STAT1006 Week 4 Cheat Sheet

Lisa Luff

8/27/2020

Contents

Resampling

- Resampling is also non-parametric
- There are many different types but we are looking at
  - Bootstrapping
  - Permutations

Resampling Assumptions

- Still needs to be random sampling
- The resamples are not independent

Pros of Resampling

- Usually it is accepted that we act "as if" the assumptions for standard statistics are satisfied
- Good for awkward or "interesting" statistics that cannot utilise standard statistics
- Don't require such a multitude of formulas for different problems
- More accurate in practice than standard methods

Bootstrapping and Permutation Tests

- We can relax some conditions by utilising the power of computers
- The fundamentals are still applied, and the answers received are in the same form
- They are conceptually simple by answering what would happen if we took many samples
- There are some limitations, but their effectiveness and range of use means they're highly utilised

Bootstrapping vs Permutation Tests

Resampling	Application	Sampling Procedure
Bootstrapping	Standard deviation, confidence interval, hypothesis testing, bias	Samples drawn at random, **with replacement**
Permutations	Hypothesis testing	Samples drawn at random, **without replacement**

Bootstrapping

- Used to draw statistical inference on a population from a sample
- The idea:
  - Original sample represents the population it is drawn from
  - So resamples should represent what we would see from many samples from the population
    - We don't need to rely on Central Limit Theorum, etc.
- The bootstrap distribution of a statistic is created from the sample statistic of the resamples
- For many resamples, the distribution will mostly be approximately the same as the population distribution
  - Where the original sample becomes the new population, with the distribution centered around the original sample statistic, rather than the parameter value
- Used to estimate a parameter (mean), and variability of the estimate (variance)
  - We don't need a formula for the test statistic or standard deviation

Steps of Bootstrapping

1. Resampling

- Create many resamples by repeatedly sampling **with replacement** from the single random sample
- Each resample is the **same size** as the original random sample
- Because we have replacement, it means resamples can have multiples of a value from the original sample

2. Bootstrap distribution

- The sampling distribution of a statistic represents the values of the statistic in all samples of the same size from the population
- The bootstrap distribution of a statistic is the values taken by the statistic in all possible resamples of the same size.
- The bootstrap distribution gives information (shape and spread) about the *sampling* distribution

3. Statistics

- Regard $X_1, X_2, ..., X_n$ as the new population
- Resample it B times with replacement, $X^*_{b1}, X^*_{b2}, ..., X^*_{bn}$, where b = 1, 2, . . . , B
- We need clear notation to distinguish between the sample statistics, and those of the resamples
  - $\hat{F}$ - The empirical distribution function of the observed data
  - Sample mean: $\hat{\theta}$
  - Resampling mean: $\hat{\theta}^*(b)$
  - Sample standard error: $\hat{se}$
  - Resampling standard error: $s(x^{*b})$
  - Bootstrap Distribution mean: $\hat{\theta}^*(\cdot)$
  - Bootstrap Distribution standard error: $\hat{se}_B$

- $\hat{\theta}^*(\cdot) = \sum_{b=1}^{B} \frac{\hat{\theta}^*(b)}{B}$
  - The mean of the bootstrap, is the mean of the mean of the resamples
- $\hat{se}_B = [\sum_{b=1}^{B} \frac{[\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)]^2}{B-1}]^{\frac{1}{2}}$
  - The standard deviation of the bootstrap, is the standard deviation of the standard deviation of the resamples
- The percentile method for confidence interval
  - Rank the values for $\hat{\theta}^*(b)$
  - For a 95% confidence interval, after ranking the bootstrapped θ coefficients, just take the 2.5% as the lower confidence limit and the 97.5% as the upper confidence limit
  - The percentile (1 - a)*100% confidence interval for a population mean is:
    - $(\hat{\theta}^*_{(\frac{a}{2})}, \hat{\theta}^*_{(1-\frac{a}{2})})$

Permutation Tests

- Significance (hypothesis) tests based on permuation resamples of the original data
- Resamples are drawn **without** replacement
- Can also be called:
  - Randomisation tests
  - Re-randomisation tests
  - Exact tests
- The idea is that if the null hypothesis that there is no difference between two samples is true, then if we create random samples of the same size as the two original samples, using a combination of data from each sample, for the two new samples, the test statistic, should not be far from the original test statistic
  - If the test statistic of the sample is as large or larger than what is considered rare for the resamples, we would reject the null hypothesis

Permutation Conditions

- Can only be used when we are able to see how to resample in a way that is consistent with the study design
  - Needs to align with testing the null hypothesis that the observations are exchangeable
- Requires equal variances **very important**

- If you cannot do a permutation test, we can calculate a bootstrap confidence interval instead

Pros and Cons of Permutation Tests

- Pros
  - You can use it for any test statistic, regardless of if the distribution is known
  - You can choose what statistic best differentiates between the null and alternative hypothesis
  - Can be used for unbalanced designs
  - Can combine dependents tests on mixtures of categorical, ordinal and metric data
- Cons
  - Limited by the conditions of use
  - Same weakness to variance as classical Student's t-test

Steps of Permutation Testing

1. Analyse the problem

- Find the hypotheses
- What distribution is the data drawn from?

2. Choose a test statistic

- To distinguish the null hypothesis from the alternative

3. Rearrange the observations

- Compute the test statistic for **all possible permutations of the data from the observations**
- Notation:
  - Sample sizes: n_x, n_y
  - Total number of samples: N
  - Samples: $X = x_1, ..., x_{n_x}$ and $Y = y_1, ..., y_{n_y}$
  - Test statistic: T
  - Sample means: $\overline{X}, \overline{Y}$
  - Sample variances: S_x^2, S_y^2
- There will be $C^N_{n_x(\text{or } n_y)} = \binom{n}{m} = \frac{n!}{m!.(n-m)!}$ permutations total to be calculated
  - Note: if both samples are the same size, you only need half the permutation due to symmetry
- $T = \frac{\overline{X}-\overline{Y}}{\sqrt{\frac{(n_x-1)S_x^2+(n_y-1)S_y^2}{(n_x-1)+(n_y-1)}}} \cdot \sqrt{\frac{n_x.n_y}{n_x+n_y}}$

4. Compute the test statistic

- For the original sample
  - Under $H_0, T \sim t_{n_x+n_y-2}$ (t-distribution with $n_x + n_y - 2df$)
  - $\hat{F}$ - The empirical distribution function of the observed data
- Compute the p-value for the observed value t of test statistic T
  - $p = 1 - P(|T| \leq |t|H_0) = 2[1 - P(|T| \leq |t|H_0)] = 2[1 - F_{t,n_x+n_y-2}(|t|)]$
    - Essentially the p-value is the number of values for t from the resamples that are higher than the sample t divided by the total number of permutations

5. Make a decision

- Reject the null hypothesis if the value of the test statitic for the original data is an extreme value in the permutation distribution of the statistic
- Otherwise accept the null hypothesis and reject the alternative
- Decision rule: rejct H_0 if p-value $\leq \alpha$

In R

- Several package available, we are using **boot** and **perm**
- Need to see in class what is happening with how to do the bootstrapping
- Permutations:
  VectorA <- c()
  VectorB <- c()
  permTS(*VectorA*, *VectorB*, alternative = '', method = 'exact.mc')

QQ Plots

- QQ plots, or Normal quantile plot
- Test for normality, the closer the points to the red line, the closer to normal
- If you have some spots quite far from the red line, there would be some concern about using a t-test
  - Likely skewed data