UPPSALA UNIVERSITY
Department of Mathematics
Rolf Larsson

Exam in Mathematical Statistics
Inference 1 1MS035
2022–01–12
Solutions

1. We have a random sample 1.7, 1.2, 2.1 from a random variable X with density function

$$f_X(x) = \begin{cases} 2\theta^{-2}x, & 0 \le x \le \theta, \\ 0, & \text{otherwise,} \end{cases}$$

where $\theta > 0$. Estimate θ using

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

Solution: At first, we calculate

$$m(\theta) = E(X) = \int_0^\theta x \cdot 2\theta^{-2} x dx = 2\theta^{-2} \int_0^\theta x^2 dx = \frac{2}{3}\theta.$$

Moreover, we observe $\bar{x} = 5/3$. Hence, the moment estimate solves

$$\frac{2}{3}\theta = \frac{5}{3},$$

and so, it is $\theta^* = 5/2 = 2.5$.

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

Solution: In general, for observations $x_1, ..., x_n$, we need to minimize

$$Q(\theta) = \sum_{i=1}^n \{x_i - m(\theta)\}^2 = \sum_{i=1}^n \left(x_i - \frac{2}{3}\theta\right)^2.$$

The first two derivatives are

$$Q'(\theta) = -\frac{4}{3}\sum_{i=1}^n \left(x_i - \frac{2}{3}\theta\right) = -\frac{4}{3}n\left(\bar{x} - \frac{2}{3}\theta\right),$$
$$Q''(\theta) = \frac{8}{9}n.$$

We find that $Q''(\theta) > 0$ for all θ, so we get a minimum by solving $Q'(\theta) = 0$. This solution is easily seen to give the estimate

$$\theta^* = \frac{3}{2}\bar{x} = \frac{5}{2} = 2.5,$$

i.e. the same as the moment estimate.

In fact, this follows directly from 'anmärkning 7.11' at p.286 in Alm-Britton.

(c) maximum likelihood. (2p)

Solution: For positive observations $x_1, ..., x_n$, the likelihood is

$$L(\theta) = \prod_{i=1}^{n} f_X(x_i) = \prod_{i=1}^{n} 2\theta^{-2} x_i I\{x_i \le \theta\},$$

where $I\{A\}$ is 1 if A is fulfilled and 0 otherwise. We may rewrite the likelihood as

$$L(\theta) = 2^n \left(\prod_{i=1}^{n} x_i \right) \theta^{-2n} I\{\max_i x_i \le \theta\}.$$

As a function of θ, $L(\theta) = 0$ for all $\theta < \max_i x_i$, and for $\theta \ge \max_i x_i$, it is monotone decreasing with its maximum at $\theta = \max_i x_i$. Hence, the maximum likelihood estimate is

$$\theta^* = \max_i x_i = 2.1,$$

which is different from the other two estimates.

2. A statistician wants to investigate the proportion of people (who are assumed to be employed) that take their car to work.

(a) Suppose that the proportion of people living in average size cities that take their car to work is p. A random sample of 1000 people from such cities is taken, among which x take their car to work. Estimate p by

$$p_1^* = \frac{x}{1000}.$$

Show that p_1^* is unbiased for p. (1p)

Solution: We have one observation x from $X \sim \mathrm{Bin}(1000, p)$. This means that for the estimator corresponding to p_1^*, $p_1^*(X) = X/1000$, we have

$$E\{p_1^*(X)\} = \frac{1}{1000} E(X) = \frac{1}{1000} 1000p = p,$$

i.e. p_1^* is unbiased.

(b) Suppose that the proportion of people who live in big cities that take their car to work is $p - a$, while the proportion that does so of people that live in small cities or at the countyside that do so is $p + a$, where $a > 0$.

Assume that in a new investigation, the sample is stratified so that a random sample of 250 people who live in big cities is taken, and a random sample of 750 people who live in small cities or at the countryside is taken. (None of the samples contains any people living in average size cities.)

In the big city sample, y take their car to work, and in the other sample, the number of people taking their car to work is z.

The estimate

$$p_2^* = \frac{3y + z}{1500}$$

is proposed.

Show that p_2^* is unbiased for p. (1p)

Solution: As in (a), we have that y and z are observations of $Y \sim \text{Bin}(250, p - a)$ and $Z \sim \text{Bin}(750, p + a)$, respectively.
The expectation of the estimator corresponding to p_2^* is

$$\begin{aligned} E\{p_2^*(Y, Z)\} &= \frac{3}{1500} E(Y) + \frac{1}{1500} E(Z) \\ &= \frac{3}{1500} 250(p - a) + \frac{1}{1500} 750(p + a) = p, \end{aligned}$$

i.e. p_2^* is unbiased.

(c) Assume that

$$p < \frac{1}{2}, \quad a > \frac{1}{2}\sqrt{p(1-p)}.$$

Then, show that p_2^* is more efficient than p_1^*. (3p)

Solution: The most efficient estimate is the one whose corresponding estimator has the smallest variance. The variances for our estimators are

$$V\{p_1^*(X)\} = \left(\frac{1}{1000}\right)^2 V(X) = \frac{1}{1000^2} 1000p(1-p) = \frac{p(1-p)}{1000}$$

and

$$\begin{aligned} &V\{p_2^*(Y, Z)\} \\ &= \left(\frac{3}{1500}\right)^2 V(Y) + \left(\frac{1}{1500}\right)^2 V(Z) \\ &= \frac{9}{1500^2} 250(p - a)\{1 - (p - a)\} + \frac{1}{1500^2} 750(p + a)\{1 - (p + a)\} \\ &= \frac{9 \cdot 250 + 750}{1500^2}\{p(1-p) - a^2\} - \frac{9 \cdot 250 - 750}{1500^2}(1 - 2p)a \\ &= \frac{2}{1500}\{p(1-p) - a^2\} - \frac{1}{1500}(1 - 2p)a. \end{aligned}$$

Now, because $p < 1/2$, we have $1 - 2p > 0$, which because $a > 0$ gives

$$V\{p_2^*(Y, Z)\} < \frac{2}{1500}\{p(1-p) - a^2\}.$$

Moreover, $a > \frac{1}{2}\sqrt{p(1-p)}$ implies $a^2 > \frac{1}{4}p(1-p)$, and so,

$$V\{p_2^*(Y, Z)\} < \frac{2}{1500} \cdot \frac{3}{4}p(1-p) = \frac{p(1-p)}{1000} = V\{p_1^*(X)\},$$

which shows that p_2^* is more efficient than p_1^*, Q.E.D.

3. In his factory at the North Pole, Santa Claus wants to make sure that the proportion of defect toy cars is not too high. He is satisfied if this proportion is lower than 1%.

   He checks if this can be true by standing at the production line, counting the number of produced toy cars until the first defective one comes along. He then observes that the first defective toy car is the 350th one.

   (a) Perform a suitable hypothesis test to conclude if Santa can be satisfied. Choose the 5% level. (2p)

   Solution: The number of toy cars that Santa needs to check including the first one that is defect, X say, is geometric distributed (Swedish: ffg) with parameter p (say). We observe $x = 350$.

   Santa is satisfied if $p < 0.01$, so it could be suitable to have this in the alternative hypothesis (what we want to 'prove'). Hence, we would like to test H_0: $p = 0.01$ vs H_1: $p < 0.01$.

   It is clear that the smaller p is, the longer it takes (on average) to find the first defect toy car, so we should reject for large x. This leads us to the P value

$$P(X \geq 350; p = 0.1) = 0.99^{349} \approx 0.03$$

   (the event $X \geq 350$ is equivalent to the event that the first 349 cars are non defect), and because $0.03 < 0.05$, we can reject H_0 at the 5% level.

   On the 5% level, we have evidence that Santa will be satisfied.

(b) Calculate the power of the test in (a) if in fact, 0.1% of the toy cars are defect. (3p)

Solution: At first, we need to calculate the critical region for the 5% test. The critical region should be on the form $\{x \geq C\}$. Here, C is the smallest integer C such that $P(X \geq C; p = 0.1) < 0.05$.
Because as in (a), $P(X \geq C; p = 0.1) = 0.99^{C-1}$, we have the inequality

$$0.99^{C-1} < 0.05,$$

i.e. $(C - 1) \ln 0.99 < \ln 0.05$, i.e.

$$C > 1 + \frac{\ln 0.05}{\ln 0.99} \approx 299.1.$$

Hence, we need to choose $C = 300$.
This means that the power for the test if $p = 0.1\% = 0.001$ is

$$P(X \geq 300; p = 0.001) = 0.999^{299} \approx 0.74,$$

so the required power is about 74%.

4. A researcher wants to find out if cows weight differently at farms that produce ecologic milk compared to farms that produce non ecologic milk. From each type of farm, ten cows are randomly selected. Their weights in tons are given in the table below.

Ecologic	1.51	1.23	0.85	1.44	0.97	1.12	1.37	1.22	0.88	1.12
Non ecologic	1.33	0.99	1.45	1.02	0.78	1.05	1.26	1.02	1.28	1.36

Hint: The sample mean and variance of the weights for the ecologic sample are 1.171 and 0.0518, respectively. For the non ecologic sample, the corresponding numbers are 1.154 and 0.0446, respectively.

(a) Calculate a 95% confidence interval for the difference of mean weights between cows on ecologic and non ecologic farms. Make sure to carefully specify all your assumptions. (4p)

Solution: We assume that we have two independent samples, $x_1, .., x_{n_1}$ from $X \sim N(\mu_1, \sigma_1^2)$ (ecologic) and $y_1, .., y_{n_2}$ from $Y \sim N(\mu_2, \sigma_2^2)$ (non ecologic), where $n_1 = n_2 = 10$. 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 $\bar{x} = 1.171$ and $\bar{y} = 1.154$, and the sample variances $s_x^2 = 0.0518$ and $s_y^2 = 0.0446$. This gives us the pooled variance

$$s_p^2 = \frac{(n_1 - 1)s_x^2 + (n_2 - 1)s_y^2}{n_1 + n_2 - 2} = \frac{9 \cdot 0.0518 + 9 \cdot 0.0446}{18} = 0.0482,$$

and we get the 95% confidence interval

$$\begin{aligned} I_{\mu_1 - \mu_2} &= \bar{x} - \bar{y} \pm t_{0.025}(n_1 + n_2 - 2) s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \\ &= 1.171 - 1.154 \pm t_{0.025}(18)\sqrt{0.0482}\sqrt{\frac{1}{10} + \frac{1}{10}} \\ &= 0.017 \pm 2.1009\sqrt{0.0482}\sqrt{\frac{2}{10}} \\ &= 0.017 \pm 0.206 = (-0.189,\ 0.223). \end{aligned}$$

where $t_{0.025}(18) = 2.1009$ is obtained from table 6.

(b) Can you conclude that cows weight differently at farms that produce ecologic milk compared to farms that produce non ecologic milk? (1p)

Solution: No, not at the risk level 5% ($= 1 - 0.95$), because $\mu_1 - \mu_2 = 0$ is included in the confidence interval.

5. In the distant ski resort Svartlien, the number of avalanches (Swedish: laviner) per year is supposed to be Poisson distributed with parameter λ. In the last ten years, the numbers of avalanches have been 3, 0, 4, 2, 2, 6, 1, 5, 6, 1.

   (a) Estimate λ. (1p)

   Solution: The moment estimate (and also the MLE) is given by $\lambda^* = \bar{x} = 3$, where $\bar{x} = 3$ is the sample mean.

   (b) Gudrun says that, according to her experience from 'old days', the mean number of avalanches per year in Svartlien is 2, no more, no less. She thinks it still ought to be this way. Check if Gudrun is right by performing a suitable hypothesis test. (4p)

   Solution: Let $X \sim \text{Po}(\lambda)$ be the number of avalanches per year. From Gudrun's claim, there is no indication of any direction of an alternative hypothesis, so we want to test H_0: $\lambda = 2$ vs H_1: $\lambda \neq 2$.

   We have observations $x_1, ..., x_n$ from X with $n = 10$. Because $2n = 20 > 15$, normal approximation is permitted, and under H_0, we have for the estimator $\bar{X}$ that $E(\bar{X}) = 2$ and $V(\bar{X}) = 2/n$. Hence, we have the test statistic

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

   We observe

$$T_{obs} = \frac{\bar{x} - 2}{\sqrt{2/n}} = \frac{3 - 2}{\sqrt{2/10}} = \sqrt{5} \approx 2.24.$$

   Testing at the 5% level, we find that $2.24 > 1.96 = \lambda_{0.025}$ and so, we may reject H_0 at the 5% level.

   On this risk level, we have evidence that Gudrun's claim is false.

   Alternative 1: Instead consider $T = (\sum X_i - 2n)/\sqrt{2n}$, leading to the same result.

   Alternative 2: We can perform an 'exact' calculation by considering that under H_0, $Y = \sum_i X_i \sim \text{Po}(20)$, which gives the P value

$$2P(Y \geq 30) = 2 \cdot \{1 - P(Y \leq 29)\} \approx 0.044 < 0.05.$$

6. Zerblatt performs 100 independent measurements of the acceleration due to gravity g, in m/s^2, at the planet QZ34. The measurements have mean 15.20 and standard deviation 0.8.

   (a) Calculate a 99% confidence interval for the true value of g at QZ34. (3p)

   Solution: Suppose that we have a random sample $x_1, ..., x_n$ where $n = 100$ from an unknown distribution with expectation g and unknown variance σ^2. Use the reference variable

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

   This gives a confidence interval with approximative confidence level 99% via

$$\bar{x} \pm t_{0.005}(99)\frac{s}{\sqrt{100}} = 15.20 \pm 2.6259\frac{0.8}{10} = 15.20 \pm 0.21 = (14.99, 15.41).$$

   (b) A law in physics say that $G = m \cdot g$, where G is the weight of an object measured in Newton (N), and where m is its mass in kg. On the planet QZ34, Zerblatt finds a rock with weight 8.0 N. Calculate a 99% confidence interval for the mass of this rock. (2p)

   Solution: We have $8 = m \cdot g$, i.e. $g = 8/m$. The interval $14.99 \leq 8/m \leq 15.41$ corresponds to $8/15.41 \leq m \leq 8/14.99$.
   Hence, a confidence interval with approximate confidence level 99% for m is

$$(8/15.41,\ 8/14.99) = (0.519,\ 0.534).$$

7. At the latest traditional Svartlien New Years eve Slalom competiton (NYS), eight participants got results (in seconds) in two runs on the same track (and with the same snow conditions), according to the table below.

Participant no	1	2	3	4	5	6	7	8
Time run 1	14.5	12.3	24.3	36.2	27.0	17.0	26.6	33.1
Time run 2	13.8	12.7	23.7	29.8	26.5	14.2	25.9	31.5

For potential competitors of this event, does the skill improve (the running time get shorter) from run 1 to run 2? Try to answer this question by performing a suitable hypothesis test.

It is not allowed to assume that the run times are normally distributed. (5p)

Lösning: We want to find evidence that the skill improves, and so, we should have this in the alternative hypothesis. Hence, we test H_0: the skill does not improve vs H_1: the skill improves.

The sign test: We have 8 differences, time run 1 minus time run 2:

0.7, −0.4, 0.6, 6.4, 0.5, 2.8, 0.7, 1.6.

If the skill improves, these differences should tend to be positive.

Let X be the number of positive differences. Under H_0, $X \sim \text{Bin}(8, 1/2)$. We observe $x = 7$. Thus, the P value is

$$P(X \geq 7) = 8 \cdot 2^{-8} + 2^{-8} = 9/2^8 = 0.035,$$

which is smaller than 5%. Conclusively, we may reject H_0 at the 5% level. At this level, we have evidence that the skill has improved.

The signed rank test: If we take absolute values of the differences, the only observed negative difference has rank 1. There are 2^8 ways to assign plus or minus to the 8 differences. We may get rank sum in at most two ways, either the observed one, which gives rank sum 1, or if all differences are positive, which gives rank sum 0. Hence, the P value is $2/2^8 \approx 0.008$, and we may reject H_0 e.g. at the 1% level. (Observe that this is not possible for the sign test here.)

Alternatively, we can use the critical region from table 10, $T \leq 5$ for a one-sided 5% test. We observe $T = 1 \leq 5$, so we can reject at the 5% level (and also at the 1% level, with critical region $T \leq 1$).

8. The numbers of people that emigrated from Sweden during the years 2009-2020 are given in the following table (data from Statistics Sweden).

Year	2009	2010	2011	2012	2013	2014
Number	39 240	48 853	51 179	51 747	50 715	51 237

Year	2015	2016	2017	2018	2019	2020
Number	55 830	45 878	45 620	46 981	47 718	48 937

Is there a trend in this material? Try to answer this question by performing a suitable hypothesis test. (5p)

Lösning: Test H_0: no trend vs H_1: trend. Possible tests (that do not assume the normal distribution) are the runs test and a test based on Spearman's rank correlation.

Runs test: the observed median equals the mean of the two 'middle' numbers when ranked, i.e.

$$\frac{48853 + 48937}{2} = 48895.$$

If we assign 0 to the numbers below the median and 1 to the numbers above, we get the sequence 00 11111 0000 1, i.e. four runs. We reject for few runs (many runs does not correspond to a trend). We have $n_1 = 6 = n_2$, and this is too small for normal approximation to be valid. We need to calculate the P value in an 'exact' manner.

If Z is the number of runs, we get

$$P(Z=2) = \frac{2\binom{5}{0}^2}{\binom{12}{6}} = \frac{2}{\binom{12}{6}}, \quad P(Z=3) = \frac{2\binom{5}{1}\binom{5}{0}}{\binom{12}{6}} = \frac{10}{\binom{12}{6}},$$
$$P(Z=4) = \frac{2\binom{5}{1}^2}{\binom{12}{6}} = \frac{50}{\binom{12}{6}},$$

which yields the P value

$$P(Z \leq 4) = \frac{2}{\binom{12}{6}} + \frac{10}{\binom{12}{6}} + \frac{50}{\binom{12}{6}} = \frac{62}{\binom{12}{6}} \approx 0.067 > 0.05.$$

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

At this level, we have no evidence of a trend.

Spearman: The ranks for the numbers are

$$1, 6, 9, 11, 8, 10, 12, 3, 2, 4, 5, 7,$$

and we get

$$S = \sum_{j=1}^{12} d_j^2 = (1-1)^2 + (6-2)^2 + ... + (7-12)^2 = 322,$$

which gives the observed Spearman rank korrelation ($n = 12$)

$$r_S = 1 - \frac{6S}{n^3 - n} = 1 - \frac{6 \cdot 322}{12^3 - 12} \approx -0.1259.$$

The observed test statistic is

$$T_{obs} = \sqrt{n-1} r_S = \sqrt{11} \cdot (-0.1259) \approx -0.42.$$

For a test on the level α, $|T_{obs}|$ is to be compared to $\lambda_{\alpha/2}$, because H_1 encompasses both a positive and negative trend, and these result in positive or negative values on the test statistic, respectively.

Because $|-0.42| < 1.96 = \lambda_{0.025}$, we may not reject H_0 at the 5% level.

At this level, we have no evidence of a trend.