NRES 798 — Statistical Methods for Ecologists
Chapter 5: Frameworks for Statistical Analysis

Oscar García

March 1, 2013

Contents

1 Frameworks

Because of time limitations we will focus on parametric classical statistics. This is by far the approach most commonly used in scientific reporting.

The Bayesian approach has already been discussed in the *Introduction* and following notes. In hypothesis testing it works in a very similar way to that of credibility intervals.

Nonparametric or *distribution-free* methods do not assume any particular population distribution. They rely on properties of the empirical distribution that are always valid. For instance, on test statistics based on ranks or other functions of the ordered observations, the so-called *order statistics*. Nonparametric Statistics includes *randomization* or *permutaion tests*, called Monte Carlo analysis in the textbook. Another example is the histogram, which is a simple nonparametric estimate of a population distribution.

Nonparametric methods are more *robust* than parametric ones, in that they are not affected by deviations from an assumed distribution. This is specially important in small samples,in large samples the Central Limit Theorem tends to make distributional assumptions more plausible. Many nonparametric tests are also computationally simpler. On the other hand, the tests are less powerful, and estimators less precise, provided that the parametric model is reasonable.

2 Example, models, sampling

The running example in this chapter is the nest density of ants in forest and field habitats (number per unit area, not probability!). The data, stored in a data frame in *R*, is:

```
> ants <- data.frame(habitat = c(rep(’forest’,6), rep(’field’,4)),
+ nests = c(9,6,4,6,7,10,12,9,12,10))
> ants
   habitat nests
1   forest     9
2   forest     6
3   forest     4
4   forest     6
5   forest     7
6   forest    10
7    field    12
8    field     9
9    field    12
10   field    10
>
```

Two ideas previously discussed are important, and should avoid being confused by irrelevant details (irrelevant to the statistical analysis):

First, we are dealing with a *model*, a simplified representation of some aspects of a real situation. For purposes of the current analysis, the relevant part of the model is a population of nest counts on an infinite number of possible quadrants. Specifically, a probability distribution of such counts, or possibly one distribution for *Forest* and another one for *Field*. Any conclusions from the analysis apply to this model. If the model is “close enough” to reality, we may be confident in applying the results to the real thing. It has been assumed that the only relevant factors are the two habitats, however defined, ignoring any seasonal variation, spatial gradients, etc. That may or may not be good enough in practice; but that is an issue of scientific modelling, not statistical analysis.

Second, it is assumed that the sample is “representative” and suitably “randomized”, obtained by simple random sampling. It applies to *this particular site*. Precisely, what this mean is that we assume that the observations are independently distributed, each with the same probability distribution as the population.

In any case, the first step should be to produce summaries and graphs in order to: (a) check for possible data errors or outliers, and (b) assess if the model assumptions appear reasonable.

3 Parametric analysis

Hypothesis testing follows the steps already discussed in the notes to the previous chapter:

1. Formulate the null hypothesis. H_0: the population has a normal distribution $N(\mu, \sigma)$, irrespective of habitat. The alternative is that *Forest* and *Field* have normal distributions with the same σ but different means[1].

2. Choose a test statistic. The F ratio, which is the ratio of a measure of variance between groups (habitats) to the variance within groups (details in Chapter 10)[2]. If there is a common distribution (H_0), F should be around 1. If the means differ, the ratio should be greater than 1.

3. Ascertain the sampling distribution of the test statistic (under H_0). It is known that if H_0 is true then F has the F distribution. If the number

[1] Actually, we know from Chapter 2 that this kind of data is more likely closer to a Poisson than to a normal distribution. And in the Poisson the variance equals the mean (both are equal to the rate parameter λ). A normal approximation might not be all that good in samples of this size. The square root transformation of Poisson variables is closer to the normal, and mean and variance are then more independent. It might be better then to do the test using the square root of the nest counts instead of the counts themselves.

[2] For two groups, like here, a common alternative test statistic is the *t-statistic*, the difference of means divided by the standard deviation. There are also versions of the F and t tests that do not assume equal variance.

of groups is m, here 2, the distribution parameters are $m - 1 = 1$ and $n - m = 8$.

4. Test. The calculated F is 8.78 (do not confuse the test statistic F with the CDF). Either

   (a) Traditional method, compare to critical values for given α. The critical value for $\alpha = 0.05$ is a value c such that $P\{F > c\} = 0.05$, or $P\{F \leq c\} = 0.95$. That is, the 0.95-quantil. It can be found from statistical tables, or computed with the quantil function $F_F^{-1}(0.95) = 5.32$. In *R* this is `qf(0.95, 1, 8)`. Similarly, the critical value for $\alpha = 0.01$ is $F_F^{-1}(0.99) = 11.26$. We reject H_0 at the 0.05 level, but not at the 0.01 level. The difference between the number of ant nests in forests and in fields is *statistically significant*, $F = 8.78^*$ (one star).

   (b) Reporting a p-value. Find $P\{F > 8.78\} = 1 - F_F(8.78)$ (suitable computer software required). In *R*, `1 - pf(8.78, 1, 8)` gives the p-value 0.018. This is less than 0.05 but greater than 0.01. Same conclusion.

In *R* the whole thing can be done by `oneway.test(nests ~ habitat, ants, var.equal=T)`.

4 A nonparametric test

To get the flavour, let's see the *sign test*, a simple nonparametric hypothesis test. It uses paired data, that is, compares pairs of observations within experimental units, or within (hopefully relatively homogeneous) groups. Assume that in the ant nests problem one has counted ant nests in one field quadrat and one forest quadrat in each of 6 different sites:

site	field	forest
1	13	7
2	8	3
3	5	7
4	14	7
5	3	10
6	7	3

H_0 says that field and forest observations within each site belong to the same population. Or at least that they have the same median. There may be differences between sites, but not between field and forest within a site. The alternative is that the number of nests tends to be higher in fields than in forests.

Compare the counts for each site. *Field* is larger than *forest* in 4 out of 6 cases. How likely would this be if H_0 were true? Ignoring ties, the probability of field $>$ forest would be 0.5, the same as that of field $<$ forest. The observations are independent. Therefore, the probability of obtaining m "$>$" out of n would be a binomial with parameters n and 0.5. The probability of 4 or more out of 6, the p-value, is `1 - pbinom(3, 6, 0.5)` $= 0.34$. Not all that unlikely, so H_0 is not rejected.

Typically this would be tested with a paired t-test, which has higher power (you can verify that here it gives $p = 0.19$). Or a two-way ANOVA. But these assumes normality. The Wilcoxon test is also nonparametric, and is more powerful than the sign test. But the sign test is simpler, and it can be used even if there are no numerical measurements, only comparisons, for instance, "A is prettier than B". In the ants study, it might be possibly to estimate visually if there are more nests in field than in forest, perhaps using photographs, without making any measurements.