Classical Inference: Point Estimation

- Let θ^* denote the unknown parameter of a random variable X (typically mean, variance, scale, shape etc) and suppose we observe i.i.d samples of X which are recorded in the dataset $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$.
- In frequentist approach, we estimate θ^*, by defining a point estimator $\hat{\Theta}$ as a function of the random samples $X_1, \ldots X_n$ as $\hat{\Theta} = h(X_1, \ldots X_n)$.
- While $\hat{\Theta}$ is a random variable, given $\mathcal{D}$ the estimator takes the value $\hat{\Theta} = h(x_1, \ldots x_n)$.
- Example : Sample mean $\hat{\mu}_n = \frac{\sum_{i=1}^{n} X_i}{n}$.

Point Estimators: Properties

- The Bias $B(\hat{\Theta})$ of an estimator $\hat{\Theta}$ is defined as

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

- Unbiased estimators are estimators with zero bias, i.e., $B(\hat{\Theta}) = 0$ and hence $E[\hat{\Theta}] = \theta^*$
- Are all unbiased estimators good ? Let $\hat{\Theta}_1 = X_1$ and $\hat{\Theta}_2 = \frac{\sum_{i=1}^{n} X_i}{n}$. Which estimator is better?
- $Var(\hat{\Theta}_1) = \sigma^2$ while $Var(\hat{\Theta}_2) = \frac{\sigma^2}{n}$.
- We need other measures to determine how good an estimator is, something that looks at the variance of these estimators.

Mean square error of Point Estimators

- The mean squared error of an estimator $\hat{\Theta}$ is defined as

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

- Note that

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

- This means that biased estimators could possibly have lower MSE error if they have extremely low variance!
- Find MSE of $\hat{\Theta}_1 = X_1$ and $\hat{\Theta}_2 = \hat{\mu}_n + 1$.
- Bias-Variance tradeoff talks a lot in machine learning!

Consistency of estimators

- What happens to estimators as the size of the data set ($|\mathcal{D}| = n$) increases? Do all estimators converge to θ^*?
- Not necessarily! For examples $\hat{\Theta}_1 = X_i$ where X_i is picked random from $\mathcal{D}$ does not converge.
- What about $\hat{\mu}_n$. Using SLLN, we see that this does.
- Let $\hat{\Theta}_1, \hat{\Theta}_2, \ldots, \hat{\Theta}_n, \ldots,$ be a sequence of point estimators of θ^* (here n denotes the size of the dataset) We say that $\hat{\Theta}_n$ is a **consistent estimator** of θ, if

$$\lim_{n\to\infty} P(|\hat{\Theta}_n - \theta^*| \geq \epsilon) = 0, \quad \text{for all } \epsilon > 0$$

- This is convergence in probability. If almost sure convergence holds, it is called strongly consistent.
- Clearly, $\hat{\Theta}_n = \hat{\mu}_n$ is strongly consistent and hence consistent.

Consistency of estimators

Theorem

Let $\hat{\Theta}_1, \hat{\Theta}_2, \ldots,$ be a sequence of point estimators of θ^*. If

$$\lim_{n \to \infty} MSE(\hat{\Theta}_n) = 0$$

then $\hat{\Theta}_n$ is a consistent estimator of θ^*

$$\begin{aligned} P(|\hat{\Theta}_n - \theta^*| \geq \epsilon) &= P(|\hat{\Theta}_n - \theta^*|^2 \geq \epsilon^2) \\ &\leq \frac{E[\hat{\Theta}_n - \theta^*]^2}{\epsilon^2} \quad \text{Markov Inequality} \\ &= \frac{MSE(\hat{\Theta}_n)}{\epsilon^2} \\ &\to 0 \text{ as } n \to \infty. \end{aligned}$$