STAT1006 Week 5 Cheat Sheet

Lisa Luff

9/17/2020

Contents

Interpreting Data

- Questions to ask:
  - What *cases* does the data describe?
    - Cases are what you collect the data from/about. Eg. When analysing student's heights, the case is the student.
  - What variables are present and how are they measured?
    - The variables are what we are testing, but sometimes you need to take into account other varibles that will impact the oucome.
    - We also need to make sure that how we measure is appropriate and consistent. Eg. With student heights, do we measure in cms or feet. Do we push down their hair? Measure against a wall or free standing?
  - What data type are the variables?
    - Categorical or numerical?
  - Do some variables explain or even cause changes in the other variables?
    - If so, how do we factor this into our analysis?

Analysing the Relationship Between Numerical Variables

- Start with a graph
  - Look for an overall Pattern and deviations from the pattern
  - Use numerical descriptions of the data and overall pattern (if appropriate)
- Define the variable types:
  - A *response* variable
    - Measures or records an outcome of a study
    - And the *response or dependent* variable is plotted on the *y axis*
  - An *explanatory* variable
    - Explains changes in the response variable
    - Typically the *explanatory or independent* variable is plotted on the *x axis*
- How to decide your graph type:
  - *Response* variable is *numerical*, and *categorical* variable is *categorical*
    - Side by side box-plots
    - Side by side histograms
    - We typically analyse this with an ANOVA
  - *Both* variables are *numerical*
    - Scatterplot
    - We typically analyse this with regression
- The type of graph used, or how you use them to represent the data will depend on the combination of data you use, or want to analyse them.

Linear Relationship Between Two Numerical Variables

- The relationship of two variables can be summarised:
  - Graphically - Scatterplot
  - Numerically - Correlation

Graphical Summary (Scatterplots)

- Interpreting scatterplots:
  - We look for an overall pattern:
    - Form(Shape) - Linear, curver, clusters, no pattern
    - Direction - Postive, negative, no direction
    - Strength - How closely the point fit the "form"
    - Outliers - Are there any deviations from the pattern
  - Direction:
    - Positive direction - High values of one variable tend to associate with high values of the other variable
    - Negative direction - Low values of one variable tend to associate with high values of the other variable
    - No relationship - The variables vary independently of one another/knowing one, doesn't tell you anything about the other
      - This can look like a horizontal line, or complete scatter
  - Strength:
    - Based on how much variation, or scatter there is around the main form
    - A strong relationship means one variable will give a pretty good estimate of the other
    - A weak relationship, any one value for one variable will give a range of values for the other variable
- If you have two numerical variables based on one categorical variable, you can do a coloured scatterplot, and interpret each category of data separately on the same scatterplot

- Scale for a scatterplot:
  - Using an inappropriate scale for a scatterplot can give an incorrect impression
  - Both variables should have a similar amount of space
    - Plot roughly square
    - Points should occupy entire plot space (no blank space)

Numerical Summary (Pearson Sample Correlation Coefficient)

- The Pearson Sample Correlation Coefficient - r
    - Measures the direction and strength of the linear relationship between two numerical variables
- How to calculate:
    - $r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)S_x S_y} = \frac{S_{xy}}{S_x S_y}$
    - Where:
        - $\bar{x}$ - Sample mean of X
        - $\bar{y}$ - Sample mean of Y
        - S_x - Sample standard deviation of X
        - S_y - Sample standard deviation of Y
        - S_{xy} - Sample covariance between X and Y
        - n - Number of observations
- Use R to calculate, we don't need to do it manually
    - cor(x, y = NULL, use = "everything", method = c("pearson", "kendall", "spearman"))
        - If method is unspecified, the default is Pearson
    - However, note that the formula consider the variation in the X variable, in relation to the variation in the Y variable
- Understanding correlation:
    - The correlation r always falls between -1 and 1
        - Positive r - Indicates *positive* direction between variables
        - Negative r - Indicates *negative* direction between variables
        - r near 0 - Indication *no relationship* between variables
    - The closer to -1 or 1, the stronger the relationship
- r does not distinguish between X and Y/they are treated symmetrically
    - Therefore swapping which variable is X and Y
- The units of both will *not* affect the value of r
    - This is because we get rid of units when we standardise (get z-scores)

Assumptions for Correlation Coefficient Methods

- Pearson's Correlation Coefficient:
    - Both variables are on equal interval/ratio scales
    - *Linear* relationship between the variables
    - Each variable has a *Normal* distribution
- Spearman's and Kendall's Correlation Coefficients
    - Independence
    - Continuity
        - $F_{X,Y}$ is a continuous distribution
        - And therefore there is a numerical element
    - However both are free of distributions

Other Types of Correlations (Spearman's and Kendall)

- Spearman:
    - Correlation Coefficients based on *ranks*
    - Ordinal(categorical) and/or non-normal distribution
- Kendall
    - To deal with data samples with *tied ranks*
    - It is known as the Kendall's tau-b coefficient and is more effective in determining whether *two-nonparametric* data samples with ties are correlated
    - Nominal(both categorical) data

Comparing Pearson's and Spearman's

- When assumptions for Pearson's cannot be met, then use Spearman's rank correlation coefficient
- Pearson's
  - Measure only the degree of *linear* association
  - Based on the assumption of bivariate Normality (Normality of $F_{X,Y}$)
- Spearman's
  - Take into account only the ranks
  - Measure the degree of *monotone* association
    - If X increases, does Y increase
  - Inferences on the rank correlation coefficients are distribution-free

Hypothesis Testing for a Linear Relationship - Pearson's

- The test of significance for ρ(Rho)(population correlation) uses the one-sample t-test for $H_0 : \rho = 0$
  - We compute the t statistic for sample size n and sample correlation coefficient r

1. Hypotheses

- $H_0 : \rho = 0$ - No correlation
- $H_A : \rho \neq 0$ - Correlation

2. Test Statistic

- cor.test(x, y, alternative = c("two.sided", "less", "greater"), method = c("pearson", "kendall", "spearman"), exact = NULL, conf.level = 0.95, continuity = FALSE, ...)
  - Uses base package - Kendall and Spearman has more precise packages
- Simple form:
  - $r = \sqrt{r^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$
- Actual form:
  - $t = \frac{r - \rho}{\sqrt{\frac{1-r^2}{n-2}}}$, where
    - $r = \sqrt{r^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$

3. Sampling Distribution

- $t \sim T$ with df (n - 2)

4. The P-Value

- The P-Value is the area under the sampling distribution $T(n-2)$ for values of T as or more extreme than t in the direction of H_A
  - $H_A : \rho > 0$ is P($T \geq t$)
  - $H_A : \rho < 0$ is P($T \leq t$)
  - $H_A : \rho \neq 0$ is 2P($T \geq |t|$)

5. Decision
6. Conclusion

Inference for Correlation

- When the hypothesis $H_0 : \rho = 0$ is:
  - Rejected - It is safe to assume that there is some sort of relationship between the variables X and Y
  - *Not* rejected - Do not assume that the variables are unrelated
    - It is possible a *Type II error* has occured
      - A Type II error is when the null hypothesis is incorrectly accepted. This can occur because the methodology needs to be more stringent, or that the sample happens to just be non-representative.
    - It is possible X and Y are related in a nonlinear way that the correlation coefficient r has no chance of detecting
      - A good way to investigate if this is the case if to examine a residual plot

Hypothesis Testing for a Linear Relationship - Spearman's

- The test of significance for ρ_s uses approximate Normal distribution for $H_0 : \rho_s = 0$
- We compute the z statistic for sample size n and sample correlation coefficient r_s

1. Hypothesis

- H_0: X and Y are independent
- H_A: X and Y are dependent

2. Test statistic

- cor.test(X, Y, method="spearman", alternative=, exact=F)
- Exact form:
  - $r_s = \frac{\sum_{i=1}^{n}(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_{i=1}^{n}(u_i - \bar{u})^2 \sum_{i=1}^{n}(v_i - \bar{v})^2}}$, where
    - u_i - Rank(x_i)
    - v_i - Rank(y_i)
    - $d_i - u_i - v_i$ = Difference in ranks
    - n - Number of pairs of X's and Y's
  - If u_i and v_i are integers, then a more convenient formula is:
    - $r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$
  - For a large sample n
    - E(r_s) $= 0$
    - Var(r_s) $= \frac{1}{n-1}$
- Normal approximation:
  - $Z = \frac{r_s - E(r_s)}{SD(r_s)} = \frac{r_s - 0}{\frac{1}{\sqrt{(n-1)}}} = \sqrt{(n-1)} \sim N(0,1)$

3. Sampling Distribution

- Exact - Based on ranks
- Approximation - Based on $Z \sim N(0,1)$

4. The P-Value

- cor.test uses:
  - $n < 10$ - Uses an Edgeworth series approximation and is considered to be exact = TRUE
  - $n < 1290$ - Uses algorithm AS 89
  - Otherwise, uses asymptotic t approximation

5. Decision
6. Conclusion

Hypothesis Testing for a Linear Relationship - Kendall's Rank Correlation

- The parameter of interest is Kendalls Population Correlation Coefficient τ(tau) also known as $\hat{\tau}$, which we can estimate using Kendalls Test Statistic $\overline{K}$
- The null hypothesis is statistical independence:
  - $F_{X,Y}(X, Y) = F_X(X)F_Y(Y)$ for all (X, Y), where
    - $F_{X,Y}$ - Joint distribution of X and Y
    - F_X - Marginal distribution of X
    - F_Y - Marginal distribution of Y
  - Or $H_0 : \tau = 0$

1. Hypothesis

- $H_0 : \tau = 0$
- H_A :
  - $\tau > 0$ - One-sided upper-tail (positively correlated)
  - $\tau < 0$ - One-sided lower-tail (negatively correlated)
  - $\tau \neq 0$ - Two-sided (correlated)

2. Test Statistic

- $\tau = 2P[(Y_2 - Y_1)(X_2 - X_1) > 0] - 1$
- $K = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} Q[(X_i, Y_i), (X_j, Y_j)]$, where
  - For all $\frac{n(n-1)}{2}$ pairs of observations (X_i, Y_i) and (X_j, Y_j) with $1 \leq i < \leq n$, calculate the *paired sign statistic* $Q[(X_i, Y_i), (X_j, Y_j)]$, where
  - $Q[(a, b), (c, d)] = \begin{cases} 1 \text{ if } (d-b)(c-a) > 0 \\ -1 \text{ if } (d-b)(c-a) < 0 \end{cases}$
- For a large sample n $(n > 50)$
  - $E(K) = 0$
  - $Var(K) = \frac{n(n-1)(2n+5)}{18}$
  - $K^* = \frac{K - E(K)}{\sqrt{Var(K)}}$
    - Which asymptotically follows a N(0, 1) distribution
- Kendall's τ Rank Correlation Coefficient - Estimates population τ using the sample estimate
  - $\hat{\tau} = \overline{K} = \frac{2K}{n(n-1)}$, given that
    - $-\frac{n(n-1)}{2} \leq K \leq \frac{n(n-1)}{2}$

3. Sampling Distribution

- Exact - Based on ranks when $n < 50$ (n - paired samples) that contain finite values and there are no ties
- Approximation - Otherwise, $\tau \sim Z \sim N(0, 1)$

4. The P-value

- cor.test(X, Y, method="Kendall", alternative=)

5. Decision
6. Conclusion