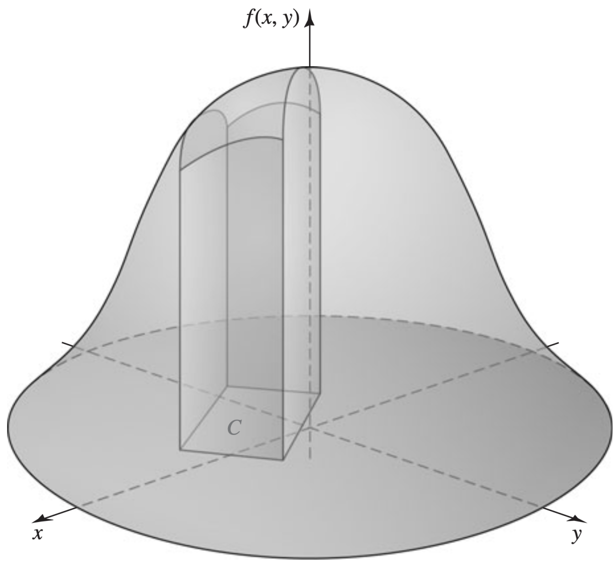

Figure 4.2: A bivariate density is a surface in three dimensional space, and probabilities are volumes beneath the curve. The total volume beneath the curve must be one.

4.2 Jointly absolutely continuous random pairs

Definition 4.6. If (X, Y) are jointly absolutely continuous, then their joint distribution can be summarized with a **joint probability density function** (**joint pdf**). This is a bivariate function $f_{XY} : \mathbb{R}^2 \to \mathbb{R}$ with the following properties:

- $f_{XY}(x, y) \geq 0$;
- $P\left((X, Y) \in \mathbb{R}^2\right) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y)\,\mathrm{d}x\mathrm{d}y = 1.$

Given the joint pdf, you can compute the joint distribution by integrating:

$$P((X, Y) \in C) = \iint_C f_{XY}(x, y)\,\mathrm{d}x\mathrm{d}y, \quad C \subseteq \mathbb{R}^2.$$

In the univariate case, probability corresponds to area under the density curve. In the bivariate case, probability corresponds to volume under the density *surface*, as in Figure 4.2.

Definition 4.7. If (X, Y) are jointly absolutely continuous, then each random variable X and Y has a **marginal probability density function** (**marginal pdf**) that describes the distribution

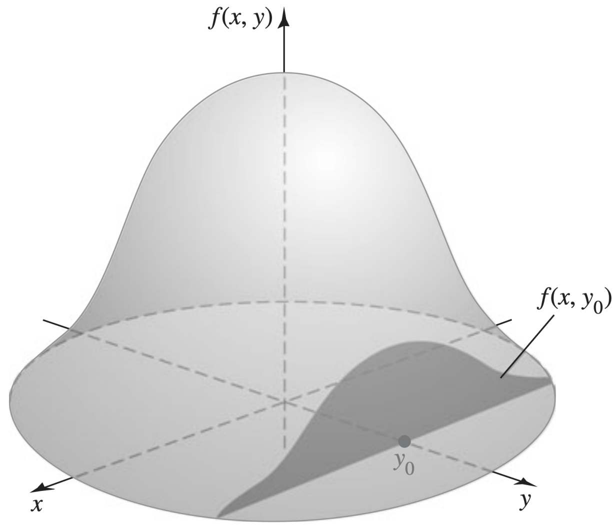

Figure 4.3: The conditional density is a renormalized cross-section of the joint density, where we hold the argument of the conditioning variable constant.

of that random variable on its own:

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)\,\mathrm{d}y \tag{4.13}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y)\,\mathrm{d}x. \tag{4.14}$$

To compute f_X, for example, we say that we *integrate y out* of the joint density.

To see where the marginal density formulas come from, consider the marginal CDF of one of the variables:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f_{XY}(t, y)\mathrm{d}y\mathrm{d}t.$$

The marginal PDF is just the derivative, so the fundamental theorem of calculus tells us that

$$f_X(x) = F_X'(x) = \frac{\mathrm{d}}{\mathrm{d}x} \int_{-\infty}^{x} \int_{-\infty}^{\infty} f_{XY}(t, y)\mathrm{d}y\mathrm{d}t = \int_{-\infty}^{\infty} f_{XY}(x, y)\mathrm{d}y.$$

Definition 4.8. The **conditional probability density function** (**conditional pdf**) of X or Y describes its revised behavior *conditional* on the event $Y = y$ or $X = x$, respectively. They are defined to be

$$f_{X|Y}(x \,|\, y) = \frac{f_{XY}(x, y)}{f_Y(y)} \tag{4.15}$$

$$f_{Y|X}(y \,|\, x) = \frac{f_{XY}(x, y)}{f_X(x)}. \tag{4.16}$$

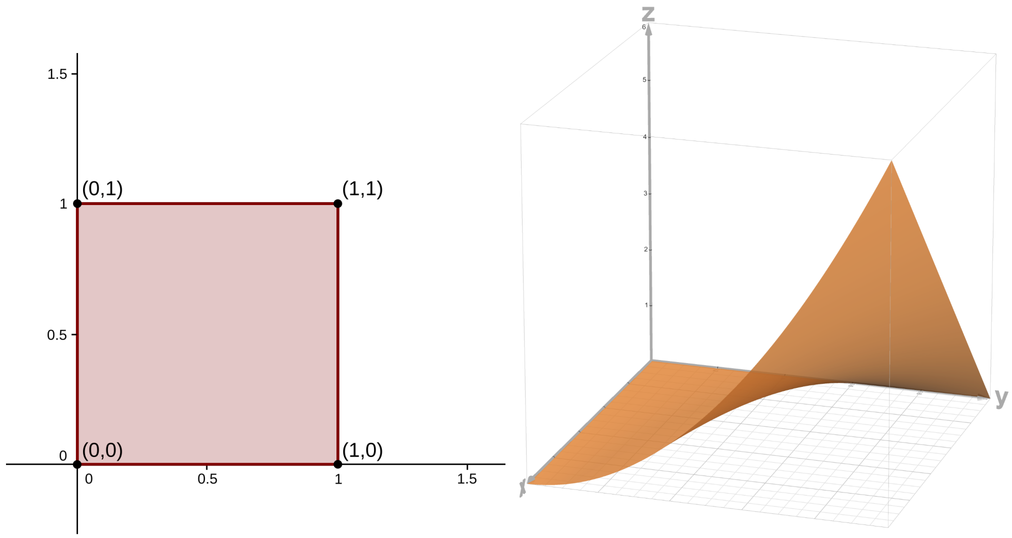

Figure 4.4: In Example 4.4, the joint range is the unit square on the left, and the joint density is the surface on the right.

So $f_{X|Y}(x \mid y)$, for example, is the cross-section of f_{XY} that holds the second argument fixed at y but lets x vary. The denominator renormalizes to ensure that the new *conditional* density integrates to 1 in x. See Figure 4.3 for a cartoon.

Theorem 4.4. (**Marginal-conditional decomposition**) From Definition 4.8, we get "two identities for the price of one:"

$$f_{XY}(x,\, y) = f_{X|Y}(x \mid y) f_Y(y) = f_{Y|X}(y \mid x) f_X(x). \tag{4.17}$$

Corollary 4.5. (**Hierarchical representation**) In order to specify a joint distribution for $(X,\, Y)$, you can do so *hierarchically*:

$$\begin{aligned} X &\sim P_X \\ Y \mid X &\sim P_{Y|X}. \end{aligned}$$

Theorem 4.4 guarantees that we can stitch these two pieces together to get a valid joint distribution.

Theorem 4.6. (**Bayes' theorem, again**) Combining the definition of conditional density with the marginal-conditional decomposition gives a version of Bayes' theorem for the conditional PDFs:

$$f_{X|Y}(x \mid y) = \frac{f_{XY}(x,\, y)}{f_Y(y)} = \frac{f_{Y|X}(y \mid x) f_X(x)}{f_Y(y)} \tag{4.18}$$

$$f_{Y|X}(y \mid x) = \frac{f_{XY}(x,\, y)}{f_X(x)} = \frac{f_{X|Y}(x \mid y) f_Y(y)}{f_X(x)}. \tag{4.19}$$

Example 4.4. Let X and Y have joint density

$$f_{XY}(x,\, y) = \begin{cases} 6xy^2, & x,\, y \in (0,\, 1) \\ 0 & \text{else.} \end{cases}$$

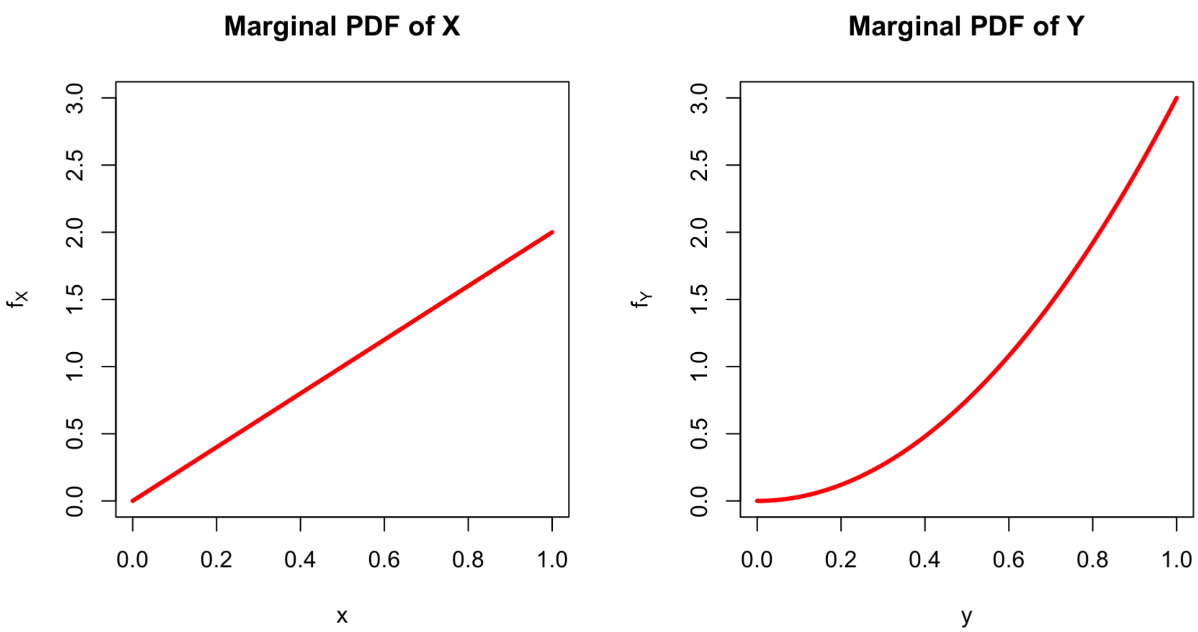

Figure 4.5: Marginal densities in Example 4.4

The joint range is $\text{Range}(X, Y) = \text{supp}(f_{XY}) = (0, 1) \times (0, 1)$, the unit square in the plane (Figure 4.4). The marginal densities are

$$
\begin{aligned}
f_X(x) &= \int_{-\infty}^{\infty} f_{XY}(x, y)\,\mathrm{d}y = \int_0^1 6xy^2\,\mathrm{d}y = [2xy^3]_0^1 = 2x, \quad x \in (0, 1).\\
f_Y(y) &= \int_{-\infty}^{\infty} f_{XY}(x, y)\,\mathrm{d}x = \int_0^1 6xy^2\,\mathrm{d}x = [3x^2y^2]_0^1 = 3y^2, \quad y \in (0, 1).
\end{aligned}
$$

The conditional densities are

$$
\begin{aligned}
f_{X|Y}(x \mid y) &= \frac{f_{XY}(x, y)}{f_Y(y)} = \frac{6xy^2}{3y^2} = 2x, \quad x \in (0, 1)\\
f_{Y|X}(y \mid x) &= \frac{f_{XY}(x, y)}{f_X(x)} = \frac{6xy^2}{2x} = 3y^2, \quad y \in (0, 1).
\end{aligned}
$$

We see that $f_{X|Y} = f_X$ and $f_{Y|X} = f_Y$, meaning that conditioning does not actually teach us anything. This motivates a definition.

Definition 4.9. A jointly absolutely continuous pair (X, Y) are **independent** if

$$f_{XY}(x, y) = f_X(x) f_Y(y) \quad \text{for all } (x, y) \in \mathbb{R}^2. \tag{4.20}$$

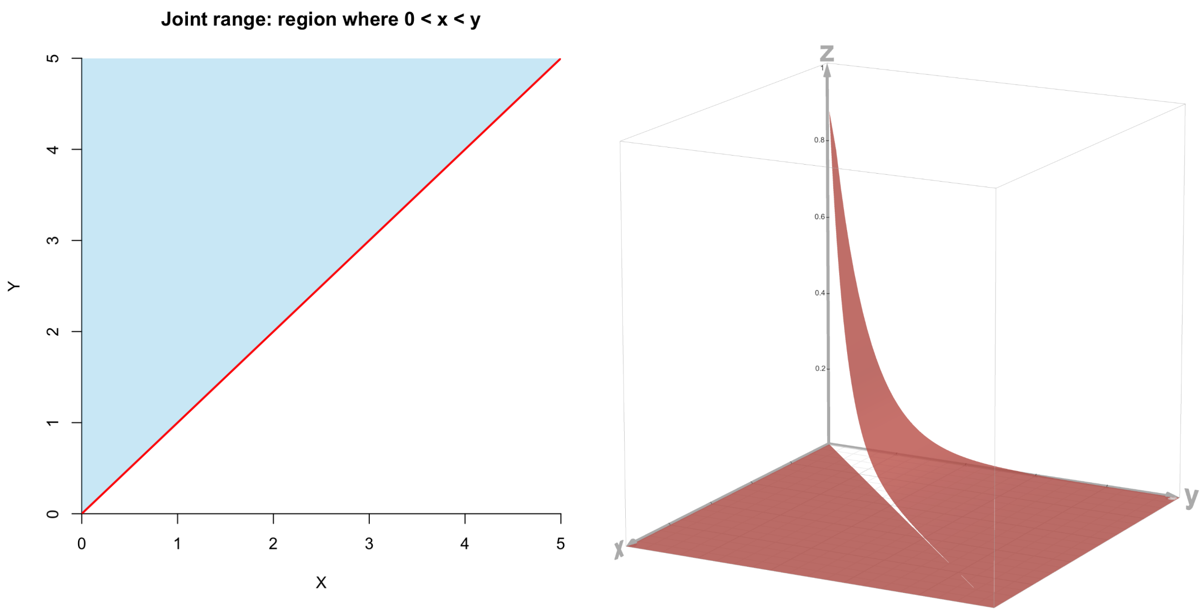

Figure 4.6: The joint density in Example 4.5.

Alternatively but equivalently, X and Y are independent if

$$f_{X|Y}(x \mid y) = f_X(x) \quad \text{for all } (x,\ y) \in \mathbb{R}^2 \tag{4.21}$$
$$f_{Y|X}(y \mid x) = f_Y(y) \quad \text{for all } (x,\ y) \in \mathbb{R}^2. \tag{4.22}$$

Example 4.5. Consider $(X,\ Y)$ with joint density

$$f_{XY}(x,\ y) = \begin{cases} e^{-y} & 0 < x < y < \infty \\ 0 & \text{else.} \end{cases}$$

The joint range and the density surface are displayed in Figure 4.6. Let's compute all of the marginals and conditionals:

$$\begin{aligned}
f_X(x) &= \int_{-\infty}^{\infty} f_{XY}(x,\ y)\,\mathrm{d}y = \int_x^{\infty} e^{-y}\,\mathrm{d}y = [-e^{-y}]_x^{\infty} = e^{-x}, && 0 < x \\
f_{Y|X}(y \mid x) &= \frac{f_{XY}(x,\ y)}{f_X(x)} = \frac{e^{-y}}{e^{-x}} = e^{-(y-x)}, && x < y \\
f_Y(y) &= \int_{-\infty}^{\infty} f_{XY}(x,\ y)\,\mathrm{d}x = \int_0^{y} e^{-y}\,\mathrm{d}x = e^{-y}\int_0^{y}\mathrm{d}x = ye^{-y}, && 0 < y \\
f_{X|Y}(x \mid y) &= \frac{f_{XY}(x,\ y)}{f_Y(y)} = \frac{e^{-y}}{ye^{-y}} = \frac{1}{y}, && 0 < x < y.
\end{aligned}$$

These are all familiar distributions:

$$\begin{aligned}
X &\sim \text{Exponential}(1) \\
Y &\sim \text{Gamma}(2,\ 1) \\
X \mid Y = y &\sim \text{Unif}(0,\ y).
\end{aligned}$$

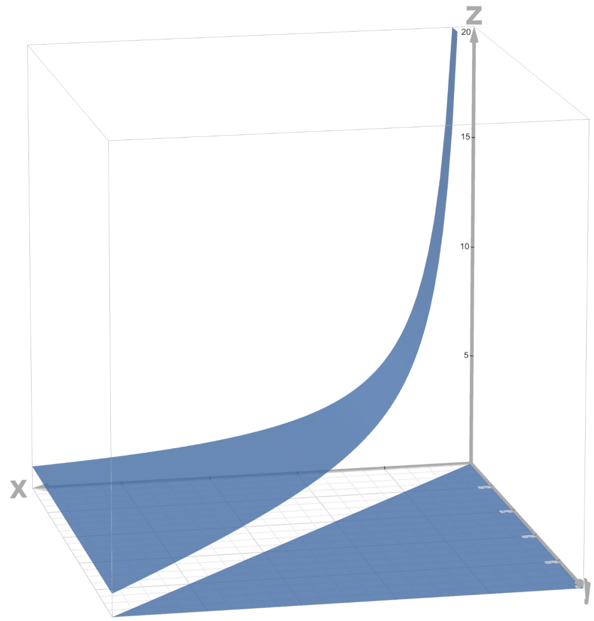

Figure 4.7: Joint density in Example 4.6.

Conditional on $X = x$, Y is Exponential(1), but shifted right by the amount x.

Example 4.6. Consider a joint distribution for (X, Y) written hierarchically:

$$\begin{aligned} X &\sim \text{Unif}(0, 1) \\ Y \mid X = x &\sim \text{Unif}(0, x). \end{aligned}$$

So

$$\begin{aligned} f_X(x) &= 1 && 0 \leq x \leq 1 \\ f_{Y|X}(y \mid x) &= \frac{1}{x} && 0 \leq y \leq x \end{aligned}$$

$$\begin{aligned} f_{XY}(x, y) &= f_{Y|X}(y \mid x) f_X(x) \\ &= \frac{1}{x} && 0 \leq y \leq x \leq 1. \end{aligned}$$

The joint density surface is displayed in Figure 4.7. The marginal density of Y is

$$\begin{aligned}
f_Y(y) &= \int_0^1 f_{XY}(x,\ y)\mathrm{d}x \\
&= \int_0^y f_{XY}(x,\ y)\mathrm{d}x + \int_y^1 f_{XY}(x,\ y)\mathrm{d}x \\
&= \int_0^y 0\,\mathrm{d}x + \int_y^1 \frac{1}{x}\mathrm{d}x \\
&= \int_y^1 \frac{1}{x}\mathrm{d}x \\
&= [\ln x]_y^1 \\
&= \ln 1 - \ln y \\
&= -\ln y, \qquad\qquad 0 < y < 1.
\end{aligned}$$

The conditional density is

$$f_{X|Y}(x \mid y) = \frac{f_{XY}(x,\ y)}{f_Y(y)} = \frac{1/x}{-\ln y}, \quad 0 \leq y \leq x \leq 1.$$

To compute the conditional expectation of X given $Y = y$, we apply the definition of expected value, but we use the *conditional* density:

$$\begin{aligned}
E(X \mid Y = y) &= \int_0^1 x f_{X|Y}(x \mid y)\,\mathrm{d}x \\
&= \int_y^1 x f_{X|Y}(x \mid y)\,\mathrm{d}x \\
&= \int_y^1 x \frac{1/x}{-\ln y}\,\mathrm{d}x \\
&= \frac{1}{-\ln y}\int_y^1 \mathrm{d}x \\
&= \frac{1-y}{-\ln y}.
\end{aligned}$$

Example 4.7. Consider the joint distribution of random variables X and Y, written in hierarchical form:

$$\begin{aligned}
X &\sim \text{Gamma}\left(\tfrac{\nu}{2},\ \tfrac{\nu}{2}\right) \\
Y \mid X = x &\sim \text{N}\,(0,\ 1/x).
\end{aligned}$$

This models a normal variable with a "random variance". To derive the marginal pdf of Y, we fix

$y \in \mathbb{R}$, and get ready for some hardcore "massage and squint":

$$
\begin{aligned}
f_Y(y) &= \int_{-\infty}^{\infty} f_{Y\,|\,X}(y\,|\,x) f_X(x)\,\mathrm{d}x \\
&= \int_{-\infty}^{0} f_{Y\,|\,X}(y\,|\,x) f_X(x)\,\mathrm{d}x + \int_{0}^{\infty} f_{Y\,|\,X}(y\,|\,x) f_X(x)\,\mathrm{d}x \\
&= \cancel{\int_{-\infty}^{0} f_{Y\,|\,X}(y\,|\,x)\cdot 0\,\mathrm{d}x} + \int_{0}^{\infty} f_{Y\,|\,X}(y\,|\,x) f_X(x)\,\mathrm{d}x \\
&= \int_0^\infty \underbrace{\frac{1}{\sqrt{2\pi\frac{1}{x}}} e^{-\frac{1}{2}\frac{y^2}{1/x}}}_{\text{pdf of N}(0,1/x)} \underbrace{\frac{\left(\frac{\nu}{2}\right)^{\nu/2}}{\Gamma\left(\frac{\nu}{2}\right)} x^{\frac{\nu}{2}-1} e^{-\frac{\nu}{2}x}}_{\text{pdf of Gamma}(\nu/2,\,\nu/2)}\,\mathrm{d}x \\
&= \frac{\nu^{\frac{\nu}{2}} 2^{-\frac{\nu}{2}} 2^{-\frac{1}{2}}}{\sqrt{\pi}\cdot\Gamma\left(\frac{\nu}{2}\right)} \int_0^\infty x^{\frac{\nu}{2}+\frac{1}{2}-1} e^{-\frac{1}{2}(\nu+y^2)x}\,\mathrm{d}x \\
&= \frac{\nu^{\frac{\nu}{2}} 2^{-\frac{\nu}{2}} 2^{-\frac{1}{2}}}{\sqrt{\pi}\cdot\Gamma\left(\frac{\nu}{2}\right)} \frac{\Gamma\left(\frac{\nu}{2}+\frac{1}{2}\right)}{\left[\frac{1}{2}(\nu+y^2)\right]^{\frac{\nu}{2}+\frac{1}{2}}} \int_0^\infty \underbrace{\frac{\left[\frac{1}{2}(\nu+y^2)\right]^{\frac{\nu}{2}+\frac{1}{2}}}{\Gamma\left(\frac{\nu}{2}+\frac{1}{2}\right)} x^{\frac{\nu}{2}+\frac{1}{2}-1} e^{-\frac{1}{2}(\nu+y^2)x}}_{\text{pdf of Gamma}\left(\alpha=\frac{\nu}{2}+\frac{1}{2},\,\beta=\frac{1}{2}(\nu+y^2)\right)}\,\mathrm{d}x \\
&= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi}\cdot\Gamma\left(\frac{\nu}{2}\right)} \nu^{\frac{\nu}{2}} 2^{-\frac{\nu}{2}} 2^{-\frac{1}{2}} \left[2^{-1}(\nu+y^2)\right]^{-\left(\frac{\nu}{2}+\frac{1}{2}\right)} \\
&= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi}\cdot\Gamma\left(\frac{\nu}{2}\right)} \nu^{\frac{\nu}{2}} \cancel{2^{-\frac{\nu}{2}} 2^{-\frac{1}{2}} 2^{\frac{\nu}{2}} 2^{\frac{1}{2}}} (\nu+y^2)^{-\left(\frac{\nu}{2}+\frac{1}{2}\right)} \\
&= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi}\cdot\Gamma\left(\frac{\nu}{2}\right)} \nu^{\frac{\nu}{2}} \left[\nu\left(1+\frac{1}{\nu}y^2\right)\right]^{-\left(\frac{\nu}{2}+\frac{1}{2}\right)} \\
&= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi}\cdot\Gamma\left(\frac{\nu}{2}\right)} \cancel{\nu^{\frac{\nu}{2}} \nu^{-\frac{\nu}{2}}} \nu^{-\frac{1}{2}} \left(1+\frac{1}{\nu}y^2\right)^{-\left(\frac{\nu}{2}+\frac{1}{2}\right)} \\
&= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\cdot\Gamma\left(\frac{\nu}{2}\right)} \left(1+\frac{1}{\nu}y^2\right)^{-\frac{\nu+1}{2}}.
\end{aligned}
$$

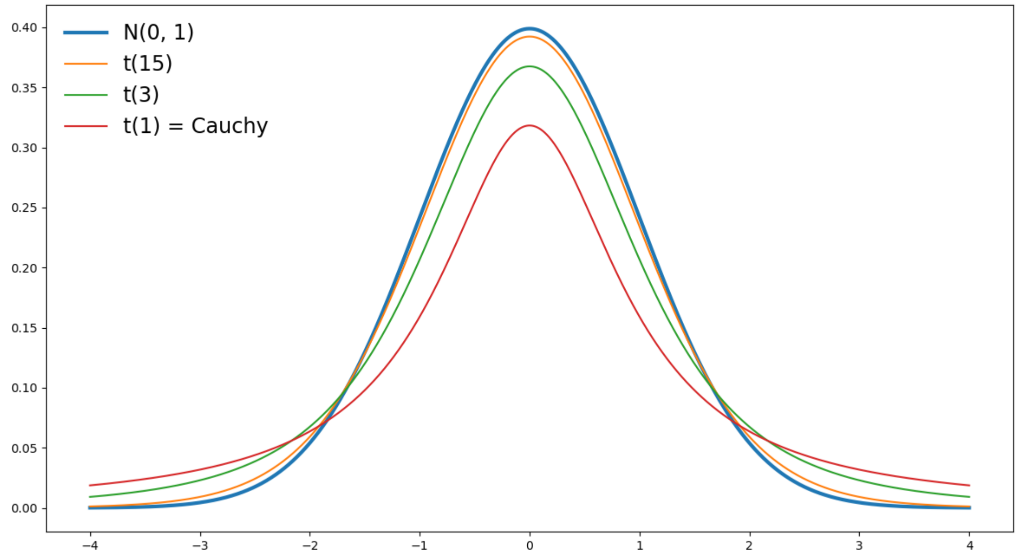

Figure 4.8

Definition 4.10. An absolutely continuous random variable X has **Student's t distribution** on $\text{Range}(X) = \mathbb{R}$ if its pdf is

$$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\cdot\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{1}{\nu}x^2\right)^{-\frac{\nu+1}{2}}, \quad x \in \mathbb{R}. \tag{4.23}$$

The parameter $\nu > 0$ is called the **degrees of freedom**. We denote this $X \sim t_\nu$.

It turns out that the Cauchy distribution is the special case of Student's t when $\nu = 1$:

$$f_X(x) = \frac{\Gamma(1)}{\Gamma(1/2)\sqrt{\pi}}(1+x^2)^{-1} = \frac{1}{\pi}\frac{1}{(1+x^2)}.$$

This follows because $\Gamma(1) = 1$ and $\Gamma(1/2) = \sqrt{\pi}$. We see in Figure 4.8 that members of this family have a bell-shaped pdf with heavier tails than the normal, but the tails get thinner as ν increases. In fact, as $\nu \to \infty$, t_ν converges to N(0, 1). Table 4.3 describes the situation with the moments. We know that for $\nu = 1$, the Cauchy distribution has no finite moments. As ν increases and the tails become less heavy, we accumulate more and more finite moments. In the extreme case of $t_\infty = \text{N}(0, 1)$, finally every moment is finite. But for finite $0 < \nu < \infty$, we eventually run out. So the mgf of Student's t distribution is undefined.

feature	**when does it exist?**	**value when it exists**
$E(X)$	$\nu > 1$	0
$\text{var}(X)$	$\nu > 2$	$\frac{\nu}{\nu-2}$
$\text{skew}(X)$	$\nu > 3$	0
$\vdots$	$\vdots$	$\vdots$

Table 4.3: As the degrees of freedom of Student's t increase, the tails become thinner, and the higher moments become finite.

Figure 4.9: William Gosset (1876 - 1937) published on the t distribution under the pseudonym *Student* while he was an employee of the Guinness brewery in Ireland.

4.3 Tower property (or, "law of total expectation")

Here's another useful shortcut for computing expected values:

Theorem 4.7. If X and Y are jointly distributed random variables, then we can compute the marginal expected value of X according to:

$$E(X) = E(E(X \mid Y)). \tag{4.24}$$

The inner expectation is a function of Y, and the outer expectation is with respect to the marginal distribution of Y. So the outer expectation is computed using LOTUS, where $g(Y) = E(X \mid Y)$ is the transformation.

Partial proof. Consider the special case where X and Y are continuous with joint density f_{XY}. The result is true even if the random variables are not all continuous, but the assumption simplifies the proof. Recall the following:

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x \\ E(X \mid Y = y) &= \int_{-\infty}^{\infty} x f_{X|Y}(x \mid y)\,\mathrm{d}x. \end{aligned}$$

So now, we treat $E(X \mid Y)$ as a transformation of Y, and we compute *its* expected value using LOTUS:

$$\begin{aligned} E(E(X \mid Y)) &= \int_{-\infty}^{\infty} E(X \mid Y = y) f_Y(y)\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X|Y}(x \mid y)\,\mathrm{d}x\, f_Y(y)\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X|Y}(x \mid y) f_Y(y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{XY}(x,\, y)\,\mathrm{d}x\,\mathrm{d}y && \text{MC decomp} \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{XY}(x,\, y)\,\mathrm{d}y\,\mathrm{d}x && \text{Fubini} \\ &= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{XY}(x,\, y)\,\mathrm{d}y\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x \\ &= E(X). \end{aligned}$$

□

Example 4.8. Recall Example 4.3 where $N \sim \text{Poisson}(\lambda)$ and $X \mid N = n \sim \text{Binom}(n,\, p)$. We saw that marginally, $X \sim \text{Poisson}(p\lambda)$, and so we know that $E(X) = p\lambda$. But if we wanted to jump directly to the marginal expected value, the tower property makes it easy:

$$E(X \mid N) = Np$$

$$\begin{aligned} E(X) &= E(E(X \mid N)) \\ &= E(Np) \\ &= pE(N) \\ &= p\lambda. \end{aligned}$$

Example 4.9. Recall Example 4.6 where $X \sim \text{Unif}(0, 1)$ and $Y \mid X = x \sim \text{Unif}(0, x)$. We saw that the marginal density of Y is $f_Y(y) = -\ln y$ for $0 < y < 1$, so the marginal expected value is

$$E(Y) = \int_0^1 -y \ln y \, \mathrm{d}y.$$

This is a very unpleasant calculation, but the tower property makes it easy:

$$E(Y \mid X) = X/2$$

$$\begin{aligned} E(Y) &= E(E(Y|X)) \\ &= E(X/2) \\ &= E(X)/2 \\ &= (1/2)/2 \\ &= 1/4. \end{aligned}$$

easy peasy lemon squeezy. Furthermore, we already know that $E(X) = 1/2$, but let's use the tower property to perform a sanity check:

$$\begin{aligned} E(X) &= E(E(X \mid Y)) \\ &= E\left(\frac{1-Y}{-\ln Y}\right) \\ &= \int_0^1 \frac{1-y}{-\ln y} f_Y(y) \, \mathrm{d}y \\ &= \int_0^1 \frac{1-y}{-\ln y}(-\ln y) \, \mathrm{d}y \\ &= \int_0^1 (1-y)\mathrm{d}y \\ &= [y - 0.5y^2]_0^1 \\ &= 1 - 1/2 \\ &= 1/2. \end{aligned}$$