Notes on Point Estimates and Confidence Intervals

Shasha Liao
Georgia Tech

October 24, 2020

In statistics, we often use sample mean and sample variance to estimate the unknown true mean and variable of a feature. In linear regression, we also estimate the coefficients of the model from data. In this way, these estimators are all of randomness and are not equal to the exact values of the unknown variables. But, they are close to the true values. How close? Or how accurate are these estimators? How confidence are we to believe that we are making a good estimate?

1 Point Estimator

Suppose we have a set of independent samples $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ from random variables X and Y. And we use $\hat{\theta} = h(x_1, x_2, ..., x_n)$ to estimate θ. Here θ is an unknown determined quantity related to X, such as $E[X]$ and $Var(X)$. Since $\hat{\theta}$ takes values from real values, we call it a **point estimator** of θ. Later, we will discuss about using an interval $[\theta_{low}, \theta_{high}]$ to estimate θ to a degree of confidence.

1.1 Bias

- A good estimator should be unbiased.
- Intuitively, this means that if we could average the estimates $\hat{\theta_1}, \hat{\theta_2}, ..., \hat{\theta_k}$ obtained from a huge number, say k, of datasets, then the average of these estimators should be exactly θ.
- Mathematically, $\hat{\theta}$ is an **unbiased estimator** of θ if

$$E[\hat{\theta}] = \theta, \text{ or } B(\hat{\theta}) = 0$$

where

$$B(\hat{\theta}) = E[\hat{\theta}] - \theta$$

is called the **bias** of the estimator $\hat{\theta}$. Here are some unbiased estimators:

- Sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is an unbiased point estimator for the mean $E[X]$.
- Sample variance $S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ is an unbiased point estimator for the variance of X.

Note that $S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$ is a biased point estimator for the standard deviation of X. Actually, it is not possible to find an estimate of the standard deviation which is unbiased for all population distributions, as the bias depends on the particular distribution.

1.2 Variance

- A good estimator should also have low variance.
- Intuitively, this means if we use k different datasets and made k different estimates $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_k$ for θ, the differences between these k estimates should not be large.
- Mathematically, the **variance** of an estimator $\hat{\theta}$ is defined as

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

  Here are a list of variance of common estimators:

  - Variance of sample mean: $Var(\bar{x}) = \frac{\sigma^2}{n}$, where $\sigma^2 = Var(X)$. (Large $n \Rightarrow$ small $Var(\bar{x})$)
  - Variance of sample variance: $Var(S^2) = \begin{cases} \frac{\mu_4}{n} - \frac{\sigma^4(n-3)}{n(n-1)} & (\text{ general case}), \\ \frac{2\sigma^4}{n-1} & (\text{ if } X \text{ is normal}). \end{cases}$

  where when X is normal, we have $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$. See here for detailed proofs.

- The standard deviation of a point estimator $\hat{\theta}$ is called the **standard error** of $\hat{\theta}$, or $SE(\hat{\theta})$. The standard error of a point estimator will play a crucial rule in estimating the confidence interval and hypothesis testing of the significance of the a point estimate.

Variance of sample mean (from here):

- The variance of the sample median depends on the distribution you are sampling from. If you know the sampling distribution you can use the distribution of order statistics to find the distribution of the median and thence its variance.
- If you don't know and don't want to make assumptions about the distribution, then you can do something like bootstrapping to estimate the variance.

1.3 Mean Squared Error(MSE)

MSE is the most common-used measure of the accuracy of an estimator:

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

A good estimator should have a low MSE.

2 Confidence Interval