MATH10282 Introduction to Statistics
Semester 2, 2019/2020
Example Sheet 4 - Solutions

1. For $\hat{Q}(0.10)$, $r = 0.10 \times 26 = 2.6$. Thus

$$\begin{aligned}\hat{Q}(0.10) &= x_{(2)} + 0.6 \times (x_{(3)} - x_{(2)}) \\ &= 0.2290 + 0.6 \times (0.6390 - 0.2290) \\ &= 0.475\end{aligned}$$

For $\hat{Q}(0.25)$, $r = 0.25 \times 26 = 6.5$. Thus,

$$\begin{aligned}\hat{Q}(0.25) &= x_{(6)} + 0.5 \times (x_{(7)} - x_{(6)}) \\ &= 1.8529 + 0.5 \times (2.3581 - 1.8529) \\ &= 2.1055\end{aligned}$$

For $\hat{Q}(0.50)$, $r = 0.5 \times 26 = 13$. Thus $\hat{Q}(0.50) = x_{(13)} = 2.8957$.
For $\hat{Q}(0.75)$, $r = 0.75 \times 26 = 19.5$. Thus

$$\begin{aligned}\hat{Q}(0.75) &= x_{(19)} + 0.5 \times (x_{(20)} - x_{(19)}) \\ &= 6.0358 + 0.5 \times (6.2642 - 6.0358) \\ &= 6.15\end{aligned}$$

For $\hat{Q}(0.90)$, $r = 0.9 \times 26 = 23.4$. Thus

$$\begin{aligned}\hat{Q}(0.90) &= x_{(23)} + 0.4 \times (x_{(24)} - x_{(23)}) \\ &= 14.3460 + 0.4 \times (14.8156 - 14.3460) \\ &= 14.53384\end{aligned}$$

If the distribution were symmetric, we would expect the distances from the median to the lower and upper quartiles to be similar: however, here $\hat{Q}(0.5) - \hat{Q}(0.25) = 0.7902$ while $\hat{Q}(0.75) - \hat{Q}(0.5) = 3.2453$, indicating a high degree of asymmetry.

2. (i) The minimum is $x_{(1)} = 29.140$.
For $\hat{Q}(0.25)$, $r = 21 \times 0.25 = 5.25$, so

$$\begin{aligned}\hat{Q}(0.25) &= x_{(5)} + 0.25 \times (x_{(6)} - x_{(5)}) \\ &= 36.571 + 0.25 \times (38.417 - 36.571) \\ &= 37.0325\end{aligned}$$

For $\hat{Q}(0.5)$, $r = 21 \times 0.5 = 10.5$, so

$$\begin{aligned}\hat{Q}(0.50) &= x_{(10)} + 0.5(x_{(11)} - x_{(10)}) \\ &= 41.042 + 0.5 \times (41.330 - 41.042) \\ &= 41.186\end{aligned}$$

For $\hat{Q}(0.75)$, $r = 21 \times 0.75 = 15.75$, so

$$\begin{aligned}\hat{Q}(0.75) &= x_{(15)} + 0.75 \times (x_{(16)} - x_{(15)}) \\ &= 44.004 + 0.75 \times (44.025 - 44.004) \\ &= 44.01975\end{aligned}$$

The maximum is 51.261. Thus the five number summary is

$$(29.140, 37.0325, 41.186, 44.01975, 51.261)\,.$$

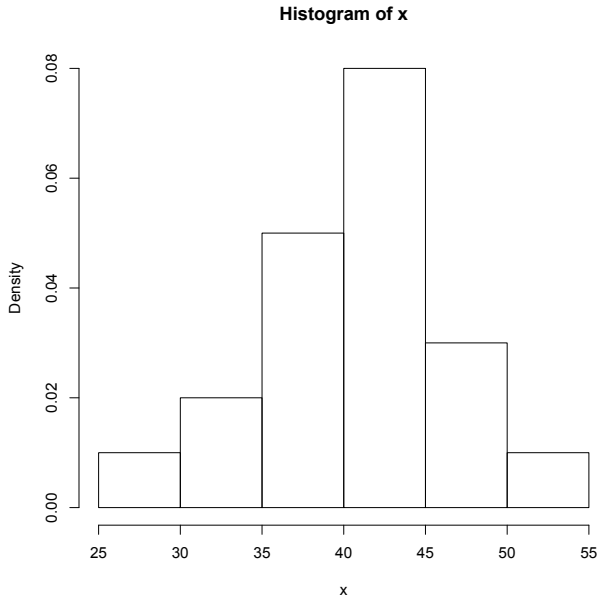
(ii) The mean and variance are

$$\bar{x} = \frac{\sum_{i=1}^{20} x_i}{20} = \frac{813.884}{20} = 40.6942$$
$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} = \frac{33700.68 - 20 \times 40.6942^2}{19} = 30.54325\,.$$

Hence the standard deviation is $s = 5.527$ (3 s.f.).

(iii)

Interval	Frequency	Height of density histogram
(25,30]	1	1/100
(30,35]	2	2/100
(35,40]	5	5/100
(40,45]	8	8/100
(45,50]	3	3/100
(50,55]	1	1/100
	20	



Looking at the five number summary, mean, and histogram the distribution appears reasonably symmetric.

3. Recall that

$$\text{Hist}(x) = \frac{\nu_k}{nh}\,, \quad x \in B_k\,.$$

where ν_k is the number of x_i in B_k, and $B_1, \ldots, B_K$ are disjoint intervals of width h covering $[x_{(1)}, x_{(n)}]$.

(a) In the question, $h = 10$. For $x = 20.0 \in (15, 25]$, $\nu_k = 142$ and so $\text{Hist}(x) = \nu_k/(nh) = 142/(500 \times 10) = 0.0284$.

For $x = 80.0 \in (75, 85]$, $\nu_k = 6$ and so $\text{Hist}(x) = \nu_k/(nh) = 6/(500 \times 10) = 0.0012$.

(b) With the new tabulation, for $x = 20.0 \in (5, 25]$ we have $\nu_k = 83 + 142 = 225$ and so $\text{Hist}(x) = 225/(500 \times 20) = 0.0225$.

For $x = 80.0 \in (65, 85]$, we have $\nu_k = 13 + 6 = 19$, so $\text{Hist}(x) = 19/(500 \times 20) = 0.0019$.

The histogram values are quite similar across the two tabulations.

4. (i) Observe that

$$\begin{aligned}
\hat{\mu}_{\text{Hist}} &= \int_{a_1}^{a_{K+1}} x \,\text{Hist}(x)\, dx \\
&= \sum_{k=1}^{K} \int_{a_k}^{a_{k+1}} x \,\text{Hist}(x)\, dx \\
&= \sum_{k=1}^{K} \int_{a_k}^{a_{k+1}} x \,\frac{\nu_k}{nh}\, dx = \sum_{k=1}^{K} \frac{\nu_k}{nh} \left[\frac{x^2}{2}\right]_{a_k}^{a_{k+1}} \\
&= \sum_{k=1}^{K} \frac{\nu_k}{2nh}(a_{k+1}^2 - a_k^2) \\
&= \sum_{k=1}^{K} \frac{\nu_k}{2nh}(a_{k+1} + a_k)(a_{k+1} - a_k) \\
&= \sum_{k=1}^{K} \frac{\nu_k}{n} \frac{(a_{k+1} + a_k)}{2}\,.
\end{aligned}$$

This is the formula that usually appears in textbooks for finding the sample mean from a grouped set of data.

(ii*) We show this using proof by contradiction. First note that the assumption that $h < \min_{i \neq j} |x_i - x_j|$ implies that

$$|x_i - x_j| > h \text{ for all } i \neq j\,. \ (*)$$

Now suppose that two different observations, x_i and x_j with $i \neq j$ are both in $B_k = (a_k, a_{k+1}]$. As B_k is of width h, the distance between x_i and x_j must be at most h, i.e. $|x_i - x_j| \leq h$. This contradicts (*) above, hence there cannot be two different observations in B_k. Thus there must be one or zero observations in B_k.

(iii*) By the above, for all k, $v_k = 0$ or $v_k = 1$. Thus, we have

$$\begin{aligned}
\hat{\mu}_{\text{Hist}} &= \frac{1}{2n} \sum_{k=1}^{K} \nu_k (a_{k+1} + a_k) \\
&= \frac{1}{2n} \sum_{\{k:v_k=0\}} v_k (a_{k+1} + a_k) + \frac{1}{2n} \sum_{\{k:v_k=1\}} v_k (a_{k+1} + a_k) \\
&= \frac{1}{2n} \sum_{\{k:v_k=0\}} 0 + \frac{1}{2n} \sum_{\{k:v_k=1\}} 1 \times (a_{k+1} + a_k).
\end{aligned}$$

However, $\{k : v_k = 1\}$ is the set of indices of bins containing precisely one observation, and this is just a re-ordering of $k_1, \ldots, k_n$, since each

observation is contained in one of the bins. Hence

$$\hat{\mu}_{\text{Hist}} = \frac{1}{2n} \sum_{\{k:v_k=1\}} (a_{k+1} + a_k) = \frac{1}{2n} \sum_{i=1}^{n} (a_{k_i+1} + a_{k_i}) \,.$$

(iv*) Since $a_{k_i+1} = a_{k_i} + h$ and $a_{k_i} < x_i$, adding h to both sides we have that $a_{k_{i+1}} < x_i + h$ and also $a_{k_i} + a_{k_{i+1}} < 2x_i + h$. Multiplying by $1/(2n)$ and summing over i, we have

$$\frac{1}{2n} \sum_{i=1}^{n} (a_{k_i} + a_{k_{i+1}}) < \frac{1}{2n} \sum_{i=1}^{n} (2x_i + h)$$

The left hand side is equal to $\hat{\mu}_{\text{Hist}}$. The right hand side satisfies $\frac{1}{2n} \sum_{i=1}^{n} (2x_i + h) = \frac{1}{n} \sum_{i=1}^{n} x_i + \frac{1}{2n} nh = \bar{x} + \frac{h}{2}$. Hence we have shown

$$\hat{\mu}_{\text{Hist}} < \bar{x} + \frac{h}{2} \,.$$

A similar argument shows that $\hat{\mu}_{\text{Hist}} > \bar{x} - h/2$.

(v*) As $h \to 0$, eventually $h < \min_{i \neq j} |x_i - x_j|$, and so we may assume that the above results hold. Now we use a sandwiching idea.

$$\begin{aligned}
&\bar{x} - \frac{nh}{2} \leq \hat{\mu}_{\text{Hist}} \leq \bar{x} + \frac{nh}{2} \\
&\implies \lim_{h \to 0} \left\{ \bar{x} - \frac{nh}{2} \right\} \leq \lim_{h \to 0} \hat{\mu}_{\text{Hist}} \leq \lim_{h \to 0} \left\{ \bar{x} + \frac{nh}{2} \right\} \qquad \text{taking limits} \\
&\implies \bar{x} \leq \lim_{h \to 0} \hat{\mu}_{\text{Hist}} \leq \bar{x} \,.
\end{aligned}$$

Hence $\lim_{h \to 0} \hat{\mu}_{\text{Hist}} = \bar{x}$.

5. (i) The calculations are as follows:

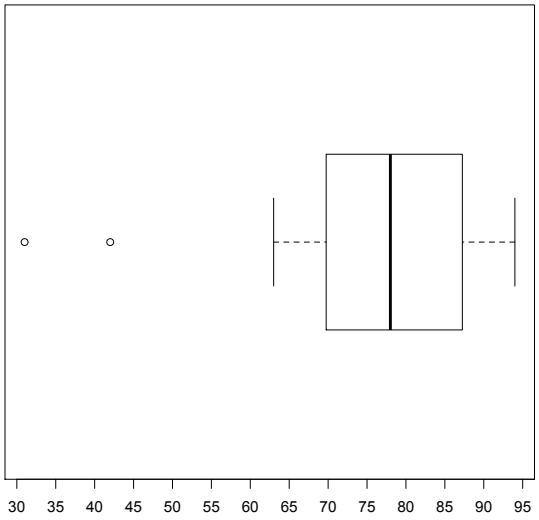
$$\begin{aligned}
x_{(1)} &= 31 \\
x_{(18)} &= 94 \\
\hat{Q}(0.25) &= x_{(4)} + 0.75(x_{(5)} - x_{(4)}) \\
&= 69 + 0.75(70 - 69) = 69.75 \\
\hat{Q}(0.5) &= x_{(9)} + 0.5(x_{(10)} - x_{(9)}) \\
&= (77 + 79)/2 = 78 \\
\hat{Q}(0.75) &= x_{(14)} + 0.25(x_{(15)} - x_{(14)}) \\
&= 87 + (1/4)(88 - 87) = 87.25
\end{aligned}$$

Thus the five number summary is

$$(31, 69.75, 78, 87.25, 94) \,,$$

and IQR = 87.25 - 69.75 = 17.5.

(ii) First we identify outliers. The thresholds are $\hat{Q}(0.25) - 1.5\text{IQR} = 43.5$, $\hat{Q}(0.75) + 1.5\text{IQR} = 113.5$. Thus the data values 31 and 42 are classified as outliers. The lower adjacent value is 63. The upper adjacent value is 94. Your hand drawn box plot should resemble the figure below.



The distribution is fairly symmetric apart from two small-valued outliers. The two whiskers seem a little shorter than one might expect under normality. There are clear outliers which also suggest a deviation from normality.

(iii) $\bar{x} = 75.06$, $s = 16.540$, $\hat{Q}(0.5) = 78.0$. Here the sample mean is less than the median. Its value is pulled down somewhat by the low-valued outliers.

6. (i) Recall the idea of standardization, i.e. if $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0, 1)$. Hence

$$F(x) = \Pr(X \leq x) = \Pr\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right)$$
$$= \Pr\left(Z \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

(ii) Here F is strictly increasing, so $Q(p) = F^{-1}(p)$. I.e. if $q = Q(p)$, then $F(q) = F(F^{-1}(p)) = p$, so

$$\Phi\left(\frac{q-\mu}{\sigma}\right) = p\,.$$

Applying Φ^{-1} to both sides,

$$\frac{q-\mu}{\sigma} = \Phi^{-1}(p)\,,$$

and so $Q(p) = q = \mu + \sigma\Phi^{-1}(p)$. The interquartile range is

$$\begin{aligned} Q(0.75) - Q(0.25) &= (\mu + \sigma\Phi^{-1}(0.75)) - (\mu + \sigma\Phi^{-1}(0.25)) \\ &= (\mu - \mu) + \sigma \times (0.6745 - (-0.6475)) = 1.349\sigma\,. \end{aligned}$$

(iii)

$$\begin{aligned} \hat{\text{IQR}} &= 2 \times 0.6745s \\ &= 22.31 \text{ in this case}\,. \end{aligned}$$

This is somewhat larger than the empirically estimated IQR,

$$\hat{Q}(0.75) - \hat{Q}(0.25) = 17.5\,,$$

again suggesting the normal distribution may not be a good fit.

7. Recall from lectures (Chapter 2 slide 6) that the median is

$$\tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & \text{if } n \text{ is even} \end{cases}$$

Meanwhile, letting $r = p(n+1)$ and r' be the integer part of r,

$$\hat{Q}(p) = \begin{cases} x_{(r)} & \text{if } r = 1, 2, \ldots, n \\ x_{(r')} + (r - r')(x_{(r'+1)} - x_{(r')}) & \text{if } 1 \le r \le n-1 \\ x_{(1)} & \text{if } r = 0 \\ x_{(n)} & \text{if } r = n \end{cases}$$

If n is odd, then $n+1$ is even and $\frac{n+1}{2}$ is an integer. Thus, with $p = 0.5$, in this case $r = p(n+1) = 0.5(n+1)$ is an integer with $1 \le r \le n-1$. Hence

$$\hat{Q}(0.5) = x_{(r)} = x_{\left(\frac{n+1}{2}\right)}\,.$$

If n is even, with $p = 0.5$, we have $r = p(n+1) = \frac{n+1}{2} = \frac{n}{2} + 0.5$ with $\frac{n}{2}$ an integer, and so $r' = \frac{n}{2}$. Also, $1 \le r' \le n-1$. Thus,

$$\begin{aligned} \hat{Q}(0.5) &= x_{(r')} + (r - r')(x_{(r'+1)} - x_{(r')}) = x_{\left(\frac{n}{2}\right)} + \frac{1}{2}\left(x_{\left(\frac{n}{2}+1\right)} - x_{\left(\frac{n}{2}\right)}\right) \\ &= \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right)\,. \end{aligned}$$

Hence for both n even and n odd, $\hat{Q}(0.5)$ agrees with the earlier expression for the sample median.