Warren Ewens Gregory Grant

Statisical Methods in Bioinformatics:

An Introduction

Second Edition

With 30 Figures

Springer

3.5 Hypothesis Testing: Examples

3.5.1 Example 1. Testing for a Mean: the One-Sample Case

A classic test in statistics concerns the unknown mean μ of a normal distribution. Suppose first that the variance σ^2 of this distribution is known. One case of this is the test of the null hypothesis $\mu = \mu_0$ against the one-sided alternative hypothesis $\mu > \mu_0$. If this test is carried out using the observed values of random variables $X_1, X_2, \ldots, X_n$ having the normal distribution in question, the statistical theory of Chapter 9 leads to the use of $\bar{X}$ as an optimal test statistic and the rejection of the null hypothesis if the observed value $\bar{x}$ of $\bar{X}$ is "too much larger" than μ_0. The random variable $\bar{X}$ has known variance σ^2/n and mean μ_0 if the null hypothesis is true. The standardization procedure described in Section 1.10.2 then shows that the random variable Z, defined by

$$Z = \frac{(\bar{X} - \mu_0)\sqrt{n}}{\sigma} \tag{3.24}$$

has the standard normal distribution when the null hypothesis is true. Since the probability that such a random variable exceeds 1.645 is 0.05, a desired Type I error 5% is achieved if the null hypothesis is rejected when

$$\frac{(\bar{x} - \mu_0)\sqrt{n}}{\sigma} \geq 1.645, \tag{3.25}$$

where $\bar{x}$ is the observed value of $\bar{X}$ once the data are obtained. Equivalently, the null hypothesis is rejected if

$$\bar{x} \geq \mu_0 + 1.645\sigma/\sqrt{n}. \tag{3.26}$$

If the alternative hypothesis had been $\mu \leq \mu_0$, the null hypothesis would be rejected if the observed value $\bar{x} \leq \mu_0 - 1.645\sigma/\sqrt{n}$. If the alternative hypothesis had been two-sided, so that no specification is made for the value of μ, the null hypothesis would be rejected if $|\bar{x} - \mu_0| \geq 1.96\sigma/\sqrt{n}$. This shows that the nature of the alternative hypothesis determines the values of the test statistic that lead to rejection of the null hypothesis. It will be shown in Section 3.7 that in some cases it can also determine the choice of the test statistic itself.

A more realistic situation arises when σ^2 is unknown, in which case a *one-sample t-test* is used. Here we estimate the unknown variance σ^2 by s^2, defined in (3.6), and use as test statistic the one-sample t statistic, defined by

$$t = \frac{(\bar{x} - \mu_0)\sqrt{n}}{s}. \tag{3.27}$$

Under the assumption that $X_1, X_2, \ldots, X_n$ are NID(μ, σ^2), the null hypothesis distribution of T, defined by

$$T = \frac{(\bar{X} - \mu_0)\sqrt{n}}{S}, \tag{3.28}$$

is well known (as the t distribution with $n-1$ degrees of freedom). The density function of T is independent of μ_0 and σ^2, being

$$f_T(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)\left(1+\frac{t^2}{n}\right)^{(n+1)/2}}, \quad -\infty < t < +\infty. \tag{3.29}$$

An outline of the derivation of this density function is given in Problem 3.7.

It is perhaps remarkable that this density function is independent of the value of σ^2. The value σ^2 is not specified under the null hypothesis, and this implies that significance points of t can be calculated no matter what the value of σ^2 might be. These significance points have been calculated from (3.29) for a variety of values of n and the chosen Type I error, and are widely available.

The t distribution (3.29) differs from standard normal distribution applying for the statistic Z, so that the significance points appropriate for Z are not appropriate for T. However the t distribution converges to the standard normal distribution as $n \to \infty$.

Since the null hypothesis distribution of T is independent of the values of μ_0 and σ^2, T is said to be a *pivotal quantity*. It is because of the pivotal nature of T that explicit significance points of the t distribution can be found, whatever the values of μ_0 and σ^2 might be.

3.5.2 Example 2. The Two-Sample t-Test

A protein coding gene is a segment of the DNA that codes for a particular protein (or proteins). In any given cell type at any given time, this protein may or may not be needed. Each cell will generate the proteins it needs, which will usually be some small subset of all possible proteins. If a protein is generated in a cell, we say that the gene coding for this protein is *expressed* in that cell type. Furthermore, any given protein can be expressed at many different levels. One cell type might need more copies of a particular protein than another cell type. When this happens we say that the gene is *differentially expressed* between the two cell types. There are several techniques for measuring the level of gene expression in a cell type. All of these methods are subject to both biological and experimental variability. Therefore, one cannot simply measure the level of expression once in each cell type to test for differential expression. Instead, one must repeat each experiment several times and perform a statistical test of the hypothesis that they are expressed at the same or different levels.

Suppose that the mean expression levels of a given gene in two cell types, for example normal and tumor (cancerous) cells, are to be compared. In statistical terms, this comparison can be framed as the test of the equality of two unknown means. For the moment we assume that the (unknown) variance of expression level in normal cells is identical to that in tumor cells. To test for equality of the two means, we plan to measure the expression levels of m cells of one type and compare these with the expression levels of n cells of another type. Suppose that, before the experiment, the measurements $X_{11}, X_{12}, \ldots, X_{1m}$ from the first cell type are thought of as m NID(μ_1, σ^2) random variables, and the measurements $X_{21}, X_{22}, \ldots, X_{2n}$ from the second cell type are thought of as n NID(μ_2, σ^2) random variables. The null hypothesis states that $\mu_1 = \mu_2 (= \mu$, unspecified). We assume for the moment that the alternative hypothesis leaves both μ_1 and μ_2 unspecified, so that our eventual test is two-sided.

N normal
I identically
D distributed

The theory in Chapter 9 shows that, under the assumptions made, the optimal test statistic is T, defined now by

$$T = \frac{(\bar{X}_1 - \bar{X}_2)\sqrt{mn}}{S\sqrt{m+n}}, \tag{3.30}$$

with S defined by

$$S^2 = \frac{\sum_{i=1}^{m}(X_{1i} - \bar{X}_1)^2 + \sum_{i=1}^{n}(X_{2i} - \bar{X}_2)^2}{m+n-2}. \tag{3.31}$$

The form of this test statistic can been understood by observing that the variance of $\bar{X}_1 - \bar{X}_2$ is $\sigma^2/m + \sigma^2/n$. If we had known the variance σ^2, we could use as test statistic the quantity Z, defined by

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma^2/m + \sigma^2/n}} = \frac{(\bar{X}_1 - \bar{X}_2)\sqrt{mn}}{\sigma\sqrt{m+n}}. \tag{3.32}$$

Since σ^2 is unknown, it is estimated by the pooled estimator S^2, using observations from both normal and tumor cells, and in general from the two groups being compared. This leads to the T statistic in (3.30).

The null hypothesis probability distribution of T is independent of both the value for the (common) mean unspecified under the null hypothesis and of the unknown variance σ^2. This implies that T (defined by (3.30)) is a pivotal quantity. The null hypothesis distribution of T is the t distribution (3.29) with $m + n - 2$ degrees of freedom, and this enables a convenient assessment of the significance of the observed value t of T, defined as

$$t = \frac{(\bar{x}_1 - \bar{x}_2)\sqrt{mn}}{s\sqrt{m+n}}, \tag{3.33}$$

with s defined by

$$s^2 = \frac{\sum_{i=1}^{m}(x_{1i} - \bar{x}_1)^2 + \sum_{i=1}^{n}(x_{2i} - \bar{x}_2)^2}{m + n - 2}. \tag{3.34}$$

For the two-sided test discussed above, significantly large positive or large negative values of t lead to the rejection of the null hypothesis. When the alternative hypothesis is $\mu_1 \geq \mu_2$, significantly large positive values of t lead to the rejection of the null hypothesis, and when the alternative hypothesis is $\mu_1 \leq \mu_2$, significantly large negative values of t lead to the rejection of the null hypothesis.

In reality, expression levels cannot generally be expected to have normal distributions, nor should the variances of the two types generally be expected to be equal. These two assumptions were made in the above t-test procedure, and the significance points of the t distribution are calculated assuming that both assumptions hold. Thus in practice it might not be appropriate to use the t-test to test for differential expression. In general, if the normal distribution assumption is unjustified we should use the non-parametric tests: these are discussed in Section 3.8.2 and in Chapter 13.

The optimality property of the two-sample t-test procedure described above derives from statistical theory – see Chapter 9. The theoretical development assumes that the variances of the random variables in the two groups considered are equal. When, as is often the case in practice, these two variances cannot reasonably be taken as being equal, the theoretical approach of Chapter 9 fails to lead to a testing procedure for which the test statistic has the same distribution for all parameter values not specified by the null hypothesis. That is, no pivotal quantity analogous to equal variance case T as defined in (3.30) exists. This implies that there is no well-defined null hypothesis probability distribution available analogous to that in (3.29) from which significance points can be obtained, whatever the unknown variances in the two groups might be. Because of this, approximate heuristic procedures are required.

One frequently used procedure is as follows. Under the null hypothesis, $\bar{X}_1$ and $\bar{X}_2$ have normal distributions with the same mean and respective variances σ_1^2/m and σ_2^2/n, so that the difference $\bar{X}_1 - \bar{X}_2$ has a normal distribution with mean zero and variance

$$\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}. \tag{3.35}$$

The variances σ_1^2 and σ_2^2 are unknown, but have estimators S_1^2 and S_2^2, where

$$S_1^2 = \frac{\sum_{i=1}^{m}(X_{i1} - \bar{X}_1)^2}{m-1}, \quad S_2^2 = \frac{\sum_{i=1}^{n}(X_{i2} - \bar{X}_2)^2}{n-1}. \tag{3.36}$$

One then computes the statistic T', defined by

$$T' = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{A+B}}, \tag{3.37}$$

where

$$A = \frac{S_1^2}{m}, \quad B = \frac{S_2^2}{n}.$$

When the null hypothesis of equal means is true, T' has an approximate t distribution with degrees of freedom given by the largest integer less than or equal to ν (see Lehmann (1986)), where ν defined by

$$\nu = \frac{(A+B)^2}{\frac{A^2}{m-1} + \frac{B^2}{n-1}}.$$

When $m = n$, T' is identical to the T statistic (3.30). However, in this case the number of degrees of freedom appropriate for t' is not equal to the number $2(n-1)$ applying when the two variances are assumed to be equal: The value of ν lies in the interval $[n-1, 2(n-1)]$, the actual value depending on the ratio of S_1^2/S_2^2.

Markowski and Markowski (1990) show for the case $m = n$ that even when the variances in the two groups differ, use of the "equal variance" t-test procedure leads to a very small error.

An important case of the two-sample t test arises if $n = m$ and the random variables X_{1i} and X_{2i} are logically paired, for example being expression levels of normal and tumor cells taken from the same person. In this "paired t-test" case the test is carried out by using the differences $D_i = X_{1i} - X_{2i}$ and basing the test entirely on these differences. This reduces the test to a one-sample t-test with test statistic T as defined in (3.28) and with X_i replaced by D_i and μ_0 set equal to 0. The test statistic is then

$$T = \frac{\bar{D}\sqrt{n}}{S_D}, \tag{3.38}$$

where S_D^2 defined by the right-hand side in (3.5), with X_i replaced by D_i and $\bar{X}$ by $\bar{D}$.

The advantage of the pairing procedure is that the variance estimate S_D^2 measures only cell type to cell type variation, and eliminates person-to-person variation. If there is significant person-to-person variation, this provides a more powerful test of cell type to cell type variation. In this procedure we see the beginnings of the concept of the Analysis of Variance (ANOVA). In an ANOVA procedure the variation in a body of data is broken down into separate components, each measuring one source of variation, and the significance of one potential source of variation can be investigated free of any influence of other potential sources of variation. The ANOVA concept is developed at length in Section 9.5.

3.5.3 Example 3. Tests on Variances

In Section 3.5.2 we considered two tests, each comparing the means of two groups of random variables. These tests differ depending on whether or not one is prepared to assume that the variances of the random variables in the two groups are equal. This makes it important to describe a test for equality of variances.

We suppose that $X_{11}, X_{12}, \ldots, X_{1m}$ are NID(μ_1, σ_1^2) and X_{21}, $X_{22}, \ldots,$ X_{2n} are NID(μ_2, σ_2^2). We wish to test the null hypothesis $\sigma_1^2 = \sigma_2^2$. To do this we consider the ratio S_1^2/S_2^2 of the two variance estimators S_1^2 and S_2^2 defined in (3.36). Under the null hypothesis this ratio has the F distribution with $(m-1, n-1)$ degrees of freedom, developed in Section 2.13, whatever values the unknown means μ_1 and μ_2 take. If for example the alternative hypothesis were $\sigma_1^2 > \sigma_2^2$, significantly large values of the observed value of this ratio would lead to rejection of the null hypothesis. Significance points of F for Type I errors arising in practice are extensively tabulated, allowing a ready evaluation of whether the observed value of the ratio is indeed significantly large.

We will meet the F test in Section 9.5 in the context of ANOVA (the analysis of variance), where (perhaps unexpectedly) it is used as a test for the equality of several means, rather than as a test for the equality of two variances.

3.5.4 Example 4. Testing for the Parameters in a Multinomial Distribution

In this example we consider a test of the null hypothesis that prescribes specific values for the probabilities $\{p_i\}$ in the multinomial distribution (2.30). The alternative hypothesis considered here is composite and leaves these probabilities unspecified. This can be used, for example, to test for prescribed probabilities for the four nucleotides in a DNA sequence.

Let Y_i be the number of observations in category i. A test statistic often used for this testing procedure is X^2, defined by

$$X^2 = \sum_{i=1}^{k} \frac{(Y_i - np_i)^2}{np_i}. \tag{3.39}$$

Sufficiently large values of the observed value

$$\sum_{i=1}^{k} \frac{(y_i - np_i)^2}{np_i} \tag{3.40}$$

of X^2 lead to rejection of the null hypothesis. The quantity (3.40) may be thought of as a measure of the discrepancy between the observed values $\{y_i\}$ and the respective null hypothesis expected values $\{np_i\}$.