UPPSALA UNIVERSITY
Department of Mathematics
Rolf Larsson

Exam in Mathematical Statistics
Inference 1 1MS035
2022–06–14
Solutions

1. Let X be a discrete random variable with probability function

$$p_X(x) = \begin{cases} 4\theta^2 & \text{if } x = 1, \\ 4\theta(1-2\theta) & \text{if } x = 2, \\ (1-2\theta)^2 & \text{if } x = 3, \\ 0 & \text{otherwise,} \end{cases}$$

where $0 \leq \theta \leq 1/2$.

We have a random sample $x_1 = 2$, $x_2 = 2$, $x_3 = 3$, $x_4 = 3$ from X.

(a) Find the moment estimate of θ. (1p)

Solution: At first, we calculate the expectation:

$$\begin{aligned} E(X) &= \sum_x x p_X(x) = 1 \cdot p_X(1) + 2 \cdot p_X(2) + 3 \cdot p_X(3) \\ &= 4\theta^2 + 2 \cdot 4\theta(1-2\theta) + 3 \cdot (1-2\theta)^2 = 3 - 4\theta. \end{aligned}$$

The moment estimate θ solves $\bar{x} = m(\theta)$ where $m(\theta) = E(X)$. We have $\bar{x} = 2.5$, and solving

$$2.5 = 3 - 4\theta,$$

gives us the moment estimate $\theta = \theta^* = 1/8 = 0.125$.

(b) Find the maximum likelihood estimate of θ. (4p)

Solution: In the discrete case, the likelihood equals the probability to obtain the sample, which in this case is

$$L(\theta) = p_X(2)^2 p_X(3)^2 = \{4\theta(1-2\theta)\}^2 \{(1-2\theta)^2\}^2 = 16\theta^2(1-2\theta)^6.$$

It is equivalent (and easier) to maximize

$$l(\theta) = \ln\{L(\theta)\} = \ln(16) + 2\ln(\theta) + 6\ln(1-2\theta),$$

which has the first two derivatives

$$\begin{aligned} l'(\theta) &= \frac{2}{\theta} - \frac{12}{1-2\theta}, \\ l''(\theta) &= -\frac{2}{\theta^2} - \frac{24}{(1-2\theta)^2}. \end{aligned}$$

We always have $l''(\theta) < 0$, which means that we get a maximum by solving $l'(\theta) = 0$, i.e. $1 - 2\theta = 6\theta$, which gives $\theta = 1/8$.
Hence, the ML estimate is $\theta^* = 1/8 = 0.125$, which is equal to the moment estimate in this case.

2. We have a random sample x_1, x_2, x_3 from the random variable X which has expectation $\mu+m$ and variance 1, a random sample y_1, y_2, y_3, y_4 from the random variable Y which has expectation m and variance 1, and one observation z of the random variable Z which has expectation μ and variance 2. The means of the first two samples are denoted by $\bar{x}$ and $\bar{y}$, respectively. We may assume that X, Y, Z are simultaneously independent.

   Two estimates of μ are proposed:

$$\mu_1^* = \frac{\bar{x} - \bar{y} + z}{2}, \quad \mu_2^* = 2(\bar{x} - \bar{y}) - z.$$

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

   Solution: We calculate the expectations of the corresponding estimators, using that $E(\bar{X}) = \mu + m$, $E(\bar{Y}) = m$ and $E(Z) = \mu$ to obtain

$$E(\mu_1^*) = \frac{1}{2}E(\bar{X}) - \frac{1}{2}E(\bar{Y}) + \frac{1}{2}E(Z) = \frac{1}{2}(\mu + m) - \frac{1}{2}m + \frac{1}{2}\mu = \mu$$

   and

$$E(\mu_2^*) = 2E(\bar{X}) - 2E(\bar{Y}) - E(Z) = 2(\mu + m) - 2m - \mu = \mu,$$

   showing that they are both unbiased.

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

   Solution: We check which estimator has the smallest variance. Using $V(\bar{X}) = 1/3$, $V(\bar{Y}) = 1/4$, $V(Z) = 2$ and independence, we obtain

$$\begin{aligned} V(\mu_1^*) &= \left(\frac{1}{2}\right)^2 V(\bar{X}) + \left(\frac{1}{2}\right)^2 V(\bar{Y}) + \left(\frac{1}{2}\right)^2 V(Z) \\ &= \frac{1}{4} \cdot \frac{1}{3} + \frac{1}{4} \cdot \frac{1}{4} + \frac{1}{4} \cdot 2 = \frac{31}{48} \approx 0.65 \end{aligned}$$

   and

$$\begin{aligned} V(\mu_2^*) &= 2^2 V(\bar{X}) + 2^2 V(\bar{Y}) + V(Z) \\ &= 4 \cdot \frac{1}{3} + 4 \cdot \frac{1}{4} + 2 = \frac{13}{3} \approx 4.33. \end{aligned}$$

   Hence, $V(\mu_1^*) < V(\mu_2^*)$, which means that μ_1^* is more efficient that μ_2^*.

3. The number of participants at the Mid summer party in the distant village Gråboda in the heart of the county of Småland in Sweden is assumed to be Poisson distributed with parameter (expectation) λ. The main organizer, Börje (who does not count as a participant) believes that $\lambda = 2$. However, his daughter Börthie (who does not count as a participant either) is more optimistic, and claims that λ must be greater than 2.

   (a) It turns out that there are five participants at the party. Is there any evidence that Börthie is right? Try to find out by testing a suitable hypothesis, on the 5% level. (2p)

   Solution: We want to test H_0: $\lambda = 2$ vs H_1: $\lambda > 2$. Rejecting H_0 gives evidence that Börthie is right.

   We use the direct method. The number of participants $X \sim \text{Po}(\lambda)$, and we observe $x = 5$. We reject for large x (the bigger the λ, the bigger the expectation). The P value is given as (e.g. from table 3)

   $$P(X \geq 5; \lambda = 2) = 1 - P(X \leq 4; \lambda = 2) \approx 1 - 0.9473 = 0.0527.$$

   Testing at the 5% level, we find that we may not reject H_0, because the P value is $0.0527 > 0.05$.

   At this risk level, we have no evidence that Börthie is right.

   (b) Calculate the power of the test in (a) in case $\lambda = 6$. (3p)

   Solution: At first, we calculate the critical region, i.e. we find the smallest x such that $P(X \geq x; \lambda = 2) < 0.05$. We saw in (a) that $x = 5$ is not good enough. But from table 3,

   $$P(X \geq 6; \lambda = 2) = 1 - P(X \leq 5; \lambda = 2) \approx 1 - 0.9834 = 0.0166 < 0.05,$$

   and so, the critical region is given by $C = \{x \geq 6\}$.

   Thus, the power for $\lambda = 6$ is obtained as

   $$P(X \geq 6; \lambda = 6) = 1 - P(X \leq 5; \lambda = 6) \approx 1 - 0.4457 = 0.5543,$$

   i.e. about 55%.

4. Eight randomly selected overweight Lagotto dogs are presented with a new diet, to see if this diet makes them lose weight. The table below gives their weights in kilos before and after one week with the diet food.

   Test a suitable hypothesis to try to conclude if the diet works or not. (5p)

Dog no.	1	2	3	4	5	6	7	8
Weight before	20.2	18.3	17.3	21.6	15.1	19.3	19.4	17.6
Weight after	18.3	18.4	14.7	19.8	16.5	17.2	18.2	17.4

   Solution: This is a paired sample. Let Z be the lost weight. We may assume that $Z \sim N(\mu, \sigma^2)$. If the diet works, then $\mu > 0$, so we want to test H_0: $\mu = 0$ vs H_1: $\mu > 0$. We can use the t test.

   The observed differences (z_i) are

$$1.9,\ -0.1,\ 2.6,\ 1.8,\ -1.4,\ 2.1,\ 1.2,\ 0.2,$$

   from which we calculate the sample mean $\bar{z} = 1.0375$ and the sample variance $s^2 \approx 1.8370$. We get the observed statistic (sample size $n = 8$, number of degrees of freedom $n - 1 = 7$)

$$t_{obs} = \frac{\bar{z}}{\sqrt{s^2/n}} = \frac{1.0375}{\sqrt{1.8370/8}} \approx 2.17 > t_{0.05}(7) \approx 1.89,$$

   and so, we can reject H_0 at the 5% level.

   On this risk level, we have evidence that the diet works.

   An alternative is the sign test. Let the number of positive differences be $U \sim \text{Bin}(8, p)$. We want to test H_0: $p = 1/2$ vs H_1: $p > 1/2$, and we observe $u = 6$. This gives us the P value (e.g. from table 2)

$$P(U \geq 6) = 1 - P(U \leq 5) \approx 1 - 0.8555 = 0.1444 > 0.05,$$

   so with this test, we can not reject H_0 at the 5% level.

   (It is no surprise that the sign test does not reject, since it uses much less information than the t test.)

   Another alternative is the signed rank test, which gives an observed rank sum for possitive differences as 31, and this is just significant at the 5% level.

5. Zlatan and Tony practice football penalty shoots on the same goal keeper, Hedvig. Among 40 shots each, Zlatan scores on 32 of them, and Tony scores on 24. Say that the probability of scoring on a penalty shot is p_1 for Zlatan and p_2 for Tony.

   (a) Calculate a 95% confidence interval for $p_1 - p_2$. (4p)

   Solution: We can view successive penalty shots as independent with the same probability of succes (goal), and this means that the number of goals is Binomially distributed. Let the number of goals by Zlatan and Tony be $X \sim \text{Bin}(40, p_1)$ and $Y \sim \text{Bin}(40, p_2)$, respectively. We have the estimates $p_1^* = x/40 = 32/40 = 0.8$ and $p_2^* = y/40 = 24/40 = 0.6$, the corresponing estimators being $X/40$ and $Y/40$. Our reference variable is

$$T = \frac{(p_1^* - p_2^*) - (p_1 - p_2)}{d},$$

   where we have the standard error

$$d = \sqrt{\frac{p_1^*(1-p_1^*)}{40} + \frac{p_2^*(1-p_2^*)}{40}}.$$

   The rule of thumb for normal approximation is fulfilled, since

$$40p_1^*(1-p_1^*) = 6.4, \quad 40p_2^*(1-p_2^*) = 9.6,$$

   which are both greater than 5. Hence, we may use that $T \approx N(0,1)$. We observe

$$d = \sqrt{\frac{0.8 \cdot 0.2}{40} + \frac{0.6 \cdot 0.4}{40}} = 0.1,$$

   which gives us the 95% confidence interval

$$I_{p_1-p_2} = p_1^* - p_2^* \pm \lambda_{0.025} d = 0.8 - 0.6 \pm 1.96 \cdot 0.1 = (0.004,\ 0.396).$$

   (b) Are Zlatan and Tony equally good at penalty shots? Try to deduce this from the result in (a). What is your conclusion? (1p)

   Solution: We test H_0: $p_1 - p_2 = 0$ vs H_1: $p_1 \neq p_2$ at the 5% level. Because 0 is not contained in the confidence interval in (a), we reject H_0.
   At this risk level, we have evidence that they are not equally good.

6. The Spanish company El Giant tests the life lengths of an electronic component. Its life length X is assumed to follow an exponential distribution with expectation μ.

   A random sample of 200 components are tested, and their mean life length is 122.3 days.

   (a) Calculate a 99% confidence interval for μ. (3p)

   Solution: Say that the random sample is $x_1, ..., x_n$ where $n = 200$. Since $E(X) = \mu$, it is natural to choose $\bar{X}$ as estimator of μ, and we get the estimate $\mu^* = \bar{x} = 122.3$. To calculate the confidence interval, because $n = 200$ is large we can use the central limit theorem, which gives that

$$\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right),$$

   where $\sigma^2 = V(X)$. From the exponential distribution, we know that $V(X) = 1/\beta^2 = \mu^2$, where $\mu = 1/\beta$. Hence, we may estimate σ^2 by $(\mu^*)^2 = \bar{x}^2$. This gives us the 99% confidence interval

$$\begin{aligned} I_\mu &= \bar{x} \pm \lambda_{0.005}\sqrt{\frac{\bar{x}^2}{n}} = 122.3 \pm 2.5758\sqrt{\frac{122.3^2}{200}} = 122.3 \pm 22.3 \\ &= (100.0,\ 144.6). \end{aligned}$$

   Alternatively, we can do as in the book, p.347, to get the confidence interval

$$\left(\frac{\bar{x}}{1 + \frac{2.5758}{\sqrt{200}}}, \frac{\bar{x}}{1 - \frac{2.5758}{\sqrt{200}}}\right) = (103.5,\ 149.5).$$

   (b) Calculate a 99% confidence interval for the intensity parameter $\beta = 1/\mu$. (2p)

   Solution: The interval $100.0 \leq \mu = \frac{1}{\beta} \leq 144.6$ corresponds to

$$0.0069 \approx \frac{1}{144.6} \leq \beta \leq \frac{1}{100.0} = 0.0100.$$

   Hence, a 99% confidence interval for μ is $(0.0069,\ 0.0100)$.

   From the alternative interval in (a), we similarly obtain $(0.0067,\ 0.0097)$.

7. In the table below, 313 female students are classified according to color of hair (blond or not) and color of eyes (blue or not). Are colors of hair and eyes independent for female students? Perform a suitable hypothesis test to find out the answer. (5p)

	Blue eye color	Other eye color
Blond hair color	64	17
Other hair color	50	182

Solution: We want to test H_0: independence vs H_1: dependence. We may use the independence/homogeneity χ^2 test. We only need to check the rule of thumb. The row sums are 81 and 232, and the column sums are 114 and 199. The expected counts under H_0 for cells (i, j), $i, j = 1, 2$, are

$$e_{11} = \frac{81 \cdot 114}{313} \approx 29.50, \quad e_{12} = \frac{81 \cdot 199}{313} \approx 51.50,$$
$$e_{21} = \frac{232 \cdot 114}{313} \approx 84.50, \quad e_{22} = \frac{232 \cdot 199}{313} \approx 147.50,$$

which are all greater than 5, hence χ^2 approximaton is permitted. The number of degrees of freedom is $(2-1)(2-1) = 1$. We get the observed test statistic

$$Q = \frac{(64-29.5)^2}{29.5} + \frac{(17-51.5)^2}{51.5} + \frac{(50-84.5)^2}{84.5} + \frac{(182-147.5)^2}{147.5}$$
$$\approx 85.6 > \chi_{0.001}(1) \approx 10.828,$$

hence we can reject H_0 at the 1% level (and also at much lower levels).

We have strong evidence of dependence between hair color and eye color.

8. At mid summers eve, 8 randomly selected children from the village Abborrviken and 8 other randomly selected children from the village Havsbyn were asked how many times they have gone for a swim outside this summer. The data is given in the table below. Test a suitable hypothesis to find out if children from Abborrviken go swimming outside as often as children from Havsbyn before mid summer.

   It is not allowed to assume that data comes from the normal distribution. (5p)

Abborrviken	4	15	7	13	9	11	12	27
Havsbyn	6	0	1	8	3	10	2	14

   Solution: This is two independent samples, and we can not assume normality. We test H_0: they go swimming equally often vs H_1: $\neg H_0$. We use the Wilcoxon rank test.

   Merging the two samples and ranking the observations, we observe the rank sum for 'Havsbyn' as

$$R = 1 + 2 + 3 + 4 + 6 + 8 + 10 + 14 = 48.$$

   If we test on the 5% level, we need to check both one-sided 2.5% test critical regions. From table 9, these are $R \leq 49$ and $R \geq 87$. Hence, our observation belongs to the critical region ($48 < 49$), and so, we can reject H_0 at the 5% level.

   At this risk level, we have evidence that they do not go swimming equally often.

   Alternatively, because the sample sizes are $8 \geq 7$, we may use normal approximation. We have

$$E(R) = \frac{8 \cdot 17}{2} = 68, \quad V(R) = \frac{8^2 \cdot 17}{12},$$

   and so, we get

$$T_{obs} = \frac{R - E(R)}{\sqrt{V(R)}} = \frac{48 - 68}{\sqrt{8^2 \cdot 17/12}} \approx -2.10 < -1.96 = -\lambda_{0.025},$$

   which again leads us to reject H_0 at the 5% level.