STAT 206B
Chapter 1: Introduction

Winter 2022

† Statistical Problems: CR 1.1

- The main purpose of statistical theory is to derive from observations of a random phenomenon an *inference* about the probability distribution underlying this phenomenon.
- A random phenomenon is directed by a parameter θ.
- Observations $x_1, \ldots, x_n$ are generated from the random phenomenon.
- Deduce an inference on θ from theses observations.

- CR Ex 1.1.5: Consider a dataset that consists of the monthly unemployment rate and the monthly number of accidents (in thousands) in Michigan from 1978 to 1897. Lenk (1999) argues in favor of a connection between these two variates, in that higher unemployment rates lead to less traffic on the roads and thus fewer accidents.

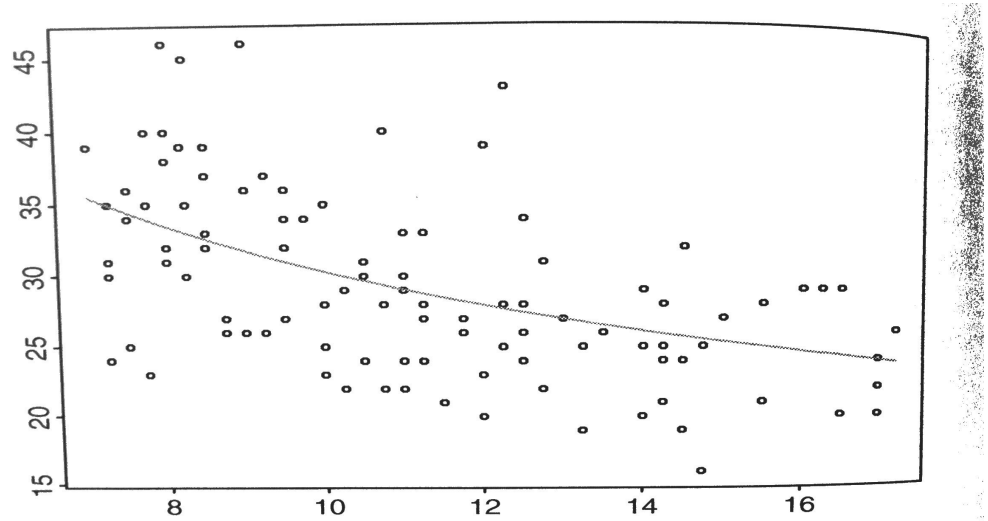


Figure 1.1.1. *Plot of monthly unemployment rate versus number of accidents (in thousands) in Michigan, from 1978 to 1987.* (Source: *Lenk (1999).*)

- Statistical inference is based on *probabilistic modeling* of the observed random phenomenon.

  $\star\star$ We need to set up a *probability model* that characterizes the behavior of the observations *conditional* on θ.

  $\star\star$ The model should be consistent with knowledge about the underlying scientific problem and the data collection process and it should provide *an adequate representation of the observed phenomenon.*

  $\star\star$ Probabilistic modeling is a necessarily *reductive* formalization step at the same time.

  "all models are wrong, but some are useful" (G. Box and N. Draper, 1987).

- CR Ex 1.1.5 (contd): Let N_i and ρ_i denote the number of accidents and the corresponding unemployment rate in month i. We may assume a parametric structure in the dependence between unemployment rates and number of accidents using the Poisson regression model,

$$N_i \mid \mu_i \overset{indep}{\sim} \text{Poi}(\mu_i), \text{where } \log(\mu_i) = \beta_0 + \beta_1 \log(\rho_i).$$

$\star\star$ The fit of the model and the implications of the resulting inference need to be evaluated; Does the model fit the data? are the substantive conclusions reasonable and how sensitive are the results to the modeling assumption?

- Statistical inference is concerned with drawing conclusions, from *quantities that we observe (numerical data)*, about *quantities that are not observed*.

  $\star\star$ *Quantities that are not observed?* e.g.

  1) Parameters that govern the hypothetical process leading to the observed data.

  2) Potentially observable quantities such as future observations of a process.

† Statistical Analysis

- Suppose $x_1, \ldots, x_n$'s are random variables (observable or only hypothetically observable), $x_i \in \mathcal{X}$.
- Identify parameters $\boldsymbol{\theta}$ describing the conditions under which the random variables are generated (unknown).

  ⋆⋆ Let Θ be the parameter space, i.e., the set of all possible values of $\boldsymbol{\theta}$.
- Specify a joint probability distribution for the observable random variables (assume a parametric function for this course);

$$f(\boldsymbol{x} \mid \boldsymbol{\theta}),$$

  where $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{\theta}$.

† Statistical Analysis – contd

- Say $x_1, \ldots, x_n$ are conditionally independent given $\boldsymbol{\theta}$;

$$f(x_1, \ldots, x_n \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i \mid \theta_i),$$

where θ_i is the parameter for the distribution of x_i.

⋆⋆ *implication:* x_j gives no additional information about x_i beyond that in knowing in $\boldsymbol{\theta}$.

- Further assume that θ_i are all equal, i.e., $\theta_1 = \ldots = \theta_n = \theta$

$\Leftrightarrow$ x_i's are conditionally independent and identically distributed (iid) from a common distribution;

$$f(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta).$$

† PF §1.2.1 Example: Suppose that we are interested in the prevalence of an infectious disease in a small city. The higher prevalence, the more public health precautions we would recommend be put into place. A small random sample of 20 individuals from the city were checked for infection. Write a joint sampling distribution.

† Statistical Problems

- **Definition 1.1.7** A parametric statistical model consists of the observation of a random variable x, distributed according to $f(x \mid \theta)$, where only the parameter is unknown and belongs to a vector space Θ of finite dimension.

- Use statistical methods to deduce from these observations an inference about θ.

  ⋆⋆ Estimation e.g. What is the value of θ?

  ⋆⋆ Testing e.g. Is θ_1 greater than θ_2?

  ⋆⋆ Prediction e.g. The distribution of a future observation y depending on x, $p(y \mid x)$

† Bayesian Paradigm (CR 1.2)

- **Definition 1.2.1** A Bayesian statistical model is made of a parametric statistical model, $f(x \mid \theta)$, and a prior distribution on the parameter $\pi(\theta)$.

- *Likelihood:* Thought of as a function of θ, refer to the joint sampling distribution as the likelihood function,

$$\ell(\theta \mid \boldsymbol{x}) = f(\boldsymbol{x} \mid \theta),$$

  where θ is unknown and depends on the observed data.

- *Priors:* The uncertainty on the parameter(s), θ is modeled through a probability distribution on the parameter space Θ, $\pi(\theta)$ or $\pi(\theta \mid \tau)$ where τ is called a hyperparameter.

† Bayesian Paradigm – contd

- We will later talk about how to construct a prior distribution (CR Chapter 3).
- The parameter θ is supported by the parameter space Θ. θ is an index to a frequentist. In Bayesian modeling, the unknown θ is treated as a **random variable**.
- Ex 1.2.2 (Bayes (1764)) A billiard ball W is rolled on a line of length one, with a uniform probability of stopping anywhere. It stops at p. A second ball O is then rolled n times under the same assumptions and X denotes the number of times the ball O stopped on the left of W.

† Bayesian Paradigm–contd

- In particular settings, parameters can be viewed as random. But, not *always.*

- unknown parameter $\rightarrow$ random parameter? *(CR p10)*

  "...as for instance, quantum physics, the parameter to be estimated cannot be perceived as resulting from a random experiment in most cases. e.g. physical quantities like the speed of light, c.... the limited accuracy of the measurement instruments implies that the true value of c will never be known, and thus that is justified to consider c as being uniformly distributed on $[c_0 - \epsilon, c_0 + \epsilon]$.... "

- So, we defend...

  $\star\star$ Using a probability distribution is still a convenient way to summarize the available information (or even lack of information) about θ.

† Prior and Posterior Distributions (CR 1.4)

- Bayesian analysis is performed by combining the prior information (through the prior distribution $\pi(\theta)$) and the sample information (through the sampling distribution $f(x \mid \theta)$) into the posterior distribution.
- All decisions and inferences are made from the posterior distribution of θ given $\boldsymbol{x}$.
- The **joint distribution** $\psi(x, \theta) = \pi(\theta) f(x \mid \theta)$
- The **marginal distribution**

$$m(x) = \int_{\Theta} \psi(x, \theta) d\theta = \int_{\Theta} \pi(\theta) f(x \mid \theta) d\theta.$$

† Prior and Posterior Distributions – contd

- The inference is based on the distribution of θ conditional on x, $\pi(\theta \mid x)$ – **posterior distribution**. For $m(x) > 0$

$$\pi(\theta \mid x) = \frac{\psi(x, \theta)}{m(x)} = \frac{\pi(\theta) f(x \mid \theta)}{m(x)} \propto \pi(\theta) f(x \mid \theta)$$

- The posterior distribution combines the prior beliefs about θ with the information about θ contained in the sample x so the posterior distribution reflects the updated beliefs about θ after observing x.

  $\Rightarrow$ Give a composite picture of the final beliefs about θ.

  $\Rightarrow$ All decisions and inferences are made from $\pi(\theta \mid x)$.

† Prior and Posterior Distributions – contd

- Suppose $Y \sim g(y \mid \theta, x)$ is to be observed. The **posterior predictive distribution** of Y, given observed $X = x$, is

$$g(y \mid x) = \int_{\Theta} g(y \mid \theta, x)\pi(\theta \mid x)d\theta.$$

- If we assume conditional independence of y from x (that is, $g(y \mid \theta, x) = g(y \mid \theta)$),

$$g(y \mid x) = \int_{\Theta} g(y \mid \theta)\pi(\theta \mid x)d\theta$$

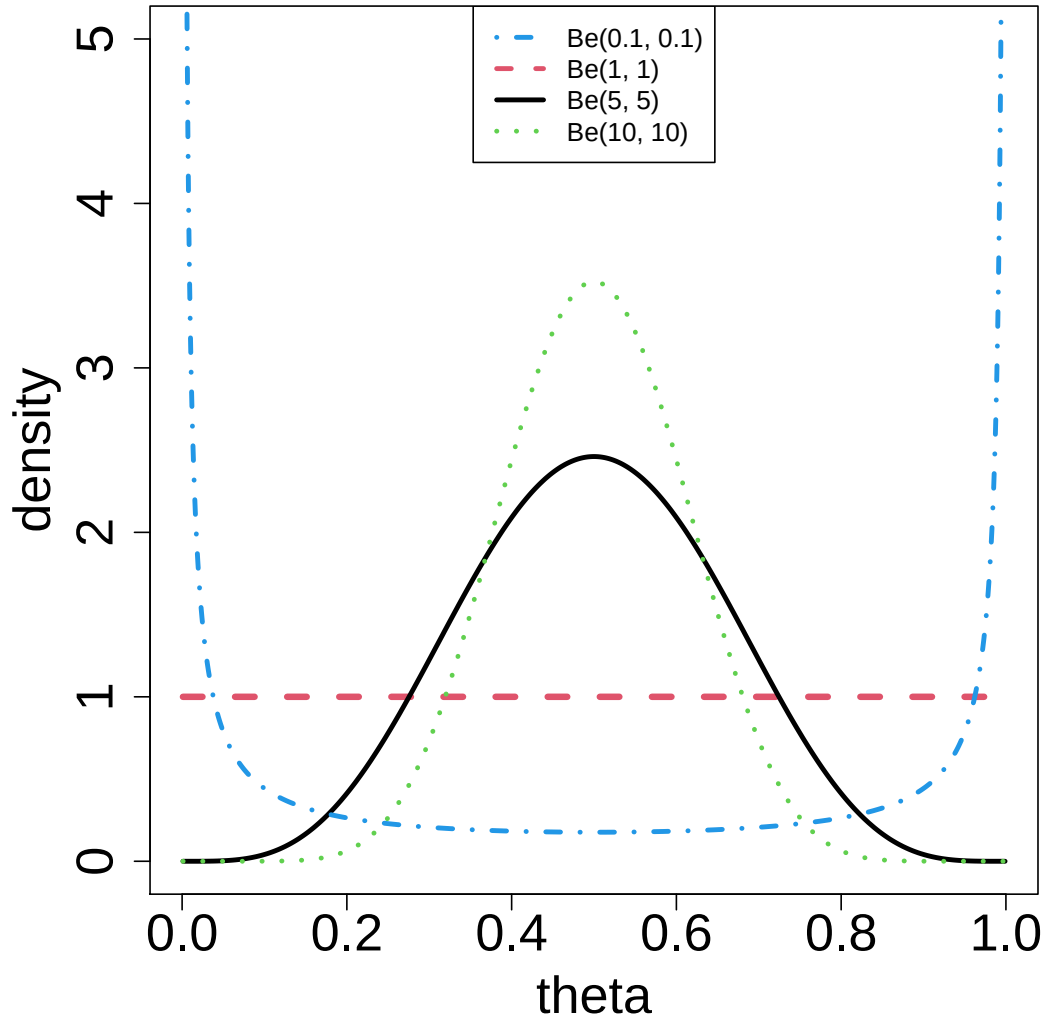
** *Note*: $m(y) = \int_{\Theta} g(y \mid \theta)\pi(\theta)d\theta$ is called the prior predictive distribution.

† PF §1.2.1 Example (contd): Suppose that the sampling model for x_i is a Bernoulli, i.e., $x_i \mid \theta \overset{iid}{\sim} \text{Ber}(\theta)$, and the prior is $\text{Be}(\alpha, \beta)$, where the hyperparameters α and β are known,

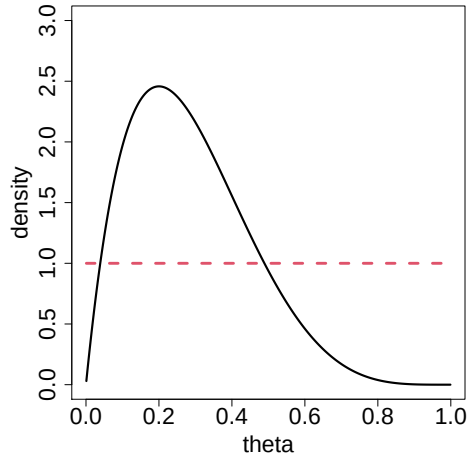
$$\pi(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}, \quad 0 < \theta < 1.$$

Find the joint, marginal, posterior, and predictive distributions.

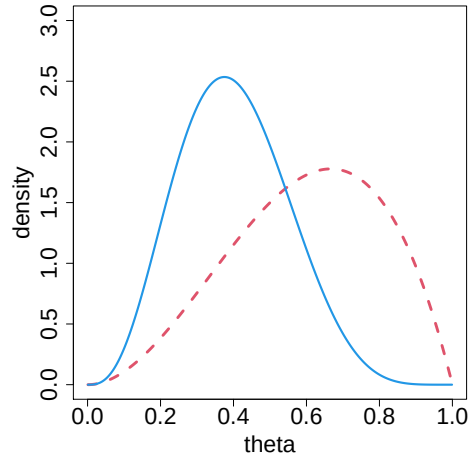
† PF §1.2.1 Example (contd): Examples of the beta density



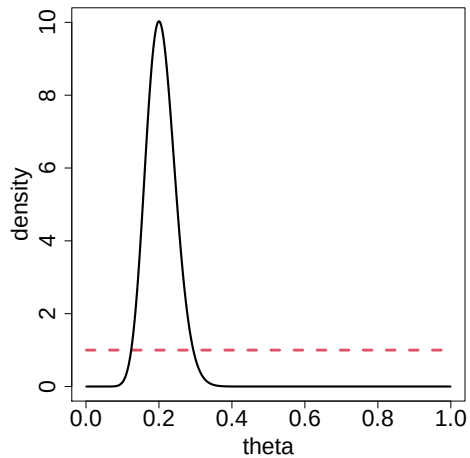
† PF §1.2.1 Example (contd):



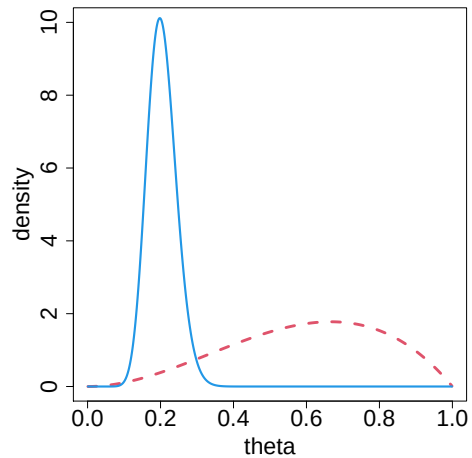
(a) Be$(1, 1)$, $n = 5, x = 1$



(b) Be$(3, 2)$, $n = 5, x = 1$

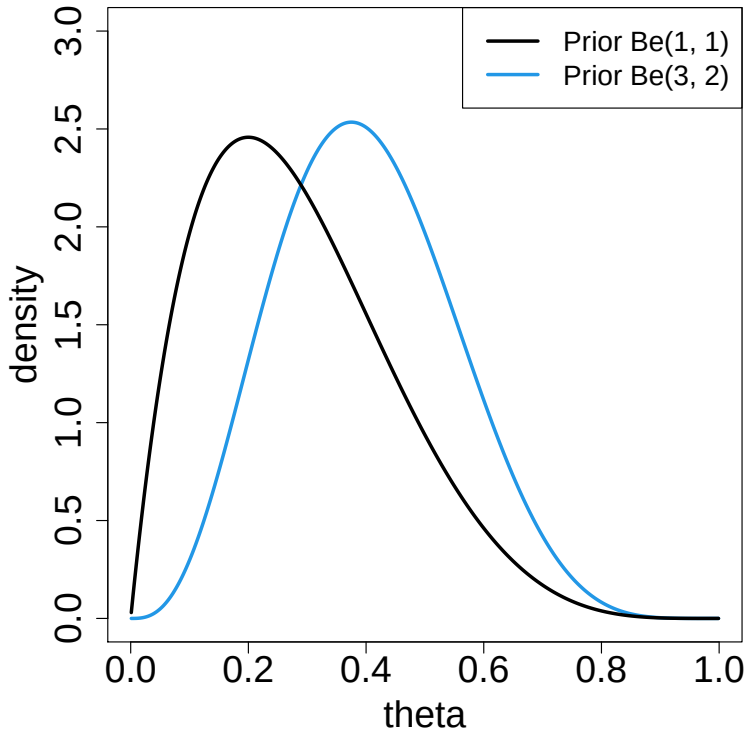


(c) Be$(1, 1)$, $n = 100, x = 20$

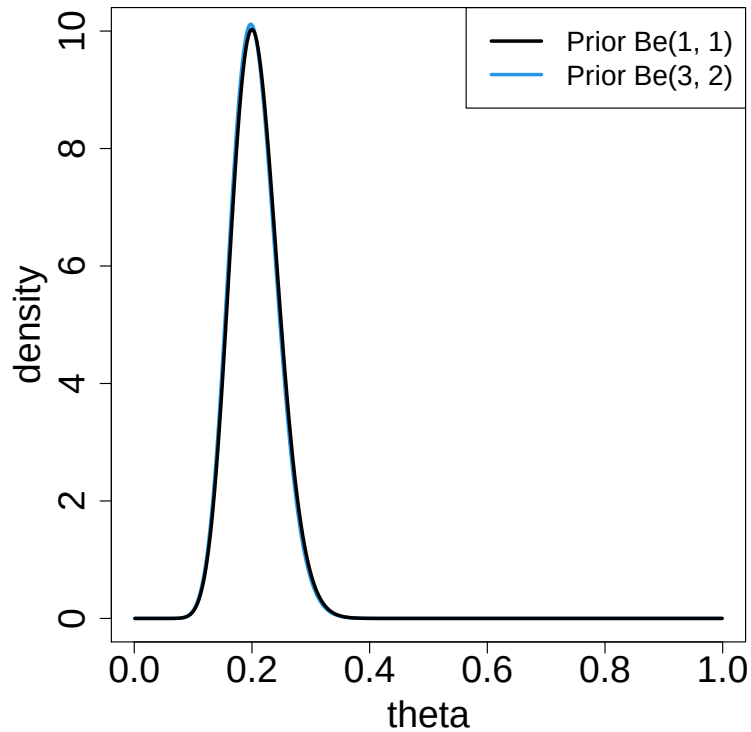


(d) Be$(3, 2)$, $n = 100, x = 20$

† PF §1.2.1 Example (contd):



(a) $n = 5, x = 1$



(b) $n = 100, x = 20$

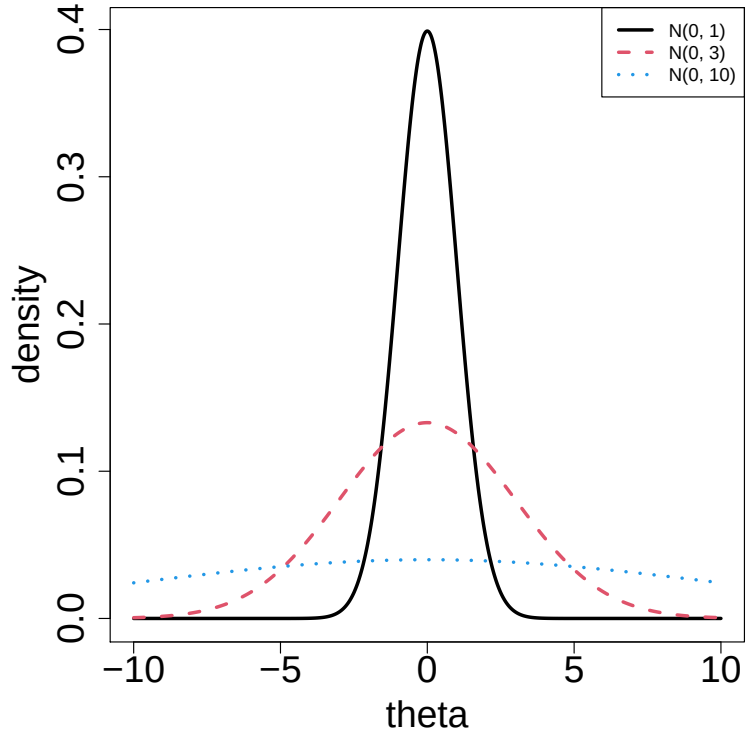
⊕ For more, read Hoff §3.1 The binomial model.

♣ Example 2 (JB Example 1 p127): Assume that observations x_i's are normally distributed with mean θ and known variance σ^2. The parameter of interest, θ also has normal distribution with parameters μ and τ^2. Find the posterior distribution of θ given $\boldsymbol{x}$. Also, find the posterior predictive distribution of a future observation y assuming conditional independence between y and $\boldsymbol{x}$ given θ.

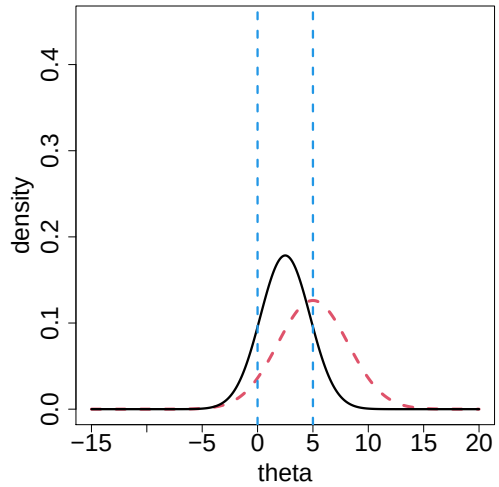
** *Note*: This example is important because the normal likelihood and normal prior combination is very common.

** Also, read Hoff §5.1-5.2.

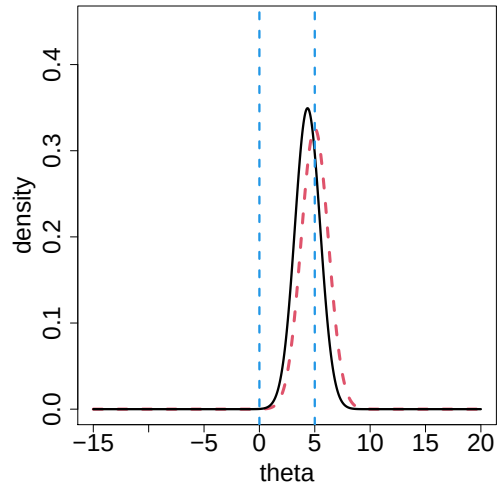
♣ Example 2(contd) Examples of the normal density



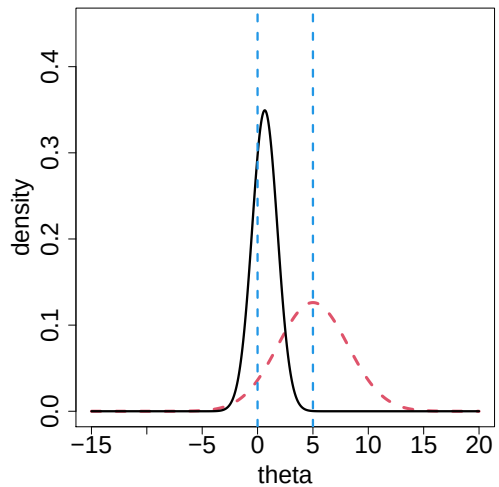
♣ Example 2(contd): Suppose $\bar{x} = 0$ with $n = 1$ and $\mu = 5$.



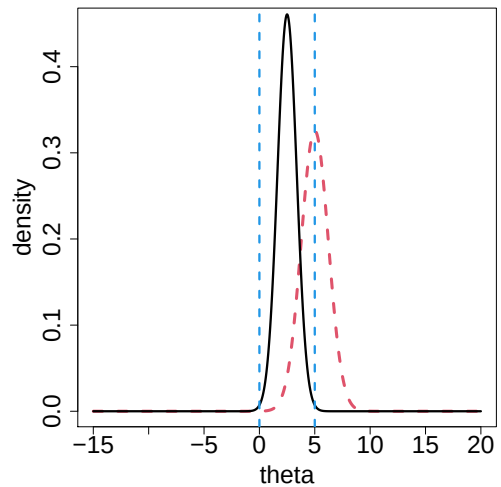
(a) $\sigma^2 = 10$ & $\tau^2 = 10$



(b) $\sigma^2 = 10$ & $\tau^2 = 1.5$

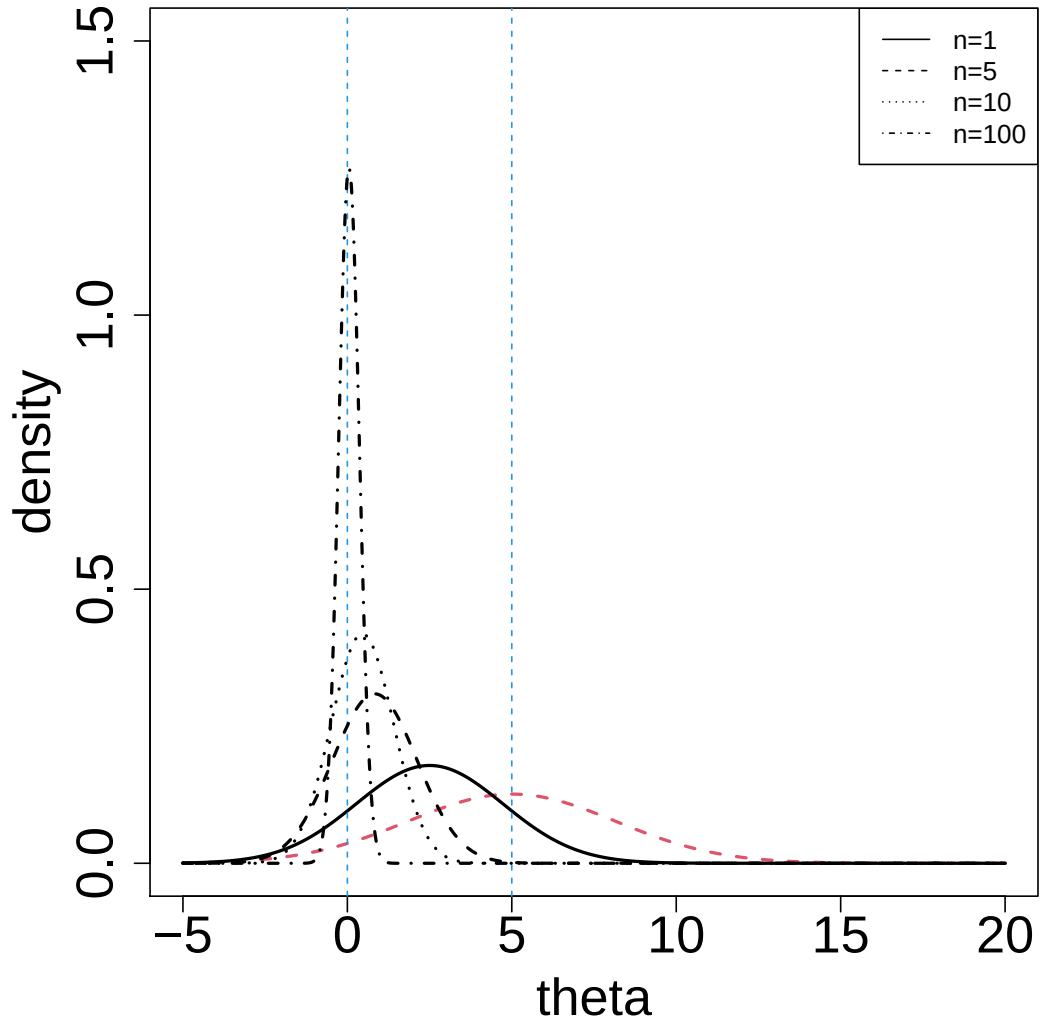


(c) $\sigma^2 = 1.5$ & $\tau^2 = 10$



(d) $\sigma^2 = 1.5$ & $\tau^2 = 1.5$

♣ Example 2(contd): Suppose $\bar{x} = 0$ and $\mu = 5$ with $\sigma^2 = \tau^2 = 10$ and vary n.



† Conjugate Priors (CR Sec 3.3)

- **Def 3.3.1:** A family $\mathcal{F}$ of probability distributions on Θ is said to be *conjugate* (or closed under sampling) for a likelihood function $f(x \mid \theta)$ if, for every $\pi \in \mathcal{F}$, the posterior distribution $\pi(\theta \mid x)$ also belong to $\mathcal{F}$.

e.g1 A beta prior distribution and a binomial sampling model lead to a beta posterior distribution. We say "The class of beta priors is conjugate for the binomial sampling distribution."

e.g2 Similarly, normal priors are a conjugate family for normal sampling distributions.

- If $\mathcal{F}$ is a conjugate family,

  obtaining the posterior $\Leftrightarrow$ updating the corresponding parameters

† Examples: Conjugate Priors

e.g1 Assume $x \mid \theta \sim \mathsf{N}(\theta, \sigma^2)$ and $\theta \sim \mathsf{N}(\mu, \tau^2)$.

$$\Rightarrow \quad \theta \mid x \sim \mathsf{N}\left(\left(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}\left(\frac{x}{\sigma^2}+\frac{\mu}{\tau^2}\right), \left(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}\right).$$

⋆⋆ Normal priors are a conjugate family for normal sampling distributions.

e.g2 Assume $X \mid \theta \sim \mathsf{Bin}(n, \theta)$ and $\theta \sim \mathsf{Be}(\alpha, \beta)$.

$$\Rightarrow \theta \mid x \sim \mathsf{Be}(\alpha + x, \beta + n - x).$$

⋆⋆ Beta priors are a conjugate family for binomial sampling distributions.

- If $\mathcal{F}$ is a conjugate family,

  obtaining the posterior $\Leftrightarrow$ updating the corresponding parameters

  i.e, data does not modify the whole structure of the distribution of θ, but simply updates its parameters.

- A classical parametric approach to build up prior distributions based on limited prior input

- A conjugate family can frequently be determined by examining the likelihood functions $\ell(\theta \mid x)$ and choosing, as a conjugate family, the class of distributions with the same functional form as these likelihood functions.

  $\Rightarrow$ often called natural conjugate priors.

† Exponential Families

- A family of pdfs or pmfs is called an *exponential family* if it can be expressed as

$$f(x \mid \boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left(R(\boldsymbol{\theta})T(x)\right).$$

⋆⋆ $h(x) \geq 0$

⋆⋆ $T(x) = [t_1(x), \ldots, t_k(x)]$ are real-valued functions of the observations x (cannot depend on $\boldsymbol{\theta}$)

∗∗ natural sufficient statistic.

∗∗ all the information about θ in the sample is summarized in $T(x)$.

⋆⋆ $c(\boldsymbol{\theta}) \geq 0$

⋆⋆ $R(\boldsymbol{\theta}) = (w_1(\boldsymbol{\theta}), \ldots, w_k(\boldsymbol{\theta}))$ are real-valued functions of the possibly vector-valued parameter $\boldsymbol{\theta}$ (cannot depend on x)

† Exponential Families (contd)

- The sufficient statistic and the parameter vectors are usually of equal length.
- These include the continuous families- normal, gamma, and beta, and the discrete families- binomial, Poisson, and negative binomial.

  ⋆⋆ consider a change of variables $\boldsymbol{z} = T(x)$ and a reparameterization $\boldsymbol{\eta} = \boldsymbol{\theta}$ (natural parameter) and rewrite

$$f(z \mid \boldsymbol{\eta}) = C^{\star}(\boldsymbol{\eta}) h^{\star}(z) \exp(\boldsymbol{\eta} \boldsymbol{z})$$

  $\Rightarrow$ the canonical form

- Show a Poisson distribution, $X \sim \mathrm{Poi}(\theta)$ with $\theta > 0$ is an exponential family.

- Show a normal distribution, $X \sim \mathrm{N}(\mu, \sigma^2)$ with $\boldsymbol{\theta} = (\mu, \sigma)$, where $\mu \in \mathbb{R}$ and $\sigma > 0$, is an exponential family.

- **Example 3.3.8** Find a conjugate prior for the Poisson $\mathrm{Poi}(\theta)$ family with $0 < \theta$. Find its posterior distribution given x.

♣ Table 3.3.1 Natural conjugate priors for some common exponential families

Table 3.3.1. *Natural conjugate priors for some common exponential families*

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Normal $\mathcal{N}(\theta, \sigma^2)$	Normal $\mathcal{N}(\mu, \tau^2)$	$\mathcal{N}(\varrho(\sigma^2\mu + \tau^2 x), \varrho\sigma^2\tau^2)$ $\varrho^{-1} = \sigma^2 + \tau^2$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	$\mathcal{G}(\alpha + x, \beta + 1)$
Gamma $\mathcal{G}(\nu, \theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	$\mathcal{G}(\alpha + \nu, \beta + x)$
Binomial $\mathcal{B}(n, \theta)$	Beta $\mathcal{B}e(\alpha, \beta)$	$\mathcal{B}e(\alpha + x, \beta + n - x)$
Negative Binomial $\mathcal{N}eg(m, \theta)$	Beta $\mathcal{B}e(\alpha, \beta)$	$\mathcal{B}e(\alpha + m, \beta + x)$
Multinomial $\mathcal{M}_k(\theta_1, \ldots, \theta_k)$	Dirichlet $\mathcal{D}(\alpha_1, \ldots, \alpha_k)$	$\mathcal{D}(\alpha_1 + x_1, \ldots, \alpha_k + x_k)$
Normal $\mathcal{N}(\mu, 1/\theta)$	Gamma $\mathcal{G}a(\alpha, \beta)$	$\mathcal{G}(\alpha + 0.5, \beta + (\mu - x)^2/2)$

† Improper Prior Distributions (CR 1.4)

- Recall that the parameter is a random variable following a probability distribution $\pi(\theta)$.
- We say the prior distribution is *improper* (or *generalized*) if

$$\int_{\Theta} \pi(\theta) d\theta = +\infty.$$

- Bayesian methods apply as long as the posterior distribution is defined.
- The posterior exists when the pseudo marginal distribution $\int_{\Theta} \pi(\theta) f(x \mid \theta) d\theta$ is well defined.

♣ Example 3: Assume that an observation, x is normally distributed with mean θ and known variance σ^2. The parameter of interest, θ has an improper prior distribution, $\pi(\theta) = c$. Check it produces a proper posterior distribution. If so, find the posterior distribution.

† Two fundamental principles for the Bayesian paradigm

- Sufficiency principle
- Likelihood principle

† Sufficient Statistics

- **Def 5.2.1 (Casella & Berger)** Let $x_1, \ldots, x_n$ be a random sample of size n from a population and let $T(x_1, \ldots, x_n)$ be a real-valued or vector-valued function whose domain includes the sample space of $(x_1, \ldots, x_n)$. Then the random variable or random vector $T(x_1, \ldots, x_n)$ is called a *statistic*. The probability distribution of $T(x_1, \ldots, x_n)$ is called the *sampling distribution* of T.

  e.g. If an independent sample $x_1, \ldots, x_n$ is taken, the sample mean $\bar{x} = \sum_{i=1}^{n} x_i/n$, the sample variance $s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)$ and the sample standard deviation $s = \sqrt{s^2}$ are statistics that are often used and provide good summaries of the sample.

- **Def 1.3.1** When $x \sim f(x \mid \theta)$, a function T of x (also called a statistic) is said to be *sufficient* if the distribution of x conditional upon $T(x)$ does not depend on θ.

- *How to show that a certain statistic $T(x)$ is or is not a sufficient statistic?* Use the **Fisher–Neyman factorization lemma**.

  Under some measure theoretic regularity conditions, the likelihood can be represented as

$$f(x \mid \theta) = g(T(x) \mid \theta)h(x \mid T(x))$$

  $\Rightarrow$ $T(x)$: a function of data which summarizes all the available *sample* information concerning θ

  $\Rightarrow$ Any additional information in the sample, besides the value of the sufficient statistic, does not contain any more information about θ.

- **Example 1.3.2** Consider $x_1, \ldots, x_n$ independent observations from a normal distribution $\mathrm{N}(\mu, \sigma^2)$ where μ and σ^2 are unknown.
  - By the factorization theorem, the pair $T(x) = (\bar{x}, s^2)$ where $\bar{x} = \sum_{i=1}^{n} x_i/n$ and $s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$ is a sufficient statistic for the parameter (μ, σ)
  - Let's make it simpler. Fix $\sigma^2 = 1$ and suppose μ is only unknown. Is $\bar{x}$ sufficient?

† Sufficiency Principle

- **Sufficiency Principle** Two observations x and y factorizing through the same value of a sufficient statistic T, that is, such that $T(x) = T(y)$, must lead to the same inference.
- If principle is adopted, all inference about θ should depend on sufficient statistics since $\ell(\theta) \propto g(T(x), \theta)$.
- Sometimes criticized since it assumes that the statistical model is the one underlying the data generation.

♣ Example 2 (contd): Assume that an observation, $x_1, \ldots, x_n$ are iid from $\mathrm{N}(\theta, \sigma^2)$, where μ and σ^2 are unknown. Consider

$$\pi(\theta, \sigma^2) = \pi_1(\theta \mid \sigma^2)\pi_2(\sigma^2),$$

where π_1 is a normal distribution $\mathrm{N}(\mu, \sigma^2/n_0)$ and π_2 is a inverse gamma distribution $\mathrm{IG}(\nu/2, s_0^2/2)$.

- Find the joint posterior distribution.
- Find the posterior distributions $\pi(\theta \mid \bar{x}, s^2, \sigma^2)$ and $\pi(\sigma^2 \mid \bar{x}, s^2)$.
- Find the marginal posterior distribution of θ, $\pi(\theta \mid \bar{x}, s^2)$.

** Also, read Hoff §5.3, BDA §3.3, and Robert §4.4.1-4.4.2.

♣ Example 2 (contd): Assume that an observation, $x_1, \ldots, x_n$ are iid from $N(\theta, \sigma^2)$, where μ and σ^2 are unknown. Consider $\tilde{\pi} = 1/\sigma^2$ (not a probability density, i.e., improper prior).

- Find the joint posterior distribution.

- Find the posterior distributions $\pi(\theta \mid \bar{x}, s^2, \sigma^2)$ and $\pi(\sigma^2 \mid \bar{x}, s^2)$.

- Find the marginal posterior distribution of θ, $\pi(\theta \mid \bar{x}, s^2)$.
  ** Also, read Hoff §5.3, BDA §3.2, and Robert §4.4.1-4.4.2.

† Likelihood Principle

- (Recall the definition of Likelihood) For the observed data, $X = x$, the function $\ell(\theta \mid x) = f(x \mid \theta)$, considered as a function of θ, is called the likelihood function.

  $\star\star$ no guarantee that $\ell(\theta \mid x)$ as a function of θ is a pdf.

  $\star\star$ *Intuitive reason for the name:* given x, the value of θ_1 is more likely to be the true parameter than θ_2 if $\ell(\theta_1 \mid x) > \ell(\theta_2 \mid x)$ (x would be more probable occurrence with θ_1).

- **Likelihood Principle** The information brought by an observation x about θ is entirely contained in the likelihood $\ell(\theta \mid x)$. Moreover, if x_1 and x_2 are two observations depending on the same parameter θ, such that there exists a constant c satisfying

$$\ell_1(\theta \mid x_1) = c\ell_2(\theta \mid x_2)$$

  for every θ, they then bring the **same information about** θ and must lead to identical inference.

- c does not depend on θ.

- **In other words,** In the inference about θ, after x is observed, all relevant experimental information is contained in the likelihood function for the observed x. Furthermore, the likelihood functions contain the **same information about** θ if they are proportional to each other.

† The Likelihood Principle is emphasized in Bayesian statistics!

- *Recall* The Bayesian approach is entirely based on the posterior distribution.

$$\pi(\theta \mid x) = \frac{\ell(\theta \mid x)\pi(\theta)}{\int_{\Theta} \ell(\theta \mid x)\pi(\theta)d\theta}.$$

That is, the posterior depends on data (x) only through $\ell(\theta \mid x)$.

- Thus, the Likelihood Principle is automatically satisfied in a Bayesian setting.

Example 15 (JB, p28)– Testing fairness (1)

Suppose we are interested in testing θ, the unknown probability of heads for a possibly biased coin. Suppose we test $H0 : \theta = 1/2$ vs $Ha : \theta > 1/2$. An experiment is conducted by flipping the coin independently in a series of trials and 9 heads and 3 tails are observed. The information is not sufficient to fully specify the model, $f(x \mid \theta)$. Let's consider classical testing.

- **Scenario 1** The number of flips is pre-determined.

Example 15 (JB, p28)– Testing fairness (2)

- **Scenario 2** The number of tails is predetermined.

Example 15 (JB, p28)– Testing fairness (3)

- How about Bayesian inference?

Example 15 (JB, p28)– Testing fairness (4)

* What does this imply?

  - We did not really need to know anything about the "series of trials". That is, the rules governing when data collection stops are irrelevant to data interpretation.

  - It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience. — Edwards, Lindman, and Savage (1963, 193)

† Few Remarks!

- The correspondence of information from proportional likelihood functions applies *only* when the two likelihood functions are for the *same* parameter.
- The likelihood principle does not say all information about θ is contained in $\ell(\theta)$, just that all *experimental* information is.
- It is of fundamental importance to get the likelihood function right (the likelihood function should be a close approximation or representation of data).
- Also, see Example 1.3.5.
- Optional: read **Example 1.3.6 & Stopping Rule Principle** CR p17 & JB Chapter 7.7

† Maximum Likelihood Estimator (MLE)

- The maximum likelihood estimation approach is a way to implement the likelihood principle.

  $\ell(\hat{\theta}^{\text{MLE}} \mid x)$ is at least as great as $\ell(\theta \mid x)$ for every $\theta \in \Theta$.

- A lot more in STAT 205(B)!

† Conditionality Principle

- **Def** If two experiments on the parameter θ, $\mathcal{E}_1$ and $\mathcal{E}_2$, are available and if one of these two experiments is selected with probability p, the resulting inference on θ should only depend on the selected experiment.

- Any inference should depend only on the outcome observed and not on any other outcome we might have observed.

- This sharply contrasts with a frequentist approach (which cares long-run behavior over repeated experiments). e.g. unbiasedness, significance levels, and power of tests, etc., violate the conditionality principle.

Example 14 (JB-p25)

- Suppose a substance to be analyzed can be sent either to a laboratory in New York or a laboratory in CA. The two labs seem equally good, so a fair coin is flipped to choose between them, with "heads" denoting that the lab in NY will be chosen. The coin is flipped and comes up tails, so the CA lab is used.

- After a while, the experimental results come back and a conclusion and report must be developed. Should this conclusion take into account the fact that the coin could have been heads and hence that the experiment in NY might have been performed instead?

- Common sense and conditional view point say **NO**, but the frequentist approach calls for averaging over all possible data, even the possible NY data.

• **Example 1.3.7:** In research laboratory, a physical quantity θ can be measured by a precise but often busy machine, which provides a measurement, $x_1 \sim \mathrm{N}(\theta, 0.1)$, with probability $p = 0.5$, or through a less precise but always available machine, which gives $x_2 \sim \mathrm{N}(\theta, 10)$. The machine being selected at random, depending on the availability of the more precise machine, the inference on θ when it has been selected should not depend on that fact that the alternative machine *could have been selected.*

- **Theorem 1.3.8** (Birnbaum, 1962) The Likelihood Principle is equivalent to the conjunction of the Sufficiency and the Conditionality Principles.

  See page 18 for the proof.

- Read Sec 1.3.4 for more about the likelihood principle.