Interim Test

MA1202, Introductory Statistics

17/03/2021

General information

Please upload your work to Blackboard as a single pdf document through the link "Upload your work here". Read the **Instructions on Scanning and Uploading handwritten work**. Please name your file *YourStudentNumber_InterimTest.pdf*.

Question 1 [10 points]

A pharmaceutical company wants to know whether an experimental drug has a stabilising effect on blood pressure. Twenty randomly selected subjects were given the drug and after sufficient time for the drug to have an impact, their systolic blood pressures were recorded.

Table 1. Systolic blood pressure measurements

106	118	118	123	128
128	129	130	132	133
136	136	137	140	140
143	147	152	161	184

i) Find the sample mean, variance and standard deviation for x. (Show your working)

ii) Find inter-quartile range. (Show your working)

iii) Construct a frequency distribution for the data, taking as your grouping intervals the following 9 intervals:

$$(0, 105], (105, 115], ...(165, 175], (175, \infty).$$

iv) You should observe that some values are too far from other values of the sample. To investigate whether they are outliers perform the following steps (called the inter-quartile fence procedure):

1. Use the IQR value, you calculated at question ii), to find the following values:

   * $Q_1 - 1.5 \cdot IQR$ and $Q_3 + 1.5 \cdot IQR$, called lower and upper *inner* fences, respectively;
   * $Q_1 - 3 \cdot IQR$ and $Q_3 + 3 \cdot IQR$, called lower and upper *outer* fences, respectively;

2. The data points, that fall inside the inner fences, are not outliers. The points, between the inner and outer fences, are *outliers*, and the values outside the outer fences, are *extreme outliers*.
   Find all the outliers and extreme outliers in the dataset.

Illustrate the positions of outer fences, inner fences, Q_1, Q_3 and the median on the real line (you do not need to preserve the scale, just indicate relative positions).

The α-trimmed mean is used to eliminate influence of outliers on the estimate of mean value.

Suggest the value for α to be used to obtain the α-trimmed mean.

Calculate the α-trimmed mean for the sample and the sample standard deviation for the trimmed sample, write your observations.

Solution:

i)(2 points) Sample mean - $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\bar{x} = \frac{1}{20}\sum_{i=1}^{20} x_i = 136.05$.

Sample variance - $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, $s^2 = \frac{1}{20-1}\sum_{i=1}^{20}(x_i - \bar{x})^2 = 281.21$ (2 d.p.).

Sample standard deviation - $s = \sqrt{s^2} = 16.77$ (2 d.p.).

ii)(1 points)

$$Q_1 = \frac{x_{(\lfloor 1+(n-1)/4 \rfloor)} + x_{(\lceil 1+(n-1)/4 \rceil)}}{2},$$

$$Q_3 = \frac{x_{(\lfloor 1+3(n-1)/4 \rfloor)} + x_{(\lceil 1+3(n-1)/4 \rceil)}}{2},$$

$$\text{IQR} = Q_3 - Q_1.$$

Hence,

$IQR = Q_3 - Q_1 = \frac{x_{(16)}+x_{(15)}}{2} - \frac{x_{(6)}+x_{(5)}}{2} = 141.5 - 128 = 13.5$, where $x_{(i)}$ is corresponding ith statistics.

iii)(3 points)

Range	Frequency	Relative Frequency
$(0, 105]$	0	0.0
$(105, 115]$	1	0.05
$(115, 125]$	3	0.15
$(125, 135]$	6	0.3
$(135, 145]$	6	0.3
$(145, 155]$	2	0.1
$(155, 165]$	1	0.05
$(165, 175]$	0	0.0
$(175, \infty)$	1	0.05

iv)(4 points)

Lower inner fence= $Q_1 - 1.5IQR = 128 - 1.5 \cdot 13.5 = 107.75$;
Lower outer fence= $Q_1 - 3IQR = 128 - 3 \cdot 13.5 = 87.5$;
Upper inner fence= $Q_3 - 1.5IQR = 141.5 + 1.5 \cdot 13.5 = 161.75$;
Upper outer fence= $Q_3 - 3IQR = 141 + 3 \cdot 13.5 = 182$;

There are no values below the Lower outer fence, however, the value (106) is outside the Lower inner fence, i.e. it is an outlier.

There is one value (184) larger than Upper outer fence and no values between Upper inner fence and Upper outer fence.

An α**-trimmed mean** - the mean of the remaining middle $100(1-2\alpha)\%$ of the observations.

$$\bar{x}_\alpha := \frac{x_{([n\alpha]+1)} + x_{([n\alpha]+2)} + ...x_{(n-[n\alpha])}}{n-2[n\alpha]} = \frac{\sum_{i=[n\alpha]+1}^{n-[n\alpha]} x_{(i)}}{n-2[n\alpha]},$$

where $[s]$ is the greatest integer $\leq s$, $x_{(i)}$ - ith order statistics.

Therefore, to remove the values 106 and 184, we can choose
$\alpha = 1/20 \cdot 100\% = 5\%$ for trimming the subset; the new data set is :

$$118, 118, 123, 128, 128, 129, 130, 132, 133, 136, 136, 137, 140, 140, 143, 147, 152, 161$$

The 5%-trimmed mean is 135.06 and the sample standard deviation is 11.17.

These values are different to the ones found in question i), which reflects the fact that sample mean and sample variance are sensitive to outliers.

Question 2 [5 points]

Let $Y_1, Y_2, ..., Y_n$ denote a random sample from the uniform distribution on the interval $[0, \theta]$ with the probability density function of the form:

$$f_Y(y) = \frac{1}{\theta}, \quad 0 \leqslant y \leqslant \theta.$$

i) Using the first two population moments, find the mean and variance of Y (Show your working);

ii) Find estimator for θ using method of moments;

iii) Show that this estimator is unbiased and consistent (Show your working).

Solution:

i)(2 points) Mean of the population:

$$\mu = E(Y) = \int_{-\infty}^{\infty} y f_Y(y) dy = \int_0^\theta y \frac{1}{\theta} dy = \frac{y^2}{2\theta}|_0^\theta = \frac{\theta}{2}$$

Variance of the population:

$$\mu_2 = E(Y^2) = \int_{-\infty}^{\infty} y^2 f_Y(y) dy = \int_0^\theta y^2 \frac{1}{\theta} dy = \frac{y^3}{3\theta}|_0^\theta = \frac{\theta^2}{3}$$

$$Var(Y) = E(Y^2) - E(Y)^2 = \frac{\theta^2}{3} - \frac{\theta^2}{4} = \frac{\theta^2}{12}$$

ii) (1 point) The first sample moment $m_1 = \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. So the method of moments estimator is:

$$\frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{\theta}{2} \quad \Rightarrow \hat{\theta} = 2\bar{Y}$$

iii) (2 points) By definition, the estimator is unbiased if $E(\hat{\theta}) = \theta$.

$$E(\hat{\theta}) = E(2\bar{Y}) = 2E(\frac{1}{n}\sum_{i=1}^{n} Y_i) = 2\frac{1}{n}\sum_{i=1}^{n} E(Y_i) = 2\frac{1}{n}n\frac{\theta}{2} = \theta,$$

hence, the estimator $\hat{\theta} = 2\bar{Y}$ is unbiased.

The estimator is consistent if it converges in probability to θ, i.e. for $\varepsilon > 0$:

$$\lim_{n\to\infty} P\left(|\hat{\theta} - \theta| < \varepsilon\right) = 1$$

Using Markov's theorem:

$$\begin{aligned} P\left(|\hat{\theta} - \theta| < \varepsilon\right) &= P\left(|2\bar{Y} - \theta| < \varepsilon\right) = 1 - P\left(|2\bar{Y} - \theta| \geq \varepsilon\right) = 1 - P\left(|\bar{Y} - \theta/2| \geq \varepsilon/2\right) \\ &= 1 - P\left((\bar{Y} - \theta/2)^2 \geq \varepsilon^2/4\right) \geq 1 - \frac{E((\bar{Y}-\theta/2)^2)}{\varepsilon^2/4} = 1 - \frac{\mathrm{Var}(\bar{Y})}{\varepsilon^2/4} \\ &= 1 - \frac{\sum_{i=1}^{n}\mathrm{Var}(Y_i)}{n^2\varepsilon^2/4} = 1 - \frac{n\cdot\theta^2/12}{n^2\varepsilon^2/4} = 1 - \frac{\theta^2}{3n\varepsilon^2} \quad \to \ 1 \ \text{ as } \ n \to \infty. \end{aligned}$$

Thus, the estimator $\hat{\theta} = 2\bar{Y}$ is consistent.

Question 3 [5 points]

The life of light bulbs manufactured at a certain plant can be assumed to be normally distributed. A sample of 30 light bulbs gives the observed sample mean $\bar{x} = 1500$ hours and the observed sample standard deviation $s = 50$ hours.

i) Determine a two-sided 95% confidence interval for the average life of the light bulbs. Interpret the result.

ii) Determine **two-sided** and **one-sided** (find upper boundary, $P(\sigma^2 < B) = 1 - \alpha$) 90% confidence intervals for the variance of life of the light bulbs.
Which interval is wider, i.e. less precise?

Solution:

i) As we assume that random sample is obtained from normal distribution the average value also has normal distribution $\bar{X} \sim N(\mu, \sigma^2/n)$ and the sample size is 30, the following random variable is a pivot for the population mean:

$$Z = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim N(0, 1).$$

A two-sided confidence interval is:

$$P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) = 1 - \alpha$$

$$P\left(-z_{\alpha/2} < \frac{\bar{X}-\mu}{S/\sqrt{n}} < z_{\alpha/2}\right) = 1-\alpha$$

$$P\left(\bar{X} - z_{\alpha/2}S/\sqrt{n} < \mu < \bar{X} + z_{\alpha/2}S/\sqrt{n}\right) = 1-\alpha$$

Hence, a 95% confidence interval for the average life of the light bulbs is:

$$P\left(1500 - 1.96 \cdot 50/\sqrt{30} < \mu < 1500 + 1.96 \cdot 50/\sqrt{30}\right) = 0.95$$

$$(1482.11 < \mu < 1517.89) \quad \Rightarrow \quad \text{approximately} \quad (1482, 1518)$$

We can be 95% confident that the the average life of the light bulbs is somewhere between 1482 and 1518 hours.

ii) The following random variable as a pivot to estimate the variance of the population:

$$T(\hat{\sigma}^2, \sigma^2) = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1},$$

where S^2 is the unbiased sample variance.

For two-sided 90% confidence interval we can find $\chi^2_{0.05,29} = 17.71$ and $\chi^2_{0.95,29} = 42.56$;

$$s^2 = 2500$$

.

Hence, the 90% confidence interval is:

$$P\left(L < \frac{(n-1)S^2}{\sigma^2} < U\right) = 0.90$$

$$P\left(\frac{29 \cdot 2500}{42.56} < \sigma^2 < \frac{29 \cdot 2500}{17.71}\right) = 0.90$$

the interval is $(1703.48, 4093.73)$ (to 2 d.p.)

For one-sided 90% confidence interval we can find $\chi^2_{0.10,29} = 19.77$.

Hence, the 90% confidence interval is:

$$P\left(L < \frac{(n-1)S^2}{\sigma^2}\right) = 0.90$$

$$P\left(\sigma^2 < \frac{29 \cdot 2500}{19.77}\right) = 0.90$$

the interval is $(0, 3667.17)$ (to 2 d.p.) (the variance cannot be less than 0)

The one-sided interval is wider, so if we do not have any particular reason which type of intervals to choose, the two-sided is more precise.