Journal of Econometrics and Statistics
Vol. 2, Issue 2, 2022, pp. 219-232
https://doi.org/10.47509/JES.2022.v02i02.05

How Large Should the Sample Size Be?

Uwe Hassler*

Goethe University Frankfurt†

Article History
Received : 19 October 2022; Revised : 12 November 2022; Accepted : 19 November 2022; Published : 15 December 2022

Abstract

In many cases, parameter estimation results in limiting normality. This allows for approximate confidence intervals and significance tests. Often, one wishes to bound the length of confidence intervals and to guarantee a certain power for tests. These issues depend on the sample size: How large does it have to be? We provide simple formulae answering this question. Numerical examples show that reasonably reliable inference requires in some cases rather large samples.

Keywords: Limiting normality; confidence intervals; effect size; power of tests.

*I am grateful to Kathrin Medert-Wagner, Marc-Oliver Pohle, Paulo Rodrigues, Nazarii Salish, Jan-Lukas Wermuth and Tanja Zahn for helpful comments. Moreover, I thank an anonymous referee for corrections and comments.

†Statistics and Econometric Methods, RuW Building, 60629 Frankfurt, Germany. Email: hassler@wiwi.uni-frankfurt.de

To cite this paper
Uwe Hassler (2022). How Large Should the Sample Size Be?. *Journal of Econometrics and Statistics.* 2(2), 219-232.

1 Introduction

Clinical trials and statistical experiments may be very costly in terms of money and time. This is one reason to keep the sample sizes n small or moderate. However, precision of estimators and power of significance tests crucially hinge on the sample size. Hence, it is helpful to have an answer to the question: "How large should n be when testing at significance level 5 % to ensure a power of 80%?" Of course you may wish to replace the numbers 5 and 80 by different figures. In fact, the answer is not difficult to come up with; the problem rather is that the question is rarely asked.

The starting point of this short note is limiting normality of some appropriately scaled estimator. This allows to, first, compute approximate confidence intervals (CI) and determine how large n has to be such that the length does not exceed a given number. Second, it allows to carry out approximate significance tests, and an evaluation of the approximate power function answers the above question in inverted commas. The issue of CIs is addressed in the next section. Section 3 is dedicated to the power analysis of significance tests. Three worked numerical examples are contained in Section 4. They show that our very simple formulae may give surprising answers to the question raised in the title of our note. The results are summarized in the final section to provide an empirical guideline.

A word on notation before we begin. Let $\xrightarrow{\mathcal{D}}$ stand for convergence in distribution as the sample size n goes off to infinity, and $\xrightarrow{p}$ signifies convergence

in probability. The cumulative distribution function of a random variable following a standard normal distribution, $\mathcal{N}(0,1)$, is denoted by $\Phi(\cdot)$, and z_p stands for the p quantile thereof. The ceiling function $\lceil x \rceil$ returns the smallest integer larger than or equal to some real x. For a bijective function $h(\cdot)$, the inverse function is written as $h^{-1}(\cdot)$.

2 Framework and Assumptions

We consider inference about some parameter θ resting on limiting normality. The parameter is estimated by $\widehat{\theta}_n$ from a sample of size n. We assume a limit theorem to hold in that

$$\sqrt{m}\left(\widehat{\theta}_n - \theta\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0,\sigma^2) \text{ as } n \to \infty\,, \tag{1}$$

where m is the rate of convergence with

$$m = h(n) \text{ and } \frac{1}{m} \to 0\,. \tag{2}$$

Note that $m = h(n) \to \infty$ implies a population of infinite size from that the sample is drawn. Typically, the constant σ^2 is unknown and has to be estimated consistently: $\widehat{\sigma}_n^2 \xrightarrow{p} \sigma^2$ as $n \to \infty$. In practice, (1) is employed to construct (approximate) CIs or to perform significance tests.

REMARK 1: For the rest of the paper we assume a known bijection $h(\cdot)$. Note that n has to be a natural number. For given m we hence pick $n = \lceil h^{-1}(m) \rceil$.

A centered confidence interval at level $1-\alpha$ is readily available from (1):

$$CI_{1-\alpha} = \left[\widehat{\theta}_n \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{m}} \right] .$$

For a given confidence level $1-\alpha$ one would like the length not to exceed a given number $\ell > 0$. This obviously depends on the variance, too. For a length to be bounded by ℓ it is required that

$$m \geq M(\ell; \alpha, \sigma) := 4\, z^2_{1-\frac{\alpha}{2}} \frac{\sigma^2}{\ell^2} . \tag{3}$$

REMARK 2: If σ^2 is unknown, we need a preliminary variance estimation to use (3). To that end, assume a preliminary, reasonably large sample of size n^* yielding $\widehat{\sigma}^2_{n^*}$ with $\widehat{\sigma}^2_{n^*} \xrightarrow{p} \sigma^2$ as $n^* \to \infty$ to be plugged in. Hence, a required minimum sample size can be determined according to (3) and Remark 1:

$$n(\ell; \alpha, \widehat{\sigma}_{n^*}) = \lceil h^{-1}(M(\ell; \alpha, \widehat{\sigma}_{n^*})) \rceil , \quad M(\ell; \alpha, \widehat{\sigma}_{n^*}) := 4\, z^2_{1-\frac{\alpha}{2}} \frac{\widehat{\sigma}^2_{n^*}}{\ell^2} .$$

Confidence intervals and significance testing are related issues. Let θ_0 denote a prespecified value under the null hypothesis, and d measures the effect size, the strength of violation under the alternative H_1:

$$H_0 : \ \theta = \theta_0 \ \text{ vs. } \ H_1 : \ \theta = \theta_0 + d \, , \ d \neq 0 \, . \tag{4}$$

In case of a two-tailed test, H_0 is rejected at significance level α if and only if $CI_{1-\alpha}$ does not cover θ_0. This is equivalent to

$$Z_0 < -z_{1-\frac{\alpha}{2}} \quad \text{or} \quad Z_0 > z_{1-\frac{\alpha}{2}}\,, \quad Z_0 := \sqrt{m}\,\frac{\widehat{\theta}_n - \theta_0}{\widehat{\sigma}_n}\,, \tag{5}$$

where we assume again that the variance is estimated consistently by $\widehat{\sigma}_n^2$.

3 Approximate Power

How large does n have to be to guarantee a power of $1-\beta$ when testing at significance level α if H_1 holds true? Here, β is the probability of a type II error, i. e. the probability not to reject the null hypothesis when the alternative holds true. This question has been addressed (for special cases) in classical textbooks such as Snedecor and Cochran (1967, Sect. 4.13) and Fleiss (1981, Ch. 2), see also Lehr (1992). The answer will depend on the relative effect size under H_1:

$$\delta := \frac{d}{\sigma}. \tag{6}$$

Let $\Pi_m(\delta;\alpha)$ be the (approximate) power function depending on δ: $\Pi_m(\delta;\alpha) := \mathrm{P}(|Z_0| > z_{1-\frac{\alpha}{2}})$. Because of (1),

$$\Pi_m(\delta;\alpha) = 1 - \Phi\left(z_{1-\frac{\alpha}{2}} - \sqrt{m}\,\delta\right) + \Phi\left(-z_{1-\frac{\alpha}{2}} - \sqrt{m}\,\delta\right)\,. \tag{7}$$

Since (1) holds only as $n \to \infty$, it would be more precise to replace "=" by

"≈" in (7); we neglect this for convenience. By definition, $\Pi_m(\delta;\alpha) = 1-\beta$ or

$$\beta \;=\; \Phi\left(z_{1-\frac{\alpha}{2}} - \sqrt{m}\,\delta\right) - \Phi\left(-z_{1-\frac{\alpha}{2}} - \sqrt{m}\,\delta\right) .$$

For small α with $\Phi\left(-z_{1-\frac{\alpha}{2}}\right)$ being small, one may approximate β. First, if $\delta > 0$, $\Phi\left(-z_{1-\frac{\alpha}{2}} - \sqrt{m}\,\delta\right)$ is even smaller than $\Phi\left(-z_{1-\frac{\alpha}{2}}\right)$. In that sense, $\beta \approx \Phi\left(z_{1-\frac{\alpha}{2}} - \sqrt{m}\,\delta\right)$. Second and similarly, for small α with $\Phi\left(z_{1-\frac{\alpha}{2}}\right)$ being large, one may approximate if $\delta < 0$: $\Phi\left(z_{1-\frac{\alpha}{2}} - \sqrt{m}\,\delta\right)$ is even larger than $\Phi\left(z_{1-\frac{\alpha}{2}}\right)$, such that $\Phi\left(z_{1-\frac{\alpha}{2}} - \sqrt{m}\,\delta\right) \approx 1$. Due to symmetry it follows that

$$\beta \approx 1 - \Phi\left(-z_{1-\frac{\alpha}{2}} - \sqrt{m}\,\delta\right) = \Phi\left(z_{1-\frac{\alpha}{2}} + \sqrt{m}\,\delta\right) = \Phi\left(z_{1-\frac{\alpha}{2}} - \sqrt{m}\,|\delta|\right) .$$

Both cases together yield $\beta \approx \Phi\left(z_{1-\frac{\alpha}{2}} - \sqrt{m}\,|\delta|\right)$, where the approximation is all the better the smaller α is. Consequently, $z_\beta \approx z_{1-\frac{\alpha}{2}} - \sqrt{m}\,|\delta|$. Neglecting the approximation error yields

$$m(\alpha,\beta;\delta) := \frac{\left(z_{1-\frac{\alpha}{2}} - z_\beta\right)^2}{\delta^2} , \tag{8}$$

such that $m \geq m(\alpha,\beta;\delta)$ ensures (approximately) a minimum power of $1-\beta$. Again, the corresponding sample size $n(\alpha,\beta;\delta)$ is determined according to Remark 1.

Remark 3: In practice, of course neither d from (4) under H_1 nor σ from (1) are known. The difference d can be replaced by $\widehat{\theta}_{n^*} - \theta_0$ for some prelim-

inary sample of size n^*, and a preliminary variance estimator $\widehat{\sigma}^2_{n^*}$ is available according to Remark 2, too. Hence, a required minimum rate m can be determined as $m(\alpha, \beta; \widehat{\delta}_{n^*})$ with $\widehat{\delta}_{n^*} := (\widehat{\theta}_{n^*} - \theta_0)/\widehat{\sigma}_{n^*}$. As in Remark 1, one may determine the required sample size $n(\alpha, \beta; \widehat{\delta}_{n^*})$ to ensure a power of $1 - \beta$.

REMARK 4: Let us briefly consider one-tailed tests at significance level α. To ensure a rejection probability of $1 - \beta$ for one-tailed tests at level α, (8) has to be replaced by

$$m(\alpha, \beta; \delta) = \frac{\left(z_{1-\alpha} - z_\beta\right)^2}{\delta^2} \text{ or } \sqrt{m(\alpha, \beta; \delta)} = \frac{z_{1-\alpha} - z_\beta}{|\delta|}\,, \tag{9}$$

where $\beta < 1 - \alpha$ is assumed. Note that the same symbol $m(\alpha, \beta; \delta)$ is used for two-tailed and one-tailed tests.

4 Three Numerical Examples

EQUALITY OF MEANS: The first example embeds a result found in Lehr (1992) when testing for equality of means from paired samples. Consider bivariate random samples, (Y_i, Z_i), $i = 1, \ldots, n$, with variances σ^2_Y and σ^2_Z and correlation ρ. We wish to test for equality of means, μ_Y and μ_Z:

$$H_0 : \ \mu_Y - \mu_Z = 0 \quad \text{vs.} \quad H_1 : \ \mu_Y - \mu_Z = d\,.$$

To match the above framework, we define the differences $X_i = Y_i - Z_i$ with

$\theta = \mu_Y - \mu_Z$, $\widehat{\theta}_n = \overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and call on the following central limit theorem:

$$\sqrt{n}\,\left(\overline{X} - (\mu_Y - \mu_Z)\right) \;\stackrel{\mathcal{D}}{\to}\; \mathcal{N}(0, \sigma^2)\,,\;\; \sigma^2 := \sigma_Y^2 + \sigma_Z^2 - 2\,\rho\,\sigma_Y\sigma_Z\,.$$

For $\rho = 0$ and $\sigma_Y^2 = \sigma_Z^2$, $m(\alpha, \beta; \delta)$ from (8) specializes to the formula in Lehr (1992, p. 1101). In the general case, consistent variance estimation is given by $\widehat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2$.

Let us assume some preliminary n^* as in Remark 3 resulting in a standard deviation $\widehat{\sigma}_{n^*}$ estimating σ. Consider the effect d as a portion of the standard deviation with $\widehat{\delta}_{n^*} = d/\widehat{\sigma}_{n^*}$ or $d = \widehat{\delta}_{n^*}\widehat{\sigma}_{n^*}$. How large does n have to be to detect $d = \widehat{\delta}_{n^*}\widehat{\sigma}_{n^*}$ with a power of 80 % when testing at a 5 % level (two-tailed)? According to (8) the answer is given by

$$n(0.05, 0.2; \widehat{\delta}_{n^*}) = \left\lceil \frac{(z_{0.975} - z_{0.2})^2}{\widehat{\delta}_{n^*}^2} \right\rceil .$$

From the following table we learn that detecting (with power of 80 %) a difference d half as large as the standard deviation requires only $n = 32$, while a detection of $d = 0.1 \cdot \widehat{\sigma}_{n^*}$ requires $n = 785$.

$\widehat{\delta}_{n^*}$	0.5	0.4	0.3	0.2	0.1
$n(0.05, 0.2; \widehat{\delta}_{n^*})$	32	50	88	197	785

Tail index: Second, consider the tail index θ indicating the highest finite

moment of a random variable. Nicolau and Rodrigues (2019, Sect. III) proposed a new estimator $\widehat{\theta}_n$. Under the assumption of independent identically distributed (iid) random variates following a Pareto distribution it holds that (Nicolau and Rodrigues (2019, Thm. 1))

$$\sqrt{n}\,\left(\widehat{\theta}_n - \theta\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)\,,\ \sigma^2 = 2\,\theta^2. \tag{10}$$

If the distributional assumptions are weakened and the random sequence does not obey an exact Pareto law but only a Pareto-type tail behaviour, see Nicolau and Rodrigues (2019, eq. (8)), then the convergence becomes slower (still under iid):

$$\sqrt{m}\,\left(\widehat{\theta}_n - \theta\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2) \quad \text{with } m = n^{\gamma}\,,\ \sigma^2 = 2\,\theta^2, \tag{11}$$

where $\gamma < 1$.

For $\theta \leq 2$, finite second moments do not exist. Hence, a null hypothesis of interest is $\theta_0 = 2$. Nicolau and Rodrigues (2019, Sect. V) analyse daily absolute returns from January 1, 1999, to May 16, 2016, which amounts to $n^* = 4,532$. For the Hong Kong dollar, they find that $\widehat{\theta}_{n^*} = 2.27$. Hence by Remark 3 and with $\widehat{\sigma}^2_{n^*} = 2\,\widehat{\theta}^2_{n^*}$, $\widehat{\delta}_{n^*} = (2.27 - 2)/(2.27\,\sqrt{2})$. For a one-tailed test at 5 % with a power of $1 - \beta = 0.8$, Remark 4 provides for (10) with (9):

$$m(0.05, 0.20; \widehat{\delta}_{n^*}) = 874.02 \quad \text{or} \quad n(0.05, 0.20; \widehat{\delta}_{n^*}) = 875\,.$$

To apply (11), Nicolau and Rodrigues (2019, Fig. 5) choose m such that $m/n = n^{\gamma-1} = 0.10$. With $n^* = 4,532$ this implies $\gamma = 0.7265$ or according to Remark 1 $n = \lceil m^{1.3765} \rceil$. Without assuming exact Pareto, an 80 % power one-tailed test according to (11) requires $n(0.05, 0.20; \widehat{\delta}_{n^*}) = 11,195$ for $\alpha = 0.05$. The example shows that reliable significance tests for tail indexes require rather large sample sizes.

Note that these computations are not directly applicable to the empirical analysis by Nicolau and Rodrigues (2019), because the simple variance expression from (10) or (11) holds only under the restrictive iid assumption. This assumption is not met by daily absolute returns. For more general stationary processes Nicolau and Rodrigues (2019, Thm. 3) establish limiting normality with a more complicated variance expression that can be handled by means of robust standard errors.

Long memory: The third example is from the realm of time series analysis. Let θ now denote the parameter of fractional integration. A fractionally integrated process or time series is stationary and has so-called short memory if $\theta \leq 0$; it is stationary and displays long memory if $0 < \theta < 0.5$; and it becomes nonstationary for $\theta \geq 0.5$. A popular estimator is the so-called local Whittle estimator explored by Robinson (1995) and Velasco (1999). It is settled in the frequency domain and relies on computation at m harmonic frequencies. It holds asymptotically for $-0.5 < \theta < 0.75$ under some technical

assumptions that

$$\sqrt{m}\left(\widehat{\theta}_n - \theta\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{1}{4}\right), \quad m = n^{0.65}. \tag{12}$$

For modified and extended versions of local Whittle and for a larger range of θ, (12) continues to hold according to Abadir, Distaso, and Giraitis (2007) and Shimotsu and Phillips (2005). In practice, m has to be chosen to balance a trade-off between finite sample bias and variance of $\widehat{\theta}_n$. The rate m is often determined as $m = n^{0.65}$ following recommendations by Abadir et al. (2007, p. 1363) and Shimotsu (2010, p. 515). One nice feature is that $\sigma = \frac{1}{2}$ in (12) does not vary with θ. This allows to use (3) without Remark 2 when setting up a CI. Since a difference in θ of absolute value 0.1 makes a sizeable difference in terms of (long) memory we want the maximum length of CI to be bounded by $\ell = 0.1$. At a 95 % confidence level, (3) provides

$$M(0.1; 0.05, 0.5) = 4 \cdot 1.96^2 \frac{0.25}{0.01} = 384.16\,.$$

Since $m = n^{0.65}$, $n(0.1; 0.05, 0.5) = \lceil M(0.1; 0.05, 0.5)^{1.5385} \rceil = 9,469$: A CI at 95 % confidence level of maximum length 0.1 requires almost 9,500 observations. This shows that, due to the slow convergence rate $m = n^{0.65}$, reliable inference about long memory requires rather large sample sizes. This will equally hold true when it comes to hypothesis testing.

Hypotheses of interest are short memory against long memory,

$$H_0 : \ \theta = 0 \ \ \text{vs.} \ \ H_1 : \ \theta = d > 0 \,,$$

or nonstationarity against stationarity:

$$H_0 : \ \theta = 0.5 \ \ \text{vs.} \ \ H_1 : \ \theta = 0.5 + d \,, \ d < 0 \,.$$

The relative effect size becomes $\delta = d/\sigma = 2\,d$. Let us assume $d = 0.10$ for the first pair of hypotheses and $d = -0.10$ for the second pair, which is from a substantive point of view considerably distant from both null hypotheses. Remark 4 gives for the one-tailed tests at 5 % level with a power of 80 % ($\beta = 0.2$) that

$$m(0.05, 0.20; 0.20) = \frac{\left(z_{0.95} - z_{0.2}\right)^2}{0.2^2} = 154.56 \,,$$

resulting in $n(0.05, 0.20; 0.20) = \lceil 154.56^{1.5385} \rceil = 2,334$. This reinforces that significant inference about long memory requires large samples.

5 Summary

The starting point is limiting normality of some parameter estimator as in (1). Centered confidence intervals at level $1 - \alpha$ follow immediately. To bound their length by some maximum length one may choose the sample size

according to (3) in connection with Remark 2. To guarantee an approximate minimum power of $1 - \beta$ when performing a two-tailed test at significance level α, one may rely on (8) together with Remark 3. For one-tailed tests one should employ (9) instead.

References

Abadir, K. M., W. Distaso, and L. Giraitis (2007). Nonstationarity-extended local Whittle estimation. *Journal of Econometrics 141*, 1353–1384.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions* (3rd ed.). Wiley.

Lehr, R. (1992). Sixteen S-squared over D-squared: A relation for crude sample size estimates. *Statistics in Medicine 11*, 1099 – 1102.

Nicolau, J. and P. M. M. Rodrigues (2019). A new regression-based tail index estimator. *The Review of Economics and Statistics 101*(4), 667 – 680.

Robinson, P. M. (1995). Gaussian semiparametric estimation of long range dependence. *Annals of Statistics 23*, 1630–1661.

Shimotsu, K. (2010). Exact local Whittle estimation of fractional integration with unknown mean and trend. *Econometric Theory 26*, 501–540.

Shimotsu, K. and P. C. B. Phillips (2005). Exact local Whittle estimation of fractional integration. *The Annals of Statistics 33*, 1890–1933.

Snedecor, G. W. and W. G. Cochran (1967). *Statistical Methods* (6th ed.). Iowa State University Press.

Velasco, C. (1999). Gaussian semiparametric estimation of non-stationary time series. *Journal of Time Series Analysis 20*, 87–127.