Statistics in Biology

Casey Dunn
Department of Ecology and Evolutionary Biology
Yale University, New Haven, CT USA
casey.dunn@yale.edu
https://dunnlab.org/

July 12, 2026

1 Descriptive statistics

Parameters describe the actual population we want to know about. In most cases, we don't know the real parameter values, and instead make estimates based on observations of samples drawn from the full population. These estimates are an approximation of the true population parameters, and much of statistics is concerned with figuring out how good parameter estimates are. This allows us to test hypotheses about the parameters, as well as compare estimates to each other.

1.1 Descriptive statistics for a single variable

Many times we just want to summarize the distribution of a single variable, which we will here call X.

The arithmetic mean μ, a parameter, is an average of all values in a population:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} \tag{1}$$

Where the total population has size N and the value of each item i in the population is X_i. It often serves as an expected value of a sample drawn from the population.

The sample arithmetic mean $\bar{X}$, an estimate, is the average of all values in a sample drawn from the population:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{2}$$

Where the sample size is n and the value of each item i in the sample is X_i. Here n (the sample size) is distinguished from N (the size of the whole population), and in general $n \leq N$. The formula is otherwise identical to that for the population mean: the mean is computed the same way whether averaging a sample or the whole population. As the sample size approaches the population size, $\bar{X}$ approaches μ.

The variance σ^2, a parameter, gives an indication of how far values in the population deviate from the mean:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N} \tag{3}$$

The estimate of variance s^2 is calculated from a sample of the population as:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} \tag{4}$$

The $n-1$ in the denominator of the estimate accounts for the sample being only a fraction of the population, which leads to an underestimation of the deviances from the mean. This underestimation decreases as the sample size grows. See Appendix 2 of Grafen and Hails (2002) for an excellent explanation of why $n-1$ results in an unbiased estimation of variance.

The standard deviation, which is in the same units as the original measure, is simply the square root of the variance:

$$\sigma = \sqrt{\sigma^2} \tag{5}$$

$$s = \sqrt{s^2} \tag{6}$$

1.2 Descriptive statistics for two variables

In the above examples, we were summarizing a single variable across a population or a sample drawn from the population. We often want to measure multiple variables, and describe potential associations between them.

If the variables X and Y are both measured for all individuals in a population, their covariance is:

$$cov(X,Y) = \frac{\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)}{N} \tag{7}$$

If the variables X and Y are both measured for individuals in a sample, their covariance is:

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} \tag{8}$$

The covariance gives an indication of the degree to which the values of two variables are associated. Variance is a special case of covariance–it is the covariance of a variable with itself. Covariance can be positive (an increase in the value of X is associated with an increase in the value of Y), 0 (there is no association between X and Y), or negative (an increase in the value of X is associated with a decrease in the value of Y). There is no limit to their magnitude, they can range from $-\infty$ to ∞.

Because the different variables may have different magnitudes and units, a covariance alone can be difficult to interpret. One way around this is to normalize the covariance by the standard deviation of each variable. This results in a unit-less correlation coefficient. For a population, the correlation coefficient ρ is:

$$\rho = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{9}$$

Note that population size N also cancels out in the correlation coefficient, since the covariance has an N in the denominator and the standard deviations each have $\sqrt{N}$ in their denominator.

For a sample, the estimate r of the correlation coefficient is:

$$r = \frac{cov(X,Y)}{s_X s_Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{1}{n-1}\sum_{i=1}^{n}(\frac{X_i - \bar{X}}{s_X})(\frac{Y_i - \bar{Y}}{s_Y}) \tag{10}$$

The correlation coefficient has a few nice properties that make it easier than the covariance to interpret when you don't know much else about the sample. For one, it varies from -1 to 1, where 1 is perfect positive association between the variables and -1 is perfect negative association between the variables. Independent variables will be uncorrelated. Dependent variables, however, may also be uncorrelated, since correlation only measures linear dependence.

Another quantity that is closely related to the covariance and correlation coefficient is the least-squares linear regression, β. Whereas covariance and correlation describe association between variables that are not treated differently from each other in any way, linear regression describes the expected change in one variable relative to the change in the other. The expected change in Y given X is described by $\beta_{Y,X}$, while the expected change in X given Y is $\beta_{X,Y}$.

If Y values are plotted against their corresponding X values, $\beta_{Y,X}$ is the slope of the line drawn through the points such that the squared Y distances

from the points to the line are minimized. This is often how the linear regression is calculated. But it can also be related to the covariance and variance in the following way:

$$\beta_{Y,X} = \frac{cov(X,Y)}{\sigma_X^2} \tag{11}$$

$$\beta_{X,Y} = \frac{cov(X,Y)}{\sigma_Y^2} \tag{12}$$

Note that the linear regressions are obtained in a very similar way as the correlation coefficient. The denominator of the linear regression contains the variance of the explanatory variable that is being used to predict the dependent variable, while the denominator of the correlation contains the product of the square roots of the variance (i.e., the standard deviation) of each variable. From this relationship, it can also be seen that:

$$cov(X,Y) = \beta_{Y,X}\sigma_X^2 = \beta_{X,Y}\sigma_Y^2 \tag{13}$$

β isn't going to be very helpful for predicting the value of the dependent variable in terms of the explanatory variable if there is no linear relationship between them. This gets to another important relationship between correlation and linear regression. The square of the correlation coefficient, r^2, provides an indication of how linear the relationship is. If $r^2 = 0$, then there is no linear relationship and β won't explain one variable in terms of the other. If $r^2 = 1$, then the relationship is perfectly linear and all variation in each variable can be explained by the variation in the other variable. More generally, r^2 is the proportion of the variance in one variable that is explained by a linear fit to the other. Equivalently, r^2 provides an indication of how close the data points are to the regression line, with higher values indicating that they are closer.

There is often considerable confusion when it comes to applying correlations versus linear regressions (Twomey and Kroll, 2008). In general, remember that correlation coefficients provide an indication of the association between the variables, while linear regression determines the linear relationship (if there is one) between the variables. Like correlation, covariance provides an indication of the association between variables, but is also impacted by the units and magnitude of the variables. If, for example, two variables are perfectly correlated ($r = 1$), but the magnitude of X varies much more than the magnitude of Y, then X will have a greater impact on the covariance.

1.3 Multivariate descriptive statistics

Covariances, correlations, and linear regressions can all be generalized to more than two variables. Let's begin with an examination of the covariance matrix. Take the column vector $\mathbf{X}$ of random scalar variables $X_1 ... X_n$. Each X_i has its own mean and variance. The covariance matrix Σ has entries i, j that describe the covariance between X_i and X_j:

$$\Sigma_{i,j} = cov(X_i, X_j) \tag{14}$$

The diagonal of the covariance matrix contains the variance of each element of $\mathbf{X}$, as the variance of a variable is its covariance with itself.

Since $cov(X_i, X_j) = cov(X_j, X_i)$, the covariance matrix Σ is symmetric.

The elements i, j of the correlation matrix can be readily derived from the covariance matrix as $\Sigma_{i,j}/(\sigma_i\sigma_j)$. The elements of the diagonal of the correlation matrix are all 1, since they are the variance of element i divided by the variance of element i. This makes sense since X_i is perfectly correlated with itself as long as its variance is nonzero.

The covariance matrix can also be used to obtain the simple linear regression of each variable against a particular variable X_k. Dividing column k of the covariance matrix by the variance σ_k^2 gives entries $\Sigma_{i,k}/\sigma_k^2 = cov(X_i, X_k)/\sigma_k^2$, each of which is the slope β_{X_i,X_k} of the simple regression of X_i on X_k. Diagonal element k will be 1, since it is σ_k^2 divided by σ_k^2. This makes sense since linear regression of X_k against itself has slope 1. Note that these are the pairwise regressions of each variable against X_k alone, *not* the multiple regression of X_k against all the other variables jointly. The multiple regression coefficients are instead $\Sigma_{XX}^{-1}\Sigma_{Xk}$, where Σ_{XX} is the covariance matrix of the explanatory variables and Σ_{Xk} is the vector of their covariances with X_k; obtaining them requires inverting the covariance matrix of the predictors rather than dividing by a single variance.

One of the most popular summaries of multivariate data is the Principal Component Analysis (PCA), which is an eigendecomposition of the covariance (or correlation) matrix.

2 Probability distributions

A probability density function indicates the relative likelihood of a variable taking on a particular value. The probability of the variable taking on a value in a particular range is given by the integral of the probability density function over that range. The total area under the probability density function is 1, i.e. it encompasses all potential outcomes.

All of the probability density functions presented here, with the exception of the uniform, t, and F distributions, are exponential family distributions. This class of distributions is central to many aspects of statistics, and you will see them in many different settings.

These distributions are used in a variety of distinct ways, and it is always important to have a clear idea of how they are being applied. Distinct types of applications include:

- A description of the distribution of actual data (such as the use of a Poisson distribution to model the number of jellyfish collected in multiple net trawls).

- Based on first principles, we expect particular statistics to fit the distribution (such as the use of the normal distribution to explain the distribution of means of repeated samples from a population).
- They have a convenient shape (such as the use of a particular distribution as a Bayesian prior, even if there is no biological reason that the data should have that particular relationship between independent and dependent variables).

2.1 Normal distribution

This distribution is extremely important both because it fits the observed values of many biological data, and also because it approximates the expected mean values of samples drawn from *any* distribution (see the central limit theorem below).

$$f(X;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(X-\mu)^2}{2\sigma^2}} \tag{15}$$

To facilitate work with the normal distribution, values are often Z-standardized before analysis:

$$Z_i = \frac{X_i - \mu}{\sigma} \tag{16}$$

This centers the distribution so that the mean is 0 and scales it so that the variance is 1. This standardized normal distribution is sometimes called the z distribution.

2.2 Exponential distribution

Imagine an event that happens at rate λ, and let's define the sojourn time as the interval between adjacent events. The sojourn times will be distributed according to an exponential distribution:

$$f(t;\lambda) = \lambda e^{-\lambda t} \tag{17}$$

2.3 Gamma distribution

Imagine that we don't want to know the density of sojourn times, but instead the density of times until the event has occurred exactly k times. This is given by the Gamma distribution:

$$f(t;\lambda,k) = \frac{\lambda^k}{\Gamma(k)} t^{k-1} e^{-\lambda t} \tag{18}$$

Where $\Gamma(k)$ is the Gamma function, not to be confused with the Gamma distribution. For integers, $\Gamma(k) = (k-1)!$

2.4 Poisson distribution

Now imagine that you are going to make observations for a particular time T, and you want to know how many events k to expect in that interval. This is given by the Poisson distribution:

$$f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{19}$$

Note that here λ is the expected number of events over the interval, rather than the per-unit-time rate used above for the exponential and Gamma distributions. If events occur at rate λ_r over an interval of length T, then $\lambda = \lambda_r T$. The Poisson distribution has the very useful property that $\mu = \sigma^2 = \lambda$. If a sample has a variance greater than the mean ($s^2 > \bar{X}$), then the values are more clumped than would be expected under a Poisson distribution. If a sample has a variance less than the mean ($s^2 < \bar{X}$), then the values are more evenly dispersed than expected from a Poisson distribution.

2.5 Binomial distribution

When a measurement can only have two outcomes (e.g. success or failure), the binomial distribution describes the probability of k successes in n independent trials, given the probability p of success for any one trial:

$$f(k;p) = \binom{n}{k} p^k (1-p)^{n-k} \tag{20}$$

2.6 Negative binomial distribution

The negative binomial gives the distribution of the number k of successes that occur before a specified number of r failures, given the probability p of a success in any one trial:

$$f(k;r,p) = \binom{k+r-1}{k} p^k (1-p)^r \tag{21}$$

The mean and variance are:

$$\mu = \frac{pr}{1-p} \tag{22}$$

$$\sigma^2 = \frac{pr}{(1-p)^2} \tag{23}$$

The negative binomial is an appropriate alternative to the Poisson distribution when the sample is overdispersed, i.e. $s^2 > \bar{X}$. In fact, the variance is greater than the mean for a negative binomial for all non-zero values of p. This can be seen by rewriting the variance in terms of the mean:

$$\sigma^2 = \frac{pr}{(1-p)^2} = \frac{\mu}{1-p} \tag{24}$$

The negative binomial distribution converges on the Poisson distribution as $r \to \infty$, and the mean is held constant (which requires that $p \to 0$). In this case, the smaller the value of r, the greater the variance.

There are alternative parameterizations to the negative binomial. Rather than specify r and p, as in the above parameterization, the mean μ and a dispersion parameter ϕ can be specified. The dispersion parameter $\phi = 1/r$.

$$f(k; \mu, \phi) = \frac{\Gamma(k + \phi^{-1})}{\Gamma(\phi^{-1})\Gamma(k+1)} \left(\frac{1}{1+\mu\phi}\right)^{\phi^{-1}} \left(\frac{\mu}{\phi^{-1}+\mu}\right)^{k} \tag{25}$$

In this case, the variation σ^2 is given by:

$$\sigma^2 = \mu + \phi\mu^2 \tag{26}$$

The negative binomial is used, among other things, to model gene counts in RNA-seq data (Robinson et al., 2009):

$$Y_{gi} \sim NB(M_i p_{gj}, \phi_g) \tag{27}$$

where Y_{gi} is the number of counts for gene g in sample i, M_i is the total number of counts for sample i, p_{gj} is the fraction of counts for gene g in treatment j (to which i belongs), and ϕ_g is the dispersion for gene g. $\mu_{gi} = M_i p_{gj}$. The dispersion can be interpreted as a coefficient of variation of biological variation for gene g across samples. Since r and ϕ are inversely related, as ϕ_g decreases r approaches infinity and the distribution approximates a Poisson distribution. When there is no biological variation in gene count between samples, ϕ_g is 0 and all technical variation is accommodated by the Poisson (Robinson et al., 2009). In practice, ϕ_g is not calculated independently for each gene, a common ϕ is calculated across all genes or information is shared across genes when calculating each ϕ_g.

2.7 χ^2 distribution

The χ^2 distribution is a probability density function for the sum of the squares of k variables, each of which is independently drawn from a normal distribution with $\mu = 0$ and $\sigma^2 = 1$.

$$f(X; k) = \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} X^{\frac{k}{2}-1} e^{-\frac{X}{2}} \tag{28}$$

2.8 F distribution

The F distribution is just the distribution of a ratio of two χ^2 variables, each divided by its degrees of freedom. That is, if $X = \frac{\chi^2_m/m}{\chi^2_n/n}$, then X follows an F distribution with m and n degrees of freedom:

$$X \sim F(m, n) = \frac{\chi^2_m/m}{\chi^2_n/n} \tag{29}$$

2.9 Uniform distribution

This distribution is a flat line over a specified range of x values, x_{min} and x_{max}. The value of y is selected such that the area under the line in the specified range is 1, i.e. $y = 1/(x_{max} - x_{min})$.

2.10 t-distribution

The t-distribution is, among other things, used to compare sample means when the variance is unknown. If the variance were known, then in many cases a normal distribution would suffice. As k (the degrees of freedom) approaches infinity, the t-distribution approaches a normal distribution with $\mu = 0$ and $\sigma^2 = 1$ (i.e., a z distribution). The tails of the t-distribution have more area than do the tails of the normal distribution, accommodating the uncertainty due to the lack of an estimate for variance.

$$f(X;k) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(\frac{k}{2})}\left(1 + \frac{X^2}{k}\right)^{-\frac{k+1}{2}} \tag{30}$$

Although the t-distribution is not part of the exponential family of distributions, it can be constructed from two of the most basic exponential family distributions, the standard normal distribution and the χ^2 distribution, as follows (Grafen and Hails, 2002). If $X = \frac{N(0,1)}{\sqrt{\chi^2_k/k}}$, with the standard normal and χ^2_k variables independent, then X follows a t-distribution with k degrees of freedom:

$$X \sim t(k) = \frac{N(0,1)}{\sqrt{\chi^2_k/k}} \tag{31}$$

Where $N(0,1)$ is the standard normal distribution. Since the mean of the χ^2_k distribution is k, the denominator goes to 1 as k goes to infinity.

Now that you have seen the relationship of the t-distribution to the normal distribution, it is also interesting to note its similarity to the F distribution. The square of the standard normal distribution is the χ^2_1 distribution. If we square the t distribution, we therefore get a ratio of two χ^2 distributions, each divided by their degrees of freedom. Sound familiar? This is an F distribution with $m = 1$. So, the square of a t-distribution is a particular F distribution. The implications of this relationship will be seen when we explore General Linear Models below.

3 Laws and Theorems

3.1 The Central Limit Theorem

While we often think of the normal distribution being useful because it fits so many observed data, one of its most powerful applications results from the Central Limit Theorem (CLT). The CLT, in its most usual form, states that

"As sample size increases, the means of samples drawn from a population of any distribution will approach the normal distribution" Sokal and Rohlf (1995). The key here is that this happens *even when the underlying distribution of the data is not normal*. The variance of the resulting normal distribution will decrease as the sample size increases, due to increased precision. This has a variety of implications:

- In large enough samples, the mean of the sample means will approximate the population mean, μ.
- The standard error of the sample means, i.e. the standard deviation of the normal distribution of sample means, will be $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ where σ is the population standard deviation and n is the sample size. Usually we don't know σ, though, so we estimate it from the data as $SE_{\bar{X}} = s/\sqrt{n}$ where s is the standard deviation of the actual sample. This is often called the standard error of the mean. Note that *the standard error SE is a completely different quantity than the observed standard deviation s of the sample*. s gives an indication of how much variation is observed within a sample. SE gives an indication of how much variation is expected across samples.
- A normal distribution with mean np and variance $np(1-p)$ can be used to approximate a binomial distribution when the number of trials n is very large.

3.2 Law of large numbers

As the sample size increases, the average of the sample will approach the average of the population (i.e., the expected value).

4 Hypothesis testing

In hypothesis testing, the data are evaluated against a null hypothesis H_0 to see how unusual they are. If there is only a small chance of the data being realized under the null hypothesis, then the null hypothesis is rejected. The null hypothesis is paired with an alternative hypothesis, H_A, that is less specific than the null hypothesis and encapsulates all alternatives to the null hypothesis (sometimes there may be a set of alternative hypotheses that describe the alternatives, rather than just one). Usually there are a variety of assumptions that are made under the null hypothesis. Violation of any of these assumptions could lead to rejection of the null, even though the investigator is often interested only in whether or not one of the assumptions are violated.

To test a null hypothesis, the distribution of data that would be expected under the null is generated, and we ask how frequently data at least as extreme as those observed would be expected under the null model. Typically a P-value, the probability, assuming the null hypothesis is true, of obtaining results at least

as extreme as those observed, is reported. If the P-value is small, i.e. less than a particular α, the null hypothesis is rejected. The comparison between the sample data and the null model is made with a test statistic that summarizes some aspect of the sample data. The test statistic could, for example, be the mean of the sample or the proportion of the observations in the sample that fall into a particular category.

4.1 Relationship to confidence intervals

In many cases the test of the null hypothesis is equivalent to examining the confidence interval of the null distribution (Whitlock and Schluter, 2008). If the test statistic derived from the observation falls out of the confidence interval that contains a specified fraction of the expected values under the null hypothesis (e.g. 95%), then the null is rejected since that outcome would be expected by chance less than 5% of the time.

4.2 General Linear Models

If you've taken an introductory class to statistics, you probably learned a bunch of different tests for examining the significance of the relationship of one variable to another, including the t-test, ANOVA, MANOVA, F tests for regression, and others. Well, it turns out that they aren't so different—they are all special cases of the general linear model. Statisticians have known this for some time, many statistical software tools make no distinction between them (see the lm() function in R, for example), yet scientists still learn them as independent methods.

Here "general linear model" refers to the classical linear model with normally distributed errors, *not* the generalized linear models of McCullagh and Nelder (1989) (in which the abbreviation "GLM" is now most often used). Generalized linear models extend the general linear model to non-normal response distributions.

For a clear and informative introduction to the general linear model, and a discussion of how it relates to tests you may already be familiar with, see the first three chapters of Grafen and Hails (2002).

5 Multiple independent tests

If multiple independent tests are made, the risk of Type I error (falsely rejecting a true null hypothesis) rapidly grows.

The Bonferroni correction is a common mechanism for addressing this issue. A new more stringent α is simply calculated as $\alpha^* = \alpha/n$ where n is the number of independent tests. While this reduces the rate of Type I error, it does so at the cost of reducing the power of each test, elevating the Type II error.

Alternatively, the False Discovery Rate (FDR) may be used.

References

Grafen, A., and R. Hails. 2002. Modern Statistics for the Life Sciences. 1st ed. Oxford University Press, USA.

McCullagh, P., and J. A. Nelder. 1989. Generalized Linear Models. Monographs on Statistics and Applied Probability, 2nd ed. Chapman and Hall/CRC.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth. 2009. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics (Oxford, England) 26:139–140.

Sokal, R. R., and F. J. Rohlf. 1995. Biometry. the principles and practice of statistics in biological research, 3rd ed. W. H. Freeman.

Twomey, P. J., and M. H. Kroll. 2008. How to use linear regression and correlation in quantitative method comparison studies. International Journal of Clinical Practice 62:529–538.

Whitlock, M. C., and D. Schluter. 2008. The Analysis of Biological Data. 1st ed. Roberts and Company Publishers.