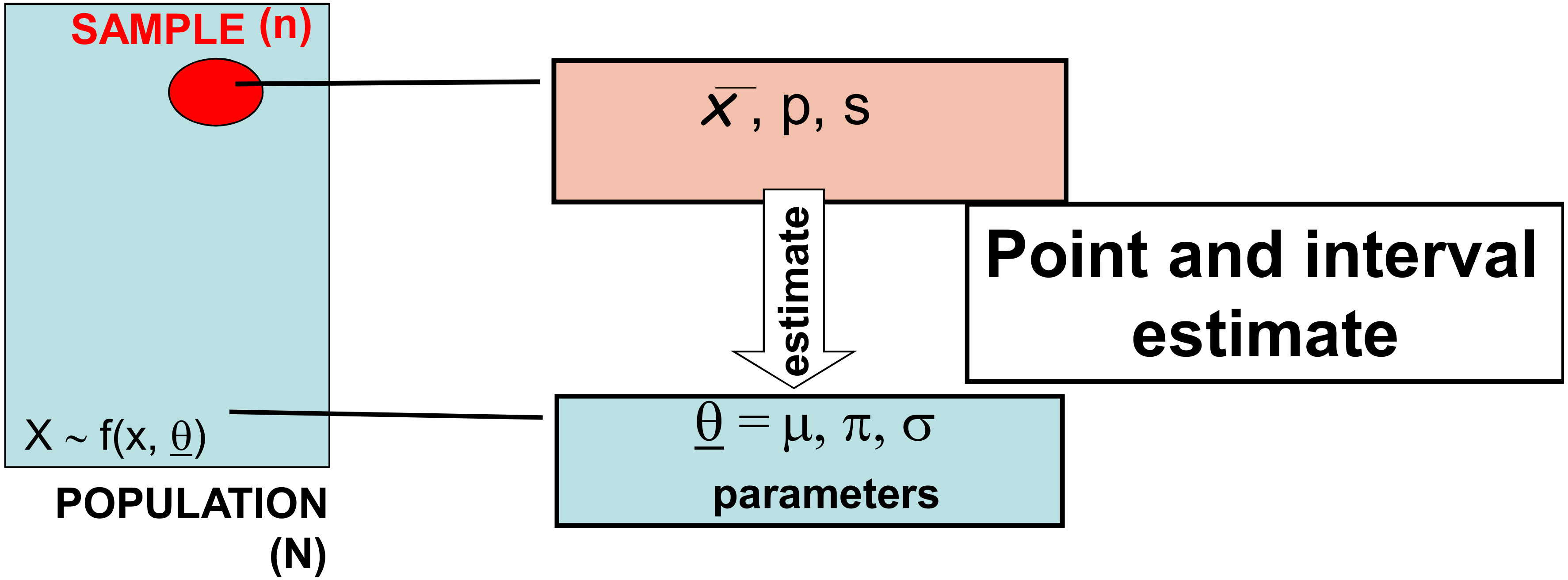
UNIVERSITA' DEGLI STUDI DI MILANO BICOCCA

DIPARTIMENTO DI MEDICINA E CHIURGIA

Sampling distribution and estimators

Inferential statistics

It face up decisional issues using a priori information and the sample data

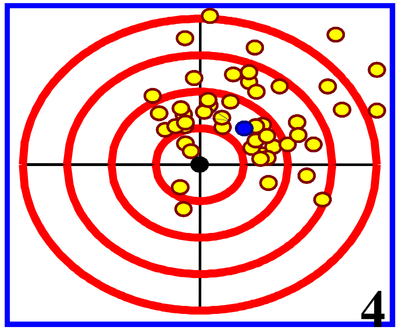

SAMPLE (n)
$\bar{x}$, p, s
estimate
Point and interval estimate
$\underline{\theta} = \mu, \pi, \sigma$
parameters
$X \sim f(x, \underline{\theta})$
POPULATION (N)

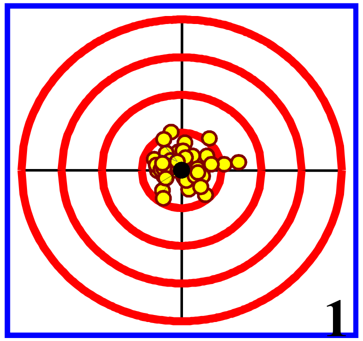
How good are these estimates???

Reliability of the estimates

The aim is to get unbiased (accurate) and precise estimates:

- Avoid systematic errors
- Minimise random errors





It is very important the plan of the experiment!

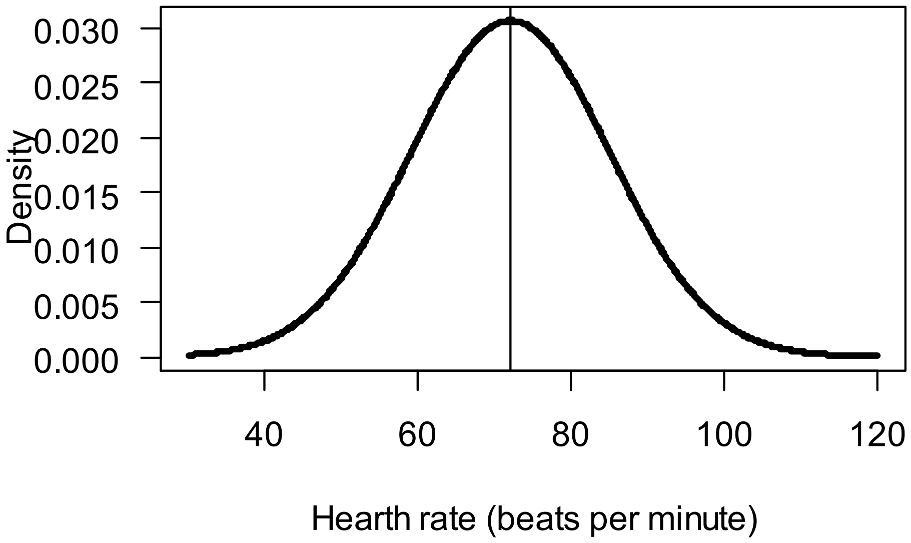
How to evaluate the reliability of the estimate?

Not possible from the single sample. We have to consider a theoretical situation in which we take from a known population all possible samples of the same size n.

Example:

Italian adults have pulse rates with a mean of 72 bpm (beats per minute), a standard deviation of 13 bpm, and a distribution that is approximately normal.

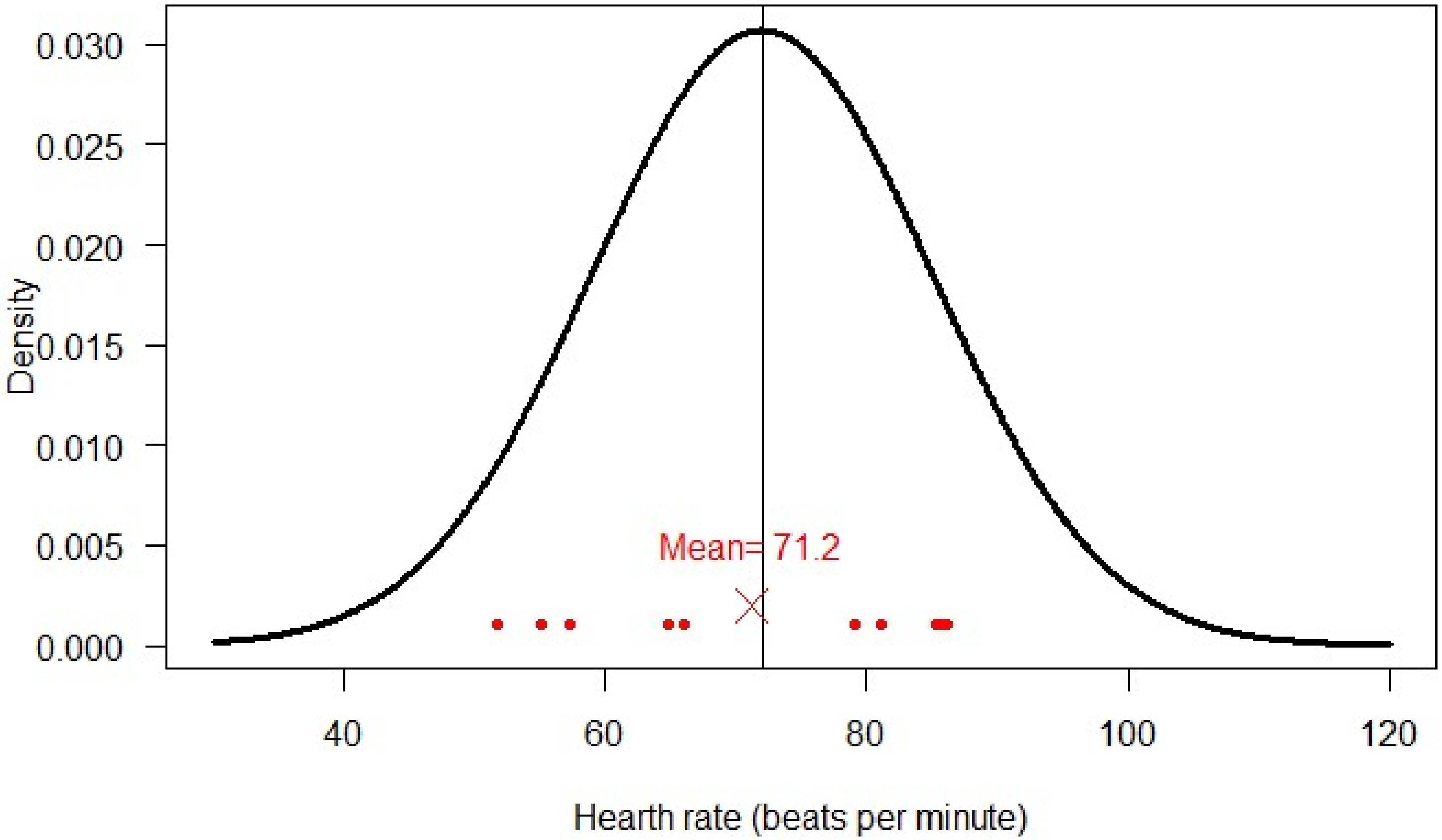
N(μ=72, σ=13)



If we select a random sample of 10 adults in Italy how do we expect to be the mean of their pulse rate?

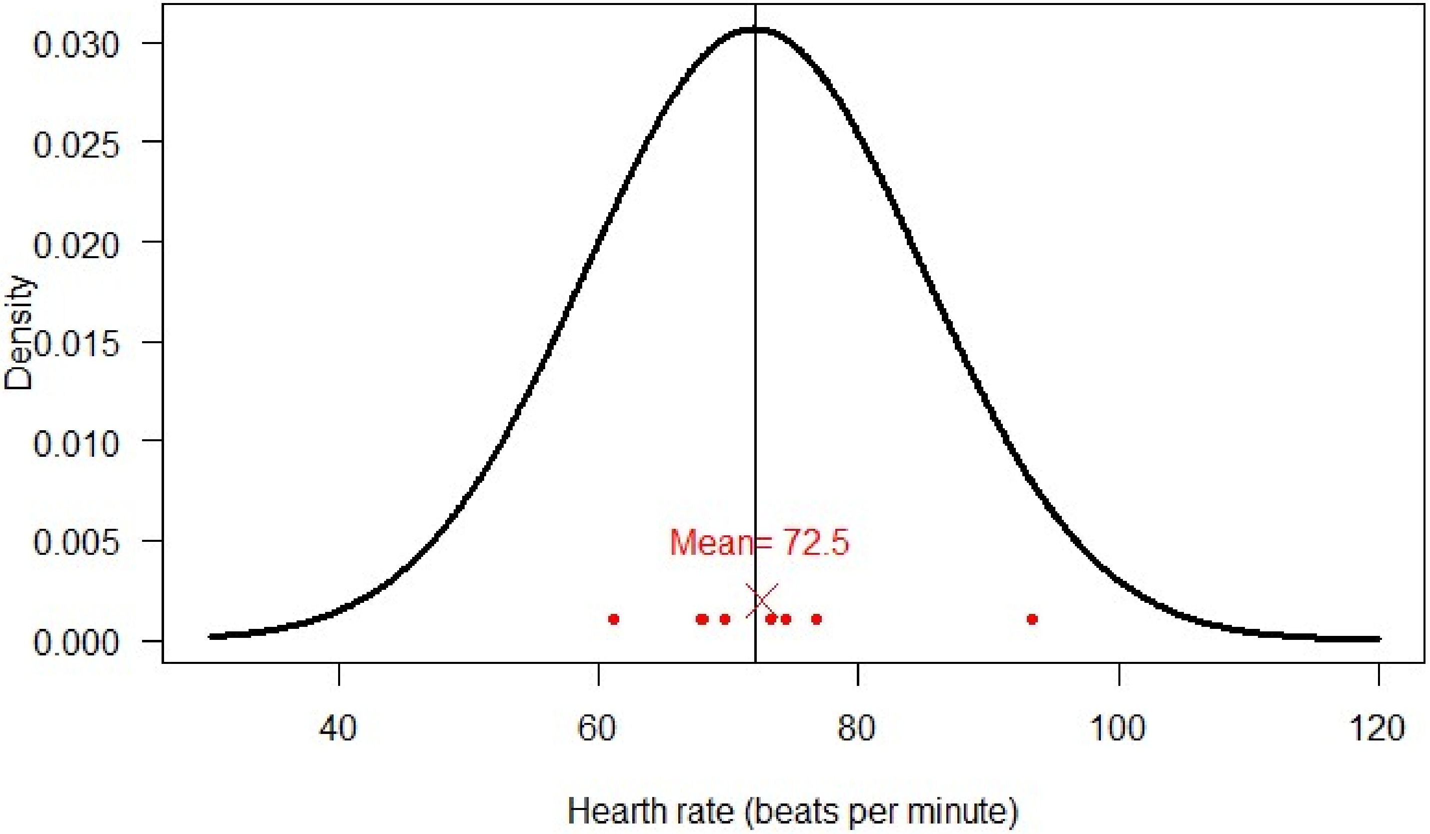
The sample mean distribution

Let’s take a random sample with n=10
Sample: 86 79 86 81 57 55 66 85 52 65
Mean=71.2



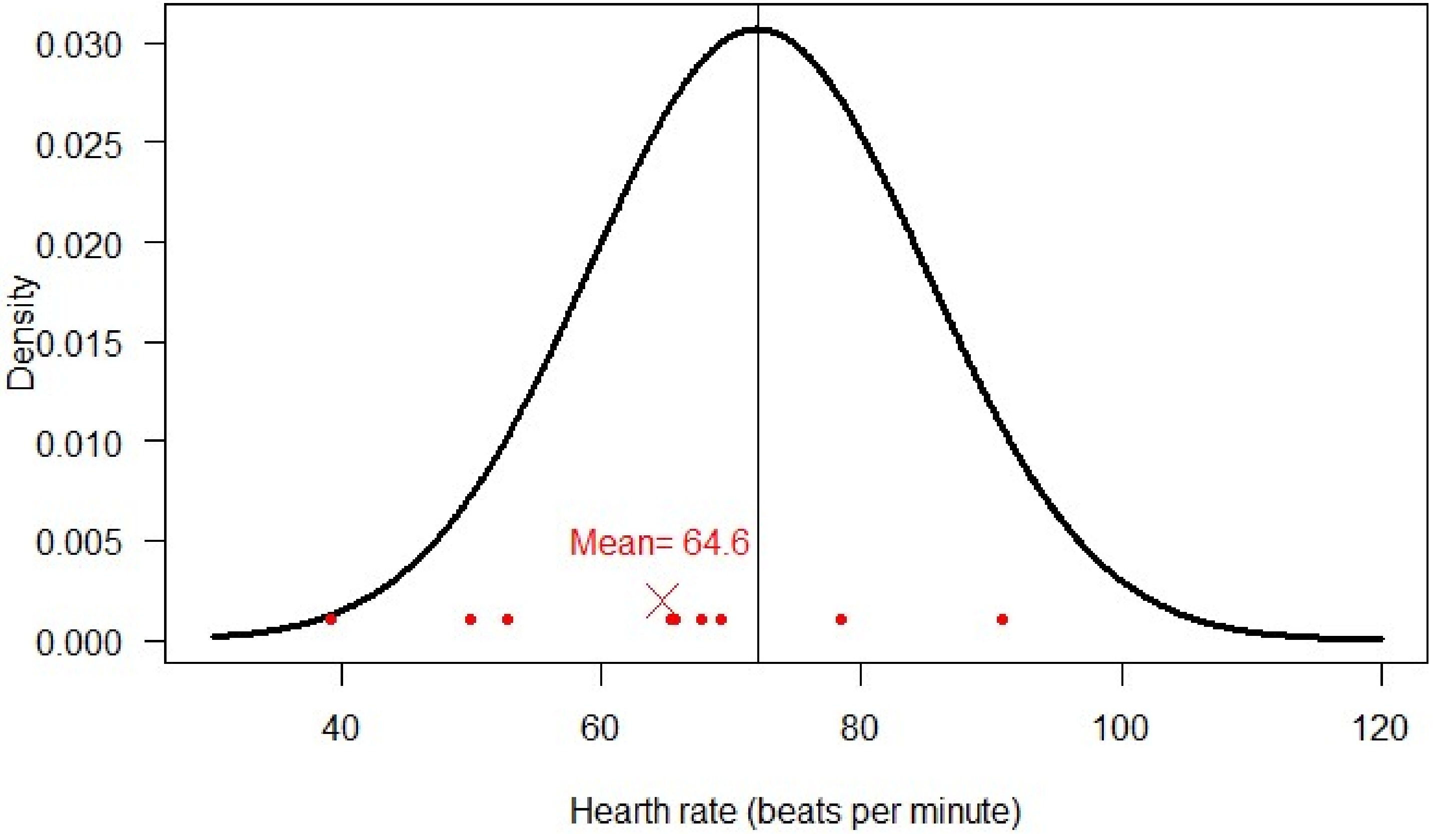
The sample mean distribution

Let’s take another random sample with n=10 (with replacement)
Sample: 73 68 61 68 74 70 93 73 68 77
Mean=72.5



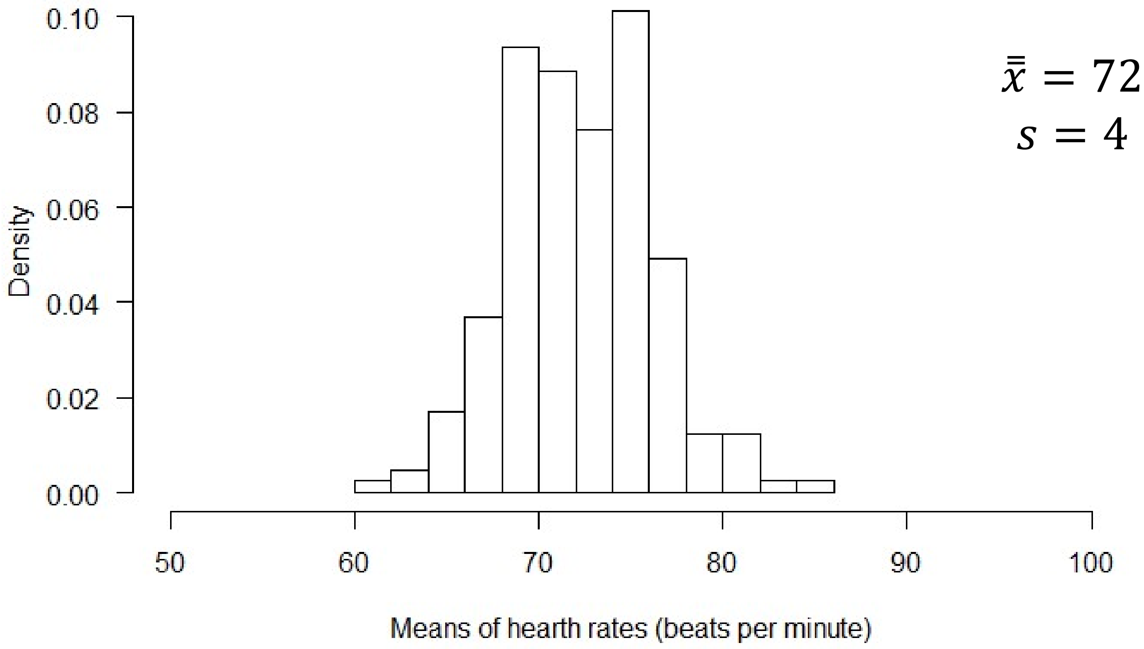
The sample mean distribution

Let’s take another random sample with n=10 (with replacement)
Sample: 91 69 50 39 66 53 65 68 78 68
Mean=64.6



The sample mean distribution

Let’s get many samples with n=10 (with replacement).
Here is the distribution of their sample mean:
They have a mean of 72 and a standard deviation of 4



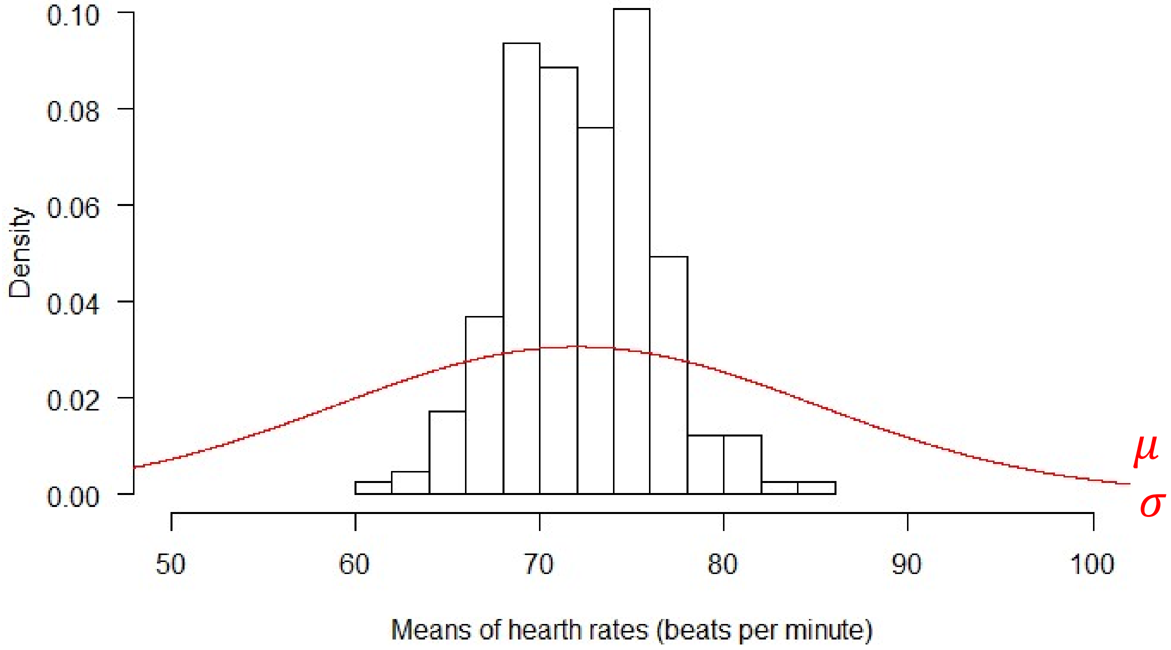
$$\bar{\bar{x}} = 72$$
$$s = 4$$

Is the mean of a (random) sample of size 10 reliable?
Is it biased? Is it precise?

Heart rate distribution vs sample mean distribution

$\bar{\bar{x}} = 72$

$s = 4$



$\mu = 72$

$\sigma = 13$

- The distribution of sample means tends to be a normal distribution.
- The sample means target the value of the population mean (unbiased).
- The standard deviation of the means is much lower than the standard deviation of the population

Standard error: the precision of the estimate

The standard deviation of the estimate is called STANDARD ERROR

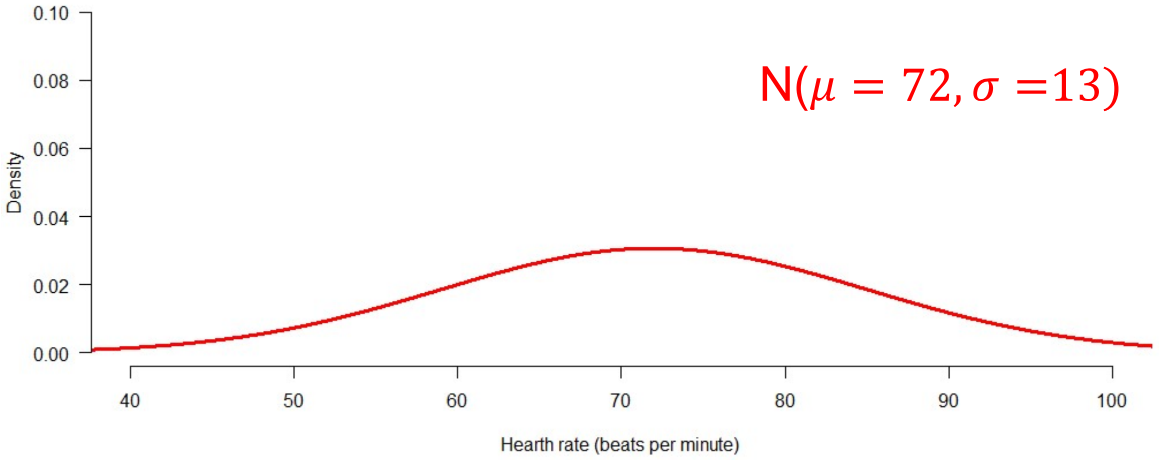
$$\text{standard error} = \sigma/\sqrt{n}$$

- It quantifies the uncertainty of the sample mean, thus its precision
- Its value depends from:
  1. The characteristics of the variable measured on the population, in particular from the degree of variability (σ);
  2. from the sample size (n);
  3. from the sampling strategy.

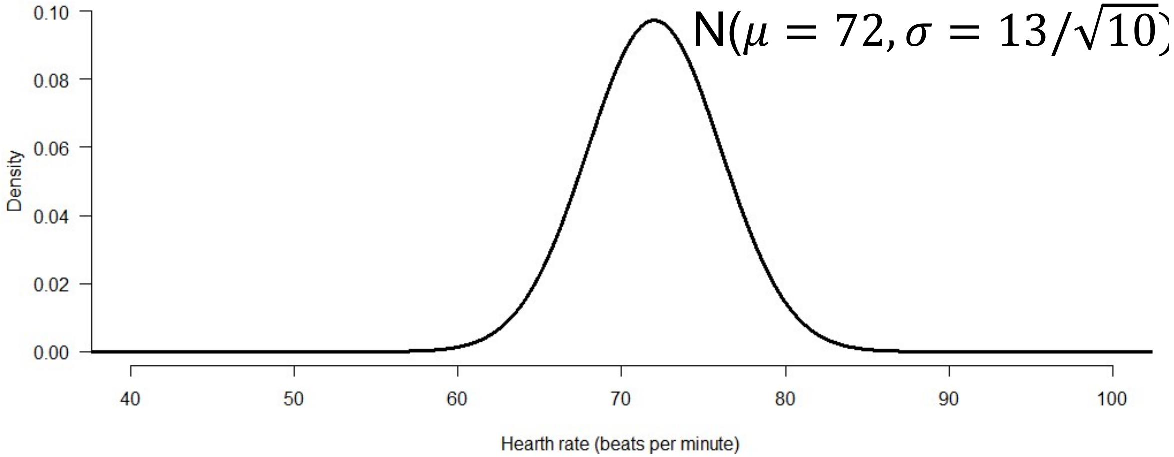
In the hearth rate example $standard\ error = \frac{\sigma}{\sqrt{n}} = \frac{13}{\sqrt{10}} = 4.1$

Standard error: the precision of the estimate

- **STANDARD DEVIATION:** Variability index of the characteristic. Indicates the difference among individuals, e.g. how different are hearth rates among Italian adults.



- **STANDARD ERROR:** Variability index of the estimate. Indicates the possible difference among different estimates, e.g. how different are sample means of hearth rates in samples of size 10.



Reliability of the estimates

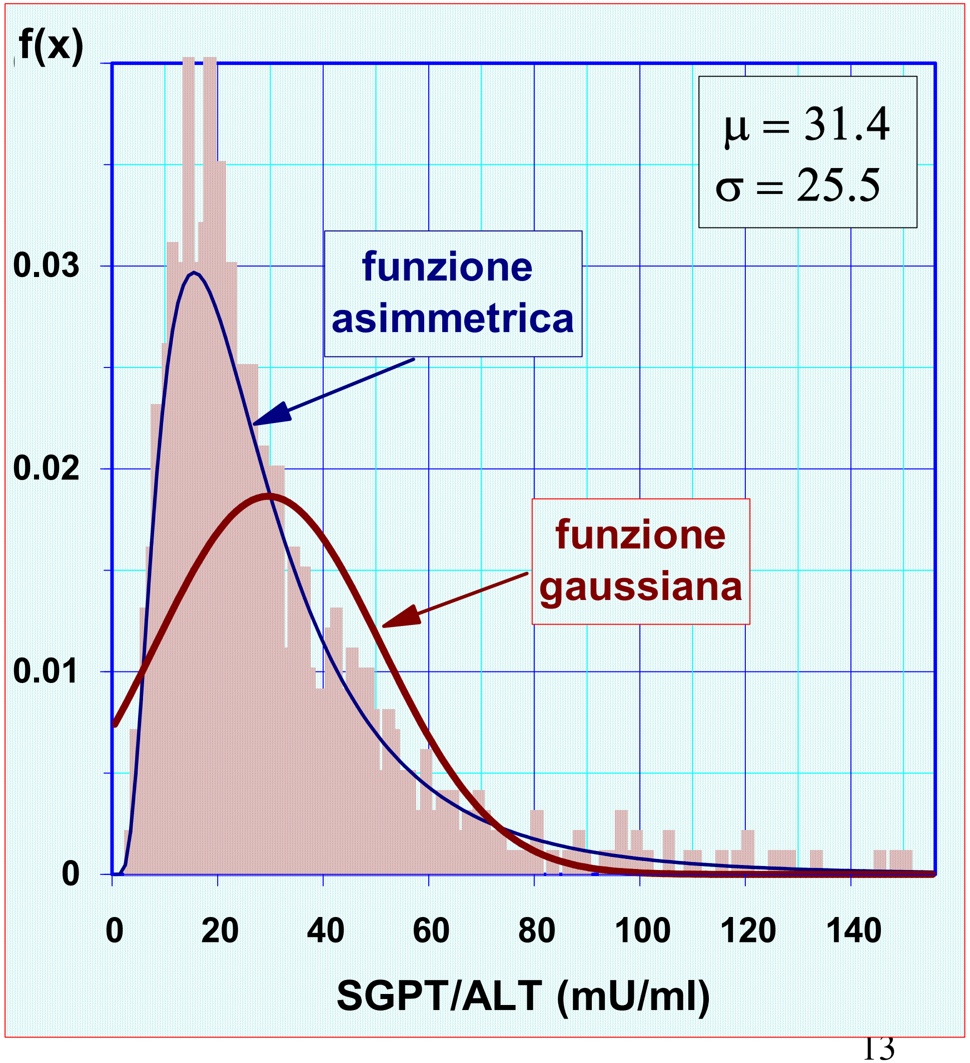
Not possible from the single sample. We have to consider a theoretical situation in which we take all possible samples of the same size n from a known population.

The inference process is thus based on knowledge on theoretical characteristics of the distribution of the sample estimator.

Example

We started from a nomal distribution (hearth rate) – but what about not normal variables?

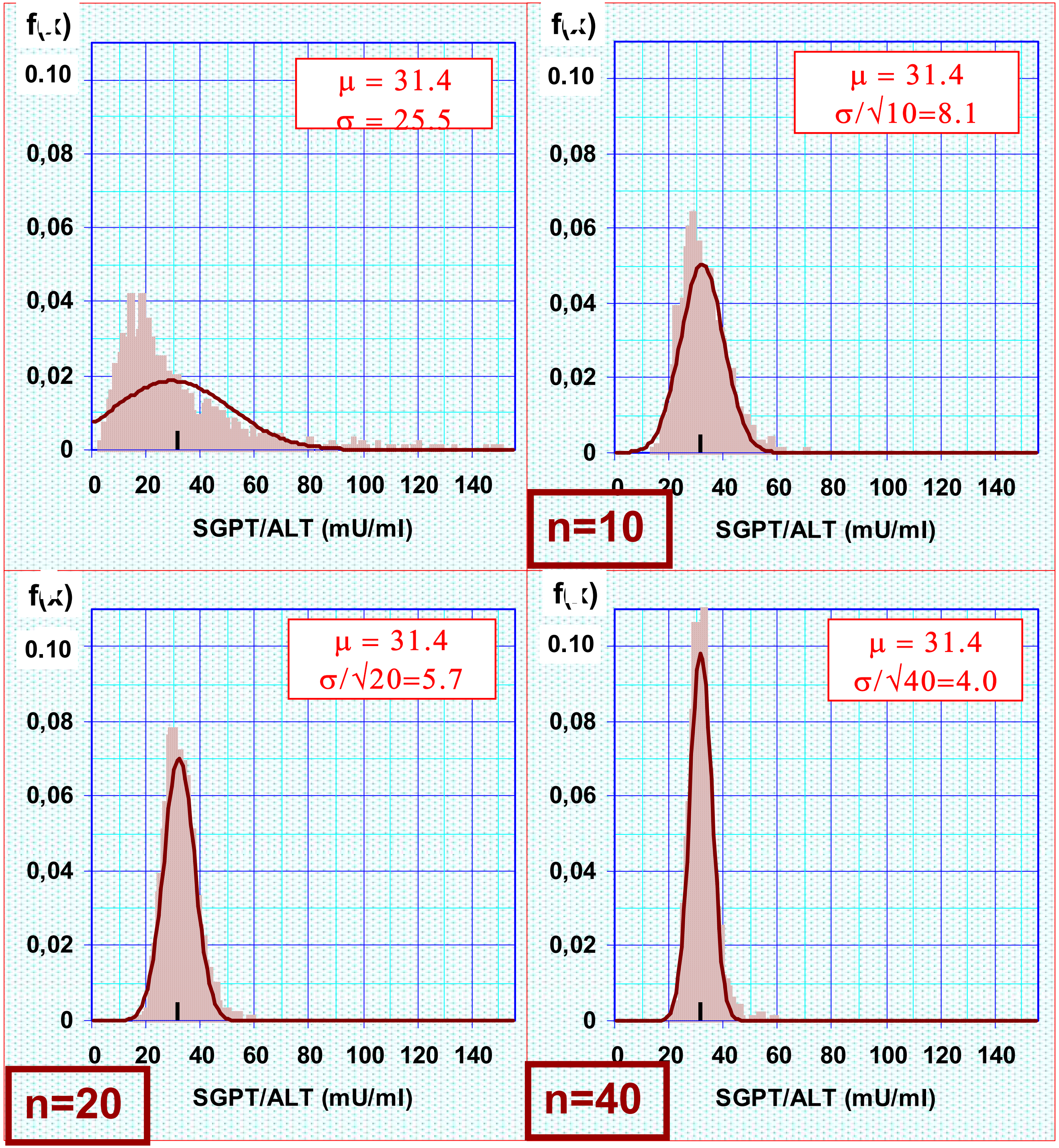
In the population of adult males, the distribution of alanine amino-transferase (ALT) is strongly asymmetric due to the presence of individuals with liver damage caused by alcohol, drugs, viral infections ...



Example

As the sample size increases, the distribution of the sample mean:

- reduces its dispersion;
- tends to become closer to a normal distribution



Central Limit Theorem

For all samples of the same size n (with n>30), the sampling distribution of the mean of a random variable with mean μ and standard deviation σ can be approximated by a normal distribution with mean μ and standard deviation σ/√n.

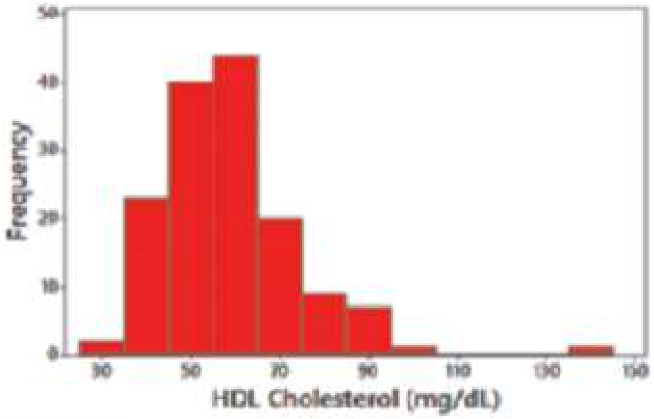


FIGURE 6-21 Nonnormal Distribution: HDL Cholesterol from 147 Women

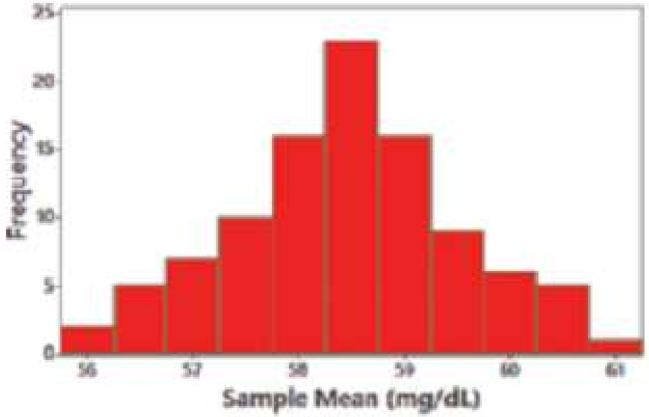


FIGURE 6-22 Approximately Normal Distribution: Means from Samples of Size *n* = 100 of HDL Cholesterol from Females

Triola, Marc M. *Biostatistics for the Biological and Health Sciences, Global Edition, 2nd Edition*. Pearson (Intl),

Key Concept

The *Central Limit Theorem* tells us that for a population with *any* distribution, the distribution of the sample means approaches a normal distribution as the sample size increases.

The procedure in this lecture form the foundation for estimating population parameters and hypothesis testing.

Central Limit Theorem

Given:

1. **The random variable *x* has a distribution (which may or may not be normal) with mean *μ* and standard deviation σ.**
2. **Simple random samples all of size *n* are selected from the population. (The samples are selected so that all possible samples of the same size *n* have the same chance of being selected.)**

Central Limit Theorem – cont.

Conclusions:

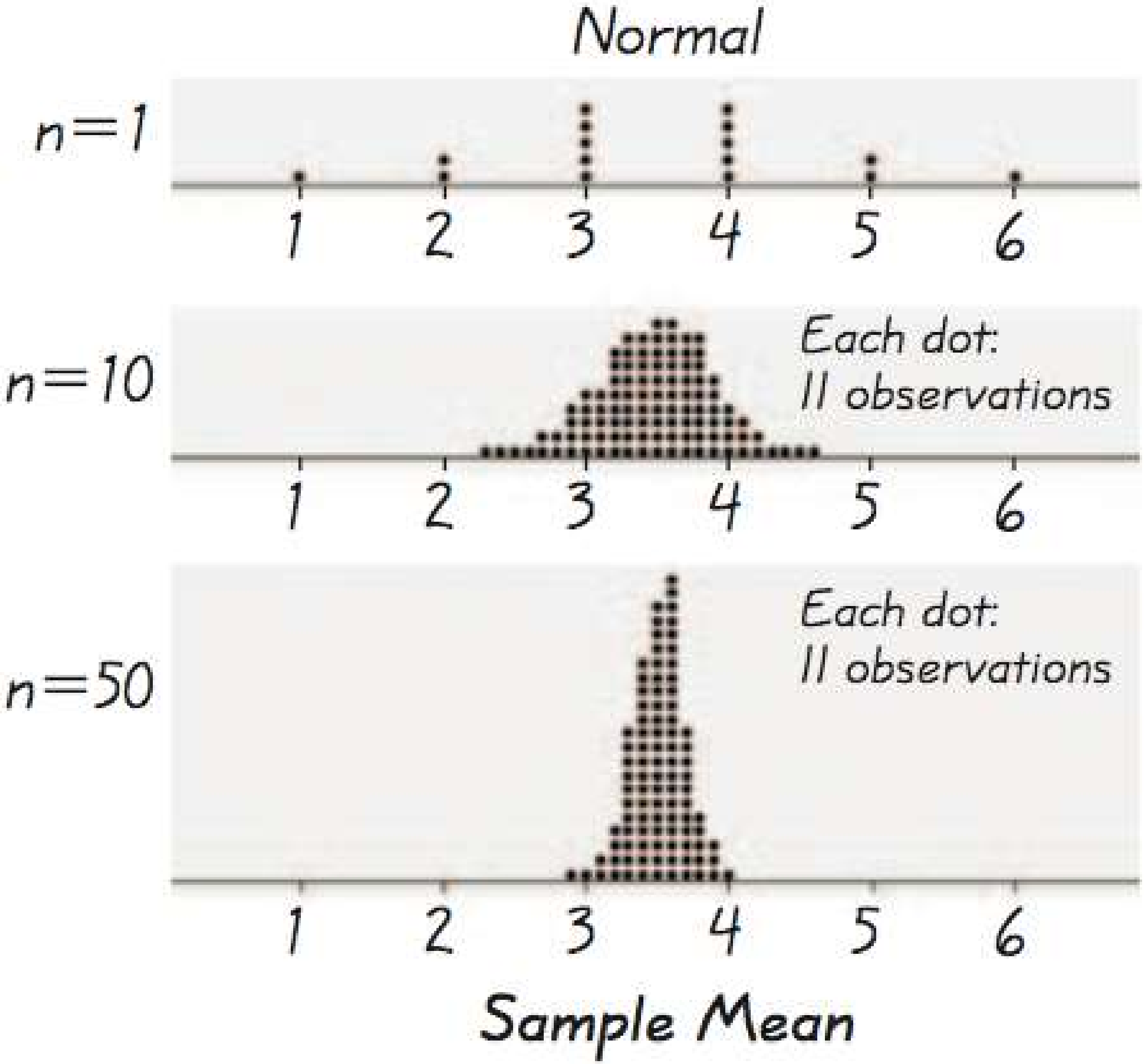
1. The distribution of sample $\bar{x}$ will, as the sample size increases, approach a normal distribution.

2. The mean of the sample means is the population mean μ.

3. The standard deviation of all sample means is $\sigma/\sqrt{n}$.

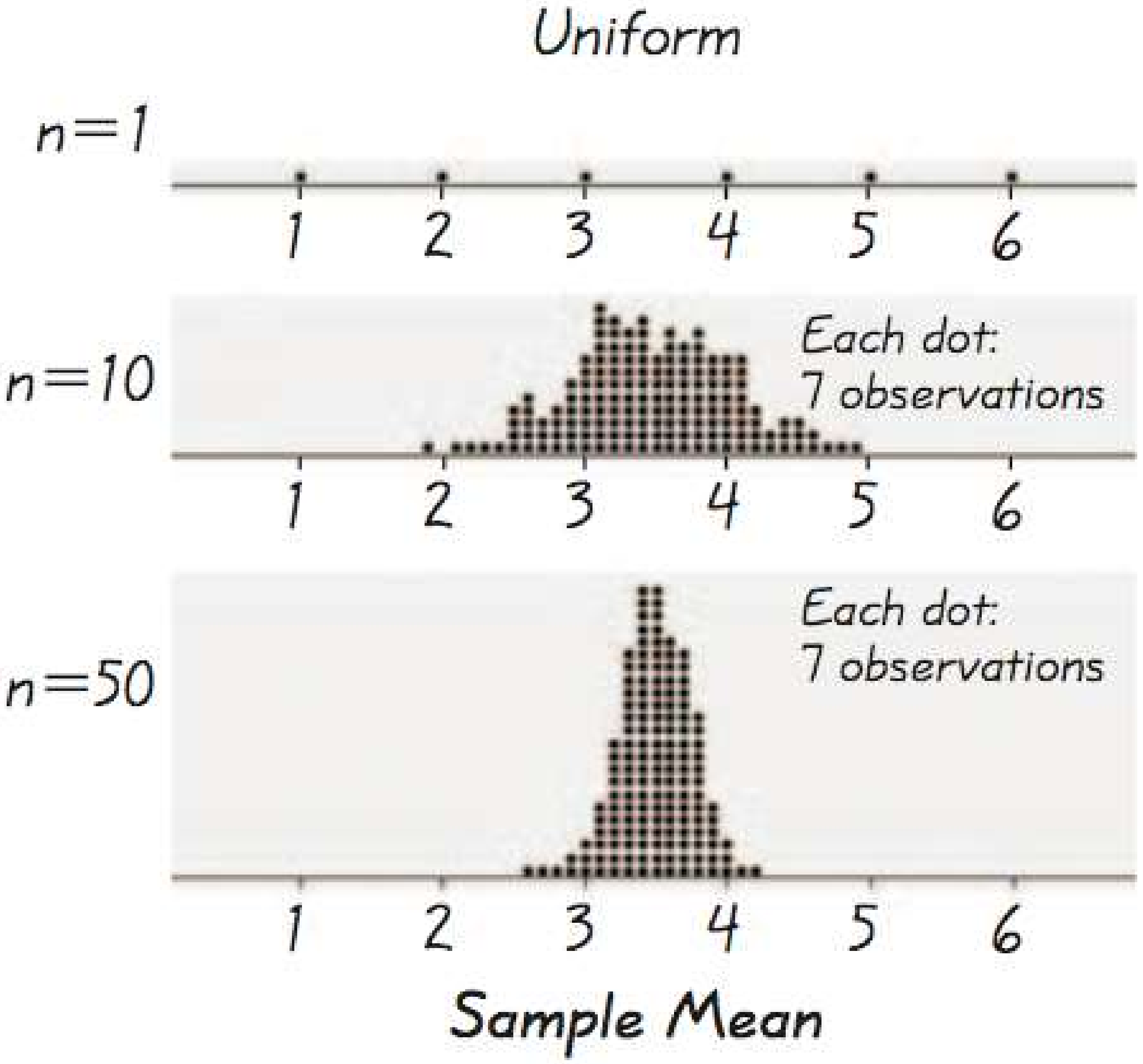
Example - Normal Distribution

As we proceed from *n* = 1 to *n* = 50, we see that the distribution of sample means is approaching the shape of a normal distribution.



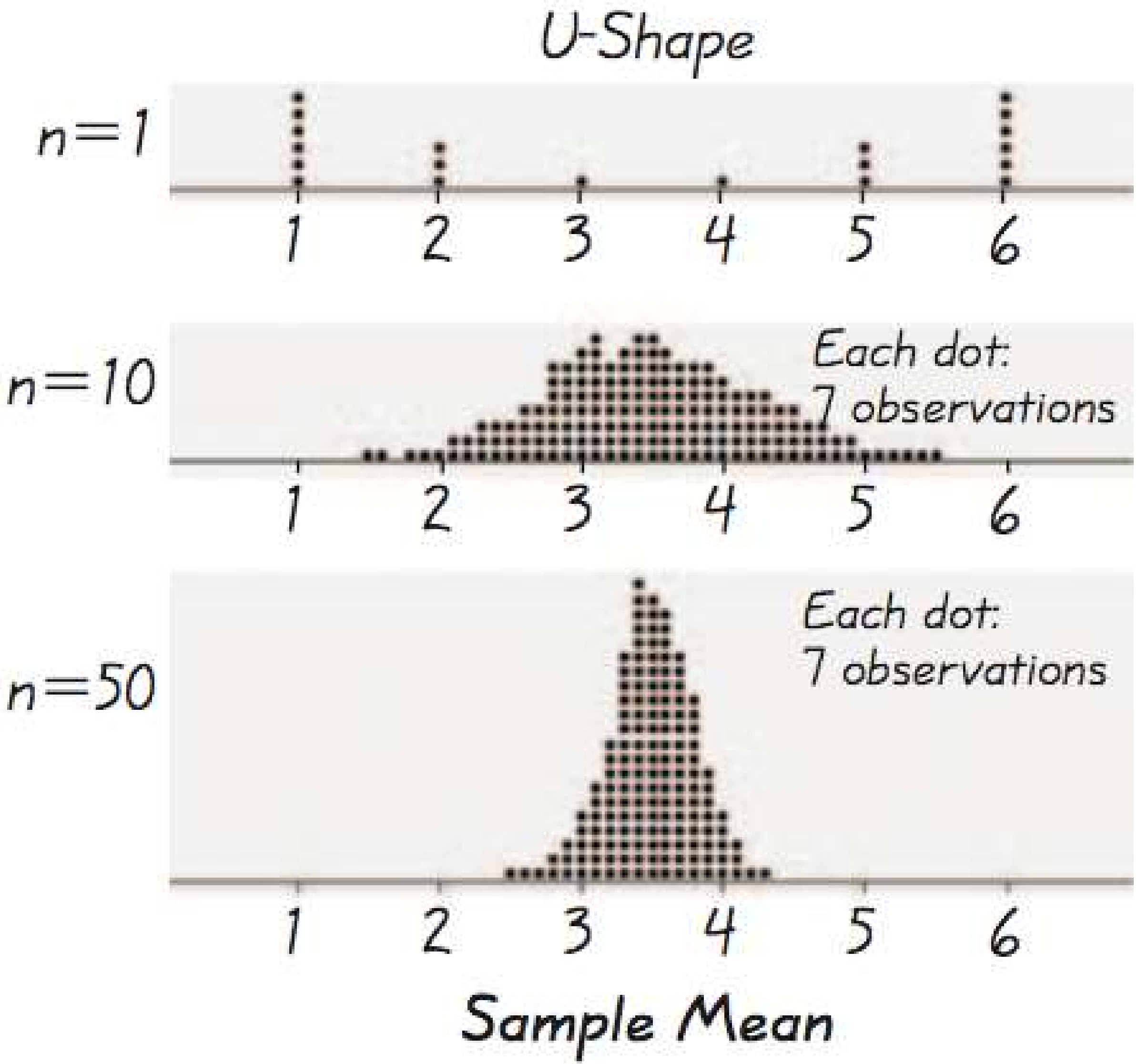
Example - Uniform Distribution

As we proceed from *n* = 1 to *n* = 50, we see that the distribution of sample means is approaching the shape of a normal distribution.



Example - U-Shaped Distribution

As we proceed from *n* = 1 to *n* = 50, we see that the distribution of sample means is approaching the shape of a normal distribution.



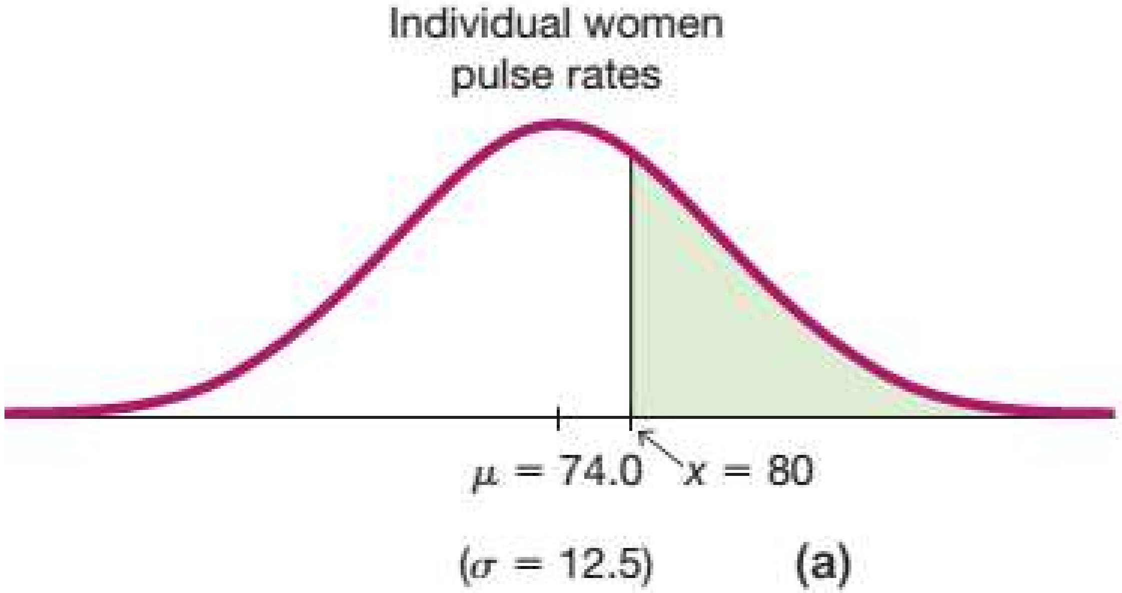
Exercise

Women have normally distributed pulse rates with a mean of 74.0 bpm and a standard deviation of 12.5 bpm.

a. Find the probability that 1 randomly selected woman has a pulse rate greater than 80 bpm.
b. Find the probability that a sample of 16 randomly selected women have a **mean** pulse rate greater than 80 bpm.
c. Find the interval of pulse rate in which we expect 95% of women
d. Find the interval of pulse rate in which we expect 95% of the means of samples of size 16

Solution

a. Find the probability that 1 randomly selected woman has a pulse rate greater than 80 bpm.

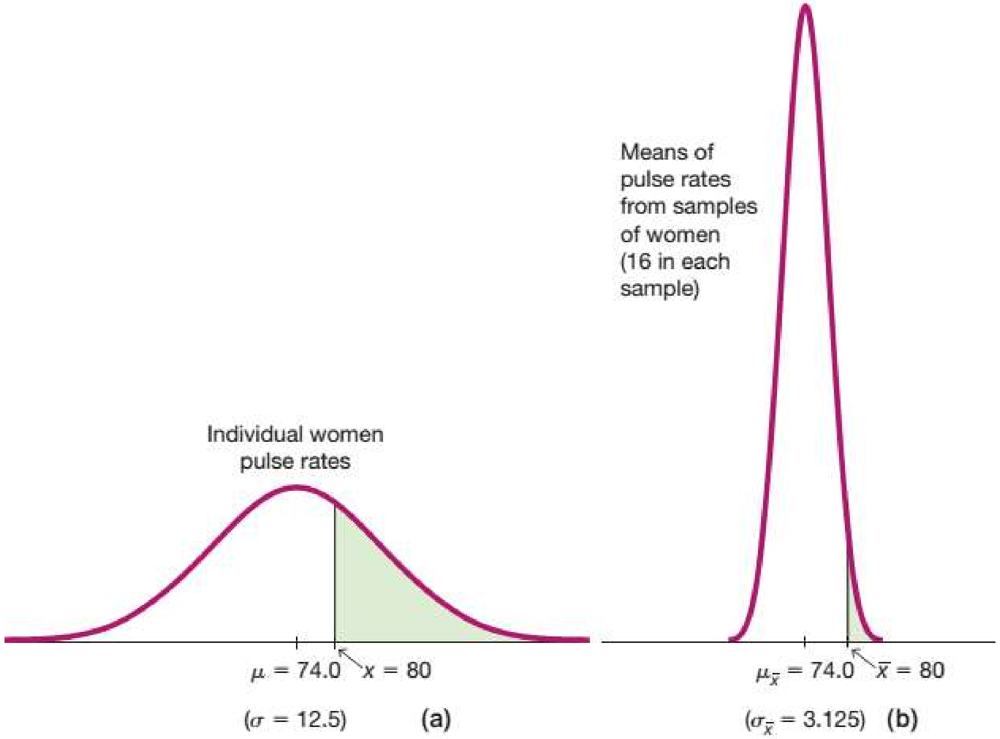


$$z = \frac{x-\mu}{\sigma} = \frac{80-74}{12.5} = 0.48$$

P(X>80)=P(Z>0.48)= 0.3156.

Solution

b. Find the probability that a sample of 16 randomly selected women have a mean pulse rate greater than 80 bpm.



$$z = \frac{x-\mu}{\sigma/\sqrt{n}} = \frac{80-74}{12.5/\sqrt{16}} = 1.92$$

$P(\bar{X} > 80) =$ P(Z>1.92)= 0.0274

Solution

c. Find the interval of pulse rate in which we expect 95% of women

$z_{0.975} = 1.96$

$z = \frac{x-\mu}{\sigma}$ -> $x = \mu + z\sigma$

$x1 = 74 - 1.96 * 12.5$ =49.5

$x2 = 74 + 1.96 * 12.5$ =98.5

95% of women are expected to have pulse rate between 49.5 and 98.5 bpm

Solution

d. Find the interval of pulse rate in which we expect 95% of the means of samples of size 16

$$z = \frac{x - \mu}{\sigma/\sqrt{n}}$$

$x = \mu + z\sigma/\sqrt{n}$

$x1 = 74 - 1.96 * 12.5/\sqrt{16}$ =67.9

$x2 = 74 + 1.96 * 12.5/\sqrt{16}$ =80.1

In samples of size 16, the mean of pulse rate of women is expected to range between 67.9 and 80.1 bpm 95% of the times

Sample proportion distribution

Can you guess the proportion of women?

How reliable is the proportion of women in this sample as compared with the proportion in the whole popuation?

Not possible evaluate it from the single sample!
We have to consider a theoretical situation in which we take all possible samples of the same size *n* from a known population .

Example:

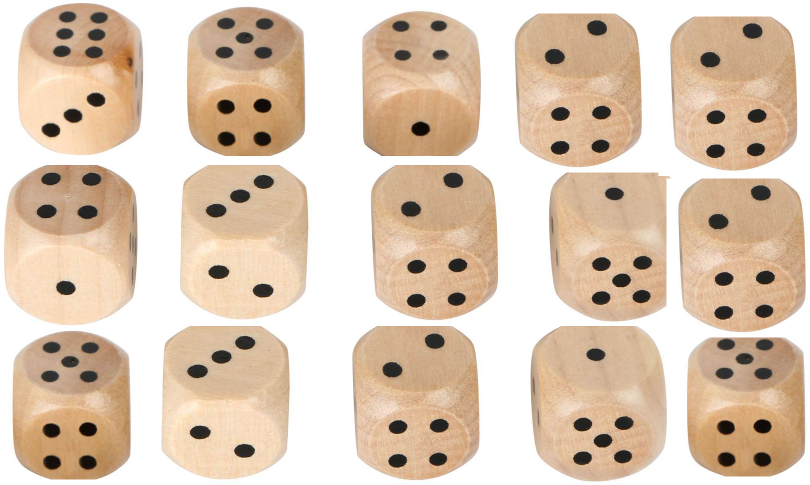
proportion of odd numbers in the faces of a die

Consider repeating this process:
Roll a die 5 times and find the proportion of odd numbers (1 or 3 or 5).

What do we know about the behavior of all sample proportions that are generated as this process continues indefinitely?

…

Example



1/5=0.2

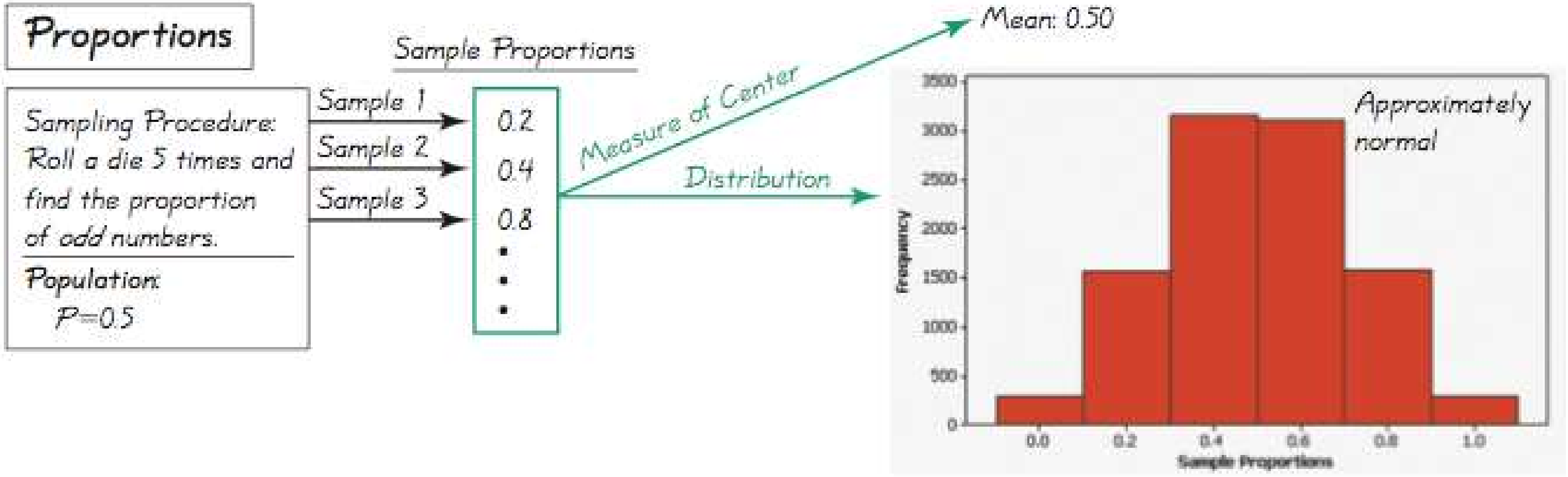
2/5=0.4

4/5=0.8

…

Example - Sampling Distributions

Specific results from 10,000 trials



All outcomes are equally likely so the population proportion of odd numbers is 0.50; the proportion of the 10,000 trials is 0.50. If continued indefinitely, the mean of sample proportions will be 0.50. Also, notice the distribution is "approximately normal."

Sample proportion distribution

The sampling distribution of the sample proportion is the distribution of sample proportions (or the distribution of the variable p) with all samples having the same sample size n taken from the same population.

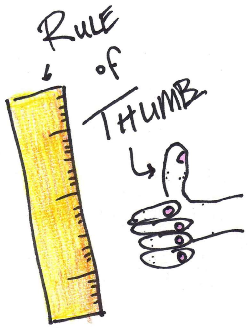
Behavior of Sample Proportions:

- the distribution of sample proportions tends to approximate a normal distribution.
- Sample proportions target the value of the population proportion in the sense that the mean of all of the sample proportions p is equal to the population proportion π; the expected value of the sample proportion is equal to the population proportion
- the standard deviation of sample proportion tends to

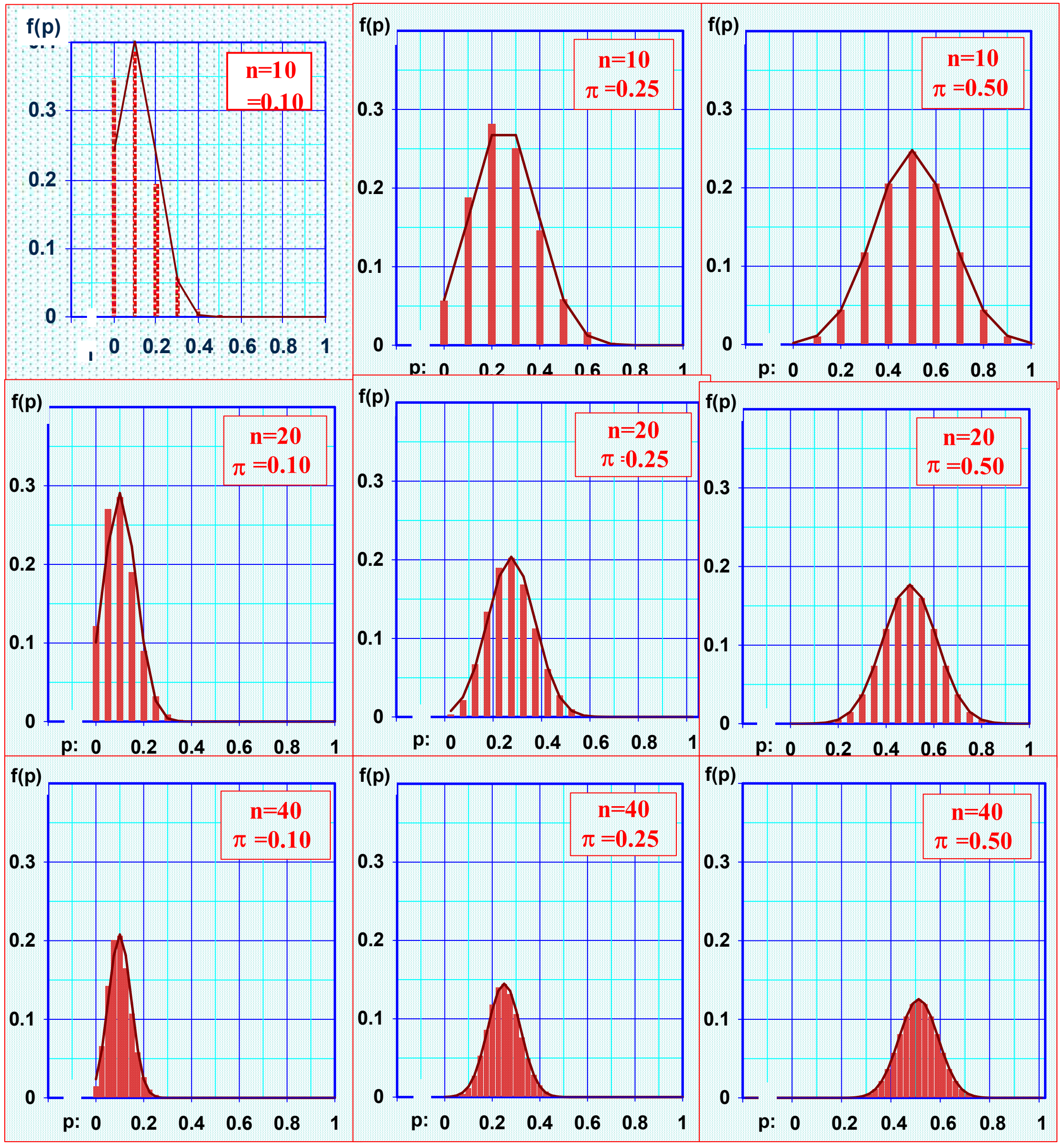
$$\frac{\sigma}{\sqrt{n}} = \sqrt{\pi(1-\pi)/n}$$

Sample proportion distribution

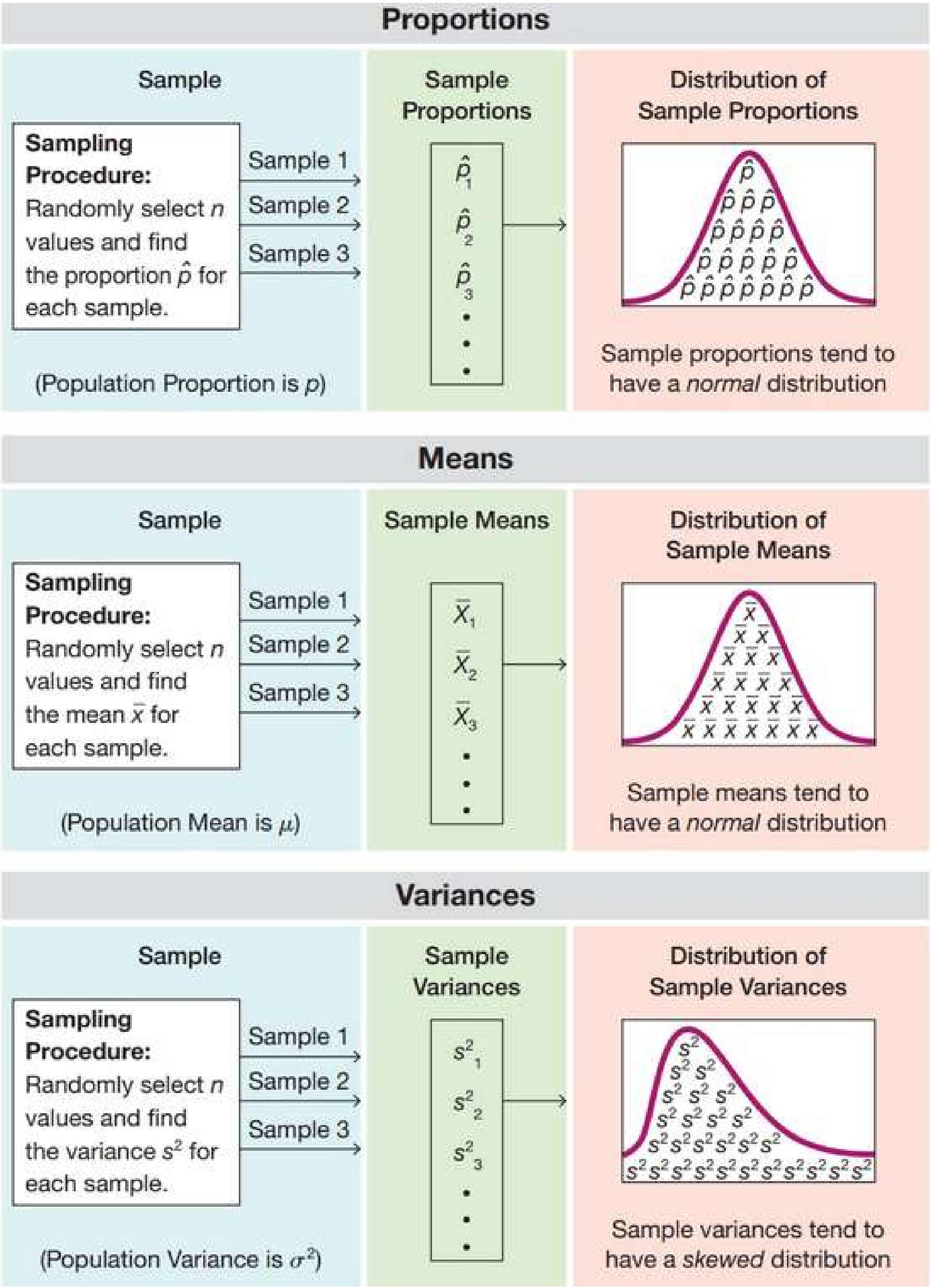
As the size (n) of the sample increases, the values of the relative frequency (p) of the event show a tendency to grow and center around the π parameter, approximating the Gaussian distribution with mean π and standard deviation $\sqrt{\pi(1-\pi)/n}$



nπ ≥5 and n(1- π) ≥5



Summary



Proportions

Sample

Sampling Procedure: Randomly select n values and find the proportion $\hat{p}$ for each sample.

Sample 1

Sample 2

Sample 3

(Population Proportion is p)

Sample Proportions

$\hat{p}_1$

$\hat{p}_2$

$\hat{p}_3$

Distribution of Sample Proportions

Sample proportions tend to have a *normal* distribution

Means

Sample

Sampling Procedure: Randomly select n values and find the mean $\bar{x}$ for each sample.

Sample 1

Sample 2

Sample 3

(Population Mean is μ)

Sample Means

$\bar{x}_1$

$\bar{x}_2$

$\bar{x}_3$

Distribution of Sample Means

Sample means tend to have a *normal* distribution

Variances

Sample

Sampling Procedure: Randomly select n values and find the variance s^2 for each sample.

Sample 1

Sample 2

Sample 3

(Population Variance is σ^2)

Sample Variances

s^2_1

s^2_2

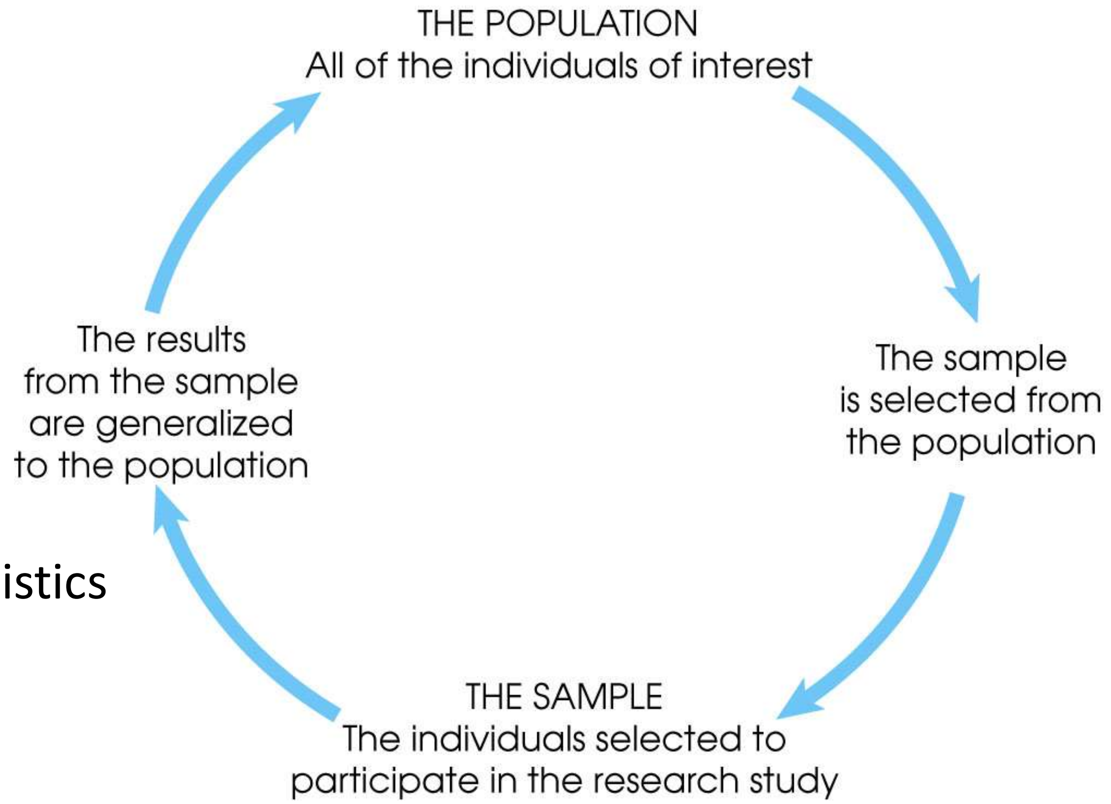
s^2_3

Distribution of Sample Variances

Sample variances tend to have a *skewed* distribution

Inferential statistics

It face up decisional issues using a priori information and the sample data



THE POPULATION
All of the individuals of interest

The sample is selected from the population

THE SAMPLE
The individuals selected to participate in the research study

The results from the sample are generalized to the population

Design of the study & sampling

Inferential statistics

Descriptive statistics on the sample