Ian Valiante

Part 2

**Introduction**

When presented with the opportunity to choose my own dataset, I thought it would be a good idea to write this report about something I’m interested in. That thing for me is bouldering, which is a discipline of rock climbing where you do not have a rope, but instead climb a wall (typically between 15-20ft high) with pads below. Most boulderers climb inside gyms that have plastic holds on walls to create boulder problems. These boulder problems can vary in difficulty, from essentially a ladder at the lowest difficulty to holding onto quarter sized holds. For most, this is training to then go outside and climb real rocks; but for some this is training for competitions.

These competitions are what we will be focusing on, as there is data present for them. The goal of bouldering competitions is to complete the most amount of boulder problems with the least number of attempts in a set period. The International Federation of Sport Climbing (IFSC) is where all the top competition climbers compete. The format for IFSC competitions is as follows:

* Different problems are set for men and women.
* The problems are reset between the Qualification, Semifinals and the Final.
* In Finals, competitors can preview the boulder problems during a collective observation time (2 minutes per Boulder) but cannot attempt the problems. In Qualifications and Semifinals, the climbers observe the problems for the first time during their first attempt (no observation prior to the round).
* Competitors are kept in an isolation room before they perform their “on sight” attempt.
* The Boulder ranking is decided by the number of problems solved. The competitor who solves the most problems wins.
* One zone hold (approximately the halfway point on a problem) is set per problem.
* The Boulder ranking is based on: 1. Number of tops reached, 2. Number of zone holds reached 3. Number of attempts to top, 4. Number of attempts to zone.

(https://www.ifsc-climbing.org/boulder/index)

The data used for this report was generated using a python script which implements Selenium to scrape the official ifsc website and generate the relevant csvs. I then used another python script to perform some data analysis to generate histograms.

**Chapter 1**

***Section 1.2***

Construct a relative frequency histogram for tops and zones in bouldering. Consider the gender of the athletes.

**A graph of different colored squares

Description automatically generated with medium confidence**

***Section 1.3***

Based off the Histogram, what is the mean number of tops for a women athletes, men? What is the mean number of zones for each sex?

Since these are weighted values, we must use the following formula:

Men Tops:

Sum of Weighted Values = (0 x 5) + (1 x 5) + (2 x 13) + (3 x 9) + (4 x 4) = 0 + 5 + 26 + 27 + 16 = 74

Total Frequency = 5 + 5 + 13 + 9 + 4 = 36

2

Women Tops (Not showing work for decreased redundance):

Men Zones:

Women Zones:

**Chapter 2**

**Section 2.3**

Suppose we are analyzing the medal counts of countries in bouldering competitions. Let the set S  represent all possible medal counts for countries:

S = {FRA, JPN, USA, KOR, SLO, GBR, BEL, CZE, GER, SRB}

Define the following subsets:

• A: The subset of countries with more than 5 medals.

• B: The subset of countries with exactly 1 medal.

• C: The subset of countries with fewer than 3 medals.

List the elements of:

• A, B, C

• A B

• A B

• A C

• A C

• B C

• B C

• C B

A = {FRA, JPN, USA}, B = {BEL, CZE, GER, SRB}, C = {GBR, BEL, CZE, GER, SRB},

These sets were determined by the following graph which was generated via a Python script: A graph of blue bars

Description automatically generated

To get the intersections and unions of the sets I used my StatsLibrary program, which gave me the following sets:

A screenshot of a computer program

Description automatically generated

**Section 2.4**

**A company tracks the number of medals won by countries in a bouldering competition. The top five countries in the competition are: FRA (12 medals), JPN (9 medals), USA (7 medals), KOR (5 medals), and SLO (3 medals).**

a. List the sample space for this experiment, S , where the sample points representcountries winning medals.

S = {FRA, JPN, USA, KOR, SLO}

b. Let A denote the event that a randomly selected country wins more than 5 medals. List the sample points in A.

S = {FRA, JPN, USA}

c. Construct a Venn diagram for the experiment, illustrating event  A .

A screenshot of a computer

Description automatically generated

d. Assign probabilities to each country based on the total number of medals won by the top five countries, ensuring the probabilities follow the axioms of probability.

The total number of medals won by the top five countries: 12 + 9 + 7 + 5 + 3 = 36.

P(FRA) = P(JPN) = P(USA) = P(KOR) = P(SLO) =

e. Find the probability of event A  (i.e., the event that a randomly selected country wins more than 5 medals).

**P(A) = P(FRA) + P(JPN) + P(USA) =**  + = = .77 = 77%

**Section 2.5**

A competition organizer wants to highlight two countries from the top five medal winners for a special feature. The countries vary in their total medals: FRA (12 medals), JPN (9 medals), USA (7 medals), KOR (5 medals), and SLO (3 medals). These medal counts are known to the organizer. Define two events A and B  as:

•A: The organizer selects the country with the highest medal count (FRA) and the country with the lowest medal count (SLO).

•B: The organizer selects at least one of the top two medal-winning countries (FRA or JPN).

Find the probabilities of these events.

|S| = )= 10

S = {{FRA, JPN}, {FRA, USA}, {FRA, KOR}, {FRA, SLO}, {JPN, USA}, {JPN, KOR}, {JPN, SLO}, {USA, KOR}, {USA, SLO}, {KOR, SLO}}.

 A = {FRA, SLO} since FRA has the most number of medals, and SLO has the least amount.

P(A) = , so a 10% chance the organizer selects the country with the highest medal count (FRA) and the country with the lowest medal count (SLO).

B = {{FRA, JPN}, {FRA, USA}, {FRA, KOR}, {FRA, SLO}, {JPN, USA}, {JPN, KOR}, {JPN, SLO}} because FRA and JPN have the most medals, we take all elements from S that include one AND OR the other.

P(B) = , so a 70% chance the organizer selects at least one of the top two medal-winning countries (FRA or JPN).

**Section 2.6**

A bouldering competition organizer decides to highlight countries based on two criteria:

•The **gender** of the medal winners (Men or Women).

•The **country** of the medal winners (FRA, JPN, USA, KOR, SLO).

How many different combinations of gender and country can the organizer choose to feature?

2 genders, so we have g = = 2

5 countries, so we have c = = 5

Since any gender can pair with any country, we do c x g = 5 x 2 = 10. The organizer can choose to feature 10 different combinations of gender and country.

**Section 2.7**

We know:

P(A) = 0.6, P(B) = 0.4, P(A B) = 0.3.

Find:

a.  P(A|B) : The probability that a selected country wins more than 5 medals, given that it is FRA or JPN.

P(A|B) = = = .75

b.  P(B|A) : The probability that a selected country is FRA or JPN, given that it wins more than 5 medals.

P(B|A) = = = .5

c.  P(A|A B) : If we know a country either wins more than 5 medals or is FRA/JPN, what is the likelihood that it specifically wins more than 5 medals?

P(A|A B) =

P(A B) = P(A)+P(B) – P(A B) = .6 + .4 - .3 = .7

P(A|A B) = = .857

So the likelihood that it wins more than 5 medals it 85.7%

d.  P(A|A B) : The probability that a selected country wins more than 5 medals, given that it satisfies both  A  and  B .

Since A implies B, P(A|A B) = 1

e.  P(A B|A B) : What is the probability that a country **wins more than 5 medals and is FRA or JPN**, given that the country either **wins more than 5 medals** or **is FRA or JPN**?

P(A B|A B) = = .429

So there’s a 42.9% that a country wins more than 5 medals and is FRA or JPN given that the country either wins more than 5 medals or is FRA or JPN.

**Section 2.8**

Can  A  (the event that a country wins more than 5 medals) and  B  (the event that a country is FRA or JPN) be mutually exclusive if:

1. P(A) = 0.6  and  P(B) = 0.4 ?

2. P(A) = 0.6  and  P(B) = 0.3 ?

In **Case 1**,  A  and  B  can be mutually exclusive if their union covers the entire sample space.

In **Case 2**,  A  and  B  can be mutually exclusive because their combined probability ( P B) = 0.9 ) is less than or equal to 1.

**Section 2.10**

In a bouldering competition, FRA won **12 medals**, and the other countries combined (JPN, USA, KOR, and SLO) won **24 medals**. It is reported that **7 of FRA’s medals** and **15 of the other countries’ medals** were won by men. A randomly selected medal is found to have been won by a man. What is the conditional probability that the medal was won by FRA?

P(FRA | Men) =

P(FRA) = = 0.333, P(Other Countries) = = = 0.667.

P(Men | FRA) = = 0.583.

P(Men | Other Countries) = = 0.625.

P(Men) = P(Men | FRA) x P(FRA) + P(Men | Other Countries) x P(Other Countries).

P(Men) = 0.194 + 0.417 = 0.611.

Now, we calculate P(FRA|Men)

P(FRA|Men) = = = .318

So, The conditional probability that the medal was won by FRA, given it was won by a man, is approximately: 0.318 or 31.8%.

**Chapter 3**

**Section 3.2**A graph of different colored squares

Description automatically generated with medium confidence

If a randomly chosen athlete from the competition is selected, given the histogram determine the following:

1.The probability distribution for  Y , the total number of **tops completed** by the athlete.

A table with numbers and letters

Description automatically generated

2.The probability distribution for  Z , the total number of **zones reached** by the athlete.

A table with numbers and lines

Description automatically generated

3.Using the data provided, calculate the mean number of tops and zones achieved by male and female athletes, separately.

**Tops:**

Mean for Men:  1.97

Mean for Women:  1.75

**Zones:**

Mean for Men:  3.15

Mean for Women:  3.29

**Section 3.4**

Consider the population of athletes in a bouldering finals competition. Based on the histogram data, suppose there are N = 500 athletes in the competition, and 60% of them reached at least one zone during the event. Define reaching at least one zone as a “success” S. The probability of S on the first trial is 0.60.

Now consider the event B, which represents the event that S  (reaching at least one zone) occurs on the second trial.  B can occur in two ways:

1. The first two trials are both successes.

2. The first trial is a failure, and the second trial is a success.

Using this context:

1. Show that P(B) = 0.60 .

2. What is P(B | the first trial is S) ?

3. Does this conditional probability differ markedly from  P(B) ?

1.

P(B) = P(Both are successes) + P(First trial is a failure, second is a success)

P(Both are successes) = P(S) x P(S) = .6 x .6 = 0.36

P(First trial is a failure and second is a success) = P(F) x P(S) = .4 x .6 = .24

So, P(B) = .36 + .24 = .6

2.

P(B | the first trial is S) = P(Second trial is a success |First trial is a success)

Since they are independent, P(B | first trial is S) = P(S) = .6

3.

No, because P(B) = .6 and P(B | the first trial is S) = .6, this is because the trials are independent.

**Section 3.5**

In a bouldering competition, 60% of athletes reached at least one zone during the event, while 40% did not. Athletes are selected at random for interviews about their performance. Find the probability that the first athlete who reached at least one zone is selected on the 4th interview.

P(S) = .6

K = 4

With these values I can use my “geometricDist” method in my StatsLibrary to calculate the geometric distribution. 

Which means there is a 3.84% the first athlete who reached at least one zone will be selected on the fourth interview.

**Section 3.6**

In a bouldering competition, 60% of athletes reached at least one zone during the event, while 40% did not. Interviews are conducted to find three athletes who reached at least one zone. What is the probability that 10 athletes must be interviewed in order to find three athletes who reached at least one zone?

Since we want 10 athletes, n = 10

Since we want athletes who reached a zone, r = 3

Probability of reaching at least one zone, p = .60

Compliment of p, q = .4

Using the “negativeBinomialDistribution” method in my StatsLibrary I determined the probability that 10 athletes must be interviews to find three athletes who reached at least one zone is: 

Or approximately 1.27%.

**Section 3.7**

In a bouldering competition, athletes are evaluated based on their number of tops completed and zones reached. From the competition:

• 9 male athletes and 7 female athletes completed exactly 3 tops.

• 14 male athletes and 18 female athletes reached exactly 4 zones.

Suppose you randomly select six athletes from this competition. The first four athletes selected are tested for their performance, and all are found to have completed 3 tops. The remaining two athletes are sold “tickets” for priority qualification and are found to have reached 4 zones.

What is the probability that the six athletes selected are such that:

• The first four all completed exactly 3 tops.

• The last two all reached exactly 4 zones.

N = 48, the total amount of athletes

n = 6, total athletes selected

y = 4, number of athletes selected from those whom completed 3 tops

r = 2, number of athletes selected from those whom reached 4 zones.

Using the hypergeometricDistribution method from my StatsLibrary I determined the probability that the six athletes selected are such that the first four all completed exactly 3 tops and the last two all reached 4 zones is:



Or approximately 7.35%.

**Section 3.7**

Athletes successfully completing tops during a bouldering finals follow a Poisson distribution, with an average of 2.5 tops per athlete in a given attempt. Based on this, calculate the following probabilities:

a. What is the probability that an athlete completes no more than 3 tops?

b. What is the probability that an athlete completes at least 2 tops?

c. What is the probability that an athlete completes exactly 5 tops?

Since the average is 2.5 tops per athlete, . We will use this for a, b and c.

1. To calculate this, we need the cumulative probability.

This means, P(X 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)

Using my Poisson distribution method from StatsLibrary I determined the cumulative probability:

A black background with white text

Description automatically generated

So, the probability that an athlete completes no more than 3 tops is 75.75%

b. To calculate this, we must calculate P(X

P(X = 1- (P(X=0) +P(X=1))

Using the same method from part a I determined this to be:

A black background with white letters

Description automatically generated

So, the probability that an athlete completes at least 2 tops is 71.27%.

c. We need P(X=5)

Using the same method I determined this to be:



So, the probability that an athlete completes exactly 5 tops is 6.68%.

**Chapter 4**

**Section 4.2**

In a bouldering competition, 18 female athletes and 14 male athletes successfully completed 4 zones. These athletes are randomly called for interviews, one at a time, until a female athlete is interviewed. Assume athletes are selected without replacement.

Let  Y  represent the number of the interview on which the first female athlete is called.

a. Find the probability function for  Y .

b. Give the corresponding distribution function.

c. What is  P(Y < 3) ?  P(Y 3) ?  P(Y = 3) ?

d. If  Y  is a continuous random variable, we argued that, for all  - < a < ,  P(Y = a) = 0 . Do any of your answers in part (c) contradict this claim? Why?

Total athletes:  N = 18 (female) + 14 (male) = 32

Female athletes: F = 18

Male athletes: M = 14

Y: The trial on which the first female athlete is interviewed

a.

P(Y=y) =

Where:

M = 14

F = 18

N = 32

Y = trial number

b.

F(Y=y) =

c.

P(Y <3) = P(Y=1)+P(Y=2)

First athlete is female, second is male

P(Y=1) = =

First athlete is male, second is female

P(Y=2) = = .2538

P(Y = 3) = .5625 + .5070 = .8163

P(Y = 3) = = .4375 x .4194 x .6 = .11

P(Y = P(Y = 1) + P(Y = 2)+ P(Y = 3)

P(Y = .5625 + .2538 + .11 = . 9263

d.

No, the discrete nature of Y ensures P(Y=a) >0, which does not contradict the continuous random variable chain.

**Section 4.4**

In a bouldering competition, climbers are scored based on their performance in completing zones and tops. Suppose a climber’s performance can be visualized on a scaled line between two markers, **Z** (representing zones reached) and **T** (representing tops completed), with the line running from **0** (no performance) to **10** (perfect performance).

1.What is the probability that a climber’s performance score is closer to **Z** (zones) than to **T** (tops)?

2.What is the probability that the climber’s score is such that the distance to **Z** is more than twice the distance to **T**?

1.

P(Closer to Z) = = = 0.5.

Or 50%

2. For the distance to  Z  to be more than twice the distance to  T :

x > 2 (10 - x)

x >

The probability is the ratio of the length of the interval (,10]  to the total length of  [0, 10] :

P(Distance to Z more than twice distance to T) =

= 10 - =

So, P(Distance to Z more than twice distance to T) = = = .3333

Or 33.33%

**Section 4.10**

In a bouldering competition, climbers are evaluated based on the number of zones they successfully complete. On average, climbers complete  = 3 zones. The competition organizers want the actual number of zones completed, Y, to be within 1 zone of ) at least 75% of the time.

What is the largest value of  , the standard deviation of  Y , that can be tolerated to meet the competition organizers’ objectives?

Chebyshev’s inequality states:

P(|Y-|) < k) ,

K: num of standard deviations within which the specified percentage of the data must lie

The organizers want P()

This corresponds to:

Solve for , k = 2

1 = k •

1 = 2 •

So, the largest value of that can be tolerated is 0.5.

**Section 4.11**

In a bouldering competition, athletes score points based on the number of attempts they make to complete a climb. The number of attempts,  Y , follows a continuous uniform distribution over the interval from 1 to 4 attempts. Athletes start with 10 points, but for every attempt they make, they lose 0.1 points.

For example:

• If an athlete takes 1 attempt, their score is  10 - 0.1(1) = 9.9  points.

• If an athlete takes 3.5 attempts, their score is  10 - 0.1(3.5) = 9.65  points.

What is the expected score for an athlete based on the number of attempts they make?

S(Y) = 10 - 0.1Y,

E[S(Y)] = ,

f(Y) =

E[S(Y)] = ,

E[S(Y)] = •29.25 = 9.75.

The expected score for an athlete based on number of attempts is 9.75 points.

**Chapter 5**

**Section 5.2**

In a bouldering competition, two distinct climbing routes are randomly assigned to one or more of three climbers: Alice (A), Bob (B), and Carol (C). Let  Y1  denote the number of routes assigned to Alice, and  Y\_2 denote the number of routes assigned to Bob. Each climber can receive 0, 1, or 2 routes.

a. Find the joint probability function for  Y1  and  Y2 .

b. Find  F(1, 0) , the cumulative probability that Alice is assigned at most 1 route and Bob is assigned 0 routes.

a.

A table with numbers and letters

Description automatically generated

b.

F(1,0) =

In a bouldering competition, climbers’ performance is measured by two metrics: the number of **tops completed** ( Y\_1 ) and the number of **zones reached** ( Y\_2 ). Assume the joint density function for  Y\_1  and  Y\_2  is:

f(y\_1, y\_2) =

\begin{cases}

1 & \text{if } 0 \leq y\_1 \leq 4 \text{ and } 0 \leq y\_2 \leq 4, \\

0 & \text{otherwise.}

\end{cases}

Answer the following questions:

a. Find the marginal density functions for  Y1  and  Y2 .

b. What is P(1 < Y1 < 3) ?  P(1 < Y2 < 3) ?

c. For what values of  y2  is the conditional density  f(y1 | y2)  defined?

d. For any y2 ,  0 \≤ y2 ≤ 4 , what is the conditional density function of  Y1  given  Y2 = y2 ?

e. Find P(1 < Y1 < 3 | Y2 = 2) .

f. Find P(1 < Y\_1 < 3 \mid Y\_2 = 3) .

g. Compare the answers you obtained in parts (a), (d), and (e). For any y2 ,  0 ≤ y2 ≤ 4 , how does  P(1 < Y1 < 3)  compare to  P(1 < Y11 < 3 | Y2 = y2) ?

a.

fy1(y1) =

fy1(y1) = = • (4-0) =

fy1(y2) =

fy1(y2) = = • (4-0) =

b. **Find P**(1 < Y1 < 3)  **and** P(1 < Y2 < 3)

P(1 < Y2 < 3) =

P(1 < Y2 < 3) = • (3-1) =

**c. For What Values of** Y2  **is** f(Y1 | Y2)  **Defined?**

The conditional density  f(Y1 | Y2)  is defined when  fY2(y2) > 0 .

From part (a),  fY2(y2) =   for  0 ≤ y2 ≤ 4 .

f(Y1 | Y2) is defined for 0 ≤ y2 ≤ 4.

d. **. Conditional Density Function** f(Y1 | Y2) **:**

The conditional density is given by:

f(Y1 | Y2) =

Substitute:

• f(Y1, Y2) = ,

• fY2(y2) = .

f(Y1 | Y2) = , 0 ≤ y1 ≤ 4, 0 ≤ y2 ≤ 4.

**e. Find P**(1 < Y1 < 3 | Y2 = 2) **:**

The conditional probability is:

P(1 < Y1 < 3 | Y2 = 2) =

Substitute

P(1 < Y1 < 3 | Y2 = 2) =

The conditional probability  P(1 < Y1 < 3 | Y2 = y2)  is equal to the marginal probability  P(1 < Y1 < 3)  because  Y1  and  Y2  are independent in this uniform distribution.