Stat 5101 Lecture Slides: Deck 5

Conditional Probability and Expectation, Poisson Process, Multinomial and Multivariate Normal Distributions

Charles J. Geyer
School of Statistics
University of Minnesota

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License (`http://creativecommons.org/licenses/by-sa/4.0/`).

Joint and Marginal Distributions

When we have two random variables X and Y under discussion, a useful shorthand calls the distribution of the random vector (X, Y) the *joint distribution* and the distributions of the random variables X and Y the *marginal distributions*.

Joint and Marginal Distributions (cont.)

The name comes from imagining the distribution is given by a table

		Y			
		grass	grease	grub	
X	red	1/30	1/15	2/15	7/30
	white	1/15	1/10	1/6	1/3
	blue	1/10	2/15	1/5	13/30
		1/5	3/10	1/2	1

In the center 3×3 table is the joint distribution of the variables X and Y. In the right margin is the marginal distribution of X. In the bottom margin is the marginal distribution of Y.

Joint and Marginal Distributions (cont.)

The rule for finding a marginal is simple.

> *To obtain a marginal PMF/PDF from a joint PMF/PDF, sum or integrate out the variable(s) you don't want.*

For discrete, this is obvious from the definition of the PMF of a random variable.

$$f_X(x) = \Pr(X = x) = \sum_y f_{X,Y}(x, y)$$

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

To obtain the marginal of X sum out y. To obtain the marginal of Y sum out x.

Joint and Marginal Distributions (cont.)

For continuous, this is a bit less obvious, but if we define

$$f_X(x) = \int f_{X,Y}(x, y)\, dy$$

We see that this works when we calculate expectations

$$\begin{aligned} E\{g(X)\} &= \int g(x) f_X(x)\, dx \\ &= \int g(x) \int f_{X,Y}(x, y)\, dy\, dx \\ &= \iint g(x) f_{X,Y}(x, y)\, dy\, dx \end{aligned}$$

The top line is the definition of $E\{g(X)\}$ if we accept f_X as the PDF of X. The bottom line is the definition of $E\{g(X)\}$ if we accept $f_{X,Y}$ as the PDF of (X, Y). They must agree, and do.

Joint and Marginal Distributions (cont.)

Because of non-uniqueness of PDF — we can redefine on a set of probability zero without changing the distribution — we can't say the marginal obtained by this rule is the unique marginal, but it is a valid marginal.

To obtain the marginal of X integrate out y. To obtain the marginal of Y integrate out x.

Joint and Marginal Distributions (cont.)

The word "marginal" is entirely dispensable, which is why we haven't needed to use it up to now.

The term "marginal PDF of X" means exactly the same thing as the the term "PDF of X".

It is the PDF of the random variable X, which may be redefined on sets of probability zero without changing the distribution of X.

"Joint" and "marginal" are just verbal shorthand to distinguish the univariate distributions (marginals) from the bivariate distribution (joint).

Joint and Marginal Distributions (cont.)

When we have three random variables X, Y, and Z under discussion, the situation becomes a bit more confusing.

By summing or integrating out one variable we obtain any of three bivariate marginals $f_{X,Y}$, $f_{X,Z}$, or $f_{Y,Z}$.

By summing or integrating out two variables we obtain any of three univariate marginals f_X, f_Y, or f_Z.

Thus $f_{X,Y}$ can be called either a joint distribution or a marginal distribution depending on context. $f_{X,Y}$ is a marginal of $f_{X,Y,Z}$, but $f_{X,Y}$ is the joint distribution of the random variables X and Y, and f_X and f_Y are marginals of it.

Joint and Marginal Distributions (cont.)

But the rule remains the same

> *To obtain a marginal PMF/PDF from a joint PMF/PDF, sum or integrate out the variable(s) you don't want.*

For example

$$f_{W,X}(w,x) = \iint f_{W,X,Y,Z}(w,x,y,z)\,dy\,dz$$

Write out what you are doing carefully like this. If the equation has the same free variables on both sides (here w and x), and the dummy variables of integration (or summation) do not appear as free variables, then you are trying to do the right thing. Do the integration correctly, and your calculation will be correct.

Joint, Marginal, and Independence

If $X_1, \ldots, X_n$ are IID with PMF/PDF f, then the joint distribution of the random vector $(X_1, \ldots, X_n)$ is

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

In short, the joint is the product of the marginals *when the variables are independent*.

We already knew this. Now we have the shorthand of "joint" and "marginals".

Conditional Probability and Expectation

The conditional probability distribution of Y given X is the probability distribution you should use to describe Y after you have seen X.

It is a probability distribution like any other. It is described in any of the ways we describe probability distributions: PMF, PDF, DF, or by change-of-variable from some other distribution.

The only difference is that the conditional distribution is a function of the observed value of X. Hence its parameters, if any, are functions of X.

Conditional Probability and Expectation (cont.)

So back to the beginning. Nothing we have said in this course tells us anything about this new notion of conditional probability and expectation.

It is yet another generalization. When we went from finite to infinite sample spaces, some things changed, although a lot remained the same. Now we go from (ordinary, unconditional) probability and expectation to conditional probability and expectation, and some things change, although a lot remain the same.

Conditional Probability and Expectation (cont.)

The conditional PMF or PDF of Y given X is written $f(y \mid x)$. It determines the distribution of the variable in front of the bar Y given a value x of the variable behind the bar X.

The function $y \mapsto f(y \mid x)$, that is, $f(y \mid x)$ thought of a a function of y for fixed x, is a PMF or PDF and follows all the rules for such.

Conditional Probability and Expectation (cont.)

In particular,

$$f(y \mid x) \geq 0, \qquad \text{for all } y$$

and in the discrete case

$$\sum_y f(y \mid x) = 1$$

and in the continuous case

$$\int f(y \mid x)\, dy = 1$$

Conditional Probability and Expectation (cont.)

From conditional PMF and PDF we define conditional expectation

$$E\{g(Y) \mid x\} = \int g(y) f(y \mid x)\, dy$$

and conditional probability

$$\begin{aligned} \Pr\{Y \in A \mid x\} &= E\{I_A(Y) \mid x\} \\ &= \int I_A(y) f(y \mid x)\, dy \\ &= \int_A f(y \mid x)\, dy \end{aligned}$$

(with integration replaced by summation if Y is discrete).

Conditional Probability and Expectation (cont.)

The variable behind the bar just goes along for the ride. It is just like a parameter.

In fact this is one way to make up conditional distributions. Compare

$$f_\lambda(x) = \lambda e^{-\lambda x}, \qquad x > 0, \ \lambda > 0$$

and

$$f(y \mid x) = x e^{-xy}, \qquad y > 0, \ x > 0$$

Conditional Probability and Expectation (cont.)

Formally, there is no difference whatsoever between a parametric family of distributions and a conditional distribution. Some people like to write $f(x \mid \lambda)$ instead of $f_\lambda(x)$ to emphasize this fact.

People holding non-formalist philosophies of statistics do see differences. Some, usually called frequentists, although this issue really has nothing to do with infinite sequences and the law of large numbers turned into a definition of expectation, would say there is a big difference between $f(x \mid y)$ and $f_\lambda(x)$ because Y is a random variable and λ is not.

More on this next semester.

Conditional Probability and Expectation (cont.)

Compare. If the distribution of X is Exp(λ), then

$$E_\lambda(X) = \int_0^\infty x f_\lambda(x)\,dx = \int_0^\infty x\lambda e^{-\lambda x}\,dx = \frac{1}{\lambda}$$

If the conditional distribution of Y given X is Exp(x), then

$$E(Y \mid x) = \int_0^\infty y f(y \mid x)\,dy = \int_0^\infty yxe^{-xy}\,dy = \frac{1}{x}$$

Just replace x by y and λ by x *in that order*.

Conditional Probability and Expectation (cont.)

Compare. If the distribution of X is Exp(λ), then

$$\Pr_\lambda(X > a) = \int_a^\infty f_\lambda(x)\,dx = \int_a^\infty \lambda e^{-\lambda x}\,dx = e^{-\lambda a}$$

If the conditional distribution of Y given X is Exp(x), then

$$\Pr(Y > a \mid x) = \int_a^\infty f(y \mid x)\,dy = \int_a^\infty x e^{-xy}\,dy = e^{-xa}$$

Just replace x by y and λ by x *in that order*.

Conditional Probability and Expectation (cont.)

Compare. If the PDF of X is

$$f_\theta(x) = \frac{x + \theta}{1/2 + \theta}, \qquad 0 < x < 1$$

then

$$E_\theta(X) = \int_0^1 x f_\theta(x)\, dx = \int_0^1 \frac{x(x+\theta)}{1/2+\theta}\, dx = \frac{2+3\theta}{3+6\theta}$$

If the conditional PDF of Y given X is

$$f(y \mid x) = \frac{y + x}{1/2 + x}, \qquad 0 < y < 1$$

then

$$E(Y \mid x) = \int_0^1 y f(y \mid x)\, dy = \int_0^1 \frac{y(y+x)}{1/2+x}\, dy = \frac{2+3x}{3+6x}$$

Conditional Probability and Expectation (cont.)

Compare. If the PDF of X is

$$f_\theta(x) = \frac{x+\theta}{1/2+\theta}, \qquad 0 < x < 1$$

then

$$\Pr_\theta(X > 1/2) = \int_{1/2}^{1} f_\theta(x)\,dx = \int_{1/2}^{1} \frac{x+\theta}{1/2+\theta}\,dx = \frac{3+4\theta}{4+8\theta}$$

If the conditional PDF of Y given X is

$$f(y \mid x) = \frac{y+x}{1/2+x}, \qquad 0 < y < 1$$

then

$$\Pr(Y > 1/2 \mid x) = \int_{1/2}^{1} f(y \mid x)\,dy = \int_{1/2}^{1} \frac{y+x}{1/2+x}\,dy = \frac{3+4x}{4+8x}$$

Conditional Probability and Expectation (cont.)

So far, everything in conditional probability theory is just like ordinary probability theory. Only the notation is different.

Now for the new stuff.

Normalization

Suppose h is a nonnegative function. Does there exist a constant c such that $f = c \cdot h$ is a PDF and, if so, what is it?

If we choose c to be nonnegative, then we automatically have the first property of a PDF

$$f(x) \geq 0, \qquad \text{for all } x.$$

To get the second property

$$\int f(x)\, dx = c \int h(x)\, dx = 1$$

we clearly need the integral of h to be finite and nonzero, in which case

$$c = \frac{1}{\int h(x)\, dx}$$

Normalization (cont.)

So

$$f(x) = \frac{h(x)}{\int h(x)\,dx}$$

This process of dividing a function by what it integrates to (or sums to in the discrete case) is called *normalization*.

We have already done this several times in homework without giving the process a name.

Normalization (cont.)

We say a function h is called an *unnormalized* PDF if it is non-negative and has finite and nonzero integral, in which case

$$f(x) = \frac{h(x)}{\int h(x)\, dx}$$

is the corresponding *normalized* PDF.

We say a function h is called an *unnormalized* PMF if it is non-negative and has finite and nonzero sum, in which case

$$f(x) = \frac{h(x)}{\sum_x h(x)}$$

is the corresponding *normalized* PMF.

Conditional Probability as Renormalization

Suppose we have a joint PMF or PDF f for two random variables X and Y.

After we observe a value x for X, the only values of the random vector (X, Y) that are possible are (x, y) where the x is the same observed value. That is, y is still a variable, but x has been fixed.

Hence what is now interesting is the function

$$y \mapsto f(x, y)$$

a function of one variable, a different function for each fixed x. That is, y is a variable, but x plays the role of a parameter.

Conditional Probability as Renormalization (cont.)

The function of two variables

$$(x, y) \mapsto f(x, y)$$

is a normalized PMF or PDF, but we are no longer interested in it.

The function of one variable

$$y \mapsto f(x, y)$$

is an unnormalized PMF or PDF, that describes the conditional distribution. How do we normalize it?

Conditional Probability as Renormalization (cont.)

Discrete case (sum)

$$f(y \mid x) = \frac{f(x, y)}{\sum_y f(x, y)} = \frac{f(x, y)}{f_X(x)}$$

Continuous case (integrate)

$$f(y \mid x) = \frac{f(x, y)}{\int f(x, y)\, dy} = \frac{f(x, y)}{f_X(x)}$$

In both cases

$$f(y \mid x) = \frac{f(x, y)}{f_X(x)}$$

or

$$\text{conditional} = \frac{\text{joint}}{\text{marginal}}$$

Joint, Marginal, and Conditional

It is important to remember the relationships

$$\text{conditional} = \frac{\text{joint}}{\text{marginal}}$$

and

$$\text{joint} = \text{conditional} \times \text{marginal}$$

but not enough. You have to remember which marginal.

Joint, Marginal, and Conditional (cont.)

The marginal is for the variable(s) behind the bar in the conditional.

It is important to remember the relationships

$$f(y \mid x) = \frac{f(x, y)}{f_X(x)}$$

and

$$f(x, y) = f(y \mid x) f_X(x)$$

Joint, Marginal, and Conditional (cont.)

All of this generalizes to the case of many variables with the same slogan.

> *The marginal is for the variable(s) behind the bar in the conditional.*

$$f(u, v, w, x \mid y, z) = \frac{f(u, v, w, x, y, z)}{f_{Y,Z}(y, z)}$$

and

$$f(u, v, w, x, y, z) = f(u, v, w, x \mid y, z) \times f_{Y,Z}(y, z)$$

Joint to Conditional

Suppose the joint is

$$f(x,y) = c(x+y)^2, \qquad 0 < x < 1,\ 0 < y < 1$$

then the marginal for X is

$$\begin{aligned} f(x) &= \int_0^1 c(x^2 + 2xy + y^2)\, dy \\ &= c\left(x^2 y + xy^2 + \frac{y^3}{3}\right)\Bigg|_0^1 \\ &= c\left(x^2 + x + \frac{1}{3}\right) \end{aligned}$$

and the conditional for Y given X is

$$f(y \mid x) = \frac{(x+y)^2}{x^2 + x + 1/3}, \qquad 0 < y < 1$$

Joint to Conditional (cont.)

The preceding example shows an important point: even though we did not know the constant c that normalizes the joint distribution, it did not matter.

When we renormalize the joint to obtain the conditional, this constant c cancels.

Conclusion: the joint PMF or PDF does not need to be normalized, since we need to renormalize anyway.

Joint to Conditional (cont.)

Suppose the marginal distribution of X is $\mathcal{N}(\mu, \sigma^2)$ and the conditional distribution of Y given X is $\mathcal{N}(X, \tau^2)$. What is the conditional distribution of X given Y?

As we just saw, we can ignore constants for the joint distribution. The unnormalized joint PDF is conditional times marginal

$$\exp(-(y-x)^2/2\tau^2)\exp(-(x-\mu)^2/2\sigma^2)$$

Joint to Conditional (cont.)

In aid of doing this problem we prove a lemma that is useful, since we will do a similar calculation many, many times.

The "e to a quadratic" lemma says that $x \mapsto e^{ax^2+bx+c}$ is an unnormalized PDF if and only if $a < 0$, in which case it the unnormalized PDF of the $\mathcal{N}(-b/2a, -1/2a)$ distribution.

First, if $a \geq 0$, then $x \mapsto e^{ax^2+bx+c}$ is bounded away from zero as either $x \to \infty$ or as $x \to -\infty$ (or perhaps both). Hence the integral of this function is not finite. So it is not an unnormalized PDF.

Joint to Conditional (cont.)

In case $a < 0$ we compare exponents with a normal PDF

$$ax^2 + bx + c$$

and

$$-\frac{(x-\mu)^2}{2\sigma^2} = -\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}$$

and we see that

$$a = -1/2\sigma^2$$
$$b = \mu/\sigma^2$$

so

$$\sigma^2 = -1/2a$$
$$\mu = b\sigma^2 = -b/2a$$

works.

Joint to Conditional (cont.)

Going back to our example with joint PDF

$$\exp\left(-\frac{(y-x)^2}{2\tau^2}-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$=\exp\left(-\frac{y^2}{2\tau^2}+\frac{xy}{\tau^2}-\frac{x^2}{2\tau^2}-\frac{x^2}{2\sigma^2}+\frac{x\mu}{\sigma^2}-\frac{\mu^2}{2\sigma^2}\right)$$

$$=\exp\left(\left[-\frac{1}{2\tau^2}-\frac{1}{2\sigma^2}\right]x^2+\left[\frac{y}{\tau^2}+\frac{\mu}{\sigma^2}\right]x+\left[-\frac{y^2}{2\tau^2}-\frac{\mu^2}{2\sigma^2}\right]\right)$$

Joint to Conditional (cont.)

we see that

$$\exp\left(\left[-\frac{1}{2\tau^2}-\frac{1}{2\sigma^2}\right]x^2+\left[\frac{y}{\tau^2}+\frac{\mu}{\sigma^2}\right]x+\left[-\frac{y^2}{2\tau^2}-\frac{\mu^2}{2\sigma^2}\right]\right)$$

does have the form e to a quadratic, so the conditional distribution of X given Y is normal with mean and variance

$$\mu_{\text{cond}}=\frac{\frac{\mu}{\sigma^2}+\frac{y}{\tau^2}}{\frac{1}{\sigma^2}+\frac{1}{\tau^2}}$$

$$\sigma^2_{\text{cond}}=\frac{1}{\frac{1}{\sigma^2}+\frac{1}{\tau^2}}$$

Joint to Conditional (cont.)

An important lesson from the preceding example is that we didn't have to do an integral to recognize that the conditional was a brand name distribution. If we recognize the functional form of $y \mapsto f(x, y)$ as a brand name PDF except for constants, then we are done. We have identified the conditional distribution.

The General Multiplication Rule

If variables X and Y are independent, then we can "factor" the joint PDF or PMF as the product of marginals

$$f(x, y) = f_X(x) f_Y(y)$$

If they are not independent, then we can still "factor" the joint PDF or PMF as as conditional times marginal

$$\begin{aligned} f(x, y) &= f_{Y|X}(y \mid x) f_X(x) \\ &= f_{X|Y}(x \mid y) f_Y(y) \end{aligned}$$

and there are two different ways to do this.

The General Multiplication Rule (cont.)

When there are more variables, there are more factorizations

$$\begin{aligned} f(x,y,z) &= f_{X|Y,Z}(x \mid y,z) f_{Y|Z}(y \mid z) f_Z(z) \\ &= f_{X|Y,Z}(x \mid y,z) f_{Z|Y}(z \mid y) f_Y(y) \\ &= f_{Y|X,Z}(y \mid x,z) f_{X|Z}(x \mid z) f_Z(z) \\ &= f_{Y|X,Z}(y \mid x,z) f_{Z|X}(z \mid x) f_X(x) \\ &= f_{Z|X,Y}(z \mid x,y) f_{X|Y}(x \mid y) f_Y(y) \\ &= f_{Z|X,Y}(z \mid x,y) f_{Y|X}(y \mid x) f_X(x) \end{aligned}$$

The General Multiplication Rule (cont.)

This is actually clearer without the clutter of subscripts

$$\begin{aligned} f(x, y, z) &= f(x \mid y, z) f(y \mid z) f(z) \\ &= f(x \mid y, z) f(z \mid y) f(y) \\ &= f(y \mid x, z) f(x \mid z) f(z) \\ &= f(y \mid x, z) f(z \mid x) f(x) \\ &= f(z \mid x, y) f(x \mid y) f(y) \\ &= f(z \mid x, y) f(y \mid x) f(x) \end{aligned}$$

and this considers only factorizations in which each "term" has only one variable in front of the bar.

Review

So far we have done two topics in conditional probability theory.

The definition of conditional probability and expectation is just like the definition of unconditional probability and expectation: variables behind the bar in the former act just like parameters in the latter.

One converts between joint and conditional with

$$\text{conditional} = \text{joint}/\text{marginal}$$
$$\text{joint} = \text{conditional} \times \text{marginal}$$

although one often doesn't need to actually calculate the marginal in going from joint to conditional; recognizing the unnormalized density is enough.

Conditional Expectations as Random Variables

An ordinary expectation is a number not a random variable. $E_\theta(X)$ is not random, not a function of X, but it is a function of the parameter θ.

A conditional expectation is a number not a random variable. $E(Y \mid x)$ is not random, not a function of Y, but it is a function of the observed value x of the variable behind the bar.

Say $E(Y \mid x) = g(x)$.

g is an ordinary mathematical function, and x is just a number, so $g(x)$ is just a number.

But $g(X)$ is a random variable when we consider X a random variable.

Conditional Expectations as Random Variables (cont.)

If we write

$$g(x) = E(Y \mid x)$$

then we also write

$$g(X) = E(Y \mid X)$$

to indicate the corresponding random variable.

Wait a minute! Isn't conditional probability about the distribution of Y when X has already been observed to have the value x and is no longer random?

Uh. Yes and no. Before, yes. Now, no.

Conditional Expectations as Random Variables (cont.)

The woof about "after you have observed X but before you have observed Y" is just that, philosophical woof that may help intuition but is not part of the mathematical formalism. None of our definitions of conditional probability and expectation require it.

For example, none of the "factorizations" of joint distributions into marginals and conditionals (slides 40–42) have anything to do with whether a variable has been "observed" or not.

So when we now say that $E(Y \mid X)$ is a random variable that is a function of X but not a function of Y, that is what it is.

Iterated Expectation

If X and Y are continuous

$$\begin{aligned} E\{E(Y \mid X)\} &= \int E(Y \mid x) f(x)\, dx \\ &= \int \left[\int y f(y \mid x)\, dy \right] f(x)\, dx \\ &= \iint y f(y \mid x) f(x)\, dy\, dx \\ &= \iint y f(x, y)\, dy\, dx \\ &= E(Y) \end{aligned}$$

The same is true if X and Y are discrete (replace integrals by sums).

The same is true if one of X and Y is discrete and the other continuous (replace one of the integrals by a sum).

Iterated Expectation Axiom

In summary

$$E\{E(Y \mid X)\} = E(Y)$$

holds for any random variables X and Y that we know how to deal with.

It is taken to be an axiom of conditional probability theory. It is required to hold for anything anyone wants to call conditional expectation.

Other Axioms for Conditional Expectation

The following are obvious from the analogy with unconditional expectation.

$$E(X + Y \mid Z) = E(X \mid Z) + E(Y \mid Z) \tag{1}$$

$$E(X \mid Z) \geq 0, \qquad \text{when } X \geq 0 \tag{2}$$

$$E(aX \mid Z) = aE(X \mid Z) \tag{3}$$

$$E(1 \mid Z) = 1 \tag{4}$$

Other Axioms for Conditional Expectation (cont.)

The "constants come out" axiom (3) can be strengthened. Since variables behind the bar play the role of parameters, which behave like constants in these four axioms, any function of the variables behind the bar behaves like a constant.

$$E\{a(Z)X \mid Z\} = a(Z)E(X \mid Z)$$

for any function a.

Conditional Expectation Axiom Summary

$$E(X + Y \mid \mathbf{Z}) = E(X \mid \mathbf{Z}) + E(Y \mid \mathbf{Z}) \quad (1)$$

$$E(X \mid \mathbf{Z}) \geq 0, \qquad \text{when } X \geq 0 \quad (2)$$

$$E\{a(\mathbf{Z})X \mid \mathbf{Z}\} = a(\mathbf{Z})E(X \mid \mathbf{Z}) \quad (3^*)$$

$$E(1 \mid \mathbf{Z}) = 1 \quad (4)$$

$$E\{E(X \mid \mathbf{Z})\} = E(X) \quad (5)$$

We have changed the variables behind the bar to boldface to indicate, that these also hold when there is more than one variable behind the bar.

We see that, axiomatically, ordinary and conditional expectation are just alike except that (3*) is stronger than (3) and the iterated expectation axiom (5) applies only to conditional expectation.

Consequences of Axioms

All the consequences we derived from the axioms for expectation carry over to conditional expectation if one makes appropriate changes of notation.

Here are some.

The best prediction of Y that is a function of X is $E(Y \mid X)$ when the criterion is expected squared prediction error.

The best prediction of Y that is a function of X is the median of the conditional distribution of Y given X when the criterion is expected absolute prediction error.

Best Prediction

Suppose X and Y have joint distribution

$$f(x, y) = x + y, \qquad 0 < x < 1,\ 0 < y < 1.$$

What is the best prediction of Y when X has been observed?

Best Prediction (cont.)

When expected squared prediction error is the criterion, the answer is

$$
\begin{aligned}
E(Y \mid x) &= \frac{\int_0^1 y(x+y)\,dy}{\int_0^1 (x+y)\,dy} \\
&= \frac{\frac{xy^2}{2} + \frac{y^3}{3}\Big|_0^1}{xy + \frac{y^2}{2}\Big|_0^1} \\
&= \frac{\frac{x}{2} + \frac{1}{3}}{x + \frac{1}{2}}
\end{aligned}
$$

Best Prediction (cont.)

When expected absolute prediction error is the criterion, the answer is the conditional median, which is calculated as follows.

First we find the conditional PDF

$$\begin{aligned} f(y \mid x) &= \frac{x+y}{\int_0^1 (x+y)\,dy} \\ &= \frac{x+y}{xy + \frac{y^2}{2}\Big|_0^1} \\ &= \frac{x+y}{x+\frac{1}{2}} \end{aligned}$$

Best Prediction (cont.)

Then we find the conditional DF. For $0 < y < 1$

$$\begin{aligned} F(y \mid x) &= \Pr(Y \le y \mid x) \\ &= \int_0^y \frac{x+s}{x+\frac{1}{2}}\, ds \\ &= \left. \frac{xs + \frac{s^2}{2}}{x+\frac{1}{2}} \right|_0^y \\ &= \frac{xy + \frac{y^2}{2}}{x+\frac{1}{2}} \end{aligned}$$

Best Prediction (cont.)

Finally we have to solve the equation $F(y \mid x) = 1/2$ to find the median.

$$\frac{xy + \frac{y^2}{2}}{x + \frac{1}{2}} = \frac{1}{2}$$

is equivalent to

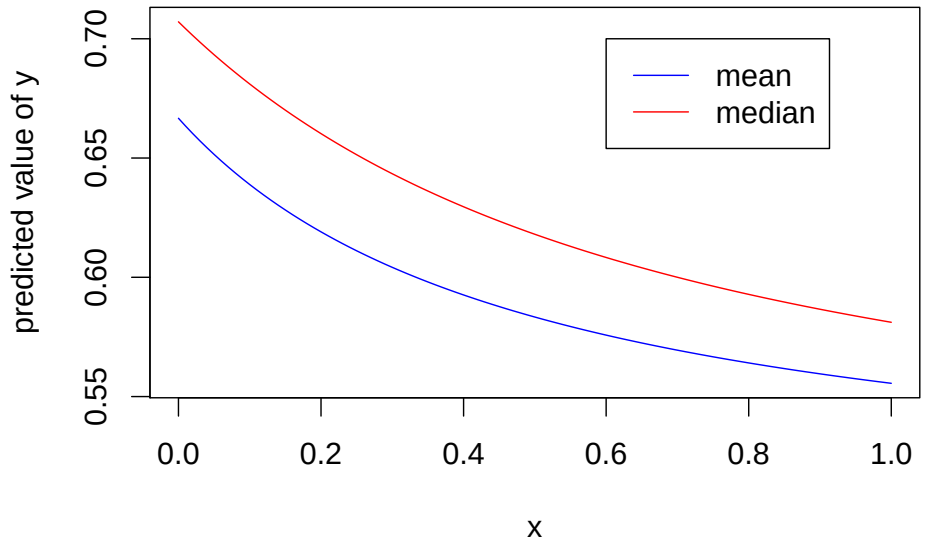
$$y^2 + 2xy - \left(x + \frac{1}{2}\right) = 0$$

which has solution

$$\begin{aligned} y &= \frac{-2x + \sqrt{4x^2 + 4\left(x + \frac{1}{2}\right)}}{2} \\ &= -x + \sqrt{x^2 + x + \frac{1}{2}} \end{aligned}$$

Best Prediction (cont.)

Here are the two types compared for this example.



Conditional Variance

Conditional variance is just like variance, just replace ordinary expectation with conditional expectation.

$$\begin{aligned} \operatorname{var}(Y \mid X) &= E\{[Y - E(Y \mid X)]^2 \mid X\} \\ &= E(Y^2 \mid X) - E(Y \mid X)^2 \end{aligned}$$

Similarly

$$\begin{aligned} \operatorname{cov}(X, Y \mid Z) &= E\{[X - E(X \mid Z)][Y - E(Y \mid Z)] \mid Z\} \\ &= E(XY \mid Z) - E(X \mid Z)E(Y \mid Z) \end{aligned}$$

Conditional Variance (cont.)

$$
\begin{aligned}
\text{var}(Y) &= E\{[Y - E(Y)]^2\} \\
&= E\{[Y - E(Y \mid X) + E(Y \mid X) - E(Y)]^2\} \\
&= E\{[Y - E(Y \mid X)]^2\} \\
&\quad + 2E\{[Y - E(Y \mid X)][E(Y \mid X) - E(Y)]\} \\
&\quad + E\{[E(Y \mid X) - E(Y)]^2\}
\end{aligned}
$$

Conditional Variance (cont.)

By iterated expectation

$$\begin{aligned} E\{[Y - E(Y \mid X)]^2\} &= E\Big(E\{[Y - E(Y \mid X)]^2 \mid X\}\Big) \\ &= E\{\mathsf{var}(Y \mid X)\} \end{aligned}$$

and

$$E\{[E(Y \mid X) - E(Y)]^2\} = \mathsf{var}\{E(Y \mid X)\}$$

because $E\{E(Y \mid X)\} = E(Y)$.

Conditional Variance (cont.)

$$
\begin{aligned}
&E\{[Y - E(Y \mid X)][E(Y \mid X) - E(Y)]\} \\
&\qquad = E\Big(E\{[Y - E(Y \mid X)][E(Y \mid X) - E(Y)] \mid X\}\Big) \\
&\qquad = E\Big([E(Y \mid X) - E(Y)]E\{[Y - E(Y \mid X)] \mid X\}\Big) \\
&\quad = E\Big(\Big[E(Y \mid X) - E(Y)\Big]\Big[E(Y \mid X) - E\{E(Y \mid X) \mid X)\}\Big]\Big) \\
&\quad = E\Big(\Big[E(Y \mid X) - E(Y)\Big]\Big[E(Y \mid X) - E(Y \mid X)E(1 \mid X)\Big]\Big) \\
&\qquad = E\Big(\Big[E(Y \mid X) - E(Y)\Big]\Big[E(Y \mid X) - E(Y \mid X)\Big]\Big) \\
&\qquad\qquad = 0
\end{aligned}
$$

Conditional Variance (cont.)

In summary, this is the iterated variance theorem

$$\text{var}(Y) = E\{\text{var}(Y \mid X)\} + \text{var}\{E(Y \mid X)\}$$

Conditional Variance (cont.)

If the conditional distribution of Y given X is Gam(X, X) and $1/X$ has mean 10 and standard deviation 2, then what is var(Y)?

First

$$E(Y \mid X) = \frac{\alpha}{\lambda} = \frac{X}{X} = 1$$
$$\text{var}(Y \mid X) = \frac{\alpha}{\lambda^2} = \frac{X}{X^2} = 1/X$$

So

$$\begin{aligned} \text{var}(Y) &= E\{\text{var}(Y \mid X)\} + \text{var}\{E(Y \mid X)\} \\ &= E(1/X) + \text{var}(1) \\ &= 10 \end{aligned}$$

Conditional Probability and Independence

X and Y are independent random variables if and only if

$$f(x, y) = f_X(x) f_Y(y)$$

and

$$f(y \mid x) = \frac{f(x, y)}{f_X(x)} = f_Y(y)$$

and, similarly

$$f(x \mid y) = f_X(x)$$

Conditional Probability and Independence (cont.)

Generalizing to many variables, the random vectors $\mathbf{X}$ and $\mathbf{Y}$ are independent if and only if the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$ is the same as the marginal distribution of $\mathbf{Y}$ (or the same with $\mathbf{X}$ and $\mathbf{Y}$ interchanged).

Bernoulli Process

A sequence X_1, X_2, ... of IID Bernoulli random variables is called a *Bernoulli process*.

The number of successes ($X_i = 1$) in the first n variables has the Bin(n, p) distribution where $p = E(X_i)$ is the success probability.

The waiting time to the first success (the number of failures before the first success) has the Geo(p) distribution.

Bernoulli Process (cont.)

Because of the independence of the X_i, the number of failures from "now" until the next success also has the Geo(p) distribution.

In particular, the numbers of failures between successes are independent and have the Geo(p) distribution.

Bernoulli Process (cont.)

Define

$$\begin{aligned} T_0 &= 0 \\ T_1 &= \min\{\, i \in \mathbb{N} : i > T_0 \text{ and } X_i = 1 \,\} \\ T_2 &= \min\{\, i \in \mathbb{N} : i > T_1 \text{ and } X_i = 1 \,\} \\ &\vdots \\ T_{k+1} &= \min\{\, i \in \mathbb{N} : i > T_k \text{ and } X_i = 1 \,\} \\ &\vdots \end{aligned}$$

and

$$Y_k = T_k - T_{k-1} - 1, \qquad k = 1, 2, \ldots,$$

then the Y_k are IID Geo(p).

Poisson Process

The Poisson process is the continuous analog of the Bernoulli process. We replace Geo(p) by Exp(λ) for the interarrival times.

Suppose T_1, T_2, ... are IID Exp(λ), and define

$$X_n = \sum_{i=1}^{n} T_i, \qquad n = 1, 2, \ldots.$$

The one-dimensional spatial point process with points at X_1, X_2, ... is called the Poisson process with rate parameter λ.

Poisson Process (cont.)

The distribution of X_n is Gam(n, λ) by the addition rule for exponential random variables.

We need the DF for this variable. We already know that X_1, which has the Exp(λ) distribution, has DF

$$F_1(x) = 1 - e^{-\lambda x}, \qquad 0 < x < \infty.$$

Poisson Process (cont.)

For $n > 1$ we use integration by parts with $u = s^{n-1}$ and $dv = e^{-\lambda s}\,ds$ and $v = -(1/\lambda)e^{-\lambda s}$, obtaining

$$\begin{aligned} F_n(x) &= \frac{\lambda^n}{\Gamma(n)} \int_0^x s^{n-1} e^{-\lambda s}\,ds \\ &= -\frac{\lambda^{n-1}}{(n-1)!} s^{n-1} e^{-\lambda s}\Big|_0^x + \int_0^x \frac{\lambda^{n-1}}{(n-2)!} s^{n-2} e^{-\lambda s}\,ds \\ &= -\frac{\lambda^{n-1}}{(n-1)!} x^{n-1} e^{-\lambda x} + F_{n-1}(x) \end{aligned}$$

so

$$F_n(x) = 1 - e^{-\lambda x} \sum_{k=0}^{n-1} \frac{(\lambda x)^k}{k!}$$

Poisson Process (cont.)

There are exactly n points in the interval $(0, t)$ if $X_n < t < X_{n+1}$, and

$$\begin{aligned}
\Pr(X_n < t < X_{n+1}) &= 1 - \Pr(X_n > t \text{ or } X_{n+1} < t) \\
&= 1 - \Pr(X_n > t) - \Pr(X_{n+1} < t) \\
&= 1 - [1 - F_n(t)] - F_{n+1}(t) \\
&= F_n(t) - F_{n+1}(t) \\
&= \frac{(\lambda t)^n}{n!} e^{-\lambda t}
\end{aligned}$$

Thus we have discovered that the probability distribution of the random variable Y which is the number of points in $(0, t)$ has the Poi(λt) distribution.

Memoryless Property of the Exponential Distribution

If the distribution of the random variable X is Exp(λ), then so is the conditional distribution of $X - a$ given the event $X > a$, where $a > 0$.

This conditioning is a little different from what we have seen before. The PDF of X is

$$f(x) = \lambda e^{-\lambda x}, \qquad x > 0.$$

To condition on the event $X > a$ we renormalize the part of the distribution on the interval (a, ∞)

$$f(x \mid X > a) = \frac{\lambda e^{-\lambda x}}{\int_a^\infty \lambda e^{-\lambda x}\, dx} = \lambda e^{-\lambda(x-a)}$$

Memoryless Property of the Exponential Distribution (cont.)

Now define $Y = X - a$. The "Jacobian" for this change-of-variable is equal to one, so

$$f(y \mid X > a) = \lambda e^{-\lambda y}, \qquad y > 0,$$

and this is what was to be proved.

Poisson Process (cont.)

Suppose bus arrivals follow a Poisson process (they don't but just suppose). You arrive at time a. The waiting time until the next bus arrives is Exp(λ) by the memoryless property. Then the interarrival times between following buses are also Exp(λ). Hence the future pattern of arrival times also follows a Poisson process.

Moreover, since the distribution of time of the arrival of the next bus after time a does not depend on the past history of the process, the entire future of the process (all arrivals after time a) is independent of the entire past of the process (all arrivals before time a).

Poisson Process (cont.)

Thus we see that for any a and b with $0 < a < b < \infty$, the number of points in (a, b) is Poisson with mean $\lambda(b - a)$, and counts of points in disjoint intervals are independent random variables.

Thus we have come the long way around to our original definition of the Poisson process: counts in nonoverlapping intervals are independent and Poisson distributed, and the expected count in an interval of length t is λt for some constant $\lambda > 0$ called the rate parameter.

Poisson Process (cont.)

We have also learned an important connection with the exponential distribution. All waiting times and interarrival times in a Poisson process have the Exp(λ) distribution, where λ is the rate parameter.

- Counts in non-overlapping intervals are independent.
- Waiting times and interarrival times are independent.
- Counts in an interval of length t are Poi(λt).
- Waiting and interarrival times are Exp(λ).

Multinomial Distribution

So far all of our brand name distributions are univariate. We will do two multivariate ones in this deck and another in Deck 8. Here is one of them.

A random vector $\mathbf{X} = (X_1, \ldots, X_k)$ is called *multivariate Bernoulli* if its components are zero-or-one-valued and sum to one. These two assumptions imply that exactly one of the X_i is equal to one and the rest are zero.

The distributions of these random vectors form a parametric family with parameter

$$E(\mathbf{X}) = \mathbf{p} = (p_1, \ldots, p_k)$$

called the *success probability* parameter vector.

Multinomial Distribution (cont.)

The distribution of X_i is Ber(p_i), so

$$E(X_i) = p_i$$
$$\text{var}(X_i) = p_i(1 - p_i)$$

for all i.

But the components of $\mathbf{X}$ are not independent. When $i \neq j$ we have $X_i X_j = 0$, because exactly one component of $\mathbf{X}$ is nonzero. Thus

$$\begin{aligned} \text{cov}(X_i, X_j) &= E(X_i X_j) - E(X_i)E(X_j) \\ &= -p_i p_j \end{aligned}$$

Multinomial Distribution (cont.)

We can write the mean vector

$$E(\mathbf{X}) = \mathbf{p}$$

and variance matrix

$$\text{var}(X) = \mathbf{P} - \mathbf{p}\mathbf{p}^T$$

where $\mathbf{P}$ is the diagonal matrix whose diagonal is $\mathbf{p}$. (The i, i-th element of $\mathbf{P}$ is the i-th element of $\mathbf{p}$. The i, j-th element of $\mathbf{P}$ is zero when $i \neq j$.)

Multinomial Distribution (cont.)

If $\mathbf{X}_1$, $\mathbf{X}_2$, ..., $\mathbf{X}_n$ are IID multivariate Bernoulli random vectors (the subscript does *not* indicate components of a vector) with success probability vector $\mathbf{p}$, then

$$\mathbf{Y} = \sum_{i=1}^{n} \mathbf{X}_i$$

has the *multinomial distribution* with *sample size* n and *success probability* vector $\mathbf{p}$, which is denoted Multi$(n, \mathbf{p})$.

Suppose we have an IID sample of n individuals and each individual is classified into exactly one of k categories. Let Y_j be the number of individuals in the j-th category. Then $\mathbf{Y} = (Y_1, \ldots, Y_k)$ has the Multi$(n, \mathbf{p})$ distribution.

Multinomial Distribution (cont.)

Since the expectation of a sum is the sum of the expectations,

$$E(\mathbf{Y}) = n\mathbf{p}$$

Since the variance of a sum is the sum of the variances when the terms are independent (and this holds when the terms are random vectors too),

$$\text{var}(\mathbf{Y}) = n(\mathbf{P} - \mathbf{p}\mathbf{p}^T)$$

Multinomial Distribution (cont.)

We find the PMF of the multinomial distribution by the same argument as for the binomial.

First, consider the case where we specify each $\mathbf{X}_j$

$$\Pr(\mathbf{X}_j = \mathbf{x}_j,\ j = 1, \ldots, n) = \prod_{j=1}^{n} \Pr(\mathbf{X}_j = \mathbf{x}_j) = \prod_{i=1}^{k} p_i^{y_i}$$

where

$$(y_1, \ldots, y_k) = \sum_{j=1}^{n} \mathbf{x}_j,$$

because in the product running from 1 to n each factor is a component of $\mathbf{p}$ and the number of factors that are equal to p_i is equal to the number of $\mathbf{X}_j$ whose i-th component is equal to one, and that is y_i.

Multinomial Coefficients

Then we consider how many ways we can rearrange the $\mathbf{X}_j$ values and get the same $\mathbf{Y}$, that is, how many ways can we choose which of the individuals are in first category, which in the second, and so forth?

The answer is just like the derivation of binomial coefficients.

The number of ways to allocate n individuals to k categories so that there are y_1 in the first category, y_2 in the second, and so forth is

$$\binom{n}{\mathbf{y}} = \binom{n}{y_1, y_2, \ldots, y_k} = \frac{n!}{y_1! \, y_2! \cdots y_k!}$$

which is called a multinomial coefficient.

Multinomial Distribution (cont.)

The PMF of the Multi($n, \mathbf{p}$) distribution is

$$f(\mathbf{y}) = \binom{n}{\mathbf{y}} \prod_{i=1}^{k} p_i^{y_i}$$

Multinomial Theorem

The fact that the PMF of the multinomial distribution sums to one is equivalent to the multinomial theorem

$$\left(\sum_{i=1}^{k} a_i \right)^n = \sum_{\substack{\mathbf{x} \in \mathbb{N}^k \\ x_1 + \cdots + x_k = n}} \binom{n}{\mathbf{x}} \prod_{i=1}^{k} a_i^{x_i}$$

of which the binomial theorem is the $k = 2$ special case.

As in the binomial theorem, the a_i do not have to be nonnegative and do not have to sum to one.

Multinomial and Binomial

However, the binomial distribution is not the $k = 2$ special case of the multinomial distribution.

If the random scalar X has the Bin(n, p) distribution, then the random vector $(X, n - X)$ has the Multi$(n, \mathbf{p})$ distribution, where $\mathbf{p} = (p, 1 - p)$.

The binomial arises when there are two categories (conventionally called "success" and "failure"). The binomial random scalar only counts the successes. A multinomial random vector counts all the categories. When $k = 2$ it counts both successes and failures.

Multinomial and Degeneracy

Because a Multi$(n, \mathbf{p})$ random vector $\mathbf{Y}$ counts all the cases, we always have

$$Y_1 + \cdots + Y_k = n$$

Thus a multinomial random vector is not truly k dimensional, since we can always write any one count as a function of the others

$$Y_1 = n - Y_2 - \cdots - Y_k$$

So the distribution of $\mathbf{Y}$ is "really" $k - 1$ dimensional at best.

Further degeneracy arises if $p_i = 0$ for some i, in which case $Y_i = 0$ always.

Multinomial Marginals and Conditionals

The short story is all the marginals and conditionals of a multinomial are again multinomial, but this is not quite right.

It is true for conditionals and "almost true" for marginals.

Multinomial Univariate Marginals

One type of marginal is trivial. If $(Y_1, \ldots, Y_k)$ has the Multi$(n, \mathbf{p})$ distribution, where $\mathbf{p} = (p_1, \ldots, p_k)$, then the marginal distribution of Y_j is Bin(n, p_j), because it is the sum of n IID Bernoullis with success probability p_j.

Multinomial Marginals

What is true, obviously true from the definition, is that collapsing categories gives another multinomial, and the success probability for a collapsed category is the sum of the success probabilities for the categories so collapsed. Suppose we have

category	Obama	McCain	Barr	Nader	Other
probability	0.51	0.46	0.02	0.01	0.00

and we decide to collapse the last three categories obtaining

category	Obama	McCain	New Other
probability	0.51	0.46	0.03

The principle is obvious, although the notation can be a little messy.

Multinomial Marginals (cont.)

Since the numbering of categories is arbitrary, we consider the marginal distribution of $Y_{j+1}, \ldots, Y_k$.

That marginal distribution is not multinomial since we need to add the "other" category, which has count $Y_1 + \cdots + Y_j$, to be able to classify all individuals.

The random vector $\mathbf{Z} = (Y_1 + \cdots + Y_j, Y_{j+1}, \ldots, Y_k)$ has the Multi$(n, \mathbf{q})$ distribution, where $\mathbf{q} = (p_1 + \cdots + p_j, p_{j+1}, \ldots, p_k)$.

Multinomial Marginals (cont.)

We can consider the marginal of $Y_{j+1}, \ldots, Y_k$ in two different ways. Define $W = Y_1 + \cdots + Y_j$. Then

$$f(w, y_{j+1}, \ldots, y_k) = \binom{n}{w, y_{j+1}, \ldots, y_k} (p_1 + \cdots + p_j)^w p_{j+1}^{y_{j+1}} \cdots p_k^{y_k}$$

is a multinomial PMF of the random vector $(W, Y_{j+1}, \ldots, Y_k)$.

But since $w = n - y_{j+1} - \cdots - y_k$, we can also write

$$f(y_{j+1}, \ldots, y_k) = \frac{n!}{(n - y_{j+1} - \cdots - y_k)!\, y_{j+1}! \cdots y_k!} \times (p_1 + \cdots + p_j)^{n - y_{j+1} - \cdots - y_k} p_{j+1}^{y_{j+1}} \cdots p_k^{y_k}$$

which is not, precisely, a multinomial PMF.

Multinomial Conditionals

Since the numbering of categories is arbitrary, we consider the conditional distribution of the $Y_1, \ldots, Y_j$ given $Y_{j+1}, \ldots, Y_k$.

$$
\begin{aligned}
f(y_1, \ldots, y_j \mid y_{j+1}, \ldots, y_k) &= \frac{f(y_1, \ldots, y_k)}{f(y_{j+1}, \ldots, y_k)} \\
&= \frac{\dfrac{n!}{y_1! \cdots, y_k!}}{\dfrac{n!}{(n - y_{j+1} - \cdots - y_k)!\, y_{j+1}! \cdots y_k!}} \\
&\quad \times \frac{p_1^{y_1} \cdots p_k^{y_k}}{(p_1 + \cdots + p_j)^{n - y_{j+1} - \cdots - y_k} p_{j+1}^{y_{j+1}} \cdots p_k^{y_k}} \\
&= \frac{(y_1 + \cdots + y_j)!}{y_1! \cdots y_j!} \prod_{i=1}^{j} \left(\frac{p_i}{p_1 + \cdots + p_j} \right)^{y_i}
\end{aligned}
$$

Multinomial Conditionals (cont.)

Thus we see that the conditional distribution of $Y_1, \ldots, Y_j$ given $Y_{j+1}, \ldots, Y_k$ is Multi$(m, \mathbf{q})$ where

$$m = n - Y_{j+1} - \ldots - Y_k$$

and

$$q_i = \frac{p_i}{p_1 + \cdots + p_j}, \qquad i = 1, \ldots, j$$

The Multivariate Normal Distribution

A random vector having IID standard normal components is called *standard multivariate normal*. Of course, the joint distribution is the product of marginals

$$\begin{aligned} f(z_1, \ldots, z_n) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-z_i^2/2} \\ &= (2\pi)^{-n/2} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} z_i^2 \right) \end{aligned}$$

and we can write this using vector notation as

$$f(\mathbf{z}) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2} \mathbf{z}^T \mathbf{z} \right)$$

Multivariate Location-Scale Families

A univariate location-scale family with standard distribution having PDF f is the set of all distributions of random variables that are invertible linear transformations $Y = \mu + \sigma X$, where X has the standard distribution. The PDF's have the form

$$f_{\mu,\sigma}(y) = \frac{1}{|\sigma|} f\left(\frac{y-\mu}{\sigma}\right)$$

A multivariate location-scale family with standard distribution having PDF f is the set of all distributions of random vectors that are invertible linear transformations $\mathbf{Y} = \boldsymbol{\mu} + \mathbf{B}\mathbf{X}$ where $\mathbf{X}$ has the standard distribution. The PDF's have the form

$$f_{\boldsymbol{\mu},\mathbf{B}}(\mathbf{y}) = f\left(\mathbf{B}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right) \cdot |\det(\mathbf{B}^{-1})|$$

The Multivariate Normal Distribution (cont.)

The family of multivariate normal distributions is the set of all distributions of random vectors that are (not necessarily invertible) linear transformations $\mathbf{Y} = \boldsymbol{\mu} + \mathbf{B}\mathbf{X}$, where $\mathbf{X}$ is standard multivariate normal.

The Multivariate Normal Distribution (cont.)

The mean vector and variance matrix of a standard multivariate normal random vector are the zero vector and identity matrix.

By the rules for linear transformations, the mean vector and variance matrix of $\mathbf{Y} = \boldsymbol{\mu} + \mathbf{B}\mathbf{X}$ are

$$\begin{aligned} E(\mathbf{Y}) &= E(\boldsymbol{\mu} + \mathbf{B}\mathbf{X}) \\ &= \boldsymbol{\mu} + \mathbf{B}E(\mathbf{X}) \\ &= \boldsymbol{\mu} \\ \text{var}(\mathbf{Y}) &= \text{var}(\boldsymbol{\mu} + \mathbf{B}\mathbf{X}) \\ &= \mathbf{B}\,\text{var}(\mathbf{X})\mathbf{B}^T \\ &= \mathbf{B}\mathbf{B}^T \end{aligned}$$

The Multivariate Normal Distribution (cont.)

The transformation $\mathbf{Y} = \boldsymbol{\mu} + \mathbf{BX}$ is invertible if and only if the matrix $\mathbf{B}$ is invertible, in which case the PDF of $\mathbf{Y}$ is

$$f(\mathbf{y}) = (2\pi)^{-n/2} \cdot |\det(\mathbf{B}^{-1})| \cdot \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T(\mathbf{B}^{-1})^T\mathbf{B}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right)$$

This can be simplified. Write

$$\text{var}(\mathbf{Y}) = \mathbf{B}\mathbf{B}^T = \mathbf{M}$$

Then

$$(\mathbf{B}^{-1})^T\mathbf{B}^{-1} = (\mathbf{B}^T)^{-1}\mathbf{B}^{-1} = (\mathbf{B}\mathbf{B}^T)^{-1} = \mathbf{M}^{-1}$$

and

$$\det(\mathbf{M})^{-1} = \det(\mathbf{M}^{-1}) = \det\Big((\mathbf{B}^{-1})^T\mathbf{B}^{-1}\Big) = \det\Big(\mathbf{B}^{-1}\Big)^2$$

The Multivariate Normal Distribution (cont.)

Thus

$$f(\mathbf{y}) = (2\pi)^{-n/2} \det(\mathbf{M})^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T \mathbf{M}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right)$$

Thus, as in the univariate case, the distribution of a multivariate normal random vector having a PDF depends only on the mean vector $\boldsymbol{\mu}$ and variance matrix $\mathbf{M}$.

It does not depend on the specific matrix $\mathbf{B}$ used to define it as a function of a standard multivariate normal random vector.

The Spectral Decomposition

Any symmetric matrix $\mathbf{M}$ has a *spectral decomposition*

$$\mathbf{M} = \mathbf{ODO}^T,$$

where $\mathbf{D}$ is diagonal and $\mathbf{O}$ is orthogonal, which means

$$\mathbf{OO}^T = \mathbf{O}^T\mathbf{O} = \mathbf{I},$$

where $\mathbf{I}$ is the identity matrix, which is equivalent to saying

$$\mathbf{O}^{-1} = \mathbf{O}^T$$

The Spectral Decomposition (cont.)

A symmetric matrix $\mathbf{M}$ is positive semidefinite if

$$\mathbf{w}^T\mathbf{M}\mathbf{w} \geq 0, \qquad \text{for all vectors } \mathbf{w}$$

and positive definite if

$$\mathbf{w}^T\mathbf{M}\mathbf{w} > 0, \qquad \text{for all nonzero vectors } \mathbf{w}.$$

The Spectral Decomposition (cont.)

Since

$$\mathbf{w}^T\mathbf{M}\mathbf{w} = \mathbf{w}^T\mathbf{O}\mathbf{D}\mathbf{O}^T\mathbf{w} = \mathbf{v}^T\mathbf{D}\mathbf{v},$$

where

$$\mathbf{v} = \mathbf{O}^T\mathbf{w} \qquad \text{and} \qquad \mathbf{w} = \mathbf{O}\mathbf{v},$$

a symmetric matrix $\mathbf{M}$ is positive semidefinite if and only if the diagonal matrix $\mathbf{D}$ in its spectral decomposition is. And similarly for positive definite.

The Spectral Decomposition (cont.)

If $\mathbf{D}$ is diagonal, then

$$\mathbf{v}^T\mathbf{D}\mathbf{v} = \sum_i \sum_j v_i d_{ij} v_j = \sum_i d_{ii} v_i^2$$

because $d_{ij} = 0$ when $i \neq j$.

Hence a diagonal matrix is positive semidefinite if and only if all its diagonal components are nonnegative, and a diagonal matrix is positive definite if and only if all its diagonal components are positive.

The Spectral Decomposition (cont.)

Since the spectral decomposition

$$\mathbf{M} = \mathbf{O}\mathbf{D}\mathbf{O}^T$$

holds if and only if

$$\mathbf{D} = \mathbf{O}^T\mathbf{M}\mathbf{O},$$

we see that $\mathbf{M}$ is invertible if and only if $\mathbf{D}$ is, in which case

$$\mathbf{M}^{-1} = \mathbf{O}\mathbf{D}^{-1}\mathbf{O}^T$$
$$\mathbf{D}^{-1} = \mathbf{O}^T\mathbf{M}^{-1}\mathbf{O}$$

The Spectral Decomposition (cont.)

If $\mathbf{D}$ and $\mathbf{E}$ are diagonal, then the i, k component of $\mathbf{DE}$ is

$$\sum_j d_{ij} e_{jk} = \begin{cases} 0, & i \neq k \\ d_{ii} e_{ii}, & i = k \end{cases}$$

Hence the product of diagonal matrices is diagonal. And the i, i component of the product is the product of the i, i components of the multiplicands.

From this, it is obvious that a diagonal matrix $\mathbf{D}$ is invertible if and only if its diagonal components d_{ii} are all nonzero, in which case $\mathbf{D}^{-1}$ is the diagonal matrix with diagonal components $1/d_{ii}$.

The Spectral Decomposition (cont.)

If $\mathbf{D}$ is diagonal and positive semidefinite with diagonal components d_{ii}, we define $\mathbf{D}^{1/2}$ to be the diagonal matrix with diagonal components $d_{ii}^{1/2}$.

Observe that

$$\mathbf{D}^{1/2}\mathbf{D}^{1/2} = \mathbf{D}$$

so $\mathbf{D}^{1/2}$ is a matrix square root of $\mathbf{D}$.

The Spectral Decomposition (cont.)

If $\mathbf{M}$ is symmetric and positive semidefinite with spectral decomposition

$$\mathbf{M} = \mathbf{O}\mathbf{D}\mathbf{O}^T,$$

we define

$$\mathbf{M}^{1/2} = \mathbf{O}\mathbf{D}^{1/2}\mathbf{O}^T,$$

Observe that

$$\mathbf{M}^{1/2}\mathbf{M}^{1/2} = \mathbf{O}\mathbf{D}^{1/2}\mathbf{O}^T\mathbf{O}\mathbf{D}^{1/2}\mathbf{O}^T = \mathbf{M}$$

so $\mathbf{M}^{1/2}$ is a matrix square root of $\mathbf{M}$.

The Multivariate Normal Distribution (cont.)

Let $\mathbf{M}$ be any positive semidefinite matrix. If $\mathbf{X}$ is standard multivariate normal, then

$$\mathbf{Y} = \boldsymbol{\mu} + \mathbf{M}^{1/2}\mathbf{X}$$

is general multivariate normal with mean vector

$$E(\mathbf{Y}) = \boldsymbol{\mu}$$

and variance matrix

$$\operatorname{var}(\mathbf{Y}) = \mathbf{M}^{1/2} \operatorname{var}(\mathbf{X}) \mathbf{M}^{1/2} = \mathbf{M}^{1/2} \mathbf{M}^{1/2} = \mathbf{M}$$

Thus every positive semidefinite matrix is the variance matrix of a multivariate normal random vector.

The Multivariate Normal Distribution (cont.)

A linear function of a linear function is linear

$$\boldsymbol{\mu}_1 + \mathbf{B}_1(\boldsymbol{\mu}_2 + \mathbf{B}_2\mathbf{X}) = (\boldsymbol{\mu}_1 + \mathbf{B}_1\boldsymbol{\mu}_2) + (\mathbf{B}_1\mathbf{B}_2)\mathbf{X}$$

thus any linear transformation of a multivariate normal random vector is multivariate normal. To figure out which multivariate normal distribution, calculate its mean vector and variance matrix.

The Multivariate Normal Distribution (cont.)

If $\mathbf{X}$ has the $\mathcal{N}(\boldsymbol{\mu}, \mathbf{M})$ distribution, then

$$\mathbf{Y} = \mathbf{a} + \mathbf{B}\mathbf{X}$$

has the multivariate normal distribution with mean vector

$$E(\mathbf{Y}) = \mathbf{a} + \mathbf{B}E(\mathbf{X}) = \mathbf{a} + \mathbf{B}\boldsymbol{\mu}$$

and variance matrix

$$\text{var}(\mathbf{Y}) = \mathbf{B}\,\text{var}(\mathbf{X})\mathbf{B}^T = \mathbf{B}\mathbf{M}\mathbf{B}^T$$

Addition Rule for Univariate Normal

If $X_1, \ldots, X_n$ are independent univariate normal random variables, then $X_1 + \cdots + X_n$ is univariate normal with mean

$$E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n)$$

and variance

$$\mathrm{var}(X_1 + \cdots + X_n) = \mathrm{var}(X_1) + \cdots + \mathrm{var}(X_n)$$

Addition Rule for Multivariate Normal

If $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independent multivariate normal random vectors, then $\mathbf{X}_1 + \cdots + \mathbf{X}_n$ is multivariate normal with mean vector

$$E(\mathbf{X}_1 + \cdots + \mathbf{X}_n) = E(\mathbf{X}_1) + \cdots + E(\mathbf{X}_n)$$

and variance matrix

$$\mathrm{var}(\mathbf{X}_1 + \cdots + \mathbf{X}_n) = \mathrm{var}(\mathbf{X}_1) + \cdots + \mathrm{var}(\mathbf{X}_n)$$

Partitioned Matrices

When we write

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \end{pmatrix}$$

and say that it is a *partitioned matrix*, we mean that $\mathbf{A}_1$ and $\mathbf{A}_2$ are matrices with the same number of columns stacked one atop the other to make one matrix $\mathbf{A}$.

Multivariate Normal Marginal Distributions

Marginalization is a linear mapping, that is, the mapping

$$\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \mapsto \mathbf{X}_1$$

is linear. Hence every marginal of a multivariate normal distribution is multivariate normal, and, of course, the mean vector and variance matrix of $\mathbf{X}_1$ are

$$\begin{aligned} \boldsymbol{\mu}_1 &= E(\mathbf{X}_1) \\ \mathbf{M}_{11} &= \mathrm{var}(\mathbf{X}_1) \end{aligned}$$

Almost Surely and Degeneracy

We say a property holds *almost surely* if it holds with probability one.

A random variable X has variance zero if and only if it is almost surely constant. This means there is an event A such that $\Pr(A) = 1$ and a constant c such that

$$X(s) = c, \qquad x \in A$$

(X may take other values on the complement of A but this does not matter since such values get multiplied by zero in computing expectations).

Multivariate Normal and Degeneracy

We say a multivariate normal random vector $\mathbf{Y}$ is *degenerate* if its variance matrix $\mathbf{M}$ is not positive definite (only positive semidefinite).

This happens when there is a nonzero vector $\mathbf{a}$ such that

$$\mathbf{a}^T\mathbf{M}\mathbf{a} = 0,$$

in which case

$$\text{var}(\mathbf{a}^T\mathbf{Y}) = \mathbf{a}^T\mathbf{M}\mathbf{a} = 0$$

and $\mathbf{a}^T\mathbf{Y} = c$ almost surely for some constant c.

Multivariate Normal and Degeneracy (cont.)

Since $\mathbf{a}$ is nonzero, it has a nonzero component a_i, and we can write

$$Y_i = \frac{c}{a_i} - \frac{1}{a_i} \sum_{\substack{j=1 \\ j \neq i}}^{n} a_j Y_j$$

This means we can always (perhaps after reordering the components) partition

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix}$$

so that $\mathbf{Y}_2$ has a nondegenerate multivariate normal distribution and $\mathbf{Y}_1$ is a linear function of $\mathbf{Y}_2$, say

$$\mathbf{Y}_1 = \mathbf{d} + \mathbf{B}\mathbf{Y}_2, \qquad \text{almost surely}$$

Multivariate Normal and Degeneracy (cont.)

Now that we have written

$$\mathbf{Y} = \begin{pmatrix} \mathbf{d} + \mathbf{B}\mathbf{Y}_2 \\ \mathbf{Y}_2 \end{pmatrix}$$

we see that the distribution of $\mathbf{Y}$ is completely determined by the mean vector and variance matrix of $\mathbf{Y}_2$, which are themselves part of the mean vector and variance matrix of $\mathbf{Y}$.

Thus we have shown that the distribution of every multivariate normal random vector, degenerate or nondegenerate, is determined by its mean vector and variance matrix.

Partitioned Matrices (cont.)

When we write

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$$

and say that it is a *partitioned matrix*, we mean that $\mathbf{A}_{11}$, $\mathbf{A}_{12}$, $\mathbf{A}_{21}$, and $\mathbf{A}_{22}$ are all matrices, that fit together to make one matrix $\mathbf{A}$.

$\mathbf{A}_{11}$ and $\mathbf{A}_{12}$ have the same number of rows.
$\mathbf{A}_{21}$ and $\mathbf{A}_{22}$ have the same number of rows.
$\mathbf{A}_{11}$ and $\mathbf{A}_{21}$ have the same number of columns.
$\mathbf{A}_{12}$ and $\mathbf{A}_{22}$ have the same number of columns.

Symmetric Partitioned Matrices

A partitioned matrix

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$$

partitioned so that $\mathbf{A}_{11}$ is square is symmetric if and only if $\mathbf{A}_{11}$ and $\mathbf{A}_{22}$ are symmetric and

$$\mathbf{A}_{21} = \mathbf{A}_{12}^T$$

Partitioned Mean Vectors

If

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$$

is a partitioned random vector, then its mean vector

$$E(\mathbf{X}) = E\left\{\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}\right\} = \begin{pmatrix} E(\mathbf{X}_1) \\ E(\mathbf{X}_2) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} = \boldsymbol{\mu}$$

is a vector partitioned in the same way.

Covariance Matrices

If $\mathbf{X}_1$ and $\mathbf{X}_2$ are random vectors, with mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, then

$$\text{cov}(\mathbf{X}_1, \mathbf{X}_2) = E\{(\mathbf{X}_1 - \boldsymbol{\mu}_1)(\mathbf{X}_2 - \boldsymbol{\mu}_2)^T\}$$

is called the *covariance matrix* of $\mathbf{X}_1$ and $\mathbf{X}_2$.

Note well that, unlike the scalar case, the matrix covariance operator is **not** symmetric in its arguments

$$\text{cov}(\mathbf{X}_2, \mathbf{X}_1) = \text{cov}(\mathbf{X}_1, \mathbf{X}_2)^T$$

Partitioned Variance Matrices

If

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$$

is a partitioned random vector, then its variance matrix

$$\begin{aligned} \text{var}(\mathbf{X}) &= \text{var}\left\{ \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \right\} \\ &= \begin{pmatrix} \text{var}(\mathbf{X}_1) & \text{cov}(\mathbf{X}_1, \mathbf{X}_2) \\ \text{cov}(\mathbf{X}_2, \mathbf{X}_1) & \text{var}(\mathbf{X}_2) \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{pmatrix} \\ &= \mathbf{M} \end{aligned}$$

is a square matrix partitioned in the same way.

Multivariate Normal Marginal Distributions (cont.)

If

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$$

is a partitioned multivariate normal random vector having mean vector

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}$$

and variance matrix

$$\text{var}(\mathbf{X}) = \begin{pmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{pmatrix}$$

then the marginal distribution of $\mathbf{X}_1$ is $\mathcal{N}(\boldsymbol{\mu}_1, \mathbf{M}_{11})$.

Partitioned Matrices (cont.)

Matrix multiplication of partitioned matrices looks much like ordinary matrix multiplication. Just think of the blocks as scalars.

$$\begin{aligned}\mathbf{AB} &= \begin{pmatrix}\mathbf{A}_{11} & \mathbf{A}_{12}\\ \mathbf{A}_{21} & \mathbf{A}_{22}\end{pmatrix}\begin{pmatrix}\mathbf{B}_{11} & \mathbf{B}_{12}\\ \mathbf{B}_{21} & \mathbf{B}_{22}\end{pmatrix}\\ &= \begin{pmatrix}\mathbf{A}_{11}\mathbf{B}_{11}+\mathbf{A}_{12}\mathbf{B}_{21} & \mathbf{A}_{11}\mathbf{B}_{12}+\mathbf{A}_{12}\mathbf{B}_{22}\\ \mathbf{A}_{21}\mathbf{B}_{11}+\mathbf{A}_{22}\mathbf{B}_{21} & \mathbf{A}_{21}\mathbf{B}_{12}+\mathbf{A}_{22}\mathbf{B}_{22}\end{pmatrix}\end{aligned}$$

Partitioned Matrices (cont.)

A partitioned matrix

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$$

is called *block diagonal* if the off-diagonal blocks are zero, that is, $\mathbf{A}_{12} = 0$ and $\mathbf{A}_{21} = 0$.

A partitioned variance matrix

$$\text{var}\left\{\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}\right\} = \begin{pmatrix} \text{var}(\mathbf{X}_1) & \text{cov}(\mathbf{X}_1, \mathbf{X}_2) \\ \text{cov}(\mathbf{X}_2, \mathbf{X}_1) & \text{var}(\mathbf{X}_2) \end{pmatrix}$$

is block diagonal if and only if $\text{cov}(\mathbf{X}_2, \mathbf{X}_1) = \text{cov}(\mathbf{X}_1, \mathbf{X}_2)^T$ is zero, in which case we say the random vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ are *uncorrelated*.

Uncorrelated versus Independent

As in the scalar case, uncorrelated does not imply independent except in the special case of joint multivariate normality. We now show that if $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are jointly multivariate normal, meaning the partitioned random vector

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix}$$

is multivariate normal, then $\text{cov}(\mathbf{Y}_1, \mathbf{Y}_2) = 0$ implies $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are independent random vectors, meaning

$$E\{h_1(\mathbf{Y}_1)h_2(\mathbf{Y}_2)\} = E\{h_1(\mathbf{Y}_1)\}E\{h_2(\mathbf{Y}_2)\}$$

for any functions h_1 and h_2 such that the expectations exist.

Uncorrelated versus Independent (cont.)

It follows from the formula for matrix multiplication of partitioned matrices that, if

$$\mathbf{M} = \begin{pmatrix} \mathbf{M}_{11} & 0 \\ 0 & \mathbf{M}_{22} \end{pmatrix}$$

and $\mathbf{M}$ is positive definite, then

$$\mathbf{M}^{-1} = \begin{pmatrix} \mathbf{M}_{11}^{-1} & 0 \\ 0 & \mathbf{M}_{22}^{-1} \end{pmatrix}$$

and

$$\begin{aligned}(\mathbf{y}-\boldsymbol{\mu})^T\mathbf{M}^{-1}(\mathbf{y}-\boldsymbol{\mu}) = (\mathbf{y}_1-\boldsymbol{\mu}_1)^T\mathbf{M}_{11}^{-1}(\mathbf{y}_1-\boldsymbol{\mu}_1) \\ + (\mathbf{y}_2-\boldsymbol{\mu}_2)^T\mathbf{M}_{22}^{-1}(\mathbf{y}_2-\boldsymbol{\mu}_2)\end{aligned}$$

where $\mathbf{y}$ and $\boldsymbol{\mu}$ are partitioned like $\mathbf{M}$.

Uncorrelated versus Independent (cont.)

$$
\begin{aligned}
f(\mathbf{y}) &= (2\pi)^{-n/2}\det(\mathbf{M})^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T\mathbf{M}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right) \\
&\propto \exp\Big(-\frac{1}{2}(\mathbf{y}_1-\boldsymbol{\mu}_1)^T\mathbf{M}_{11}^{-1}(\mathbf{y}_1-\boldsymbol{\mu}_1) \\
&\qquad -\frac{1}{2}(\mathbf{y}_2-\boldsymbol{\mu}_2)^T\mathbf{M}_{22}^{-1}(\mathbf{y}_2-\boldsymbol{\mu}_2)\Big) \\
&= \exp\Big(-\frac{1}{2}(\mathbf{y}_1-\boldsymbol{\mu}_1)^T\mathbf{M}_{11}^{-1}(\mathbf{y}_1-\boldsymbol{\mu}_1)\Big) \\
&\qquad \times \exp\Big(-\frac{1}{2}(\mathbf{y}_2-\boldsymbol{\mu}_2)^T\mathbf{M}_{22}^{-1}(\mathbf{y}_2-\boldsymbol{\mu}_2)\Big)
\end{aligned}
$$

Since this is a function of $\mathbf{y}_1$ times a function of $\mathbf{y}_2$, the random vectors $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are independent.

Uncorrelated versus Independent (cont.)

We have now proved that if the blocks of the nondegenerate multivariate normal random vector

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix}$$

are uncorrelated, then they are independent.

If $\mathbf{Y}_1$ is degenerate and $\mathbf{Y}_2$ nondegenerate, we can partition $\mathbf{Y}_1$ into a nondegenerate block $\mathbf{Y}_3$ and a linear function of $\mathbf{Y}_3$, so

$$\mathbf{Y} = \begin{pmatrix} \mathbf{d}_3 + \mathbf{B}_3\mathbf{Y}_3 \\ \mathbf{Y}_3 \\ \mathbf{Y}_2 \end{pmatrix}$$

If $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are uncorrelated, then $\mathbf{Y}_3$ and $\mathbf{Y}_2$ are uncorrelated, hence independent, and that implies the independence of $\mathbf{Y}_1$ and $\mathbf{Y}_2$ (because $\mathbf{Y}_1$ is a function of $\mathbf{Y}_3$).

Uncorrelated versus Independent (cont.)

Similarly if $\mathbf{Y}_2$ is degenerate and $\mathbf{Y}_1$ nondegenerate,

If $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are both degenerate we can partition $\mathbf{Y}_1$ as before and partition $\mathbf{Y}_2$ similarly so

$$\mathbf{Y} = \begin{pmatrix} \mathbf{d}_3 + \mathbf{B}_3\mathbf{Y}_3 \\ \mathbf{Y}_3 \\ \mathbf{d}_4 + \mathbf{B}_4\mathbf{Y}_4 \\ \mathbf{Y}_4 \end{pmatrix}$$

If $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are uncorrelated, then $\mathbf{Y}_3$ and $\mathbf{Y}_4$ are uncorrelated, hence independent, and that implies the independence of $\mathbf{Y}_1$ and $\mathbf{Y}_2$ (because $\mathbf{Y}_1$ is a function of $\mathbf{Y}_3$ and $\mathbf{Y}_2$ is a function of $\mathbf{Y}_4$).

Uncorrelated versus Independent (cont.)

And that finishes all cases of the proof that, if $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are random vectors that are jointly multivariate normal and uncorrelated, then they are independent.

Multivariate Normal Conditional Distributions

Suppose

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$$

is a partitioned multivariate normal random vector and

$$E(\mathbf{X}) = \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}$$

$$\text{var}(\mathbf{X}) = \mathbf{M} = \begin{pmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{pmatrix}$$

and $\mathbf{X}_2$ is nondegenerate, then

$$(\mathbf{X}_1 - \boldsymbol{\mu}_1) - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2)$$

is independent of $\mathbf{X}_2$.

Multivariate Normal Conditional Distributions (cont.)

The proof uses uncorrelated implies independent for multivariate normal.

$$
\begin{aligned}
&\text{cov}\{\mathbf{X}_2, (\mathbf{X}_1 - \boldsymbol{\mu}_1) - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2)\} \\
&\qquad = E\{[\mathbf{X}_2 - \boldsymbol{\mu}_2][(\mathbf{X}_1 - \boldsymbol{\mu}_1) - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2)]^T\} \\
&= E\{(\mathbf{X}_2 - \boldsymbol{\mu}_2)(\mathbf{X}_1 - \boldsymbol{\mu}_1)^T\} - E\{(\mathbf{X}_2 - \boldsymbol{\mu}_2)[\mathbf{M}_{12}\mathbf{M}_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2)]^T\} \\
&= E\{(\mathbf{X}_2 - \boldsymbol{\mu}_2)(\mathbf{X}_1 - \boldsymbol{\mu}_1)^T\} - E\{(\mathbf{X}_2 - \boldsymbol{\mu}_2)(\mathbf{X}_2 - \boldsymbol{\mu}_2)^T\mathbf{M}_{22}^{-1}\mathbf{M}_{12}^T\} \\
&= E\{(\mathbf{X}_2 - \boldsymbol{\mu}_2)(\mathbf{X}_1 - \boldsymbol{\mu}_1)^T\} - E\{(\mathbf{X}_2 - \boldsymbol{\mu}_2)(\mathbf{X}_2 - \boldsymbol{\mu}_2)^T\}\mathbf{M}_{22}^{-1}\mathbf{M}_{12}^T \\
&\qquad = \text{cov}(\mathbf{X}_2, \mathbf{X}_1) - \text{cov}(\mathbf{X}_2, \mathbf{X}_2)\mathbf{M}_{22}^{-1}\mathbf{M}_{12}^T \\
&\qquad = \mathbf{M}_{21} - \mathbf{M}_{22}\mathbf{M}_{22}^{-1}\mathbf{M}_{12}^T \\
&\qquad = \mathbf{M}_{21} - \mathbf{M}_{12}^T \\
&\qquad = 0
\end{aligned}
$$

Multivariate Normal Conditional Distributions (cont.)

Thus, conditional on $\mathbf{X}_2$, the conditional distribution of

$$(\mathbf{X}_1 - \boldsymbol{\mu}_1) - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2)$$

is the same as its marginal distribution, which is multivariate normal with mean vector zero and variance matrix

$$\begin{aligned}
&\operatorname{var}(\mathbf{X}_1) - \operatorname{cov}(\mathbf{X}_1, \mathbf{X}_2)\mathbf{M}_{22}^{-1}\mathbf{M}_{12}^T - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\operatorname{cov}(\mathbf{X}_2, \mathbf{X}_1) \\
&\qquad + \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\operatorname{var}(\mathbf{X}_2)\mathbf{M}_{22}^{-1}\mathbf{M}_{12}^T \\
&= \mathbf{M}_{11} - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{21} - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{21} \\
&\qquad + \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{22}\mathbf{M}_{22}^{-1}\mathbf{M}_{21} \\
&= \mathbf{M}_{11} - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{21}
\end{aligned}$$

Multivariate Normal Conditional Distributions (cont.)

Since

$$(\mathbf{X}_1 - \boldsymbol{\mu}_1) - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2)$$

is independent of $\mathbf{X}_2$, its expectation conditional on $\mathbf{X}_2$ is the same as its unconditional expectation, which is zero.

$$\begin{aligned} 0 &= E\{(\mathbf{X}_1 - \boldsymbol{\mu}_1) - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2) \mid \mathbf{X}_2\} \\ &= E(\mathbf{X}_1 \mid \mathbf{X}_2) - \boldsymbol{\mu}_1 - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2) \end{aligned}$$

(because functions of $\mathbf{X}_2$ behave like constants in the conditional expectation). Hence

$$E(\mathbf{X}_1 \mid \mathbf{X}_2) = \boldsymbol{\mu}_1 + \mathbf{M}_{12}\mathbf{M}_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2)$$

Multivariate Normal Conditional Distributions (cont.)

Thus we have proved that, in the case where $\mathbf{X}_2$ is nondegenerate, the conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_2$ is multivariate normal with

$$\begin{aligned} E(\mathbf{X}_1 \mid \mathbf{X}_2) &= \boldsymbol{\mu}_1 + \mathbf{M}_{12}\mathbf{M}_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2) \\ \text{var}(\mathbf{X}_1 \mid \mathbf{X}_2) &= \mathbf{M}_{11} - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{21} \end{aligned}$$

because, conditional on $\mathbf{X}_2$, the distribution of $\mathbf{X}_1 - E(\mathbf{X}_1 \mid \mathbf{X}_2)$ is multivariate normal and functions of $\mathbf{X}_2$ behave like constants, so $\mathbf{X}_1$ is a linear function of multivariate normal, hence multivariate normal.

It is important, although we will not use it until next semester, that the conditional expectation is a linear function of the variables behind the bar and the conditional variance is a constant function of the variables behind the bar.

Multivariate Normal Conditional Distributions (cont.)

In case $\mathbf{X}_2$ is degenerate, we can partition it

$$\mathbf{X}_2 = \begin{pmatrix} \mathbf{d} + \mathbf{B}\mathbf{X}_3 \\ \mathbf{X}_3 \end{pmatrix}$$

where $\mathbf{X}_3$ is nondegenerate. Conditioning on $\mathbf{X}_2$ is the same as conditioning on $\mathbf{X}_3$, because fixing $\mathbf{X}_3$ also fixes $\mathbf{X}_2$.

Hence the conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_2$ is the same as the conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_3$, which is multivariate normal with mean vector

$$E(\mathbf{X}_1 \mid \mathbf{X}_3) = \boldsymbol{\mu}_1 + \mathbf{M}_{13}\mathbf{M}_{33}^{-1}(\mathbf{X}_3 - \boldsymbol{\mu}_3)$$

and variance matrix

$$\operatorname{var}(\mathbf{X}_1 \mid \mathbf{X}_3) = \mathbf{M}_{11} - \mathbf{M}_{13}\mathbf{M}_{33}^{-1}\mathbf{M}_{31}$$