MTL 390 (Statistical Methods)
Minor Examination Assignment 1 Report

Name: Arpit Saxena Entry Number: 2018MT10742

1. Descriptive Statistics

   Consider data taken from a bivariate distribution which is listed in the following table:

x_1	x_2
4.597941	-0.082831
3.199353	1.097550
1.680648	0.410578
-1.885425	0.513452
-0.207406	-1.094364
3.604727	-0.528722
2.043016	0.232730
4.392772	-0.806960
0.923505	-3.016915
1.334248	0.789666
4.421921	-0.763320
2.965662	1.642863
2.223469	-2.652278
1.801136	1.881806
-2.200512	0.049598

   Find the biased and unbiased sample covariance matrices for the given data. Using that, or otherwise, find the kurtosis and excess kurtosis of the given sample. Do the same thing treating each variable as an individual sample from a distribution.

   Answer

   Let $\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}$ be n samples from the data, then the biased sample covariance matrix is given by:

$$S_1 = \frac{1}{n} \sum_{i=1}^{n} \left[(\boldsymbol{x_i} - \overline{\boldsymbol{x}})(\boldsymbol{x_i} - \overline{\boldsymbol{x}})^T \right]$$

   Now, substituting values from the given data and calculating, we get

$$S_1 = \begin{bmatrix} 4.174 & -0.018 \\ -0.018 & 1.803 \end{bmatrix}$$

   Now, the unbiased matrix is given by

$$S_2 = \frac{n}{n-1} S_1$$

$$= \frac{15}{14} S_1$$
$$= \begin{bmatrix} 4.472 & -0.019 \\ -0.019 & 1.932 \end{bmatrix}$$

The sample kurtosis is then given by:

$$\text{Kurtosis} = \frac{1}{n} \sum_{i=1}^{n} \left[(\boldsymbol{x_i} - \overline{\boldsymbol{x}})^T S_1^{-1} (\boldsymbol{x_i} - \overline{\boldsymbol{x}}) \right]^2 \qquad \text{(From [1])}$$

We first calculate the sample mean from the data as:

$$\overline{\boldsymbol{x}} = (1.926, -0.155)^T$$

Then, we calculate the kurtosis using the given formula and get

$$\text{Kurtosis} = 5.93$$

For a multivariate normal distribution distribution of p variables, the expected kurtosis is (from [1]) $p(p+2)$. Thus, the expected kurtosis for bivariate normal distribution is $2 * 4 = 8$. Therefore, excess kurtosis equals $5.93 - 8 = -2.07$ and our given sample is mesokurtic.

Treating the columns individually, we get

	Kurtosis	Excess Kurtosis
x_1	2.865	-0.135
x_2	3.302	0.302

2. Descriptive Statistics

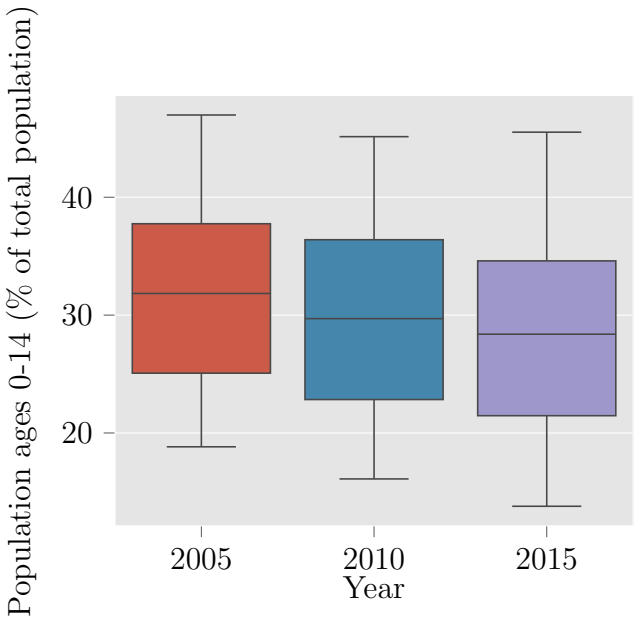
Given data is from world bank and contains values for the indicator "Population ages 0-14 (% of population)" for some 15 randomly selected countries and for 3 years, namely: 2005, 2010 and 2015. Visualize this data and draw observations.

Country Code	2005	2010	2015
KOR	18.827339	16.101602	13.781719
MEX	31.837208	29.507688	27.635969
BOL	36.326591	34.489515	32.404954
BDI	46.977947	45.140314	45.520683
VCT	28.554330	25.833449	23.719170
OSS	34.171454	31.320387	30.379485
SLB	41.314868	40.751637	40.430065
CUB	19.201050	17.741081	16.674866
HPC	45.152638	44.694382	43.648154
PNG	39.179415	38.302255	36.796778
VEN	31.689552	29.907486	28.380399
FJI	30.534296	29.000440	29.847679
NOR	19.612168	18.811563	17.962317
LTE	21.593859	19.838022	19.213434
SAU	33.873896	29.703454	25.822845

Answer

The given data is basically showing changing of distributions over time. We can visualize one distribution by plotting a box plot and then we could plot multiple box plots over the years to see how this data varies.

(The following plot was exported from python to tikz)



From the given box plots, we can observe that the mean percentage population aged 0-14 has been creasing over the years. While the minimum value in this sample has been decreasing, the maximum value is roughly the same. Thus there is an increased variance in the value of this indicator. While there are some countries in which the percentage of children has come down (could indicate aging population) while in other countries of the world it has either increased or stayed the same.

3. Sampling Distributions

At an organization, there are 10 systems and each one has a different processing power. Suppose there are 10 jobs whose loads (as defined by some metric) are randomly picked from the interval $(0, 100)$. One job is assigned to one system, and the assignment is done by arranging the systems and jobs in an increasing order (of processing power and load respectively) and assign each job to the system at the same index in the other list.

Find the probability distribution of the load at each system. Hence or otherwise, find the expected load of the job ending up on the third system (arranged in increasing order).

Answer

Let $X_1, \ldots, X_{10}$ be samples from the distribution $U(0, 1)$. Then $100X_i \sim U(0, 100)\, \forall i = 1, \ldots, 10$ would represent the job loads. The ordering amongst X_i's and $100X_i$'s is the same, so we work with X_i's to make our life easier.

Suppose X_i's arranged in ascending order are $X_{(1)}, X_{(2)}, \ldots, X_{(10)}$.

Let $k \in \{1, \ldots, 10\}$ and $n = 10$. Also, let $\epsilon > 0$ be small. Then:

$$\begin{aligned} P(X_{(k)} \in [x, x+\epsilon]) = P(&\text{one of } X_1, \ldots, X_{10} \text{ lies in } [x, x+\epsilon] \\ &\text{and exactly } k-1 \text{ are less than } x) \end{aligned}$$

(Now using the fact that X_i's are independent)

$$\begin{aligned} &= \binom{n}{1} P(\text{one lies in } [x, x+\epsilon]) P(\text{exactly } k-1 \text{ are less than } x) \\ &= n\epsilon \binom{n-1}{k-1} P(X_1 \leq x)^{k-1} P(X_1 > x)^{n-k} \qquad \text{(Since pdf is } f(x) = 1) \\ &= n\epsilon \frac{(n-1)!}{(k-1)!\,(n-k)!} x^{k-1}(1-x)^{n-k} \\ \implies f_{X_{(k)}}(x) &= \frac{x^{k-1}(1-x)^{n-k}}{\frac{(k-1)!\,(n-k)!}{n!}} \\ &= \frac{x^{k-1}(1-x)^{n-k}}{\frac{\Gamma(k)\Gamma(n-k+1)}{\Gamma(n+1)}} \\ &= \frac{x^{k-1}(1-x)^{n-k}}{B(k, n-k+1)} \qquad \text{(where } B \text{ is the Beta function)} \end{aligned}$$

We note that

$$X \sim B(\alpha, \beta) \implies f_X(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$$

where $B(\cdot, \cdot)$ is the Beta-distribution.

Comparing the equations of Beta distribution's pdf and the pdf we obtained for $X_{(k)}$, we conclude that

$$X_{(k)} \sim B(k, n-k+1) \, \forall k \in \{1, \ldots, 10\}$$

Now, we find the expected value of $X_{(3)}$ as:

$$\begin{aligned}
E(X_{(k)}) &= \int_0^1 x \times \frac{x^{k-1}(1-x)^{n-k}}{B(k, n-k+1)} dx \\
\implies E(X_{(3)}) &= \int_0^1 x \times \frac{x^2(1-x)^8}{B(3,8)} dx && (\because n = 10) \\
&= \frac{1}{B(3,8)} \int_0^1 x^3(1-x)^8 dx \\
&= \frac{1}{B(3,8)} \int_0^1 (1-x)^3 x^8 dx \\
&= \frac{1}{B(3,8)} \int_0^1 x^8(1-3x+3x^2-x^3) dx \\
&= \frac{\Gamma(11)}{\Gamma(3)\Gamma(8)} \int_0^1 (x^8 - 3x^9 + 3x^{10} - x^{11}) dx \\
&= \frac{10!}{2!\,7!} \left[\frac{1}{9} - \frac{3}{10} + \frac{3}{11} - \frac{1}{12} \right] \\
&= 360 \left[\frac{1}{36} - \frac{3}{110} \right] \\
&= 0.182
\end{aligned}$$

Therefore, the expected load of the job ending up at the 3rd system is $100E(X_{(k)}) = 18.2$

4. Sampling Distributions

Consider that any insurance claim can be modelled as coming from an exponential distribution. Two companies had 30 insurance claims over some period of time. Suppose that claims for the companies come from exponential distributions with means 10 and 12 (in lakhs) respectively. Then what is the probability that total insurance claims made to company 1 are greater than claims made to company 2? Solve this using two different approaches and compare answers.

Answer

Insurance claims to companies 1 and 2 are distributed as $\text{Exp}(\frac{1}{10})$ and $\text{Exp}(\frac{1}{12})$ respectively.

Then one approach could be that sum of all insurance claims would be distributed as Gamma$(30, \frac{1}{10})$ and Gamma$(30, \frac{1}{12})$ for company 1 and company 2 respectively. Let $X \sim$ Gamma(n, λ_1), $Y \sim$ Gamma(n, λ_2) where $n = 30, \lambda_1 = \frac{1}{10}, \lambda_2 = \frac{1}{12}$. Then,

$$\begin{aligned}
P(X > Y) &= \int_{y=0}^{\infty} \int_{x=y}^{\infty} f_X(x) f_Y(y)\, dx\, dy \\
&= \int_0^{\infty} f_Y(y) \int_y^{\infty} f_X(x)\, dx\, dy \\
&= \int_0^{\infty} \frac{\lambda_2^n y^{n-1} e^{-\lambda_2 y}}{\Gamma(n)} \frac{\Gamma(n, \lambda_1 y)}{\Gamma(n)} dy \\
&\quad \text{(where } \Gamma(\cdot, \cdot) \text{ is the upper incomplete Gamma function)} \\
&= \int_0^{\infty} \frac{\left(\frac{1}{12}\right)^{30} y^{29} e^{-\frac{y}{12}}}{29!} \frac{\Gamma(30, \frac{y}{10})}{29!} dy
\end{aligned}$$

Now, solving using mathematical computational tools, we get

$$P(X > Y) = 0.2411 \tag{1}$$

The other approach is that since the exponentially distributed random variables are iid and there are 30 of them, we can use CLT to approximate their sum. Claims for first company have mean 10 and variance $10^2 = 100$, for the second company they have mean 12 and variance $12^2 = 144$. Let $\overline{X}, \overline{Y}$ be random variables denoting sample average claims made to companies 1 and 2 respectively.

Then, by CLT we have:

$$\begin{aligned}
\frac{\overline{X} - 10}{\frac{10}{\sqrt{30}}} &\sim \mathcal{N}(0, 1) \\
\frac{\overline{Y} - 12}{\frac{12}{\sqrt{30}}} &\sim \mathcal{N}(0, 1)
\end{aligned}$$

Let $Z_1 = \frac{\overline{X} - 10}{\frac{10}{\sqrt{30}}}$ and $Z_2 = \frac{\overline{Y} - 12}{\frac{12}{\sqrt{30}}}$. Then Z_1, Z_2 are standard normal random variables and are also independent since they are made of samples which are independent. Now,

$$\begin{aligned}
& \overline{X} > \overline{Y} \\
\implies & \frac{\overline{X} - 10}{\frac{10}{\sqrt{30}}} \frac{10}{\sqrt{30}} + 10 > \frac{\overline{Y} - 12}{\frac{12}{\sqrt{30}}} \frac{12}{\sqrt{30}} + 12 \\
\implies & \frac{10}{\sqrt{30}} Z_1 + 10 > \frac{12}{\sqrt{30}} Z_2 + 12
\end{aligned}$$

$$\implies \frac{10Z_1 - 12Z_2}{\sqrt{30}} > 2$$

$$\implies Z_1 > \frac{2\sqrt{30} + 12Z_2}{10}$$

$$\begin{aligned}
\therefore P(\overline{X} > \overline{Y}) &= P\left(Z_1 > \frac{2\sqrt{30} + 12Z_2}{10}\right) \\
&= \int_{z_2=-\infty}^{\infty} \int_{z_1=\frac{2\sqrt{30}+12z_2}{10}}^{\infty} f_{Z_2}(z_2) f_{Z_1}(z_1) dz_1 dz_2 \\
&= \int_{-\infty}^{\infty} f_{Z_2}(z_2) \int_{z_1=\frac{2\sqrt{30}+12z_2}{10}}^{\infty} f_{Z_1}(z_1) dz_1 dz_2 \\
&= \int_{-\infty}^{\infty} f_{Z_2}(z_2) \left\{1 - \Phi\left(\frac{2\sqrt{30} + 12z_2}{10}\right)\right\} dz_2 \\
&\qquad \text{(where } \Phi \text{ is the standard normal CDF)} \\
&= \int_{-\infty}^{\infty} f_{Z_2}(z_2) \left\{\frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{2\sqrt{30} + 12z_2}{10\sqrt{2}}\right)\right\} dz_2 \\
&\qquad \text{(Where erf is the Gauss error function)} \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z_2^2}{2}} \left\{\frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{2\sqrt{30} + 12z_2}{10\sqrt{2}}\right)\right\} dz_2
\end{aligned}$$

The conversion from Φ to erf helps us in using computational tools, which give the result as:

$$P(\overline{X} > \overline{Y}) = 0.2416 \tag{2}$$

Now comparing (1) and (2), we observe the probabilities computes as 0.2411 and 0.2416 and these probabilities are very close.

5. Point and Interval Estimations

For a distribution with k unknown parameters, method of moments uses k moments to form a system of equations and solves it to find estimates for the parameters. This throws away information contained in higher order moments. To remedy that, the **Generalized Method of Moments (GMM)** takes $q(> k)$ moments and minimizes the sum of squares of difference between sample moments and moments calculated from the distribution.

Consider the following samples taken from a Poisson distribution with unknown λ. Find an estimate for the parameter using both method of moments as well as generalized method of moments (with 3 moments)

30	21	24	18
28	25	24	25
26	19	19	21
22	34	22	15
22	25	16	22

Answer

We first calculate expressions of three moments $E[X], E[X^2]$ and $E[X^3]$ for $X \sim P(\lambda)$. Using the MGF of the Poisson distribution, we find moments around 0:

$$\begin{aligned}
& M_X(t) = exp(\lambda(e^t - 1)) \\
\implies & M'_X(t) = \lambda e^t exp(\lambda e^t - \lambda) \\
\implies & M''_X(t) = (\lambda e^t)^2 exp(\lambda e^t - \lambda) + \lambda e^t exp(\lambda e^t - \lambda) \\
\implies & M'''_X(t) = (\lambda e^t)^3 exp(\lambda e^t - \lambda) + 2(\lambda e^t)^2 exp(\lambda e^t - \lambda) \\
& \qquad + (\lambda e^t)^2 exp(\lambda e^t - \lambda) + \lambda e^t exp(\lambda e^t - \lambda) \\
& \qquad = (\lambda e^t)^3 exp(\lambda e^t - \lambda) + 3(\lambda e^t)^2 exp(\lambda e^t - \lambda) + \lambda e^t exp(\lambda e^t - \lambda)
\end{aligned}$$

Using these, we calculate the moments as:

$$\begin{aligned}
E[X] &= M'_X(0) = \lambda \\
E[X^2] &= M''_X(0) = \lambda^2 + \lambda \\
E[X^3] &= M'''_X(0) = \lambda^3 + 3\lambda^2 + \lambda
\end{aligned}$$

Next we calculate the sample moments. Let the samples be written as $x_1, ..., x_{20}$. Then:

$$\begin{aligned}
m_1 &= \sum_{i=1}^{20} x_i = 22.9 \\
m_2 &= \sum_{i=1}^{20} {x_i}^2 = 544.4 \\
m_3 &= \sum_{i=1}^{20} {x_i}^3 = 13424.5
\end{aligned}$$

Using the method of moments, we get:

$$\widehat{\lambda}_1 = m_1 = 22.9$$

For the generalized method of moments, we note that we can take any weighting of the sample moments. In fact we can also take a positive definite matrix and define the cost function that way. Suppose we somehow decided to keep the weights for m_1, m_2, m_3 to be $100, 10, 1$ respectively. Then, we have the function:

$$\begin{aligned} Q(\lambda) &= 100(m_1 - E[X])^2 + 10(m_2 - E[X^2])^2 + 1(m_3 - E[X^3])^2 \\ &= 100(22.9 - \lambda)^2 + 10(544.4 - \lambda^2 - \lambda)^2 + (13424.5 - \lambda^3 - 3\lambda^2 - \lambda)^2 \end{aligned}$$

Then the estimator is given by:

$$\widehat{\lambda}_2 = \operatorname*{argmin}_{\lambda} Q(\lambda)$$

We note that $Q(\lambda)$ is a polynomial in lambda with degree 6, so it's not practical to calculate the minimum by hand. Using computer tools, we find:

$$\widehat{\lambda}_2 = 22.79$$

6. Point and Interval Estimations

   Find an unbiased estimator for the parameter λ of an exponential distribution $\text{Exp}(\lambda)$ as a multiple of the sample mean. Given that the estimator found is the UMVUE, show that it does not achieve equality in the Cramer-Rao inequality, consequently showing that achieving the Cramer Rao Lower Bound is not a necessary condition for being UMVUE.

   Customers arrive at a checkout counter with an average time of 10 minutes as observed from 30 customers. What is the time in which their order should be processed so that 70% of the customers don't find a line at the checkout counter.

   Answer

   Let $X_1, \ldots, X_n \sim \text{Exp}(\lambda)$ be n samples from an exponential distribution with parameter λ. Now, let

$$\begin{aligned} T(X) &= \frac{c}{\overline{X}} \\ &= \frac{cn}{\sum_{i=1}^{n} X_i} \end{aligned}$$

   Using the fact that sum of exponentially distributed variables follows the Gamma distribution, we define $Z = \sum_{i=1}^{n} X_i$ and note that $Z \sim \text{Gamma}(n, \lambda)$. Then $T(X) = \frac{cn}{Z}$

$$E(T(X)) = E\left(\frac{cn}{Z}\right)$$

$$
\begin{aligned}
&= cnE\left(\frac{1}{Z}\right) \\
&= cn\int_0^\infty \frac{1}{z}\frac{\lambda^n z^{n-1}e^{-\lambda z}}{\Gamma(n)}dz && (\because Z \sim Gamma(n,\lambda)) \\
&= \frac{cn\lambda}{n-1}\underbrace{\int_0^\infty \frac{\lambda^{n-1}z^{n-2}e^{-\lambda z}}{\Gamma(n-1)}dz}_{=1} && (\because \Gamma(n) = (n-1)\Gamma(n-1)) \\
&= \frac{cn\lambda}{n-1}
\end{aligned}
$$

To make this estimator unbiased, we have

$$
\begin{aligned}
E(T(X)) &= \lambda \\
\implies \frac{cn\lambda}{n-1} &= \lambda \\
\implies c &= \frac{n-1}{n} \\
\therefore T(X) &= \frac{n-1}{\sum_{i=1}^n X_i} && \text{Since } T(X) = \tfrac{cn}{\sum_{i=1}^n X_i}
\end{aligned}
$$

Thus, we have found an unbiased estimator for T(X).
Now, for Cramer-Rao inequality:

$$
\begin{aligned}
E\left(T(X)^2\right) &= E\left(\left(\frac{n-1}{\sum_{i=1}^n X_i}\right)^2\right) \\
&= (n-1)^2 E\left(\frac{1}{Z^2}\right) \\
&= (n-1)^2\int_0^\infty \frac{1}{z^2}\frac{\lambda^n z^{n-1}e^{-\lambda z}}{\Gamma(n)}dz && (\because Z \sim \text{Gamma}(n,\lambda)) \\
&= \frac{(n-1)^2\lambda^2}{(n-1)(n-2)}\underbrace{\int_0^\infty \frac{\lambda^{n-2}z^{n-3}e^{-\lambda z}}{\Gamma(n-2)}dz}_{=1} && (\because \Gamma(n) = (n-1)\Gamma(n-1)) \\
&= \frac{(n-1)\lambda^2}{n-2}
\end{aligned}
$$

Using this, we find the variance:

$$
\begin{aligned}
Var(T(X)) &= E(T(X)^2) - (E(T(X)))^2 \\
&= \frac{(n-1)\lambda^2}{n-2} - \lambda^2
\end{aligned}
$$

$$= \lambda^2 \left(\frac{n-1}{n-2} - 1 \right)$$

$$\implies Var(T(X)) = \frac{\lambda^2}{n-2} \tag{3}$$

Now, we find the information for X_1:

$$\begin{aligned}
I_1(\lambda) &= E\left[\left(\frac{\partial}{\partial \lambda} \log f(X_1; \lambda) \right)^2 \right] \\
&= E\left[\left(\frac{\partial}{\partial \lambda} \log \left\{ \lambda e^{-\lambda X} \right\} \right)^2 \right] \\
&= E\left[\left(\frac{\partial}{\partial \lambda} \left\{ \log \lambda - \lambda X \right\} \right)^2 \right] \\
&= E\left[\left(\frac{1}{\lambda} - X \right)^2 \right] \\
&= Var(X) && (\because E(X) = \tfrac{1}{\lambda}) \\
&= \frac{1}{\lambda^2}
\end{aligned}$$

From that we get total information of the sample as:

$$I(\lambda) = \frac{n}{\lambda^2} \tag{4}$$

$$\text{LHS of Cramer Rao inequality} = Var(T(X)) = \frac{\lambda^2}{n-2} \qquad \text{From (3)}$$

$$\text{RHS of Cramer Rao inequality} = \frac{1}{I(\lambda)} = \frac{\lambda^2}{n} \qquad \text{Since } E(T(X)) = \lambda \text{ and from (4)}$$

Thus, LHS of Cramer Rao inequality $\neq$ RHS of Cramer Rao inequality

Now, assuming that the inter-arrival time of customers at the checkout counter is independent, we can model the inter-arrival time as an exponential distribution.

We find an estimate for the parameter λ using the estimator derived previously as:

$$\widehat{\lambda} = \frac{n-1}{\sum_{i=1}^{n} X_i} = \frac{n-1}{n\overline{X}} = \frac{29}{30} \times 6 \text{ hour}^{-1} = 5.8 \text{ hour}^{-1}$$

Let t be the time to process one customer's order at the counter, then we want $P(X > t) >= 0.7$ where $X \sim Exp(\widehat{\lambda}) = Exp(5.8)$.

$$
\begin{aligned}
& P(X > t) \geq 0.7 \\
\implies & e^{-\widehat{\lambda} t} \geq 0.7 \\
\implies & \widehat{\lambda} t \leq -\log 0.7 \\
\implies & t \leq -\frac{\log 0.7}{\widehat{\lambda}} \\
\implies & t \leq 0.062 \text{ hours} = 3.72 \text{ minutes}
\end{aligned}
$$

Thus, each order should be processed in not more than 3.72 minutes for 70% of the customers to not have to wait in line at the counter.

References

[1] Cain, M.K., Zhang, Z. & Yuan, KH. Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. Behav Res 49, 1716–1735 (2017). https://doi.org/10.3758/s13428-016-0814-1