Bayesian Inference

Learning Objectives

- Understand the basic principles underlying Bayesian modeling methodology
- Understand how to use Bayesian inference for real-world problems
- Understand the computational techniques required
  - (how to turn the Bayesian crank)
- Appreciate the need for sensitivity analysis, model checking and comparison, and the potential dangers of Bayesian methods

Part 0:
What is Bayesian stats all about?

Bayesian inference

- Probabilities numerically represent beliefs about unknown quantities
- Bayes rule provides a rational method for updating those beliefs in light of new information
  - This inductive learning is Bayesian inference
- Bayesian methods are data analysis tools derived from the principles of Bayesian inference

Bayesian methods provide

- models for rational, quantitative learning
- parameter estimates with good statistical properties
- estimators that work for small and large sample sizes
- parsimonious descriptions of data, predictions for missing data, and forecasts for future data
- a coherent computational framework for model estimation, selection and validation
- methods for generating statistical procedures in complicated problems

An alternative to?

- Maximum likelihood (ML)
- Likelihood ratio tests:
  - t-test, F-tests, Chi-squared tests
- Complicated frequency arguments where we must imagine re-applying (ML) inference on “similar” data
- Appeals to “asymptopia” and the central limit theorem
- the Bootstrap, etc.

Essence of Bayesian inference

is ....

- The inductive process of learning about the general characteristics $\theta \in \Theta$ of a population $\mathcal{Y}$ from a subset $y \in \mathcal{Y}$
- *Both* θ and y are uncertain
- The information obtained in a particular data set y can be used to decrease our uncertainty about θ
- Quantifying this change is the purpose of Bayesian inference: this is Bayesian learning or updating

Bayesian learning

... begins with a numerical formulation of joint beliefs about θ and y expressed in terms of probability distributions over Θ and $\mathcal{Y}$

- For each numerical value $\theta \in \Theta$, our prior distribution $p(\theta)$ describes our belief that θ represents the true population characteristics
- For each $\theta \in \Theta$ and $y \in \mathcal{Y}$, our sampling model $p(y|\theta)$ describes our belief that y would be the outcome of our study if we knew θ to be true

Bayesian learning

Once we obtain the data y, the last step is to update our beliefs about θ

- For each numerical value $\theta \in \Theta$, our posterior distribution $p(\theta|y)$ describes our belief that θ is the true value, having observed the data set y

The posterior distribution is obtained from the prior distribution and sampling model via Bayes' rule

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_\Theta p(y|\tilde{\theta})p(\tilde{\theta})\, d\tilde{\theta}}$$

Why Bayes?

- The mathematical results of Cox (1946, 1961) and Savage (1954, 1972) prove that if $p(\theta)$ and $p(y|\theta)$ represent a rational person's beliefs, then Bayes' rule is an optimal method of updating this person's beliefs about θ given the new information y
  - justifies using Bayes' rule for quantitative learning
- In practice it can be hard to precisely mathematically formulate prior beliefs
  - $p(\theta)$ often chosen in an ad hoc manor, or for reasons of computational tractability

Why Bayes?

So how can we justify using Bayesian data analysis?

- "All models are wrong, but some are useful." (Box and Draper, 1987, pp. 424)
- if $p(\theta)$ approximates our prior beliefs then $p(\theta|y)$ shall approximate our posterior beliefs

In many complicated statistical problems there are no obvious non-Bayesian methods of inference

- Bayes' rule can be used to generate estimators
- performance evaluated with non-Bayesian criteria

Example:
Estimating the probability of a rare event

Suppose we are interested in the prevalence of an infectious disease in a small city. A small random sample of 20 individuals will be checked for infection

- Interest is in the fraction of infected individuals

$$\theta \in \Theta = [0, 1]$$

- The data records the number infected individuals

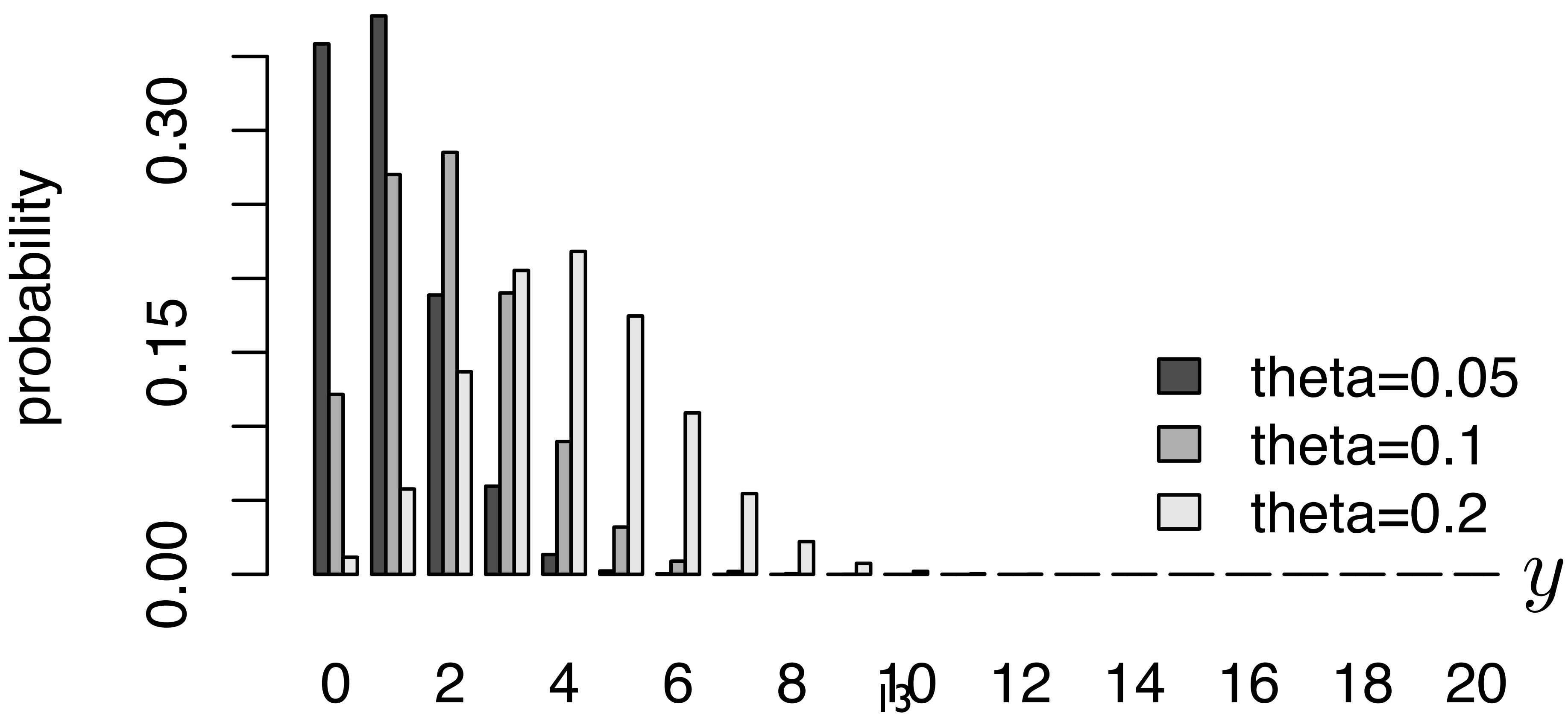
$$y \in \mathcal{Y} = \{0, 1, \ldots, 20\}$$

Example: sampling model

Before the sample is obtained, the number of infected individuals is unknown

- Let Y denote this to-be-determined value
- If θ were known, a sensible sampling model is

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

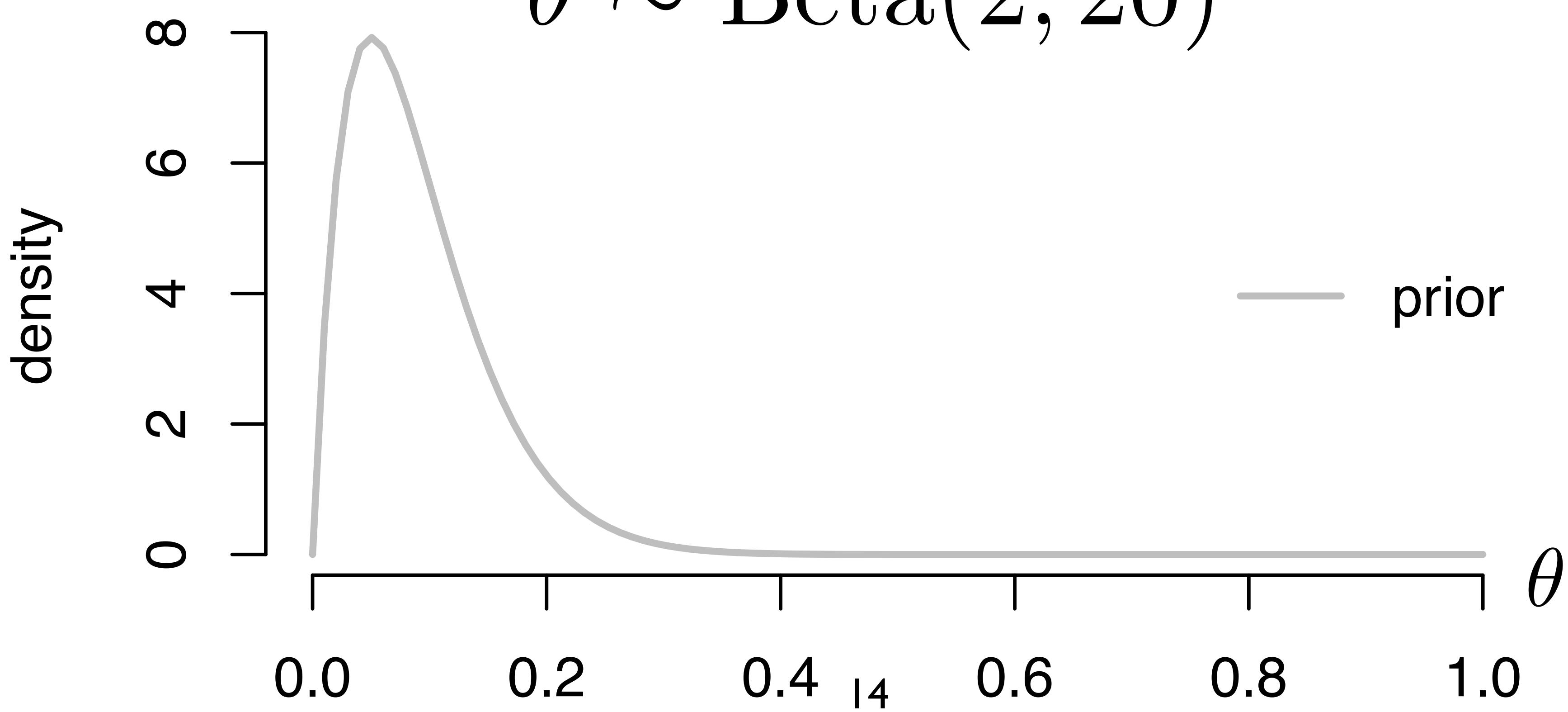


Example: prior

Other studies from various parts of the country indicate that the infection rate ranges from about 0.05 to 0.20, with an average prevalence of 0.1

- Moment matching from a beta distribution (a convenient choice) gives the prior

$$\theta \sim \mathrm{Beta}(2, 20)$$



Example: posterior

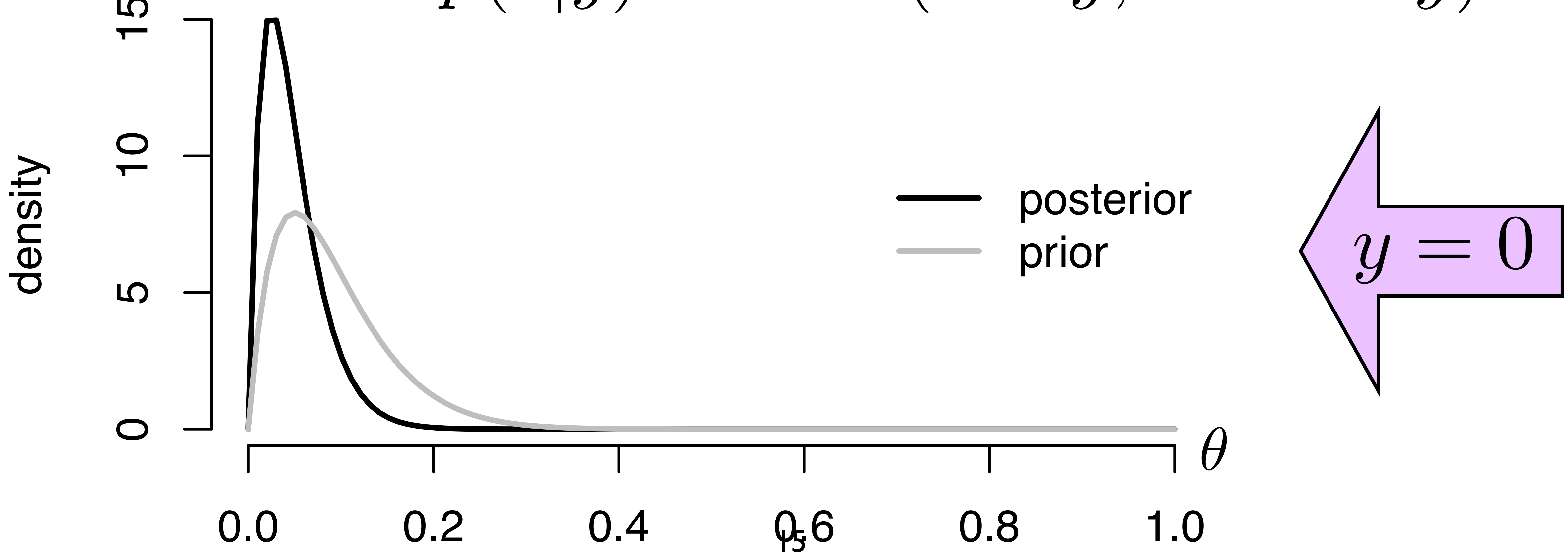
The prior and sampling model combination

$$\theta \sim \text{Beta}(a, b)$$

$$Y|\theta \sim \text{Bin}(n, \theta)$$

and an observed y (the data), leads to the posterior

$$p(\theta|y) = \text{Beta}(a + y, b + n - y)$$



Example: sensitivity analysis

How influential is our prior?

- The posterior expectation is

$$\mathbb{E}\{\theta|Y = y\} = \frac{n}{w+n}\bar{y} + \frac{w}{w+n}\theta_0$$

a weighted average of the sample mean and the prior expectation

$\theta_0 = \frac{a}{a+b}$ ⇨ prior expectation (or guess)

$w = a + b$ ⇨ prior confidence

Example: A non-Bayesian approach

A standard estimate of a population proportion θ is the sample mean $\bar{y} = y/n$, the fraction of infected people in the sample

- If $y = 0$, this gives zero, so reporting the sampling uncertainty is crucial (e.g., for reporting to health officials)

- A popular 95% confidence interval for a population proportion θ is the Wald interval:

$$\bar{y} \pm 1.96\sqrt{\bar{y}(1-\bar{y})/n}$$

which has the correct *asymptotic* coverage (i.e., for large n), but notice $y = 0$ is still problematic!

Example: A non-Bayesian approach, ctd.

People have suggested a variety of alternatives to the Wald interval in hopes of avoiding this type of behavior, e.g., the "adjusted" Wald interval (Agresti and Coull, 1998):

$$\hat{\theta} \pm 1.96\sqrt{\hat{\theta}(1-\hat{\theta})/n}, \text{ where}$$

$$\hat{\theta} = \frac{n}{n+4}\bar{y} + \frac{4}{n+4}\frac{1}{2}$$

Part 1: Fundamentals

Conditional probability

We usually denote $P(A \cap B) \equiv P(A, B)$

$$P(A|B) \equiv \frac{P(A, B)}{P(B)}$$

is the conditional probability of A given B

Law of total probability

Suppose that $\{E_1, \ldots, E_K\}$ is a partition of Ω

- i.e., $E_1, \ldots, E_K$ disjoint and $\bigcup_{i=1}^{K} E_i = \Omega$

then

$$P(A) = \sum_{i=1}^{K} P(A, E_i) \qquad \text{(LTP)}$$

(by conditional probability)

$$= \sum_{i=1}^{K} P(A|E_i)P(E_i)$$

Bayes' rule

In its simplest form:

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

Given a partition $\{E_1, \ldots, E_K\}$ of Ω :

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

(LTP) $$= \frac{P(A|E_i)P(E_i)}{\sum_{i=1}^{K} P(A|E_i)P(E_i)}$$

Independence

Two events A and B are independent if

$$P(A, B) = P(A)P(B)$$

Independence implies that $P(A|B) = P(A)$

Two events A and B are conditionally independent given C if

$$P(A, B|C) = P(A|C)P(B|C)$$

Likewise, conditional independence implies that

$$P(A|B, C) = P(A|C)$$

Probability mass function

The event that the outcome Y of our survey has the value y is expressed as $\{Y = y\}$

For each $y \in \mathcal{Y} \equiv \Omega_y$ we use the shorthand notation

$$P(Y = y) = p(y)$$

- this is the probability mass function or PMF
- the PMF has the following properties

$$0 \leq p(y) \leq 1 \quad \text{for all } y \in \mathcal{Y}$$

$$P(y \in A) = \sum_{y \in A} p(y) \Rightarrow P(y \in \mathcal{Y}) = \sum_{y \in \mathcal{Y}} p(y) = 1$$

Example: Binomial Distribution

Let $\mathcal{Y} = \{0, 1, 2, \ldots, n\}$ for some positive integer n

- The uncertain quantity $Y \in \mathcal{Y}$ has a binomial distribution with probability θ if

$$P(Y = y|\theta) = p(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

$$= \frac{n!}{y!(n-y)!} \theta^y (1-\theta)^{n-y}$$

- To evaluate the mass in R we use `dbinom(y,n,theta)`

Uncountable Sample Spaces

If $\mathcal{Y}$ is uncountable then we cannot work with probabilities of discrete events

- the event $\{Y = 5\}$ say, for $\mathcal{Y} \subseteq \mathbb{R}$ cannot have any probability "mass" since 5 is a singleton in $\mathbb{R}$
- Likewise $P(Y \leq 5) = \sum_{y \leq 5} p(y)$ does not make sense

So we must work directly with the cumulative distribution function (CDF) $F(y) = P(Y \leq y)$ instead

$F(\infty) = 1, F(-\infty) = 0$ & $F(b) \leq F(a)$ if $b < a$

Giving:

$$P(Y > a) = 1 - F(a)$$

$$P(a < Y \leq b) = F(b) - F(a)$$

Continuous RVs & PDFs

If F is continuous, then Y is a continuous RV

- For every continuous CDF F there exists a positive function $f(y)$ such that

$$F(a) = \int_{-\infty}^{a} f(y)\, dy \qquad \text{i.e.,} \qquad F'(y) = f(y)$$

This function is called the probability density function (PDF) of Y

Probability density

The properties of the PDF are similar to the PMF

E.g.,

1. $0 \leq f(y), \text{ for all } y \in \mathcal{Y}$
2. $P(y \in A) = \int_{y \in A} f(y)\, dy \Rightarrow \int_{y \in \mathcal{Y}} f(y)\, dy = 1$

In fact, we will often write $p(y) \equiv f(y)$. However,

- Unlike a PMF, the PDF may be greater than one, and
- $p(y)$ is not "the probability that $Y = y$"

Still, if $p(y_1) > p(y_2)$ we will sometimes informally say that y_1 "has higher probability" [density] than y_2

Example: Normal Distribution

Suppose that we are sampling from a population on $\mathcal{Y} = (-\infty, \infty)$, and we know that the mean of the population is μ and the variance is σ^2

- Then the distribution that has the most "spread", or is the most "diffuse" is the normal distribution: $\mathcal{N}(\mu, \sigma^2)$

$$P(Y < y|\mu, \sigma^2) = F(y) = \int_{-\infty}^{y} \underbrace{\frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\}}_{f(y)} dy$$

(CDF)

(PDF) $f(y)$

- To evaluate the CDF & PDF in R we use `pnorm(y,mu,sigma)` & `dnorm(y,mu,sigma)`

Expectation

The mean or expectation of an unknown quantity Y is

$$\mathbb{E}\{Y\} = \sum_{y \in \mathcal{Y}} y p(y) \qquad \text{if } Y \text{ is discrete}$$

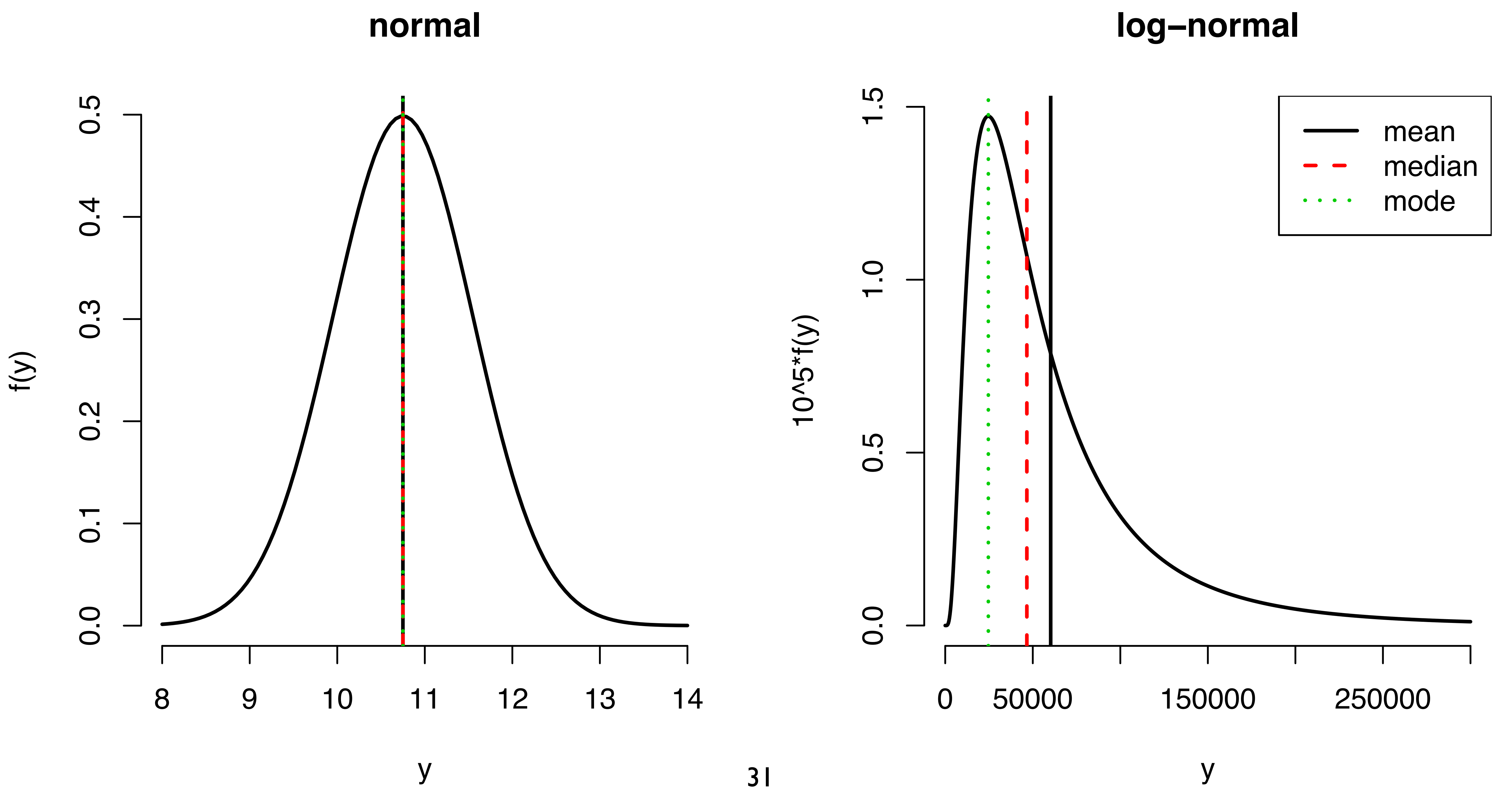
$$\mathbb{E}\{Y\} = \int_{y \in \mathcal{Y}} y f(y)\, dy \qquad \text{if } Y \text{ is continuous}$$

The mean is the center of mass of the distribution. However, it is not in general equal to either of

- the mode: "the most probable value of Y", or
- the median: "the value of Y in the middle of the distribution

Expectation

For skewed distributions (e.g., for income), the mean can be far from a typical sample value



Variance

The most popular measure of how spread out the distribution is is the variance

$$\mathrm{Var}[Y] = \mathbb{E}\{(Y - \mathbb{E}\{Y\})^2\}$$
$$= \mathbb{E}\{Y^2\} - \mathbb{E}\{Y\}^2$$

- it is the average squared distance that a sample value Y will be from the population mean $\mathbb{E}\{Y\}$
- the standard deviation is the square root of the variance
  - so it is on the same scale of Y

Quantiles

Alternative measures of the spread of a distribution are based on quantiles

- for a continuous, strictly increasing CDF F, the α-quantile is the value y_α such that $F(y_\alpha) = \alpha$
- The interquartile range of a distribution is the interval $(y_{0.25}, y_{0.75})$ which contains 50% of the mass of the distribution
- Similarly, the interval $(y_{0.025}, y_{0.975})$ contains 95% of the mass of the distribution

Joint distributions

Let Y_1, Y_2 be two random variables taking values in $\mathcal{Y}_1, \mathcal{Y}_2$

Joint beliefs about Y_1 and Y_2 can be represented with probabilities. E.g.,

- for subsets $A \subset \mathcal{Y}_1$ and $B \subset \mathcal{Y}_2$,

  $$P(\{Y_1 \in A\}, \{Y_2 \in B\})$$

  represents our belief that Y_1 is in A and Y_2 is in B

Marginals & Conditionals

As in the discrete case,

- The marginal density of Y_1 can be computed from the joint

$$f_{Y_1}(y_1) \overset{\text{(LTP)}}{=} \int_{-\infty}^{\infty} f_{Y_1,Y_2}(y_1, y_2)\, dy_2$$

- The conditional density of Y_2 given $\{Y_1 = y_1\}$ can be computed from the joint and marginal densities

$$f_{Y_2|Y_1}(y_2|y_1) = \frac{f_{Y_1,Y_2}(y_1, y_2)}{f_{Y_1}(y_1)} \qquad \text{(cond. prob.)}$$

Joint mean and covariance

For a vector RV $Y = (Y_1, \ldots, Y_n)^\top$, the expression for the mean is still

$$\mathbb{E}\{Y\} = \int y p(y)\, dy$$

so that $\mathbb{E}\{Y\} = (\mathbb{E}\{Y_1\}, \ldots, \mathbb{E}\{Y_n\})^\top$

The covariance matrix is defined as

$$\mathrm{Cov}\{Y\} = \int (y - \mathbb{E}\{Y\})(y - \mathbb{E}\{Y\})^\top p(y)\, dy$$

The diagonal of $\mathrm{Cov}\{Y\}$ is $(\mathrm{Var}[Y_1], \ldots, \mathrm{Var}[Y_n])$

Bayes' rule and estimation

Let:

$\theta =$ proportion of people in a large population who have a certain characteristic

$Y =$ number of people in a small random sample from the population who have the characteristic

Then we might treat θ as continuous and Y as discrete

Bayesian estimation of θ derives from the calculation $p(\theta|y)$, where y is the observed value of Y

This calculation first requires that we have a joint "density" $p(y,\theta)$ representing our beliefs about θ and the survey outcome Y

Prior and sampling model

Often it is natural to construct this joint density from

- $p(\theta)$, beliefs about θ
- $p(y|\theta)$, beliefs about Y for each value of θ

Having observed $\{Y = y\}$, we need to compute our updated beliefs about

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)}$$

This conditional "density" is called the posterior density of θ

A ratio of posteriors

Suppose θ_a and θ_b are two possible numerical values of the true value of θ

The posterior probability (density) of θ_a relative to θ_b, conditional on $\{Y = y\}$, is

$$\frac{p(\theta_a|y)}{p(\theta_b|y)} = \frac{p(y|\theta_a)p(\theta_a)}{p(y|\theta_b)p(\theta_b)}$$

This means that to evaluate the relative posterior probabilities of θ_a and θ_b, we do not need to compute $p(y)$

Independent & Identical

Under independence, the joint density is given by

$$p(y_1, \ldots, y_n|\theta) = \prod_{i=1}^{N} p_{Y_i}(y_i|\theta)$$

If $Y_1, \ldots, Y_n$ are all generated from a common process

- then the marginal densities are all the same

$$p(y_1, \ldots, y_n|\theta) = \prod_{i=1}^{N} p(y_i|\theta)$$

In this case we say that $Y_1, \ldots, Y_n$ are conditionally independent and identically distributed (IID)

the shorthand is: $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} p(y|\theta)$

Likelihood

Suppose that Y has sampling model $p(y|\theta)$ for $y \in \mathcal{Y} \subseteq \mathbb{R}^n$ and $\theta \in \Theta \subseteq \mathbb{R}^d$

The likelihood function is a function of θ for each fixed y given by

$$L(\theta) \equiv L(\theta; y) = p(y|\theta)$$

and simplifications often arise under IID assumptions

Classical stats is concerned with the log-likelihood

$$\ell(\theta) \equiv \ell(\theta; y) = \log p(y|\theta)$$