STAT W4201 001, Homework 3

Brian Weinstein (bmw2148)

Feb 17, 2016

Code is attached here and also posted at https://github.com/BrianWeinstein/advanced-data-analysis. Where relevant, code snippets and output are are included in-line.

Problem 1: Ramsey 4.30

Problem 2: Ramsey 4.32

Problem 3: Ramsey 5.19

(a) The pooled estimate of the variance s_p^2 is given by

$$\begin{aligned}
s_p^2 &= \frac{\sum_{i=1}^{I}(n_i - 1)s_i^2}{\sum_{i=1}^{I}(n_i - 1)} \\
&= \frac{(127-1)(0.4979)^2 + (44-1)(0.4235)^2 + (24-1)(0.3955)^2 + \cdots}{(127-1) + (44-1) + (24-1) + \cdots} \\
&\qquad \frac{\cdots + (41-1)(0.3183)^2 + (18-1)(0.3111)^2 + (16-1)(0.4649)^2 + \cdots}{\cdots + (41-1) + (18-1) + (16-1) + \cdots} \\
&\qquad \frac{\cdots + (11-1)(0.2963)^2 + (7-1)(0.3242)^2 + (6-1)(0.5842)^2}{\cdots + (11-1) + (7-1) + (6-1)} \\
&= 0.1919.
\end{aligned}$$

(b) To construct an ANOVA table to test for species differences, we first compute SS_W, the sum of squared residuals of the "within" group.

$$\begin{aligned}
SS_W &= \sum_{i=1}^{9}\left[\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y_i}\right)^2\right] = \sum_{i=1}^{9}\left[(n_i - 1)\mathrm{SE}(Y_i)\right] \\
&= (127-1)(0.4979) + (44-1)(0.4235) + \cdots + (6-1)(0.5842) \\
&= 122.8658.
\end{aligned}$$

Given the sample standard deviation of all 294 observations as one group is $\mathrm{SD}_{Total} = 0.4962$, the total sum of squared residuals SS_{Total} is

Should SS total be -9 (-I) instead of -1 ?

$$SS_{Total} = \sum_{i=1}^{9} \left[\sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y} \right)^2 \right] = (n_1 + n_2 + \cdots + n_9 - 1)(\text{SD}_{Total})$$
$$= (127 + 44 + \cdots + 6 - 1)(0.4962) = (293)(0.4962)$$
$$= 145.3866.$$

Therefore, the sum of squared residuals of the "between" group, SS_B is:

$$SS_B = SS_{Total} - SS_W = 145.3866 - 122.8658 = 22.5208.$$

The Total, Within, and Between degrees of freedom are $\text{df}_{Total} = (n-1) = 293$, $\text{df}_W = (n - I) = (294 - 9) = 285$, and $\text{df}_B = \text{df}_{Total} - \text{df}_W = 8$, respectively.

The Mean Square for the Within and Between groups is the [(Sum of Squares) / df], and the F-statistic is defined as the [(Between Mean Square) / (Within Mean Square)].

Thus the ANOVA table to test for species differences is:

Source of Variation	Sum of Squares	d.f.	Mean Square	F-Statistic	p-Value
Between Groups	22.5208	8	2.8151	6.5300	8.0551×10^{-8}
Within Groups	122.8658	285	0.4311		
Total	145.3866	293			

Where the p-value for the F-Statistic is given by the CDF of an F distribution with 8 and 285 degrees of freedom.

```
> # p-value of F-statistic
> pf(q=6.5300, df1=8, df2=285, lower.tail=FALSE)
[1] 8.055137e-08
```

(c) The ANOVA method for calculating SS_B yields the same answer as the formula:

$$SS_B = \sum_{i=1}^{I} \left(n_i \overline{Y_i}^2 \right) - n\overline{Y}^2 = \sum_{i=1}^{9} \left(n_i \overline{Y_i}^2 \right) - n\overline{Y}^2.$$

The overall mean $\overline{Y} = 7.4746$ is computed as a weighted average of the group means and $n = 294$, thus

17.45402 != 22.5208 : maybe my SS total calc is incorrect?

$$SS_B = \sum_{i=1}^{9} \left(n_i \overline{Y_i}^2 \right) - n\overline{Y}^2$$
$$= (127)(7.347)^2 + (44)(7.368)^2 + (24)(7.418)^2 + (41)(7.487)^2 + (18)(7.563)^2 + \cdots$$
$$\cdots + (16)(7.568)^2 + (11)(8.214)^2 + (7)(8.272)^2 + (6)(8.297)^2 - (294)(7.4746)^2$$
$$= 17.45402$$

(d) If the first 6 species have one common mean and the last 3 have another common mean, we find that

finish 3d

Problem 4:

Consider the Bumpuss data in Chapter 2, compute the power of the two-sided two sample t-test of size 0.05 (i.e., reject the null hypothesis if the absolute value the t-statistic is greater than or equal to 2), under the alternative that $\mu_x - \mu_y = \overline{x} - \overline{y} = 0.01$ and $\sigma = s_p = 0.0214$.

Problem 5:

Show that the two-sided two sample t-test is equivalent to the anova F-test, if the number of groups is two.

For $I = 2$ groups, the F-statistic is given by

$$\text{F-statistic} = \frac{SS_B / \left[(n-1) - (n-I)\right]}{SS_W/(n-I)},$$

where n_1 and n_2 are the sizes of samples 1 and 2, respectively, $n = n_1 + n_2$ is the total sample size, SS_B is the "between groups" sum of squared residuals, and SS_W is the "within groups" sum of squared residuals.

Simplifying, we find

$$\text{F-statistic} = \frac{SS_B/(I-1)}{SS_W/(n-I)} = \frac{SS_B/(2-1)}{SS_W/(n-2)} = \frac{SS_B/1}{SS_W/(n-2)}.$$

If the observations from group 1 are $\sim \mathrm{N}(\mu_1, \sigma^2)$ and the observations from group 2 are $\sim \mathrm{N}(\mu_2, \sigma^2)$, we know that

$$\text{F-statistic} \sim \mathrm{F}_{1,n-2} \text{ , which is equivalent to } t^2_{n-2}.$$

i.e., an F distribution with a numerator degrees of freedom of 1 and a denominator degrees of freedom of $n-2$ is equivalent to the square of a t distribution with $n-2$ degrees of freedom. Finish problem 5

Problem 6:

Consider $X_1, \ldots, X_{10}$ are i.i.d. N$(0, \sigma^2)$, $Y_1, \ldots, Y_{10}$ are i.i.d. N(μ, σ^2) and hypothesis testing:

$$H_0 : \mu = 0$$
$$H_A : \mu \neq 0.$$

Compute the power of a two sided two sample t-test of size 0.05 when $\sigma^2 = 1$ and μ = 0.1, 0.5, 1, and 2. Plot the power as a function of μ. Then, increase the sample size in each group to 20 and draw the power function in the same plot as that of the sample size 10.

Problem 7:

Under the setting of the previous problem, show that, under the null hypothesis, the p-value follows the uniform distribution on the interval [0, 1] and perform simulations to confirm it.

Todo list