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 - Advanced Statistics
Instructors: Roberto Monte & Massimo Regoli
Final Test - 2020-02-12 - Statistics

Problem 1 *Let X and Y two real random variables each of which represents a certain trait of a population. Assume that X and Y are independent and such that*

$$X \sim N\left(\mu_X, \sigma_X^2\right), \qquad Y \sim N\left(\mu_Y, \sigma_Y^2\right),$$

where $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2$ are known. Consider a simple random sample $X_1, \ldots, X_m$ [resp. $Y_1, \ldots, Y_n$] of size m [resp. n] drawn from X [resp. Y].

1. *Compute $\mathbf{P}\left(x \leq S_n^2(Y) \leq y\right)$ in terms of given $x, y \in \mathbb{R}_+$.*
2. *Determine x such that $\mathbf{P}\left(\bar{X}_m > x\right) = 0.25$.*
3. *Compute $\mathbf{P}\left(\bar{X}_m - x > \bar{Y}_n + y\right)$ in terms of given $x, y \in \mathbb{R}_+$.*

Solution. .

Problem 2 *Let $\theta > 0$ and let X be an absolutely continuous real random variable with density function $f_X : \mathbb{R} \to \mathbb{R}_+$ given by*

$$f_X(x) \stackrel{def}{=} \frac{1}{2} e^{-|x-\theta|}, \quad \forall x \in \mathbb{R}.$$

1. *Apply the method of moments to determine the estimator $\hat{\theta}_n^M$ for θ.*
2. *Check whether $\hat{\theta}_n^M$ is unbiased, consistent in probability, and consistent in square mean.*
3. *Can you "guess" the result of the method of maximum likelihood to determine the estimator $\hat{\theta}_n^{ML}$ for θ?*

   Hint: take for granted that the random variable X has finite moment of the first order; recall that an estimator $\hat{\theta}_n$ for the true value of a parameter θ is said to be consistent in probability [resp. in square mean] if

   $$\hat{\theta}_n \xrightarrow{\mathbf{P}} \theta \quad \text{[resp. } \hat{\theta}_n \xrightarrow{\mathbf{L}^2} \theta\text{]},$$

   as $n \to \infty$.

Solution. .

Problem 3 *Let X be a standard Bernoulli random variable with unknown success parameter p. Let $X_1, \ldots, X_n$ be a simple random sample of size n drawn from X and let $Z_n \equiv \sum_{k=1}^n X_k$ be the sample sum. It is well known that $Z_n \sim Bin(n, p)$. In addition, when n is large ($np \geq 10$ and $n(1-p) \geq 10$) the sample sum has approximately a normal distribution.*

1. *Determine a confidence interval for the parameter p with confidence level approximately $100(1-\alpha)\%$.*

2. *Determine the size n of the sample $X_1, \dots, X_n$ which allows a confidence interval for the parameter p with confidence level approximately $100\,(1-\alpha)\,\%$ and width w, where both α and w are given in advance.*

Solution. .

Problem 4 *Let $X_1, \dots, X_n, X_{n+1}$ 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 $X_1, \dots, X_n$ and we want use the observed values $x_1, \dots, x_n$ 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. *what is the distribution of the statistic $S_n^2 \equiv \frac{1}{n-1}\sum\limits_{k=1}^{n}\left(X_k - \bar{X}_n\right)^2$?*
4. *are the statistics $X_{n+1} - \bar{X}_n$ and $S_n^2 \equiv \frac{1}{n-1}\sum\limits_{k=1}^{n}\left(X_k - \bar{X}_n\right)^2$ independent? Why?*
5. *what is the distribution of the statistic $\left(X_{n+1} - \bar{X}_n\right)/S_n\sqrt{1+1/n}$?*
6. *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.

Solution. .

Problem 5 *Let X [resp. Y] be a Gaussian distributed random variables with (unknown) mean $\mu_X \in \mathbb{R}$ [resp. $\mu_Y \in \mathbb{R}$] and variance $\sigma_X^2 > 0$ [resp. $\sigma_Y^2 > 0$]. Assume that X describes a trait of some population before a treatment and Y describes the same trait of the same population after a treatment (for instance a power training period). Let $X_1, \dots, X_n$ be a simple random sample drawn by X and let $Y_1, \dots, Y_n$ be the corresponding sample drawn from Y. Note that we can still assume that $Y_1, \dots, Y_n$ is a simple random sample but we cannot assume that the samples $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$ are independent. Actually, there is no reason at all to think that the random variables X and Y are independent. However, it is still reasonable to assume that the random variable $D \equiv Y - X$ is Gaussian distributed and that $D_1 \equiv Y_1 - X_1, \dots, D_n \equiv Y_n - X_n$ is a simple random sample drawn from D. Assume to have measured*

$$\begin{array}{lllllll} x_1 = 73.80 & x_2 = 62.80 & x_3 = 73.40 & x_4 = 63.50 & x_5 = 71.90 & x_6 = 74.30 & x_7 = 63.10 \\ y_1 = 75.70 & y_2 = 63.70 & y_3 = 74.70 & y_4 = 64.40 & y_5 = 70.50 & y_6 = 74.90 & y_7 = 64.90 \end{array}.$$

1. *Should we reject the null hypothesis $H_0 = \mu_Y = \mu_X$ against the alternatives $H_1 = \mu_Y \neq \mu_X$ and $H_1 = \mu_Y > \mu_X$ at the significance level $\alpha = 0.05$? Consider both the rejection region method and the p-value method.*
2. *Assume that $\sigma_X = 5.51$ and that $\rho_{X,Y} = 0.98$. Should we reject the null hypothesis $H_0 = \sigma_Y^2 = \sigma_X^2$ against the alternatives $H_1 = \sigma_Y^2 \neq \sigma_X^2$ and $H_1 = \sigma_Y^2 < \sigma_X^2$ at the significance level $\alpha = 0.05$?*

Solution. .