STATISTICAL SAMPLE
(and statistical population)

a population . . .

is the set of all the objects (or the totality of observation results with which we are concerned) that posses the property X (our RV), whether their number be finite or infinite.
The statistical inference consists in arriving at (quantitative) conclusions concerning a population where it is impossible or impractical to examine the entire set of observations that make up the population. Instead, we depend on a **subset** of observations — a **sample**.

Property (RV) X — has the pdf: $f(X)$
We may form n samples, each of size m:

$$\begin{array}{rccccc} \mathrm{RV} X: & X_1 & X_2 & \dots & X_m \\ \text{Sample No: } 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 2 & x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots & \dots \\ j & x_{j1} & x_{j2} & \dots & x_{jm} \\ \dots & \dots & \dots & \dots & \dots \\ n & x_{n1} & x_{n2} & \dots & x_{nm} \end{array}$$

every $j-$th sample has a pdf: $g_j = g_j(x_{j1}, x_{j2}, \dots, x_{jm})$

A RANDOM SAMPLE...

...is a sample for which:

1. All sample constituents X_{jl} are independent of each other:

$$g_j(x_{j1}, x_{j2}, \ldots, x_{jm}) = g_{j1}(x_{j1}) g_{j2}(x_{j2}) \ldots g_{jm}(x_{jm})$$

2. have the same pdf as the X RV:

$$g_{j1}(x_{j1}) = g_{j2}(x_{j2}) = \ldots = g_{jm}(x_{jm}) = f(x)$$

A **RANDOM SAMPLE**, cntd.

A **RANDOM SAMPLE**, cntd.

The distribution (density) function of the RV X in our population should be written not as $f(x)$ but rather as

$$f = f(x; \boldsymbol{\lambda})$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_p)$ — denotes the set of p *distribution parameters* (e.g.: μ, σ, etc.)

How do we go about finding λ ?

How do we go about finding $\boldsymbol{\lambda}$?

Any function of the random variables constituting a random sample that is used for **estimation** of unknown distribution parameters $\boldsymbol{\lambda}$ is called a

statistic: $\mathsf{S} = \mathsf{S}(X_1, X_2, \ldots, X_N)$

BUT

How do we go about finding $\boldsymbol{\lambda}$?

Any function of the random variables constituting a random sample that is used for **estimation** of unknown distribution parameters $\boldsymbol{\lambda}$ is called a **statistic**: $\mathsf{S} = \mathsf{S}(X_1, X_2, \ldots, X_N)$

BUT — ANY STATISTIC S is also a RV!

STATISTIC, cntd.

$$\lambda_i \stackrel{?}{=} \mathsf{S}(X_1, X_2, \ldots, X_N) \quad \text{NO!}$$

$$\lambda_i = E[\mathsf{S}(X_1, X_2, \ldots, X_N)] \equiv \hat{\mathsf{S}}$$

more precisely:

$$\hat{\mathsf{S}} \to \hat{\mathsf{S}}_N \ \text{ or simply } \ \hat{\lambda}_N .$$

We say: the estimated value of a statistic $\hat{\mathsf{S}}$ is said to be **estimator** of the parameter(s) λ; the estimation is carried out on the basis of an N–element sample.

PROPERTIES OF ESTIMATORS

a good estimator should be

1. **consistent**:

$$\vee \varepsilon > 0 \qquad \lim_{N\to\infty} \mathcal{P}(|\hat{\boldsymbol{\lambda}}_N - \boldsymbol{\lambda}| < \varepsilon) = 1$$

2. **unbiased**:

$$\vee N \quad E(\boldsymbol{\lambda}_N) = \boldsymbol{\lambda} \quad \text{or}$$

$$\text{the bias:} \;\; B_N(\boldsymbol{\lambda}) = E(\boldsymbol{\lambda}_N) - \boldsymbol{\lambda} = 0.$$

The concept of **asymptotically unbiased estimators**:

$$\lim_{N\to\infty} B_N(\boldsymbol{\lambda}) = 0.$$

3. **efficient**:

two estimators: $\hat{\lambda}_N^*$ and $\hat{\lambda}_N^{**}$ and we have

$$\mathrm{VAR}\left(\hat{\lambda}_N^*\right) < \mathrm{VAR}\left(\hat{\lambda}_N^{**}\right)$$

then: $\hat{\lambda}_N^*$ is more efficient than $\hat{\lambda}_N^{**}$. The estimator which has the lowest variance VAR — is called the most efficient ESTIMATOR