ENGRD 2700: Basic Engineering Probability and Statistics
Fall 2019
Homework 6 Solutions

Due **Friday Nov 15** at 11:59 pm. Submit to Gradescope by clicking the name of the assignment. See `https://people.orie.cornell.edu/yudong.chen/engrd2700_2019fa.html#homework` for detailed submission instructions.

The same stipulations from Homework 1 (e.g., independent work, computer code, etc.) still apply.

1. Compute the covariance and correlation of X and Y, when these continuous random variables have joint density function given by

$$f_{X,Y}(x,y) = \begin{cases} 1/8 & x \in [0,1), y \in [0,1) \\ 1/8 & x \in [1,2), y \in [1,2) \\ 3/8 & x \in [0,1), y \in [1,2), \text{ or } x \in [1,2), y \in [0,1) \\ 0 & \text{otherwise.} \end{cases}$$

Since the joint density is the same if we swap x and y in this particular case, the marginal distributions of X, Y will be the same. (This is not true in general, and then we have to compute each of them separately.)

$$f_X(x) = \int_0^2 f_{X,Y}(x,y)dy = \begin{cases} \frac{1}{2} & x \in [0,2) \\ 0 & \text{otherwise.} \end{cases}$$

(Notice that this is the density of a uniform random variable on $[0,2]$.) Now, let's calculate $E(X)$ and $Var(X)$

$$E(X) = \int_0^2 x f_X(x)\, dx = \int_0^2 \frac{x}{2}\, dx = 1$$

$$E(X^2) = \int_0^2 \frac{x^2}{2}\, dx = \frac{4}{3}$$

$$Var(X) = E(X^2) - (E(X))^2 = \frac{4}{3} - 1^2 = \frac{1}{3}$$

To untangle $Cov(X,Y)$, we first try to obtain $E(XY)$.

$$\begin{aligned} E(XY) &= \int_0^2 \int_0^2 xy f_{X,Y}(x,y)\, dxdy \\ &= \int_0^1 \int_0^1 xy \frac{1}{8}\, dxdy + \int_0^1 \int_1^2 xy \frac{3}{8}\, dxdy \\ &\quad + \int_1^2 \int_0^1 xy \frac{3}{8}\, dxdy + \int_1^2 \int_1^2 xy \frac{1}{8}\, dxdy \\ &= \frac{1}{32} + \frac{9}{32} + \frac{9}{32} + \frac{9}{32} \\ &= \frac{7}{8}. \end{aligned}$$

$$Cov(X,Y) = E(XY) - E(X)E(Y) = \frac{7}{8} - 1 = -\frac{1}{8}$$

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = -\frac{3}{8}$$

2. Suppose there are two stocks X and Y. The annual returns of the stocks are normally distributed, with the same mean $\mu_1 = \mu_2 = 12$ and variances $\sigma_1^2 = 9$ and $\sigma_2^2 = 16$, respectively.

(a) Assume that X and Y are independent. Suppose you hold one share of X and one share of Y. Compute the probability that your total annual return is greater than 25.

Similarly $X+Y$ follows the normal distribution we'll find its mean and variance.

$$\begin{aligned}
\mu &= E(X+Y) = E(X) + E(Y) = 24 \\
\sigma^2 &= Var(X+Y) = Var(X) + Var(Y) \\
&= \sigma_1^2 + \sigma_2^2 \\
&= 9 + 16 = 25
\end{aligned}$$

Thus

$$P(X+Y > 25) = 1 - \Phi(\frac{25-\mu}{\sigma}) = 1 - \Phi(\frac{25-24}{5}) = 0.4207$$

(b) Now assume that X and Y are negatively correlated with correlation $\rho = -0.6$. It turns out that the sum of two (possibly correlated) normal random variables is still normallly distributed. Compute the probability in (a).

We need compute $P(X+Y > 25)$. Since we know $X+Y$ follows the normal distribution we'll first find its mean and variance.

$$\begin{aligned}
\mu &= E(X+Y) = E(X) + E(Y) = 24 \\
\sigma^2 &= Var(X+Y) = \mathrm{Cov}(X+Y, X+Y) = Var(X) + Var(Y) + 2 \cdot \mathrm{Cov}(X,Y) \\
&= \sigma_1^2 + \sigma_2^2 + 2 \cdot \rho \cdot \sigma_1 \cdot \sigma_2 \\
&= 9 + 16 + 2 \cdot (-0.6) \cdot 3 \cdot 4 = 10.6
\end{aligned}$$

Thus

$$P(X+Y > 25) = 1 - \Phi\left(\frac{25-\mu}{\sigma}\right) = 1 - \Phi\left(\frac{25-24}{\sqrt{10.6}}\right) = 0.3782$$

(c) Still assume that $\rho = -0.6$. Suppose you can only purchase one share of X and Y in total. That is, your returns from the two stocks are wX and $(1-w)Y$, where $0 \le w \le 1$. What choice of w minimizes the variance of your total return?

$$\begin{aligned}
Var(wX + (1-w)Y) &= w^2 Var(X) + (1-w)^2 Var(Y) + 2w(1-w)Cov(X,Y) \\
&= 9w^2 + 16(1-w)^2 - 14.4w(1-w) \\
&= 39.4w^2 - 46.4w + 16
\end{aligned}$$

The optimal w is $-\frac{-46.4}{2\times 39.4} = 0.589$

(d) Now assume that X and Y are perfectly negatively correlated (i.e., $\rho = -1$), what value of w minimizes the variance of your total return?

$$\begin{aligned}
Var(wX + (1-w)Y) &= w^2 Var(X) + (1-w)^2 Var(Y) + 2w(1-w)Cov(X,Y) \\
&= 9w^2 + 16(1-w)^2 - 24w(1-w) \\
&= 49w^2 - 56w + 16
\end{aligned}$$

The optimal w is $-\frac{-56}{2*49} = \frac{4}{7}$.

(e) Would you prefer to invest under the conditions of part (c) or part (d)? Why?

In both cases, the mean return is 12. On the other hand, while in part (c) the minimal variance is 2.33, in part (d) the minimal variance is 0. In other words, under the conditions of part (d) we can guarantee the profit of 10 without any risk (which is called arbitrage in finance). You may prefer the former situation if you are risk-seeking, or the latter situation if you are risk-averse. You are indifferent if you are risk-neutral.

3. When we have sample data we can compute the *sample* covariance and the *sample* correlation. In particular, suppose we have data pairs $((X_i, Y_i), i = 1, 2, \ldots, n)$. The sample covariance is defined as

$$q_{X,Y} = \frac{1}{n-1}\sum_{i=1}^{n}[(X_i - \bar{X})(Y_i - \bar{Y})],$$

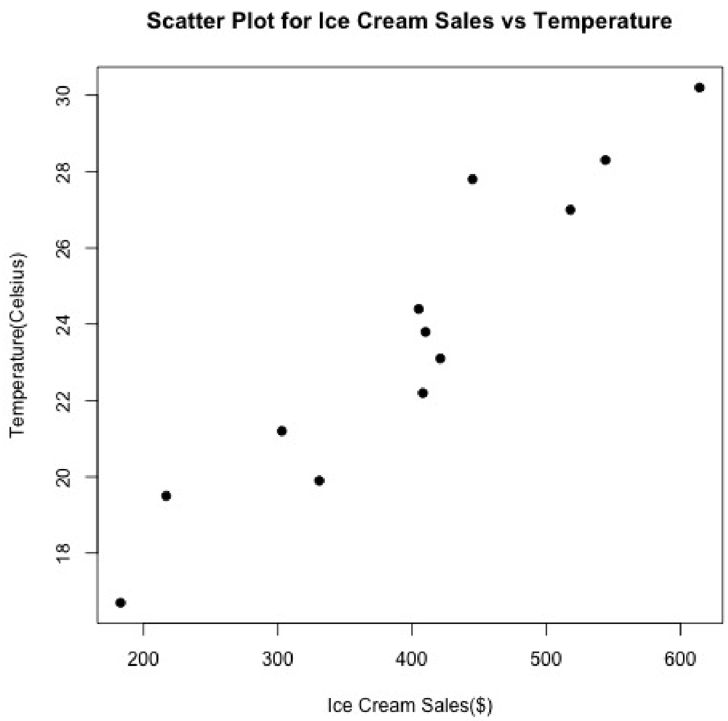
and the sample correlation is defined as

$$r = \frac{q_{X,Y}}{s_X s_Y},$$

where s_X^2 and s_Y^2 are the sample variances of the X_i's and Y_i's, respectively.

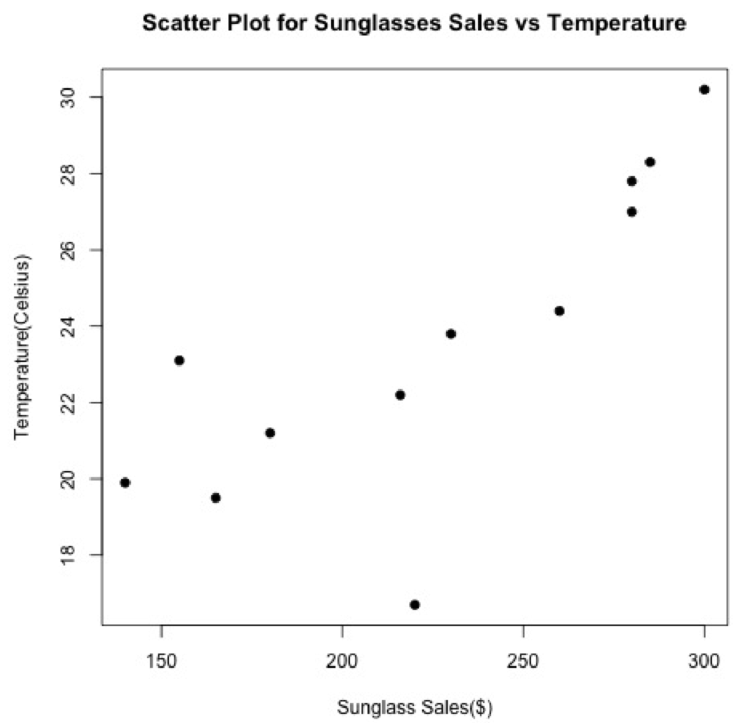
(a) Consider the dataset `DataForSunglasses.csv`. Generate a scatter plot for ice cream sales vs temperature, and compute their sample correlation. Do the same thing for sunglasses sales vs temperature, and for sunglasses sales vs ice cream sales. (That is, you need to generate three plots and compute three correlations.)

Ice cream sales vs temperature:



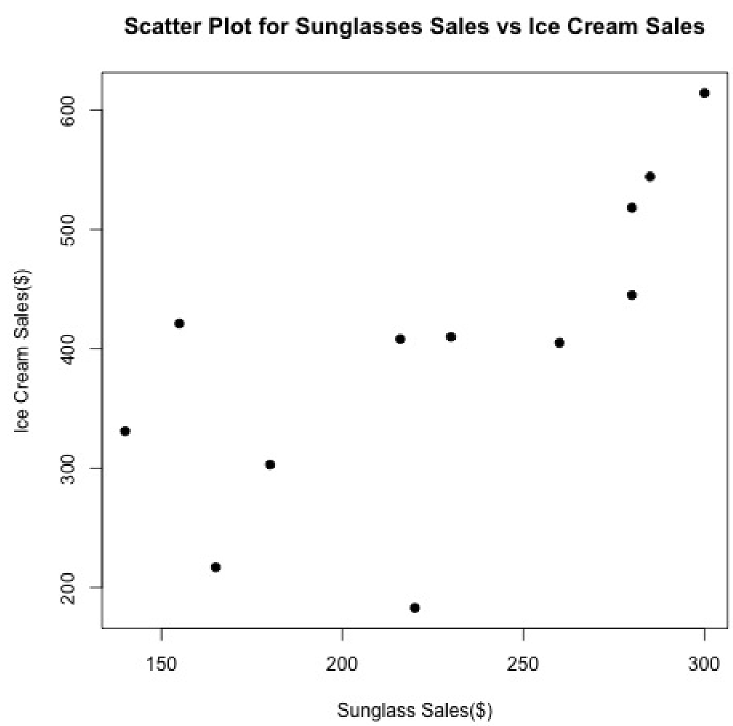
The sample correlation for Ice cream sales vs temperature is 0.95.

Sunglasses sales vs temperature:



The sample correlation for sunglasses sales vs temperature is 0.79.

Sunglasses sales vs ice cream sales:



The sample correlation for sunglasses sales vs ice cream sales is 0.70.
The R code is as follows. (Take ice cream sales vs temperature for example):

```
> t <- Temperature.Celsius.
> ic <- Ice.Cream.Sales...
```

```
> s <- Sunglass.Sales...
> plot(ic, t, pch=19, main = "Scatter Plot for Ice Cream Sales
vs Temperature", xlab="Ice Cream Sales($)",ylab="Temperature(Celsius)")
> cor(ic,t)
```

(b) You should see a positive correlation between sales of sunglasses and ice cream sales. Does this mean that ice cream makes people sensitive to sunlight? Explain in one sentence.
No. A positive correlation between sunglasses sales and ice cream sales only means both variables tend to move in the same direction, but not that there is a causal relationship between the two. It seems more likely that both ice-cream sales and sunglasses sales depend on temperature and so both tend to be high on hot days (but we can't conclude that either —- just suggesting it).

4. Suppose you flip a fair coin 100 times

(a) Use the Central Limit Theorem to approximate the probability that heads appears at most 46 times.
Let $X_1, ..., X_{100}$ be i.i.d $Bernouli(p)$ random variables, where X_i represents the outcome of the i^{th} flip. We have $E(X_i) = 0.5$ and $Var(X_i) = E(X_i^2) - E(X_i)^2 = E(X_i) - E(X_i)^2 = 0.25$

Let $S_{100} = \sum_{i=1}^{100} X_i$ be the number of heads in 100 flips, which has mean 50 and variance 25. Applying the CLT, we obtain the approximation

$$P(S_{100} \leq 65) = P(\frac{S_{100} - 50}{\sqrt{25}} \leq \frac{46 - 50}{\sqrt{25}}) \approx \Phi(-0.8) = 0.2119$$

(b) Write down an expression for the exact probability in part (a), and compute it. (In R, you can use `pbinom`.)
$S_{100} \sim Binomial(100, 0.5)$

$$P(S_{100} \leq 46) = \sum_{i=1}^{46} P(S_{100} = i) = \sum_{i=1}^{46} C_i^{100} 0.5^i \times 0.5^{100-i} = \sum_{i=1}^{46} C_i^{100} 0.5^{100}$$

type "pbinom(46,100,0.5)" in R and find $P(S_{100} \leq 46) = 0.2421$

(c) Our approximation from part (a) is not very accurate. This is primarily because we are using a continuous random variable (the normal) to approximate a discrete random variable (the binomial). To improve our approximation, we can use the fact that

$$P(S_{100} \leq 46) = P(S_{100} \leq 46 + c)$$

for any constant $0 \leq c < 1$, since S_{100} can only take integer values. It turns out that $c = 0.5$ works very well in practice, and so to approximate the probability of seeing at most 46 heads, we can apply the Central Limit Theorem to $P(S_{100} \leq 46.5)$ instead. Do this, and compare the approximation you obtain here to the one in part (a). The procedure is referred to as a *continuity correction*.

$$P(S_{100} \leq 46.5) = P(\frac{S_{100} - 50}{\sqrt{25}} \leq \frac{46.5 - 50}{\sqrt{25}}) \approx \Phi(-0.7) = 0.2420$$

We can see the result in (c) gives a better approximation to the exact probability than (a).

5. An insurance company looks at the records for millions of homeowners and conclude the probability of fire in a year is 0.01 for each house and the loss should a fire occur is \$10,000. Thus the expected loss from fire for each house is \$100. Assume the fires are independent. The company plans to sell fire insurance for \$120 (which is the expected loss plus \$20). If a house owner purchases the insurance and his/her house catches fire, the company will cover for the loss.

(a) If the company sells the insurance policy to 10 houses, what is the expected total profit of the company?
let X_i be the profit/loss for selling policy to the i-th house. Then with probability 0.99, $X_i = 120$ and with probability 0.01, $X_i = 120 - 10000 = -9880$, Thus $E(X_i) = 20$

$$E(S_{10}) = \sum_{i=1}^{10} E(X_i) = 200$$

(b) Compute the probability of bankruptcy for the company, that is, when total profit is negative.

The total money collected is $120 \times 10 = 1200$. if any of the house catches fire this year, the company has to pay 10000 dollars and go bankruptcy. Therefore,

$$P(\text{bankruptcy}) = 1 - P(\text{no fires for all household}) = 1 - 0.99^{10} = 9.56\%$$

which is way too high for a company.

(c) Use the Central Limit Theorem to approximately compute the probability of bankruptcy if the company sells the policy to 1 million houses.

The total amount of money collected is $120 \times 1,000,000 = 120,000,000$. Let L_i be the loss from fire for i^{th} household, and $L = \sum_i^{1,000,000} L_i$ be the total losses from fire the company has to pay. The L_i's are i.i.d., so

$$E(L_i) = 10000 \times 0.01 + 0 \times 0.99 = 100$$
$$Var(L_i) = E(L_i^2) - E(L_i)^2 = 990,000$$

Thus

$$E(L) = 100 \times 1,000,000 = 100,000,000$$
$$sd(L) = \sqrt{Var(L)} = \sqrt{Var(L_i) * 1,000,000} = 994,987$$

By CLT, L can be approximated by $Normal(100,000,000,\ 994,987^2)$

$$P(L \geq 120,000,000) \approx 1 - \Phi\left(\frac{120,000,000 - 100,000,000}{994,987}\right) = 1 - \Phi(20.1) \approx 0$$

6. A light bulb has a lifetime that is exponentially distributed with rate parameter $\lambda = 5$. Let L be a random variable denoting the sum of the lifetimes of 50 such bulbs. Assume that the bulbs are independent.

(a) Compute $E[L]$ and $\text{Var}(L)$.

Let L_i be the lifetime of light bulb i. Since $L_i \sim \text{Exponential}(5)$, it follows that

$$E[L_i] = \frac{1}{5} \qquad \text{Var}(L_i) = \frac{1}{5^2} = \frac{1}{25}.$$

Since $L = \sum_{i=1}^{50} L_i$, it follows that

$$E[L] = \sum_{i=1}^{50} E[L_i] = 10 \qquad \text{Var}(L) = \sum_{i=1}^{50} \text{Var}(L_i) = 2.$$

(b) Use the Central Limit Theorem to approximate $P(8 \leq L \leq 12)$.

Leveraging our work from part (a) and applying the Central Limit Theorem, we find

$$\begin{aligned} P(8 \leq L \leq 12) &= P(L \leq 12) - P(L \leq 8) \\ &= P\left(\frac{L-10}{\sqrt{2}} \leq \frac{12-10}{\sqrt{2}}\right) - P\left(\frac{L-10}{\sqrt{2}} \leq \frac{8-10}{\sqrt{2}}\right) \\ &= \Phi(\sqrt{2}) - \Phi(-\sqrt{2}) \\ &\approx 0.843. \end{aligned}$$

(c) Use the Central Limit Theorem to find an interval (a, b), *centered at* $E[L]$, such that

$$P(a \leq L \leq b) = 0.95.$$

That is, your interval should be of the form $(E[L] - c, E[L] + c)$, for some constant $c > 0$.

By the Central Limit Theorem, L is approximately normally distributed. Since the PDF of the normal random variable is symmetric about its mean, it follows that we should find c satisfying

$$P(L \leq E[L] - c) = P(L \geq E[L] + c) = 0.025.$$

We'll work with the left-most term. Using what we found in part (a) yields

$$\begin{aligned} P(L \leq E[L] - c) &= P(L \leq 10 - c) \\ &= P\left(\frac{L - 10}{\sqrt{2}} \leq \frac{10 - c - 10}{\sqrt{2}}\right) \\ &\approx \Phi\left(\frac{-c}{\sqrt{2}}\right) \end{aligned}$$

By a property shown in lecture, $\Phi(-1.960) = 1 - \Phi(1.960) = 0.025$, and so we want c to satisfy

$$\frac{-c}{\sqrt{2}} = -1.960.$$

Solving yields $c = 1.960\sqrt{2} \approx 2.772$, and we obtain the interval $[7.228, 12.772]$.