Indian Statistical Institute, Bangalore
B.Math (II)
First Semester 2012-2013
Mid-Semester Examination : Statistics (I)
Answer Script

1. The data below show IQ scores for 30 sixth graders.

088	102	126	095	109	099
102	151	115	097	092	107
081	119	094	090	109	099
102	117	098	093	105	084
114	122	087	094	101	081

(a) Make a stem and leaf plot of these data.

(b) Find the sample mean $\overline{X}$.

(c) Find the sample standard deviation S.

(d) Find the sample median M.

(e) Find $100p$-th percentile for $p = 0.25$ and $p = 0.75$.

(f) Find the second quartile Q_2.

(g) What proportion of the data lies within $\overline{X} \pm 3S$.

(h) Draw the box plot and identify the outliers.

(i) Decide on trimming fraction just enough to eliminate the outliers and obtain the trimmed mean $\overline{X}_{\mathrm{T}}$.

(j) Also obtain the trimmed standard deviation S_{T}.

(k) Between the box plot and the stem and leaf plot what do they tell us about the above data set ? In very general terms what can you say about the population from which the data have arrived.

Solution.

(a) A stem and leaf plot for the above data is given below. The stem gives the tens place and the leaf gives the ones place of the given data respectively.

8	1	1	4	7	8					
9	0	2	3	4	4	5	7	8	9	9
10	1	2	2	2	5	7	9	9		
11	4	5	7	9						
12	2	6								
13										
14										
15	1									

(b) Let, the above data of IQ scores for 30 sixth grader is denoted as $\mathbf{x} = (x_1, x_2, \ldots, x_{30})$. The sample mean of $\mathbf{x}$ is

$$\overline{X} = \frac{1}{30}\sum_{i=1}^{30} x_i = 102.4333.$$

(c) The sample standard deviation of $\mathbf{x}$ is

$$S = \sqrt{\frac{1}{30}\sum_{i=1}^{30}(x_i - \overline{X})^2} = 14.74151.$$

(d) Let us first sort the sample $\mathbf{x}$ in ascending order.

81	81	84	87	88	90	92	93	94	94	95	97	98	99	99
101	102	102	102	105	107	109	109	114	115	117	119	122	126	151

We denote the above sorted sample as $\mathbf{x}^{(s)}$. As the sample size is an even number (30), the sample median M is calculated as

$$M = [(30/2)^{\text{th}} \text{ entry in } \mathbf{x}^{(s)} + (30/2+1)^{\text{th}} \text{ entry in } \mathbf{x}^{(s)}]/2 = (99+101)/2 = 100.$$

(e) The rank of $100p$-th percentile for $p = 0.25$ i.e the rank of 25^{th} percentile is $\lceil 30 \times 0.25 \rceil = \lceil 7.5 \rceil = 8$. The 8^{th} entry in sorted sample $\mathbf{x}^{(s)}$ is 93. The proportion of values below 93 is $7/30 = 0.23\,(< 0.25)$ and the proportion of values less than or equal to 93 is $8/30 = 0.27\,(> 0.25)$. So the 25^{th} percentile is 93.

We find the $100p$-th percentile for $p = 0.75$ similarly. The rank of $100p$-th percentile for $p = 0.75$ i.e the rank of 75^{th} percentile is $\lceil 30 \times 0.25 \rceil = \lceil 22.5 \rceil = 23$. The 23^{rd} entry in sorted sample $\mathbf{x}^{(s)}$ is 109. The proportion of values below 109 is $21/30 = 0.70\,(< 0.75)$ and the proportion of values less than or equal to 109 is $23/30 = 0.77\,(> 0.75)$. So the 75^{th} percentile is 109.

(f) As the sample size is an even number (30), the rank of the second quartile Q_2 is $30 \times (2/4) = 15$ (integer). The 15^{th} entry in sorted sample $\mathbf{x}^{(s)}$ is 99. The proportion of values below 99 is $13/30 = 0.43\,(< 0.50)$ and the proportion of values less than or equal to 99 is $15/30 = 0.50$. Now the proportion of values above 101 is $14/30 = 0.47\,(< 0.50)$ the proportion of values greater than or equal to 101 is $15/30 = 0.50$. Both 15^{th} entry (99) and 16^{th} entry (101) in $\mathbf{x}^{(s)}$ satisfy the conditions of second quartile. We compute second quartile as $(99+101)/2 = 100$, which is same as the median. Although any value between 99 and 101 could be taken as second quartile Q_2.

(g) First compute the boundaries

$$\overline{X} - 3S = 58.20879,\ \overline{X} + 3S = 146.65787.$$

Other than the 8^{th} entry in $\mathbf{x}$ (151), all the other 29 entries lie within $\overline{X} \pm 3S$, so the proportion of the data lies within $\overline{X} \pm 3S$ is $29/30 = 0.9667$.

(h) The box plot for the given data on IQ scores is shown in Figure 1. The box has three vertical lines at first quartile (Q_1), second quartile (Q_2) and third quartile

(Q_3) respectively. Interquartile range is defined as ($Q_3 - Q_1$), here $109 - 93 = 16$. Whiskers are the smallest and largest data points lying within the the interval $(Q_1 - 1.5 \times IQR,\ Q_3 + 1.5 \times IQR) = (69, 133)$. The two whiskers for the given data are 81 and 126 (also shown in the box plot below). Outlier is defined as any data point either less than the lower fence or greater than the upper fence. There is only one outlier in the given data, which is 151.

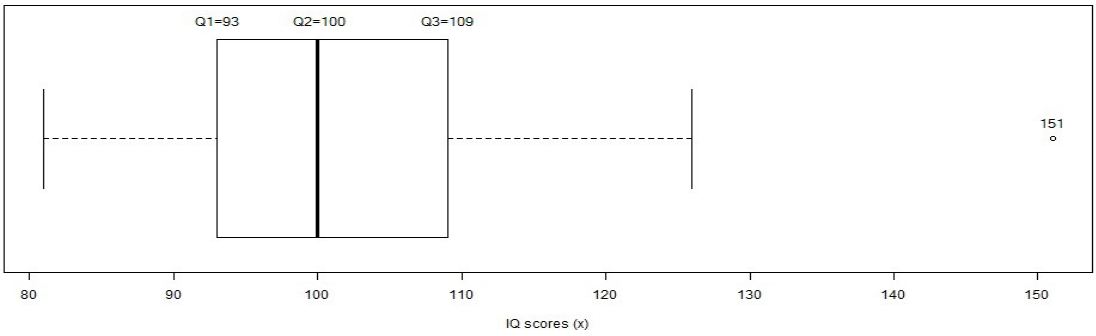

Figure 1: Boxplot of IQ scores

(i) We take trimming fraction to be 0.04, which excludes one observation from each end of $\mathbf{x}^{(s)}$. The trimmed mean is calculated as

$$\overline{X}_{\mathrm{T}} = \frac{1}{28}\sum_{i=2}^{29} x_i^{(s)} = 101.4643.$$

(j) The trimmed standard deviation with trimming fraction as 0.04 is

$$S_{\mathrm{T}} = \sqrt{\frac{1}{28}\sum_{i=2}^{29}(x_i^{(s)} - \overline{X}_{\mathrm{T}})^2} = 11.45638.$$

(k) The stem and leaf plot shows that the number of data points lie below 100 is 15, same as the number of data points lie above 100 . But, the number of rows in stem and leaf plot for data points below 100 is only two, whereas the number of rows for data point above 100 is 6 (although there is not data points in the rows of 130s and 140s, number of rows with data points are still more than the number of rows below 100). This shows a increase in spread above the median. The box plot shows that, the distance between Q_3 and Q_2 $(109 - 100 = 9)$ is more than the distance between Q_1 and Q_2 $(100 - 93 = 7)$, which indicates positive skewness (a longer right tail). Based on the given data, the population from which the data have arrived can be taken as positively skewed.

2. For random variables X and Y define the correlation coefficient ρ_{XY}. If the joint density function of the two random variables X and Y, for $a > 0$, $\lambda > 0$, $b \in \mathbb{R}$, is given by

$$f(x, y|\lambda) = \frac{1}{a+\lambda};\ 0 < x < 1,\ ax + b < y < ax + b + a + \lambda;$$

then find ρ_{XY}. Find $\lim_{\lambda \to 0} \rho_{XY}$ and $\lim_{\lambda \to \infty} \rho_{XY}$.

Solution. The *correlation coefficient* between two random variables X and Y, denoted by ρ_{XY}, is defined as

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}},$$

provided $\mathrm{Var}(X)$ and $\mathrm{Var}(Y)$ are both positive.

The mean of X is given by

$$E(X) = \int_0^1 \int_{ax+b}^{ax+b+a+\lambda} x\, f(x,y|\lambda)\, dy\, dx = \int_0^1 x\, dx = \frac{1}{2}.$$

The mean of Y is given by

$$\begin{aligned} E(Y) &= \int_0^1 \int_{ax+b}^{ax+b+a+\lambda} y\, f(x,y|\lambda)\, dy\, dx \\ &= \int_0^1 \frac{(ax+b+a+\lambda)^2 - (ax+b)^2}{2(a+\lambda)}\, dx \\ &= a + b + \frac{\lambda}{2}. \end{aligned}$$

Similarly we find

$$\begin{aligned} E(X^2) &= \int_0^1 \int_{ax+b}^{ax+b+a+\lambda} x^2\, f(x,y|\lambda)\, dy\, dx = \int_0^1 x^2\, dx = \frac{1}{3}, \\ E(Y^2) &= \int_0^1 \int_{ax+b}^{ax+b+a+\lambda} y^2\, f(x,y|\lambda)\, dy\, dx \\ &= \int_0^1 \frac{(ax+b+a+\lambda)^3 - (ax+b)^3}{3(a+\lambda)}\, dx \\ &= 7a^2/6 + b^2 + \lambda^2/3 + 2ab + 7a\lambda/6 + b\lambda, \\ E(XY) &= \int_0^1 \int_{ax+b}^{ax+b+a+\lambda} xy\, f(x,y|\lambda)\, dy\, dx \\ &= \int_0^1 x\, \frac{(ax+b+a+\lambda)^2 - (ax+b)^2}{2(a+\lambda)}\, dx \\ &= 7a/12 + b/2 + \lambda/4. \end{aligned}$$

Using this quantities we get

$$\begin{aligned} &\mathrm{Var}(X) = E(X^2) - (E(X))^2 = 1/12, \\ &\mathrm{Var}(Y) = E(Y^2) - (E(Y))^2 = (2a^2 + 2a\lambda + \lambda^2)/12, \\ &\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y) = a/12. \end{aligned}$$

Note that, both $\mathrm{Var}(X)$ and $\mathrm{Var}(Y)$ are positive. Finally we obtain the *correlation coefficient* between X and Y as

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = \frac{a}{\sqrt{2a^2 + 2a\lambda + \lambda^2}}$$

with $\lim_{\lambda\to 0} \rho_{XY} = \dfrac{a}{\lim_{\lambda\to 0}\sqrt{2a^2 + 2a\lambda + \lambda^2}} = \dfrac{1}{\sqrt{2}}$ and $\lim_{\lambda\to\infty} \rho_{XY} = \dfrac{a}{\lim_{\lambda\to\infty}\sqrt{2a^2 + 2a\lambda + \lambda^2}} = 0.$

3. Let $X_1, X_2, \ldots, X_n$ be a random sample from the distribution with *pdf* given by

$$f(x|\theta_1, \theta_2) = \frac{1}{\theta_2 - \theta_1} I_{(\theta_1, \theta_2)}(x);\ \theta_1 < \theta_2 \in \mathbb{R}.$$

Find *maximum likelihood estimators (MLE)* for θ_1 and θ_2.

Solution. Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)'$ be a random sample from $Uniform(\theta_1, \theta_2)$. To derive *MLEs* for θ_1 and θ_2 a suitable version of the *pdf* is chosen,

$$f(x|\theta_1, \theta_2) = \frac{1}{\theta_2 - \theta_1} I_{[\theta_1, \theta_2]}(x);\ \theta_1 < \theta_2 \in \mathbb{R}.$$

The likelihood function of the random sample $\mathbf{X}$ is

$$L(\theta_1, \theta_2|\mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\theta_2 - \theta_1} I_{[\theta_1, \theta_2]}(x_i) = \Big(\frac{1}{\theta_2 - \theta_1}\Big)^n I_{(-\infty, x_{(1)}]}(\theta_1) \times I_{[x_{(n)}, \infty)}(\theta_2),$$

Here $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ are the *order statistic* and $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ are the *ordered sample values*. Thus we observe that the likelihood function $L(\theta_1, \theta_2|\mathbf{x})$ is a decreasing function of $(\theta_2 - \theta_1)$. Observe that, $(\theta_2 - \theta_1)$ is minimized subject to $I_{(-\infty, x_{(1)}]}(\theta_1) \times I_{[x_{(n)}, \infty)}(\theta_2) = 1$, at $\theta_1 = x_{(1)}$ and $\theta_2 = x_{(n)}$. Thus the *MLEs* of θ_1 and θ_2 are given by

$$\hat{\theta}_{1\text{MLE}} = x_{(1)},\ \hat{\theta}_{2\text{MLE}} = x_{(n)}.$$

4. Let Y be distributed as $\varepsilon xp(\lambda)$, $\lambda > 0$. Define $X = \alpha e^Y$, $\alpha > 0$. Obtain the distribution F_X of X. Suppose we have a random sample $X_1, X_2, \ldots, X_n$ from F_X, $\alpha > 0$ and $\lambda > 0$ both unknown. Obtain *method of moments (MOM)* estimators for α and λ based on $X_1, X_2, \ldots, X_n$. Also obtain *maximum likelihood estimators (MLE)* for α and λ

Solution. Y is distributed as $\varepsilon xp(\lambda)$, $\lambda > 0$ with density function

$$\begin{aligned} f(y|\lambda) &= \lambda e^{-\lambda y},\ \text{ for } y > 0 \\ &= 0,\ \text{ for } y \leq 0. \end{aligned}$$

The *cdf* of Y is

$$\begin{aligned} F_Y(y) = P(Y \leq y) &= 1 - e^{-\lambda y},\ \text{ for } y > 0 \\ &= 0,\ \text{ for } y \leq 0. \end{aligned}$$

The *cdf* of $X = \alpha e^Y$, $\alpha > 0$ is

$$\begin{aligned} F_X(x) = P(X \leq x) = P(Y \leq \log(x/\lambda)) &= 1 - (\alpha/x)^\lambda,\ \text{ for } x > \alpha \\ &= 0,\ \text{ for } x \leq \alpha. \end{aligned}$$

The corresponding *pdf* of X can be found by applying differentiation on $F_X(x)$ with respect to x,

$$\begin{aligned} f_X(x|\alpha, \lambda) &= \lambda \alpha^\lambda x^{-(\lambda+1)},\ \text{ for } x > \alpha \\ &= 0,\ \text{ for } x \leq \alpha. \end{aligned}$$

The raw population moment of order m is

$$E(X^m) = \int_\alpha^\infty x^m f_X(x|\alpha,\lambda)\,dx = \frac{\lambda\alpha^m}{\lambda - m},$$

provided $\lambda > m$. We have a random sample X_1, X_2, ..., X_n from F_X. We define

$$m_1' = \bar{x} = \frac{1}{n}\sum_{i=1}^n x_i,\ m_2' = \frac{1}{n}\sum_{i=1}^n x_i^2,\ s^2 = m_2' - (m_1')^2 = \sum_{i=1}^n (x_i - \bar{x})^2.$$

To find *method of moments (MOM)* estimators we equate the first and second order sample raw moments to the corresponding population raw moments, that is,

$$m_1' = \frac{\lambda\alpha}{\lambda - 1},\ m_2' = \frac{\lambda\alpha^2}{\lambda - 2}.$$

For $\lambda > 2$, these two equations have unique set of solution for α and λ

$$\hat{\alpha} = \frac{m_1'\sqrt{1 + m_1'/s^2}}{1 + \sqrt{1 + m_1'/s^2}},\ \hat{\lambda} = 1 + \sqrt{1 + m_1'/s^2}.$$

$\hat{\alpha}$ and $\hat{\lambda}$ are *method of moments (MOM)* estimators for α and λ.

The likelihood function of the random sample $X_1, X_2, \ldots, X_n$ is

$$L(\alpha,\lambda|\mathbf{x}) = \prod_{i=1}^n f_X(x_i|\alpha,\lambda) = \lambda^n\alpha^{n\lambda}\prod_{i=1}^n x_i^{-(\lambda+1)}.$$

After taking logarithm on $L(\alpha,\lambda|\mathbf{x})$ we have

$$\log L(\alpha,\lambda|\mathbf{x}) = n\log\lambda + n\lambda\log\alpha - (\lambda+1)\sum_{i=1}^n \log x_i.$$

We differentiate $\log L(\alpha,\lambda|\mathbf{x})$ with respect to λ

$$\frac{d}{d\lambda}\log L(\alpha,\lambda|\mathbf{x}) = \frac{n}{\lambda} + n\log\alpha - \sum_{i=1}^n \log x_i$$

and equate it with 0 to obtain *maximum likelihood estimator (MLE)* of λ

$$\hat{\lambda}_{MLE} = \frac{n}{\sum_{i=1}^n \log(x_i/\hat{\alpha}_{MLE})},$$

where $\hat{\alpha}$ is *MLE* of α. Note that, *MLE* of α can not obtained by differentiating $\log L(\alpha,\lambda|\mathbf{x})$ with respect to α since $\log L(\alpha,\lambda|\mathbf{x})$ is unbounded with respect to α. But since $\alpha \le x_i$ for each $i = 1, 2, \ldots, n$, we may maximize $\log L(\alpha,\lambda|\mathbf{x})$ subject to the constraint $\alpha \le \min_i x_i$. The likelihood $L(\alpha,\lambda|\mathbf{x})$ is maximized with respect to α subject to $\alpha \le \min_i x_i$ at $\hat{\alpha}_{MLE} = \min_i x_i$, which is *MLE* of α. Thus the *MLEs* of α and λ are given by

$$\hat{\alpha}_{MLE} = \min_i x_i,\ \hat{\lambda}_{MLE} = \frac{n}{\sum_{i=1}^n \log(x_i/\hat{\alpha}_{MLE})}.$$