UPPSALA UNIVERSITY
Department of Mathematics
Rolf Larsson

Exam in Mathematical Statistics
Inference 1 1MS035
2021–08.16
Solutions

1. We have a random sample 2,1,1,3 from a random variable X with probability function

$$P(X=k)=\begin{cases} 1-6\theta, & k=0, \\ \theta k, & k=1,2,3, \\ 0, & \text{otherwise}, \end{cases}$$

where $0 \le \theta \le 1/6$. Estimate θ using

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

Solution: We observe $\bar{x}=(2+1+1+3)/4=7/4$, and we have

$$m(\theta)=E(X)=\sum_{k=0}^{3} kP(X=k)=1*\theta+2*(2\theta)+3*(3\theta)=14\theta.$$

Hence, solving $7/4=14\theta$ gives the moment estimate $\theta^*=1/8=0.125$.

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

Solution: Say that the observations are $x_1,...,x_n$. Because $m(\theta)=14\theta$, we need to minimize

$$Q(\theta)=\sum_{i=1}^{n}(x_i-14\theta)^2.$$

To this end, we calculate the derivatives

$$Q'(\theta)=-28\sum(x_i-14\theta)=-28n\bar{x}+392n\theta$$

and

$$Q''(\theta)=392n.$$

Because $Q''(\theta)>0$, we get a minimum by solving $0=Q'(\theta)$,
i.e. $\theta=\bar{x}/14$.
Hence, in our case, the LSE is $\theta^*=(7/4)/14=1/8$, the same as the moment estimate.

(c) maximum likelihood. (2p)

Solution: Because we have a discrete distribution, the likelihood equals the probability of observing the sample. Hence, it is, by independence,

$$\begin{aligned} L(\theta) &= P(X_1=2,X_2=1,X_3=1,X_4=3) \\ &= (2\theta)\cdot\theta\cdot\theta\cdot(3\theta)=6\theta^4. \end{aligned}$$

Observe that by assumption, $0\le\theta\le 1/6$. Since the likelihood is an increasing function of θ, we must have that its maximum in this interval is attained at its right endpoint, i.e. at $\theta=1/6$. Hence, the MLE is $\theta^*=1/6\approx 0.17$.

2. We have a random sample x_1, x_2, x_3, x_4, x_5 from a random variable X with expectation μ and variance 10, and another random sample y_1, y_2 from a random variable Y with expectation 2μ and variance 1. We may assume that X and Y are independent. The sample means are denoted by $\bar{x}$ and $\bar{y}$.

   The following estimates of μ are proposed:

$$\mu_1^* = \frac{\bar{x} + \bar{y}}{3}, \quad \mu_2^* = \frac{\bar{x} + 2\bar{y}}{5}.$$

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

   Solution: The expectations of the corresponding estimators are

$$E(\mu_1^*) = E\left(\frac{1}{3}\overline{X} + \frac{1}{3}\overline{Y}\right) = \frac{1}{3}E(\overline{X}) + \frac{1}{3}E(\overline{Y}) = \frac{1}{3} \cdot \mu + \frac{1}{3} \cdot 2\mu = \mu$$

   and

$$E(\mu_2^*) = E\left(\frac{1}{5}\overline{X} + \frac{2}{5}\overline{Y}\right) = \frac{1}{5}E(\overline{X}) + \frac{2}{5}E(\overline{Y}) = \frac{1}{5} \cdot \mu + \frac{2}{5} \cdot 2\mu = \mu.$$

   Hence, both are unbiased.

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

   Solution: Observe that

$$V(\overline{X}) = \frac{V(X)}{5} = \frac{10}{5} = 2$$

   and

$$V(\overline{Y}) = \frac{V(Y)}{2} = \frac{1}{2}.$$

   Hence, the variances of the estimators are

$$\begin{aligned} V(\mu_1^*) &= V\left(\frac{1}{3}\overline{X} + \frac{1}{3}\overline{Y}\right) = \left(\frac{1}{3}\right)^2 V(\overline{X}) + \left(\frac{1}{3}\right)^2 V(\overline{Y}) \\ &= \frac{1}{9} \cdot 2 + \frac{1}{9} \cdot \frac{1}{2} = \frac{5}{18} \approx 0.28 \end{aligned}$$

   and

$$\begin{aligned} V(\mu_2^*) &= V\left(\frac{1}{5}\overline{X} + \frac{2}{5}\overline{Y}\right) = \left(\frac{1}{5}\right)^2 V(\overline{X}) + \left(\frac{2}{5}\right)^2 V(\overline{Y}) \\ &= \frac{1}{25} \cdot 2 + \frac{4}{25} \cdot \frac{1}{2} = \frac{4}{25} = 0.16. \end{aligned}$$

   Hence, $V(\mu_2^*) < V(\mu_1^*)$, and so, μ_2^* is most efficient.

3. Burt sells umbrellas at a summer party. The number of umbrellas that he sells until it starts raining, X, is supposed to be a discrete random variable with distribution function

$$F(x) = P(X \leq x) = 1 - (1 - \theta)^x,$$

where $x = 1, 2, ...$ and $0 < \theta < 1$. For this distribution, it holds that $E(X) = 1/\theta$, so the smaller the θ, the bigger we expect X to be.

Burt's friend, the meterologist Börthie, claims that θ is less than 0.1.

At this particular occation, Burt sells 32 umbrellas before it starts to rain.

(a) Is Börthie right? Investigate this by testing a suitable hypothesis. Test at the significance level 5%. (2p)

Solution: It is most natural to put want we want to 'prove' under the alternative hypothesis, so we want to test H_0: $\theta = 0.1$ vs H_1: $\theta < 0.1$. The smaller θ is, the larger does X tend to be, so the critical region should be on the form $x \geq c$, i.e. we reject for large x.

We observe $x = 32$. The direct method gives us the P value

$$\begin{aligned} P(X \geq 32; \theta = 0.1) &= 1 - P(X \leq 31; \theta = 0.1) = 1 - F(31; \theta = 0.1) \\ &= 1 - (1 - 0.9^{31}) = 0.9^{31} = 0.038 < 0.05, \end{aligned}$$

and so, we may reject H_0 at the 5% level. On this level, we have evidence that Börthie is right.

(b) Calculate the critical region for the test. (1p)

Solution: As we have seen, the critical region is on the form $\{x \geq c\}$. We get c as the smallest c such that $P(X \geq c; \theta = 0.1) < 0.05$. By (a), we realize that $c \leq 32$. Testing gives

$$\begin{aligned} P(X \geq 31; \theta = 0.1) &= 1 - F(30; \theta = 0.1) = 0.9^{30} = 0.043 < 0.05, \\ P(X \geq 30; \theta = 0.1) &= 1 - F(29; \theta = 0.1) = 0.9^{29} = 0.047 < 0.05, \\ P(X \geq 29; \theta = 0.1) &= 1 - F(28; \theta = 0.1) = 0.9^{28} = 0.052 > 0.05, \end{aligned}$$

and so, $c = 30$. The critical region is $\{x \geq 30\}$.

(c) In case $\theta = 0.01$, what is the power of your test? (2p)

Solution: The power is the probability to reject H_0 for a certain value of the parameter under the alternative hypothesis. In this case, we have the power

$$P(X \geq 30; \theta = 0.01) = 1 - F(29; \theta = 0.01) = 0.99^{29} = 0.747,$$

i.e. about 75%.

4. On a calm summer day, Lennart and Helga are out fishing. They weigh the fish that they catch and note the weight (in kilograms) before they let them back into the water. They go on until each of them has caught five fish. The results are given in the table below.

   The heavier the fish, the better.

   In terms of this, are Lennart and Helga equally good at fishing? Try to find this out by testing a suitable hypothesis. Be careful to specify all assumptions that you make. (5p)

Lennart's fish	Helga's fish
1.7	2.3
1.6	1.2
0.7	0.8
3.2	0.5
0.2	2.7

Solution: We have two random samples that are not related pairwise. It is reasonable to assume that the weights are normally distrubuted.

So, we have a random sample $x_1, ..., x_5$ from $X \sim N(\mu_1, \sigma_1^2)$ and a random sample $y_1, ..., y_5$ from $Y \sim N(\mu_2, \sigma_2^2)$, where X and Y are independent. The variances σ_1^2 and σ_2^2 are unknown but may be considered equal, i.e.
$\sigma_1^2 = \sigma_2^2 = \sigma^2$, where σ^2 is unknown.

We want to test H_0: $\mu_1 = \mu_2$ vs H_1: $\mu_1 \neq \mu_2$. (There is nothing in the formulation of the problem that indicates that we should use a one-sided test.)
We estimate s^2 by the pooled variance s_p^2, where since $s_x^2 = 1.317$ and
$s_y^2 = 0.915$, we get $(n_1 = n_2 = 5)$

$$s_p^2 = \frac{(n_1 - 1)s_x^2 + (n_2 - 1)s_y^2}{n_1 + n_2 - 2} = \frac{4 \cdot 1.317 + 4 \cdot 0.915}{8} = 1.116.$$

Moreover, $\bar{x} = 1.48$ and $\bar{y} = 1.50$.

The observed test statistic is

$$T_{obs} = \frac{\bar{x} - \bar{y}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{1.48 - 1.50}{\sqrt{1.116}\sqrt{\frac{1}{5} + \frac{1}{5}}} \approx -0.030.$$

Since we have $n_1 - n_2 - 2 = 8$ degrees of freedom and

$$|T_{obs}| = 0.030 < t_{0.025}(8) = 2.3060,$$

we may not reject H_0 at the 5% level.

We have no evidence at the 5% level that they are not equally good.

5. Amy picks 200 strawberries on a field, without taking notice of if they are rotten or not. Let the probability that a strawberry is rotten be p.

   After picking, Amy gives all the strawberries to Sheldon. Sheldon checks them, and finds out that 15 are rotten.

   (a) Calculate a 95% confidence interval for p. (4p)

   Solution: We may assume that there are many strawberries in the field, and that Amy picks the strawberries at random. This means that X, the number of rotten strawberries that Amy picks, is $\text{Bin}(200, p)$.
   We observe $x = 15$, and get the estimate $p^* = 15/200 = 0.075$. The rule of thumb for normal approximation is fulfilled since

$$np^*(1-p^*) = 200 \cdot 0.075 \cdot (1 - 0.075) = 13.875 > 5.$$

   A confidence interval for p with approximate confidence level 95% is given by

$$\begin{aligned} I_p &= p^* \pm \lambda_{0.025}\sqrt{\frac{p^*(1-p^*)}{n}} = 0.075 \pm 1.96\sqrt{\frac{0.075 \cdot (1-0.075)}{200}} \\ &= 0.075 \pm 0.0365 = (0.04,\ 0.11). \end{aligned}$$

   (b) The owner of the field claims that $p = 0.02$. In the light of your result in (a), what do you think about this statement? (1p)

   Solution: We find that $p = 0.02$ is not included in the confidence interval, which means that (at the level 5%) there is evidence that the owner is wrong.

6. We have a random sample of 100 observations from $X \sim N(\mu, \sigma^2)$, where it is known that $\sigma^2 = 4$, while μ is unknown. The mean of the sample is $\bar{x} = 50$.

   (a) Calculate a 99% confidence interval for μ. (2p)

   Solution: We can base the confidence interval on the reference variable

$$\frac{\bar{X} - \mu}{\sigma} \sim N(0, 1),$$

   where $\sigma = 2$. With $\lambda_{0.005} = 2.5758$, this gives us the 99% confidence interval

$$\begin{aligned} I_\mu &= \bar{x} \pm \lambda_{0.005} \frac{\sigma}{\sqrt{100}} = 50 \pm 2.5758 \cdot \frac{2}{10} \\ &= 50 \pm 0.51516 = (49.485,\ 50.515). \end{aligned}$$

   (b) Calculate a 99% confidence interval for $P(X \leq 49.5)$. (3p)

   Solution: Let $p = P(X \leq 49.5)$. We have

$$p = P\left(\frac{X - \mu}{2} \leq \frac{49.5 - \mu}{2}\right) = \Phi\left(\frac{49.5 - \mu}{2}\right),$$

   where Φ is the distribution function of a standard normal random variable. This is a decreasing function of μ. Hence, the interval in (a) for μ corresponds to an interval for p with left endpoint

$$p_l = \Phi\left(\frac{49.5 - 50.515}{2}\right) \approx \Phi(-0.51) = 1 - \Phi(0.51) \approx 1 - 0.70 = 0.30,$$

   and right endpoint

$$p_u = \Phi\left(\frac{49.5 - 49.485}{2}\right) \approx \Phi(0.01) \approx 0.50.$$

   Hence, the desired confidence interval is

$$I_p = (p_l, p_u) = (0.30,\ 0.50).$$

7. A light bulb of a certain brand is supposed to have a lifelength that is exponentially distributed with expectation μ. A random sample of 200 such light bulbs was taken. Their mean lifelength was 3.2 years.

   The manifacturer claims that the average lifelength of a light bulb of this brand is at least 4 years. Perform a hypothesis test to check if the manifacturer is wrong. (5p)

   Solution: We have a random sample $x_1, ..., x_n$, where $n = 200$, from X which is exponential with expectation μ, i.e. parameter β with $\mu = 1/\beta$, with the notation from the book. We then know that $V(X) = 1/\beta^2 = \mu^2$.

   In the manufacturer is wrong, then $\mu < 4$. We want to test H_0: $\mu = 4$ vs H_1: $\mu < 4$, so that we have what we want to 'prove' in the alternative hypothesis.

   Let $\bar{X}$ be the mean of the random variables corresponding to the sample. Then, $E(\bar{X}) = \mu$ and $V(\bar{X}) = \mu^2/n$. By the central limit theorem, we have that

   $$T = \frac{\bar{X} - \mu}{\sqrt{\mu^2/n}} \approx N(0, 1).$$

   We want to reject H_0 for negative values of T far from zero. We may insert the H_0 value $\mu = 4$ and the observed $\bar{x} = 3.2$ to get

   $$T_{obs} = \frac{3.2 - 4}{\sqrt{4^2/200}} \approx -2.83 < -\lambda_{0.01} = -2.3263.$$

   Hence, we can reject H_0 at the 1% level.

   At this level, we have evidence that the manifacturer is wrong.

8. Data is available for selling of 20 houses in Milwaukee, Wisconsin. The price in thousands of dollars is given by y, and the living space (i hundreds of square feet) is given by x. The data is plotted in the figure below.

   The object is to estimate the regression model

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

   for $i = 1, 2, ..., 20$, where the ε_i are independent $N(0, \sigma^2)$.

   (a) By looking at the plot, try to judge which pair of estimated β (β^*) and coefficient of determination R^2 that we have for our data, out of the following alternatives. Motivate your answer. (3p)

      i. $\beta^* = 2.68$, $R^2 = 2\%$.

      ii. $\beta^* = 1.03$, $R^2 = 75\%$.

      iii. $\beta^* = 2.75$, $R^2 = 83\%$.

      iv. $\beta^* = 1.12$, $R^2 = 3\%$.

   Solution: From the plot, it is seen that the observations are approximately on a line that goes from about $(x, y) = (14, 70)$ to $(x, y) = (26, 100)$, giving a slope of about

$$\frac{100 - 70}{26 - 14} = 2.5.$$

   Of course, this is a very crude estimate, but it seems to outrule alternatives ii and iv.

   We are left with alternatives i and iii. Comparing R^2 values, we note that R^2 is extremely small for alternative i, while for alternative iii, it indicates a decent model fit. In the plot, the agreement with a line is also fairly decent, which leads us to believe in alternative iii.

   So, the answer is alternative iii.

   (b) The observed means are $\bar{x} = 16.22$ and $\bar{y} = 76.55$.

   What is the estimated α? (2p)

   Solution: We have the formula

$$\alpha^* = \bar{y} - \beta^* \bar{x},$$

   which, in our case (for alternative iii), gives us

$$\alpha^* = 76.55 - 2.75 \cdot 16.22 = 31.94.$$

Appendix: figures

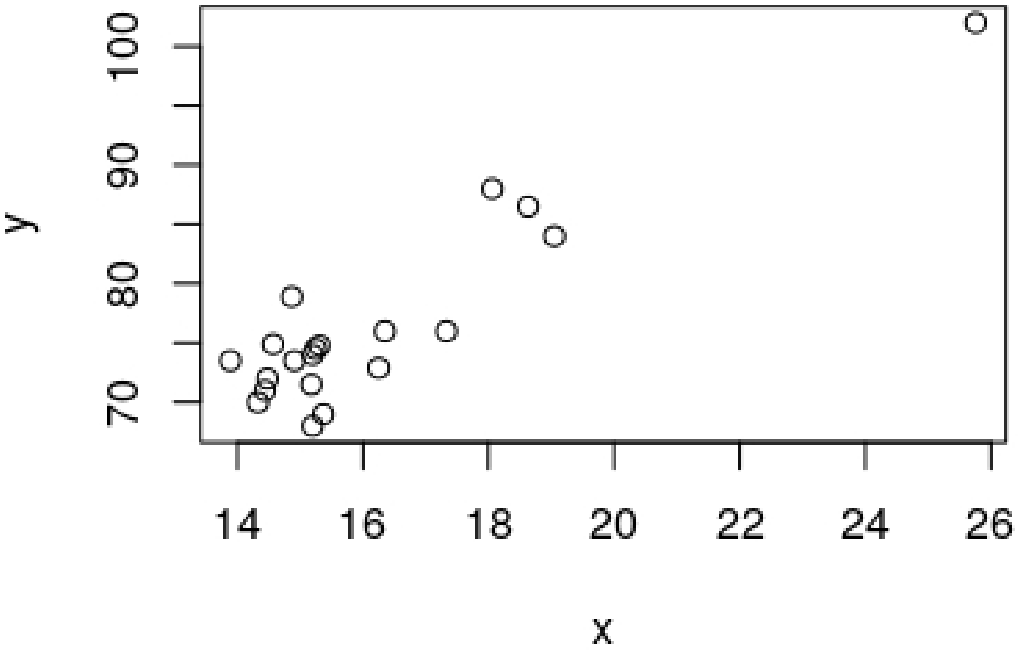

Figure 1: Plot for problem 8.