**Goodreads Dataset Report – So Many Books!**

Written by: Mia Watts

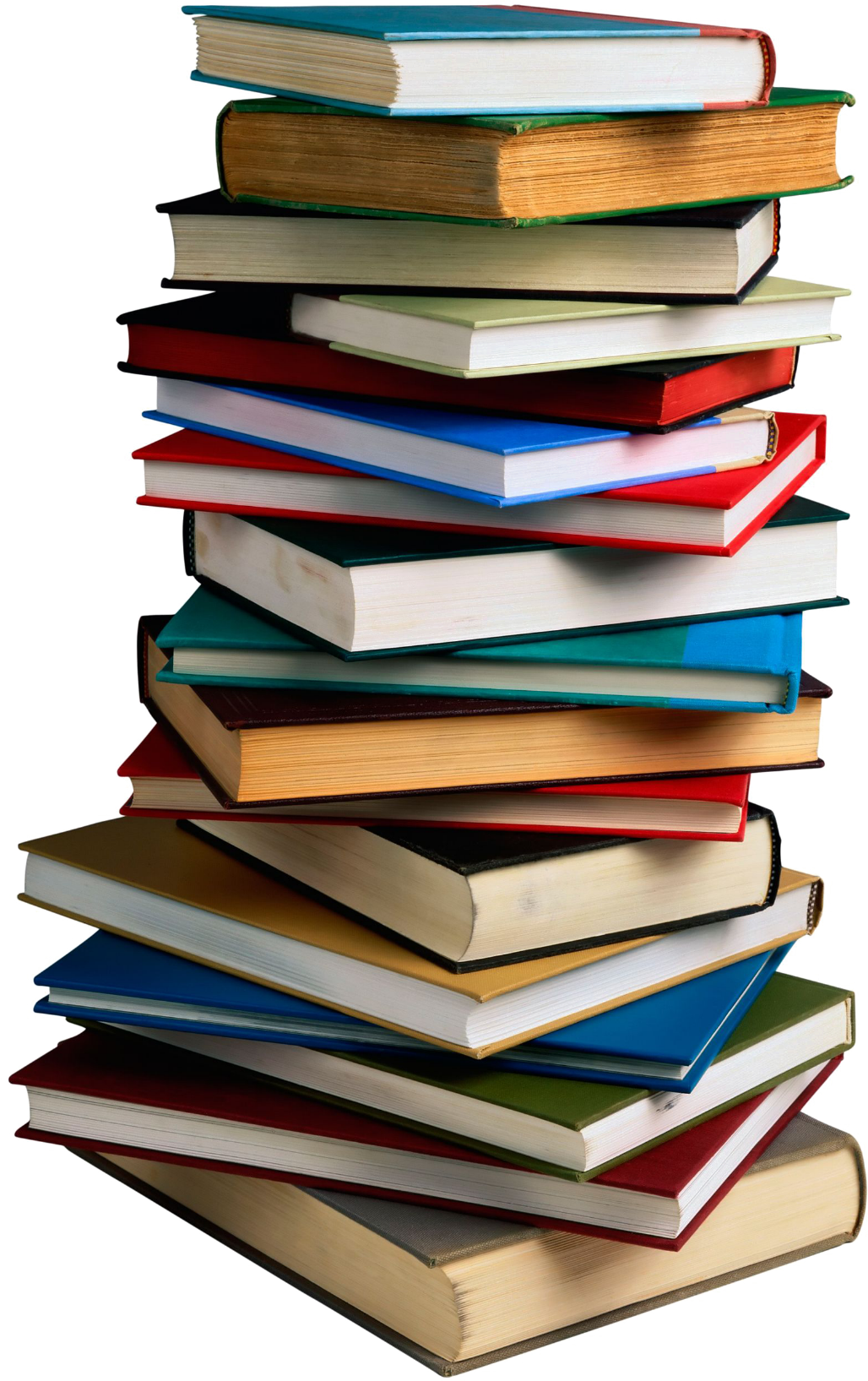


Table of Contents

[Introduction 3](#_Toc165215543)

[Chapter 1 – What Is Statistics? 4](#_Toc165215544)

[Section 1.2 4](#_Toc165215545)

[Section 1.3 5](#_Toc165215546)

[Chapter 2 – Probability 5](#_Toc165215547)

[Section 2.3 5](#_Toc165215548)

[Section 2.4 6](#_Toc165215549)

[Section 2.5 6](#_Toc165215550)

[Section 2.6 6](#_Toc165215551)

[Section 2.7 7](#_Toc165215552)

[Section 2.8 7](#_Toc165215553)

[Section 2.10 7](#_Toc165215554)

[Chapter 3 – Discrete Random Variables and Their Probability Distributions 8](#_Toc165215555)

[Section 3.2 8](#_Toc165215556)

[Section 3.4 8](#_Toc165215557)

[Section 3.5 8](#_Toc165215558)

[Section 3.6 9](#_Toc165215559)

[Section 3.7 9](#_Toc165215560)

[Section 3.8 9](#_Toc165215561)

[Section 3.11 9](#_Toc165215562)

[Chapter 4 – Continuous Variables and Their Probability Distributions 10](#_Toc165215563)

[Section 4.2 10](#_Toc165215564)

[Section 4.3 10](#_Toc165215565)

[Section 4.4 10](#_Toc165215566)

[Section 4.6 10](#_Toc165215567)

[Chapter 5 – Multivariate Probability Distributions 11](#_Toc165215568)

[Section 5.2 11](#_Toc165215569)

[Section 5.3 11](#_Toc165215570)

[Section 5.4 11](#_Toc165215571)

[Conclusion 12](#_Toc165215572)

[References 13](#_Toc165215573)

# Introduction

Goodreads is a book cataloguing site that was created in 2007. It allows users to add books to their digital bookshelves, log books being read or purchased, and see books that friends are reading (Goodreads, 2024). The site has a vast selection of books to choose from, each having a review section where users who have read the book can either encourage others to read or stay away from the book in question. Individuals using the site, either with or without an account, can search for books they’d like to read and can add the book to their “TBR” section, which stands for “To Be Read.” The site has ever-growing collections of over 10,000 books, most of which have become fan favorites.

On Kaggle, a website that houses a databasing section filled with interesting datasets that can be used for reports such as this one, a Goodreads dataset can be found that has over 11,000 entries. The Goodreads dataset, when published, had a complete collection of all books found within the site, logging their ISBNs, authors, titles, languages, and more (Soumik, 2020). Datasets such as the Goodreads one can help interested individuals learn more about the books found within the Goodreads site and can launch research projects such as this one that can tell stories about the data found inside the downloadable .csv data file.

It is important to note that with most, if not all, data, there is a story to be told. In this case, that story will be told through statistics. Problems from each section where homework was assigned in the Probability and Applied Statistics course at Stockton University will be created from scratch and will be solved using data from the Goodreads dataset, possible teaching readers more about the dataset, how Goodreads categorizes their books, and which books are more common than others. Traversing through the first five chapters of the course textbook entitled *Mathematical Statistics with Applications* by Dennis D. Wackerly, William Mendenhall III, and Richard L. Scheaffer, figures, equations, and statements will be created and made to visually or textually display the data made solely from books logged on the Goodreads website.

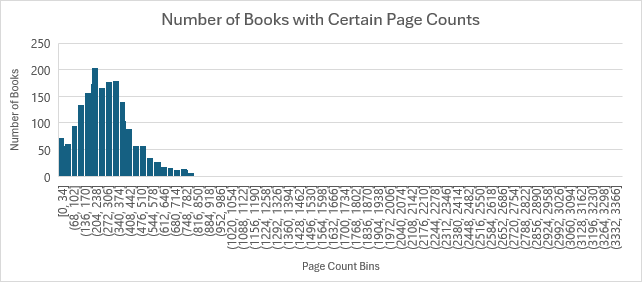


# Chapter 1 – What Is Statistics?

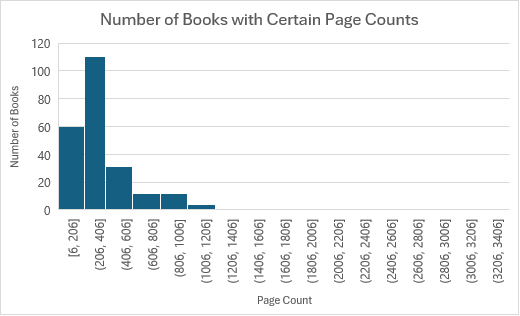
In the first chapter introduced in the textbook, different methods of graphical visualization and analyses are performed, including histogram and bar graph comparisons.

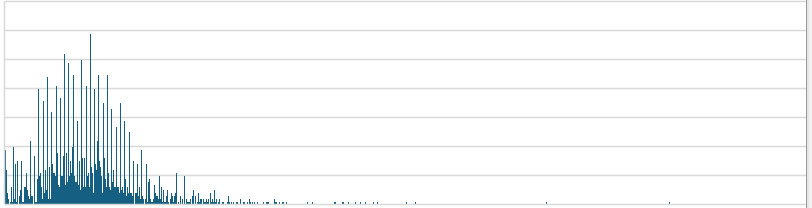
## Section 1.2

The following is a histogram created to display the number of books that have a certain page count (i.e., books with similar page counts are grouped together).



It is difficult to see the total amount of books because there are over 11,000 books that need to fit on the graph, so two more histograms can be created. The first is a smaller version of the above histogram, and the second has a large bin count (1,000) and separates books in a different way.





## Section 1.3

Given that the first three books in the dataset have review counts of 27591, 29221, and 244, find the mean review count, variance, and standard deviation.

* Mean = = 19018.67
* Variance =2
  + (27591 – 19018.67)2 + (29221 – 19018.67)2 + (244 – 19018.67)2 = 530060612.7
  + 530060612.7 \* = 26503036.3
* Standard Deviation =
  + = 16279.75142

# Chapter 2 – Probability

Chapter 2 focuses mainly on probability as the title of the chapter suggests. This includes, and is not limited to, conditional probability, independence and dependence, set notation (including union, intersection, and complement), Bayes’ rule, and random sampling. These all help with determining the probability that an event may or may not occur.

## Section 2.3

In the Goodreads .csv file, two sets can be defined:

* A: Books written by Carla Lynn Stockton
* B: Books written by Teresa Edgerton

List the points in .

* The points that would comprise of the union of A and B would be the following: *The Queen’s Necklace, Goblin Moon (The Goblin Moon Duology #1), The Work of the Sun (The Green Lion Trilogy #3), The Castle of the Silver Wheel, The Moon in Hiding (The Green Lion Trilogy #2), and Cliffs Notes on Shakespeare’s Hamlet*.

## Section 2.4

On the Goodreads website, books can be rated with 1 star, 2 stars, 3 stars, 4 stars, and 5 stars. In addition, books listed on the site can be written by one author or multiple authors. Given this information, list the sample space (all possible points) for the experiment.

* We can define the ratings as 1, 2, 3, 4, and 5 and can define the author count as O (for one) or M (for more or multiple). The sample space for this experiment would be O1, O2, O3, O4, O5, M1, M2, M3, M4, and M5. A book written by one author can have ratings from 1-5 stars as can a book written by multiple authors.

## Section 2.5

Two out of three Harry Potter books written by J.K. Rowling approved for presentations at a book signing conference are randomly selected to be shown at the next conference. One book has a 3-star rating and the other two have 4-star ratings. List the possible outcomes for this situation and assign probabilities to each.

* First, let’s define 41 as the first book with a 4-star rating and 42 as the second book with a 4-star rating. From this, we can define 3 as the book with the 3-star rating.
  + Possible outcomes = 4142, 413, and 423
  + Probabilities for each would be . There is a 33.33% that any of these combinations are chosen, and there is a 66.67% chance that the book with the 3-star rating will be chosen for the next conference.

## Section 2.6

1. 48 books written by Agatha Christie and 49 books written by P.G. Wodehouse are displayed on a large table. How many ways can someone select one book written by Agatha Christie and one book written by P.G. Wodehouse?

Using the rule, we can multiply 48 \* 49 and see that there are 2352 ways to select one book from Agatha Christie’s collection and one book from P.G. Wodehouse’s collection.

1. Out of the 48 Agatha Christie books, how many ways can you select only 4?

Using combinations, n = 48 and r = 4. Thus, with the combinations formula being , of the total number of Agatha Christie books in the dataset (48), there exist 194580 ways of choosing 4 books.

## Section 2.7

From the Goodreads dataset, a librarian found that around 0.7% of Goodreads books have the word “moon” in their titles and around 51.4% of Goodreads books have the word “the” in their titles. The intersected result of these two events was found by the librarian to be 0.25%. If we define having the word “moon” in a book’s title as P(Moon) and having the word “the” as P(The), find P(Moon | The).

* P(Moon | The) = = 0.0049. Thus, the probability that a book has the word “moon” in its title given it has the word “the” in the title is 0.49%.

## Section 2.8

Librarian Lisa is putting books away during a slow day at the county library. She notices that 15% of books were published in 2006 (P(A)). She also notices that 2.5% of books in the library were published on the twelfth day of any month (P(B)).

1. Find P().

0.3% of books were published both on the twelfth of any month and in 2006.

1. Are A and B independent?

Close, but no. They are dependent. The probability that the book was published in 2006 and on the twelfth of any month is 0.34%. The probability that the book was only published in 2006 is 15%, and the probability that the book was only published on the twelfth of any month is 2.5%. Converting these back to decimals, we have 0.0034 = 0.15 \* 0.025. This is close, because the right side ends up being 0.0038, but they are not equal, so they are dependent.

## Section 2.10

86% of books in the Goodreads database are written in American English, and the other 14% of books are written in a) other languages or b) different forms of English. 12% of books are written in English but classify as a different form. Find the conditional probability of this situation using Bayes’ Theorem and conditional probability formulas.

* = 0.0195. This means there is a 1.95% chance that given the book is written in a different form of English, it is specifically written in US English.

# Chapter 3 – Discrete Random Variables and Their Probability Distributions

To continue, this chapter discusses probability distributions which include binomial, geometric, negative binomial, and hypergeometric probability distributions. Other important formulas from the chapter include Tchebysheff’s (Chebyshev’s) Theorem and Poisson Distribution, along with how to calculate expected values, variances, and standard deviations of a certain distribution. All distributions introduced in the chapter are associated with discrete random variables as shown in the chapter’s title.

## Section 3.2

Given there are 19 books written by or about Ayn Rand in the Goodreads dataset and you can pick 2 books from the 19 in 171 ways, find Y = 1 and Y = 2 if Y represents the number of books with red covers. 10 books have red covers, and 9 have blue ones.

* p(1) = P(Y = 1) = = 0.52. There is a 52% chance that you will choose a book with a red and blue cover.
* p(2) = P(Y = 2) = = 0.26. There is a 26% chance that you will choose two books with red covers and no books with blue covers.

## Section 3.4

The library is doing a raffle, and to win, you must pick, out of a bag, a miniature version of a book by Ayn Rand. A total of 67 miniature books are in the bag, 48 of them being from Agatha Christie and 19 of them being from Ayn Rand. If you were to pick a book out of the bag 10 times, what is the probability you would draw one book from Ayn Rand from the bag?

* p = .28, q = 0.72, n = 10, y = 1
* = 0.15. There is a 15% chance that you will pick a winning book out of the bag.

## Section 3.5

From the previous problem, we saw that p = 0.28. Find the expected value assuming geometric distribution with this success value.

* = = 3.57. This means the expected value is 3.57.

## Section 3.6

Given the same success and failure probabilities as Section 3.4, what is the probability that you would pick our second winning book on the fifth draw?

* p = 0.28, q = 0.72, y = 5, r = 2
* = 0.12. There is a 12% chance that you would pick a second winning book on the fifth draw.

## Section 3.7

Continuing with the previous examples, there are 19 Ayn Rand books, 48 Agatha Christie books, and a random sample of 15 books is drawn. What is the probability that exactly 5 of the 15 drawn will be Agatha Christie books?

* N = 67, y = 5, n = 15, r = 48, N – r = 19, n – y = 10
* = 0.000457. There is a 0.0457% chance that exactly 5 of the 15 books drawn will be Agatha Christie books.

## Section 3.8

An individual who frequents the library and uses Goodreads to post reviews reads, on average, 2 books by Billy Collins once every couple of weeks. What is the probability that the individual will read only 1 book by Billy Collins once every couple of weeks?

* = 0.27. There is a 27% chance that the individual will read only 1 book by Billy Collins once every couple of weeks.

## Section 3.11

In Section 1.3’s problem, the mean of the first three review counts could be rounded to 19019 and the standard deviation could be rounded 16280. What percentage of the review values are between 2019 and 36019?

* k = 1.04
* 1 - = 0.08. 8% of the values fall between 2019 and 36019.

# Chapter 4 – Continuous Variables and Their Probability Distributions

Next, instead of focusing on discrete variables, this chapter focuses on continuous variables and their subsequent probability distributions. These distributions include Uniform Probability Distribution, Normal Probability Distribution, and Gamma Probability Distribution.

## Section 4.2

After reviewing Goodreads data and producing equations, the boss of a library wants to find a value such that the integral equals 1. For what value of a does this work?

* After calculating what is shown above, c must be , or .

## Section 4.3

Using the same function above with the new c-value, what is the expected value of the result?

* With c-value =
* = 1.6
* The expected value is 1.6.

## Section 4.4

Given that two random books have rating counts of 2088 and 1266, find the uniform probability distribution of the interval.

* = 0.0012, or 0.12% uniform probability distribution.

## Section 4.6

Given that the mean of two rating counts is 1677, find the expected value of a Gamma Distribution if alpha is 5.

* 1677 \* 5 = 8385. The expected value is 8385.

# Chapter 5 – Multivariate Probability Distributions

Lastly, this chapter introduces Bivariate and Multivariate Probability Distributions, independence within functions, distributions, densities, and probabilities, and Marginal and Conditional Probability Distributions. Many of these require knowledge of integrals, specifically double integrals where one value is integrated first followed by a second value.

## Section 5.2

The local library mobile has 20 total books to give to people who visit while it is open. There are 10 books published by Scholastic, 7 books published by Crown, and 3 books published by Bantam Books. Four books are selected. Calculate the probability function from section 5.2 if x = Scholastic and y = Crown and you choose one Scholastic book and one Crown book.

* = 0.0433.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Y1 | | | | | |
| Y2 |  | 0 | 1 | 2 | 3 | 4 |
| 0 | 0 | 0.0014 | 0.013 | 0.022 | 0.00722 |
| 1 | 0.0021 | 0.043 | 0.195 | 0.074 | 0 |
| 2 | 0.03 | 0.13 | 0.195 | 0 | 0 |
| 3 | 0.074 | 0.0722 | 0 | 0 | 0 |
| 4 | 0.0433 | 0 | 0 | 0 | 0 |

## Section 5.3

Using the above table from Section 5.2, what is P(X = 0 | Y = 1)?

* = = 0.0066857. This means that there is a 0.7% chance that zero scholastic books will be selected and only one Crown book will be selected (alluding to the fact that the other three books are all Bantam).

## Section 5.4

Using the above table from Section 5.2, is p(0, 1) = p1(0)p2(1)? (In other words, prove that the statement is either dependent or independent).

* p(0, 1) = 0.0021.
* p1(0)p2(1) = 0.0065.
* Because the values are not equal, X and Y are dependent.

# Conclusion

There are many books logged within the Goodreads dataset, and from a quick statistical analysis of the data found within, many conclusions can be drawn. From the histograms shown in the sections from Chapter 1, most books in the Goodreads dataset have around 200-400 pages, with some books having more. Chapter 1 also assists in showing calculations related to mean, variance, and standard deviation, concluding that for the first three books in the set, these values are fairly high when discussing the review counts of each. In the sections shown in Chapter 2, probabilities were concluded based on the combinations of ways one may select a book at a conference or at a book tabling event. It was found that a small percentage of books both have the word “moon” and “the” in their titles and that there are ten ways by which to order books that can have a 1-5 star rating and be written by either one author or multiple authors. These values indicate the type of data with which is being worked, showing how large the dataset is and how few books may occasionally fall within a certain category.

Continuing with Chapter 3, Tchebysheff’s Theorem, Poisson distribution, and binomial, hypergeometric, geometric, and negative binomial distributions were either used or referenced to figure out the percentages of books that had certain cover colors or the percent chance of selecting a certain book given they were written by different authors. In the case of the examples where a raffle was created where individuals had to pick books out of a bag, there was a 15% chance that someone would win by picking a book written by Ayn Rand after ten tries given there were fewer winning books than there were losing books. Expected values of these problems were considered as well, and it was found that the expected value of a certain distribution was 3.57 provided geometric distribution was used while following the same raffle problem as before. Along with this information, it was found that 8% of books have review counts between 2019 and 36019, showing that most review counts do not fall within this interval.

Later, Chapter 4 introduced Uniform Distribution, Gamma Distribution, and integrals which were important when attempting to find a specific c-value. It was found that a c-value for a given equation created by a library’s boss was 0.25 and the expected value of the equation was 1.6. After these problems, a result was found for the Uniform Distribution of two rating counts. With one book having 1266 reviews and another having 2088, the distribution of these books regarding their reviews was 12%. Lastly, Gamma Distribution was used to find the expected value given an alpha value of 5 and a mean review count of 1677, which was 8385. These problems can help others discover more about the Goodreads dataset, once again showing the scale of the dataset and how many books fall into certain categories.

Lastly, Chapter 5 focused more on multivariate probability distributions, looking primarily at the probability distribution of publishers. The book publishers chosen were Scholastic, Crown, and Bantam Books, and for the case of problems within Chapter 5’s sections, there were ten books published by Scholastic, seven by Crown, and three by Bantam Books. It was found that there would be a 0.7% chance that, when selecting two books, none would be Scholastic, with choosing a Scholastic book (X) and choosing a Crown book (Y) being dependent events overall.

# References

Amazon. (2024). *About goodreads*. Goodreads. https://www.goodreads.com/about/us.

Soumik. (2020, March 9). *Goodreads-books*. Kaggle. https://www.kaggle.com/datasets/jealousleopard/goodreadsbooks.

Wackerly, D. D., Mendenhall, W., & Scheaffer, R. L. (2012). *Mathematical statistics with applications*. Brooks/Cole.