# At what sample size do correlations stabilize?

Felix D. Schönbrodt[a], Ludwig-Maximilians-Universität München, Germany

Marco Perugini[b], University of Milan - Bicocca, Italy

**Paper in press at the Journal of Research in Personality (June 2013)**

**doi:10.1016/j.jrp.2013.05.009**

## Author Note

[a]Department of Psychology, Ludwig-Maximilians-Universität, Leopoldstr. 13, 80802 München, Germany. Phone: +49 89 2180 5217. Fax: +49 89 2180 99 5214. Email: felix@nicebread.de. Correspondence concerning this article should be addressed to Felix Schönbrodt.

[b]Marco Perugini, Department of Psychology, University of Milan - Bicocca, Piazza dell'Ateneo Nuovo 1 (U6), 20126 Milan, Italy. Email: marco.perugini@unimib.it.

Acknowledgements: Ideas for this simulation study partly originated on a summer school on robust statistics in 2011 (supported by EAPP and ISSID), and during discussions with Jens Asendorpf.

## *Abstract*

Sample correlations converge to the population value with increasing sample size, but the estimates are often inaccurate in small samples. In this report we use Monte-Carlo simulations to determine the critical sample size from which on the magnitude of a correlation can be expected to be stable. The necessary sample size to achieve stable estimates for correlations depends on the effect size, the width of the corridor of stability (i.e., a corridor around the true value where deviations are tolerated), and the requested confidence that the trajectory does not leave this corridor any more. Results indicate that in typical scenarios the sample size should approach 250 for stable estimates.

*Keywords:* correlation, accuracy, sample size, simulation

## *Highlights:*

- Sample correlations converge to true value *ρ*, but are inaccurate in small samples
- From which sample size on do correlations only show minor fluctuations around *ρ*?
- Monte-Carlo simulations were used to determine the "point of stability" (*POS*)
- Necessary sample size depends on effect size, tolerable fluctuations, and confidence
- In typical scenarios *n* should approach 250 for stable estimates

At what sample size do correlations stabilize?

## 1. Introduction

Most research in psychology seems to be concerned with the endeavor to determine the sign of an effect with some confidence, using the null hypothesis significance testing (NHST) procedure. Several authors, however, have argued that any field of science should move from binary decisions derived from the NHST procedure towards giving a more precise point estimate of the magnitude of an effect (Edwards & Berry, 2010; Kelley & Maxwell, 2003). Consider, for example, a correlation of $r = .40$ in a sample of 25 participants. This correlation is significantly different from zero ($p = .047$). Hence, it might be concluded with some confidence that there is "something > 0" in the population, and the study would be counted as a success from the NHST perspective. However, plausible values of the true correlation $\rho$, as expressed by a 90% confidence interval, range from .07 to .65. The estimate is quite unsatisfactory from an accuracy point of view – in any scenario beyond the NHST ritual it will make a huge difference whether the true correlation in the population is .07, which would be regarded as a very small effect in most research contexts, or .65, which would be a very large effect in many contexts. Moreover, precise point estimates are relevant for a priori sample size calculations. Given the huge uncertainty in the true magnitude of the effect, it is hard to determine the necessary sample size to replicate the effect (e.g., for an intended power of 80% and $\rho = .07$: $n = 1599$; $\rho = .40$: $n = 46$; $\rho = .65$: $n = 16$).

In this contribution we deal with a related question of practical importance in personality research: At which sample size does a correlation stabilize? Many researchers might have observed that the magnitude of a correlation is pretty unstable in small samples, as the following empirical example demonstrates. Multiple questionnaire scales have been administered in an open online study (Schönbrodt & Gerstenberg, 2012; Study 3). The thick black line in Figure 1 shows the evolution of the correlation between two scales, namely "hope of power" and "fear of losing control" when after each new participant the correlation is recalculated. It can be seen that the correlation evolved from $r = .69$ ($n = 20$, $p < .001$) to $r = .26$ ($n = 274$, $p < .001$). From a visual inspection, the trajectory did not stabilize up to a sample size of around 150. Data have not been rearranged – it is simply the order how participants dropped into the study. Some other correlations in this data set evolved from significantly negative to non-significant, others changed from one significant direction into the significant opposite, and some correlations were stable right from the beginning with only few fluctuations around the final estimate. But how do we get to know when a correlation estimate is sufficiently stable?

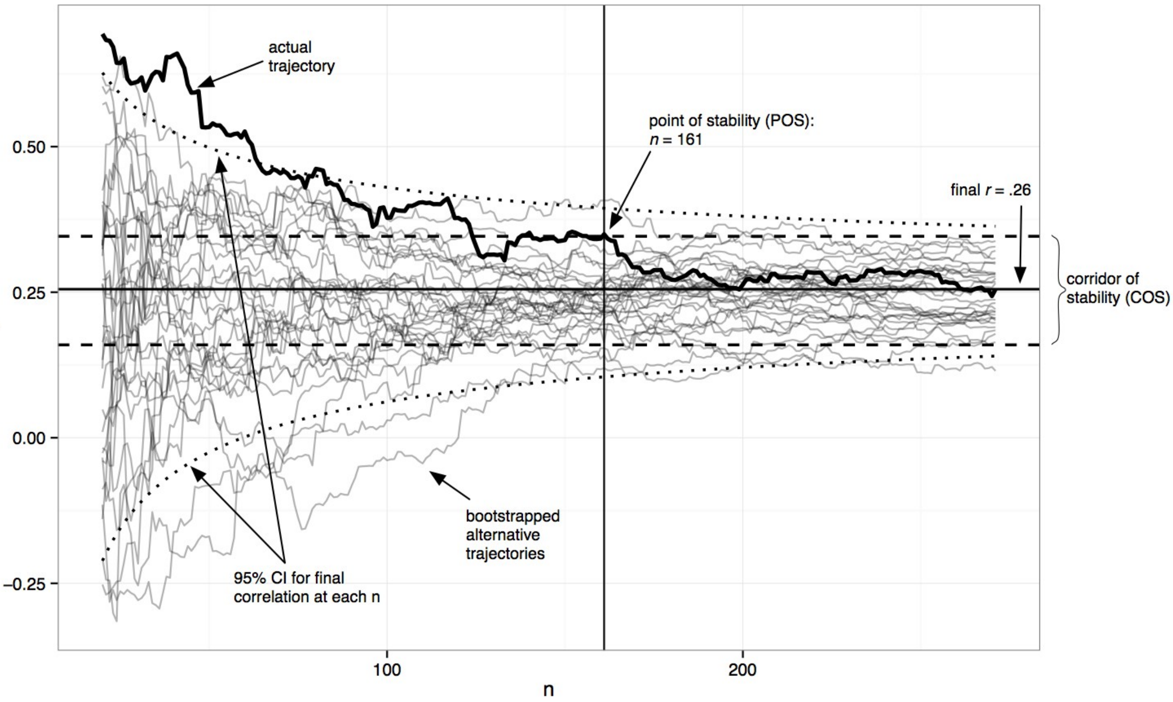

*Figure 1:* Actual (thick black line) and bootstrapped (thin gray lines) trajectories of a correlation. The dotted curved lines show the 95% confidence interval for the final correlation of $r = .26$ at each $n$. Dashed lines show the ± .1 corridor of stability (*COS*) around the final correlation. The point of stability (*POS*) is at $n = 161$. After that sample size the actual trajectory does not leave the *COS*.

## 2. Definition and Operationalization of Stability

Suppose a true correlation of $\rho$ = .40. When the estimate in the sample is .41 or .38, most researchers would agree that this is a rather trivial deviation from the true value. A stronger deviation like .26 or .57 could be deemed more problematic, depending on the research setting. And even stronger deviations like .10 or .65 (which would still be within the 95% CI at $n$ = 25) probably would be judged unacceptable from a substantial point of view.

When talking about stability, minor fluctuations around the true value can be tolerated, but not large deviations. Hence, a *corridor of stability* (*COS*) around the true value can be defined, where all deviations within that corridor are classified as being acceptable. As the confidence interval around correlations partly depends on the magnitude of the correlation, the corridor is defined in units of $q$, an effect size measure for correlations that only depends on sample size (Cohen, 1988). For that purpose, $\rho$ is Fisher-$r$-to-$Z$-transformed and the desired width of the corridor, $w$, is both subtracted from and added to that value. Therefore, $w$ denotes the half-width of the *COS*. These upper and lower boundaries then are back-transformed to a correlation metric. The desired width of the corridor depends on the specific research context (see Figure 1 for a *COS* with $w$ = .10). In this paper, three widths are used: ± .10, ± .15, and ± .20. Following the rules of thumb proposed by Cohen (1992), a value of .10 for $w$ corresponds to a small effect size. Hence, if the sample correlation $r$ stays within a corridor with $w$ = ± .10, the resulting deviations only have a small effect size. The half-width of a corridor is denoted in the subscript of *COS*. As an example, for $\rho$ = .40 the $COS_{.10}$ is [.313; . 481], $COS_{.15}$ *is* [.267; .518], and $COS_{.20}$ is [.220; .554].

With increasing sample size the sample correlation approaches the true value with a continuously decreasing confidence interval (see Figure 1). The *point of stability* (*POS*) is defined as that sample size from which value on the trajectory of the correlation does not leave the *COS* any more. For sample sizes with $n$ > *POS*, the estimate only shows tolerable fluctuations around the true value.

To assess the variability of possible trajectories, bootstrap samples of the final sample size can be drawn from the original raw data, and the evolutions of correlation for the new data sets are calculated. Figure 1 shows some exemplary bootstrapped trajectories. It can be seen that some trajectories start well above the final value (as the original trajectory), some start even with a significant negative value, and some start already within the *COS* without ever leaving it. With increasing sample size, all trajectories converge into the *COS*. Each of these trajectories has its own *POS*, and if several thousand bootstrap trajectories are sampled *a distribution of POS* can be obtained. Computing percentiles of this distribution allows pinpointing the critical *POS*, $POS_{crit,}$ from which value on at least, for example, 80% of all trajectories do not leave the *COS* any more. Henceforward, these percentiles are called the *confidence* in the stability.

With these definitions, the answer to the research question can be formulated more precisely: We are interested in the critical sample size $POS_{crit}$, from which value on the estimate of a correlation does not leave the $COS_w$ with a confidence of 80% (90%, 95%).

## 3. Method and Results

To compute a distribution of *POS* values, Monte-Carlo simulations have been run in the *R* environment for statistical computing (R Development Core Team, 2012). The complete source code for the computations can be downloaded from the online supplementary material. The following steps have been performed for the simulation:

- Simulate a bivariate Gaussian distribution with 1'000'000 cases and a specified correlation $\rho$ (the "population").

- Draw $B$ = 100'000 bootstrap samples[1] with $n_{max}$ = 1000 cases from this population
- For each bootstrap sample calculate the correlation for every sample size $n$, starting from $n_{min}$ = 20 up to $n_{max}$ = 1000, with a step size of 1 ("trajectory of the correlation").
- Calculate the *POS* for each bootstrapped trajectory. For that purpose trace back the trajectory from $n_{max}$ (99.98 % of all trajectories terminated within the *COS* at $n$ = 1000) until it breaks the *COS* for the first time. The sample size of this break is recorded as the *POS* for this trajectory.

Seven different correlations were used for the populations ($\rho$s = .1, .2, .3, .4, .5, .6, and . 7), and three different half-widths of the *COS* ($w$ = .10, .15, and .20). Correlations were imposed using a method proposed by Ruscio and Kaczetow (2008). As a result of this procedure, a sample of 100'000 *POS* values was obtained for each $\rho$ and each *COS* width. Subsequently, three percentiles of these *POS* distributions were calculated: 80%, 90%, and 95%.

Table 1 shows $POS_{crit}$ for each experimental condition. Not surprisingly, all other things being equal, the required sample size increases with smaller $w$ and larger confidence. Furthermore, larger correlations stabilize earlier.

[1] 100'000 bootstrap samples might seem an unusually high number. We ran so many replications because the estimation of extreme quantiles is less accurate than that of quantiles near the median, in particular for long-tailed distributions (Wilcox, 2005). Hence, a large number of bootstrap replications is needed to obtain accurate estimates for the 95% confidence condition and low $\rho$s. For example, in the current simulation the width of the 95% CI of the 95th quantile ($w$ = .1, $\rho$ = .1) is 99 for $B$ = 1000, 27 for $B$ = 10'000, and 10 for $B$ = 100'000.

| | Critical point of stability (*POS*$_{crit}$): Level of confidence | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 80% | | | 90% | | | 95% | | |
| ρ | *w* = .10 | *w* = .15 | *w* = .20 | *w* = .10 | *w* = .15 | *w* = .20 | *w* = .10 | *w* = .15 | *w* = .20 |
| .1 | 252 | 110 | 61 | 362 | 158 | 88 | 470 | 209 | 116 |
| .2 | 238 | 104 | 57 | 341 | 150 | 83 | 446 | 197 | 109 |
| .3 | 212 | 93 | 51 | 304 | 134 | 75 | 403 | 177 | 99 |
| .4 | 181 | 78 | 43 | 260 | 114 | 63 | 342 | 152 | 84 |
| .5 | 143 | 62 | 34 | 208 | 90 | 50 | 275 | 121 | 68 |
| .6 | 104 | 45 | 25 | 150 | 66 | 37 | 202 | 89 | 51 |
| .7 | 65 | 28 | 20 | 96 | 42 | 24 | 129 | 58 | 35 |

Table 1: *The critical points of stability (*POS$_{crit}$*) for different widths (*w*) of the corridor of stability (*COS*), different levels of confidence, and different ρs.*

To explore the robustness of these results, two ancillary analyses were performed. First, the conditions for a break were made harder by requiring that the trajectory has to leave the *COS* for two or three *consecutive* instances to qualify for a break (i.e., if only for a single sample size the trajectory is outside the *COS*, but several sample sizes before and after that point are within, this would not be classified as a break). For two consecutive breaks, the *POS* was reduced on average by 3 cases (maximum: 8 cases), for three consecutive breaks on average by 6 cases (maximum: 13). These reductions are rather negligible in comparison to the overall magnitude of the typical *POS*.

Second, we explored the impact of non-normal distributions with skewness, heavy tails, and outliers. In an analysis of 440 large-scale real world data sets in psychology only 4.3% could be considered as reasonable approximations to a Gaussian normal distribution (Micceri, 1989). Hence, deviations from normality are rather the rule than the exception in psychology.

To explore the impact of these deviations on the $POS_{crit}$ values, we used four real world data sets provided by T. Micceri[2] as marginal distributions and imposed the specified population correlations (Ruscio & Kaczetow, 2008). We constrained our analysis to typical non-normal distributions found in psychology (i.e., some skewness and somewhat heavier tails)[3]. Results for these variables were comparable to the Gaussian simulated data set. In non-normal distributions the *POS* had a median increase of 1.7% compared to the normal case (i.e., on average the correlations stabilized slightly later), and 90% of the differences between the non-normal and normal *POS* were smaller than 6%.

## 4. Discussion

It has been argued that for a cumulative growth of knowledge accurate estimates of the magnitude of an effect would be more fruitful than simple binary decision derived from NHST. Previous approaches concerned with the accuracy of estimates focused on confidence intervals around the point estimates. By defining the aspired level of accuracy one can compute the necessary sample size (Algina & Olejnik, 2003; Maxwell, Kelley, & Rausch, 2008).

The current report extends this literature by applying a sequential sampling perspective, and answers the question: How many participants do we have to sample to be confident that the correlation has *stabilized* within a reasonable corridor? The answer, reported in Table 1, depends on the size of the true correlation, the accuracy that is requested, and the confidence that the researcher wants to have in the decision. Precise and stable estimates within a corridor of +/- .05 need large samples beyond $n$ = 1000, as has been noted before (Hunter & Schmidt,

[2] The data sets can be downloaded from http://www.freewebs.com/tedstats/Files/Real_Data.zip. In our analysis we included all pairwise combinations from four variables (ACT_composite, GRE_quantitative, GRE_verbal, and Cumulative_GPA).

[3] We did not explore the effect of extreme deviations from normality, which could be expected to have more impact on the results under some special conditions.

2004; Maxwell et al., 2008). But this level of precision can only be achieved by a relatively small number of high-budget studies. In practice, most research is done with budget and logistical constraints and therefore a sensible pragmatic question is when we can consider an estimate of a correlation coefficient reasonably stable.

If Table 1 should be boiled down to simple answers, one can ask what effect size typically can be expected in personality. In a meta-meta-analysis summarizing 322 meta-analyses with more than 25'000 published studies in the field of personality and social psychology, Richard, Bond, and Stokes-Zoota (2003) report that the average published effect is $r = .21$, less than 25% of all meta-analytic effects sizes are greater than .30, and only 5.28% of all effects are greater than .50. Hence, without any specific prior knowledge it would be sensible to assume an effect size of .21[4]. Further let's assume that a confidence level of 80% is requested (a level that is typically used for statistical power analyses), and only small effect sizes ($w < .10$) are considered as acceptable fluctuations. By applying these values on Table 1 the required sample size is around $n = 238$. Of course, what is a meaningful or expected correlation can vary depending on the research context and questions. In some research contexts even small correlations of .10 might be meaningful and with consequential implications. In this case, larger samples are needed for stable correlations.  In other research contexts the expected correlation can be greater (e.g., convergent validity between different measures of the same trait) or the researcher is willing to accept a slightly less stable estimate, perhaps compensating with an increased level of confidence. This would reduce the necessary sample size. But even under these conditions there are few occasions in which it may be justifiable to go below $n = 150$ and for typical research scenarios reasonable trade-offs between accuracy and confidence start to be achieved when $n$ approaches 250.

[4] As a reviewer pointed out, meta-analyses tend to overestimate the effect size as only published studies are incorporated into the analysis. Hence, a cautious researcher might even lower the expectation.

# References

Algina, J., & Olejnik, S. (2003). Sample size tables for correlation analysis with applications in partial correlation and multiple regression analysis. *Multivariate Behavioral Research*, *38*, 309–323. doi:10.1207/S15327906MBR3803_02

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New Jersey, US: Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. doi:10.1037/0033-2909.112.1.155

Edwards, J. R., & Berry, J. W. (2010). The presence of something or the absence of nothing: Increasing theoretical precision in management research. *Organizational Research Methods*, *13*, 668–689. doi:10.1177/1094428110380467

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, CA: SAGE.

Kelley, K., & Maxwell, S. E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods*, *8*, 305–321.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537–563. doi:10.1146/annurev.psych.59.103006.093735

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156–166. doi:10.1037/0033-2909.105.1.156

R Development Core Team. (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*, 331–363. doi:10.1037/1089-2680.7.4.331

Ruscio, J., & Kaczetow, W. (2008). Simulating multivariate nonnormal data using an iterative algorithm. *Multivariate Behavioral Research*, *43*, 355–381.

doi:10.1080/00273170802285693

Schönbrodt, F. D., & Gerstenberg, F. X. R. (2012). An IRT analysis of motive questionnaires: The Unified Motive Scales. *Journal of Research in Personality*, *6*, 725–742. doi:10.1016/j.jrp.2012.08.010

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed.). San Diego: Academic Press.