Contents

1 Test 20.12.2011
2 Test 19.01.2012
3 Test 09.02.2012
4 Test 18.06.2012
5 Test 20.07.2012
6 Test 07.09.2012
8 Test 22.01.2013
9 Test 19.02.2013
10 Test 17.06.2013
11 Test 18.07.2013
12 Test 11.09.2013
14 Test 21.01.2014
15 Test 12.02.2014
16 Test 12.06.2014
17 Test 24.07.2014
18 Test 12.09.2014
19 Test 19.12.2014
20 Test 22.01.2015
21 Test 12.02.2015
22 Test 24.07.2015
23 Test 11.09.2015
24 Test 22.12.2015
25 Test 27.01.2016
26 Test 26.02.2016
27 Test 13.06.2016
28 Test 19.07.2016
29 Test 06.09.2016
31 Test 24.01.2017
32 Test 15.02.2017
33 Test 16.06.2017
34 Test 25.07.2017

July 26, 2017
1 Test 20.12.2011

Text

Framework: A research group, currently active in a Brazilian area named Maranhão, is working on behalf of an International Non-Governmental Organization (NGO).

Exercise 1. The research group aims at evaluating family expenses for food and, in particular, how they are related to the number of components of the family. The following table summarizes some statistics computed on a sample of 62 families (Legend: $E = \ln(average\ daily\ expense\ for\ food)$, where expenses are in $; C = number\ of\ components$; the * indicates standard deviations and covariances computed by using the sample size at the denominator).

<table>
<thead>
<tr>
<th>$\bar{E}$</th>
<th>$\bar{C}$</th>
<th>$s_E^*$</th>
<th>$s_C^*$</th>
<th>$s_{EC}^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.15</td>
<td>6.3</td>
<td>3.34</td>
<td>0.77</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Formulate a sensible statistical model and answer the following questions:

(A) Estimate all parameters of the model with OLS.

(B) Does the size of the family, in terms of number of components, influence significantly the average daily expense for food? Answer the question by using the p-value and commenting the result (in making computations, approximate the distribution of the test statistic with the Normal distribution).

(C) (No for students with 6 CFU) Compute the 98% confidence interval for the standard deviation of the error component.

(D) Compute fitted value and residual for a family with an average daily expense for food of 17.09$ and 6 components.

Exercise 2. An educational program, directed to young mothers, aims at increasing their working abilities. Being currently in the exploratory phase, the educational program involves a small sample of mothers, whose average hourly wages (in $) has been evaluated before and after the program. The following table summarize the data.

<table>
<thead>
<tr>
<th>Ana</th>
<th>Beatriz</th>
<th>Camila</th>
<th>Débora</th>
<th>Erika</th>
<th>Fernanda</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before</td>
<td>1.61</td>
<td>1.78</td>
<td>1.35</td>
<td>1.73</td>
<td>1.4</td>
</tr>
<tr>
<td>After</td>
<td>1.64</td>
<td>1.98</td>
<td>1.81</td>
<td>1.64</td>
<td>1.75</td>
</tr>
</tbody>
</table>

(A) (No for students with 3 CFU) Is the educational program effective in increasing the mean of the hourly wages? ($\alpha = 0.1$).

(B) (No for students with 6 CFU) The NGO aims at extending the program to a larger sample. How many mothers we would recruit in order to get a confidence interval, regarding the mean difference between the wages (before and after), with size 0.06 at a 0.9 confidence level? In answering, exploit the information coming from the pilot survey, if needed.

(C) (Only for students with 6 CFU) Is the standard deviation of the difference between the wages, before and after, significantly different from 0.24$? ($\alpha = 0.1$)

Exercise 3. (No for students with 3 CFU) The research group is located near the local hospital and sometimes the researchers collaborate with its staff. The patients entering the hospital’s First Aid are classified as affected by infectious disease symptoms or by other kind of symptoms. Assume that: the number of patients arriving at the First Aid every 5 minutes is Poisson distributed, with mean 1.4 for the infections disease symptoms group and 2.2 for the other one; the correlation between the two Poisson r.v.’s is 0.35.

(A) Compute the probability that, during 5 minutes, more than 2 patients with infectious disease symptoms arrive at the First Aid.

(B) Compute the mean and the standard deviation of the total number of patients arriving, every 5 minutes, at the First Aid.
Solution

Exercise 1. Simple Linear Regression Model $y = \beta_0 + \beta_1 x + u$ with the classical assumptions on $u.$
(A) $\hat{\beta}_1 = 1.1806, \hat{\beta}_0 = -5.288, \hat{\sigma}^2 = 10.6735.$
(B) Test for $H_0: \beta_1 = 0$ against $H_1: \beta_1 \neq 0;$ test statistic: $(\hat{\beta}_1 - \beta_1)/se(\hat{\beta}_1)$ with distribution $T(n-2)$ approximated with a $N(0,1)$ because of the exercise indications; sample value of the test statistic (under $H_0$): $t_{camp} = 2.191; \text{p-value} = 2P[(\hat{\beta}_1 - 0)/se(\hat{\beta}_1) > |t_{camp}| |H_0| \approx 2P[Z > 2.191 | H_0] = 2 \times 0.014225 = 0.02845$ that needs to be compared with the typical values of $\alpha$ (as a reference, p-value = 0.0323 if one uses the $T(n-2 = 60)$ distribution).
(C) Pivot for $\sigma^2$: $\hat{\sigma}^2(n-2)/\sigma^2 \sim \chi^2(n-2);$ interval at $(1 - \alpha) = 0.98$ for $\sigma^2$: $[\hat{\sigma}^2(n-2)/c_2, \hat{\sigma}^2(n-2)/c_1] = [7.2461, 17.0844]$, where $c_1 = 37.4849$ and $c_2 = 88.3794;$ corresponding interval for $\sigma$: $[2.6919, 4.1333].$
(D) $x_i = 6$ implies $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = 1.7958; y_i = \ln(17.09) = 2.8385$ implies $u_i = y_i - \hat{y}_i = 1.0427.$

Useful formulas and values: $n = 62, \bar{x} = 6.3, \bar{y} = 2.15, \text{dev}(\bar{x}) = 36.7598, \text{dev}(\bar{y}) = 691.6472, \text{codev}(\bar{x}, \bar{y}) = 43.4, \hat{\beta}_1 = \text{codev}(\bar{x}, \bar{y})/\text{dev}(\bar{x}), \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \hat{\sigma}^2 = \text{dev}(\text{RES})/(n-2) = \left[\text{dev}(\bar{y}) - \hat{\beta}_1^2 \text{dev}(\bar{x})\right]/(n-2)$, $se(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2/\text{dev}(\bar{x})} = \sqrt{0.2904} = 0.5388,$ $t_{camp} = (\hat{\beta}_1 - 0)/se(\hat{\beta}_1).$

Exercise 2. Situation of paired data: $X =$’wages before’, $Y =$’wages after’, $D = Y - X \sim N(\mu_D, \sigma_D^2).$

<table>
<thead>
<tr>
<th>$x_i$</th>
<th>Ana</th>
<th>Beatriz</th>
<th>Camila</th>
<th>Débora</th>
<th>Erika</th>
<th>Fernanda</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_i$</td>
<td>1.61</td>
<td>1.78</td>
<td>1.35</td>
<td>1.73</td>
<td>1.4</td>
<td>1.36</td>
<td></td>
</tr>
<tr>
<td>$d_i$</td>
<td>0.03</td>
<td>0.2</td>
<td>0.46</td>
<td>-0.09</td>
<td>0.35</td>
<td>0.38</td>
<td>1.33</td>
</tr>
<tr>
<td>$d_i^2$</td>
<td>0.04</td>
<td>0.04</td>
<td>0.21</td>
<td>0.09</td>
<td>0.35</td>
<td>0.38</td>
<td>0.5275</td>
</tr>
</tbody>
</table>

(A) Test for $H_0: \mu_d = 0$ against $H_1: \mu_d > 0;$ test statistic: $(\bar{D} - \mu_d)/se(\bar{D}) \sim T(n-1);$ rejection region for $\alpha = 0.1$: $R = (t_{crit} = 1.6759, \infty);$ sample value of the test statistic (under $H_0$): $t_{camp} = 2.517.$
(B) $n = (2\sigma_D^2/A)^2 = (2 \times 1.6449 \times 0.2157/0.06)^2 = 139.9 \approx 140$ ($\alpha = 0.1$).
(C) Test for $H_0: \sigma_D = 0.24$ against $H_1: \sigma_D \neq 0.24;$ test statistic: $S_D^2(n-1)/\sigma_D^2 \sim \chi^2(n-1);$ acceptance region for $\alpha = 0.1$: $A = (c_1 = 1.1455, c_2 = 11.0705);$ sample value of the test statistic (under $H_0$): $c_{camp} = 4.0936.$

Useful formulas and values: $\bar{d} = \sum_{i=1}^{n} d_i/n = 0.2217, s_D^2 = \left[\sum_{i=1}^{n} d_i^2 - n\bar{d}^2\right]/(n-1) = 0.0465, s_D = 0.2157, se(\bar{d}) = s_D/\sqrt{n} = 0.0881.$

Exercise 3. $X_1 =$’number of patients with infections disease symptoms’ $\sim Po(\lambda_1 = 1.4);$ $X_2 =$’number of patients with other symptoms’ $\sim Po(\lambda_2 = 2.2);$ $\rho(X_1, X_2) = 0.35.$
(A) $P(X_1 > 2) = 1 - P(X_1 \leq 2) = 1 - (0.2466 + 0.34524 + 0.24167) = 0.1665.$
(B) $E(X_1 + X_2) = E(X_1) + E(X_2) = \lambda_1 + \lambda_2 = 3.6, V(X_1 + X_2) = V(X_1) + V(X_2) + 2C(X_1, X_2) = \lambda_1 + \lambda_2 + 2\rho\sqrt{\lambda_1}\sqrt{\lambda_2} = 4.8285, \sigma(X_1 + X_2) = 2.19738.$
Exercise 1. A farmers’ cooperative in the Manabi province produces bananas for a Fair Trade circuit. The issue of interest is if, among the associate producers, the yields per unit of land are related to the firm size or not. The following sample statistics are computed on 2010 data:

<table>
<thead>
<tr>
<th>Farm 1</th>
<th>Farm 2</th>
<th>Farm 3</th>
<th>Farm 4</th>
<th>Farm 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yields per ha (ton/ha)</td>
<td>29</td>
<td>27</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>Cultivated area (ha)</td>
<td>19</td>
<td>18</td>
<td>13</td>
<td>10</td>
</tr>
</tbody>
</table>

Formulate a sensible statistical model and answer the following questions:
(A) Compute estimates and standard errors of the two regression parameters.
(B) Is the unit yield significantly related to the size of the cultivated area? ($\alpha = 0.01$)
(C) Decompose the variance of the dependent variable in its components and comment the resulting goodness-of-fit index.

Exercise 2. A relevant point concerns the possible price difference with bananas commerced outside Fair Trade circuits. The following table summarize a sample of data collected in 2010.

<table>
<thead>
<tr>
<th>observations</th>
<th>mean(price per ton)</th>
<th>sd(price per ton)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fair Trade circuit</td>
<td>13</td>
<td>300</td>
</tr>
<tr>
<td>Normal Trade circuit</td>
<td>15</td>
<td>75</td>
</tr>
</tbody>
</table>

Assuming that the prices are normally distributed (assumption not completely justified) in both trade circuits and with equal variances, answer the following questions.
(A) (No for students with 3 CFU) Estimate the mean difference of prices for bananas commerced in the two different trade circuits. Compute the standard error of the point estimator employed.
(B) (No for students with 6 CFU) Compute the 95% confidence interval for the mean difference of prices.
(C) (Only for students with 6 CFU) Are the mean prices payed within the Fair Trade circuit significantly higher than the other ones? ($\alpha = 0.01$)
(D) (No for students with 6 CFU) Compute the 95% confidence interval for the common standard deviation.

Exercise 3. (No for students with 3 CFU) Data concerning the 2010 yields of the farmers’ cooperative indicate that 13.2% of the whole production of bananas has important defects (and thus cannot be sold). Unfortunately, tomorrow we will need to spend some hours for controlling a simple random sample of 200 bananas. Referring to such a sample:
(A) What is the probability of finding less than 14.8% of bananas with defects?
(B) What is the probability of the same event in A) knowing that at least 13.2% of bananas of the sample has defects?

Solution

Exercise 1. Simple Linear Regression Model $y = \beta_0 + \beta_1 x + u$ with the classical assumptions on $u$.
(A) $\hat{\beta}_1 = 0.8357, \hat{\beta}_0 = 12.1993, \text{se}(\hat{\beta}_1) = 0.175, \text{se}(\hat{\beta}_0) = 2.6227$.
(B) Test for $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$; test statistic: $(\hat{\beta}_1 - \beta_1)/\text{se}(\hat{\beta}_1)$ with distribution $T(n-2)$; sample value of the test statistic (under $H_0$): $t_{\text{amp}} = 4.7752$; Acceptation region for $\alpha = 0.01$: $[-5.8409, 5.8409]$.
(C) $\text{dev}(y) = 45.2, \text{dev}(\text{REG}) = 39.9448, \text{dev}(\text{RES}) = 5.2552; R^2 = 0.8837$ so that the regression model explains 88.37% of the deviance of the dependent variable.
Useful formulas and values:

<table>
<thead>
<tr>
<th></th>
<th>Farm 1</th>
<th>Farm 2</th>
<th>Farm 3</th>
<th>Farm 4</th>
<th>Farm 5</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_i$</td>
<td>29</td>
<td>27</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>122</td>
</tr>
<tr>
<td>$x_i$</td>
<td>19</td>
<td>18</td>
<td>13</td>
<td>10</td>
<td>13</td>
<td>73</td>
</tr>
<tr>
<td>$y_i^2$</td>
<td>841</td>
<td>729</td>
<td>484</td>
<td>484</td>
<td>484</td>
<td>3022</td>
</tr>
<tr>
<td>$x_i^2$</td>
<td>361</td>
<td>324</td>
<td>169</td>
<td>100</td>
<td>169</td>
<td>1123</td>
</tr>
<tr>
<td>$x_iy_i$</td>
<td>551</td>
<td>486</td>
<td>286</td>
<td>220</td>
<td>286</td>
<td>1829</td>
</tr>
</tbody>
</table>

$n = 5, \bar{x} = \sum_{i=1}^{n} x_i/n = 14.6, \bar{y} = \sum_{i=1}^{n} y_i/n = 24.4, dev(x) = \sum_{i=1}^{n} x_i^2 - nx^2 = 57.2, dev(y) = \sum_{i=1}^{n} y_i^2 - ny^2 = 45.2, codev(x,y) = \sum_{i=1}^{n} x_iy_i - nx\bar{y} = 47.8, \hat{\beta}_1 = codev(x,y)/dev(x), \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \hat{\sigma}^2 = dev(RES)/(n-2) = 1.7517, se(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2/\text{dev}(x)}, se(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2(1/n + \bar{x}^2/\text{dev}(x))}; t_{\text{comp}} = (\hat{\beta}_1 - 0)/se(\hat{\beta}_1); dev(REG) = \beta_1^2 \text{dev}(x), dev(RES) = dev(y) - dev(REG).

**Exercise 2.** $X_1 = \text{‘prices for the Fair Trade circuit’} \sim N(\mu_1, \sigma_1), X_2 = \text{‘prices for other circuits’} \sim N(\mu_2, \sigma_2)$; samples from $X_1$ and $X_2$ are independent; we assume $\sigma_1 = \sigma_2 = \sigma$ (see the text). We need to make inference on $\mu_1 - \mu_2$.

(A) The point estimator of the quantity of interest is $\bar{X}_1 - \bar{X}_2$; point estimate: $\bar{x}_1 - \bar{x}_2 = 225$; standard error (under the assumption of a common standard deviation): $se(\bar{X}_1 - \bar{X}_2) = s_{\text{pooled}}\sqrt{1/n_1 + 1/n_2} = 8.8437$.

(B) Pivot for $\mu_1 - \mu_2$: $[\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)]/(s_{\text{pooled}}\sqrt{1/n_1 + 1/n_2})$ with distribution $T(n_1 + n_2 - 2)$; confidence interval for $\alpha = 0.05$: $[206.8216, 243.1784]$ (value from the $T(26)$ table: $t = 2.0555$).

(C) Test for $H_0 : \mu_1 - \mu_2 = 0$ against $H_1 : \mu_1 - \mu_2 > 0$; test statistic: $\frac{\bar{X}_1 - \bar{X}_2}{(s_{\text{pooled}}\sqrt{1/n_1 + 1/n_2})}$ with distribution $T(n_1 + n_2 - 2)$; rejection region for $\alpha = 0.01$: $(2.4786, \infty)$; sample value of the test statistic (under $H_0$): $t_{\text{comp}} = 25.4419$.

(D) Pivot for $\sigma^2$: $S_{\text{pooled}}^2(n_1 + n_2 - 2)/\sigma^2$ with distribution $\chi^2(n_1 + n_2 - 2)$; confidence interval for $\sigma^2$ with $\alpha = 0.05$: $[337.8008, 1022.9541]$ (values from the $\chi^2(26)$ table: $c_1 = 13.8439, c_2 = 41.9232$); confidence interval for $\sigma$: $[18.3794, 31.9837]$.

Useful formulas and values: $n_1 = 13, n_2 = 15, \bar{x}_1 = 300, \bar{x}_2 = 75, s_1^2 = 605.16, s_2^2 = 492.84, s_{\text{pooled}}^2 = [s_1^2(n_1 - 1) + s_2^2(n_2 - 1)]/(n_1 + n_2 - 2) = 544.68, s_{\text{pooled}} = 23.3384$.

**Exercise 3.** $X = \text{‘a single banana has defects’} \sim Be(p = 0.132); Y = \text{‘number of bananas with defects in a s.r.s. of 200’} \sim Bi(n = 200, p = 0.132) \approx N(np = 26.4, npq = 22.9152)$ (note that $npq \geq 10$); $W = \text{‘proportion of bananas with defects in a s.r.s. of 200’} \approx N(p = 0.132, pq/n = 0.0239^2)$.

(A) $P(W < 0.148) = P(Z < 0.6685) = 0.74809$.

(B) $P(W < 0.148|W \geq 0.132) = P(W < 0.148, W \geq 0.132)/P(W \geq 0.132) = P(0.132 \leq W < 0.148)/P(W \geq 0.132) = (0.74809 - 0.5)/(1 - 0.5) = 0.49617$. 
3 Test 09.02.2012

Text

Framework: A social program in Bangladesh.

Exercise 1. **No for students with 3 CFU.** The SOKE social program, directed to disadvantage young people, aims at facilitating the entry of enrolled individuals on the job market. Specifically, enrolled people is trained to assemble and install small solar kits for energy production. After 3 years the program concludes and trained people have to find themselves a job. About 2 years after the individual terminated the program, the company monitors their job status. The following table compares the job status of a sample of individuals involved in the SOKE with the corresponding situation of similar programs.

<table>
<thead>
<tr>
<th></th>
<th>SOKE program</th>
<th>Similar programs</th>
</tr>
</thead>
<tbody>
<tr>
<td>People surveyed</td>
<td>113</td>
<td>121</td>
</tr>
<tr>
<td>Of whom workers</td>
<td>50.3%</td>
<td>38.3%</td>
</tr>
</tbody>
</table>

(A) Does the SOKE program give a probability of finding a job significantly higher than other similar programs? (**α = 0.01**)

(B) Compute the power of the test at point (A) if the alternative hypothesis is “the SOKE program gives a probability of finding a job of 15.1 percentage points higher that other similar programs”.

(C) Compute the p-value of the test at point (A) and comment the result.

Exercise 2. **(No for students with 3 CFU).** The producer guarantees for the solar kit at least 6900 hours of duration (whereas not relevant for the exercise, a plant is considered functioning if it works at least at 85% if its maximum capabilities).

(A) Assume that the duration of a kit is a \( N(\mu = 8090, \sigma = 940) \) random variable. Considering a simple random sample of 14 units, compute the probability that at least 13 of them give the duration guaranteed.

(B) Assume that the duration of the kit is a Normal random variable with \( \sigma \) as in point (A). Compute the mean duration \( \mu \) in order a randomly selected kit guarantees at least the required duration with probability 0.951.

Exercise 3. The Dhaka stock (the main Bangladeshi stock market) is an emerging financial market attracting an increasing number of investors. For assessing its risk level, the monthly returns of the Dhaka stock index are regressed against the corresponding S&P 500 returns (the S&P500, the index of the 500 largest US companies, is taken as a benchmark). The regression, concerning 52 observations, has given the following statistics:

<table>
<thead>
<tr>
<th>( \beta_0 )</th>
<th>( \beta_1 )</th>
<th>( \sigma )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>-0.0057</td>
<td>1.0708</td>
</tr>
<tr>
<td>Standard error</td>
<td>0.011606</td>
<td>0.236</td>
</tr>
</tbody>
</table>

with a covariance between \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) estimated to \(-0.00001346\). The financial literature indicates that if the \( \beta_1 \) coefficient is lower/equal/higher than 1, then the market in the dependent variable is less/equally/more riskier than the market in the independent variable. By adopting the usual assumptions for the linear regression model, answer the following questions.

(A) Can we deduce that the Dhaka market has a risk level significantly different from the benchmark? (**α = 0.05**)

(B) **No for students with 6 CFU.** Compute the confidence interval for \( \sigma \) at the confidence level 0.99.

(C) **Only for students with 6 CFU.** Compute the prediction of the conditional mean corresponding to a S&P500 return of \(-0.092\); compute also the standard error of the corresponding estimator. (Hint: Derive the standard error using the properties of linear transformations of random variables applied to the formula of the prediction)

(D) **No for students with 6 CFU.** Compute the confidence interval for the conditional mean, corresponding to a S&P500 return of \(-0.092\), at a 99% confidence level. (Hint: follow the hint of point (C))
Solution

Exercise 1. $X_1$ = ’a young person trained by SOKE works’ $\sim Be(p_1)$, $X_2$ = ’a young person trained by other programs works’ $\sim Be(p_2)$; samples from $X_1$ and $X_2$ are independent. We need to make inference on $p_1 - p_2$.

(A) Test for $H_0 : p_1 - p_2 = 0$ against $H_1 : p_1 - p_2 > 0$; test statistic: $(\hat{p}_1 - \hat{p}_2 - 0)/se$ with approximate distribution $N(0,1)$; rejection region for $\alpha = 0.01$: $(2.3263, \infty)$; sample value of the test statistic (under $H_0$):

$$z_{\text{camp}} = 1.8593.$$ 

(B) Power of the test for $H_1 : p_1 - p_2 = 0.151$. $\gamma = P(sample \in R\mid H_1) = P[(\hat{p}_1 - \hat{p}_2 - 0)/se > z_{\text{crit}}\mid H_1] = P[\hat{p}_1 - \hat{p}_2 > z_{\text{crit}}se\mid H_1] = P[(\hat{p}_1 - \hat{p}_2 - 0.151)/se > (z_{\text{crit}}se - 0.151)/se\mid H_1] = P(Z > -0.0153\mid H_1) = 0.50531.$$

(C) $p - value = P((\hat{p}_1 - \hat{p}_2 - 0)/se > z_{\text{camp}}\mid H_0) = P(Z > 1.8593\mid H_0) = 0.03149.$

Useful formulas and values: $n_1 = 113$, $\hat{p}_1 = 0.503$, $\hat{q}_1 = 0.497$, $n_2 = 121$, $\hat{p}_2 = 0.383$, $\hat{q}_2 = 0.617$, $se = \sqrt{\hat{p}_1\hat{q}_1/n_1 + \hat{p}_2\hat{q}_2/n_2} = \sqrt{0.00221231 + 0.00195298} = 0.06454.$

Exercise 2. $X$ = ’duration of one kit’ $\sim N(\mu = 8090, \sigma = 940)$.

(A) $Y$ = ’one kit gives the guaranteed duration’ $\sim Be(p = 0.8972)$, where $p = P(Y = 1) = P(X > 6900) = P(Z > -1.266)$; $W$ = ’number of kits, within a s.r.s of 14 giving the guaranteed duration’ $\sim Bi(n = 14, p = 0.8972)$. $P(W \geq 13) = P(W = 13) + P(W = 14) = 0.35136 + 0.21913 = 0.57049.$

(B) $X \sim N(\mu, \sigma = 940)$. We want to find $\mu$ such that $0.951 = P(X \geq 6900) = P((X - \mu)/\sigma \geq (6900 - \mu)/940) = P(Z > -1.6546)$. By equating $(6900 - \mu)/940 = -1.6546$ we get $\mu = 8455.4$.

Exercise 3. Simple Linear Regression Model $y = \beta_0 + \beta_1 x + u$ with the classical assumptions on $u$.

(A) Test for $H_0 : \beta_1 = 1$ against $H_1 : \beta_1 \neq 1$; test statistic: $(\hat{\beta}_1 - 1)/se(\hat{\beta}_1)$ with distribution $T(n - 2)$; sample value of the test statistic (under $H_0$): $t_{\text{camp}} = 0.3$; Acceptation region for $\alpha = 0.05$: $[-2.0086, 2.0086]$.

(B) Pivot for $\sigma^2$: $\hat{\sigma}^2(n - 2)/\sigma^2 \sim \chi^2(n - 2)$; interval at $\alpha = 0.01$ for $\sigma^2$: $[\hat{\sigma}^2(n - 2)/c_2, \hat{\sigma}^2(n - 2)/c_1] = [0.0044066, 0.0125143]$; where $c_1 = 27.9907$ and $c_2 = 79.49$; corresponding interval for $\sigma$: $[0.0664, 0.1119]$.

(C) Taking $x_0 = -0.092$, we denote the conditional mean as $\mu_0 = E(y|x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$. As a consequence: $\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = -0.1042$; $V(\hat{\mu}_0) = V(\hat{\beta}_0) + x_0^2V(\hat{\beta}_1) + 2x_0C(\hat{\beta}_0, \hat{\beta}_1)$, so that by replacing variances and covariances with the corresponding sample estimates we have $se(\hat{\mu}_0) = \sqrt{\hat{V}(\hat{\beta}_0)} = 0.00060859 = 0.02467$.

(D) We use the notation of point (C). Pivot for $\mu_0$: $(\hat{\mu}_0 - \mu_0)/se(\hat{\mu}_0) \sim T(n - 2)$; confidence interval for $\mu_0$ at $\alpha = 0.01$: $[\hat{\mu}_0 - tse(\hat{\mu}_0), \hat{\mu}_0 + tse(\hat{\mu}_0)] = [-0.1703, -0.0382] (t = 2.6778)$.

Useful formulas and values: $n = 52$, $df = n - 2 = 50$, $\hat{\beta}_0 = -0.0057$, $\hat{\beta}_1 = 1.0708$, $\hat{\sigma} = 0.0837$, $\hat{V}(\hat{\beta}_0) = 0.0116062 = 0.00013470$, $\hat{V}(\hat{\beta}_1) = 0.2362 = 0.05569600$, $\hat{C}(\hat{\beta}_0, \hat{\beta}_1) = -0.00001346$. 

7
4 Test 18.06.2012

Text

Framework: The topic of interest is the possible relationship between fertility rate (number of children per woman, married or in union, in her lifetime) and contraceptive prevalence (percentage of women, married or in union, using contraception).

Exercise 1. A study evaluates the possible relationship between fertility rate, taken as dependent variable, and the contraceptive prevalence in different developing countries.

<table>
<thead>
<tr>
<th>Country</th>
<th>Afghanistan</th>
<th>Nigeria</th>
<th>Namibia</th>
<th>Mauritania</th>
<th>Nicaragua</th>
<th>Yemen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fertility rate</td>
<td>7.1</td>
<td>5.3</td>
<td>3.2</td>
<td>4.4</td>
<td>2.8</td>
<td>5.5</td>
</tr>
<tr>
<td>Contraceptive prevalence</td>
<td>5</td>
<td>13</td>
<td>44</td>
<td>8</td>
<td>69</td>
<td>21</td>
</tr>
</tbody>
</table>

Formulate a sensible statistical model and answer the following questions.

(A) No for students with 6 CFU. Compute the confidence interval for $\beta_1$ at the confidence level 0.99.

(B) Compute the $R^2$ index and interpret the value obtained.

(C) Compute the regression residual for country Mauritania.

Exercise 2. (No for students with 3 CFU). Since the Republic of Mali has one of the largest fertility rates in the world, the women of the Sikasso region (one of the most populated areas of the country) has bee involved in an NGO program for diffusing contraception among women. The following table compares the results of two surveys lead in 2011 and before the beginning of the program.

<table>
<thead>
<tr>
<th></th>
<th>2009 Survey</th>
<th>2011 Survey</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size</td>
<td>525</td>
<td>609</td>
</tr>
<tr>
<td>Contraceptive prevalence (%)</td>
<td>11.9</td>
<td>16.5</td>
</tr>
</tbody>
</table>

(A) Has the program been effective in increasing the contraceptive prevalence? ($\alpha = 0.01$)

(B) Compute the p-value of the test at point (A).

(C) Compute the power of the test at point (A) if the alternative hypothesis is “the NGO program increases the contraceptive prevalence of 5.9 percentage points”.

Exercise 3. (No for students with 3 CFU). The number of children per married woman (in her reproductive life) follows a Poisson distribution with parameter 2.5 if a woman uses contraceptive methods and parameter 4.6 if she does not; moreover, 20% of married women use contraceptive methods.

(A) Knowing that a married woman (randomly drawn from the population) had 5 children in her reproductive life, compute the probability that she used contraceptive methods.

(B) Compute the probability that a married woman (randomly drawn from the population) in her reproductive life had 5 children and did not use contraceptive methods.

Solution

Exercise 1. Simple Linear Regression Model $y = \beta_0 + \beta_1 x + u$ with the classical assumptions on $u$.

(A) Pivot for $\beta_1$: $(\hat{\beta}_1 - \beta_1)/se(\hat{\beta}_1) \sim T(n - 2)$; interval at $\alpha = 0.01$ for $\hat{\beta}_1$: $[\hat{\beta}_1 - t \cdot se(\hat{\beta}_1), \hat{\beta}_1 + t \cdot se(\hat{\beta}_1)] = [-0.1347,0.0289]$, where $t = 4.6041$.

(B) $R^2 = \text{dev}(\text{REG})/\text{dev}(y) = 0.6893$.

(C) Mauritania corresponds to $i = 4$; thus $\hat{y}_4 = \hat{\beta}_0 + \hat{\beta}_1 x_4 = 5.7043$, $u_4 = y_4 - \hat{y}_4 = -1.3043$.

Useful formulas and values:
Exercise 2. \( X_1 \) = ‘a women of the Sikasso region used contraception (2009)’ \( \sim \) \( Be(p_1) \), \( X_2 \) = ‘a women of the Sikasso region used contraception (2011)’ \( \sim \) \( Be(p_2) \); samples from \( X_1 \) and \( X_2 \) are independent. We need to make inference on \( p_2 - p_1 \).

(A) Test for \( H_0 : p_2 - p_1 = 0 \) against \( H_1 : p_2 - p_1 > 0 \); test statistic: \( (\hat{p}_2 - \hat{p}_1 - 0)/se \) with approximate distribution \( N(0, 1) \); rejection region for \( \alpha = 0.01 \): (2.3263, \( \infty \)); sample value of the test statistic (under \( H_0 \)): \( z_{camp} = 2.2289 \).

(B) \( p-value = P(\hat{p}_2 - \hat{p}_1 - 0)/se > z_{camp}|H_0) = P(Z > 2.2289|H_0) = 0.01291 \).

(C) Power of the test for \( H_1 : p_2 - p_1 = 0.059 \). \( \gamma = P(\text{sample} \in \text{R}|H_1) = P(\hat{p}_2 - \hat{p}_1 - 0)/se > z_{crit}|H_1) = P(\hat{p}_2 - \hat{p}_1 > z_{crit}se|H_1) = P(\hat{p}_2 - \hat{p}_1 - 0.059)/se > (z_{crit}se - 0.059)/se|H_1) = P(Z > -0.5325|H_1) = 0.7028 \).

Useful formulas and values: \( n_1 = 525 \), \( \hat{p}_1 = 0.119 \), \( \hat{q}_1 = 0.881 \), \( n_2 = 609 \), \( \hat{p}_2 = 0.165 \), \( \hat{q}_2 = 0.835 \), \( se = \sqrt{\hat{p}_1\hat{q}_1/n_1 + \hat{p}_2\hat{q}_2/n_2} = \sqrt{0.00019969 + 0.00022623} = \sqrt{0.00042592} = 0.02064 \).

Exercise 3. \( X \) = ‘number of children per woman during her reproductive life’; \( U \) = ‘a woman uses contraception’. \( X|U \sim Po(\lambda_1 = 2.5) \), \( X|\overline{U} \sim Po(\lambda_2 = 4.6) \), \( P(\overline{U}) = 0.2 \). Let us denote the event \( (X = 5) \) as \( A \).

(A) Then \( P(U|X = 5) = P(U|A) = P(A|U)P(U)/P(A) = 0.08826 \)

(B) \( P(U, X = 5) = P(U \cap A) = P(A|\overline{U})P(\overline{U}) = 0.13802 \).

Useful formulas and values: By using the p.m.f. of the Poisson distribution we get \( P(A|U) = P(X = 5|U) = 0.0668 \), \( P(A|\overline{U}) = P(X = 5|\overline{U}) = 0.17253 \); \( P(A) = P(A|U)P(U) + P(A|\overline{U})P(\overline{U}) = 0.15138 \).
5 Test 20.07.2012

Text
Framework: Soil erosion.

Exercise 1. (No for students with 3 CFU). Soil erosion in a degraded mountain area progresses every 5-years (in mm) with a mean of 19.7 and a standard deviation of 10.9. Assuming normality, independence (in the sense that erosion over different periods are independent) and additivity (in the sense that erosion over many periods is the sum of the corresponding quantities in the subperiods):
(A) Compute the probability that, in 20 years, soil erosion lies between 82 and 109 mm.
(B) Compute the probability of the event in (A) knowing that soil erosion in the same period is greater than 91 mm.

Exercise 2. (No for students with 3 CFU). La Paz administration (Bolivia) is monitoring soil erosion of its mountain area. The following table summarize a simple random sample of measures of 5-years soil erosion in different La Paz sites.

<table>
<thead>
<tr>
<th>Site 1</th>
<th>Site 2</th>
<th>Site 3</th>
<th>Site 4</th>
<th>Site 5</th>
<th>Site 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Erosion (mm)</td>
<td>2.1</td>
<td>8.1</td>
<td>3.7</td>
<td>3.5</td>
<td>3.9</td>
</tr>
</tbody>
</table>

Assuming that the measures of soil erosion follow a Normal distribution:
(A) Is the mean soil erosion significantly higher than 3 mm? (\(\alpha = 0.1\))
(B) Compute the pdf of the first observation in the sample at the maximum likelihood estimates of the parameters.

Exercise 3. In areas affected by this phenomenon, an important contribution to soil erosion is provided by runoff from melting snow. In order to evaluate such a contribution, a research institution has collected some sample information (RE = winter-adjusted Rainfall Erosivity index, WR = 5-years Winter Runoff) from 32 different climatological stations, aiming at modeling the rainfall erosivity index as a function of the winter runoff.

<table>
<thead>
<tr>
<th>RE</th>
<th>WR</th>
</tr>
</thead>
<tbody>
<tr>
<td>145</td>
<td>107</td>
</tr>
<tr>
<td>346000</td>
<td>172000</td>
</tr>
<tr>
<td>234000</td>
<td></td>
</tr>
</tbody>
</table>

(A) (No for students with 6 CFU) Compute the confidence interval (\(\alpha = 0.02\)) for the standard deviation of the error component.
(B) Hydrological theory prescribes that at WR equal to zero, RE also should be zero. Do data confirm this prescription? Answer the question by setting the problem as a statistical hypothesis test (\(\alpha = 0.02\)).
(C) Decompose the deviance of RE in their regression and residual components providing the three values; interpret the values obtained.
(D) (No for students with 6 CFU) Historical data for La Paz Department give a WR of 153. By means of the estimated model, compute the confidence interval (confidence level = 0.99) for the corresponding conditional expectation of the RE index.

Solution

Exercise 1. \(X_i\), 'soil erosion (in mm) in the \(i\)-th period (1 period = 5 years)' \(\sim N(\mu = 19.7, \sigma^2 = 118.81)\). \(X_i\) for different periods are independent and additive. \(Y = \sum_{i=1}^{n} X_i\) (soil erosion over \(n = 4\) periods) \(\sim N(n\mu = 78.8, n\sigma^2 = 475.24)\) (\(\sqrt{n}\sigma = 21.8\)).
(A) Then \(P(82 < Y < 109) = P(0.147 < Z < 1.385) = 0.35867.\)
Exercise 2. \(X = \) soil erosion in 5-years (mm) \(N(\mu, \sigma)\).

(A) Test of \(H_0 : \mu = 3\) against \(H_1 : \mu > 3\); test statistic: \((\bar{x} - 3)/\text{se}\) with distribution \(T(n-1 = 5)\); rejection region for \(\alpha = 0.1\): \((1.4759, \infty)\); sample value of the test statistic (under \(H_0\)): \(t_{\text{samp}} = 0.8219\).

(B) \(f(x_1 = 2.1; \mu = 3.78333, \sigma^2 = 4.54139) = 0.13703\) where we set \(\mu\) and \(\sigma^2\) at their ML estimates \(\hat{\mu}_{\text{ML}} = \bar{x}\), \(\hat{\sigma}^2_{\text{ML}} = s^2\) and \(f(x; \mu, \sigma^2)\) denotes the formula of the p.d.f. of the normal distribution.

Useful formulas and values:

\[
\begin{array}{cccccc}
   & \text{Site 1} & \text{Site 2} & \text{Site 3} & \text{Site 4} & \text{Site 5} & \text{Site 6} \\
   x_i & 2.1 & 8.1 & 3.7 & 3.5 & 3.9 & 1.4 \\
   x_i^2 & 4.41 & 65.61 & 13.69 & 12.25 & 15.21 & 1.96 \\
   \text{Sum} & 22.7 & & & & & \\
\end{array}
\]

\(n = 6, \bar{x} = \sum_{i=1}^{n} x_i/n = 3.78333, \text{dev}(\bar{x}) = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 27.24833, s^2 = \text{dev}(\bar{x})/(n - 1) = 5.44967, s = 2.33445, \text{se} = s/\sqrt{n} = 0.95304; s^2 = \text{dev}(\bar{x})/n = 4.54139\).

Exercise 3. Simple Linear Regression Model \(y = \beta_0 + \beta_1 x + u\) with the classical assumptions on \(u\), where \(y = RE\) and \(x = WR\).

(A) Pivot for \(\sigma^2\): \((n - 2)\hat{\sigma}^2/\sigma^2 \sim \chi^2(n - 2)\); interval at \(\alpha = 0.02\) for \(\sigma^2\): \([\hat{(n - 2)\hat{\sigma}^2/c_2}, (n - 2)\hat{\sigma}^2/c_1]\) = [543.33, 1849.15]; corresponding interval for \(\sigma\): \([23.3094, 43.0017]\) \(c_1 = 14.9535, c_2 = 50.8922\).

(B) Test of \(H_0 : \beta_0 = 0\) against \(H_1 : \beta_0 \neq 0\); test statistic: \((\hat{\beta}_0 - 0)/\text{se}(\hat{\beta}_0)\) with distribution \(T(n - 2)\); acceptance region for \(\alpha = 0.02\): \([-2.4573, 2.4573]\); sample value of the test statistic (under \(H_0\)): \(t_{\text{samp}} = -0.06\).

(C) \(\text{dev}(y) = 346000, \text{dev}(REG) = 318348.84, \text{dev}(RES) = 27651.16\); the regression model explains 92.01% of the variability of \(y\).

(D) Pivot for \(\mu_0 = E(y|x_0 = 153)\): \((\hat{\mu}_0 - \mu_0)/\text{se}(\hat{\mu}_0) \sim T(n - 2)\); interval at \(\alpha = 0.01\) for \(\mu_0\): \([\hat{\mu}_0 - t \cdot \text{se}(\hat{\mu}_0), \hat{\mu}_0 + t \cdot \text{se}(\hat{\mu}_0)]\) = [190.158, 225.0048].

Useful formulas and values:

\(n = 32, df = n - 2 = 30, \hat{\beta}_1 = \text{cod}(\bar{x}, y)/\text{dev}(\bar{x}) = 1.3605, \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = -0.5698, \text{dev}(RES) = \text{dev}(y) - \hat{\beta}_1^2 \text{dev}(\bar{x}) = 27651.16, \hat{\sigma}^2 = \text{dev}(RES)/df = 921.7054, \hat{\sigma} = 30.3596, \text{se}(\hat{\beta}_0) = \hat{\sigma} \sqrt{1/n + \bar{x}^2/\text{dev}(\bar{x})} = 9.495, \text{dev}(REG) = \hat{\beta}_1^2 \text{dev}(\bar{x}) = 318348.84, \hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = 207.5814, \text{se}(\hat{\mu}_0) = \hat{\sigma} \sqrt{1/n + (x_0 - \bar{x})^2/\text{dev}(\bar{x})} = 6.3358\).
6 Test 07.09.2012

Text

Framework: One of the most difficult tasks of building and managing a global portfolio is to assess the risk of foreign investments.

Exercise 1. Some researchers use country credit ratings (CR) for predicting the country annualized risk (AR), that represents the quantity of interest. The following table summarizes some statistics on a sample of 28 countries.

\[
\begin{array}{cccc}
\frac{1}{n} \sum_{i=1}^{n} CR_i & \frac{1}{n} \sum_{i=1}^{n} (CR_i - \bar{CR})^2 & \frac{1}{n} \sum_{i=1}^{n} AR_i & \frac{1}{n} \sum_{i=1}^{n} (AR_i - \bar{AR})^2 \\
63 & 417 & 33 & 224 & -167 \\
\end{array}
\]

Formulate an appropriate statistical model and answer the following questions.

(A) Test whether the independent variable is significant (\( \alpha = 0.01 \)).

(B) (No for students with 6 CFU) Compute the 99% confidence interval for the standard deviation of the error component.

(C) The observation corresponding to South Africa, \((CR = 75, AR = -14)\), has been removed because considered an outlier (an outlier is here defined as an observation whose residual is outside the interval \( \pm 3 \) times the standard deviation of the error component). Compute the residual corresponding to the given observation and indicate if it is an outlier.

(D) A general rule is to include a country in the portfolio only if its predicted (conditional mean) annual risk is significantly below 46. Knowing that the Ukrainian credit rating is 73.42, can it be included in the portfolio? (\( \alpha = 0.02 \)).

Exercise 2. (No for students with 3 CFU). Two agencies have evaluated the credit rating of a set of potentially interesting countries.

<table>
<thead>
<tr>
<th>Credit rating by AZ agency</th>
<th>Algeria</th>
<th>Bangladesh</th>
<th>Egypt</th>
<th>Kenya</th>
<th>South Africa</th>
<th>Zimbabwe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Credit rating by BY agency</td>
<td>95</td>
<td>97</td>
<td>75</td>
<td>60</td>
<td>80</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>58</td>
<td>56</td>
<td>68</td>
<td>65</td>
<td>61</td>
<td>79</td>
</tr>
</tbody>
</table>

Assuming that the credit rating follows a Normal distribution answer the following questions.

(A) Do the means of the credit ratings proposed by the two agencies differ significantly? (\( \alpha = 0.01 \))

(B) How many units we would need in the sample in order to build a confidence interval for the difference between the two means of size 9.5 for \( \alpha = 0.01 \)?

Exercise 3. (No for students with 3 CFU). Two random variables have the following joint p.m.f.:

\[
\begin{array}{c|cc}
x & 1 & 3 \\
\hline
-2 & 0.07 & 0.27 \\
0 & 0.12 & 0.16 \\
2 & 0.06 & 0.32 \\
\end{array}
\]

(A) Compute \( E(X|Y = 3) \) and \( V(X|Y = 3) \).

(B) Compute \( E(X - Y) \).

Solution

Exercise 1. Simple Linear Regression Model \( y = \beta_0 + \beta_1 x + u \) with the classical assumptions on \( u \), where \( y = AR \) and \( x = CR \).
(A) Test of $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$; test statistic: $(\hat{\beta}_1 - 0)/se(\hat{\beta}_1)$ with distribution $T(n-2)$; acceptation region for $\alpha = 0.01$: $(-2.779, 2.779)$; sample value of the test statistic (under $H_0$): $t_{samp} = -3.32675$.

(B) Pivot for $\sigma^2$: $(n-2)\hat{\sigma}^2/\sigma^2 \sim \chi^2(n-2)$; interval at $\alpha = 0.01$ for $\sigma^2$: $[(n-2)\hat{\sigma}^2/c_2, (n-2)\hat{\sigma}^2/c_1] = [91.1, 394.2]$; corresponding interval for $\sigma$: $[9.54, 19.85]$ ($c_1 = 11.1602$, $c_2 = 48.2899$).

(C) Fitted value: $\hat{y}_{SA} = \hat{\beta}_0 + \hat{\beta}_1 x_{SA} = 28.1942$; residual: $u_{SA} = y_{SA} - \hat{y}_{SA} = -42.1942$, so that it is an outlier.

(D) Denoting $\mu_0 = E(y|x_0 = 73.42)$, we need to test $H_0 : \mu_0 = 46$ against $H_1 : \mu_0 < 46$; test statistic: $(\hat{\mu}_0 - \mu_0)/se(\hat{\mu}_0)$ with distribution $T(n-2)$; rejection region for $\alpha = 0.02$: $(-\infty, -2.162)$; sample value of the test statistic (under $H_0$): $t_{samp} = -6.2225$.

Useful formulas and values: $n = 28$, $df = n - 2 = 26$, $\hat{\beta}_1 = \text{cod}(x,y)/\text{dev}(x) = -0.40048$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 58.2302$, $\text{dev}(\text{RES}) = \text{dev}(y) - \hat{\beta}_1^2 \text{dev}(x) = 4399.3573$, $\hat{\sigma}^2 = \text{dev}(\text{RES})/df = 169.2061$, $\hat{\sigma} = 13.0079$, $se(\hat{\beta}_1) = \hat{\sigma}/\sqrt{\text{dev}(x)} = 0.1204$, $\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = 28.827$, $se(\hat{\mu}_0) = \hat{\sigma}/\sqrt{1/n + (x_0 - \bar{x})^2/\text{dev}(x)} = 2.7598$.

**Exercise 2.** $X =$ 'credit rating by AZ'; $Y =$ 'credit rating by BY'. Since the sample units are the same in the two samples, we can handle the two questions as paired data, with $D = X - Y \sim N(\mu_D, \sigma_D^2)$.

(A) Test of $H_0 : \mu_D = 0$ against $H_1 : \mu_D \neq 0$; test statistic: $(\bar{d} - 0)/se(\bar{d})$ with distribution $T(n-1)$; acceptation region for $\alpha = 0.01$: $(-4.0321, 4.0321)$; sample value of the test statistic (under $H_0$): $t_{samp} = 1.9805$.

(B) $n = (2z\hat{\sigma}_D/A)^2 = 115.16 \approx 116$, where $z = 2.5758$ ($\alpha = 0.01$), $A = 9.5$.

Useful formulas and values:

<table>
<thead>
<tr>
<th>$x_i$</th>
<th>Algeria</th>
<th>Bangladesh</th>
<th>Egypt</th>
<th>Kenya</th>
<th>South Africa</th>
<th>Zimbabwe</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_i$</td>
<td>95</td>
<td>97</td>
<td>75</td>
<td>60</td>
<td>80</td>
<td>76</td>
<td></td>
</tr>
<tr>
<td>$d_i$</td>
<td>37</td>
<td>41</td>
<td>7</td>
<td>-5</td>
<td>19</td>
<td>-3</td>
<td>96</td>
</tr>
<tr>
<td>$d_i^2$</td>
<td>1369</td>
<td>1681</td>
<td>49</td>
<td>25</td>
<td>361</td>
<td>9</td>
<td>3494</td>
</tr>
</tbody>
</table>

$n = 6$, $df = n - 1 = 5$, $\overline{d} = 16$, $\text{dev}(\overline{d}) = \sum_{i=1}^{n} d_i^2 - n \overline{d}^2 = 1958$, $s_D^2 = \text{dev}(\overline{d})/df = 391.6$, $s_D = 19.78889$, $se(\overline{d}) = \sqrt{s_D^2/n} = \sqrt{65.2667} = 8.07878$.

**Exercise 3.**

(A) $E(X|Y = 3) = 0.13333$, $V(X|Y = 3) = 3.12889$.

(B) $E(X - Y) = E(X) - E(Y) = 0.08 - 2.5 = -2.42$.

Useful formulas and values: $f(x|Y = 3) = f(x, 3)/f_Y(3)$,

| $x$ | $f(x|Y = 3)$ | $-2$ | 0.36 | 0.21333 | 0.42667 | Sum |
|-----|--------------|------|------|---------|---------|-----|
| $x f(x|Y = 3)$ | -0.72 | 0 | 0.85333 | 0.13333 | | |
| $x^2 f(x|Y = 3)$ | 1.44 | 0 | 1.70667 | 3.14467 | | |

<table>
<thead>
<tr>
<th>$x$</th>
<th>-2</th>
<th>0</th>
<th>2</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f(x)$</td>
<td>0.34</td>
<td>0.28</td>
<td>0.38</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$y$</th>
<th>1</th>
<th>3</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f(y)$</td>
<td>0.25</td>
<td>0.75</td>
<td>1</td>
</tr>
</tbody>
</table>

| $y f(y)$ | 0.25 | 2.25 | 2.5 |
Text

Framework: The topic of interest is the intergenerational mobility in earnings.

Exercise 1. **(No for students with 3 CFU)**. Let $X$ be a $N(\mu, \sigma^2)$ r.v. with a known $\sigma$ and let $\mathbf{X} = (X_1, X_2, X_3)$ be a simple random sample from $X$. Consider the following estimators of $\mu$:

$$T_1 = \frac{4X_1 + 5X_2 - 3X_3}{6} \quad T_2 = \frac{3X_1 + 5X_2}{8}$$

(A) Which one is more efficient between $T_1$ and $T_2$? Motivate the answer.

**(No for students with 6 CFU)** (B) We are interested to interval estimation of $\mu$. By exploiting the best estimator at point (A), propose a pivot for $\mu$ for which you have got statistical tables. Explain why the proposed quantity satisfies the properties of a pivot.

Exercise 2. M. Corak (University of Ottawa) in recent work analyzes if the intergenerational mobility in earnings is related to the wealth inequality. The following table displays some statistics relative to the sample of 18 countries considered in the work (N.B.: standard deviations and covariances are computed considering the number of observations as denominator).

<table>
<thead>
<tr>
<th>Intergenerational mobility index</th>
<th>mean</th>
<th>standard deviation</th>
<th>covariance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inequality index</td>
<td>3.046</td>
<td>1.602</td>
<td>-9.958</td>
</tr>
<tr>
<td></td>
<td>37.056</td>
<td>9.618</td>
<td></td>
</tr>
</tbody>
</table>

Formulate an appropriate statistical model and answer the following questions.

(A) Test whether the wealth inequality index is significant in the regression ($\alpha = 0.02$).

(B) Decompose the total variability of the dependent variable in their regression and residual components and comment the results.

(C) **(No for students with 6 CFU)** Compute the confidence interval ($\alpha = 0.02$) for the intercept of the model.

(D) Compute the residual for Italy (intergenerational mobility = 2, wealth inequality = 32). Comment the result.

Exercise 3. **(No for students with 3 CFU)** Two surveys, lead on two independent random samples of families, aimed at investigate if the level of intergenerational mobility (a continuous variable evaluated as explained above) has changed in US between 2005 and 2010. The following table reports some selected statistics from the two surveys.

<table>
<thead>
<tr>
<th></th>
<th>observations</th>
<th>mean</th>
<th>median</th>
<th>$s$</th>
<th>$s_*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>430</td>
<td>2.318</td>
<td>2.229</td>
<td>1.294</td>
<td>1.296</td>
</tr>
<tr>
<td>2005</td>
<td>209</td>
<td>2.099</td>
<td>2.003</td>
<td>1.247</td>
<td>1.25</td>
</tr>
</tbody>
</table>

(A) Is the average level of intergenerational mobility increased significantly between the two surveys ($\alpha = 0.05$)?

(B) Compute the probability of the second kind error of the test at point (A) by considering an alternative hypothesis under which the mean level of intergenerational mobility in 2010 is 0.31 points higher than in 2005.

---

1. Inequality from generation to generation: the United States in Comparison. The intergenerational mobility is evaluated as the reciprocal of the intergenerational elasticity in earnings, where this last quantity is the percentage difference in earnings in the child’s generation associated with the percentage difference in the parental generation (for example, an intergenerational elasticity in earnings of 0.6 tells us that if one father makes 100% more than another then the son of the high income father will, as an adult, earn 60% more than the son of the relatively lower income father). The wealth inequality is measured by means of the Gini index.
Solution

Exercise 1. Assumptions: \( X \sim N(\mu, \sigma^2) \) (\( \sigma \) known); \( X = (X_1, X_2, X_3) \) s.r.s. from \( X \). We at making inference on \( \mu \).

(A) By applying the properties of expected values we see that:

- \( E(T_1) = \mu \) and \( E(T_2) = \mu \) (both are unbiased estimators of \( \mu \));
- \( V(T_1) = 50/36\sigma^2 \) and \( V(T_2) = 34/64\sigma^2 \).

This implies that \( T_2 \) is more efficient.

(B) By exploiting \( T_2 \), the pivot for \( \mu \) with the required attributes is \( \frac{T_2 - \mu}{\sqrt{34/64}} \), whose distribution is \( N(0,1) \).

Exercise 2. Simple Linear Regression Model \( y = \beta_0 + \beta_1x + u \) with the classical assumptions on \( u \), where \( y = \text{'intergenerational mobility index'} \) and \( x = \text{'inequality index'} \).

(A) Test of \( H_0: \beta_1 = 0 \) against \( H_1: \beta_1 \neq 0 \); test statistic: \( (\hat{\beta}_1 - 0)/se(\hat{\beta}_1) \) with distribution \( T(n-2) \); acceptation region for \( \alpha = 0.02 \): \(-2.583, 2.583\); sample value of the test statistic (under \( H_0 \)): \( t_{samp} = -3.38771 \).

(B) \( dev(y) = 46.1953 \), \( dev(\text{REG}) = 19.2951 \), \( dev(\text{RES}) = 26.9002 \); the regression model explains 41.77% of the variability of \( y \).

(C) Pivot for \( \beta_0: \left( \hat{\beta}_0 - \beta_0 \right)/se(\hat{\beta}_0) \sim T(n-2) \); interval at \( \alpha = 0.02 \) for \( \beta_0: [3.892, 10.178] \) (\( t = 2.5835 \)).

(D) Fitted value: \( \hat{y}_{Ita} = \hat{\beta}_0 + \hat{\beta}_1x_{Ita} = 3.5903 \); residual: \( u_{Ita} = y_{Ita} - \hat{y}_{Ita} = -1.5903 \). The Italian level of intergenerational mobility is somewhat lower than its theoretical level (conditional mean) corresponding the value of the independent variable.

Useful formulas and values: \( n = 18, df = n - 2 = 16, \hat{\beta}_1 = \text{cod}(x, y)/\text{dev}(x) = -0.10765, \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 7.035, \)

\( \text{dev}(\text{RES}) = \text{dev}(y) - \hat{\beta}_1^2/\text{dev}(x) = 26.9002, \sigma^2 = \text{dev}(\text{RES})/df = 1.6813, \sigma = 1.2966, se(\hat{\beta}_1) = \sigma/\sqrt{\text{dev}(x)} = 0.0318, se(\hat{\beta}_0) = \sigma \sqrt{1/n + \bar{x}^2/\text{dev}(x)} = 1.2165 \).

Exercise 2. Assumptions: \( X_1 = \text{'intergenerational mobility in 2005'} \sim N(\mu_1, \sigma_1^2), X_2 = \text{'intergenerational mobility in 2010'} \sim N(\mu_2, \sigma_2^2) \). Since the sample sized are large enough, the normality assumptions are not strictly needed.

(A) Test of \( H_0: \mu_2 - \mu_1 = 0 \) against \( H_1: \mu_2 - \mu_1 > 0 \); test statistic: \((\bar{X}_2 - \bar{X}_1 - 0)/se \) with approximated distribution \( N(0,1) \) under \( H_0 \); rejection region for \( \alpha = 0.05 \): \( z_{crit} = 1.6449, +\infty \); sample value of the test statistic (under \( H_0 \)): \( z_{samp} = 2.0571 \).

(B) \( H_1: \mu_2 - \mu_1 = 0.31, \beta = P(\text{test statistic} \in \text{A}|H_1) = P((\bar{X}_2 - \bar{X}_1)/se < z_{crit}|H_1) = P(\bar{X}_2 - \bar{X}_1 < z_{crit}se|H_1) = P((\bar{X}_2 - \bar{X}_1 - 0.31)/se < (z_{crit}se - 0.31)/se|H_1) = P(Z < -1.267) = 0.1026 \).

Useful formulas and values:

\( \bar{n}_1 = 209, n_2 = 430, \bar{x}_1 = 2.099, \bar{x}_2 = 2.318, s_1^2 = 1.555009, s_2^2 = 1.674436, \bar{x}_2 - \bar{x}_1 = 0.219, se = se(\bar{X}_2 - \bar{X}_1) = \sqrt{s_2^2/n_2 + s_1^2/n_1} = \sqrt{0.011334} = 0.1065 \).
8  Test 22.01.2013

Text

Framework: The perception of corruption among different countries.

Exercise 1. *Transparency international*\(^2\) publishes each year an international comparison on the level of perceived corruption. We aim at investigating if this variable is related to the deficiencies of the judicial system in terms of length of proceedings. The following table, computed on 17 European Union countries, summarizes some statistics concerning the variables ln(CPI) (as measure the perceived corruption in 2012) and ln(V/P) (as proxy for the possible deficiencies of the judicial system during 1999 – 2006). Legend: CPI = Corruption Perceptions Index (the values of the original work have been transformed in order to simplify interpretation); V = number of violations verified by the European Court of Human Rights concerning the *length of proceedings*; P = total population in millions people.\(^3\)

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>standard deviation</th>
<th>correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ln(CPI)</td>
<td>3.415</td>
<td>0.588</td>
<td>0.609</td>
</tr>
<tr>
<td>ln(V/P)</td>
<td>0.614</td>
<td>1.577</td>
<td></td>
</tr>
</tbody>
</table>

Formulate an appropriate statistical model and answer the following questions.

(A) Test whether the variable related to the weaknesses of judicial system is significant in the regression (\(\alpha = 0.02\)).

(B) Compute the residual for Italy (CPI = 58, V/P = 15.23) and its standardized transformation.

(C) (No for students with 6 CFU) Compute the confidence interval (\(\alpha = 0.01\)) for the standard deviation of the error component.

(D) (No for students with 6 CFU) Compute the confidence interval (\(\alpha = 0.02\)) for the conditional mean of CPI for a country having V/P = 4.07.

Exercise 2. (No for students with 3 CFU). *Transparency international* gives the following statistics (referred to 2012) on CPI:

<table>
<thead>
<tr>
<th>Country</th>
<th>Observations</th>
<th>Mean</th>
<th>Standard error of the sample mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Italy</td>
<td>7</td>
<td>58</td>
<td>2.4</td>
</tr>
<tr>
<td>Austria</td>
<td>8</td>
<td>31</td>
<td>2.4</td>
</tr>
</tbody>
</table>

Assuming that the two samples are independent and that the corresponding random variables are Normally distributed with the same variance, answer the following questions:

(A) Estimate the mean difference of CPI between the two countries; complement the estimate with the standard error of the estimator used.

(B) Test whether the Italian level of perceived corruption is significantly higher that in the other country (\(\alpha = 0.02\)).

(C) Compute the log-likelihood for the last observation of the Austria sample, namely CPI = 25, considering parameters that can be retrieved from the reported statistics.

(D) (No for students with 6 CFU). Exploiting the information available for Austria, compute how many observations would be necessary for getting a confidence interval for the mean of size 2.81 at a confidence level 0.98.

Solution

Exercise 1. Simple Linear Regression Model \(y = \beta_0 + \beta_1 x + u\) with the classical assumptions on u, where \(y = \ln(CPI)\) and \(x = \ln(V/P)\).
(A) Test of $H_0: \beta_1 = 0$ against $H_1: \beta_1 \neq 0$; test statistic: $(\hat{\beta}_1 - 0)/se(\hat{\beta}_1)$ with distribution $T(n-2)$; acceptation region for $\alpha = 0.02$: $(-2.602, 2.602)$; sample value of the test statistic (under $H_0$): $t_{samp} = 2.9737$.

(B) Fitted value: $\hat{y}_{Ita} = \hat{\beta}_0 + \hat{\beta}_1 x_{Ita} = 3.894$; residual: $\hat{u}_{Ita} = y_{Ita} - \hat{y}_{Ita} = 0.1665$ ($x_{Ita} = \ln(15.23) = 2.72327$, $y_{Ita} = \ln(58) = 4.06044$); standardized residual: $\hat{u}_{Ita}/\hat{\sigma} = 0.3353$

(C) Pivot for $\sigma^2$: $\hat{\sigma}^2/(n-2)/\sigma^2 \sim \chi^2(n-2)$; interval at $(1 - \alpha) = 0.99$ for $\sigma^2$: $[\hat{\sigma}^2/(n-2)/c_2, \hat{\sigma}^2(n-2)/c_1] = [0.1127, 0.8037]$, where $c_1 = 4.6009$ and $c_2 = 32.8013$; corresponding interval for $\sigma$: $[0.3358, 0.8965]$.

(D) Taking $x_0 = \ln(4.07) = 1.4036$, we denote the conditional mean as $\mu = E(y|x_0) = \beta_0 + \beta_1 x_0$. Pivot for $\mu_0$: $(\hat{\mu}_0 - \mu_0)/se(\hat{\mu}_0) \sim T(n-2)$; interval at $(1 - \alpha) = 0.98$ for $\mu_0$: $[\hat{\mu}_0 - tse(\hat{\mu}_0), \hat{\mu}_0 + tse(\hat{\mu}_0)] = [3.2438, 3.9448]$, where $t = 2.6025$.

Useful formulas and values: $n = 17$, $df = n - 2 = 15$, $\bar{x} = 0.614$, $\bar{y} = 3.415$, $dev(x) = 42.2778$, $dev(y) = 5.8776$, $cod(x, y) = 9.6001$, $\hat{\beta}_1 = \text{cod}(x, y)/\text{dev}(x) = 0.22707$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 3.2756$, $dev(RES) = dev(y) - \hat{\beta}_1^2 dev(x) = 3.6977$, $\hat{\sigma}^2 = dev(RES)/n = 0.2465$, $\sigma = 0.4965$, $se(\hat{\beta}_1) = \hat{\sigma}/\sqrt{\text{dev}(x)} = 0.0764$, $\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = 3.5943$, $se(\hat{\mu}_0) = \hat{\sigma}/\sqrt{1/n + (x - x_0)^2/\text{dev}(x)} = 0.1347$.

**Exercise 2.** Assumptions: $X_1 = \ln(CPI_{Austria}) \sim N(\mu_1, \sigma^2)$, $X_2 = \ln(CPI_{Austria}) \sim N(\mu_2, \sigma^2)$; independent samples. Because of assumptions, for making inference on $\mu_1 - \mu_2$ we use $\overline{X}_1 - \overline{X}_2$ whose distribution is $[(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)]/se \sim T(n_1 + n_2 - 2)$, where $se = S_{pooled}\sqrt{1/n_1 + 1/n_2}$ and $S_{pooled}^2 = [(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2]/(n_1 + n_2 - 2)$.

(A) $\mu_1 - \mu_2$ can be estimated by $\overline{X}_1 - \overline{X}_2 = 27$; the corresponding standard error is $se = S_{pooled}\sqrt{1/n_1 + 1/n_2} = 3.4104$.

(B) Test of $H_0: \mu_1 - \mu_2 = 0$ against $H_1: \mu_1 - \mu_2 > 0$; test statistic: $(\overline{X}_1 - \overline{X}_2 - 0)/se$ with distribution $T(n_1 + n_2 - 2)$ under $H_0$; rejection region for $\alpha = 0.02$: $(t_{crit} = 2.2816, +\infty)$; sample value of the test statistic (under $H_0$): $t_{samp} = 7.917$.

(C) $\ln f(x_{Austria}) = 25; \mu = 31, \sigma^2 = 46.08 = \ln(0.03977) = -3.2248$, where $f(x; \mu, \sigma^2)$ denotes the formula of the p.d.f. of the normal distribution.

(D) $n = (2\sigma/A)^2 = 11.2397^2 = 126.3308 \approx 127$, where $\alpha = 0.02$, $z = 2.3263$, $\sigma = 6.78823$, $A = 2.81$.

Useful formulas and values: $n_1 = 7, n_2 = 8, \overline{X}_1 = 58, \overline{X}_2 = 31, s_1 = 6.3498, s_2 = 6.7882, s_1^2 = 40.32, s_2^2 = 46.08, s_{pooled} = \sqrt{43.4215} = 6.5895$, $se = \sqrt{1.6309} = 3.4104$. 
9 Test 19.02.2013

Text

Framework: Transparency international publishes each year statistics concerning the Corruption Perception Index (CPI score: higher values indicate less corruption): the statistics cover almost all countries allowing international comparisons. Although coming from a poor situation, the Italian condition is further deteriorated in the last few years, moving from a rank 40 in 2007 to a rank 65 in 2011 (considering only countries surveyed in both years). The issue of interest is to check if the Italian situation got worse because the other countries improved their situation or, on the opposite, because the level of perceived corruption is really increased within the country.

Exercise 1. The first step of the analysis is to investigate if the perception of corruption has globally changed between the two years considered. We do this with a linear regression model in which the 2011 CPI score is regressed against the corresponding 2007 values: if nothing has changed (on average) between the two years, the regression line should be the bisector of the 1-st quadrant. Considering the statistics displayed in the following table (122 sample observations) answer the questions.

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>standard deviation</th>
<th>correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPI score 2007</td>
<td>4.013</td>
<td>2.165</td>
<td>0.975</td>
</tr>
<tr>
<td>CPI score 2011</td>
<td>4.05</td>
<td>2.165</td>
<td></td>
</tr>
</tbody>
</table>

(A) (B) Formulate appropriately the null hypotheses (and the corresponding alternative) needed for testing if nothing has changed in the period considered; then do the appropriate tests ($\alpha = 0.05$). Comment the results if at least one of the tests leads to a rejection.

(C) By means of an appropriate index, evaluate how well the model fit the data and interpret the result obtained.

Exercise 2. (No for students with 3 CFU). The second step of the analysis is to check if the Italian situation made worse between the two years considered. The following table summarizes a few statistics (the surveys are lead independently; the CPI score is an average among the different surveys used):

<table>
<thead>
<tr>
<th>Year</th>
<th>Surveys used</th>
<th>CPI score</th>
<th>Confidence interval of the CPI mean at 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>6</td>
<td>5.2</td>
<td>[4.7, 5.7]</td>
</tr>
<tr>
<td>2011</td>
<td>9</td>
<td>3.9</td>
<td>[3.2, 4.8]</td>
</tr>
</tbody>
</table>

Assuming that CPI is Normally distributed, answer the following questions:

(A) Estimate the standard deviation in each one of the two years considered.

(B) (Only for students with 9 CFU). Compute the confidence interval ($\alpha = 0.01$) for the standard deviation in 2011.

(C) Test whether the Italian perception of corruption is significantly increased in the period ($\alpha = 0.05$).

Exercise 3. (No for students with 3 CFU) The following table reports the number of components elected in 2008 in the two major parties of the Italian Parliament (Camera dei deputati + Senato), together with the corresponding number of convicted (prescribed or not) or under judicial investigation.

<table>
<thead>
<tr>
<th></th>
<th>Elected</th>
<th>Convicted or under judicial investigation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partito delle Libertà (PdL)</td>
<td>421</td>
<td>59</td>
</tr>
<tr>
<td>Partito Democratico (PD)</td>
<td>336</td>
<td>15</td>
</tr>
</tbody>
</table>

(A) Do the proportions of people convicted or under judicial investigation differ significantly between the two parties ($\alpha = 0.01$)?

(B) Compute the probability of the second kind error of the test at point (A) by considering an alternative hypothesis under which the two percentages differ of 0.063 points.

\[http://www.transparency.org/\]
\[http://www.ilfattoquotidiano.it/2012/09/30/cento-parlamentari-condannati-imputati-indagati-o-prescritti/368539/\]
\[http://it.wikipedia.org/wiki/Senatori_della_XVI_Legislatura_della_Repubblica_Italiana\]
\[http://it.wikipedia.org/wiki/Deputati_della_XVI_Legislatura_della_Repubblica_Italiana\]
Solution

Exercise 1. Simple Linear Regression Model \( y = \beta_0 + \beta_1 x + u \) with the classical assumptions on \( u \), where \( y = \text{CPI score 2011} \) and \( x = \text{CPI score 2007} \).

(A) Test \( H_0 : \beta_0 = 0 \) against \( H_1 : \beta_0 \neq 0 \); test statistic: \( \hat{\beta}_0 - 0 / \text{se}(\hat{\beta}_0) \) with distribution \( T(n-2) \); acceptance region for \( \alpha = 0.05 \): \((-1.98, 1.98)\); sample value of the test statistic (under \( H_0 \)): \( t_{\text{amp}} = 1.48472 \).

(B) Test \( H_0 : \beta_1 = 1 \) against \( H_1 : \beta_1 \neq 1 \); test statistic: \( \hat{\beta}_1 - 1 / \text{se}(\hat{\beta}_1) \) with distribution \( T(n-2) \); acceptance region for \( \alpha = 0.05 \): \((-1.98, 1.98)\); sample value of the test statistic (under \( H_0 \)): \( t_{\text{amp}} = -1.23247 \).

(C) \( R^2 = \text{dev}(\text{REG}) / \text{dev}(\bar{y}) = 0.950625 \).

Useful formulas and values: \( n = 122, df = n - 2 = 120, \bar{x} = 4.013, \bar{y} = 4.05, \text{dev}(\bar{x}) = 571.8415, \text{dev}(\bar{y}) = 571.8415, \text{cod}(\bar{x}, \bar{y}) = 557.5454, \hat{\beta}_1 = \text{cod}(\bar{x}, \bar{y}) / \text{dev}(\bar{x}) = 0.975, \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 0.1373, \text{dev}(\text{RES}) = \text{dev}(\bar{y}) - \hat{\beta}_1 \text{dev}(\bar{x}) = 28.2347, \hat{\sigma}^2 = \text{dev}(\text{RES}) / df = 0.2353, \hat{\sigma} = 0.4851, \text{se}(\hat{\beta}_1) = \hat{\sigma} / \sqrt{\text{dev}(\bar{x})} = 0.0203, \text{se}(\hat{\beta}_0) = \hat{\sigma} \sqrt{1 / n + \bar{x}^2 / \text{dev}(\bar{x})} = 0.0925, \text{dev}(\text{REG}) = \text{dev}(\bar{y}) - \text{dev}(\text{RES}) = 543.6068 \).

Exercise 2. Assumptions: \( X_1 = \text{'Italian CPI in 2007'} \sim N(\mu_1, \sigma^2), X_2 = \text{'Italian CPI in 2011'} \sim N(\mu_2, \sigma^2) \); independent samples.

(A) \( \sigma^2_1 \) can be estimated by means of the unbiased sample variance \( s^2_1 \), so that the estimate of \( \sigma_1 \) is \( s_1 = 0.4764 \); the same for \( s_2 = 1.0408 \) as estimate of \( \sigma_2 \). The computation is as follows (we omit indices): 1) compute the size of the confidence interval \( A \) from the data; 2) this size is related directly with \( s \) by means of the relationship \( A = 2s / \sqrt{n} \), so that the wished estimate is \( s = A \sqrt{n} / (2t) \). Useful formulas and values: \( n_1 = 6, A_1 = 1, t_1 = 2.5706, n_2 = 9, A_2 = 1.6, t_2 = 2.306 \).

(B) Interval for \( \sigma^2 \) in 2011 (for sake of simplicity we omit the index 2). Pivot \( (n - 1)S^2 / \sigma^2 \) with distribution \( \chi^2(n - 1) \); confidence interval at \( \alpha = 0.01 \) for \( \sigma^2 \): [0.3947, 6.4455]; corresponding confidence interval for \( \sigma \): [0.6282, 5.3538].

(C) Test of \( H_0 : \mu_2 - \mu_3 = 0 \) against \( H_1 : \mu_2 - \mu_3 > 0 \); assuming \( \sigma^2_1 = \sigma^2_2 \), the test statistic is \( (\bar{X}_1 - \bar{X}_2 - 0) / \text{se} \) (\( \text{se} = \text{S}_{\text{pooled}} \sqrt{1/n_1 + 1/n_2} \)) and \( S^2_{\text{pooled}} = [(n_1 - 1)S^2_1 + (n_2 - 1)S^2_2] / (n_1 + n_2 - 2) \) with distribution \( T(n_1 + n_2 - 2) \) under \( H_0 \); rejection region for \( \alpha = 0.05 \): \( \{t_{\text{crit}} = 1.7709, +\infty\} \); sample value of the test statistic (under \( H_0 \)): \( t_{\text{amp}} = 2.8408 \). Useful formulas and values: \( \bar{X}_1 = 5.2, \bar{X}_2 = 3.9, \text{S}_{\text{pooled}} = \sqrt{0.7539} = 0.8683, \sqrt{\text{se}} = \sqrt{0.2094} = 0.4576 \).

Exercise 3. Assumptions: \( X_1 = \text{'Convicted or under judicial investigation in PdL'} \sim Be(p_1), X_2 = \text{'Convicted or under judicial investigation in PD'} \sim Be(p_2) \); independent samples.

(A) Test of \( H_0 : p_1 - p_2 = 0 \) against \( H_1 : p_1 - p_2 \neq 0 \); the test statistic is \( (\bar{X}_1 - \bar{X}_2 - 0) / \text{se} \) (\( \text{se} = \sqrt{\bar{p}_1q_1/n_1 + \bar{p}_2q_2/n_2} \)) with distribution \( N(0, 1) \) under \( H_0 \); rejection region for \( \alpha = 0.01 \): \((-z_{\text{crit}} = -2.576, z_{\text{crit}} = 2.576)\); sample value of the test statistic (under \( H_0 \)): \( z_{\text{amp}} = 4.6983 \).

(B) \( \beta \) of the test for \( H_1 : p_1 - p_2 = 0.063 \). \( \beta = P(\text{sample} \in A|H_1) = P[z_{\text{crit}} \leq (\hat{p}_1 - \hat{p}_2 - 0) / \text{se} \leq z_{\text{crit}}|H_1] = P[z_{\text{crit}} \leq \hat{p}_1 - \hat{p}_2 \leq z_{\text{crit}}|H_1] = P[-z_{\text{crit}} \leq \hat{p}_1 - \hat{p}_2 \leq z_{\text{crit}}|H_1] = P[-z_{\text{crit}} \leq \hat{p}_1 - \hat{p}_2 \leq z_{\text{crit}}|H_1] = P[-0.524|H_1] = 0.30028 - 0 = 0.30028 \).

Useful formulas and values: \( n_1 = 421, \hat{p}_1 = 0.14014, \hat{q}_1 = 0.85986, n_2 = 336, \hat{p}_2 = 0.04464, \hat{q}_2 = 0.95536, \text{se} = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} = \sqrt{0.00029 + 0.00013} = \sqrt{0.00041} = 0.2033 \).
Text
Framework: The school dropout is one important Italian social issue: within the EU27, in 2012 Italy has the fourth highest school dropout rate.

Exercise 1. We aim at investigating if the school dropout (taken as dependent variable) is related to level of wealth, proxied by means of the per-capita GNP. The following table gives some statistics computed in 2007 on 18 Italian regions (pcGNP = per-capita GNP (Euro); DR = dropout rate (Percentage); ln is the natural logarithm).

<table>
<thead>
<tr>
<th>Mean vector</th>
<th>Variance - covariance matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>ln(pcGNP)</td>
<td>ln(pcGNP)</td>
</tr>
<tr>
<td>10.058</td>
<td>0.07045</td>
</tr>
<tr>
<td>DR</td>
<td>DR</td>
</tr>
<tr>
<td>18.294</td>
<td>-0.647</td>
</tr>
</tbody>
</table>

(A) Test whether the dropout rate is significantly related to the per-capita GNP ($\alpha = 0.05$).
(B) Compute the confidence interval for ($\alpha = 0.05$) for the standard deviation of the error component.
(C) Compute the 3 deviances involved in the deviance decomposition in a regression framework.
(D) Compute fitted value and residual for Lombardia (Per-capita GDP = 31848 Euro, School dropout rate = 18.3%).

Exercise 2. Although high in comparison with other countries, the Italian school dropout tended to decrease in recent years. The following table reports some statistics (the last row reports statistics computed on the corresponding difference) computed on the same 18 regions.

<table>
<thead>
<tr>
<th>Period</th>
<th>Mean</th>
<th>Unbiased sample variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>20.756</td>
<td>5.754</td>
</tr>
<tr>
<td>2011</td>
<td>17.083</td>
<td>4.252</td>
</tr>
<tr>
<td>(2011 − 2004)</td>
<td>−3.672</td>
<td>2.928</td>
</tr>
</tbody>
</table>

(Only for students with 9 CFU) Assuming that the dropout rate is Normally distributed (a not completely justified assumption), answer the following questions:
(A) Test whether the dropout rate has decreased, in mean, more than 2 percentage points ($\alpha = 0.05$).
(B) Estimate the correlation coefficient of the dropout rate between the two years.

Exercise 3. (No for students with 3 CFU) Two college classes, the first one from a Liceo Scientifico (S), the second one from a Liceo Classico (C), are composed by 20 students each one. The probability that a students drops out is 7.6% in the Liceo Scientifico and 14.4% in the Liceo Classico. Dropouts in the two classes are independent.

(A) Compute the mean and the variance of the sum of dropouts in the two classes.
(B) Compute the probability of observing exactly 3 dropouts in the same class.

Solution

Exercise 1. Simple Linear Regression Model $y = \beta_0 + \beta_1 x + u$ with the classical assumptions on $u$, where $y = DR$ and $x = pcGNP$ by region.

(A) Test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$; test statistic: $(\hat{\beta}_1 - 0)/se(\hat{\beta}_1)$ with distribution $T(n - 2)$; acceptance region for $\alpha = 0.05$: $(-2.12, 2.12)$; sample value of the test statistic (under $H_0$): $t_{samp} = -2.19847$.

(B) Pivot for $\sigma^2$: $\hat{\sigma}^2(n - 2)/\sigma^2 \sim \chi^2(n - 2)$; interval at $(1 - \alpha) = 0.95$ for $\sigma^2$: $[\hat{\sigma}^2(n - 2)/c_2, \hat{\sigma}^2(n - 2)/c_1] = [12.2745, 51.2563]$, where $c_1 = 6.9077$ and $c_2 = 28.8454$; corresponding interval for $\sigma$: $[3.5035, 7.1593]$. 
(C) $dev(y) = n \cdot \text{var}(y) = 461.016$, $dev(\text{REG}) = \hat{\beta}_1^2 \cdot dev(\bar{x}) = 106.9547$, $dev(\text{RES}) = dev(y) - dev(\text{REG}) = 354.0613$.

(D) Fitted value: $\hat{y}_{\text{Lombardia}} = \hat{\beta}_0 + \hat{\beta}_1 \ln(x_{\text{Lombardia}}) = 15.4403$; residual: $u_{\text{Lombardia}} = y_{\text{Lombardia}} - \hat{y}_{\text{Lombardia}} = 2.8597$, where $(x_{\text{Lombardia}} = 31848, y_{\text{Lombardia}} = 18.3)$.

Useful formulas and values: $n = 18$, $df = n - 2 = 16$, $\bar{x} = 10.058$, $\bar{y} = 18.294$, $dev(\bar{x}) = 1.2681$, $dev(y) = 461.016$, $\text{cod}(\bar{x}, y) = -11.646$, $\hat{\beta}_1 = \text{cod}(\bar{x}, y) / dev(\bar{x}) = -9.18382$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 110.6648$, $dev(\text{RES}) = dev(y) - \hat{\beta}_1^2 \cdot dev(\bar{x}) = 354.0613$, $\hat{\sigma}^2 = dev(\text{RES}) / df = 22.1288$, $\hat{\sigma} = 4.7041$, $\text{se}(\hat{\beta}_1) = \hat{\sigma} / \sqrt{dev(\bar{x})} = 4.1774$.

Exercise 2. Assumptions: $X_1$ = 'Italian dropout rate in 2004', $X_2$ = 'Italian dropout rate in 2011'. Because the sample is the same at two different times we work on paired data $D = X_2 - X_1 \sim N(\mu_D, \sigma_D^2)$.

(A) Test $H_0 : \mu_D = -2$ against $H_1 : \mu_D < -2$; test statistic: $(\bar{d} - -2) / \text{se}(\bar{d})$ with distribution $T(n - 1)$; acceptance region for $\alpha = 0.05$: $(-\infty, -1.7396)$; sample value of the test statistic (under $H_0$): $t_{\text{samp}} = -2.4227$.

Useful formulas and values: $n = 18$, $\bar{d} = -3.672$, $s_D = 2.928$, $\text{se}(\bar{d}) = s_D / \sqrt{n} = 0.6901$.

(B) Since $V(D) = V(X_2 - X_1) = V(X_2) + V(X_1) - 2C(X_1, X_2)$ then $C(X_1, X_2) = [V(X_2) + V(X_1) - V(D)] / 2$. But $C(X_1, X_2) = \rho / (\sigma(X_1)\sigma(X_2))$ and then $\rho = [V(X_2) + V(X_1) - V(D)] / (2\sigma(X_1)\sigma(X_2)) = [4.252^2 + 5.754^2 - 5.754^2] / (2 \cdot 5.754 \cdot 4.252) = 0.8709$.

Exercise 3. Assumptions: $X_1$ = 'Number of dropouts in the LS' $\sim \text{Bin}(n_1 = 20, p_1 = 0.076)$, $X_2$ = 'Number of dropouts in the LC' $\sim \text{Bin}(n_2 = 20, p_2 = 0.144)$, $X_1$ and $X_2$ independent.

(A) $E(X_1 + X_2) = E(X_1) + E(X_2) = 1.52 + 2.88 = 4.4$, $V(X_1 + X_2) = V(X_1) + V(X_2) = 1.4045 + 2.4653 = 3.8698$, where $E(X_i) = n_i p_i$ and $V(X_i) = n_i p_i (1 - p_i)$ for $i = 1, 2$.

(B) Let $c = 3$. Then $P((X_1 = c, X_2 = 0) \cup (X_1 = 0, X_2 = c)) = P(X_1 = c, X_2 = 0) + P(X_1 = 0, X_2 = c) = P(X_1 = c) P(X_2 = 0) + P(X_1 = 0) P(X_2 = c) = 0.13055 \cdot 0.04461 + 0.2058 \cdot 0.24213 = 0.05565$, where the probabilities are computed with the Binomial p.m.f.
11  Test 18.07.2013

Text

Framework: A public administration planned to improve the management of the municipal waste.

Exercise 1. (No for students with 3 CFU)

An important issue is the variability of the quantity stored in the dumpers (or near them): if too high, this indicates that the emptying trucks can find too few (and in this case they are inefficient) or too much waste (and in this case further travels are needed). The following data report some statistics computed on a sample of dumpers of the same size (including also waste laid down – measures in thousands liters).

<table>
<thead>
<tr>
<th>Observations</th>
<th>Min</th>
<th>Median</th>
<th>Mean</th>
<th>Max</th>
<th>Variance (unbiased)</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
<td>0.163</td>
<td>1.639</td>
<td>1.687</td>
<td>3.72</td>
<td>0.723</td>
</tr>
</tbody>
</table>

Assuming that the quantity of waste is, approximately, Normally distributed, answer the following questions:

(A) Test the hypothesis that the standard deviation is significantly larger than 0.69 ($\alpha = 0.05$).

(B) (No for students with 6 CFU) Compute the confidence interval for the standard deviation at the same $\alpha$.

Exercise 2. (No for students with 3 CFU)

The administration decorated some garbage dumpsters of the humid fraction with the idea that this could incentivate the separate collection of this portion of waste. The following table summarizes one year of data (values are the humid fraction in percentage) collected on two different samples.

<table>
<thead>
<tr>
<th>With decorated dumpsters</th>
<th>37</th>
<th>22</th>
<th>35</th>
<th>31</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>With standard dumpsters</td>
<td>24</td>
<td>34</td>
<td>33</td>
<td>20</td>
<td>31</td>
</tr>
</tbody>
</table>

Assuming that the percentage fraction is Normally distributed without any assumption on the variances, answer the following questions:

(A) Compute a point estimate of the difference of the percentage fraction between the different dumpsters; complement the estimate with the standard error of the estimator employed.

(B) Test whether the mean humid fraction is significantly different between the two samples ($\alpha = 0.1$).

(C) (No for students with 6 CFU) Using the sample information of decorated dumpsters, compute how many sample units are needed for getting a confidence interval for the mean fraction of size 1.7 at the confidence level 90%.

Exercise 3. (No for students with 3 CFU)

Let $X$ a continuous r.v. such that $X \sim \text{Uniform}(1, 4)$.

(A) Compute the p.d.f. of $X$ at 2.6 and 4.7.

(B) Compute the c.d.f. of $X$ at the same values.

(C) Compute the probability of $X > 3.3$ knowing that $X \in [2.5, 3.6]$.

Solution

Exercise 1. Assumptions: $X = \text{‘Quantity of waste in one dumper’} \sim N(\mu, \sigma^2)$.

(A) Test $H_0 : \sigma = 0.69$ against $H_1 : \sigma > 0.69$; test statistic: $(n - 1)S^2/\sigma_0^2$, where $\sigma_0 = 0.69$, with distribution $\chi^2(n - 1)$; rejection region for $\alpha = 0.05$: $(43.773, +\infty)$; sample value of the test statistic (under $H_0$): $\chi^2_{\text{samp}} = 45.5577$.

(B) Pivot for $\sigma^2$: $\hat{\sigma}^2(n - 1)/\sigma^2 \sim \chi^2(n - 1)$; interval at $(1 - \alpha) = 0.95$ for $\sigma^2$: $[\hat{\sigma}^2(n - 1)/c_2, \hat{\sigma}^2(n - 1)/c_1] = [0.4617, 1.2918]$, where $c_1 = 16.7908$ and $c_2 = 46.9792$; corresponding interval for $\sigma$: $[0.6795, 1.1366]$.

Useful formulas and values: $n = 31$, $s^2 = 0.723$.  

22
Exercise 2. Assumptions: $X_1 =$ fraction in decorated $N(\mu_1, \sigma_1^2)$, $X_2 =$ fraction in standard $N(\mu_2, \sigma_2^2)$, $X_1$ and $X_2$ independent.

(A) Point estimator of $\mu_1 - \mu_2$: $\overline{X}_1 - \overline{X}_2$; corresponding estimate $\bar{p}_1 - \bar{p}_2 = 3$; standard error: $se = \sqrt{\overline{V}(\overline{X}_1 - \overline{X}_2)} = \sqrt{se_1^2 + se_2^2} = \sqrt{14.12} = 3.7577$.

(B) Test $H_0 : \mu_1 - \mu_2 = 0$ against $H_1 : \mu_1 - \mu_2 \neq 0$; test statistic by Satterthwaite-Welsh: $(\overline{X}_1 - \overline{X}_2 - 0)/se$ with distribution $T(df \backsimeq 8)$; acceptance region for $\alpha = 0.1$: $(-1.8595, 1.8595)$; sample value of the test statistic (under $H_0$): $t_{samp} = 0.7984$.

(C) $n = (2z\sigma/A)^2 = 11.1668^2 = 124.6984 \backsimeq 125$, where $z = 1.645$ ($\alpha = 0.1$), $\sigma$ is estimated by $s_1 = 5.7706$ and $A = 1.7$.

Useful formulas and values:

<table>
<thead>
<tr>
<th>$x_{1,i}$</th>
<th>37</th>
<th>22</th>
<th>35</th>
<th>31</th>
<th>32</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_{1,i}^2$</td>
<td>1369</td>
<td>484</td>
<td>1225</td>
<td>961</td>
<td>1024</td>
<td>5063</td>
</tr>
<tr>
<td>$x_{2,i}$</td>
<td>24</td>
<td>34</td>
<td>33</td>
<td>20</td>
<td>31</td>
<td>142</td>
</tr>
<tr>
<td>$x_{2,i}^2$</td>
<td>576</td>
<td>1156</td>
<td>1089</td>
<td>400</td>
<td>961</td>
<td>4182</td>
</tr>
</tbody>
</table>

$n_1 = 5$, $n_2 = 5$, $\bar{p}_1 = 31.4$, $\bar{p}_2 = 28.4$, $dev_1 = 133.2$, $dev_2 = 149.2$, $s_1^2 = 33.3$, $s_2^2 = 37.3$, $s_1 = 5.7706$, $s_2 = 6.1074$, $se_1 = s_1/\sqrt{n_1} = 2.5807$, $se_1 = s_2/\sqrt{n_2} = 2.7313$, $se_1^2 = 6.66$, $se_2^2 = 7.46$, $df = \frac{(se_1^2 + se_2^2)^2}{\frac{se_1^2}{n_1 - 1} + \frac{se_2^2}{n_2 - 1}} = 7.9744 \backsimeq 8$.

Exercise 3. Assumptions: $X \sim$ Uniform$(1,4)$.

(A) $f(2.6) = 0.3333$, $f(4.7) = 0$.

(B) $F(2.6) = 0.5333$, $F(4.7) = 1$.

(C) Let $c_1 = 3.3$, $c_2 = 2.5$, $c_3 = 3.6$.

$P(X > c_1|c_2 \leq X \leq c_3) = P(X > c_1 \cup c_2 \leq X \leq c_3)/P(c_2 \leq X \leq c_3) = P(c_1 < X \leq c_3)/P(c_2 \leq X \leq c_3) = (P(X \leq c_3) - P(X \leq c_1))/P(X \leq c_3) = 0.2727$, where $P(X \leq c_1) = 0.7667$, $P(X \leq c_2) = 0.5$, $P(X \leq c_3) = 0.8667$. 

23
12 Test 11.09.2013

Text

Framework: An international program labeled MPAC (Music Practice Against Crime) aims at reducing the youth crime rate by means of an educational program based on music practice.

Exercise 1. (No for students with 3 CFU)
The following table shows some statistics relative to two samples of towns after some years since the beginning of the MPAC program. The variable of interest is the number of crimes, every 1000 habitants, committed by people aged less than 18 years.

<table>
<thead>
<tr>
<th>Towns inserted in MPAC</th>
<th>5.3</th>
<th>5.7</th>
<th>7.4</th>
<th>8.7</th>
<th>9.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other towns</td>
<td>13.7</td>
<td>11.8</td>
<td>11</td>
<td>13.2</td>
<td>16.5</td>
</tr>
</tbody>
</table>

Assuming that the variable of interest is, approximately, Normally distributed with equal standard deviation, answer the following questions:

(A) Compute a point estimate of all parameters.

(B) Compute the value of the log-likelihood of the first observation of the sample at the estimates of point A).

(C) (No for students with 6 CFU) Compute an interval estimation of the common standard deviation ($\alpha = 0.02$).

(D) Test whether the program has been effective in reducing the crime rate ($\alpha = 0.05$).

Exercise 2.
Another interesting point is to quantify the contribution of the money invested in the MPAC program in terms of reduction of the crime rate (taken as dependent variable).

<table>
<thead>
<tr>
<th>Investments ($ per habitant)</th>
<th>14.7</th>
<th>12.8</th>
<th>8.8</th>
<th>11</th>
<th>7.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crimes (number per 1000 habitants)</td>
<td>5.3</td>
<td>5.7</td>
<td>7.4</td>
<td>8.7</td>
<td>9.1</td>
</tr>
</tbody>
</table>

(A) Test whether the number of crimes depends significantly on the money invested in ($\alpha = 0.05$). How much each dollar invested in (per habitant) reduces, an average, the number of crimes (per 1000 habitants)?

(B) (No for students with 6 CFU) Compute the confidence interval at $\alpha = 0.01$ for the conditional mean of the dependent variable for a value of the independent equal to 13.2.

(Hint: note that the data on the number of crimes is the same as in exercise 1).

Exercise 3. (No for students with 3 CFU)
The joint p.m.f. of $X$ and $Y$ is given by

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y = 3$</th>
<th>$Y = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.08</td>
<td>0.36</td>
</tr>
<tr>
<td>2</td>
<td>0.3</td>
<td>0.26</td>
</tr>
</tbody>
</table>

(A) Compute the correlation coefficient between the two variables.

(B) Compute the mean and the standard deviation of $Y|X = 0$.

Solution

Exercise 1. Assumptions: $X_1 =$’Crimes (per 1000 habitants) in towns inserted in MPAC’~ $N(\mu_1, \sigma^2)$; $X_2 =$’Crimes (per 1000 habitants) in other towns’~ $N(\mu_2, \sigma^2)$; independent samples.

(A) $\hat{\mu}_1 = \overline{x}_1 = 7.24, \hat{\mu}_2 = \overline{x}_2 = 13.24, \hat{\sigma} = s_{pooled} = 1.9263$.

(B) log-likelihood of 1-st observation for parameters equal to their estimated values = $l(x_{1,1}; \hat{\mu}_1, \hat{\sigma}^2) = -\frac{1}{2} \ln(2\pi\hat{\sigma}^2) - \frac{1}{2} \frac{(x_{1,1} - \hat{\mu}_1)^2}{\hat{\sigma}^2} = -2.08168$, where $x_{1,1} = 5.3$ and the parameter estimated are given above.
(C) Pivot for $\sigma^2$: $S_{pooled}^2(n_1 + n_2 - 2) / \sigma^2 \sim \chi^2(n_1 + n_2 - 2)$; interval at $(1 - \alpha) = 0.98$ for $\sigma^2$: $[S_{pooled}^2(n_1 + n_2 - 2) / c_2, S_{pooled}^2(n_1 + n_2 - 2) / c_1] = [1.4775, 18.0286]$, where $c_1 = 1.6465$ and $c_2 = 20.0902$; corresponding interval for $\sigma$: $[1.2155, 4.246]$.

(D) Test $H_0: \mu_1 - \mu_2 = 0$ against $H_1: \mu_1 - \mu_2 < 0$; test statistic: $(\bar{X}_1 - \bar{X}_2 - 0) / se$ with distribution $T(n_1 + n_2 - 2)$; rejection region for $\alpha = 0.05$: $(-\infty, -1.8595)$; sample value of the test statistic (under $H_0$): $t_{samp} = -4.925$.

Useful formulas and values:

<table>
<thead>
<tr>
<th>$x_{1,i}$</th>
<th>5.3</th>
<th>5.7</th>
<th>7.4</th>
<th>8.7</th>
<th>9.1</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_{1,i}^2$</td>
<td>28.09</td>
<td>32.49</td>
<td>54.76</td>
<td>75.69</td>
<td>82.81</td>
<td>273.84</td>
</tr>
<tr>
<td>$x_{2,i}$</td>
<td>13.7</td>
<td>11.8</td>
<td>11</td>
<td>13.2</td>
<td>16.5</td>
<td>66.2</td>
</tr>
<tr>
<td>$x_{2,i}^2$</td>
<td>187.69</td>
<td>139.24</td>
<td>121</td>
<td>174.24</td>
<td>272.25</td>
<td>894.42</td>
</tr>
</tbody>
</table>

$n_1 = 5$, $n_2 = 5$, $\bar{x}_1 = 7.24$, $\bar{x}_2 = 13.24$, $dev_1 = 11.752$, $dev_2 = 17.932$, $s_1^2 = 2.938$, $s_2^2 = 4.483$, $s_{pooled}^2 = [(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2] / (n_1 + n_2 - 2) = 3.7105$, $s_{pooled} = 1.9263$ se = $s_{pooled} \sqrt{1/n_1 + 1/n_2} = 1.2183$.

Exercise 2. Assumptions: linear regression model $y = \beta_0 + \beta_1 x + u$, where $y =$ number of crimes (per 1000 habitants), $x =$ $\$ invested (per habitant)' and the error term $u$ satisfies the usual assumptions.

(A) Test $H_0: \beta_1 = 0$ against $H_1: \beta_1 \neq 0$; test statistic: $(\hat{\beta}_1 - 0) / se(\hat{\beta}_1)$ with distribution $T(n - 2)$; acceptance region for $\alpha = 0.05$: $[-3.1824, 3.1824]$; sample value of the test statistic (under $H_0$): $t_{samp} = -2.714$.

(B) Pivot for $\mu_0 = E(y|x = x_0 = 13.2)$: $(\hat{\mu}_0 - \mu_0) / se(\hat{\mu}_0) \sim T(n - 2)$ (where $\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1x_0 = 6.1364$ and $se(\hat{\mu}_0) = \hat{\sigma}\sqrt{1/n + (x_0 - \bar{x})^2 / dev(x)} = 0.6262$); interval at $(1 - \alpha) = 0.99$ for $\mu_0$: $[\hat{\mu}_0 - se(\hat{\mu}_0)t, \hat{\mu}_0 + se(\hat{\mu}_0)t] = [2.4791, 9.7938]$, where $t = 5.8409$.

<table>
<thead>
<tr>
<th>$x_i$</th>
<th>14.7</th>
<th>12.8</th>
<th>8.8</th>
<th>11</th>
<th>7.1</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_i^2$</td>
<td>216.09</td>
<td>163.84</td>
<td>77.44</td>
<td>121</td>
<td>50.41</td>
<td>628.78</td>
</tr>
<tr>
<td>$x_i y_i$</td>
<td>79.91</td>
<td>72.96</td>
<td>65.12</td>
<td>95.7</td>
<td>64.61</td>
<td>376.3</td>
</tr>
</tbody>
</table>

$(y_i$ and $y_i'$ are given above as $x_{1,i}$ and $x_{1,i}'$, respectively $n = 5$, $\bar{x} = 10.88$, $\bar{y} = 7.24$, $dev(x) = 36.098$, $dev(y) = 11.752$, $cod(x, y) = -17.556$, $\beta_1 = cod(x, y) / dev(x) = -0.4757$, $\beta_0 = \bar{y} - \bar{x} x = 12.4153$, $dev(RES) = dev(y - \hat{\beta}_1 x) / (n - 2) = 3.4012$, $\hat{\beta}_2 = dev(RES) / (n - 2) = 1.1337$, $se(\hat{\beta}_1) = \hat{\sigma} / \sqrt{dev(x)} = 0.1753$.

Exercise 3.

(A) $\rho(X, Y) = Cov(X, Y) / (\sigma(X) \sigma(Y)) = -0.3619$, where $E(X) = 1.12$, $E(Y) = 4.24$, $V(X) = E(X^2) - E(X)^2 = 2.24 - 1.12^2 = 0.9856$, $V(Y) = E(Y^2) - E(Y)^2 = 18.92 - 4.24^2 = 0.9424$, $\sigma(X) = \sqrt{V(X)} = 0.9928$, $\sigma(Y) = \sqrt{V(Y)} = 0.9708$, $Cov(X, Y) = E(XY) - E(X)E(Y) = 4.4 - 1.12 * 4.24 = -0.3488$.

(B) $E(Y|X = 0) = 4.6364$, $V(Y|X = 0) = E(Y^2|X = 0) - E(Y|X = 0)^2 = 22.0909 - 4.6364^2 = 0.595$, $\sigma(Y|X = 0) = 0.7714$. 


Text

Framework: An NGO is studying fire as an important problem of the Amazon Forest.

Exercise 1.

A substantially homogeneous portion of the Pará region (Brazil) has been divided into 63 regular cells. The following table summarizes some sample statistics relative to runaway fires counted in each cell during 2012.

<table>
<thead>
<tr>
<th>Statistic Value</th>
<th>min</th>
<th>1-st quartile</th>
<th>median</th>
<th>3-rd quartile</th>
<th>max</th>
<th>mean</th>
<th>unbiased variance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>1.1587</td>
<td>1.297</td>
</tr>
</tbody>
</table>

(A) Specify the variable of interest and choose an appropriate statistical model for it; motivate such a choice by both theoretical and empirical considerations.

(B) Estimate all parameters of the model by maximum likelihood and compute the corresponding standard errors.

(C) Compute the log-likelihood for the simple random sample $x = (1, 3)$ (only two observations), at the parameter values estimated in (B).

(D) An expert asserted at the beginning of 2012 that the mean number of fires by cell should be lower than 1.33. Test such an opinion at a significance level of 10%.

(E) Compute the type II error probability of the test at point (D) assuming that, under the alternative, the mean number of fires by cell is 1.05.

Exercise 2.

The University of California, Irvine provides yearly predictions of the fire risk during the dry season in the Amazon Forest. The dependent variable is FSS (Fire Season Severity index, i.e. the count of fires detected in the season by satellites); the two independent variables, representing the sea surface temperature anomalies in Pacific and Atlantic, are ONI (Ocean Nino Index) and AMO (Atlantic Multidecadal Oscillation index).

We consider here a simplified analysis including AMO only as independent variable. The following table reports some statistics computed on a sample of 25 observations (sd denotes the square root of the unbiased sample variance).

<table>
<thead>
<tr>
<th>mean(AMO)</th>
<th>mean(FSS)</th>
<th>sd(AMO)</th>
<th>sd(FSS)</th>
<th>slope coefficient estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0772</td>
<td>216.29</td>
<td>0.7799</td>
<td>85.07</td>
<td>83.47</td>
</tr>
</tbody>
</table>

(A) Test whether AMO is significant in explaining FSS ($\alpha = 0.05$).

(B) (No for students with 6 CFU) Compute the confidence interval for the standard deviation of the error component considering a confidence level of 90%.

(C) (No for students with 6 CFU) According to current projections, the AMO for 2013 is estimated at 0.79. Compute the confidence interval at $\alpha = 0.1$ for the corresponding conditional mean of FSS.

Solution

Exercise 1.

(A) Assumptions: $X = \text{`Runaway fires in one cell'~ Po(}$\lambda$)$. The assumption is justified both theoretically, because the Poisson is suitable for counting data, and empirically, given that the sample mean and variance are close.

(B) $\hat{\lambda} = \overline{x} = 1.1587$, $se(\hat{\lambda}) = \sqrt{\hat{\lambda}/n} = 0.1356$.

(C) The log-likelihood for the Poisson is the joint log-pmf of the sample, in general expressed by \( l(\lambda) = \sum_{i=1}^{n} \ln(\lambda \cdot x_i!) \). Taking \( \lambda = \hat{\lambda} = 1.1587 \), \( n = 2 \), \( x_1 = 1 \) and \( x_2 = 3 \) we get \( l(\lambda) = \ln(0.3637) + \ln(0.8141) = -3.51996 \).

(D) Test \( H_0 : \lambda = 1.33 \) against \( H_1 : \lambda < 1.33 \); test statistic: \( \overline{X} - 1.33 \)/\( \sigma_x \) (where \( \overline{X} \) denotes the standard error of \( \overline{X} \) under \( H_0 \)) with approximated distribution \( N(0,1) \); rejection region for \( \alpha = 0.1 \): \( -\infty, z_{crit} = -1.2816 \); sample value of the test statistic (under \( H_0 \)): \( z_{samp} = -1.179 \).

(E) Power of the test \( H_0 : \lambda = 1.33 \) against \( H_1 : \lambda = 1.05 \). For simplicity, we take \( 1.33 = \lambda_0 \) and \( 1.05 = \lambda_1 \). \( P(\overline{X} < z_{crit}|H_1) = P(\overline{X} > \lambda_1 + z_{crit} \sigma_x|H_1) = P(\overline{X} > \lambda_1 + z_{crit} \sigma_x|H_1) = P(\overline{X} > 3.7265) = 0.2338 \), where \( \sigma_x = \sqrt{1.05/n} \) denotes the standard error of \( \overline{X} \) under \( H_1 \).

Useful formulas and values: \( n = 63, \overline{X} = 1.1587, \sigma_x = \sqrt{\lambda_0/n} = 0.1453, \sigma_x = \sqrt{\lambda_1/n} = 0.1291 \)

Exercise 2.
Assumptions: linear regression model \( y = \beta_0 + \beta_1 x + u \), where \( y = \text{FSS} \), \( x = \text{AMO} \) and the error term \( u \) satisfies the usual assumptions.

(A) Test \( H_0 : \beta_1 = 0 \) against \( H_1 : \beta_1 \neq 0 \); test statistic: \( (\hat{\beta}_1 - 0)/\hat{\sigma}_x \) with distribution \( T(n-2) \); acceptance region for \( \alpha = 0.05 \): \( -2.0687, 2.0687 \); sample value of the test statistic (under \( H_0 \)): \( t_{samp} = 5.7008 \).

(B) Pivot for \( \sigma^2 \): \( (\overline{X}_n - 2)\hat{\sigma}/\sigma^2 \sim \chi^2(n-2) \); interval at \( 1 - \alpha = 0.9 \) for \( \sigma^2 \): \( [\hat{\sigma}^2(n-2)/c_2, \hat{\sigma}^2(n-2)/c_1] = [2046.4572, 5498.5569] \), where \( c_1 = 13.0905, c_2 = 35.1725 \); corresponding interval for \( \sigma^2 \): \( [45.2378, 74.1523] \).

(C) Pivot for \( \mu_0 = E(y|x = x_0 = 0.79) \): \( (\hat{\mu}_0 - \mu_0)/\hat{\sigma}(\mu_0) \sim T(n-2) \) (where \( \hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = 275.7874 \) and \( \hat{\sigma}(\mu_0) = \hat{\sigma} \sqrt{1/n + (x_0 - \overline{X})^2/\sigma(\mu_0) = 15.3005} \); interval at \( 1 - \alpha = 0.9 \) for \( \mu_0 \): \( [\hat{\mu}_0 - \hat{\sigma}(\mu_0)t, \hat{\mu}_0 + \hat{\sigma}(\mu_0)t] = [249.5644, 302.0105] \), where \( t = 1.7139 \).

Useful formulas and values: \( n = 25, \overline{X} = 0.0772, \overline{Y} = 216.29, \sigma^2 = s^2_X(n-1) = 14.5979, \sigma^2 = s^2_Y(n-1) = 173685.7176, \hat{\beta}_1 = 83.47, \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} = 209.8461, \sigma(\hat{\beta}_1) = \sigma^2 X \sigma^2 X, \sigma(\hat{\beta}_1) = 71978.9366, \hat{\beta}_1 = 71978.9366, \sigma^2 = \sigma^2_x(n-2) = 3129.519, \sigma^2 = \sigma^2_x(n-2) = 3129.519, \sigma_x = \sigma_x = 14.6418 \).
Framework: A research team at SNU (Seoul National University) is focusing on the Korea economy.

Exercise 1.
A classical analysis aims at checking the Okun Law on the empirical ground. The version of the law considered, assumes that $\Delta u = \alpha + \beta \Delta y$, where $y$ denotes the GDP, $u$ the unemployment rate and the $\Delta$ operator applied to a symbol indicates the 1-year change in the corresponding variable. Summary statistics on a sample of 20 years: $\Delta u = 0.049$, $s(\Delta u) = 1.183$, $\Delta y = 5.099$, $s(\Delta y) = 3.582$, $s(\Delta u, \Delta y) = -3.424$, where $s(\cdot)$ and $s(\cdot, \cdot)$ denote the sample standard deviation and covariance, respectively, both computed using the usual $n - 1$ degrees of freedom at denominator.

Translate the law into an appropriate statistical model and answer the following questions.

(A) Compute the confidence interval for the conditional mean of the 1-year change of the unemployment rate when the GDP remains unchanged ($\alpha = 0.1$).

(B) Test whether the independent variable is significant in the regression ($\alpha = 0.1$).

(C) Compute fitted value and residual for an observation having $(\Delta u_i = -0.311, \Delta y_i = 3.682)$.

Exercise 2.
The SNU team surveyed two independent simple random samples of Korean workers in services and in other industries, respectively. The following table summarizes some statistics concerning the expectations of the workers about their economical situation in the next three years.

<table>
<thead>
<tr>
<th>Industry</th>
<th>Positive</th>
<th>Neutral</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Services</td>
<td>115</td>
<td>145</td>
<td>83</td>
</tr>
<tr>
<td>Other</td>
<td>118</td>
<td>205</td>
<td>82</td>
</tr>
</tbody>
</table>

Answer the following questions.

(A) Compute a point estimate of the difference, between the two industries, in the proportion of people expecting that their economical situation is getting better; complement the point estimate with the corresponding standard error.

(B) Test whether the expectation about an improvement of the economical situation is statistically higher in Services ($\alpha = 0.1$).

(C) Compute the power of the test at point (B) assuming that $H_1: p_{\text{Services}} - p_{\text{Other}} = 0.049$, where $p_\ast$ denotes the probability of an improvement in the industry $\ast$.

Exercise 3.
Considering the table in exercise (2), transform industry to the random variable $X$ (mapping Services to 1 and Other to $-1$) and expectation to the random variable $Y$ (mapping the three outcomes to 1, 0, $-1$ respectively). Assume in addition that the probability of one cell is identical to the corresponding relative frequency.

(A) Draw the cdf of $Y | X = -1$.

(B) Compute the variance of $Y | X = -1$.

Solution

Exercise 1.
Assumptions: $y = \beta_0 + \beta_1 x + \varepsilon$, where $\varepsilon$’s are i.i.d. and $\sim N(0, \sigma^2)$, $y = \Delta u$, $x = \Delta y$. 


(A) When $x = 0$, the conditional mean of $y$ is $\beta_0$. Pivot for $\beta_0$: $(\hat{\beta}_0 - \beta_0)/se(\hat{\beta}_0) \sim T(n - 2)$; interval at $(1 - \alpha) = 0.9$ for $\beta_0$: $[\hat{\beta}_0 - t \cdot se(\hat{\beta}_0), \hat{\beta}_0 + t \cdot se(\hat{\beta}_0)] = [0.9183, 1.9012]$, where $t = 1.7341$.

(B) Test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$; test statistic: $(\hat{\beta}_1 - 0)/se(\hat{\beta}_1)$ with distribution $T(n - 2)$; acceptance region for $\alpha = 0.1$: $[-1.7341, 1.7341]$; sample value of the test statistic (under $H_0$): $t_{samp} = -5.8188$.

(C) Residual for $(x_i = 3.682, y_i = -0.311)$: $\hat{\epsilon}_i = y_i - \hat{y}_i = -0.7381$ where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = 0.4271$.

Useful formulas and values:

\[ n = 20, \bar{x} = 5.099, \bar{y} = 0.049, \text{dev}(\bar{x}) = s^2_x (n - 1) = 243.7838, \text{dev}(\bar{y}) = s^2_y (n - 1) = 26.5903, \text{cod}(\bar{x}, \bar{y}) = -65.056, \hat{\beta}_1 = -0.2669, \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 1.4097, \text{dev}(\text{RES}) = \text{dev}(\bar{y}) - \hat{\beta}_1^2 \text{dev}(\bar{x}) = 9.2295, \hat{\sigma}^2 = \text{dev}(\text{RES})/(n - 2) = 0.5127, \text{se}(\hat{\beta}_1) = \hat{\sigma}/\sqrt{\text{dev}(\bar{x})} = 0.04586, \text{se}(\hat{\beta}_0) = \hat{\sigma}\sqrt{1/n + \bar{x}^2/\text{dev}(\bar{x})} = 0.2834 \]

Exercise 2.

Assumptions: $X_1$ =’Positive expectation for one worker in Services’ ~ $Be(p_1)$; $X_2$ =’Positive expectation for one worker in other industries’ ~ $Be(p_2)$; $X_1$ and $X_2$ are independent as well as the corresponding samples.

(A) Estimator of $\theta = p_1 - p_2$: $\hat{\theta} = \hat{p}_1 - \hat{p}_2$; corresponding estimate: $\hat{\theta} = \hat{p}_1 - \hat{p}_2 = 0.04392$ with $se = se(\hat{\theta}) = se(\hat{p}_1 - \hat{p}_2) = \sqrt{\hat{p}_1q_1/n_1 + \hat{p}_2q_2/n_2} = \sqrt{0.00116} = 0.034052$.

(B) We use the notation in (A). Test of $H_0 : \theta = 0$ against $H_1 : \theta > 0$; test statistic: $(\hat{\theta} - 0)/se$ with approximated distribution $N(0,1)$; rejection region for $\alpha = 0.1$: ($z_{crit} = 1.2816, \infty$); sample value of the test statistic (under $H_0$): $z_{samp} = 1.2898$.

(C) We use the notation in (A). Power of the test $H_0 : \theta = 0$ against $H_1 : \theta = 0.049$. We denote 0.049 as $\theta_1$. $\gamma = P(\text{sample } \in R|H_1) = P((\hat{\theta} - 0)/se > z_{crit}|H_1) = P(\hat{\theta} > z_{crit} \cdot se|H_1) = P((\hat{\theta} - \theta_1)/se > (z_{crit} \cdot se - \theta_1)/se|H_1) = P(Z > -0.15741) = 0.56254$.

Useful formulas and values:

\[ n_1 = 343, n_2 = 465, \hat{p}_1 = 0.3353, \hat{p}_2 = 0.2914 \]

Exercise 3.

<table>
<thead>
<tr>
<th>X</th>
<th>1</th>
<th>0</th>
<th>-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1537</td>
<td>0.1939</td>
<td>0.111</td>
</tr>
<tr>
<td>-1</td>
<td>0.1578</td>
<td>0.2741</td>
<td>0.1096</td>
</tr>
</tbody>
</table>

(A) (B)

<table>
<thead>
<tr>
<th>y</th>
<th>-1</th>
<th>0</th>
<th>1</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f(y</td>
<td>X = -1)$</td>
<td>0.2025</td>
<td>0.5062</td>
<td>0.2914</td>
</tr>
<tr>
<td>$F(y</td>
<td>X = -1)$</td>
<td>0.2025</td>
<td>0.7086</td>
<td>1</td>
</tr>
<tr>
<td>$y f(y</td>
<td>X = -1)$</td>
<td>-0.2025</td>
<td>0</td>
<td>0.2914</td>
</tr>
<tr>
<td>$y^2 f(y</td>
<td>X = -1)$</td>
<td>0.2025</td>
<td>0</td>
<td>0.2914</td>
</tr>
</tbody>
</table>

For the cdf $F(y|X = -1)$, only the values at the jumps are reported.

\[ V(Y|X = -1) = E(Y^2|X = -1) - E(Y|X = -1)^2 = 0.4938 - 0.0889^2 = 0.4859 \]
Text

Framework: A research team is involved in a program concerning young women in the Gaza strip.

Exercise 1.

A training program has involved women aged less than 35, with at least 2 children, without a job but wishing to work. Such women have been divided in two subsamples trained by two different teams, each one with its own method. The following table reports some statistics on the scores reported by the enrolled women in the final test.

<table>
<thead>
<tr>
<th></th>
<th>observations</th>
<th>sum</th>
<th>sum of squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>7</td>
<td>406</td>
<td>23642</td>
</tr>
<tr>
<td>Group 2</td>
<td>7</td>
<td>340</td>
<td>17490</td>
</tr>
</tbody>
</table>

Assuming that the scores are normally distributed and making no assumptions on the size of the variances, answer the following questions.

(A) Did the two samples provide significantly different average scores (significance level 0.1)?

(B) A value of $\sigma$ higher than 9.7 is believed to express a too large heterogeneity in the enrollment of women. Test whether this is true in the sample with the largest estimated standard deviation ($\alpha = 0.1$).

(C) Compute how many women should be enrolled in the first group for estimating the mean score with a confidence interval of size 1.59 at the 0.95 confidence level.

(D) Assuming to identify the parameters with the corresponding estimated values, compute the probability that a woman, randomly drawn among the 14 in the two samples, has a score greater than 58.

Exercise 2.

In order to contextualize above training program, a statistical analysis has concerned a sample of women, aged between 30 and 35. The aim is to evaluate the possible relationship between the number of children, taken as dependent variable, and the number of years of school education. The following table reports some statistics ($c =$ number of children, $s =$ number of years of school education; the standard deviation, $sd(\cdot)$, and the correlation coefficient $cor(\cdot)$ are computed using the sample size at denominator).

<table>
<thead>
<tr>
<th>n</th>
<th>min(c)</th>
<th>max(c)</th>
<th>$\bar{c}$</th>
<th>sd(c)</th>
<th>min(s)</th>
<th>max(s)</th>
<th>$\bar{s}$</th>
<th>sd(s)</th>
<th>cor(c, s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>158</td>
<td>0</td>
<td>8</td>
<td>0.766</td>
<td>1.623</td>
<td>5</td>
<td>14</td>
<td>7.766</td>
<td>2.044</td>
<td>-0.5162</td>
</tr>
</tbody>
</table>

Specify an appropriate simple linear regression model with normally distributed errors and answer the following questions.

(A) Build the confidence interval for the model intercept at 0.99 confidence level.

(B) Test whether, according to expectations, a higher level of education is associated to a smaller number of children ($\alpha = 0.01$).

(C) Compute how much of the total variability of the dependent variable is explained by the model.

(D) Something is wrong in applying a simple linear regression model to these data. What? Discuss this point.

Solution

Exercise 1.

Assumptions: $X_1 =$‘Score of one woman in Group 1’ $\sim N(\mu_1, \sigma_1)$; $X_2 =$‘Score of one woman in Group 2’ $\sim N(\mu_2, \sigma_2)$; $X_1$ and $X_2$ are independent as well as the corresponding samples.

(A) Test of $H_0: \mu_1 - \mu_2 = 0$ against $H_1: \mu_1 - \mu_2 \neq 0$; test statistic (Satterthwaite-Welsh, under $H_0$): $(\hat{\mu}_1 - \hat{\mu}_2 - 0)/se$ with approximated distribution $T(df = 7.145 \approx 7)$; acceptance region for $\alpha = 0.1$: $[-1.8946, 1.8946]$; sample value of the test statistic (under $H_0$): $t_{samp} = 1.8683$. 

30
Exercise 2

We consider now the second sample removing index 2 from the different symbols. Test of $H_0 : \sigma = 9.7$ against $H_1 : \sigma > 9.7$; test statistic (under $H_0$): $S^2 df/9.7^2$ with distribution $\chi^2(df = 6)$; rejection region for $\alpha = 0.1$: $[10.6446, \infty)$; sample value of the test statistic (under $H_0$): $\chi^2_{samp} = 10.37$.

We consider now the first sample removing index 1 from the different symbols. $n = (2\hat{\sigma}^2 z/A)^2 = 9.7582^2 = 95.2223 \approx 96$, where $A = 1.59, \alpha = 0.05, z = 1.96, \hat{\sigma} = 3.9581$.

We set $X|G_1 \sim N(\mu_1 = 58, \sigma_1 = 3.9581), X|G_2 \sim N(\mu_2 = 48.5714, \sigma_2 = 12.7522)$, where $G_1 = \text{Group 1}, G_2 = \text{Group 2}; c = 58$. $P(X > c) = P(X > c|G_1)P(G_1) + P(X > c|G_2)P(G_2) = 0.3649$, where $P(G_1) = n_1/(n_1 + n_2) = 0.5, P(G_2) = n_2/(n_1 + n_2) = 0.5, P(X > c|G_1) = P((X - \mu_1)/\sigma_1 > (c - \mu_1)/\sigma_1|G_1) = P(Z > 0|G_1) = 0.5, P(X > c|G_2) = P((X - \mu_2)/\sigma_2 > (c - \mu_2)/\sigma_2|G_2) = P(Z > 0.7394|G_2) = 0.22984$.

Useful formulas and values:

- $n_1 = 7, n_2 = 7, df_1 = n_1 - 1 = 6, df_2 = n_2 - 1 = 6, \mu_1 = \tau_1 = \sum_{i=1}^{n_1} x_{1,i}/n_1 = 58, \mu_2 = \tau_2 = \sum_{i=1}^{n_2} x_{2,i}/n_2 = 48.5714, dev_1 = \sum_{i=1}^{n_1} x_{1,i}^2 - n_1 \tau_1^2 = 94, dev_2 = \sum_{i=1}^{n_2} x_{2,i}^2 - n_2 \tau_2^2 = 975.7143, s_1^2 = dev_1/df_1 = 15.6666, s_2^2 = df_2 = 162.6186, s_1 = 3.9581, s_2 = 12.7522, se_1 = s_1/\sqrt{n_1} = 1.496, se_2 = s_2/\sqrt{n_2} = 4.8199, se^2 = 2.2381, se^2 = 23.2313, se = \sqrt{se^2 + se^2} = \sqrt{25.46939} = 5.0467, df = (se^2 + se^2)/cn + se^2/df_2 = 7.1454$.

Exercise 2

Assumptions: $y = \beta_0 + \beta_1 x + u$, where $u$’s are i.i.d. and $N(0, \sigma^2), y = c, x = s$.

(A) Pivot for $\beta_0$: $[\hat{\beta}_0 - \beta_0]/se(\hat{\beta}_0) \sim T(n - 2)$; interval at $(1 - \alpha) = 0.99$ for $\beta_0$: $[\hat{\beta}_0 - t \cdot se(\hat{\beta}_0), \hat{\beta}_0 + t \cdot se(\hat{\beta}_0)] = [2.8089, 5.0893]$, where $t = 2.6077$.

(B) Test for $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 < 0$; test statistic: $(\hat{\beta}_1 - 0)/se(\hat{\beta}_1)$ with distribution $T(n - 2)$; rejection region for $\alpha = 0.01$: $(-\infty, -2.3053]$; sample value of the test statistic (under $H_0$): $t_{samp} = -7.5278$.

(C) $R^2 = 0.266$. Then the model explains 26.6% of the total variability of $y$.

(D) The dependent variable $c$ assumes only integer values in a narrow domain. The assumption, implicit in the regression model formulated, that $y|x \sim N(\beta_0 + \beta_1 x, \sigma^2)$ is not reasonable in this case. In fact $y$ can assume not integer values and, for suitable values of $s$ can go outside the domain of the dependent variable (for which $0$ is the minimum).

Useful formulas and values:

- $n = 158, \tau = 7.766, \bar{y} = 0.766, dev(x) = s^2_n = 660.1139, dev(y) = s^2_n = 416.1924, cod(x,y) = \rho s_x s_y n = -270.5668, \hat{\beta}_1 = -0.4099, \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \tau = 3.9491, dev(REG) = \hat{\beta}_1^2 dev(x) = 110.8996, dev(REG) = dev(y) - dev(REG) = 365.2927, \hat{\sigma}^2 = dev(REG)/(n - 2) = 1.957, se(\hat{\beta}_1) = \hat{\sigma}/\sqrt{\text{dev}(x)} = 0.0544, se(\hat{\beta}_0) = \hat{\sigma}/\sqrt{1/n + \tau^2/\text{dev}(x)} = 0.4372$.
Test 12.06.2014

Text
Framework: A research program deals with climate changes and the possible impact on this of the human activities.

Exercise 1.
One important element to understand the climate of a given area is the so called eddy flux, the exchange rate of specific gases between natural ecosystems or between areas.
By using specific information, the researchers built a simple linear model aimed at predicting eddy flux of carbon in a forestry African area. The following table reports the sample values of the eddy flux of carbon, taken as dependent variable, together with the corresponding fitted values coming from the estimated model.

<table>
<thead>
<tr>
<th>Observation</th>
<th>Fitted value</th>
</tr>
</thead>
<tbody>
<tr>
<td>72</td>
<td>1</td>
</tr>
<tr>
<td>47</td>
<td>...</td>
</tr>
<tr>
<td>−69</td>
<td>−22</td>
</tr>
<tr>
<td>−5</td>
<td>9</td>
</tr>
<tr>
<td>55</td>
<td>72</td>
</tr>
</tbody>
</table>

(A) Compute the value of the missing fitted value, specifying the property used to derive it. (Hint: If you are unable to compute the missing fitted value, replace it by the corresponding observation in order to answer the following points.)

(B) Compute the confidence interval for $\sigma$ at $\alpha = 0.01$.

(C) Evaluate the $R^2$ index of the regression and interpret the result.

Exercise 2.
Soil erosion is one important concern for many African areas. The following table reports measures relative to two different samples, useful to evaluate the possible impact of the human intervention on soil erosion.

<table>
<thead>
<tr>
<th>Managed forest</th>
<th>226 218 239 272 203</th>
</tr>
</thead>
<tbody>
<tr>
<td>Undisturbed forest</td>
<td>188 189 205 199 205</td>
</tr>
</tbody>
</table>

Assuming that the measures of soil erosion follow a Normal distribution with no assumptions on the variances:

(A) Compute a point estimate of the mean impact of the human intervention on the soil erosion; complement the estimate with the corresponding standard error.

(B) Does the human intervention increase significantly the mean soil erosion? ($\alpha = 0.01$)

(C) Perform the same test at point (B) under the assumption of equal variances and compare the two outcomes.

Exercise 3.
Consider the previous exercise and assume that the measures of soil erosion have the distribution indicated with parameters identical to the corresponding estimates. Knowing that 51% of forests are undisturbed, answer the following questions:

(A) Compute the probability that a randomly chosen forest has a soil erosion measure greater than 203.

(B) Knowing that a measure of erosion is greater than 203, compute the probability that it belongs to an undisturbed forest.

Solution

Exercise 1.
(A) The missing fitted value (bolded in the table) can be computed by remembering that, in a simple linear model with intercept, the residuals have zero sum: \( \sum_{i=1}^{n} \hat{u}_i = 0. \) This implies that \( \hat{y}_2 = -\sum_{i \neq 2} \hat{u}_i = 7. \) Remembering that in general \( \hat{u}_i = y_i - \hat{y}_i, \) we have \( \hat{y}_2 = y_2 - \hat{u}_2 = 47 - 7 = 40. \)

(B) Pivot for \( \sigma^2: \) \( \hat{\sigma}^2(n-2)/\sigma^2 \) with distribution \( \chi^2(n-2); \) interval for \( \sigma^2 \) at \( 1 - \alpha = 0.99: \) \([\hat{\sigma}^2(n-2)/\sigma_2, \hat{\sigma}^2(n-2)/\sigma_1]\) = [606.3176, 108530.4992] where \( c_1 = 0.0717, c_2 = 12.8382, \hat{\sigma}^2 = dev(Res)/(n - 2) = 2594.6667; \) corresponding interval for \( \sigma: \) [24.6235, 329.4397].

(C) \( R^2 = \frac{dev(Reg)}{dev(y)} = 1 - \frac{dev(Res)}{dev(y)} = 1 - 7784/13204 = 0.4105. \) Then the model explains 41% of the total variability of \( y. \)

Exercise 2.
Assumptions: \( X_1 = \\text{‘Measure of soil erosion in managed forest’} \sim N(\mu_1, \sigma_1); \) \( X_2 = \\text{‘Measure of soil erosion in undisturbed forest’} \sim N(\mu_2, \sigma_2); \) \( X_1 \) and \( X_2 \) are independent as well as the corresponding samples.

(A) Point estimator of \( \mu_1 - \mu_2: \) \( \overline{X}_1 - \overline{X}_2; \) corresponding point estimate: \( \overline{X}_1 - \overline{X}_2 = 231.6 - 197.2 = 34.4; \) corresponding standard error: \( se(\overline{X}_1 - \overline{X}_2) = \sqrt{V(\overline{X}_1 - \overline{X}_2)} = \sqrt{s_1^2/n_1 + s_2^2/n_2} = \sqrt{149.9} = 12.2434. \)

(B) Test of \( H_0 : \mu_1 - \mu_2 = 0 \) against \( H_1 : \mu_1 - \mu_2 > 0; \) test statistic (Satterthwaite-Welsh, under \( H_0)): \( (\hat{\mu}_1 - \hat{\mu}_2 - 0)/se \) with approximated distribution \( T(df = 4.805 \approx 5); \) rejection region for \( \alpha = 0.01: \) [3.3649, \( \infty \)]; sample value of the test statistic (under \( H_0): \) \( t_{samp} = 2.8097. \)

(C) Test of the same hypotheses as in (B) but assuming \( \sigma_1 = \sigma_2 = \sigma. \) Test statistic (under \( H_0): \) \( (\hat{\mu}_1 - \hat{\mu}_2 - 0)/\sqrt{s^2(1/n_1 + 1/n_2)} \) with distribution \( T(n_1 + n_2 - 2 = 8); \) rejection region for \( \alpha = 0.01: \) [2.8965, \( \infty \)]; sample value of the test statistic (under \( H_0): \) \( t_{samp} = 2.8097. \)

Useful formulas and values:
\( n_1 = 5, n_2 = 5, df_1 = n_1 - 1 = 4, df_2 = n_2 - 1 = 4, \hat{\mu}_1 = \overline{X}_1 = \sum_{i=1}^{n_1} x_{1,i}/n_1 = 231.6, \hat{\mu}_2 = \overline{X}_2 = \sum_{i=1}^{n_2} x_{2,i}/n_2 = 197.2, \) \( dev_1 = \sum_{i=1}^{n_1} x_{1,i}^2/n_1 - n_1\overline{X}_1^2 = 2721.2, dev_2 = \sum_{i=1}^{n_2} x_{2,i}^2/n_2 - n_2\overline{X}_2^2 = 276.8, s_1^2 = dev_1/df_1 = 680.3, s_2^2 = dev_2/df_2 = 69.2, s_1 = 26.0826, s_2 = 8.3187, se_1 = s_1/\sqrt{n_1} = 11.6645, se_2 = s_2/\sqrt{n_2} = 3.7202, \)
\( se_1^2 = 136.06, se_2^2 = 13.84, se = \sqrt{se_1^2 + se_2^2} = \sqrt{149.9} = 12.2434, df = \frac{(se_1^2 + se_2^2)^2}{se_1^2/df_1 + se_2^2/df_2} = 4.8054, s_p^2 = (s_1^2df_1 + s_2^2df_2)/(df_1 + df_2) = 374.75, \sqrt{s_p^2(1/n_1 + 1/n_2)} = \sqrt{149.9} = 12.2434. \)

Exercise 3. Let \( X = \text{‘Measure of soil erosion’}, M = \text{‘managed forest’}, U = \text{‘undisturbed forest’}; X|M \sim N(\mu_1 = 231.6, \sigma_1 = 26.0826); X|U \sim N(\mu_2 = 197.2, \sigma_2 = 8.3187). \)

(A) \( P(X > 203) = P(X > 203|M)P(M) + P(X > 203|U)P(U) = 0.547, \) where \( P(X > 203|M) = P(Z > -1.0965) = 0.8636, P(X > 203|U) = P(Z > 0.6972) = 0.2428, P(U) = 0.51. \)

(B) \( P(U|X > 203) = P(X > 203|U)P(U)/P(X > 203) = 0.2264. \)
Text

Framework: A research group analyzes health data in different countries by using information coming from OECD statistics (http://www.oecd.org/statistics/).

Exercise 1.

Health expenses (health, for short), considered as dependent variable, are expected to be related to the level of GDP (gdp, for short) of the Country. The following table summarizes some statistics coming from a sample of 25 OECD Countries (original values in millions US$) in 2007:

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>√Biased Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>ln(gdp)</td>
<td>12.84</td>
<td>1.454</td>
</tr>
<tr>
<td>ln(health)</td>
<td>10.33</td>
<td>1.531</td>
</tr>
</tbody>
</table>

The correlation coefficient between the two log-variables is equal to 0.991. Formulate an appropriate statistical model and answer the following questions.

(A) Test the null hypothesis of a unit slope coefficient (equivalent to one-to-one proportionality between the two log-variables) for $\alpha = 0.01$.

(B) Compute the confidence interval for the standard deviation of the error component at $\alpha = 0.01$.

(C) Compute the residual for the observation $(gdp_{Netherlands}, health_{Netherlands}) = (615570, 61329)$ and interpret the result.

(D) Compute the point and the interval estimate ($\alpha = 0.02$) of the conditional mean of a Country whose gdp is 192836.

Exercise 2.

According to a 2012 survey, the percentage of Italians aged between 45 and 64 judging itself in a good or very good status (simply good, henceforth) was 68.3% among females and 73.9% among males.

Knowing that the two simple random samples of females and males (with sizes equal to 1096 and 1000, respectively) was sampled independently:

(A) Test whether males perceive to be more healthy than females ($\alpha = 0.02$).

(B) Compute the power of the test at point (A) when the alternative is “the difference males minus females about their perceived good status is 6.4 percentage points”.

Exercise 3.

We know that, within some population of females, the percentage of women judging itself in a good health status is the same as indicated in exercise 2.

(A) Drawing randomly 6 women, compute the probability that there are at least 2 with a not good health status.

(B) Knowing that, among the 6 extracted, there are at least 2 with a not good health status, compute the probability of getting exactly 2 in the same condition.

Solution

Exercise 1.

Simple linear model $y_i = \beta_0 + \beta_1 x_i + u_i$ with the usual assumptions, where $y = \ln(health)$ and $x = \ln(gdp)$.

(A) Test of $H_0 : \beta_1 = 1$ against $H_1 : \beta_1 \neq 1$; test statistic: $(\hat{\beta}_1 - 1)/se(\hat{\beta}_1)$ with distribution $T(n - 2 = 23)$; acceptance region for $\alpha = 0.01$: $[-2.8073, 2.8073]$; sample value of the test statistic (under $H_0$): $t_{samp} = 1.4794$. 

Let $X$ where the numerator is computed by means of the Binomial p.m.f. $eta(B)$ probabilities are computed by means of the Binomial p.m.f. $P(T)$ distribution $(D)$ Point estimate of $(B)$ Pivot for $(A)$ Statistic (under $(H)$ that $\beta(\hat{\beta}) = \beta(0) + \beta(1)x_0 = 9.6034$. Pivot for $\mu_0$: $(\hat{\mu}_0 - \mu_0)/se(\hat{\mu}_0)$ with distribution $T(n-2 = 23)$; interval for $\mu_0$ at $1 - \alpha = 0.98$: $[\hat{\mu}_0 - tse(\mu_0), \hat{\mu}_0 + tse(\mu_0)] = [9.5128, 9.7481]$ where $t = 2.4999$.

Useful formulas and values:

$$n = 25, \overline{x} = 12.84, \overline{y} = 10.33, \text{dev}(\overline{x}) = n\sigma^2(\overline{x}) = 52.8529, \text{dev}(\overline{y}) = n\sigma^2(\overline{y}) = 58.509, \text{cod}(\overline{x}, \overline{y}) = n(\overline{x})\sigma(\overline{y})\rho(\overline{x}, \overline{y}) = 55.151, \hat{\beta}_1 = \text{cod}(\overline{x}, \overline{y})/\text{dev}(\overline{x}) = 1.0435, \hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} = -3.0683, \text{dev}(RES) = \text{dev}(y) - \hat{\beta}_1\text{dev}(x) = 1.05, \hat{\sigma}^2 = \text{dev}(RES) = 0.0457, se(\hat{\beta}_1) = \hat{\sigma}/\sqrt{\text{dev}(\overline{x})} = 0.0294, se(\hat{\mu}_0) = \hat{\sigma}\sqrt{1/n + (x_0 - \overline{x})^2/\text{dev}(\overline{x})} = 0.0471$.

Exercise 2.

Assumptions: $X_1$ =’A female has a good health perception’ ~ $Be(p_1)$; $X_2$ =’A male has a good health perception’ ~ $Be(p_2)$; $X_1$ and $X_2$ are independent as well as the corresponding samples.

(A) Test of $H_0: p_2 - p_1 = 0$ against $H_1: p_2 - p_1 > 0$; test statistic (under $H_0$): $(\hat{p}_1 - \hat{p}_2 - 0)/se$ with approximated distribution $N(0, 1)$; rejection region for $\alpha = 0.02$: $[z_{crit} = 2.0537, \infty)$; sample value of the test statistic (under $H_0$): $\hat{z}_{samp} = 2.8341$.

(B) Power of above test considering $H_1: p_2 - p_1 = 0.064$. Let $0.064 = c$. $\gamma = P(\text{sample} \in R|H_1) = P(\hat{p}_1 - \hat{p}_2 > z_{crit}|H_1) = P(\hat{p}_1 - \hat{p}_2 > z_{crit} \cdot se|H_1) = P((\hat{p}_1 - \hat{p}_2 - c)/se > (z_{crit} \cdot se - c)/se|H_1) = P(Z > -1.1853) = 0.88204$.

Useful formulas and values:

$$n_1 = 1096, n_2 = 1000, \hat{p}_1 = 0.683, \hat{p}_2 = 0.739, se = \sqrt{\hat{p}_1\hat{q}_1/n_1 + \hat{p}_2\hat{q}_2/n_2} = \sqrt{0.00039043} = 0.0198$.

Exercise 3.

Let $X$ =’A woman has a poor perception of her health’ ~ $Be(p = 0.317)$. 

(A) Drawing $n = 6$ independent observations from $X$ we get $Y = \sum_{i=1}^{6} X_i \sim Bi(n = 6, p = 0.317)$. Hence, $P(Y \geq 2) = 1 - P(Y < 2) = 1 - [P(Y = 0) + P(Y = 1)] = 1 - (0.10151 + 0.28269) = 0.61579$, where the two probabilities are computed by means of the Binomial p.m.f.

(B) $P(Y = 2|Y \geq 2) = P(Y = 2, Y \geq 2)/P(Y \geq 2) = P(Y = 2)/P(Y \geq 2) = 0.32801/0.61579 = 0.53267$, where the numerator is computed by means of the Binomial p.m.f.
Text

Framework: Rwanda Government asserts that when its electrical energy plan will be completed, it will cover only 70% of its population...

Exercise 1.

It is a common economical belief that, especially for developing Countries, to made energy more available enhances economic development. A regression analysis, lead on 20 Rwanda areas, attempted to evaluate GDP as function of electrical consumptions (both taken per-capita). The following table summarizes some statistics from a simple regression model where the two variables (GDP per-capita in Dollars, electrical consumption per-capita in Kwh; both are referred to one year) are expressed in natural logarithm

<table>
<thead>
<tr>
<th></th>
<th>( \hat{\beta}_0 )</th>
<th>( \hat{\beta}_1 )</th>
<th>( \hat{\sigma} )</th>
<th>( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>3.82</td>
<td>1.07</td>
<td>0.158</td>
<td>0.6721</td>
</tr>
<tr>
<td>Standard error</td>
<td>0.53</td>
<td>0.18</td>
<td>0.03</td>
<td></td>
</tr>
</tbody>
</table>

The correlation coefficient between the estimators of the two beta coefficients is equal to \(-0.997759\). Formulate an appropriate statistical model and answer the following questions.

(A) Test whether the elasticity coefficient (\( \beta_1 \)) is significantly larger than 0.5 (\( \alpha = 0.05 \)).

(B) Compute the confidence interval for the standard deviation of the error component for a confidence level of 98%.

(C) Estimate the conditional mean of the dependent variable for an electrical consumption equal to 19.7 and provide the corresponding standard error. (Hint: Derive the standard error using the properties of linear transformations of random variables).

(D) Test whether the conditional mean of the dependent variable for an electrical consumption equal to 19.7 is significantly different from 7 (significance level 98%).

(E) Compute the fitted value and the residual for the observation (electrical consumption = 23.9, GDP = 1734).

Exercise 2.

Two experts are asked to formulate predictions on the percentage increase, over the next three years, of electrical per-capita consumptions in different Rwanda areas.

<table>
<thead>
<tr>
<th>Expert</th>
<th>Byumba</th>
<th>Gisenyi</th>
<th>Butare</th>
<th>Cyangugu</th>
<th>Kyunguo</th>
<th>Gabiro</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mwambe</td>
<td>6.8</td>
<td>2.7</td>
<td>6.5</td>
<td>5.1</td>
<td>7.1</td>
<td>5.3</td>
</tr>
<tr>
<td>Ndhoro</td>
<td>5.9</td>
<td>1.8</td>
<td>6.3</td>
<td>7.1</td>
<td>0.2</td>
<td>5.1</td>
</tr>
</tbody>
</table>

(A) Test whether the expert opinions are, on average, significantly different (\( \alpha = 0.05 \)).

(B) By using the information coming from the sample, compute how many areas we would have in the sample for getting a confidence interval for mean difference of size 0.59 at \( \alpha = 0.01 \).

(C) According to the Mwambe expert, estimate via Maximum Likelihood the percentage increase of electrical per-capita consumptions in the areas considered; provide the corresponding standard error.

Solutions

Exercise 1.

Simple linear model \( y_i = \beta_0 + \beta_1 x_i + u_i \) with the usual assumptions, where \( y = \ln(GDP) \) and \( x = \ln(\text{electrical consumption}) \).

(A) Test \( H_0 : \beta_1 = 0.5 \) against \( H_1 : \beta_1 > 0.5 \); test statistic: \( (\hat{\beta}_1 - 0.5)/\text{se}(\hat{\beta}_1) \) with distribution \( T(n-2 = 18) \); rejection region for \( \alpha = 0.05 \): (1.7341, \( \infty \)); sample value of the test statistic (under \( H_0 \)): \( t_{\text{samp}} = 3.1667 \).

(B) Pivot for \( \sigma^2 \): \( \hat{\sigma}^2(n-2)/\sigma^2 \) with distribution \( \chi^2(n-2 = 18) \); interval for \( \sigma^2 \) at \( 1 - \alpha = 0.98 \): \([\hat{\sigma}^2(n-2)/\epsilon_2, \hat{\sigma}^2(n-2)/\epsilon_1] = [0.0129, 0.0641]\) where \( \epsilon_1 = 7.0149 \), \( \epsilon_2 = 34.8053 \), \( \hat{\sigma}^2 = 0.158^2 = 0.025 \); corresponding
interval for $\sigma$: [0.1136, 0.2531].

(C) Point estimate of $\mu_0 = E(y|x_0 = \ln(19.7) = 2.9806) = \beta_0 + \beta_1 x_0$: $\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = 7.0093$; the corresponding standard error is the estimated value of the square root of $V(\hat{\mu}_0) = V(\hat{\beta}_0) + x_0^2 V(\hat{\beta}_1) + 2x_0\text{C}(\hat{\beta}_0, \hat{\beta}_1)$, where $V(\hat{\beta}_0)$ and $V(\hat{\beta}_1)$ are estimated by the squares of the corresponding standard errors and $\text{C}(\hat{\beta}_0, \hat{\beta}_1) = \rho(\hat{\beta}_0, \hat{\beta}_1)/\sigma(\hat{\beta}_0)\sigma(\hat{\beta}_1)$ is estimated by replacing the corresponding sample values. $se(\hat{\mu}_0) = \sqrt{0.0013} = 0.0363$, $\text{C}(\hat{\beta}_0, \hat{\beta}_1) = -0.0952$.

(D) Test $H_0 : \mu_0 = 7$ against $H_1 : \mu_0 \neq 7$; test statistic: $(\hat{\mu}_0 - 7)/se(\hat{\mu}_0)$ with distribution $T(n - 2 = 18)$; rejection region for $\alpha = 0.02$: $[-2.5524, 2.5524]$; sample value of the test statistic (under $H_0$): $t_{samp} = 0.2552$.

(E) The observation considered is $(x_i = \ln(23.9) = 3.1739, y_i = \ln(1734) = 7.4582)$. Fitted value: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = 7.216$; residual: $\hat{u}_i = y_i - \hat{y}_i = 0.2421$.

Exercise 2.
Assumptions: $X_1 = \text{‘MWambe’s opinion’} \sim N(\mu_1, \sigma_1^2)$; $X_2 = \text{‘Ndhoró’s opinion’} \sim N(\mu_2, \sigma_2^2)$.

(A) Since the two samples are paired, we work on $D = X_1 - X_2 \sim N(\mu_D, \sigma_D^2)$. Test $H_0 : \mu_D = 0$ against $H_1 : \mu_D \neq 0$; test statistic (under $H_0$): $(\bar{d} - 0)/se(\bar{d})$ with distribution $T(n - 1)$; acceptation region for $\alpha = 0.05$: $[-2.5706, 2.5706]$; sample value of the test statistic (under $H_0$): $t_{samp} = 0.3674$.

(B) $n = \left(\frac{2sd_D}{A}\right)^2 = 9.7022^2 = 94.1326 \approx 95$, where $A = 0.59, z = 2.5758 (\alpha = 0.01)$.

(C) ML estimator of $\mu_1$: $\bar{X}_1$; corresponding estimate: $\bar{x}_1 = 4.5667$; corresponding standard error: $s_1/\sqrt{n_1} = 0.9265$.

Useful formulas and values:

<table>
<thead>
<tr>
<th>$i$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_{1,i}$</td>
<td>6.8</td>
<td>2.7</td>
<td>6.5</td>
<td>5.1</td>
<td>1</td>
<td>5.3</td>
<td>27.4</td>
</tr>
<tr>
<td>$x_{2,i}$</td>
<td>5.9</td>
<td>1.8</td>
<td>6.3</td>
<td>7.1</td>
<td>0.2</td>
<td>5.1</td>
<td></td>
</tr>
<tr>
<td>$x_{1,i}$</td>
<td>46.24</td>
<td>7.29</td>
<td>42.25</td>
<td>26.01</td>
<td>1</td>
<td>28.09</td>
<td>150.88</td>
</tr>
<tr>
<td>$d_i$</td>
<td>0.9</td>
<td>0.9</td>
<td>0.2</td>
<td>-2</td>
<td>0.8</td>
<td>0.2</td>
<td>1</td>
</tr>
<tr>
<td>$d_i^2$</td>
<td>0.81</td>
<td>0.81</td>
<td>0.04</td>
<td>4</td>
<td>0.64</td>
<td>0.04</td>
<td>6.34</td>
</tr>
</tbody>
</table>

$n = 6, \bar{d} = 0.1667, dev(\bar{d}) = \sum_{i=1}^n d_i^2 - n\bar{d}^2 = 6.1733, s_D^2 = dev(\bar{d})/(n - 1) = 1.2347, s_D = 1.1112, se(\bar{d}) = s_D/\sqrt{n} = 0.4536$.

$n_1 = 6, \bar{x}_1 = 4.5667, dev(\bar{x}_1) = \sum_{i=1}^n x_{1,i}^2 - n_1\bar{x}_1^2 = 25.7533, s_1^2 = dev(\bar{x}_1)/(n_1 - 1) = 5.1507, s_1 = 2.2695$. 
Text

Framework: Shale gas is a natural gas found in shale, a non-porous rock which does not allow the gas to escape. Fracking (hydraulic fracturing) is a process in which a high-pressure fluid (usually chemicals and sand suspended in water, with approximated proportions of 0.5%, 9.5% and 90%, respectively) is pumped into shale to create narrow fractures which allow the gas to be released and captured. In many countries, such a technique is at the center of the current public opinion debate because of the possible damages caused to the environment.

Exercise 1. The following table reports the result of two UK polls about supporting (or not) fracking.

<table>
<thead>
<tr>
<th>Year</th>
<th>Observations</th>
<th>Support (%)</th>
<th>Oppose (%)</th>
<th>Uncertain (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>893</td>
<td>25.08</td>
<td>20.6</td>
<td>54.31</td>
</tr>
<tr>
<td>2014</td>
<td>855</td>
<td>22.46</td>
<td>25.73</td>
<td>51.81</td>
</tr>
</tbody>
</table>

In this exercise you are working as a consultant for the opponents, and they asked you to remove completely undecided people from the analysis (round the counts if required). Once this is done, answer the following questions.

(A) Test whether the percentage of unfavorable people is increased ($\alpha = 0.05$). Explicit the model employed, the $H_0/H_1$ hypotheses, the test statistics employed and its distribution under $H_0$.

(B) Compute the type II error probability of the test at point (A) considering the alternative hypothesis ‘the percentage of people unfavorable to fracking is increased of 8.6 percentage points’.

(C) Assume that, in 2014, the variable ‘Do you oppose to fracking?’ follows a Bernoulli distribution with parameter equal to the value estimated from the sample. Extracting independently 178 people, compute the first and the third quartiles of the number of opponents.

Exercise 2. Fracking requires a huge amount of water but this issue is shared by other energy production methods. The following table compares the the full-cycle amount of water consumed in producing shale gas against the corresponding amount needed by a ‘green’ fuel type (data are expressed in $\log_{10}(\text{liters}/\text{MWh})$, $sd = \sqrt{\text{unbiased sample variance}}$; $mean \mp 1.64 \times sd$ represents the most likely interval for the water consumed).

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Shale gas</th>
<th>Cellulosic ethanol</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean $\mp 1.64 \times sd$</td>
<td>[3.4, 4.23]</td>
<td>[3.38, 4.2]</td>
</tr>
</tbody>
</table>

Assuming that the water consumption is Normally distributed without any assumption on the variances, answer the following questions.

(A) Test whether the two fuel types require a significantly different mean consumption of water ($\alpha = 0.01$).

(B) Compute the value of the log-likelihood, at the estimated values of the parameters, for the last observation ($x_{16} = 3.61$) of the shale gas sample.

(C) Compute an efficient estimate of the common standard deviation if one assumes that the two populations have the same variance; specify the distribution of the corresponding estimator.

Exercise 3. (Only students with 6 CFU) The following table compares the water consumed against the corresponding gas production in a small sample of Texas wells.

| Consumption of water ($\log_{10}(\text{liters})$) | 10.6 | 8.4 | 9.8 | 7.7 | 7.3 |
| Production of gas ($\log_{10}(\text{MWh})$) | 9.4 | 7.2 | 8.3 | 6.7 | 6.2 |

Formulate an appropriate model where: 1) the conditional mean of the gas production is a linear function of the water consumption; 2) the relationship is assumed homoskedastic.

(A) Compute the confidence interval at $\alpha = 0.1$ for the standard deviation of the error component. Explicit the pivot and its distribution.

(B) Compute the point prediction and the corresponding interval at $\alpha = 0.1$ for the conditional mean of the gas production (in the original scale) when the water consumed is 76078000 liters.

1. [http://www.telegraph.co.uk/news/earth/energy/11028321/]
Solutions

Exercise 1. Once undecided people are removed we have

<table>
<thead>
<tr>
<th>Observations</th>
<th>Supporters</th>
<th>Opponents</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>408</td>
<td>224</td>
</tr>
<tr>
<td>2014</td>
<td>412</td>
<td>220</td>
</tr>
</tbody>
</table>

Assumptions: $X_1$: ‘Unfavorable in 2013’ $\sim$ Be($p_1$), $X_2$: ‘Unfavorable in 2014’ $\sim$ Be($p_2$), independent samples.

(A) Test $H_0 : p_2 - p_1 = 0$ against $H_1 : p_2 - p_1 > 0$; test statistic: $(\hat{p}_2 - \hat{p}_1 - 0)/se$ with (approximated) null distribution $N(0, 1)$; rejection region for $\alpha = 0.05$: $(z_{crit} = 1.6449, \infty)$; sample value of the test statistic (under $H_0$): $z_{samp} = 2.38525$.

(B) Consider $H_0 : p_2 - p_1 = 0$ against $H_1 : p_2 - p_1 = 0.086$. Setting 0.086 = c we have $\beta = P(sample \in A|H_1) = P((\hat{p}_2 - \hat{p}_1)/se \leq z_{crit}|H_1) = P((\hat{p}_2 - \hat{p}_1 - c)/se \leq (se \times z_{crit} - c)/se|H_1) = P(Z \leq -0.827|H_1) = 0.20423$.

(C) $X =$ ‘Fracking opponent (in 2014)?’ $\sim$ Be($p = 0.534$) implies $Y =$ ‘Number of supporters among 178’ $\sim Bi(n = 178, p = 0.534)$. The quantities required can be computed using the $N(np = 95.0485, npq = 44.2945)$ approximation to the Binomial: $y_{0.25} = 90.5595 \pm 91, y_{0.75} = 99.5375 \pm 100$.

Useful formulas and values: $\hat{p}_1 = 0.451; \hat{p}_2 = 0.534; se = \sqrt{\hat{p}_1(1 - \hat{p}_1) + \hat{p}_2(1 - \hat{p}_2)} = 0.034797$, where $\hat{q}_1 = 1 - \hat{p}_1 = 0.549, \hat{q}_2 = 1 - \hat{p}_2 = 0.466$.

Exercise 2. Assumptions: $X_1 =$ ‘water consumption by shale gas’ $\sim N(\mu_1, \sigma_1^2)$; $X_2 =$ ‘water consumption by cellulosic ethanol’ $\sim N(\mu_2, \sigma_2^2)$; independent samples. Since $l_1 = \overline{x} - 1.648, l_2 = \overline{x} + 1.648$, the original data can be converted to more familiar quantities as: $\overline{x} = (l_1 + l_2)/2, s = (l_2 - l_1)/3.28$. Then $n_1 = 16, n_2 = 24, \overline{x}_1 = 3.815, \overline{x}_2 = 3.79, s_1 = 0.25305, s_2 = 0.25$.

(A) Test $H_0 : \mu_2 - \mu_1 = 0$ against $H_1 : \mu_2 - \mu_1 \neq 0$; test statistic: $(\overline{x}_2 - \overline{x}_1)/se$ (Satterthwaite-Welsh), with (approximated) null distribution $T(df \approx 32)$; acceptance region for $\alpha = 0.01: [-2.7385, 2.7385]$; sample value of the test statistic (under $H_0$): $t_{samp} = -0.3076$. $se_1 = s_1/\sqrt{n_1} = 0.06326, se_2 = s_2/\sqrt{n_2} = 0.05103$, $se = \sqrt{s_1^2/n_1 + s_2^2/n_2} = 0.08128; df = \frac{(se_1^2 + se_2^2)^2}{\frac{se_1^4}{n_1-1} + \frac{se_2^4}{n_2-1}} = 32.03$.

(B) $\ln f(x_{16}) = 3.61; \mu = 3.815, \sigma = 0.25305$ = $\ln \left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(x_{16}-\mu)^2}{\sigma^2}\right)\right) = \ln(1.13552) = 0.12709$.

(C) $s_p = \sqrt{(s_1^2(n_1 - 1) + s_2^2(n_2 - 1))/(n_1 + n_2 - 2)} = \sqrt{0.0631T} = 0.25121. (n_1 + n_2 - 2)s_p^2/\sigma^2 \sim \chi^2(n_1 + n_2 - 2)$

Exercise 3. $y = \beta_0 + \beta_1 x + u$ with the usual assumptions on $u$.

(A) Pivot for $\sigma^2$: $(n - 2)\hat{\sigma}^2/\sigma^2$ with distribution $\chi^2(n - 2)$; interval for $\sigma^2$ at $\alpha = 0.1: [0.0106, 0.23541]$, where $c_1 = 0.3518$ and $c_2 = 7.8147$; corresponding interval for $\sigma$: $[0.10925, 0.4852]$

(B) Estimate and confidence interval for $\mu_0 = E(y|x_0) = \beta_0 + \beta_1 x_0$. The point estimator is $\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$; the corresponding estimate is $\hat{\mu}_0 = 6.75624$. The pivot is $(\hat{\mu}_0 - \mu_0)/se(\hat{\mu}_0)$ with distribution $T(n-2)$; the corresponding interval at $\alpha = 0.1$ is $[6.54265, 6.96983]$, where $t = 2.3534$.

Useful formulas and values:

<table>
<thead>
<tr>
<th>$i$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_i$</td>
<td>10.6</td>
<td>8.4</td>
<td>9.8</td>
<td>7.7</td>
<td>7.3</td>
<td>43.8</td>
</tr>
<tr>
<td>$y_i$</td>
<td>9.4</td>
<td>7.2</td>
<td>8.3</td>
<td>6.7</td>
<td>6.2</td>
<td>37.8</td>
</tr>
<tr>
<td>$x_i y_i$</td>
<td>112.36</td>
<td>70.56</td>
<td>96.04</td>
<td>59.29</td>
<td>53.29</td>
<td>391.54</td>
</tr>
<tr>
<td>$y_i^2$</td>
<td>88.36</td>
<td>51.84</td>
<td>68.89</td>
<td>44.89</td>
<td>38.44</td>
<td>292.42</td>
</tr>
<tr>
<td>$x_i y_i$</td>
<td>99.64</td>
<td>60.48</td>
<td>81.34</td>
<td>51.59</td>
<td>45.26</td>
<td>338.31</td>
</tr>
</tbody>
</table>

$n = 5, df = n - 2 = 3, \overline{x} = 8.76, \overline{y} = 7.56, dev(x) = 7.852, dev(y) = 6.652, codev(x,y) = 7.182, \hat{\beta}_1 = codev(x, y)/dev(x) = 0.9147, \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = -0.4525, dev(Res) = \sum_{i=1}^{n} \hat{u}_i^2 = dev(y) - \hat{\beta}_1^2 dev(x) = 0.0828, \hat{\sigma}^2 = dev(Res)/df = 0.0276, se(\hat{\mu}_0) = \hat{\sigma}/\sqrt{1/n + (\overline{x} - x_0)^2/\overline{x}} = 0.0908, x_0 = \log_{10}(76078000) = 7.88126.$
Framework: A longtime debated issue concerns the level of crime in urban areas.

Exercise 1. The urban area assigned to the 3rd Police District of New York has been divided into 35 homogeneous portions. The following table reports some statistics concerning the number of auto thefts that have taken place in such portions during the last month.

<table>
<thead>
<tr>
<th>Min</th>
<th>Median</th>
<th>Max</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
<td>6</td>
<td>1.74</td>
<td>1.559</td>
</tr>
</tbody>
</table>

Considering the 35 portions as a simple random sample, choose the appropriate model (among Bernoulli, Poisson, Normal and Gamma) and answer the following questions.

(A) Estimate the parameters of the model by Maximum Likelihood; complement the estimates with the corresponding standard errors.

(B) Compute the value of the log-likelihood, at the estimated values of the parameters, for a sample composed only by \((x_1 = 3, x_2 = 2)\).

(C) Test whether the mean of the distribution assumed is significantly different from 1.1 \((\alpha = 0.01)\).

(D) Compute the power of the test at point (C) considering the alternative hypothesis ‘the mean of the distribution chosen is 1.7’.

Exercise 2. (Only students with 9 CFU) An interesting study concerns the possible relationship between the crime level (taken as dependent variable) and the amount of policemen. The following table reports some statistics concerning the percentage changes in the number of policeman \((\Delta MP)\) and of larcenies \((\Delta L)\) in different New York precincts.

<table>
<thead>
<tr>
<th>(\Delta MP)</th>
<th>(\Delta L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2.8</td>
<td>-2</td>
</tr>
<tr>
<td>3.7</td>
<td>0</td>
</tr>
<tr>
<td>-8.8</td>
<td>15.3</td>
</tr>
<tr>
<td>-11</td>
<td>4.4</td>
</tr>
<tr>
<td>-10.2</td>
<td>3.6</td>
</tr>
</tbody>
</table>

Formulate a convenient statistical model and answer the following questions.

(A) Test whether \(\Delta L\) is significantly associated to \(\Delta MP\) \((\alpha = 0.1)\).

(B) Compute the \(R^2\) index and interpret the result.

Exercise 3. A random variable \(X\) has a distribution with mean \(\mu\) and standard deviation \(\sigma\). Two different estimators of \(\mu\) are compared,

\[
T_1 = \frac{1X_1 - 2X_2 + 3X_3}{2} \quad T_2 = \frac{1X_1 + 2X_2 + 3X_3}{6}
\]

where \((X_1, \ldots, X_n)\) is a simple random sample drawn from \(X\).

(A) Which on is more efficient between \(T_1\) and \(T_2\)? Motivate the answer.

(B) Are the two estimator consistent? Motivate the answer.

Solution

Exercise 1.

Assumptions: Let \(X = \)‘Number of auto thefts by urban area portion’. Considering the characteristics of \(X\), the possible alternatives and the sample values, the only sensible choice is \(X \sim Po(\lambda)\).

(A) ML estimator of \(\lambda\): \(\hat{\lambda} = \overline{X}\); corresponding sample value \(\hat{\lambda} = \overline{x} = 1.74\); corresponding standard error: \(\sqrt{\frac{\lambda}{n}} = \sqrt{0.0497} = 0.223\).

(B) \(\ln f(x_1 = 3, x_2 = 2; \lambda = 1.74) = \ln \left(e^{-\lambda x_1/x_1!} + e^{-\lambda x_2/x_2!}\right) = \ln(0.15411) + \ln(0.2657) = -3.19548\).
(C) Test $H_0: \lambda = 1.1$ against $H_1: \lambda \neq 1.1$; test statistic: $\hat{\lambda} - 1.1)/se_0$ with (approximated) null distribution $N(0, 1)$; acceptance region for $\alpha = 0.01$: $-z_{crit} = -2.5758, z_{crit} = 2.5758$; sample value of the test statistic (under $H_0$): $z_{samp} = 3.61009$.

(D) Consider $H_0: \lambda = 1.1$ against $H_1: \lambda = 1.7$ and set $1.1 = \lambda_0$, $1.7 = \lambda_1$. We have $\gamma = P(sample \in R(H_1)) = 1 - P(sample \in \lambda_0|H_1)\leq (\lambda - \lambda_0)/se_0 \leq z_{crit}|H_1) = 1 - P(-z_{crit} \times se_0 + \hat{\lambda} \leq z_{crit} \times se_0 + \lambda_0|H_1) = 1 - P(-z_{crit} \times se_0 + \lambda_0 - \lambda_1)/se_1 \leq (\lambda - \lambda_1)/se_1 \leq (\hat{\lambda} - \lambda_1)/se_1|H_1) = 1 - P(-4.79445 \leq Z \leq -0.65046/H_1) = 1 - 0.2577 = 0.7423$.

Useful formulas and values:

$\lambda = 1.74; se_0 = \sqrt{1.1/n} = 0.1773, se_1 = 1.7/n = 0.2204$.

Exercise 2. $y = \beta_0 + \beta_1 x + u$ with the usual assumptions on $u$.

(A) Test $H_0: \beta_1 = 0$ against $H_1: \beta_1 \neq 0$; test statistic: $(\hat{\beta}_1 - 0)/se$ with null distribution $T(n - 2)$; acceptance region for $\alpha = 0.1$: $[-2.3534, 2.3534]$; sample value of the test statistic (under $H_0$): $t_{samp} = -1.1444$.

(B) $R^2 = dev(Reg)/dev(y) = 0.2928$. The model explains 29.28% of the total variability of $y$.

Useful formulas and values:

\[
\begin{array}{|c|c|c|c|c|c|c|}
\hline
i & 1 & 2 & 3 & 4 & 5 & Sum \\
\hline
x_i & -2.8 & 3.7 & -8.8 & -11 & -10.2 & -29.1 \\
y_i & -2 & 0 & 15.3 & 4.4 & 3.6 & 21.3 \\
x_i^2 & 7.84 & 13.69 & 77.44 & 121 & 104.04 & 324.01 \\
y_i^2 & 4 & 0 & 234.09 & 19.36 & 12.96 & 270.41 \\
x_i y_i & 5.6 & 0 & -134.64 & -48.4 & -36.72 & -214.16 \\
\hline
\end{array}
\]

$n = 5, df = n - 2 = 3, \overline{x} = -5.82, \overline{y} = 4.26, dev(x) = 154.65, dev(y) = 179.67, codev(x, y) = -90.19, \hat{\beta}_1 = codev(x, y)/dev(x) = -0.583, \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 0.866, dev(Res) = \sum_{i=1}^{n} (\overline{y} - \hat{\beta}_1 \overline{x})^2 = dev(\overline{y} - \beta_1 \overline{x}) = 127.0689, \hat{\sigma}^2 = dev(Res)/df = 42.3563, dev(Reg) = dev(y) - dev(Res) = 52.6031$.

Exercise 3. $X_i \sim [\mu, \sigma]$.

(A) The most efficient estimator is the one with the smallest $MSE = Variance + Bias^2$. The two estimators are linear combinations of the random variables involved, then: $E(T_1) = \mu \rightarrow bias(T_1) = E(T_1) - \mu = \mu - \mu = 0; V(T_1) = 14/4\sigma^2; MSE(T_1) = 14/4\sigma^2$. $E(T_2) = \mu \rightarrow bias(T_2) = E(T_2) - \mu = \mu - \mu = 0; V(T_2) = 14/36\sigma^2; MSE(T_2) = 14/36\sigma^2$. Then $T_2$ is the most efficient one.

(B) Both estimators are clearly not consistent since their MSE, irrespective of the sample size, remains fixed at the value computed at point (A).
Exercise 1. Only for students with 9 CFU
Economic ability is sometimes believed as one of the major determinants of varying public expenditures. The following table reports some statistics, drawn from a 1960 dataset, concerning 22 U.S. Eastern States (\(PE = \) ‘Per capita state and local public expenditures’; \(EAI = \) ‘Economic ability index’, in which income, retail sales, and the value of output (manufactures, mineral, and agricultural) per capita are equally weighted).

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Median</th>
<th>Max</th>
<th>Mean</th>
<th>(\sqrt{\text{biased variance}})</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\ln(PE))</td>
<td>5.209</td>
<td>5.602</td>
<td>5.924</td>
<td>5.585</td>
<td>0.172</td>
<td>0.7257</td>
</tr>
<tr>
<td>(\ln(EAI))</td>
<td>4.177</td>
<td>4.594</td>
<td>4.801</td>
<td>4.566</td>
<td>0.176</td>
<td></td>
</tr>
</tbody>
</table>

Specify a convenient model and answer the following questions.

(A) Test whether \(\ln(PE)\) is significantly related to the independent variable (\(\alpha = 0.1\)).

(B) How much of the total variability of the dependent variable is explained by the model?

(C) Compute the shortest interval that, with a probability 0.9, includes the true value of the conditional mean of the dependent variable for \(EAI = 85.5\).

Exercise 2.
The dataset mentioned in the previous exercise includes also the data of the U.S. Western states. It is sometimes believed that, on average, the PE in Western states (again considered on the ln scale) exceed those of Eastern states. The following table reports some statistics.

<table>
<thead>
<tr>
<th></th>
<th>(n)</th>
<th>(\bar{x})</th>
<th>(s)</th>
<th>(s_*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Western</td>
<td>22</td>
<td>5.7097</td>
<td>0.2096</td>
<td>0.2</td>
</tr>
<tr>
<td>Eastern</td>
<td>22</td>
<td>5.585</td>
<td>0.1758</td>
<td>0.1678</td>
</tr>
</tbody>
</table>

Assuming that \(\ln(PE)\) has the same standard deviation in the two group of states:

(A) Test whether the belief expressed above is statistically significant (\(\alpha = 0.1\)).

(B) Compute the confidence interval for the common standard deviation at the confidence level 90%.

(C) Only for students attending in 2014-2015 and with 9 CFU
Test whether the assumption of equal standard deviation is supported by the data (\(\alpha = 0.1\)).

(D) No for students attending in 2014-2015 and with 9 CFU
Compute the value of the log-likelihood, at the maximum likelihood estimates of the parameters, for the sub-sample \((x_1 = 5.29, x_2 = 5.55)\) in the Eastern group.

Exercise 3. Only students with 6 CFU

\(\text{Economic ability to pay taxes}\) is the economic concept according to which those who have more wealth, or earn higher incomes, should pay more taxes; this principle, that is also related to income redistribution, motivate a \textit{progressive tax system}. The following table reports the people that agree/disagree with this concept, as emerges from a specific question of a survey.

<table>
<thead>
<tr>
<th></th>
<th>Agree</th>
<th>Disagree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Western state</td>
<td>679</td>
<td>547</td>
</tr>
<tr>
<td>Eastern state</td>
<td>549</td>
<td>367</td>
</tr>
</tbody>
</table>

(A) Test whether the percentage of agreeing people is statistically different between the two group of states (\(\alpha = 0.05\)).

(B) Compute the probability of the type II error of the test at point (A) considering the alternative hypothesis ‘the percentage of agreeing people in the West is 4.5 points below that in the East’.

Exercise 4.
Let $X_1$, $X_2$ and $X_3$ be random variables with mean $\mu$ and standard deviation $\sigma$. Define the new variable

$$P = +1.6X_1 + 1.7X_2 + 1.2X_3$$

(A) Compute the mean and the standard deviation of $P$ if we assume that the three variables are independent.

(B) Compute the mean and the standard deviation of $P$ if we assume that the correlation coefficient between each couple of variables is $\rho$. In this case, the standard deviation of $P$ is larger, equal or smaller than in the case at point (A)? Motivate the answer.

Solution

Exercise 1.
Assumptions: $y = \beta_0 + \beta_1 x + u$ with the usual assumptions on $u$, where $y = \ln(PE)$ and $x = \ln(EAIT)$.

(A) Test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$; test statistic: $(\hat{\beta_1} - 0)/\hat{\sigma}$ with null distribution $T(n - 2)$; acceptation region for $\alpha = 0.1$: $[-1.7247, 1.7247]$; sample value of the test statistic (under $H_0$): $t_{samp} = 4.7171$.

(B) $R^2 = dev(Reg)/dev(y) = 0.5266$. The model explains 52.66% of the total variability of $y$.

(C) Pivot for $\mu_0 = E(y|x_0 = \ln(85.5)) = 4.4845$: $(\hat{\mu_0} = \mu_0)/\hat{\sigma} \sim T(n - 2)$; confidence interval for $\mu_0$ at $\alpha = 0.1$: $[5.4468, 5.5566]$, tabulated $t = 1.7247$.

Useful formulas and values:

$$n = 22, df = n - 2 = 20, \bar{x} = 4.57, \bar{y} = 5.58, dev(x) = 0.6815, dev(y) = 0.6508, codev(x,y) = 0.4833, \hat{\beta_1} = codev(x,y)/dev(x) = 0.7092, \hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x} = 2.3468, dev(Res) = \sum_{i=1}^n \hat{\sigma_i}^2 = dev(y) - \hat{\beta_1}^2 dev(x) = 0.3081, \hat{\sigma^2} = dev(Res)/df = 0.0154, \hat{\sigma} = 0.1241, se(\hat{\beta_1}) = \hat{\sigma}/\sqrt{dev(x)} = 0.1503, dev(Reg) = dev(y) - dev(Res) = 0.3428, \hat{\mu_0} = \hat{\beta_0} + \hat{\beta_1}x_0 = 5.5017, se(\hat{\mu_0}) = \hat{\sigma}\sqrt{1/n + (x_0 - \bar{x})^2/dev(x)} = 0.0318$.

Exercise 2.
Assumptions: $X_1 = \ln(PE)$ of a Western state $\sim N(\mu_1, \sigma_1), X_2 = \ln(PE)$ of a Eastern state $\sim N(\mu_2, \sigma_2)$. The sample are of course independent.

(A) Test $H_0 : \mu_2 - \mu_1 = 0$ against $H_1 : \mu_2 - \mu_1 > 0$; test statistic: $(\bar{X}_2 - \bar{X}_1 - 0)/se_p$ ($se_p$ is the standard error computed under $\sigma_1 = \sigma_2$) with null distribution $T(n_1 + n_2 - 2)$; rejection region for $\alpha = 0.1$: $(1.302, +\infty)$; sample value of the test statistic (under $H_0$): $t_{samp} = 2.138$.

(B) Pivot for $\sigma^2_2 = \sigma_2^2$: $S_p^2(n_1 + n_2 - 2)/\sigma^2$ with distribution $\chi^2(n_1 + n_2 - 2)$; confidence interval at $\alpha = 0.1$ for $\sigma^2$: $[0.027, 0.0558]$ ($c_1 = 28.144, c_2 = 58.124$); corresponding interval for $\sigma$: $[0.1644, 0.2363]$.

(C) Test $H_0 : \sigma_1^2 / \sigma_2^2 = 1$ against $H_0 : \sigma_1^2 / \sigma_2^2 \neq 1$; test statistic: $S_1^2/S_2^2$ with null distribution $F(n_1 - 1, n_2 - 1)$; acceptance region for $\alpha = 0.1$: $[0.4798, 2.0842]$; sample value of the test statistic (under $H_0$): $f_{samp} = 1.4215$.

(D) $\ln(f(x_1) = 5.29, x_2 = 5.55; \mu = 5.585, \sigma = 0.1678) = ln(0.50696) + ln(2.32633) = 0.165$, where the two addends are computed as $ln(x_1; \mu, \sigma)$ = $ln \left( \frac{1}{\sqrt{2\pi\sigma}} e^{-(x_1 - \mu)^2/2\sigma^2} \right)$ replacing the values of $x_1, x_2, \mu, \sigma$.

Useful formulas and values:

$$n_1 = 22, n_2 = 22, \bar{x}_1 = 5.7097, \bar{x}_2 = 5.585, s_1^2 = 0.0439, s_2^2 = 0.0309, s_p^2 = [(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]/(n_1 + n_2 - 2) = 0.03742, sp = 0.19344, se_p = sp\sqrt{1/n_1 + 1/n_2} = 0.05832$$

Exercise 3.
Assumptions: $X_1 = \ln(PE)$ of a Western state agrees $\sim Be(p_1), X_2 = \ln(PE)$ of an Eastern state agrees $\sim Be(p_2)$. The samples are independent.

(A) Test $H_0 : p_2 - p_1 = 0$ against $H_1 : p_2 - p_1 \neq 0$; test statistic: $(\bar{X}_2 - \bar{X}_1 - 0)/se$ with null distribution $N(0,1)$; acceptance region for $\alpha = 0.05$: $[-z_{crit} - 1.96, z_{crit} = 1.96]$; sample value of the test statistic (under $H_0$): $z_{samp} = 2.1135$.

(B) Consider $H_0 : p_2 - p_1 = 0$ against $H_1 : p_2 - p_1 = 0.045$ and set $0.045 = c$. We have $\beta = P(sample \in A|H_1) = P(-z_{crit} \leq (\bar{X}_2 - \bar{X}_1 - 0)/se \leq z_{crit}|H_1) = P(-z_{crit} \times se \leq \bar{X}_2 - \bar{X}_1 \leq z_{crit} \times se|H_1) = P((-z_{crit} \times se - c)/se \leq (\bar{X}_2 - \bar{X}_1 - c)/se \leq (z_{crit} \times se - c)/se|H_1) = P(-4.0497 \leq Z \leq -0.1298|H_1) = 0.4483$. 

43
Useful formulas and values: $n_1 = 1226$, $n_2 = 916$, $\bar{x}_1 = 0.55383$, $\bar{x}_2 = 0.59934$, $se = \sqrt{\bar{x}_1(1 - \bar{x}_1)/n_1 + \bar{x}_2(1 - \bar{x}_2)/n_2} = 0.02153$.

Exercise 4.
Assumptions: $X_i \sim (\mu, \sigma)$ for $i = 1, 2, 3$, $P = c_1 X_1 + c_2 X_2 + c_3 X_3$ where $c_1 = +1.6$, $c_2 = +1.7$, $c_3 = +1.2$.
Using the properties of linear combinations of random variables we have, in general, that:

$$E(P) = c_1 E(X_1) + c_2 E(X_2) + c_3 E(X_3)$$
$$V(P) = c_1^2 V(X_1) + c_2^2 V(X_2) + c_3^2 V(X_3) + 2c_1c_2 C(X_1, X_2) + 2c_1c_3 C(X_1, X_3) + 2c_2c_3 C(X_2, X_3).$$

- About $E(P)$: In both (A) and (B), using [1] we get: $E(P) = 4.5\mu$.
- About $\sigma(P)$: The contribution of the variances is the same in both (A) and (B), while the contribution of the covariances is different:
  - In case (A), the random variables are assumed independent, so that $C(X_i, X_j) = 0$ for any $i \neq j$. This implies $V(P) = 6.89\sigma^2$, $\sigma(P) = 2.62488\sigma$;
  - In case (B), the random variables are such that $\rho(X_i, X_j) = \rho$ for any $i \neq j$, so that $C(X_i, X_j) = \rho(X_i)\sigma(X_i)\sigma(X_j) = \rho\sigma = \rho\sigma^2$. This implies $V(P) = 6.89\sigma^2 + 2 \times 6.68\rho\sigma^2$, $\sigma(P) = \sqrt{6.89 + 13.66\rho\sigma}$. $\sigma(P)$ in (B) is larger/equal/smaller than in (A) when $\rho$ is greater/equal/smaller than zero.
Framework: Quality of Life in European countries.

Exercise 1. Only for students with 9 CFU
A statistical analysis aims at relating the Quality of Life, as evaluated by a specific index\(^1\) to the Gini Index\(^2\) in the same year. The hypothesis to investigate is whether the quality of life (taken as dependent variable) is associated with the level of inequality. The following table reports some statistics (\(Q =\) Quality of Life; \(G =\) Gini Coefficient) for some European countries in 2013.

<table>
<thead>
<tr>
<th>(n)</th>
<th>(\frac{1}{n} \sum_{i=1}^{n} q_i)</th>
<th>(\frac{1}{n-1} \sum_{i=1}^{n} (q_i - \overline{q})^2)</th>
<th>(\frac{1}{n} \sum_{i=1}^{n} g_i)</th>
<th>(\frac{1}{n-1} \sum_{i=1}^{n} (g_i - \overline{g})^2)</th>
<th>(\frac{1}{n-1} \sum_{i=1}^{n} (g_i - \overline{g})(q_i - \overline{q}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>122.7</td>
<td>3175.4</td>
<td>31.6</td>
<td>22.7</td>
<td>-67.8</td>
</tr>
</tbody>
</table>

Specify a convenient model and answer the following questions.

(A) Test the hypothesis in the text (\(\alpha = 0.1\)).

(B) Compute the shortest interval that, with probability 0.98, includes the true value of the standard deviation of the error component. Is it guaranteed that the true value of the parameter lies within the interval? Explain.

(C) An observation can be considered an outlier if the corresponding standardized residual lies outside the \((-2.5, 2.5)\). Is the observation \((q_i = 31.86, g_i = 25.6)\) an outlier? Motivate the answer.

Exercise 2.
Some experts support the idea that the quality of life has changed between 2012 and 2015 essentially as a consequence of the Pollution Index. The following table show some statistics for a group of Mediterranean countries.

<table>
<thead>
<tr>
<th></th>
<th>France</th>
<th>Italy</th>
<th>Croatia</th>
<th>Spain</th>
<th>Greece</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>61.8</td>
<td>85.5</td>
<td>49.9</td>
<td>85.6</td>
<td>87.8</td>
</tr>
<tr>
<td>2015</td>
<td>45.9</td>
<td>57.8</td>
<td>32.8</td>
<td>46.3</td>
<td>44.9</td>
</tr>
</tbody>
</table>

(A) Test whether there has been a significant decrease in the mean level of pollution (\(\alpha = 0.01\)).

(B) Only students with 6 CFU Compute the confidence interval for the standard deviation in the 2015 data at \(1 - \alpha = 95\%\).

(D) Compute how many observations we should have in the 2015 sample to get a confidence interval for the mean having half-size 2.16 at \(\alpha = 0.05\).

Exercise 3.
Within Europe, Italy is one of the countries with the largest improvement in the pollution index between 2012 and 2015. A simple random sample of Italian citizens has been asked about whether they perceive such an improvement:

<table>
<thead>
<tr>
<th></th>
<th>Agree</th>
<th>Disagree</th>
<th>Undecided</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>84</td>
<td>275</td>
<td>202</td>
</tr>
<tr>
<td>Female</td>
<td>137</td>
<td>265</td>
<td>168</td>
</tr>
</tbody>
</table>

(A) Only students with 6 CFU Estimate the percentage of undecided people in the population by Maximum Likelihood; complement the estimate with the corresponding standard error.

(B) Only students with 6 CFU Compute the log-likelihood for the subsample \((x_1 = 1, x_2 = 0, x_3 = 0)\) at the ML estimate of the parameters (point A).

(C) Test whether the percentage of undecided people is statistically different between the two genders by using the p-value. Comment the result.

---

Exercise 4.

A specific analysis proposes two scenarios for the pollution index in 2020: the optimistic is characterized by a mean level of 29 and is reached with a probability of 60.8%; the alternative is characterized by a mean equal to 37. Assuming that the index follows a Normal distribution with standard deviation 4.3 in both cases.

(A) Ignoring the scenario that will take place in 2020, compute the probabilities to have a value of the pollution index: 1) less than 23.4; 2) equal to 23.4; 3) larger that 23.4.

(B) Assuming to observe a value of the index larger than 23.4, compute the probability that the optimistic scenario happened.

Solution

Exercise 1.

Assumptions: $y = \beta_0 + \beta_1 x + u$ with the usual assumptions on $u$, where $y = Q$ and $x = G$.

(A) Test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$; test statistic: $(\hat{\beta}_1 - 0)/se$ with null distribution $T(n-2)$; acceptation region for $\alpha = 0.1$: $[-1.7247, 1.7247]$; sample value of the test statistic (under $H_0$): $t_{samp} = -1.1672$.

(B) Pivot for $\sigma^2$: $\hat{\sigma}^2/(n-2)/\sigma^2 \sim \chi^2(n-2)$; confidence interval for $\sigma^2$ at $\alpha = 0.02$: $[1661.8865, 7557.8458]$; corresponding interval for $\sigma$: $[40.7662, 86.9359]$ (tabulated values: $c_1 = 8.2604$, $c_2 = 37.5662$).

(C) Standardized residual corresponding to $(y_i = 31.86, x_i = 25.6)$: $\hat{u}_i/\hat{\sigma} = -1.9466$ ($\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = 140.6207$, $\hat{u}_i = \hat{y}_i - y_i = -108.7607$). The observation is not an outlier.

Useful formulas and values:

\[
n = 22, \ df = n - 2 = 20, \ \bar{x} = 31.6, \ \bar{y} = 122.7, \ dev(x) = 476.7, \ dev(y) = 66683.4, \ codev(x, y) = -1423.8, \ \\
\hat{\beta}_1 = codev(x, y)/dev(x) = -2.9868, \ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 217.0824, \ dev(Res) = \sum_{i=1}^{n} \hat{u}_i^2 = dev(y) - \hat{\beta}_1^2 dev(x) = 62430.8167, \hat{\sigma}^2 = dev(Res)/df = 3121.5408, \hat{\sigma} = 55.8708, \ se(\hat{\beta}_1) = \hat{\sigma}/\sqrt{dev(x)} = 2.559.
\]

Exercise 2.

Assumptions: $X_1 =$ ‘Pollution index 2012’ $\sim N(\mu_1, \sigma_1)$, $X_2 =$ ‘Pollution index 2015’ $\sim N(\mu_2, \sigma_2)$.

<table>
<thead>
<tr>
<th></th>
<th>France</th>
<th>Italy</th>
<th>Croatia</th>
<th>Spain</th>
<th>Greece</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>61.8</td>
<td>85.5</td>
<td>49.9</td>
<td>85.6</td>
<td>87.8</td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td>45.9</td>
<td>57.8</td>
<td>32.8</td>
<td>46.3</td>
<td>44.9</td>
<td></td>
</tr>
</tbody>
</table>

\[
d_1 = 15.9 \quad d_2 = 252.81 \\
x_1 = 45.9 \quad \chi_1 = 2106.81 \\
x_2 = 2106.81 \\
\frac{x_2}{\chi_1} = 3340.84 \\
\frac{x_2}{\chi_1} = 1075.84 \\
\frac{x_2}{\chi_1} = 2143.69 \\
\frac{x_2}{\chi_1} = 2016.01 \\
\frac{x_2}{\chi_1} = 10683.19
\]

(A) Assuming $D = X_1 - X_2 \sim N(\mu_D, \sigma_D)$, test $H_0 : \mu_D = 0$ against $H_1 : \mu_D > 0$; test statistic: $\bar{D}/se(\bar{D})$ with null distribution $T(n-1)$; rejection region for $\alpha = 0.01$: $(3.7469, \infty)$; sample value of the test statistic (under $H_0$): $t_{samp} = 5.161$.

(B) We assume $X_2 = X_1 \sim N(\mu, \sigma)$, pivot for $\sigma^2$: $S^2(n-1)/\sigma^2 \sim \chi^2(n-1)$; confidence interval at $\alpha = 0.05$ for $\sigma^2$: $[28.1544, 64.1746]$ corresponding interval for $\sigma$: $[5.3061, 25.4489]$ (tabulated values: $c_1 = 0.4844$, $c_2 = 11.1433$).

(C) $n = (2\pi\hat{\sigma}/A)^2 = 8.0361^2 = 64.5784 \approx 65$ where $z = 1.96 (\alpha = 0.05)$, $A = 2 \times 2.16 = 4.32, \hat{\sigma} = 8.8562$.

Useful formulas and values: $n_1 = n_2 = n = 5, \bar{d} = 28.58, s_D = 12.38273, se(\bar{D}) = s_D/\sqrt{n} = 5.53773$.

Exercise 3.

Assumptions: $X_1 =$ ‘A male citizen is undecided’ $\sim Be(p_1)$, $X_2 =$ ‘A female citizen is undecided’ $\sim Be(p_2)$. The samples are independent.

(A) We need to pool males and females considering $X =$ ‘A citizen is undecided’ $\sim Be(p)$. ML estimator of $p$: $\hat{X}$, corresponding point estimate: $\bar{x} = 370/1131 = 0.32714$; associated standard error: $se = \sqrt{\bar{x}(1-\bar{x})}/n = \sqrt{0.00019} = 0.1395$. 

46
(B) \( l = \ln f(x_1 = 1, x_2 = 0, x_3 = 0; p = 0.32714) = \ln \prod_{i=1}^{3} p^{x_i}(1 - p)^{1-x_i} = \sum_{i=1}^{3} \ln [p^{x_i}(1 - p)^{1-x_i}] = \ln(0.32714) + \ln(0.67286) + \ln(0.67286) = -1.9098.\)

(C) Test \( H_0 : p_2 - p_1 = 0 \) against \( H_1 : p_2 - p_1 \neq 0 \); test statistic: \( (\bar{X}_1 - \bar{X}_2)/se \) with null distribution \( N(0, 1) \); sample value of the test statistic (under \( H_0 \)): \( z_{samp} = 2.34626 \), p-value = 0.01896.

Useful formulas and values: \( n_1 = 561, n_2 = 570, \bar{x}_1 = 0.36007, \bar{x}_2 = 0.29474, se = \sqrt{\bar{x}_1(1 - \bar{x}_1)/n_1 + \bar{x}_2(1 - \bar{x}_2)/n_2} = 0.02785. \)

Exercise 4.
Assumptions: \( X = \)’Pollution index in 2020’; \( X|O \sim M(\mu_O = 29, \sigma_O = 4.3) \), \( X|A \sim M(\mu_A = 37, \sigma_A = 4.3) \); \( P(O) = 0.608. \)
Setting \( c = 23.4 \) we have:

(A)
- \( P(X < c) = P(X < c|O)P(O) + P(X < c|A)P(A) = 0.05892 \), where \( P(X < c|A) = 0.0964 \) and \( P(X < c|O) = 0.00078; \)
- \( P(X = c) = 0; \)
- \( P(X > c) = 1 - P(X \leq c) = 0.94108. \)

(B) \( P(O|X \geq c) = P(X \geq c|O)P(O)/P(X \geq c) = 0.58378. \)
Text

Framework: One of the consequences of the global warming is the increase of frequency and intensity of heat waves.

Exercise 1.

The variable temperature anomaly is computed as ‘current temperature’ minus ‘long-term (1951-1980) average temperature’. The following table reports some statistics computed using daily observations of this variable, during the period 22 July – 10 August, in the northern hemisphere (although not completely reasonable, assume that the two set of measures are independent).

<table>
<thead>
<tr>
<th>Period</th>
<th>min</th>
<th>max</th>
<th>mean</th>
<th>std.dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1981 – 1990</td>
<td>-2.657</td>
<td>2.922</td>
<td>0.38</td>
<td>1.024</td>
</tr>
<tr>
<td>2001 – 2010</td>
<td>-1.887</td>
<td>4.073</td>
<td>1.224</td>
<td>1.126</td>
</tr>
</tbody>
</table>

(A) Only students with 9 CFU in 2014-2015. Test whether the two standard deviations are significantly different ($\alpha = 5\%$).

(B) Test whether there has been a significant increase in the mean temperature anomaly using the p-value; comment the outcome against usual values of the significance level.

(C) Compute the confidence interval ($\alpha = 5\%$) for the standard deviation in the 2001 – 2010 sample.

(B) Only students with 6 CFU or with 9 CFU not in 2014-2015. How many observations would be needed in the 2001 – 2010 sample to get a confidence interval for the mean of size 0.171 for $\alpha = 0.1$?

Exercise 2. Only for students with 9 CFU

The following table (computed on 12 France cities during August 1-19, 2003) summarizes the results of a simple linear model on the relationship between mortality (taken as dependent variable) and temperature (precise definition of the variables: Mortality = ln(deaths 2003/average deaths 1999-2002), DeltaTemp = Average 2003 temperature – Average 1999-2002 temperature).

<table>
<thead>
<tr>
<th></th>
<th>$\beta_0$</th>
<th>$\beta_1$</th>
<th>$\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>-0.308</td>
<td>0.118</td>
<td>0.12684</td>
</tr>
<tr>
<td>Std Error</td>
<td>0.1946</td>
<td>0.032</td>
<td>0.05672</td>
</tr>
<tr>
<td>Estimated correlation</td>
<td>-0.982</td>
<td>–</td>
<td></td>
</tr>
</tbody>
</table>

Specify a convenient model and answer the following questions.

(A) Estimate the conditional mean of the dependent variable for a value of the independent equal to 7.02 and provide the corresponding standard error. (Hint: Derive the standard error using the properties of linear transformations of random variables).

(B) Compute the confidence intervals at $\alpha = 0.1$ for both the conditional mean and the value of the dependent variable when DeltaTemp is 7.02. Explain the difference in the size of the two intervals.

Exercise 3.

Let $X$ be a random variable with p.d.f.

$$f_X(x) = \begin{cases} 
  1 + x & \text{if } x \in (-1, 0] \\
  1 - x & \text{if } x \in (0, 1] \\
  0 & \text{otherwise}
\end{cases}$$

(A) Explain why $X$ is a well defined random variable.

(B) Draw a graph of the p.d.f. of $X$.

(C) Compute $F_X(0.57)$ ($F_X(x)$ denotes the c.d.f. of $X$ at $x$).
Solution

Exercise 1.
Assumptions: \( X_1 = ‘\text{Temperature anomaly 1981-1990}’ \sim N(\mu_1, \sigma_1), X_2 = ‘\text{Temperature anomaly 2001-2010}’ \sim N(\mu_2, \sigma_2) \). Samples assumed independent.

(A) Test of \( H_0: \sigma_1^2 / \sigma_2^2 = 1 \) against \( H_1: \sigma_1^2 / \sigma_2^2 \neq 1 \); test statistic: \( S_1^2 / S_2^2 \) with null distribution \( F(n_1 - 1, n_2 - 1) \); acceptance region for \( \alpha = 0.05 \): \( (0.7568, 1.3214) \); sample value of the test statistic (under \( H_0 \)): \( f_{\text{Samp}} = 0.827 \).

(B) Test of \( H_0: \mu_2 - \mu_1 = 0 \) against \( H_1: \mu_2 - \mu_1 > 0 \); test statistic: \( (X_2 - X_1) / se \) with null distribution \( N(0, 1) \) (because of the large samples); \( z_{\text{Samp}} = 7.8423 \), p-value = \( P(Z > z_{\text{crit}}|H_0) = 0 \).

(C) Changing notation, we assume \( X_2 = X \sim N(\mu, \sigma) \). Pivot for \( \sigma^2 \): \( S^2(n-1)/\sigma^2 \) with distribution \( \chi^2(n-1) \); confidence interval at \( \alpha = 0.05 \) for \( \sigma^2 \): \( [1.0515, 1.5591] \) corresponding interval for \( \sigma \): \( [1.0254, 1.2486] \) (tabulated values: \( c_1 = 161.83, c_2 = 239.96 \)).

Exercise 2.
Assumptions: \( y = \beta_0 + \beta_1 x + u \) with the usual assumptions on \( u \), where \( y = \text{Mortality} \) and \( x = \text{DeltaTemp} \).

(A) Point estimate of \( \mu_0 = E(y|x_0 = 7.02) = \beta_0 + \beta_1 x_0 \); \( \hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = 0.52036 \); the corresponding standard error is the estimated value of the square root of \( V(\hat{\mu}_0) = V(\hat{\beta}_0) + x_0^2 V(\hat{\beta}_1) + 2 x_0 C(\hat{\beta}_0, \hat{\beta}_1) \), where \( V(\hat{\beta}_0) \) and \( V(\hat{\beta}_1) \) are estimated by the squares of the corresponding standard errors and \( C(\hat{\beta}_0, \hat{\beta}_1) = \rho(\hat{\beta}_0, \hat{\beta}_1) / (\sigma(\hat{\beta}_0) \sigma(\hat{\beta}_1)) \) is estimated by replacing the corresponding sample values. \( se(\hat{\mu}_0) = \sqrt{0.00248} = 0.04976 \), \( \hat{C}(\hat{\beta}_0, \hat{\beta}_1) = -0.00612 \).

(B) Pivot for \( H_0: (\hat{\mu}_0 - \mu_0) / se(\hat{\mu}_0) \) with distribution \( T(n-2) \); confidence interval for \( \mu_0 \) at \( \alpha = 0.1 \): \( [0.43017, 0.61055] \).

Exercise 3.

(A) \( X \) is a well defined r.v. because:

1. \( f_X(x) \geq 0 \quad \forall x \in \mathcal{R} \)

2. \( \int_{-\infty}^{\infty} f_X(x) dx = 1 \) (easily computed because the p.d.f. has a triangular shape)

(B) Not reported.

(C) \( F_X(0.57) = 1/2 + 0.40755 = 0.90755 \) (the second addend is the area of a trapezoid).
Text

Framework: The Global Peace Index (GPI) is a measure of Countries’ peacefulness. It is a synthesis of 22 indicators produced by the Institute for Economics and Peace (IEP – \[http://economicsandpeace.org/\]) and developed in consultation with an international panel of peace experts. The index ranges between one and five, where less means more peaceful.

Exercise 1.
The following table shows some summary statistics on GPI computed, in two different years, on the same sample (82 countries, consider the sample size “large”\[1\]).

<table>
<thead>
<tr>
<th>Year</th>
<th>min</th>
<th>max</th>
<th>mean</th>
<th>variance-covariance matrix</th>
<th>N. increasing GPI over 5 years before</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>1.343</td>
<td>3.37</td>
<td>2.053</td>
<td>0.193</td>
<td>0.18</td>
</tr>
<tr>
<td>2014</td>
<td>1.398</td>
<td>3.65</td>
<td>2.152</td>
<td>0.18</td>
<td>0.234</td>
</tr>
</tbody>
</table>

(A) Test whether the mean GPI has significantly changed in the last five years. Answer comparing the p-value with typical values of the significance level.

(B) Assume a Normal model. Estimate the parameters needed at point (A) by Maximum Likelihood (ML); write the corresponding values and standard errors.

(C) Dominican Republic had GPI equal to 2.004 in 2009 and 2.093 in 2014. Compute the contribution of this country to the log-likelihood at point (B) when the parameters take the ML estimates.

Exercise 2.
Another way to check whether the GPI has significantly changed, is to consider, for each country, the variable increase/not increase of the index between 2009 and 2014.

(A) Has the majority of countries increased its GPI? Answer by using test of hypothesis ($\alpha = 0.05$).

(B) Compute the power of the test at point (A) when the alternative is ‘the percentage of countries increasing GPI between 2009 and 2014 is 64.4%’.

(C) Only students with 6 CFU. Considering the sample given, compute the value of the score function for the model involved in the exercise when the corresponding parameter is 0.68.

(D) Only students with 6 CFU. Draw randomly and independently 5 countries: compute the probability that more than 3 increased their GPI taking parameters equal to the estimated values.

Exercise 3. Only students with 9 CFU
We aim at analyzing whether GPI is associated with the Literacy Rate (LR, evaluated on a $[0,1]$-scale) taken as independent variable. The following table summarizes some sample statistics.

\[
\begin{array}{cccccc}
\sum_{i=1}^{82} GPI_i & \sum_{i=1}^{82} GPI_i^2 & \sum_{i=1}^{82} LR_i & \sum_{i=1}^{82} LR_i^2 & \sum_{i=1}^{82} GPI_i \cdot LR_i \\
176.45 & 398.87 & 70.02 & 62.33 & 147.54 \\
\end{array}
\]

Specify a simple linear model and answer the following questions.

(A) Estimate the coefficients and the variance-covariance matrix of the parameters involved in the conditional mean.

(B) Compute the confidence interval for the standard deviation of the error component ($\alpha = 0.05$).

(C) Only students with 9 CFU following before 2015–2016. Compute the goodness of fit index of the regression model and interpret the result.

(D) Only students with 9 CFU following in 2015–2016. The estimates of the skewness and of the kurtosis of the residuals gave 1.057 and 4.059, respectively. Can we conclude that the distribution of the error component is normal? ($\alpha = 0.05$)

\[1\] For variances and covariances, the denominator is the sample size; the last columns indicates the number of countries in a worse situation than 5 years earlier.
Solution

Exercise 1.
Assumptions: $X_1$ = ‘GPI in 2009’, $X_2$ = ‘GPI in 2014’. Since the two sample are paired (same countries in two different years), assume $D = X_2 - X_1 \sim (\mu_D, \sigma_D^2)$.

(A) Test of $H_0: \mu_D = 0$ against $H_1: \mu_D \neq 0$; test statistic: $(\overline{D} - 0)/se(\overline{D})$ with large sample null distribution $N(0, 1)$; sample value of the test statistic (under $H_0$): $z_{samp} = 3.4422$, $p$-value $= 2 \cdot P((\overline{D} - 0)/se(\overline{D}) > |z_{samp}| |H_0) = 0.0058$.

(B) Assumption: $D = X_2 - X_1 \sim N(\mu_D, \sigma_D^2)$. ML estimate of $\mu_D$: $\overline{d} = 0.099$; ML estimate of $\sigma_D^2$: $s_D^2 = 0.067$; $se(\overline{D}) = \sqrt{\overline{V}(\overline{D})} = \sqrt{s_D^2/n} = \sqrt{0.000817} = 0.028584$; $se(S_{1D}^2) = \sqrt{\overline{V}(S_{1D}^2)} = \sqrt{2s_D^4(n-1)/n^2} = \sqrt{0.000181526} = 0.0104$.

(C) $d_i = x_{2,i} - x_{1,i} = 2.093 - 2.004 = 0.089$; $l_i = \ln f(d_i = 0.089); \mu = 0.099, \sigma^2 = 0.0678) = \ln(1.5401) = 0.4318$.

Useful formulas and values: $n_1 = n_2 = n = 82$, $\overline{x}_1 = 2.053$, $\overline{x}_2 = 2.152$, $s_1^2 = 0.193$, $s_2^2 = 0.234$, $s_{1,2} = 0.18$; $\overline{d} = \overline{x}_2 - \overline{x}_1 = 0.099$. $s_D^2 = s_{1,2}^2 + s_1^2 - 2s_{1,2} = 0.067$, $s_D^2 = n/(n-1)s_{1,2}^2 = 0.06783$, $s_D = 0.26044$, $se(\overline{D}) = s_D/\sqrt{n} = 0.02876$.

Exercise 2.
Assumptions X =‘Country increased its GPI ~ Be(p); simple random sample.

(A) Test of $H_0: p = 0.5$ against $H_1: p > 0.5$; test statistic: $(\hat{p} - 0.5)/se_0$ with large sample null distribution $N(0, 1)$; acceptance region for $\alpha = 0.05$: $(-\infty, 1.645)$; sample value of the test statistic (under $H_0$): $z_{samp} = 3.75467$.

(B) Power for $H_0: p = 0.5$ against $H_0: p > 0.644$; denote the critical value as $z$ and $0.644 = p_1$. $\gamma = P(sample \in R(H_1)) = P(\hat{p} > z \cdot se_0 > 0.5|H_1) = P(\hat{p} > 0.5|se_1 > (z \cdot se_0 + 0.5 - p_1)/se_1|H_1) = P(Z > -1.0057|H_1) = 1 - 0.157278 = 0.842722$.

(C) After some algebra, the score function can be written as $(\overline{x} - p)n/(npq) = 10.2941$, where $p = 0.68$.

(D) $Y$ = ‘number increasing GPI over 5 ~ Bi(n = 5, p > 0.7073). Thus $P(Y > 3) = P(Y \geq 4) = 0.3663 + 0.177 = 0.5433$

Useful formulas and values: $n = 82$, $\hat{p} = \overline{x} = 0.7073$, $se_0 = \sqrt{0.5 \cdot 0.5/n} = 0.0552$, $se_1 = \sqrt{0.644 \cdot (1 - 0.644)/n} = 0.0529$.

Exercise 3.
Assumptions: $y = \beta_0 + \beta_1 x + u$, where $y$ = GPI, $x$ = LR and the classical simple linear regression assumptions on $u$ are taken.

(A) $\hat{\beta}_1 = \text{cod}(\overline{x}, y)/\text{dev}(x) = -1.23283$, $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 3.20455$, $\hat{V}(\hat{\beta}_1) = \hat{\sigma}^2 / \text{dev}(x) = 0.0754$, $\hat{V}(\hat{\beta}_0) = \hat{\sigma}^2 (1/n + \overline{x}^2/\text{dev}(x)) = 0.05731$, $\hat{C}(\hat{\beta}_0, \hat{\beta}_1) = -\hat{\sigma}^2 \overline{x}/\text{dev}(x) = -0.06838$.

(B) Pivot for $\sigma^2$: $df \cdot \hat{\sigma}^2/\sigma^2$ with distribution $\chi^2(df)$; confidence interval for $\sigma^2$ at $\alpha = 0.05$: $[0.1437, 0.2688]$, corresponding confidence interval for $\sigma$: $[0.379, 0.5177]$. Tabulated values: $c_1 = 57.1532, c_2 = 106.6286$.

(C) $R^2 = \text{dev}(\text{REG})/\text{dev}(y) = 0.2013$; the regression model explains 20.13% of the variability of $y$.

(D) Test $H_0: u \sim N \text{ vs } H_1: u \sim \text{N}$; test statistics $JB = n/6 (sk(\overline{u})^2 + (ku(\overline{u}) - 3)^2)/4 = 19.1008$, with null distribution $\chi^2(2)$; critical value at $\alpha = 0.05$: $5.9915$.

Useful formulas and values: $n = 82$, $df = n - 2 = 80$, $\overline{x} = 0.8539$, $\overline{y} = 2.1518$, $\text{dev}(x) = 2.5398$, $\text{dev}(y) = 19.1797$, $\text{cod}(\overline{x}, y) = -3.1311$, $\text{dev(RES)} = \hat{\beta}_2^2 \text{dev}(x) = 3.8601$, $\text{dev(RES)} = \text{dev}(y) - \text{dev(RES)} = 15.3196$, $\hat{\sigma}^2 = \text{dev(RES)}/df = 0.1915$, $sk(\overline{u}) = 1.057$, $ku(\overline{u}) = 4.059$. 

51
Exercise 1. A number of Consumers International (CI) Members have been interviewed on the state of the consumer protection in their respective countries. One of the questions was their opinion about the effectiveness of consumer protection enforcement in the food sector. The following table summarizes some statistics (EU = European Union, NA = North America; assume the sample sizes ‘large’).

<table>
<thead>
<tr>
<th>Opinion</th>
<th>EU &amp; NA Member</th>
<th>African Member</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effective</td>
<td>36</td>
<td>34</td>
</tr>
<tr>
<td>Ineffective</td>
<td>29</td>
<td>31</td>
</tr>
</tbody>
</table>

(A) Test whether the proportion of Members supporting effectiveness is significantly different between the two groups of countries ($\alpha = 0.02$).

(B) Compute the power of the test at point (A) considering the alternative hypothesis “the proportion supporting effectiveness in Africa is lower than in EU & NA by 28 percentage points”.

Exercise 2. Considering again the data in the previous exercise, focus only on African CI members.

(A)- (B) (Only students attending 2015-2016)

Test whether the percentage supporting effectiveness is significantly different from 61% using Likelihood Ratio, Wald and Score test statistics ($\alpha = 0.05$); then compare results. (Some useful formulas: log-Likelihood = $n (\bar{X} \ln p + (1 - \bar{X}) \ln(1 - p));$ Score = $n (\bar{X} - p) / (pq);$ Outer Product of Gradients = Fisher Information = $1/pq$).

(C) (Only students attending before 2015-2016)

Compute the joint contribution to the maximum of the log-likelihood of the first three observations in the sample (0, 1, 0).

(D) (Only students attending before 2015-2016)

Compute how many observations should be included in the sample to get a confidence interval for the proportion of African Members supporting effectiveness having half size 0.094 for $\alpha = 0.05$.

Exercise 3. (Only students with 6 CFU)

Under some experimental conditions, a group of researchers measured the generation time of Escherichia coli bacteria, getting the following statistics (values in minutes; assume the sample size ‘large’).

<table>
<thead>
<tr>
<th>$n$</th>
<th>min</th>
<th>median</th>
<th>max</th>
<th>mean</th>
<th>standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>71</td>
<td>4</td>
<td>12</td>
<td>46</td>
<td>15.2</td>
<td>10.5</td>
</tr>
</tbody>
</table>

Assuming a Gamma model with parameters $\alpha$ and $\beta$, answer the following questions.

(A) Estimate the model parameters by Method of Moments (i.e. matching convenient theoretical expectations of the distribution to the corresponding sample values).

(B) The standard errors associated to the previous estimates are $se(\hat{\alpha}) = 0.37827$ and $se(\hat{\beta}) = 0.0276$. Test whether the distribution of the generation time is Exponential (answer comparing the p-value with typical values of the significance level).

Exercise 4. (Only students with 6 CFU)

Let A and B two production lines of a fish based product. For this product, the law limit for the number of coliform bacteria is 500 (per g of product), but the company planned to stay at a safer value, equal to half the law limit. The number of coliform bacteria is approximately Normally distributed, with mean 205 and variance 294 in line A, mean 209 and variance 294 in line B. Knowing that 28.2% of the production comes from line A:

(A) Compute the probability that a product has a number of coliform bacteria above the firm safety value.

\[http://www.consumersinternational.org\]
Exercise 5. (Only students with 9 CFU)
A guesswork presumes that the number of proved food frauds (F), by country, tends to increase with the level of GDP. The following table reports some statistics (year 2013; consider the sample size ‘large’).

<table>
<thead>
<tr>
<th>n</th>
<th>(\sum_{i=1}^{n} \ln(GDP_i))</th>
<th>(\sum_{i=1}^{n} \ln(GDP_i)^2)</th>
<th>(\sum_{i=1}^{n} \ln(F_i))</th>
<th>(\sum_{i=1}^{n} \ln(F_i)^2)</th>
<th>(\sum_{i=1}^{n} \ln(GDP_i) \cdot \ln(F_i))</th>
</tr>
</thead>
<tbody>
<tr>
<td>92</td>
<td>1004.7</td>
<td>11323.3</td>
<td>393.2</td>
<td>2012.4</td>
<td>4526.5</td>
</tr>
</tbody>
</table>

Specify a convenient simple linear model and answer the following questions.

(A) Test whether the guesswork is supported by the data. Answer comparing the p-value with typical significance levels.

(B) Compute the confidence interval \((\alpha = 0.02)\) for the standard deviation of the error component.

(C) Compute the confidence interval \((\alpha = 0.02)\) for the conditional mean of the number of frauds (in log) for a GDP = 3104000.

(D) Only students attending before 2015–2016. Compute the residual for Denmark (number of frauds = 45, GDP= 336701).

(E) Only students attending in 2015–2016. By doing some model diagnostics, we linearly regressed the squared residuals on the independent variable: the corresponding slope coefficient is estimated at \(-0.18866\) with standard error 0.13662. This results can be used to check a specific model assumption. Specify which one and test it by using the information provided \((\alpha = 0.01)\).

Solution

Exercise 1.
Assumptions: \(X_1 = \text{‘Effective in EU & NA’} \sim Be(p_1), X_2 = \text{‘Effective in Africa’} \sim Be(p_2)\); simple random ‘large’ samples.

(A) Test of \(H_0 : p_1 - p_2 = 0\) against \(H_1 : p_1 - p_2 \neq 0\); test statistic: \((\hat{p}_1 - \hat{p}_2 - 0)/se\) (Wald type statistics) with (approximated) null distribution \(N(0,1)\); corresponding value of the test statistic: \(z_{\text{stat}} = 0.352\); acceptation region for \(\alpha = 0.02\): \([-2.3263, 2.3263]\).

(B) Power for \(H_0 : p_1 - p_2 = 0\) against \(H_1 : p_1 - p_2 = 0.28\); denote critical values as \(-z\) and \(z\) and 0.28 = \(c\). 
\[\beta = P(\text{sample} \in A|H_1) = P(z \leq (\hat{p}_1 - \hat{p}_2)/se) = z_{\text{stat}}^2 = P(\sum_{i=1}^{n} \frac{\ln(F_i)}{\ln(GDP_i) + \ln(F_i)} \leq \gamma) = P(\gamma \approx 0.9098)\]

Useful formulas and values: \(n_1 = 65, n_2 = 65, \hat{p}_1 = 0.5538, \hat{p}_2 = 0.5231, se = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = \sqrt{0.00763951} = 0.0874\)

Exercise 2.
Assumptions: \(X = \text{‘Effective in Africa’} \sim Be(p)\); simple random ‘large’ sample.

(A)–(B) Test of \(H_0 : p = 0.61\) against \(H_1 : p \neq 0.61\) for \(\alpha = 0.05\); we set \(p_0 = 0.61\). Likelihood test: statistic \(2\log(\hat{p} - (\hat{p}_0)) = 2[-44.9853 - 45.9959] = 2.0213\); corresponding (approximated) sample null distribution \(\chi^2(1)\); acceptation region \([0, 3.8415]\). Wald test: statistic \((\hat{p} - 0.61)/\sqrt{pq/n} = -1.4368\); corresponding (approximated) sample null distribution \(N(0,1)\); acceptation region \([-1.96, 1.96]\). Score test: statistic \((\hat{p} - 0.61)/\sqrt{pq/n} = -0.1902\); corresponding (approximated) sample null distribution \(N(0,1)\); acceptation region \([-1.96, 1.96]\).

(C) \(l_1(\hat{p}) + l_2(\hat{p}) + l_3(\hat{p}) = -0.7404 + 0.648 + 0.7404 = -2.1288\), where \(l_i(\hat{p}) = x_i \ln \hat{p} + (1 - x_i) \ln (1 - \hat{p})\).

(D) \(n = \hat{p}q(z/B)^2 = 108.4562 \approx 109\), where \(z = 1.96\) \((\alpha = 0.02)\), \(B = 0.094\).

Useful formulas and values: \(n = 65, \hat{p} = 0.5231, \hat{q} = 1 - \hat{p} = 0.4769, l(p) = n (X \ln p + (1 - X) \ln(1 - p))\); \(l'(p) = n (X - p) / (pq)\); \(OPG = 1/pq\).

Exercise 3.
Assumptions: \(X = \text{'Generation time'} \sim G(\alpha, \beta)\).

(A) The mean/variance relationships, \(E(X) = \alpha/\beta, \ Var(X) = \alpha/\beta^2\), imply \(\hat{\beta} = \bar{x}/s^2 = 0.1379, \hat{\alpha} = \bar{x}\hat{\beta} = 2.0956\).

(B) The exponential distribution is a particular case of the Gamma: \(Exp(\beta) = G(\alpha = 1, \beta)\). Then \(H_0: \alpha = 1\ vs \ H_1: \alpha \neq 1\); \(z_{\text{stat}} = (\hat{\alpha} - 1)/se(\hat{\alpha}) = 2.8963; \) corresponding (approximated) null distribution \(N(0,1); \) p-value \(= 2P(Z > |z_{\text{stat}}|H_0) = 2P(Z > 2.8963|H_0) = 0.0038\).

Useful formulas and values:
\[
\bar{x} = 15.2, \ s^2 = 110.25, \ se(\hat{\alpha}) = 0.37827.
\]

Exercise 4.
Assumptions: \(X = \text{'Number of coliform bacteria'}\); \(X|A \sim N(\mu_A = 205, \sigma_A^2 = 294), \ X|B \sim N(\mu_B = 209, \sigma_B^2 = 294); \) \(P(A) = 0.282\). Law limit for \(X\): \(l = 500\); company’s safety limit: \(s = 250\).

(A) \(P(X > s) = P(X > s|A)P(A) + P(X > s|B)P(B) = 0.00725\).

(B) \(P(A|X > s) = P(X > s|A)P(A)/P(X > s) = 0.16872, \ P(B|X > s) = 1 - P(A|X > s) = 0.83128\).

Useful formulas and values:
\(P(X > s|A) = 0.00434, \ P(X > s|B) = 0.0084, \ P(A) = 0.282\).

Exercise 5.
Assumptions: \(y = \beta_0 + \beta_1 x + u\), where \(x = \ln(GDP)\), \(y = \log(F)\).

(A) Test \(H_0: \beta_1 = 0 \ vs \ H_1: \beta_1 > 0\); test statistics \((\hat{\beta}_1 - 0)/se(\hat{\beta}_1)\), with (approximated) null distribution \(N(0,1); z_{\text{stat}} = 8.81947, \ p\text{-value} = P(Z > z_{\text{stat}}|H_0) = 0\).

(B) Pivot for \(\sigma^2\): \(df \cdot \hat{\sigma}^2/\sigma^2\) with distribution \(\chi^2(df)\); confidence interval for \(\sigma^2\) at \(\alpha = 0.02\): \([1.4344, 2.8829]\); corresponding confidence interval for \(\sigma\): \([1.1977, 1.6979]\). Tabulated values: \(c_1 = 61.7541, c_2 = 124.1163\).

(C) \(x_0 = \ln(3104000) = 14.9482\), pivot for \(\mu_0 = E(y|x_0)\): \((\hat{\mu}_0 - \mu_0)/se(\hat{\mu}_0)\) with (approximated) distribution \(N(0,1); \) confidence interval for \(\mu_0\) at \(\alpha = 0.02\): \([6.1578, 7.7207]\). Tabulated value: \(z = 2.3263\).

(D) \(x_1 = \ln(336701) = 12.727, \ y_1 = \ln(45) = 3.8067, \hat{\gamma}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_1 = 5.4693, \hat{u}_1 = y_1 - \hat{\gamma}_1 = -1.6626\).

(E) A significant slope coefficient in this regression cast doubt on the assumption of homoskedasticity. The corresponding \(t\text{-statistics\)} is \(-0.18866/0.13662 = -1.3809\), that have to be compared with the acceptance region, for \(\alpha = 0.01\, \text{given by} \ [\pm 2.5758, 2.5758]\).

Useful formulas and values:
\(n = 92, \ df = n - 2 = 90, \ \bar{x} = 10.9207, \ \bar{y} = 4.2739, \ \text{dev}(x) = 351.3208, \ \text{dev}(y) = 331.8974, \ \text{cod}(x, y) = 232.4996; \ \hat{\beta}_1 = \text{cod}(x, y)/\text{dev}(x) = 0.6618, \ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = -2.9532, \ \text{dev}(\text{REG}) = \hat{\beta}_1^2 \text{dev}(x) = 153.8652, \ \text{dev}(\text{RES}) = \text{dev}(y) - \text{dev}(\text{REG}) = 178.0322; \ \hat{\sigma}^2 = \frac{\text{dev}(\text{RES})}{df} = 1.9781, \ se(\hat{\beta}_1) = \hat{\sigma}/\sqrt{\text{dev}(\bar{x})} = 0.075, \ \hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = 6.9393, \ se(\hat{\mu}_0) = \hat{\sigma} \sqrt{1/n + (\bar{x} - x_0)/\text{dev}(\bar{x})} = 0.3359\).
26  Test 26.02.2016

Text


Exercise 1. (Only students with 9 CFU)
In order to understand the future costs of the photovoltaic energy, the Institute analyzed the relationship between the module price (in €2014/Wp; Wp = Watt peak), taken as dependent, and the cumulated producer capacity for this kind of energy (in GW). The following table displays some selected statistics (German data, years 1982 – 2013; the variance-covariance is computed with denominator n – 1).

<table>
<thead>
<tr>
<th>Variable</th>
<th>mean vector</th>
<th>variance-covariance matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>log_{10}(Price)</td>
<td>0.265</td>
<td>0.361 – 0.682</td>
</tr>
<tr>
<td>log_{10}(Capacity)</td>
<td>0.207</td>
<td>– 0.682 1.36</td>
</tr>
</tbody>
</table>

Formulate a convenient regression model and answer the following questions.
(A) Compute a goodness of fit measure of the model and interpret the result.
(B) According to one scenario, the power installed in 2050 will be 30000 GW. Compute the corresponding interval estimate of the conditional mean of the module price in €2014/Wp (α = 0.05).

Exercise 2. Some experts have been asked about the future evolution (horizon 2050) of total photovoltaic system costs. Their answers are expressed in terms of cagr (compound annual growth rate) of the module price in €2014/Wp.

<table>
<thead>
<tr>
<th>Expert</th>
<th>Ackermann</th>
<th>Brodbeck</th>
<th>Dirchs</th>
<th>Fiedler</th>
<th>Gunther</th>
</tr>
</thead>
<tbody>
<tr>
<td>cagr (%)</td>
<td>–13.4</td>
<td>–12.3</td>
<td>–14.1</td>
<td>–8</td>
<td>–11.7</td>
</tr>
</tbody>
</table>

Assume that experts’ opinions on cagr are normally distributed.
(A) Estimate the model parameters with good unbiased estimators; complement such estimates by the corresponding standard errors. (Note: Point estimates are useful in Exercises 3 to 5 too)
(B) Compute the value of the log-likelihood function at the parameter values estimated above; compute the same at ML estimates and compare results.
(C) Compute the probability that, in 2050, the module price decreases below 0.0019 €2014/Wp. Answer considering the parameter values estimated from the data and the fact that the module price was 0.2343 €2014/Wp in 2013.

Useful formulas. Log-likelihood: \( l(\mu, \sigma) = -1/2 \left[ n \ln(2\pi) + n \ln \sigma^2 + \sum_{i=1}^{n} (x_i - \mu)^2 / \sigma^2 \right] \). Annual compounding law: \( p_t = p_0(1 + x/100)^t \), where \( p_t \) is the amount after \( t \) years (0 is the start), \( x \) is the cagr in percentage terms.

Exercise 3. (Only students attending before 2015-2016) Consider the data and the setup in Exercise 2.
(A) Test the null hypothesis that the mean parameter is equal to –17 (α = 0.02).
(B) Compute the confidence interval (α = 0.1) for the standard deviation parameter of the model.
(C) Compute how many experts should be interviewed to get a confidence interval for the mean of size 1.2 at a confidence level 0.95.

Exercise 4. (Only students attending in 2015-2016) Consider the data and the setup in Exercise 2.
(A) Test the null hypothesis that the mean parameter is equal to –17 (α = 0.02) by means of the Wald and the Score tests; use exact test distributions, if possible, and compare the results.
(B) Test the null hypothesis that the standard deviation parameter is equal to 2.1 (α = 0.02) by means of the Score test computed according to the asymptotic theory.

Exercise 5. (Only students with 6 CFU)
Future predictions over large horizons are very uncertain. To take into account this point, different scenarios are considered. Data in Exercise 2 are surveyed assuming an intermediate scenario: the corresponding mean and the standard deviation for the percentage cagr can be estimated from such data. The optimistic scenario reduces the mean value of the intermediate by further 1.2 percentage points; on the opposite, the pessimistic adds 2.5 percentage points to the mean value in the intermediate. Assume that, under each scenario, cagr follows a Normal distribution with the mean indicated above and standard deviation identical in the three cases; in addition, assume probabilities equal to 0.32 and 0.22 for the optimistic and pessimistic scenarios, respectively.

(A) Compute the probability that cagr will be larger than −11.4.
(B) Compute the probability of each scenario if cagr would be larger than −11.4.

Solution

Exercise 1.
Assumptions: \( y = \beta_0 + \beta_1 x + u \), where \( y = \log_{10}(\text{Price}) \), \( x = \log_{10}(\text{Capacity}) \).

(A) \( R^2 = 0.9474 \). The model explains 94.7% of the variability of the module price taken in \( \log_{10} \).

(B) \( x_0 = \log_{10}(30000) = 4.4771 \), pivot for \( \mu_0 = E(y|x_0) \): \( (\mu_0 - \hat{\mu}_0)/se(\hat{\mu}_0) \) with distribution \( T(n-2) \); confidence interval for \( \mu_0 \) at \( \alpha = 0.05 \): \([-2.071, -1.681]\). To convert it into an interval for the conditional mean of the module price, use the relationship \( a = \log_{10}(b) \Rightarrow b = 10^a \); this leads to \([0.00849, 0.02082]\) Tabulated value: \( t = 2.042 \).

Useful formulas and values:
\[
\begin{align*}
n &= 32, \quad df = n - 2 = 30, \quad \overline{x} = 0.207, \quad \overline{y} = 0.265, \quad dev(x) = 42.16, \quad dev(y) = 11.191, \quad cod(x,y) = -21.142; \\
\hat{\beta}_1 &= cod(x,y)/dev(x) = -0.50147, \quad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 0.3688, \quad dev(\text{REG}) = \hat{\beta}_1^2 dev(x) = 10.60209, \quad dev(\text{RES}) = dev(y) - dev(\text{REG}) = 0.58991; \quad \hat{\sigma}^2 = dev(\text{RES})/df = 0.01963, \quad R^2 = dev(\text{REG})/dev(y) = 0.94738, \quad \hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = -1.87634, \quad se(\hat{\mu}_0) = \hat{\sigma} \sqrt{1/n + (\overline{x} - x_0)^2/dev(x)} = 0.09541.
\end{align*}
\]

Exercise 2.
Assumptions: \( X \sim \text{cagr} \sim N(\mu, \sigma^2) \).

(A) Point estimators: \( \hat{\mu} = \overline{X} \), \( \sigma^2 = S^2 \); point estimates: \( \hat{\mu} = -11.9 \), \( \hat{\sigma}^2 = s^2 = 5.625 \); standard errors: 
\[
se(\overline{X}) = \sqrt{s^2/n} = \sqrt{1.125} = 1.06066, \quad se(S^2) = \sqrt{2s^4/(n-1)} = \sqrt{15.82031} = 3.97748.
\]

(B) \( l(\mu = \pi, \sigma^2 = s^2) = -1/2 \left[ n \ln(2\pi) + n \ln s^2 + \sum_{i=1}^{n}(x_i - \overline{x})^2/s^2 \right] = -1/2 \left[ n \ln(2\pi) + n \ln s^2 + (n - 1) \right] = -10.91275, \quad l(\mu = \pi, \sigma^2 = S^2) = -1/2 \left[ n \ln(2\pi) + n \ln s^2 + \sum_{i=1}^{n}(x_i - \overline{x})^2/S^2 \right] = -1/2 \left[ n \ln(2\pi) + n \ln s^2 + n \right] = -10.85489.
\]

(C) According to the compounding law, \( P_1 = p_0(1 + X/100)^t \), where \( p_0 = 0.2343 \), \( t = 2050 - 2013 = 37 \), \( X \) is the cagr. Taking \( c = 0.0019 \) we have \( P(F_1 < c) = P(p_0(1 + X/100)^t < c) = P(X < ((c/p_0)^{1/t} - 1) \cdot 100) = P(X < -12.20173) = P(Z < -0.12722) = 0.49398 \).

Useful formulas and values:
\[
\begin{array}{c|cccccc}
\hline
i & 1 & 2 & 3 & 4 & 5 & \text{sum} \\
\hline
x_1 & -13.4 & -12.3 & -14.1 & -8 & -11.7 & -59.5 \\
x_1^2 & 179.56 & 151.29 & 198.81 & 64 & 136.89 & 730.55 \\
\hline
n = 5, \quad \overline{x} = -11.9, \quad dev(\overline{x}) = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 = 22.5, \quad s^2 = dev(\overline{x})/(n-1) = 5.625, \quad s^2 = dev(x)/n = 4.5. \\
\end{array}
\]

\( ^\dagger \)Since \( E(\log X) \neq \log E(X) \) the interval computed is not the confidence interval for the conditional mean of the module price, but is the best we can do with the tools provided in the course.
Exercise 3.

(A) Test of $H_0 : \mu = -17$ against $H_1 : \mu \neq -17$ for $\alpha = 0.02$; test statistic (under $H_0$) $(\bar{X} - -17)/se(\bar{X})$ with null distribution $T(n-1)$; sample value of the test statistic 4.8083; acceptance region $[-3.7469, 3.7469]$. 

(B) Pivot for $\sigma^2$: $(n-1)S^2/\sigma^2$ with distribution $\chi^2(n-1)$; confidence interval for $\sigma^2$ at $\alpha = 0.1$: $[2.3715, 31.6579]$; corresponding interval for $\sigma$: $[1.54, 5.6265]$. 

(C) $n = (2\bar{z}/A)^2 = 7.7474^2 = 60.0228 \approx 61$, where $z = 1.96$ ($\alpha = 0.05$), $\bar{z} = 2.3717$, $A = 1.2$.

Exercise 4.

(A) Test of $H_0 : \mu = -17$ against $H_1 : \mu \neq -17$ for $\alpha = 0.02$. Wald test: statistic (under $H_0$) $(\bar{X} - -17)/se(\bar{X}) = 4.8083$; corresponding exact sample null distribution $T(n-1)$; acceptance region $[-3.7469, 3.7469]$. Arranging formulas we get that in this case the Score statistic is identical to the Wald statistic.

(B) Test of $H_0 : \sigma = 2.1$ against $H_1 : \sigma \neq 2.1$ for $\alpha = 0.02$. Mixing the Score and the OPG for $\sigma^2$, after some simplifications we get the following expression for the Score statistic: $\sqrt{n}/2 \left[ \sum_{i=1}^{n} (X_i - \mu)^2 / (n\sigma^2) - 1 \right]$; the parameters specified by $H_0$ have to be replaced by the corresponding value; nuisance parameters have to be estimated; in this way the previous formula becomes $\sqrt{n}/2 (S^2/2.1^2 - 1)$ whose asymptotic null distribution is $N(0,1)$; sample value: $0.0323$, acceptance region $[-2.3263, 2.3263]$.

(C) Test of $H_0 : \mu = -17, \sigma = 2.1$ against $H_1 : \mu \neq -17, \sigma \neq 2.1$. We set $\mu_0 = -17, \sigma_0 = 2.1$. Likelihood test statistic: $2[\ell(\mu = \bar{X}, \sigma^2 = s^2) - \ell(\mu_0, \sigma_0^2)] = 2[-10.8549 - -25.6003] = 29.4908$; corresponding (approximated) sample null distribution $\chi^2(2)$; acceptance region ($\alpha = 0.1$) $(0.4.6052]$.

Exercise 5.

Assumptions: $X \sim \text{cagd}^r; X|I \sim N(\mu_I = -11.9, \sigma_I^2 = 5.625), X|O \sim N(\mu_O = -13.1, \sigma_O^2 = 5.625), X|P \sim N(\mu_P = -9.4, \sigma_P^2 = 5.625), P(O) = 0.32, P(P) = 0.22$. Set $c = -11.4$.

(A) $P(X > c) = P(X > c|P)P(P) + P(X > c|I)P(I) + P(X > c|O)P(O) = 0.44346$.

(B) $P(P|X > c) = P(X > c|P)P(P)/P(X > c) = 0.17084$, $P(I|X > c) = P(X > c|I)P(I)/P(X > c) = 0.43205$, $P(O|X > c) = 1 - P(P|X > c) - P(I|X > c) = 0.39711$.

Useful formulas and values:

$P(X > c|P) = P(Z > 0.71678) = 0.23675$, $P(X > c|I) = P(Z > 0.21082) = 0.41651$, $P(X > c|O) = P(Z > -0.84327) = 0.80046$. 

57
Text


Exercise 1. United Kingdom (UK) is going toward a referendum on EU membership (end of June 2016). One important issue is whether Britain’s exit from the EU (Brexit for short) will affect the UK economy and the income of their citizens. A small sample of experts has been asked to quantify the percentage variation of the GDP.

<table>
<thead>
<tr>
<th>Change in UK GDP (%)</th>
<th>A. Z</th>
<th>B. Y</th>
<th>C. X</th>
<th>D. W</th>
<th>E. V</th>
<th>F. U</th>
<th>G. T</th>
</tr>
</thead>
<tbody>
<tr>
<td>−1.7</td>
<td>−1.2</td>
<td>−1.5</td>
<td>−0.5</td>
<td>−3.8</td>
<td>−0.5</td>
<td>−2.1</td>
<td></td>
</tr>
</tbody>
</table>

Choose an appropriate model for the variable of interest and answer the following questions.

(A) Compute the confidence interval for the standard deviation at a 99% confidence level.

(B) Is the mean of the predicted change in GDP significantly negative? (α = 0.01).

(C) Only students with 6 CFU. How many academics should be interviewed to get a confidence interval for the mean of size 0.613 with α = 0.05?

Exercise 2. Only students with 9 CFU

Is there any relationship between per capita GDP and the Brexit bent? A researcher collected data of 29 UK regions and regressed the Brexit propensity (in %) on the corresponding GDP per capita (log_{10}(thousands Euros)). Below the table with the main estimates.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std Error</th>
<th>Estimated correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>β₀</td>
<td>54.8456</td>
<td>1.2584</td>
<td>−0.9245</td>
</tr>
<tr>
<td>β₁</td>
<td>−0.1528</td>
<td>0.0372</td>
<td>−</td>
</tr>
<tr>
<td>σ²</td>
<td>6.6768</td>
<td>1.8172</td>
<td>−</td>
</tr>
</tbody>
</table>

Answer the following questions.

(A) Is the Brexit proportion significantly related to the GDP per capita? (α = 0.01)

(B) Compute the residual for the region with GDP per capita = 48000 Euros and the 48.9% of Brexit favorables.

(C) Compute the three deviances involved in the deviance decomposition and derive a goodness of fit index of the regression model. Hint: The three deviances can be obtained from \( se(\hat{\beta}_1) \), \( \hat{\beta}_1 \) and \( \hat{\sigma}^2 \).

Exercise 3. Eurobarometer publishes periodically survey statistics concerning the public opinion on different themes in the EU. This is the opinion of Italians about the future of EU (the two samples are independent).

<table>
<thead>
<tr>
<th></th>
<th>Interviewed</th>
<th>Optimistic</th>
<th>Pessimistic</th>
<th>Don’t know</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>1067</td>
<td>41.2%</td>
<td>50.9%</td>
<td>7.9%</td>
</tr>
<tr>
<td>2014</td>
<td>1068</td>
<td>47.6%</td>
<td>46.3%</td>
<td>6.2%</td>
</tr>
</tbody>
</table>

Answer the following questions.

(A) Test whether the percentage of pessimists has significantly changed (α = 0.01). Explicit the model employed, the \( H_0/H_1 \) hypotheses, the test statistics employed and its distribution under \( H_0 \).

(B) Compute the power of the test at point (A) considering the alternative hypothesis ‘the percentage of pessimists decreased by 6.2 percentage points’.

(C) Assume that, in 2014, the variable ‘Are you pessimistic about the future of EU?’ follows a Bernoulli distribution with parameter equal to the value estimated from the sample. Extracting independently 100 people, compute the eighth decile of the number of pessimists in this group.

Solution

Exercise 1.
Assumptions: $X$ = ‘Change in UK GDP’ $\sim N(\mu, \sigma^2)$.

(A) Pivot for $\sigma^2$: $(n-1)S^2/\sigma^2$ with distribution $\chi^2(n-1)$; confidence interval for $\sigma^2$ at $\alpha = 0.01$: [0.4145, 11.3782]; corresponding interval for $\sigma$: [0.6438, 3.3732]; tabulated values: $c_1 = 0.6757, c_2 = 18.5476$.

(B) Test of $H_0: \mu = 0$ against $H_1: \mu < 0$ for $\alpha = 0.01$; test statistic (under $H_0$) $(\bar{X} = 0)/se(\bar{X})$ with null distribution $T(n-1)$; sample value of the test statistic $-3.773$; rejection region $(-\infty, -3.1427)$.

(C) $n = (2\hat{\sigma}/\text{size})^2 = 52.3998 \approx 53$, where $z = 1.96 (\alpha = 0.05), \hat{\sigma} = 1.132, \text{size} = 0.613$.

Useful formulas and values:

<table>
<thead>
<tr>
<th>$i$</th>
<th>$x_i^2$</th>
<th>$x_i^2$</th>
<th>$x_i^2$</th>
<th>$x_i^2$</th>
<th>$x_i^2$</th>
<th>$x_i^2$</th>
<th>$x_i^2$</th>
<th>sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1.7</td>
<td>-1.2</td>
<td>-1.5</td>
<td>-0.5</td>
<td>-3.8</td>
<td>-0.5</td>
<td>-2.1</td>
<td>-11.3</td>
</tr>
<tr>
<td>2</td>
<td>2.89</td>
<td>1.44</td>
<td>2.25</td>
<td>0.25</td>
<td>14.44</td>
<td>0.25</td>
<td>4.41</td>
<td>25.93</td>
</tr>
</tbody>
</table>

$n = 7, df = n - 1 = 6, \bar{X} = -1.6143, \text{dev}(\bar{X}) = \sum_{i=1}^{n} x_i^2 - n\bar{X}^2 = 7.6886, s^2 = \text{dev}(\bar{X})/(n - 1) = 1.2814, se(\bar{X}) = \hat{\sigma}/\sqrt{n} = 0.4279$.

Exercise 2.

Assumptions: $y = \beta_0 + \beta_1x + u$, where $y = \%$ of Brexit supporters, $x = \log_{10}(GDP/1000)$.

(A) Test of $H_0: \beta_1 = 0$ against $H_1: \beta_1 \neq 0$ for $\alpha = 0.01$; test statistic (under $H_0$) $(\hat{\beta}_1 - 0)/se(\hat{\beta}_1)$ with null distribution $T(n - 2)$; sample value of the test statistic $-4.1075$; acceptance region $(-2.7707, 2.7707)$.

(B) Residual for (percentage) $y_i = 48.9, GDP_i = 48000$: $y_i - \hat{y}_i = -5.6887$ where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1x_i = 54.5887$ and $x_i = \log_{10}(48000/1000) = 1.6812$.

(C) $se(\hat{\beta}_1) = \hat{\sigma}/\sqrt{\text{dev}(\bar{X})} \Rightarrow \text{dev}(\bar{X}) = \hat{\sigma}^2/\text{se}(\hat{\beta}_1)^2 = 4824.8352; \text{dev}(\text{REG}) = \hat{\beta}_1^2\text{dev}(\bar{X}) = 112.6495, \hat{\sigma}^2 = \text{dev}(\text{RES})/df \Rightarrow \text{dev}(\text{RES}) = \hat{\sigma}^2 \cdot df = 180.2736, \text{dev}(y) = d(\text{REG}) + d(\text{RES}) = 292.9231$. Then $R^2 = \text{dev}(\text{REG})/\text{dev}(y) = 0.3846$.

Useful formulas and values:

$n = 29, df = n - 2 = 27$.

Exercise 3. Assumptions: $X_1$ = ‘Pessimist in 2013?’ $\sim Be(p_1), X_2$ = ‘Pessimist in 2014?’ $\sim Be(p_2)$, independent samples.

(A) Test of $H_0: p_2 - p_1 = 0$ against $H_1: p_2 - p_1 \neq 0$ for $\alpha = 0.01$; test statistic $(\hat{p}_2 - \hat{p}_1)/se$ with approximated null distribution $N(0, 1)$; sample value of the test statistic: $-2.1286$; acceptance region $[-2.5758, 2.5758]$.

(B) Power of the test at point (A) considering $H_1: p_2 - p_1 = -0.062$. Denoting $z = 2.5758$ and $c = -0.062$ we have $\beta = P(\text{sample} \in A|H_1) = P(-z < (\hat{p}_2 - \hat{p}_1)/se < z|H_1) = P(-z \cdot se < \hat{p}_2 - \hat{p}_1 < z \cdot se|H_1) = P(-z \cdot se - c)se < (\hat{p}_2 - \hat{p}_1 - c)/se < (z \cdot se - c)/se|H_1) = P(-z - c/se < Z < z - c/se|H_1) = P(Z < z-c/se) - P(Z < -z-c/se) = P(Z < 5.4447) - P(Z < 0.2931) = 0.6153 = 0.3847$. Then $\gamma = 1 - \beta = 0.6153$.

(C) $X = \%$ of pessimists in 2014 $\sim Be(p = 0.463)$. $Y = \%$ of pessimists among 100’ $\sim Bi(n = 100, p = 0.463)$. Using the Normal approximation $N(np = 46.3, npq = 24.8631)$ we need to find $c$ such that $0.8 = P(Y \leq c): 0.8 = P(Y \leq c) = P((Y - np)/\sqrt{npq} \leq (c - np)/\sqrt{npq}) = P(Z \leq z)$. Finding $z = 0.8416$ from the Normal table we get $(c - np)/\sqrt{npq} = z$ and then $c = np + z\sqrt{npq} = 50.4966$, that could be rounded to 50.
Exercise 1.

The following table reports some statistics of cannabis wholesale and retail prices (in Euros/Gram) computed on a sample of 21 Western-Central European countries (source: UNODC = United Nations Office on Drugs and Crime\textsuperscript{1}; denominators of variances and covariances are \( n-1 \)).

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \log_{10}(\text{wholesale price}) )</td>
<td>0.732</td>
<td>0.248</td>
<td></td>
</tr>
<tr>
<td>( \log_{10}(\text{retail price}) )</td>
<td>1.144</td>
<td>0.203</td>
<td>0.773</td>
</tr>
<tr>
<td>Difference</td>
<td>0.412</td>
<td>0.158</td>
<td></td>
</tr>
</tbody>
</table>

(Only students with 9 CFU from here on out.) Take the log of the retail price as dependent and answer the following questions.

(A) United Kingdom had wholesale price = 1.553 and retail price = 4.4. Compute fitted value and residuals for this country.

(B) Compute the confidence interval for the standard deviation of the error component at \( \alpha = 0.01 \).

(C) Under the proportionality hypothesis, one unit increase in the independent increases the dependent by the same amount. Test such an hypothesis (\( \alpha = 0.02 \)).

(D) Compute the confidence interval (\( \alpha = 0.05 \)) for the value of the retail price in a country whose wholesale price is 7.763.

Exercise 2. (Only students with 6 CFU)

Consider the premise and the data of the previous exercise. Assuming that log-prices follow a normal distribution, answer the following questions.

(A) Test whether difference between the means of the log-prices (retail minus wholesale) is significantly different from one (\( \alpha = 0.01 \)).

(B) Compute the confidence interval at 95\% for the standard deviation of the difference between the means.

Exercise 3.

The Quinnipiac University Poll\textsuperscript{2} at the question Do you think that the use of marijuana should be made legal in the United States, or not? (May 2016), got the answers summarized in the following table. Warnings: Total is not the sum of the first two columns because of removal of independent (neither Republican nor Democrat) people; the last Total refers to a comparable poll in November 2012.

<table>
<thead>
<tr>
<th></th>
<th>Republicans</th>
<th>Democrats</th>
<th>Total</th>
<th>Total (Nov. 2012)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Should</td>
<td>196</td>
<td>427</td>
<td>810</td>
<td>714</td>
</tr>
<tr>
<td>Should not</td>
<td>370</td>
<td>194</td>
<td>673</td>
<td>658</td>
</tr>
</tbody>
</table>

(A) What’s the difference in the proportion of legalization supporters between Republicans and Democrats? Specify a symbol for each parameter involved in the required quantity; estimate each of such symbols by Maximum Likelihood (ML) and write the formulas of the corresponding estimator; provide the estimate of the required quantity together with the corresponding standard error.

(B) Focusing on the two totals only, test whether the percentage of legalization supporters has significantly changed (\( \alpha = 0.05 \)). Explicit the model employed, the \( H_0/H_1 \) hypotheses, the test statistics employed and its distribution under \( H_0 \).

(C) Do the test at point (B), for the same significance level, by applying the Likelihood Ratio test. (Hints: for a single \( X \sim Be(p) \) with sample size \( n \), the log-lik is \( l(p) = n(\pi \ln p + (1 - \pi) \ln(1 - p)) \); samples from the

\textsuperscript{1}https://www.unodc.org/unodc/index.html
\textsuperscript{2}http://www.pollingreport.com/drugs.htm
different parties are independent; under $H_0$ the common $p$ can be estimated pooling the samples).

(D) Compute the power of the test at point (B) considering the alternative hypothesis ‘the percentage of legalization supporters increased by 4.7 percentage points’.

(E)-(F) (Only students with 6 CFU). Interviewing a group of 5 people supporting the same party (Republican or Democratic), we knew that 4 of them favor legalization. Assuming that prior and likelihood probabilities can be estimated from the data, compute the posterior probabilities of supporting each party.

Solution

Exercise 1.

Assumptions: $y = \beta_0 + \beta_1 x + u$, where $x = \log_{10}$ (retail price) $y = \log_{10}$ (wholesale price).

(A) Residual for United Kingdom, $(x_i = \log_{10}(1.553) = 0.1912, y_i = \log_{10}(4.4) = 0.6435)$: $y_i - \hat{y}_i = -0.1583$ where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = 0.8018$.

(B) Pivot for $\sigma^2$: $(n-2)\hat{\sigma}^2/\sigma^2$ with distribution $\chi^2(n-2)$; confidence interval for $\sigma^2$ at $\alpha = 0.01$: $[0.008597, 0.048467]$; corresponding interval for $\sigma$: $[0.0927, 0.2202]$; tabulated values: $c_1 = 6.844, c_2 = 38.5823$.

(C) Test of $H_0$: $\hat{\beta}_1 = 1$ against $H_1$: $\hat{\beta}_1 \neq 1$ for $\alpha = 0.02$; test statistic (under $H_0$) $(\hat{\beta}_1 - 1)/se(\hat{\beta}_1)$ with null distribution $T(n-2)$; sample value of the test statistic $-3.0828$; acceptation region $(-2.5395, 2.5395)$.

(D) Consider for the moment $y_0 = y(x_0)$ where $x_0 = \log_{10}(7.763) = 0.89$. Pivot for $y_0$: $(y_0 - \hat{y}_0)/se(\hat{y}_0)$ with distribution $T(n-2)$, where $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ and $se(\hat{y}_0) = \hat{\sigma}/\sqrt{n} + (\pi - x_0)^2/\text{dev}(x) + 1$; corresponding interval: $[0.9582, 1.5298]$, tabulated value 2.093. On the other side, the quantity of interest is retail price $= 10^{y_0}$, so that the searched interval is $[10^{0.9582} = 9.0824, 10^{1.5298} = 33.8673]$.

Useful formulas and values:

\[ n = 21, \ df = n - 2 = 19, \ \bar{x} = 0.732, \ \bar{y} = 1.144, \ \text{dev}(x) = 1.2301, \ \text{dev}(y) = 0.8242, \ \text{codev}(x, y) = 0.7783, \ \hat{\beta}_1 = 0.6327, \ \hat{\beta}_0 = 0.6808, \ \text{dev(Res)} = \text{dev}(y) - \hat{\beta}_1^2 \text{dev}(x) = 0.3317, \ \hat{\sigma}^2 = \text{dev(Res)}/df = 0.0175, \ \text{se}(\hat{\beta}_1) = \hat{\sigma}/\sqrt{\text{dev}(x)} = 0.1191, \ \hat{y}_0 = 1.244, \ \text{se}(\hat{y}_0) = 0.1365. \]

Exercise 2.

Assumptions: $X_1 = \log_{10}$ (wholesale price), $X_2 = \log_{10}$ (retail price), $D = X_2 - X_1 \sim N(\mu_D, \sigma_D^2)$.

(A) Test of $H_0$: $\mu_D = 1$ against $H_1$: $\mu_D \neq 1$ for $\alpha = 0.01$; test statistic (under $H_0$) $(\hat{\mu}_1 - 1)/se(\hat{\mu}_1)$ with null distribution $T(n-1)$; sample value of the test statistic $-17.0541$; acceptation region $(-2.8453, 2.8453)$.

(B) Pivot for $\sigma_D^2$: $(n-1)S_D^2/\sigma_D^2$ with distribution $\chi^2(n-1)$; confidence interval for $\sigma_D^2$ at $\alpha = 0.05$: $[0.0146, 0.0521]$, corresponding interval for $\sigma_D^2$: $[0.1209, 0.2282]$, tabulated values: $c_1 = 9.5908, c_2 = 34.1696$.

Useful formulas and values:

\[ n = 21, \ df = n - 1 = 20, \ \bar{d} = 0.412, \ s_D = 0.158, \ \text{se}(\bar{d}) = \hat{\sigma}_D/\sqrt{n} = 0.0345. \]

Exercise 3.

Assumptions at point (A): $X_1 = ‘Does a Republican support legalization?’ \sim \text{Be}(p_1)$, $X_2 = ‘Does a Democrat support legalization?’ \sim \text{Be}(p_2)$, independent samples. Assumptions at points (B)-(D): $X_1 \sim \text{Support legalization in 2016?’} \sim \text{Be}(p_1)$, $X_2 \sim \text{Support legalization in 2012?’} \sim \text{Be}(p_2)$, independent samples.

(A) Estimate $p_2 - p_1$ by $\hat{p}_2 - \hat{p}_1$, where $\hat{p}_1 = \sum_{i=1}^{n_1} X_{1,i}/n_1$, $\hat{p}_2 = \sum_{j=1}^{n_2} X_{2,i}/n_2$ (ML estimators); $\hat{p}_2 - \hat{p}_1 = 0.3463 - 0.6876 = 0.3413$, $se = \sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2} = \sqrt{0.000746} = 0.02731$.

(B) Test of $H_0$: $p_1 - p_2 = 0$ against $H_1$: $p_1 - p_2 \neq 0$ for $\alpha = 0.05$; test statistic $(\hat{p}_2 - \hat{p}_1)/se$ with approximated null distribution $N(0,1)$; sample value of the test statistic: $1.38$, acceptation region $[-1.96, 1.96]$; $\hat{p}_1 - \hat{p}_2 = 0.5462 - 0.5204 = 0.0258, se = \sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2} = \sqrt{0.000349} = 0.01868$.

(C) Test of $H_0$: $p_1 - p_2 = 0$ against $H_1$: $p_1 - p_2 \neq 0$ for $\alpha = 0.05$; test statistic $2(\hat{L}(H_0))$ with approximated null distribution $\chi^2(1)$; sample value of the test statistic: $1.9036$; rejection region $(3.8415, \infty)$; $\hat{L} = l_1(\hat{p}_1) + l_2(\hat{p}_2) = n_1(\bar{x}_1 \ln \hat{p}_1 + (1-\bar{x}_1) \ln(1-\hat{p}_1)) + n_2(\bar{x}_2 \ln \hat{p}_2 + (1-\bar{x}_2) \ln(1-\hat{p}_2)) = -1021.6002 + 949.8548 = -1971.4549$; $\hat{L}(H_0) = l_1(\hat{p}) + l_2(\hat{p}) = n_1(\bar{x}_1 \ln \hat{p} + (1-\bar{x}_1) \ln(1-\hat{p})) + n_2(\bar{x}_2 \ln \hat{p} + (1-\bar{x}_2) \ln(1-\hat{p})) = -1022.0581 + 950.3486 = -1972.4068$, $\hat{p} = (810 + 714)/(1483 + 1372) = 0.5338$. 61
(D) Power of the test at point (B) considering $H_1 : p_1 - p_2 = 0.047$. Denoting $z = 1.95996398454005$ and $c = 0.047$ we have $\beta = P(sample \in A|H_1) = P(-z < (\hat{p}_1 - \hat{p}_2)/se < z|H_1) = P(-z \cdot se < \hat{p}_1 - \hat{p}_2 < z \cdot se|H_1) = P(-(z \cdot se - c)/se < (\hat{p}_1 - \hat{p}_2 - c)/se < (z \cdot se - c)/se|H_1) = P(-(z - c)/se < Z < z - c/se|H_1) = P(Z < z - c/se) - P(Z < -z - c/se) = P(Z < -0.5557) - P(Z < -4.4756) = 0.289206 - 4e-06 = 0.289202$. Then $\gamma = 1 - \beta = 0.710798$.

(E) Assumptions: $X =$‘Number in favor of legalization over 5’; $R =$‘Republican’, $D =$‘Democratic’; $X|R \sim Bi(n = 5, p_R = 0.3463), X|D \sim Bi(n = 5, p_D = 0.6876)$; $P(R) = 0.4768, P(D) = 0.5232$. Denote $x = 4$.

$P(R|X = x) = P(X = x|R)P(R)/P(X = x) = 0.1093, P(D|X = x) = 1 - P(R|X = x) = 0.8907, P(X = x) = P(X = x|R)P(R) + P(X = x|D)P(D) = 0.2051, P(X = x|R) = 0.047, P(X = x|D) = 0.3492$. 

62
Framework: If the bee disappeared off the face of the Earth, man would have no more than four years left to live (ascribed to Albert Einstein). CCD (Colony Collapse Disorder) is the phenomenon that occurs when the majority of worker bees in a colony disappear and leave behind the queen, plenty of food and a few nurse bees to care for the remaining immature bees and the queen. While such disappearances have occurred throughout the history of apiculture, latest years registered a rise in the number of disappearances of honey bee colonies. Several possible causes have been proposed, but no single proposal has gained widespread acceptance among the scientific community.

Exercise 1.

The following table summarizes the outcome of two surveys lead in different time periods on different samples of beekeepers. The variable of interest is the yearly percentage loss due to CCDs.

<table>
<thead>
<tr>
<th></th>
<th>2007-2011</th>
<th>2012-2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of beekeepers</td>
<td>109</td>
<td>93</td>
</tr>
<tr>
<td>Yearly percentage loss (mean ±1 sd)</td>
<td>0.34 ± 0.15</td>
<td>0.46 ± 0.18</td>
</tr>
</tbody>
</table>

Assuming that the yearly percentage losses are normally distributed and considering the two sample sizes as large, answer the following questions.

(A) By using the p-value, test whether the mean percentage loss due to CCD is different between the two periods. Comment the result.

(B) Do the test at point (A) by applying the Likelihood Ratio test and considering $\alpha = 0.001$. (Hints: for a single $X \sim N(\mu, \sigma^2)$ with sample size $n$, the log-lik is $l(\mu, \sigma^2) = -0.5n \left(\ln(2\pi\sigma^2) + \left(s^2 + (\bar{x} - \mu)^2\right)/\sigma^2\right)$, where $\bar{x}$ and $s^2$ are the sample mean and the biased sample variance, respectively; samples from the different periods are independent; under $H_0$, the common $\mu$ can be estimated pooling the samples).

(C) Assume that the standard deviations in the two periods are equal. Estimate the common standard deviation and compute the corresponding confidence interval for $\alpha = 0.05$.

Exercise 2. (Only students with 9 CFU)

One of the supposed causes of CCD is the excess of monoculture diet, while, according to this hypothesis, bees should receive food from a variety of sources/plants. The following table reports data from a sample of 28 beekeepers in Pennsylvania (the percentage in the second variable is computed as “Number of CCD episodes / (Number of beehives × Number of years)” multiplied by 100).

<table>
<thead>
<tr>
<th>Number of cultures in the diet</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\log_{10}(\text{Percentage of CCD episodes})$</td>
<td>1.929</td>
<td>0.9</td>
<td>-0.502</td>
</tr>
<tr>
<td>1.641</td>
<td>0.108</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Answer the following questions.

(A) How much of the variability of the dependent variable is explained by the model?

(B) Is the independent variable significant in explaining the dependent variable? Use $\alpha = 0.02$.

(C) Compute the confidence interval ($\alpha = 0.01$) for the expectation of the dependent variable considering a beekeeping with 4 cultures in the diet.

(D) Estimate how much the dependent variable changes (on average) when the Number of cultures in the diet moves from 1 to 2; provide the corresponding standard error.

Exercise 3.

According to another hypothesis, CCDs are mainly caused by the varroa distructor, a mite infesting beehives. Some researchers experimented a treatment based on powdered sugar (unfortunately, a very mild treatment), counting the parasites fallen on a plate. To avoid to count all mites, we divided the plate in 529 squares of the same size and we counted the parasites on a sample of such squares. The following table reports the outcome.
Assume that the number of varroas counted per square follows a Poisson distribution and answer the following questions.

(A) Estimate all parameters of the assumed distribution by Maximum Likelihood; provide the corresponding standard errors.

(B) (Only students with 6 CFU). Compute the joint contribution of the first two observations to the total log-likelihood in its maximum.

(C) (Only students with 6 CFU). Test whether the mean number of varroas by square is significantly different from 3 using the score test ($\alpha = 0.02$). (Useful formulas: log-likelihood $= n(\ln \lambda - \ln \sigma) + \sum_{i=1}^{n} \ln x_i!$, score $= n(\bar{x}/\lambda - 1)$, Information $= n/\lambda$, where $n$ is the sample size, $\lambda$ is the Poisson parameter; $\bar{x}$ is the sample mean)

(D) (Only students with 6 CFU). Use the inference at point (A) to estimate the total number of mites in the plate and the corresponding standard error.

(E) (Only students with 6 CFU). Use the inference at point (C) to get a confidence interval at $\alpha = 0.05$ for the total number of parasites.

Solution

Exercise 1.
Assumptions: $X_1 = \text{‘yearly percentage loss 2007-2011’} \sim N(\mu_1, \sigma_1^2)$, $X_2 = \text{‘yearly percentage loss 2012-2013’} \sim N(\mu_2, \sigma_2^2)$; independent, large samples.

(A) Test of $H_0 : \mu_2 - \mu_1 = 0$ against $H_1 : \mu_2 - \mu_1 \neq 0$; test statistic (under $H_0$) $\left(\bar{X}_2 - \bar{X}_1\right)/\text{se}(\bar{X}_2 - \bar{X}_1)$ with approximated null distribution $N(0,1)$; sample value of the test statistic 5.094599; p-value $= 2P(Z > |5.094599|) = 0$.

(B) Same test as in (A) for $\alpha = 0.001$; test statistic $2(\tilde{t} - \tilde{t}(H_0))$ with approximated null distribution $\chi^2(1)$; sample value of the test statistic: 27.0895; rejection region (10.8276, $\infty$); $\tilde{t} = t_1(\hat{\mu}_1, \hat{\sigma}_1^2) + t_2(\hat{\mu}_2, \hat{\sigma}_2^2) = 52.6241 + 27.515 = 80.1391$, where $t_j(\hat{\mu}_j, \hat{\sigma}_j^2) = l_j(\bar{x}_j, s_j^2) = -0.5n_j \left(\ln(2\pi s_j^2) + \left(s_j^2 + (\bar{x}_j - x_j)^2\right)/s_j^2\right)$ for $j = 1, 2$; $\tilde{t}(H_0) = l_1(\hat{\mu}_1, \hat{\sigma}_1^2) + l_2(\hat{\mu}_2, \hat{\sigma}_2^2) = 45.1623 + 21.432 = 66.5943$, where $l_j(\hat{\mu}_j, \hat{\sigma}_j^2) = l_j(\bar{x}_j, s_j^2) = -0.5n_j \left(\ln(2\pi s_j^2) + \left(s_j^2 + (\bar{x}_j - x_j)^2\right)/s_j^2\right)$ for $j = 1, 2$.

(C) Assuming $\sigma_1 = \sigma_2 = \sigma$ we can estimate the common $\sigma$ using the estimator $S_{\text{pooled}}^2 = \sqrt{(df_1S_1^2 + df_2S_2^2)/df}$ with $df_1 = n_1 - 1$, $df_2 = n_2 - 1$, $df = df_1 + df_2 = n_1 + n_2 - 2$; it gives $\tilde{\sigma} = s_{\text{pooled}} = \sqrt{0.0271} = 0.1645$. Pivot for $\sigma^2$: $dfS_{\text{pooled}}^2/\sigma^2$ with distribution $\chi^2(df)$; confidence interval for $\sigma^2$ at $\alpha = 0.05$: [0.0224, 0.0333]; corresponding confidence interval for $\sigma$: [0.1498, 0.1823]; tabulated values $c_1 = 102.728$, $c_2 = 241.0579$.

Useful formulas and values:

\begin{align*}
n_1 = 109, n_2 = 93, \bar{x}_1 = 0.64, \bar{x}_2 = 0.46, s_1 = 0.15, s_2 = 0.18, s_1^2 = 0.0225, s_2^2 = 0.0324, \text{se}(\bar{X}_2 - \bar{X}_1) = \sqrt{s_1^2/n_1 + s_2^2/n_2} = \sqrt{0.00055} = 0.02355, \bar{X} = (n_1\bar{x}_1 + n_2\bar{x}_2)/(n_1 + n_2) = 0.39525.
\end{align*}

Exercise 2.
Assumptions: $y = \beta_0 + \beta_1 x + u$, where $y = \log_{10}(\text{Percentage of CCD episodes})$, $x =$number of cultures in the diet.

(A) $R^2 = 0.252$: the model explains 25.2% of the variability of $y$.

(B) Test of $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ for $\alpha = 0.02$; test statistic (under $H_0$) $(\hat{\beta}_1 - 0)/\text{se}(\hat{\beta}_1)$ with null distribution $T(n-2)$; sample value of the test statistic $-2.9597$; acceptance region $(-2.4786, 2.4786)$.

(C) Pivot for $\mu_0 = E(y|x_0 = 4) = \beta_0 + \beta_1 x_0$; $(\hat{\mu}_0 - \hat{\mu}_0)/\text{se}(\hat{\mu}_0)$ with distribution $T(n-2)$; confidence interval for $\mu_0$ at $\alpha = 0.01$: [1.3889, 1.6436]; tabulated value: $t = 2.7787$.

(D) The value to estimate is the difference between $E(y|x = 2) = \beta_0 + \beta_1 2$ and $E(y|x = 1) = \beta_0 + \beta_1 1$, a difference that is equal to $\beta_1$, whose estimate and standard error (below) are already computed to do the test at point (B).
Useful formulas and values:
\[ n = 28, \ df = n - 2 = 26, \ \bar{x} = 1.929, \ \bar{y} = 1.641, \ dev(x) = 21.87, \ dev(y) = 0.3149, \ codev(x, y) = -1.3174, \]
\[ \hat{\beta}_1 = -0.0602, \ \hat{\beta}_0 = 1.7572, \ dev(Res) = dev(y) - \hat{\beta}_1^2 dev(x) = 0.2356, \ \hat{\sigma}^2 = dev(Res)/df = 0.0091, \ se(\hat{\beta}_1) = \hat{\sigma}/\sqrt{dev(x)} = 0.0204, \]
\[ \hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = 1.5162, \ se(\hat{\mu}_0) = \sqrt{\hat{\sigma}^2(1/n + (x - x_0)^2/dev(x))} = \sqrt{0.0021} = 0.0458. \]

Exercise 3. Assumptions: \( X = \text{’Number of varroas in a square’} \sim Po(\lambda). \)

(A) Maximum Likelihood estimator of \( \lambda \): \( \hat{\lambda} = X \); corresponding estimate: \( \hat{\lambda} = 26/10 = 2.6 \); corresponding standard error: \( se(\hat{\lambda}) = \sqrt{\hat{\lambda}/n} = \sqrt{0.26} = 0.5099. \)

(B) The contribution of \( x_1 \) and \( x_2 \) to the log-likelihood is \( (x_1 \ln \lambda - \lambda - \ln x_1!) + (x_2 \ln \lambda - \lambda - \ln x_2!) \). At the maximum, \( \lambda = \hat{\lambda} \), so that the two addends give \( -1.52523 + -1.52523 = -3.05045. \)

(C) Test of \( H_0 : \lambda = 3 \) against \( H_0 : \lambda \neq 3 \) for \( \alpha = 0.02 \); test statistic: \( score(\lambda_0)/\sqrt{n} = n(\hat{\lambda}/\lambda_0 - 1)/\sqrt{n/\lambda_0} = (\hat{\lambda} - \lambda_0)/\sqrt{\lambda_0/n} \) with approximated null distribution \( N(0, 1) \); sample value of the test statistic: \( -0.7303 \); acceptance region \([-2.3263, 2.3263]\].

(D) True quantities are related as follow: \( mean = total/529; \) then \( total = mean \cdot 529. \) Since an estimator of the true mean is \( \bar{X} \), an estimator of the true total is \( \hat{total} = \bar{X} \cdot 529. \) Then \( \hat{total} = \hat{\bar{X}} \cdot 529 = 1375.4 \) and \( se(\hat{total}) = se(\bar{X}) \cdot 529 = 269.7381. \)

(E) From the (approximated) distribution of \( \bar{X} \) we can derive the (approximated) distribution of \( total. \) Then, pivot for \( total: (total - \hat{total})/se(\hat{total}) \) with approximated distribution \( N(0, 1); \) corresponding confidence interval for \( \alpha = 0.05 \): \([846.723, 1904.077]\); tabulated value \( z = 1.96. \)

Text

Framework: Kenya has an unemployment rate of 40% but, since many years, produces the best athletes for long distance competitions (like marathon). Different scientific works tried to catch the secret.

Exercise 1.

One of the most important determinants of endurance capacities is $\dot{V}O_2\text{max}$ (maximal oxygen consumption) which reflects the aerobic physical fitness of an individual. A paper reported the following statistics on sedentary vs active Kenyan boys (sd, the standard deviation, is computed using degrees of freedom at the denominator).

<table>
<thead>
<tr>
<th></th>
<th>sample size</th>
<th>median</th>
<th>mean ± sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sedentary</td>
<td>11</td>
<td>51.7</td>
<td>52.6 ± 3.4</td>
</tr>
<tr>
<td>Active</td>
<td>19</td>
<td>55.4</td>
<td>54.4 ± 5.2</td>
</tr>
</tbody>
</table>

Assuming that $\dot{V}O_2\text{max}$ is normally distributed, not necessarily with equal variances, answer the following questions.

(A) Specify the formula of the joint log-likelihood of the two samples and compute its maximum.

(Hints: For a single $X \sim N(\mu, \sigma^2)$, the log-lik is $l(\mu, \sigma^2) = -n \left( \ln(2\pi\sigma^2) + \left( s^2 + (x - \mu)^2 \right) / \sigma^2 \right) / 2$, where $n$, $\mu$ and $s^2$ are the sample size, the sample mean and the biased sample variance, respectively; samples from different groups are independent).

(B) Test whether the two sets have a significantly different mean level of $\dot{V}O_2\text{max}$ ($\alpha = 0.02$). In addition, specify the model, the hypotheses, the test statistic and its distribution under the null.

(C) (Only students attending in 2014-2015 and 2016-2017) Test whether the two variances could be assumed equal ($\alpha = 0.01$). In addition, specify the model, the hypotheses, the test statistic and its distribution under the null.

(D) (Only students attending since 2015-2016) Do the test at point (A) (at the same significance level) by applying the Likelihood Ratio test. In addition, specify the test statistic and its distribution under the null.

(Hints: consider hints at point (A); under the null hypothesis, the common $\mu$ can be estimated pooling the samples).

Exercise 2. (Only students with 9 CFU)

Another paper investigated the relationship between $\dot{V}O_2\text{max}$ (taken as dependent variable) and the Body Mass Index ($\text{BMI} = \text{mass} / \text{height}^2$) in a sample of 20 male adolescents of the Nandi district. The OLS inference of the simple linear model (assuming normal errors) gave

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimated parameter</th>
<th>Estimated Variance-Covariance matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>95.0857</td>
<td>48.8251 -3.1386</td>
</tr>
<tr>
<td>BMI</td>
<td>-1.9947</td>
<td>-3.1386 0.2036</td>
</tr>
</tbody>
</table>

and $\hat{\sigma}^2 = 8.6349$. Answer the following questions.

(A) Test whether the BMI is significant in the regression ($\alpha = 0.02$). In addition, specify the hypotheses, the test statistic and its distribution under the null.

(B) Compute the confidence interval at ($\alpha = 0.02$) for the variance of the error term. In addition, specify the pivot and its distribution.

(C) Compute how much of the variability of $\dot{V}O_2\text{max}$ is explained by the model.

(Hint: The needed quantities have to be derived from the statistics provided).

(D) Compute the confidence interval for the conditional mean of $\dot{V}O_2\text{max}$ for a boy with BMI equal to 16.91 ($\alpha = 0.05$). In addition, specify the pivot and its distribution.

(Hint: The quantity required is a linear combination of the estimated coefficients).

Exercise 3. (Only students with 6 CFU).
To give an idea of the physical activity of Nandi adolescents, the following table reports the distance village-school (in km) of a simple random sample of boys.

| 10.6 | 12.3 | 13.6 | 7.3 | 11.5 | 9.7 |

Assuming that the distance traveled is normally distributed, answer the following questions.

(A) Can we conclude that, as a mean, the Nandi adolescents travel more than 2.9 = 18 km each day to attend school? (α = 0.05)

(B) Compute the power of the test at point (A) considering the alternative hypothesis 'to attend school, the boys of the district travel, as a mean, 2·10 = 20 km each day (approximate the distribution of the test statistics with the Normal).

(C) Assume that X is a Normal random variable, with unknown mean µ and variance σ² equal to the value estimated from the data; draw a simple random sample of 10 observations from X. If we estimate µ by the sample mean, compute the probability that the absolute value of the estimation error is less that 0.599.

(D) Compute the sample size needed in order the absolute value of the estimation error is lower than the error margin given at point (C) with probability equal to 0.9.

Solution

Exercise 1.
Assumptions: $X_1 = \bar{V}O_2 \max \text{ of sedentary boys' } \sim N(\mu_1, \sigma_1^2), X_2 = \bar{V}O_2 \max \text{ of active boys' } \sim N(\mu_2, \sigma_2^2)$; independent samples.

(A) Independence implies that the joint likelihood is the product of the likelihoods of the two samples; this, in turn, implies that the joint log-likelihood is the sum of the two log-likelihoods: $l(\mu_1, \sigma_1^2; \mu_2, \sigma_2^2) = l_1(\mu_1, \sigma_1^2) + l_2(\mu_2, \sigma_2^2) = -n_1 \ln(2\pi \sigma_1^2) + \frac{(\bar{x}_1^2 - (\mu_1)^2)}{\sigma_1^2}/2 - n_2 \ln(2\pi \sigma_2^2) + \frac{(\bar{x}_2^2 - (\mu_2)^2)}{\sigma_2^2}/2$. Joint maximum: $l(\hat{\mu}_1 = \bar{x}_1, \sigma_1^2 = \bar{x}_1^2 - \bar{x}_2^2, \mu_2 = \bar{x}_2, \sigma_2^2 = \bar{x}_2^2) = -n_1 \ln(2\pi \sigma_1^2) - n_2 \ln(2\pi \sigma_2^2) + (\mu_1 - \mu_2)^2/2 = -28.54565 + 57.77071 - 86.31636$.

(B) Test $H_0: \mu_1 - \mu_2 = 0$ against $H_1: \mu_1 - \mu_2 \neq 0$ for $\alpha = 0.02$; test statistic (under $H_0$) $(\bar{x}_1 - \bar{x}_2)/se(\bar{x}_1 - \bar{x}_2)$ with approximated null distribution $T(df = 27.4531 \approx 27)$; sample value of the test statistic $-1.14437$; acceptance region: $[-2.47266, 2.47266]$.

(C) Test $H_0: \sigma_1^2/\sigma_2^2 = 1$ against $H_0: \sigma_1^2/\sigma_2^2 \neq 1$ for $\alpha = 0.01$; test statistic (under $H_0$) $S_1^2/S_2^2$ with null distribution $F(df_1 = 10, df_2 = 18)$; sample value of the test statistic 0.42751; acceptance region: $[0.18726, 4.03046]$.

(D) Same test as in (B) for the same $\alpha$; test statistic $2(\bar{x} - \hat{\mu}(H_0))$ with approximated null distribution $\chi^2(1)$; sample value of the test statistic $2(max(l - max_{H_0}(l))) = 1.68339$; rejection region (5.41189, $\infty$). Using the notation at point (A): max($l(\hat{\mu}_1 = \bar{x}, \sigma_1^2 = \bar{x}^2, \mu_2 = \bar{x}, \sigma_2^2 = \bar{x}^2)$) = $-29.2258 + 57.93225$ = 87.15005.

Useful formulas and values:

\begin{align*}
n_1 = 11, n_2 = 19, df_1 = 10, df_2 = 18, \bar{x}_1 = 52.6, \bar{x}_2 = 54.4, s_1 = 3.4, s_2 = 5.2, s_1^2 = 11.56, s_2^2 = 27.04, s_1^2 = 10.50099, s_2^2 = 25.61684, se_1 = s_1/\sqrt{n_1} = 1.02514, se_2 = s_2/\sqrt{n_2} = 1.19296, se(\bar{x}_2 - \bar{x}_1) = \sqrt{s_1^2/n_1 + s_2^2/n_2} = \sqrt{2.47407} = 1.57292, Satterthwaite-Welsh df = (s_1^2 + s_2^2)^2/(s_1^2/df_1 + s_2^2/df_2) = 27.45314 \approx 27, \bar{x} = (n_1 \bar{x}_1 + n_2 \bar{x}_2)/(n_1 + n_2) = 53.74.
\end{align*}

Exercise 2.
Assumptions: $y = \beta_0 + \beta_1 x + u$, where $y = \bar{V}O_2 \max, x = BMI$.

(A) Test $H_0: \beta_1 = 0$ against $H_1: \beta_1 \neq 0$ for $\alpha = 0.02$; test statistic (under $H_0$) $(\hat{\beta}_1 - 0)/se(\hat{\beta}_1)$ with null distribution $T(n - 2)$; sample value of the test statistic $-4.42068$; acceptance region $(-2.55238, 2.55238)$.

(B) Pivot for $s^2: \sigma^2(n - 2)/\sigma^2$ with distribution $\chi^2(n - 2)$; confidence interval for $\sigma^2$ at $\alpha = 0.02$: $[4.46565, 22.15683]$; corresponding confidence interval for $\sigma$: $[2.11321, 4.70711]$; tabulated values: $c_1 = 7.01491, c_2 = 34.80531$.

(C) $R^2 = dev(REG)/dev(y) = 0.5205$: the model explains 52.05% of the variability of $y$.

(D) Pivot for $\mu_0 = E(y|x_0 = 16.91) = \beta_0 + \beta_1 x_0$: $(\hat{\mu}_0 - \mu_0)/se(\hat{\mu}_0)$ with distribution $T(n - 2)$; confidence interval for $\mu_0$ at $\alpha = 0.05$: $[59.3659, 63.3448]$; tabulated value: $t = 2.1009$. 
Exercise 3. Assumptions: $X = \text{Distance village-school and back} \sim N(\mu, \sigma^2)$.

<table>
<thead>
<tr>
<th>$x_i$</th>
<th>10.6</th>
<th>12.3</th>
<th>13.6</th>
<th>7.3</th>
<th>11.5</th>
<th>9.7</th>
<th>65</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_i^2$</td>
<td>112.36</td>
<td>151.29</td>
<td>184.96</td>
<td>53.29</td>
<td>132.25</td>
<td>94.09</td>
<td>728.24</td>
<td></td>
</tr>
</tbody>
</table>

Then $n = 6$, $df = n - 2 = 5$, $\bar{x} = 10.8333$, $dev(x) = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 24.0733$, $s^2 = dev(x)/df = 4.8147$, $s = 2.1942$, $se(\bar{x}) = s/\sqrt{n} = 0.8958$.

(A) Test $H_0: \mu = 18$ against $H_1: \mu > 18$ for $\alpha = 0.05$; test statistic: $(\bar{X} - \mu_0)/se(\bar{X})$ with null distribution $T(n-1)$; sample value of the test statistic: $-8.0004$; rejection region ($t_{crit} = 2.015, \infty$).

(B) Power for $H_0: \mu = \mu_0 = 18$ against $H_1: \mu = \mu_1 = 20$: $\gamma = P(\text{sample} \in R|H_1) = P((\bar{X} - \mu_0)/se(\bar{X}) > t_{crit}|H_1) = P(\bar{X} > \mu_0 + t_{crit} \cdot se(\bar{X})|H_1) = P((\bar{X} - \mu_1)/se(\bar{X}) > (\mu_0 + t_{crit} \cdot se(\bar{X}) - \mu_1)/se(\bar{X})|H_1) = P(Z > -0.2176|H_1) = 0.5861$ (computed using the normal instead of the $T$ distribution).

(C) Let the error margin be $c = 0.599$. Under the assumptions we have $\bar{X} \sim N(\mu, \sigma^2/n)$, so that $P(|\bar{X} - \mu| < c) = P(-c < \bar{X} - \mu < c) = P(-c/se(\bar{X}) < (\bar{X} - \mu)/se(\bar{X}) < c/se(\bar{X})) = P(-0.8633 < Z < 0.8633) = 0.612$ ($se(\bar{X}) = \sigma/\sqrt{n} = 0.6939$).

(D) $0.9 = P(|\bar{X} - \mu| < c) = P(-c < \bar{X} - \mu < c) = P(-c/se(\bar{X}) < (\bar{X} - \mu)/se(\bar{X}) < c/se(\bar{X})) = P(-z < Z < z)$ implies $z = 1.6449$; then $z = c/se(\bar{X}) = c/(\sigma/\sqrt{n})$ implies $n = (z\sigma/c)^2 = 36.3051 \approx 37$.  

68
Text

Framework: Violence against women.

Exercise 1. (Only students with 9 CFU)

Violence against women could be related to a more general violence climate in the country. The following table reports some statistics relative to a sample of $n = 9$ African and Asian developing countries: $PV$ is the percentage of women subjected to at least one violence episode (physical and/or sexual) during the whole lifetime; $GPI$ is the Global Peace Index (higher means less peaceful; minimum possible value equal to 1).

<table>
<thead>
<tr>
<th>$\sum_{i=1}^{n} PV_i$</th>
<th>$\sum_{i=1}^{n} PV_i^2$</th>
<th>$\sum_{i=1}^{n} GPI_i$</th>
<th>$\sum_{i=1}^{n} GPI_i^2$</th>
<th>$\sum_{i=1}^{n} PV_i \times GPI_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>372.3</td>
<td>17351.71</td>
<td>11.07</td>
<td>15.3</td>
<td>499.87</td>
</tr>
</tbody>
</table>

Formulate a convenient simple linear regression model with errors $\epsilon \sim N(0, \sigma^2)$ and answer the following questions. (Note: a linear regression model is not completely justified with these data.)

(A) Test whether $GPI$ is significant in the regression ($\alpha = 0.1$). In addition, specify the hypotheses, the test statistic and its distribution under the null.

(B) Compute the confidence interval for the conditional mean of $PV$ when $GPI$ takes its minimum possible value ($\alpha = 0.05$). In addition, specify the pivot used and its distribution.

(C) Compute how much of the variability of the dependent variable is explained by the model.

(D) (Only students attending before 2015-2016) Compute fitted value and residual for Vietnam ($PV = 34.4$, $GPI = 0.666$).

(E) (Only students attending since 2015-2016) Test whether the Normality assumption on the error term is justified ($\alpha = 0.1$) considering the following sample statistics on residuals: $n^{-1} \sum_{i=1}^{n} \hat{u}_i^3/\hat{\sigma}^3 = -0.1401$, $n^{-1} \sum_{i=1}^{n} \hat{u}_i^4/\hat{\sigma}^4 = 0.7621$. In addition, specify the hypotheses, the test statistic and its distribution under the null.

Exercise 2.

DHS (Demographic and Health Surveys, Domestic Violence Module country reports) reported the following survey statistics about physical and/or sexual violence in Uganda (percentage = percentage of women subjected to at least one episode during the whole lifetime):

<table>
<thead>
<tr>
<th>Year</th>
<th>sample size</th>
<th>percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>221</td>
<td>51.6</td>
</tr>
<tr>
<td>2006</td>
<td>193</td>
<td>60.6</td>
</tr>
</tbody>
</table>

(A) Specify the formula of the joint log-likelihood of the two samples and compute its maximum. (Hints: For a single $X \sim Be(p)$, the log-lik is $l(p) = n[\overline{x} \ln p + (1 - \overline{x}) \ln q]$, where $n$ and $\overline{x}$ are the sample size and mean, respectively, and $q = 1 - p$; samples from different groups are independent).

(B) Test whether, in the 2006–2011 period, there has been a statistically significant decrease in the percentage of women suffering violences ($\alpha = 0.1$). In addition, specify the test statistic and its distribution under the null.

(C) Compute the type 2 error probability of the test at point (B) when the alternative is “the probability to be subject to a violence in 2011 is lower than in 2006 by 11.3 percentage points”.

(D) (Only students attending since 2015-2016) Do the test at point (B) (at the same significance level) by applying the Likelihood Ratio test.

(Hints: consider hints at point (A); under the null hypothesis, the common $p$ can be estimated pooling the samples).

Exercise 3.

Violence (physical and/or sexual) against a woman can have an impact on her ensuing working career. The following table reports some sample statistics on the hourly wage of two samples of US women coming from a similar condition 5 years before (the standard deviations are computed using degrees of freedom at denominator).

<table>
<thead>
<tr>
<th>Suffering violence</th>
<th>Sample size</th>
<th>Average</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not suffering violence</td>
<td>21</td>
<td>9.74</td>
<td>2.25</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>14</td>
<td>2.54</td>
</tr>
</tbody>
</table>

Assuming a Normal distribution of wages with equal standard deviations, answer the following questions.

(A) (Only students with 6 CFU or with 9 CFU and attending before 2015-2016)
Compute the confidence interval at $\alpha = 0.01$ for the difference between the means of the two groups of women.

(B) (Only students with 6 CFU attending in 2016-2017)
Test whether the assumption of equal standard deviations is supported by the data ($\alpha = 0.05$).

(C) (Only students with 6 CFU)
We know that, in a large population of US women in a condition comparable with the sample, 3.91% of them have been victims of a violence. Assuming that the wage of a woman is Normally distributed with parameters equal to the estimated values (depending on its group), compute the probability that a woman gains less than 8.378 per hour.

(D) (Only students with 6 CFU)
Under the conditions of point (C), compute the probability that a woman gaining less than 8.378 per hour has been subjected to a violence.

Solution

Exercise 1.
Assumptions: $y = \beta_0 + \beta_1 x + u$, where $y = PV$, $x = GPI$.

(A) Test $H_0: \beta_1 = 0$ against $H_1: \beta_1 \neq 0$ for $\alpha = 0.1$; test statistic (under $H_0$) $(\hat{\beta}_1 - 0)/se(\hat{\beta}_1)$ with null distribution $T(n-2)$; sample value of the test statistic 2.8405; acceptance region ($-1.8946, 1.8946$).

(B) Pivot for $\mu_0 = E(y|x = x_0) = \beta_0 + \beta_1 x_0$ where $x_0 = 1$: $(\hat{\mu}_0 - \mu_0)/se(\hat{\mu}_0)$ with distribution $T(n-2)$; confidence interval for $\mu_0$ at $\alpha = 0.05$: [25.4805, 45.7956] (tabulated value = 2.3646).

(C) $R^2 = dev(REG)/dev(y) = 0.5355$: the model explains 53.55% of the variability of $y$.

(D) $\hat{\gamma}_{Vietnam} = y_{Vietnam} - \hat{\gamma}_{Vietnam} = 34.4 - 27.3191 = 7.0809$, where $\hat{\gamma}_{Vietnam} = \hat{\beta}_0 + \hat{\beta}_1 x_{Vietnam}$.

(E) Test $H_0: u \sim N$ against $H_0: u \sim N$: test statistics $(\hat{\beta}_1^2 + (\hat{\mu}_u - 3)/2)n/6$ with null distribution $\chi^2(2)$; sample value of the test statistics: 1.9075; rejection region: (4.6052, $\infty$).

Useful formulas and values:

$n = 9$, $df = n - 2 = 7$, $\bar{y} = 41.3667$, $dev(x) = 1.6839$, $dev(y) = 1950.9$, $cod(x, y) = 41.941$, $\hat{\beta}_1 = 24.9071$, $\hat{\beta}_0 = 10.731$, $dev(REG) = 1044.627$, $dev(RES) = 906.273$, $\hat{\sigma}^2 = 129.4676$, $se(\hat{\beta}_1) = \sqrt{76.8855} = 8.7684$, $\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = 35.638$, $se(\hat{\mu}_0) = \sqrt{\hat{\sigma}^2(1/n + (\bar{x} - x_0)^2)/dev(x)} = \sqrt{18.4525} = 4.2956$.

Exercise 2. Assumptions: $X_1$ is ‘Subject to violence in 2011?‘ $\sim Be(p_1)$, $X_2$ is ‘Subject to violence in 2006?’ $\sim Be(p_2)$; independent samples.

(A) Independence implies that the joint likelihood is the product of the likelihoods of the two samples; this, in turn, implies that the joint log-likelihood is the sum of the two log-likelihoods: $l(p_1, p_2) = l_1(p_1) + l_2(p_2) = n_1[\overline{X}_1 \ln p_1 + (1 - \overline{X}_1) \ln q_1] + n_2[\overline{X}_2 \ln p_2 + (1 - \overline{X}_2) \ln q_2]$. Joint maximum: $l(\hat{p}_1, \hat{p}_2) = n_1[\overline{X}_1 \ln \hat{p}_1 + (1 - \overline{X}_1) \ln(1 - \hat{p}_1)] + n_2[\overline{X}_2 \ln \hat{p}_2 + (1 - \overline{X}_2) \ln(1 - \hat{p}_2)] = -153.0724 + -129.4072 = -282.4796$.

(B) Test $H_0: p_1 - p_2 = 0$ against $H_1: p_1 - p_2 < 0$ for $\alpha = 0.1$; test statistic (under $H_0$) $(\overline{X}_1 - \overline{X}_2)/se(\overline{X}_1 - \overline{X}_2)$ with approximated null distribution $N(0, 1)$; sample value of the test statistic $-1.8498$; rejection region: ($-\infty, z_{crit} = -1.2816$).
(C) $\beta$ for $H_0 : p_1 - p_2 = 0$ against $H_1 : p_1 - p_2 = -0.113$. Taking $c = -0.113$, $\beta = P(sample \in A|H_1) = P((X_1 - X_2)/se \geq z_{crit}|H_1) = P((X_1 - X_2)/se \geq z_{crit}) = P((X_1 - X_2 - c)/se \geq (z_{crit} \times se - c)/se|H_1) = P(Z \geq z_{crit} - c/|se|H_1) = P(Z \geq 1.041) = 0.1489$. 

(D) Same test as in (B) for the same $\alpha$; test statistic $2(l - \hat{l}(H_0))$ with approximated null distribution $\chi^2(1)$; sample value of the test statistic: $2(max(l) - max_{H_0}(l)) = 3.3917$; rejection region $(2.7055, \infty)$. Using the notation at point (A): max$(l) = l(\tilde{p}_1 = \tilde{p}_1, \tilde{p}_2 = \tilde{p}_2) = -282.4796$, max$_{H_0}(l) = l(\tilde{p}_1 = \tilde{p}_1, \tilde{p}_2 = \tilde{p}_2) = -153.8568 + -130.3186 = -284.1755$.

Useful formulas and values:

\[ n_1 = 221, n_2 = 193, \tilde{p}_1 = \tilde{p}_1 = 0.516, \tilde{p}_2 = \tilde{p}_2 = 0.606, se(\bar{X}_2 - \bar{X}_1) = \sqrt{p_1q_1/n_1 + p_2q_2/n_2} = \sqrt{0.0023672} = 0.04865. \]

Exercise 3. Assumptions: $X_1$ = ‘Hourly wage of a woman suffering violence’ $\sim N(\mu_1, \sigma_1^2)$, $X_2$ = ‘Hourly wage of a woman not suffering violence’ $\sim N(\mu_2, \sigma_2^2)$; independent samples.

(A) Let us assume $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Pivot for $\mu_1 - \mu_2$: $[(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)]/se$ with distribution $T(n_1 + n_2 - 2)$, where $se = \sqrt{S_p^2(1/n_1 + 1/n_2)}$ and $S_p^2$ is the pooled variance; corresponding confidence interval at $\alpha = 0.01$ for $\mu_1 - \mu_2$: $[-6.2626, -2.2574]$ (tabulated value = 2.7045).

(B) Test $H_0 : \sigma_1^2/\sigma_2^2 = 1$ against $H_0 : \sigma_1^2/\sigma_2^2 \neq 1$ for $\alpha = 0.05$; test statistic (under $H_0$) $S_1^2/S_2^2$ with null distribution $F(df_1 = 20, df_2 = 20)$; sample value of the test statistic 0.7847; acceptance region: $[0.4058, 2.4645]$.

(C) We re-express the assumptions: $X/V \sim N(\mu_1 = 9.74, \sigma_1^2 = 5.0625)$, $X/V \sim N(\mu_2 = 14, \sigma_2^2 = 6.4516)$. We set $c = 8.37$. Then $P(X < c) = P(X < c|V)P(V) + P(X < c|V)P(V) = 0.0234$.

(D) Then $P(V|X < c) = P(X < c|V)P(V)/P(X < c) = 0.4531$.

Useful formulas and values:

\[ n_1 = 21, n_2 = 21, df_1 = 20, df_2 = 20, \bar{X}_1 = 9.74, \bar{X}_2 = 14, s_1^2 = 2.25^2 = 5.0625, s_2^2 = 2.54^2 = 6.4516, s_p^2 = (df_1 s_1^2 + df_2 s_2^2)/(df_1 + df_2) = 5.7571, se(\bar{X}_1 - \bar{X}_2) = \sqrt{s_p^2(1/n_1 + 1/n_2)} = \sqrt{0.5483} = 0.7405. \]

\[ P(X < c|V) = P((X - \mu_1)/\sigma_1 < (c - \mu_1)/\sigma_1|V) = P(Z < -0.6089) = 0.2713, P(X < c|V) = P((X - \mu_2)/\sigma_2 < (c - \mu_2)/\sigma_2|V) = P(Z < -2.2165) = 0.0133, P(V) = 0.0391, P(V) = 0.9609. \]
Text

Framework: Global warming and \(CO_2\) emissions.

Exercise 1. (Only students with 9 CFU)

It is well known that developed countries contribute more to the global warming. The question to be answered is: “can we estimate a relationship between \(CO_2\) emissions (taken as dependent variable) and \(GDP\)?”. The following table reports some statistics computed from World Bank data\(^1\) in 2013 (\(C = CO_2\) emissions per capita (metric tons); \(G = GDP\) per capita (constant 2010 US$)).

<table>
<thead>
<tr>
<th>(n)</th>
<th>(\sum_{i=1}^{n} \ln(G_i))</th>
<th>(\sum_{i=1}^{n} \ln(G_i)^2)</th>
<th>(\sum_{i=1}^{n} \ln(C_i))</th>
<th>(\sum_{i=1}^{n} \ln(C_i)^2)</th>
<th>(\sum_{i=1}^{n} \ln(C_i) \times \ln(G_i))</th>
</tr>
</thead>
<tbody>
<tr>
<td>140</td>
<td>1197.53</td>
<td>10553.08</td>
<td>95.2</td>
<td>388.87</td>
<td>1092.54</td>
</tr>
</tbody>
</table>

Formulate an appropriate simple linear regression model between the log variables and answer the following questions (consider the sample size large enough to use asymptotic properties).

(A) Compute the confidence interval at 90\% for each of the two regression parameters.

(B) Test whether the slope coefficient is significantly different from 1 using the p-value. Specify the hypotheses, the test statistic and its distribution under the null.

(C) Compute how much of the variability of the dependent variable is explained by the model.

(D) Observations giving absolute standardized residuals larger than 2.5 are identified as outliers. Is \((C_{LUX} = 18.7, G_{LUX} = 102242)\) an outlier? Motivate the answer.

Exercise 2.

We aim at checking whether world countries are really committed to reduce \(CO_2\) emissions. The following table reports some statistics computed on the same sample of 140 countries (the variable is \(CO_2\) emissions per capita expressed in metric tons; the standard deviations are computed using degrees of freedom at denominator).

<table>
<thead>
<tr>
<th>Year</th>
<th>average</th>
<th>standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>4.67</td>
<td>5.44</td>
</tr>
<tr>
<td>2013</td>
<td>4.45</td>
<td>5.05</td>
</tr>
<tr>
<td>Difference 2013 − 2011</td>
<td>−0.22</td>
<td>1.42</td>
</tr>
</tbody>
</table>

We also know that, during the same period, 68 countries reduced their emissions.

(A) Test whether the mean level has been significantly reduced in the period by using the p-value (consider the sample size large enough to use asymptotic properties).

(B) “Less than 58\% of countries reduced their emissions”. Test this assertion for \(\alpha = 0.05\).

(C) Compute the type 2 error probability of the test at point (B) when the alternative is “49\% of countries reduced emissions”.

(D) (Only students attending since 2015-2016) Do the test at point (B) (at the same significance level) by applying the Likelihood Ratio test (when \(X \sim Be(p)\), the likelihood function is \(L(p) = p^n\bar{x}(1 − p)^n(1−\bar{x})\), where \(\bar{x}\) is the sample mean).

Exercise 3. (Only students with 6 CFU)

To get an idea, we list some typical \(CO_2\) concentrations in different environments:

- 250 – 350 ppm in outdoor ambients;
- 350 – 1000 ppm in occupied indoor spaces with good air exchange;
- 1000 – 2000 ppm give complaints of drowsiness and poor air;

\(^1\)http://databank.worldbank.org/data
• 2000 – 5000 ppm causes headaches, sleepiness and stuffy air;
• 5000 ppm is the workplace exposure limit in most US jurisdictions.

The Police installed a CO₂ detector in a typically crowded room of the headquarter, resulting in the following sample of daily measures (the measures are expressed in natural logarithm since this transformation is more in line with the Normal distribution).

\[ \ln(\text{ppm}) \begin{array}{cccccc} 8.11 & 8.23 & 7.94 & 8.68 & 8.46 & 7.94 \end{array} \]

Assume that \( \ln(\text{ppm}) \) (the CO₂ concentration in natural logarithm) is Normally distributed and answer the following questions.

(A) (Only students attending before 2015-2016) Estimate the parameters with good unbiased estimators and provide the corresponding standard errors.

(B) Compute the value of the log-likelihood at the estimated values of the parameters. Is the value computed and provide the corresponding standard errors.

(C) Test against \( 0 \): if \( \text{CO}_2 \) concentration in natural logarithm) is Normally distributed and answer the following questions.

Exercise 1.

Exercise assumptions: \( y = \beta_0 + \beta_1 x + u \), where \( y = \ln C, x = \ln G \).

(A) Pivot for \( \beta_1 \): \((\hat{\beta}_1 - \beta_1)/se(\hat{\beta}_1)\) with approximate distribution \( N(0,1) \); confidence interval for \( \alpha = 0.1: [0.8299, 0.967] \).

(B) Test \( H_0: \beta_1 = 0 \) vs \( H_1: \beta_1 \neq 0 \); test statistics \((\hat{\beta}_1 - 1)/se(\hat{\beta}_1)\) with null distribution approximately \( N(0,1) \); sample value of the test statistics: \(-2.43751; \ p - value = 2P(Z > | - 2.43751|) = 2P(Z > 2.43751) = 2 \times 0.00739 = 0.01479 \).

(C) \( R^2 = \text{dev}(\text{REG})/\text{dev}(y) = 0.7712 \): the model explains 77.12% of the variability of \( y \).

(D) \( x_{\text{LUX}} = \ln(102242) = 11.535, y_{\text{LUX}} = \ln(18.7) = 2.929 \). Then \( \hat{y}_{\text{LUX}} = y_{\text{LUX}} - \hat{\beta}_1x_{\text{LUX}} \); standardized residual: |\( \hat{u}_{\text{LUX}}/\hat{\alpha} \)| = | - 0.5866 | 0.5866 \( \Rightarrow \) not outlier.

Useful formulas and values:
\( n = 140, \ df = n - 2 = 138, \ \bar{x} = 8.5538, \ \bar{y} = 0.68, \ \text{dev}(x) = 309.665, \ \text{dev}(y) = 324.134, \ \text{cod}(\bar{x}, \bar{y}) = 278.2196, \ \hat{\beta}_1 = 0.8985, \ \hat{\beta}_0 = -7.0052, \ \text{dev}(\text{REG}) = 249.9674, \ \text{dev}(\text{RES}) = 74.1666, \ \hat{\sigma}^2 = 0.5374, \ se(\hat{\beta}_1) = \sqrt{0.0017} = 0.0417, \ se(\hat{\beta}_0) = \sqrt{0.1308} = 0.3617 \).

Exercise 2.

(A) Assumptions: \( D = X_2 - X_1 \sim N(\mu_D, \sigma_D^2) \), where \( X_1 = \text{CO}_2 \text{ emissions in 2011}' \), \( X_2 = \text{CO}_2 \text{ emissions in 2013}' \) (paired samples). Test \( H_0: \mu_D = 0 \) vs \( H_1: \mu_D < 0 \); test statistic (under \( H_0 \)) \((\bar{D} - 0)/se(\bar{D})\) with approximated null distribution \( N(0,1) \); sample value of the test statistics: \(-1.8332; \ p - value = P(Z < -1.8332) = 0.0334 \).

(B) Test \( H_0: p = 0.58 \) against \( H_1: p < 0.58 \) for \( \alpha = 0.05 \); test statistic (under \( H_0 \)) \((\bar{X} - 0.58)/se_0 \) with approximated null distribution \( N(0,1) \); sample value of the test statistic \(-2.2603; \ rejection \text{ region: } (\infty, z_{crit} = -1.6449) \).

(C) \( \beta \) for \( H_0: p = 0.58 \) against \( H_1: p = 0.49 \). Take \( p_0 = 0.58 \) and \( p_1 = 0.49 \). \( \beta = P(\text{sample} \in A|H_1) = P((\bar{X} - p_0)/se_0 \geq z_{crit}|H_1) = P(\bar{X} \geq p_0 + z_{crit} \times se_0|H_1) = P((\bar{X} - p_1)/se_1 \geq (z_{crit} \times se_0 + p_0 - p_1)/se_1|H_1) = P(Z \geq (z_{crit} \times se_0 + p_0 - p_1)/se_1|H_1) = P(Z \geq 0.5062) = 0.3063 \).
(D) Same test as in (B) for the same $\alpha$; test statistic $2(\hat{l} - \hat{l}(H_0))$ with approximated null distribution $\chi^2(1)$; sample value of the test statistic: $2(\text{max}(l) - \text{max}(H_0(l))) = 5.0361$; rejection region $(3.8415, \infty)$. Using the notation at point (B): \( l(p) = n(\pi \ln \pi + (1-\pi) \ln (1-p)) \), \( \text{max}(l) = l(\pi) = n(\pi \ln \pi + (1-\pi) \ln (1-\pi)) = -96.9835 \), \( \text{max}\text{log}(H_0(l) = l(p_0) = n(\pi \ln p_0 + (1-\pi) \ln (1-p_0)) = -99.5015 \).

Useful formulas and values:

- $n = 140, \bar{d} = -0.22, s_D = 1.42, se(\bar{D}) = \sqrt{s_D^2/n} = \sqrt{0.0144} = 0.12,$
- $p_0 = 0.58, p_1 = 0.49, se_0 = \sqrt{p_0 q_0/n} = \sqrt{0.00174} = 0.04171, se_1 = \sqrt{p_1 q_1/n} = \sqrt{0.00178} = 0.04225.$

**Exercise 3.** Assumptions: $X = \ln(\text{ppm}) \sim N(\mu, \sigma^2)$.

<table>
<thead>
<tr>
<th>$i$</th>
<th>$x_i$</th>
<th>$x_i^2$</th>
<th>$x_i^2$</th>
<th>$x_i^2$</th>
<th>$x_i^2$</th>
<th>$x_i^2$</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8.11</td>
<td>65.7721</td>
<td>67.7329</td>
<td>63.0436</td>
<td>75.3424</td>
<td>71.5716</td>
<td>406.5062</td>
</tr>
<tr>
<td>2</td>
<td>8.23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>7.94</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>8.68</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>8.46</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>7.94</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(A) Unbiased estimator of $\mu$: $\overline{X}$; unbiased estimator of $\sigma^2$: $S^2$; estimate of $\mu$: $\pi = 8.22667$ with standard error $se(\overline{X}) = \sqrt{s^2/n} = \sqrt{0.0146} = 0.12082$; estimate of $\sigma^2$: $s^2 = 0.08759$ with standard error $se(S^2) = \sqrt{2s^4/(n-1)} = \sqrt{0.00307} = 0.05539$.

(B) $l(\mu, \sigma^2) = -0.5 \left[ n \ln(2\pi\sigma^2) + \sigma^{-2} \sum_{i=1}^{n} (x_i - \mu)^2 \right]$. Log-likelihood at the estimated values of parameters:

\[
l(\pi, s^2) = -0.5 \left[ n \ln(2\pi s^2) + s^{-2} \sum_{i=1}^{n} (x_i - \pi)^2 \right] = -0.5 \left[ n \ln(2\pi s^2) + s^{-2}(n-1)s^2 \right] = -0.5 \left[ n \ln(2\pi s^2) + n - 1 \right] = -0.70825.
\]

The value computed is not the maximum possible value of the log-likelihood, since the ML estimator of $\sigma^2$ is $s^2$; not $s^2$. $l(\pi, s^2) = -0.5 \left[ n \ln(2\pi s^2) + s^{-2} \sum_{i=1}^{n} (x_i - \pi)^2 \right] = -0.5 \left[ n \ln(2\pi s^2) + s^{-2}ns^2 \right] = -0.5 \left[ n \ln(2\pi s^2) + n \right] = -0.66129.$

(C) Assumptions: $X = \ln(\text{ppm}) \sim N(\mu = 8.22667, \sigma^2 = 0.08759)$. $P(ppm > 5000) = P(\ln(ppm) > \ln 5000) = P(X > \ln 5000) = P((X - \mu)/\sigma > (\ln 5000 - 8.22667)/0.29595) = P(Z > 0.98167) = 0.16313.$

(D) $X_i$ ‘ppm at day $i > 5000?’ \sim Be(p = 0.16313).$ Then $Y$ ‘number of days with $ppm > 5000$ in 1 year’ \sim $Bi(n = 21, p = 0.16313).$ $P(Y \geq 2) = 1 - P(Y \leq 2) = 1 - P(Y = 0) - P(Y = 1) = 1 - 0.02376 - 0.09725 = 0.87899.$

Useful formulas and values:

\[n = 6, df = 5, \bar{x} = 8.22667, dev(\bar{x}) = 0.43793, s^2 = 0.08759, s^2 = 0.07299.\]
Exercise 1.

The following table reports sample statistics concerning the variable $X = \text{daily number of frauds}$ (assume the sample a simple random sample).

<table>
<thead>
<tr>
<th></th>
<th>$\sum_{i=1}^{n} x_i$</th>
<th>$\sum_{i=1}^{n} x_i^2$</th>
<th>$\sum_{i=1}^{n} x_i^3$</th>
<th>$\sum_{i=1}^{n} \ln(x_i!)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>119</td>
<td>237</td>
<td>717</td>
<td>2619</td>
<td>130.99</td>
</tr>
</tbody>
</table>

Assuming that $X$ follows a Poisson distribution, answer the following questions.

(A) Estimate the parameter of the model by maximum likelihood; provide the corresponding standard error.

(B) Compute the value of the log-likelihood when the parameter is $2.3$.

(C) Test whether the parameter of the model is significantly different from the value at point (B) using the likelihood ratio statistic ($\alpha = 0.01$).

(D) If the test of the hypothesis at point (C) is done using the usual score statistic $(X - \hat{\lambda})/\sqrt{\hat{\lambda}/n}$, where $\hat{\lambda}$ is the value of the parameter under the null, the rejection region is $(-\infty, -2.576) \cup (2.576, \infty)$. Compute the power of this test when the alternative is $\lambda = 2.7$.

(E) Assuming that the parameter is equal to the estimated value, compute the probability that, in two days, less than 2 frauds are discovered.

(Hints: the contribution to the log-likelihood of the $i$-th sample observation is $x_i \ln(\lambda) - \lambda - \ln(x_i!)$).

Exercise 2. (Only students with 9 CFU)

When a firm asks for a loan, it must motivate its request. To contrast excessive demands, the institution wants to use a statistical model to estimate a fair level of the amount to finance, where fair is intended in comparison to the motivations received. The model (for the moment in a prototype stage) is a simple linear model where the dependent variable is the amount requested in natural logarithm (the independent variable is not disclosed for confidentiality reasons).

The following table reports the values of the dependent variable and the corresponding fitted values,

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_i$</td>
<td>1071</td>
<td>1809</td>
<td>1972</td>
<td>1298</td>
<td>889</td>
</tr>
<tr>
<td>$\hat{y}_i$</td>
<td>1162</td>
<td>1934</td>
<td>1669</td>
<td>1439</td>
<td>...</td>
</tr>
</tbody>
</table>

while the estimated parameters resulted $\hat{\beta}_0 = 493.8$ (s.e. = 248.4) and $\hat{\beta}_1 = 943.9$ (s.e. = 236.5).

(A) Compute the missing fitted value and the model residuals.

(B) Compute the confidence interval at $\alpha = 0.01$ for the standard deviation of the error component.

(C) Compute the percentage of deviance explained by the model.

Exercise 3. (Only students with 6 CFU)

Considering defaulted loans, the financial institution is comparing the loss ratio (loss/loan) of two ethnic groups. The following table show some statistics of two independent random samples (s.d. indicates the square root of the unbiased sample variance).

<table>
<thead>
<tr>
<th>Ethnic group</th>
<th>size</th>
<th>mean</th>
<th>s.d.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>25</td>
<td>0.21</td>
<td>0.07</td>
</tr>
<tr>
<td>B</td>
<td>16</td>
<td>0.27</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Assuming that the loss ratio follows a Normal distribution answer the following questions.

(A) Test whether the two groups have a significantly different standard deviation ($\alpha = 0.05$).
(B) Compute the confidence interval at $\alpha = 0.01$ for the difference between the two means. Take care to use a test procedure consistent with the outcome of point (A).

(C) Image to select randomly one observation from the two samples but you ignore the group whether it belongs. Assuming a Normal distribution with parameters equal to the estimated values, compute the probability that such an observation has loss ratio greater than 0.32.

Solution

**Exercise 1.** Assumptions: $X$ = ‘number of frauds discovered in a day’ Po($\lambda$).

(A) The Maximum Likelihood (ML) estimator of $\lambda$ is the sample mean $\bar{X}$ and has variance $V(\bar{X}) = \lambda/n$. Thus, $\hat{\lambda} = \bar{X} = 237/119 = 1.9916$, s.e.(X) = $\sqrt{\lambda/n} = \sqrt{0.0167} = 0.1294$.

(B) Log-likelihood: $l(\lambda) = \sum_{i=1}^{n} x_i \ln \lambda - n\lambda - \sum_{i=1}^{n} \ln(x_i!) = -207.2905$, where $\lambda = 2.3$, $\sum_{i=1}^{n} x_i = 237$, $n = 119$,

$$n \sum_{i=1}^{n} \ln(x_i!) = 130.99.$$  

(C) Test $H_0 : \lambda = 2.3$ against $H_1 : \lambda \neq 2.3$; test statistic, $2 [l(\hat{\lambda}) - l(\lambda_0)] = 2(-204.712 - (-207.2905)) = 5.157$ where $\lambda_0 = 2.3$; rejection region, $R = (6.6349, \infty)$.

(D) Power of the test $H_0 : \lambda = 2.3$ against $H_1 : \lambda = 2.7$ under the alternative, using the score statistic: $\gamma = P(\text{sample} \in R|H_1) = P(\{(X - \lambda_0)/se_0 < -z|H_1\} + P(\{(X - \lambda_0)/se_0 > z|H_1\} = P(X < -z \cdot se_0 + \lambda_0 | H_1) + P(X > z \cdot se_0 + \lambda_0 | H_1) = P((X - \lambda_1)/se_1 < (-z \cdot se_0 + \lambda_0 - \lambda_1)/se_1 | H_1) + P((X - \lambda_1)/se_1 > (z \cdot se_0 + \lambda_0 - \lambda_1)/se_1 | H_1) = P(Z < -0.0329|H_1) + P(Z > 0.2781|H_1) = 1/2 + 0.0696 = 0.6096.

(E) Exploiting additivity of the Poisson we have $X = X_1 + X_2 \sim Po(2 \lambda = 3.9832)$. Then $P(X < 2) = P(X \leq 1) = P(X = 0) + P(X = 1) = 0.0186 + 0.0742 = 0.0928$.

Useful formulas and values:

$$se_0 = \sqrt{\lambda_0/n} = 0.13902, \quad se_1 = \sqrt{\lambda_1/n} = 0.15063.$$  

**Exercise 2.** Assumptions: $y = \beta_0 + \beta_1 x + u$, where $y = \ln(\text{amount})$ and $x = \cdot$.

(A) From the properties of OLS estimators of linear regression parameters we know that $\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} y_i$, or, equivalently, $\sum_{i=1}^{n} \hat{u}_i = 0$. The missing fitted value can be computed using that property.

<table>
<thead>
<tr>
<th>$\hat{y}_i$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1071</td>
<td>1809</td>
<td>1972</td>
<td>1298</td>
<td>889</td>
<td>7039</td>
</tr>
</tbody>
</table>

| $\hat{u}_i = y_i - \hat{y}_i$ | 1162| 1934| 1669| 1439| 835 | 7039 |

| $\hat{u}_i^2$ | 8281| 15625| 91809| 19881| 2916| 138512 |

| $\hat{y}_i^2$ | 1147041| 3272481| 3888784| 1684804| 790321| 10783431 |

(B) Pivot for $\sigma^2$: $\hat{\sigma}^2(n-2)/\sigma^2$ with distribution $\chi^2(n-2)$; interval at $\alpha = 0.01$ for $\sigma^2$: $[10789.088, 1031240.5584]$; corresponding interval for $\sigma$: $[103.8705, 1389.6908]$.

(C) $R^2 = dev(REG)/dev(y) = 0.8415$: the model explains 84.1506% of the variability of $y$.

Useful formulas and values:

$$n = 5, \quad df = n - 2 = 3, \quad dev(RES) = \sum_{i=1}^{n} \hat{u}_i^2 = 138512, \quad \hat{\sigma}^2 = dev(RES)/(n-2) = 46170.667, \quad \bar{y} = 1407.8, \quad dev(y) = 873926.80, \quad dev(REG) = dev(y) - dev(RES) = 735414.80.$$  

**Exercise 3.** Assumptions: $X_1$ = ‘loss ratio in A’ $\sim N(\mu_1, \sigma^2_1)$, $X_2$ = ‘loss ratio in B’ $\sim N(\mu_2, \sigma^2_2)$, independent samples.

(A) Test $H_0 : \sigma_1^2/\sigma_2^2 = 1$ vs $H_1 : \sigma_1^2/\sigma_2^2 \neq 1$; test statistic (under $H_0$) $S_1^2/S_2^2$ with null distribution $F(n_1 - 1, n_2 - 1)$; sample value of the test statistics: 0.76563; acceptance region for $\alpha = 0.05$: $(0.41027, 2.70064)$.  

76
(B) Since the test at point (A) leads to accept $H_0$, we compute the interval at $\alpha = 0.01$ for $\mu_1 - \mu_2$ using the pivot $[(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)]/\sqrt{S_p^2(1/n_1 + 1/n_2)}$, with distribution $T(n_1 + n_2 - 2)$; confidence interval: $[-0.12416, 0.00416]$.  

(C) $P(X > c) = P(X > c|A)P(A) + P(X > c|B)P(B) = 0.13919$, where $P(X > c|A) = P((X - \mu_1)/\sigma_1 > (c - \mu_1)/\sigma_1|A) = P(Z > 1.57143) = 0.05804$, $P(X > c|B) = P((X - \mu_2)/\sigma_2 > (c - \mu_2)/\sigma_2|B) = P(Z > 0.625) = 0.26599$, where $c = 0.32$, $\mu_1 = 0.21$, $\sigma_1 = 0.07$, $\mu_2 = 0.27$, $\sigma_2 = 0.08$.  

Useful formulas and values:

$n_1 = 25$, $n_2 = 16$, $\bar{x}_1 = 0.21$, $\bar{x}_2 = 0.27$, $s_1^2 = 0.0049$, $s_2^2 = 0.0064$, $s_p^2 = [s_1^2(n_1 - 1) + s_2^2(n_2 - 1)]/(n_1 + n_2 - 2) = 0.00548$, $\sqrt{S_p^2(1/n_1 + 1/n_2)} = \sqrt{0.00056} = 0.02369$. 

77
Text

Framework: Health perception in OECD countries is surveyed according to three categories: 1) Bad/very bad, 2) Fair (not good, not bad), 3) Good/very good. In what follows such categories are labeled as bad, fair and good, respectively.

Exercise 1. (Only students with 9 CFU)

We aim at evaluating whether health perception is related to the economic condition. The following table reports some statistics on the percentage of adults believing itself in bad health and the per capita GDP in US$ (both variables are taken in natural logarithm) computed in 2014 on 28 countries.

\[
\begin{align*}
\sum_{i=1}^{n} \ln PB_i & \quad \sum_{i=1}^{n} (\ln PB_i)^2 & \quad \sum_{i=1}^{n} \ln GDP_i & \quad \sum_{i=1}^{n} (\ln GDP_i)^2 & \quad \sum_{i=1}^{n} \ln PB_i \times \ln GDP_i \\
62.774 & \quad 146.375 & \quad 295.354 & \quad 3118.752 & \quad 659.113
\end{align*}
\]

Considering a simple linear regression model where \( \ln PB \) is the dependent variable, answer the following questions.

(A) Test whether the independent variable is significant in the regression (\( \alpha = 0.001 \)). Specify the null and the alternative hypotheses, the test statistics and its distribution under the null.

(B) Compute how much of the variability of the dependent variable is explained by the model.

(C) (Only students attending since 2015-2016) Test whether the Normality assumption on the error term is justified (\( \alpha = 0.01 \)) considering the following sample statistics on residuals \( \hat{u}_i : n^{-1} \sum_{i=1}^{n} \hat{u}_i / \hat{\sigma} = 0, n^{-1} \sum_{i=1}^{n} \hat{u}_i^2 / \hat{\sigma}^2 = 0.929, n^{-1} \sum_{i=1}^{n} \hat{u}_i / \hat{\sigma}^3 = 0.031, n^{-1} \sum_{i=1}^{n} \hat{u}_i^2 / \hat{\sigma}^4 = 0.929 \). In addition, specify the hypotheses, the test statistic and its distribution under the null.

(D) (Only students attending before 2015-2016) Compute the regression residual for the observation \( (PB_{Iceland} = 6.3, GDP_{Iceland} = 44290) \).

(E) Compute the prediction interval for the conditional mean of the dependent variable for a country having 56980 per capita GDP (\( \alpha = 0.01 \)).

Exercise 2.

The following table reports some sample statistics on the health perception in Italy during 2014.

<table>
<thead>
<tr>
<th></th>
<th>size</th>
<th>percentage in bad status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>1158</td>
<td>10.3</td>
</tr>
<tr>
<td>Female</td>
<td>964</td>
<td>13.7</td>
</tr>
</tbody>
</table>

Assuming the two simple random samples are independent answer the following questions.

(A) Compute the maximum value of the joint log-likelihood for this sample.

(B) Is there some significant difference in health perception between the two genders? Test by comparing the p-value of the usual Wald statistic with typical values of the significance level. Specify the null and the alternative hypotheses, the test statistics and its distribution under the null.

(C) Do the same test at point (B) but using the likelihood ratio statistic and \( \alpha = 0.05 \). Under the null, estimate the common probability using a pooled estimator.

(D) If the test of the hypothesis at point (B) is done using the usual Wald statistic, the rejection region at some significant level is \( (-\infty, -1.96) \cup (1.96, \infty) \). Compute the power of this test when the alternative is \( H_1 = p_M - p_F = 0.038 \) (\( p_M \) and \( p_F \) are the probabilities of bad health perception for the two genders).

(E) (Only students with 6 CFU) Consider a population composed by 35.8% males and 64.2% females with percentages of people in bad health identical to above table. Drawn randomly two observations from one of the two genders (unfortunately, we ignore the group whether they belong), none of them resulted in a bad health status. Compute the probability of such an event (namely, none between two extracted is in bad health status).

(F) (Only students with 6 CFU) Given the information that none of two extracted is in a bad health status, compute the probability that the two extracted are males.
Hint: The log-likelihood for a random variable $X \sim Be(p)$ is $l(p) = n(\ln p + (1 - \ln(1 - p)))$, where $n$ is the sample size.

**Exercise 3. (Only students with 6 CFU)**

A parameter $\theta > 0$ can be estimated using two different estimators, say $T_1$ and $T_2$. The former has mean $\theta n/(n + 1)$ and standard deviation $\theta/\sqrt{n} + 1$; the latter has mean $\theta$ and standard deviation $\theta/\sqrt{n}$.

(A) Which one is more efficient when the sample size is $n = 17$? Motivate the answer.

(B) Are $T_1$ and $T_2$ consistent? Which one is asymptotically more efficient? Motivate the answers.

**Solution**

**Exercise 1.**

Assumptions: $y = \beta_0 + \beta_1 x + u$, where $y = \ln PB$ and $x = \ln GDP$.

(A) Test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$; test statistic, $(\hat{\beta}_1 - 0)/\hat{\sigma} \cdot \hat{\beta}_1$ with null distribution $T(n - 2)$; sample value of the test statistic, $-5.1704$; acceptation region for $\alpha = 0.001$, $A = (-3.7066, 3.7066)$.

(B) $R^2 = dev(\text{REG})/\text{dev}(y) = 0.5069$: the model explains 50.69% of the variability of $y$.

(C) Test $H_0 : u \sim N$ against $H_0 : u \sim N$; test statistics $(\hat{\beta}^2 + (\hat{K}u - 3)^2/4)n/6$ with null distribution $\chi^2(2)$; sample value of the test statistics: 5.0084; rejection region for $\alpha = 0.01$: (9.2103, $\infty$).

(D) $(x)_{\text{Iceland}} = \ln(44290) = 10.6985, y_{\text{Iceland}} = \ln(6.3) = 1.8405$. Then $\hat{y}_{\text{Iceland}} = y_{\text{Iceland}} - \hat{y}_{\text{Iceland}} = 1.8405 - 2.1011 = -0.2606$, where $\hat{y}_{\text{Iceland}} = \hat{\beta}_0 + \hat{\beta}_x x_{\text{Iceland}}$.

(E) Confidence interval for $\mu_0 = E(y|x_0) = \hat{\beta}_0 + \hat{\beta}_x x_0$ where $x_0 = \ln(56980) = 10.9505$. The point estimator is $\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_x x_0$; the corresponding estimate is $\hat{\mu}_0 = 1.8649$. The pivot is $(\hat{\mu}_0 - \mu_0)/\hat{\sigma}(\hat{\mu}_0)$ with distribution $T(n - 2)$; the corresponding interval at $\alpha = 0.01$ is [1.5993, 2.1305], where $t = 2.7787$.

Useful formulas and values:

$n = 28$, $df = n - 2 = 26$, $\tau = 10.54836$, $\bar{y} = 2.24193$, $\text{dev}(x) = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 3.25252$, $\text{dev}(y) = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 5.64018$ codev($\bar{x}, \bar{y}$) = $\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = 3.04957$, $\hat{\beta}_1 = -0.9376$, $\hat{\beta}_0 = 12.13208$, $\text{dev}(\text{RES}) = \text{dev}(y) - \hat{\beta}_1^2 \text{dev}(x) = 2.78089$, $\hat{\sigma}^2 = \text{dev}(\text{RES})/df = 0.10696$, $se(\hat{\beta}_1) = \hat{\sigma}/\sqrt{\text{dev}(x)} = 0.18134$, $dev(\text{REG}) = \hat{\beta}_1^2 \text{dev}(x) = 2.85928$, $\hat{\mu}_0 = 1.86492$, $se(\hat{\mu}_0) = \hat{\sigma}/\sqrt{1/n + (x_0 - \bar{x})^2/\text{dev}(x)} = 0.09559$.

**Exercise 2.** Assumptions: $X_1 =$‘Male in bad health status?’, $X_2 =$‘Female in bad health status?’ $\sim Be(p_1)$; independent samples.

(A) Independence implies that the joint likelihood is the product of the likelihoods of the two samples; this, in turn, implies that the joint log-likelihood is the sum of the two log-likelihoods: $l(p_1, p_2) = l_1(p_1) + l_2(p_2) = n_1[\ln p_1 + (1 - \bar{x}_1) \ln q_1] + n_2[\ln p_2 + (1 - \bar{x}_2) \ln q_2]$.

Joint maximum: $l(\hat{p}_1, \hat{p}_2) = n_1[\ln \hat{p}_1 + (1 - \bar{x}_1) \ln(1 - \hat{p}_1)] + n_2[\ln \hat{p}_2 + (1 - \bar{x}_2) \ln(1 - \hat{p}_2)] = -384.0218 + -385.0987 = -769.1206$.

(B) Test $H_0 : p_1 - p_2 = 0$ against $H_1 : p_1 - p_2 \neq 0$; test statistic (under $H_0$) $(\hat{X}_1 - \hat{X}_2)/se(\hat{X}_1 - \hat{X}_2)$ with approximated null distribution $N(0, 1)$; sample value of the test statistic $-2.38968$; p-value = $2P((\hat{X}_1 - \hat{X}_2)/se(\hat{X}_1 - \hat{X}_2) > | -2.38968|/H_0) = 2P(Z > 2.38968|H_0) = 2 \times 0.00843 = 0.01686$.

(C) Same test as in (B) for $\alpha = 0.05$; test statistic $2(\hat{t} - \tilde{t}(H_0))$ with approximated null distribution $\chi^2(1)$; sample value of the test statistic, $5.79771$; rejection region (3.8146, $\infty$). Using the notation at point (A): $\tilde{t} = \max(l) = l(\hat{p}_1 = \bar{x}_1, \hat{p}_2 = \bar{x}_2) = -769.12058$, $\tilde{t}(H_0) = \max_{H_0}(l) = l(\hat{p}_1 = \bar{x}, \hat{p}_2 = \bar{x}) = -385.39814 + -386.6213 = -772.01944$.

(D) $\gamma$ for $H_0 : p_1 - p_2 = 0$ against $H_1 : p_1 - p_2 = 0.038$. Taking $c = 0.038$, $\beta = P(\text{sample} \in A|H_1) = P(\cdot) \leq \gamma \leq z_{\text{crit}} \cdot \sigma \leq \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \leq z_{\text{crit}} \cdot \sigma \leq \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = 3.04957$, $\beta = P(\hat{\sigma}/\sqrt{\text{dev}(x)} = 0.18134$, $P(\hat{\beta}_1^2 \text{dev}(x) = 2.85928$, $\hat{\mu}_0 = 1.86492$, $se(\hat{\mu}_0) = \hat{\sigma}/\sqrt{1/n + (x_0 - \bar{x})^2/\text{dev}(x)} = 0.09559$.

79
(E) Notation: \( G = 'Gender', X = 'How many with a bad health in two randomly drawn individuals?' \). Then \((X|G = M) \sim Bi(p_M = 0.103, n = 2), (X|G = F) \sim Bi(p_F = 0.137, n = 2)\) where \( P(M) = 0.358, P(F) = 0.642 \).

\[
P(X = 0) = P(X = 0|G = M)P(G = M) + P(X = 0|G = F)P(G = F) = 0.80461 \times 0.358 + 0.74477 \times 0.642 = 0.76619.
\]

(F) \( P(G = M|X = 0) = P(X = 0|G = M)P(G = M)/P(X = 0) = 0.80461 \times 0.358/0.76619 = 0.37595. \)

**Exercise 3.**

The efficiency of two estimators can be compared looking at their \( MSE = bias^2 + V \):

\[
MSE(T_1) = bias(T_1)^2 + V(T_1) = \theta^2(1^2/(n+1)^2 + \theta^2/(n+1)) = \theta^2(1^2 + 1 + n)/(n+1)^2, \quad MSE(T_2) = bias(T_2)^2 + V(T_2) = 0 + \theta^2/n = \theta^2/n.
\]

(A) For \( n = 17 \), \( MSE_1 = \theta^2 \times 0.058642, MSE_2 = \theta^2 \times 0.058824 \), hence \( T_1 \) is more efficient because its MSE is smaller.

(B) Both estimators are consistent since their \( MSE \) goes to zero for \( n \to \infty \). They have the same asymptotic efficiency since their \( MSE \) goes to zero (for \( n \to \infty \)) with the same speed. Formally, we can prove the result showing that the ratio between the two MSE goes to 1 for \( n \to \infty \).