BST 675 – Fall 2011 – Dr. Charnigo

Unit IV: Multiple Random Variables

a. Motivating Case Study #1: Modeling a distribution of blood pressure and cholesterol scores

Let X denote systolic blood pressure and Y denote total serum cholesterol. Suppose that, within a certain population, X is normally distributed with mean $\mu_X := 130$ and standard deviation $\sigma_X := 15$. Suppose, moreover, that Y is normally distributed with mean $\mu_Y := 180$ and standard deviation $\sigma_Y := 25$.

High blood pressure and high cholesterol are both considered risk factors for heart disease. If we agree to define high blood pressure as greater than 140 and high cholesterol as greater than 200, then let us consider the question of how many people in the population have at least one of these risk factors.

From Unit I we recall that

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

for generic events A and B. Letting $A := \{X > 140\}$ and $B := \{Y > 200\}$, we have in the present example

$$P(\text{at least one risk factor}) = P(X > 140) + P(Y > 200) - P(X > 140 \cap Y > 200).$$

We can calculate $P(X > 140)$ and $P(Y > 200)$ directly in R using the code

```
1-pnorm(140,mean=130,sd=15)
1-pnorm(200,mean=180,sd=25)
```

or we can employ a Z table such as appears in the back of an introductory methods textbook:

$$P(X > 140) = P(Z > (140-130)/15)) = P(Z > 0.667) = 1-P(Z \leq 0.667) \quad \text{and}$$

$$P(Y > 200) = P(Z > (200 - 180)/25)) = P(Z > 0.800) = 1 - P(Z \leq 0.800).$$

Either way, we get 0.252 and 0.212.

But now we must evaluate $P(X > 140 \ \cap \ Y > 200)$. If X and Y were independent, we would have

$$P(X > 140 \ \cap \ Y > 200) = P(X > 140)P(Y > 200) = 0.053$$

and then

$$P(\text{at least one risk factor}) = 0.252 + 0.212 - 0.053 = 0.411.$$

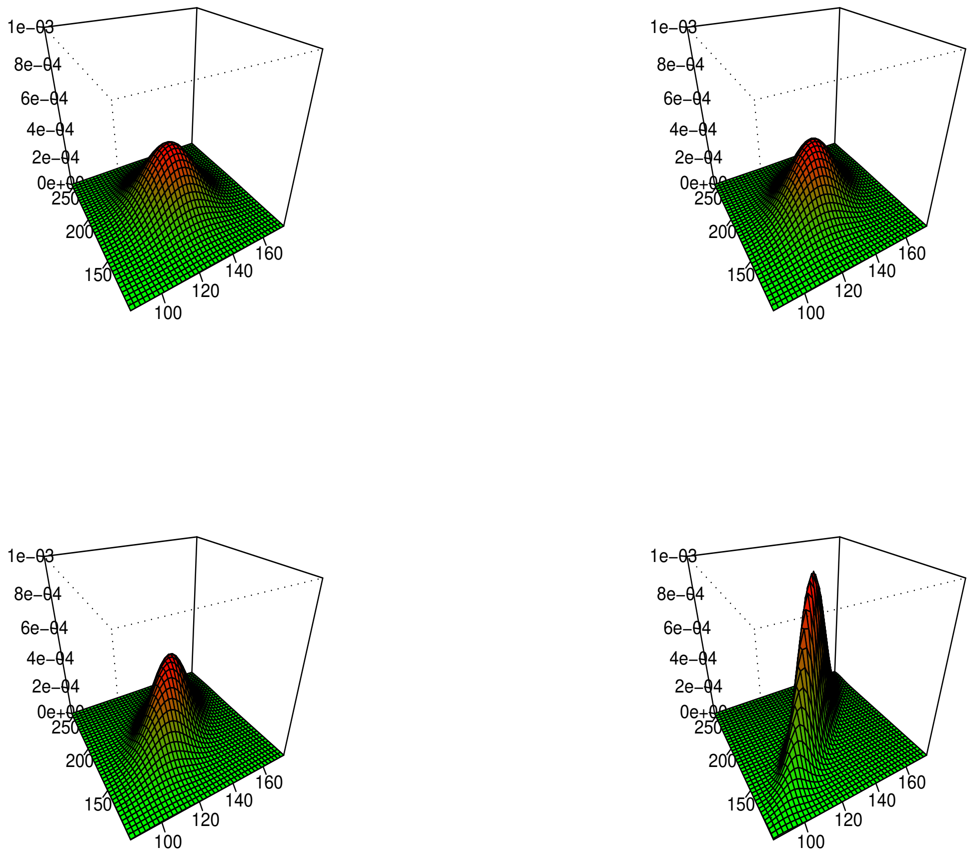
Yet, an independence assumption seems unrealistic. Indeed, we would anticipate that high blood pressure would occur disproportionately often among people with high cholesterol and vice versa. As such, $P(X > 140 \ \cap Y \ > 200)$ should be [choose one: lower, higher] than 0.053 and P(at least one risk factor) should be [choose one: lower, higher] than 0.411.

Apart from saying that our answer for P(at least one risk factor) should be shifted by some amount no more than 0.159, because $0.053 + 0.159 = 0.212$ and $P(X > 140 \ \cap Y \ > 200)$ cannot be larger than $P(Y > 200)$, we cannot really do much more with the information given. Indeed, we need another piece of information: how strongly are X and Y correlated?

As you know from your introductory methods course, correlation is a measure of linear association between two continuous random variables. We will revisit the definition of correlation later in Unit IV and address its implications both for calculating probabilities involving normal distributions and for calculating variances of sums of random variables. For now, I leave you with two remarks.

First, many people speak loosely and use correlation as a synonym for association. For instance, they may write in their manuscripts that "clinical outcomes such as death correlate with treatment". When you are collaborating with them on manuscripts, you will have an opportunity to help them avoid such imprecise statements. I encourage you not to settle for the status quo: bring the manuscripts on which you are participating to a higher standard, not just in the analyses employed but also in the precision of the statements describing the analyses and their results! (Another pet peeve of mine is a statement such as "OR = 1.49 [95% CI 1.23 to 1.81]". What's wrong with that?)

Second, we know that the distribution of X can be visualized in terms of its probability density function, a bell curve, and similarly for Y. How can we visualize the distribution of X and the distribution of Y simultaneously or, if you will, the "joint distribution" of X and Y? Instead of portraying the probability that X falls in a certain interval or the probability that Y falls in a certain interval as the area under a curve, we can portray the probability that X and Y together fall in a certain region as the volume under a surface known as the "joint probability density function" of X and Y. The figure below illustrates this surface in four cases: zero correlation between X and Y, correlation of 0.3, correlation of 0.6, and correlation of 0.9. A formal mathematical treatment of joint probability density functions will, however, require the concept of a "double integral". We will develop that concept in part c and then resolve this motivating case study in part h.



b. Motivating Case Study #2: Finding an appropriate nonlinear transformation in regression

Recall the simple linear regression model

$$Y_i = \alpha + \beta x_i + \epsilon_i \quad \text{for } i \in \{1, \ldots, n\}$$

from your introductory methods course. The error terms $\epsilon_1, \ldots, \epsilon_n$ are assumed to be normally distributed with mean 0 and common variance $\sigma^2 \in (0, \infty)$. (In practice, why do we care whether this assumption holds?) As such, given the values $x_1, \ldots, x_n$, we should also have that $Y_1, \ldots, Y_n$ are normally distributed with different means $\alpha + \beta x_1, \ldots, \alpha + \beta x_n$ but common variance σ^2.

If the values $x_1, \ldots, x_n$ themselves conform to a bell curve, then despite having different means $Y_1, \ldots, Y_n$ should also conform to a bell curve. Hence, a simple diagnostic for non-normality of the error terms in linear regression is to examine a histogram of $Y_1, \ldots, Y_n$ and note whether it conforms to a bell curve. If not, then you reach one of two conclusions: (i) the values $x_1, \ldots, x_n$ have a strange configuration; or, (ii) the error terms have a strange distribution.

If your conclusion is (ii), then you can employ a robust regression technique that does not assume normally distributed errors (Cf. Lecture 3 of CPH 931, Fall 2009). Alternatively, you can attempt a nonlinear transformation of $Y_1, \ldots, Y_n$. For example, if $Y_1, \ldots, Y_n$ are positive, then you can let $W_i := \log[Y_i]$ for $i \in \{1, \ldots, n\}$ and then consider a modified linear regression model

$$W_i = \alpha^* + \beta^* x_i + \epsilon_i^* \quad \text{for } i \in \{1, \ldots, n\}.$$

With a "good" transformation, the error terms $\epsilon_1^*, \ldots, \epsilon_n^*$ in the modified linear regression model may end up being approximately normally distributed. But with a "bad" transformation, you may be even worse off than before! (Given this challenge, why do data analysts bother with transformations at all? Why do they not automatically resort to robust regression techniques?)

So, you may ask, what sort of transformation is anticipated to work well under a given set of circumstances? We will return to that question when we resolve this motivating case study in part h.

c. Math tutorial: partial derivatives and double integrals

Consider a function of one variable, say $y = F(x)$. Recall that the derivative of $F(x)$ is defined as

$$\lim_{\delta \to 0} \frac{F(x+\delta) - F(x)}{\delta},$$

provided that the limit exists. (The derivative is said not to exist if the limit does not exist.) The derivative will, in general, depend on x. Hence, the derivative is itself a function of x, which we may denote $F'(x)$. Sometimes we also write $\frac{dy}{dx}$ or $\frac{d}{dx}F(x)$ for the derivative.

In your first calculus course, after a few painful exercises in which you directly used the limit definition to calculate derivatives, you eventually learned several practical techniques for calculating derivatives. These included general principles such as the product rule and the chain rule as well as specific formulas like $\frac{d}{dx}x^n = nx^{n-1}$ and $\frac{d}{dx}\exp[cx] = c\exp[cx]$.

Now consider a function of two variables, say $z = F(x, y)$. The partial derivative of $F(x, y)$ with respect to x is defined as

$$\lim_{\delta \to 0} \frac{F(x+\delta, y) - F(x, y)}{\delta},$$

provided that the limit exists. The partial derivative will, in general, depend on both x and y. Hence, the partial derivative is itself a function of x and y, which we may denote $F_x(x, y)$ or $\frac{\partial z}{\partial x}$ or $\frac{\partial}{\partial x}F(x, y)$. The partial derivative of $F(x, y)$ with respect to y is similarly defined as

$$\lim_{\delta \to 0} \frac{F(x, y+\delta) - F(x, y)}{\delta}$$

and denoted $F_y(x, y)$ or $\frac{\partial z}{\partial y}$ or $\frac{\partial}{\partial y}F(x, y)$.

Practically speaking, taking a partial derivative with respect to x (resp., y) is not really any different from taking a derivative as you did in your first calculus course. Just treat y (resp., x) as fixed. For example, suppose that $z = F(x, y) := 2x^3y^2 - x^4y^3$. Then

$\frac{\partial}{\partial x}F(x, y) =$ and $\frac{\partial}{\partial y}F(x, y) =$

We can also differentiate partial derivatives themselves. The partial derivative of $F_x(x, y)$ with respect to x is defined as

$$\lim_{\delta \to 0} \frac{F_x(x+\delta, y) - F_x(x, y)}{\delta}$$

and denoted $F_{xx}(x, y)$ or $\frac{\partial^2 z}{\partial x^2}$ or $\frac{\partial^2}{\partial x^2} F(x, y)$, while the partial derivative of $F_x(x, y)$ with respect to y is defined as

$$\lim_{\delta \to 0} \frac{F_x(x, y+\delta) - F_x(x, y)}{\delta}$$

and denoted $F_{xy}(x, y)$ or $\frac{\partial^2 z}{\partial y \partial x}$ or $\frac{\partial^2}{\partial y \partial x} F(x, y)$.

The partial derivatives of $F_y(x, y)$ are defined analogously. The differentiated partial derivatives are sometimes called second-order partial derivatives, while the original partial derivatives are called first-order partial derivatives. Of course, we can also define third- and higher-order partial derivatives.

Continuing from the preceding example, we have

$\frac{\partial^2}{\partial x^2} F(x, y) =$ and $\frac{\partial^2}{\partial y \partial x} F(x, y) =$

as well as

$\frac{\partial^2}{\partial x \partial y} F(x, y) =$ and $\frac{\partial^2}{\partial y^2} F(x, y) =$

Consider a function of one variable, say $y = f(x)$. Let $a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$ be real numbers. A corresponding Riemann sum of $f(x)$ over $[a, b]$ is

$$\sum_{i=1}^{n} f(x_i)(x_i - x_{i-1}) = \sum_{i=1}^{n} f(x_i) \mathit{length}(I_i),$$

where I_i is defined as the interval $[x_{i-1}, x_i]$. Let $n \to \infty$ with $\max_{i \in \{1,\ldots,n\}} \mathit{length}(I_i) \to 0$. Under fairly general conditions, a sequence of Riemann sums of $f(x)$ over $[a, b]$ will converge to a number, denoted

$$\int_a^b f(x)\, dx$$

and called the (Riemann) integral of $f(x)$ over $[a, b]$. Geometrically, the integral of $f(x)$ over $[a, b]$ is the (net signed) area under the curve of $f(x)$ for $x \in [a, b]$, while a Riemann sum is an approximation to that area based on rectangles of height $f(x_i)$ and width $length(I_i)$.

In your first calculus course, you learned that (often) the best way to evaluate $\int_a^b f(x)\, dx$ is not to appeal directly to the definition as a limit of Riemann sums but rather to invoke the Fundamental Theorem of Calculus. More specifically, if you can find a function $F(x)$ such that $F'(x) = f(x)$, then $\int_a^b f(x)\, dx = F(b) - F(a)$. (Integrals over infinite intervals, say $\int_a^\infty f(x)\, dx$, can also be defined as limits of integrals over finite intervals, say $\lim_{b\to\infty} \int_a^b f(x)\, dx$.)

Now consider a function of two variables, say $z = f(x, y)$. Let R be a subset of $\mathbb{R}^2$ with finite area that does not have any "holes" in it. Mathematicians call such a subset "simply connected". One example is a rectangle such as $[a, b] \times [c, d]$ for real numbers $a < b$ and $c < d$. Let R itself be partitioned into n smaller simply connected subsets, which we will call $R_1, \ldots, R_n$. A corresponding Riemann sum of $f(x, y)$ over R is

$$\sum_{i=1}^{n} f(x_i, y_i)\, area(R_i),$$

where (x_i, y_i) is any point inside or on the boundary of R_i. Let $n \to \infty$ with $\max_{i\in\{1,\ldots,n\}} area(R_i) \to 0$. Under fairly general conditions, a sequence of Riemann sums of $f(x, y)$ over R will converge to a number, denoted

$$\int\int_R f(x, y)\, dx\, dy \tag{1}$$

and called the (Riemann) integral of $f(x, y)$ over R. Geometrically, the integral of $f(x, y)$ over R is the (net signed) volume under the surface of $f(x, y)$ for $(x, y) \in R$. (Integrals over regions with infinite area can also be defined as limits of integrals over regions with finite area.)

A practical procedure for evaluating (1) entails expressing R, if possible, in one of three ways:

1. R consists of all points (x, y) satisfying $a \le x \le b$ and $c \le y \le d$;

2. R consists of all points (x, y) satisfying $a \le x \le b$ and $c(x) \le y \le d(x)$; or,

3. R consists of all points (x, y) satisfying $a(y) \le x \le b(y)$ and $c \le y \le d$.

Above, quantities like a, $a(y)$, etc., are either constants or functions of the indicated variable. In the first case, we have

$$\int\int_R f(x,y)\;dx\;dy = \int_a^b \left[\int_c^d f(x,y)\;dy\right]\;dx = \int_c^d \left[\int_a^b f(x,y)\;dx\right]\;dy.$$

In the second case, we have

$$\int\int_R f(x,y)\;dx\;dy = \int_a^b \left[\int_{c(x)}^{d(x)} f(x,y)\;dy\right]\;dx.$$

In the third case, we have

$$\int\int_R f(x,y)\;dx\;dy = \int_c^d \left[\int_{a(y)}^{b(y)} f(x,y)\;dx\right]\;dy.$$

Thus, evaluating the "double integral" (1) reduces to evaluating two "single integrals" in succession.

For example, let us evaluate (1) with $R := \{(x, y) : x \ge 0, y \ge 0, x + y \le 1\}$ and $f(x, y) := \exp[(1 - x)^4]y^2$. Noting that R consists of all points (x, y) satisfying $0 \le x \le 1$ and $0 \le y \le (1 - x)$, we compute

$$\begin{aligned}
\int\int_R f(x,y)\;dx\;dy &= \int_0^1 \left[\int_0^{(1-x)} \exp[(1-x)^4]y^2\;dy\right]\;dx \\
&= \int_0^1 \left[\exp[(1-x)^4]\int_0^{(1-x)} y^2\;dy\right]\;dx \\
&= \int_0^1 \exp[(1-x)^4](1-x)^3/3\;dx \\
&= -(1/12)\int_0^1 -\exp[(1-x)^4]4(1-x)^3\;dx \\
&= (1/12)(\exp[1]-1).
\end{aligned}$$

Note that, in this example, we could also express R as consisting of all points (x, y) satisfying $0 \le x \le (1 - y)$ and $0 \le y \le 1$. However, we would become stuck if we attempted to integrate in dx first.

d. Math tutorial: vector and matrix computations

A matrix is just a two-dimensional array of real numbers, such as

$$\begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \\ 1 & 4 & 9 & 16 \end{bmatrix}.$$

A matrix having more than one column is typically denoted with a capital Roman letter in bold font such as $\mathbf{A}$. The individual entries of a matrix $\mathbf{A}$ may be represented symbolically with doubly subscripted lower case Roman letters such as a_{21} for the entry in the second row and first column. A matrix with m rows and n columns is referred to as an $m \times n$ matrix.

The transpose of a matrix $\mathbf{A}$, denoted $\mathbf{A}'$ or $\mathbf{A}^T$, is obtained by reversing the roles of the rows and columns in $\mathbf{A}$: the first row of $\mathbf{A}$ becomes the first column of $\mathbf{A}^T$, the first column of $\mathbf{A}$ becomes the first row of $\mathbf{A}^T$, the second row of $\mathbf{A}$ becomes the second column of $\mathbf{A}^T$, and so forth. A matrix equal to its own transpose is called symmetric. Necessarily, any symmetric matrix has the same numbers of rows and columns. A matrix with the same numbers of rows and columns is called square. Thus, all symmetric matrices are square, although the converse is not true. (What are the transposes of the matrices given at the top of this page? Are these matrices square? Are they symmetric?)

A matrix having just one column is called a vector and is typically denoted with a lower case Roman letter in bold font such as $\mathbf{b}$. The individual entries of a vector $\mathbf{b}$ may be represented symbolically with singly subscripted lower case Roman letters such as b_2 for the entry in the second row. Because a one-column matrix consumes much space on the printed page, we sometimes express a vector as the transpose of a one-row matrix. For instance, $(-1, 1)^T$ represents

$$\begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

The sum of two $m \times n$ matrices is an $m \times n$ matrix whose entries are the sums of the corresponding entries in the matrices being added. For example,

$$\begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 3 & -1 \\ -1 & 2 \end{bmatrix} = \begin{bmatrix} 4 & -2 \\ -1 & 3 \end{bmatrix}.$$

The product of a single real number (often called a "scalar") with an $m \times n$ matrix $\mathbf{A}$ is an $m \times n$ matrix whose entries are the products of the real number with the corresponding entries in $\mathbf{A}$. For example,

$$2\begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & -2 \\ 0 & 2 \end{bmatrix}.$$

An $m \times n$ matrix $\mathbf{A}$ can be multiplied by an $n \times p$ matrix $\mathbf{B}$. We write this product by juxtaposition with no intervening multiplication symbol, $\mathbf{AB}$. The product is an $m \times p$ matrix $\mathbf{C}$ such that $c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}$. In words, the entries in row i of matrix $\mathbf{A}$ are multiplied by the entries in column j of matrix $\mathbf{B}$, and then these products are added. For example,

$$\begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} (1 \times 3) + (-1 \times 2) \\ (0 \times 3) + (1 \times 2) \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix},$$

while

$$\begin{bmatrix} 3 \\ 2 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}$$

is undefined because the number of columns in the first matrix is not the same as the number of rows in the second matrix. Thus, unlike ordinary multiplication, matrix multiplication is not in general commutative.

For the rest of part d in Unit IV, all matrices (except vectors denoted in lower case Roman letters) are assumed to be square. We say that $\mathbf{A}$ is invertible if there exists a matrix $\mathbf{B}$ such that $\mathbf{AB} = \mathbf{BA} = \mathbf{I}$, where $\mathbf{I}$ is defined to be the identity matrix with 1 entries on the diagonal and 0 entries elsewhere. For example, the 2×2 identity matrix is

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The identity matrix is so named because $\mathbf{IC} = \mathbf{CI} = \mathbf{C}$ for any matrix $\mathbf{C}$. The matrix $\mathbf{B}$ referred to above is then called the inverse of $\mathbf{A}$ and denoted $\mathbf{A}^{-1}$. A matrix $\mathbf{A}$ that does not have an inverse is called singular.

A question of both theoretical and practical importance is: when does a matrix $\mathbf{A}$ have an inverse? If $\mathbf{A}$ is 2×2, then the answer is simple: an inverse exists if and only if $a_{11}a_{22} - a_{12}a_{21} \neq 0$, in which case the inverse is given by

$$\left(\frac{1}{a_{11}a_{22} - a_{12}a_{21}}\right)\begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}.$$

For matrices larger than 2×2, matters are much more complicated. Even when they do exist, inverses are not readily computed by pencil and paper. Software (for example, R with its "solve" command) can be helpful, with certain caveats discussed in texts on numerical analysis. So, when do inverses exist for matrices larger than 2×2? To address this question, we need to introduce the concepts of eigenvalues and eigenvectors.

Suppose that we can find a nonzero vector $\mathbf{x}$ (i.e., a vector with at least one nonzero entry) such that $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ for some real number λ. We refer to $\mathbf{x}$ as an eigenvector of $\mathbf{A}$ and to λ as an eigenvalue of $\mathbf{A}$. For example, we have

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = 3\begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}\begin{bmatrix} 1 \\ -1 \end{bmatrix} = 1\begin{bmatrix} 1 \\ -1 \end{bmatrix},$$

so that the 2×2 matrix with first row $(2, 1)$ and second row $(1, 2)$ has two eigenvalues 3 and 1 corresponding to eigenvectors $(1, 1)^T$ and $(1, -1)^T$. Note that neither one of these two eigenvectors is a scalar multiple of the other.

More generally, we say that a set of eigenvectors is linearly independent if no one eigenvector in the set can be expressed as a linear combination of the others. A famous theorem of linear algebra is that any symmetric $m \times m$ matrix has m eigenvalues corresponding to m linearly independent eigenvectors, although the eigenvalues are not necessarily distinct. For example, the 3×3 identity matrix has three copies of 1 as its eigenvalues corresponding to the linearly independent eigenvectors $(1, 0, 0)^T$, $(0, 1, 0)^T$, and $(0, 0, 1)^T$.

For the balance of part d in Unit IV, we now make the further assumption that all matrices (except vectors denoted in lower case Roman letters) are not only square but also symmetric. The determinant of a matrix $\mathbf{A}$ is the product of its eigenvalues. (Some people define the determinant differently and then state this as a theorem, but I find it easier just to take this as the definition.) A famous theorem from linear algebra states that a matrix has an inverse if and only if its determinant is nonzero. An obvious corollary is that a matrix has an inverse if and only if all of its eigenvalues are nonzero. (So, when numerical analysts say that a matrix is near singular, what they mean is that it has an inverse but that at least one of the eigenvalues is very close to zero.)

We say that a matrix $\mathbf{A}$ is nonnegative definite if $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$ for any vector $\mathbf{x}$. We say that a matrix $\mathbf{A}$ is positive definite if $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$ for any nonzero vector $\mathbf{x}$. (Why require that $\mathbf{x}$ be nonzero?) For example, the 2×2 identity matrix is positive definite because $\mathbf{x}^T\mathbf{I}\mathbf{x} = x_1^2 + x_2^2 > 0$ whenever $\mathbf{x} \neq (0,0)^T$. Yet another famous theorem from linear algebra states that a positive definite matrix has all positive eigenvalues. An obvious corollary is that a positive definite matrix has an inverse.

Here are two (among many) applications of vector and matrix computations:

1. Let $\mathbf{y} := (Y_1, \ldots, Y_n)^T$, $\mathbf{b} := (\beta_0, \beta_1, \ldots, \beta_p)^T$, and $\mathbf{e} := (\epsilon_1, \ldots, \epsilon_n)^T$. Also, let $\mathbf{X}$ be an $n \times (p+1)$ matrix whose j^{th} row is $(1, x_{j1}, x_{j2}, \ldots, x_{jp})$. The linear regression model $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$ can be represented compactly using vector and matrix computations as $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$. More importantly, the solution to the least squares problem can be represented compactly as $\widehat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, assuming that the inverse of $\mathbf{X}^T\mathbf{X}$ exists.

2. We define the covariance between two random variables X and Y (here I am using capital letters — but not in bold font — to represent random variables, not matrices) to be $E[(X - E[X])(Y - E[Y])]$, assuming that all expectations in sight exist finitely. Let $X_1, \ldots, X_p$ be a collection of random variables. Their covariance matrix $\mathbf{C}$ is defined by setting c_{jk} to be the covariance between X_j and X_k for $j, k \in \{1, 2, \ldots, p\}$. (What is the covariance between X_j and itself?) The variance of $a_1X_1 + a_2X_2 + \cdots + a_pX_p$ is then given by $\mathbf{a}^T\mathbf{C}\mathbf{a}$, where $\mathbf{a} := (a_1, \ldots, a_p)^T$. Since the variance of any quantity is nonnegative, this shows that any covariance matrix is nonnegative definite. Moreover, a covariance matrix is positive definite as long as there is no nonzero vector $\mathbf{a}$ such that $a_1X_1 + a_2X_2 + \cdots + a_pX_p$ is constant with probability one.

e. Joint and marginal probability mass/density functions (Cf. Larsen and Marx pp. 203-220)

Recall that a random variable is a function from an underlying sample space S to the set of real numbers $\mathbb{R}$. We now define a k-dimensional random vector, k a positive integer, to be a function from S to the k-fold Cartesian product $\mathbb{R}^k$. For now we fix $k = 2$. The random vector may be denoted $\mathbf{X}$, or we may write X and Y to represent its two components. Note that X and Y are themselves random variables. (Caution: In this context, a capital letter in bold font does not represent a matrix!)

If $\mathbf{X}$ is a discrete random vector, which is to say that $\mathbf{X}$ realizes only countably many values with positive probability, then X and Y may be described by a joint probability mass function

$$f_{X,Y}(x, y) := P(X = x, Y = y)$$

such that

$$P((X, Y)^T \in A) = \sum_{\{(x,y)^T \in A \cap \mathcal{S}\}} f_{X,Y}(x, y)$$

for any set $A \subset \mathbb{R}^2$, where

$$\mathcal{S} := \{(x, y)^T \in \mathbb{R}^2 : f_{X,Y}(x, y) > 0\}$$

is the support set of $\mathbf{X}$. The introduction of $\mathcal{S}$ here avoids the question of how to define an uncountable sum. Summation over the empty set is defined as 0. Note that we must have $f_{X,Y}(x, y) \geq 0$ and $\sum_{\{(x,y)^T \in \mathcal{S}\}} f_{X,Y}(x, y) = 1$.

For example, suppose that X is the number of phone messages received in the next hour and has the Poisson distribution with mean $\lambda \in (0, \infty)$, while Y is the number of text messages received in the next hour and has the Poisson distribution with mean $\mu \in (0, \infty)$. If phone messages arrive independently of text messages, then for any nonnegative integers x and y we have

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

This example illustrates that, in some cases, the joint probability mass function for X and Y is simply the product of the probability mass function for X with the probability mass function for Y. However, this is not always true. Suppose that for $\{(x,y)^T \in \mathbb{R}^2 : x \in \{0,1\}, y \in \{0,1\}\}$ we have

$$f_{X,Y}(x,y) = (x + 2y + 1)/10.$$

Clearly, $(x+2y+1)/10$ cannot be written as a product of a function of x with a function of y. Yet,

$$\sum_{\{(x,y)^T \in \mathcal{S}\}} (x + 2y + 1)/10 = 1/10 + 2/10 + 3/10 + 4/10 = 1,$$

so that $f_{X,Y}(x,y)$ is a valid joint probability mass function.

Since X and Y are themselves random variables, we may be interested in describing their distributions individually. If $\mathbf{X}$ is a discrete random vector, this can be accomplished by summing the joint probability mass function over appropriate sets. Explicitly, let u be any real number and put $A := \{(x,y)^T \in \mathbb{R}^2 : x = u\}$. Then $\sum_{\{(x,y)^T \in A \cap \mathcal{S}\}} f_{X,Y}(x,y)$ is simply $P(X = u)$ and may be labeled $f_X(u)$. In this way we recover the probability mass function of X, which is called a marginal probability mass function for a reason that will become clear presently. The marginal probability mass function of Y may be recovered similarly. While marginal probability mass functions are uniquely determined from the joint probability mass function, the reverse is not true.

If the joint probability mass function was obtained by multiplying the marginal probability mass functions, then presumably there is no need to derive the marginal probability mass functions using the approach indicated in the last paragraph. However, suppose that the joint probability mass function is not the product of a function of x with a function of y, as in the previous example. Then the approach indicated in the last paragraph is useful. Indeed, with $f_{X,Y}(x,y) = (x+2y+1)/10$ for $\{(x,y)^T \in \mathbb{R}^2 : x \in \{0,1\}, y \in \{0,1\}\}$ we have

$f_X(0) =$ $\qquad\qquad\qquad$ $f_X(1) =$

$f_Y(0) =$ $\qquad\qquad\qquad$ $f_Y(1) =$

When the joint probability mass function $f_{X,Y}(x,y)$ is listed in tabular form as below, the marginal probability mass functions $f_X(x)$ and $f_Y(y)$ may be read off the margins of the table.

$f_{X,Y}(x,y)$	$y=0$	$y=1$	$f_X(x)$
$x=0$	$1/10$	$3/10$	
$x=1$	$2/10$	$4/10$	
$f_Y(y)$			1

Can you exhibit a different joint probability mass function that is compatible with the above marginal probability mass functions?

We can define continuous random vectors to be those for which the joint cumulative distribution function

$$F_{X,Y}(x,y) := P(X \le x, Y \le y)$$

is continuous. However, a more practical (but stringent) working definition is that there exist a joint probability density function $f_{X,Y}(x,y)$ such that

$$P(\mathbf{X} \in A) = \int\int_A f_{X,Y}(x,y)\ dx\ dy$$

for any set $A \subset \mathbb{R}^2$. In particular,

$$F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u,v)\ du\ dv$$

and, if $f_{X,Y}(x,y)$ is continuous,

$$\frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y) = f_{X,Y}(x,y).$$

Note that we must have $\int\int_{\mathbb{R}^2} f_{X,Y}(x,y)\ dx\ dy = 1$ and that we may as well require $f_{X,Y}(x,y) \ge 0$.

Suppose that $\mathbf{X}$ is a continuous random vector with joint probability density function

$$f_{X,Y}(x,y) = 8xy1_{\{0<x<y<1\}}.$$

To verify that this is a valid joint probability density function, note that

$$\int\int_{\mathbb{R}^2} f_{X,Y}(x,y)\ dx\ dy = \int_0^1 \left\{ \int_0^y 8xy\ dx \right\}\ dy$$

$$= \int_0^1 8y\{x^2/2\}_0^y\ dy = \int_0^1 4y^3\ dy = \{y^4\}_0^1 = 1.$$

Suppose that we want to find $F_{X,Y}(x,y)$. Consider five cases:

1. $x \le 0$ or $y \le 0 \implies F_{X,Y}(x,y) = 0$.
2. $x \ge 1$ and $y \ge 1 \implies F_{X,Y}(x,y) = 1$.
3. $x \ge y$ and $0 \le y \le 1 \implies F_{X,Y}(x,y) = \int_0^y \left\{\int_0^v 8uv\ du\right\}\ dv$
$= \int_0^y 4v^3\ dv = y^4$.
4. $y \ge 1$ and $0 \le x \le 1 \implies F_{X,Y}(x,y) = \int_0^x \left\{\int_u^1 8uv\ dv\right\}\ du$
$= \int_0^x 4(u-u^3)\ du = 2x^2 - x^4$.
5. $0 \le x \le y \le 1 \implies F_{X,Y}(x,y) = \int_0^x \left\{\int_u^y 8uv\ dv\right\}\ du$
$= \int_0^x 4(uy^2 - u^3)\ du = 2x^2y^2 - x^4$.

Just as a marginal probability mass function is obtained by summing a joint probability mass function, a marginal probability density function is obtained by integrating a joint probability density function. Explicitly, we have

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dy \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dx.$$

An easy way to remember which is which is to note that the marginal probability density function of X must depend on x, so the y should be integrated out.

Continuing from the previous example, we have, for $x \in (0,1)$,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dy = \int_x^1 8xy\ dy$$

$$=$$

and, for $y \in (0,1)$,

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dx = \int_0^y 8xy\ dx$$

$$=$$

For a discrete random vector $\mathbf{X}$, the expected value $E[g(X,Y)]$ is defined as

$$\sum_{\{(x,y)^T \in \mathcal{S}\}} g(x,y) f_{X,Y}(x,y),$$

provided that the sum is absolutely convergent. For a continuous random vector $\mathbf{X}$, the expected value $E[g(X,Y)]$ is

$$\int\int_{\mathbb{R}^2} g(x,y) f_{X,Y}(x,y)\ dx\ dy,$$

provided that the integral is absolutely convergent. Even though we are now studying random vectors, expected value has the same linearity and monotonicity properties discussed earlier this semester.

Continuing from the previous example, put $g(X,Y) := X^\gamma Y^\delta$ for constants $\gamma \in [0,\infty)$ and $\delta \in [0,\infty)$. We have

$E[g(X,Y)] =$

Starting a new example, suppose that $f_{X,Y}(x,y)$ has the form $f_X(x)f_Y(y)$, which is to say that the joint probability density function is the product of the marginal probability density functions. I claim that, in this case, $Var[X+Y] = Var[X] + Var[Y]$ if all of these quantities exist as finite numbers. To prove my claim, I begin by noting that

$$E[XY] = \int\int_{\mathbb{R}^2} xy f_{X,Y}(x,y)\ dx\ dy = \int_{\mathbb{R}} x f_X(x)\ dx \int_{\mathbb{R}} y f_Y(y)\ dy = E[X]E[Y].$$

Then, putting $\mu := E[X]$ and $\nu := E[Y]$, I have

$$\begin{aligned}
Var[X+Y] &= E[(X+Y)^2] - (E[X+Y])^2 \\
&= E[X^2 + 2XY + Y^2] - (\mu^2 + 2\mu\nu + \nu^2) \\
&= E[X^2] - \mu^2 + E[Y^2] - \nu^2 + 2E[XY] - 2\mu\nu \\
&= E[X^2] - \mu^2 + E[Y^2] - \nu^2 + 2E[X]E[Y] - 2\mu\nu \\
&= E[X^2] - \mu^2 + E[Y^2] - \nu^2 \\
&= Var[X] + Var[Y].
\end{aligned}$$

You will learn next week that X and Y are called independent if their joint probability density function decomposes into the product of their marginal probability density functions.

f. Conditional probability mass/density functions (Cf. Larsen and Marx pp. 249-257)

Let $\mathbf{X}$ be a discrete random vector with components X and Y. Let $f_{X,Y}(x, y)$ denote the joint probability mass function of X and Y, $f_X(x)$ the marginal probability mass function of X, and $f_Y(y)$ the marginal probability mass function of Y.

For any $x \in \mathbb{R}$ at which $f_X(x) > 0$, we define

$$f_{Y|X}(y|x) := f_{X,Y}(x, y)/f_X(x) = P(Y = y, X = x)/P(X = x) = P(Y = y|X = x)$$

to be the conditional probability mass function of Y given that $X = x$.

The interpretation of $f_{Y|X}(y|x)$ is that, for any set $A \subset \mathbb{R}$,

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

where $S_{Y|x} := \{y \in \mathbb{R} : f_{Y|X}(y|x) > 0\}$.

For any function $g(y)$, we define the conditional expectation of $g(Y)$ given that $X = x$ as

$$E[g(Y)|X = x] = \sum_{y \in \mathcal{S}_{Y|x}} g(y) f_{Y|X}(y|x).$$

To illustrate, suppose that for $\{(x, y)^T \in \mathbb{R}^2 : x \in \{0, 1\}, y \in \{0, 1\}\}$ we have

$$f_{X,Y}(x, y) = (x + 2y + 1)/10.$$

We have $f_X(x) = (4 + 2x)/10$ for $x \in \{0, 1\}$, so that

$$f_{Y|X}(y|0) =$$

$$f_{Y|X}(y|1) =$$

$$E[Y^2|X = 0] =$$

Let $\mathbf{X}$ be a continuous random vector with components X and Y. Let $f_{X,Y}(x,y)$ denote the joint probability density function of X and Y, $f_X(x)$ the marginal probability density function of X, and $f_Y(y)$ the marginal probability density function of Y.

For any $x \in \mathbb{R}$ at which $f_X(x) > 0$, we define

$$f_{Y|X}(y|x) := f_{X,Y}(x,y)/f_X(x)$$

to be the conditional probability density function of Y given that $X = x$.

The interpretation of $f_{Y|X}(y|x)$ is that, for any set $A \subset \mathbb{R}$,

$$P(Y \in A|X = x) = \int_A f_{Y|X}(y|x)\ dy.$$

For any function $g(y)$, we define the conditional expectation of $g(Y)$ given that $X = x$ as

$$E[g(Y)|X = x] = \int_{\mathbb{R}} g(y) f_{Y|X}(y|x)\ dy.$$

To illustrate, suppose that

$$f_{X,Y}(x,y) = 8xy1_{\{0<x<y<1\}}.$$

We have $f_X(x) = 4(x - x^3)1_{\{0<x<1\}}$, so that

$$f_{Y|X}(y|x) =$$

$$P(Y \le 3/4|X = 1/2) =$$

$$E[Y^2|X = 1/2] =$$

We say that X and Y are independent random variables if

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

for any sets $A, B \subset \mathbb{R}$. If $\mathbf{X}$ is a discrete random vector, then as a special case we may take $A := \{x\}$ and $B := \{y\}$ to obtain the probability mass decomposition

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

for any $(x, y)^T \in \mathbb{R}^2$. Likewise, the probability density decomposition

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

may be used to characterize independence if $\mathbf{X}$ is a continuous random vector, with the technical caveat that the probability density decomposition is not required for all $(x, y)^T \in \mathbb{R}^2$ but only for $(x, y)^T \in C \subset \mathbb{R}^2$ with $P(\mathbf{X} \in C) = 1$.

Suppose that X and Y are independent (and continuous, for simplicity in the calculations to follow). Then, whenever all quantities below exist as finite numbers, we have

$$\begin{aligned}
M_{X+Y}(t) &= E[\exp\{t(X+Y)\}] \\
&= E[\exp\{tX\}\exp\{tY\}] \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} \exp[tx]\exp[ty] f_{X,Y}(x, y)\ dx\ dy \\
&= \int_{\mathbb{R}} \exp[tx] f_X(x)\ dx \int_{\mathbb{R}} \exp[ty] f_Y(y)\ dy \\
&= E[\exp\{tX\}]E[\exp\{tY\}] \\
&= M_X(t)M_Y(t).
\end{aligned}$$

This result can be used to prove that the sum of two independent normal random variables is normal, among many other useful relations (Cf. Written Assignment 5, BST 675, Fall 2010).

g. Univariate and bivariate transformations

Suppose that X is a continuous random variable with probability density function $f_X(x)$. Let g be a function of one real variable and define $Y := g(X)$. Let

$$\mathcal{X} := \{x \in \mathbb{R} : f_X(x) > 0\}$$

denote the support of X and put

$$\mathcal{Y} := \{y \in \mathbb{R} : y = g(x) \text{ for some } x \in \mathcal{X}\}.$$

We will identify some circumstances under which Y will be a continuous random variable and obtain its probability density function $f_Y(y)$.

First suppose that g is strictly increasing on $\mathcal{X}$, in that $u > v$ implies $g(u) > g(v)$ whenever $u, v \in \mathcal{X}$. Then, for any $y \in \mathcal{Y}$ there exists a unique $x \in \mathcal{X}$ such that $g(x) = y$. We refer to this x as $g^{-1}(y)$. Since $g(X) \leq y$ is equivalent to $X \leq g^{-1}(y)$, we have

$$F_Y(y) = F_X(g^{-1}(y)).$$

Moreover, if $f_X(x)$ is continuous on $\mathcal{X}$ and $g^{-1}(y)$ has a continuous derivative on $\mathcal{Y}$, then Y is a continuous random variable with probability density function

$$f_Y(y) = f_X(g^{-1}(y))\frac{d}{dy}g^{-1}(y)1_{\{y \in \mathcal{Y}\}}.$$

Next suppose that g is strictly decreasing on $\mathcal{X}$, in that $u > v$ implies $g(u) < g(v)$ whenever $u, v \in \mathcal{X}$. Again, for any $y \in \mathcal{Y}$ there exists a unique $x \in \mathcal{X}$ such that $g(x) = y$. We refer to this x as $g^{-1}(y)$. Since $g(X) \leq y$ is equivalent to $X \geq g^{-1}(y)$, we have

$$F_Y(y) = 1 - F_X(g^{-1}(y)).$$

Moreover, if $f_X(x)$ is continuous on $\mathcal{X}$ and $g^{-1}(y)$ has a continuous derivative on $\mathcal{Y}$, then Y is a continuous random variable with probability density function

$$f_Y(y) = -f_X(g^{-1}(y))\frac{d}{dy}g^{-1}(y)1_{\{y \in \mathcal{Y}\}}.$$

Both cases above can be subsumed into the single formula

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| 1_{\{y \in \mathcal{Y}\}},$$

which we refer to as the "univariate transformation formula".

As an example, suppose that X has probability density function

$$f_X(x) := \exp[-x] 1_{\{x>0\}}$$

with corresponding cumulative distribution function

$$F_X(x) := (1 - \exp[-x]) 1_{\{x>0\}}.$$

Put $g(x) := 1 - \exp[-x]$, which is strictly increasing and which maps $\mathcal{X} = (0, \infty)$ to $\mathcal{Y} = (0, 1)$. We have $g^{-1}(y) = -\log(1 - y)$, from which we find that the probability density function of Y is

Caution: The univariate transformation formula does not apply to, for example, $g(x) := x^2$ if $\mathcal{X}$ is an interval containing both positive and negative real numbers because g is neither strictly increasing nor strictly decreasing on such an interval. A modified version of the univariate transformation formula is available for such cases, but a simpler approach is usually to reason out what is the cumulative distribution function of Y and then differentiate the cumulative distribution function of Y to get the probability density function of Y.

Now let $\mathbf{X}$ be a continuous random vector with components X and Y, and let $f_{X,Y}(x, y)$ denote the joint probability density function of X and Y. Suppose that g_1 and g_2 are functions defined on the support of $\mathbf{X}$ that define a one-to-one bivariate transformation, in the sense that

$$g_1(x, y) = g_1(w, z),\ g_2(x, y) = g_2(w, z) \quad \text{implies} \quad x = w,\ y = z.$$

Then the equations $u = g_1(x, y), v = g_2(x, y)$ can be solved for x and y, say $x = h_1(u, v)$ and $y = h_2(u, v)$. Put $U := g_1(X, Y)$ and $V := g_2(X, Y)$. Assuming the existence of all partial derivatives referenced below, the joint probability density function of U and V is

$$f_{U,V}(u,v) = f_{X,Y}(h_1(u,v), h_2(u,v)) \left| Det \begin{bmatrix} \frac{\partial h_1(u,v)}{\partial u} & \frac{\partial h_1(u,v)}{\partial v} \\ \\ \frac{\partial h_2(u,v)}{\partial u} & \frac{\partial h_2(u,v)}{\partial v} \end{bmatrix} \right| 1_{\{(u,v)^T \in S_{U,V}\}},$$

where Det returns the determinant of a square matrix (assumed nonzero here) and

$$S_{U,V} := \{(u,v)^T \in \mathbb{R}^2 : \exists (x,y)^T \in \mathbb{R}^2 \text{ with } u = g_1(x,y), v = g_2(x,y), f_{X,Y}(x,y) > 0\}.$$

We refer to the above formula for $f_{U,V}(u,v)$ as the "bivariate transformation formula".

Remarks: The symbol $\exists$ is mathematical shorthand for "there exists". While the determinant of an $m \times m$ matrix does not have a simple computational formula for arbitrary m, for $m = 2$ there is a simple computational formula: the product of the diagonal elements minus the product of the off-diagonal elements.

To illustrate, suppose that X has the chi-square distribution on 2 df,

$$f_X(x) = (1/2) \exp[-x/2] 1_{\{x>0\}},$$

and that, independently, Y has the uniform distribution on $(-\pi/2, \pi/2)$,

$$f_Y(y) = (1/\pi) 1_{\{-\pi/2<y<\pi/2\}}.$$

Put $g_1(x,y) := \sqrt{x} \cos y$ and $g_2(x,y) := \sqrt{x} \sin y$ for $x \in (0,\infty)$ and $y \in (-\pi/2, \pi/2)$. Let $U := g_1(X,Y)$ and $V := g_2(X,Y)$. What is the joint distribution of U and V? What are the marginal distributions of U and V?

<u>Step 1. Find the support of U and V.</u> Since $\cos y$ must be positive when $-\pi/2 < y < \pi/2$ while $\sin y$ can be positive or negative or zero, we have $S_{U,V} = \{(u,v)^T \in \mathbb{R}^2 : u > 0\}$.

<u>Step 2. Verify that the transformation is one-to-one.</u> With $u = \sqrt{x} \cos y$ and $v = \sqrt{x} \sin y$, we have $x = u^2 + v^2$ and $\tan[y] = v/u$. Since $-\pi/2 < y < \pi/2$,

$\tan[y] = v/u$ has the unique solution $y = \arctan[v/u]$. So put $h_1(u,v) := u^2+v^2$ and $h_2(u,v) := \arctan[v/u]$. The fact that we were able to solve for y and x not only implies that the transformation is one-to-one but also provides useful results for the next step.

Step 3. Evaluate the matrix determinant. We have

$$\frac{\partial h_1(u,v)}{\partial u} = 2u, \quad \frac{\partial h_1(u,v)}{\partial v} = 2v,$$

$$\frac{\partial h_2(u,v)}{\partial u} = \frac{1}{1+(v/u)^2}\frac{\partial(v/u)}{\partial u} = \frac{-v}{u^2+v^2}, \quad \text{and}$$

$$\frac{\partial h_2(u,v)}{\partial v} = \frac{1}{1+(v/u)^2}\frac{\partial(v/u)}{\partial v} = \frac{u}{u^2+v^2}.$$

So the matrix determinant is

Step 4. Report the joint probability density function. We have

$f_{X,Y}(h_1(u,v), h_2(u,v)) =$

so that

$f_{U,V}(u,v) =$

Step 5. Report the marginal probability density functions. Since $f_{U,V}(u,v)$ can be written in the form $g(u)h(v)$, the kernels of $f_U(u)$ and $f_V(v)$ are obvious. All we need to do is determine the normalizing constants, but this is not difficult. Since V is obviously a standard normal random variable, we must have

$$f_V(v) = (2\pi)^{-1/2}\exp[-v^2/2].$$

This implies that

$$f_U(u) = 2(2\pi)^{-1/2}\exp[-u^2/2]1_{\{u>0\}}.$$

How would you describe the distribution of U?

h. Resolution of Motivating Case Studies

To resolve our first motivating case study, we propose that systolic blood pressure (X) and total serum cholesterol (Y) have the bivariate normal distribution with joint probability density function

$$f_{X,Y}(x,y) = (2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2})^{-1}\times$$

$$\exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right\}\right],$$

where $\mu_X, \mu_Y \in (-\infty,\infty)$, $\sigma_X, \sigma_Y \in (0,\infty)$, and $\rho \in (-1,1)$.

Some properties of the bivariate normal distribution are as follows:

• The marginal distribution of X is normal with mean μ_X and variance σ_X^2. You can prove this with a gruesome computation that involves completing a square in the integrand and invoking the kernel method.

• Likewise, the marginal distribution of Y is normal with mean μ_Y and variance σ_Y^2.

• The correlation between X and Y, defined generally as

$$E[(X-\mu_X)(Y-\mu_Y)]/\sqrt{E[(X-\mu_X)^2]E[(Y-\mu_Y)^2]},$$

equals ρ. Note that, for a bivariate normal distribution, a zero correlation between the component random variables is equivalent to independence. However, such an equivalency does not hold for a generic bivariate distribution; two random variables can be uncorrelated without being independent.

• The conditional distribution of Y given that $X = x$ is normal with mean $\mu_Y + \rho(\sigma_Y/\sigma_X)(x-\mu_X)$ and variance $\sigma_Y^2(1-\rho^2)$. Note, then, that simple linear regression is essentially estimating the parameters of a bivariate normal distribution from data. In particular, $\rho(\sigma_Y/\sigma_X)$ plays the role of slope, and since we can figure out reasonable estimates for each of these parameters, we can also figure out a reasonable estimate of the slope.

Now, returning to our specific question about $P(X > 140 \ \cap \ Y > 200)$, the answer is

$$\int_{140}^{\infty}\int_{200}^{\infty} f_{X,Y}(x,y)\; dx\; dy$$

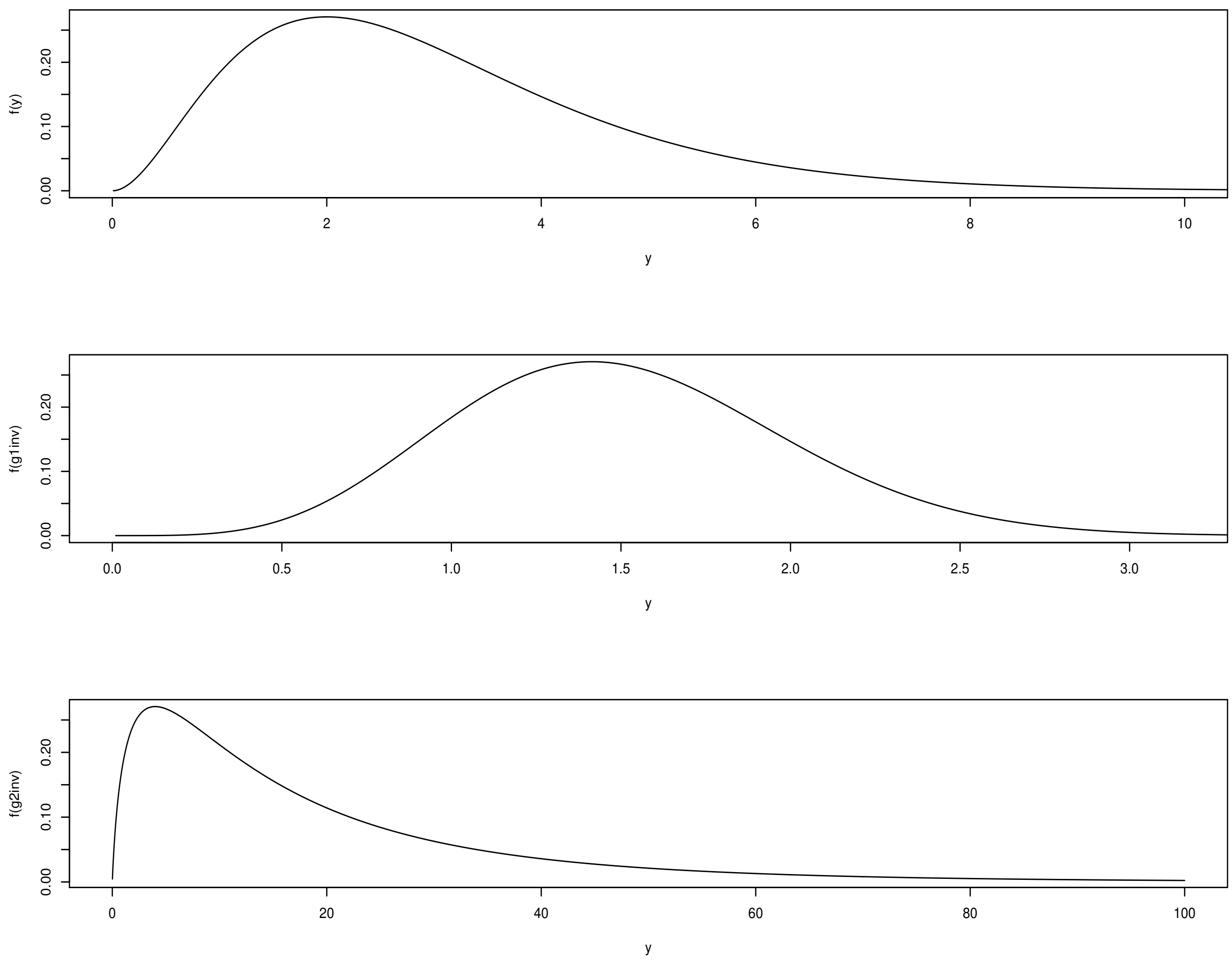
with $f_{X,Y}(x,y)$ as on the preceding page and $\mu_X = 130$, $\sigma_X = 15$, $\mu_Y = 180$, $\sigma_Y = 25$. We must also assume a value for ρ. For concreteness, let us assume that $\rho = 0.4$. Then the double integral may be approximated numerically with the following R code.

```
x<- rep((55:205), 251)
y<- sort( rep((55:305), 151))
rho <- 0.4
fxy <- (1/(2*pi*15*25*sqrt(1-rho^2)))*
exp(-(1/(2*(1-rho^2))) * ( ((x-130)/15  )^2
-2*rho*((x-130)/15)*((y-180)/25 )+ ( (y-180)  / 25)^2  ) )
Event <- (x > 140)*(y > 200)
sum(Event*fxy)
```

The result is 0.090, which is noticeably higher than the 0.053 that would have been obtained under the (unrealistic) assumption that X and Y were independent. Hence, the proportion of people with at least one of the two risk factors is

$$\begin{aligned} P(X > 140 \ \cup \ Y > 200) &= P(X > 140) + P(Y > 200) - P(X > 140 \ \cap \ Y > 200) \\ &= 0.252 + 0.212 - 0.090 \\ &= 0.374. \end{aligned}$$

To resolve our second motivating case study, let us suppose that Y has marginal probability density function $f_Y(y) := y^2 \exp[-y]/2$ for $y > 0$. The first panel of the figure on the next page, created with the R code also shown on the next page, depicts $f_Y(y)$. Clearly, any data arising from $f_Y(y)$ will have a right-skewed distribution.




```
postscript("C:/Documents and Settings/Rich/
Desktop/Rich/BST675F10/Dilation.ps")
y <- (1:10000)/100
g1inv <- y^2
g2inv <- sqrt(y)
f <- function(y) { y^2*exp(-y)/2  }
par(mfrow=c(3,1))
plot(y,f(y),type="l",xlim=c(0,10))
plot(y,f(g1inv),type="l",xlim=c(0,sqrt(10)))
plot(y,f(g2inv),type="l",xlim=c(0,100))
dev.off()
```

What will happen if we transform Y? Consider two possibilities. First, suppose that $Z := g_1(Y)$ with $g_1(y) := y^{1/2}$ for $y > 0$. Then $g_1^{-1}(y) = y^2$ and the probability density function of Z will be $f_Y(y^2)\frac{d}{dy}y^2$ for $y > 0$.

The second panel of the figure on the previous page shows $f_Y(y^2)$, which is a horizontal compression of $f_Y(y)$. However, the compression is not uniform, and precisely because of the non-uniform compression $f_Y(y^2)$ ends up with a roughly bell-shaped appearance. Of course, $f_Y(y^2)$ is not itself the probability density function of Z, as we need the factor of $\frac{d}{dy}y^2$ to ensure that the probability density function of Z integrates to one.

Even so, the take-home message is clear: An inverse transformation with positive second derivative reduces right skewness, and therefore a forward transformation with negative second derivative reduces right skewness. This is why the square root (and the logarithm) are often suggested as transformations to reduce right skewness.

Second, suppose that $W := g_2(Y)$ with $g_2(y) := y^2$ for $y > 0$. Then $g_2^{-1}(y) = y^{1/2}$ and the probability density function of W will be $f_Y(y^{1/2})\frac{d}{dy}y^{1/2}$ for $y > 0$.

The third panel of the figure shows $f_Y(y^{1/2})$, which is a horizontal dilation of $f_Y(y)$. However, the dilation is not uniform, and indeed the non-uniform dilation works against us.

Again, the take-home message is clear: An inverse transformation with negative second derivative exacerbates right skewness, and therefore a forward transformation with positive second derivative exacerbates right skewness. This is why the square (and the exponential) are never suggested as transformations to reduce right skewness.

Of course, the situation is more complicated when Y is the dependent variable in a linear regression model. There, our interest is in whether the conditional distribution of Y given $X = x$ is approximately normal. However, this is usually the case if the marginal distribution of Y is approximately normal.