STATISTICA, anno LXXXII, n. 2, 2022

REDISCOVERING A LITTLE KNOWN FACT ABOUT THE *T*-TEST AND THE *F*-TEST: ALGEBRAIC, GEOMETRIC, DISTRIBUTIONAL AND GRAPHICAL CONSIDERATIONS

Jennifer A. Sinnott[1]
Department of Statistics, The Ohio State University, Columbus, Ohio, USA
Steven N. MacEachern
Department of Statistics, The Ohio State University, Columbus, Ohio, USA
Mario Peruggia
Department of Statistics, The Ohio State University, Columbus, Ohio, USA

SUMMARY

We discuss the role that the null hypothesis should play in the construction of a test statistic used to make a decision about that hypothesis. To construct the test statistic for a point null hypothesis about a binomial proportion, a common recommendation is to 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 consistent with the recommendation. We provide geometric intuition regarding this equivalence and we consider extensions to testing nested hypotheses in Gaussian linear models. We discuss an application to graphical residual diagnostics where the form of the test statistic makes a practical difference. By examining the formulation of the test statistic from multiple perspectives in this familiar example, we provide simple, concrete illustrations of some important issues that can guide the formulation of effective solutions to more complex statistical problems.

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, students may be given data on the sexes of a sample of n babies born during

[1] Corresponding Author. E-mail: jsinnott@stat.osu.edu

a certain time period. They may 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 0.5, for example[2]. Typically, for large n, the distribution of the sample proportion is approximated by $\widehat{p} \sim N(p, p(1-p)/n)$, and two slightly different procedures are introduced. For estimation and confidence interval construction, $\widehat{p}$ is commonly plugged into the variance formula, and a $100(1-\alpha)\%$ confidence interval is calculated as

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

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

$$\frac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}. \tag{2}$$

Although many different approaches to both testing and interval estimation have been proposed — and many commonly used statistical software packages allow the user to apply continuity corrections to these formulas to improve the asymptotic approximation (e.g., by setting the argument `correct = TRUE` in the `R` function `prop.test`) — 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 $\widehat{p}$ or p_0 to estimate the variance, commenting on their relative merits in terms of a comparison of their power functions. For further discussions of procedures used in the one-sample proportion setting, see, e.g., Agresti and Coull (1998) and Yang and Black (2019).

Also among the first procedures taught are estimation and hypothesis testing for the mean μ of a normal $N(\mu, \sigma^2)$ population with unknown variance σ^2. For example, students 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, $\bar{Y} \sim N\left(\mu, \sigma^2/n\right)$, and a confidence interval is constructed analogously to (1), as

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

where

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \tag{3}$$

[2] There is 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 σ with S). To test $H_0 : \mu = \mu_0$ for a pre-specified μ_0, we can, analogously to (2), invoke the null. When H_0 holds, we know $\mu = \mu_0$ but still need to estimate σ^2. Since μ is known, the most efficient estimator of σ^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, that we think will be of interest to colleagues who teach this material as well as those interested in the development and implications of some of our most fundamental statistical tools.

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.

To begin, we note 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,

$$T = \frac{\sqrt{n-1}\,T_0}{\sqrt{n - T_0^2}}, \tag{4}$$

over the interval $(-\sqrt{n}, \sqrt{n})$, which is the set of possible values for T_0. Specifically, for any fixed α, with $0 \le \alpha \le 1$, let $c_\alpha \ge 0$ be the critical value of the size α test based on T_0.

The rejection region of this test is

$$R_{T_0} = \{\mathbf{y} = (y_1, \ldots, y_n)^\intercal : |T_0(\mathbf{y})| \geq c_\alpha\}.$$

Because the transformation in Equation (4) is monotonic increasing on $[0, \sqrt{n})$, the set

$$R_T = \{\mathbf{y} = (y_1, \ldots, y_n)^\intercal : |T(\mathbf{y})| \geq (\sqrt{n-1}\, c_\alpha)/(\sqrt{n - c_\alpha^2})\}$$

satisfies $R_T = R_{T_0}$. It follows that the test that rejects if and only if

$$|T(\mathbf{y})| \geq (\sqrt{n-1}\, c_\alpha)/(\sqrt{n - c_\alpha^2})$$

has the exact same rejection region (in sample space) as the test that rejects when $|T_0(\mathbf{y})| \geq c_\alpha$. The two tests must then have the same size and power function and are therefore equivalent.

As noted by a colleague, a simple way to establish Equation (4) is to recognize that the one sample t-test can be derived as a likelihood ratio test that rejects $H_0 : \mu = \mu_0$ when the ratio

$$\lambda(\mathbf{Y}) = \frac{\sup_{\sigma^2} L(\mu_0, \sigma^2 | \mathbf{Y})}{\sup_{\mu, \sigma^2} L(\mu, \sigma^2 | \mathbf{Y})}$$

is small or, equivalently, when the ratio of sums of squares under the null and full model,

$$R = \frac{\sum_{j=1}(Y_j - \mu_0)^2}{\sum_{j=1}^n (Y_j - \bar{Y})^2}, \tag{5}$$

is large. This ratio can be expressed as

$$R = \frac{\sum_{j=1}(Y_j - \bar{Y})^2 + n(\bar{Y} - \mu_0)^2}{\sum_{j=1}^n (Y_j - \bar{Y})^2} = 1 + \frac{T^2}{n-1}$$

or as

$$R = \frac{\sum_{j=1}(Y_j - \mu_0)^2}{\sum_{j=1}^n (Y_j - \mu_0)^2 - n(\bar{Y} - \mu_0)^2} = \frac{1}{1 - T_0^2/n}.$$

The former expression leads to the standard t-test based on T, while the latter leads to the test based on T_0. Equating these two expressions yields the identity of Equation (4).

This relationship between T and T_0 is, of course, not new: for example, 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. $H_A : \mu \neq \mu_0$ (Lehmann, 1986). 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. Lefante, Jr. and Shah, 1986; Good, 1986; Shah and Lefante Jr, 1987; Shah and Krishnamoorthy, 1993; LaMotte, 1994). However, we feel that the equivalence is worth revisiting, both in the context of the t-test and in the more general setting of nested linear models, where an analogous equivalence holds. The geometric interpretation of the equivalence, not described in these earlier references, provides an interesting addition to the geometric interpretation of linear models. Moreover, despite the test statistics leading to identical conclusions in the linear models setting, one choice naturally leads a practitioner to consider so-called studentized residuals while the other leads to so-called standardized residuals—and these sets of residuals do have different properties and, when plotted, may lead to different visual interpretations. We expand on these remarks in subsequent sections.

3. The geometric point of view

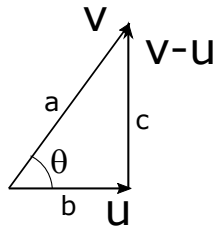
Interestingly, the equivalence of T_0 and T can be understood geometrically because they can both be viewed as trigonometric functions of the same angle, and it is possible to express any trigonometric function in terms of any other trigonometric function, up to sign. To see the geometric relationship,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

$$\begin{aligned} & & ||\mathbf{v}||^2 &= ||\mathbf{u}||^2 &+ \quad & ||\mathbf{v}-\mathbf{u}||^2, \\ \text{i.e.,} & & \sum_{i=1}^{n}(Y_i-\mu_0)^2 &= n(\bar{Y}-\mu_0)^2 &+ \quad & \sum_{i=1}^{n}(Y_i-\bar{Y})^2, \\ \text{i.e.,} & & \text{SSTO} &= \text{SST} &+ \quad & \text{SSE}, \end{aligned}$$

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:

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

A stylized, two-dimensional representation of the essence of these geometric relationships is presented in Figure 1. Using basic trigonometric expressions it is easy to derive



$$
\begin{aligned}
a &= \|\mathbf{v}\| = \sqrt{\text{SSTO}},\\
b &= a\cos\theta = \|\mathbf{u}\| = \sqrt{\text{SST}},\\
c &= a\sin\theta = \|\mathbf{v}-\mathbf{u}\| = \sqrt{\text{SSE}},\\
T_0^2 &= n\,(b^2/a^2) = n\cos^2\theta,\\
T^2 &= (n-1)(b^2/c^2) = (n-1)\cot^2\theta.
\end{aligned}
$$

Figure 1 – Geometric representation of the test statistics T_0 and T.

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).

4. Extension to linear models

The results presented in the previous sections 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.

4.1. Nested models

For instance, consider the standard linear model

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

where $\mathbf{Y} = (Y_1,\ldots,Y_n)^\intercal$ is a vector of observations, $\mathbf{X}_{n\times p}$ is a design matrix of rank $p < n$, $\beta = (\beta_1,\ldots,\beta_p)^\intercal$ is a vector of regression parameters, and $\boldsymbol{\epsilon} = (\epsilon_1,\ldots,\epsilon_n)^\intercal$ is an error vector with elements $\epsilon_i \overset{\text{iid}}{\sim} N(0,\sigma^2)$. Suppose we wish to determine if a specific collection of p_2 covariates in $\mathbf{X}$ does not significantly contribute to the prediction of $\mathbf{Y}$ in the linear model. We can formulate this question as a testing problem in which the null hypothesis states that the p_2 regression coefficients for these covariates 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

$$\mathbf{Y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \boldsymbol{\epsilon},$$

where $\mathbf{X} = [\mathbf{X}_1|\mathbf{X}_2]$ and $\beta = (\beta_1^\intercal, \beta_2^\intercal)^\intercal$, with β_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

$$H_0 : \beta_2 = 0 \text{ vs. } H_A : \beta_2 \neq 0.$$

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 $\beta_2 = 0$.

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

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

and

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

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

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

the vector of residuals is

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

and the residual sum of squares is

$$\mathrm{SSE}_1 = \mathbf{Y}^\intercal\mathbf{Q}_1^\intercal\mathbf{Q}_1\mathbf{Y} = \mathbf{Y}^\intercal\mathbf{Q}_1\mathbf{Y}.$$

Similarly, under the full model, the vector of predicted values is

$$\widehat{\mathbf{Y}}_{12} = \mathbf{P}_{12}\mathbf{Y},$$

the vector of residuals is

$$\mathbf{r} = \mathbf{Y} - \widehat{\mathbf{Y}}_{12} = \mathbf{Q}_{12}\mathbf{Y},$$

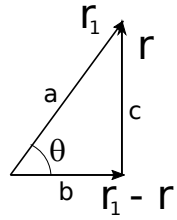
and the residual sum of squares is

$$\mathrm{SSE}_{12} = \mathbf{Y}^\intercal\mathbf{Q}_{12}\mathbf{Y}.$$

The reduction in sums of squares ensuing from fitting the larger model is given by

$$\mathrm{SS}_{2|1} = \mathrm{SSE}_1 - \mathrm{SSE}_{12} = \mathbf{Y}^\intercal(\mathbf{Q}_1 - \mathbf{Q}_{12})\mathbf{Y} = \mathbf{Y}^\intercal(\mathbf{P}_{12} - \mathbf{P}_1)\mathbf{Y}.$$

The "traditional" procedure compares $\mathrm{SS}_{2|1}$ to SSE_{12}, the error sum of squares for the full model, while the "null hypothesis" procedure compares $\mathrm{SS}_{2|1}$ to $\mathrm{SSE}_1 = \mathrm{SS}_{2|1} + \mathrm{SSE}_{12}$,



$$
\begin{aligned}
a &= \|\mathbf{r}_1\| = \sqrt{\mathrm{SSE}_1},\\
b &= a\cos\theta = \|\mathbf{r}_1 - \mathbf{r}\| = \sqrt{\mathrm{SS}_{2|1}},\\
c &= a\sin\theta = \|\mathbf{r}\| = \sqrt{\mathrm{SSE}_{12}},\\
F_{\mathrm{null}} &= [(n-p_1)/p_2](b^2/a^2) = [(n-p_1)/p_2]\cos^2\theta,\\
F_{\mathrm{trad}} &= [(n-p)/p_2](b^2/c^2) = [(n-p)/p_2]\cot^2\theta.
\end{aligned}
$$

Figure 2 – Geometric representation of the decomposition of the sums of squares for testing a nested hypothesis in the general linear model.

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

$$F_{\mathrm{trad}} = \frac{\mathrm{SS}_{2|1}/p_2}{\mathrm{SSE}_{12}/(n-p)}$$

and

$$F_{\mathrm{null}} = \frac{\mathrm{SS}_{2|1}/p_2}{\mathrm{SSE}_1/(n-p_1)} = \frac{\mathrm{SS}_{2|1}/p_2}{(\mathrm{SS}_{2|1}+\mathrm{SSE}_{12})/(n-p_1)},$$

respectively.

4.2. *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 a stylized, two-dimensional display in Figure 2, along with the relationships between its various elements. Algebraic and trigonometric manipulations similar to those outlined in Section 2 show that F_{trad} is a one-to-one, increasing function of F_{null} over $(0,(n-p_1)/p_2)$, the set of possible values for F_{null}:

$$F_{\mathrm{trad}} = \frac{(n-p)F_{\mathrm{null}}}{n-p_1-p_2F_{\mathrm{null}}}. \tag{6}$$

Thus, as in the case of the t-test, tail-area tests using F_{trad} and F_{null} are identical. Note that, when $p=1$, $p_1=0$, and $p_2=1$, the relationship between F_{trad} and F_{null} given in Equation (6) 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,

$$
\begin{aligned}
b^2/\sigma^2 &= \mathrm{SS}_{2|1}/\sigma^2 \sim \chi^2_{p_2},\\
c^2/\sigma^2 &= \mathrm{SSE}_{12}/\sigma^2 \sim \chi^2_{n-p},
\end{aligned}
$$

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.

4.3. *Does the difference ever matter?*

While the test procedures based on F_{trad} and F_{null} produce identical inferences, the realized values of the test statistics are different. In this section we consider a situation in which, arguably, it is preferable to work with one of the two statistics rather than the other.

Residual plots are effective graphical devices for assessing the quality of the fit of a linear regression model and for detecting potential outliers. As noted in Section 9.4.1 of Weisberg (2014), a simple test for determining if observation i is an outlier in a regression model that includes p_1 predictors is to include an additional predictor which is an indicator of the observation in question (i.e., a 0-1 vector whose only element equal to 1 is the i-th one) and to test if the regression coefficient of the indicator is equal to zero.

Assuming normal errors for the regression model and letting $p_2 = 1$, it is natural to cast this problem into the framework of Section 4.1 and compare the full model with $p = p_1 + p_2$ predictors (the original predictors and the indicator of observation i) and the nested model that omits the indicator variable. Observation i is declared an outlier if the null hypothesis that the coefficient of its indicator variable is zero is rejected.

The traditional statistic for this problem is F_{trad}, which has an $F_{1,n-p}$ distribution under the null. The square root of F_{trad} (with sign matching the sign of the regression residual for observation i) is the usual t statistic for outlier detection described by Weisberg (2014). It is also a quantity known as the *studentized residual* for observation i, a normalized version of the raw residual, $\hat{e}_i$, computed using an estimate of the error variance, $\hat{\sigma}^2_{(i)}$, that *omits* observation i from the calculation. Conceptually, this point of view is appealing because, if the null hypothesis were violated and observation i were indeed an outlier, its inclusion in the calculation would inflate the estimate of the error variance. As stated in Weisberg (2014), the studentized residual can be expressed as

$$t_i = \frac{\hat{e}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}},$$

where h_{ii} denotes the leverage of observation i given by the i-th diagonal element of the projection (or hat) matrix $\mathbf{P}_{12}$ for the full model.

On the other hand, as seen in Section 4.1, the same test could also be performed using the statistic F_{null}. The signed square root of F_{null} turns out to be what is called the *standardized residual* for observation i, a normalized version of the raw residual, $\hat{e}_i$, computed using an estimate of the error variance, $\hat{\sigma}^2$, that uses all observations, *including* observation i. This would be the natural calculation to perform if one were to assume that the null hypothesis were true. As stated in Weisberg (2014), the standardized residual can be expressed as

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{ii}}},$$

and the deterministic relationship between studentized and standardized residuals is given by

$$t_i = r_i \sqrt{\frac{n-p}{n-p+1-r_i^2}}.$$

This deterministic relationship mirrors, on the square root scale, the deterministic relationship between F_{trad} and F_{null}. Ultimately, because of the deterministic relationships relating F_{trad}, F_{null}, and the two residual test statistics, an outlier test based on any of these four statistics leads to the same decision.

Residual plots are often used to conduct an exploratory assessment of the fit of the regression model. In this type of analysis, the plots are scanned visually for the existence of identifiable patterns and idiosyncratic features that might reveal violations of the modeling assumptions. With regard to outlier detection specifically, plots of residuals vs. fitted values are inspected to reveal the presence of unusually large residuals. We argue that, owing to the nonlinearity of the transformation that relates standardized residuals to studentized residuals, a studentized residual plot is better suited than a standardized residual plot to achieve this goal.

We illustrate this point with an example based on a subset of the data on brain and body weights for 100 species of placental mammals reported in Sacher and Staffeldt (1974). Here, for the measurements on the 21 species of primates included in the data set, we consider the simple linear regression of the natural logarithm of brain weight on the natural logarithm of body weight. Standardized and studentized residual plots are presented in the top row of Figure 3. Two species stand out: *Homo Sapiens* (with large positive residuals) and *Gorilla Gorilla* (with large negative residuals). Both are flagged as outliers at the 0.05 level with respective p-values of 0.0034 and 0.0301 (unadjusted for multiplicity of comparisons).

The extent to which these two species lie out compared to the other 19 species is clearly different. As evidenced visually in both plots, the residual for *Homo Sapiens* is further removed from the bulk of the residuals than the residual for *Gorilla Gorilla* and this impression is more notably accentuated in the studentized residual plot. This is due to the nonlinear relationship between standardized and studentized residuals which

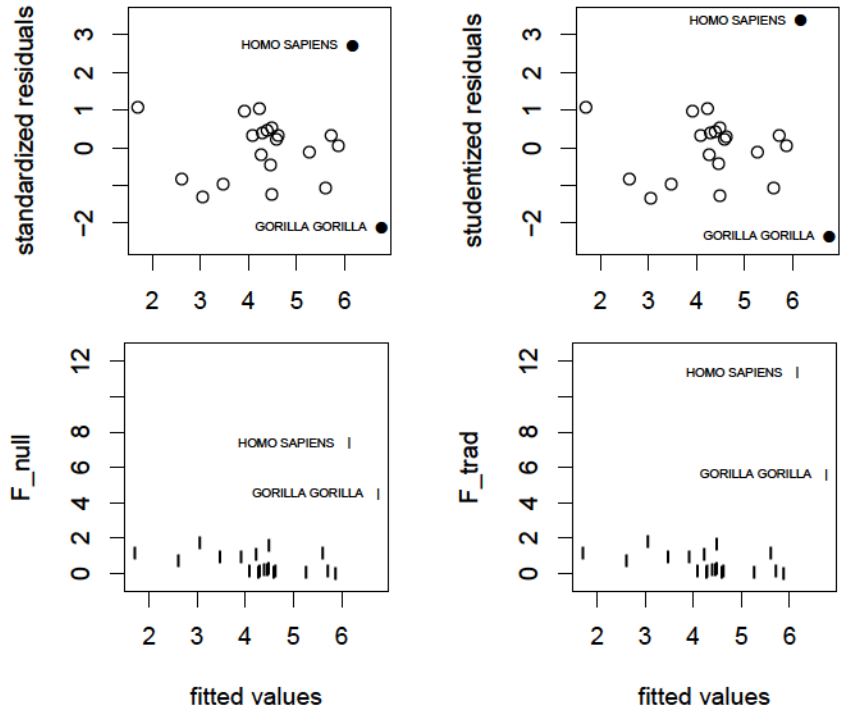


Figure 3 – Standardized and studentized residuals vs. fitted values for the primates data (top row) and their squared counterparts (bottom row).

causes the difference in absolute size between the two to increase monotonically as the absolute size of the standardized residual goes from 1 to infinity. In particular, as shown in Figure 4, the size of such difference becomes very noticeable when the absolute value of the standardized residual exceeds a value of about 2.5.

In our example, the absolute difference between studentized and standardized residuals is 0.6563 (very noticeable) for *Homo Sapiens*, 0.2394 (noticeable) for *Gorilla Gorilla*, and between 0.0011 and 0.0273 (hardly noticeable) for all other species. The displays in the bottom line of Figure 3, being based on F_{null} and F_{trad} which are the squared versions of the standardized and studentized residuals, emphasize even more the features just described. In summary, the displays based on the studentized residuals and on F_{trad} can focus the analyst's attention on the most extreme cases more effectively than those based on the standardized residuals and on F_{null}.

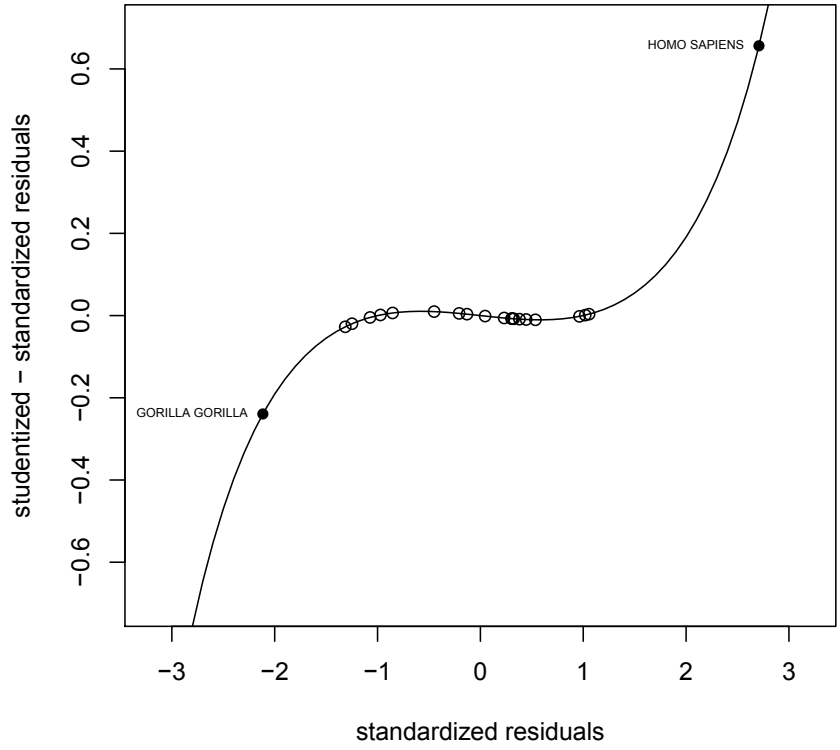


Figure 4 – Differences between studentized and standardized residuals vs. standardized residuals for the primates data. The solid line traces the deterministic relationship linking the plotted quantities.

5. The role of the null hypothesis in the construction of a test statistic

The fundamental question raised by the examples we presented in 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 something 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.

We consider first the case of an i.i.d. sample of size n from $f(x \mid \theta)$, a distribution indexed by a single parameter, θ, and rely on the results and examples presented in Casella

and Berger (2002). We denote by $L(\theta|\mathbf{X}) = f(\mathbf{X}|\theta)$ the likelihood function.

The score is defined as $S(\mathbf{X}|\theta) = d/d\,\theta \log f(\mathbf{X}|\theta)$. It can be shown that, for all θ, $ES(\mathbf{X}|\theta) = 0$ and $\mathrm{Var}S(\mathbf{X}|\theta) = I_n(\theta)$, the expected Fisher information in the sample. The point null hypothesis $H_0 : \theta = \theta_0$ is tested using the score test statistic $S(\mathbf{X}|\theta_0)/\sqrt{I_n(\theta_0)}$, which has mean 0 and variance 1 for all n, and, under appropriate regularity conditions, converges in distribution under the null to a standard normal as n goes to infinity, enabling the derivation of approximate cut-off values. Equivalently, the test can be based on the square of the score test statistic which has an asymptotic χ^2_1 distribution.

For n independent Bernoulli(p) observations yielding y successes, $\widehat{p} = y/n$ and the resulting score test statistic for testing $H_0 : p = p_0$ is the one given in Eq. (2). Its squared version is therefore

$$U_S(y, n; p_0) = \frac{(\widehat{p} - p_0)^2}{p_0(1-p_0)/n}.$$

Suppose that, for all θ, $W_n(\mathbf{X})$ is a consistent sequence of estimators of θ, having standard error $S_n(\mathbf{X})$. The Wald statistic for testing $H_0 : \theta = \theta_0$ is constructed as $(W_n(\mathbf{X}) - \theta_0)/S_n(\mathbf{X})$ and, if asymptotic normality holds, approximate cut-off values can again be derived under the null based on the quantiles of a standard normal. If the square of the Wald statistic is used for testing, approximate cut-offs should be based on the quantiles of a χ^2_1 distribution. Often $W_n(\mathbf{X})$ is taken to be the maximum likelihood estimator of θ, with $S_n(\mathbf{X}) = 1/\sqrt{I_n(W_n(\mathbf{X}))}$. Upon observing y successes out of n independent Bernoulli(p) trials, this recipe yields the statistic of formula (2), but with p_0 replaced by $\hat{p} = y/n$ in the denominator of that expression. The squared version of the statistic is therefore

$$U_W(y, n; p_0) = \frac{(\widehat{p} - p_0)^2}{\widehat{p}(1-\widehat{p})/n}.$$

The likelihood ratio test statistic for testing $H_0 : \theta = \theta_0$ is defined as

$$\lambda(\mathbf{X}) = \frac{L(\theta_0|\mathbf{X})}{\sup_\theta L(\theta|\mathbf{X})}.$$

Assuming appropriate regularity conditions, $-2\log\lambda(\mathbf{X})$ has an asymptotic χ^2_1 distribution under the null that can be used to obtain approximate cut-offs for the test. For the case of n independent Bernoulli(p) observations, denoting by y the total number of successes, the resulting likelihood ratio test will reject for large values of

$$U_L(y, n; p_0) = -2\log\left(\frac{p_0^y(1-p_0)^{n-y}}{\hat{p}^y(1-\hat{p})^{n-y}}\right).$$

Engle (1984) defines these three types of tests for the more general situation in which the parameter vector is multidimensional, including the case in which only a subset of

the parameters are of inferential interest while the remaining ones are regarded as nuisance parameters. A detailed recount of the insightful results presented there is beyond the scope of this article, but an important message is that, quite generally, the three types of tests will behave asymptotically similarly under the null and under local alternatives, although the asymptotic behavior for alternative values away from θ_0 will typically differ.

For finite samples the three statistics may yield different tests. The reason for this is illustrated in Figure 5 which presents scatter plots of the squared score, U_S, and squared Wald, U_W, statistics against the log-likelihood statistic, U_L, and of the squared score statistics, U_S, against the squared Wald statistic, U_W, for $n = 30$ and $p_0 = 1/3$. While these statistics are, separately, related monotonically for $\hat{p} \leq 1/3$ and $\hat{p} > 1/3$, the overall relationships are not monotonic. An examination of the rejection regions for these tests shows that the order in which the total number of successes enters the rejection region (as the size of the tests increase) differs among them. This is a situation in which the choice of which statistic to use matters.

As an example of a multidimensional situation including parameters of inferential interest and nuisance parameters, consider again the problem of testing a nested reduced model against the full model in the Gaussian linear model setting. There, the likelihood ratio test rejects the null hypothesis that the reduced model holds when the ratio

$$\lambda(\mathbf{Y}, \mathbf{X}) = \frac{\sup_{\beta_1, \sigma^2} L(\beta_1, \sigma^2 | \mathbf{Y}, \mathbf{X}_1)}{\sup_{\beta, \sigma^2} L(\beta, \sigma^2 | \mathbf{Y}, \mathbf{X})}$$

is small, or, equivalently, when the ratio $\mathrm{SSE}_1/\mathrm{SSE}_{12}$ of the error sum of squares under the reduced (null) model and the full model is large, ultimately leading to the equivalent tests based on F_{null} (a multiple of the score statistic as defined in Engle, 1984) and F_{trad} (a multiple of the Wald statistic as defined in Engle, 1984). This structure of the likelihood ratio test for nested models had already been noticed for the special case presented in Section 2, when discussing the derivation of the t-test in its two equivalent forms based on the ratio of Equation (5). Using the multiparameter definitions of the three types of test statistics, their deterministic functional relationships, and considering their asymptotic and finite sample distributions, Engle (1984) shows that the resulting tests are, in this case, equivalent both asymptotically and in finite samples.

6. 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,

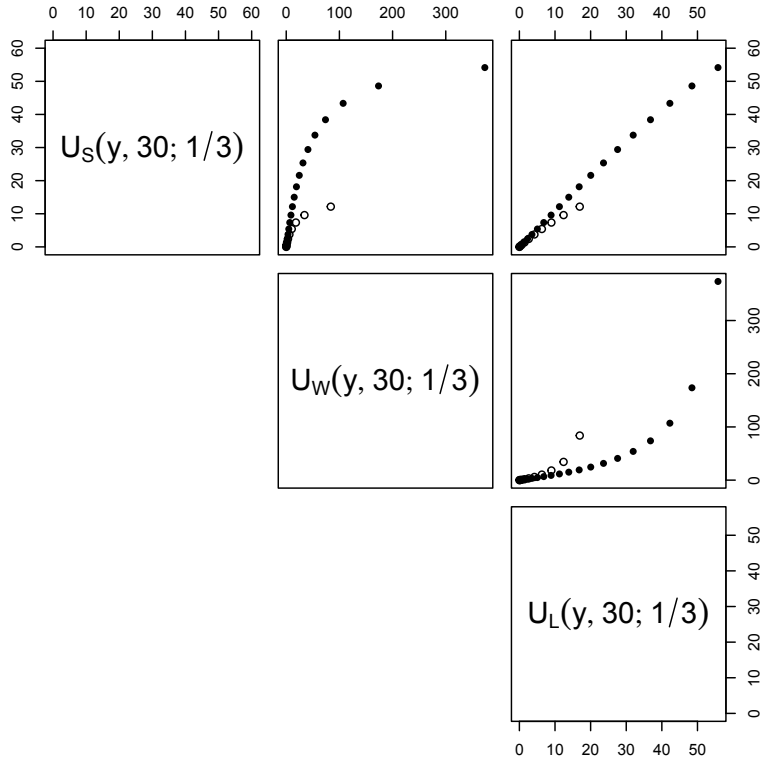


Figure 5 – Relationships between the squared score, U_S, squared Wald, U_W, and log-likelihood, U_L, test statistics for the case of independent Bernoulli data with $n = 30$ and $p_0 = 1/3$. The open plotting symbols correspond to values of y such that $\widehat{p} \leq 1/3$. The solid plotting symbols correspond to values of y such that $\widehat{p} > 1/3$. The statistics are not plotted for $y = 0$ and $y = 30$ to avoid cases where the Wald statistic is undefined.

and is, to our knowledge, not typically made in introductory statistics classes, even at the graduate level. We have also noted that the 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 calcu-

lation 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_{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).

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_{trad}, namely that its numerator and denominator are independent. Another feature worth noting is that the denominator of F_{trad} does not depend on the particular reduced model under consideration while the denominator of F_{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. Further, the graphical example of Section 4.3 illustrates that when the value of the statistic itself is of interest, rather than the formal testing decision, there may be practical reasons for preferring the use of one statistic over the other.

In Section 5 we reviewed three popular methods for building test statistics (the score, Wald, and likelihood ratio methods), discussing the different emphasis that they place on the null and alternative hypotheses. For all cases examined in this paper, the three methods yield asymptotically equivalent procedures while emphasizing different features of the testing problem. As noted in Engle (1984) this is related to the different metrics used to evaluate discrepancy between the null and the alternative. The Wald test accounts directly for differences in the parameter values, the likelihood ratio test measures differences in the log-likelihoods, and the score test assesses how steep the slope of the log-likelihood is at the null value. While under very general conditions the three methods yield procedures that are asymptotically equivalent, we have noticed that the resulting finite sample tests may differ for independent Bernoulli data. Engle (1984) presents additional examples where finite-sample conclusions might differ, comments on the different insight that the various formulations might bring to bear for specific models, and suggests that potential computational considerations might induce the analyst to opt for one of the tests over the other two.

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, are worth ruminating on, as they help us better see what considerations lead to the preference of one statistical procedure over another.

Choosing the right test statistic for a particular problem can be somewhat of an art, and understanding the similarities, differences, advantages, and disadvantages of the choice in the simple settings we considered may be helpful when turning to more complicated settings.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grants No. SES-1424481, No. DMS-1613110, and No. SES-1921523. The authors would like to thank the referee, who provided a very thorough review which improved the paper.

REFERENCES

A. AGRESTI, B. A. COULL (1998). *Approximate is better than "exact" for interval estimation of binomial proportions*. The American Statistician, 52, no. 2, pp. 119–126.

G. CASELLA, R. BERGER (2002). *Statistical Inference*. Duxbury-Thomson Learning, Pacific Grove.

F. CHAO, P. GERLAND, A. R. COOK, L. ALKEMA (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, no. 19, pp. 9303–9311.

R. F. ENGLE (1984). *Chapter 13 Wald, likelihood ratio, and Lagrange multiplier tests in econometrics*. Elsevier, vol. 2 of *Handbook of Econometrics*, pp. 775–826.

I. GOOD (1986). *Comments, conjectures, and conclusions: C258 editorial note on c257 regarding the t-test*. Journal of Statistical Computation and Simulation, 25, no. 3-4, pp. 296–297.

L. R. LAMOTTE (1994). *A note on the role of independence in t statistics constructed from linear statistics in regression models*. The American Statistician, 48, no. 3, pp. 238–240.

J. J. LEFANTE, JR., A. K. SHAH (1986). *C257. a note on the one-sample t-test*. Journal of Statistical Computation and Simulation, 25, no. 3-4, pp. 295–296.

E. L. LEHMANN (1986). *Testing Statistical Hypotheses*. John Wiley & Sons, New York.

D. S. MOORE, G. P. MCCABE, B. A. CRAIG (2012). *Introduction to the Practice of Statistics*. WH Freeman, New York.

G. A. SACHER, E. F. STAFFELDT (1974). *Relation of gestation time to brain weight for placental mammals: implications for the theory of vertebrate growth*. The American Naturalist, 108, no. 963, pp. 593–615.

A. K. SHAH, K. KRISHNAMOORTHY (1993). *Testing means using hypothesis-dependent variance estimates*. The American Statistician, 47, no. 2, pp. 115–117.

A. K. SHAH, J. J. LEFANTE JR (1987). *C293. a note on using a hypothesis-dependent variance estimate*. Journal of Statistical Computation and Simulation, 28, no. 4, pp. 347–349.

S. WEISBERG (2014). *Applied Linear Regression*. Wiley Series in Probability and Statistics. Wiley, New York. URL https://books.google.com/books?id=FHt-AwAAQBAJ.

S. YANG, K. BLACK (2019). *Using the standard wald confidence interval for a population proportion hypothesis test is a common mistake*. Teaching Statistics, 41, no. 2, pp. 65–68.