In class topics/quizes

Question 1

Suppose A,B are events with O-P(B)+1.

P(A|B)=0then P(A|B)-1.

TRUE

Question 2

Suppose A,B are events with both OP(A)=1 and 0-PB-1.

If P (BA) - 1 then P(A|B) — 1.

FALSE

Question 3

Suppose A,B are events with both OP(A)=1 and 0-PB-1.

IP(BA")-1 then P(A|B)-1

TRUE

Question 4

Suppose A,B are events with both OP(A)=1 and 0-PB-1.

Suppose both P(B)-1 and P(BA)-1. Then P(A) = P(B).

TRUE

Consider the sample space for the random experiment "flip a coin twice", that is = {HH, HT, TH,TT}.

Now consider the following two random variables defined on the sample space N.

X(w) = number of Hs in w.

Y(w) = number of Ts in w.

Question 1

The random variables X and Y are the same. That is, X (w) = Y(w) for all w € .

False

Question 2

The random variables X and Y have the same distribution. That is, their probability mass functions are the same.

True

Suppose X, Y are both Bernoulli trials, that is, they take values 0, 1.

If cov(X,Y) = 0 then X and Y are independent.

TRUE

Suppose X is a random variable that takes values 0, 1.

Suppose Y is a random variable that takes values -1, 0, 1.

If cov(X,Y) = 0 then X and Y are independent.

FALSE

WEEK2

The questions below refer to the following hypothetical situation.

Albert and Bob are old friends who frequently play with each other some game (chess, ping-pong, blackjack, checkers,...).

Their skills are such that for any one round of the game Albert wins with probability 0.45 or equivalently Bob wins with probability 0.55.

Question 1

Suppose Albert and Bob play three rounds of the game. What is the probability that Bob wins at least two rounds?

0.5747

Question 2

Suppose Albert and Bob play five rounds of the game. What is the probability that Bob wins at least three rounds?

0.593

Question 3

Suppose Albert and Bob play 11 rounds of the game. What is the probability that Bob wins at least 6 rounds?

0.633

Question 4

Suppose Albert and Bob play 101 rounds of the game. What is the probability that Bob wins at least 51 rounds?

0.843

The following two questions refer to the following context. Suppose X is a continuous random variable with probability density function f(x) and cumulative distribution function F(x).

The cumulative distribution function F(x) is never strictly larger than one.

TRUE

The probability density function f(x) is never strictly larger than one.

FALSE

Suppose X is uniformly distributed in the interval [5,25].

What is the median of X?

15

The questions below refer to the following hypothetical situation.

Albert and Bob are old friends who frequently play with each other some game (chess, ping-pong, blackjack, checkers...).

Their skills are such that for any one round of the game Albert wins with probability 0.45 or equivalently Bob wins with probability 0.55.

Question 1

Suppose Albert and Bob play three rounds of the game. What is the probability that Bob wins at least two rounds?

0.5747

Question 2

Suppose Albert and Bob play five rounds of the game. What is the probability that Bob wins at least three rounds?

0.593

Question 3

Suppose Albert and Bob play 11 rounds of the game. What is the probability that Bob wins at least 6 rounds?

0.633

Question 4

Suppose Albert and Bob play 101 rounds of the game. What is the probability that Bob wins at least 51 rounds?

0.843

The following two questions refer to the following context. Suppose X is a continuous random variable with probability density function f(x) and cumulative distribution function F(x).

Question 1

The cumulative distribution function F(x) is never strictly larger than one.

[✓ True]

False

Question 2

The probability density function f(x) is never strictly larger than one.

True

[✓ False]

Question 1

Suppose Z is a standard normal random variable.

Find the median of Z.

0

Question 2

Suppose Z is a standard normal random variable.

Find the 0.975-quantile of Z

1.96

Question 3

Suppose Z is a standard normal random variable.

Find the 0.995-quantile of Z.

2.576

Question 4

Suppose X is a normal random variable such that 0.995-quantile(X) = 40.909 and

0.975-quantile(X) -33.519.

Find the mean of X.

10

The two questions below refer to the following context suppose X-B(m,p) and Y- Bin,p) where 0<p<1 and m< n are positive integers.

The variables X/m and Y/n have the same expected value

True

Question 2

The variables X/m and Y/n have the same variance.

False

Question 1

Suppose X1,..., X, are independent normal variables. Then their "sample mean"

is normal.

True

Question 2

Suppose X1,...,X, are independent standard normal variables. Then their "sample

mean"

is standard normal.

False

The three questions below refer to the following setup.

Suppose X1,...,X, are id Bernoulli trials with probability of success 0.5 and Observe that nX~ B(n, 0.5) and thus

x= PX-0.50.1) = P(0.4≤0.6)

Question 1

Find P(X -0.50.1) form = 10.

Correct!

0.656

Correct Answers

0.656 (with margin: 0.01)

Question 2

Find P(X -0.5 <0.1) form = 100.

Correct!

0.964

Correct Answers

0.9648 (with margin: 0.01)

Question 3

Find P(X -0.5 <0.1) form=200.

Correct!

0.996

Correct Answers 0.9963 (with margin: 0.001)

(0.4n≤ n ≤ 0.6m)

0.34/0.34 pts

0.33/0.33 pts

0.33/0.33 pts

Question 1

The sum of niid random normal variables is normal.

True

Question 2

The sum of niid uniform random variables is uniform.

False

Question 3

The sum of niid exponential random variables is exponential.

False

Question 1

Suppose a distribution has mean μ and standard deviation o

Recall that for or Є (0, 1) the (1-0) acceptance interval for the sample mean of

observations is

Therefore the SMALLER the a the WIDER the acceptance interval.

True

Correct!

Question 2

Suppose a distribution has mean and standard deviation o

Recall that for or Є (0, 1) the (1-0) acceptance interval for the sample mean of

observations is

Therefore the LARGER the number of observations the WIDER the acceptance interval.

False

The three questions below refer to the following setup.

Suppose X1,...,X, are id Bernoulli trials with probability of success 0.5 and X+-+x Observe that nX~ B(n,0.5) and thus

x= P(X -0.50.1) = P(0.4≤ X ≤ 0.6) = P(0.4n ≤ n ≤ 0.6m)

Question 1

Find P(X -0.5 <0.1) form = 10.

0.656

0.656 (with margin: 0.01)

Question 2

Find P(X-0.50.1) for 100.

0.964

0.9648 (with margin: 0.01)

Question 3

Find PX-0.50.1) form 200.

Correct!

0.996

Question 1

The sum of niid random normal variables is normal.

True

Question 2

The sum of niid uniform random variables is uniform.

False

Question 3

The sum of n id exponential random variables is exponential.

False

Question 1

0.5/0.5 pts

Suppose Z has standard normal distribution and T hast distribution with (n-1) degrees of freedom for some sample size $n > 10$. Then $P(-1<Z<1)<P(-1<T<1)$

False

Question 2

Suppose we are analyzing a random sample of size > 10 and or € (0,1). Then

True

Question 1

According to a recent Gallup poll, 42% of Americans approve of the job Biden is doing as president. This sample proportion is based on a random sample of 812 adults living in the US.

Compute the UCL (upper confidence limit) of the 95% confidence interval for the population proportion of adults in the US who approve of the job Biden is doing as president.

The answer is a number between 0 and 1. Please use three digits of accuracy.

0.454

0.454 (with margin: 0.01)

Question 2

In the question above, what is the margin of error at the 95% confidence level? The answer is a number between 0 and 1. Please use three digits of accuracy.

0.034

0.034 (with margin: 0.001)

46-880 Introduction to Probability and Statistics, mini-1 2023
Solution to Little Test 1

1. Suppose you roll three dice. What is the probability that at least two of the rolls are the same?

   Solution. It is easier to compute the complement:

   $$\mathbb{P}(\text{the three rolls are different}) = \frac{6 \cdot 5 \cdot 4}{6^3} = \frac{20}{36}.$$

   Hence

   $$\mathbb{P}(\text{at least two rolls are the same}) = 1 - \mathbb{P}(\text{the three rolls are different}) = 1 - \frac{20}{36} = \frac{16}{36}.$$

2. Suppose 80% of all statisticians are shy, whereas only 15% of economists are shy. Suppose that 90% of the people at a large gathering are economists and the other 10% are statisticians.

   You meet a random person at the gathering. If that person is shy, what is the probability that the person is a statistician?

   Solution.

   Consider the following events

   A : the random person is a statistician.

   B : the random person is shy.

   We know $\mathbb{P}(A) = 0.1$ and also $\mathbb{P}(B|A) = 0.8$ and $\mathbb{P}(B|A^c) = 0.15$. We want $\mathbb{P}(A|B)$. Apply Bayes' Theorem:

   $$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A) \cdot \mathbb{P}(A)}{\mathbb{P}(B|A) \cdot \mathbb{P}(A) + \mathbb{P}(B|A^c) \cdot \mathbb{P}(A^c)} = \frac{0.8 \cdot 0.1}{0.8 \cdot 0.1 + 0.15 \cdot 0.9} = \frac{0.08}{0.215} = 0.372$$

3. Suppose 80% of all statisticians are shy, whereas only 15% of economists are shy. Suppose that 90% of the people at a large gathering are economists and the other 10% are statisticians.

   You meet a random person at the gathering. What is the probability that the person is shy?

   Solution. We want

   $$\mathbb{P}(B) = \mathbb{P}(B|A) \cdot \mathbb{P}(A) + \mathbb{P}(B|A^c) \cdot \mathbb{P}(A^c) = 0.8 \cdot 0.1 + 0.15 \cdot 0.9 = 0.215.$$

4. Suppose X is a binomial random variable with 50 trials and probability of success 0.44, that is, $X \sim B(50, 0.44)$.

   What is the numerical value that X takes with highest probability?

   Solution. The possible values of X are $0, 1, \cdots, 50$. Using a spreadsheet we can compute

   $$\mathbb{P}(X = x) = \texttt{binom.dist}(\texttt{x}, 50, 0.44, 0) \text{ for } x = 0, 1, \ldots, 20.$$

   It is then evident that the highest probability value is 22 with probability

   $$\texttt{binom.dist(22, 50, 0.44, 0)} = \texttt{0.113}.$$

5. A hotel has 50 rooms. Assume that a reservation is a “no-show” (that is, it does not show up) with probability 0.1. Consequently, the hotel routinely accepts more than 50 reservations. What is the largest number of reservations that the hotel can accept so that the probability that the hotel will be overbooked is less than 0.2?

   (The hotel is overbooked if it does not have enough rooms for the reservations who show up.)

   Solution. Let n = number of reservations that the hotel accept and X = number of reservations that show up. Then $X \sim B(n, 0.9)$ and the hotel is overbooked when $X \geq 51$. Hence the probability that the hotel is overbooked is

   $$\mathbb{P}(X \geq 51) = 1 - \mathbb{P}(X \leq 50) = 1 - \texttt{binom.dist}(50, \texttt{n}, 0.9, 1).$$

   We have the following values of this probability for different values of n

n	51	52	53	54	55	56
$\mathbb{P}(X \geq 51)$	0.00464	0.02829	0.0898	0.1985	0.3451	0.5065

   Thus the largest n can be so that $\mathbb{P}(X \geq 51) < 0.2$ is $n = 54$.

6. Suppose calls arrive at a 1-800 service line according to a Poisson process with an average rate of arrival of 12 calls per hour.

   What is the probability that at least one call arrives within the next 10 minutes?

   Solution. Let X = number of calls that arrive in the next 10 minutes. Thus $X \sim \text{Pois}(12/6)$ and we want $\mathbb{P}(X \geq 1)$. That is

   $$\mathbb{P}(X \geq 1) = 1 - \mathbb{P}(X = 0) = 1 - \texttt{poisson.dist}(0, 12/6, 0) = 0.86466.$$

Solution to Little Test 2

1. Suppose X is a random variable with cumulative distribution function $F(x)$. If a and b are numbers with $a \leq b$, then $F(a) \leq F(b)$.

   Solution. TRUE: Since $\{X \leq a\} \subseteq \{X \leq b\}$ it follows that

   $$F(a) = \mathbb{P}(X \leq a) \leq \mathbb{P}(X \leq b) = F(b).$$

2. Suppose X is a continuous random variable with density function $f(x)$. If a and b are numbers with $a \leq b$, then $f(a) \leq f(b)$.

   Solution. FALSE: Suppose $X \sim U(0,1)$, that is, X is uniformly distributed between 0 and 1. Then $f(0.5) = 1$ but $f(2) = 0 < f(0.5)$.

3. If the random variables X, Y are uncorrelated then $\mathrm{var}(X+Y) = \mathrm{var}(X-Y)$ then X, Y.

   Solution. TRUE: We know that

   $$\mathrm{var}(X+Y) = \mathrm{var}(X) + \mathrm{var}(Y) + 2\mathrm{cov}(X,Y) = \mathrm{var}(X) + \mathrm{var}(Y)$$

   and

   $$\mathrm{var}(X-Y) = \mathrm{var}(X) + \mathrm{var}(Y) - 2\mathrm{cov}(X,Y) = \mathrm{var}(X) + \mathrm{var}(Y).$$

   Hence $\mathrm{var}(X+Y) = \mathrm{var}(X) + \mathrm{var}(Y) = \mathrm{var}(X-Y)$.

4. If the random variables X, Y are such that $\mathrm{var}(X+Y) = \mathrm{var}(X-Y)$ then X, Y are uncorrelated.

   Solution. TRUE: We know that

   $$\mathrm{var}(X+Y) = \mathrm{var}(X) + \mathrm{var}(Y) + 2\mathrm{cov}(X,Y)$$

   and

   $$\mathrm{var}(X-Y) = \mathrm{var}(X) + \mathrm{var}(Y) - 2\mathrm{cov}(X,Y).$$

   Hence $\mathrm{var}(X+Y) = \mathrm{var}(X-Y)$ implies that

   $$\mathrm{var}(X) + \mathrm{var}(Y) + 2\mathrm{cov}(X,Y) = \mathrm{var}(X) + \mathrm{var}(Y) - 2\mathrm{cov}(X,Y).$$

   Hence $4\mathrm{cov}(X,Y) = 0$ and thus $\mathrm{cov}(X,Y) = 0$. Therefore X, Y are uncorrelated.

5. Suppose X is a normal random variable such that 0.995-quantile$(X) = 40.909$ and 0.975-quantile$(X) = 33.519$.

   Find the mean of X.

   Solution. Suppose $\mu = \mathbb{E}(X)$ and $\sigma = \mathrm{stdev}(X)$. From the properties of the normal distribution we know that

   $$0.995\text{-quantile}(X) = \mu + 2.576\sigma$$

and

$$0.975\text{-quantile}(X) = \mu + 1.96\sigma.$$

Hence we have the two equations

$$\mu + 2.576\sigma = 40.909$$
$$\mu + 1.96\sigma = 33.519$$

If we multiply the first one by 1.96 and the second one by 2.576 and subtract, we get

$$0.616\mu = (2.576 - 1.96)\mu = 2.576 \cdot 33.519 - 1.96 \cdot 40.909 = 6.1633.$$

Thus $\mu = 10.0536$.

6. Suppose the random variables $X_1, \dots, X_n$ are iid standard normal and consider their sample mean

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n}.$$

Find the smallest integer n such that $|\bar{X}_n| \leq 0.1$ with probability at least 0.95. In other words, find the smallest integer n such that $\mathbb{P}(|\bar{X}_n| \leq 0.1) \geq 0.95$.

Solution. The properties of normal distributions, iid, and sample mean imply that $\bar{X}_n \sim N(0, 1/n)$ or equivalently $\sqrt{n}\bar{X}_n \sim N(0,1)$. Thus we have

$$\mathbb{P}(|\bar{X}_n| \leq 0.1) = \mathbb{P}(|\sqrt{n}\bar{X}_n| \leq 0.1\sqrt{n}).$$

We know that for $Z \sim N(0,1)$ we have $\mathbb{P}(|Z| \leq 1.96) = 0.95$. Since $\sqrt{n}\bar{X}_n \sim N(0,1)$ to have $\mathbb{P}(|\sqrt{n}\bar{X}_n| \leq 0.1\sqrt{n}) \geq 0.95$ we must have $0.1\sqrt{n} \geq 1.96$, that is

$$n \geq 19.6^2 = 384.16.$$

Thus $n = 385$.

Introduction to Probability and Statistics
Solution to Problem Set 1

1. (Sample Spaces & Probability Rules)

   A standard 52-card deck is divided into 13 ranks and 4 suits. The ranks are *A, 2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K.* The suits are *Spades, Clubs, Diamonds, and Hearts.* Each card has a rank and a suit. Additionally each suit has a color; Clubs and Spades are *black* while Diamonds and Hearts are *red*. King, Queen and Jack are *face* cards. So, there are 12 face cards in the deck of 52 playing cards. From a well shuffled pack of 52 cards, we draw two random cards with replacement, that is, after we draw the first card (and examine it) we replace it in the deck and then we draw the second card (after reshuffling of course).

   (a) Describe the sample space for this experiment. How many elements are there in the sample space?

   Solution. The sample space has $52 \times 52 = 2704$ elements. The first drawn card can be any of the 52 cards, as well as the second one. We can enumerate the sample space as $S = \{$(SA,SA), (SA,S2), (SA,S3), $\dots$, (SA,SK), (HA, HA), $\dots$, (HA,HK), $\dots$(HK,HK)$\}$ where A is ace, S is spades and H is hearts and so on.

   (b) What is the probability that at least one card is black?

   Solution. This probability can be expressed as

   $$\begin{aligned} P\{\text{At least one card is black}\} &= 1 - P\{\text{Both cards are red}\} \\ &= 1 - 1/2 \times 1/2 = 3/4 \\ &= 0.75. \end{aligned}$$

   (c) What is the probability that both cards are aces?

   Solution. The probability of picking an ace from 52 cards is $4/52 = 1/13$. The probability of picking an ace twice (since there is replacement) from 52 cards is $1/13^2 = 0.006$.

   (d) What is the probability that both cards have the same suit?

   Solution. The probability that one will pick the same suit as the first one is $1/4 = 0.25$.

   (e) What is the probability that the cards are of different ranks?

   Solution. This can be seen as:

   $$\begin{aligned} P\{\text{Cards have different ranks}\} &= 1 - P\{\text{Picking the same rank cards}\} \\ &= 1 - 1/13 = 0.923. \end{aligned}$$

   (f) How do the answers to (a) - (e) change if instead, we draw two random cards without replacement, that is, after we draw the first card we do not replace it in the deck.

   Solution.

   a. Since there is no replacement now, we have $52 \times 51 = 2652$ elements in the sample space. The sample space loses all duplicate elements such as (C2,C2) , (H4,H4) , $\dots$, (SA,SA) .

   b. This probability can be expressed as

   $$\begin{aligned} P\{\text{At least one card is black}\} &= 1 - \mathbb{P}\{\text{Both cards are red}\} \\ &= 1 - 1/2 \times 25/51 = 0.755 \end{aligned}$$

c. Since there is no replacement, we have that

$$\mathbb{P}\{\text{Both are aces}\} = 4/52 \times 3/51 = 0.005.$$

d. Probability that the second card will have the same suit with first one is $12/51 = 0.235$.

e. This probability can be expressed as

$$\begin{aligned}\mathbb{P}\{\text{Cards have different ranks}\} &= 1 - \mathbb{P}\{\text{Picking the same rank cards}\} \\ &= 1 - 3/51 = 48/51 = 0.941\end{aligned}$$

2. (Conditional Probability)

Consider the same scenario as in Problem 1. From a well shuffled pack of 52 cards, we draw a random card.

(a) Describe the sample space for this experiment. How many elements are there in the sample space?
Solution. The sample space has 52 elements,
$S = \{$SA, S2, . . . , SK, CA, . . . ,CK, DA, . . . , DK, HA, . . . , HK$\}$.

(b) What is the probability that the drawn card is the ace of diamonds given that it is an ace?
Solution. We have that

$$\begin{aligned}\mathbb{P}\{\text{Card is ace of diamonds}|\text{Card is ace}\} &= \frac{\mathbb{P}\{\text{Card is both ace and ace of diamonds}\}}{\mathbb{P}\{\text{ card is ace}\}} \\ &= \frac{1/52}{4/52} = 1/4 = 0.25.\end{aligned}$$

(c) What is the probability that the drawn card is a Queen given that it is a face card?
Solution.

$$\begin{aligned}\mathbb{P}\{\text{Card is a Queen}|\text{Card is a face card}\} &= \frac{\mathbb{P}\{\text{Card is face card of Queens}\}}{\mathbb{P}\{\text{Card is face card}\}} \\ &= \frac{4/52}{12/52} = 1/3 = 0.333.\end{aligned}$$

(d) What is the probability that the drawn card is a Spade given that it is black?
Solution.

$$\begin{aligned}\mathbb{P}\{\text{Card is a Spade}|\text{Card is a black card}\} &= \frac{\mathbb{P}\{\text{Card is a Spade}\}}{\mathbb{P}\{\text{Card is black}\}} \\ &= \frac{13/52}{1/2} = 1/2 = 0.5.\end{aligned}$$

3. (Joint & conditional probability)

This exercise is in the spirit of *market basket analysis* technique widely used by retailers. Assume the coffee shop in the Tepper building sells coffee and bagels in addition to other products. The coffee shop has observed that 80% of its customers buy coffee, 70% buy bagels, and 60% buy both coffee and bagels.

(a) Based on the above data, complete the following probability table concerning the behavior of random customers at the coffee shop:

	bagels	no-bagels	
coffee	0.6		0.8
no-coffee			
	0.7		1.00

Solution. Fill in the blanks so that the numbers in each row and column add up properly:

	bagels	no-bagels	
coffee	0.6	0.2	0.8
no-coffee	0.1	0.1	0.2
	0.7	0.3	1.00

(b) What is the probability that a customer at the coffee shop buys neither coffee nor bagels?
Solution. Let A = customer buys coffee, B = customer buys bagels. We need $\mathbb{P}(\overline{A} \cap \overline{B})$. From the numbers in part (a) we get

$$\mathbb{P}(\overline{A} \cap \overline{B}) = 0.1.$$

(c) Suppose a customer buys coffee. What is the probability that that customer does not buy bagels?
Solution. Keeping the notation in (b) now we need $\mathbb{P}(\overline{B}|A)$:

$$\mathbb{P}(\overline{B}|A) = \frac{\mathbb{P}(A \cap \overline{B})}{\mathbb{P}(A)} = \frac{0.2}{0.8} = 0.25.$$

(d) Are the events "a customer buys coffee" and "a customer buys bagels" independent?
Solution. No, because

$$\mathbb{P}(A \cap B) = 0.6 \neq 0.8 \cdot 0.7 = \mathbb{P}(A) \cdot \mathbb{P}(B).$$

4. (Bayes' theorem)

A crucial game of the Los Angeles Lakers basketball team depends on their key player who is recovering from injury. According to his doctor's report, there is a 50% chance that he will be able to play, and hence a 50% he will not be able to play. The coach has estimated the team's chance of winning at 70% if the player is able to play and 40% if he is unable to play.

(a) What is the probability that the Lakers will win the game?
Solution. We first list the probabilities that are given in the question.

$$\begin{aligned} \mathbb{P}(\texttt{Play}) &= 0.5, & \mathbb{P}(\texttt{Doesn't Play}) &= 0.5 \\ \mathbb{P}(\texttt{Win}|\texttt{Play}) &= 0.7, & \mathbb{P}(\texttt{Lose}|\texttt{Play}) &= 0.3 \\ \mathbb{P}(\texttt{Win}|\texttt{Doesn't Play}) &= 0.4, & \mathbb{P}(\texttt{Lose}|\texttt{Doesn't Play}) &= 0.6 \end{aligned}$$

We have

$$\begin{aligned} \mathbb{P}(\texttt{Win}) &= \mathbb{P}(\texttt{Win}|\texttt{Play}) \times \mathbb{P}(\texttt{Play}) + \mathbb{P}(\texttt{Win}|\texttt{Doesn't Play}) \times \mathbb{P}(\texttt{Doesn't Play}) \\ &= 0.7 \times 0.5 + 0.4 \times 0.5 = 0.550. \end{aligned}$$

(b) You have just heard that the Lakers won the game. What is the probability that the key player was able to play in the game?

Solution. We have

$$\begin{aligned}\mathbb{P}(\texttt{Play}|\texttt{Win}) &= \frac{\mathbb{P}(\texttt{Win}|\texttt{Play}) \times \mathbb{P}(\texttt{Play})}{\mathbb{P}(\texttt{Win})}\\ &= \frac{0.7 \times 0.5}{0.55} = 0.636.\end{aligned}$$

(c) You have just heard that the Lakers lost the game. What is the probability that the key player was unable to play in the game?

Solution. We have that

$$\begin{aligned}\mathbb{P}(\texttt{Doesn't Play}|\texttt{Lose}) &= \frac{\mathbb{P}(\texttt{Lose}|\texttt{Doesn't Play}) \times \mathbb{P}(\texttt{Doesn't Play})}{\mathbb{P}(\texttt{Lose})}\\ &= \frac{0.6 \times 0.5}{0.45} = 0.667.\end{aligned}$$

(d) Should the numerical answers in parts (b) and (c) add up to one? Explain.

Solution. No, not necessarily. They are not complementary events since the conditional probabilities are conditioned on different events.

5. (Random variables)

Suppose you roll two fair, six-sided dice. Let X and Y denote the values of the two possible outcomes.

(a) Describe the probability distribution of the random variable $Z = \min(X, Y)$. To that end, first determine the possible values of Z and then the probability that Z attains each of them:

value						
probability						

Solution.

value	1	2	3	4	5	6
prob	$\frac{11}{36} = 0.3055$	$\frac{9}{36} = 0.25$	$\frac{7}{36} = 0.1944$	$\frac{5}{36} = 0.1388$	$\frac{3}{36} = 0.0833$	$\frac{1}{36} = 0.0277$

(b) What are the expected value and variance of Z?

Solution.

$$\mathbb{E}(Z) = \frac{11}{36}\cdot 1 + \frac{9}{36}\cdot 2 + \frac{7}{36}\cdot 3 + \frac{5}{36}\cdot 4 + \frac{3}{36}\cdot 5 + \frac{1}{36}\cdot 6 = \frac{11 + 18 + 21 + 20 + 15 + 6}{36} = \frac{91}{36} = 2.5277$$

and

$$\begin{aligned}\mathrm{var}(Z) &= \mathbb{E}(Z^2) - \mathbb{E}(Z)^2\\ &= \frac{11}{36} \cdot 1^2 + \frac{9}{36} \cdot 2^2 + \frac{7}{36} \cdot 3^2 + \frac{5}{36} \cdot 4^2 + \frac{3}{36} \cdot 5^2 + \frac{1}{36} \cdot 6^2 - \left(\frac{91}{36}\right)^2\\ &= 1.97145.\end{aligned}$$

(c) Repeat (a) and (b) for the random variable $W = \max(X, Y)$.

Solution.

value	1	2	3	4	5	6
prob	$\frac{1}{36} = 0.0277$	$\frac{3}{36} = 0.0833$	$\frac{5}{36} = 0.1388$	$\frac{7}{36} = 0.1944$	$\frac{9}{36} = 0.25$	$\frac{11}{36} = 0.3055$

Hence

$$\begin{aligned}\mathbb{E}(W) &= \frac{1}{36} \cdot 1 + \frac{3}{36} \cdot 2 + \frac{5}{36} \cdot 3 + \frac{7}{36} \cdot 4 + \frac{9}{36} \cdot 5 + \frac{11}{36} \cdot 6 \\ &= \frac{161}{36} \\ &= 4.4722.\end{aligned}$$

and

$$\begin{aligned}\mathrm{var}(W) &= \mathbb{E}(W^2) - \mathbb{E}(W)^2 \\ &= \frac{1}{36} \cdot 1^2 + \frac{3}{36} \cdot 2^2 + \frac{5}{36} \cdot 3^2 + \frac{7}{36} \cdot 4^2 + \frac{9}{36} \cdot 5^2 + \frac{11}{36} \cdot 6^2 - \left(\frac{161}{36}\right)^2 \\ &= 1.97145.\end{aligned}$$

(d) Determine if each of the following claims is true or false:

(i) The random variables Z and $7 - W$ are the same, that is for every element $\omega \in \Omega$ of the sample space we have

$$Z(\omega) = 7 - W(\omega).$$

Solution. FALSE. For example, when both rolls are 1 we have

$$Z = W = 1 \neq 7 - W.$$

(ii) The random variables Z and $7 - W$ have the same probability mass function.
Solution. TRUE. That is evident from the tables in (a) and (c).

(iii) The random variables $X + Y$ and $Z + W$ are the same.
Solution. TRUE. This follows because for any two numbers $x, y \in \mathbb{R}$ we always have

$$x + y = \min(x, y) + \max(x, y).$$

(iv) The random variables $X + Y$ and $Z + W$ have the same probability mass function.
Solution. TRUE. This immediately follows from (iii).

46-880 Introduction to Probability and Statistics

Problem Set 2, mini-1 2023

Refrain from using generative artificial intelligence tools (such as ChatGPT) to complete this assignment. At this stage the use of these kinds tools will most likely be a hindrance to your learning.

You are encouraged to discuss the problems below with your classmates before you proceed with the online submission of your answers.

Please submit your answers online between 10am on Friday September 8 and 9am on Monday September 11. The online submission must be completed individually. Please do not discuss details of your online submission with other students before 9am on Monday September 11.

1. (Binomial Distribution)

The World Series of baseball is to be played by team A and team B. To that end, seven games are scheduled and the team that wins four games or more wins the series.

Suppose that team A is the better team, in the sense that the probability is 0.6 that team A will win any specific game. Assume also that the result of any game is independent of that of any other and every game is played until there is a winner (no ties).

(a) What is the probability that team A will win the series?

Solution.

Let X be the number of games won by team A. Its distribution is Binomial(0.6, 7). A wins in all cases except $X \leq 3$, so the probability they win is:

```
1 - pbinom(3, 7, 0.6)
```

```
## [1] 0.710208
```

(b) Observe that if a team wins the first four games, then no more games are needed to determine the winner. Similarly if a team wins four of the first five games no more games are needed.

What is the probability that the winner is determined in the first four games? In other words, what is the probability that one of the teams wins the first four games?

Solution.

We are interested in the joint event "either A wins the first four games, or B wins the first four games." These events are mutually exclusive, so we can just add their probabilities.

```
0.6^4 + 0.4^4
```

```
## [1] 0.1552
```

(c) What is the probability that all seven games are needed to determine the winner? In other words, what is the probability that neither team wins four or more of the first six games?

Solution.

To make matters easier, focus only on team A. The seventh game is unnecessary if A wins four or more of the first six games, but also if it wins two or less. So we want A to win exactly three of the first six games.

```
dbinom(3, 6, 0.6)
```

```
## [1] 0.27648
```

(d) Suppose that each team wins two of the first four games. What is the probability that team A will win the series?

Solution.

This is the probability that A wins two or three of the remaining three games.

```
1 - pbinom(1, 3, 0.6)
```

```
## [1] 0.648
```

2. (Expectation and Variance)

A manager supervises three experimental projects. Each has a 50% chance of producing an innovative, successful new product. The teams that are developing each project work separately, so the success of one project is independent of the success of other projects. A successful project returns 6 times the amount invested; that is, if a budget B is invested in one project, the profit would be $6B - B = 5B$ if successful, and $-B$ otherwise. The manager has $900,000 budget and three options: invest her full budget in one project, split it evenly across two projects, split it evenly across all three.

(a) What is the expected value and the standard deviation of total profits if she invests her full budget in one project?

Solution.

The profit values are $5B$ and $-B$ each with probability 1/2. Hence the expected value is $\frac{1}{2}(5B) + \frac{1}{2}(-B)$, or $2B$, that is, 1.8 million dollars.

The standard deviation is $\sqrt{\frac{1}{2}(5B - 2B)^2 + \frac{1}{2}(-B - 2B)^2}$. This comes out to $3B$, or 2.7 million dollars.

(b) What is the expected value and the standard deviation of total profits if she splits her budget evenly across two projects?

Solution.

Now there are three possible profit values depending on how many projects (2, 1, or 0) succeed: $5B = 2.5B + 2.5B$ with probability 1/4, $2B = 2.5B - 0.5B$ with probability 1/2, and $-B = -0.5B - 0.5B$ with probability 1/4. Hence the expected value is $\frac{1}{4}(5B) + \frac{1}{2}(2B) + \frac{1}{4}(-B) = 2B$, that is, 1.8 million dollars.

The standard deviation is $\sqrt{\frac{1}{4}(5B - 2B)^2 + \frac{1}{4}(-B - 2B)^2 + \frac{1}{2}(2B - 2B)^2} = \frac{3B}{\sqrt{2}}$. This is about 1.909 million dollars.

(c) What is the expected value and the standard deviation of total profits if she splits her budget evenly across all three projects?

Now the possible profit values are $5B = 3 \cdot \frac{5B}{3}$, $3B = 2 \cdot \frac{5B}{3} - \frac{B}{3}$, $B = \frac{5B}{3} - 2 \cdot \frac{B}{3}$,$-B = 3 \cdot \frac{-B}{3}$ if 3, 2, 1, or 0 projects succeed respectively, and they occur with probabilities $\frac{1}{8}, \frac{3}{8}, \frac{3}{8}, \frac{1}{8}$ respectively.

Similar calculations to the ones above show that the expected value is $2B = 1.8$ million and the standard deviation is $\frac{3B}{\sqrt{3}} = 1.5588$ million.

(d) Compare her options in terms of expected profits. Which option should the manager choose if she prefers higher expected profits?

Solution.

They all have the same expected profit.

(e) Compare her options in terms of risk, measured by the standard deviation of the profits. Which option should the manager choose if she prefers lower risk?

Solution.

The one that splits the budget across all three projects has the lowest standard deviation.

3. (Poisson Distribution)

An online shop sells video game consoles at an average rate of 2 game consoles per day.

(a) What is the expected number of game consoles that the shop sells in a 30-day period?

Solution.

Let X denote the number of game consoles that the shop sells in a 30-day period. We have $X \sim \text{Poisson}(30 \cdot 2) = \text{Poisson}(60)$. Hence, $E(X) = 60$.

(b) Find the probability that the shop sells between 8 and 17 game consoles in a 5-day period. Equivalently, find $P(8 \leq X \leq 17)$ where X = number of game consoles that the shop sells in a 5-day period.

Solution.

Let X = number of game consoles that the shop sells in a 5-day period. Then $X \sim \text{Poisson}(10)$ and thus $P(8 \leq X \leq 17) = P(X \leq 17) - P(X \leq 7)$ is

```
ppois(17,10) - ppois(7,10)
```

```
## [1] 0.7655017
```

(c) Suppose the shop sold 14 game consoles in the last 7 days. What is the probability that the shop sells 4 or fewer consoles in the next 2 days?

Solution.

Let X = number of game consoles that the shop sells in the next two days. Then $X \sim \text{Poisson}(4)$ regardless of what happened in the last 7 days. We want $P(X \leq 4)$, that is,

```
ppois(4,4)
```

```
## [1] 0.6288369
```

(d) A new console model will be released soon. Thus the shop would like to sell their entire current inventory within the next few days. What is the largest number of consoles that the shop should have in stock so that the shop will sell all of them within the next 7 days with probability 0.95 or higher?

Solution.

Let X = number of consoles that the shop sells in a 7-day period and let k = number of consoles in stock. Then $X \sim \text{Poisson}(14)$ and the probability of selling all of their stock within the next 7 days is $P(X \geq k) = 1 - P(X \leq k - 1)$. To find the largest k we do some trial and error:

```r
kvector = 1:10
probs = 1-ppois(kvector-1,14)
probs >= 0.95
```

```
## [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE
```

We see that the largest k for which $P(X \geq k) \geq 0.95$ is $k = 8$.

4. (Uniform Distribution)

The incomes of all families in a particular suburb is uniformly distributed. It is known that the median income for all families in this suburb is $70,000 and that 10% of all families in the suburb have incomes above $110,000.

(a) For a randomly chosen family, what is the probability that its income will be between $70,000 and $110,000?

Solution.

We know that 50% of the families have income above $70,000. Hence 40% = 50% - 10% of the families have income between $70,000 and $110,000. Thus the probability that the income of a randomly chosen family is between $70,000 and $110,000 is 0.4.

(b) What are the minimum and maximum incomes in the suburb?

Solution.

Since the income is uniform, it should be uniformly distributed in some interval $[a, b]$. Since the median is $70,000 we have

$$\frac{a+b}{2} = 70,000.$$

On the other hand, from part (a) we know that

$$\frac{40,000}{b-a} = 0.4$$

Thus $b - a = 100,000$ and $b + a = 140,000$. It thus follows that $a = 20,000$ and $b = 120,000$.

(c) What is the probability that a randomly chosen family has an income above $60,000?

Solution.

$$\frac{120,000 - 60,000}{100,000} = 0.6.$$

(d) Find the probability that a randomly chosen family has an income above $110,000 given that its income is above $80,000.

Solution.

Let A = income is above $\$110,000$ and B= income is above $\$80,000$. We have $P(A) = 0.1$ and $P(B) = 0.4$ and want

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} = \frac{0.1}{0.4} = 0.25.$$

5. (Poisson as limit of binomial)

The purpose of this exercise is to illustrate an interesting connection between the binomial and Poisson distributions.

Let $\lambda = 4$ and $n = 10$.

(a) Suppose $X_n \sim B(n, \lambda/n)$. Use Excel or R to compute the probability mass function of X_n. That is, compute $\mathbb{P}(X_n = x)$ for $x = 0, 1, \ldots, n$.

Solution.

```
lambda = 4
n = 10
one_to_n = 1:n
pmf_n = dbinom(one_to_n,n,lambda/n)
print(pmf_n)
```

```
##  [1] 0.0403107840 0.1209323520 0.2149908480 0.2508226560 0.2006581248
##  [6] 0.1114767360 0.0424673280 0.0106168320 0.0015728640 0.0001048576
```

(b) Suppose $X \sim \text{Pois}(\lambda)$. Use Excel or R to compute the probability mass function of X for values up to n. More precisely, compute $\mathbb{P}(X = x)$ for $x = 0, 1, \ldots, n$.

Solution.

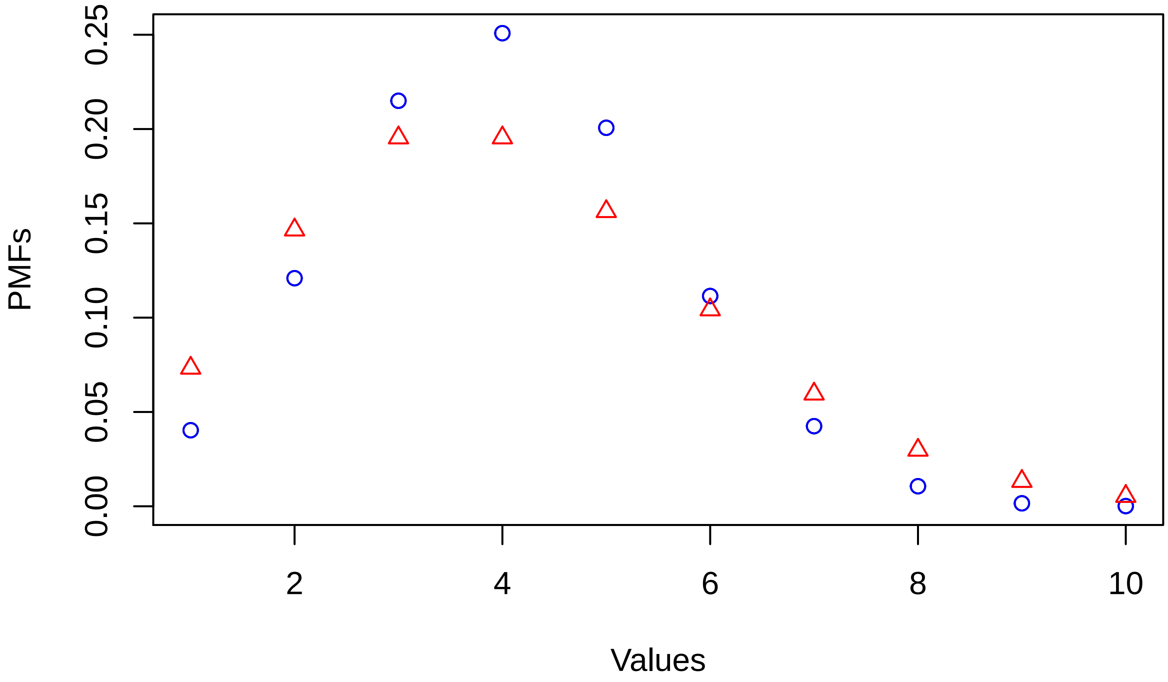
```
lambda = 4
n = 10
one_to_n = 1:n
pmf = dpois(one_to_n,lambda)
print(pmf)
```

```
##  [1] 0.073262556 0.146525111 0.195366815 0.195366815 0.156293452 0.104195635
##  [7] 0.059540363 0.029770181 0.013231192 0.005292477
```

(c) Generate scatter charts to compare the probability mass functions of X_n and of X that you computed in parts (a) and (b) above. What do you observe?

Solution.

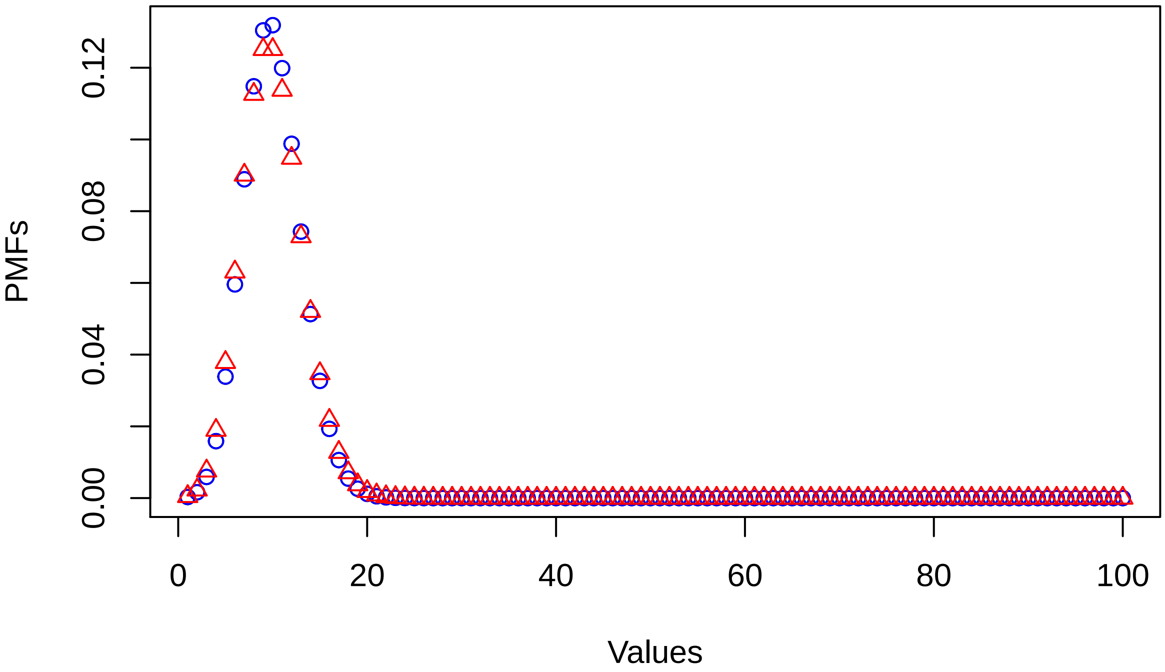
```
plot(one_to_n,pmf_n,xlab = "Values", ylab = "PMFs")
points(one_to_n,pmf_n, col = "blue", pch = 1)
points(one_to_n,pmf, col = "red", pch = 2)
```



(d) Repeat your comparison and chart visualization for other values of λ and n. For example, $n = 20, 50, 100$ and $\lambda = 2, 10, 20$.

Solution.

```
lambda = 10
n = 100
one_to_n = 1:n
pmf_n = dbinom(one_to_n,n,lambda/n)
pmf = dpois(one_to_n,lambda)
plot(one_to_n,pmf_n,xlab = "Values", ylab = "PMFs")
points(one_to_n,pmf_n, col = "blue", pch = 1)
points(one_to_n,pmf, col = "red", pch = 2)
```



As n gets larger, the two distributions become more similar.

46-880 Introduction to Probability and Statistics

Problem Set 3, mini-1 2023

Refrain from using generative artificial intelligence tools (such as ChatGPT) to complete this assignment. At this stage the use of these kinds tools will most likely be a hindrance to your learning.

You are encouraged to discuss the problems below with your classmates before you proceed with the online submission of your answers.

Please submit your answers online between 10am on Friday September 15 and 9am on Monday September 18. The online submission must be completed individually. Please do not discuss details of your online submission with other students before 9am on Monday September 18.

1. (Exponential distribution)

A small private ambulance service in Western Pennsylvania has determined that the time between emergency calls is exponentially distributed with a mean of 75 minutes. The company needs to decide how many ambulances to keep. Before doing so, the owners would like to have answers to the following questions.

(a) What is the probability that an emergency call is received within 45 minutes?

Solution.

```
lambda <- 1/75
pexp(45, rate = lambda)
```

```
## [1] 0.4511884
```

(b) What is the probability that at least two hours elapse between an emergency call and the following one?

Solution.

```
1 - pexp(120, rate = lambda)  # 120 minutes = 2 hours
```

```
## [1] 0.2018965
```

(c) The unit just received an emergency call. What is the probability that the next emergency call will arrive between 45 minutes and two hours from now?

Solution.

```
pexp(120, rate = lambda) - pexp(45, rate = lambda)
```

```
## [1] 0.3469151
```

(d) What is the distribution of the *number* of emergency calls received in 24 hours? What is the expected number of emergency calls received in 24 hours?

Solution.

```
#The distribution of the number of emergency calls received in 24 hours is Poisson.

expected_24hrs <- 1440/75
expected_24hrs
```

```
## [1] 19.2
```

2. (Poisson & Exponential)

On average, the state police catch eight speeders per hour at a certain location on the Pennsylvania Turnpike. Assume that the arrival of speeders follows a Poisson process.

(a) What is the expected time (in hours) between successive speeders?

Solution.

```
expected_time <- 1 / 8
expected_time
```

```
## [1] 0.125
```

(b) What is the rate parameter λ of speeders per hour?

Solution.

```
lambda <- 8
lambda
```

```
## [1] 8
```

(c) What is the probability that the state police wait less than 10 minutes for the next speeder?

Solution.

```
 pexp(10/60, rate = lambda)
```

```
## [1] 0.7364029
```

(d) What is the probability that the state police wait between 15 and 20 minutes for the next speeder?

Solution.

```
pexp(20/60, rate = lambda) - pexp(15/60, rate = lambda)
```

```
## [1] 0.06585183
```

(e) What is the probability that the state police wait more than 55 minutes for the next speeder?

Solution.

```
1 - pexp(55/60, rate = lambda)
```

```
## [1] 0.000653392
```

3. (Normal Distribution)

Scores on an achievement test are known to be normally distributed with a mean of 420 and a standard deviation of 80.

(a) For a randomly chosen person taking the test, what is the probability of a score more than 500?

Solution.

```
mean <- 420
sd <- 80

1 - pnorm(500, mean = mean, sd = sd)
```

```
## [1] 0.1586553
```

(b) For a randomly chosen person taking the test, what is the probability of a score between 400 and 480?

Solution.

```
pnorm(480, mean = mean, sd = sd) - pnorm(400, mean = mean, sd = sd)
```

```
## [1] 0.372079
```

(c) What is the minimum test score needed to be in the top 10% of all people taking the test?

Solution.

```
qnorm(0.90, mean = mean, sd = sd)
```

```
## [1] 522.5241
```

(d) For a randomly chosen individual, state, without doing the calculations, in which of the following ranges his score is most likely to be: 400–440, 440–480, 480–520, or 520–560?

Solution.

```
# 400--440
# As for a normally distributed variable, scores are most densely populated near the mean.
```

4. (Normal distribution and combinations of random variables)

Financial portfolio theory commonly assumes that asset returns are normally distributed. Assume the annual returns of government bonds, corporate bonds, and common stocks are normally distributed with expected value and standard deviation as indicated in the following table

	government bonds	corporate bonds	common stocks
expected return	3%	6%	11%
standard deviation	3%	8%	15%

Assume that the annual returns of these three asset classes are independent.

(a) Which one of the three asset classes is least likely to attain a negative return over the next year?

Solution. Government bonds. This follows from the calculation of P(return <= 0) for each asset class:

```
muvec = c(0.03,0.06,.11)
sigmavec = c(0.03,0.08,0.15)
print("Probs of loss for government, corporate, and common stocks:")
```

```
## [1] "Probs of loss for government, corporate, and common stocks:"
```

```
pnorm(numeric(3),muvec,sigmavec)
```

```
## [1] 0.1586553 0.2266274 0.2316776
```

(b) Suppose an investor has an annual target return of 7%. Which one of the three asset classes is most likely to attain or exceed this target return?

Solution. Common stocks. This follows from the calculation of P(return >= 0.07) for each asset class:

```
muvec = c(0.03,0.06,.11)
sigmavec = c(0.03,0.08,0.15)
print("Probs of return > 7% for government, corporate, and common stocks:")
```

```
## [1] "Probs of return > 7% for government, corporate, and common stocks:"
```

```
1-pnorm(0.07*c(1,1,1),muvec,sigmavec)
```

```
## [1] 0.09121122 0.45026178 0.60513709
```

(c) What is the probability that the annual return of corporate bonds is higher than the annual return of government bonds?

Solution. Consider the random variable $Y - X$ where X = return of government bonds, Y = return of corporate bonds. Since X and Y are independent, the difference $Y - X$ is normally distributed with expected value $0.06 - 0.03 = 0.03$ and variance $(0.03)^2 + (0.08)^2 = 0.0073$.

Therefore the probability that the annual return of corporate bonds is higher than the annual return of government bonds is P(Y-X > 0):

```
print(paste("Prob return of corporate bonds > return of government bonds =",1-pnorm(0,0.03,0.0854)))
```

```
## [1] "Prob return of corporate bonds > return of government bonds = 0.637313872768523"
```

(d) Suppose an asset management firm offers three investment plans (“mutual funds’’). Each of the plans is a portfolio of the three asset classes with allocations as described in the following table

	government bonds	corporate bonds	common stocks
plan A	40%	40%	20%
plan B	20%	30%	50%
plan C	10%	20%	70%

For each of the above three plans, determine the expected return and standard deviation. (You are strongly advised to use a spreadsheet or vector operations for these calculations.)

Solution. The variance calculations use the independence of the three variables:

```
planA = c(0.4,0.4,0.2)
planB = c(0.2,0.3,0.5)
planC = c(0.1,0.2,0.7)
muplanA = sum(planA*muvec); varplanA = sum((planA^2)*(sigmavec^2)); sigmaplanA = varplanA^(0.5)
muplanB = sum(planB*muvec); varplanB = sum((planB^2)*(sigmavec^2)); sigmaplanB = varplanB^(0.5)
muplanC = sum(planC*muvec); varplanC = sum((planC^2)*(sigmavec^2)); sigmaplanC = varplanC^(0.5)
```

	expected return	standard deviation
plan A	5.8%	4.548%
plan B	7.9%	7.897%
plan C	9.2%	10.625%

(e) Suppose an investor has an annual target return of 3%. Among above three plans, which one is most likely to attain or exceed this target return?

Solution. Plan B. This follows from the calculation of P(return >= 0.07) for each plan:

```
muplans = c(muplanA,muplanB,muplanC)
sigmaplans = c(sigmaplanA,sigmaplanB,sigmaplanC)
print("Probs of return > 3% for plan A, plan B, plan C:")
```

```
## [1] "Probs of return > 3% for plan A, plan B, plan C:"
```

```
1-pnorm(0.03*c(1,1,1),muplans,sigmaplans)
```

```
## [1] 0.7309601 0.7325199 0.7202234
```

46-880 Introduction to Probability and Statistics

Problem Set 4 Solutions, mini-1 2023

You are encouraged to discuss the problems below with your classmates before you proceed with the online submission of your answers.

Please submit your answers online between 10am on Friday September 23 and 9am on Monday September 26. The online submission must be completed individually. Please do not discuss details of your online submission with other students before 9am on Monday September 26.

1. (Law of Large Numbers and Central Limit Theorem)

The goal of this exercise is to illustrate both the Law of Large Numbers and Central Limit Theorem. Let $n = 100$ and $p = 0.5$.

(a) Suppose $X \sim B(n, p)$ and $Y = X/n$, that is, X has binomial distribution with n trials and probability of success p and Y is the sample mean of n Bernoulli trials. Use the binomial distribution and the fact that $X = nY$ to compute $P(|Y - p| \leq 0.01)$ which is the same as $P(|X - np| \leq 0.01n)$.

Solution.

Since $X \sim B(n, p)$ we have $P(|X - np| \leq 0.01n) = P(np - 0.01n \leq X - np \leq np + 0.01n) = P(X - np \leq np + 0.01n) - P(X \leq np - 0.01n - 1)$, which can be computed as follows

```
n=100
p=0.5
prob_exact = pbinom(n*p+0.01*n,n,p)-pbinom(n*p-0.01*n-1,n,p)
print(prob_exact)
```

```
## [1] 0.2356466
```

(b) Compute the mean μ and variance σ^2 of Y.

Solution.

Since $X \sim B(n, p)$ we have $E(X) = np$, $\text{var}(X) = np(1-p)$ and thus $\mu = E(X)/n = p$ and $\sigma^2 = \text{var}(X)/n^2 = p(1-p)/n$.

```
mu = p
sigma = sqrt(p*(1-p)/n)
print(paste("mu =",mu))
```

```
## [1] "mu = 0.5"
```

```
print(paste("sigma =",sigma))
```

```
## [1] "sigma = 0.05"
```

(c) Suppose $W \sim N(\mu, \sigma^2)$ for the μ, σ^2 you found in part (b). Compute $P(|W - p| \leq 0.01)$ and compare with the value $P(|Y - p| \leq 0.01)$ that you obtained in part (a).

Solution.

Since $W \sim N(\mu, \sigma^2)$ we have $P(|W - p| \leq 0.01) = P(W \leq p + 0.01) - P(W \leq p - 0.01)$ which can be computed as follows

```
mu = p
sigma = sqrt(p*(1-p)/n)
prob_norm = pnorm(p+0.01,mu,sigma)-pnorm(p-0.01,mu,sigma)
print(prob_norm)
```

```
## [1] 0.1585194
```

We can see that $P(|W-p| \leq 0.01) = 0.16$ underestimates $P(|Y-p| \leq 0.01) = 0.24$

(d) Repeat Repeat (a), (b) and (c) for other values of n and p, for instance $n = 1000, 10000$ and $p = 0.3, 0.7$. What do you observe?

Solution. As n gets larger the two probabilities gets closer to each other and also get closer to 1.

```
n=10000
p=0.3
delta = 0.01
prob_exact = pbinom(n*p+delta*n,n,p)-pbinom(n*p-delta*n-1,n,p)
print(prob_exact)
```

```
## [1] 0.9717054
```

```
mu = p
sigma = sqrt(p*(1-p)/n)
prob_norm = pnorm(p+delta,mu,sigma)-pnorm(p-delta,mu,sigma)
print(prob_norm)
```

```
## [1] 0.9709037
```

(e) Repeat (a), (b) and (c) but to compare $P(|Y-p| \leq 0.1)$ with $P(|W-p| \leq 0.1)$ and to compare $P(|Y-p| \leq 0.2)$ with $P(|W-p| \leq 0.2)$. What do you observe?

```
n=100
p=0.5
prob_exact = pbinom(n*p+0.1*n,n,p)-pbinom(n*p-0.1*n-1,n,p)
print(prob_exact)
```

```
## [1] 0.9647998
```

```
mu = p
sigma = sqrt(p*(1-p)/n)
prob_norm = pnorm(p+0.1,mu,sigma)-pnorm(p-0.1,mu,sigma)
print(prob_norm)
```

```
## [1] 0.9544997
```

```
n=100
p=0.5
prob_exact = pbinom(n*p+0.2*n,n,p)-pbinom(n*p-0.2*n-1,n,p)
print(prob_exact)
```

```
## [1] 0.9999678
```

```
mu = p
sigma = sqrt(p*(1-p)/n)
prob_norm = pnorm(p+0.2,mu,sigma)-pnorm(p-0.2,mu,sigma)
print(prob_norm)
```

```
## [1] 0.9999367
```

The two probabilities are much closer to each other and to one.

2. (Moment generating functions)

Recall that the moment generating function (mgf) of a random variable X is the function ψ_X defined via $\psi_X(t) = \mathbb{E}(e^{tX})$.

(a) Suppose $Y = aX + b$ for some constants a, b and suppose $\psi_X(t)$ exists for all $t \in \mathbb{R}$. Write down ψ_Y in terms of ψ_X. For example, is it true that $\psi_Y(t) = a\psi_X(t) + b$?

Solution.

$\psi_Y(t) = \mathbb{E}(e^{tY}) = \mathbb{E}(e^{t(aX+b}) = \mathbb{E}(e^{tb} \cdot e^{(ta)X}) = e^{tb} \cdot \mathbb{E}(e^{(ta)X}) = e^{tb}\psi_X(at)$.

(b) Suppose $X \sim U(0,1)$, that is, X is uniformly distributed on the interval $[0,1]$. Show that $\psi_X(t) = (e^t - 1)/t$ for $t \neq 0$ and $\psi_X(0) = 1$.

Solution.

For $t = 0$ we have $\psi_X(t) = \mathbb{E}(e^0) = \mathbb{E}(1) = 1$.

For $t \neq 0$ we have $\psi_X(t) = \mathbb{E}(e^{tX}) = \int_0^1 e^{tx} dx = \frac{e^{tx}}{t}|_{x=0}^{x=1} = \frac{e^t - 1}{t}$.

(c) We say that a continuous random variable has *triangular* distribution with lower limit 0, upper limit 2 and mode 1 if it has the following density function

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 2 - x & \text{if } 1 \leq x \leq 2 \\ 0 & \text{if } 2 < x \end{cases}$$

Observe that the graph of this density is a triangle. (The triangular distribution is similarly defined for lower limit ℓ, upper limit u, and mode m for any constants $\ell < m < u$.)

Show that the moment generating function of a random variable X with the above triangular distribution is $\psi_X(t) = (e^t - 1)^2/t^2$ for $t \neq 0$ and $\psi_X(0) = 1$.

Solution.

For $t = 0$ we have $\psi_X(t) = \mathbb{E}(e^0) = \mathbb{E}(1) = 1$.

For $t \neq 0$ we have

$$\begin{aligned}
\psi_X(t) = \mathbb{E}(e^{tX}) &= \int_0^1 xe^{tx}dx + \int_1^2 (2-x)e^{tx}dx \\
&= \left[\frac{xe^{tx}}{t} - \frac{e^{tx}}{t^2}\right]_{x=0}^{x=1} + 2\frac{e^{tx}}{t}|_{x=1}^{x=2} - \left[\frac{xe^{tx}}{t} - \frac{e^{tx}}{t^2}\right]_{x=1}^{x=2} \\
&= \left(\frac{e^t}{t} - \frac{e^t}{t^2} + \frac{1}{t^2}\right) + 2\frac{e^{2t} - e^t}{t} - \left(\frac{2e^2t}{t} - \frac{e^2t}{t^2} - \frac{e^t}{t} + \frac{e^t}{t^2}\right) \\
&= \frac{e^{2t} - 2e^t + 1}{t^2} = \frac{(e^t - 1)^2}{t^2}
\end{aligned}$$

(d) Use **(b)** and **(c)** to prove that the sum of two independent $U(0,1)$ variables has triangular distribution with lower limit 0, upper limit 2 and mode 1.

Solution

Since X, Y are independent and $U(0,1)$ from part (b) it follows that the mgf of $X + Y$ is

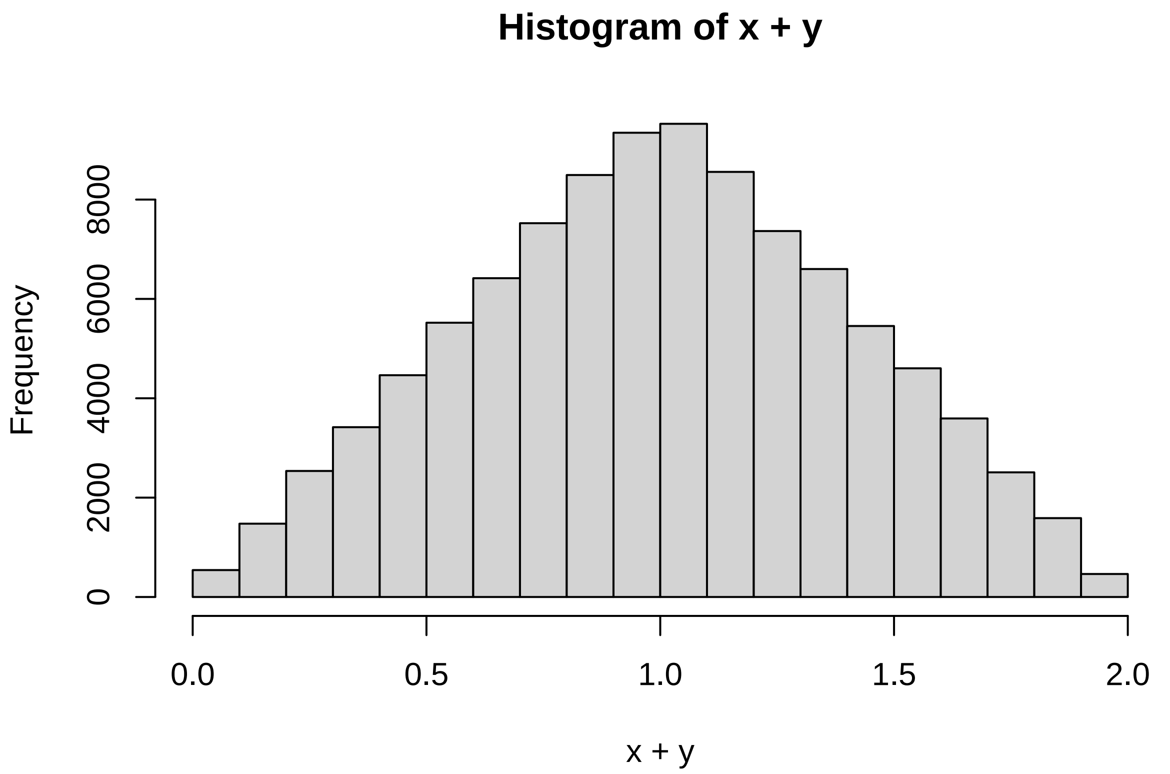
$$\psi_{X+Y}(t) = \psi_X(t) \cdot \psi_Y(t) = \frac{(e^t - 1)^2}{t^2}.$$

This is the mgf of the triangular distribution in part (c) and thus $X + Y$ has that distribution.

(e) Write some R code to confirm that **(d)** indeed holds by generating histograms of 100000 random draws of the sum of two independent $U(0,1)$ variables.

Solution

```
x = runif(100000,0,1)
y = runif(100000,0,1)
hist(x+y)
```

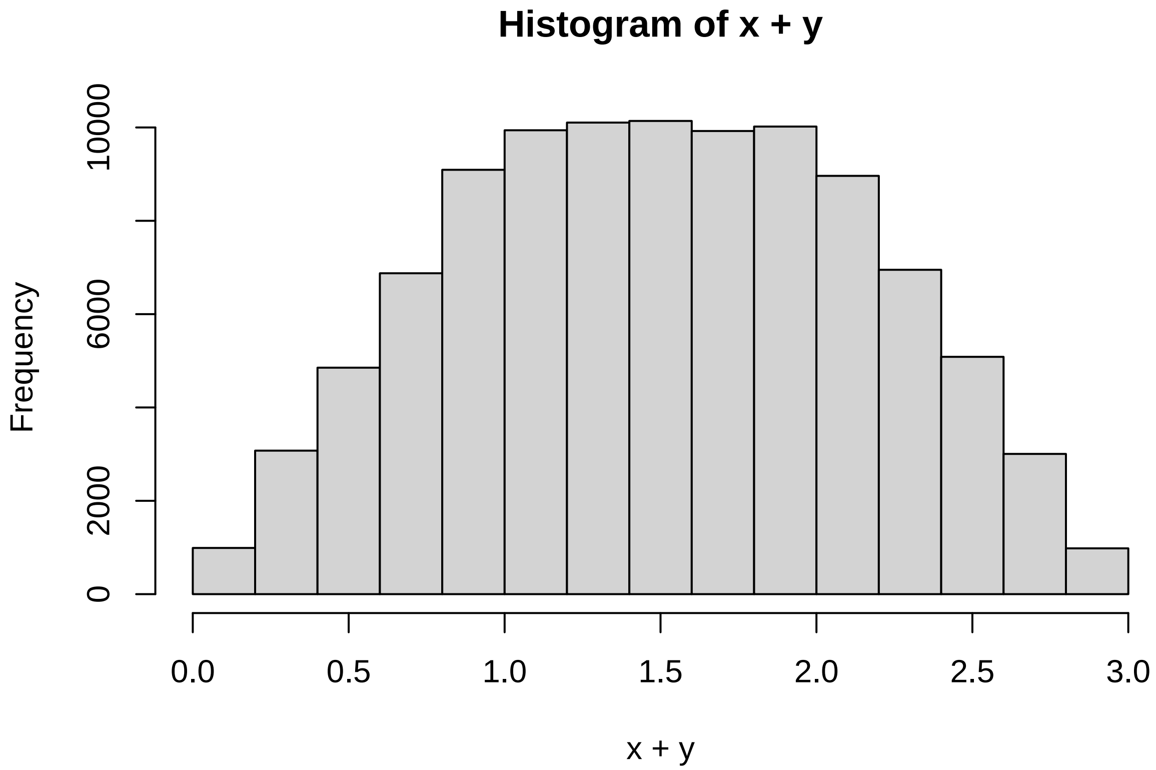


The histogram indeed looks triangular.

(f) Modify your R code in part (e) to check if the following conjecture is true: if $X \sim U(0,1)$ and $Y \sim U(0,2)$ are independent then $X+Y$ has triangular distribution (with suitable lower limit, upper limit, and mode).

Solution

```
x = runif(100000,0,1)
y = runif(100000,0,2)
hist(x+y)
```

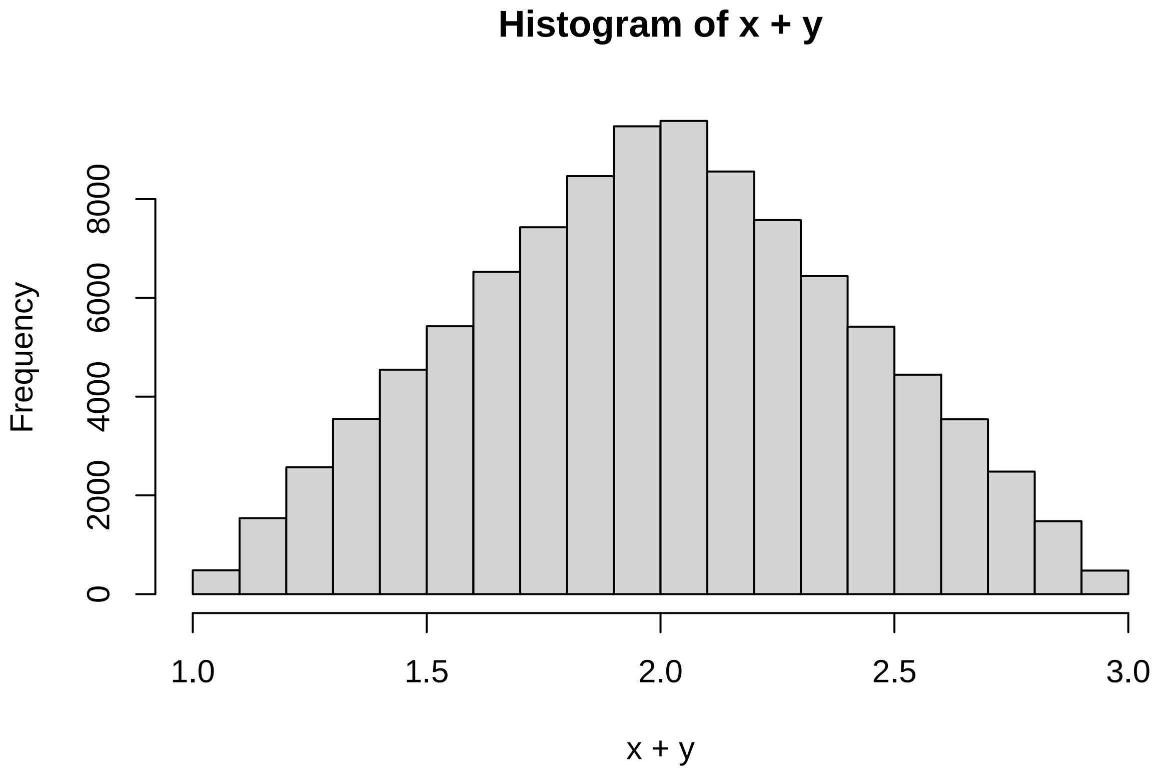


This second histogram does not look triangular. Thus the conjecture is false.

(g) Use any of the above parts to determine if the following conjecture is true: if $X \sim U(0,1)$ and $Y \sim U(1,2)$ are independent then $X + Y$ has triangular distribution (with suitable lower limit, upper limit, and mode).

Solution

```
x = runif(100000,0,1)
y = runif(100000,1,2)
hist(x+y)
```



This third histogram looks triangular. That suggests that the conjecture is true. We next prove that via the mgf.

Observe that $Y = W + 1$ where $W \sim U(0, 1)$. Thus part (a) implies that $\psi_Y(t) = e^t \psi_W(t)$ with both $X, W \sim U(0, 1)$ and independent. Consequently $\psi_{X+Y}(t) = \psi_X(t)\psi_Y(t) = e^t\psi_X(t)\psi_W(t) = e^t\psi_{X+W}(t)$. By part (a) again the latter mgf is the mgf of $X + W + 1$ which is evidently triangular with lower limit $= 1$, upper limit $= 3$ and mode $= 2$.

3. (Sample Mean)

A manufacturer designed an electronic component aimed to have an electrical resistance of 60 ohm. The production process introduces some variation in the actual resistance of the components: The population resistance is normally distributed with mean $\mu = 60$ ohm and standard deviation $\sigma = 3$ ohm. A client who recently purchased a batch of these electrical components tests 16 of them for their resistance.

(a) What is the probability that the sample mean resistance will be less than 59 ohm?

Solution: Let X be the random variable denoting the actual electrical resistance of a component. $X \sim N(\mu = 60, \sigma^2 = 3^2)$. Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, denoting the sample mean. According to the Central Limit Theorem:

$$\bar{X} \sim N(\mu, \frac{1}{n}\sigma^2)$$

```
n=16
mu=60
sigma = 3
sample_mean_sd = 1/sqrt(n)*sigma
pnorm(59,mu,sample_mean_sd)
```

```
## [1] 0.09121122
```

(b) What is the probability that the sample mean resistance differs from the population mean by more than 1 ohm?

Solution:

$$\begin{aligned} P(|\bar{X}-\mu|>1) &= P(\bar{X}-\mu \le -1) + P(\bar{X}-\mu \ge 1) \\ &= P(\bar{X}-\mu \le -1) + \big(1 - P(\bar{X}-\mu \le 1)\big) \\ &= P(\bar{X} \le \mu - 1) + 1 - P(\bar{X} \le \mu + 1) \\ &= P(\bar{X} \le 59) + 1 - P(\bar{X} \le 61) \end{aligned}$$

This probability is calculated as

```
pnorm(mu-1,mu,sample_mean_sd)+1-pnorm(mu+1,mu,sample_mean_sd)
```

```
## [1] 0.1824224
```

(c) Find and interpret a 95% acceptance interval for the sample mean resistance.

Solution:

```
q=1.96 # Here we use the convention that 95% confidence interval is [mean - 1.96*sd, mean+ 1.96*sd]
lower_bound = mu - q*sample_mean_sd
upper_bound = mu + q*sample_mean_sd
sprintf("The 95%% acceptance interval interval for the sample mean resistance is [%.3f,%.3f].", lower_bo
```

```
## [1] "The 95% acceptance interval interval for the sample mean resistance is [58.530,61.470]."
```

(d) Find the resistance r (in ohm) such that the sample mean resistance is less than r with probability 0.2.

Solution:

$$P(\bar{X} \le r) = 0.2$$

The value of r is

```
qnorm(0.2, mu,sample_mean_sd)
```

```
## [1] 59.36878
```

(e) The client takes a second (independent) random sample of n components for some $n > 16$. Without doing any calculations, state how your answers in parts (a), (b), (c) and (d) would change for the second sample.

Solution: When n increases, the sample mean, $\bar{X}$, will follow a normal distribution with the same mean and a smaller variance. In other words, the sample mean is more likely to fall close to the theoretical mean. Therefore,

1. the probability in part (a) will decrease.
2. the probability in part (b) will decrease.
3. The 95% acceptance interval in part (c) will be narrower
4. The threshold in part (d) will increase.

4. (Confidence Interval for Population Mean with Known Variance)

A college admission officer for an MBA program has determined that historically applicants have undergraduate grade point averages that are distributed with standard deviation 0.45. From a random sample of 25 applications from the current year, the sample mean grade point average is 2.90.

(a) Calculate the standard error of the sample mean.

Solution. The standard error of the mean is given by:

$$\frac{\sigma}{\sqrt{n}} = \frac{0.45}{\sqrt{25}} = 0.09$$

```
sigma = 0.45
n = 25
se = sigma/n^0.5
print(paste("se = ",se))
```

```
## [1] "se =  0.09"
```

(b) Calculate the margin of error of 90% and 99% confidence intervals for the population mean.

Solution. The margin of error for the $(1-\alpha)$-confidence interval is

$$z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

$z_{\alpha/2} = 1 - \alpha/2$ quantile of the standard normal.

```
print(paste("MOE for 90% CI",qnorm(.95)*se))
```

```
## [1] "MOE for 90% CI 0.148036826425632"
```

```
print(paste("MOE for 99% CI",qnorm(.995)*se))
```

```
## [1] "MOE for 99% CI 0.231824637319401"
```

(c) Find a 95% confidence interval for the population mean.

Solution. The $(1-\alpha)$-confidence interval for the population mean is given by:

$$\bar{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

where $z_{\alpha/2} = 1 - \alpha/2$ quantile of the standard normal. Since we have $\bar{X} = 2.9, \sigma = 0.45$, and $\alpha = 0.05$ we get

```
z = qnorm(.975)
barX = 2.9
sigma = 0.45
n = 25
LCL = barX-z*sigma/n^0.5
UCL = barX+z*sigma/n^0.5
print(paste("95% confidence interval = [",LCL,UCL,"]"))
```

```
## [1] "95% confidence interval = [ 2.7236032413914 3.0763967586086 ]"
```

(d) Based on the sample results, a statistician computes for the population mean a confidence interval extending from 2.81 to 2.99. Find the confidence level associated with this interval.

Solution.

The margin of error of the confidence interval is half its width. Thus

$$z_{\alpha/2} \cdot \frac{0.45}{\sqrt{25}} = \frac{2.99 - 2.81}{2} = 0.09 \Rightarrow z_{\alpha/2} = 1$$

Therefore the confidence $1 - \alpha$ can be computed as follows

```
alpha_over_2 = 1-pnorm(1)
print(paste("Confidence level = ",1-2*alpha_over_2))
```

```
## [1] "Confidence level =  0.682689492137086"
```

(e) Suppose the same sample mean had been obtained but with a larger sample of applications. State, without calculation, if the 99% confidence interval for the population mean would be wider or narrower than that in part (b).

Solution. The confidence interval would be narrower than that in part (b). The width of the $(1-\alpha)$-confidence interval is twice its margin of error

$$2 \cdot z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Thus the width of the confidence interval decreases as the sample size increases. In other words, as the sample size increases, the confidence interval becomes narrower.

46-880 Introduction to Probability and Statistics, Problem Set 5

Refrain from using generative artificial intelligence tools (such as ChatGPT) to complete this assignment. At this stage the use of these kinds tools will most likely be a hindrance to your learning.

You are encouraged to discuss the problems below with your classmates before you proceed with the online submission of your answers.

Please submit your answers online between 10am on Friday September 29 and 9am on Monday October 2. The online submission must be completed individually. Please do not discuss details of your online submission with other students before 9am on Monday October 2.

1. (Confidence Interval for Population Mean with Known Variance)

   A college admission officer for an MBA program has determined that historically applicants have undergraduate grade point averages that are distributed with standard deviation 0.45. From a random sample of 25 applications from the current year, the sample mean grade point average is 2.90.

   (a) Calculate the standard error of the mean.

   (b) Calculate the margin of error of 90% and 99% confidence intervals for the population mean.

   (c) Find a 95% confidence interval for the population mean.

   (d) Based on the sample results, a statistician computes for the population mean a confidence interval extending from 2.81 to 2.99. Find the confidence level associated with this interval.

   (e) Suppose the same sample mean had been obtained but with a larger sample of applications. State, without calculation, if the 99% confidence interval for the population mean would be wider or narrower than that in part (b).

2. (Confidence Interval and Hypothesis Testing for Sample Proportion)

   The president of a large country is contemplating proposing a change to the constitution so that he can rule for one more term. In a random sample of 100 citizens of that country, 56 expressed opposition to this proposal.

   (a) Calculate the margin of error of a 95% confidence interval for the population proportion of citizens opposed to the proposed change.

   (b) Find a 90% confidence interval for the proportion of all citizens opposed to the proposed change.

   (c) Suppose we want to find a 90% confidence interval for the population proportion with margin of error at most 0.04. How large a sample is needed?

   Hint: the largest value of $\hat{p}(1-\hat{p})$ *for* $\hat{p} \in [0,1]$ *is attained at* $\hat{p} = 0.5$.

   (d) Test the null hypothesis that one-half of all citizens oppose this proposal at the 5% significance level.

3. (Confidence Intervals and Hypothesis Testing with Unknown Population Variance)

The cost of capital for companies is determined by the market and represents the degree of perceived risk by investors. A recent report claims that the average cost of capital for big companies in the US is 7.30%. An analyst wants to test whether this claim is valid or not. She collects a sample of 42 randomly selected firms, which is given in the Excel file `MSBAProblemSet5.xlsx`.

(a) Use the data in the file `MSBAProblemSet5.xlsx` to compute the sample mean, sample variance, and sample standard deviation.

(b) Calculate the margin of error of a 95% confidence interval for the population mean.

(c) Test the hypothesis that the average cost of capital is 7.30%. Use a t-test and $\alpha = 0.05$. Clearly state your null hypothesis, the test statistic, critical values, decision rule and your conclusion.

(d) Does the conclusion in part (c) change if you use instead $\alpha = 0.01$?

(e) Find the p-value corresponding to the t-test used in part (c). Can you find any relationship between your results in parts (c) and (d)?

46-880 Introduction to Probability and Statistics

Solution to Problem Set 5, mini-1 2023

1. (Confidence Interval for Population Mean with Known Variance)

(a)

```r
sigma <- 0.45
n <- 25
SE <- sigma / sqrt(n)
SE
```

```
## [1] 0.09
```

(b)

```r
Z_90 <- qnorm(0.95)  # For two-tailed 90% confidence, you'd use the 95th percentile of the standard nor
Z_99 <- qnorm(0.995) # For two-tailed 99% confidence, you'd use the 99.5th percentile

ME_90 <- Z_90 * SE
ME_99 <- Z_99 * SE

ME_90
```

```
## [1] 0.1480368
```

```r
ME_99
```

```
## [1] 0.2318246
```

(c)

```r
Z_95 <- qnorm(0.975) # For two-tailed 95% confidence
ME_95 <- Z_95 * SE
CI_lower_95 <- 2.90 - ME_95
CI_upper_95 <- 2.90 + ME_95

CI_lower_95
```

```
## [1] 2.723603
```

```r
CI_upper_95
```

```
## [1] 3.076397
```

(d)

```r
observed_ME <- (2.99 - 2.81) / 2
observed_Z <- observed_ME / SE
# Use the pnorm function to get the probability for the observed Z-value
confidence_prob <- pnorm(observed_Z)
confidence_level <- (confidence_prob - 0.5) * 2  # Multiply by 2 for two-tailed
```

```
confidence_level
```

```
## [1] 0.6826895
```

(e)

```
#The 99% confidence interval would be narrower with a larger sample size.
```

2. (Confidence Interval and Hypothesis Testing for Sample Proportion)

(a)

```
n <- 100
p_hat <- 56/n
Z_95 <- qnorm(0.975) # Z value for 95% confidence (two-tailed)
ME_95 <- Z_95 * sqrt(p_hat * (1-p_hat) / n)
ME_95
```

```
## [1] 0.09729005
```

(b)

```
Z_90 <- qnorm(0.95) # Z value for 90% confidence (two-tailed)
ME_90 <- Z_90 * sqrt(p_hat * (1-p_hat) / n)
CI_lower_90 <- p_hat - ME_90
CI_upper_90 <- p_hat + ME_90

CI_lower_90
```

```
## [1] 0.4783516
```

```
CI_upper_90
```

```
## [1] 0.6416484
```

(c)

```
ME_target <- 0.04
p <- 0.5
n_required <- (Z_90 * sqrt(p * (1-p)) / ME_target)^2
ceiling(n_required) # to ensure a whole number for sample size
```

```
## [1] 423
```

(d)

```
p_0 <- 0.5
test_statistic <- (p_hat - p_0) / sqrt(p_0 * (1-p_0) / n)
p_value <- 2 * (1 - pnorm(abs(test_statistic)))

p_value
```

```
## [1] 0.2301393
```

We can’t reject the null hypothesis.

3. (Confidence Intervals and Hypothesis Testing with Unknown Population Variance)

(a)

```
#load your data

# Read the data from the Excel file
library(readxl)
data = read_excel("MSBAProblemSet5.xlsx",sheet="costofcapital")

# Compute sample statistics
sample_mean <- mean(data$`Cost of capital in US$`)
sample_variance <- var(data$`Cost of capital in US$`)
sample_std_dev <- sd(data$`Cost of capital in US$`)

sample_mean
```

```
## [1] 0.0801619
```

```
sample_variance
```

```
## [1] 0.0003262473
```

```
sample_std_dev
```

```
## [1] 0.01806232
```

(b)

```
n <- length(data$`Cost of capital in US$`)
alpha <- 0.05
t_value <- qt(1 - alpha/2, df = n-1)

margin_of_error <- t_value * (sample_std_dev / sqrt(n))
margin_of_error
```

```
## [1] 0.005628614
```

(c)

```
# Using t-test
population_mean <- 0.073
t_statistic <- (sample_mean - population_mean) / (sample_std_dev / sqrt(n))
critical_value <- qt(1 - alpha/2, df = n-1)

# Decision rule
if (abs(t_statistic) > critical_value) {
    decision <- "Reject H0"
} else {
    decision <- "Fail to reject H0"
}

t_statistic
```

```
## [1] 2.569684
```

```
critical_value
```

```
## [1] 2.019541
```

```
decision
```

```
## [1] "Reject H0"
```

(d)

```
alpha_new <- 0.01
critical_value_new <- qt(1 - alpha_new/2, df = n-1)

# Decision rule
if (abs(t_statistic) > critical_value_new) {
    decision_new <- "Reject H0"
} else {
    decision_new <- "Fail to reject H0"
}

critical_value_new
```

```
## [1] 2.701181
```

```
decision_new
```

```
## [1] "Fail to reject H0"
```

```
1
```

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

(e)

```
p_value <- 2 * (1 - pt(abs(t_statistic), df = n-1))
p_value
```

```
## [1] 0.01391233
```

The p-value provides a continuous measure of the evidence against a null hypothesis. In this case, p-value suggests that there's evidence against the null hypothesis at the 0.05 level, but not strong enough evidence at the 0.01 level.

46-880 Introduction to Probability and Statistics

Practice Set 1 Solutions, mini-1 2023

1. (Sample Spaces & Probability Rules)

Consider the following experiment: you roll a fair, six-sided die, and toss two fair coins.

(a) Describe the sample space for this experiment. How many elements are there in the sample space?

Solution.

Let X be the outcome of the die. The possible set of values for X is {1,2,3,4,5,6}. Similarly, let Y and Z be the outcome of the first and the second coin, respectively. The possible set of values for each of them is {H,T}. Hence the sample space for the experiment of rolling a die and tossing two coins is:

$$\begin{aligned}\{(i,j,k) : i = 1,\ldots,6,\ j,k = H,T\} \quad = \quad & \{(1,H,H),(1,H,T),(1,T,H),(1,T,T),\\ & (2,H,H),(2,H,T),(2,T,H),(2,T,T),\\ & (3,H,H),(3,H,T),(3,T,H),(3,T,T),\\ & (4,H,H),(4,H,T),(4,T,H),(4,T,T),\\ & (5,H,H),(5,H,T),(5,T,H),(5,T,T),\\ & (6,H,H),(6,H,T),(6,T,H),(6,T,T)\}.\end{aligned}$$

```
n_X = 6 # number of possible values for X
n_Y = 2 # number of possible values for Y
n_Z = 2 # number of possible values for Z
n = n_X * n_Y * n_Z
print(paste0("The sample space has ", n, " elements."))
```

```
## [1] "The sample space has 24 elements."
```

(b) What is the probability that you roll an odd number and two tails come up?

Solution.

An odd number and two tails come up for the triples (1,T,T), (3,T,T), (5,T,T). Since the die and the coins are fair, each of the 24 triples is equally likely.

```
n_b = 3 # number of of possible triples in the event
p_b = n_b / n
print(paste0("P(roll an odd number and two tails come up) = ", p_b))
```

```
## [1] "P(roll an odd number and two tails come up) = 0.125"
```

(c) What is the probability that at most one tail comes up?

Solution.

It is easier to compute the probability of the complement event. More than one tail come up for the triples (1,T,T), (2,T,T), (3,T,T), (4,T,T), (5,T,T), (6,T,T).

```r
n_c_complement = 6 # number of of possible triples in the complement event
p_c_complement = n_c_complement / n # probability of the complement event
p_c = 1 - p_c_complement
print(paste0("P(at most one tail comes up) = ", p_c))
```

```
## [1] "P(at most one tail comes up) = 0.75"
```

(d) What is the probability that the roll of die is larger than the number of heads that come up?

Solution.

It is easier to compute the probability of the complement event. The roll of die is not larger than the number of heads that come up for the triples (1,H,H), (1,H,T), (1,T,H), (2,H,H).

```r
n_d_complement = 4 # number of of possible triples in the complement event
p_d_complement = n_d_complement / n # probability of the complement event
p_d = 1 - p_d_complement
print(paste0("P(the roll of die is larger than the number of heads that come up) = ", p_d))
```

```
## [1] "P(the roll of die is larger than the number of heads that come up) = 0.833333333333333"
```

2. (Conditional Probability)

Consider the following experiment: a fair coin is tossed four times.

(a) Describe the sample space for this experiment. How many elements are there in the sample space?

Solution.

Let X, Y, Z, and W be the outcome of the first, second, third, and fourth toss, respectively. The possible set of values for each of them is {H,T}. Hence the sample space for the experiment of tossing a coin four times is:

$$
\begin{aligned}
\{(i,j,k,l) : i,j,k,l = H,T\} \quad = \quad & \{(H,H,H,H),(H,H,H,T),(H,H,T,H),(H,H,T,T),\\
& (H,T,H,H),(H,T,H,T),(H,T,T,H),(H,T,T,T),\\
& (T,H,H,H),(T,H,H,T),(T,H,T,H),(T,H,T,T),\\
& (T,T,H,H),(T,T,H,T),(T,T,T,H),(T,T,T,T)\}.
\end{aligned}
$$

```r
n_X = 2 # number of possible values for X
n_Y = 2 # number of possible values for Y
n_Z = 2 # number of possible values for Z
n_W = 2 # number of possible values for W
n = n_X * n_Y * n_Z * n_W
print(paste0("The sample space has ", n, " elements."))
```

```
## [1] "The sample space has 16 elements."
```

(b) What is the probability that at most three heads occur in the four tosses given that the first two tosses were heads?

Solution.

Let B = the first two tosses are heads, and A = at most three heads occur in the four tosses. The first two tosses are heads for the tuples (H,H,H,H), (H,H,H,T), (H,H,T,H), (H,H,T,T). Since the coin is fair, each of the 16 triples is equally likely.

```
n_B = 4 # number of elements in B
p_B = n_B / n
print(paste0("P(B) = ", p_B))
```

```
## [1] "P(B) = 0.25"
```

Similarly, at most three heads occur in the four tosses for all the tuples except for (H,H,H,H). The tuples (H,H,H,T), (H,H,T,H), (H,H,T,T) are common among A and B.

```
n_AandB = 3 # number of elements in the intersection of A and B
p_AandB = n_AandB / n
print(paste0("P(A and B) = ", p_AandB))
```

```
## [1] "P(A and B) = 0.1875"
```

We need $P(A|B) = \frac{P(A \cap B)}{P(B)}$.

```
p_AgivenB = p_AandB / p_B
print(paste0("P(A|B) = ", p_AgivenB))
```

```
## [1] "P(A|B) = 0.75"
```

(c) What is the probability that exactly three tails occur in the four tosses given that the first two tosses were tails?

Solution.

Let D = the first two tosses are tails, and C = exactly three tails occur in the four tosses. The first two tosses are tails for the tuples (T,T,H,H), (T,T,H,T), (T,T,T,H), (T,T,T,T). Since the coin is fair, each of the 16 tuples is equally likely.

```
n_D = 4 # number of elements in D
p_D = n_D / n
print(paste0("P(D) = ", p_D))
```

```
## [1] "P(D) = 0.25"
```

Similarly, exactly three tails occur in the four tosses for the tuples (T,T,T,H), (T,T,H,T), (T,H,T,T), (H,T,T,T). The tuples (T,T,H,T) and (T,T,T,H) are common among C and D.

```
n_CandD = 2 # number of elements in the intersection of C and D
p_CandD = n_CandD / n
print(paste0("P(C and D) = ", p_CandD))
```

```
## [1] "P(C and D) = 0.125"
```

We need $P(C|D) = \frac{P(C \cap D)}{P(D)}$.

```
p_CgivenD = p_CandD / p_D
print(paste0("P(C|D) = ", p_CgivenD))
```

```
## [1] "P(C|D) = 0.5"
```

(d) What is the probability that at least three heads occur in the four tosses given that the first toss was a head?

Solution.

Let F = the first toss is a head, and E = at least three heads occur in the four tosses. The first toss is a head for half of the tuples. Since the coin is fair, each of the 16 tuples is equally likely.

```
n_F = 8 # number of elements in F
p_F = n_F / n
print(paste0("P(F) = ", p_F))
```

```
## [1] "P(F) = 0.5"
```

Similarly, at least three heads occur in the four tosses for the tuples (H,H,H,H), (H,H,H,T), (H,H,T,H), (H,T,H,H), (T,H,H,H). The tuples (H,H,H,H), (H,H,H,T), (H,H,T,H), (H,T,H,H) are common among E and F.

```
n_EandF = 4 # number of elements in the intersection of E and F
p_EandF = n_EandF / n
print(paste0("P(E and F) = ", p_EandF))
```

```
## [1] "P(E and F) = 0.25"
```

We need $P(E|F) = \frac{P(E \cap F)}{P(F)}$.

```
p_EgivenF = p_EandF / p_F
print(paste0("P(E|F) = ", p_EgivenF))
```

```
## [1] "P(E|F) = 0.5"
```

3. (Bayes' Theorem)

An analyst predicts that there is a 0.40 probability that the U.S. economy will perform well. If the U.S. economy performs well, then there is a 0.80 probability that Asian countries will also perform well. On the other hand, if the U.S. economy does not perform well, then the probability that Asian countries will perform well goes down to 0.30.

(a) What is the probability that both the U.S. economy and the Asian countries will perform well?

Solution.

Let A = the U.S. economy will perform well, B = Asian countries will perform well. We first list the probabilities that are given in the question:

$$\begin{aligned} P(A) &= 0.40 \\ P(B|A) &= 0.80 \\ P(B|\bar{A}) &= 0.30 \end{aligned}$$

We need $P(A \cap B) = P(A)P(B|A)$.

```
p_A = 0.4 # probability that the U.S. economy will perform well
p_BgivenA = 0.8 # probability that Asian countries will perform well
# given that the U.S. economy performs well
p_AandB = p_A * p_BgivenA
print(paste0("P(A and B) = ", p_AandB))
```

```
## [1] "P(A and B) = 0.32"
```

(b) What is the probability that the Asian countries will perform well?

Solution.

We need P(B). We can use the total probability rule to write $P(B) = P(A \cap B) + P(\bar{A} \cap B)$. We already have the first of these terms. To compute the second one use the following: $P(\bar{A} \cap B) = P(B|\bar{A})P(\bar{A})$

```
p_NotA = 1 - p_A
p_BgivenNotA = 0.3 # probability that Asian countries will perform well
# given that the U.S. economy does not perform well
p_NotAandB = p_NotA * p_BgivenNotA
p_B = p_AandB + p_NotAandB
print(paste0("P(B) = ", p_B))
```

```
## [1] "P(B) = 0.5"
```

(c) What is the probability that the U.S. economy will perform well, given that the Asian countries perform well?

Solution.

We need P(A|B). We can use the Bayes' theorem to write $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$.

```
p_AgivenB = p_BgivenA * p_A / p_B
print(paste0("P(A|B) = ", p_AgivenB))
```

```
## [1] "P(A|B) = 0.64"
```

(d) What is the probability that the U.S. economy will not perform well, given that the Asian countries do not perform well?

Solution.

We need $P(\bar{A}|\bar{B})$. We can use the Bayes' theorem to write $P(\bar{A}|\bar{B}) = \frac{P(\bar{B}|\bar{A})P(\bar{A})}{P(\bar{B})}$ where $P(\bar{B}) = 1 - P(B)$ and $P(\bar{B}|\bar{A}) = 1 - P(B|\bar{A})$.

```
p_NotB = 1 - p_B
p_NotBgivenNotA = 1 - p_BgivenNotA
p_NotAgivenNotB = p_NotBgivenNotA * p_NotA / p_NotB
print(paste0("P(not A|not B) = ", p_NotAgivenNotB))
```

```
## [1] "P(not A|not B) = 0.84"
```

4. (Random Variables)

Investment advisors recommend risk reduction through international diversification. International investing allows you to take advantage of the potential for growth in foreign economies, particularly in emerging markets. An investor is considering investment in either Europe or Asia. The investor has studied these markets and believes that both markets will be influenced by the U.S. economy, which has the following three possible states and associated probabilities: a "good" state that will occur with probability 0.20, a "fair" state that will occur with probability 0.50, and a "poor" state that will occur with probability 0.30.

The probability distributions of the returns for investments in Europe and Asia are given in the accompanying table.

State of the U.S. Economy	Returns in Europe (in %)	Returns in Asia (in %)	Probability of State
Good	10	18	0.20
Fair	6	10	0.50
Poor	-6	-12	0.30

(a) Find the expected value and the standard deviation of returns in Europe.

Solution.

Let X = returns in Europe. We need

$$\begin{aligned} E(X) &= P(Good) * 10 + P(Fair) * 6 + P(Poor) * (-6) \\ var(X) &= P(Good) * (10 - E(X))^2 + P(Fair) * (6 - E(X))^2 + P(Poor) * (-6 - E(X))^2 \\ std(X) &= \sqrt{var(X)}. \end{aligned}$$

```
p = c(0.2, 0.5, 0.3) # probabilities of the states
x = c(10, 6, -6) # returns in Europe for each state
E_Europe = sum(p * x)
var_Europe = sum(p * (x - E_Europe)^2)
std_Europe = sqrt(var_Europe)
print(paste0("E(X) = ", E_Europe, "%, std(X) = ", std_Europe, "%"))
```

```
## [1] "E(X) = 3.2%, std(X) = 6.20966987850401%"
```

(b) Find the expected value and the standard deviation of returns in Asia.

Let Y = returns in Asia. We need

$$\begin{aligned} E(Y) &= P(Good) * 18 + P(Fair) * 10 + P(Poor) * (-12) \\ var(Y) &= P(Good) * (18 - E(Y))^2 + P(Fair) * (10 - E(Y))^2 + P(Poor) * (-12 - E(Y))^2 \\ std(Y) &= \sqrt{var(Y)}. \end{aligned}$$

```
y = c(18, 10, -12) # returns in Asia for each state
E_Asia = sum(p * y)
var_Asia = sum(p * (y - E_Asia)^2)
std_Asia = sqrt(var_Asia)
print(paste0("E(Y) = ", E_Asia, "%, std(Y) = ", std_Asia, "%"))
```

```
## [1] "E(Y) = 5%, std(Y) = 11.5325625946708%"
```

(c) Which of the two investments (Europe or Asia) will this investor pick if the investor prefers higher expected returns and is indifferent to standard deviation? Such kind of investor is called "risk-neutral".

Solution.

Since E(X) < E(Y), this "risk-neutral" investor will pick the investment in Asia.

(d) Which of the two investments (Europe or Asia) will this investor pick if the investor prefers lower standard deviation and is indifferent to expected returns? Such kind of investor is called "risk-averse".

Remark: Investment decisions typically involve a tradeoff between risk and return. You will see a lot on this theme during your time at Tepper.

Solution.

Since std(X) < std(Y), this "risk-averse" investor will pick the investment in Europe.

Solution to Little Test 3

1. Suppose a population is normally distributed with mean 20 and variance 4. Compute a 90% acceptance interval for the sample mean of random samples of size 36.

   Solution. The $(1-\alpha)$ acceptance interval is
   $$\mu \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$
   We have $\mu = 20, \sigma = 2, \alpha = 0.1$, and $n = 36$. Thus the 90% acceptance interval is
   $$20 \pm 1.6448 \cdot \frac{2}{\sqrt{36}} = 20 \pm 0.54828$$

2. Suppose a poll of 3000 American adults reveals that 1400 of them approve of the job that Joe Biden is doing as president. Thus the estimate of the proportion of the population who approve Joe Biden is $1400/3000 = 0.467$

   Find the margin of error of this estimate at the 0.95 confidence level. (That is, the radius of the 0.95 confidence interval for the population proportion.)

   Please enter the number with four digits of accuracy. Note that the number is between 0.0001 and 0.1000.

   Solution. We have $n = 3000$ and $\hat{p} = 1400/3000 = 0.467$. Thus the MOE at the 0.95 confidence level is
   $$z_{0.025} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{3000}} = 1.96 \cdot \sqrt{\frac{0.467 \cdot 0.533}{3000}} = 0.017853$$

3. A hypothesis test for population mean based on a random sample of size n rejects the null hypothesis at the significance level if the p-value of the test is less than α.

   Solution. TRUE: This is one of the rules of the t-test.

4. Suppose the sample correlation of 200 observations of the variables X and Y is 0.2. Then we reject the null hypothesis that the correlation of X and Y is zero at the 0.05 significance level.

   Solution. TRUE: Compute the t-stat. We have $n = 200$ and $r = 0.2$. Hence
   $$\text{t-stat} = 0.2 \cdot \sqrt{\frac{198}{1-0.2^2}} = 2.8722$$
   which is larger the the critical t-value:
   $$t_{198,0.025} = \texttt{T.INV}(\texttt{0.975}, \texttt{198}) = 1.972.$$

5. A hypothesis test for population mean based on a random sample of size n rejects the null hypothesis at the significance level α if $|\text{t-stat}|$ is less than $t_{n-1,\alpha/2}$.

   Solution. FALSE: The rule for the t-test is that we reject at the significance level α if $|\text{t-stat}|$ is LARGER than $t_{n-1,\alpha/2}$.

6. The estimate of a regression coefficient is statistically significant at level α when the $(1-\alpha)$ confidence interval of that coefficient does not contain the value zero.

   Solution. TRUE: The estimate is statistically significant at level α when we reject the null hypothesis that it is zero at level α. The above statement is one of the rejection rule for the t-test.

46-880 Introduction to Probability and Statistics

Practice Set 2 Solutions, mini-1 2023

1. (Random Variables)

This problem is based on a "doubling-up system" devised by gamblers in an attempt to win at casinos in Las Vegas. Suppose you win a dollar for each dollar you bet if the flip of a fair coin comes out heads. You start by betting \$1. If you win in the first round, you quit with a \$1 net profit. If you lose in the first round, then you could enter the gamble for a second round and bet \$2. Hence if you win in the second round, you recover the loss of the first round and quit with a net \$1 profit. If you lose in the second round again, then you could enter the gamble for a third round and bet \$4. Hence if you win in the third round, you recover the losses of the first and second rounds and quit with a net \$1 profit. You could keep going always doubling your bet until the coin flip finally comes out heads at which point you quit with a net \$1 profit.

(a) Suppose you use the above doubling-up system but only up to three rounds. In other words, you always quit after the third round if you make it that far. Find the probability of each of the following events:

A = you win in the first round

B = you lose in the first round and win in the second round

C = you lose in the first and second rounds and win in the third round

D = you lose in all three rounds

Solution.

The events can be described in terms of coin flips as follows:

A = outcome of the first coin flip is H

B = outcome of the first and second coin flips is (T,H)

C = outcome of the first, second, and third coin flips is (T,T,H)

D = outcome of the first, second, and third coin flips is (T,T,T)

Therefore, P(A) = 0.5, P(B) = 0.5 * 0.5 = 0.25, P(C) = P(D) = 0.5 * 0.5 * 0.5 = 0.125.

(b) Let X be the number of rounds you end up playing. Given the above system, X can take the three values 1, 2, and 3. Describe the probability distribution of X:

Value	1	2	3
Probability			

Solution.

The values of X are respectively 1, 2, and 3 for the mutually exclusive and collectively exhaustive events A, B, C∪D. Hence from part (a) we get the following probability distribution for X:

X Value	1	2	3
Probability	0.5	0.25	0.25

(c) Let Y be the value of your final net profit or loss (after at most three rounds). Describe the probability distribution of Y and the value of E(Y). To that end, first determine the possible values of Y and then the probability that Y attains each of them.

Solution.

The possible values of Y are respectively 1 and -7 = - 1 - 2 - 4 for the mutually exclusive and collectively exhaustive events A∪B∪C and D. Hence from part (a) we get the following probability distribution for Y:

Y Value	1	- 7
Probability	0.875	0.125

We need E(Y).

```
p_y = c(1 - (0.5)^3, (0.5)^3) # Y probabilities
y = c(1, -7) # Y values
E_Y = sum(p_y * y)
print(paste0("E(Y) = ", E_Y))
```

```
## [1] "E(Y) = 0"
```

(d) Suppose you use the above doubling-up system but this time you go up to five rounds. Let Z be the value of your final net profit or loss (after at most five rounds). Describe the probability distribution of Z and the value of E(Z).

Solution.

The possible values of Z are respectively 1 and -31 = - 1 - 2 - 4 - 8 - 16. The latter occurs when all five coin flips are tails which occurs with probability 0.5 * 0.5 * 0.5 * 0.5 * 0.5 = 0.03125. Hence we get the following probability distribution for Z:

Z Value	1	- 31
Probability	0.96875	0.03125

We need E(Z).

```
p_z = c(1 - (0.5)^5, (0.5)^5) # Z probabilities
z = c(1, -31) # Z values
E_Z = sum(p_z * z)
print(paste0("E(Z) = ", E_Z))
```

```
## [1] "E(Z) = 0"
```

(e) (***Optional***) Repeat part (d) assuming you go up to n rounds.

Solution.

Similarly, the possible values are respectively 1 and - (2^n - 1) = - 1 - 2 - 4 - . . . - 2^{n-1}. The latter occurs when all n coin flips are tails which occurs with probability $(0.5)^n$. Hence we get the following distribution:

Value	1	- (2^n - 1)
Probability	1 - $(0.5)^n$	$(0.5)^n$

The expectation is equal to

$$(1 - (0.5)^n) * 1 + (0.5)^n * (1 - 2^n) = 1 - (0.5)^n + (0.5)^n - (0.5)^n * 2^n = 0.$$

You can play with the following chunk of code changing the value of n and checking that you always get zero as the expectation.

```
n = 100 # change this value and check the expectation
p_n = c(1 - (0.5)^n, (0.5)^n) # probabilities
v_n = c(1, 1 - 2^n) # values
Expectation = sum(p_n * v_n)
print(paste0("Expectation = ", Expectation))
```

```
## [1] "Expectation = 0"
```

2. (Binomial Distribution)

According to a survey by Transamerica Center for Health Studies, 15% of Americans still have no health insurance. Suppose five individuals are randomly selected.

(a) What is the probability that all five have health insurance?

Solution.

Let X denote the number of individuals who have health insurance. We have that X~B(5, 0.85). We need P(X = 5).

```
n = 5 # number of individuals
p = 0.85 # probability of an individual having health insurance
p_5 = dbinom(5, n, p)
print(paste0("P(X = 5) = ", p_5))
```

```
## [1] "P(X = 5) = 0.4437053125"
```

(b) What is the probability that no more than two have health insurance?

Solution.

We need P(X <= 2).

```
p_atmost2 = pbinom(2, n, p)
print(paste0("P(X <= 2) = ", p_atmost2))
```

```
## [1] "P(X <= 2) = 0.026611875"
```

(c) What is the probability that at least four have health insurance?

Solution.

We need P(X >= 4) = 1 - P(X <= 3).

```
p_atleast4 = 1 - pbinom(3, n, p)
print(paste0("P(X >= 4) = ", p_atleast4))
```

```
## [1] "P(X >= 4) = 0.83521"
```

(d) What is the expected number of individuals who have health insurance?

Solution.

We need E(X) = n * p.

```
E_X = n * p
print(paste0("E(X) = ", E_X))
```

```
## [1] "E(X) = 4.25"
```

(e) Calculate the variance and the standard deviation for the probability distribution of the number of individuals who have health insurance.

Solution.

We need var(X) = n * p * (1 - p).

```
var_X = n * p * (1 - p)
std_X = sqrt(var_X)
print(paste0("var(X) = ", var_X, ", std(X) = ", std_X))
```

```
## [1] "var(X) = 0.6375, std(X) = 0.798435971133566"
```

3. (Binomial Distribution)

A tour operator has a bus that can accommodate 30 tourists. The operator knows that tourists may not show up, so he sells 35 tickets. The probability that an individual tourist will not show up is 0.16, independent of all other tourists. Each ticket costs $60, and is non-refundable if a tourist fails to show up. If a tourist shows up and a seat is not available, the tour operator has to pay $50 penalty to the tourist (in addition to a refund).

(a) What is the probability that there will be a seat available for every tourist who shows up? In other words, what is the probability that the number of tourists who show up is at most 30?

Solution.

Let X denote the total number of tourists (out of the 35 who bought tickets) who are no-show. We have that X~B(35, 0.16). We need P(X >= 5) = 1 - P(X <= 4).

```
n = 35 # number of tourists
p = 0.16 # probability of no-show
p_atleast5 = 1 - pbinom(4, n, p)
print(paste0("P(X >= 5) = ", p_atleast5))
```

```
## [1] "P(X >= 5) = 0.679117485922592"
```

(b) What is the probability that the tour will be overbooked? In other words, what is the probability that at least 31 out of the 35 reservations end up showing up?

Solution.

We need P(X <= 4).

```
p_atmost4 = pbinom(4, n, p)
print(paste0("P(X <= 4) = ", p_atmost4))
```

```
## [1] "P(X <= 4) = 0.320882514077408"
```

(c) What is the probability that the tour will be overbooked by exactly 2 tourists? In other words, what is the probability that exactly 32 out of the 35 reservations end up showing up?

Solution.

We need P(X = 3).

```
p_3 = dbinom(3, n, p)
print(paste0("P(X = 3) = ", p_3))
```

```
## [1] "P(X = 3) = 0.101206114748936"
```

(d) What is the probability that the tour will be overbooked by 3 or more tourists?

Solution.

We need P(X <= 2).

```
p_atmost2 = dbinom(2, n, p)
print(paste0("P(X <= 2) = ", p_atmost2))
```

```
## [1] "P(X <= 2) = 0.0483029184029012"
```

(e) What is the probability that there will be a seat available for every tourist who shows up and the tour starts with at least 27 passengers?

Solution.

We need P(5 <= X <= 8) = P(X <= 8) - P(X <= 4).

```
p_5to8 = pbinom(8, n, p) - pbinom(4, n, p)
print(paste0("P(5 <= X <= 8) = ", p_5to8))
```

```
## [1] "P(5 <= X <= 8) = 0.583701040170801"
```

(f) Let Y denote the value of the total penalty that the tour operator ends up paying. Find the distribution of Y. To that end, determine all possible values of Y and the probability of each of them.

Solution.

Observe that Y takes the values $250, $200, $150, $100, $50, and zero when X = 0, X = 1, X = 2, X = 3, X = 4, and X >= 5, respectively.

```
x = c(0, 1, 2, 3, 4)
y = c(250, 200, 150, 100, 50, 0)                # values that Y can take
p_y = c(dbinom(x, n, p), 1 - pbinom(4, n, p)) # compute probability of each value of Y
print(paste0("P(Y = ", y, ") = ", p_y))        # print the pmf of Y
```

```
## [1] "P(Y = 250) = 0.00223756166131086" "P(Y = 200) = 0.0149170777420724"
## [3] "P(Y = 150) = 0.0483029184029012"  "P(Y = 100) = 0.101206114748936"
## [5] "P(Y = 50) = 0.154218841522188"    "P(Y = 0) = 0.679117485922592"
```

Hence we get the following probability distribution for Y:

Y Value	250	200	150	100	50	0
Probability	0.0022	0.0149	0.0483	0.1012	0.1542	0.6791

(g) Find the expected value E(Y) and variance var(Y) of the total penalty the tour operator has to pay.

Solution.

We need E(Y) and var(Y).

```
E_Y = sum(p_y * y)
var_Y = sum(p_y * (y - E_Y)^2)
print(paste0("E(Y) = $", E_Y, ", var(Y) = ", var_Y))
```

```
## [1] "E(Y) = $28.6197972751803, var(Y) = 2401.86183280251"
```

4. (Poisson Distribution)

A local pharmacy administers an average of 84 Covid-19 vaccines per week. The vaccine shots are evenly administered across all days. Suppose the number of Covid-19 vaccine shots administered at a local pharmacy follows a Poisson distribution with an average of 84 vaccines administered per week.

(a) What is the expected number of vaccine shots administered during a day, during a weekend (Saturday through Sunday), and during a week (Monday through Sunday)?

Solution.

Let X, Y, and Z denote the number of vaccine shots administered during a week, during a day, and during a weekend, respectively. We have X~Poisson(84), Y~Poisson(84/7), and Z~Poisson(84*2/7). We need E(X), E(Y), and E(Z).

```r
lambda_X = 84 # vaccine rate per week
lambda_Y = 84 / 7 # vaccine rate per day
lambda_Z = 84 * 2 / 7 # vaccine rate per weekend
print(paste0("E(X) = ", lambda_X, ", E(Y) = ", lambda_Y, ", and E(Z) = ", lambda_Z))
```

```
## [1] "E(X) = 84, E(Y) = 12, and E(Z) = 24"
```

(b) Find the probability that no vaccine shots are administered during a day.

Solution.

We need P(Y = 0).

```r
p_Y_0 = dpois(0, lambda_Y)
print(paste0("P(Y = 0) = ", p_Y_0))
```

```
## [1] "P(Y = 0) = 6.14421235332821e-06"
```

(c) Find the probability that the number of vaccine shots administered during a week (Monday through Sunday) is more than 70.

Solution.

We need P(X > 70) = 1 - P(X <= 70).

```r
p_X_more70 = 1 - ppois(70, lambda_X)
print(paste0("P(X > 70) = ", p_X_more70))
```

```
## [1] "P(X > 70) = 0.932749722518648"
```

(d) Find the probability that the number of vaccine shots administered during a weekend (Saturday through Sunday) is more than 20 but less than 30.

Solution.

We need P(20 < Z < 30) = P(Z <= 29) - P(Z <= 20).

```r
p_Z_20to30 = ppois(29, lambda_Z) - ppois(20, lambda_Z)
print(paste0("P(20 < Z < 30) = ", p_Z_20to30))
```

```
## [1] "P(20 < Z < 30) = 0.625237727991275"
```

(e) Find the probability that exactly 70 vaccine shots are administered during the next week (Monday through Sunday) given that 90 vaccine shots were administered during the last week.

Solution.

A property of the Poisson distribution is that the numbers of occurrences of the event in disjoint time intervals are mutually independent. We need P(X = 70).

```
p_X_70 = dpois(70, lambda_X)
print(paste0("P(X = 70) = ", p_X_70))
```

```
## [1] "P(X = 70) = 0.0138168330109938"
```

46-880 Introduction to Probability and Statistics

Practice Set 3 Solutions, mini-1 2023

1. (Uniform Distribution)

A worker at a landscape design center uses a machine to fill bags with potting soil. Assume that the quantity put in each bag follows the continuous uniform distribution with low and high filling weights of 10 pounds and 12 pounds, respectively.

(a) Calculate the expected value and the standard deviation of this distribution.

Solution.

Let X denote the weight of a bag. We have X~U[10, 12]. We need E(X) and std(X).

```r
a = 10 # lowest possible value of X
b = 12 # highest possible value of X
E_X = (a + b) / 2
var_X = (b - a)^2 / 12
std_X = sqrt(var_X)
print(paste0("E(X) = ", E_X, ", std(X) = ", std_X))
```

```
## [1] "E(X) = 11, std(X) = 0.577350269189626"
```

(b) Find the probability that the weight of a randomly selected bag is no more than 11 pounds.

Solution.

We need P(X <= 11).

```r
p_b = punif(11, 10, 12)
print(paste0("P(X <= 11) = ", p_b))
```

```
## [1] "P(X <= 11) = 0.5"
```

(c) Find the probability that the weight of a randomly selected bag is at least 10.5 pounds.

Solution.

We need P(X >= 10.5) = 1 - P(X < 10.5).

```r
p_c = 1 - punif(10.5, 10, 12)
print(paste0("P(X >= 10.5) = ", p_c))
```

```
## [1] "P(X >= 10.5) = 0.75"
```

(d) Find the probability that the weight of a randomly selected bag is between 10.4 and 11.8 pounds.

Solution.

We need P(10.4 <= X <= 11.8) = P(X <= 11.8) - P(X < 10.4).

```
p_d = punif(11.8, 10, 12) - punif(10.4, 10, 12)
print(paste0("P(10.4 <= X <= 11.8) = ", p_d))
```

```
## [1] "P(10.4 <= X <= 11.8) = 0.7"
```

(e) Find the probability that the weight of a randomly selected bag is one standard deviation or more below its expected value.

Solution.

We need P(E(X) - X >= std(X)) = P(X <= E(X) - std(X)).

```
p_e = punif(E_X - std_X, 10, 12)
print(paste0("P(E(X) - X >= std(X)) = ", p_e))
```

```
## [1] "P(E(X) - X >= std(X)) = 0.211324865405187"
```

(f) If the weight of a randomly selected bag is at least 10.25 pounds, what is the probability that its weight is at most 11.5 pounds?

Solution.

We need P(X <= 11.5|X >= 10.25) = $\frac{P(10.25<=X<=11.5)}{P(X>=10.25)}$ where P(10.25 <= X <= 11.5) = P(X <= 11.5) - P(X < 10.25) and P(X >= 10.25) = 1 - P(X < 10.25).

```
p_num = punif(11.5, 10, 12) - punif(10.25, 10, 12)
p_den = 1 - punif(10.25, 10, 12)
p_f = p_num / p_den
print(paste0("P(X <= 11.5|X >= 10.25) = ", p_f))
```

```
## [1] "P(X <= 11.5|X >= 10.25) = 0.714285714285714"
```

2. (Exponential Distribution)

Prior to placing an order, the amount of time (in minutes) that a driver waits in line at a Starbucks drive-thru follows an exponential distribution with a probability density function of f(x) = 0.2 $e^{-0.2x}$.

(a) What is the mean waiting time (in minutes) that a driver faces prior to placing an order?

Solution.

Let X denote the waiting time (in minutes). We have X~Exp(0.2). We need E(X).

```
lambda = 0.2 # rate of service
E_X = 1 / lambda
print(paste0("E(X) = ", E_X, " minutes"))
```

```
## [1] "E(X) = 5 minutes"
```

(b) What is the rate parameter λ of this distribution? What is the standard deviation of this distribution?

Solution.

The rate parameter is $\lambda = 0.2$. We need std(X).

```
var_X = 1 / lambda^2
std_X = sqrt(var_X)
print(paste0("std(X) = ", std_X))
```

```
## [1] "std(X) = 5"
```

Note that var(X) = E(X).

(c) What is the probability that a driver spends more than the mean waiting time before placing an order?

Solution.

We need P(X > E(X)) = 1 - P(X <= E(X)).

```r
p_moreE = 1 - pexp(E_X, lambda)
print(paste0("P(X > E(X)) = ", p_moreE))
```

```
## [1] "P(X > E(X)) = 0.367879441171442"
```

(d) What is the probability that a driver spends less than 10 minutes before placing an order?

Solution.

We need P(X < 10).

```r
p_less10 = pexp(10, lambda)
print(paste0("P(X < 10) = ", p_less10))
```

```
## [1] "P(X < 10) = 0.864664716763387"
```

(e) What is the probability that a driver spends between 4 and 6 minutes before placing an order?

Solution.

We need P(4 <= X <= 6) = P(X <= 6) - P(X < 4).

```r
p_4to6 = pexp(6, lambda) - pexp(4, lambda)
print(paste0("P(4 <= X <= 6) = ", p_4to6))
```

```
## [1] "P(4 <= X <= 6) = 0.148134752205019"
```

(f) A driver has already waited in line for 5 minutes. What is the probability that she will spend at least 10 more minutes before placing an order?

Solution.

We need P(X >= 15|X >= 5). Using the memoryless property of the exponential distribution, we simply need P(X >= 15 - 5) = P(X >= 10) = 1 - P(X < 10).

```r
p_more10 = 1 - pexp(10, lambda)
print(paste0("P(X >= 10) = ", p_more10))
```

```
## [1] "P(X >= 10) = 0.135335283236613"
```

3. (Poisson & Exponential Distributions)

In a local law office, jobs to a printer are sent at a rate of 8 jobs per hour. Suppose that the number of jobs sent to a printer follows the Poisson distribution.

(a) What is the expected time between successive jobs?

Solution.

Let X denote the time between successive jobs in hours. We have X~Exp(8). We need E(X).

```
lambda_X = 8
E_X = 1 / lambda_X
print(paste0("E(X) = ", E_X, " hours"))
```

```
## [1] "E(X) = 0.125 hours"
```

(b) A job was just sent to the printer. What is the probability that the next job will be sent within five minutes?

Solution.

We need P(X <= 5/60).

```
p_X_b = pexp(5/60, lambda_X)
print(paste0("P(X <= 5/60) = ", p_X_b))
```

```
## [1] "P(X <= 5/60) = 0.486582880967408"
```

Alternatively, let Y denote the number of jobs sent to the printer in 5 minutes. We have Y~Poisson(8*5/60). we need P(Y >= 1) = 1 - P(Y = 0).

```
lambda_Y = 8 * 5 / 60
p_Y_b = 1 - dpois(0, lambda_Y)
print(paste0("P(Y >= 1) = ", p_Y_b))
```

```
## [1] "P(Y >= 1) = 0.486582880967408"
```

(c) What is the probability that the time between successive jobs is between 15 and 20 minutes?

Solution.

We need P(15/60 <= X <= 20/60) = P(X <= 20/60) - P(X < 15/60).

```
p_X_c = pexp(20/60, lambda_X) - pexp(15/60, lambda_X)
print(paste0("P(15/60 <= X <= 20/60) = ", p_X_c))
```

```
## [1] "P(15/60 <= X <= 20/60) = 0.0658518320138112"
```

(d) What is the probability that 5 or more jobs will be sent in the next 40 minutes?

Solution.

Let Z denote the number of jobs sent to the printer in 40 minutes. We have Z~Poisson(8*40/60). We need P(Z >= 5) = 1 - P(Z <= 4).

```
lambda_Z = 8 * 40 / 60
p_Z_d = 1 - ppois(4, lambda_Z)
print(paste0("P(Z >= 5) = ", p_Z_d))
```

```
## [1] "P(Z >= 5) = 0.615929624153269"
```

(e) What is the probability that no jobs will be sent in the next ten minutes?

Solution.

We need P(X > 10/60) = 1 - P(X <= 10/60).

```
p_X_e = 1 - pexp(10/60, lambda_X)
print(paste0("P(X > 10/60) = ", p_X_e))
```

```
## [1] "P(X > 10/60) = 0.263597138115727"
```

Alternatively, let W denote the number of jobs sent to the printer in 10 minutes. We have W~Poisson(8*10/60). We need P(W = 0).

```
lambda_W = 8 * 10 / 60
p_W_e = dpois(0, lambda_W)
print(paste0("P(W = 0) = ", p_W_e))
```

```
## [1] "P(W = 0) = 0.263597138115727"
```

(f) What is the probability that at most 8 jobs will be sent in the next hour given that 6 jobs were sent in the last hour?

Solution.

A property of the Poisson distribution is that the numbers of occurrences of the event in disjoint time intervals are mutually independent. Let V denote the number of jobs sent to the printer in an hour. We have V~Pois(8). We need P(V <= 8).

```
lambda_V = 8
p_V_f = ppois(8, lambda_V)
print(paste0("P(V <= 8) = ", p_V_f))
```

```
## [1] "P(V <= 8) = 0.592547341437591"
```

4. (Normal Distribution)

Suppose that the miles-per-gallon (mpg) rating of passenger cars is a normally distributed random variable with mean = 33.8 mpg and standard deviation = 3.5 mpg.

(a) What is the probability that a randomly selected passenger car gets at least 40 mpg?

Solution.

Let X denote the mpg of a randomly selected passenger car. X has normal distribution with mean $\mu = 33.8$ and variance $\sigma^2 = 3.5^2 = 12.25$, that is, X~N(33.8, 12.25). We need P(X >= 40) = 1 - P(X <= 40).

```
mu = 33.8 # mean of the distribution
sigma = 3.5 # standard deviation of the distribution
p_more40 = 1 - pnorm(40, mu, sigma)
print(paste0("P(X >= 40) = ", p_more40))
```

```
## [1] "P(X >= 40) = 0.038244730448449"
```

(b) What is the probability that a randomly selected passenger car gets between 30 and 35 mpg?

Solution.

We need P(30 <= X <= 35) = P(X <= 35) - P(X <= 30).

```
p_30to35 = pnorm(35, mu, sigma) - pnorm(30, mu, sigma)
print(paste0("P(30 <= X <= 35) = ", p_30to35))
```

```
## [1] "P(30 <= X <= 35) = 0.495344323481789"
```

(c) An automobile manufacturer wants to build a new passenger car with an mpg rating that improves upon 95% of existing cars. What is the minimum mpg that would achieve this goal?

Solution.

We want to compute L such that P(X <= L) = 0.95 . That is, the minimum value for the mpg that will place it in the top 5% of the distribution is the 0.95-quantile of mpgs.

```
L = qnorm(0.95, mu, sigma)
print(paste0("0.95-quantile = ", L))
```

```
## [1] "0.95-quantile = 39.5569876943301"
```

(d) What is the probability that a randomly selected passenger car gets an mpg rating that exceeds the mean mpg by more than 2 standard deviations? Does your answer depend on the actual values of the mean and standard deviation?

Solution.

We need P(X - $\mu > 2\sigma$) = P(X > $\mu + 2\sigma$) = 1 - P(X <= $\mu + 2\sigma$).

```
p_d = 1 - pnorm(mu + 2*sigma, mu, sigma)
print(paste0("P(X > ", mu + 2*sigma, ") = ", p_d))
```

```
## [1] "P(X > 40.8) = 0.0227501319481792"
```

The answer does not depend on actual values of mean and standard deviation. This is a property of normal distribution. More generally, for any given value of m, P(X > $\mu + m\sigma$), or equivalently P(X < $\mu + m\sigma$), is fixed regardless of the actual values of μ and σ.

(e) Two passenger cars are chosen at random. What is the probability that at least one of them gets more than 40 mpg?

Solution.

From part (a), we know that the probability of an mpg more than 40 for a randomly chosen passenger car is approximately 0.0382. Let Y denote the number of passenger cars (out of 2) which get more than 40 mpg. We have Y~B(2, 0.0382). We need P(Y >= 1) = 1 - P(Y = 0).

```
p_Y_more1 = 1 - dbinom(0, 2, p_more40)
print(paste0("P(Y >= 1) = ", p_Y_more1))
```

```
## [1] "P(Y >= 1) = 0.0750268014898234"
```

5. Normal Distribution and Combination of Random Variables

Stock A has an annual expected return of 8% with a standard deviation of 6%. Stock B has an annual expected return of 12% with a standard deviation of 10%. Assume the returns of these two stocks are independent and normally distributed.

(a) Which one of the stocks is more likely to attain a negative return over the next year?

Solution.

Let A and B denote the annual returns of stock A and stock B, respectively. We need to compare P(A < 0) and P(B < 0).

```
mu_A = 0.08
sigma_A = 0.06
mu_B = 0.12
sigma_B = 0.10
p_Aless0 = pnorm(0, mu_A, sigma_A)
p_Bless0 = pnorm(0, mu_B, sigma_B)
print(paste0("P(A < 0) = ", p_Aless0, ", P(B < 0) = ", p_Bless0))
```

```
## [1] "P(A < 0) = 0.0912112197258678, P(B < 0) = 0.115069670221708"
```

Therefore, stock B is more likely to attain a negative return over the next year.

Alternatively, we could say stock B is more likely to attain a negative return since for this one the difference from 0 to the mean is smaller when measured in standard deviations.

(b) Suppose an investor has an annual target return of 14%. Which one of the stocks is more likely to attain or exceed this target return over the next year??

Solution.

We need to compare P(A >= 0.14) = 1 - P(A < 0.14) and P(B >= 0.14) = 1 - P(B < 0.14).

```
target = 0.14
p_Amore14 = 1 - pnorm(0.14, mu_A, sigma_A)
p_Bmore14 = 1 - pnorm(0.14, mu_B, sigma_B)
print(paste0("P(A >= 0.14) = ", p_Amore14, ", P(B >= 0.14) = ", p_Bmore14))
```

```
## [1] "P(A >= 0.14) = 0.158655253931457, P(B >= 0.14) = 0.420740290560897"
```

Therefore, stock B is more likely to attain or exceed this target return over the next year.

Alternatively, we could say stock B is more likely to attain or exceed this target return since for this one the difference from target to the mean is smaller when measured in standard deviations.

(c) What is the probability that the annual return of stock A is higher than the annual return of stock B?

Solution.

Let C = A - B. We have E(C) = E(A) - E(B) and var(C) = var(A) + var(B) since A and B are independent. Thus, C~N(E(C), var(C)) since A and B are normally distributed. We need p(A > B) = P(A - B > 0) = P(C > 0) = 1 - P(C <= 0).

```
mu_C = mu_A - mu_B
var_C = sigma_A^2 + sigma_B^2
sigma_C = sqrt(var_C)
p_Cmore0 = 1 - pnorm(0, mu_C, sigma_C)
print(paste0("P(A > B) = ", p_Cmore0))
```

```
## [1] "P(A > B) = 0.365800294479951"
```

(d) What is the probability that the annual return of stock B is at least 2% higher than the annual return of stock A?

Solution.

We need p(B - A >= 0.02) = P(C <= -0.02).

```
p_Cless2 = pnorm(-0.02, mu_C, sigma_C)
print(paste0("P(B - A >= 0.02) = ", p_Cless2))
```

```
## [1] "P(B - A >= 0.02) = 0.568084128572637"
```

(e) Consider three possible allocations of investment capital:

Allocation 1: 40% stock A and 60% stock B

Allocation 2: 60% stock A and 40% stock B

Allocation 3: 50% stock A and 50% stock B

In finance, the *Sharpe ratio* measures the performance of an investment such as a security or portfolio compared to a risk-free asset, after adjusting for its risk. Assuming the expected return of the risk-free asset is 0, Sharpe ratio is the ratio

$$\frac{\text{expected return}}{\text{standard deviation}}.$$

Determine the Sharpe ratio for each of the three possible allocations. Which of the three allocations has the highest Sharpe ratio?

Solution.

Let A1, A2, and A3 denote the annual returns of Allocation 1, Allocation 2, and Allocation 3, respectively. A1, A2, and A3 follow normal distributions.

```
mu_A1 = 0.4 * mu_A + 0.6 * mu_B
sigma_A1 = sqrt(0.4^2 * sigma_A^2 + 0.6^2 * sigma_B^2)
Sharpe_A1 = mu_A1 / sigma_A1
print(paste0("E(A1) = ", mu_A1, ", std(A1) = ", sigma_A1, ", Sharpe ratio = ", Sharpe_A1))
```

```
## [1] "E(A1) = 0.104, std(A1) = 0.064621977685614, Sharpe ratio = 1.60935959753445"
```

```
mu_A2 = 0.6 * mu_A + 0.4 * mu_B
sigma_A2 = sqrt(0.6^2 * sigma_A^2 + 0.4^2 * sigma_B^2)
Sharpe_A2 = mu_A2 / sigma_A2
print(paste0("E(A2) = ", mu_A2, ", std(A2) = ", sigma_A2, ", Sharpe ratio = ", Sharpe_A2))
```

```
## [1] "E(A2) = 0.096, std(A2) = 0.0538144961882948, Sharpe ratio = 1.7839059509932"
```

```
mu_A3 = 0.5 * mu_A + 0.5 * mu_B
sigma_A3 = sqrt(0.5^2 * sigma_A^2 + 0.5^2 * sigma_B^2)
Sharpe_A3 = mu_A3 / sigma_A3
print(paste0("E(A3) = ", mu_A3, ", std(A3) = ", sigma_A3, ", Sharpe ratio = ", Sharpe_A3))
```

```
## [1] "E(A3) = 0.1, std(A3) = 0.058309518948453, Sharpe ratio = 1.71498585142509"
```

Therefore Allocation 2 has the highest Sharpe ratio.

46-880 Introduction to Probability and Statistics

Practice Set 4 Solutions, mini-1 2023

1. Normal Distribution and Combination of Random Variables

Consider the following hypothetical investment universe:

- The universe has 100 stocks
- The stocks have returns that are i.i.d. normal. Each of stock returns has mean 10% and standard deviation 15%.

(a) Find the probability of a loss (a return less than zero) if you invest your capital in one of the 100 stocks.

Solution.

Let X denote the return of a portfolio that is fully invested in a single stock. We have X~N(10, 15^2). We need P(X $<$ 0).

```
mu_X = 10 # expected return of each stock
sigma_X = 15 # standard deviation of return of each stock
p_X_less0 = pnorm(0, mu_X, sigma_X)
print(paste0("P(X < 0) = ", p_X_less0))
```

```
## [1] "P(X < 0) = 0.252492537546923"
```

(b) Find the mean and standard deviation of return (in percentage points) of a portfolio that is evenly invested (half and half) in two of the 100 stocks.

Solution.

Let Y denote the return of a portfolio that is evenly invested in two stocks. We have Y~N(E(Y), var(Y)) where E(Y) = 0.5 E(X) + 0.5 E(X) = E(X) = 10% and var(Y) = $(0.5)^2$ var(X) + $(0.5)^2$ var(X). We need std(Y).

```
mu_Y = mu_X
var_Y = 0.5^2 * sigma_X^2 + 0.5^2 * sigma_X^2
sigma_Y = sqrt(var_Y)
print(paste0("std(Y) = ", sigma_Y, "%"))
```

```
## [1] "std(Y) = 10.6066017177982%"
```

(c) Find the probability of a loss (a return less than zero) for a portfolio that is evenly invested (half and half) in two of the 100 stocks.

Solution.

We need P(Y $<$ 0).

```
p_Y_less0 = pnorm(0, mu_Y, sigma_Y)
print(paste0("P(Y < 0) = ", p_Y_less0))
```

```
## [1] "P(Y < 0) = 0.17288929307558"
```

(d) Find the mean and standard deviation of return (in percentage points) of a portfolio that is evenly invested (fractions of 1/100 each) in the 100 stocks.

Solution.

Let Z denote the return of a portfolio that is evenly invested in the 100 stocks. We have Z~N(E(Z), var(Z)) where E(Z) = 0.01 E(X) + $\ldots$ + 0.01 E(X) = E(X) = 10% and var(Z) = $(0.01)^2$ var(X) + $\ldots$ + $(0.01)^2$ var(X). We need std(Z).

```
mu_Z = mu_X
var_Z = 100 * 0.01^2 * sigma_X^2
sigma_Z = sqrt(var_Z)
print(paste0("std(Z) = ", sigma_Z, "%"))
```

```
## [1] "std(Z) = 1.5%"
```

(e) Find the probability of a loss (a return less than zero) for a portfolio that is evenly invested (fractions of 1/100 each) in the 100 stocks.

Solution.

We need P(Z $<$ 0).

```
p_Z_less0 = pnorm(0, mu_Z, sigma_Z)
print(paste0("P(Z < 0) = ", p_Z_less0))
```

```
## [1] "P(Z < 0) = 1.3083924686053e-11"
```

2. (Sample Mean)

In a particular year, the rates (in percentage points) of return of U.S. common stock mutual funds had a normal distribution with a mean of 14.9 and a standard deviation of 6.4. Suppose you take a random sample of sixteen of these mutual funds.

```
n1 = 16 # sample size
mu_rate = 14.9 # population mean
sigma_rate = 6.4 # population standard deviatoin
```

(a) What are the mean and standard deviation (in percentage points) of the sample mean of the rates of return of the sixteen mutual funds?

Solution. We have a sample of size n = 16. Hence the sample mean $\bar{X}$ is normal with mean 14.9 and standard deviation

```
sigma_Xbar_rate = sigma_rate / sqrt(n1)
print(sigma_Xbar_rate)
```

```
## [1] 1.6
```

In other words, $\bar{X}$~N(14.9, 2.56).

(b) What is the probability that the sample mean rate of return is more than 18.0?

Solution. We need P($\bar{X}$ $>$ 18.0) = 1 - P($\bar{X}$ $<=$ 18.0).

```
prob_b = 1 - pnorm(18.0, mean = mu_rate, sd = sigma_Xbar_rate)
print(paste0("P(Xbar > 18.0) = ", prob_b))
```

```
## [1] "P(Xbar > 18.0) = 0.0263421266891415"
```

(c) What is the probability that the sample mean rate of return is between 11.8 and 18.0?

Solution. We need P(11.8 $<$ $\bar{X}$ $<$ 18.0) = P($\bar{X}$ $<$ 18.0) - P($\bar{X}$ $<=$ 11.8).

```
prob_c = pnorm(18.0, mean = mu_rate, sd = sigma_Xbar_rate) -
         pnorm(11.8, mean = mu_rate, sd = sigma_Xbar_rate)
print(paste0("P(11.8 < Xbar < 18.0) = ", prob_c))
```

```
## [1] "P(11.8 < Xbar < 18.0) = 0.947315746621717"
```

(d) Find the rate R (in percentage points) such that the sample mean rate of return is less than R with probability 0.25.

Solution. The rate R is the 0.25-th quantile, that is, P($\bar{X}$ $<$ R) = 0.25$.

```
R = qnorm(0.25, mean = mu_rate, sd = sigma_Xbar_rate)
print(paste0("R = ", R))
```

```
## [1] "R = 13.8208163996863"
```

(e) What is the probability that the sample mean rate of return differs from the population mean by more than 2 percentage points?

Solution. We want

$$\begin{aligned} P(|\bar{X} - 14.9| > 2) &= 1 - P(-2 < \bar{X} - 14.9 < 2) \\ &= 1 - P(12.9 < \bar{X} < 16.9) \\ &= 1 - [P(\bar{X} < 16.9) - P(\bar{X} \leq 12.9)]. \end{aligned}$$

```
prob_e = 1 - pnorm(16.9, mean = mu_rate, sd = sigma_Xbar_rate) +
             pnorm(12.9, mean = mu_rate, sd = sigma_Xbar_rate)
print(paste0("P(|Xbar - 14.9| > 2) = ", prob_e))
```

```
## [1] "P(|Xbar - 14.9| > 2) = 0.211299547333711"
```

(f) Suppose the random sample includes more than sixteen funds. Without doing any calculations, determine whether the probability that the sample mean rate of return differs from the population mean by more than 2 percentage points is larger or smaller than that found in (e).

Solution. Increasing the sample size, decreases the standard deviation of the sample mean distribution without changing its mean. Thus the density of the bell-shaped distribution increases in the center. Hence the probability that the sample mean rate of return differs from the population mean by more than 2 percentage points, P($|\bar{X}$ - 14.9$|$ $>$ 2), is smaller than that found in part (e).

(g) Find a symmetric interval around the population mean of 14.9 that includes 95% of the sample means based on samples of 16 mutual funds. What is this interval called?

Solution. Based on part (a), $\bar{X}$ is normally distributed with mean 14.9 and standard deviation 1.6. A symmetric 95% interval around the population mean of 14.9 excludes 2.5% of the probability distribution from both sides. Therefore, finding L and U such that P($\bar{X}$ $<=$ L) = 0.025 and P($\bar{X}$ $<=$ U) = 0.975 will give us the desired interval [L,U].

```
L95_rate = qnorm(0.025, mean = mu_rate, sd = sigma_Xbar_rate)
U95_rate = qnorm(0.975, mean = mu_rate, sd = sigma_Xbar_rate)
print(paste0("95% Confidence Interval: [", L95_rate, ", ", U95_rate,"]"))
```

```
## [1] "95% Confidence Interval: [11.7640576247359, 18.0359423752641]"
```

This is called the **acceptance interval** for the sample mean rate of return. Therefore, we could alternatively calculate the 95% acceptance interval using $\mu \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

```r
L95_rate2 = mu_rate - qnorm(0.975) * sigma_rate / sqrt(n1)
U95_rate2 = mu_rate + qnorm(0.975) * sigma_rate / sqrt(n1)
print(paste0("95% Confidence Interval: [", L95_rate2, ", " ,U95_rate2,"]"))
```

```
## [1] "95% Confidence Interval: [11.7640576247359, 18.0359423752641]"
```

3. (Confidence Interval for Population Mean with Known Variance)

A safety officer is concerned about speeds on a certain section of the New Jersey Turnpike. He records the speeds of 40 cars on a Saturday afternoon. The worksheet "HighwaySpeeds" in the Excel file *MSBAProblemSet4.xlsx* contains the data. Assume that the population standard deviation is 5 mph.

```r
n2 = 40 # sample size
sigma = 5 # population standard deviation
```

(a) Use the dataset to compute the sample mean of speeds.

Solution.

```r
#install.packages("readxl") # Run this line if readxl is not installed.
library(readxl)
```

```r
Highway_Speeds = read_excel("MSBAPracticeSet4.xlsx", sheet = "HighwaySpeeds")
Xbar_speed = mean(Highway_Speeds$Speed)
print(paste0("Sample mean of speeds = ", Xbar_speed))
```

```
## [1] "Sample mean of speeds = 66"
```

(b) Calculate the standard error of the sample mean of speeds.

Solution. The standard error of the sample mean is given by $\frac{\sigma}{\sqrt{n}}$.

```r
SE_speed = sigma / sqrt(n2)
print(paste0("Standard error of the sample mean of speeds = ", SE_speed))
```

```
## [1] "Standard error of the sample mean of speeds = 0.790569415042095"
```

(c) For the 90% confidence interval, determine the margin of error in estimating the mean speed of all cars on that section of the turnpike. (Remember that the margin of error is half the width of the whole confidence interval.)

Solution. The margin of error of the $(1-\alpha)$-confidence interval for the population mean is given by $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

```r
MOE90_speed = qnorm(0.95) * SE_speed
print(paste0("MOE of the 90% CI = ", MOE90_speed))
```

```
## [1] "MOE of the 90% CI = 1.30037096968889"
```

(d) What is the smallest sample size that would be required to reduce the margin of error of the 90% confidence interval by 25% (i.e., to three quarters of its current size)? Does the answer depend on the confidence level?

Solution. We are looking for the smallest integer m such that

$$
\begin{aligned}
z_{\alpha/2}\frac{\sigma}{\sqrt{m}} &\leq 0.75 z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \\
\frac{1}{\sqrt{m}} &\leq 0.75\frac{1}{\sqrt{n}} \\
\sqrt{m} &\geq \frac{1}{0.75}\sqrt{n} \\
m &\geq \frac{1}{0.75^2}n
\end{aligned}
$$

```
m = ceiling(1 / 0.75^2 * n2)
print(paste0("m = ", m))
```

```
## [1] "m = 72"
```

The answer does not depend on the confidence level since $z_{\alpha/2}$ is being canceled out in the inequality.

(e) Construct the 95% confidence interval for the mean speed of all cars on that section of the turnpike.

Solution. The $(1-\alpha)$-confidence interval for the population mean is given by $\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

```
L95_speed = Xbar_speed - qnorm(0.975) * SE_speed
U95_speed = Xbar_speed + qnorm(0.975) * SE_speed
print(paste0("95% Confidence Interval = [", L95_speed, ", ", U95_speed,"]"))
```

```
## [1] "95% Confidence Interval = [64.4505124192386, 67.5494875807614]"
```

(f) Suppose the same sample mean had been obtained but with a larger sample of cars. State, without calculation, if the 95% confidence interval for the population mean would be wider or narrower than that in part (e).

Solution. The confidence interval would be narrower than that in part (e). The width of the $(1-\alpha)$-confidence interval is twice its margin of error, i.e., $2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. Thus the width of the confidence interval decreases as the sample size increases. In other words, as the sample size increases, the confidence interval becomes narrower.

46-880 Introduction to Probability and Statistics

Problem Set 5 Solutions, mini-1 2022

1. (Confidence Interval for Population Mean with Known Variance)

A safety officer is concerned about speeds on a certain section of the New Jersey Turnpike. He records the speeds of 40 cars on a Saturday afternoon. The worksheet "HighwaySpeeds" in the Excel file *PracticeSet5.xlsx* contains the data. Assume that the population standard deviation is 5 mph.

```r
n2 = 40 # sample size
sigma = 5 # population standard deviation
```

(a) Use the dataset to compute the sample mean of speeds.

Solution.

```r
#install.packages("readxl") # Run this line if readxl is not installed.
library(readxl)
```

```r
Highway_Speeds = read_excel("PracticeSet5.xlsx", sheet = "HighwaySpeeds")
Xbar_speed = mean(Highway_Speeds$Speed)
print(paste0("Sample mean of speeds = ", Xbar_speed))
```

```
## [1] "Sample mean of speeds = 66"
```

(b) Calculate the standard error of the sample mean of speeds.

Solution. The standard error of the sample mean is given by $\frac{\sigma}{\sqrt{n}}$.

```r
SE_speed = sigma / sqrt(n2)
print(paste0("Standard error of the sample mean of speeds = ", SE_speed))
```

```
## [1] "Standard error of the sample mean of speeds = 0.790569415042095"
```

(c) For the 90% confidence interval, determine the margin of error in estimating the mean speed of all cars on that section of the turnpike. (Remember that the margin of error is half the width of the whole confidence interval.)

Solution. The margin of error of the $(1-\alpha)$-confidence interval for the population mean is given by $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

```r
MOE90_speed = qnorm(0.95) * SE_speed
print(paste0("MOE of the 90% CI = ", MOE90_speed))
```

```
## [1] "MOE of the 90% CI = 1.30037096968889"
```

(d) What is the smallest sample size that would be required to reduce the margin of error of the 90% confidence interval by 25% (i.e., to three quarters of its current size)? Does the answer depend on the confidence level?

Solution. We are looking for the smallest integer m such that

$$\begin{aligned} z_{\alpha/2}\frac{\sigma}{\sqrt{m}} &\leq 0.75 z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \\ \frac{1}{\sqrt{m}} &\leq 0.75\frac{1}{\sqrt{n}} \\ \sqrt{m} &\geq \frac{1}{0.75}\sqrt{n} \\ m &\geq \frac{1}{0.75^2}n \end{aligned}$$

```
m = ceiling(1 / 0.75^2 * n2)
print(paste0("m = ", m))
```

```
## [1] "m = 72"
```

The answer does not depend on the confidence level since $z_{\alpha/2}$ is being canceled out in the inequality.

(e) Construct the 95% confidence interval for the mean speed of all cars on that section of the turnpike.

Solution. The $(1-\alpha)$-confidence interval for the population mean is given by $\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

```
L95_speed = Xbar_speed - qnorm(0.975) * SE_speed
U95_speed = Xbar_speed + qnorm(0.975) * SE_speed
print(paste0("95% Confidence Interval = [", L95_speed, ", ", U95_speed,"]"))
```

```
## [1] "95% Confidence Interval = [64.4505124192386, 67.5494875807614]"
```

(f) Suppose the same sample mean had been obtained but with a larger sample of cars. State, without calculation, if the 95% confidence interval for the population mean would be wider or narrower than that in part (e).

Solution. The confidence interval would be narrower than that in part (e). The width of the $(1-\alpha)$-confidence interval is twice its margin of error, i.e., $2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. Thus the width of the confidence interval decreases as the sample size increases. In other words, as the sample size increases, the confidence interval becomes narrower.

2. (Confidence Interval for Population Mean with Unknown Variance)

Many of today's leading companies, including Google, Microsoft, and Facebook, are based on technologies developed within universities. A business school professor believes that a university's research expenditure (*Research* in $ millions) and the age of its technology transfer office (*Duration* in years) are major factors that enhance innovation. She wants to know what the average values are for the *Research* and *Duration* variables. She collects data from 143 universities on these variables. The worksheet "'Startups" in the Excel file *PracticeSet5.xlsx* contains the data.

```
n3 = 143 # sample size
```

(a) Use the dataset to compute the sample mean of *Research* and *Duration*.

Solution.

```
Startups = read_excel("PracticeSet5.xlsx", sheet = "Startups")
Xbar_Research = mean(Startups$Research)
Xbar_Duration = mean(Startups$Duration)
print(paste0("Sample mean of Research = ", Xbar_Research, ", Sample mean of Duration = ", Xbar_Duration
```

```
## [1] "Sample mean of Research = 302.466887412587, Sample mean of Duration = 20.5034965034965"
```

(b) Use the dataset to compute the sample standard deviation of *Research* and *Duration*.

Solution.

```
s_Research = sd(Startups$Research)
s_Duration = sd(Startups$Duration)
print(paste0("Sample standard deviation of Research = ", s_Research, ", Sample standard deviation of Du
```

```
## [1] "Sample standard deviation of Research = 429.6348007808, Sample standard deviation of Duration =
```

(c) What is the standard error of the sample mean of *Research*?

Solution. The standard error of the sample mean is given by $\frac{s}{\sqrt{n}}$.

```
SE_Research = s_Research / sqrt(n3)
print(paste0("Standard error of the sample of Research = ", SE_Research))
```

```
## [1] "Standard error of the sample of Research = 35.9278669373932"
```

(d) Calculate the margin of error of the 99% confidence interval for the population mean of *Duration*.

Solution. The margin of error of the $(1-\alpha)$-confidence interval for the population mean is given by $t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$.

```
SE_Duration = s_Duration / sqrt(n3)
MOE99_Duration = qt(0.995, n3-1) * SE_Duration
print(paste0("MOE of the 99% CI for Duration = ", MOE99_Duration))
```

```
## [1] "MOE of the 99% CI for Duration = 2.77553880136709"
```

(e) Construct and interpret 95% confidence interval for the mean research expenditure of all universities.

Solution. The $(1-\alpha)$-confidence interval for the population mean is given by $\bar{X} \pm t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$.

```
L95_Research = Xbar_Research - qt(0.975,n3-1) * SE_Research
U95_Research = Xbar_Research + qt(0.975,n3-1) * SE_Research
print(paste0("95% Confidence Interval for Research = [", L95_Research, ", ", U95_Research,"]"))
```

```
## [1] "95% Confidence Interval for Research = [231.44428507297, 373.489489752205]"
```

With probability 0.95, this interval brackets the mean research expenditure of all universities.

(f) Construct and interpret 95% confidence interval for the mean age of the technology transfer office of all universities.

Solution. The $(1-\alpha)$-confidence interval for the population mean is given by $\bar{X} \pm t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$.

```
L95_Duration = Xbar_Duration - qt(0.975,n3-1) * SE_Duration
U95_Duration = Xbar_Duration + qt(0.975,n3-1) * SE_Duration
print(paste0("95% Confidence Interval for Duration = [", L95_Duration, ", ", U95_Duration,"]"))
```

```
## [1] "95% Confidence Interval for Duration = [18.4020274683404, 22.6049655386526]"
```

With probability 0.95, this interval brackets the mean age of technology transfer office of all universities.

3. (Confidence Interval for Population Proportion)

A study examined “sidewalk rage” in an attempt to find insight into anger’s origins and offer suggestions for anger-management treatments. “Sidewalk ragers” tend to believe that pedestrians should behave in a certain way. One possible strategy for sidewalk ragers is to avoid walkers who are distracted by other activities such as smoking and tourism. Sample data were obtained from 50 pedestrians in Lower Manhattan. It was noted if the pedestrian was smoking (equaled 1 if smoking, 0 otherwise) or was a tourist (equaled 1 if tourist, 0 otherwise). The worksheet “Pedestrians” in the Excel file *PracticeSet5.xlsx* contains the data.

```
n4 = 50 # sample size
```

(a) Use the dataset to compute the sample proportion of pedestrians who smoke while walking and sample proportion of pedestrians who are tourists.

Solution.

```
Pedestrians = read_excel("PracticeSet5.xlsx", sheet = "Pedestrians")
phat_Smoking = mean(Pedestrians$Smoking)
phat_Tourist = mean(Pedestrians$Tourist)
print(paste0("phat for smoking = ", phat_Smoking, ", phar for tourist = ", phat_Tourist))
```

```
## [1] "phat for smoking = 0.2, phar for tourist = 0.4"
```

(b) What is the standard error of the sample proportion of pedestrians who smoke while walking?

Solution. The standard error of the sample proportion is given by $\frac{\sqrt{p(1-p)}}{\sqrt{n}}$.

```
SE_Smoking = sqrt(phat_Smoking * (1-phat_Smoking)) / sqrt(n4)
print(paste0("Standard error of the sample proportion for smoking = ", SE_Smoking))
```

```
## [1] "Standard error of the sample proportion for smoking = 0.0565685424949238"
```

(c) Calculate the margin of error of the 99% confidence interval for the population proportion of pedestrians who are tourists.

Solution. The margin of error of the $(1-\alpha)$-confidence interval for the population proportion is given by $z_{\alpha/2}\frac{\sqrt{p(1-p)}}{\sqrt{n}}$.

```
SE_Tourist = sqrt(phat_Tourist * (1-phat_Tourist)) / sqrt(n4)
MOE99_Tourist = qnorm(0.995) * SE_Tourist
print(paste0("MOE of the 99% CI for tourist = ", MOE99_Tourist))
```

```
## [1] "MOE of the 99% CI for tourist = 0.178458689014858"
```

(d) Suppose we want to find the 99% confidence interval for the population proportion of pedestrians who are tourists with margin of error at most 0.1. How large a sample is needed?

Hint: The largest value of $\hat{p}(1-\hat{p})$ for $\hat{p} \in [0,1]$ is attained at $\hat{p} = 0.5$.

Solution We have $\alpha = 1 - 0.99 = 0.01$ and we want m such that

$$
\begin{aligned}
MOE = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{m}} &= z_{0.005}\sqrt{\frac{\hat{p}(1-\hat{p})}{m}} \leq 0.1 \\
\Rightarrow \sqrt{m} &\geq \frac{z_{0.005}}{0.1}\sqrt{\hat{p}(1-\hat{p})} \\
\Rightarrow m &\geq \left(\frac{z_{0.005}}{0.1}\right)^2 \hat{p}(1-\hat{p})
\end{aligned}
$$

The largest possible value of $\hat{p}(1-\hat{p})$ is 0.25 so we should choose m so that

$$m \geq 0.25\left(\frac{z_{0.005}}{0.1}\right)^2$$

```
m = ceiling((qnorm(0.995) / 0.1)^2 * 0.25)
print(paste0("m = ", m))
```

```
## [1] "m = 166"
```

(e) Construct and interpret the 95% confidence interval for the proportion of pedestrians in Lower Manhattan who smoke while walking.

Solution. The $(1-\alpha)$-confidence interval for the population proportion is given by $\bar{X} \pm z_{\alpha/2}\frac{\sqrt{p(1-p)}}{\sqrt{n}}$.

```
L95_Smoking = phat_Smoking - qnorm(0.975) * SE_Smoking
U95_Smoking = phat_Smoking + qnorm(0.975) * SE_Smoking
print(paste0("95% Confidence Interval for smoking = [", L95_Smoking, ", ", U95_Smoking,"]"))
```

```
## [1] "95% Confidence Interval for smoking = [0.0891276940520258, 0.310872305947974]"
```

With probability 0.95, this interval brackets the proportion of pedestrians in Lower Manhattan who smoke while walking.

(f) Construct and interpret the 95% confidence interval for the proportion of pedestrians in Lower Manhattan who are tourists.

Solution. The $(1-\alpha)$-confidence interval for the population mean is given by $\bar{X} \pm z_{\alpha/2}\frac{\sqrt{p(1-p)}}{\sqrt{n}}$.

```
L95_Tourist = phat_Tourist - qnorm(0.975) * SE_Tourist
U95_Tourist = phat_Tourist + qnorm(0.975) * SE_Tourist
print(paste0("95% Confidence Interval for toursit = [", L95_Tourist, ", ", U95_Tourist,"]"))
```

```
## [1] "95% Confidence Interval for toursit = [0.264209711910859, 0.535790288089141]"
```

With probability 0.95, this interval brackets the proportion of pedestrians in Lower Manhattan who are tourists.

4. (One-sided vs. Two-sided Hypothesis Testing)

A pharmaceutical manufacturer is concerned that the impurity concentration in pills should not exceed 3%. A random sample of 64 pills from a production run was checked. The sample mean impurity concentration of this sample was 3.07% and the sample standard deviation was 0.4%.

(a) Test the null hypothesis $H_0 : \mu = 3\%$ against the two-sided alternative $H_1 : \mu \neq 3\%$ at the at the 5% significance level.

Solution We have $\bar{X} = 3.07$ and $S = 0.4$. To test the null hypothesis $H_0 : \mu = 3\%$ against the two-sided alternative $H_1 : \mu \neq 3\%$ compute the t-statistic $\frac{3.07-3}{0.4/\sqrt{64}}$

```
t_statistic = (3.07-3)/(0.4/64^0.5)
print(t_statistic)
```

```
## [1] 1.4
```

We need to compare with the critical t-value $t_{n-1,\alpha/2} = t_{63,0.025} = 1.9983$.

```
critical_tvalue = qt(0.975,63)
print(critical_tvalue)
```

```
## [1] 1.998341
```

We do not reject the null hypothesis because 1.4 = absolute value of t-statistic is not larger than 1.998 = critical t-value.

(b) Determine the t-statistic for the test in part (a).

Solution We already computed the t-statistic 1.4 it in part (a). Here it is again:

```
t_statistic = (3.07-3)/(0.4/64^0.5)
print(t_statistic)
```

```
## [1] 1.4
```

(c) Find the p-value for the test in part (a).

Solution The p-value is $P(|t_{63}| \geq |\text{t-stat}|)$. That is

```
p_value = 2*(1-pt(1.4,63))
print(p_value)
```

```
## [1] 0.1664188
```

(d) Repeat the above for the one-sided hypothesis $H_0 : \mu \leq 3\%$ against the one-sided alternative $H_1 : \mu > 3\%$.

Solution Since this is a one-sided test, we reject if t-stat $> t_{n-1,\alpha} = t_{63,0.05}$. The latter critical value is

```
oneside_critical_tvalue = qt(0.95,63)
print(oneside_critical_tvalue)
```

```
## [1] 1.669402
```

Since 1.4 is not larger than 1.6694, we do not reject the above one-sided hypothesis.

(e) In the context of this problem, explain why a one-sided alternative hypothesis is more appropriate than a two-sided alternative.

Solution Because in this case what matters is that the impurity concentration does not exceed 3%.

46-880 Introduction to Probability and Statistics

Solution to Problem Set 6, mini-1 2023

1. (Inference about Correlation)

The worksheet `Drinks` in the Excel file `ProblemSet6.xlsx` shows data for Per Capita Real (inflation adjusted) Income (`rpinc`), Real Price of Coffee (`rpcofe`), Real Price of Carbonated Beverages (`rpcarb`) for 1955 through 2015.

(a) Find the sample covariance of (`rpinc`, `rpcofe`) and of (`rpinc`, `rpcarb`).

Solution.

```
library(readxl)
Drinks = read_excel("MSBAProblemSet6.xlsx",sheet = "Drinks")
print(paste("cov(rpcinc,rpcofe) = ",cov(Drinks$rpcinc,Drinks$rpcofe)))
```

```
## [1] "cov(rpcinc,rpcofe) =  -46.6217030248449"
```

```
print(paste("cov(rpcinc,rpcarb) = ",cov(Drinks$rpcinc,Drinks$rpcarb)))
```

```
## [1] "cov(rpcinc,rpcarb) =  -308.24041357135"
```

(b) Find the sample correlation of (`rpinc`, `rpcofe`) and of (`rpinc`, `rpcarb`).

Solution.

```
print(paste("correl(rpcinc,rpcofe) = ",cor(Drinks$rpcinc,Drinks$rpcofe)))
```

```
## [1] "correl(rpcinc,rpcofe) =  -0.232763180505237"
```

```
print(paste("correl(rpcinc,rpcarb) = ",cor(Drinks$rpcinc,Drinks$rpcarb)))
```

```
## [1] "correl(rpcinc,rpcarb) =  -0.942568341770843"
```

(c) Consider the pair (`rpinc`, `rpcofe`). Test, at the 5% significance level, the null hypothesis that the correlation between `rpinc` and `rpcofe` is zero against the alternative of non-zero correlation. In other words, test the hypothesis $H_0 : \rho = 0$, against the alternative $H_1 : \rho \neq 0$.

Solution. We can do it two ways. Compte t-stat and compare with critical t-value

```
r = cor(Drinks$rpcinc,Drinks$rpcofe)
tstat = r*sqrt(59/(1-r^2))
print(paste("tstat=",tstat))
```

```
## [1] "tstat= -1.83838193250958"
```

```
print(paste("critical t-value at 5% significance level =",qt(0.975,59)))
```

```
## [1] "critical t-value at 5% significance level = 2.00099537808827"
```

Thus we do not reject at the 5% significance level. Alternatively, we can use `cor.test`:

```
cor.test(Drinks$rpcinc,Drinks$rpcofe)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Drinks$rpcinc and Drinks$rpcofe
## t = -1.8384, df = 59, p-value = 0.07104
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.45775292  0.02024416
## sample estimates:
##        cor
## -0.2327632
```

Again we reach the same conclusion: do not reject, as the p-value is larger than 5%.

(d) Repeat part (c) for the pair (`rpinc`, `rpcarb`).

Solution. We can do it two ways. Compte t-stat and compare with critical t-value

```
r = cor(Drinks$rpcinc,Drinks$rpcarb)
tstat = r*sqrt(59/(1-r^2))
print(paste("tstat=",tstat))
```

```
## [1] "tstat= -21.6757944239505"
```

```
print(paste("critical t-value at 5% significance level =",qt(0.975,59)))
```

```
## [1] "critical t-value at 5% significance level = 2.00099537808827"
```

Thus we do reject at the 5% significance level. Alternatively, we can use `cor.test`:

```
cor.test(Drinks$rpcinc,Drinks$rpcarb)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Drinks$rpcinc and Drinks$rpcarb
## t = -21.676, df = 59, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9652735 -0.9057301
## sample estimates:
##        cor
## -0.9425683
```

Again we reach the same conclusion: reject, as the p-value is smaller than 5%.

(e) Find the sample correlation of (`rpcofe`, `rpcarb`)? Does the sign of the correlation support the fact that carbonated beverages are considered a substitute good for coffee?

Solution.

```
print(paste("correl(rpcofe,rpcarb) = ",cor(Drinks$rpcofe,Drinks$rpcarb)))
```

```
## [1] "correl(rpcofe,rpcarb) =  0.111269146897356"
```

The correlation is low so it does not support that fact. Indeed, we do not reject the null hypothesis that the correlation between the price of carbonated beverages and the price of coffee is zero at the 5% significance level:

```r
r = cor(Drinks$rpcofe,Drinks$rpcarb)
tstat = r*sqrt(59/(1-r^2))
print(paste("tstat=",tstat))
```

```
## [1] "tstat= 0.860014962299231"
```

```r
print(paste("critical t-value at 5% significance level =",qt(0.975,59)))
```

```
## [1] "critical t-value at 5% significance level = 2.00099537808827"
```

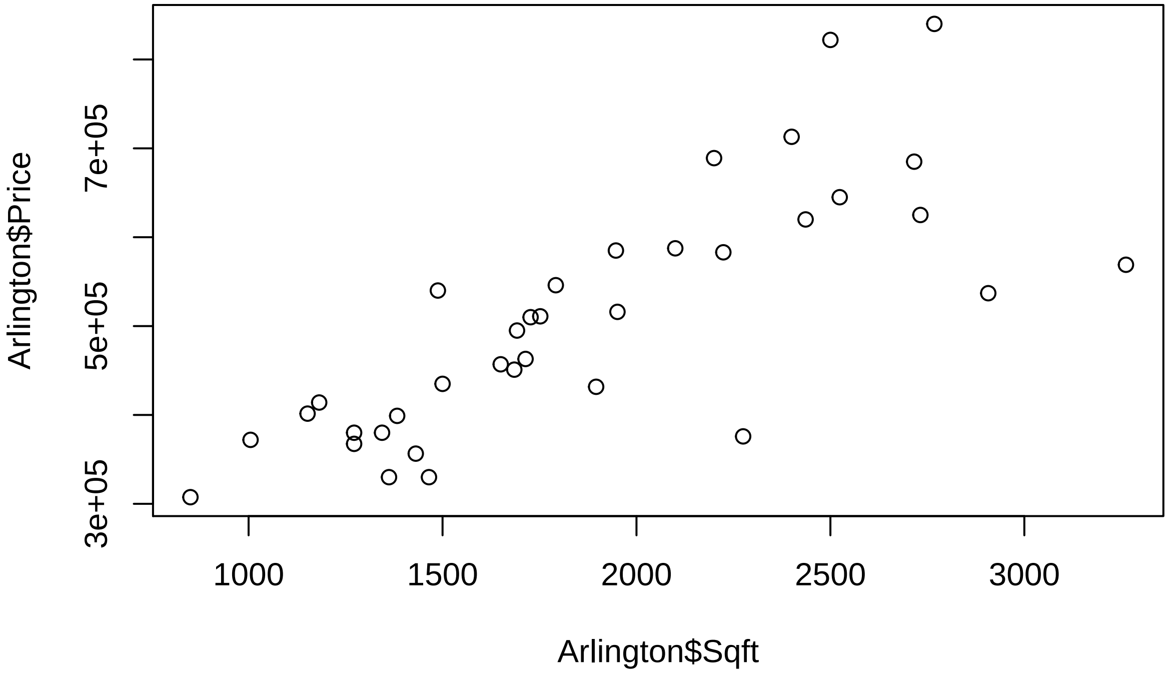
2. (Simple Linear Regression)

A realtor in Arlington, Massachusetts, is analyzing the relationship between the sale price of a home (Price in $), its square footage (Sqft), the number of bedrooms (Beds), and the number of bathrooms (Baths). She collects data on 36 sales in Arlington in the first quarter of 2009 for the analysis. The worksheet `Arlington Homes` in the Excel file `MSBAProblemSet6.xlsx` shows the corresponding data.

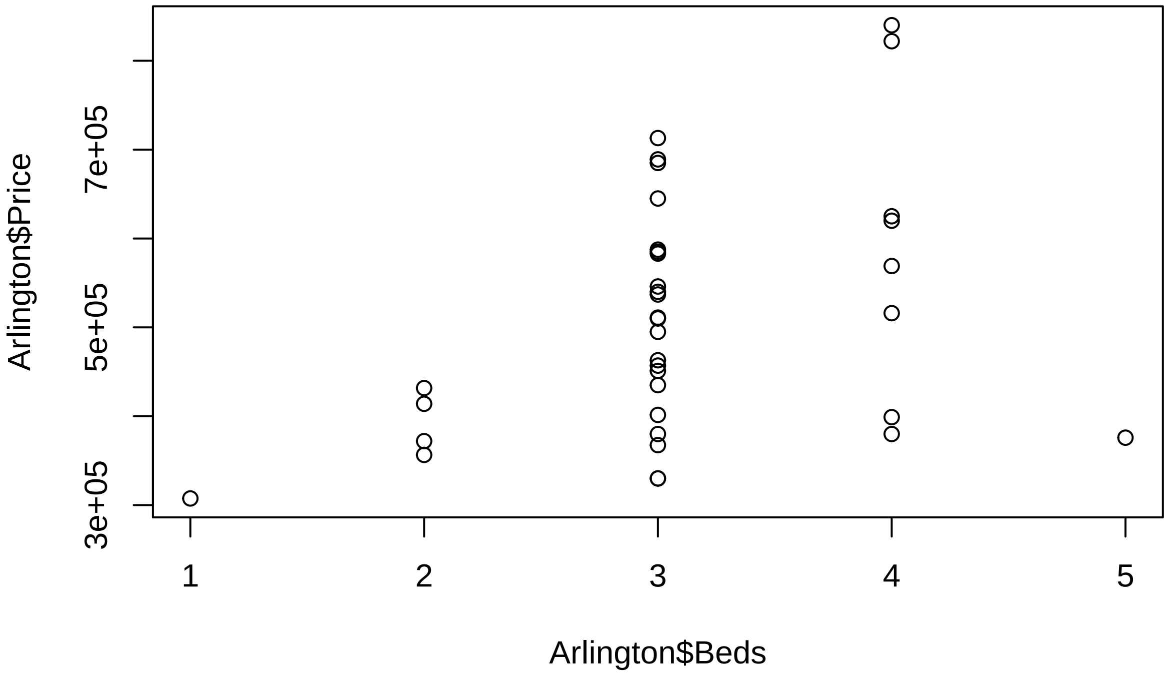
(a) Prepare three scatter plots illustrating Price vs. Sqft, Beds, or Baths, separately. In each case, does the plot provide evidence that the independent variable has a positive effect on Price?

Solution.

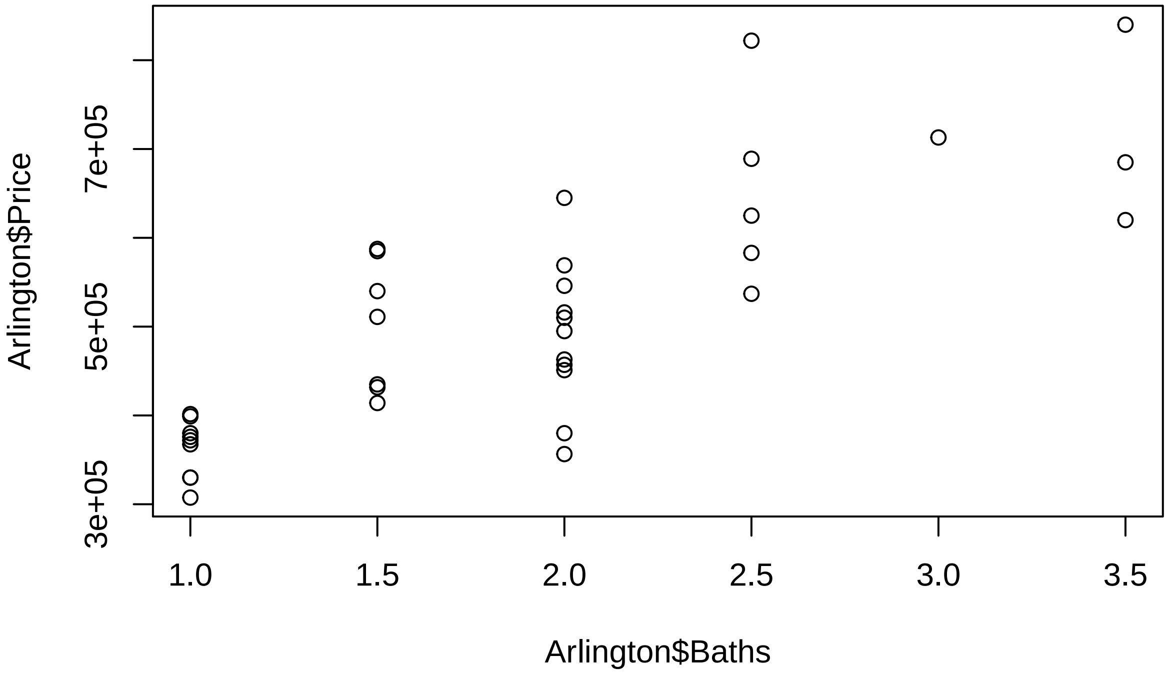
```r
library(readxl)
Arlington = read_excel("MSBAProblemSet6.xlsx",sheet = "Arlington Homes")
plot(Arlington$Sqft, Arlington$Price)
```



```r
plot(Arlington$Beds, Arlington$Price)
```



```
plot(Arlington$Baths, Arlington$Price)
```



In each case, indeed the plot suggests that the independent variable has a positive effect on Price.

(b) Estimate three simple linear regression models that use Price as the response variable with Sqft, Beds, or Baths as the explanatory variable.

Solution.

```
model1 = lm(Price~Sqft, Arlington)
print(summary(model1))
```

```
##
```

```
## Call:
## lm(formula = Price ~ Sqft, data = Arlington)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -202981  -31801    3719   29697  202812
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 171331.46   48909.96   3.503  0.00131 **
## Sqft           179.14      24.89   7.198 2.51e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 87110 on 34 degrees of freedom
## Multiple R-squared:  0.6038, Adjusted R-squared:  0.5921
## F-statistic: 51.81 on 1 and 34 DF,  p-value: 2.513e-08
```

```
model2 = lm(Price~Beds, Arlington)
print(summary(model2))
```

```
##
## Call:
## lm(formula = Price ~ Beds, data = Arlington)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -257518  -68056   -3416   64408  273243
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   300114      93139   3.222   0.0028 **
## Beds           66661      29131   2.288   0.0285 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 128800 on 34 degrees of freedom
## Multiple R-squared:  0.1335, Adjusted R-squared:  0.108
## F-statistic: 5.236 on 1 and 34 DF,  p-value: 0.02846
```

```
model3 = lm(Price~Baths, Arlington)
print(summary(model3))
```

```
##
## Call:
## lm(formula = Price ~ Baths, data = Arlington)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -169418  -51922  -10507   40589  222420
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   231272      38164   6.060 7.20e-07 ***
## Baths         147323      18967   7.767 4.89e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 83080 on 34 degrees of freedom
## Multiple R-squared:  0.6396, Adjusted R-squared:  0.629
## F-statistic: 60.33 on 1 and 34 DF,  p-value: 4.892e-09
```

(c) Which model provides a better fit for the price? Explain.

Solution.

The third model that uses `Baths` as the independent variable provides the best fit for `Price`. This one has the highest R-square and both coefficients are statistically significant at a very low significance level 10^{-7}.

(d) State, without using the data and just based on the regression results from part (b), the sample correlation between Price and each of the independent variables.

Solution.

We just have to take the square root of R-square:

```
print(paste("correl(Price,Sqft) = ",sqrt(0.6038)))
```

```
## [1] "correl(Price,Sqft) =  0.777045687202496"
```

```
print(paste("correl(Price,Beds) = ",sqrt(0.1335)))
```

```
## [1] "correl(Price,Beds) =  0.365376518128902"
```

```
print(paste("correl(Price,Baths) =",sqrt(0.6396)))
```

```
## [1] "correl(Price,Baths) = 0.799749960925288"
```

As a sanity check, we could also compute the sample correlations directly (even though we did not need to):

```
print(paste("correl(Price,Sqft) = ",cor(Arlington$Price,Arlington$Sqft)))
```

```
## [1] "correl(Price,Sqft) =  0.777020674762787"
```

```
print(paste("correl(Price,Beds) = ",cor(Arlington$Price,Arlington$Beds)))
```

```
## [1] "correl(Price,Beds) =  0.365316173785397"
```

```
print(paste("correl(Price,Baths) = ",cor(Arlington$Price,Arlington$Baths)))
```

```
## [1] "correl(Price,Baths) =  0.799730122334478"
```

(e) For each of the simple regression models in part (b), determine if the independent variable is statistically significant at the 5% significance level. In other words, for each regression model, test, at the 5% significance level, against the two sided alternative, the null hypothesis that the slope coefficient is zero.

Solution.

In all three models the independent variable is statistically significant at the 5% significance level.

3. (Hypothesis Testing for Regression Coefficients)

The vice president of purchasing for a large national retailer has asked you to prepare an analysis of retail sales by state. He wants to know if either the employment rate or the per capita disposable income are related to per capita retail sales. The worksheet `Economic Activity` in the Excel file `MSBAProblemSet6.xlsx` shows the corresponding data.

(a) Develop two regression models to predict per capita retail sales, that is, per capita retail sales is the response variable. For the first model use employment rate as the predictor variable. For the second model use per capita disposable income as the predictor variable.

Solution.

```
library(readxl)
Economic = read_excel("MSBAProblemSet6.xlsx",sheet = "Economic Activity")
model1 = lm(Economic$`Per Capita Retail Sales (dollars)`~Economic$`Employment Rate (percentage)`)
print(summary(model1))
```

```
##
## Call:
## lm(formula = Economic$`Per Capita Retail Sales (dollars)` ~ Economic$`Employment Rate (percentage)`)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5451 -0.8645  0.0628  1.0372  5.2449
##
## Coefficients:
##                                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                              1.04284    4.21949   0.247   0.8058
## Economic$`Employment Rate (percentage)`  0.20582    0.06638   3.101   0.0032 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.929 on 49 degrees of freedom
## Multiple R-squared:  0.164,  Adjusted R-squared:  0.147
## F-statistic: 9.614 on 1 and 49 DF,  p-value: 0.003198
```

```
model2 = lm(`Per Capita Retail Sales (dollars)`~`Per Capita Disposable Income (dollars)`,Economic)
print(summary(model2))
```

```
##
## Call:
## lm(formula = `Per Capita Retail Sales (dollars)` ~ `Per Capita Disposable Income (dollars)`,
##     data = Economic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6358 -1.2793 -0.2379  0.9060  6.3159
##
## Coefficients:
##                                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                               1.514e+01  1.863e+00   8.126 1.23e-10
## `Per Capita Disposable Income (dollars)` -3.646e-05  6.466e-05  -0.564    0.575
##
## (Intercept)                              ***
## `Per Capita Disposable Income (dollars)`
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.102 on 49 degrees of freedom
## Multiple R-squared:  0.006448,   Adjusted R-squared:  -0.01383
## F-statistic: 0.318 on 1 and 49 DF,  p-value: 0.5754
```

(b) Based on the first model, predict the effect of a one-percentage-point increase in employment rate on per capita retail sales. Similarly, based on the second model, predict the effect of a one-thousand dollar increase in per capita disposable income on per capital retail sales.

Solution.

A one-percentage-point increase in employment rate would lead to an increase of 0.2058 dollars in per capita retail sales.

(c) Compute a 95% confidence interval for the slope coefficients in each regression equation.

Solution.

```
print(confint(model1,level=0.95))
```

```
##                                               2.5 %    97.5 %
## (Intercept)                             -7.43654082 9.5222223
## Economic$`Employment Rate (percentage)`  0.07242664 0.3392099
```

```
print(confint(model2,level=0.95))
```

```
##                                                  2.5 %       97.5 %
## (Intercept)                              11.3932461925 1.888003e+01
## `Per Capita Disposable Income (dollars)` -0.0001664015 9.347593e-05
```

(d) For each regression model, determine if the slope of the predictor variable is statistically significant at the 5% significance level. In other words, for each regression model, test, at the 5% significance level, test the null hypothesis that the slope coefficients is zero.

Solution.

For model 1 the slope is statistically significant at the 5% significance level because the above 95% confidence interval does not contain zero, or alternatively because the p-value is $0.0032 < 0.05$.

For model 2 the slope is not statistically significant at the 5% significance level because the above 95% confidence interval containd zero, or alternatively because the p-value is $0.575 > 0.05$.