

Percentile bootstrap of the median paired delta

mare_mark stats design note

This note supports the stats design page. It states what `bootstrap_interval` computes, proves the invariants the page relies on, and derives the exact bootstrap distribution of the median so that a small case can be checked by hand.

1. Setting

Let $d = (d_1, \dots, d_n)$ be the paired deltas, $d_i = c_i - b_i$, and $d_{(1)} \leq \dots \leq d_{(n)}$ their order statistics. For $0 \leq p \leq 1$ the type-7 quantile of a finite sample y of size k is

$$Q_y(p) = y_{(\lfloor h \rfloor + 1)} + (h - \lfloor h \rfloor)(y_{(\lfloor h \rfloor + 2)} - y_{(\lfloor h \rfloor + 1)}), \quad h = (k-1)p, \quad (1)$$

with the convention that the second term vanishes when h is an integer. `bootstrap_interval(d, seed, B, gamma)` performs the following steps.

1. For $b = 1, \dots, B$, draw indices $j_1, \dots, j_n$ from the seeded generator and set $\theta_b^* = Q_{d^*}(1/2)$ with $d^* = (d_{j_1}, \dots, d_{j_n})$.
2. Sort $\theta_1^*, \dots, \theta_B^*$ and return $L = Q_{\theta^*}(\alpha/2)$, $U = Q_{\theta^*}(1 - \alpha/2)$ with $\alpha = 1 - \gamma/100$.

2. Invariants

Lemma 1 (range). For every sample y and every p , $\min_i y_i \leq Q_y(p) \leq \max_i y_i$, and $p \mapsto Q_y(p)$ is non-decreasing.

Proof. $Q_y(p)$ is $(1-w)y_{(i)} + wy_{(i+1)}$ with $w \in [0, 1]$, a convex combination of two order statistics, which lie in $[y_{(1)}, y_{(k)}]$. On each interval $[i/(k-1), (i+1)/(k-1)]$ the map is affine with slope $(k-1)(y_{(i+1)} - y_{(i)}) \geq 0$, and it is continuous at the nodes. $\square$

Proposition 2 (bounds). For $0 < \gamma < 100$, $\min_i d_i \leq L \leq U \leq \max_i d_i$.

Proof. Each d^* consists of values of d , so by Lemma 1 $\theta_b^* \in [\min d^*, \max d^*] \subset [\min d, \max d]$. Applying Lemma 1 to the sample θ^* places L and U in $[\min_b \theta_b^*, \max_b \theta_b^*]$. Finally $\alpha/2 < 1 - \alpha/2$ because $\alpha < 1$, and monotonicity gives $L \leq U$. $\square$

3. The exact bootstrap distribution of the median

Proposition 3. Let $n = 2m + 1$ and let the deltas be distinct. Under resampling with replacement, the resampled median θ^* takes only the values $d_{(1)}, \dots, d_{(n)}$, and

$$\hat{G}(d_{(k)}) = P^*(\theta^* \leq d_{(k)}) = \sum_{j=m+1}^n \binom{n}{j} \left(\frac{k}{n}\right)^j \left(1 - \frac{k}{n}\right)^{n-j}. \quad (2)$$

Proof. For odd n the type-7 median of d^* is its middle order statistic, which is one of the drawn values. The event $\theta^* \leq d_{(k)}$ holds exactly when at least $m+1$ of the n draws are among the k smallest values. Draws are independent and each falls among them with probability k/n , so the count is $\text{Bin}(n, k/n)$. $\square$

For even $n = 2m$ the resampled median is the midpoint of the m -th and $(m+1)$ -th resampled order statistics, so θ^* takes values in the set of midpoints $(d_{(i)} + d_{(j)})/2$ with $i \leq j$. In both cases $\hat{G}$ is a step function, and the interval endpoints are observed deltas or midpoints of two of them.

As a worked case, take $n = 5$. Proposition 3 gives $\hat{G}(d_{(1)}) = 0.058$, $\hat{G}(d_{(2)}) = 0.317$, $\hat{G}(d_{(3)}) = 0.683$, $\hat{G}(d_{(4)}) = 0.942$ and $\hat{G}(d_{(5)}) = 1$. Since $\hat{G}(d_{(1)}) > 0.025$, the exact 95 % percentile interval is $[d_{(1)}, d_{(5)}]$: with five blocks the interval for the median is the whole range of the deltas.

4. Coverage of the percentile interval

Proposition 4 (Efron). Suppose there is an increasing map φ and a distribution H , symmetric about 0, such that $W = \varphi(\hat{\theta}) - \varphi(\theta) \sim H$ and $\varphi(\theta^*) - \varphi(\hat{\theta}) \sim H$ under resampling. Then the percentile interval $[\hat{G}^{-1}(\alpha/2), \hat{G}^{-1}(1 - \alpha/2)]$ covers θ with probability exactly $1 - \alpha$.

Proof. From the second assumption, $\hat{G}(x) = H(\varphi(x) - \varphi(\hat{\theta}))$, so $\hat{G}^{-1}(q) = \varphi^{-1}(\varphi(\hat{\theta}) + H^{-1}(q))$. Since φ is increasing, $\theta \leq \hat{G}^{-1}(1 - \alpha/2) \iff -W \leq H^{-1}(1 - \alpha/2) \iff W \geq H^{-1}(\alpha/2)$ by symmetry, and likewise $\theta \geq \hat{G}^{-1}(\alpha/2) \iff W \leq H^{-1}(1 - \alpha/2)$. The probability of both is $H(H^{-1}(1 - \alpha/2)) - H(H^{-1}(\alpha/2)) = 1 - \alpha$. $\square$

When the assumption holds only asymptotically, the coverage error of the two-sided percentile interval is of order $n^{-1/2}$ in general (Efron and Tibshirani, *An Introduction to the Bootstrap*, 1993, ch. 14). The discreteness of Proposition 3 adds an error that does not vanish with B ; it only vanishes as n grows.

5. Monte Carlo error

The returned endpoints are quantiles of B draws from $\hat{G}$, not of $\hat{G}$ itself. Let q be the p -quantile of $\hat{G}$ and N the number of the B draws that are at most q . Then $N \sim \text{Bin}(B, \hat{G}(q))$, so the fraction of draws below the true endpoint has standard deviation $\sqrt{\hat{G}(q)(1 - \hat{G}(q))/B}$. For $p = 0.025$ and $B = 2000$ that is 0.0035: the reported 95 % endpoints are the true bootstrap endpoints of an interval whose tail masses are 2.5 ± 0.35 %. Because $\hat{G}$ is a step function, the endpoint usually equals the exact one once B exceeds a few hundred.

6. Index generation

Indices are $j = x \bmod n$ for 64-bit states x . Write $2^{64} = qn + r$ with $0 \leq r < n$. Exactly r residues are produced by $q + 1$ states and the others by q states. For a uniformly distributed state,

$$|P(j) \cdot n - 1| \leq \frac{n}{2^{64}}, \quad (3)$$

which is below 10^{-13} for $n \leq 10^6$.