Probability foundations

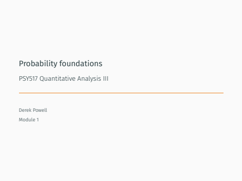
PSY517 Quantitative Analysis III

Derek Powell

Module 1

Probability foundations

2021-08-31



Probability foundations

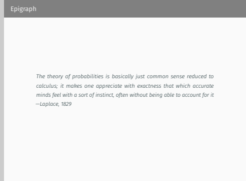
PSY517 Quantitative Analysis III

Derek Powell

Module 1

Epigraph

The theory of probabilities is basically just common sense reduced to calculus; it makes one appreciate with exactness that which accurate minds feel with a sort of instinct, often without being able to account for it
—Laplace, 1829



Epigraph

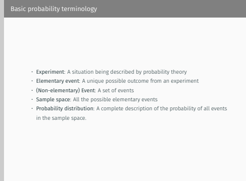
The theory of probabilities is basically just common sense reduced to calculus; it makes one appreciate with exactness that which accurate minds feel with a sort of instinct, often without being able to account for it
—Laplace, 1829

- many ways to view probability: one way is to see it as the inductive version of formal logic
- Laplace is proposing a real psychological theory here
- Now if he were totally right, we wouldn’t need probability. Nor logic
  - so don’t worry if you are feeling you do not have an “accurate mind”
  - but hopefully you will begin to develop an intuition.

Basic probability terminology

- **Experiment**: A situation being described by probability theory
- **Elementary event**: A unique possible outcome from an experiment
- **(Non-elementary) Event**: A set of events
- **Sample space**: All the possible elementary events
- **Probability distribution**: A complete description of the probability of all events in the sample space.

└─Basic probability terminology



Basic probability terminology

- **Experiment**: A situation being described by probability theory
- **Elementary event**: A unique possible outcome from an experiment
- **(Non-elementary) Event**: A set of events
- **Sample space**: All the possible elementary events
- **Probability distribution**: A complete description of the probability of all events in the sample space.

- If we “conduct” or “observe” the experiment, only one elementary event will happen. they are mutually exclusive

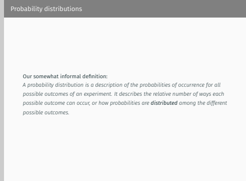
in the pants example

- The experiment is picking pants to wear
- Elementary events are each pair of pants
- Non-elementary events might be “wearing jeans” (set of 3 elementary events)
- Sample space = all the pants
- Probability distribution - … next slide

Probability distributions

Our somewhat informal definition:

A probability distribution is a description of the probabilities of occurrence for all possible outcomes of an experiment. It describes the relative number of ways each possible outcome can occur, or how probabilities are **distributed** *among the different possible outcomes.*



Probability distributions

Our somewhat informal definition:

A probability distribution is a description of the probabilities of occurrence for all possible outcomes of an experiment. It describes the relative number of ways each possible outcome can occur, or how probabilities are **distributed** *among the different possible outcomes.*

- the main thing to understand from this lecture
- when we say P(x) we will usually mean a distribution

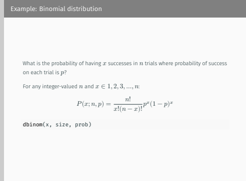
Example: Binomial distribution

What is the probability of having x successes in n trials where probability of success on each trial is p?

For any integer-valued n and $x \in 1, 2, 3, ..., n$:

$$P(x; n, p) = \frac{n!}{x!(n-x)!} p^x (1-p)^x$$

```
dbinom(x, size, prob)
```



Example: Binomial distribution

What is the probability of having x successes in n trials where probability of success on each trial is p?

For any integer-valued n and $x \in 1, 2, 3, ..., n$:

$$P(x; n, p) = \frac{n!}{x!(n-x)!} p^x (1-p)^x$$

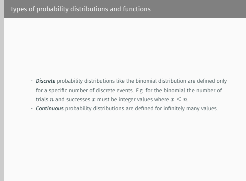
dbinom(x, size, prob)

- aka, how many ways (out of all the possible ways) can you get x successes on n trials with prob p
- same as globe tossing model, pulling balls out of urn, n flips of a coin, etc.
- this is binomial pmf

Types of probability distributions and functions

- ***Discrete*** probability distributions like the binomial distribution are defined only for a specific number of discrete events. E.g. for the binomial the number of trials n and successes x must be integer values where $x \leq n$.
- ***Continuous*** probability distributions are defined for infinitely many values.

└─Types of probability distributions and functions



Types of probability distributions and functions

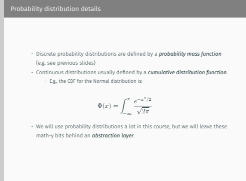
- ***Discrete*** probability distributions like the binomial distribution are defined only for a specific number of discrete events. E.g. for the binomial the number of trials n and successes x must be integer values where $x \leq n$.
- ***Continuous*** probability distributions are defined for infinitely many values.

Probability distribution details

- Discrete probability distributions are defined by a ***probability mass function*** (e.g. see previous slides)
- Continuous distributions usually defined by a ***cumulative distribution function***.
  - E.g., the CDF for the Normal distribution is:

$$\Phi(x) = \int_{-\infty}^{x} \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

- We will use probability distributions a lot in this course, but we will leave these math-y bits behind an ***abstraction layer***.



Probability distribution details

- Discrete probability distributions are defined by a ***probability mass function*** (e.g. see previous slides)
- Continuous distributions usually defined by a ***cumulative distribution function***.
  - E.g., the CDF for the Normal distribution is:

$$\Phi(x) = \int_{-\infty}^{x} \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

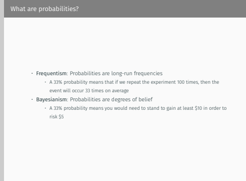
- We will use probability distributions a lot in this course, but we will leave these math-y bits behind an ***abstraction layer***.

- We will let them be a black box, will describe them with samples (as we'll get to)
- Abstraction layer/black boxes are important tools.
- You will have permission, at least for most of this course, to not care about most of these details

What are probabilities?

- **Frequentism**: Probabilities are long-run frequencies
  - A 33% probability means that if we repeat the experiment 100 times, then the event will occur 33 times on average
- **Bayesianism**: Probabilities are degrees of belief
  - A 33% probability means you would need to stand to gain at least $10 in order to risk $5

└─What are probabilities?



What are probabilities?

- **Frequentism**: Probabilities are long-run frequencies
  - A 33% probability means that if we repeat the experiment 100 times, then the event will occur 33 times on average
- **Bayesianism**: Probabilities are degrees of belief
  - A 33% probability means you would need to stand to gain at least $10 in order to risk $5

- frequentism - actually means if we repeatedinfinitely many times, would be 33% of the times
- Bayesianism doesn't disagree with frequentism, also thinks the long run frequencies will result—it's just that isn't the DEFINITION of probability
- One way to see the difference is for things that will only happen (or not happen) once. What is the probability it will rain in PHX today?

Bayes’ rule

Bayes’ rule equation

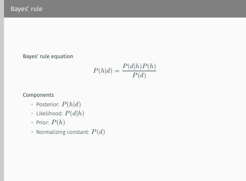
$$P(h|d) = \frac{P(d|h)P(h)}{P(d)}$$

Components

- Posterior: $P(h|d)$
- Likelihood: $P(d|h)$
- Prior: $P(h)$
- Normalizing constant: $P(d)$

2021-08-31

Bayes’ rule



Bayes’ rule

Bayes’ rule equation

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)}$$

Components

- Posterior: $P(h|d)$
- Likelihood: $P(d|h)$
- Prior: $P(h)$
- Normalizing constant: $P(d)$

- take time with this, explain each term
- the key idea is it’s just a simple fact of probability, nothing fancy. But it’s USEFUL
- Ask if people want the math, show if they do or skip

Bayes rule with simple probabilities

Suppose we are vampire hunters and our lab has just processed a positive blood test for vampirism. We can calculate the probability that the suspect is a vampire as:

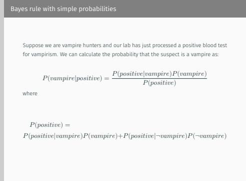
$$P(vampire|positive) = \frac{P(positive|vampire)P(vampire)}{P(positive)}$$

where

$$P(positive) =$$
$$P(positive|vampire)P(vampire)+P(positive|\neg vampire)P(\neg vampire)$$

2021-08-31

└─Bayes rule with simple probabilities



Bayes rule with simple probabilities

Suppose we are vampire hunters and our lab has just processed a positive blood test for vampirism. We can calculate the probability that the suspect is a vampire as:

$$P(vampire|positive) = \frac{P(positive|vampire)P(vampire)}{P(positive)}$$

where

$$P(positive) =$$
$$P(positive|vampire)P(vampire)+P(positive|\neg vampire)P(\neg vampire)$$

Diagnosing vampirism

$$P(vampire|positive) = \frac{P(positive|vampire)P(vampire)}{P(positive)}$$

Suppose the test successfully detects vampirism 95% of the time, and only issues a false-positive 1% of the time. Of course, vampires are rare, only 0.1% of the population.

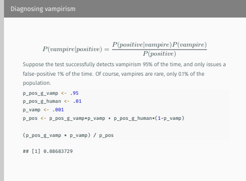
```
p_pos_g_vamp <- .95
p_pos_g_human <- .01
p_vamp <- .001
p_pos <- p_pos_g_vamp*p_vamp + p_pos_g_human*(1-p_vamp)

(p_pos_g_vamp * p_vamp) / p_pos
```

```
## [1] 0.08683729
```

Probability foundations

2021-08-31

└─Diagnosing vampirism



Diagnosing vampirism

$$P(vampire|positive) = \frac{P(positive|vampire)P(vampire)}{P(positive)}$$

Suppose the test successfully detects vampirism 95% of the time, and only issues a false-positive 1% of the time. Of course, vampires are rare, only 0.1% of the population.

```
p_pos_g_vamp <- .95
p_pos_g_human <- .01
p_vamp <- .001
p_pos <- p_pos_g_vamp*p_vamp + p_pos_g_human*(1-p_vamp)

(p_pos_g_vamp * p_vamp) / p_pos
```

```
## [1] 0.08683729
```

Bayes rule with probability distributions

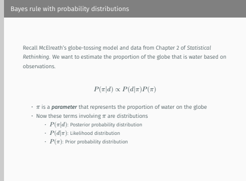
Recall McElreath's globe-tossing model and data from Chapter 2 of *Statistical Rethinking*. We want to estimate the proportion of the globe that is water based on observations.

$$P(\pi|d) \propto P(d|\pi)P(\pi)$$

- π is a ***parameter*** that represents the proportion of water on the globe
- Now these terms involving π are distributions
  - $P(\pi|d)$: Posterior probability distribution
  - $P(d|\pi)$: Likelihood distribution
  - $P(\pi)$: Prior probability distribution

2021-08-31

└─Bayes rule with probability distributions



Bayes rule with probability distributions

Recall McElreath's globe-tossing model and data from Chapter 2 of *Statistical Rethinking*. We want to estimate the proportion of the globe that is water based on observations.

$$P(\pi|d) \propto P(d|\pi)P(\pi)$$

- π is a ***parameter*** that represents the proportion of water on the globe
- Now these terms involving π are distributions
  - $P(\pi|d)$: Posterior probability distribution
  - $P(d|\pi)$: Likelihood distribution
  - $P(\pi)$: Prior probability distribution

Now each of these things is a distribution. note likelihood distribution is technically not a probability distribution but

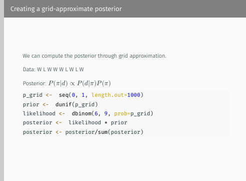
Creating a grid-approximate posterior

We can compute the posterior through grid approximation.

Data: W L W W W L W L W

Posterior: $P(\pi|d) \propto P(d|\pi)P(\pi)$

```
p_grid <-  seq(0, 1, length.out=1000)
prior <-  dunif(p_grid)
likelihood <-  dbinom(6, 9, prob=p_grid)
posterior <-  likelihood * prior
posterior <- posterior/sum(posterior)
```



Creating a grid-approximate posterior

We can compute the posterior through grid approximation.

Data: W L W W W L W L W

Posterior: $P(\pi|d) \propto P(d|\pi)P(\pi)$

p_grid <- seq(0, 1, length.out=1000)
prior <- dunif(p_grid)
likelihood <- dbinom(6, 9, prob=p_grid)
posterior <- likelihood * prior
posterior <- posterior/sum(posterior)

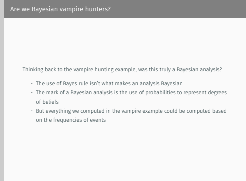
π is a rate or proportion, which is continuous it can fall anywhere between 0 and 1. We don't want to do calculus so we will approximate the posterior of π with this code. Coding means never having to do calculus (mostly).

`posterior` is a vector of probability densities. We could write our own density function using it, to look up the closest grid value and return the probability value.

Are we Bayesian vampire hunters?

Thinking back to the vampire hunting example, was this truly a Bayesian analysis?

- The use of Bayes rule isn't what makes an analysis Bayesian
- The mark of a Bayesian analysis is the use of probabilities to represent degrees of beliefs
- But everything we computed in the vampire example could be computed based on the frequencies of events


Are we Bayesian vampire hunters?

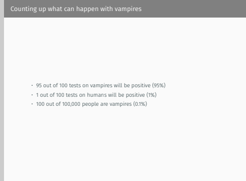
Thinking back to the vampire hunting example, was this truly a Bayesian analysis?

- The use of Bayes rule isn't what makes an analysis Bayesian
- The mark of a Bayesian analysis is the use of probabilities to represent degrees of beliefs
- But everything we computed in the vampire example could be computed based on the frequencies of events

- Bayes simultaneously gets too much credit and not enough. Not enough because Bayes theorem has become fundamental to all modern statistics, including non-bayesian approaches.
- Too much because Laplace is really the one who began developing the math to actually apply Bayes' rule in statistical inference.
- Key insight is that we can decompose the joint probability distribution in a really useful way

Counting up what can happen with vampires

- 95 out of 100 tests on vampires will be positive (95%)
- 1 out of 100 tests on humans will be positive (1%)
- 100 out of 100,000 people are vampires (0.1%)



Counting up what can happen with vampires

- 95 out of 100 tests on vampires will be positive (95%)
- 1 out of 100 tests on humans will be positive (1%)
- 100 out of 100,000 people are vampires (0.1%)

Counting up what can happen with R

```r
population <- tibble(
  species = rep(c("vampire", "human"), times = c(100, 99900)),
  test_result =  c(
    rep(c("positive", "negative"), times=c(95, 5)),
    rep(rep(c("positive", "negative"), times=c(99900*.01, 99900*.99)))
  )
)

population %>%
  filter(test_result=="positive") %>%
  count(species) %>%
  mutate(proportion = n/sum(n))
```

```
## # A tibble: 2 x 3
##   species       n proportion
##   <chr>     <int>      <dbl>
## 1 human       999     0.913
## 2 vampire      95     0.0868
```

└─Counting up what can happen with R

Counting up what can happen with R

```r
population <- tibble(
  species = rep(c("vampire", "human"), times = c(100, 99900)),
  test_result =  c(
    rep(c("positive", "negative"), times=c(95, 5)),
    rep(rep(c("positive", "negative"), times=c(99900*.01, 99900*.99)))
  )
)

population %>%
  filter(test_result=="positive") %>%
  count(species) %>%
  mutate(proportion = n/sum(n))
```

```
## # A tibble: 2 x 3
##   species       n proportion
##   <chr>     <int>      <dbl>
## 1 human       999     0.913
## 2 vampire      95     0.0868
```

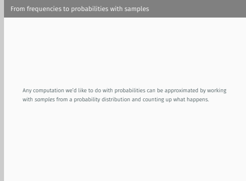
Everything we just did could be based on the frequencies of events

- 100 out of 100,000 people are vampires (0.1%)
- 95 out of 100 tests on vampires will be positive (95%)
- 1 out of 100 tests on humans will be positive (1%)

From frequencies to probabilities with samples

Any computation we’d like to do with probabilities can be approximated by working with *samples* from a probability distribution and counting up what happens.

└─From frequencies to probabilities with samples



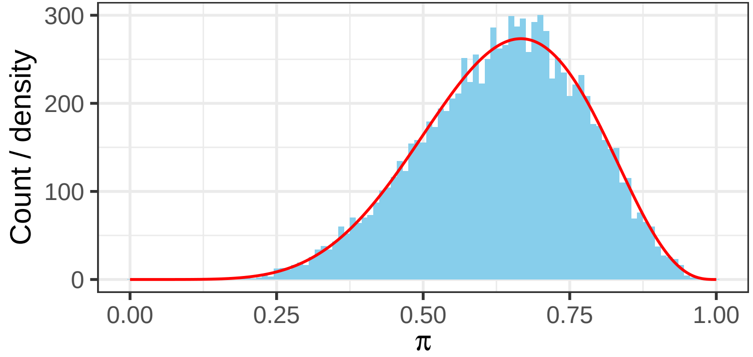
From frequencies to probabilities with samples

Any computation we’d like to do with probabilities can be approximated by working with *samples* from a probability distribution and counting up what happens.

Sampling from a grid-approximate posterior distribution

We can represent the globe tossing posterior distribution (or any distribution) by collecting samples from it.

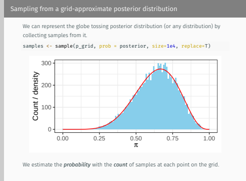
```
samples <- sample(p_grid, prob = posterior, size=1e4, replace=T)
```



We estimate the ***probability*** with the ***count*** of samples at each point on the grid.

2021-08-31

└Sampling from a grid-approximate posterior distribution


Sampling from a grid-approximate posterior distribution

We can represent the globe tossing posterior distribution (or any distribution) by collecting samples from it.

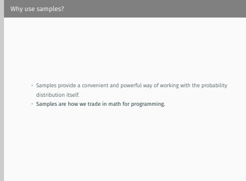
samples <- sample(p_grid, prob = posterior, size=1e4, replace=T)

We estimate the ***probability*** with the ***count*** of samples at each point on the grid.

- should unpack the sample() function more maybe, show it off with a smaller vector

Why use samples?

- Samples provide a convenient and powerful way of working with the probability distribution itself.
- **Samples are how we trade in math for programming.**



Why use samples?

- Samples provide a convenient and powerful way of working with the probability distribution itself.
- **Samples are how we trade in math for programming.**

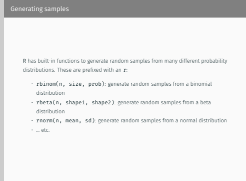
Generating samples

R has built-in functions to generate random samples from many different probability distributions. These are prefixed with an `r`:

- `rbinom(n, size, prob)`: generate random samples from a binomial distribution
- `rbeta(n, shape1, shape2)`: generate random samples from a beta distribution
- `rnorm(n, mean, sd)`: generate random samples from a normal distribution
- … etc.

2021-08-31

└─Generating samples



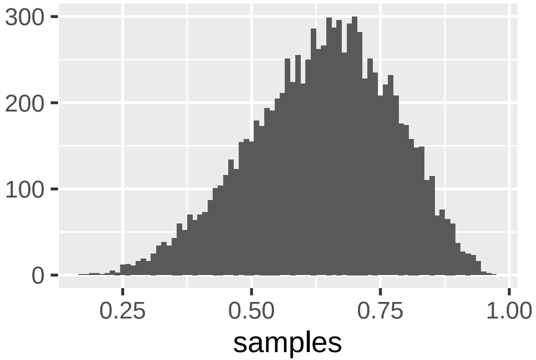
Generating samples

R has built-in functions to generate random samples from many different probability distributions. These are prefixed with an `r`:

- `rbinom(n, size, prob)`: generate random samples from a binomial distribution
- `rbeta(n, shape1, shape2)`: generate random samples from a beta distribution
- `rnorm(n, mean, sd)`: generate random samples from a normal distribution
- … etc.

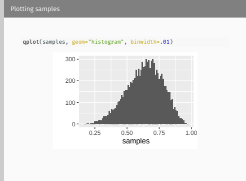
Plotting samples

```
qplot(samples, geom="histogram", binwidth=.01)
```



2021-08-31

└Plotting samples


Plotting samples

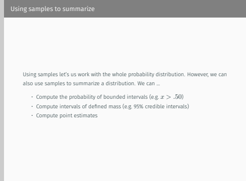
qplot(samples, geom="histogram", binwidth=.01)

- qplot interface to ggplot
- can be nice for simple plots

Using samples to summarize

Using samples let's us work with the whole probability distribution. However, we can also use samples to summarize a distribution. We can ...

- Compute the probability of bounded intervals (e.g. $x > .50$)
- Compute intervals of defined mass (e.g. 95% credible intervals)
- Compute point estimates



Using samples to summarize

Using samples let's us work with the whole probability distribution. However, we can also use samples to summarize a distribution. We can ...

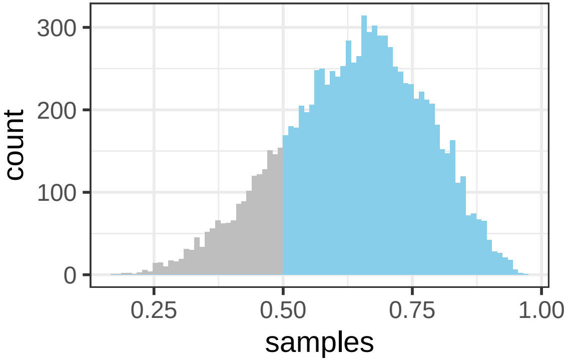
- Compute the probability of bounded intervals (e.g. $x > .50$)
- Compute intervals of defined mass (e.g. 95% credible intervals)
- Compute point estimates

Estimating bounded intervals

What is the probability that more than half the globe is water?

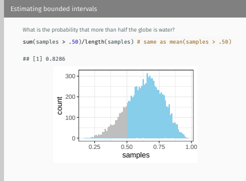
```r
sum(samples > .50)/length(samples) # same as mean(samples > .50)
```

```
## [1] 0.8286
```



2021-08-31

└─Estimating bounded intervals


Estimating bounded intervals

What is the probability that more than half the globe is water?

sum(samples > .50)/length(samples) # same as mean(samples > .50)

## [1] 0.8286

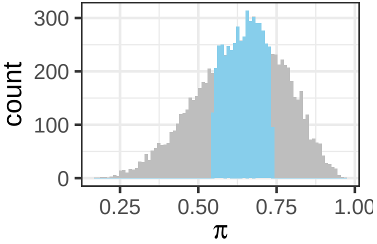
Estimating intervals of defined mass

“Credible Intervals” (CI)

Below is a 50% credible interval.

```
quantile(samples, c(.25, .75))
```

```
##       25%       75% 
## 0.5435435 0.7367367
```

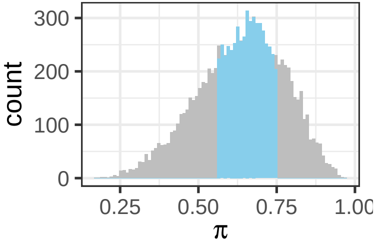


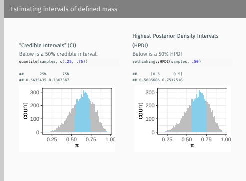
Highest Posterior Density Intervals (HPDI)

Below is a 50% HPDI

```
rethinking::HPDI(samples, .50)
```

```
##      |0.5      0.5| 
## 0.5605606 0.7517518
```





- CI = center of probability mass
- HPDI = narrowest interval containing specified probability mass
  - narrowest interval containing 56% of probability

Point estimates

Maximum A-Posteriori Estimate (MAP)

- Equivalent to posterior mode

```
Mode(samples)
```

```
## [1] 0.6736737
```

Posterior mean

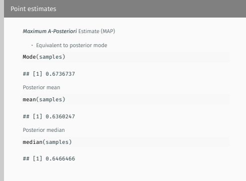
```
mean(samples)
```

```
## [1] 0.6360247
```

Posterior median

```
median(samples)
```

```
## [1] 0.6466466
```



Point estimates

Maximum A-Posteriori Estimate (MAP)

- Equivalent to posterior mode

```
Mode(samples)
```

```
## [1] 0.6736737
```

Posterior mean

```
mean(samples)
```

```
## [1] 0.6360247
```

Posterior median

```
median(samples)
```

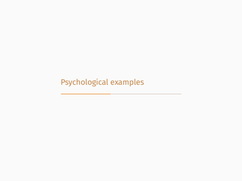
```
## [1] 0.6466466
```

I created my own `Mode()` function b/c there isn't one built in. Will provide this for you when you may need it in the future assignments.

Our posterior is a little lopsided, so you can see that the MAP is a bit higher than the median and mean.

Psychological examples

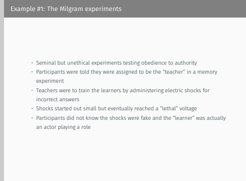
2021-08-31



Psychological examples

Example #1: The Milgram experiments

- Seminal but unethical experiments testing obedience to authority
- Participants were told they were assigned to be the “teacher” in a memory experiment
- Teachers were to train the learners by administering electric shocks for incorrect answers
- Shocks started out small but eventually reached a “lethal” voltage
- Participants did not know the shocks were fake and the “learner” was actually an actor playing a role



Example #1: The Milgram experiments

- Seminal but unethical experiments testing obedience to authority
- Participants were told they were assigned to be the “teacher” in a memory experiment
- Teachers were to train the learners by administering electric shocks for incorrect answers
- Shocks started out small but eventually reached a “lethal” voltage
- Participants did not know the shocks were fake and the “learner” was actually an actor playing a role

The Beta-Binomial model

Can estimate the proportion of people in population who comply with authority with the ***Beta-binomial*** model: a model for a rate or proportion π.

$$y \sim Binomial(n, \pi)$$

$$\pi \sim Beta(\alpha, \beta)$$

Components of the model

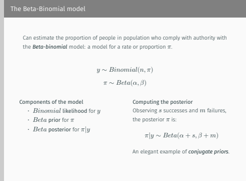
- $Binomial$ likelihood for y
- $Beta$ prior for π
- $Beta$ posterior for $\pi|y$

Computing the posterior

Observing s successes and m failures, the posterior π is:

$$\pi|y \sim Beta(\alpha + s, \beta + m)$$

An elegant example of ***conjugate priors***.



The Beta-Binomial model

Can estimate the proportion of people in population who comply with authority with the ***Beta-binomial*** model: a model for a rate or proportion π.

$$y \sim Binomial(n, \pi)$$

$$\pi \sim Beta(\alpha, \beta)$$

Components of the model

- $Binomial$ likelihood for y
- $Beta$ prior for π
- $Beta$ posterior for $\pi|y$

Computing the posterior

Observing s successes and m failures, the posterior π is:

$$\pi|y \sim Beta(\alpha + s, \beta + m)$$

An elegant example of ***conjugate priors***.

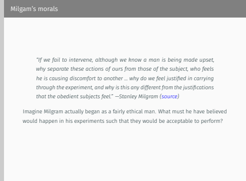
- McElreath’s “globe tossing” model is an example

Milgam’s morals

> *“If we fail to intervene, although we know a man is being made upset, why separate these actions of ours from those of the subject, who feels he is causing discomfort to another … why do we feel justified in carrying through the experiment, and why is this any different from the justifications that the obedient subjects feel.” —Stanley Milgram (source)*

Imagine Milgram actually began as a fairly ethical man. What must he have believed would happen in his experiments such that they would be acceptable to perform?

2021-08-31



Milgam’s morals

“If we fail to intervene, although we know a man is being made upset, why separate these actions of ours from those of the subject, who feels he is causing discomfort to another … why do we feel justified in carrying through the experiment, and why is this any different from the justifications that the obedient subjects feel.” —Stanley Milgram (source)

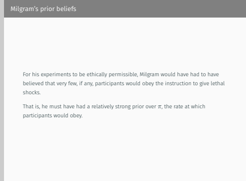
Imagine Milgram actually began as a fairly ethical man. What must he have believed would happen in his experiments such that they would be acceptable to perform?

It’s hard to know what’s in another person’s heart. We do know that Milgram would go on to conduct his study on hundreds of naive subjects without ever publicly expressing regret at the harm he caused them. However, he did privately acknowledge his precarious moral position

Milgram's prior beliefs

For his experiments to be ethically permissible, Milgram would have had to have believed that very few, if any, participants would obey the instruction to give lethal shocks.

That is, he must have had a relatively strong prior over π, the rate at which participants would obey.

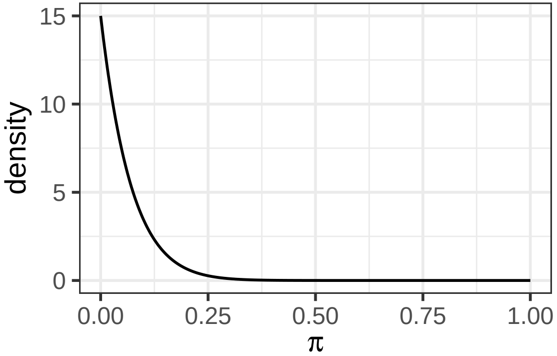

Milgram's prior beliefs

For his experiments to be ethically permissible, Milgram would have had to have believed that very few, if any, participants would obey the instruction to give lethal shocks.

That is, he must have had a relatively strong prior over π, the rate at which participants would obey.

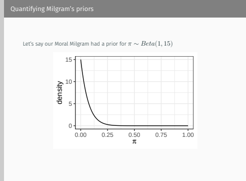
Quantifying Milgram’s priors

Let’s say our Moral Milgram had a prior for $\pi \sim Beta(1, 15)$



2021-08-31

Quantifying Milgram’s priors


Quantifying Milgram’s priors

Let’s say our Moral Milgram had a prior for $\pi \sim Beta(1, 15)$

Investigating Milgram’s priors

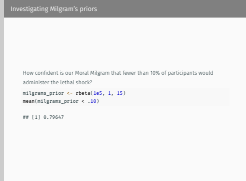
How confident is our Moral Milgram that fewer than 10% of participants would administer the lethal shock?

```
milgrams_prior <- rbeta(1e5, 1, 15)
mean(milgrams_prior < .10)
```

```
## [1] 0.79647
```

2021-08-31

Probability foundations
└─Psychological examples
└─Investigating Milgram’s priors



Investigating Milgram’s priors

How confident is our Moral Milgram that fewer than 10% of participants would administer the lethal shock?

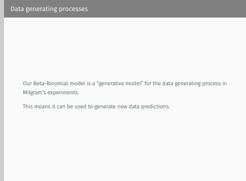
```
milgrams_prior <- rbeta(1e5, 1, 15)
mean(milgrams_prior < .10)
```

```
## [1] 0.79647
```

Data generating processes

Our Beta-Binomial model is a “generative model” for the data generating process in Milgram’s experiments.

This means it can be used to generate new data predictions.



Data generating processes

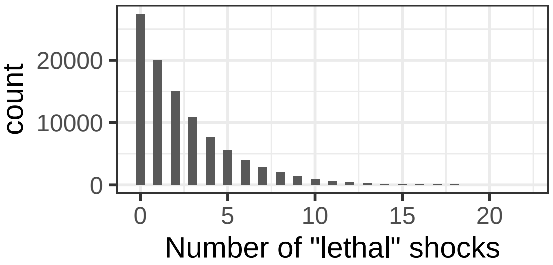
Our Beta-Binomial model is a “generative model” for the data generating process in Milgram’s experiments.

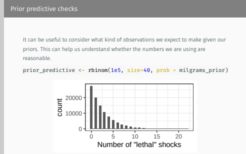
This means it can be used to generate new data predictions.

Prior predictive checks

It can be useful to consider what kind of observations we expect to make given our priors. This can help us understand whether the numbers we are using are reasonable.

```
prior_predictive <- rbinom(1e5, size=40, prob = milgrams_prior)
```




Prior predictive checks

It can be useful to consider what kind of observations we expect to make given our priors. This can help us understand whether the numbers we are using are reasonable.

prior_predictive <- rbinom(1e5, size=40, prob = milgrams_prior)

- Generated samples from the prior predictive
- we pass Milgram’s prior on π into `rbinom()` to sample from a binomial with those probabilities
- now beta and binomial work together nicely so there are no surprises here, the shape is very similar to the beta prior
- Tho maybe our milgram isn’ so moral and a stronger prior would really be needed to justify his decisions

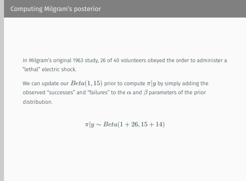
Computing Milgram's posterior

In Milgram's original 1963 study, 26 of 40 volunteers obeyed the order to administer a "lethal" electric shock.

We can update our $Beta(1, 15)$ prior to compute $\pi|y$ by simply adding the observed "successes" and "failures" to the α and β parameters of the prior distribution.

$$\pi|y \sim Beta(1 + 26, 15 + 14)$$

2021-08-31



Computing Milgram's posterior

In Milgram's original 1963 study, 26 of 40 volunteers obeyed the order to administer a "lethal" electric shock.

We can update our $Beta(1, 15)$ prior to compute $\pi|y$ by simply adding the observed "successes" and "failures" to the α and β parameters of the prior distribution.

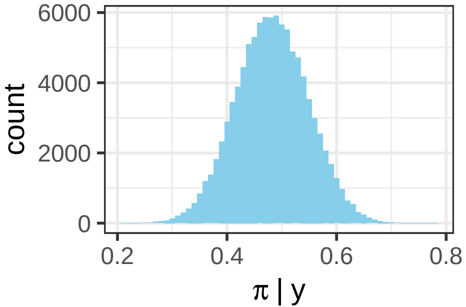
$$\pi|y \sim Beta(1 + 26, 15 + 14)$$

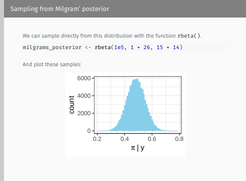
Sampling from Milgram’ posterior

We can sample directly from this distribution with the function `rbeta()`.

```r
milgrams_posterior <- rbeta(1e5, 1 + 26, 15 + 14)
```

And plot these samples:




Sampling from Milgram’ posterior

We can sample directly from this distribution with the function rbeta().

milgrams_posterior <- rbeta(1e5, 1 + 26, 15 + 14)

And plot these samples:

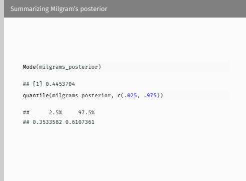
Summarizing Milgram’s posterior

```
Mode(milgrams_posterior)
```

```
## [1] 0.4453704
```

```
quantile(milgrams_posterior, c(.025, .975))
```

```
##      2.5%     97.5% 
## 0.3533582 0.6107361
```



Summarizing Milgram’s posterior

```
Mode(milgrams_posterior)
```

```
## [1] 0.4453704
```

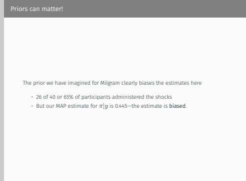
```
quantile(milgrams_posterior, c(.025, .975))
```

```
##      2.5%     97.5% 
## 0.3533582 0.6107361
```

Priors can matter!

The prior we have imagined for Milgram clearly biases the estimates here

- 26 of 40 or 65% of participants administered the shocks
- But our MAP estimate for $\pi|y$ is 0.445—the estimate is **biased**.



Priors can matter!

The prior we have imagined for Milgram clearly biases the estimates here

- 26 of 40 or 65% of participants administered the shocks
- But our MAP estimate for $\pi|y$ is 0.445—the estimate is **biased**.

But likelihoods often matter more

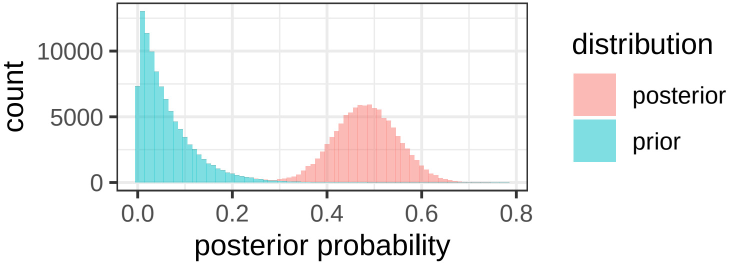
- With enough data, the likelihood comes to “dominate” the prior
- After 40 observations the posterior has shifted a great deal despite our strong starting prior

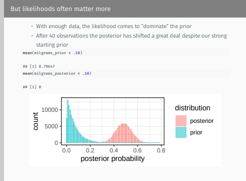
```
mean(milgrams_prior < .10)
```

```
## [1] 0.79647
```

```
mean(milgrams_posterior < .10)
```

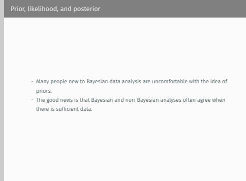
```
## [1] 0
```




But likelihoods often matter more
- With enough data, the likelihood comes to “dominate” the prior
- After 40 observations the posterior has shifted a great deal despite our strong starting prior
mean(milgrams_prior < .10)
## [1] 0.79647
mean(milgrams_posterior < .10)
## [1] 0

Prior, likelihood, and posterior

- Many people new to Bayesian data analysis are uncomfortable with the idea of priors.
- The good news is that Bayesian and non-Bayesian analyses often agree when there is sufficient data.



Prior, likelihood, and posterior

- Many people new to Bayesian data analysis are uncomfortable with the idea of priors.
- The good news is that Bayesian and non-Bayesian analyses often agree when there is sufficient data.

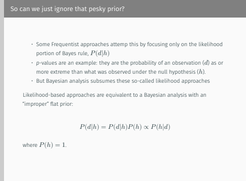
So can we just ignore that pesky prior?

- Some Frequentist approaches attemp this by focusing only on the likelihood portion of Bayes rule, $P(d|h)$
- *p*-values are an example: they are the probability of an observation (d) as or more extreme than what was observed under the null hypothesis (h).
- But Bayesian analysis subsumes these so-called likelihood approaches

Likelihood-based approaches are equivalent to a Bayesian analysis with an “improper” flat prior:

$$P(d|h) = P(d|h)P(h) \propto P(h|d)$$

where $P(h) = 1$.



So can we just ignore that pesky prior?

- Some Frequentist approaches attemp this by focusing only on the likelihood portion of Bayes rule, $P(d|h)$
- *p*-values are an example: they are the probability of an observation (d) as or more extreme than what was observed under the null hypothesis (h).
- But Bayesian analysis subsumes these so-called likelihood approaches

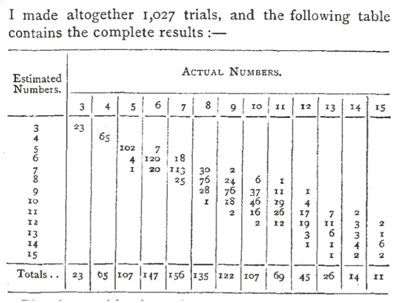
Likelihood-based approaches are equivalent to a Bayesian analysis with an “improper” flat prior:

$$P(d|h) = P(d|h)P(h) \propto P(h|d)$$

where $P(h) = 1$.

Example #2: The power of numerical discrimination

- Jevons (1871) took a handful of black beans and tossed them at a white shallow container surrounded by black cloth
- On each toss, a certain number of beans would land in the square and he would then look and try to immediately call out their number
- Then he counted how many beans there actually were in the container, and recorded the result.

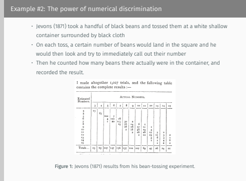


I made altogether 1,027 trials, and the following table contains the complete results :—

Estimated Numbers.	ACTUAL NUMBERS.												
	3	4	5	6	7	8	9	10	11	12	13	14	15
3	23												
4		65											
5			102	7									
6			4	120	18								
7			1	20	113	30	2						
8					25	76	24	6	1				
9						28	76	37	11	1			
10						1	18	46	19	4			
11							2	16	26	17	7	2	
12								2	12	19	11	3	2
13										3	6	3	1
14										1	1	4	6
15											1	2	2
Totals . .	23	65	107	147	156	135	122	107	69	45	26	14	11

Figure 1: Jevons (1871) results from his bean-tossing experiment.

Example #2: The power of numerical discrimination



Example #2: The power of numerical discrimination
- Jevons (1871) took a handful of black beans and tossed them at a white shallow container surrounded by black cloth
- On each toss, a certain number of beans would land in the square and he would then look and try to immediately call out their number
- Then he counted how many beans there actually were in the container, and recorded the result.
Figure 1: Jevons (1871) results from his bean-tossing experiment.

He made 1027 trials, so he did this basically until he got bored.

A bean-tossing model

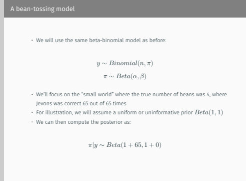
- We will use the same beta-binomial model as before:

$$y \sim Binomial(n, \pi)$$

$$\pi \sim Beta(\alpha, \beta)$$

- We'll focus on the "small world" where the true number of beans was 4, where Jevons was correct 65 out of 65 times
- For illustration, we will assume a uniform or uninformative prior $Beta(1, 1)$
- We can then compute the posterior as:

$$\pi | y \sim Beta(1 + 65, 1 + 0)$$



A bean-tossing model

- We will use the same beta-binomial model as before:

$$y \sim Binomial(n, \pi)$$

$$\pi \sim Beta(\alpha, \beta)$$

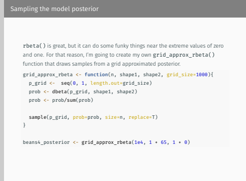
- We'll focus on the "small world" where the true number of beans was 4, where Jevons was correct 65 out of 65 times
- For illustration, we will assume a uniform or uninformative prior $Beta(1, 1)$
- We can then compute the posterior as:

$$\pi | y \sim Beta(1 + 65, 1 + 0)$$

Sampling the model posterior

`rbeta()` is great, but it can do some funky things near the extreme values of zero and one. For that reason, I'm going to create my own `grid_approx_rbeta()` function that draws samples from a grid approximated posterior.

```
grid_approx_rbeta <- function(n, shape1, shape2, grid_size=1000){
  p_grid <-  seq(0, 1, length.out=grid_size)
  prob <- dbeta(p_grid, shape1, shape2)
  prob <- prob/sum(prob)

  sample(p_grid, prob=prob, size=n, replace=T)
}

beans4_posterior <- grid_approx_rbeta(1e4, 1 + 65, 1 + 0)
```



Sampling the model posterior

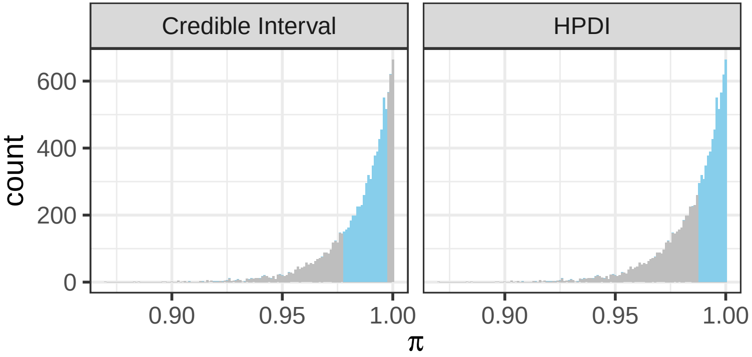
rbeta() is great, but it can do some funky things near the extreme values of zero and one. For that reason, I'm going to create my own grid_approx_rbeta() function that draws samples from a grid approximated posterior.

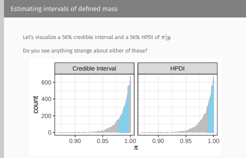
```
grid_approx_rbeta <- function(n, shape1, shape2, grid_size=1000){
  p_grid <-  seq(0, 1, length.out=grid_size)
  prob <- dbeta(p_grid, shape1, shape2)
  prob <- prob/sum(prob)

  sample(p_grid, prob=prob, size=n, replace=T)
}

beans4_posterior <- grid_approx_rbeta(1e4, 1 + 65, 1 + 0)
```

Estimating intervals of defined mass

Let’s visualize a 56% credible interval and a 56% HPDI of $\pi|y$.

Do you see anything strange about either of these?




Estimating intervals of defined mass

Let’s visualize a 56% credible interval and a 56% HPDI of $\pi|y$.

Do you see anything strange about either of these?

- McElreath likes to use 89 and other “odd” intervals as examples, to illustrate the arbitrariness of the standard 95 confidence interval. So this is giving the man his due.
- strangeness is that 95 Cred Interval does not contain MAP value
- Generally, if the choice of interval matters then you shouldn’t use intervals!

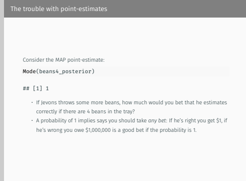
The trouble with point-estimates

Consider the MAP point-estimate:

```
Mode(beans4_posterior)
```

```
## [1] 1
```

- If Jevons throws some more beans, how much would you bet that he estimates correctly if there are 4 beans in the tray?
- A probability of 1 implies says you should take *any bet*: If he’s right you get $1, if he’s wrong you owe $1,000,000 is a good bet if the probability is 1.



The trouble with point-estimates

Consider the MAP point-estimate:

Mode(beans4_posterior)

## [1] 1

- If Jevons throws some more beans, how much would you bet that he estimates correctly if there are 4 beans in the tray?
- A probability of 1 implies says you should take *any bet*: If he’s right you get $1, if he’s wrong you owe $1,000,000 is a good bet if the probability is 1.

As before we can create point estimates to summarize the posterior distribution using our samples.

- One way of thinking about probability is in terms of the gambles you’d be willing to make
- Hopefully it’s obvious that that is not really a good bet, and that the MAP is a poor summary point-estimate here.

Choosing point estimates

In this case, the posterior mean or median make for more reasonable summaries:

```
mean(beans4_posterior)
```

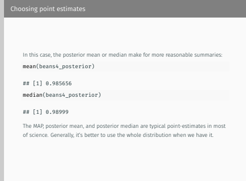
```
## [1] 0.985656
```

```
median(beans4_posterior)
```

```
## [1] 0.98999
```

The MAP, posterior mean, and posterior median are typical point-estimates in most of science. Generally, it's better to use the whole distribution when we have it.

2021-08-31



Choosing point estimates

In this case, the posterior mean or median make for more reasonable summaries:

```
mean(beans4_posterior)
## [1] 0.985656
median(beans4_posterior)
## [1] 0.98999
```

The MAP, posterior mean, and posterior median are typical point-estimates in most of science. Generally, it's better to use the whole distribution when we have it.

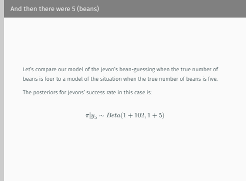
if you're in a situation where the choice of point estimate matters, try not to use point estimates!

And then there were 5 (beans)

Let’s compare our model of the Jevon’s bean-guessing when the true number of beans is four to a model of the situation when the true number of beans is five.

The posteriors for Jevons’ success rate in this case is:

$$\pi | y_5 \sim Beta(1 + 102, 1 + 5)$$



And then there were 5 (beans)

Let’s compare our model of the Jevon’s bean-guessing when the true number of beans is four to a model of the situation when the true number of beans is five.

The posteriors for Jevons’ success rate in this case is:

$$\pi | y_5 \sim Beta(1 + 102, 1 + 5)$$

Comparing posteriors with summaries

Four beans

```
mean(beans4_posterior)
```

```
## [1] 0.985656
```

```
quantile(beans4_posterior, c(.025, .975))
```

```
##      2.5%     97.5%
## 0.9459459 1.0000000
```

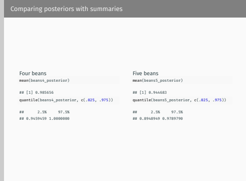
Five beans

```
mean(beans5_posterior)
```

```
## [1] 0.944683
```

```
quantile(beans5_posterior, c(.025, .975))
```

```
##      2.5%     97.5%
## 0.8948949 0.9789790
```



Comparing posteriors with summaries

Four beans

mean(beans4_posterior)

## [1] 0.985656

quantile(beans4_posterior, c(.025, .975))

##   2.5%   97.5%
## 0.9459459 1.0000000

Five beans

mean(beans5_posterior)

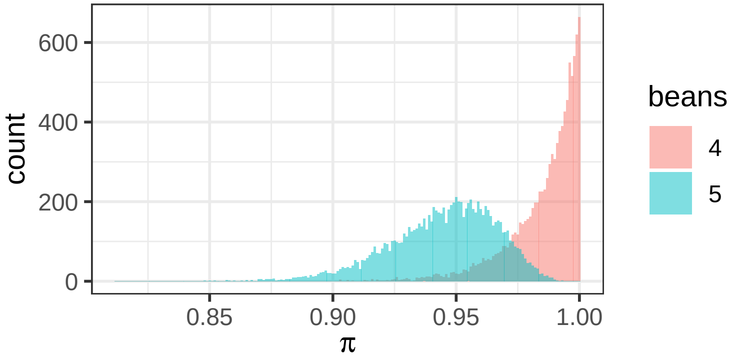
## [1] 0.944683

quantile(beans5_posterior, c(.025, .975))

##   2.5%   97.5%
## 0.8948949 0.9789790

- what can you conclude from this? (not much really)
- we can kind of see the rate is lower with 5, but we’re feeling like we need to do some kind of statistical test to know if they are really different

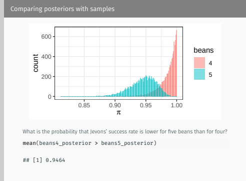
Comparing posteriors with samples



What is the probability that Jevons’ success rate is lower for five beans than for four?

```
mean(beans4_posterior > beans5_posterior)
```

```
## [1] 0.9464
```


Comparing posteriors with samples

What is the probability that Jevons’ success rate is lower for five beans than for four?

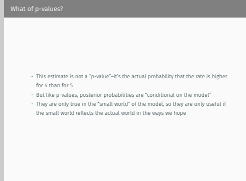
mean(beans4_posterior > beans5_posterior)

## [1] 0.9464

- It’s very easy to compare things directly with posterior samples. We can directly compute the probability we want.
- This is not a “p-value”—it’s the actual probability that the rate is higher for 4 than for 5 (assuming our small world model is trustworthy)
- probabilities we compute are conditional on the model

What of p-values?

- This estimate is not a “p-value”–it’s the actual probability that the rate is higher for 4 than for 5
- But like p-values, posterior probabilities are “conditional on the model”
- They are only true in the “small world” of the model, so they are only useful if the small world reflects the actual world in the ways we hope



What of p-values?

- This estimate is not a “p-value”–it’s the actual probability that the rate is higher for 4 than for 5
- But like p-values, posterior probabilities are “conditional on the model”
- They are only true in the “small world” of the model, so they are only useful if the small world reflects the actual world in the ways we hope

Computing new quantities with samples

What is the difference in Jevon's success rates are when there are four beans compared with when there are five? i.e. what is the distribution of $\pi|y_4 - \pi|y_5$?

```
success_diff <- beans4_posterior - beans5_posterior
Mode(success_diff)
```

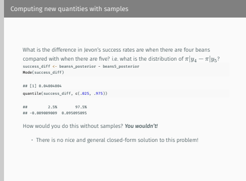
```
## [1] 0.04004004
```

```
quantile(success_diff, c(.025, .975))
```

```
##         2.5%        97.5%
## -0.009009009  0.095095095
```

How would you do this without samples? ***You wouldn't!***

- There is no nice and general closed-form solution to this problem!



Computing new quantities with samples

What is the difference in Jevon's success rates are when there are four beans compared with when there are five? i.e. what is the distribution of $\pi|y_4 - \pi|y_5$?

```
success_diff <- beans4_posterior - beans5_posterior
Mode(success_diff)
```

```
## [1] 0.04004004
```

```
quantile(success_diff, c(.025, .975))
```

```
##         2.5%        97.5%
## -0.009009009  0.095095095
```

How would you do this without samples? ***You wouldn't!***

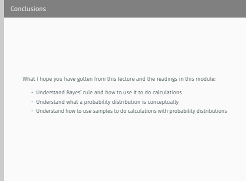
- There is no nice and general closed-form solution to this problem!

Conclusions

What I hope you have gotten from this lecture and the readings in this module:

- Understand Bayes’ rule and how to use it to do calculations
- Understand what a probability distribution is conceptually
- Understand how to use samples to do calculations with probability distributions

2021-08-31



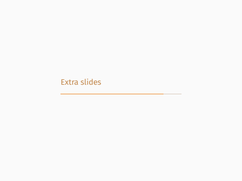
Conclusions

What I hope you have gotten from this lecture and the readings in this module:

- Understand Bayes’ rule and how to use it to do calculations
- Understand what a probability distribution is conceptually
- Understand how to use samples to do calculations with probability distributions

Extra slides

2021-08-31



Extra slides

Mode() function definition

Don’t sweat what is happening in here, but given a numerical vector it returns the mode (most frequent value).

```
Mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}
```

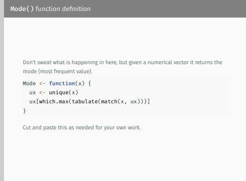
Cut and paste this as needed for your own work.

2021-08-31

Probability foundations

└─Extra slides

└─Mode() function definition



Mode() function definition

Don’t sweat what is happening in here, but given a numerical vector it returns the mode (most frequent value).

```
Mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}
```

Cut and paste this as needed for your own work.

Inspecting the prior predictive

1. If Milgram enrolls 40 participants in his study, what is the single most probable number of people administering the lethal shock according to Milgram's priors?

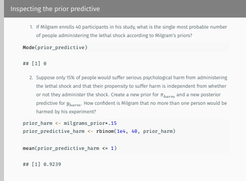
```
Mode(prior_predictive)
```

```
## [1] 0
```

2. Suppose only 15% of people would suffer serious psychological harm from administering the lethal shock and that their propensity to suffer harm is independent from whether or not they administer the shock. Create a new prior for π_{harm} and a new posterior predictive for y_{harm}. How confident is Milgram that no more than one person would be harmed by his experiment?

```
prior_harm <- milgrams_prior*.15
prior_predictive_harm <- rbinom(1e4, 40, prior_harm)

mean(prior_predictive_harm <= 1)
```

```
## [1] 0.9239
```



Inspecting the prior predictive

1. If Milgram enrolls 40 participants in his study, what is the single most probable number of people administering the lethal shock according to Milgram's priors?

Mode(prior_predictive)

## [1] 0

2. Suppose only 15% of people would suffer serious psychological harm from administering the lethal shock and that their propensity to suffer harm is independent from whether or not they administer the shock. Create a new prior for π_{harm} and a new posterior predictive for y_{harm}. How confident is Milgram that no more than one person would be harmed by his experiment?

prior_harm <- milgrams_prior*.15
prior_predictive_harm <- rbinom(1e4, 40, prior_harm)

mean(prior_predictive_harm <= 1)

## [1] 0.9239

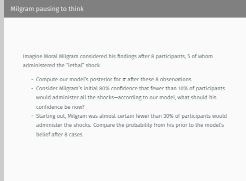
```
Mode(prior_predictive)
prior_harm <- milgrams_prior*.15
prior_predictive_harm <- rbinom(1e4, 40, prior_harm)
mean(prior_predictive_harm <= 1)
```

Milgram pausing to think

Imagine Moral Milgram considered his findings after 8 participants, 5 of whom administered the “lethal” shock.

- Compute our model’s posterior for π after these 8 observations.
- Consider Milgram’s initial 80% confidence that fewer than 10% of participants would administer all the shocks—according to our model, what should his confidence be now?
- Starting out, Milgram was almost certain fewer than 30% of participants would administer the shocks. Compare the probability from his prior to the model’s belief after 8 cases.

2021-08-31



Milgram pausing to think

Imagine Moral Milgram considered his findings after 8 participants, 5 of whom administered the “lethal” shock.

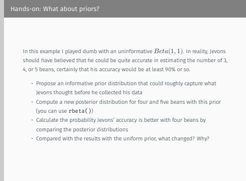
- Compute our model’s posterior for π after these 8 observations.
- Consider Milgram’s initial 80% confidence that fewer than 10% of participants would administer all the shocks—according to our model, what should his confidence be now?
- Starting out, Milgram was almost certain fewer than 30% of participants would administer the shocks. Compare the probability from his prior to the model’s belief after 8 cases.

A Moral Milgram would stop and think about his experiment as he conducted it.

Hands-on: What about priors?

In this example I played dumb with an uninformative $Beta(1, 1)$. In reality, Jevons should have believed that he could be quite accurate in estimating the number of 3, 4, or 5 beans, certainly that his accuracy would be at least 90% or so.

- Propose an informative prior distribution that could roughly capture what Jevons thought before he collected his data
- Compute a new posterior distribution for four and five beans with this prior (you can use `rbeta()`)
- Calculate the probability Jevons' accuracy is better with four beans by comparing the posterior distributions
- Compared with the results with the uniform prior, what changed? Why?



Hands-on: What about priors?

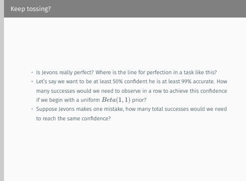
In this example I played dumb with an uninformative $Beta(1, 1)$. In reality, Jevons should have believed that he could be quite accurate in estimating the number of 3, 4, or 5 beans, certainly that his accuracy would be at least 90% or so.

- Propose an informative prior distribution that could roughly capture what Jevons thought before he collected his data
- Compute a new posterior distribution for four and five beans with this prior (you can use `rbeta()`)
- Calculate the probability Jevons' accuracy is better with four beans by comparing the posterior distributions
- Compared with the results with the uniform prior, what changed? Why?

a good prior would be something like Beta(50,1) the posterior of 4beans gets shifted more than posterior of 5beans and both distributions get more narrow, so we get a bit more confident (tho not that different)

Keep tossing?

- Is Jevons really perfect? Where is the line for perfection in a task like this?
- Let's say we want to be at least 50% confident he is at least 99% accurate. How many successes would we need to observe in a row to achieve this confidence if we begin with a uniform $Beta(1,1)$ prior?
- Suppose Jevons makes one mistake, how many total successes would we need to reach the same confidence?



Keep tossing?

- Is Jevons really perfect? Where is the line for perfection in a task like this?
- Let's say we want to be at least 50% confident he is at least 99% accurate. How many successes would we need to observe in a row to achieve this confidence if we begin with a uniform $Beta(1,1)$ prior?
- Suppose Jevons makes one mistake, how many total successes would we need to reach the same confidence?

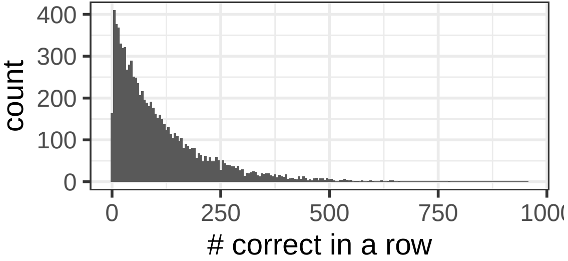
- this may sound like p-hacking! but in the Bayesian framework this kind of thing is fine (for this kind of question)

70 is a good answer with no failures 167 is a good answer with 1 failure

Simulating other kinds of outcomes

- If Jevons makes one mistake on trial 65, he would need 102 correct in a row to reach 50% confidence he is at least 99% accurate.
- The code below simulates 1,000,000 bean-tossing trials assuming that Jevons is actually 99.1% accurate.

```
x <- rbernoulli(1e6, .991)
runs <- rle(x)
run_lengths <- runs$lengths[runs$values==TRUE]
```

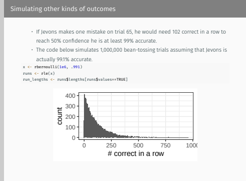


Probability foundations

Extra slides

2021-08-31

Simulating other kinds of outcomes


Simulating other kinds of outcomes

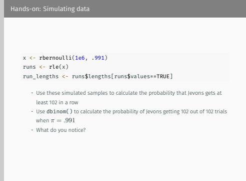
- If Jevons makes one mistake on trial 65, he would need 102 correct in a row to reach 50% confidence he is at least 99% accurate.
- The code below simulates 1,000,000 bean-tossing trials assuming that Jevons is actually 99.1% accurate.

```
x <- rbernoulli(1e6, .991)
runs <- rle(x)
run_lengths <- runs$lengths[runs$values==TRUE]
```

Hands-on: Simulating data

```
x <- rbernoulli(1e6, .991)
runs <- rle(x)
run_lengths <- runs$lengths[runs$values==TRUE]
```

- Use these simulated samples to calculate the probability that Jevons gets at least 102 in a row
- Use `dbinom()` to calculate the probability of Jevons getting 102 out of 102 trials when $\pi = .991$
- What do you notice?



Hands-on: Simulating data

```
x <- rbernoulli(1e6, .991)
runs <- rle(x)
run_lengths <- runs$lengths[runs$values==TRUE]
```

- Use these simulated samples to calculate the probability that Jevons gets at least 102 in a row
- Use `dbinom()` to calculate the probability of Jevons getting 102 out of 102 trials when $\pi = .991$
- What do you notice?

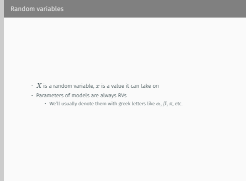
```
mean(run_lengths >= 102)
dbinom(102, 102, .991)
```

- This is just to show you that all is right with the universe when we use **probabilities**

Random variables

- X is a random variable, x is a value it can take on
- Parameters of models are always RVs
    - We'll usually denote them with greek letters like α, β, π, etc.

2021-08-31

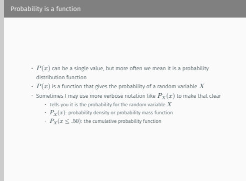


Random variables

- X is a random variable, x is a value it can take on
- Parameters of models are always RVs
    - We'll usually denote them with greek letters like α, β, π, etc.

Probability is a function

- $P(x)$ can be a single value, but more often we mean it is a probability distribution function
- $P(x)$ is a function that gives the probability of a random variable X
- Sometimes I may use more verbose notation like $P_X(x)$ to make that clear
    - Tells you it is the probability for the random variable X
    - $P_X(x)$: probability density or probability mass function
    - $P_X(x \leq .50)$: the cumulative probability function



Probability is a function

- $P(x)$ can be a single value, but more often we mean it is a probability distribution function
- $P(x)$ is a function that gives the probability of a random variable X
- Sometimes I may use more verbose notation like $P_X(x)$ to make that clear
    - Tells you it is the probability for the random variable X
    - $P_X(x)$: probability density or probability mass function
    - $P_X(x \leq .50)$: the cumulative probability function

hate focusing too much on notation but we may need some. Generally I will try to make things go down as smooth as possible without writing down anything truly incorrect.