Exercises Set 10 - Solution

1 A new post-it chemical

The calculation is the same as in last week's pizza exercise, yet now we do not have access to the individual datapoints. We only have the aggregated values of the mean and the standard deviation calculated by the laboratory. This means we do not have the full information about the data, and cannot do all analyses, for example we cannot calculate the Total SS directly as in the previous exercise.

Each of the 3 groups has 40 samples, hence $\nu_E = 117$ and $\nu_B = 2$. $SS_1 = \sum_{j=1}^{40}(X_{1,j} - \bar{X}_j)^2 = 39 * Var(X_1) = 39 * 4.5^2 = 789.8$. Analogously, we find:

$$SS_1 = 789.8, SS_2 = 1179.8, SS_3 = 1312, SS_E = 3281.5$$

The total mean is $\bar{X}_T = 19.1$. The SSB we compute directly as $SS_B = \sum_{i=1}^{3} N_{Si} * (\bar{X}_i - \bar{X}_T)^2 = 405.2$. We cannot check SST, hence we compute it by adding SSE and SSB.

This yields the following ANOVA table:

Source	ν	SS	MS	F
Group/Between	2	405.2	202.6	7.23
Error/Within	117	3281.5	28.04	
Total	119	3686.7		

We compute the $\alpha = 0.01$ percentile of the F-distribution as $qF_{2,117}(99\%) = 4.79$. The experimental F exceeds the critical one, hence we reject H_0 and state that there is significant difference between the groups. Given that the EPFL glue has the highest mean, and a comparable standard deviation, it looks like the EPFL glue is also statistically significantly the best, but we would have to show that separately.

Some more info on how the F-distribution works. On the level of the sum of squares, the group SSB is much smaller than the error SSE. So superficially, it may seem that the model does not explain much of the variance. However, the large numbers $N_{Si} = 40$ pull the MSE down, such that the total variance *per datapoint* is much lower within the groups than between them.

2 Is size a good predicator of weight? (a linear regression)

We want to find a linear function to express the weight with respect to the height. The linear regression should minimize the sum of square error SSE between the actual weight (w_i), and $\hat{w}_i = \hat{a} + \hat{b}h_i$, the

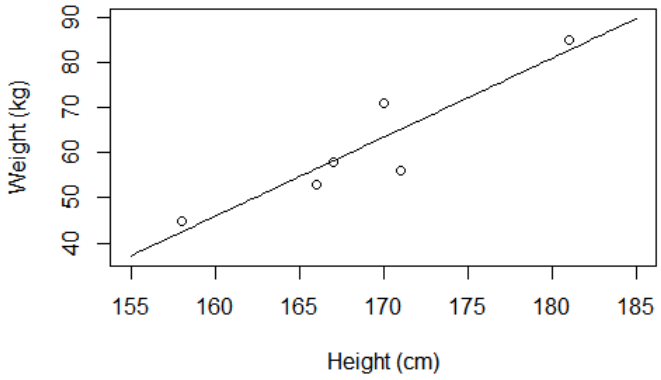
predicted one. The two model's parameters a and b are estimated with $\hat{a}$ and $\hat{b}$.

$$\begin{aligned} SSE &= \sum_{i=1}^{n}(w_i - \hat{b}h_i - \hat{a})^2 \\ \frac{\partial SSE}{\partial \hat{b}} &= -2 \cdot \sum_{i=1}^{n}(w_i - \hat{b}h_i - \hat{a})h_i = -2 \cdot \left(\sum_{i=1}^{n} w_i h_i - \hat{b} \cdot \sum_{i=1}^{n} h_i^2 - n\hat{a}\bar{h}\right) = 0 \\ \frac{\partial SSE}{\partial \hat{a}} &= -2 \cdot \sum_{i=1}^{n}(w_i - \hat{b}h_i - \hat{a}) = -2 \cdot (n\bar{w} - n\hat{b}\bar{h} - n\hat{a}) = 0 \end{aligned}$$

This is exactly the same equation as derived in the lecture. The parameters are obtained after solving these equations:

$$\hat{b} = \frac{n\bar{h}\bar{w} - \sum_{i=1}^{n} w_i h_i}{n\bar{h}^2 - \sum_{i=1}^{n} h_i^2} = 1.748 \qquad \hat{a} = \bar{w} - \hat{b}\bar{h} = -233.75$$

The linear regression $\hat{w} = f(h) = \hat{a} + \hat{b}h$ is shown in the figure below.



The ANOVA table is:

	Sum of Squares	Degree of Freedom ν	Mean Squares
Model	$SSM = \sum_i(\hat{w}_i - \bar{w})^2 = 837.0$	1	$MSM = SSM/1 = 837.0$
Error	$SSE = \sum_i(\hat{w}_i - w_i)^2 = 165.3$	$n - 2 = 4$	$MSE = SSE/4 = 41.33$
Total	$SST = \sum_i(w_i - \bar{w})^2 = 1029.3$	$n - 1 = 5$	-

The Fisher coefficient is $F = MSM/MSE = 20.9$. Since this is bigger than $qF_{1,4}(95\%) = 7.709$, we have to reject the hypothesis that $b = 0$, so there is a relation between the height and the weight.

The error variance is $\hat{\sigma}^2 = MSE = 41.33$.

The regression coefficient $R^2 = SSM/SST = 0.813$. The closer this coefficient is to 1, the better are the points fall onto the regression line.

4 A new milkshake recipe

a) [1 pt] $\bar{X}_1 = 6.77$, $s_1^2 = \frac{1}{5}\sum_{i=1}^{6}(X_i - \bar{X})^2 = \frac{1}{5}\sum_{i=1}^{6} X_i^2 - 2\bar{X}X_i + \bar{X}^2 = 6.56$, so $s_1 = 2.56$.

b) [1 pt] H_0: The population means of group 1 and 3 are the same. $H_0 : \mu_1 = \mu_2$. All differences are statistical fluctuations.

c) [2 pt] Welch's T-statistic is computed as $T_{1,3} = \frac{\bar{X}_1 - \bar{X}_3}{\sqrt{\frac{s_1^2+s_3^2}{6}}} = 2.56$.

d) [5 pt] To compute the degrees of freedom, df, we need to compute $a_{13} = (\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2 = 3.90$ and $b = \frac{1}{n-1}((\frac{s_1^2}{n_1})^2 + (\frac{s_2^2}{n_2})^2) = 0.40$. Then $df = round(a/b) = round(9.87) \approx 10$. Looking up in the quantile table, $qt(p = 0.975, df = 10) = 2.228$. As $T_{1,3} > qt(p = 0.975, df = 10)$, we have statistically significant evidence that your shake is better than Mueller's.

a12=4.66, a13=3.90, a23=3.72, b12=0.47, b13=0.40, b23=0.38, all $\nu \sim 10$.

All other T-statistics are in the acceptance region, so we have no statistically significant evidence that our shake is different from Emmis.

e) [1 pt] As we have seen in the lecture, the 1-factor/n-level ANOVA model is given by $\alpha_i = \bar{X}_i - \bar{X}$. We need to compute the global average, $\bar{X} = \frac{1}{3}(\bar{X}_1 + \bar{X}_2 + \bar{X}_3) = 5.28$. Thus,
$\alpha_1 = 1.48$, $\alpha_2 = 0.63$, $\alpha_3 = -2.12$.

f) [5 pt] The 1-factor/3-levels ANOVA table describes this problem:

	df	SS	MS	F
Model	2	$6 * \sum_{i=1}^{3} \alpha_i^2 = 42.49$	21.25	F=MSB/MSE=3.51
Error	15	$\sum_{i=1}^{3}\sum_{j=1}^{6}(X_{i,j} - \mu - \alpha_i)^2 = 90.92$	6.06	
Total	17	$\sum_{i=1}(X_{i,j} - \bar{X})^2 = 133.41$		

g) [2 pt] The F-statistic of this model is 3.51. From the provided table, we find $qF_{2,15}(p = 0.95) = 3.682$. The critical region is $K_c = [3.682, \infty]$, therefore we have to accept the null hypothesis. We cannot find significant evidence for an effect.