Chapter 12

Statistical inference

12.1 Introduction

The development of statistical ideas in the first half of the 20th century, including the 'classical' theory of statistical inference and hypothesis testing, was arguably one of the greatest and possibly undersung intellectual achievements of that period. The powerful ideas developed – endowing the treatment of problems involving uncertainty with precision – are seductive and the 'New Archaeologists', from the 1960s on who promoted quantitative methodology, were accordingly seduced (see Chapter 1 of Baxter, 2003).

The statistical landscape has changed since statistics acquired its cachet in some archaeological circles, particularly with respect to the development of computing power and the exploitation of computer-intensive methodologies. The effect of these more recent computer-intensive techniques on *statistical* practice in archaeology merits a review, though this is not attempted here.

The main aim of this chapter is to provide a brief review of the ideas that underpin hypothesis testing which, with variation in emphasis, form one of the staples of quantitative archaeology texts. Their application using `R` is illustrated.

Central to the statistical theory is the idea of drawing inferences about populations from random samples. As discussed in many texts on quantitative methods in archaeology, serious questions arise about the nature of the 'population' sampled, the extent to which samples can be treated as 'random', and the implications this has for the applicability of formal methods of statistical inference to archaeological data. Attitudes have ranged from enthusiastic promotion of statistical inferential methods to scepticism about its practical value – this latter among archaeologists otherwise sympathetic to what statistics has to offer.

This ambivalence can be traced back to the early flourishing of quantitative methodology in archaeology. Archaeologists who embraced statistical inferential

ideas sometimes explicitly embedded their thinking within an overtly positivist framework, promoting the approach as 'scientific' and 'objective'. 'Over-selling' of the ideas predictably generated negative reactions. At one extreme, rejection of the New Archaeology led to wholesale rejection of statistical methodology. This is not a logically defensible position; the use of statistical analysis can be decoupled from any overarching philosophy attached to it. As Brandon later commented, in the preface to Westcott and Brandon (2000), 'during its heyday, statistics had been waved above archaeologists' heads as an "answer" to dealing with a multitude of archaeological problems' but 'after much yelling and arm-waving, most agreed that statistics were not an answer in themselves but ... an extremely important tool available for archaeological use'.

Doran and Hodson (1975), in a text sympathetic to the exploration of statistical ideas, explicitly distanced themselves from New Archaeology, finding its claims 'greatly exaggerated and therefore dangerous' and 'a bizarre mixture of naivety and dogmatism' (Doran and Hodson 1975: 5). Their treatment of statistical inference was fairly short and 'theoretical' with few examples of applications; they suggested that, compared to the classical statistical approach, 'a rather different rationale for disciplined inference in archaeology is required' and that this 'remains to be devised' (Doran and Hodson, 1975: 94–95).

Later textbooks have accorded more space to hypothesis testing and inference. Shennan (1988, 1997) and Drennan (1996, 2009) are the most widely used. The more recent text of VanPool and Leonard (2010) contains the most extensive treatment of inferential procedures in introductory archaeology texts I know of but is flawed in some respects (footnote 10).

The normal distribution is fundamental to the statistical theory. Though not used directly for hypothesis testing as much as one might think, it leads to the development of other distributions (t, F, chi-squared etc.) that are widely used in practice. These are the subject of Sections 12.2.3 and 12.2.4. Sections 12.2.1 and 12.2.2 use the normal distribution to introduce ideas central to statistical inference. The chapter misses out a lot; some omissions are noted in Section 12.4

12.2 Common hypothesis tests

12.2.1 The normal distribution

The intention here is to provide a condensed summary of some common hypothesis tests, and associated ideas, along with their implementation in `R`. It is assumed that the reader is acquainted with the idea of the normal distribution. Computational formulae are not provided; they are widely available if needed. Software such as `R` does the computation, leaving users to concentrate on underlying ideas and

interpretation.

Mathematically, the normal distribution defines a curve with a total area of 1 beneath it. Distributions are characterized by *parameters*; the normal depends on only two, its *mean*, μ, and *standard deviation*, σ. The distribution is symmetrical about its mean, and 'bell-shaped', so the mean is also the median and mode. About 95% of the distribution lies within two standard deviations from the mean, and 68% within one standard deviation.

If a random variable X has a normal distribution the notation

$$X \sim N(\mu, \sigma^2)$$

is used to express this, where σ^2 is the *variance*. Given a *random sample* of size n from a normal distribution denote the *estimated* mean and standard deviation by $\bar{x}$ and s; the estimated standard error of the mean is $s/\sqrt{n}$. If σ is known $\sigma/\sqrt{n}$ can be used rather than its estimate.

With this notation in place, assuming known σ,

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

has the *standard normal distribution* with mean zero and unit variance.

Many applications involve finding the probability in the tail of the distribution, defined by some value of z. Output in `R` for analyses based on the normal and other distributions often include the information needed, but functions exist allowing their separate calculation if necessary.

Thus, the `pnorm` function, for $z = -2.36$ using `pnorm(-2.36)`, returns the value 0.009137468, the probability in the lower-tail of the distribution written as P(z < -2.36) = 0.009. Replacing z with + 2.36 gives 0.9908625, or P(z $<$ $+2.36$) = 0.991. These are examples of a p-values.To get the upper-tail probability P(z $> +2.36$) either subtract this from 1 or, more directly, use `pnorm(2.36, lower.tail = FALSE)`. In either case, given the symmetry of the distribution, 0.009137468 is obtained[1].

These are examples of lower-tail and upper-tail probabilities. Often interest lies in assessing how extreme the observed result is; that is, what is the (combined) tail probability of getting a value either more negative than -2.36 or more positive than +2.36. This is obtained by doubling the lower- or upper-tailed probability calculated and is written as the two-sided p-value P($|$z$|$ > 2.36) = 0.018, or slightly less than 2%. The p-value needs to be interpreted and this brings us on to the subject of inference.

[1]This renders redundant the detailed tables of the normal and other distributions – once essential – still to be found in the appendices of introductory texts.

12.2.2 Inference

The value of z derived from a sample can be used in two ways. If μ and σ are known for some population of interest, assumed to be normal, the p-value can be used to assess if it is plausible that the sample is selected from the population specified. If the p-value is 'small' it would be concluded that the sample is unlikely to be from the population. In practice this kind of application is limited.

More usually, and the main idea in the tests used here, a value is *assumed* for the mean μ, and the observed data are used, via the calculated z statistic, to assess the plausibility of this assumption. A simple way of thinking about the more basic inferential procedures is that they are employed in order to say something useful about the true values of the unknown parameters; this includes the use of confidence intervals (Section:12.2.5).

The *null hypothesis* states that μ equals some specified value, μ_0. The notation used to express this is $\mathrm{H}_0 : \mu = \mu_0$. The z-statistic can be calculated using this assumption. If it is 'unusual' as measured by the p-value then we *reject* the null hypothesis, otherwise we do not reject it[2].

Decisions need to be made on the part of the user. One is whether to use one- or two-sided p-values. The choice amounts to specifying an *alternative hypothesis*; for example, if $\mathrm{H}_0 : \mu = 0$ possible alternative hypotheses are $\mathrm{H}_1 : \mu < 0$, $\mathrm{H}_1 : \mu > 0$ or $\mathrm{H}_1 : \mu \neq 0$. The first two of these are examples of one-sided alternative hypotheses where it is believed that, if the null hypothesis is incorrect, the true value differs in a particular direction from that assumed; the third possibility, an instance of a two-sided test, specifies that if the null hypothesis is incorrect the true value of μ could lie either side of that hypothesized. Two-sided tests will be used below unless otherwise stated.

The important question is 'what constitutes a 'small' or 'unusual' p-value?'. Some researchers are content to report the exact p-value, allowing others to decide on their own interpretation. Commonly, though, a *decision rule* is used. These are arbitrary; conventionally the use of rules based on p-values of 0.05 and 0.01 are widespread. If the p-value is less than 0.05, using a two-sided test, the null hypothesis is rejected at the 5% level of significance or, more concisely is *significant at the 5% level*. If the p-value is also less than 0.01 the null hypothesis is rejected at the 1% level, providing stronger evidence against the null hypothesis.

[2]It is tempting to say, in the latter case, that the null hypothesis is 'accepted' and the temptation should be resisted. A null hypothesis may not be rejected for a host of reasons, of which the possibility that it is 'true' is only one. Small sample sizes and/or large sample variability can both lead to non-rejection, even when the null hypothesis is false. Non-rejection implies that there is not enough evidence with the data to hand to reject the null hypothesis; the term 'accepted' carries connotations of establishing that the null hypothesis is 'correct' which is not the case.

As a 'rule-of-thumb', using p-values based on 0.05, 0.01, and 0.000, and rejecting the null hypothesis, implies 'strong', 'very strong' and 'overwhelming' evidence against the null hypothesis. For the normal distribution the 5% and 1% levels of significance correspond to values of $|z| > 1.96$ and $|z| > 2.575$. This means that, given a z-score one looks to see if its modulus exceeds 1.96 or 1.575. These are called *critical values*.

As a notational aside, α can be used for the associated p-values, with $z_{\alpha/2}$ being the z-score that 'cuts off' $\alpha/2$ of the probability in the upper-tail. Thus, $z_{.025} = 1.96$ defines a tail probability of 0.025 which, for a two-sided test, is doubled to get the significance level of 0.05 (5%), that is, $\mathrm{P}(|z| > 1.96) = 0.05$. Some help is available in choosing α. It is the probability of incorrectly rejecting a correct null hypothesis; this is a *Type I error* and the user controls the probability of this in setting the decision rule. If it is important not to make this sort of error a relatively small value of α would be set.

A *Type II error*, occurs when an incorrect null hypothesis is not rejected. Call the probability of this β; the *power* of a test is defined as $(1 - \beta)$ and is the probability of correctly rejecting an incorrect null hypothesis. A balancing act between the two types of error is implicit since reducing the size of α increases the size of β and hence also reduces the power. Usually the focus of interest is on controlling α and one 'lives with' the associated power; increasing the sample size n is one way of improving the power. Other things (i.e. α) being equal, competing tests that do the same job can be compared in terms of their relative power.

12.2.3 Tests of means – t-tests

The z-test is of limited use for practical purposes because of the 'σ known' assumption; it usually isn't. This is remedied in a straightforward way at the expense of a little extra complexity. The z-statistic can be expressed in a general form as

$$z = \frac{\text{difference}}{\text{SE}}$$

where 'difference' is the observed difference between $\bar{x}$ and μ_0, and SE is the standard error which scales the statistic so it is dimensionless. The 'obvious' modification if σ is unknown is to use the estimated standard error, $s/\sqrt{n}$ resulting in the t-statistic

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}.$$

The distribution of this (assuming independent random sampling from a normal distribution) looks much like the normal but with thicker tails, the thickness of which depends on the sample size, n, through a quantity known as the *degrees of*

freedom (DF) which is $(n-1)$ for single-sample tests. As n gets larger $s \to \sigma$ and the t-distribution approaches the normal distribution.

The t-statistic is therefore most valuable when dealing with small samples. It was for precisely this problem that the statistic was developed by W. G. Gosset, published under the pseudonym 'Student' in the journal *Biometrika* in 1908 (hence the terminology 'Student's t-test' sometimes seen). The ideas of the previous section carry through, but the critical value that varies with n is now denoted $t_{\alpha/2}$.

In `R` suppose that, for a sample size $n = 7$ with 6 DF, $t = -2.36$ is obtained. Using the `pt` function, `pt(-2.31, 6)` returns 0.0281, where the second argument in the call to the function, 6, is the degrees of freedom. Doubling this to get the two-sided p-value gives 0.056 so the null hypothesis would not be rejected at the 5% level of significance; for $n = 17$, 37 and 137 the respective probabilities are 0.031. 0.024, and 0.020. If a critical value, $t_{\alpha/2}$ is required use the `qt` function; for a sample size of 7 and $\alpha = .025$, `qt(0.025, 6)` gives $|t_{\alpha/2}| = 2.45$.

One-sample hypothesis tests are often uninteresting. The specification of a realistic null hypothesis can be 'artificial' (where does it come from?). It is difficalt to find examples in the literature where real data (i.e. with contextual information) is confronted with a realistic problem requiring a one-sample test (see, also, Section 12.2.5). Practically, the comparison of samples from two populations is more likely to be of interest. Here there is a 'natural' null hypothesis to test.

The two-sample t-test has a similar structure to those already discussed. The difference between the two sample means is $(\bar{x_1} - \bar{x_2})$ and the natural specification of the null hypothesis is that the difference in population means is $(\mu_1 - \mu_2) = 0$ (note the use of subscripts to distinguish between the two populations/samples). Extra complexity is introduced because the SE can be estimated in two ways. The first of these – that usually described in introductory texts and the default in many software packages – assumes that the population variances are equal, $\sigma_1^2 = \sigma_2^2$, and the estimate of the SE is based on a weighted average of their estimates, with DF $= (n_1 + n_2 - 2)$. In the second case no such assumption is made and the SE is estimated as the square-root of the sum of the squared SE estimates of the two samples. In this case the DF needs to be approximated, and an approximate t-test is carried out using this.

It's possible to test the hypothesis that the variances are equal in advance of a t-test – the subject of the following section. More simply, it is straightforward to carry out tests with and without the assumption (Section 12.3.2). Minor numerical differences apart results, in terms of the conclusions to be drawn, will often be the same; if not there may be problems with the samples used that need addressing before applying any formal test about the means. The difference in the definitions of the estimated SEs is spelled out in the footnote; it is highly improbable that

you will ever need to calculate these 'by hand'[3].

12.2.4 Tests of variances

The sum of squares of n independent identically distributed normal random variables has the chi-squared (χ^2) distribution, which is asymmetrical and bounded below by zero. The ratio of two chi-squared variables has the F-distribution; these can be used directly for one- and two-sample hypothesis tests but arise in other contexts as well.

The null hypothesis $\mathrm{H}_0 : \sigma^2 = \sigma_0^2$ can be tested using a chi-squared statistic, but I cannot recall seeing any realistic application of this in the archaeological literature and will not discuss it further. Chi-squared tests do have a useful role in testing the hypothesis of no association between two categorical variables and this is illustrated in Section 12.3.3.

The F-test is based on the ratio of two chi-squared random variables and may be used for testing hypotheses about two population variances where the 'natural' null hypothesis is $\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2$, or $\mathrm{H}_0 : \sigma_1^2/\sigma_2^2 = 1$. This can be applied in advance of a two-sample t-test to see if the equal variances assumption is reasonable. The hypothesis is tested using the statistic

$$\mathrm{F} = s_1^2/s_2^2$$

which under the null hypothesis, follows the F-distribution with $(n_1 - 1), (n_2 - 1)$ DF, a little more complicated than the previous tests since it depends on two separate degrees of freedom.

[3] In the first case the estimated SE is

$$s\sqrt{(1/n_1 + 1/n_2)}$$

where

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

. In the second case the estimated SE is

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

None of the quantitative archaeology texts I am familiar with give the formula for the approximate DF in the second case so, for the record, here it is.

$$DF = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_1 - 1)}.$$

As commented you ought never need to have to calculate this yourself; you should know of the two possibilities since the second case is the default in `R`.

It is convenient to label the samples so that $s_1^2 > s_2^2$ and F > 1 so only the upper-tail of the distribution is of concern. The decision rule for a 5% level of significance requires that the critical values of $F_{0.025}$ or $F_{0.05}$ be determined, depending on whether a two- or one-sided test is being used. For a problem where F = 2, $n_1 = 7$ and $n_2 = 9$ there are (6, 8) DF. Using the `qf` function, `qf(.025, 6,8, lower.tail = F)` gives $F_{0.025} = 4.652$, so the null hypothesis is not rejected at the 5% level of significance using a two-sided test. Replacing 0.025 with 0.05 gives $F_{0.05} = 3.581$, leading to the same conclusion as for a one-sided test. Replace 0.025 and 0.05 with 0.005 and 0.01 for tests at the 1% level.

The function `pf`, analogous to `pnorm` and `pt` is also available if one wishes to avoid strict adherence to a decision rule. Thus, supposing F = 4.2 with (6, 8) DF, `pf(4.2, 6,8, lower.tail = F)` gives a p-value of 0.033, which is significant at the 5% level, but not at 1%, for a one-sided test. On doubling it, it is not significant at either level for a two-sided test.

12.2.5 Confidence intervals

An 'objection' of sorts to the tests that have been described, already voiced – even when the assumptions needed are valid – is that they are 'uninteresting'. Ultimately all they do is tell you whether the single value defined by the null hypothesis is 'plausible' or not. No information is provided about what other values are, or are not, plausible, or how precisely the sample statistics estimate the population values. Confidence intervals provide the same information as an hypothesis test, and much more besides. Attention is confined to problems involving means using t-tests; similar ideas can be applied to problems concerning variances.

In a slight extension of notation let $t_{crit} = |t_{df,\alpha/2}|$ be the critical value of t for a (two-sided) test at the α (or $100\alpha\%$) level of significance, with df the degrees of freedom. The formulae for both one- and two-sample tests can be rearranged to get an expression for μ or $\mu_1 - \mu_2$ which can be evaluated at the positive and negative values of t_{crit} to obtain a $100(1-\alpha)\%$ confidence interval.

For the single-sample problem this leads to

$$\bar{x} - SE \times t_{crit} \leq \mu \leq \bar{x} - SE \times t_{crit}$$

and for the two-sample problem

$$(\bar{x}_1 - \bar{x}_2) - SE \times t_{crit} \leq ((\mu_1 - \mu_2) \leq (\bar{x}_1 - \bar{x}_2) + SE \times t_{crit}$$

where SE and df are the appropriate standard error and degrees of freedom that would be used for the associated hypothesis test. The following may be noted.

1. There is no commitment to a particular null hypothesis.

2. Any value contained in a $100(1-\alpha)\%$ confidence interval would not be rejected at the $100\alpha\%$ level of significance; if a value of μ is contained in a 95% confidence interval it would not be rejected as a null hypothesis at the 5% level of significance[4].

3. A consequence of the above is that conclusions about any null hypothesis of interest can be 'read off' from the confidence interval, so testing the null hypothesis in the manner described in earlier sections is redundant.

4. The width of a confidence interval provides some indication of how good an estimate the sample mean (or difference in means) is of the population means (or their difference). A very narrow confidence interval, for example, 'pin-points' the true value of the parameter of interest very well. An hypothesis test, without elaboration, provides no such information.

Together these amount to a powerful argument for preferring confidence intervals rather than hypothesis tests in many practical applications.

12.3 Examples of R use

12.3.1 Data

The data used are from Shennan (1997: 103). Twenty-four observations are available on the areas (square meters) of *marae* ceremonial enclosures of the Society Islands in the Pacific, located in two different valleys. Interest lies in whether or not there are differences between the valleys in terms of the mean size of the enclosures as measured by area.

The data are given in Table 12.1 and have been reconstructed from Shennan who provides log-transformed values because 'a preliminary check indicated that [area] was very skewed so it has been logged' (Shennan, 1997: 102).

The reconstructed data for the two valleys are shown in the stripcharts/dotplots of Figure 12.1, along with those for the log-transformed data.

[4]This is possibly the simplest way of thinking about the interpretation of confidence intervals. Another way is that in a hypothetically large number of repeated samples the confidence intervals will cover the true value 95% of the time; this is illustrative of the *frequentist* way of thinking about probability which not all statisticians like (to put it mildly!). In practice one has a single confidence interval to deal with, and it either does or doesn't cover the true value – you don't know. It is tempting, and understandable, to say that you are 95% 'confident' that your interval contains the true value, and is wrong. This use of the term 'confidence' can be taken to imply a conception of probability as a 'degree of belief'; this is outwith the frequentist paradigm though employed in other approaches to inference.

	Area (square meters)													
Valley 1	2.87	2.52	2.35	1.48	1.92	1.99	1.84	2.01	2.16	2.56	1.95	2.00	2.35	1.87
Valley 2	1.94	2.17	2.05	2.22	2.31	2.14	1.95	2.40	2.81	2.74				

Table 12.1: *The area (square meters) of* marae *ceremonial enclosures from two regions (adapted from Shennan, 1997: 103).*

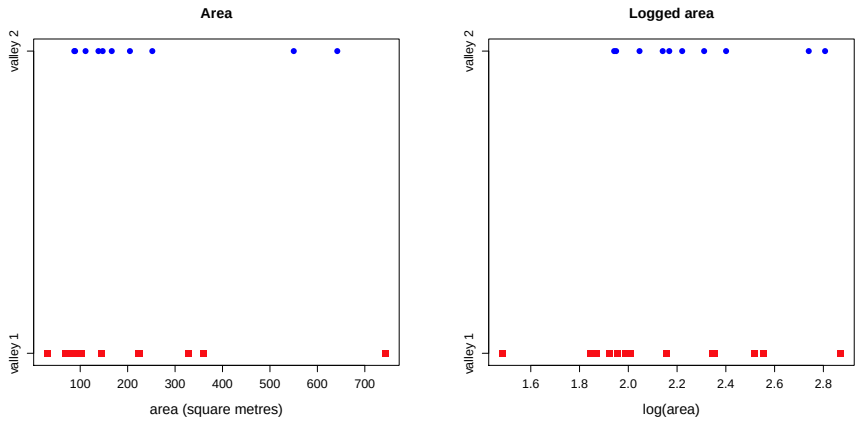

Figure 12.1: *Stripcharts of the area and log-transformed area of* marae *enclosures from two separate locations.*

The most obvious feature of the plot for the untransformed data is the two clear outliers for Valley 2, and an outlier for Valley 1 which might be interpreted as evidence of skewness; the log-transform appears to 'cure' this problem for the first valley, but not entirely so for the second[5].

Looking at the plot for the log-transformed areas there is little obvious evidence of serious differences in the location and spread of the data for the two valleys and an experienced analyst might stop at this point; we shall proceed with more formal tests by way of illustration. The estimated means and variances for the log-transformed data are $\bar{x}_1 = 2.13$, $\bar{x}_1 = 2.27$, $s_1^2 = 0.128$ and $s_2^2 = 0.091$. For illustrating one-sample t-tests the data for Valley 2 is used and it is assumed that $\mathrm{H}_0 : \mu = 2.10$ is of interest[6].

[5]Since area is a squared quantity a square-root transformation is another possibility and produces results very similar to the log-transformation.

[6]This kind of assumption is sometimes motivated in texts by assuming that previous study, involving a large sample or a population, has suggested the null hypothesis.

12.3.2 t-tests

For both one- and two-sample t-tests in R the `t.test` function is used. Defaults are to carry out a two-sided test, reporting a 95% confidence interval. For tests about the equality of variances the `var.test` function is used. The following six tests were applied; the first three involve one-sample t-tests and illustrate typical output and the use of arguments to the function. The fourth and fifth examples illustrate the two versions of the two-sample t-test, and the sixth example involves the F-test.

```
t.test(valley_2, mu = 2.1)
t.test(valley_2, mu = 2.0)
t.test(valley_2, mu = 2.0, conf.level = 0.99)
t.test(valley_1, valley_2)
t.test(valley_1, valley_2, var.equal = TRUE)
var.test(valley_1, valley_2)
```

The log-transformed data for the two sites are contained in the objects `valley_1` and `valley_2`. For the one-sample tests the null hypothesis is specified using the argument `mu`; for the two-sample t-tests equality of the population means is the default null hypothesis, without assuming equality of population variances; for `var.test` equality of population variances is the default null hypothesis (these can be varied if necessary – use the help facility for details). The confidence interval can be varied using the `conf.level` argument and will be illustrated; one-sided tests are not shown but the arguments `alternative = "less"` or `alternative = "greater"` are available, depending on which of $\mu < \mu_0$ or $\mu > \mu_0$ is of interest.

For the first test shown, where $\mathrm{H}_0 : \mu = \mu_0$ with $\mu_0 = 2.1$, the following output is obtained.

```
        One Sample t-test

data:  valley_2
t = 1.8119, df = 9, p-value = 0.1034
alternative hypothesis: true mean is not equal to 2.1
95 percent confidence interval:
 2.057083 2.488317
sample estimates:
mean of x
   2.2727
```

The p-value 0.1034 is greater than 0.05 so is not significant at the 5% or even 10% level (if it is not significant at 5% it cannot be significant at 1%). The 95% confidence interval is (2.06, 2.49). Any value ouside this range will be rejected as

a null hypothesis at the 5% level; this is illustrated in the second example where $\mu_0 = 2.0$ is assumed for the null hypothesis. The following output excludes some of the information identical to that from the first analysis.

```
t = 2.861, df = 9, p-value = 0.01875
alternative hypothesis: true mean is not equal to 2
95 percent confidence interval:
 2.057083 2.488317
```

The t- and p-values change, the latter to 0.019. This can be reported by simply noting it; observing that the test is (just) significant at the 2% level; or that the null hypothesis is rejected at the 5% but not the 1% level. Information on the confidence interval does not change and is shown here precisely to emphasize this point; it does not depend on the value used for the null hypothesis. The third example is identical to the second except for the confidence level

```
99 percent confidence interval:
 1.962942 2.582458
```

where the 99% confidence interval, (1.96. 2.58) is (inevitably) wider than the 95% interval. Note that the interval contains the value 2.0, another way of showing that the null hypothesis $\mathrm{H}_0 : \mu_0 = 2.0$ would not be rejected at the 1% level of significance.

For the first of the two-sample t-tests, not assuming equal variances, the output is

```
        Welch Two Sample t-test

data:  valley_1 and valley_2
t = -1.026, df = 21.309, p-value = 0.3164
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4193637  0.1421065
sample estimates:
mean of x mean of y
 2.134071  2.272700
```

In the call to the `t.test` function separate listing of the data for the two groups as the first two arguments alerts the function to the fact that a two-sample test is intended.

The additional argument `var.equal = TRUE` in the fifth analysis specifies that equal variances are to be assumed. The differences from the results from the analysis that does not make this assumption are

```
        Two Sample t-test

data:  valley_1 and valley_2
t = -0.9959, df = 22, p-value = 0.3301
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4273128  0.1500556
```

There are minor differences in the numerical output but the conclusion that there is no evidence of a difference in population means, with p-values of more than 0.3, is the same in both cases. The 'Welch' in the heading to the output in the first analysis pays tribute to one of the scholars responsible for the theory associated with the unequal variances problem.

The results are so similar that the simpler and more widely advertized equal-variances version is clearly acceptable for reporting purposes, and implies there is no evidence to contradict this assumption. An experienced analyst familiar with F-tests would recognize this by just looking at the variance estimates. More formally, using the `var.test` function we get

```
        F test to compare two variances

data:  valley_1 and valley_2
F = 1.4132, num df = 13, denom df = 9, p-value = 0.6124
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3689233 4.6805505
sample estimates:
ratio of variances
          1.413196
```

The p-value is so 'large' that the conclusion would often be expressed – a little loosely – as 'not significant', with specific levels such as at 10%, 5% and 1% left implicit.

12.3.3 Chi-squared tests

The chi-squared test arises naturally in the context of testing hypotheses about single variances, but these are of limited interest and not discussed. The test is of greater importance for testing the hypothesis of no association between categorical variables displayed in a two-way contingency table. The left-hand side of Table 12.2 (Table 2 in VanPool *et al.*, 2000), is an example. The entries show the frequencies

	Observed				Expected				
	Room 1	Room 2	Room 3	Plaza	Room 1	Room 2	Room 3	Plaza	Total
Chert	86	21	38	97	81	23	31	106	242
Chalcedony	39	10	12	13	25	7	10	33	74
Obsidian	4	3	3	8	6	2	2	8	14
Quartzite	16	8	7	6	12	4	4	16	37
Igneous	217	63	81	353	238	69	93	314	714
Total	362	105	143	477	362	105	143	477	1085

Table 12.2: *The observed and expected frequencies of flaked stone artifacts from three room blocks and a plaza area at Galeana, a large pueblo-like settlement in north-western Mexico. (Source: Table 2 in VanPool* et al.*, 2000.)*

of flake stone artifacts, categorized by material and location, found within a pre-Hispanic settlement site in Mexico.

The chi-squared test is almost invariably covered in introductory texts so only an outline is given here, the focus being on practical application. The table is of size $I \times J$, in this instance 5×4. There are thus 20 observations or *cells* in the table with entries that will be denoted by O_{ij} for row i and column j. Another way of expressing the null hypothesis of no association between material and location is that the distribution of frequencies across location (expressed as a percentage) should be similar apart from random variation. Exploratory methods for investigating this include barplots (Chapter 4) and correspondence analysis (Chapter 9); the chi-squared test is a more formal means of investgation.

To proceed, the *expected values*, E_{ij}, under the null hypothesis are needed and these are given (rounded to integers for easier comparison) to the right-hand side of Table 12.2, by

$$\frac{\text{row total} \times \text{column total}}{\text{overall total}}.$$

The chi-squared test statistic is then calculated as

$$X^2 = \sum \frac{(O-E)^2}{E}$$

which follows the chi-squared distribution (χ^2) (approximately) with $(I-1)(J-1)$ DF if the null hypothesis is true[7]. If the null hypothesis is incorrect some of the O will differ noticeably from E so X^2 will be 'large' and a one-sided test is called for in using χ^2 to make this assessment. It is fairly obvious from tabular inspection (particularly if row percentages are used) that the distribution of materials across locations varies, so a 'significant' result rejecting the hypothesis of no association is expected.

[7]I use X^2 for the test statistic to distinguish it from the theoretical χ^2 value – this usage is not universal.

Using `chisq.test(flakes, correct = F)`, where `flakes` is the name of the data set, confirms this

```
        Pearson's Chi-squared test

data:  flakes
X-squared = 49.1023, df = 12, p-value = 2.007e-06

Warning message:
In chisq.test(flakes) : Chi-squared approximation may be incorrect
```

The small p-value leads to clear rejection of the null hypothesis of no association. This can also be obtained using `pchisq(49.103, 12, lower.tail = F)`. If a decision rule needs to be explicitly stated (as some journals require) `qchisq(0.01, 12, lower.tail = F)` gives the 1% critical value as 26.22.

Rather than relying on the p-value alone more detailed inspection of output can be helpful. Use `chisq.test(flakes, correct = F)$expected` to obtain expected values; for residuals replace `expected` with `residuals` and for standardized residuals use `stdres` instead of `expected`. The (Pearson) residuals are defined as $(O-E)/\sqrt{E}$ and the standardized residuals as $(O-E)/s$ where s is an estimate of the standard deviation of $(O-E)$.

Values for the residuals, with and without standardization, are given in Table 12.3. As a rule-of-thumb, absolute values of the standardized residuals in excess of 2 draw attention to cells where departure from the null hypothesis is most obvious. Several such values stand out in the table; these are mostly associated with the 'Chalcedony' and 'Igneous' categories, the 'Plaza' also standing out. Examining Table 12.2 shows that 'Chalcedony' is under-represented in the Plaza and over-represented in the rooms compared to the predictions of the hypothesis of no association; the opposite is true for the 'Igneous' material.

	Residuals				Standardized residuals			
	Room 1	Room 2	Room 3	Plaza	Room 1	Room 2	Room 3	Plaza
Chert	0.59	-0.5	1.17	-0.91	0.81	-0.6	1.42	-1.38
Chalcedony	2.88	1.06	0.77	-3.42	3.66	1.16	0.85	-4.74
Obsidian	-0.82	0.95	0.43	0.03	-1.01	1.01	0.47	0.04
Quartzite	1.04	2.34	1	-2.55	1.3	2.5	1.09	-3.46
Igneous	-1.37	-0.73	-1.22	2.21	-2.88	-1.32	-2.24	5.04

Table 12.3: *Residuals, raw and standardized, from a chi-squared test of the data in Table 12.2.*

The warning message alerts you to the fact that some of the expected values are small (less than 5) which may invalidate the approximation of the distribution of the test statistic to χ^2. The sample size for obsidian of 14 is quite small and

two of the expected values are noticeably less than 5. It is not a problem here; the cells affected do not contribute greatly to the highly significant value of X^2 as the residuals in Table 12.3 show. This is not an inevitable outcome and small expected values can be particularly problematic with small overall sample sizes[8].

Small samples can be dealt with in a variety of ways. The example so far used the argument `correct = F` in the call to `chisq.test`, which calculates X^2 as defined. The default for 2×2 tables is actually `correct = T` which applies Yates's *continuity correction*, replacing $(O-E)$ in the definition with $|O-E|-0.5$ with the intention of improving the approximation to χ^2. Another alternative, particularly useful for 2×2 tables, is the *Fiser exact test* the details of which are not entered into here. For illustrative purposes the data from the rows for 'Chert' and 'Igneous' and columns for 'Room 2' and 'Room 3' from Table 12.2, divided by 10 and rounded, will be used, giving

	Room 2	Room 3	Total
Chert	4	8	12
Igneous	12	16	28
Total	16	24	40

Table 12.4: *Artificial data. The observed frequencies of flaked stone artifacts have been adapted from Table 12.2 using two rows and columns only and rescaling by dividing by 10.*

The p-values for the chi-squared test with and without the continuity correction gives p-values of 0.83 and 0.57, both tests issuing a warning about the validity of the chi-squared approximation. The Fisher test, implemented with the `fisher.test` function, `fisher.test(flakes)`, returns a p-value of 0.67. None of these values are significant at the levels usually employed, but the example is sufficient to demonstrate that they can give rise to different p-values so can potentially lead to different conclusions about the null hypothesis.

It is difficult to advise on which test is 'best'; opinions differ. For something as apparently simple as a 2×2 table there is a considerable statistical literature on the subject[9]. In `R` it is possible to use simulation to obtain a p-value without assuming a chi-square distribution with the `simulation.p.value = T` argument.

[8]Some of the test statistics described so far are obviously dependent on sample size and increase as n gets larger, resulting eventually and inevitably in the rejection of any null hypothesis. The same is true, less obviously, for X^2. If the observed values of O are scaled by a factor of k the expected values follow suit and $X^2 \rightarrow kX^2$. An implication of this is that such tests are of limited value for very large samples, the 'significance' of any differences being a matter of expert judgment rather than formal testing.

[9]Some of this is to do with practical performance; some to do with theoretical matters con-

For the 5×4 Table 12.2 a p-value is 0.0005, which is different from that previously obtained; for the 2×2 example with and without continuity correction the p-values are 0.72 and 0.74. If simulation is used, continuity correction is only available for the 2×2 case; the p-value will vary slightly if simulation is repeated, hence the use of the term 'a p-value'.

12.3.4 F-tests and ANOVA

Analysis of variance (ANOVA) techniques are important in a wide range of statistical applications. They are not mentioned much, if at all, in quantitative archaeology texts (VanPool and Leonard, 2010, is an exception), possibly because the applications are often 'complex' enough to be beyond the ambitions of such texts. It is also the case that applications in the archaeological literature are not profuse (though this is also true of less complex tests that are accorded space).

The general idea underpinning ANOVA is discussed briefly; the only application considered in any detail is the problem of comparing more than two sample means using one-way ANOVA. Models for data, where ANOVA is relevant, typically assume that the data can be modeled as the sum of systematic and random (or error) components, that is

$$\text{Data} = \text{Systematic} + \text{Random}$$

with a focus on whether or not the systematic component is, in some sense, 'important' compared to the random component. In general the systematic element can itself consist of component parts, and interest may focus on whether or not only a subset of these are required to explain variation in the data.

To test this, the total variation in the data is broken down into the contributions of systematic variation and random variation or sums of squares (SS); the SS are converted to variances by division by the appropriate degrees of freedom so MS = SS/DF where MS is the *mean square* – call these MS_S and MS_R for the systematic and random components – and their ratio is calculated as

$$\text{F} = \text{MS}_S/\text{MS}_R$$

where F is an F-statistic with degrees of freedom determined by the context. Under the null hypothesis, which will also depend on context, and assuming normality of the error term this should follow an F-distribution and the approach outlined for

cerning 'experimental design'. Commonly Fisher's test has been recommended when observed values are small, but it has also been noticed that it often performs similarly to a chi-squared test with a continuity correction. A preference for chi-squared without the continuity correction is possibly a minority opinion among statisticians but sometimes the only version presented in introductory texts. The continuity-corrected version is the default in `R` but other software can differ.

comparing two sample variances in Section 12.2.4 and at the end of Section 12.3.2 can be used.

To show what R output looks like, we revisit the two-sample t-test of Sections 12.2.3 and 12.3.2, reformulated as an ANOVA problem. The log-transformed area data for both valleys need to be stacked as a single column of data (`log_area`), with a second column of equal length providing information on site location (`valley`).

Different ways of conducting the ANOVA are available of which the function `oneway.test` is simplest. The test may be carried out with or without assuming equal error variances within samples. The latter is the default. In the following output the last two lines were obtained using `var.equal = TRUE` in the function call. The p-values and their DF are identical to those from the `t.test` analysis.

```
oneway.test(log_area ~ valley, var.equal = FALSE)

        One-way analysis of means (not assuming equal variances)
data:  log_area and valley
F = 1.0527, num df = 1.000, denom df = 21.309, p-value = 0.3164

# Using var.equal = TRUE
        One-way analysis of means
F = 0.9918, num df = 1, denom df = 22, p-value = 0.3301
```

The more general problem of comparing $p > 2$ means involves the null hypothesis

$$\mathrm{H}_0 \colon \mu_1 = \mu_2 = \ldots = \mu_p.$$

For a second example data on the maximum flake length (mm) of 10 unbroken flakes for each of four raw material types from Cerro del Diablo, a Late Archaic Mexican site, given in Table 10.1 of VanPool and Leonard (2010), are reproduced in Table 12.5.

Assuming these can be treated as random samples for the four material types the null hypothesis is $\mathrm{H}_0 \colon \mu_1 = \mu_2 = \mu_3 = \mu_4$ (i.e. the mean population lengths are the same) [10]. It is obvious that H_0 will be rejected (compare the lengths for obsidian and rhyolite). A formal test of this is not really needed; the comparison

[10] It was noted in the introduction that VanPool and Leonard (2010) deal with hypothesis tests at greater length than competing texts. It was also suggested that the treatment is flawed, and this is an apposite point at which to elaborate. The fundamental problem is that, having alerted the reader to the importance of distinguishing between population and sample quantities, the authors ignore their own advice. The treatment of the way null hypotheses are expressed is wrong. Thus, in the context of two-sample t-tests, the usual null hypothesis that the population means are the same, $\mathrm{H}_0 \colon \mu_1 - \mu_2 = 0$ in the example, is expressed incorrectly in terms of equality of the *sample means*, $\mathrm{H}_0 \colon \bar{X}_1 - \bar{X}_2 = 0$. This is a statement that the sample means are the same; the observed data are used to test hypotheses about unknown population quantities and it does

Chert	Obsidian	Rhyolite	Silicified wood
41	30	135	113
110	53	141	111
73	45	138	97
52	34	175	70
176	105	143	117
61	102	132	48
69	51	130	134
40	47	109	115
64	71	125	103
48	58	120	106

Table 12.5: *The maximum flake length (mm) of unbroken flakes for four raw material types (Source: Table 10.1 in Van Pool and Leonard, 2010.)*

between pairs of material types is of more interest and can be achieved using the `aov` function, which is more general than `oneway.test`.

In the first instance it is sensible to look at the data graphically, which can be done using boxplots as in Figure 12.2. That there are differences between material types, implying that the null hypothesis will be rejected, is obvious. Flakes made of rhyolite and silicified wood clearly tend to be longer than those of chert and obsidian; it is less clear if chert and obsidian differ significantly (though probably not) or if rhyolite and wood differ. There are some clear outliers within material types that will be temporarily ignored in order to illustrate the mechanics of application.

The basic analysis is as follows. The data are held in a data frame `aov.data` with the columns labeled `length` and `material`. It is obvious from the p-values (and the cues provided in the output) that the null hypothesis is comprehensively rejected.

```
aov.example <- aov(length ~ material, data = aov.data)
summary(aov.example)
            Df Sum Sq Mean Sq F value    Pr(>F)
material     3  33156 11051.9  13.373 5.085e-06 ***
Residuals   36  29751   826.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

not makes sense to express the null hypothesis about unknowns in terms of known sample values. This notational problem, arguably a conceptual one as well, is pervasive in the several chapters dealing with hypothesis testing problems, including those on ANOVA and chi-squared tests.

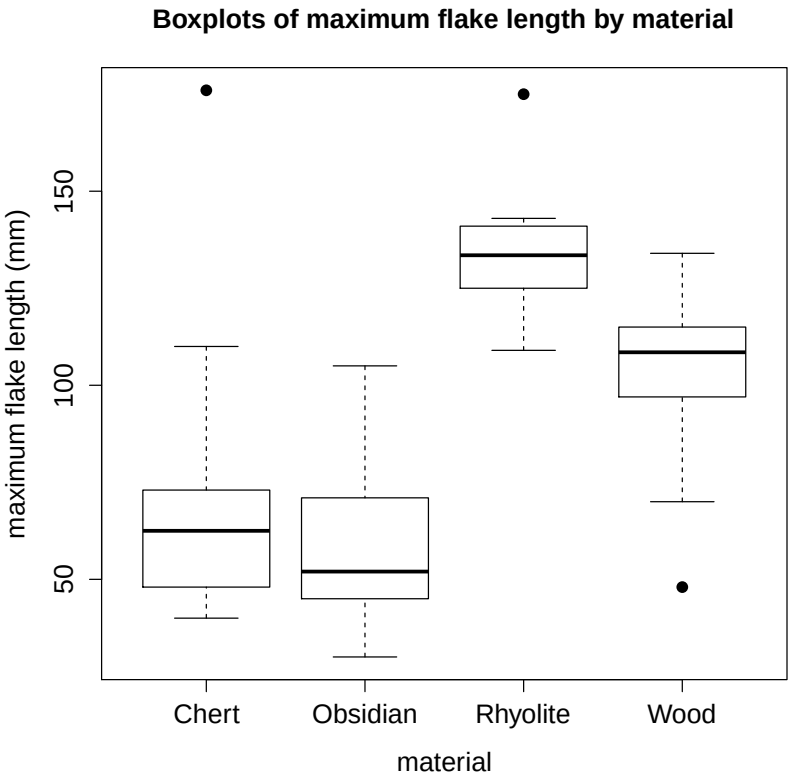

Figure 12.2: *Boxplots of maximum flake lengths by material using the data of Table 12.5.*

To investigate which material types differ in their typical length in a formal way the obvious thing to do is to undertake all possible pairwise comparisons; this involves a *multiple comparison test*. The two-sample t-test is a candidate for such testing but inflates the probability of finding a significant difference. There are different approaches to 'correcting' for this; a popular one is Tukey's HSD (Honestly Significant Difference) test which, given the fitted `aov` model, can be implemented using the `TukeyHSD` function so `TukeyHSD(aov.example)` returns

```
$material
                    diff        lwr        upr     p adj
Obsidian-Chert     -13.8 -48.424703  20.824703 0.7076807
Rhyolite-Chert      61.4  26.775297  96.024703 0.0001685
Wood-Chert          28.0  -6.624703  62.624703 0.1488159
Rhyolite-Obsidian  75.2  40.575297 109.824703 0.0000064
Wood-Obsidian       41.8   7.175297  76.424703 0.0127440
Wood-Rhyolite      -33.4 -68.024703   1.224703 0.0620233
```

Here `diff` is the difference in means and `lwr` and `upr` are lower and upper values of a 95% confidence interval (the default). Superficially the p-values do not entirely concur with the expectations raised by preliminary data analysis. In particular, the wood-chert comparison is not significant at the 10% level, the wood-obsidian comparison is significant at the 5% but not the 1% level, and the wood-rhyolite comparison, with the least 'predictable' outcome, is not significant at the 5% level.

The term 'superficially' is used because the ANOVA assumes that the samples within material types are from a normal distribution, and equal variances are assumed. So far nothing has been done about the outliers, that for chert at 176 mm being particularly prominent. It is possible to test for equal variances using `bartlett.test(length ~ material, data = aov.data)` which gives a p-value of 0.09 so, if a decision rule of 5% is used, the variances do not differ significantly at this level. The problem here is that outliers will inflate the variances, possibly quite considerably, and their effect on the test outcome is unpredictable.

If nothing else, the outliers also call into question the normality assumption. This can, if one wishes, be tested for formally. Many such tests have been proposed a large number of which are rarely used, if at all; the Shapiro-Wilk test, implemented with the `shapiro.test` function in `R`, is widely-regarded as one of the best. For chert, the first ten values in the stacked data set, produces a p-value of 0.005 with `shapiro.test(aov.data$length[1:10])` which is very strong evidence against the normality assumption; repeating this for other materials gives non-significant results at the 10% level, so chert is the main problem.

In the absence of archaeological knowledge that might dictate that the chert outlier be treated separately, the sensible thing is to omit it to see if it affects substantive conclusions. The Bartlett test now gives a p-value of 0.68 – much larger than that originally used, so no hint that the equal variances assumption may be wrong. For Tukey's HSD test all the p-values change as follows.

```
$material
                    diff       lwr        upr     p adj
Obsidian-Chert      -2.4 -30.54352  25.743516 0.9956328
Rhyolite-Chert      72.8  44.65648 100.943516 0.0000002
Wood-Chert          39.4  11.25648  67.543516 0.0031873
Rhyolite-Obsidian   75.2  47.80711 102.592887 0.0000001
Wood-Obsidian       41.8  14.40711  69.192887 0.0012254
Wood-Rhyolite      -33.4 -60.79289  -6.007113 0.0117777
```

Other than the obsidian-chert comparison the p-values are reduced; most noticeably the wood-chert comparison is now significant at the usual levels compared to its previous non-significance, and wood-rhyolite now differs significantly at the 5% level.

If the more radical, and possibly less justifiable, omission of all three outliers suggested by Figure 12.2 is contemplated there is relatively little change, the only one worth noting being the fact that the wood-rhyolite comparison is now only significant at the 9% level. This is not surprising since outliers that inflate the mean for rhyolite and deflate it for wood are now omitted, so the mean difference is reduced by about 10 mm to 20 mm.

This mixture of formal and informal analysis shows that there are highly significant differences between length for material types; you don't need the formal testing to arrive at this conclusion. An outlier apart you can accept, quite happily, a conclusion that chert and obsidian could be sampled from populations with the same mean length, and the boxplots show that their dispersion is also similar. All other pairwise comparisons, silicified wood and rhyolite apart, suggest highly *statistically significant* differences. Conclusions about the wood-rhyolite comparison are equivocal – it depends on the attitude towards outliers, but the evidence for a difference is not overwhelmingly significant whatever treatment is adopted. The more important issue is whether or not the observed differences matter much in terms of the archaeological aims driving the analysis. Is a difference of 20–30 mm in the mean length of silified wood and rhyolite of any consequence, regardless of statistical significance. If not, there is little point in worrying about the statistical significance and you are spared the effort of further data collection, necessary if differences of this magnitude are potentially important and you want to assert that they are 'real'.

12.4 Some omitted topics

As already noted, this chapter misses out a considerable amount. The focus has been on 'classical' statistical inference, the 'objectivity' and 'scientific rigor' of which first attracted the New Archaeologists and which, at an introductory level, is the approach most archaeologists will have been exposed to in texts on quantitative methodology currently in use.

Arguments about 'theory' in statistics have raged as much, and as fiercely, as they have have in archaeology. Competing 'theories', of which Bayesian inference is the most prominent, reject much of the conceptual machinery that underpins classical theory, in the way that probability is to be understood, for example, and how data should be interrogated and inferences drawn from them.

Bayesian thinking had its early advocates in archaeology (Cowgill, 1977b: 361–62; Orton, 1980: 220; Orton, 1992: 139) but, except at a basic level, very little was done about exploring the methodology. There is a good reason why this state of affairs arose, which is that for the practical application of Bayesian ideas the necessary computational power needed to be developed. This happened; Buck

et al. (1996) was the first book-length treatment of the *Bayesian Approach to Interpreting Archaeological Data* (to give the book its full title) and remains the standard text. Nevertheless – with one major exception – Bayesian methods have not yet come to be routinely used in statistical analyses of archaeological data.

The major exception – and it is difficult to over-emphasize its importance – is in application to dating problems, and especially the calibration of radiocarbon dates and their interpretation. Without going into detail, the software in common use to provide these dates typically depends on Bayesian calculations, even if the user does not always appreciate this. As examples of what has been achieved, recent programmes of dating using Bayesian methods have produced important revisions of the previously accepted chronology for the British Neolithic and Anglo-Saxon periods (Whittle *et al.*, 2011; Bayliss *et al.*, 2013; see, also, Section 9.5). Despite its ubiquity there is the suspicion that not all 'consumers' of radiocarbon dates fully understand how they are to be interpreted, notwithstanding the literature that exists to explain this; it is something that might usefully be included in future texts on quantitative methodology.

Within the 'classical' paradigm more could have been said about significance testing in the context of regression models – noted briefly in Section 5.1.1 and 5.1.4 and with a more detailed illustration in the third example of Section 5.2 – and more complex ANOVA models (though these have had little archaeological use). Other modeling methodologies, within the class of generalized linear models that depend on inferential ideas and have attracted some archaeological use, have also not been discussed. These include log-linear models (Lewis, 1986; Shennan, 1997: 201-13; Baxter, 2003: 131–36) and logistic regression (Baxter, 2003: 60–62, 162–63).

Another topic omitted – on grounds of length rather than complexity – is that of non-parametric hypothesis testing methods. Such methods do not assume an underlying probability distribution for the sampled population, removing the dependency on the normality assumption. Although parametric tests, such as the t-test, are more powerful if the normality assumption is valid, non-parametric tests that can be used as an alternative can also be competitive in terms of power and would seem to be an attractive alternative when normality is in doubt.

Given this, I am a little surprised at the relative lack of space given to such methods in the standard quantitative archaeology textbooks. For those wishing to explore this further the `R` function `wilcox.test` does a similar job to `t-test` using the Wilcoxon one- or two-sample tests (the latter also known as the Mann-Whitney test). The Kruskal-Wallis test, `kruskal.test`, is the non-parametric analog of `oneway.test`.

12.5 Discussion

The question that motivated this chapter was 'how useful are the 'classical' methods of statistical inference for archaeological purposes?'. I take it as axiomatic that the ideas and methodologies of statistical inference are important. On the 'horses for courses' principle, however, it is not axiomatic that the methodologies as originally developed are equally applicable to different domains of study.

There are at least two aspects involved; one concerns the correct use of methodology and the other its usefulness. Much of the critical commentary that emerged in the 1970s and 1980s falls into the former category. Cowgill (1977b), for example, identified several areas of mis-use and mis-understandings and offered corrective advice. His appraisal of the value of statistical inferential methods in archaeology was less negative than that of, for example, Doran and Hodson (1975).

In a spirit of what might seem deliberate provocation, Cowgill used the adjectives 'mind-boggling' and 'riduculous' to characterize *some* uses of significance testing. He commented (Cowgill, 1977b: 365) that it 'seems so much more useful that it seems incredible that the estimation approach [i.e. confidence intervals] is not used more often [than significance] tests' – a still pertinent view, endorsed in this chapter. The emphasis on significance testing at the expense of estimation was attributed to 'tradition' in the social sciences, and an 'uncritical' acceptance of the hypothesis testing framework. This, also, remains pertinent[11].

The view that significance testing is frequently uninteresting and not of much use has already been expressed. It is 'usefulness' that I would use as the main, and pragmatic, criterion for judging the merits, in practice, of any particular method of data analysis. Judgments need to be divorced from any irritation with abuse of methodology, and not confused with 'philospophical stances' that involve the wholesale rejection of 'scientific' methodology.

Generalization should be approached with trepidation. Mine would be that, with some exceptions, I have been struck, whenever I have reviewed the literature, by the paucity of widespread and what I'd regard as convincing uses of the methods discussed in this chapter[12]. Some methods have been used hardly at all, or not to the extent you might assume from their textbook treatments.

For example, VanPool and Leonard (2010: Chapter 10) observe that ANOVA,

[11] As an aside, at the time of writing, the journal *Basic and Applied Social Psychology* has just 'banned' the use of the null hypothesis significance test (and confidence intervals) from its pages on the grounds that they are 'invalid'. What is meant by 'invalid' is not very clear; it seems to stem more from concerns about the misuse and misinterpretation of significance tests, a concern many statisticians share.

[12] An exception that springs to mind is sampling theory, where random sampling (of regions using test-pits, for example) leads naturally to the use of standard inferential ideas. This is not typical of the manner in which archaeological 'samples' are usually acquired. The subject is quite a specialized one; Orton (2000) is a thorough and accessible account for archaeologists.

one of the 'most powerful tools in the statistician's toolkit', 'hasn't been applied as widely as it deserves to be in archaeological analysis'. Two-sample t-tests are simpler but still not extensively used; graphical methods will often make it clear whether there are substantively important differences or not, obviating the need for formal tests. Should the data merit further scrutiny after graphical analysis then, even if one has no qualms about the assumptions involved, sample-size effect need to be borne in mind. With large enough samples any difference, however small, will be found to be significant. The main merit of testing is for small samples where a non-significant result guards against reading too much into apparent differences.

None of these observations are new; commentators such as Cowgill (1997b) were saying this kind of thing from an early stage. I would add that it is probably difficult to put together a convincing collection of case-studies based on the t-test that lead to insights not more readily attainable by other means. The same is true, only more so, of one-sample tests. It is difficult to think of examples, particularly those in textbooks, that are other than illustrative and artificial.

Chi-squared tests for no association in contingency tables have probably been more widely used than tests of means. Analyses often do not go beyond reporting a statistically significant association or its lack. If detection of 'significance' is the sole reason for analyzing a table it is tempting to suggest that much can be achieved by judicious tabular inspection. This includes the scaling of rows or columns to percentages, with the ordering chosen to highlight similarities or differences where there is not a natural ordering. This is, of course, done, but not necessarily consistently. The suspicion exists that, despite the ubiquity of tabular presentation in archaeology, the widespread use of ill-chosen `Excel` bar- and pie-charts in preference to such direct interpretation (Sections 4.2 and 4.3) testifies to a certain discomfort with the latter.

The last few paragraphs express a view about the value of the standard and 'classical' methods of statistical methods for archaeological purposes. The importance of the theory *per se* is unquestionable, as is the beneficial impact it has had in many areas of practical application. Whether archaeology is one of these areas is the interesting question.