

An Introduction To Bayesian Statistics



Simon Thornewill von Essen

Data Analyst, Goodgame Studios

@sthornewillve



“[Bayesian Stats.] is one of those ideas that seems hard when you first encounter it. Then, at some point there is a breakthrough and then it seems obvious.

Once you’ve got it, it’s such a beautiful idea that it changes how you see everything.”

- Prof. Allan Downey,
Data Framed, 2018

Data
Framed

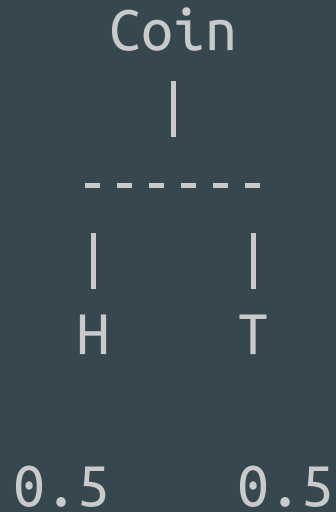
BY  DataCamp



How do we estimate probability?

- **Classical:** By considering equal outcomes
- **Frequentist:** Relative Frequency over time
- **Bayesian:** By quantifying our uncertainties

Coin Toss: Classical Est.



Dice: Classical Est.



Classical Stats

- Requirements
 - All Outcomes are known
 - Outcomes are assumed to be equally likely
- Advantages
 - Fast Estimation
 - Easy to understand
- Disadvantages
 - Outcomes must be known
 - Often created overly simplified models when applied to complex phenomena

How do we estimate probability?

- ~~Classical~~

- Frequentist



- Take measurements over time
- Measurements will eventually approximate the parameter we want to measure

- Bayesian

Frequentist Est.

- Check to see if thermometer is properly calibrated

Frequentist Approach:

Take many readings and use the expectation value (mean) and std for sample



Calculate the probability of your data given your data following some parameter.



Frequentist Stats

- Requirements
 - Possibility to perform experiments indefinitely
 - Parameters are assumed to be fixed
 - Able to estimate params given enough experiments
- Advantages
 - Works well with simulations
 - “Objective”
- Disadvantages
 - Requires large sample size to be meaningful
 - Does not allow for integration of domain knowledge
 - P-values and confidence intervals are unintuitive
 - Difficult to communicate to non-statisticians

Thermometer Calibration: Test

- 1. If P-value = 0.001 (highly significant), is the probability of getting this result or a more extreme one given our data 0.001? 
- 2. For a given confidence interval, does the parameter lie within it 95% of the time? 

Thermometer Calibration: Test

- 1. If P-value = 0.001 (highly significant), is the probability of getting this result or a more extreme one given our data 0.001?

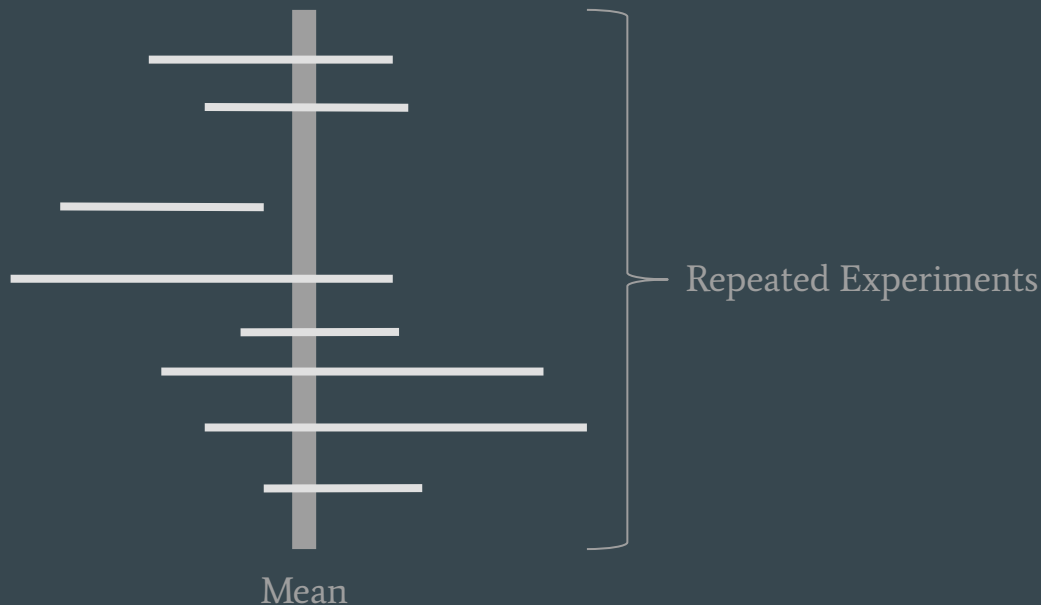
Probability of getting this result or a more extreme one given a null hypothesis (i.e. no difference in experimental groups) is 0.001

- 2. For a given confidence interval, does the parameter lie within it 95% of the time?

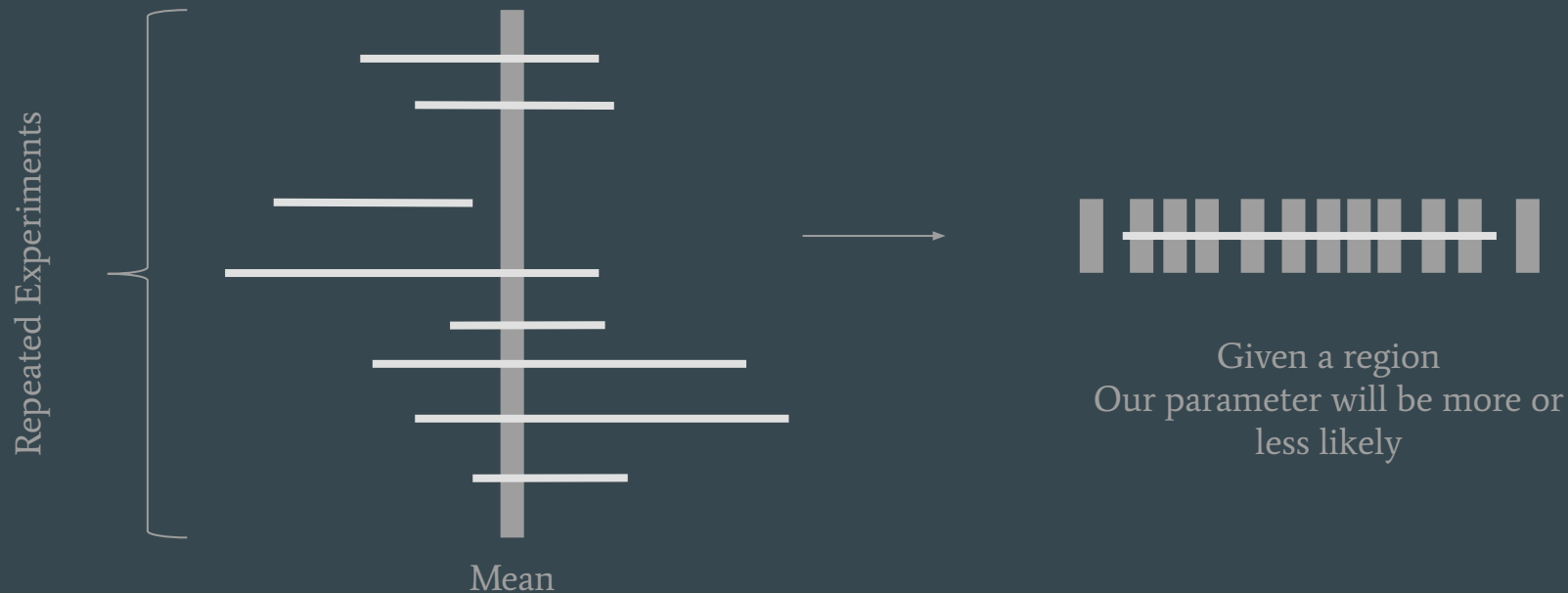
Intervals of repeated experiments will contain the parameter 95% of the time

Thermometer Calibration: Test Learnings

i Frequentism expects that parameters exist and are fixed, the probabilities are the likelihood of our data given these expectations



i “Wait, wasn’t this was we were doing with frequentism?”



Thermometer Calibration: Test Learnings



Child doesn't move, your repeated photos contain them 95% of the time

Frequentist Stats Disav. Cont.

What if?

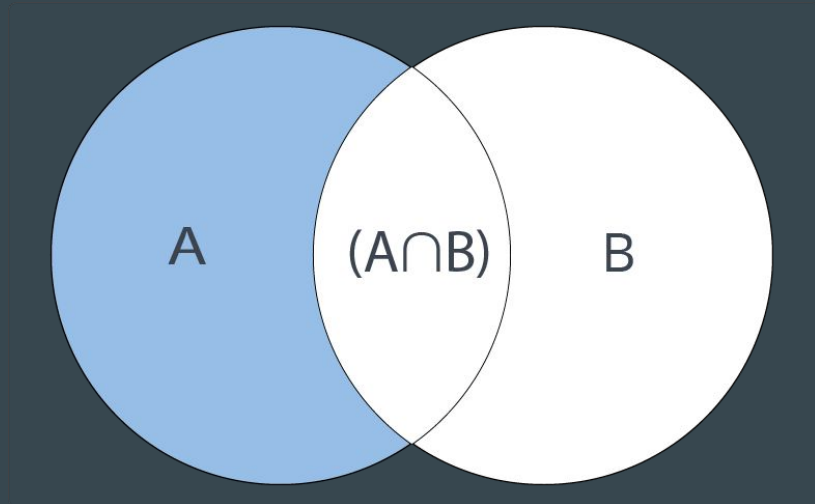
- Amount of data you have is limited? ✓
- You have relevant and applicable prior information ✓
- “Infinite” experiments are not possible? (Cost, feasibility) ✓
- Stakeholders have a hard time understanding frequentist logic? ✓
- Children never stay still and assuming they do is blasphemy ✓

How do we estimate probability?

- ~~Classical~~
- ~~Frequentist~~
- Bayesian

Bayes Theorem: Derivation

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



Bayes Theorem: Derivation

The Same

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

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



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

Bayes Theorem: Alternate View

θ = Parameter,

X = Data

$$\overset{\text{posterior}}{p(\theta|X)} = \frac{\overset{\text{likelihood}}{p(X|\theta)}\overset{\text{prior}}{p(\theta)}}{\underset{\text{normalisation}}{p(X)}}$$

Bayes Theorem: Alternate View

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)}$$

The equation shows the posterior probability $p(\theta|X)$ as the product of the likelihood $p(X|\theta)$ and the prior $p(\theta)$, divided by the marginal likelihood $p(X)$. A checkmark is placed above the likelihood term, and question marks are placed above the prior and marginal likelihood terms to indicate they are the focus of the problem.

? Problem:

- How to calculate $p(X)$
- How to calculate $p(\theta)$

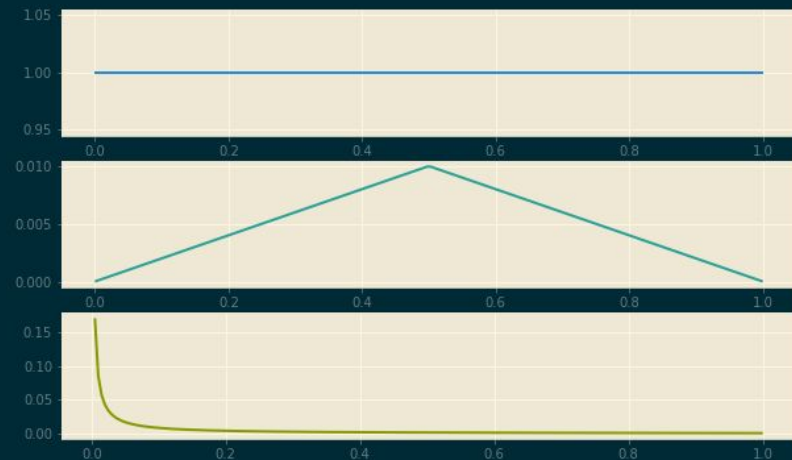
Bayes Theorem: How to Calculate $P(\theta)$?

1. Create Your Own
2. Take Previous posterior

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)^{?}}{p(X)}$$

Bayes Theorem: How to Calculate $P(\theta)$?

1. Create Your Own
2. Take Previous posterior



Bayes Theorem: How to Calculate $P(\theta)$?

Problem: Are you Baking your biases into your model?



Bayes Theorem: How to Calculate $P(\theta)$?

Might as well have your explicit and tangible biases.

As the sample size increases, priors get washed out. (As long as you are “reasonable”)

- Low Sample Size: Frequentist Stats is börked anyway, so why not?
- High Sample Size: Prior Doesn't matter
- “Reasonable” = Cromwell's Rule

Bayes Theorem: How to Calculate P(X)?

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)} \text{ ?}$$

1. Sum over all possible hypotheses
2. This is the hard part
3. Can be ignored if comparing inside of the same distribution

$$p(X) = \sum_{i=0}^n p(X|\theta_i)p(\theta_i)$$
$$p(X) = \int p(X|\theta)p(\theta)d\theta$$

How!?: Doctor's Diagnosis

- Doctor's Diagnosis:
 - You are suspicious you have a rare disease
 - Disease affects 0.1% of the population
 - Test is 99% accurate
 - What is the probability you have this disease given that you tested positive?

How!?: Doctor's Diagnosis

$$p(d \mid pos.) = \frac{p(pos. \mid d)p(d)}{p(pos.)}$$

$$p(d \mid pos.) = \frac{p(pos. \mid d)p(d)}{p(pos. \mid d)p(d) + p(pos. \mid \neg d)p(\neg d)}$$

$$p(d \mid pos.) = \frac{0.99 \times 0.001}{0.99 \times 0.001 + 0.01 \times 0.999}$$

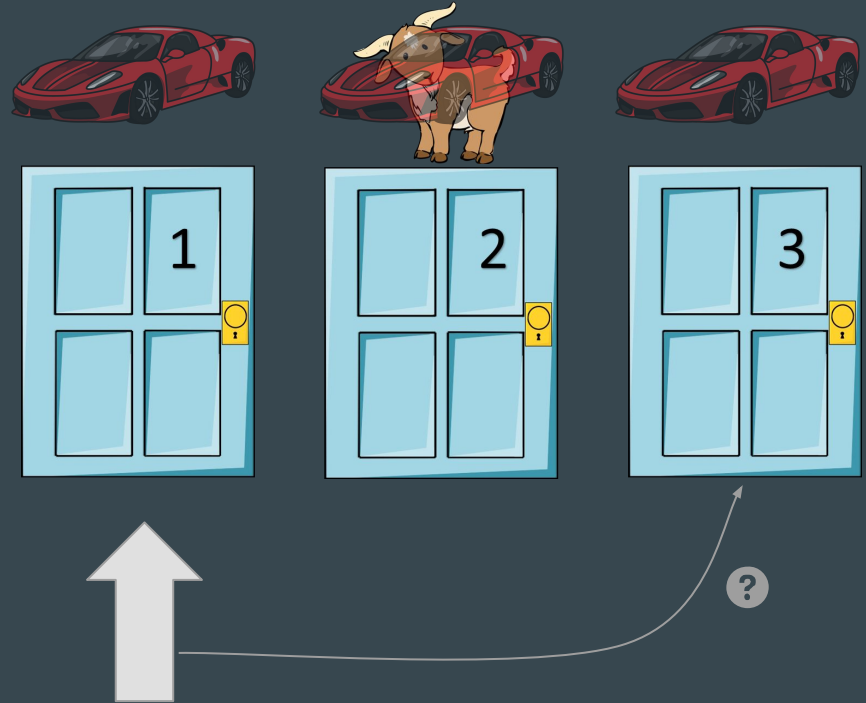
$(9.9 \cdot 10^{-4})$ $(9.99 \cdot 10^{-3})$

How!?: Doctor's Diagnosis

- Doctor's Diagnosis:
 - Works out to be roughly 9% chance of having disease
 - Rate is lower than expected because disease is so rare
 - 99% is a high accuracy, but not *that* high
 - Application: Think about this example next time you build a classifier

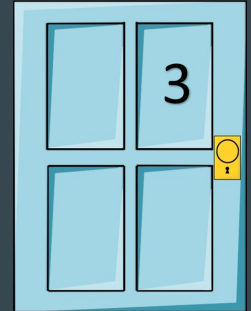
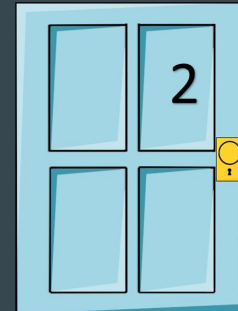
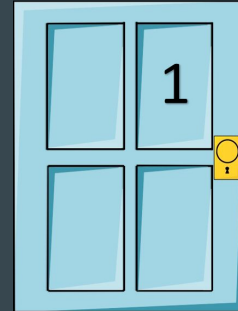
How!?: Monty Hall Problem

- The Monty Hall Problem:
 - You Pick Door 1
 - Monty opens door 2 to reveal a goat
 - Should you switch to door 3?



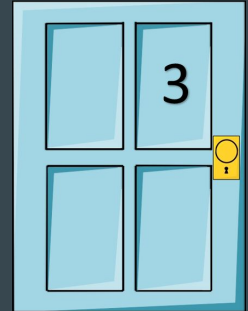
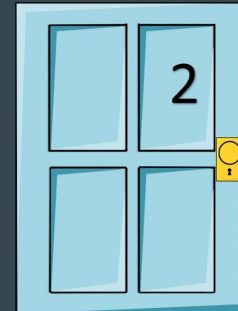
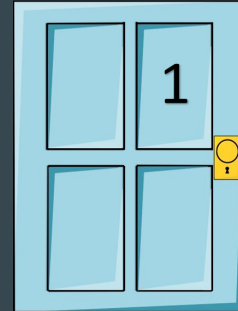
How!?: Monty Hall Problem

Hypothesis i	Prior $P(\theta_i)$
Car Behind 1	$1/3$
Car Behind 2	$1/3$
Car Behind 3	$1/3$



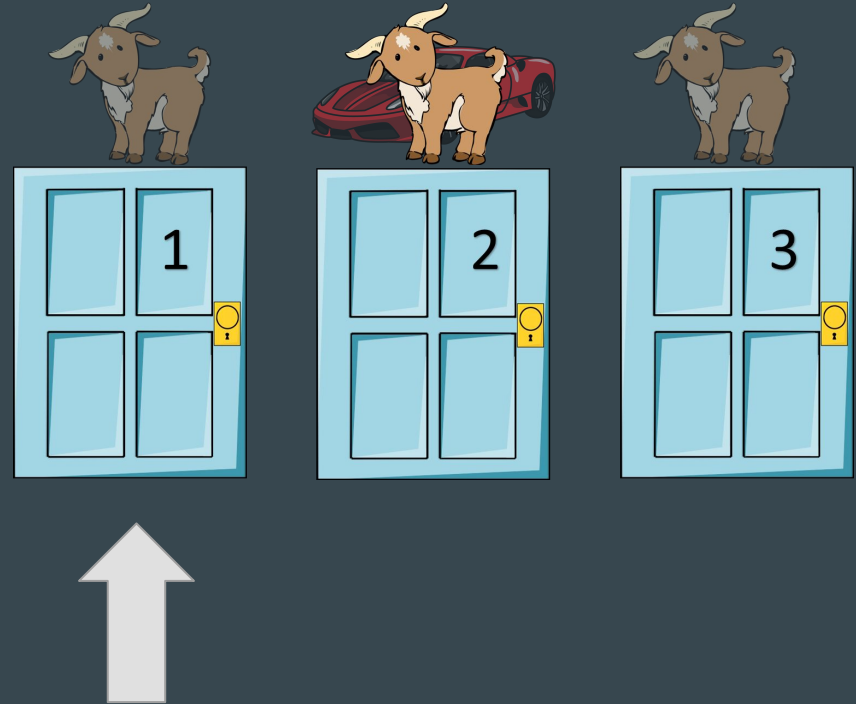
How!?: Monty Hall Problem

Hypothesis i	Likelihood $p(X \theta_i)$
Car Behind 1	
Car Behind 2	
Car Behind 3	



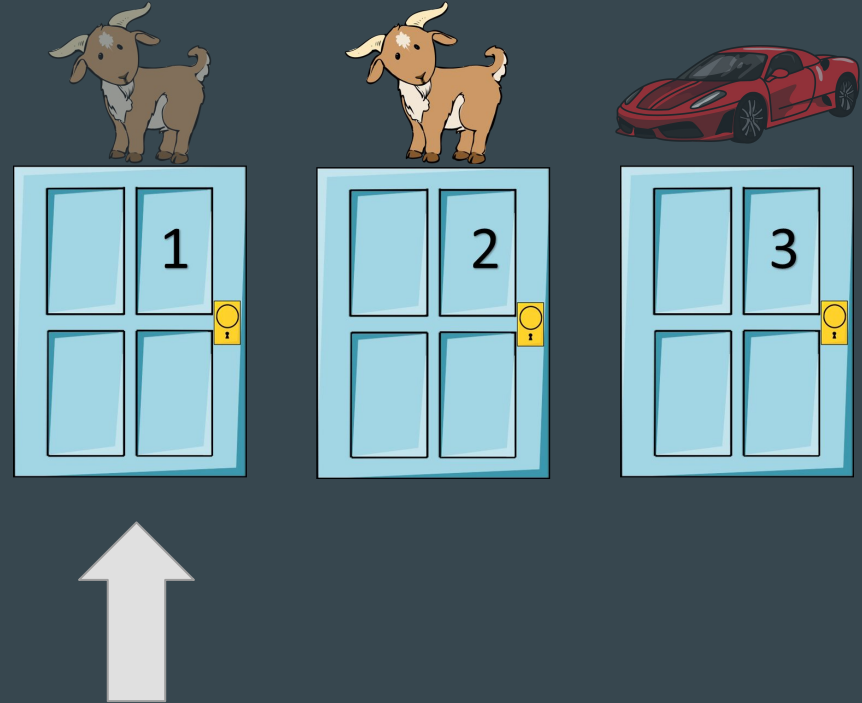
How!?: Monty Hall Problem

Hypothesis i	Likelihood $p(X \theta_i)$
Car Behind 1	$1/2$
Car Behind 2	
Car Behind 3	



How!?: Monty Hall Problem

Hypothesis i	Likelihood $p(X \theta_i)$
Car Behind 1	$1/2$
Car Behind 2	0.0
Car Behind 3	



How!?: Monty Hall Problem

Hypothesis i	Prior $p(\theta_i)$	Likelihood $p(X \theta_i)$
Car Behind 1	1/3	1/2
Car Behind 2	1/3	0.0
Car Behind 3	1/3	1.0

How!?: Monty Hall Problem

Hypothesis i	Prior $p(\theta_i)$	Likelihood $p(X \theta_i)$	Prior * Likelihood
Car Behind 1	1/3	1/2	1/6
Car Behind 2	1/3	0.0	0.0
Car Behind 3	1/3	1.0	2/6

$$\begin{aligned} P(X) &= \sum_{i=0}^n p(X|\theta_i)p(\theta_i) = 1/6 + 0 + 2/6 \\ &= 3/6 \\ &= 1/2 \end{aligned}$$

* This is the Dot Product of $p(\theta_i)$ and $p(X | \theta_i)$

How!?: Monty Hall Problem

Hypothesis i	Prior $p(\theta_i)$	Likelihood $p(X \theta_i)$	Prior * Likelihood	Posterior
Car Behind 1	1/3	1/2	1/6	1/3
Car Behind 2	1/3	0.0	0.0	0.0
Car Behind 3	1/3	1.0	2/6	2/3

Key to problem:

Monty does not choose doors at random and so opening a door provides you with information

How!?: Train Analysis

- You see a train labeled 60 (labels in ascending order of creation)
- How many trains does this company own given that you've seen #60?



How!?: Train Analysis

Hypothesis i	Prior $p(\theta_i)$	Likelihood $p(X \theta_i)$	Prior * Likelihood	Posterior
1 Train	1/1000	0.0	0.0	post_1
2 Trains	1/1000	0.0	0.0	post_2
...
60 Trains	1/1000	1/60	$1/(6 \cdot 10^4)$	post_{60}
...
1000 Trains	1/1000	1/1000	$1/10^6$	post_{1000}

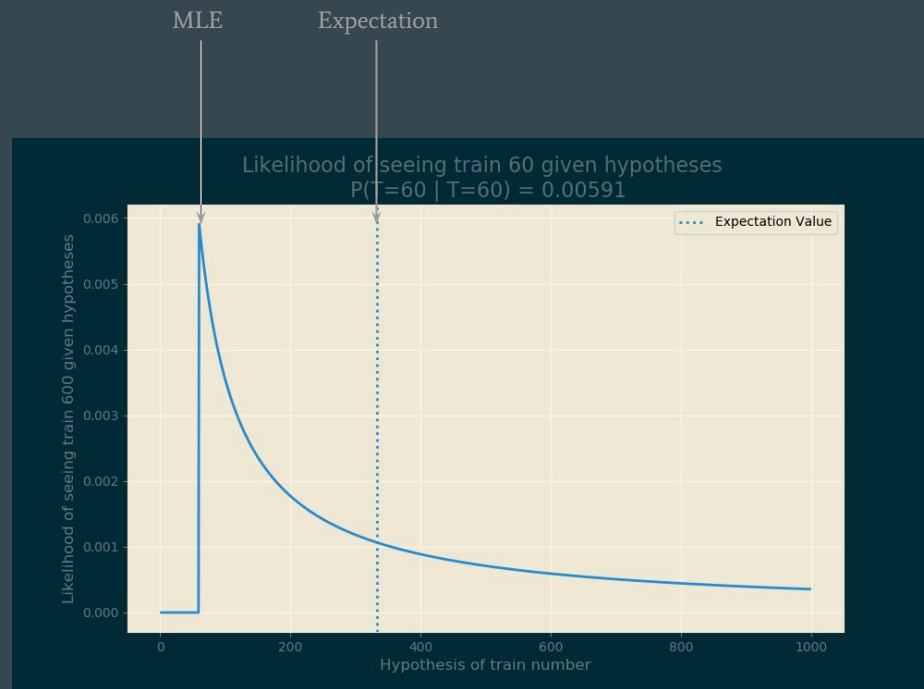
$$\Sigma = P(\text{Train})$$

How!?: Train Analysis

```
1  n = 1000
2
3  hypo = [1/n for i in range(n)]
4
5  likelihood = [1/i for i in range(n)]
6
7  post_not_norm = np.array(hypo) * np.array(likelihood)
8
9  normalisation = post_not_norm.sum()
10
11 posterior = post_not_norm / normalisation
12
```

How!?: Train Analysis

- What if we change priors?
 - Posterior changes
- What if we increase the max number of trains
 - Posterior changes



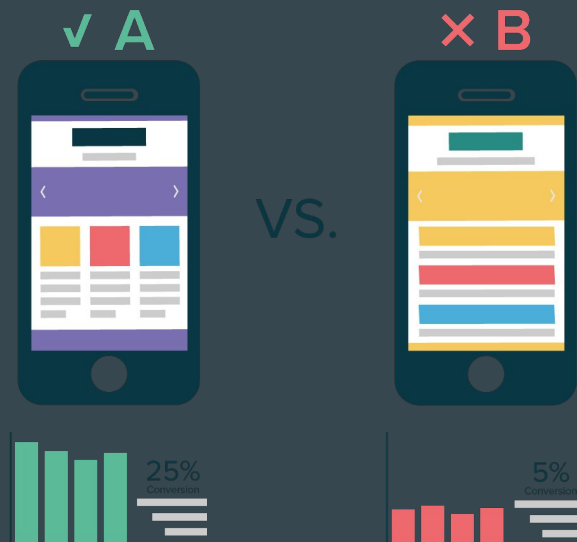
How!?: Part 1 - Discrete Case Recap

- Remember the table calculation!
- Steps:
 - Pick Prior (Often Uniform)
 - Multiply by Frequentist Likelihood
 - Divide by Normalisation constant

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{\sum_{i=0}^n p(X|\theta_i)p(\theta_i)}$$

How!?: Part 2 - Continuous Case

1. AB Testing Revisited:
 - a. Two variants
 - b. What is the probability of A being better than B?
2. Time for Bayesian Statistics!



How!?: Part 2 - Continuous Case

Challenges w/ Freq. AB Tests:

- Test needs to reach pre-defined sample size
- Need to adjust α for multiple tests (Bonferroni Corrections)
- People start accepting null hypotheses
- Can only reject/fail to reject null hypothesis, (leads to p-hacking)
- People peek at tests before tests are over, (moar p-hacking)

How!?: Part 2 - Continuous Case

AB Test:

- For people randomly placed in control/test
- Track conversions (1/0)
- What is our Likelihood?
 - Bernoulli

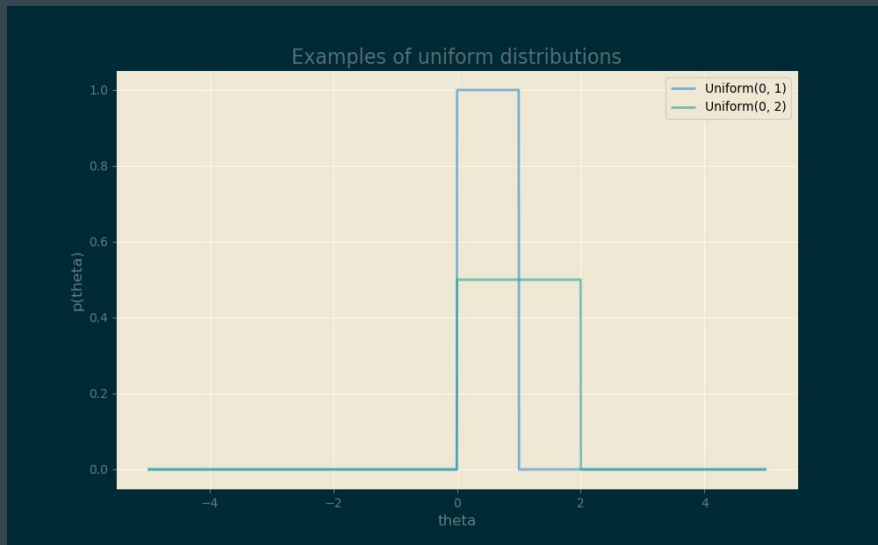
$$P(X|\theta) = \theta^{\sum y_i} (1 - \theta)^{n - \sum y_i}$$

How!?: Part 2 - Continuous Case

Prior?:

- Uninformed Prior
- Uniform distribution
- Represented by
Indicator Function

$$P(\theta) = I_{\{0 \leq \theta \leq 1\}}$$



How!?: Part 2 - Continuous Case

$$P(\theta|X) \propto [\theta^{\sum y_i} (1-\theta)^{n-\sum y_i}] [I_{\{0 \leq \theta \leq 1\}}]$$

How!?: Part 2 - Continuous Case

$$P(\theta|X) = \frac{\theta^{\sum y_i} (1 - \theta)^{n - \sum y_i} I_{\{0 \leq \theta \leq 1\}}}{\int_0^1 \theta^{\sum y_i} (1 - \theta)^{n - \sum y_i} I_{\{0 \leq \theta \leq 1\}} d\theta}$$

How!?: Part 2 - Continuous Case

$$P(\theta|X) = \frac{\theta^{\sum y_i} (1 - \theta)^{n - \sum y_i} I_{\{0 \leq \theta \leq 1\}}}{\boxed{A^{-1}} \int_0^1 \boxed{A} \theta^{\sum y_i} (1 - \theta)^{n - \sum y_i} I_{\{0 \leq \theta \leq 1\}} d\theta}$$

$$A = \frac{\Gamma(\sum n + 2)}{\Gamma(\sum y_i + 1) \Gamma(\sum n - y_i + 1)}$$

How!?: Part 2 - Continuous Case

$$\begin{aligned} P(\theta|X) &= A(\theta^{\sum y_i} (1 - \theta)^{n - \sum y_i} I_{\{0 \leq \theta \leq 1\}}) \\ &= \textit{Beta}(\alpha, \beta) \end{aligned}$$

How!?: Part 2 - Continuous Case

$$P(\theta|X) = \text{Beta}(\alpha, \beta)$$

$$\alpha = 1 + \sum y_i,$$

$$\beta = n - 1 + \sum y_i$$



Conjugate Priors

- Beta distribution is example of conj. Prior
- Use it and you will get the same distribution in posterior
- Once the math is done, never do it again
- Update functions using data as it appears

$$P(\theta|X) = \textit{Beta}(\alpha, \beta)$$

Conjugate Priors

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters	Interpretation of hyperparameters ^[note 1]	Posterior predictive ^[note 2]
Bernoulli	p (probability)	Beta	α, β	$\alpha + \sum_{i=1}^n x_i, \beta + n - \sum_{i=1}^n x_i$	$\alpha - 1$ successes, $\beta - 1$ failures ^[note 1]	$p(\tilde{x} = 1) = \frac{\alpha'}{\alpha' + \beta'}$

https://en.wikipedia.org/wiki/Conjugate_prior#Table_of_conjugate_distributions
 (Just Google “conjugate priors table wikipedia”)

Conjugate Priors

Poisson	λ (rate)	Gamma	k, θ	$k + \sum_{i=1}^n x_i, \frac{\theta}{n\theta + 1}$	k total occurrences in $\frac{1}{\theta}$ intervals	$\text{NB}(\tilde{x} \mid k', \theta')$ (negative binomial)
			α, β [note 3]	$\alpha + \sum_{i=1}^n x_i, \beta + n$	α total occurrences in β intervals	$\text{NB}\left(\tilde{x} \mid \alpha', \frac{1}{1 + \beta'}\right)$ (negative binomial)
Exponential	λ (rate)	Gamma	α, β [note 3]	$\alpha + n, \beta + \sum_{i=1}^n x_i$	α observations that sum to β [6]	$\text{Lomax}(\tilde{x} \mid \beta', \alpha')$ (Lomax distribution)

https://en.wikipedia.org/wiki/Conjugate_prior#Table_of_conjugate_distributions
(Just Google “conjugate priors table wikipedia”)

How!?: Part 2 - Continuous Case Recap

- Steps:
 - Pick Prior (Often Uniform)
 - Multiply by Frequentist Likelihood
 - Divide by Normalisation constant
 - Integral over all possible hypotheses
 - (Tips and tricks may be required)
 - If you're lucky, you can use a conjugate prior for updates

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{\int p(X|\theta)p(\theta)d\theta}$$

Demo: Conjugate Priors

<https://seeing-theory.brown.edu/bayesian-inference/index.html>

Posteriors - What Now?

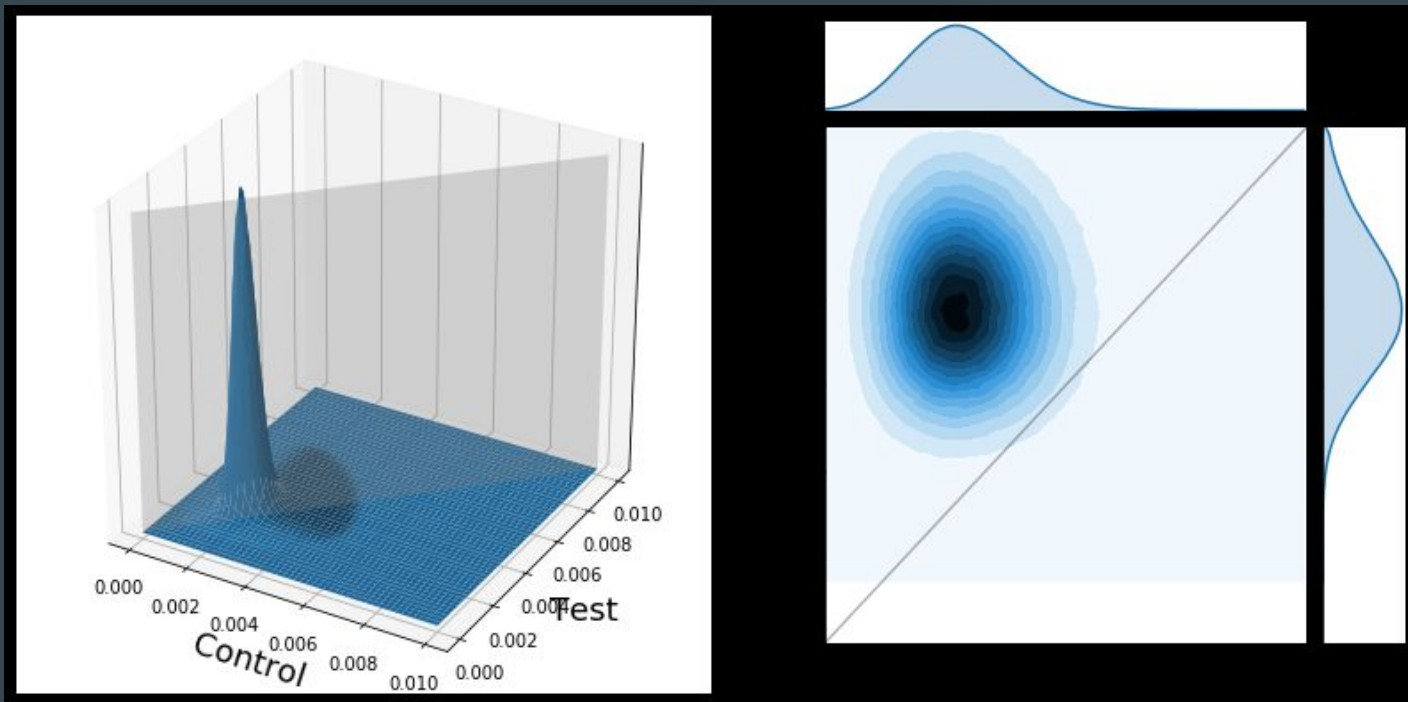
Do we calculate P-values now?

- No need, just calculate $P(\theta_2 > \theta_1)$
- Takes some calculation, but the result is nicer

Can we peek at results or stop at any time?

- Yes!

Posteriors - What Now?



Demo: Bayesian AB Testing

Bayesian Stats

- Advantages

- a. Estimation in the case of little data (specific circumstances)
- b. Parameters are distributions
- c. Probabilistic answers answer questions people tend to have

- Disadvantages

- a. Steeper learning curve
- b. Integrals are hard (MCMC computationally expensive)
- c. Criticisms of being less “objective” due to use of priors
- d. Point estimates become the same as frequentist estimations with high sample sizes

How do we estimate probability?

- **Classical:** By considering equal outcomes
- **Frequentist:** Relative Frequency over time
- **Bayesian:** By quantifying our uncertainties

How do we estimate probability?

- ~~Classical: By considering equal outcomes~~
- ~~Frequentist: Relative Frequency over time~~
- ~~Bayesian: By quantifying our uncertainties~~
- There is only one field of statistics

Conclusion and “Call to Action”

Understanding Bayes vs Freq. is key to understanding a lot of the field stats

- Beginner scientists: Check my sources as a jump off point
- Experienced scientists: Come and tell me how you applied this knowledge
- Decision-makers: Be aware of these kinds of analyses and when the strengths benefit your use-case.
- General Advice: Try to think about conditional probability more often (doctor's diagnosis) in relation to drawing conclusions

Resources for further learning

- Mathematical Understanding:
“Bayesian Stats: From Concept to Data Analysis”,
U of Santa Cruz
- Intuition between Bayesianism & Frequentism:
“Frequentism and Bayesianism”,
Scipy - Jake Vander Plas
- Examples of Real World Applications:
“Think Bayes”,
Allan Downey
- Further reading into MCMC and pyMC:
“Bayesian Methods for Hackers”,
Cameron Davidson-Pilon

Find Slides on Github

<https://cutt.ly/zGqux9>

A screenshot of the GitHub profile page for Simon Thornewill von Essen. The profile includes a profile picture, a bio, location (Hamburg, Germany), email (simon@thornewill.org), and a link to his LinkedIn profile. The page also shows a list of pinned repositories, including 'Udacity-DataScience-Nanodegree', 'Pet-Project---Bodybuilding-WFPB-Diet', 'Pet-Project---Tygem-Fuseki-Web-Scraper-using-Python', 'Udacity-DataAnalyst-Nanodegree', 'Hamburg_DS_Meetup_Bayesian-Stats_Intro', and 'Udemy_LazyProgrammer_Courses'. At the bottom, there is a contribution graph showing 213 contributions in the last year.

Simon Thornewill von Essen
SThornewillE

[Edit profile](#)

J. Data Analyst
Goodgame Studios
Hamburg, Germany
simon@thornewill.org
<https://www.linkedin.com/in/sthornew...>

Overview Repositories (11) Projects (0) Stars (1) Followers (32) Following (7)

Pinned [Customize your pins](#)

- Udacity-DataScience-Nanodegree**
Udacity Data Science Nanodegree Repository. Contains lecture notes, and dummy scripts as well as projects undertaken for the nanodegree.
HTML ★ 8 11
- Udacity-DataAnalyst-Nanodegree**
Contains lecture notes, and dummy scripts as well as projects undertaken for the nanodegree.
HTML ★ 1
- Pet-Project---Bodybuilding-WFPB-Diet**
Use Linear Programming to minimise diet costs while maximising nutritional value
Jupyter Notebook
- Hamburg_DS_Meetup_Bayesian-Stats_Intro**
Slides, notes, and code for Bayesian Stats. talk
Python
- Pet-Project---Tygem-Fuseki-Web-Scraper-using-Python**
Pet Project 01: Tygem Fuseki Web Scraper using Python
Python ★ 2
- Udemy_LazyProgrammer_Courses**
Exercise attempts for Lazy Programmer courses on Udemy
HTML ★ 1

213 contributions in the last year [Contribution settings](#)

Learn how we count contributions.

Less More

Fin!



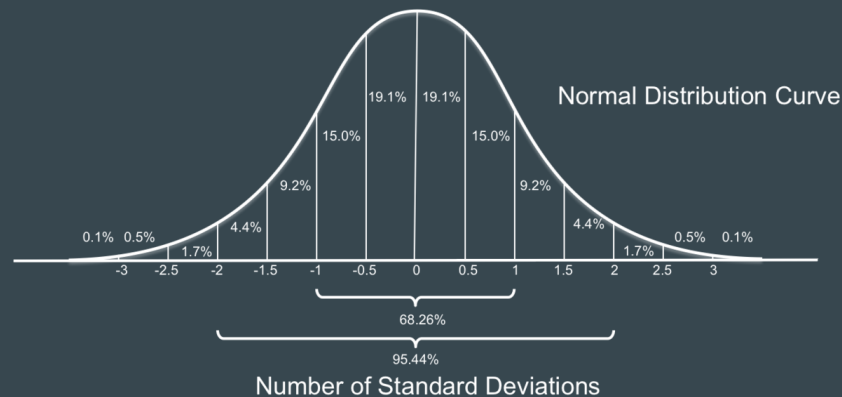
@sthorneville



Frequentist Est.: Confidence Intervals

Confidence Interval:

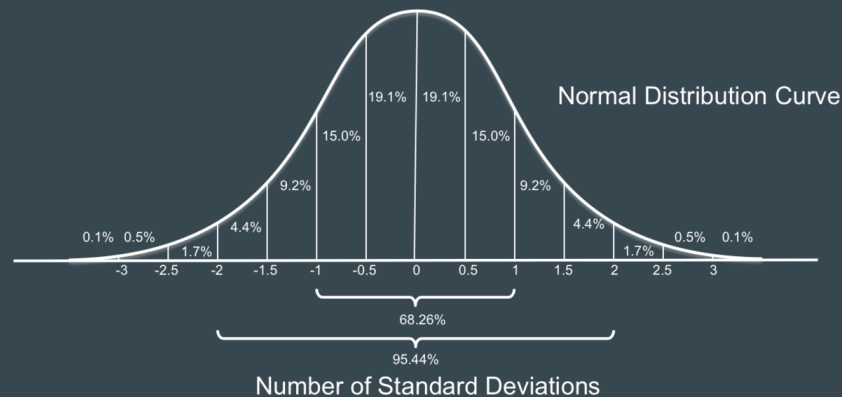
- From sample mean and standard deviation, calculate an interval
- Upon repeated experiments, intervals contain the true parameter x% of the time



Frequentist Est.: Confidence Intervals

CI Intuition::

- bootstrap CI n times -> intervals would contain the mean of population 95% of the time



Frequentist Est.: P-Values

P-value:

- Probability of seeing data given a parameter

