Assignment 5 solutions

EMATM0061: Statistical Computing and Empirical Methods, TB1, 2022

Dr. Rihuan Ke

Introduction

This is the fifth assignment for Statistical Computing and Empirical Methods (Unit EMATM0061) on the MSc in Data Science & MSc in Financial Technology with Data Science. This assignment is mainly based on Lectures 11, 12 and 13 (see the Blackboard).

The submission deadline for this assignment is 23:59, 04 November 2022. Note that this assignment will not count towards your final grade. However, it is recommended that you try to answer the questions to gain a better understanding of the concepts.

Create an R Markdown for the assignment

It is a good practice to use R Markdown to organize your code and results. You can start with the template called `Assignment05_Template.Rmd` which can be downloaded via Blackboard.

You may also want to use R Markdown to organize your solutions. If you are considering submitting your solutions, please generate a PDF file. For example, you can choose the "PDF" option when creating the R Markdown file (note that this option may require Tex to be installed on your computer), or use R Markdown to output an HTML and print it as a PDF file in a browser, or use your own way of creating a PDF file that contains your solutions.

Only a PDF file will be accepted in the submission of this assignment. To submit the assignment, please visit the "Assignment" tab on the Blackboard page, where you downloaded the assignment.

Load packages

Some of the questions in this assignment require the tidyverse package. If it hasn't been installed on your computer, please use `install.packages()` to install them first.

To road the tidyverse package:

```
library(tidyverse)
```

1. Conditional probability, Bayes rule and independence

Recall that Bayes theorem helps to "invert" conditional probabilities, and the law of total probability allows us to write an (unconditional) probability in terms of a collection of conditional probabilities.

Bayes theorem

Suppose we have a probability space $(\Omega, \mathcal{E}, \mathbb{P})$. Given events $A, B \in \mathcal{E}$ with $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$, we have

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(B) \cdot \mathbb{P}(A \mid B)}{\mathbb{P}(A)}.$$

The law of total probability

Suppose we have a probability space $(\Omega, \mathcal{E}, \mathbb{P})$, and $A_1, A_2, \cdots \in \mathcal{E}$ forms a partition of Ω. For any event $B \in \mathcal{E}$, we have

$$\mathbb{P}(B) = \sum_i \mathbb{P}(A_i \cap B) = \sum_{\{i:\mathbb{P}(A_i)>0\}} \mathbb{P}(B \mid A_i) \cdot \mathbb{P}(A_i).$$

Independent and dependent events

Let $(\Omega, \mathcal{E}, \mathbb{P})$ be a probability space.

1. A pair of events $A, B \in \mathcal{E}$ are said to be independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$.
2. A pair of events $A, B \in \mathcal{E}$ are said to be dependent if $\mathbb{P}(A \cap B) \neq \mathbb{P}(A) \cdot \mathbb{P}(B)$.

1.1 Bayes theorem

(Q1) Let A be the event that it rains next week and B the event that the weather forecaster predicts that there will be rain next week.

Let's suppose that the probability of rain next week is $\mathbb{P}(A) = 0.9$.

Suppose also that the conditional probability that there is a forecast of rain, given that it really does rain, is $P(B|A) = 0.8$.

On the other hand, the conditional probability that there is a forecast of dry weather, given that there really isn't any rain is $P(B^c|A^c) = 0.75$.

Now suppose that there is a forecast of rain. What is the conditional probability of rain, given the forecast of rain $P(A|B)$?

Answer

First, we compute the probability of B (the weather forecaster predicts that there will be rain next week), using the law of total probability:

$$\begin{aligned}
\mathbb{P}(B) &= \mathbb{P}(B \mid A) \cdot \mathbb{P}(A) + \mathbb{P}(B \mid A^c) \cdot \mathbb{P}(A^c) \\
&= \mathbb{P}(B \mid A) \cdot \mathbb{P}(A) + [1 - \mathbb{P}(B^c \mid A^c)] \cdot [1 - \mathbb{P}(A)] \\
&= 0.8 \cdot 0.9 + (1 - 0.75) \cdot (1 - 0.9) \\
&= 0.745
\end{aligned}$$

Second, we apply the Bayes Theorem

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A) \cdot \mathbb{P}(B \mid A)}{\mathbb{P}(B)} = \frac{0.9 \times 0.8}{0.745} \approx 0.996$$

```r
p_A<-0.9
p_B_given_A<-0.8
p_not_B_given_not_A<-0.75
p_B<-p_B_given_A*p_A+(1-p_not_B_given_not_A)*(1-p_A)
p_A_given_B<-p_B_given_A*p_A/p_B
p_A_given_B
```

```
## [1] 0.966443
```

1.2 Conditional probabilities

(Q1) Suppose we have a probability space $\Omega, \mathcal{E}, \mathbb{P}$.

1. Suppose that $A, B \in \mathcal{E}$ and $A \subseteq B$ and $\mathbb{P}(B) \neq 0$. Give an expression for $\mathbb{P}(A|B)$ in terms of $\mathbb{P}(A)$ and $\mathbb{P}(B)$. What about when $\mathbb{P}(B \backslash A) = 0$?

2. Suppose that $A, B \in \mathcal{E}$ with $A \cap B = \emptyset$. Give an expression for $\mathbb{P}(A \mid B)$. What about when $\mathbb{P}(A \cap B) = 0$?
3. Suppose that $A, B, \in \mathcal{E}$ with $B \subseteq A$. Give an expression for $\mathbb{P}(A|B)$. What about when $\mathbb{P}(B \backslash A) = 0$?
4. Suppose that $A \in \mathcal{E}$. Give an expression for $\mathbb{P}(A|\Omega)$ in terms of $\mathbb{P}(A)$?
5. Show that given three events $A, B, C \in \mathcal{E}$ we have $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A \mid B \cap C) \cdot \mathbb{P}(B \mid C) \cdot \mathbb{P}(C)$.
6. Show that given three events $A, B, C \in \mathcal{E}$ and $\mathbb{P}(B \cap C) \neq 0$, we have $\mathbb{P}(A|B \cap C) = \frac{\mathbb{P}(B|A \cap C) \cdot \mathbb{P}(A|C)}{\mathbb{P}(B|C)}$.

Answer

1. By the definition of conditional probability:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)}{\mathbb{P}(B)}$$

   If $\mathbb{P}(B \backslash A) = 0$, then

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)}{\mathbb{P}(A) + \mathbb{P}(B \backslash A)} = 1/1 = 1.$$

2. $$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = 0$$

   Note that this includes the case $\mathbb{P}(A \cap B) = 0$.

3. $$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1.$$

   Note that this includes the case $\mathbb{P}(B \backslash A) = 0$

4. $$\mathbb{P}(A \mid \Omega) = \frac{\mathbb{P}(A \cap \Omega)}{\mathbb{P}(\Omega)} = \frac{\mathbb{P}(A)}{1} = \mathbb{P}(A).$$

5. By the definition of conditional probability:

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A|B \cap C)\mathbb{P}(B \cap C) = \mathbb{P}(A|B \cap C)\mathbb{P}(B|C)\mathbb{P}(C)$$

6. We substitute in the definition of each of the constituting conditional probabilities to see that

$$\begin{aligned} \frac{\mathbb{P}(B \mid A \cap C) \cdot \mathbb{P}(A|C)}{\mathbb{P}(B|C)} &= \left\{\frac{\mathbb{P}(B \cap A \cap C)}{\mathbb{P}(A \cap C)}\right\}\left\{\frac{\mathbb{P}(A \cap C)}{\mathbb{P}(C)}\right\}\left\{\frac{\mathbb{P}(B \cap C)}{\mathbb{P}(C)}\right\}^{-1} \\ &= \frac{\mathbb{P}(B \cap A \cap C)}{\mathbb{P}(B \cap C)} = \mathbb{P}(A|B \cap C) \end{aligned}$$

(Q2) Consider a flight from Bristol to Paris.

1. If it is windy, then the probability of the flight being cancelled is 0.3.
2. If it is not windy, then the probability of the flight being cancelled is 0.1.

The probability that it is windy is 0.2. Calculate the probability that the flight is not cancelled.

Answer

Let A be the event that the flight is not cancelled, and B be the event that it is windy.

Then $\mathbb{P}(A|B) = 1 - 0.3 = 0.7$ and $\mathbb{P}(A|B^c) = 1 - 0.1 = 0.9$.

By the law of total probability, $\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c) = 0.7 \cdot 0.2 + 0.9 \cdot 0.8 = 0.86$.

1.3 Mutual independence and pair-wise independent

(Q1) Consider a simple probability space $(\Omega, \mathcal{E}, \mathbb{P})$ with $\Omega = \{(0,0,0),(0,1,1),(1,0,1),(1,1,0)\}$. Since $(\Omega, \mathcal{E}, \mathbb{P})$ is a simple probability space containing four elements we have

$$\mathbb{P}(\{(0,0,0)\}) = \mathbb{P}(\{(0,1,1)\}) = \mathbb{P}(\{(1,0,1)\}) = \mathbb{P}(\{(1,1,0)\}) = 1/4.$$

Consider the events $A := \{(1,0,1),(1,1,0)\}, B := \{(0,1,1),(1,1,0)\}$ and $C := \{(0,1,1),(1,0,1)\}$.

Verify that $\mathbb{P}(A\cap B) = \mathbb{P}(A)\cdot\mathbb{P}(B)$, $\mathbb{P}(A\cap C) = \mathbb{P}(A)\cdot\mathbb{P}(C)$ and $\mathbb{P}(C\cap B) = \mathbb{P}(C)\cdot\mathbb{P}(B)$. Hence, we deduce that the events A, B, C are pair-wise independent.

What is $A\cap B\cap C$? What is $\mathbb{P}(A\cap B\cap C)$? Are the events A, B, C mutually independent?

Answer

Since $\mathbb{P}(\{(0,0,0)\}) = \mathbb{P}(\{(0,1,1)\}) = \mathbb{P}(\{(1,0,1)\}) = \mathbb{P}(\{(1,1,0)\}) = 1/4$, we have

$$\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = 1/2$$

Since each of $A\cap B$, $A\cap C$ and $B\cap C$ has only one element,

$$\mathbb{P}(A\cap B) = 1/4 = (1/2)\cdot(1/2) = \mathbb{P}(A)\cdot\mathbb{P}(B)$$
$$\mathbb{P}(A\cap C) = 1/4 = (1/2)\cdot(1/2) = \mathbb{P}(A)\cdot\mathbb{P}(C)$$
$$\mathbb{P}(C\cap B) = 1/4 = (1/2)\cdot(1/2) = \mathbb{P}(C)\cdot\mathbb{P}(B)$$

So A, B, C are pairwise independent.

On the other hand, $A\cap B\cap C = \emptyset$ and $\mathbb{P}(A\cap B\cap C) = 0$, so

$$\mathbb{P}(A\cap B\cap C) \neq \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C) = 1/8.$$

Therefore, A, B and C are not mutually independent.

1.4 The Monty hall problem(*)

This is an **optional** question. You might want to return to this after completing the others.

(Q1)

Consider the following game:

At a game show there are three seemingly identical doors. Behind one of the doors is a car, and behind the remaining two is a goat.

1. The contestant of the game first gets to choose one of the three doors. The host then opens one of the other two doors to reveal a goat.
2. The contestant now gets a chance to either (a) switch their choice to the other unopened door or (b) stick to their original choice.
3. The host then opens the door corresponding to the contestant's final choice. They get to keep whatever is behind their final choice of door.

Question: does the contestant improve their chances of winning the car if they switch their choice?

For clarity, we make the following assumptions:

1. The car is assigned to one of the doors at random with equal probability for each door.
2. The assignment of the car and the initial choice of the contestant are independent.
3. Once the contestant makes their initial choice, the host always opens a door which (a) has a goat behind it and (b) is not the contestant's initial choice. If there is more than one such door (i.e. when the contestant's initial choice corresponds to the door with a car behind it) the host chooses at random from the two possibilities with equal probability.

To formalise our problem we introduce the following events for $i = 1, 2, 3$:

- A_i denotes the event that car is placed behind the i-th door;
- B_i denotes the event that contestant initially chooses the i-th door;
- C_i denotes the event that the host opens the i-th door to reveal a goat.

Consider a situation in which the contestant initially selects the first door (B_1) and then the host opens the second door to reveal a goat (C_2). What is $\mathbb{P}(A_3 \mid B_1 \cap C_2)$?

What does this suggest about a good strategy? Should we switch choices?

Answer

We shall use the

$$\mathbb{P}(A_3|B_1 \cap C_2) = \frac{\mathbb{P}(C_2 \mid A_3 \cap C_1) \cdot \mathbb{P}(A_3|B_1)}{\mathbb{P}(C_2|B_1)}$$

To evaluate this expression we first need to compute

$$\begin{aligned}
\mathbb{P}(C_2|B_1) &= \frac{\mathbb{P}(B_1 \cap C_2)}{\mathbb{P}(B_1)} \\
&= \frac{\mathbb{P}(A_1 \cap B_1 \cap C_2) + \mathbb{P}(A_2 \cap B_1 \cap C_2) + \mathbb{P}(A_2 \cap B_1 \cap C_2)}{\mathbb{P}(B_1)} \\
&= \frac{\mathbb{P}(C_2|A_1 \cap B_1) \cdot \mathbb{P}(A_1 \cap B_1) + \mathbb{P}(C_2|A_2 \cap B_1) \cdot \mathbb{P}(A_2 \cap B_1) + \mathbb{P}(C_2|A_3 \cap B_1) \cdot \mathbb{P}(A_3 \cap B_1)}{\mathbb{P}(B_1)} \\
&= \frac{\mathbb{P}(C_2|A_1 \cap B_1) \cdot \mathbb{P}(A_1) \cdot \mathbb{P}(B_1) + \mathbb{P}(C_2|A_2 \cap B_1) \cdot \mathbb{P}(A_2) \cdot \mathbb{P}(B_1) + \mathbb{P}(C_2|A_3 \cap B_1) \cdot \mathbb{P}(A_3) \cdot \mathbb{P}(B_1)}{\mathbb{P}(B_1)} \\
&= \mathbb{P}(C_2|A_1 \cap B_1) \cdot \mathbb{P}(A_1) + \mathbb{P}(C_2|A_2 \cap B_1) \cdot \mathbb{P}(A_2) + \mathbb{P}(C_2|A_3 \cap B_1) \cdot \mathbb{P}(A_3) \\
&= \frac{1}{3}(\mathbb{P}(C_2|A_1 \cap B_1) + \mathbb{P}(C_2|A_2 \cap B_1) + \mathbb{P}(C_2|A_3 \cap B_1)) \\
&= \frac{1}{3}\{1/2 + 0 + 1\} \\
&= \frac{1}{2}
\end{aligned}$$

Here we use the fact that (1) $\mathbb{P}(C_2|A_1 \cap B_1) = \mathbb{P}(C_3|A_1 \cap B_1) = 1/2$, since if the contestant's initial guess is correct then the host picks either of the two remaining doors with equal probability and (2) $\mathbb{P}(C_2|A_2 \cap B_1) = 0$ and $\mathbb{P}(C_2|A_3 \cap B_1) = 1$ since whenever the contestant's initial guess corresponds to a door with a goat behind it then the host has a single choice of door to reveal.

Moreover, we also have $\mathbb{P}(C_2|A_1 \cap B_1) = 1$ since the host must choose a door with a goat behind it (so can't choose the third door if A_3 holds), and must choose a door other than the contestant's initial guess (so can't choose the first door if B_1 holds). We also have $\mathbb{P}(A_3|B_1) = \mathbb{P}(A_3) = 1/3$ by using the independence of the choice of the door for the car and the contestant's initial choice, combined with the fact that the initial assignment of the car is random with equal probabilities.

Hence, we have

$$\mathbb{P}(A_3|B_1 \cap C_2) = \frac{\mathbb{P}(C_2|A_3 \cap B_1) \cdot \mathbb{P}(A_3|B_1)}{\mathbb{P}(C_2|B_1)} = \frac{1 \cdot \frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}.$$

By the same argument we obtain $\mathbb{P}(A_{i_3}|B_{i_1} \cap C_{i_2}) = 2/3$ for any $i_1, i_2, i_3 \in \{1, 2, 3\}$ such that $i_1, i_2, i_3 = \{1, 2, 3\}$. Hence, the contestant always increases their chances by switching choice (from i_1 to i_3).

2. Random variables and discrete random variables

This section covers some of the concepts from Lectures 12 and 13.

Random variables and discrete random variables Suppose we have a probability space $(\Omega, \mathcal{E}, \mathbb{P})$. A random variable is a mapping $X : \Omega \to \mathbb{R}$, such that

$$\text{for every } a, b \in \mathbb{R}, \{\omega \in \Omega : X(\omega) \in [a, b]\} \text{ is an event in } \mathcal{E}$$

A discrete random variable is a random variable $X : \Omega \to \mathbb{R}$ whose distribution is supported on a discrete (and hence finite or countably infinite) set $A \subseteq \mathbb{R}$

Expectation

The expectation $\mathbb{E}(X)$ of the random variable X is defined by $\mathbb{E}(X) := \sum_{x \in \mathbb{R}} x \cdot p_X(x)$.

Linearity of Expectation: Given random variables $X_1, X_2, \cdots X_n$ and numbers $\alpha_1, \alpha_2, \cdots, \alpha_n$, we have

$$\mathbb{E}(\sum_{i=1}^{n} \alpha_i X_i) = \sum_{i=1}^{K} \alpha_i \mathbb{E}(X_i)$$

Equivalent condition for independent random variables

Let $X_1, \cdots, X_k : \Omega \to \mathbb{R}$ be a sequence of random variables. Then $X_1, \cdots, X_k$ are independent if and only if the following relationship holds for every sequence of well-behaved function $f_1, f_2, \cdots, f_k$,

$$\mathbb{E}(f_1(X_1) \cdots f_k(X_k)) = \mathbb{E}(f_1(X_1)) \cdots \mathbb{E}(f_k(X_k)).$$

2.1 Expectation and variance

(Q1) Suppose that we have random variables X and Y. Recall that the covariance between X and Y is defined by

$$\mathrm{Cov}(X, Y) := \mathbb{E}\big[\big(X - \bar{X}\big) \cdot \big(Y - \bar{Y}\big)\big]$$

where $\bar{X}$ and $\bar{Y}$ are the expectations of X and Y, respectively.

Now, suppose X and Y are independent. Apply the linearity of expectation and the equivalent condition for independent random variables described above, to prove that $\mathrm{Cov}(X, Y) = 0$. To apply the above condition for independent random variables, you may choose proper functions $f_1, f_2, \cdots$ when necessary.

Answer

Firstly, we can write the covariance into the following form

$$\begin{aligned}\mathrm{Cov}(X, Y) &:= \mathbb{E}\big[\big(X - \bar{X}\big) \cdot \big(Y - \bar{Y}\big)\big] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].\end{aligned}$$

Secondly, since X and Y are independent, letting $f_1(x) := x$, $f_2(x) := x$, we have

$$\mathbb{E}[XY] = \mathbb{E}[f_1(X) f_2(Y)] = \mathbb{E}[f_1(X)] \cdot \mathbb{E}[f_2(Y)] = \mathbb{E}[X]\mathbb{E}[Y].$$

Therefore,

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0.$$

2.2 Distributions

Suppose that $\alpha, \beta \in [0, 1]$ with $\alpha + \beta \leq 1$ and let X be a discrete random variable with distribution supported on $\{0, 3, 10\}$. Suppose that $\mathbb{P}(X = 3) = \alpha$ and $\mathbb{P}(X = 10) = \beta$ and $\mathbb{P}(X \notin \{0, 3, 10\}) = 0$.

Given the random variable X, answer the following questions (Q1), (Q2), . . . , Q(4).

(Q1) Expectation and variance of a discrete random variable

1. What is the probability mass function p_X for X?
2. What is the expectation of X?
3. What is the variance of X?
4. What is the standard deviation of X?

Answer

The probability mass function is

$$p_X(x) = \begin{cases} 1 - \alpha - \beta & \text{if } x = 0, \\ \alpha & \text{if } x = 3, \\ \beta & \text{if } x = 10, \\ 0 & \text{otherwise .} \end{cases}$$

The expectation is

$$\mathbb{E}(X) = 3 \times \alpha + 10 \times \beta = 3\alpha + 10\beta$$

The variance is

$$\mathrm{Var}(X) = \mathbb{E}\big(X^2\big) - \mathbb{E}(X)^2 = (3^2 \times \alpha + 10^2 \times \beta) - (3\alpha + 10\beta)^2 = 9\alpha + 100\beta - 9\alpha^2 - 100\beta^2 - 60\alpha\beta$$

The standard deviation is

$$\sigma(X) = \sqrt{\mathrm{Var}(X)} = \sqrt{9\alpha + 100\beta - 9\alpha^2 - 100\beta^2 - 60\alpha\beta}.$$

(Q2) Distribution and distribution function.

Recall that the distribution of a random variable maps a subset S of $\mathbb{R}$ to a real number.

1. Write down the distribution P_X of X. You can use the indicator function $\mathbf{1}_S$ to "indicate" if a number is in the set S.
2. Write down the distribution F_X of X

Answer

The distribution is given by

$$P_X(S) = (1 - \alpha - \beta)\mathbf{1}_S(0) + \alpha\mathbf{1}_S(3) + \beta\mathbf{1}_S(10).$$

The distribution function is given by

$$F_X(x) = \begin{cases} 0, & \text{if } x < 0 \\ 1 - \alpha - \beta, & \text{if } 0 \leq x < 3 \\ 1 - \beta, & \text{if } 3 \leq x < 10 \\ 1, & \text{if } 10 \leq x \end{cases}$$

(Q3) Variance and Covariance.

Define a new random variance $Y = X_1 + X_2 + \cdots + X_n$ where $X_1, X_2, \cdots, X_n$ are independent random variables, each of which has the same distribution as the random variable X.

Derive an expression for the variance of Y. You can use the conclusion from Question 2.1 (Q1).

Answer

Since $X_1, \cdots, X_n$ are independent, $\mathrm{Cov}(X_i, X_j) = 0$ for any $i \neq j$.

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2 \sum_{1 \le i < j \le n} \mathrm{Cov}(X_i, X_j) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = n \cdot (9\alpha + 100\beta - 9\alpha^2 - 100\beta^2 - 60\alpha\beta).$$

(Q4) Explore the distribution of the sum of n independent random variables when n is large.

Let Y be defined above, and additionally, let $\alpha = 0.2$ and $\beta = 0.3$. Recall that $X_1, X_2, \cdots, X_n$ are discrete random variables. Explain why the random variable Y is discrete.

Now let's explore the probability mass function of Y using **`R` programming**, and see its behaviour when n is large.

(Step 1). First, Y can be viewed as a generalization of the binomial random variables. A binomial random variable can be written as the sum of a sequence of Bernoulli random variables (which take values from $\{0, 1\}$). Here Y can be written as the sum of X_i, which takes values from $\{0, 3, 10\}$ (hence the distributions of X_i are known as generalized Bernoulli distributions). In `R`, we can obtain samples of Y by using the function `rmultinom()`.

For example, letting $n = 7$, run the command,

```
rmultinom(2, 7, prob=c(0.5, 0.2, 0.3))
```

```
##      [,1] [,2]
## [1,]    3    4
## [2,]    2    1
## [3,]    2    2
```

We then get a matrix, consisting of two samples (i.e., two columns), and three rows corresponding to the number of $\{X_i\}$ taking the three possible values $\{0, 3, 10\}$, respectively. For example, the three rows in the second column are $(4, 1, 2)$, meaning that we have a sample where

- 4 of $\{X_1, \cdots, X_7\}$ are equal to 0,
- 1 of $\{X_1, \cdots, X_7\}$ is equal to 3,
- and 2 of $\{X_1, \cdots, X_7\}$ are equal to 10.

Then we can compute the value of Y having these values of X_i.

Now let $n = 3$ and generate 50000 samples of $\{X_1, X_2, X_3\}$ using the `rmultinom()` function. Store the samples in an object called `samples_Xi`. Then based on `samples_Xi`, create 50000 samples of Y and store it in a data-frame called `samples_Y`, consisting of a single column called `Y`.

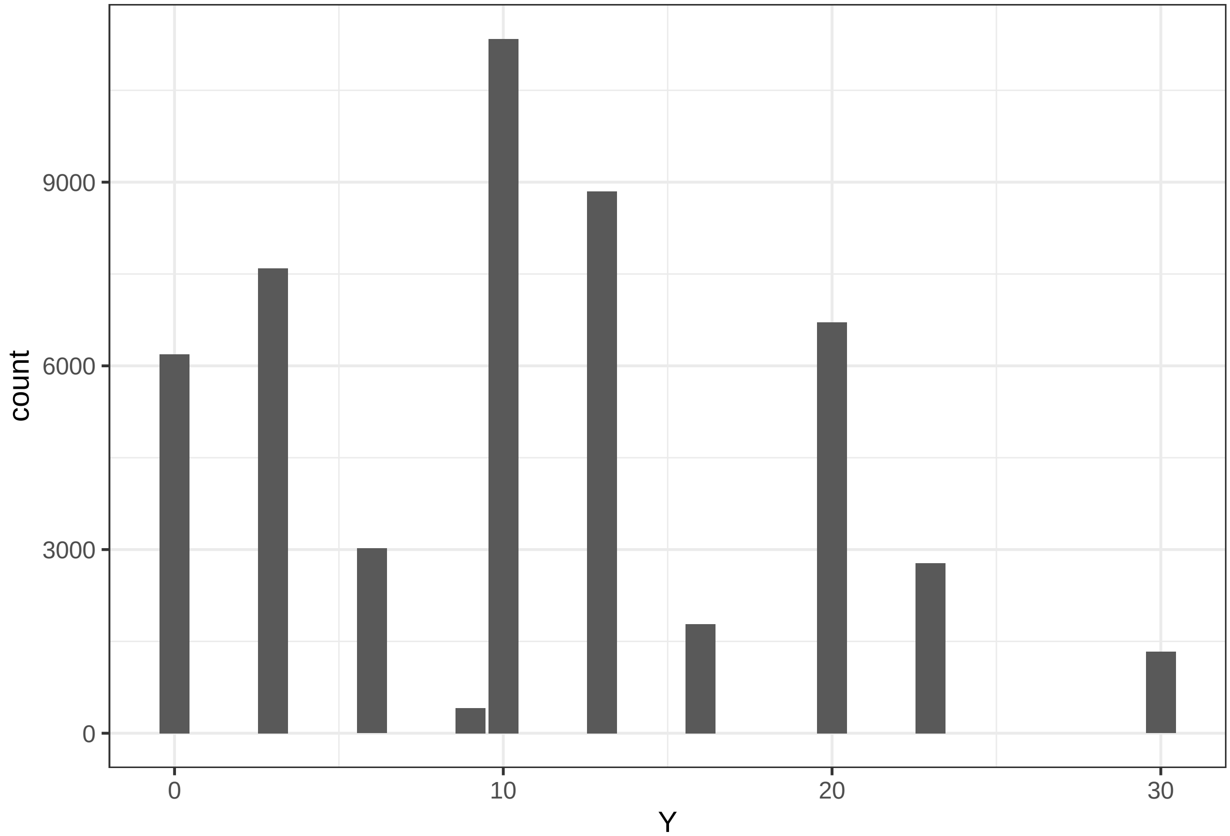
Answers

```
samples_Xi <- rmultinom(50000, 3, c(0.5, 0.2, 0.3) )
samples_Y = data.frame(Y=0*samples_Xi[1,] + 3*samples_Xi[2,] + 10*samples_Xi[3,])
```

(Step 2) Use the `ggplot` and `geom_bar()` function to create a bar plot for the samples of Y.

Answer

```r
ggplot(samples_Y, aes(Y)) + geom_bar() + theme_bw()
```



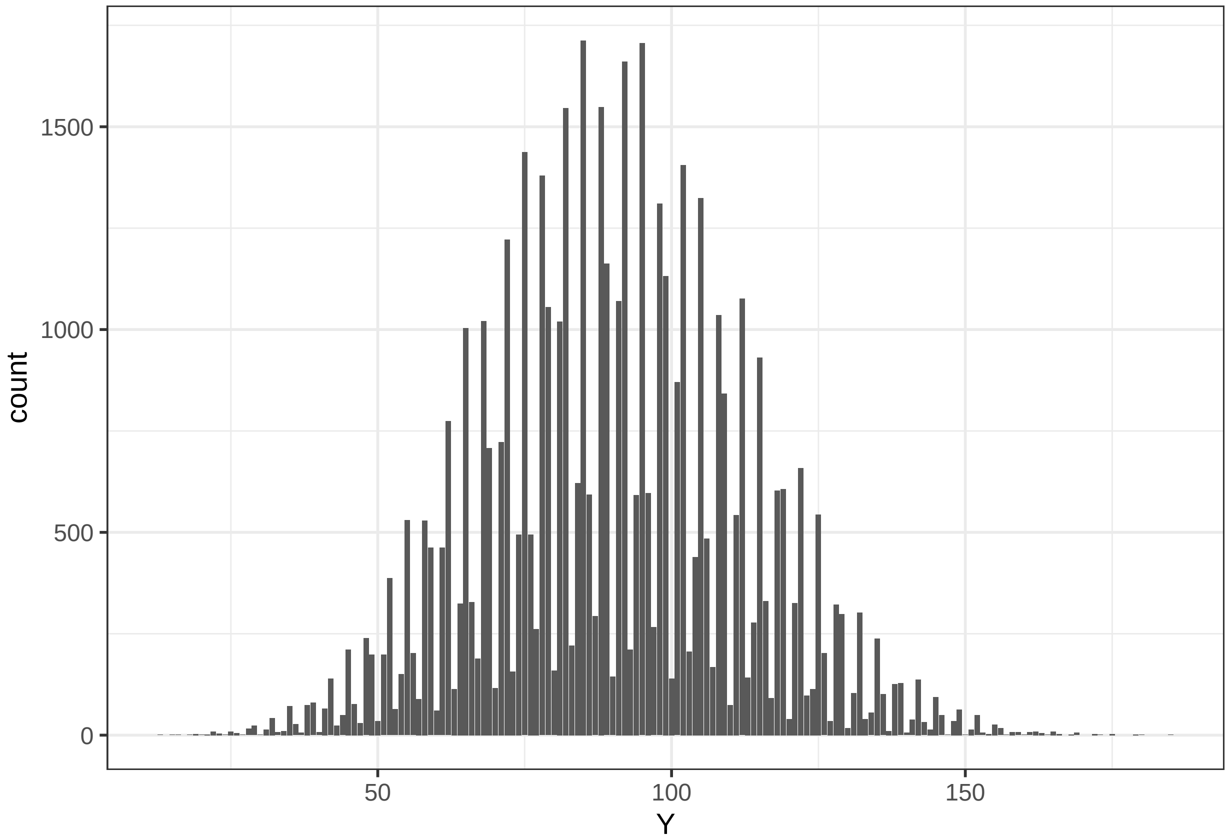
Of course your results might look different because of the randomness when using the `rmultinom()` function. Notice that Y takes values from a subset of $[0, 30]$.

(Step 3)

Now, increase the values of n by setting $n = 20$, and repeat Step 1 and Step 2 to create a new bar plot for the samples of Y. What are the maximum and minimum values of the samples? What is the sample range?

Answer

```r
samples_Xi <- rmultinom(50000, 25, c(0.5, 0.2, 0.3) )
samples_Y = data.frame(Y=0*samples_Xi[1,] + 3*samples_Xi[2,] + 10*samples_Xi[3,])
ggplot(samples_Y, aes(Y)) + geom_bar() + theme_bw()
```



Answer The minimum value, maximum value, and range of the sample is

```
print(range(samples_Y))
```

```
## [1]  13 185
```

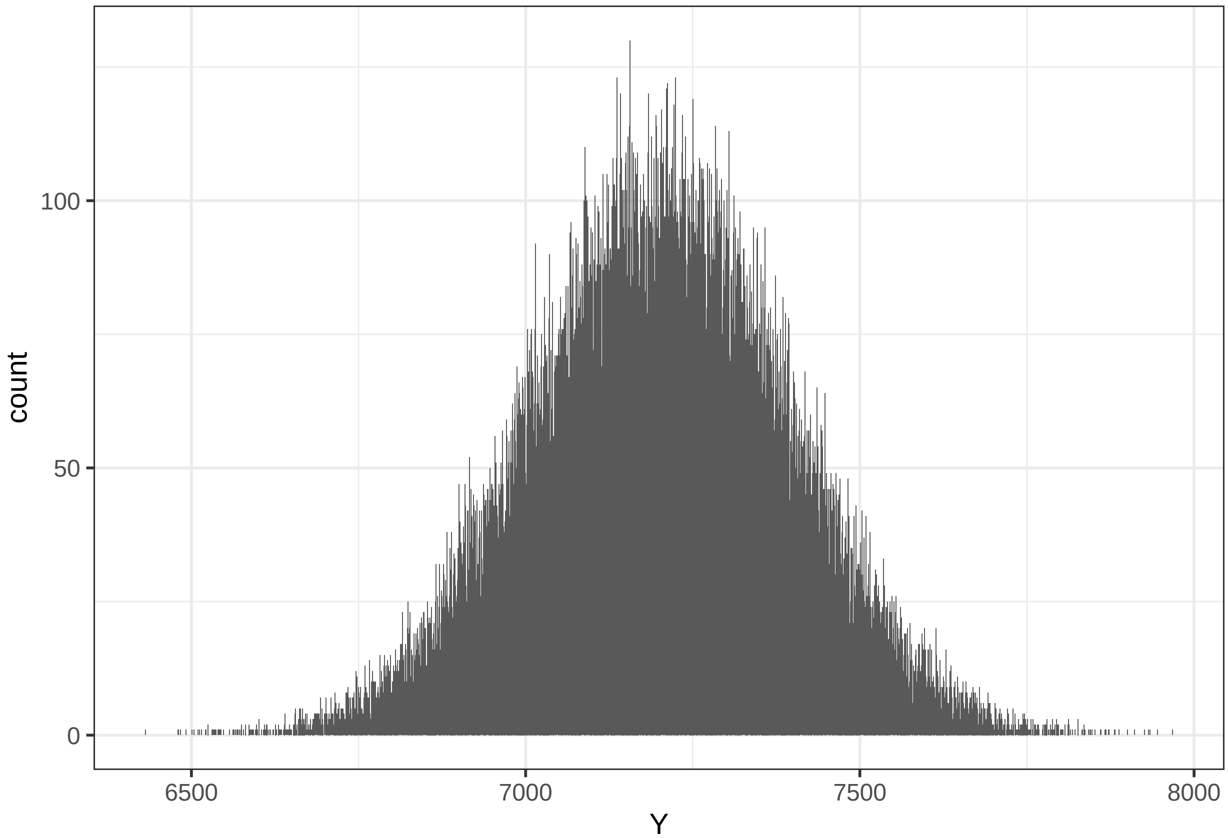
```
print(diff(range(samples_Y)))
```

```
## [1] 172
```

(Step 4) Next, increase n to 2000 and do the plot again.

Answer

```
samples_Xi <- rmultinom(50000, 2000, c(0.5, 0.2, 0.3) )
samples_Y = data.frame(Y=0*samples_Xi[1,] + 3*samples_Xi[2,] + 10*samples_Xi[3,])
ggplot(samples_Y, aes(Y)) + geom_bar() + theme_bw()
```



Notice that as we increase n, the distribution of Y tends to follow a bell-like shape. While Y is a discrete random variable (no matter how large n is), the distribution looks closer to the distribution of a continuous random variable, which is known as a Gaussian random variable. We will explore this behaviour in our future lectures.