

01/06

HW#1: Due 01/20 (Th)

- A lot more not mentioned: ~~t -distribution~~, Laplace (double-exponential) distribution, F -distribution
- Distributions can be parameterized in different ways. Please be careful when working on problems from JB since JB uses a parameterization different from that in CB.

Bayesian Lasso by Park & Casella (2012) in JASA

- Simulating Random Samples from R

- ★★ Use built-in functions. e.g.; rmnorm, dnorm, pnorm, qnorm...

- ★★ Use relationships between distributions.

- ★★ Use relationship $p(x, y) = \underline{p(x)p(y \mid x)}$ to simulation from a joint distribution when possible

$$\tilde{\mathbf{x}} = (\tilde{x}_1, \dots, \tilde{x}_p)$$

$$\tilde{x}^{(b)}, \quad b=1, \dots, B$$

• Example 1: Dirichlet distribution

$$a_p > 0$$

★★ Obtain a random sample from a Dirichlet distribution $\mathbf{x} = (x_1, \dots, x_k) \sim \text{Dir}(a_1, \dots, a_k)$.

$$\begin{cases} 0 \leq x_p \leq 1 \\ \sum_{p=1}^k x_p = 1 \end{cases}$$

★★ (Step 1:) Simulate $\tilde{x}_p \sim \text{Gamma}(\underline{a}_p, \underline{c})$, $p = 1, \dots, k$, where \tilde{x}_p 's are ^{mutually} independent and $c > 0$ is an arbitrary constant. Then let $\underline{x}_p = \tilde{x}_p / \sum_{p'=1}^k \tilde{x}_{p'}$, $p = 1, \dots, k$. $(x_1, \dots, x_k) \sim$

★★ (Step 2:) Repeat until the target sample size is met.

- Example 1: Dirichlet distribution (contd)

★★ Simulate $\mathbf{x} \sim \text{Dirichlet}(3, 1, 2)$

$$\sum x_p = 1$$

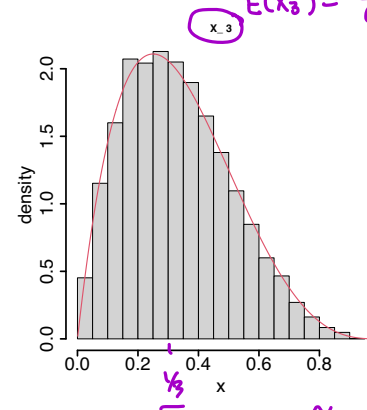
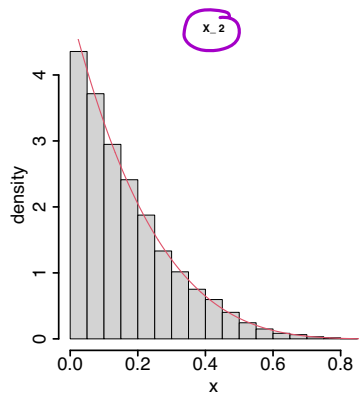
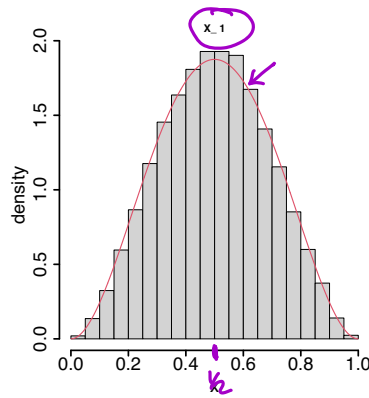
$$\mathbf{x} = (x_1, x_2, x_3), \quad K=3$$

$$E(x_1) = \frac{3}{3+1+2} = \frac{1}{2}$$

$$E(x_2) = \frac{1}{6}$$

$$E(x_3) = \frac{2}{6} = \frac{1}{3}$$

$$E(x_p) = \frac{\alpha_p}{\sum_{p=1}^K \alpha_p}$$



$$x_1 \sim \text{Be} \left(\begin{matrix} \alpha_1 \\ \alpha_2 + \alpha_3 \end{matrix} \right) = \text{Be} \left(\begin{matrix} 3 \\ 3 \end{matrix} \right)$$

$$\tilde{x}_1, \tilde{x}_2, \tilde{x}_3$$

$$\tilde{y} = \tilde{x}_2 + \tilde{x}_3$$

$$\frac{\tilde{x}_1}{\tilde{x}_1 + \tilde{y}} = x_1$$

- Example 2: IG distribution

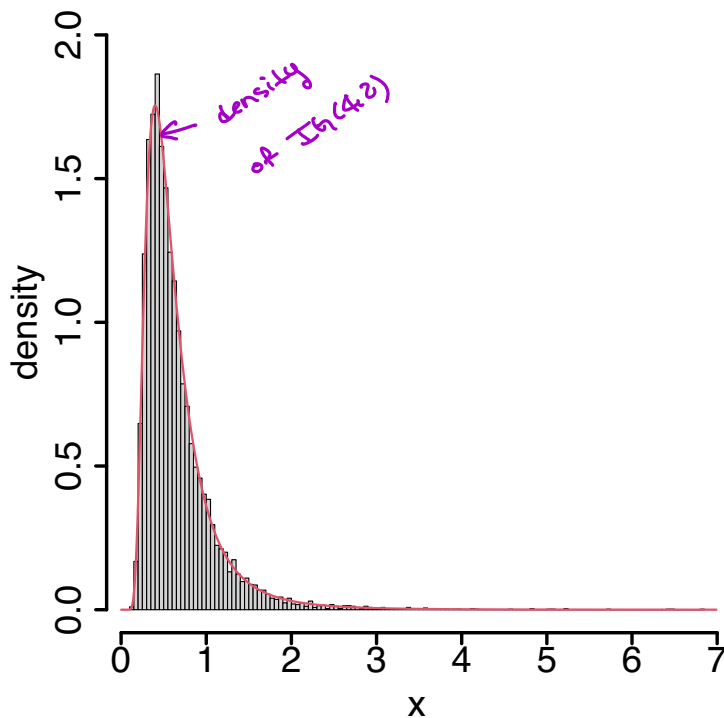
- ★★ Obtain a random sample from an inverse Gamma distribution, $x \sim \text{IG}(a, b)$.
- ★★ (Step 1:) Simulate $\tilde{x} \sim \text{Gamma}(a, b)$, where b is a rate parameter (so $E(\tilde{x}) = a/b$). Then let $x = 1/\tilde{x}$.
- ★★ (Step 2:) Repeat until the target sample size is met.

- Example 2: IG (contd)

★★ Simulate $\mathbf{x} \sim \text{IG}(4, 2)$

$$\tilde{\mathbf{x}} \sim \text{Ga}(\mathbf{4}, 2)$$

$$\mathbf{x} = \frac{1}{\tilde{\mathbf{x}}}$$



- Example 3: Normal \times IG distribution

★★ Suppose we have

$$\begin{aligned}\underline{p(x, y)} &= \underline{p(x)p(y | x)} \\ &= \underbrace{\frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp\left(-\frac{\beta}{x}\right)}_{\text{IG}(x | \alpha, \beta)} \underbrace{\frac{1}{\sqrt{2\pi x}} \exp\left(-\frac{(y-m)^2}{2x}\right)}_{\text{N}(y | m, x)}.\end{aligned}$$

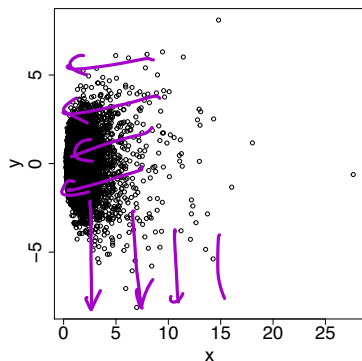
★★ Obtain a random sample of (x, y) from their joint $p(x, y)$.

★★ (Step 1:) Simulate $x \sim \text{IG}(x | \alpha, \beta)$ and $y | x \sim \text{N}(y | m, x)$.

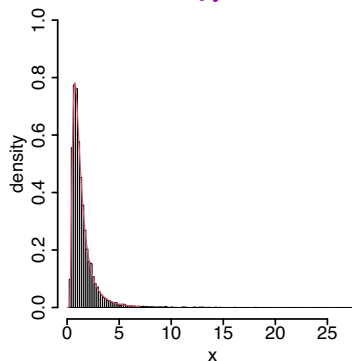
★★ (Step 2:) Repeat until the target sample size is met.

- Example 3: Normal \times IG distribution (contd)

★★ Simulate $(x, y) \sim \text{IG}(x \mid 3, 3) \text{N}(y \mid 0, x)$

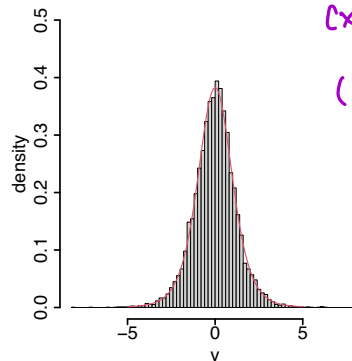


(a) (x, y)



(b) x

$\text{IG}(3, 3)$



(c) y

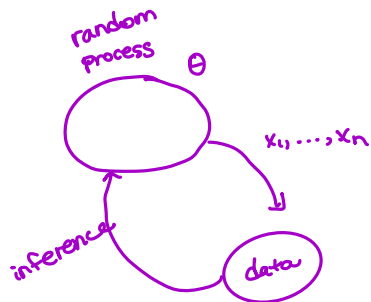
t_6

\mathcal{B}
 $x^{(b)} \sim \text{IG}(3, 3)$
 $y^{(b)} \mid x^{(b)} \sim \text{N}(0, x^{(b)})$
 $(x^{(1)}, y^{(1)})$
 $(x^{(2)}, y^{(2)})$
 $(x^{(3)}, y^{(3)})$
 \vdots
 $(x^{(B)}, y^{(B)})$

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, \dots, x_n are generated from the random phenomenon.
- Deduce an inference on θ from these 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 1987. 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.

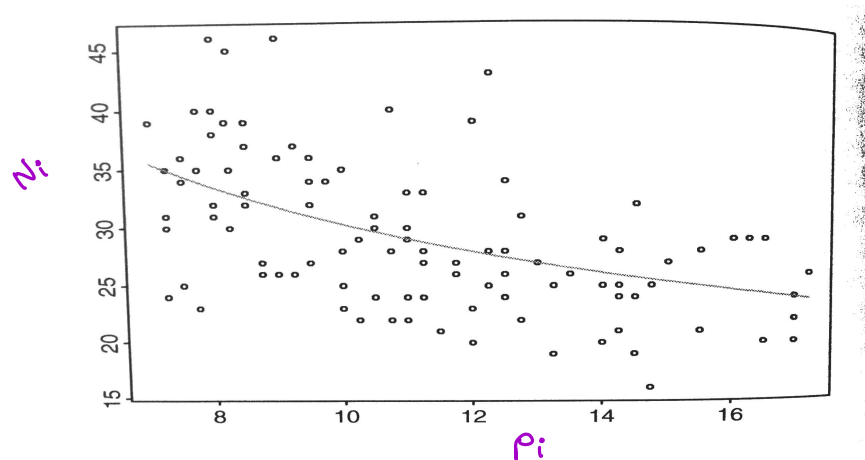


Figure 1.1.1. Plot of monthly unemployment rate versus number of accidents (in thousands) in Michigan, from 1978 to 1987. (Source: Lenk (1999).)

x

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

$f(x|\theta)$

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

★★ 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*.

★★ 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,

$$\mu_i = \exp(\beta_0 + \beta_1 \log(\rho_i)) \quad \beta_0, \beta_1 \in \mathbb{R}$$

$\mu_i \in \mathbb{R}^+$

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

★★ 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*.

★★ *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.

prediction

† Statistical Analysis

- Suppose x_1, \dots, x_n 's are random variables (observable or only hypothetically observable), $x_i \in \mathcal{X}$.
- Identify parameters θ 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 θ .
- Specify a joint probability distribution for the observable random variables (assume a parametric function for this course);

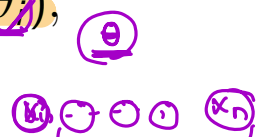
$$\underline{f(\mathbf{x} \mid \theta)},$$

where $\mathbf{x} = (x_1, \dots, x_n)$ and θ .

† Statistical Analysis – contd

- Say x_1, \dots, x_n are **conditionally independent given θ** ;

$$f(x_1, \dots, x_n | \theta) = \prod_{i=1}^n f(x_i | \theta) = \prod_{i=1}^n f(x_i | \theta_i),$$

$\{\theta = (\theta_1, \dots, \theta_n)\}$ $\theta_j \rightarrow x_i \quad i \neq j$


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 θ .

- Further assume that θ_i are all equal, i.e., $\theta_1 = \dots = \theta_n = \theta$
 $\Leftrightarrow x_i$'s are **conditionally independent and identically distributed (iid)** from a common distribution;

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

† 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.

$x_i, \quad i=1, \dots, 20$: assume independence among x_i ,

$$x_i = \begin{cases} 0 & \text{no infection} \\ 1 & \text{infection} \end{cases}$$

$$\mathbf{x} = (x_1, \dots, x_{20})$$

$$\begin{aligned} f(\mathbf{x} | \theta) &= \prod_{i=1}^{20} f(x_i | \theta) \\ &= \prod_{i=1}^{20} \theta^{x_i} (1-\theta)^{1-x_i} \\ &= \theta^{\sum x_i} (1-\theta)^{20 - \sum x_i} \end{aligned}$$

$x_i | \theta_i \overset{\text{indep}}{\sim} \text{Ber}(\theta_i) \quad \theta_i \in (0, 1)$

θ_i

$$\theta_1 = \theta_2 = \dots = \theta_{20} = \theta$$

$$\Rightarrow x_i | \theta \overset{\text{iid}}{\sim} \text{Ber}(\theta), \quad \text{given } \theta \in (0, 1)$$

† Statistical Problems

- **Definition 1.1.7** A parametric statistical model consists of the observation of a random variable x , distributed according to $f(x | \theta)$, 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 , $p(y | x)$

† Bayesian Paradigm (CR 1.2)

- **Definition 1.2.1** A Bayesian statistical model is made of a parametric statistical model, $f(\mathbf{x} \mid \theta)$, 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,

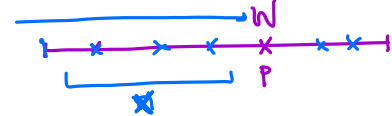
$$\underbrace{\ell(\theta \mid \mathbf{x})}_{\text{likelihood}} = \underbrace{f(\mathbf{x} \mid \theta)}_{\text{joint sampling distribution}},$$

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 to summarize the available information about θ** .



- 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 . *Given X , what inference can we make on p ?*

$$X \in \{0, \dots, n\}$$

$$\rightarrow p \sim \text{Unif}(0, 1) = \mathcal{B}(1, 1)$$

$$\rightarrow X | p \sim \text{Bin}(n, p) \quad p \in (0, 1)$$

$$\rightarrow p | X \sim \text{---}$$

† 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...

★★ 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 $f(x | \theta)$) into the posterior distribution.
- The **joint distribution** $\psi(x, \theta) = \pi(\theta)f(x | \theta)$
- The **marginal distribution**

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

prior probability distribution →

† Prior and Posterior Distributions – contd

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

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

- 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 .

⇒ Give a composite picture of the final beliefs about θ .

⇒ All decisions and inferences are made from $\pi(\theta | x)$.

† Prior and Posterior Distributions – contd

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

$$\underline{g(y | x)} = \int_{\Theta} \overbrace{g(y | \theta, x) \pi(\theta | x)}^{h(y, \theta | x)} d\theta.$$

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

$$\underline{g(y | x)} = \int_{\Theta} \overbrace{g(y | \theta) \pi(\theta | x)}^{= g(y | \theta, x)} d\theta$$

**** Note:** $m(y) = \int_{\Theta} g(y | \theta) \pi(\theta) d\theta$ 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, $\theta \in (0, 1)$, $\alpha > 0, \beta > 0$

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

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