Gov 2001 Section 7, 2025

Review of Estimation

Prove/explain why the following statements are true, or give a counterexample if false:

1. There exists an estimator $\hat{\theta}$ for some parameter θ for which $\text{MSE}(\hat{\theta}) = (\text{Bias}(\hat{\theta}))^2$.
2. If $\hat{\theta}$ is an unbiased estimator for θ, all other estimators for θ are biased.
3. The sample mean is an unbiased estimator under all models for which a mean exists.
4. The squared error loss of an estimator for its estimand $L(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2$ is an r.v.

Important Inequalities

1. **Markov's inequality.** Let X be a nonnegative rv such that $E(X)$ exists. For any $t > 0$,

$$P(X \geq t) \leq \frac{E[X]}{t}.$$

   Proof: By LOTP, $E[X] = E[X|X \geq t]P(X \geq t) + E[X|X < t]P(X < t)$. The second term is positive, while $E[X|X \geq t] \geq t$ by the definition of conditional expectation.

2. **Chebyshev's inequality.** Let $X \sim [\mu, \sigma^2]$. We use this notation to express that r.v. X has mean μ and variance σ^2 (but it is not necessarily normal). For any $t > 0$,

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

   Proof: Let $Z = (X - \mu)/\sigma \sim [0, 1]$. So $E(Z^2) = 1$. Then

$$P(|X - \mu| \geq t) = P\left(\left(\frac{X - \mu}{\sigma}\right)^2 \geq \frac{t^2}{\sigma^2}\right) = P\left(Z^2 \geq \frac{t^2}{\sigma^2}\right) \leq \frac{E(Z^2)}{t^2/\sigma^2} = \frac{\sigma^2}{t^2}.$$

3. **Cauchy-Schwartz inequality.** If X, Y have finite variances, then

$$E|XY| \leq \sqrt{E[X^2]E[Y^2]}.$$

4. **Jensen's inequality.** If g is a convex function such that $E[g(X)]$ exists, then

$$E[g(X)] \geq g(E[X]).$$

Convergence

1. Convergence in probability: $X_n \xrightarrow{p} X$ if for all $\varepsilon > 0$,

$$P(|X_n - X| \geq \varepsilon) \xrightarrow{n \to \infty} 0.$$

You can interpret $|X_n - X|$ as the distance between two r.v.s, and $|X_n - X| \geq \varepsilon$ is the event that X_n is outside the circle of which X is the center and ε is the radius. Convergence in probability indicates that when you have more data (n goes to infinity), even for an extremely small ε, it becomes very unlikely that X_n is outside the circle (the probability goes to zero). In other words, X_n moves arbitrarily close to X.

Notice that though in most cases we use $X_n \xrightarrow{p} c$ for constant c, which implies that the variance vanishes when $n \to \infty$, X_n *can* also converge to a random variable. For example, let the sequence $X_n = X + \frac{1}{n}Z$, where r.v. $Z \sim \mathcal{N}(0,1)$, independent of r.v. X. Then $X_n \xrightarrow{p} X$.

Note: Convergence in probability can usually be proved by Markov's inequality or Chebyshev's inequality. For example, let $X_i \overset{i.i.d}{\sim} [\mu, \sigma^2]$ for $i = 1, \ldots, n$, and $\bar{X}_n = \sum_{i=1}^n X_i/n$. Then, for any $\varepsilon > 0$,

$$P(|\bar{X}_n - \mu| > \varepsilon) \leq \frac{\mathrm{Var}(\bar{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0$$

2. Convergence in distribution: $X_n \xrightarrow{d} X$ if the CDF of X_n converges pointwise to the CDF of X, i.e.,

$$\lim_{n\to\infty} F_{X_n}(u) = F_X(u)$$

Notice that convergence in probability is stronger than convergence in distribution: $X_n \xrightarrow{p} X \Rightarrow X_n \xrightarrow{d} X$ but $X_n \xrightarrow{d} X \not\Rightarrow X_n \xrightarrow{p} X$.

Example: Let $X_n \sim \mathcal{N}(1/n, 1)$ for $n \in \mathbb{N}$; and $X \sim \mathcal{N}(0,1)$, where $X, X_1, X_2, \ldots, X_n$ are mutually independent. As always, denote the CDF of $\mathcal{N}(0,1)$ by $\Phi(\cdot)$. By the continuity of Φ, we have

$$\mathbb{P}(X_n \leq x) = \Phi(x - 1/n) \to \Phi(x) = \mathbb{P}(X \leq x).$$

So, $X_n \xrightarrow{d} X$, which means their distributions are close for n large. However, it does not mean that $|X_n - X|$ is close to zero with high probability. Indeed, $X_n \not\xrightarrow{p} X$. Since X_n and X are independent normal distribution, $X_n - X \sim \mathcal{N}(1/n, 2)$. So for any fixed $\varepsilon > 0$,

$$\mathbb{P}(|X_n - X| > \varepsilon) = 2\left\{1 - \Phi\left(\frac{\varepsilon - 1/n}{\sqrt{2}}\right)\right\} \not\to 0.$$

3. Interpretation in estimation: usually, our statistic or estimator (remember: they are functions of the data $X_1, \ldots, X_n$) is treated as a sequence, written as T_n. We are interested in how they behave as we add a sufficiently large amount of data ($n \to \infty$). For example, the sample mean is $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, and this forms a sequence when n increases. Both modes of convergence try to assert that "some sort of estimation error" is small:

   (a) $X_n \xrightarrow{p} X$ tells that getting large error, i.e., $|X_n - X| \geq \varepsilon$ is unlikely.

   (b) $X_n \xrightarrow{d} X$ tells that the error in the shape of $x \mapsto P(X_n \leq x)$ relative to $x \mapsto P(X \leq x)$ is small.

Tools for Asymptotics

1. Weak Law of Large Numbers (WLLN): Let $X_1, \cdots, X_n \overset{i.i.d.}{\sim}$ with mean μ (and variance σ^2). Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, then as $n \to \infty$,

$$\bar{X}_n \xrightarrow{p} \mu$$

2. Central Limit Theorem (CLT): Let $X_1, \cdots, X_n \overset{i.i.d.}{\sim}$ with mean μ and variance σ^2. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, then as $n \to \infty$,

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \xrightarrow{d} \mathcal{N}(0,1) \quad \text{or} \quad \sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{d} \mathcal{N}(0,\sigma^2)$$

   For n sufficiently large (often $n \approx 30$ in practice), this implies $\bar{X}_n \dot{\sim} \mathcal{N}(\mu, \frac{\sigma^2}{n})$. But notice that in principle, you should not write

$$\bar{X}_n - \mu \xrightarrow{d} \mathcal{N}(0, \sigma^2/n)$$

   It is technically incorrect because the right hand side still depends on n.

3. Slutsky's theorem: If $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ are sequences of random variables, such that

$$X_n \xrightarrow{d} X \quad \text{and} \quad Y_n \xrightarrow{p} c \quad (c \text{ a constant}),$$

   then

   (a) $X_n + Y_n \xrightarrow{d} X + c$,

   (b) $X_n Y_n \xrightarrow{d} cX$,

   (c) if $c \neq 0$, then $X_n/Y_n \xrightarrow{d} X/c$.

   Important: In general, $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{d} Y$ does NOT imply that $X_n + Y_n \xrightarrow{d} X + Y$

4. Continuous mapping theorem: If $X_1, X_2, \ldots$ are sequences of random variables and g is a continuous function, then

   (a) if $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$,

   (b) if $X_n \xrightarrow{p} X$, then $g(X_n) \xrightarrow{p} g(X)$.

5. Delta method: Suppose that

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

   as $n \to \infty$ and g is a continuously differentiable function. Then as $n \to \infty$,

$$\sqrt{n}(g(\hat{\theta}) - g(\theta)) \xrightarrow{d} \mathcal{N}\left(0, \left(\frac{\partial g(\theta)}{\partial \theta}\right)^2 \omega^2\right)$$

   Example: If we have $\bar{X}_n \dot{\sim} \mathcal{N}(\mu, \frac{\sigma^2}{n})$, then

$$g(\bar{X}_n) \dot{\sim} \mathcal{N}\left(g(\mu), (g'(\mu))^2 \frac{\sigma^2}{n}\right)$$

Consistency

An estimator $T(Y) = \hat{\theta}$ is said to be consistent for the estimand θ if, as $n \to \infty$, $\hat{\theta} \xrightarrow{p} \theta^*$, i.e., it converges in probability to the estimand's true value. To prove consistency, there are several approaches you may consider:

1. Mean Squared Error: A sufficient (but not necessary) condition for consistency of $\hat{\theta}$ for θ is $\text{MSE}(\hat{\theta}, \theta) \to 0$.
2. LLN (sample mean): For i.i.d. X_i with finite expectation $\mathbb{E}(X_i) = \mu$, $\bar{X} \xrightarrow{p} \mu$.
3. Continuous mapping theorem (function of sample mean): Suppose g is continuous, if $X_n \xrightarrow{p} \mu$, then $g(X_n) \xrightarrow{p} g(\mu)$.

Interval Estimator

1. **Definition:** Let $L(Y)$ and $U(Y)$ be functions of the data Y such that $L(y) \leq U(y)$ for all y. The random interval $[L(Y), U(Y)]$ is an interval estimator of a parameter θ if upon observing $Y = y$, the inference $L(y) \leq \theta \leq U(y)$ is made.

   Remark. Notice that frequentists treat the parameter θ as fixed. The randomness comes from where the ends of the interval are positioned as a function of the observed data.

2. **Coverage:** The coverage probability of an interval estimator is the probability that the true parameter lies within the interval, i.e.,

   $$\mathbb{P}(\theta \in [L(Y), U(Y)]).$$

   Note that the coverage probability is a function of θ.

3. **Confidence interval:** An interval estimator with coverage probability at least $1 - \alpha$ for all possible values of θ is called a $100(1 - \alpha)\%$ confidence interval (CI). We call $1 - \alpha$ the level of CI, and call the half-width $0.5(U(Y) - L(Y))$ the margin of error.

   Remarks.

   - We often use $\alpha = 0.05$ and work with 95% confidence intervals.
   - Confidence intervals are widely misinterpreted. Do **not** interpret a 95% CI as there being a 95% probability of a random θ being between two values. In this frequentist setting, we treat θ as fixed, so we should say "the probability that the random interval generated by repeated draws of the data contains the fixed θ^* is 0.95." In other words, if we generate the n data points 100 times and construct the CI for each dataset, then on average 95 of the CIs will contain the true θ^*.
   - A 95% CI is not unique. We can choose between CIs based on the following criteria:
     - Shortest expected width: minimize $\mathbb{E}(U(Y) - L(Y))$.
     - Equal-tailed: require that $\mathbb{P}(\theta < U(Y)) = \mathbb{P}(\theta > L(Y))$.

– Centered on estimator: the CI is $[\hat{\theta} - C(Y), \hat{\theta} + C(Y)]$, where $\hat{\theta}$ is some estimator of θ.

4. **Asymptotic confidence interval:** $[L_n, U_n]$ such that

$$\lim_{n\to\infty} \mathbb{P}(\theta \in [L_n, U_n]) = 1 - \alpha$$

To derive the asymptotic confidence interval of an estimator (particularly for those that are constructed as the mean of data), use CLT and then calculate the upper and lower bounds of the parameter θ, which should be the functions of $\hat{\theta}$.

$$\frac{\hat{\theta} - \theta}{\text{se}(\hat{\theta})} \xrightarrow{d} \mathcal{N}(0, 1)$$

Since we know $\Phi(-1.96) = 1 - \Phi(1.96) = 0.025$, we get the symmetric 95% asymptotic confidence interval:

$$\mathbb{P}\left(\hat{\theta} - 1.96 \cdot \text{se}(\hat{\theta}) \leq \theta \leq \hat{\theta} + 1.96 \cdot \text{se}(\hat{\theta})\right) \to 0.95$$

Practice Questions

1. For $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$ with both parameters unknown, show that sample variance is unbiased and consistent for estimand σ^2:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

2. All probability distributions have *moments*, which are standard expressions that define its shape in ways you've already heard of and other more nuanced ways (the variance, the skew, kurtosis, etc.). Describing a population distribution (or empirical sample distribution) in terms of its moments is really useful in social science (e.g., the skew of income in the U.S. population is positive).

Specifically, the nth central moment of a random variable X is defined as $\mathbb{E}[(X - \mathbb{E}[X])^n]$, but it is more common to work with the nth moment defined as $\mathbb{E}[X^n]$ (getting rid of the $\mathbb{E}[X]^n$ part).

Suppose the random variable X for your population has the following first four moments:

$$\mathbb{E}[X] = \frac{1}{2}, \quad \mathbb{E}[X^2] = \frac{1}{2}, \quad \mathbb{E}[X^3] = \frac{3}{4}, \quad \mathbb{E}[X^4] = \frac{3}{2}.$$

Suppose you took an i.i.d. sample $\{X_1, \ldots, X_{20}\}$ of size 20 from this distribution. Let

$$T = \frac{1}{20}(X_1^2 + X_2^2 + \cdots + X_{20}^2) = \bar{X}^2,$$

an estimator of the second moment.

(a) What are $\mathbb{E}[T]$ and $\text{Var}(T)$? Be sure to explain why.

(b) Use CLT to approximate the probability that $T \leq 1$. Leave your answer as a function of standard normal CDF Φ.

3. We have learned that if the variance of a sequence of random variables with finite mean goes to zero as $n \to \infty$, then the sequence will converge in probability to some value. But this is a **sufficient** condition, not a **necessary** one.

   To see this, consider the sequence of random variables X_n with the following probability distribution:

$$X_n = \begin{cases} 0 & \text{with probability } 1 - \frac{1}{n}, \\ n & \text{with probability } \frac{1}{n}. \end{cases}$$

   (a) Find $\mathbb{E}[X_n]$.

   (b) Use the definition of convergence in probability to show that $X_n \xrightarrow{p} 0$.

4. Let $X_i \overset{\text{iid}}{\sim} \text{Unif}[a, a+1]$, for $i = 1, 2, \ldots$, where $0 < a < \infty$. What is the probability limit of the arithmetic mean (AM), geometric mean (GM), and harmonic mean (HM):

$$A_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad G_n = \left(\prod_{i=1}^{n} X_i\right)^{1/n}, \quad H_n = \frac{n}{\sum_{i=1}^{n} 1/X_i}.$$

   Hint: use log-transformation and then the Continuous mapping theorem.