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 $\nu_B = 2$. $SS_1 = \sum_{j=1}^{40}(x_{1,j} - \bar{x}_1)^2 = 39 * Var(X_1) = 39 * 4.5^2 = 789.8$. 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 SS_B we compute directly as $SS_B = \sum_{i=1}^{3} n_i * (\bar{x}_i - \bar{x}_T)^2 = 405.2$. We cannot check SS_T, hence we compute it by adding SS_E and SS_B.

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 SS_B is much smaller than the error SS_E. So superficially, it may seem that the model does not explain much of the variance. However, the large numbers $n_i = 40$ pull the MS_E 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 SS_E between the actual weight (w_i), and $\hat{w}_i = \hat{a} + \hat{b}h_i$, the

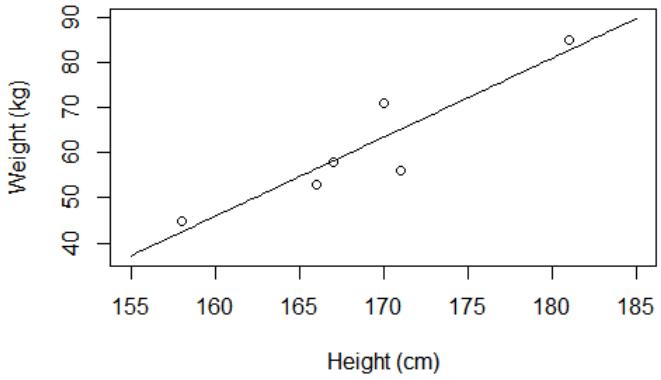
predicted one. The two model's parameters a and b are estimated with $\hat{a}$ and $\hat{b}$.

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

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	F
Model	$SS_M = \sum_i(\hat{w}_i - \bar{w})^2 = 864.0$	1	$MS_M = SS_M/1 = 864$	20.9
Error	$SS_E = \sum_i(\hat{w}_i - w_i)^2 = 165.3$	$n - 2 = 4$	$MS_E = SS_E/4 = 41.33$	
Total	$SS_T = \sum_i(w_i - \bar{w})^2 = 1029.3$	$n - 1 = 5$	-	

The Fisher coefficient is $F = MS_M/MS_E = 20.9$. 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 $\hat{\sigma}^2 = MS_E = 41.33$.

The regression coefficient $R^2 = SS_M/SS_T = 0.84$. The closer this coefficient is to 1, the better are the points fall onto the regression line.

3 Exam question from 2024

The correction for the exam question is given in the following pages as was used to correct last year's exams.

a) $P(A \cap \bar{B}) = P(A) \cdot P(\bar{B} \mid A)$

$= P(A) \cdot [1 - P(B \mid A)]$

$= 0.2 \cdot [1 - 0.6]$

$= 0.2 \cdot 0.4$

$= 0.08 = 8\%$

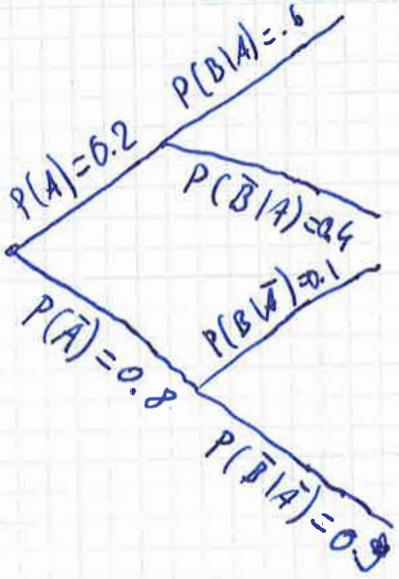
1b) $P(A) = 0.2$ $P(B|A) = 0.6$ ~~$P(\bar{A}) = 0.8$~~ $P(\bar{A} \cap \bar{B}) = 0.72$

$P(\bar{B}|A) = 0.4$

$P(\bar{A} \cap \bar{B}) = P(\bar{A}) \cdot P(\bar{B}|\bar{A})$

$\Rightarrow P(\bar{B}|\bar{A}) = \frac{P(\bar{A} \cap \bar{B})}{P(\bar{A})} = \frac{0.72}{0.8} = 0.9$ } 1 Pt.

~~$P(\bar{B}|\bar{A})$~~ $P(B|\bar{A}) = 1 - P(\bar{B}|\bar{A}) = 0.1$



1c) $$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|\bar{A}) \cdot P(\bar{A})}$$

$$= \frac{0.6 \cdot 0.2}{0.6 \cdot 0.2 + 0.1 \cdot 0.8} = \frac{0.12}{0.12 + 0.08}$$

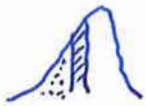
$$= \frac{0.12}{0.2} = \underline{\underline{0.6}} = 60\%$$

$\mu = 150 \quad \sigma = 10$

1 d) $P(R < 160) = \Phi\left(\frac{160 - \mu}{\sigma}\right) = \Phi\left(\frac{160 - 150}{10}\right) = \Phi(1)$

$\rightsquigarrow$ z-table $\Phi(1) = 0.841 \approx \underline{\underline{84\%}}$

1) e) $P(130 \leq R < 140) = P(R < 140) - P(R < 130)$



$= \Phi\left(\frac{140-150}{10}\right) - \Phi\left(\frac{130-150}{10}\right) =$

$= \Phi(-1) - \Phi(-2)$

$= [1 - \Phi(1)] - [1 - \Phi(2)]$

$= [1 - 0.841] - [1 - 0.9772]$

$= 0.159 - 0.0228$

$= 0.1362$

≈ 0.14 or $\approx 14\%$ or $= 13.62\%$

$N_S = 25$

1f) $P(\bar{R} < 155) = \Phi\left(\frac{155 - \mu}{\sigma/\sqrt{N_S}}\right)$

$= \Phi\left(\frac{155 - 150}{10/\sqrt{25}}\right)$

$= \Phi\left(\frac{5}{2}\right) = \Phi(2.5)$

$= 0.9938 \approx 0.99$

$\approx 99\%$

Opts if they did not use $\sigma/\sqrt{N_S}$

$$\bar{x} = \frac{\sum_i x_i}{\sum_i 1} = \frac{\sum_i x_i}{N_S}$$

1 g) $\bar{x} = \frac{21 + 19 + 22 + 26}{4} = \frac{88}{4} = \underline{\underline{22}}$ } 0.5 pt.

$$s^2 = \frac{1}{N_S - 1} \sum_i (x_i - \bar{x})^2$$

$$= \frac{1}{4-1} \left((-1)^2 + (-3)^2 + (0)^2 + (4)^2 \right)$$

$$= \frac{1}{2} \left(1 + 9 + 0 + 16 \right) = \underline{\underline{\frac{26}{3}}} \approx 8.7$$

or 8.67

if calculating s (std.) instead of s^2 (variance): lose 0.5 pts.

12) A one-sided t-test. (because σ unknown & small sample & comparing to given value).

one-sided t-test

↓

if not mentioned but done correctly in 11) -> still get the half point.

$\nu = N_s - 1 = 4 - 1 = 3$ (degrees of freedom, dof) 0.5 pt

1c)

$$t = \frac{\bar{x} - 20}{\sqrt{s^2/N_s}} = \frac{22 - 20}{\sqrt{\frac{26}{3}/4}} = \frac{2}{\sqrt{2.16}} = 1.35873\ldots$$

2 pt

either: $q\,t_{99\%}(\nu = 3) = 4.541$ 0.5 pt

as $t \leq q\,t_{99\%}(\nu = 3)$

1 pt (one statement is enough)

→ Cannot reject null hypothesis (the mean is larger than 20K only because of random fluctuations) ⟸ Cannot prove "the mean is large than 20K at 99% confidence".