Lecture 2.
Discrete Bayesian inference and conjugate distributions for proportions and Poisson means

Summary

1. Bayesian inference for discrete parameters - binomial example
2. Bayesian direct probability statements about parameters
3. Prior distributions for proportions
4. Predictions for binomial data
5. Prior distributions for Poisson means
6. Predictions for Poisson data

Bayesian inference for discrete parameters

- Have observable quantities y, *i.e.* the data
- Unknown quantity θ taking on one of a discrete set of values $\theta_i, i = 1, .., I$
- Specify a sampling model $p(y \mid \theta)$
- Specify a **prior distribution** $p(\theta_i)$
- Together define $p(y, \theta_i) = p(y|\theta_i)p(\theta_i)$: a **'full probability model'**

Then use Bayes theorem to obtain conditional probability distribution for unobserved quantities of interest given the data:

$$p(\theta_i \mid y) = \frac{p(y \mid \theta_i)\, p(\theta_i)}{\sum_k p(y \mid \theta_k)\, p(\theta_k)} \propto p(y \mid \theta_i)\, p(\theta_i)$$

when considering $p(y \mid \theta_i)$ as a function of θ_i: ie the *likelihood*.

$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

Inference about a discrete parameter

Suppose I have 3 coins in my pocket,

1. *biased 3:1 in favour of heads*
2. *a fair coin,*
3. *biased 3:1 in favour of tails*

I randomly select one coin and toss it once, observing a head. What is the probability that I have chosen coin 3?

- Let $y = 1$ denote the event that I observe a head
- θ denote the probability of a head: $\theta \in (0.25, 0.5, 0.75)$
- Prior: $p(\theta = 0.25) = p(\theta = 0.5) = p(\theta = 0.75) = 0.33$
- Sampling distribution: $p(y|\theta) = \theta^y (1 - \theta)^{(1-y)}$

Coin	θ	Prior $p(\theta)$	Likelihood $p(y=1\|\theta)$	Un-normalised Posterior $p(y=1\|\theta)p(\theta)$	Normalised Posterior $\frac{p(y=1\|\theta)p(\theta)}{p(y)^{\dagger}}$
1	0.25	0.33	0.25	0.0825	0.167
2	0.50	0.33	0.50	0.1650	0.333
3	0.75	0.33	0.75	0.2475	0.500
	Sum	1.00	1.50	0.495	1.000

† The normalising constant can be calculated as $p(y) = \sum_i p(y|\theta_i)p(\theta_i)$

So observing a head on a single toss of the coin means that there is now a 50% probability that the chance of heads is 0.75 and only a 16.7% probability that the chance of heads in 0.25.

Bayesian inference - how did it all start?

In 1763, Reverend Thomas Bayes of Tunbridge Wells wrote

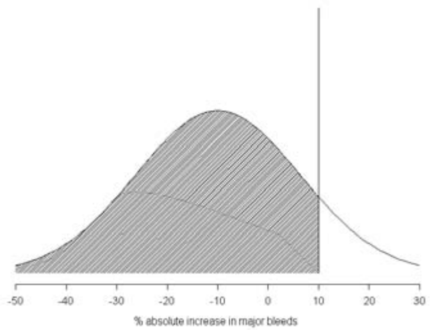
PROBLEM.

Given the number of times in which an unknown event has happened and failed: *Required* the chance that the probability of its happening in a fingle trial lies fomewhere between any two degrees of probability that can be named.

In modern language, given $r \sim$ Binomial(θ, n),
what is Pr$(\theta_1 < \theta < \theta_2 | r, n)$?

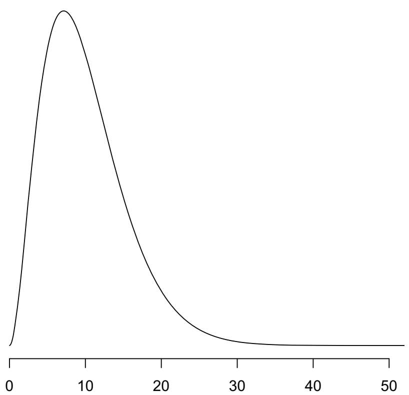
Basic idea: Direct expression of uncertainty about unknown parameters

eg "There is an 89% probability that the absolute increase in major bleeds is less than 10 percent with low-dose PLT transfusions" (Tinmouth et al, 2004)



Example: surgical

- Suppose a hospital is considering a new high-risk operation
- Experience in other hospitals indicate that the risk θ for each patient is expected to be around 10%
- it would be fairly surprising (all else being equal) to be less than 3% or more than 20%



Probability that mortality risk above 15% = $\Pr(\theta > 0.15) = 0.17$
$\Pr(0.03 < \theta < 0.20) = 0.90)$

Why a direct probability distribution?

1. Tells us what we want: what are plausible values for the parameter of interest?
2. No *P-values*: just calculate relevant tail areas
3. No (difficult to interpret) *confidence intervals*: just report, say, central area that contains 95% of distribution
4. Easy to make predictions (see later)
5. Fits naturally into decision analysis / cost-effectiveness analysis / project prioritisation
6. There is a procedure for adapting the distribution in the light of additional evidence: i.e. *Bayes theorem* allows us to learn from experience

And what about disadvantages?

1. Requires the specification of what we thought *before* new evidence is taken into account: the *prior* distribution
2. Explicit allowance for quantitative subjective judgement in the analysis,
3. Analysis may be more complex than a traditional approach
4. Computation may be more difficult
5. Currently no established standards for Bayesian reporting

- Vital importance of *accountability* to scientific community, journal editors, policy-makers, FDA etc etc.
- Strong need for *transparency* - not easy given complexity!

Inference on proportions

What is a reasonable form for a prior distribution for a proportion?

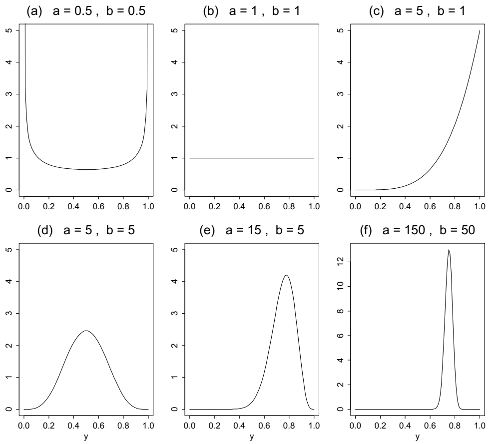
$\theta \sim \text{Beta}[a, b]$ represents a beta distribution with properties:

$$\begin{aligned} p(\theta|a,b) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\,\theta^{a-1}\,(1-\theta)^{b-1}; \quad \theta \in (0,1) \\ \mathrm{E}(\theta|a,b) &= \frac{a}{a+b} \\ \mathrm{V}(\theta|a,b) &= \frac{ab}{(a+b)^2(a+b+1)}: \end{aligned}$$

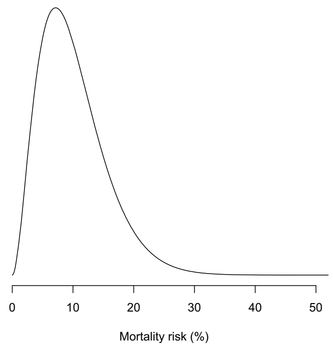
$$\Rightarrow \int_0^1 \theta^{a-1}\,(1-\theta)^{b-1}\,d\theta = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

$(\Gamma(a) = (a-1)!$ if a integer; $\Gamma(1) = 0! = 1)$

Beta distributions



Surgical example



- A Beta[3,27] proportional to $\theta^2(1-\theta)^{26}$
- Mean = 3/(3+27) = 0.1, standard deviation 0.054, variance 0.003, median 0.091, mode 0.071.
- An equi-tailed 90% interval is (0.03, 0.20) which has width 0.17, but a narrower 'Highest posterior density' interval is (0.02, 0.18) with width 0.16

Predictions with unknown parameters

- Suppose we assume a parametric sampling distribution $p(y|\theta)$
- willing to express our uncertainty about the parameter θ as a distribution $p(\theta)$
- before observing a future quantity Y, we can integrate out the unknown parameter to produce a predictive distribution
$$p(y) = \int p(y|\theta)p(\theta)d\theta :$$
- for discrete parameter distributions this takes the form
$$p(y) = \sum_i p(y|\theta_i)p(\theta_i).$$
- Such predictions are useful in, for example, cost-effectiveness models, design of studies, checking whether observed data is compatible with expectations, and so on

The mean and variance of a predictive distribution can be obtained using standard formulae:

$$\begin{aligned} \boldsymbol{E}[Y] &= \boldsymbol{E}_\Theta[\boldsymbol{E}[Y|\theta]] \\ \boldsymbol{V}[Y] &= \boldsymbol{E}_\Theta[\boldsymbol{V}[Y|\theta]] + \boldsymbol{V}_\Theta[\boldsymbol{E}[Y|\theta]] \end{aligned}$$

Θ subscripts emphasises when the expectation or variance is with respect with the distribution for θ

In certain cases we can obtain an algebraic expression for the predictive distribution.

Standard identities

For 2 random variables X and Y with joint distribution $p(x,y)$, then

$$\boldsymbol{E}[Y] = \boldsymbol{E}_X[\boldsymbol{E}[Y|x]]; \quad \boldsymbol{V}[Y] = \boldsymbol{E}_X[\boldsymbol{V}[Y|x]] + \boldsymbol{V}_X[\boldsymbol{E}[Y|x]]$$

Proof (assuming regularity conditions to reverse order of integration)

$$\begin{aligned}\boldsymbol{E}[Y] &= \int_Y y\, p(y)dy = \int_Y \int_X y\,[p(y|x)p(x)dx]\,dy \\ &= \int_X \left[\int_Y y\, p(y|x)dy\right] p(x)dx = \boldsymbol{E}_X[\boldsymbol{E}[Y|x]].\end{aligned}$$

$$\begin{aligned}\boldsymbol{V}[Y] &= \int_Y (y - \boldsymbol{E}[Y])^2 p(y)dy = \int_X \left[\int_Y (y-\boldsymbol{E}[Y])^2 p(y|x)dy\right] p(x)dx \\ &= \int_X \left[\int_Y (y - \boldsymbol{E}[Y|x] + \boldsymbol{E}[Y|x] - \boldsymbol{E}[Y])^2 p(y|x)dy\right] p(x)dx \\ &= \int_X \left[\int_Y (y - \boldsymbol{E}[Y|x])^2 p(y|x)dy\right] p(x)dx + \int_X (\boldsymbol{E}[Y|x] - \boldsymbol{E}[Y])^2 \left[\int_Y p(y|x)dy\right] p(x)dx \\ &= \boldsymbol{E}_X[\boldsymbol{V}[Y|x]] + \boldsymbol{V}_X[\boldsymbol{E}[Y|x]]\end{aligned}$$

Predictions for Binomial data

Suppose

$$\begin{aligned}\theta &\sim \text{Beta}(a,b) \\ Y &\sim \text{Binomial}(\theta, n).\end{aligned}$$

The exact predictive distribution for Y is known as the **Beta-Binomial** with

$$p(y) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \binom{n}{y} \frac{\Gamma(a+y)\Gamma(b+n-y)}{\Gamma(a+b+n)}.$$

If $a = b = 1$, *i.e.* the prior distribution is uniform between 0 and 1, $p(y)$ is uniform over 0,1,...,n

(This was the noted by Bayes in 1761)

$$\boldsymbol{E}[Y] = \boldsymbol{E}_\Theta[\boldsymbol{E}[Y|\theta]] = \boldsymbol{E}_\Theta[n\theta] = n\frac{a}{a+b}$$

$$\begin{aligned}\boldsymbol{V}[Y] &= \boldsymbol{E}_\Theta[\boldsymbol{V}[Y|\theta]] + \boldsymbol{V}_\Theta[\boldsymbol{E}[Y|\theta]] \\ &= \boldsymbol{E}_\Theta[n\theta(1-\theta)] + \boldsymbol{V}_\Theta[n\theta] = n\frac{ab}{(a+b)}\frac{(n+a+b)}{(1+a+b)}\end{aligned}$$

Surgical: continued. Suppose our hospital was going to do 20 operations next year - how many deaths might we expect, and what is the chance there will be at least 6 deaths?

Let Y be the number of deaths next year

$\theta \sim$ Beta$(3, 27)$ and $Y \sim$ Binomial$(\theta, 20)$ and so Y is beta-binomial with mean $0.1 \times 20 = 2$, variance 2.90 and standard deviation 1.70

We can also calculate $\Pr(Y \geq 6) = 0.04$

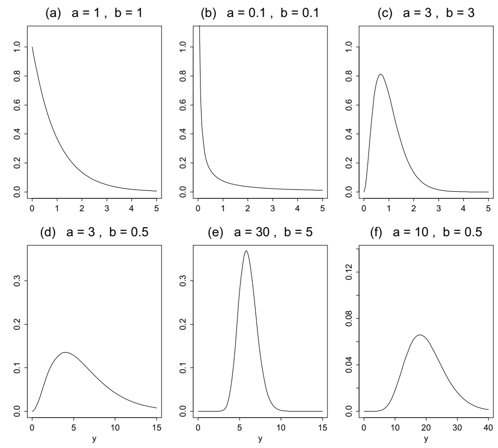
The gamma distribution

Flexible distribution for positive quantities. If $Y \sim \text{Gamma}[a, b]$

$$
\begin{aligned}
p(y|a,b) &= \frac{b^a}{\Gamma(a)} y^{a-1}\, e^{-by}; \quad y \in (0, \infty) \\
\mathsf{E}(Y|a,b) &= \frac{a}{b} \\
\mathsf{V}(Y|a,b) &= \frac{a}{b^2}.
\end{aligned}
$$

WinBUGS notation: `y ~ dgamma(a,b)`

Gamma distributions



- Gamma[1,b] distribution is exponential with mean $1/b$
- Gamma[$\frac{v}{2}, \frac{1}{2}$] is a Chi-squared χ^2_v distribution on v degrees of freedom
- $Y \sim$ Gamma(ϵ, ϵ) approximates $p(y) \propto 1/y$, or that $\log Y \approx$ Uniform
- Used as conjugate prior distribution for Poisson means and inverse variances (precisions)
- Used as sampling distribution for skewed positive valued quantities (alternative to log normal) — MLE of mean is sample mean

Predictions for Poisson data

Suppose

$$
\begin{aligned}
\theta &\sim \text{Gamma}(a, b) \\
Y &\sim \text{Poisson}(\theta).
\end{aligned}
$$

The exact predictive distribution for Y is known as the **Negative-Binomial** with

$$p(y) = \frac{\Gamma(a+y)}{\Gamma(a)\Gamma(y+1)} \frac{b^a}{(b+1)^{a+y}}.$$

$$\boldsymbol{E}[Y] = \boldsymbol{E}_{\Theta}[\boldsymbol{E}[Y|\theta]] = \boldsymbol{E}_{\Theta}[\theta] = \frac{a}{b}$$

$$\begin{aligned}\boldsymbol{V}[Y] &= \boldsymbol{E}_{\Theta}[\boldsymbol{V}[Y|\theta]] + \boldsymbol{V}_{\Theta}[\boldsymbol{E}[Y|\theta]] \\ &= \boldsymbol{E}_{\Theta}[\theta] + \boldsymbol{V}_{\Theta}[\theta] = \frac{a}{b} + \frac{a}{b^2} = \frac{a(b+1)}{b^2}\end{aligned}$$