Lezione 9 - Statistic (Inferential Statistics)

Observed Sample

x_i is the realisation of a random variable.

The collection of x_i is the observed sample.

The observed sample is the realisation of the random sample $(X_1, X_2, \dots, X_n \overset{iid}{\sim})$

Statistical Inference

Point Estimation → first problem with inferential statistics
↳ estimation on a population parameter.

Why is it useful to consider the random sample and not the observed values?

Because if we make observations at different times the result might be different.

Using the random variables we can control the uncertainty.

$n = 4$

$X_1, X_2, X_3, X_4 \overset{iid}{\sim} N(\mu, 0{,}8^2)$

We can estimate μ as:

$$\text{Estimator } (\hat{\mu}) = \frac{X_1 + X_2 + X_3 + X_4}{4}$$

Suppose we get 4 realised values, x_1, x_2, x_3, x_4, we get the estimate of μ

$$\underbrace{\text{Estimate } (\hat{\mu})}_{\substack{\text{Realisation of} \\ \text{the estimator}}} = \frac{x_1 + x_2 + x_3 + x_4}{4}$$

What happens if we repeat the experiment, but one of the is different from the ones before?

The ~~est~~ estimator is the sample but the estimate is different

We can check the probability that the estimation error is high:

$$P(\underbrace{|\bar{X}_4 - \mu|}_{} > 1) \simeq 1{,}2\%$$

Probability the estimate is different from the true value by more than 1.

None of the values of μ that we have found are the real values, finding the true value of μ is near impossible

$$\bar{X}_4 \sim N\left(\mu, \frac{0{,}8^2}{4}\right)$$

Because $|\cdot|$

$$\Rightarrow P(|\bar{X}_4 - \mu| > 1) = P\left(\left|\frac{\bar{X}_4 - \mu}{\sqrt{\frac{0{,}8^2}{4}}}\right| > \frac{1}{\sqrt{\frac{0{,}8^2}{4}}}\right) = 2\left(1 - \phi\left(\frac{2}{0{,}8}\right)\right) = 2 \cdot (1 - \phi(2{,}5))$$

It's convenient to measure more, as it reduces the probability of our $\bar{X}_n$ being far from μ.

$X_1 - \mu \sim N(0, (0,8)^2)$ will have more error as we only have one measure.

The more measurements, the more precise will our estimator be.

A parameter, θ, is a characteristic (unknown) of e.g. the mean or variance.

An estimator of θ: $\hat{\theta} = U = g(X_1, \ldots, X_n)$ (random sample)

An estimate of θ: $\hat{\theta} = u = g(x_1, \ldots, x_n)$ (real number)

↳ same g

Unbiased Estimator

An estimator is a statistic $T = g(X_1, \ldots, X_n)$ where task is to estimate the unknown parameter θ

Unbiasedness → first measure of reliability of estimator.

An estimator of θ is unbiased for θ if

$$E(T) = \theta \quad \forall \theta$$

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} f(\theta) \rightarrow T = g(X_1, \ldots, X_n) \sim \tilde{f}(\theta)$$

T is unbiased if $E(T) = \theta \quad \forall \theta$

$E(T - \theta) = 0$

The sample mean is an unbiased estimator for the population mean.

Since $E(\bar{X}_n) = \mu$, since $E(\bar{X}_n) = E(X_i) = \mu$

↳ if the expectation of f is denoted by μ

<u>Prop.</u>

$X_1, X_2, \dots, X_n \overset{iid}{\sim} f$

Suppose we are interested in estimating σ^2

$$S_n^2 = \text{sample variance} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

$$\Rightarrow E(S_n^2) = \sigma^2$$

<u>Proof</u> We are first proving that $\sum_1^n (X_i - \bar{X}_n)^2 = \sum_1^n (X_i - \mu)^2 - n(\bar{X}_n - \mu)^2$

$\mu = E(X_i)$

$$\sum_{i=1}^{n} (X_i - \bar{X}_n)^2 = \sum_{i=1}^{n} (X_i - \mu + \mu + \bar{X}_n)^2 = \sum \{ (X_i - \mu)^2 + (\bar{X}_n - \mu)^2 - 2(\bar{X}_n - \mu)(X_i - \mu)$$

$$= \sum_{i=1}^{n} (X_i - \mu)^2 + n(\bar{X}_n - \mu) - 2(\bar{X}_n - \mu) \underbrace{\sum_{i=1}^{n} (X_i - \mu)}_{n(\bar{X}_n - \mu)}$$

$$= \sum_{i=1}^{n} (X_i - \mu)^2 - n(\bar{X}_n - \mu)^2 - 2n(\bar{X}_n - \mu)^2 \, n(\bar{X}_n - \mu)$$

$$= \sum_1^n (X_i - \mu)^2 - n(\bar{X}_n - \mu)^2 \rightarrow \text{Proof}$$

<u>Proof $E(S_n^2) = \sigma^2$</u>

$$E(S_n^2) = E\left(\frac{1}{n-1} \sum_1 (X_1 - \bar{X}_n)^2\right] = \frac{1}{n-1} E\left[\sum_1 (X_i - \mu)^2 - n(\bar{X}_n - \mu)^2\right]$$

$$= \frac{1}{n-1} \left\{ E\left[\sum_{i=1}^{n} (X_i - \mu)^2\right] - n E\left[(\bar{X}_n - \mu)^2\right]\right.$$

$$\overset{!}{=} \frac{1}{n-1}\, n\sigma^2 - \cancel{n} \cdot \frac{\sigma^2}{\cancel{n}} = \frac{1}{\cancel{n-1}} \cancel{(n-1)}\, \sigma^2 = \sigma^2 \rightarrow \text{proof}$$

These p the sample mean is an unbiased estimator.