UPPSALA UNIVERSITY
Department of Mathematics
Rolf Larsson

Exam in Mathematical Statistics
Inference 1 1MS035
2023–08–23
Solutions

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

$$p_X(x) = \begin{cases} 9\theta^2 & \text{if } x = 2, \\ 6\theta(1-2\theta) & \text{if } x = 4, \\ (1-3\theta)^2 & \text{if } x = 6, \\ 0 & \text{otherwise,} \end{cases}$$

where $0 \leq \theta \leq 1/3$.

We have a random sample $x_1 = 4$, $x_2 = 2$, $x_3 = 6$ from X.

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

Solution: The expectation is

$$E(X) = \sum_x x p_X(x) = 2 \cdot 9\theta^2 + 4 \cdot 6\theta(1-3\theta) + 6 \cdot (1-3\theta)^2 = 6 - 12\theta.$$

The moment estimate is the θ that solves $\bar{x} = m(\theta)$, where $m(\theta) = E(X)$. We have $\bar{x} = 4$, and we need to solve

$$4 = 6 - 12\theta,$$

which gives the moment estimate $\theta = \theta^* = 1/6 \approx 0.167$.

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

Solution: In the discrete case, the likelihood is the probability to get the observed sample, which in our case is

$$L(\theta) = p_X(4)p_X(2)p_X(6) = 6\theta(1-3\theta) \cdot 9\theta^2 \cdot (1-3\theta)^2 = 54\theta^3(1-3\theta)^3.$$

For maximization, it is equivalent (and most often easier) to maximize

$$l(\theta) = \ln\{L(\theta)\} = \ln(54) + 3\ln(\theta) + 3\ln(1-3\theta),$$

which has the first two derivatives

$$l'(\theta) = \frac{3}{\theta} - \frac{9}{1-3\theta},$$
$$l''(\theta) = -\frac{3}{\theta^2} - \frac{27}{(1-3\theta)^2}.$$

We note that we always have $l''(\theta) < 0$, i.e. we get a maximum by solving $l'(\theta) = 0$, i.e. $3 - 9\theta = 9\theta$, i.e. $\theta = 1/6$.

Hence, we get the ML estimate $\theta^* = 1/6 \approx 0.167$, in this case the same as the moment estimate.

2. We have a random sample x_1, x_2, x_3, x_4 from the random variable X which has expectation 2μ and variance 2, and a random sample y_1, y_2 from the random variable Y which has expectation μ and variance 1. 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:

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

   (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 $E(\bar{X}) = \mu$ and $E(\bar{Y}) = 2\mu$, to obtain

$$E(\mu_1^*) = E(\bar{X}) - E(\bar{Y}) = 2\mu - \mu = \mu$$

   and

$$E(\mu_2^*) = \frac{1}{3}\{E(\bar{X}) + E(\bar{Y})\} = \frac{1}{3}(2\mu + \mu) = \mu,$$

   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 $V(\bar{X}) = V(X)/4 = 2/4 = 1/2$, $V(\bar{Y}) = V(Y)/2 = 1/2$, and independence, we obtain (observe: $V(aX + bY) = a^2V(X) + b^2V(Y)$)

$$V(\mu_1^*) = V(\bar{X} - \bar{Y}) = V(\bar{X}) + V(\bar{Y}) = \frac{1}{2} + \frac{1}{2} = 1$$

   and

$$\begin{aligned} V(\mu_2^*) &= V\left(\frac{1}{3}\bar{X} + \frac{1}{3}\bar{Y}\right) = \frac{1}{9}V(\bar{X}) + \frac{1}{9}V(\bar{Y}) \\ &= \frac{1}{9} \cdot \frac{1}{2} + \frac{1}{9} \cdot \frac{1}{2} = \frac{1}{9}. \end{aligned}$$

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

3. Following the so called "Ling's law", the time in days (not necessarily an integer) after the last of June until it is possible to pick lingonberries in the forest may, to a good approximation, be described by a random variable X with cumulative distribution function

$$F(x) = 1 - \exp\left(-\frac{\theta x^2}{10000}\right),$$

where $x > 0$, and otherwise, $F(x) = 0$. The parameter θ is unknown.

(a) Linga believes that $\theta = 9$. Her friend Lingberth thinks that $\theta < 9$. One year, it is possible to pick lingonberries in the forest 60 days after the last of June. Is Lingberth right? Try to find out by testing a suitable hypothesis. Use the significance level 5%. (2p)

Hint: A smaller θ corresponds to a larger X.

Solution: Because we want to find evidence for what Lingberth says, it is natural to test H_0: $\theta = 9$ vs the alternative H_1: $\theta < 9$. By the hint, we reject for large values. Hence, calculating under H_0, we get the P value

$$P(X \geq 60) = 1 - F(60) = \exp\left(-\frac{9 \cdot 60^2}{10000}\right) \approx 0.0392.$$

Since $0.0392 < 0.05$, we may reject H_0 at the 5% level.
At this level, we have found evidence that Lingberth is right.

(b) What is the power of the test in (a) for $\theta = 2$? (3p)

Solution: The critical region is on the form $C = \{x > K\}$, where $0.05 = P(X > K)$ calculated under H_0. To find K, we need to solve

$$0.05 = P(X > K) = 1 - F(K) = \exp\left(-\frac{9K^2}{10000}\right),$$

i.e. $K = 100\sqrt{-\ln(0.05)/9} \approx 57.7$.
The power that we want to calculate is given by the probability that $X \in C$ given that $\theta = 2$, i.e.

$$\exp\left(-\frac{2 \cdot 57.7^2}{10000}\right) \approx 0.51 = 51\%.$$

4. Ellen prepares for a big party for vegans. At the supermarket CIA, she buys 10 cauliflowers (Swedish: blomkål). This sample of cauliflowers has average weight 520 grams and standard deviation 46.5.

   Her friend Lena wants to help out. She goes to the supermarket POOC and buys 8 cauliflowers. This sample of cauliflowers has average weight 560 grams and standard deviation 53.2.

   We may assume that the weight of a cauliflower is normally distributed. Let the expected weight of a cauliflower from CAI be μ_1, and correspondingly, the expected weight of a cauliflower from POOC is μ_2.

   (a) Calculate a 95% confidence interval for $\mu_1 - \mu_2$. Be careful to state all your assumptions. (4p)

   Solution: Assume that we have one random sample (CAI) $x_1, ..., x_{10}$ from $X \sim N(\mu_1, \sigma_1^2)$ and one random sample (POOC) $y_1, ..., y_8$ from $Y \sim N(\mu_2, \sigma_2^2)$. The parameters σ_1^2 and σ_2^2 are unknown, but may be assumed to be equal, i.e. $\sigma_1^2 = \sigma_2^2 = \sigma^2$ for some unknown σ^2. We estimate σ_1^2 and σ_2^2 by $s_x^2 = 46.5^2$ and $s_y^2 = 53.2^2$, respectively.
   This gives us the estimated σ^2 (the pooled variance estimate) as

$$s_p^2 = \frac{9 \cdot 46.5^2 + 7 \cdot 53.2^2}{16} \approx 2454.5.$$

   Now, we have the reference variable

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{10} + \frac{1}{8}}} \sim t(16),$$

   which gives us the 95% confidence interval ($\bar{x} = 520$, $\bar{y} = 560$)

$$\begin{aligned} I_{\mu_1-\mu_2} &= \bar{x} - \bar{y} \pm t_{0.025}(16) s_p \sqrt{\frac{1}{10} + \frac{1}{8}} \\ &= 520 - 560 \pm 2.1199\sqrt{2454.5}\sqrt{\frac{1}{10} + \frac{1}{8}} \\ &= -40 \pm 49.8 = (-89.8,\ 9.8). \end{aligned}$$

   (b) Is there evidence that the expected weights of cauliflowers from CAI are different from the expected weights of cauliflowers from POOC? (1p)

   Solution: No, because 0 is included in the confidence interval for $\mu_1 - \mu_2$. This means that H_0: $\mu_1 = \mu_2$ is not rejected vs H_1: $\mu_1 \neq \mu_2$ at the 5% level.

5. In the male elite division of bandy in Sweden 2021-2022, 2158 goals were scored in 240 matches. Assume that the number of goals in a match is Poisson distributed, and that the numbers of goals in different matches are independent and identically distributed.

   (a) Calculate a 99% confidence interval for the expected number of goals in a random match. (3p)

   Solution: Let the number of goals in a match be $X \sim \text{Po}(\mu)$. We have a random sample $x_1, ..., x_n$ from X, where $n = 240$, with observed sum $\sum_{i=1}^{240} x_i = 2158$. We estimate μ by $\mu^* = \bar{x} = 2158/240 \approx 8.9917$.
   Because $n\mu^* = 2158 > 15$, we may use normal approximation. We have the reference variable

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

   from which we get the 99% confidence interval

$$\begin{aligned} I_\mu &= \bar{x} \pm \lambda_{0.005}\sqrt{\frac{\mu^*}{n}} = 8.9917 \pm 2.5758\sqrt{\frac{8.9917}{240}} \\ &= 8.9917 \pm 0.4986 = (8.4931,\ 9.4903). \end{aligned}$$

   (b) Calculate a 99% confidence interval for the expected time in minutes between two goals. (Assume that every match takes exactly 90 minutes.) (2p)

   Solution: From probability theory, it is known that if the number of goals in a match is $\text{Po}(\mu)$, then the time between goals is exponentially distributed with expectation $1/\mu$. So the expected number of matches between two goals is $1/\mu$, and if every match lasts 90 minutes, the expected number of minutes between two goals is $\theta = 90/\mu$. We want a 99% confidence interval for θ.
   To this end, we can just transform the interval in (a),

$$8.4931 \leq \mu \leq 9.4903,$$

   into the interval

$$9.5 \approx \frac{90}{9.4903} \leq \frac{90}{\mu} = \theta \leq \frac{90}{8.4931} \approx 10.6,$$

   so we get the confidence interval $(9.5,\ 10.6)$.

6. A machine chops boards (Swedish: brädor) in pieces of θ centimeters (not necessarily integer). After each board has been chopped, there is a remaining piece of board with a length that can be assumed to be uniformly distributed on the interval $(0, \theta)$.

   The number θ is unknown. One day, there were 120 remaining pieces with mean length 37.4 centimeters.

   (a) Estimate θ by using the method of moments. (1p)

   Solution: Let the length of a remaining piece of board be $X \sim \text{Re}(0, \theta)$. We have $E(X) = \theta/2$, and we observe a random sample $x_1, ..., x_n$ with $n = 120$ and $\bar{x} = 37.4$.

   The moment estimate solves $\theta/2 = \bar{x}$, so we get

   $$\theta^* = 2\bar{x} = 2 \cdot 37.4 = 74.8.$$

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

   Solution: We know that $V(X) = \theta/12$. Hence, $V(\bar{X}) = \theta/(12n)$, and the variance of our estimate is

   $$D^2 = V(2\bar{X}) = 4V(\bar{X}) = \frac{\theta}{3n} = \frac{\theta}{360}.$$

   Our reference variable is (we have many observations, so we can use normal approximation)

   $$\frac{2\bar{X} - \theta}{d} \approx N(0, 1),$$

   with standard error

   $$d = \sqrt{\frac{\theta^*}{360}} \approx 0.4558,$$

   and we get the 95% confidence interval

   $$I_\theta = 2\bar{x} \pm \lambda_{0.025} d = 74.8 \pm 1.96 \cdot 0.4558 = 74.8 \pm 0.9 = (73.9,\ 75.7).$$

7. The ice cream company SAI has the following slogan:
Our ice cream is the best, pretty tasteless is the rest!

During the summer, a random sample of 200 people tried chocolate ice cream of two brands, BG and SAI. Out of these 200 people, 112 said that the SAI chocolate ice cream tasted better than the BG chocolate ice cream, and the remaining 88 said the opposite (BG tasted better than SAI).

Does chocolate ice cream from SAI taste better than chocolate ice cream from BG? Perform a suitable hypothesis test to find out. (5p)

Solution: Because of the question, a one-sided test is most appropriate. We test H_0: they taste equally well vs H_1: chocolate ice cream from SAI tastes better.

We may use the sign test. Let X be the number of people that prefer SAI. Then, $X \sim \text{Bin}(n, p)$ with $n = 200$, and we want to test H_0: $p = 1/2$ vs H_1: $p > 1/2$. We observe $x = 112$. The rule of thumb holds, because under H_0, $np(1-p) = 200 \cdot (1/2)^2 = 50 > 5$.

We have the estimate X/n and the test statistic

$$T = \frac{X/n - 1/2}{\sqrt{(1/2)^2/n}} = 2\sqrt{n}\left(\frac{X}{n} - \frac{1}{2}\right) \sim N(0,1),$$

under H_0. We observe

$$T_{obs} = 2\sqrt{200}\left(\frac{112}{200} - \frac{1}{2}\right) \approx 1.697 > 1.6449 = \lambda_{0.05},$$

which means that we can reject H_0 at the 5% level.

At this level, we have evidence that chocolate ice cream from SAI tastes better than chocolate ice cream from BG.

8. After summer, a school class meets to compare their salaries in SKr per hour for their respective summer jobs. A random selection of four boys and four girls recieved hourly salaries according to the table below.

   Is there evidence that boys and girls recieve different hourly salaries at summer jobs? Perform a suitable hypothesis test to find out.

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

Girls	102	97	98	130
Boys	100	90	75	92

Solution: In the text, there is no hint on the direction of the alternative, so the appropriate thing is to test H_0: equal salaries for boys and girls vs H_1: $\neg H_0$. We use the Wilcoxon two sample test.

By ranking all the salaries, for boys we observe the rank sum

$$R_{obs} = 6 + 2 + 1 + 3 = 12.$$

For a 5% test, we compare this to the one-sided 2.5% critical regions in table 9 for $n_1 = n_2 = 4$, and this yields $R \leq 10$ and $R \geq 26$. Since R_{obs} is not contained in any of these regions, we may not reject H_0 at the 5% level.

At this level, we have no evidence that the salaries are not equal.

The P value may also be calculated directly, by listing all cases that are as most as extreme as the observed one. These are the cases with observed rank sums $1+2+3+4 = 10$, $1+2+3+5 = 11$ and $1+2+3+6 = 1+2+4+5 = 12$, i.e. there are four of them. Hence, the P value is (multiply by two since the test is two-sided)

$$2 \cdot \frac{4}{\binom{8}{4}} = \frac{8}{70} = \frac{4}{35} \approx 0.11.$$

Hence, because $0.11 > 0.05$, we may not reject H_0 at the 5% level (and not at the 10% level either).