Module 11: ANOVA - Analysis of Variance

ANOVA can be used to compare the means of several populations with continuous populations simultaneously. The population variance of the dependent variable must be equal in all groups.

Recall that

$$t = \frac{\overline{d}}{\sqrt{S_d^2/n}}$$

Which is the difference in two means over the standard error. When comparing multiple independent samples it is easier to use a pooled variance, but to do so the variances must be equal.

Equality of Variances

The equation for pool variance:

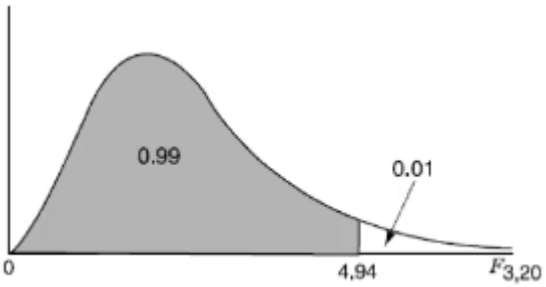
$$S_p^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}$$

Assumption for pooled variance is that variances in the two groups are equal. We can test this with H0 = $\sigma_1 = \sigma_2$ and use the F distribution which is indexed by the denominator df and the numerator df; choose the larger estimated variance to be numerator and the smaller estimated variance to be the denominator.

Test statistic:

$$\mathrm{F} = \frac{s_1^2}{s_2^2}$$

If F is greater or smaller than critical values for a given significance level the null hypothesis is rejected and we can conclude there is evidence the two population variances are not equal.



The F distribution is not symmetric, which makes it hard to look up critical values. It can be done in R: pf(F, df1, df2, lower=F)

The F test is not always appropriate as it is sensitive to departures from normality. Examine variability in the two groups by comparing sample variances using boxplots to help decide which standard error is appropriate. In the case where variances are unequal we use the same procedure but SE is estimated as:

$$SE(\overline{X}_1 - \overline{X}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Using the n-1 degrees of freedom from whichever sample is smaller as an approximate (SAS or R would figure out the exact value).

ANOVA

Terminology:

- Factor - category/grouping variable
- Level - individual group of the factor
- Balanced design - same number of individuals in each level

The general data configuration is we have k population groups each with n_k observations, which can be the same or different.

Assumptions:

- Observations are independent
- Data are random samples from k independent populations
- Within each population the dependent variable is normally distributed
- The population variance of the dependent variable is equal in all groups.

H0: The k populations means are equal

Ha : The k populations means are not all equal or at least one is not equal

Recall that variance as a function of Y is:

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

The numerator is the "Total variability" or the "Total sum of squares" (SST)

In ANOVA we split the SST into two components:

1. Variability due to differences between the groups (SS Between Groups)
2. Variability due to differences between individual y values within the groups (SS Within Group)

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{i,j} - \bar{y})^2$$

Which can also be expressed as:

$$\sum_{i=1}^{N}(y_{ij} - \bar{y})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2$$

SS Total = SS Within Groups + SS Between Groups

ANOVA TABLE

Source	Sums of Squares	df	Mean Square	F
Between Groups (Model)	SS Between	k-1	SSB / (k-1)	MSB / MSW
Within Groups (Error)	SS Within	N-k	SSW / (N-k)	
Total	SS Total	N-1		

R^2 is the proportion of variability explained by the difference between groups:

R^2 = SS Between / SS Total

Adjustment Procedures

- Tukey's adjustment is appropriate when comparing pairs of means and is among the most powerful
  - Provides exact P-values when groups are equal sizes

Means with the same letter are not significantly different.				
Tukey Grouping		Mean	N	group
	A	7.5000	6	1
	A			
B	A	6.1667	6	5
B	A			
B	A	5.1667	6	4
B	A			
B	A	5.0000	6	2
B				
B		4.3333	6	3

In the above Turkey procedure we observe group 1 and 3 are significantly different

- Scheffe's adjustment is appropriate for general contrasts
- Bonferroni's adjustment is appropriate for any situation but can be too conservative

Parametric vs Non-parametric Tests

Tests are **parametric** because they make assumptions about the distribution of the data.

Non-parametric methods make fewer and more generic assumptions about the distribution of the data. These tests are generally more friendly toward non-normal distributions and small sample sizes.

Situation	Parametric test	Nonparametric test
One sample or paired-matched	T-test	Wilcoxon Signed-Rank Test
2 groups	T-test *Equal variances *Unequal variances	Wilcoxon Rank-Sum Test
> 2 group	ANOVA	Kruskal-Wallis test

Sign Test

Simplest non-parametric test. Analyze only the signs of the differences:

Child	Before Treatment	After 1 Week of Treatment	Difference (Before-After)	Sign
1	85	75	10	+
2	70	50	20	+
3	40	50	-10	-
4	65	40	25	+
5	80	20	60	+
6	75	65	10	+
7	55	40	15	+
8	20	25	-5	-

H0: The median difference is zero (half the signs are positive and half are negative)

Ha: The median difference is not zero

Wilcoxon Signed-Rank Test

Paired-sample t-test equivalent. Uses information on the relative magnitude of the paired differences as well as their signs.

Assumption:

- Independent observations
- Continuous or ordinal observations
- Symmetric distribution

H0: The median difference is zero

Ha: The median difference is not zero

1. Rank the magnitude of the differences (ignoring the signs)
2. Attach the signs to the ranks to form signed ranks
3. Calculate the test statistic, R, which is the sum of the positive ranks.
4. $n \geq 20 \rightarrow$ normal approximation
The values of the test statistic will range from 0 to $N(N+1)/2$, with a mean value of $N(N+1)/4$

Wilcoxon Rank-Sum Test

For two independent samples

Assumption:

- Independent observations
- Continuous or ordinal observations
- If using as a test of median, samples must have the same shape

H0: Distributions of populations from which the two groups are samples are the same.

Ha: Distributions of populations from which the two groups are samples are not the same.

1. Combine the two groups into one large sample and rank the observations from smallest to largest. Tied observations are assigned a average rank to all measurements with the same value.
2. Sum the ranks of each of the original groups.
3. The Wilcoxon sum rank test statistic (R) is the sum of the ranks in the group with the smallest sum

Kruskal-Wallis Test

A non-parametric analogue to one-way ANOVA. It's an extension of the Wilcoxon rank-sum test for more than two groups. Based on ranked data. The Kruskal Wallis test will tell you if there is a significant difference between groups. However, it won’t tell you which groups are different.

Assumption:

- Use when normality assumption is violated.
- Samples drawn from population are random
- Observations are independent
- The dependent variable is at least ordinal
- All groups have the same distribution shape

H0: The population medians are equal in all groups

Ha: At least one of the groups population median is different

Step 1: Rank data without respect to group (rank the data from 1 to N ignoring group membership).
rij = rank of yij
Tie values obtain the rank of the average rank they would receive if not tied
Step 2: For each group i, compute

$$R_i = \sum_{j=1}^{n_i} r_{ij}$$

Step 3: Calculate h*

$$h^* = \frac{12}{N(N+1)} * \left(\sum_{i=1}^{k} \frac{R_i^2}{n_i} \right) - 3(\mathrm{N}+1)$$

When H0 is true, h* has an approximate chi-square distribution with k-1 degrees of freedom

When we have tied values, need to make additional adjustments to h.

$$h = \frac{h^*}{\left(1 - \frac{C}{N(N^2-1)}\right)}$$

If g is the categories of tied values, the Lth category can be described as $c_L = m * (m^2 - 1)$, where m is the number of ties values for the Lth category. The correction factor is C = c1 + c2 + ... + cg

---

Revision #5
Created 29 August 2022 14:48:07 by Elkip
Updated 30 August 2022 00:44:33 by Elkip