STAT1006 Week 1 Cheat Sheet

Lisa Luff

8/6/2020

Contents

Revision

Statistical Inference

- Select from:
  - A statistic is a property of the sample
  - Statistics of a sample:
    - Mean: $\bar{(x)}$
    - Standard Deviation: s
    - Proportion: $\hat{(p)}$
- Which is a representation of:
  - A parameter is a property of the population
  - Parameters of a population are:
    - Mean: μ
    - Standard Deviation: σ
    - Proportion: p
- Statistical inference is:
  - Hypothesis testing
  - Estimation:
    1. Point estimation
    2. Confidence interval

Hypothesis Testing

1. State the hypotheses (null and alternative)
2. Calculate the test statistic
3. Determine the sampling distribution of the test statistic

- Eg. The sampling distribution is t, with $df = x$

4. Find the p-value
5. Make a decision
6. State your conclusion

Testing a population mean: One Sample T-Test

- σ is the population standard deviation
- μ is the population mean
- s is the sample standard deviation
- $\bar{x}$ is the sample mean
- n is the sample size
- When σ is known - Z test
  - Test statistic: $z = \frac{(\bar{x}-\mu)}{\frac{\sigma}{\sqrt{n}}}$
  - Which follows the Standard Normal Distribution, N(0, 1)
- When σ is unknown - t-test
  - Test statistic: $t = \frac{\bar{x}-\mu}{(\frac{s}{\sqrt{n}})}$
  - Which follows Student's T-Distribution with df = n - 1

Robustness of T-Tests

- If a significance test is robust, the p-value does not change very much when the conditions for use of the procedure are violated
  - For t-tests
    - Except for small samples, simple random sampling is more important than the condition that the population distribution is Normal

    a) Sample size at least 15: t procedures can be used, except in the presence of strong outliers or skewness

    b) Sample size less than 15: Use t procedures if the data appear close to Normal

    c) Large samples(Central Limit Theorum), roughly n $\geq$ 40: The t procedures can be used even for clearly skewed distributions

Continuity Correction

- A continuity correction factor is used when you use a continuous function to approximate a discrete one
  - eg. When you use a Normal distribution to approximate a binomial
- You just add or subtract 0.5
  - Deciding whether to add or subtract 0.5:
    - $P(X = n)$ use $P(n - 0.5 < X < n + 0.5)$
    - $P(X > n)$ use $P(X > n + 0.5)$
    - $P(X < n)$ use $P(X < n - 0.5)$
    - $P(X \geq n)$ use $P(X > n - 0.5)$
    - $P(X \leq n)$ use $P(X > n + 0.5)$

Which Statistical Modelling/Inference?

Which statistical modelling/inference?

Type of objective	Type of data	Statistical method/model (PARAMETRIC)	Statistical method/model (NONPARAMETRIC)
RELATIONSHIPS	2 numerical (1 or more than 1 numerical explanatory, 1 numerical response)	Simple Linear Regression Multiple Linear Regression (Weeks 5-12)	
COMPARISONS	Numerical – 1 sample from 1 population	One sample t-test	Wilcoxon Signed Rank Test (Week 1)
	Numerical – 1 sample paired	Paired t-test	Wilcoxon Signed Rank Test (Week 2)
	Numerical – 2 samples from 2 independent populations	Two sample t-test	Wilcoxon Rank Sum Test (Week 2)
	Numerical – 3 samples or more from 3 or more independent populations	ANOVA	Kruskal Wallis Test (Week 3)

STAT1006 Semester 2 2020 15

New Topics

Non-Parametric (Distribution Free) Testing

- Extension of t and F tests for cases where the usual assumption of Normality may not be tenable
  - Assumptions made for single sample t-tests:
    1. The data is a random sample
    2. The distribution of the sample means are Normal
       - This can mean:
         a) The underlying data distribution is Normal
         b) The sample size is large enough for the Central Limit Theorum to give a Normal distribution for the sample means
  - Assumptions made for non-parametric methods (usually):
    1. Independence
    2. Population distributions are continuous
- Methods are based on ranks, which, after calculating, the actual observations are not used
  - Disadvantage: Can be less efficient because not all the information in the observations is used

Addressing non-Normal data

- Is the lack of normality due to outliers?
  - If the outlier appears to be "real" data, you have to leave it in
  - If you have reason to think the outlier is an error, you may be able to remove it
- Try transforming the data
  - Try a logarithm for right-skewed data that are positive numbers
- Try another standard distribution
  - Other procedures can replace the t procedures if data (especialy right-skewed data) fit another distribution
- Use modern bootstrap methods and permutation tests
  - Although often more computationally intensive than t procedures, such methods avoid the requiement of a specific type of distribution for the population
- Use non-parametric methods, as discussed here

Non-Parametric Hypothesis Testing

- If a sample size is small and the population is non-Normal, non-parametric tests provide an alternative
- Mean is replaced by median, denoted as $\tilde{\mu}$ or η
  - We use median because population might be skewed
- How to apply this as a sign test (rank test is an extension of this):
  1. State hypotheses in terms of $\tilde{\mu}_0$ and $\tilde{\mu}_A$ and compare with $\tilde{\mu}$
  2. Replace each value in the data set with + if it is greater than $\tilde{\mu}_0$, or - if it is less
  3. If H_0 is true, then the number of plus signs should be approximately equal to the number of minus signs
  - If we observe more of one sign than the other than would occur by chance alone, reject H_0

Rank Transformations

- Based on the ranks, or order, of the data and not the actual values
- Focuses on the center of the population(s)
  - If a Normal distribution: mean
  - If a skewed distribution: median
    - Ranking data is essential for median, as you must order the values to find which is at the center

- To rank observations, order them from smallest to largest, then assign a value to each starting with 1 for the smallest
  - Moving the values to their ranks retains only the ordering of the observations and makes no other use of their numerical values
- Dealing with **ties** in rank tests:
  - If two or more values are the same as each other
    - Each will be assigned the average of the corresponding ranks
    - Then you continue ranking with the next value
      - EG: 3rd and 4th values are equal, $(3 + 4)/2 = 3.5$, so both the 3rd and 4th values will be assigned as 3.5, and then the next value will the 5th.

The Wilcoxon Signed Rank Test

- The exact distribution for the Wilcoxon test W applies only data that aren't tied
- Most statistical software will detect ties and make the necessary adjustments

1. Draw a simple random sample from the population of size n, with values of $X_1, X_2, ..., X_n$
2. Set the hypotheses in terms of $\tilde{\mu}$
3. Transform each value of X_i into $X_i - \tilde{\mu}_0 = X_i'$
4. Remove ties (this would be when $X_i' = 0$)
5. Rank the **absolute** values of X_i'

- If there are ties between $|X_i'|$ values, then statistical software will be needed to evaluate, as it will be able to make the necessary adjustments

6. The sum W^+ of the ranks for the postive differences (used to evaluate $\tilde{\mu} < \tilde{\mu}_0$) is the Wilcoxon signed rank statistic (Not positive because they're absolute, only those whose value X_i' was positive)

- This can also be done to calculate W_- (sum of negative values), when evaluating $\tilde{\mu} > \tilde{\mu}_0$
- Or to calculate W $(min(W^+, W_-))$, when evaluating $\tilde{\mu} \neq \tilde{\mu}_0$

7. Calculate the p-value for the Wilcoxon signed rank statistic using the Normal approximation with the continuity correction (assuming n $\geq$ 15)

- R uses this approximation
- If the distribution of the responses is not affected by the different treatments within pairs, then W^+ has:
  - Mean: $\mu_{W^+} = \frac{n(n+1)}{4}$
  - Standard deviation $= \sigma_{W^+} = \sqrt{\frac{n(n+1)(2n+1)}{24}}$
- Under H_0, the sampling distribution of $z = \frac{(W^+(\pm 0.5 \text{ continuity correction}) - \mu_{W^+})}{\sigma_{W^+}}$ is approximately Standard Normal

8. The Wilcoxon signed rank test rejects the hypothesis that there are no systematic differences between X_i' when the rank sum W^+ is far from its mean

Linear Models and Linear Regression

- Linear models are the basis of many common methods of statistical analysis
  - eg. t-tests (one or two means), f-tests (several means) and multivariate methods