

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.



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 the probability?

- ~~Classical~~

- Frequentist



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

- Bayesian

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

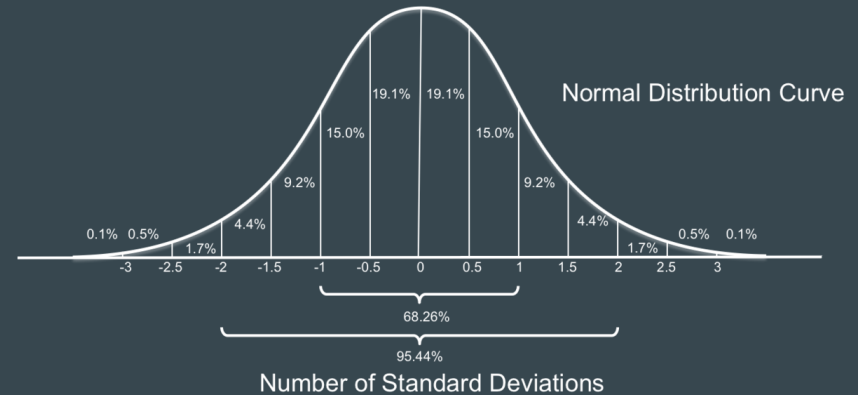
Calculate the probability of your data given your data following some parameter.



Thermometer Calibration: Frequentist Est.

Confidence Interval:

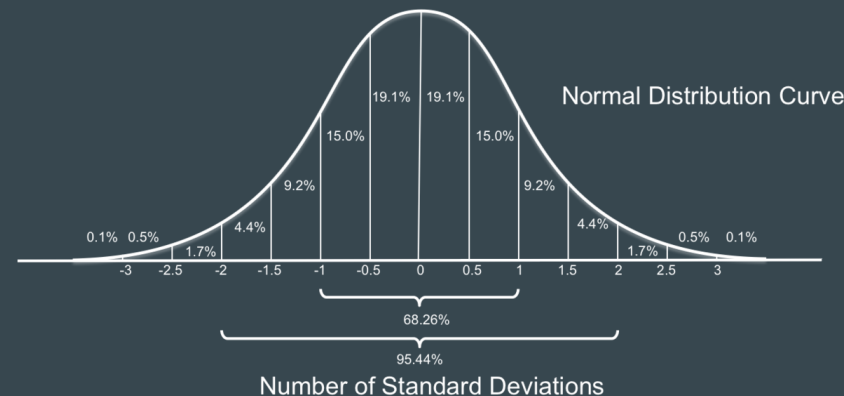
- From sample mean and standard deviation, calculate an interval
- “Interval that contains the true parameter some percent of the time upon repeated experiments”



Thermometer Calibration: Frequentist Est.

Confidence Interval:

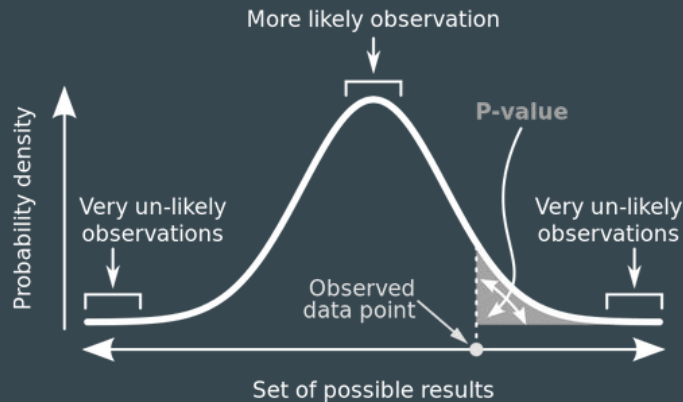
- Intuition:
 - If you were to bootstrap the confidence interval n times
 - Interval would contain the mean of population 95% of the time





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”
- $P(X | \mu)$



Thermometer Calibration: Test

- 1. If P-value = 0.001 (highly significant), is the probability of getting this result given our data 0.001? 
- 2. Does a 95% Confidence interval contain the true value 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?

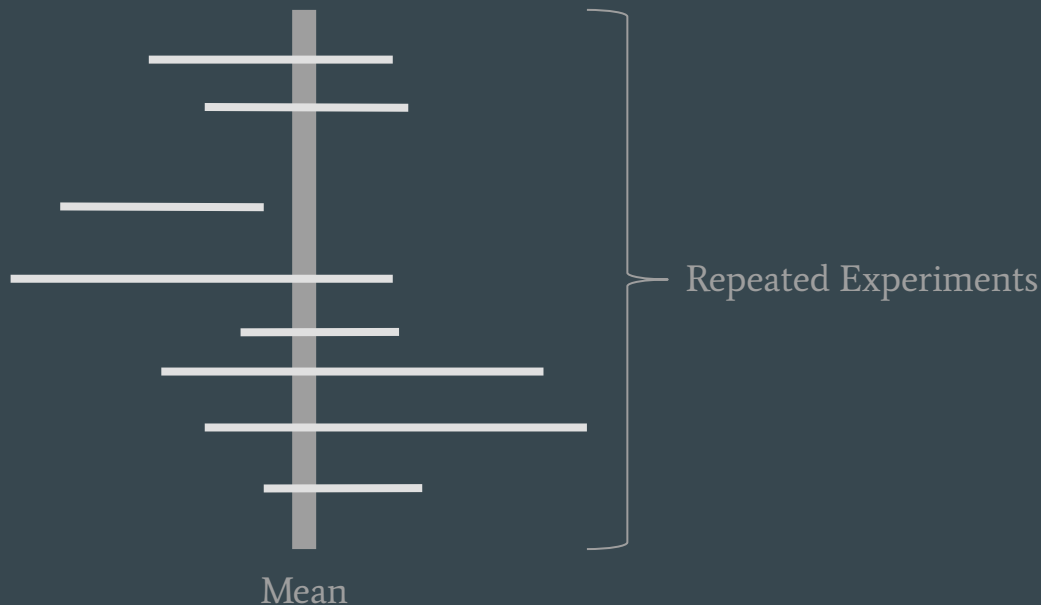
Probability of getting this result given no difference in experimental groups is 0.001

- 2. Does a 95% Confidence interval contain the true value 95% of the time?

Interval with 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



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 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

$$\overset{\text{posterior}}{p(B \mid A)} = \frac{\overset{\text{likelihood}}{p(A \mid B)} \overset{\text{prior}}{p(B)}}{\underset{\text{normalisation}}{p(A)}}$$

Bayes Theorem: Alternate View

θ = Parameter,

X = Data

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

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

Bayes Theorem: Inverting our “view”

- i Bayesianism expects that parameters can vary, what is certain is our data



Bayes Theorem: Inverting our “view”

- i If you say, “wait, wasn’t this was we were doing with frequentism?”, you understand why people don’t like it.



Bayes Theorem: Alternate View

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

The equation is annotated with a green checkmark above the numerator and yellow question marks next to the terms $p(X|\theta)p(\theta)$ and $p(X)$.

❓ 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 $P(\theta)$?



Bayes Theorem: How to Calculate P(X)?

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

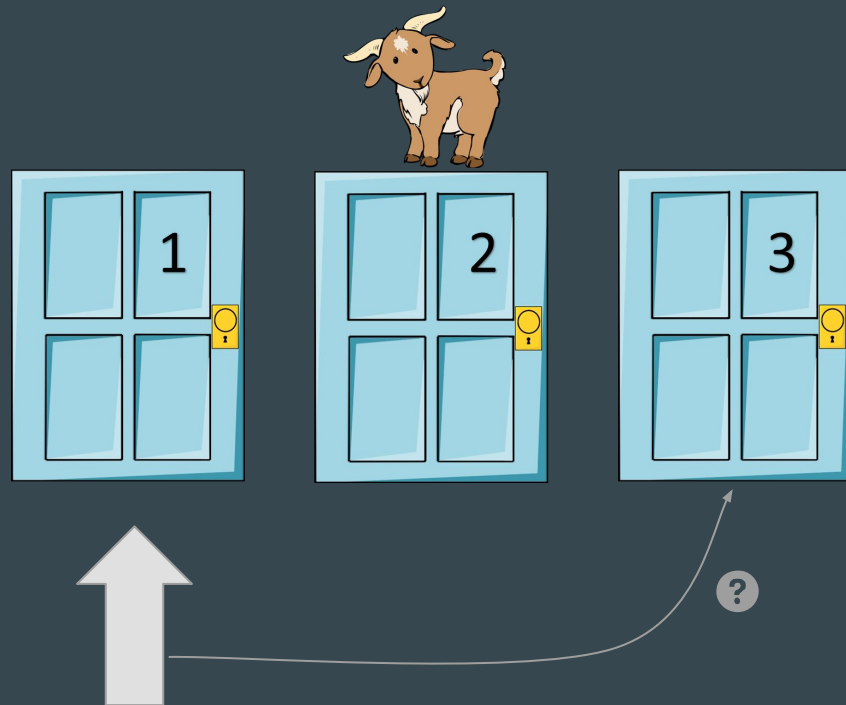
1. Sum of all possible numerators
2. Yes, this can get difficult


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

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

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 $p(X \theta_i)$
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	1/3

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

* 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
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

How!?: Train Analysis

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

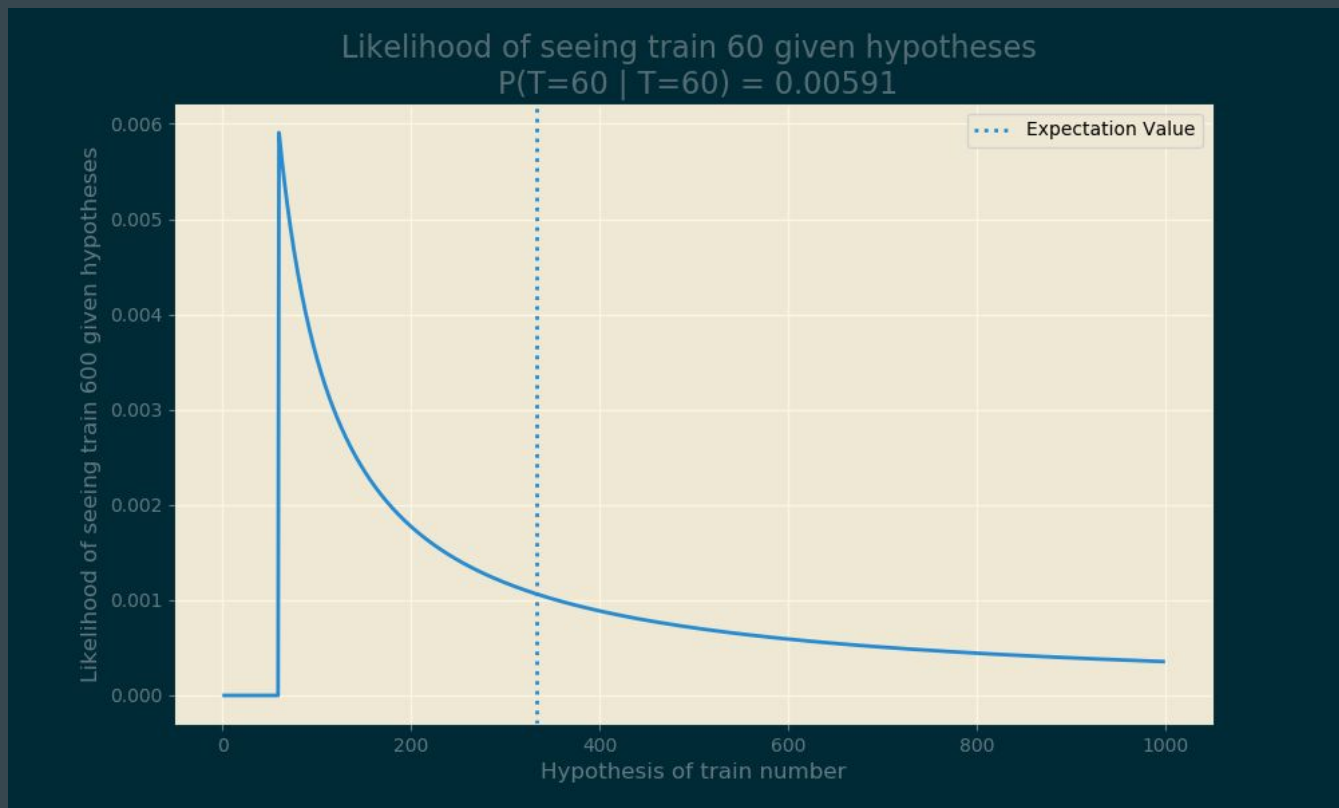


How!?: Train Analysis

Hypothesis i	Prior $p(\theta_i)$	Likelihood $p(X \theta_i)$	Prior * Likelihood	Posterior
1 Train	1/N	0	0	post_1
2 Trains	1/N	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



How!?: Train Analysis

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



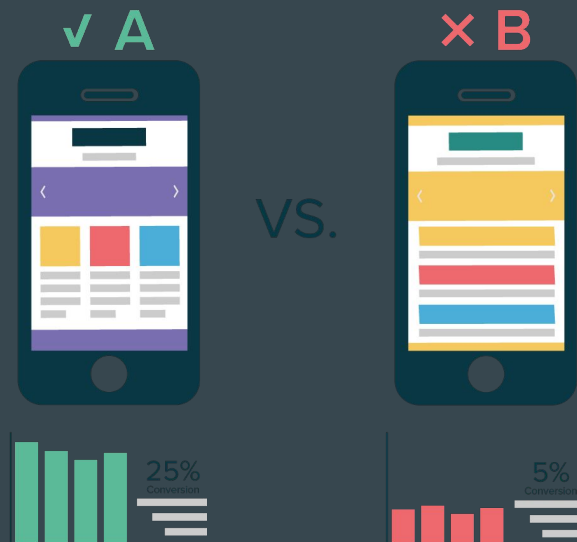
How!?: Part 1 - Discrete Case Recap

- 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 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

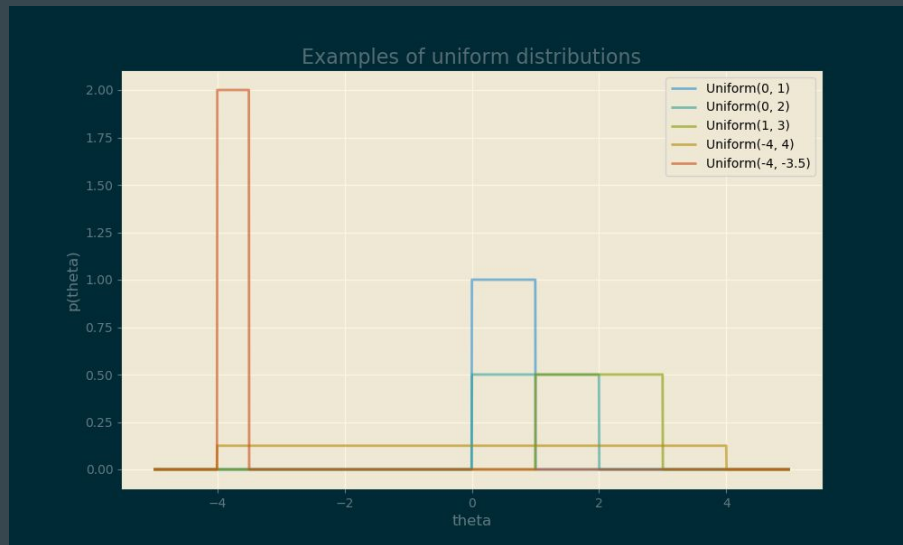
$$P(X|\theta) = \theta^{\sum y_i} (1 - \theta)^{n - \sum y_i}$$

How!?: Part 2 - Continuous Case

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

Prior?:

- Uninformed Prior
- Uniform distribution
- Represented by
Indicator Function



How!?: Part 2 - Continuous Case

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

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

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

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

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



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)

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

When is it okay not to perform Normalisation?

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

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

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

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

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”)

Posteriors - What Now?

- Estimation of Parameters
- Credible Intervals
- Check priors

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

Demo: Conjugate Priors

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

- Disadvantages

- Steeper learning curve
- Integrals are hard
- Workarounds can be computationally expensive
- Criticisms of being less “objective” due to use of priors
- Uniform prior gets special criticism
- Estimations become the same as frequentist estimations with high sample sizes (redundant)

Foreshadow: MCMC

- Integrating is hard
- Is it possible to estimate an integral without complicated calculations?

Conclusion and “Call to Action”

- Remember differences between F and B to understand both
- Understand that it's not as difficult as it looks
- Practice makes perfect
- Sources to get started
- Remember when you should consider it
- Might not be necessary

Resources for further learning

The screenshot shows a GitHub repository page for 'SThorneWillvE / Hamburg_DS_Meetup_Bayesian-Stats_Intro'. The repository has 0 Watchers, 0 Stars, and 0 Forks. The main navigation bar includes links for Code, Issues (0), Pull requests (0), Projects (0), Wiki, Security, Insights, and Settings. The current view is for the file 'sources / bayesian-stats_sources.md' on the 'master' branch. A commit by 'SThorneWillvE' is shown with the message 'Add wikipedia conjugate prior source' and commit hash '912b045' dated 'on May 31'. The file statistics show '60 lines (30 sloc)' and '2.42 KB'. The file content is displayed in a light gray box with buttons for 'Raw', 'Blame', 'History', and icons for opening a new file, editing, and deleting. The content of the file is as follows:

Bayesian Stats Sources

In this subdirectory of the repository I would like to outline the major sources that I have used to inform my understanding of bayesian statistics and should be used as a starting point.

I will also outline the strengths and weaknesses of each source so that people can more easily choose which sources are most relevant for them.

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

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, and a list of pinned repositories. The pinned repositories are: 'Udacity-DataScience-Nanodegree', 'Udacity-DataAnalyst-Nanodegree', 'Pet-Project---Bodybuilding-WFPB-Diet', 'Pet-Project---Tygem-Fuseki-Web-Scraper-using-Python', 'Udemy_LazyProgrammer_Courses', and 'Hamburg.DS.Meetup.Bayesian-Stats_Intro'. The 'Hamburg.DS.Meetup.Bayesian-Stats_Intro' repository is highlighted with a yellow background. Below the pinned repositories, there is a section for '178 contributions in the last year' with a calendar grid showing contributions from July to July. The grid shows a high density of contributions in the latter half of the year, particularly in the last few months.

Find me on Social Media!



@sthornewillve



Extra Slides

...

Bayes Theorem: Derivation

The Same

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



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