Chapter 4: Distributions

Prerequisite: Chapter 1

4.1 *The Algebra of Expectations and Variances*

In this section we will make use of the following symbols:

${}_n\mathbf{a}_1$ is a random variable
${}_n\mathbf{b}_1$ is a random variable
${}_n\mathbf{c}_1$ is a constant vector
${}_m\mathbf{D}_n$ is a constant matrix, and
${}_n\mathbf{F}_m$ is a constant matrix.

Now we define the *expectation of a continuous random variable*, such that

$$E(a_i) = \int_{-\infty}^{\infty} f(a_i)a_i da_i, \tag{4.1}$$

where $f(a_i)$ is the density of the probability distribution of a_i. Given that $f(a_i)$ is a density function, it must therefore be the case that

$$E(a_i) = \int_{-\infty}^{\infty} f(a_i)da_i = 1.$$

Often in this book, $f(a_i)$ will be taken to be normal, but not always. In fact, in some instances, a_i will be discrete rather than continuous. In that case,

$$E(a_i) = \sum_j^J \Pr(a_i = j) \cdot j \tag{4.2}$$

where there are J discrete possible outcomes for a_i. We call $E(\cdot)$ the *expectation operator*. Regardless as to whether **a** and **b** are normal, the following set of theorems apply. First, we note that the expectation of a constant is simply that constant itself:

$$E(\mathbf{c}) = \mathbf{c}. \tag{4.3}$$

The expectation of a sum is equal to the sum of the expectations:

$$E(\mathbf{a} + \mathbf{b}) = E(\mathbf{a}) + E(\mathbf{b}). \tag{4.4}$$

The expectation of a linear combination comes in two flavors; one for premultiplication and one for postmultiplication:

$$E(\mathbf{Da}) = \mathbf{D}E(\mathbf{a}). \tag{4.5}$$

$$E(\mathbf{a}'\mathbf{F}) = E(\mathbf{a}')\mathbf{F}. \tag{4.6}$$

You can see from the above two equations that a constant matrix can pass through the expectation operator, which often simplifies our algebra greatly. All of these theorems will be important in enabling statistical inference and in trying to understand the average of various quantities.

We now define the *variance operator*, V(·), such that

$$\mathrm{V}(\mathbf{a}) = \mathrm{E}\{[\mathbf{a} - \mathrm{E}(\mathbf{a})][\mathbf{a} - \mathrm{E}(\mathbf{a})]'\}. \tag{4.7}$$

We could note here that if E(**a**) = **0**, that is if **a** is mean centered, the variance of **a** simplifies to E(**aa**').

Whether **a** is mean centered or not we also have the following theorems:

$$\mathrm{V}(\mathbf{a} + \mathbf{c}) = \mathrm{V}(\mathbf{a}). \tag{4.8}$$

Equation (4.8) shows that the addition (or subtraction) of a constant vector does not modify the variance of the original random vector. That fact will prove useful to us quite often in the chapters to come. But now it is time to look at what is arguably the most important theorem of the book. At least it is safe to say that it is the most referenced equation in the book:

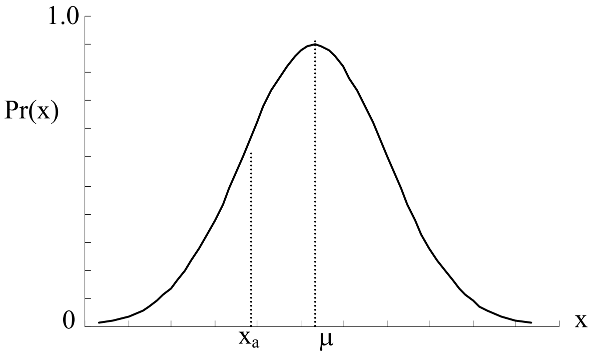
$$\mathrm{V}(\mathbf{Da}) = \mathbf{D}\mathrm{V}(\mathbf{a})\mathbf{D}' \tag{4.9}$$

$$\mathrm{V}(\mathbf{a}'\mathbf{F}) = \mathbf{F}'\mathrm{V}(\mathbf{a})\mathbf{F} \tag{4.10}$$

Equation (4.9), that shows that the variance of a linear combination is a quadratic form based on that linear combination, will be extremely useful to us, again and again in this book.

4.2 *The Normal Distribution*

The normal distribution is widely used in both statistical reasoning and in modeling marketing processes. It is so widely used that a short-hand notation exists to state that the variable x is normally distributed with mean μ and variance σ^2: x ~ N(μ, σ^2). We will start out by discussing the *density function* of the normal distribution even though the distribution function is somewhat more fundamental (it is, after all, called the normal distribution) and in fact the density is derived from the distribution function rather than vice versa. In any case, the density gives the probability that a variable takes on a particular value. We plot this probability as a function of the value:



The equation that sketches out the bell shaped curve in the figure is

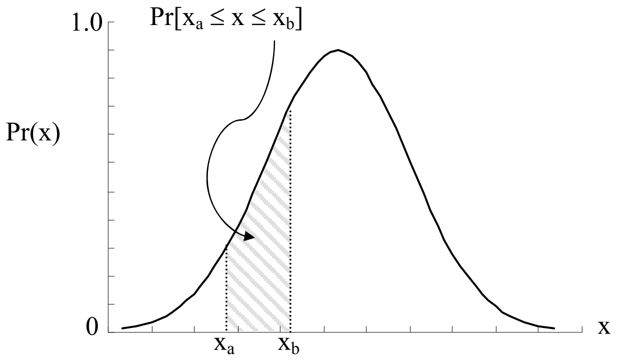
$$\Pr(x) \equiv \Pr(x = x_a) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-(x_a - \mu)^2}{2\sigma^2}\right]. \qquad (4.11)$$

Most of the "action" takes place in the exponent [and here we remind you that exp(x) = e^x]. In fact, the constant $1/\sqrt{2\pi}\sigma$ is needed solely to make sure that the total probability under the curve equals one, or in other words, that the function integrates to 1. You might also note that the σ is not under the radical sign. Alternatively you can include a σ^2 under the radical. When we standardize such that $\mu = 0$ and $\sigma^2 = 1$ we generally rename x_a to z_a and then

$$\Pr(z) \equiv \Pr(z = z_a) = \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-z_a^2}{2}\right] = \phi(z_a) \quad . \qquad (4.12)$$

Note that $\phi(\cdot)$ is a very widely used notational convention to refer to the standard normal density function. This will show up in many places in the chapters to follow.

In statistical reasoning, we are often interested in the probability that a normal variable falls between two particular values, say x_a and x_b. We can picture this situation as below:



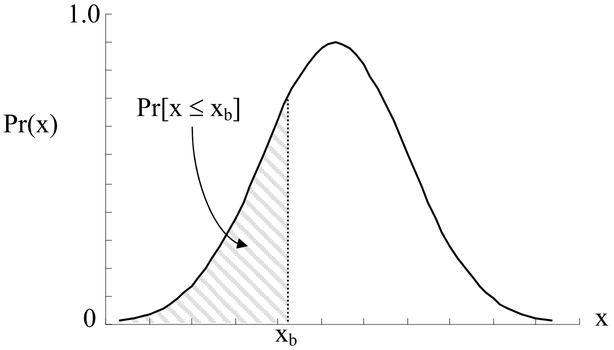
We can derive the probability by integrating the area under the curve from x_a to x_b. There is no analytic answer – that is to say no equation will allow you to calculate the exact value – so the only way you can do it is by a brute force computer program that creates a series of tiny rectangles between x_a and x_b. If the bases of these rectangles become sufficiently small, even though the top of the function is obviously not flat, we can approximate this probability to an arbitrary precision by adding up the areas of these rectangles. We write this area using the integral symbol as below:

$$\Pr[x_a \le x \le x_b] = \frac{1}{\sqrt{2\pi}\sigma} \int_{x_a}^{x_b} \exp\left[\frac{-(x_a - \mu)^2}{2\sigma^2}\right] dx.$$

We can standardize, using the calculus change-of-variables technique, and then move the constant under the integral, all of which yields the same probability as above. This is shown next:

$$\Pr[z_a \le z \le z_b] = \int_{z_a}^{z_b} \phi(z)\, dz \; .$$

We are now ready to define the *normal distribution function*, which means the probability that x is less than or equal to some value, like x_b. This is pictured below:



Here, to calculate this probability, we must integrate the left tail of the distribution, starting at -∞ at ending up at x_b. This will give us the probability that a normal variate x is less than x_b:

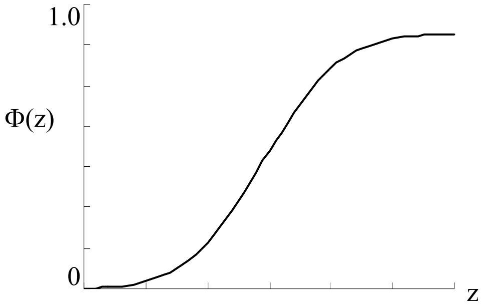
$$\Pr[x \le x_b] = \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{x_b} \exp\left[\frac{-(x_a - \mu)^2}{2\sigma^2}\right] dx \quad \text{or} \tag{4.13}$$

$$= \int_{-\infty}^{z_b} \phi(z)\, dz = \Phi(z_b)\,. \tag{4.14}$$

Note the notation $\Phi(z_b)$ implies the probability that $z \le z_b$ The symbol Φ is an uppercase phi while ϕ is the lowercase version of that Greek letter. It is traditional to use a lower case letter for a function, while the integral of that function is signified with the upper case version of that letter. Note also that

$$\frac{\partial \Phi(z)}{\partial z} = \phi(z)\,. \tag{4.15}$$

A graphical representation of Φ(z) is show below:



The curve pictured above is often called an *ogive*.

In many cases, for example cases having to do with choice probabilities in Chapter 12, we wish to know that probability that a random variate is greater than 0:

$$\Pr(x \geq 0) = \Phi(\mu / \sigma) \equiv \Phi\left[E(x) \middle/ \sqrt{V(x)}\right] . \tag{4.16}$$

4.3 *The Multivariate Normal Distribution*

For purposes of comparison, let us take the normal distribution as presented in the previous section,

$$\Pr(x) \equiv \Pr(x = x_a) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[\frac{-(x_a - \mu)^2}{2\sigma^2}\right]$$

and rewrite it a little bit. For one thing, $\sqrt{a} = a^{1/2}$. In that case, rewriting the above gives us

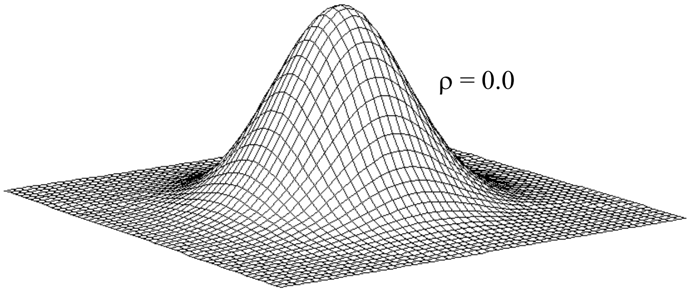
$$\Pr(x = x_a) = \frac{1}{(2\pi)^{1/2}\left(\sigma^2\right)^{1/2}} \exp\left[\frac{-(x_a - \mu)^2}{2\sigma^2}\right].$$

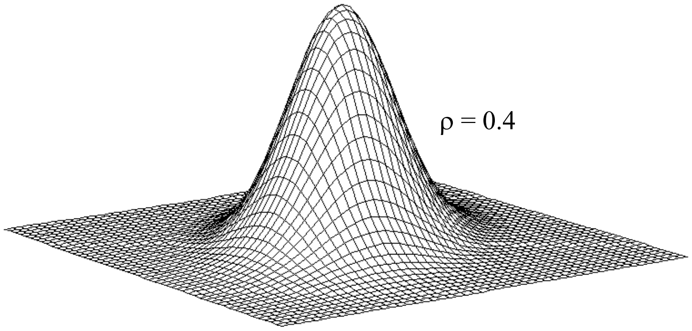
Now lets say we have a column vector of p variables, **x**, and that **x** follows the multivariate normal distribution with mean vector **μ** (which is also p by 1), and variance matrix **Σ** (which is a symmetric p by p matrix). In that case, the probability that the random vector **x** takes on the set of m values that we will call $\mathbf{x}_a$ is given by

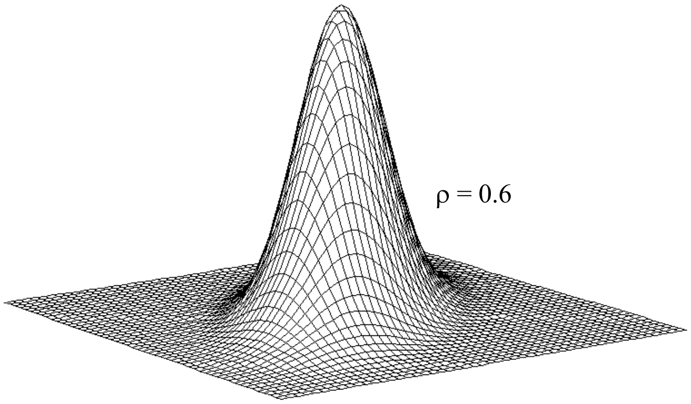
$$\Pr(\mathbf{x} = \mathbf{x}_a) = \frac{1}{(2\pi)^{p/2} \, |\boldsymbol{\Sigma}|^{1/2}} \exp\left[(\mathbf{x}_a - \boldsymbol{\mu})' \, \boldsymbol{\Sigma}^{-1} (\mathbf{x}_a - \boldsymbol{\mu}) / 2\right]. \tag{4.17}$$

We would ordinarily use a short-hand notation for Equation (4.17), saying that **x** ~ N(**μ**, **Σ**).

Making some analogies, in the univariate expression σ^2 appears in the denominator (of the exponent) while in the multivariate case we have $\boldsymbol{\Sigma}^{-1}$ filling the same role. You might also notice that in the fraction before the exponent, we see σ in the univariate case, but $|\boldsymbol{\Sigma}|^{1/2}$ shows up in the multivariate case, the square root of the determinant of the variance matrix. In the univariate case there is the square root of 2π, in the multivariate we see the (p/2)th root of 2π. A picture of the bivariate normal density function appears below for three different values of the correlation $\rho = \sigma_{12} / \sigma_1 \sigma_2$.







4.4 *Chi Square*

We have already seen that the scalar y, where $y \sim N(\mu, \sigma^2)$, can be converted to a z score, $z \sim N(0, 1)$ where $z = \frac{y - \mu}{\sigma}$. If I square that z score I end up with a chi square variate with one degree of freedom, i. e.

$$z^2 = \chi_1^2 .$$

More generally, if I have a vector $\mathbf{y} = [y_1 \quad y_2 \quad \cdots \quad y_n]'$ and if **y** is normally distributed with mean vector

$$\boldsymbol{\mu} = \begin{bmatrix} \mu \\ \mu \\ \cdots \\ \mu \end{bmatrix}$$

and variance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2\mathbf{I}$$

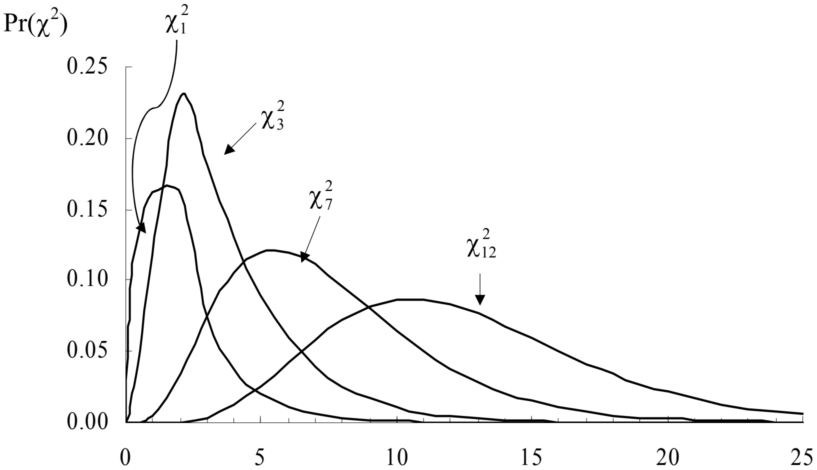
we of course say that y ~ N($\boldsymbol{\mu}$, $\sigma^2\mathbf{I}$). Converting each of the y_i to z scores, that is

$$z_i = \frac{y_i - \mu}{\sigma}$$

for all i, 1, 2, ···, n; we have $\mathbf{z} = [z_1 \quad z_2 \quad \cdots \quad z_n]'$. We can say that the vector $\mathbf{z}$ ~ N(0, $\mathbf{I}$). In that case,

$$\mathbf{z}'\mathbf{z} = \sum_i^n z_i^2 \sim \chi_n^2.$$

The Chi Square density function is approximated in the following figure, using several different degrees of freedom to illustrate the shape.



With small degrees of freedom, the distribution looks like a normal for which the left tail has been folded over the right. This is more or less what happens when we square something - we fold the negative half over the positive. With larger degrees of freedom, the Chi Square begins to resemble the normal again, and in fact, as can be seen in the graph, the similarity is already quite striking at 12 degrees of freedom. This similarity is virtually complete by 30 degrees of freedom.

4.5 *Cochran's Theorem*

For any n · 1 vector $\mathbf{z} \sim N(0, \mathbf{I})$ and for any set of n · n matrices $\mathbf{A}_i$ where $\sum_i^n \mathbf{A}_i = \mathbf{I}$, then

$$\sum_i^n \mathbf{z}'\mathbf{A}_i\mathbf{z} = \mathbf{z}'\mathbf{z} \tag{4.18}$$

which, as we have just seen, is distributed as χ^2_n. Further, if the rank (see Section 3.7) of $\mathbf{A}_i$ is r_i we can say that

$$\sum_i^n r_i = n \quad \text{and} \tag{4.19}$$

$$\mathbf{z}'\mathbf{A}_i\mathbf{z} \sim \chi^2_{r_i}. \tag{4.20}$$

Each quadratic form $\mathbf{z}'\mathbf{A}_i\mathbf{z}$ is an independent Chi Square. The sum of independent Chi Square values is also a Chi Square variable with degrees of freedom equal to the sum of the component's degrees of freedom. This allows us to test nested models, such as those found in Chapters 9 and 10 as well as Chapters 12 and 13. In addition, multiple degree of freedom hypothesis testing for the linear model is based on this theorem as well. Defining $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\mathbf{M} = \mathbf{I} - \mathbf{P}$, then since

$$\mathbf{y}'\mathbf{y} = \mathbf{y}'\mathbf{I}\mathbf{y} = \mathbf{y}'\mathbf{P}\mathbf{y} + \mathbf{y}'\mathbf{M}\mathbf{y},$$

we have met the requirements of Cochran's Theorem and we can form an F ratio using the two components, $\mathbf{y}'\mathbf{P}\mathbf{y}$ and $\mathbf{y}'\mathbf{M}\mathbf{y}$. In addition, the component $\mathbf{y}'\mathbf{P}\mathbf{y}$ can be further partitioned using the hypothesis matrix $\mathbf{A}$ or restricted models.

4.6 *Student's* t-*Statistic*

Like the normal distribution, the Chi Square is derived with a known value of σ. The formula for Chi Square on n degrees of freedom is

$$\chi^2_n = \sum_i^n \frac{(y_i - \mu)^2}{\sigma^2} = \sum_i^n \frac{[(y_i - \bar{y}) + (\bar{y} - \mu)]^2}{\sigma^2}. \tag{4.21}$$

You will note in the numerator of the right hand piece, a $\bar{y}$ has been added and subtracted. Now we will square the numerator of that right hand piece which yields

$$\chi^2_n = \frac{1}{\sigma^2} \sum_i^n \frac{(y_i - \bar{y})^2 + 2y_i\bar{y} - 2y_i\mu - 2\bar{y}^2 + 2\bar{y}\mu + \bar{y}^2 - 2\bar{y}\mu + \mu^2}{\sigma^2}. \tag{4.22}$$

At this time, we can modify Equation (4.22) by distributing the Σ addition operator, canceling some terms, and taking advantage of the fact that

$$\sum_{i}^{n} y_i = n\bar{y}.$$

Doing so, we find that

$$\chi_n^2 = \sum_{i}^{n} \frac{(y_i - \bar{y})^2}{\sigma^2} + \frac{n(\bar{y} - \mu)^2}{\sigma^2}. \tag{4.23}$$

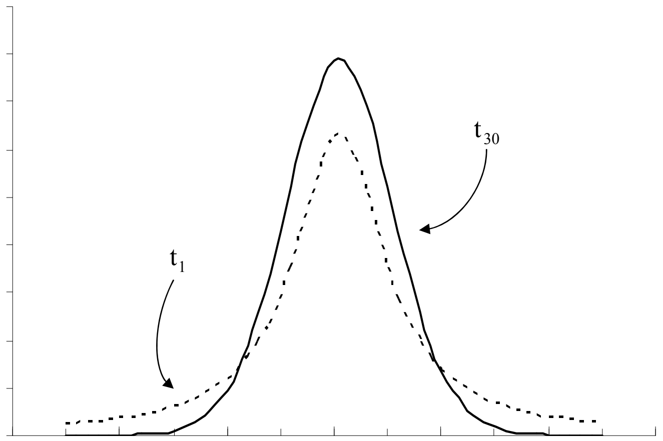
You might note that at this point Equation (4.23) shows the decomposition of an n degree of freedom Chi Square into two components which Cochran's Theorem shows us are both themselves distributed as Chi Square. But the numerator of the summation on the right hand side, that is $\sum_{i}^{n} (y - \bar{y})^2$, is the corrected sum of squares and as such it is equivalent to (n - 1)s^2. Rewriting both components slightly we have

$$\chi_n^2 = \frac{(n-1)s^2}{\sigma^2} + \frac{(\bar{y} - \mu)^2}{\sigma^2/n} \tag{4.24}$$

which leaves us with two Chi Squares. The one on the right is a z-score squared and has one degree of freedom. The reader might recognize it as a z score for the arithmetic mean, $\bar{y}$. The Chi Square on the left has n - 1 degrees of freedom. At this point, to get the unknown value σ^2 to vanish we need only create a ratio. In fact, to form a *t*-statistic, we do just that. In addition, we divide by the n - 1 degrees of freedom in order to make the *t* easier to tabulate:

$$\begin{aligned} t &= \sqrt{\frac{(\bar{y} - \mu)^2}{\sigma^2/n} \bigg/ \frac{(n-1)s^2}{\sigma^2(n-1)}} \\ &= \frac{\bar{y} - \mu}{s/\sqrt{n}} \end{aligned} \tag{4.25}$$

The more degrees of freedom a *t* distribution has, the more it resembles the normal. The resemblance is well on its way by the time you reach 30 degrees of freedom. Below you can see a graph that compares the approximate density functions for *t* with 1 and with 30 df.



The 1 df function has much more weight in the tails, as it must be more conservative.

4.7 *The F Distribution*

With the F statistic, a ratio is also formed. However, in the case of the F, we do not take the square root, and the numerator χ^2 is not restricted to one degree of freedom:

$$F_{r_1,r_2} = \frac{\chi^2_{r_1}/r_1}{\chi^2_{r_2}/r_2} \quad .$$