UPPSALA UNIVERSITY
Department of Mathematics
Rolf Larsson

Exam in Mathematical Statistics
Inference 1 1MS035
2023–01–03
Solutions

1. We have a random sample 0, 1, 1, 2, 0, 0, 1, 2 from a discrete random variable X with probability function

$$p_X(k) = \begin{cases} 2(1-\theta)/3, & k = 0, \\ (1+\theta)/3, & k = 1, \\ \theta/3, & k = 2, \\ 0, & \text{otherwise}, \end{cases}$$

where $0 < \theta < 1$. Estimate θ using

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

Solution: The expectation is

$$m(\theta) = E(X) = \sum_k k p_X(k) = 0 \cdot \frac{2(1-\theta)}{3} + 1 \cdot \frac{1+\theta}{3} + 2 \cdot \frac{\theta}{3} = \frac{1}{3} + \theta.$$

The sample mean is $\bar{x} = 7/8$, and the moment estimate solves

$$\frac{7}{8} = \bar{x} = m(\theta) = \frac{1}{3} + \theta,$$

and we get the estimate

$$\theta^* = \frac{7}{8} - \frac{1}{3} = \frac{13}{24} \approx 0.54.$$

(b) the least squares method, (1p)

Solution: Since $m'(\theta) = 1 \neq 0$, by 'anmärkning 7.11' in Alm-Britton, p.286, we have that the least squares estimate equals the moment estimate in this case, i.e. it is $\theta^* = 13/24 \approx 0.54$.

(c) maximum likelihood. (3p)

Solution: We have three observed zeros, three observed ones and two observed twos, which gives the likelihood as the probability of the observed sample:

$$L(\theta) = \left\{\frac{2(1-\theta)}{3}\right\}^3 \left(\frac{1+\theta}{3}\right)^3 \left(\frac{\theta}{3}\right)^2 = C(1-\theta)^3(1+\theta)^3\theta^2,$$

where C is a constant. Maximizing $L(\theta)$ is equivalent to maximizing the log likelihood

$$l(\theta) = \ln\{L(\theta)\} = \ln C + 3\ln(1-\theta) + 3\ln(1+\theta) + 2\ln\theta.$$

We get the derivatives

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

Since $l''(\theta) < 0$ for all $\theta \in (0,1)$, we obtain a maximum by solving $0 = l'(\theta)$, which gives us

$$0 = -3\theta(1+\theta) + 3\theta(1-\theta) + 2(1-\theta)(1+\theta) = 2 - 8\theta^2.$$

The solutions are

$$\theta = \pm\sqrt{\frac{1}{4}} = \pm\frac{1}{2},$$

and the only one belonging to the interval $(0,1)$ is the maximum likelihood estimate

$$\theta^* = \frac{1}{2} = 0.5.$$

2. We have a random sample x_1, x_2, x_3, x_4, x_5 of the random variable X which has expectation $\mu + m$ and variance 2, a random sample y_1, y_2 of the random variable Y which has expectation $2m - \mu$ and variance 1, and a random sample z_1, z_2 of the random variable Z which has expectation m and variance 1. The means of the samples are denoted by $\bar{x}$, $\bar{y}$ and $\bar{z}$, respectively. We may assume that X, Y, Z are simultaneously independent.

   Two estimates of μ are proposed:

$$\mu_1^* = \frac{\bar{x} - \bar{y} + \bar{z}}{2}, \quad \mu_2^* = \frac{5\bar{x} - 4\bar{y} + 3\bar{z}}{9}.$$

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

   Solution: The corresponding estimators are

$$\mu_1^* = \frac{1}{2}\bar{X} - \frac{1}{2}\bar{Y} + \frac{1}{2}\bar{Z}, \quad \mu_2^* = \frac{5}{9}\bar{X} - \frac{4}{9}\bar{Y} + \frac{3}{9}\bar{Z},$$

   where $\bar{X}$, $\bar{Y}$ and $\bar{Z}$ are the random variables that correspond to $\bar{x}$, $\bar{y}$ and $\bar{z}$.

   Because $E(\bar{X}) = E(X) = \mu + m$, and similarly $E(\bar{Y}) = 2m - \mu$ and $E(\bar{Z}) = m$, we get

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

   and

$$\begin{aligned} E(\mu_2^*) &= \frac{5}{9}E(\bar{X}) - \frac{4}{9}E(\bar{Y}) + \frac{3}{9}E(\bar{Z}) \\ &= \frac{5}{9}(\mu + m) - \frac{4}{9}(2m - \mu) + \frac{3}{9}m = \mu, \end{aligned}$$

   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

$$V(\bar{X}) = \frac{V(X)}{5} = \frac{2}{5}, \quad V(\bar{Y}) = \frac{V(Y)}{2} = \frac{1}{2}, \quad V(\bar{Z}) = \frac{V(Z)}{2} = \frac{1}{2}.$$

This gives

$$\begin{aligned} V(\mu_1^*) &= \left(\frac{1}{2}\right)^2 V(\bar{X}) + \left(-\frac{1}{2}\right)^2 V(\bar{Y}) + \left(\frac{1}{2}\right)^2 V(\bar{Z}) \\ &= \frac{1}{4} \cdot \frac{2}{5} + \frac{1}{4} \cdot \frac{1}{2} + \frac{1}{4} \cdot \frac{1}{2} = \frac{14}{40} = \frac{7}{20} = 0.35, \end{aligned}$$

and

$$\begin{aligned} V(\mu_2^*) &= \left(\frac{5}{9}\right)^2 V(\bar{X}) + \left(-\frac{4}{9}\right)^2 V(\bar{Y}) + \left(\frac{3}{9}\right)^2 V(\bar{Z}) \\ &= \frac{25}{81} \cdot \frac{2}{5} + \frac{16}{81} \cdot \frac{1}{2} + \frac{9}{81} \cdot \frac{1}{2} = \frac{45}{162} = \frac{5}{18} \approx 0.28, \end{aligned}$$

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

3. In his factory at the North Pole, Santa Claus wants to check if the production of baby dolls works as it should. He is satisfied if the average weight of all produced dolls, μ say, is greater than 0.5 kg.

   He picks a random sample of 20 baby dolls, and measures the weight of all dolls in this sample. The average weight of the dolls in the sample is 0.54 kg.

   Santa assumes that the weight of a doll is normally distributed with unknown expectation μ and (from previous studies) known standard deviation 0.1 kg.

   (a) Perform a suitable hypothesis test to conclude if Santa can be satisfied. Choose the 5% level. (1p)

   Solution: We have a random sample $x_1, ..., x_n$ from $N(\mu, \sigma^2)$, where $n = 20$ and $\sigma = 0.1$. Santa is satisfied if $\mu > 0.5$, so it is natural to have that under the alternative, and test H_0: $\mu = 0.5$ vs H_1: $\mu > 0.5$.
   Let the estimator be $\bar{X}$, the random variable corresponding to $\bar{x}$. We observe $\bar{x} = 0.54$. As test variable, we can take

$$T = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0.1).$$

   Hence, with inserted values (under H_0), we get

$$T_{obs} = \frac{0.54 - 0.50}{0.1/\sqrt{20}} \approx 1.79 > \lambda_{0.05} = 1.6449,$$

   and so, we may reject H_0 at the 5% level.
   At this level, we have evidence that Santa can be satisfied.

   (b) Calculate the P value of the test in (a). (1p)

   Solution: The P value is the probability under H_0 to obtain a value at least as extreme as the observed one for the test statistic, i.e.

$$\begin{aligned} P(T \geq 1.79) &= 1 - P(T < 1.79) = 1 - \Phi(1.79) = 1 - 0.9633 \\ &= 0.0367 \approx 0.04, \end{aligned}$$

   where $\Phi(\cdot)$ is the distribution function of a standard normal variable.
   As expected, the P value is smaller than 0.05.

(c) Calculate the power of the test in (a) if $\mu = 0.56$. (2p)

Solution: The power of the test is the probability to reject for a parameter value that belongs to H_1, here $\mu = 0.56$. At first, we need the critical region for the test, and as we see from (a), this region is given by (for $\mu = 0.5$)

$$T = \frac{\bar{X} - 0.5}{\sigma/\sqrt{n}} > \lambda_{0.05} = 1.6449.$$

The requested power is the probability for this event when $\mu = 0.56$, which is

$$\begin{aligned}
&P\left(\frac{\bar{X} - 0.5}{\sigma/\sqrt{n}} > \lambda_{0.05}\right) \\
&= P\left(\frac{\bar{X} - 0.56}{\sigma/\sqrt{n}} + \frac{0.56 - 0.5}{\sigma/\sqrt{n}} > \lambda_{0.05}\right) \\
&= P\left(\frac{\bar{X} - 0.56}{\sigma/\sqrt{n}} > \lambda_{0.05} - \frac{0.56 - 0.5}{\sigma/\sqrt{n}}\right) \\
&= 1 - P\left(\frac{\bar{X} - 0.56}{\sigma/\sqrt{n}} \leq \lambda_{0.05} - \frac{0.56 - 0.5}{\sigma/\sqrt{n}}\right) \\
&= 1 - \Phi\left(\lambda_{0.05} - \frac{0.56 - 0.5}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(1.6449 - \frac{0.56 - 0.5}{0.1/\sqrt{20}}\right) \\
&\approx 1 - \Phi(-1.04) = \Phi(1.04) \approx 0.8508 \approx 0.85.
\end{aligned}$$

(d) Santa is not satisfied with the power of the test, and regrets that he didn't take a bigger random sample. How many dolls should there be in the sample in order for the power for $\mu = 0.56$ to be at least 0.99? (1p)

Solution: From the derivation in (c), Santa would need a sample size n such that

$$1 - \Phi\left(\lambda_{0.05} - \frac{0.56 - 0.5}{\sigma/\sqrt{n}}\right) \geq 0.99,$$

or equivalently

$$\Phi\left(\sqrt{n}\frac{0.06}{\sigma} - \lambda_{0.05}\right) \geq 0.99 = \Phi(\lambda_{0.01}).$$

Since $\Phi(x)$ is monotone increasing in x, this means that

$$\sqrt{n}\frac{0.06}{\sigma} - \lambda_{0.05} \geq \lambda_{0.01},$$

i.e., with inserted values,

$$n \geq \sigma^2\frac{(\lambda_{0.01} + \lambda_{0.05})^2}{0.06^2} = 0.1^2\frac{(2.3263 + 1.6449)^2}{0.06^2} \approx 43.8.$$

Thus, Santa needs a sample size of at least $n = 44$.

4. At the North Pole factory, the pixies (tomtenissar) try out a new method to paint chess boards, which according to its inventor, the chief pixen Flirpo, is quicker than the old painting method. Five pixies, that are choosen randomly by Flirpo, get to paint two chess boards. Each of them uses the old method to paint one board and the new method to paint another. The painting skill may vary among pixies, but the skill is not assumed to increase with practice. (It is just the choice of method that might make a difference.)

   The results (in seconds) are given in the table below. Assume that the painting times are independent and normally distributed.

Pixie	Flirpa	Florp	Firp	Furpie	Furp
Old method	2.34	1.45	5.67	0.67	1.58
New method	2.12	1.15	5.34	0.55	1.28

   Test a suitable hypothesis to conclude if the new metod is better (quicker) than the old one. (5p)

   Solution: We can model this as paired sample (x_i, y_i) for $i = 1, 2, 3, 4, 5$, where the corresponding random variables (X_i, Y_i) are independent, with distributions $X_i \sim N(\mu_i, \sigma_1^2)$ and $Y_i \sim N(\mu_i + \Delta, \sigma_2^2)$, all parameters unknown. This means that

$$Z_i = Y_i - X_i \sim N(\Delta, \sigma^2),$$

   where $\sigma^2 = \sigma_1^2 + \sigma_2^2$ is unknown. Hence, the differences $z_i = y_i - x_i$ form a random sample from $N(\Delta, \sigma^2)$. The z_i are -0.22, -0.30, -0.33, -0.12, -0.30, from which we obtain $\bar{z} = -0.254$, $s_z^2 = 0.00728$ ($s_z \approx 0.0853$).

   We want to test H_0: $\Delta = 0$ vs H_1: $\Delta < 0$. (What we want to 'prove' is that the time is shorter for the new method, corresponding to $\Delta < 0$.) We can use the reference variable,

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

   for which we observe (under H_0)

$$T_{obs} = \frac{\bar{z} - 0}{s_z/\sqrt{5}} = \frac{-0.254}{\sqrt{0.00728}/\sqrt{5}} \approx -6.66 < -t_{0.01}(4) = -3.7469.$$

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

   On this risk level, we have evidence that the new method is better than the old one.

5. A director of studies at the North Pole Technical University (NPTU) examines the starting salaries for newly examined students. She takes a random sample of 100 students. The mean monthly salary for these students is 23.5 North Pole Dollars (NPD) with sample standard deviation 2.5.

   It is not permitted to assume that salaries are normally distributed.

   (a) Calculate a 95% confidence interval for the expected salary of a newly examined student from NPTU. (4p)

   Solution: We have a random sample $x_1, ..., x_n$, where $n = 100$, from a random variable X with expectation μ and variance σ^2. Both μ and σ are unknown, but estimated by $\bar{x} = 23.5$ and $s = 2.5$.

   Let $\bar{X}$ and S^2 be the random variables corresponding to $\bar{x}$ and s^2. Because n is large, it is permitted to use the approximation

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \approx t(n-1).$$

   Taking this as the reference variable, it follows that we get the 95% confidence interval (use $t_{0.025}(99) \approx t_{0.025}(100) = 1.9840$)

$$\begin{aligned} I_\mu &= \bar{x} \pm t_{0.025}(99)\frac{s}{\sqrt{100}} = 23.5 \pm 1.9840 \cdot \frac{2.5}{10} \\ &= 23.5 \pm 0.496 = (23.0,\ 24.0). \end{aligned}$$

   (b) Recently, the vice-chancellor of NPTU has claimed that the expected salary of a newly examined student from NPTU is 25 NPD. Given the result in (a), what can you say about this claim? (1p)

   Solution: The value 25 is not contained in the 95% confidence interval, so we have evidence that the vice-chancellor is not correct at the 5% level.

6. We have a random sample

$$4.2,\ 2.3,\ 0.4,\ 5.0,\ 3.7,\ 3.2,\ 0.7,\ 4.5,\ 2.6,\ 1.1,$$

from $X \sim \text{Re}(0, \theta)$.

Calculate a 95% confidence interval for θ. (5p)

Hint: Start by considering the estimator $\max_{1 \leq i \leq 10} X_i$, where $X_1, ..., X_{10}$ are random variables corresponding to the random sample of observations.

Solution: Let $Y = \max_{1 \leq i \leq 10} X_i$, where $X_1, ..., X_{10}$ are independent $\text{Re}(0, \theta)$. The distribution function of X may be written as

$$F_X(x) = \frac{x}{\theta} I\{x \leq \theta\},$$

where $\theta > 0$ and $I\{x \leq \theta\} = 1$ if $x \leq \theta$ and 0 otherwise.

It follows that, by independence, the distribution function of Y is given by

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(X_1 \leq y, ..., X_{10} \leq y) \\ &= \{P(X \leq y)\}^{10} = \{F_X(y)\}^{10} = \left(\frac{y}{\theta}\right)^{10} I\{y \leq \theta\}. \end{aligned}$$

Now, define the reference variable $Z = Y/\theta$. We find that the distribution function of Z is

$$F_Z(z) = P(Z \leq z) = P\left(\frac{Y}{\theta} \leq z\right) = P(Y \leq \theta z) = F_Y(\theta z) = z^{10} I\{z \leq 1\}.$$

We get the (upper) a quantile z_a by solving

$$a = P(Z > z_a) = 1 - F(z_a),$$

i.e.

$$1 - a = F(z_a) = z_a^{10} I\{z_a \leq 1\},$$

with solution $z_a = (1 - a)^{1/10}$. Hence, putting a equal to 0.975 and 0.025 respectively, we find

$$0.95 = P\left(0.025^{1/10} \leq Z = \frac{Y}{\theta} \leq 0.975^{1/10}\right),$$

i.e.

$$0.95 = P\left(\frac{Y}{0.975^{1/10}} \leq \theta \leq \frac{Y}{0.025^{1/10}}\right).$$

Inserting the observed sample maximum $y = 5.0$, this gives us the 95% confidence interval

$$I_\theta = \left(\frac{5}{0.975^{1/10}},\ \frac{5}{0.025^{1/10}}\right) = (5.01,\ 7.23).$$

7. A General Society Survey in the US cross-classified degree of fundamentalism of religious beliefs by the highest degree of education according to the table below.

   Are religious beliefs independent of education? Perform a suitable hypothesis test to find out. (5p)

	Religious Beliefs		
Education	Fundamentalist	Moderate	Liberal
High school/Junior college or lower	748	786	550
Bachelor or graduate	138	252	252

Solution: Test H_0: religious beliefs are independent of education vs H_1: $\neg H_0$. The row sums are $o_{1\cdot} = 748 + 786 + 550 = 2084$, $o_{2\cdot} = 138 + 252 + 252 = 642$, and the column sums are $o_{\cdot 1} = 748 + 138 = 886$, $o_{\cdot 2} = 786 + 252 = 1038$ and $o_{\cdot 3} = 550 + 252 = 802$. The total sum is $2084 + 642 = 2726$.

The expected counts for cells (i, j), $i = 1, 2$, $j = 1, 2, 3$, are

$$\begin{aligned} e_{11} &= \frac{2084 \cdot 886}{2726} = 677.34, \quad e_{12} = \frac{2084 \cdot 1038}{2726} = 793.54, \\ e_{13} &= \frac{2084 \cdot 802}{2726} = 613.12, \quad e_{21} = \frac{642 \cdot 886}{2726} = 208.66, \\ e_{22} &= \frac{642 \cdot 1038}{2726} = 244.46, \quad e_{23} = \frac{642 \cdot 802}{2726} = 188.88. \end{aligned}$$

All $e_{ij} > 5$, which means that χ^2 approximation is valid. The number of degrees of freedom is $(2 - 1)(3 - 1) = 2$.

We get the observed test statistic

$$\begin{aligned} Q &= \frac{(748 - 677.34)^2}{677.34} + \frac{(786 - 793.54)^2}{793.54} + \frac{(550 - 613.12)^2}{613.12} \\ &\quad + \frac{(138 - 208.66)^2}{208.66} + \frac{(252 - 244.46)^2}{244.46} + \frac{(252 - 188.88)^2}{188.88} \\ &\approx 59.2 > 13.816 = \chi^2_{0.001}(2), \end{aligned}$$

and so, we can reject H_0 at the 0.1% level.

At this level, we have evidence that religious beliefs and education are dependent.

8. Apartments of the same size are sold in two parts of a city, East and West. Two random samples are taken, one for each part. For these samples, the prices in millions of Swedish krones are given in the following table.

East	6.6	11.0	5.3	7.2	5.8
West	7.6	7.9	13.0		

It is not reasonable to assume that apartment prices are normally distributed.

Are the prices for this kind of apartments equally distributed for East and West? Try to answer this question by performing a suitable hypothesis test.

(5p)

Solution: Test H_0: the prices are equally distributed vs H_1: $\neg H_0$.

We can use either the Wilcoxon test or the permutation test.

For the Wilcoxon test, we rank the obervations (e.g. from lowest to highest). The rank sum for 'West' is

$$R = 5 + 6 + 8 = 19.$$

We can use table 9 of AB to find the critical values. Say that the significance level is 5%. Observe that the test is two-sided, so the critical region is the union of two one-sided critical regions at levels 0.025. We have $n_1 = 3$, $n_2 = 5$, and we get the critical region as the union of $\{R \leq 6\}$ and $\{R \geq 21\}$. We observe $R = 19$, and this is not contained in the critical region. Hence, we may not reject H_0 at the 5% level.

At the 5% level, we have no evidence that the prices are unequally ditributed for the two city parts.

We can also calculate the P value directly as two times (because the test is two sided) the fraction of possible rank sums for a group of three observations that are at most 22. These are $5 + 8 + 9 = 22$, $6 + 7 + 9 = 22$, $6 + 8 + 9 = 23$, $7 + 8 + 9 = 24$, so there are four of them. The number of possible rank sums is $\binom{8}{3} = 56$, hence we get the P value

$$2 \cdot \frac{4}{56} = \frac{1}{7} \approx 0.14 > 0.05.$$

Alternatively, we can use the permutation test. For this test, there is no table, but we can calculate the P value in the style as above. As the test statistic T, we can take the sum of observations for 'West'. (This is equivalent to take the statistic which is the difference of means for the two samples.) The observed sum is

$$T = 7.6 + 7.9 + 13.0 = 28.5.$$

As for the Wilcoxon test, the P value is two times the fraction of possible sums of three out of our nine observations that are at most 28.5. These are $7.6 + 7.9 + 13.0 = 28.5$, $5.8 + 11 + 13 = 29.8$, $6.6 + 11 + 13 = 30.6$,

$7.2 + 11 + 13 = 31.2$, $7.6 + 7.9 + 13.0 = 28.5$, $7.6 + 11 + 13 = 31.6$,
$7.9 + 11 + 13 = 31.9$, i.e. there are seven of them. This gives us the P value

$$2 \cdot \frac{7}{56} = \frac{1}{4} = 0.25 > 0.05.$$

Hence also with the permutation test, we may not reject H_0 at the 5% level.

At this level, we have no evidence that the prices are unequally distributed for the two city parts.