Bayesian Statistics for Genetics

Lecture 2: Binomial Sampling, part 1

June, 2024

Outline

Important ideas we will recap:

- Bayes' Theorem – a statement of conditional probability
- Bayesian inference – using probability to describe **belief**

In this session:

- More formal analysis of the ACE study's binomial model
- What to do with a posterior distribution?

Bayes theorem: conditional probability

For a *partition* $\{H_1, \ldots, H_K\}$, the axioms of probability imply the following:

- Rule of total probability:

$$\sum_{k=1}^{K} \mathbb{P}[\, H_k \,] = 1$$

- Rule of marginal probability:

$$\mathbb{P}[\, A \,] \;=\; \sum_{k=1}^{K} \mathbb{P}[\, A \text{ and } H_k \,] = \sum_{k=1}^{K} \mathbb{P}[\, A | H_k \,] \mathbb{P}[\, H_k \,]$$

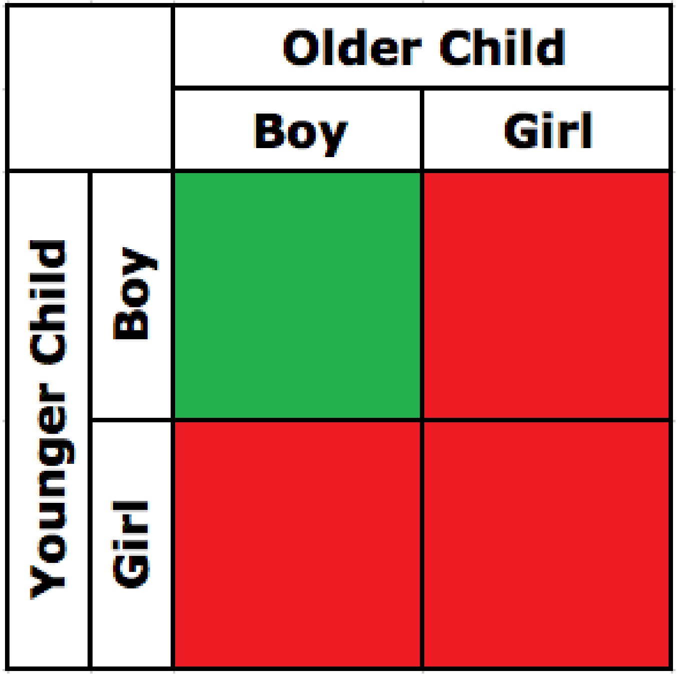
Simple case: $K = 2$ with $H_1 = B$ and $H_2 = B^c$ (the complement of B):

$$\begin{aligned} \mathbb{P}[\, A \,] &= \mathbb{P}[\, A \text{ and } B \,] + \mathbb{P}[\, A \text{ and } B^c \,] \\ &= \mathbb{P}[\, A | B \,] \mathbb{P}[\, B \,] + \mathbb{P}[\, A | B^c \,] \mathbb{P}[\, B^c \,]. \end{aligned}$$

Bayes' Theorem: conditional probability

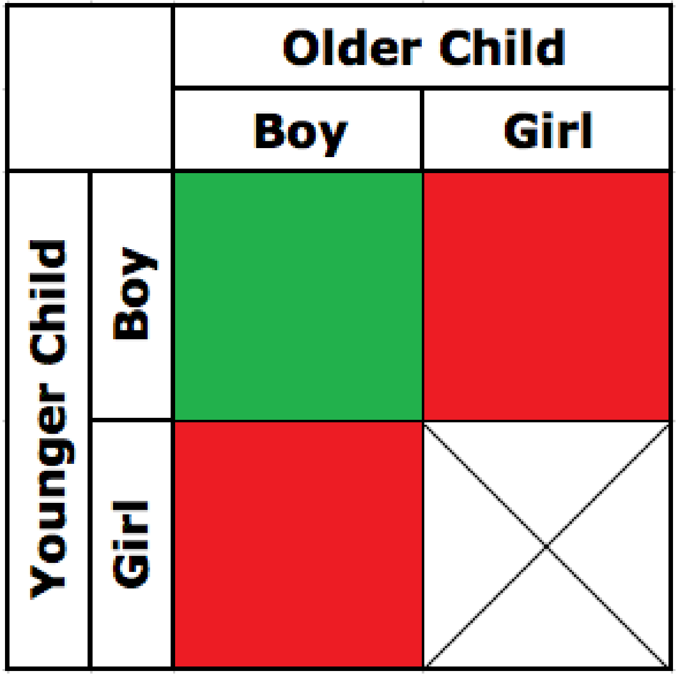
Some genetics! Jo* — a randomly-chosen father of two with at least one boy — has two kids. **Given that** at least one is a boy; what's the probability he has two boys?

Unconditional



$\mathbb{P}[2 \text{ Boys}] = 1/4 = 0.25$

Conditional



$\mathbb{P}[2 \text{ Boys} | 1 \text{ Boy}] = 1/3 \approx 0.33$

Bayes' Theorem: conditional probability

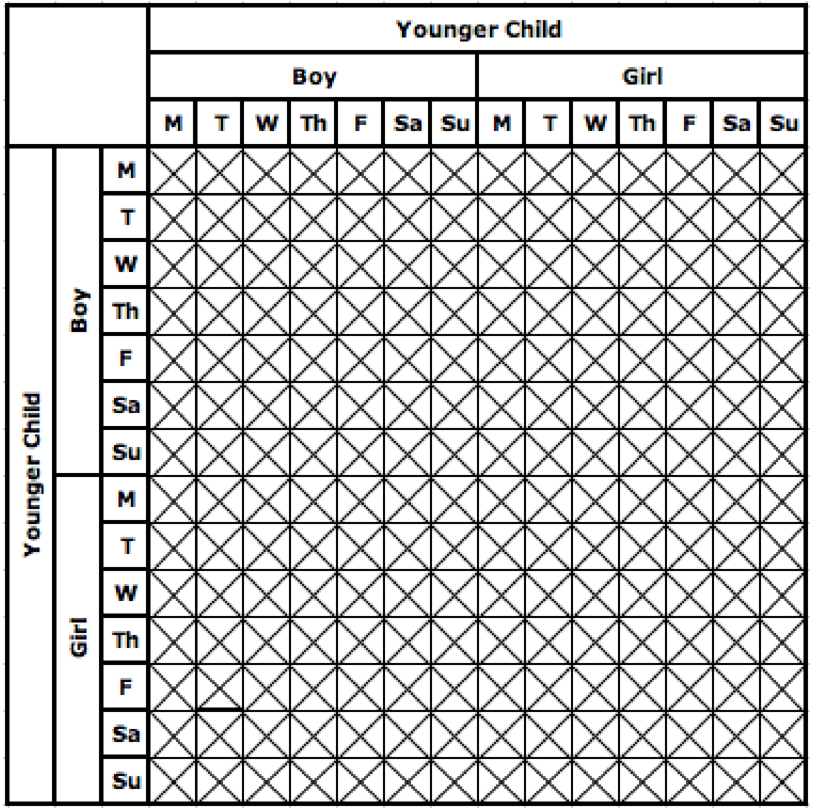
Now a problem – not a trick! – to show that conditional probability can be non-intuitive, and careful reasoning is needed;

Q. Jo has two children. **Given that** at least one is a *boy who was born on a Tuesday*; what's the probability he has two boys?

- The 'obvious' (but wrong!) answer is to stick with 1/3. What can Tuesday possibly have to do with it?
- It may help your intuition, to note that a boy being born on a Tuesday is a (fairly) rare event;
  - Having two sons would give Jo two chances of experiencing this rare event
  - Having only one would give him one chance
  - 'Conditioning' means we **know** this event occurred, i.e. Jo was 'lucky' enough to have the event
- **Easier Q.** Is $\mathbb{P}[\text{2 Boys}|\text{1 Tues Boy}] > 1/3$? or $< 1/3$?

Bayes' Theorem: conditional probability

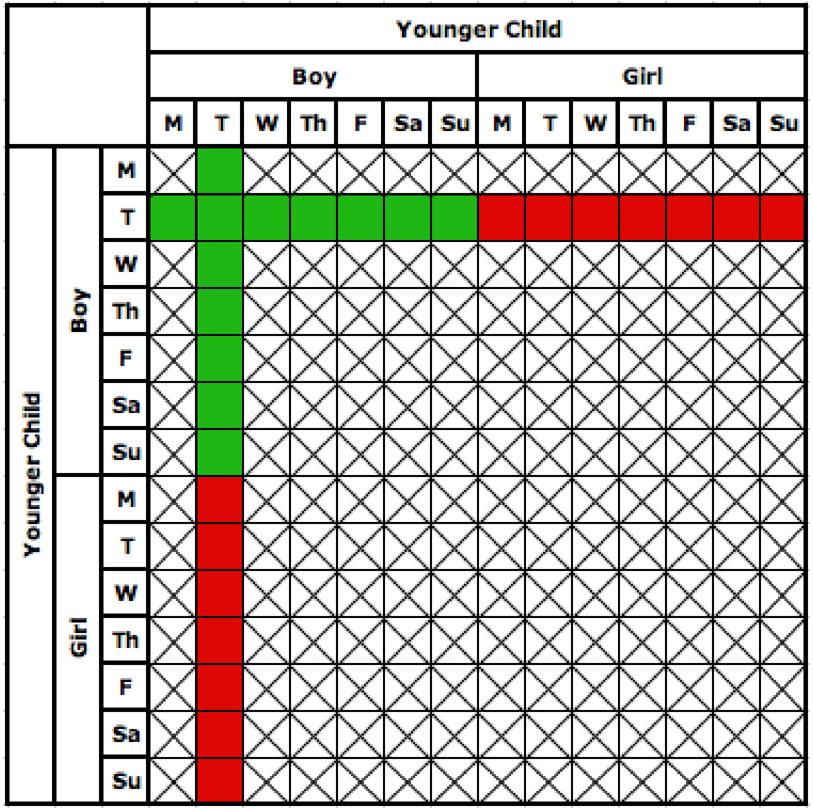
All the possible births and sexes;



Q. When we condition, which row and column are we considering?

Bayes' Theorem: conditional probability

Conditioning on at least one Tuesday-born boy;



... giving $\mathbb{P}[2 \text{ Boys}|1 \text{ Tues Boy}]=13/27\approx 0.48$, **quite different** from $1/3\approx 0.33$.

Bayes' Theorem: conditional probability

Formal example: Let $B =$ Female and $B^c =$ Male . Suppose in a given population over the age of 18:

$$\mathbb{P}[\,B\,] = 0.55, \qquad \mathbb{P}[\,B^c\,] = 0.45.$$

Event of interest: $A =$being diagnosed with diabetes.

In the US in 2018, for over 18 year olds, $\mathbb{P}[\,A|B\,] = 0.095$ and $\mathbb{P}[\,A|B^c\,] = 0.11$, so

$$\begin{aligned}\mathbb{P}[\,A\,] &= \mathbb{P}[\,A|B\,]\mathbb{P}[\,B\,] + \mathbb{P}[\,A|B^c\,]\mathbb{P}(B^c\,] \\ &= 0.095 \times 0.55 + 0.11 \times 0.45 \\ &= 0.05225 + 0.0495 \\ &= 0.10175\end{aligned}$$

So 10.2% of the population have diabetes.

Bayes theorem: Flipping around the conditioning

$$\text{Bayes theorem}: \ \mathbb{P}(H_j|E) \ = \ \frac{\overbrace{\mathbb{P}(E|H_j)}^{\text{"Likelihood"}} \ \overbrace{\mathbb{P}(H_j)}^{\text{"Prior"}}}{\underbrace{\mathbb{P}(E)}_{\text{Normalizing Constant}}} = \frac{\mathbb{P}(E|H_j)\mathbb{P}(H_j)}{\sum_{k=1}^{K} \mathbb{P}(E|H_k)\mathbb{P}(H_k)}$$

for $j = 1, \ldots, K$.

Anticipating Bayesian inference:

- One begins with (prior) belief about events H_j, $\mathbb{P}(H_j)$, and...
- ...updates it to (posterior) belief $\mathbb{P}(H_j|E)$, given that event E occurs.

Note that the likelihood, on its own, doesn't generally describe beliefs.

Bayes theorem: Flipping around the conditioning

What's the probability that a person with diabetes is female?

In probability speak:

$$\begin{aligned}\mathbb{P}(B|A) &= \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)} \\ &= \frac{0.095 \times 0.55}{0.10175} \\ &= 0.514\end{aligned}$$

So there is a 0.514 chance that a randomly sampled person with diabetes is female.

This is *updated* from our prior probability of being female $\mathbb{P}(B) = 0.55$ – it's a slight reduction since males are more likely to have diabetes.

Conditional independence

Conditional independence is a key concept when constructing statistical models – we start by describing *independence*.

For events A and B, it is always true that,

$$\mathbb{P}(\, A \;\text{ and }\; B\,) = \mathbb{P}(\, A \,|\, B\,) \times \mathbb{P}(\, B\,).$$

Bayes theorem:

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)}.$$

Viewed in a Bayesian way, knowledge that A occurs has *updated our beliefs* about B.

How about when we **don't** learn anything from B's occurrence?

Conditional independence

Then

$$\mathbb{P}(\,B\,|\,A\,) = \mathbb{P}(\,B\,)$$

or equivalently

$$\mathbb{P}(\,A \text{ and } B\,) = \mathbb{P}(\,A\,) \times \mathbb{P}(\,B\,).$$

- The events A and B are said to be *independent*.
- Knowledge that A occurs does not affect our beliefs about B.
- Knowledge that B occurs does not affect our beliefs about A, i.e., this implies $\mathbb{P}(A|B) = \mathbb{P}(A)$.

If diabetes risk was the same in females and males, then knowing diabetes status, A, would not tell us anything about the sex of the person, B, i.e., $\mathbb{P}(B|A) = \mathbb{P}(B)$.

Conditional independence

In statistical modeling, independence is rarely relevant, but conditional independence is ubiquitous.

Extending this idea, events F and G are *conditionally independent given* H, if

$$\mathbb{P}(\,F \text{ and } G \,|\, H\,) \;=\; \mathbb{P}(\,F\,|\,H\,) \times \mathbb{P}(\,G\,|\,H\,),$$

Or written another way:

$$\mathbb{P}(\,F\,|\,G, H\,) \;=\; \mathbb{P}(\,F\,|\,H\,).$$

Given H, knowledge that G occurred does not alter our beliefs in F occurring.

Conditional Independence: Example

Data:

Suppose we know events:

$F = \{$ a patient develops cancer $\}$
$G = \{$ patient's parent's genotype $\}$
$H = \{$ patient's genotype $\}$

Informal statement:

If we know the patient's genotype H, does knowledge of the parents' genotype G give any additional information?

Formal statement:

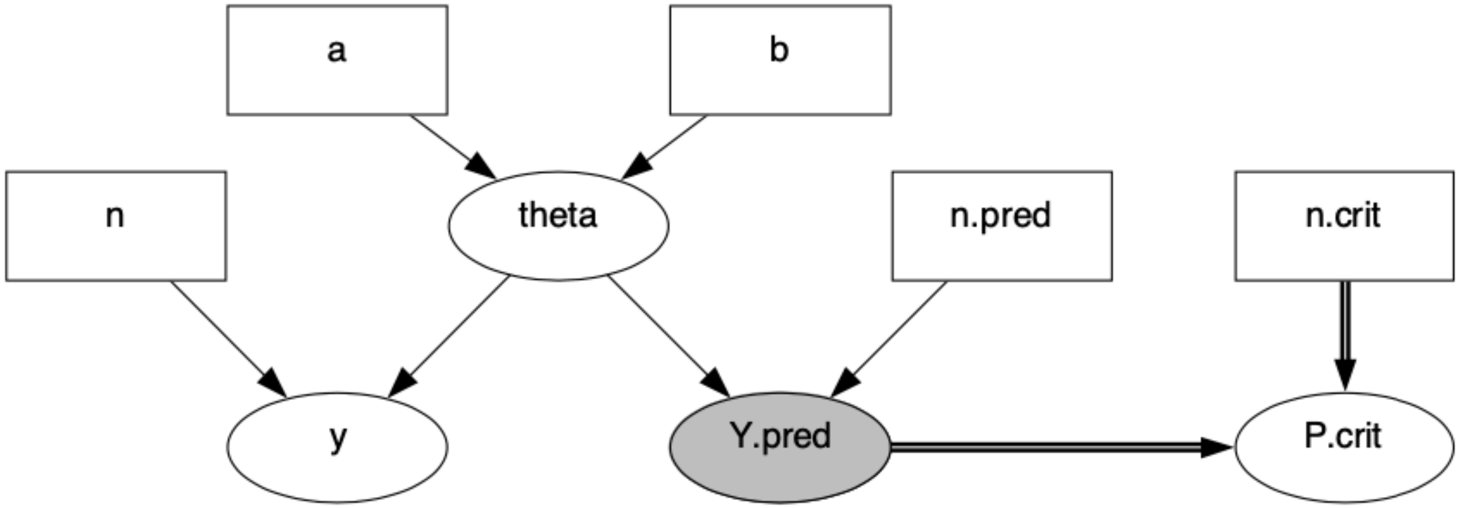
Does

$$\mathbb{P}(\, F \,|\, H \,) = \mathbb{P}(\, F \,|\, G, H \,)?$$

Answer: In general, conditional independence will hold, but not on all occasions; in genomic imprinting genes are expressed in a parent-of-origin-specific manner, i.e., the expression of the gene depends upon the parent who passed on the gene.

Conditional Independence: Example

Conditional independencies can be neatly expressed through graphs, as in this example from the BUGS book (Lunn *et al* 2013)



Conditioning on a connecting node 'blocks' the path between other variables. (This format may also be familiar from causal analysis)

Conditional Independence: Example

In likelihood-based inference, conditional independence is *very* widely-used.

For example, the sampling model for data $\boldsymbol{y} = [y_1, \ldots, y_n]^T$ is often taken to be:

$$\begin{aligned}
p(\boldsymbol{y}|\boldsymbol{\theta}) &= p(y_1, \ldots, y_n|\boldsymbol{\theta}) \\
&= p(y_1|\boldsymbol{\theta}) \times p(y_2|y_1, \boldsymbol{\theta}) \times \cdots p(y_n|y_{n-1}, \ldots, y_1, \boldsymbol{\theta}) \\
&= p(y_1|\boldsymbol{\theta}) \times p(y_2|\boldsymbol{\theta}) \times \cdots p(y_n|\boldsymbol{\theta}) \\
&= \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta})
\end{aligned}$$

where we have assumed conditional independence, i.e., given $\boldsymbol{\theta}$, the observations are independent.

Example: For coin tosses, the outcomes are conditionally independent, given the probability of a head θ. (But what happens if we have > 1 coin?)

Overview of Bayesian Inference

At a high level, with a model specified and data available, Bayes is automatic. (Examples follow!) But it's worth noting that integration, i.e. averaging, in some form, is usually the biggest hurdle. Bayesian approaches to:

- Estimation: marginal posterior distributions on parameters of interest – similar approaches permit testing. Need to integrate over the other parameters
- Prediction: via the predictive distribution, integrating over parameter uncertainty
- Hypothesis Testing: Bayes factors give the relative support for different ranges of θ – and a different form of testing. Need to average over different submodels

We'll describe all three in the context of a *binomial model* – in general we focus on estimation and prediction.

Elements of Bayes Theorem for a Binomial Model

Suppose the data consist of N Bernoulli (i.e. 0/1) responses y_i, $i = 1, \ldots, N$.

We assume these responses are conditionally independent, given a common "success" probability θ.

Under this conditional independence assumption, the distribution of the total $y = \sum_{i=1}^{N} y_i$ has to be a *binomial* distribution, in which

$$\mathbb{P}[\,Y = y \,|\, \theta\,] = \binom{N}{y} \theta^y \, (1-\theta)^{N-y}$$

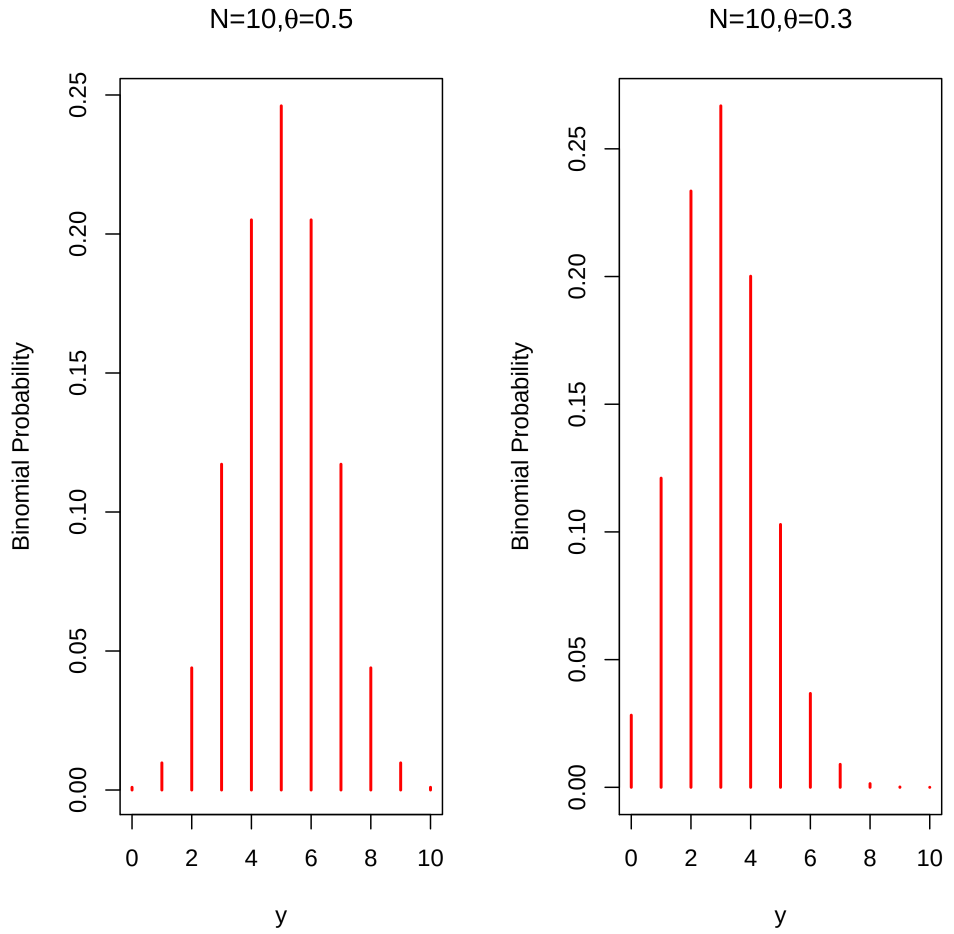
is the probability of seeing $Y = y$, for the permissible values $y = 0, 1, \ldots, N$ **given** the probability θ.

Elements of Bayes Theorem for a Binomial Model

Binomial distributions (right) for two values of θ with $N = 10$.

Fixing y, we may view the probability of the data as a function of θ – when it is known as the likelihood function:

$$L(\theta) = \theta^y (1-\theta)^{N-y}.$$



Elements of Bayes Theorem for a Binomial Model

The maximum likelihood estimate (MLE) is the proportion of successes:

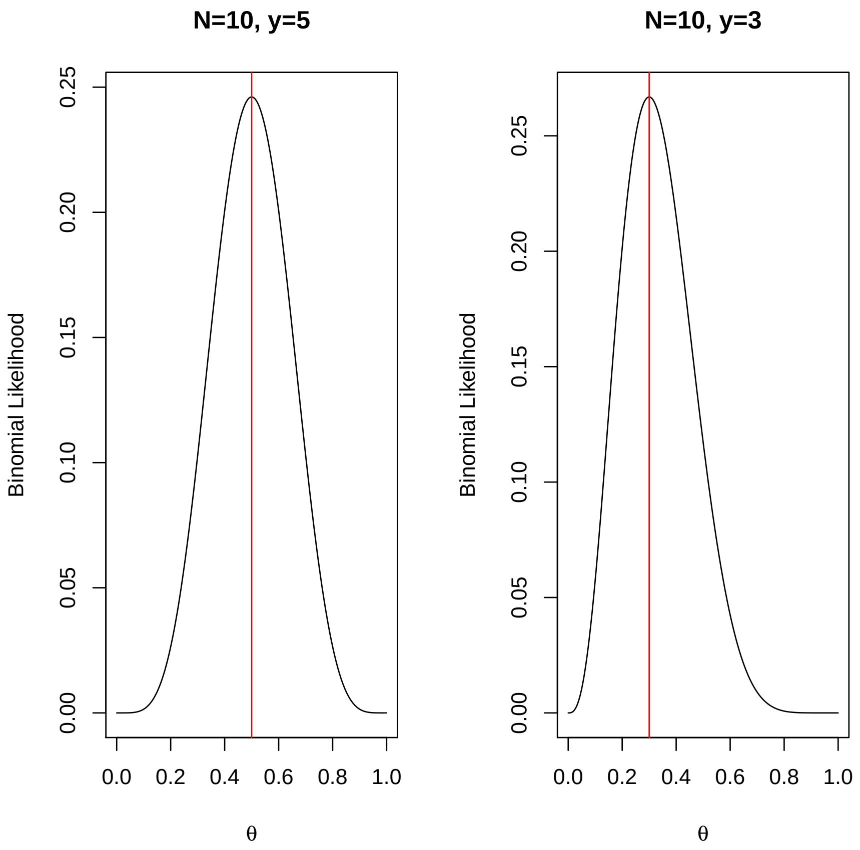
$$\widehat{\theta} = \frac{y}{N} = \overline{y},$$

and gives the highest probability to the observed data, i.e., maximizes the likelihood function. The standard error of this estimate is

$$\sqrt{\theta(1-\theta)/N}.$$

which we approximate by

$$\sqrt{\widehat{\theta}(1-\widehat{\theta})/N}.$$



Binomial likelihoods for $y = 5$ (left) and $y = 10$ (right), with $N = 10$. The MLEs are indicated in red.

Bayes and frequentist estimates for binomial

If $y = 0$ ($y = N$), we get estimate $\widehat{\theta} = 0$ (=1) and a standard error of 0, which is clearly problematic.

Agresti & Coull (1998) give a famous workaround, the "Adjusted Wald interval": with estimate

$$\tilde{\theta} = \frac{4}{N+4}\frac{1}{2} + \frac{N}{N+4}\overline{y},$$

to give the interval:

$$\tilde{\theta} \ \pm \ 1.96\sqrt{\tilde{\theta}(1-\tilde{\theta})/N}.$$

It works well in practice, but what might be a more convincing justification for it?

Beta priors for Binomial θ

Recall Bayes Theorem: $p(\theta|y) \propto p(y|\theta) \times p(\theta)$.

- Bayes theorem requires the *likelihood*, which we have already specified as binomial, and a *prior*.
- For a probability $0 < \theta < 1$ an obvious candidate prior is the uniform (i.e. flat) distribution on (0,1): but this is too restrictive for general use.
- The beta distribution, Beta(a, b), is more flexible. (The uniform distribution is a special case with $a = b = 1$.) We specify a and b in advance, i.e., *a priori*.
- The form of the beta distribution is

$$p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

for $0 < \theta < 1$, where $\Gamma(\cdot)$ is the gamma function*.

*$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$

Beta priors for Binomial θ

- The Beta(a, b) distribution is valid† for $a > 0, b > 0$.
- How can we think about specifying a and b?
- As you may know, the Normal distribution is specified by its mean (μ) and variance (σ^2), but the beta distribution's a and b are less simple.
- The mean and variance are:

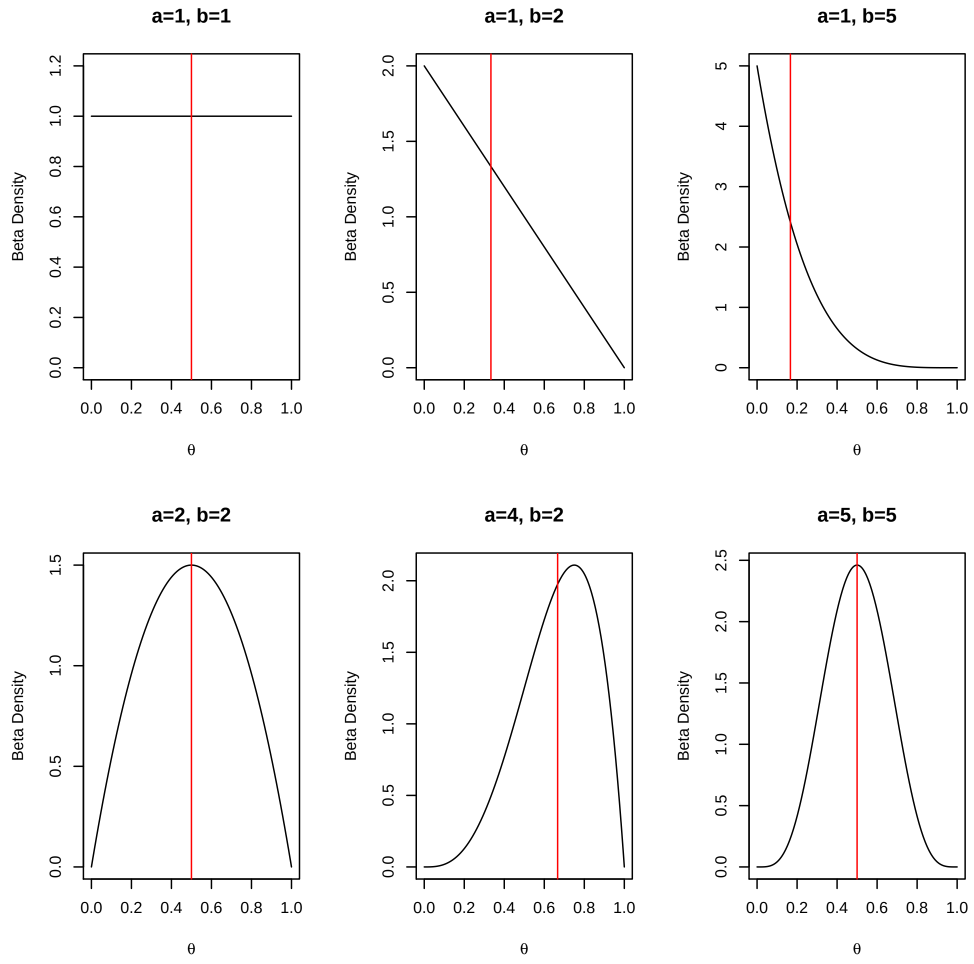
$$\begin{aligned}\mathbb{E}[\theta] &= \frac{a}{a+b}\\ \text{Var}[\theta] &= \frac{\mathbb{E}[\theta](1-\mathbb{E}[\theta])}{a+b+1}.\end{aligned}$$

Hence, increasing a and b concentrates the distribution about the mean.

†A distribution is valid if it is non-negative and integrates to 1

Beta priors for Binomial θ

The quantiles, e.g. the median or the 10% and 90% points, are not available as a simple formula, but are easily obtained within software – in R we use the function `qbeta(p,a,b)`.

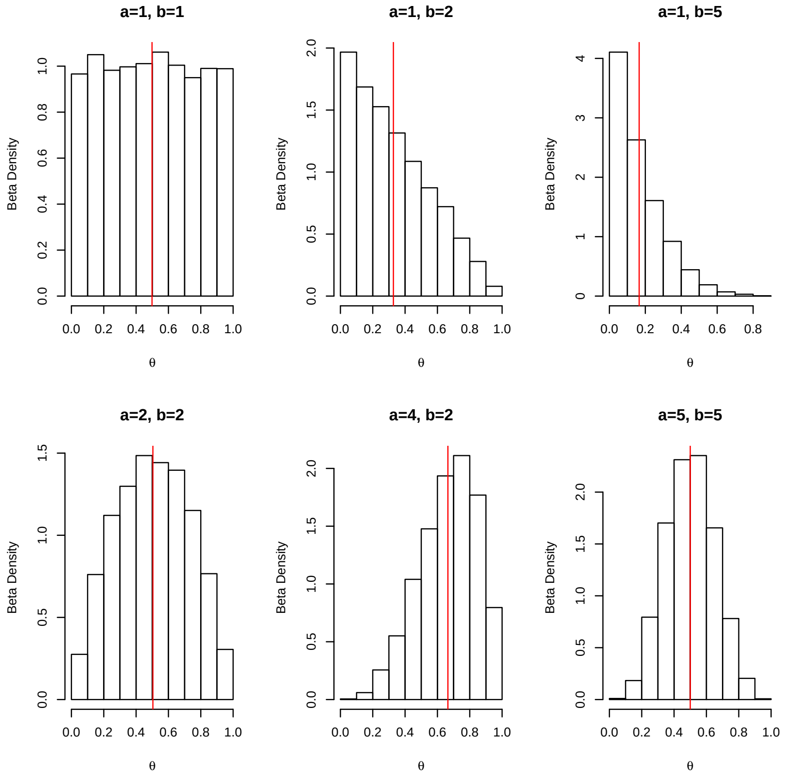


Beta distributions, Beta(a, b) (right). The red lines indicate the means.

Samples to Summarize Beta Distributions

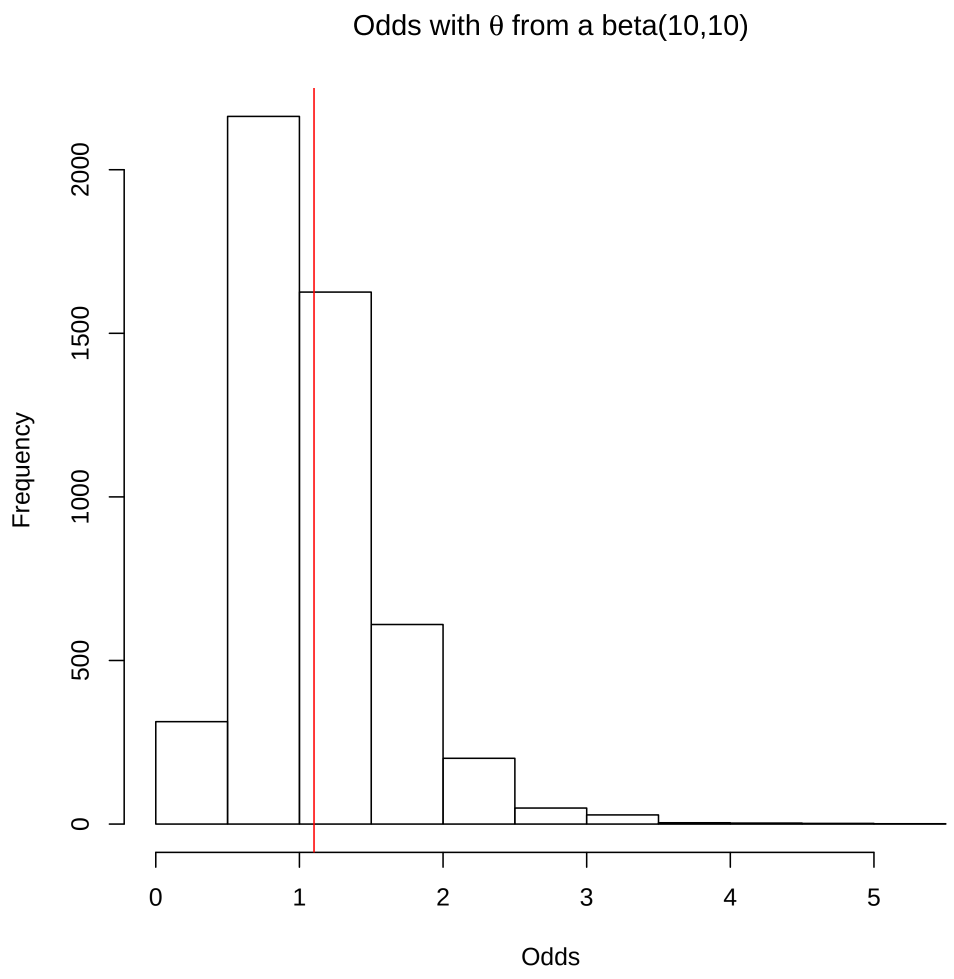
Probability distributions and samples from distributions are equivalent, in a sense: given a probability distribution we can generate samples, and given a big-enough sample we can reconstruct their probability distribution. (More on this later!)

- Probability distributions can be investigated by generating samples from them, and then examining histograms, moments and quantiles
- Right, some histograms of samples from beta distributions for different choices of a and b, with sample means in red
- Compare with previous slide to see the duality



Samples for Describing Weird Parameters

- Generating samples for e.g. a Beta's mean seems overkill – recall 2.22
- But for functions of the probability θ, such as the odds $\theta/(1-\theta)$, sampling is the easiest method
- Once we have samples for θ we can simply transform the samples to the functions of interest.
- We may have clearer prior opinions about the odds, than the probability.
- Right: samples from the prior on the odds $\theta/(1-\theta)$ with $\theta \sim \text{Beta}(10, 10)$. The red line indicates the sample mean.



Issues with Uniform Priors

If we have little prior information about a parameter, we might think that a uniform prior, i.e. a prior $p(\theta) \propto$ const reflects this ignorance. But there are two problems:

1. We can't be uniform on all scales since, if $\phi = g(\theta)$:

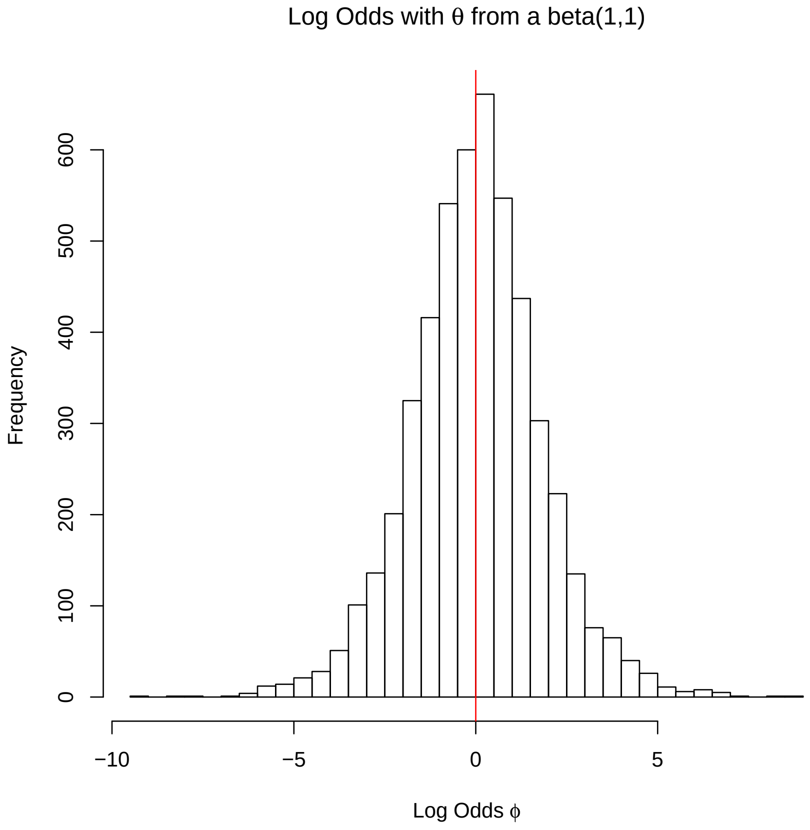
$$\underbrace{p_\phi(\phi)}_{\text{Prior for } \phi} = \underbrace{p_\theta(g^{-1}(\phi))}_{\text{Prior for } \theta} \times \underbrace{\left|\frac{d\theta}{d\phi}\right|}_{\text{Jacobian}}$$

   and so if $g(\cdot)$ is a nonlinear function, the Jacobian will be a function of ϕ and hence not uniform.

2. If the parameter is not on a finite range, an improper distribution will result (that is, the form will not integrate to 1). This can lead to an improper posterior distribution, and without a proper posterior we can't do inference.

Issues with Uniform Priors

- For example, what does a flat prior on Binomial θ imply about log odds $\phi = \log\left(\frac{\theta}{1-\theta}\right)$? (Both are arguable 'natural' choices)
- The answer (right) is a very **non**-uniform distribution



Not being uniform on all scales need not be a problem, but do be aware of it, and cautious with 'flat' priors. They don't describe ignorance – often the opposite.

Posterior Derivation: The Quick Way

When we want to identify a particular probability distribution we *only* need to concentrate on terms that involve the random variable.

For example: as seen in 2.21, the form of the beta distribution is

$$p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

But if we just knew the density was proportional to $\theta^{a-1}(1-\theta)^{b-1}$, we could work out the other terms – all they do is ensure $p(\theta)$ integrates to 1.

(We haven't yet looked at Normal distrbutions, but for random variable X with density of the form $p(x) \propto \exp(c_1 x^2 + c_2 x)$ for constants c_1 and c_2, then we *know* that the random variable X *must* have a Normal distribution.)

Posterior Derivation: The Quick Way

For the binomial model with a beta prior, the posterior is

$$\begin{aligned} p(\theta|y) &= \mathbb{P}(y|\theta) \times p(\theta) \\ &= \binom{N}{y}\theta^y(1-\theta)^{N-y} \times \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \end{aligned}$$

but all we need to focus on is the terms in θ:

$$\begin{aligned} p(\theta|y) &\propto \theta^y(1-\theta)^{N-y} \times \theta^{a-1}(1-\theta)^{b-1} \\ &= \theta^{y+a-1}(1-\theta)^{N-y+b-1}. \end{aligned}$$

From this form, we know the posterior **must** be a Beta$(y+a, N-y+b)$ distribution – and so can work out its mean, quantiles etc, just like we did for Beta priors.

This is an example of a **conjugate** Bayesian analysis, in which the prior is in the same family as the posterior.

Agresti and Coull's adjusted interval

Recall, from earlier, the *adjusted Wald interval*:

$$\begin{aligned}\tilde{\theta} &\pm 1.96\sqrt{\tilde{\theta}(1-\tilde{\theta})/N}\ , \text{ where}\\ \tilde{\theta} &= \frac{1}{2}\frac{4}{N+4} + \overline{y}\frac{N}{N+4}.\end{aligned}$$

Notice the link with the adjusted Wald interval for the 0 successes case, the estimate is equal to the posterior mean with a Beta(a, b) prior with $a = b = 2$.

Posterior Summaries

- Reporting a point estimate (e.g. posterior mean, or median) alone is rare
- **Credible intervals** – regions that capture a fixed proportion of the posterior support (usually 95%) are the standard way to describe uncertainty.
- These also permit a form of testing, by reporting whether a 95% interval contain the value $\theta_0 = 0.5$
- A typical way to construct a 90% posterior credible interval (θ_L, θ_U) is to solve

$$
\begin{aligned}
0.05 &= \int_0^{\theta_L} p(\theta|y)\ d\theta \\
0.95 &= \int_0^{\theta_U} p(\theta|y)\ d\theta
\end{aligned}
$$

Posterior Summaries

- The quantiles of a beta are not available in closed form, but are easy to evaluate in `R`:

```
y <- 7; N <- 10; a <- b <- 1
qbeta(c(0.05,0.5,0.95),y+a,N-y+b)
[1] 0.4356258 0.6761955 0.8649245
```

- ...so the posterior median is 0.68 and a 90% credible interval is [0.44,0.86].
- Compare this to the MLE of 0.70 and asymptotic 90% confidence interval of $0.70 \pm 1.645 \times \sqrt{0.7 \times 0.3/10} = [0.46, 0.94]$.

Bayes and Frequentist Estimates for Binomial

Example: $N = 10, y = 0$ gives

$$\tilde{\theta} = \frac{4}{10+4}\frac{1}{2} + \frac{10}{10+4}\overline{y} = \frac{4}{28} = 0.14$$

with adjusted standard error

$$\sqrt{\tilde{\theta}(1-\tilde{\theta})/10} = \sqrt{\frac{4}{28}\left(1 - \frac{24}{28}\right)/10} = 0.11$$

... but 0.14 $\pm$ 1.96$\times$0.11 goes negative! Using Bayes instead with a Beta(2,2) prior for θ:

```
y <- 0; N <- 10; a <- b <- 2; apost <- a+y; bpost <- b+(N-y)
qbeta(p=c(0.025,0.975), apost, bpost)
[1] 0.01920667 0.36029744
```

So a Bayesian 95% credible interval is (0.019,0.36).

A more challenging example, from COVID

Suppose a seroprevalence test is carried out with

- Sensitivity, $\mathbb{P}[$ +ve test | disease $]$ denoted δ and assumed known
- Specificity, $\mathbb{P}[$ -ve test | no disease $]$ denoted γ and assumed known
- True prevalence denoted π – this is what's of interest

We test n people and y are recorded as having the disease. Our inital model is

$$y|p \sim \text{Binomial}(N, p)$$

where p is the probability of a +ve test result, with

$$\begin{aligned} p &= \mathbb{P}(\text{ +ve test }) \\ &= \mathbb{P}(\text{ +ve test } | \text{ disease })\mathbb{P}(\text{ disease }) \\ &\quad + \mathbb{P}(\text{ +ve test } | \text{ no disease })\mathbb{P}(\text{ no disease }) \\ &= \delta\pi + (1-\gamma)(1-\pi) = \pi(\delta + \gamma - 1) + (1-\gamma) \end{aligned}$$

A more challenging example, from COVID

With this binomial model the MLE is (exercise!):

$$\widehat{\pi} = \frac{y - N(1-\gamma)}{N(\delta + \gamma - 1)}.$$

This estimate, and approximate confidence intervals, don't do a good job of avoiding negative prevalences.

A Bayesian model is

$$\begin{aligned} y|\pi &\sim \text{Binomial}(N, \pi(\delta + \gamma - 1) + (1-\gamma)) \\ \pi &\sim \text{Beta}(a, b) \end{aligned}$$

Not conjugate!

However, a simple rejection algorithm (Gelfand & Smith 1992) can be implemented that simulates samples from the posterior $p(\pi|y)$.

A more challenging example, from COVID

We'll use a *rejection algorithm* to generate samples from the posterior. For unknown parameter $\boldsymbol{\theta}$ with likelihood $p(\boldsymbol{y} \mid \widehat{\boldsymbol{\theta}})$ with maximum value $M = p(\boldsymbol{y} \mid \widehat{\boldsymbol{\theta}})$ for MLE $\widehat{\boldsymbol{\theta}}$, the algorithm has two steps:

1. Generate $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$ from the prior

2. Generate $U \sim U(0, 1)$ and if

$$U < \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta})}{M},$$

   accept that $\boldsymbol{\theta}$ – otherwise return to 1.

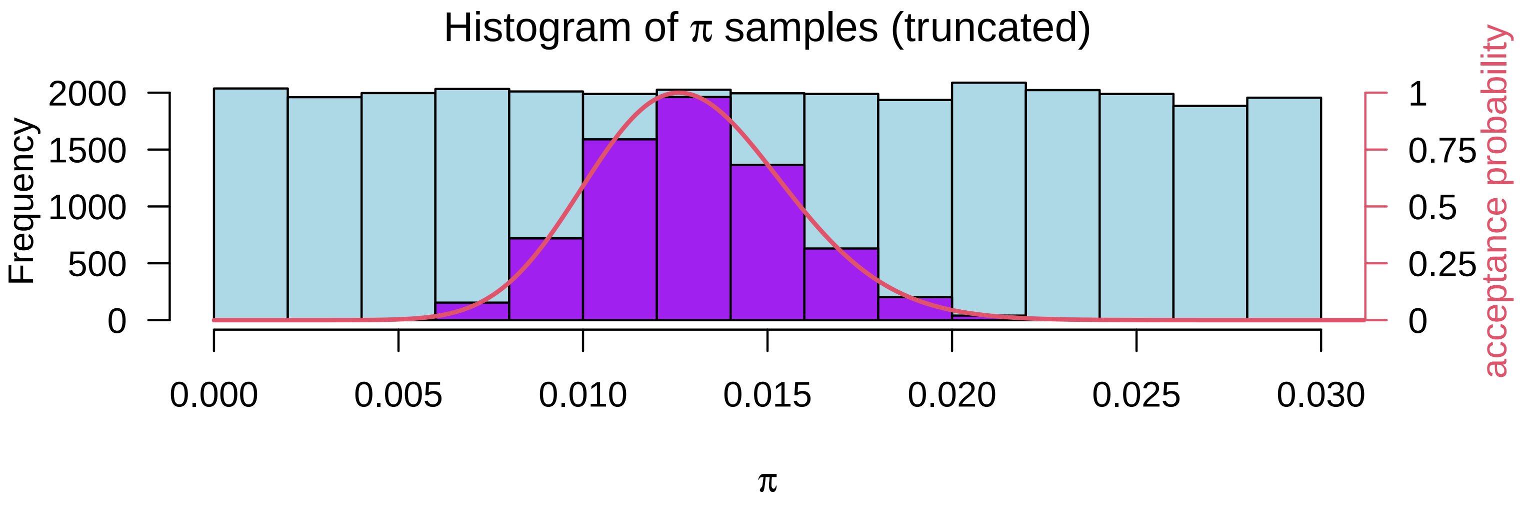
The probability that a point is accepted is given by

$$p_a = \frac{\int p(\boldsymbol{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}{M} = \frac{p(\boldsymbol{y})}{M}.$$

A more challenging example, from COVID

In early April, 2020, Bendavid et al recruited n=3330 residents of Santa Clara County, California and tested them for COVID-19 antibodies. With y =50 positive tests, the naïve estimate is 1.50%. We'll assume sensitivity is $\delta = 0.8$ and specificity is $\gamma = 0.995$, and use a flat prior parameters with $a = b = 1$;

Prior and posterior samples for prevalence π. The posterior median is 1.28% and a 90% interval is (0.87%,1.77%).



See Gelman & Carpenter 2020 for a more comprehensive Bayesian analysis

A more challenging example, from COVID

R code to do the analysis:

```
lik <- function(pi){ dbinom(y, n, pi*(delta+gamma-1) + (1-gamma) ) } # likelihood
M   <- dbinom(y, n, y/n)          # likelihood at MLE

set.seed(4) # random number seed
bigB    <- 1E6 # number of step 1 samples to take
many.pi <- rbeta(bigB, 1,1) # samples from prior
many.u  <- runif(bigB)      # samples from uniform

post.pi <- subset( many.pi, many.u < lik(many.pi)/M ) # evaluation step

# summarize the posterior
length(post.pi)
[1] 6677
quantile(post.pi, c(0.5, 0.05, 0.95))
        50%          5%         95%
0.012841460 0.008695393 0.017657390
```

This method works (eventually!) for any bounded likelihood.

Summary

Conjugate analyses are computationally convenient but rarely available in practice.

Historically, the philosophical standpoint of Bayesian statistics was emphasized, now pragmatism is taking over.

Benefits of a Bayesian approach:

- Inference is based on probability and output is very intuitive
- Framework is flexible, and so complex models can be built
- Can incorporate prior knowledge
- If the sample size is large, prior choice is less crucial (generally!)

Summary

Challenges of a Bayesian analysis:

- Requires a likelihood and a prior, and inference is only as good as the appropriateness of these choices.
- Computation can be daunting, though software is becoming more user-friendly and flexible; later we will describe and illustrate a number of approaches including `INLA` and `Stan`.
- One should be wary of models becoming too elaborate – we have the technology to contemplate complicated models, but do the data support complexity?

Posterior Derivation: The Long Way

- The posterior can also be calculated by keeping in all the normalizing constants:

$$
\begin{aligned}
p(\theta|y) &= \frac{\mathbb{P}(y|\theta) \times p(\theta)}{\mathbb{P}(y)} \\
&= \frac{1}{\mathbb{P}(y)} \begin{pmatrix} N \\ y \end{pmatrix} \theta^y (1-\theta)^{N-y} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}.
\end{aligned}
$$

- The normalizing constant is

$$
\begin{aligned}
\mathbb{P}(y) &= \int_0^1 \mathbb{P}(y|\theta) \times p(\theta) d\theta \\
&= \begin{pmatrix} N \\ y \end{pmatrix} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \theta^{y+a-1}(1-\theta)^{N-y+b-1} d\theta \\
&= \begin{pmatrix} N \\ y \end{pmatrix} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(y+a)\Gamma(N-y+b)}{\Gamma(N+a+b)}
\end{aligned}
$$

- The integrand on line 2 is a Beta$(y+a, N-y+b)$ distribution, up to a normalizing constant, and so we know what this constant has to be.

Posterior Derivation: The Long Way

- The normalizing constant is therefore:

$$\mathbb{P}(y) = \binom{N}{y} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(y+a)\Gamma(N-y+b)}{\Gamma(N+a+b)}$$

- This is a probability distribution, i.e. $\sum_{y=0}^{N} \mathbb{P}(y) = 1$ with $\mathbb{P}(y) > 0$, for $y = 0, 1, \ldots, N$.
- For a particular y value, this expression tells us the probability of that value given the model, i.e. the likelihood and prior we have selected: this will reappear later in the context of hypothesis testing.
- Substitution of $\mathbb{P}(y)$ into (1) and canceling the terms that appear in the numerator and denominator gives the posterior:

$$p(\theta|y) = \frac{\Gamma(N+a+b)}{\Gamma(y+a)\Gamma(N-y+b)} \theta^{y+a-1}(1-\theta)^{N-y+b-1}$$

which is a Beta$(y+a, N-y+b)$.

The Posterior Mean: A Summary of the Posterior

- Recall the mean of a Beta(a, b) is $a/(a+b)$.
- The posterior mean of a Beta$(y+a, N-y+b)$ is therefore

$$
\begin{aligned}
\mathbb{E}[\theta|y] &= \frac{y+a}{N+a+b} \\
&= \frac{y}{N+a+b} + \frac{a}{N+a+b} \\
&= \frac{y}{N} \times \frac{N}{N+a+b} + \frac{a}{a+b} \times \frac{a+b}{N+a+b} \\
&= \text{MLE} \times \text{W} + \text{Prior Mean} \times \text{(1-W)}.
\end{aligned}
$$

- The weight W is

$$
\text{W} = \frac{N}{N+a+b}.
$$

- As N increases, the weight tends to 1, so that the posterior mean gets closer and closer to the MLE.

The Posterior Mean: A Summary of the Posterior

- Notice that the uniform prior $a = b = 1$ gives a posterior mean of

$$\mathbb{E}[\theta|y] = \frac{y+1}{N+2}.$$