This time: ANOVA

Next time: (1 Dec) ANOVA, catagorical data

Read: LN pp. L-283 → L-301 This time: L-269 →

Note: (24) on Pg. 30 helps explain how accurate predictions are

Continuing the poplar tree study: 4 inde pendent samples, continuous

new notation: Y_{ij} subgroups need 2 indexes (which of the 4 groups, which individual tree) $(S_1 S_2 S_3 S_4)$

Basic model: assume $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4$ so use σ for all

4 groups have 6 possible ways to compare 2 at a time

Multiple Comparisons Problem: making more than 1 CI means more errors

interval method is better than strait p-value calculations

weighted mean = "votes" according to how many observations there are

n = total number of subjects, est. of μ (grand mean)

Measure of the extent to which $\bar{y}_i$ is NOT close to $\bar{y}$:

SS_B (between groups sum of squares) (JMP calls it "treatment" here)

$\sum_{i=1}^{I} n_i (\bar{y}_i - \bar{y})^2$ (the answer is squared)

Things that increase SS_B even if null is true:

1. As I (number of groups) increases

easy fix: divide by ~I → $\frac{1}{I-1} \sum_{i=1}^{I} n_i (\bar{y}_i - \bar{y})^2 = \frac{SS_B}{DF_B} = MS_B$ mean square between groups

degrees of freedom between

2. (worse!) changing units can make SS_B arbitrarily bigger/smaller

fix: create $\frac{\text{signal}}{\text{noise}}$ ratio ← MS_B get a unitless quantity F-ratio (Fisher)

can you put any number to get that mean?

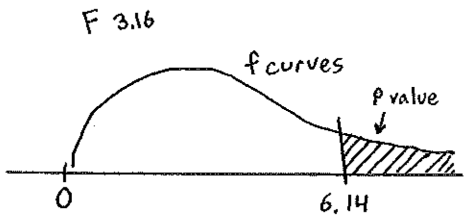
$\begin{bmatrix} 1 \\ 2 \\ 9 \end{bmatrix}$ n = 3 $\begin{bmatrix} \bigcirc ✓ \\ \bigcirc ✓ \\ \times \end{bmatrix}$ so 2 degrees of freedom!

mean $\bar{y} = 4$ mean 4

with 2 obs., there are (n-1) df for spread

$\sum_{i=1}^{I} \sum_{j=1}^{n_i}$ ← double summation called SS_W (within-group sum of squares, JMP calls it "error")

ratio is $\hat{\sigma}^2 = \frac{SS_W}{DF_W} = MS_W =$ with group mean square



Long run hist of F if null true

& all pop. hist groups follow normal curve well

$P = 0.0056 \doteq 0.6\%$ so null looks bad. diffs are statsig. (greater than 0.5%)

ANOVA Decomposition (breakdown)

SST = total sum of squares (JMP calls it $\overset{\text{corrected}}{\text{C. total}}$)

Check pp. L-272 for annotated JMP, table on R-30 to understand JMP labels in analysis of variance

Step 1: decide how many comparisons to make, K

Step 2: decide the level of confidence $100(1-\alpha)\%$

Step 3: pairwise comparisons will be the form $(\bar{y}_i - \bar{y}_j)$ & will have estimated standard error:

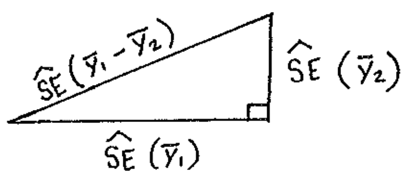
$$\widehat{SE}(\bar{y}_i - \bar{y}_j) = \hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

2 indep. samples:

group	n	$\bar{y}$	s
1	n_1	$\bar{y}_1$	s_1
2	n_2	$\bar{y}_2$	s_2

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

$$= \sqrt{\left(\frac{s_1}{\sqrt{n_1}}\right)^2 + \left(\frac{s_2}{\sqrt{n_2}}\right)^2}$$



$$SE(\bar{y}_1) = \frac{\sigma_1}{\sqrt{n_1}}$$

$$SE(\bar{y}_2) = \frac{\sigma_2}{\sqrt{n_2}}$$

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

now suppose $\sigma_1 = \sigma_2 = \sigma$

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} = \sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$= \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

root mean squared error = $\sqrt{MS_W}$ ($\hat{\sigma}^2 = MS_W$ = "pooled variance estimate")

3. How wide should variables be?

fix: use a bigger t number to <u>widen</u> all the intervals

$$\boxed{t_{n-I}^{1-\frac{\alpha}{k}}}$$

4. $\widehat{SE}$ of est. diff

$\hat{\sigma}$ = RMSE $\qquad \widehat{SE} = \hat{\sigma}\sqrt{\frac{1}{n-1} + \frac{1}{n-1}}$

5. Look to see which CI's don't include 0 - these are statsig. pairwise differences.

summarize & interpret the results

The F tables are in the notes (we won't be doing many calculations, we'll be given JMP values w/p-values already computed)

Homework #5: fecundity = reproductive sucess drosophilia = fruit flies

outcome variable: close to continuous, 3 groups, independent an ANOVA problem