Stat inf examen januari 2024

16 augustus 2024

Exercise 1. Consider a random sample $X_1, \ldots, X_n$ from the statistical model $(\mathbb{R}, B, \{P_\sigma; \sigma > 0\})$ where P_σ has density function

$$f_\sigma = \frac{1}{2\sigma} e^{-\frac{|x|}{\sigma}}.$$

For $X \sim P_\sigma$ it holds that $E[X] = 0, E[|X|] = \sigma, Var[X] = 2\sigma^2$ and the characteristic function is given by $\phi_X(t) = \frac{1}{1+\sigma^2 t^2}$.

(a) Calculate the MLE for σ.

(b) Use the properties of the MLE to determine its asymptotic distribution.

(c) Construct an estimator for the variance of X and find its asymptotic distribution.

(d) Construct an approximate $100 \times (1-\alpha)\%$ confidence interval for σ based on its MLE.

(e) Conider the hypothesis test problem

$$H_0 \colon \sigma = \frac{1}{2}$$
$$H_1 \colon \sigma \neq \frac{1}{2}$$

Construct thee Wald and score test for this problem. Are they different or not?

Exercise 2. Three treatments are compared to each other by applying them in random order to each of n independent subjects. This results in a random $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ distributed as $\boldsymbol{X} = (X_1, X_2, X_3)^T$ with X_i the effect of treatment i for $i = 1, 2, 3$. Assume that $\boldsymbol{X} \sim N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ both unknown.

(a) Construct a hypothesis test for the null hypothesis of no difference in effect (i.e. $\mu_1 = \mu_2 = \mu_3$) based on the random sample.

(b) Construct simultaneous confidence intervals for $\mu_1 - \mu_2$ and $\mu_3 - \mu_2$ with joint confidence level $1 - \alpha$ based on the random sample.

(c) Construct a likelihood ratio test for the null hypothesis of independence of the components $X_1, ..., X_3$ (against the general alternative that the components are not all independent) based on the random sample.

Exercise 3. Consider $\boldsymbol{X}_{11}, \ldots, \boldsymbol{X}_{1n_1}$ a random sample from a distribution F_1 with mean $\boldsymbol{\mu}_1$ and covariance matrix $\boldsymbol{\Sigma}_1$ and $\boldsymbol{X}_{21}, \ldots, \boldsymbol{X}_{2n_2}$ a random sample from a distribution F_2 with mean $\boldsymbol{\mu}_2$ and covariance matrix $\boldsymbol{\Sigma}_2$. Both samples are independent of each other.

(a) Show that in this case the between sum of squares matrix can be written as

$$\mathbf{B}_p = \frac{n_1 n_2}{n} (\overline{\boldsymbol{X}}_{1n_1} - \overline{\boldsymbol{X}}_{2n_2})(\overline{\boldsymbol{X}}_{1n_1} - \overline{\boldsymbol{X}}_{2n_2})^T.$$

(b) Let $\mathbf{W}_p$ denote the within sum of squares matrix and $\mathbf{C}_p$ the pooled sample covariance matrix. Show that

$$\max_{\mathbf{a} \neq \mathbf{0}} \frac{\mathbf{a}^T \mathbf{B}_p \mathbf{a}}{\mathbf{a}^T \mathbf{W}_p \mathbf{a}} = \frac{n_1 n_2}{n(n-2)} (\overline{\boldsymbol{X}}_{1n_1} - \overline{\boldsymbol{X}}_{2n_2})^T \mathbf{C}_p^{-1} (\overline{\boldsymbol{X}}_{1n_1} - \overline{\boldsymbol{X}}_{2n_2}).$$

(c) Which vector $\mathbf{a}$ yields the maximum above?

(d) Let $\overline{Y}_{1n_1} = \mathbf{a}^T\overline{\boldsymbol{X}}_{1n_1}$ and $\overline{Y}_{2n_2} = \mathbf{a}^T\overline{\boldsymbol{X}}_{2n_2}$, and consider a new observation $\boldsymbol{X}_0$ and its projection $Y_0 = \mathbf{a}^T\boldsymbol{X}_0$. Show that classifying observation $\boldsymbol{X}_0$ to group 2 if the Euclidean distances satisfy $d^2(Y_0, \overline{Y}_{1n_1}) > d^2(Y_0, \overline{Y}_{2n_2})$ is equivalent to classifying $\boldsymbol{X}_0$ to group 2 if $Y_0 \geq m = \frac{\overline{Y}_{1n_1}+\overline{Y}_{2n_2}}{2}$.

Exercise 4. Data of 20 insurance firms was collected to relate the speed with which particular insurance innovation is adopted by an insurance firm to the size of the firm an d the type of the firm (stock or mutual). The variables in this dataset are:

variable name	description
Months	time elapsed (in months between the introduction of the innovation and the moment that the firm adopted the innovation)
Size	size of the firm, measured by the amount of total assets, in million dollar
Type	type of firm: either stock company or mutual company

The following model was fit in R to understand the relation between the speed with which a particular insurance innovation is adopted by an insurance firm (Months) and the two covariates Size and Type.

```
Call:
lm(formula = Months ~ Size * Type, data = firm)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7440 -1.7406 -0.4557  1.9311  6.3259

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    33.83837    2.44065  13.864  2.47e-10 ***
Size           -0.10153    0.01305  -7.779  7.79e-07 ***
TypeStock       8.13125    3.65405   2.225   0.0408 *
Size:TypeStock -0.00042    0.01833  -0.023   0.9821
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: ? on ? degrees of freedom
Multiple R-squared:  0.8951,    Adjusted R-squared:  ?
F-statistic: ? on ? and ? DF,  p-value: 4.075e-08
```

[Fig 2: tekening Size vs months met regressielijnen]

(a) Give the expression of the regression line for mutual and stock companies respectively.

(b) Formulate the hypothesis test that is considered by the F-statistic on the last line in the output above. Calculate the value of the F-statistic and specify both degrees of freedom.

(c) Calculate the R^2_{adj} for the above model.

(d) As alternative also the following model is considered.

```
Call:
lm(formula = Months ~ Size, data = firm)

Residuals:
     Min       1Q   Median       3Q      Max
-10.4620  -4.7236   0.7912   4.3427   7.9053

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.48211    2.84425  12.827  1.71e-10 ***
Size        -0.09394    0.01426  -6.589  3.45e-06 ***
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Multiple R-squared:  0.7069,    Adjusted R-squared:  0.6906
```

Which model is preferred based on their adjusted coefficient of determination?

(e) We want to perform a hypothesis test to investigate where there is significant difference between the two models. Formulate a hypothesis and a test statistic that can be used and its distribution under the null hypothesis. Calculate the value of the test statistic based on the outputs above. What is your conclusion if p-value=0.00027?

(f) Which assumptions need to hold for the above inforensic result to be valid? Explain for each of your assumptions how they can be verified for the model in (a) based on the plots in figure 2 and formulate your conclusion.