ORIGIN ≡ 0

The Formal Logic of Statistical Tests

The biological literature is full of scientific research papers in which the data are presumably random samples of larger populations. From these, sample descriptive statistics are calculated and summarized. The authors then proceed to advance one or more *hypotheses* concerning the problem under study. From this, usually in the Results section or associated tables, these hypotheses or related *derived statistics* are judged either to be statitically *significant* or *insignificant,* and often Probability values and/or confidence intervals are reported. All of this, regarding hypotheses, significance and confidence intervals, falls under the rubric of *Inferential Statistics.*

As an associate editor of a major journal and frequent reviewer, I very often receive papers to evaluate that include inferential statistics. It is depressingly common to see results summarized in an incoherent fashion. Usually, incompletely labeled tables are presented that are strikingly similar to the output of one or another statistical "black box" with significance levels indicated by ** etc. However, it remains unclear just what the author(s) had in mind, or just what conclusions they or the reader are supposed to draw from the output. In reading the Material & Methods section of the paper, these authors are often very precise about the software utilized (e.g., SPSS vers. xxx, such-and-such a procedure with whatever options chosen) but frustratingly vague about WHY a particular technique was chosen given their data, or WHAT their statistical hypotheses might have been, or even HOW the results derived from the "black box" relate to the conclusions they are trying to draw. Sometimes I come the the conclusion that the authors know what they are doing but are simply unclear in their presentation. In other instances, however, the authors are definitely relying too much on the "black box" to do the thinking for them. (As an aside, I tend to have fewer problems of this type with authors who use R. My guess is that in order to use R, one usually has to spend a little more time learning proper statistical technique...)

In conducting inferential statistics in biological research, therefore, it is very important to be consider carefully and be explicit about the logic of what one is doing, and to provide readers of your papers with sufficient information that they can fill in the gaps where necessary. Most textbooks in statistics present this logic reasonably well at least the first time encountered in the book. Many, including Zar, become a little sloppy thereafter because they assume that they have already told you how the logic works (which they have) and are subsequently trying to add new issues into the mix along with, perhaps, an intuitive rationale. Also, in the case of Zar, the author is attempting to be comprehensive, necessitating brevity within the extended narrative of the book.

In my opinion, biologists conducing statistical analysis have the following multi-part problem:

1). First, one must state clearly just what biological hypothesis, or hypotheses (one at a time), constitute the subject of study. Such hypotheses must be independent, preferably stated prior to, data collection.

2). Given the biological hypothesis, one must find an appropriate statistical procedure, or perhaps several, with underlying assumptions that qualify them as most readily applicable.

3). Data must be collected and analyzed in a way that is consistent with all of the assumptions of the chosen statistical procedure(s). The procedures typically follow a specific logic that must be understood and strictly followed.

4). Results then need to be presented in a way that repects the logic of each statistical test and allows for reconstruction of missing steps, when necessary, by potential readers.

5). Finally, and most importantly, there must be an explicit consideration of whether any of the statistical results actually *mean anything* as far as the original biological hypotheses were concerned.

Logic of Statistical Tests:

Here's an excellent framework to follow in conducting a statistical test (Example comes from a One-Sample t-test of the mean. We'll see this shortly):

Assumptions:

- Observed values X_1, X_2, X_3, ... X_n are a random sample from $\sim N(\mu,\sigma^2)$.

- Variance σ^2 of the population is *unknown.*

^ Each statistical test is only applicable to specific kinds of samples drawn from a population with specific properties. In this case, data values X are a properly drawn random sample from a population that has a Normal Distribution with population parameters mean=μ and variance=σ^2 that are unknown. The researcher needs to verify whether the data at hand might be drawn from a Normal Distribution. If so, then one can proceed. If not, the test is formally inapplicable. In many instances, however, tests may be robust to violations of one or more assumptions. For example, the t-test is reasonably *robust* to the assumption of population normal distribution, so usually one can proceed as long as the sample isn't wildly non-Normal.

Hypotheses:

H_0: $\mu = \mu_0$ **< μ_0 is a specified value for μ**
H_1: $\mu \neq \mu_0$ **< Two sided test**

or

H_0: $\mu = \mu_0$ **< μ_0 is a specified value for μ**
H_1: $\mu < \mu_0$ **< One sided test**

^ Biological hypotheses are restated formally in a statistical test as *statistical hypotheses.* Statistical hypotheses consist of a matched pair of hypotheses that together comprise *all possible events* (i.e., outcomes) in the *sample space* (i.e., set of all possible outcomes - See *Biostatistics* Worksheet 050). In other words, the probability of the Union of hypotheses is exactly 1.0. The pair of hypotheses consist of:

the *null hypothesis* H_0 - a biologically "uninteresting" hypothesis often indicating no effect for treatments, random behavior, or otherwise non-biological results

the *alternative hypothesis* H_1. - a biologically "interesting" hypothesis perhaps indicating a value or difference for a biological treatment, etc.

The general strategy of a statistical test is to use a probability distribution to determine whether H_0 is likely or unlikely. If unlikely, we can reject H_0 and in turn accept H_1. Acceptance of H_1 would then be a *statistical decision* based on the fact that H_1 is the only alternative hypothesis presented in the test. Consideration of the biological interpretion of the test, and multiple possible alternative explanations, comes later.

In some instances, statistical hypotheses are termed "two-sided" if two distinct possibilities are implicit in H_1. For instance, in the *two-sided* statement of hypotheses above, H_1 says that $\mu < \mu_0$ or $\mu > \mu_0$. By contrast, the *one-sided* statement of hypotheses above allows for only one possibility $\mu < \mu_0$.

Test Statistic:

$t := \frac{X_{bar} - \mu_0}{\frac{s}{\sqrt{n}}}$ < t is the normalized distance between means Xbar and μ_0

^ A test statistic is a number calculated from the sample used in making a statistical decision between H_0 and H_1. Test statistics are usually calculated so that one may consult a well-known statistical distribution. In this case the value of the *test statistic* t will be compared with the *t-distribution* to find P(t). Note that statistic t and the t-distribution are different things.

Sampling Distribution:

If Assumptions hold and H_0 is true, then $t \sim t_{(n-1)}$

^ Test statistics X are carefully chosen to have probabilities P(X) and cumulative probabilities Φ(X) that are understood. In this case if H_0 is true, then the t statistic is distributed according to Student's t-distribution t(t) with (n-1) degrees of freedom.

Critical Value of the Test:

$\alpha := 0.05$ < Probability of Type I error must be explicitly set

^ See below for the definition of "Type 1" or "α" error. This is a criterion for how stringent the test will be. Stringency, however, is a tradeoff with "Type 2" or "β" error as described below. Both types of error are dependent on the number of observations n.

$C := qt(\alpha, n - 1)$

^ A probability P(X) is set above by α. From this, one needs to find the quantile, that is, an X value for which one has $P(X)=\alpha$ under some probability distribution. Standard statistical tables provide a way to find X from P(X) as do explicit functions built into modern statistical software. In both cases one typically works with the cumulative probability function Φ(X). In the example above, since $t \sim t(n-1)$ we use the inverse cumulative t function qt() to find the Critical Value C. Note: C is a quantile - a cut off value of the test statistic t.

Decision Rule:

IF $t > C$ THEN REJECT H_0, OTHERWISE ACCEPT H_0 < One-way case

IF $|t| > C$ THEN REJECT H_0, OTHERWISE ACCEPT H_0 < Two-way case

^ The decision rule compares the calculated test statistic of the sample with the critical value C. If the rule is determined to be true, then H_0 is *rejected,* and the alternative H_1 *accepted* for statistical purposes. Of course, upon rejecting H_0, deciding whether H_1 is the *only viable biological* hypothesis comes later.

Probability Value:

$P = \Phi_t(t)$ < probability of finding test statistic t given the Assumptions *and* if H_0 is true.

^ Although not part of the formal statistical test, it is common practice to provide a probability value P(X) for the test statistic X calculated in the test assuming H_0 to be true. In the case above, since $t \sim t_{(n-1)}$ and we have statistic t, we use the cumulative probability density function pt(t) to find $\Phi(t)$.

Common attributions for P:

IF	0.001	< P		then the results are *statistically very highly significant.*
IF	0.001	< P <	0.01	then the results are *statistically highly significant.*
IF	0.01	< P <	0.05	then the results are *statistically signifcant.*
IF	0.05	< P		then the resulst are *NOT statistically signifcant.*

Note that most statistical software reports probability P rather than critical values C. In order to use this information, one must recast the Decision Rule as follows:

Alternative Decision Rule based on Probability:

IF $P < \alpha$ THEN REJECT H_0, OTHERWISE ACCEPT H_0 < Two-way case

^ note the direction of arrow here!

Confidence Interval

$$\left(X_{bar} + C_1 \cdot \frac{s}{\sqrt{n}} \quad X_{bar} + C_2 \cdot \frac{s}{\sqrt{n}}\right)$$

< CI for the mean X_{bar}

^ Although not part of the formal statistical test, Confidence Intervals (CI) are commonly reported.

Truth Table of Statistical Tests:

	If H0 is true	If H0 is false
If H0 is rejected	Type 1 error α	No error (1-β) = **POWER**
If H0 is accepted	No error	Type 2 error β

^ Two kinds of error are possible in using a statistical test's Decision Rule. The probability for *type 1 error* is explicitly set at α. *Type 2 errors* with probability β are minimized by proper choice of a statistical test. The complement of Type 2 error with probability (1-β) is termed the POWER of the test. In general, greater stringency in α (such as choosing $\alpha = 0.01$ instead of $\alpha = 0.05$) decreases the power of a test. However, the number of observation in the sample (n) influences both α and power of a test. More data is always better within limits. As a result, one often adjusts initial design of a test by first calculating estimates of potential power for a test with different values of n prior to collecting data.