

Real Statistics for Policy-Makers: Exercises in the Queensland Context



Image courtesy: www.fraudtechwire.com and The Peace and Collaborative Development Network

June 2019

This manual has been prepared by Dr Alisher Ergashev of Industry Analysis,
Department of Agriculture and Fisheries.

© State of Queensland, 2019.

The Queensland Government supports and encourages the dissemination and exchange of its information. The copyright in this publication is licenced under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence.



Under this licence you are free, without having to seek our permission, to use this publication in accordance with the licence terms.

You must keep intact the copyright notice and attribute the State of Queensland as the source of the publication.

Note: Some content in this publication may have different licence terms as indicated.

For more information on this licence, visit <https://creativecommons.org/licenses/by/4.0/>.

The information contained herein is subject to change without notice. The Queensland Government shall not be liable for technical or other errors or omissions contained herein. The reader/user accepts all risks and responsibility for losses, damages, costs and other consequences resulting directly or indirectly from using this information.

Table of Contents

Table of Contents.....	3
List of Tables, Acronyms and Abbreviations	4
Foreword	5
Acknowledgements	6
1. Key Statistics Terms and Definitions	7
Case Study.....	7
Theoretical Summary	8
Practical Exercises	9
2. Sampling	11
Case Studies	11
Theoretical Summary	12
Practical Exercises	13
3. Summary Statistics.....	14
Case Study.....	14
Theoretical Summary	15
Practical Exercises	16
4. Statistical Graphics.....	17
Case Studies	17
Theoretical Summary	18
Practical Exercises	19
5. Hypothesis Testing	20
Case Studies	20
Theoretical Summary	21
Practical Exercises	22
6. Linear Regression Analysis.....	23
Case Study.....	23
Theoretical Summary	24
Practical Exercises	26
References.....	28

List of Tables, Acronyms and Abbreviations

TABLES

Table 1: Key Statistics Terms and Definitions	8
Table 2: A Summary of the 2016-17 AAGIS	10
Table 3: Data Sampling: Terms and Definitions.....	12
Table 4: Summary Statistics: Measures of Location, Spread and Shape.....	15
Table 5: A Forecast of the Gross Value of Production for Queensland Primary Industries	16
Table 6: Main Features of Statistical Graphics	18
Table 7: Lengths (cm) of Sample Fish in the Gulf of Carpentaria Inshore Fin Fish Fishery	19
Table 8: Tomato Production in Australia, 1961-2016.....	19
Table 9: One-Sample Hypothesis Testing Procedure	21
Table 10: Simulation results.....	22
Table 11: Assumptions and Statistical Tests in Linear Regression Modelling	24
Table 12: Analysis of Variance (ANOVA) for Multiple Linear Regression	25
Table 13: Genomic Breeding Value of Cows Based on a Genetic Marker Expression Level ...	26
Table 14: Selected Macro-Economic Indicators of Queensland Economy, 1994-2017.....	27

ACRONYMS AND ABBREVIATIONS

AAGIS	Australian Agricultural and Grazing Industries Survey
ABARES	Australian Bureau of Agricultural and Resource Economics and Sciences
ABS	Australian Bureau of Statistics
AE	Adult equivalent
ANOVA	Analysis of variance
CaneLCA	Sugarcane life cycle assessment (tool)
CPI	Consumer Price Index
C-D	the Cobb-Douglas (production function)
DAF	Queensland Department of Agriculture and Fisheries
DAWR	Australian Department of Agriculture and Water Resources
GVP	Gross values of production
I-O	Input-output (model)
LW	Liveweight
MLR	Multiple linear regression
MSA	Meat Standards Australia
QATC	Queensland Agricultural Training Colleges
RIFA	Red imported fire ant
SA2	Statistical Area 2
SD	Standard deviation
SRS	Simple random sampling
TAFE	Technical and Further Education (institute)

Foreword

This statistics refresher for policy officers was borne out of frustration. Of all people, we cannot obey the dictum attributed to Bismarck that “Laws are like sausages. It is better not to see them being made”. Indeed, it is our responsibility to prepare policy decisions to the best of our ability, as difficult as it may be at times.

The use of robust information is vital for making evidence-based recommendations that guide policy preparation, monitoring, and evaluation ([Banks, 2009; ABS, 2010](#)). Without statistically reliable data, “the development process is blind: policy-makers cannot learn from their mistakes, and the public cannot hold them accountable” ([World Bank, 2000](#)).

A policy analyst is often constrained by data deficiencies that handicap evidence-based policy-making. This can lead to reliance on 'quick and dirty' surveys, or the overuse of focus groups, instead of applying statistical principles to make the most of available information.

Even apparently good and plentiful data are no guarantee of correct conclusions if such statistical concepts as statistical significance, sample representativeness, causality attribution and hypothesis testing are applied incorrectly.

Real Statistics for Policy-Makers: Exercises in the Queensland Context is therefore designed to help practitioners handle data in a statistically robust manner, even if not conducting formal statistical analysis. Basic concepts are illustrated through practical applications, case studies and exercises based on the actual information from the open-access data web libraries and focused on Australian and Queensland economies. Examples require the use of Microsoft Excel or similar statistical software package.

For each practical exercise, there is a detailed solution in Excel available upon further email request to the author: Alisher.Ergashev@daf.qld.gov.au (re: Statistics Exercises).

Acknowledgements

The author wishes to acknowledge the assistance and valued insights of the colleagues from Queensland Department of Agriculture and Fisheries. In particular, suggestions and comments from **Angela Anderson** (Agri-Science Queensland), **Tichaona Pfumayaramba** and **Caleb Connolly** (Rural Economic Development), **George Antony** and **Ken Smith** (Industry Analysis) are gratefully considered.

1. Key Statistics Terms and Definitions



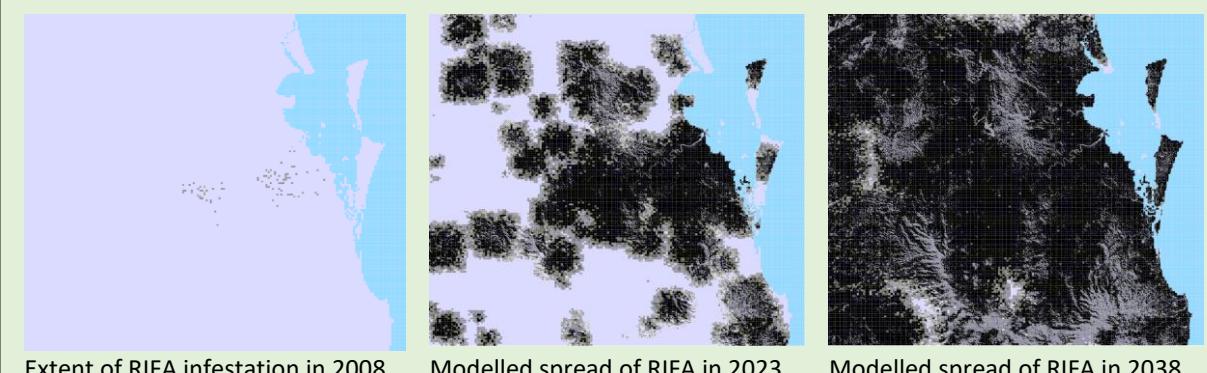
Statistics is a branch of mathematics dealing with data collection, organisation, analysis, interpretation and presentation. We encounter statistics in our daily lives more often than we probably realise and from many different sources, like the TV news, posts on Instagram, or Twitter.

Case Study

[Antony et al. \(2009\)](#) of the then-called Queensland Department of Primary Industries and Fisheries established quantitative indicators of the potential environmental impact by the red imported fire ant (RIFA) in Australia:

The area with some RIFA present is 8,405 ha in 2008. In 15 years, spread is expected to reach the boundaries of the modelled square, and by 2038, RIFA practically fills the 180 km square area, with some variation in density. Out of the total land mass of 2,624,828 ha RIFA are predicted to occur on 2,585,076 ha, of which infestation is 'dense' on 2,114,924 ha, 'common' on 79,313 ha, 'present' on 390,839 ha, leaving only 39,752 ha 'free'.

Figures below illustrate the spatial spread in 15-year intervals



What types of data are used in this example? Which variables can you identify?



Image courtesy: Pros an Cons

Theoretical Summary

Table 1: Key Statistics Terms and Definitions

Term	Definition																
Statistics	A collection of methods for collecting, displaying, analysing, and drawing conclusions from data																
Descriptive Statistics	The branch of statistics that involves organising, displaying and describing data																
Inferential Statistics	The branch of statistics that involves drawing conclusions about a population based on information from a sample taken from that population																
Population	Any specific collection of objects of interest																
Sample	Any sub-set or sub-collection of the population, including the case that the sample consists of the whole population, in which case it is termed a census																
<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: left; width: 15%;">Denotation</th><th style="text-align: left; width: 15%;">Definition</th><th style="text-align: left; width: 30%;">Calculation Method</th><th style="text-align: left; width: 40%;">Function in Excel</th></tr> </thead> <tbody> <tr> <td>Observations</td><td>n Number of observations, or the sample size</td><td>Count number of all values in a sequence</td><td>=COUNT("cell range")</td></tr> <tr> <td>Minimum</td><td>x_{min} The smallest value in the data set</td><td>The first number, when a sequence is sorted in ascending order</td><td>=MIN("cell range")</td></tr> <tr> <td>Maximum</td><td>x_{max} The largest value in the data set</td><td>The first number, when a sequence is sorted in descending order</td><td>=MAX("cell range")</td></tr> </tbody> </table>		Denotation	Definition	Calculation Method	Function in Excel	Observations	n Number of observations, or the sample size	Count number of all values in a sequence	=COUNT("cell range")	Minimum	x_{min} The smallest value in the data set	The first number, when a sequence is sorted in ascending order	=MIN("cell range")	Maximum	x_{max} The largest value in the data set	The first number, when a sequence is sorted in descending order	=MAX("cell range")
Denotation	Definition	Calculation Method	Function in Excel														
Observations	n Number of observations, or the sample size	Count number of all values in a sequence	=COUNT("cell range")														
Minimum	x_{min} The smallest value in the data set	The first number, when a sequence is sorted in ascending order	=MIN("cell range")														
Maximum	x_{max} The largest value in the data set	The first number, when a sequence is sorted in descending order	=MAX("cell range")														
Statistic	A number that represents a property of the sample; it is computed from the sample data. A statistic is a numerical summary of a sample																
Parameter	A numerical characteristic of the whole population that can be estimated by a statistic. A parameter is a numerical summary of a population																
Variable	A measurement (numerical or categorical) that can be determined for each member. You can think of the variable as kind of like a question																
Data	The actual values (numbers or words) of the variable. You can think of the data points as the answers to that Variable question																
Qualitative Data	A data concerned with descriptions, which can be observed but cannot be computed. The classification of objects is based on attributes/properties																
<i>Binary Data</i>	A type of data that place things in one of two mutually exclusive categories: right/wrong, true/false, or accept/reject																
<i>Nominal Data</i>	"Labelled" or "named" data which can be divided into various groups that do not overlap																
<i>Ordinal Data</i>	Ordinal data, unlike nominal data, involves some order; ordinal numbers stand in relation to each other in a ranked fashion																
Quantitative Data	The type of data, which can be measured and expressed numerically. It focuses on numbers and mathematical calculations																
<i>Discrete Data</i>	A set of data that can take only certain values. Discrete data result when the number of possible values is either a finite number, or a countable number																
<i>Continuous Data</i>	A set of data, which values falls in a continuous sequence (can take on any value within a finite or infinite interval)																

Practical Exercises

Exercise 1

Please determine the correct data type (quantitative or qualitative) and indicate what subtype the data belongs to.

Hint: Data that are discrete often start with the words "the number of"

- a. The number of commercial fishing licences in Queensland as of August 2018
- b. The type of crop protection products used by Queensland horticultural farmers
- c. The distance from Biosecurity Sciences Laboratory at Coopers Plains to the nearest prawn farm
- d. The number of agricultural vocational education and training courses provided by Technical and Further Education (TAFE) Queensland in 2016-17
- e. The type of tomato production – processing or fresh consumption
- f. Carcase weight of a MSA (Meat Standards Australia) trade steer sold across Queensland saleyards in 2016-17

Exercise 2

Determine what the key terms refer to in the following example:

Australian Bureau of Agricultural and Resource Economics and Sciences (ABARES) conducted the 2016-17 Australian Agricultural and Grazing Industries Survey (AAGIS) to analyse physical characteristics of the beef, grain and lamb-producing farmers in Australia and in each state, including Queensland.

Fill in the letter of the phrase that best describes each of the items below. Using the information from [Table 2](#), provide with actual values, where applicable.

- | | | |
|---------------|--------------------------|---|
| 1. Population | <input type="checkbox"/> | a) all agricultural farms formally registered in Australia in 2016-17 |
| 2. Statistic | <input type="checkbox"/> | b) the total area irrigated of a specialist beef farm in Queensland in 2016-17 |
| 3. Parameter | <input type="checkbox"/> | c) 10,830; 55; 0.86 |
| 4. Sample | <input type="checkbox"/> | d) a randomly selected group of lamb-producing farms formally registered in Australia |
| 5. Variable | <input type="checkbox"/> | e) the average age of owner manager of all registered mixed enterprise sheep farms in Australia |
| 6. Data | <input type="checkbox"/> | f) all registered beef farms in Australia |
| | | g) the average area operated by the randomly selected Queensland lamb-producing farms |

Table 2: A Summary of the 2016-17 AAGIS

	Population		Sample		Total area irrigated (ha)		Area operated at 30 June (ha)		Age of owner manager (years)	
	Queensland	Australia	Queensland	Australia	Queensland	Australia	Queensland	Australia	Queensland	Australia
Specialist beef farms	6,760	18,456	276	560	1 (36)	2 (53)	16,787 (11)	13,576 (18)	63 (2)	63 (2)
Mixed enterprise beef farms	643	6,698	45	264	7 (75)	9 (27)	8,766 (23)	7,752 (46)	59 (4)	61 (2)
Beef industries combined	7,403	25,154	321	824	1 (40)	4 (25)	16,091 (11)	12,025 (17)	63 (2)	63 (1)
Specialist sheep farms	185	13,631	13	389	0	2 (27)	19,231 (24)	3,300 (17)	57 (7)	62 (1)
Mixed enterprise sheep farms	420	10,817	25	448	1 (110)	14 (40)	9,215 (26)	3,555 (9)	55 (7)	57 (2)
Sheep industries combined	605	24,448	38	837	0 (110)	7 (35)	12,290 (18)	3,413 (10)	56 (5)	60 (1)
More than 20% receipts from prime lamb sales	77	7,765	4	275	ns	2 (29)	ns	1,988 (16)	ns	61 (1)
More than zero receipts from prime lamb sales	114	9,454	12	419	1 (115)	14 (43)	8,706 (18)	5,253 (14)	50 (6)	58 (1)
Lamb industries combined	191	17,219	16	694	1 (115)	9 (38)	10,830 (27)	3,780 (12)	55 (8)	59 (1)

Note: The figures in brackets () are relative standard errors. ns = not supplied because of insufficient sample size.

Source: Author's compilation based on [DAWR \(2018\)](#).



Image courtesy: The Habitat Advocate, Australian Wool Innovation, Humane Society International, Meat & Livestock Australia

2. Sampling



Sampling is the process of selecting a representative subset of data points. This data from the sample can then be analysed to identify patterns and trends in the larger data set being examined. Using the appropriate method of sampling allows us to avoid bias in our results.

Case Studies

Please refer to selected studies and think about the methods they used for collecting data:

- A study on the main factors that determine food prices in Australia conducted by [Spencer \(2016\)](#) for the Rural Industries Research and Development Corporation

Our assessment of the visibility and transparency of food commodity and category prices along chains is reflective of the limited availability of representative and consistent prices in many circumstances. This study has necessarily focused in areas where pricing data is available - mostly changes in farm-gate prices across most sectors over time, and on wholesale and retail prices in grocery channels. This means certain retail and foodservice channels are not covered by the analysis, as the scope of this study prevents more detailed work, which would be required in such cases.

- A study on the consumers' trust in the Queensland vegetable supply chain members and their behavioural responses conducted by [Ariyawardana et al. \(2017\)](#)

With a view to obtaining variability in the demographic profile of the respondents, 17 organisations were selected to represent a wide range of suburbs within south-west Brisbane, Queensland. The gate keepers of the contact organisations were requested to distribute questionnaires to their membership and they were then requested to distribute additional questionnaires to their friends and colleagues. Persons who are responsible for food/grocery shopping and who were over the age of 18 were invited to respond to the survey. By adopting a snowball sampling technique, a total of 1,370 questionnaires were administered in 2013 and 869 questionnaires were returned with a response rate of 63%. After data cleaning, 854 survey responses were analysed by using descriptive statistics and an ordered logistic regression model.

- A study by [Renouf et al. \(2018\)](#) on the effectiveness of the sugarcane life cycle assessment (CaneLCA) tool for evaluating the environmental implications in Australian sugarcane sector

The data required for the CaneLCA analyses were collected by agricultural extension officers during consultation sessions with farmers who had implemented the best management practices, with some follow-up clarification by phone and email. The analyses were used to construct a representative profile of cane growing practices in the Wet Tropics region before and after the practice changes. The representative case was developed by defining from the sample average practices and calculating farming inputs and yields based on production-weighted averages. The derived data were entered into CaneLCA to generate results for the representative case.



Image courtesy: Reading Craze

Theoretical Summary

Table 3: Data Sampling: Terms and Definitions

Term	Definition
Random Sampling	A method of selecting a sample that gives every member of the population an equal chance of being selected. Its usefulness is that it remains a ‘blind’ study, because no assumptions are made
Simple Random Sampling (SRS)	A straightforward method for selecting a random sample; give each member of the population a number. Use a random number generator to select a set of labels, which will identify the members of your sample
Systematic Sampling	A type of random sampling method in which sample members from a larger population are selected according to a random starting point and a fixed, periodic interval
Cluster Sampling	A method for selecting a random sample and dividing the population into groups (clusters); use simple random sampling to select a set of clusters. Every individual in the chosen clusters is included in the sample
Stratified Sampling	A method of random sampling where subgroups of the population are represented adequately. Divide the population into groups and use SRS to identify a proportionate number of individuals from each stratum
Quota Sampling	A type of stratified sampling in which selection within the strata is non-random; it ensures that there will be a representative sample of the population for specified criteria or strata
Convenience Sampling	A non-random method of selecting a sample; this method selects individuals that are easily accessible and may result in biased data. In this method, no inclusion criteria identified prior to the selection of subjects
Purposive Sampling	A non-random technique of the deliberate choice of a participant due to the qualities he/she possesses with regards to the research goals. It is effective if only limited number of subjects can be primary data sources
Sampling with Replacement	Once a member of the population is selected for inclusion in a sample, that member is returned to the population for the selection of the next individual
Sampling without Replacement	A member of the population may be chosen for inclusion in a sample only once. If chosen, the member is not returned to the population before the next selection
Sampling Error	The natural variation that results from selecting a sample to represent a larger population; this variation decreases as the sample size increases, so selecting larger samples reduces sampling error
Non-sampling Error	An issue that affects the reliability of sampling data; it includes a variety of human errors such as poor study design, biased sampling method, inaccurate information from participants, data entry errors, poor analysis
Representative Sample	A subset of the population that has the same characteristics as the population. A representative sample should be an unbiased reflection of what the population is like
Agricultural Census	The basic source of agricultural commodity statistics that is conducted every five years (with sample surveys in inter-censal years) covering all businesses with the Estimated Value of Agricultural Operations of greater than \$40,000

Practical Exercises

Exercise 3

Please answer the following questions, based on this hypothetical example:

An economist from Queensland Department of Agriculture and Fisheries (DAF) wishes to estimate the proportion of all commercial farmers in Tully Statistical Area 2 (SA2) region who had their property – fully or partially – affected by tropical cyclone Yasi in 2011. In a random sample of 100 banana farmers, 89 were somehow affected.

- a. What is the population of interest?
- b. What is the parameter of interest?
- c. What is the statistic involved?
- d. Based on this sample, do we know the proportion of all commercial farmers who were affected by Yasi?

Exercise 4

Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience), based on this hypothetical example:

A study is done by a DAF Economist to determine the average tuition that a student of the Queensland Agricultural Training Colleges (QATC) pays per semester. Each student in the following samples is asked how much tuition he or she paid for the current semester.

What is the type of sampling in each case?

- a. A sample of 100 QATC students is taken by organizing the students' names by qualification (Certificate II, Certificate III, Diploma, Bachelor), and then selecting 25 students from each.
- b. A random number generator is used to select a student from the alphabetical listing of all QATC students in the current semester. Starting with that student, every 50th student is chosen until 75 students are included in the sample.
- c. A completely random method is used to select 75 students. Each QATC student in the current semester has the same probability of being chosen at any stage of the sampling process.
- d. The Certificate II, Certificate III, Diploma, and Bachelor students are numbered one, two, three, and four, respectively. A random number generator is used to pick two of those qualifications. All students in those two qualifications are in the sample.
- e. A DAF Economist stands in front of the main building at Emerald Agricultural College one Wednesday to ask the first 100 students he/she encounters what they paid for tuition this semester. Those 100 students are the sample.

3. Summary Statistics



Summary statistics are used to summarise a set of observations, in order to communicate the largest amount of information as simply as possible. Summary statistics are particularly useful for comparing one project to another, or to a response before and after an event.

Case Study

In *Queensland Innovation Survey* conducted by [Verreyne \(2011\)](#) for the then-called Department of Employment, Economic Development and Innovation the measurement of innovation by Queensland firms was undertaken:

Queensland firms operating in the financial/insurance services and rental, hiring/real estate services industry reported the largest mean number of serious competitors compared to other industries ($M=262.43$, $SE=213.55$, $SD=2,070.41$). However, the large standard error and standard deviation indicates large variability in the number of competitors across firms. Queensland firms operating in the construction industry reported the largest median number of serious competitors ($Med=8.00$). Using the median is more accurate in this case as it is less affected by extreme scores.

Number of serious competitors reported by Queensland firms

	Mean	Standard Error of Mean	Standard Deviation	Median	Minimum	Maximum	Number of firms
Agriculture, Forestry and Fishing	6.29	1.306	6.396	3.00	1	23	24
Mining	6.36	1.533	5.085	5.00	2	15	11
Manufacturing	9.75	2.636	19.011	4.00	2	100	52
Electricity, Gas, Water and Water Services	7.50	3.331	12.464	4.00	1	50	14
Construction	49.58	26.181	190.602	8.00	1	1,000	53
Wholesale Trade, Retail trade, Accommodation and Food Services	10.81	2.209	35.408	4.00	1	500	257
Transport, Postal and Warehousing	10.00	5.047	21.997	4.00	1	100	19
Information Media and Telecommunications, Professional, Scientific and Technical Services	34.47	14.293	128.634	5.00	1	1,000	81
Financial and Insurance Services, Rental, Hiring and Real Estate Services	262.43	213.546	2,070.406	5.00	1	20,000	94
Other	21.36	9.495	98.678	4.00	1	1,000	108
Total	51.11	28.357	757.177	5.00	1	20,000	713

What can you say about the differences in presence of serious competitors between industries?

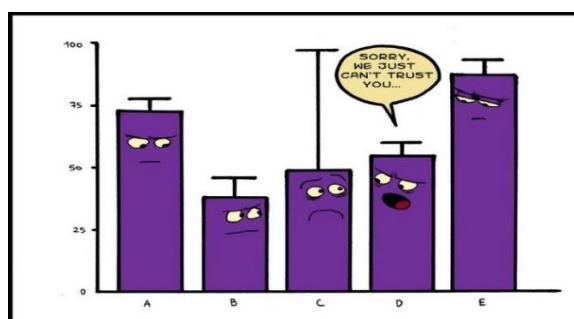


Image courtesy: Perdomics Facebook page

Theoretical Summary

Table 4: Summary Statistics: Measures of Location, Spread and Shape

Statistic	Denotation	Definition	Formula / Calculation method	Function in Excel
MEASURES OF LOCATION (CENTRAL TENDENCY) (summarise a list of numbers by a "typical" value)				
Mean	\bar{x} ("x-bar"), or μ	The sum of all the values in the data set divided by the number of values in the data set	$\bar{x} = \frac{\sum x}{n}$	=AVERAGE("cell range")
Mode	<i>Mode</i>	The value that occurs most often	The value which frequency is the largest in a sequence	=MODE("cell range")
Median (Middle Quartile)	\tilde{x} ("x-tilde"), or Q2	The middle score for a set of data that has been arranged in order of magnitude	The middle number, when a sequence has an odd number of values. The average of the two middle numbers, when the sample is even.	=MEDIAN("cell range")
Lower (First) Quartile	Q1	The median of the lower half of the data set	After arranging the values in ascending order: $Q1 = (n + 1)/4$	=QUARTILE.EXC("cell range",1)
Upper (Third) Quartile	Q3	The median of the upper half of the data set	After arranging the values in ascending order: $Q3 = 3(n + 1)/4$	=QUARTILE.EXC("cell range",3)
MEASURES OF SPREAD (VARIATION) (summarise how much members of a list of numbers differ from each other)				
Range	<i>Range</i>	The difference between the largest and smallest values	$Range = x_{max} - x_{min}$	=MAX("cell range")-MIN("cell range")
Inter-quartile Range	IQR	The difference between the 3rd quartile (75th percentile) and the 1st quartile (25th percentile)	$IQR = Q3 - Q1$	=QUARTILE.EXC("cell range",3)-QUARTILE.EXC("cell range",1)
Sample Variance	s^2	The average of the squares of the deviations	$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$	=VAR.S("cell range")
Standard Deviation	s , σ , or SD	A number that measures how far data values are from their mean	$s = \sqrt{s^2} = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$	=STDEV("cell range")
Z-Scores	Z	A measure of deviation for a single individual, as opposed to a group of scores	$Z_i = \frac{x_i - \bar{x}}{s}$	=("cell value"-AVERAGE("cell range"))/STDEV("cell range")
Standard Error of the Mean	σ_x ("sigma sub x-bar"), or SE	The standard deviation of the error in the sample mean with respect to the true mean	$\sigma_{\bar{x}} = \frac{s}{\sqrt{n}}$	=STDEV("cell range")/SQRT(count("cell range"))
MEASURES OF SHAPE (describe the distribution (or pattern) of the data within a dataset)				
Normal Distribution	$N(\bar{x}, s)$	A true symmetric distribution of observed values, with mode, median and mean being the same	3-sigma rule: ≈68.3% of values lie within one SD away from the mean; ≈95.4% lie within two SDs; and ≈99.7% are within three SDs	=NORM.DIST("cell value", mean, standard_dev, cumulative)
Skewness	G_1	The tendency for the values to be more frequent around the high or low ends of the x-axis	$G_1 = \frac{n}{(n - 1)(n - 2)} \frac{\sum(x - \bar{x})^3}{s^3}$	=SKEW("cell range")
Kurtosis	G_2	A measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution	$G_2 = \frac{n(n + 1)}{(n - 1)(n - 2)(n - 3)} \frac{\sum(x - \bar{x})^4}{s^4} - \frac{3(n - 1)^2}{(n - 2)(n - 3)}$	=KURT("cell range")

Practical Exercises

Exercise 5

Using the information in [Table 5](#), present the summary statistics of the 2017-18 gross value of production (GVP) forecasts for Queensland primary industries (\$ million).

Advanced: Please manually calculate the summary statistics and compare with those automatically generated by Excel, using the following path:

Data -> Data Analysis -> Descriptive Statistics -> Summary Statistics

Hint: If your Excel workbook does not have a “Data Analysis” tab, you can upload it this way:

1. Click the **File** tab, click **Options**, and then click the **Add-Ins** category.
2. In the **Manage** box, select **Excel Add-ins** and then click **Go**.
3. In the **Add-Ins** box, check the **Analysis ToolPak** check box, and then click **OK**.

Table 5: A Forecast of the Gross Value of Production for Queensland Primary Industries

Commodity	GVP
Cattle and calves	5,379
Poultry	640
Pigs	232
Other livestock	41
Sheep and lambs	11
Eggs	237
Milk (all purpose)	225
Wool	75
Bananas	580
Other fruit and nuts	272
Avocados	226
Strawberries	160
Macadamias	126
Mangoes	113
Mandarins	107
Apples	93
Pineapples	70
Table grapes	65
Tomatoes	298
Other vegetables	231
Capsicums & chillies	128
Beans	77
Mushrooms	70
Sweet potatoes	64
Melons (rock & cantaloupe)	59
Lettuce	56
Commodity	GVP
Potatoes	52
Sweet Corn	44
Zucchini & Button squash	41
Melons (watermelon)	37
Pumpkin	32
Carrots	27
Onions	26
Nurseries	907
Turf	327
Cut flowers	161
Sugar Cane	1,125
Cotton (raw)	884
Other crops	134
Chickpeas	406
Wheat	282
Grain sorghum	276
Other cereal grains	188
Maize	64
Barley	34
Crustaceans	107
Finfish	67
Mollusc	4
Recreational Fishing	94
Aquaculture	105
Forestry and logging	270

Note: As forecasted for 2017-18 by Industry Analysis (DAF) in April 2018. The gross value of production at farm gate excludes first-stage processing.

Source: [DAF \(2018\)](#).

Exercise 6

Consider these questions about measures of spread:

- a. What is the biggest problem with the range?
- b. What makes the interquartile range a better measure of spread?
- c. Why is the variance better than both range and interquartile range?
- d. What makes the standard deviation (SD) better than the variance?

4. Statistical Graphics

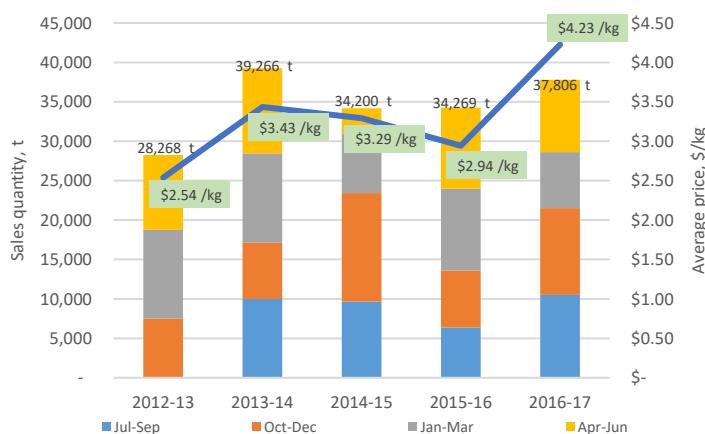


Statistical graphics, also known as graphical techniques, are graphics in the field of statistics used to visualise quantitative data. A well-structured and suitably labelled graph can summarise a variable (or set of variables) far more efficiently than any text format.

Case Studies

What types of the graphs can you see in the following examples?

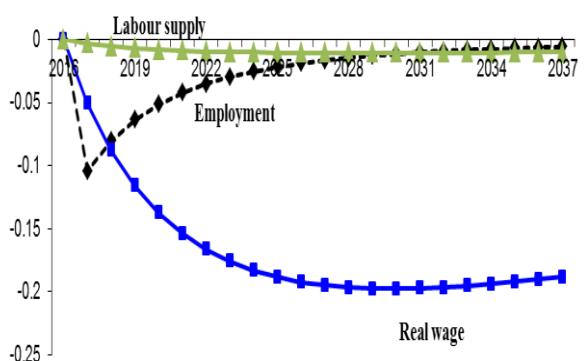
- [Ergashev \(2018\)](#) analysed a quarterly trade of tomatoes in Brisbane Markets



While annual quantities remain relatively consistent, seasonal volumes of tomatoes sold in the Brisbane wholesale markets do not follow a set pattern. For example, 7.2 thousand tonnes in Oct-Dec 2015 compared to 13.8 thousand tonnes in the same quarter of previous financial year, and 10.4 thousand tonnes in the following quarter of Jan-Mar 2016. The average tomato price for the last six years is \$3.3 per kilo, ranging from \$2.5 in 2012-13 to \$ 4.2 in 2016-17.

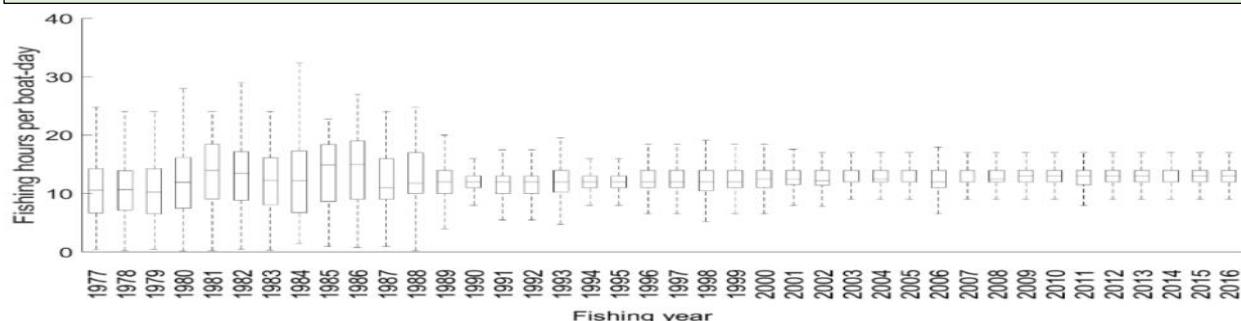
- [Wyttwer \(2016\)](#) modelled impact of a 25% reduction in Queensland sugarcane production

The chart shows the labour market impacts of the scenario. Closing 25% of sugarcane plantations and decommissioning the land reduces the available amount of capital and land. In the year of the closure, real wages adjust little. Therefore, adjustment in the labour market proceeds mainly through job losses. State-wide job losses are around 2,200 relative to base, plus there are additional job losses interstate. In succeeding years, the real wage falls, which brings labour demand back towards labour supply. Once labour demand matches labour supply, real wages stop falling.



- [Yang et al. \(2016\)](#) summarised available data in Queensland saucer scallop trawl sector

The chart shows the distribution (box and whisker plot) of fishing hours boat-day⁻¹ in 1977-2016. The records with missing hours were imputed using the Poisson model from the linear mixed model 1988-2016.



Theoretical Summary

Table 6: Main Features of Statistical Graphics

Type of Graph	Main Features	When to Use
Bar Charts	Simple to create. The height of the bar represents the frequency or relative frequency of that category. Can be horizontal as well as vertical	Used on the frequencies of categorical or ordinal variables
Histograms	A bit more complicated than bar charts, as you need to decide on a number of cut-off points (or bins), and count the number of observations within each bin	Used on the frequencies of quantitative variables only
Box Plots	The central line is the median, whilst the edges of the box are the quartiles, and the whiskers are the extremes. It is constructed from five values: the minimum value, the first quartile, the median, the third quartile, and the maximum value	A very comprehensive way of comparing multiple groups side-by-side
Pie Charts	They force us to compare areas (or angles), which is pretty hard. Avoid creating pie charts!	Used to tell a story about the parts-to-whole aspect of a set of data
Scatter Plot	Each observation is a point on the graph. The plot shows the direction of a relationship between the variables: the closer the data points come when plotted to making a straight line, the higher the linear correlation between the two variables	A more advanced way to represent two quantitative variables simultaneously

Hint: Good statistical graphics should:

- be self-explanatory
- relate directly to the argument
- be clearly labelled
- include a caption
- be finished with extra touches, where appropriate



Image courtesy: StockCharts

Practical Exercises

Exercise 7

Random samples, each size $n = 10$, were taken of the lengths in centimetres of three kinds of commercial fish in the Gulf of Carpentaria inshore fin fish fishery ([Table 7](#)):

Table 7: Lengths (cm) of Sample Fish in the Gulf of Carpentaria Inshore Fin Fish Fishery

Barramundi (<i>Lates calcarifer</i>)	108 105	100 107	99 105	125 119	87 118
Spot-tail shark (<i>Carcharhinus sorrah</i>)	133 145	140 160	152 138	142 139	137 138
King threadfin (<i>Polydactylus macrochir</i>)	82 74	60 79	83 82	82 80	82 80

From this hypothetical example, use the **Histogram** function in **Data Analysis Add-In** to:

- Find the frequency by grouping the measures in bins: 51-60, 61-70, ..., 151-60, 161-170
- Plot the frequency histogram with Bins on X-axis and Frequency on Y-axis
- Provide manual solution to questions (a) and (b), and compare your results:
- In each sample (n), calculate the frequency (f) and relative frequency ($rf_i = \frac{f_i}{n}$)
- Plot a histogram as a column chart, by using the bins and the **FREQUENCY** function

Exercise 8

Using the data from [Table 8](#), plot the scatter diagram with Production as the independent variable and Area harvested as the dependent variable. Comment on the any linear trend.

Table 8: Tomato Production in Australia, 1961-2016

Year	Production tonnes	Area ha	Year	Production tonnes	Area ha	Year	Production tonnes	Area ha
1961	142,591	7,003	1980	196,922	8,450	1999	394,371	8,549
1962	142,591	7,003	1981	216,836	9,057	2000	413,617	8,322
1963	131,115	6,680	1982	228,390	9,083	2001	556,240	9,582
1964	137,995	6,619	1983	224,077	8,714	2002	424,950	8,477
1965	149,556	6,602	1984	258,281	9,118	2003	364,368	7,309
1966	162,270	6,760	1985	270,475	9,292	2004	474,220	8,460
1967	175,741	7,200	1986	249,400	7,508	2005	407,867	7,806
1968	155,770	6,987	1987	266,019	8,547	2006	450,459	7,750
1969	156,794	7,074	1988	282,551	8,353	2007	296,035	7,293
1970	162,912	7,219	1989	318,618	9,140	2008	362,286	6,788
1971	178,464	7,444	1990	322,060	9,604	2009	440,093	6,789
1972	189,985	8,317	1991	364,108	10,071	2010	471,883	7,734
1973	172,353	7,656	1992	330,549	9,006	2011	301,719	8,244
1974	132,736	7,081	1993	279,762	8,554	2012	371,514	7,415
1975	165,441	7,868	1994	327,221	8,903	2013	455,654	6,290
1976	162,151	7,917	1995	340,033	8,657	2014	326,189	6,139
1977	178,071	8,595	1996	370,913	8,580	2015	389,205	5,847
1978	182,454	8,548	1997	393,117	8,830	2016	405,167	5,430
1979	172,639	8,172	1998	380,130	8,023			

Source: [FAOSTAT \(2017\)](#).

5. Hypothesis Testing



Hypothesis testing allows us to carry out inferences about population parameters using data from a sample. A hypothesis test evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data.

Case Studies

Please refer to selected fertiliser studies conducted in Central Queensland:

Cox et al. (2010):

Empirical and simulation results from three crop rotations incorporating cereals, pulses and nitrogen (N) fertiliser application were examined over four years in a subtropical environment, central Queensland, Australia. The **hypothesis** was that pulse crops in rotation with cereals would be a viable alternative to applying N fertilisers and would improve farm business economic performance provided the yield potential of pulses were not compromised by planting into very low soil water situations.

Sands and Lester (2016):

There is some soil testing evidence to suggest that nutrient stratification of non-mobile nutrients such as phosphorous (P) and potassium (K) is occurring across a range of Central Queensland soil types. Our **hypothesis** is that sulphur (S) could be starting to be a more limiting nutrient over time at this site — although the lack of an S response when applied alone suggests that it is still not the most limiting nutrient.

Lester et al. (2016):

The wetting and drying pattern of the soil in the northern region means that some of the subsoils have become largely depleted of nutrients, as the moist soil deeper in the profile is exploited by plant roots especially in winter crops where in-crop rainfall is lower. In response to these challenges, the **hypothesis** was developed that relatively high rates of nutrients could be placed in the subsoil (10-30 cm) to provide for several crop phases. The initial application would see some disturbance, with the duration of the responses uncertain.

Paungfoo-Lonhienne et al. (2017):

Recent studies under controlled conditions found that certain plant species such as peanut, sorghum and grasses release phytochemicals from roots that inhibit activities of soil nitrifying microorganisms. We **hypothesised** that compared to continuous mono-cropping or bare fallow, legume crop rotation may influence soil microbial community composition and the abundance of nitrifiers by altering soil N status and other bio-physico-chemical properties in the rhizosphere.

Can you think of any alternate hypotheses for each of the study under question?



Image courtesy: Everchem Fertilizer Company, Cultivate Colorado, East Grain

Theoretical Summary

Table 9: One-Sample Hypothesis Testing Procedure

STEP 1: STATE THE HYPOTHESES		STEP 4: STATE THE DECISION RULE		
Null Hypothesis (H_0) is the statement about the value of a population parameter	Alternative Hypothesis (H_A) is the statement that is accepted if evidence proves null hypothesis to be false	The decision rule is the instruction that states the conditions under which the null hypothesis will be accepted or rejected		
$H_0: \mu \leq \text{value}$	$H_A: \mu > \text{value}$ (right-tail test)	Reject H_0 , if the observed value of the test statistic is in the critical region	Fail to reject H_0 , if the observed value of the test statistic is outside of the critical region	
$H_0: \mu \geq \text{value}$	$H_A: \mu < \text{value}$ (left-tail test)			
$H_0: \mu = \text{value}$	$H_A: \mu \neq \text{value}$ (two-tail test)			
STEP 2: SELECT THE APPROPRIATE TEST STATISTIC AND ITS PROBABILITY DISTRIBUTION		STEP 5: COMPUTE THE OBSERVED VALUE OF THE APPROPRIATE TEST STATISTIC		
Test statistic is a random variable that is calculated from sample data to determine whether or not to reject the null hypothesis	Probability distribution is a function that describes the likelihood of obtaining the possible values that a random variable can assume			
1. Test a hypothesis of a proportion (p)				
z-statistic	normal distribution	$z = \frac{\hat{p} - p}{\sqrt{pq}}$		
2. Test a hypothesis of a mean (μ)				
a. If the population standard deviation (σ) is known and the sample size $n > 30$		$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$		
z-statistic	normal distribution			
b. If the population standard deviation (σ) is unknown and/or sample size $n < 30$		$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$		
t-statistic	t-distribution			
STEP 3: CHOOSE THE LEVEL OF SIGNIFICANCE		STEP 6: MAKE THE STATISTICAL DECISION		
Significance level (α) is the probability of rejecting the null hypothesis when it is true (Type I error). Suggested α include:		Compare the computed test statistic with critical value. If the computed value is within the rejection region(s), we reject the null hypothesis; otherwise, we do not reject the null hypothesis		
$\alpha=0.01$ (1%) for quality assurance work projects				
$\alpha=0.05$ (5%) for consumer research				
$\alpha=0.10$ (10%) for political polling				
STEP 7: MAKE THE ECONOMIC DECISION		Based on the decision in Step 6, we state a conclusion in the context of the original problem		

Practical Exercises

Exercise 9

A DAF Economist wants to analyse a horticultural labour market in Mareeba, Queensland. To do so, they need to take into account the real wage of mango pickers in the area. Although the official statistics says it is \$25 per hour, the DAF Economists suspects that this size may be too low in real terms. It is then decided to take a sample of 30 mango pickers in five different locations to see if the average per hour salary is significantly greater than \$25.

What are the appropriate hypotheses (H_0 and H_A) for their significance test? Please explain.

- $H_0: \rho = \$25 | H_A: \rho > \25 , where ρ is the proportion of the mango picker's salary
- $H_0: \rho = \$25 | H_A: \rho < \25 , where ρ is the proportion of the mango picker's salary
- $H_0: \mu = \$25 | H_A: \mu > \25 , where μ is the average size of the mango picker's salary
- $H_0: \mu = \$25 | H_A: \mu < \25 , where μ is the average size of the mango picker's salary

Exercise 10

In order to test the effectiveness of a new vaccine, the DAF researchers want to analyse the post-treatment effect in a random sample of chickens. The rule of thumb is that the vaccine should be regarded effective if the portion of sample chickens with disease is less than 10 per cent three months after the treatment.

The researchers randomly selected a sample of 400 chickens from available flock and performed necessary tests, after which they concluded that 14 per cent of the sample had disease after treatment. The [Table 10](#) sums up the results of 1,000 simulations, each simulating a sample of 400 chickens, assuming there are 10 per cent those with disease.

Using this hypothetical example, please answer the following questions:

- Formulate a hypothesis (H_0 and H_A) for testing the effectiveness of a new vaccine
- According to the simulations, what is the probability of getting a sample with 14 per cent or more of chickens with disease?
- What should we conclude regarding the hypothesis?

Table 10: Simulation results

Measured % of chickens with disease	Frequency
6	7
7	40
8	93
9	173
10	327
11	253
12	73
13	27
14	7



Image courtesy: Everchem

6. Linear Regression Analysis



Linear regression is used for finding linear relationship between target and one or more predictors. The major conceptual limitation of all regression techniques is that you can only ascertain relationships, but never be sure about underlying causal mechanism.

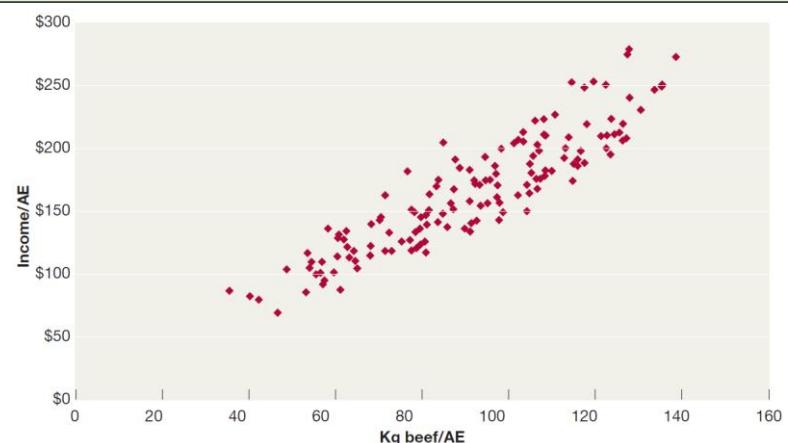
Case Study

[McLean and Holmes \(2015\)](#) analysed the performance of the northern beef industry:

Beef prices are an important component, but usually get more attention than they deserve. The general expectation is that higher beef prices mean higher income and higher profits. However when different herds, or groups of herds, are compared this is not the case. Figure below graphs the average income per kg of liveweight (LW) and corresponding income per adult equivalent (AE) of beef herds from the Northern Beef Report data and shows there is no obvious relationship. This seems counter intuitive and is a critical point in understanding what drives the profit of a beef business.



The income of a beef business is determined by how much beef it produces and therefore this is where attention should be focused. The productivity of the herd is measured in terms of kilograms of beef produced per AE per year (kg beef/AE). Figure below graphs the relationship between kg beef/AE and income/AE using exactly the same data as in Figure above. Here the relationship is clear, as herd productivity increases along the bottom axis, herd income increases along the vertical axis.



Can we say that the increases in beef producers' income are caused by higher beef prices?

Theoretical Summary

Table 11: Assumptions and Statistical Tests in Linear Regression Modelling

Assumptions	Meaning	Statistical Test
Linearity	It is assumed that the relationship between variables is linear. Look at bivariate scatterplot of the variables of interest. If curvature in the relationships is evident, you may consider either transforming the variables, or explicitly allowing for nonlinear components	No "general" test for linearity
(Absence of) Outliers	It is assumed that there are no outliers, which are observations that appear to deviate markedly from other observations in the sample. The box plot and histogram are useful in identifying them. Rather than exclude outliers (unless they are "bad data"), use a robust method of regression	Grubbs, Tietjen-Moore, Dixon, Generalized Extreme Studentized Deviate, Interquartile Range
(Absence of) Autocorrelation (or serial correlation, independence of residuals)	It is assumed that there is no autocorrelation in the data, which occurs when the residuals are not independent from each other. The independence of residuals is observed, if collected data represents a random sample from the relevant population	Durbin-Watson, Breusch-Godfrey
Normality	It is assumed that the residuals (predicted minus observed values) follow the normal distribution. You can produce histograms for the residuals as well as normal probability plots in order to inspect the distribution of the residual values	Shapiro-Wilk, Chi-square, D'Agostino-Pearson, Kolmogorov-Smirnov, Jarque-Berre
(Absence of) Multicollinearity	It is assumed that there is no multicollinearity, which occurs when the independent variables are too highly correlated with each other. It can be checked by inspecting a correlation matrix (if correlations are above 0.80, then there is a problem)	Farrar-Glauber, Variance Inflation Factor
Homoscedasticity (or equality of variances, homogeneity of variances)	It is assumed that residuals do not vary systematically with the predicted values. It can be checked by plotting the residuals against the values predicted by the regression model. There should be no clear cone-shaped pattern in the distribution	Bartlett, Levene, Fligner-Killeen, Goldfeld-Quandt
Stationarity (for time series only)	In time series techniques, it is assumed that the data are stationary meaning its statistical properties (mean, variance and autocorrelation) do not vary with time. To test whether a given time series is stationary or not, an indirect test for the existence of a unit root is applied	(Augmented) Dickey-Fuller, Kwiatkowski-Phillips-Schmidt-Shin, Priestley-Subba Rao, Wavelet Spectrum

Table 12: Analysis of Variance (ANOVA) for Multiple Linear Regression

Statistic	Denotation	Definition	Formula
Response (dependent, criterion) Variable	y	A variable in a functional relation whose value is determined by the values assumed by other variable(s) in the relation	$y = f(x_1, x_2, \dots, x_k)$ $y_i = \beta_0 + \beta x_i + \epsilon_i$
Predictor (independent, explanatory) Variable	x	A variable in a functional relation whose value determines the value(s) of other variables. There are k predictor variables	x_1, x_2, \dots, x_k $k \geq 1$
Observed Value	y_i	The value that is actually observed (what actually happened)	(from sample)
Error (disturbance)	ϵ_i , or e_i	Deviation of the observed value from unobservable <i>true</i> value of a quantity of interest (population mean or sample mean)	$\epsilon_i = y_i - \mu; e_i = y_i - \bar{y}$
Residual	r_i	Difference between observed value of dependent variable and predicted value. It tells if the prediction was too high or too low	$r_i = y_i - \hat{y}_i$
Predicted (fitted) Value	\hat{y}_i ("y hat")	Prediction of the mean response value when you input the values of the predictors, factor levels, components in the model	$\hat{y}_i = \alpha + \beta x_i$
Standard (standardised) Residuals	r_{st}	The residual divided by its standard deviation	$r_{st} = \frac{r_i}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n r_i^2}}$
ANOVA Summary, Part 1			
Multiple Correlation Coefficient	R	Measures the strength of a linear relationship (1 means a perfect positive relationship, 0 - no relationship at all)	$R = \sqrt{R^2}$
R Square	R^2	Coefficient of determination that gives the percentage variation in y explained by x -variables	$R^2 = 1 - \frac{RSS}{TSS}$
Adjusted R-Squared	R^2_{adj}	R Square that adjusts for the number of terms in a model. Useful in comparing the explanatory power of different models	$R^2_{adj} = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$
Standard Error of the Regression (or Estimate)	S , or SER	Precision that regression coefficient is measured (the smaller the "S" value, the closer the values are to regression line)	$SER = \sqrt{\frac{\sum r_i^2}{n - k - 1}}$
Observations	N , or n	Total number of observations, total sample size	(from sample)
ANOVA Summary, Part 2			
Regression Degrees of Freedom	DFR	The number of independent pieces of information that went into calculating the estimate	$DFR = k$
Residual Degrees of Freedom	RDF	The more variables you add, the more you erode your ability to test the model	$RDF = n - k - 1$
Total Degrees of Freedom	TDF	The total degrees of freedom equals $N - 1$	$TDF = n - 1$
Regression Sum of Squares	SSR	Sum of squares of the deviations of the predicted values from the mean value. Shows how well a model represents the data	$SSR = \sum (\hat{y}_i - \bar{y})^2$
Residual Sum of Squares	RSS	Sum of squares of residuals. Measures the discrepancy between the data and an estimation model	$RSS = \sum (y_i - \hat{y}_i)^2$
Total Sum of Squares	TSS	Sum of the "within-samples" sum of squares and "between-samples" sum of squares	$TSS = SSR + RSS$
Regression Mean Squared	MSR	Sum of squares divided by its associated degrees of freedom	$MSR = \frac{SSR}{DFR}$
Mean Squared Error	MSE	Measures the average squared difference between the estimated values and what is estimated	$MSE = \frac{RSS}{RDF}$
F-statistic	F	The statistic which measures if the means of different samples are significantly different or not	$F = \frac{MSR}{MSE}$
Significance F		Significance associated P-Value which is determined by comparing F-statistic to F critical value from F-Distribution Table	Excel function $=F.DIST.RT(F, DFR, RDF)$
ANOVA Summary, Part3			
Intercept (or Constant)	β_0 ("beta zero")	Determines where the line intersects the Y-axis. It indicates the mean of the distribution of Y when X s are equal to 0	$\beta = (X^T X)^{-1} X^T Y$, where X is a $(n \times k)$ design matrix, $Y = X \beta + \epsilon$ is a response matrix
(Partial Regression) Coefficients	β ("beta")	Slope of linear relationship between response variable and the part of a predictor var that is independent of other predictors	
Standard Error of Coefficients	$SE(\beta)$	An estimate of the standard deviation of the coefficient, the amount it varies across cases	$SE(\beta) = \sqrt{MSE(X^T X)^{-1}}$
t Statistic	t	The T Statistic for the null hypothesis vs. the alternate hypothesis. It is the coefficient divided by its standard error	$t = \frac{\beta}{SE(\beta)}$
P-value	P -Value	The probability that the regression coefficient is not statistically significant (i.e. not different from zero)	

Practical Exercises

Exercise 11

A genomic breeding value of cows based on a genetic marker expression level is summarised in [Table 13](#). Run a linear regression with Breeding Value as the response variable and Expression as the explanatory variable.

Using this hypothetical example, you should be able to:

- Create a scatter plot
- Find the correlation between the two variables

Hint: There are three ways to calculate the correlation in Excel:

1. Use the **CORREL** function,
2. Select **Data -> Data Analysis -> Correlation**, and
3. Calculate the correlation manually, using the correlation coefficient formula

- Find the residual standard error (i.e. how variable the residuals are)
- Investigate the normality of residuals via plotting the residuals vs fitted values and the normal probability plot
- Find the best value of the slope and the intercept
- Find out whether the intercept and slope are significantly different from 0 or not
- Using the regression equation, describe what this tells us about the relationship between Breeding Values and Expression

Advanced: Provide manual solution to questions (b), (c) and (e), and compare your results.

Table 13: Genomic Breeding Value of Cows Based on a Genetic Marker Expression Level

Expression	Breeding Value
1.58	5.43
0.09	3.54
0.82	5.14
0.58	3.6
1.91	8.07
1.88	6.91
1.15	5.08
1.1	4.13
1.06	4.65
1.8	7.44
0.91	2.85
0.66	6.11
0.91	5.52
0.08	4.38
0.21	3.82
1.91	5.68
1.36	4.69
0.49	4.09
1.77	5.91
1.78	4.87



Image courtesy: Irongate Equine Clinic

Exercise 12

Productivity gains can be measured by multi-factor productivity growth or the Solow residual, which indicates the increased efficiency of labour (L) and capital (K) inputs of production as they transform into output (Y).

The Cobb-Douglas (C-D) production function can be written as follows: $Y = A K^\alpha L^\beta$

If we transform the data by taking natural logarithms of Y , L and K , and then perform a multiple regression, the coefficient of $\log(K)$ will be our estimate of α and the coefficient of $\log(L)$ will be our estimate of β :

$$\log Y = \log A + \alpha \log K + \beta \log L$$

By analysing the Queensland growth accounting, using the data in [Table 14](#), you should be able to run a regression with \log of gross state product ($\ln Y$) as the dependent variable, and \log of Capital input ($\ln K$) and \log of Labour input ($\ln L$) as independent variables:

- Perform data diagnostics (linearity, outliers, normality, homoscedasticity, multicollinearity)

Advanced: Since the data are time series, test for stationarity. Transform the data, if required

- Check for the significance of the model and its goodness-of-fit. Interpret your results
- Using the resulted coefficients, write a standard C-D model and interpret your findings
- Check if the output exhibits constant, increasing or decreasing returns to scale

Advanced: Provide manual solution to questions (a) and (b), and compare your results.

Table 14: Selected Macro-Economic Indicators of Queensland Economy, 1994-2017

Fiscal year	Gross state product ¹ (\$ Millions)	Net capital stock ² (\$ Millions)	Hours worked in all jobs ³ ('000 Hours)
	Chain volume measures	Chain volume measures	Seasonally Adjusted
1994-95	135,092	501,925	2,664,889.33
1995-96	140,416	520,436	2,714,697.12
1996-97	147,739	541,648	2,730,430.29
1997-98	153,659	561,361	2,794,823.28
1998-99	163,224	582,826	2,838,422.70
1999-2000	170,633	605,260	2,926,841.59
2000-01	175,598	620,679	2,938,662.47
2001-02	187,555	640,099	2,972,262.79
2002-03	193,607	666,598	3,095,679.51
2003-04	205,764	695,674	3,193,888.12
2004-05	216,944	728,059	3,378,561.31
2005-06	228,285	766,034	3,460,406.23
2006-07	243,422	810,422	3,659,814.85
2007-08	254,761	860,848	3,757,479.42
2008-09	258,293	909,163	3,852,928.35
2009-10	262,036	947,907	3,838,911.52
2010-11	263,597	988,109	3,868,380.99
2011-12	278,078	1,044,884	3,961,185.53
2012-13	285,936	1,100,188	3,925,235.71
2013-14	292,155	1,151,576	4,014,448.12
2014-15	295,602	1,182,414	4,004,076.73
2015-16	303,352	1,199,920	4,041,744.16
2016-17	308,709	1,217,565	4,045,393.94

¹[ABS Cat. No. 5220.0 - Australian National Accounts: State Accounts, 2016-17. Table 1: Gross State Product, Chain volume measures and current prices.](#)

²[ABS Cat. No. 5220.0 - Australian National Accounts: State Accounts, 2016-17. Table 23: Queensland Capital Stock by Type of asset, Institutional sector and Industry.](#)

³[ABS Cat. No. 6202.0 - Labour Force, Australia, July 2018. Table 19: Monthly hours worked in all jobs by Employed full-time, part-time and Sex and by State and Territory.](#)

References

- Australian Bureau of Statistics [ABS]. (2010). A guide for using statistics for evidence based policy. Cat. No. 1500.0. Canberra: Australian Bureau of Statistics.
- Australian Bureau of Statistics [ABS]. (2017). Australian National Accounts: State Accounts, 2016-17. Cat. No. 5220.0 Canberra: Australian Bureau of Statistics.
- Australian Bureau of Statistics [ABS]. (2018). Labour Force, Australia, July 2018. Cat. No. 6202.0. Canberra: Australian Bureau of Statistics.
- Antony, G., Scanlan, J., Francis, A., Kloessing, K., & Nguyen, Y. (2009). Revised benefits and costs of eradicating the red imported fire ant. In: 53rd Annual Conference of the Australian Agricultural and Resource Economics Society, 10–13 February, Cairns. *Queensland Department of Primary Industries and Fisheries, Brisbane*.
- Ariyawardana, A., Ganegodage, K., & Mortlock, M. Y. (2017). Consumers' trust in vegetable supply chain members and their behavioural responses: A study based in Queensland, Australia. *Food Control*, 73, 193-201.
- Banks, G. (2009). *Evidence-based policy making: What is it? How do we get it?* ANU Public Lecture Series, presented by ANZSOG, Canberra, February 2009. Productivity Commission: Canberra.
- Cox, H. W., Kelly, R. M., & Strong, W. M. (2010). Pulse crops in rotation with cereals can be a profitable alternative to nitrogen fertiliser in central Queensland. *Crop and Pasture Science*, 61(9), 752-762.
- Queensland Department of Agriculture and Fisheries [DAF]. (2018). *Queensland AgTrends 2017-18: Update April 2018*. Brisbane: Queensland Department of Agriculture and Fisheries.
- Australian Department of Agriculture and Water Resources [DAWR]. (2018). Farm survey data for the beef, slaughter lambs and sheep industries. Canberra: Australian Department of Agriculture and Water Resources.
- Ergashev, A. (2018). *Tomato Supply Chain in Queensland*. Unpublished technical report. Brisbane: Queensland Department of Agriculture and Fisheries.
- Statistics Division of the Food and Agriculture Organization of the United Nations [FAOSTAT]. (2017, December 15). Crops. Retrieved January 16, 2018, from Data: <http://www.fao.org/faostat/en/#data/QC>.
- Lester, D. W., Bell, M., Graham, R., Sands, D., & Brooke, G. (2016). P & K, day degrees & dual purpose crops concurrent session (Day 2) Phosphorus and potassium nutrition. *GRDC Grains Research Update*, 157.
- McLean, I. & Holmes, P. (2015). Improving the performance of northern beef enterprises (2nd Edition). Meat & Livestock Australia Limited.
- Paungfoo-Lonhienne, C., Wang, W., Yeoh, Y. K., & Halpin, N. (2017). Legume crop rotation suppressed nitrifying microbial community in a sugarcane cropping soil. *Scientific reports*, 7(1), 16707.
- Renouf, M. A., Poggio, M., Collier, A., Price, N., Schroeder, B. L., & Allsopp, P. G. (2018). Customised life cycle assessment tool for sugarcane (CaneLCA)—a development in the evaluation of alternative agricultural practices. *The International Journal of Life Cycle Assessment*, 1-15.
- Sands, D. & Lester, D. (2016). *Chickpea: production five years after deep PKS application—Capella*. Queensland Grain Research – 2016, Regional Agronomy. Queensland Department of Agriculture and Fisheries.
- Spencer, S. (2016). *From farm to retail – how food prices are determined in Australia*. Canberra: Australian Government Rural Industries Research and Development Corporation.
- Verreynne, M. L. (2011). Queensland innovation survey 2011. Brisbane: University of Queensland Business School.
- Wittwer, G. (2016). *Closure of 25% of Queensland's Sugarcane Plantations*. Unpublished technical report. Centre of Policy Studies, Victoria University.
- World Bank. (2000). *World Development Indicators 2000 (English)*. World Bank Publication 20372. Washington, D.C. The World Bank.
- Yang, W. H., Wortmann, J., Robins, J. B., Courtney, A. J., O'Neill, M. F., & Campbell, M. J. (2016). *Quantitative assessment of the Queensland saucer scallop (Amusium balloti) fishery*. Technical Report. Brisbane: Queensland Department of Agriculture and Fisheries.