2002.2 Estimation

Part of @Stats Index

Point Estimation

Inference is an attempt to guess about the properties of a population given a sample.

An *estimator* is a function of sample data. We use it to learn about *estimands*. Using an estimator, you can create *estimates*.

This can get pretty theoretical. For example, an experiment can be conceived as sampling from a world in which everyone is treated.

Setup

Assume a population with an unknown distribution, mean μ and variance σ^2 .

A sample gives you observations $Y_1, \ldots, Y_n$. If the sample is truly random, then the observations are i.i.d. (a key assumption). We can write the sample mean as $\bar{Y}$.

In this case $E(Y_i) = \mu$, and $V(Y_i) = \sigma^2$

Imagine drawing samples over and over, and applying an estimator to each sample. This would create a distribution of estimates, known as that estimator's *sampling distribution*. In reality, we probably have only one sample - so we have to make inferences about the sampling distribution.

Finite-Sample Properties of Estimators

We want an estimator to be *unbiased*: $E[\hat{\mu}] = \mu$ over repeated samples. The degree of bias is $E[\hat{\mu} - \mu] = E[\hat{\mu}] - \mu$.

We also prefer estimators that are more *efficient*. $\hat{\mu}_1$ is more efficient than $\hat{\mu}_2$ if $V(\hat{\mu}_1) < V(\hat{\mu}_2)$ - this being the variance of the sampling distribution.

The square root of the variance of the sampling distribution is the *standard error*.

Relative Estimator Quality

One way to measure the tradeoff between bias and error is the mean squared error.

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$$

Can also be written as $\text{Bias}(\hat{\theta})^2 + V(\hat{\theta})$

Equivalence Derivation

Expanding,

$$\begin{aligned} MSE(\hat{\theta}) &= E[(\hat{\theta} - \theta)^2] \\ &= E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + \theta^2 \\ &= E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + \theta^2 + E[\hat{\theta}]^2 - E[\hat{\theta}]^2 \\ &= E[\hat{\theta}]^2 - 2\theta E[\hat{\theta}] + \theta^2 + E[\hat{\theta}^2] - E[\hat{\theta}]^2 \\ &= \text{Bias}(\hat{\theta})^2 + V(\hat{\theta}) \end{aligned}$$

Estimators of variance

Is the 'Plug-in sample of variance', $\overline{X^2} - \bar{X}^2$, an unbiased estimator of population variance? No.

But Why?

Let $Z = X^2$. Then $E[\overline{X^2}] = E[\bar{Z}] = E[Z] = E[X^2]$. So far so good.

We can write

$$\begin{aligned} E[\bar{X}^2] &= E[E(\bar{X})^2 + \bar{X}^2 - E(\bar{X})^2] \\ &= E(\bar{X})^2 + E(\bar{X}^2) - E(\bar{X})^2 \\ &= E(\bar{X})^2 + V(\bar{X}) \\ &= E(\bar{X})^2 + \sigma^2/n \end{aligned}$$

So what is $E[\overline{X^2} - \bar{X}^2]$?

$$\begin{aligned} E[\overline{X^2} - \bar{X}^2] &= E(\overline{X^2}) - E(\bar{X}^2) \\ &= E[X^2] - E(\bar{X})^2 + -\sigma^2/n \\ &= E[X^2] - E(X)^2 - \sigma^2/n \\ &= \sigma^2 - \sigma^2/n \\ &= \frac{n-1}{n}\sigma^2 \end{aligned}$$

We use a correction, the sample variance -

$$\hat{V}(X) = \hat{\sigma}^2 = \frac{n}{n-1}[\overline{X^2} - \bar{X}^2]$$

Asymptotic Properties of Estimators

Consider the behavior of $\hat{\theta}_1, \dots \hat{\theta}_n$, increasing in sample size. How does this sequence behave?

If $\exists \lim_{n\to\infty} \hat{\theta}_n$ then we have 'stochastic convergence'

There are two types of convergence:

- Convergence in probability: X_n converges in probability to a, or $X_n \xrightarrow{p} a$ if

$$\lim_{n\to\infty} P(|X_n - a| < \varepsilon) \; \forall \; \varepsilon > 0$$

- Convergence in distribution: limit of the CDFs of X_n is equal to CDF of A at every point where the CDF of A is continuous. We write this as $X_n \xrightarrow{d} A$. Convergence in probability is a special case of convergence in distribution where $V(A) = 0$.

A sufficient but unnecessary criterion: $E(X_n) \to a$, $V(X_n) \to 0$ (collapse to a spike)
Ex. sample mean converges in probability to population mean.

Consistency

An estimator is consistent if $\hat{\theta}_1, \dots, \hat{\theta}_n$ converges in probability to true value θ.

A pretty minimal requirement. No guarantee of finite sample performance. Not the same as unbiasedness - e.g. $\bar{X} + 5$ is biased but consistent.

Weak Law of Large Numbers

If $X_1, \dots, X_n$ are i.i.d sequence of RV, each with finite mean μ, then $\bar{X}_n \xrightarrow{p} \mu$ as $n \to \infty$
.

In other words, the sample mean converges in probability to the expected value as sample size grows.

Central Limit Theorem and Asymptotic Normality

My third-favorite theorem, and for good reason.

Allows us to characterize the sampling distribution:

Now we can understand the Central Limit Theorem:

Let $X_1, \ldots, X_n$ be a sequence of i.i.d. RVs with finite mean μ and variance σ^2. Then, regardless of how X_i are distributed:

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

In other words, as the sample size grows, the sampling distribution becomes more normal around the population mean. Specifically - the sampling distribution for $n = 1$ is the population distribution. As n grows, it becomes more and more like a normal.

The CLT also means that standardized sample mean - z-score - converges to standard normal.

Can use this property to make claims about the uncertainty in our estimators - something we will use for hypothesis testing.

Interval Estimation

In addition to point estimates, it'd be nice to talk about uncertainty around estimates.

Thanks to the Central Limit Theorem, we can use standard errors (the sd. of the sampling distribution) in two ways: confidence intervals and p-values (later).

The confidence interval is a 'range of possible values within which we estimate the true value to fall.'

Conf. Intervals

Supposing we're in CLT-land, we can estimate the shape of our sampling distribution: $\mathcal{N}(\hat{\theta}, \hat{\sigma}^2/n)$.

If that sampling distribution were real - and it's our best guess - we can make claims about how frequently we'd expect to see certain results.

Drawing from a regular normal distribution, how likely is it that we'd get something within $\pm\sigma$ of the mean? We can get this from CDF differences: $F(\mu + \sigma) - F(\mu - \sigma)$. In R, this is `pnorm(1) - pnorm(-1)`, which is 68%.

We can do the same for our sampling distribution by standardizing it. Turns out that this follows a Student's t-distribution with parameter $n-1$, which converges to normality as n grows.

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \sim \tau(n-1)$$

CI Interpretation

A 95% CI does **NOT** mean a 95% estimated probability that the true value is in this interval. Rather, it means if we were to do this experiment 100 times, we would expect - based on our data - that this interval would contain the true value 95 times or so.

The intuition from this comes from thinking of the 'true' sampling distribution of 95% CIs. If we drew large-enough-sample CIs over and over from a 'true' sampling distribution centered on μ with SD $\hat{\sigma}/\sqrt{n}$, then, 5% of the time, our sample mean would be more than 1.96 standard errors from the true mean.