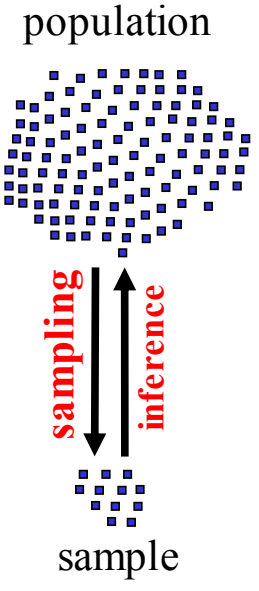
Drawing samples from the population and sampling distribution of means

Prof. Giuseppe Verlato
Unit of Epidemiology & Medical Statistics,
Department of Diagnostics & Public Health,
University of Verona



Sampling = procedures for drawing samples from the population.

Statistical inference = procedures of drawing conclusions about the source population from the sample. Generalizing the results obtained from the sample back to the source population.

The sample should be **representative** of the source population.

If passers-by are interviewed in a given street, is the sample representative of the source population ?

No, it isn't. Indeed, the probability of being interviewed changes from one subject (statistical unit) to another as a function of individual habits and it is unknown.

In order to get a representative sample, we need a **sampling frame** *(lista di campionamento)* from which to draw the sample.

We need to exactly know the chances that every unit has of being selected (**probability sampling).**

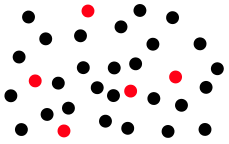
Sampling fraction = Ratio of sample size to population size**.**

In the night after 2004 European elections, a politician disputed the outcome of a sample survey performed by a statistical research organization.

Indeed the politician's party had got 31.5% of votes according to sample-based generalization made by the research organization. However the proportion was slightly higher (34.7%) after counting 20% of total votes.

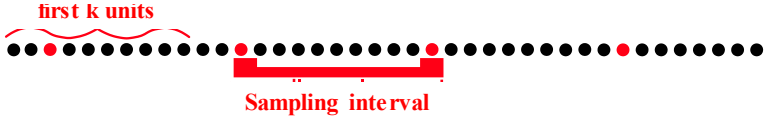
The politician marked the difference between statistical projections and real data. The statistician answered back that the 20% of counted votes were not representative of the Italian population.

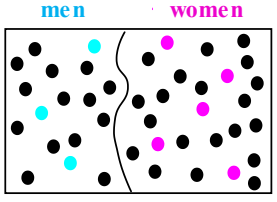
On the following day newspapers reported that the politician's party had actually got 31.5% of votes.



Simple random sampling: all population units have an equal probability of being selected (sampling fraction = **n**/N).

Systematic sampling: the researcher uses natural ordering of the sampling frame, randomly selects one of the **first k** units, and then selects **every k-th unit**. In other words an arbitrary starting point is chosen and then units are drawn at a preselected interval.

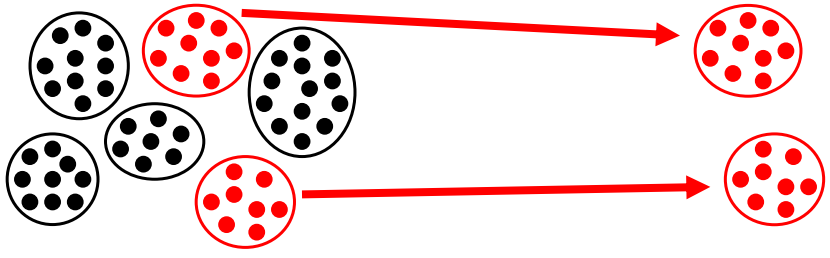




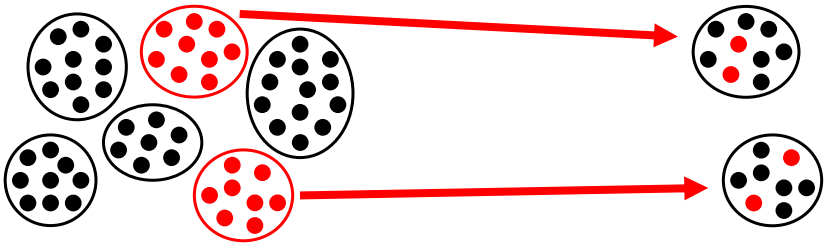
Stratified sampling: a simple random sampling is performed in each population group. Optimal allocation of subsample size n_i **proportional to** $N_i\sigma_i$

CLUSTER SAMPLING:

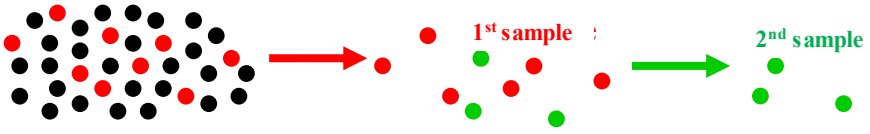
Some 2° level units are randomly selected. All 1° level units enclosed within the extracted 2nd level units are considered for the study.



TWO STAGE SAMPLING: some 2° level units are randomly selected, from which 1° level units are randomly drawn.



TWO PHASE SAMPLING: A 1° sample is randomly drawn which will undergo cheaper investigations (postal questionnaire). From the 1st sample a 2nd sample is randomly drawn which will undergo more expensive investigations (clinical visits, lab tests).



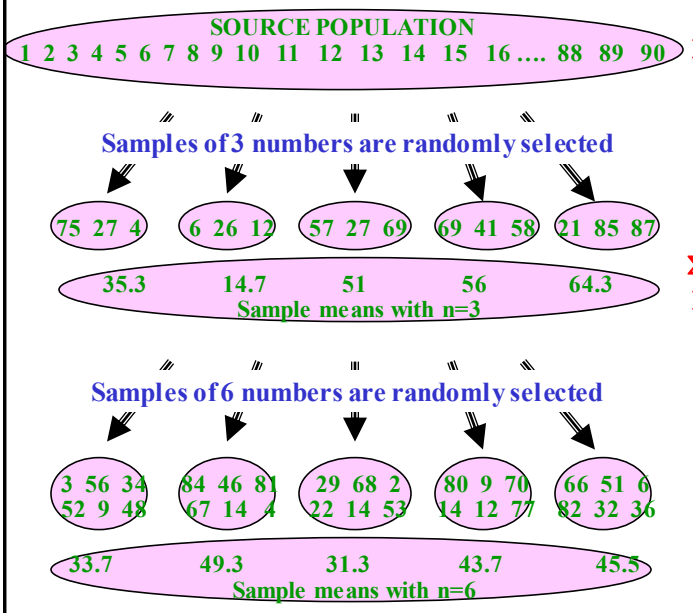
SAMPLING DISTRIBUTION of MEANS

Sample mean is a **random variable**, as:

1) it changes from one sample to another one.

2) the sample is drawn at random.

EXPERIMENTAL ASSESSMENT of the DISTRIBUTION of SAMPLE MEAN - 1



$\Sigma x=4095$ $\mu=45.5$
$\Sigma x^2=247065$ $\sigma=25.98$
uniform distribution

$\Sigma x=221.3$ $\bar{x}_x=44.3$
$\Sigma x^2=11339.3$ $s_x=19.6$

$\Sigma x=203.5$ $\bar{x}_x=40.7$
$\Sigma x^2=8526$ $s_x=7.8$

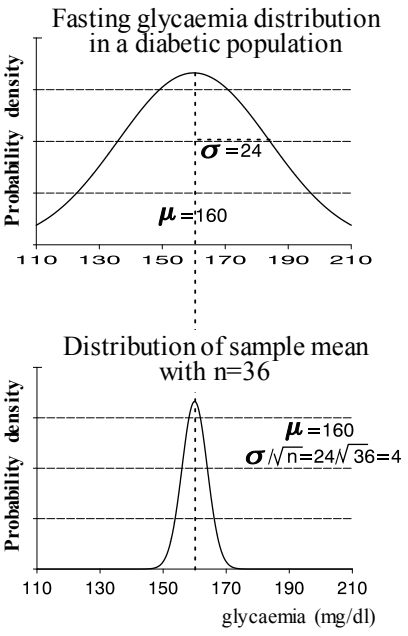
EXPERIMENTAL ASSESSMENT of the DISTRIBUTION of SAMPLE MEAN - 2

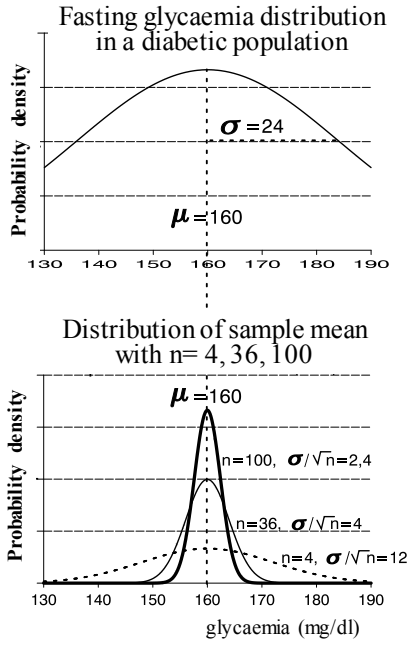
1) The mean of means (44.27 with n=3 and 40.7 with n=6) is approximately equal to the mean of the source population (45.5)

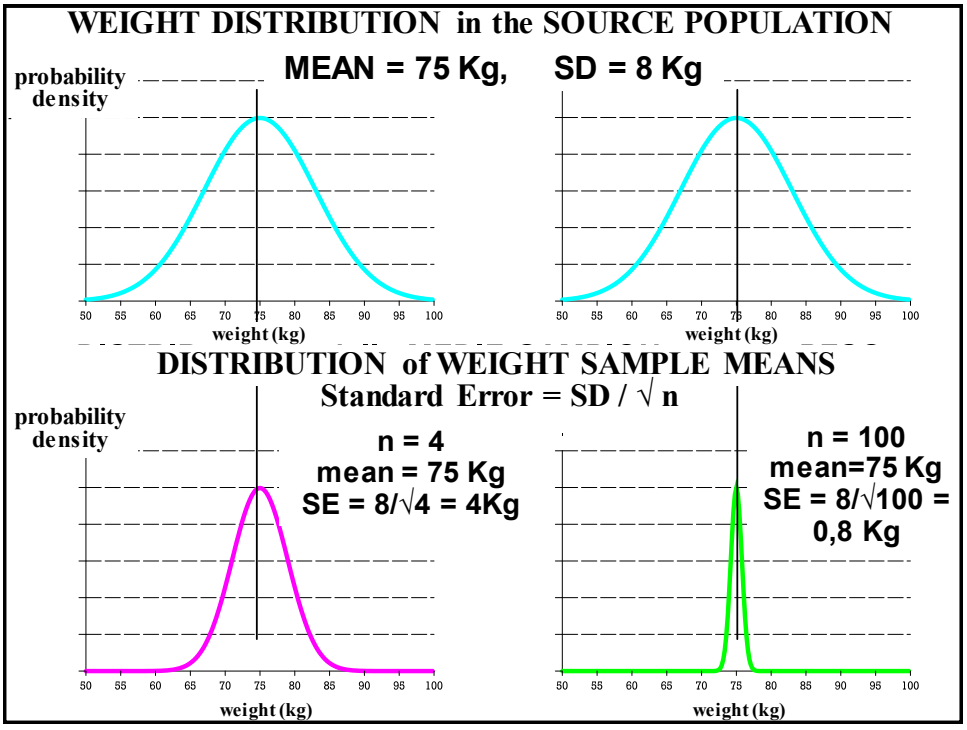
2) The variability of sample means is lower than the variability recorded in the source population: SD_x= Standard Error = $\sigma/\sqrt{n}$

Sample size	observed value	expected value
n = 3	SD_x = 19.63	$25.98/\sqrt{3}=15.00$
n = 6	SD_x = 7.80	$25.98/\sqrt{6}=10.61$

3) As sample size increases, the distribution of sample mean tends to become normal irrespective of the distribution of the variable under study.


Fasting glycaemia distribution
in a diabetic population
Probability density
σ =24
μ =160
110 130 150 170 190 210
Distribution of sample mean
with n=36
Probability density
μ =160
σ /√n=24/√36=4
110 130 150 170 190 210
glycaemia (mg/dl)


Fasting glycaemia distribution
in a diabetic population
Probability density
σ =24
μ =160
130 140 150 160 170 180 190
Distribution of sample mean
with n= 4, 36, 100
Probability density
μ =160
n=100, σ /√n=2,4
n=36, σ /√n=4
n=4, σ /√n=12
130 140 150 160 170 180 190
glycaemia (mg/dl)


WEIGHT DISTRIBUTION in the SOURCE POPULATION
probability density
MEAN = 75 Kg,
SD = 8 Kg
50 55 60 65 70 75 80 85 90 95 100
weight (kg)
50 55 60 65 70 75 80 85 90 95 100
weight (kg)
DISTRIBUTION of WEIGHT SAMPLE MEANS
Standard Error = SD / √ n
probability density
n = 4
mean = 75 Kg
SE = 8/√4 = 4Kg
50 55 60 65 70 75 80 85 90 95 100
weight (kg)
n = 100
mean=75 Kg
SE = 8/√100 = 0,8 Kg
50 55 60 65 70 75 80 85 90 95 100
weight (kg)