Statistiek (WISB263)

Sketch of Solutions of the Final Exam

January 28, 2019

Schrijf uw naam op elk in te leveren vel. Schrijf ook uw studentnummer op blad 1.
(The exam is an *open–book* exam: notes and book are allowed. The scientific calculator is allowed as well).
The maximum number of points is 110 (10 extra BONUS points!!).
Grade= $min(100, \text{points})$.
Points distribution: 25–16–12-15–32 (+10 extra BONUS points!!)

1. Let $\mathbf{Y} = \{Y_1, \ldots, Y_N\}$ be a random sample such that $Y_i \overset{i.i.d.}{\sim} \text{Pois}(\theta)$ for $i \in \{1, \ldots, N\}$.

   (a) [7pt] Find the Maximum Likelihood Estimator (MLE) of θ and its asymptotic distribution.
   Solution.
   The log-likelihood is:
   $$\ell(\theta; \mathbf{Y}) = \sum_{i=1}^{N} (Y_i \ln(\theta) - \theta - \ln(Y_i!))$$
   so that
   $$\ell'(\theta; \mathbf{Y}) = \frac{1}{\theta} \sum_{i=1}^{N} Y_i - N$$
   Hence
   $$\hat{\theta}_{\text{MLE}} = \overline{Y}_N$$
   since
   $$\ell''(\theta; \mathbf{Y}) = -\frac{1}{\theta^2} \sum_{i=1}^{N} Y_i < 0.$$
   By the CLT, since $\text{Var}(Y_i) = \theta$ we have:
   $$\frac{\sqrt{N}(\overline{Y}_N - \theta)}{\sqrt{\theta}} \xrightarrow{d} N(0, 1)$$
   so that $\overline{Y}_N \approx N(\theta, \theta/N)$.

   (b) [7pt] Find the MLE of e^θ and show that it is biased.
   Solution.
   By the invariance principle, the MLE of e^θ is $e^{\hat{\theta}_{\text{MLE}}} = e^{\overline{Y}_N}$. Since the exponential function is a strictly convex function, by Jensen inequality we have:
   $$\mathbb{E}_\theta \left(e^{\overline{Y}_N} \right) > e^{\mathbb{E}_\theta(Y_N)} = e^\theta$$

   (c) [11pt] In case we are able to measure only the first n ($n < N$) observations, and of the other $N - n$ observations we know the value x of the sum, determine the MLE of θ. Compare this result with point (a).
   Solution.
   If we define:
   $$U := \sum_{i=n+1}^{N} Y_i$$
   we have that $U \sim \text{Pois}((N - n)\theta)$. Therefore, the log-likelihood is in this case:
   $$\ell(\theta; \mathbf{Y}) = \sum_{i=1}^{n} (Y_i \ln(\theta) - \theta - \ln(Y_i!)) + x \ln((N - n)\theta) - (N - n)\theta - \ln(x!)$$

$$\ell'(\theta;\mathbf{Y}) = \frac{1}{\theta}\sum_{i=1}^{n} Y_i - n + \frac{1}{\theta}x - (N-n) = \frac{1}{\theta}(\sum_{i=1}^{n} Y_i + x) - N = \frac{1}{\theta}N\overline{Y}_N - N$$

so that:

$$\hat{\theta}^{\star}_{\text{MLE}} = \overline{Y}_N = \hat{\theta}_{\text{MLE}}$$

2. Let X be a discrete random variable whose probability mass function $p(x) = \mathbb{P}(X = x)$ under H_0 and H_1 is given by:

x	1	2	3	4	5	6	7
$p(x\|H_0)$	0.01	0.01	0.01	0.01	0.01	0.01	0.94
$p(x\|H_1)$	0.06	0.05	0.04	0.03	0.02	0.01	0.79

(a) [8pt] Find the most powerful test for testing H_0 versus H_1 with significance level $\alpha = 0.04$.
Solution.
By the Neyman-Pearson Lemma, the most powerful test rejects for small values of the ratio $p(x|H_0)/p(x|H1)$. Computing this ratio we obtain:

x	1	2	3	4	5	6	7
$\Lambda(x) = \frac{p(x\|H_0)}{p(x\|H_1)}$	1/6	1/5	1/4	1/3	1/2	1	1.19

that is increasing in x. So rejecting for small values of Λ corresponds to rejecting for small values of x. Since $\alpha = 0.04$, we need to choose c such that $0.04 = \mathbb{P}(\Lambda(X) < c|H_0) = \mathbb{P}(X \leq \bar{c}|H_0)$. Therefore, $\bar{c} = 4$.

(b) [8pt] Compute the power for this test.
Solution.

$$\pi = \mathbb{P}(X \leq \bar{c}|H_1) = 0.18$$

3. Consider the oneway ANOVA model:

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

with $i \in \{1,\ldots,I\}$, $j \in \{1,\ldots,J\}$ and $\epsilon_{ij} \overset{i.i.d.}{\sim} N(0,\sigma^2)$.

(a) [6pt] Show that set of statistics $(\overline{Y}_{1\cdot}, \overline{Y}_{2\cdot}, \ldots \overline{Y}_{I\cdot}, S_p^2)$ is sufficient for $(\mu_1, \mu_2, \ldots, \mu_I, \sigma^2)$, where $\overline{Y}_{i\cdot} := \frac{1}{J}\sum_{j=1}^{J} Y_{ij}$ and $S_p^2 := \frac{SS_W}{I(J-1)}$.
Solution.
Under ANOVA assumptions Y_{ij} are independent random variables such that $Y_{ij} \sim N(\mu_i, \sigma^2)$. The joint likelihood function can be written:

$$\begin{aligned} L(\boldsymbol{\mu},\sigma^2;\mathbf{Y}) &= \prod_{i=1}^{I}\prod_{j=1}^{J} \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{(Y_{ij}-\mu_i)^2}{2\sigma^2}} = \frac{1}{(2\pi\sigma^2)^{IJ/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i,j}(Y_{ij}-\mu_i)^2\right) \\ &= \frac{1}{(2\pi\sigma^2)^{IJ/2}} \exp\left(-\frac{1}{2\sigma^2}J\sum_{i=i}^{I}\mu_i^2\right) \exp\left(-\frac{1}{2\sigma^2}\sum_{i,j}Y_{ij}^2 + \frac{2J}{2\sigma^2}\sum_{i=1}^{I}\mu_i\overline{Y}_{i\cdot}\right) \\ &= \frac{1}{(2\pi\sigma^2)^{IJ/2}} \exp\left(-\frac{1}{2\sigma^2}J\sum_{i=i}^{I}\mu_i^2\right) \exp\left(-\frac{I(J-1)S_p^2}{2\sigma^2} + \frac{J}{\sigma^2}\sum_{i=1}^{I}\mu_i\overline{Y}_{i\cdot}\right) \end{aligned}$$

Therefore, by the Factorization Theorem, $(\overline{Y}_{1\cdot}, \ldots, \overline{Y}_{I\cdot}, S_p^2)$ is a sufficient statistic for $(\mu_1, \ldots, \mu_I, \sigma^2)$.

(b) [6pt] Show that S_p^2 is independent of each $\overline{Y}_{i\cdot}$.
Solution.
Since

$$S_p^2 = \frac{J}{I(J-1)}\sum_{i=1}^{I} S_i^2$$

and by $(Y_{ij} - \overline{Y}_{i\cdot}) \perp Y_{ij}$ for all i,j, the result easily follows.

4. An algorithm is developed for generating pseudo-random numbers. We want to test now this algorithm by performing an experiment in which it produces $n = 10000$ digits (i.e. each digit is an integer in $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$). Suppose the following frequencies are observed:

Digit	0	1	2	3	4	5	6	7	8	9
Frequency	1007	987	928	986	1010	1029	987	1006	1034	1026

(a) [15pt] Choose an appropriate test for testing at $\alpha = 0.05$ level of significance whether the algorithm is a proper pseudo-random number generator. Explain clearly your choice and perform the test.
Solution.

Since the large sample size we apply the Pearson χ^2 goodness of fit test for a multinomial distribution (that we proved to be asymptotically equivalent to the GLRT). Under the null hypothesis $p_i = 1/10$, so that the realization of the χ^2 test statistic χ^2 distributed with 9 degrees of freedom is:

$$\chi^2 = \frac{1}{1000}\sum_{i=1}^{10}(x_i - 10000/10)^2 = 8.576$$

Since $p = 0.48 > 0.05$ we do not reject H_0.

5. Consider the multiple linear regression model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

with $\boldsymbol{\beta}^\top = (\beta_0, \beta_1, \beta_2)$, $\mathbf{Y}^\top = (Y_1, \ldots, Y_n)$, where $n = 40$ is the sample size and $\mathbf{e}^\top = (\epsilon_1, \ldots, \epsilon_n)$ with ϵ_i i.i.d. $N(0, \sigma^2)$.
We obtain the following estimates of the least squares estimators and for the residual sum of squares: $\hat{\boldsymbol{\beta}}^\top = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = (3, 4, -3)$; $\hat{\mathbf{e}}^\top\hat{\mathbf{e}} = 37$, with $\hat{\mathbf{e}} := \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. Moreover, we know that:

$$(\mathbf{X}^\top\mathbf{X})^{-1} = \begin{pmatrix} 3 & -2 & 1 \\ -2 & 4 & 0 \\ 1 & 0 & 3 \end{pmatrix}$$

(a) [10pt] Find 95% CIs for each of $\beta_0, \beta_1, \beta_2$.
Solution.
An unbiased estimator of σ^2 is given by:

$$s^2 = \frac{\hat{\mathbf{e}}^\top\hat{\mathbf{e}}}{n-p} = \frac{37}{37} = 1$$

so that the estimated covariance matrix $\Sigma = \mathrm{Cov}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}})$ is given by:

$$\Sigma = \begin{pmatrix} 3 & -2 & 1 \\ -2 & 4 & 0 \\ 1 & 0 & 3 \end{pmatrix}$$

We know that

$$\frac{\hat{\beta}_i - \beta_i}{s_{\hat{\beta}_i}} \sim t_{37}$$

with $s_{\hat{\beta}_i} = s(\mathbf{X}^\top\mathbf{X})^{-1}_{ii}$. Therefore a 95% CI for β_0 is 3 ± 3.46, for β_1 is 4 ± 4.05 and for β_2 is -3 ± 3.46

(b) [10pt] Test the following hypotheses at 0.05 level of significance and state the conclusion:

$$\begin{cases} H_0: & \beta_0 + 2\beta_1 = 10, \\ H_1: & \beta_0 + 2\beta_1 \neq 10, \end{cases}$$

Solution.
Let us calculate a 95% CI for $\beta_0 + 2\beta_1$ and then apply the duality theorem. We notice that:

$$\frac{\hat{\beta}_0 + 2\hat{\beta}_1 - (\beta_0 + 2\beta_1)}{\sqrt{\text{Var}(\hat{\beta}_0 + 2\hat{\beta}_1)}} \sim t_{37}$$

Since

$$\text{Var}(\hat{\beta}_0 + 2\hat{\beta}_1) = \text{Var}(\hat{\beta}_0) + 4\text{Var}(\hat{\beta}_1) + 4\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \Sigma_{1,1} + 4\Sigma_{2,2} + 4\Sigma_{1,2} = 3 + 16 - 8 = 11$$

we have that a 95% CI for $\beta_0 + 2\beta_1$ is $11 \pm 2.02 \times \sqrt{11}$. Since 10 lies inside the CI we do not reject H_0.

(c) [5pt] Given the fitted linear model, we want now to make a prediction. We are interested indeed at the predicted value $\hat{Y}$ corresponding to the values of the predictors: $x_1 = 1; x_2 = -1$. Give a 95% CI for the *true* prediction (i.e. $\beta_0 + \beta_1 x_1 + \beta_2 x_2$).
Solution:
The prediction is:

$$\hat{y} = 3 + 4 + 3 = 10$$

The variance of the prediction:

$$\begin{aligned} \text{Var}\hat{Y} &= \text{Var}\hat{\beta}_0 + x_1^2\text{Var}\hat{\beta}_1 + x_2^2\text{Var}\hat{\beta}_2 + 2x_1\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) + 2x_2\text{Cov}(\hat{\beta}_0, \hat{\beta}_2) + 2x_1x_2\text{Cov}(\hat{\beta}_0, \hat{\beta}_2) \\ &= \text{Var}\hat{\beta}_0 + \text{Var}\hat{\beta}_1 + \text{Var}\hat{\beta}_2 + 2\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) - 2\text{Cov}(\hat{\beta}_0, \hat{\beta}_2) - 2\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) \\ &= \Sigma_{1,1} + \Sigma_{2,2} + \Sigma_{3,3} + 2\Sigma_{1,2} - 2\Sigma_{1,3} - 2\Sigma_{2,3} \\ &= 3 + 4 + 3 - 4 - 2 + 0 = 4 \end{aligned}$$

Therefore a 95% CI for the prediction is:

$$10 \pm t_{37,0.025}\sqrt{\text{Var}\hat{Y}} = 10 \pm 3.92$$

(d) [7pt] Suppose now that σ^2 is known and that $\sigma^2 = 1$. Are the least square estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ independent? Prove your statement.
Solution:
For multivariate normal distributed RVs null covariance implies independence. Since $\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) = \Sigma_{2,3} = 0$, the independence follows.

6. **BONUS Ex.1** [5pt]: By using properly the CLT, calculate the following limit:

$$\lim_{n \to +\infty} e^{-n} \sum_{k=0}^{n+\sqrt{n}} \frac{n^k}{k!}$$

7. **BONUS Ex.2** [5pt]: Let $\mathbf{X} = \{X_1, \ldots, X_n\}$ be a random sample such that $X_i \overset{i.i.d.}{\sim} \text{Unif}(\frac{1}{\theta}, \theta)$ with $\theta > 1$.

(a) Find the Maximum Likelihood Estimator (MLE) of θ and state conditions on the sample $\mathbf{X}$ for guaranteeing its existence.