An Introduction To Bayesian Statistics

•••

Simon Thornewill von Essen

Data Analyst, Goodgame Studios

`@sthornewillve` 🐍

How do we estimate probability?

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

Coin Toss: Classical Est.

```
     Coin
       |
    -------
    |     |
    H     T

 0.5     0.5
```

Dice: Classical Est.

```
            Dice
              |
 -------------------------
 |    |    |    |    |    |
 1    2    3    4    5    6

0.16 0.16 0.16 0.16 0.16 0.16
```

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
- Bayesian

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

Thermometer Calibration: Frequentist Est.

- Check to see if thermometer is properly calibrated

Frequentist Approach:

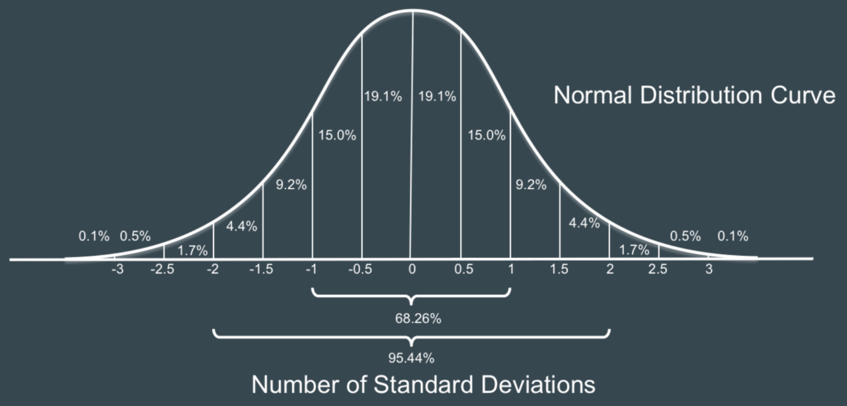
Take many readings and use the expectation value (mean) and std for sample

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

Thermometer Calibration: Frequentist Est.

Confidence Interval:

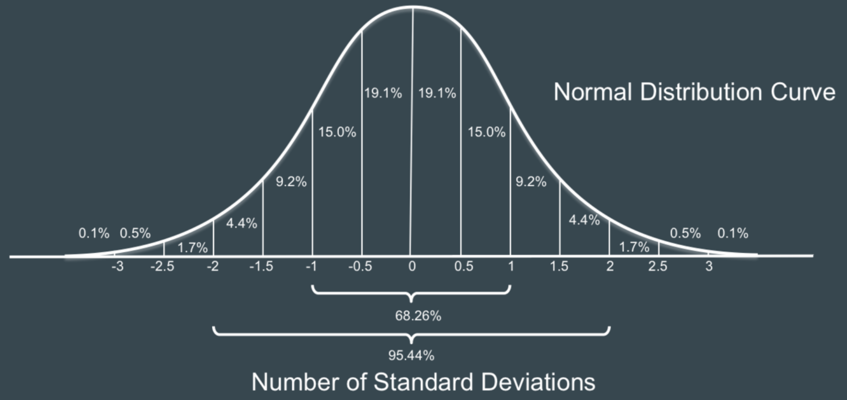
- From sample mean and standard deviation, calculate an interval
- Interval contains the true parameter x% of the time upon repeated experiments



Thermometer Calibration: Frequentist Est.

Confidence Interval:

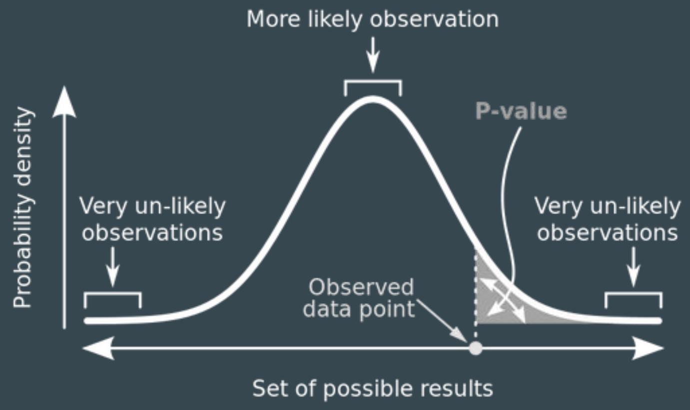
- Intuition:
  - If you were to bootstrap the confidence interval n times
  - Interval would contain the mean of population 95% of the time



Probability of Rain: Frequentist Est.

P-value:

- Probability of data given a parameter
- “The probability that outcome is due to random chance given that there is no difference between experimental groups”
- $P(X \mid \mu)$



Thermometer Calibration: Test

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

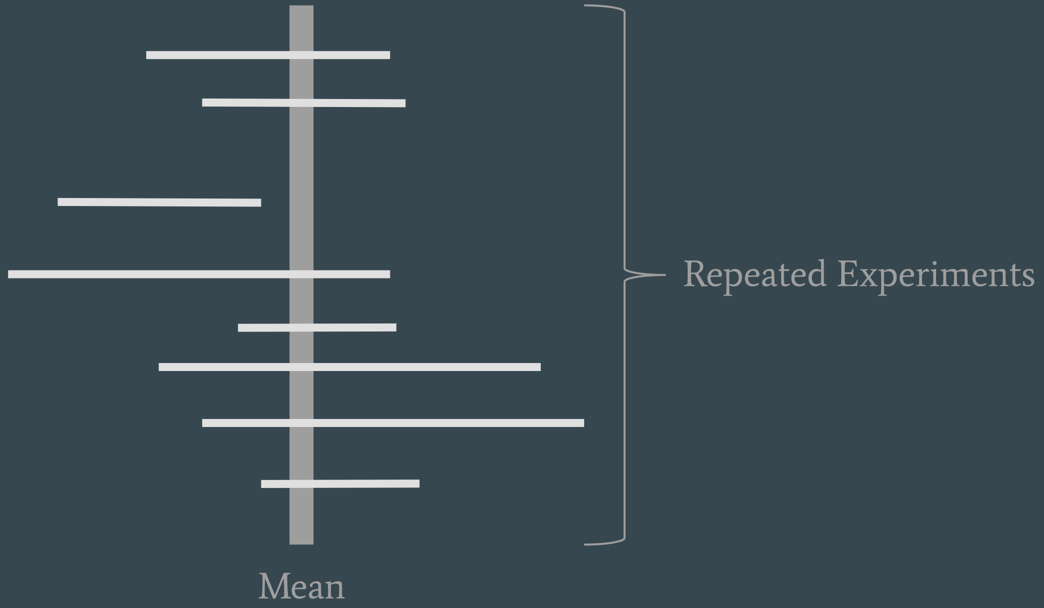
   Probability of getting this result <u>given no difference in experimental groups</u> is 0.001

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

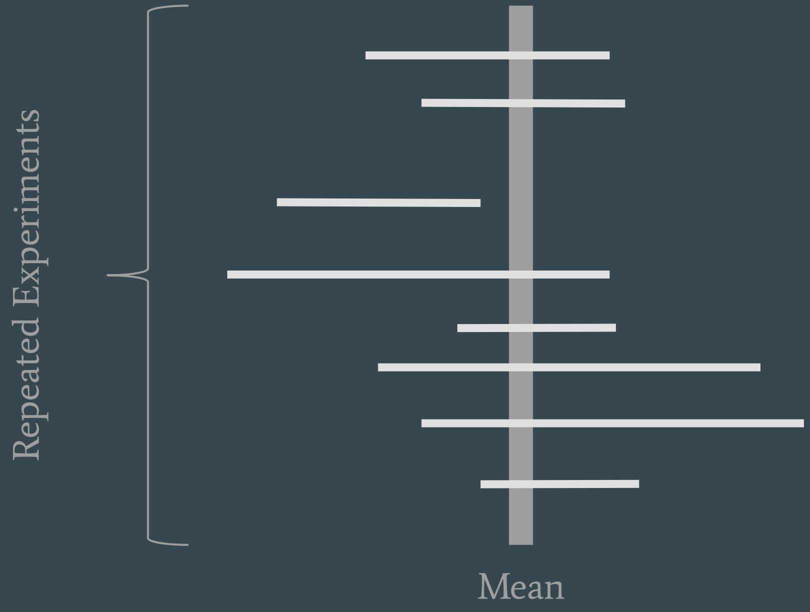
   <u>Intervals of repeated experiments</u> will contain the parameter 95% of the time

Thermometer Calibration: Test Learnings

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



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



Given a region
Our parameter will be more or less likely

Example from J. Vanderplas’s 2014 Scipy Talk

Thermometer Calibration: Test Learnings

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

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

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

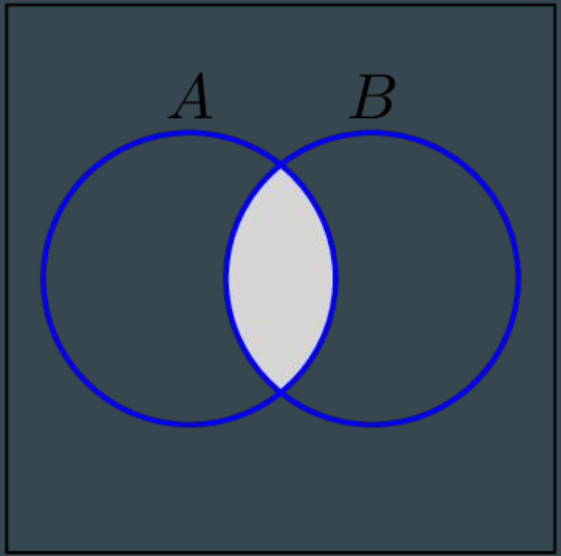
Bayes Theorem

- Goal: Invert a likelihood

$$\underbrace{p(B \mid A)}_{\text{posterior}} = \frac{\overbrace{p(A \mid B)}^{\text{likelihood}}\ \overbrace{p(B)}^{\text{prior}}}{\underbrace{p(A)}_{\text{normalisation}}}$$

Bayes Theorem: Derivation

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



Bayes Theorem: Derivation

The Same

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

$$\therefore \; p(B \mid A) = \frac{p(A \mid B)\, p(B)}{p(A)}$$

Bayes Theorem: Alternate View

θ = Parameter,

X = Data

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

Bayes Theorem: Alternate View

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

? Problem:

- How to calculate p(X)
- How to calculate p(θ)

Bayes Theorem: How to Calculate P(θ)?

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(θ)?

Problem: Are you Baking your biases into your model?

Bayes Theorem: How to Calculate P(θ)?

Might as well have your explicit and tangible biases.

As the sample size increases, priors get washed out.

- Low Sample Size: Frequentist Stats is börked anyway, so why not?
- High Sample Size: Prior Doesn't matter

Bayes Theorem: How to Calculate P(X)?

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

?

1. Sum of all possible numerators
2. Yes, this can get difficult

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

$$p(X) = \int p(X|\theta)p(\theta)d\theta$$

Bayes Theorem: How to Calculate P(X)?

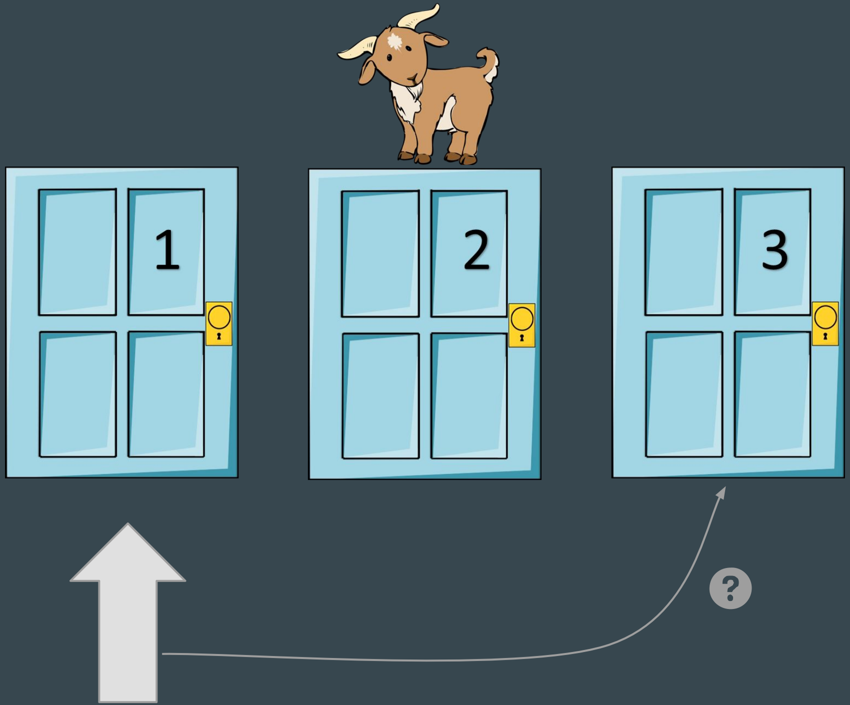
You can ignore P(X) if you are comparing posteriors for the same distributions

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

?

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	Prior $p(\theta_i)$	Likelihood $p(X \mid \theta_i)$
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 $p(X \mid \theta_i)$	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	1/3

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

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

How!?: Monty Hall Problem

Hypothesis i	Prior $p(\theta_i)$	Likelihood $p(X \mid \theta_i)$	Prior * Likelihood	Posterior
Car Behind 1	1/3	1/2	1/6	1/3
Car Behind 2	1/3	0.0	0.0	0
Car Behind 3	1/3	1.0	1/3	2/3

Key to problem:

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

"Table Method" from A. Downey's *Think Bayes*

How!?: Train Analysis

- You see a train labeled 60
- What was the probability of seeing 60 given that you saw it?

How!?: Train Analysis

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

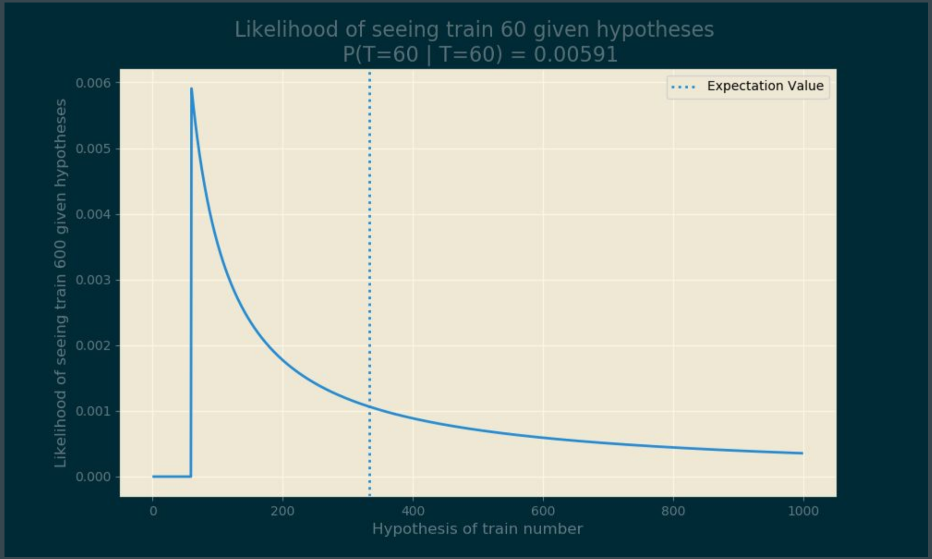
Σ =
P(Train)

How!?: Train Analysis

How!?: Train Analysis

- What if we change priors?
  - Posterior changes

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



Example from A. Downey's "Think Bayes"

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

AB Test:

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

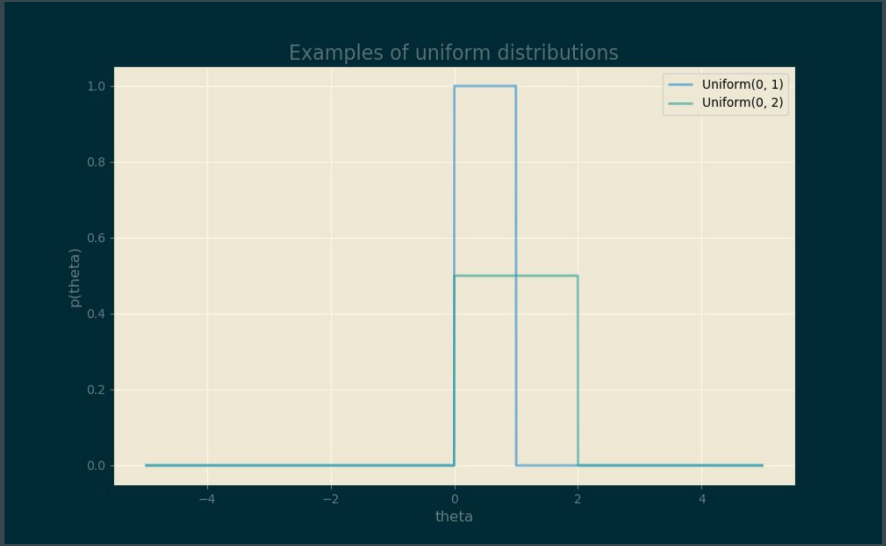
$$P(X|\theta) = \theta^{\sum y_i}(1-\theta)^{n-\sum y_i}$$

How!?: Part 2 - Continuous Case

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

<u>Prior?</u>:

- Uninformed Prior
- Uniform distribution
- Represented by Indicator Function



How!?: Part 2 - Continuous Case

$$P(\theta|X) \propto [\theta^{\sum y_i}(1-\theta)^{n-\sum y_i}][I_{\{0\leq\theta\leq1\}}]$$

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

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

How!?: Part 2 - Continuous Case

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

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

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)

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

When is it okay not to perform Normalisation?

- When you are comparing two values inside of the same set that creates p(P)

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

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

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

See more Conj. Priors in U. of Santa Cruz's Coursera Course

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	$\mathrm{NB}(\tilde{x} \mid k', \theta')$ (negative binomial)
			α, β[note 3]	$\alpha + \sum_{i=1}^{n} x_i, \ \beta + n$	α total occurrences in β intervals	$\mathrm{NB}\left(\tilde{x} \mid \alpha', \frac{1}{1+\beta'}\right)$ (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]	$\mathrm{Lomax}(\tilde{x} \mid \beta', \alpha')$ (Lomax distribution)

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

Posteriors - What Now?

- Estimation of Parameters
- Credible Intervals
- Vary priors to see effects
- Calculate P(Θ=x | X)
- etc.

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

Demo:
Conjugate Priors

Posteriors - What Now?

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)

Posteriors - What Now?

Do we calculate P-values now?

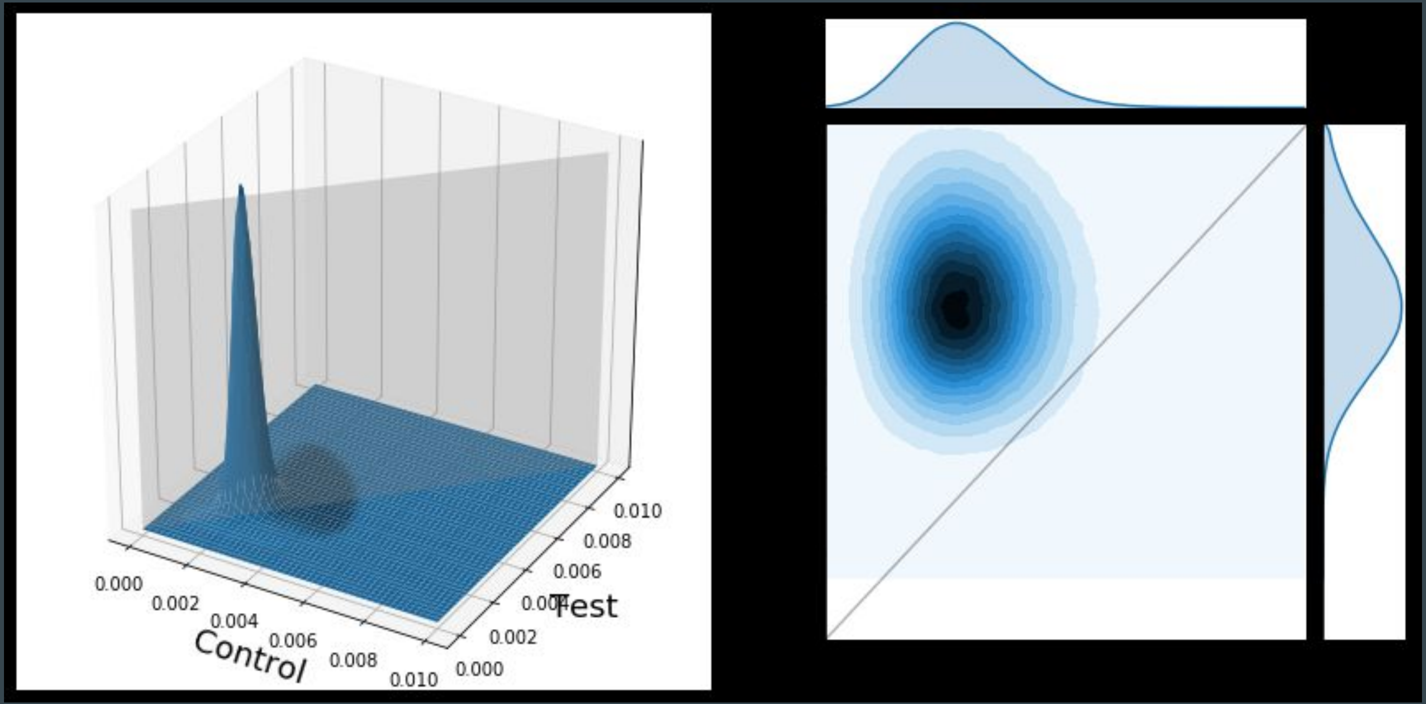
- No need, just calculate $P(\theta_2 > \theta_1)$
- Takes some calculation, but the result is nicer

Can we peek or stop at any time?

- Yes!

More from Nora Neumann @ Pydata 2016 on YouTube

Posteriors - What Now?



More from Vincenzo Lavorini @ “Bayesian A/B Testing with Python: the easy guide” Towards Data Science

Bayesian Stats

- Advantages
  - Incorporation of Domain Knowledge
  - Estimation in the case of little data (specific circumstances)
  - Allows for models of as little or high complexity as necessary
  - Parameters are distributions
  - Easier to communicate with more interpretable answers

- Disadvantages
  - Lots and lots of theory
  - Integrals are hard, MCMC methods aren't easy either
  - MCMC can be computationally expensive
  - Criticisms of being less "objective" due to use of priors
  - 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

Conclusion and "Call to Action"

- Understanding Bayes vs Freq. is key to understanding a lot of stats
- For scientists: Check my sources as a jump off point
- For decision-makers: Consider these kinds of analyses when little data is available

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:

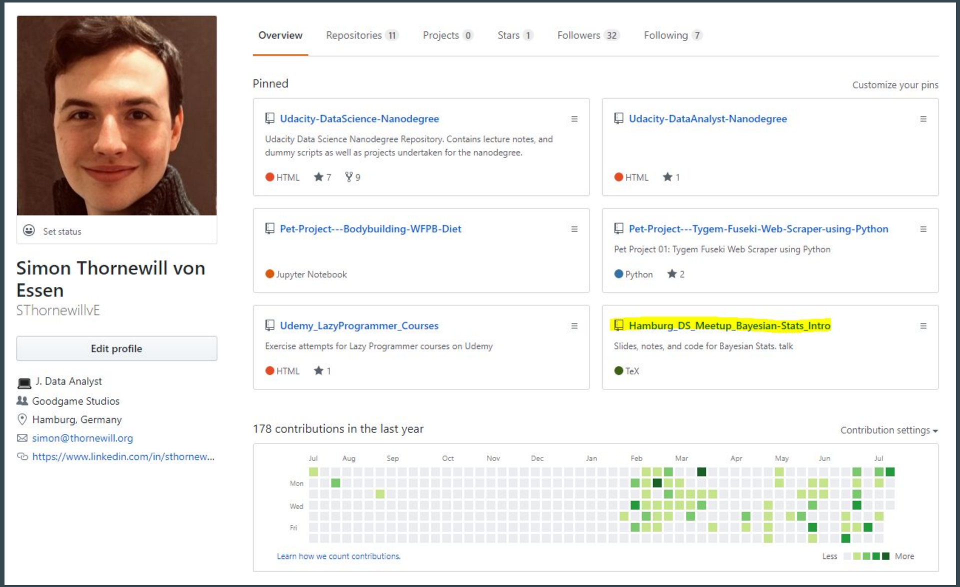
  “Think Bayes”,
  Allan Downey

- Further reading into MCMC and pyMC:

  “Bayesian Methods for Hackers”,
  Cameron Davidson-Pilon

Find Slides on Github

https://cutt.ly/zGqux9



Fin!

•••

`@sthornewillve`