1. *The probability mass function (pmf) of $NegBin(r,\theta)$ is given by*

$$f(x|r,\theta) = \binom{r+x-1}{x}\theta^r(1-\theta)^x \ ; x = 0, 1, 2, \cdots .$$

Obtain mean and variance of $NegBin(r,\theta)$. Consider a sequence of negative binomial distribution $NegBin(r_n,\theta_n)$, $n \geq 1$. Let $r_n \to \infty$, and $\theta_n \to 1$ as $n \to \infty$ such that $\frac{r_n(1-\theta_n)}{\theta_n} \to \lambda > 0$ as $n \to \infty$. Let the mean and variance of $NegBin(r,\theta)$ be denoted respectively by μ_n and σ_n^2. Find $\underset{n\to\infty}{lim}\,\mu_n$ and $\underset{n\to\infty}{lim}\,\sigma_n^2$. Also obtain limit of pmf of negative binomial distribution $NegBin(r_n,\theta_n)$, as $n \to \infty$ and identify the limiting distribution.

Solution: Let $X \sim NegBin(r,\theta)$. The mean is:

$$\begin{aligned} E(X) &= \sum_{x=0}^{\infty} \frac{(r+x-1)!}{x!(r-1)!} x(1-\theta)^x\theta^r = \sum_{y=0}^{\infty} \frac{(r+y)!}{y!(r-1)!}(1-\theta)^{y+1}\theta^r \\ &= \frac{r(1-\theta)}{\theta}\sum_{y=0}^{\infty}\frac{(r+1+y-1)!}{y!r!}(1-\theta)^y\theta^{r+1} = \frac{r(1-\theta)}{\theta}. \end{aligned}$$

$$\begin{aligned} E(X(X-1)) &= \sum_{x=0}^{\infty} \frac{(r+x-1)!}{x!(r-1)!} x(x-1)(1-\theta)^x\theta^r = \sum_{y=0}^{\infty} \frac{(r+y+2-1)!}{y!(r-1)!}(1-\theta)^{y+2}\theta^r \\ &= \frac{r(r+1)(1-\theta)^2}{\theta^2}\sum_{y=0}^{\infty}\frac{(r+2+y-1)!}{y!(r+1)!}(1-\theta)^y\theta^{r+2} = \frac{r(r+1)(1-\theta)^2}{\theta^2}. \end{aligned}$$

The variance is:

$$\begin{aligned} Var(X) &= E(X(X-1)) + E(X) - [E(X)]^2 \\ &= \frac{r(1-\theta)}{\theta}\Big[\frac{1-\theta}{\theta}+1\Big] = \frac{r(1-\theta)}{\theta^2}. \end{aligned}$$

From the above and under the given assumptions, we have

$$\underset{n\to\infty}{lim}\,\mu_n = \underset{n\to\infty}{lim}\frac{r_n(1-\theta_n)}{\theta_n} = \lambda \text{ and } \underset{n\to\infty}{lim}\,\sigma_n^2 = \underset{n\to\infty}{lim}\frac{r_n(1-\theta_n)}{\theta_n^2} = \lambda.$$

The probability mass function (pmf) of $NegBin(r_n,\theta_n)$ is given by

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

Under the assumptions $r_n \to \infty$, and $\theta_n \to 1$ as $n \to \infty$ such that $\frac{r_n(1-\theta_n)}{\theta_n} \to \lambda > 0$ as $n \to \infty$, we have

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

The limiting distribution of $NegBin(r_n, \theta_n)$ 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 given by*

$$f(x|\theta) = 2(m+1)(x-\theta)^{2m+1} \ ;\theta < x < \theta+1, -\infty < \theta < \infty.$$

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

Solution: The r^{th} raw population moment is:

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

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

$$E(X_1) = \frac{2(m+1)}{2m+3} + \theta.$$

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

$$\hat{\theta}_{MOM,n} = \bar{X}_n - \frac{2(m+1)}{2m+3}.$$

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

$$L(\theta|\mathbf{x}) = L(\theta|x_1, \cdots, x_n) = \prod_{i=1}^{n} f(x_i|\theta) = (2(m+1))^n \prod_{i=1}^{n} (x_i - \theta)^{2m+1} \text{ for } \theta < x_1, \cdots, x_n < \theta + 1.$$

$$= (2(m+1))^n \prod_{i=1}^{n} (x_i - \theta)^{2m+1} \text{ for } x_{(1)} > \theta, x_{(n)} - 1 < \theta,$$

and is 0 otherwise.

The likelihood function is decreasing in θ in the given interval. Thus, the MLE of θ is $X_{(n)} - 1$.

□

3. *Let $X_1 \sim \chi^2_m$ and $X_2 \sim \chi^2_n$ be independent. Define $Y = X_1$ and $W = \frac{n}{m}\frac{X_1}{X_2}$. Obtain the joint density function of $f_{YW}(y, w)$ of Y and W. Hence obtain the marginal density function $f_W(w)$ of W and identify it.*

Solution: The joint density of X_1 and X_2 is given by

$$f_{X_1,X_2}(x_1, x_2) = \frac{exp(-\frac{x_1+x_2}{2})}{2^{\frac{m+n}{2}}\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} x_1^{\frac{m}{2}-1} x_2^{\frac{n}{2}-1}, \text{ for } 0 \leq x_1, x_2 < \infty.$$

Let us make the following transformations:
$y = x_1$ and $w = \frac{nx_1}{mx_2}$, so that $x_1 = y$ and $x_2 = \frac{ny}{mw}$.
The Jacobian of the transformation is

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial w} & \frac{\partial x_1}{\partial y} \\ \frac{\partial x_2}{\partial w} & \frac{\partial x_2}{\partial y} \end{vmatrix} = \begin{vmatrix} 0 & 1 \\ -\frac{ny}{mw^2} & \frac{n}{mw} \end{vmatrix} = \frac{ny}{mw^2}.$$

Thus, the joint pdf of Y and W is

$$f_{Y,W}(y, w) = \frac{exp(-\frac{y}{2}(1 + \frac{n}{mw}))}{2^{\frac{m+n}{2}}\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} y^{\frac{m}{2}-1} (\frac{ny}{mw})^{\frac{n}{2}-1} \frac{ny}{mw^2}, \text{ for } 0 \leq y, w < \infty$$

$$= \frac{exp(-\frac{y}{2}(1 + \frac{n}{mw}))}{2^{\frac{m+n}{2}}\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} y^{\frac{m+n}{2}-1} (\frac{n}{m})^{\frac{n}{2}} \frac{1}{w^{\frac{n}{2}+1}}.$$

The marginal distribution of W is

$$f_W(w) = \int_0^\infty f_{Y,W}(y, w) dy$$

$$= (\frac{n}{m})^{\frac{n}{2}} \frac{1}{2^{\frac{m+n}{2}}\Gamma(\frac{m}{2})\Gamma(\frac{n}{2}) w^{\frac{n}{2}+1}} \int_0^\infty y^{\frac{m+n}{2}-1} exp(-\frac{y}{2}(1 + \frac{n}{mw})) dy$$

$$= \frac{\Gamma\big(\frac{m+n}{2}\big)}{\Gamma\big(\frac{m}{2}\big)\Gamma\big(\frac{n}{2}\big)} \frac{\big(\frac{mw}{n}\big)^{\frac{m}{2}-1}\big(\frac{m}{n}\big)}{\big(\frac{mw}{n} + 1\big)^{\frac{m+n}{2}}}, \text{ for } 0 \leq w < \infty.$$

W follows the F-distribution with parameters (m, n).

□

4. *Following is the data set of daily minimum temperature at a hill station recorded* $^\circ F$ *during the month of April.*

$$77, 80, 82, 68, 65, 59, 61, 57, 50, 62, 61, 70, 69, 64, 67$$
$$70, 62, 65, 65, 73, 76, 87, 80, 82, 83, 79, 79, 71, 80, 77.$$

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

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

(c) *Find the* $100p$ *percentile for* $p = 0.25, 0.50$ *and* 0.75.

(d) *Find the first quartile* Q_1, *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 mean* $\bar{X}_T$.

(g) *Explain 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

$$50, 57, 59, 61, 61, 62, 62, 64, 65, 65, 65, 67, 68, 69, 70,$$
$$70, 71, 73, 76, 77, 77, 79, 79, 80, 80, 80, 82, 82, 83, 87.$$

(a) Following is the steam and leaf plot for the daily minimum temperatures.

Table 1: Stem and leaf plot, Stem: tens digits (° F), Leaf: ones digits (°F).

5	0	7	9								
6	1	1	2	2	4	5	5	5	7	8	9
7	0	0	1	3	6	7	7	9	9		
8	0	0	0	2	2	3	7				

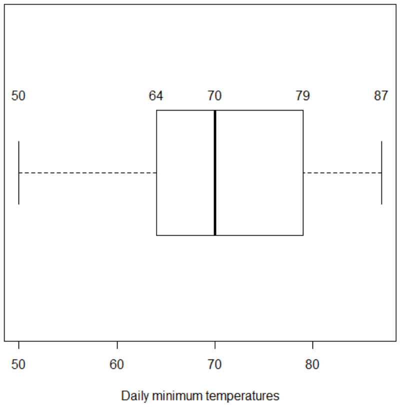
(b) Let $x_1, \cdots, x_n$ denote the sample values, $n = 30$. Then, the sample mean is

$$\frac{1}{30}\sum_{i=1}^{30} x_i = 70.7.$$

(c) The index of 25th percentile is $\lceil 0.25 * 30 \rceil = 8$ in the sorted sample. 64 is the 8^{th} entry in the sorted sample. Hence, the 25^{th} percentile is 64. The 50^{th} percentile is the mean of the 15^{th} $(0.5 * 30 = 15)$ and the 16^{th} entry in the sorted sample *i.e.* the 50^{th} percentile is 70 $((70 + 70)/2)$. The index of 75th percentile is $\lceil 0.75 * 30 \rceil = 23$ in the sorted sample. 79 is the 23^{rd} entry in the sorted sample. Hence, the 75^{th} percentile is 79.

(d) As Q_1, median and Q_3 are same as the 25^{th}, 50^{th} and the 75^{th} percentile, $Q_1 = 64$, median is 70 and $Q_3 = 79$.

Figure 1: Box plot of the 30 daily minimum temperatures (°F)



(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 interquartile range is $IQR = Q_3 - Q_1 = 15$. 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, the upper and lower whiskers coincide with maximum and the minimum values of the data. None of the data points are outside the interval $(Q_1 - 1.5 * IQR, Q_3 + 1.5 * IQR)$. Hence, there are no outliers.

(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 are no outliers, we do not eliminate any observations.

(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 70 are 14 and the number of data points greater than 70 are also 14. 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 $Q_3 - Q_2 = 9$ is more than the distance between $Q_2 - Q_1 = 6$. Based on the data, the population from which it arrived from can be taken as positively skewed.

□