1. *Consider a sequence of binomial distributions $Bin(n, \theta_n)$, $n \geq 1$. Let $\theta_n \to 0$ as $n \to \infty$, such that $n\theta_n \to \lambda > 0$ as $n \to \infty$. Let the mean and variance of $Bin(n, \theta_n)$ be denoted respectively by μ_n and σ_n^2. Find $\lim_{n\to\infty} \mu_n$ and $\lim_{n\to\infty} \sigma_n^2$. Also obtain limit of pmf of binomial distribution, $Bin(n, \theta_n)$ as $n \to \infty$ and identify the limiting distribution.*

   Solution: Let $X_n \sim Bin(n, \theta_n)$. The mean is:

$$E(X_n) = \sum_{x=0}^{n} \frac{n!}{x!(n-x)!} x\theta_n^x (1-\theta_n)^{(n-x)} = n\theta_n \sum_{y=0}^{n-1} \frac{(n-1)!}{y!(n-1-y)!} \theta_n^y (1-\theta_n)^{n-1-y} = n\theta_n.$$

$$\begin{aligned} E(X_n(X_n-1)) &= \sum_{x=2}^{\infty} \frac{n!}{(x-2)!(n-x)!} \theta_n^x (1-\theta_n)^{(n-x)} \\ &= n(n-1)\theta_n^2 \sum_{y=0}^{\infty} \frac{(n-2)!}{y!(n-2-y)!} \theta_n^y (1-\theta_n)^{(n-2-y)} = n(n-1)\theta_n^2. \end{aligned}$$

   The variance is:

$$\begin{aligned} Var(X_n) &= E(X_n(X_n-1)) + E(X_n) - [E(X_n)]^2 \\ &= n(n-1)\theta_n^2 + n\theta_n - n^2\theta_n^2 = n\theta_n(1-\theta_n). \end{aligned}$$

   From the above and under the given assumptions, we have

$$\lim_{n\to\infty} \mu_n = \lim_{n\to\infty} n\theta_n = \lambda \text{ and } \lim_{n\to\infty} \sigma_n^2 = \lim_{n\to\infty} n\theta_n(1-\theta_n) = \lambda.$$

   The probability mass function (pmf) of $Bin(n, \theta_n)$ is given by

$$f_n(x|\theta_n) = \binom{n}{x} \theta_n^x (1-\theta_n)^{(n-x)} \ ; x = 0, 1, 2, \cdots, n.$$

   Under the assumptions $\theta_n \to 0$ as $n \to \infty$, such that $n\theta_n \to \lambda > 0$ as $n \to \infty$, we have

$$\begin{aligned} \lim_{n\to\infty} f_n(x|\theta_n) &= \lim_{n\to\infty} \binom{n}{x} \theta_n^x (1-\theta_n)^{(n-x)} \\ &= \lim_{n\to\infty} \frac{n(n-1)\cdots(n-x+1)!}{n^x} \frac{(n\theta_n)^x}{x!} \Big(1 - \frac{n\theta_n}{n}\Big)^n \\ &= \frac{1}{x!}\lambda^x \lim_{n\to\infty} \Big(1 - \frac{n\theta_n}{n}\Big)^n \\ &= \frac{\lambda^x}{x!} exp(-\lambda), x = 0, 1, \cdots. \end{aligned}$$

   The limiting distribution of $Bin(n, \theta_n)$ as $n \to \infty$ under the given assumptions is Poisson distribution with mean λ.

   □

2. *Let $X_1, X_2, \cdots, X_n$ be a random sample from the distribution whose probability density function (pdf) is proportional to*

$$f(x|\theta) = (x-\theta)^{a-1}(1+\theta-x)^{b-1} I_{(\theta,\theta+1)}(x) \ ; -\infty < \theta < \infty \text{ unknown and } a, b > 0 \text{ known.}$$

Obtain $E[X_1^r]$. Find the method of moments (MOM) estimator for θ. Find maximum likelihood estimator (MLE) for θ.

Solution: The pdf of X_1 is

$$f(x|\theta) = \frac{1}{\beta(a,b)}(x-\theta)^{a-1}(1-(x-\theta))^{b-1}, \text{ for } \theta < x < \theta + 1.$$

The r^{th} raw population moment is:

$$\begin{aligned}
E(X_1^r) &= \frac{1}{\beta(a,b)} \int_\theta^{1+\theta} x^r (x-\theta)^{a-1}(1-(x-\theta))^{b-1} dx \\
&= \frac{1}{\beta(a,b)} \int_0^1 (y+\theta)^r y^{a-1}(1-y)^{b-1} dy \\
&= \frac{1}{\beta(a,b)} \Big[\int_0^1 y^{a-1}(1-y)^{b-1} \big(y^r + \binom{r}{1} y^{r-1}\theta + \cdots + \theta^r \big) \Big] dy \\
&= \frac{1}{\beta(a,b)} \Big[\beta(r+a,b) + r\theta\beta(r+a-1,b) + \binom{r}{2}\theta^2\beta(r+a-2,b) + \cdots + \theta^r\beta(a,b) \Big].
\end{aligned}$$

To find the method of moments (MOM) estimator for θ, we equate the first sample raw moment to the first population raw moment.

The sample mean (first sample raw moment) is $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and the population first moment is

$$E(X_1) = \frac{\beta(a+1,b)}{\beta(a,b)} + \theta = \frac{a}{a+b} + \theta.$$

The MOM estimator for θ, say $\hat{\theta}_{MOM,n}$, is

$$\hat{\theta}_{MOM,n} = \bar{X}_n - \frac{a}{a+b}.$$

Let $x_{(1)}, \cdots, x_{(n)}$ be the ordered sample values. The Likelihood function is defined as

$$\begin{aligned}
L(\theta|\mathbf{x}) = L(\theta|x_1, \cdots, x_n) &= \prod_{i=1}^n f(x_i|\theta) \\
&= \Big[\frac{1}{\beta(a,b)}\Big]^n \prod_{i=1}^n (x_i-\theta)^{a-1}(1+\theta-x_i)^{b-1} \text{ for } x_{(1)} > \theta, x_{(n)} - 1 < \theta,
\end{aligned}$$

and is 0 otherwise.

(i) Let $a = b = 1$, the likelihood function becomes

$$L(\theta|\mathbf{x}) = L(\theta|x_1, \cdots, x_n) = \prod_{i=1}^n f(x_i|\theta) = I[x_{(n)} - 1 < \theta < x_{(1)}].$$

The MLE of θ in this case is not unique. The likelihood is constant in the above given interval. Any point in the interval $(x_{(n)}-1, x_{(1)})$ can be taken as a ML estimate for θ. For example, a possible choice for MLE for θ can be $(X_{(1)}+X_{(n)}-1)/2$.

(ii) Let $a=1, b>1$, the likelihood is increasing in θ. The MLE in this case is $X_{(1)}$.

(iii) Let $a>1, b=1$, the likelihood is decreasing in θ. The MLE in this case is $X_{(n)}-1$.

(iv) Let $a<1, b<1$, then the MLE can be taken as either $X_{(1)}$ or $X_{(n)}-1$.

(v) Let $a>1, b>1$. The MLE is unique and is the root of the following equation

$$\frac{\partial log(L(\theta|\mathbf{x}))}{\partial\theta} = -(a-1)\sum_{i=1}^{n}\frac{1}{x_i-\theta} + (b-1)\sum_{i=1}^{n}\frac{1}{1+\theta-x_i} = 0.$$

which lies in the above interval $(X_{(n)}-1, X_{(1)})$.

□

3. *A discrete model that is often used for the waiting time X to failure of an item is given by the pmf*

$$f(k|\theta) = \theta^{(k-1)}(1-\theta); k = 1, 2, \cdots; 0<\theta<1.$$

Suppose that we only record the time of failure, if the failure occurs on or before r and otherwise just note that the item has atleast lived $r+1$ periods. Let Y denote this censored waiting time. Write down the pmf of Y. If $Y_1, \cdots, Y_n$ is a random sample from this censored waiting time distribution, obtain method of moments (MOM) estimator for θ.

Solution: $Y_1 = X_1$, if $X_1 \leq r$ and $Y_1 = r+1$, if $X_1 > r$. The pmf was Y_1 is

$$f_{Y_1}(x|\theta) = \begin{cases} \theta^{(x-1)}(1-\theta) & \text{if } x = 1, 2, \cdots, r \\ \theta^r & x = r+1. \end{cases}$$

Note that, $P(X_1 > r) = \sum_{j=r+1}^{\infty}\theta^{j-1}(1-\theta) = \theta^r$. To find the method of moments (MOM) estimator for θ, we equate the first sample raw moment to the first population raw moment.

The sample mean (first sample raw moment) is $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n}Y_i$ and the population first moment is

$$E(Y_1) = \sum_{x=1}^{r} x\theta^{(x-1)}(1-\theta) + (r+1)\theta^r = 1+\theta+\theta^2+\cdots+\theta^r = \frac{(1-\theta^{r+1})}{(1-\theta)}.$$

The MOM estimator for θ, say $\hat{\theta}_{MOM,n}$, is the viable solution to the equation

$$\frac{(1-\hat{\theta}^{r+1}_{MOM,n})}{(1-\hat{\theta}_{MOM,n})} = \bar{Y}_n.$$

□

4. *Consider the following data set*

$$127, 144, 93, 184, 79, 53, 195, 57, 202, 80, 95, 161, 204, 108, 124,$$
$$67, 102, 310, 177, 89, 146, 110, 141, 94, 118, 69, 63, 100, 207, 160.$$

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

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

(c) Find the $100p$ percentile for $p = 0.75$.

(d) Find the median M and the third quartile Q_3.

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

(f) Explain how to obtain the trimmed mean $\bar{X}_T$. Decide on trimming fraction just enough to eliminate the outliers and obtain the trimmed median $\bar{M}_T$.

(g) Explain (need not compute) how to obtain the trimmed standard deviation S_T.

(h) 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 arrived?

Solution: The sorted sample data in increasing order is

$$53, 57, 63, 67, 69, 79, 80, 89, 93, 94, 95, 100, 102, 108, 110, 118, 124, 127, 141, 144, 146,$$
$$160, 161, 177, 184, 195, 202, 204, 207, 310.$$

(a) Following is the steam and leaf plot for the data set.

Stem and leaf plot, Stem: tens and hundred's digits, Leaf: ones digits.

$$\begin{array}{r|l} 5 & 3, 7 \\ 6 & 3, 7, 9 \\ 7 & 9 \\ 8 & 0, 9 \\ 9 & 3, 4, 5 \\ 10 & 0, 2, 8 \\ 11 & 0, 8 \\ 12 & 4, 7 \\ 13 & \\ 14 & 1, 4, 6 \\ 15 & \\ 16 & 0, 1 \\ 17 & 7 \\ 18 & 4 \\ 19 & 5 \\ 20 & 2, 4, 7 \\ 21 & \\ \vdots & \\ 31 & 0 \end{array}$$

(b) Let $x_1, \cdots, x_n$ denote the sample values, $n = 30$. Then, the sample mean is

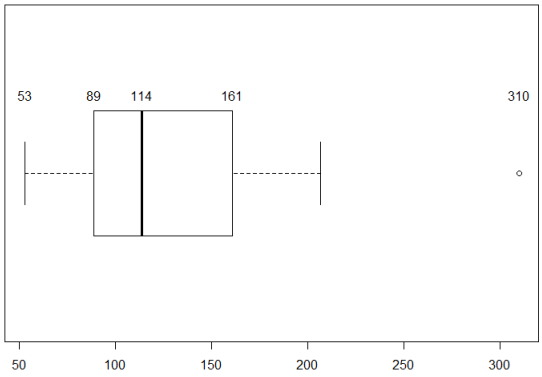
$$\frac{1}{30}\sum_{i=1}^{30} x_i = 128.6333.$$

(c) The index of 75th percentile is $\lceil 0.75 * 30 \rceil = 23$ in the sorted sample. 161 is the 23^{rd} entry in the sorted sample. Hence, the 75^{th} percentile is 161.

(d) The median M and the third quartile Q_3 (or the 75^{th} percentile) are 114 and 161, respectively.

(e) The box plot of the data is given in Figure 1. The 3 vertical lines give the first quartile Q_1, the median and the third quartile Q_3. The index of 25th percentile is $\lceil 0.25 * 30 \rceil = 8$ in the sorted sample. Q_1 or the 25^{th} percentile is 89. The interquartile range is $IQR = Q_3 - Q_1 = 72$. The upper whisker is located at $min(max(x_1, \cdots, x_n), Q_3 + 1.5 * IQR)$ and the lower whisker is located at $max(min(x_1, \cdots, x_n), Q_1 - 1.5 * IQR)$. Here, only the lower whisker coincides with the minimum of the data values. There is one data point lying outside the interval $(Q_1 - 1.5 * IQR, Q_3 + 1.5 * IQR)$. Hence, the outlier is 310.

Figure 1: Box plot of the 30 data points



(f) Sort the data in increasing order. Remove $T\%$ of the observations from each end. Calculate the sample mean of the remaining observations. The resulting quantity is the Trimmed mean. Let the sorted observations be denoted as $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$. Say, $T\%$ observations are removed from each end of the sorted sample, *i.e.* we remove $t = \lfloor nT/100 \rfloor$ (greatest integer less than or equal to $nT/100$) observations from each end. Then, the trimmed mean is

$$\bar{X}_T = \frac{1}{n-2t}\sum_{i=t+1}^{n-t} x_{(i)}.$$

As there is one outlier, we choose $t = 1 = \lfloor n \times 4/100 \rfloor$. Hence, the trimmed mean is

$$\bar{X}_T = \frac{1}{n-2}\sum_{i=2}^{n-2} x_{(i)} = 124.8571.$$

The trimmed median is the mean of the $\lceil 0.5 * 28 \rceil = 14^{th}$ and the 15^{th} entries of the sorted sample after removing the first and the last observation in it. This is same as M, *i.e.* $M_T = 114$.

(g) The trimmed standard deviation S_T is obtained as follows. Then, the trimmed standard deviation S_T is

$$S_T = \sqrt{\frac{1}{n-2t} \sum_{i=t+1}^{n-t} (x_{(i)} - \bar{X}_T)^2}.$$

(h) The stem and leaf plot shows that the number of points less than 114 are 15 and the number of data points greater than 114 are also 15. The data is more spread above the median. The box plot indicates postive skewness (the right side tail of the distribution is longer than the left). The distance between $Q_3 - M = 47$ is more than the distance between $M - Q_1 = 25$. Based on the data, the population from which it arrived from can be taken as positively skewed.

□