Joint Probability Distributions and Random Samples (Devore Chapter Five)

1016-345-01: Probability and Statistics for Engineers*

Spring 2013

Contents

Tuesday 16 April 2013

1 Joint Probability Distributions

Consider a scenario with more than one random variable. For concreteness, start with two, but methods will generalize to multiple ones.

1.1 Two Discrete Random Variables

Call the rvs X and Y. The generalization of the pmf is the *joint probability mass function*, which is the probability that X takes some value x *and* Y takes some value y:

$$p(x, y) = P((X = x) \cap (Y = y)) \tag{1.1}$$

Since X and Y have to take on some values, all of the entries in the joint probability table have to sum to 1:

$$\sum_x \sum_y p(x, y) = 1 \tag{1.2}$$

We can collect the values into a table: Example: problem 5.1:

		y		
$p(x,y)$		0	1	2
	0	.10	.04	.02
x	1	.08	.20	.06
	2	.06	.14	.30

This means that for example there is a 2% chance that $x = 1$ and $y = 2$. Each combination of values for X and Y is an outcome that occurs with a certain probability. We can combine those into events; e.g., the event $(X \leq 1) \cap (Y \leq 1)$ consists of the outcomes in which (X, Y) is $(0, 0)$, $(0, 1)$, $(1, 0)$, and $(1, 1)$. If we call this set of (X, Y) combinations A, the probability of the event is the sum of all of the probabilities for the outcomes in A:

$$P((X, Y) \in A) = \sum_{(x,y)\in A} \sum p(x, y) \tag{1.3}$$

So, specifically in this case,

$$P((X \leq 1) \cap (Y \leq 1)) = .10 + .04 + .08 + .20 = .42 \tag{1.4}$$

The events need not correspond to rectangular regions in the table. For instance, the event $X < Y$ corresponds to (X, Y) combinations of $(0, 1)$, $(0, 2)$, and $(1, 2)$, so

$$P(X < Y) = .04 + .02 + .06 = .12 \tag{1.5}$$

Another event you can consider is $X = x$ for some x, regardless of the value of Y. For example,

$$P(X = 1) = .08 + .20 + .06 = .34 \tag{1.6}$$

But of course $P(X = x)$ is just the pmf for X alone; when we obtain it from a joint pmf, we call it a marginal pmf:

$$p_X(x) = P(X = x) = \sum_y p(x, y) \tag{1.7}$$

and likewise

$$p_Y(y) = P(Y = y) = \sum_x p(x, y) \tag{1.8}$$

For the example above, we can sum the columns to get the marginal pmf $p_Y(y)$:

y	0	1	2
$p_Y(y)$	.24	.38	.38

or sum the rows to get the marginal pmf $p_X(x)$:

x	$p_X(x)$
0	.16
1	.34
2	.50

They're apparently called marginal pmfs because you can write the sums of columns and rows in the margins:

$p(x,y)$		y 0	1	2	$p_X(x)$
	0	.10	.04	.02	.16
x	1	.08	.20	.06	.34
	2	.06	.14	.30	.50
	$p_Y(y)$	.24	.38	.38	

1.1.1 Independence of Random Variables

Recall that two events A and B are called independent if (and only if) $P(A \cap B) = P(A)P(B)$. That definition extends to random variables:

Two random variables X and Y are independent if and only if the events $X = x$ and $Y = y$ are independent for all choices of x and y, i.e., if $p(x, y) = p_X(x)p_Y(y)$ for all x and y.

We can check if this is true for our example. For instance, $p_X(2)p_Y(2) = (.50)(.38) = .19$ while $p(2, 2) = .30$ so $p(2, 2) \neq p_X(2)p_Y(2)$, which means that X and Y are *not* independent. (If X and Y were independent, we'd have to check that by checking $p(x, y) = p_X(x)p_Y(y)$ for each possible combination of x and y.)

1.2 Two Continuous Random Variables

We now consider the case of two continuous rvs. It's not really convenient to use the cdf like we did for one variable, but we can extend the definition of the pdf by considering the probability that X and Y lie in a tiny box centered on (x, y) with sides Δx and Δy. This probability will go to zero as either Δx or Δy goes to zero, but if we divide by $\Delta x\, \Delta y$ we

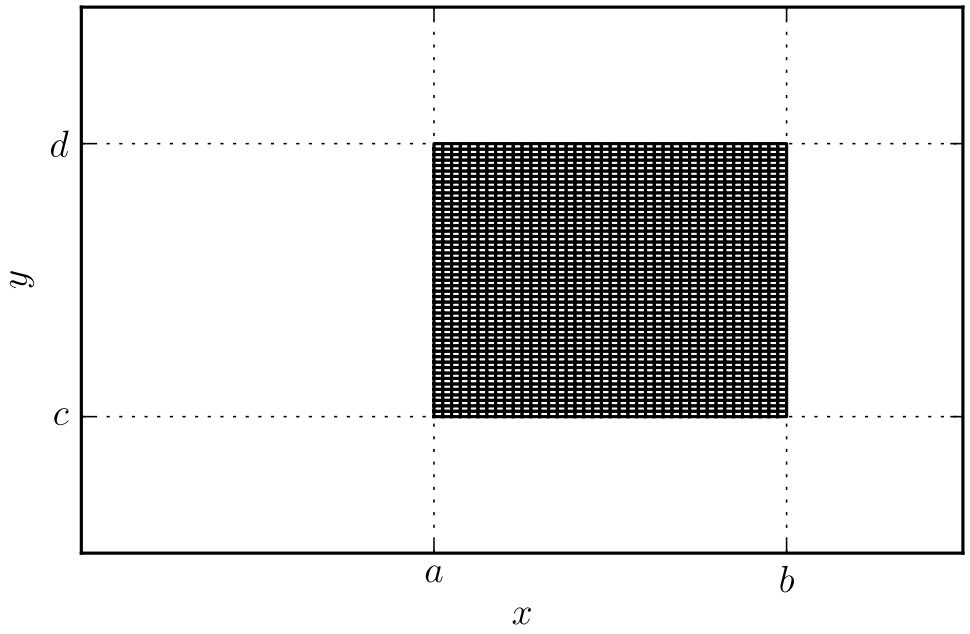
get something which remains finite. This is the *joint pdf*:

$$f(x,y) = \lim_{\Delta x, \Delta y \to 0} \frac{P\left(\left(x - \frac{\Delta x}{2} < X < x + \frac{\Delta x}{2}\right) \cap \left(y - \frac{\Delta y}{2} < Y < y + \frac{\Delta y}{2}\right)\right)}{\Delta x \Delta y} \tag{1.9}$$

We can construct from this the probability that (X, Y) will lie within a certain region

$$P((X,Y) \in A) \approx \sum_{(x,y)\in A} \sum f(x,y)\,\Delta x\,\Delta y \tag{1.10}$$

What is that, exactly? First, consider special case of a rectangular region where $a < x < b$ and $c < y < d$. Divide it into M pieces in the x direction and N pieces in the y direction:

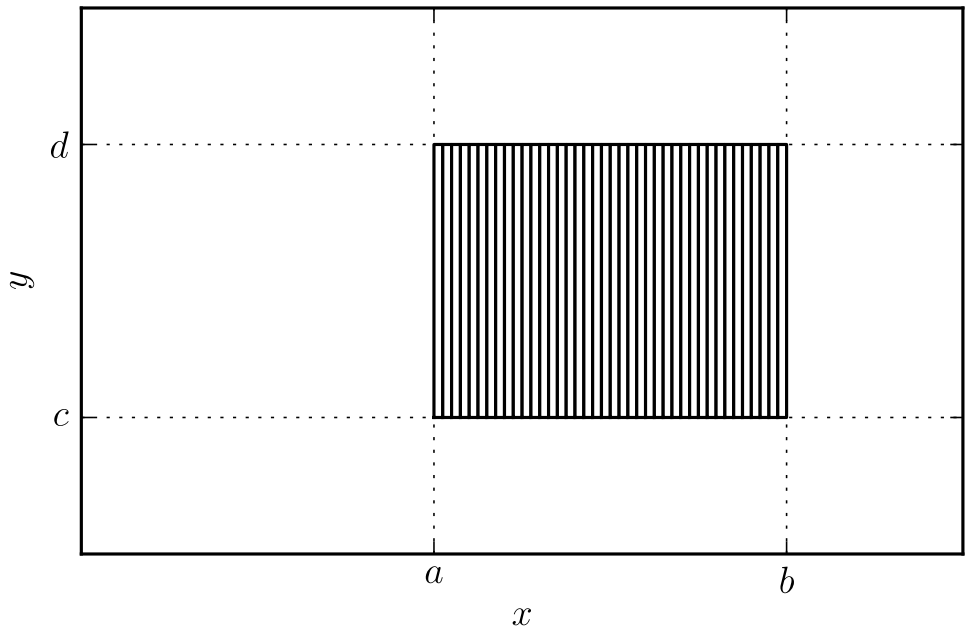


$$P((a < X < b) \cap (c < Y < d)) \approx \sum_{i=1}^{M} \sum_{j=1}^{N} f(x_i, y_j)\,\Delta x\,\Delta y \tag{1.11}$$

Now,

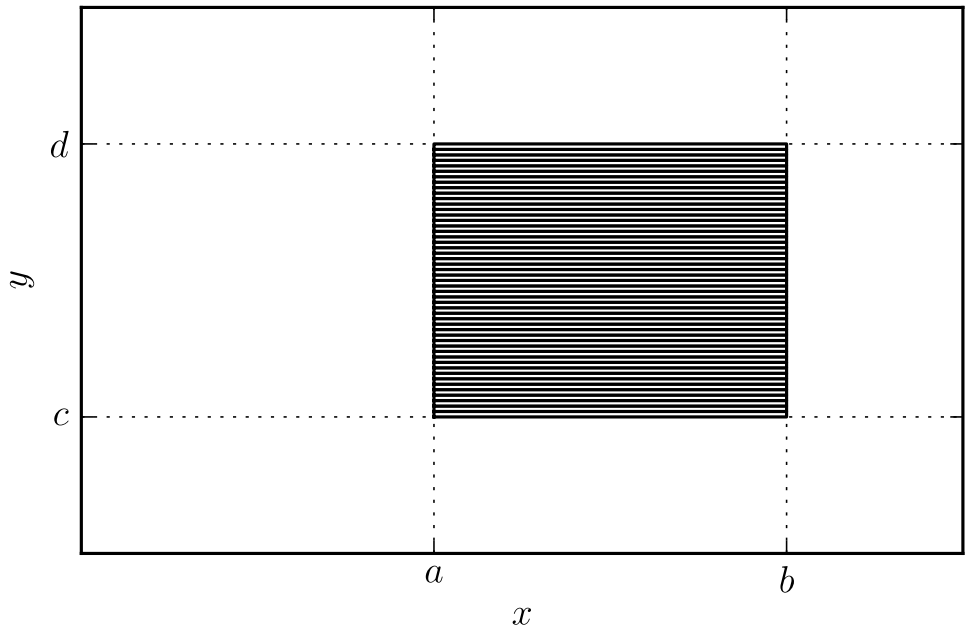
$$\sum_{j=1}^{N} f(x_i, y_j)\,\Delta y \approx \int_c^d f(x_i, y)\,dy \tag{1.12}$$

so



$$P((a < X < b) \cap (c < Y < d)) \approx \sum_{i=1}^{M} \int_c^d f(x_i, y)\, dy\, \Delta x \approx \int_a^b \left(\int_c^d f(x, y)\, dy \right) dx \quad (1.13)$$

This is a double integral. We integrate with respect to y and then integrate with respect to x. Note that there was nothing special about the order we approximated the sums as integrals. We could also have taken the limit as the x spacing went to zero first, and then found



$$P((a < X < b) \cap (c < Y < d)) \approx \sum_{j=1}^{N} \sum_{i=1}^{M} f(x_i, y_j)\, \Delta x\, \Delta y \approx \sum_{j=1}^{N} \int_a^b f(x, y_j)\, dx\, \Delta y \approx \int_c^d \left(\int_a^b f(x, y)\, dx \right) dy \tag{1.14}$$

As long as we integrate over the correct region of the x, y plane, it doesn't matter in what order we do it.

All of these approximations get better and better as Δx and Δy go to zero, so the expression involving the integrals is actually exact:

$$P((a < X < b) \cap (c < Y < d)) = \int_a^b \left(\int_c^d f(x, y)\, dy \right) dx = \int_c^d \left(\int_a^b f(x, y)\, dx \right) dy \tag{1.15}$$

The normalization condition is that the probability density for all values of X and Y has to integrate up to 1:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1 \tag{1.16}$$

1.3 Collection of Formulas

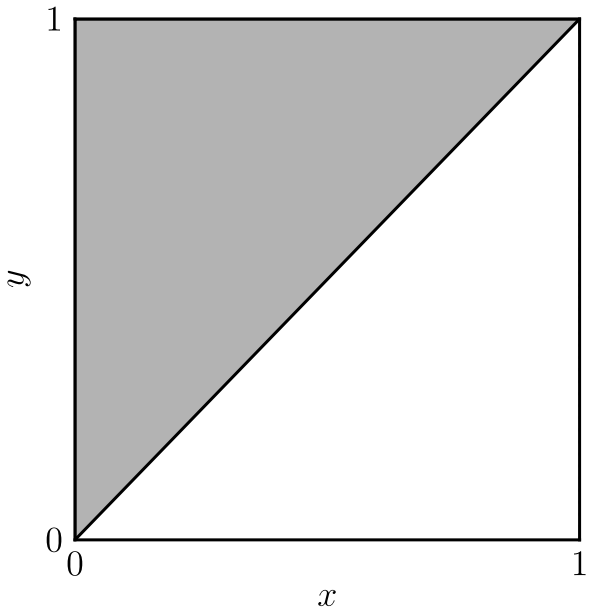
	Discrete	Continuous
Definition	$p(x, y) = P([X = x] \cap [Y = y])$	$f(x, y) \approx \frac{P\left(\left[X = x \pm \frac{\Delta x}{2}\right] \cap \left[Y = y \pm \frac{\Delta y}{2}\right]\right)}{\Delta x\, \Delta y}$
Normalization	$\sum_x \sum_y p(x, y) = 1$	$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$
Region A	$P((X, Y) \in A) = \sum\sum_{(x,y) \in A} p(x, y)$	$P((X, Y) \in A) = \int\int_{(x,y) \in A} f(x, y)\, dx\, dy$
Marginal	$p_X(x) = \sum_y p(x, y)$	$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$
X & Y indep *iff*	$p(x, y) = p_X(x) p_Y(y)$ for all x, y	$f(x, y) = f_X(x) f_Y(y)$ for all x, y

1.4 Example of Double Integration

Last time, we calculated the probability that a pair of continuous random variables X and Y lie within a rectangular region. Now let's consider how we'd integrate to get the probability that (X, Y) lie in a less simple region, specifically $X < Y$. For simplicity, assume that the joint pdf $f(x, y)$ is non-zero only for $0 \leq x \leq 1$ and $0 \leq y \leq 1$. So we want to integrate

$$P(X < Y) = \int\int_{x<y} f(x, y)\, dx\, dy \qquad (1.17)$$

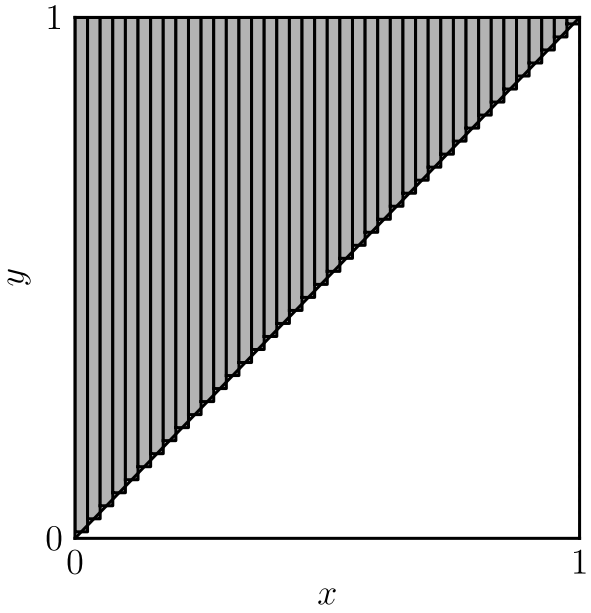
The region we want to integrate over is shown in the figure at right:



If we do the y integral first, that means that, for each value of x, we need work out the range of y values. This means slicing up the triangle along lines of constant x and integrating in the y direction for each of those (see right)
The conditions for a given x are $y \geq 0$, $y > x$, and $y \leq 1$. So the minimum possible y is x and the maximum is 1, and the integral is

$$P(X < Y) = \int_0^1 \left(\int_x^1 f(x, y)\, dy \right) dx \qquad (1.18)$$

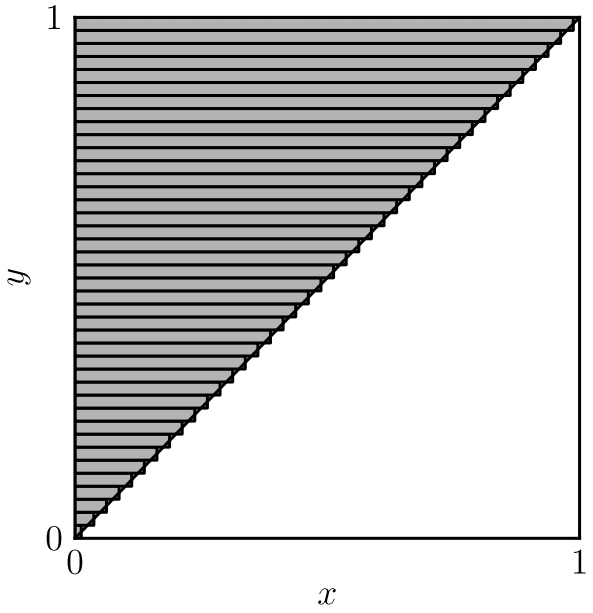
We integrate x over the full range of x, from 0 to 1, because there is a strip present for each of those xs. Note that the limits of the y integral depend on x, which is fine because the y integral is inside the x integral.



If, on the other hand, we decided to do the x integral first, we'd be slicing up into slices of constant x and considering the range of x values for each possible y (see right).
Now, given a value of y, the restrictions on x are $x \geq 0$, $x < y$, and $x \leq 1$, so the minimum x is 0 and the maximum is y, which makes the integral

$$P(X < Y) = \int_0^1 \left(\int_0^y f(x,y)\, dx \right) dy \quad (1.19)$$

The integral over y is over the whole range from 0 to 1 because there is a constant-y strip for each of those values.



Note that the limits on the integrals are different depending on which order the integrals are done. The fact that we get the same answer (which should be apparent from the fact that we're covering all the points of the same region) is apparently called Fubini's theorem.

Practice Problems

5.1, 5.3, 5.9, 5.13, 5.19, 5.21

Thursday 18 April 2013

2 Expected Values, Covariance and Correlation

We can extend further concepts to the realm of multiple random variables. For instance, the expected value of any function $h(X,Y)$ which can be constructed from random variables X and Y is taken by multiplying the value of the function corresponding to each outcome by the probability of that outcome:

$$E(h(X,Y)) = \sum_x \sum_y h(x,y)\, p(x,y) \quad (2.1)$$

In the case of continuous rvs, we replace the pmf with the pdf and the sums with integrals:

$$E(h(X,Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,y)\, f(x,y)\, dx\, dy \quad (2.2)$$

(All of the properties we'll show in this section work for both discrete and continuous distributions, but we'll do the explicit demonstrations in one case; the other version can be produced via a straightforward conversion.)

In particular, we can define the mean value of each of the random variables as before

$$\mu_X = E(X) \tag{2.3a}$$
$$\mu_Y = E(Y) \tag{2.3b}$$

and likewise the variance:

$$\sigma_X^2 = V(X) = E((X-\mu_X)^2) = E(X^2) - \mu_X^2 \tag{2.4a}$$
$$\sigma_Y^2 = V(Y) = E((Y-\mu_Y)^2) = E(Y^2) - \mu_Y^2 \tag{2.4b}$$

Note that each of these is a function of only one variable, and we can calculate the expected value of such a function without having to use the full joint probability distribution, since e.g.,

$$E(h(X)) = \sum_x \sum_y h(x)\, p(x,y) = \sum_x h(x) \sum_y p(x,y) = \sum_x h(x)\, p_X(x)\ . \tag{2.5}$$

Which makes sense, since the marginal pmf $p_X(x) = P(X = x)$ is just the probability distribution we'd assign to X if we didn't know or care about Y.

A useful expected value which is constructed from a function of both variables is the *covariance*. Instead of squaring $(X - \mu_X)$ or $(Y - \mu_Y)$, we multiply them by each other:

$$\mathrm{Cov}(X,Y) = E([X-\mu_X][Y-\mu_Y]) = E(XY) - \mu_X\mu_Y \tag{2.6}$$

It's worth going explicitly through the derivation of the shortcut formula:

$$\begin{aligned} E([X-\mu_X][Y-\mu_Y]) &= E(XY - \mu_X Y - \mu_Y X + \mu_X\mu_Y) \\ &= E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X\mu_Y \\ &= E(XY) - \mu_X\mu_Y - \cancel{\mu_Y\mu_X} + \cancel{\mu_X\mu_Y} \end{aligned} \tag{2.7}$$

We can show that the covariance of two independent random variables is zero:

$$\begin{aligned} E(XY) &= \sum_x \sum_y x\, y\, p(x,y) = \sum_x \sum_y x\, y\, p_X(x)\, p_Y(y) = \sum_x \left(x\, p_X(x) \left(\sum_y y\, p_Y(y) \right) \right) \\ &= \sum_x x\, p_X(x) \mu_Y = \mu_Y \sum_x x\, p_X(x) = \mu_X\mu_Y \qquad \text{if } X \text{ and } Y \text{ are independent} \end{aligned} \tag{2.8}$$

The converse is, however, not true. We can construct a probability distribution in which X and Y are not independent but their covariance is zero:

$p(x,y)$		y = −1	0	1	$p_X(x)$
	−1	0	.2	0	.2
x	0	.2	.2	.2	.6
	1	0	.2	0	.2
	$p_Y(y)$	.2	.6	.2	

From the form of the marginal pmfs, we can see $\mu_X = 0 = \mu_Y$, and if we calculate

$$E(XY) = \sum_x \sum_y x\, y\, p(x, y) \tag{2.9}$$

we see that for each x, y combination for which $p(x, y) \neq 0$, either x or y is zero, and so $\mathrm{Cov}(X, Y) = E(XY) = 0$.

Unlike variance, covariance can be positive or negative. For example, consider two rvs with the joint pmf

		y			
$p(x,y)$		-1	0	1	$p_X(x)$
	-1	.2	0	0	.2
x	0	0	.6	0	.6
	1	0	0	.2	.2
	$p_Y(y)$	.2	.6	.2	

Since we have the same marginal pmfs as before, $\mu_X \mu_Y = 0$, and

$$\mathrm{Cov}(X, Y) = E(XY) = \sum_x \sum_y x\, y\, p(x, y) = (-1)(-1)(.2) + (1)(1)(.2) = .4 \tag{2.10}$$

The positive covariance means X tends to be positive when Y is positive and negative when Y is negative.

On the other hand if the pmf is

		y			
$p(x,y)$		-1	0	1	$p_X(x)$
	-1	0	0	.2	.2
x	0	0	.6	0	.6
	1	.2	0	0	.2
	$p_Y(y)$	.2	.6	.2	

which again has $\mu_X \mu_Y = 0$, the covariance is

$$\mathrm{Cov}(X, Y) = E(XY) = \sum_x \sum_y x\, y\, p(x, y) = (-1)(1)(.2) + (1)(-1)(.2) = -.4 \tag{2.11}$$

One drawback of covariance is that, like variance, it depends on the units of the quantities considered. If the numbers above were in feet, the covariance would really be $.4\,\mathrm{ft}^2$; if you converted X and Y into inches, so that 1 ft became 12 in, the covariance would become 57.6 (actually $57.6\,\mathrm{in}^2$). But there wouldn't really be any increase in the degree to which X and Y are correlated; the change of units would just have spread things out. The same thing would happen to the variances of each variable; the covariance is measuring the spread of each variable along with the correlation between them. To isolate a measure of how

correlated the variables are, we can divide by the product of their standard deviations and define something called the *correlation coëfficient*:

$$\rho_{X,Y} \equiv \mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\,\sigma_Y} \tag{2.12}$$

To see why the product of the standard deviations is the right thing, suppose that X and Y have different units, e.g., X is in inches and Y is in seconds. Then μ_X is in inches, σ_X^2 is in inches squared, and σ_X is in inches. Similar arguments show that μ_Y and σ_Y are in seconds, and thus $\mathrm{Cov}(X,Y)$ is in inches times seconds, so dividing by $\sigma_X\sigma_Y$ makes all of the units cancel out.

Exercise: work out σ_X, σ_Y, and $\mathrm{Corr}(X,Y)$ for each of these examples.

Practice Problems

5.25, 5.27, 5.31, 5.33, 5.37, 5.39

Tuesday 23 April 2013

3 Statistics Constructed from Random Variables

3.1 Random Samples

We've done explicit calculations so far on pairs of random variables, but of course everything we've done extends to the general situation where we have n random variables, which we'll call X_1, X_2, $\ldots X_n$. Depending on whether the rvs are discrete or continuous, there is a joint pmf or pdf

$$p(x_1, x_2, \ldots, x_n) \qquad \text{or} \qquad f(x_1, x_2, \ldots x_n) \tag{3.1}$$

from which probabilities and expected values can be calculated in the usual way.

A special case of this is if the n random variables are independent. Then the pmf or pdf factors, e.g., for the case of a continuous rv,

$$X_1, X_2, \ldots X_n \text{ independent means } f(x_1, x_2, \ldots x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n) \tag{3.2}$$

An even more special case is when each of the rvs follows the same distribution, which we can then write as e.g., $f(x_1)$ rather than $f_{X_1}(x_1)$. Then we say the n rvs are *independent and identically distributed* or iid. Again, writing this explicitly for the continuous case,

$$X_1, X_2, \ldots X_n \text{ iid means } f(x_1, x_2, \ldots x_n) = f(x_1) f(x_2) \cdots f(x_n) \tag{3.3}$$

This is a useful enough concept that it has a name, and we call it a *random sample* of size n drawn from the distribution with pdf $f(x)$. For example, if we roll a die fifteen times, we can make statements about those fifteen numbers taken together.

3.2 What is a Statistic?

Often we'll want to combine the n numbers X_1, X_2, $\ldots X_n$ into some quantity that tells us something about the sample. For example, we might be interested in the average of the numbers, or their sum, or the maximum among them. Any such quantity constructed out of a set of random variables is called a *statistic*. (It's actually a little funny to see "statistic" used in the singular, since one thinks of the study of statistics as something like physics or mathematics, and we don't talk about "a physic" or "a mathematic".) A statistic constructed from random variables is itself a random variable.

3.3 Mean and Variance of a Random Sample

Remember that if we had a specific set of n numbers $x_1, x_2, \ldots x_n$, we could define the sample mean and sample variance as

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.4a}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{3.4b}$$

On the other hand, if we have a single random variable X, its mean and variance are calculated with the expected value:

$$\mu_X = E(X) \tag{3.5a}$$

$$\sigma_X^2 = E([X - \mu_X]^2) \tag{3.5b}$$

So, given a set of n random variables, we could combine them in the same ways we combine sample points, and define the mean and variance of those n numbers

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3.6a}$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{3.6b}$$

Each of these is a statistic, and each is a random variable, which we stress by writing it with a capital letter. We could then ask about quantities derived from the probability distributions of the statistics, for example

$$\mu_{\overline{X}} = E(\overline{X}) \tag{3.7a}$$

$$(\sigma_{\overline{X}})^2 = V(\overline{X}) = E([\overline{X} - \mu_{\overline{X}}]^2) \tag{3.7b}$$

$$\mu_{S^2} = E(S^2) \tag{3.7c}$$

$$(\sigma_{S^2})^2 = V(S^2) = E([S^2 - \mu_{S^2}]^2) \tag{3.7d}$$

Of special interest is the case where the $\{X_i\}$ are an iid random sample. We can actually work the means and variances just by using the fact that the expected value is a linear operator. So

$$\mu_{\overline{X}} = E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu_X = \mu_X \tag{3.8}$$

which is kind of what we'd expect: the average of n iid random variables has an expected value which is equal to the mean of the underlying distribution. But we can also consider what the variance of the mean is. I.e., on average, how far away will the mean calculated from a random sample be from the mean of the underlying distribution?

$$(\sigma_{\overline{X}})^2 = V(\overline{X}) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = E\left(\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \mu_X\right]^2\right) \tag{3.9}$$

It's easy to get a little lost squaring the sum, but we can see the essential point of what happens in the case where $n = 2$:

$$\begin{aligned}(\sigma_{\overline{X}})^2 &= E\left(\left[\frac{1}{2}(X_1 + X_2) - \mu_X\right]^2\right) = E\left(\left[\frac{1}{2}(X_1 - \mu_X) + \frac{1}{2}(X_2 - \mu_X)\right]^2\right) \\ &= E\left(\frac{1}{4}(X_1 - \mu_X)^2 + \frac{1}{2}(X_1 - \mu_X)(X_2 - \mu_X) + \frac{1}{4}(X_2 - \mu_X)^2\right)\end{aligned} \tag{3.10}$$

Since the expected value operation is linear, this is

$$\begin{aligned}(\sigma_{\overline{X}})^2 &= \frac{1}{4}E([X_1 - \mu_X]^2) + \frac{1}{2}E([X_1 - \mu_X][X_2 - \mu_X]) + \frac{1}{4}E([X_2 - \mu_X]^2) \\ &= \frac{1}{4}V(X_1) + \frac{1}{2}\,\mathrm{Cov}(X_1, X_2) + \frac{1}{4}V(X_2)\end{aligned} \tag{3.11}$$

But since X_1 and X_2 are independent random variables, their covariance is zero, and

$$(\sigma_{\overline{X}})^2 = \frac{1}{4}\sigma_X^2 + \frac{1}{4}\sigma_X^2 = \frac{1}{2}\sigma_X^2 \tag{3.12}$$

The same thing works for n iid random variables: the cross terms are all covariances which equal zero, and we get n copies of $\frac{1}{n^2}\sigma_X^2$ so that

$$(\sigma_{\overline{X}})^2 = \frac{1}{n}\sigma_X^2 \tag{3.13}$$

This means that if you have a random sample of size n, the sample mean will be a better estimate of the underlying population mean the larger n is. (There are assorted anecdotal examples of this in Devore.)

Note that these statements can also be made about the sum

$$T_o = X_1 + X_2 + \cdots + X_n = n\,\overline{X} \tag{3.14}$$

rather than the mean, and they are perhaps easier to remember:

$$\mu_{T_o} = E(T_o) = n\mu_X \tag{3.15a}$$
$$(\sigma_{T_o})^2 = V(T_o) = n\sigma_X^2 \tag{3.15b}$$

3.4 Sample Variance Explained at Last

We now have the machinery needed to show that $\mu_{S^2} = \sigma_X^2$. This result is what validates using the factor of $\frac{1}{n-1}$ rather than $\frac{1}{n}$ in the definition of sample variance a couple of weeks ago, so it's worth doing as a demonstration of the power of the formalism of random samples.

The sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{3.16}$$

is a random variable. Its mean value is

$$E(S^2) = E\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = \frac{1}{n-1}\sum_{i=1}^{n} E\left([X_i - \overline{X}]^2\right) \tag{3.17}$$

If we write

$$X_i - \overline{X} = (X_i - \mu_X) - (\overline{X} - \mu_X) \tag{3.18}$$

we can look at the expected value associated with each term in the sum:

$$\begin{aligned} E\left([X_i - \overline{X}]^2\right) &= E\left(\left[(X_i - \mu_X) - (\overline{X} - \mu_X)\right]^2\right) \\ &= E\left([(X_i - \mu_X)]^2\right) - 2E\left([X_i - \mu_X]\left[\overline{X} - \mu_X\right]\right) + E\left(\left[\overline{X} - \mu_X\right]^2\right) \end{aligned} \tag{3.19}$$

The first and last terms are just the variance of X_i and $\overline{X}$, respectively:

$$E\left([(X_i - \mu_X)]^2\right) = V(X_i) = \sigma_X^2 \tag{3.20}$$
$$E\left(\left[(\overline{X} - \mu_X)\right]^2\right) = V(\overline{X}) = \frac{\sigma_X^2}{n} \tag{3.21}$$

To evaluate the middle term, we can expand out $\overline{X}$ to get

$$E\left([X_i - \mu_X]\left[\overline{X} - \mu_X\right]\right) = \frac{1}{n}\sum_{j=1}^{n} E\left([X_i - \mu_X]\,[X_j - \mu_X]\right) \tag{3.22}$$

Since the rvs in the sample are independent, the covariance between *different* rvs is zero:

$$E\left([X_i - \mu_X]\,[X_j - \mu_X]\right) = \mathrm{Cov}(X_i, X_j) = \begin{cases} 0 & i \neq j \\ \sigma_X^2 & i = j \end{cases} \tag{3.23}$$

Since there is only one term in the sum for which $j = i$,

$$\sum_{j=1}^{n} E\left([X_i - \mu_X]\,[X_j - \mu_X]\right) = \sigma_X^2 \tag{3.24}$$

and

$$E\left([X_i - \mu_X]\left[\overline{X} - \mu_X\right]\right) = \frac{\sigma_X^2}{n} \tag{3.25}$$

Putting it all together we get

$$E\left([X_i - \overline{X}]^2\right) = \sigma_X^2 - 2\frac{\sigma_X^2}{n} + \frac{\sigma_X^2}{n} = \left(1 - \frac{1}{n}\right)\sigma_X^2 = \frac{n-1}{n}\sigma_X^2 \tag{3.26}$$

and thus

$$E(S^2) = \frac{1}{n-1}\sum_{i=1}^{n}\frac{n-1}{n}\,\sigma_X^2 = \sigma_X^2 \tag{3.27}$$

Q.E.D.!

Practice Problems

5.37, 5.39, 5.41, 5.45, 5.49

Thursday 25 April 2013

3.5 Linear Combinations of Random Variables

These results about the means of a random sample are special cases of general results from section 5.5, where we consider a general linear combination of *any* n rvs (not necessarily iid)

$$Y = a_1X_1 + a_2X_2 + \cdots a_nX_n = \sum_{i=1}^{n} a_iX_i \tag{3.28}$$

and show that

$$\mu_Y = E(Y) = a_1E(X_1) + a_2E(X_2) + \cdots + a_nE(X_n) = a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n \tag{3.29}$$

and

$$\text{if } X_1, \ldots, X_n \text{ independent,} \qquad V(Y) = {a_1}^2V(X_1) + {a_2}^2V(X_2) + \cdots + {a_n}^2V(X_n) \tag{3.30}$$

The first result follows from linearity of the expected value; the second can be illustrated for $n = 2$ as before:

$$\begin{aligned} V(a_1X_1 + a_2X_2) &= E\left([(a_1X_1 + a_2X_2) - \mu_Y]^2\right) = E\left([(a_1X_1 + a_2X_2) - (a_1\mu_1 + a_2\mu_2)]^2\right) \\ &= E\left([a_1(X_1 - \mu_1) + a_2(X_2 - \mu_2)]^2\right) \\ &= E\left({a_1}^2(X_1 - \mu_1)^2 + 2a_1a_2(X_1 - \mu_1)(X_2 - \mu_2) + {a_2}^2(X_2 - \mu_2)^2\right) \\ &= {a_1}^2V(X_1) + 2a_1a_2\,\mathrm{Cov}(X_1, X_2) + {a_2}^2V(X_2) \end{aligned} \tag{3.31}$$

and then the cross term vanishes if X_1 and X_2 are independent.

3.6 Linear Combination of Normal Random Variables

To specify the distribution of a linear combination of random variables, beyond its mean and variance, you have to know something about the distributions of the individual variables. One useful result is that a linear combination of independent normally-distributed random variables is itself normally distributed. We'll show this for the special case of the sum of two independent standard normal random variables:

$$Y = Z_1 + Z_2 \tag{3.32}$$

How do we get the pdf $p(Y)$? Well, for one rv, we can fall back on our trick of going via the cumulative distribution function

$$F(y) = P(Y \leq y) = P(Z_1 + Z_2 \leq y) \tag{3.33}$$

Now, $Z_1 + Z_2 \leq y$ is just an event which defines a region in the z_1, z_2 plane, so we can evaluate its probability as

$$F(y) = \iint\limits_{z_1+z_2\leq y} f(z_1, z_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{y-z_2} \frac{e^{-{z_1}^2/2}}{\sqrt{2\pi}}\frac{e^{-{z_2}^2/2}}{\sqrt{2\pi}}\, dz_1\, dz_2 \tag{3.34}$$

Where the limits of integration for the z_1 integral come from the condtion $z_1 \leq y - z_2$. Since z_1 and z_2 are just integration variables, let's rename them to w and z to reduce the amount of writing we'll have to do:

$$F(y) = \int_{-\infty}^{\infty}\int_{-\infty}^{y-z} \frac{e^{-w^2/2}}{\sqrt{2\pi}}\frac{e^{-z^2/2}}{\sqrt{2\pi}}\, dw\, dz = \int_{-\infty}^{\infty} \Phi\left(y - z\right)\frac{e^{-z^2/2}}{\sqrt{2\pi}}\, dz \tag{3.35}$$

We've used the usual definition of the standard normal cdf to do the w integral, but now we're sort of stuck with the z integral. But if we just want the pdf, we can differentiate with respect to y, using

$$\frac{d}{dy}\Phi\left(y - z\right) = \frac{e^{-(y-z)^2/2}}{\sqrt{2\pi}} \tag{3.36}$$

so that

$$f(y) = F'(y) = \int_{-\infty}^{\infty} \frac{e^{-(y-z)^2/2}}{\sqrt{2\pi}}\frac{e^{-z^2/2}}{\sqrt{2\pi}}\, dz = \int_{-\infty}^{\infty} \frac{1}{2\pi}\exp\left(\frac{-(y-z)^2 - z^2}{2}\right) dz \tag{3.37}$$

The argument of the exponential is

$$\frac{-(y-z)^2 - z^2}{2} = -z^2 + yz - \frac{1}{2}y^2 = -\left(z - \frac{y}{2}\right)^2 + \frac{1}{4}y^2 - \frac{1}{2}y^2 \tag{3.38}$$

If we define a new integration variable u by

$$\frac{u}{\sqrt{2}} = z - \frac{y}{2} \tag{3.39}$$

so that $dz = du/\sqrt{2}$, the integral becomes

$$\begin{aligned} f(y) &= \int_{-\infty}^{\infty} \frac{1}{2\pi} \exp\left(-\frac{u^2}{2} - \frac{y^2}{4}\right) du = \frac{1}{\sqrt{2}\sqrt{2\pi}}\, e^{-y^2/4} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du \\ &= \frac{1}{\sqrt{2}\sqrt{2\pi}}\, e^{-y^2/(2[\sqrt{2}]^2)} \end{aligned} \tag{3.40}$$

which is just the pdf for a normal random variable with mean $0+0 = 0$ and variance $1+1 = 2$.

The demonstration a linear combination of standard normal rvs, or for the sum of normal rvs with different means, is similar, but there's a little more algebra to keep track of the different σs.

4 The Central Limit Theorem

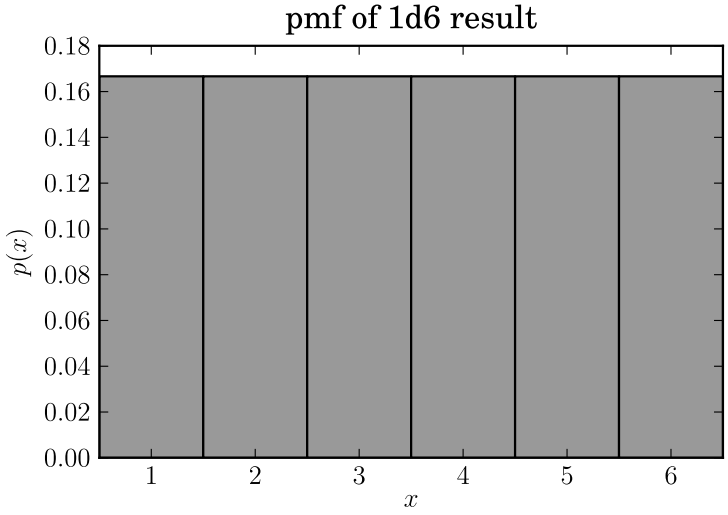
We have shown that if you add a bunch of independent random variables, the resulting statistic has a mean which is the sum of the individual means and a variance which is the sum of the individual variances. There is remarkable theorem which means that if you add up enough iid random variables, the mean and variance are all you need to know. This is known as the *Central Limit Theorem*:

> If X_1, X_2, $\ldots X_n$ are independent random variables each from the same distribution with mean μ and variance σ_X^2, their sum $T_o = X_1 + X_2 + \cdots + X_n$ has a distribution which is approximately a normal distribution with mean $n\mu$ and variance $n\sigma_X^2$, with the approximation being better the larger n is.

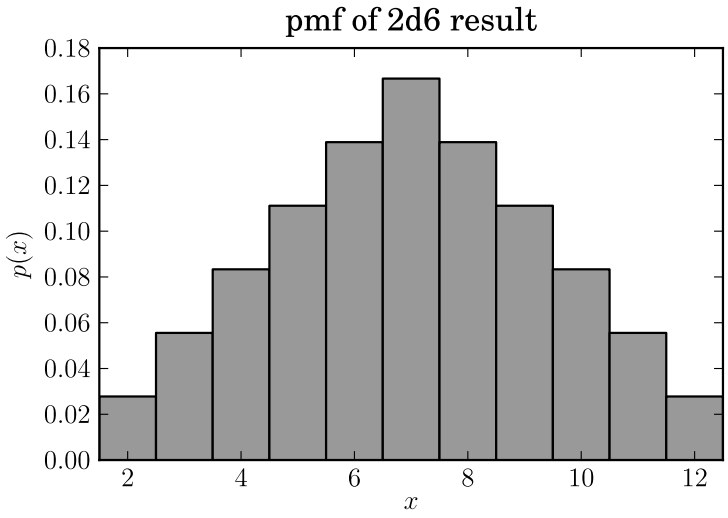
The same can also be said for the sample mean $\overline{X} = T_o/n$, but now the mean is μ and the variance is σ_X^2/n. As a rule of thumb, the central limit theorem applies for $n \gtrsim 30$.

We have actually already used the central limit theorem when approximating a binomial distribution with the corresponding normal distribution. A binomial rv can be thought of as the sum of n Bernoulli random variables.

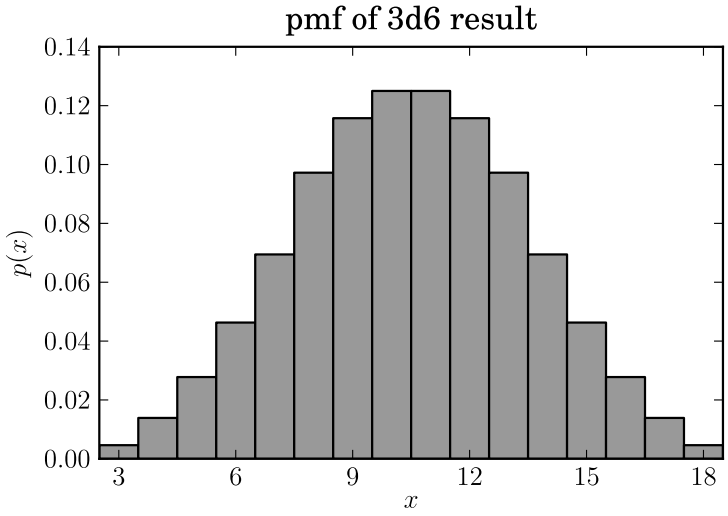
You may also have seen the central limit theorem in action if you've considered the pmf for the results of rolling several six-sided dice and adding the results. With one die, the results are uniformly distributed:



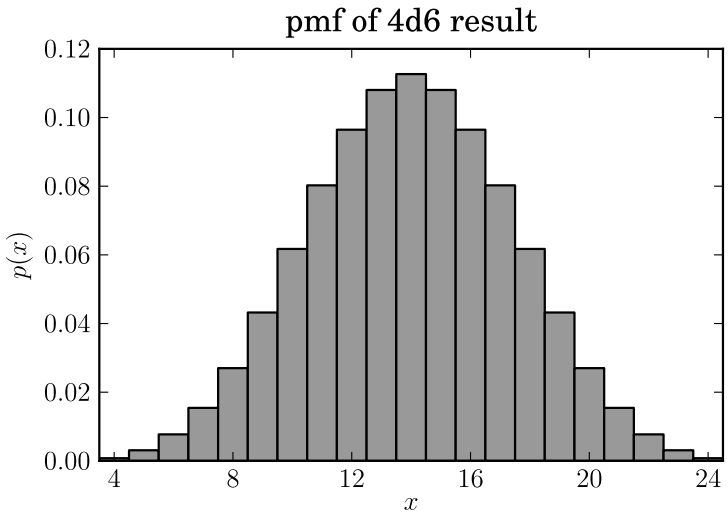
If we roll two dice and add the results, we get a non-uniform distribution, with results close to 7 being most likely, and the probability distribution declining linearly from there:

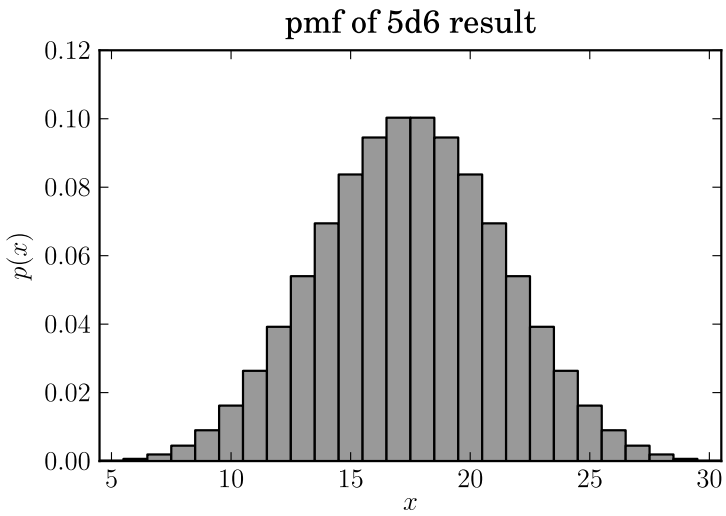


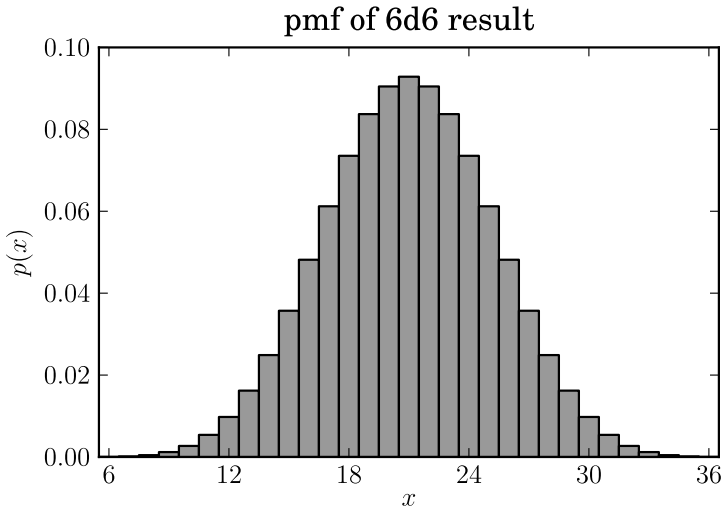
If we add three dice, the distribution begins to take on the shape of a "bell curve" and in fact such a random distribution is used to approximate the distribution of human physical properties in some role-playing games:



Adding more and more dice produces histograms that look more and more like a normal distribution:







5 Summary of Properties of Sums of Random Variables

Property	When is it true?
$E(T_o) = \sum_{i=1}^{n} E(X_i)$	Always
$V(T_o) = \sum_{i=1}^{n} V(X_i)$	When $\{X_i\}$ independent
T_o normally distributed	Exact, when $\{X_i\}$ normally distributed
	Approximate, when $n \gtrsim 30$ (Central Limit Theorem)

Practice Problems

5.55, 5.57, 5.61, 5.65, 5.67, 5.89