Mathematics Bootcamp

Part III: Probability and Distribution Theory

Brian Cozzi[1] Michael Valancius[1] Graham Tierney[1]
Becky Tang[1]

[1]Department of Statistical Science
Duke University

Graduate Orientation, August 2019

Outline

Random Variables

Random Variables

Definition: A *random variable* is a function from the sample space to the real numbers.
Suppose we want to perform a survey, run an experiment, do some quantitative study of a population of interest...

Let Ω be the set of all possible outcomes of a study.

Let $\omega \in \Omega$ be a particular outcome unit.

$Y = Y(\omega)$ is a function of ω and a random variable.

Random Variables - Example

The Experiment: 2 Dice are rolled together
The Sample Space: All pairs of numbers from 1 through 6
The Random Variable: The sum of the numbers

Density and Mass Functions of Random Variables

Related to any random variable X is the concept of probability *density* and probability *mass* functions. Specifically, a *probability mass function* (PMF) for a discrete random variable is defined as:

$$f_X(x) = P(X = x); \forall x$$

and the *probability density function* (PDF) for a continuous random variable is defined as a function that satisfies the following relationship:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\mathrm{d}t; \forall x$$

Density and Mass Functions of Random Variables - Example

An example of a density function for a Geometric Random variable from the coin tossing example earlier:

$$f_X(x) = P(X = x) = (1-p)^{x-1}p \cdot \mathbf{I}(x \in 1, 2, 3, \ldots)$$

Notice that we can use the PMF (and analogously the PDF) to derive the CDF:

$$P(X \leq b) = \sum_{k=1}^{b} f_X(k) = F_X(b)$$

This partial sum is what we had used to reach the Geometric CDF presented earlier

Cumulative Distribution Functions of Random Variables

Definition: The cumulative distribution function (CDF) or a random variable denoted by $F_X(x)$ is defined as:

$$F_X(x) = P_X(X \leq x); \quad \forall x$$

A function is a CDF if and only if the following are true:

- $\lim_{x \to -\infty} F(x) = 0$ and $lim_{x \to \infty} F(x) = 1$
- $F(x)$ is a non-decreasing function of x
- $F(x)$ is right continuous i.e. for every number $x_0, \lim_{x \to x_0^+} F(x) = F(x_0)$

An important implication of CDFs: *A random variable X is continuous if $F_X(x)$ is a continuous function of x. A random variable is discrete if $F_X(x)$ is a step function of x.*

Cumulative Distribution Functions of Random Variables - Example

If p denotes the probability of getting a head on any toss, and the experiment consists of tossing a coin until a head appears, then we define the random variable $X =$ the number of tosses required until a head. The CDF of this random variable is given as:

$$P(X \leq x) = \sum_{i=1}^{x} (1-p)^{i-1} p$$

Exponential Families

Exponential Families

Random variables belong to the exponential family if their PMFs/PDFs can be expressed in the form:

$$f_X(x|\theta) = h(x)c(\theta)\exp\left\{\sum_{i=1}^{k}\omega_i(\theta)t_i(x)\right\}$$

Where:

$$h(x) \geq 0$$
$$c(\theta) \geq 0$$
$$\omega_1(\theta), \ldots, \omega_k(\theta) \in \mathbb{R}$$
$$t_1(x), \ldots t_k(x) \in \mathbb{R}$$

We will show later some of the convenient properties of the exponential family distributions.

Exponential Families - Example

Consider the binomial PMF for a random variable $X \sim \text{Binomial}(n, p)$

$$P(X = x) = \frac{n!}{(n-x)!x!} p^x (1-p)^{n-x}$$

This is an exponential family PMF. We can show this by re-expressing terms:

$$P(X = x) = \frac{n!}{(n-x)!x!} (1-p)^n \exp\{x \log\left(\frac{p}{1-p}\right)\}$$

$$h(x) = \frac{n!}{(n-x)!x!} \mathbb{I}_{x=0,\ldots,n}$$

$$c(p) = (1-p)^n$$

$$\omega_1(p) = \log\left(\frac{p}{1-p}\right)$$

$$t_1(x) = x$$

Exponential Families - Exercise

Consider the following normal PDF for $X \sim \mathrm{Normal}(\mu, \sigma^2)$

$$f_X(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$$

Show that this is an exponential family PDF

Exponential Families - Exercise Cont.

Consider the following PDF

$$f_X(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$$

Expanding the exponential yields the following

$$f_X(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{\mu^2}{2\sigma^2}\} \exp\{-\frac{x^2}{2\sigma^2} + \frac{\mu x}{\sigma^2}\}$$

$$h(x) = 1$$

$$c(\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{\frac{-\mu^2}{\sigma^2}\}; \mu \in \mathbb{R} \quad \sigma^2 > 0$$

$$\omega_1(\mu,\sigma) = \frac{1}{\sigma^2} \quad \omega_2 = \frac{\mu}{\sigma^2}$$

$$t_1(x) = \frac{-x^2}{2} \quad t_2(x) = x$$

Discrete Distributions

A random variable X is *discrete* if the range of X, the sample space, is countable. In most situations, the random variable has integer valued outcomes

Some examples of discrete distributions:

- Binomial Distribution
- Poisson Distribution
- Negative Binomial Distribution

Binomial Distribution

This distribution counts the the number of successes in n independent trials all with the same fixed probability p of success

$$X \sim \text{Binomial}(n, p)$$

$$P(X = x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

$$\mathbb{E}[X] = np$$

$$\mathbb{V}[X] = np(1-p)$$

Poisson Distribution

This distribution is used for counting the number of events over some time horizon based on an intensity parameter λ

$$X \sim \text{Poisson}(\lambda)$$

$$P(X = x) = \frac{\exp^{-\lambda} \lambda^x}{x!}$$

$$\mathbb{E}[X] = \mathbb{V}[X] = \lambda$$

Poisson Distribution - Exercise

Prove that $\mathbb{E}[X] = \lambda$ if $X \sim \text{Poisson}(\lambda)$

Poisson Distribution - Exercise Cont.

We need to compute the following:

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!} \\ &= \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!}\end{aligned}$$

Now recognize the following result from the Taylor series expansion on $\exp\{y\} = \sum_{i=0}^{\infty} \frac{y^i}{i!}$. Use this result with a clever substitution:

Negative Binomial Distribution

This distribution counts the the number successful trials k that occur before the rth failed trial, where each trial has fixed probability p of success

$$X \sim \mathrm{NB}(p, r)$$

$$P(X = n) = \frac{(k + r - 1)!}{k!(r - 1)!} p^k (1 - p)^r$$

$$\mathbb{E}[X] = \frac{pr}{1 - p}$$

$$\mathbb{V}[X] = \frac{pr}{(1 - p)^2}$$

Is highly related to the Poisson and Gamma Distributions

Continuous Distributions

A random variable X is *continuous* if the range of X, the sample space, takes on an uncountably infinite number of values. In most instances the random variable has real-valued outcomes.

Some examples of Continuous Distributions

- Normal Distribution
- Chi-Squared Distribution
- Exponential Distribution
- Gamma Distribution
- Inverse-Gamma Distribution
- Student-t Distribution
- F Distribution
- Beta Distribution

Normal Distribution

A random variable $X \sim \mathrm{Normal}(\mu, \sigma^2)$ with PDF:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$$

We also sometimes express this in terms of a *precision* parameter, rather than a variance, $X \sim \mathrm{Normal}(\mu, \phi^{-1})$ which becomes useful when performing Bayesian inference. If $Z \sim \mathrm{Normal}(0, 1)$ then the distribution of Z is standard normal.

Chi-Squared Distribution

If $Z_1, Z_2, \ldots, Z_k$ are independent, standard normal random variables, then

$$\sum_{j=1}^{k} Z_j^2 \sim \chi_k^2$$

follows a Chi-Squared distribution with k degrees of freedom. This is a special case of the Gamma distribution, discussed on the next slide.

Gamma Distribution

A random variable $X \sim \mathrm{Gamma}(\alpha, \beta)$ with PDF:

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp\{-x\beta\}$$

$$\mathbb{E}[X] = \frac{\alpha}{\beta}$$

$$\mathbb{V}[X] = \frac{\alpha}{\beta^2}$$

$$\alpha, \beta > 0$$

$$x \in (0, \infty)$$

Gamma Distribution - Important Properties

Here are some important tricks that will be useful in **711** and **601**

- if $\alpha = 1$ and then $X \sim \text{Exponential}(\lambda = \beta)$
- if $\alpha = \frac{\nu}{2}$ and $\beta = \frac{1}{2}$ then $X \sim \chi^2_\nu$
- if $X \sim \text{Gamma}(\alpha_1, \beta)$ and $Y \sim \text{Gamma}(\alpha_2, \beta)$ then $X + Y \sim \text{Gamma}(\alpha_1 + \alpha_2, \beta)$
- if $X \sim \text{Gamma}(k, \theta)$, then $\frac{1}{X} \sim \text{Inverse} - \text{Gamma}(k, \frac{1}{\theta})$

Student's-t Distribution

A random variable T follows a Student's-t distribution if

$$T = \frac{Z}{\sqrt{V/\nu}},$$
$$Z \sim N(0,1),$$
$$V \sim \chi^2_\nu$$

and Z and V are independent.

F Distribution

A random variable X follows a F-distribution with numerator degrees of freedom ν_1 and denominator degrees of freedom ν_2 if

$$X = \frac{V_1/\nu_1}{V_2/\nu_2}$$

where V_1 and V_2 are independent chi-squared random variables with degrees of freedom equal to ν_1 and ν_2 respectively.

Beta Distribution

A random variable $X \sim \text{Beta}(\alpha, \beta)$ with PDF:

$$f_X(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}$$

$$\mathbb{V}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

$$\alpha, \beta > 0$$

$$x \in (0, 1)$$

Very useful for eliciting probability distributions for proportions.

Cool distributional relationships

Transformations of Random Variables using the Change of Variables Formula

Assume that X has a pdf $f_X(x)$ and that $Y = g(X)$ where g is a monotone function. Suppose that $f_X(x)$ is continuous on $\mathcal{X}$, and that g^{-1} has a continuous derivative on $\mathcal{Y}$ where $\mathcal{X}, \mathcal{Y}$ are such that $\mathcal{X} = \{x : f_X(x) > 0\}$ and $\mathcal{Y} = \{y : y = g(x)\}$. Then the pdf of Y is given as follows:

$$f_Y(y) = f_X(g^{-1}(y))|\frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y)|$$

Transformations of Random Variables - Example

Assume that $X \sim f_X(x) = 1$ i.e. $X \sim \text{Uniform}(0, 1)$. Furthermore, $Y = -\log(X)$. What is the PDF of Y?

First note that $g(X) = Y = -\log(X) \rightarrow g^{-1}(Y) = e^{-Y}$.
Therefore, using the formulation from earlier:

$$f_Y(y) = 1 \cdot |-e^{-Y}| = e^{-Y}$$
$$Y \sim \text{Exponential}(\lambda = 1)$$

Transformations of Random Variables - Exercise

Assume that $X \sim \mathrm{Normal}(0, 1)$. Let $Y = X^2$. What is the distribution of Y?

The PDF of the standard normal distribution is given as follows:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{x^2}{2}\}$$

Transformations of Random Variables - Exercise Cont.

Consider that $Y = g(X) = X^2 \rightarrow g^{-1}(Y) = \mp\sqrt{Y}$. Hence, consider that we can partition the support of X into two pieces $S_1 = (-\infty, 0)$ and $S_2 = (0, \infty)$ where the function $g(X)$ is monotone. Note that $\mathcal{Y} = (0, \infty)$. Use the change of variables formulation over the two partitions and sum:

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} \exp\{\frac{-(-\sqrt{Y})^2}{2}\}|-\frac{1}{2\sqrt{y}}| + \frac{1}{\sqrt{2\pi}} \exp\{\frac{-(\sqrt{Y})^2}{2}\}|\frac{1}{2\sqrt{y}}|$$
$$= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{Y}} \exp\{\frac{-y}{2}\}$$

Hence, we get that $Y \sim \chi^2_{df=1}$

Moments

Expectations and Variances of Random Variables

The expectation of any random variable can be computed as follows:

- $\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) \mathrm{d}x$ when X is continuous
- $\mathbb{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x) f_X(x) = \sum_{x \in \mathcal{X}} g(x) \mathbb{P}(X = x)$ when X is discrete

The variance can be computed using the expectations as follows:

$$\mathbb{V}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

You will need to do some calculus to find each of these quantities

Kernel Tricks for Computing Expectations - Example

If we say that $X \sim \text{Exponential}(\lambda)$, with PDF $f_X(x) = \lambda \exp\{-\lambda x\}$. In order to find $\mathbb{E}[X]$, you must find:

$$\mathbb{E}[X] = \int_0^\infty x\lambda \exp\{-\lambda x\} \mathrm{d}x$$

- Integration by parts
- Something a bit more clever?

Kernel Tricks for Computing Expectations - Example Cont.

First, notice that if we say that $X \sim \text{Gamma}(\alpha, \beta)$ and PDF $g_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$

In instances when $\alpha = 1$ then this an Exponential random variable with $\lambda = \beta$.

The integrand from the previous slide, is *almost* like a Gamma PDF with $\alpha = 2$. Hence, you can complete it by some clever multiplication and division:

$$\mathbb{E}[X] = \frac{\Gamma(2)}{\lambda} \int_0^\infty \frac{\lambda}{\Gamma(2)} x^{2-1} \lambda \exp\{-\lambda x\} \mathrm{d}x = \frac{1}{\lambda} \cdot 1 = \frac{1}{\lambda}$$

Kernel Tricks for Computing Expectations - Exercise

Use the kernel trick for Exponential random variables to find $\mathbb{V}[X]$

Kernel Tricks for Computing Expectations - Exercise

The first step in this process is to find $\mathbb{E}[X^2]$ which you can calculate using the kernel trick as follows:

$$\mathbb{E}[X^2] = \frac{\Gamma(3)}{\lambda^2} \int_0^\infty \frac{\lambda^2}{\Gamma(3)} x^{3-1} \lambda \exp\{-\lambda x\} \mathrm{d}x = \frac{2}{\lambda^2} \cdot 1 = \frac{2}{\lambda^2}$$

Plug this result into the variance formula presented earlier using the expectation from earlier:

$$\mathbb{V}[X] = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

Note: You will not have to use a lot of integration here. Always try these tricks first

Properties of Expectations and Variances

Let X be a random variable and a a scalar constant, then:

$$\mathbb{E}[aX] = a\mathbb{E}[X]$$
$$\mathbb{V}[aX] = a^2\mathbb{V}[X]$$

Variances also have nice properties. Consider two random variables X and Y.

$$\mathbb{V}[X \mp Y] = \mathbb{V}[X] + \mathbb{V}[Y] \mp 2\mathbb{C}(X, Y),$$

where $\mathbb{C}(X, Y)$ denotes the covariance between X and Y. These extend to multivariate random variables as well.

Total Expectation and Total Variance Laws

In many examples, you are interested in marginal moments from conditional distributions. Your first option of course if to find the joint distribution, do some marginalization and then integrate, but I do not like calculus so **instead**:

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]]$$
$$\mathbb{V}[Y] = \mathbb{V}[\mathbb{E}[Y|X]] + \mathbb{E}[\mathbb{V}[Y|X]]$$

Total Expectation and Total Variance Laws - Example

Assume that we have the following relationship:

$$X|N \sim \text{Binomial}(N, p)$$

$$N \sim \text{NegativeBinomial}(\tau = \frac{1}{1+\beta}, r = 1)$$

Find $\mathbb{E}[X]$ and $\mathbb{V}[X]$

Tip: $\mathbb{E}[N] = \frac{r\tau}{1-\tau}$ and $\mathbb{V}[N] = \frac{\tau r}{(1-\tau)^2}$

Total Expectation and Total Variance Laws - Example Cont.

First, we iterate to find the expectation

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[\mathbb{E}[X|N]] \\ &= \mathbb{E}[Np] \\ &= p\frac{\frac{1}{1+\beta}}{1-\frac{1}{1+\beta}} \\ &= \frac{p}{\beta}\end{aligned}$$

Next, we proceed with finding the variance

$$\begin{aligned}\mathbb{V}[X] &= \mathbb{E}[\mathbb{V}[X|N]] + \mathbb{V}[\mathbb{E}[X|N]] \\ &= \mathbb{E}[Np(1-p)] + \mathbb{V}[Np] \\ &= \frac{p(1-p)}{\beta} + p^2\frac{1+\beta}{\beta^2}\end{aligned}$$

Total Expectation and Total Variance Laws - Exercise

$$X|P \sim \text{Binomial}(n, P)$$
$$P \sim \text{Beta}(a, b)$$

Find the $\mathbb{E}[X]$ and $\mathbb{V}[X]$

Tip:

$$\mathbb{E}[P] = \frac{a}{a+b}$$
$$\mathbb{V}[P] = \frac{ab}{(a+b)^2(a+b+1)}$$

Total Expectation and Total Variance Laws - Exercise Cont.

We can start by finding the marginal expectation first:

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|P]] = \mathbb{E}[nP] = n\mathbb{E}[P] = n\frac{a}{a+b}$$

And then the marginal variance:

$$\begin{aligned}
\mathbb{V}[X] &= \mathbb{V}[\mathbb{E}[X|P]] + \mathbb{E}[\mathbb{V}[X|P]] \\
&= \mathbb{V}[nP] + \mathbb{E}[nP(1-P)] \\
&= n^2\mathbb{V}[P] + n\mathbb{E}[P - P^2] \\
&= n^2\frac{ab}{(a+b)^2(a+b+1)} + n\frac{a}{a+b} \\
&- n(\frac{ab}{(a+b)^2(a+b+1)}) - n(\frac{a}{a+b})^2 \\
&= n\frac{ab(a+b+n)}{(a+b)^2(a+b+1)}
\end{aligned}$$

Moment Generating Functions

- The moment generating function (MGF)

$$M_x(t) = \mathbb{E}[e^{tX}]$$

  uniquely defines the distribution of a random variable
- So for X discrete, $M_x(t) = \sum_{x \in S_x} e^{tx} P(X = x)$
- And for X continuous, $M_x(t) = \int_{x \in S_x} e^{tx} f_X(x) dx$
- Why do we care? Besides the uniqueness property, MGFs are extremely helpful for determining distributions of sums of independent random variables

MGF Example: Gamma Distribution

Suppose $X \sim Gamma(\alpha, \beta)$. Then

$$
\begin{aligned}
M_x(t) &= \int_0^\infty e^{xt} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} dx \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-x(\beta-t)} dx \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x \cdot x^{(\alpha-1)-1} e^{-x(\beta-t)} dx \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha-1)}{(\beta-t)^{\alpha-1}} \times \\
&\quad \int_0^\infty x \cdot \frac{(\beta-t)^{\alpha-1}}{\Gamma(\alpha-1)} x^{(\alpha-1)-1} e^{-x(\beta-t)} dx \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha-1)}{(\beta-t)^{\alpha-1}} \mathbb{E}[Y], \quad Y \sim Ga(\alpha-1, \beta-t)
\end{aligned}
$$

MGF Example: Gamma Distribution

Recalling the expectation of a Gamma and simplifying terms:

$$\begin{aligned} M_x(t) &= \frac{\beta^\alpha}{(\alpha-1)!}\frac{(\alpha-2)!}{(\beta-t)^{\alpha-1}} \cdot \frac{\alpha-1}{\beta-t} \\ &= \left(\frac{\beta}{\beta-t}\right)^\alpha, \quad \text{for } t < \beta \end{aligned}$$

MGF of Chi-Squared

Rather than brute-force deriving the MGF for the χ^2, we can use the MGF of a Gamma random variable. We previously noted that if $Y \sim Gamma(k/2, 1/2)$, then Y is also distributed as χ^2_k.
So the MGF of $Y \sim \chi^2_k$ is

$$M_y(t) = \left(\frac{1/2}{1/2 - t}\right)^{k/2} = \left(\frac{1}{1 - 2t}\right)^{k/2} = (1 - 2t)^{-k/2}$$

Random Matrices and Multivariate Statistics

Random Vectors

If we have d random variables $X_1, X_2, \ldots, X_d$, each defined on the real line, we can write them as the d dimensional column vector

$$\mathbf{X} = (X_1, \cdots X_d)^T$$

which we call a d-dimensional **random vector**. The joint distribution function of the random vector $\mathbf{X}$ is

$$\begin{aligned} F_X(\mathbf{x}) &= F_X(x_1, \ldots, x_d) \\ &= P(X_1 \leq x_1, \ldots, X_d \leq x_d) \\ &= P(\mathbf{X} \leq \mathbf{x}) \end{aligned}$$

If F_X is absolutely continuous, then the joint density function f_X of $\mathbf{X}$ is

$$f_X(\mathbf{x}) = f_X(x_1, \ldots, x_d) = \frac{\partial^d F_X(x_1, \ldots, x_d)}{\partial x_1 \cdots \partial x_d}$$

Random Vectors

To find the marginal density of a subset of the d variables, you can just integrate the others out. For example, if we have a joint bivariate density $f_{X_1,X_2}(x_1, x_2)$, then

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, x_2)dx_2 \qquad f_{X_2}(x_2) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, x_2)dx_1$$

The components of a random vector $\mathbf{X}$ are **independent** if the joint distribution function is a product of the marginal distribution functions

$$F_X(\mathbf{x}) = \prod_{i=1}^{d} F_i(x_i)$$

In addition, the joint density is the product of marginals

$$f_X(\mathbf{x}) = \prod_{i=1}^{d} f_i(x_i)$$

Expectation and Covariance

If $\mathbf{X}$ is a random vector with values in $\mathbb{R}^d$, then its expected value is given by the d dimensional vector

$$\mu_X = E(\mathbf{X}) = (E(X_1), \cdots, E(X_d)) = (\mu_1, \cdots, \mu_d)^T$$

and the $d \times d$ **covariance matrix** of $\mathbf{X}$ is

$$\begin{aligned}
\mathbf{\Sigma}_{XX} &= \text{cov}(\mathbf{X}, \mathbf{X}) \\
&= E[(\mathbf{X} - \mu_X)(\mathbf{X} - \mu_X)^T] \\
&= E[(X_1 - \mu_1, \cdots, X_d - \mu_d)(X_1 - \mu_1, \cdots, X_d - \mu_d)^T] \\
&= \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{pmatrix}
\end{aligned}$$

Correlation Matrix

The **correlation matrix** of $\mathbf{X}$ can be obtained by from $\mathbf{\Sigma}_{XX}$ by dividing the ith row by σ_i and the jth column by σ_j. The $d \times d$ matrix is then

$$P_{XX} = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1d} \\ \rho_{21} & 1 & \cdots & \rho_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{d1} & \rho_{d2} & \cdots & 1 \end{pmatrix}$$

where

$$\rho_{ij} = \rho_{ji} = \begin{cases} \frac{\sigma_{ij}}{\sigma_i \sigma_j} & i \neq j \\ 1 & \text{otherwise} \end{cases}$$

is the pairwise correlation coefficient between X_i and X_j. The correlation coefficient will always lie between -1 and 1 and is a measure of association between X_i and X_j.

Linear Functions of Random Vectors

If $\mathbf{Y}$ is a linear function of $\mathbf{X}$ such that

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$$

the mean vector and covariance matrix of $\mathbf{Y}$ is given by

$$\mu_Y = \mathbf{A}\mu_x + \mathbf{b}$$
$$\mathbf{\Sigma}_{YY} = \mathbf{A}\mathbf{\Sigma}_{\mathbf{XX}}\mathbf{A}^T$$

Multivariate Normal Distribution

The form of the multivariate normal looks similar to that of the univariate normal. A random d vector $\mathbf{X}$ follows a multivariate normal distribution with mean vector μ and positive definite symmetric covariance matrix $\mathbf{\Sigma}$ if it has the density function

$$f(\mathbf{x}|\mu, \mathbf{\Sigma}) = (2\pi)^{-d/2}|\mathbf{\Sigma}|^{-1/2}e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)}$$

We notationally denote a d dimensional normal distribution as

$$\mathbf{X} \sim N_d(\mu, \mathbf{\Sigma})$$

Multivariate Normal Distribution

The **Mahalanobis distance** from $\mathbf{x}$ to μ is given by the quadratic form

$$\Delta = \sqrt{(\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mu)}$$

An important result is that a random vector $\mathbf{X}$ follows a multivariate distribution if and only if every linear function of $\mathbf{X}$ follows a univariate normal distribution.

In linear models, we often assume that $\mathbf{\Sigma} = \sigma^2 \mathbf{I_d}$, in which case the density function reduces to

$$f(\mathbf{x}|\mu, \sigma) = (2\pi\sigma)^{-d/2} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T(\mathbf{x}-\mu)}$$

Partitioned Random Vectors

Suppose we have two random vectors $\mathbf{X}$ and $\mathbf{Y}$, where $\mathbf{X}$ has d_1 components and $\mathbf{Y}$ had d_2 components. Let $\mathbf{Z}$ be the random $d_1 + d_2$ vector

$$\mathbf{Z} = \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix}$$

Then the expected value and covariance matrix of $\mathbf{Z}$ is given by

$$\begin{aligned} \mu_Z = E[\mathbf{Z}] &= \begin{pmatrix} E[\mathbf{X}] \\ E[\mathbf{Y}] \end{pmatrix} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} \\ \mathbf{\Sigma}_{ZZ} &= \begin{pmatrix} cov(\mathbf{X}, \mathbf{X}) & cov(\mathbf{X}, \mathbf{Y}) \\ cov(\mathbf{Y}, \mathbf{X}) & cov(\mathbf{Y}, \mathbf{Y}) \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{\Sigma_{XX}} & \mathbf{\Sigma}_{XY} \\ \mathbf{\Sigma}_{YX} & \mathbf{\Sigma}_{YY} \end{pmatrix} \end{aligned}$$

where $\mathbf{\Sigma}_{XY} = \mathbf{\Sigma}_{YX}^T$.

Marginal/Conditional Normal Distribution

The marginal distribution of $\mathbf{Y}$ is

$$\mathbf{Y} \sim N_{d_2}(\mu_y, \mathbf{\Sigma}_{YY})$$

The conditional distribution of $\mathbf{Y}$ given that $\mathbf{X} = \mathbf{x}$ is multivariate normal with mean vector and covariance matrix given by

$$\mu_{Y|X} = \mu_Y + \mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}(\mathbf{x} - \mu_X)$$
$$\mathbf{\Sigma}_{Y|X} = \mathbf{\Sigma}_{YY} - \mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{XY}$$

Wishart Distribution

Given n i.i.d. d-dimensional vectors

$$\mathbf{X}_i \sim N_d(\mu, \mathbf{\Sigma}), \quad i = 1, 2, \ldots, n$$

we say that the random positive-definite, symmetric matrix

$$\mathbf{W} = \sum_{i=1}^{n} \mathbf{X_i} \mathbf{X_i}^T$$

follows a **Wishart distribution** with n degrees of freedom and matrix $\mathbf{\Sigma}$. We denote the Wishart distribution by

$$\mathbf{W} \sim \mathcal{W}_d(n, \mathbf{\Sigma})$$

You can think of the Wishart as a randomly drawn covariance matrix multiplied by the degrees of freedom n, since $E[\mathbf{W}] = n\mathbf{\Sigma}$. As $n \to \infty$, $\mathbf{W}/n \to \mathbf{\Sigma}$.

Properties of the Wishart Distribution

1. Let $\mathbf{W}_j \sim \mathcal{W}_d(n_j, \mathbf{\Sigma})$ be independent. Then $\sum_{j=1}^{m} \mathbf{W}_j \sim \mathcal{W}_d(\sum_{j=1}^{m} n_j, \mathbf{\Sigma})$
2. Suppose $\mathbf{W} \sim \mathcal{W}_d(n, \mathbf{\Sigma})$ and let $\mathbf{A}$ be a constant matrix having full row rank. Then $\mathbf{AWA}^T \sim \mathcal{W}_d(n, \mathbf{A\Sigma A}^T)$.
3. Suppose $\mathbf{W} \sim \mathcal{W}_d(n, \mathbf{\Sigma})$ and let $\mathbf{a}$ be a fixed d dimensional vector. Then $\mathbf{a}^T\mathbf{Wa} \sim (\mathbf{a}^T\mathbf{\Sigma a})\chi^2_n$.

You can think of the Wishart as a multidimensional chi-square distribution. If $\mathbf{W}$ follows a Wishart distribution, then $\mathbf{W}^{-1}$ follows an **inverse Wishart distribution**.

Review Exercises

Given that

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim N_3(\mu, \mathbf{\Sigma})$$

where

$$\mu = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix} \qquad \mathbf{\Sigma} = \begin{pmatrix} 3 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 2 \end{pmatrix}$$

1. Find the correlation matrix ρ of $\mathbf{X}$
2. Find the marginal distribution of X_2.
3. Find the marginal distribution of $\{X_1, X_3\}$.
4. Find the conditional distribution of $X_1 | X_3 = -1$.
5. Find the conditional distribution of $X_1 | \{X_2 = 1, X_3 = -1\}$
6. Are $\{X_1, X_3\}$ and X_2 independent?
7. Are $X_1 + X_2$ and $X_1 - X_2$ independent?

Solutions

1. Find the correlation matrix ρ of $\mathbf{X}$

$$\rho = \begin{pmatrix} 1 & 0 & 1/\sqrt{6} \\ 0 & 1 & 0 \\ 1/\sqrt{6} & 0 & 1 \end{pmatrix}$$

2. Find the marginal distribution of X_2.

$$X_2 \sim N(1, 1)$$

3. Find the marginal distribution of $\{X_1, X_3\}$.

$$\{X_1, X_3\} \sim N\left(\begin{pmatrix} 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}\right)$$

Solutions

4. Find the conditional distribution of $X_1|X_3 = -1$.

   Using the conditional distribution formula, the conditional distribution of $\{X_1, X_2\}$ given $X_3 = -1$ is

$$\mu = \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \end{pmatrix} (1/2)(0) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$\mathbf{\Sigma} = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 1 \\ 0 \end{pmatrix} (1/2) \begin{pmatrix} 1 \\ 0 \end{pmatrix}^T = \begin{pmatrix} 5/2 & 0 \\ 0 & 1 \end{pmatrix}$$

   So looking at the marginal, the conditional distribution of X_1 is $N(0, 5/2)$.

5. Find the conditional distribution of $X_1|\{X_2 = 1, X_3 = -1\}$
   $\{X_1, X_2\}$ given $X_3 = -1$ is
   You can do this using the conditional distribution formula or note that X_1 is independent of X_2 (from the next question). So the answer will be the same as above.

Solutions

6. Are $\{X_1, X_3\}$ and X_2 independent?
   Yes. Since they are multivariate normally distributed and the pairwise correlation between X_2 and $\{X_1, X_3\}$ is 0, they are independent.
7. Are $X_1 + X_2$ and $X_1 - X_2$ independent?
   No, the covariance between $X_1 + X_2$ and $X_1 - X_2$ is nonzero. Also, both terms involve X_1 and X_2 so there's no reason to expect them to be independent.

Probability Theory

Distribution Functions for Multivariate Random Variables

There are three types of distribution functions that we will cover:

- Joint Distribution
- Marginal Distribution
- Conditional Distribution

Joint Distribution - Bivariate Case

Joint PDF: A function $f(x, y)$ from $\mathbb{R}^2 \to \mathbb{R}$ is called a joint PDF of the random vector (X, Y) if for every $A \subset \mathbb{R}^2$

$$\mathbb{P}((X, Y) \in A) = \int_A \int f_{X,Y}(x, y) \mathrm{d}x \mathrm{d}y$$

Joint PMF: The function $f(x, y)$ from $\mathbb{R}^2 \to \mathbb{R}$ defined by $f_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y)$ is the joint PMF of X, Y. Then for every $A \subset \mathbb{R}^2$

$$\mathbb{P}((X, Y) \in A) = \sum_{(x,y) \in A} f_{X,Y}(x, y)$$

Joint Distribution - Exercise

Assume that X and Y have the joint PDF:

$$f_{X,Y}(x,y) = 4xy$$
$$0 < x < 1$$
$$0 < y < 1$$

Find $P(Y < X)$ * Show that this is proper *

Joint Distribution - Exercise Cont.

We can set up the double integral required for this probability as follows:

$$
\begin{aligned}
p(Y < X) &= \int_0^1 \int_0^x 4xy \mathrm{d}y \mathrm{d}x \\
&= \int_0^1 [4x\frac{y^2}{2}]\Big|_0^x \mathrm{d}x \\
&= \int_0^1 2x^3 \mathrm{d}x = \frac{1}{2}
\end{aligned}
$$

Marginal Distribution

Given the joint PDF or joint PMF, you can find the marginal PDF or PMF:

Marginal PDF:

$$f_X(x) = \int_Y f_{X,Y}(x, y)\mathrm{d}y$$

Marginal PMF:

$$f_Y(y) = \sum_x f_{X,Y}(x, y)$$

Conditional Probability and Independence

Starting with something familiar. Consider two events A and B with the sample space Ω.

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

Furthermore, consider the following notion of independence for the same two events. A and B are independent if:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$

Conditional Probability and Independence - Continued

Conditional Probability for more than two events. Let $A_1, A_2, \ldots$ be a partition of the sample space and let B be any set, then for $i = 1, 2, \ldots$:

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{i=1}^{\infty} \mathbb{P}(B|A_j)\mathbb{P}(A_j)}$$

We can similarly extend the definition of independence to cases with more then two events. A collection of events $A_1, \ldots, A_n$ are considered mutually independent if for any subcollection $A_{i_1}, \ldots, A_{i_K}$ we have that:

$$\mathbb{P}(\cap_{j=1}^{K} A_{i_j}) = \prod_{j=1}^{K} \mathbb{P}(A_{i_j})$$

Conditional Distribution

Assume that $X, Y \sim f_{X,Y}(x, y)$, then we can employ Bayes' rule for distributions:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

Conditional Probability - Example

In morse code, information is represented as dots and dashes. Assume the following:

$$\mathbb{P}(\textit{dot sent}) = \frac{3}{7}; \quad \mathbb{P}(\textit{dash sent}) = \frac{4}{7}$$

Furthermore, we also know that $\mathbb{P}(\textit{dot received} | \textit{dot sent}) = \frac{7}{8}$. Find $\mathbb{P}(\textit{dot sent} | \textit{dot received})$.

Conditional Probability - Example Cont.

In order to use Bayes Rule, we first need $\mathbb{P}(dot \quad received)$.

$$\mathbb{P}(dot \quad received) = \mathbb{P}(dot \quad received \cap dot \quad sent) +$$
$$\mathbb{P}(dot \quad received \cap \quad dash \quad sent) \qquad = \frac{7}{8}\frac{3}{7}$$
$$+ \left(\frac{1}{8}\right)\left(\frac{4}{7}\right) = \frac{25}{26}$$

Applying Bayes Rule:

$$\mathbb{P}(dot \quad sent | dot \quad received) = \frac{\mathbb{P}(dot \quad sent \cap dot \quad received)}{\mathbb{P}(dot \quad sent)}$$
$$= \frac{\left(\frac{7}{8}\right)\left(\frac{3}{7}\right)}{\frac{25}{26}}$$

Conditional Probability - Exercise

In the population the probability of an infectious disease is $\mathbb{P}(D) = 0.01$. The probability of testing positive if the disease is present is $\mathbb{P}(+|D) = 0.95$. The probability of a negative test given the disease is not present is $\mathbb{P}(-|ND) = 0.95$. What is the probability of the disease being present if the test is positive i.e. $\mathbb{P}(D|+)$?

Conditional Probability - Exercise Cont.

First find the probability of a positive test:

$$\begin{aligned}\mathbb{P}(+) = \mathbb{P}(+|D)P(D) + \mathbb{P}(+|ND)P(ND) &= 0.01 \cdot 0.95 + 0.05 \cdot 0.99 \\ &= 0.059\end{aligned}$$

Next, we can invoke Bayes Rule:

$$\mathbb{P}(D|+) = \frac{\mathbb{P}(D \cap +)}{\mathbb{P}(+)} = \frac{0.01 \cdot 0.95}{0.059} \approx 0.161$$

Conditional Distributions - Exercise

Assume that (X, Y) are a continuous random vector with joint pdf given by:

$$f_{X,Y}(x, y) = \exp\{-y\} \quad 0 < x < y < \infty$$

Find the marginal distribution of X and the conditional distribution $Y|X$

Conditional Distributions - Example Cont.

We start by finding the marginal distribution of X:

$$f_X(x) = \int_x^\infty \exp\{-y\}\mathrm{d}y = e^{-x}$$
$$X \sim \mathrm{Exponential}(\lambda = 1)$$

Now use the results on conditional distributions given earlier, to find:

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{\exp\{-y\}}{\exp\{-x\}}\mathbb{I}(x < y)$$

Independence - Example

Consider an experiment of tossing two dice. The sample space is therefore:

$$\Omega = \{(1,1),(1,2),\ldots(1,6),(2,1),\ldots,(2,6),\ldots,(6,6)\}$$

Further, we define the events:

$$A = \{\text{doubles appear}\}$$
$$B = \{\text{the sum is between 7 and 10}\}$$
$$C = \{\text{the sum is 2 or 7 or 10}\}$$

Are the events A, B, C mutually independent?

Independence - Example Cont.

Note that the following can be found by enumeration:

$$\mathbb{P}(A) = \frac{1}{6}; \quad \mathbb{P}(B) = \frac{1}{2}; \quad \mathbb{P}(C) = \frac{1}{3}$$

Furthermore:

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(\text{sum is } 10, \text{ comprised of doubles}) = \frac{1}{36}$$
$$= \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C) = \frac{1}{6} \cdot \frac{1}{2} \cdot \frac{1}{3}$$

But notice that $\mathbb{P}(B \cap C) = \frac{11}{36} \neq \mathbb{P}(B)\mathbb{P}(C)$. Therefore we do not have pairwise independence and hence claims of mutual independence cannot be made.

Independence - Exercise

Consider the following sample sample that consists of the 3! permutations of $\{a, b, c\}$ along with triples of each letter:

$$\Omega = \{aaa, bbb, ccc, abc, bca, cba, acb, bac, cab\}$$

Each element in Ω is assumed to have probability $\frac{1}{9}$. Define the event A_i:

$$A_i = \{i^{th} \text{ place in the triple is occupied by } a\};$$
$$i = 1, 2, 3$$
$$\mathbb{P}(A_i) = \frac{1}{3}$$

Are the events A_i mutually independent?

Independence - Exercise Cont.

Pairwise independence is satisfied:

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1 \cap A_3) = \mathbb{P}(A_2 \cap A_3) = \frac{1}{9}$$

But the joint event:

$$\mathbb{P}(A_1 \cap A_2 \cap A_3) = \frac{1}{9} \neq \mathbb{P}(A_1)P(A_2)\mathbb{P}(A_3)$$

Hence, the events are **not** mutually independent

Conditional Independence

True independence is pretty rare in most applications, so we generally rely on *conditional independence*.

Suppose we have three random variables Y_1, Y_2, Y_3 that we believe are *"independent and identically distributed (iid)"*. Does our knowledge about the value of one inform our knowledge about another?

$$Pr(Y_1 = y_1 \mid Y_2 = y_2, Y_3 = y_3) = P(Y_1 = y_1)?$$

Conditional Independence

Suppose Y_1, Y_2, Y_3 are *Conditionally Independent* given some parameter vector θ. This means that

$$Pr(Y_1 = y_1 \mid \theta, Y_2 = y_2, Y_3 = y_3) = Pr(Y_1 = y_1 \mid \theta)$$

Now we can say,

$$P(Y_1, Y_2, Y_3 \mid \theta) = P(Y_1 \mid \theta)P(Y_2 \mid \theta)P(Y_3 \mid \theta)$$

Suppose we want to find the value of θ that makes these data most likely... See MLE

Bayesian Analysis

Different Interpretations of Probability

In general, there are two main interpretations of probability, both of which are consistent with the axioms of probability discussed to this point.

- **Frequentists** posit that the probability of an event is its relative frequency over time, i.e., its relative frequency of occurrence after repeating a process a large number of times under similar conditions.
- The **Bayesian** interpretation, gives the notion of probability a subjective status by regarding it as a measure of the 'degree of belief' of the individual assessing the uncertainty of a particular situation.

Importance of the Distinction

There are some key consequences to the Bayesian interpretation of probability:

- No assumption about the randomness of a particular event. Instead, probability is a measure of our own uncertainty.
- The importance of this distinction is that probability statements can be made about a much larger class of objects.

Namely, in parametric models, the parameters are often assumed to have "true" (albeit unknown) values, Bayesian methods can use probability to describe the uncertainty about parameters. In this general framework, anything unknown can be described by a probability distribution.

Bayes Rule

Bayesian inference uses Bayes' theorem to update probabilities after more evidence or data is obtained.

Two quantities of interest:

1. $y \in \mathcal{Y}$: the data is a member of $\mathcal{Y}$, the *sample space* or the set of all possible datasets;
2. $\theta \in \Theta$: the parameter (Θ: the parameter space), expressing the population characteristics.

Bayesian Analysis: the Basics

Three Distributions:

1. For each numerical value $\theta \in \Theta$, the **prior distribution** $p(\theta)$ describes our belief that θ represents the true population characteristics;
2. For each $\theta \in \Theta$ and $y \in \mathcal{Y}$, the **sampling model** $p(y|\theta)$ describes our belief that y would be the outcome of the study if we knew θ to be true;
3. For each numerical value of $\theta \in \Theta$, the **posterior distribution** $p(\theta|y)$ describes our belief that θ is the true value, having observed dataset y.

Posterior Distribution

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

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

Bayes' rule tells us how our beliefs should change after seeing new information.
In practice, however, since evaluating $\int_\Theta p(y|\tilde{\theta})p(\tilde{\theta})d\tilde{\theta}$ is often intractable, the posterior is instead obtained by

$$p(\theta|y) \propto p(y|\theta)p(\theta),$$

and the form of the right hand side can help us determine $p(\theta|y)$.

Prior Distribution

Generally, the prior distribution for a parameter θ is a probability distribution that reflects our uncertainty about θ before data (or, if updating, new data) is taken into account.

The prior distribution is the choice of the person conducting the analysis and ideally provides useful information that might be known about θ a priori. For example, if we are interested in describing the probability that the US Women's National team defeating Thailand in a soccer match, we might a priori have more belief that the probability θ is closer to 1 than to 0.

Conjugacy

Definition

A class $\mathcal{P}$ of prior distributions for θ is called conjugate for a sampling model $p(y|\theta)$ if

$$p(\theta) \in \mathcal{P} \Rightarrow p(\theta|y) \in \mathcal{P}$$

- For example, as seen in the following exercise, the beta distribution is conjugate for the binomial sampling model. Thus, if $p(\theta) \sim Beta(a, b)$ and $p(y|\theta) \sim Binomial(\theta)$, then $p(\theta|y) \sim Beta(c, d)$.
- Conjugate priors have a practical advantage: they provide computational convenience and interpretability since the posterior will follow a known parametric form.
- However, they might not always be flexible enough, and for more complicated or higher dimensional problems they quickly become impossible to use.

Binomial Model

Let $Y \sim Binomial(n, \theta)$, where $Y \in \{0, 1, ...n\}$. Having observed $Y = y$, we conduct posterior inference: $p(\theta|y)$. Using Bayes rule:

$$\begin{aligned} p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{p(y)} \\ &= \frac{\binom{n}{y}\theta^y(1-\theta)^{n-y}p(\theta)}{p(y)} \\ &\propto c(y)\theta^y(1-\theta)^{n-y}p(\theta) \end{aligned}$$

If we choose a conjugate prior $p(\theta)$ to our likelihood, then we will have a closed form expression for the posterior.

Finding Conjugate Prior

For a binomial sampling model:

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

Thus, a conjugate prior is a distribution of θ such that $p(\theta) \propto \theta^{c_1}(1-\theta)^{c_2}$ as a function of θ. The Beta distribution satisfies this requirement, as is illustrated in the following example.

Binomial Model Continued

Recall that if $\theta \sim Beta(a, b)$, then $p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$.

$$
\begin{aligned}
p(\theta|y) &\propto c(y)\theta^{y}(1-\theta)^{n-y}p(\theta) \\
&= c(y)\theta^{y}(1-\theta)^{n-y}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \\
&= \left(c(y)\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right)\theta^{y+a-1}(1-\theta)^{n+b-y-1} \\
&= c_2(y)\theta^{y+a-1}(1-\theta)^{n+b-y-1}
\end{aligned}
$$

How do we confirm that the posterior distribution is also beta? Since the posterior distribution is a proper probability distribution, it integrates to 1. This fact, combined with some algebra, reveals the posterior distribution to be Beta(y + a, n + b - y).

MSE Example: Tedious Algebra

$$
\begin{aligned}
1 &= \int_0^1 c_2(y)\theta^{y+a-1}(1-\theta)^{n+b-y-1}d\theta \\
&\implies 1 = c_2(y)\int_0^1 \theta^{y+a-1}(1-\theta)^{n+b-y-1}d\theta \\
&\implies 1 = c_2(y)\frac{\Gamma(y+a)\Gamma(n+b-y)}{\Gamma(n+a+b)} \\
&\implies c_2(y) = \frac{\Gamma(n+a+b)}{\Gamma(y+a)\Gamma(n+b-y)} \\
&\implies p(\theta|y) = \frac{\Gamma(n+a+b)}{\Gamma(y+a)\Gamma(n+b-y)}\theta^{y+a-1}(1-\theta)^{n+b-y-1} \\
&\implies p(\theta|y) = Beta(y+a, n+b-y)
\end{aligned}
$$

A Binomial Example

A survey is carried out to study the support rate θ $(0 < \theta < 1)$ of a policy. 100 people are surveyed, and a binary response Y_i is obtained from each person i $(i = 1, 2, \ldots, 100)$, $Y_i \sim \text{Bernoulli}(\theta)$ (that is, $Y = \sum_{i=1}^{100} Y_i \sim \text{Binomial}(100, \theta)$).

Before the survey, we believe that $\theta \sim \text{Beta}(5, 5)$, while the result of the survey is $Y = 60$. We'd like to obtain the posterior distribution of θ given the survey outcome.

A Binomial Example (Cont'd)

The prior distribution is $\theta \sim \text{Beta}(5, 5)$, that is

$$p(\theta) = \frac{\theta^{5-1}(1-\theta)^{5-1}}{B(5,5)} \propto \theta^{5-1}(1-\theta)^{5-1}$$

The sampling distribution is $Y \sim \text{Binomial}(100, \theta)$, that is, for each $\theta \in (0, 1)$ and $y = 0, 1, \ldots, 100$,

$$P(Y = y|\theta) = \binom{100}{y} \theta^y (1-\theta)^{100-y}.$$

Using Bayes' rule, the posterior distribution of θ given that $Y = 60$ is

$$\begin{aligned} p(\theta|Y = 60) &\propto p(Y = 60|\theta)p(\theta) \\ &= \theta^{60}(1-\theta)^{100-60}\theta^{5-1}(1-\theta)^{5-1} \\ &= \theta^{65-1}(1-\theta)^{45-1}, \end{aligned}$$

which has the form of the p.d.f. of a $\text{Beta}(65, 45)$ distribution. Thus, we have $\theta|Y = 60 \sim \text{Beta}(65, 45)$.

Bayesian Updating

Bayesian inference provides a framework for updating beliefs upon observing data. There is an initial belief, described by the prior distribution. Data is observed. Following Bayes Rule, the beliefs are updated into what is called the posterior distribution.

Question: If we observe data $D_1 = (x_1 ... x_n)$ and find $p(\theta|x_1 ... x_n)$ and then later observe more data $D_2 = (x_{n+1} ... x_{n+m})$, is $p(\theta|D_1, D_2) \propto p(D_2|\theta)p(\theta|D_1)$?

In other words, can the first posterior we derived after observing D_1 be used as the prior for conducting posterior inference when new data D_2 is observed?

Bayesian Updating Continued

So long as D_1 and D_2 are treated as conditionally independent given θ, the answer is yes. Thus, Bayesian inference gives us a powerful tool for repeatably updating a model every time more data is observed. The former posterior distribution becomes the new prior once more data is observed.

Example: In our previous example, we found that for a Binomial(100,θ) model with a prior of $\theta \sim$Beta(5,5) and observed data of $Y = 60$, the posterior distribution is $\theta|Y = 60 \sim$Beta(65,45). Now suppose we observe 100 more surveys, this time with $Y_2 = 55$. How do our beliefs change?

$$\begin{aligned} p(\theta|Y_2 = 55) &\propto p(Y_2 = 55|\theta)p(\theta|Y_1 = 60) \\ &= \theta^{55}(1-\theta)^{100-55}\theta^{65-1}(1-\theta)^{45-1} \\ &= \theta^{120-1}(1-\theta)^{90-1} \end{aligned}$$

Thus, our updated posterior is Beta(120,90).

Bayesian Updating Continued

What if, instead of updating, we restarted the analysis with the original prior of $Beta(5,5)$ and now had $n = 100 + 100 = 200$ and $Y = Y_1 + Y_2 = 115$?

$$\begin{aligned} p(\theta|Y) &\propto p(Y = 115|\theta)p(\theta) \\ &= \theta^1 15(1-\theta)^{200-115}\theta^{5-1}(1-\theta)^{5-1} \\ &= \theta^{120-1}(1-\theta)^{90-1} \end{aligned}$$

This, as before, has the form of Beta(120,90).

Reference Guide

- *Statistical Inference* - Casella and Berger
- *Mathematical Statistics* - Bickel and Doksum
- *A First Course in Bayesian Statistical Methods - Hoff*
- *Bayesian Computation with*R - Albert