

Probability & Data

Week 1 – Intro to Probability

DR NICK FEWSTER-YOUNG

Topics to be covered

- Welcome and Course Structure (Housekeeping)
- Why study Probability & Data?
- Statistical Software
- Introduction to Probability & Set Theory
- Examples

Course Coordinator

- Dr Nick Fewster-Young
Y 3-51, City West Campus
Email: nick.fewster-young@unisa.edu.au

Course Information

- Course Website – Lecture notes, exercises and data.
- Course Outline – What this course expects, assessments and course overview.
- References – See course website
- Software – R (download it, its free). We may use Matlab and Excel.

Teaching Arrangements

- Workshop (2 hours)
 - A mixture of theory, examples and implementation in R.
 - There will be a combination of lecture style and practically working on problems as an individual or group.
- Tutorial/Consultation
 - Each week, there is 1 hour dedicated to us working out problems and having consultation.

Assessments

- Continuous Assessment (Specific Questions chosen from tutorial questions each week)
- Assignment 1
 - Probability and Set Theory
 - Discrete Distributions
 - Real world problem
- Assignment 2
 - Applying Statistical methods to solve a real world problem
 - Producing a report
- Examination
 - Examines topics from the whole semester.

Why Probability and Data?

- Data drives the world is the basis for good decision making both for the government and companies.
- Probability utilises data to make inform and make good decisions.
- Statistics is concerned with the [collection, analysis, and interpretation of data](#), as well as the effective communication and presentation of results relying on data.
- The vital key to understanding and passing statistics is [knowing](#) when to use what. There are so many different tests and each can only be used in certain circumstances.
- It's easy to get confused and make mistakes with [the large number](#) of statistical methods we will learn.

Probability vs. Statistics

- Different subjects: both about random processes
- Probability
 - Logically self-contained
 - A few rules for computing probabilities
 - One correct answer
- Statistics
 - Messier and more of an art
 - Get experimental data and try to draw probabilistic conclusions
 - No single correct answer

Statistics

- **Statistics** is a very *practical discipline*, concerned about *real problems* in the real world! In the real world you will need to *quantitatively* describe numerical data.
- You will have to be able to *analyse large amounts of data*. More and more massive databases are being compiled.
- Answers provided by statistical approaches can provide the basis for making decisions or choosing actions.
- Sometimes you can find what you're looking for. Other times you will find something you never knew existed.
- Statistics applies to almost any field.

How can I do well?

- Keep on top of everything and work regularly on problems and implementation in R.
- Stay Positive – there is going to be times where the data or problem may not like you, so stay positive.
- Follow up on any questions you have!
- Start work on the Assignments early!

What we will learn in the next few weeks

- Learn how to define an outcome 'space'.
- Learn the basics of probability and data.
- Learn how to identify the different types of data: discrete, continuous, categorical, binary.
- Learn how statistical programs can construct histograms for discrete & continuous data.
- Learn about the various discrete and continuous probability distributions
- Expectation and Variance
- Dealing with Data sets and applying theory to applications

DATA - Collection

➤ We need to understand our data before we play with it! So can you answer these questions about the data collection process?

- How was it collected?
- Is it a sample?
- Was it properly sampled?
- Was the dataset transformed in any way?
- Are there some known problems on the dataset?

Data - Structure

A recommended and standard way to structure data is to

- Each variable forms a column and contains values
- Each observation forms a row
- Each type of observational data forms a table

If we follow such an approach then it will speed up your analyse since it is easy to visualise and compatible with many statistical tools and libraries.

Example

Course Website – datasets: [beer.csv](#) and [breweries.csv](#) (courtesy of Kaggle)

The structure of the dataset is:

Beers:

ID: Unique identifier of the beer.

Name: Name of the beer.

ABV: Alcohol by volume of the beer.

IBU: International Bittering Units of the beer.

Style: Style of the beer.

Ounces: Ounces of beer.



Example

The other dataset's structure:

Breweries:

ID: Unique identifier of the brewery.

Name: Name of the brewery.

City: City where the brewery is located.

State: State where the brewery is located.

R

- Free open source package.
- Very easy to use and install.
- There is a large amount of documentation, help and forums on how to learn, use and implement R.
- Let's download from the course website 😊.
- Let's bring up the beers and breweries datasets in R.

Importing & reading data into R

- Rstudio has a GUI interface that allows you to import datasets very nicely!
- The commands that we will see are the following:
 - `read_csv("filename")` – read csv file
 - `dir()` – working directory
 - `str(dataset)` – structure of the dataset
 - `read_delim("filename")` -- read txt file
 - `read_table("filename")` – read table (more complex)
 - `read_xlsx("filename")` – read xlsx file

Data Types

➤ Quantitative and Qualitative Data Types

- Quantitative Data Types

1. Discrete, e.g. number of children, number of students in a class
2. Continuous, e.g. weight, height, currency, time, distance

- Qualitative Data Types

1. Nominal, e.g. $S = \{\text{yes, no}\}$, eye color = (brown, blue, hazel, green)
2. Ordinal/Categorical, e.g. grades = {F2, F1, P2, P1, C, D, HD}.

Example: Beers

Each of the columns in the dataset are variables. Each of the variables have a specific data type:

Type: `str(beers)` – obtain the structure of `beers.csv`

Beers:

ID: Discrete

Name: Nominal

ABV: Continuous

IBU: Discrete

Style: Categorical

Ounces: Categorical

Brewery_id: Categorical

Probability - Introduction

We know that the concept of probability is a number between **0 and 1**.

It relates to the chance of an observation or event occurring!

$$P(A) = 0$$

- Means that the event A is nearly impossible to occur; while

$$P(A) = 1$$

- Means that the event A will nearly certainly occur.
- How does an event get assigned a particular probability value? Well, there are three ways of doing so:
 - The personal opinion approach;
 - The relative frequency approach;
 - The classical approach.

Examples

- ❖ At which end of the probability scale would you put the probability that:
 - a. You will win the lottery some day?
 - b. A randomly selected student will get an A in this course?
 - c. *You* will get an A in this course?

Probability – Concepts and Counting

A simple example of probability which we see everyday is in counting and the chance an event occurs.

Formally, the **probability** of the **event A** occurring in **S** is the number of events comprising of A occurs over the total number of events in S, that is

$$P(A) = \frac{\text{number of events comprising } A}{\text{total number of events in } S}$$

For example,

- What is the probability of getting exactly 1 heads in 3 tosses of a fair coin?

Heads or Tails

- What is the probability of getting exactly 1 heads in 3 tosses of a fair coin?

Let's toss a single coin then this produces "outcomes", so in this single experiment, we have
 $\{Head, Tail\}$.

If we make three tosses then the experiment has the following outcomes,
 $\{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$.

- Therefore, there are eight outcomes and there are three outcomes where there is exactly one Head.
So.....

- Answer $\rightarrow P(1\ Head) = \underline{\hspace{2cm}}$

- ✓ That was an easy example, let's try our luck with cards.

Time to play cards -- Decks

- Suppose you have a deck of 50 cards with five players consisting:
 - 5 ranks: Rainbow Kitten, Tacocat, Hairy Potato, Watermelon Cat and Bearded Cat [4 of each]
 - 5 Action Cards: Attack, Skip, Favour, Nope, Vision (See the Future) [4 each]
 - Defuse Cards: 6 Defuse cards
 - Exploding Kitten: There are 4 of these.
- The Game:
 - Each player is dealt four cards and one defuse card (No exploding kittens)
 - The exploding kittens and extra defuse card are placed randomly somewhere in the deck.
 - Simply, each player picks up at the end of their turn and blows up if they get an exploding kitten unless they can defuse it.
- Calculate the probability:
 1. You pick up an exploding kitten straight up.

Sets

➤ A **set** is a collection of distinct observations or events.

- All the different card types make up a set (S) in the Exploding Kitten card game (represented by letters):

$S := \{E, D, R, T, H, W, B, A, S, F, N, V\}$

- The action cards can be represented as a set (or subset):

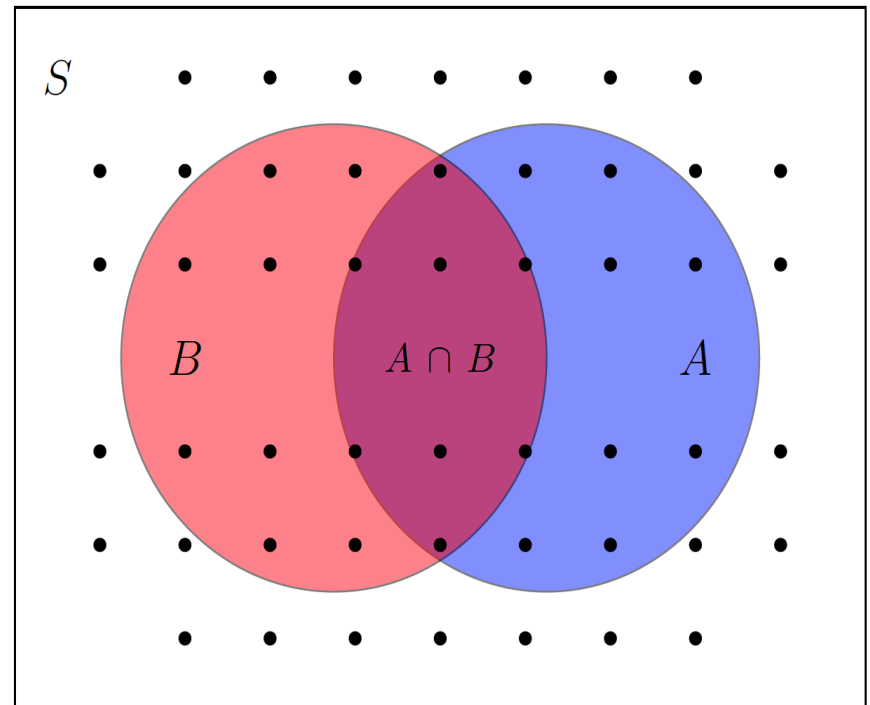
$\text{Action} := \{A, S, F, N, V\}$

- And...so do the defuse cards:

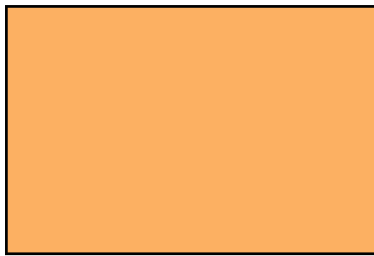
$\text{Defuse} := \{D\}$

- We could have both Action **OR** Defuse together:

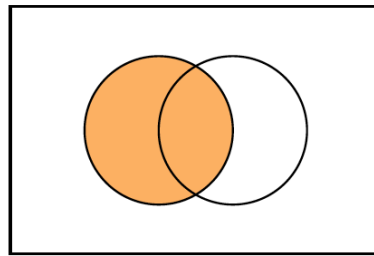
$\text{Action} \cup \text{Defuse} := \{A, S, F, N, V, D\}$



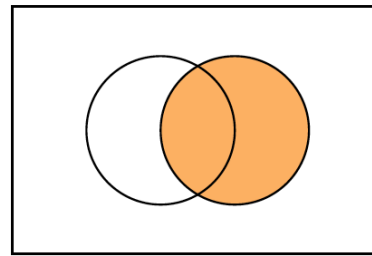
Notation



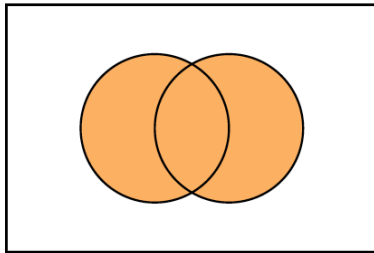
S



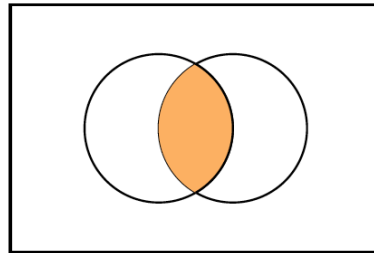
$Action$



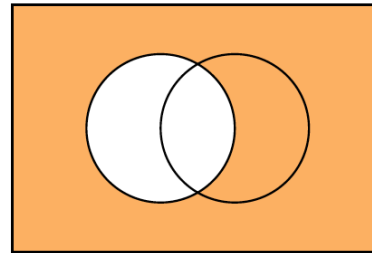
$Defuse$



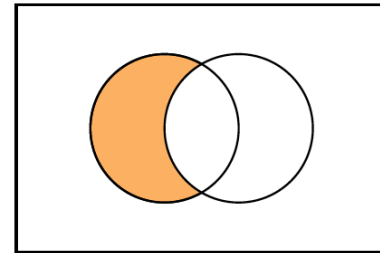
$Action \cup Defuse$



$Action \cap Defuse$



$Action^c$



$Action - Defuse$

A quick set question

➤ A data science team consists of statisticians and computer scientists.

10 people are computer scientists

4 people are statisticians

2 do both

- How many people are in the team?

A quick set question

➤ A data science team consists of statisticians and computer scientists.

10 people are computer scientists

4 people are statisticians

2 do both

- How many people are in the team?
 - Let A = Computer Scientists, B = Statisticians. See that $A \cap B = 2$.
 - Therefore, $|A \cup B| = |A| + |B| - |A \cap B| = 10 + 4 - 2 = 12$.

Permutations

Lining things up.

How many ways can you do it? 'abc' and 'cab' are different permutations of {a, b, c}

Permutations of k from a set of n

- Give all permutations of 3 things out of {a, b, c, d}
 - abc abd acb acd adb adc
bac bad bca bcd bda bdc
cab cad cba cbd cda cdb
dab dac dba dbc dca dcb

- Would you want to do this for 7 from a set of 10?

No way!

Combinations of k from a set of n (distinct)

➤ Give all combinations of 3 things out of $\{a, b, c, d\}$.

- Answer: $\{a,b,c\}$, $\{a,b,d\}$, $\{a,c,d\}$, $\{b,c,d\}$

Permutations vs. Combinations

- abc abd acb acd adb adc
bac bad bca bcd bda bdc
cab cad cba cbd cda cdb
dab dac dba dbc dca dcb

{a, b, c}

{a, b, d}

{a, c, d}

{b, c, d}

Permutation

Math:
$$\frac{n!}{(n-r)!} = \frac{n(n-1)(n-2)\dots 3 \times 2 \times 1}{(n-r)(n-r-1)\dots 3 \times 2 \times 1}$$

Eg. $4! = 4 \times 3 \times 2 \times 1 = 24$

Combination

Four "choose" three : $C(n, r) = {}^nC_r = {}_nC_r = \binom{n}{r} = \frac{n!}{r!(n-r)!}$

$$\binom{n}{k} = \binom{4}{3} = \frac{4!}{3!(4-3)!} = 4$$

The Lottery 😊

The numbers are drawn one at a time, and if we have the lucky numbers (no matter what order) we win!

So let's say there are 36 distinct numbers and we pick 8 numbers!

Assume that the order does not matter (ie combinations).

What is the probability that we win?

Well there are 30260340 different distinct combinations! Ohh...no! 8 out of 36 looked good odds.

But to win, that is now 1 out of 30260340 \longrightarrow $Probability(win) = \frac{1}{30260340} = 0.000000033$

Betting odds and Horses

From the beginning of the Bayesians to present day, odds have existed. People place on bets on winning.

In the present, we now are using large data sets and probability to determine the odds.

Let's look at a simple example from the Adelaide Cup. In 2017, a few horses pulled out and left only 8 horses in the race. Let's have a look at the odds and suppose you put \$2 down on each horse. What is the potential maximum loss and gain?

Horse	Win	Horse	Win
#1	\$3	#5	\$9.5
#2	\$3	#6	\$10
#3	\$5	#7	\$30
#4	\$6	#8	\$100

Betting odds and Horses

- Where did those odds come from and what are the probabilities of a certain horse winning?
- Assume you bet \$2 on each of the eight horses and the track wants to take 20%, from the amount collected. What would be the probability of Horse #1 and Horse #8?
- Suppose Horse #1 won then the Payout = \$6.

The track would take 20%, which leaves you with the absolute payout = $6/0.8 = \$7.5$.

An odd is a relative percentage or ratio based on your payout to bet, so

$$Odd = \frac{Payout - Bet}{Bet} = \frac{7.5 - 2}{2} = \frac{5.5}{1}.$$

It looks like good odds, (remember 5:1 odd means 5 chances of losing and 1 chance to win) the probability that this horse wins would be

$$P(\text{Horse \#1 wins}) = \frac{1}{Odd + 1} = 0.157.$$

So a 15% chance of winning if we bet on Horse #1.

- Repeat for Horse #8 and see that we get 0.8% chance.

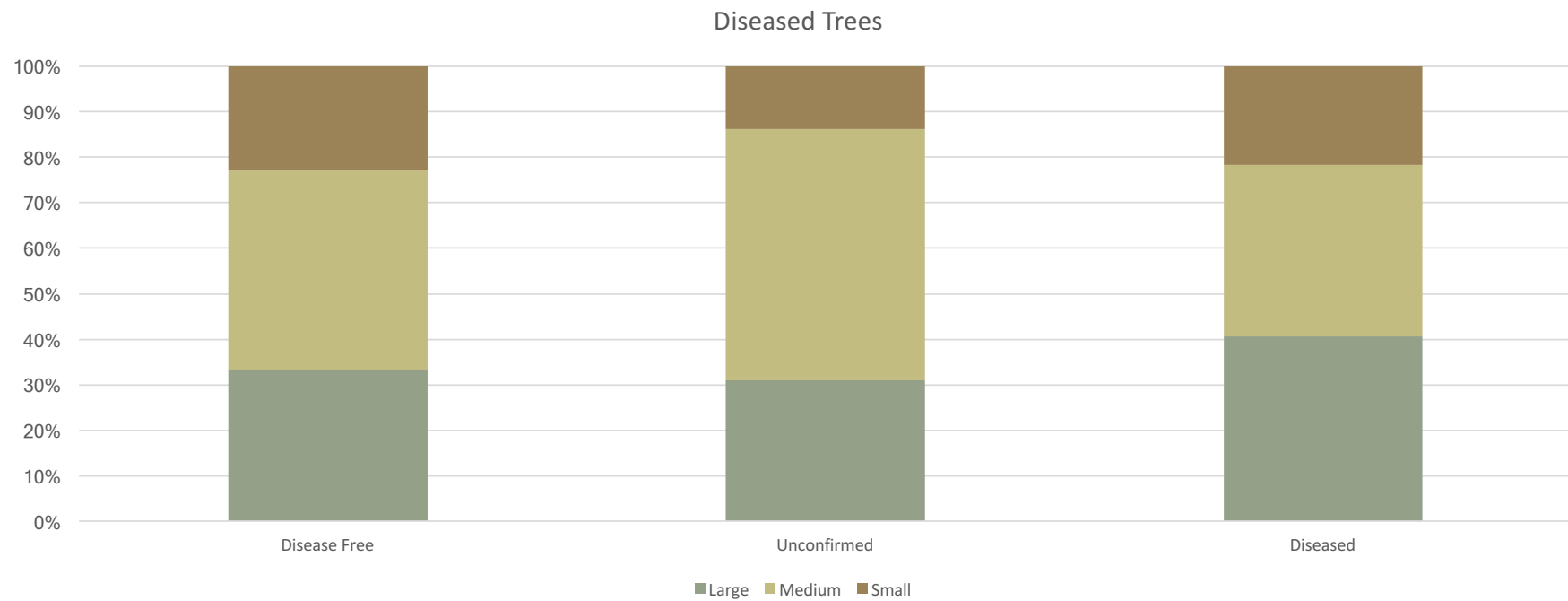
Probability & Data – Tree Disease Example

- A random sample of 200 trees of various sizes was examined in the Adelaide Hills yielding the following results:


Type	Disease Free	Unconfirmed	Diseased	Total
Large	35	18	15	68
Medium	46	32	14	92
Small	24	8	8	40
Total	105	58	37	200

- This is a **table of counts** – it counts number of occurrences for each joint observation.
- What does this data say???

100% Stacked Column Chart



Diseased Trees

- a. What is the probability that one tree selected at random is large?
 - b. What is the probability that one tree selected at random is diseased?
 - c. What is the probability that one tree selected at random is both small and diseased?
 - d. What is the probability that one tree selected at random is either small or disease-free?
 - e. What is the probability that one tree selected at random from the population of medium trees is unconfirmed of disease?
- 

Diseased Trees

- a. What is the probability that one tree selected at random is large?

$$P(X = large) = \frac{68}{200}$$

- b. What is the probability that one tree selected at random is diseased?

$$P(Y = Diseased) = \frac{37}{200}$$

- c. What is the probability that one tree selected at random is both small and diseased?

$$P(small \cap diseased) = \frac{8}{200}$$

- d. What is the probability that one tree selected at random is either small or disease-free?

$$P(small \cup disease - free) = \frac{35 + 46 + 24 + 8 + 8}{200} = \frac{121}{200}$$

- e. What is the probability that one tree selected at random from the population of medium trees is doubtful of disease or disease-free?

$$P(medium \cap (doubtful \cup disease - free)) = \frac{46 + 32}{200} = \frac{78}{200}$$

Probability Axioms

Let's revisit some axioms and more on set theory – only briefly though.

- Complement Rule

$$P(A) = 1 - P(A^C)$$

- Empty Set

$$P(\emptyset) = 0$$

- If A subset of B then

$$P(A) \leq P(B)$$

- Set Rule #1

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Example

Example: If 7% of the population uses walking as a mode of transport, 28% of the population uses a car as a mode of transport, and 5% of the population uses both, what percentage of the population neither uses walking or a car as a means of transport?

- $P(\text{walking}) = 0.07$
- $P(\text{Car}) = 0.28$
- $P(\text{walking} \cap \text{car}) = 0.05$

Solution:

$$\begin{aligned} P\left((\text{walking} \cup \text{car})^c\right) &= 1 - P(\text{walking} \cup \text{car}) \\ &= 1 - P(\text{walking}) - P(\text{car}) + P(\text{walking} \cap \text{car}) = 1 - 0.3 = 0.7 \end{aligned}$$

Example

- A company has bid on two large construction projects. The company president believes that the probability of winning the first contract is 0.6, the probability of winning the second contract is 0.4, and the probability of winning both contracts is 0.2.
 - a. What is the probability that the company wins at least one contract?
 - b. What is the probability that the company wins the first contract but not the second contract?
 - c. What is the probability that the company wins neither contract?
 - d. What is the probability that the company wins exactly one contract?

Example

- A company has bid on two large construction projects. The company president believes that the probability of winning the first contract is 0.6, the probability of winning the second contract is 0.4, and the probability of winning both contracts is 0.2.

- a. What is the probability that the company wins at least one contract?

$$P(1st\ win) = 0.6, \ P(2nd\ win) = 0.4, \ P(wins\ both) = 0.2. \text{ Therefore,} \\ P(At\ least\ one) = P(1st\ win \cup 2nd\ win) = 0.6 - 0.4 + 0.2 = 0.8$$

- b. What is the probability that the company wins the first contract but not the second contract?

$$P(1st\ win \cap (2nd\ win)^c) = P(1st\ win) - P(1st\ win \cap 2nd\ win) = 0.6 - 0.2 = 0.4$$

- c. What is the probability that the company wins neither contract?

$$P((at\ least\ one)^c) = 1 - P(at\ least\ one) = 0.2$$

- d. What is the probability that the company wins exactly one contract?

$$P(1st\ win \cup 2nd\ win) - P(1st\ win \cap 2nd\ win) = 0.8 - 0.2 = 0.6$$

Multiple Events

- In most circumstances there is always the question, if one type of event happens and then another type of event occurs, what is the chance?
- For example, Dr Nick wants to win a game and know his probability for different scenarios. There is a 6-sided dice which he tosses, and 4 sided dice (with commands: A, B, C, D) as well.
 - The first dice presents a 6 and the second presents a B. What is the probability that this happened?
 - $P(X=6) = 1/6$ and $P(Y=B) = 1/4$, therefore for both to happen, we multiply the probabilities together.
 - $P(X=6 \text{ and } Y=B) = P(X=6) \times P(Y=B) = 1/24$.
 - Still wondering why? Write down all the events that could happen and see how many times both will occur.

Example – Case Study

- The company that you are working for is expanding and they need to place a security pass system in. The boss decides that a bi-generated username system is the best and most personal approach. Your boss assigns you with the task of determining how many username ids can be generated based off a 5 character username where the first three are letters and the last 2 are numbers. E.g.

N F Y 0 7

The idea is the same as multiplying. If write out the options for each character –

A	A	A	0	0
B	B	B	0	0
Z	Z	Z	9	9

We see that there are 26 possible choices for the letters and 10 for the numbers.

Example cont.

This means that the total number of outcomes is

$$26 \times 26 \times 26 \times 10 \times 10 = 1757600.$$

We now could calculate different probabilities for combinations of username ids!

Example in R – Back to Beers

- ❖ If that is all the beers in America, then what is the probability (proportion) of “American Pale Ale (APA)” over all beers?
- ❖ (H) Using the `breweries.csv`, which brewery produces the largest variety in beer and what is the proportion (probability) of picking a beer from this brewery randomly?

Solutions to R work will appear next week.

That wraps up Week 1! Next week.....

- Conditional Probability
- Independence

