Bayesian Reasoning

Librarian or Farmer

This is a slightly modified abstract from Daniel Kahneman's Thinking, Fast and Slow

> An individual has been described by a neighbor as follows: “Steve is very shy and withdrawn. 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 farmer?

So, is Steve more likely to be a librarian or a farmer?

Lets give some numbers to your intuitions

Let's say that

- the ratio of librarians that meet this description is high, 0.8
- the ratio of farmers that meet this description is very low, 0.05

Lets give some numbers to your intuitions

From that assumption, we get that

$$P(steve = librarian) = \frac{0.8}{0.8 + 0.05} \approx 0.94$$

What about priors?

What if we consider that there are 20 times more farmers than librarians?

- Does it change your belief that Steve is a Librarian?
- If so, by how much?

Bayes' Rule

Let's rephrase Bayes' rule as

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} = \frac{P(E|H)P(H)}{P(E|H)P(H) + P(E|\neg H)P(\neg H)}$$

- H : Hypothesis
- E : Evidence
- $\neg$: Not
- $P(H|E)$: Probability after incorporating new evidence (`posterior`)
- $P(H)$: Probability before incorporating new evidence (`prior`)

Bayes' Rule

For our problem

- H: Steve is a Librarian
- E: Steve's description
- $P(E|H)$: ratio librarians that fit the description = 0.8
- $P(E|\neg H)$ = ratio of farmers that fit the description = 0.05
- $P(H)$ = ratio of librarians to farmers (without reading the description). 1/21

Bayesian' Rule

Replacing our values on Bayes' Rule

$$P(Steve = Librarian|text) = \frac{0.8 * 1/21}{0.8 * 1/21 + 0.05 * 20/21}$$
$$= \frac{0.8}{0.8 + 1} = 0.\hat{4} \approx 0.44$$

What about additional evidence?

What if we are then told that Steve lives near the city

Let's say that

- the ratio of librarians that live near the city is moderate, 0.5
- the ratio of farmers that live near the city is low, 0.1

How do we incorporate this new evidence?

Considering new evidence

- H: Steve is a Librarian
- E: Steve lives near the city
- $P(E|H)$: ratio librarians that live near the city = 0.5
- $P(E|\neg H)$ = ratio of farmers that live near the city = 0.1
- $P(H)$ = Our previous believe of Steve being a librarian = $0.\hat{4}$

$$P(Steve = Librarian|city) = \frac{0.5 * 0.\hat{4}}{0.5 * 0.\hat{4} + 0.1 * 0.\hat{5}} = 0.80$$

Excercise

Prove that applying Bayes' Rule considering the description of Steve and then his living location (as we have done) results in the same probabilities as conditioning on both evidences at once

Bayesian Inference

Method of statistical inference that uses Bayes' Rule to update the probability of a rv (hypothesis) as more information becomes available.

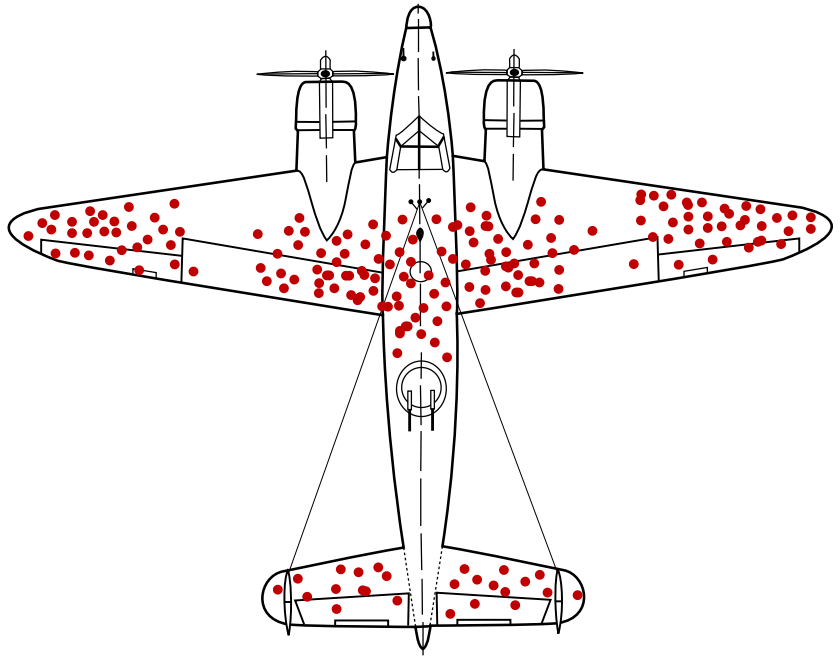
Evidence analysis

WWII Aircrafts

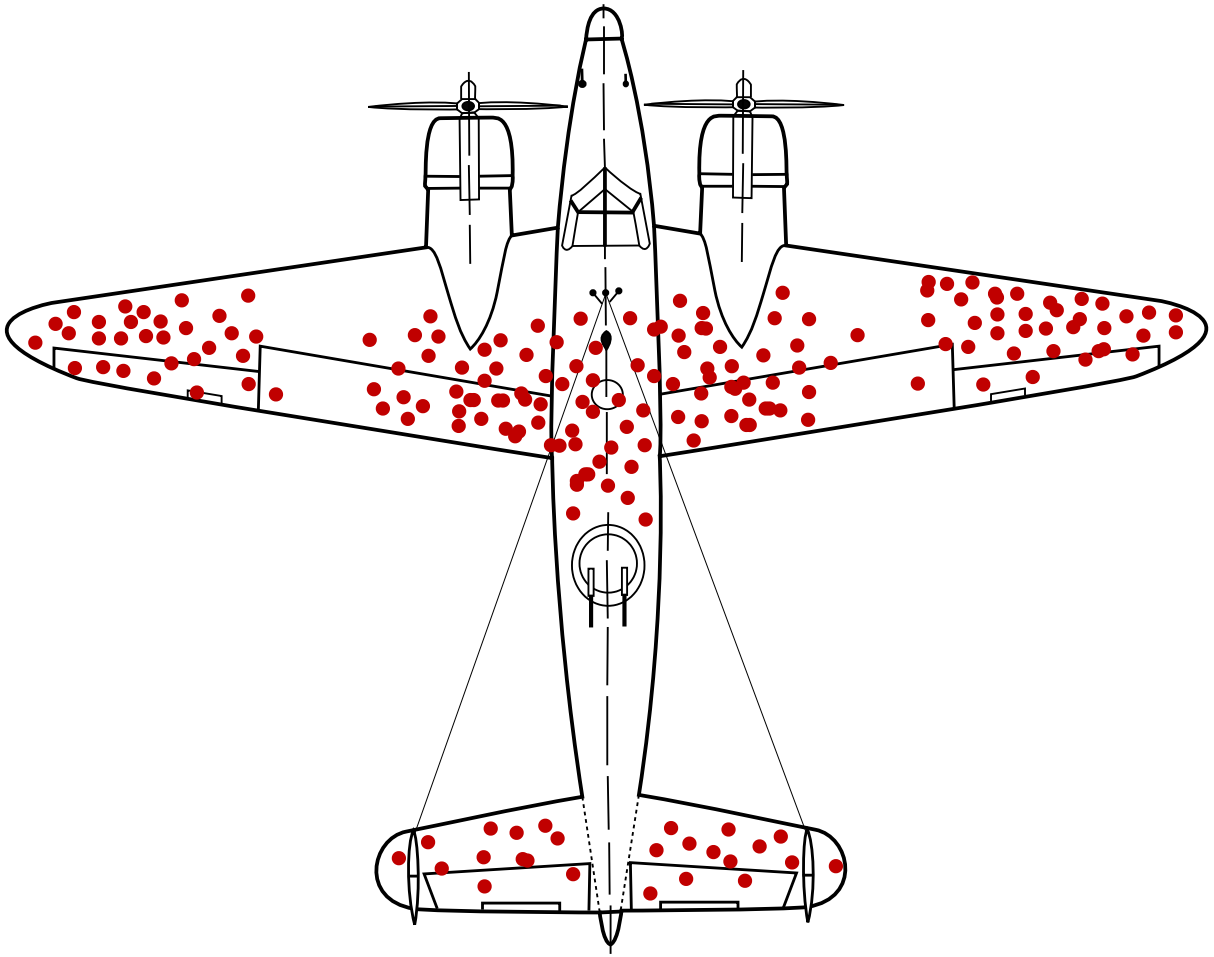
During WWII the US Navy conducted a study to determine where aircraft should be armored to better survive enemy fire.

They collected data from aircraft that had been deployed to conflict areas.

The red dots represent the locations where aircrafts had been shot the most.



Where should armor be placed to improve the chances of aircraft surviving enemy fire?



Recommended Videos

3Blue1Brown: Bayes theorem, the geometry of changing beliefs `15:45 min`

Veritasium: How To Update Your Beliefs Systematically - Bayes' Theorem `10:36 min`

Julia Galef: A visual guide to Bayesian thinking `11:24 min`

Bayes' Rule for continuous distributions

Bayes' Rule

Let's rewrite Bayes' rule for estimating a continuous rv X, given some observation o

$$p(x|o) = \frac{p(o|x)p(x)}{\int_X p(o|x)p(x)dx}$$

$p(x|o)$: `posterior`

$p(o|x)$: `likelihood`

$p(x)$: `prior`

$\int_X p(o|x)p(x)dx$: `normalization constant`

Normalization constant

$$\int_X p(o|x)p(x)dx$$

is referred to as the normalization constant, as it guarantees that the posterior probability is a proper probability density function (as guarantees it integrates to 1)

This integral can make Bayesian inference intractable as it is hard to compute for arbitrary functions.

To keep integrals trackable, it should be possible to use ***parametric functions***, as their integrals would be well defined.

`Note` **Instead of using parametric functions, some approaches go around the integral tractability issue using ratios/proportionality. We will later study about Markov Chain Monte Carlo, which is one such method**

Bayes' Update

Lets consider we want to update our posterior using $\text{likelihood}_{(0)}$,
Using Bayes (ommiting normalization constant for now),

$$\text{posterior}_{(0)} \propto \text{likelihood}_{(0)} * \text{prior}_{(0)}$$

If we then want to update it using $\text{likelihood}_{(1)}$, we get

$$\text{posterior}_{(1)} \propto \text{likelihood}_{(1)} * \text{posterior}_{(0)}$$

And so on,

$$\text{posterior}_{(2)} \propto \text{likelihood}_{(2)} * \text{posterior}_{(1)}$$

What we can notice is that unless the posterior's and the likelihood*prior's function are the same, or cyclic, it would not be possible to keep using a finite set of distributions, as we would need to change the distribution each time we updated the posteriors.

Conjugate Distributions/Priors

When posteriors and priors belong to the same parametric function after Bayesian update,

- The posterior and prior are called **conjugate distributions**, and
- The family of distributions of the prior is called the **conjugate prior** of the family of distributions of the likelihood.

Commonly used Conjugate Priors

Beta Distribution is a conjugate prior to:

- Bernoulli (Binomial) Distribution
- Geometric Distribution

Gamma Distribution is a conjugate prior to:

- Gamma Distribution
- Exponential Distribution
- Gaussian Distribution

Gaussian Distribution is a conjugate prior to:

- Gaussian Distribution

Product of two Gaussians

Having two Gaussians, $f(\cdot)$ and $g(\cdot)$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma_f} \exp\left(-\frac{(x-\mu_f)^2}{2\sigma_f^2} \right)$$
$$g(x) = \frac{1}{\sqrt{2\pi}\sigma_g} \exp\left(-\frac{(x-\mu_g)^2}{2\sigma_g^2} \right)$$

The product $f(\cdot)g(\cdot)$ is

$$f(x)g(x) = \frac{1}{2\pi\sigma_f\sigma_g} \exp\left(-\left(\frac{(x-\mu_f)^2}{2\sigma_f^2} + \frac{(x-\mu_g)^2}{2\sigma_g^2} \right) \right)$$

Focusing on the exponential

$$\frac{(x-\mu_f)^2}{2\sigma_f^2}+\frac{(x-\mu_g)^2}{2\sigma_g^2}$$

we can expand the power of two and rearrange it as,

$$\frac{x^2(\sigma_f^2+\sigma_g^2)-2x(\mu_f\sigma_g^2+\mu_g\sigma_f^2)+\mu_f^2\sigma_g^2+\mu_g^2\sigma_f^2}{2\sigma_f^2\sigma_g^2}$$

$$\frac{x^2-2x\frac{\mu_f\sigma_g^2+\mu_g\sigma_f^2}{\sigma_f^2+\sigma_g^2}+\frac{\mu_f^2\sigma_g^2+\mu_g^2\sigma_f^2}{\sigma_f^2+\sigma_g^2}}{2\frac{\sigma_f^2\sigma_g^2}{\sigma_f^2+\sigma_g^2}}$$

which has the same form of the exponential for a Gaussian with

$$\mu_{fg} = \frac{\mu_f \sigma_g^2 + \mu_g \sigma_f^2}{\sigma_f^2 + \sigma_g^2}, \quad \sigma_{fg} = \sqrt{\frac{\sigma_f^2 \sigma_g^2}{\sigma_f^2 + \sigma_g^2}}$$

By simple algebraic manipulation we can then get

$$f(x)g(x) \propto \frac{1}{\sqrt{2\pi}\sigma_{fg}} \exp\left(-\frac{(x - \mu_{fg})^2}{2\sigma_{fg}^2} \right)$$

`Note` **There is a residual leftover in the exponential that does not depend on x, so it can be taken out as a factor that scales the resulting Gaussian. The product of two Gaussians is in fact not a normalized Gaussian (does not integrate to 1), but a scaled one. When replaced into Bayes, the normalization constant will be equal to the multiplicative inverse of this scaling factor, as the resulting distribution HAS to integrate to 1 to be a valid distribution, as per definition.**

Excercise

Make a similar proof showing the Beta Distribution is the conjugate prior of the Binomial distribution

Hint:
The results of a Beta Distribution multiplied by a Binomial is

$$\mathbf{Beta}(\theta|\alpha+1,\beta) \propto \mathbf{Bernoulli}(x=1|\theta)\,\mathbf{Beta}(\theta|\alpha,\beta)$$

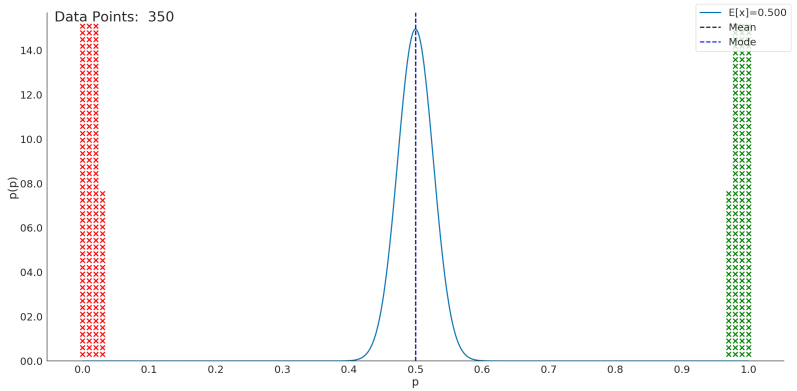
$$\mathbf{Beta}(\theta|\alpha,\beta+1) \propto \mathbf{Bernoulli}(x=0|\theta)\,\mathbf{Beta}(\theta|\alpha,\beta)$$

Coin Toss

Using a Beta Distribution as prior, specifically $\mathrm{Beta}(p|\alpha = 1, \beta = 1) = \mathcal{U}(p|0, 1)$, and a Binomial Distribution as likelihood

Results

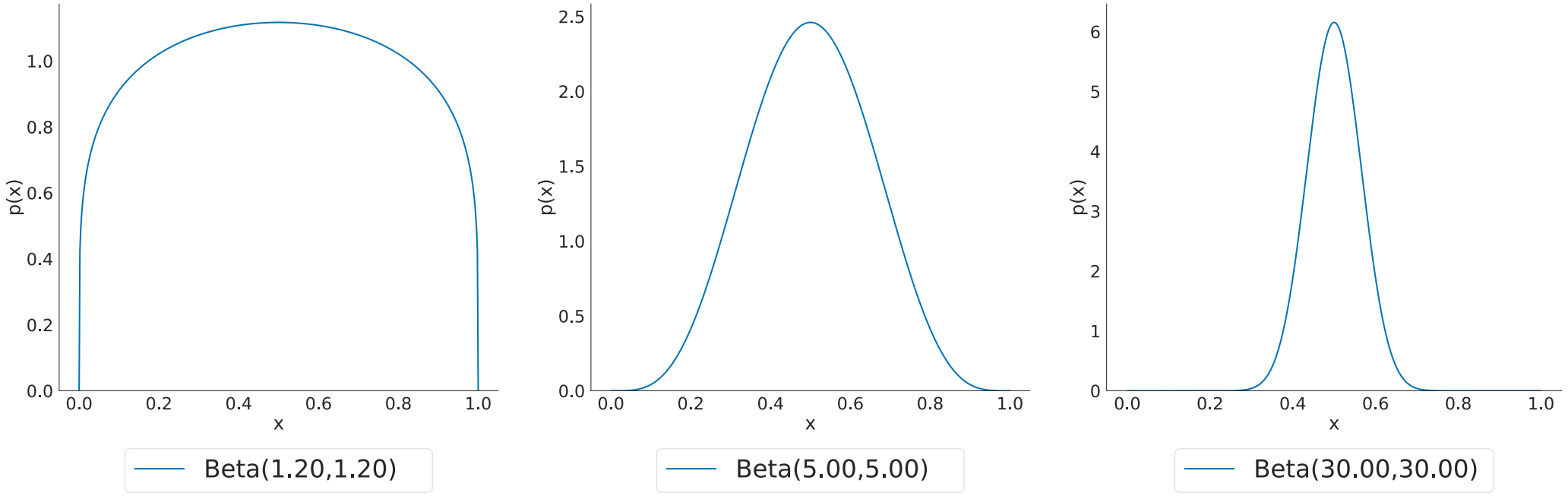
After running the previous simulation we obtain:



What would be the result of the simulation?

- Probability of getting heads/tails = 50% ?
- This would be an estimate of the distribution shown above (its mean)

The following distributions have the same point estimate as the previous one ($p = 0.5$)

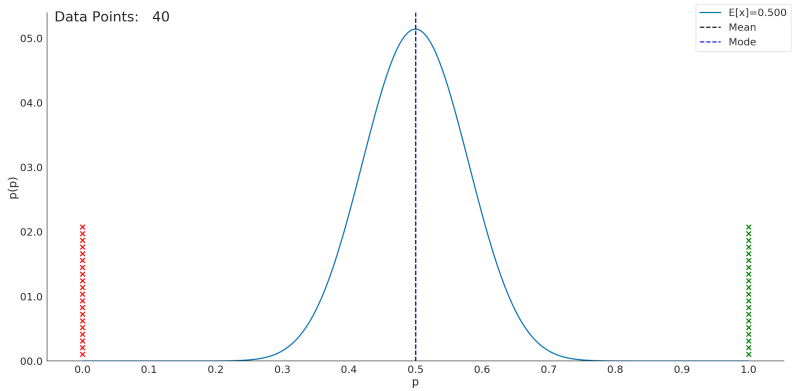


Are they the same?

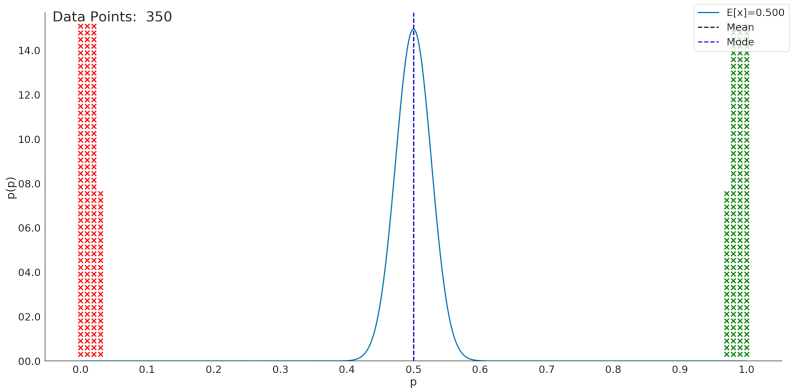
Distributions vs Estimates

Distributions give us more information.

If we are interested in the confidence of predictions in a range of values we can compute the confidence that p is between 0.45 and 0.55



(a)



(b)

$$P(0.45 < p < 0.55) = \int_{0.45}^{0.55} \text{Beta}$$

- 0.48 (a)
- 0.94 (b)

Distributions vs Estimates

We can also compute p_0 and p_1, so that $P(p_0 \leq p \leq p_1) = 0.95$ (95% credible interval)

$$p_0 \ : \int_0^{p_0} \text{Beta}(p) = 0.025, \qquad p_1 \ : \int_p^{p_1} \text{Beta}(p) = 0.975$$

which are 0.351 to 0.649 for the left distribution and 0.448 to 0.552.
That is,

- For (a) there is a 95% probability that p falls within 0.351 and 0.649
  That is, $0.351 < p < 0.649$ is the 95% credible interval.
- For (b) there is a 95% probability that p falls between 0.448 and 0.552
  That is, $0.448 < p < 0.552$ is the 95% credible interval.

Cumulative density function

For a rv X, the cumulative density function $CDF(x)$ gives the probability that X will take a value less than or equal to x

$$\text{CDF}(x) = \int_{-\infty}^{x} p(x)$$

Using the CDF to compute the probability of a range

$$P(0.45 < p < 0.55) = \mathrm{CDF}(0.55) - \mathrm{CDF}(0.45)$$

Quantile function

(a.k.a Percent Point Function)

For a rv X, the quantile function $Q(p)$ gives the value at which the probability of the rv will take a value less than or equal to the given probability

So, Q is the inverse of the CDF

$$Q(p) = \text{CDF}^{-1}$$

Using the Q to compute the values for a 95% credible range

$$p_0 = Q(0.025), \qquad p_1 = Q(0.975)$$

`Side note`

Why we like parametric functions (specially Gaussians)

For most density functions there are no closed formulas to compute the CDF, however, given its common use, most software has efficient implementations for its calculation

Why we use parametric functions (specially Gaussians)

Likewise (or more so), there are no closed formulas to compute Q, but its common to have efficient implementations
A particular case is for Gaussian distributions (known as the `68-95-99.7 rule`).

For a Gaussian distribution, an approximate credible interval of 68%, 95% and 99.7% can be easily found by correspondingly adding/substracting 1, 2 and 3 standard deviations. So

- $\mu \pm \sigma$ corresponds to approximately a 68% credible interval
- $\mu \pm 2\sigma$ corresponds to approximately a 95% credible interval
- $\mu \pm 3\sigma$ corresponds to approximately a 99.7% credible interval