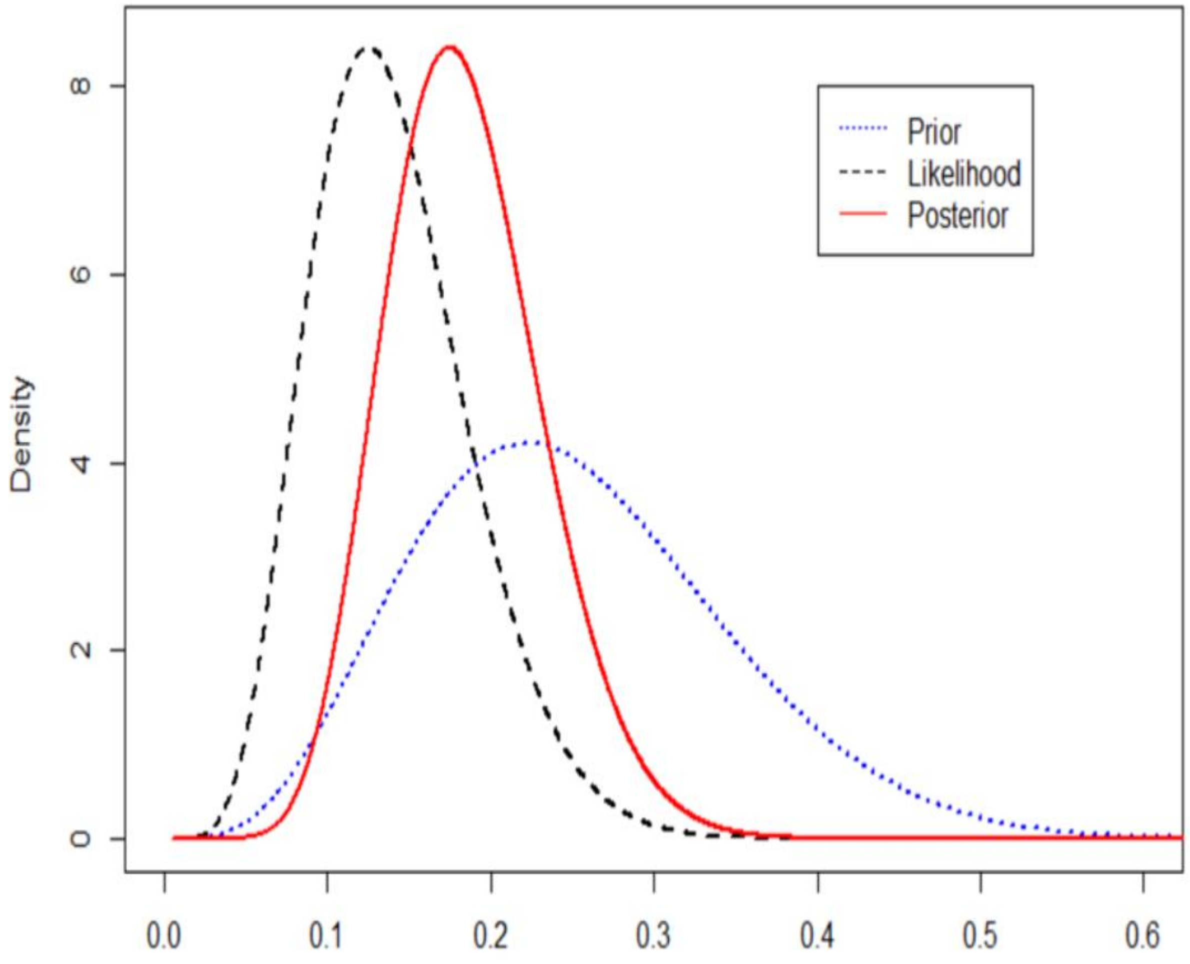


Applied Bayesian Statistics

Unit 4

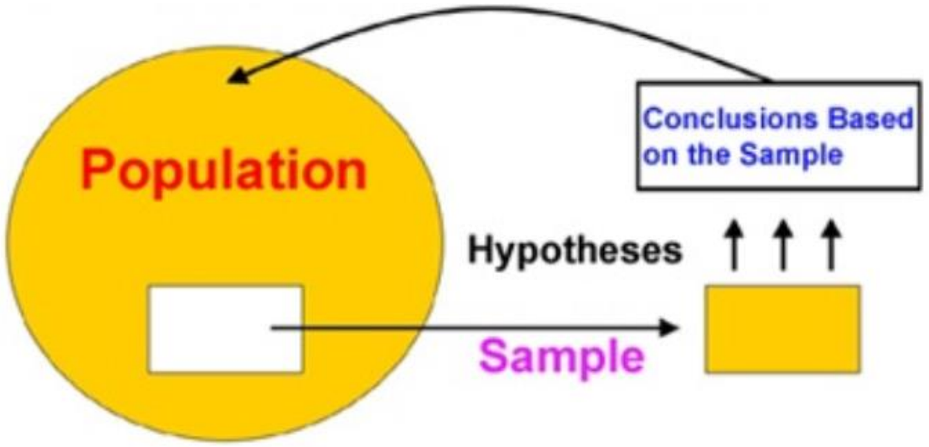
Inference for a Population Proportion

Hamdy F. F. Mahmoud, PhD
Collegiate Assistant Professor
Statistics Department @ VT
ehamdy@vt.edu

This Unit covers

- **What do we mean by statistical inference?**
- **Estimation and Testing: Frequentist Approach**
  - Maximum Likelihood Estimation
  - Frequentist Confidence Intervals
  - Frequentist Hypothesis Testing
- **Estimation and Testing: Bayesian Approach**
  - The Posterior Mean
  - Other Bayesian Point Estimates
  - Bayesian Confidence Intervals
  - Bayesian Hypothesis Testing
- **Posterior Predictive Distributions**
- What is to come?

❑ What do we mean by statistical inference?



Statistical inference is drawing conclusions about an entire population based on data in a sample drawn from that population. From both *frequentist* and *Bayesian* perspectives, there are three main goals of inference: *estimation*, *hypothesis testing*, and *prediction*.

Estimation and Testing: Frequentist Approach

Maximum Likelihood Estimation

Frequentist approach uses the data distribution to estimate an unknown population parameter.

For our motivating example, our data consisting of independent yes/no responses and the data distribution is the binomial probability mass function:

$$\boldsymbol{p(y|\pi) = \binom{n}{y}\pi^{y}(1-\pi)^{n-y},\ \ y = 0, 1, \dots., n}$$

The frequentist approach does not treat the parameter as a random variable and does not specify a prior distribution to it.

- ○ The goal of the frequentist approach is to obtain a point estimate of a population parameter and one of the frequentist methods is the maximum likelihood method (MLE).

Maximum Likelihood Estimation of a population parameter

Step 1: take the log of the likelihood function

$$l(\pi) = log\binom{n}{y} + ylog(\pi) + (n-y)\log(1-\pi)\,, \qquad\qquad 0 < \pi < 1$$

Step 2: take the 1st derivative of the log likelihood function

$$\frac{dl(\pi)}{d\pi} = \frac{y}{\pi} - \frac{n-y}{1-\pi}$$

Step 3: set the 1st derivative equal to 0 and solve for π

$$\frac{y}{\pi} - \frac{n-y}{1-\pi} = 0$$

Step 4: find the maximum likelihood estimate of π

$$\hat{\pi} = \frac{y}{n}$$

Maximum Likelihood Estimation

To check if this is a maximum or minimum, we take the 2nd derivative of the likelihood function. If it is negative, it is a maximum and if it is positive, it is a minimum.

$$\frac{d^2 l(\pi)}{d\pi^2} = -\frac{y}{\pi^2} - \frac{n-y}{(1-\pi)^2}$$

Evaluating this at $\hat{\pi} = \frac{y}{n}$ gives $\frac{-n^3}{y(n-y)} < 0$

For our example, the maximum likelihood estimation is

$$\hat{\pi} = \frac{7}{50} = 0.14$$

Frequentist Confidence Interval

- The point estimate is very unlikely to be exactly equal to the true population parameter.

- The fundamental idea behind the frequentist confidence interval is repeated sampling.

- **For example**, a 95% confidence interval means that if we selected randomly all the possible samples of size *n* and calculated a 95% confidence interval for each sample, we find 95% of the calculated confidence intervals will contain the population parameter and 5% will not contain the population parameter.

Let us visualize this idea for the population proportion, click on this Applet!

Frequentist Confidence Interval

- **In reality,** we select only one sample and calculate only one confidence interval. We have no way of knowing for sure whether our particular confidence interval is one of the lucky 95% of confidence intervals that contain the true value of π , or whether we instead by bad luck got one of the 5% of samples that produce 95% confidence intervals that don't contain the true value.

- Let use a built-in function in R to estimate the confidence interval of π based on the frequentist approach that uses the maximum likelihood method!

- The R function name is *binom.test*

Frequentist Confidence Interval

- Go to R or RStudio!
- Write help(binom.test) in R to see the syntax of this function.

Write the following code to get the point estimate, 95% confidence interval, and test of hypotheses for π based on our data with 7 successes in 50 trials:

```
> binom.test(7, 50, p=0.2, conf.level = .95)
```

```
Exact binomial test

data:  7 and 50
number of successes = 7, number of trials = 50, p-value = 0.3765
alternative hypothesis: true probability of success is not equal to 0.2
95 percent confidence interval:
 0.0581917   0.2673960
sample estimates:
probability of success
                   0.14
```

Frequentist Confidence Interval

- What is the point estimate and 95% confidence interval from the output?

- Can we say that the probability that the interval (0.058, 0.267) contains the population proportion, π, is 0.95? And why?

Frequentist Hypothesis Testing

Hypothesis testing is appropriate when different courses of action would be taken given different values of an unknown population parameter. If we were convinced that the proportion was more substantial, say larger than 10%, we might indeed want to go before the regents to argue against the tuition increase.

A statistical hypothesis is a statement about an unknown population parameter. Hypotheses testing involve setting up two such statements, which are mutually exclusive.

Thus, we might want to test the following hypotheses regarding π:

$$H_0 : \pi \leq 0.1 \text{ vs } H_a : \pi > 0.1$$

Frequentist Hypothesis Testing

- The frequentist uses the p-value to evaluate the evidence in the data against the null hypothesis, H_0 , and in favor of the alternative hypothesis, H_a.

- The concept of p-value again is rooted in the question of what would happen under repeated sampling, specifically, *assuming that H_0 is true,* if we draw many, many simple random samples, what is the probability of getting a sample with as much evidence against H_0 as our actual sample has, or even more.

- So p-value is *not* the probability that H_0 is true. Small p-value indicates that the data is inconsistent with the null hypothesis. The smaller the p-value, the less likely it would have been to draw a sample like ours if H_0 were true.

Frequentist Hypothesis Testing

- Go to R or RStudio!

The code below is to see which hypotheses the data support about the population proportion, π, based on our data with 7 successes in 50 trials:

```
> binom.test( 7, 50, p = .1, alternative="greater")

Exact binomial test

data: 7 and 50
number of successes = 7, number of trials = 50,
p-value = 0.2298
alternative hypothesis: true probability of success is greater than 0.1
95 percent confidence interval:
0.0675967 1.0000000
sample estimates:
probability of success
0.14
```

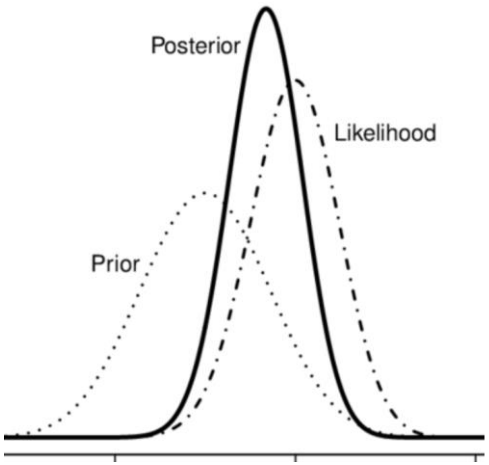
- What is your conclusion?

Questions?

Estimation and Testing: Bayesian Approach

Bayesian Inference

- The **posterior distribution** contains all the information about the unknown parameter, so all Bayesian inference is based on it.

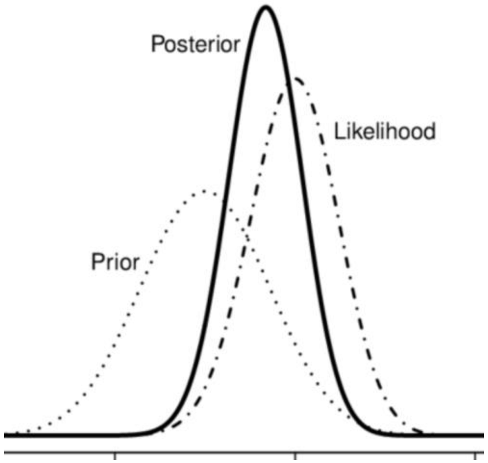


- Although the plot of the posterior density gives a full graphical description, numeric summaries of the posterior are needed as well, such as posterior mean, median, mode, standard deviation, 5-number summary, … etc.

- To estimate a confidence interval for a population parameter or to do hypotheses testing, the posterior distribution is used.

The Posterior Mean

- The posterior mean is often used as the Bayesian point estimate of a parameter.



- For a beta prior and binomial likelihood, the posterior distribution is $Beta(\alpha + y, \beta + n - y)$ and its mean is

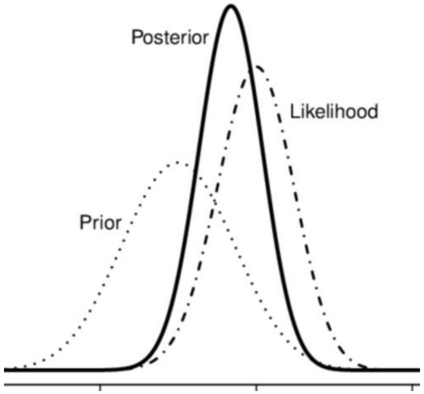
$$E(\pi|y) = \frac{(\alpha + y)}{(\alpha + y) + (\beta + n - y)} = \frac{\alpha + y}{\alpha + \beta + n}$$

- With a sample of 50 and 7 successes, and beta(10,40) prior

$$E(\pi|y) = \frac{10 + 7}{10 + 40 + 50} = \frac{17}{100} = 0.17$$

The Posterior Mean

Notes on the posterior mean



- It is always between the prior mean and maximum likelihood estimate

For $Beta(10,40)$ prior, the mean is $E(\pi) = \frac{10}{10+40} = 0.2$ and the maximum likelihood estimate (MLE) from the data function is $\frac{y}{n} = \frac{7}{50} = 0.14$.

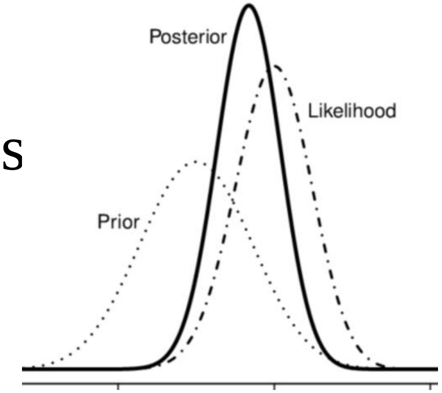
- The posterior mean is the weighted average of the prior mean and MLE.

$$\mu_{posterior} = \frac{\alpha + y}{\alpha + \beta + n} = w\frac{\alpha}{\alpha + \beta} + (1 - w)\frac{y}{n}$$

where $w = \frac{\alpha+\beta}{\alpha+\beta+n}$

Other Bayesian Point Estimates

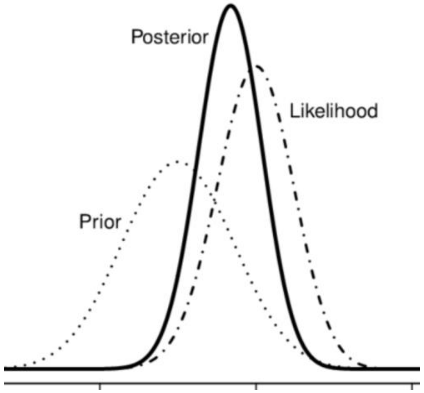
The posterior median and posterior mode are sometimes are used instead of the posterior mean as Bayesian point estimates.



- The mode for beta distribution is $\frac{\alpha-1}{\alpha+\beta-2}$, see the distributions table.
- For Beta(10, 40) prior, the posterior distribution mode is $\frac{\alpha+y-1}{\beta+n-y-2} = \frac{10+7-1}{40+50-7-2} = \frac{16}{81} = 0.198$

Other Bayesian Point Estimates

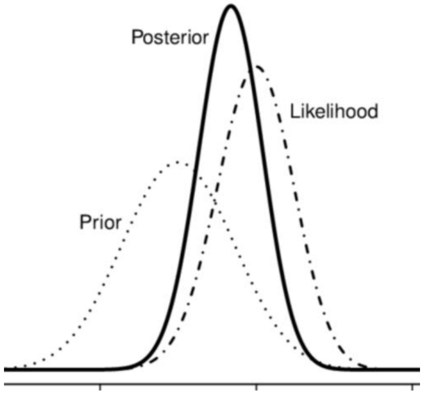
- Estimate the **posterior mode** using U(0,1) prior, and 7 successes and 43 failures?



- Estimate the **posterior median** for these two priors: U(0,1) and Beta(10,40).

Other Bayesian Point Estimates

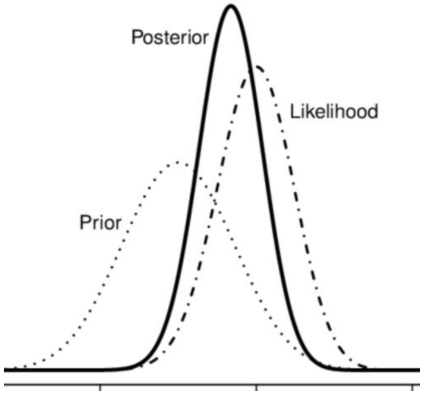
The posterior variance is one summary of the spread of the posterior distribution. The larger the posterior variance, the more uncertainty we still have about the parameter, even after learning from the current data.



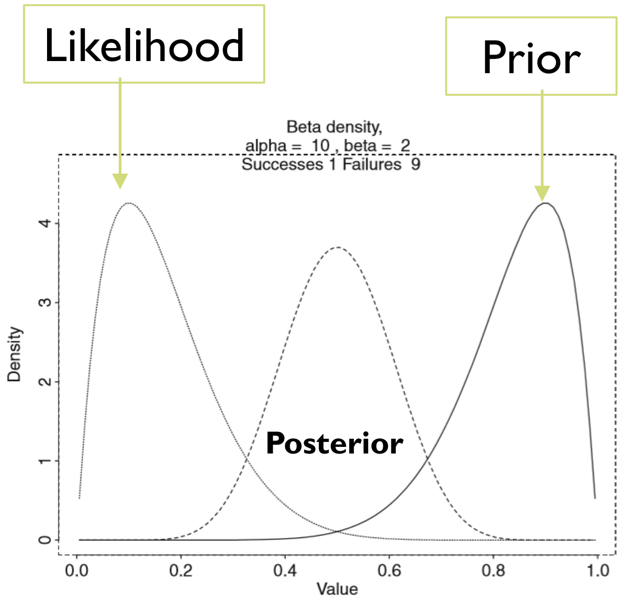
- The variance for beta distribution is $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$, see the table of distributions.
- In our school-quitting example, with Beta(10, 40) prior, the posterior distribution is $Beta(\alpha+y,\beta+n-y)$. So the posterior variance is $\frac{(\alpha+y)(\beta+n-y)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}=\frac{(10+7)(40+50-7)}{(10+40+50)^2(10+40+50+1)}=\frac{17*83}{100^2*101}=0.00014$

Other Bayesian Point Estimates

- The posterior variance is smaller with an informative prior than with a noninformative prior.



- The posterior variance is smaller than the prior variance, except when the prior and the data are in direct conflict.



The beta prior density has mean 0.17, while the sample proportion in the data is 0.9. The posterior density is a compromise between the prior and the likelihood, but the posterior variance is not smaller than the prior variance.

Other Bayesian Point Estimates

- Calculate the prior variance and posterior variance in case beta(10,40) prior is used for the school-quitting example.

- Calculate the prior variance and posterior variance in case uniform(0,1) prior is used for the our school-quitting example.

- Are the posterior variance and prior variance different in the two cases above?

Bayesian Confidence Interval, credible set

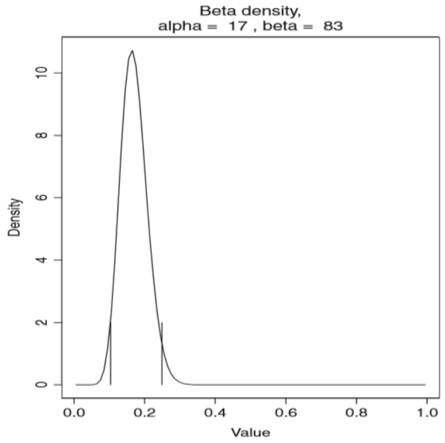
Bayesian intervals called “credible sets” are used as numeric posterior summaries. There are two types of credible intervals commonly used:

1) **Equal-Tail Posterior Credible Sets**

For example, the endpoints of a *95% equal-tail credible set* are the 0.025 and the 0.975 quantiles of the posterior distribution.

We can use built-in R functions to calculate them. For the quitting school problem with the beta(10,40) prior, the posterior density is beta(17, 83), and the qbeta function in R can be used as follows:

```
> qbeta(c(0.025, 0.975), 17, 83 )
> [1] 0.1033333 0.2491463
```

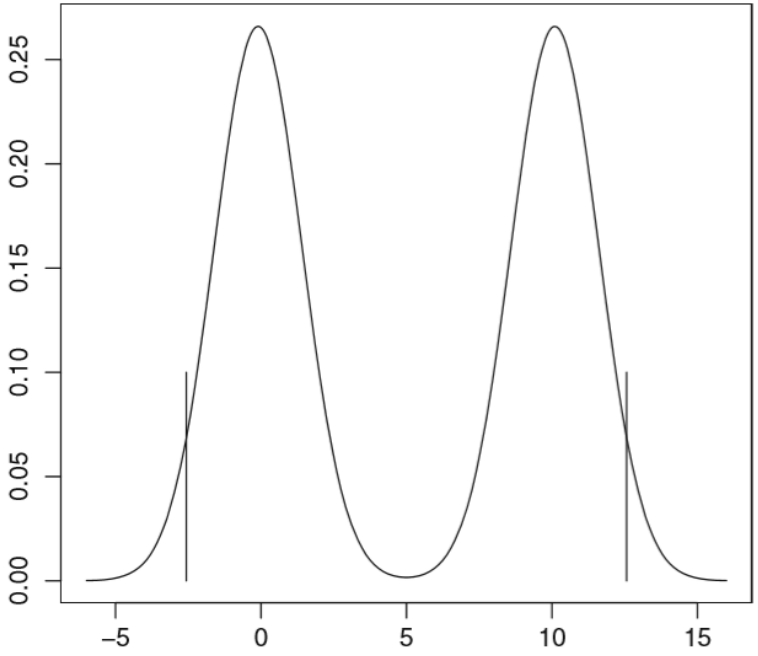


Bayesian Confidence Interval, credible set

- Calculate a 99% equal-tail credible set if instead a uniform prior is used and show which one is narrower, and why?

Bayesian Confidence Interval, credible set

Equal-tail intervals are easy to compute and easy to understand, but its disadvantage is that, unless the posterior density is symmetric and unimodal, there may be points outside the interval that have higher posterior density than some points inside the interval.

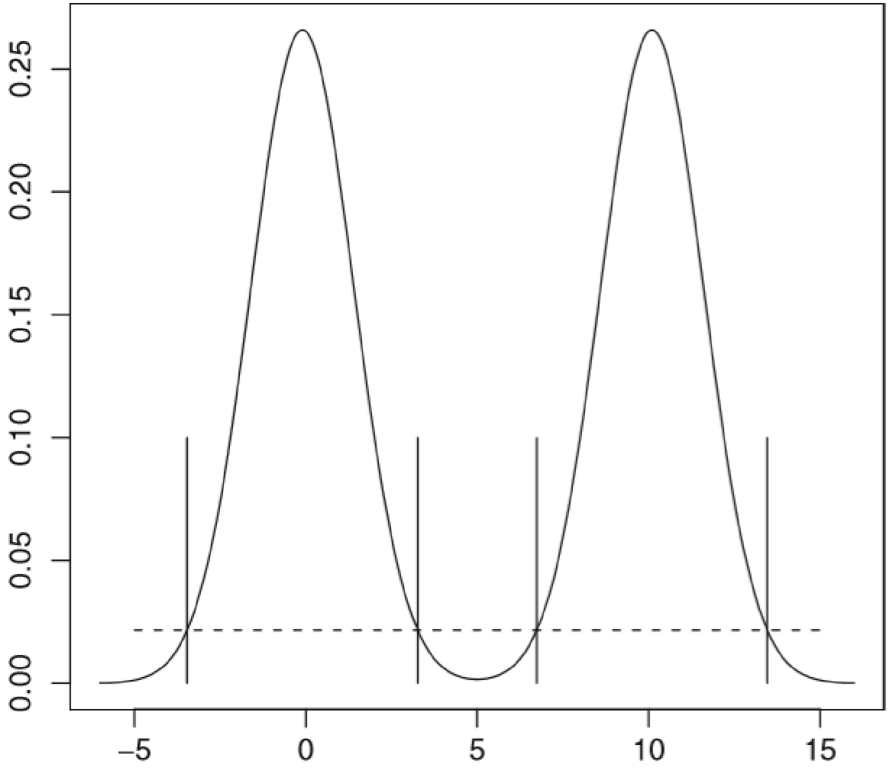


The posterior density above is bimodal with widely separated modes. The equal tail credible set includes the region around 5, where the posterior density is much smaller than it is immediately outside the interval.

Bayesian Confidence Interval, credible set

2) **Highest Posterior Density Region HPD**

The posterior density at any point inside such an HPD region is greater than the density at any point outside it.



HPD region for a bimodal density is two disjoint intervals—the interval between the two left more vertical lines and the one between the two right more vertical lines

Bayesian Confidence Interval, credible set

Interpretation of Bayesian Intervals

- The posterior distribution represents our updated subjective probability distribution for an unknown parameter. **So the interpretation of the 95% credible set is that the probability that the true population proportion, π, is in that interval is 0.95.**

- **For example,** if the beta(10,40) is our prior beliefs or knowledge about the parameter, π , then after seeing our data, we would believe that

$$P(0.103 < \pi < 0.249) = 0.95$$

- This is precisely the kind of statement that **the frequentist cannot make** about confidence intervals. This is one of the main differences between the frequentist approach and Bayesian approach.

Bayesian Hypothesis Testing

How do we use Bayesian inference to test the following hypotheses?

$$\boldsymbol{H_0 : \pi \leq 0.1} \text{ vs } \boldsymbol{H_a : \pi > 0.1}$$

We simply need the posterior probabilities of these two ranges of values for π . Suppose that the Beta(10, 40) had been our prior, so our posterior distribution is Beta(17, 83). We can use a built-in R function to obtain P(π ≤ 0.1|y) as follows:

```
> pbeta(.1, 17, 83)
> [1] 0.01879825
```

- With this prior, we would conclude that P(π ≤ 0.1| y) = 0.019.
- Compare this result to the test of hypotheses of the frequentist approach, and state your conclusion?

Note: The interpretation of a Bayesian posterior probability is totally different from that of a frequentist p-value. Recall that a frequentist p-value is the probability, evaluated under the assumption that the null hypothesis is true, of drawing a random sample that contained as much evidence against the null as, or more than, the dataset we actually have. A frequentist p-value cannot be interpreted as the probability that the null hypothesis is true.

Bayesian Hypothesis Testing

- If instead unifrom prior is used, calculate $P(\pi \leq 0.1|y)$ and what is your conclusion?

- What is the probability that H_a is correct?

Questions?

Posterior Predictive Distributions

Posterior Predictive Distribution

- In many studies, the research question of interest is predicting values of a future sample from the same population.

- For example, suppose we are considering interviewing another sample of 50 VT students in the hope to get an idea of how it is likely to turn out before we go to the trouble of doing so!

- More generally, we are considering a new sample of sample size *n*∗ and want to estimate the probability of some particular number *y*∗ of successes in this sample. If we actually knew the true value π, we would just use the binomial probability:

$$P(y^*|\pi) = \binom{n^*}{y^*} \pi^{y^*}(1-\pi)^{n^*-y^*}, y^*=1,2,\ldots, n^*$$

  - But we can not use this. Why?

Posterior Predictive Distribution

- Thus, the **posterior predictive probability** of getting some particular value of y^* in a future sample of size n^* is

$$p(y^*|y) = \int_0^1 p(y^*|\pi)p(\pi|y)d\pi, \qquad y^* = 0,1, \ldots \ldots, n$$

where y denotes the data from the original sample and $p(\pi|y)$ is the posterior distribution.

- Using a *beta*(α, β) prior and a dataset with *y* successes in a sample of size *n*, then the posterior density *p*(π|*y*) is *beta*(α_{post}, β_{post}), where $\alpha_{post} = \alpha + y$ and $\beta_{post} = \beta + n - y$, and the predictive probability of getting y^* successes in a future sample of size n^* is

$$p(y^* \mid y) = \int_0^1 \binom{n^*}{y^*} \pi^{y^*} (1-\pi)^{n^*-y^*} \frac{\Gamma(\alpha_{post}+\beta_{post})}{\Gamma(\alpha_{post})\Gamma(\beta_{post})} \pi^{\alpha_{post}-1}(1-\pi)^{\beta_{post}-1} d\pi$$

$$= \binom{n^*}{y^*} \frac{\Gamma(\alpha_{post}+\beta_{post})}{\Gamma(\alpha_{post})\Gamma(\beta_{post})} \int_0^1 \pi^{y^*+\alpha_{post}-1}(1-\pi)^{n^*-y^*+\beta_{post}-1} d\pi$$

Posterior Predictive Distribution

$$p(y^* \mid y) = \int_0^1 \binom{n^*}{y^*} \pi^{y^*} (1-\pi)^{n^*-y^*} \frac{\Gamma(\alpha_{post}+\beta_{post})}{\Gamma(\alpha_{post})\Gamma(\beta_{post})} \pi^{\alpha_{post}-1} (1-\pi)^{\beta_{post}-1} d\pi$$

$$= \binom{n^*}{y^*} \frac{\Gamma(\alpha_{post}+\beta_{post})}{\Gamma(\alpha_{post})\Gamma(\beta_{post})} \int_0^1 \pi^{y^*+\alpha_{post}-1} (1-\pi)^{n^*-y^*+\beta_{post}-1} d\pi$$

- The expression inside the integral is the kernel of another beta density, $beta(y^* + \alpha_{post.}, n^* - y^* + \beta_{post.})$, so we can easily find to what it will integrate.

$$p(y^* \mid y) = \binom{n^*}{y^*} \frac{\Gamma(\alpha_{post}+\beta_{post})}{\Gamma(\alpha_{post})\Gamma(\beta_{post})} \frac{\Gamma(y^*+\alpha_{post})\Gamma(n^*-y^*+\beta_{post})}{\Gamma(\alpha_{post}+\beta_{post}+n^*)}$$

This distribution is known as the Beta-Binomial distribution

Posterior Predictive Distribution

- What is the posterior predictive probability of getting 5 successes, y*=5, when we select a sample of 20, n^*=20, when $\alpha_{post} = 3$, $\beta_{post} = 4$?

Posterior Predictive Distribution

- The function *pbeta**p*** in the R package *LearnBayes* calculates predictive probabilities.
- Go to R!

- For example, suppose our posterior distribution is beta(17, 83) and we are planning a new survey with sample size n^* = 25. We can get the probabilities of obtaining 3, 4, 5, or 6 “yesses” in that future sample by entering

```
library(LearnBayes)
> pbetap(c( 17, 83), 25, 3:6 )
> [1] 0.1795188 0.1898756 0.1625925 0.1168966
```

Posterior Predictive Distribution

- What is the most possible number of successes in a random sample of 25?

  Try this `plot(pbetap(c(17, 83), 25, 0:25 ) )` !!

Practice

Go to canvas, files, and find the problem sheet Hands-on #4 that is in “Hands-on Sheets” folder.