NANYANG TECHNOLOGICAL UNIVERSITY

SPMS/DIVISION OF MATHEMATICAL SCIENCES

2016/17 Semester 1 MH2500 Probability and Introduction to Statistics Tutorial 10

For the tutorial on 27 October, let us discuss

- Ex. 4.7.42, 45, 49, 50, 54, 60.

Ex. 4.7.42. Let X be an exponential random variable with standard deviation σ. Find $P(|X - E(X)| > k\sigma)$ for $k = 2, 3, 4$, and compare the results to the bounds from Chebyshev's inequality.

[Solution:] First, we compute the mean and standard deviation of X. Integrating by parts,

$$\begin{aligned}
E(X) &= \int_0^\infty t\lambda e^{-\lambda t} dt \\
&= \left[-te^{-\lambda t}\right]_0^\infty + \int_0^\infty e^{-\lambda t} dt \\
&= \frac{1}{\lambda}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
E(X^2) &= \int_0^\infty t^2\lambda e^{-\lambda t} dt \\
&= \left[-t^2 e^{-\lambda t}\right]_0^\infty + \int_0^\infty 2te^{-\lambda t} dt \\
&= 0 + \frac{2}{\lambda} E(X) \\
&= \frac{2}{\lambda^2}.
\end{aligned}$$

Hence the mean is $\frac{1}{\lambda}$ and the variance is $\frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$. Hence $\sigma = 1/\lambda$.

Next, we recall that the cdf of X is $F_X(t) = 1 - e^{-\lambda t}$ for $t \geq 0$. Hence

$$\begin{aligned}
P(|X - E(X)| > k\sigma) &= P\left(X > \frac{k+1}{\lambda}\right) + P\left(0 < X < \frac{1-k}{\lambda}\right) \\
&= P\left(X > \frac{k+1}{\lambda}\right) + 0 \\
&= 1 - (1 - e^{-(k+1)}) = e^{-k+1}.
\end{aligned}$$

Hence for $k = 2, 3, 4$, the probabilities are $e^{-3} \approx 0.0498$, $e^{-4} \approx 0.0183$, $e^{-5} \approx 0.00673$, respectively.

By Chebyshev's inequality, for $k = 2, 3, 4$, respectively,

$$P(|X - E(X)| > k\sigma) \leq \frac{1}{k^2} = \frac{1}{4}, \frac{1}{9}, \frac{1}{16} \approx 0.25, 0.111, 0.0625.$$

Ex. 4.7.45. Find the covariance and correlation of N_i and N_j, where $N_1, N_2, \ldots, N_r$ are multinomial random variables. (Hint: Express them as sums.)

[Solution:] Suppose each N_i has a probability of success of p_i. Then each N_i is binomial with n and p_i. The mean of N_i is np_i and the variance is $np_i(1-p_i)$. Next, we evaluate $E(N_iN_j)$ by considering it as trinomial sum with a success in N_i, b success in N_j and $(n-a-b)$ failures. The joint probability mass function is given by

$$\frac{n!}{a!b!(n-a-b)!}p_i^a p_j^b (1-p_i-p_j)^{n-a-b}$$

and so the expected value is

$$E(N_iN_j) = \sum_{\substack{a,b\in\mathbb{Z}\\0\le a+b\le n}} ab\frac{n!}{a!b!(n-a-b)!}p_i^a p_j^b (1-p_i-p_j)^{n-a-b}$$

(noting that when $a=0$ or $b=0$, the summand is zero,)

$$= \sum_{\substack{a,b\in\mathbb{Z}\\2\le a+b\le n}} \frac{n!}{(a-1)!(b-1)!(n-a-b)!}p_i^a p_j^b (1-p_i-p_j)^{n-a-b}.$$

Replacing a by $c+1$ and b by $d+1$, the sum is

$$\begin{aligned}E(N_iN_j) &= \sum_{\substack{c,d\in\mathbb{Z}\\0\le c+d\le n-2}} \frac{n!}{c!d!(n-c-d-2)!}p_i^{c+1} p_j^{d+1} (1-p_i-p_j)^{n-c-1-d-1}\\ &= n(n-1)p_ip_j \sum_{\substack{c,d\in\mathbb{Z}\\0\le c+d\le n-2}} \frac{n-2!}{c!d!(n-2-c-d)!}p_i^{c} p_j^{d} (1-p_i-p_j)^{n-2-c-d}\\ &= n(n-1)p_ip_j,\end{aligned}$$

since the sum represents the sum of all probabilities of a trinomial distribution.

Therefore, the covariance is

$$\begin{aligned}\mathrm{Cov}(N_i,N_j) &= E(N_iN_j) - E(N_i)E(N_j)\\ &= n(n-1)p_ip_j - n^2p_ip_j\\ &= -np_ip_j,\end{aligned}$$

and the correlation is

$$\begin{aligned}\rho &= \frac{\mathrm{Cov}(N_i,N_j)}{\sqrt{\mathrm{Var}(N_i)}\sqrt{\mathrm{Var}(N_j)}}\\ &= \frac{-np_ip_j}{\sqrt{np_i(1-p_i)}\sqrt{np_j(1-p_j)}}\\ &= -\sqrt{\frac{p_ip_j}{(1-p_i)(1-p_j)}}.\end{aligned}$$

Ex. 4.7.49. Two independent measurements, X and Y, are taken of a quantity μ. Suppose $E(X) = E(Y) = \mu$ but σ_X and σ_Y are unequal. The two measurements are combined by means of a weighted average to give

$$Z = \alpha X + (1-\alpha)Y$$

where α is a scalar and $0 \le \alpha \le 1$.

a. Show that $E(Z) = \mu$.
b. Find α in terms of σ_X and σ_Y to minimize $\mathrm{Var}(Z)$.
c. Under what circumstances is it better to use the average $(X+Y)/2$ than either X or Y alone?

[Solution:]
a. $E(Z) = E(\alpha X + (1-\alpha)Y) = \alpha E(X) + (1-\alpha)E(Y) = \alpha\mu + (1-\alpha)\mu = \mu$.
b.

$$\begin{aligned}\mathrm{Var}(Z) &= \mathrm{Var}(\alpha X + (1-\alpha)Y)\\ &= \alpha^2\mathrm{Var}(X) + (1-\alpha)^2\mathrm{Var}(Y)\\ &= \alpha^2\sigma_X^2 + (1-\alpha)^2\sigma_Y^2\\ &= (\sigma_X^2+\sigma_Y^2)\alpha^2 - 2\sigma_Y^2\alpha + \sigma_Y^2.\end{aligned}$$

Completing the square, we find that

$$\begin{aligned}\mathrm{Var}(Z) &= (\sigma_X^2+\sigma_Y^2)\left(\alpha - \frac{\sigma_Y^2}{\sigma_X^2+\sigma_Y^2}\right)^2 - \frac{\sigma_Y^4}{\sigma_X^2+\sigma_Y^2} + \sigma_Y^2\\ &= (\sigma_X^2+\sigma_Y^2)\left(\alpha - \frac{\sigma_Y^2}{\sigma_X^2+\sigma_Y^2}\right)^2 + \frac{\sigma_X^2\sigma_Y^2}{\sigma_X^2+\sigma_Y^2}.\end{aligned}$$

Hence $\mathrm{Var}(Z)$ has minimum value $\sigma_X^2\sigma_Y^2/(\sigma_X^2+\sigma_Y^2)$ when $\alpha = \sigma_Y^2/(\sigma_X^2+\sigma_Y^2)$.

c. Using $(X+Y)/2$ means using $\alpha = 1/2$ while using X or Y alone means using $\alpha = 1$ or 0. It is better to use $\alpha = 1/2$ if (using the equation $\mathrm{Var}(Z) = \alpha^2\sigma_X^2 + (1-\alpha)^2\sigma_Y^2$),

$$\frac{1}{4}\sigma_X^2 + \frac{1}{4}\sigma_Y^2 < \sigma_X^2 \qquad \text{and} \qquad \frac{1}{4}\sigma_X^2 + \frac{1}{4}\sigma_Y^2 < \sigma_Y^2$$

i.e.,

$$\sigma_Y^2 < 3\sigma_X^2 \qquad \text{and} \qquad \frac{1}{3}\sigma_X^2 < \sigma_Y^2,$$

which can be combined to give

$$\frac{1}{3}\sigma_Y^2 < \sigma_X^2 < 3\sigma_Y^2.$$

Hence $(X+Y)/2$ is better than X or Y alone if $\frac{1}{3} < \sigma_X^2/\sigma_Y^2 < 3$.

Ex. 4.7.50. Suppose that X_i, where $i = 1, \ldots, n$ are independent random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$. Let $\overline{X} = n^{-1}\sum_{i=1}^n X_i$. Show that $E(\overline{X}) = \mu$ and $Var(\overline{X}) = \sigma^2/n$.

[Solution:] Since $\overline{X}$ is a linear combination of X_i's,

$$E(\overline{X}) = E\left(\sum_{i=1}^n \frac{X_i}{n}\right) = \sum_{i=1}^n E\left(\frac{X_i}{n}\right) = \frac{1}{n}\sum_{i=1}^n E(X_i) = \frac{1}{n}\cdot n\cdot\mu = \mu.$$

Similarly,

$$\begin{aligned}
\operatorname{Var}(\overline{X}) &= \operatorname{Var}\left(\sum_{i=1}^{n} \frac{X_i}{n}\right) \;=\; \sum_{i=1}^{n} \operatorname{Var}\left(\frac{X_i}{n}\right) \\
&= \frac{1}{n^2} \sum_{i=1}^{n} \operatorname{Var}(X_i) \qquad \text{(since the } X_i\text{'s are independent)} \\
&= \frac{1}{n^2} \cdot n \cdot \sigma^2 \\
&= \frac{\sigma^2}{n}.
\end{aligned}$$

Ex. 4.7.54. Let X, Y, and Z be uncorrelated random variables with variances σ_X^2, σ_Y^2, and σ_Z^2, respectively. Let

$$\begin{aligned}
U &= Z + X \\
V &= Z + Y.
\end{aligned}$$

Find $\operatorname{Cov}(U, V)$ and ρ_{UV}.

[Solution:] We are given that X, Y, and Z are uncorrelated, i.e., $\operatorname{Cov}(X, Y) = \operatorname{Cov}(X, Z) = \operatorname{Cov}(Y, Z) = 0$. By theorem from lecture (Handout 7 slide 15/16),

$$\begin{aligned}
\operatorname{Cov}(U, V) &= \operatorname{Cov}(Z + X, Z + Y) \\
&= \operatorname{Cov}(Z, Z) + \operatorname{Cov}(Z, Y) + \operatorname{Cov}(X, Z) + \operatorname{Cov}(X, Y) \\
&= \operatorname{Var}(Z) + 0 + 0 + 0 \\
&= \sigma_Z^2.
\end{aligned}$$

Next,

$$\operatorname{Var}(U) \;=\; \operatorname{Var}(Z + X) = \operatorname{Var}(Z) + \operatorname{Var}(X) + 2\operatorname{Cov}(Z, X) \;=\; \sigma_Z^2 + \sigma_X^2.$$

Similarly, $\operatorname{Var}(V) = \sigma_Z^2 + \sigma_Y^2$. Therefore,

$$\rho_{UV} = \frac{\operatorname{Cov}(U, V)}{\sqrt{\operatorname{Var}(U)\operatorname{Var}(V)}} = \frac{\sigma_Z^2}{\sqrt{(\sigma_Z^2 + \sigma_X^2)(\sigma_Z^2 + \sigma_Y^2)}}.$$

Ex. 4.7.60. Let Y have a density that is symmetric about zero and let $X = SY$, where S is an independent random variable taking on the values $+1$ and -1 with probability $\frac{1}{2}$ each. Show that $\operatorname{Cov}(X, Y) = 0$, but that X and Y are not independent.

[Solution:] First, $E(S) = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot (-1) = 0$. Next,

$$\begin{aligned}
\operatorname{Cov}(X, Y) &= \operatorname{Cov}(SY, Y) = E(SY^2) - E(SY)E(Y) \\
&= E(S)E(Y^2) - E(S)E(Y)E(Y) \qquad \text{(since } S \text{ and } Y \text{ are independent)} \\
&= 0.
\end{aligned}$$

Since $X = SY$, we see that

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{2} f_Y(y), & x = y; \\ \frac{1}{2} f_Y(-y), & x = -y; \\ 0, & \text{otherwise.} \end{cases}$$

Since Y is a continuous random variable, we may assume that there are at least two different values y_1, and y_2, where $y_1 \neq \pm y_2$ and $f_Y(y_1), f_Y(y_2) > 0$. Then

$$f_{X,Y}(y_1, y_2) = 0 \neq f_X(y_1) f_Y(y_2).$$

This shows that X and Y are not independent.