A behavioural biologist believes that performance of a laboratory rat on an intelligence test depends, to a large extent, on the amount of protein in the rat’s daily diet. To check out the theory, he accumulated the following data after working with 10 rats.

<table>
<thead>
<tr>
<th>Rat</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of units of protein daily(x)</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>15</td>
<td>18</td>
<td>20</td>
</tr>
<tr>
<td>Score on standard test(y)</td>
<td>20</td>
<td>19</td>
<td>23</td>
<td>25</td>
<td>23</td>
<td>25</td>
<td>26</td>
<td>24</td>
<td>28</td>
<td>26</td>
</tr>
</tbody>
</table>

\[
\sum x_i^2 = 1608, \quad \sum y_i^2 = 5781, \quad \sum x_i y_i = 2950.
\]

(a) Calculate the sample correlation coefficient \(r\) and comment on the result.
(b) Do the data provide sufficient evidence to indicate that the amount of protein and the score on standard test are dependent?

**Solution:**

We have that \(n = 10\) and

\[
\bar{x} = 12, \quad \bar{y} = 23.9
\]

\[
S_{xx} = 1608 - 10 \times 12^2 = 168
\]

\[
S_{yy} = 5781 - 10 \times 23.9^2 = 68.9
\]

\[
S_{xy} = 2950 - 10 \times 12 \times 23.9 = 82.
\]

(a) The sample correlation coefficient \(r\) is given by

\[
r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = 0.7621661.
\]

This yields that \(r^2 = 0.581\) of the variability in the score (y values) results from the fitted line

\[
\hat{y} = \hat{\alpha} + \hat{\beta} x = 18.04286 + 0.4880952 x.
\]

(b) The hypothesis to be tested is

\[
H : \rho = 0 \quad vs \quad H_A : \rho \neq 0.
\]

A observed value of the test statistic is given by

\[
t_0 = r \times \frac{\sqrt{n-2}}{\sqrt{1-r^2}} = 3.329924.
\]
The corresponding $p$-value is

$$p = 2P(t_8 \geq 3.329924) = 0.01038558.$$ 

Conclusion: we have strong evidence that the amount of protein and the score on standard test are dependent.

2. Portable personal computers, sometimes called laptops, represent a fast-growing segment of the PC market. According to the Market Intelligence Research Company, the use of portable laptop computers can be classified into the following user segments.

<table>
<thead>
<tr>
<th>User Segment</th>
<th>1988 Percentages</th>
<th>Frequency</th>
<th>Current Survey</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business-Professional</td>
<td>69%</td>
<td>102</td>
<td></td>
</tr>
<tr>
<td>Goverment</td>
<td>21%</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>7%</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Home</td>
<td>3%</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>100%</td>
<td>150</td>
<td></td>
</tr>
</tbody>
</table>

Do the data provide sufficient evidence to indicate that the figures obtained in the 1998 study are not accurate today?

**Solution:** The hypothesis to be tested is

$$H : p_1 = 0.69, \ p_2 = 0.21, \ p_3 = 0.07 \ p_4 = 0.03 \ \text{vs} \ \ H_A : \ \text{At least one of the equalities does not hold.}$$

Calculations for the $\chi^2$-test are summarized in the following table.

<table>
<thead>
<tr>
<th>User Segment</th>
<th>Freq. (Obs. Freq.)</th>
<th>Prob. under $H(p_{i0})$</th>
<th>Exp. Freq.($np_{i0}$)</th>
<th>$(n-np_{i0})^2$</th>
<th>$\frac{(n-np_{i0})^2}{np_{i0}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business-Prof.</td>
<td>102</td>
<td>.69</td>
<td>103.5</td>
<td>0.0217</td>
<td></td>
</tr>
<tr>
<td>Goverment</td>
<td>32</td>
<td>.21</td>
<td>31.5</td>
<td>0.0079</td>
<td></td>
</tr>
<tr>
<td>Education</td>
<td>12</td>
<td>.07</td>
<td>10.5</td>
<td>0.214</td>
<td></td>
</tr>
<tr>
<td>Home</td>
<td>4</td>
<td>.03</td>
<td>4.5</td>
<td>0.056</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>150</td>
<td>1</td>
<td>150</td>
<td>0.2995</td>
<td></td>
</tr>
</tbody>
</table>

The corresponding $p$-value with $k - 1 = 3$ is

$$p = P(\chi^2_3 \geq 0.2995) = 0.96.$$ 

**Conclusion:** The data is consistent with the null hypothesis, that is, we cannot conclude that the data given are inaccurate today.
3. The personnel manager of a consumer product company asked a random sample of employees how they felt about the work they were doing. The following table gives a breakdown of their responses by gender. Do the data provide sufficient evidence to conclude that the level of job satisfaction is related to gender? Use $\alpha = 0.10$.

<table>
<thead>
<tr>
<th>Response</th>
<th>Gender</th>
<th>Very interesting</th>
<th>Fairly interesting</th>
<th>Not interesting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>70</td>
<td>41</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>35</td>
<td>34</td>
<td>11</td>
<td></td>
</tr>
</tbody>
</table>

**Solution:** The table is

\[
\begin{array}{cccc|c}
  i = 1 & n_{ij} & n_i \times n_j & \frac{(n_{ij} - n_i \times n_j)^2}{n_i \times n_j} & n_i \\
  j = 1 & 70 & 63 & 0.778 & 120 \\
  j = 2 & 41 & 45 & 0.356 & 12 \\
  j = 3 & 9 & 12 & 0.750 & 12 \\
  \hline
  i = 2 & n_{ij} & n_i \times n_j & \frac{(n_{ij} - n_i \times n_j)^2}{n_i \times n_j} & n_j \\
  j = 1 & 35 & 42 & 1.167 & 80 \\
  j = 2 & 34 & 30 & 0.533 & 12 \\
  j = 3 & 11 & 8 & 1.125 & 12 \\
  \hline
  n & n_i \times n_j & \frac{(n_{ij} - n_i \times n_j)^2}{n_i \times n_j} & \sum_{i,j} \frac{(n_{ij} - n_i \times n_j)^2}{n_i \times n_j} & n \\
  105 & 75 & 20 & 200 \\
  75 & 30 & 8 & 80 \\
  20 & 20 & 2 & 20 \\
  \hline
  \sum_{i,j} & & & 4.708 & 200 \\
\end{array}
\]

The 10% significant level Chi-square GOF test for independence is

1. **Hypothesis:** $H_0$: $p_{ij} = p_i \cdot p_j$ vs $H_1$: At least one equality does not hold.

2. **Test statistic:** $\chi^2_0 = \sum_{i,j} \frac{(n_{ij} - n_i \times n_j/n)^2}{n_i \times n_j/n} = 4.7083$

3. **Rejection region:** $\chi^2_1 > \chi^2_{k-1, \alpha} = \chi^2_{2, 0.1} = 4.605$

4. **P-value:** $P(\chi^2 > 4.7083) = 0.095 < 0.1$

5. **Conclusion:** Reject $H_0$ and conclude that there is sufficient evidence against the null hypothesis at $\alpha = 0.05$. 

3
1. Consider the data frame `fuel.frame`, which has information on makes of cars taken from the April 1990 issue of Consumer Reports. This practice continues the discussion in week 11 about the relationship between weight and fuel recorded on the `fuel.frame`.

(a) Create two vectors \( x \) and \( y \) whose elements correspond to Weight (in pounds) and Fuel (in gallons per 100 miles) on the data `fuel.frame` respectively.

(b) Calculate \( S_{xx} \), \( S_{yy} \) and \( S_{xy} \), and store your results as \( S_{xx} \), \( S_{yy} \) and \( S_{xy} \) respectively.

(c) Test whether the weight has an influence on the fuel using the F-test.

(d) Check your calculations for the regression ANOVA table using `summary`.

(e) Find the regression line for Fuel on Weight using `lsfit`, and then find an estimate of average fuel in gallons per 100 miles corresponding to cars with weight 2500 pounds and a 90% confidence interval for this estimate.

(f) Obtain a 90% confidence interval for the prediction of the fuel in gallons per 100 miles of a car with weight 2500 pounds.

(g) Calculate the sample correlation coefficient \( r \) between weight and fuel, comment on the result and check if \( r \) is significant using `cor.test`.

(h) Check if \( x \) follows normal distribution.

1. Plot a histogram of \( x \) with 9 classes using `hist(x,nclass=9)`.
2. Output summary of \( x \), divide the range of \( x \) into 9 equal intervals, calculate their width and set a vector of cut-off points `Inter`.
3. Count the frequency of \( x \) in each interval into the vector say, `freqx` and check its sum to be 60.
4. Combine interval with frequency less than 5. The revised `Inter` and `freqx` may be called `I` and `freq` respectively.
5. Calculate the vector of expected probabilities `pr` and hence the chi-square test statistics and \( p \)-value from the vector `ch`.

Solution:

\[
\text{# (a)} \text{ }
\]
\[
x<- fuel.frame[,1] \text{       # give the Weight}
y<- fuel.frame[,4] \text{       # give the Fuel}
\]

\[
\text{# (b)} \text{ }
\text{# calculate Sxx, Syy and Sxy}
\]
\[
> c(mean(x),mean(y))
\]
> Sxx<-sum((x-mean(x))*(x-mean(x)))
> Sxy<-sum((x-mean(x))*(y-mean(y)))
> Syy<-sum((y-mean(y))*(y-mean(y)))
> c(Sxx, Sxy, Syy)
[1] 1.450711e+07 1.909687e+04 3.385687e+01

# (c)

# The hypothesis to be test is
# \beta=0 vs \beta\not= 0

RegSS<-Sxy^2/Sxx  # give the Regression SS
n<-length(x)  # give the sample size
ResSS < -(Syy -Sxy^2/Sxx)/(n-2)  # give the Residual SS

F0<- RegSS/ResSS  # give a observed value of the F-statistic
p<-1-pf(F0, 1, n-2)  # give the p-value
> p
[1] 0

# Conclusion: There are very strong evidence that
# the weight of a car has an influence on the fuel.

# (d)

> summary(aov(y~x))

Df Sum of Sq Mean Sq  F Value Pr(F)
  x    1 25.13875 25.13875 167.2433 0
Residuals 58 8.71812 0.15031

# The result is the same as in (c)

# (e)
```r
> lsfit(x,y)$coef
   Intercept        X
  0.3914324  0.00131638

  # The regression line is given by
  y = 0.3914324 + 0.00131638 x.

  # An estimate of average fuel:
  > haty <- 0.3914324 + 0.00131638 * 2500
  > haty
  [1] 3.682382

  # To get the confidence interval, we first obtain t_{0.05} and s
  t0 <- qt(0.95, n-2) # Recall n is the sample size
  s <- sqrt(ResSS)

  # then get the left and right bounds (with x0=2500)
  > bleft <- haty - t0 * s * sqrt(1/n + (2500 - mean(x))^2/Sxx)
  > bleft
  [1] 3.574442

  > bright <- haty + t0 * s * sqrt(1/n + (2500 - mean(x))^2/Sxx)
  > bright
  [1] 3.790323

  # Now we get the 90% confidence interval
  [3.574442, 3.790323]

  # (f)
  # The left and right bounds of fuel
  # for a special car with weight x0=2500

  > bleft1 <- haty - t0 * s * sqrt(1 + 1/n + (2500 - mean(x))^2/Sxx)
  > bleft1
  [1] 3.025391

  > bright1 <- haty + t0 * s * sqrt(1 + 1/n + (2500 - mean(x))^2/Sxx)
  > bright1
  [1] 4.339374

  # Now we get the 90% confidence interval
  [3.025391, 4.339374]
```
# (g)

r<-Sxy/sqrt(Sxx*Syy)

> r
[1] 0.8616848

> cor.test(x,y,alt="two.sided")

Pearson's product-moment correlation
data: x and y
t = 12.9323, df = 58, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.7779857 0.9153320
sample estimates:
cor
0.8616848

#This implies that r^2= 0.7425006 of the variability in the fuel (y values)
# results from the fitted line
# y=0.3914324+ 0.00131638 x.
#The test for the relationship between x and y is significant.

# (h) 1
> par(mfrow = c(1, 1))
> hist(x, nclass = 9)

# (h) 2
> summary(x)
Min. 1st Qu. Median Mean 3rd Qu. Max.
1845 2571 2885 2901 3231 3855

> mi = min(x)
> ma = max(x)
> nclass = 9
> width = (ma - mi)/nclass
> width
[1] 223.3333

> inter = seq(mi, ma, width)
> inter
[1] 1845.000 2068.333 2291.667 2515.000 2738.333 2961.667 3185.000 3408.333
[9] 3631.667 3855.000

# (h) 3
> freqx = c()
> freqx[1] = length(x[x <= mi + width])
> for (t in 2:nclass) {
+ freqx[t] = length(x[x <= mi + t * width]) - length(x[x <=
+ mi + (t - 1) * width])

7
> freqx
[1]  2  5  7  8 13  6  8  5  6
> sum(freqx)
[1]  60

# (h) 4
> I = inter[-2]
> I
[1]  1845.000 2291.667 2515.000 2738.333 2961.667 3185.000 3408.333 3631.667
[9] 3855.000
> nclass1 = nclass - 1
> freq = c()
> for (i in 2:nclass1) {
+    freq[i] = freqx[i + 1]
+ }
> freq
[1]  7  7  8 13  6  8  5  6

# (h) 5
> mu = mean(x)
> mu
[1] 2900.833
> si = sqrt(var(x))
> si
[1] 495.8661

> pr = c()
> for (i in 1:nclass1) {
+    pr[i] = pnorm((I[i + 1] - mu)/si) - pnorm((I[i] - mu)/si)
+ }
> pr
[1] 0.09301557 0.10862358 0.15331061 0.17725445 0.16788168 0.13025340 0.08278408
[8] 0.04309861
> spr = sum(pr)
> spr
[1] 0.956222
# Note that the sum(pr) is not 1 as it considers only the probabilities over the range of x.
> ch = c()
> for (i in 1:nclass1) {
+    ch[i] = (freq[i] - n * pr[i])^2/(n * pr[i])
+ }
> ch
[1] 0.3608261178 0.0357332247 0.1561893458 0.13025340 0.08278408 0.04309861 1.6468465263
> chi2 = sum(ch)
> chi2
[1] 7.237458
> pvalue = 1 - pchisq(sum(ch), nclass1 - 1 - 2)
> pvalue
[1] 0.2035708
# Since the p-value is 0.204, we conclude that x follows normal distribution.