# qbs120_ps3_gibran

Gibran Erlangga

10/2/2021

## Question 1

(Based on Rice 4.54) Let X, Y, and Z be independent RVs with variances $\sigma_X^2, \sigma_Y^2, \sigma_Z^2$. Let:

$$U = Z - X$$

$$V = Z - Y$$

Answer the following questions:
(a) Find $Cov(U, V)$ and $\rho_{U,V}$.

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

Note that X, Y, and Z are independent. Hence its Covariance is 0.

$$\rho_{U,V} = \frac{Cov(U,V)}{\sqrt{Var(U)\ Var(V)}}$$

With $Var(U)$ and $Var(V)$ as:

$$\begin{aligned} Var(U) &= Var(Z-X) \\ &= Var(Z) - Var(X) + 2Cov(Z,X) \\ &= \sigma_Z^2 - \sigma_X^2 \end{aligned}$$

Similarly,

$$Var(V) = Var(Z-Y) = \sigma_Z^2 - \sigma_Y^2$$

Hence,

$$\begin{aligned}\rho_{U,V} &= \frac{Cov(U,V)}{\sqrt{Var(U)\ Var(V)}} \\ &= \frac{\sigma_Z^2}{\sqrt{(\sigma_Z^2-\sigma_X^2)(\sigma_Z^2-\sigma_Y^2)}}\end{aligned}$$

(b)If $U = Z + X$ and $V = Z + Y$, do the values of $Cov(U, V)$ and $\rho_{U,V}$ computed in part a) change? Explain.

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

With $Var(U)$ and $Var(V)$ as:

$$\begin{aligned}Var(U) &= Var(Z+X) \\ &= Var(Z)+Var(X)+2Cov(Z,X) \\ &= \sigma_Z^2+\sigma_X^2\end{aligned}$$

Similarly,

$$Var(V) = Var(Z+Y) = \sigma_Z^2+\sigma_Y^2$$

Hence,

$$\begin{aligned}\rho_{U,V} &= \frac{Cov(U,V)}{\sqrt{Var(U)\ Var(V)}} \\ &= \frac{\sigma_Z^2}{\sqrt{(\sigma_Z^2+\sigma_X^2)(\sigma_Z^2+\sigma_Y^2)}}\end{aligned}$$

Therefore, $Cov(U, V)$ stays the same while $\rho_{U,V}$ changes in its denominator.

(c) How does $\rho_{U,V}$ change if $\sigma_Z^2$ is much larger than $\sigma_X^2$ or $\sigma_Y^2$?
As $\sigma_Z^2$ exists in both numerator and denominator, when its value gets much larger, then the value of $\rho_{U,V}$ will be close to 1.

(d) How does $\rho_{U,V}$ change if $\sigma_Z^2$ is much smaller than $\sigma_X^2$ or $\sigma_Y^2$?
As $\sigma_Z^2$ exists in both numerator and denominator, when its value gets much smaller, then the value of $\rho_{U,V}$ will be close to 0.

(e) How do the answers for parts c) and d) relate to variable standardization?

Variable standardization is basically an effort to rescale your data to have a fixed range of 0 to 1, which is what the answers in c) and d) show when we increase or decrease the value of $\sigma_Z^2$ in the equation.

## Question 2

(Based on Rice 4.64) Let X and Y be jointly distributed RVs with correlation $\rho_{X,Y}$; define: the standardized random variables $\bar{X}$ and $\bar{Y}$ as:

$$\bar{X} = (X - E[X])/\sqrt{Var(X)}$$

$$\bar{Y} = (Y - E[Y])/\sqrt{Var(Y)}$$

(a) Show that $Cov(\bar{X}, \bar{Y})$ and $\rho_{X,Y}$.

$$\begin{aligned} Cov(\bar{X}, \bar{Y}) &= Cov(\frac{X - E[X]}{\sqrt{Var(X)}}, \frac{Y - E[Y]}{\sqrt{Var(Y)}}) \\ &= \frac{1}{\sqrt{Var(X)}\sqrt{Var(Y)}} Cov(X - E[X], Y - E[Y]) \\ &= \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} \\ Cov(\bar{X}, \bar{Y}) &= \rho_{X,Y} \end{aligned}$$

(b) Principal component analysis (PCA) is normally defined by the eigenvalue decomposition of the sample covariance matrix for multivariate data. Look at the R PCA function prcomp(). What is the impact of setting center=T and scale=T when calling prcomp()? When might this be desirable?

The impact of setting up center=T and scale=T when calling pcomp() in R is to have the data normalized before the principal component extraction is conducted. It is generally a good idea to have your variables scaled, to ensure that the result does not get affected by the scale differences each variable has, unless you are 100% sure that all your variables are recorded in the same scale.

## Question 3

(Based on Rice 4.74) The number of offspring of an organism is a discrete random variable with mean $\mu$ and variance $\sigma^2$. Each of its offspring reproduces in the same manner. Hint: use the Law of Total Expectation.

(a) Find the expected number of offspring in the third generation.
We know that $E[T_1] = \mu$ and $E[T_2|T_1] = T_1\mu$, hence $E[T_3|T_2] = T_2\mu$

To get the expected value, we can simply take expectations over the conditional expectations:

$$\begin{aligned} E[E[T_2|T_1]] &= E[T_2] \\ &= E[T_1\mu] \\ &= \mu\ E[T_1] \\ &= \mu^2 \end{aligned}$$

Similarly,

$$\begin{aligned} E[E[T_3|T_2]] &= E[T_3] \\ &= E[T_2\mu] \\ &= \mu\ E[T_2] \\ &= \mu^3 \end{aligned}$$

(b) Find the variance of the number of offspring in the third generation.

We know that $Var(T_1) = \sigma^2$. Hence:

$$Var(T_2|T_1) = T_1\sigma^2$$
$$Var(T_3|T_2) = T_2\sigma^2$$

Similar to previous method, we take expected values pver each conditional variance:

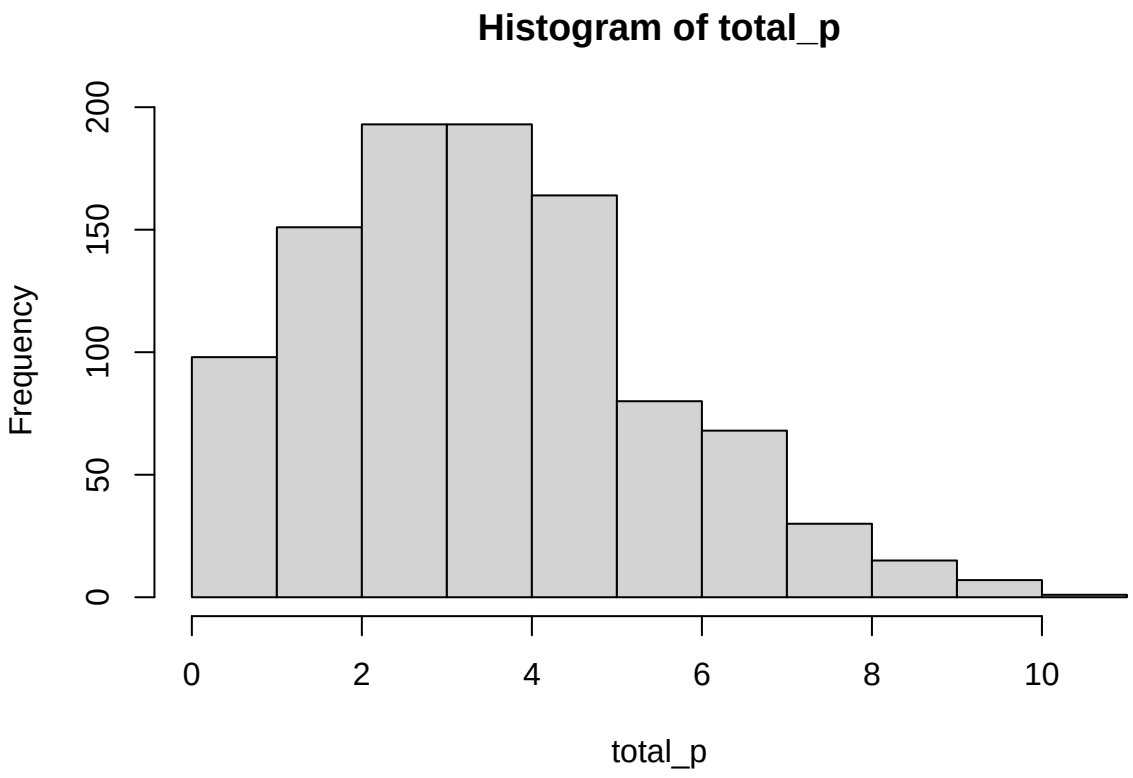
$$\begin{aligned} E[Var(T_2|T_1)] &= Var(T_2) \\ &= E[T_1\sigma^2] \\ &= \sigma^2 \ E[T_1] \\ &= \sigma^2\mu \end{aligned}$$

Similarly,

$$\begin{aligned} E[Var(T_3|T_2)] &= Var(T_3) \\ &= E[T_2]\sigma^2] \\ &= \sigma^2\mu^2 \end{aligned}$$

(c) Validate your answers to a) and b) via simulation with the number of offspring represented by a Poisson RV with = 2. Create 1000 separate populations that each include 3 generations and use a histogram to visualize the empirical distribution of the number of offspring in the third generation. Estimate the expected number of 3rd generation offspring using the average across all 1000 simulations and estimate the variance of the number using the R var() function (we will learn the basis for these estimates in Chapter 8). Compare these estimates with the values computed according to the results in part a) and b).

```
set.seed(26)

# get value for 2nd and 3rd gen for 1000 diff populations
for (i in 1:1000){
  second_gen_offsp <- rpois(i,2)
  for (j in length(second_gen_offsp)){
    third_gen_offsp <- rpois(j, 2)
  }
}

# get sum value for each population from 1000 diff populations
total_p <- c()
for (i in 1:1000){
  each_p <- sum(second_gen_offsp[i], third_gen_offsp[i])
  total_p <- append(total_p, each_p)
}

hist(total_p)
```




```
mean(total_p)
```

```
## [1] 3.965
```

```
var(total_p)
```

```
## [1] 4.059835
```

## Question 4

(Optional - Based on Rice 4.81) Find the moment-generating function of a Bernoulli RV and use it to find the mean, variance and third central moment.

Using the definition of moment generating function, we get:

$$\begin{aligned} M_X(t) &= E[e^t X] \\ &= \sum_{x e R_X} e^{tX} pX(x) \\ &= e^{t*1} pX(1) + e^{t*0} pX(0) \\ &= 1 - p + pe^t \end{aligned}$$

Hence, the moment generating function of a Bernoulli RV X is defined for any t in R:

$$M_X(t) = 1 - p + p\ e^t$$

Asked for mean, variance and its third central moment.

## Question 5

(Based on Rice 5.1) Let $X1, X2, ...$ be a sequence of independent random variables with $E[Xi] = \mu$ and $Var(Xi) = \sigma_i^2$. Show that if $n^{-2} \sum_{i=1}^{n} \sigma_i^2 \to 0$, then $\bar{X} \to \mu$ in probability.

Chebyshev's Inequality states that if X be an RV with $E[X] = \mu$ and $Var(X) = \sigma^2$. For any $\epsilon > 0$:

$$P(|\hat{X} - \mu| > \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$$

$$\begin{aligned} E(\bar{X}) &= E[\frac{1}{n}\sum_{i=1}^{n} x_i] \\ &= \frac{1}{n}\sum_{i=1}^{n} E[x_i] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu \\ &= \mu \end{aligned}$$

$$\begin{aligned} Var(\bar{X}) &= Var(\frac{1}{n}\sum_{i=1}^{n} x_i) \\ &= \frac{1}{n^2} Var(\sum_{i=1}^{n} x_i) \\ &= \frac{1}{n^2} Var(\sum_{i=1}^{n} \sigma^2) \end{aligned}$$

$$\begin{aligned} 0 \leq P(|\hat{X} - \mu| > \epsilon) &\leq \frac{\sigma^2}{\epsilon^2} \\ 0 \leq P(|\hat{X} - \mu| > \epsilon) &\leq \frac{\sigma^2}{\epsilon^2} \\ \lim_{n\to\infty} P(|\hat{X} - \mu| > \epsilon) &\leq \frac{1}{\epsilon^2}\frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 \\ \lim_{n\to\infty} P(|\hat{X} - \mu| > \epsilon) &= 0 \end{aligned}$$

## Question 6

(Optional - Based on Rice 5.5) Using moment-generating functions, show that as $n \to \infty$, $p \to 0$, and $np \to \lambda$, the binomial distribution with parameters n and p tends to the Poisson distribution.

## Question 7

(Based on Rice 5.16) Suppose that $X_1, ..., X_{20}$ are independent random variables with density functions $f(x) = 3x^2, 0 \leq x \leq 1$. Let $S = X_1 + ... + X_{20}$.

(a) Use the central limit theorem to approximate $P(S \leq 14)$. Fornuka

$$\begin{aligned} P(S \leq 14) &= P(\frac{S - \frac{60}{4}}{\sqrt{6/8}} \leq \frac{14 - \frac{60}{4}}{\sqrt{6/8}}) \\ &= P(\frac{S - \frac{60}{4}}{\sqrt{6/8}} \leq -1.16) \\ \omega \approx &= 1 - 0.877 = 0.123 \end{aligned}$$

(b) If you are instead asked to approximate $P(S \leq 15)$, what simplifications can be made in the calculation?

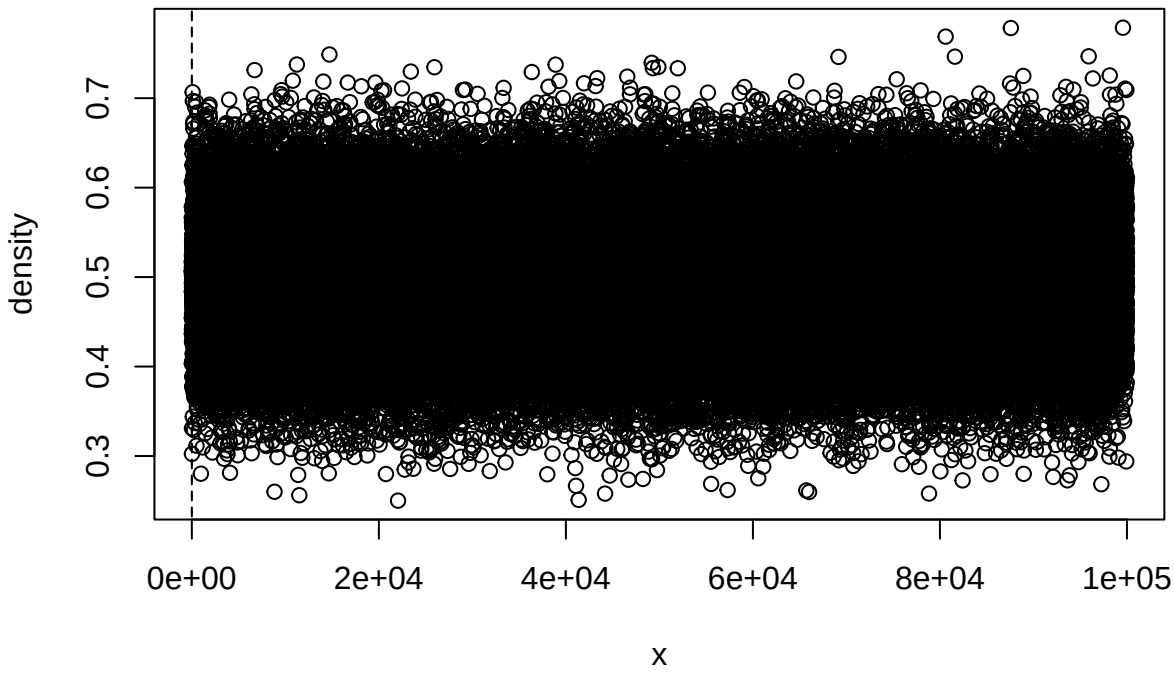
$$\begin{aligned} P(S \leq 15) &= P(\frac{S - \frac{60}{4}}{\sqrt{6/8}} \leq \frac{15 - \frac{60}{4}}{\sqrt{6/8}}) \\ &= P(\frac{S - \frac{60}{4}}{\sqrt{6/8}} \leq 0) \\ &= 0 \end{aligned}$$

the value goes to negative direction.

(c) Validate the approximation by plotting the CLT-based density (compute this using dnorm()) and true density of S. Use the inverse CDF method to simulate from the true density and plot using a kernel density estimate (R code plot(kernel()), we'll learn the details of kernel density estimation later in the course).

```
set.seed(26)

n_sim = 100000
n_val = 20

u_mean = matrix(nrow = n_sim, ncol = 4)
b_mean = matrix(nrow = n_sim, ncol = 4)
u_val = c()
b_val = c()
norm_proba = c()

func = function(x) {
  return(3*x^2)
}

for (i in 1:length(n_val)) {
  n = n_val[i]
```

```
  u_val = matrix(runif(n*n_sim), nrow = n_sim, ncol = n)
  b_val = matrix(func(n*n_sim), nrow = n_sim, ncol = n)
  u_mean[, i] = apply(u_val, 1, mean)
  b_mean[, i] = apply(b_val, 1, mean)
}

plotCLT = function(avg, mean, var) {
  x_val = seq(from=0, to=1, by=0.01)
  norm_proba = dnorm(x_val, mean=mean, sd=sqrt(var))
  plot(avg, xlab="x", ylab="density")
  lines(x_val, norm_proba, type="l", lty="dashed")
}

plotCLT(avg=u_mean[, 1], mean=0.75, var=(3/(80*n_val[1])))
```



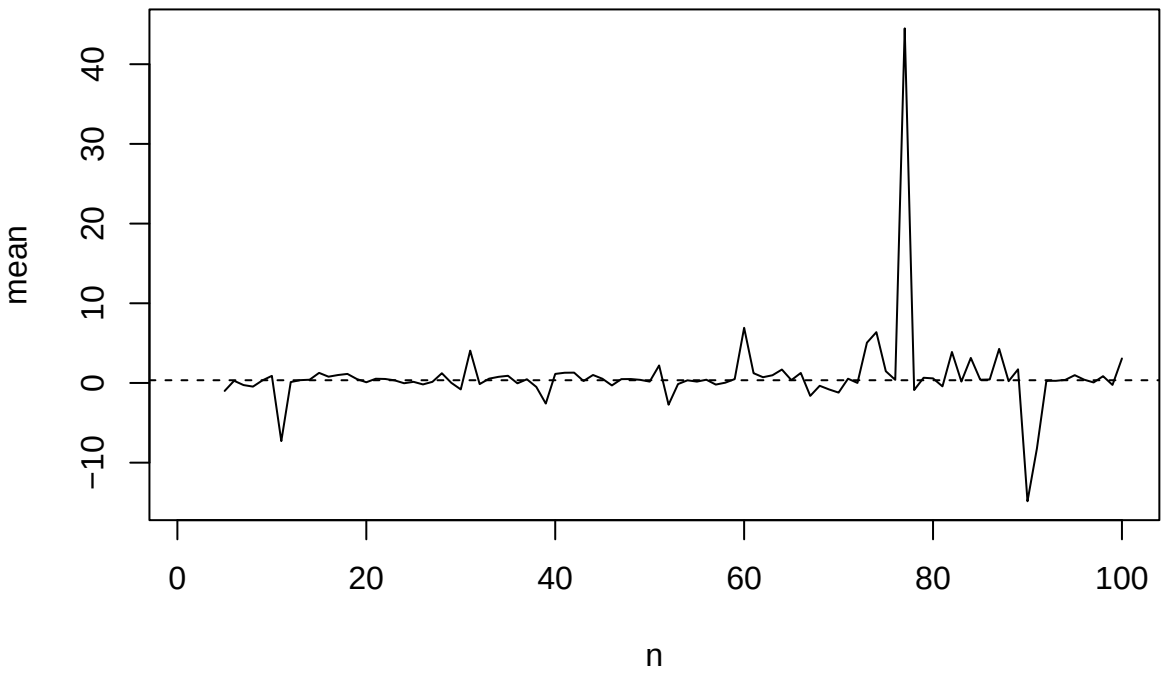

## Question 8

(Based on Rice 5.21) We wish to evaluate the integral $I(f) = \int_a^b f(x)\,dx$ using a numerical estimate. Let g be a density function on $[a, b]$. Generate $X_1, ..., X_n$ from g and estimate I by $I(f) = 1/n \sum_{i=1}^{n} \frac{f(X_i)}{g(X_i}$.
(a) Show that E(I(f)) = I(f)

$$
\begin{aligned}
E(\hat{I}(f)) &= E(1/n \sum_{i=1}^{n} \frac{f(X_i)}{g(X_i)}) \\
&= 1/nE(\sum_{i=1}^{n} \frac{f(X_i)}{g(X_i)}) \\
&= 1/nnE(\sum_{i=1}^{n} \frac{f(X_i)}{g(X_i)}) \\
&= E(\frac{f(X_i)}{g(X_i)}) \\
&= \int_a^b \frac{f(X_i)}{g(X_i)} g(X_i)\ dx \\
&= \int_a^b f(X_i)\ dx \\
E(\hat{I}(f)) &= I(f)
\end{aligned}
$$

(b) Demonstrate the result in a) via simulation with f(x) the density of the standard normal, a=0, b=1 and g(x) the density of the standard uniform distribution. Evaluate for n = 5,...,100. Plot I(f) as a function of n and include a horizontal line at I(f).

```
set.seed(26)

n=5:100
mean = c()
x_val = runif(n, 0, 1)
for (i in n) {
  f_val = density(dnorm(x_val, 0, 1))
  g_val = density(runif(n, 0, 1))
  mean[i] = mean(f_val$x / g_val$x)
}

plot(mean, type="l", xlab="n", ylab="mean")
I_g = pnorm(1) - pnorm(0)
abline(h=I_g, lty="dashed")
```




(c) (optional) Can this estimate be improved by choosing g to be other than uniform? Repeat the simulation in b) using a different choice of g (one you think will improve the estimate) and generate a new plot of I(f) vs. n that includes the estimates from both g functions. Discuss the relative estimation performance.