Exam Statistical Models

December 22, 2021

Department of Mathematics, Vrije Universiteit Amsterdam

You may use a simple calculator provided it is not part of a communicating device. Motivate your answers. Write your solutions clearly, using consistent notation.

You can use the following quantiles: $F_{1,54;0.95} = 4.02$, $F_{1,60;0.95} = 4.00$, $F_{2,54;0.95} = 3.17$, $F_{2,60;0.95} = 3.15$, $F_{3,54;0.95} = 2.78$, $F_{3,60;0.95} = 2.76$, $t_{98;0.975} = 1.98$, $F_{1,98;0.95} = 3.94 = t^2_{98,0.975}$, $\chi^2_{1;0.95} = 3.84$, $\chi^2_{2;0.95} = 5.99$, $\chi^2_{3;0.95} = 7.81$. The significance level is always $\alpha = 0.05$.

1. Three competing factories (Factory 1, Factory 2 and Factory 3) produce the same type of steel cable using different manufacturing processes. The raw material used to construct the cable comes from two different locations (Location 1 and Location 2). At each of the three factories, 10 spools of cable made of material coming from Location 1 and 10 spools of cable made of material coming from Location 2 were randomly selected. Each spool was subjected to a strength test that involved increasing the weight on the cable (in increments of one kilogram) until it broke. For each spool, the critical weight (in kg) on the cable at the moment it broke was recorded. The following table shows the average critical weight (in kg) for each factory and for each location:

	Factory 1	Factory 2	Factory 3	row average
Location 1	759.8	786.1	706.8	750.9
Location 2	730.8	785.9	708.9	741.9
column average	745.3	786.0	707.8	746.4

   (i) (5) Give the appropriate two-way ANOVA model that can be applied to investigate the effects of the factors Location and Factory (and their interaction) on the cable strength. Specify the model assumptions and the constraints needed to make the model identifiable. Give the least squares estimates of the main effect corresponding to Location 2 and the interaction effect between Factory 3 and Location 1.

   (ii) (8) After fitting the ANOVA model to the data, an ANOVA table is obtained. This table is partially presented below. Provide the missing information where possible. Round to two decimal places if necessary.

```
                   Df  Sum Sq  Mean Sq  F value
Location           __  _______  _______     1.49
Factory            __  _______    30561  _______
Location:Factory   __     3017  _______  _______
Residuals          __    44327  _______
```

   (iii) (7) Based on the completed ANOVA table of part (ii), carry out a two-way ANOVA for the two factors and their interaction.

   (iv) (7) Based on the results of fitting the full model, one decides to fit a one-way ANOVA model instead. Which factor should then be used? Describe what the corresponding incidence matrix for this one-way ANOVA model should be and present schematically the corresponding one-way ANOVA table, provide the numbers in the column `Df` (degrees of freedom).

2. The Michaelis–Menten model for enzyme kinetics may be written as $y = f(x, \theta_1, \theta_2) = \dfrac{\theta_1 x}{\theta_2 + x}$, where y is the reaction rate, x is the concentration of a substrate, $\theta_1, \theta_2 > 0$. Let data $(Y_1, x_1), \ldots, (Y_n, x_n)$ from n observations be obtained, from which one wishes to estimate $\boldsymbol{\theta} = (\theta_1, \theta_2)$ by fitting the model $Y_i = f(x_i, \theta_1, \theta_2) + \varepsilon_i$, $i = 1, \ldots, n$, where $\varepsilon_1, \ldots, \varepsilon_n$ are independent random errors with mean zero and variance σ^2. Suppose $n = 100$.

   (i) (3) Give the normal equations used for calculating the LSE of $\boldsymbol{\theta} = (\theta_1, \theta_2)$.

   (ii) (6) Suppose that one obtained the LSE $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2) = (4, 3)$ for the parameter $\boldsymbol{\theta}$, and the estimate
   $$\widehat{\mathrm{Cov}(\hat{\boldsymbol{\theta}})} = \hat{\sigma}^2 (\hat{V}^T \hat{V})^{-1} \approx \begin{pmatrix} 1 & 0 \\ 0 & 0.16 \end{pmatrix}$$
   for the covariance matrix of $\hat{\boldsymbol{\theta}}$. Construct an approximate 95% confidence interval for θ_2 and test the hypothesis H_0: $\theta_2 = 2$ against H_1: $\theta_2 \neq 2$.

   (iii) (8) Construct an approximate 95% confidence interval for the expected response $f(1, \theta_1, \theta_2)$.

   (iv) (7) Suppose that the residual sum of squares is $S(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} [Y_i - f(x_i, \hat{\theta}_1, \hat{\theta}_2)]^2 = 49$, and suppose that also the LSE $\tilde{\theta}_1$ for the reduced (nested) model $Y_i = f(x_i, \theta_1, 1) + \varepsilon_i$, $i = 1, \ldots, n$, with the corresponding residual sum of squares $S(\tilde{\theta}) = \sum_{i=1}^{n} [Y_i - f(x_i, \tilde{\theta}_1, 1)]^2 = 51$ were obtained. Estimate the parameter σ^2 and test the hypothesis: `"the reduced model fits well"`.

3. Let $Y_1, \ldots, Y_n$ be independent random variables with Y_i having a probability density function given by
   $$f(y; \lambda_i, \kappa) = \Big(\frac{\lambda_i y}{\kappa}\Big)^{1/\kappa} \frac{e^{-\lambda_i y/\kappa}}{y\Gamma(1/\kappa)} \quad \text{for } \ y > 0,$$
   where $\lambda_i > 0$, $i = 1, \ldots, n$ are unknown parameters, the common parameter κ is assumed to be known, and $\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx$ is the gamma function.

   (i) (12) The general form of the exponential family is $f(y, \theta_i) = \exp\left\{\frac{y\theta_i - b(\theta_i)}{\phi/A_i} + c(y, \phi/A_i)\right\}$. Show that the distribution of Y_i can be written in this form with an appropriate function $h(\lambda_i) = \theta_i$. Identify the functions b, c, and the parameters ϕ and A_i,

   (ii) (4) Derive the canonical link function $g(\mu)$.

   (iii) (8) For any random variable Y, the *coefficient of variation* is defined by $v = \frac{[\mathrm{Var}(Y)]^{1/2}}{\mathbb{E}(Y)}$. Show that the coefficient of variation of Y_i does not depend on i.

4. Let $\{Z_t\}$ denote a white noise time series with variance σ^2.

   (i) (5) Let the time series $\{X_t\}$ be given by $X_t = Y + Z_{t-1} - 2Z_t$, where Y is some random variable, independent of $\{Z_t\}$, with $\mathrm{E}Y = 1$ and $\mathrm{Var}(Y) = 1$. Is $\{X_t\}$ weakly stationary?

   (ii) (6) Let the time series $\{X_t\}$ be given by $X_t = Z_1 + Z_{t-1} - 2Z_t$. Is $\{X_t\}$ weakly stationary?

   (iii) Let $\{X_t\}$ be the ARMA time series given by
   $$X_t = -0.1 X_{t-1} + Z_t + Z_{t-1}.$$

      (a) (5) Consider the time series $\{Y_t\}$ given by $Y_t = \nabla X_t = X_t - X_{t-1}$. Show that $\{Y_t\}$ follows an ARMA(p, q) model and identify the values of the parameters p and q.

      (b) (9) Assuming stationarity of $\{X_t\}$, derive $\mathrm{E}X_t$ and $\gamma_X(0) = \mathrm{Var}(X_t)$ in terms of σ^2.