

APPENDIX

PROBABILITY AND HYPOTHESIS TESTING IN BIOLOGY

ESTIMATING PROBABILITY

In crossing his pea plants, Mendel obtained genetic ratios that were in excellent agreement with his model of factor segregation. Other workers soon obtained comparable results. The favorable agreement of data with theory reflected a basic property of Mendel's model: if Mendelian factors are considered independent, then the probability of observing any one factor segregating among progeny must simply reflect its "frequency," the proportion with which it occurs among the gametes. Frequencies are probabilities seen in the flesh. Mendel clearly understood this, and it was for this reason he sought large sample sizes. If you are flipping a coin, and you *know* the probability of "heads" to be $1/2$, the way to get your observed frequency of "heads" to approximate the expected value most reliably is to flip the coin many times.

But what if you are a human, raising a family? Families of several hundred children are not common. When one has only four children, the children may not exhibit a Mendelian ratio, just because of random chance. Mendel could not have deduced his model working with family sizes of four.

However, current geneticists are in a more fortunate position than was Mendel. Thanks to his work, and a large amount of subsequent investigation, we now have in hand reliable models of segregation behavior—we know what to expect. In a cross between two heterozygotes (Aa) we expect a 3:1 phenotypic ratio, dominant to recessive, among the progeny. That is to say, possessing a model of Mendelian segregation, *we know what the probabilities are*. In our cross, each individual among the progeny has a $1/4$ probability of being homozygous recessive (aa) and showing the recessive trait. Because we know the explicit probabilities of Mendel's segregation model, we can make ready predictions about what segregation patterns to expect in families of small size. Imagine, for instance, that you choose to have three children. What are the odds that you will have a boy, a girl, and a boy, in that order? The probability of the first child being a boy is $1/2$. When the second child comes, its sex does *not* depend on what happened before, and the probability of it being a girl is also $1/2$. Similarly, the probability of a male third child is $1/2$. Because the three children represent *independent* Mendelian events, simple probability theory applies: "the probability of two independent events occurring together is equal to the product of their individual properties." In this case, the probability $P = 1/2 \times 1/2 \times 1/2 = 1/8$. It is just this process we use in employing Punnett squares. Of course, P need not equal $1/2$. If one asks what is the probability that two parents heterozygous for albinism will produce one normal, one albino, and one normal child, in that order, $P = 3/4 \times 1/4 \times 3/4 = 9/64$.

The principal difficulty in applying a known model to any particular situation is to include all the possibilities in one's estimate. For instance, what if one had said above, "what is the probability P of obtaining two male children and one female child in a family of three?" In this case, the order is *not* specified, and so the three births cannot be considered independently. Imagine, for example, that the first two births turn out to be boys. The answer to the question is $P = 1/2!$ P in this case is a *conditional probability*. When there is more than one way in which an event can occur, each alternative must be taken into account. What one does is calculate the probability of each alternative, and then sum them up. Estimating the probability that two of three children will be male, there are three ways that this can occur: F, M, M; M, F, M; and M, M, F. Summing the probabilities gives us:

$$P = (1/2 \times 1/2 \times 1/2) + (1/2 \times 1/2 \times 1/2) + (1/2 \times 1/2 \times 1/2)$$

or

$$P = 3(1/8) = 3/8$$

In the case of parents heterozygous for albinism, the probability of one albino child in three is calculated similarly:

$$P = 3(9/64) = 27/64$$

BINOMIAL DISTRIBUTIONS

As you can see, things are rapidly getting out of hand, and we have only been considering families with three children. Fortunately, it is possible to shorten the analysis considerably. Hidden within the pattern above is one of the greatest simplicities of mathematics. Let's go back and reexamine the example of one girl in three births. Let the probability of obtaining a boy at any given birth be p , and the probability of obtaining a girl be q . We can now describe all the possibilities for this family of three:

| <u>Composition of family</u> | <u>Order of birth</u> | <u>Calculation</u> | <u>Probability</u> |
|------------------------------|-----------------------|---------------------|---|
| 3 boys | ♂ ♂ ♂ | $p \cdot p \cdot p$ | p^3 |
| 2 boys and 1 girl | ♀ ♂ ♂ | $q \cdot p \cdot p$ | $\left. \begin{array}{l} p^2q \\ p^2q \\ p^2q \end{array} \right\} 3p^2q$ |
| | ♂ ♀ ♂ | $p \cdot q \cdot p$ | |
| | ♂ ♂ ♀ | $p \cdot p \cdot q$ | |
| 1 boy and 2 girls | ♀ ♀ ♂ | $q \cdot q \cdot p$ | $\left. \begin{array}{l} pq^2 \\ pq^2 \\ pq^2 \end{array} \right\} 3pq^2$ |
| | ♀ ♂ ♀ | $q \cdot p \cdot q$ | |
| | ♂ ♀ ♀ | $p \cdot q \cdot q$ | |
| 3 girls | ♀ ♀ ♀ | $q \cdot q \cdot q$ | q^3 |

Because these are all the possibilities (two objects taken three at a time = $2^3 = 8$), the sum of them must equal unity, or 1. Therefore we can state, for families of three, a general rule for two-alternative traits:

$$P = p^3 + 3 p^2q + 3 pq^2 + q^3$$

This will be true whatever the trait. To estimate the probability of two boys and one girl, with $p = 1/2$ and $q = 1/2$, one calculates that $3 p^2q = 3/8$. To estimate the probability of one albino in three from heterozygous parents, $p = 3/4$, $q = 1/4$, so that $3 p^2q = 27/64$.

This is where the great simplification comes in. $p^3 + 3 p^2q + 3 pq^2 + q^3$ is known as a binomial series. It represents the result of raising (expanding) the sum of two factors (a binomial) to a power, n . Simply said, $p^3 + 3 p^2q + 3 pq^2 + q^3 = (p + q)^3$. The reason we find this power series nested within Mendelian segregation derives again from the Mendelian models of segregation that we are using: independent events have multiplicative probabilities. For two alternative phenotypes, p and q , and three segregated events, $n = 3$, it will always be true under Mendel's model that the segregational possibilities may be described as $(p + q)^3$. And this will be true for any value of n . The expansion is a natural consequence of the basic assumption of independence.

Binomial expansions have distinct mathematical properties. Consider the values of n from 1 to 6:

| <u>n</u> | <u>Binomial</u> | <u>Expanded binomial</u> |
|----------|-----------------|--|
| 1 | $(A + B)$ | $a + b$ |
| 2 | $(A + B)^2$ | $a^2 + 2ab + b^2$ |
| 3 | $(A + B)^3$ | $a^3 + 3a^2b + 3ab^2 + b^3$ |
| 4 | $(A + B)^4$ | $a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$ |
| 5 | $(A + B)^5$ | $a^5 + 5a^4b + 10a^3b^2 + 10a^2b^3 + 5ab^4 + b^5$ |
| 6 | $(A + B)^6$ | $a^6 + 6a^5b + 15a^4b^2 + 20a^3b^3 + 15a^2b^4 + 6ab^5 + b^6$ |

- The expanded binomial always has $n + 1$ terms.
- For each term, the sum of the exponents = n .
- For each expansion, the sum of the coefficients = the number of possible combinations.
- If the numerical coefficient of any term is multiplied by the exponent of a in that term, then divided by the number (position) of the term in the series, the result is the coefficient of the next following term.
- The coefficients form a symmetrical distribution (Pascal's magic triangle): the coefficient of any term is the sum of the two coefficients to either side on the line above.

Now it is easy to do calculations of probabilities for small-sized families. Just select the appropriate term of the indicated binomial. The probability of four boys and one girl in a family of five is $5a^4b$, or $5(1/2)^4(1/2)$, or $5/32$. The probability of three albinos from heterozygous parents in a family of five is $10a^2b^3$, or $10(3/4)^2(1/4)^3$, or $45/512$.

The binomial series does not always have to be expanded to find the term of interest. Because of the symmetry implicit in the "magic triangle," one can calculate the numerical value for the coefficient of any term directly:

$$\text{The coefficient of } (a)^x (b)^{N-x} = \frac{N!}{X! (N-X)!}$$

For any binomial term, the two exponents add up to N , so if a 's exponent is X , then b 's exponent must be $(N - X)$. The exclamation mark is a particularly appropriate symbol: $N!$ is read as "N factorial," and stands for the product of n and all smaller whole numbers (thus $13! = (13)(12)(11)(10)(9)(8)(7)(6)(5)(4)(3)(2)(1)$). So to calculate the probability of three albino children from heterozygous parents in a family of five, the exponent is first calculated:

$$\text{The exponent of } (a)^2(b)^3 = \frac{5!}{2!3!}$$

or

$$\frac{(5)(4)(3)(2)(1)}{(2)(1)(3)(2)(1)} = 10$$

The appropriate term is therefore $10a^2b^3$, and the probability is, as before:

$$10(3/4)^2(1/4)^3, \text{ or } 45/512.$$

What if a cross has three progeny phenotypes, or four? The same sort of reasoning applies as for two: the expansion is now a multinomial. For a trinomial (imagine lack of dominance in a trait A , and a cross of $Aa \times Aa$ —you would expect a phenotypic ratio of 1:2:1, $AA:Aa:aa$), the appropriate expansion is

$(p + q + r)^n$. To calculate a particular trinomial expansion, one proceeds in a fashion analogous to the binomial:

$$\frac{N!}{w!x!y!} p^w q^x r^y$$

Here, w , x , and y are the numbers of offspring in each class, with the probabilities p , q , and r , respectively. Thus the probability of getting exactly 1AA, 2Aa, and 1aa among a total of four progeny is:

$$\frac{4!}{1!2!1!} (1/4)^1 (1/2)^2 (1/4)^1 = 3/16$$

EXPECTED RESULTS VS. OBSERVED RESULTS

So far we have been concerned with predicting the results of a cross, given a certain expectation based upon the Mendelian model of segregation. How do we compare the results we actually obtain with expectation? At what point does an observed ratio no longer “fit” the Mendelian prediction? Making such decisions is an essential element of genetic analysis. Most of the reason why we study patterns of inheritance is that deviations from Mendelian proportions often reveal the action of some other factor operating to change what we see.

The most important aspect of “testing hypotheses” by comparing expectation with observation is so obvious it is often overlooked. It, however, lies at the heart of most statistical problems in data interpretation. It is this: one cannot test a hypothesis without explicitly knowing the expected result. If one flips a coin six times, what is the expected result? Do you see the difficulty? There is no simple answer to the question because it is too vaguely worded. The most likely result is three heads, three tails (the *maximum likelihood expectation* or *epsilon* (ϵ), but it would not be unreasonable to get two heads and four tails. Every now and then you would even get six heads! The point is that there is a spectrum of possible outcomes distributed around the most likely result. Any test of a hypothesis and any decisions about the goodness-of-fit of data to prediction must take this spectrum into account. A coin-flipping model does not predict three heads and three tails, but rather a distribution of possible results due to random error, around $\epsilon = 3$ and 3. A hypothesis cannot be tested without knowing the underlying distribution.

What then about Mendelian segregation? What is the expected distribution in a Mendelian cross? Go back and look at the “magic triangle” of expanded binomials, and you will see the answer. The answer lies in the coefficients. They represent the frequency of particular results, and the spectrum of coefficients is the distribution of probabilities. For the example of flipping a coin six times, $\epsilon = 3$ and 3 and the probability distribution is 1:6:15:20:15:6:1. The probability of ϵ , of getting precisely three heads and three tails, is $20/(\epsilon$ coefficients) or $20/64$. But all of the other possibilities have their probabilities as well, and each must be taken into account in assessing results. In this case the probability is $44/64$ that you will not get exactly three heads and three tails. Would you reject the hypothesis of 50:50 probability heads vs. tails because of such a result? Certainly you should not.

So how does one characterize the expected distribution? Look at the behavior of the probability spectrum as you flip the coin more and more times:

flips

1 1 + 1
2 1 + 2 + 1
3 1 + 3 + 3 + 1
4 1 + 4 + 6 + 4 + 1
5 1 + 5 + 10 + 10 + 5 + 1
6 1 + 6 + 15 + 20 + 15 + 6 + 1
7 1 + 7 + 21 + 35 + 35 + 21 + 7 + 1
8 1 + 8 + 28 + 56 + 70 + 56 + 28 + 8 + 1
9 1 + 9 + 36 + 84 + 126 + 126 + 84 + 36 + 9 + 1
10 1 + 10 + 45 + 120 + 210 + 252 + 210 + 120 + 45 + 10 + 1

As the coin is flipped more and more times, the results increasingly come to fit a smooth curve! Because in this case $a = b$ (probability of heads and tails is equal), the curve is symmetrical. Such a random-probability curve is known as a random or *normal* distribution. Note that as n increases, $P(\)$ actually goes *down*. Do you see why?

To test a hypothesis, replicate experiments are analyzed and the distribution of results obtained are compared to the distribution of results originally expected.

THE NORMAL DISTRIBUTION

In comparing experimental data, with prediction, our first task is to ascertain the nature of the underlying distribution. We have seen that the binomial distribution generates a bell-shaped distribution of possibilities centered around the most likely result. Many genetic characteristics have been found to fit this same *normal curve* (height or weight in humans, for example). In general, any property varying at random will also exhibit a “normal” distribution. Thus, experimental errors, when not due to some underlying systematic bias, are expected to be normally distributed.

The likelihood that a given data set fits a normal distribution may be characterized in terms of four simple *statistics*:

1. *Mean*. The arithmetic mean, or average value, is the most useful general measure of central tendency. It is defined as:

$$\bar{X} = \frac{\sum X_i}{N}$$

or the sum () of the individual measurements (X_i) divided by the number of measurements. For normal distributions, the mean value equals the *mode*, the value that occurs at highest frequency (e.g., X will =).

2. *Variation*. The degree to which data are clustered around the mean is usually estimated as *the standard deviation*, sigma (). For continuously varying traits such as height, is defined as the square root of the mean of the squared deviations:

$$\sigma = \sqrt{\frac{\sum (X_i - \bar{X})^2}{N - 1}}$$

The factor ($N - 1$) is used rather than N as a correction because the data are only an estimate of the entire sample. When sample sizes are large, N may be used instead. The square of the standard deviation has particular significance in statistics and is called the *variance*. The variance is a

particularly useful statistic because variances are additive. If one source of error contributes a certain amount of variance to the data, and another source of error contributes an additional amount, then the total variance seen in the data is equal to the sum of the individual error contributions. By partitioning variance, one may assess particular contributions to experimental error.

For discontinuous traits like albinism, such a definition has no meaning (what is the “mean” of a 3:1 segregation pattern?), and the standard deviation is defined instead in terms of the frequencies of alternative alleles:

$$\sigma = \sqrt{\frac{pq}{N}}$$

For normally distributed data, 68 percent of the data lie within one standard deviation of the mean, 95 percent within two standard deviations, and 99 percent within three standard deviations.

3. *Symmetry*. Lack of symmetry, or *skew*, is usually measured as a third order statistic (standard deviation was calculated in terms of σ^2 , the square), as the average of the cubed deviations from the mean divided by the cube of the standard deviation:

$$\alpha_3 = \frac{\frac{1}{N} \left[\sum (X_i - \bar{X})^3 \right]}{\sigma^3}$$

For a symmetrical distribution, $\alpha_3 = 0$. It is important to know whether or not a particular set of data has a symmetrical distribution in attempting to select the proper statistical distribution with which to compare it.

4. *Peakedness*. The degree to which data are clustered about the mean, *kurtosis*, is measured by a fourth order statistic, the mean of the fourth powers of the deviations from the mean divided by the fourth power of the standard deviation:

$$\alpha_4 = \frac{\frac{1}{N} \left[\sum (X_i - \bar{X})^4 \right]}{\sigma^4}$$

For a normal distribution, α_4 is always equal to 3. Values greater than 3 indicate a more peaked distribution (*leptokurtic*), while values less than 3 indicate a flatter distribution (*platykurtic*).

THE *t* DISTRIBUTION

It is important to understand that the distribution within a data set will not always resemble that of the population from which the sample was taken, even when the overall population has a normal distribution. Even though 95 percent of the individuals of a real population may fall within ± 2 standard deviations of the mean, the actual sample may deviate from this figure due to the effects of small sample size.

When N is less than 20, a family of statistics is usually employed that takes the effect of small population size into account. The standard deviation is corrected for sample size as the *standard error*, s , which is basically an estimate of the degree to which sample mean approximates overall mean:

$$\bar{s} = \frac{\sigma}{\sqrt{N}}$$

Data of a small sample may be related to the overall population in terms of the difference of their two means, divided by the standard error:

$$t = \frac{\bar{X} - \mu}{\bar{s}}$$

thus, solving the equation for mu (the real mean of the overall population), the real mean equals the estimated mean (\bar{X}) +/- the factor ts :

$$\mu = \bar{X} \pm t\bar{s}$$

t measures the deviation from the normal distribution attributable to sample size. t has its own distribution, which is fully known. One may thus inquire, for any experimental data (especially of small sample size), whether the variability in the data is or is not greater than predicted by the t distribution. Imagine, for example, a data set concerning adult human height in inches:

Individual height

60

65

66

68

68

69

69

70

71

74

$$N = 10$$

$$\bar{X} = 68$$

$$\sigma = 3.77$$

$$\bar{s} = 1.19$$

(for $N = 10$ and $p = .95$,
we see that $t = 2.228$)

The t distribution tells us that 95 percent of all estimates would be expected to exhibit a mean of mu equal to $\bar{X} \pm ts$. In this case, $\mu = 68 \pm (2.228)(1.19)$, or 68 ± 2.65 . Thus, 95 percent of all estimations of mean height would be expected to fall within the range of 65 to 71. The probability that the two values falling outside of this range represent the same underlying distribution (belong to the same cluster of points) is less than 5 percent.

THE POISSON DISTRIBUTION

Recall that the binomial expansion $(p + q)^n$ yields a symmetrical distribution only when $p = q$ (as was the case for flipping coins, when the probabilities of heads and tails were equal). Often, however, the probabilities of two alternatives are not equal, as in the case of our example of albinism, where $p = 3/4$. In this case, the proper binomial expansion is $(3/4 + 1/4)^2$, and the three possible genotypes are in the proportions $1(3/4)(3/4) + 2(3/4)(1/4) + 1(1/4)(1/4)$ or 0.56 AA; 0.37 Aa; 0.06 aa, a very lopsided distribution. The skew reflects the numerical difference between the values of p and q .

For data where p and q represent the frequencies of alternative alleles, the deviation from symmetry can be very significant, although it is minimized by large sample sizes (n). When the difference in the two frequencies is so great that one of them is of the order $1/n$, then the various combinations of p and q will exhibit an extremely skewed distribution, the *Poisson distribution*.

The Poisson distribution, like the t distribution, is known explicitly. It is possible, for any data set, to compare “observed” with “expected.” One generates the “expected” result by multiplying sample sizes by the probability that the Poisson distribution indicates for each class:

$$\text{Poisson probability} = e^{-m} \left(\sum 1, m, \frac{m^2}{2!}, \frac{m^3}{3!}, \dots, \frac{m^i}{i!} \right)$$

Because the Poisson distribution is known, one may look up values of e^{-m} (the natural log of the mean value of the distribution) and so calculate the expected probability of obtaining data in each of the classes m, m^2 , etc. Imagine, for instance, searching for rare enzyme variants in different populations of humans:

| ① # Variant Enzyme Types Observed | ② # Populations | ③ Total # Observations | ④ Poisson Probability | ⑤ Predicted # |
|---|--------------------|------------------------------|-----------------------------------|------------------|
| | | (① × ②) | | (④ × N) |
| 0 | 110 | 0 | (.712)(1) = .712 | 105 |
| 1 | 28 | 28 | (.712)(.34) = .242 | 35 |
| 2 | 6 | 12 | $(.712) \frac{(.34)^2}{2} = .041$ | 6 |
| 3 | 2 | 6 | $(.712) \frac{(.34)^3}{6} = .005$ | 1 |
| 3 | 1 | 4 | .0001 | 0 |
| | $N = 147$ | $\bar{X} = 50$ | | |

m , the average number of variants per population, is $50/147$, or 0.340 . Looking up this number in the Poisson distribution table (table of e^m), we obtain $e^m = 0.712$. Now substitute the values of m and e^m into the formula for expected probability to obtain the values predicted of the assumption of an underlying Poisson distribution.

The Poisson distribution has the property that its variance ($\sigma^2 = \mu$) is equal to its mean. In the above example, the mean should be taken for this purpose as the total sample observations, 50 . If one accepts these data as fitting a Poisson distribution, then σ^2 also = 50 , and $\sigma = 7.07$. For random errors (which are normally distributed), two variances encompass 95 percent of the estimates, so that the “true” mean of these data has a 95 percent chance of lying within $\pm 2(7.07)$ of 50 , or between 36 and 64 for a sample of 147 populations.

LEVEL OF SIGNIFICANCE

Knowledge of the underlying distribution permits the investigator to generate a hypothetical data set—data that would be predicted under the hypothesis being tested. The investigator is then in a position to compare the predicted data with the experimental data already obtained. How is this done? At what point is the similarity not good enough? If 50 progeny of a cross of rabbits heterozygous for albinism are examined, the expected values would be $(3/4 \times 50)$ normal: $(1/4 \times 50)$ albino, or $37:13$ normal:albino. What if the observed result is actually 33 normal and 17 albino? Is that good enough?

What is needed is an arbitrary criterion, some flat rule that, by convention, everybody accepts. Like table manners, there is no law of nature to govern behavior in judgments of similarity, just a commonly agreed-to criterion. The criterion is derived from the normal distribution, the one most often encountered in genetic data. Recall that the normal distribution has a shape such that ± 2 alpha encompasses 95 percent of the

data. Quite arbitrarily, that is taken as the critical point. Any data falling more than 2 alpha from the mean are taken as not representative of the mean. More generally, for any data set of whatever distribution, 95 percent confidence intervals are the criteria for hypothesis rejections. Less than 5 percent of the time is such a deviation from expectation predicted on the basis of chance alone.

THE CHI-SQUARE DISTRIBUTION

Now the results of the rabbit cross previously described can be assessed. We know the underlying distribution of Mendelian data is normally distributed, we have a set of experimental data and a corresponding set of predicted values, and we have a criterion for the desired goodness-of-fit of experiment to production.

What is the procedure? The most direct way is to generate the predicted probability distribution using the coefficients of the binomial expansion. This, however, is a rather formidable task, as the desired expansion is $(3/4 + 1/4)^{50}$!

To reduce problems of calculation, a different tack is usually taken. Rather than directly comparing the observed and expected distributions in a one-to-one fashion, the investigator compares a property of the distributions, one very sensitive to differences in underlying distribution shape. What is compared is the dependence of chance deviations on sample size. This dependence is estimated by the statistic X^2 , or *chi-squared*, which is defined as the sum of the mean square deviations:

$$X^2 = \sum \left[\frac{(X_{\text{obs.}} - X_{\text{predicted}})^2}{X_{\text{predicted}}} \right]$$

When sample sizes are small, or there are only two expected classes, X^2 is calculated as:

$$X^2 = \text{sum} \frac{(1(\text{observed \#}) - (\text{expected \#}) - 1/2)^2}{\text{expected \#}}$$

The reduction of the absolute value of the deviation by 1/2 is known as the Yates Correction, and is carried out when the number of any of the expected classes is less than ten, or, as we shall see, when there is only one degree of freedom (d. f. = the number of expected classes, 2 in this case, minus 1). Chi-square tests are normally not applied to any set of data containing a class with less than five members.

The distribution of the X^2 statistic is known explicitly. Calculating a value for X^2 , one can inquire whether a value as large as calculated would be expected on the basis of chance alone 5 percent of the time. If not, then by our arbitrary 95 percent level of significance, the deviation of observation from prediction is significant and the hypothesis used to generate the prediction is significant and the hypothesis used to generate the prediction should be rejected.

For the case of the rabbit cross discussed previously:

| | Dominant | Recessive | Total |
|-----------------------------|----------|-----------|---------------|
| Observed | 33 | 17 | 50 |
| Predicted | 37 | 13 | 50 |
| Δ | -4 | +4 | 0 |
| Δ^2 | 16 | 16 | |
| $\Delta^2/\text{predicted}$ | 0.432 | 1.231 | $X^2 = 1.663$ |

Note carefully that we use the raw data in calculating X^2 . This is because X^2 concerns itself with the dependence of deviations on sample size. When data are reduced to percentages, this normalizes them to sample size, removing the differences we are attempting to test and making the comparison meaningless. Always use real data in a X^2 test.

Now what is done with the X^2 value of 1.663? Before assessing its significance, we need to allow for the effect of different numbers of classes in the outcome. Because there are more chances for deviations when there are more classes, the predicted values of X^2 are greater when more classes are involved in the test. For this reason, X^2 tables are calculated completely for each *potentially varying* class number. This last point is particularly important: if there are four classes of offspring among a total of 100, and you observe 22, 41, and 17 for the first three classes, what sort of options are available for the members that may occur in the final class? None at all (it must contain 20). So, given that the total is fixed, there are only three potentially varying classes, or three *degrees of freedom*. Degrees of freedom are defined as the number of independently varying classes in the test. For X^2 tests, the degrees of freedom are $(n - 1)$, one less than the number of independent classes in the text.

We may now, at long last, assess the probability that our rabbit result could be so different from the expected 3:1 ratio due just to chance. For a X^2 of 1.663 and one degree of freedom, the probability is 21 percent that a deviation this great could result from chance alone. Thus we do not reject the hypothesis of a 3:1 segregation ratio based upon these data (the X^2 value would have to have been >3.84 for rejection). As you can see, the 5 percent rejection criterion is very conservative. Data must be very far from prediction before a hypothesis is rejected outright.

Note that failure to reject the 3:1 segregation ratio hypothesis for the rabbit data does not in any sense establish that this hypothesis is correct. It says only that the experiment provides no clear evidence for rejecting it. What about other alternatives? The data (33 dominant, 17 recessive) fit a 2:1 ratio very well indeed. Are we then free to choose the 2:1 segregation ratio hypothesis as the more likely? No. There is no *evidence* for rejecting the 3:1 ratio hypothesis. *Based on the data*, either hypothesis is tenable.

It isn't necessary to stop here, of course. The obvious thing to do in a situation like this is to go out and collect more data. With a sample size of 200 and the same proportion (135 dominant to 65 recessive), a clear choice is possible between the two hypotheses:

| | Hypothesis I (3:1 ratio) | | | Hypothesis II (2:1 ratio) | | |
|----------------------------------|-----------------------------|-----------|-------------|------------------------------|-----------|-------------|
| | Dominant | Recessive | Total | Dominant | Recessive | Total |
| Observed | 135 | 65 | 200 | 135 | 65 | 200 |
| Predicted | 150 | 50 | 200 | 133 | 67 | 200 |
| (Obs.-pred.) | -15 | 15 | 0 | 2 | -2 | 0 |
| (Obs.-pred.) ² | 225 | 225 | | 4 | 4 | |
| (Obs.-pred.) ² /pred. | 1.5 | 4.5 | $X^2 = 6.0$ | .03 | .06 | $X^2 = .09$ |

While the fit of hypothesis II (2:1 ratio) is very good (a greater than 70 percent chance that the deviation from prediction is due solely to chance), the fit of hypothesis I (3:1 ratio) is terrible (only a 1 percent chance that the deviation from the prediction of the 3:1 hypothesis is due to chance), far exceeding the 5 percent limits required for rejection. The investigator can now state that there is enough objective evidence for rejecting the hypothesis that the traits are segregating in a 3:1 ratio.

There is nothing magic about a 3:1 ratio, no reason why it must be observed. It represents chromosomal segregation behavior, while the investigator is observing realized physiological phenotypes. Perhaps in this case the homozygous dominant is a lethal combination:

| | | |
|---|--------|----|
| | A | a |
| A | (dies) | Aa |
| a | Aa | aa |

One would, in such a circumstance, predict just such a 2:1 segregation ratio. Employing statistical tests can never verify the validity of a hypothesis. They are properly employed to reject hypotheses that are clearly inconsistent with the observed data.

The application of X^2 tests of goodness-of-fit is not limited to data that are normally distributed. The Poisson-distributed data discussed previously could be compared to the values predicted by the Poisson distribution (column 2 vs. column 5) using a X^2 analysis, if there were at least five members in each class.

The X^2 test finds its most common application in analyzing the results of genetic crosses. In a Mendelian dihybrid cross, for example, a Mendelian model of segregation predicts a segregation ratio of 9:3:3:1. Actual data may be compared to the data one would have expected to obtain if the progeny indeed segregate in these proportions. Any deviation from prediction suggests that something is going on to alter the proportions we see, and thus can point the way to further investigation. In Mendel's dihybrid cross of yellow and wrinkled peas, the X^2 test is as follows:

| | Smooth Yellow | Smooth Green | Wrinkled Yellow | Wrinkled Green | Total |
|-----------------------------|---------------|--------------|-----------------|----------------|---------------|
| Observed | 315.0 | 108.0 | 101.0 | 32.0 | 556 |
| Predicted | 312.7 | 104.3 | 104.3 | 34.7 | 556 |
| Δ | +2.3 | +3.7 | -3.3 | -2.7 | 0 |
| Δ^2 | 5.29 | 13.69 | 10.89 | 7.29 | |
| $\Delta^2/\text{predicted}$ | .017 | 0.131 | 0.104 | 0.210 | $X^2 = 0.462$ |

$(n - 1) = 3$ degrees of freedom, so there is a greater than 90 percent probability that the deviation we see from prediction is due to chance. This is a very good fit of data to prediction.

By contrast, other traits in peas exhibit quite different behavior in dihybrid crosses:

| | Purple Long | Purple Round | Red Long | Red Round | Total |
|-----------------------------|-------------|--------------|----------|-----------|---------------|
| Observed | 296 | 19 | 27 | 87 | 429 |
| Predicted | 242 | 80 | 80 | 27 | 429 |
| Δ | +54 | -61 | -53 | +60 | 0 |
| Δ^2 | 2916 | 3721 | 2809 | 3600 | |
| $\Delta^2/\text{predicted}$ | 12.0 | 46.5 | 35.1 | 133.3 | $X^2 = 226.9$ |

Clearly the hypothesis that these dihybrid progeny are segregating in a 9:3:3:1 ratio should be rejected.

TESTING INDEPENDENT ASSORTMENT

Many situations arise in genetic analysis where the critical issue is whether or not genes are acting independently. An example is provided by dihybrid matings, which upon chi-square analysis prove to differ significantly in their segregation from 9:3:3:1. What conclusions can be drawn from this? The deviation may arise because at least one of the genes is not segregating in a Mendelian 3:1 ratio. An alternative possibility is that both genes are segregating normally, but not independently of each other. Such situations arise when genes are located close to one another on the chromosome. Such *linkage* can be detected by

what is known as a *contingency test*. The simplest of these 2×2 contingency tests, chi-squares distributed with one degree of freedom, allow the calculation of X^2 directly. The test has the important property that abnormal segregation of one (or both) of the genes does not affect the test for independent assortment. Even if one of the genes is not observed to segregate in a 3:1 fashion due to some sort of phenotypic interaction, the two genes might still be linked.

To examine two genes for linkage (or lack of independent assortment), the data are arrayed in 2×2 matrix, and marginal totals are examined.

| | <i>Y</i> | <i>y</i> | Total |
|----------|--------------|--------------|-------------------------|
| <i>X</i> | <i>a</i> | <i>b</i> | <i>a + b</i> |
| <i>x</i> | <i>c</i> | <i>d</i> | <i>c + d</i> |
| Total | <i>a + c</i> | <i>b + d</i> | <i>N(a + b + c + d)</i> |

$$X^2 = \frac{(lad - bcl - (1/2)N)^2 N}{(a + b)(a + c)(c + d)(b + d)}$$

The formula for X^2 looks complicated, but it is actually quite simple. Consider again the dihybrid cross in pea plants (which happens to be the first reported case of linkage, in 1908 by William Bateson and R. C. Punnett):

| | Obs. | Predicted on a Hypothesis of 9:3:3:1 |
|---------------|------|--------------------------------------|
| Purple, long | 296 | 242 |
| Purple, round | 19 | 80 |
| Red, long | 27 | 80 |
| Red, round | 87 | 27 |

Is the obvious deviation (recall earlier calculation of the X^2 as 226!) due to one of the genes segregating in a non-Mendelian manner, or is it due to a lack of independence in assortment? The test is as follows:

| | <i>P</i> | <i>p</i> | Totals |
|----------|------------------|-----------------|----------------|
| <i>L</i> | 296(<i>PL</i>) | 27(<i>pL</i>) | 323 |
| <i>l</i> | 19(<i>Pl</i>) | 87(<i>pl</i>) | 106 |
| Total | 315 | 114 | <i>N = 429</i> |

$$X^2 = \frac{([25752 - 513] - (1/2)429)^2 429}{(323)(315)(106)(114)} = 145.3$$

As this 2×2 contingency chi-square test has only one degree of freedom, the critical X^2 value at the 5 percent level is 3.84. The traits are clearly linked.

As an alternative to carrying out contingency analyses, one may investigate aberrant 9:3:3:1 segregations by conducting a further test cross of F_1 hybrid individuals back to the recessive parent. As all of the four genotypes may be scored unambiguously in a test cross, one simply uses a standard chi-square test of goodness-of-fit of the results to the predicted 1:1:1:1 ratio.

TESTING POOLED DATA FOR HOMOGENEITY

Another problem that often arises in genetic analysis is whether or not it is proper to “pool” different data sets. Imagine, for example, that data are being collected on the segregation of a trait in corn plants, and that the plant is growing on several different farms. Would the different locations have ecological differences that may affect segregation to different degrees? Is it fair to pool these data, or should each plot be analyzed separately? For that matter, what evidence is there to suggest that it is proper to pool the progeny from any two individual plants, even when growing next to one another?

The decision as to whether or not it is proper to pool several data sets is basically a judgment of whether the several data sets are homogeneous—whether they represent the same underlying distribution. To make a decision, a homogeneity test with a chi-square distribution is carried out. This test is carried out in four stages:

1. First, a standard chi-square analysis is performed on each of the individual data sets (the Yates correction is *not* used). In each case, the observed data are compared to the prediction based on the hypothesis being tested (such as a 3:1 Mendelian segregation).
2. The individual X^2 values are added together, and the degrees of freedom are also summed. This value is the *total chi-square*.
3. To estimate that component of the total chi-square due to statistical deviation from prediction, the *pooled chi-square* is calculated for the summed data of all samples. The degrees of freedom are $n - 1$, one less than the number of phenotypic classes. Again, the Yates correction is not used.
4. If there is no difference between the individual samples, then the two X^2 values calculated in steps 2 and 3 will be equal. If, however, the individual data sets are not homogenous, then step two's X^2 will be greater than step three's X^2 by that amount. So to estimate the homogeneity chi-square, subtract the pooled X^2 from the total X^2 . In parallel, subtract the “pooled” X^2 degrees of freedom from the “total” X^2 degrees of freedom. The value obtained, the homogeneity X^2 , with its associated degrees of freedom, is used to consult a X^2 table to determine whether this value of X^2 exceeds the 5 percent value for the indicated degrees of freedom. If it does, then this constitutes evidence that the data sets are heterogeneous and should not be pooled.