MA20033 - Solution Sheet Four

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

2004/05 Semester I

1. **Calculate the mean, median and once trimmed mean for the following data:**

   3, 12, 13, 14, 15, 15, 17, 19, 20, 24

   We find that

$$\begin{aligned} \overline{x} &= \frac{1}{10}\sum_{i=1}^{10} x_{(i)} = \frac{152}{10} = 15.2, \\ med(x) &= \frac{x_{(5)} + x_{(6)}}{2} = \frac{15+15}{2} = 15, \\ \overline{x}_1 &= \frac{1}{8}\sum_{i=2}^{9} x_{(i)} = \frac{125}{8} = 15.625. \end{aligned}$$

   Now suppose that the observation with value 3 is replaced by a new observation with value x, where x takes some value between 0 and 30. Derive expressions for the three estimates calculated above as a function of x, and comment on their robustness.

   The revised mean is

$$\overline{x} = \frac{1}{10}\sum_{i=1}^{10} x_{(i)} = \frac{149+x}{10} \in [14.9, 17.9].$$

   For the median, note that if $x \leq 15$ then $x_{(5)}$ and $x_{(6)}$ are unchanged. If $x = 16$ then $x_{(5)} = 15$ and $x_{(6)} = 16$ while if $x \geq 17$ then $x_{(5)} = 15$ and $x_{(6)} = 17$. Thus,

$$med(x) = \begin{cases} 15 & 0 \leq x \leq 15, \\ 15.5 & x = 16, \\ 16 & 17 \leq x \leq 30. \end{cases}$$

   In this case, $med(x) \in [15, 16]$. For the once-trimmed mean, x and 24 are trimmed for $0 \leq x \leq 12$, 3 and 24 are trimmed for $13 \leq x \leq 24$ and for $25 \leq x \leq 30$ we trim 3 and x. Hence,

$$\overline{x}_1 = \begin{cases} 15.625 & 0 \leq x \leq 12, \\ \frac{113+x}{8} & 13 \leq x \leq 24, \\ 17.125 & 25 \leq x \leq 30. \end{cases}$$

Thus, $\overline{x}_1 \in [15.625, 17.125]$. The mean is affected by the value of x in all cases, with the effect being more noticeable when x is in either extreme of the data, as illustrated by the wide range of values $\overline{x}$ can take. The affect of the extreme values of x is moderated in the once-trimmed mean, with x only affecting the statistic for values in the main body of the data. Hence, the range of possible values is much reduced. The median is the least affected of the three, as illustrated by its narrow range of possible values.

2. **Use your statistical tables to find the following quantities.**

   (a) **The values which enclose the central 90% of a $N(0,1)$ distribution.**

   Let $Z \sim N(0,1)$. Then $P(Z < 1.645) = 0.95$ and $P(Z < -1.645) = 0.05$. Thus,

   $$P(-1.645 < Z < 1.645) \;=\; P(Z < 1.645) - P(Z < -1.645) \;=\; 0.90.$$

   (b) **The values which enclose the central 90% of a $N(10,5)$ distribution.**

   If $X \sim N(10,5)$ then $(X-10)/\sqrt{5} \sim N(0,1)$. Hence,

   $$\begin{aligned} &P(-1.645 < (X-10)/\sqrt{5} < 1.645) \\ &\quad = \;\; P(10 - 1.645\sqrt{5} < X < 10 + 1.645\sqrt{5}) \;=\; 0.90. \end{aligned}$$

   (c) **The values which enclose the central 95% of a χ^2_{15} distribution.**

   From tables, $P(\chi^2_{15} > 6.262) = 0.975$ and $P(\chi^2_{15} > 27.488) = 0.025$. Thus,

   $$P(6.262 < \chi^2_{15} < 27.488) \;=\; P(\chi^2_{15} > 6.262) - P(\chi^2_{15} > 27.488) \;=\; 0.95.$$

3. **To estimate the mean 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, $\overline{x} = 60.69$ and

   $$s^2 \;\;=\;\; \frac{1}{14}\left\{\sum_{i=1}^{15} x_i^2 - 15\overline{x}^2\right\} \;\;=\;\; \frac{1}{14}\left\{55254.35 - 15(60.69)^2\right\} \;\;=\;\; 0.8055.$$

   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 95% 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,

   $$\begin{aligned} &P(-1.96 < \sqrt{15}(\overline{X} - \mu) < 1.96|\mu) \\ &\quad = \;\; P(\overline{X} - 1.96/\sqrt{15} < \mu < \overline{X} + 1.96/\sqrt{15}|\mu) \;=\; 0.95. \end{aligned}$$

   Thus, $(\overline{X} - 1.96/\sqrt{15}, \overline{X} + 1.96/\sqrt{15})$ is a random interval which contains μ with probability 0.95. The 95% confidence interval for μ is a realisation of this, $(\overline{x} - 1.96/\sqrt{15}, \overline{x} + 1.96/\sqrt{15})$. In this instance, the 95% confidence interval for μ is $(60.18, 61.20)$.

(b) **Evaluate a 95% confidence interval for σ^2 (assumed unknown).**

In this instance, $(15-1)S^2/\sigma^2 \sim \chi^2_{15-1}$. Now $P(\chi^2_{14} > 5.629) = 0.975$ and $P(\chi^2_{14} > 26.119) = 0.025$ so that

$$P(5.629 < 14S^2/\sigma^2 < 26.119|\sigma^2) \;=\; 0.95.$$

Solving for σ^2 gives $P(14S^2/26.119 < \sigma^2 < 14S^2/5.629|\sigma^2) = 0.95$. Thus, $(14S^2/26.119, 14S^2/5.629)$ is a random interval that contains σ^2 with probability 0.95. The 95% confidence interval for σ^2 is a realisation of this, $(14s^2/26.119, 14s^2/5.629)$. In this instance, the 95% confidence interval for σ^2 is $(0.43, 2.00)$.

4. **Sometimes it is the case that a one-sided rather than a two-sided confidence interval is required, which means that we want the realisation of a random (half) interval of the form either**

$$P\{\theta > g_1(X_1, \ldots, X_n)|\theta\} \;=\; 1-\alpha,$$

(a one-sided lower $(1-\alpha)100\%$ random interval for θ), or

$$P\{\theta < g_2(X_1, \ldots, X_n)|\theta\} \;=\; 1-\alpha,$$

(a one-sided upper $(1-\alpha)100\%$ random interval for θ). Derive and evaluate a 95% upper confidence interval for σ^2 using the dog data set of question 3.

Note that $P(\chi^2_{14} > 6.571) = 0.95$ so that

$$P(14S^2/\sigma^2 > 6.571|\sigma^2) = 0.95$$

and hence $P(\sigma^2 < 14S^2/6.571|\sigma^2) = 0.95$. $(0, 14S^2/6.571)$ is a one-sided upper random interval for σ^2 with probability 0.95 and $(0, 14s^2/6.571)$ is a one-sided 95% upper confidence interval for σ^2. In this instance, the interval is $(0, 1.716)$.

5. **Suppose we wish to find an interval estimator for the parameter p when we have a random variable $X \sim Bin(n,p)$.**

(a) **By considering the Normal approximation to the Binomial distribution, show that $\frac{X/n-p}{\sqrt{p(1-p)/n}}$ is an approximate pivot for p, and state its sampling distribution.**

The Normal approximation to the Binomial distribution tells us that, provided p is not too close to 0 or 1 and n is larger than about 15, by matching the mean and variance of the distributions

$$\begin{aligned} X \sim Bin(n,p) \;&\Rightarrow\; X \sim N(np, np(1-p)) \text{ approximately} \\ &\Rightarrow\; \frac{X}{n} \sim N(p, \frac{p(1-p)}{n}) \text{ approximately} \\ &\Rightarrow\; \frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1) \text{ approximately} \end{aligned}$$

(using properties of Normal distributions). In other words, $\frac{X/n-p}{\sqrt{p(1-p)/n}}$ is an approximate pivot for p, and it follows a $N(0,1)$ distribution.

(b) **Write down a random interval which contains p with probability (approximately) 0.95 (to do this you will need to make a further approximation of the variance of the Normal using a point estimator of p).**

In order to write down a random interval which contains p with probability (approximately) 0.95, we use the same ideas as in question 3(a) for deriving a random interval for the mean of a Normal distribution,

$$P\left(\frac{X}{n} - 1.96\sqrt{\frac{p(1-p)}{n}} < p < \frac{X}{n} + 1.96\sqrt{\frac{p(1-p)}{n}}\,\middle|\, p\right) \approx 0.95.$$

However, this still has the problem that the variance of the approximating Normal involves the parameter p, and so we make one further approximation and estimate the variance by plugging in the estimator of p, X/n, to give the approximate interval

$$P\left(\frac{X}{n} - 1.96\sqrt{\frac{X/n(1-X/n)}{n}} < p < \frac{X}{n} + 1.96\sqrt{\frac{X/n(1-X/n)}{n}}\,\middle|\, p\right) \approx 0.95.$$

6. **Suppose $X_1, \ldots, X_n$ are iid $U(0,\theta)$ random quantities, and that we wish to find an interval estimator for θ.**

(a) **Recalling the cumulative distribution function of $M = \max\{X_1, \ldots, X_n\}$ (see question 3 on Question Sheet Two), find a pivot for θ by considering a particular linear transformation of M.**

Recall that if $X \sim U(a,b)$ then $X = (b-a)U + a$ where $U \sim U(0,1)$. Thus, let $T = M/\theta$. then $T = \max\{X_1/\theta, \ldots, X_n/\theta\}$ and each $X_i/\theta \sim U(0,1)$ so do not depend upon θ. Hence, T does not depend on θ and is a pivot for θ. Alternatively,

$$\begin{aligned} P(T \le t) &= P(M \le \theta t) \\ &= \begin{cases} 0 & \theta t < 0, \\ \left(\frac{\theta t}{\theta}\right)^n & 0 \le \theta t \le \theta, \\ 1 & \theta t \ge \theta \end{cases} = \begin{cases} 0 & t < 0, \\ t^n & 0 \le t \le 1, \\ 1 & t \ge 1 \end{cases} \end{aligned}$$

which shows us that T is a pivot for θ as the cdf of T does not depend on θ.

(b) **Now using the cumulative distribution function of the pivot you have derived, find a random interval which contains θ with probability 0.95.** We seek c_1 and c_2 such that

$$P(c_1 < T < c_2) = P(T < c_2) - P(T < c_1) = 0.95.$$

There are infinite choices for c_2 and c_1. One choice is to set $P(T < c_2) = 0.975$ and $P(T < c_1) = 0.025$. Hence, $c_2 = 0.975^{1/n}$ and $c_1 = 0.025^{1/n}$. Recalling that $T = M/\theta$ we have $P(0.025^{1/n} < M/\theta < 0.975^{1/n}|\theta) = 0.95$. Hence,

$$P(M/0.975^{1/n} < \theta < M/0.025^{1/n}|\theta) = 0.95.$$

Thus, $(M/0.975^{1/n}, M/0.025^{1/n})$ is a random interval which contains θ with probability 0.95.

(c) **Does this interval contain the maximum likelihood estimator of θ?**
The maximum likelihood estimator of θ is M but $M/0.975^{1/n} > M$ so M is not in this random interval.