UPPSALA UNIVERSITY
Department of Mathematics
Rolf Larsson

Exam in Mathematical Statistics
Inference 1 1MS035
2021–06–12
Solutions

1. We have a random sample 2.3, 1.2, 0.2, 6.7, 3.1 from a continuous random variable X with density function

$$f_X(x) = \frac{3x^2}{\theta^3} \exp\left(-\frac{x^3}{\theta^3}\right),$$

where $x > 0$, and 0 otherwise. Assume that $\theta > 0$.
Without proof, you may use that $E(X) = \theta\Gamma(4/3) \approx 0.893\theta$.

Estimate θ using

(a) the method of moments, (1p)

Solution: We have $m(\theta) = E(X) \approx 0.893\theta$ and $\bar{x} = 2.7$. Solving $m(\theta) = \bar{x}$ yields the moment estimate

$$\theta^* = \frac{2.7}{0.893} \approx 3.0.$$

(b) the least squares method, (2p)

Solution: Write $c = \Gamma(4/3) \approx 0.893$ and $n = 5$, and let the data be $x_1, ..., x_n$. We want to minimize

$$Q(\theta) = \sum_{i=1}^{n} \{x_i - m(\theta)\}^2 = \sum_{i=1}^{n} (x_i - c\theta)^2.$$

Differentiation yields

$$\begin{aligned} Q'(\theta) &= -2c \sum_{i=1}^{n} (x_i - c\theta) = -2cn(\bar{x} - c\theta), \\ Q''(\theta) &= 2c^2 n. \end{aligned}$$

We find that $Q''(\theta) > 0$, which means that we obtain a minimum by solving $Q'(\theta) = 0$, which gives the estimate $\theta^* = \bar{x}/c \approx 3.0$ as above.

(c) maximum likelihood. (2p)

Solution: With C as a constant (not depending on θ), we get the likelihood

$$L(\theta) = \prod_{i=1}^{n} f_X(x_i; \theta) = C\theta^{-3n} \exp\left(-\theta^{-3} \sum_{i=1}^{n} x_i^3\right).$$

Maximizing $L(\theta)$ is equivalent to maximizing

$$l(\theta) = \log L(\theta) = \log C - 3n \log \theta - \theta^{-3} \sum_{i=1}^{n} x_i^3.$$

We get the first two derivatives

$$\begin{aligned} l'(\theta) &= -3n\theta^{-1} + 3\theta^{-4} \sum_i x_i^3, \\ l''(\theta) &= 3n\theta^{-2} - 12\theta^{-5} \sum_i x_i^3. \end{aligned}$$

For $\theta > 0$, solving $l'(\theta) = 0$ is equivalent to solving

$$0 = -n + \theta^{-3} \sum_i x_i^3,$$

implying

$$\theta = \theta^* = \left(n^{-1} \sum_i x_i^3\right)^{1/3} \approx 4.1.$$

This yields a maximum, because

$$l''(\theta^*) = 3n\theta^{*-2} - 12\theta^{*-5} \sum_i x_i^3 = 3n\theta^{*-2} - 12\theta^{*-5} n\theta^{*3} = -9n\theta^{*-2} < 0.$$

Hence, $\theta^* \approx 4.1$ is the MLE of θ.

2. We have a random sample x_1, x_2, x_3, x_4 from a random variable X with expectation $\mu - m$ and variance 4, and another random sample $y_1, y_2, ..., y_5$ from a random variable Y with expectation m and variance 1. Moreover, we have one observation z from a random variable Z with expectation m and variance 1. We may assume that X, Y and Z are independent. The sample means are denoted $\bar{x}$ and $\bar{y}$.

   The following estimates of μ are proposed:

$$\mu_1^* = \bar{x} + \bar{y}, \quad \mu_2^* = \bar{x} + 5\bar{y} - 4z.$$

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

   Solution: Denote by $\overline{X}$ the random variable corresponding to $\bar{x}$, etcetera. For the estimators μ_1^* and μ_2^*, we have

$$E(\mu_1^*) = E(\overline{X} + \overline{Y}) = E(\overline{X}) + E(\overline{Y}) = E(X) + E(Y) = (\mu - m) + m = \mu$$

   and

$$\begin{aligned} E(\mu_2^*) &= E(\overline{X} + 5\overline{Y} - 4Z) = E(\overline{X}) + 5E(\overline{Y}) - 4E(Z) \\ &= E(X) + 5E(Y) - 4E(Z) = (\mu - m) + 5m - 4m = \mu, \end{aligned}$$

   showing that both estimates are unbiased.

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

   Solution: We calculate the variances of the estimators:

$$\begin{aligned} V(\mu_1^*) &= V(\overline{X}) + V(\overline{Y}) = \frac{V(X)}{4} + \frac{V(Y)}{5} = \frac{4}{4} + \frac{1}{5} = 1.2, \\ V(\mu_2^*) &= V(\overline{X}) + 5^2 V(\overline{Y}) + (-4)^2 V(Z) \\ &= \frac{V(X)}{4} + 25\frac{V(Y)}{5} + 16V(Z) = \frac{4}{4} + 5 + 16 = 22. \end{aligned}$$

   Because μ_1^* has the smallest variance, it is the most efficient of the two.

3. The time in days (and fractions of days) after the first of June until the water temperature in the lake 'Blåvattnet' exceeds 20 degrees (so that it is suitable to go swimming there) is exponentially distributed with expectation μ.

   Eskil believes that $\mu = 20$.

   (a) One year, Eskil has to wait 35 days from the first of June until the water temperature in the lake exceeds 20 degrees. Perform a suitable hypothesis test to find out if this in accord with $\mu = 20$. (2p)

   Solution: From the formulation of the problem, there is no particular reason why we should have $\mu > 20$ rather than $\mu < 20$ or the opposite. Hence, the natural thing to do is to test $H_0 : \mu = 20$ vs $H_1 : \mu \neq 20$.

   Let the time until the water temperature exceeds 20 degrees, X, be exponential with parameter $\beta = 1/\mu$. The distribution function is

   $$F(t; \beta) = P(X \leq t; \beta) = 1 - e^{-\beta t}.$$

   Using the direct method, we get that half the p value is (the test is double sided)

   $$P(X \geq 35; \beta = 0.05) = 1 - F(35; 0.05) = e^{-35*0.05} = 0.17,$$

   which is greater than 0.025, so we may not reject H_0 at the 5% level. We have no evidence that $\mu \neq 20$.

   (b) For $\mu = 40.0$, calculate the power of the test in (a). (3p)

   Solution: At first, we need to calculate the critical region of the test. A test a level 5% rejects H_0 if $X < c_1$ or $X > c_2$, for c_1 and c_2 such that $P(X \leq c_1; \beta = 0.05) = 0.025$ and $P(X \geq c_2; \beta = 0.05) = 0.025$.

   This gives $0.025 = 1 - e^{-0.05c_1}$, i.e. $c_1 = -20\ln(0.975) = 0.506$, and $0.025 = e^{-0.05c_2}$, i.e. $c_2 = -20\ln(0.025) = 73.8$.

   Now, for $\mu = 40.0$, i.e. $\beta = 1/40 = 0.025$, we get
   $P(X < 0.506) = 1 - e^{-0.025 \cdot 0.506} = 0.013$ and
   $P(X > 73.8) = e^{-0.025 \cdot 73.8} = 0.158$.

   Thus, the power for $\mu = 40$, i.e. the probability to reject H_0 for this μ, is $0.013 + 0.158 = 0.171$, so about 17%.

4. A group of five sprinters tests two types of running shoes, Neki and Pamu. One day, they run 100 meters with the Neki shoes, and the next day they run with Pamu. The weather conditions and the shapes of the sprinters are the same for both days. Their times are given in the table below.

Sprinter:	Ed	Beth	Sue	Bob	Usain
Neki	10.27	11.52	11.22	10.75	9.82
Pamu	10.12	11.12	11.27	10.50	9.72

In terms of achieving the shortest time, are the types of shoes equally good? Try to answer this question by performing a suitable statistical test. Make sure to specify all your assumptions. (5p)

Solution: This is a paired sample. We assume that for sprinter i, where $i = 1, 2, ..., 5$, the decrease in running time $Z \sim (\Delta, \sigma^2)$, where σ^2 is unknown. The sample of decreases, $z_1, z_2, ..., z_5$, is 0.15, 0.40, -0.05, 0.25, 0.10. The test statistic is

$$T = \frac{\overline{Z} - \Delta}{s/\sqrt{5}} \sim t(4),$$

where $\overline{Z} = \frac{1}{5}\sum_{i=1}^{5} Z_i$, where Z_i is the random variable corresponding to the observation z_i.

We want to test H_0: $\Delta = 0$ vs H_1: $\Delta \neq 0$ (from the outset, there is no suspicion that one particular type of shoe is better than the other).

We have $\bar{z} = 0.17$ and $s \approx 0.1681$. Hence, under H_0, we observe

$$T_{obs} = \frac{0.17 - 0}{0.1681/\sqrt{5}} \approx 2.26.$$

Because $2.26 < t_{0.025}(4) = 2.7764$, we may not reject H_0 at the 5% level.

At the 5% level, there is no evidence that the types of shoes are not equally good.

5. During autumn, a researcher randomly selects 20 moose bulls (Swedish: älgtjurar) in the forest "Storskogen", and measures their weights. She then does the same with 15 randomly selected moose bulls in the forest "Lillskogen". The weights of the selected moose bulls in "Storskogen" have mean 495 kg, and the standard deviation is 10. For "Lillskogen", the corresponding mean is 450 kg, with standard deviation 8.

   Let μ_1 be the expected weight for a moose bull in "Storskogen", and let μ_2 be the corresponding for "Lillskogen".

   (a) Calculate a 99% confidence interval for $\mu_1 - \mu_2$. Make sure to specify all your assumptions. (4p)

   Solution: We assume that we have two independent samples, $x_1, .., x_{n_1}$ from $X \sim N(\mu_1, \sigma_1^2)$ ("Storskogen") and $y_1, .., y_{n_2}$ from $Y \sim N(\mu_2, \sigma_2^2)$ ("Lillskogen"), where $n_1 = 20$, $n_2 = 15$. The parameters μ_1, μ_2, σ_1^2 and σ_2^2 are considered unknown. Moreover, assume that $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (say). We have observed the means $\bar{x} = 495$ and $\bar{y} = 450$, and the sample standard deviations $s_x = 10$ and $s_y = 8$. This gives us the pooled variance

$$s_p^2 = \frac{(n_1 - 1)s_x^2 + (n_2 - 1)s_y^2}{n_1 + n_2 - 2} = \frac{19 \cdot 10^2 + 14 \cdot 8^2}{33} \approx 84.727,$$

   and we get the 99% confidence interval

$$\begin{aligned} I_{\mu_1 - \mu_2} &= \bar{x} - \bar{y} \pm t_{0.005}(n_1 + n_2 - 2) s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \\ &= 495 - 450 \pm t_{0.005}(33)\sqrt{84.727}\sqrt{\frac{1}{20} + \frac{1}{15}} \\ &= 45 \pm 2.73\sqrt{84.727}\sqrt{\frac{1}{20} + \frac{1}{15}} \\ &= 45 \pm 8.6 = (36.4,\ \ 53.6). \end{aligned}$$

   where $t_{0.005}(33) \approx 2.73$ is obtained by "interpolation" from table 6.

   (b) On the significance level 1%, can you conclude that the expected weights μ_1 and μ_2 are different for the two forests? Motivate your answer. (1p)

   Solution: Yes, because $\mu_1 - \mu_2 = 0$ is not included in the 99% confidence interval.

6. In the November 2020 opinion poll by Statistics Sweden, the liberal party (L) got 3.8% of the sympathies. The number of respondents was 4692 people.

   In case of an election, L will have to leave the Swedish parliament if they get less than 4% of the votes.

   In case of an election in November 2020, perform a statistical test to judge if L would have had to leave the Swedish parliament. Motivate your answer. (5p)

   Solution: Let the number of L voters in the sample be X. Because the population is large, we may assume that $X \sim \mathrm{Bin}(n, p)$, where $n = 4692$ and p is the (unknown) proportion of L voters in the population. We want to see if $p < 0.04$. A natural thing is to test H_0: $p = 0.04$ vs H_1: $p < 0.04$. Then, if we reject H_0, we have some evidence that $p < 0.04$, causing L to leave the parliament in case of an election.

   The rule of thumb for normal approximation is fulfilled since

$$n \cdot 0.04 \cdot (1 - 0.04) = 4692 \cdot 0.04 \cdot 0.96 \approx 180 > 5.$$

   We have the test variable

$$T = \frac{X/n - 0.04}{\sqrt{0.04 \cdot (1 - 0.04)/n}} \approx N(0, 1).$$

   We have observed $x/n = 0.0038$. Hence, we get the observed test statistic

$$T_{obs} = \frac{0.038 - 0.04}{\sqrt{0.04 \cdot 0.96/4692}} \approx -0.70.$$

   Since $|T_{obs}| = 0.70 < \lambda_{0.05} = 1.6449$, we may not reject H_0 at the 5% level.

   So at this level, we have no evidence that L will have to leave the parliament.

7. The number of thunderstorms (Swedish: åskskurar) in Upptuna during summer is assumed to be Poisson distributed with parameter λ.

   One summer, there were 20 thunderstorms in Upptuna.

   (a) Calculate a 95% confidence interval for λ. (2p)

   Solution: Let the number of thunderstorms be $X \sim \text{Po}(\lambda)$. We observe $x = 20$. Hence, our estimate is $\lambda^* = 20$, which is greater than 15, so normal approximation is allowed. We have $V(X) = \lambda$ and the reference variable

$$\frac{X - \lambda}{\sqrt{\lambda^*}} \approx N(0, 1).$$

   In the usual way, this gives us the 95% confidence interval

$$I_\lambda = x \pm \lambda_{0.025}\sqrt{\lambda^*} = 20 \pm 1.96\sqrt{20} = 20 \pm 8.765 = (11.235,\ 28.765).$$

   In fact, this is not entirely correct since the lower limit of the interval is $11.235 < 15$, so it does not meet the rule of thumb for normal approximation. A better approach is the following:

   Let the 95% confidence interval consist of those $\lambda = \lambda_0$ that do not cause rejection of H_0: $\lambda = \lambda_0$ vs H_1: $\lambda \neq \lambda_0$ on the 5% level. We observe $x = 20$. The λ_0 that do not cause rejection are those who fulfill $P(X \leq 20; \lambda = \lambda_0) \geq 0.025$ and $P(X \geq 20; \lambda = \lambda_0) \geq 0.025$.

   By trial and error, we find $P(X \leq 20; \lambda = 30.89) \approx 0.025$ and $P(X \geq 20; \lambda = 12.22) \approx 0.025$.

   This leads to the 95% confidence interval $I_\lambda = (12.22,\ 30.89)$.

   (b) Calculate a 95% confidence interval for the probability of 25 thunderstorms or more during one summer in Upptuna. (3p)

   Solution: Let $p = P(X \geq 25)$. We want a confidence interval for p. By normal approximation with continuity correction,

$$p = 1 - P(X \leq 24) = 1 - P(X \leq 24.5) = 1 - P\left(\frac{X - \lambda}{\sqrt{\lambda}} \leq \frac{24.5 - \lambda}{\sqrt{\lambda}}\right)$$
$$\approx 1 - \Phi\left(\frac{24.5 - \lambda}{\sqrt{\lambda}}\right),$$

   where Φ is the distribution function of a standard normal variable.

   Since p is strictly increasing in λ, we have that the interval from (a), $11.235 \leq \lambda \leq 28.735$, corresponds to $p_l \leq p \leq p_u$, where

$$p_l \approx 1 - \Phi\left(\frac{24.5 - 11.235}{\sqrt{11.235}}\right) \approx 1 - \Phi(3.96) \approx 0.00004 \approx 0,$$
$$p_u \approx 1 - \Phi\left(\frac{24.5 - 28.735}{\sqrt{28.735}}\right) \approx 1 - \Phi(-0.79) = \Phi(0.79) \approx 0.785,$$

   and so, we get the confidence interval

$$I_p = (0,\ 0.785).$$

If we don't want to rely on the asymptotic approximation, we calculate

$$P(X \geq 25; \lambda = 12.22) \approx 0.0009$$

and

$$P(X \geq 25; \lambda = 30.89) \approx 0.88$$

By the same reasons as above, this gives us the confidence interval

$$I_p = (0.0009,\ 0.88).$$

8. Consider the regression model

$$Y_i = \alpha + \beta x_i + \varepsilon_i,$$

where $i = 1, 2, ..., n$ and all ε_i are independent $N(0, \sigma^2)$.

Data were simulated from this model four times, each time with different values on the parameters α, β and σ^2. The number of observations was $n = 200$. Afterwards, the data were plotted, the parameters were estimated and the coefficients of determination (R^2) were calculated. The plots are shown in figures 1-4 below.

Match figures 1-4 with the estimated models and corresponding R^2 values. Motivate your solution. (5p)

Models:

$$\begin{aligned} y_i^* &= 10.4 + 2.0x_i, \quad R^2 = 96\% \quad (a) \\ y_i^* &= 10.3 - 2.0x_i, \quad R^2 = 97\% \quad (b) \\ y_i^* &= 11.4 + 2.0x_i, \quad R^2 = 60\% \quad (c) \\ y_i^* &= 11.4 - 2.0x_i, \quad R^2 = 60\% \quad (d) \end{aligned}$$

Solution: At first, we find positive correlations (giving positive slopes of regression lines, when fitted) in figures 1 and 2, and negative correlations in figures 3 and 4.

For figures 1 and 2, the correlation is higher, corresponding to a higher R^2, in figure 2. This is what we see in model (a), so we have that model (a) corresponds to figure 2. Hence, figure 1 must correspond to the positive slope model with lower R^2, and this is model (c).

Simularly, in the negative correlation figures 3 and 4, the linear relationship is stronger (corresponding to a higher R^2) in figure 4, which must then correspond to model (b). Finally, this means that figure 3 corresponds to model (d).

Thus, the answer is a-2, b-4, c-1, d-3.

Appendix: figures

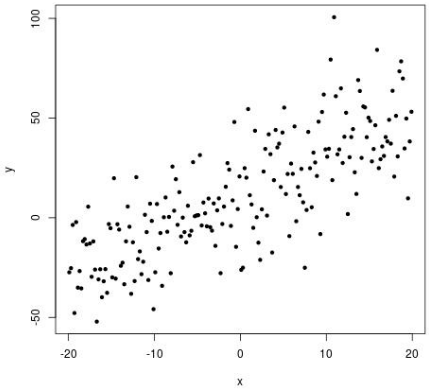

Figure 1: Plot for problem 8.

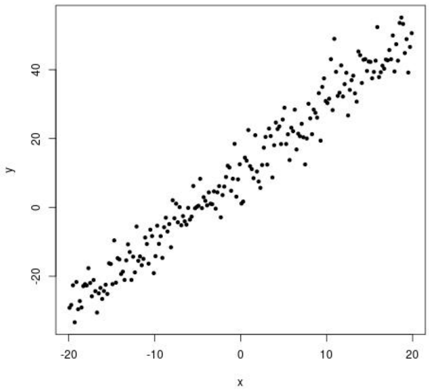

Figure 2: Plot for problem 8.

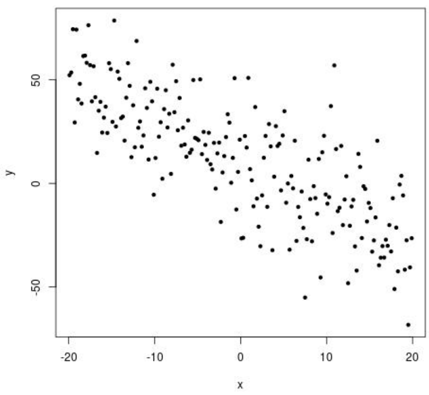

Figure 3: Plot for problem 8.

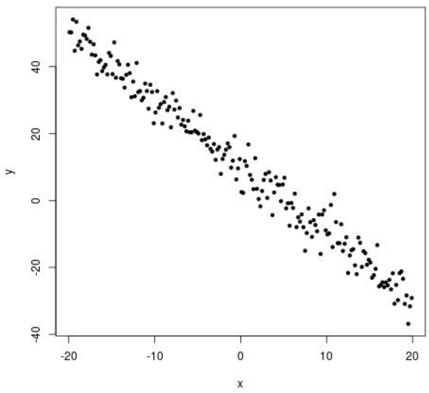

Figure 4: Plot for problem 8.