One-parameter Models
Part1 - discrete observation
(part of chapter 1& part of chapter3)

Outline

- I). Binomial distribution
- II). Beta distribution
- III). Estimating the Binomial proportion
- IV). Poisson distribution
- V). Posterior Inference for Poisson mean

I). Binomial Distribution

- 5 Conditions that need to be met:
  - 1. There is a series of N trials
  - 2. For each trial there are only two possible outcomes
  - 3. For each trial, the two outcomes are mutually exclusive
  - 4. Independence between the outcomes of each trial
  - 5. The probability for each outcome, stays the same from trial to trial

Example: Flipping a Coin

- Each flip is a particular trial
- Only 2 possible outcomes
- Mutually exclusive – only a head or tail can occur
- Independent – outcome of one flip doesn't affect any other flips

Binomial example

Take the example of 5 coin tosses. What's the probability that you flip exactly 3 heads in 5 coin tosses?

Binomial distribution

Solution:

One way to get exactly 3 heads: HHHTT

What's the probability of this exact arrangement?

P(heads)xP(heads) xP(heads)xP(tails)xP(tails)
$=(1/2)^3$ *x* $(1/2)^2$

Another way to get exactly 3 heads: THHHT

Probability of this exact outcome = $(1/2)^1$ *x* $(1/2)^3$ *x* $(1/2)^1$ = $(1/2)^3$ *x* $(1/2)^2$

Binomial distribution

In fact, $(1/2)^3$ *x* $(1/2)^2$ is the probability of each unique outcome that has exactly 3 heads and 2 tails.

So, the overall probability of 3 heads and 2 tails is:

$(1/2)^3 \; x \; (1/2)^2 + (1/2)^3 \; x \; (1/2)^2 + (1/2)^3 \; x \; (1/2)^2$ + ….. for as many unique arrangements as there are—but how many are there??

$\binom{5}{3}$ ways to arrange 3 heads in 5 trials

$= 5!/3!(5-3)! = 10$

Outcome	Probability
THHHT	$(1/2)^3 \times (1/2)^2$
HHHTT	$(1/2)^3 \times (1/2)^2$
TTHHH	$(1/2)^3 \times (1/2)^2$
HTTHH	$(1/2)^3 \times (1/2)^2$
HHTTH	$(1/2)^3 \times (1/2)^2$
THTHH	$(1/2)^3 \times (1/2)^2$
HTHTH	$(1/2)^3 \times (1/2)^2$
HHTHT	$(1/2)^3 \times (1/2)^2$
THHTH	$(1/2)^3 \times (1/2)^2$
HTHHT	$(1/2)^3 \times (1/2)^2$

10 arrangements $\times (1/2)^3 \times (1/2)^2$

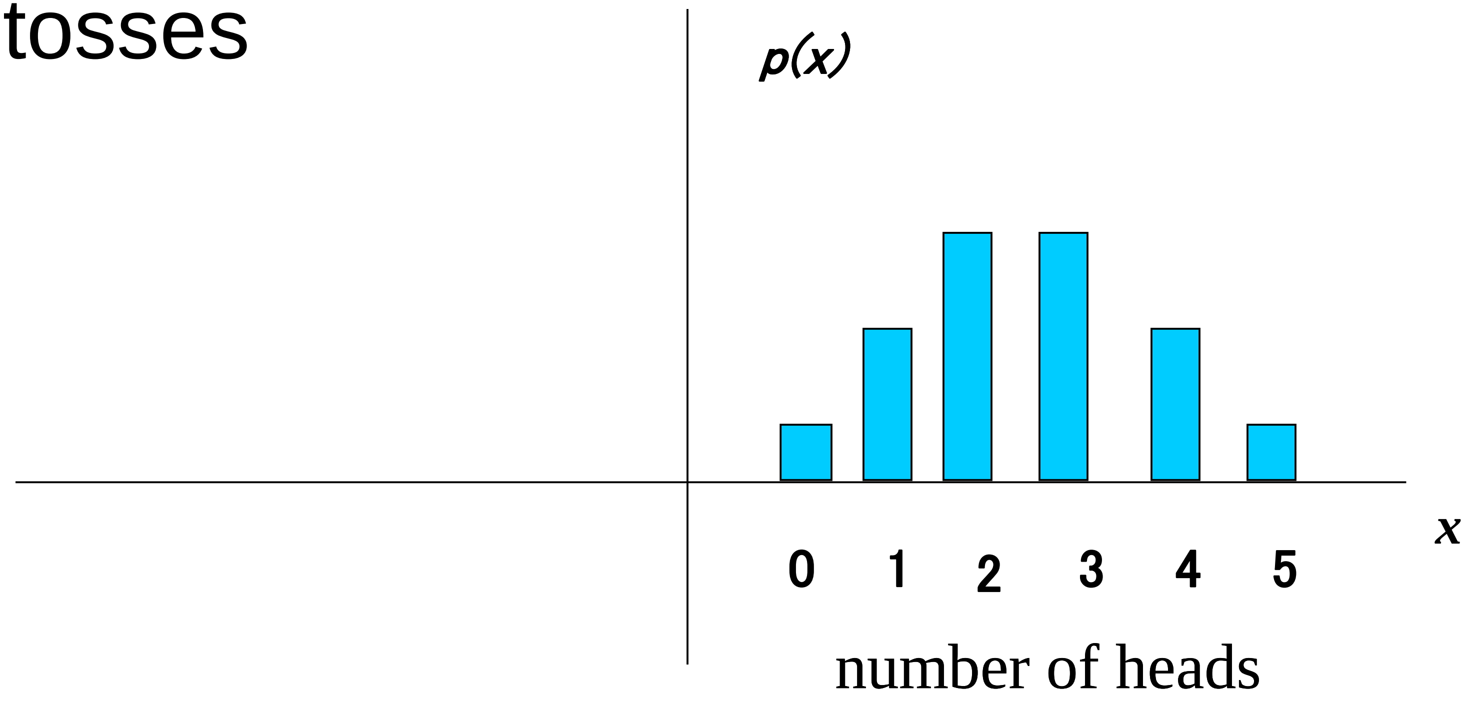
The probability of each unique outcome (note: they are all equal)

$\therefore$ **P(3 heads and 2 tails) =** $\binom{5}{3}$ ***x P(heads)^3 x P(tails)^2 =***

$$10 \times (½)^{5} = 31.25\%$$

Binomial distribution function:

X= the number of heads tossed in 5 coin tosses



Binomial distribution, generally

Note the general pattern emerging → if you have only two possible outcomes (call them 1/0 or yes/no or success/failure) in *n* independent trials, then the probability of exactly *X* "successes"=

n = number of trials

$$\binom{n}{X} p^{X} (1-p)^{n-X}$$

1-p = probability of failure

X = # successes out of *n* trials

p = probability of success

Definitions: Binomial

- **Binomial:** Suppose that *n* independent experiments, or trials, are performed, where *n* is a fixed number, and that each experiment results in a "success" with probability *p* and a "failure" with probability *1-p*. The total number of successes, *X*, is a binomial random variable with parameters *n* and *p*.
- We write: ***X ~ Bin (n, p)*** *{reads: "X is distributed binomially with parameters n and p}*
- And the probability that $X=r$ (i.e., that there are <u>exactly</u> *r* successes) is:

$$\boxed{P(X=r)=\binom{n}{r}p^{r}(1-p)^{n-r}}$$

Newton's Binomial Theorem

$$(a + b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{(n-k)}$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Special case: Bernouilli

Bernouilli trial: If there is only 1 trial with probability of success *p* and probability of failure *1-p*, this is called a Bernouilli distribution. (special case of the binomial with *n*=1)

Probability of success:

$$P(X=1)=\binom{1}{1}p^{1}(1-p)^{1-1}=p$$

Probability of failure:

$$P(X=0)=\binom{1}{0}p^{0}(1-p)^{1-0}=1-p$$

Binomial distribution: example

- If I toss a coin 20 times, what's the probability of getting exactly 10 heads?

$$\binom{20}{10}(.5)^{10}(.5)^{10} = .176$$

Binomial distribution: example

- If I toss a coin 20 times, what's the probability of getting of getting 2 or fewer heads?

$$\binom{20}{0}(.5)^0(.5)^{20} = \frac{20!}{20!0!}(.5)^{20} = 9.5x10^{-7} +$$

$$\binom{20}{1}(.5)^1(.5)^{19} = \frac{20!}{19!1!}(.5)^{20} = 20x9.5x10^{-7} = 1.9x10^{-5} +$$

$$\binom{20}{2}(.5)^2(.5)^{18} = \frac{20!}{18!2!}(.5)^{20} = 190x9.5x10^{-7} = 1.8x10^{-4}$$

$$= 2.0x10^{-4}$$

Two main characteristics of probability distributions: expected value and variance:

If X follows a binomial distribution with parameters n and p: $\boldsymbol{X \sim Bin\ (n, p)}$

Then:

$\mu_X = E(X) = np$

$\sigma_X^2 = Var\ (X) = np(1-p)$

$\sigma_X = SD\ (X) = \sqrt{np(1-p)}$

Note: the variance will always lie between

0*N and .25 *N

Because:

p(1-p) reaches maximum at p=.5

=> p(1-p)=.25

Binomial mean and variance

$$\mu_X = E(X) = np$$

$$\sigma_X^2 = Var(X) = np(1-p)$$

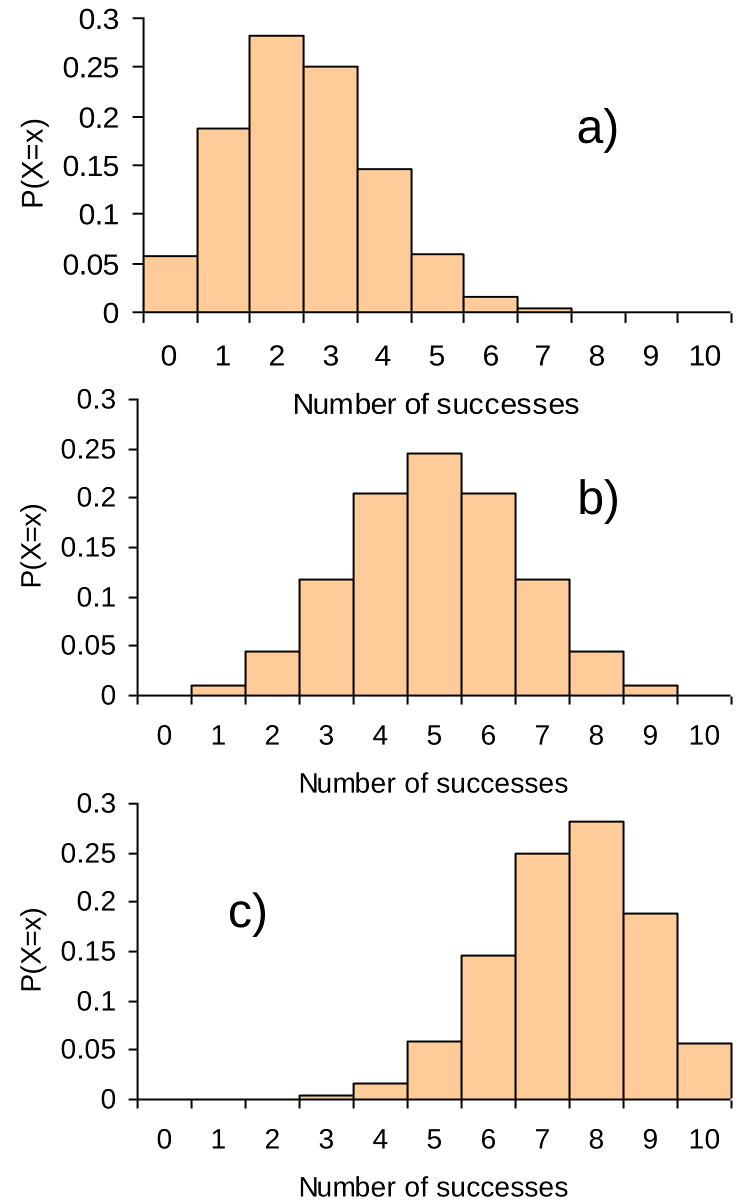
Effect of changing *p* when *n* is fixed.

a) $n = 10$, $p = 0.25$

b) $n = 10$, $p = 0.5$

c) $n = 10$, $p = 0.75$

For small samples, binomial distributions are skewed when p is different from 0.5.



Characteristics of Bernouilli distribution

For Bernouilli ($n=1$)

$E(X) = p$

$Var(X) = p(1-p)$

Variance Proof

For Y~Bernouilli (p)

$$\begin{cases} Y=1 \text{ if yes} \\ Y=0 \text{ if no} \end{cases}$$

$$Var(Y) = E(Y^2) - E(Y)^2$$

$$= [1^2 p + 0^2 (1-p)] - [1p + 0(1-p)]^2$$

$$= p - p^2$$

$$= p(1-p)$$

For X~Bin (N,p)

$$X = \sum_{i=1}^{n} Y_{Bernouilli}; Var(Y) = p(1-p)$$

$$= Var(X) = Var(\sum_{i=1}^{n} Y) = \sum_{i=1}^{n} Var(Y) = np(1-p)$$

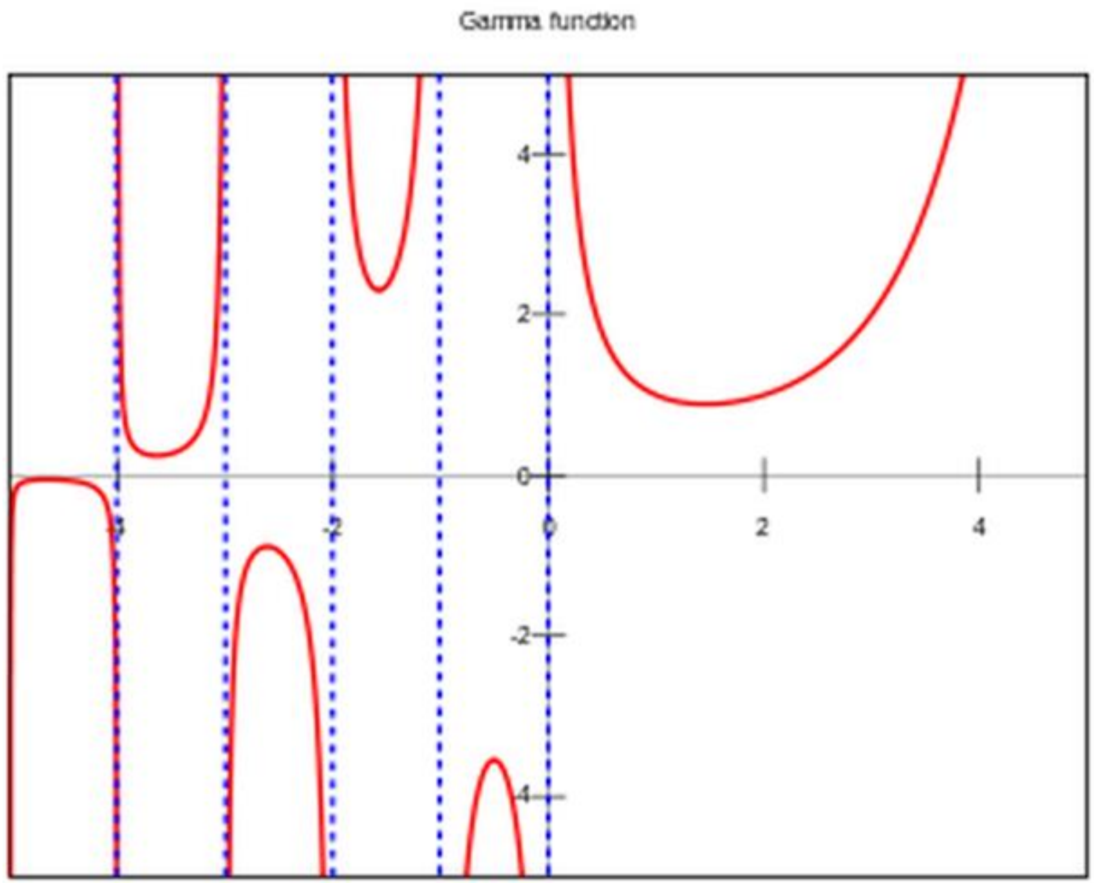
II). Beta Distribution

Gamma function

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x}\, dx.$$

$$\Gamma(t+1) = t\Gamma(t)$$

$$\Gamma(n) = (n-1)!$$



Beta Distribution

- Distribution over $\mu \in [0, 1]$.

$$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$

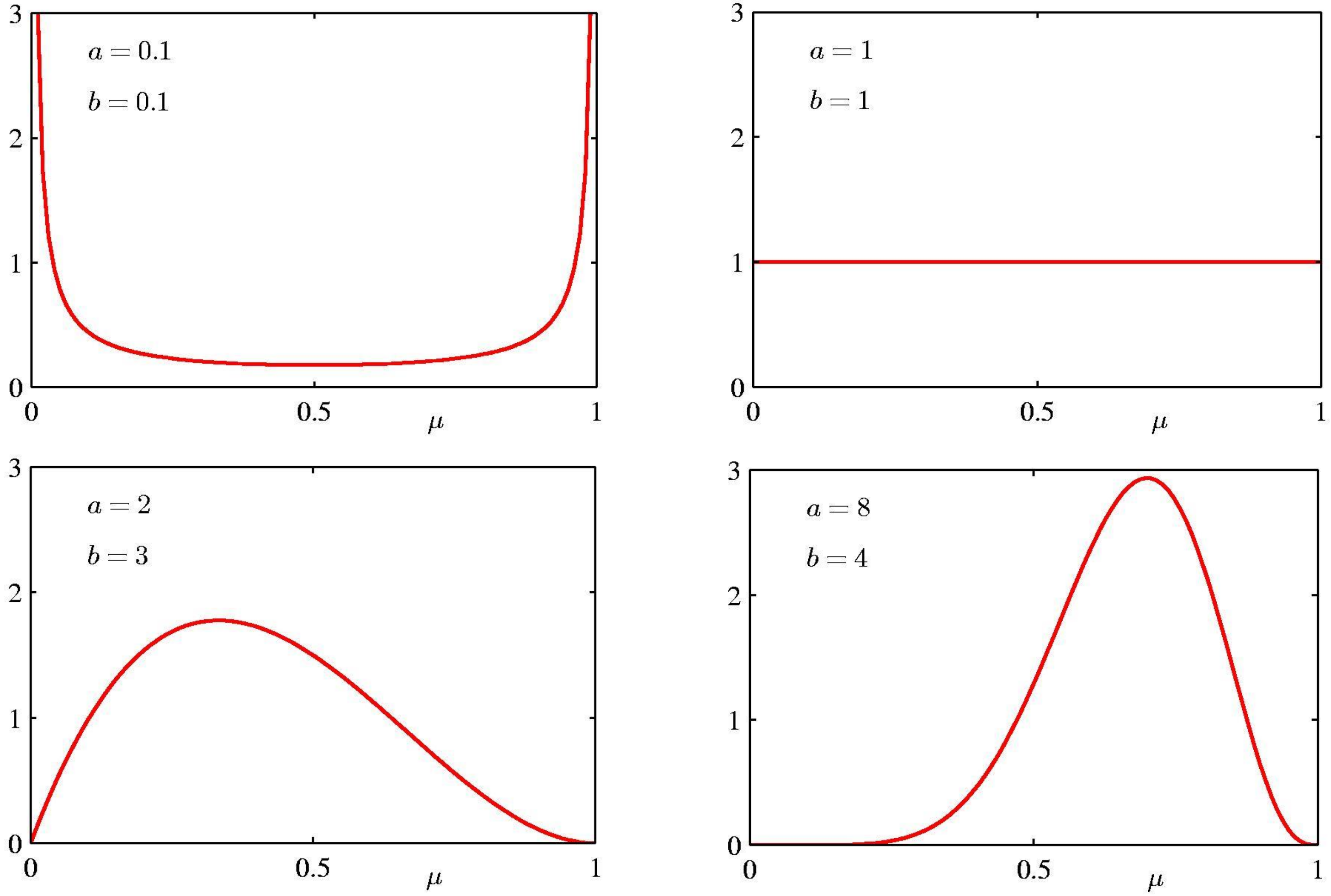
$$\mathbb{E}[\mu] = \frac{a}{a+b}$$

$$\mathrm{var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)}$$

$$\mathrm{mode}\,[\mu] = \frac{a-1}{a+b-2}$$ for a>1 and b>1

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

Beta Distribution



III). Estimating the Binomial proportion

- Given information: we tossed a loaded coin for N times, and we observed m Heads
- Task: what is the probability that this coin gives the Head side?

III.a) Frequentist's Parameter Estimation: Maximum Likelihood Estimation (MLE)

$$\mathcal{D} = \{x_1, \ldots, x_N\}, \; m \text{ heads } (1), \; N - m \text{ tails } (0)$$

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n} (1-\mu)^{1-x_n}$$

$$\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} \ln p(x_n|\mu) = \sum_{n=1}^{N} \{x_n \ln \mu + (1 - x_n) \ln(1-\mu)\}$$

The value to maximize the likelihood/log-likelihood is:

$$\mu_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} x_n = \frac{m}{N}$$

Problem of MLE:

- Example: $\mathcal{D} = \{1, 1, 1\} \rightarrow \mu_{\mathrm{ML}} = \frac{3}{3} = 1$
- Prediction: future tosses will always land heads up
  - =>Overfitting to D

- Example: $\mathcal{D} = \{0,0,0\} \rightarrow \mu_{\mathrm{ML}} = \frac{0}{3} = 0$
- Prediction: future tosses will never land heads up
  - =>Overfitting to D

Problem of MLE:

- **Central Limit Theorem** (CLT)

Draw a SRS of size n from a population which has mean μ

and standard deviation σ

Even if the population is not normal distribution but n is large, then we still approximately have:

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

- **In the proportion case**

$$\bar{X} = \frac{m}{N} \sim N(p, \frac{\sqrt{Np(1-p)}}{\sqrt{N}})$$

Thus the 95% Confidence Interval is: $\frac{m}{N} \pm 1.96\sqrt{\frac{\frac{m}{N}(1-\frac{m}{N})}{N}} = (0,0)$

III.b)

Bayesian estimation of the proportion

A real example with details

Setting

- We are interested in the prevalence of an infectious disease in a small city.

- The higher the prevalence, the more public health precautions we would recommend be put into place.

- A small random sample of 20 individuals from the city will be checked for infection.

Parameter and sample spaces

- θ: The fraction of infected individuals in the city which we are interested in.
- Parameter space includes all numbers between zero and one: $\Theta=[0,1]$

- y: The total number of people in the sample who are infected.
- Sample spaces: $y=\{0,1,\ldots,20\}$

III.b.1). Sampling Model provides the likelihood

- Before the sample is obtained the number of infected individuals in the sample is unknown.
- Let the variable Y denote this to-be-determined value.
- If the value of θ were known sampling model for Y would be a binomial(20,θ) probability distribution:

$$Y|\theta \sim \text{binomial}(20,\theta)$$

i.e. If the true infection rate θ is 0.05, then the probability that there will be zero infected individuals in the sample (Y = 0) is 36%.

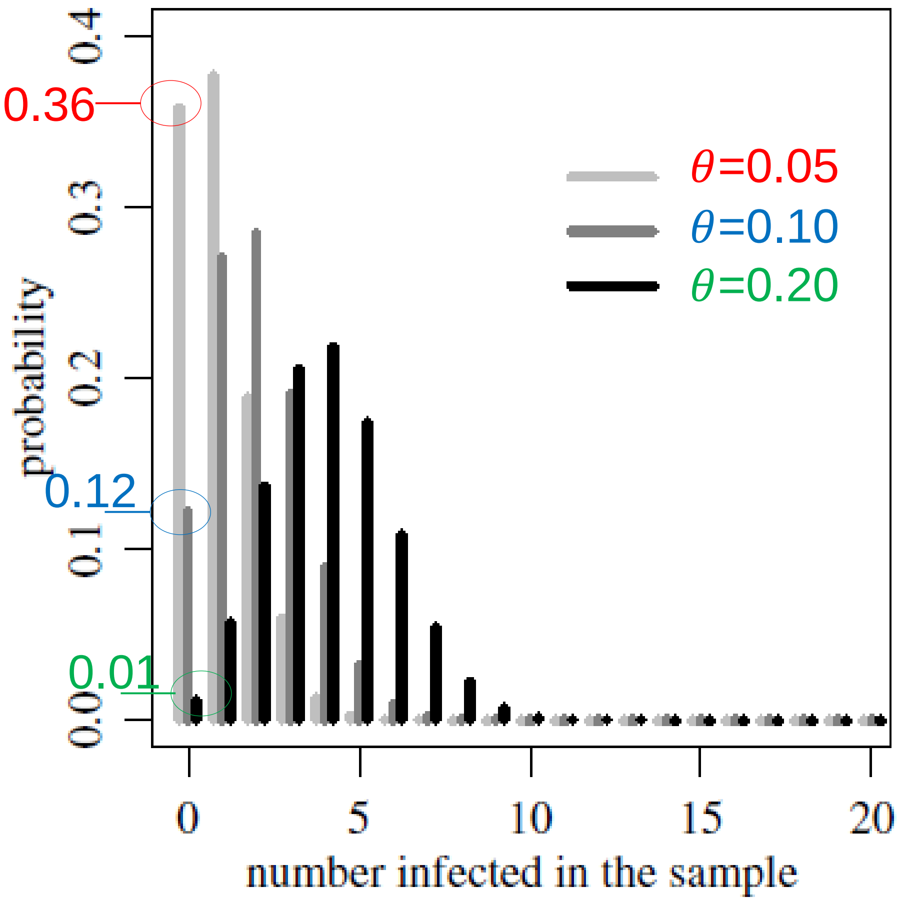


Figure: binomial(20,θ) distribution for 3 values for θ

III.b.2). **Prior distribution**

- Other studies from various parts of the country indicate that the infection rate in comparable cities ranges from about 0.05 to 0.20, with an average prevalence of 0.10.

- This prior information suggests that we use a prior distribution p(θ) that assigns a substantial amount of probability to the interval (0.05, 0.20), and that the expected value of under p(θ) is close to 0.10.

Prior distribution

- However, there are infinitely many probability distributions that satisfy these conditions
- It is **not clear** that we can discriminate among them with our limited amount of prior information.
- We will therefore use a prior distribution p(θ) that has the characteristics described above, but whose particular mathematical form is chosen for reasons of computational convenience.
- Specifically, we will encode the prior information using a member of the family of Beta distributions

Beta distribution

- A beta distribution has two parameters which we denote as a and b .
- If θ has a beta(a, b) distribution, then the expectation of θ : $E[\theta]=\frac{a}{(a+b)}$ and
- the most probable value of θ :
  $\mathrm{mode}[\theta]=\frac{(a-1)}{(a-1+b-1)}$

- For our problem where θ is the infection rate, we will represent our prior information about θ with a beta(2,20) probability distribution. i.e. θ ~Beta(2,20)

Prior distribution

- The expected value of θ for this prior distribution : E[θ]=$\frac{2}{(2+20)}$ = 0.09

- The curve of the prior distribution is highest at mode[θ]=$\frac{(2-1)}{(2-1+20-1)}$= 0.05

- About two-thirds of the area under the curve occurs between 0.05 and 0.20.
  Pr($0.05 < \theta$<0.20)=0.66

- The prior probability that the infection rate is below 0.10 is 64%.
  Pr(θ<0.10)=0.64

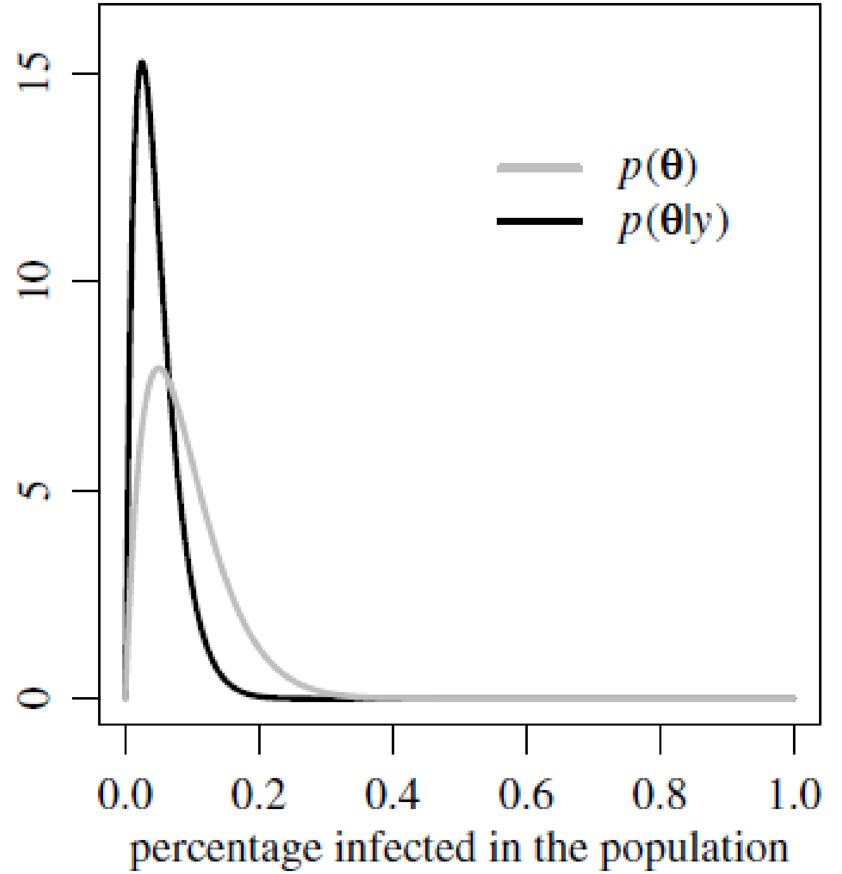


Figure: gray line show the prior distribution

III.b.3). **the posterior**

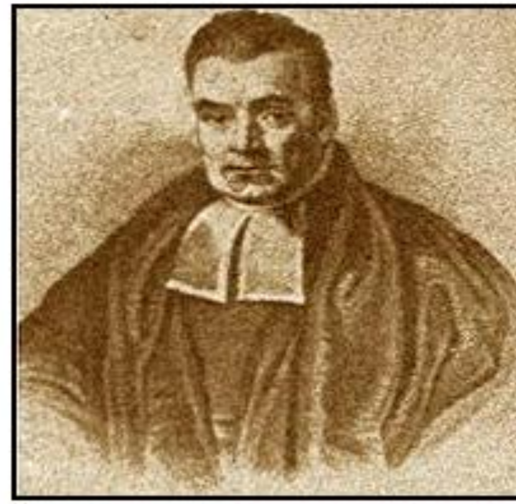
- Bayes' formula

$$P(\theta \mid Y) = \frac{P(Y, \theta)}{P(Y)} = \frac{P(Y \mid \theta) P(\theta)}{P(Y)}$$

- Bayesian statistics

  prior + evidence => posterior probability

- Bayesian estimation



Reverend Thomas Bayes 1702-1761

$$p(\theta \mid Y) \propto p(\theta)\, p(Y \mid \theta)$$

Posterior distribution

Prior distributio n

Likelihood

Deriving the posterior

prior (Beta(a,b))

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

likelihood (binomial)

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

posterior

$$p(\theta \mid y) = C \cdot \theta^{a-1}(1-\theta)^{b-1}\theta^{y}(1-\theta)^{n-y}$$

$$\propto \theta^{a+y-1}(1-\theta)^{b+n-y-1} = Beta(a+y, b+n-y)$$

Posterior distribution

- If $Y|\theta$ ~ binomial(n,θ) and θ~beta(a,b), when we observe a numeric value y of Y , the posterior distribution is a Beta(a+y, b+n−y) distribution

- For out study, suppose a value of Y=0 is observed, i.e. none of the sample individuals are infected.
- The posterior distribution of is then a beta(2, 40) distribution.
  $\theta|\{Y=0\}$ ~beta(2,40)

Posterior distribution

- The density of posterior distribution is further to the left than the prior distribution, and more peaked as well.

- It is to the left of p(θ) because the observation that Y = 0 provides evidence of a low value of θ .

- It is more peaked than p(θ) because it combines information from the data and the prior distribution, and thus contains more information than in p(θ) alone.

- The peak of this curve is mode[θ|Y=0]= 0.025
- The posterior expectation E[$\theta | \mathrm{Y} = 0$] = 0.048.
- The posterior probability that $\theta < 0.10$ is 93%. Pr(θ <0.10|Y=0)=0.93

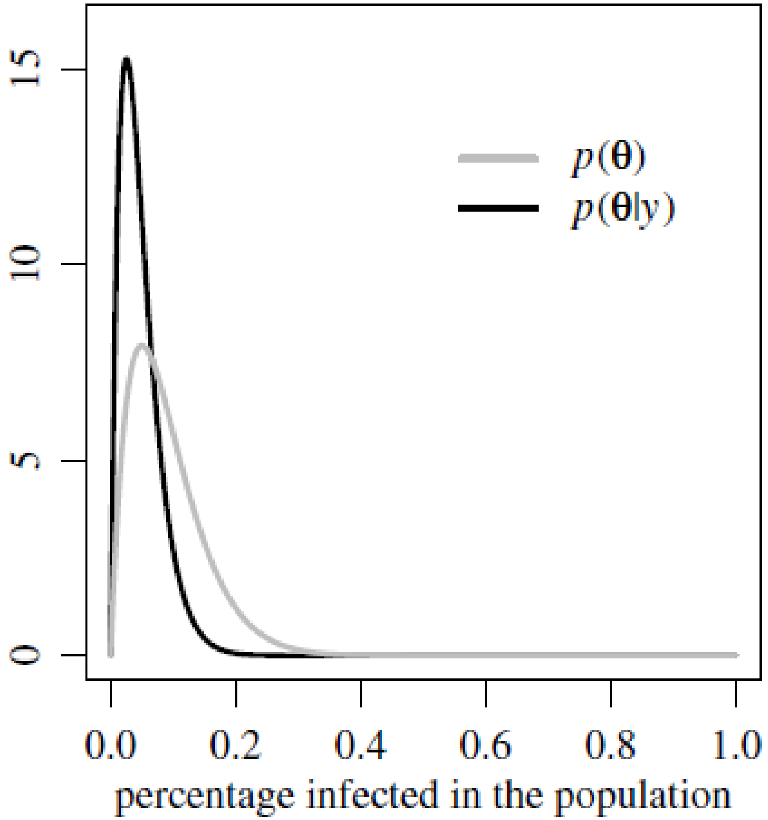


Figure : Black line show the density of posterior distribution

Posterior distribution

- The posterior distribution p(θ|Y = 0) provides us with a model for learning about the city-wide infection rate .
- From a theoretical perspective, a rational individual whose prior beliefs about θ were represented by a beta(2,20) distribution now has beliefs that are represented by a beta(2,40) distribution.
- As a practical matter, if we accept the beta(2,20) distribution as a reasonable measure of prior information, then we accept the beta(2,40) distribution as a reasonable measure of posterior information.

III.b.4). Relationship between the prior and the posterior

Situation

- Consider prior beliefs represented by Beta(a, b) distributions for values of (a, b)
- If $\theta \sim$Beta(a,b) then given Y = y the posterior distribution of θ is Beta(a + y, b + n − y).

Posterior Distribution

- If $\theta|\{Y = y\}$~beta(a + y, b + n − y)
- Then

  E[θ|y]=$\frac{a+y}{a+b+n}$
-
  mode[θ|y]=$\frac{a+y-1}{a+b+n-2}$
-
  Var[θ|y]=$\frac{E[\theta|y]\ E\ [1-\theta|y]}{a+b+n+1}$

Posterior expectation

- The posterior expectation:
  $E[\theta|Y=y] = \frac{a+y}{a+b+n} = \frac{n}{a+b+n}\frac{y}{n} + \frac{a+b}{a+b+n}\frac{a}{a+b}$

  $= \frac{n}{w+n}\bar{y} + \frac{w}{w+n}\theta_0$

- Posterior expectation is a weighted average of the sample mean $\bar{y}$ and the prior expectation θ_0

- Where $\theta_0 = \frac{a}{a+b}$ is the prior expectation of θ which represents our prior guess at the true value of θ and

- $w = a + b$ represents our confidence in this guess, expressed on the same scale as the sample size.

Posterior expectation

- The posterior expectation:

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

$$= \frac{w}{w+n} \times \text{prior expectation} + \frac{n}{w+n} \times \text{data average}$$

- The posterior expectation is a weighted average of the prior expectation and the sample average which is recognized as a combination of prior and data information .

Combining information

- Compare by row: we can see the effect of the sample size

- Compare different sample size with prior~beta(1,1) and prior~(3,2), the shape of posterior distribution become narrower.

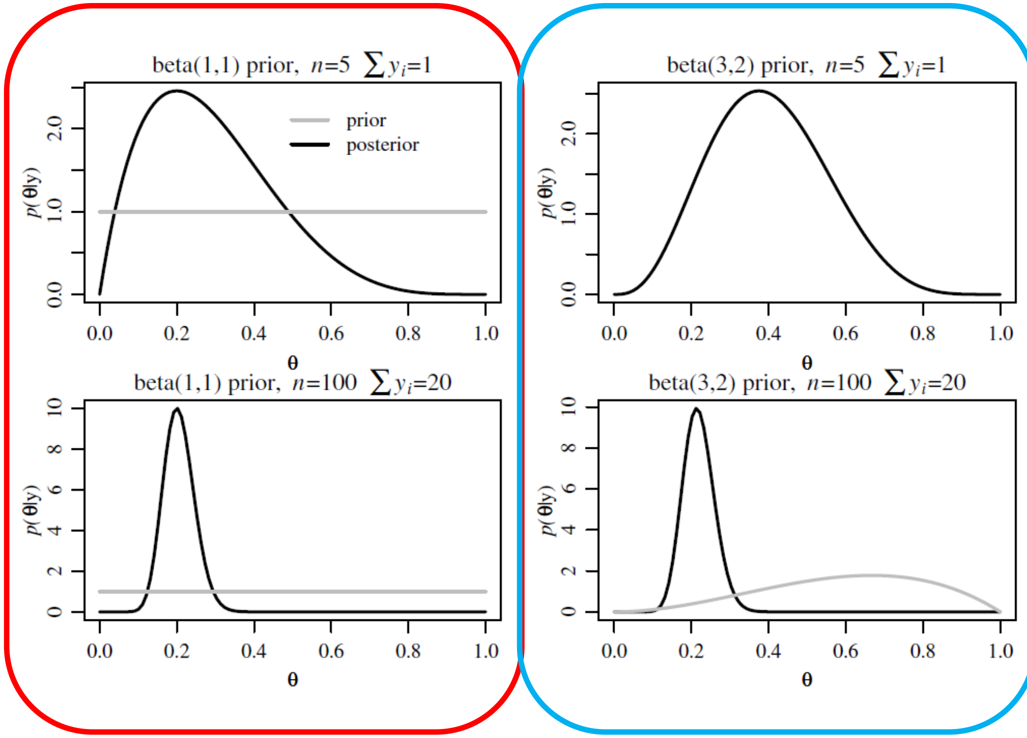


Figure:Beta posterior distributions under two different sample sizes and two different prior distributions.

Combining information

- Compare by column:
we can see the effect of the prior distribution.

- Compare different prior distribution
- when n=5(row1 in both columns), the shape of posterior distribution are different .

- when n=100(row2 in both columns), the shape of both of them are similarly.
Since the sample size n is much larger than our prior sample size a + b, a majority of our information about θ should be coming from the data as opposed to the prior distribution
i.e. n≫a+b then
$\frac{a+b}{a+b+n} \approx 0$, E[$\theta$|Y=y]≈ $\frac{y}{n}$

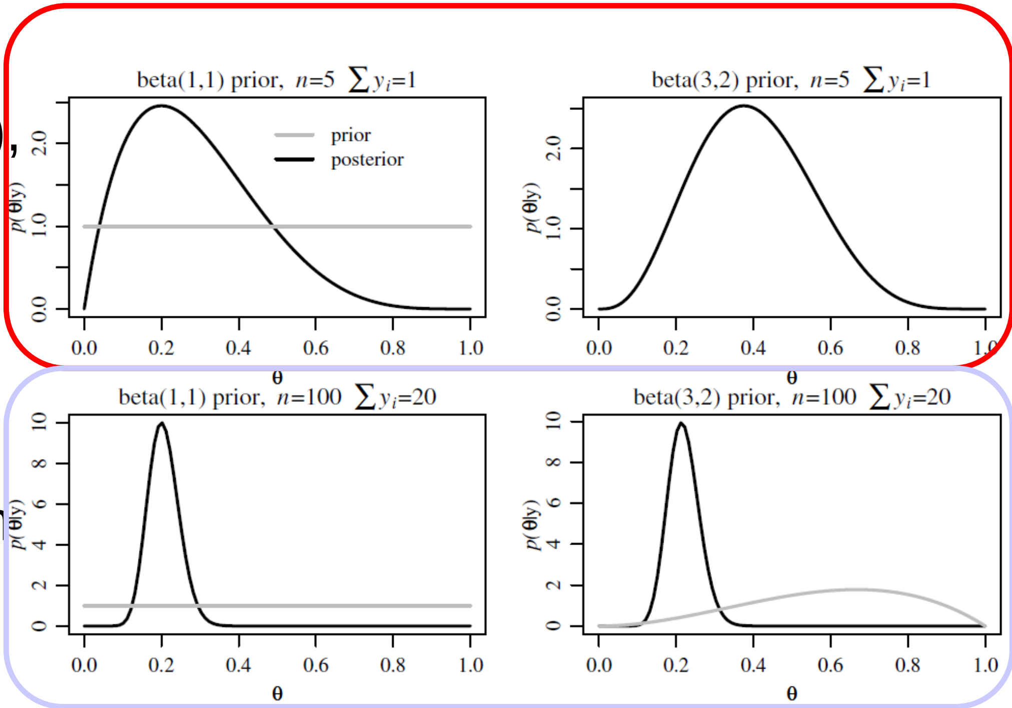
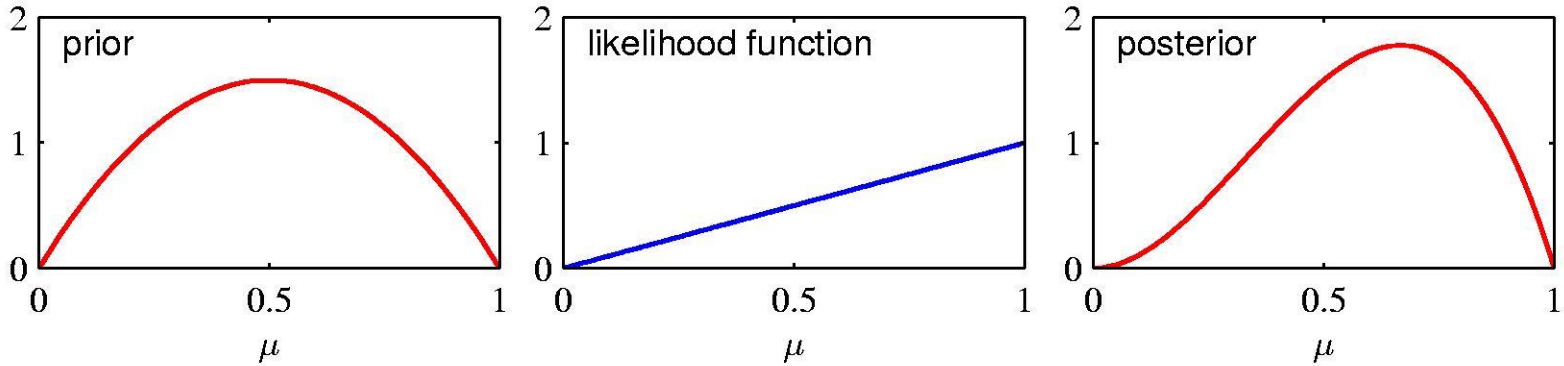


Figure:Beta posterior distributions under two different sample sizes and two different prior distributions.

- Summary:

General property of posterior distributions:

Prior · Likelihood = Posterior



Posterior as compromise between data and prior

- Prior vs. Posterior Expectation:

$E(\theta)=E(E(\theta|y))$

Prior expectation equals posterior expectation averaged over all (prior predictive) data

- Prior vs. Posterior Variance

$var(\theta)=E(var(\theta|y))+var(E(\theta|y))$

Prior variance >= expected posterior var.

Knowing more is usually better.

Not always better. But on average at least as good.

Check the effect of the Prior: Sensitivity analysis

- If we have a prior guess θ_0 and a degree of confidence w , we can approximate the prior beliefs about θ with beta(a, b) which a=$w\theta_0$, b=$w(1-\theta_0)$
- Then the approximate posterior beliefs: beta($w\theta_0 + \text{y}, w(1-\theta_0) + n - y$)

- We can compute the posterior distribution for a wide range of θ_0 and w values to perform a sensitivity analysis

Sensitivity analysis

- Sensitivity analysis is an exploration of how posterior information is affected by differences in prior opinion.
- Contour plot of two posterior quantities such as Figure can explores the effect of θ_0 and w on the posterior distribution .
- e.g. With the prior expectation equal to 0.2,The posterior expectation at most equal to 0.12($E[\theta|Y=0] \leq 0.12$) with variety value of w

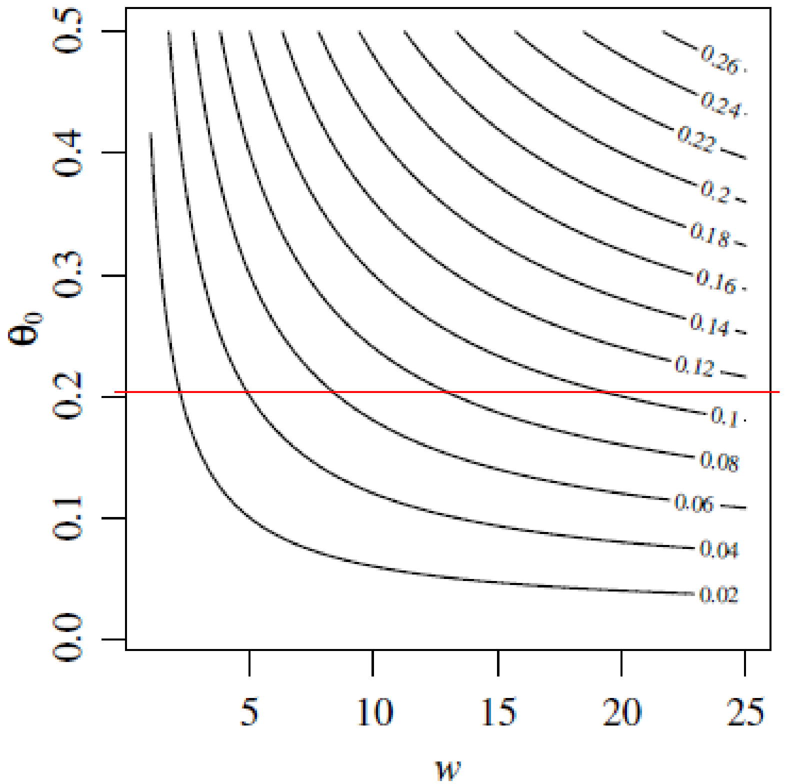


Figure: Contours of the posterior expectation $E[\theta|Y=0]$

Sensitivity analysis

- Example:
  If the city officials would like to recommend a vaccine to the general public unless they were reasonably sure that the current infection rate was less than 0.10.

When prior expectation (θ_0) is low, with different value of w , the $\Pr(\theta < 0.10|Y=0)$ at least have 0.9 or more

When prior belief (w) is low, with different value of θ_0, the $\Pr(\theta < 0.10|Y=0)$ at least have 0.9

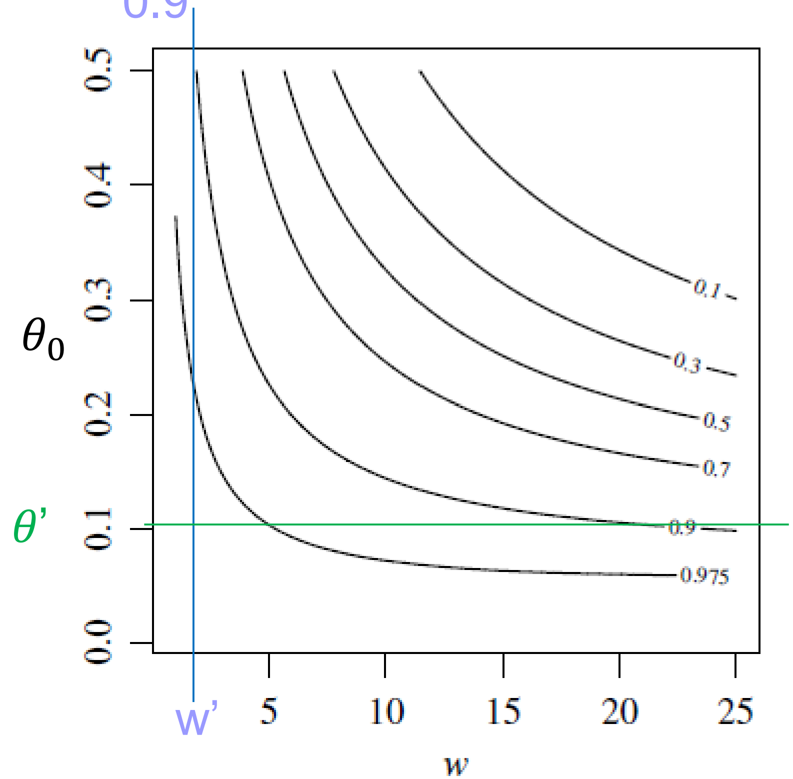


Figure: Contours of the posterior probability $\Pr(\theta < 0.10|Y=0)$

- People with weak prior beliefs w'(low values of w) or low prior expectations θ' are generally 90% or more certain that the infection rate is below 0.10

III.b.5). How to summarize the posterior?

- Point estimate:

1. posterior mean
2. Maximum a posteriori (MAP) estimate

- Interval estimate:

1.quantile-based interval

2.Highest posterior density(HPD) region

Quantile-based interval

- The easiest way to obtain a confidence interval
- To make a $100\times(1-\alpha)\%$ quantile-based confidence interval, find numbers $\theta_\alpha < \theta_{1-\alpha/2}$ such that
  1. $\Pr(\theta < \theta_\alpha | Y = y) = \frac{\alpha}{2}$
  2. $\Pr(\theta > \theta_{1-\alpha/2} | Y = y) = \frac{\alpha}{2}$
- $\theta_\alpha, \theta_{1-\alpha/2}$ are the $\alpha/2$ and 1-α/2 posterior quantiles of θ

- However, there are θ -values outside the quantile-based interval that have higher probability (density) than some points inside the interval.
- This suggests a more restrictive type of interval: Highest posterior density region

Highest posterior density(HPD) region

- A $100 \times (1-\alpha)\%$ HPD region consists of a subset of the parameter space, s(y)⊂Θ such that
  1. $\Pr(\theta \in s(y)|Y=y) = 1-\alpha$
  2. If $\theta_a \in s(y)$ and $\theta_b \notin s(y)$
     then p$(\theta_a|Y=y) > p(\theta_b|Y=y)$

- All points in an HPD region have a higher posterior density than points outside the region

- However if the posterior density is multimodal (having multiple peaks), the HPD region might not be an interval.

Highest posterior density(HPD) region

- Procedure to obtain HPD region:

1. Move a horizontal line down across the density, including in the HPD region all θ-values having a density above the horizontal line.
2. Stop moving the line down when the posterior probability of the θ-values in the region reaches $(1-\alpha)$.

- Figure1 shows the 95% HPD region is [0.04, 0.48], which is narrower (more precise) than the 95 % quantile-based interval[0.06,0.52], although both contain 95% of the posterior probability.

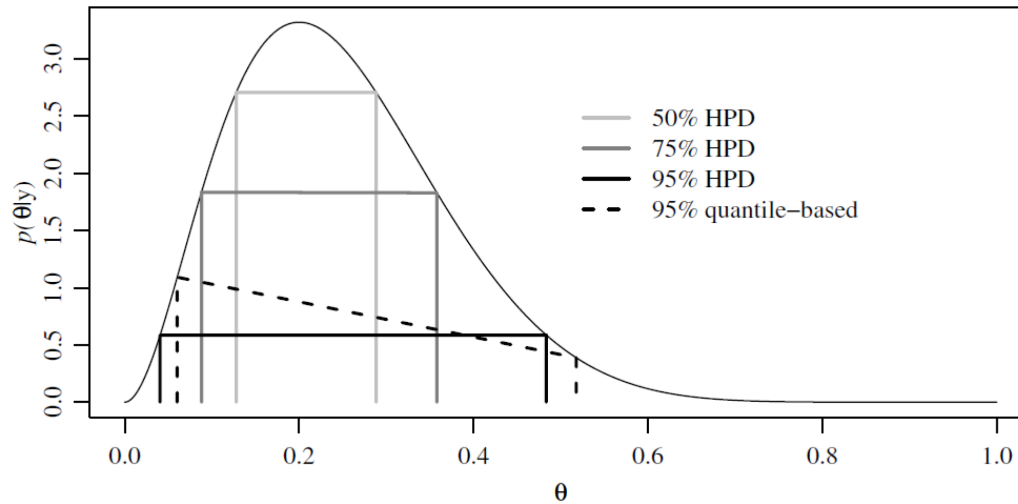


Figure1: Highest posterior density regions of varying probability content.

III.b.6) Conjugate Priors

- Prior $\theta \sim Beta(\alpha, \beta)$
- Likelihood $p(y|\theta) \propto \theta^{y} (1-\theta)^{n-y}$
- Posterior $\theta|y \sim Beta(\alpha+y, \beta+n-y)$

$E(\theta|y) = (\alpha+y)/(\alpha+\beta+n)$

as $n \rightarrow +\infty$, prior has diminishing influence on the posterior

$Var(\theta|y)=E(\theta|y)*(1-E(\theta|y))/(\alpha+\beta+n+1)$

$\rightarrow 0$ as $n \rightarrow +\infty$

Conjugate Priors

- Conjugacy: posterior dist'n has the same parametric form as prior.
- Beta distribution is conjugate to binomial likelihood
- Exponential families have natural conjugate priors

$$f_X(x;\theta) = h(x)\exp\left(\sum_{i=1}^{s}\eta_i(\boldsymbol{\theta})T_i(x) - A(\boldsymbol{\theta})\right)$$

Examples: normal, exponential, gamma, chi-square, beta, Weibull (if the shape parameter is known), Dirichlet, Bernoulli, binomial, multinomial, Poisson, negative binomial, and geometric distributions

Sequential Analysis with conjugate prior

$\text{beta}(\alpha,\beta)$

$\big\downarrow$ data: y1,n1

$\text{beta}(\alpha+y1,\beta+n1-y1)$

$\big\downarrow$ data: y2,n2

$\text{beta}(\alpha+y1+y2, \beta+n1+n2-y1-y2)$

$\big\downarrow$

……

- Informative conjugate priors has a “pseudo-data” interpretation

IV). Poisson distribution

Poisson Distribution

- Poisson distribution is for counts—if events happen at a constant rate over time, the Poisson distribution gives the probability of X number of events occurring in time T.

Poisson Mean and Variance

- Mean

$$\boxed{\mu = \lambda}$$

- Variance and Standard Deviation

$$\boxed{\sigma^2 = \lambda}$$

$$\boxed{\sigma = \sqrt{\lambda}}$$

For a Poisson random variable, the variance and mean are the same!

where λ = expected number of hits in a given time period

Poisson Distribution, example

The Poisson distribution models counts, such as the number of new cases of SARS that occur in women in Hong Kong next month.

The distribution tells you the probability of all possible numbers of new cases, from 0 to infinity.

If X= # of new cases next month and *X* ~ Poisson (λ), then the probability that *X*=*k* (a particular count) is:

$$\boxed{p(X=k)=\frac{\lambda^k e^{-\lambda}}{k!}}$$

Example

- For example, if new cases of Ebola Virus in Africa are occurring at a rate of about 2 per month, then these are the probabilities that: 0,1, 2, 3, 4, 5, 6, to 1000 to 1 million to… cases will occur in Africa in the next month:

Poisson Probability table

X	P(X)
0	$\frac{2^0 e^{-2}}{0!}$ =.135
1	$\frac{2^1 e^{-2}}{1!}$ =.27
2	$\frac{2^2 e^{-2}}{2!}$ =.27
3	$\frac{2^3 e^{-2}}{3!}$ =.18
4	$\frac{2^4 e^{-2}}{4!}$ =.09
5	
…	…

Example: Poisson distribution

Suppose that a rare disease has an incidence of 1 in 1000 person-years. Assuming that members of the population are affected independently, find the probability of k cases in a population of 10,000 (followed over 1 year) for k=0,1,2.

The expected value (mean) =λ = .001*10,000 = 10
10 new cases expected in this population per year→

$$P(X=0)=\frac{(10)^0 e^{-(10)}}{0!}=.0000454$$

$$P(X=1)=\frac{(10)^1 e^{-(10)}}{1!}=.000454$$

$$P(X=2)=\frac{(10)^2 e^{-(10)}}{2!}=.00227$$

more on Poisson…

"Poisson Process" (rates)

Note that the Poisson parameter λ can be given as the mean number of events that occur in a defined time period OR, equivalently, λ can be given as a rate, such as λ=2/month (2 events per 1 month) that must be multiplied by *t*=time (called a "Poisson Process") →

X ~ Poisson (λ)

$$P(X = k) = \frac{(\lambda t)^k e^{-\lambda t}}{k!}$$

E(X) = λt

Var(X) = λt

Example

For example, if new cases of Ebola in Africa are occurring at a rate of about 2 per month, then what's the probability that exactly 4 cases will occur in the next 3 months?

X ~ Poisson (λ=2/month)

$$\mathrm{P}(\mathrm{X}=4 \text{ in } 3 \text{ months})=\frac{(2*3)^{4}e^{-(2*3)}}{4!}=\frac{6^{4}e^{-(6)}}{4!}=13.4\%$$

Exactly 6 cases?

$$\mathrm{P}(\mathrm{X}=6 \text{ in } 3 \text{ months})=\frac{(2*3)^{6}e^{-(2*3)}}{6!}=\frac{6^{6}e^{-(6)}}{6!}=16\%$$

Practice problems

1a. If calls to your cell phone are a Poisson process with a constant rate λ=2 calls per hour, what's the probability that, if you forget to turn your phone off in a 1.5 hour movie, your phone rings during that time?

1b. How many phone calls do you expect to get during the movie?

Answer

1a. If calls to your cell phone are a Poisson process with a constant rate λ=2 calls per hour, what's the probability that, if you forget to turn your phone off in a 1.5 hour movie, your phone rings during that time?

X ~ Poisson (λ=2 calls/hour)

$P(X\geq 1)=1 - P(X=0)$

$$P(X=0)=\frac{(2*1.5)^0 e^{-2(1.5)}}{0!}\frac{(3)^0 e^{-3}}{0!}=e^{-3}=.05$$

$\therefore P(X\geq 1)=1 - .05 = 95\%$ chance

1b. How many phone calls do you expect to get during the movie?

$E(X) = \lambda t = 2(1.5) = 3$

V). Posterior Inference for Poisson mean

Poisson Distribution

- A random variable Y has a Poisson distribution with mean θ if
$$\Pr(Y = y|\theta) = \frac{\theta^y e^{-\theta}}{y!}$$
for $y \in \{0,1,2,\ldots\}$
- E[Y|θ]=θ Var[Y|θ]=θ
- Poisson family of distributions has a “mean-variance relationship” because if one Poisson distribution has a larger mean than another, it will have a larger variance as well.

Posterior inference

- If we model $Y_1, \dots, Y_n$ as i.i.d Poisson with mean θ , then the joint likelihood of sample data:

$$\Pr(Y_1 = y_1, \dots, Y_n = y_n | \theta) = \prod_{i=1}^{n} p(y_i|\theta)$$

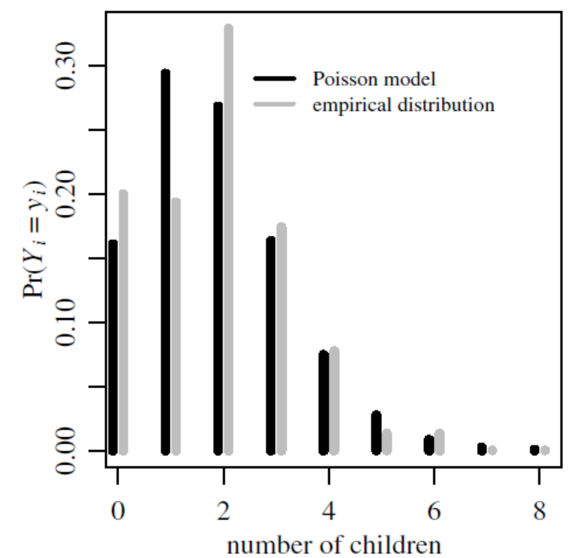
$$= \prod_{i=1}^{n} \frac{1}{y_i!} \theta^{y_i} e^{-\theta}$$

$$= c(y_1, \dots, y_n) \theta^{\sum y_i} e^{-n\theta}$$

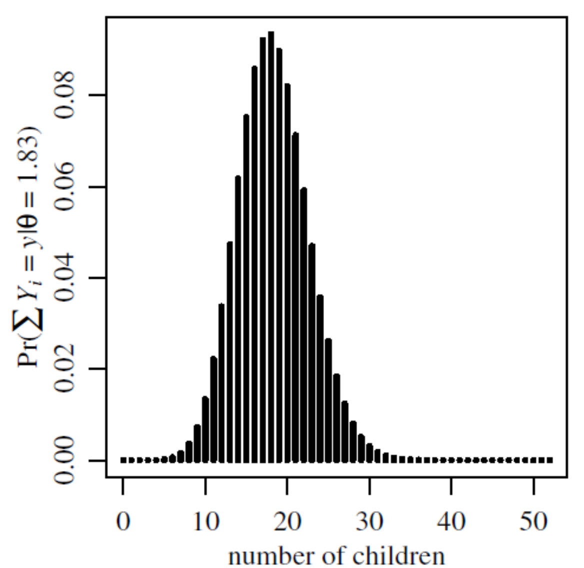
- $\sum_{i=1}^{n} Y_i$ contains all the information about θ that is available in the data.
- $\sum_{i=1}^{n} Y_i$ is a sufficient statistic and $\{\sum_{i=1}^{n} Y_i | \theta\} \sim$ Poisson($n\theta$)

Posterior inference

- Figure1 shows a Poisson distribution with mean of 1.83, with the empirical distribution of the number of children of women of age 40 from the GSS during the 1990s.



- Figure2 shows the distribution of the sum of 10 i.i.d. Poisson random variables with mean 1.83 which is the same as a Poisson distribution with mean 18.3
$\{\sum_{i=1}^{10} Y_i | \theta = 1.83\} \sim$ Poisson(10*1.83=18.3)



Posterior for conjugate prior

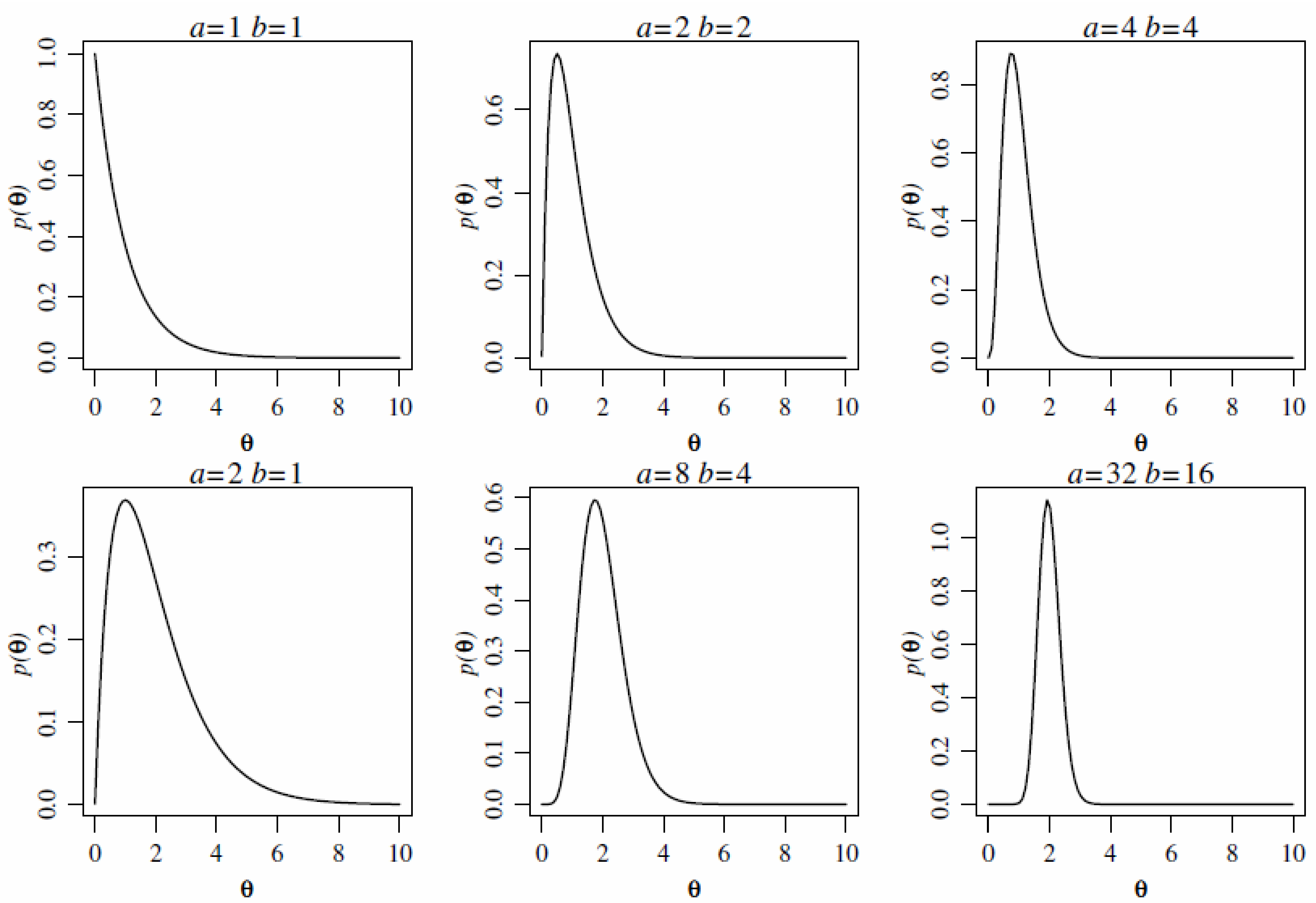
$$
\begin{aligned}
p(\theta|y_1,\ldots,y_n) &\propto p(\theta) \times p(y_1,\ldots,y_n|\theta) \\
&\propto p(\theta) \times \theta^{\sum y_i} e^{-n\theta}.
\end{aligned}
$$

- Its conjugate prior is a Gamma distribution:

$$p(\theta) = \mathrm{dgamma}(\theta, a, b) = \frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta}, \quad \text{for } \theta,\ a,\ b > 0.$$

$$
\begin{aligned}
&\mathrm{E}[\theta] = a/b; \\
&\mathrm{Var}[\theta] = a/b^2; \\
&\mathrm{mode}[\theta] = \begin{cases} (a-1)/b & \text{if } a > 1 \\ 0 & \text{if } a \le 1 \end{cases}.
\end{aligned}
$$

Gamma distributions



Posterior

$$\begin{aligned} p(\theta|y_1,\ldots,y_n) &= p(\theta)\times p(y_1,\ldots,y_n|\theta)/p(y_1,\ldots,y_n) \\ &= \left\{\theta^{a-1}e^{-b\theta}\right\}\times\left\{\theta^{\sum y_i}e^{-n\theta}\right\}\times c(y_1,\ldots,y_n,a,b) \\ &= \left\{\theta^{a+\sum y_i-1}e^{-(b+n)\theta}\right\}\times c(y_1,\ldots,y_n,a,b). \end{aligned}$$

Thus:

$$\left.\begin{array}{c} \theta\sim\mathrm{gamma}(a,b) \\ Y_1,\ldots,Y_n|\theta\sim\mathrm{Poisson}(\theta) \end{array}\right\}\Rightarrow\{\theta|Y_1,\ldots,Y_n\}\sim\mathrm{gamma}(a+\sum_{i=1}^n Y_i, b+n)$$

Analysis of the posterior

$$\begin{aligned} \mathrm{E}[\theta|y_1,\ldots,y_n] &= \frac{a+\sum y_i}{b+n} \\ &= \frac{b}{b+n}\frac{a}{b} + \frac{n}{b+n}\frac{\sum y_i}{n} \end{aligned}$$

- b is interpreted as the number of prior observations;
- a is interpreted as the sum of counts from b prior observations
- For large n, the information from the data dominates the prior information

$$n >> b \Rightarrow \mathrm{E}[\theta|y_1,\ldots,y_n] \approx \bar{y},\ \mathrm{Var}[\theta|y_1,\ldots,y_n] \approx \bar{y}/n$$

Posterior predictive distribution

$$
\begin{aligned}
p(\tilde{y}|y_1,\ldots,y_n) &= \int_0^\infty p(\tilde{y}|\theta,y_1,\ldots,y_n)p(\theta|y_1,\ldots,y_n)\ d\theta \\
&= \int p(\tilde{y}|\theta)p(\theta|y_1,\ldots,y_n)\ d\theta \\
&= \int \mathrm{dpois}(\tilde{y},\theta)\mathrm{dgamma}(\theta,a+\sum y_i,b+n)\ d\theta \\
&= \int \left\{\frac{1}{\tilde{y}!}\theta^{\tilde{y}}e^{-\theta}\right\}\left\{\frac{(b+n)^{a+\sum y_i}}{\Gamma(a+\sum y_i)}\theta^{a+\sum y_i-1}e^{-(b+n)\theta}\right\}\ d\theta \\
&= \frac{(b+n)^{a+\sum y_i}}{\Gamma(\tilde{y}+1)\Gamma(a+\sum y_i)}\int_0^\infty \theta^{a+\sum y_i+\tilde{y}-1}e^{-(b+n+1)\theta}\ d\theta\,.
\end{aligned}
$$

- it turns out to be a negative binomial distribution:

$$
p(\tilde{y}|y_1,\ldots,y_n) = \frac{\Gamma(a+\sum y_i+\tilde{y})}{\Gamma(\tilde{y}+1)\Gamma(a+\sum y_i)}\left(\frac{b+n}{b+n+1}\right)^{a+\sum y_i}\left(\frac{1}{b+n+1}\right)^{\tilde{y}}
$$

$$
\begin{aligned}
\mathrm{E}[\tilde{Y}|y_1,\ldots,y_n] &= \frac{a+\sum y_i}{b+n} = \mathrm{E}[\theta|y_1,\ldots,y_n]; \\
\mathrm{Var}[\tilde{Y}|y_1,\ldots,y_n] &= \frac{a+\sum y_i}{b+n}\frac{b+n+1}{b+n} = \mathrm{Var}[\theta|y_1,\ldots y_n]\times(b+n+1) \\
&= \mathrm{E}[\theta|y_1,\ldots,y_n]\times\frac{b+n+1}{b+n}.
\end{aligned}
$$

Example:

- Comparing the women with college degrees to those without in terms of their numbers of children.
- We have data on the educational attainment and number of children of 155 women who were 40 years of age at the time of their participation in the survey.

- Let $Y_{1,1} \ldots, Y_{n_1,1}$ denote the numbers of children for the $n_1$women without college degrees

$$Y_{1,1} \ldots, Y_{n_1,1} | \theta_1 \sim i.i.d. Poisson(\theta_1)$$

- Let $Y_{1,2} \ldots, Y_{n_2,2}$ be the data for women with degrees.

$$Y_{1,2} \ldots, Y_{n_2,2} | \theta_2 \sim i.i.d. Poisson(\theta_2)$$

Posterior distribution

- Less than bachelor's:
  $n_1 = 111, \sum_{i=1}^{n_1} Y_{i,1} = 217$,
  $\bar{Y}_1 = 1.95$

- Bachelor's or higher:
  $n_2 = 44, \sum_{i=1}^{n_2} Y_{i,2} = 66$,
  $\bar{Y}_2 = 1.50$

- $\{\theta_1, \theta_2\}$~i.i.d gamma(a=2.b=1)
- Posterior distribution:
  $\theta_1 | \{n_1 = 111, , \sum_{i=1}^{n_1} Y_{i,1} = 217\}$~
  gamma(2+217,1+111)=gamma(219,112)

  $\theta_2 | \{n_2 = 44, \sum_{i=1}^{n_2} Y_{i,2} = 66\}$~
  gamma(2+66,1+44)=gamma(68,45)

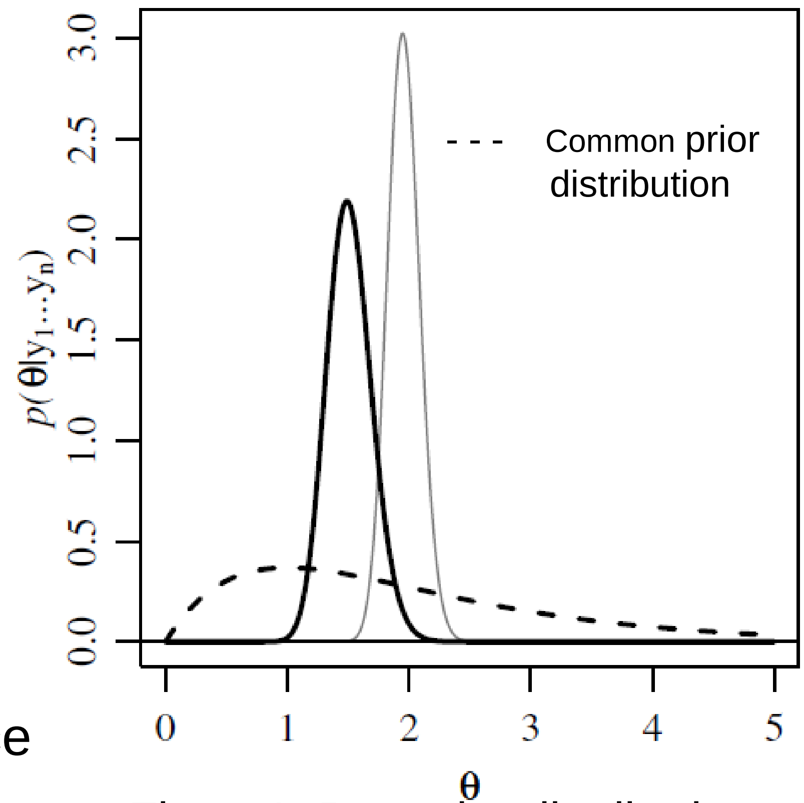


Figure1: Posterior distribution of mean birth rates of the two groups

- The posterior indicates substantial evidence that $\theta_1 > \theta_2$.
  i.e. $\Pr(\theta_1 > \theta_2 | , \sum_{i=1}^{n_1} Y_{i,1} = 217, \sum_{i=1}^{n_2} Y_{i,2} = 66)$

  $= \int_0^\infty \int_0^{\theta_1} p(\theta_1, \theta_2 | y_{1,1}, \ldots, y_{n,2}) d\theta_2 d\theta_1$ =0.97

Posterior predictive distribution

- Consider two randomly sampled individuals, one from each of the two populations.
- The posterior predictive distributions for $\tilde{Y}_1$ and $\tilde{Y}_2$ are both negative binomial distributions
- There is much more overlap between these two distributions than between the posterior distributions of θ_1 and θ_2
- i.e. $\Pr(\tilde{Y}_1 > \tilde{Y}_2 | , \sum_{i=1}^{n_1} Y_{i,1} = 217, \sum_{i=1}^{n_2} Y_{i,2} = 66\ ) =$0.48

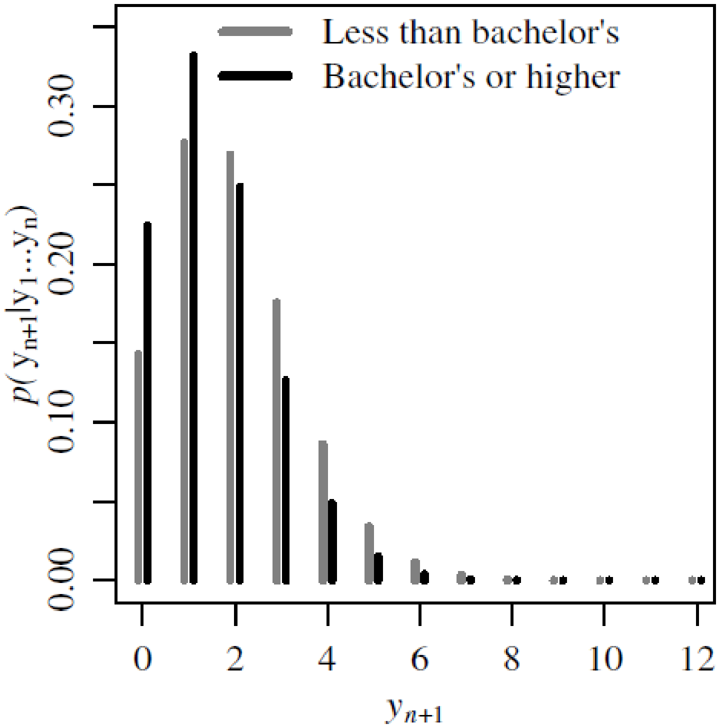


Figure2: Posterior predictive distributions for number of children