UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

UNIVARIATE SAMPLING BOOTSTRAP PROCEDURES USING PRIOR INFORMATION

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

WILLIAM HOWARD BEASLEY IV
Norman, Oklahoma
2010

UNIVARIATE SAMPLING BOOTSTRAP PROCEDURES
USING PRIOR INFORMATION

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF PSYCHOLOGY

BY

_____
Dr. Joseph Rodgers, Chair

_____
Dr. Jorge Mendoza

_____
Dr. Larry Toothaker

_____
Dr. Robert Terry

_____
Dr. Scott Gronlund

_____
Dr. Kevin Grasse

Table of Contents

List of Tables

List of Figures

Abstract

Analyses that test nonzero correlations and incorporate prior information can help accumulate knowledge and advance research at a faster pace than typical analyses that disregard previous studies and continue to test unreasonable nil hypotheses. The performance of several bootstrap and parametric procedures are evaluated using populations that had varying degrees of correlation and nonnormality. With correlated heteroscedastic variables, the parametric procedures produced robust point estimates, but showed liberal error rates that worsened as sample sizes grew to N_{Obs} = 1,000. This paper proposes two Bayesian univariate sampling bootstrap procedures (the SlotHI and SlotOI) that exhibit much better error rates across all evaluated populations and prior distributions. Based on this simulation, we suggest that the univariate sampling bootstraps are preferred when testing nonzero correlations in nonnormal populations, regardless if prior information is considered.

Title: Univariate Sampling Bootstrap Procedures using Prior Information

Background

The procedures described below build directly upon two previous papers. Lee and Rodgers (1998) developed a univariate sampling bootstrap where each X value is independently combined with each Y to produce a sampling frame of N_{Obs}^2 bivariate points. This approach blends bootstrap characteristics (e.g., scores are resampled with replacement) with permutation test characteristics (e.g., the sampling frame reflects the hypothesis, not the observed sample).

Beasley et al. (2007) described how the sampling frame can be diagonalized to an arbitrary correlation value, which then can be exploited in two ways. In one procedure, the imposed correlation reflects a nonzero hypothesis value; thus the Lee and Rodgers bootstrap was generalized to accommodate a nonzero null hypothesis. In a second procedure, the imposed correlation mimics the observed value; in a sense, the typical bivariate sampling bootstrap is expanded to consider N_{Obs}^2 scores instead of N_{Obs}. The relevant conceptual and procedural details are in the 'Likelihood Distributions' section.

Bayesian statistics can address some important questions that Frequentist statistics cannot. The current paper describes how the univariate sampling bootstrap can accommodate Bayesian analysis. The bootstrap's sampling distribution represents the probability of observing a correlation r_{Obs}, after the statistician has assumed the population has a fixed hypothesized correlation of ρ_{Hyp}. This distribution is also known as a likelihood distribution, $p(r_{Obs} \mid \rho_{Hyp})$, and is the foundation of Frequentist inference.

The new bootstrap procedures partition the range of ρ_{Hyp} into many non-overlapping sets. When this discretized likelihood information is combined with prior information, $p(\rho_{Hyp})$, the product is a Bayesian posterior distribution, $p(\rho_{Hyp} \mid r_{Obs})$. The relevant conceptual and procedural details are in the 'Posterior Distributions' section. The motivations for using a bootstrap for Bayesian inference are described next.

Leveraging Previous Research

Building on previous research can be beneficial for a field, but contemporary psychological research misses at least two opportunities, both of which are related to statistical issues.

Incorporating Prior Information

One opportunity is missed when a statistical analysis neglects to incorporate prior knowledge formally into an experiment's analysis. The consideration of existing research can stabilize the field's collective opinion. Howard, Maxwell & Fleming (2000, p. 316) wrote, "it is rarely the case in psychology that a single study can be viewed as providing a definitive test of a scientific hypothesis. Instead, multiple studies are almost always necessary. However, a serious limitation of the doctrinaire NHST [null hypothesis significance testing] approach is that it does not provide a useful foundation for accumulating evidence over multiple studies."

In contrast to Frequentist NHST, Bayesian analyses can coherently synthesize recent experimental data with subjective expectations and previously observed data. This is beneficial because (a) spuriously strong findings can be dampened and (b) correlations that corroborate previous findings, but are small or moderate, are more likely to be significant because the CIs are narrower.

Testing Non-Nil Hypotheses

Psychologists typically miss a second opportunity to advance existing knowledge. Frequently, a study tries to establish only that there is *some* relationship among variables, and "this goal accounts for most of the explicit formal use of statistics in psychological research" (Krantz, 1999, p. 1376). In the context of *t* tests and ANOVAs, this "nil" hypothesis states that there is no difference between groups; in the context of correlations and regressions, the nil hypothesis is rejected when the CI excludes $\rho = 0$ (Cohen, 1994). We believe that it is important to evaluate this "no effect" hypothesis initially, but believe that the field can further benefit by testing specific values subsequently. As a field matures, the ability to test non-nil hypotheses becomes not only beneficial, but critical or even mandatory.

Suppose a developmental psychologist has observed a correlation of $r = .5$ between vocabulary and intelligence among some subgroup (say, males on welfare). After some evidence is found that the population correlation is most likely positive and doesn't include zero, the analysis usually stops. But the data can be further leveraged, and the knowledge can be sharpened if subsequent experiments address questions such as "does this subgroup have a correlation that is stronger than the general population?" or "is this relationship stronger than the relationship between reading and intelligence?"[1]

We don't believe that nil hypotheses and Frequentist inference should never be used, but rather that they shouldn't always be used exclusively –in many situations, psychology can capitalize by supplementing conventional analysis with Bayesian inference and non-nil hypotheses. This paper develops and evaluates two procedures that

[1] For the sake of illustration, we are assuming there general population correlations are known well enough that sampling error can be ignored.

allow a statistician to (1) incorporate prior information and (2) test specific nonzero correlations, while (3) providing more robust inferences than the conventional parametric procedures.

Robustness

Parametric procedures have been used in Bayesian analysis and to test non-nil values for decades (Jeffreys, 1939; Fisher, 1915). However, these procedures are not always robust when the assumption of normality is violated. Beasley et al. (2007) reported that conventional parametric procedures had acceptable (Frequentist) Type I error rates when nonzero correlations were tested if the variables were normally distributed, but not when the variables were skewed.

Within Bayesian practices, Boos and Monahan (1986) remarked that while the prior distribution receives a lot of discussion and scrutiny, the appropriateness of the likelihood distribution is often conceded without explicit concern. They suggested that the assumptions of the likelihood model deserved increased attention, and developed a bootstrap that was more robust to violations than its parametric counterparts.

When the sample size is small, prior information is most influential (and arguably at its most useful). Knowledge about the populations' distributional characteristics is less certain using small samples, so it is even more important to use robust procedures that are more protected from deviations from normality and homoscedasticity.

Scenario 1: Expertise in a field with no previous data.

Suppose your research team has an expert with years of clinical experience. Although data haven't been collected in your novel experiment, she believes that it is likely that a *weak* relationship exists between the two considered variables

(operationalized with questionnaires). Fifteen subjects are piloted and a *strong* correlation is observed. Furthermore, the scores seem to suggest that the population joint distribution is not bivariate normal (but this is difficult to determine with only 15 points).

Team members discuss what the population correlation might be before writing the upcoming grant proposal. This important conversation will influence the power analysis, and therefore influence the budget for recruiting subjects. It might even dictate whether different questionnaires should be included. Although the sample correlation is strong, your group remains skeptical of the unexpected results and decides that the population correlation is more likely to be *moderate* than strong. This paper describes statistical procedures that allow the inference to reflect the two sources of information, while being robust to violations of nonnormality.

Scenario 2: Incorporating prior research.

The prior information in the previous scenario came from personal judgments. Previously observed data are a valid source of prior information as well. Suppose you revive a study that your lab conducted several years ago. The expensive protocol hasn't changed, so you want to take advantage of the previous information. For various reasons, you hesitate combining the 24 previous subjects with the 30 new subjects without any adjustment.

One compromise is to include the previous subjects as prior information in your current analyses. The influence of the previous subjects could be reduced by, say, 50%. In effect, the information of 12 subjects is being added to your current experiment of 30 (the relationship between the prior distributions and the previously observed sampled size is discussed later).

Alternatively, suppose that you want to use the information from all subjects in a previously published study, but you don't have access to their raw scores, only the sample correlation. The procedures evaluated here allow a researcher to do this, while testing non-nil hypotheses.

Likelihood Distributions

Univariate Sampling Bootstraps. In previous work, Beasley et al. (2007) evaluated correlation bootstrap procedures that allowed nonzero point hypotheses to be tested with Frequentist inference. We briefly review this procedure, before we describe its application with Bayesian inference.

The **hypothesis imposed** univariate sampling bootstrap (**HI**) allows a researcher to create a confidence interval (CI) to compare to a point hypothesis, ρ_{Hyp}. The five stages are described below and represented in the second column in Figure 1.

1) Collect a sample of N_{Obs} bivariate points: $((X_1, Y_1),\ldots,(X_{Nobs}, Y_{Nobs}))$.

2) Construct the *sampling frame*:

   a) Combine every x_i with every y_i.

   b) Standardize the X and Y variables in this rectangular sampling.

   $x_i' = (x_i - \bar{x})/s_x; y_i' = (y_i - \bar{y})/s_y.$

   c) Impose the value ρ_{Hyp} on the N_{obs}^2 points to create a diagonalized sampling frame. For example[2] transform y_i' to

   $y_i'' = x_i' \times \rho_{Hyp} + y_i' \sqrt{1 - \rho_{Hyp}^2}.$

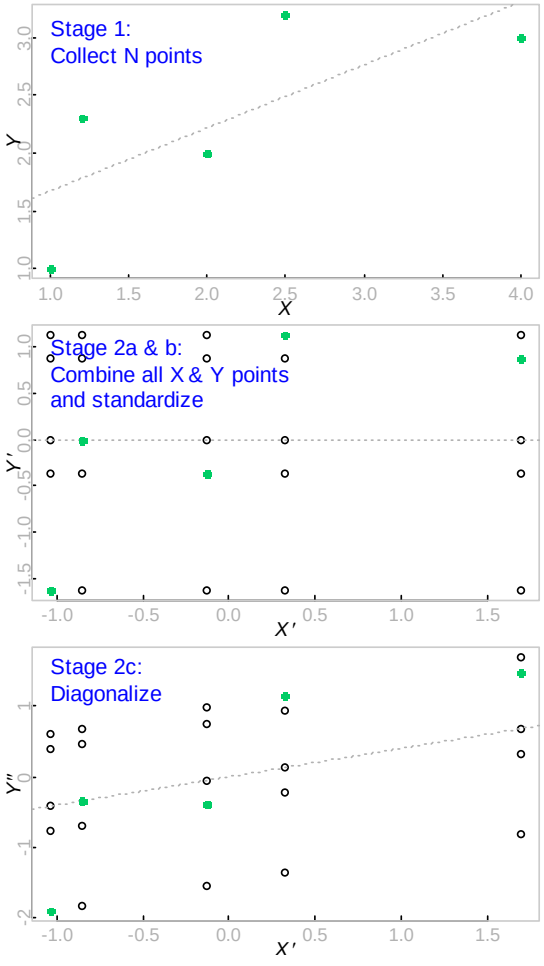
This sampling frame now has a correlation of ρ_{Hyp}, and is used to represent the hypothesized population when creating the bootstrap distribution.

Stage 3 is repeated for B cycles.

3) Draw N_{Obs} points randomly with replacement to create one bootstrap sample.

4) Calculate r^* for each of the B bootstrap samples, which forms a bootstrap distribution of B bootstrap statistics.

Figure 1. Stages 1 and 2 of the SlotHI. The observed points are solid green. The created points are open circles.



[2] Kasier & Dickman (1962) show that any decomposition of the correlation matrix will impose the designed correlation on the sample, assuming it has a mean of zero and unit variance. For this paper, we have chosen a bivariate simplification of the Cholesky decomposition. See the discussion for alternative diagonalization approaches.

This bootstrap distribution can estimate $p(r_{Obs} \mid \rho_{Hyp})$, which is the probability of drawing r_{Obs} from the population, given a population correlation of ρ_{Hyp}. If ρ_{Hyp} represents a null hypothesis, a Frequentist hypothesis can be tested in a fifth stage by comparing r_{Obs} to the 95% CI of $[r^*_{(.025)}, r^*_{(.975)}]$, (these endpoints are the 2.5th and 97.5th percentiles of the bootstrap distribution). More relevant to this paper is that the bootstrap distribution also can be used as the likelihood distribution in Bayesian inference. We will return to this point in the next section.

The construction of the **observed imposed** univariate sampling bootstrap (**OI**) differs from the HI in two ways. In Stage 2c, the rectangular sampling frame has the observed correlation, r_{Obs} imposed upon it (instead of ρ_{Hyp}). Consequently, (instead of r_{Obs}) the value of ρ_{Hyp} is compared to $[r^*_{(.025)}, r^*_{(.975)}]$ in the fifth stage when testing a Frequentist hypothesis.[3]

The inferential performance of the HI and OI is very similar. In previous research, the OI performed slightly better than the HI, although both were preferable to parametric procedures when nonzero Frequentist hypotheses were tested in samples drawn from nonnormal populations (Beasley et al., 2007). The OI's bootstrap distribution also will be used as a likelihood distribution later in the paper.

We believe the term 'univariate *sampling* bootstrap' is more appropriate than 'univariate bootstrap'. The distinguishing feature isn't the number of variables considered in a dataset, but the independence of the sampling of (the multivariate) points.

[3] Even though the procedure of the OI fixes the diagonal to r_{Obs} and creates r^*s that are compared to ρ, this provides a Bayesian posterior of $p(\rho \mid r_{Obs})$ only if the researcher (implicitly or explicitly) assigns a uniform prior, which is discussed below.

Other Bootstraps. The **bivariate** sampling bootstrap (**Biv**) was introduced in the initial bootstrap article (Efron, 1979). Only one procedural difference distinguishes it from the OI. In Stage 2, the observed sample serves as the sampling frame; no values are recombined, standardized or diagonalized. Like the OI, its angle is r_{Obs}; unlike the OI, there are only N_{Obs} points in the sampling frame.

The five stages of the **parametric** bootstrap are identical to the OI and Biv, except that the sampling frame is constructed differently in Stage 2. Parametric assumptions must be made about the population and typically it is assumed that X and Y have a bivariate normal distribution with a correlation of r_{Obs} (Efron, 1982, Section 5.2; Efron & Tibshirani, 1993, Section 6.5).

$$\binom{X}{Y} \sim N\left(\binom{0}{0}, \begin{pmatrix} 1 & r_{Obs} \\ r_{Obs} & 1 \end{pmatrix}\right)$$

This convention is followed in our simulation, but we note that another valid variation could be to set the correlation to ρ_{Hyp} in Stage 2, and compare r_{Obs} to the CI in Stage 5. The distinction between these variations would mirror the distinction between the OI and HI. If the parametric sampling frame had been diagonalized to ρ_{Hyp}, the procedure could be considered a conventional Monte Carlo procedure (see Beasley & Rodgers, 2009, p. 368-372 for more discussion of these relationships).

With apologies to Efron, we will refer to the parametric bootstrap as the **posit** bootstrap (**Pos**, in the sense that the set sampling frame reflects some type of judgment). Our study includes two other procedures that are also 'parametric', so we use 'posit' for distinction. We are not proposing that this term needs to be changed in other contexts.

In Beasley et al. (2007), the inferential performance of the Biv was noticeably worse than the HI or OI, especially at $N_{Obs} = 5$. One explanation could be that the HI and

OI sampling frames provide 118,755 (= *nCr* for ($n = N^2+N$-1, and $r = N$)) unique bootstrap samples, while the Biv sampling frame provides only 126 (= *nCr* for ($n = N+N$-1, and $r = N$)) unique samples (Tucker, 1984, p. 188, Theorem 2). With small samples, this discrepancy means that the Biv bootstrap distribution is more jagged and discontinuous than the HI and OI bootstrap distributions. If the number of unique samples is much smaller than *B*, the Pos will be the smoothest of all the bootstraps, because the number of unique samples with continuous variables is theoretically infinite.

Parametric Procedures. Parametric procedures also are available to model the sampling distribution of a correlation. Fisher's *r*-to-*z* transformation uses the function $\tanh^{-1}$ to transform correlation values into a pivotal statistic on the *z* scale (Fisher, 1915; 1919)[4]. This approximation permits the Gaussian distribution to describe the probability of observing the statistic, given a point hypothesis of ρ_{Hyp}. The hypothesis is typically tested by calculating a standardized test statistic,

$$z_{Test} = (z_{Obs} - \zeta_{Hyp}) / \sigma_{Obs} \quad (1),$$

where $z_{Obs} = \tanh^{-1}(r_{Obs})$, $\zeta_{Hyp} = \tanh^{-1}(\rho_{Null})$, and $\sigma_{Obs} = 1/\sqrt{(N_{Obs} - 3)}$. The *p*-value is the area in a $N(0, 1)$ distribution that is more extreme than z_{Test} (Hays, 1994, Section 14.21). Later we describe a procedure where the relationship is restated slightly: the probability is the area in a $N(z_{Obs}, \sigma^2)$ distribution that is more extreme than ζ_{Hyp}.

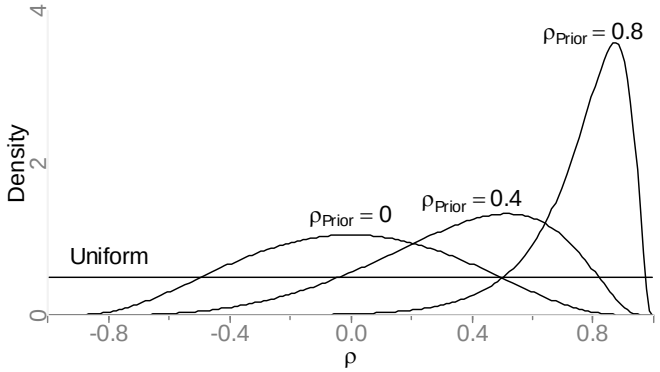
[4] Beasley et al. (2007) evaluated a similar procedure that describes the sampling distribution with a non-central *t* (instead of a normal distribution). Although it performed slightly better than Fisher's procedure with small samples, we are not including it for two reasons. First, Fisher's transformation performs acceptably with small samples. Second, when the normal distribution is used, several aspects of the study are more familiar and more concisely described to readers.

Prior Distributions

The simulation includes the four proper priors shown in Figure 2. One is a bounded uniform distribution, $U(-1, 1)$. The other three are Normal distributions that are later transformed to the ρ metric.[5] On the *z* metric, the three distributions are $N(0, 7^{-.5})$, $N(\tanh^{-1}(.4), 7^{-.5})$ and $N(\tanh^{-1}(.8), 7^{-.5})$. On the *r* metric, these priors are centered on different correlations (i.e., 0, .4, and .8) and are skewed toward zero. We refer to these priors as Gauss00, Gauss04, and Gauss08.

Recall that the standard deviation in Fisher's *r*-to-*z* transformation is $1/\sqrt{(N - 3)}$. The value $7^{-.5}$ was chosen to be equivalent to an observed sample size of $N = 10$. In other words, when $N_{Obs} = N_{Prior} = 10$, r_{Obs} is equally as influential as ρ_{Prior}, and their corresponding *z*s are simply averaged in the analytic procedure (described below).

Figure 2. Prior distributions examined. One uniform and three informative distributions were evaluated with each procedure.



The Slot procedures described below can accommodate any prior distribution. A (transformed) normal prior was used to provide a meaningful comparison with the existing analytic parametric procedure.

[5] A nonuniform scaling factor accounted for the nonlinear *r*-to-*z* transformation. See the *H*( ) function later in the SlotParametric description.

We are intentionally not calling the uniform prior *the* reference prior. The decision that all values of ρ are equal can be considered a prior judgment in itself –and in many situations, it's not plausible that $p(\rho = -.97) = p(\rho = 0)$. Furthermore, this prior is uniformly distributed on the ρ metric, but unimodal on the ζ metric. If we had used a prior that was uniform on ζ, it would be "U" shaped when transformed to ρ; we didn't desire a "reference" prior that implied $p(\rho = .8) > p(\rho = 0)$.

Posterior Distributions

The bootstrap and parametric approaches described above are commonly used to produce a sampling distribution, which is an essential component in Frequentist inference. For example if $r_{Obs} = .672$, a bootstrap distribution can provide the *p*-value of a one-tailed hypothesis, $p(.672 \leq r_{Obs} \leq 1 \mid \rho_{Hyp} = .4)$ by counting the proportion of r^* values that are greater than r_{Obs}.

As stated before, a Frequentist inference fixes ρ_{Hyp} to a hypothetical null point value and the resulting likelihood sampling distribution describes the probability of observing r_{Obs} (or a range of potential r_{Obs} values). In contrast, Bayesian inference conditions on a fixed r_{Obs}, and the resulting posterior distribution describes the probability of observing a ρ_{Hyp} (or a range of ρ_{Hyp}). The posterior distribution, which merges information from the likelihood sampling distribution and the prior distribution, can provide information such as the probability of a one-tailed hypothesis, $p(-1 \leq \rho_{Hyp} \leq .4 \mid r_{Obs} = .672)$, or a two-tailed hypothesis. We evaluate two approaches of creating posterior distributions: a numerical integration approach and an analytical approach.

Bayesian CIs, which are formed from the posterior, have a different and more intuitive interpretation than Frequentist CIs; we defer to other sources for a more complete explanation.[6] As with Frequentist inference, when a hypothesized value falls outside as 95% CI, it is equivalent to having a (Bayesian) *p*-value less than .05.[7]

Slot: Numerical Integration. The range of ρ (i.e., -1 to +1) is divided into *S* nonoverlapping intervals we call *slots*. *S* was fixed to 200, so the slots' midpoints were -.995, -.985,…, .985, .995. Each slot has a specific prior and likelihood value; these are multiplied to produce the slot's value in the posterior distribution.

After the posterior distribution is discretized, the proportional relationship is $p(\rho_i \mid r_{Obs}) \propto p(r_{Obs} \mid \rho_i) \times p(\rho_i)$, where $i = 1,\ldots,200$ and $\rho_i = -.995, -.985,\ldots, .995$. In other words, the slot's posterior probability is proportional to the product of its likelihood and prior probability. If that value is divided by the sum of the products of all 200 slots, the relationship becomes an equality:

$$p(\rho_i|r_{Obs}) = \frac{p(r_{Obs}|\rho_i)p(\rho_i)}{\sum_{k=1}^{S}[p(r_{Obs}|\rho_k)p(\rho_k)]} \quad (2).$$

Because the distributions are discretized, the value of r_{Obs} actually is not a point, but a small slot too, which we informally call the 'observed slot'. For instance, if $r_{Obs} = .607$,

---

[6] Introductions to Bayesian analysis are available in several recent books that are accessible and balanced, such as Carlin & Louis (2009), Albert (2009) and Gill (2008). In short, a Bayesian 95% CI contains the population parameter with 95% probability (given the observed sample and the prior information). In contrast, a Frequentist 95% CI represents an interval that should contain the population parameter 95% percent of the time for similarly constructed CIs (given the observed sample). Pruzek (1999, p. 288) points out, "this conclusion leaves out reference to the specific numerical interval obtained with the extant sample."

[7] Again the Bayesian approach has a different and more intuitive interpretation here. For example, see the "posterior predictive *p*-values" sections in Gelman, Carlin, Stern & Rubin (2004, p. 162 & p. 175-176). Bayesian CIs are also called 'credible intervals' or 'credible sets'. Also see Carlin & Louis (2009, Section 2.3.2) for the related 'highest posterior density' credible set.

the relevant likelihood is $p(.60 < r_{Obs} \leq .61 \mid \rho_i)$. In other words, when the HI sampling frame is diagonalized to ρ_i, the estimated likelihood is the number of bootstrapped stats between .60 and .61.

When this numerical integration approach is combined with one of the likelihood approaches described above, we prepend 'Slot' to the term. The Slot procedures we evaluated are the SlotHI, SlotOI, SlotBiv, SlotPos, and SlotParametric.

SlotHI. Suppose $S = 200$, $r_{Obs} = .607$, and so the observed slot is (.60, .61].

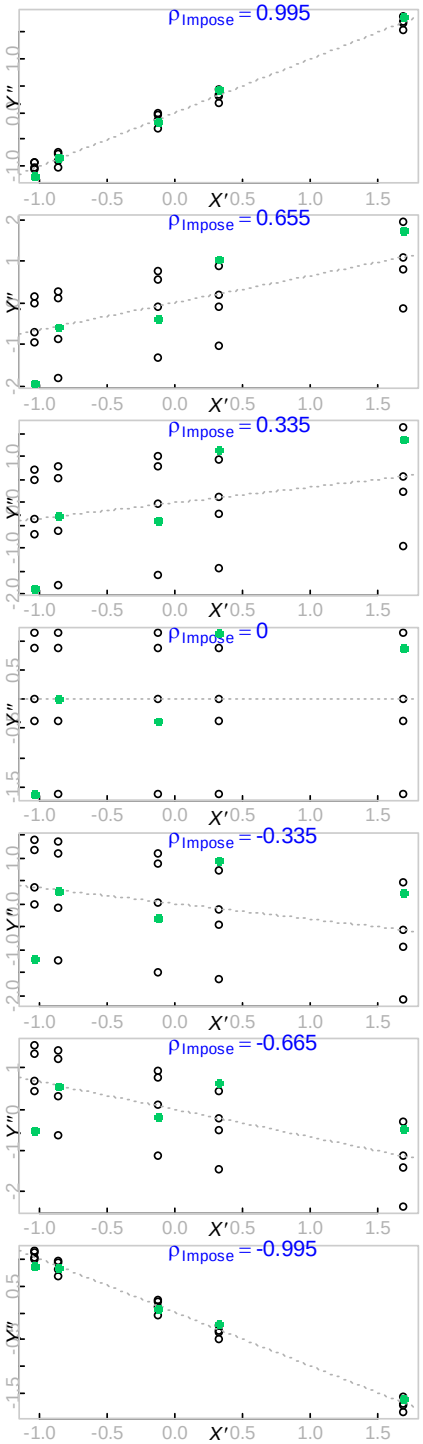
A) Partition ρ into S mutually exclusive *slots*. The i^{th} slot's midpoint is ρ_i.

B) Retrieve the nonnegative prior probability associated with each slot. When a uniform prior is used, the value is a constant .5 (= 1/(1– -1)).

C) Create the S sampling frames and S bootstrap distributions. Each sampling frame has a different imposed value of ρ_i: -.995, -.985,…, .995. Figure 3 portrays seven of these sampling frames (also see Stage 2 in the HI algorithm). From each sampling frame, draw B samples of size N_{Obs}. (In other words, repeat HI Stages 2-4 for each slot).

D) Calculate the likelihood associated with each slot (i.e., $p(r_{Obs} \mid \rho_i)$). For the SlotHI, this is the proportion of r^*s falling within the observed slot.

Figure 3. Step C in the SlotHI algorithm. The range of ρ is divided into S slots. Each slot has a sampling frame that diagonalized to a different value of ρ_{Hyp}. The original 5 observed points are green.



E) If desired, adjust the bootstrap's likelihood distribution after correcting for bias and acceleration. See Stages 1-3 in the section below, "Bootstrap Likelihood Adjustments".

F) Calculate the product of the prior and likelihood for each slot, by dividing each product by the sum of the *S* products (see Equation 2).

G) If a 95% CI is desired, find the two values of ρ that create 2.5% areas in each tail. We used an interpolation technique to compensate for the discrete character of the posterior distribution, described in the 'Interpolating CI Bounds' section.

There are two notable differences between the current SlotHI and the previous HI (Beasley et al., 2007); one difference is procedural and the other involves the resulting sampling distribution. The procedural difference is that the HI uses one sampling frame, and considers the number of r^*s in the entire tail; the SlotHI uses many sampling frames, and considers the number of r^*s falling within the bounds of only the observed slot. In a sense, the HI's CI is directly accessing the bootstrap distribution's CDF (in Stage 5), while the SlotHI is initially using the PDF (in Step D). Later, this PDF will be combined with the prior to produce the posterior (Step F), and the posterior's CDF is accessed by the CI (Step G).

The locations of the procedures' sampling distributions are different. The HI's distribution is roughly centered at ρ_{Null}. The SlotHI's likelihood distribution is roughly centered at r_{Obs} (and its posterior distribution is roughly centered at ρ_{Post}). The likelihoods and posteriors of the OI, Biv, Pos, SlotOI, SlotBiv, and SlotPos are also centered at these values.

SlotOI, SlotBiv, and SlotPos. These three procedures are executed identically to the SlotHI, except for two differences in Step C and one in Step D. First in Step C, whereas the SlotHI constructs S sampling frames (i.e., one for each slot), the other procedures construct only one. Second in Step C, the SlotHI's sampling frames are diagonalized to ρ_i, while the sampling frame of the other procedures has a correlation of r_{Obs}.

Each procedure has a different reason why their sampling frame has a correlation of r_{obs}. The SlotOI explicitly diagonalizes its rectangular sampling frame to r_{Obs}. The SlotPos generates bivariate normal scores that have a r_{Obs} correlation. Finally, the SlotBiv's sampling frame is the observed sample, which of course has a correlation of r_{Obs}.

After their sampling frame is constructed, the three procedures follow the same steps as each other again, and evaluate the likelihood of each slot in Step D. For $i = 1$, the number of r^*s falling between -1 and -.99 are counted. For $i = 2$, the number of r^*s between -.99 and -.98 are counted, and this is repeated a total of S times. Notice that these three procedures construct 1 sampling frame and bootstrap distribution, but consider all S slots. Whereas the SlotHI constructs S sampling frames and bootstrap distributions, but considers only 1 slot (i.e., the slot that r_{obs} falls in). This is a considerable computational disadvantage for the SlotHI (however see the Appendix for optimizations that reduce the discrepancy).

SlotParametric. The SlotParametric is conceptually very similar to the SlotHI, but its likelihood is a Gaussian PDF instead of a bootstrap PDF. Substituting Equation 1 for the likelihood probability in Equation 2 produces

$$p(\rho_i|r_{\text{Obs}}) = \frac{H\big((z_{\text{Obs}} - \zeta_i)/\sigma\big)p(\rho_i)}{\sum_{k=1}^{S}\big[H\big((z_{\text{Obs}} - \zeta_k)/\sigma\big)p(\rho_k)\big]}$$

where $\zeta_k = \tanh^{-1}(\rho_k)$. If *r* and *z* were linearly related, the *H*() function would be the standard normal PDF (e.g., $\varphi(0) = e^0(2\pi)^{-.5}$). Since the relationship is nonlinear, the different slot widths have to be accounted for with the standard normal CDF, Φ(). On the *r* scale, a slot's midpoint is ρ_k, and its upper and lower bounds are $\rho_{k,u}$ and $\rho_{k,l}$, while on the *z* scale, they're ζ_k, $\zeta_{k,u}$, and $\zeta_{k,l}$. The likelihood term is then $H((z_{\text{Obs}} - \zeta_k)/\sigma)$ $= \Phi[(z_{\text{Obs}} - \zeta_{k,u})/\sigma] - \Phi[(z_{\text{Obs}} - \zeta_{k,l})/\sigma]$.

The SlotParametric and SlotPos conceptually bridge the analytic to the bootstrap procedures. The SlotPos generates bootstrap samples and statistics like the SlotHI, SlotOI and SlotBiv, but it assumes a bivariate normal distribution like the SlotParametric and the analytic procedures assume. The SlotParametric is even closer to the analytic procedure because it uses the *r*-to-*z* transformation, but it discretizes ρ like the slot procedures.

Interpolating CI Bounds. The bounds of the CI can be estimated after the slots' posterior probabilities are calculated. To find the lower bound, the probabilities are accumulated until the sum exceeds 2.5%. Then the *CI* boundary is interpolated after comparing the cumulative probability on the *slot's* left and right boundary. For instance, suppose the cumulative probability is 2.1% for $\rho_{141,l} = .40$, and 2.6% for $\rho_{141,u} = .41$. The estimated critical value would be .4 + (.025 - .021)/(.026 -.0 21) × (.41-.40) = .408. In other words, the cumulative probability of .025 is 80% of the distance between .021 and .026, so the interpolated critical value is 80% of the distance between .4 and .41. The process is mirrored for the CI's upper bound.

Analytic Procedure

The *r*-to-*z* transformation accommodates Bayesian inference with three steps. First, the correlation values are transformed to the *z* metric. Second, the prior and observed information are synthesized. Third, the CI is determined and transformed back to the *r* metric. This is the only procedure in the simulation that doesn't partition ρ into slots.

Step 1: The values are transformed to the *z* metric. The two location values are $z_{Obs} = \tanh^{-1}(r_{Obs})$, and $\zeta_{Prior} = \tanh^{-1}(\rho_{Prior})$. The two standard deviations are $\sigma_{Obs} = 1/\sqrt{(N_{Obs} - 3)}$ and $\sigma_{Prior} = 1/\sqrt{(N_{Prior} - 3)}$. The interpretation of N_{Prior} was discussed in the subsection, 'Prior Distributions'. Later equations are cleaner if *standard deviations* are converted into *precisions*. The precision[8] is commonly defined as the inverse of the variance, so $\tau_{Obs} = (1/\sqrt{(N_{Obs} - 3)})^{-2} = N_{Obs} - 3$ and $\tau_{Prior} = N_{Prior} - 3$.

Step 2: After the transformation, the likelihood and prior distributions assume a normal distribution. A closed form equation describes the conjugate relationship. The posterior distribution is $\sim N(z_{Post}, 1/\tau_{Post})$, where the mean and precision are:

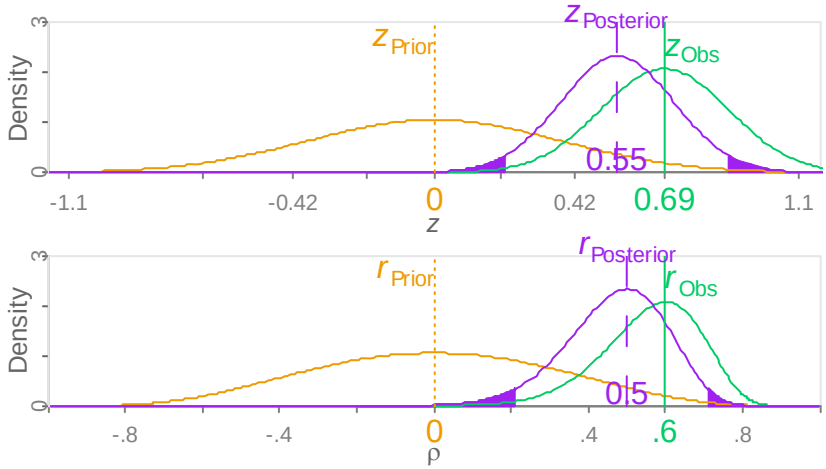
$$z_{Post} = \frac{(z_{Obs})\tau_{Obs} + (z_{Prior})\tau_{Prior}}{\tau_{Obs} + \tau_{Prior}} \text{ and } \tau_{Post} = \tau_{Obs} + \tau_{Prior} = 1 / \sigma_{Post}^2 \qquad (3).$$

The posterior mean is the average of the observed and prior mean, after weighting by their precisions (Carlin & Louis, 2009, Equations 2.3-2.4). In a sense, z_{Obs} and σ_{Obs} in Equation 1 is replaced by z_{Post} and σ_{Post}. A uniform prior distribution is accommodated by reducing N_{Prior} to 3, so that the prior precision is zero (and the prior variance is infinite). When the prior is uniform, the equation for z_{Post} reduces to z_{Obs}.

[8] If N_{Obs} (or N_{Prior}) were less than 3, we would restrict τ_{Obs} (or τ_{Prior}) to zero. However N_{Obs} is never less than 5 in the simulation.

Step 3: On the z scale, the 95% CI is $z_{Post} \pm 1.96/\sqrt{\tau_{Post}}$. On the r scale, the interval becomes $\tanh(z_{Post} \pm 1.96/\sqrt{\tau_{Post}})$. A hypothesis can be tested by comparing these intervals to ζ_{Hyp} and ρ_{Hyp}.

Figure 4. Example of the analytic procedure. The purple posterior is a compromise between the orange prior and green likelihood distributions. In this case, $\rho_{Prior} = 0$, $r_{Obs} = .6$, $N_{Prior} = 10$, and $N_{Obs} = 30$. The top panel is on the z scale, while the bottom is on the r scale. A hypothesis is rejected if ζ_{Hyp} (or ρ_{Hyp}) falls in the purple tails of the posterior distribution.



Bootstrap Likelihood Adjustments

Effort had been dedicated to improving the accuracy and correctness of the bootstrap CIs (for a summary, see Efron & Tibshirani, 1993, Section 22.2). The same principles that adjusted the endpoints of the CI can be applied to the entire bootstrap likelihood distribution.

The bootstrap procedures described have implicitly been using the *percentile* method: the quantile of the bootstrap distribution maps directly to the quantile of the inferred population. When the r^*s are sorted, they can be viewed as an empirical cumulative distribution function (CDF_{Perc}), $\hat{G}$, whose values are bounded by [0, 1].

Suppose c is a valid correlation value. If 86% of the r^*s are smaller than c, then $\mathrm{CDF}_{\mathrm{Perc}}(c) = \hat{G}(c) = .86$. The critical values for a (1 – α)% CI can be expressed as

$$\left(\hat{G}^{-1}\left(\frac{\alpha}{2}\right), \hat{G}^{-1}\left(1-\frac{\alpha}{2}\right)\right) = \left(\hat{r}^{\left(\frac{\alpha}{2}\right)}_{\mathrm{LowerCrit}}, \hat{r}^{\left(1-\frac{\alpha}{2}\right)}_{\mathrm{UpperCrit}}\right)$$

However the percentile method's direct mapping does not lead to the best population inference in most conditions. Efron developed two successive adjustments, the bias-corrected method (BC; 1982) and the bias-corrected and accelerated method (BC$_a$; 1987). The BC considers the bias in the bootstrap distribution, z_0. The BC$_a$ additionally considers the acceleration, a, which adjusts for heteroscedasticity in the statistic[9]. The CDFs for the two adjustments are:[10]

$$\begin{aligned} \mathrm{CDF}_{\mathrm{BC}}(c) &= \Phi\left[\Phi^{-1}\left(\hat{G}(c)\right) - 2z_0\right] \\ \mathrm{CDF}_{\mathrm{BC}_a}(c) &= \Phi\left[\frac{\Phi^{-1}\left(\hat{G}(c)\right)(1-az_0)-2z_0}{\Phi^{-1}\left(\hat{G}(c)\right)(a)+1}\right] \end{aligned} \quad (4).$$

The adjustments are made to the likelihood distribution, before the prior is applied.

1) The bootstrap distribution is accumulated to create a cumulative distribution:

   $\mathrm{CDF}_{\mathrm{Perc}}(c) = \mathrm{CDF}^{(i)}_{\mathrm{Perc}} = \sum_{k=1}^{i} p(r_{\mathrm{Obs}}|\rho_k)$, for i = 1, 2,…, S.

   The $\mathrm{CDF}_{\mathrm{Perc}}$ values corresponding to the left and right boundary of the i^{th} slot are $\mathrm{CDF}^{(i-1)}_{\mathrm{Perc}}$ and $\mathrm{CDF}^{(i)}_{\mathrm{Perc}}$. There are $S + 1$ CDF values, beginning and ending with $\mathrm{CDF}^{(0)}_{\mathrm{Perc}} = 0$, and $\mathrm{CDF}^{(S)}_{\mathrm{Perc}} = 1$.

2) After substituting $\mathrm{CDF}_{\mathrm{Perc}}$ for $\hat{G}$ in Equation 4, $\mathrm{CDF}_{\mathrm{BC}}$ and $\mathrm{CDF}_{\mathrm{BC}a}$ are calculated.

[9] Positive acceleration indicates that, as ρ increases, so does its standard error (Efron & Tibshirani, 1993, Sections 14.3 & 22.2)

[10] This uses the standard Gaussian CDF and its inverse (e.g., $\Phi(1.96) \approx .975$ and $\Phi^{-1}(.975) \approx 1.96$).

3) The CDFs are transformed back into PDFs:

$\mathrm{PDF}_{\mathrm{Perc}}(\rho_i) = p(r_{\mathrm{Obs}}|\rho_i)$

$\mathrm{PDF}_{\mathrm{BC}}(\rho_i) = \mathrm{CDF}_{\mathrm{BC}}(\rho_i) - \mathrm{CDF}_{\mathrm{BC}}(\rho_{i-1})$

$\mathrm{PDF}_{\mathrm{BC}_a}(\rho_i) = \mathrm{CDF}_{\mathrm{BC}_a}(\rho_i) - \mathrm{CDF}_{\mathrm{BC}_a}(\rho_{i-1})$, for $i = 1,\ldots,S$.

These steps are unnecessary for the percentile CI, because $\mathrm{PDF}_{\mathrm{Perc}}$ is simply the likelihood produced by the bootstrap. When the estimated bias and acceleration are zero, $\mathrm{CDF}_{\mathrm{BC}}$ and $\mathrm{CDF}_{\mathrm{BC}a}$ reduce to $\mathrm{CDF}_{\mathrm{Perc}}$.

We experimented with a second type of BC_a we tentatively call the BC_a straddle (BC_{as}). The CDF equation is the same (Equation 4), but the acceleration term is estimated differently. Details are found in the Appendix.

Notice that there are two unrelated occasions during the Slot algorithm that a PDF is accumulated to produce a CDF. The first occasion is in Step E, when the *likelihood* PDF is transformed into a likelihood CDF, which facilitates the BC and BC_a adjustments. In Step G, the *posterior* PDF is transformed into a posterior CDF, which determines the CI boundaries.

Point Estimates

In addition to constructing CIs, the posterior distributions were used to estimate ρ_{Post}. The analytic procedure's estimate is r_{Post} (= $\tanh^{-1}(\zeta_{\mathrm{Post}})$). The slot procedures' estimate is the value of each slot, weighted by the slot's posterior probability, $r_{\mathrm{Post}} = \sum_{k=1}^{S}[\rho_k \times p(\rho_k|r_{\mathrm{Obs}})]$. The bias and mean squared error (MSE) were assessed for these estimates.

The correlation sampling distribution is biased when $\rho \neq 0$; the bias decreases as N_{Obs} grows. In addition to r_{Post}, we evaluated an adjustment for r (Olkin & Pratt, 1958,

Equation 2.6) that is intended to decrease the bias. However in our simulations, it typically produced estimates with larger bias and MSE than the unadjusted r_{Post}, so the results are not presented here.

Distinction from Rubin's Bayesian Bootstrap

The procedures described in this paper are operationally different than the procedure Rubin broadly called a Bayesian bootstrap (1981; Efron, 1982, Section 10.7). Rubin's procedure assigns a prior distribution to the *observed scores*. However the Slot procedures assign a prior distribution to the *parameters* (as do the procedures in Boos & Monahan, 1986; Efron & Tibshirani, 1993, p. 358, also mentioned combining a bootstrap with a parameter's prior distribution).

We propose that direct interaction with the parameters is frequently more useful to behavioral research than interaction with the observed scores. Theoretically, if a well-defined prior distribution of the observations exists, a prior distribution of the parameter(s) can be derived; however we don't think this approach practical. Rubin himself wasn't fond of the Bayesian bootstrap. The final section of his 1981 article, which could be described as a deliberate reduction to absurdity, intentionally argues against his procedure's relevance. We feel that the researcher is more likely to have helpful and reliable prior information about the parameters than the observed scores, especially when multivariate questions are addressed.

Evaluations of Inferential Procedures, not of Inferential Philosophies

The procedures included in this study are tools that can be used by different schools of inference. Our goal is not to evaluate the arguments in these debates (e.g., CIs *vs*. *p*-values, Bayesian *vs.* Frequentist, or hypothesis tests *vs.* point estimation), but to

evaluate the relevant statistical properties of the procedures. From a statistical perspective, the conclusions of the different philosophies were equivalent sometimes (e.g., the procedures with desirable Type I error rates also will have desirable CI coverage). The conclusions were related sometimes (e.g., the procedures with robust Bayesian inferences typically had robust Frequentist inferences). And other times the conclusions were divergent (e.g., the procedure that produced the best point estimates sometimes produced the least robust CIs).

To address the characteristics relevant to these different philosophies, we evaluated six statistical procedures using twenty-seven unimodal populations. Five outcomes are summarized: bias, MSE, Type I error rates, Type P error rates and statistical power.

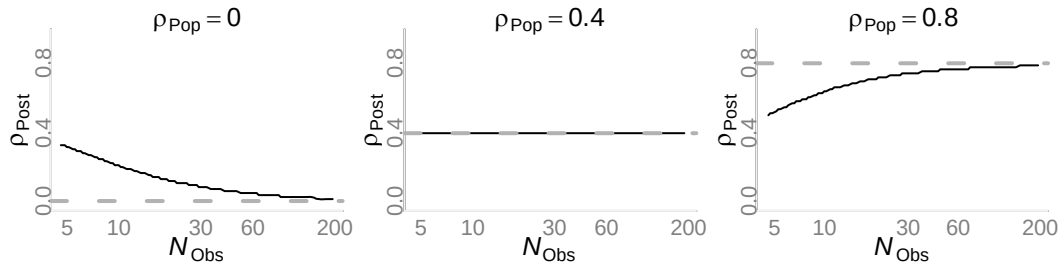
Method

Before describing the factors used in the simulation, we first describe how the procedures were evaluated.

Bias and MSE

Bias and MSE are assessed by comparing the estimated posterior mean, r_{Post}, to the ideal value, ρ_{Post}. The value of ρ_{Post} is a combination of ρ_{Pop} and ρ_{Prior}; it is calculated using the closed-form parametric procedure, except r_{Obs} is replaced with the population value, ρ_{Pop}. With a uniform prior, ρ_{Post} will equal ρ_{Pop}. Three Gaussian priors are shown in Figure 2. Figure 5 shows the relationship between ρ_{Post}, ρ_{Pop}, and N_{Obs} when $\rho_{Prior} = .4$. When the prior distribution is centered on the population value, the ρ_{Pop} and ρ_{Post} are equal for all sample sizes (e.g., the middle panel of Figure 5).

Figure 5. The value of ρ_{Post} as a function of N_{Obs} and ρ_{Pop}. The prior is Gaussian with $\rho_{Prior} = .4$ and $N_{Prior} = 10$. As N_{Obs} increases, ρ_{Post} (solid line) approaches ρ_{Pop} (dashed line).



Type I and Type P Error Rates

Type I error rate is the proportion of times the ρ_{Pop} is incorrectly excluded by a CI, which reflects the likelihood distribution only. Its nominal value is α, and α_{Obs} is estimated with simulation studies. However, in this evaluation of Bayesian procedures,

we want to identify procedures that reliably incorporate observed *and prior* information with nonnormal populations. We are tentatively calling the incorrect rejection of ρ_{Post} the 'Type P error rate', and its nominal and simulation values are γ and γ_{Obs}.

The value of γ asymptotes to α as N_{Obs} grows infinitely large and N_{Prior} is held constant to a positive value. It also will equal α when no prior information is considered (i.e., $N_{Prior} \leq 3$ and $\tau_{Prior} = 0$). However γ will be zero when there is no observed information[11] ($N_{Obs} \leq 3$ and $\tau_{Obs} = 0$). This is because sampling variability does not exist, and the CI already is centered at $\rho_{Prior} = \rho_{Post}$.

For values of N_{Obs} and N_{Prior} between 3 and infinity, the false rejection rate of ρ_{Post} is:

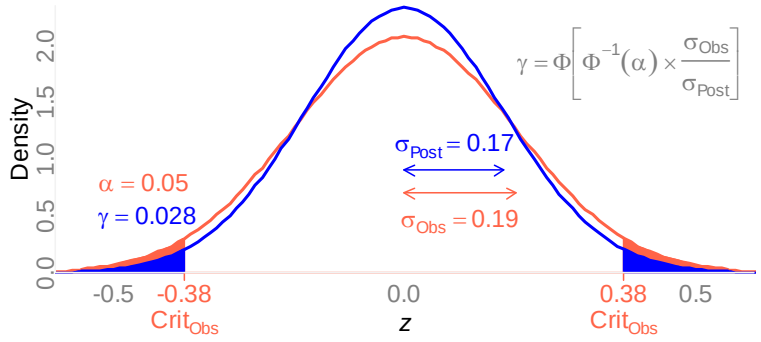
$$\gamma = \Phi\left[\Phi^{-1}(\alpha) \times \frac{\sigma_{Obs}}{\sigma_{Post}}\right] = \Phi\left[\Phi^{-1}(\alpha) \times \sqrt{\frac{\tau_{Post}}{\tau_{Obs}}}\right] = \Phi\left[\Phi^{-1}(\alpha) \times \sqrt{\frac{\tau_{Obs}+\tau_{Prior}}{\tau_{Obs}}}\right]$$

When prior information is considered, the posterior distribution will have greater precision (i.e., be narrower) than the likelihood distribution. Therefore $\pm Crit_{Obs}$, which cuts off α of the likelihood distribution, will cut off less than α of the posterior distribution (in Figure 6, compare the larger orange tails of the likelihood distribution to the smaller blue tails of the posterior distribution).

When a uniform prior is used, the variability in the posterior and likelihood are equal, so $\gamma = \alpha$ and therefore Type P and Type I error rates are equal. Another perspective is that $\rho_{Post} = \rho_{Pop}$ when a prior is uniform; because Type P is the rejection o f ρ_{Post} and Type I is the rejection of ρ_{Pop}, the two error rates are equivalent.

[11] When both N_{Obs}, $N_{Prior} \leq 3$, the Type P rate and ρ_{Post} are undefined, since the denominator of Equation 3 is zero. However, you could argue that in the combination of no observed information and a uniform prior distribution (which weights all value of ρ equally), $\rho_{Post} = 0$.

Figure 6. Comparison of Type I and Type P error rates (α and γ) when N_{Obs} = 30, N_{Prior} = 10, and $\rho_{Prior} = \rho_{Pop} = 0$.



Simulation Factors

Overall Design. Our experimental design consists of six completely crossed factors and one partially nested factor. Each factor has 4 to 8 levels, with a total of 11,648 cells. The factors are: (1) population distribution (8 levels), (2) population correlation: ρ_{Pop} (4 levels), (3) hypothesized correlation: ρ_{Hyp} (3 levels), (4) observed sample size: N_{Obs} (3 levels), (5) statistical procedure (6 levels), (6) CI method (4 levels), and (7) prior distribution (4 levels).

Population Distribution (Factor 1). Population scores were generated by an approach developed by Headrick (2010). The populations of the primary simulation were built from combinations of a normal distribution, two skewed distributions ($\chi^2(df=1)$ and $\chi^2(df=3)$), and a negatively kurtotic distribution (Beta(α=2,β=2)). The seven bivariate distributions were (a) Normal*X*Normal*Y*, (b) Chi1*X*Chi1*Y*, (c) Normal*X*Chi1*Y*, (d) Chi1*X*Normal*Y*, (e) Chi3*X*Chi3*Y*, (f) Beta22*X*Beta22*Y*, (g) Normal*X*Beta22*Y*, and (h) Beta22*X*Normal*Y*. Figures 7 and 8 show the marginal distributions and some of the joint distributions.

We initially examined a pool of 11 univariate (and 27 bivariate) distributions in a small simulation with 20,000 replications. These were chosen to represent the range of

nonnormal unimodal distributions identified by Micceri (1989) that applied researchers commonly encountered. Their distributional properties and Frequentist performance are summarized in Beasley et al. (2007, Table 6).

From the pool, 4 univariate (and 8 bivariate) distributions were selected as the most challenging. In other words, they revealed the liberal behavior of the examined procedures; the 18 distributions that were not selected for the primary simulation had rejection rates closer to α and γ than those that were selected. It is not surprising that Chi-Square(*df* = 1) has the largest skew and positive kurtosis, while Beta(2,2) has the largest negative kurtosis of the pool. The heteroscedastic Chi1*X*Chi1*Y* distribution was very problematic for some procedures, so we included a less severe heteroscedastic distribution that used Chi-Square(*df*=3). The procedures were robust with sharply peaked symmetric distributions like the Laplace (also called the double exponential).

Figure 7. The standardized univariate distributions of the simulated populations: Normal, Chi1, Chi3 and Beta22.

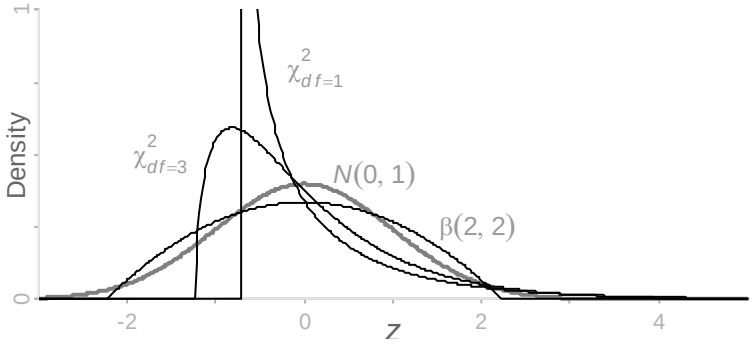


Figure 8. Simulated populations. For each cell, the contours were calculated from 10^5 points and 10^3 points are plotted. The calculated and theoretical correlations are (almost indistinguishable) dashed lines, while the loess curve is solid. Axes of Normal marginals extend from -3.5 to 3.5. Axes of the *standardized* Chi-Square(1) and Beta(2, 2) extend from-1 to 6 and from -2.5 to 2.5, respectively.

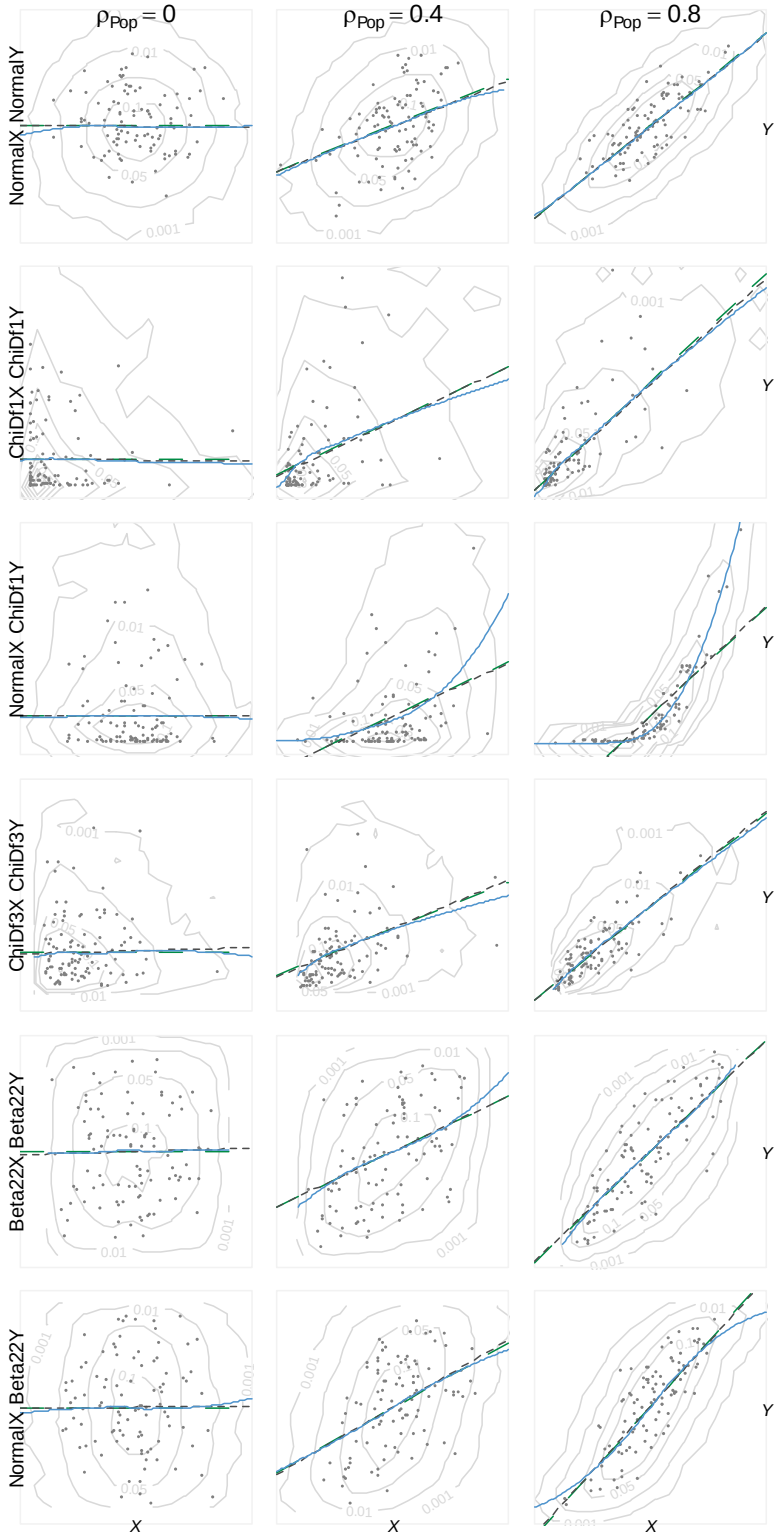


Table 1
Summary of the different definitions of correlations used in the paper.

ρ_{Prior}	Center of the prior (on the *z* scale); Factor 7. *Specified by researcher.*
ρ_{Pop}	Theoretical and generated value in the population; Factor 2. *Unknown to researcher.*
ρ_{Post}	Theoretical center of posterior (synthesis of ρ_{Prior} and ρ_{Pop}); determined by Factors 2 & 7. *Unknown to researcher.*
ρ_{Hyp}	Hypothesized value that is compared to the CI; Factor 3. *Specified by researcher.*
r_{Obs}	Estimate of ρ_{Obs}. *Calculated by researcher.*
r_{Post}	Estimate of ρ_{Post} (synthesis of r_{Prior} and r_{Pop}). *Calculated by researcher (it is influenced by the specified ρ_{Prior}).*

Population ρ_{Pop} (Factor 2) and Hypothesized ρ_{Hyp} (Factor 3). Four correlations were used to generate population distributions, ρ_{Pop} = 0, 0.4, 0.6, and 0.8. Four correlations were simulated as the null hypothesis, ρ_{Hyp} = 0.0, 0.4, 0.6, and 0.8. We note that results would be symmetric (except for sampling error) for negative ρ_{Pop}s. These are two of the six different types of correlations defined in this paper; they are summarized in Table 1.

Sample Size (Factor 4). Eight different observed sample sizes were generated, N_{Obs} = 5, 10, 15, 30, 60, 200, 500, 1,000. We initially included only $N_{Obs} \leq 60$, because they represent experiments and nested studies with small-to-moderate sample sizes in which statistical procedures are thought to be most vulnerable to violations of the normality assumption. However some of the error rates remained very liberal at N_{Obs} = 60 and we were then interested if α_{Obs} would stabilize and approach α with larger samples. All 8 samples sizes are displayed in Figures 9-15, while Table 2 contains only N_{Obs} = 10, 60, and 1,000.

Statistical Procedure (Factor 5). This factor is the core feature of the study. Six procedures were examined: (a) SlotHI, (b) SlotOI, (c) SlotBiv, (d) SlotPos, (e) SlotParametric, and (f) the parametric. Our basic goal is to allow these procedures to compete in their power and control of Type I and Type P errors, as well as bias and MSE. These procedures tested the same batch of generated samples, which provides better comparisons between the procedures. For instance, all procedures tested the same 500,000 samples of size $N_{Obs} = 15$ drawn from the Normal$X$Skew$Y$ population where $\rho_{Pop} = 0.4$.

Statistical and CI Methods (Factor 6). This is the only (partially) nested factor of the design. Some bootstrap procedures were not crossed with all four CI methods. The SlotOI and SlotBiv were tested with the percentile, BC, BC$_a$ and BC$_{as}$. The SlotPos used the percentile and BC, while the SlotHI used just the percentile. No adjustments were used with the two parametric procedures.

The estimated values informing the BC, BC$_a$, and BC$_{as}$ (i.e., z_0 and a) are calculated from a sampling frame. Currently it is not clear to us how these estimates are conceptually related to the SlotHI, because it has a different sampling frame for each of its slot.[12]

To minimize the stochastic effects of bootstrapped distributions, all CIs were calculated from the same bootstrap distribution for a given procedure and sample, which provides better comparisons between the CI methods. To reduce the complexity of the results, only the best CI method is reported for each procedure. The BC was the best for

[12] One possible solution is to adjust each slot (and thus each estimated likelihood for a given ρ) independently. Another possibility is using each Slot (which represents values of ρ from -1 to 1) to estimate how the standard error of ρ changes linearly with its value. Ultimately, the SlotHI was found to be very robust, even with a percentile CI.

SlotBiv (although the BC_a and BC_{as} were very similar) and for SlotPos. The BC_{as} was best for the SlotOI.

Prior Distributions (Factor 7). The four prior distributions described in the Introduction were examined with each procedure (see Figure 2).

Apparatus / Computer Architecture

The simulation code was written in C# 3.0, which called the bootstrap code written in C++ with SSE4.1 intrinsics. Up to 14 instances of the simulation ran independently on 5 single-socket processors. Results were saved to Microsoft SQL Server 2008. The simulation took 30 days to create 500,000 replications in each cell. For the SlotOI, SlotPos and SlotBiv, each bootstrap distribution contained 9,999 r^*s. For the SlotHI when $N_{Obs} \leq 60$, each of its 200 slots contained 1,999 r^*s; when $N_{Obs} \geq 200$, each of its 400 slots contained 4,999 r^*s.[13] Regarding variability in the simulation estimates, nominal error rates of $\alpha_{Obs} = .05$ have a 95% CI of [.0494, .0506] (i.e., $.05 \pm 1.96 \times \sqrt{(.05 \times .95 / 500{,}000)}$).

Two types of random number generators (RNG) were used. Bootstrap routines used a 59-bit multiplicative congruential generator (Intel, 2009) because the randomness and period length requirements for selecting 9,999 indices (or 2×9,999 continuous values for each SlotPos) are not very demanding. A cryptographically strong RNG generated the population scores; unlike conventional RNGs, it does not accept a seed and does not

[13] As the width of the observed slot decreases, we suspect it's important to increase B in the SlotHI. For example, when $S = 200$ and $r_{Obs} = .672$, the likelihood is partially determined by the proportion of r^*s falling in the interval (.67, .68]. Increasing S to 400 shrinks the bounds of the observed slot to (.670, .675] and potentially decreases the stability of the estimate $p(r_{Obs}| \rho)$ (which is operationalized as $p(.670 \leq r_{Obs} < .675 | \rho)$).

produce a predictable sequence of values (Toub & Farkas, 2007). Values from this RNG produced seeds that initialized the bootstrap RNGs.

Results

The procedures were evaluated by their performance on Type I error rates, Type P error rates, power, bias and MSE. Each statistic is based on the average across 500,000 replications. For example, Type I, Type P, and power statistics are based on average rejection rates across 500,000 replications.

Type I (and Type P) Error with a Uniform Prior

Figure 9 shows Type I error with a uniform prior. The upper left panel indicates the rate of incorrectly rejecting $\rho_{Post} = 0$ in a Normal*X*Normal*Y* population. The nominal α was set to .05. Because the procedures are exposed to many nonideal situations (such as small samples and correlated, nonnormal populations), their performance can be expected to deviate from the ideal .05. Under these nonideal conditions, we consider a procedure's error desirable if it is less than .075. Each panel in Figure 9 has a gray band covering $.025 < \alpha < .075$. A thin white line marks $\alpha = .05$. Recall that when the prior is uniform, $\rho_{Post} = \rho_{Pop}$, so Type P and Type I error rates are equal.

In the top left panel, no procedure exceeds the upper limit of the gray band, indicating none are excessively liberal for this specific condition. In a sense, this panel is the 'easiest' test for an inferential procedure, and thus good performance is expected. Moving down a column changes the population distribution. For the first column, ρ_{Pop} and ρ_{Hyp} are zero, so the vertical axis is the proportion of times the researcher incorrectly rejects the hypothesis that there is no linear relationship in the population. In the second column, $\rho_{Pop} = \rho_{Hyp} = 0.4$, so this panel represents incorrectly rejecting the hypothesis that there is a moderate correlation. The third and fourth columns represent $\rho_{Pop} = \rho_{Hyp} = 0.6$ and $\rho_{Pop} = \rho_{Hyp} = 0.8$.

Consistent with Beasley et al. (2007), all evaluated procedures behave desirably when the population is bivariate normal (i.e., panels in the first row) or when the variables are uncorrelated (i.e., panels in the first column). Otherwise, control of Type I and Type P error is not assured for the parametric procedures. They are very liberal in heteroscedastic populations, such as Chi3*X*Chi3*Y* and Chi1*X*Chi1*Y*, always exceeding $\alpha = \gamma = .09$ when $N_{Obs} \geq 10$. The liberalness is evident with correlations as low as $\rho_{Pop} = .4$. The error rates did not return to acceptable levels as sample size increased –the control deteriorated further, in fact. When $N_{Obs} = 1{,}000$ in the Chi3 and Chi1 distributions, α_{obs} reached .14 and .27 (the values in Figures 9-10 are truncated at a ceiling of $\gamma = .2$).

The error rates of the SlotHI and SlotOI were usually within .01 of each other and under .075. The notable exception occurred when Beta22XBeta22*Y* was strongly correlated at $\rho_{Pop} = .8$. For the SlotHI, α peaked at .114 before falling to .086 when $N_{Obs} = 1{,}000$; the SlotOI stayed slightly above $\alpha_{Obs} = .10$ for $N_{Obs} \geq 60$. When $\rho_{Pop} \leq .6$, neither had an α_{Obs} exceed .083.

The SlotBiv performed desirably in the heteroscedastic populations, but exhibited serious problems in populations with strong nonlinear relationships (e.g., Normal*X*Chi1*Y* and Normal*X*Chi3*Y*).

The procedures with parametric distribution assumptions (i.e., the analytic, SlotParametric, and SlotPos) performed very similarly. In Figures 9-12, the red cross, the blue cross, and the purple diamond usually are on top of each other. This suggests the Slot's discretization was an acceptable approximation of the continuous parameter.

Figure 9. Type P error with a uniform prior. Because the prior is uniform, these are also Type I error rates. The columns represent population correlations, while the rows are the population distribution. Vertical locations are truncated to α = .2. The panel replaced by the legend is replicated in the middle panel of Figure 11.

Based on the results above, we narrowed our attention to the analytic and SlotHI procedures, which are shown in Table 2. These are the same values used to create the plots in Figure 9. Rates that exceeded .075 (marked by the top of the gray band in Figure 9), are indicated in blue. Rates that exceeded .090 are red. If liberal error rates are a concern, these results suggest the SlotHI is more robust than the parametric procedure.

Table 2

Rates for incorrectly rejecting ρ_{Post} (and ρ_{Pop}) for the two best performing procedures. The nominal rate is .050. Rates exceeding .075 are blue, and rates exceeding .10 are red; the corresponding row headers are bolded.

NormalXNormalY	ρ_{Post}			
N_{Obs} = 10	.0	.4	.6	.8
Analytic	.051	.050	.050	.048
SlotHI	.032	.038	.052	.075
N_{Obs} = 60				
Analytic	.050	.051	.050	.050
SlotHI	.047	.050	.053	.060
N_{Obs} = 1,000				
Analytic	.050	.050	.050	.050
SlotHI	.056	.058	.048	.060

NormalXChi1Y	ρ_{Post}			
N_{Obs} = 10	.0	.4	.6	.8
Analytic	.051	.046	.037	.039
SlotHI	.033	.025	.018	.002
N_{Obs} = 60				
Analytic	.051	.041	.029	.023
SlotHI	.047	.022	.008	.004
N_{Obs} = 1,000				
Analytic	.050	.041	.030	.034
SlotHI	.056	.016	.003	.001

Chi3XChi3Y	ρ_{Post}			
N_{Obs} = 10	.0	.4	.6	.8
Analytic	.052	**.078**	**.090**	**.101**
SlotHI	.044	.035	.048	**.077**
N_{Obs} = 60				
Analytic	.050	**.093**	**.111**	**.128**
SlotHI	.060	.058	.051	.053
N_{Obs} = 1,000				
Analytic	.050	**.099**	**.120**	**.137**
SlotHI	.058	.058	.045	.036

Beta22XBeta22Y	ρ_{Post}			
N_{Obs} = 10	.0	.4	.6	.8
Analytic	.052	.052	.053	.054
SlotHI	.034	.041	.056	**.082**
N_{Obs} = 60				
Analytic	.050	.052	.053	.055
SlotHI	.047	.058	.073	**.103**
N_{Obs} = 1,000				
Analytic	.050	.052	.053	.056
SlotHI	.056	.062	.061	**.086**

ChiXChi1Y	ρ_{Post}			
N_{Obs} = 10	.0	.4	.6	.8
Analytic	.059	**.128**	**.165**	**.194**
SlotHI	.068	.031	.047	.079
N_{Obs} = 60				
Analytic	.048	**.167**	**.211**	**.243**
SlotHI	.080	.068	.051	.048
N_{Obs} = 1,000				
Analytic	.049	**.195**	**.242**	**.275**
SlotHI	.058	.075	.044	.026

NormalXBeta22Y	ρ_{Post}			
N_{Obs} = 10	.0	.4	.6	.8
Analytic	.051	.050	.049	.047
SlotHI	.032	.039	.050	.072
N_{Obs} = 60				
Analytic	.051	.049	.048	.044
SlotHI	.047	.051	.059	.066
N_{Obs} = 1,000				
Analytic	.050	.049	.048	.044
SlotHI	.056	.058	.047	.070

Type P Error with Nonuniform Priors

Figure 10 shows the rejection rates of ρ_{Post} when a nonuniform prior is used. The gray band marks the values of γ that correspond to $.025 < \alpha < .075$. Recall that as N_{Obs} grows, the precision of the likelihood approaches the precision of the posterior, so γ asymptotes to α. The white lines in the center of the gray band designate the γ corresponding to $\alpha = .05$.

The results of the nonuniform priors were fairly consistent with the uniform prior. Typically the Bayesian inference was liberal only if the Frequentist inference was liberal (and we mention the exceptions below). Furthermore, the Type P error rates of the three Gaussian prior distributions closely resembled each other. Gauss04 was chosen to represent the nonuniform priors in Figure 10 partly because it is the worst-case scenario for the SlotHI. We think it controlled γ and α well with Gauss04, and it was controlled even better with the other Gaussian distributions.

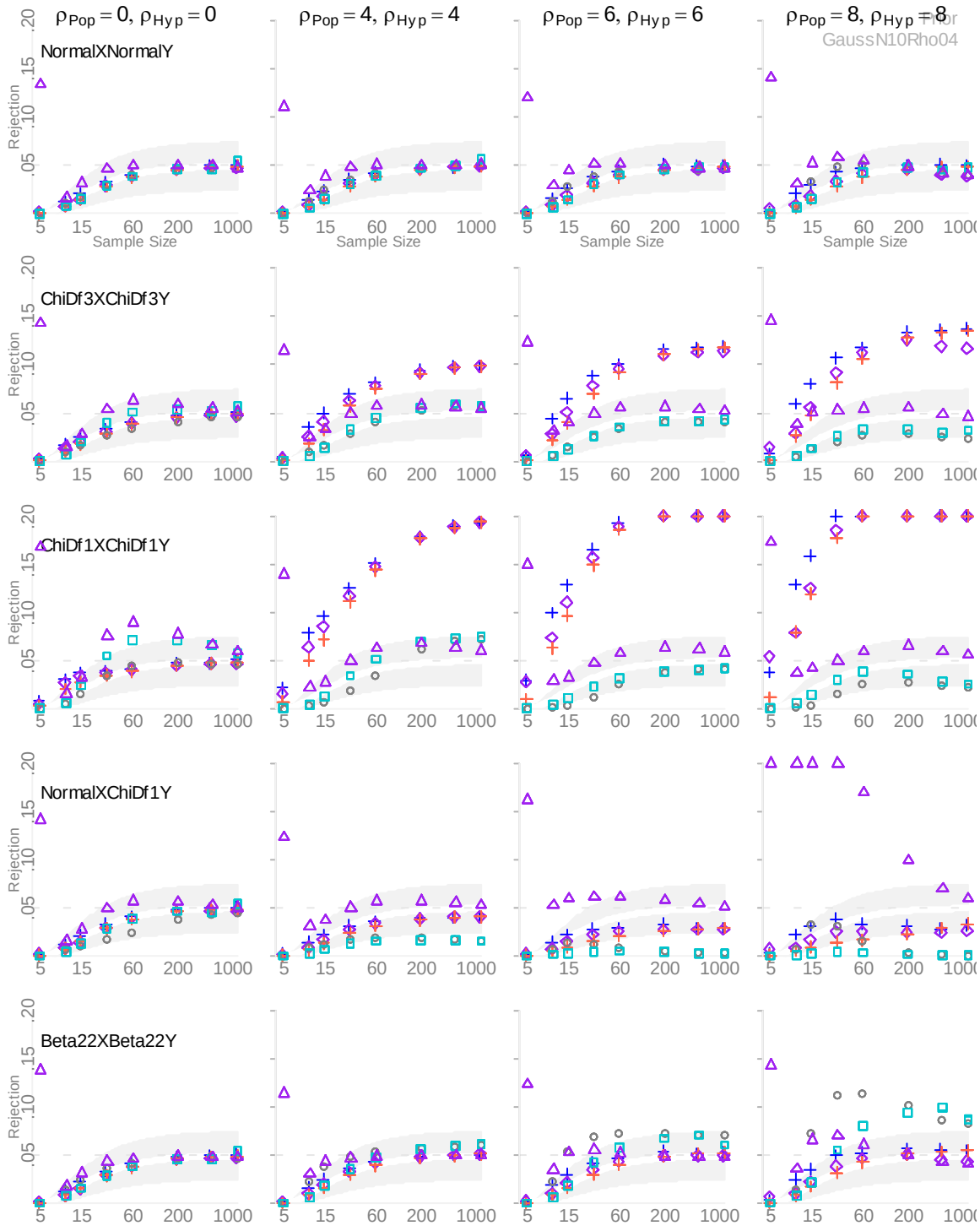
Among the univariate sampling bootstraps, there were two panels where γ exceeded the curved gray bands. With an uncorrelated Chi1*X*Chi1*Y*, γ_{Obs} was .055 and .071 (instead of .053 and .064). A larger violation occurred with a Beta22*X*Beta22*Y* when $\rho_{Pop} = .8$. The Type P error of the SlotOI (gray circle in Figure 10) jumped to a much higher rate with the nonuniform prior. It peaked at .11 when N_{Obs} was 30 and 60, and slowly declined to .083 when N_{Obs} grew to 1,000. This cell was also the weakest performance for the SlotHI, whose rates reached .100.

The error rates of the three parametric procedures were very similar with Gaussian priors as they were with a uniform prior. Error rates were very inflated in

heteroscedastic populations with nonzero correlations. Otherwise, their control of γ was very good, and arguably even better than with a uniform prior.

The SlotBiv again was very liberal when the Normal*X*Chi1*Y* population was strongly correlated. In all populations at $N_{Obs} = 5$, the SlotBiv had much more trouble with the Gaussian priors than with the uniform prior. We don't have a good explanation for this; perhaps it's related to its unsmooth bootstrap distribution with small samples.

Figure 10. Type P error with Gaussian prior of $\rho_{Prior} = .4$. The columns represent population correlations, while the rows are the population distribution. See the Figure 9 legend.



Power

Statistical power is the probability of correctly rejecting ρ_{Pop}. In our opinion, there was no procedure that consistently showed stronger or weaker power than others. In all conditions, power was virtually 1 when $N_{Obs} \geq 200$; in some panels it reached 1 by $N_{Obs} = 15$. Figures 11 and 12 display the rates of rejecting ρ_{Hyp}. Each row represents a different ρ_{Hyp} value, and each column is a different ρ_{Pop}. The diagonal panels (i.e., when $\rho_{Hyp} = \rho_{Pop}$) contain Type I error rates; gray bands cover $.025 < \alpha < .075$. The off-diagonal panels (i.e., when $\rho_{Hyp} \neq \rho_{Pop}$) contain power rates; gray bands cover rates above .95. The diagonal cells in Figure 11 are scaled re-expressions of Figure 9's first row (the vertical axis now extends from 0 to 1.0). The diagonal cells in Figure 12 re-express Figure 9's fourth row.

Figure 11 contains Normal*X*Normal*Y* rejection rates. When $N_{Obs} \geq 10$ and $\rho_{Hyp} = 0$, the procedures perform very similarly; the analytic (red plus) has barely more power than the SlotHI (turquoise square). Dropping to the second row when $\rho_{Hyp} = .4$, the SlotHI has more power than the analytic when $\rho_{Pop} = 0$, but the analytic is stronger when $\rho_{Pop} = .8$. This pattern continues for the third row, where the SlotHI is again more powerful than the analytic when the population correlation is smaller than the hypothesized correlation.

Figure 12 contains Normal*X*Chi1*Y* rejection rates, which follow the previous pattern. The analytic is stronger than the SlotHI when $\rho_{Hyp} > \rho_{Pop}$, but then the SlotHI is more powerful when $\rho_{Hyp} < \rho_{Pop}$.

In the conditions where Type P and Type I error rates were low (such as with the nonlinear relationships in Normal*X*Chi1*Y* with $\rho \geq .6$), the power was not substantially

lower than in Normal*X*Normal*Y* (in Figures 11 and 12, compare the cells in the 3rd column). Incidentally, the other procedures had substantially higher power than in the equivalent Normal*X*Normal*Y* cells (although SlotBiv's power comes at the expense of liberal error rates). We examined the likelihood distributions from the different procedures. When $\rho_{Pop} = .8$ in the Normal*X*Chi1*Y* population, it appears the SlotHI consistently keeps the CI's left boundary too low.

Figure 11. Rejection rates with a uniform prior and a Normal*X*Normal*Y* population. Columns represent different population correlations, while rows are the tested correlations. The diagonal panels are Type I error rates; a gray band indicates .025 < α < .075. The off-diagonal panels are power rates; a gray band indicates rates above .95. See the Figure 9 legend.

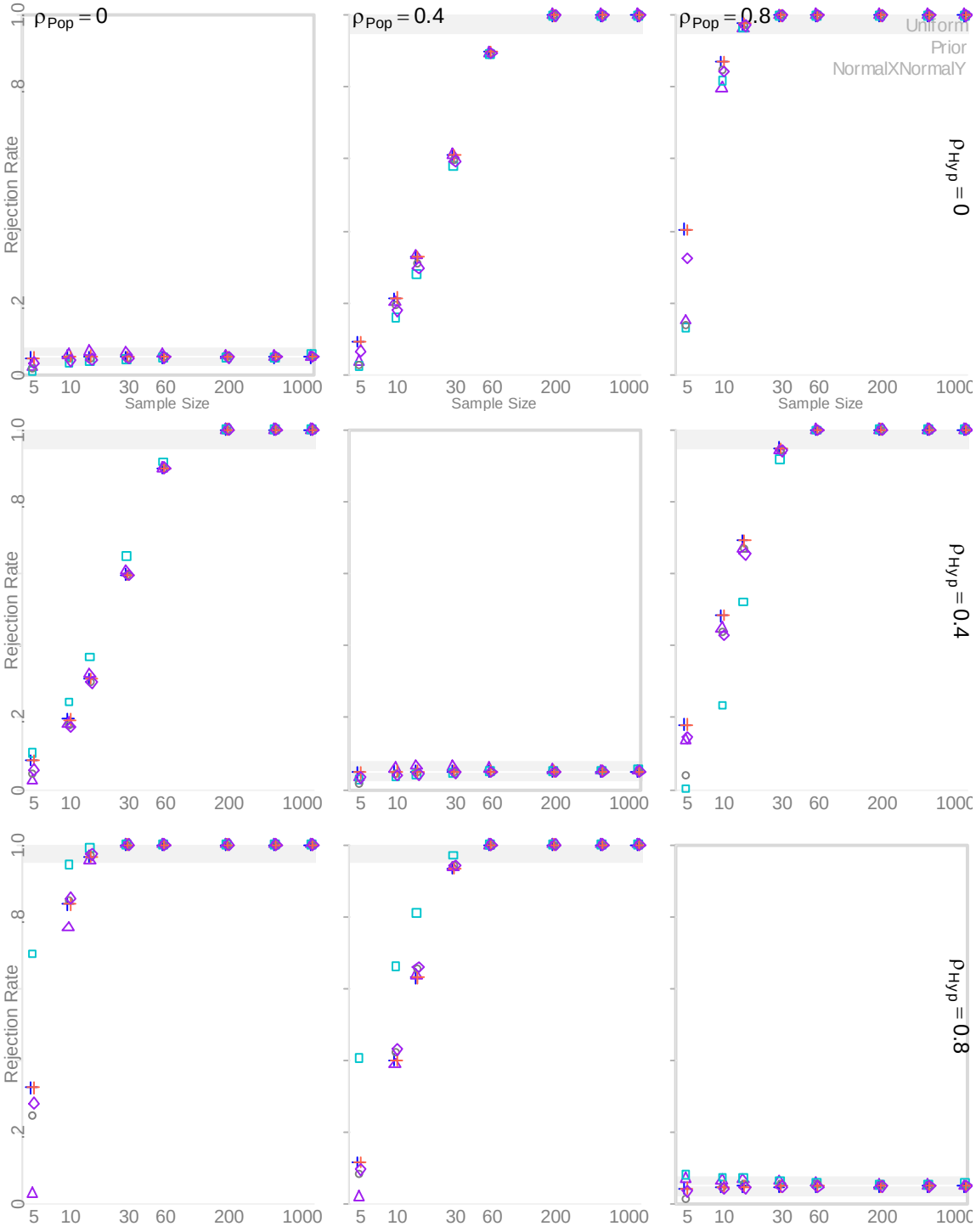
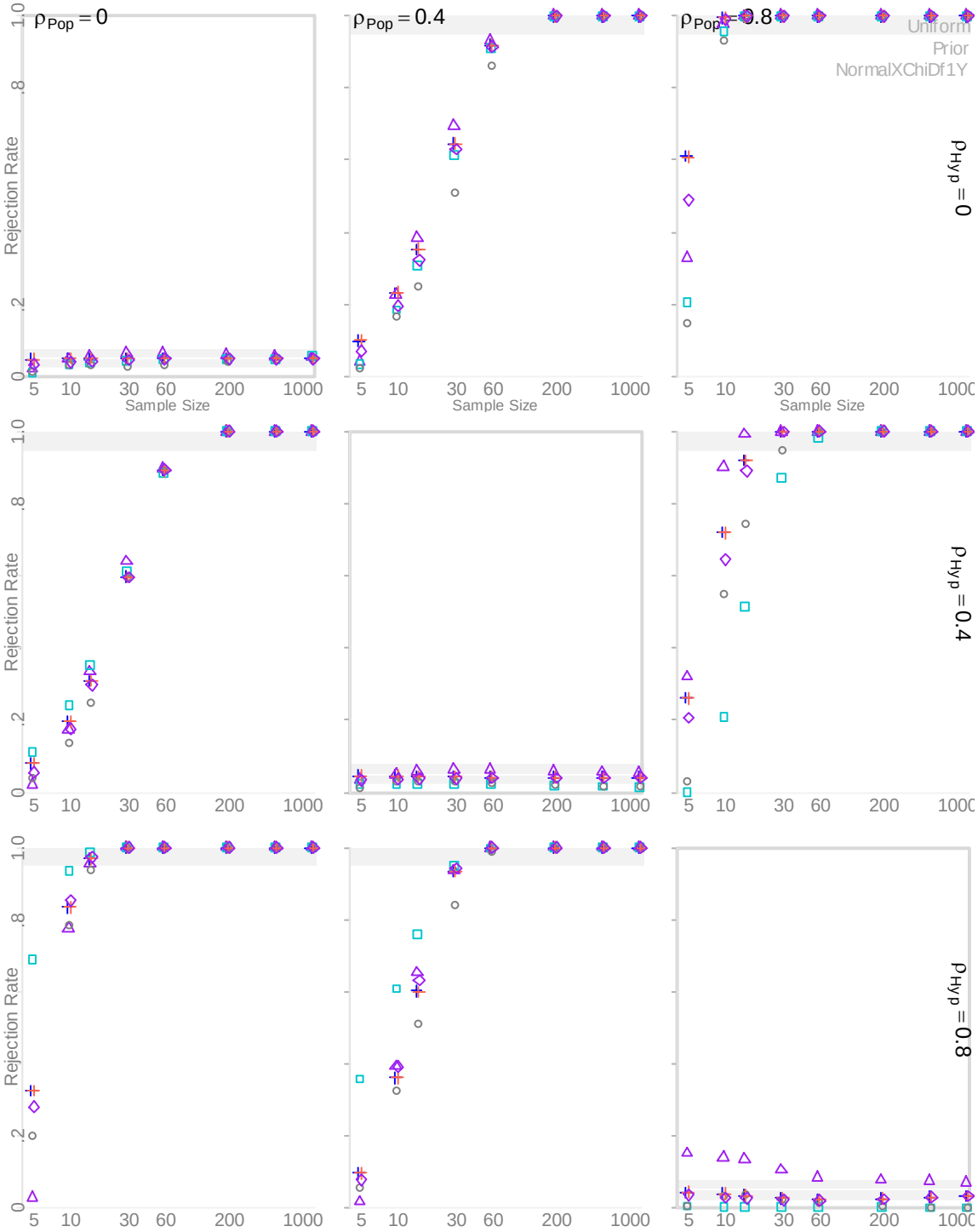


Figure 12. Rejection rates with a uniform prior and a Normal*X*Chi1*Y* population. See the Figure 11 caption.



Bias

The bias results are more difficult to summarize than the error and power results. In our opinion, no procedures were remarkably worse than others. From our perspective, no procedure was consistently negatively or consistently positively biased, and all quickly approached zero as N_{Obs} increased.

With a uniform prior, the reported procedures showed very little bias when $\rho_{Pop} = 0$. The *r* statistic itself is biased when ρ_{Pop} is nonzero (Olkin & Pratt, 1958), and this is evident in the last two columns of Figure 13 (although there are a few exceptions in the nonlinear population).

Surprisingly, poor Type I and P performance did not usually correspond to large biases. For example, the worst Type I error was seen when the parametric procedures tested samples from the Chi1*X*Chi1*Y* population. However in the same condition, their bias was remarkably good and outperformed the bootstrap procedures. Similarly, although the SlotHI showed its worst control of α in the strongly correlated Beta22*X*Beta22*Y* population, its bias in that population was comparable to the parametric procedures.

MSE

Despite differences in error, power and bias, the procedures had very remarkably similar MSEs, regardless of the prior. Results with a Gauss04 prior are shown in Figure 15.

Figure 13. Bias with a uniform prior.

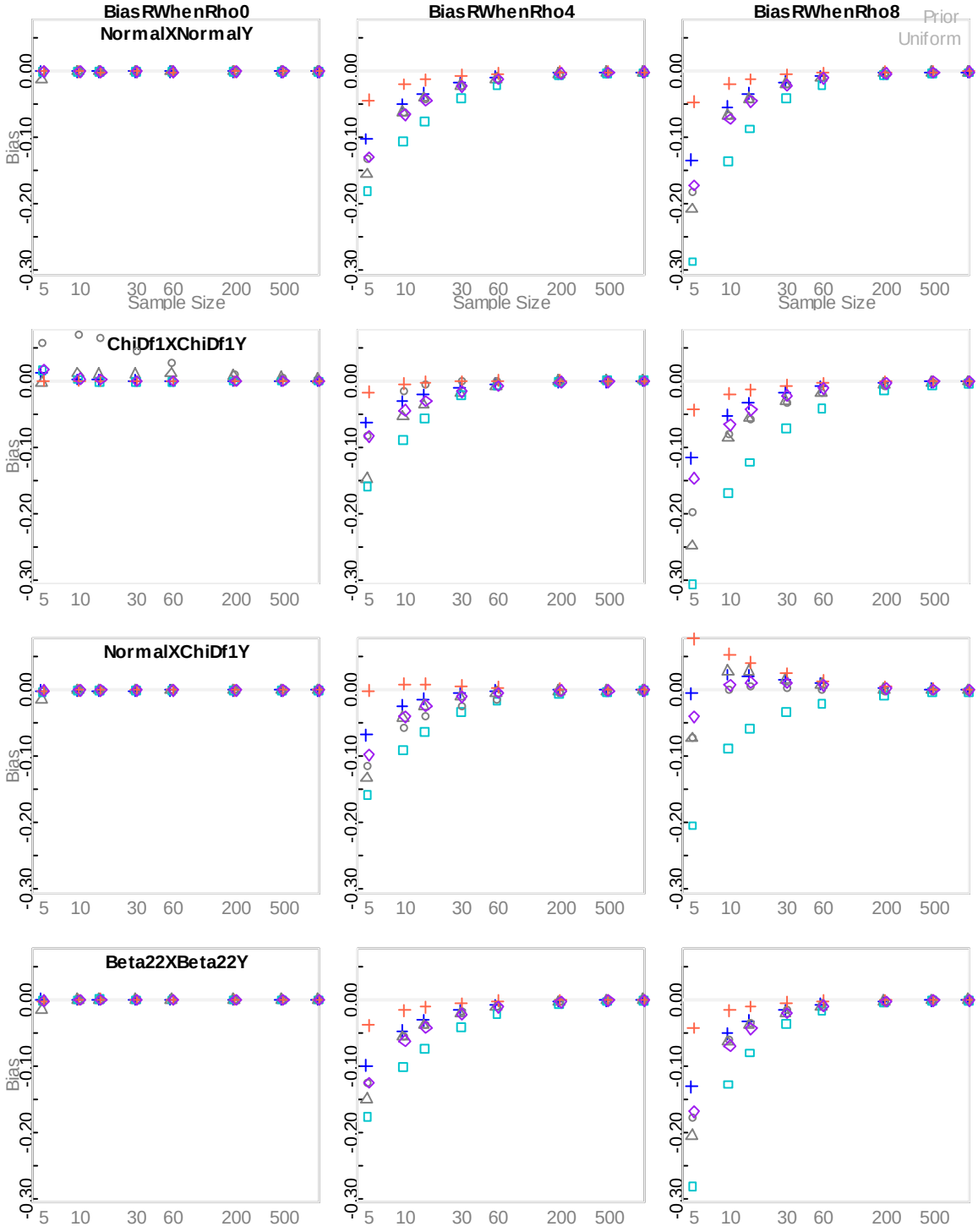


Figure 14. Bias with a Gauss04 prior.

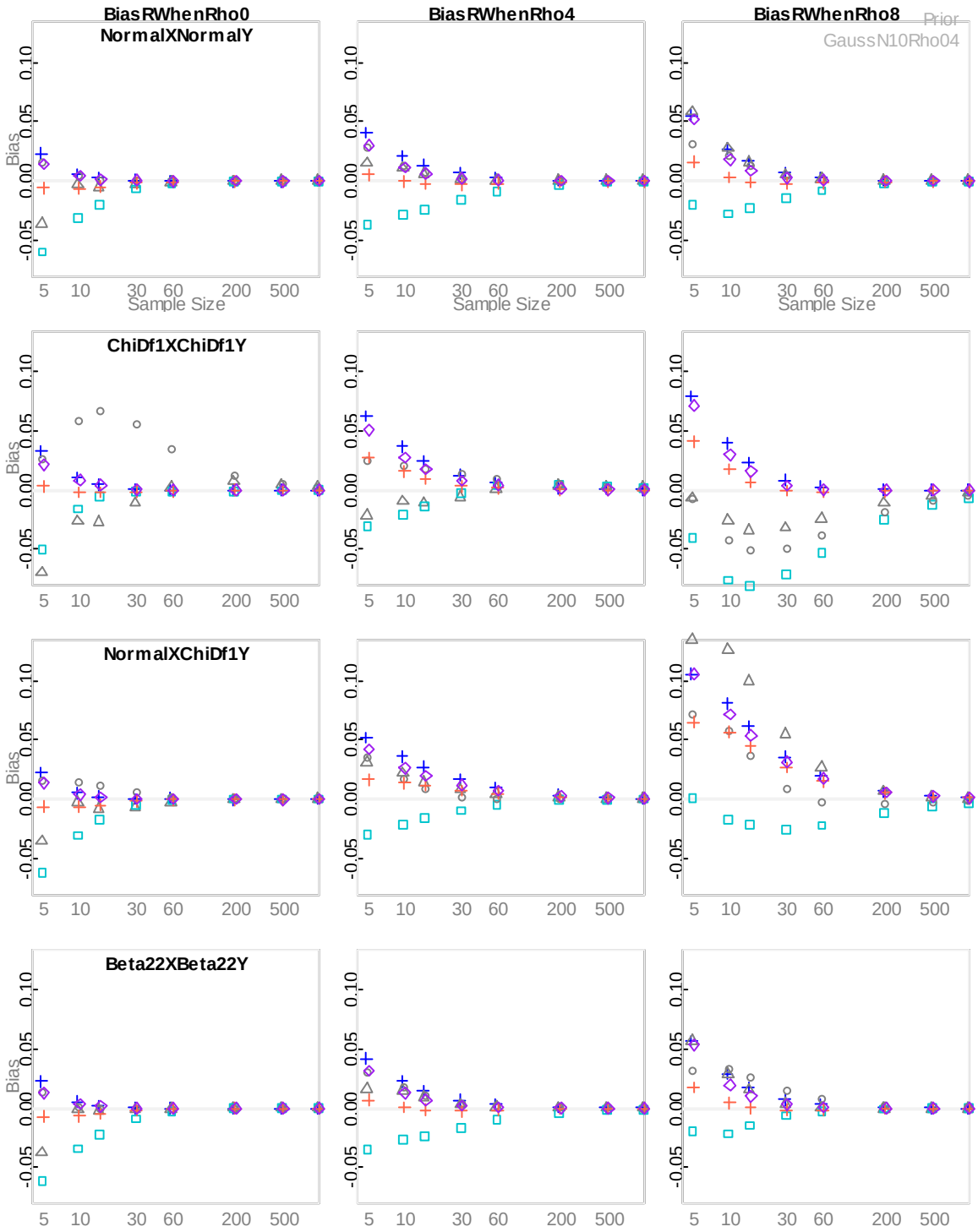
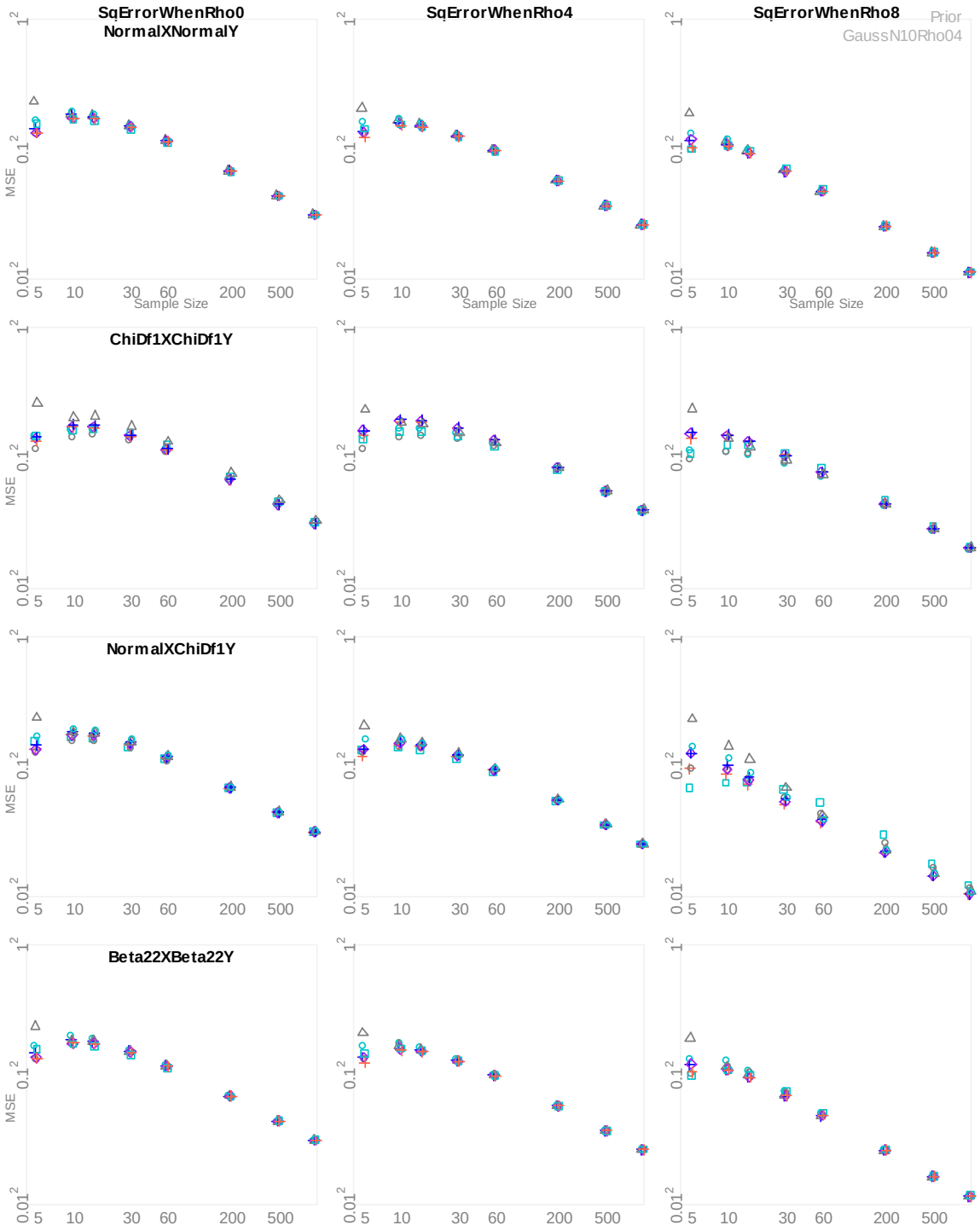


Figure 15. MSE with a Gauss04 prior.



Discussion

Performance: Power and Error Rates

Our impressions are similar to Beasley et al. (2007): the analytic procedure (i.e., the closed-form parametric procedure) is recommended if $\rho_{Hyp} = 0$ or if bivariate normality is assured. Otherwise, the SlotHI and SlotOI should be strongly considered because they are more robust to nonnormality, especially heteroscedasticity. When the parametric procedures tested nonzero correlations in heteroscedastic populations, their inferences were very liberal, with Type I and Type P error rates consistently exceeding .15. Unfortunately, this problem always worsened with larger sample sizes, and even appeared with correlations as low as $\rho_{Pop} = .4$. Skewed distributions like the Chi1 (skew = 2.8) and Chi3 (skew = 1.6) are not uncommon. As of 1989, Micceri estimated that 11% of studies used distributions where the skew exceeded 2.0.

In the conditions where we do recommend the analytic procedure, its advantage is primarily related to convenience and software availability, and not related to power. As seen in the top row of Figures 11 and 12, the power of the SlotHI (turquoise box) was competitive with the analytic (red plus) when evaluating $\rho_{Hyp} = 0$ in normal populations. This surprised us; we expected the parametric procedures to perform better than the bootstraps when the parametric assumptions were true. When rejecting nonzero values of ρ_{Hyp}, the power of the SlotHI won three out of the four cells (see the four off-diagonal cells in the last two rows of Figures 11 and 12).

We also were surprised to see the SlotHI outperform the HI[14] noticeably, even when the SlotHI used a uniform prior. We currently do not have a thorough explanation. Procedurally, the only difference is the SlotHI considered 200 different sampling frames to estimate the likelihood distribution, while the Frequentist HI considered only one. In other words, one SlotHI sampling frame is responsible for a distance of .01 (e.g., $.34 < \rho \leq .35$), while the single HI sampling frame covers a distance of 2 (i.e., $-1 \leq \rho \leq 1$).

The SlotHI and SlotOI typically had comparable power and error rates, and our decision to focus on the SlotHI was almost arbitrary. Their control of Type I and Type P error was roughly equivalent, while the SlotHI arguably had better power. They had one weak area in our simulation: Beta22*X*Beta22*Y*, when $\rho_{Pop} = .8$. Type P error of the SlotOI reached ~.13, while the SlotHI reached ~.11. When $\rho_{Pop} \leq .6$ and $N_{Obs} \geq 10$, the Type I error for both procedures never exceeded .083.

The SlotBiv was very liberal with the Normal*X*Chi1*Y* population when $\rho_{Pop} = 0.8$ in small and moderate sample sizes ($\gamma_{Obs} \sim .30$). However its liberalness fell to acceptable levels when $N_{Obs} \geq 500$. In the other populations, its performance was liberal, but acceptable. Its Type I error frequently was between .05 and .075.

Performance: Bias and MSE

The parametric procedures' point estimates apparently are much more robust than their CIs. Regarding bias, they typically outperformed bootstrap procedures, even in the heteroscedastic correlated populations (Figures 13 and 14).

Regarding MSE, the procedures were almost indistinguishable (Figure 15). Other conditions, such as sample size and prior distribution, were much more influential than

[14] The non-Bayesian version that is described in the 'Building Likelihood Distributions' section and in Beasley et al., 2007.

the choice of procedure. We are very comfortable recommending parametric procedures when the researcher requires point estimates.

Extensions: Sampling Frame

The current versions of the HI and OI do not treat the *X* and *Y* variables interchangeably, because the Cholesky imposes the correlation on only the *Y* variable. We think it is a small issue for bivariate conditions, because the difference in performance was never substantial in the conditions we tested (e.g., the results for the Normal*X*Chi1*Y* and Chi1*X*Normal*Y* populations were similar). However, in applied research the decision of designating the *X* and *Y* variable is often arbitrary, and we are concerned that this arbitrary decision could be more influential when generalized to multivariate relationships. If there are *q* parameters, there are (*q* - 1)! arbitrary decisions.

Although the *X* marginal is reproduced in the diagonalized sampling frame (Figure 1, Stage 2c), the *Y* marginal is altered (between Stages 2b and 2c; technically, the *Y''* marginal is altered). As the imposed correlation grows stronger, the *Y* marginal becomes more similar to the *X* marginal. When the imposed correlation is 1, then *Y'' = X'*, so the *Y''* marginal is identical to the *X'* values in the observed sample.

Unlike Cholesky decomposition, spectral decomposition treats the variables interchangeably. However, spectral decomposition does not reproduce the marginal of either *X* or *Y*. Unfortunately, the spectral's performance was disappointing in a previous study (Beasley et al., 2007, Appendix). The SlotPos didn't reproduce the sample's marginals (and in a sense, neither do the SlotParametric and analytic procedures); their failures with the heteroscedastic population might suggest that this feature is important.

However, spectral decomposition is the only method we are aware of that treats variables interchangeably when imposing a *linear* relationship.

A potential solution is to impose a *nonlinear* relationship in the sampling frame, while simultaneously respecting the observed marginals. Methods by Headrick (2010) and Ruscio and Kaczetow (2008) are two good candidates for future versions of the SlotHI and SlotOI. For reference, look at the Normal*X*Chi1*Y* row in Figure 8. When the points are forced to have normal and skewed marginals, while simultaneously exhibiting the imposed linear correlation (i.e., the dashed line), the curvilinear relationship is produced. If a sample can reliably estimate the correlation and marginals, the resulting bootstrap inference may be reliable in multivariate settings.

If a multivariate extension of the SlotHI or SlotOI becomes as robust as the bivariate version, we hope a new class of large-sample research scenarios can be served. In smaller datasets, robustness is important because little is known about the population. Sampling variability not only affects the correlation inference, but also obscures the validity of the parametric assumptions. If a researcher is misled about the population and chooses a poor transformation, the inference could be more misleading than if the scores had not been altered. Choosing a bad transformation is less likely in large datasets. The marginal and joint distributions are better defined, so the researcher has better information when selecting an appropriate transformation.

Applications like mediation analyses could benefit as well. Suppose there is heteroscedasticity in the linear relationship between *X* and *Y*, but their relationships with *M* are well behaved. The analyst is reluctant to transform *X* and *Y* (with a log or square root for example), because their relationship with *M* would become nonlinear. Of course

if a linear model becomes inappropriate, there are more advanced procedures that address this new issue, such as a nonlinear regression. Alternatively, a generalized linear model (with something other than an identity link) can be considered if the variables are not transformed.

However, options like these can introduce additional assumptions and require larger samples for estimations that are equally reliable. If the multivariate SlotHI or SlotOI is robust to heteroscedasticity, the analyst can retain the linear model and identity link, and won't have to increase the model's complexity. Furthermore, there is typically some prior knowledge about the total effect and the direct effect (commonly labeled *c* and *c*'), so the Bayesian capabilities of these procedures could assist mediation analyses too (Yuan & MacKinnon, 2009).

Extensions: Adaptive Slot Widths

The widths of the slots do not have to be equal. If either the prior or likelihood distribution is sharply peaked (which will happen if N_{Obs} is very large or the prior information is very specific), the inference's accuracy may benefit if smaller intervals are concentrated near the peaks. In the situations we explored, 200 slots seemed adequate when $N_{Obs} \leq 200$. However when the sample was larger, the SlotHI's pattern of error rates became less stable –the values showed an unsmooth pattern that alternated between too liberal and too conservative. As a result, we increased *S* from 200 to 400 and *B* from 1,999 to 4,999 (recall the other bootstraps used $B = 9{,}999$) and the error rates stabilized. The power rates were unaffected, because they were already near the asymptote of 1. Bias and MSE were unaffected as well.

The instability arises because the entire posterior distribution is contained in a small number of slots. For example, with N_{Obs} = 1,000, ρ_{True} = .8, and a Beta22*X*Beta22*Y* population, the posterior typically is entirely contained in (.77, .85]. It is difficult to obtain a reliable resolution of 2.5% when 100% of the distribution is contained in only 8 of the 200 slots. The posterior values in the other slots are zero, so in a sense, 96% of the SlotHI's computation was wasted in this situation. However, the practical importance of this issue may be small. Bias and MSE appear to be stable, and the width of the CI is very small when N_{Obs} = 1,000 –it can easily discriminate a ρ_{Hyp} of .75 from a ρ_{Pop} of .80.

The previously described interpolation technique alleviated much of the Type I instability. But this linear approximation might be insufficient when the distribution is contained in fewer than 8 slots. The problem cannot be solved by only increasing *B*. If it's not feasible to increase the number of slots, or adapt the width of the slots, it could be beneficial to smooth the posterior points and allow the neighbors to inform each other. Boos and Monahan (1986) used kernel smoothing in this situation; a variation of Romberg integration might be another alternative.

Extensions: Regression

A correlation centers the variables to have a mean of zero and standardizes the slope based on the standard deviations. In this sense, a correlation ignores information that is captured by a regression. Despite our appeals to incorporate more information, we think these correlation procedures have three uses. First, in the early stages of research, an expert's prior knowledge may be more accurately expressed as a correlation than an unstandardized regression slope. Second, there may be occasions where different means and variances should be ignored. Third, this procedure is a good starting point;

and SlotHI and SlotOI regression procedures probably can be generalized and evaluated in later research.

The procedural extension might be as simple as adding a Stage 2d that unstandardizes the *X* and *Y* variables in the diagonalized sampling frame and restores the variable's mean and variance. For instance, the diagonalized *X* and *Y* sampling frame values are first multiplied by sd(*X*) and sd(*Y*) and then added to mean(*X*) and mean(*Y*), respectively. There might need to be some small adjustment because sd(*X*) was calculated with N_{Obs} scores, but the sampling frame has N_{Obs}^2 scores. Later, a 2-parameter bootstrap distribution would be created as the slope and intercept are calculated for each bootstrap sample.

Conclusions

Although psychology has fewer than some other disciplines, we do have many accepted population point estimates, such as the American mean IQ is very close to 100 (Lichtenberger & Kaufman, 2009), the comorbidity of depression and substance abuse is roughly 25% (Kessler, et al., 2003), and the correlation between social desirability and the first component of personality scales is at least .8 (see for a list of these studies, Backstrom, Bjorklund & Larsson, 2009, p. 335). If it is difficult to think of an example in a field, it does not necessarily mean that non-nil testing won't be beneficial. It could mean that previous research hasn't been motivated to form a consensus about a value, but only about a direction. "But if all we, as psychologists, learn from research is that A is larger than B ($p < .01$), we have not learned very much. And this is typically all we learn." (Cohen, 1994).

Regardless if an informative prior is considered, the SlotHI and SlotOI produced CIs that were more robust than the parametric procedures when ρ is nonzero. Their Type I error rates were usually under .06, and they never exceeded .083 when $-.6 \leq \rho \leq .6$.

In the situations that the SlotHI or the analytic procedure is recommended, we are not advocating that all other procedures should be avoided. We do not think a *p*-value from a single procedure can adequately describe a nontrivial research question. A better understanding is much more likely when multiple procedures are considered (as well as multiple models, graphs, and priors).

Not only can a study use multiple procedures, but it can consider multiple priors as well. A reference prior can be reported to represent the researcher's best guess of the population correlation if no other research is available. But usually some subjective or objective information is available, and in these cases, a field might advance more quickly if both a reference and informative priors are considered and reported.

References

Albert, J. (2009). *Bayesian computation with R* (2nd ed.). Springer: New York.

Bäckström, M., Björklund, F., & Larsson, M. R. (2008). Five-factor inventories have a major general factor related to social desirabilitywhich can be reduced by framing items neutrally. *Journal of Research in Personality, 43,* 335–344.

Beasley, W. H., DeShea, L., Toothaker, L. E., Mendoza, J. L., Bard, D. E., & Rodgers, J. L. (2007). Bootstrapping to test for nonzero population correlation coefficients using univariate sampling. *Psychological Methods, 12*(4), 414-433.

Beasley, W. H., & Rodgers, J. L. (2009). Resampling methods. In R. E. Millsap & A. Maydeu-Olivares (Eds.) *Quantitative methods in psychology* (pp. 362-403). Thousand Oaks, CA: Sage.

Boos, D. D., & Monahan, J. F. (1986). Bootstrap methods using prior information. *Biometrika, 73*(1), 77-83.

Carlin, B. P., & Louis, T. A. (2009). *Bayesian methods for data analysis* (3rd ed.). Chapman & Hall/CRC: Boca Raton.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*(12), 997–1003.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7*(1)*,* 1-26.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans.* Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association, 82*(397)*,* 171-185.

Efron, B., & Tibshirani, R.J. (1993). *An introduction to the bootstrap.* Chapman & Hall/CRC: Boca Raton.

Fisher, R. A. (1915) Frequency distribution of the values of the correlations coefficient in samples from an indefinitely large population. *Biometrika, 10*(4)*,* 507-521.

Fisher, R. A. (1919). The genesis of twins. *Genetics, 4,* 489-499.

Gelman, A., Carlin, J B., Stern, H. S. & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Chapman & Hall/CRC: Boca Raton.

Gill, J. (2008). *Bayesian methods* (2nd ed.). Chapman & Hall/CRC: Boca Raton.

Headrick, T. C. (2010). *Statistical simulation: Power method polynomials and other transformations*. Chapman & Hall: Boca Raton.

Hays, W. L. (1994). *Statistics*. Belmont, CA: Wadsworth.

Howard, G. S., Maxwell, S. E., & Fleming, K. J. (2000). The proof is in the pudding: an illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. *Psychological Methods, 5*(3), 315-332.

Intel Corporation (2010). *Math Kernel Library: Vector Statistical Library notes*. Retrieved March 15, 2010, from http://software.intel.com/sites/products/documentation/hpc/mkl/vsl/vslnotes.pdf.

Jeffreys, H. (1939). *Theory of Probability*. Oxford: University Press.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K. R., Rush, A. J., Walters, E. E., & Wang, P. S. (2003). The epidemiology of major depressive disorder: Results from the National Comorbidity Survey Replication (NCS-R). *The Journal of the American Medical Association, 289*(23), 3095-3105.

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association, 44*(448), 1372-1381.

Lee, W., & Rodgers, J. L. (1998). Bootstrapping correlation coefficients using univariate and bivariate sampling. *Psychological Methods, 3*(1)*,* 91-103.

Lichtenberger, E. O., & Kaufman, A. S. (2009). *Essential of WAIS-IV assessment.* Wiley: New Jersey.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1)*,* 156-166.

Netlib (2010). *dxapy*. Retrieved March 11, 2010, from http://www.netlib.org/blas/daxpy.f.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (2003). *Numerical recipes in C++* (2nd ed.) Cambridge, UK: Cambridge University Press.

Pruzek, R. M. (1999). Significance testing introduction and overview. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (Ch. 11). Mahwah, NJ: Erlbaum.

Raudenbush, S. W. & Bryk, A. S. (2004). *Hierarchical linear models: Application and data analysis methods*. Thousand Oaks, CA: Sage.

Rubin, D. B. (1981). The Bayesian bootstrap. *Annals of Statistics, 9*(1), 130-134.

Ruscio, J., & Kaczetow, W. (2008). Simulating multivariate nonnormal data using an iterative algorithm. *Multivariate Behavioral Research, 43*(3), 355-381.

Toub, S. & Farkas, S, (2007, September). Tales from the CryptoRandom. *MSDN Magazine, 22*(9), 125-130.

Tucker, A. (1984). *Applied combinatorics*. New York: Wiley.

Yuan, Y., & MacKinnon, D. P. (2009). Bayesian Mediation Analysis. *Psychological Methods, 14*(4), 301-322.

Appendix

Computational Optimizations

Diagonalization Shortcuts. The Cholesky decomposition can be expressed as $y_i' = x_i \times \rho + y_i\sqrt{1-\rho^2} = \sqrt{1-\rho^2}\left(x_i \times \left(\rho/\sqrt{1-\rho^2}\right) + y_i\right)$. The ρ is constant for the sampling frame of the HI and OI, so $\rho/\sqrt{1-\rho^2}$ can be calculated once and saved as a fixed slope, b. This avoids repeating the expensive division and square root operations. Furthermore, the scaling factor $\sqrt{1-\rho^2}$ outside the parentheses can be ignored because the correlation coefficient is invariant to linear transformations of Y. The reduced equation ($y_i' = bx_i + y_i$) is much cheaper to computer, and even could be completed in one operation if the hardware supports a fused multiply-accumulate ('FMA3'; also see the 'daxpy' routine in BLAS, Netlib, 2010).

Reusing Sum$X$ and (N_{Obs}·SumX^2 - (Sum$X$)^2). When S = 200, the SlotHI must account for 200 different sampling frames and bootstrap distributions, while the SlotOI, SlotBiv and SlotPar need to account for only 1. However, the SlotHI doesn't have to take 200 times longer. When a Cholesky decomposition diagonalizes the sampling frame, the X values are not affected and therefore can be reused. When the correlation is expressed as

$$r = \frac{N\cdot\text{Sum}XY - \text{Sum}X\cdot\text{Sum}Y}{\sqrt{(N\cdot\text{Sum}X^2 - (\text{Sum}X)^2)(N\cdot\text{Sum}Y^2 - (\text{Sum}Y)^2)}},$$

only the three terms involving Y (i.e., Sum$Y$, SumY^2, and SumXY) need to be calculated more than once. (For the appendix, the term N is used instead of N_{Obs}).

The same strategy was applied to the jackknives in the BC_a and BC_{as}. The five sums (i.e., Sum$X$, SumX^2, Sum$Y$, SumY^2, SumXY) were calculated once for each sample

of N scores, and then reused for the N jackknife samples. When the i^{th} score was excluded, the i^{th} jackknifed statistic is calculated as

$$r_{-i} = \frac{\{N-1\}\{\mathrm{Sum}XY - x_i y_i\} - \{\mathrm{Sum}X - x_i\}\{\mathrm{Sum}Y - y_i\}}{\sqrt{\left(\{N-1\}\left\{\mathrm{Sum}X^2 - x_i^2\right\} - \{\mathrm{Sum}X - x_i\}^2\right)\left(\{N-1\}\left\{\mathrm{Sum}Y^2 - y_i^2\right\} - \{\mathrm{Sum}Y - y_i\}^2\right)}}.$$

Without this shortcut, the N jackknifed sums of XY require $N \times (N-1)$ multiplications and $N \times (N-1) - 1$ additions for the bivariate sampling bootstrap; with this shortcut, this term requires $2N$ multiplications and $2N - 1$ additions. The discrepancy grows even larger for a univariate sampling bootstrap, because it uses N^2 jackknife samples of N^2 - 1 scores. When the sample contains 1,000 observations, the shorter routine uses roughly 1 arithmetic operation for every 500,000 operations used by the less efficient routine. Also, we expect the shortcut produces fewer misses in the CPU's lower cache levels, which further shorten the routine's duration.

BC_a and BC_{as}

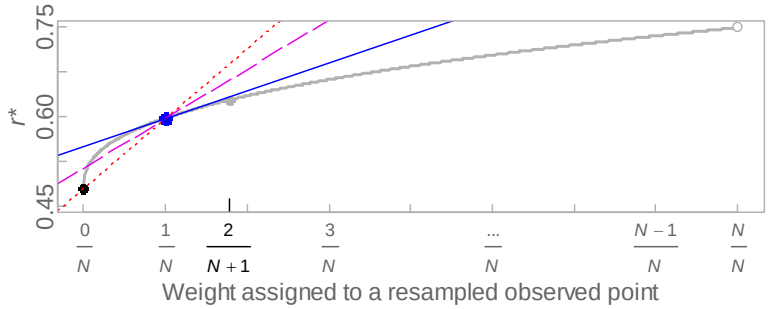
Two types of jackknifes are relevant to the BC_{as}. The typical jackknife excludes one observation and calculates the plug-in statistic on the samples remaining $N - 1$ observations. This repeats for every observation, producing a jackknife distribution of N resampled statistics. For this paper, the correlation coefficient is the plug-in statistic. The *positive* jackknife is similar to the typical jackknife, but it *duplicates* the i^{th} point and calculates the statistic on the $N + 1$ points.

The acceleration term in the BC_a, a, estimates the rate of change of the standard error of ρ, as ρ increases[15]. To understand acceleration, it helps to think of each observed

[15] This is described in depth in Efron & Tibshirani (1993, Ch. 22). In short, the typical BC_a uses a jackknife to estimate the direction of a line. This line, which is orthogonal to level curves in an N-dimensional space, reduces the space to the one-parameter 'least favorable family' in order to estimate a.

point having a resampling weight in a bootstrap sample; the weight is a proportion ranging from 0 to almost 1. There is a relationship between the resampled statistic and the resampled weight for the i^{th} observed point. A weight of zero means the point wasn't included in a bootstrap sample and the remaining N - 1 points have a weight of N / (N - 1) when calculating r^*. In contrast, a weight close to 1 means the point was repeatedly sampled almost N times and the remaining points have a weight close to 0. Figure 16 displays a hypothetical relationship between the resampling weight of the i^{th} point, and the resulting value of r^* (see also Efron & Tibshirani, 1993, Figure 20.6). The blue dot is located at (x, y) = ($1/N$, r_{obs}), which represents the observed sample. For reasons described in Efron & Tibshirani, Section 22.5, it is important to find the slope of the function for each observation. The slope for the i^{th} point, shown with a blue line in Figure 16, is called the empirical influence component, U_i, and can be difficult to calculate analytically with a nonparametric bootstrap.

Figure 16. Illustration of different acceleration estimates. The gray line is the value of r^* as the resampling weight of an observation is increased from 0 to the limit of 1.



When analytically calculating the slope is difficult, Efron & Tibshirani (1993, Equation 14.15 and p. 290) suggest approximating this slope with $U_i = (N-1)(r_{(\cdot)} - r_{(-i)})$, where $r_{(-i)}$ is calculated from a sample that excludes the i^{th} point (otherwise known as the

i^{th} jackknifed statistic). The mean of the N values of $r_{(-i)}$ is $r_{(-\cdot)}$, which should be close to r_{Obs}. When the mean is the jackknifed statistic, instead of the correlation, $r_{(-\cdot)} = r_{Obs}$. The estimated acceleration equation is

$$\hat{a} = \frac{\sum_{i=1}^{N} U_i^3}{6\left(\sum_{i=1}^{N} U_i^2\right)^{3/2}} \quad (5),$$

and there are different equations for U_i, as we will discuss.

This approximated slope using the jackknife is shown with the red dotted line in Figure 16. This approach is sometimes called a 2-point formula, as it passes through the jackknifed statistic and the observed statistic (shown as the black dot and blue dot). A 2-point formula can be expressed as

$$\text{slope} = [f(x_0) - f(x_0 - h)] / [x_0 - (x_0 - h)] = [f(x_0) - f(x_0 - h)] / h \quad (6),$$

where f is the function of interest, x_0 is the tangent point, and h is the horizontal distance between the two points. In our case, x_0 is $1/N$ and $x_0 - h$ is $0/N$; $f(x_0)$ is r_{Obs} (or $r_{(-\cdot)}$), and $f(x_0 - h)$ is approximately the jackknife estimate. The size, but not the sign, of any multiplicative factor that is constant for all N influence components (like h in the denominator of Equation 6) can be ignored when estimating acceleration, because it cancels itself in Equation 5. The commonly used formula for U_i in Equation 5 is

$$U_i = r_{(-\cdot)} - r_{(-i)} \quad (7).$$

Another 2-point formula finds the slope between the observed statistic and the *positive* jackknifed statistic (the blue dot and gray dot). The empirical influence component using the positive jackknife is $U_i = (N + 1)(r_{(+i)} - r_{Obs})$ (Efron & Tibshirani, 1993, Equation 20.22). Although it shouldn't make much practical difference, we don't understand why the positive jackknife's empirical influence component equation uses

r_{obs}, but the jackknife's equation uses $r_{(\cdot)}$. In this case, h is negative in Equation 6, because the second point is to the right of $1/N$:

$$U_i = (r_{Obs} - r_{(+i)}) / -1 = r_{(+i)} - r_{Obs} \quad (8).$$

The slope estimate is potentially more accurate when informed by points that straddle r_{Obs}, also known as a 3-point formula. The purple dashed line in Figure 16 uses three points to calculate its slope: the first uses the jackknife estimate $(x_0 - h_1, f(x_0 - h_1))$, the second uses r_{Obs} $(x_0, f(x_0))$, and the third uses the positive jackknife estimate $(x_0 + h_2, f(x_0 + h_2))$. When $h_1 = h_2 = h$, the 3-point formula reduces to $[f(x_0 + h) - f(x_0 - h)] / 2h$. Because the denominator is equal for all N points, it can be ignored and the influence component becomes

$$U_i = (r_{(+i)} - r_{Obs}) - (r_{Obs} - r_{(-i)}) = r_{(+i)} - r_{(-i)} \quad (9).$$

Equation 9 reflects at least three approximations that theoretically decrease accuracy. First, the positive jackknife isn't positioned exactly above $2/N$; because a resampled sample has $N + 1$ points, the weight is actually $2/(N + 1)$. Ideally, the weights of the remaining observations should be reduced by $(N - 1)/N$ so that $h_1 = h_2$. Second, the typical (or negative) jackknife uses samples of $N - 1$ points. If we understand the concepts correctly, weights of the remaining observations should be increased by $N/(N - 1)$, but apparently its developers believe this discrepancy is small enough that it can be ignored. Third, the x_0 value for Equation 7 is the mean of the jackknifed statistics, $r_{(-\cdot)}$, but x_0 for Equation 8 is r_{Obs} (instead of the mean of positive jackknifed statistics, $r_{(+\cdot)}$). We simplified the routine by replacing $r_{(-i)}$ with r_{Obs} in Equation 9, which removes r_{Obs} from the equation altogether. These three approximations may lead to less accurate values in ideal conditions, but we feel the resulting values won't be noticeably worse, and

might more numerically stable when calculated in nonideal conditions with finite precision.

The BC_a and Equation 5 were introduced generally, without being connected exclusively to the jackknife (Efron, 1987). Therefore the BC_{as} is not a new CI adjustment, but rather one more type of BC_a. The empirical difference between the two types slightly favored the BC_{as} in our simulations; but the difference was small enough that we recommend a practitioner use the typical BC_a if an existing implementation is easily available (and has been thoroughly tested). The differences in the two estimates of a were very small, and judging from the small difference between the BC and BC_a's performance, the influence of a on the BC_a was small to begin. The recommendation could change when a different statistic is bootstrapped. We would expect the BC_{as} to do better as the second derivative grows (at the blue point in Figure 16). See Efron & Tibshirani (1993, Section 20.6) for more discussion of the different advantages of the jackknife and positive jackknife.