Characterizing the Power of the t-test for Heavy Tailed Data

Samuel P. Engle*

December 4, 2021

Abstract

The t-test is a standard inferential procedure in economics and finance. When the data exhibit heavy tails, the t-test may have low power. This paper characterizes the rate at which power converges to 1 for data in a particular class of heavy tailed distributions. While classical results on the rate of convergence of power focus on exponential rates, we find the rate to be a much slower polynomial rate when the data have heavy tails. We compare these results with other results on the efficiency of the t-test in the literature, and use empirically-calibrated simulation evidence to demonstrate how our results make good finite-sample predictions.

1 Introduction

Since its introduction in Student (1908), the usual t-test for inference about the mean has played a ubiquitous role in theory and practice in econometrics and statistics. Initially motivated as the optimal test in the canonical inference problem with Gaussian observations, asymptotic arguments extend the application of the t-test to scenarios in which the data are not normally distributed. Heavy-tailed data are a particular departure from normality that has been of increasing interest. This paper develops new results characterizing the power of the t-test when the underlying distributions have heavy tails.

The main contribution of this paper is establishing the rate at which the power of the t-test converges to 1 under a fixed alternative when the data exhibit heavy tails. In classical settings, when the moment generating function exists the type-II error rate disappears exponentially quickly in sample size for any fixed alternative. We show that when the moment

*email: sengle2@wisc.edu. website: `https://www.samuelpengle.com`

generating function does not exist, under a different set of regularity conditions type-II error disappears at a polynomial rate. Our baseline results also apply in stylized regression and simultaneous equation settings.

Our results complement existing results in the statistical and econometric literature on using the t-test with heavy-tailed data. Recently, Müller (2019) and Müller (2020) establish slow rates of convergence of the t-test statistic to a standard normal random variable under the null hypothesis when the data exhibit Pareto-like tails. We find that similar slow-convergence results hold when looking at the type-II error rate under a fixed alternative. Shephard (2020) proposes an alternative estimator in regression settings with heavy-tailed data. Young (2021) argues that not only are heavy-tailed data prevalent in economic applications, but heteroskedastic-robust inference is particularly sensitive to heavy-tailed data.

We also contribute to a long history in the statistical literature on the properties of the t-test statistic in heavy-tailed settings. Under symmetry conditions, Efron (1969) shows that the t-test will tend to be asymptotically conservative when the tails of the data are sufficiently heavy. Giné et al. (1997) provide necessary and sufficient conditions for the t-test statistic to be asymptotically standard-normal or subgaussian. Shao (1999) establishes large-deviation results for the t-test statistic.

This paper also contributes to the literature on asymptotic efficiency of the t-test. Hodges and Lehmann (1956a) was the first paper to propose the relative efficiency measure we adopt in this paper, and they derived the efficiency of the t-test when the observations are normally distributed. Recently, He and Shao (1996) derived the Bahadur efficiency of closely-related normalized score tests. Their results show that t-tests are reasonably robust to heavy-tails when considering the behavior of p-values. Our results show that heavy-tails lead to slow convergence of the type-II error rate to zero, and therefore provide a new and different perspective.

For this paper we focus on the cases where we have an i.i.d. sample $X_1, \ldots, X_n$, with $\mathbb{E}\, X_i = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$. The tests we consider are hypothesis tests of the form:

$$H_0 : \mu = \mu_0 \qquad \text{v.s.} \qquad H_1 : \mu > \mu_0 \tag{1}$$

We will mainly focus on the behavior of the t-test, however we will also present results for the z-test for comparison. In each case the null hypothesis is rejected for sufficiently large values of the test statistic

$$Z_n := \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma} \qquad T_n := \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} \tag{2}$$

where $\bar{X}$ is the sample mean and S^2 is the sample variance. T_n is the classic t-test statistic, and Z_n is what we will refer to as the z-test statistic. In both cases, we reject the null hypothesis when the test statistic is sufficiently large. Note that the z-test statistic is generally not available, however relative to the t-test statistic, the properties of Z_n are easier to derive.

This paper proceeds as follows. In Section 2, we present our main result and compare with other common relative efficiency results in this setting. In Section 3 we show how these results could be used when estimating linear models. In Section 4 we provide some simulation evidence. In Section 5 we conclude.

2 Efficiency of the t-test

We first present the main results of this paper: a characterization of the asymptotic type-II error rate for tests using Z_n and T_n when the observations have heavy tails. We then compare these results to relative efficiency comparisons based on local asymptotic power and Bahadur relative efficiency.

2.1 Hodges-Lehmann Relative Efficiency

The Neyman-Pearson testing paradigm associates low type-II error rates at a fixed level α with good testing performance. Most conventional tests result in the type-II error rate converging to 0 asymptotically, which is part of the inherent challenge in comparing tests using asymptotic methods. One way to avoid this technical roadblock is to evaluate the rate at which the type-II error rate of a test converges to 0. This is precisely the motivation for the relative efficiency measure first proposed in Hodges and Lehmann (1956b). Consider a test using test statistic W_n for testing (1), where the null hypothesis is rejected when $W_n > C_W$ for some C_W. The Hodges-Lehmann (HL) relative efficiency is typically defined as

$$\lim_{n\to\infty} -\frac{1}{n}\log P_\mu(W_n < C_W) \tag{3}$$

where P_μ denotes that the probability is computed under the alternative $\mathbb{E}\, X_i = \mu$. In many cases of interest, (3) is zero because the convergence rate is not exponential.

Assumption 1. *The tails of X_i are regularly varying: there exists a slowly-varying function*[1] *L and $\gamma > 2$ such that*

$$\lim_{x\to\infty} \frac{P(|X_i| > x)}{L(x)x^{-\gamma}} = 1 \tag{4}$$

[1] A function L is slowly varying (at infinity) if for all $t > 0$, $L(tx)/L(x) \to 1$ as $x \to \infty$.

Further, the X_i satisfy a tail balance condition:

$$\lim_{x \to \infty} \frac{P(X_i > x)}{P(X_i < -x)} \in (0, \infty) \tag{5}$$

An interpretation of Assumption 1 is that the tails of X_i are eventually well approximated by a power law, as one looks sufficiently far out in the tail. Furthermore, the assumption allows for skewness but does not allow for the rate of decay of the tails to be different.

Theorem 1. *When the X_i are i.i.d. and Assumption 1 is satisfied, we have that:*

$$\lim_{n \to \infty} \frac{n^{\gamma-1}}{L(n)} P(Z_n < C_\alpha) = (\Delta/\sigma)^{-\gamma} \tag{6}$$

$$\lim_{n \to \infty} \frac{n^{\gamma-1}}{L(n)} P(T_n < C_\alpha) = (\Delta/\sigma)^{-\gamma} \tag{7}$$

Remark 1. For each fixed alternative, the rate at which the type-II error converges to 0 for both the z-test and t-test as a power of n. This implies that efficiency comparisons based on (**??**) do not capture asymptotic behavior for a broad class of data generating processes, and in fact treats them as equivalent (the limit in (**??**) is equal to zero whenever Assumption 1 holds). In addition, the polynomial rate of convergence implies that larger samples are required for small type-II error rates than is assumed when the convergence rate is exponential.

Remark 2. The type-II error rate asymptotically obeys a power law in the alternative. This implies that even in large samples, the t-test might not be able to detect differences of practical significance if they are too small.

Remark 3. The proof is based on results from Cline and Hsing (1989) and Mikosch and Nagaev (1998). In those papers, large deviation results for sums of i.i.d. random variables with heavy tails. The essential idea is that when the tails of the X_i are heavy, asymptotically large-deviation probabilities of a sum are equal to large deviations of the maximum.

2.2 Comparison With Local Asymptotic Power

The most common approximation of the asymptotic power of tests is local asymptotic power. Under the i.i.d. and finite variance assumptions, we have that

$$Z_n \Rightarrow \mathcal{N}(0,1), \qquad T_n \Rightarrow \mathcal{N}(0,1) \tag{8}$$

under the null hypothesis. To compute the local power approximation, we introduce a sequence of alternatives μ_n, such that $\mathbb{E}\, X_i = \mu_n = \mu_0 + \delta/\sqrt{n}$. Under this sequence of alternatives, we have that the test statistics converge to shifted normal random variables:

$$Z_n \Rightarrow \mathcal{N}(\delta/\sigma, 1), \qquad T_n \Rightarrow \mathcal{N}(\delta/\sigma, 1) \tag{9}$$

An implication of (9) is that under a local asymptotic power comparison, the power properties of each test are the same for all distributions with the same variance σ^2. By contrast, in Theorem 1, not only does the asymptotic power depend on the variance σ, it also depends on the tail parameter γ. Thus, our results provide for a finer distinction relative to local asymptotics in this setting.

2.3 Bahadur Relative Efficiency

Another notion of relative efficiency, due to Bahadur (1960), is to compare the rate at which the p-values of a test converge to 0 under a fixed alternative. If we denote the sequence of distribution functions of the test statistic W_n under the null as G_n, then the sequence of p-values is given by:

$$1 - G_n(W_n) \tag{10}$$

In He and Shao (1996), it is shown that the self-normalized sum obeys a large deviation result which implies exponential convergence of the p-values under a fixed alternative. The self-normalized sum in the context of hypothesis testing for the mean is:

$$S_n := \frac{\sum_{i=1}^{n} X_i - \mu_0}{\sqrt{\sum_{i=1}^{n} (X_i - \mu_0)^2}} \tag{11}$$

S_n and T_n are related by a 1-to-1 transformation, and thus it can be shown that the results in He and Shao (1996) can be adapted to show that under Assumption 1,

$$\lim_{n\to\infty} \frac{1}{n} \log(1 - G_n(T_n)) = \log\left(\sup_{c\geq 0} \inf_{t\geq 0} \mathbb{E} \exp\left\{2tcX_i - \frac{\Delta}{\sigma} \frac{t(c^2 + X_i^2)}{\sqrt{1 + \Delta^2/\sigma^2}}\right\}\right) \tag{12}$$

This result implies that for the t-test statistic, p-values converge to 0 at an exponential rate. Notice that this suggests that T_n should be used rather than Z_n: for Z_n, under our heavy-tailed assumptions, we have that, as a direct application of Proposition 3.1 in Mikosch and Nagaev (1998), we have that under Assumption 1

$$\lim_{n\to\infty} \frac{n^{\gamma-1}}{L(n)} \left(1 - G_n\left(Z_n\right)\right) = \nu^{-\gamma} \tag{13}$$

Thus, when using Bahadur relative efficiency to compare Z_n and T_n, T_n appears to have additional robustness of well-controlled p-values. This contrasts with our Theorem 1, where under Hodges-Lehmann relative efficiency, the two tests have the same properties with respect to Type-II error.

3 Application to Linear IV Models

Consider the simplest linear IV model:

$$y_i = x_i \theta + \varepsilon_i$$
$$x_i = z_i \pi + \upsilon_i$$

where $y_i,\ x_i,\ z_i \in \mathbb{R}$. The conditional moment condition is satisfied so that $\mathbb{E}((\varepsilon_i,\ \upsilon_i)'|z_i) = 0$. We will also assume strong identification: $\pi^2 \geq C > 0$. We can accommodate additional control variables by regressing y_i, x_i, and z_i on the controls, then proceeding by residual regression. With this in mind, we also assume $\mathbb{E}\, z_i = 0$. We conjecture that our results in this section extend to the case when x_i and z_i are possibly vector-valued and the model is possibly over-identified, however we leave such results to future work.

We consider a variant of the standard Wald test in this setting to test

$$H_0 : \theta = \theta_0 \qquad \text{v.s.} \qquad H_1 : \theta > \theta_0 \tag{14}$$

We consider a slightly different heteroskedastic robust test statistic. Let $W_i := (y_i - x_i\theta_0) z_i$, and let $U_i := (y_i - x_i\theta_0)\operatorname{sgn}(z_i)$. Note that testing (14) is equivalent to testing $\mathbb{E}\, W_i = 0$ or $\mathbb{E}\, U_i = 0$ against a one-sided alternative. Our test statistics we consider are

$$R_n := \frac{\sqrt{n}\bar{W}}{S_W} \qquad Q_n := \frac{\sqrt{n}\bar{U}}{S_U} \tag{15}$$

which are the t-tests formed from W_i and U_i respectively. We assume that Assumption 1 is satisfied for W_i and U_i with γ_W and γ_U respectively. Notice we must have that $\gamma_U > \gamma_W$. If $\gamma_U < \gamma_W$, this implies that there exists an $\epsilon > 0$ such that $\mathbb{E}\,|U_i|^{\gamma_U+\epsilon} = \infty$ but $\mathbb{E}\,|W_i|^{\gamma_U+\epsilon} < \infty$. Clearly, U_i will have more moments than W_i. Note that in the case of linear regression, where $z_i = x_i$, using W_i is similar to the Wald test statistic using the OLS estimator and heteroskedastic robust standard errors, while U_i is similar to using the estimator proposed in Shephard (2020). In any case, since $\gamma_U > \gamma_W$, we have that by Theorem 1, Q_n is more efficient than R_n when the observations have heavy tails. This potentially contradicts the ordering provided by local asymptotic power. In the case of homskedasticity, z_i is the optimal instrument, and therefore R_n is an efficient test under local asymptotic power comparisons. For relatively weak forms of heteroskedasticity, R_n will still be preferred to Q_n under local comparisons. This suggests that when power is close to zero, R_n will have more power than Q_n. Our results in this paper imply that power will converge to 1 at a faster rate for Q_n than for R_n. Thus, determining how to choose a test statistic relates to the region practitioners want higher power in the alternative space.

4 Simulations

We follow Shephard (2020) and provide simulation results calibrated to arithmetic returns from the SPDR S&P 500 ETF Trust (SPY), using data from August 1st 2018 to August 4th 2020. We simulate data from the following DGP:

$$\begin{aligned} y_i &= (z_i - \psi)\beta_1 + \varepsilon_i \\ z_i &= \psi + V_\nu \sigma_Z \sqrt{\frac{\nu - 2}{\nu}},\ V_\nu \sim t_\nu \\ \varepsilon_i &\sim \mathcal{N}(0, (1 + |z_i - \psi|^\zeta)C^2) \end{aligned} \tag{16}$$

Here, z_i are calibrated to match the weekly returns. We set $\psi = 0.21$ to match the average weekly returns, and likewise set $\sigma_Z = 3.24$ to match the sample standard deviation. Fitting the normalized z_i to a student-t distribution leads to an implied estimate of ν of 2.16; as in Shephard (2020), we set this to 2.4, for slightly different reasons. Unlike that paper, we include conditional heteroskedasticity. The form of (16) is motivated so that for some choices of ζ, R_n and U_n as defined in (15) both lead to asymptotically valid inference under the null. We choose $\zeta = 0.3$ with this in mind. C is chosen so that $\mathbb{E}\,\varepsilon_i^2 = 4$, as in Shephard (2020). We set $\beta_{1,0} = 1$. To remove issues with the Wald-statistic based on the OLS estimator having poor size control, we first compute the finite-sample critical values to lead to valid inference. We set the sample size to $n = 100$.

We use two different test statistics to test the point null $H_0 : \beta_1 = 1$. The first is R_n, which is similar to the standard heteroskedastic-robust Wald test using the OLS estimator. The second test statistic is Q_n, which is similar to the test statistic proposed in Shephard (2020)

Figure 1 plots the estimated power for a two sided test that $\beta_1 = 1.0$ based on 10000 Monte Carlo draws. On the left we see power, and on the right we see the power loss from using R_n instead of Q_n, on both an absolute and relative scale.

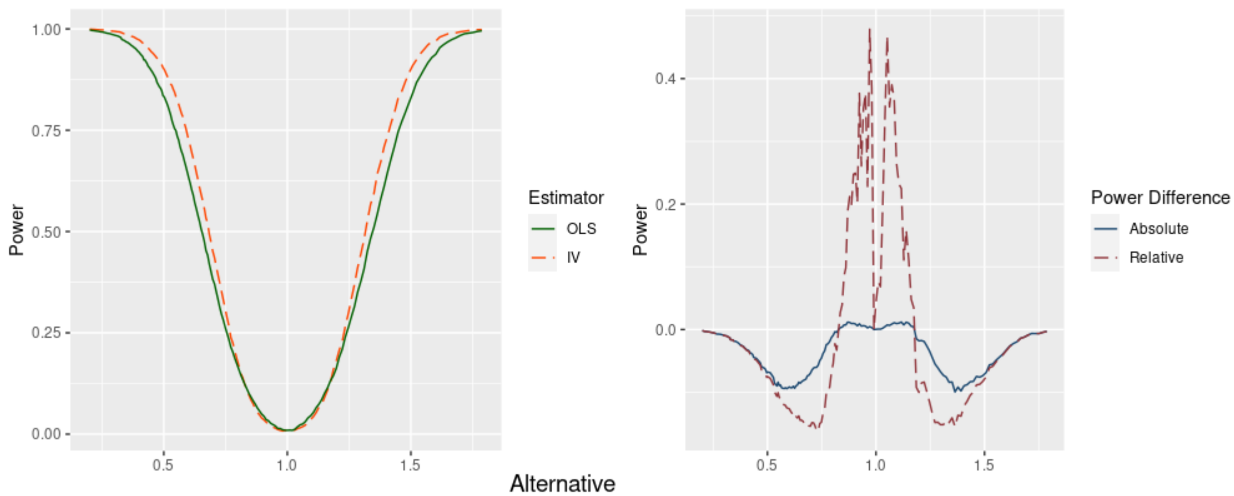

Figure 1: Comparison of OLS and IV estimators for heavy-tailed data

In the left panel, we see that Monte Carlo estimates of the power curves for the test based on R_n (the solid line based on the OLS estimator) and Q_n (the dashed line based on the IV estimator from Shephard (2020)). We highlight here that there is potentially a tradeoff between the two procedures over regions in the alternative space where high-power is desired. In this context, the asymptotic variance of $\hat{\beta}_{OLS}$ is 1.2759, and thee asymptotic variance of $\hat{\beta}_{IV}$, the estimator proposed in Shephard (2020), is 1.4973. This implies that under local power comparisons, R_n is preferred to Q_n. Local to the null $\beta_{1,0} = 1$, we see in the right panel of Figure 1 that R_n outperforms Q_n, as the solid line and dashed lined are both above 0 in a region local to 1. This is consistent with traditional comparisons based on local asymptotics, but notice that the absolute different is fairly small. Farther from the null, Q_n performs better, which is where we expect our theory to provide better predictions.

5 Conclusion

This paper provides new results on the efficiency of t-tests when the data have heavy tails. These extensions highlight that using classical efficiency comparisons for heavy tailed data might not provide accurate information about the performance of test statistics under fixed alternatives. Our results complement recent work studying the performance of test statistics under the null hypothesis when the data exhibit heavy tails.

It would be desirable to extend the main results to cases in which the observations have one thin tail and one heavy tail, and see whether the convergence rate equivalence of the z-test and t-test still holds in that scenario. Other useful extensions include extending this results to classical Wald statistics and more generally to GMM-type statistics. It would also be interested to consider how to use Theorem 1 to conduct power analysis; currently, the limiting type-II error rate diverges as the alternative approaches the null, implying that higher order terms and knowledge of the function L might be useful in practice.

References

BAHADUR, R. R. (1960): "Stochastic Comparison of Tests," *The Annals of Mathematical Statistics*, 31, 276–295.

CLINE, D. B. AND T. HSING (1989): "Large Deviation Probabilities for Sums of Random Variables with Heavy or Subexponential Tails," .

EFRON, B. (1969): "Student's t-Test under Symmetry Conditions," *Journal of the American Statistical Association*, 64, 1278–1302.

GINÉ, E., F. GÖTZE, AND D. M. MASON (1997): "When Is the Student t -Statistic Asymptotically Standard Normal?" *The Annals of Probability*, 25, 1514–1531.

HE, X. AND Q.-M. SHAO (1996): "Bahadur Efficiency and Robustness of Studentized Score Tests," *Annals of the Institute of Statistical Mathematics*, 48, 295–314.

HODGES, J. AND E. LEHMANN (1956a): "The Efficiency of Some Nonparametric Competitors of the T-Test," *Annals of Mathematical Statistics*, 27, 324–335.

——— (1956b): "The Efficiency of Some Nonparametric Competitors of the T-Test," *Annals of Mathematical Statistics*, 27, 324–335.

MIKOSCH, T. AND A. V. NAGAEV (1998): "Large Deviations of Heavy-Tailed Sums with Applications in Insurance," *Extremes. Statistical Theory and Applications in Science, Engineering and Economics*, 1, 81–110.

MÜLLER, U. K. (2019): "Refining the Central Limit Theorem Approximation via Extreme Value Theory," *Statistics & Probability Letters*, 155, 108564.

——— (2020): "A More Robust T-Test," *arXiv*.

SHAO, Q.-M. (1999): "A Cramér Type Large Deviation Result for Student's t-Statistic," *Journal of Theoretical Probability*, 12, 385–398.

SHEPHARD, N. (2020): "An Estimator for Predictive Regression : Reliable Inference for Financial Economics," .

Student (1908): "The Probable Error of a Mean," *Biometrika*, 1–25.

Young, A. (2021): "Leverage, Heteroskedasticity and Instrumental Variables in Practical Application".","," *Manuscript, London School of Economics*.