# TEACHING STATISTICAL INFERENCE WITHOUT NORMALITY

Christian Hafner

LIDAM Discussion Paper ISBA
2021 / 27

UCLouvain

LIDAM
ISBA | Louvain Institute of Data Analysis and Modeling in economics and statistics

# Teaching statistical inference without normality

Christian M. Hafner*

June 2021

**Abstract**

Teaching undergraduate statistics has long been dominated by a parallel approach, where large and small sample inference is taught side-by-side. Necessarily, the set of assumptions for the two cases is different, with normality and homoskedasticity being the main ingredients for the popular t- and F-tests in small samples. This paper advocates a change of paradigm and the exclusive presentation of large sample inference in introductory classes, for three reasons: First, it weakens and simplifies the assumptions. Second, it reduces the number of sampling distributions and therefore simplifies the presentation. Third, it may give better performance in many small sample situations where the assumptions such as normality are violated. The detection of these violations in small samples is inherently difficult due to low power of normality or homoskedasticity tests. Many numerical examples are given. In the era of big data, it is anachronistic to deal with small sample inference in introductory statistics classes, and this paper makes the case for a change.

**Keywords:** sampling distributions, simplification, big data

*Louvain Institute of Data Analysis and Modeling in economics and statistics, and ISBA, UCLouvain, Louvain-la-Neuve, Belgium. christian.hafner@uclouvain.be

# 1 Introduction

> *Reports that say that something hasn't happened are always interesting to me, because as we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns—the ones we don't know we don't know. And if one looks throughout the history of our country and other free countries, it is the latter category that tends to be the difficult ones.* D. Rumsfeld, U.S. Secretary of Defense, 2002.

Standard textbooks dealing with statistical hypothesis tests, confidence intervals, and other methods for statistical inference, are typically structured in the following way. Take, for example, the problem of testing the mean of some random variable $Y$, for which an i.i.d. sample $Y_1, \ldots, Y_n$ is available. The problem is to test the null hypothesis $H_0 : \mu = \mu_0$ against the alternative hypothesis $H_1 : \mu \neq \mu_0$. The test statistic, i.e. the famous "t-test", is constructed as the standardized sample mean $\bar{Y}$, i.e.

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

where $S$ is the sample standard deviation, and two cases are distinguished: (i) when $n$ is "large", $H_0$ is rejected at level $\alpha$ if $|T| > z_{1-\alpha/2}$, where $z_q$ is the $q$-quantile of the standard normal distribution, and (ii), when $n$ is "small", we need to add the normality assumption for $Y$, and $H_0$ is rejected at level $\alpha$ if $|T| > t_{n-1,1-\alpha/2}$, where $t_{n-1,q}$ is the $q$-quantile of the student-t distribution with $n-1$ degrees of freedom.

Millions of students have suffered to understand this procedure, which, to say the least, is misleading, and that for various reasons. First, it creates an artificial distinction between "large" and "small", with the necessity to define a threshold, e.g. $n = 30$[1]. Second, it is overly complicated adding two "small sample distributions" (student-t and Fisher) to the "large sample distributions" (normal and chi-square), so that students have

[1]Making students think that it is some kind of statistical law that for $n = 30$ (or larger) one has to take normal quantiles, and for $n = 29$ (or smaller) one has to take student-t quantiles.

to deal with four rather than two distributions. Third, the small sample approach may perform worse than the asymptotic approximation. In the above example, when $n = 20$ and $Y$ has a student-t distribution with two degrees of freedom, the "t-test" with student-t quantiles is undersized, i.e. the empirical size is about 0.035 for a nominal size of 0.05, while the test based on standard normal quantiles is correctly sized. Note that standard normality tests do not have much power to detect the underlying non-normality, due to the small sample size. In the given example, the often used Jarque-Bera (1980) test only has a power of about 45% at level $\alpha = 0.05$.

The argument becomes even more compelling when considering tests for the variance, because in that case, the statistics involve fourth moments. They are either implicitly assumed to be known (if $Y$ is assumed to be normally distributed), or estimated in the case of asymptotic tests. The asymptotic tests therefore explicitly take into account a potential deviation from normality, ensuring consistency even under non-normality. While in the testing of the mean problem there is only one statistic, the "t-test", we now have two different statistics in the testing of the variance problem. Under non-normality, only the asymptotic test is consistent. It may even give better performances in small samples, as we will show.

The "unknown unknowns" in small sample inference correspond to the unawareness of the consequences of false assumptions. Our recommendation is to ban small sample inference from statistics classes altogether. The necessary assumptions such as normality or homoskedasticity are much too strong, rarely hold in practice, and can hardly be verified. Only working with asymptotic tests involves weaker assumptions, simplifies the presentation and reduces the number of considered distributions. In the era of "big data", it appears anachronistic to still insist on small sample inference using twentieth century textbooks. The emperor has no clothes. It is time for statistics teaching to embrace the twenty-first century.

## 2 Estimation of the variance

Suppose $Y_1, \ldots, Y_n$ is an i.i.d. sample of $n$ observations drawn from a population random variable $Y$ with mean $\mu$ and variance $\sigma^2$. One of the first things we learn in statistics is that the common estimator of the variance $\sigma^2$, i.e.

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

where $\bar{Y}$ is the sample mean, has the denominator $n-1$, and not, as one would intuitively expect, $n$. The teacher then usually spends much time and effort to explain why it should be $n-1$, introducing the notions of bias, degrees of freedom, and perhaps even projection onto subspaces. This is also usually the point where students become suspicious about statistics.[2]

In fact, the alternative, natural, estimator that divides by $n$, let us call it $S_*^2$,

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

is not any worse than $S^2$. This is not only because, obviously, in large samples the difference does not matter[3] but also because $S_*^2$ may have better properties than $S^2$ in small samples. It depends on what criteria we use, and on the underlying distribution of $Y$.

Yes, $S^2$ is the minimum variance unbiased estimator of $\sigma^2$ under normality. No, it is not the most efficient estimator even under normality, because $S_*^2$ has a smaller mean squared error (MSE) than $S^2$. Assuming a finite kurtosis $\kappa$, straightforward calculations show that

$$MSE(S^2) = \frac{\sigma^4}{n-1}\left\{\kappa - 1 - \frac{\kappa - 3}{n}\right\} \quad (1)$$

$$MSE(S_*^2) = \frac{\sigma^4}{n^2} + \left(\frac{n-1}{n}\right)^2 MSE(S^2) \quad (2)$$

[2]The $n-1$-paradigm has even become a commonplace in society, where e.g. school calculators are equipped with $\sigma_{n-1}$-buttons. The R functions sd and var, by default, divide by $n-1$.

[3]it is $O_p(n^{-1})$, so the difference vanishes quickly as $n$ increases

| $n$ | 5 | 10 | 15 | 20 | 25 | 30 |
|---|---|---|---|---|---|---|
| $\kappa = 3$ | 1.524 | 1.221 | 1.139 | 1.102 | 1.080 | 1.066 |
| $\kappa = 30$ | 1.559 | 1.234 | 1.147 | 1.107 | 1.084 | 1.069 |

Table 1: Ratio of the MSE of $S^2$ wrt $S_*^2$.

and we see that $MSE(S^2) \geq MSE(S_*^2)$ uniformly over $\sigma^2$ and $\kappa$.

In small samples, the difference in efficiency in terms of MSE is striking, see Table 1. For example, for $n = 10$ and a kurtosis of 3, the relative efficiency loss of $S^2$ wrt $S_*^2$ is 22.1%. This is quite robust wrt the kurtosis, so e.g. for a kurtosis of 30, the efficiency loss is 23.4%. In even smaller samples, e.g. $n = 5$, the efficiency loss increases to 52.4% for a kurtosis of 3 and 55.9% for a kurtosis of 30.

Moreover, the proportion of the squared bias of $S_*^2$ to its MSE is negligible, and becomes even smaller for higher values of the kurtosis. For $n = 10$ and $\kappa = 3$, the proportion is 1.1% and for $\kappa = 30$ it is only 0.0008%. Thus, in relative terms, the squared bias of $S_*^2$ is negligible with respect to its variance.

The conclusion of this is that in large samples, the difference between $S^2$ and $S_*^2$ does not matter, while in small samples, $S_*^2$ is (much) more efficient. It is not true that $S^2$ has generally better small sample properties than $S_*^2$, this only holds for the bias. As soon as we take efficiency criteria such as MSE, the reverse is true. It is inconceivable to use, for example, Ridge regression, which introduces a small bias in order to reduce the MSE, when at the same time we stick to the unbiasedness paradigm of the variance estimator. Introducing degrees-of-freedom corrections (such as replacing $n$ by $n - 1$) can be counterproductive in small samples, and these are exactly the situations for which these corrections are designed. In fact, more generally, this extends to other degrees-of-freedom corrections, which might be more harmful than beneficial, see e.g. the section on linear regression.

# 3 Testing the mean

Turning to the genuine problem of statistical inference, we first discuss the classical test for a single mean, and then the test for the difference between the means of two populations.

## 3.1 Tests for a single mean

The problem is to test the mean $\mu$ of a random variable $Y$. We alternatively make the following assumptions.

(A1) $Y \sim N(\mu, \sigma^2)$

(A2) $E[Y] = \mu$, $Var(Y) = \sigma^2$, $\kappa := E[(Y-\mu)^4]/\sigma^4 < \infty$

Obviously, (A1) implies (A2), but not vice versa, so Assumption (A1) is (much) stronger than Assumption (A2).

We want to test $H_0 : \mu = \mu_0$. Under Assumption (A1),

$$T = \sqrt{n}\frac{\bar{Y} - \mu_0}{S} \sim t_{n-1} \tag{3}$$

by the definition of the student-t distribution.

Under Assumption (A2),

$$T_* = \sqrt{n}\frac{\bar{Y} - \mu_0}{S_*} \to_d N(0,1) \tag{4}$$

by the CLT and Slutsky's lemma.

If Assumption (A1) is correct, then $T \sim t_{n-1}$, and the test will have exact size and optimal power properties[4]. If Assumption (A1) is not correct, then $T$ has an unknown distribution. Using the $t_{n-1}$ distribution is an approximation, just as using the asymptotic $N(0,1)$, and it is not a priori clear which one has the better performance in terms of size and power.

[4]Under (A1), the t-test is equivalent to a likelihood ratio test, which is uniformly most powerful against simple or one-sided alternatives.

Even in small samples, the asymptotic approximation can be better than the exact distribution under presumed normality. Consider, for example, the case of an underlying student-t distribution with two degrees of freedom. Thus, the mean of $Y$ is $\mu$, but the variance is infinite. Even in this quite extreme case, normality tests have low power in small samples, see the lower panel of Figure 1. In other words, in practical situations with only ten or twenty observations, it is quite difficult to find out that the variance may not exist, or that the data are non-Gaussian. In this situation, the classical t-test in (3) is strongly undersized, as shown in the upper panel of Figure 1, which shows the estimated size of the tests based on 100,000 replications. The test based on (4), on the other hand, is remarkably close to the nominal size for the considered sample sizes. Of course, both tests are inconsistent in this case in the sense that the empirical size does not converge to the nominal size as $n$ increases to infinity. However, this example shows that situations exist where the test based on asymptotic distributions has better small sample properties than a classical test designed for small samples. Small sample inference pretends to be more precise than the asymptotic approximations, while often the opposite is true.

## 3.2 Testing the difference between the means of two populations

Consider now the problem of comparing the means of two populations, $Y_1$ and $Y_2$, of which two independent samples have been drawn, each of size $n$. We make the following two alternative assumptions.

(A3) $Y_1 \sim N(\mu_1, \sigma_1^2)$, $Y_2 \sim N(\mu_2, \sigma_2^2)$, and $\sigma_1^2 = \sigma_2^2$.

(A4) $E[Y_i] = \mu_i$, $Var(Y_i) = \sigma_i^2$, $\kappa_i := E[(Y_i - \mu_i)^4]/\sigma_i^4 < \infty$, $i = 1, 2$.

Note the second part of (A3) which assumes variance equality, or homoskedasticity. Thus, (A3) is stronger than (A4) not only in terms of the distribution, but also because (A4) allows for heteroskedasticity, which (A3) does not. In this section, we want to investigate whether or not homoskedasticity is innocuous for our testing problem, and to isolate this feature we take normality for granted, only for illustration and as a benchmark scenario.

The null hypothesis to be tested is $H_0 : \mu_1 = \mu_2$. Under (A3) and $H_0$,

$$\sqrt{n}\,\frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{S_1^2 + S_2^2}} \sim t_{2(n-1)} \tag{5}$$

Here, not only is normality crucial for (5) to hold, but also the equivalence of the two variances, $\sigma_1^2$ and $\sigma_2^2$, which are not tested and not of interest. If $\sigma_1^2 \neq \sigma_2^2$, then the test based on (5) is oversized, see Figure 2.

Alternatively, under the much weaker Assumption (A4), and without needing to impose $\sigma_1^2 = \sigma_2^2$, we have

$$\sqrt{n}\,\frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{S_{1*}^2 + S_{2*}^2}} \to_d N(0,1). \tag{6}$$

It is true that a test based on (6) is oversized in small samples, but simple size adjustments can be made without affecting the validity of the asymptotic distribution. For example, define the variance ratio $R$ as the ratio of the larger variance to the smaller, i.e. $R = \max(S_{1*}^2/S_{2*}^2, S_{2*}^2/S_{1*}^2)$. Then a size adjusted statistic would be given by

$$\sqrt{n}\,\frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{(S_{1*}^2 + S_{2*}^2)(1 + 3R^{-1/2}\log(R)/n)}} \to_d N(0,1). \tag{7}$$

The additional term in the denominator, $3R^{-1/2}\log(R)/n$, is $O_p(n^{-1})$ and hence does not affect the asymptotic distribution. Figure 2 shows the small sample performances of the classical test in (5) and of the size adjusted test in (7) for the case of a normal distribution and a variance ratio of 4, where the size of (7) is much closer to the nominal level. This also holds for other variance ratios[5]. The statistic in (7) is not meant to be a universally applicable test statistic. The size adjustment is certainly depending on the underlying distribution, which is unknown in practice. For the Gaussian case, however, this test has a remarkable performance and clearly outperforms the classical test for a wide range of variance ratios.

The lower panel of Figure 2 shows the power of a corresponding test for variance equality[6], i.e. $F := R$, and $H_0 : \sigma_1^2/\sigma_2^2 = 1$. Under normality and $H_0$, $F \sim F_{n-1,n-1}$.

[5] I have investigated variance ratios ranging from 1 to 25 and found excellent properties for the size-adjusted test.

[6] Here we use the classical F-test because in this section we took normality for granted in order to focus on the homoskedasticity assumption. For more general variance tests, see the next section.

The true variance ratio is 4, and the variance equality test has power larger than 80% in samples of size 20 or larger. So in practice one will usually be alerted not to use the classical test in (5) but rather use some form of size adjustment.

The conclusion of this is that the classical textbook statistic for comparing means of two populations is underperforming a size-adjusted test based on the asymptotic distribution in small samples for the case where homoskedasticity fails. In teaching, we should refrain from presenting the classical test as an "exact" small sample test, as the necessary assumptions (normality and homoskedasticity) are very unlikely to hold in most practical situations. Rather, my recommendation is to use the asymptotic test in (6), which requires neither normality nor homoskedasticity. Then, in the small sample case, size adjustments will be necessary anyway, depending on the underlying distribution and variance ratio, but this should not be presented in an introductory statistics class.

# 4 Tests for the variance

In the case $Y \sim N(\mu, \sigma^2)$, the textbook statistic is based on the result

$$V_1 := \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \tag{8}$$

where $\chi^2_{n-1}$ is the chi-square distribution with $n-1$ degrees of freedom. The normality assumption for $Y$ is crucial for this result.

Alternatively, one could consider the asymptotic result

$$V_2 := \sqrt{n/(\hat{\kappa}-1)}\ \left(S^2_*/\sigma^2 - 1\right) \to_d N(0,1), \quad n \to \infty \tag{9}$$

where $\hat{\kappa}$ is the sample kurtosis of $Y$ (without degrees of freedom corrections). The test in (9) does not require normality of $Y$ but only finite fourth moments. The proof of (9) uses the CLT, the law of large numbers (which ensures that $\hat{\kappa} \to_p \kappa$), and Slutsky's lemma.

Here we have two different test statistics. In small samples, where normality tests have no power, many scenarios exist where the test based on $V_2$ will have a better performance than $V_1$ in terms of size and power. In large samples and non-normality of $Y$,

blindly applying $V_1$ leads to biased and inconsistent tests for the variance, while $V_2$ is asymptotically unbiased and consistent.

We perform a small Monte Carlo study with 100,000 replications and an underlying Student-t distribution with five degrees of freedom. That is, the true variance is $\sigma^2 = 5/3$, and we apply the test statistics $V_1$ and $V_2$ to test $H_0 : \sigma^2 = 5/3$. Results are shown in Figure 3. Not surprisingly, the normality tests Shapiro-Wilk and Jarque-Bera (lower panel) have even less power than in Figure 1, as the $t_5$ distribution is closer to a normal than a $t_2$. For example, at a sample size of $n = 20$, the power of both tests is only about 20% at a nominal level of 5%. This means that in practice it will be quite unlikely to uncover non-normality in small samples should the true distribution be a $t_5$.

The upper panel of Figure 3 shows the estimated size of the two test statistics, $V_1$ and $V_2$. Both tests are quite severely oversized. However, only the $V_2$ is consistent, since the estimated size approaches the nominal level as $n$ increases. Moreover, the size distortion of $V_2$ is smaller than that of $V_1$ for all $n \geq 16$, and the gap widens as $n$ increases. Clearly, $V_2$ is preferable although size adjustments are necessary in small samples.

Consider now the problem of comparing the variances of two sub-populations, $Y_1$ and $Y_2$, of which two independent samples have been drawn, each of size $n$. The usual statistic for testing $H_0 : \sigma_1^2 = \sigma_2^2$ is the following. Under Assumption (A3),

$$\frac{S_1^2}{S_2^2} \sim F_{n-1,n-1}$$

where $S_1^2$ and $S_2^2$ are the sample variances of the two sub-samples. To emphasize, this only holds under normality of the two populations.

Alternatively, only assuming Assumption (A4),

$$\sqrt{\frac{n}{S_1^4(\hat{\kappa}_1 - 1) + S_2^4(\hat{\kappa}_2 - 1)}} \left(S_1^2 - S_2^2\right) \to_d N(0, 1)$$

where $\hat{\kappa}_1$ and $\hat{K}_2$ are the sample kurtosis coefficients of the two sub-samples.

# 5 Linear regression

When students are asked about the assumptions in a linear regression model, the usual answer is "the error term has to be normally distributed". When asked why, the response is "well, because otherwise the tests and confidence intervals are invalid". In worse cases, even the unbiasedness of the OLS estimator, or the Gauss-Markov theorem, are doubted. All of this is not true, of course.

The normality paradigm is pervasive in statistics, and it is particularly so in regression models. Countless times, students have estimated a model, tested normality of the residuals, rejected normality, and then tried to solve the "problem" by tinkering with the model specification, or transforming the response and/or explanatory variables, often making the situation worse. As we will see, inference about the regression model based on the asymptotic distribution, without the normality assumption, may even be more robust with respect to other phenomena such as heteroskedasticity.

The analogue of the "t-test" in regression models is the "F-test", another big monster of small sample inference, and every student of introductory statistics is afraid of it. It adds an additional layer of complexity, having two degrees of freedom parameters instead of just one. What is worse, it pretends to yield exact small sample inference, and hence be more precise than asymptotic approximations, while in fact many scenarios exist where the opposite is true.

To make the point, consider the classical linear model

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon \tag{10}$$

where $X_1, \ldots, X_p$ are explanatory variables and $\varepsilon$ is the error term. We then collect an i.i.d. sample of $(X_i, Y_i)_{i=1}^n$. In the case of stochastic explanatory variables, the usual assumption is that with probability one, the rank of the regressor matrix $X$ is equal to $p + 1$. The assumptions about the error term, in analogy to the testing of the mean problem, are as follows.

(A5) $\varepsilon | X \sim N(0, \sigma^2)$

(A6) $E[\varepsilon|X] = 0$, $Var(\varepsilon|X) = \sigma^2$, $\kappa := E[\varepsilon^4|X]/\sigma^4 < \infty$

Obviously, (A5) implies (A6), but not vice versa, so Assumption (A5) is (much) stronger than Assumption (A6).

We now want to test linear restrictions of the model, put in the general form

$$H_0 : A\beta = c$$

where $A$ is a fixed $J \times (p+1)$ matrix of rank $J$, $\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$, and $c$ is a fixed $(p+1)$-vector. Important special cases are, e.g., $\beta_1 = \ldots = \beta_p = 0$, or $\beta_j = 0$ for some $j, 0 \leq j \leq p$.

The restricted and unrestricted models are estimated by OLS. Call $SSR$ the sum of squared residuals of the unrestricted model, and call $SSR_0$ the sum of squared residuals of the restricted model under $H_0$. Then, we have the following results. First, the classical F-test, is obtained using Assumption (A5), for which under $H_0$,

$$F := \frac{SSR_0 - SSR)/J}{SSR/(n-p-1)} \sim F_{J,n-p-1} \tag{11}$$

Alternatively, using Assumption (A6), we obtain the asymptotic result under $H_0$,

$$H := \frac{SSR_0 - SSR}{SSR/n} \rightarrow_d \chi^2_J \tag{12}$$

To emphasize the difference, (11) is an exact finite sample result which only holds under normality of $\varepsilon$, while (12) is an asymptotic result which holds much more generally. The denominator of $H$ could of course be the same as in $F$, using a degrees of freedom correction, which does not change the asymptotic distribution. In that case, the $H$-statistic would simply be $J$ multiplied by the $F$-statistic. But that is not what I am advocating here, at least not for teaching undergraduates, where simplicity is most important.

Note that both tests rely on the assumption of homoskedasticity and absence of autocorrelation in the error terms. If this assumption is incorrect, then both tests are asymptotically invalid. In large samples this is typically not a problem, as the form of heteroskedasticity/autocorrelation can be detected, and the test statistics can be corrected to take this into account. In small samples, however, it may be very difficult to spot violations of this assumption, and diagnostic tests often have low power.

In the following I give an example where the H-test, based on the asymptotic distribution, has better properties in terms of size and power, than the F-test, in situations where homoskedasticity tests have low power. It is therefore misleading to pretend that the F-test is generally preferable to the simpler asymptotic test when homoskedasticity is not rejected.

Suppose we have a heteroskedastic linear model for $Y$ with three explanatory variables, i.e. $Y = \beta_0 + \sum_{i=1}^{3} \beta_i X_i + \varepsilon$, and suppose we want to test this model against the reduced model $Y = \beta_0 + \beta_1 X_1 + \varepsilon$, i.e. under $H_0$, $\beta_2 = \beta_3 = 0$. We set up a small simulation study to evaluate the size properties of both tests in small samples. The model under $H_0$ is generated with $\beta_0 = 0$ and $\beta_1 = 1$, and conditional on $X$, $\varepsilon$ is drawn from a normal distribution with mean zero and variance $1/X_2^2$. The explanatory variables $X$ are drawn from a tri-variate standard normal distribution. We generate 20,000 independent replications to estimate the distribution of the test statistics.

Figure 4 shows the estimated level of the F-test in (11) of $H_0 : \beta_2 = \beta_3 = 0$ as a function of the sample size $n$ vs the estimated level of the H-test in (12), at a nominal level of 5%. Both tests are inconsistent as they ignore the underlying heteroskedasticity. As a consequence, they are both undersized in larger samples. In small samples, however, the estimated size of the H-test is much closer to the nominal level of 5%. For example, for $n = 20$, the size of the F-test is 2.1% while that of the H-test is 4.9%.

The lower panel of Figure 4 shows the estimated power of the White test for heteroskedasticity as a function of the sample size $n$. Several versions of the White test exist, I chose the one that includes the explanatory variables and their squares into the auxiliary regression,

$$e_i^2 = \alpha_0 + \sum_{i=1}^{3} \alpha_i X_i + \sum_{i=1}^{3} \gamma_i X_i^2 + u_i \tag{13}$$

where $e_i$, $i = 1, \ldots, n$, are the OLS residuals of the regression (10). The White test statistic is then obtained as $nR^2$, where $R^2$ is the coefficient of determination of the auxiliary regression in (13). Under $H_0 : \alpha_i = 0, \gamma_i = 0, i = 1, 2, 3$, this statistic has an asymptotic $\chi^2$ distribution with 6 degrees of freedom. Including additionally cross-products of ex-

planatory variables does not change the results substantially. We see from Figure 4 that the White test has very little power against the present form of heteroskedasticity. At $n = 30$, it is still less than 10%.

In other words, in the present scenario of heteroskedasticity, which is very difficult to detect in small samples, we find that the simpler asymptotic chi-square test has better properties than the F-test. It is more robust against heteroskedasticity, at least in this case, and many other cases including error autocorrelation can be found.

# 6 About normality and homoskedasticity

The normal distribution plays a central role in statistics, and for good reasons. It is, however, not, as many people tend to think, a universal law for the distribution of independent random events.[7] Counterexamples are numerous. In fact, natural phenomena such as turbulent wind speeds or the distribution of sand deposits are better described by the generalized hyperbolic distribution, as has been shown by Barndorff-Nielsen (1977). In finance, it is well known since Mandelbrot (1963) and Fama (1965) that stock returns are not normally distributed, with Mandelbrot advocating the stable Paretian class of distributions. Prior to Mandelbrot normality was indeed the common model based on the work of Bachelier (1900).

One of the first things I teach the students in my Quantitative Risk Management class is the simple toy model for a random variable $Y$,

$$Y = \sigma Z, \quad Z \sim N(0, 1)$$

where the scale $\sigma$ is a positive random variable, independent of $Z$. It is then a simple exercise to show that the kurtosis of $Y$ is given by

$$\kappa = 3 + 3\frac{Var(\sigma^2)}{E[\sigma^2]^2}$$

[7] It is therefore inadequate to speak of a "normal" distribution, and we should rather use the term "Gaussian" distribution.

where the second term is the excess kurtosis and is strictly positive, unless $\sigma$ is a constant. In other words, whenever there is some randomness in the scale of a random event, leptokurtic distributions will be the rule rather than the exception.

In teaching statistics, it should be clearly distinguished between the modelling of distributions of natural phenomena, and sampling distributions of statistics or estimators. For the latter, at least for the classical problems, normality has its justification by the CLT. For the former, this is not the case and it very much depends on the type of data at hand. A priori, no normality assumption should be made for the distribution of phenomena unless one has a clear reason for it.[8]

This paper is not an obituary for the Fisher or the student-t distribution. Both have their merits in the modelling of random phenomena. The student-t is an often used model for data exhibiting fat tails, while the Fisher distribution is more flexible than the chi-square due to its additional parameter. The multivariate F distribution, for example, has been used as a model for realized covariance matrices, being more flexible than the Wishart distribution which is the multivariate generalization of a chi-square.

This paper claims to be, however, an obituary for the Fisher and the student-t distribution as devices for small sample inference, where the parameters are determined by the "degrees-of-freedom" of some test statistic.[9] In practice, better choices are available, so that we should not teach our students to work under restrictive, unrealistic assumptions.

# 7 Conclusions

As a child of the 20th century, my education in statistics and econometrics was dominated by textbooks written in an era where computers just emerged as a new technology and

[8] One such reason might be the use of aggregated data, for example economic time series that are aggregated over time, say at a monthly or quarterly frequency. If at the data generating frequency the data is i.i.d. or martingale difference, then the aggregated series will be approximately Gaussian due to the CLT.

[9] In fact, we should stop calling the parameters of a student-t and an F-distribution "degrees-of-freedom".

where numerical examples consisted of few observations in low dimensions. No high-dimensional data analysis, no machine learning, no data mining. Skimming through undergraduate textbooks today, my impression is that not much has changed.

This paper proposes a simplification in the way introductory statistics is taught. We have seen that in many small sample situations, the classical textbook statistics should be treated with care, because the assumptions are too strong and alternative devices may show better performances. Of course, the cases considered in this paper are not exhaustive, but rather serve as illustrations of the main idea. The same type of arguments apply to other inference problems, for example in analysis of variance (ANOVA), or in time series models. However, this paper has not covered randomization based inference, as surveyed e.g. by Rosenbaum (2002), which is a completely different approach and therefore not affected by our critique.

For the teaching of undergraduate statistics classes, I draw the following conclusions and recommendations:

1. Simplify the presentation of statistical inference
2. Do not use degree-of-freedom corrections
3. Do not assume normality
4. Do not use small sample statistics such as the t-test or the F-test.
5. Instead, assume finite fourth moments (much weaker) and use asymptotic tests.
6. The Student-t and Fisher distributions have their merits for modelling the distributions of random phenomena, but should not be used for teaching statistical inference.

Of course, my perception of statistical inference today has been strongly influenced by the "big data" environment in which we are living. Unless there are restrictions for data privacy, governments, companies and private persons are usually drowning in data. Even for public health issues such as the Covid-19 pandemic, huge datasets are available

for everyone to download. Why should we care about our textbook example with $n = 13$ observations?

Beyond teaching, this approach raises the natural question what to do in practice whenever only a small sample is available, as it may occur in clinical trials or surveys, and it is not possible to increase the sample. Again, my recommendation is to be extremely careful with strong assumptions, which might be convenient but whose consequences are often not understood. Rather, I would suggest to consider alternative devices, under different assumptions. If all approaches hint in the same direction, one can be more confident with the decision. If they are different, however, then the results are inconclusive, and in these particular cases one should refrain from using statistics as a decision-making tool.

Because decision making in small sample situations is so sensitive with respect to the assumptions, it should be reserved to advanced statistics classes and research centers. It has no room in introductory statistics classes. There, the gain in time could be used in many ways, depending on the curriculum, for example paving the way for more advanced statistics, data mining or machine learning classes.

## Acknowledgments

I would like to thank Helmut Herwartz, Oliver Linton, and Johan Segers for helpful discussions and comments on a preliminary version of this paper.

# Bibliography


Bachelier, L. (1900). Theorie de la speculation, Annales de l'Ecole Normale Superieure, 17, 21-86.

Barndorff-Nielsen, Ole (1977). Exponentially decreasing distributions for the logarithm of particle size, Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences. The Royal Society. 353 (1674): 401–409.

Fama, E.F. (1965). The behavior of stock market prices, Journal of Business, 38, 34-105.

Jarque, C.M. and Bera, A.K. (1980). Efficient tests for normality, heteroskedasticity and serial independence of regression residuals, *Economics Letters*, 6, 255-259.

Mandelbrot, B. (1963). The variation of certain speculative prices, Journal of Business, 36, 394-419.

Rosenbaum P.R., Observational Studies. Springer; 2002.

Shapiro, S. S.; Wilk, M. B. (1965). "An analysis of variance test for normality (complete samples)". Biometrika. 52 (3–4): 591–611.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica*, 48, 817-838.

# Appendix: Figures

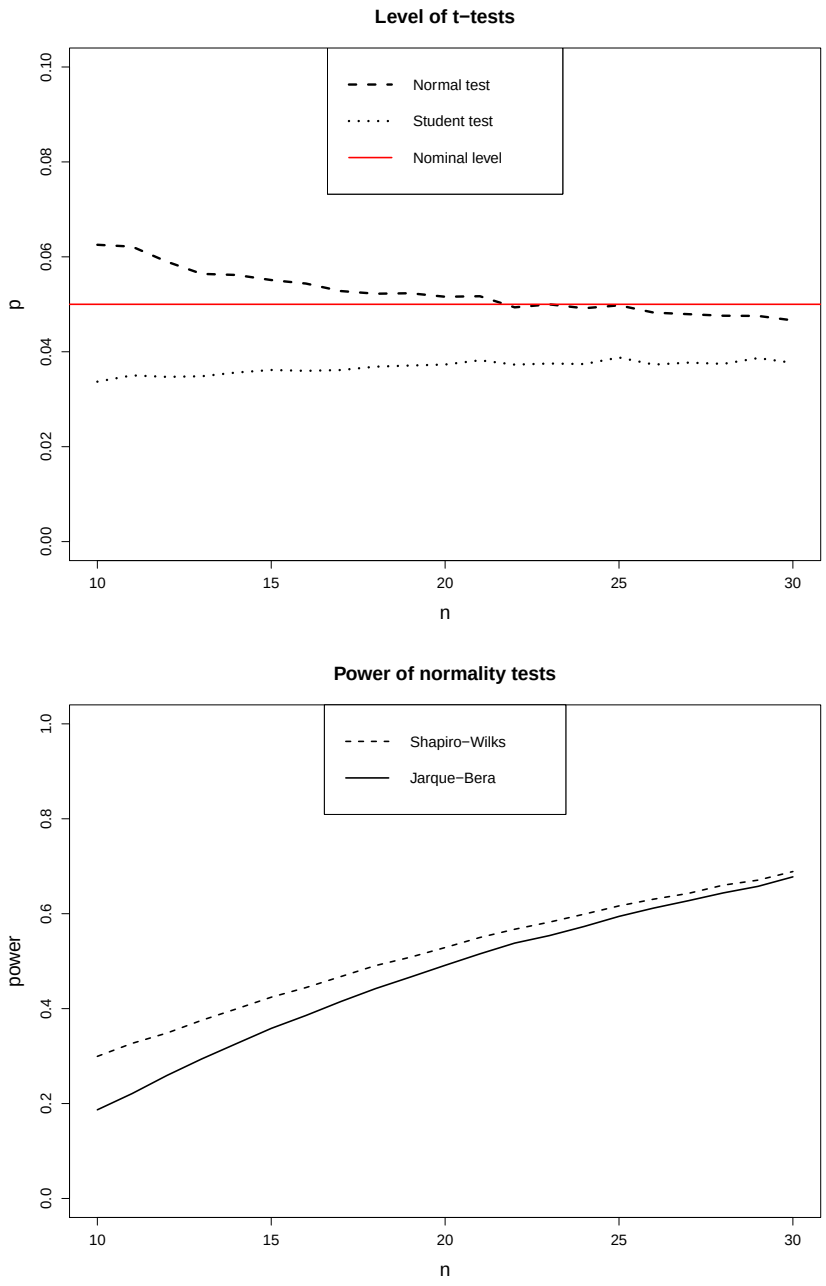


Figure 1: Upper panel: Estimated level of the t-test as a function of the sample size $n$ using student-t quantiles (dotted line) vs standard normal quantiles (dashed line), at a nominal level of 5%. Lower panel: Estimated power of the Shapiro Wilk (1965, dashed line) and Jarque Bera (1980, solid line) normality tests as a function of the sample size $n$. The underlying distribution is a student-t with 2 degrees of freedom.

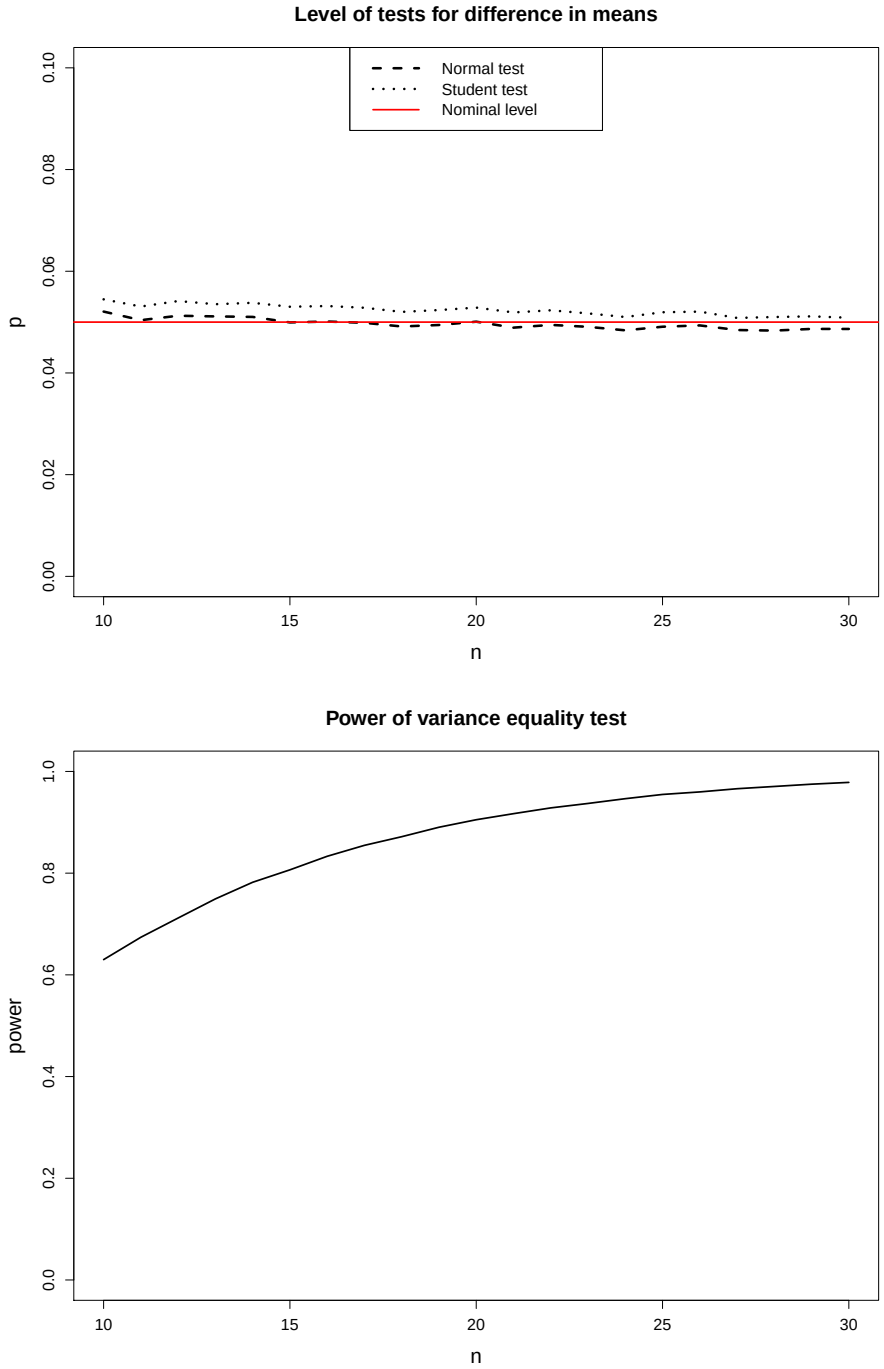


Figure 2: Upper panel: Estimated level of the test based on (5) as a function of the sample size $n$ (dotted line) vs the test based on (7) (dashed line), at a nominal level of 5%. Lower panel: Estimated power of the Fisher variance equality test as a function of the sample size $n$. The underlying distribution is a standard normal.

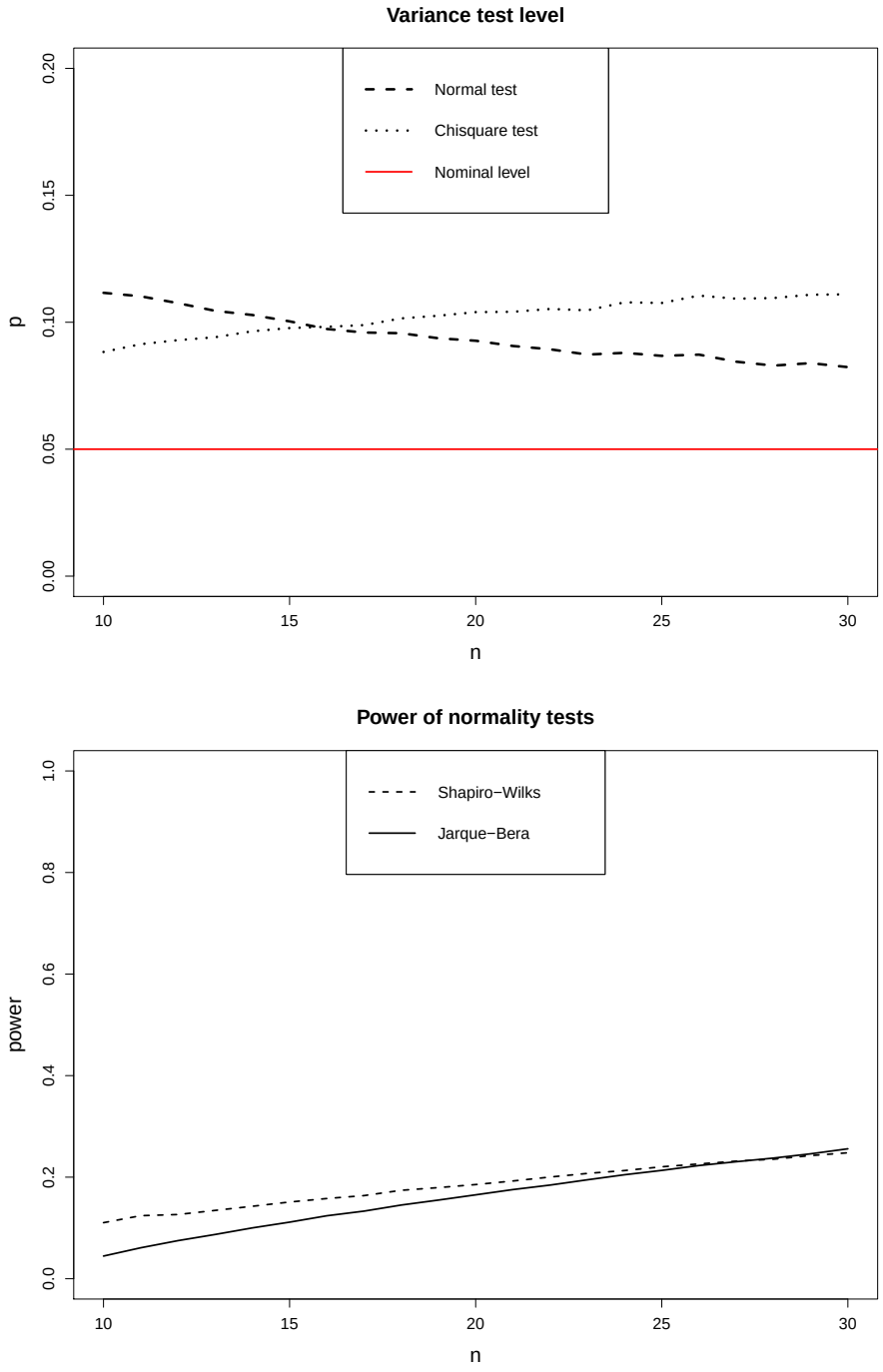


Figure 3: Upper panel: Estimated level of the chi-square test $V_1$ based on (8) as a function of the sample size $n$ (dotted line) vs the test $V_2$ based on (9) (dashed line), at a nominal level of 5%. Lower panel: Estimated power of the Shapiro Wilk (dashed line) and Jarque Bera (solid line) normality tests as a function of the sample size $n$. The underlying distribution is a student-t with 5 degrees of freedom.

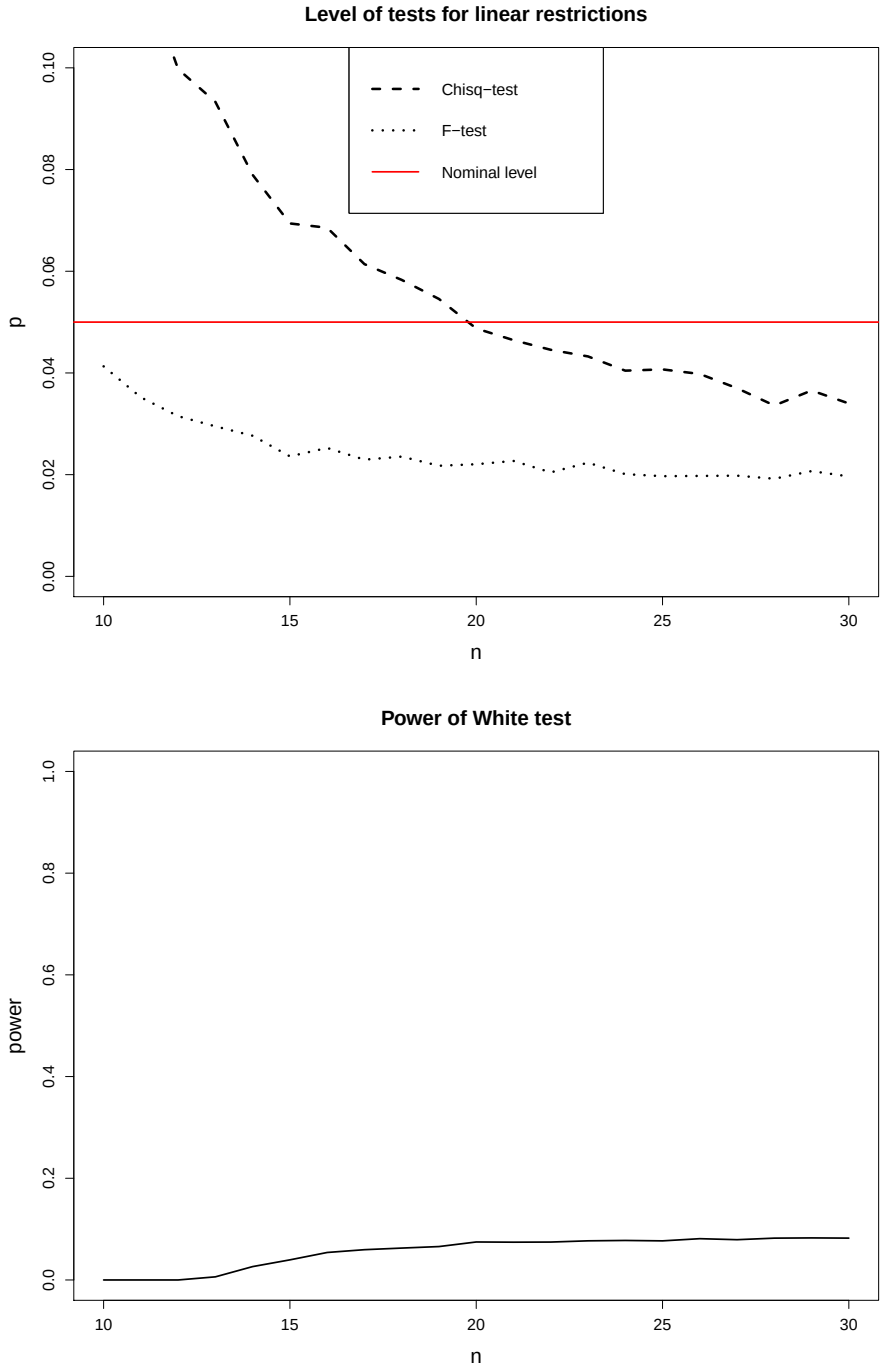


Figure 4: Upper panel: Estimated level of the F-test of $H_0 : \beta_2 = \beta_3 = 0$ in the model $Y = \beta_0 + \sum_{i=1}^{3} \beta_i X_i + \varepsilon$ as a function of the sample size $n$ (dotted line) vs the chi-square test (dashed line), at a nominal level of 5%. Conditional on $X$, $\varepsilon$ has a normal distribution with mean zero and variance $1/X_2^2$. The model under $H_0$ is generated with $\beta_0 = 0$ and $\beta_1 = 1$. $X$ is drawn from a tri-variate standard normal distribution. Lower panel: Estimated power of the White (1980) test for heteroskedasticity as a function of the sample size $n$.