COMP245: Probability and Statistics 2016 - Problem Sheet 2 Solutions

Numerical Summaries

S1) **Purpose: Link a measure of location with a measure of dispersion.**

To find minima and maxima, we start by differentiating wrt m,

$$\frac{d}{dm}\sum_{i=1}^{n}(x_i - m)^2 = -2\sum_{i=1}^{n}(x_i - m) = 2(mn - \sum_{i=1}^{n} x_i).$$

Setting this equal to zero yields the stationary point of $m = \sum_{i=1}^{n} x_i/n = \bar{x}$. To check this is a minimiser, differentiate again wrt m,

$$\frac{d^2}{dm^2}\sum_{i=1}^{n}(x_i - m)^2 = 2n$$

which is positive for all m. Therefore, $\bar{x}$ is a minimiser of $\sum_{i=1}^{n}(x_i - m)^2$.

S2) **Purpose: Link a measure of location with a measure of dispersion.**

As suggested in the hint, assume all samples are ordered so $x_1 \leq x_2 \leq \ldots \leq x_n$.

The case of $n = 1$ is trivial, and for $n = 2$

$$\sum_{i=1}^{n}|x_i - m| = |x_1 - m| + |x_2 - m| \geq x_2 - x_1$$

with equality attained $\forall m$ in the range $x_1 \leq m \leq x_2$, which includes the median.

Suppose the result holds for all samples up to size n, and now consider an ordered sample of size $n + 2$. First note that the median of $x_2, x_3, \ldots, x_n, x_{n+1}$ is equal to the median of $x_1, x_2, x_3, \ldots, x_n, x_{n+1}, x_{n+2}$ (since in the larger sample we have simply appended a data point on either side), but that the former is a sample of size n. So we wish to show that the median of $x_2, x_3, \ldots, x_n, x_{n+1}$ is a minimiser of $\sum_{i=1}^{n+2}|x_i - m|$. Then

$$\sum_{i=1}^{n+2}|x_i - m| = |x_1 - m| + |x_{n+2} - m| + \sum_{i=2}^{n+1}|x_i - m| \geq x_{n+2} - x_1 + \sum_{i=2}^{n+1}|x_i - m|$$

with equality attained $\forall m$ in the range $x_1 \leq m \leq x_{n+2}$; and clearly the median of $x_2, x_3, \ldots, x_n, x_{n+1}$ lies within this range and is also a minimiser of $\sum_{i=2}^{n+1}|x_i - m|$ by the inductive hypothesis.

S3) **Purpose: This time you need to provide the measure of dispersion.**
A corresponding measure of dispersion would be

$$\sum_{i=1}^{n} I(x_i \neq m).$$

If m is our measure of location of the data, then this measure of dispersion counts how many of the sample take some different value. This would be minimised by the mode.

S4) **Purpose: Practice computing the mean and median of data sets. Also, making you wary of skewed data sets.**
Median $= 110$, mean $= 138.6$.

Because of the right skew.

Because it will be sensitive to the outlying value of 414.

S5) **Purpose: Practice computing the covariance and correlation for ordered pairs of a real data set.**
Differences are: -25.3, -20.5, -10.3, -24.4, -17.5, -30.6, -11.8, -12.9, -3.8, -20.6, -28.4.

Mean $= -18.74$, median $= -20.5$, sd $= 7.94$ (or 8.33).

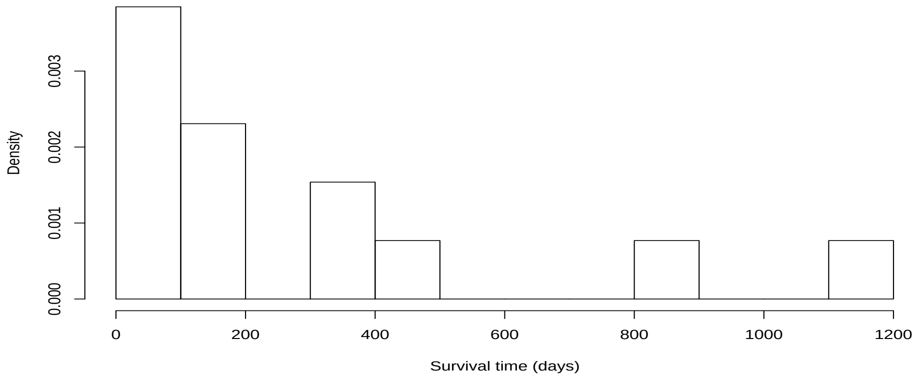
Covariance $= 19.24$, correlation $= 0.51$.

S6) **Purpose: Practice computing interquartile range and median.**
The lower quartile LQ $= x_{((n+1)/4)} = x_{(23/4)}$ which is three quarters of the way between $x_{(5)} = 26.39$ and $x_{(6)} = 27.08$. Hence LQ $= 26.39 + (27.08 - 26.39) \times 3/4 = 26.908$.

Similarly, the upper quartile UQ $= x_{((n+1)\times 3/4)} = x_{(69/4)}$, which is one quarter of the way between $x_{(17)} = 33.28$ and $x_{(18)=33.40}$. Hence, UQ $= 33.28 \times 3/4 + 33.40 \times 1/4 = 33.31$.

The median is $x_{((n+1)/2)} = x_{(23/2)} = 28.69/2 + 29.36/2 = 29.0$.

S7) **Purpose: Practice plotting a histogram and transforming skewed data.**



Mean = 286, sd = 332.72 (or 346.3 for $\frac{1}{n-1}$ formula).

Because of the skewness of the data.

Skewness = 1.43 (or similar); Skewness of log transformed data = 0.26 (or similar).

S8) **Purpose: Example using the harmonic mean.**

The car travels a total of 20 miles in (10/30 hours plus 10/60 hours). That is, 20 miles in 0.5 hours. That is, 40 miles per hour. (Not (30+60)/2.)

This can be most simply calculated using the harmonic mean,

$$\frac{2}{\frac{1}{30}+\frac{1}{60}} = \frac{2}{\frac{3}{60}} = 40.$$