Empirical Methods

– Exam –

Werner Nutt

10 September 2019

Name: ____

Registration Number: ____________________________________

Question	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
Points	23	13	23	15	18	8	100
Reached							

Instructions

- The exam comprises 6 questions, which consist of several subquestions. You will have 2 hours time to answer the questions.

- If numerical answers are requested it suffices to write them down as arithmetic expressions, like e.g., $\frac{7 \cdot 3+5}{30}$. Expressions should be simplified as much as possible, though.

- There is a total of 100 points that can be achieved in this exam. Marking will be out of 90. That is, to achieve a final mark of 30, it will suffice to obtain 90 points.

- Please, write down the answers to your questions in the present exam booklet handed out to you.

- For drafts use the blank paper provided by the university.

- If the space in the booklet turns out to be insufficient, please use the university paper for additional answers and return them with the booklet.

- **No questions** will be answered during the exam. If you are not sure about interpreting a question, you may write down additional assumptions you made in order to proceed with your solution.

1 Drawing Balls from an Urn (28 Points)

Suppose we have an infinite supply of balls, colored either red or blue. The probability of drawing a red ball from the supply is equal to the probability of drawing a blue one. We draw two balls from the supply and put them into an urn. Now, we repeatedly draw a ball from the urn, note its colour, and then return to the urn.

(i) What is the probability that the first ball drawn is red?

(ii) What is the probability that the first two balls drawn are red?

Now, suppose that the first two balls drawn are in fact red.

(iii) What is the probability that both balls in the urn are coloured red?

(iv) What is the probability that the next ball drawn will be red?

(v) Generalizing, what is the probability that the next n balls drawn will be red?

Justify your answers.

Hint: Bayes' Theorem may be useful for this question.

Sample Solution

(i) What is the probability that the first ball drawn is red?

Answer: 1/2

There are two possible arguments. The first is that anyway blue and red balls are equally likely to end up in the urn. If we take a ball from the urn, this is as good as directly taking a ball from the infintite supply.

The other one considers the three possible cases for the urn: (1) two red balls, (2) two blue balls, (3) one red and one blue ball. The probabilities for these cases are 1/4, 1/4, and 1/2, respectively. The probabilities for drawing a red ball in these cases are 1, 0, and 1/2, respectively. The overall probability therefore is

$$1/4 \times 1 + 1/4 \times 0 + 1/2 \times 1/2 = 1/4 + 0 + 1/4 = 1/2.$$

(ii) What is the probability that the first two balls drawn are red?

Answer: 3/8

Now we continue the second argument above. In the first case, the probability for two red balls is 1, in the second 0, and in the third $1/2 \times 1/2 = 1/4$. The overall probability therefore is

$$1/4 \times 1 + 1/4 \times 0 + 1/2 \times 1/4 = 1/4 + 0 + 1/8 = 3/8.$$

(iii) What is the probability that both balls in the urn are coloured red?

For this and the next question we apply Bayes' Rule. We consider two events:

$\mathcal{A} =$ both balls in the urn are red;

$C =$ the first two balls chosen are red.

Then Bayes' Rule says

$$P(\mathcal{A} \mid C) = \frac{P(C \mid \mathcal{A})\, P(\mathcal{A})}{P(C)}.$$

We list the three quantities on the right:

$P(\mathcal{A}) = 1/4$, as found earlier;

$P(C) = 3/8$, which is the answer to (ii);

$P(C \mid \mathcal{A}) = 1$, which is obvious.

Filling into Bayes' formula gives

$$P(\mathcal{A} \mid C) = \frac{P(C \mid \mathcal{A})\, P(\mathcal{A})}{P(C)} = \frac{1 \times 1/4}{3/8} = 1/4 \times 8/3 = 2/3.$$

(iv) What is the probability that the next ball drawn will be red.

Clearly, after two red balls have bee drawn, it is impossible that the urn contains only blue balls. Hence, either all balls in the urn are red, or there is one red and one blue ball. Denoting the case "one ball in the urn is red and the other one blue" as $\mathcal{B}$, we conclude $P(\mathcal{B} \mid C) = 1/3$. Let $\mathcal{R}$ denote the event "the next ball drawn is red" as $\mathcal{R}$. Then

$$\begin{aligned} P(\mathcal{R} \mid C) &= P(\mathcal{R}\mathcal{A} \mid C) + P(\mathcal{R}\mathcal{B} \mid C) \\ &= P(\mathcal{R} \mid \mathcal{A}C) \times P(\mathcal{A} \mid C) + P(\mathcal{R} \mid \mathcal{B}, C) \times P(\mathcal{B} \mid C) \\ &= 1 \times 2/3 + 1/2 \times 1/3 \\ &= 2/3 + 1/6 (= 5/6). \end{aligned}$$

(v) Generalizing, what is the probability that the next n balls drawn will be red?

We continue the preceding argument and denote "the next n balls are red" as $\mathcal{R}_n$. Then the calculation is as above, replacing $\mathcal{R}$ with $\mathcal{R}_n$:

$$\begin{aligned} P(\mathcal{R}_n \mid C) &= P(\mathcal{R}_n \mathcal{A} \mid C) + P(\mathcal{R}_n \mathcal{B} \mid C) \\ &= P(\mathcal{R}_n \mid \mathcal{A}C) \times P(\mathcal{A} \mid C) + P(\mathcal{R}_n \mid \mathcal{B}, C) \times P(\mathcal{B} \mid C) \\ &= 1 \times 2/3 + (1/2)^n \times 1/3 \\ &= 2/3 + 1/(3 \times 2^n). \end{aligned}$$

2 Waiting Time **(13 Points)**

The waiting time (in minutes) until an event is observed is distributed as an exponential distribution with parameter $\lambda = 0.5$.

(i) On average, how much time do we have to wait to observe the event?

(ii) If the event is not observed in the first 2 minutes, how much more do you expect to wait for it?

(iii) If we run the experiment twice in parallel, find the probability that the minimum waiting time is at least 3 minutes.

Justify your answers, mentioning all assumptions made.

Hint: Remember that the density of an exponential distribution is $\lambda e^{-\lambda x}$ for $x \geq 0$.

Sample Solution

(i) On average, how much time do we have to wait to observe the event?

The expected value of an exponential with parameter λ is $1/\lambda$; so we expect to wait 2 minutes.

(ii) If the event is not observed in the first minute, how much more do you expect to wait for it?

The exponential is memoryless, so if it has not been observed in the first 2 minutes, we expect to wait 2 minutes more.

(iii) If we run the experiment twice in parallel, find the probability that the maximum waiting time is at most 3 minutes.

Assume that the two experiments running in parallel are independent of each other; call them $\mathcal{X}$ and $\mathcal{Y}$. Then

$$\begin{aligned} P[\min\{\mathcal{X},\mathcal{Y}\} \geq 3] &= P[\mathcal{X} \geq 3, \mathcal{Y} \geq 3] = P[\mathcal{X} \geq 3]P[\mathcal{Y} \geq 3] \\ &= P[\mathcal{X} \geq 3]^2 = (1-(1-e^{-1.5}))^2 = e^{-3} \end{aligned}$$

3 Joint Distribution **(23 Points)**

The joint probability density function of $\mathcal{X}$ and $\mathcal{Y}$ is

$$f(x, y) = cxe^{-x-2y} \qquad 0 < x;\ 0 < y.$$

(i) Find the value c.

(ii) Compute the density function of $\mathcal{X}$.

(iii) Compute the density function of $\mathcal{Y}$.

(iv) Are $\mathcal{X}$ and $\mathcal{Y}$ independent? (Explain your answer.)

(v) Find $P(\mathcal{Y} < \mathcal{X})$.

Show your computations.

Sample Solution

(i) Find the value c.

$$\begin{aligned} 1 = \int_0^\infty \int_0^\infty f(x, y)\,dx\,dy &= \int_0^\infty \int_0^\infty cxe^{-x-2y}\,dx\,dy \\ &= c\int_0^\infty xe^{-x}\,dx \int_0^\infty e^{-2y}\,dy \end{aligned}$$

We now compute the two integrals separately:

$$\begin{aligned} \int_0^\infty xe^{-x}\,dx &= \Big[- xe^{-x} \Big]_0^\infty - \int_0^\infty (-e^{-x})\,dx \\ &= 0 + \int_0^\infty e^{-x}\,dx \\ &= 0 + \Big[- e^{-x} \Big]_0^\infty = 1 \end{aligned}$$

$$\int_0^\infty e^{-2y}\,dy = \Big[- \frac{1}{2}e^{-2y} \Big]_0^\infty = \frac{1}{2}.$$

We see that $c\frac{1}{2} = 1$, so $c = 2$.

(ii) Compute the density function of $\mathcal{X}$.

$$\begin{aligned} f_{\mathcal{X}}(x) &= 2\int_0^\infty xe^{-x}e^{-2y}\,dy \\ &= xe^{-x}\,2\int_0^\infty e^{-2y}\,dy = xe^{-x} \end{aligned}$$

(iii) Compute the density function of $\mathcal{Y}$.

$$\begin{aligned} f_{\mathcal{Y}}(y) &= 2\int_0^\infty xe^{-x}e^{-2y}\,dx \\ &= 2e^{-2y}\int_0^\infty xe^{-x}\,dx = 2e^{-2y} \end{aligned}$$

(iv) Are $\mathcal{X}$ and $\mathcal{Y}$ independent?

We see that $f(x,y) = f_{\mathcal{X}}(x)f_{\mathcal{Y}}(y)$, therefore $\mathcal{X}$ and $\mathcal{Y}$ are independent.

(v) Find $P(\mathcal{Y} < \mathcal{X})$.

$$\begin{aligned} P(\mathcal{Y} < \mathcal{X}) &= \int_0^\infty \int_0^x f(x,y)\,dy\,dx \\ &= \int_0^\infty xe^{-x}\int_0^x 2e^{-2y}\,dy\,dx \\ &= \int_0^\infty xe^{-x}\Big[-\frac{2}{2}e^{-2y}\Big]_0^x\,dx \\ &= \int_0^\infty xe^{-x}\Big(1-e^{-2x}\Big)\,dx \\ &= \int_0^\infty x\Big(e^{-x}-e^{-3x}\Big)\,dx \\ &= \Big[x\,(\frac{1}{3}e^{-3x}-e^{-x})\Big]_0^\infty - \int_0^\infty \frac{1}{3}e^{-3x}-e^{-x}\,dx \\ &= 0 + \int_0^\infty e^{-x}-\frac{1}{3}e^{-3x}\,dx \\ &= 1-\Big[-\frac{1}{9}e^{-3x}\Big]_0^\infty \\ &= 1-(0+\frac{1}{9}) = \frac{8}{9} \end{aligned}$$

4 Testing Car Tyres (15 Points)

The manufacturer of a new car tyre claims that its average life will be at least 60,000 km.

To verify this claim a sample of 25 tyres is tested. The outcome of the test is a sample mean of 54,000 km and a sample standard deviation of 12,000 km.

(i) Find a value c such that, with probability 99%, the true mean is less than c.

(ii) Compute the p-value for the hypothesis $H_0 : \mu \geq 60,000$. Approximate it as well as you can from the probability tables provided.

(iii) What would you need to change in your calculations if we knew that the *population* standard deviation is 12,000 km? What would be the values for c in (i) and the p-value in (ii)?

Justify your answers.

Sample Solution

(i) Find a value c such that, with probability 99%, the true mean is less than c.

We want a 99% lower CI, which can be easily computed using the table. That is:

$$c = \overline{X} + t_{\alpha,n-1}\frac{S}{\sqrt{n}},$$

where in this case $\overline{X} = 54,000$, $\alpha = 0.01$, $S = 12,000$, and $n = 25$. Thus,

$$c = 54,000 + t_{0.01,24}\frac{12,000}{5} = 54,000 + 2.492 \times 2,400 \approx 54,000 + 6,000 = 60,000.$$

(ii) Compute the p-value for the hypothesis $H_0 : \mu \geq 60,000$. Approximate it as well as you can from the probability tables provided.

The test statistic for this hypothesis is

$$v = \frac{\sqrt{n}}{S}(\mu_0 - \overline{X}) = \frac{\sqrt{25}}{12,000}(60,000 - 54,000) = 5\,\frac{6,000}{12,000} = 2.5$$

From the table of t_{24}, the closest significance level for rejecting the hypothesis is 0.01. So, we can say that the p-value is approximately 0.01.

(iii) What would you need to change in your calculations if we knew that the population standard deviation is 12,000 km? What would be the values for c in (i) and the p-value in (ii)?

All the computations will be analogous, but knowing the population standard deviation allows us to use a normal distribution instead of a t-distribution. First, we review the computation of the lower CI:

$$\begin{aligned} c &= \overline{X} + z_{\alpha}\frac{\sigma}{\sqrt{n}} \\ &= 54,000 + z_{0.01}\frac{12,000}{5} \\ &\approx 54,000 + 2.33 \times 2,400 = 54,000 + \frac{7}{3}2,400 = 54,000 + 5,600 = 59,600. \end{aligned}$$

Then we recalculate the p-value. The test statistic v is analogous:

$$v = \frac{\sqrt{n}}{\sigma}(\mu_0 - \overline{X}) = \frac{\sqrt{25}}{12,000}(60,000 - 54,000) = 2.5.$$

We see from the z-table that the $P(Z <= 2.5) = 0.9938$. Hence the corresponding p-value is 0.0062 instead of 0.01.

5 Low and High Throws With a Fair Die (18 Points)

We throw a fair die repeatedly. If the die falls on 1, 2, or 3, we say that it is a low throw (otherwise, it is a high throw). Approximate the probability of:

(i) seeing at most 55 low throws, in 100 throws of the die,

(ii) seeing at least 215 highs, in 400 throws.

Suppose now that you win 1 Euro for each low throw, and lose 1 Euro for each high throw.

(iii) What is the probability of winning more than 10 Euros after 100 throws?

(iv) How many throws do you need to guarantee that the probability of winning at least 10 Euros is greater than 0.99?

Justify your answers.

Sample Solution

Making a low throw on a fair die is a Bernoulli experiment with parameter $p = 0.5$. Making n throws is then a binomial variable with parameters (n, p), which has mean $np = n/2$ and variance $np(1 - p) = n/4$. Call these variables $\mathcal{X}_n$. Since n is large enough (100 or 400), we can use the Central Limit Theorem, and approximate this binomial through a normal $\mathcal{N}(n/2, n/4)$.

(i) Approximate the probability of seeing at most 55 low throws, in 100 throws of the die:

$$P[\mathcal{X}_{100} \leq 55] = P[\mathcal{X}_{100} < 55.5] \approx P[Z < \frac{55.5 - 50}{\sqrt{25}}] = P[Z < 1.1] = 0.8643$$

(ii) Approximate the probability of seeing at least 215 highs, in 400 throws:

$$\begin{aligned} P[\mathcal{X}_{400} \geq 215] &= P[\mathcal{X}_{400} > 214.5] \approx P[Z > \frac{214.5 - 200}{\sqrt{100}}] \\ &= P[Z > 1.45] = 1 - 0.9265 = 0.0735 \end{aligned}$$

(iii) What is the probability of winning more than 10 Euros after 100 throws?

To win more than 10 euros after 100 throws, the difference between the number of low throws and high throws should be more than 10; that is, we must see more than 55 low throws. So, from the answer 1 we know that the probability of winning more than 10 euros is 1 - 0.8643 = 0.1357.

(iv) How many throws do you need to guarantee that the probability of winning at least 10 Euros is greater than 0.9?

From the use of a normal distribution, we can see that the probability of winning at least 10 Euros is always < 0.5, regardless of the number of throws used. So, it is impossible to ever get this probability to 0.9.

6 Example Distribution (8 Points)

Provide an example distribution covered in the course (other than the exponential distribution, which occurred in the exam).

(i) Give the name and the density function of the distribution.

(ii) Briefly explain where it is applied.