MA20033 - Solution Sheet Five

Simon Shaw
s.c.shaw@maths.bath.ac.uk

2004/05 Semester I

1. **The following data, which are assumed to come from a normal distribution with mean μ, representing "the passage time of light", and variance σ^2, may be regarded as Newcomb's measurements of "the passage time of light".**

28	26	33	24	34	−44	27	16	40	−2
29	22	24	21	25	30	23	29	31	19

   (a) **Given that $\sum_{i=1}^{20} x_i = 435$ and $\sum_{i=1}^{20} x_i^2 = 15365$, find an unbiased estimate of the mean and the variance.**

   Unbiased estimators of μ and σ^2 are $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ respectively (we showed this on question 2.(c) of Question Sheet Two). In this case, the estimates are

$$\overline{x} = \frac{1}{20}\sum_{i=1}^{20} x_i = \frac{435}{20} = 21.75,$$

$$s^2 = \frac{1}{19}\sum_{i=1}^{20}(x_i - \overline{x})^2$$

$$= \frac{1}{19}\left(\sum_{i=1}^{20} x_i^2 - 20\overline{x}^2\right) = \frac{1}{19}\{15365 - 20(21.75)^2\} = 310.7237.$$

   (b) **Construct a 95% confidence interval for the "true" passage time of light.**

   As both μ and σ^2 are unknown then $(\overline{X} - \mu)/(S/\sqrt{20}) \sim t_{19}$. Now $P(t_{19} > 2.093) = 0.025$ so that $P(-2.093 < t_{19} < 2.093) = 0.95$. Substituting t_{19} by $(\overline{X} - \mu)/(S/\sqrt{20})$, we have

$$P\left(-2.093 < \frac{\overline{X} - \mu}{S/\sqrt{20}} < 2.093 \middle| \mu, \sigma^2\right)$$

$$= P\left(\overline{X} - 2.093\frac{S}{\sqrt{20}} < \mu < \overline{X} + 2.093\frac{S}{\sqrt{20}} \middle| \mu, \sigma^2\right) = 0.95.$$

Thus, $(\overline{X} - 2.093\frac{S}{\sqrt{20}}, \overline{X} + 2.093\frac{S}{\sqrt{20}})$ is a random interval which contains μ with probability 0.95. The 95% confidence interval for μ, the "true" passage time of light, is a realisation of this, $(\overline{x} - 2.093\frac{s}{\sqrt{20}}, \overline{x} + 2.093\frac{s}{\sqrt{20}})$. In this case, with $\overline{x} = 21.75$ and $s = \sqrt{310.7237} = 17.6274$, the interval is $(13.5002, 29.9998)$.

(c) **Omitting the two smallest data values, recalculate the unbiased estimate of the mean and variance and the corresponding 95% confidence interval. Comment briefly on differences in results compared with those using all 20 data values.**

This time, we note that $(\overline{X} - \mu)/(S/\sqrt{18}) \sim t_{17}$. Now $P(t_{17} > 2.110) = 0.025$ so that $P(-2.110 < t_{17} < 2.110) = 0.95$ and $(\overline{X} - 2.110\frac{S}{\sqrt{18}}, \overline{X} + 2.110\frac{S}{\sqrt{18}})$ is a random interval which contains μ with probability 0.95. The 95% confidence interval for μ is a realisation of this, $(\overline{x} - 2.110\frac{s}{\sqrt{18}}, \overline{x} + 2.110\frac{s}{\sqrt{18}})$. In this case,

$$\begin{aligned}
\overline{x} &= \frac{1}{18}\sum_{i=1}^{18} x_i = \frac{435 - (-2) - (-44)}{18} = \frac{481}{18} = 26.7222, \\
s^2 &= \frac{1}{17}\sum_{i=1}^{18}(x_i - \overline{x})^2 \\
&= \frac{1}{17}\left(\sum_{i=1}^{18} x_i^2 - 18\overline{x}^2\right) = \frac{1}{17}\{(15365 - (-2)^2 - (-44)^2) - 18(26.7222)^2\} \\
&= \frac{1}{17}\{13425 - 18(26.7222)^2\} = 33.6242.
\end{aligned}$$

Hence $s = \sqrt{33.6242} = 5.7986$; the 95% confidence interval is $(23.8384, 29.6061)$.

The removal of the "outliers" -44, -2 affect the mean, standard deviation and confidence interval appreciably. We note, again, a lack of robustness to summaries based upon averaging. The confidence intervals were constructed using the assumption of normality. A stem-and-leaf plot of the data indicates that without the "outliers" this seems reasonable. However, if the outliers are included, one would have to question this.

2. **Recall question 3. of Question Sheet Four. To estimate the gestation period of domestic dogs, 15 randomly selected dogs are observed during pregnancy. Their gestation periods, in days, are:**

62.0	61.4	59.8	62.2	60.3
60.4	59.4	60.2	60.4	60.8
61.8	59.2	61.1	60.4	60.9

Letting x_i denote the observed gestation period of the ith dog, we find that $\sum_{i=1}^{15} x_i = 910.3$ and $\sum_{i=1}^{15} x_i^2 = 55254.35$. Thus, $\overline{x} = 60.6867$ and

$$s^2 = \frac{1}{14}\left\{\sum_{i=1}^{15} x_i^2 - 15\overline{x}^2\right\} = \frac{1}{14}\left\{55254.35 - 15(60.6867)^2\right\} = 0.8055.$$

Once more, we will make the assumption that these 15 observations are realisations from a population which may be modelled by a $N(\mu, \sigma^2)$ distribution.

(a) **Evaluate a 99% confidence interval for μ when σ^2 is known to be 1.**

In this case, $\overline{X} \sim N(\mu, 1/15)$ so that $\sqrt{15}(\overline{X} - \mu) \sim N(0,1)$. Thus,

$$\begin{aligned} &P(-2.576 < \sqrt{15}(\overline{X} - \mu) < 2.576|\mu) \\ &= \; P(\overline{X} - 2.576/\sqrt{15} < \mu < \overline{X} + 2.576/\sqrt{15}|\mu) \; = \; 0.99. \end{aligned}$$

Thus, $(\overline{X} - 2.576/\sqrt{15}, \overline{X} + 2.576/\sqrt{15})$ is a random interval which contains μ with probability 0.99. The 99% confidence interval for μ is a realisation of this, $(\overline{x} - 2.576/\sqrt{15}, \overline{x} + 2.576/\sqrt{15})$. In this instance, the 99% confidence interval for μ is $(60.0249, 61.3551)$.

(b) **Evaluate a 99% confidence interval for μ when σ^2 is assumed unknown.**

When σ^2 is unknown then the random interval of the form $(\overline{X} - 2.576\sigma/\sqrt{15}, \overline{X} + 2.576\sigma/\sqrt{15})$ is of no benefit for we cannot evaluate a realisation of it (as we don't know σ). Instead, we must use that $(\overline{X} - \mu)/(S/\sqrt{15}) \sim t_{14}$. Now, $P(t_{14} > 2.977) = 0.005$ so that $P(-2.977 < t_{14} < 2.977) = 0.99$. Substituting t_{14} by $(\overline{X} - \mu)/(S/\sqrt{15})$, we have

$$\begin{aligned} &P\left(-2.977 < \frac{\overline{X} - \mu}{S/\sqrt{15}} < 2.977 \middle| \mu, \sigma^2\right) \\ &= \; P\left(\overline{X} - 2.977\frac{S}{\sqrt{15}} < \mu < \overline{X} + 2.977\frac{S}{\sqrt{15}} \middle| \mu, \sigma^2\right) \; = \; 0.99. \end{aligned}$$

Thus, $(\overline{X} - 2.977\frac{S}{\sqrt{15}}, \overline{X} + 2.977\frac{S}{\sqrt{15}})$ is a random interval which contains μ with probability 0.99. The 99% confidence interval for μ is a realisation of this, $(\overline{x} - 2.977\frac{s}{\sqrt{15}}, \overline{x} + 2.977\frac{s}{\sqrt{15}})$. In this case, with $\overline{x} = 60.6867$ and $s = \sqrt{0.8055} = 0.8975$, the interval is $(59.9968, 61.3765)$.

(c) **Comment on the differences or similarities of these intervals.**

The two intervals are very similar. The observed sample standard deviation was 0.8975 which is close to an assumed standard deviation of 1 (we will shortly investigate how to test whether this data suggests that $\sigma^2 = 1$). Notice that had we merely substituted s for σ in the confidence interval derived from the Normal distribution, yielding $(\overline{x} - 2.576\frac{s}{\sqrt{15}}, \overline{x} + 2.576\frac{s}{\sqrt{15}})$, then we would have only an approximate 95% confidence interval which is narrower than the true 95% confidence interval, $(\overline{x} - 2.977\frac{s}{\sqrt{15}}, \overline{x} + 2.977\frac{s}{\sqrt{15}})$, as the latter incorporates the uncertainty in not knowing σ^2.