

Online Appendix to “Partial Identification of Spread Parameters”

Jörg Stoye

New York University

September 12, 2005

This appendix contains closed-form missing-data bounds on some functionals of mean and variance. Technically, the results are corollaries of proposition 1. All notations are as in the main paper.

1 Worst-Case Bounds on the t-Ratio

Consider a researcher who observes a random sample from some population and wishes to test the null hypothesis $H_0 : E(X) = \theta$. The standard test statistic for this situation is the t-ratio $t \equiv \sqrt{N-1} \frac{E_N(X) - \theta}{\sqrt{V_N(X)}}$, where $E_N(X)$ and $\sqrt{V_N(X)}$ are the sample mean and standard deviation respectively. The statistic is compared to Student’s t-distribution if X is known to be normal and to a (limiting) normal distribution otherwise.

Imagine however that the sample suffered from potentially selective missing data. The observed t-ratio would then be misleading, but if the original sample was selected randomly, its t-ratio would be meaningful. Therefore, the researcher might be interested in the range of possible values of the full sample’s t-ratio, given her partial observation of the sample.

This problem does not involve population statistics, but is an application of theorem 1 because the full sample’s distribution can be written as

$$F_N = p_N F_{1N} + (1 - p_N) F_{0N},$$

where F_N is the sample analog of F etc. Of the above quantities, F_{1N} and p_N are observed, whereas F_{0N} and, by implication, F_N are not. Hence, theorem 1 delivers worst-case bounds on the full sample’s

expectation and variance, $E_N(X)$ and $V_N(X)$, and by implication the following bounds on t .

Corollary 1 *Bounds on the t-Ratio*

Consider the setup of proposition 1 but replace F by a partially observed sample distribution F_N . The t -ratio, $t_N \equiv \left(\frac{N-1}{V_N(X)}\right)^{\frac{1}{2}} (E_N(X) - \theta)$, is bounded by

$$\begin{aligned} & \left(\frac{N-1}{p_N E_{1N}(X^2) + (1-p_N)\alpha - (p_N E_{1N}(X) + (1-p_N)\alpha)^2} \right)^{1/2} (p_N E_{1N}(X) + (1-p_N)\alpha - \theta) \\ & \leq t_N \leq \\ & \left(\frac{N-1}{p_N E_{1N}(X^2) + (1-p_N)x^2 - (p_N E_{1N}(X) + (1-p_N)x)^2} \right)^{1/2} (p_N E_{1N}(X) + (1-p_N)x - \theta), \end{aligned}$$

where

$$\begin{aligned} \alpha &= \begin{cases} 0 & \text{if } \alpha^* < 0 \\ \alpha^* & \text{if } 0 \leq \alpha^* \leq 1 \\ 1 & \text{if } \alpha^* > 1 \end{cases} \\ \alpha^* &= \frac{2p_N E_{1N}(X^2) + \theta - p_N E_{1N}(X)(1+2\theta)}{(1-p_N)(2\theta-1)} \end{aligned}$$

and

$$\begin{aligned} x &= \begin{cases} 0 & \text{if } x^* < 0 \\ x^* & \text{if } 0 \leq x^* \leq 1 \\ 1 & \text{if } x^* > 1 \end{cases} \\ x^* &= \frac{E_{1N}(X^2) - E_{1N}(X)\theta}{E_{1N}(X) - \theta}. \end{aligned}$$

These bounds are sharp.

Proof. The proof begins by displaying general FOC's for extremizing the t-ratio. For notational convenience, drop the subscripts N . Let $P(X)$ be parameterized by some parameter, say i , and define the functions $\mu(i) \equiv E(X)$ and $V(i) \equiv V(X)$. Then extremization of the t-ratio amounts to solving the problem

$$\max / \min \frac{\mu(i) - \theta}{\sqrt{V(i)}} \tag{1}$$

$$s.t. (\mu(i), V(i)) \in H((\mu(i), V(i))), \tag{2}$$

where dropping $N^{1/2}$ is obviously inconsequential.

The objective function is strictly monotonic in $V(i)$ and therefore maximized [minimized] somewhere on the lower [upper] boundary of $H((\mu(i), V(i)))$. From the proof of theorem 1, I have descriptions of these boundaries that parameterize the upper boundary by $\alpha \in [0, 1]$ and the lower one by $x \in [0, 1]$. The problem can therefore be solved by replacing $H((\mu(i), V(i)))$ in [2] with the relevant boundary. To further simplify matters, I first assume an interior solution with respect to α respectively x and then consider the possibility that this fails.

The First-Order Condition for [1] is

$$\begin{aligned} \frac{1}{V(i)} \left[-\frac{1}{2}(\mu(i) - \theta)(V(i))^{-1/2}V'(i) + (V(i))^{1/2}\mu'(I) \right] &= 0 \\ \implies V'(i)(\mu(i) - \theta) &= 2V(i)\mu'(i). \end{aligned} \quad (3)$$

To solve the minimization problem, use the parameterization of the upper boundary of $H((\mu(i), V(i)))$ to write

$$\begin{aligned} \mu(\alpha) &= p\mu_1 + (1-p)\alpha \\ \mu'(\alpha) &= 1-p \\ V(\alpha) &= pE_1(X^2) + (1-p)\alpha - (\mu(\alpha))^2 \\ V'(\alpha) &= 1-p - 2\mu'(\alpha)\mu(\alpha) = (1-p)(1-2\mu(\alpha)). \end{aligned}$$

Substituting into the FOC and dividing through by $(1-p)$ gives

$$\begin{aligned} (\mu(\alpha) - \theta)(1 - 2\mu(\alpha)) &= 2[pE_1(X^2) + (1-p)\alpha - (\mu(\alpha))^2] \\ \implies (\mu(\alpha) - \theta) + 2\theta\mu(\alpha) &= 2[pE_1(X^2) + (1-p)\alpha] \\ \implies [p\mu_1 + (1-p)\alpha](1 + 2\theta) - \theta &= 2pE_1(X^2) + 2(1-p)\alpha, \end{aligned}$$

which is solved by

$$\alpha^* = \frac{2pE_1(X^2) + \theta - p\mu_1(1 + 2\theta)}{(1-p)[2\theta - 1]}.$$

It is easily seen that the objective function is concave in α , so that α^* indeed describes a minimum.

As before, if $\alpha^* \notin [0, 1]$, then the optimal α is the nearest feasible α .

To maximize the t-ratio, use the parameterization of the identification set's lower boundary to write

$$\begin{aligned}
\mu(x) &= p\mu_1 + (1-p)x \\
\mu'(x) &= 1-p \\
V(x) &= pE_1(X^2) + (1-p)x^2 - (\mu(x))^2 \\
V'(x) &= 2(1-p)(x - p\mu_1 - (1-p)x) = 2p(1-p)(x - \mu(x)).
\end{aligned}$$

Substituting into [3] and dividing by $2(1-p)$ yields

$$\begin{aligned}
(\mu(x) - \theta)(x - \mu(x)) &= pE_1(X^2) + (1-p)x^2 - (\mu(x))^2 \\
\implies \mu(x)(x + \theta) - \theta x &= pE_1(X^2) + (1-p)x^2 \\
\implies (p\mu_1 + (1-p)x)(x + \theta) - \theta x &= pE_1(X^2) + (1-p)x^2 \\
\implies p\mu_1(x + \theta) - p\theta x &= pE_1(X^2) \\
\implies x^* &= \frac{E_1(X^2) - \mu_1\theta}{\mu_1 - \theta}.
\end{aligned}$$

Here, observe cancellation of $-(\mu(x))^2$ in the first step, substitution for $\mu(x)$ in the second step, and cancellation of $(1-p)x^2$ as well as cancellation of θx within the LHS in the third one.

It remains to show that x^* is a maximum. To derive this from Second-Order Conditions is tedious. Notice however that in $(\mu(x), \sigma(x))$ -space, it is easy to show that the t-ratio has linear isoquants and the set H is convex (its boundaries are parabolic segments). This implies that x^* describes a local optimum. The problem's Lagrangian is continuous because objective function as well as constraint are and the objective's denominator cannot be zero if $V_1(X) > 0$. Being the only critical point, x^* must then be a global optimum. ■

An application of this result would be the following: assume that a random sample with nonrandom missing data generated an observed t-ratio that corresponds to a p-value of 0.001. Then the bounds imply the minimum t-ratio, and hence maximum p-value, that this sample could have generated had the missing data been observed. If this maximum p-value is below 5%, then the sample is known to reject the null hypothesis in question at a 5% significance level without any assumptions on the missing data process.

The corollary does not establish that if bounds on the t-ratio are used in this way, the resulting decision procedure has a pre-assigned nominal risk of type I errors in the presence of missing data. Such a generalization of t-ratios is the subject of ongoing research.

2 Worst-Case Bounds on the Coefficient of Variation

Closed-form bounds on the coefficient of variation, $CV(X) \equiv \sqrt{V(X)}/E(X)$, are as follows.

Corollary 2 *In the same setup as above, the coefficient of variation is bounded by*

$$\frac{\left[pE_1(X^2) (E_1(X))^2 + (1-p) (E_1(X))^2 - \left[p(E_1(X))^2 + (1-p)E_1(X^2) \right]^2 \right]^{1/2}}{p(E_1(X))^2 + (1-p)E_1(X^2)} \leq CV(X) \leq \frac{\left[pE_1(X^2) + (1-p)\alpha - [pE_1(X) + (1-p)\alpha]^2 \right]^{1/2}}{(pE_1(X) + (1-p)\alpha)},$$

where

$$\alpha = \begin{cases} 0 & \text{if } \alpha^* < 0 \\ \alpha^* & \text{if } 0 \leq \alpha^* \leq 1 \\ 1 & \text{if } \alpha^* > 1 \end{cases}$$

$$\alpha^* = \frac{pE_1(X) - 2pE_1(X^2)}{1-p}.$$

These bounds are sharp.

Proof. This follows from the previous corollary by omitting the factor $(N-1)$, setting θ equal to 0, and exchanging minimization and maximization. Minimization of CV is achieved by choosing $x^* = E_1(X^2)/E_1(X) \in [0, 1]$, i.e. no adjustment for the case that $x^* < 0$ or $x^* > 1$ is needed; indeed, x^* has been substituted for in the above expression. ■