EE 131A
Probability
Instructor: Lara Dolecek

Homework 8 Solutions
Due: March 14, 2023
TAs: Jayanth Shreekumar, Rushi Bhatt

Chapters 7.1 - 7.3 & 8 of *Probability, Statistics, and Random Processes* by A. Leon-Garcia

1. Consider the jointly Gaussian random variables X and Y that have the following joint PDF:

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_X^2}+\frac{y^2}{\sigma_Y^2}-\frac{2\rho xy}{\sigma_X\sigma_Y}\right)\right].$$

(a) Prove that Y is a Gaussian random variable by deriving its marginal PDF, $f_Y(y)$. Find the mean and variance of Y.

Solution:

The marginal PDF of Y, $f_Y(y)$ is derived as follows:

$$\begin{aligned} f_Y(y) &= \int_{x=-\infty}^{\infty} f_{X,Y}(x,y)dx \\ &= \int_{x=-\infty}^{\infty} \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_X^2}+\frac{y^2}{\sigma_Y^2}-\frac{2\rho xy}{\sigma_X\sigma_Y}\right)\right] dx. \end{aligned}$$

To perform this integral, we need to complete a square inside the argument of the exponential.

$$\begin{aligned} &-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_X^2}+\frac{y^2}{\sigma_Y^2}-\frac{2\rho xy}{\sigma_X\sigma_Y}\right) \\ &= -\frac{1}{2(1-\rho^2)}\left(\left[\frac{x}{\sigma_X}-\frac{\rho y}{\sigma_Y}\right]^2-\frac{\rho^2y^2}{\sigma_Y^2}+\frac{y^2}{\sigma_Y^2}\right) \\ &= -\frac{1}{2(1-\rho^2)}\left[\frac{x}{\sigma_X}-\frac{\rho y}{\sigma_Y}\right]^2-\frac{1}{2(1-\rho^2)}\frac{(1-\rho^2)y^2}{\sigma_Y^2} \\ &= -\frac{1}{2(1-\rho^2)\sigma_X^2}\left[x-\frac{\rho\sigma_X y}{\sigma_Y}\right]^2-\frac{y^2}{2\sigma_Y^2}. \end{aligned}$$

Substituting this exponential argument in the integral of $f_Y(y)$ gives us:

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma_Y}\exp\left[-\frac{y^2}{2\sigma_Y^2}\right]\int_{x=-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma_X\sqrt{1-\rho^2}}\exp\left[-\frac{\left[x-\frac{\rho\sigma_X y}{\sigma_Y}\right]^2}{2\sigma_X^2(1-\rho^2)}\right]dx$$

The value of this integral is 1 because it is the pdf of a Gaussian with mean $\frac{\rho\sigma_X y}{\sigma_Y}$ and variance $\sigma_X^2(1-\rho^2)$. Thus,

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma_Y}\exp\left[-\frac{y^2}{2\sigma_Y^2}\right], \quad -\infty < y < -\infty,$$

which proves that Y is a Gaussian random variable with mean 0 and variance σ_Y^2.

(b) Prove that $f_{X|Y}(x|y)$ corresponds to another Gaussian random variable by determining its closed form equation, then find its mean and variance.
Solution:
The conditional PDF $f_{X|Y}(x|y)$ is derived as follows:

$$\begin{aligned}
f_{X|Y}(x|y) &= f_{X,Y}(x,y)/f_Y(y) \\
&= \frac{1}{\sqrt{2\pi}\sigma_X\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_X^2}+\frac{y^2}{\sigma_Y^2}-\frac{2\rho xy}{\sigma_X\sigma_Y}\right)+\frac{y^2}{2\sigma_Y^2}\right].
\end{aligned}$$

One more time, we operate on the exponential argument:

$$\begin{aligned}
&-\frac{1}{2(1-\rho^2)}\left(\frac{x^2}{\sigma_X^2}+\frac{y^2}{\sigma_Y^2}-\frac{2\rho xy}{\sigma_X\sigma_Y}\right)+\frac{y^2}{2\sigma_Y^2} \\
&= -\frac{1}{2(1-\rho^2)}\left[\frac{x}{\sigma_X}-\frac{\rho y}{\sigma_Y}\right]^2+\frac{1}{2(1-\rho^2)}\left[\frac{\rho^2y^2}{\sigma_Y^2}-\frac{y^2}{\sigma_Y^2}\right]+\frac{y^2}{2\sigma_Y^2} \\
&= -\frac{1}{2(1-\rho^2)}\left[\frac{x}{\sigma_X}-\frac{\rho y}{\sigma_Y}\right]^2-\frac{y^2}{2\sigma_Y^2}+\frac{y^2}{2\sigma_Y^2} \\
&= -\frac{1}{2\sigma_X^2(1-\rho^2)}\left[x-\frac{\rho\sigma_X y}{\sigma_Y}\right]^2.
\end{aligned}$$

Consequently, we conclude that:

$$f_{X|Y}(x|y) = \frac{1}{\sqrt{2\pi}\sigma_X\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2\sigma_X^2(1-\rho^2)}\left[x-\frac{\rho\sigma_X y}{\sigma_Y}\right]^2\right],$$

where $-\infty < x < \infty$. This proves that $f_{X|Y}(x|y)$ corresponds to another Gaussian random variable with mean $\rho\sigma_X y/\sigma_Y$, and variance $\sigma_X^2(1-\rho^2)$.

2. Assume that $X_1, X_2, ..., X_n$ are independent random variables with possibly different distributions and let S_n be their sum. Let $m_k = E(X_k)$, $\sigma_k^2 = VAR(X_k)$, and $M_n = m_1 + m_2 + \cdots + m_n$. Assume that $\sigma_k^2 < R$ and $m_k < T$ for all k. Prove that, for any $\epsilon > 0$,

$$P(|\frac{S_n}{n} - \frac{M_n}{n}| < \epsilon) \to 1$$

as $n \to \infty$ using Chebyshev's inequality.
Solution:

$$\begin{aligned}
\frac{S_n}{n} &= \frac{X_1+X_2+\cdots+X_n}{n} \\
E[\frac{S_n}{n}] &= \frac{m_1+m_2+\cdots+m_n}{n} = \frac{M_n}{n} \\
VAR[\frac{S_n}{n}] &= \frac{VAR(X_1+X_2+\cdots+X_n)}{n^2} = \frac{\sigma_1^2+\sigma_2^2+\cdots+\sigma_n^2}{n^2} \le \frac{nR}{n^2} = \frac{R}{n},
\end{aligned}$$

using the Chebyshev's inequality we get:

$$
\begin{aligned}
P[|\frac{S_n}{n} - \frac{M_n}{n}| \geq \epsilon] &\leq \frac{Var(\frac{S_n}{n})}{\epsilon^2} \\
&\leq \frac{R}{n\epsilon^2} \to 0 \text{ as } n \to \infty
\end{aligned}
$$

Thus,

$$P(|\frac{S_n}{n} - \frac{M_n}{n}| < \epsilon) \to 1 \text{ as } n \to \infty.$$

3. *Application of CLT.*

   (a) A fair coin is tossed 100 times. Estimate the probability that the number of heads is between 40 and 60. Estimate the probability that the number is between 60 and 80.

   Solution:
   Let $n = 100$ and let $X_i = 1$ if i^{th} toss is a head, else 0. Then $S_n = X_1 + \ldots + X_n$. We have $\mu = nE[X_i] = np = 50$ and $\sigma^2 = nVAR[X_i] = np(1-p) = 25$. The central limit theorem gives:

   $$
   \begin{aligned}
   P[40 \leq S_n \leq 60] &= P[\frac{40-50}{\sqrt{25}} \leq \frac{S_n - \mu}{\sigma} \leq \frac{60-50}{\sqrt{25}}] \\
   &= Q(-2) - Q(2) = 0.9544
   \end{aligned}
   $$

   Similarly,

   $$
   \begin{aligned}
   P[60 \leq S_n \leq 80] &= P[\frac{60-50}{\sqrt{25}} \leq \frac{S_n - \mu}{\sigma} \leq \frac{80-50}{\sqrt{25}}] \\
   &= Q(2) - Q(6) \approx Q(2) = 0.0228
   \end{aligned}
   $$

   (b) Repeat part (a) for if we toss the coin 10000 times and for the intervals [4000,6000] and [6000,8000].

   Solution:
   We have $n = 10000$, $\mu = 5000$, $\sigma^2 = 2500$. The central limit theorem gives:

   $$
   \begin{aligned}
   P[4000 \leq S_n \leq 6000] &= P[\frac{4000-5000}{\sqrt{2500}} \leq \frac{S_n - \mu}{\sigma} \leq \frac{6000-5000}{\sqrt{2500}}] \\
   &= Q(-20) - Q(20) \approx 1
   \end{aligned}
   $$

   $$
   \begin{aligned}
   P[6000 \leq S_n \leq 8000] &= P[\frac{6000-5000}{\sqrt{2500}} \leq \frac{S_n - \mu}{\sigma} \leq \frac{8000-5000}{\sqrt{2500}}] \\
   &= Q(20) - Q(60) \approx 0
   \end{aligned}
   $$

4. The sum of a list of 108 real numbers is to be computed. Suppose that the numbers are rounded off to the nearest integer so that each number has an error that is uniformly distributed in the interval (-0.5,0.5). Use the central limit theorem to estimate the probability that the absolute value of the total error in the sum of the 108 numbers exceeds 2.

   Hint: Assign a random variable to each of the rounding errors and use the CLT on their sum.

   Solution:

   Let $X_1, X_2, \dots X_{108}$ be the 108 rounding errors. Let $S = \sum_{i=1}^{108} X_i$ be the total error. Since X_i is uniformly distributed in the interval (-0.5,0.5), the PDF of X_i is

$$f_{X_i}(x) = \begin{cases} 1 & -0.5 < x < 0.5 \\ 0 & \text{otherwise.} \end{cases}$$

   Hence,

$$E(X_i) = \int_{-0.5}^{0.5} x dx = 0.$$

$$\begin{aligned} VAR(X_i) &= E(X_i^2) - E(X_i)^2 \\ &= E(X_i^2) \\ &= \int_{-0.5}^{0.5} x^2 dx \\ &= \frac{1}{12}. \end{aligned}$$

   Thus,

$$E(S) = E(\sum_{i=1}^{108} X_i) = 0$$

   and

$$\begin{aligned} VAR(S) &= \sum_{i=1}^{108} VAR(X_i) \\ &= 108 \times \frac{1}{12} \\ &= 9 \end{aligned}$$

Now we apply the central limit theorem to S:

$$\begin{aligned}
P(|S| > 2) &= P(S > 2) + P(S < -2) \\
&= P(\frac{S - E(S)}{\sqrt{VAR(S)}} > \frac{2 - E(S)}{\sqrt{VAR(S)}}) + P(\frac{S - E(S)}{\sqrt{VAR(S)}} < \frac{-2 - E(S)}{\sqrt{VAR(S)}}) \\
&= P(\frac{S - E(S)}{\sqrt{VAR(S)}} > \frac{2}{3}) + P(\frac{S - E(S)}{\sqrt{VAR(S)}} < -\frac{2}{3}) \\
&= 2Q(\frac{2}{3}) \\
&\approx 0.5
\end{aligned}$$

5. Suppose $X_1, \cdots, X_n$ are i.i.d random variables with mean $E(X) = \mu$ and variance $VAR(X) = \sigma^2$. *Sample variance* is defined as

$$V_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - M_n)^2,$$

where M_n is the *sample mean*. Show that the expected value of V_n is given by:

$$\mathbb{E}[V_n] = \frac{n-1}{n}\sigma^2$$

Hint: Manipulate V_n into the form:

$$V_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (M_n - \mu)^2$$

Solution:

$$\begin{aligned}
V_n &= \frac{1}{n}\sum_{i=1}^{n}(X_i - M_n)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu + \mu - M_n)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\{(X_i - \mu)^2 + (\mu - M_n)^2 + 2(X_i - \mu)(\mu - M_n)\}
\end{aligned}$$

Thus, we reach that:

$$\begin{aligned}
V_n &= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 + \frac{1}{n}\sum_{i=1}^{n}(\mu-M_n)^2 + \frac{2}{n}\sum_{i=1}^{n}(X_i-\mu)(\mu-M_n) \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 + (\mu-M_n)^2 + \frac{2}{n}(\mu-M_n)\sum_{i=1}^{n}(X_i-\mu) \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 + (\mu-M_n)^2 - 2(\mu-M_n)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - (M_n-\mu)^2
\end{aligned}$$

$$\begin{aligned}
E(V_n) &= E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2 - (M_n-\mu)^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}E((X_i-\mu)^2) - E((M_n-\mu)^2) \\
&= VAR(X) - VAR(M_n) \\
&= \sigma^2 - \frac{\sigma^2}{n} \\
&= \frac{n-1}{n}\sigma^2.
\end{aligned}$$

The reason is that the variance will always be smaller when calculated using the sum of squared distances to the sample mean (called sample variance), compared to using the sum of squared distances to the population mean (called population variance). They will be equal only when the sample mean and the population mean coincide. This can also be explained in terms of degrees of freedom: we are losing one degree of freedom when performing the calculations shown above.

As a concrete example. consider that the population mean is 2050, but we only have a limited number of samples from this population which are: $2051, 2053, 2055, 2050, 2051$. We calculate the sample mean as $\frac{2051+2053+2055+2050+2051}{5} = 2052$. Now, if we use the sample mean to calculate the variance (sample variance) using the formula $\frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}[(X_i - \mathbb{E}[X])^2]$ we obtain 3.2. But if we use the population mean to calculate the variance (population variance), we obtain 7.2. Therefore, to obtain an unbiased estimate for the variance, we multiply the above expression with $\frac{n}{n-1}$. This is called Bessel's correction.