04 - Binomial distribution

HCI/PSYCH 522
Iowa State University

January 27, 2022

Overview

- Random variables
  - Bernoulli distribution
    - Model for success/failure
  - Binomial distribution
    - Model for success/failure counts
- Inference for success/failure counts
  - Estimating 1 probability of success
  - Comparing 2 probabilities of success
  - Comparing 3+ probabilities of success

Random variables

Suppose you will run a study (any data collection) and you will have some outcome. A random variable is any numerical summary of the outcome of that study.

We may know the following quantities for random variables:

- Distribution:
  - Image, i.e. the possible values for X
  - For discrete random variables, probability mass function (pmf) $P(X = x)$.
  - Cumulative distribution function (cdf), $P(X \leq x)$.
- Expectation (average value), $E[X]$
- Variance (variability), $Var[X]$
- Standard deviation (variability), $\sqrt{Var[X]}$

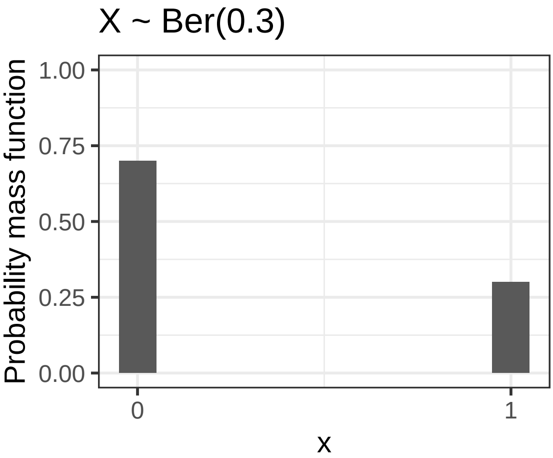
Bernoulli

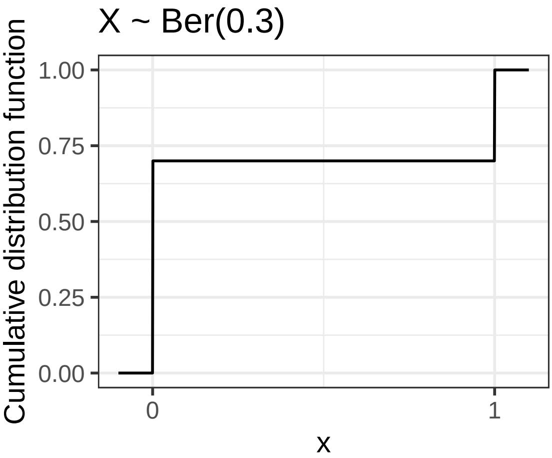
Suppose we are interested in recording the success or failure. By convention, we code 1 as a success and 0 as a failure and call this value X.

If $X \sim Ber(p)$, then X is a Bernoulli random variable with probability of success p and

- $P(X = 1) = p$,
- $P(X = 0) = (1 - p)$,
- $E[X] = p$,
- $Var[X] = p(1 - p)$, and
- $SD[X] = \sqrt{p(1 - p)}$.

Bernoulli


X ~ Ber(0.3)
Probability mass function
1.00
0.75
0.50
0.25
0.00
0
1
x


X ~ Ber(0.3)
Cumulative distribution function
1.00
0.75
0.50
0.25
0.00
0
1
x

6-sided die example

Let X be an indicator that a 1 was rolled on a 6-sided die. More formally

$$X = \begin{cases} 1 & \text{if a 1 is rolled} \\ 0 & \text{if anything else is rolled.} \end{cases}$$

Then we write $X \sim Ber(1/6)$ and know

- $P(X = 1) = 1/6$,
- $P(X = 0) = 1 - 1/6 = 5/6$,
- $E[X] = 1/6$,
- $Var[X] = 1/6 \times (1 - 1/6) = 1/6 \times 5/6 = 5/36$, and
- $SD[X] = \sqrt{5/36} = \sqrt{5}/6$.

Binomial

Suppose we count the number of successes in n attempts with a common probability of success p where each attempt is independent and call this count Y.

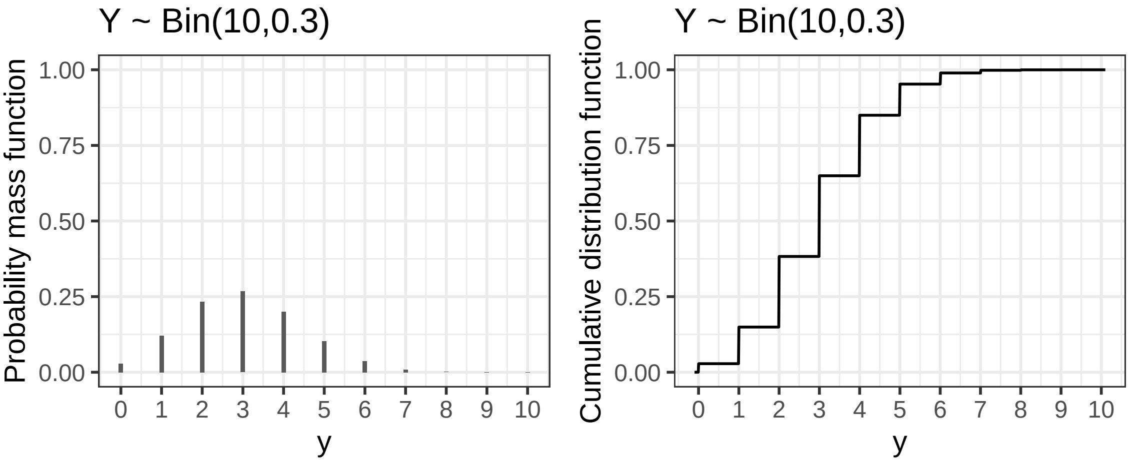
If $Y \sim Bin(n, p)$, then Y is a binomial random variable with n attempts and probability of success p and

- $P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}$ for $y = 0, 1, \ldots, n$
- $E[Y] = np$
- $Var[Y] = np(1-p)$

We can use R to calculate the probability mass function values, e.g. if $Y \sim Bin(10, 1/6)$ and we want to calculate $P(Y = 2)$ we use

```
n <- 10; p <- 1/6; y <- 2
dbinom(y, size = n, prob = p)

## [1] 0.29071
```

Binomial


Y ~ Bin(10,0.3)
Probability mass function
1.00
0.75
0.50
0.25
0.00
0 1 2 3 4 5 6 7 8 9 10
y
Y ~ Bin(10,0.3)
Cumulative distribution function
1.00
0.75
0.50
0.25
0.00
0 1 2 3 4 5 6 7 8 9 10
y

6-sided die example

Suppose we roll a 6-sided die 10 times and record the number of times we observed a 1. Assume independence between our roles, we have $Y \sim Bin(10, 1/6)$ and we know

- $E[Y] = 10 \times 1/6 = 10/6$,
- $Var[Y] = 10 \times 1/6 \times (1 - 1/6) = 10/6 \times 5/6 = 50/36$, and
- $SD[Y] = \sqrt{10 * 5/36} = \sqrt{50}/6$.

Unknown probability

Suppose you run a study where

- you have n attempts,
- each trial is independent,
- each trial has probability of success θ,

and you are interested in θ.

Let Y be the number of success observed in n attempts and assume $Y \sim Bin(n, \theta)$. A common point estimate is

$$\hat{\theta} = y/n$$

where y is the observed number of successes.

Examples

Suppose you run a study to see how many students correctly register for class using the new Workday system. Since the probability of success might differ depending on what classes need to be registered, you give each student the same list of classes.

- You randomly sample 10 ISU undergraduate students at 8 are successful. Our estimate of the probability of success is $\hat{\theta} = 8/10 = 0.8$.
- You randomly sample 100 ISU undergraduate students at 80 are successful. Our estimate of the probability of success is $\hat{\theta} = 80/100 = 0.8$.
- You randomly sample 1000 ISU undergraduate students at 800 are successful. Our estimate of the probability of success is $\hat{\theta} = 800/1000 = 0.8$.

Although the point estimate is the same, clearly we should have more certainty about the last estimate compared to the first. We need some way to quantify our uncertainty about the true value θ.

Bayesian estimation

Bayes' Rule

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \propto p(y|\theta)p(\theta)$$

where

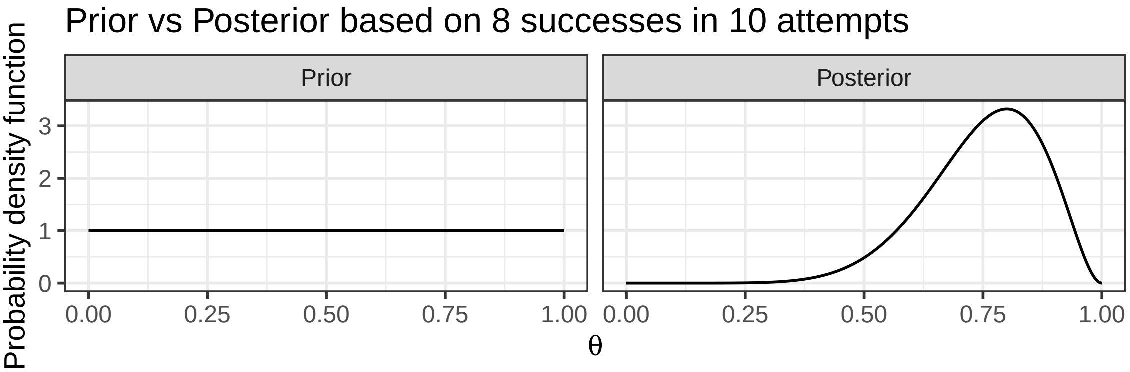
- y is our data,
- θ are our unknowns, e.g. probability of success,
- $p(y|\theta)$ comes from our model, e.g. binomial, (sometimes referred to as the likelihood),
- $p(\theta)$ is our prior belief, and
- $p(\theta|y)$ is our posterior belief.

Thus Bayesian estimation provides a mathematical mechanism to learn about the world using data, e.g.

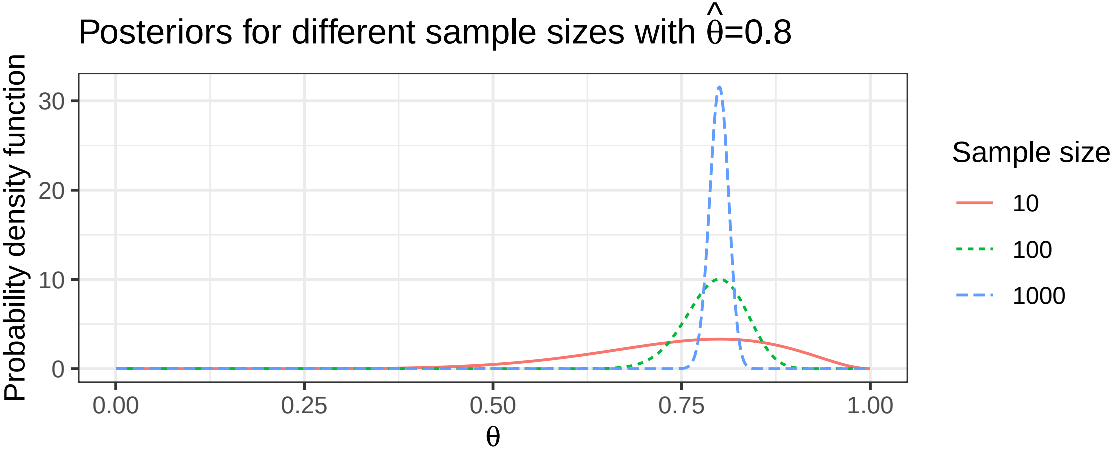
$$p(\theta) \longrightarrow p(\theta|y).$$

Bayesian estimation for probability of success

If we know nothing about our probability of success θ, our prior belief is reasonably represented by a uniform distribution between 0 and 1, i.e. $\theta \sim Unif(0,1)$. When we obtain data y, then our posterior belief is represented by a Beta distribution, i.e. $\theta|y \sim Be(1+y, 1+n-y)$.

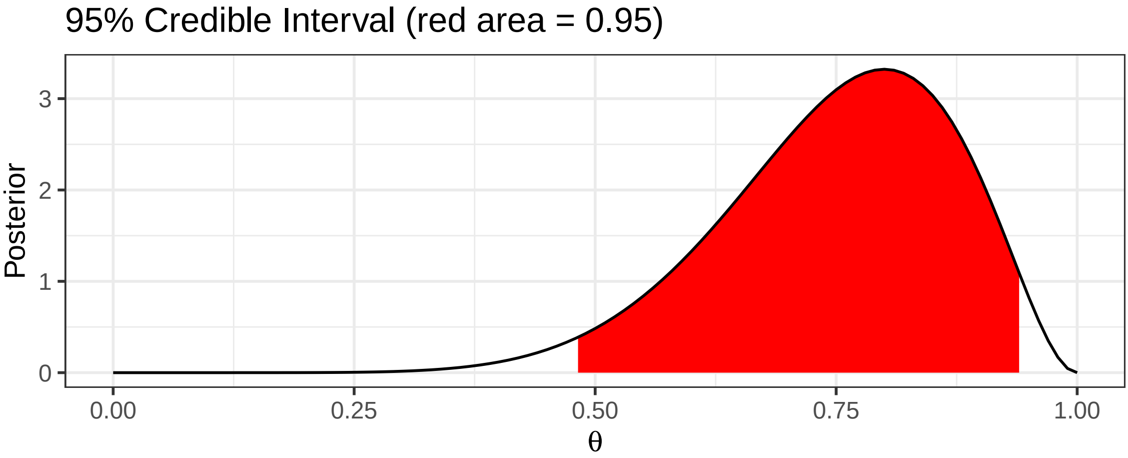


Comparison of posteriors



Credible intervals

A 95% credible interval for θ is the interval such that the area under the posterior is 0.95.



95% Credible Intervals in R

```
a <- 1 - 0.95 # for 95\% CIs

y <- 8; n <- 10
qbeta(c(a/2, 1-a/2), shape1 = 1+y, shape2 = 1+n-y)

## [1] 0.4822441 0.9397823

y <- 80; n <- 100
qbeta(c(a/2, 1-a/2), shape1 = 1+y, shape2 = 1+n-y) %>% round(2)

## [1] 0.71 0.87

y <- 800; n <- 1000
qbeta(c(a/2, 1-a/2), shape1 = 1+y, shape2 = 1+n-y) %>% round(2)

## [1] 0.77 0.82
```

Multiple probabilities

If we are collecting success/failure data under multiple conditions, then we can estimate multiple probabilities.

Let Y_i be the success count in condition i out of n_i attempts for conditions $i = 1, \ldots, \mathrm{I}$. If we assume

- all observations are independent and
- the probability of success within a condition is constant,

then our model is

$$Y_i \stackrel{ind}{\sim} Bin(n_i, \theta_i).$$

If we assume ignorance about θ_i, then we have

$$\text{Prior: } \theta_i \stackrel{ind}{\sim} Unif(0,1) \qquad \longrightarrow \qquad \text{Posterior: } \theta_i|y_i \stackrel{ind}{\sim} Be(1+y_i, 1+n_i-y_i).$$

Example

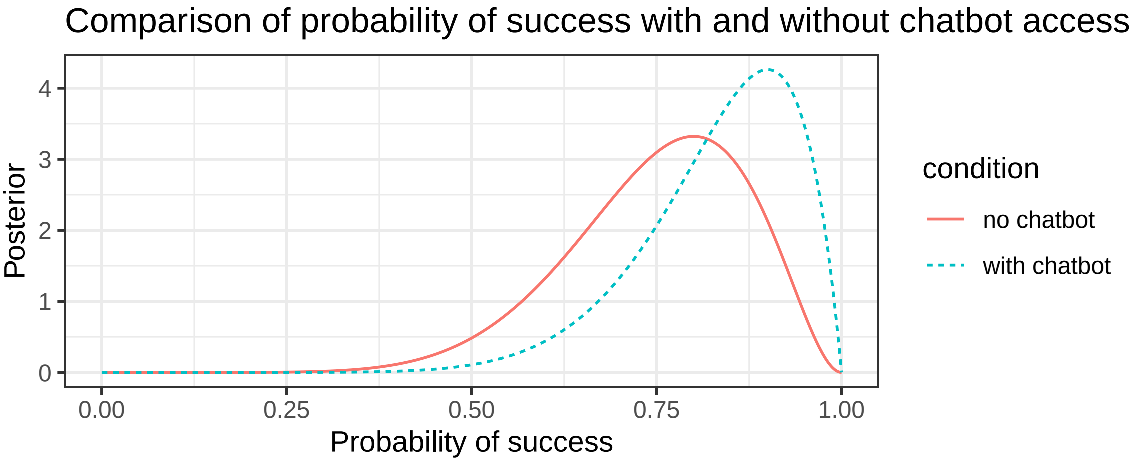
Consider the Workday registration example where we have two conditions:

- no chatbot help and
- with chatbot help.

Research question: How does the chatbot help affect the probability of success in registering for classes?

We randomly select 20 undergraduate students and randomly assign each one a chatbot or no chatbot help condition such that each condition has 10 students (balanced). When we collect the data, we find that 8/10 successfully register without access to chatbot help and 9/10 successfully register with access to chatbot help.

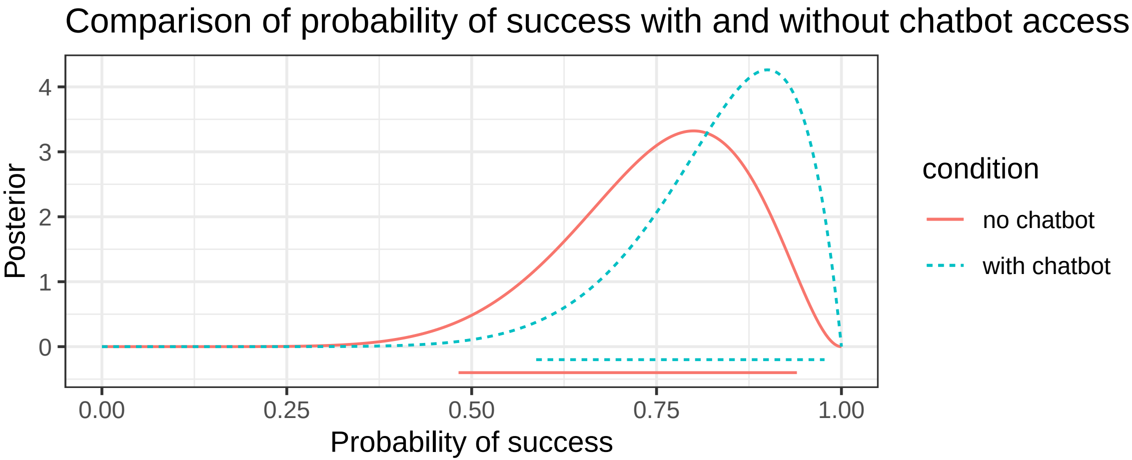
Posterior distributions



95% Credible intervals

```r
a = 1-0.95

# no chatbot access
y <- 8
n <- 10
qbeta(c(a/2, 1-a/2), 1+y, 1+n-y) %>% round(2)

## [1] 0.48 0.94

# with chatbot access
y <- 9
n <- 10
qbeta(c(a/2, 1-a/2), 1+y, 1+n-y) %>% round(2)

## [1] 0.59 0.98
```

Plotting credible intervals



Calculating probabilities

Suppose we are interested in calculate

$$P(\theta_{\text{with chatbot}} > \theta_{\text{no chatbot}}|y)$$

where y generally means "all the data".

We can use a Monte Carlo (or simulation) approach:

```
n_reps = 1e5 # some large number
theta_nochatbot   <- rbeta(n_reps, 1+8, 1+10-8)
theta_withchatbot <- rbeta(n_reps, 1+9, 1+10-9)
mean(theta_withchatbot > theta_nochatbot)

## [1] 0.70493
```