Rediscovering a little known fact about the 
t-test: algebraic, geometric and 
distributional considerations

Jennifer A. Sinnott, Steven MacEachern and Mario Peruggia
Department of Statistics, The Ohio State University, Columbus, OH

July 31, 2019

Abstract
In this article we discuss the role that the null hypothesis should play in the construction of a test statistic used to make a decision concerning the truth of that hypothesis. Motivated by the common recommendation that, to construct the test statistic for testing a point null hypothesis about a binomial proportion, one should act as if the null hypothesis is true, we argue that, on the surface, the one-sample t-test of a point null hypothesis about a Gaussian population mean does not appear to follow the recommendation. We show how simple algebraic manipulations of the usual t-statistic lead to an equivalent test procedure that is consistent with the recommendation, we provide geometric intuition regarding this equivalence, we consider extensions to the problem of testing nested hypotheses in Gaussian linear models. Noting that the equivalence between these tests is largely unknown and is periodically represented in the literature, we argue that these issues should be discussed in advanced undergraduate and graduate courses.

Keywords: Binomial proportion; F-test; nested models; null hypothesis; orthogonal sum of squares decomposition; test statistic.

1 Introduction

Among the first procedures taught in an introductory statistics class are hypothesis testing and confidence interval estimation for a proportion (see, e.g., [Moore et al. 2012]). For example, a student may be given data on the sexes of a sample of $n$ babies born during a certain time period and be asked either to estimate the true proportion $p$ of babies born
male and provide a confidence interval, or to test whether the proportion is equal to, for example, 0.5. Typically, for large $n$, the distribution of the sample proportion is approximated by $\hat{p} \sim N(p, p(1 - p)/n)$, and two slightly different procedures are introduced. For estimation and confidence interval construction, $\hat{p}$ is commonly plugged into the variance formula, and a $100(1 - \alpha)\%$ confidence interval is calculated as

$$\hat{p} \pm z_{\alpha/2} \sqrt{\hat{p}(1 - \hat{p})/n}.$$  

(1)

For testing $H_0 : p = p_0$ for a pre-specified $p_0$, students are advised to act as though the null were true, and use the null to construct the test statistic. As a result, $p_0$ is plugged into the variance formula, producing the test statistic

$$\frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}.$$  

(2)

Although many different approaches to both testing and interval estimation have been proposed — and many commonly used statistical software packages do not use these formulas exactly — in the authors’ experience, the above methods are still frequently taught for hand calculation in introductory statistics classes of various levels. For instance, Example 10.3.5 in Casella and Berger (2002) discusses precisely two test procedures based on test statistics that use $\hat{p}$ or $p_0$ to estimate the variance, commenting on their relative merits in terms of a comparison of their power functions.

Also among the first procedures taught are estimation and hypothesis testing for the mean $\mu$ of a normal $N(\mu, \sigma^2)$ population with unknown variance $\sigma^2$. For example, a student may be given data on the heights of a random sample of U.S. women and be asked to estimate the true mean height, or test whether it is equal to some specified value. If our data consist of a random sample $Y_1, \ldots, Y_n$ from the $N(\mu, \sigma^2)$ population, $\bar{Y} \sim N(\mu, \sigma^2/n)$, and a confidence interval is constructed analogously to (1), as

$$\bar{Y} \pm t_{n-1, \alpha/2} S/\sqrt{n}$$

where

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$  

(3)

---

1 There is ample evidence that this proportion is larger than 0.5 in most of the world; see, e.g., Chao et al. (2019)
is the sample variance. (This follows from observing that $T := (\bar{Y} - \mu)/(S/\sqrt{n})$ has a $t$ distribution with $n - 1$ degrees of freedom, accounting for the replacement of $\sigma$ with $S$) To test $H_0 : \mu = \mu_0$ for a pre-specified $\mu_0$, we can, analogously to [2], invoke the null. When $H_0$ holds, we know $\mu = \mu_0$ but still need to estimate $\sigma^2$. Since $\mu$ is known, the most efficient estimator of $\sigma^2$ is:

$$S_0^2 := \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mu_0)^2.$$ 

Our test statistic would thus be:

$$T_0 := \frac{\bar{Y} - \mu_0}{S_0/\sqrt{n}}.$$ 

But, of course, people do not use this test statistic! Instead, they construct a statistic that ignores the information that $\mu = \mu_0$ provided by $H_0$, and perform the standard one-sample $t$-test using the test statistic

$$T = \frac{\bar{Y} - \mu_0}{S/\sqrt{n}}.$$ 

At first glance, one might suspect that using this test statistic would be less efficient than using $T_0$, since its denominator has $n - 1$ degrees of freedom rather than $n$.

We are thus led to wonder why information provided by the null is discarded in constructing the one-sample $t$-test. In the remainder of the paper we clarify this question and present a more general perspective.

## 2 Establishing the connection

The connection between the two methods proposed at the end of the previous section can be established from an algebraic and from a geometric point of view. We look at these two approaches separately.

### 2.1 The algebraic point of view

The first point to make is that any intuition that a test based on $T_0$ rather than $T$ could be more efficient is wrong: a tail-area test based on $T_0$ and one based on $T$ produce identical answers. This is because $T$ is a one-to-one, increasing function of $T_0$,

$$T = \frac{\sqrt{n - 1} T_0}{\sqrt{n - T_0^2}},$$ (4)
over the interval \((-\sqrt{n}, \sqrt{n})\), which is the set of possible values for \(T_0\). This relationship is not new: it arises substantively in Lehmann’s approach for demonstrating that the one sample \(t\)-test is a uniformly most powerful (UMP) unbiased test of \(H_0 : \mu = \mu_0\) vs. \(H_A : \mu \neq \mu_0\) \cite{lehmann1986}. The full details of the argument are best left to Lehmann, but, very briefly, for parameters in exponential family distributions, Lehmann’s Theorem 1 in Chapter 5 gives a set of conditions about the form of a test statistic in relation to the family’s sufficient statistics. When these conditions are satisfied, a test based on the test statistic is UMP unbiased. The set of conditions Lehmann provides is satisfied by \(T_0\) rather than \(T\), and the UMP unbiasedness of the \(t\)-test is then established by exhibiting that \(T\) is a one-to-one function of \(T_0\).

Interestingly, this equivalence does not seem to be widely known (at least based on our informal surveying of several colleagues). This is somewhat surprising. In fact, in addition to appearing in Lehmann’s book, the algebraic equivalence of the test statistics is periodically mentioned in the literature (see, e.g., \cite{lefante1986, good1986, shah1987, shah1993, lamotte1994}). Therefore, we believe students should be routinely exposed to this equivalence and given more insight into its essence, emphasizing not only the algebraic aspect, but also the geometric and distributional implications.

### 2.2 The geometric point of view

The second point is that the equivalence of \(T_0\) and \(T\) can be understood geometrically because they can both be viewed as trigonometric functions of the same angle, and it is possible to express any trigonometric function in terms of any other trigonometric function, up to sign. To see the geometric relationship, define the vectors \(\mathbf{v} = (Y_1 - \mu_0, Y_2 - \mu_0, \ldots, Y_n - \mu_0)^\top\) and \(\mathbf{1} = (1, 1, \ldots, 1)^\top\). Then, the orthogonal projection of \(\mathbf{v}\) onto \(\mathbf{1}\) is \(\mathbf{u} = (\bar{Y} - \mu_0)\mathbf{1}\), and the Pythagorean Theorem implies:

\[
\|\mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v} - \mathbf{u}\|^2,
\]

i.e.,

\[
\sum_{i=1}^{n}(Y_i - \mu_0)^2 = n(\bar{Y} - \mu_0)^2 + \sum_{i=1}^{n}(Y_i - \bar{Y})^2,
\]

i.e.,

\[
\text{SSTO} = \text{SST} + \text{SSE},
\]
where we introduce analysis of variance terminology, with SSTO, SST, and SSE indicating the Sums of Squares for Total, Treatment, and Error, respectively. Thus, if we define θ to be the angle between $\mathbf{1}$ and $\mathbf{v}$, then:

$$T_0^2 = n \frac{\text{SST}}{\text{SSTO}} = n \cos^2 \theta \quad \text{and} \quad T^2 = (n - 1) \frac{\text{SST}}{\text{SSE}} = (n - 1) \cot^2 \theta.$$ 

These geometric relationships are illustrated in Figure 1. Using basic trigonometric expressions it is easy to derive the stated algebraic relationship between $T$ and $T_0$. In fact,

$$T^2 = (n - 1) \cot^2 \theta = (n - 1) \frac{\cos^2 \theta}{\sin^2 \theta} = (n - 1) \frac{\cos^2 \theta}{1 - \cos^2 \theta}.$$ 

Substituting $\cos^2 \theta = T_0^2 / n$ into this expression and taking square roots on both sides (making sure the signs match, as they should) yields Equation (4).

### 3 Extension to general linear models

The results presented in the previous section are not specific to the $t$-test setting. In fact, constructing a test statistic by invoking the null hypothesis and constructing it in the “traditional” way produces equivalent test procedures across a range of linear models. This connection can be established by rewriting the two statistics as functions of different terms in the orthogonal decomposition of the sum of squares.

For instance, consider the standard linear model

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon,$$
where $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ is a vector of observations, $\mathbf{X}_{n \times p}$ is a design matrix of rank $p < n$, $\mathbf{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a vector of regression parameters, and $\mathbf{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ is an error vector with elements $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$. Suppose we wish to test the hypothesis that the parameters in a certain subset of size $p_2$ are all zero. Without loss of generality we can assume that the parameters of interest are the last $p_2 < p$ and rewrite the model as

$$\mathbf{Y} = \mathbf{X}_1\mathbf{\beta}_1 + \mathbf{X}_2\mathbf{\beta}_2 + \mathbf{\epsilon},$$

where $\mathbf{X} = [\mathbf{X}_1 | \mathbf{X}_2]$ and $\mathbf{\beta} = (\mathbf{\beta}_1^T, \mathbf{\beta}_2^T)^T$, with $\mathbf{\beta}_i$ of dimension $p_i$ for $i = 1, 2$, and $p_1 + p_2 = p$.

The testing problem concerning the nested model can then be stated as

$$H_0 : \mathbf{\beta}_2 = \mathbf{0} \quad \text{vs.} \quad H_A : \mathbf{\beta}_2 \neq \mathbf{0}.$$ 

Both the “traditional” and the “null hypothesis” testing procedures try to quantify the importance of the reduction in error sums of squares that ensues from entertaining the full model rather than the reduced model, but they differ in the comparison yardstick they use. The “traditional” procedure uses a yardstick based on the full model. The “null hypothesis” procedure uses a yardstick based on the reduced model with $\mathbf{\beta}_2 = \mathbf{0}$.

Geometrically, the statistics arise from a sequence of projections. Specifically, define:

$$\mathbf{P}_1 = \mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T, \quad \mathbf{Q}_1 = \mathbf{I} - \mathbf{P}_1,$$

and

$$\mathbf{P}_{12} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T, \quad \mathbf{Q}_{12} = \mathbf{I} - \mathbf{P}_{12}.$$ 

The matrix $\mathbf{P}_1$ operates an orthogonal projection onto the space spanned by the columns of the reduced design matrix $\mathbf{X}_1$ and the matrix $\mathbf{P}_{12}$ operates an orthogonal projection onto the space spanned by the columns of the full design matrix $\mathbf{X}$. Under the reduced model, the vector of predicted values is

$$\hat{\mathbf{Y}}_1 = \mathbf{P}_1\mathbf{Y},$$

the vector of residuals is

$$\mathbf{r}_1 = \mathbf{Y} - \hat{\mathbf{Y}}_1 = \mathbf{Q}_1\mathbf{Y},$$

and the residual sum of squares is

$$\text{SSE}_1 = \mathbf{Y}^T\mathbf{Q}_1^T\mathbf{Q}_1\mathbf{Y} = \mathbf{Y}^T\mathbf{Q}_1\mathbf{Y}.$$
Similarly, under the full model, the vector of predicted values is
\[ \hat{Y}_{12} = P_{12}Y, \]
the vector of residuals is
\[ r = Y - \hat{Y}_{12} = Q_{12}Y, \]
and the residual sum of squares is
\[ SSE_{12} = Y^TQ_{12}Y. \]

The reduction in sums of squares ensuing from fitting the larger model is given by
\[ SS_{2|1} = SSE_1 - SSE_{12} = Y^T(Q_1 - Q_{12})Y = Y^T(P_{12} - P_1)Y. \]

The “traditional” procedure compares \( SS_{2|1} \) to \( SSE_{12} \), the error sum of squares for the full model, while the “null hypothesis” procedure compares \( SS_{2|1} \) to \( SSE_1 = SS_{2|1} + SSE_{12} \), the error sum of squares for the reduced model envisioned to hold under the null. After adjusting for the degrees of freedom of the various sums of squares, the resulting test statistics are
\[ F_{\text{trad}} = \frac{SS_{2|1}/p_2}{SSE_{12}/(n - p)}, \]
and
\[ F_{\text{null}} = \frac{SS_{2|1}/p_2}{SSE_1/(n - p_1)} = \frac{SS_{2|1}/p_2}{(SS_{2|1} + SSE_{12})/(n - p_1)}, \]
respectively.

### 3.1 Algebra, geometry, and distributional results

The orthogonal decomposition at play in this setting is analogous to the one presented in Section 2 and is described in Figure 2 along with the relationships between its various elements. Algebraic and trigonometric manipulations similar to those outlined at the end of Section 2 show that \( F_{\text{trad}} \) is a one-to-one, increasing function of \( F_{\text{null}} \) over \((0, (n - p_1)/p_2)\), the set of possible values for \( F_{\text{null}} \):
\[ F_{\text{trad}} = \frac{(n - p)F_{\text{null}}}{n - p_1 - p_2F_{\text{null}}}. \]
Thus, as in the case of the *t*-test, tail-area tests using $F_{\text{trad}}$ and $F_{\text{null}}$ are identical. Note that, when $p = 1$, $p_1 = 0$, and $p_2 = 1$, the relationship between $F_{\text{trad}}$ and $F_{\text{null}}$ given in Equation (5) reduces to the relationship between $T^2$ and $T_0^2$ implied by Equation (4).

The implementation of either test procedure requires knowledge of the distribution of the corresponding test statistic under the null hypothesis. Using the notation introduced in Figure 2, standard distributional results imply that, under the null hypothesis,

$$b^2 / \sigma^2 = \frac{SS_{2|1}}{\sigma^2} \sim \chi_{p_2}^2,$$
$$c^2 / \sigma^2 = \frac{SSE_{12}}{\sigma^2} \sim \chi_{n-p}^2,$$

with $b^2$ independent of $c^2$.

Then,

$$F_{\text{trad}} = \frac{b^2 / p_2}{c^2 / (n - p)} \sim F_{p_2, n-p},$$

as it is the ratio of two independent chi-square random variables divided by their degrees of freedom. Also,

$$\frac{p_2}{n-p_1} F_{\text{null}} = \frac{b^2}{b^2 + c^2} \sim \text{Beta} \left( \frac{1}{2} p_2, \frac{1}{2} (n - p) \right),$$

as it is the ratio between a chi-square random variable and the sum of that chi-square random variable and an independent chi-square random variable.
4 Discussion

The idea of constructing a test statistic by pretending that the null hypothesis is true is routinely presented as a general guideline when using binomial data for testing the hypothesis that a population proportion is equal to a given value. Yet, this guideline is not followed, at least on the surface, when normal data are used to build the $t$-test for testing the hypothesis that the population mean is equal to a given value. As we noted in the paper, the $t$-test is actually equivalent to a procedure based on a test statistic derived by following the guideline, but making the connection requires a little algebra, and is, to our knowledge, not typically made in introductory statistics classes. We have also noted that the same considerations presented for the $t$-test extend to the use of the $F$-test for testing hypotheses concerning nested linear models with Gaussian errors.

So, we are left to speculate why, in the case of the $t$-test and of the $F$-test, the “traditional” procedure is preferred to the “null hypothesis” procedure. If a formal comparison is required, there is no clear distributional advantage of one approach over the other. For the comparison of nested linear models, under the null, the “traditional” procedure requires calculation of the tail area of an $F$ distribution and the “null hypothesis” procedure requires calculation of the tail area of a Beta distribution. If a power calculation has to be performed under some alternative, it can be based on the non-central $F$-distribution for the traditional procedure and on the Type I non-central Beta distribution for the “null hypothesis” procedure, again with no clear advantage of one approach over the other. Similar considerations apply to the case of the $t$-test.

An appealing aspect of the “traditional” procedures is that the $t$-statistic $T$ and the $F$-statistic $F_{\text{trad}}$ are both constructed as ratios of independent quantities. Because, in both cases, the decision rule is based on an assessment of the relative size of the numerator and denominator, it is conceivable that independence may have been a key factor in establishing the tradition, as an informal comparison of independent quantities is easier. Under the null, the denominators of the “null hypothesis” test statistics are more efficient estimators of variability (have more degrees of freedom) than their “traditional” counterparts. However, this gain in efficiency is offset by the dependence between numerator and denominator (see LaMotte 1994 for a related discussion).
The fundamental question raised by the examples we presented is this article concerns the role that the null hypothesis should play in the testing paradigm. By assumption, the null hypothesis is assumed true in order to assess statistical significance, but to what extent should one rely on it to construct the test statistic? When confronted with a new statistical model and a new parameter of interest, it can be somewhat of an art to determine a good choice of test statistic. Three common “automatic” approaches for constructing test statistics from likelihoods privilege the null differently: score tests are typically built under the null; Wald tests are typically built under the alternative; and likelihood ratio tests compare the null and the alternative somewhat equally.

However, the test statistic resulting from direct application of one of these principles may not be the one that is ultimately used in common practice. Consider again the case of testing a nested reduced model against the full model in the Gaussian linear model setting. The $F$-test based on the statistic $F_{\text{trad}}$ is often presented as a likelihood ratio test, that is, a test that rejects the null hypothesis when the ratio

$$\lambda(Y, X) = \frac{\sup_{\beta_1, \sigma^2} L(\beta_1, \sigma^2|Y, X_1)}{\sup_{\beta, \sigma^2} L(\beta, \sigma^2|Y, X)}$$

is small. This is, of course, true. Yet, the statistic arising from direct calculation of the likelihood ratio is $\text{SSE}_1/\text{SS}_2|1$, a constant multiple of $F_{\text{null}}^{-1}$, and further algebraic manipulations are required to obtain the test procedure that rejects for large values of $F_{\text{trad}}$.

In addition to the basic guiding principles, other considerations may be at play when a certain tradition is established of preferring one form of a test procedure over another for a given problem. For the nested model comparison, we already noted one desirable feature exhibited by $F_{\text{trad}}$, namely that its numerator and denominator are independent. Another feature worth noting is that the denominator of $F_{\text{trad}}$ does not depend on the particular reduced model under consideration while the denominator of $F_{\text{null}}$ does. Although this is not much of a computational burden, it is intuitively appealing to be able to use the same yardstick in the denominator when testing different nested models against the same full model.

In sum, while we do not have a conclusive explanation as to why certain traditions have established themselves as the standard of practice for specific problems, we believe that these issues, often overlooked, should be brought to the attention of statistics students in
advanced undergraduate and graduate courses. The examples we discussed provide results for common testing problems that can be used to focus the in-class discussion.

Acknowledgments. This material is based upon work supported by the National Science Foundation under Grants No. SES-1424481 and No. DMS-1613110.

References

Casella, G. and Berger, R. (2002). *Statistical Inference*. Duxbury-Thomson Learning, Second edition.

Chao, F., Gerland, P., Cook, A. R., and Alkema, L. (2019). Systematic assessment of the sex ratio at birth for all countries and estimation of national imbalances and regional reference levels. *Proceedings of the National Academy of Sciences*, 116(19):9303–9311.

Good, I. (1986). Comments, conjectures, and conclusions: C258 editorial note on c257 regarding the t-test. *Journal of statistical computation and simulation*, 25(3-4):296–297.

LaMotte, L. R. (1994). A note on the role of independence in t statistics constructed from linear statistics in regression models. *The American Statistician*, 48(3):238–240.

Lefante Jr, J. J. and Shah, A. K. (1986). C257. a note on the one-sample t-test. *Journal of statistical computation and simulation*, 25(3-4):295–296.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses*. John Wiley & Sons.

Moore, D. S., McCabe, G. P., and Craig, B. A. (2012). Introduction to the practice of statistics.

Shah, A. K. and Krishnamoorthy, K. (1993). Testing means using hypothesis-dependent variance estimates. *The American Statistician*, 47(2):115–117.

Shah, A. K. and Lefante Jr, J. J. (1987). C293. a note on using a hypothesis-dependent variance estimate. *Journal of Statistical Computation and Simulation*, 28(4):347–349.