NRES 798 — Statistical Methods for Ecologists
Chapter 4: Framing and Testing Hypotheses

Oscar García

February 24, 2013

Contents

1 Scientific Method

All this is controversial, and opinions vary. It is useful to be aware of the various views anyway, maybe where your journal reviewer or editor is coming from. See the first part of the notes, *Statistics: Introduction and overview*, for a similar although slightly different account.

2 Testing Statistical Hypothesis

See Section 5 in *Statistics: Introduction and overview* for a short description. More detail follows.

2.1 Formulation

Let's use the previous example of the dispersion of seedlings in a field. A simple model is the Poisson point process, where seedlings are located "at random", independently and homogeneously (constant rate λ). We might suspect that this is not realistic, that in fact the locations tend to be clumped in some way. We count seedlings in a number of quadrats, and calculate the coefficient of dispersion $s^2/\overline{Y}$ (Chapter 3 notes, section 4.5). In the Poisson model both the population variance and the mean are λ, so that the coefficient should be around 1. A value greater than 1 would indicate clumping. But we could also obtain such a value by pure chance even if there is no clumping. One can find a threshold or critical value such that the probability of finding a coefficient larger than that is small, say less than 5%. If the observed coefficient is larger than that, we may take it as evidence of clumping. In other words, we reject the hypothesis of a Poisson process as unlikely.

That is basically it. All (classical) hypothesis testing follows this reasoning. The steps are:

1. Dream-up a simple model, the *null hypothesis* H_0, usually what we would like to prove wrong[1]. Null: "nothing interesting". In this case a "random" or Poisson dispersion. It is the opposite of the *alternative hypothesis*, what we might suspect to be the truth (clumping, "something going on"). We cannot actually prove that H_0 is wrong, but if we are lucky we might be able to show that it is "highly unlikely".

2. Choose a *test statistic*, a statistic (function of the sample) that can discriminate between H_0 and the alternative. The coefficient of dispersion in the example[2].

3. Find the distribution of the test statistic, based on H_0 (the *sampling distribution*, remember?).

4. Choose a suitably small *significance level*[3] α. For instance, $\alpha = 0.05$. Partition the possible values of the test statistic into a *rejection region* or *critical region* for which the probability (assuming H_0) is α, and an *acceptance region* having probability $1 - \alpha$. In the example there is a

[1] Not allways. E.g., one might test for normality before further analysis that depends on that assumption.

[2] This is a hypothetical example, other test statistics might be better.

[3] Also called a *p-value*, although see Section 2.2 for a possible distinction.

critical value c such that $P\{s^2/\overline{Y} > c\} = 0.05$, and $P\{s^2/\overline{Y} \leq c\} = 0.95$.

5. If the observed test statistic lies in the rejection region, then "reject H_0", otherwise "accept H_0". "There must be something" *vs.* "maybe not".

Observations:

- This roundabout approach has a big advantage: only a simple hypothetical model for H_0 is needed, a complicated model for a realistic alternative is not required. In the example, the Poisson process model is relatively simple, a model describing clumping would be much more complicated. In fact, we do not even need to be very explicit about the alternative, exactly what kind of clumping are we talking about. The alternative can be just "not H_0".
- Most of the time the acceptance region is a simple interval, and the rejection region consists of either one or two intervals. With the alternative of *clumping*, each region is a simple interval, we are ignoring the possibility of the seedlings having a spatial pattern more regular than random. It is a *one-sided* or *one-tailed* test. If the alternative were "either clumping or uniformity", we would need two critical values, and H_0 would be rejected if the coefficient of dispersion is either too large or too small: a *two-sided* or *two tails* test.
- This is essentially the same as a confidence interval for the test statistic. Most of what was said before about those, and the lab practice, apply to hypothesis testing too.
- Most likely, you will always be using standard tests, previously developed for certain typical situations. The test statistic is given, and the distributions are already worked out (t-, χ^2, or F distributions). All you need to do is to calculate or look-up critical values or p-values[4].
- The probability of acceptance/rejection depends on sample size, besides population variability and other things. Larger samples may succeed in rejecting H_0, where a smaller sample failed. Note however that the significance level for a sequence of tests is not the same anymore. On average, one in twenty studies will reject a true H_0 at the 0.05 level. Often the one that gets published!

[4] Watch out for some terminology mix-up in the textbook: *critical value* is used for critical or significance levels, aka p-values, e.g., on page 97.

2.2 P-values, reporting

For better or for worse, most scientific reporting has standardized on significance levels (p-values) of 0.05 (1-in-20, “statistically significant”, one *), and 0.01 (“very” or “highly” significant, two **). This is the same for hypothesis testing and for confidence intervals.

These particular levels are partly a carry-over from the time of statistical tables (not so long ago). Critical values could only be practically tabulated for a few significance levels, and those two were standard. The way that works is to compare the observed test statistic to the tabulated critical value for the given level. It gets a bit messier with double-sided tests, more on that later.

More recently, it has become accepted/fashionable in some journals to report instead the p-value calculated for the observed test statistic. This can then be compared to 0.05, 0.01, or any other level. The practice is made feasible by suitable statistical software. In this case we might consider *p-value* not to be synonimous with *significance level*.

2.3 Type I and type II errors

It is conventional to define two possible types of error in hypothesis testing:

Type I: Reject H_0 when H_0 is true. This is what we have mainly been discussing. The probability of committing this error is the significance level α. It is like convicting the innocent.

Type II: Accept H_0 when H_0 is false. The guilty escapes conviction. I mentioned that this can happen if the sample is not large enough. The probability of this *not* happening is called the *power* of the test, denoted by β. In most cases it is difficult to calculate. Statisticians try to devise tests that are powerful. Simple users cannot do much about it, apart from gathering more measurements. Good to be aware of these errors, but power calculations may not be of much use to them, except perhaps when planning certain kinds of experiments.

3 Hypothesis testing, estimation, prediction

Very frequently, testing hypothesis produces the right answer to the wrong question. One may not be interested in showing that there is *some* effect, but rather in assessing how large that effect might be. In publications it is common to see an analysis of variance, for instance, where the relevant question is one of estimation. Still, a hypothesis test might be a reasonable first step, no point in estimating a more complex model if a simple one will do.

In some situations the parameters of a model may have an intrinsic interest, they may have an important interpretation. More commonly, parameter estimation is an intermediate step to prediction, using the model to estimate the values of future observations.

I CAN'T BELIEVE SCHOOLS ARE STILL TEACHING KIDS ABOUT THE NULL HYPOTHESIS.
I REMEMBER READING A BIG STUDY THAT CONCLUSIVELY DISPROVED IT *YEARS* AGO.

http://xkcd.com/892/