Geno Copertino

Dataset Problems

I am going to be building and solving problems revolving around probability and applied statistics. The book used as a reference point was *Mathematical Statistics with Applications 7th Edition* by Wackerly, Mendenhall III, and Shaeaffer. The book itself covers a variety of topics and I will be making problems around the first five chapters. These five chapters include probability, discrete random variables distributions, continuous variables distributions and multivariate distributions.

The dataset I am using is https://www.kaggle.com/datasets/dimitrisangelide/speedtest-data-by-ookla which is a list of a lot of different countries download/upload speed, ping, and how many subsequent devices/recorded attempts for those countries according to Ookla. I also used https://www.kaggle.com/datasets/stetsondone/video-game-sales-by-genre for some other questions that needed data easier to implement for certain distributions. This is a dataset of video game genre sales across different nations.

The first chapter consisted mainly of using graphical and numerical methods to characterize a set of measurements. This consisted of understanding relative frequency histograms and understanding how to calculate things like standard deviation given a certain dataset. The second chapter lays some of the groundwork for probability as a whole. It gives a review on set notation; gives you tools for counting sample points and jumps into the important law of total probability and Bayes' rule.

Chapter 3 gets into the meat of the book as it goes over many useful distributions. It covers Binomial, Geometric, Hypergeometric, and Poisson probability distributions. These were all topics that had potential to be on the exams for Probability and Applied statistics and half of them wound up finding their way on there. It also covered the cool sounding Tchbysheff's Theorem. Essentially, they all deal with different ways to calculate probability depending on what information you are given. For example, Binomial deals with finding the number of successes GIVEN the fixed number of trials. Geometric deals with giving you a fixed number of successes and FINDING the number of trials needed to obtain the first one. They are all extremely interesting and useful as they genuinely can be applied in different practical ways in life.

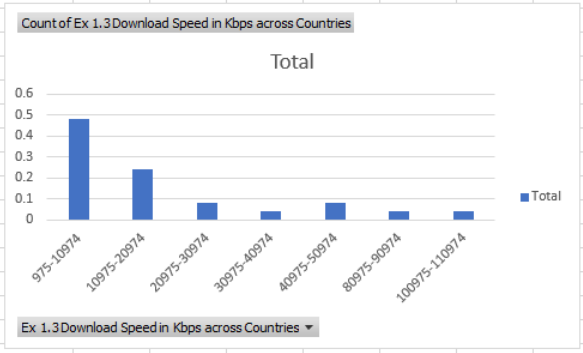
Chapters 4 and 5 deal with Continuous Variables Probability Distributions and Multivariate Probability Distributions. Most of the topics covered in these chapters involve the introduction of integrals in their formulas. Chapter 4 deals with single integral problems and how to find expected (mean) and variance of a density function. Chapter 5 deals with using double

integrals to solve bivariate and multivariate probability distributions and testing if random variables are dependent on one another.

Section 1.2 Characterizing a Set of Measurements: Graphical Methods:

I used 25 countries from my dataset and made a relative frequency histogram based off of their download speeds in kbps. United States sits at 102,337 Kbps for reference.

Row Labels	Count of Ex 1.3 Download Speed in Kbps across Countries
975-10974	48.0%
10975-20974	24.00%
20975-30974	8.00%
30975-40974	4.00%
40975-50974	8.00%
80975-90974	4.00%
100975-110974	4.00%
Grand Total	**100.00%**



Section 1.3 Characterizing a Set of Measurements: Numerical Methods:

In my dataset it tells me the mean for the download speed in Kbps is 29,000 and the standard deviation is 28,000. What fraction of countries have download speeds in the following intervals?

a. 1,000 to 56,000 Kbps = 0.68 or 68% of countries. Via the empirical rule $\mu \pm \sigma$ contains approximately 68% of the measurements. $\mu + \sigma = 29{,}000 + 28{,}000 = 56000$ and $\mu - \sigma = 29{,}000 - 28{,}000 = 1{,}000$.

b. 1,000 to 84,000 = .815

Section 2.3: A Review of Set Notation

In Overwatch 2 (action video game) there are three roles: tank, damage, and support. In an arcade mode called 3v3, you can have any combination of 3 of these roles. Define the super set and subsets.

A: if one support player in is in the game

B: if one tank is in the game

Find the union and intersection of A and B.

S = {(t, t, t), (t, t, d), (t, d, d,), (d, d, d,), (d, d, s,), (d, s, s,), (s, s, s,), (s, s, t), (s, t, t), (t, d, s)

A = {(d, d, s), (s, t, t), (t, d, s)

B = {(t, d, d,), (s, s, t), (t, d, s)

A ∪ B = {(d, d, s), (s, t, t), (t, d, s), (t, d, d), (s, s, t)

A ∩ B = {(t, d, s)

Section 2.4 A Probabilistic Model for an Experiment: The Discrete Case

	Owns a device	
Lives in China	Yes	No
Yes	0.0016	0.9984
No	0.0083	0.9917

a. lives in china: 100%
b. lives in china but doesn't own a device: 99.84%
c. Owns a device but doesn't live in China: 0.0083%

Section 2.5 Calculating the Probability of an Event: The Sample-Point Method:

The data I linked reports the median upload speed across all of the countries listed was 17.3k Kbps. That is half of the countries had incomes exceeding this amount of upload speed and half had less than or equal to below this amount. If four random countries are surveyed and each one reveals whether its average upload speed is above 17.3k Kbps.

a. List the points in the sample space: The sample space is {GGGG, GGGN, GGNG, GNGG, NGGG, GGNN, GNGN, GNNG, NGGN, NGNG, NNGG, GNNN, NGNN, NNGN. NNNG, NNNN} where G = greater and N = not greater.
b. A: 11/16
   B: 3/8
   C:1/4
c. P(A): 69%
   P(B): 38%
   P(C): 25%

Section 2.6: Tools for Counting Sample Points:

Mohamed needs to get to the airport to move to Singapore as he wants to achieve better Internet. A fleet of ten taxis is dispatched to three airports. Five go to airport A, four go to airport B and one goes to airport C. How many distinct ways does Mohamed have to get to the airport?

(10!)/(5!*4!*1!) = 1260

Section 2.7 Conditional Probability and the Independence of Events:

We have the sales of action and fighting games from Japan and North America. Suppose that a single sale is selected from the data.

A: The sale came from Japan.

B: The sale is an action game.

Find the following:

a. P(A): .14
b. P(B): .81
c. P(A ∩ B): .13
d. P(A ∪ B): .82
e. P(A Bar): .86
f. P(A ∪ B BAR): .18
g. P(A ∩ B BAR): .87
h. P(A|B): .68
i. P(B|A): .13/.81 = .16

	Sales		
Nation	Action	Fight	Total
Japan	.13	0.07	14
NA	0.68	.17	86
Total	.81	.19	1.00

Section 2.8: Two Laws of Probability

A study was conducted and determined0 24% of gamers like sports games and 16% like MMOs. If the two events are independent, find P(A ∪ B), P(A' ∩ B'), P(A' ∪ B').

$P(A \cup B) = .24 + .16 - (.24 * .16) = .0384$

P(A' ∩ B') = P(A U B)' = .6326

P(A' ∪ B') = .76 + .84 - .6384 = 0.9616

Section 2.9: Calculating the Probability of an Event: The Event-Composition Method:

PS5's produced in a factory—40% come from line I and 60% from line II. Line I has a defect rate of 3%, whereas line II has a defect rate of 5%. If an item is chosen at random from a day's production, find the probability that it will not be defective.

I = items from line I

II = items from line II

N = not defective

$P(N) = P(N \cap (I \cup II)) = P(N \cap I) + P(N \cap II) = .97(.4) + .95(.6) = .958$

Section 2.10 The Law of Total Probability and Bayes' Rule:

In the U.S.A. 80% of females reported good internet quality while 50% of males reported good internet quality. A group of 20 people, 15 female and 5, were asked to give their opinion on the internet quality. A response picked at random was positive. What is the probability that it was of a male.

P(F) = 15/20 = ¾

P(M) = 5/20 = ¼

P(N|F) = 1 – 80% = .2

P(N|M) = 1- 50% = .5

P(M|N) = .5 * (1/4) / .2 (3/4) + .5 (1/4) = 0.45 = 45%

Section 3.2: The probability Distribution for a Discrete Random Variable

A gamer decides he wants to pick up and try out some new games. There are 5 action games and 10 sports games. The gamer decides he will pick up 3 of these games. If he selects games at random, what are the odds that a single sports game will be picked up? Two? Three? None?

$$P(3) = \frac{\binom{5}{0}\binom{10}{3}}{\binom{15}{3}} = \frac{120}{455} = .26$$

$$P(2) = \frac{\binom{5}{1}\binom{10}{2}}{\binom{15}{3}} = \frac{225}{455} = .49$$

$$P(1) = \frac{\binom{5}{2}\binom{10}{1}}{\binom{15}{3}} = \frac{100}{455} = .22$$

$$P(0) = \frac{\binom{5}{3}\binom{10}{0}}{\binom{15}{3}} = \frac{10}{455} = .02$$

Section 3.3: The Expected Value of a Random Variable or a Function of a Random Variable

Using our answer(s) from above, find the mean, variance, and standard deviation of Y.

Mean = (0)(.02) + 1(.22) + 2(.49) + 3(.26) = 1.98

Variance = (0 – 1.98)^2(0.02) + (1 – 1.98)^2(.22) + (2-1.98)^2(.49) + (3-1.98)^2(.26) = 0.56

Standard Deviation = sqrt(.56) = 0.74

Section 3.4 The Binomial Probability Distribution:

In terms of sales in North America between Racing and Puzzle games, 25% of people favored Puzzle games. Identify the event favors Puzzle as a success S. It is evident that the probability of S on trial 1 is .25. Consider the event B that S occurs on the second trial. Then B can occur two ways: The first two trials are both successes or the first trial is a failure and the second is a success. Show that P(B) = .25. What is P(B| the first trial is S)? Does this conditional probability differ markedly from P(B)?

The probability of first trial being a failure is 1 - .25 = .75

The probability of the second trial becoming a success is .25.

The probability of back-to-back successes is = .0625

The probability of a failure then success back-to-back is .75 * .25 = .1875

Total probability of two trials becoming both successes or the first trial is a failure and second is a success is .0625 + .1875 = .25

P(B) does equal .25. The conditional does not really differ from P(B).

Section 3.5 Geometric Probability Distribution:

27% of people from Japan prefer Role-Playing video games. If we rounded up Japanese people sequentially and asked them what genre they preferred what is the probability that the first person prefers Role-Playing video games is on the fifth asked person total.

P(X = 5) = .27 * (1 - .27)^5-1

= .27 * .73^4

= .07668

Section 3.7 Hypergeometric Probability Distribution:

A lot of 7 gamers are selected by a company to test a new video game from a group of 20 total gamers. Of the total gamers, 15 were fans of Role-Playing games and 5 were fans of Action games. The lot was selected randomly and only 1 Action game enjoyer was selected. Should we have any doubt on the randomness of the selection process?

$P(Y \leq 1) = \frac{\binom{15}{7}\binom{5}{0}}{\binom{20}{7}} + \frac{\binom{15}{6}\binom{5}{1}}{\binom{20}{6}} = 0.40$

Since the probability is not small there is no reason to doubt the randomness of the selection.

Section 3.8 Poisson Probability Distribution:

Customers arriving at a video game store at an average of ten per hour. During a given hour what are the probabilities that

a. no more than three customers arrive?: $P(X \leq 3) = P(X = 0) + P(x = 1) + P(X = 2) + P(x = 3) = \frac{e^{-10} * 10^0}{0!} + \frac{e^{-10} * 10^1}{1!} + \frac{e^{-10} * 10^2}{2!} + \frac{e^{-10} * 10^3}{3!} = 0.01033605$
b. at least two customers arrive?: $P(X \geq 2) = 1 - \big(P(x = 0) + P(X = 1)\big) = 1 - \frac{e^{-10} * 10^0}{0!} - \frac{e^{-10} * 10^1}{1!} = 1 - 11 * e^{-10} = 0.99950060$
c. exactly five customers arrive?: $P(X = 4) = \frac{e^{-10} * 10^4}{4!} = 0.01891663$

Section 3.11 Tchebysheff's Theorem:

25% of Japanese gamers prefer Role-Playing video games. If the same proportion of Americans preferred those games and 120 Americans were interviewed on what genre they preferred,

a. what is the expected number of Americans who prefer Role-Playing video games?
   p = .25 n = 120
   E(x) = np = 120(.25) = 30
b. what is the standard deviation of the number Y who would prefer Role-Playing games?
   $$\sigma = \sqrt{V(x)} = \sqrt{npq} = \sqrt{120(.25)(.75)} = 4.74$$
c. is it likely that the number of Americans who preferred Role-Playing games exceeds 33? It is likely since the standard deviation is 4.74. 30 + 4.74 ≈ 34.74

Section 4.2 Probability Distribution of a Continuous Random Variable:

In the past 40 years (in millions) the Puzzle games could sell at most 121 copies in a given region. The total amount of copies sold in a 40 year span is a random variable Y with a probability density function given by:

$$\begin{cases} y, & 0 \leq y \leq 1 \\ 40 - y & 1 \leq y \leq 2 \\ 0, & elsewhere \end{cases}$$

a. Find F(y): For 0 < y < 1 $F(y) = \int_0^y t dt = \frac{y^2}{2}$

   For 1 ≤ y < 2, $F(y) = \int_0^1 t dt + \int_1^y (40 - t) dt = 40y - \frac{y^2}{2} - 1.$

b. Find the probability that the sales will be between 101 and 141 copies in a 40 year span.

$$P(.83 \leq 1.17) = F(1.17) - F(.83) = 44.89$$

c. Given that there were more than 121 million copies sold, find the probability that there were more than 181.5 million copies sold:
   P(Y > 1.5 | Y > 1) = P(Y > 1.5)/P(Y > 1) = .125/.5 = .25.

Section 4.3 Expected Values for Continuous Random Variables:

If Y has a density function $\begin{cases} \left(\frac{1}{2}\right)(2 - y), & 0 \leq y \leq 1 \\ 0, & elsewhere \end{cases}$

Find the mean and variance of Y.

Mean = $E(Y) = \int_0^1 .5y(2 - y) dy = \frac{y^2}{2} + \frac{y^3}{6} = 2/3$

E(Y^2) = $\int_0^1 .5y^2(2 - y) dy = \frac{y^3}{3} + \frac{y^4}{8} = 11/24$

V(Y) = 2/3 – (11/24)^2 = .457

Section 4.4 Uniform Probability Distribution:

In the action video game Assassin's Creed Brotherhood you can buy parachutes and use them. If the main characters lands at a random point on a line between buildings A and B, find the probability that he is closer to A than to B.

If he is closer to A, he landed in the interval (A, $\frac{(A+B)}{2}$) This would be one half of the total interval length, so the probability is .5. If the distance to A is more than three times his distance to B, he landed in the interval ($\frac{3B+A}{4}, B$). This is one quarter of the total length, so the probability is .25.

Section 5.2 Bivariate and Multivariate Probability Distributions:

PS5's are to be stocked on the shelves once at the beginning of each week at Gamestop and then sold to individual customers. Let Y1 denote the proportion of the capacity of the stock of games available after the shelves are stocked at the beginning of the week. PS5's are in extremely limited quantity so Y1 varies from week to week. Y2 denotes the proportion of the capacity of PS5's that is sold during the week. Y1 and Y2 are proportions so both are bounded from 0 to 1. The amount of PS5 sold cannot exceed the amount stocked for obvious reasons. Suppose the joint density function is:

$$f(x,y) = \begin{cases} 2x, & 0 \le y \le x \le 1 \\ 0, & elsewhere \end{cases}$$

Find the probability that less than one-half of the PS5's will be stock and more than one-quarter will be sold.

$P(0 \le X \le .5, .25 \le Y) = \int_{1/4}^{1/2} \int_{1/4}^{x} 2x dy x$ = 5/192

Section 5.3 Marginal and Conditional Probability Distributions

Using information from the last exercise, find the marginal density function for Y.

$$f_2(y) = \int_y^1 2x dx = 1 - y^2, 0 \le y \le 1$$

Section 5.4 Independent Random Variables

Is the amount in stock earlier, independent of the amount sold. Since the joint density is not the product of the marginal densities the amount in stock is dependent.