Asymptotic Convergence of Thinned Subsampling

Lucinda Khalil

December 7, 2023

Let $\mathbf{x} = (x_1, x_2, ..., x_n)$ be the observed data, which are independent and identically distributed according to some unknown distribution with finite variance. Let $\hat{\theta} = \hat{\theta}(\mathbf{x})$ be an estimator of θ which is asymptotically normal with a limiting variance of $\sigma^2 < \infty$. Then by the central limit theorem:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \sigma^2)$$

We consider estimators which can be expanded in the following form

$$\hat{\theta} = \theta + \frac{1}{n} \sum_{i=1}^{n} \phi_F(x_i) + R_n \tag{1}$$

where ϕ_F is measurable in x with $E_F(\phi_F(X)) = 0$, $0 < E_F(\phi_F^2(X)) = \sigma^2 < \infty$, and the remainder term satisfies

$$\sqrt{n} R_n \xrightarrow{p} 0$$

The condition that ϕ_F should be measurable ensures that the resulting $\phi_F(x_i)$ terms are all also random variables, since random variables are defined to be measurable functions (Fristedt and Gray (1997)), and the composition of measurable functions is also measurable (Strichartz (2010)). We can therefore describe estimators satisfying (1) as having a linear part which is $O(n^{-1/2})$ and a remainder term with a lower order.

Consider $\mathbf{x_s}$ to be a subsample of $\mathbf{x}$ which is obtained by thinning, (sometimes described as Poisson subsampling as in Hájek (1960)). This is done by selecting each data point in $\mathbf{x}$ with probability p. Here we will consider the case where the probability assigned to each x_i is the same, this method is sometimes called Bernoulli subsampling (Wang and Zou (2021)). Thinning is equivalent to simple random sampling without replacement where the size of the subsample, r, is a random variable with a binomial distribution. In order to obtain an estimate of the distribution of θ, the statistic is calculated on B thinned samples where B is large, so this size of each thinned subsample is distributed as $r_s \sim Bin(n,p)$ for $s = 1, 2, ..., B$. Also set $d_s = n - r_s$ to be the number of deleted samples. Note that if r_s is sampled to be 0 or n, we either have an empty subsample, or one equivalent to the original observed sample, so we discard this and resample r_s. Both of these cases are unlikely when p is not close to 0 or 1, and the probability of this happening coverges to 0 as $n \to \infty$. This result is proven on page 151 of Feller (1968) and will be discussed in more detail later in this proof.

We want to show that this method of subsampling results in estimates which are distributed is the same way as the true statistic asymptotically. Since we have assumed that the true θ is asymptotically normal, we need to show that the distribution of the thinned $\hat{\theta}_s$ also obeys the central limit theorem and is asymptotically normal.

Note that thinning can also be reframed as a delete-d Jackknife where d, the number of x_i deleted for each subsample is instead variable. The following proof is based on the asymptotic normality of the distribution of the Jackknife estimates presented in Theorem 2 (iii) of Wu (1990).

Similarly, we define the Jackknife histogram

$$J(t) = P\left\{\left(\frac{nr_s}{d_s}\right)^{1/2}\frac{\hat{\theta}_s - \hat{\theta}}{\hat{\sigma}} \leq t\right\}$$

where $\hat{\sigma}^2$ is a consistent estimator of the limiting variance σ^2. Since $\hat{\theta}$ is assumed to be normal, the aim is to prove that $\hat{\theta}_s$ is also $\forall s \in 1, 2, ..., B$.

Theorem 1. *If $d_s/n \geq \lambda$ for some $\lambda > 0$ and $r_s \to \infty$ for all $s = 1, 2, ..., B$, then*

$$\sup_t |J(t) - \Phi(t)| \xrightarrow{p} 0$$

For this proof, we will use the assumed form of the estimator seen in (1) to obtain:

$$\hat{\theta}_s = \theta + \frac{1}{r_s}\sum_{i \in x_s} \phi_F(x_i) + R_{n,s}$$

$$\hat{\theta}_s - \hat{\theta} = \underbrace{\frac{1}{r_s}\sum_{i \in x_s} \phi_F(x_i) - \frac{1}{n}\sum_{i=1}^{n} \phi_F(x_i)}_{A} + \underbrace{R_{n,s} - R_n}_{B}$$

Firstly, we will show in Theorem 2 that the term linear in ϕ_F, A, converges asymptotically to a normal distribution, and then in Theorem 3 that the remainder term, B, is asymptotically negligible.

To show the convergence of the linear part, define $\hat{\theta}_s^L$ and $\hat{\theta}^L$ to be the truncated versions of $\hat{\theta}_s$ and $\hat{\theta}$ respectively by removing the remainders:

$$\hat{\theta}_s^L = \theta + \frac{1}{r_s}\sum_{i \in x_s} \phi_F(x_i)$$

$$\hat{\theta}^L = \theta + \frac{1}{n}\sum_{i=1}^{n} \phi_F(x_i)$$

Let the corresponding Jackknife histogram be:

$$J_L(t) = P\left\{\left(\frac{nr_s}{d_s}\right)^{1/2}\frac{\hat{\theta}_s^L - \hat{\theta}_s}{\hat{\sigma}} \leq t\right\}$$

Theorem 2. *If $0 < E[(X - \mu)^2] < \infty$, $r_s \to \infty$ and $d_s \to \infty$ for all $s = 1, 2, ..., B$, then*

$$\sup_t |J_L(t) - \Phi(t)| \xrightarrow{a.s.} 0$$

where Φ is the standard normal cdf.

Proof. We will begin by showing that the following is satisfied for a thinned set x_s of size r_s:

$$\lim_{n \to \infty} \frac{1}{(n-1)\hat{\sigma}^2}\sum_{i=1}^{n}(x_i - \hat{\theta})^2 \; 1\left(|x_i - \hat{\theta}| \geq \tau\sqrt{\frac{r_s d_s}{n}}\hat{\sigma}\right) = 0 \qquad \tau > 0 \tag{2}$$

Note that $\hat{\sigma}^2 r_s d_s / n$ is the variance of $\hat{\theta}_s$ as seen in Hájek (1960).

Since we have assumed $\hat{\theta} \to \theta$ and $\hat{\sigma} \to \sigma < \infty$ almost surely, (2) becomes:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (x_i - \theta)^2 \; 1 \left(|x_i - \theta| \geq \tau \sqrt{\frac{r_s d_s}{n}} \right) \tag{3}$$

where the coefficient of the sum has been simplified but remains $O(1/n)$ and the σ inside the indicator is taken into the constant τ.

Now we bound $r_s d_s/n$ above and below. Without loss of generality, let $\min(r_s, d_s) = d_s$, such that $\hat{p} = r_s/n \geq \frac{1}{2}$, then

$$\begin{aligned} \frac{r_s d_s}{n} &= n\hat{p}(1-\hat{p}) = d_s \hat{p} > \frac{1}{2} \min(r_s, d_s) \\ \frac{r_s d_s}{n} = \frac{(n-d_s)d_s}{n} &= d_s - \frac{d_s^2}{n} < d = \min(r_s, d_s) \\ \implies \frac{1}{2} \min(r_s, d_s) &< \frac{r_s d_s}{n} < \min(r_s, d_s) \end{aligned}$$

This result shows $\min(r_s, d_s) \to \infty \iff r_s d_s/n \to \infty$ as $n \to \infty$.This means that given any $k > 0$, we can choose an m large enough such that for $n \geq m$, $r_s d_s/n > k^2$, then (3) can be bounded above:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\theta})^2 \; 1 \left(|x_i - \theta| \geq \tau k \right)$$

which is arbitrarily small since k is arbitrarily large. So we have shown that (2) holds, then clearly

$$\lim_{n\to\infty} \frac{1}{(n-1)\hat{\sigma}^2} \sum_{i \in x_s} (x_i - \hat{\theta})^2 \; 1 \left(|x_i - \hat{\theta}| \geq \tau \sqrt{\frac{r_s d_s}{n}} \hat{\sigma} \right) = 0 \qquad \tau > 0$$

since we have only removed d_s terms from the sum here. By Theorem 3.1 in Hájek (1960), this implies that $\hat{\theta}_s^L$ has an asymptotically normal distribution with mean θ and variance $\hat{\sigma}^2 r_s d_s/n$. Then Pólya (1920) implies the initial claim:

$$\sup_t |J_L(t) - \Phi(t)| \xrightarrow{p} 0$$

Therefore the claim holds.

□

We now show that the remainder term, B, is asymptotically negligible.

Theorem 3. *If $\sqrt{n} R_n \xrightarrow{p} 0$, $\frac{d_s}{n} > \lambda$ for some $\lambda > 0$ for all $s = 1, 2, ..., B$ and $r_s \to \infty$, then*

$$Q(\epsilon) = P\left\{ \left(\frac{n r_s}{d_s} \right)^{1/2} |R_{n,s} - R_n| \leq \epsilon \right\} \xrightarrow{p} 0$$

Proof. $Q(\epsilon)$ is clearly non-negative, and we can use the constant λ to bound it above:

$$\left(\frac{n r_s}{d_s} \right)^{1/2} |R_{n,s} - R_n| \leq \left(\frac{r_s}{\lambda} \right)^{1/2} |R_{n,s} - R_n|$$

Then using the assumption that $\sqrt{n} R_n \xrightarrow{p} 0$:

$$\lim_{n\to\infty} \left(\frac{r_s}{\lambda} \right)^{1/2} |R_{n,s} - R_n| = \lim_{n\to\infty} |\sqrt{r_s} R_{n,s} - \sqrt{r_s} R_n|$$

where the first term clearly converges to 0 since $R_{n,s}$ is of size r_s, and the second term is bounded above by $\sqrt{n} R_n$ which also converges to 0. Therefore, the claim in proven. □

Note that the requirement that $\frac{d_s}{n} > \lambda$ for some $\lambda > 0$ may not always be met since d_s is a random variable, however as n becomes large we can see using Feller (1968) that for $p < 0.9$, the probability of d_s less than $0.1n$ for example converges to 0:

$$P(d_s \leq 0.1n) \leq \frac{(n - 0.1n)(1-p)}{(n(1-p) - 0.1n)^2} = \frac{0.9(1-p)}{n(0.9-p)^2} \to 0$$

This means that asymptotically, given that we have chose $p < 0.9$, we can assume $\frac{d_s}{n} > 0.1$ and more generally that this term is $O(n)$.

Finally, combining Theorems 2 and 3 prove the desired result stated in Theorem 1.

References

William Feller. *An introduction to probability theory and its applications*, page 150–152. John Wiley, 1968.

B. Fristedt and L. Gray. *A Modern Approach to Probability Theory.* Birkhauser, Boston, MA, 1997.

Jaroslav Hájek. Limiting distributions in simple random sampling from a finite population. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 5:361–374, 1960.

G. Pólya. Über den zentralen grenzwertsatz der wahrscheinlichkeitsrechnung und das momentenproblem. *Mathematische Zeitschrift*, 8:171–181, 1920. URL `http://eudml.org/doc/167598`.

Robert S. Strichartz. *The way of analysis.* World Publishing Corporation, 2010.

HaiYing Wang and Jiahui Zou. A comparative study on sampling with replacement vs poisson sampling in optimal subsampling. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 289–297. PMLR, 13–15 Apr 2021. URL `https://proceedings.mlr.press/v130/wang21a.html`.

C. F. J. Wu. On the asymptotic properties of the jackknife histogram. *The Annals of Statistics*, 18(3):1438–1452, 1990. ISSN 00905364. URL `http://www.jstor.org/stable/2242062`.