DV is linear
- Errors are normally distributed with a mean of 0 and constant variance σ^2
- Errors are independent of each other (there is no relationship between observations)
- Errors have homoscedasticity (error variance is the same regardless of the level of IVs)
- Errors and IVs are uncorrelated (there are separate methods when IVs are correlated)

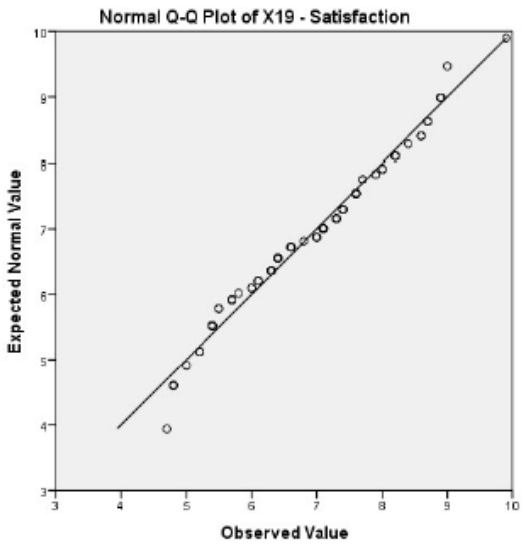
Test for linearity

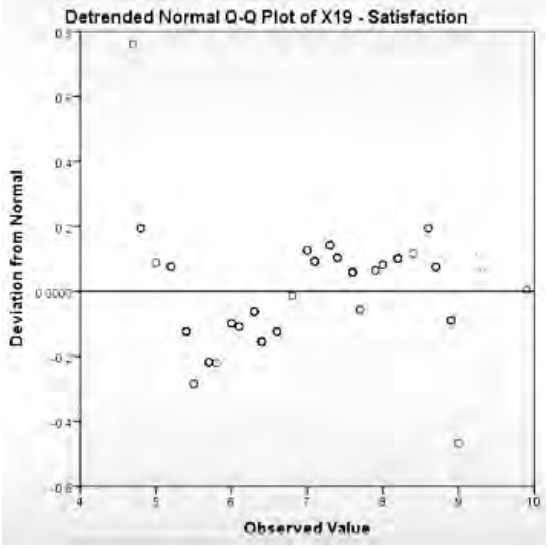
- Use a scatter plot of the IV and DV
- If the relationship is non-linear, the model will have a poor fit.

Test for normality

- Examine histogram and normal P-P or Q-Q plot of the DV. A straight line pattern on the normal plot supports normality.







the data is clustered around 0, so it is normally distributed

- It is not essential that errors are normally distributed. This is because of the law of large numbers (only if you have a large number of observations).
- However, if any outliers exist, their effects on the regression results must be assessed and appropriate action taken.
- T and F tests are robust, as long as the sample is large enough. If the sample size is small and errors are not normal, the t-test is invalid. However, the estimates are still valid if other assumptions hold.

Bootstrapping

- Bootstrapping – used to derive CIs and test hypotheses when the sample size is small
- The sample data is treated as though it is the pop. Repeated resampling gives a series of bootstrap estimates, from which the sampling distribution of the estimator is derived. From this sampling distribution, you can calc the mean and st dev.

Bootstrap for Coefficients

Model		B	Bootstrap[a]				
			Bias	Std. Error	Sig. (2-tailed)	BCa 95% Confidence Interval	
						Lower	Upper
1	(Constant)	3.676	-.059	.610	.001	2.413	4.648
	X6 - Product Quality	.415	.007	.075	.001	.272	.596

Test for heteroscedasticity

- Heteroscedasticity – when the error variance differs across levels of the IV
- To test for linearity and heteroscedasticity, plot the standardised residuals against the predicted Y values (in x axis).