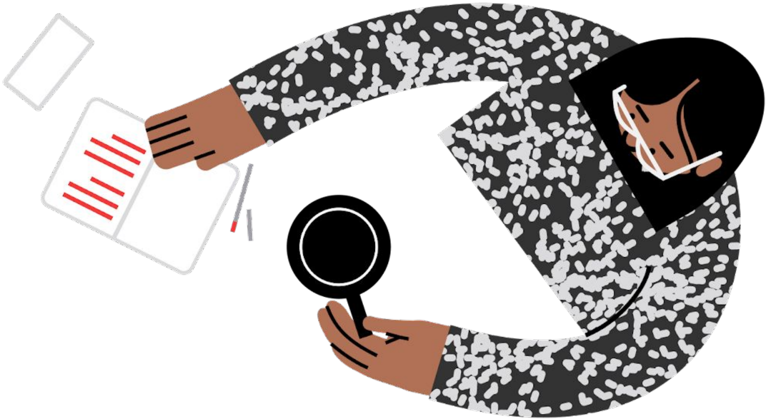


# Practical Statistics for Running Experiments

Annie Guo
Applied Scientist @Yelp

# What you will learn

In this section we'll go through a few easy rules of thumb that you can use in work & life.

- The power of small samples: **Rule of Five**
- How to read ratings: **Laplace's Rule of Succession**
- When to double anything: **Rule of 72**
- Estimate of the chance of something that hasn't happened yet: **Rule of Three**
- Change your belief based on new data: **Bayes Theorem**

# Case 1:
# The Power of Small Samples

# Time to File Reports

Imagine that you work for a company with 10,000 employees, and you are tasked to find out **how much time they spend filing expense reports per day.**

How would you do that?

# Time to File Reports

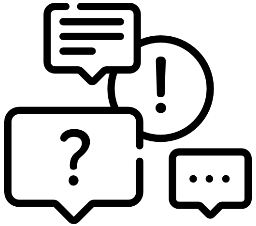
Census

# Time to File Reports

Now we will pick 5 random employees and ask them what their median report filing time is.

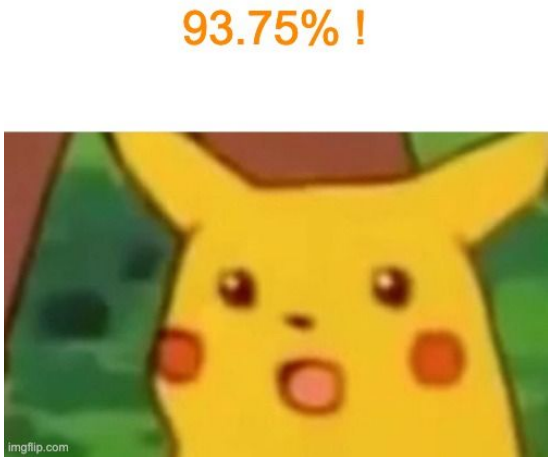
Let's imagine the results we get are 10, 20, 45, 15, 60 minutes.

*Can you use this sample of only 5 to estimate the median of every employee's filing time?*

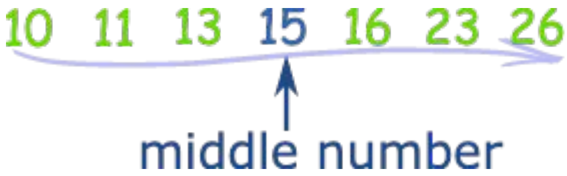


# Rule of Five

There is a 93.75% chance that the median of a population is between the smallest and largest values in any random sample of five from that population.



# Rule of Five – The Proof




- The probability of randomly picking a value above the **median** = 50%.
- The probability of randomly picking 5 values which are all ABOVE the median = (50%) ^ 5 = 3.125%
- Similarly, the probability of randomly picking 5 values which are all BELOW the median = (50%) ^ 5 = 3.125%

Probability of the median value being between the highest and lowest of 5 randomly picked values

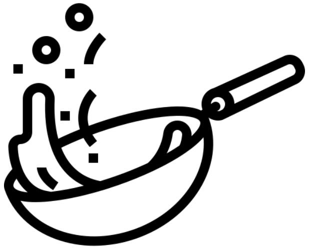
= 1 - 3.125% * 2

= 93.75%

# Time to Cook Dinner

We’ll pick 5 random attendees in the workshop for the following question:

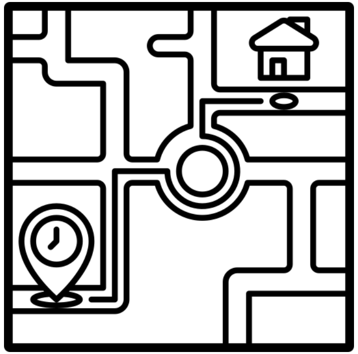
*How long do you spend cooking dinner every day?*



# Time to Commute

We’ll pick 5 random attendees in the workshop for the following question:

*Prior to Covid, how long do you spend on commuting every day (one-way)?*

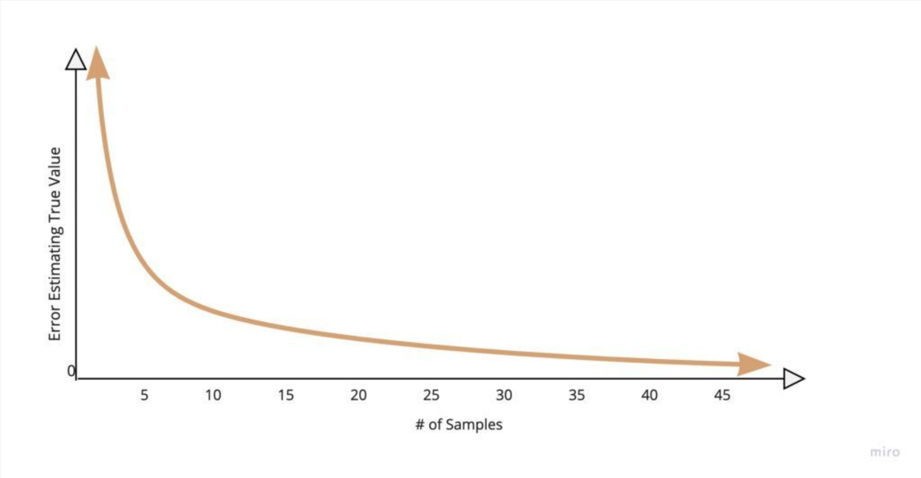


# Rule of Five

Rule of Five can be applied to samples from **any distribution**, and it can be extended if you have more than 5 data points.

Depending on your sample size, you can take nth smallest and largest values in the sample and get a even tighter bound for the estimates.

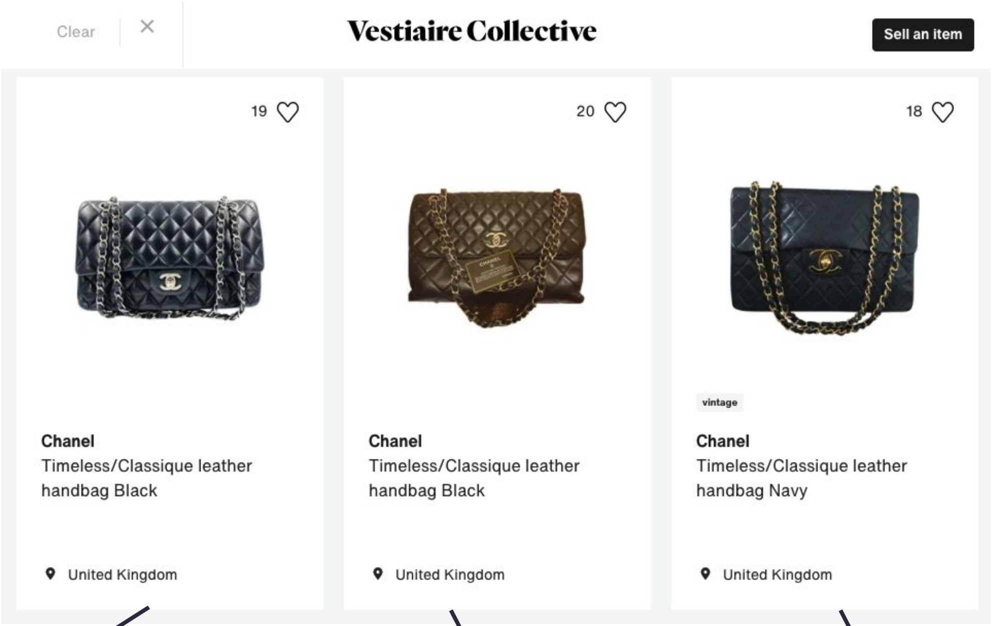
# Takeaway




You need far less data than you think.

# Case 2:
# How to read ratings

# Buying a Handbag




## Which one should you buy from?

100% positive rating,
10 reviews

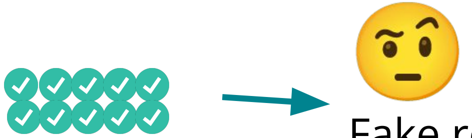
96% positive rating,
50 reviews

93% positive rating,
200 reviews

# Our Instinct

The more data we see, the more confidence we have in a given rating.

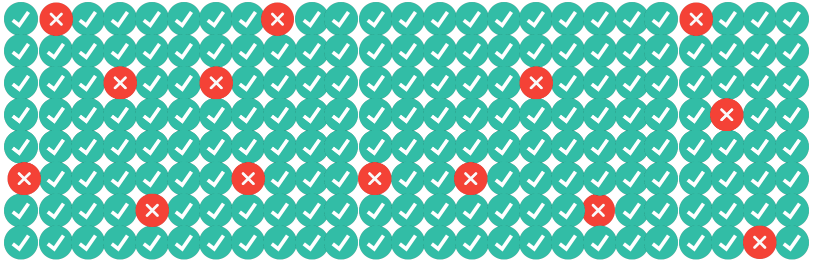
100%
10 reviews



Fake reviews?
Got lucky?

96%
50 reviews

93%
200 reviews



# How to read a rating – a simple solution

$$\frac{10}{10} \;\text{->}\; \frac{11}{12} \simeq 91.7\%$$

$$\frac{a+1}{b+2}\text{where a = number of positive ratings and b = number of total ratings}$$

# Read Ratings again

100%
10 reviews

$$\frac{11}{12} \simeq 91.7\%$$

96%
50 reviews

$$\frac{49}{52} \simeq 94.2\%$$

Based on this rule, your best bet is on the second seller with 50 reviews.

93%
200 reviews

$$\frac{187}{202} \simeq 92.5\%$$

# Laplace’s Rule of Succession

Suppose that every time there is an opportunity for an event to happen, then it occurs with unknown probability $p$. Laplace's law of succession states that, if before we observed any events we thought all values of $p$ were equally likely, then after observing $k$ events out of $n$ opportunities, a good estimate of $p$ is $\hat{p} = (k+1)/(n+2)$

$$P(X_{n+1} = 1 \mid S_n = k) = \frac{k+1}{n+2}.$$

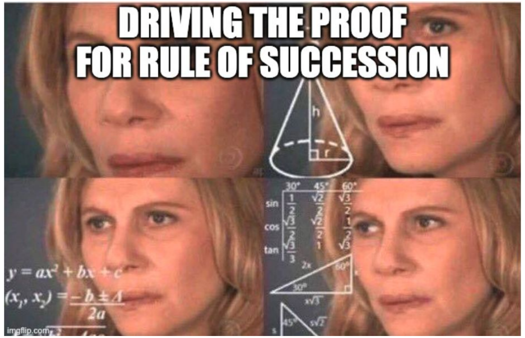
# Laplace's Rule of Succession – the Proof

The proof for the Laplace's Rule of Succession involves 4 statistical methods:

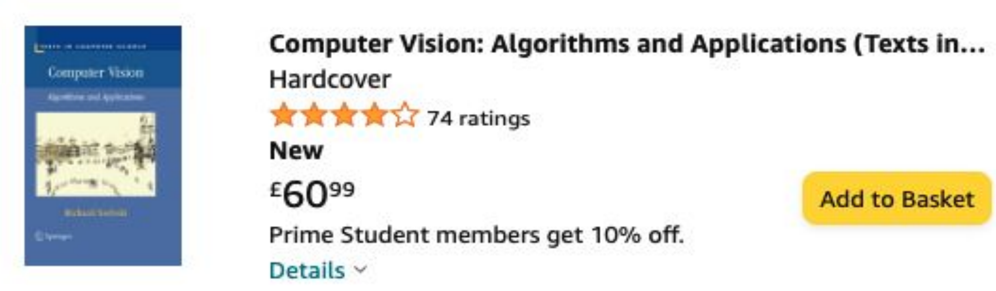
1. *Law of Total Probability*
2. *Bayes's Theorem*
3. *Binomial Probability Distribution*
4. *Beta Distribution*

Reference is in the Appendix if you are interested in deriving this formula by hand.

$$
\begin{aligned}
P(X_{n+1} = 1 \mid S_n = k) &= \int_0^1 p \frac{(n+1)!}{k!(n-k)!} p^k (1-p)^{n-k} dp \\
&= \frac{(n+1)!}{k!(n-k)!} \int_0^1 p^{k+1} (1-p)^{n-k} dp \\
&= \frac{(n+1)!}{k!(n-k)!} B(k+1, n-k) \\
&= \frac{(n+1)!}{k!(n-k)!} \frac{(k+1)!(n-k)!}{(k+1+n-k+1)!} \\
&= \frac{k+1}{n+2}.
\end{aligned}
$$



# Buying Books




*Group Exercise: which seller should you buy the book from?*

- Seller 1: 100% positive reviews, 14 reviews
- Seller 2: 96% positive reviews, 500 reviews
- Seller 3: 95% positive rating, 2000 reviews

# Case 3:
# How to estimate growth

# Estimating Growth

Your company is launching a new product and you are its dedicated product manager.

The product has been in beta mode for 3 months and currently have 100 users. Based on historical data, it has achieved an average user growth rate of 12% per month:

| | **Month 0** | **Month 1** | **Month 2** |
|---|---|---|---|
| **Number of Users** | 100 | 112 | 125 |

The Director of Product asked you during standup:

- *How long will it take for the number of users to double?*

# Rule of 72

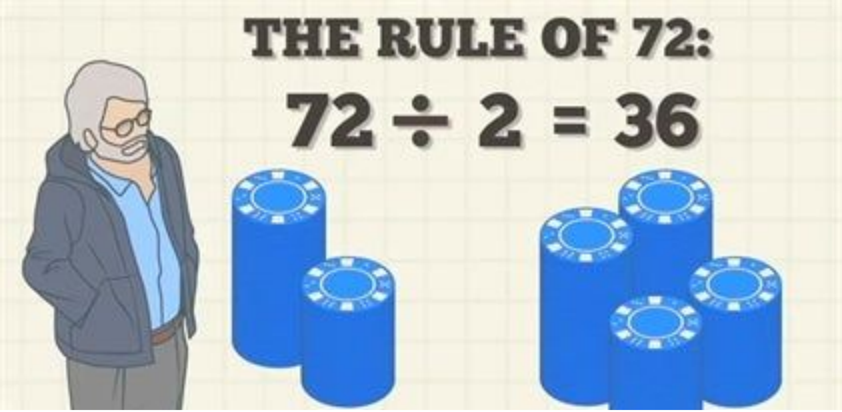
Widely used in finance, Rule of 72 is an easy way to estimate the doubling time of an object.

For example, if you save £10 in a savings account and the interest rate is 2%, then it would take 36 years (72/2 = 36) to grow to £20.

$$t \approx \frac{72}{r}$$

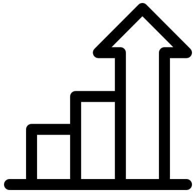
$t$ = number of periods required to double an investment's value

$r$ = interest rate per period, as a percentage



# Rule of 72 – New Product Example

- *How long will it take for the number of users to double?*
  - Average User Growth rate per month = 12%
  - Time to double number of users = $\frac{72}{12}$ = 6
    - This means that based on current growth, you can expect to double your user base every 6 months.



# Rule of 72

- Rule of 72 is applied on exponential growth and can be used in a variety of way for both work and life decisions:
  - The projected growth of a new product
  - When will the company's revenue double
  - How long you need to save vs. invest to get enough deposit for your dream home

- There are several variants of this rule:

  - Rule of 72
  - Rule of 70
  - Rule of 69.3

$$t = \frac{\ln(2)}{\ln(1 + r/100)} \approx \frac{72}{r},$$

  - The choice of the numerator is mostly a matter of preference: 69.3 is the most accurate, but 72 works well in most day-to-day lives and is more easily divisible.

# Case 4:
# Estimate the chance of something that hasn’t happened yet

# Selling into a new area

Suppose you are a solution engineer and you are pitching your company's product to clients in a new country, Wakanda.

If you've pitched to 20 clients and found 7 who are interested, you may estimate that the chance of getting a deal in Wakanda is 7/20 (35%).

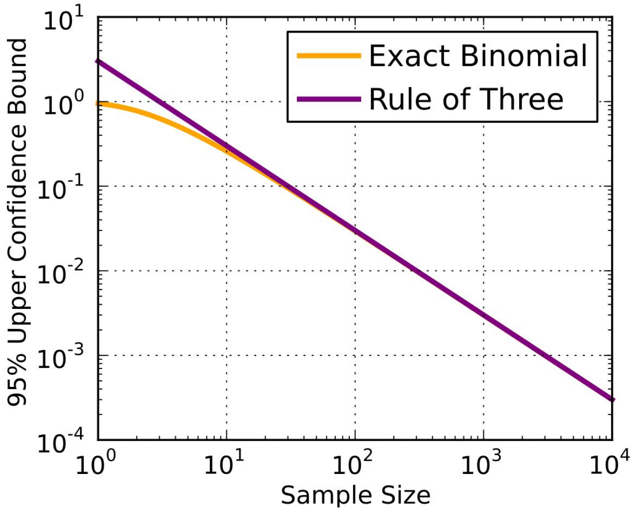
But what if you've pitched to 20 clients and none of them is interested? Are you willing to say that the chance of getting a deal is 0/20, i.e. there is absolutely no deals in Wakanda?

# Rule of Three

Rule of Three gives a quick and easy way to estimate this type of probabilities.

If you've tested *N* cases and still haven't encountered what you were looking for, then the probability of it happening is less than *3/N*.

*P(deal | having done 20 pitches) = 3/ 20 = 15%*




**Assumptions:**

- Underlying data is binary, i.e. yes/no for whether the client is interested in your product
- The probability of a successful event, i.e. a client being interested in your product, is fixed

**Usage:**

- Great for cases when the target events are rare
  - E.g. quality assurance, frauds, anomalies

# Case 5:
# Cause & Effect

# Conditional Probability Formula

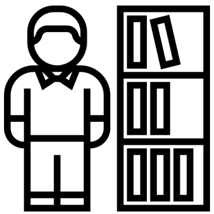
$$\underset{\text{Probability of } A \text{ given } B}{P(A \mid B)} = \frac{\overset{\text{Probability of } A \text{ and } B}{P(A \cap B)}}{\underset{\text{Probability of } B}{P(B)}}$$

No need to memorise - just want to introduce this notation here

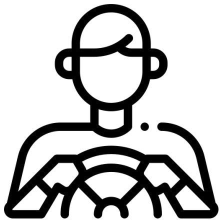
# Who is Steve?

Steve is very shy and withdrawn, invariably helpful but with very little interest in people or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail.

Is Steve more likely to be a librarian or a lorry driver?

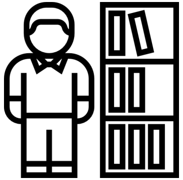


VS

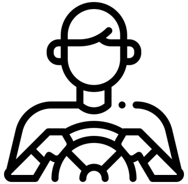


# Our Instinct

*“Steve is more likely to be a librarian because librarians are more likely to be shy and meek. ”*



$P\ (\textit{Steve fits the description} \mid \textit{Steve is a librarian})$ = 40%



$P\ (\textit{Steve fits the description} \mid \textit{Steve is a lorry driver})$ = 10%

However, this answer is incorrect because the problem we need to solve is:

$P\ (\textit{Steve is a lorry driver or librarian} \mid \textit{Steve fits the description})$

In other words, we need to ***flip*** the two sides of the conditional probability!

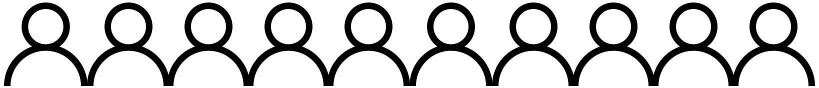
# Who is Steve – with a bit more data

When answering the question, most of us forget to take into account additional information that we have around the question.

There are way more lorry drivers than librarians in the UK: for every librarian there are 10 lorry drivers.

Will this change your conclusion?

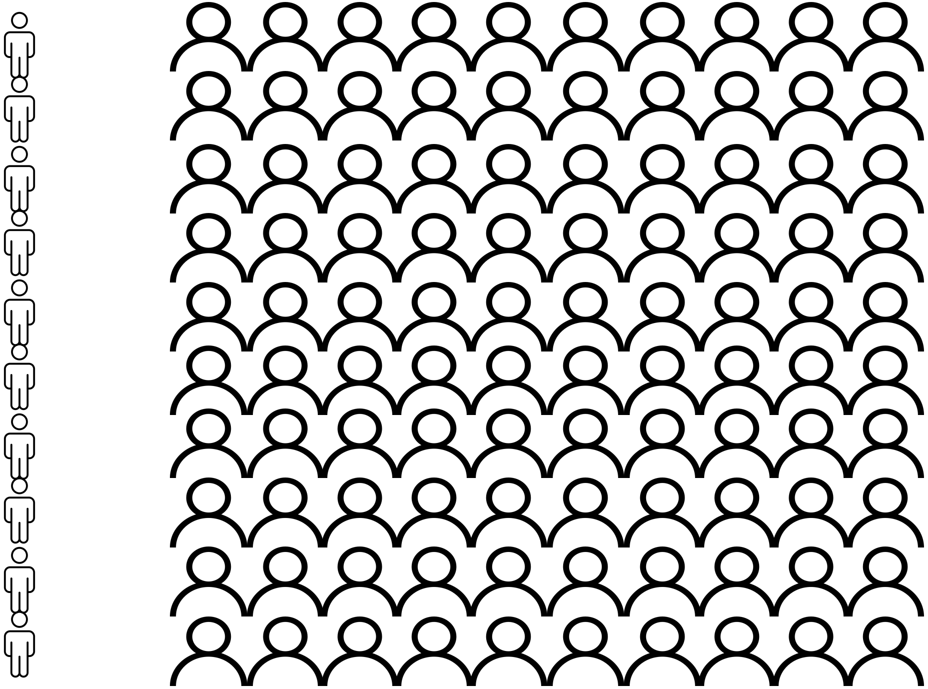
# Who is Steve – visualised



Librarian

Lorry driver

# Who is Steve – visualised



Multiple both groups by 10:

- 10 librarians
- 100 lorry drivers

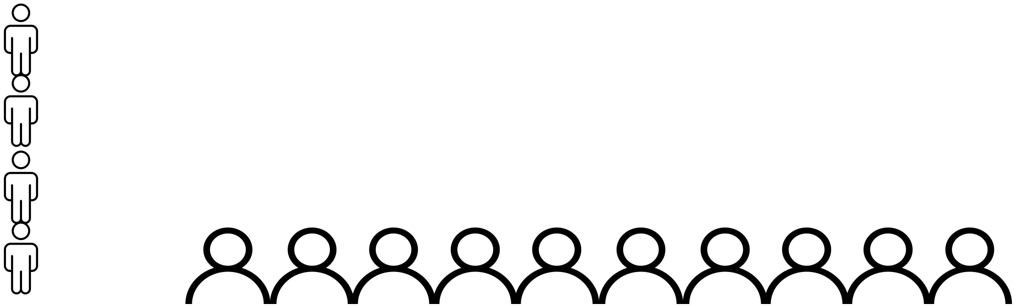
$P\ (fits\ the\ description|librarian)$ = 40%

$P\ (fits\ the\ description|lorry\ driver)$ = 10%

# Who is Steve – visualised

Assuming:

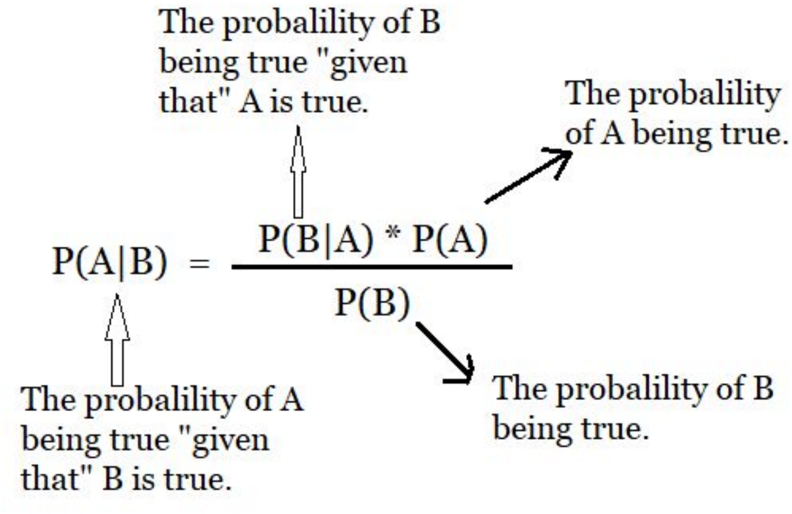
- Number of shy librarians = 10 * 40% = 4
- Number of shy lorry drivers = 100 * 10% = 10

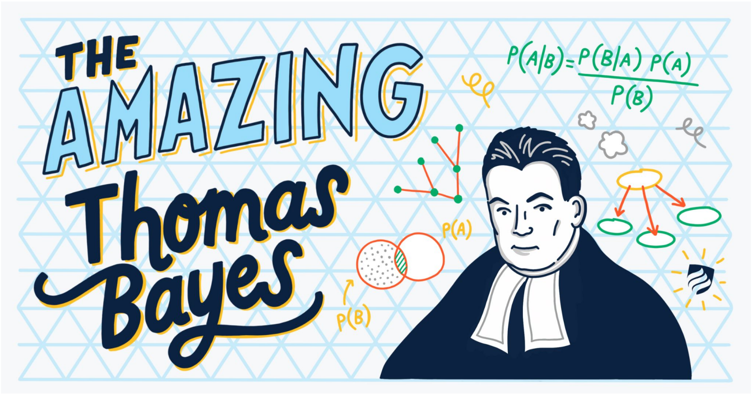


*P (Librarian | fits the description)* = 4 / (4 + 10) ≈ **28%**
*P (Lorry Driver | fits the description)* = 10/ (4 + 10) ≈ **71%**

*It is much more likely for Steve to be a lorry driver than a librarian!*

# Bayes' Theorem




Bayes' Theorem allows you to "flip" the two sides of the conditional probability and update your current belief once new data comes in.

# Who's Steve – using Bayes' Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

*P (Librarian | fits the description) = P (fits the description | librarian) * P (librarian) / P(fits the description) = 40% * (10/110) / (14/ 110) ≈* **28%**

*P (Lorry driver | fits the description) = P (fits the description | lorry driver) * P (lorry driver) / P(fits the description) = 10% * (100/110) / (14/ 110) ≈* **71%**

# Group Exercise: Internet Doctor

I discovered a new mole on my arm and I did some search on the internet.

Now I think that I might have skin cancer 🤨

Assuming the following:

- P( mole | skin cancer ) = 70%
- P (growing a new mole) = 10%
- P (skin cancer for under 50s) = 0.02%

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

$A, B$ = events

$P(A|B)$ = probability of A given B is true

$P(B|A)$ = probability of B given A is true

$P(A), P(B)$ = the independent probabilities of A and B

*What is the probability that I have skin cancer, given that I have a new mole?*

*P (skin cancer | mole)*

# Questions?