CHAPTER 3

SOME UNIVARIATE DISTRIBUTIONS

> I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the 'law of error.' A savage, if he could understand it, would worship it as a god. It reigns with severity in complete self-effacement amidst the wildest confusion. The huger the mob and the greater the anarchy the more perfect is its sway.
>
> FRANCIS GALTON *J. Anthrop. Institute*, **15**, 487-499 (1886)

> Experimenters imagine that it [the normal distribution] is a theorem of mathematics, and mathematicians believe it to be an experimental fact.
>
> GABRIEL LIPPMANN, attributed

3.1 Introduction

In this chapter we describe some of the more commonly encountered probability distributions, giving both pdf's and cdf's. We mentioned "exploratory data analysis" as what we do when we first look at data; part of this can be finding a pdf to model the data. If we can use a well-known pdf, we can then use the many well-established results on how to analyze data with that pdf; if we have to develop a whole new probability model, we are in for a lot more work. In Section 2.6.3 we discussed some ways to see if data match a given pdf, something we revisit more rigorously in Section 6.5.1; for now we simply offer a guide to some of the most useful probability density functions. We also describe how to generate a series of numbers that mimic a random variable with a given pdf, since (Chapter 5) there are statistical methods that depend on being able to do this.

Most discussions of pdf's are unduly complicated because, for historical reasons, different names and symbols are used for parameters with the same role – and some of the names (notably "number of degrees of freedom") are not very informative.

To minimize this complexity we note that for continuously distributed random variables nearly all pdf's have formulas like

$$\frac{1}{cA_r(s)}\phi_s\left(\frac{x-l}{c}\right) \qquad \text{for} \quad L_b \le x \le L_c \tag{3.1}$$

In this expression the L's give the range of the variable: often from $-\infty$ to ∞, or from 0 to ∞, but sometimes over a finite range. The function ϕ_s gives the actual shape of the pdf; the constant $A_r(s)$ is the area under the function, included to normalize the integral of ϕ to unity. We call s the **shape parameter**; not all pdf's have one. But almost all pdf's do have two others:

1. A **location parameter** l, which sets the location of the pdf on the x-axis. This parameter appears mostly, though not always, for pdf's on $(-\infty,\infty)$.

2. A **scale parameter** c, which expands or contracts the scale of the x-axis. For the pdf to remain properly normalized, c also has to scale the size of the pdf, and so multiplies $A_r(s)$.

We will give each pdf in three forms: (1) a stripped-down form, usually with l set to 0 and c to 1; (2) in the full form, but with all the parameters, and using the L, c, and s symbols, and (3) in the full form, but with all the parameters given their conventional symbols. Though the first two expressions are not equal, we will use the equal sign to connect them. You should try to see how equation 3.1 above can be applied to the stripped-down form to produce the conventional one.

3.2 Uniform Distribution

The simplest probability density function is the **uniform** or **rectangular** distribution, which has the probability density function (pdf)

$$\phi(x) = \begin{array}{ll} 1 & 0 \le x \le 1 \\ 0 & x < 0 \quad \text{or} \quad x > 1 \end{array} \tag{3.2}$$

and the cumulative distribution function (cdf)

$$\Phi(x) = \begin{array}{ll} 0 & x < 0 \\ x & 0 \le x \le 1 \\ 1 & x > 1 \end{array} \tag{3.3}$$

In this case the "conventional" formula is the one without scaling, and we indicate that a random variable X has this distribution by writing $X \sim U(0,1)$. We can apply l and c to move the nonzero part to any location, and make it of any finite length, in which case we have

$$\phi(x) = c^{-1} \quad \text{for} \quad l \le x \le l + c$$

This distribution is used mostly in simulation methods, as the basis for computing rv's with other pdf's; most computer software contains a function named `ran` (or some similar name), repeated calls to which are supposed to produce numbers with this distribution. We say "supposed to" because the numbers produced by actual software can depart from this ideal distribution in two ways:

- The first departure is that any computer function is deterministic; so what it actually produces is a set of numbers designed to "look random"; hence these are called **pseudorandom** numbers. Almost always, `ran` includes an argument (called a **seed**); different seeds create different sets of random numbers, so if we let the seed depend on (say) the time, we can get random sets that vary every time. But if we want to debug a program it is very useful to always be able to use the same sequence.

- The second departure can be very harmful: many computer routines in fact do not produce a collection of uniformly distributed and independent random numbers. If your work depends on the output of `ran` being very like a uniform distribution, you should use a known algorithm, not just the default function for your system. *Press et al.* [1992] provide a good basic discussion, and *L'Ecuyer and Simard* [2007] gives a full review, which notes that the "Mersenne twister" of *Matsumoto and Nishimura* [1998], passes all the tests known at the time. Another algorithm, adjustable to different levels of randomness, is given by *Lüscher* [1994] and *James* [1994].

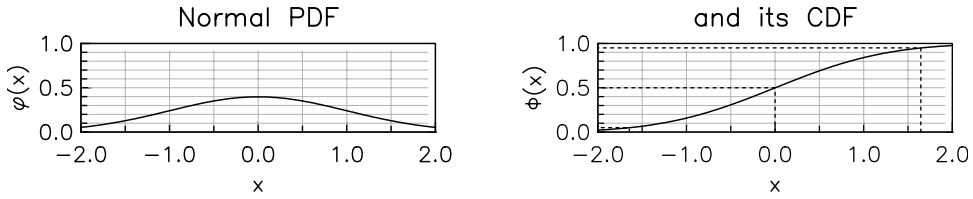

Figure 3.1: A Normal, or Gaussian, pdf and its cdf, plotted for zero mean and unit variance. In this figure, unlike most of the others, the y axis is not exaggerated relative to the x axis.

x	± 1.00	± 1.65	± 1.96	± 2.58	± 3.29	± 3.90
Mass fraction	0.68	0.90	0.95	0.99	0.999	0.9999

3.3 The Normal (Gaussian) Distribution

We have already met this pdf,[1] but present it again to illustrate our different ways of writing a pdf:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \qquad = \frac{1}{c\sqrt{2\pi}} e^{-(x-l)^2/2c^2} \qquad = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \tag{3.4}$$

where the location parameter (conventionally μ) is called the **mean** and the scale parameter (conventionally σ) is called the **standard deviation**. Figure 3.1 shows the pdf and cdf for this distribution, with the dashed lines showing where Φ attains the values 0.05, 0.50, and 0.95. The 0.5 and 0.95 lines bound 0.9 (90%) of the area under the pdf (often called, since this is a "density function," the mass). Table 3.3 gives some values of the mass as a function of different x-values (plus and minus). We would, for example, expect that out of 1000 rv's with this distribution, no more than one would be more than 3.3σ away from the central location μ. We described in Section 2.12 why this pdf is special, namely that the central limit theorem says that (for example) sums of rv's approach this distribution somewhat irrespective of their own distribution. It very often simplifies the theory to assume Gaussian behavior (we will see many examples of this), and many datasets are normally distributed, at least approximately. But you should never casually assume it.

This is a good place to introduce another term of statistical jargon, namely what a **standardized random variable** is. We say that any normally distributed random variable $X \sim N(\mu, \sigma)$ may be transformed to one

[1] For the history of the names, see *Stigler* [1980].

with the standard normal distribution (with $\mu = 0$ and $\sigma = 1$ by creating the new **standardized random variable** $Z = (X - \mu)/\sigma$, so that $Y \sim N(0,1)$. We will see other examples of such transformations.

3.3.1 Generating Normal Deviates

The title of this section contains one of those terms (like Love waves) liable to bring a smile until you get used to it; but "deviates" is the standard term for what we have called a collection of random numbers. There are quite a few ways of producing random numbers with a Gaussian pdf; for very large simulations we need methods that are fast and also reliably produce the relatively rare large values [*Thomas et al.*, 2007]. We describe two related methods; both allow us to give some examples of the procedures developed in Section 2.10 for finding the pdf's of functions of random variables.

In both methods, we start by getting two rv's, with an independent and identical distribution, namely a uniform one; as noted in Section 3.2, this is usually easy given the usual `ran` function in software. These two variables can viewed as specifying a point in a square, with $0 \le x_1 \le 1$ and $0 \le x_2 \le 1$, and the pdf describing the distribution of these points is a uniform bivariate one, with $\phi(x_1, x_2) = 1$ over this square region. We term these two rv's U_1 and U_2, though we use x_1 and x_2 for the actual numbers.

The first method, known as the **Box-Muller transform**, computes two values given by

$$y_1 = \sqrt{-2\ln x_1}\cos 2\pi x_2 \qquad \text{and} \qquad y_2 = \sqrt{-2\ln x_1}\sin 2\pi x_2$$

which (we assert) each has a Normal distribution. To show this, we have to find the pdf's of the rv's

$$Y_1 = \sqrt{-2\ln U_1}\cos 2\pi U_2 \qquad \text{and} \qquad Y_2 = \sqrt{-2\ln U_1}\sin 2\pi U_2$$

where we have the same expressions but with random variables replacing the conventional ones.

First we write the joint pdf in polar coordinates, as $\phi_Y(r,\theta)$, the distribution in angle θ is uniform because U_2 is. The pdf in radius is found from equation (2.22) in Chapter 2, with $W = g(R) = \sqrt{-2\ln R}$, with R (like U_1) uniformly distributed. The inverse is $g^{-1}(w) = e^{-w^2/2}$; applying equation (2.22) and using the uniformity of R, gives

$$\phi_W = \phi_R(g^{-1}(w))\left|\frac{d}{dw}g^{-1}(w)\right| = we^{-w^2/2} \tag{3.5}$$

If Y_1 and Y_2 are (as we claim) iid variables with a Gaussian distribution, the joint pdf will be the product

$$\phi_Y = \frac{1}{2\pi} e^{-(x_1^2+x_2^2)/2} = \frac{1}{2\pi} e^{-r^2/2}$$

Although this only contains r, it remains a joint pdf. If we integrate over all θ we get a pdf in r alone:

$$\phi'_R = \int_0^{2\pi} \phi_Y r \, d\theta = r e^{-r^2/2} \tag{3.6}$$

Equations (3.5) and (3.6) give the same result: the pdf found for the transformed variable $W = g(R) = \sqrt{-2\ln R}$ for R uniformly distributed matches that for the radius of the bivariate Gaussian. So we see that taking this function of a uniformly distributed rv, and then multiplying the result by the unit vector in a random direction (which is what $(\cos 2\pi U_2, \sin 2\pi U_2)$ is) will produce a pair of Gaussian rv's.

Another way to get Normal deviates from uniform ones is to note that if the points are uniformly distributed over the square, they will also be uniformly distributed over the circle inscribed within it, which they can be limited to if we remove all pairs for which $r^2 = x_1^2 + x_2^2 \geq 1$. Then we form

$$y_1 = \frac{x_1}{r}\sqrt{-2\ln r^2} = x_1\sqrt{\frac{-2\ln r^2}{r^2}} \quad \text{and} \quad y_2 = \frac{x_2}{r}\sqrt{-2\ln r^2} = x_2\sqrt{\frac{-2\ln r^2}{r^2}}$$

where the second form requires only one square root. This method avoids calls to trigonometric functions, and thus is usually faster despite requiring 27% more calls to the uniform random number generator. It is equivalent to the Box-Muller transform because x_1/r and x_2/r are equivalent to a sine and cosine, and will be distributed uniformly in angle, while r^2 is also uniformly distributed, making it a valid replacement for x_1.

3.4 Point Processes

The next set of pdf's are all used for describing point processes, which we introduced in Section 1.2; so we start by giving a more formal description of the Poisson process, which we used as a simple model for geomagnetic reversals. In this process we suppose that the probability of some event occurring is equal over all time, and is described by a **rate** (also called an

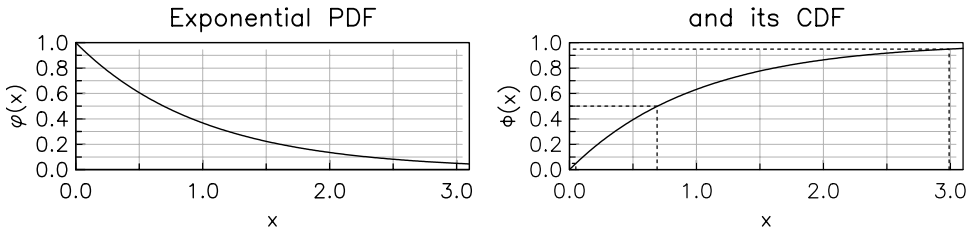

Figure 3.2: Exponential pdf and its cdf.

intensity), which has the dimensions of probability over time. The conditions for a Poisson process with rate λ are that, as a small time interval h approaches zero:

- The number of events in disjoint time intervals are independent: the number in $(0, t]$ is independent of (does not affect) the number in $(t,\ t+h)$.

- The probability of an event in an interval of length h is, as $h \to 0$, approximately proportional to h plus a remainder that is $o(h)$. (Remember that a function $g(h)$ is $o(h)$ if $\lim_{h\to 0} g(h)/h = 0$.)

- the probability of more than one event in an interval of length h is $o(h)$.

The Poisson process is **memoryless**: what happens at any time does not depend on earlier history. This makes the Poisson process the simplest point process, described by only one parameter and fundamentally unpredictable. In earthquake statistics and other areas the Poisson process often serves as a kind of "least interesting" model, against which we compare data to see if there is anything more complicated going on.

A more general form of point process is the **renewal process**, in which the probability of an event depends on the time since the last event. The Poisson process might be called a renewal process that isn't, since in its case the probability is constant, and the time of the next event is uninfluenced by the time of the previous one. A renewal process might be appropriate for geomagnetic field reversals, since we might expect that immediately following a geomagnetic reversal, while the geodynamo recovers its stable polarity, a second reversal might be less likely, and short intervals less probable, than under a pure Poisson model. Conversely, for earthquakes a

second event is more likely right after one has occurred, so the events are clustered. Renewal models can imitate both types of behaviors.

3.4.1 Exponential Distribution

For a Poisson process it can be shown that the interval between successive occurrences has an **exponential** distribution. The pdf for this is defined over $[0,\infty)$:

$$\phi(x) = e^{-x} = ce^{-x/c} = \lambda e^{-\lambda x}$$

and the cumulative distribution function is

$$\Phi(x) = 1 - e^{-\lambda x} \tag{3.7}$$

This pdf is peaked towards zero, so even though the probability of occurrence does not vary with time, short intervals have a much higher probability of occurring than long ones do. Figure 3.2 shows this distribution; again, the dotted lines show the values for Φ equal to 0.05, 0.50, and 0.95: much less symmetric than the same points for the Normal.

Producing random numbers for this distribution is very easy, and illustrates a method that can be applied to some other pdf's. We can think of taking uniformly distributed rv's and placing them on the y-axis of the cdf; then if we map these into the x-axis through the inverse cdf function, the result will have the distribution we want. This result is general; how usable it is depends on how easy it is to compute the inverse function for the cdf. In this case it is easy; from equation (3.7) the inverse cdf function is

$$\Phi^{-1}(y) = \frac{-\ln(1-y)}{\lambda} = \frac{-\ln(y)}{\lambda}$$

where the last quality is gotten from the observation that if $1-Y$ is uniformly distributed between 0 and 1, Y will be also. This method depends on having an easy way to compute the inverse cdf; in Section 3.3.1 we needed more complicated methods because the cdf for the Gaussian does not have an easily-computed inverse.

3.4.2 Poisson Distribution

We get another distribution from the Poisson process by considering a different random variable: the number of events, k, occurring in a time interval of length T. The distribution of $k(T)$ is called a **Poisson distribution**

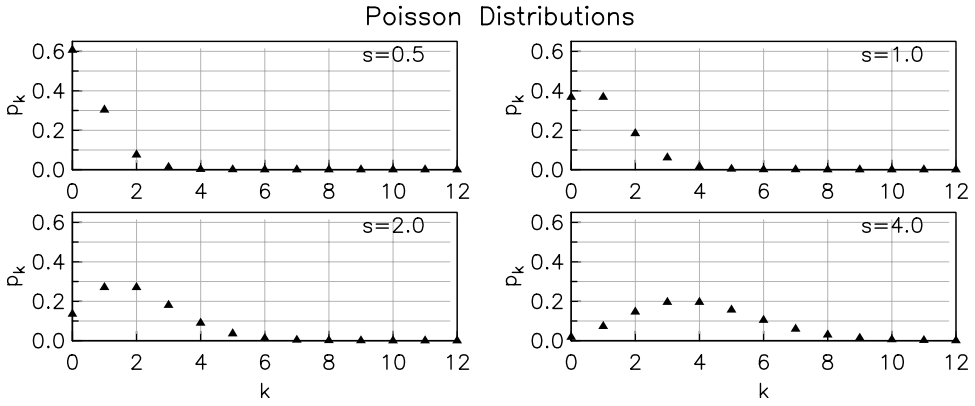

Figure 3.3: Poisson distributions for different values of s.

(Figure 3.3); the probability of getting k events is

$$p_k = s^k \frac{e^{-s}}{k!}$$

Here s is the shape parameter; this distribution does not have a scale or location parameter. For a Poisson process with rate λ, observed over an arbitrary interval of length T, the shape parameter is $s = \lambda T$.

This distribution is discrete, so we write the pdf using delta functions:

$$\phi(x) = \sum_{k=0}^{\infty} p_k \delta(x-k)$$

As s becomes large, the distribution starts to look like a discrete version of the Normal. The expected value (first moment) of the distribution is just the shape parameter s.

This distribution arises whenever we have a small probability of something happening in each of a large number of instances, and want to know the distribution of the number of "somethings" in a given instance. Magnetic reversals have a small probability of happening in each of many years; another example, not arising from a point process, would be deaths from being struck by lightning; if we took a number of groups of (say) 100,000 people, we would expect the number of such deaths in each group to be Poisson-distributed.[2]

[2] This has the same structure as the classic example for this distribution: the number of soldiers, in each corps of the Prussian army, who were killed each year by being kicked by a horse.

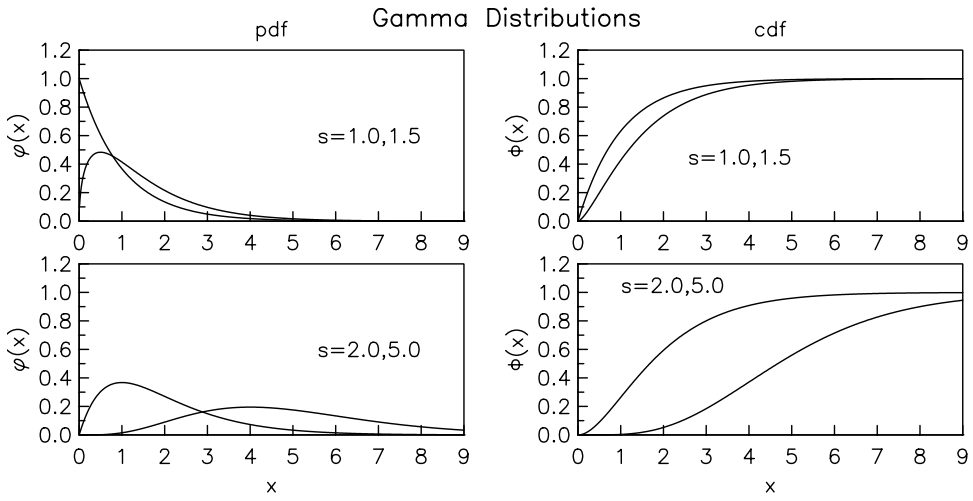

Figure 3.4: Gamma distributions for different values of s.

3.4.3 Gamma Distribution

A renewal process is often modeled by assuming that the interval lengths follow a **gamma distribution**; this has a pdf on $[0,\infty)$ that looks rather like a continuous version of the Poisson distribution:

$$\phi(x) = \frac{1}{\Gamma(s)} x^{s-1} e^{-x} = \frac{1}{c\Gamma(s)} \left(\frac{x}{c}\right)^{s-1} e^{-x/c} = \frac{\lambda^s}{\Gamma(s)} x^{s-1} e^{-\lambda x} \tag{3.8}$$

where the Γ function (used to normalize the distribution) is defined by

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u}\, du$$

As in the previous example we have followed the convention for dealing with point processes and written the scale parameter as $\lambda = c^{-1}$. Equation (3.8) shows that the exponential distribution is a gamma distribution with $s = 1$.

Figure 3.4 shows the gamma density function for different values of s. For a renewal point process governed by this kind of probability density function, λ describes the rate of events well after each event, while s controls the shape of the probability function immediately following each event. Values of $s \le 1$ correspond to an enhanced probability (relative to a Poisson process with the same value of λ) of another event immediately following one that has just occurred. Values of $s \ge 1$ indicate a diminished

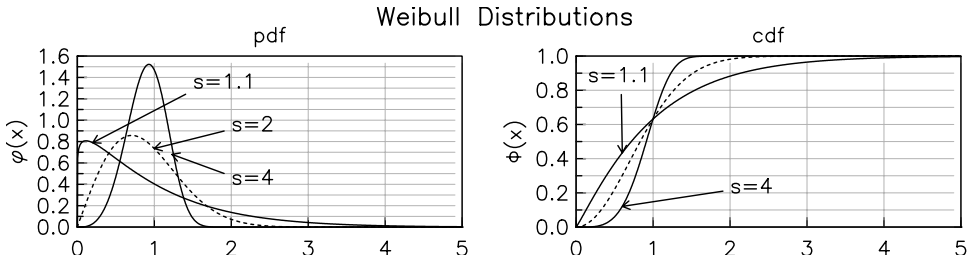

Figure 3.5: Weibull distributions with different shape parameters.

probability of another event immediately following any given one. For geomagnetic reversals, using the renewal model gives $s \gg 1$. The physical interpretation of this is controversial: it may be because the geological record does not record short polarity intervals adequately, or it may reflect something fundamental about the geodynamo [*McFadden*, 1984; *McFadden and Merrill*, 1984]. Gamma distributions are also used in statistical seismology, since the existence of aftershocks shows that this is a process with memory, and an enhanced probability of another earthquake immediately following any given one. Even after removing obvious aftershocks, earthquakes often appear to cluster in time, for which one probability model is a renewal process with interevent times following a gamma distribution with $s < 1$.

3.4.4 Weibull Distribution

This distribution (Figure 3.5) was invented to describe failure rates, and so is another choice for modeling renewal processes. The pdf is

$$\phi(x) = x^{s-1}e^{-x^s} = \left(\frac{s}{c}\right)\left(\frac{x-l}{c}\right)^{s-1} e^{-((x-l)/c)^s} \quad \text{for} \quad x \geq l$$

which makes the cdf relatively simple:

$$\Phi(x) = 1 - e^{-(x)^s} = 1 - e^{-((x-l)/c)^s}$$

The shape and scale parameters are sufficient to provide a flexible distribution, so a nonzero location parameter is less often used. The exponential distribution is a special case for $s = 1$.

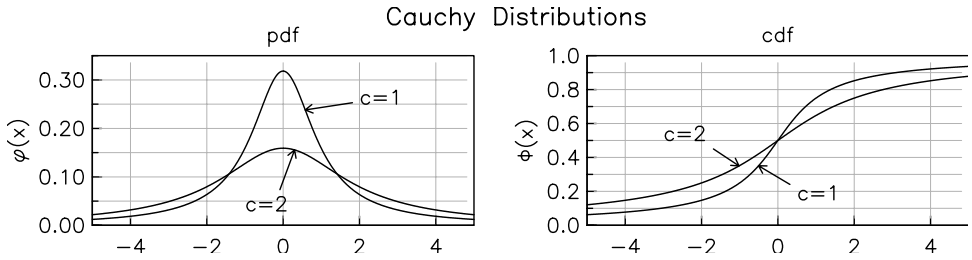

Figure 3.6: Cauchy distributions of different width.

3.5 Distributions Derived from the Normal

We now consider a collection of pdf's which have in common that they apply to some combination of one or more Normal rv's. Some of these pdf's, are not much used to represent data; rather, they are part of various statistical tests which we will discuss in Chapter 6.

3.5.1 Cauchy Distribution

Our first example is, however, meant to represent data – or perhaps we should say, meant to represent data that we hope we will never see, because such an rv has such a wide variation that we could not say much useful. In Section 2.11.1 we found the pdf for the sum of two Normal random variables, namely another Normal. Suppose instead that we consider the ratio of two Normal variables, making our new random variable $Y = X_1/X_2$, with $X_i \sim N(0,\sigma)$. It is easy to see that it would be relatively common for the denominator X_2 to be small, and hence for Y to be large; so we would expect the pdf of Y to be much more heavy-tailed than the pdf of X.

We can use equation (2.27) to get the actual distribution, namely

$$\phi(x) = \frac{1}{2\pi}\int_{-inf}^{\infty} |x| e^{-y^2/2} e^{-x^2y^2/2}\, dx = \frac{1}{\pi}\int_0^{\infty} x e^{-y^2(x^2+1)/2}\, dx$$

A change of variables to $u = y^2$ makes this into the integral of an exponential in u:

$$\frac{1}{2\pi}\int_0^{\infty} e^{-u(x^2+1)/2}\, du = \frac{1}{\pi(1+x^2)} = \frac{c}{\pi(c^2+(x-l)^2)}$$

which is the pdf for the **Cauchy distribution**, shown in Figure 3.6. This distribution is integrable (it has to be, to be a pdf), but the first and higher

moments do not exist (that is to say, are infinite): rv's with a Cauchy distribution do not have a mean or variance – though they do have a median and an interquartile range. This pdf has the heaviest tails possible for a pdf; it shows that shows that it may not always be possible to follow even such a standard procedure as finding a mean.

3.5.2 Chi-Squared Distribution

If, instead of taking the ratio of Normal rv's, we take the product, we get the χ^2 distribution, one of several that are mostly used for statistical tests. We start by squaring a random variable with a normal distribution; that is, if X is a random variable distributed with a normal pdf with mean 0 and standard deviation 1 (i.e., $X \sim N(0,1)$, then the distribution of the random variable $Z = X^2$ is conventionally written as χ^2_1, and is conventionally called the **chi-square distribution with one degree of freedom**, [3].

The pdf can be derived by the procedure described in Section 2.11:

$$\Phi_z(z) = \Pr(Z \le z) = \Pr(-\sqrt{z} \le X \le \sqrt{z})$$

where the second expression comes from using the inverse function to x^2, namely the square root. Rewriting the rightmost expression gives

$$\Phi_z(z) = \Phi_x(\sqrt{z}) - \Phi_x(-\sqrt{z})$$

and differentiating with respect to $z = x^2$ using the chain rule gives

$$\phi_z(z) = \frac{d\Phi(z)}{dz} = \frac{1}{\sqrt{z}}\left(\frac{d\Phi_x(\sqrt{z})}{dz} - \frac{d\Phi_x(-\sqrt{z})}{dz}\right) = \frac{1}{2\sqrt{z}}\left(\phi_x(\sqrt{z}) + \phi_x(-\sqrt{z})\right)$$

And finally, since the pdf for ϕ_x is Normal, we get

$$\phi_z(z) = \frac{1}{\sqrt{2\pi}}\frac{e^{-z/2}}{\sqrt{z}}$$

which is a special case of the gamma distribution, with $\lambda = s = \frac{1}{2}$.

Next, consider n random variables $Z_1, Z_2 \ldots Z_n$, which are independent and each distributed as χ^2 with one degree of freedom. The distribution of $Y = Z_1 + Z_2 + \cdots + Z_n$ is called the **chi-square distribution with n degrees**

[3] However unevocative you find the "degree of freedom" terminology, it is standard, so you had best get used to it.

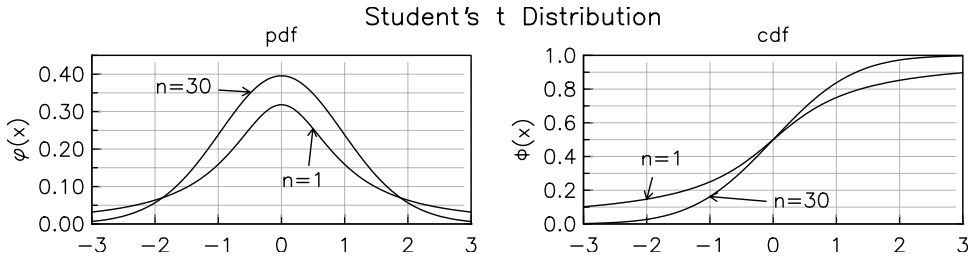

Figure 3.7: Student's t distribution for different values of n.

of freedom, denoted χ^2_n (that is, $Y \sim \chi^2_n$). Each of the Z_i's has a gamma distribution; the sum of n iid rv's with a gamma distribution (all with the same λ) is also gamma distributed, so the χ^2 distribution with n degrees of freedom is a gamma distribution with $s = n/2$ and $\lambda = ½$:

$$\phi(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{(n/2)-1} e^{-x/2}$$

The expected value of the χ^2_n distribution is n, and the variance is $2n$. We will see later that a useful quantity is the **reduced chi-square**, χ^2_n/n, which has an expected value of one, independent of n.

The χ^2 distribution finds widespread application in model fitting. If we have n observations o_i, with predicted values c_i and measurement errors σ_i, then we can form the **standardized residuals**

$$r_i^2 = \frac{(o_i - c_i)^2}{\sigma_i^2}$$

where the "standardized" part, as in the discussion in Section 3.3.1, comes from scaling the residuals $o - c$ by their errors. Then, if the residuals are distributed according to a Normal pdf, the **sum of squared residuals** (abbreviated **ssr**), $\sum_i r_i^2$, has a χ^2_n distribution; the **reduced ssr**, which is the ssr divided by n, would then be distributed as the reduced chi-square. So we would hope that the reduced ssr would be close to one – and closer as n becomes larger, as we discuss in more detail in Section 6.5.2.

3.5.3 Student's t Distribution

If X is normally distributed ($X \sim N(0,1)$), and $Z \sim \chi^2_n$ with Z and X independent, then the distribution of $X/\sqrt{Z/n}$ is the **Student's t distribution**

with n degrees of freedom.[4] This has the pdf (Figure 3.7):

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

so that, as with χ^2, the shape factor is an integer. The t distribution is symmetric about zero. As the number of degrees of freedom, n, approaches infinity, the t distribution tends to the Normal distribution. The t distribution is used in testing whether samples have statistically distinguishable means – we will, again, discuss this fully when we turn to hypothesis testing in Chapter 6, specifically in Sections 6.4.2 and 6.4.3.

3.5.4 F Distribution

Next, suppose X is a random variable distributed as χ^2 with m degrees of freedom; then X/m is a similarly-distributed rv which has been standardized to make its expected value unity. Take Z to be another rv, independent of X, and distributed as χ^2 with n degrees of freedom. Now consider the random variable Y that is the ratio of X and Z when both have been normalized by their degrees of freedom:

$$Y = \frac{X/m}{Z/n}$$

This variable will be distributed according to what is called the F distribution with m and n degrees of freedom, denoted $F_{m,n}$. The probability density function of this is given by

$$\phi(x) = \frac{\Gamma[(m+n)/2]}{\Gamma(m/2)\Gamma(n/2)}\left(\frac{m}{n}\right)^{m/2}\cdot x^{m/2-1}\left(1+\frac{m}{n}x\right)^{-(m+n)/2}$$

for x over $[0,\infty)$. We do not plot this because of the complications of having two variables. Like the t and χ^2 distributions, the F distribution is used, not as a pdf for modeling data, but rather in testing whether two sets of data (or, more often, residuals) have meaningfully different variances. This use comes from the rv Y being a ratio of reduced χ^2's; if the fit of two models is equally good, we would expect the reduced χ^2 for the residuals to be about the same, and Y to be near one; if Y is not it may indicate than one model fits better than the other.

[4] The name "Student" is the pseudonym that F. S. Gosset used for his statistical publications, to avoid problems with his employer, the Guinness brewery.

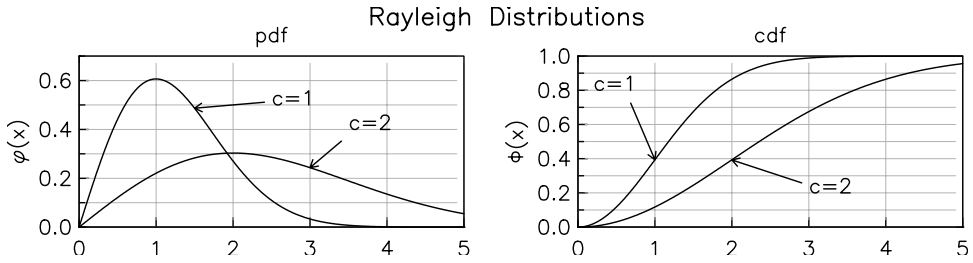

Figure 3.8: Rayleigh distributions of different widths.

3.5.5 Rayleigh Distribution

Another distribution is for a random variable that is the square of the sum of squares of two Normal rv's. An example would be the pdf of the amplitude of a two-dimensional vector, each of whose components was normally distributed – at the origin this amplitude is zero. We already derived this pdf in Section 3.3.1; another derivation comes from realizing that the rv for the square of the amplitude (call this Z) is distributed as χ^2_2, with cdf $\Phi_0(x) = 1 - e^{-z/2}$. Then, we can find the cdf of $X = Z^{1/2}$ by the same kind of procedure as we used in Section 3.5.2; taking $\Phi(x)$ to be the cdf of X, we have

$$\Phi(x) = \Pr(X \le x) = \Pr(Z \le x^2) = \Phi_0(x^2) = 1 - e^{-x^2}$$

so that the pdf is

$$\phi(x) = 2xe^{-x^2} = \left(\frac{x}{c^2}\right) e^{-x^2/2c^2}$$

This is the pdf of the **Rayleigh distribution**, which is used in the theory of random walks. Note that if we add two orthogonal vectors described by normal distributions, the expected value of the result is the origin; but the pdf of the amplitude is not zero: the most probable amplitude does not correspond to the most probable location.

3.6 "Normal" Distributions on Other Intervals

We now turn to another class of distributions related to the Normal, though only by analogy: these are pdf's for rv's on other intervals than the entire real line.

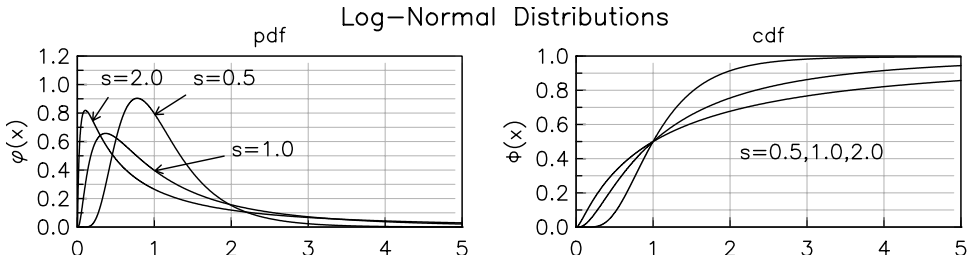

Figure 3.9: Log-normal distributions with different shape factors.

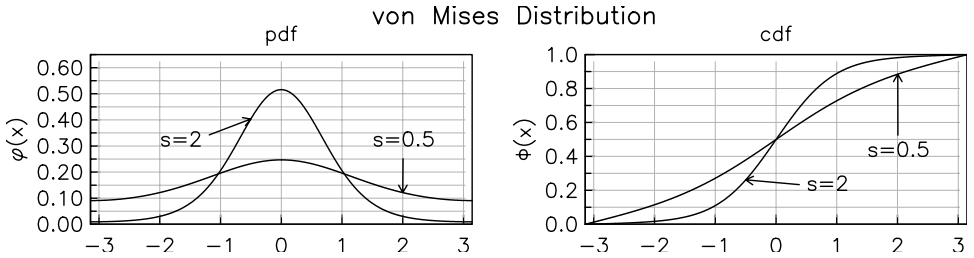

Figure 3.10: Distribution functions for the von Mises distribution, for different widths.

3.6.1 Log-Normal Distribution

Consider an rv given by $Y = e^X$ where X is Normal; Y is said to be **log-normally** distributed, since $X = \ln Y$. The exponent maps the real line (for X) into the positive numbers (for Y), which is appropriate for variables that are intrinsically positive. The pdf (Figure 3.9) is

$$\phi(x) = \frac{1}{s\sqrt{2\pi}} \frac{e^{-(\ln(x))^2/2s^2}}{x} = \frac{1}{s\sqrt{2\pi}} \frac{e^{-(\ln((x-l)/c))^2/2s^2}}{x-l} \qquad \text{for} \quad x \geq l$$

By the Central Limit Theorem, the Normal is appropriate if we have an rv that is the sum of many others; but if we take exponentials, the sum becomes a product. The log-normal is thus a possibility when we think the process can be modeled as a series of multiplications (or divisions) of positive rv's. For this reason this pdf is sometimes used to model the distribution of sediment grain sizes: the grains are produced by repeated splitting of bigger grains. Similarly, volumes of equally-magnetized materials in rocks might arise from repeated divisions, so the log-normal is used for the distribution of the magnetization of basaltic rocks.

3.6.2 von Mises Distribution

The next distribution, though rarely discussed in elementary statistics, is important for **circular data**: that is, directions in two dimensions, for which the variable range is range $[-\pi, \pi)$, with the value of the density function being the same at each end. We can scale the range to make this distribution apply to any variable which is modulo some number: for example, the fractional part of a real-valued quantity, which would fall in $[0, 1)$. The earthquake times of day in Section 1.2 are in this category.

For such rv's the von Mises distribution (Figure 3.10) is analogous to the Normal. Its pdf is

$$\phi(x) = \frac{1}{2\pi I_0(s)} e^{s\cos(x-l)}$$

where I_0 is a modified Bessel function. Note that in this case the shape factor modifies the width of the function, as the scale factor does for the Normal. As $s \to 0$, ϕ approaches a uniform distribution.

3.6.3 Fisher Distribution

The next distribution is, strictly speaking, for two dimensions, not one: but the pdf has only one argument. This **Fisher distribution** is again analogous to the Normal, but this time for the domain of the random variable being the surface of a sphere. So it is used for modeling directions on a sphere, for example namely the distribution of poles found when directions of rock magnetiation are interpreted as coming from dipole fields with different orientations; this is actually what the distribution was invented for, The pdf is

$$\phi(\Delta, \theta) = \frac{s}{4\pi \sinh s} e^{s\cos\Delta} = \frac{\kappa}{4\pi \sinh \kappa} e^{\kappa \cos\Delta}$$

This expression omits a location parameter; or rather, makes it implicit through the coordinates used for Δ, which is the angular distance from the maximum of the pdf. The distribution is circularly symmetric about its maximum, hence there is no dependence on the azimuthal variable θ. The location of the maximum value, with $\Delta = 0$, is also the expected value of the rv. Both the Fisher and von Mises distributions have only a shape parameter, which determines the width; for the Fisher this is called the **concentration parameter**. If $\kappa = 0$ the distribution is uniform over the sphere; as $\kappa \to \infty$ becomes large, the pdf (for Δ small) approximates a Normal rotated

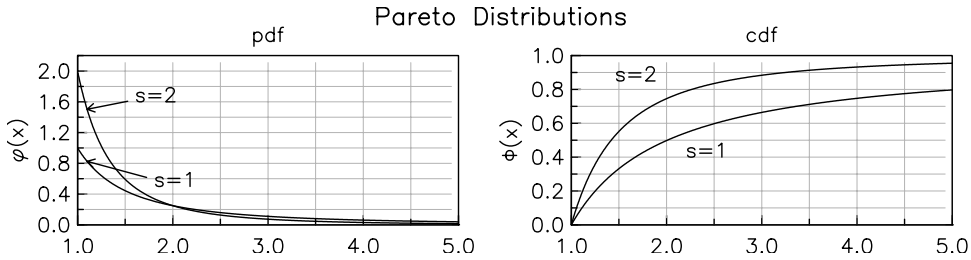

Figure 3.11: Pareto distributions for different shape factors.

about the maximum. This version of the pdf gives the probability density over a unit angular area.[5]

A related distribution is the probability density per unit of angular distance from the pole, averaged over all values of θ; this pdf is

$$\phi(\Delta) = \frac{\kappa}{2\sinh\kappa} e^{\kappa\cos\Delta}\sin\Delta$$

This is analogous to the Rayleigh distribution (Section 3.5.5), and likewise goes to zero at the origin. Neither of these distributions incorporates possible variation in azimuth, as might be caused by unequal variances; representing such variation is possible, but calls for more complicated pdfs.

3.7 Miscellaneous Distributions

We now discuss some distributions that are useful in geophysics but do not seem to fall into any of the categories above.

3.7.1 Pareto Distribution

The Pareto is a curiosity, since it is both applicable to many geophysical phenomena but few geophysicists have heard of it. Pareto-distributed variables range over half of the real line, though the range is $[l,\infty)$ not $[0,\infty)$. The most notable feature of the Pareto is that the tail is much heavier than it is for the Normal or the Exponential, so there is a much higher probability of very large values. The pdf (Figure 3.11) is

$$\phi(x) = sl^s x^{-(s+1)}$$

[5] Angular areas on the sphere are measured in steradians, with the area of the unit sphere being 4π.

where $s \geq 0$; the location parameter determines the lower end of the range. Another term for this pdf is "power-law distribution", for the obvious reason that the value, for large x, decreases as a power of x.

To see how this leads to a famous geophysical result, we first find the cdf and its complement

$$\Phi(x) = 1 - l^s x^{-s} \quad \text{and} \quad 1 - \Phi(x) = (l/x)^s$$

where the complement gives $\Pr(X > x)$, this being the way in which the result was originally expressed by Pareto. Now suppose a dataset has n occurrences of something, with the values having a Pareto distribution. Then let $n(x)$ be the number of occurrences greater than x; this number is an rv with the pdf $n[1 - \Phi(x)] = nl^s x^{-s}$. Taking the logarithm of this gives

$$\ln[n(x)] = \ln(nls) - s\ln(x) \tag{3.9}$$

If we remember that earthquake magnitude M is suppsed to be a logarithmic function of earthquake size (ignoring for now what "size" refers to), we see that equation 3.9 has exactly the form of the Gutenberg-Richter relationship for the number of earthquakes with magnitude greater than magnitude M:

$$log_{10}(n) = a - bM$$

Thus the "b-value" of a set of earthquakes, something much discussed in seismicity studies, is related to the shape factor for the Pareto distribution of earthquake size.

Note that we can show the relationship (3.9) without binning the data, If we sort the data values to produce **rank-ordered** data

$$x_{(1)} \geq x_{(2)} \geq x_{(3)} \geq \quad \ldots \quad x_{(n-1)} \geq x_{(n)}$$

then $x_{(k)}$ is a value such that k values are greater than or equal to this, which means that $n(x_{(k)}) = k$. If the data have a Pareto distribution plotting $\ln k$ against $\ln(x_{(k)})$ will give points that fall on a straight line. Showing k against x (sorted) on a log-log plot, is sometimes called a **rank-frequency** plot because it was first used in cases where x was the relative frequency of occurrence of words in a body of text.[6] You should be aware, though, that such a plot can mislead you into thinking that data are Pareto-distributed even when they are not [*Clauset et al.*, 2009].

[6] This plotting method was pioneered by G. K. Zipf, and the Pareto distribution is sometimes called "Zipf's law".

3.8 A General Method for Producing Random Deviates

We now describe a general method for constructing random variates for any pdf. If the inverse of the cumulative distribution is easy to compute, we only need to apply it to uniformly distributed rv's; but what if this computation is difficult? Then we can use a general, though less efficient, procedure called the **rejection method**.

Suppose we want to produce random variates with a pdf $\phi_1(x)$. First, find a function f such that $f(x) \geq \phi_1(x)$ wherever $\phi_1 > 0$, with $f(x) \geq 0$ elsewhere. Since $f(x)$ is everywhere zero or positive, normalizing it creates a pdf $\phi_2(x)$:

$$\phi_2(x) = \frac{f(x)}{\int f(x)\,dx} \tag{3.10}$$

where the limits of the integral are whatever is appropriate for ϕ_1. We also choose f so its inverse uses only standard functions, so we can easily generate random variates with pdf $\phi_2(x)$.

Having found $f(x)$, we generate two random variables: F, distributed according to ϕ_2 and U, uniformly distributed between 0 and 1. If

$$U \leq \frac{\phi_1(F)}{f(F)} \tag{3.11}$$

we accept F as one of our X's; if not, we find another F and U, and repeat.

The probability of getting X within a specified range is

$$\Pr[x \leq X \leq x + \delta x] = \Pr[x \leq F \leq x + \delta x| \text{ accept}]$$

which by Bayes' theorem is

$$\frac{\Pr[\text{accept}|x \leq F \leq x + \delta x]\Pr(x \leq F \leq x + \delta x)}{\Pr[\text{ accept}]} \tag{3.12}$$

Now, given the way we defined acceptance in equation (3.11), we have

$$\Pr[\text{ accept}|x \leq F \leq x + \delta x] = \frac{\phi_1(F)}{f(F)}$$

while the total probability of getting an acceptance is this integrated over all F; since

$$\int \phi_1(x)\,dx = 1$$

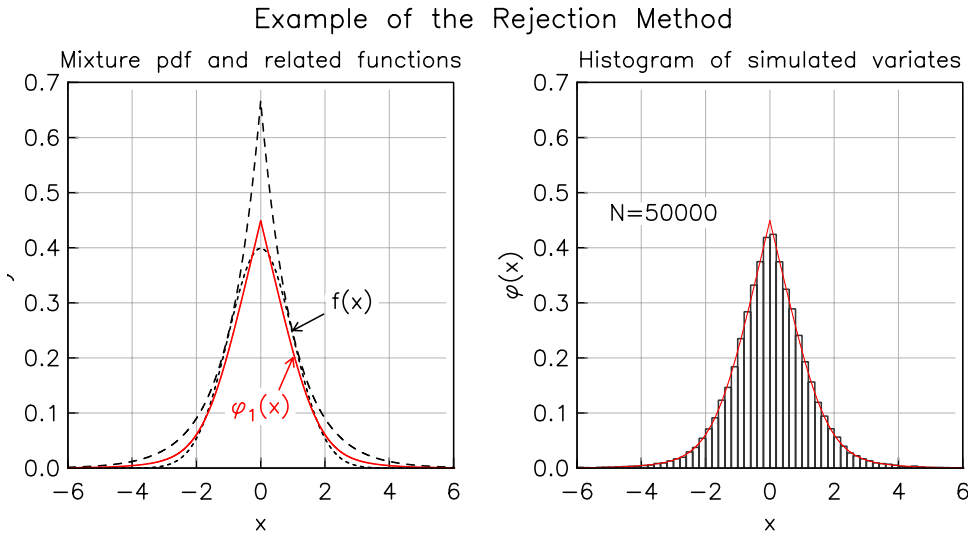

Figure 3.12: In the left panel, the dotted line shows a Gaussian pdf; the solid red line is the mixture pdf we wish to generate variates for, and the dashed line is the function used to produce variates for rejection. The right panel compares the target pdf (again in red) with a histogram of 50,000 variates produced using the rejection method.

this is

$$\frac{1}{\int f(x)dx}$$

Putting all these expressions into equation (3.12), we find

$$\Pr[x \le X \le x+\delta x] = \frac{\phi_1(X)}{f(X)}\phi_2(X)\int f(x)dx = \phi_1(X)$$

by the definition (3.10). The closer we can get $f(x)$ to $\phi_1(x)$, the smaller the fraction of trial values that will be rejected.

For example, consider generating random variables for the pdf shown in Figure 3.12, which is

$$\phi_1(x) = .5N(0,1) + .5\left(\frac{e^{-|x|}}{2}\right)$$

This is a combination of two pdf's (what is known as a **mixture model**), and is designed to be roughly Gaussian near the center, while having heavier tails. A suitable function for $f(x)$ is $1.33 \times e^{-|x|}$; it is easy to gener-

ate variates appropriately distributed for this pdf,[7] and it is close to $\phi_1(x)$, which minimizes the fraction of variates that are not accepted. Figure 3.12 also shows the result for 50000 generated variates, each of which took an average of 3×1.34 calls to `ran` to produce. The pdf matches the histogram very well. Thinking in terms of histograms may help you to see how the method works: we can think of starting with data whose histogram is that of a double exponential, and then trimming each bin by the ratio of ϕ_1/f.

[7] Take $-\ln(1-U)$ where U is a uniform variate to get an exponential distribution, and to get a two-sided exponential distribution take a second uniform variate, and change the sign of the exponential variate if it is less than $\frac{1}{2}$.