Exam Statistical Models

16 December 2015

You may use a simple calculator provided it is not part of a communicating device, and 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) Write the appropriate two-way ANOVA model that can be applied to investigate the effects of factors Location and Factory (and their interaction) on the cable strength. Specify model assumptions and 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) Let the significance level $\alpha = 0.05$. Based on the completed ANOVA table in part (ii), carry out a two-way ANOVA for the both 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 is then to use? 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 $y = f(x, \theta_1, \theta_2) = \frac{\theta_1 x}{\theta_2 + x}$, where y is the reaction rate, x is the concentration of a substrate, $\theta_1, \theta_2 > 0$. We obtain data $\{(Y_1, x_1), \ldots, (Y_n, x_n)\}$ from n observations, and wish 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 common mean zero and common 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 we obtained the LSE $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2) = (4, 3)$ for the parameter $\boldsymbol{\theta}$ and an estimator for the covariance matrix of $\hat{\boldsymbol{\theta}}$

$$\widehat{\mathrm{Cov}(\hat{\boldsymbol{\theta}})} = \hat{\sigma}^2 (\hat{V}^T \hat{V})^{-1} \approx \begin{pmatrix} 1 & 0 \\ 0 & 0.16 \end{pmatrix}.$$

Construct a 95% (approximate) confidence interval for θ_2 and test the hypothesis H_0 : $\theta_2 = 2$ against $H_1 : \theta_2 \neq 2$.

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

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

3. Suppose $Y_1, \ldots, Y_n$ are independent random variables, with Y_i having the gamma distribution with positive parameter λ_i, $i = 1, \ldots, n$. The probability density function of Y_i is

$$f(y; \lambda_i, \kappa) = \left(\frac{\lambda_i y}{\kappa}\right)^{1/\kappa} \frac{e^{-\lambda_i y/\kappa}}{y\Gamma(1/\kappa)} \quad \text{for } y > 0,$$

where $\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx$ is the gamma function and κ is a known positive constant.

(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 $v = \frac{\mathrm{Var}(Y)}{[\mathbb{E}(Y)]^2}$. 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) (7) 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?

(i) (7) 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

$$X_t = -0.1X_{t-1} + Z_t + Z_{t-1}.$$

(a) (6) Assuming stationarity of $\{X_t\}$, derive $\mathrm{E}X_t$ and $\gamma_X(0) = \mathrm{Var}(X_t)$ in terms of σ^2.

(b) (5) By using (a), derive $\gamma_X(1) = \mathrm{Cov}(X_t, X_{t+1})$ in terms of σ^2.

1(i) $Y_{ijk} = \mu+\alpha_i+\beta_j+\gamma_{ij}+\epsilon_{ijk}$, $\epsilon_{ijk} \overset{\text{ind}}{\sim} N(0,\sigma^2)$, $i = 1,2,3$, $j = 1,2$, $k = 1,\ldots,10$. $\sum_{i=1}^{3}\alpha_i = 0$; $\sum_{j=1}^{2}\beta_j = 0$; $\sum_{i=1}^{3}\gamma_{ij} = 0$, $j = 1,2$; $\sum_{j=1}^{2}\gamma_{ij} = 0$, $i = 1,2,3$. Next, $\hat{\beta}_2 = \overline{Y}_{2\cdot\cdot} - \overline{Y}_{\cdots} = 741.9 - 746.4 = -4.5$, $\hat{\gamma}_{31} = \overline{Y}_{31\cdot} - \overline{Y}_{3\cdot\cdot} - \overline{Y}_{\cdot 1\cdot} + \overline{Y}_{\cdots} = 706.8 - 707.8 - 750.9 + 746.4 = -5.5$.

1(ii) As $I = 2$, $J = 3$, $K = 10$, $n = IJK = 60$.

	Df	Sum Sq	Mean Sq	F value
Location	$I-1=1$	$1224\times 1=1224$	$1.49*820.87=1224$	1.49
Factory	$J-1=2$	$30561\times 2=61122$	$820.87\times 37.23=30561$	37.23
Interaction	$(I-1)(J-1)=2$	3017	$3017/2=1508.5$	$1508.5/820.87=1.84$
Residuals	$n-IJ=54$	44327	$44327/54=820.87$	

1(iii) Since $1.84 < F_{2,54;0.95} = 3.17$, we do not reject $H_0 : \gamma_{ij} = 0$. Thus, there is no interaction effect. Since $1.49 < F_{1,54;0.95} = 4.02$, we do not reject $H_A : \alpha_i = 0$. Thus, there is no effect Location. Since $37.23 > F_{2,54;0.95} = 3.17$, we reject $H_B : \beta_j = 0$. Thus, the effect Factory is important.

1(iv) Since the F-statistic corresponding to the Factory effect (37.23) is much larger than the critical value of $\mathcal{F}_{2,54;0.95} = 3.17$, while the F statistic corresponding to the Location effect (1.49) is smaller than the critical value of $\mathcal{F}_{1,54;0.95} = 4.02$. The only significant factor of the two is Factory and the one-way ANOVA model should use Factory as the single factor.

The incidence matrix for the one-factor model with Factor Factory consists of 3 columns of length 60. The first column consists of the first 20 ones and the next 40 zeros, in the second column the second 20-tuple are ones and the rest are zeros, etc.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Factory	2	*61122*	*30561*	*35.87*	*
Residuals	57	*48560*	*851.93*		

The values between stars (like *61122* etc.) are actually not asked (but can in principle be derived) in the exam problems, students do not have to compute those values.

2(i) The two normal equations for calculating the LSE of $\boldsymbol{\theta}$ are

$$\sum_{i=1}^{n}\frac{x_i}{\theta_2+x_i}\left(Y_i-\frac{\theta_1 x_i}{\theta_2+x_i}\right)=0,\quad \sum_{i=1}^{n}\frac{\theta_1 x_i}{(\theta_2+x_i)^2}\left(Y_i-\frac{\theta_1 x_i}{\theta_2+x_i}\right)=0.$$

2(ii) Since $\hat{\theta}_k \pm t_{n-p;1-\alpha/2}\sqrt{\left[\widehat{\mathrm{Cov}(\hat{\boldsymbol{\theta}})}\right]_{kk}}$ is a $(1-\alpha)$-confidence interval for θ_k, $k = 1,\ldots,p$, in our case we obtain that $\hat{\theta}_2 \pm t_{98;0.975}\sqrt{0.16} = 3 \pm 1.98 \cdot 0.4 \approx (2.2, 3.8)$ is the approximate 0.95-confidence intervals for θ_2. The $H_0 : \theta_2 = 2$ is rejected since 2 is not contained by the 95%-confidence interval for θ_2.

2(iii) Compute $\hat{v}_x = (\frac{\partial f}{\partial \theta_1}(1,4,3), \frac{\partial f}{\partial \theta_2}(1,4,3))^T = (1/4, -1/4)$ and compute next $\hat{v}_x^T \widehat{\mathrm{Cov}(\hat{\boldsymbol{\theta}})}\hat{v}_x = \frac{1}{16} - 0.01 = 0.0525 = (0.23)^2$. So, the approximate 0.95%-confidence intervals for $f(1,\theta_1,\theta_2)$ is $f(1,\hat{\theta}_1,\hat{\theta}_2) \pm t_{98;0.975}\sqrt{\hat{v}_x^T \widehat{\mathrm{Cov}(\hat{\boldsymbol{\theta}})}\hat{v}_x} = 1 \pm 1.98 \cdot 0.23 = [0.54, 1.46]$.

2(iv) An estimator of σ^2 is $\hat{\sigma}^2 = S(\hat{\boldsymbol{\theta}})/(n-p) = \frac{49}{98} = 0.5$. To test the hypothesis: `the reduced model fits well`, we compare the statistics $\frac{(S(\tilde{\theta})-S(\hat{\boldsymbol{\theta}}))/(2-1)}{S(\hat{\boldsymbol{\theta}})/(100-2)} = \frac{51-49}{49/98} = 4 > F_{1,98;0.95} = 3.94$ so that this hypothesis is rejected, i.e., the reduced model does not fit well.

3(i) $f(y;\lambda_i,\kappa) = \exp\left\{\frac{\log(\lambda_i y/\kappa)}{\kappa} - \frac{\lambda_i y}{\kappa} - \log[y\Gamma(1/\kappa)]\right\} = \exp\left\{\frac{\lambda_i y - \log(\lambda_i)}{-\kappa} + \frac{\log(y/\kappa)}{\kappa} - \log[y\Gamma(1/\kappa)]\right\} = \exp\left\{\frac{y\theta_i - b(\theta_i)}{\phi/A_i} + c(y,\phi/A_i)\right\}$, where $\theta_i = \lambda_i$, $b(\theta_i) = \log\theta_i$, $\phi/A_i = -\kappa$ and $c(y,-\kappa) = \frac{\log(y/\kappa)}{\kappa} - \log[y\Gamma(1/\kappa)]$. This is the canonical form of the exponential family.

3(ii) The canonical link function is $g(\mu) = b'^{-1}(\mu) = \frac{1}{\mu}$.

3(iii) For the gamma data we have $\mathbb{E}(Y_i) = b'(\theta_i) = 1/\theta_i$ and $\mathrm{Var}(Y_i) = b''(\theta_i)\phi/A_i = (-1/\theta_i^2)(-\kappa) = \kappa/\theta_i^2$. Hence the coefficient of variation is $v = \mathrm{Var}(Y)/[\mathbb{E}(Y)]^2 = (\kappa/\theta_i^2)/(1/\theta_i)^2 = \kappa$, which does not depend on i.

4(i) The time series $\{X_t\}$ is weakly stationary, since $\mathrm{E}X_t = 1$ and $\mathrm{Cov}(X_t, X_{t+h}) = 1 + \gamma_Y(h)$, where $\gamma_Y(h)$ is the autocovariance function of the stationary MA time series $Y_t = Z_{t-1} - 2Z_t$.

4(ii) The time series $\{X_t\}$ is not weakly stationary, since $\mathrm{Var}(X_t) = 6\sigma^2$, if $t \neq 1, 2$; $\mathrm{Var}(X_1) = 2\sigma^2$ if $t = 1$, $\mathrm{Var}(X_1) = 8\sigma^2$ if $t = 2$, i.e., the variance of X_t depends on t.

4(iii) As $\{X_t\}$ is stationary, $\mathrm{E}(X_t) = \mu$ must be constant so that by taking $\mathrm{E}(X_t)$ of both sides of the equation describing the time series $\{X_t\}$, we obtain that $\mathrm{E}(X_t) = 0$. Next, as $\{X_t\}$ is stationary, $\mathrm{Var}(X_t) = \gamma_X(0)$ must be constant so that by taking Var of both sides of the equation for $\{X_t\}$, we obtain $\gamma_X(0) = 0.01\gamma_X(0) + \mathrm{var}(Z_t) + \mathrm{var}(Z_{t-1}) - 2\mathrm{Cov}(0.1X_{t-1}, Z_{t-1}) = 0.01\gamma_X(0) + 2\sigma^2 - 0.2\sigma^2$ which implies $\gamma_X(0) = 180\sigma^2/99$.

As $\mathrm{E}X_t = 0$, $\gamma_X(1) = \mathrm{E}(X_t X_{t-1})$. Using the equation describing the time series $\{X_t\}$, we derive $\mathrm{E}(X_t X_{t-1}) = -0.1\mathrm{E}(X_{t-1}^2) + \mathrm{E}(Z_t X_{t-1}) + \mathrm{E}(Z_{t-1}X_{t-1}) = -0.1\gamma_X(0) + \mathrm{E}(Z_{t-1}^2) = 117\sigma^2/99$.