Probability Lecture II (August, 2006)

1 More on Named Distribution

1.1 Normal distribution

A random variable X has normal(μ, σ^2) distribution, if the probability density function of X is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2/\sigma}; \text{-}\infty < x < \infty. \tag{1}$$

In the case where $\mu = 0$ and $\sigma = 1$, the distribution is called standard normal distribution. It can be showed that if X has normal(μ, σ^2), then $E(X) = \mu$, and $\text{Var}(X) = \sigma^2$.

Properties The normal density curve has the following properties:

1. it is symmetric about μ with a bell-shape curve, concave on either side of μ.
2. The areas under the curve within $1, 2$, and 3 σ from μ are 68%, 95% and 99.7%, respectively.
3. $X = \sigma Z + \mu$ has normal(μ, σ^2) distribution, where Z is the standard normal, i.e., normal$(0, 1)$, variable and $\sigma \geq 0$.

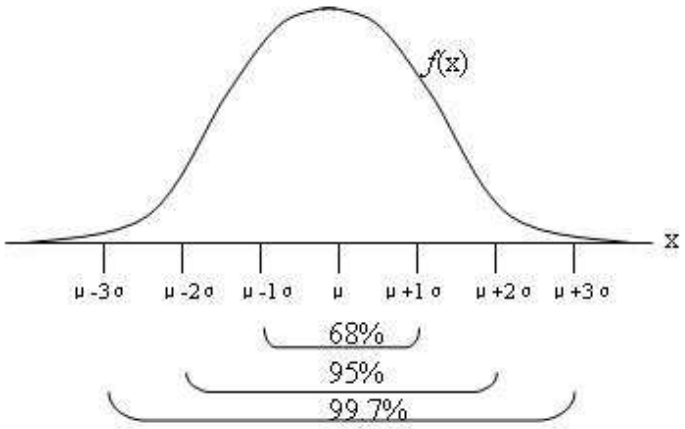

Figure 1: The normal(μ, σ^2) density curve.

2 The Moment-Generating Function

Definition 1 *The moment-generating function (mgf) of a random variable X is $M(t) = E(e^{tX})$ provided the expectation is defined.*

Remark 2 *Note that the expectation, and therefore the moment-generating function, might not exist for all values of t.*

Remark 3 *If $X_1, \cdots, X_n$ are independent random variables with moment generating functions $M_{X_1}, \cdots, M_{X_n}$, then $X_1 + \cdots + X_n$ has moment generating function given by*

$$M_{(X_1+\cdots+X_n)}(t) = \prod_{i=1}^{n} M_{X_i}(t). \tag{2}$$

Theorem 4 *If the mgf exists for t in an open interval containing zero, it uniquely determines the probability distribution.*

Theorem 5 *If the mgf exists in an open interval containing zero, then $M^{(r)}(0) = E(X^r)$ where $M^{(r)}(0)$ is the r^{th} derivative of M at 0.*

The advantage of Theorem (5) is that when the moment of a variable (which involves integration) is difficult to calculate, we can differentiate the mgf to achieve the same result, and differentiation is just mechanical.

Example 6 *(Gamma Distribution) The gamma(α, λ) density function depends on two parameters, $\alpha > 0$ and $\lambda > 0$, and has density function*

$$g_{\alpha,\lambda}(t) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\lambda t}, & t \geq 0 \\ 0 & t < 0 \end{cases} \quad \text{where } \Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du, \ x > 0$$

Remark 7 *It follows by integration by parts that, for all $p > 0$, $\Gamma(p) = p\Gamma(p)$ and that $\Gamma(k) = (k-1)!$ for positive integers k.*

Remark 8 *If $\alpha = 1$, the gamma density coincides with the exponential density. The parameter α is called a shape parameter for the gamma density, and λ is called a scale parameter. Varying α changes the shape of the density, whereas varying λ changes the scale of the density.*

We will find $E(X)$ and $\text{Var}(X)$ for a gamma variable X. The mgf of a gamma distribution is

$$\begin{aligned} M(t) &= \int_0^\infty e^{tx} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{(t-\lambda)x} dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \left(\frac{\Gamma(\alpha)}{(\lambda - t)^\alpha} \right) \quad \text{since } \int_0^\infty x^{\alpha-1} e^{(t-\lambda)x} dx \text{ converages for } t < \lambda \\ &\quad \text{and can be calculated by relating it to the gamma density with } \alpha \text{ and } \lambda - t \\ &= \left(\frac{\lambda}{\lambda - t} \right)^\alpha \end{aligned} \tag{3}$$

Therefore,

$$EX = M^{(1)}(0) = \frac{\alpha}{\lambda}$$

$$EX^2 = M^{(2)}(0) = \frac{\alpha(\alpha+1)}{\lambda^2}$$

and

$$\begin{aligned}\text{Var}(X) &= EX^2 - [EX]^2 \\ &= \frac{\alpha(\alpha+1)}{\lambda^2} - \frac{\alpha^2}{\lambda^2} \\ &= \frac{\alpha}{\lambda^2}.\end{aligned}$$

□

Example 9 *Suppose X has a gamma(α_1, λ) distribution , and independently Y has a gamma(α_2, λ) distribution. By (2), the mgf of $X + Y$ is*

$$\left(\frac{\lambda}{\lambda - t}\right)^{\alpha_1} \left(\frac{\lambda}{\lambda - t}\right)^{\alpha_2} = \left(\frac{\lambda}{\lambda - t}\right)^{\alpha_1 + \alpha_2}, \ t < \lambda.$$

Since $\left(\frac{\lambda}{\lambda - t}\right)^{\alpha_1+\alpha_2}$ is the mgf of a gamma distribution with parameters $\alpha_1 + \alpha_2$ and λ, we see that the sum of n independent exponential(λ) random variables–since exponential(λ) is the special case gamma$(1, \lambda)$–follows a gamma distribution with parameters n and λ. Thus, the time between n consecutive events of a Poisson process follows a gamma distribution. □

3 Joint Distribution

3.1 For Random Variables

The table below summarizes the formulae for calculating some quantities related to joint distributions. However, the table only serves as an outline. When doing actual calculations, one should, instead of relying on the formulae, use reasonings and the basic conditional probability concepts we developed in the first lecture. The examples below demonstrate some of these skills.

	for discrete variables X and Y	for continuous variables X and Y
Probability on a set B	$P((X,Y) \in B) = \sum_{(x,y)\in B} P(x,y)$	$P((X,Y) \in B) = \int\int_B f(x,y)dxdy$
Marginals	$P(X = x) = \sum_{\text{all } y} P(x,y)$	$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy$
	$P(Y = y) = \sum_{\text{all } x} P(x,y)$	$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx$
For indepdent X, Y	$P(x,y) = P(X = x)P(Y = y)$	$f(x,y) = f_X(x)f_Y(y)$
Expectation of $g((X,Y))$	$E(g(X,Y)) = \sum_{all\ x}\sum_{all\ y} g(x,y)P(x,y)$	$E(g(X,Y)) = \int\int g(x,y)f(x,y)dxdy$

Table 1: Table for Joint Distribution formulae

3.1.1 For Independent Variables

Example 10 *Let $X_1, X_2, ..., X_n$ be a collection of independent random variables with cdf $F_1, F_2, ..., F_n$, respectively. The cdf of either the maximum, or the minimum of the $X's$ can be found easily as the following.*

$$\begin{aligned}F_{\max}(x) &= P(X_{\max} \le x) \\ &= P(X_1 \le x, X_2 \le x, ..., X_n \le x) \\ &= P(X_1 \le x)P(X_2 \le x)...P(X_n \le x) \text{ since } X_1, X_2, ..., X_n \text{ are independent} \\ &= F_1(x)F_2(x)...F_n(x)\end{aligned}$$

and

$$\begin{aligned} F_{\min}(x) &= P(X_{\min} \le x) \\ &= 1 - P(X_{\min} > x) \\ &= 1 - P(X_1 > x, X_2 > x, ..., X_n > x) \\ &= 1 - [1 - F_1(x)][1 - F_2(x)]...[1 - F_n(x)]. \end{aligned}$$

Example 11 *Minimum of independent exponential variable is exponential.*

Let $X_1, X_2, ..., X_n$ be independent random variables, and X_i has exponential distribution with rate λ_i, $i = 1, 2, ..., n$. Find the distribution of $X_{\min}$.

For $i = 1, 2, ..., n$, the cdf of X_i is

$$F_i(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-\lambda_i x} & \text{if } x \ge 0 \end{cases}$$

Since the $X's$ are non-negative, so is their minimum. So $X_{\min}$ has cdf $F_{\min}(x) = 0$ for $x < 0$. For $x \ge 0$,

$$\begin{aligned} F_{\min}(x) &= 1 - e^{-\lambda_1 x} e^{-\lambda_2 x} ... e^{-\lambda_n x} \\ &= 1 - e^{-(\lambda_1 + \lambda_2 + ... + \lambda_n)x} \end{aligned}$$

which is the cdf of the exponential distribution with rate $\lambda_1 + \lambda_2 + ... + \lambda_n$. □

Example 12 *Suppose X and Y are independent uniform $(0, 1)$ random variables.*

a)

$$P(X^2 + Y^2 \le 1) = \frac{\pi}{4}$$

b)

$$\begin{aligned} P(X^2 + Y^2 \le 1 | X + Y \ge 1) &= \frac{P(X^2 + Y^2 \le 1, X + Y \ge 1)}{P(X + Y \ge 1)} \\ &= \frac{\pi/4 - 1/2}{1/2} \\ &= \pi/2 - 1 \end{aligned}$$

c)

$$P(Y \le X^2) = \int_0^1 x^2 dx = \frac{1}{3} x^3 |_0^1 = \frac{1}{3}$$

d)

$$P(|X - Y| \le 0.5) = 1 - 1/4 = 3/4$$

e)

$$P(|\frac{X}{Y} - 1| \le 0.5) = P(\frac{2}{3} X \le Y \le 2X) = 1 - \frac{1}{2}(\frac{1}{2} + \frac{2}{3}) = \frac{5}{12}$$

f)

$$P(Y \ge X | Y \ge \frac{1}{2}) = (\frac{1}{2} - \frac{1}{8}) / \frac{1}{2} = \frac{3}{4}.$$

□

Example 13 *let X and Y be independent exponentially distributed random variables with parameter λ and μ, respectively. Find $P(X < Y)$.*

Since X and Y are independent,

$$f(x,y) = (\lambda e^{-\lambda x})(\mu e^{-\mu y}) = \lambda\mu e^{-\lambda x - \mu y}.$$

Then,

$$\begin{aligned} P(X < Y) &= \int\int_{x<y} \lambda\mu e^{-\lambda x-\mu y} dx\ dy \\ &= \int_{x=0}^{\infty} dx \int_{y=x}^{\infty} \lambda\mu e^{-\lambda x-\mu y} dy \\ &= \int_{x=0}^{\infty} \lambda e^{-\lambda x-\mu y} dx \\ &= \frac{\lambda}{\lambda+\mu} \end{aligned}$$

□

Exercise 14 *Suppose $U_{(1)} < U_{(2)} < ... < U_{(5)}$ are the order statistics of 5 independent uniform $(0,1)$ variable $U_1, U_2, ..., U_5$, so $U_{(i)}$ is the i^{th} smallest of $U_1, U_2, ..., U_5$.(See Pitman[1993] P352, example 3)*

a) Find the joint density of $U_{(2)}$ and $U_{(4)}$.
b) Find $P(U_{(2)} > 1/4$ and $U_{(4)} > 1/2)$.

Independent Normal Variables

Theorem 15 *Linear combination of independent normal variables are always normally distributed. In addition, if X and Y are independent with normal(λ, σ^2) and normal(μ, τ^2) distributions, then $X + Y$ has normal $(\lambda + \mu, \sigma^2 + \tau^2)$ distribution.*

The proof of the theorem makes use of the rotational symmetry of the joint distribution of independent standard normal random variable X and Y. See Pitman [1993].

Example 16 *For $\sigma = 1, 2, 3$ suppose $X_\sigma$has normal $(0, \sigma^2)$ distribution, and these three random variables are independent.*

a) Find $P(X_1 + X_2 + X_3 < 4)$.
Let $S = X_1 + X_2 + X_3$. Then S has normal $(0, 1^2 + 2^2 + 3^2)$ distribution, and if $Z = S/\sqrt{14}$ is S standardized, the problem is to find

$$P(S < 4) = P(Z < \frac{4-0}{\sqrt{1^2+2^2+3^2}}) = P(Z < \frac{4}{\sqrt{14}}) \approx 0.857$$

b) Find $P(4X_1 - 10 < X_2 + X_3)$.

$P(4X_1 - 10 < X_2 + X_3) = P(4X_1 - X_2 - X_3 < 10) = P(L < 10)$ where $L = 4X_1 - X_2 - X_3$.
Then, L has normal distribution with mean 0 and variance $4^2 \times 1^2 + (-1)^2 \times 2^2 + (-1)^2 \times 3^2 = 29$, the probability is
$P(L < 10) = P(Z < \frac{10}{\sqrt{29}}) \approx 0.968$. □

Remark 17 *If X and Y are independent with density functions $f_X(x)$ and $f_Y(y)$ in the plane $\mathbf{R}^2$, then, a formula for the joint density function $f_{X+Y}(z)$, where $Z = X + Y$, is*

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x) dx.$$

This is the *convolution formula.*

Exercise 18 *Suppose that X and Y are independent and normally distributed with mean 0 and variance 1. Find the distribution of $\frac{X}{Y}$. (See discussion and solution in Pitman[1993].)*

$\chi^2, t,$ and F Distribution

Definition 19 *If Z is a standard normal random variable, the distribution of $U = \sum_{i=1}^{n} Z_i^2$ is called the chi-square distribution with n degree of freedom, denoted χ_n^2.*

It can be shown that χ_1^2 is a special case of the gamma distribution with parameters $\frac{1}{2}$ and $\frac{1}{2}$. In example 9, we see that the sum of independent gamma random variables sharing the same value of λ follows a gamma distribution. Thus, χ_n^2 is a gamma distribution with $\alpha = \frac{n}{2}$ and $\lambda = \frac{1}{2}$.

Definition 20 *If $Z \sim N(0,1)$ and $U \sim \chi_n^2$ and Z and U are independent, then the distribution of $\frac{Z}{\sqrt{U/n}}$ is called the t distribution with n degrees of freedom.*

The density function of the t distribution with n degrees of freedom is

$$f(t) = \frac{\Gamma[(n+1)/2]}{\sqrt{n\pi}\Gamma(n/2)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}$$

which can be obtained by using a method similar with the exercise (18) above for the density of a quotient of two independent variables.

The shape of the density function for t distribution is very similar with that for the normal distribution, except that the curve for the t distribution has heavier tails on two sides. However, as n increases the density curve for the t distribution gets closer and closer to that for the normal. When $n = 30$, the density curves for the t and the normal distribution are almost indistinguishable.

Theorem 21 *Show that if $X_1, X_2, ...$ are independent $N(\mu, \sigma^2)$ variables, then $\overline{X}$ and S^2 are independent, and $\overline{X}$ is $N(\mu, \sigma^2/n)$ and $(n-1)S^2/\sigma^2$ is χ_{n-1}^2, where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1}(X_i - \overline{X})^2$.*

Proof. (From Rice[1995]) The proof of the statement is built on the fact that $\overline{X}$ and the vector of random variables $(X_1 - \overline{X}, X_2 - \overline{X}, ..., X_n - \overline{X})$ are independent. We will not prove this fact here; the interested readers are referred to Rice [1995] for a treatment using the moment-generating function.

■

Since S^2 is a function of $(X_1 - \overline{X}, X_2 - \overline{X}, ..., X_n - \overline{X})$, and functions of independent vectors are also independent, we can conclude that $\overline{X}$ and S^2 are independent.

Since $X_1, X_2, ...$ are independent $N(\mu, \sigma^2)$, an exdension of theorem (15) shows that $\sum_{i=1}^{n} X_i$ is $N(n\mu, n\sigma^2)$. Thus, deviding a constant n, makes $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ a normal variable with mean $E(\overline{X}) = \frac{n\mu}{n} = \mu$ and variance $\mathrm{Var}(\overline{X}) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$. In addition, $\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2$ follows χ_1^2 by definition (19).

To see that $(n-1)S^2/\sigma^2$ is χ_{n-1}^2, note that

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2=\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2\sim\chi^2_n, \text{ since } \frac{X_i-\mu}{\sigma}\sim N(0,1)$$

and

$$\begin{aligned}\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2&=\frac{1}{\sigma^2}\sum_{i=1}^{n}[(X_i-\overline{X})+(\overline{X}-\mu)]^2\\&=\frac{1}{\sigma^2}\left\{\sum_{i=1}^{n}(X_i-\overline{X})^2+2\sum_{i=1}^{n}(X_i-\overline{X})(\overline{X}-\mu)+\sum_{i=1}^{n}(\overline{X}-\mu)^2\right\}\\&=\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2+\sum_{i=1}^{n}\left(\frac{\overline{X}-\mu}{\sigma}\right)^2\\&=\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2+\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2\end{aligned}\tag{4}$$

Let $W=\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$, $U=\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\overline{X})^2$ and $V=\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2$, (4) says that $W=U+V$. Since U is a function of $(X_1-\overline{X},X_2-\overline{X},...,X_n-\overline{X})$, and V is a function of $\overline{X}$, U and V are independent by the fact we mentioned at the beginning of the proof.

So far, we have showed that $W\sim\chi^2_n$, and $V\sim\chi^2_1$. Let $M_W(t)$ be the mgf for W and so on. Then,

$$\begin{aligned}M_U(t)&=\frac{M_W(t)}{M_V(t)}\\&=\frac{(1-2t)^{-n/2}}{(1-2t)^{-1/2}}\\&=(1-2t)^{-(n-1)/2}\end{aligned}$$

which is the mgf of a random variable with a χ^2_{n-1} distribution.

□

Definition 22 *Let U and V be independent chi-square random variables with m and n degrees of freedom, respectively. The distribution of*

$$W=\frac{U/m}{V/n}$$

is called the F distribution with m and n degree of freedom and is denoted by $F_{m,n}$.

3.1.2 For Dependent Variables

Example 23 *(Gamma and uniform) Suppose X has gamma $(2,\lambda)$ distribution, and that given $X=x, Y$ has uniform $(0,x)$ distribution. Find the joint density of X and Y.*

By the definition of the gamma distribution

$$f_X(x) = \begin{cases} \lambda^2 x e^{-\lambda x}, & x > 0 \\ 0, & x \le 0 \end{cases}$$

and from the uniform $(0, x)$ distribution of Y given $X = x$

$$f_Y(y|X = x) = \begin{cases} 1/x, & 0 < y < x \\ 0, & \text{otherwise} \end{cases}$$

So by the multiplication for densities

$$f(x, y) = f_X(x) f_Y(y|X = x) = \begin{cases} \lambda^2 e^{-\lambda x}, & 0 < y < x \\ 0, & \text{otherwise} \end{cases}$$

Example 24 *Find the marginal density of Y.*

Integrating out x in the joint density gives the marginal density of Y: for $y > 0$

$$f_Y(y) = \int_0^\infty f(x, y)dx = \int_y^\infty \lambda^2 e^{-\lambda x} dx = \lambda e^{-\lambda y}$$

The density is of course 0 for $y \le 0$. That is to say, Y has exponential (λ) distribution.

	Disceret Case	Continuous Case
Multiplication rule	$P(X = x, Y = y) = P(X = x)P(Y = y\|X = x)$	$f(x, y) = f_X(x) f_Y(y\|X = x)$
Cond. dist. of $(Y\|X = x)$	$P(Y \in B\|X = x) = \sum_{y \in B} P(Y = y\|X = x)$	$P(Y \in B\|X = x) = \int_B f_Y(y\|X = x)dy$
Average cond. expectation	$E(Y) = \sum_{\text{all } x} E(Y\|X = x)P(X = x)$	$E(Y) = \int E(Y\|X = x) f_X(x)dx$

Table 2: Table for Conditioning formulae

4 Conditioning by a Random Vector

4.1 Discrete Case

If $\mathbf{Y}$ and $\mathbf{Z}$ are discrete random vectors possibly of different dimensions, we want to study the conditional probability structure of $\mathbf{Y}$ given that $\mathbf{Z}$ has taken on a particular value $\mathbf{z}$.

Definition 25 *Define the conditional probability mass function $p(\cdot \mid \mathbf{z})$ of $\mathbf{Y}$ given $\mathbf{Z} = \mathbf{z}$ by*

$$p(\mathbf{y} \mid \mathbf{z}) = P[\mathbf{Y} = \mathbf{y} \mid \mathbf{Z} = \mathbf{z}] = \frac{p(\mathbf{y}, \mathbf{z})}{p_{\mathbf{Z}}(\mathbf{z})} \tag{5}$$

where p and $p_{\mathbf{Z}}$ are the probability mass functions of $(\mathbf{Y}, \mathbf{Z})$ and $\mathbf{Z}$. The conditional probability mass function p is defined only for values of $\mathbf{z}$ such that $p_{\mathbf{Z}}(\mathbf{z}) > 0$. With this definition it is clear that $p(\cdot \mid \mathbf{z})$ is the mass function of a probability distribution because

$$\sum_{\mathbf{y}} f(\mathbf{y} \mid \mathbf{z}) = \frac{\sum_{\mathbf{y}} p(\mathbf{y}, \mathbf{z})}{p_{\mathbf{Z}}(\mathbf{z})} = \frac{p_{\mathbf{Z}}(\mathbf{z})}{p_{\mathbf{Z}}(\mathbf{z})} = 1$$

This probability distribution is called the conditional distribution of $\mathbf{Y}$ given that $\mathbf{Z} = \mathbf{z}$.

Example 26 *Let $\mathbf{Y} = (Y_1, \cdots, Y_n)$, where the Y_i are the indicators of a set of n Bernoulli trials with success probability p. Let $Z = \sum_{i=1}^n Y_i$, the total number of successes. Then $\mathbf{Z}$ has a binomial, $\mathcal{B}(n, p)$, distribution and*

$$p(\mathbf{y} \mid \mathbf{z}) = \frac{P[Y = \mathbf{y}, Z = \mathbf{z}]}{\binom{n}{z} p^z (1-p)^{n-z}} = \frac{p^y (1-p)^{n-z}}{\binom{n}{z} p^z (1-p)^{n-z}} = \frac{1}{\binom{n}{z}}.$$

Thus, if we are told we obtained k successes in n binomial trials, then these successes are as likely to occur on one set of trials as on any other. □

4.1.1 Bayes' Rule

Let $q(\mathbf{z} \mid \mathbf{y})$ denote the conditional probability mass function of $\mathbf{Z}$ give $\mathbf{Y} = \mathbf{y}$. Then,

$$p(\mathbf{y}, \mathbf{z}) = p(\mathbf{y} \mid \mathbf{z}) f_{\mathbf{Z}}(\mathbf{z})$$

$$p(\mathbf{y} \mid \mathbf{z}) = \frac{q(\mathbf{z} \mid \mathbf{y}) p_{\mathbf{Y}}(\mathbf{y})}{\sum_{\mathbf{y}} q(\mathbf{z} \mid \mathbf{y}) p_{\mathbf{Y}}(\mathbf{y})} \text{ Bayes' Rule}$$

whenever the denominator of the right-hand side is positive.

4.1.2 Conditional Expectation for Discrete Variables

Definition 27 *Suppose that Y is a random variable with $E(|Y|) < \infty$. Define the conditional expectation of Y given $\mathbf{Z} = \mathbf{z}$, written $E(Y \mid \mathbf{Z} = \mathbf{z})$, by*

$$E(Y \mid \mathbf{Z} = \mathbf{z}) = \sum_{\mathbf{y}} y p(y \mid \mathbf{z}).$$

Note that if $p_{\mathbf{Z}}(\mathbf{z}) > 0$,

$$E(|Y| \ \ |\mathbf{Z} = \mathbf{z}|) = \sum_{\mathbf{y}} |y| \, p_Y(y|z) \leq \sum_{\mathbf{y}} |y| \, \frac{p_Y(y)}{p_{\mathbf{Z}}(\mathbf{z})} = \frac{E\left(|Y|\right)}{p_{\mathbf{Z}}(\mathbf{z})}.$$

The inequality is because that $\{y \cap \mathbf{z}\} \subseteq \{y\}$. Thus, when $p_{\mathbf{Z}}(\mathbf{z}) > 0$, the conditional expected value of Y is finite whenever the expected value is finite.

Definition 28 *Let $g(z) = E(Y \mid \mathbf{Z} = \mathbf{z})$. The random variable $g(\mathbf{Z})$ is written $E(Y \mid \mathbf{Z})$ and is called the conditional expectation of Y given $\mathbf{Z}$.*

Example 29 *As an example we calculate $E(Y_1 \mid Z)$ where Y_1 and $\mathbf{Z}$ are given in Example 26. We have*

$$E(Y_1 \mid Z = i) = P[Y_1 = 1 \mid Z = i] = \frac{\binom{n-1}{i-1}}{\binom{n}{i}} = \frac{i}{n}.$$

The first of these equalities holds because Y_1 is an indicator. The second follows from the equation in Example 26 because $\binom{n-1}{i-1}$ is just the number of ways i successes can occur in n Bernoulli trials with the first trial being a success. Therefore,

$$E(Y_1 \mid Z) = \frac{Z}{n}.$$

Exercise 30 *Let X_1 and X_2 be the numbers on two independent fair-die rolls. Let X be the minimum and Y the maximum of X_1 and X_2. Calculate: $E(Y|X = x)$ and $E(X|Y = y)$.*

Exercise 31 *Repeat the last exercise with X_1 and X_2 two draws without replacement from $\{1, 2, ..., n\}$.*

Properties of Conditional Expected Values In the context of previous lecture, the conditional distribution of a random vector $\mathbf{Y}$ given $\mathbf{Z} = \mathbf{z}$ corresponds to a single probability function $P_{\mathbf{z}}$ on (Ω, A). Specifically, define for $A \in \mathcal{A}$,

$$P_{\mathbf{z}}(A) = P(A \mid [\mathbf{Z} = \mathbf{z}]) \text{ if } p_{\mathbf{Z}}(\mathbf{z}) > 0.$$

This $P_{\mathbf{z}}$ is just the conditional probability function on (Ω, A) mentioned before. Now the conditional distribution of $\mathbf{Y}$ given $\mathbf{Z} = \mathbf{z}$ is the same as the distribution of $\mathbf{Y}$ if $P_{\mathbf{z}}$ is the probability function on (Ω, A). Therefore, the conditional expectation is an ordinary expectation with respect to the probability function $P_{\mathbf{z}}$.

Properties. It follows that all the properties of the expectation given before hold for the conditional expectation given $\mathbf{Z} = \mathbf{z}$. Thus, for any real-valued function $r(\mathbf{Y})$ with $E\,|r(\mathbf{Y})| < \infty$,

1. $E(r(\mathbf{Y}) \mid \mathbf{Z} = \mathbf{z}) = \sum_y r(\mathbf{y})p(\mathbf{y} \mid \mathbf{z})$ identically in $\mathbf{z}$ for any Y_1, Y_2 such that $E(|Y_1|),\ E(|Y_2|)$ are finite.
2. $E(\alpha Y_1 + \beta Y_2 \mid \mathbf{Z}) = \alpha E(Y_1 \mid \mathbf{Z}) + \beta E(Y_2 \mid \mathbf{Z})$, since $E(\alpha Y_1 + \beta Y_2 \mid \mathbf{Z} = \mathbf{z}) = \alpha E(Y_1 \mid \mathbf{Z} = \mathbf{z}) + \beta E(Y_2 \mid \mathbf{Z} = \mathbf{z})$ holds for all $\mathbf{z}$.
3. $E(Y \mid \mathbf{Z}) = E(Y)$ if Y and $\mathbf{Z}$ are independent.
4. $E(h(\mathbf{Z}) \mid \mathbf{Z}) = h(\mathbf{Z})$.
5. $E(q(\mathbf{Y}, \mathbf{Z}) \mid \mathbf{Z} = \mathbf{z}) = E(q(\mathbf{Y}, \mathbf{z}) \mid \mathbf{Z} = \mathbf{z})$ (subsitution theorem for conditional expectation)
6. $E(E(Y \mid \mathbf{Z})) = E(Y)$

Property 6 is true because

$$E(E(Y \mid \mathbf{Z})) = \sum_{\mathbf{z}} P_{\mathbf{Z}}(\mathbf{z})[\sum_y yp(y \mid \mathbf{z})] = \sum_{y,\mathbf{z}} yp(y \mid \mathbf{z})p_{\mathbf{Z}}(\mathbf{z}) = \sum_{y,\mathbf{z}} yp(y, \mathbf{z}) = E(Y).$$

The interchange of summation used is valid because the finiteness of $E(|Y|)$ implies that all sums converge absolutely.

As an illustration, we check $E(E(Y \mid \mathbf{Z})) = E(Y)$ for $E(Y_1 \mid \mathbf{Z}) = \frac{Z}{n}$ given before. In this case,

$$E(E(Y_1 \mid \mathbf{Z})) = E\binom{Z}{n} = \frac{np}{n} = p = E(Y_1)$$

4.2 Continuous Case

Definition 32 *Suppose $(\mathbf{Y}, \mathbf{Z})$ is a continuous random vector having coordinates that are themselves vectors and having density function $p(\mathbf{y}, \mathbf{z})$. In analogy to (5), the conditional density function of $\mathbf{Y}$ given $\mathbf{Z} = \mathbf{z}$ is*

$$p(\mathbf{y} \mid \mathbf{z}) = \frac{p(\mathbf{y}, \mathbf{z})}{p_{\mathbf{Z}}(\mathbf{z})}$$

if $p_{\mathbf{Z}}(\mathbf{z}) > 0$.

4.2.1 Bayes' Rule

The Bayes' rule for the continuous case is defined as

$$p(\mathbf{y} \mid \mathbf{z}) = \frac{p_{\mathbf{Y}}(\mathbf{y})q(\mathbf{z} \mid \mathbf{y})}{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p_{\mathbf{Y}}(\mathbf{t})q(\mathbf{z} \mid \mathbf{t})dt_1 \cdots dt_n},$$

where q is the conditional density of $\mathbf{Z}$ given $\mathbf{Y} = \mathbf{y}$.

Remark 33 *If $\mathbf{Y}$ and $\mathbf{Z}$ are independent, the conditional distributions equal the marginals as in the discrete case.*

Remark 34 *If $E(|Y|) < \infty$, we denote the conditional expectation of Y given $\mathbf{Z} = \mathbf{z}$ in analogy to the discrete case as the expected value of a random variable with density $p(y \mid \mathbf{z})$. More generally, if $E(|r(\mathbf{Y})|) < \infty$, the conditional expectation of $r(\mathbf{Y})$ given $\mathbf{Z} = \mathbf{z}$ can be obtained from*

$$E(r(\mathbf{Y}) \mid \mathbf{Z} = \mathbf{z}) = \int_{-\infty}^{\infty} r(\mathbf{y})p(\mathbf{y} \mid \mathbf{z})d\mathbf{y}.$$

5 Transformation of a Random Vector

We have encountered the change of variable formula for the case involving random variables. In this section, we will see a more general case: transformation of a random vector.

Let $\mathbf{h} = (h_1, ...h_k)^T$, where each h_i is a real-valued function on R^k. Thus, $\mathbf{h}$ is a transformation from R^k to R^k. Recall that the *Jacobian* $J_h(t)$ of h evaluated at $\mathbf{t} = (t_1, \cdots, t_k)^T$ is by definition the determinant

$$J_h(\mathbf{t}) = \begin{vmatrix} \frac{\partial}{\partial t_1}h_1(\mathbf{t}) & \cdots & \frac{\partial}{\partial t_1}h_k(\mathbf{t}) \\ \vdots & & \vdots \\ \frac{\partial}{\partial t_k}h_1(\mathbf{t}) & \cdots & \frac{\partial}{\partial t_k}h_k(\mathbf{t}) \end{vmatrix}.$$

Theorem 35 *Let $\mathbf{X}$ be continuous and let S be an open subset of R^k such that $P(\mathbf{X} \in S) = 1$. If $\mathbf{g} = (g_1, \cdots, g_k)^T$ is a transformation from S to R^k such that g and S satisfy the conditions:*

1. $\mathbf{g}^{-1}$ has continous first partial derivatives on S.
2. $\mathbf{g}^{-1}$ is one-to-one on S.
3. The Jacobian of h does not vanish on S.

Then, the density of $\mathbf{Y} = \mathbf{g}(\mathbf{Y})$ is given by

$$p_{\mathbf{Y}}(\mathbf{y}) = p_{\mathbf{X}}(g^{-1}(\mathbf{y}))\left|J_{\mathbf{g}^{-1}}(\mathbf{y})\right| \tag{6}$$

for $\mathbf{y} \in \mathbf{g}(S)$..

Example 36 *Suppose $\mathbf{X} = (X_1, X_2)^T$ where X_1 and X_2 are independent with $\mathcal{N}(0, 1)$ and $\mathcal{N}(0, 4)$ distributions, respectively. What is the joint distribution of $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$? Here,*

$$p_{\mathbf{X}}(x_1, x_2) = \frac{1}{4\pi} \exp\left(-\frac{1}{2}\left[x_1^2 + \frac{1}{4}x_2^2\right]\right).$$

In this case, $S = R^2$. Also note that $g_1(\mathbf{x}) = x_1 + x_2, g_2(\mathbf{x}) = x_1 - x_2, g_1^{-1}(\mathbf{y}) = \frac{1}{2}(y_1 + y_2), g_2^{-1}(\mathbf{y}) = \frac{1}{2}(y_1 - y_2)$, that the range $\mathbf{g}(S)$ is R^2 and that

$$J_{\mathbf{g}^{-1}}(\mathbf{y}) = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = -\frac{1}{2}.$$

Upon substituting these quantities in (6), we obtain

$$
\begin{aligned}
p_{\mathbf{Y}}(y_1, y_2) &= \frac{1}{2} p_{\mathbf{X}}\left(\frac{1}{2}(y_1 + y_2), \frac{1}{2}(y_1 - y_2)\right) \\
&= \frac{1}{8\pi} \exp - \left[\frac{1}{4}(y_1 + y_2)^2 + \frac{1}{16}(y_1 - y_2)^2\right] \\
&= \frac{1}{8\pi} \exp - \frac{1}{32}\left[5y_1^2 + 5y_2^2 + 6y_1 y_2\right].
\end{aligned}
$$

This is an example of bivariate normal density.

Gamma and Beta Distribution A random variable X has Beta(r, s) distribution if it has density

$$
b_{r,s}(x) = \frac{x^{r-1}(1-x)^{s-1}}{B(r,s)}, \text{ for } 0 < x < 1,
$$

where $B(r,s) = \frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)}$ is the beta function and $\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du$ as in the gamma distribution.

Example 37 *If X_1 and X_2 are independent random variables with gamma(p, λ) and gamma(q, λ) distributions, respectively, then $Y_1 = X_1 + X_2$ and $Y_2 = \frac{X_1}{(X_1+X_2)}$ are independent and have, respectively, gamma$(p+q, \lambda)$ and Beta(p, q) distribution.*

If $\lambda = 1$, the joint density of X_1 and X_2 is

$$
p(x_1, x_2) = \frac{e^{-(x_1+x_2)} x_1^{p-1} x_2^{q-1}}{\Gamma(p)\Gamma(q)}, \text{ for } x_1 > 0,\ x_2 > 0.
$$

Let

$$
\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \mathbf{g}(y_1, y_2) = \begin{pmatrix} x_1 + x_2 \\ \frac{x_1}{x_1+x_2} \end{pmatrix}.
$$

Then, $\mathbf{g}$ is one-to-one on $S = \{\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} : x_1 > 0, x_2 > 0\}$ and its range is $S_1 = \{\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} : y_1 > 0, 0 < y_2 < 1\}$. We note that on S_1

$$
\mathbf{g}^{-1}(y_1, y_2) = \begin{pmatrix} y_1 y_2 \\ y_1 - y_1 y_2 \end{pmatrix} \tag{7}
$$

Therefore,

$$
J_{\mathbf{g}^{-1}}(y_1, y_2) = \begin{vmatrix} y_2 & 1 - y_2 \\ y_1 & -y_1 \end{vmatrix} = -y_1 \tag{8}
$$

Subsititute (7) and (8) into (6) and get for the density of $\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \mathbf{g}(X_1, X_2)$,

$$
p_{\mathbf{Y}}(y_1, y_2) = \frac{e^{-y_1}(y_1 y_2)^{p-1}(y_1 - y_1 y_2)^{q-1} y_1}{\Gamma(p)\Gamma(q)}, \text{ for } y_1 > 0, 0 < y_2 < 1. \tag{9}
$$

Simplifying (9) gives

$$
p_{\mathbf{Y}}(y_1, y_2) = g_{p+q,1}(y_1) b_{p,q}(y_2).
$$

Thus, the statement is proved for $\lambda = 1$. If $\lambda \neq 1$, define $X_1' = \lambda X_1$ and $X_2' = \lambda X_2$. Now X_1' and X_2' are independent $\Gamma(p, 1)$, $\Gamma(q, 1)$ variables respectively. Because $X_1' + X_2' = \lambda(X_1 + X_2)$ and $X_1'(X_1' + X_2')^{-1} = X_1(X_1 + X_2)^{-1}$, the statement follows. □

6 Markov Chain

6.1 Discrete case

In this section, we consider a stochastic process $\{X_n, n = 0, 1, 2, ...\}$ that takes on a finite or countable number of possible values. Unless otherwise mentioned, this set of possible values of process will be denoted by the set of non-negative integers $\{0, 1, 2, ...\}$. If $X_n = i$, then the process is said to be in state i at time n. We suppose that whenever the process is in state i, there is a fixed probability P_{ij} that it will next be at state j. That is, we suppose that

$$P\{X_{n+1} = j|X_n = i, X_{n-1} = i_{n-1}, ..., X_1 = i_1, X_0 = i_0\} = P_{ij}$$

for all states $i_0, i_1, ..., i_{n-1}, i, j$ and all $n \geq 0$. Such a stochastic process is known as a Markov Chain.

Example 38 *(A random walk model) A Markov chain whose state space is given by the integers $i = 0, \pm1, \pm2, ...$ is said to be a random walk if , for some number $0 < p < 1$,*

$$P_{i,i+1} = p = 1 - P_{i,i-1} \ ; \ i = 0, \pm1, \pm2, ...$$

The preceding Markov chain is called a random walk for we may think of it as being a model for an individual walking on a straight line who at each point of time either takes one step to the right with probability p or one step to the left with probability $1 - p$.

Chapman-Kolmogorov Equations

We have defined the one-step transition probability $P_{ij.}$ We now define the n-step transition probabilities P_{ij}^n to be the probability that a process in state i will be in state j after n additional transitions. That is,

$$P_{ij}^n = P\{X_{n+k} = j|X_k = i\}, \ n \geq 0, i, j \geq 0$$

The Chapman-Kolmogorov equations provide a method for computing these n-step transition probabilities. These equations are

$$P_{ij}^{n+m} = \sum_{k=0}^{\infty} P_{ik}^n P_{kj}^m \text{ for all } n, m \geq 0, \text{all } i, j$$

and are most easily understood by noting that $P_{ik}^n P_{kj}^m$ represents the probability that starting in i the process will go to state j in $n + m$ transitions through a path which takes it into state k at the n^{th} transition. Hence, summing over all intermediate states k yields the probability that the process will be in state j after $n + m$ transitions. Formally, we have

$$\begin{aligned} P_{ij}^{n+m} &= P\{X_{n+m} = j|X_0 = i\} \\ &= \sum_{k=0}^{\infty} P\{X_{n+m} = j, X_n = k|X_0 = i\} \\ &= \sum_{k=0}^{\infty} P\{X_{n+m} = j|X_n = k, X_0 = i\} P\{X_n = k|X_0 = i\} \\ &= \sum_{k=0}^{\infty} P_{ik}^n P_{kj}^m \end{aligned}$$

6.2 Continuous case

Suppose we have a continuous-time stochastic process $\{X(t), t \geq 0\}$ taking on values in the set of non-negative integers. In analogy with the definition of a discrete-time Markov chain, we say that the process $\{X(t), t \geq 0\}$ is a continous-time Markov chain if for all $s, t \geq 0$ and non-negative integers $i, j, x(u), 0 \leq u < s$

$$\begin{aligned} P\{X(t+s) = j | X(s) = i, X(u) = x(u), 0 \leq u < s\} \\ = P\{X(t+s) = j | X(s) = i\} \end{aligned}$$

In other words, a continous-time Markov chain is a stochastic process having the Markovian property that the conditional distribution of the future $X(t+s)$ given the present $X(s)$ and the past $X(u)$, $0 \leq u < s$, depends only on the present and is independent of the past. If in addition,

$$P\{X(t+s) = j | X(s) = i$$

is independent of s, then the continuous-time Markov chain is said to have stationary or homogeneous transition probabilities.

Example 39 *Suppose the service in a particular barbra shop consists of two procedures: A customer upon arrival goes initially to chair 1 where his/her hair will be raised by an assistant; after this is done the customer moves on to chair 2 where his/her hair will be cut by the stylist. The service time at the two steps are assumped to be independent random variable that are exponentially distributed with respective rates μ_1 and μ_2. Suppose that the potential customers arrive in accordance with a Poisson process having rate λ, and that a potential customer will enter the system only if both chairs are empty.*

This problem can be modeled as a continuous-time Markov chain. Since a potential customer will enter the shop only if there is no other customers in the shop, there will always be either 0, or 1 customers in the shop. If there is 1 customer in the shop, then we would need to know which chair the customer is on. Therefore, an appropriate stat space would consists of three states: 0 =shop is empty, 1 =chair 1 is occupied and 2 =chair 2 is occupied.

7 Delta Method

(The followings are taken from Rice[1995])

Suppose that we know the expectation and the variance of a random variable X, but not the entire distribution and that we are interested in the mean and variance of $Y = g(X)$ for some fixed function g. For example, we might be able to measure X and determine its mean and variance, but really be interested in Y, which is related to X in a known way. We might wish to know $Var(Y)$, at least approximately, in order to assess the accuracy of the indirect measurement process. From the results given in this section. we cannot in general find $E(Y) = \mu_Y$ and $Var(Y) = \sigma_Y^2$ from $E(X) = \mu_X$ and $Var(X) = \sigma_X^2$, unless the function g is linear. However, if g is nearly linear in a range in which X has high probability, it can be approximated by a linear function and approximate moments of Y can be found.

In proceeding as just described, we follow a tack often taken in applied mathematics: When confronted with a nonlinear problem that we cannot solve, we linearize. In probability and statistics, this method is called **propagation of error**, or the δ **method**. Linearization is carried out through a Taylor series expansion of g about μ_X. To the first order,

$$Y = g(X) \approx g(\mu_X) + (X - \mu_X) g'(\mu_X)$$

We have expressed Y as approximately equal to a linear function of X. Recalling that if $U = a+bV$, then $E(U) = a + bE(V)$ and $Var(U) = b^2 Var(V)$, we find

$$\mu_Y \approx g(\mu_X)$$
$$\sigma_Y^2 \approx \sigma_X^2 [g'(\mu_X)]^2$$

We know that in general $E(Y) \neq g(E(X))$, as given by the approximation. In fact, we can carry out the Taylor series expansion to the second order to get an improved approximation of μ_Y

$$Y = g(X) \approx g(\mu_X) + (X - \mu_X)g'(\mu_X) + \frac{1}{2}(X - \mu_X)^2 g''(\mu_X)$$

Taking the expectation of the right-hand side, we have, since $E(X - \mu_X) = 0$,

$$E(Y) \approx g(\mu_X) + \frac{1}{2}\sigma_X^2 g''(\mu_X)$$

How good such approximations are depends on how nonlinear g is in a neighborhood of μ_X and on the size of σ_X. From Chebyshev's inequality we know that X is unlikely to be many standard deviations away from μ_X; if g can be reasonably well approximated in this range by a linear function, the approximations for the moments will be reasonable as well.

Example 40 *The relation of voltage, current, and resistance is $V = IR$. Suppose that the voltage is held constant at a value V_0 across a medium whose resistance fluctuates randomly as a result, say, of random fluctuations at the molecular level. The current therefore also varies randomly. Suppose that it can be determined experimentally to have mean $\mu_I = 0$ and variance σ_I^2. We wish to find the mean and variance of the resistance, R, and since we do not know the distribution of I, we must resort to an approximation. We have*

$$R = g(I) = \frac{V_0}{I}$$
$$g'(\mu_I) = -\frac{V_0}{\mu_I^2}$$
$$g''(\mu_I) = \frac{2V_0}{\mu_I^3}$$

Thus,

$$\mu_R \approx \frac{V_0}{\mu_I} + \frac{V_0}{\mu_I^3}\sigma_I^2$$
$$\sigma_R^2 \approx \frac{V_0^2}{\mu_I^4}\sigma_I^2$$

We see that the variability of R depends on both the mean level of I and the variance of I. This makes sense, since if I is quite small, small variations in I will result in large variations in $R = V_0/I$, whereas if I is large, small variations will not affect R as much. The second-order correction factor for μ_R also depends on μ_I and is large if μ_I is small. In fact, when I is near zero, the function $g(I) = V_0/I$ is quite nonlinear, and the linearization is not a good approximation.

References

[1] Peter J. Bickel and Kjell A. Doksum (2001) Mathematical Statistics: Basic Ideas and Selected Topics, Vol. I, 2nd Edition. *Prentice Hall*

[2] Geoffrey Grimmett and David Stirzaker (2002) Probability and Random Processes, 3rd Edition, *Oxford University Press.*

[3] Jim Pitman (1993) Probability, *Springer-Verlag New York, Inc.*

[4] John A. Rice (1995) Mathematical Statistics and Data Analysis, 2nd Edition, *Duxbury Press.*

[5] Sheldon M. Ross (2003) Introduction to Probability Models, *Academic Press.*