II Università di Roma, Tor Vergata
Dipartimento d'Ingegneria Civile e Ingegneria Informatica
LM in Ingegneria dell'Informazione e dell'Automazione
Complementi di Probabilità e Statistica
Homework - 2019-12-06

Problem 1 *Let $(\Omega, \mathcal{E}, \mathbf{P}) \equiv \Omega$ be a probability space and let $(X_n)_{n\geq 1}$ be a sequence of independent identically distributed Bernoulli random variables with success probability p. Set*

$$Z_n \overset{def}{=} \textstyle\sum_{k=1}^{n} X_n, \qquad \bar{X}_n = \frac{1}{n}\textstyle\sum_{n=1}^{n} X_k.$$

Assume that n is large. What you can say about the distributions, expectation, and variance of Z_n and $\bar{X}_n$? Consider the case $n = 100,000$ and $p = 1/2$. Use both the Central Limit Theorem and the Tchebychev inequality to estimate the probability that Z_n lies between $49,500$ and $50,500$. What you can say about the distributions, expectation, and variance of Z_n and $\bar{X}_n$ if $(X_n)_{n\geq 1}$ is a sequence of independent and Poisson distributed random variables with the same rate parameter λ?

Solution. Independently of n, the random variable Z_n has the binomial distribution with parameters n and p. In symbols

$$Z_n \sim Bin\,(n, p)\,.$$

As a consequence,

$$\mathbf{E}\,[Z_n] = np \quad \text{and} \quad \mathbf{D}^2\,[Z_n] = np\,(1-p)\,.$$

By virtue of the Central Limit Theorem, we know that

$$\frac{Z_n - np}{\sqrt{np\,(1-p)}} \overset{\mathbf{w}}{\to} N\,(0,1)\,.$$

Otherwise saying: as n is large, the distribution of $(Z_n - np)\,/\sqrt{np\,(1-p)}$ is approximately the standard normal distribution. On the other hand, we can write

$$\sqrt{\frac{p\,(1-p)}{n}}\frac{Z_n - np}{\sqrt{np\,(1-p)}} = \frac{Z_n}{n} - p = \bar{X}_n - p,$$

that is

$$\bar{X}_n = \sqrt{\frac{p\,(1-p)}{n}}\frac{Z_n - np}{\sqrt{np\,(1-p)}} + p$$

This implies that, as n is large, the distribution of $\bar{X}_n$ is approximately normal with

$$\mathbf{E}\left[\bar{X}_n\right] = p \quad \text{and} \quad \mathbf{D}^2\left[\bar{X}_n\right] = \frac{p\,(1-p)}{n}.$$

In the case $n = 100,000$ and $p = 1/2$, thanks to the Central Limit Theorem, we can write

$$\begin{aligned}
\mathbf{P}\left(49,500 \leq Z_n \leq 50,500\right) &= \mathbf{P}\left(49,500 - np \leq Z_n - np \leq 50,500 - np\right) \\
&= \mathbf{P}\left(\frac{49,500 - np}{\sqrt{np\left(1-p\right)}} \leq \frac{Z_n - np}{\sqrt{np\left(1-p\right)}} \leq \frac{50,500 - np}{\sqrt{np\left(1-p\right)}}\right) \\
&\simeq \Phi\left(\frac{50,500 - np}{\sqrt{np\left(1-p\right)}}\right) - \Phi\left(\frac{49,500 - np}{\sqrt{np\left(1-p\right)}}\right) \\
&= \Phi\left(\frac{500}{\sqrt{25,000}}\right) - \Phi\left(\frac{-500}{\sqrt{25,000}}\right) \\
&= 2\Phi\left(\sqrt{10}\right) - 1 \\
&= 2 \cdot 0.9992 - 1 \\
&= 0.9984.
\end{aligned}$$

where Φ is the distribution function of the standard normal. Instead, with the goal of applying the Tchebychev inequality, we can write

$$\begin{aligned}
\mathbf{P}\left(49,500 \leq Z_n \leq 50,500\right) &= \mathbf{P}\left(\frac{49,500 - np}{\sqrt{np\left(1-p\right)}} \leq \frac{Z_n - np}{\sqrt{np\left(1-p\right)}} \leq \frac{50,500 - np}{\sqrt{np\left(1-p\right)}}\right) \\
&= \mathbf{P}\left(-\sqrt{10} \leq \frac{Z_n - 50,000}{50\sqrt{10}} \leq \sqrt{10}\right) \\
&= \mathbf{P}\left(\left|\frac{Z_n - 50,000}{50\sqrt{10}}\right| \leq \sqrt{10}\right) \\
&= 1 - \mathbf{P}\left(\left|\frac{Z_n - 50,000}{50\sqrt{10}}\right| > \sqrt{10}\right) \\
&= 1 - \mathbf{P}\left(\left|\frac{Z_n - 50,000}{50\sqrt{10}}\right| \geq \sqrt{10}\right).
\end{aligned}$$

On the other hand, by the Tchebychev inequality we have

$$\mathbf{P}\left(\left|\frac{Z_n - 50,000}{50\sqrt{10}}\right| \geq \sqrt{10}\right) \leq \frac{\mathbf{D}^2\left[\frac{Z_n - 50,000}{50\sqrt{10}}\right]}{10} = \frac{1}{10}.$$

Therefore,

$$\mathbf{P}\left(49,500 \leq Z_n \leq 50,500\right) \geq 1 - \frac{1}{10} = \frac{9}{10} = 0.9.$$

This shows that the central limit aproach provides a sharper bound for the desired probability than the Tchebychev inequality approach.

Problem 2 *Suppose that a random variable X, which represents the reaction time at some stimulus, has a uniform distribution on an interval $[0, \theta]$, where the parameter $\theta > 0$ is unknown. An investigator wants to estimate θ on the basis of a simple random sample $X_1, \ldots, X_n$ of reaction times. Since θ is the largest possible time in the entire population of reaction times, the investigator consider as a first estimator for the parameter θ the largest sample reaction time. That is to say, the investigator consider as a first estimator the statistic*

$$\hat{\theta}_1 \equiv \check{X}_n \equiv \max\left(X_1, \ldots, X_n\right).$$

1. *Is $\check{X}_n$ unbiased? In case $\check{X}_n$ is not unbiased, is it possible to derive from $\check{X}_n$ an unbiased estimator of θ?*

2. *As a second estimator, the investigator consider the statistic*

$$\hat{\theta}_2 \equiv \bar{X}_n \equiv \frac{1}{n}\textstyle\sum_{k=1}^n X_k.$$

Is $\bar{X}_n$ unbiased? In case $\bar{X}_n$ is not unbiased, is it possible to derive from $\bar{X}_n$ an unbiased estimator of θ?

3. *In the investigator's shoes, what estimator would you prefer among those considered?*

Solution.

1. Writing $F_{\check{X}_n} : \mathbb{R} \to \mathbb{R}$ for the distribution function of the statistic $\check{X}_n$, we have

$$\begin{aligned} F_{\check{X}_n}(x) &= \mathbf{P}\left(\check{X}_n \le x\right) = \mathbf{P}\left(X_1 \le x, \dots, X_n \le x\right) = \textstyle\prod_{k=1}^n \mathbf{P}\left(X_k \le x\right) \\ &= \textstyle\prod_{k=1}^n \mathbf{P}\left(X \le x\right) = \mathbf{P}\left(X \le x\right)^n = F_X(x)^n. \end{aligned}$$

for every $x \in \mathbb{R}$, where $F_X : \mathbb{R} \to \mathbb{R}$ is the distribution function of the random variable X. On the other hand, since X is uniformly distributed on $[0, \theta]$, it has density $f_X : \mathbb{R} \to \mathbb{R}$ given by

$$f_X(x) \stackrel{\text{def}}{=} \frac{1}{\theta} 1_{[0,\theta]}(x), \quad \forall x \in \mathbb{R}.$$

Hence,

$$\begin{aligned} F_X(x) &= \int_{(-\infty,x]} f_X(u)\ d\mu_L(u) = \int_{(-\infty,x]} \frac{1}{\theta} 1_{[0,\theta]}(u)\ d\mu_L(u) = \frac{1}{\theta}\int_{(-\infty,x]\cap[0,\theta]} d\mu_L(u) \\ &= \begin{cases} \frac{1}{\theta}\int_{\varnothing} d\mu_L(u) = 0 & \text{if } x < 0 \\ \frac{1}{\theta}\int_{[0,x]} d\mu_L(u) = \frac{x}{\theta} & \text{if } 0 \le x \le \theta \\ \frac{1}{\theta}\int_{[0,\theta]} d\mu_L(u) = 1 & \text{if } \theta < x \end{cases} \\ &= \frac{x}{\theta} 1_{[0,\theta]}(x) + 1_{(\theta,+\infty)}(x) \end{aligned}$$

It then follows,

$$F_{\check{X}_n}(x) = F_X(x)^n = \frac{x^n}{\theta^n} 1_{[0,\theta]}(x) + 1_{(\theta,+\infty)}(x),$$

for every $x \in \mathbb{R}$. Now, we have

$$F'_{\check{X}_n}(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{nx^{n-1}}{\theta^n} & \text{if } 0 < x < \theta \\ 0 & \text{if } \theta < x \end{cases}.$$

Note that $F_{\check{X}_n}$ is not everywhere differentable. Eventually, is not differentiable in the point $x = \theta$. However, considering the function $f_{\check{X}_n} : \mathbb{R} \to \mathbb{R}$ given by

$$f_{\check{X}_n}(x) \stackrel{\text{def}}{=} \frac{nx^{n-1}}{\theta^n} 1_{(0,\theta)}(x), \quad \forall x \in \mathbb{R},$$

a straightforward computation shows that

$$F_{\check{X}_n}(x) = \int_{(-\infty,x]} f_{\check{X}_n}(u)\ d\mu_L(u),$$

for every $x \in \mathbb{R}$. This implies that $\check{X}_n$ is absolutely continuous with density $f_{\check{X}_n}$. As a consequence,

$$\mathbf{E}\left[\check{X}_n\right] = \int_{\mathbb{R}} x f_{\check{X}_n}(x)\ d\mu_L(x) = \int_{\mathbb{R}} x \frac{nx^{n-1}}{\theta^n} 1_{(0,\theta)}(x)\ d\mu_L(x) = \frac{n}{\theta^n} \int_{(0,\theta)} x^n\ d\mu_L(x)$$
$$= \frac{n}{\theta^n} \int_0^\theta x^n\ dx = \frac{n}{\theta^n} \frac{x^{n+1}}{n+1}\bigg|_0^\theta = \frac{n}{n+1}\theta.$$

We can conclude that $\check{X}_n$ is not a unbiased estimator of θ but $\frac{n+1}{n}\check{X}_n$ is an unbiased estimator of θ.

2. We have

$$\mathbf{E}\left[\bar{X}_n\right] = \mathbf{E}[X] = \int_{\mathbb{R}} x f_X(x)\ d\mu_L(x) = \int_{\mathbb{R}} \frac{x}{\theta} 1_{[0,\theta]}(x)\ d\mu_L(x)$$
$$= \frac{1}{\theta} \int_{[0,\theta]} x\ d\mu_L(x) = \frac{1}{\theta} \int_0^\theta x\ dx = \frac{1}{\theta} \frac{x^2}{2}\bigg|_0^\theta = \frac{\theta}{2}.$$

Hence, $\bar{X}_n$ is not a unbiased estimator of θ but $2\bar{X}_n$ is an unbiased estimator of θ.

3. From 1. and 2. we know that

$$\mathbf{E}\left[\frac{n+1}{n}\check{X}_n\right] = \theta \qquad \text{and} \qquad \mathbf{E}\left[2\bar{X}_n\right] = \theta.$$

Hence, both $\frac{n+1}{n}\check{X}_n$ and $2\bar{X}_n$ are unbiased estimators of the parameter θ. To choose which is preferable between them, we consider

$$\mathbf{D}^2\left[\frac{n+1}{n}\check{X}_n\right] \qquad \text{and} \qquad \mathbf{D}^2\left[2\bar{X}_n\right].$$

We have

$$\mathbf{E}\left[\check{X}_n^2\right] = \int_{\mathbb{R}} x^2 f_{\check{X}_n}(x)\ d\mu_L(x) = \int_{\mathbb{R}} x^2 \frac{nx^{n-1}}{\theta^n} 1_{(0,\theta)}(x)\ d\mu_L(x) = \frac{n}{\theta^n} \int_{(0,\theta)} x^{n+1}\ d\mu_L(x)$$
$$= \frac{n}{\theta^n} \int_0^\theta x^{n+1}\ dx = \frac{n}{\theta^n} \frac{x^{n+2}}{n+2}\bigg|_0^\theta = \frac{n}{n+2}\theta^2.$$

Therefore,

$$\mathbf{D}^2\left[\check{X}_n\right] = \mathbf{E}\left[\check{X}_n^2\right] - \mathbf{E}\left[\check{X}_n\right]^2 = \frac{n}{n+2}\theta^2 - \frac{n^2}{(n+1)^2}\theta^2 = \frac{n}{(n+1)^2(n+2)}\theta^2.$$

As a consequence,

$$\mathbf{D}^2\left[\frac{n+1}{n}\check{X}_n\right] = \left(\frac{n+1}{n}\right)^2 \mathbf{D}^2\left[\check{X}_n\right] = \left(\frac{n+1}{n}\right)^2 \frac{n}{(n+1)^2(n+2)}\theta^2 = \frac{\theta^2}{n(n+2)}.$$

On the other hand,

$$\mathbf{D}^2\left[2\bar{X}_n\right] = 4\mathbf{D}^2\left[\bar{X}_n\right] = \frac{4}{n}\mathbf{D}^2[X] = \frac{4}{n}\frac{\theta^2}{12} = \frac{\theta^2}{3n}.$$

Now, for any $n > 1$ we clearly have

$$\mathbf{D}^2\left[\frac{n+1}{n}\check{X}_n\right] < \mathbf{D}^2\left[2\bar{X}_n\right].$$

It follows that the estimator $\frac{n+1}{n}\check{X}_n$ is preferable to $2\bar{X}_n$.

Exercise 3 *Let X be a binomially distributed real random variable with number of trials parameter m and unknown success parameter p. An investigator wants to estimate p on the basis of a simple random sample $X_1, \ldots, X_n$ of size n drawn from X.*

1. *Assume the investigator applies the method of moments. What is the estimator $\hat{p}_{M,n}$?*
2. *Assume the investigator applies the likelihood method. What is the estimator $\hat{p}_{ML,n}$?*

Solution. .

Exercise 4 *Let X be a normally distributed random variable with unknown mean μ and variance σ^2. An investigator wants to estimate μ and σ^2 on the basis of a simple random sample $X_1, \ldots, X_n$ of size n drawn from X.*

1. *Assume the investigator applies the likelihood methods. What are the estimator $\hat{\mu}_{LM}$ and $\hat{\sigma}^2_{LM}$?*
2. *Assume the investigator applies the method of moments. What are the estimators $\hat{\mu}_{MM}$ and $\hat{\sigma}^2_{MM}$? Hint: guess what $\hat{\sigma}^2_{MM}$ could be and get it!*

Solution. .

Exercise 5 *Let X a random variable representing a characteristic of a certain population. Assume that X has a density $f_X : \mathbb{R} \to \mathbb{R}$ given by*

$$f_X(x) \stackrel{def}{=} \frac{1}{\theta} e^{-\frac{x-3}{\theta}} 1_{[3,+\infty)}(x), \quad \forall x \in \mathbb{R},$$

where θ is a positive parameter.

1. *Apply the method of moments to find the estimator $\hat{\theta}_M$ of the parameter θ.*
2. *Apply the maximum likelihood method to find the estimator $\hat{\theta}_{ML}$ of the parameter θ.*
3. *Use the estimators $\hat{\theta}_M$ and $\hat{\theta}_{ML}$ to build estimators for $\mathbf{E}[X]$ and $\mathbf{D}^2[X]$.*

Solution. .

Exercise 6 *Assume that the returns of a stock in a financial market are normally distributed with unknown mean μ and variance σ^2. Let X be the normal random variable representing the realization of the returns and let $X_1, \ldots, X_n$ be a simple random sample of size n drawn from X. Assume that $n = 5$ and the realizations of the sample are*

$$x_1 \equiv -1.5, \quad x_2 \equiv -0.5, \quad x_3 \equiv 1.5, \quad x_4 \equiv 2.0, \quad x_5 \equiv 2.5$$

1. *Determine a 99% confidence interval for the mean μ.*
2. *Find the confidence for an interval of width 0.1.*

3. *Determine a* 90% *confidence interval for the standard deviation* σ.

Solution. .

Exercise 7 *Assume that a library master believes that the mean duration in days of the borrowing period is* $20d$. *However, the library master selects a simple random sample of* 100 *books in the library and discovers that the sample mean and variance of the borrowing days are* $18d$ *and* $8d^2$, *respectively. Determine a* 99% *confidence interval for the mean duration of the borrowing days to check whether library master's initial guess is correct.*

Solution. .

Exercise 8 *The mark of a infamous exam of Probability and Statistics are normally distributed with standard deviation* $\sigma = 2$. *A simple random sample of nine students is selected end the following evaluations are computed*

$$\sum_{k=1}^{9} x_k = 237 \quad \textit{and} \quad \sum_{k=1}^{9} x_k^2 = 6295.$$

1. *Find a* 90% *confidence interval for the mean mark.*
2. *Discuss, without computation, whether the lenght of a* 95% *confidence interval would be smaller, greater or equal than the lenght of the interval previously determined.*
3. *How large the minimum sample size should be to obtain a* 90% *confidence interval for the mean mark with width equal to* 3*? Besides the confidence interval method is it possible to apply the Tchebychev inequality?*

Solution. .

Exercise 9 *Let* $X_1, \ldots, X_n, X_{n+1}$ *be a simple random sample of size* $n+1$ *drawn from a Gaussian distributed random variable* X *with unknown mean* μ *and variance* σ^2. *Assume that we have observed* $X_1, \ldots, X_n$ *and we want use the observed values* $x_1, \ldots, x_n$ *to determine a confidence interval for the prediction of* X_{n+1}. *To this goal give detailed answers to the following questions:*

1. *what is the distribution of the statistic* $\bar{X}_n$*?*
2. *what is the distribution of the statistic* $\left(X_{n+1} - \bar{X}_n\right) / \sigma\sqrt{1 + 1/n}$*?*
3. *are the statistics* $X_{n+1} - \bar{X}_n$ *and* $S_n^2 \equiv \frac{1}{n-1} \sum_{k=1}^{n} \left(X_k - \bar{X}_n\right)^2$ *independent?*
4. *what is the distribution of the statistic* $\left(X_{n+1} - \bar{X}_n\right) / S_n\sqrt{1 + 1/n}$*?*

   After answering the above questions, build an interval in which the random variable X_{n+1} *takes its values with probability* α *and determine the corresponding confidence interval for the prediction of* X_{n+1}. *In the end, assume that* $n = 7$ *and we have*

   $$x_1 = 7005, \quad x_2 = 7432, \quad x_3 = 7420, \quad x_4 = 6822, \quad x_5 = 6752, \quad x_6 = 5333, \quad x_7 = 6552.$$

   compute the 95% *confidence interval for the prediction of* X_8.

Exercise 10 *Let X be a Gaussian random variable with unknown mean μ_X and variance σ_X^2 representing a certain characteristic of a population. Assume that testing the sample mean $\bar{X}_n$ and the sample standard deviation S_n of a simple random sample $X_1, \ldots, X_n$ of size $n \equiv 9$ drawn from X we obtain the value $\bar{X}_n(\omega) \equiv \bar{x}_n = 251.50cm$ and $S_n(\omega) \equiv s_n = 2.30cm$.*

1. *Considering both the rejection region method and the p-value method, should the null hypothesis $H_0 : \mu_X = 250cm$ be rejected against the alternative $H_a : \mu_X \neq 250cm$ at the significance level $\alpha = 0.1$?*
2. *Considering both the rejection region method and the p-value method, should the null hypothesis $H_0 : \sigma_X^2 = 4$ be rejected against of the alternative $H_a : \sigma_X^2 > 4$ at the significance level $\alpha = 0.05$? Calculate the probability $\beta(5)$ of a II type error.*

Solution. .

Exercise 11 *Let X be a Gaussian random variable with unknown mean μ and variance σ^2 representing a certain characteristic of a population and let $X_1, \ldots, X_n$ be a simple random sample of size n drawn from X. Assume that $n = 25$ and that the realizations $x_1, \ldots, x_{25}$ of the sample give an information summarized by*

$$\sum_{k=1}^{25} x_k = 100 \quad and \quad \sum_{k=1}^{25} x_k^2 = 560$$

1. *Considering both the rejection region method and the p-value method, should the null hypothesis $H_0 : \sigma^2 = 4$ be rejected against of the alternative $H_1 : \sigma^2 > 4$ with a significance level $\alpha = 0.05$? Calculate the probability $\beta(5)$ of a II type error.*
2. *Considering both the rejection region method and the p-value method, should the null hypothesis $H_0 : \sigma^2 = 4$ be rejected against of the alternative $H_1 : \sigma^2 \neq 4$ with a significance level $\alpha = 0.05$? Calculate the probability $\beta(5)$ of a II type error.*

Solution. .

Exercise 12 *In order to measure the depencence between two random variables X and Y a simple sample of size 10 is drawn from the random vector (X, Y) and the following quantities are computed*

$$\bar{x}_{10} = 49.6670, \quad \bar{y}_n = -0.4333, \quad s_x^2 = 236.1390, \quad s_y^2 = 0,1750, \quad \gamma_{x,y} = 5.8072.$$

1. *May you find the equation of the regression live of Y against X?*
2. *May you find the estimated mean square error?*
3. *What value of y would you predict for a corresponding value of x?*