UPPSALA UNIVERSITY
Department of Mathematics
Rolf Larsson

Exam in Mathematical Statistics
Inference 1 1MS035
2023–06–15
Solutions

1. We have a random sample 0.3, 0.7, 1.5, 0.8, 0.6 from a continuous random variable X with density function

$$f(x) = \frac{1}{2\theta^3} x^2 e^{-x/\theta},$$

where $x > 0$ and $\theta > 0$.

(a) Estimate θ by using the method of moments. (1p)
Hint: Without proof, you may use that $E(X) = 3\theta$.

Solution: The sample mean is $\bar{x} = 0.78$. The moment estimate solves

$$0.78 = \bar{x} = m(\theta) = E(X) = 3\theta,$$

so we get the estimate $\theta^* = 0.78/3 = 0.26$.

(b) Estimate θ by using the method of maximum likelihood. (4p)

Solution: For a random sample $x_1, ..., x_n$, the likelihood is

$$\begin{aligned} L(\theta) = f(x_1) \cdot ... \cdot f(x_n) &= \frac{1}{2\theta^3} x_1^2 e^{-x_1/\theta} \cdot ... \cdot \frac{1}{2\theta^3} x_n^2 e^{-x_n/\theta} \\ &= 2^{-n} x_1^2 \cdot ... \cdot x_n^2 \theta^{-3n} \exp\left(-\frac{1}{\theta} \sum_{i=1}^n x_i \right). \end{aligned}$$

We want to maximize the likelihood w.r.t. θ. This is equivalent to maximizing the log likelihood

$$l(\theta) = \ln\{L(\theta)\} = C - 3n \ln \theta - \frac{1}{\theta} \sum_{i=1}^n x_i,$$

where C is a constant not depending on θ.
Differentiation yields

$$\begin{aligned} l'(\theta) &= -\frac{3n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^n x_i, \\ l''(\theta) &= \frac{3n}{\theta^2} - \frac{2}{\theta^3} \sum_{i=1}^n x_i. \end{aligned}$$

We solve $l'(\theta) = 0$ to get

$$\theta = \frac{1}{3n} \sum_{i=1}^n x_i = \frac{\bar{x}}{3},$$

which gives a maximum because

$$l''\left(\frac{\bar{x}}{3}\right) = \frac{3n}{(\bar{x}/3)^2} - \frac{2}{(\bar{x}/3)^3} n\bar{x} = -\frac{3^3 n}{\bar{x}^2} < 0.$$

Then, by inserting the numbers we get the ML estimate

$$\theta^* = \frac{\bar{x}}{3} = 0.26,$$

which equals the moment estimate in this case.

2. We have a random sample x_1, x_2, x_3 of the random variable X which has expectation μ and variance 1, and a random sample y_1, y_2, y_3, y_4 of the random variable Y which has expectation 2μ and variance 4. The means of the samples are denoted by $\bar{x}$ and $\bar{y}$, respectively. We may assume that X and Y are independent.

   Two estimates of μ are proposed:

$$\mu_1^* = \frac{2\bar{x} + \bar{y}}{4}, \quad \mu_2^* = \frac{3\bar{x} + 2\bar{y}}{7}.$$

   (a) Show that μ_1^* and μ_2^* are both unbiased for μ. (2p)

   Solution: The corresponding estimators are

$$\mu_1^* = \frac{1}{2}\bar{X} + \frac{1}{4}\bar{Y}, \quad \mu_2^* = \frac{3}{7}\bar{X} + \frac{2}{7}\bar{Y},$$

   where $\bar{X}$ and $\bar{Y}$ are the random variables that correspond to $\bar{x}$ and $\bar{y}$. Because $E(\bar{X}) = E(X) = \mu$, and similarly $E(\bar{Y}) = 2\mu$, we get

$$E(\mu_1^*) = \frac{1}{2}E(\bar{X}) + \frac{1}{4}E(\bar{Y}) = \frac{1}{2} \cdot \mu + \frac{1}{4} \cdot 2\mu = \mu,$$

   and

$$E(\mu_2^*) = \frac{3}{7}E(\bar{X}) + \frac{2}{7}E(\bar{Y}) = \frac{3}{7} \cdot \mu + \frac{2}{7} \cdot 2\mu = \mu,$$

   showing that both estimates are unbiased for μ.

(b) Which one of μ_1^* and μ_2^* is most efficient? Motivate your answer. (3p)

Solution: At first, we note that

$$V(\bar{X}) = \frac{V(X)}{3} = \frac{1}{3}, \quad V(\bar{Y}) = \frac{V(Y)}{4} = 1.$$

This gives

$$\begin{aligned} V(\mu_1^*) &= \left(\frac{1}{2}\right)^2 V(\bar{X}) + \left(\frac{1}{4}\right)^2 V(\bar{Y}) \\ &= \frac{1}{4} \cdot \frac{1}{3} + \frac{1}{16} \cdot 1 = \frac{7}{48} \approx 0.1458, \end{aligned}$$

and

$$\begin{aligned} V(\mu_2^*) &= \left(\frac{3}{7}\right)^2 V(\bar{X}) + \left(\frac{2}{7}\right)^2 V(\bar{Y}) \\ &= \frac{9}{49} \cdot \frac{1}{3} + \frac{4}{49} \cdot 1 = \frac{7}{49} = \frac{1}{7} \approx 0.1429, \end{aligned}$$

which means that $V(\mu_2^*) < V(\mu_1^*)$.
Hence, μ_2^* is most efficient of the two.

3. The number of power failures at Donald's summer house follows a Poisson distribution with parameter (mean) λ. During one summer, the Larsson family rents Donald's summer house. Donald tells the family that λ is at most 1.

   (a) The family suspects that there are more power failures at the summer house than what Donald claims. In fact, it turns out that during the summer when they rent it, there are three power failures in total.

   Test a suitable hypothesis to try to check if Donald is telling the truth. (2p)

   Solution: Test H_0: $\lambda = 1$ vs H_1: $\lambda > 1$. (This means that if we reject H_0, we have evidence that Donald does not tell the truth.)

   Let X be the number of power failures. Under H_0, we have $X \sim \text{Po}(1)$. We observe $x = 3$. We reject H_0 for 'at least as extreme' values on x, which gives the P value

$$P(X \geq 3) = 1 - P(X \leq 2) = 1 - e^{-1} - \frac{1^1}{1!}e^{-1} - \frac{1^2}{2!}e^{-1} = 1 - \frac{5}{2}e^{-1} \approx 0.08,$$

   and since $0.08 > 0.05$ we find that we may not reject H_0 at the 5% level. On this risk level, we have no evideince that Donald is not telling the truth.

   (b) Calculate the power of the test in (a) if in fact, $\lambda = 5$. (3p)

   Solution: At first, we calculate the critical region $\{x \geq C\}$, where C is such that we (just) may reject H_0 at the 5% level. We have seen that $C = 3$ is not good enough. Analogous to (a), we find

$$P(X \geq 4) = 1 - P(X \leq 3) \approx 1 - 0.98 = 0.02,$$

   so we may take $C = 4$.

   Hence, the power when $\lambda = 5$ equals

$$P(X \geq 4; \lambda = 5) = 1 - P(X \leq 3; \lambda = 5) \approx 1 - 0.265 = 0.735,$$

   i.e. 73.5%.

4. Seasonal ranges (in hectares) for alligators were monitored by biologists on a lake in Florida. Five alligators monitored in the spring showed ranges of 8.0, 12.1, 8.1, 18.2, 31.7. Four different alligators monitored in the summer showed ranges of 102.0, 81.7, 54.7, 50.7.

   Estimate the difference between mean spring and summer ranges, with a 95% confidence interval. Be careful to state your assumptions. (5p)

   Solution: We assume that we have two independent samples, $x_1, .., x_{n_1}$ from $X \sim N(\mu_1, \sigma_1^2)$ (spring) and $y_1, .., y_{n_2}$ from $Y \sim N(\mu_2, \sigma_2^2)$ (summer), where $n_1 = 5$, $n_2 = 4$. The parameters μ_1, μ_2, σ_1^2 and σ_2^2 are considered unknown.

   We have observed the means $\bar{x} = 15.620$ and $\bar{y} = 72.275$, and the sample variances $s_x^2 = 98.057$ and $s_y^2 = 582.2558$. These variances are very distinct from each other, so it does not seem appropriate to assume that $\sigma_1^2 = \sigma_2^2$.

   To calculate the confidence interval, we at first need to solve

$$\frac{1}{f} = \frac{1}{n_1 - 1} \frac{(n_2 s_x^2)^2}{(n_2 s_x^2 + n_1 s_y^2)^2} + \frac{1}{n_2 - 1} \frac{(n_1 s_y^2)^2}{(n_2 s_x^2 + n_1 s_y^2)^2}$$

   for f, and by inserting numbers we so obtain $f \approx 3.81 \approx 4$. We get the 95% confidence interval

$$\begin{aligned} I_{\mu_1 - \mu_2} &= \bar{x} - \bar{y} \pm t_{0.025}(4) \sqrt{\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}} \\ &= 15.620 - 72.275 \pm 2.7764 \sqrt{\frac{98.057}{5} + \frac{582.2558}{4}} \\ &= -56.655 \pm 35.682 = (-92.3,\ -21.0). \end{aligned}$$

   where $t_{0.025}(4) = 2.7764$ is obtained from table 6.

   Assuming equal variances, the confidence interval is $(-84.4,\ -28.9)$, hence considerably more narrow.

5. We have one observation $x = 1.2$ of the random variable X, which is exponentially distributed with parameter β, i.e. it has density function $f(x) = \beta e^{-\beta x}$ for $x > 0$ and 0 otherwise.

   Calculate a 90% confidence interval for β. (5p)

   Solution: As a reference variable, take $R = \beta X$, which is Exponentially distributed with parameter 1 (Exp(1)) because it has distribution function

$$\begin{aligned} F_R(t) &= P(R \leq t) = P(\beta X \leq t) = P\left(X \leq \frac{t}{\beta}\right) \\ &= 1 - \exp\left(-\beta \cdot \frac{t}{\beta}\right) = 1 - \exp(-t). \end{aligned}$$

   The quantiles r_α for $R \sim \text{Exp}(1)$ are given by

$$\alpha = P(R > r_\alpha) = \exp(-r_\alpha),$$

   i.e. $r_\alpha = -\ln \alpha$. Now,

$$\begin{aligned} 1 - \alpha &= P(r_{1-\alpha/2} < R < r_{\alpha/2}) = P\left(r_{1-\alpha/2} < \beta X < r_{\alpha/2}\right) \\ &= P\left(\frac{r_{1-\alpha/2}}{X} < \beta < \frac{r_{\alpha/2}}{X}\right) = P\left(\frac{-\ln(1-\alpha/2)}{X} < \beta < \frac{-\ln(\alpha/2)}{X}\right), \end{aligned}$$

   which with $\alpha = 0.10$ and $x = 1.2$ yields the 90% confidence interval

$$I_\beta = \left(\frac{-\ln(0.95)}{1.2}, \ \frac{-\ln(0.05)}{1.2}\right) = (0.043, \ 2.946).$$

6. Reaction times were measured for a random sample $x_1, ..., x_n$ of $n = 200$ car drivers. The mean reaction time in the sample was $\bar{x} = 1.1$ seconds, and the standard deviation was $s = 0.3$.

   (a) Calculate a 95% confidence interval for the mean reaction time for car drivers in the whole population. (4p)

   Solution: Let the reaction time be described by the random variable X, with $E(X) = \mu$. We want to calculate a 95% confidence interval for μ.

   Because n is large, the central limit theorem motivates the reference variable

$$R = \frac{\bar{X} - \mu}{s/\sqrt{n}} \approx N(0, 1),$$

   where $\bar{X}$ is the mean of $X_1, ..., X_n$, the random variables corresponding to the sample $x_1, ..., x_n$. It follows that a 95% confidence interval for μ is given by

$$I_\mu = \bar{x} \pm \lambda_{0.025} \frac{s}{\sqrt{n}} = 1.1 \pm 1.96 \cdot \frac{0.3}{\sqrt{200}} = 1.1 \pm 0.0416 = (1.0584,\ 1.1416).$$

   (b) Somebody claims that the mean reaction time for the whole population of car drivers is 1.0 seconds. Can you find support for this claim? (1p)

   Solution: We find that $\mu = 1.0$ is not included in the confidence interval in (a). This is equivalent to rejecting $H_0 : \mu = 1.0$ vs H_1: $\mu \neq 1.0$ at the 5% level. Hence, at this risk level, we have evidence that the claim does not hold.

   Aternatively, calculate the observed test statistic

$$\frac{\bar{x} - 1.0}{s/\sqrt{n}} = \frac{1.1 - 1.0}{0.3/\sqrt{200}} \approx 4.71 > 1.96 = \lambda_{0.025},$$

   showing that we may reject H_0 at the 5% level. (In fact, H_0 may also be rejected at much smaller levels.)

7. At the University of Falnarp, a genetic experiment was conducted. A particular type of beans were cultivated. Upon harvest, it was expected that four genetic variants, numbered 1,2,3,4, should occur according to the proportions 1:3:3:9 (i.e. the first variant occurs in 1/16 of the cases, the second variant occurs in 3/16 of the cases etcetera). The results of the experiment are given in the table below.

Variant	1	2	3	4
Frequency	6	21	30	71

Does the experiment confirm the expectations about proportions? Try to answer this question by performing a suitable hypothesis test. (5p)

Solution: Let p_j be that probability that the bean belongs to variant j for $j = 1, 2, 3, 4$. We want to test H_0: $p_1 = 1/16$, $p_2 = 3/16$, $p_3 = 3/16$, $p_4 = 9/16$ vs H_1: $\neg H_0$. We use the χ^2 goodness of fit test ('test av anpassning').

The total number of beans in the experiment is $6 + 21 + 30 + 71 = 128$. This gives us the expected frequencies under H_0 as $e_j = 128p_j$, i.e.

$$e_1 = 128 \cdot \frac{1}{16} = 8,\ e_2 = 128 \cdot \frac{3}{16} = 24 = e_3,\ e_4 = 128 \cdot \frac{9}{16} = 72.$$

We find that all $e_j \geq 5$ and so χ^2 approximation is permitted. The observed test variable is

$$Q = \frac{(6-8)^2}{8} + \frac{(21-24)^2}{24} + \frac{(30-24)^2}{24} + \frac{(71-72)^2}{72} \approx 2.39.$$

The number of degrees of freedom is the number of cells minus one, $4 - 1 = 3$, and we have

$$Q \approx 2.39 < \chi^2_{0.05}(3) = 7.8147,$$

implying that we may not reject H_0 at the 5% level.

On this risk level, we have no evidence that the expected proportions are not correct.

8. The table below gives the total consumption of nuclear power electricity (in thousands of Tera Joule) in Sweden during the years 2000-2021 (data from Statistics Sweden, SCB). Is there any trend in this material? Check this by performing a suitable hypothesis test.

   It is not allowed to assume that the numbers are normally distributed. (5p)

Year	Consumption
2000	597
2001	771
2002	723
2003	717
2004	818
2005	755
2006	683
2007	662
2008	673
2009	540
2010	599
2011	603
2012	647
2013	662
2014	650
2015	559
2016	640
2017	643
2018	683
2019	682
2020	497
2021	541

Solution: Test H_0: trend vs H_1: no trend. We may use either the runs test or a test based on Spearman's rank correlation.

The runs test: The median in the material is 656 (the mean of the two 'middle' observations in terms of rank, 650 and 662). We assign 0 to observations less than the median och 1 to those greater than the median, giving the sequence
0 1 1 1 1 1 1 1 1 0 0 0 0 1 0 0 0 0 1 1 0 0.
This sequence has 7 runs.

We may use normal approximation, because $n_0 = n_1 = 11$, giving $\min(n_0, n_1) = 11 \geq 10$. We have $E(R) = 12$, $V(R) = 12 \cdot 11/21$. Thus, the observed test variable is

$$T = \frac{7 - 12}{\sqrt{12 \cdot 11/21}} \approx -2.18,$$

which is smaller than i.e. $-\lambda_{0.05} = -1.6449$, leading us to reject H_0 at the 5% level. (Observe that we only reject for few runs, since this is what corresponds to a trend.)

On this risk level, we have evidence of a trend.

Spearman's rank correlation: We may calculate the rank correlation (the correlation between the ranks of the data and the years) as $r_s = -0.5192$. Because the number of observations is $n = 22 \geq 10$, we may use normal approximation. We observe the test variable

$$T = \sqrt{21} \cdot (-0.5192) \approx -2.38 < -\lambda_{0.025} = -1.96,$$

and so, we may reject H_0 at the 5% level. Observe that the test is two-sided here, because the trend may be either positive or negative, yielding different signs on T.