Amy Braverman

(Jet Propulsion Laboratory)

Subsampling

Sponsored by the Jet Propulsion Laboratory and by the Keck Institute for Space Studies, California Institute of Technology

Outline

Whether Frequentist or Bayesian, we need to know how the sampling distribution of $\hat{\theta}$ depends on the (realized) value (of $\Theta =$) θ.

Can we get this information without making assumptions about the distribution*, $f_{\hat{\theta}|\Theta}(\hat{\theta}|\theta)$?

* Please forgive the abuse of notation: using $\hat{\theta}$ to represent both a random variable and its realized value.

Outline

Introduction to subsampling:

- Subsampling.
- Comments and cautions.

Material in this section based substantially on notes provided by Professors Anirban DasGupta of Purdue University and Charlie Geyer of the University of Minnesota, and on the Politis, Romano, and Wolf 1999 book, *Subsampling*.

Subsampling

- A subsample is a sample of size $M < N$ from $Y_1, \ldots, Y_N$ drawn *without* replacement. Denote the subsample by $V_1, \ldots, V_M$.
- $V_1, \ldots, V_M$ is a sample from the *true* distribution, F.
- Compute $\hat{\theta}^* = g(V_1, \ldots, V_M)$. Recall that $\hat{\theta} = g(Y_1, \ldots, Y_N)$.
- Repeat the subsampling B times and obtain $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$.
- The CDF of $\hat{\theta}$ is estimated by $L_{Sub,M}(t)$:

$$L_{Sub,M}(t) = P_F(\sqrt{M}(\hat{\theta}^* - \hat{\theta}) \leq t) \approx \frac{1}{B} \sum_{b=1}^{B} 1\left[\sqrt{M}\left(\hat{\theta}^*_b - \hat{\theta}\right) \leq t\right].$$

Subsampling

Only two things are required:

- We must assume that $N^{\alpha}\left(\hat{\theta}-\theta\right)$ converges (in probability) to J where J is some non-degenerate distribution, and α is some exponent. We don't have to know either of them- just that they exist.

- We have to choose M such that

$$M \longrightarrow \infty, \quad \frac{M}{N} \longrightarrow 0 \quad \text{as } N \longrightarrow \infty.$$

If so, then $L_{Sub,M}(t)$ converges (in probability) to $J(t)$ as $N \longrightarrow \infty$. *

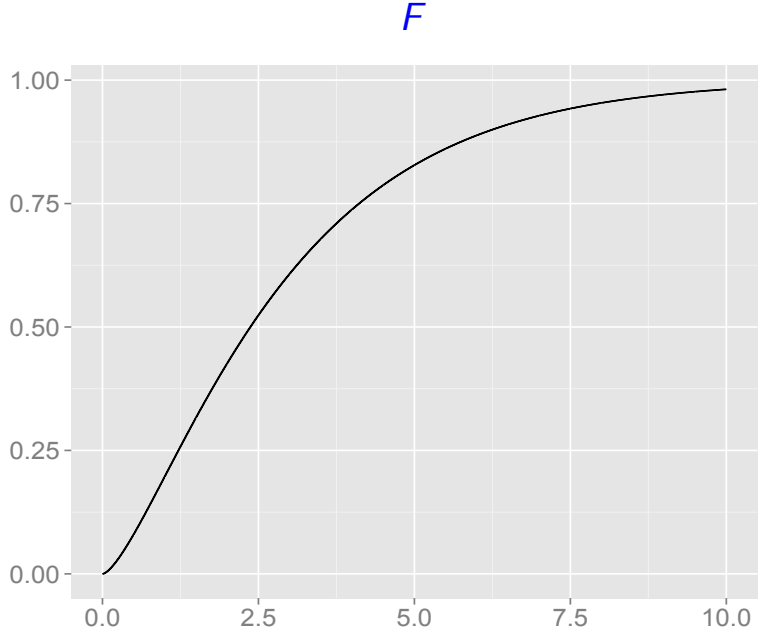
* Stated imprecisely and without all technical conditions. See Politis, Romano, and Wolfe (1999) page 43 for details.

Subsampling

Remarks:

- Assumptions less restrictive than required for the Bootstrap.
- “Works" for situations where the Bootstrap won’t: stationary m-dependent sequences (time series), extremes, etc.
- The catch: must choose M to be big, but not too big. How?

Subsampling



Example of the concept: $M = 3$

$$\mathbf{y} = (2.82, 1.71, 2.07, 4.60, 3.39)^T,\ \hat{\theta}(\mathbf{Y}) = 2.82,$$

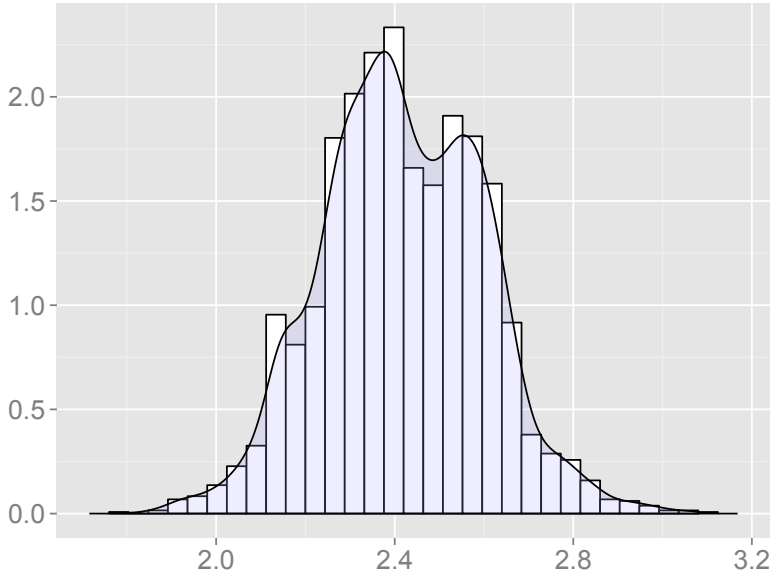
$$\begin{aligned}
\mathbf{v}_1 &= (4.60, 1.71, 2.07)^T, \quad \hat{\theta}_1^*(\mathbf{Y}) = 2.07,\\
\mathbf{v}_2 &= (2.07, 2.82, 1.71)^T, \quad \hat{\theta}_2^*(\mathbf{Y}) = 2.07,\\
\mathbf{v}_3 &= (3.39, 1.71, 2.82)^T, \quad \hat{\theta}_3^*(\mathbf{Y}) = 2.82,\\
&\vdots\\
\mathbf{v}_B &= (1.71, 3.39, 2.82)^T, \quad \hat{\theta}_B^*(\mathbf{Y}) = 2.82.
\end{aligned}$$

Yes, it's true that it all depends on the sample, but we are taking expectations over all possible samples.

Subsampling

Subsampling distribution of the sample median, $\hat{\theta}^*$

$N = 3000, B = 3000, M = \lfloor N^{2/3} \rfloor$

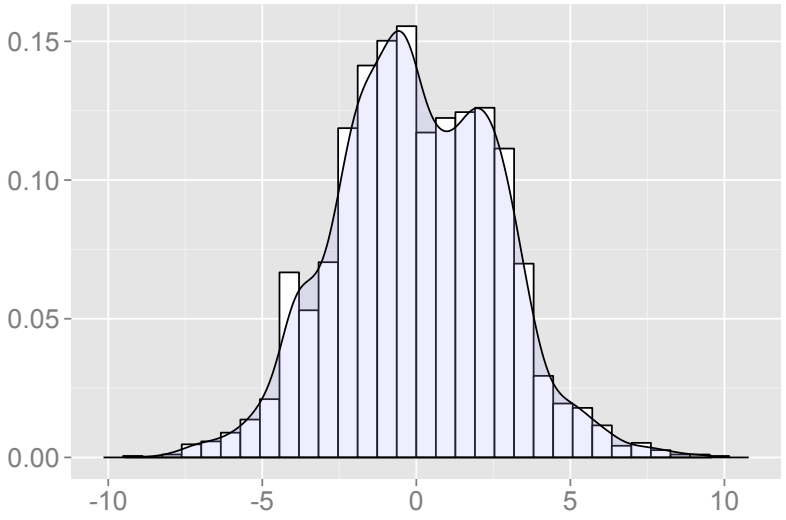


$E(\hat{\theta}^*) = 2.418,$

$var(\hat{\theta}^*) = .033$

Subsampling distribution of the sample median, $\sqrt{M}\left(\hat{\theta}^* - \hat{\theta}\right)$

$N = 3000, B = 3000, M = \lfloor N^{2/3} \rfloor$



$E\left(\sqrt{M}(\hat{\theta}^* - \hat{\theta})\right) = .015,$

$var\left(\sqrt{M}(\hat{\theta}^* - \hat{\theta})\right) = 6.798$

Subsampling

	Expected value	Variance
True: $\sqrt{N}(\hat{\theta} - q_{.50})$	-.049	7.118
Bootstrap: $\sqrt{N}(\hat{\theta}^* - \hat{\theta})$	.257	9.063
M/N bootstrap: $\sqrt{M}(\hat{\theta}^* - \hat{\theta})$	.111	7.342
Subsampling: $\sqrt{M}(\hat{\theta}^* - \hat{\theta})$	.015	6.798

So, which one to use?

Subsampling

Quoting/paraphrasing Charlie Geyer:

- The Right Thing is samples of the right size, N, from the right distribution, F.
- The Politis and Romano thing (subsampling) is samples of the wrong size, M, from the right distribution, F.
- The Efron thing (bootstrapping) is samples of the right size, N, from the wrong distribution, $\hat{F}_N$.
- Both Wrong Things are wrong. We would like to do the right thing, but we can't because we have only one sample.
- Which Wrong Thing do we want to do?