3.5 Conditional distributions

Let's first consider random pairs (X, Y).

Often with random pairs (X, Y), knowing something about Y tells us something about X. We focus on the special case where we *observe* $Y = y$. That is, we don't just know "something" about Y, we know *everything* about Y.

For example, if Y is age, we might observe $Y = 63.7$ years. We might be interested in how X, the the quality of life of a person undergoing chemotherapy on a 10-point scale, is distributed knowing $Y = 63.7$. We might expect that the distribution of $X|Y = y$ might change for $X|Y = 20$ versus $X|Y = 50$ or $X|Y = 70$.

Knowing something about Y, say that Y is in the set B, might tell us something about X. By definition,

$$P(X \in A | Y \in B) = \frac{P(X \in A \cap Y \in B)}{P(Y \in B)}.$$

Now let (X, Y) be continuous with pdf $f(x, y)$. We can let B be very small, e.g. $B = [y, y + dy)$ where dy is really, really tiny. This is *almost* like seeing $Y = y$. Then

$$\begin{aligned} P(X \in A | y \leq Y < y + dy) &= \frac{P(X \in A \cap y \leq Y < y + dy)}{P(y \leq Y < y + dy)} \\ &\approx \frac{\left[\int_A f(x, y) dx\right] dy}{f_Y(y) dy} \\ &= \int_A \left[\frac{f(x, y)}{f_Y(y)}\right] dx. \end{aligned}$$

So given we know $Y = y$, $\frac{f(x,y)}{f_Y(y)}$ is what we integrate over the set A to get $P(X \in A | Y = y)$. We take the limit as $dy \to 0$ to make this precise.

def'n: We define the conditional pdf of X given that $Y = y$ as

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

This conditional pdf has a range that may depend on the value $Y = y$ we observe, $R_{X|y} = \{x : f(x,y) > 0\}$.

Remember: we **see** $Y = y$. Y is *no longer random*, it's a number, like $Y = 2.73$ milligrams or $Y = \$80,000$.

---

def'n: We define the conditional pdf of Y given that $X = x$ as

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} \text{ with range } R_{Y|x} = \{y : f(x,y) > 0\}.$$

def'n: We define the conditional pdf of X given that $Y = y$ as

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)} \text{ with range } R_{X|y} = \{x : f(x,y) > 0\}.$$

Problem 9, page 108.

The random pair (X, Y) has pdf $f(x, y) = k$ on
$R_{X,Y} = \{(x, y) : -1 < x < 1,\ 0 < y < 1 - x^2\}$.

We are asked to find the marginal pdf's $f_X(x)$ and $f_Y(y)$ and conditional pdf's $f_{X|Y}(x,y)$ and $f_{Y|X}(y|x)$. By looking at the region $R_{X,Y}$ we can see $R_X = (-1, 1)$ and $R_Y = (0, 1)$.

Let's first find k.

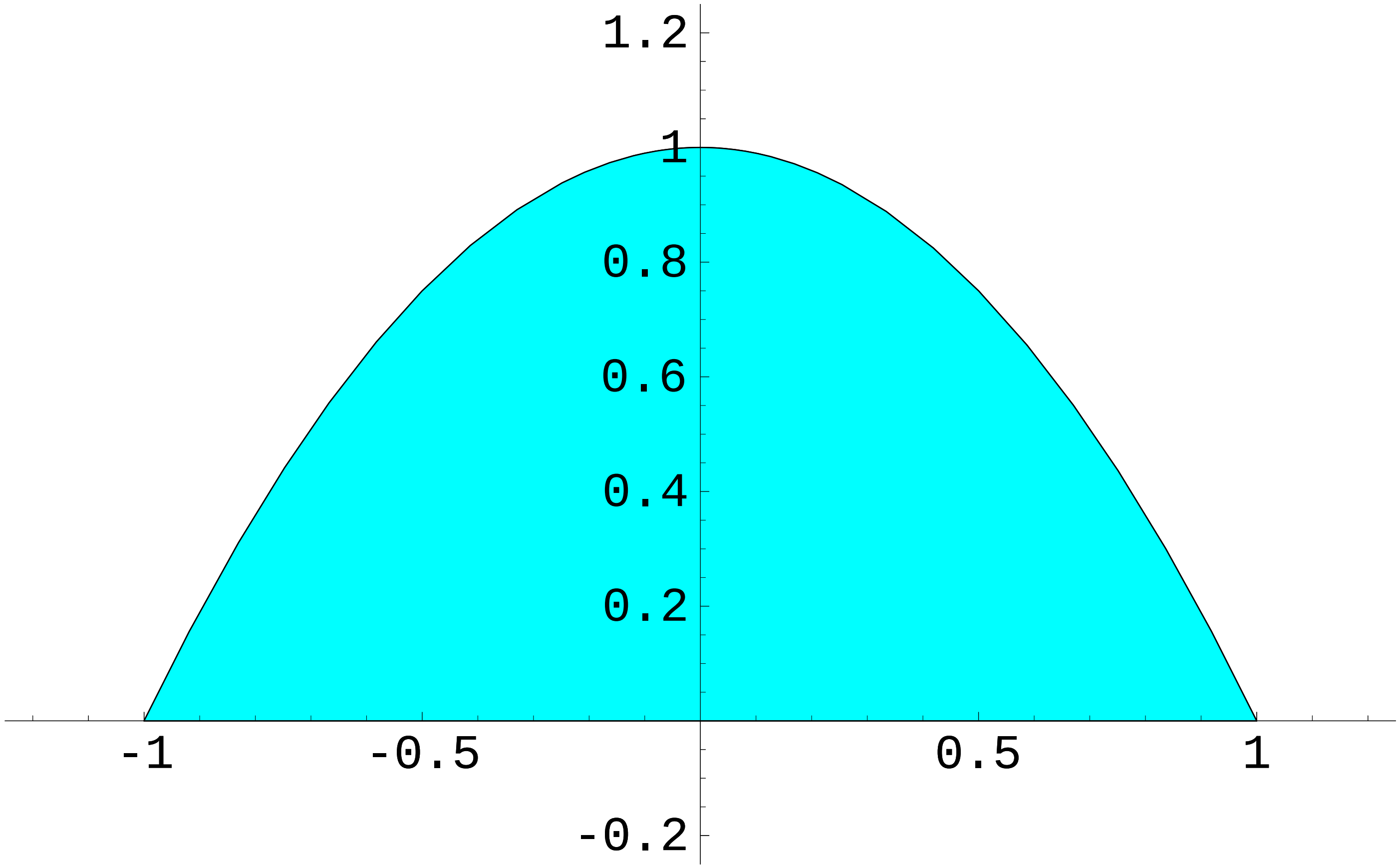

Figure 1: Possible outcomes of experiment (X, Y), namely $R_{X,Y}$, are shaded.

$$
\begin{aligned}
\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dydx &= \int_{-1}^{1}\int_{0}^{1-x^2} kdydx \\
&= k\int_{-1}^{1}\left[[y]_0^{1-x^2}\right]dx \\
&= k\int_{-1}^{1}[1-x^2]dx \\
&= k\left[x-\frac{1}{3}x^3\right]_{-1}^{1} \\
&= k\left[1-\frac{1}{3}-(-1+\frac{1}{3})\right] = \frac{k4}{3}
\end{aligned}
$$

Setting this integral to one yields $k = 0.75$. So $f(x,y) = 0.75$ over $R_{X,Y} = \{(x,y) : -1 < x < 1,\ 0 < y < 1-x^2\}$.

Let $x \in R_X$. The marginal pdf $f_X(x)$ is

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x,y)dy \\ &= \int_0^{1-x^2} 0.75dy \\ &= 0.75[y]_0^{1-x^2} \\ &= 0.75(1-x^2). \end{aligned}$$

The conditional pdf $f_{Y|X}(y|x)$ is

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{0.75}{0.75(1-x^2)} = \frac{1}{1-x^2},$$

on $R_{Y|x} = \{y : 0 < y < 1-x^2\}$. That is, $Y|X = x \sim U(0, 1-x^2)$.

Let $y \in R_Y$. The marginal pdf $f_Y(y)$ is

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f(x,y)dx \\ &= \int_{-\sqrt{1-y}}^{\sqrt{1-y}} 0.75dx \\ &= 0.75[x]_{-\sqrt{1-y}}^{\sqrt{1-y}} \\ &= 1.5\sqrt{1-y}. \end{aligned}$$

The conditional pdf $f_{X|Y}(x|y)$ is

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)} = \frac{0.75}{1.5\sqrt{1-y}} = \frac{0.5}{\sqrt{1-y}},$$

on $R_{X|y} = \{x : -\sqrt{1-y} < x < \sqrt{1-y}\}$. That is,
$X|Y = y \sim U(-\sqrt{1-y}, \sqrt{1-y})$.

So why are these useful?

Let's say that we only care about Y. We know marginally Y has pdf $f_Y(y) = 1.5\sqrt{1-y}$ over possible values $R_Y = (0,1)$. So, for example,

$$P(Y \le 0.5) = \int_0^{0.5} 1.5\sqrt{1-y}dy \approx 0.647.$$

This is if we know nothing about X. However, what if we know $X = -0.7$? Adding this bit of information now gives us the pdf $f_{Y|X}(y|-0.7) = (1-(-0.7)^2)^{-1} = 1/0.51$ on $R_{Y|-0.7} = (0, 1-(-0.7)^2) = (0, 0.51)$. So

$$P(Y \le 0.5|X = -0.7) = \int_0^{0.5} \frac{1}{0.51} dx \approx 0.980.$$

That is, knowing that $X = -0.7$ increases the probability that $Y \le 0.5$ dramatically.

In this example, are X and Y independent?

No they can't be because the range of $X|Y=y$ depends on y. Similarly the range of $Y|X=x$ depends on x.

Note that another way to check for independence is to check that $f_{X|Y}(x|y)=f_X(x)$ or $f_{Y|X}(y|x)=f_Y(y)$. That is, X and Y are independent if knowing the value one one doesn't affect the distribution of the other. This in fact happens if $f(x,y)=f_X(x)f_Y(y)$; or $p(x,y)=p_X(x)p_Y(y)$ for discrete pairs.

Recap: conditional probabilities of the form $P(X \in A|Y = y)$ and $P(Y \in y|X = x)$ are obtained from conditional pdf's:

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

and

$$f_{Y|X} = \frac{f(x, y)}{f_X(x)}.$$

Conditional pmf's for discrete (X, Y) are defined similarly:

$$p_{X|Y}(x|y) = P(X = x|Y = y) = \frac{P(X = x \cap Y = y)}{P(Y = y)} = \frac{p(x, y)}{p_Y(y)}.$$

$$p_{Y|X}(y|x) = P(Y = y|X = x) = \frac{P(Y = y \cap X = x)}{P(X = x)} = \frac{p(x, y)}{p_X(x)}.$$

In all cases, the range of possible values will depend on the value of the variable being conditioned on.

Example: Let X be the age in in months a child speaks his/her first word and let Y be the Gesell adaptive score, a measure of a child's aptitude (observed later on). Are X and Y related? How does the child's aptitude *change* with how long it takes them to speak?

Assuming a particular model, the PDF for (X, Y) is estimated to be

$f(x,y) = \exp(-60.22+1.3006x-0.0134x^2+0.9520y-0.0098xy-0.0043y^2)$

based on data collected on 21 children in a UCLA study. We can further find

$$f_Y(y) = \exp(-3.557 - 0.00256(y - 93.67)^2),$$

and

$$\begin{aligned} f_{Y|X}(y|x) &= \exp(-55.59 + 1.073x - 0.0055x^2 + \\ &\quad 0.9520y - 0.0098xy - 0.0043y^2). \end{aligned}$$

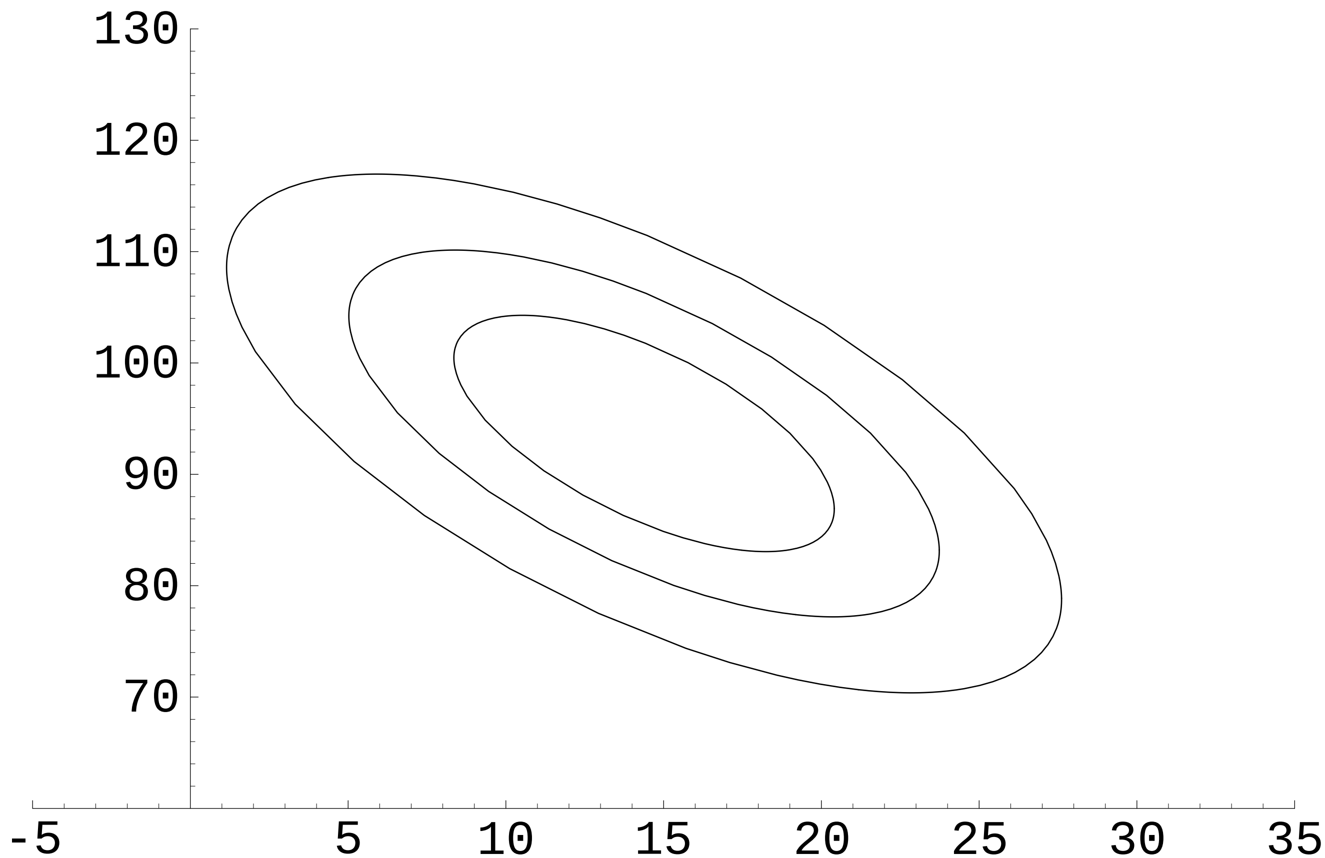

Figure 2: Level curves (25%,50%,75%) of $f(x, y)$ for (X, Y).

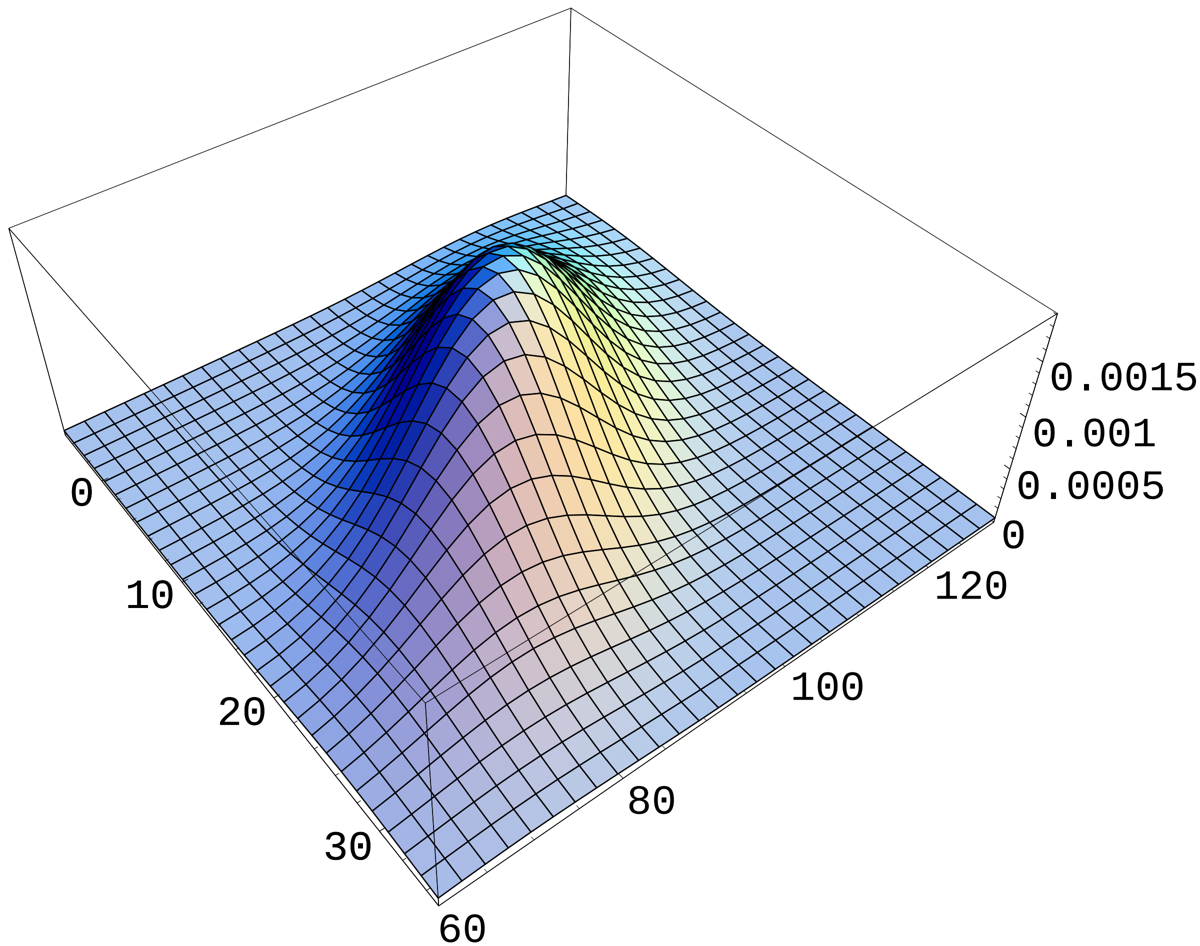

Figure 3: 3D plot of $f(x, y)$ for (X, Y).

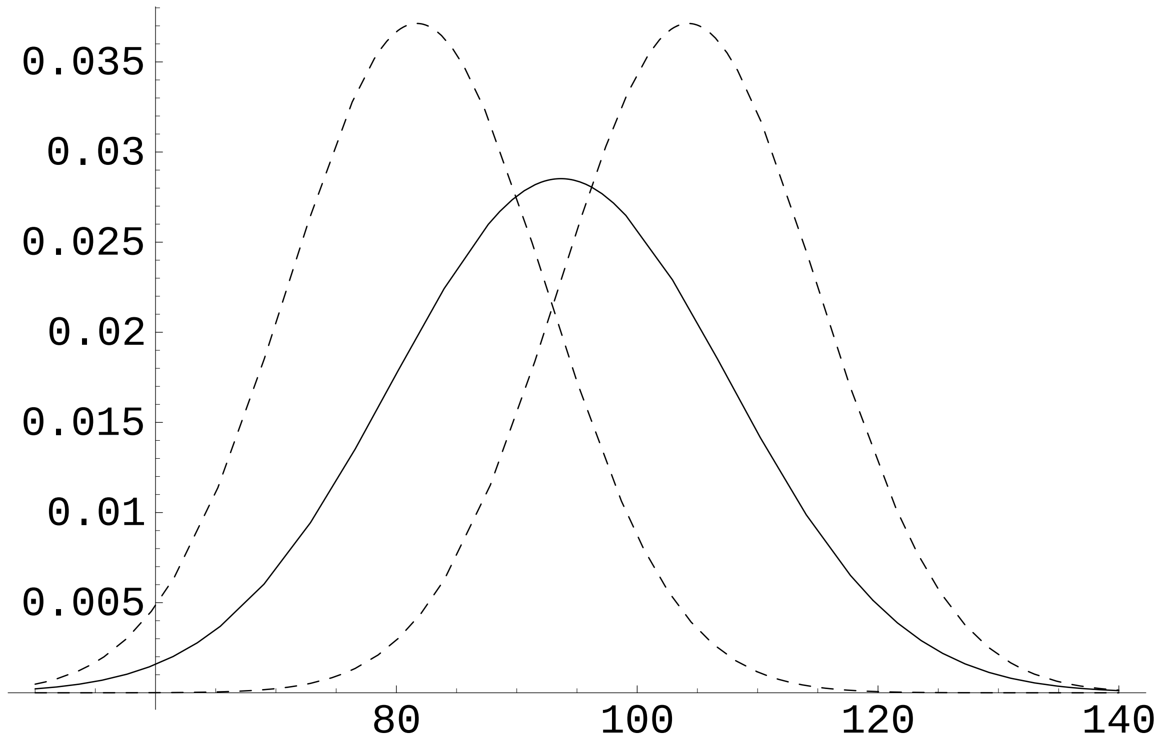

Figure 4: Solid is $f_Y(y)$; left dashed is $f_{Y|X}(y|25)$ the right dashed is $f_{Y|X}(y|10)$. As the age in months of first words $X = x$ increases, the distribution of Gesell Adaptive Scores Y decreases.

Jointly defining (X, Y) through one marginal and one conditional pdf

By definition,

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

Multiplying both sides by $f_Y(y)$ gives us

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

This is the continuous random variable analogue of

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A).$$

Similarly for discrete (X, Y),

$$p(x, y) = p_{X|Y}(x|y) p_Y(y) = p_{Y|X}(y|x) p_X(x).$$

Example: say X is the amount of rain in inches an apple tree gets over a summer. Assume $X \sim \exp(0.1)$. Conditional on $X = x$, the yield of apples in pounds is $Y|X = x \sim \exp\left(\frac{1}{2x}\right)$. Then the joint pdf of (X, Y), (rain, apples), is

$$f(x, y) = f_{Y|X}(y|x) f_X(x) = \frac{1}{2x} e^{-\frac{y}{2x}} 0.1 e^{-0.1x} = \frac{1}{20x} e^{-0.1x - \frac{y}{2x}},$$

on $R_{X,Y} = (0, \infty) \times (0, \infty)$.

This idea extends to a continuous X and a discrete $Y|X = x$ as well, a more advanced topic that we'll skip.

Question: how would we obtain the *marginal* distribution of X, the yield of apples in pounds? The joint density is averaged over rainfall:

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx = \text{no closed form solution!}$$

Random vectors

The random vector $\mathbf{X}$ and fixed vector $\mathbf{x}$ are often denoted

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \in \mathbb{R}^n \text{ and } \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n,$$

rather than $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ and $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. In the former notation the vectors are *column* vectors, the latter notation writes them as *row* vectors. More on this later.

Say that $(X_1, \ldots, X_n)$ is continuous and therefore has a density $f(x_1, \ldots, x_n) \geq 0$ on $\mathbb{R}^n$. In vector notation we say $\mathbf{X}$ is continuous with pdf $f(\mathbf{x}) \geq 0$.

There are several marginal and conditional distributions one can obtain from $f(x_1, \ldots, x_n)$, For the random pair (X, Y) there were only four: $f_X(x)$, $f_Y(y)$, $f_{X|Y}(x|y)$, and $f_{Y|X}(y|x)$.

We list them all for $n = 3$, all obtained from the joint pdf $f(x_1, x_2, x_3)$. Think of an outcome such as $\mathbf{X} = (X_1, X_2, X_3)$ where X_1 is lifetime in years, X_2 is average packs of cigarettes smoked per week, and X_3 is average weekly alcohol consumption in drinks/week. It's a good bet that the three variables will be related.

First the possible marginal distributions are obtained by averaging over the variable(s) we are not interested in:

$$\begin{aligned}
f_{X_1,X_2}(x_1,x_2) &= \int_{-\infty}^{\infty} f(x_1,x_2,x_3)dx_3 \\
f_{X_1,X_3}(x_1,x_3) &= \int_{-\infty}^{\infty} f(x_1,x_2,x_3)dx_2 \\
f_{X_2,X_3}(x_2,x_3) &= \int_{-\infty}^{\infty} f(x_1,x_2,x_3)dx_1 \\
f_{X_1}(x_1) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1,x_2,x_2)dx_2dx_3 \\
f_{X_2}(x_2) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1,x_2,x_2)dx_1dx_3 \\
f_{X_3}(x_3) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1,x_2,x_2)dx_1dx_2
\end{aligned}$$

There are *many* possible conditional distributions we can consider:

$$
\begin{aligned}
f_{X_1|X_2,X_3}(x_1|x_2,x_3) &= f(x_1,x_2,x_3)/f_{X_2,X_3}(x_2,x_3) \\
f_{X_2|X_1,X_3}(x_2|x_1,x_3) &= f(x_1,x_2,x_3)/f_{X_1,X_3}(x_1,x_3) \\
f_{X_3|X_1,X_2}(x_3|x_1,x_2) &= f(x_1,x_2,x_3)/f_{X_1,X_2}(x_1,x_2) \\
f_{X_1,X_2|X_3}(x_1,x_2|x_3) &= f(x_1,x_2,x_3)/f_{X_3}(x_3) \\
f_{X_1,X_3|X_2}(x_1,x_3|x_2) &= f(x_1,x_2,x_3)/f_{X_2}(x_2) \\
f_{X_2,X_3|X_1}(x_2,x_3|x_1) &= f(x_1,x_2,x_3)/f_{X_1}(x_1) \\
f_{X_1|X_2}(x_1|x_2) &= f_{X_1,X_2}(x_1,x_2)/f_{X_2}(x_2) \\
f_{X_1|X_3}(x_1|x_3) &= f_{X_1,X_3}(x_1,x_3)/f_{X_3}(x_3) \\
f_{X_2|X_1}(x_2|x_1) &= f_{X_1,X_2}(x_1,x_2)/f_{X_1}(x_1) \\
f_{X_2|X_3}(x_2|x_3) &= f_{X_2,X_3}(x_2,x_3)/f_{X_3}(x_3) \\
f_{X_3|X_1}(x_3|x_1) &= f_{X_1,X_3}(x_1,x_3)/f_{X_1}(x_1) \\
f_{X_3|X_2}(x_3|x_2) &= f_{X_2,X_3}(x_2,x_3)/f_{X_2}(x_2)
\end{aligned}
$$

Recall that (X_1, X_2, X_3) are *independent* iff $f(x_1, x_2, x_3) = f_{X_1}(x_1) f_{X_2}(x_2) f_{X_3}(x_3)$. When this is true, the marginal and conditional distributions simplify markedly (this holds for continuous **and** discrete random vectors, just replace pdf's by pmf's):

$$
\begin{aligned}
f_{X_1,X_2}(x_1, x_2) &= f_{X_1}(x_1) f_{X_2}(x_2) \\
f_{X_1,X_3}(x_1, x_3) &= f_{X_1}(x_1) f_{X_3}(x_3) \\
f_{X_2,X_3}(x_2, x_3) &= f_{X_2}(x_2) f_{X_3}(x_3) \\
f_{X_1}(x_1) &= f_{X_1}(x_1) \\
f_{X_2}(x_2) &= f_{X_2}(x_2) \\
f_{X_3}(x_3) &= f_{X_3}(x_3)
\end{aligned}
$$

$$
\begin{aligned}
f_{X_1|X_2,X_3}(x_1|x_2,x_3) &= f_{X_1}(x_1) \\
f_{X_2|X_1,X_3}(x_2|x_1,x_3) &= f_{X_2}(x_2) \\
f_{X_3|X_1,X_2}(x_3|x_1,x_2) &= f_{X_3}(x_3) \\
f_{X_1,X_2|X_3}(x_1,x_2|x_3) &= f_{X_1}(x_1)f_{X_2}(x_2) \\
f_{X_1,X_3|X_2}(x_1,x_3|x_2) &= f_{X_1}(x_1)f_{X_3}(x_3) \\
f_{X_2,X_3|X_1}(x_2,x_3|x_1) &= f_{X_2}(x_2)f_{X_3}(x_3) \\
f_{X_1|X_2}(x_1|x_2) &= f_{X_1}(x_1) \\
f_{X_1|X_3}(x_1|x_3) &= f_{X_1}(x_1) \\
f_{X_2|X_1}(x_2|x_1) &= f_{X_2}(x_2) \\
f_{X_2|X_3}(x_2|x_3) &= f_{X_2}(x_2) \\
f_{X_3|X_1}(x_3|x_1) &= f_{X_3}(x_3) \\
f_{X_3|X_2}(x_3|x_2) &= f_{X_3}(x_3)
\end{aligned}
$$

Note that for independent random variables, the conditional pdf's are simply the marginal pdf's. This implies, for example, that $P(X_2 \in A | X_1 = x_1) = P(X_2 \in A)$, that X_2 only cares about its marginal distribution; information on X_1 does not change probabilities concerning X_2.

In other words, if X_1 is independent of X_2, once we know $f_{X_1}(x_2)$, X_2 tells us nothing about X_1.

However, say X_1 and X_2 are independent and both distributed Bern(p) but *we do not know* p. Then X_1 and X_2 cannot be independent.

This ties into Bayesian statistics, sufficiency, and other ideas we'll explore later on.

We are often given a *random sample* – a collection of $(X_1, \ldots, X_n)$ that all have the same marginal distribution $f_X(x)$ and are independent. Finding joint, conditional, and marginal distributions is easy for a random sample. For example let X_1, X_2, X_3 be a random sample from $f_X(x) = \lambda e^{-\lambda x}$ on $x \geq 0$. Then

$$f(x_1, x_2, x_3) = f_X(x_1) f_X(x_2) f_X(x_3) = \lambda^3 e^{-\lambda(x_1+x_2+x_3)}$$

for $x_1 \geq 0$, $x_2 \geq 0$, $x_3 \geq 0$. This property, that the joint distribution of a random sample is the product of the marginal distributions, will come in handy later on when we compute the *likelihood* function $\mathcal{L}(\lambda)$ given the values (x_1, x_2, x_3) of a random sample.

Independence among X_1, X_2, and X_3 tells us

$$f_{X_1|X_2,X_3}(x_1|x_2, x_3) = f_X(x_1) = \lambda e^{-\lambda x_1},$$

for $x_1 \geq 0$. X_1 doesn't care about X_2 and X_3!

3.6 Functions of jointly distributed random variables

Let (X, Y) be jointly defined. We may be interested in functions such as X/Y, $X+Y$, XY, or X^Y. For discrete (X, Y) obtaining the distribution of such functions can be straightforward. For (X, Y) continuous this is typically *not* the case unless we make more stringent assumptions, for example that X and Y are independent and have the same marginal distribution.

On pp. 101-103 there is a general method for continuous random vectors, sometimes termed the “Jacobian method” or “transformation of variables” in calculus, that is a generalization of the the method we derived for simple functions $Y = g(X)$ considered in the last chapter. We’ll look through this in class.

For our purposes, we’ll go over a few important results.

Result: Let X and Y be independent. Let $g(x)$ and $h(y)$ be any functions. Then $g(X)$ is independent of $h(Y)$.

Example: Let (X, Y) be the height X in inches of a randomly selected female from the U. of M. graduate program, and let Y be the mileage on the car she drives. Assume X is independent of Y Then $X/12$, how tall she is in feet, is independent of $5280Y$, the number of feet her car has been driven.

Example: Let $X \sim N(10, 2)$ independent of $Y \sim N(-5, 8)$. Then $X - 10 \sim N(0, 2)$ is independent of $Y/5 \sim N(-1, 8/25)$.

Moment generating functions are useful tools for finding the distribution of *sums* of independent random variables, and other functions (Properties C and D, pp. 158-159).

Result: Let $X_1, \ldots, X_n$ be independent. Furthermore, let $X_j \sim N(\mu_j, \sigma_j^2)$ for $j = 1, \ldots, n$. Then

$$X_1 + X_2 + \cdots + X_n \sim N(\mu_1 + \mu_2 + \cdots + \mu_n, \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2).$$

Example: Let $X \sim N(10, 2)$ independent of $Y \sim N(-5, 8)$. Then $X + Y \sim N(10 - 5, 2 + 8) = N(5, 10)$.

Also: $X - 10 + Y/5 \sim N(0 - 1, 2 + 8/25) = N(-1, 58/25)$ from the previous slide and the fact that $X - 10$ is independent of $Y/5$.

Example: Say $X \sim N(0,1)$. What is the distribution of $10X$? How about $X+3$? How about $10X+3$? How about $10(X+3)$?

Example: Say $X \sim N(70.8, 6.5)$. What is the distribution of $\frac{X-70.8}{\sqrt{6.5}}$?

Example: Say $X_1 \sim N(\mu_1, \sigma^2/n_1)$ independent of $X_2 \sim N(\mu_2, \sigma^2/n_2)$. What is the distribution of $X_1 - X_2$?

Example: $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ independently, for $i = 1, \ldots, n$. What is the density of $\mathbf{Y} = (Y_1, \ldots, Y_n)$?

Example: $Y_i \sim \text{Bern}\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)$ independently, for $i = 1, \ldots, n$. What is the pmf of $\mathbf{Y} = (Y_1, \ldots, Y_n)$?

Example: Say $Y_1, Y_2, \ldots, Y_n$ are independent and all distributed $N(\mu, \sigma^2)$. What is the distribution of

$$\bar{Y} = \frac{Y_1 + Y_2 + \cdots + Y_n}{n}?$$

Example: Let (X, Y) be jointly distributed with pdf

$$f(x,y) = \begin{cases} 12x^2y^3 & \text{for } \begin{array}{c} 0 \leq x \leq 1 \\ 0 \leq y \leq 1 \end{array} \\ 0 & \text{otherwise} \end{cases}$$

1. Does the joint pdf factor into $f(x,y) = g(x)h(y)$? If so, what are these functions?
2. Are X and Y independent? Why or why not? What are the marginal densities of X and Y?
3. Find the conditional pdf's $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$. HINT: If X and Y are independent, this is easy...

Example:

Let $W \sim N(10, 2)$ be the time in minutes spent waiting for the bus to come, $R \sim N(20, 4)$ be the time spent riding on the bus, and $S \sim N(5, 1)$ be the time spent strolling to my office after riding the bus.

1. What is the distribution of T, the total time it takes to get to work?
2. It's 8am and I've stupidly scheduled a meeting at 8:30am. What is the probability that I will be on time for the meeting?

Example: Let $D = 0, 1$ denote whether or not a randomly selected pig truly has toxoplasmosis (no, yes). Let $T = 0, 1$ indicate that the pig tests negative or positive for toxoplasmosis. Say (D, T) are jointly distributed with pmf

d	t	$p(d, t)$
0	0	0.5980
0	1	0.0520
1	0	0.0945
1	1	0.2555

1. What is the prevalence of toxoplasmosis $P(D = 1)$?
2. What is the sensitivity of the test $P(T = 1|D = 1)$? What is the specificity $P(T = 0|D = 0)$?
3. If a pig tests positive, what is the probability the pig has the disease $P(D = 1|T = 1)$?
4. Are D and T independent? Why or why not?