

Bayesian Statistics: A Primer for Data Scientists - PART I

antonieta.mira@usi.ch

Alan Agresti, Maria Kateri and Antonietta Mira

PLS DO NOT CIRCULATE THESE SLIDES

Notation

- *Upper-case letters for RVs: Y, X*
- *Lower-case letters for their observed values: y, x*
- Y for a RV that represents a **response** variable.
- x for the observed value of an **explanatory** variable.
- We use Greek letters for parameters: θ, π, μ
- We use **boldface** letters for vectors: **Y, y, θ**

Abbreviations

- RV = Random variable
- pdf = probability density function (continuous RV): $f(y|\theta)$
- pmf = probability mass function (discrete RV): $f(y|\theta)$
- LHD = likelihood function: $l(y|\theta) = l(\theta)$

Classical (Frequentist) versus Bayesian Statistics

Classical approach

- Parameters are unknown but regarded as taking fixed values.
- They are not RVs, so they do not have probabilities for their possible values.
- Probabilities apply to the RVs that yield the data, summarized by their probability distributions.

Example - Confidence intervals

Randomly samples 200 people who recently graduated from a particular state university.

Y = annual income in first job after graduation.

Goals include estimating the mean annual income μ and comparing the means of various groups (gender / race / subject major).

- Sample mean, in thousands of dollars, of $\bar{y} = 63.2$,
- 95% confidence interval for μ is (61.8, 64.6).

How is this confidence interval interpreted?

Example - Confidence intervals

Randomly samples 200 people who recently graduated from a particular state university.

Y = annual income in first job after graduation.

Goals include estimating the mean annual income μ and comparing the means of various groups (gender / race / subject major).

- Sample mean, in thousands of dollars, of $\bar{y} = 63.2$,
- 95% confidence interval for μ is (61.8, 64.6).

How is this confidence interval interpreted?

With the classical approach, it is *not* correct to say that the *probability* is 0.95 that μ falls between 61.8 and 64.6 thousand dollars.

The RV in the study, to which probabilities apply, is not μ but rather the sample mean \bar{Y} before we observe the data.

Once we observe the data and construct a particular confidence interval, that interval either *does* or *does not* contain μ .

We do not know which is the case for the specific sample at hand.

A probability such as 0.95 applies to the random interval, $\bar{Y} \pm$ (margin of error), *before* observing the data.

Different realizations of random samples have different observed \bar{y} values and yield different bounds for confidence intervals.

The interpretation of the confidence interval (61.8, 64.6) is that if one could repeatedly conduct this study, each time selecting a random sample of size 200 from the population of all recent graduates of the university and each time constructing a 95% confidence interval, then in the long run 95% of such confidence intervals would contain the unknown population mean μ .

Example: Hypothesis testing

In comparing the population mean incomes μ_1 of humanities graduates and μ_2 of science graduates, suppose the P -value in a t -test of $H_0: \mu_1 = \mu_2$ against $H_1: \mu_1 \neq \mu_2$ is 0.04.

How is the P -value defined?

Example: Hypothesis testing

In comparing the population mean incomes μ_1 of humanities graduates and μ_2 of science graduates, suppose the P -value in a t -test of $H_0: \mu_1 = \mu_2$ against $H_1: \mu_1 \neq \mu_2$ is 0.04.

How is the P -value defined?

Then 0.04 is the probability that the t test statistic based on the difference between the sample means $\bar{Y}_1 - \bar{Y}_2$ for the two groups divided by its standard error, takes a value like the one observed or even more extreme (i.e., even farther from 0 in either direction), in repeated samples of the same size, presuming that H_0 is true.

That is, the P -value applies to potential samples, when H_0 is true.

It is incorrect to interpret 0.04 as the probability that the null hypothesis H_0 is true, because H_0 deals with parameters

H_0 is either true or false. If we make a decision about H_0 using 0.05 significance level, then 0.05 represents the long-run proportion of times that we would mistakenly reject H_0 , if it were actually true.

Summary

- With the classical, so-called frequentist approach to statistical inference, probability statements do not apply to parameters.
- Probabilities cannot be stated for **hypotheses**, because the hypotheses refer to parameter values, and likewise probabilities do not apply to constructed **confidence intervals**.
- Classical statistics calculates probabilities about RVs and statistics such as test statistics that vary randomly from sample to sample, not about parameters.
- Statistics have sampling distributions, parameters do not. Those sampling distributions have *frequentist* interpretations in terms of what would happen in hypothetical repeated sampling of the same type, but that sampling does not actually occur.
- Because of this, interpretations of classical inferential tools such as P -values can be difficult for scientists and laypeople to understand, and they are often misinterpreted.

Bayesian Statistics: Probability Distributions for Observations *and* Parameters

- The Bayesian approach treats parameters as RVs.
- Because of this, statistical inferences can make probability statements directly about the parameters.
- Bayesians can make statement of this type with reference to a population mean μ : *“The probability is 0.95 that μ falls between 61.8 and 64.6 thousand dollars.”*
- A Bayesian inference for comparing population mean incomes μ_1 and μ_2 of humanities majors and science majors might be *“The probability is 0.02 that $\mu_1 > \mu_2$ and 0.98 that $\mu_1 < \mu_2$.”*
- Interpretations of Bayesian intervals and probabilities are **simpler** and more **natural** than the interpretations of frequentist confidence intervals and P -values.

- Overall **practical conclusions** derived by the classical frequentist and the Bayesian approaches to statistical inference are **usually substantively the same** as sample size increases.
- However the interpretations is quite different bwn the two approaches.

Concepts of Probability: Frequentist and Subjective

The difference in interpretation bwn frequentist and Bayesian methods stems from a difference in the meaning that the two approaches give to *probability*.

For studies that use *randomization* for gathering data, such as a randomized experiment or a random sample survey, frequentist statistical methods employ a definition relating to the relative frequency of that outcome in unobserved but like situations.

Frequentist definition of probability

For an observation of a random phenomenon, the ***probability*** of a particular outcome is the proportion of times that outcome would occur in an indefinitely long sequence of like observations under the same conditions.

This definition of probability is not always applicable.

An alternative definition of probability is subjective:

Subjective definition of probability

The probability of a particular outcome is a measure of the degree of belief of that outcome, based on all of the available information.

Bayesian statistics adopts the subjective definition of probability as its foundation

We now review some basic definitions and rules used with both frequentist and subjective definitions of probability.

Bayes' Theorem expresses the conditional probability of the event of interest, given a second event, in terms of a known conditional probability of the second event, given the event of interest.

Bayes' Theorem provides a basis for the methods of Bayesian statistics, which find conditional probabilities about parameter values, given the data, using conditional probabilities for the data, given parameter values.

Probabilities and Conditional Probabilities of Events

- **Sample space** = \mathcal{S} = set of all the possible outcomes
- **Event** = any subset of a sample space.
- For two events A and B , let $A \cup B = \text{union}$ = outcomes that are in A or in B or in both
- $AB = \text{intersection}$ = outcomes that are in A and in B .
- $P(A)$ = probability of an event A .

Probabilities of events satisfy rules based on three *axioms*:

- $P(A) \geq 0$
- $P(\mathcal{S}) = 1$
- For A and B disjoint
 $P(A \cup B) = P(A) + P(B)$.

The **conditional probability** of an event A , given that an event B occurred, is the fraction of the event B that is also in event A :

$$P(A | B) = \frac{P(AB)}{P(B)} = \frac{P(A \text{ and } B)}{P(B)},$$

provided that $P(B) > 0$.

The **multiplicative law of probability** re-expresses the conditional probability as

$$P(AB) = P(A | B)P(B) = P(B | A)P(A).$$

It follows that

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

We will see that:

$$\text{Posterior} = \frac{\text{LHD} \times \text{Prior}}{\text{Evidence}}$$

$A \Rightarrow$ parameter

$B \Rightarrow$ data

General form of Bayes Theorem

The diagnostic test example applied this result with B as the event D of having the disease and A as the event + of a positive diagnosis.

Bayes' Theorem generalizes to a **partition** of the sample space \mathcal{S} into $c \geq 2$ disjoint events $\{B_1, B_2, \dots, B_c\}$, that is, events such that $S = B_1 \cup B_2 \cup \dots \cup B_c$ with B_i and B_j disjoint for each pair.

For each $k = 1, \dots, c$,

$$P(B_k | A) = \frac{P(AB_k)}{P(A)} = \frac{P(A | B_k)P(B_k)}{P(A | B_1)P(B_1) + P(A | B_2)P(B_2) \cdots P(A | B_c)P(B_c)}.$$

Bayes' Theorem generalizes from events to RV.

For **discrete** RVs, the outcomes are the distinct, separate values that the RV can assume, usually integers.

Continuous RVs have an infinite continuum of possible values. Their probability distributions assign probabilities to *intervals* of real numbers rather than individual values.

When X and Y are both continuous RVs, a version of Bayes' Theorem finds a *probability density function (pdf)* $g(x|y)$ from the *pdf* $f(y|x)$.

With *joint pdf* $f(x,y)$ and *marginal pdfs* $f_1(x)$ for X and $f_2(y)$ for Y :

$$g(x|y) = \frac{f(x,y)}{f_2(y)} = \frac{f(y|x)f_1(x)}{\int_{\mathcal{X}} f(y|\tilde{x})f_1(\tilde{x})d\tilde{x}}$$

For discrete RVs X and Y , each function in the theorem is a *probability mass function (pmf)*, and the denominator has a sum instead of an integral.

Another version of Bayes' Theorem permits one of (X, Y) to be discrete and one to be continuous.

The Bayesian Approach to Statistical Inference

Bayesian statistical inferences apply Bayes' Theorem to find probabilities about the parameter values, conditional on the data, using probabilities for the data, conditional on the parameter values.

Bayesian Prior and Posterior Distributions

- The Bayesian approach to statistical inference assumes a **prior distribution** for the parameters that reflects information available about the parameters before we observe the data.
- That information might be based on other studies or on beliefs of “experts”.
- The prior distribution may be relatively uninformative, so that inferential results are less subjective, based almost entirely on the observed data.
- The prior distribution combines with the information that the data provide to generate a **posterior distribution** for the parameters.
- Bayesian statistical inferences are based on the posterior distribution.

Posterior distribution: computation

Ingredients:

- A parameter θ with parameter space Θ for its possible values
- For an observation y : $f(y | \theta)$ denote the *probability function*, for a given value of the parameter. This is a *probability density function (pdf)* or a *probability mass function (pmf)*, according to whether Y is a continuous or a discrete RV.
- For n observations $\mathbf{y} = (y_1, \dots, y_n)$, let $f(\mathbf{y} | \theta)$ denote their *joint probability function*, given the parameter value.
- For independent observations, such as obtained with a simple random sample or an experiment employing randomization,

$$f(\mathbf{y} | \theta) = f(y_1 | \theta)f(y_2 | \theta)\cdots f(y_n | \theta).$$

- Let $p(\theta)$ denote the *pdf* for the *prior distribution* of θ .
- Given the n observations $\mathbf{y} = (y_1, \dots, y_n)$, we let $g(\theta | \mathbf{y})$ denote the probability function for the *posterior distribution* of θ , given the data.

Bayes' Theorem to obtain the posterior

- By *Bayes' Theorem*, combining the joint probability function $f(\mathbf{y} | \theta)$ with the prior pdf $p(\theta)$ we obtain the posterior of θ

$$g(\theta | \mathbf{y}) = \frac{f(\mathbf{y} | \theta)p(\theta)}{f(\mathbf{y})} = \frac{f(\mathbf{y} | \theta)p(\theta)}{\int_{\Theta} f(\mathbf{y} | \tilde{\theta})p(\tilde{\theta})d\tilde{\theta}}. \quad (1)$$

- The denominator $f(\mathbf{y})$ is the marginal probability function of the data, obtained by integrating out the parameter.
- In terms of θ , $g(\theta | \mathbf{y})$ is proportional to $f(\mathbf{y} | \theta)p(\theta)$:
 $g(\theta | \mathbf{y}) \propto f(\mathbf{y} | \theta)p(\theta) = \ell(\theta)p(\theta)$
- That product of LHD (likelihood function), $\ell(\theta)p(\theta)$ and prior determines the posterior distribution of θ , because the denominator $f(\mathbf{y})$ does not involve θ .
- Except in a few simple cases we *cannot* identify the posterior distribution from the product $f(\mathbf{y} | \theta)p(\theta)$ and will thus use simulation methods.

When the prior is relatively flat, the posterior has similar shape as the LHD.

Beta-Binomial example

Bernoulli distribution: the *pmf* for each Y_i is $f(y_i | \pi) = \pi^{y_i}(1 - \pi)^{1-y_i}$, which yields the probabilities π when $y_i = 1$ and $1 - \pi$ when $y_i = 0$.

The joint *pmf* for the n independent and identically distributed (iid) Bernoulli observations is

$$f(\mathbf{y} | \pi) = \prod_{i=1}^n \pi^{y_i}(1 - \pi)^{1-y_i} = \pi^{\sum_i y_i} (1 - \pi)^{n - \sum_i y_i}.$$

Binomial Likelihood with Uniform Prior Induces Beta Posterior Distribution

- Goal: estimate the population proportion π when the data consist of n independent observations of a binary response variable with “success” and “failure” outcomes.
- The number of successes y in n observations = binomial RV
- View the LHD in terms of the part of $\ell(\theta) = f(\mathbf{y} | \theta)$ that involves θ .
- The binomial LHD $\ell(\pi) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}$ is proportional to $\pi^y (1 - \pi)^{n-y}$, so it simplifies to $\ell(\pi) = \pi^y (1 - \pi)^{n-y}$, ignoring $\binom{n}{y}$.
- Since π is a probability, a Bayesian prior distribution for π is defined over the interval $[0, 1]$ of its possible values.

- Lacking any prior information about π , a data scientist could use a standard **uniform distribution**, which has *pdf* that is uniformly spread but positive only over that interval,

$$p(\pi) = 1, \quad 0 \leq \pi \leq 1.$$

- Combining the uniform prior distribution with the binomial likelihood function, from Bayes Theorem the posterior *pdf* of π is

$$g(\pi | y) \propto \ell(\pi)p(\pi) = [\pi^y(1 - \pi)^{n-y}] \cdot 1, \quad 0 \leq \pi \leq 1.$$

- This prior distribution is relatively uninformative, in the sense that the posterior distribution has exactly the same shape as the binomial likelihood function, for $0 \leq \pi \leq 1$.

- This posterior distribution is a special case of the **beta distribution**.
- Its *pdf* $g(\pi | \alpha_1, \alpha_2)$ is proportional to

$$g(\pi | \alpha_1, \alpha_2) \propto \pi^{\alpha_1-1} (1 - \pi)^{\alpha_2-1}, \quad 0 \leq \pi \leq 1, \quad (2)$$

for $\alpha_1 > 0$ and $\alpha_2 > 0$ and a proportionality constant (ignored here) involving α_1 and α_2 so that $g(\pi | \alpha_1, \alpha_2)$ integrates to 1 over the interval $[0, 1]$.

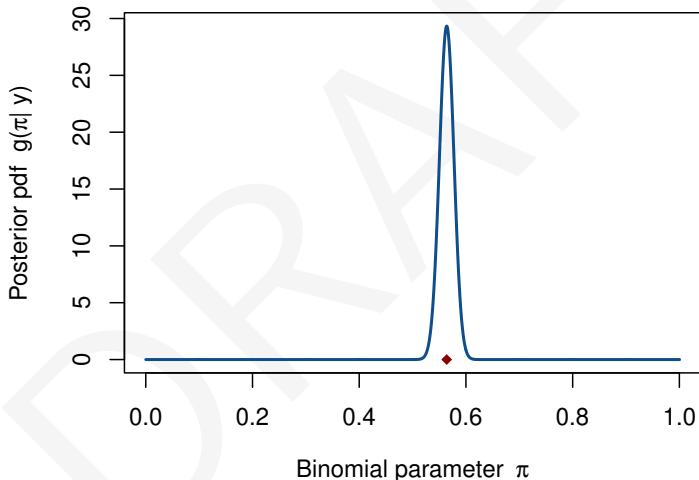
- Since α_1 and α_2 are not the parameters of main interest for the analysis but merely determine the shape of the beta distribution, they are referred to as **hyperparameters**.
- The uniform distribution is the special case of a beta distribution with $\alpha_1 = \alpha_2 = 1$.
- The mean of the beta distribution is $\alpha_1 / (\alpha_1 + \alpha_2)$, which equals $1/2$ when $\alpha_1 = \alpha_2$.

- With the uniform prior distribution for π , the posterior *pdf* $g(\pi | y) \propto \pi^y(1 - \pi)^{n-y}$ for $0 \leq \pi \leq 1$ has the form of a beta distribution.
- Equating $y = \alpha_1 - 1$ and $n - y = \alpha_2 - 1$ in (2), we find that the beta hyperparameters are $\alpha_1 = y + 1$ and $\alpha_2 = n - y + 1$.
- The mean of the posterior distribution is $\alpha_1 / (\alpha_1 + \alpha_2) = (y + 1) / (n + 2)$.
- With a relatively large sample size n , this beta distribution is approximately bell-shaped with mean close to y/n , which is the sample proportion for the observed data.

Example: Proportion Supporting a Woman's Choice about Abortion

- One item in the 2021 GSS asked whether a woman should be able to get an abortion if she wants it for any reason.
- Of 1328 respondents, 749 said *yes* and 579 said *no*.
- The binomial outcome is $y = 749$ in $n = 1328$ observations.
- The sample proportion is $y/n = 749/1328 = 0.564$.
- With the uniform prior distribution, the Bayesian posterior *pdf* $g(\pi | y)$ of π is beta with hyperparameters $\alpha_1 = y + 1 = 750$ and $\alpha_2 = n - y + 1 = 580$.
- Figure ?? shows this beta *pdf*.

Beta posterior *pdf* with hyperparameters $\alpha_1 = 750$ and $\alpha_2 = 580$ for population proportion who believe a woman should be able to get an abortion for any reason. The MLE is located at the red diamond



According to this posterior distribution, the range of plausible values for π is narrow, quite close to the sample proportion 0.564. The posterior $P(\pi < 0.50) = 0.0000015$.

This value can be computed with R.

We can cumulate the probability at 0.50 for beta distribution with hyperparameters 750, 580 and obtain 0.0000015

```
> pbeta(0.50, 750, 580)
[1] 1.510934e-06
```

A Criticism: Bayesian Inferences Are Subjective

- A criticism to the Bayesian approach was that it is *subjective* rather than *objective*, because of the need to choose a prior
- However, classical frequentist methods also make assumptions, such as the choice of probability distribution $p(y | \theta)$
- The nature of the data usually suggests natural choices and we can use the data to check such assumptions
- See for example the assumptions in a regression analysis
- With large samples, the choice of prior distribution is usually not crucial, because the posterior relies mainly on the LHD
- When the sample size is not large, one can usually select relatively objective prior distribution, which is typically highly disperse
- Supporters of the Bayesian approach argue that the resulting statistical inferences are more natural than frequentist inferences in being able to make probability statements directly about parameter values.

Bayesian Point Estimates

- Bayesian point estimators of a parameter θ use summary measures of central tendency of the posterior distribution
- A possible Bayesian estimator is the point at which the posterior distribution is maximized, which is its *mode*. However, it is more common to use the **posterior mean**,

$$E(\theta | \mathbf{y}) = \int_{\Theta} \theta g(\theta | \mathbf{y}) d\theta.$$

- The **posterior median** is another possible Bayesian estimator. The posterior probability falling above that value and the posterior probability falling below it both equal 1/2.
- With large n , posterior distributions are typically **approximately normal**.
Then, such posterior summaries are all close to each other, and they usually take similar values as the MLE.

Example: General Social Survey

From the GSS data we found the posterior distribution of the population proportion π in the U.S. who believe a woman should have the right to an abortion for any reason.

With a uniform prior and the survey results in which $y = 749$ said *yes* and $n - y = 579$ said *no*, the posterior distribution of π is a beta with hyperparameters $\alpha_1 = y + 1 = 750$ and $\alpha_2 = n - y + 1 = 580$.

The posterior mean estimate of π is

$E(\pi | y) = \alpha_1 / (\alpha_1 + \alpha_2) = (y + 1) / (n + 2) = 0.5639$, the same to three decimal places as the MLE, $\hat{\pi} = y/n = 749/1328 = 0.5640$.

The posterior mode is $(\alpha_1 - 1) / (\alpha_1 + \alpha_2 - 2) = y/n$, the same as the MLE.

With R software, using the quantile function for beta distributions, we find that the posterior median is also the same as the MLE to three decimal places:

The 0.50 quantile (median) for beta dist. with hyperparameters 750, 580 is obtained with the R command

```
> qbeta(0.50, 750, 580)
[1] 0.5639418
```

Shrinkage with a Bayesian Posterior Mean Estimate

For Bayesian estimation of a binomial parameter π using a uniform prior distribution we found that the posterior mean estimate of π is $E(\pi | y) = (y + 1)/(n + 2)$.

We can express this estimate as

$$E(\pi | y) = \frac{y + 1}{n + 2} = \left(\frac{n}{n + 2} \right) \frac{y}{n} + \left(\frac{2}{n + 2} \right) \frac{1}{2}.$$


This is a weighted average of the MLE (the sample proportion), $\hat{\pi} = y/n$, and the mean of the uniform prior distribution, $1/2$.

The estimate shrinks the sample proportion toward 0.50. The weight $n/(n + 2)$ given to the sample proportion increases toward 1 as n increases.

For the example on opinions about a woman's choice regarding abortion, the Bayes estimate with the uniform prior gives weight $n/(n+2) = 1328/1330 = 0.9985$ to the sample proportion.

The Bayes estimate corresponds to a sample mean of $n+2$ observations of Bernoulli RVs, combining the n sample observations with 2 imaginary prior observations, of which 1 is a success and 1 is a failure.

This behavior is typical of Bayesian posterior mean estimates. For independent samples from the most commonly used probability distributions with their standard prior distributions,¹ this estimate is a weighted average of the MLE and the mean of the prior distribution, with relatively more weight given to the MLE as the sample size increases.

¹Technically, the *natural exponential family* of distributions 

Decision-Theoretic Evaluation of Estimators

The posterior mean and posterior median estimators of a parameter can be justified as minimizing the expected value of a measure of the distance that the estimator falls from the parameter that it estimates.

This is a main result of ***statistical decision theory***, which is concerned with how to make optimal decisions in the face of uncertainty.

A basic tool of statistical decision theory is the ***loss function***: a function of the parameter θ and the estimator $\hat{\theta}$.

A common one is the *squared error* loss function,

$$L(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2.$$

For example, to evaluate the sample mean \bar{Y} as an estimator of a population mean μ , $L(\mu, \bar{Y}) = (\bar{Y} - \mu)^2$.

The loss function refers to a single sample and is itself a RV. We can evaluate an estimator $\hat{\theta}$ by the expected value of this loss function with respect to the probability function $f(\mathbf{y} | \theta)$ of the observations \mathbf{y} , given the value of the parameter θ ,

$$R_{\hat{\theta}}(\theta) = E_{\mathbf{y}}[L(\theta, \hat{\theta})] = \int_{\mathbf{y}} L(\theta, \hat{\theta}(\mathbf{y})) f(\mathbf{y} | \theta) d\mathbf{y},$$

called the **risk function**.

The expectation here is taken over the space \mathcal{Y} of possible values of \mathbf{y} with θ held fixed, so the risk function $R_{\hat{\theta}}(\theta)$ is a function of θ .

For the squared-error loss function,

$$R_{\hat{\theta}}(\theta) = E(\hat{\theta} - \theta)^2 = \text{MSE},$$

the *mean squared error* for the estimator.

For example, with n independent observations, using \bar{Y} as an estimator of μ in a population having variance σ^2 , the risk function is $R_{\bar{Y}}(\mu) = E(\bar{Y} - \mu)^2 = \sigma^2/n$. This is the variance of the estimator and is the same for all μ .

Example For a RV Y having a Binom(n, π) distribution, we now compare the risk functions with a squared-error loss function for the MLE $\hat{\pi} = Y/n$ and the Bayes estimator $\tilde{\pi} = (Y + 1)/(n + 2)$ based on a uniform prior distribution.

The risk function for $\hat{\pi}$ is

$$R_{\hat{\pi}}(\pi) = E(\hat{\pi} - \pi)^2 = \text{var}(\hat{\pi}) = \frac{\pi(1 - \pi)}{n}.$$

You can derive² that the risk function for $\tilde{\pi}$ is

$$R_{\tilde{\pi}}(\pi) = E(\tilde{\pi} - \pi)^2 = \left(\frac{n}{n+2}\right)^2 \frac{\pi(1-\pi)}{n} + \left(\frac{2}{n+2}\right)^2 \left(\pi - \frac{1}{2}\right)^2. \quad (3)$$

At $\pi = 0.5$, $R_{\tilde{\pi}}(\pi) = [n/(n+2)]^2(1/4n) < 1/4n = R_{\hat{\pi}}(\pi)$.

At $\pi = 0$ or $\pi = 1$, $R_{\tilde{\pi}}(\pi) = 1/(n+2)^2 > 0 = R_{\hat{\pi}}(\pi)$.

These suggest that the Bayes estimator is better in the middle of the range of possible π values whereas the MLE is better when π is close to 0 or 1.

Figure ?? illustrates, showing the two risk functions when $n = 10$ and when $n = 50$.

²From the decomposition of MSE into variance plus squared bias, to be reviewed in Section ??.

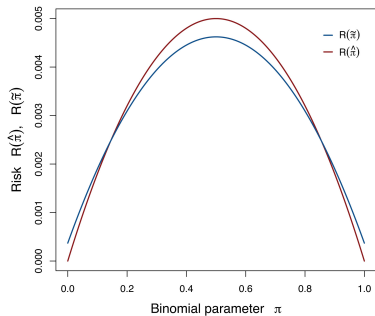
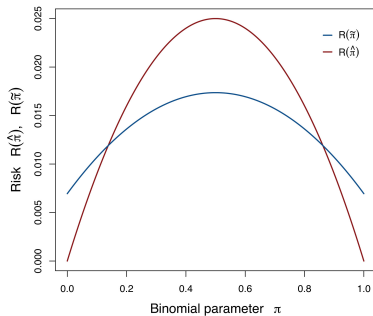


Figure: Risk functions for MLE $\hat{\pi}$ (red) and Bayes estimator $\tilde{\pi}$ (blue), for estimating binomial parameter π when $n = 10$ (left) and when $n = 50$ (right).

In a Bayesian framework, with a parameter θ treated as a RV rather than fixed, the overall evaluation of an estimator $\hat{\theta}$ averages the risk function with respect to the prior distribution $p(\theta)$ for θ . This yields the **Bayesian risk**,

$$r_p(\hat{\theta}) = E_{\theta}[R_{\hat{\theta}}(\theta)] = E_{\theta}\{E_{\mathbf{y}}[L(\theta, \hat{\theta})]\} = \int_{\Theta} \int_{\mathbf{y}} L(\theta, \hat{\theta}(\mathbf{y})) f(\mathbf{y} | \theta) p(\theta) d\mathbf{y} d\theta.$$

Any estimator that minimizes the Bayes risk for some loss function is called a **Bayes estimator**.

Recall that the posterior *pdf* $g(\theta | \mathbf{y}) \propto f(\mathbf{y} | \theta) p(\theta)$, and in fact it can be proved that a Bayesian estimator minimizes the *posterior expected loss*

$$E_{\theta}[L(\theta, \hat{\theta}) | \mathbf{y}] = \int_{\Theta} L(\theta, \hat{\theta}(\mathbf{y})) g(\theta | \mathbf{y}) d\theta.$$

For the squared-error loss function, the posterior expected loss is

$$\begin{aligned} \int_{\Theta} L(\theta, \hat{\theta}(\mathbf{y}))g(\theta | \mathbf{y})d\theta &= \int_{\Theta} (\theta - \hat{\theta})^2 g(\theta | \mathbf{y})d\theta = \\ &= \hat{\theta}^2 - 2\hat{\theta} \int_{\Theta} \theta g(\theta | \mathbf{y})d\theta + \int_{\Theta} \theta^2 g(\theta | \mathbf{y})d\theta. \end{aligned}$$

To minimize this, we differentiate with respect to $\hat{\theta}$ and equate to 0, obtaining

$$2\hat{\theta} - 2 \int_{\Theta} \theta g(\theta | \mathbf{y})d\theta = 0.$$

Since the second derivative is positive, we obtain the minimum of the posterior expected loss at $\hat{\theta} = \int_{\Theta} \theta g(\theta | \mathbf{y})d\theta$. That is, for the squared-error loss function, the Bayes estimator of θ is the *posterior mean* $E(\theta | \mathbf{y})$.

For the absolute-error loss function

$$L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|,$$

one can show that the Bayes estimator of θ is the *posterior median*.

Bayesian Posterior Intervals

Using the posterior distribution, we can form a Bayesian interval estimate of the parameter of interest θ that is analogous to a frequentist confidence interval.

For example, analogous to the frequentist 95% confidence interval is an interval that contains 95% of the mass of the posterior *pdf* $g(\theta | \mathbf{y})$.

Such an interval is called a ***posterior interval*** or ***credible interval***.

Percentile Intervals and Highest Posterior Density (HPD) Intervals

A simple way to form a posterior interval uses percentiles of $g(\theta | \mathbf{y})$ with equal tail probabilities.

For example, the 95% posterior equal-tail **percentile interval** for θ is the region between the 2.5 and 97.5 percentiles (0.025 and 0.975 quantiles) of the posterior distribution.

An alternative Bayesian posterior interval contains the desired probability such that the posterior *pdf* is higher over all values in the interval than over all the values not in it.

This posterior interval is called a **highest posterior density** (HPD) interval.

The HPD interval for a parameter is the shortest possible interval having the desired probability. It is *identical to the percentile-based interval when the posterior distribution is unimodal and symmetric*. It is usually preferred to the percentile-based interval when the posterior distribution is highly skewed.

Proportion Supporting a Woman's Choice Revisited

For the survey on abortion the posterior distribution of the population proportion π who support a woman being able to get an abortion whenever she wants it is the beta distribution with hyperparameters $\alpha_1 = 750$ and $\alpha_2 = 580$.

The 95% **posterior equal-tail percentile interval** has as its endpoints the 0.025 and 0.975 quantiles of that posterior.

The following R output shows that this interval is (0.537, 0.590).

Here we utilize the `binom` package in R, which can provide a wide variety of intervals for the binomial parameter.

The output also displays the **frequentist large-sample 95%** confidence interval $\hat{\pi} \pm 1.96\sqrt{\hat{\pi}(1 - \hat{\pi})/n}$ that uses the maximum likelihood estimate $\hat{\pi} = y/n$ and its estimated standard error.

The output also shows that the **HPD interval**, which is the default provided with the `binom.bayes` function in the `binom` package, is (0.537, 0.591).

Since the sample size $n = 1328$ is quite large, the Bayesian intervals are similar to each other and to the frequentist confidence interval.

Quantiles of beta to get the 95% posterior percentile interval

```
> qbeta(c(0.025, 0.975), 750, 580)
```

```
[1] 0.5371820 0.5904555
```

```
> install.packages("binom")
```

```
> library(binom)
```

Frequentist 95% confidence interval:

```
> binom.confint(749, 1328, conf=0.95, method="asymptotic")
```

```
0.5373355 0.5906765
```

HPD interval with uniform prior dist. (which is beta with
alpha1 = alpha2 = 1)

```
> binom.bayes(749,1328,conf=0.95, alpha1=1.0, alpha2=1.0)
```

```
method x n shape1 shape2 mean lower upper
```

```
bayes 749 1328 750 580 0.5639098 0.5372464 0.5905193
```


If the posterior distribution had not been a standard one available in software, we could have found an excellent approximation of the percentile interval by simulating a huge number of observations from the posterior distribution and then finding the relevant sample percentiles (quantiles).

Simulate 50 million observations from posterior beta distribution
approximate Bayesian 95% percentile interval with quantiles of
the sampled observations

histogram of posterior dist. (not shown here)

```
> postvalues <- rbeta(50000000,750,580)
```

```
> quantile(postvalues, c(0.025, 0.975))
```

```
2.5% 97.5%
```

```
0.5371813 0.5904545
```

```
> hist(postvalues)
```

In some cases, the **posterior *pdf* takes highest value at the boundary of the parameter space**, such as at the value 0 or the value 1 for a population proportion π , and the posterior *pdf* is monotone decreasing as one moves away from the boundary.

In such cases, the **HPD interval is more appropriate than the percentile interval**, because the percentile interval would not contain the parameter values that have the highest posterior density values.

Interpretation: Bayesian Posterior Intervals versus Frequentist Confidence Intervals

In the example of estimating the population proportion π in the U.S. supporting the right of a woman to have an abortion, the Bayesian point estimate and 95% posterior intervals induced by a uniform prior are nearly identical to the MLE and frequentist 95% confidence interval.

This reflects that n is large and that the prior distribution is highly disperse and thus carries little *a priori* information.

As a consequence, nearly all the information contained in the posterior distribution comes from the data and is summarized by the likelihood function.

Although the frequentist and Bayesian results are nearly identical, the interpretations are quite different.

In the **frequentist approach**, the parameter π is fixed, not a RV. It either *is* or *is not* in the 95% confidence interval (0.537, 0.591). We do not know which is the case.

Our 95% confidence refers to a probability when the data (not the parameter) are viewed as the RV, that is, before they are observed. It has the frequentist interpretation that if we used this method repeatedly with independent hypothetical samples, in the long run 95% of the confidence intervals would contain the true π value.

The probability applies to possible data in future samples that we will not observe, rather than to the parameter.

By contrast, with the **Bayesian approach**, π is itself a RV and has a probability distribution.

After observing the data and constructing the posterior interval (0.537, 0.590), we can conclude that the probability is 0.95 that π takes value between 0.537 and 0.590.

Although their interpretations differ, Bayesian and frequentist approaches usually lead to the **same practical conclusions**, because the LHD is the foundation of each.

The set of parameter values regarded as plausible in a frequentist inference are usually similar to those regarded as plausible with a Bayesian inference.

The resemblance increases as n increases or as the variance of the prior distribution increases, because the posterior distribution then has maximum and shape increasingly similar to the maximum and shape of the likelihood function.

Bayesian Significance Testing and Prediction

Besides point and interval estimation, the other principal frequentist statistical inferential method is significance testing of a null hypothesis about the parameter value.

Prediction of future observations is also important in many applications.

We next present a Bayesian analog of significance testing and a Bayesian method for predicting future observations.

Significance Testing based on Posterior Probabilities for Parameter Regions

With a continuous prior distribution, the posterior distribution is also continuous.

Then, the posterior probability of any single value for a parameter such as a population proportion π is zero.

This accords with intuition in most applications that null hypothesis conditions such as $\pi = 0.50$ *exactly* are implausible.

It is usually more relevant to summarize the evidence that $\pi < 0.50$ versus $\pi > 0.50$.

We can do this by reporting relevant posterior tail probabilities. When neither posterior $P(\pi > 0.50 | y)$ nor $P(\pi < 0.50 | y)$ is close to 0, we regard 0.50 as one of the plausible values for π .

Ways exist of setting prior distributions that are a mixture of continuous and discrete so that conditions such as a point null hypothesis $H_0: \pi = 0.50$ have positive posterior probability.

But that posterior probability can then depend strongly on the choice of prior distribution.

An alternative approach to be introduced summarizes information about two hypotheses by forming a “Bayes factor” based on the ratio of the posterior probabilities of the two hypotheses.

Example: Proportion Supporting a Woman's Choice Revisited

For the abortion data with a uniform prior, the posterior of the population proportion π supporting a woman's right to have an abortion is beta with hyperparameters $\alpha_1 = 750$ and $\alpha_2 = 580$.

We now use this posterior to find the posterior $P(\pi > 0.50 \mid y)$ and $P(\pi < 0.50 \mid y)$, to analyze whether we can conclude that a majority or a minority of the population support this right.

We can determine these by using R to find the cumulative probability of the posterior distribution at the value 0.50.

Cumulative probability at 0.50 for beta posterior distribution

```
> pbeta(0.50, 750, 580)  
[1] 1.510934e-06
```

The posterior $P(\pi < 0.50 \mid y) = 0.0000015$.

Thus, $P(\pi > 0.50 \mid y) = 0.9999985$, indicating extremely **strong evidence that a majority of the population believe that a woman should be able to get an abortion for any reason.**

A corresponding result with the **frequentist approach** uses the P -value for testing

$H_0: \pi = 0.50$ (implicitly $H_0: \pi \leq 0.50$) against

$H_1: \pi > 0.50$.

Since $y = 749$, the P -value is the probability of observing $Y \geq 749$ out of $n = 1328$ observations when actually H_0 is true, which is $1 - P(Y \leq 748 \mid \pi = 0.50) = 0.0000017$.

The one-sided (right-tail) P -value for binomial distribution equals 1 - cumulative probability and can be obtained in R as:

```
> 1 - pbinom(748, 1328, 0.50)
```

```
[1] 1.712424e-06
```

If we conducted a formal frequentist significance test with a probability of Type I error such as 0.05, the result would be in line with the Bayesian one in terms of concluding that $\pi > 0.50$.

However, the interpretations of the probabilities are quite different.

The value 0.0000015 is the Bayesian posterior probability that $\pi < 0.50$, whereas the frequentist P -value of 0.0000017 merely relates to hypothetical samples if H_0 were true, that is, the probability of a sample like observed or even more extreme if actually $\pi = 0.50$.

Predicting Future Observations

Sometimes the main goal of a statistical analysis is to predict future observations.

A Bayesian **posterior predictive distribution** is the probability distribution for a future observation Y_f .

Given the data, the posterior predictive *pdf* is

$$\begin{aligned}h(y_f | \mathbf{y}) &= \int_{\Theta} f(y_f | \theta, \mathbf{y}) g(\theta | \mathbf{y}) d\theta \\ &= \int_{\Theta} f(y_f | \theta) g(\theta | \mathbf{y}) d\theta,\end{aligned}$$

where $f(y_f | \theta, \mathbf{y}) = f(y_f | \theta)$ because Y_f is independent of the sample data \mathbf{y} , given θ . So, we obtain $h(y_f | \mathbf{y})$ by averaging the probability function $f(y_f | \theta)$ for known θ with respect to the posterior distribution $g(\theta | \mathbf{y})$ of θ . For binary data with $\pi = P(Y_f = 1 | \pi)$, we can prove that the posterior predictive $P(Y_f = 1 | \mathbf{y})$ is the mean of the posterior distribution of π .

With the frequentist approach to statistical inference, it is not possible to use such integration to obtain $h(y_f | \mathbf{y})$, because θ is not a RV with its own distribution.

Except in a few cases, making probabilistic predictions about future observations is not straightforward.

With large n , a simple approximation uses $f(y_f | \hat{\theta})$ as a predictive distribution, acting as if θ equals its MLE $\hat{\theta}$.

We will show that Bayesian posterior predictive distributions are also useful for checking assumptions of the model on which the analyses are based.

One such check compares simulated observations from the posterior predictive distribution with the observed data.

The simulated data should look like the observed data in terms of central tendency and variability.

Noninformative Prior Distributions

We now introduce two types of priors used when we want to have little influence on the posterior and thus on statistical inferences.

To eliminate the subjective aspect of the choice of the prior distribution, Bayesians can use a *noninformative prior distribution*, containing only vague information about the parameter.

An example is a **uniform** distribution for the parameter, for which the posterior distribution is merely the LHD re-scaled to integrate out to 1. We used this approach for the binomial parameter.

If the parameter can take any real-number value, such as the mean μ of a normal distribution, we could take the prior distribution to be uniform over the entire real line, that is, $f(\mu) = c$ for $-\infty < \mu < \infty$.

However, this is not a legitimate *pdf*, because its integral over the entire real line is infinite rather than 1.

A function such as this is called an ***improper prior distribution***.

An improper prior distribution can also be non-uniform. For instance, when the parameter can take any positive value, such as the standard deviation σ of a normal distribution, some Bayesian methods use $f(\sigma) = 1/\sigma$ for $\sigma > 0$, which also has infinite integral over the possible values.

Bayesian statistical methods can use an improper prior distribution as long as the posterior distribution that it induces is proper.

When we use a uniform prior distribution over the parameter space, the posterior distribution is the LHD re-scaled.

Bayesian inferences are then very similar to frequentist statistical inferences. For example, posterior intervals may be identical or nearly identical to frequentist confidence intervals, although the interpretation is simpler for Bayesian intervals.

Another commonly-used noninformative prior distribution is the ***Jeffreys prior distribution***.

It is specified so that posterior results are equivalent regardless of the scale of measurement for the parameter.