# Sample properties

Often times, we do not know the underlying distribution (and thus the pdf) of a given dataset/population and therefore we cannot determine population properties.

$$E(X) = \mu = \int_{-\infty}^{\infty} x f(x)\, dx$$

If $f(x)$ is unknown, we can't compute $\mu, \sigma^2, \sigma$, etc.

An estimator $(\hat{\theta})$ is an expression used to estimate a statistical quantity $\theta$, eg: population mean or variance.

In statistical terms, one can define the bias of an estimator.

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

bias of the estimator

expected value of the estimator

true value of the estimand.

$B(\hat{\theta}) = 0$

(if $E(\hat{\theta}) = \theta$)

$\hat{\theta}$ is called an UNBIASED ESTIMATOR.

$B(\hat{\theta}) \neq 0$

(if $E(\hat{\theta}) \neq \theta$)

$\hat{\theta}$ is called a BIASED ESTIMATOR.

(1) Sample mean $\longrightarrow$ unbiased estimator for the population

Given $n$ measurements constituting a sample, the sample mean is the arithmetic average of the values obtained for the random variable.

$n$ values: $x_1, x_2, \ldots, x_n$

$$\boxed{\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)}$$

$$\bar{X} = \boxed{\mu} = \int_{-\infty}^{\infty} x f(x)\, dx.$$

true value of the population mean (estimand)

$$E(\bar{x}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} \times n\mu = \boxed{\mu}$$

each measurement is a random variable with the same underlying distribution

$$B(\bar{x}) = E(\bar{x}) - \mu = \mu - \mu = \boxed{0}$$

(2) Sample variance: is used as an estimator for the population variance.

$$\rightarrow s^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$\bar{x}$ → sample mean.

$s^2$ is an unbiased estimator of $\sigma^2$ (population variance).

The factor of $(n-1)$ appears as the number of degrees of freedom after defining the sample mean is $(n-1)$.

$\tilde{s}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ : uncorrected sample variance (biased estimator)

③ Sample standard deviation is the square root of the sample variance.

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2} \qquad \tilde{s} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

(stdev.p) (stdev) in Excel

④ Sample covariance

For two random variables $X$ and $Y$, if there are $n$ datapoints $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

$$\boxed{q_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}$$

$$\begin{array}{l|cccc|}
1^{st}\text{ datapoint} \rightarrow & x_{11} & x_{12} & \cdots & x_{1p} \\
 & x_{21} & x_{22} & \cdots & x_{2p} \\
 & \vdots & & & \\
n^{th}\text{ datapoint} \rightarrow & x_{n1} & x_{n2} & \cdots & x_{np} \\
 & \uparrow & & & \uparrow \\
 & 1^{st}\text{ variable} & & & n^{th}\text{ variable.}
\end{array}$$

$$\vec{\bar{x}} = (\bar{x}_1 \;\; \bar{x}_2 \;\; \cdots \;\; \bar{x}_p) = \left(\frac{1}{n}\sum_{i=1}^{n} x_{i1} \;\; \frac{1}{n}\sum_{i=1}^{n} x_{i2} \;\; \cdots \;\; \frac{1}{n}\sum_{i=1}^{n} x_{ip}\right) = \frac{1}{n}\sum_{i=1}^{n}(x_{i1} \;\; x_{i2} \;\; \cdots$$

$$\boxed{\vec{\bar{x}} = \frac{1}{n}\sum_{i=1}^{n}\vec{x}_i}$$

$$\boxed{s_{jk} = s_{kj}} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$$

$$j, k = 1, 2, 3, \ldots, \boxed{p}.$$

$$S = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & & & \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix}$$

Sample covariance matrix

diagonal entries are the sample variances

symmetric matrix since $s_{jk} = s_{kj}$.

$$S = \begin{pmatrix} \frac{1}{n-1}\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2 & \frac{1}{n-1}\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i1}-\bar{x}_2) & \cdots & \frac{1}{n-1}\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{ip}-\bar{x}_p) \\ \vdots & & & \\ & & & \frac{1}{n-1}\sum_{i=1}^{n}(x_{ip}-\bar{x}_p)^2 \\ \frac{1}{n-1}\sum_{i=1}^{n}(x_{ip}-\bar{x}_p)(x_{i1}-\bar{x}_1) & \cdots & & \end{pmatrix}_{p\times p}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} \begin{pmatrix} (x_{i1}-\bar{x}_1) \\ (x_{i2}-\bar{x}_2) \\ \vdots \\ (x_{ip}-\bar{x}_p) \end{pmatrix}_{p\times 1} \begin{pmatrix} (x_{i1}-\bar{x}_1) & (x_{i2}-\bar{x}_2) & \cdots & (x_{ip}-\bar{x}_p) \end{pmatrix}_{1\times p}$$

$$\boxed{S} = \boxed{\frac{1}{n-1}\sum_{i=1}^{n}(\vec{x}_i - \vec{\bar{x}})^T \cdot (\vec{x}_i - \vec{\bar{x}})}$$

Sample covariance matrix.