

# The beta-binomial model

Giorgio Corani - (IDSIA, SUPSI)

Bayesian Data Analysis and Probabilistic  
Programming

- The Beta-Binomial model: Ch. 3 of Bayes Rules! An Introduction to Applied Bayesian Modeling
  - <https://www.bayesrulesbook.com/chapter-3.html#chapter-3>
  - Alicia A. Johnson, Miles Q. Ott, Mine Dogucu

# The bias $\theta$ of a coin

- A coin falls tails with probability  $\theta \in (0, 1)$
- $\theta$  is the *bias* of the coin
  - $\theta = 0$ : it always lands tails
  - $\theta = 1$ : it always lands heads
- $\theta \in (0, 1)$  is a continuous parameter

## The bias $\theta$ of a coin

- First we choose a model of our prior beliefs for each possible value of  $\theta$  (*prior*).
- Then we collect some data and we express the probability of observing the data given each value of  $\theta$  (*likelihood*).
- Eventually we use Bayes' rule to obtain the posterior distribution of  $\theta$  given the data.

# The coin problem

- The methodology shown in the following can be used in applications such as estimating:
  - the proportion of supporters of a political party
  - the click-through rate of an online advertisement
  - etc.

## Prior density for a continuous parameter

- The prior for a continuous parameter is specified by a *probability density function* (pdf), denoted by  $f(\theta)$ .
- The pdf specifies all possible values of  $\theta$  and the relative plausibility of each.
- It accounts for all possible values of the parameter and it integrates to 1.
- For  $\theta$ , the pdf is limited on  $(0,1)$

## Properties of $f(\theta)$

- $f(\theta) \geq 0$
- $\int f(\theta) d\theta = 1$
- $P(a < \theta < b) = \int_a^b f(\theta) d\theta$
- The underlying area between  $a$  and  $b$  is the probability of  $\theta$  being in this range.

## Density vs probability

- A continuous pdf is not a probability; we can also have  $f(\theta) > 1$  in some points.
- Probabilities are obtained by integrating the pdf over an interval.
- $f(\theta)$  is used to compare the plausibility of different values of  $\theta$ 
  - the greater  $f(\theta)$ , the more plausible the corresponding value of  $\theta$ .



## The Beta pdf

- Beta( $a, b$ ), is a pdf restricted to the  $[0, 1]$  interval.
- Its parameters are  $a > 0$  and  $b > 0$ . Parameters used in prior models are referred to as *hyperparameters*.
- The pdf is:

$$f(\theta) = \frac{1}{\underbrace{B(a, b)}_{\text{normalizing constant}}} \theta^{a-1} (1 - \theta)^{b-1} \propto \theta^{a-1} (1 - \theta)^{b-1}$$

$a, b > 0$

- $\theta$  is raised to the power of  $a - 1$  (not  $a$ )
- $1 - \theta$  is raised to the power of  $b - 1$  (not  $b$ )

- The **mean** or **expected value** of  $\theta$  is a weighted average: each possible  $\theta$  value is weighted by its pdf:

$$E[\theta] = \int x \cdot f(x) dx$$

- The **mode** is the value of  $\theta$  at which the pdf is highest.

$$\text{Mode}(\theta) = \arg \max_{\theta} f(\theta)$$

- The variance measures the expected squared distance of possible  $\theta$  values from their mean:

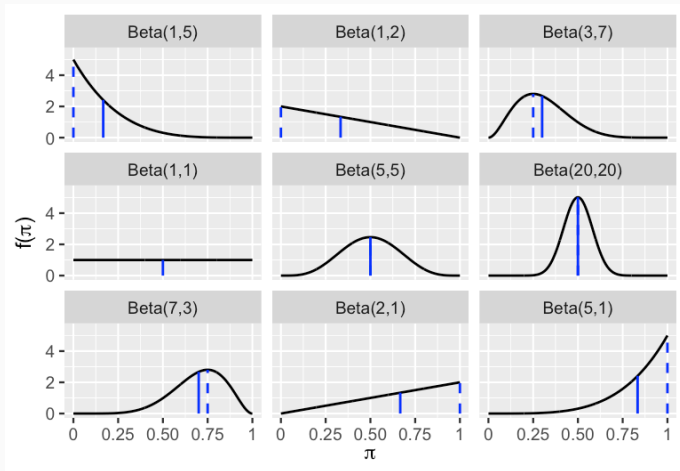
$$\text{Var}(\theta) = E((\theta - E(\theta))^2) = \int (\theta - E(\theta))^2 \cdot f(\theta) d\theta.$$

## Standard deviation

- The variance has squared units; the standard deviation, which measures the typical unsquared distance of  $\theta$  values from  $E(\theta)$ , is easier to interpret.
- The standard deviation measures the expected distance of possible  $\theta$  values from their mean:

$$SD(\theta) := \sqrt{\text{Var}(\theta)}$$

## How the density changes with $a$ and $b$



**Figure 1:** Mean: solid. Mode: dashed.

## Central tendency measures of the Beta

$$E(\theta) = \frac{a}{a+b}$$
$$\text{Mode}(\theta) = \frac{a-1}{a+b-2} \quad \text{when } a, b > 1.$$

## Variability measures for Beta pdf

$$\text{VAR}(\theta) = \frac{ab}{(a+b)^2(a+b+1)}$$

$$\text{SD}(\theta) = \sqrt{\frac{ab}{(a+b)^2(a+b+1)}}$$

## Quiz yourself

- When  $b$ , the pdf is:
  - Right-skewed, with a mode smaller than 0.5.
  - Symmetric with mode 0.5.
  - Left-skewed with mode greater than 0.5.
- Using the same options as above, discuss the pdf when  $a > b$ .
- Which pdf has greater variability: Beta(20,20) or Beta(5,5)?



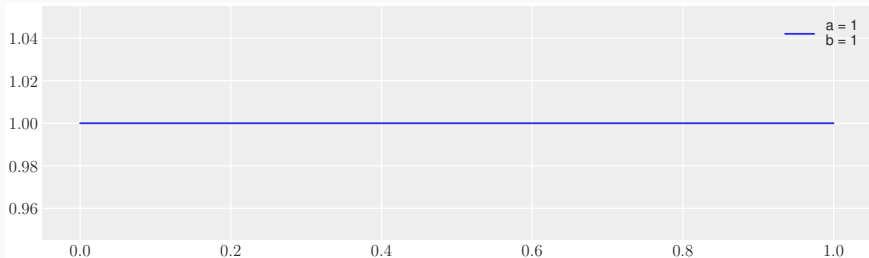
## Effect of $a$ and $b$

- $a > b$ : the distribution is right-skewed, the mode is larger than 0.5; vice versa for  $b > a$ .
- $a = b$ : symmetric distribution with mean 0.5.
- Increasing  $a$  and  $b$  decreases the variance.

## Uniform distribution: $a = b = 1$

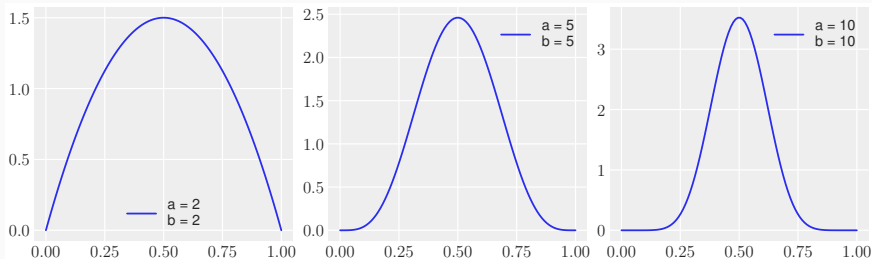
$$\begin{aligned}f(\theta) &\propto \theta^{a-1}(1-\theta)^{b-1} \\&= \theta^0(1-\theta)^0 \\&= 1\end{aligned}$$

- This a *uniform* distribution: all values in  $(0, 1)$  are equally probable.
- $E(\theta) = \frac{a}{a+b} = 0.5$ .



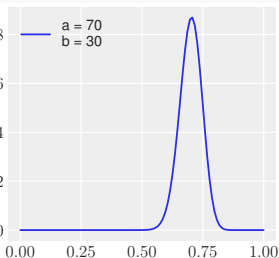
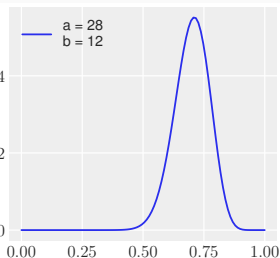
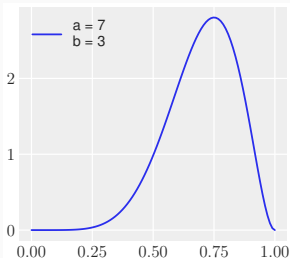
## Increasing $a$ and $b$ the prior becomes more concentrated

- We increase both  $a$  and  $b$  satisfying  $a = b$ .
- The pdf becomes more concentrated around the expected value  $\theta = 0.5$ .



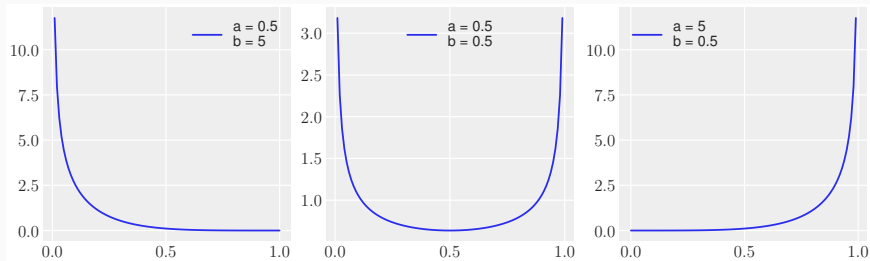
## If we think the coin is rigged

- If we suspect the coin has 70% chance of landing heads, we set  $a = \frac{7}{3}b$ .
- We represent more confidence in this statement by setting  $a = \frac{7}{3}b$  and increasing  $b$ .



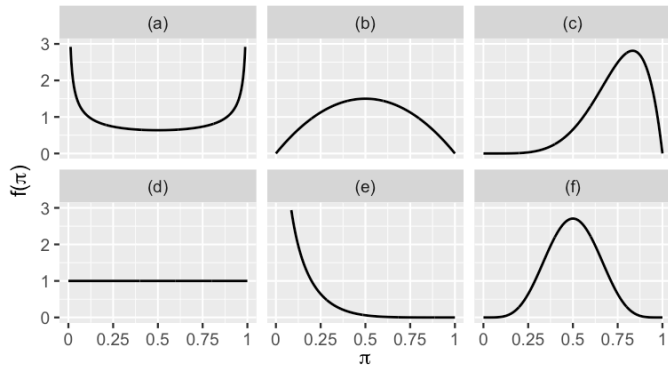
## $a$ and/or $b < 1$

- The density becomes concave (everywhere, or only on part of the domain)



# Quiz

- Recognize Beta(0.5,0.5), Beta(1,1), Beta(2,2), Beta(6,6), Beta(6,2), Beta(0.5,6).



## Tuning $a$ and $b$

- The support for a politician is at about 70 percentage points, though he recently polled between 45 and 90 points.
- We set the ratio  $a/b$  as follows:

$$\frac{a}{a+b} = .7$$
$$a = \frac{7}{3}b$$

- Feasible pairs of values are for instance (7,3), (14,6), etc.

## Tuning $a$ and $b$

- We check how the 5-th and the 95-th quantile vary with  $a$  and  $b$

```
from scipy.stats import beta
q1 = beta.ppf(q=[0.05,0.95],a=7, b=3, loc=0, scale=1)
q2 = beta.ppf(q=[0.05,0.95],a=28, b=12, loc=0, scale=1)
q1
## array([0.45035835, 0.90225319])
q2
## array([0.57661174, 0.81188104])
```



## Tuning $a$ and $b$

- We try different couples  $(a, b, a = \frac{7}{3}b)$  to match the variance.

(a,b)	(7, 3)	(28, 12)	(70, 30)
5-th quantile	0.45	0.58	0.62
95-th quantile	0.90	0.81	0.77

- The choice (7, 3) captures the mean and the variability of the polls in this example.

## Tune a Beta prior!

- Tune a Beta prior for the cases below:
  - John applies to a job. He thinks I has a 40% chance of getting the job, but he is pretty unsure; he expresses his uncertainty by putting his chance between 20% and 60%.
  - A scientist has created a new test for a disease. He expects that the test is accurate 80% of the time with a variance of 0.05.
- Usually there is no single right answer, but multiple reasonable answers.

# The Binomial data model

- After having defined the pdf, the second step of our Bayesian analysis is to collect data.
- We also define the likelihood function, to be used within Bayes' rule.
- In our example, the data collection is done by tossing the coin  $n$  times and observing the number  $y$  of heads.

## Likelihood: assumptions

- Each observation takes a binary value (head or tail; also referred to as *success* and *insuccess*)
- The *success* usually refer to the rarer event among the two.
- The flips are independent: the probability of *heads* at the next flip does not depend on the outcome of the previous flips.
- The success probability  $\theta$  is constant in all flips.

# The binomial likelihood

Given  $\theta$ , a single flip takes:

- *heads* with probability  $\theta$
- *tails* with probability  $1 - \theta$
- Assuming a constant  $\theta$  and the independence of the flips, the sequence

$H \quad T \quad T \quad H \quad H$

has probability

$$\theta(1 - \theta)(1 - \theta)\theta\theta = \theta^2(1 - \theta)^3$$

- In general, a sequence containing  $y$  heads in  $n$  flips has probability

$$\theta^y(1 - \theta)^{n-y}$$

## Binomial likelihood

- We can get  $\binom{n}{y} = \frac{n!}{k!(n-y)!}$  sequences containing  $y$  successes in  $n$  trials.

- The probability of observing  $y$  successes in  $n$  trials is:

$$P(y \mid \theta) = \binom{n}{y} \theta^y (1 - \theta)^{1-y}$$

- This is probability of the observing  $y$  tails within  $n$  flips, given the value of  $\theta$ .

# The Beta-binomial model

$$\theta \sim \text{Beta}(a, b).$$

$$y|\theta \sim \text{Bin}(n, \theta)$$

- This model applies to any setting where parameter  $\theta$  lies in  $[0,1]$ 
  - requires tuning of a Beta prior
  - assumes data  $y$  to be the number of “successes” in  $n$  fixed, independent trials with constant probability of success  $\theta$ .

## Binomial likelihood

- Assume we observe  $y=6$  in  $n=10$  flips.
- The likelihood measures the relative compatibility of the observed data with different  $\theta \in [0, 1]$ .
- According to the data  $\theta=0.6$  is ten times more plausible than  $\theta=0.3$ :

$$\text{Bin}(y = 6, n = 10, \theta = 0.6) = \binom{10}{6} 0.6^6 (0.4)^4 = 0.35$$

$$\text{Bin}(y = 6, n = 10, \theta = 0.3) = \binom{10}{6} 0.3^6 (0.7)^4 = 0.037$$



## Binomial likelihood

$$P(y \mid \theta) = \binom{n}{y} \theta^y (1 - \theta)^{1-y}$$

- This a *likelihood* function if interpreted in this way:
  - the probability is a function of  $\theta$ .
  - the observation  $y$  are fixed
- The likelihood function shows how the probability of the observed data varies with  $\theta$ .
- It does not integrate to 1 over all values of  $\theta$ !
- It integrates to 1 if we keep  $\theta$  fixed and we integrate over possible outcomes  $y$ . But this would not be a likelihood function!

- Adopting a beta prior and a binomial *likelihood*, Bayes' rule yields a beta *posterior* density with updated parameters.
- We use  $f$  rather than  $p$  as it represents a density.

$$f(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$$

Beta prior

$$f(y \mid \theta) \propto \theta^y(1-\theta)^{n-y}$$

Binomial likelihood

$$f(\theta \mid y) \propto \theta^{y+a-1}(1-\theta)^{n-y+b-1}$$

Beta posterior

The beta prior is *conjugate* with the binomial likelihood, as we obtain a posterior Beta pdf.

- The Beta-binomial model is **conjugate**.
- The prior is conjugated with the likelihood if the posterior has the same functional form of the prior.
- Historically, problems in Bayesian statistics were restricted to the use of conjugate priors, because of mathematical tractability.
- Modern computational techniques allow Bayesian analysis without conjugacy, allowing the resurgence of Bayesian statistics in recent years.

# The posterior is a compromise of prior and likelihood

- Given the prior  $\text{Beta}(a,b)$ , the prior mean of  $\theta$  is:

$$\frac{a}{a+b}$$

- Having observed  $y$  tails in  $n$  flips, the posterior pdf of  $\theta$  is  $\text{Beta}(y+a, n-y+b)$ .

- The posterior mean of  $\theta$  is:

$$E_{\text{post}}[\theta] = \frac{a+y}{a+y+b+n-y} = \frac{a+y}{a+b+n}$$

# The posterior is a compromise of prior and likelihood

## ■ Rearranging:

$$\underbrace{\frac{a + y}{a + b + n}}_{\text{posterior}} = \underbrace{\frac{y}{n}}_{\text{observed proportion}} \underbrace{\frac{n}{n + a + b}}_{\text{weight}} + \underbrace{\frac{a}{a + b}}_{\text{prior mean of } \theta} \underbrace{\frac{a + b}{n + a + b}}_{\text{weight of the prior}}$$

- The posterior mean is a weighted average of the prior mean and the observed proportion.
- The weight of the observed proportion increases with  $n$ ; the weight of the prior mean increases with  $a$  and  $b$ .

## The posterior is a compromise of prior and likelihood

$$\underbrace{\frac{a + y}{a + b + n}}_{\text{posterior}} = \underbrace{\frac{y}{n}}_{\text{observed proportion}} \underbrace{\frac{n}{n + a + b}}_{\text{weight}} + \underbrace{\frac{a}{a + b}}_{\text{prior mean of } \theta} \underbrace{\frac{a + b}{n + a + b}}_{\text{weight of the prior}}$$

- We can interpret the prior as representing an imaginary sample, containing  $a$  successes and  $b$  insuccesses.
- The larger  $a$  and  $b$ , the larger the imaginary sample; thus our confidence in the prior increases.

## Test yourself!

- Let  $\theta$  denote the proportion of people that prefer dogs to cats.
- You express your prior beliefs by a Beta(7, 2) model.
- According to your prior, what are reasonable values for  $\theta$  ?
- In a survey 19 out of 20 people prefer dogs.
- How would that change your understanding about the mean and the certainty of  $\theta$ ?

## Sequential updating

- Based on some theoretical studies, a scientist summarizes its belief in the chance  $\theta$  of a new drug being able to cure a disease as Beta(1,10) distribution.
- In an experimental trial, the drug cures 13/20 persons.
- What's the posterior distribution of  $\theta$  after the first experiment?
- In a second experiment, the drug cures 20/40 persons.
- What's the posterior distribution of  $\theta$  after the second experiment?



## Sequential updating

- Prior:  $\text{Beta}(1,10)$ ,  $E[\theta] = \frac{1}{11} = 0.09$
- After first experiment:
  - $f(\theta|D_1) = \text{Beta}(1 + 13, 10 + 20)$
  - $E[\theta] = \frac{14}{44} = 0.32$
  - Thus  $\text{Beta}(14,30)$  becomes the prior before analyzing the data of the second experiment.

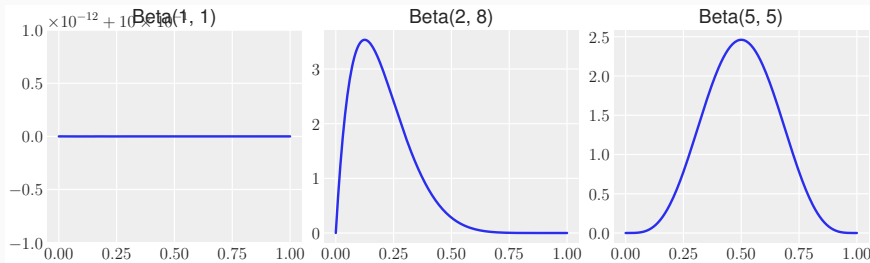
- After second experiment:

- $\text{Beta}(14+20, 30+40)$

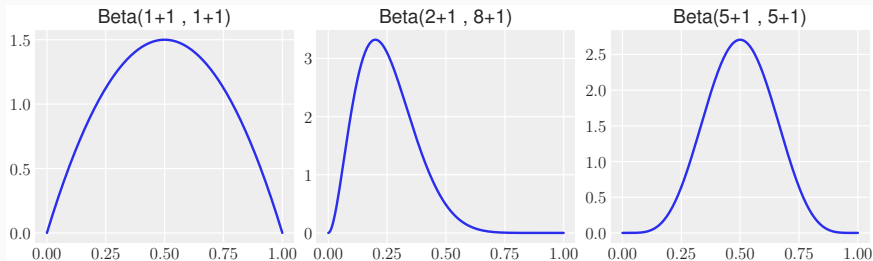
- $E[\theta] = \frac{34}{104} = 0.33$

# Impact of the prior on the posterior

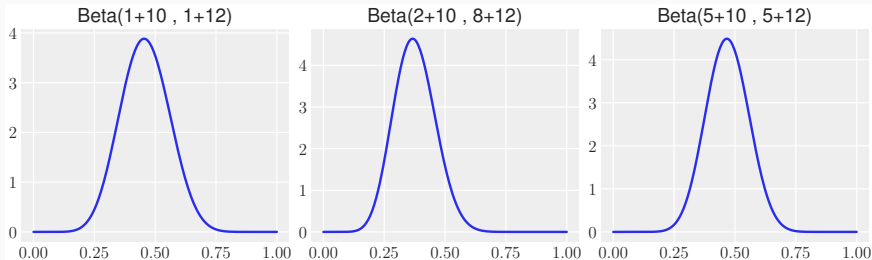
- It is useful to consider different priors: priors encode domain expertise, and different experts provide you with reasonable but different assessment.
- For instance:



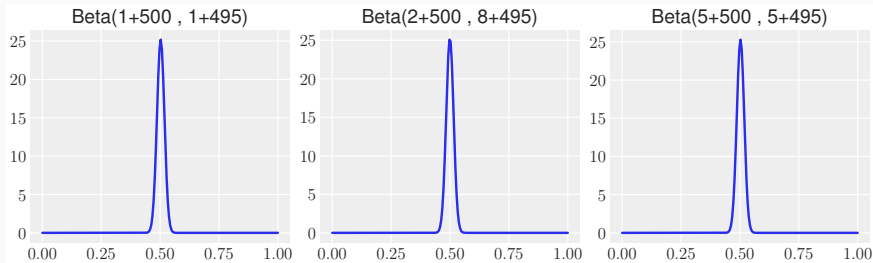
## The posterior is prior-sensitive with small data (example: $y=1$ , $n=2$ )



## The posterior becomes similar with more data (10 tails, 12 heads)



# When data is larger, the posterior is the same whatever the prior



Posterior means  $E_{\text{post}}[\theta]$  obtained from different priors:

- $\frac{500+1}{500+1+495+1} = \frac{501}{997} = 0.502$

- $\frac{500+2}{500+2+495+8} = \frac{502}{1005} = 0.499$

- $\frac{500+5}{500+5+495+5} = \frac{505}{1005} = 0.502$

- Also the posterior variances are practically identical.

## Test your self!

For each scenario of the next slide, identify whether

- the prior has more influence on the posterior
- the data has more influence on the posterior
- the posterior is an equal compromise between the data and the prior.

## Test your self!

■ Prior:  $\theta \sim \text{Beta}(1,4)$ , data:  $y=8, n=10$

■ Prior:  $\theta \sim \text{Beta}(20,3)$ , data:  $y=0, n=1$

■ Prior:  $\theta \sim \text{Beta}(4,2)$ , data:  $y=1, n=3$

■ Prior:  $\theta \sim \text{Beta}(20,2)$ , data:  $y=10, n=200$

git # The posterior mean is just part of the information

■ Bayesian analysis yields the posterior distribution of  $\theta$ , **not** a single value.

■ The dispersion of the posterior is a measure of our uncertainty.

■ The uncertainty decreases when we have more data.



## Sensitivity to the prior

- With a large amount of data, the posterior is practically the same with any prior, but how much data is needed varies with the problem.
- If we only have few data, the posterior can differ depending on the adopted prior; it makes sense to repeat the analysis with different priors (*sensitivity*).
- This is sensible: the prior encodes our previous knowledge and different experts could have different priors.

## Discussion

- Priors and likelihood are assumptions which are part of the model.
- Flat priors provide no information (uninformative priors) and should be avoided.
- *Slightly informative* priors are recommended.
- In many cases we know that the parameter can only be positive, or its order of magnitude, etc.
- For instance a  $\text{Beta}(1,1)$  prior is flat but limits the possible values of  $\theta$  between 0 and 1.

## Priors need a broad support

- In the following slides we discuss some problem which arise if the support of the prior is too small.
- The *support* of a pdf is the set of points where the pdf is  $>0$ .

## Priors need a broad support

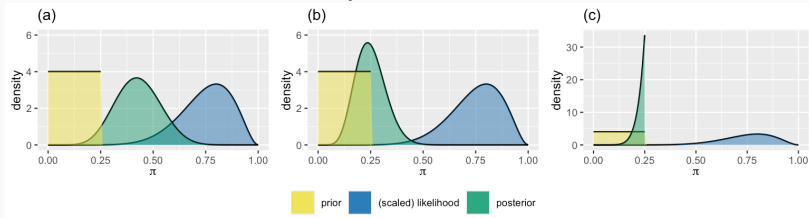
- Bayesian analysis loses its benefits if the prior pdf has a too small support, i.e. it assigns a prior probability of zero also to plausible parameter values.
- For instance a priori we assume  $\pi$  to be equally likely anywhere between 0 and 0.25 and that surely it doesn't exceed 0.25:

$$\pi \sim \text{Unif}(0, 0.25)$$

- And assume to observe  $y=8$  successes in  $n=10$  trials.

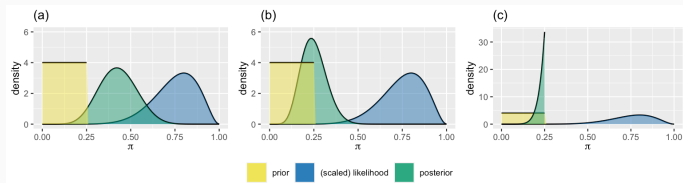
# Priors need a broad support

- The prior pdf, the scaled likelihood and the posterior are shown below. Which is correct plot?



# Priors need a broad support

- The correct plot is the third.



- The support of the posterior is inherited from the support of the prior.
- Thus both prior and posterior assigns zero probability to any  $\pi > 0.25$ . the posterior model must also assign zero probability to any value in that range.
- No matter how much evidence we will collect, the posterior pdf will be truncated beyond the 0.25 cap.

## How to avoid a regrettable prior

- Let  $\pi$  be the parameter of interest.
- Be sure to assign non-0 pdf to every *possible* value of  $\pi$ .
- For example, if  $\pi$  is a proportion which can range from 0 to 1, the prior model should be defined on this range.

- We have seen how Bayesian inference works when Bayes' rule can be solved analytically (conjugacy).
- Only simple likelihood functions have conjugate priors.
- Complex models have no conjugate priors and requires numerical Markov chain Monte Carlo (MCMC) to get the posterior.