Comparing Several Means

Some slides from R. Pruim

STA303/STA1002: Methods of Data Analysis II, Summer 2016

Michael Guerzhoy

The Dating World of Swordtail Fish

- In some species of swordtail fish, males develop brightly coloured swordtails
- Southern Platyfish do not
- Want to know: will female Southern Platyfish prefer males with artificial brightly-coloured swordtails?
  - If they do, that’s evidence that males in other species evolved as a result of female preference
- Experiment: multiple pair of males, one with a transparent artificial tail, one with a bright yellow artificial swordtail. Measure the percentage of time the female spends courting with the male with the yellow tail. There are 84 females in total.

Platyfish

- Eventually, we would like to know whether females spent more time with the yellow-swordtailed males. But we would like to first investigate whether there is anything else going on in the data that might affect our conclusions
- Question: Do the (means of) the quantitative variables depend on which group (given by categorical variable) the individual is in?

- (the fish, in R)

Computing Group Means with Linear Regression

- Fit a linear regression:
- $Y \sim a_0 + a_{g1} I_{g1} + a_{g2} I_{g2} + \cdots + a_{gN} I_{gN}$
- Y: the percentage of time the female spends with the yellow-tailed male
- I_{gk}: 1 if the case involves Group k, 0 otherwise
- Regression:
  - Minimize $\sum_i \left(Y_i - (a_0 + a_{g1} I_{i,g1} + a_{g2} I_{i,g2} + \cdots + a_{gN} I_{i,gN})\right)^2$

Computing Group Means with Linear Regression

- $\sum_i \left(Y_i - (a_0 + a_{g1} I_{i,g1} + a_{g2} I_{i,g2} + \cdots + a_{gN} I_{i,gN})\right)^2$
$$= \sum_{group} \sum_{i \in group} \left(Y_i - a_{group}\right)^2$$

- $\sum_{i \in group} \left(Y_i - a_{group}\right)^2$ is minimized when a_{group} =? (show how to do this)

Computing Group Means with Linear Regression

- $\left(\sum_{i\in group}\left(Y_i - a_{group}\right)^2\right)' = 0$

$$-2\sum_{i\in group}\left(Y_i - a_{group}\right) = 0$$

$$\sum_{i\in group} Y_i = \sum_{i\in group} a_{group}$$

$$\mathrm{a}_{group} = \frac{\sum_{i\in group} Y_i}{N_{group}}$$

Computing the Means with R

- (in R)

Are the Pairs Different from Each Other?

- If we had just two pairs in which we're interested, we could simply use a t-Test
  - Estimate the pooled variance from the entire dataset
    - $s_p^2 = \frac{(N_{g1}-1)s_1^2+\cdots+(N_{gN}-1)s_N^2}{(N_{g1}-1)+\cdots+(N_{gN}-1)}$ $((N_{g1}-1)+\cdots+(N_{gN}-1)$ d.f.)
    - $\frac{mean_{g1}-mean_{g2}}{s_p\sqrt{\frac{1}{N_{g1}}+\frac{1}{N_{g2}}}} \sim t((N_{g1}-1)+\cdots+(N_{gN}-1))$
- But we're interested in whether *any* pair is different from *any other pair*

ANOVA

- Null Hypothesis: the means of all the groups are equal
- Notation:
  - N: number of individuals/observation all together
  - $\bar{X}$*:* mean for entire data set is
- Group *i:*
  - N_i: number of individuals in group *i*
  - X_{ij}: value for individual *j* in group *i*
  - $\bar{X}_i$: mean for group *i*

ANOVA: Idea

- If all the group means are the same, the average variation *within* the groups should be almost as large as the average variation within the entire dataset (why *almost?*)
- Variation BETWEEN groups:
  - For each data value look at the difference between its group mean and the overall mean: $\sum_i \; N_i(\bar{X}_i - \bar{X})^2$
- Variation WITHIN groups:
  - For each data value look at the difference between the value and the group mean: $\sum_i \; \sum_j \left(X_{ij} - \bar{X}_i\right)^2$

ANOVA: Idea

- SSReg (Regression Sum of Squares, variation across groups) : $\sum_i \; N_i(\bar{X}_i - \bar{X})^2$ (d.f.: Ngroups-1)
- RSS (Residual Sum of Squares, variation within groups): $\sum_i \; \sum_j \left(X_{ij} - \bar{X}_i\right)^2$ (d.f.: Npoints-Ngroups)
- Compute the ratio of the averages:

  - $F = \frac{\sum_i \; N_i(\overline{X_i} - \bar{X})^2}{Ngroups-1} / \frac{\sum_i \; \sum_j \left(X_{ij} - \bar{X}_i\right)^2}{Npoints-Ngroups}$

ANOVA: Idea

- $F = \frac{\sum_i \ (\overline{X_i} - \bar{X})^2}{Ngroups - 1} / \frac{\sum_i \ \sum_j \left(X_{ij} - \bar{X}_i\right)^2}{Npoints - ngroups}$
- If "average" between-group variation is not larger than "average" within-group variation (i.e., the Null Hypothesis is true), $F \approx 1$
- If between-group variation is larger than within-group variation (i.e., the means for the different groups are different), $F > 1$
- $\frac{\sum_i \ N_i (\overline{X_i} - \bar{X})^2}{\sigma^2} \sim \chi^2 (Ngroups - 1)$
- $\frac{\sum_{ij} \left(X_{ij} - \overline{X_i}\right)^2}{\sigma^2} \sim \chi^2 (Npoints - Ngroups)$
- $F \sim F(Ngroups - 1, Npoints - Ngroups)$

The F distribution

- If $W_1 \sim \chi^2(k_1)$ and $W_2 \sim \chi^2(k_2)$, then

$$F = \frac{W_1}{W_2} \sim F(k_1, k2)$$

ANOVA: the model

- Constant variance σ^2, (possibly) different means μ_i for the different groups
$$X_{ij} \sim N(\mu_i, \sigma^2)$$
- Null Hypothesis: $\mu_1 = \mu_2 = \cdots = \mu_{Ngroups}$
- F statistic: $\mathrm{F} = \frac{\sum_i \; N_i(\overline{X_i} - \bar{X})^2}{Ngroups-1} \Big/ \frac{\sum_i \; \sum_j \left(X_{ij} - \bar{X}_i\right)^2}{Npoints-ngroups}$
- F-test: $P_{\mu_1 = \cdots = \mu_{Ngroups}}\,(F > f)$
  - If the Null Hypothesis is true,

$$F \sim F(Ngroups - 1, Npoints - Ngroups)$$

ANOVA table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Pair	5	938.7	187.75	0.7858	0.563
Residuals	78	18636.7	238.93		

1 less than # of groups

# of data values - # of groups

(equals df for each group added together)

ANOVA table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Pair	5	938.7	187.75	0.7858	0.563
Residuals	78	18636.7	238.93		

$$\sum_{ij}(X_{ij} - \bar{X}_i)$$

$$\sum_i N_i(\bar{X}_i - \bar{X})^2$$

ANOVA table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Pair	5	938.7	187.75	0.7858	0.563
Residuals	78	18636.7	238.93		

$$MSG = \frac{SSG}{DFG}$$
$$MSE = \frac{SSE}{DFE}$$

F = MSG / MSE

$P(F > f) \sim F(DFG < DFE)$

ANOVA table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Pair	5	938.7	187.75	0.7858	0.563
Residuals	78	18636.7	238.93		

The p-value for the F-statistic. A measure of how compatible the data is with the hypothesis that

$$\mu_1 = \cdots = \mu_{Ngroups}$$

Pairwise t-Tests

- Suppose we find (using an F-test) that there *are* differences between the different means. That still doesn't tell us what the differences are
- Naively, we can run a t-Test for every pair of groups
- (in R)

Problem with Multiple Comparisons

- If we are computing a p-value and the Null Hypothesis is true, we'd get a false positive 5% of the time (1 time out of 20)
  - False positive: p-value<.05, but the Null Hypothesis is true
- If we are computing 20 p-values and the Null Hypothesis is true, what percent of the time will we get at least one false positive?

Problem with Multiple Comparisons

- If we are computing 20 p-values and the Null Hypothesis is true, what percent of the time will we get at least one false positive?

$$1 - (1 - 0.05)^{20} \approx 64\%$$

- If we have 7 groups, and compare each mean to each other mean, how many comparisons do we make?
  - (Show in R)

Problem with Multiple Comparisons

- N variables to do pairwise comparison on:

$$\binom{N}{2} = N(N-1)/2 \text{ comparisons}$$

- Intuition:
  - See the table in R
  - For each coefficient (N) of them, compare it to every other (N-1): N(N-1) comparisons. But we compared each pair twice, so divide by two: N(N-1)/2

Bonferroni correction

- Boole's inequality: the probability of any one of the events $E_1, E_2, \dots, E_n$ happening is smaller than $\sum_i P(E_i)$:
    - $P(\cup_i E_i) \leq \sum_i P(E_i)$
    - Idea: the probability is largest when the events are mutually exclusive, in which case the probability is $\sum_i P(E_i)$
- $P\left(\cup_{i=1\dots n}^{n}\left(p_i \leq \frac{\alpha}{n}\right)\right) \leq \sum_{i=1}^{n} P\left(p_i \leq \frac{\alpha}{n}\right) = \frac{n\alpha}{n} = \alpha$

Bonferroni correction

- If we want the *familywise* p-value threshold to be α, make the individual p-value threshold be $\frac{\alpha}{n}$, where *n* is the number of groups
- Generally, *very* conservative
  - Why?

Tukey's Honest Significant Differences (HSD)

- Tukey's HSD is a method of adjusting the SE estimate based on the range of the data
  - Not as conservative as using the Bonferroni correction

Confidence Intervals -- Bonferroni

- If the statistic is t-distributed:

$$\hat{\theta} \pm t_{df,1-\frac{\alpha}{k}} \cdot SE(\hat{\theta})$$

- (In R)

Summary: F-test and Pairwise Comparisons

- Assuming (and checking) normal distributions with constant variance in different groups:
  - Run F-test to see if any of the means are different
  - Can follow up and check pairwise differences
- If you have a hypothesis about which group means are different *ahead of time*, that's like running multiple studies
  - Some of your multiple studies might be wrong, of course
  - Still, okay not to adjust as long as you report that you had lots of hypotheses about which means might be different
    - Of course, if you have lots of hypotheses, people might think you're a little bit scatterbrained