Chapter 1

Bootstrapping with replacement is treating the sample as if it was the population. A bootstrap confidence interval for a parameter can be thought of a range of believable values for the true parameter value.

Interpretation: *It is a fairly safe bet that the true value of the parameter is one of the values in this confidence interval.*

Non-sampling errors:

Impossible to correct and impossible to tell how badly it will affect the result.

Selection bias – Population sampled not the population of interest
Non-response bias – People chosen to do survey do not respond
Self-selection bias – People volunteer themselves to do the survey, anyone can take part
Question effects – Variations in wordings can affect responses
Interviewer effects – Different interviewers asking the same questions can get different results
Behavioural considerations – People answer questions to what is socially desirable
Transferring findings – Taking data from 1 population and transferring results to another
Survey format effects – Question order, survey layout, interview in person/phone/email

Observational studies:

Useful for identifying possible causes of effects, but they cannot reliable establish causation.

1) Cross- sectional – A study that observes a group at a point in time. It provides a "snapshot"
2) Longitudinal – A study that observes a group over a long period of time. Comprised of a series of cross-sectional studies

Experiments:

Only well-designed and well-executed experiments can be used to reliably establish causation

1) Completely randomised design – Treatments are allocated entirely by chance. It is an attempt to keep the treatment groups as similar as possible.
2) Randomised block design - Group units by some known factor and then randomise within each block in an attempt to balance out unknown factors.

Guidelines for assessing 'Chance alone'

When the **tail proportion** of the re-randomised distribution is **less than 10%** then:

- The observed difference would be unlikely when chance is acting alone; therefore it is a fairly safe bet chance isn't acting alone.
- We have evidence **against** 'chance is acting alone'
- We have evidence that **chance is not acting alone**

When the **tail proportion** of the re-randomised distribution is **bigger than 10%** then:

- The observed difference is not unusual when chance is acting alone, therefore chance **COULD** be acting alone.
- We have **NO** evidence **against** 'chance is acting alone'

- 'Chance' **COULD** be acting alone **OR** something else as well as 'chance' **COULD** also be acting.

Chapter 2

Types of variables:

Quantitative

- Continuous (measurements, no gaps between values)
- Discrete (usually counts, there are gaps between each value)

Qualitative

- Categorical (no order, eg male and female)
- Ordinal (natural order eg income group)

Plots for continuous variables:

- Dot plot (small sample size, shows clusters, groups and outliers)
- Stem and leaf plot (moderate sample size, displays density and shape of distribution, shows outliers
- Box plot (moderate to large sample size, good for comparing several sets of data, displays centre, spread, skewness and outliers.
- Histogram (large sample size, displays density and shape of distribution)

Plots for discrete and qualitative variables:

- Frequency tables
- Bar graph

Things that are sensitive to outliers:

- Sample mean
- Range (very)
- Sample standard deviation

Chapter 3

Relationships between 2 Quant variables:

Exploratory tool: ***Scatter plot***

Features to look for:

- Trend (linear or non linear)
- Scatter (constant or non-constant)
- Outliers
- Strength of relationship (strong or weak)
- Association (positive or negative)
- Groupings

Relationships between a Quant and a Qual variable:

Exploratory tool: ***Side by side plots of the same scale (dot or box) sample size indicator of which plot to use.***

Features to look for:

- Any group differences:
  * averages
  * variability (spread)
  * shapes (skewness, modes)
- Details of individual groups:
  * Outliers, gaps, clusters, groupings

Relationships between 2 Qual variables:

Exploratory tools: **Two-way table of counts OR Bar graphs of proportions.**

Features to look for:

- Most common and least common combinations
- Differences in distributions (row or column bar graphs)

Chapter 4

Conditional probability:

- Key words that indicate conditional situations are: **given that, of those, if, assuming that**

Statistical independence:

- Events A and B are statistically independent if **pr(A|B) = pr(A)**
- If A and B are statistically independent, then **pr(A and B) = pr(A) x pr(B)**
- If the *n* events are mutually independent then you can multiply them pr(A1) x pr(A2)….etc

Chapter 5

Confidence intervals:

An approximate 95% (normality based) confidence interval for a population mean, μ, has the form: $\bar{x} \pm t \times se(\bar{x})$

Standard error:
A measure of the variability of the estimate. The larger the standard error, the greater the variation there is in our estimation process. Vice versa. We're looking for small standard errors so our intervals are narrower. Roughly the average distance of the sample estimate from their mean.

Standard deviation:
Roughly the average distance of observations from their mean.

Margin of error:
The number subtracted from the point estimate. t multiplier and standard error together.
eg: $\bar{x} \pm t \times se(\bar{x})$ **= 12.5±2.201x1.654 = 12.5±3.64**

The margin of error is HALF the width of the confidence interval.

Confidence intervals for a single mean, μ

2 assumptions:

- We have to have random samples
- That the data has come from a normal distribution
-

t-multiplier:

The t-multiplier *t* **decreases** as the degrees of freedom, df **increases.**

For single means, as the sample size *n* increases, the standard error decreases, df increases which leads to the t multiplier decreasing and therefore the width of the interval decreases.

T multiplier and standard error ONLY affects the width of the confidence interval.

Difference between two proportions p1-p2

3 sampling situations:

1) Situation A: *Proportions from two independent samples*
2) Situation B: *One sample of size n, several response categories*
3) Situation C: *One sample of size n, many 'Yes/No' items*

Chapter 6

We always test the **null hypothesis.** We can **never** show or prove that H_0 **is true.**

4 Parameters we test:

- Single mean
- Single proportion
- Difference between two means
- Difference between two proportions

t-test hypotheses:

The null hypothesis, H_0 is always: H_0: *parameter = hypothesised value (a number)*
H_0 is always equals!

The alternative hypothesis H_1 has the following forms:

- H_1*: parameter ≠ hypothesised value (2 sided hypothesis)*
- H_1: *parameter > hypothesised value (1 sided hypothesis)*
- H_1: *parameter < hypothesised value (1 sided hypothesis)*

T-test statistic:

It is a measure of discrepancy between what we see in the data and what we would expect to see if the null hypothesis was true.

The t-test statistic, t_0, is the distance between the data estimate and the hypothesised value in terms of standard errors. It tells us how many standard errors the data estimate is away from the hypothesised value.

The larger the t-test statistic, the smaller the P-value

P-value:

The P-value is the probability of getting data like ours or worse.

The P-value measures the strength of evidence **against** the null hypothesis. The **smaller** the P-value, the stronger the evidence against H_0 **SMALL P's ARE SIGNIFICANT!**

Strength of P values against H_0

> 0.12 = None
= 0.10 = Weak
= 0.05 = Some
= 0.01 = Strong
< 0.001 = Very strong

Also, testing at the 5% level of significance:
< 0.05 = significant, reject H_0 in favour of H_1
> 0.05 = Not significant, do not reject H_0

Statistical significance vs practical significance:

Statistical significance relates to the **P-value.**
Practical significance relates to the **size of an effect.**

The size of an effect is estimated with a **confidence interval**. Look at the confidence interval when determining the practical significance of an effect.

Chapter 7

Assumptions for single mean t-procedures:

- Observations in the sample are independent (random sample) **CRITICAL**
- No clusters or multi modes
- 15 – 40 sample size guide.
  Small <15 – no outliers, at most slight skewness
  Medium 15 < n < 40 – no outliers, not strongly skewed
  Large > 40 – no gross outliers, may be strongly skewed

The normality assumption isn't critically important because the t-test is robust against departures from normality. As the sample gets **bigger** the t-test gets more robust. Two tailed tests are more robust than one tailed tests.

Paired data comparisons:

- The two sets of data are related (dependent)
- We ANALYSE THE DIFFERENCES
- A paired data t-test is the same as an one sample t-test applied to the differences

Non-parametric tests (sign test):

Advantage: No assumptions about the underlying distributions.
Disadvantage: parametric tests are more powerful; they are more likely to reject the null when the null is actually false. NO CONFIDENCE INTERVAL. If assumptions are met, use parametric.
*Also assumes independence

Sign test:

- Hypothesis statements use medians not means
- Observations above the hypothesised value is given a positive sign +
- Observations below the hypothesised value is given a negative sign –
- Observations that are the same as the hypothesised value are ignored
- Evidence against the null will be imbalance of signs

Assumptions for two independent mean t-procedures:

- Requires independence **WITHIN** each of the samples **CRITICAL**
- Requires independence **BETWEEN** the two samples (sample 1 is independent of sample 2) **CRITICAL**
- Normality assumption
- Sample size guidelines, n1+n2 gives total sample size. 15 – 40 guide applies

F-test for one-way ANOVA

We test: H_0= all of the underlying means are the same.
H_1 = Not all of the underlying means are the same

Evidence against the null hypothesis:
The data gives evidence against the null hypothesis when the variability **between** the sample means is large relative to the variability **within** the samples. When Sw is larger than Sb you will get a larger F-statistic and smaller P-value = evidence against the null.

Assumptions for the F-test:

- Samples are random **CRITICAL**
- Samples are independent **CRITICAL**
- The underlying distributions are normal
- The standard deviations of the underlying distributions are equal
  Largest SD/smallest SD < 2 as a guide

The F-test is robust against departures from normality like the 2 sample t-test.
The F-test is reasonably robust with the equal standard deviations assumption, but the Tukey pairwise confidence intervals are not.

Chapter 8

Two kinds of Chi-square tests:

- If we have **qualitative** data presented in a **one way table of counts** we used chi-square for **goodness of fit**
- If it is presented in a **two way table of counts** we use chi-square to test for **independence**

One way tables:

How to calculate expected count for one way table:
Total of the table x the expected probability of the column.

Degrees of freedom:
df = J-1 J = number of categories

Hypotheses: H_0**=** the data comes from the specified distribution
H_1 = the data does not come from the specified distribution

Two way tables:

How to calculate expected count for two way table:
Roman Catholics over nuns. Ri x Cj/ n Ri and Cj are the row and column totals. N is the total for the table.

Degrees of freedom:
df – (I-1)(J-1) I = number of rows J = number of columns

Hypotheses: H_0 = the distribution of variable 1 is the same for each level of variable 2
H_1 = the distribution of variable one is not the same at all levels of variable 2
The null hypothesis is often written as a statement of 'sameness' (Independence is also used)

Validity of Chi-square tests:

- At least 80% of **expected counts** must be **5 or more**
- Each **expected count** must be **greater than 1**

Chapter 9

Regression: using one or more quantitative variable to predict/explain/describe the behaviour of a second variable

Residuals: deviations of points from the line. (distance)

Regression equation: Y = c+mx c= constant (y intercept) m=slope

Residual: observed value – predicted value y - $\hat{y}$

Equation of a straight line: $\boldsymbol{y = ß_0 + ß_1 x}$

- $ß_0$ is the Y intercept
- $ß_1$ is the slope of the line
- They are both never known, they are the TRUE intercept and TRUE slope
- If they have ^ hats they are the sample parameters
- $ß_1$ describes how much does Y change when X changes by 1 unit