STAT1006 Week 3 Cheat Sheet

Lisa Luff

8/19/2020

Contents

The χ^2 and F distributions are both right-skewed, which means they are one way. So we are only looking at the distance of the statistic from the left side. You cannot ignore whether it is negative or positive because they are not symmetrical. So the larger the statistic, the stronger the arguement to reject the null hypothesis. You must minus from 1 if looking at whether something is greater than.

This means that because we are testing whether the significance of the difference is greater than there being no difference, we need to find 1 - the probability of the test statistic

ANOVA

- Parametric
- ANOVA is used to analyse completely randomised experiments
  - Due to being random, it is assumed subjects are homogenous
  - Also called analysis of variance
  - OR analysis of variance F test
- Different populations will experience sampling variability
- Inference is the mean response
- **We are comparing means, not variance**
- ANOVA asks if a set of means give evidence for differences among the population means
  - It's not about how far about the means are, but how far relative to the variability of the individual observations
  - Looking for statistical significance of differences between means

ANOVA Basics

- Data = factor effect + error OR Data = fit + residual
  - Random sampling will always have chance variations
    - So any differences will be the actual effect of the factor PLUS chance variation
- $X_{ij} = \mu_i + \varepsilon_{ij}$ for $i = 1, ..., K$ and $j = 1, ..., n_i$
  - Where ε_{ij} is variance

ANOVA Assumptions

- That we have K **independent SRS's** from each population
- We measure the same response from each sample
- Each population is Normally distributed (using usual rules to test if Normal)
  - With an unknown mean μ
  -
    - For ANOVA models, chance variations are assumed to be Normally distributed

- All the populations have the same standard deviation σ, which is unknown
  - Use Levene Test to check if the variance amongst samples are equal
    - The Levene Test is based on the null hypothesis that the variances are equal, and will give a p-value indicating the probability that they are
  - Use Welch ANOVA test for unequal variances instead of F test

One vs Two-Way ANOVA

- One-Way ANOVA is when only one factor/or there is only one way to classify the populations of interest
  - One factor - Factor = IQ (only thing tested), Populations = Students from several universities (multiple)
  - One way to classify populations - Factor = tyre tread lifetimes (multiple tyre tread types tested), populations = mix of tyre brands (number of populations grouped together to make a single population)
- Two-Way ANOVA is used for two factors
  - Two factors - For students, testing IQ and average grade

ANOVA vs t-test

- Use t-test for 2 groups, as the F test is more limited
- Use the F test for more than two groups

ANOVA Parameters

- The model parameters are population means $\mu_1, \mu_2, ..., \mu_K$ and standard deviation σ
- The unknown parameters are all K population means(μ_i) and the common population standard deviation σ
- To estimate μ_i, use the sample mean:
  $\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$
- To estimate σ, we use the pooled(or common) standard deviation:
  $s_p = \sqrt{\frac{(n_i-1)s_1^2+(n_2-1)s_2^2+...+(n_K-1)s_K^2}{(n_1-1)+(n_2-1)+...+(n_K-1)}}$

ANOVA Hypotheses

- Null hypothesis is that there are no differences between the means
  $H_0 = \mu_1 = \mu_2 = ... = \mu_K$
- The alternative hypothesis is that there is **some difference**
  $H_A = \mu_1 \neq \mu_2 \neq ... \neq \mu_K$

ANOVA Partition of Total Variation

- Total variation = Variation due to treatment + variation due to random sampling
  - Total variation is also known as:
    - Sum of squares total
  - Variation due to treatment is also known as:
    - Sum of squares among (SSA)
    - Sum of squares group (SSG)
  - Variation due to random sampling is also known as:
    - Sum of squares error (SSE)
    - Sum of squares within (SSW)

Total Variation (SST)

- X_{ij} - The $i-th$ observation in group j
- n_j - The number of observations in group j
- n - The total number of observations in all groups
- K - The number of groups
- $\overline{\overline{X}}$ - The overall or grand mean
  $\overline{\overline{X}} = \frac{\sum_{j=1}^{K}\sum_{i=1}^{n_j} X_{ij}}{K}$
  SST $= \frac{\sum_{j=1}^{K}\sum_{i=1}^{n_j} X_{ij}}{K}(X_{ij} - \overline{\overline{X}})^2$
  OR
  SST $= (X_{11} - \overline{\overline{X}})^2 + (X_{21} - \overline{\overline{X}})^2 + ... + (X_{n_K K} - \overline{\overline{X}})^2$

Among Group Variation (SSA/SSG)

- $\overline{X}_j$ - The sample mean of group j
  SSA/SSG $= \sum_{j=1}^{K} n_j(\overline{X}_j - \overline{\overline{X}})^2$
  OR
  SSA/SSG $= n_1(\overline{X}_1 - \overline{\overline{X}})^2 + n_2(\overline{X}_2 - \overline{\overline{X}})^2 + ... + n_K(\overline{X}_K - \overline{\overline{X}})^2$

Within Group Variation (SSE/SSW)

- X_{ij} - The $i-th$ observation in group j
  SSE/SSW $= \sum_{j=1}^{K}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_j)^2$
  OR
  SSE/SSW $= (X_{11} - \overline{X}_1)^2 + (X_{21} - \overline{X}_1)^2 + ... + (X_{n_K K} - \overline{X}_K)$

ANOVA F Test Statistic

F $= \frac{\text{Variation among the sample means}}{\text{Variation among individuals in the same sample}}$

- The measures of variation in the numerator and denominator are *mean squares*
  - Numerator - Mean square for groups (MSG)
    MSG $= \frac{SSA}{K-1}$, where $K-1$ is the degrees of freedom
  - Denominator - Mean square for error (MSE)
    MSE $= \frac{SSW}{n-K}$, where $n-K$ is the degrees of freedom
  - For K = 2, the is the pooled - variance in the t-test
  - MSE is also knows as the *pooled/common sample variance* written as s_p^2
    - s_p is the *pooled/common* standard deviation
    - s_p^2 estimates the common variance σ^2

Checking Standard Deviations in ANOVA

- The results of the ANOVA F test are approximately correct when the largest sample standard deviation is no more than twice as large as the smallest sample standard deviation

F Distributions

- F dsitributions are a family of right-skewed distributions, with values only greater than 0
- A specific F distribution is determined by its degrees of freedom using the numerator and denominator of the F statistic
  - Always give the degrees of freedom first when describing an F distribution
  - Notation is F(df1, df2)
    - Where df1 is given by the numerator
    - And df2 is given by the denominator
- Degrees of freedom for the F test
  - When comparing K populations, with an SRS of n_i from the i^{th} population
  - The total number of observations is
    - $N = n_1 + n_2 + ... + n_K$
- If the null hypothesis is true, the ANOVA F statistic has the F distribution with
  - $K - 1$ degrees of freedom in the numerator
  - AND $N - K$ degrees of freedom in the denominator

6 Step ANOVA Hypothesis Testing

1. Hypothesis testing

- $H_0 = \mu_1 = \mu_2 = ... = \mu_K$
- $H_A = \mu_1 \neq \mu_2 \neq ... \neq \mu_K$

2. Test statistic

- $F_0 = \frac{MSG}{MSE}$

3. The sampling distribution

- F distributed as F with df = (K - 1, N - K)

4. p-value

- P(F(K - 1, N - K) $>$ F_0)

5. and 6. Decision and Conclusion

ANOVA in R

variablename.aov <- aov(*Factor~Populations*)
summary.aov(*variablename*.aov)

The Kruskal-Wallis Test

- Non-parametric
- A rank test that can replace the ANOVA F test
- Used to analyse completely randomised experimental designs
- Rank all the responses from all groups combined, and then apply one-way ANOVA to the ranks, rather than the original observations
- Kruskal-Wallis test statistic is basically SSG for ranks
- Extension of Wilcoxon Rank Test for more than 2K population **medians**
- Distribution-free test procedure

Kruskal-Wallis Assumptions

- Independent random samples
- Continuous dependent variable
- Data may be ranked both within and among samples
- Populations have same variability
- Populations have same shape
- **Robust** in regard to the alst two
  - Use F test in completely randomised designs and when the more stringent assumptions hold

Kruskal-Wallis Test vs ANOVA F Test

- Kruskal-Wallis can replace ANOVA F test
  - You can relax the assumption about Normality
  - Independent random sampling is still important!

Kruskal-Wallis Hypotheses

- H_0 = response has same distribution in all groups
  $H_0 = \tilde{\mu}_1 = ... = \tilde{\mu}_K$ (If distributions have the same shape)

- H_A = response is systematically different in some groups than in others
  $H_A = \tilde{\mu}_1 \neq ... \neq \tilde{\mu}_K$ (If distributions have the same shape)

- When H is large, we reject the null hypothesis that all populations have the same distribution

Chi-Squared Approximation for Kruskal-Wallis

- When the sample sizes are large and all populations have the same continuous distribution, H has approximately the chi-square (χ^2) distribution with K - 1 degrees of freedom
  - Use χ^2 distribution to approximate if each sample group
    - Size = $n_i > 5$
    - Df = K - 1

Kruskal-Wallis Test Statistic

- Draw independent SRS's of sizes $n_1, n_2, ..., n_K$ from K populations
- There are N observations total
- Rank all N observations and let R_i be the sum of the ranks for the i^{th} sample
- The Kruskal-Wallis statistic is
  $H(\text{or } KW) = \frac{12}{N(N+1)} \sum \frac{R_i^2}{n_i} - 3(N+1)$
- When the sample sizes n_i are moderately large and all K populations have the same continuous distribution,
  $H(\text{or } KW) \sim \chi^2_{K-1}$

6 Step Kruskal-Wallis Hypothesis Testing

1. Hypothesis testing

- $H_0 =$ response has same distribution in all groups
- $H_A =$ response is systematically different in some groups than in others
- OR if distributions have same shape:
  - $H_0 = \tilde{\mu}_1 = ... = \tilde{\mu}_K$
  - $H_A = \tilde{\mu}_1 \neq ... \neq \tilde{\mu}_K$

2. Test statistic

- $H = \frac{12}{N(N+1)} \sum \frac{R_i^2}{n_i} - 3(N+1)$

3. The sampling distribution

- $H \sim \chi^2_{K-1}$

4. p-value

- P$(\chi^2_{K-1} > H)$

5. and 6. Decision and Conclusion