In general, any procedure that seeks to determine whether a set of data could reasonably have originated from some given probability distribution, or *class* of probability distributions, is called a *goodness-of-fit* test. The principle behind the particular goodness-of-fit test we will look at is very straightforward: First the observed data are grouped, more or less arbitrarily, into k classes; then each class's "expected" occupancy is calculated on the basis of the presumed model. If it should happen that the set of observed and expected frequencies shows considerably more disagreement than sampling variability would predict, our conclusion will be that the supposed $p_X(k)$ or $f_Y(y)$ was incorrect.

In practice, goodness-of-fit tests have several variants, depending on the specificity of the null hypothesis. Section 10.3 describes the approach to take when both the form of the presumed data model and the values of its parameters are known. More typically, we know the form of $p_X(k)$ or $f_Y(y)$, but their parameters need to be estimated; these are taken up in Section 10.4.

A somewhat different application of goodness-of-fit testing is the focus of Section 10.5. There, the null hypothesis is that two random variables are *independent.* In more than a few fields of endeavor, tests for independence are among the most frequently used of all inference procedures.

10.2 The Multinomial Distribution

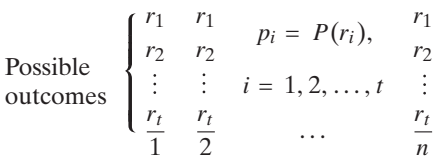
Their diversity notwithstanding, most goodness-of-fit tests are based on essentially the same statistic, one that has an asymptotic chi square distribution. The underlying structure of that statistic, though, derives from the *multinomial distribution,* a direct extension of the familiar *binomial*. In this section we define the multinomial and state those of its properties that relate to goodness-of-fit testing.

Given a series of n independent Bernoulli trials, each with success probability p, we know that the pdf for X, the total number of successes, is

$$P(X=k)=p_X(k)=\binom{n}{k}p^k(1-p)^{n-k}, \quad k=0,1,\ldots,n \tag{10.2.1}$$

One of the obvious ways to generalize Equation 10.2.1 is to consider situations in which at each trial, one of t outcomes can occur, rather than just one of two. That is, we will assume that each trial will result in one of the outcomes $r_1, r_2, \ldots, r_t$, where $p(r_i)=p_i, i=1,2,\ldots,t$ (see Figure 10.2.1). It follows, of course, that $\sum_{i=1}^{t} p_i = 1$.

Figure 10.2.1



In the binomial model, the two possible outcomes are denoted s and f, where $P(s)=p$ and $P(f)=1-p$. Moreover, the outcomes of the n trials can be nicely summarized with a single random variable X, where X denotes the number of successes. In the more general multinomial model, we will need a random variable to count the number of times that *each* of the r_i's occurs. To that end, we define

$$X_i = \text{number of times } r_i \text{ occurs}, \quad i = 1, 2, \ldots, t$$

For a given set of n trials, $X_1 = k_1, X_2 = k_2, \ldots, X_t = k_t$ and $\sum_{i=1}^{t} k_i = n$.

Theorem 10.2.1 *Let X_i denote the number of times that the outcome r_i occurs, $i = 1, 2, \ldots, t$, in a series of n independent trials, where $p_i = P(r_i)$. Then the vector $(X_1, X_2, \ldots, X_t)$ has a multinomial distribution and*

$$p_{X_1, X_2, \ldots, X_t}(k_1, k_2, \ldots, k_t) = P(X_1 = k_1, X_2 = k_2, \ldots, X_t = k_t)$$
$$= \frac{n!}{k_1! k_2! \cdots k_t!} p_1^{k_1} p_2^{k_2} \cdots p_t^{k_t},$$
$$k_i = 0, 1, \ldots, n; \quad i = 1, 2, \ldots, t; \quad \sum_{i=1}^{t} k_i = n$$

Proof Any particular sequence of k_1 r_1's, k_2 r_2's, ..., and k_t r_t's has probability $p_1^{k_1} p_2^{k_2} \ldots p_t^{k_t}$. Moreover, the total number of outcome sequences that will generate the values $(k_1, k_2, \ldots, k_t)$ is the number of ways to permute n objects, k_1 of one type, k_2 of a second type, ..., and k_t of a tth type. By Theorem 2.6.2 that number is $n!/k_1!k_2! \ldots k_t!$, and the statement of the theorem follows. □

Depending on the context, the r_i's associated with the n trials in Figure 10.2.1 can be either single numerical values (or categories) or ranges of numerical values (or categories). Example 10.2.1 illustrates the first type; Example 10.2.2, the second. The only requirements imposed on the r_i's are (1) they must span all of the outcomes possible at a given trial and (2) they must be mutually exclusive.

Example 10.2.1 Suppose a loaded die is tossed twelve times, where

$$p_i = P(\text{Face } i \text{ appears}) = ci, \quad i = 1, 2, \ldots, 6$$

What is the probability that each face will appear exactly twice?

Note that

$$\sum_{i=1}^{6} p_i = 1 = \sum_{i=1}^{6} ci = c \cdot \frac{6(6+1)}{2}$$

which implies that $c = \frac{1}{21}$ (and $p_i = i/21$). In the terminology of Theorem 10.2.1, the possible outcomes at each trial are the $t = 6$ faces, 1 ($= r_1$) through 6 ($= r_6$), and X_i is the number of times face i occurs, $i = 1, 2, \ldots, 6$.

The question is asking for the probability of the vector

$$(X_1, X_2, X_3, X_4, X_5, X_6) = (2, 2, 2, 2, 2, 2)$$

According to Theorem 10.2.1,

$$P(X_1 = 2, X_2 = 2, \ldots, X_6 = 2) = \frac{12!}{2!\,2! \cdots 2!} \left(\frac{1}{21}\right)^2 \left(\frac{2}{21}\right)^2 \cdots \left(\frac{6}{21}\right)^2$$
$$= 0.0005$$

■

Example 10.2.2

Five observations are drawn at random from the pdf

$$f_Y(y) = 6y(1-y), \quad 0 \le y \le 1$$

What is the probability that one of the observations lies in the interval [0, 0.25), none in the interval [0.25, 0.50), three in the interval [0.50, 0.75), and one in the interval [0.75, 1.00]?

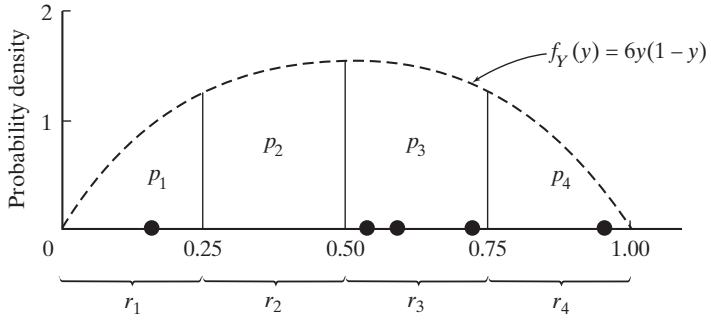


Figure 10.2.2

Figure 10.2.2 shows the pdf being sampled, together with the *ranges* r_1, r_2, r_3, and r_4, and the intended disposition of the five data points. The p_i's of Theorem 10.2.1 are now areas. Integrating $f_Y(y)$ from 0 to 0.25, for example, gives:

$$\begin{aligned} p_1 &= \int_0^{0.25} 6y(1-y)\,dy \\ &= 3y^2 \Big|_0^{0.25} - 2y^3 \Big|_0^{0.25} \\ &= \frac{5}{32} \end{aligned}$$

By symmetry, $p_4 = \frac{5}{32}$. Moreover, since the area under $f_Y(y)$ equals 1,

$$p_2 = p_3 = \frac{1}{2}\left(1 - \frac{10}{32}\right) = \frac{11}{32}$$

Let X_i denote the number of observations that fall into the ith range, $i = 1, 2, 3, 4$. The probability associated with the multinomial vector (1, 0, 3, 1), then, is *0.0198*:

$$\begin{aligned} P(X_1 = 1, X_2 = 0, X_3 = 3, X_4 = 1) &= \frac{5!}{1!\,0!\,3!\,1!}\left(\frac{5}{32}\right)^1\left(\frac{11}{32}\right)^0\left(\frac{11}{32}\right)^3\left(\frac{5}{32}\right)^1 \\ &= 0.0198 \end{aligned}$$

■

A Multinomial/Binomial Relationship

Since the multinomial pdf is conceptually a straightforward generalization of the binomial pdf, it should come as no surprise that each X_i in a multinomial vector is, itself, a binomial random variable.

Theorem 10.2.2

Suppose the vector $(X_1, X_2, \ldots, X_t)$ is a multinomial random variable with parameters $n, p_1, p_2, \ldots$, and p_t. Then the marginal distribution of $X_i, i = 1, 2, \ldots, t$, is the binomial pdf with parameters n and p_i.

Proof To deduce the pdf for X_i we need simply to dichotomize the possible outcomes at each of the trials into "r_i" and "not r_i." Then X_i becomes, in effect, the number of "successes" in n independent Bernoulli trials, where the probability of success at any given trial is p_i. By Theorem 3.2.1, it follows that X_i is a binomial random variable with parameters n and p_i. □

Comment Theorem 10.2.2 gives the pdf for any given X_i in a multinomial vector. Since that pdf is the binomial, we also know that the mean and variance of each X_i are $E(X_i) = np_i$ and $\text{Var}(X_i) = np_i(1 - p_i)$, respectively.

Example 10.2.3

A physics professor has just given an exam to fifty students enrolled in a thermodynamics class. From past experience, she has reason to believe that the scores will be normally distributed with $\mu = 80.0$ and $\sigma = 5.0$. Students scoring ninety or above will receive A's, between eighty and eighty-nine, B's, and so on. What are the expected values and variances for the numbers of students receiving each of the five letter grades?

Let Y denote the score a student earns on the exam, and let r_1, r_2, r_3, r_4, and r_5 denote the ranges corresponding to the letter grades A, B, C, D, and F, respectively. Then

$$\begin{aligned} p_1 &= P(\text{Student earns an A}) \\ &= P(90 \le Y \le 100) \\ &= P\left(\frac{90-80}{5} \le \frac{Y-80}{5} \le \frac{100-80}{5}\right) \\ &= P(2.00 \le Z \le 4.00) \\ &= 0.0228 \end{aligned}$$

If X_1 is the number of A's that are earned,

$$E(X_1) = np_1 = 50(0.0228) = 1.14$$

and

$$\text{Var}(X_1) = np_1(1 - p_1) = 50(0.0228)(0.9772) = 1.11$$

Table 10.2.1 lists the means and variances for all the X_i's. Each is an illustration of the Comment following Theorem 10.2.2.

Table 10.2.1

Score	Grade	p_i	$E(X_i)$	$\text{Var}(X_i)$
$90 \le Y \le 100$	A	0.0228	1.14	1.11
$80 \le Y < 90$	B	0.4772	23.86	12.47
$70 \le Y < 80$	C	0.4772	23.86	12.47
$60 \le Y < 70$	D	0.0228	1.14	1.11
$Y < 60$	F	0.0000	0.00	0.00

■