MATH10282 Introduction to Statistics
Semester 2, 2020/2021
Example Sheet 4

Attempt the calculations for these questions by hand, using a calculator at most. When calculating the sample p-quantile, use the form given in lectures i.e. $\hat{Q}(p) = x_{(p(n+1))}$, with interpolation as appropriate.

1. The following data ($n = 25$) are a random sample of observations from a continuous distribution. They have been ordered from smallest to largest for your convenience.

```
 0.1686  0.2290  0.6390  0.9597  1.7250  1.8529
 2.3581  2.4002  2.5168  2.5394  2.5557  2.6999
 2.8957  2.9106  3.4549  3.4900  4.6079  5.7313
 6.0358  6.2642  7.4736 13.3279 14.3460 14.8156
16.2938
```

Calculate the sample quantiles $\hat{Q}(p)$ for $p = 0.10, 0.25, 0.50, 0.75, 0.90$. Use this information to deduce whether the data are sampled from a symmetric distribution or not.

2. The following data ($n = 20$) are a random sample from a particular continuous distribution.

```
29.140 31.131 34.307 35.352 36.571 38.417 38.495 39.713
40.286 41.042 41.330 41.983 43.631 43.845 44.004 44.025
45.007 46.499 47.845 51.261
```

They have been ordered from smallest to largest for your convenience.

(i) Compute the five number summary for these data.

(ii) Calculate the sample mean and standard deviation. (Expressions for $\overline{x}$ and s are in Chapter 2 of the course notes).

(iii) Obtain a frequency table of the data using equally-sized intervals of length $h = 5.0$ in the range $[25, 55]$. Hence, plot a density histogram for the data.

(iv) Using the results of parts (i) - (iii) above, discuss whether you think the underlying distribution of the data is symmetric.

3. This question concerns the income data given on p.3 of Chapter 1 of the course notes. The data are reproduced below for convenience.

Intervals	Frequencies	Percents
5 to 15	83	16.6
15 to 25	142	28.4
25 to 35	90	18.0
35 to 45	79	15.8
45 to 55	46	9.2
55 to 65	28	5.6
65 to 75	13	2.6
75 to 85	6	1.2
85 to 95	4	0.8
95 to 105	3	0.6
105 to 115	0	0.0
115 to 125	2	0.4
125 to 135	0	0.0
135 to 145	0	0.0
145 to 155	1	0.2
155 to 165	0	0.0
165 to 175	1	0.2
175 to 185	1	0.2
185 to 195	1	0.2
Totals	500	100.0

Table: Manchester adults gross income data.

(a) For a histogram based on the tabulation of the income data above, i.e. with intervals of width 10 starting at 5, calculate the height of the density histogram at $x = 20.0$ and $x = 80.0$.

(b) Now consider the histogram based on intervals of width $h = 20$ starting at $x = 5.0$. Re-tabulate the data accordingly and repeat the calculations of part (a) for the new histogram. Compare your results with part (a).

4. Suppose we have a random sample of observations $x_1, \ldots, x_n$, from a continuous variable X with unknown p.d.f $f_X(x)$. Suppose that the range $(a_1, a_{K+1}]$ contains all the data and is divided into K subintervals $B_k = (a_k, a_{k+1}]$ each of width h. Let ν_k denote the number of observations in B_k.

Recall that the density histogram $\text{Hist}(x)$ based on the bins B_k can be thought of as a data-based estimate of $f_X(x)$. This suggests that the mean of X,

$$\mu = \int_{-\infty}^{\infty} x\, f_X(x) dx\,,$$

could be estimated via

$$\hat{\mu}_{\text{Hist}} = \int_{a_1}^{a_{K+1}} x\, \text{Hist}(x) dx\,.$$

(i) Show that $\hat{\mu}_{\text{Hist}}$ can be expressed as

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

[*Part (i) above has appeared on a previous exam. Parts (ii)–(v) below have not, and are included as challenging optional exercises, indicated by a star.*]

Assume now that there are no repeated values in the data, and that $h < \min_{i \neq j} |x_i - x_j|$.

(ii*) Show that under the above assumption, each B_k contains either one observation or no observations.

(iii*) Under the above assumption, let $B_{k_i} = (a_{k_i}, a_{k_i+1}]$ denote the bin to which x_i belongs. Show that

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

(iv*) Hence show that

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

(v*) Hence find $\lim_{h \to 0} \hat{\mu}_{\text{Hist}}$.

5. The following ordered data are the scores obtained in a particular examination by a random sample of $n = 18$ students:

```
31, 42, 63, 69, 70, 72, 75, 75, 77,
79, 80, 82, 83, 87, 88, 91, 93, 94
```

(i) Calculate the five number summary for these data and also the interquartile range (IQR).

(ii) Use the results of part (i) to produce a boxplot for these data. Comment on the shape of the empirical distribution. Does a normal distribution look like a plausible model for the data?

(iii) Calculate the sample mean, $\bar{x}$ and the sample standard deviation, s for these data. Compare the value of the sample mean with that of the sample median and comment on any discrepancy between the two.

6. (i) [Revision from MATH10141 Probability I] Show that the c.d.f. of a $N(\mu, \sigma^2)$ distribution is

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right) \,,$$

where Φ is the standard normal c.d.f., i.e. the c.d.f. of a $N(0, 1)$ distribution.

(ii) Hence show that the quantile function of a $N(\mu, \sigma^2)$ distribution is

$$Q(p) = \mu + \sigma\Phi^{-1}(p) \ ,$$

and that the interquartile range of a $N(\mu, \sigma^2)$ distribution is

$$\mathrm{IQR}_N = 1.349\sigma \, .$$

The above result suggests that the interquartile range can be estimated using data via

$$\widehat{\mathrm{IQR}}_N = 1.349s \, ,$$

where s denotes the sample standard deviation.

(iii) Using the dataset in Question 5 above, calculate the value of the normal-based estimate, $\widehat{\mathrm{IQR}}_N$, from part (ii) and compare its value with the empirical IQR calculated in Q5(i). Comment on any discrepancy in the results.

7. By considering the case of odd and even n separately, show that $\hat{Q}(0.5)$ is equal to the sample median for the dataset $x_1, \ldots, x_n$.