Problem Set 3

Statistics 104

Due February 27, 2020 at 11:59 pm

Problem set policies. *Please provide concise, clear answers for each question. Note that only writing the result of a calculation (e.g., "SD = 3.3") without explanation is not sufficient. For problems involving* R*, be sure to include the code in your solution.*

Please submit your problem set via Canvas as a PDF, along with the R Markdown source file.

We encourage you to discuss problems with other students (and, of course, with the course head and the TFs), but you must write your final answer in your own words. Solutions prepared "in committee" are not acceptable. If you do collaborate with classmates on a problem, please list your collaborators on your solution.

Problem 1.

Let X be a random variable with the following probability mass function:

$X = x$	0	1	2	3
$P(X = x)$	0.10	0.20	0.30	0.40

a) Find $P(X \geq 2)$.

b) Find $P(X \geq 2 | X \geq 1)$.

c) Find $E(X)$.

d) Find Var(X).

Problem 2.

A man buys a racehorse for \$20,000 and enters it in two races. He hopes to make a profit from selling the horse after the two races. If the horse wins both races, its value will jump to \$100,000. If it wins one of the races, it will be worth \$50,000. If, however, it loses both races, it will be worth only \$10,000. The man believes there is a 20% chance that the horse will win the first race and a 30% chance it will win the second one.

a) Assuming the races are independent, calculate the man's expected profit and the standard deviation of the man's profit.

b) Is it reasonable to assume the races are independent? Explain your answer.

Problem 3.

Recently, mumps outbreaks have become more common, with many occurring among individuals 18-24 years of age living on college campuses. Two doses of the measles-mumps-rubella (MMR) vaccine are recommended for protection from mumps. Herd immunity refers to the proportion of individuals that must be immune to effectively prevent the spread of disease through a population. In order to prevent the spread of mumps, at least 96% of people in a community must have received two doses of the MMR vaccine.

a) Suppose that 94% of undergraduate students in the United States report having received two doses of the MMR vaccine. What is the probability that in one upperclassman House at Harvard, enough students are vaccinated to achieve herd immunity? There are approximately 400 students in any given House.

b) Calculate the probability that herd immunity is achieved in all 12 Houses.

c) Discuss the validity of the assumptions required to make the calculation in part i.

Problem 4.

Assume the annual returns on a stock portfolio are normally distributed with a mean of 14.7% and a standard deviation of 33%. A return of 0% indicates the value of the portfolio does not change.

a) What is the probability that in any given year the portfolio will lose money?

b) What is the probability that in any given year the portfolio will have at least a 50% return?

c) What is the probability that in any given year the portfolio will have a return between 25% and 75%?

d) Calculate the return value that marks off the lowest 10% of annual returns for this portfolio.

e) What is the probability that four of the next ten years will have a return greater than 50%? Comment on the validity of any assumptions required to make this calculation.

Problem 5.

An earnings announcement is an official statement of a company's profitability for a specific time period. Whether a stock beats earnings expectations or not typically influences share price. This problem considers the value of Facebook stock.

Suppose that Facebook beats earnings expectations 75% of the time. When the stock beats earnings expectations, the value of returns is normally distributed with mean 10% and standard deviation 5%. When the stock fails to beat earnings expectations, the value of returns is normally distributed with mean -5% and standard deviation of 8%.

Ahead of the earnings announcement, calculate the following probabilities:

a) The probability that the stock has a return greater than 5%.

b) The probability that the stock has a return less than 5%.

Problem 6.

This problem explores an interesting property of the sum of independent normal random variables via the use of the `rnorm( )` function. This function is used to draw a specified number of observations from a normal distribution and has the generic structure

```
rnorm(n, mean = 0, sd = 1)
```

where `n` is the number of observations, `mean` is the parameter μ, and `sd` is the parameter σ. For example, the following code draws 10 observations from the standard normal, Z.

```
set.seed(2019)  #set seed for pseudo-random sampling
rnorm(10, mean = 0, sd = 1)
```

```
##  [1]  0.7385227 -0.5147605 -1.6401813  0.9160368 -1.2674820  0.7382478
##  [7] -0.7826228  0.5092959 -1.4899391 -0.3191793
```

Suppose X is normally distributed with mean 15 and standard deviation 2, while Y is normally distributed with mean 10 and standard deviation 9.

a) Generate 10,000 observations each from X and Y (NOTE: It is *not* necessary to print out all the values as part of your solution). Using graphical and numerical summaries, confirm that the sets of observations appear normally distributed with the expected parameters.

b) Suppose we are interested in the distribution of the random variable $X + Y$.

   i. Using the simulated observations in part a), directly calculate $E(X+Y)$ and $\text{Var}(X+Y)$.

   ii. Plot the distribution of $X + Y$ and describe its shape.

c) Suppose we are interested in the distribution of the random variable $X - Y$. Using the simulated observations in part a), directly calculate $E(X-Y)$ and $\text{Var}(X-Y)$.

d) From theory, calculate $E(X+Y)$, $\text{Var}(X+Y)$, $E(X-Y)$, and $\text{Var}(X-Y)$. Compare the theoretical values to the answers from parts b) and c).

e) From the simulated observations, estimate $P(X < Y)$.

f) If two independent random variables X and Y are normally distributed, then their sum is also normally distributed. Calculate $P(X < Y)$ and compare this value to the estimate from part e).

Problem 7.

A large insurance agency services a number of customers who have purchased both a homeowner's policy and an automobile policy from the agency. For each type of policy, a deductible amount must be specified. For an automobile policy, the choices are \$100 and \$250, whereas for a homeowner's policy, the choices are \$0, \$100, and \$300.[1]

Suppose an individual with both types of policy is selected at random from the agency's files. Let X = the deductible amount on the auto policy, and Y = the deductible amount on the homeowner's policy.

Suppose the following table represents the joint distribution of X and Y:

	$Y = 0$	$Y = 100$	$Y = 300$
$X = 100$	0.05	0.23	0.20
$X = 250$	0.30	0.20	0.02

a) What are the mean and standard deviation of X?

b) What are the mean and standard deviation of Y?

c) What is the covariance between X and Y?

d) What is the correlation between X and Y?

e) Calculate $E(X+Y)$.

f) Calculate $\text{Var}(X+Y)$.

g) Are X and Y independent? Justify your answer.

h) What is the expected value of Y given $X = 250$?

[1] *Editorial note—this problem requires a lot of (tedious) algebra. However, the process of thinking through these calculations can be instructive for understanding distributions for pairs of random variables.*

Problem 8.

In target marketing, the goal is to target certain customers for promotions. A promotion might consist of mailing a special catalogue to a customer. Firms maintain large databases of information on their customers. One of the most useful variables is frequency of purchases; i.e., how often a customer makes a purchase.

A customer is randomly chosen from the record of existing customers and sent a special catalogue.

Let N represent a random variable that takes value 1 if a customer makes a new purchase and value 0 if otherwise. Let F be a random variable representing purchase frequency, where F takes on values $\{1, 2, 3, 4\}$.

- A value of $F = 1$ indicates that 1 purchase was made within the last year.
- A value of $F = 2$ indicates that 2 - 10 purchases were made within the last year.
- A value of $F = 3$ indicates that 11 - 20 purchases were made within the last year.
- A value of $F = 4$ indicates that more than 20 purchases were made within the last year.

The marketing research department has determined that the joint probability distribution of (N, F) is given by the following table:

	$N = 0$	$N = 1$
$F = 1$	0.08	0.02
$F = 2$	0.36	0.24
$F = 3$	0.10	0.10
$F = 4$	0.02	0.08

a) Calculate $P_{NF}(1, 2)$.

b) Calculate and interpret $p_N(1)$.

c) Calculate the marginal distribution of F, $P_F(f)$.

d) Calculate the conditional distribution of F given $N = 1$.

e) Calculate the conditional distribution of N given $F = 4$.

f) Calculate the conditional expectation $E(N|F = 4)$.

g) Consider the relationship between N and F.

   i. Would you expect N and F to be independent? Explain your answer.

   ii. Based on the joint probability distribution, are N and F independent? Explain your answer.

   iii. Briefly explain a plausible strategy for targeting customers for promotions, based on the work done in this problem. Use language accessible to someone who has not taken a statistics course and limit your explanation to at most five sentences.