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
Solved Problems on Confidence and Prediction Intervals 2023-01-17

Problem 1 *Assume that the log-returns of a stock in a financial market are Gaussian distributed with unknown mean μ and variance σ^2. Let X be the normal random variable representing the realization of the log-returns and let $X_1, \dots, X_n$ 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.

1. From data we obtain

$$\bar{x}_5 \equiv \frac{1}{5}\sum_{k=1}^{5} x_k = 0.8$$

and

$$s_{X,5}^2 = \frac{1}{4}\sum_{k=1}^{5} (x_k - \bar{x}_5)^2 = 2.95 \Rightarrow s_{X,5} = 1.72$$

Now, since X is Gaussian distributed with unknown variance, to determine a $100\,(1-\alpha)\,\%$ confidence interval for the mean μ the statistic to be considered is

$$\frac{\bar{X}_n - \mu}{S_{X,n}/\sqrt{n}} \sim t_{n-1}.$$

The achievement of a $100\,(1-\alpha)\,\%$ confidence interval requires to use the $\alpha/2$ upper and lower critical values $t^+_{n-1,\alpha/2} \equiv t_{n-1,1-\alpha/2}$ and $t^-_{n-1,\alpha/2} \equiv t_{n-1,\alpha/2} = -t_{n-1,1-\alpha/2}$ of t_{n-1} for $\alpha = 0.01$, where $t_{n-1,1-\alpha/2}$ [resp. $t_{n-1,\alpha/2}$] denotes the $(1-\alpha/2)$ [resp. $\alpha/2$]-quantile. In fact, we have

$$-t_{n-1,1-\alpha/2} < \frac{\bar{X}_n - \mu}{S_{X,n}/\sqrt{n}} < t_{n-1,1-\alpha/2} \Leftrightarrow -\left(\bar{X}_n + t_{n-1,1-\alpha/2}\frac{S_{X,n}}{\sqrt{n}}\right) < -\mu < -\left(\bar{X}_n - t_{n-1,1-\alpha/2}\frac{S_{X,n}}{\sqrt{n}}\right)$$
$$\Leftrightarrow \bar{X}_n - t_{n-1,1-\alpha/2}\frac{S_{X,n}}{\sqrt{n}} < \mu < \bar{X}_n + t_{n-1,1-\alpha/2}\frac{S_{X,n}}{\sqrt{n}},$$

which implies

$$\begin{aligned} 1-\alpha &= \mathbf{P}\left(-t_{n-1,1-\alpha/2} < \frac{\bar{X}_n - \mu}{S_{X,n}/\sqrt{n}} < t_{n-1,1-\alpha/2}\right) \\ &= \mathbf{P}\left(\bar{X}_n - t_{n-1,1-\alpha/2}\frac{S_{X,n}}{\sqrt{n}} < \mu < \bar{X}_n + t_{n-1,1-\alpha/2}\frac{S_{X,n}}{\sqrt{n}}\right). \end{aligned}$$

It follows that the desired confidence interval for μ is given by the random interval

$$\left(\bar{X}_n - t_{n-1,1-\alpha/2}\frac{S_{X,n}}{\sqrt{n}},\ \bar{X}_n + t_{n-1,1-\alpha/2}\frac{S_{X,n}}{\sqrt{n}}\right)$$

A realization of such a confidence interval is then given by

$$\left(\bar{x}_n - t_{n-1,1-\alpha/2}\frac{s_{X,n}}{\sqrt{n}},\ \bar{x}_n + t_{n-1,1-\alpha/2}\frac{s_{X,n}}{\sqrt{n}}\right).$$

In the case considered, since $t_{n-1,1-\alpha/2} \equiv t_{4,0.995} = 4.60$, $\bar{x}_n \equiv \bar{x}_5 = 0.80$, $s_{X,n} \equiv s_{X,5} = 1.72$, the realization of the confidence interval becomes

$$(-2.74,\ 4.34).$$

2. From 1. it is clearly seen the width w of a $100(1-\alpha)\%$ confidence interval is given by

$$w = 2t_{n-1,1-\alpha/2}\frac{S_{X,n}}{\sqrt{n}}.$$

As a consequence, the $100(1-\alpha)\%$ confidence of an interval having width $w = 0.1$ is given by the number $\alpha \in (0,1)$ such that

$$\frac{w\sqrt{n}}{2s_{X,n}} = t_{n-1,1-\alpha/2}.$$

With reference to our sample characterized by $n = 5$, $s_{X,n} = 1.72$ we need to determine α such that

$$\frac{0.1 \cdot \sqrt{5}}{2 \cdot 1.72} \approx 0.065 = t_{4,1-\alpha/2}.$$

Now, since

$$\mathbf{P}(T_4 \leq 0.065) \approx 0.524,$$

where T_4 is the Student random variable with 4 degrees of freedom, setting

$$1 - \alpha/2 = 0.524,$$

we obtain

$$\alpha = 0.952,$$

which yields a

$$100(1-0.952)\% = 4.8\%$$

confidence interval. Note that we could equivalently have considered the equation

$$\mathbf{P}(T_4 \geq 0.065) \approx 0.476.$$

In this case, setting

$$\alpha/2 = 0.476,$$

we still would have ended up with

$$\alpha = 0.952.$$

3. Again, since X is Gaussian distributed, to determine a $100(1-\alpha)\%$ confidence interval for the standard deviation σ the statistic to be considered is

$$\frac{(n-1)S_{X,n}^2}{\sigma^2} \sim \chi_{n-1}^2.$$

Since χ_{n-1}^2 is not symmetric, the achievement of a $100(1-\alpha)\%$ confidence interval requires to use the $\alpha/2$ and the $1-\alpha/2$ critical value $\chi_{n-1,\alpha/2}^{2,-} \equiv \chi_{n-1,\alpha/2}^2$ and $\chi_{n-1,\alpha/2}^{2,+} = \chi_{n-1,1-\alpha/2}^2$ of χ_{n-1}^2

for $\alpha = 0.1$, where $\chi^2_{n-1,\alpha/2}$ [resp. $\chi^2_{n-1,1-\alpha/2}$] is the $\alpha/2$-quantile [$1-\alpha/2$-quantile] of the χ^2_{n-1} distribution. In fact, we have

$$\chi^{2,-}_{n-1,\alpha/2} < \frac{(n-1)\, S^2_{X,n}}{\sigma^2} < \chi^{2,+}_{n-1,\alpha/2} \Leftrightarrow \frac{1}{\chi^{2,-}_{n-1,\alpha/2}} > \frac{\sigma^2}{(n-1)\, S^2_{X,n}} > \frac{1}{\chi^{2,+}_{n-1,\alpha/2}}$$
$$\Leftrightarrow \frac{(n-1)\, S^2_{X,n}}{\chi^{2,+}_{n-1,\alpha/2}} < \sigma^2 < \frac{(n-1)\, S^2_{X,n}}{\chi^{2,-}_{n-1,\alpha/2}},$$

which implies

$$1-\alpha = \mathbf{P}\left(\chi^{2,-}_{n-1,\alpha/2} < \frac{(n-1)\, S^2_{X,n}}{\sigma^2} < \chi^{2,+}_{n-1,\alpha/2}\right) = \mathbf{P}\left(\frac{(n-1)\, S^2_{X,n}}{\chi^{2,+}_{n-1,\alpha/2}} < \sigma^2 < \frac{(n-1)\, S^2_{X,n}}{\chi^{2,-}_{n-1,\alpha/2}}\right).$$

It follows that the desired confidence interval for the variance σ^2 is given by

$$\left(\frac{(n-1)\, S^2_{X,n}}{\chi^{2,+}_{n-1,\alpha/2}} < \sigma^2 < \frac{(n-1)\, S^2_{X,n}}{\chi^{2,-}_{n-1,\alpha/2}}\right) = \left(\frac{(n-1)\, S^2_{X,n}}{\chi^{2}_{n-1,1-\alpha/2}},\ \frac{(n-1)\, S^2_{X,n}}{\chi^{2}_{n-1,\alpha/2}}\right).$$

In the case considered, since $\chi^2_{n-1,\alpha/2} \equiv \chi^2_{4,0.5} = 0.71$, $\chi^2_{n-1,1-\alpha/2} \equiv \chi^2_{4,0.95} = 9.49$, $\bar{x}_n \equiv \bar{x}_5 = 0.80$, $s^2_{X,n} \equiv s^2_{X,5} = 2.95$, a realization of the confidence interval is given by

$$\left(\frac{4s^2_{X,n}}{\chi^2_{4,0.95}},\ \frac{4s^2_{X,n}}{\chi^2_{4,0.05}}\right) = \left(\frac{4\cdot 2.95}{9.49},\ \frac{4\cdot 2.95}{0.71}\right) = (1.24,\ 16.62)\,.$$

As a consequence the $100\,(1-0.1)\,\%$ confidence interval for the standard deviation σ is

$$(1.11,\ 4.08)\,.$$

This completes the solution.

Problem 2 *Assume that a library master believes that the mean duration in days of the borrowing period is* $20d$. *However, the library master selects randomly 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. Note that the distribution of the random variable X representing the duration in days of the borrowing period is unknown. However, are known the sample mean and variance realizations referred to a simple sample of size $n = 100$, which may be considered a large sample. In this case the statistic to be considered is given by

$$\frac{\bar{X}_n - \mu}{S_{X,n}/\sqrt{n}},$$

which is approximatively distributed as a standard Gaussian random variable $Z \sim N(0,1)$. The achievement of a $100\,(1-\alpha)\,\%$ confidence interval requires to use the $\alpha/2$ upper and lower critical values $z^+_{\alpha/2} \equiv z_{1-\alpha/2}$ and $z^-_{\alpha/2} \equiv z_{\alpha/2} = -z_{1-\alpha/2}$ of Z for $\alpha = 0.01$, where $z_{1-\alpha/2}$ [resp. $z_{\alpha/2}$] denotes the $(1-\alpha/2)$ [resp. $\alpha/2$]-quantile. In fact, we have

$$-z_{1-\alpha/2} < \frac{\bar{X}_n - \mu}{S_{X,n}/\sqrt{n}} < z_{1-\alpha/2} \Leftrightarrow -\left(\bar{X}_n + z_{1-\alpha/2}\frac{S_{X,n}}{\sqrt{n}}\right) < -\mu < -\left(\bar{X}_n - z_{1-\alpha/2}\frac{S_{X,n}}{\sqrt{n}}\right)$$
$$\Leftrightarrow \bar{X}_n - z_{1-\alpha/2}\frac{S_{X,n}}{\sqrt{n}} < \mu < \bar{X}_n + z_{1-\alpha/2}\frac{S_{X,n}}{\sqrt{n}},$$

whuihc implies

$$1-\alpha \approx \mathbf{P}\left(-z_{1-\alpha/2} < \frac{\bar{X}_n-\mu}{S_{X,n}/\sqrt{n}} < z_{1-\alpha/2}\right)$$
$$= \mathbf{P}\left(\bar{X}_n - z_{1-\alpha/2}\frac{S_{X,n}}{\sqrt{n}} < \mu < \bar{X}_n + z_{1-\alpha/2}\frac{S_{X,n}}{\sqrt{n}}\right)$$

It follows that the desired confidence interval for μ is given by

$$\left(\bar{X}_n - z_{1-\alpha/2}\frac{S_{X,n}}{\sqrt{n}},\ \bar{X}_n + z_{1-\alpha/2}\frac{S_{X,n}}{\sqrt{n}}\right).$$

In the case considered,

$$n = 100, \quad \bar{x}_n \equiv \bar{x}_{100} = 18, \quad s_{X,n} \equiv s_{X,100} = \sqrt{8}, \quad z_{1-\alpha/2} \equiv 2.58.$$

Therefore, a realization of the confidence interval is given by

$$\left(\bar{x}_n - z_{1-\alpha/2}\frac{s_{X,n}}{\sqrt{n}},\ \bar{x}_n + z_{1-\alpha/2}\frac{s_{X,n}}{\sqrt{n}}\right) = \left(18 - 2.58\cdot\frac{\sqrt{8}}{\sqrt{100}}, 18 + 2.58\cdot\frac{\sqrt{8}}{\sqrt{100}}\right) = (17.27,\ 18.73)\,.$$

It follows that library master's initial guess is not supported by data. Note that this problem can be tackled also exploiting the hypothesis test method. In fact, assume as the null hypothesis that library master's assumption is correct, that is $H_0 : \mu = \mu_0$, and as the alternative hypothesis that library master's assumption is wrong, that is $H_0 : \mu \neq \mu_0$. The same consideration as above on the available information on the random variable X led to consider the rejection region

$$R = \left\{\frac{\bar{X}_n-\mu}{S_{X,n}/\sqrt{n}} < z_{\alpha/2}\right\} \cup \left\{\frac{\bar{X}_n-\mu}{S_{X,n}/\sqrt{n}} > z_{1-\alpha/2}\right\} = \left\{\frac{\bar{X}_n-\mu}{S_{X,n}/\sqrt{n}} < -2.58\right\} \cup \left\{\frac{\bar{X}_n-\mu}{S_{X,n}/\sqrt{n}} > 2.58\right\},$$

where $\mu = 20$. Computing the statistic $\frac{\bar{X}_n-\mu}{S_{X,n}/\sqrt{n}}$ for the available realization, we obtain

$$\frac{\bar{x}_n-\mu}{s_{X,n}/\sqrt{n}} = \frac{18-20}{\sqrt{8}/10} = -7.07 \in R.$$

Hence, the library master's assumption has to be rejected.

Problem 3 *A sample of* 60 *cars of the same model are tested for gasoline consumption, expressed as litres per* 100 *kilometers ($L/100km$). The result of the test yield a mean consumption μ of* 9.4 $L/100km$ *and a standard deviation σ of* 1.5 $L/100km$.

1. *Determine a* 99% *confidence interval for the mean consumption. Is it necessary to make any assumption on the consumption distribution?.*

2. *Assume that the consumption is normally distributed with variance $\sigma^2 = 2$ $L/100km$. How large has to be the sample if, with the same confidence, we want the maximum error to be* 0.25 $L/100km$*? Apply the method of the confidence interval and the Chebychev inequality.*

Solution. .

Problem 4 *Let X [resp. Y] 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 population before a treatment and Y describes the same population after a treatment. 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.*

1. *Given $\alpha > 0$, can you build a $100(1-\alpha)\%$ confidence interval for the difference $\mu_Y - \mu_X$?*
2. *Assume to have measured*

$$\begin{array}{ccccccc} x_1 = 3.85 & x_2 = 2.82 & x_3 = 3.44 & x_4 = 3.48 & x_5 = 1.92 & x_6 = 4.39 & x_7 = 3.12 \\ y_1 = 5.73 & y_2 = 3.84 & y_3 = 4.78 & y_4 = 4.40 & y_5 = 1.91 & y_6 = 4.98 & y_7 = 4.94 \end{array}.$$

Determine a realization of the 95% confidence interval built above.

Solution. We clearly have

$$\mu_D \equiv \mathbf{E}[D] \equiv \mathbf{E}[Y - X] = \mathbf{E}[Y] - \mathbf{E}[X] = \mu_Y - \mu_X.$$

Therefore, introducing the sample mean of size n drawn from D, that is

$$\bar{D}_n = \frac{1}{n}\sum_{k=1}^{n} D_k = \frac{1}{n}\sum_{k=1}^{n}(Y_k - X_k) = \frac{1}{n}\sum_{k=1}^{n} Y_k - \frac{1}{n}\sum_{k=1}^{n} X_k = \bar{Y}_n - \bar{X}_n$$

and the unbiased sample variance of size n drawn from D, that is

$$S_{n,D}^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(D_k - \bar{D}_n\right)^2,$$

under the assumptions considered, we have that the statistic

$$\frac{\bar{D}_n - \mu_D}{S_{n,D}/\sqrt{n}},$$

has the Student distribution with $n-1$ degrees of freedom. As a consequence, given $\alpha > 0$ a $100(1-\alpha)\%$ confidence interval for μ_D is given by

$$\left(\bar{D}_n - t^-_{\frac{\alpha}{2},n-1}\frac{S_{n,D}}{\sqrt{n}}, \bar{D}_n + t^+_{\frac{\alpha}{2},n-1}\frac{S_{n,D}}{\sqrt{n}}\right).$$

The realization of such a confidence interval are of the form

$$\left(\bar{d}_n - t_{\frac{\alpha}{2},n-1}\frac{s_{n,D}}{\sqrt{n}}, \bar{d}_n + t_{\frac{\alpha}{2},n-1}\frac{s_{n,D}}{\sqrt{n}}\right),$$

where $\bar{d}_n$ [resp. $s_{n,D}$] is the value taken by the sample mean estimator $\bar{D}_n$ [resp. unbiased sample standard deviation estimator $S_{n,D}$] on the available realizations $d_1, \dots, d_n$ of the sample $D_1, \dots, D_n$. Since in our case $n = 7$ and $\alpha \equiv 0.05$, we have

$$t_{\frac{\alpha}{2},n-1} \equiv t_{0.025,6} = 2.45$$

Furthermore,

$$\bar{d}_n \equiv \bar{d}_7 = 1.08 \quad \text{and} \quad s_{n,D} \equiv s_{7,D} = 0.67$$

Therefore,

$$\bar{d}_n - t_{\frac{\alpha}{2},n-1}\frac{s_{n,D}}{\sqrt{n}} = 1.08 - 2.45\frac{0.67}{\sqrt{7}} = 0.46$$

and

$$\bar{d}_n + t_{\frac{\alpha}{2},n-1}\frac{s_{n,D}}{\sqrt{n}} = 1.08 + 2.45\frac{0.67}{\sqrt{7}} = 1.70.$$

Thus the realization of the confidence interval is

$$(0.46,\ 1.70)\,.$$

Problem 5 *Two independent groups of* 50 *people, say group A and group B, are affected by a disease. A drug is given to the individuals of group A (*treatment group*) but not to the individuals of group B (*control group*). After a week since the administration of the drug a medical check shows that* 45 *[resp.* 30*] people of group A [resp. B] got recovered from the disease. Let p_A [resp. p_B] the proportion of inividuals of group A [resp. B] who got recovered.*

1. *Given $\alpha > 0$, can you build a $100\,(1-\alpha)\,\%$ confidence interval for the true value of difference $p_A - p_B$?*
2. *Determine the realization of the* 95% *confidence interval built above and use it to comment on the efficacy of the drug.*

Solution. .

Problem 6 *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 $X_1, \ldots, X_n$ 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 7 *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_X and variance σ_X^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. *what is the distribution of the statistic* $S^2_{X,n} \equiv \frac{1}{n-1} \sum_{k=1}^{n} \left(X_k - \bar{X}_n\right)^2$*?*

4. *are the statistics* $X_{n+1} - \bar{X}_n$ *and* $S^2_{X,n}$ *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.

1. Since $X_1, \ldots, X_n$ is a simple random sample of size n drawn from a Gaussian distributed random variable X, the sample mean $\bar{X}_n$ is Gaussian distributed with mean $\mu_{\bar{X}_n} = \mu_X$ and variance $\sigma^2_{\bar{X}_n} = \frac{1}{n}\sigma^2_X$. In symbols, $\bar{X}_n \sim N\left(\mu_X, \frac{1}{n}\sigma^2_X\right)$.

2. Since X_{n+1} is independent of $X_1, \ldots, X_n$ and has the same distribution of X, the random variable $U \equiv X_{n+1} - \bar{X}_n$ is Gaussian distributed with mean $\mu_U = \mu_{X_{n+1}} - \mu_{\bar{X}_n} = \mu_X - \mu_X = 0$ and variance $\sigma^2_U = \sigma^2_{X_{n+1}} + \sigma^2_{\bar{X}_n} = \sigma^2_X + \frac{1}{n}\sigma^2_X = (1+1/n)\,\sigma^2_X$. In symbols, $U \sim N\left(0, \left(1+\frac{1}{n}\right)\sigma^2_X\right)$. It then follows that the statistic

$$Z \equiv U/\sigma_X\sqrt{1+1/n} = \left(X_{n+1} - \bar{X}_n\right)/\sigma_X\sqrt{1+1/n}$$

is standard Gaussian distributed. In symbols, $Z \sim N(0,1)$.

3. Since $X_1, \ldots, X_n$ is a simple random sample of size n drawn from a Gaussian distributed random variable X, we know that the statistic

$$W \equiv (n-1)\,S^2_{X,n}/\sigma^2_X$$

has a chi-square distribution with $n-1$ degrees of freedom. In symbols, $W \equiv (n-1)\,S^2_{X,n}/\sigma^2_X \sim \chi^2_{n-1}$.

4. Since X_{n+1} is independent of $X_1, \ldots, X_n$, it is also independent of $S^2_{X,n}$, which is a function of $X_1, \ldots, X_n$, Furthermore, we know that $\bar{X}_n$ and $S^2_{X,n}$ are independent. It then follow that $X_{n+1} - \bar{X}_n$, which is a function of X_{n+1} and $\bar{X}_n$ is independent of $S^2_{X,n}$.

5. Since $U \equiv X_{n+1} - \bar{X}_n$ is independent of $S^2_{X,n}$, also $Z \equiv U/\sigma_X\sqrt{1+1/n}$ is independent of $W \equiv (n-1)\,S^2_{X,n}/\sigma_X$. We then obtain that the statistic

$$\frac{Z}{\sqrt{W/(n-1)}} = \frac{U/\sigma_X\sqrt{1+1/n}}{\sqrt{(n-1)\,S^2_{X,n}/\sigma^2_X/(n-1)}} = \frac{X_{n+1} - \bar{X}_n}{S_{X,n}\sqrt{1+1/n}}$$

has the Student distribution with $n-1$ degrees of freedom. In symbols, $\frac{X_{n+1} - \bar{X}_n}{S_{X,n}\sqrt{1+1/n}} \sim t_{n-1}$. As a consequence, we can write

$$\mathbf{P}\left(\left|\frac{X_{n+1} - \bar{X}_n}{S_{X,n}\sqrt{1+\frac{1}{n}}}\right| < t_{\alpha/2,n-1}\right) \geq 1 - \alpha.$$

where $t_{\alpha/2,n-1}$ is the upper tail critical value of level $\alpha/2$ of the Student distribution with $n-1$ degrees of freedom. On the other hand,

$$\left|\frac{X_{n+1}-\bar{X}_n}{S_{X,n}\sqrt{1+\frac{1}{n}}}\right| < t_{\alpha/2,n-1} \Leftrightarrow -t_{\alpha/2,n-1} < \frac{X_{n+1}-\bar{X}_n}{S_{X,n}\sqrt{1+\frac{1}{n}}} < t_{\alpha/2,n-1}$$

$$\Leftrightarrow \bar{X}_n - t_{\alpha/2,n-1}S_{X,n}\sqrt{1+\frac{1}{n}} < X_{n+1} < \bar{X}_n - t_{\alpha/2,n-1}S_{X,n}\sqrt{1+\frac{1}{n}}.$$

Therefore, a $100\,(1-\alpha)\,\%$ confidence interval for X_{n+1} is given by the random interval

$$\left(\bar{X}_n - t_{\alpha/2,n-1}S_{X,n}\sqrt{1+\frac{1}{n}},\ \bar{X}_n - t_{\alpha/2,n-1}S_{X,n}\sqrt{1+\frac{1}{n}}\right).$$

A realization of such an interval is given by

$$\left(\bar{x}_n - t_{\alpha/2,n-1}s_{X,n}\sqrt{1+\frac{1}{n}},\ \bar{x}_n - t_{\alpha/2,n-1}s_{X,n}\sqrt{1+\frac{1}{n}}\right)$$

where, in the case considered, $n=7$, $\bar{x}_n = 6759.429$, $s_{X,n} = 710.716$, $\alpha = 0.05$, $t_{\alpha/2,n-1} = 2.447$. We then obtain

$$(4900.229,\ 8618.629)$$

as prediction interval for x_8.

Problem 8 *A test on the reaction time measured in seconds to a sudden emergency has lead to the following results in* 10 *people:*

$$t_1 = 0.77,\quad t_2 = 0.75,\quad t_3 = 0.70,\quad t_4 = 0.72,\quad t_5 = 0.70,$$
$$t_6 = 0.69,\quad t_7 = 0.67,\quad t_8 = 0.79,\quad t_9 = 0.64,\quad t_{10} = 0.72.$$

Assume that the reaction time can be modelled by a normal random variable $T \sim N\left(\mu,\sigma^2\right)$ with unknown μ and σ^2.

1. *Compute the sample mean and variance referred to the above sample.*
2. *Find the confidence interval for μ [resp. σ^2] at the confidence level* 95%.
3. *What would the confidence interval for μ be if the variance were known and we had $\sigma^2 = 0.0025$?*
4. *What should the size of the sample be to achieve a precision of* 10*?*

Solution. .

Problem 9 *Assume we need to measure the same trait X in two different population and the results of our measurement for a sample of size $n_1 = 15$ [resp. $n_2 = 20$] of the first [resp. second] population gives a value $\bar{x}_{n_1} = 24.0$ [resp. $\bar{x}_{n_2} = 26.0$] of the sample mean μ_X [resp. μ_Y] of the characteristic under investigation, with a sample variance $s^2_{n_1} = 4.5$ [resp. $s^2_{n_2} = 5.0$]. Assume that the trait is normally distributed with the same unknown variance σ^2_X. Computed a confidence interval for the value of the difference $\mu_X - \mu_Y$ with a* 95% *confidence level.*

Solution. .