Chapter 2. Measuring Probability Distributions

The full specification of a probability distribution can sometimes be accomplished quite compactly. If the distribution is one member of a parametric family, the full description requires only the specification of the values of the parameters. For example, if we state that a random variable X has a Binomial $(5, .3)$ distribution, we are saying its probability distribution is given by $P(X = x) = \binom{5}{x}(.3)^x(.7)^{5-x}$ for $x = 0, 1, 2, 3, 4, 5$. Or if we state X has an Exponential (2.3) distribution, we mean its probability density function is $f_X(x) = 2.3e^{-2.3x}$, for $x \geq 0$. If we remain within a single parametric family, it is usually convenient to continue to use some version of the parametrization of that family to describe that distribution, and to compare different members of the family by comparing the values of the parameters. But for more general purposes, comparing members of different families or even comparing discrete distributions with continuous distributions, we will require summary measures that make sense beyond a single parametric family, even though they themselves do not fully specify the distribution. The most basic such summary measure is *expectation*.

2.1 Expectation

The expectation of a random variable is a weighted average of its possible values, weighted by its probability distribution. Mathematically, we define the *expectation of X,* denoted $E(X)$, as follows:

For the discrete case:

$$E(X) = \sum_{\text{all x}} x p_X(x). \tag{2.1}$$

For the continuous case:

$$E(X) = \int_{-\infty}^{\infty} x f_X(x) dx \tag{2.2}$$

We shall use two alternative and completely equivalent terms for the expectation $E(X)$, referring to it as *expected value of* X or the *mean of* X. The notation $E(X)$ is potentially misleading; it suggests that we are talking about a function or transformation of the random variable, which would of course itself be a random variable. In fact, $E(X)$ is a number and is properly described as a function of the probability distribution of X, rather than a function of X.

The expectation of X summarizes the probability distribution of X by describing its center. It is an average of the possible values, with attention paid to how likely they are to occur. If we adopt the interpretation of the probability distribution of X as being a distribution of a unit mass along the real line, then $E(X)$ is the center of gravity.

Example 2.A. If X has the discrete distribution

x	1	2	3
$p_X(x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$

then

$$E(X) = 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{4} = \frac{7}{4} = 1.75$$

[Figure 2.1]

Example 2.B. If X is a continuous random variable with probability density function

$$\begin{aligned} f_X(x) &= 12x^2(1-x) \quad \text{for } 0 \leq x \leq 1 \\ &= 0 \quad \text{otherwise ,} \end{aligned}$$

then

$$\begin{aligned} E(X) &= \int_0^1 x \cdot 12x^2(1-x)dx \\ &= 12 \cdot \int_0^1 (x^3 - x^4)dx \\ &= 12\left[\frac{x^4}{4} - \frac{x^5}{5}\right]_0^1 = \frac{12}{20} = .6 \end{aligned}$$

For any symmetric probability distribution, the expectation is at the point of symmetry.

Example 2.C. If X has a Binomial $\left(8, \frac{1}{2}\right)$ distribution (Figure 1.5(a)), then $E(X) = 4$.

Example 2.D. If X has a Uniform $(0, 1)$ distribution (Figure 2.3), then $E(X) = \frac{1}{2}$.

[Figure 2.2]

We shall see that the expectation enjoys statistical properties that make it uniquely valuable, but it is far from being the only possible way of describing the center of a probability distribution. For example, we could use the *mode,* which is the most probable value in the discrete case and the value with the highest density in the continuous case. The modes in Examples 2.A and 2.B are 1 and 2/3, respectively, the latter found by solving the equation

$$\frac{d}{dx} f_X(x) = 0 \tag{2.3}$$

for x. (Since the logarithm is a monotonic function, you would find exactly the same solution by solving the equation

$$\frac{d}{dx} \quad \log_e f_X(x) = 0$$

for x. Sometimes the calculation is easier in this form.) One of the drawbacks of the mode is that it may not be uniquely defined (consider the Uniform distribution, Figure 2.3); another is that it may be quite sensitive to even minor changes in the distribution in ways that belie a general statistical usefulness. For example, minor changes to the Uniform density could produce a mode at any point in its range. Another possible measure of the center of a probability distribution is the *median,* which may be described as that value ν for which $P(X \leq \nu) = P(X \geq \nu) = 1/2$, when such a value exists. (When no such value exists, we may take any ν satisfying

$$P(X \leq \nu) \geq \frac{1}{2}$$

and

$$P(X \geq \nu) \geq \frac{1}{2}$$

as *a median.*) For Example 2.B the median can be found by solving the equation

$$F_X(x) = \frac{1}{2}. \tag{2.4}$$

Since for that density, for $0 \leq x \leq= 1$,

$$\begin{aligned} F_X(x) &= \int_0^x 12u^2(1-u)du \\ &= x^3(4 - 3x) \quad \text{for } 0 \leq x \leq 1, \end{aligned}$$

We can solve (2.4) numerically to find $\nu = .614$. For Example 2.A, any value of ν in the interval $1 \leq \nu \leq 2$ is a median.

Example 2.E. The Beta (α, β) distribution. The probability density function $f_X(x)$ of Example 2.B is a member of a parametric family of distributions that will be particularly useful to us when we discuss inference. But in order to introduce that family we must first define a generalization of n factorial that will appear later in several contexts, the *Gamma function.* The Gamma function is a definite integral that was studied by the mathematician Leonhard Euler in the 1700's. The *Gamma function* is given for $a > 0$ by

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx. \tag{2.5}$$

If a is an integer, this integral can be evaluated by repeated integration-by-parts to give

$$\Gamma(n) = (n-1)! = 1 \cdot 2 \cdot 3 \ldots (n-1), \tag{2.6}$$

and hence it can be viewed as a generalization of the factorial function, as a way of defining $(a-1)!$ for non-integer values of a. In fact, the Gamma function can be shown (again by integration-by-parts) to have the factorial property

$$\Gamma(a) = (a-1)\Gamma(a-1) \tag{2.7}$$

for all $a > 1$. Euler showed that

$$\Gamma(.5) = \sqrt{\pi}; \tag{2.8}$$

this fact, together with (2.7) and (2.6), permits the iterative evaluation of $\Gamma(a)$ for $a = 1, 1.5, 2, 2.5, \ldots$. Tables of $\Gamma(a)$ for $0 < a < 1$ and (2.7) permit the evaluation of $\Gamma(a)$ more generally. Also, Stirling's formula (Section 1.4) applies as well to $\Gamma(a)$ as to $n!$; for example,

$$\Gamma(a+1) \sim \sqrt{2\pi}\, \frac{a^{a+1}}{2} e^{-a}. \tag{2.9}$$

The family of *Beta* (α, β) *distributions* is given by the probability density functions

$$\begin{aligned} f_X(x) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot x^{\alpha-1}(1-x)^{\beta-1} \quad \text{for } 0 \le x \le 1 \\ &= 0 \quad \text{otherwise,} \end{aligned} \tag{2.10}$$

where the parameters α, β are positive numbers, $\alpha > 0, \beta > 0$. For example, if $\alpha = 3$ and $\beta = 2$, we have

$$\Gamma(\alpha) = \Gamma(3) = 2! = 2$$

$$\Gamma(\beta) = \Gamma(2) = 1! = 1$$

$$\Gamma(\alpha+\beta) = \Gamma(5) = 4! = 24,$$

and

$$f_X(x) = 12x^2(1-x) \quad \text{for } 0 \le x \le 1,$$

the density of Example 2.B. For another example, if $\alpha = \beta = 1$, we have $\Gamma(\alpha) = \Gamma(\beta) = 0! = 1$, $\Gamma(\alpha+\beta) = 1! = 1$, and

$$f_X(x) = 1 \quad \text{for } 0 \le x \le 1,$$

the Uniform $(0, 1)$ distribution.

[Figure 2.3]

Figure 2.3 illustrates these and other cases. Note that all Beta distributions (other than the case $\alpha = \beta = 1$) have a single mode; by solving the equation (2.3) it is easy to show that if $\alpha > 1$ and $\beta > 1$, the mode is at $x = \frac{\alpha-1}{\alpha+\beta-2}$. If $\alpha = \beta$, the density is symmetric about $1/2$; the larger α and β are, the more concentrated the distribution is around its center.

If α and β are integers, then the Gamma functions in (2.10) can be written as factorials, and we have

$$\begin{aligned}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} &= \frac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!} \\ &= \frac{(\alpha+\beta-2)!(\alpha+\beta-1)}{(\alpha-1)!(\beta-1)!} \\ &= \binom{\alpha+\beta-2}{\alpha-1}(\alpha+\beta-1),\end{aligned}$$

so (2.10) can in this case be written

$$f_X(x) = (\alpha+\beta-1)\binom{\alpha+\beta-2}{\alpha-1}x^{\alpha-1}(1-x)^{\beta-1}, \quad \text{for } 0 \le x \le 1. \tag{2.11}$$

The Beta distributions derive their name from the fact that in classical analysis the definite integral

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx \tag{2.12}$$

viewed as a function of two arguments, has been named the *Beta function,* and it has been shown to be related to the Gamma function by

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}. \tag{2.13}$$

Thus $B(\alpha,\beta)$ is just the reciprocal of the coefficient in (2.10), and the relations (2.12) and (2.13) can be seen as reflecting the fact that

$$\int_0^1 f_X(x)dx = 1,$$

for the Beta density (2.10), as would be true for any distribution whose possible values are $0 \le x \le 1$.

We have noted that the mode of a Beta density has a simple expression in terms of α and β as long as $\alpha > 1$ and $\beta > 1$, namely $(\alpha-1)/(\alpha+\beta-2)$. The expectation has an even simpler expression,

$$E(X) = \frac{\alpha}{\alpha+\beta}, \tag{2.14}$$

and one way to show this illustrates a technique based upon the elementary fact that $\int_{-\infty}^{\infty} f_X(x)dx = 1$, a technique that can frequently be useful in evaluating integrals that arise in statistics. By definition,

$$\begin{aligned}E(X) &= \int_0^1 x \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}dx \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 x^{\alpha}(1-x)^{\beta-1}dx,\end{aligned}$$

The technique involves manipulating the integrand by rescaling it so that it becomes a probability density, and of course multiplying the entire expression by the reciprocal of the scaling factor used, to preserve its value. In the present case, the integrand can be easily rescaled to become a Beta $(\alpha+1,\beta)$ density:

$$E(X) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} \cdot \int_0^1 \frac{\Gamma(\alpha+\beta+1)}{\Gamma(\alpha+1)\Gamma(\beta)}x^{\alpha}(1-x)^{\beta-1}dx$$

The integral here equals 1, since it is the integral of the Beta density (2.10) with $\alpha+1$ substituted for α. This gives

$$\begin{aligned}E(X) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+1)} \cdot \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} \\ &= \frac{\alpha}{\alpha+\beta},\end{aligned}$$

using the factorial property (2.7): $\Gamma(\alpha+1)=\alpha\Gamma(\alpha)$, and $\Gamma(\alpha+\beta+1)=(\alpha+\beta)\Gamma(\alpha+\beta)$.

For example, if $\alpha=3$ and $\beta=2$ we find $E(X)=3/(3+2)=.6$, the same result we found in Example 2.B. And if $\alpha=\beta=3$ we have $E(X)=\frac{1}{2}$, which is also obvious from the symmetry of the density in that case. Indeed, $E(X)=\frac{1}{2}$ for any Beta(α,α) density.

The expectation is defined for any random variable for which the sum (2.1) or integral (2.2) exists and is well-defined. If the range of X is finite, then $E(X)$ will necessarily exist, but if X has an infinite range it may not.

Example 2.F. Suppose X has probability density

$$\begin{aligned} f_X(x) &= \frac{1}{x^2} \quad \text{for } x \geq 1 \\ &= 0 \quad \text{otherwise} \end{aligned}$$

[Figure 2.4]

It can be easily verified that this is a density: $\int_1^\infty x^{-2}dx=1$. Then formally,

$$E(X)=\int_1^\infty x\cdot\frac{1}{x^2}dx=\int_1^\infty \frac{1}{x}dx=\log_e(x)|_1^\infty.$$

a divergent integral. We could (and will) stretch a point and define $E(X)=\infty$ for such a case, but even that device is not available for the next example.

Example 2.G. Suppose X has probability density

$$\begin{aligned} f_X(x) &= \frac{1}{2x^2} \quad \text{for } |x| \geq 1 \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Then, formally,

$$E(X)=\int_1^\infty \frac{1}{2x}dx+\int_{-\infty}^{-1}\frac{1}{2x}dx,$$

which is the difference of two divergent integrals, and is not well-defined even in the extended sense of Example 2.F. We shall encounter other such examples later and see that they do arise naturally in statistically meaningful situations; they are not merely mathematical curiosities. If we wish to reconcile such phenomena with our intuition, we could think of the density of Example 2.F as spreading its unit mass so widely to the extreme right that wherever we attempt to balance the mass, it tilts down to the right. In the case of Example 2.G, the mass is symmetrically distributed about zero (so we might think it should balance at zero), but it is so extremely spread that when we attempt to balance it at zero, the axis breaks; a balance is not attainable!

2.2 Expectations of Transformations

Barring cases such as those in Example 2.G where the expectation is not well-defined, we can calculate the expectation of any discrete or continuous random variable from the definitions (2.1) and (2.2). This applies in particular to a random variable that is a transformation of another random variable: if $Y=h(X)$, we can first use the techniques of Section 1.8 to find the probability distribution of Y, and then apply the appropriate definition to find $E(Y)$. In the continuous case with a monotone transformation $h(x)$ with inverse $g(y)$, this would involve finding (using (1.29))

$$f_Y(y)=f_X(g(y))|g'(y)|$$

and then calculating (using definition (2.2))

$$E(Y)=\int_{-\infty}^\infty yf_Y(y)dy.$$

Or, combining the two steps into one,

$$E(Y) = \int_{-\infty}^{\infty} y f_X(g(y))|g'(y)|dy. \tag{2.15}$$

For nonmonotone transformations such as $h(X) = X^2$, the procedure would be even more cumbersome. Fortunately, there is a route that is usually simpler, that accomplishes the work in one step and works for nonmonotone transformations as well. That is to calculate:

(2.16) *Discrete Case:* For any function h,

$$E(h(X)) = \sum_{\text{all } x} h(x) p_X(x);$$

(2.17) *Continuous Case:* For any function h,

$$E(h(X)) = \int_{-\infty}^{\infty} h(x) f_X(x) dx,$$

(Of course, we assume implicitly that the sum or integral is well-defined, for example that it is not divergent as in Example 2.G.)

To see why these simpler formulae work (that is, give the same answer as the longer route), consider the continuous case with a strictly monotonic increasing transformation h, where $E(Y) = E(h(X))$ is given by (2.15) with $|g'(y)| = g'(y)$. Making the change of variable $x = g(y)$, $y = h(x)$, $dx = g'(y)dy$, we have

$$\begin{aligned} E(Y) &= \int_{-\infty}^{\infty} y f_X(g(y)) g'(y) dy \\ &= \int_{-\infty}^{\infty} h(x) f_X(x) dx. \end{aligned}$$

Which is to say the equality of (2.15) and (2.17) is no more than a consequence of a change-of-variable in an integral. The same idea works more generally, to show that (2.16) and (2.17) are valid for even quite irregular nonmonotonic transformations $Y = h(X)$.

Example 2.H. Consider the simple discrete distribution for X,

x	0	1	2
$p_X(x)$	.5	.3	.2

and the transformation $Y = 1/(X + 1)$. The long route would have us first find the probability function of Y. Since the inverse of $h(x) = 1/(x + 1)$ is $x = g(y) = \frac{1}{y} - 1$, this is $p_Y(y) = p_X(g(y))$, given by

y	$\frac{1}{3}$	$\frac{1}{2}$	1
$p_Y(y)$	.2	.3	.5

Then, using the definition (2.1),

$$E(Y) = \frac{1}{3}(.2) + \left(\frac{1}{2}\right)(.3) + 1(.5) = \frac{43}{60} = .72.$$

But working with (2.16) we can find directly that

$$\begin{aligned} E(Y) = E(h(X)) &= \sum_{\text{all } x} \frac{1}{(x+1)} p_X(x) \\ &= \left(\frac{1}{0+1}\right)(.5) + \left(\frac{1}{1+1}\right)(.3) + \left(\frac{1}{2+1}\right)(.2) \\ &= 1(.5) + \frac{1}{2}(.3) + \frac{1}{3}(.2) = \frac{43}{60} = .72. \end{aligned}$$

In any case, note that simply substituting $E(X)$ for X in $h(X)$ gives an incorrect answer: $E(X) = 1.75$, and $1/(1.75+1) = .36 \neq .72$. In general, $E(h(X)) \neq h(E(X))$, although we will encounter one class of cases (linear transformations) where this would work.

Example 2.I. Suppose X has the standard normal distribution, with density

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad \text{for } -\infty < x < \infty.$$

What is the expectation of $Y = X^2$? We have previously (Example 1.M) found the distribution of Y to be the Chi Square (1 d.f.) distribution, with density

$$\begin{aligned} f_Y(y) &= \frac{1}{\sqrt{2\pi y}} e^{\frac{-y}{2}} \quad \text{for } y > 0 \\ &= 0 \quad \text{for } y \leq 0. \end{aligned}$$

We could then find

$$\begin{aligned} E(Y) &= \int_0^\infty y \cdot \frac{1}{\sqrt{2\pi y}} e^{\frac{-y}{2}} dy \\ &= \frac{1}{\sqrt{2\pi}} \cdot \int_0^\infty \sqrt{y} e^{\frac{-y}{2}} dy. \end{aligned}$$

This integral may be evaluated either with the aid of a table of definite integrals, or by making the change of variable $z = y/2$, $dz = \frac{1}{2}dy$, and finding

$$\begin{aligned} \int_0^\infty \sqrt{y} e^{\frac{-y}{2}} dy &= 2\sqrt{2} \int_0^\infty z^{\frac{1}{2}} e^{-z} dz \\ &= 2\sqrt{2}\Gamma(1.5) \\ &= 2\sqrt{2} \cdot (.5)\Gamma(.5) \\ &= \sqrt{2} \cdot \sqrt{\pi} = \sqrt{2\pi}, \end{aligned}$$

where we have used (2.7) and (2.8). In any event

$$E(Y) = \frac{1}{\sqrt{2\pi}} \cdot \sqrt{2\pi} = 1. \tag{2.18}$$

Alternatively, the same result could be obtained by calculating

$$E(X^2) = \int_{-\infty}^\infty x^2 \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx,$$

which involves a similar change of variable or a table, which would have been much the simpler route had we not already had $f_Y(y)$ available. It should be emphasized that the answer will be the same whichever way it is found; we are free to use the easiest way we can.

Example 2.J. Linear change of scale. The simplest transformation is a linear transformation; it is also the most used. We have already encountered $Y = h(X) = aX + b$ in Example 1.N. The need for the expectation of this transformation is so common we single it out as a theorem.

Theorem: For any constants a and b,

$$E(aX + b) = aE(X) + b. \tag{2.19}$$

Proof (for the continuous case): Using (2.17) we have

$$\begin{aligned} E(aX + b) &= \int_{-\infty}^{\infty} (ax + b) f_X(x) dx \\ &= \int_{-\infty}^{\infty} (ax f_X(x) + b f_X(x)) dx \\ &= a \int_{-\infty}^{\infty} x f_X(x) dx + b \int_{-\infty}^{\infty} f_X(x) dx \\ &= aE(X) + b \cdot 1. \end{aligned}$$

Thus we used the facts that an integral of a sum is the sum of integrals, and that $\int_{-\infty}^{\infty} f_X(x)dx = 1$ for any density. The proof for the discrete case is similar.

Two particular cases are worth singling out: If $b = 0$, then

$$E(aX) = aE(X),$$

and if $a = 0$, then

$$E(b) = b. \tag{2.21}$$

This last equation is intuitively obvious; "b" represents the "random variable" that always equals the number b, so of course this must be its expected value. For example, $E(3.1416) = 3.1416$. Note that $E(X)$ is just a number too, so

$$E(E(X)) = E(X). \tag{2.22}$$

To help us remember that $E(X)$ is a number (and not a transformation of X), we will often adopt the notation

$$E(X) = \mu_X \tag{2.23}$$

or

$$E(X) = \mu,$$

when there is no confusion as to which random variable is meant. Then (2.22) becomes $E(\mu_X) = \mu_X$ or $E(\mu) = \mu$.

2.3 Variance

Expectation was introduced as a summary measure of the center of a probability distribution. Another statistically important aspect of a probability distribution is how spread out it is. For example, the two densities of Figure 2.5 agree as to their center, but one represents a more widely dispersed mass distribution than does the other. If the probability distributions represented uncertainty about a physical constant, the greater dispersion of A would represent greater uncertainty.

In some respects, the most natural measure of dispersion of the distribution of a random variable X would involve asking, how far can we expect X to be, on average, from the center of its distribution? What is the expected value of $X - E(X)$, or $X - \mu_X$, measured without regard to sign? That is, what is $E|X - \mu_X|$?

Example 2.K: Suppose X has a Standard Normal distribution. Then

$$\mu_X = E(X) = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx.$$

From the symmetry of the density about zero we expect $\mu_X = 0$, as long as the integral does not diverge. In fact, its convergence will follow from our next computation. Taking $\mu_X = 0$, we have

$$E|X - \mu_X| = E|X|.$$

This may be found using (2.17) with $h(x) = |x|$, or by noting that, if $Y = X^2$, then $|X| = +\sqrt{Y}$. Then again (as in Examples 1.M and 2.I) using the fact that Y has a Chi Square (1 d.f.) distribution,

$$\begin{aligned} E|X| &= E(+\sqrt{Y}) \\ &= \int_0^\infty \sqrt{y}\frac{1}{\sqrt{2\pi y}}e^{-\frac{y}{2}}dy \\ &= \frac{1}{\sqrt{2\pi}} \cdot \int_0^\infty e^{-\frac{y}{2}}dy \\ &= \sqrt{\frac{2}{\pi}} \cdot \int_0^\infty \frac{1}{2}e^{-\frac{y}{2}}dy = \sqrt{\frac{2}{\pi}}, \end{aligned}$$

since we recognize the latter integrand as an Exponential $\left(\theta = \frac{1}{2}\right)$ density, whose integral must equal 1. Thus we expect (on average) that X will be $\sqrt{\frac{2}{\pi}} \approx .8$ units from its expectation, in absolute value. Note that by finding $E|X|$ to be finite, we show that the integral defining μ_X is not divergent, since (using an alternative expression for $E|X|$)

$$\begin{aligned} E|X| &= \int_{-\infty}^\infty |x|\phi(x)dx \\ &= \int_0^\infty |x|\phi(x)dx + \int_{-\infty}^0 |x|\phi(x)dx \\ &= \int_0^\infty x\phi(x)dx + \int_0^\infty x\phi(x)dx \\ &= 2\int_0^\infty x\phi(x)dx, \end{aligned}$$

because $\phi(x) = \phi(-x)$. Since we have shown this integral to be convergent $\left(\text{to } \sqrt{\frac{2}{\pi}}\right)$,

$$\begin{aligned} \mu_X &= \int_{-\infty}^\infty x\phi(x)dx \\ &= \int_0^\infty x\phi(x)dx + \int_{-\infty}^0 x\phi(x)dx \end{aligned}$$

is the sum of two convergent integrals, one the negative of the other.

The measure $E|X - \mu_X|$ may seem the most natural way of measuring dispersion, but for two reasons we adopt a different measure for most purposes. One reason is mathematical convenience; the other is that curiously, a measure based upon the squared dispersion arises most naturally from theoretical considerations.

The measure we shall make most use of is called the *variance;* it is the expectation of the transformation $h(x) = (x - \mu_X)^2$, the expected squared deviation of the random variable X from its expectation. It will be denoted

$$\text{Var}(X) = E[(X - \mu_X)^2] \tag{2.24}$$

or, alternatively, by σ_X^2 (or σ^2 when there is no confusion over which random variable is intended.) The variance is a measure of spread in much the same way that $E|X - \mu_X|$ is a measure of spread, but on a squared-units scale.

Example 2.L. Consider the discrete X with distribution

x	$-a$	0	a
$p_X(x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Then

$$\mu_X = E(X) = (-a)\cdot\frac{1}{4} + 0\cdot\frac{1}{2} + a\cdot\frac{1}{4} = 0$$

and; from (2.16),

$$\begin{aligned}\text{Var}(X) &= \sum_{\text{all x}} (x-0)^2 p_X(x)\\ &= (-a-0)^2\cdot\frac{1}{4} + (0-0)^2\cdot\frac{1}{2} + (a-0)^2\cdot\frac{1}{4}\\ &= a^2\cdot\frac{1}{4} + a^2\cdot\frac{1}{4} = \frac{a^2}{2}\end{aligned}$$

[Figure 2.6]

The larger a is, the larger the variance is.

It is difficult to interpret the variance because it is expressed in squared units, and so we shall define another measure in terms of Var(X). The *standard deviation* of a random variable X is given by

$$\sigma_X = \sqrt{\text{Var}(X)} = \sqrt{E(X-\mu_X)^2}. \tag{2.25}$$

Notice that in going to the standard deviation we are almost back to the first measure of spread we introduced. The standard deviation is given by $\sqrt{E(X-\mu_X)^2}$; our earlier measure was $E|X-\mu_X|$, which is equal to $E(\sqrt{(X-\mu_X)^2})$. These give different answers, however.

Example 2.L (continued). Here the standard deviation is given by

$$\sigma_X = \sqrt{\text{Var}(X)} = \frac{a}{\sqrt{2}},$$

while

$$\begin{aligned}E|X-\mu_X| &= \frac{1}{4}\cdot|-a-0| + \frac{1}{2}|0-0| + \frac{1}{4}|a-0|\\ &= \frac{a}{2}.\end{aligned}$$

Both are measures of spread that increase linearly with a, but they differ by a factor of $\sqrt{2}$. The preference for σ_X will be explained later, on statistical grounds.

Example 2.M. The Standard Normal Distribution. Let X have the standard normal density $\phi(x)$ (as in Examples 1.M and 2.K). In Example 2.K we found that

$$\mu_X = 0.$$

Then the variance is

$$\begin{aligned}\text{Var}(X) &= E(X-\mu_X)^2\\ &= E(X^2).\end{aligned}$$

But in Example 2.I we found that $E(Y) = E(X^2) = 1$, and so

$$\text{Var}(X) = 1$$

and also

$$\sigma_X = 1.$$

The calculation of variances is often simplified by the following device:

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

This latter step can be justified directly from (2.16) or (2.17); it also is a special case of a more general result we shall encounter in the next chapter, namely that the expected value of any sum of random variables is the sum of their expectations. Now

$$\begin{aligned}E(-2\mu X) &= -2\mu E(X) \\ &= -2\mu \cdot \mu = -2\mu^2,\end{aligned}$$

and

$$E(\mu^2) = \mu^2,$$

using (2.20) and (2.21), and so

$$\mathrm{Var}(X) = E(X^2) - \mu^2. \tag{2.26}$$

Example 2.N. Suppose X is a continuous random variable with probability density function

$$\begin{aligned}f_X(x) &= 2x \; 0 \le x \le 1 \\ &= 0 \quad \text{otherwise.}\end{aligned}$$

This is a particular case of a Beta density, with $\alpha = 2$ and $\beta = 1$; this fact and (2.14), or direct calculation of $\int_0^1 x \cdot 2x dx$, gives us $E(X) = \mu_X = 2/3$. To find $\mathrm{Var}(X)$ we could calculate from the definition,

$$\mathrm{Var}(X) = \int_0^1 \left(x - \frac{2}{3}\right)^2 2x dx,$$

but it is simpler to use (2.26). Find

$$E(X^2) = \int_0^1 x^2 \cdot 2x dx = 2\int_0^1 x^3 dx = \frac{1}{2},$$

Then

$$\mathrm{Var}(X) = \frac{1}{2} - \left(\frac{2}{3}\right)^2 = \frac{1}{18}.$$

Example 2.O. Discrete Uniform Distribution. If we choose a ticket at random from an urn containing $n+1$ tickets numbered sequentially from 0 through n tickets, the selected number X is a discrete random variable with probability distribution

$$\begin{aligned}p_X(x) &= \frac{1}{n+1} \quad \text{for } x = 0, 1, 2, \ldots, n \\ &= 0 \quad \text{otherwise.}\end{aligned}$$

Then we can use two elementary formulae for sums to find $E(X)$ and $\mathrm{Var}(X)$.

$$\sum_{x=0}^{n} x = \frac{n(n+1)}{2} \tag{2.27}$$

$$\sum_{x=0}^{n} x^2 = \frac{n(n+1)(2n+1)}{6}. \tag{2.28}$$

We have

$$\begin{aligned} E(X) &= \sum_{x=0}^{n} x \cdot \frac{1}{(n+1)} \\ &= \frac{1}{(n+1)} \cdot \frac{n(n+1)}{2} = \frac{n}{2}, \end{aligned} \tag{2.29}$$

$$\begin{aligned} E(X^2) &= \sum_{x=0}^{n} x^2 \cdot \frac{1}{(n+1)} \\ &= \frac{1}{(n+1)} \cdot \frac{n(n+1)(2n+1)}{6} = \frac{n(2n+1)}{6}. \end{aligned}$$

Then

$$\mathrm{Var}(X) = \frac{n(2n+1)}{6} - \left(\frac{n}{2}\right)^2 = \frac{n(n+2)}{12}. \tag{2.30}$$

These results will be useful later, when we discuss nonparametric tests.

2.4 Linear Change of Scale

As we have mentioned before (Examples 1.N and 2.J), the most common transformation is a linear transformation, a linear change of scale $Y = aX + b$. We may wish to translate results from one "ruler" to another, or we may wish to define a scale in such a way as to simplify the expression of our results. In Example 2.J we found the expectation of Y,

$$E(aX + b) = aE(X) + b. \tag{2.19}$$

or, in alternative notation,

$$E(aX + b) = a\mu_X + b. \tag{2.31}$$

The variance of Y has a simple form also.

Theorem: For any constants a and b,

$$\mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X). \tag{2.32}$$

Proof: By definition, $\mathrm{Var}(aX + b)$ is the expectation of

$$\begin{aligned} [(aX + b) - E(aX + b)]^2 &= [aX + b - (a\mu_X + b)]^2 \quad \text{(using (2.31))} \\ &= (aX - a\mu_X)^2 \\ &= (a^2(X - \mu_X)^2). \end{aligned}$$

But then (2.19) (or (2.20), in particular), tells us that

$$\begin{aligned} \mathrm{Var}(aX + b) &= E[a^2(X - \mu_X)^2] \\ &= a^2E(X - \mu_X)^2 = a^2\,\mathrm{Var}(X). \end{aligned}$$

Note that (2.19) and (2.32) could be written alternatively as

$$\mu_{aX+b} = a\mu_X + b \tag{2.33}$$

$$\sigma^2_{aX+b} = a^2\sigma^2_X,$$

and that we can immediately deduce the corresponding formula for the standard deviation,

$$\sigma_{aX+b} = |a|\sigma_X. \tag{2.35}$$

One important point to note is the obvious one that the variance and standard deviation of $aX + b$ do not depend on b, that the dispersion of a distribution is unaffected by the shift of the origin by b units, that $\text{Var}(aX + b) = \text{Var}(aX)$ and $\sigma_{aX+b} = \sigma_{aX}$.

An important use of these formulae is that they permit us to choose a scale of measurement with an eye toward simplification. By a suitable linear change of scale we can arrange to have a random variable expressed in what we will call *standardized form,* with expectation zero and variance (and standard deviation) one. This can be accomplished by transforming X by subtracting its expectation and dividing by its standard deviation:

$$W = \frac{X - \mu_X}{\sigma_X}. \tag{2.36}$$

Note that $W = aX + b$, for the special choices $a = 1/\sigma_X$ and $b = -\mu_X/\sigma_X$. Then

$$E(W) = \frac{1}{\sigma_X} \cdot \mu_X - \frac{\mu_X}{\sigma_X} = 0$$

and

$$\text{Var}(W) = \left(\frac{1}{\sigma_X}\right)^2 \sigma^2_X = 1;$$

That is

$$\mu_W = 0,$$

$$\sigma^2_W = \sigma_W = 1.$$

The statistical usefulness of expressing measurements in this standardized form, in standard deviation units with μ_X as the origin, will be apparent later.

2.5 The Normal (μ, σ^2) distributions

In Example 1.M we introduced the Standard Normal distribution as that of a continuous random variable X with density

$$\phi(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}} \quad \text{for } -\infty < x < \infty, \tag{1.32}$$

and we went on to define the *Normal* (μ, σ^2) *distribution* as the distribution of $Y = \sigma X + \mu$, which we found had density

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \quad \text{for } -\infty < y < \infty. \tag{1.38}$$

In Examples 2.K and 2.M we found the expectation and variance of X to be

$$\mu_X = 0,$$

$$\sigma^2_X = 1.$$

We can now apply (2.33) and (2.34) to see that $E(Y) = \sigma E(X) + \mu = \mu$, while $Var(Y) = \sigma^2\,\text{Var}(X) = \sigma^2$, or

$$\mu_Y = \mu,$$

$$\sigma^2_Y = \sigma^2,$$

$$\sigma_Y = \sigma.$$

These expressions explain our choice of notation and terminology; μ is indeed the mean or expectation of Y, and σ is the standard deviation of Y. We can also consider the inverse of the transformation that defines Y,

$$X = \frac{Y - \mu}{\sigma} \ ,$$

and see that X may be viewed as Y expressed in standardized form, hence the name "Standard Normal". We shall follow statistical custom and usually abbreviate "Normal(μ, σ^2)" to "$N(\mu, \sigma^2)$", which may be read as "the Normal distribution with mean (or expectation) μ and variance σ^2." Then $N(0,1)$ denotes the Standard Normal distribution.

By a linear change of scale, all probabilities connected with any Normal distribution can be reduced to the use of the table for the Standard Normal cumulative distribution function $\Phi(x)$.

Example 2.P. Suppose Y has a $N(5,16)$ distribution. What is the probability that Y is between 3 and 10? We can write

$$\begin{aligned} P(3 < Y < 10) &= P(Y < 10) - P(Y \leq 3) \\ &= P(Y \leq 10) - P(Y \leq 3). \end{aligned}$$

Now for any number a, if Y has a $N(\mu, \sigma^2)$ distribution,

$$\begin{aligned} P(Y \leq a) &= P(\sigma X + \mu \leq a) \\ &= P\left(X \leq \frac{a - \mu}{\sigma}\right) \\ &= \Phi\left(\frac{a - \mu}{\sigma}\right). \end{aligned}$$

Thus

$$\begin{aligned} P(3 < Y < 10) &= \Phi\left(\frac{10 - 5}{\sqrt{16}}\right) - \Phi\left(\frac{3 - 5}{\sqrt{16}}\right) \\ &= \Phi(1.25) - \Phi(-.5) \end{aligned}$$

The Table for $\Phi(x)$ gives $\Phi(1.25) = .8944$, and using the symmetry relation $\Phi(-x) = 1 - \Phi(x)$, $\Phi(-.5) = 1 - \Phi(.5) = 1 - .6915 = .3085$, so

$$P(3 < Y < 10) = .8944 - .3085 = .5859$$

[Figure 2.7]

As a mental check on any calculations involving a Normal distribution, it is helpful to keep in mind a few basic facts. For the Standard Normal distribution, $N(0,1)$,

$$\begin{aligned} P(-1 \leq X \leq 1) &= .6926 \approx 2/3 \\ P(-2 \leq X \leq 2) &= .9545 \approx .95 \\ P(-3 \leq X \leq 3) &= .9973 \approx 1 \end{aligned}$$

This means for example that for any Normal density, about 95% of the area is included within 2 standard deviations of the mean. In Figure 2.7, this is the range from -3 to 13, while the range from 1 to 9 includes about 2/3 of the area.

*2.6 Stieltjes Integrals

In defining expectation we were hampered by the necessity of defining $E(X)$ separately for discrete and continuous random variables, because we had used different mathematical devices (probability distribution functions p_X and density functions f_X) to describe their distributions. A mathematically more advanced treatment that permits a unified treatment of expectation takes advantage of the fact that the cumulative distribution function has the same definition for both cases,

$$F_X(x) = P(X \leq x),$$

and defines

$$E(X) = \int_{-\infty}^{\infty} x dF_X(x),$$

where the integral is a Stieltjes integral (properly defined in advanced calculus either as a Riemann-Stieltjes or as a Lebesgue-Stieltjes integral.) For the discrete or continuous cases, the meaning of $E(X)$ is unchanged (as are the ways it is computed), but the theoretical handing of $E(X)$ is simplified by being able to treat both cases together, as well as others that properly belong to neither, for example mixtures of the two cases.

Example 2.Q. Suppose X has the cumulative distribution function

$$\begin{aligned} F_X(x) &= 1 - \frac{1}{2}e^{-x} \quad \text{for } x \leq 0 \\ &= 0 \quad \text{for } x < 0. \end{aligned}$$

[Figure 2.8]

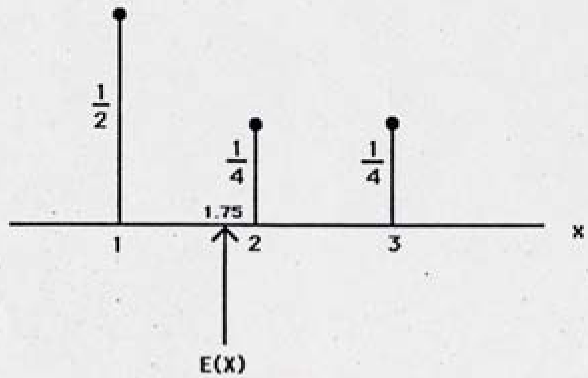
The random variable X is neither discrete nor continuous: While it has a jump at $x = 0$ it is not a jump function; while it has a nonzero derivative for $x > 0$,

$$F_X'(x) = \frac{1}{2}e^{-x} \quad \text{for } x > 0,$$

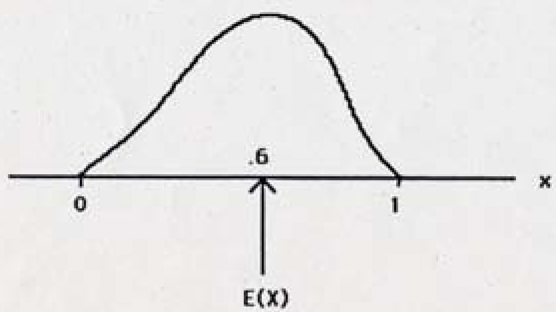
that derivative integrates to $\frac{1}{2}$, and so is not a proper density. Rather, F_X describes the distribution of a random variable which with probability $\frac{1}{2}$ equals zero (the discrete part) and with probability $\frac{1}{2}$ has an Exponential ($\theta = 1$) density. It is as if we select a lightbulb at random from two bulbs, one of which has an exponentially distributed lifetime, the other of which is dead already. We shall discuss how this description can be used to calculate $E(X)$ in the next chapter, but it can also be found as a Stieltjes integral:

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x dF_X(x) \\ &= 0 \cdot \frac{1}{2} + \int_0^{\infty} x \cdot F_X'(x) dx \\ &= \int_0^{\infty} x \cdot \frac{1}{2}e^{-x} dx = \frac{1}{2}. \end{aligned}$$

We shall not have much need of this unifying device, but it can be helpful to know of its availability, and it is essential in probability theory.



(a) E(X) for the discrete case of example 2A.



(b) E(X) for the continuous case of example 2B

Figure 2.1

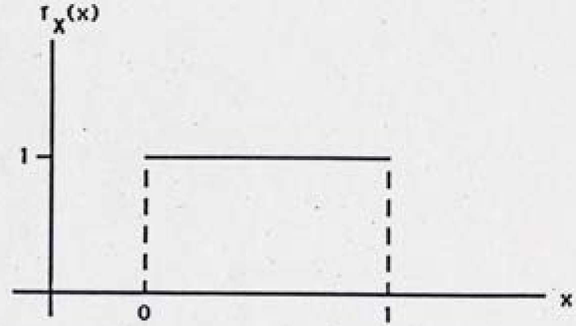


Figure 2.2. The Uniform(0, 1) density.

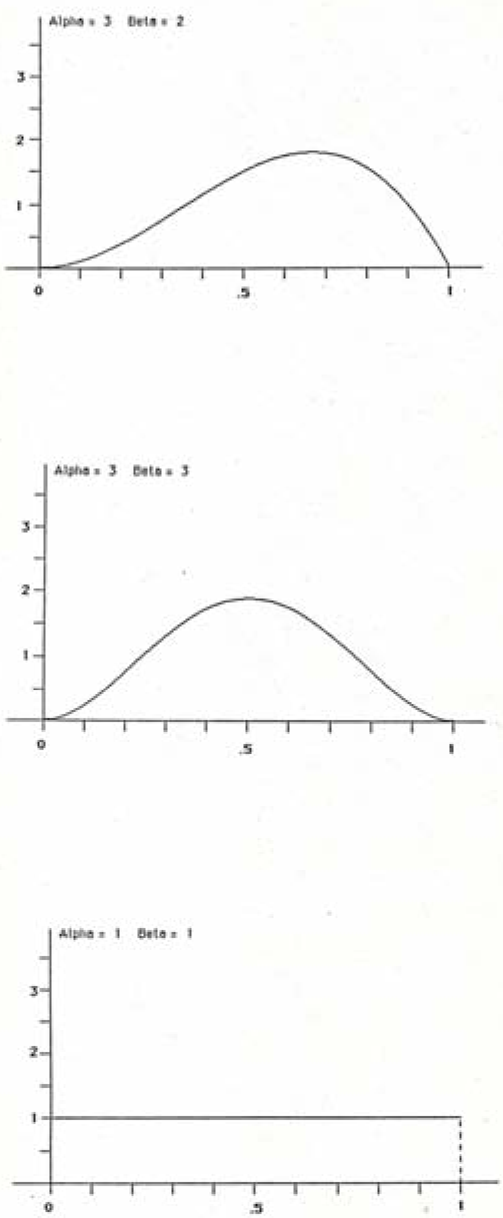


Figure 2.3. Examples of Beta densities.

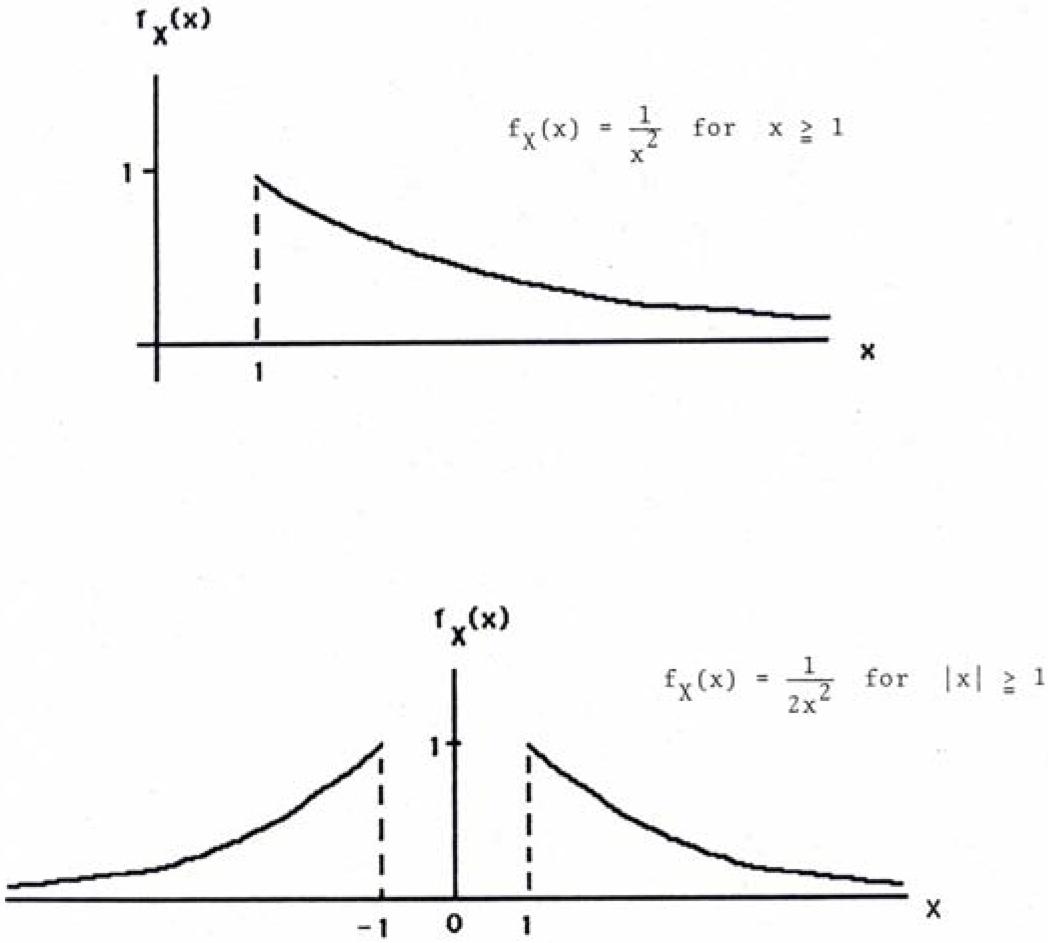
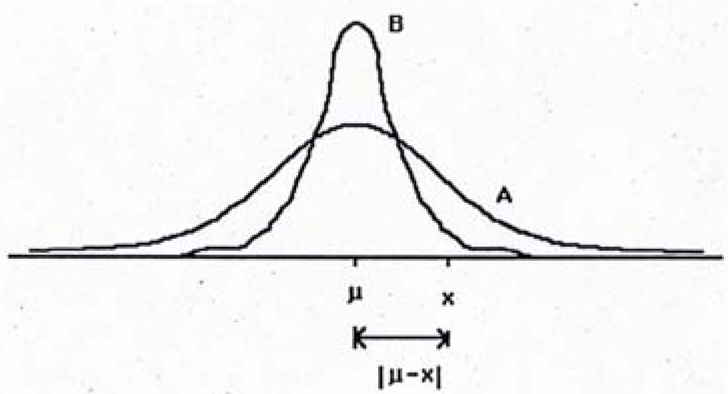
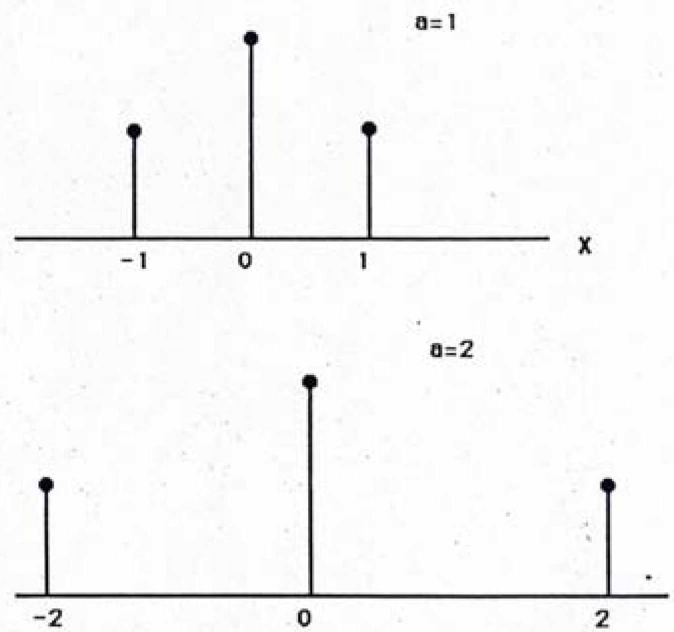


Figure 2.4



<u>Figure 2.5</u>. Two probability densities with equal expectations and unequal dispersion.



<u>Figure 2.6</u>

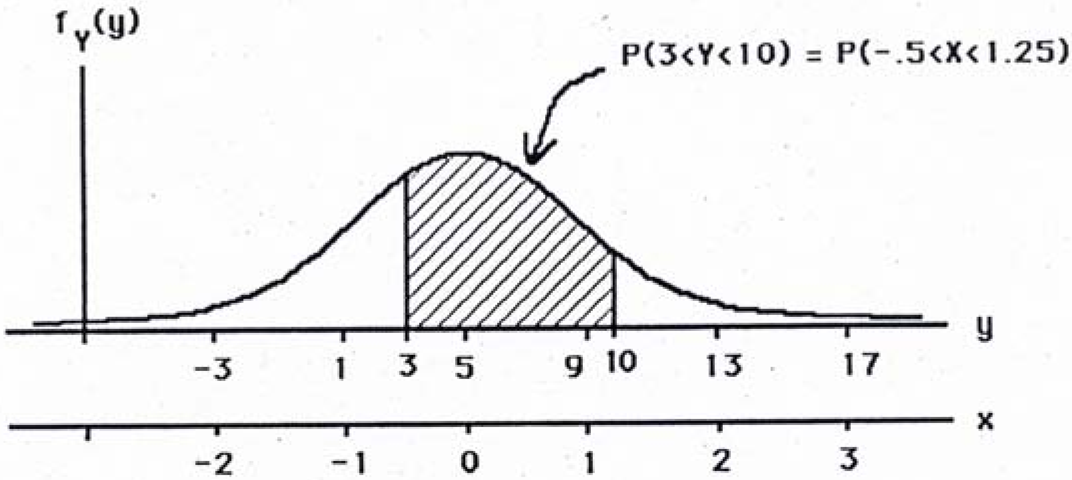


Figure 2.7

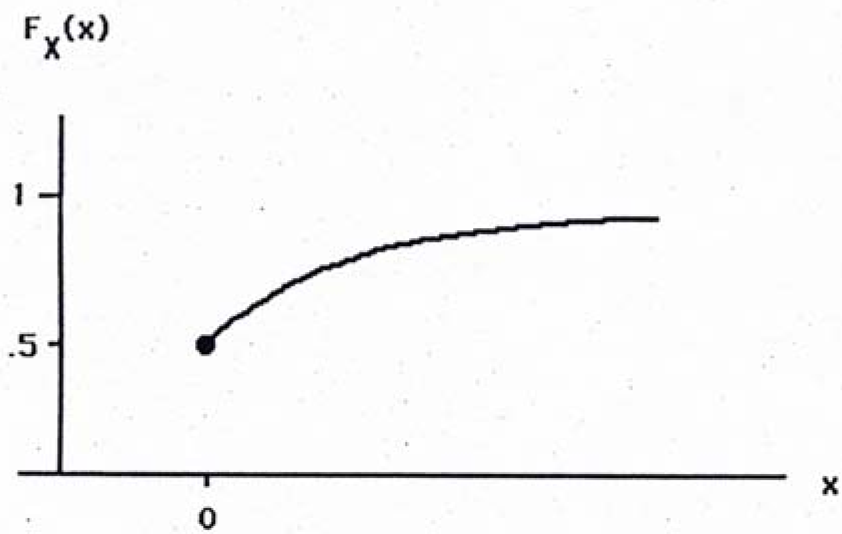


Figure 2.8