An Introduction To Bayesian Statistics

Simon Thornewill von Essen
Data Analyst, Goodgame Studios

`@sthornewillve`

How do we estimate the probability?

- Classical: By considering equal outcomes
- Frequentist: Relative Frequency over time
- Bayesian: By updating our beliefs for each obs.

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
  - High Bias
  - Outcomes must be known
  - Cannot create sophisticated (high variance) models

How do we estimate the probability?

- ~~Classical~~
- Frequentist
- Bayesian

Thermometer Calibration: Frequentist Est.

- Calibrating Thermometer to show accurate values
- Follows a Normal Distribution

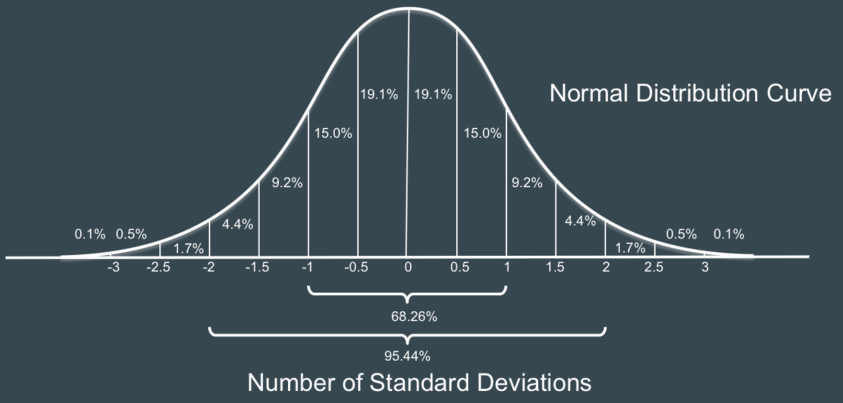
Frequentist Approach:

Take many readings and use the expectation value (mean) to find value over time.

Thermometer Calibration: Frequentist Est.

Confidence Interval:

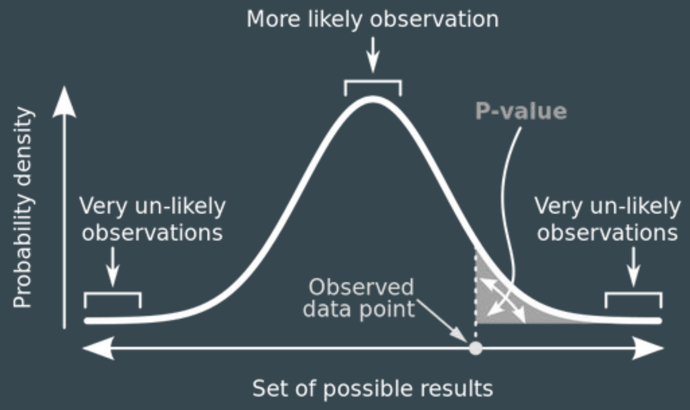
- From sample mean and standard deviation, calculate an interval
- "Interval that contains the true parameter some percent of the time"



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. Mean thermometer temp is higher than assumed param,
P-value = 0.001 (highly significant),

Does this mean that the probability of mean thermometer temp is 0.999? ❌

❓ 2. 95% Confidence interval is [98°C, 102 °C] and mean = 100°C,

Does this mean that 100°C will fall inside this interval 95% of the time? ❌

Thermometer Calibration: Test

1. Mean thermometer temp is higher than assumed param,
   P-value = 0.001 (highly significant),

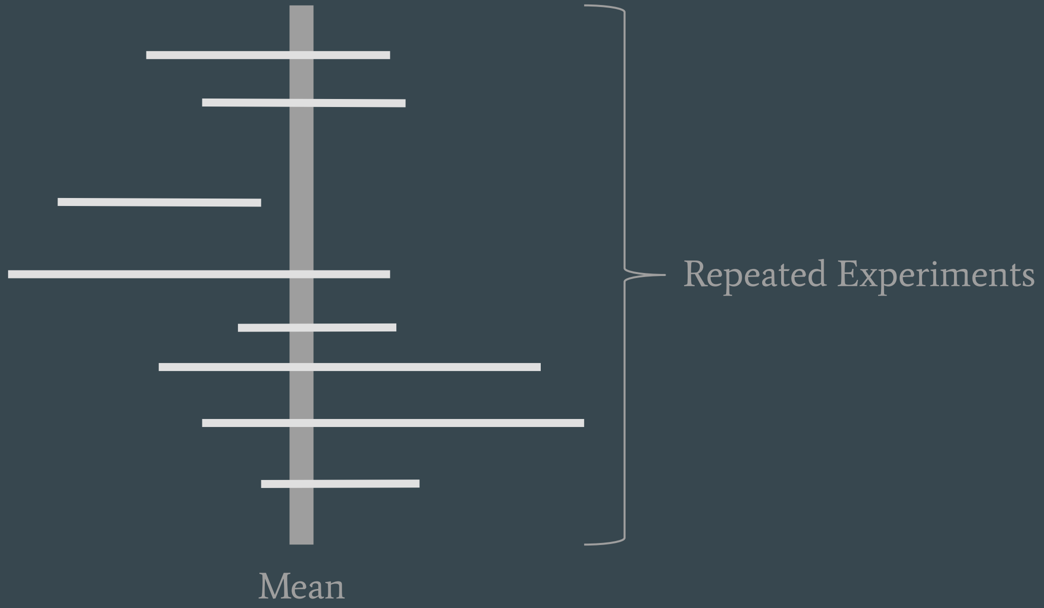
   Probability of getting this result given no difference in experimental groups is 0.001

2. 95% Confidence interval is [98°C, 102 °C] and mean = 100°C,

   Interval will contain the parameter 95% of the time

Thermometer Calibration: Test Learnings

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



Thermometer Calibration: Test Learnings

Child doesn’t move, but you will only take a picture of them 95% of the time

Frequentist Stats

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

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 don’t is blasphemy ✓

How do we estimate the probability?

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

Bayes Theorem

- Goal:

Invert a likelihood

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

Bayes Theorem: Derivation

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(θ | X): Prob. Param given Dat.
- p(B): Prior
- p(A | B): Freq. Likelihood
- p(A): Normalisation Const.

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

Problem:

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

Bayes Theorem: How to Calculate P(X)?

1. What is p(X)?
2. Sum of all possible numerators
3. Yes, this can get difficult

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

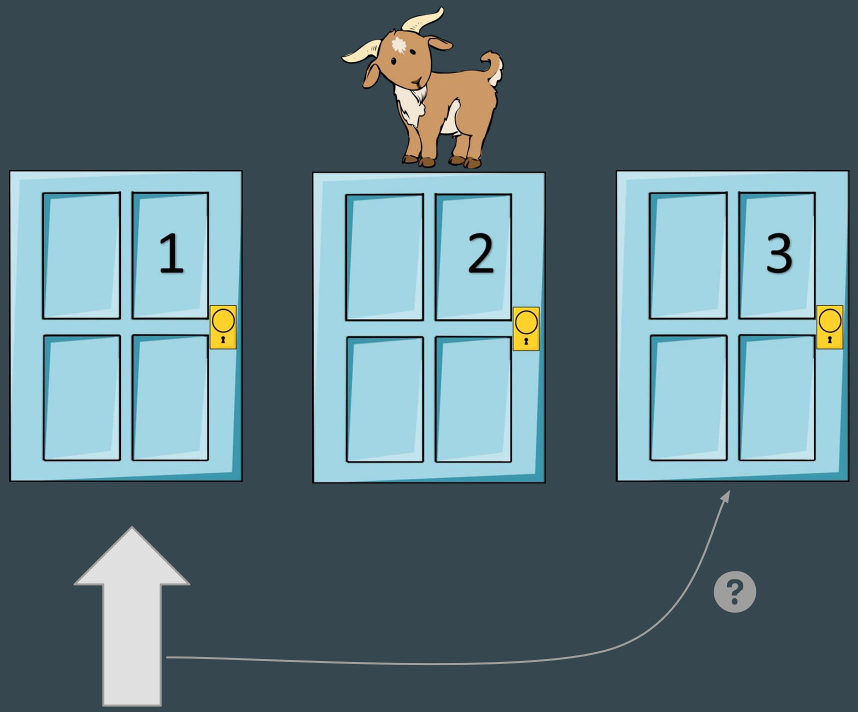
$$p(X) = \int p(X|\theta)p(\theta)d\theta$$

Bayes Theorem: How to Calculate P(θ)?

1. Create Your Own
2. Take Previous P(θ | X)

How!?: Part 1 - Discrete Case

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



How!?: Part 1 - Priors

Hypothesis i	Prior p(θ_i)
Car Behind 1	1/3
Car Behind 2	1/3
Car Behind 3	1/3

How!?: Part 1 - Likelihoods Given Priors

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!?: Part 1 - Likelihoods Given Priors

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!?: Part 1 - Likelihoods Given Priors

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

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

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

How!?: Part 1 - Discrete Case Recap

- 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 the parameters for each variant given the data?
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

Prior?:

- Uninformed Prior
- Indicator Function

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

How!?: Part 2 - Continuous Case

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

How!?: Part 2 - Continuous Case

$$P(\theta|X) = \frac{\theta^{\sum y_i}(1-\theta)^{n-\sum y_i} I_{\{0\leq\theta\leq1\}}}{\int_0^1 \theta^{\sum y_i}(1-\theta)^{n-\sum y_i} I_{\{0\leq\theta\leq1\}} d\theta}$$

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

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

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

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

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

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	$k,\ \theta$	$k + \sum_{i=1}^{n} x_i,\ \frac{\theta}{n\theta + 1}$	k total occurrences in $\frac{1}{\theta}$ intervals	$\mathrm{NB}(\tilde{x} \mid k', \theta')$ (negative binomial)
			$\alpha,\ \beta$[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	$\alpha,\ \beta$[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")

Demo:
Conjugate Priors

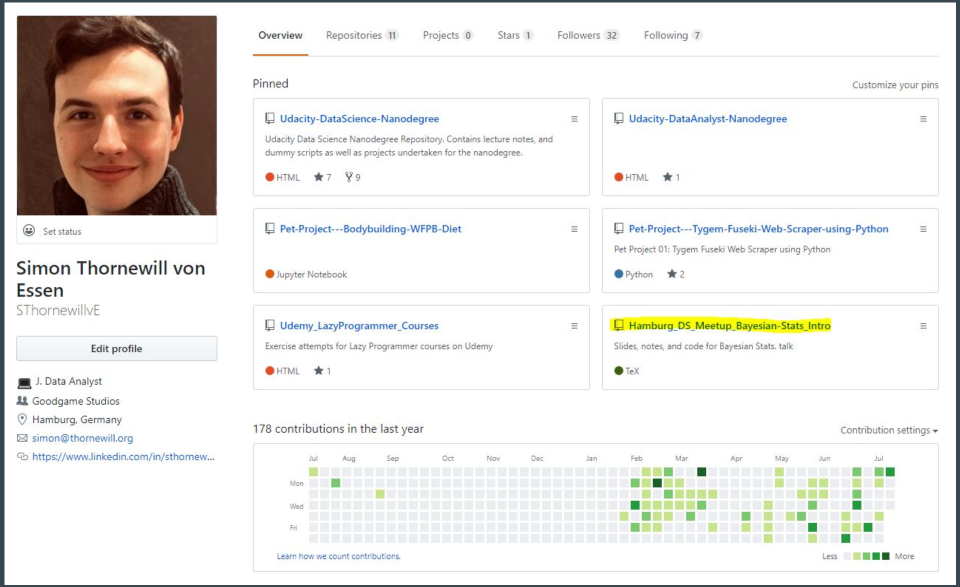
Foreshadow: MCMC

- Integrating is hard

Conclusion and “Call to Action”

Find Slides on Github

https://cutt.ly/zGqux9



Find me on Social Media!

...

`@sthornewillve`