Asking Questions in Biology

A Guide to Hypothesis-testing, Experimental Design and Presentation in Practical Work and Research Projects

Fifth edition

Chris Barnard
Francis Gilbert
Peter McGregor

Pearson

Asking Questions in Biology

Transformations.

So what do we do if we suspect our residuals may not be normally distributed? Happily, and as long as our sample size is big enough (> 50 as a rough guide) to make a comparison meaningful, the wide range of statistical packages now available for personal computers makes the answer simple. Test it! There are some well-established significance tests that allow comparisons between frequency distributions of data and various theoretical distributions, of which the normal is the commonest. Those used most commonly are the one-sample Kolmogorov–Smirnov test (for large samples, where $n > 2000$), chi-squared and, for small- to medium-sized samples where n lies between 3 and 2000, the Shapiro–Wilk test. R® makes the process of checking for normality particularly easy (*see* Box 3.1). If we test our distribution and find it does not differ significantly from normal, then we're at liberty to use any appropriate parametric test at our disposal. If it does differ, we can do one of two things. We can abandon the idea of using parametric statistics and choose an appropriate *non-parametric* test instead (*see below*), or we can *transform* the data to see whether they can be normalised. Several transformations are available but the most widely used are probably *logarithmic* [log(x) or, where there are zeros in the untransformed data, $\log(x + 1)$] or *square-root* transformations. Simply log or take the square root of each data value, and test for normality again. Where percentages or proportions stray below about 30 per cent (or 0.3) or above 70 per cent (or 0.7), an *arcsine square-root* transformation (calculated by taking the square root of the proportion – so divide percentages by 100 first – then the inverse sine ($\sin^{-1}$ on many calculators) of the result) will stretch out the truncated tails and prevent undue violation of the normality assumption of parametric tests (Box 3.1 shows how to transform data in R®).

BOX 3.1 Testing whether data conform to a normal distribution

Testing for normality as part of a test for differences

Assume that we are interested in whether the mean sizes (response variable) of seven groups (the predictor, a grouping variable called a *factor*) of grasshoppers are different, and that we have 100 measurements for each group. One of the assumptions of the test is that the residuals of the response variable, the differences of each value from the mean for its group, are normally distributed. Another assumption is that the residuals of each group all have the same normal distribution (i.e. the variances of the groups all have the same value – referred to as the 'homogeneity of variances' assumption): we shall repeat the test for this assumption in the context of exactly how to carry out tests of difference, later on (*see* Box 3.3a on p.77). For now we are concerned simply with testing the assumptions.

In R®, as with nearly all statistical packages, the data for the response variable (`size`) are in a single column, and a second column (`grassh`) indexes the group to which each value belongs (here running from `A` to `G`). Such group-membership variables are called factors or, more generally, predictors. Notice that factors are nominal variables. Here the values are alphanumeric so R® assumes automatically that the variable is a nominal factor: hence we do not need to declare it as such before running the model.

We now 'fit the model', here by running a General Linear Model, or *glm,* saving it in the object 'm1':

```
> m1 <- glm(size ~ grassh)
```

We will see standard tests for a difference among groups later in detail (Box 3.3a). If not asked, R® produces no output, so we are not bothered by details of results from the test of differences. However, R® holds all the required details in memory (in `m1`) that we need for testing for normality. In particular, the residuals from the fitted model are held in `resid(m1)`. To test for normality, we should do three checks: (i) a statistical test of normality; (ii) visually plot the residuals together with a normal distribution; and (iii) look at the quantile–quantile plot (or q–q plot), which plots the ranked residuals against a similar number of ranks produced from a normal distribution – the result should be a straight line, with systematic deviations from this interpretable as various kinds of non-normality. If there is evidence of skew, then we can test for that as well.

For (i), the first check, we simply type/paste:

```
> shapiro.test(resid(m1))
```

and R® gives us the result:

```
Shapiro-Wilk normality test
data: resid(m1)
W = 0.9974, p-value = 0.3316
```

Here the null hypothesis is that the distribution is normal, and since the probability is not less than 0.05, we cannot reject this on the basis of these data. This looks fine.

We can also easily do (ii), visualising the distribution of the residuals together with the expected normal distribution. First we plot the histogram of residuals, here running from −6 to +6 with a bar width of 0.5:

```
> hist(resid(m1), breaks=seq(-6,6,0.5))
```

Then we construct a set of *x*-values running from −6 to +6.

```
> xv <- seq(-6,6,0.1)
```

and then generate the normal curve for a mean of zero and the observed standard deviation of our residual, using the probability density function `dnorm`. The height of our frequency distribution depends on how many data points there are, so we have to add a scaling factor. A rough guide to the correct scaling factor is the number of data points multiplied by the chosen bar width:

```
> hist.ht <- length(resid(m1))*0.5
> yv  <-  dnorm(xv,mean=0.0,sd=sqrt
  (var(resid(m1))))*hist.ht
```

and add this to the plot:

```
> lines(xv,yv)
```

From Fig. (i) we see that the fit is really pretty good.

Obtaining (iii), the q–q plot, involves plotting the data (`qqnorm`) and then the line (`qqline`), which is a dashed (`lty=2`) rather than a solid line (`lty=1`). Thus we type:

```
> qqnorm(resid(m1))
> qqline(resid(m1),lty=2)
```

In Fig. (ii) we can see a very good straight line apart from some minor deviations right at the ends. Thus all seems well: our data are not significantly different from a normal distribution.

Testing whether the normal distribution is the same in every group

A very important assumption of parametric tests (i.e. those based on a particular distribution of residuals, usually the normal) is

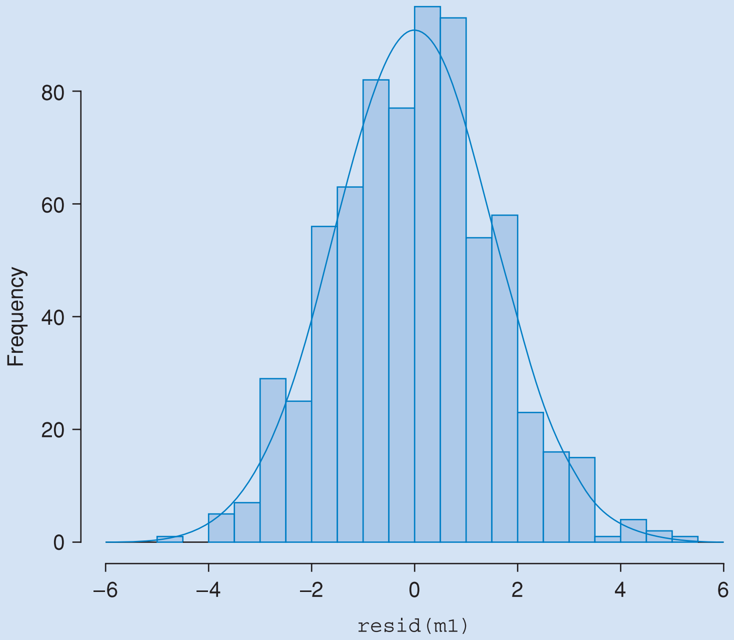


Figure (i) Distribution of the residuals from the model: clearly they are close to normal.

that of constant variance, i.e. that the normal distribution is the same in every group. There are two possible tests for this, Bartlett's and the Fligner–Killeen:

```
> bartlett.test(size ~ grassh)
```

```
Bartlett test of homogeneity of
variances
data: size and grassh
Bartlett's K-squared = 4.7749,
df = 6, p-value = 0.573
```

or

```
> fligner.test(size~grassh)
```

```
Fligner-Killeen test of homogeneity of
variances
data: size by grassh
Fligner-Killeen:med chi-squared =
4.4846,
df = 6, p-value = 0.6114
```

Either way we get the same result, which is that there is no evidence to reject the null hypothesis of homogeneity of variances among these groups.

A quick way of producing a set of four diagnostic graphs to help in assessing the assumptions of our test (or 'model') is to fit the model and then use '`plot(model)`':

```
> m1 <- glm(size ~
  grassh)
> plot(m1)
```

and R® produces a set of four diagnostic graphs, obtainable in sequence by pressing the <Enter> button. One is the q–q plot. Others help us to assess outliers and the assumption of homogeneity of variance, and the row numbers of the outliers are helpfully identified on the plots. It is a good idea to use this routinely whenever we fit any kind of model (*see* Box 3.14). In fact, R® standardises the residuals automatically in different ways according to the assumed distribution of the residuals (normal, Poisson, binomial, etc.), and hence the linearity of the q–q plot is always an important part of checking your model assumptions.

If this routine identifies some outliers (e.g. rows 33 and 54 of 99 data points), then it is easy to remove them by weighting them out of the analysis. Thus we set up a weighting variable of 1s and 0s to identify the values we want removed (the 0s). First make a variable containing the outlier rows:

```
> sel <- c(33,54)
```

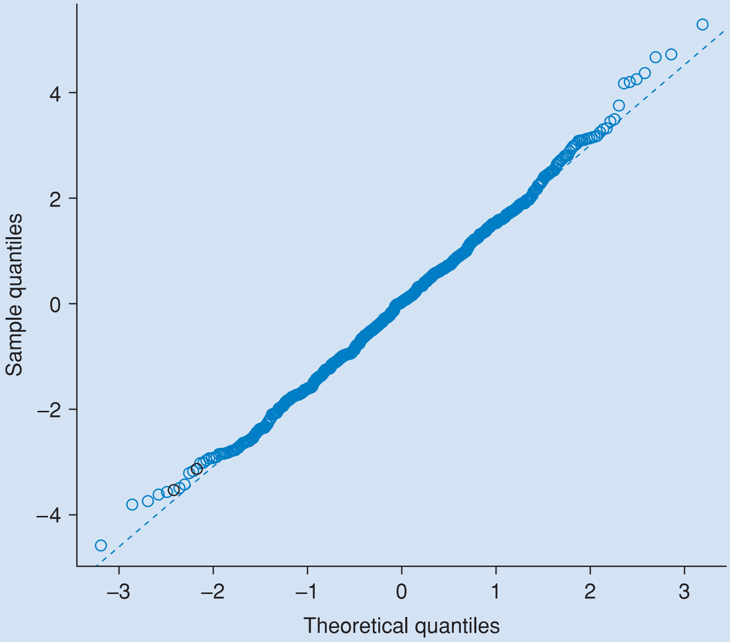


Figure (ii) The *q*-*q* plot for the model, showing the straight line expected for the normal distribution.

then a variable of all the row numbers of the data:

```
> num <- c(1:99)
```

and then a weighting variable that has 0 for the rows that match the numbers in ‘`sel`’, and 1 for all other rows:

```
> wt <- ifelse(is.na(match
  (num,sel)),1,0)
```

Then re-run the analysis weighting out those rows, and check for normality again:

```
> m2 <- glm(size ~ grassh, weights = wt)
```

If you want permanently to exclude these rows from all subsequent analyses, then copy all but the excluded rows into a new dataframe, remove the old one, and attach the new:

```
> dtfr2 <- dtfr
  [dtfr$wt==1,]
> remove (dtfr)
> attach (dtfr2)
```

Transforming the response variable is very easy in R®:

```
> Lsize <- log10(size)
  transforms to base10
  logarithms
> Rsize <- sqrt(size)
  transforms using
  square root
> asn_trans(var) trans-
  forms a percentage
  variable using the
  arc-sine square-root
  transformation
```

Testing for normality as part of a test for a trend

In R®, the way you test for a trend is very similar to the way you test for a difference. The only distinction lies in the predictor, which is a continuous or constant-interval variable, rather than a factor. The same set of residuals are produced, which you can test for normality in an exactly similar manner. You can also test for homogeneity of variances in exactly the same way.

Of course, even transformation may not succeed in normalising our data, in which case we must seriously consider using non-parametric statistics. Indeed, we may not even get as far as worrying about the normality of our response variable before opting for a non-parametric approach. Among their various other requirements, parametric tests based on the normal distribution demand that measurements are of the constant-interval kind, so in principle cannot deal with the other types of measurement we might be forced to use (although in practice they often can). Non-parametric tests are much less restrictive here.

Non-parametric tests.

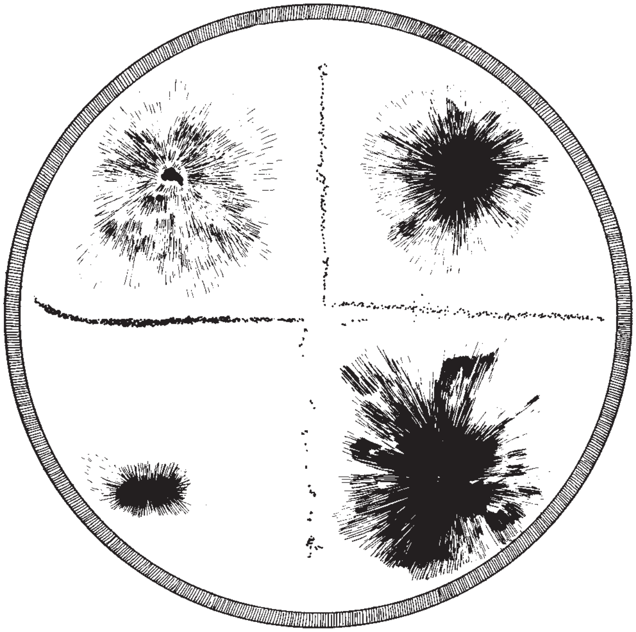
Non-parametric tests are sometimes referred to as *distribution-free*, *ranked* or *ranking* tests because they do not rely on residuals being distributed normally, and generally work on the ranks of the data values rather than the data values themselves. While they may be distribution-free, however, they are not entirely assumption-free. They assume the response data have *some* basic properties, such as independence of measurement (*see later*) and a degree of underlying continuity (*see* Martin & Bateson, 1993): crucially, they also make the same assumption of equal dispersion among groups as do parametric tests (which assume equal variances among groups). In most cases, however, these assumptions are easily met. In the jargon of statisticians, non-parametric tests are thus more *robust* because they are capable of dealing with a much wider range of data sets than their parametric equivalents. While they can deal with the same constant-interval measurements as parametric tests, they can also cope with ordinal (ranking) and nominal (classificatory) measurements. Non-parametric tests are especially useful when sample sizes are small and assumptions about underlying normality particularly troublesome. There are a couple of drawbacks, however. The first, arguably overstated (*see* Martin & Bateson, 1993), weakness is that non-parametric tests are generally slightly less powerful (*power* here meaning the probability of properly rejecting the null hypothesis – we shall return to this shortly) than their parametric equivalents. The second, which is slowly being addressed (*see for example* Meddis, 1984), is that the range of tests for more complex analyses involving several variables at the same time is very limited. Sophisticated multivariate analysis is still the undisputed province of parametric statistics. Nevertheless, for our purposes, and with a few exceptions, there are perfectly good parametric and non-parametric equivalents, and we shall introduce them both in our discussion of significance testing.

3 One-tailed versus two-tailed, and general versus specific tests

The third important issue we must consider relates to the prediction we are trying to test. Suppose we are predicting a difference between two sets of data for a particular response variable, say a difference in the rate of growth of a bacterial culture on agar medium containing two different nutrients (the predictor, a factor). We could make two kinds of prediction. On the one hand, we could predict a difference without implying anything about which culture should grow faster. In this case, we wouldn't care whether culture A grew faster than culture B or vice versa. This is a *general* prediction. On the other hand, based on some prior knowledge or theory we might predict that one particular culture would grow faster than the other, e.g. A would grow faster than B; this is a *specific* prediction. Which of these kinds of prediction we make affects the way we test the predictions.

The same distinction arises with trend predictions. Imagine we want to know whether there is a trend between the size of a male cricket and the number of fights he wins over the course of a day. We can make a general prediction (there will be a trend, positive or negative), or we can make a specific prediction (larger males will win more fights; i.e. the trend will be positive). We can think of the general prediction as incorporating both positive and negative trends; either would be interesting. The specific prediction is concerned with only one of these.

In cases like those above, where there are only two possible specific predictions within the general one, we can use the same significance test for either general or specific predictions, but with different threshold probability levels for the test statistic (*see below*) to be significant. Because here the specific and general predictions are concerned with one and two directions of effect respectively, the threshold value of the test statistic at the 5 per cent level in the general test becomes the threshold value at the 10 per cent level in the specific test. In statisticians' jargon, we thus do either a one-tailed (specific) or a two-tailed (general) version of the same test. There is, of course, an obvious, and dangerous, trap for the unwary here. The trap is this: if the value of a test statistic just fails to meet the 5 per cent threshold in a two-tailed test, there is a sore temptation to pretend the test is really one-tailed so that the test statistic becomes significant. *It must be stressed that this is tantamount to cheating. A one-tailed test is legitimate* ***only*** *when the prediction is made in* ***advance*** *of obtaining the result* ***and*** *when results in the opposite direction can reasonably be regarded as equivalent to no difference or trend at all.*[†] *It is completely inadmissible as a fallback when a two-tailed test fails to yield a significant outcome.* A one-tailed test should thus be used only when there are genuine reasons for predicting the direction of a difference or trend *in advance.*



[†] If you predict from theory that A will have a greater mean value than B (i.e. H_1 is $A > B$), you are also assuming that both of the alternative results ($A < B$ and $A = B$) are *equivalent* and *together* form the null hypothesis, H_0. Thus if you find that, contrary to your prediction, the mean value of B is much greater than that of A, and would have been significant had you framed your hypothesis as a general one (i.e. H_1 is $A \neq B$), you are not allowed to conclude *anything* other than that the null hypothesis has *not* been rejected.