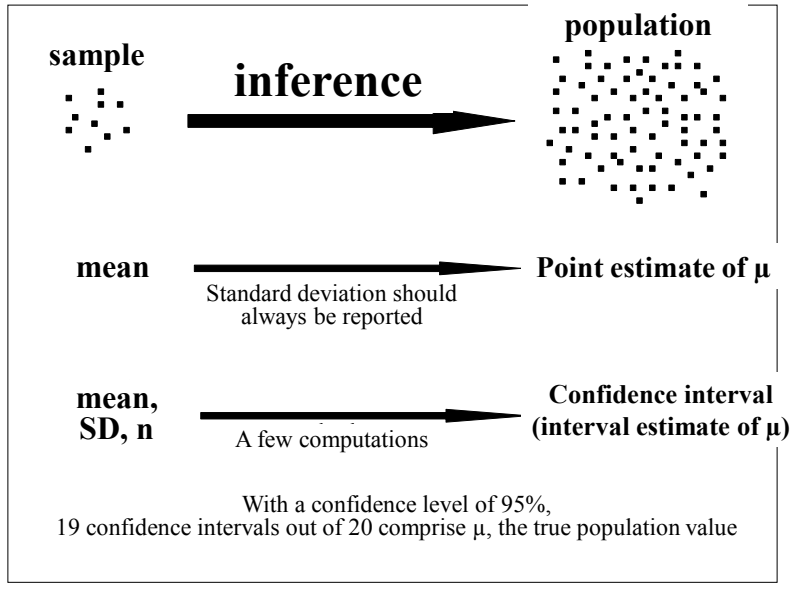
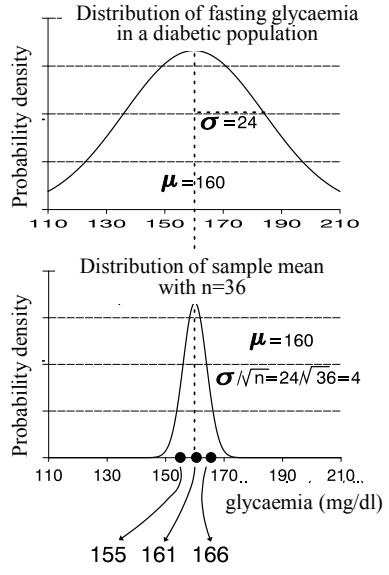
Confidence interval

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

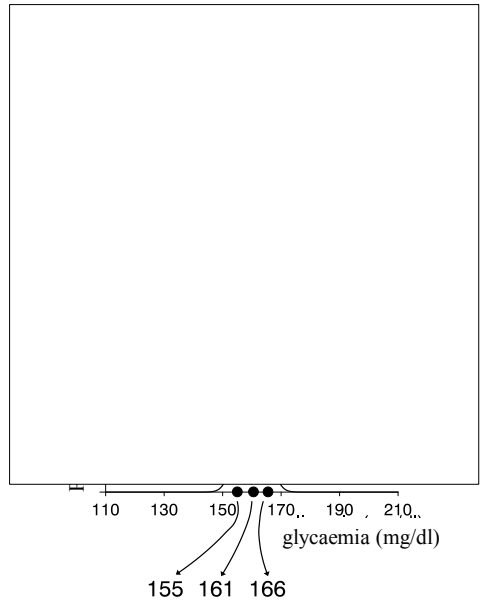


Sampling theory



Point estimates of μ

Statistical inference



Point estimates of μ

Given that a sample is randomly selected from a source population, conclusions drawn from a sample can always be wrong.

Statistical inference is done with "awareness of inherent limitations":

1) One tries to estimate the probability to be wrong.

2) One tries to limit the probability to be wrong.

"*As you are a human being, never state what tomorrow will bring*", Semonides of Amorgos.

"*State contente umane genti al quia, che se possuto aveste veder tutto, mestier non era parturir Maria*", Dante Alighieri

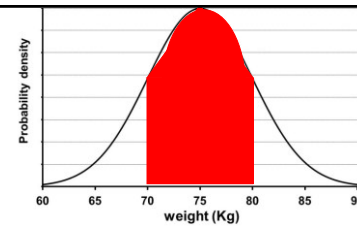
Current scientific approach: We cannot find the truth, but we can approach it.

To make statistical inference, we need to recall previous information:

1) The normal distribution, in particular which interval comprises 95% of the distribution.

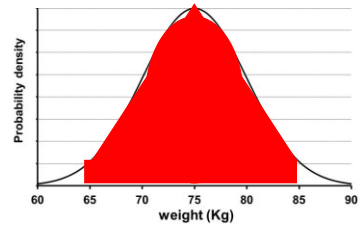
2) The sampling distribution of the mean.

mean ± 1 standard dev. = 75 ± 5 = 70-80 Kg



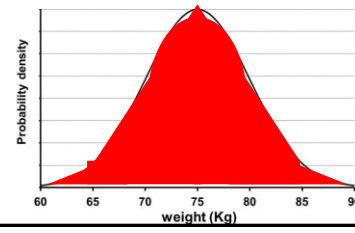
68.26%

mean ± 2 standard dev.= 75 ± 10 = 65-85 Kg

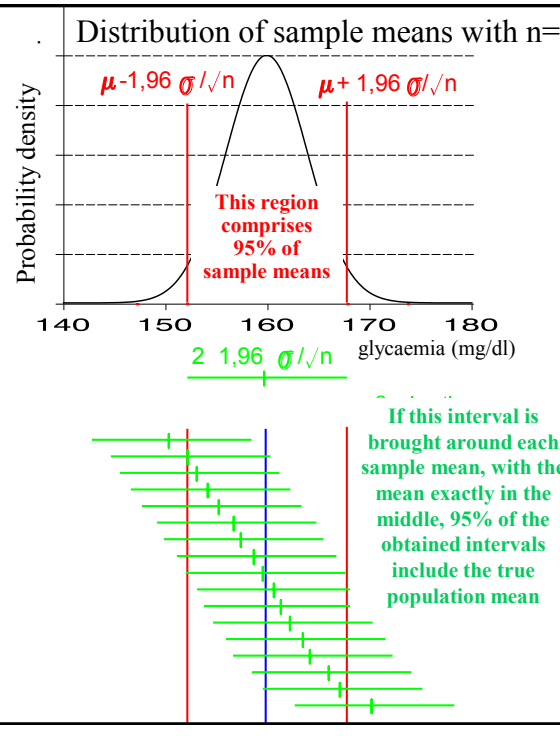


95.44%

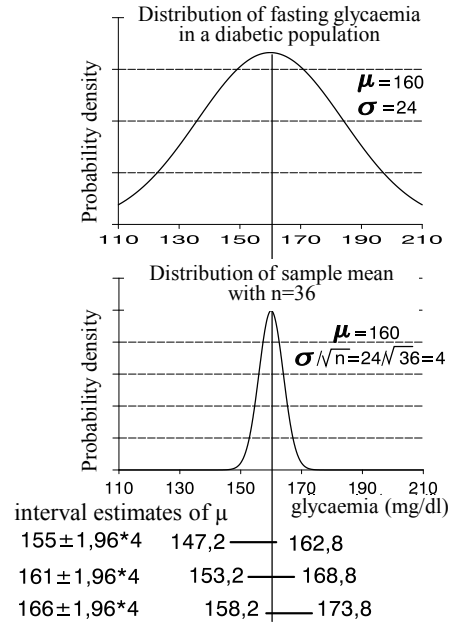
mean ± 3 standard dev.= 75 ± 15 = 60-90 Kg



99.74%



$$\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$$



Point estimate consists in a single value. However:

1) This value nearly never coincides with the true value (parameter) of the population;

2) Different samples yield different point estimates.

Interval estimate consists in an interval:

1) This interval has a given probability (usually 95%) to comprise the true value (parameter) of the population;

2) Intervals, obtained from different samples, are at least partly superimposed.

DEFINITION of CONFIDENCE INTERVAL

The confidence interval of the population parameter Θ is an interval which has a given probability $(1-\alpha)$ to comprise the true population parameter:

$$p(L_{low} < \Theta < L_{up}) = 1-\alpha$$

where:

L_{low} = lower limit; L_{up} = upper limit

$1-\alpha$ = level of confidence; α = error probability

MATHEMATICALLY deriving the 95% Confidence Interval of the mean from the sampling distribution of the mean (σ known)

$$\Pr\left(\mu - 1.96 * \sigma/\sqrt{n} < \bar{x} < \mu + 1.96 * \sigma/\sqrt{n}\right) = 0.95$$

$$\mu - 1.96 * \sigma / \sqrt{n} < \bar{x} < \mu + 1.96 * \sigma / \sqrt{n}$$

$$\Downarrow \; -\mu$$

$$-1.96 * \sigma / \sqrt{n} < \bar{x} - \mu < 1.96 * \sigma / \sqrt{n}$$

$$\Downarrow \; -\bar{x}$$

$$-\bar{x} - 1.96 * \sigma / \sqrt{n} < -\mu < -\bar{x} + 1.96 * \sigma / \sqrt{n}$$

$$\Downarrow \; \text{Multiply by -1}$$

$$\bar{x} + 1.96 * \sigma / \sqrt{n} > \mu > \bar{x} - 1.96 * \sigma / \sqrt{n}$$

$$\bar{x} - 1.96 * \sigma / \sqrt{n} < \mu < \bar{x} + 1.96 * \sigma / \sqrt{n}$$

A confidence interval decreases if:

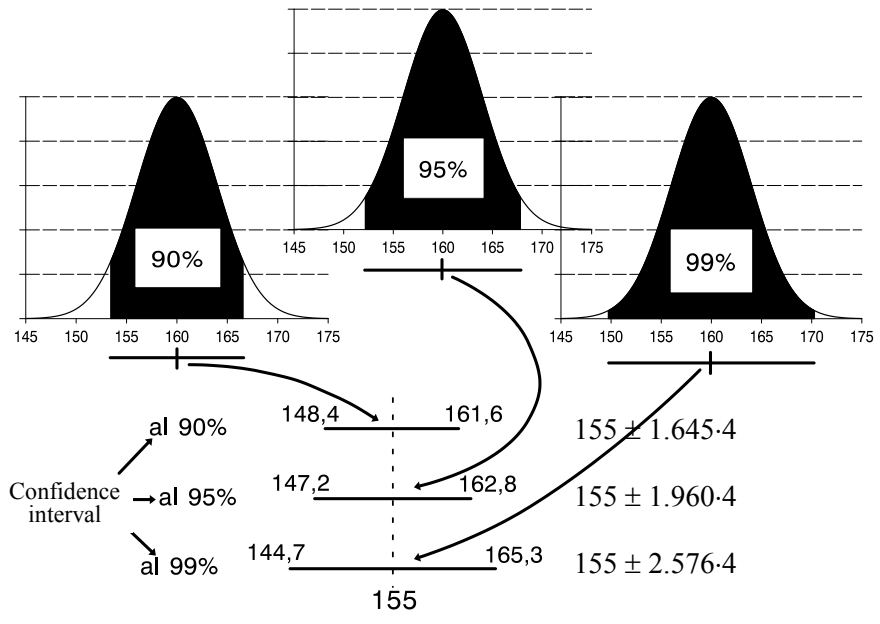
1) the **confidence level** (1-α) decreases
(from 99% to 95% and further to 90%)

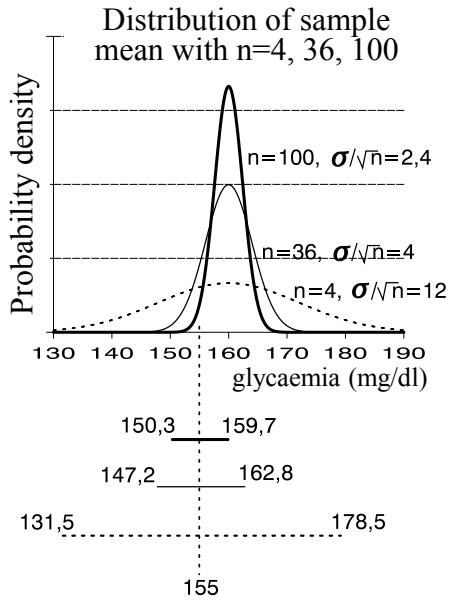
2) the **variability in the source population** decreases
(from σ=48 to σ=24 and further to σ=12)

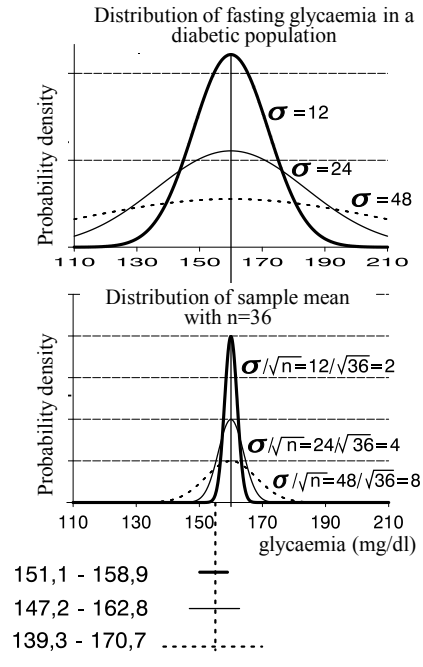
3) the **sample size** increases
(from n=4 to n=36 and further to n=100)

$$\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$$

The most appropriate method is to increase sample size




Distribution of sample mean with n=4, 36, 100
Probability density
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)
150,3
159,7
147,2
162,8
131,5
178,5
155


Distribution of fasting glycaemia in a diabetic population
Probability density
σ =12
σ =24
σ =48
110 130 150 170 190 210
Distribution of sample mean with n=36
Probability density
σ/√n=12/√36=2
σ/√n=24/√36=4
σ/√n=48/√36=8
110 130 150 170 190 210
glycaemia (mg/dl)
151,1 - 158,9
147,2 - 162,8
139,3 - 170,7

Example: Computing the confidence interval of a population mean

Problem: Which is the 95% confidence interval of the mean weight of a population, if a sample of 16 subjects has a mean of 75 Kg? Weight is normally distributed in that population with a standard deviation of 12 Kg.

Data: $x = 75$ Kg $\quad \sigma = 12$ Kg $\quad n = 16$ $\quad 1-\alpha = 95\%$ $\quad z_{\alpha/2} = 1.96$

Equation to be used: $95\%\ C.I. = x \pm z_{\alpha/2} \cdot \sigma/\sqrt{n} = x \pm z_{\alpha/2} \cdot S.E.$

1st step: computing standard error

$S.E. = \sigma/\sqrt{n} = 12/\sqrt{16} = 12/4 = 3$ Kg

2nd step: computing confidence interval

$95\%\ C.I. = x \pm z_{\alpha/2} \cdot S.E. = 75 \pm 1.96 \cdot 3 =$ [80.88 Kg, 69.12 Kg

The interval ranging from 69.12 Kg (lower limit) to 80.88 Kg (upper limit) has 95 chances in 100 to comprise the true population mean.

What about if we don't know σ, population standard deviation?

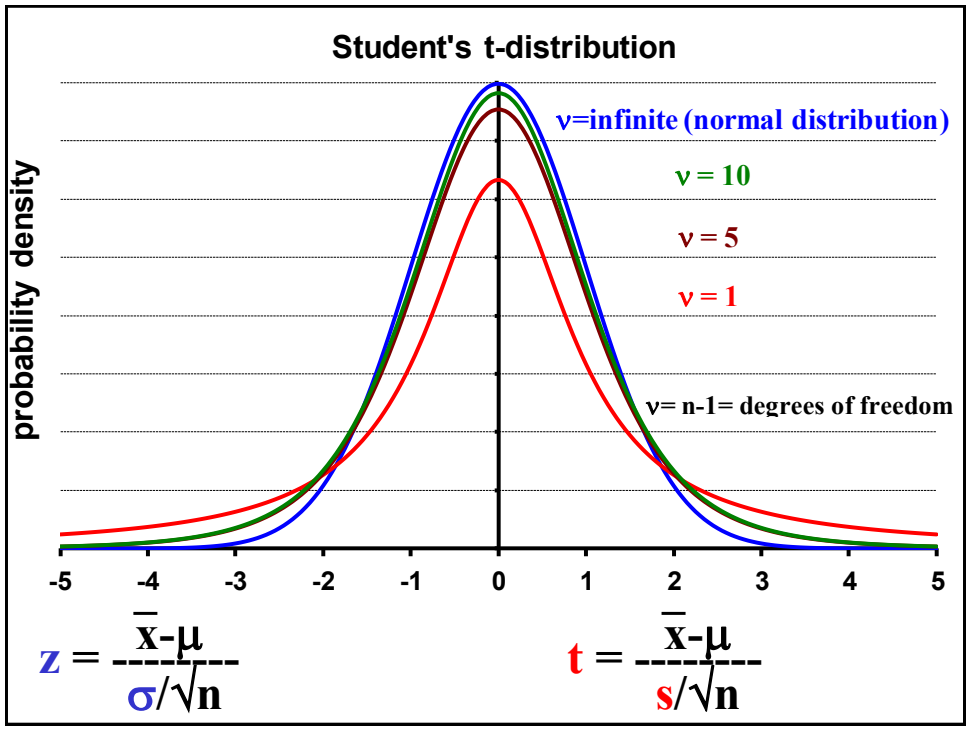
We can use s (sample standard deviation) as an estimate of σ

If sample size is larger enough (n≥60), s is a precise estimate of σ.

$$C.I. = \bar{x} \pm Z_{\alpha/2} * s / \sqrt{n}$$

If sample size is small (n<60), estimating σ from s adds another source of sample variability

We cannot use the z-distribution anymore. We must rather use another distribution, the t-distribution, which has a larger variability than the z-distribution.



In summary:

$$z = \frac{x-\mu}{\sigma} \qquad z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \qquad t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$$

known σ $\Rightarrow$ $\bar{x} \pm Z_{\alpha/2} * \sigma / \sqrt{n}$

unknown σ $\Rightarrow$ $\bar{x} \pm t_{\alpha/2,\,\nu} * s / \sqrt{n}$

Before computer use spread throughout the world, the normal approximation was adopted whenever possible. Nowadays it is very easy to perform exact computations, so that the following equation, based on the normal approximation, is no longer used in statistical practice:

unknown σ and $n \geq 60$ $\Rightarrow$ $\bar{x} \pm Z_{\alpha/2} * s / \sqrt{n}$

Example: Computing the confidence interval of a population mean

Problem: Which is the 95% confidence interval of the mean weight of a population, if a sample of 16 subjects has a mean of 75 Kg and a standard deviation 12 Kg?

Data: $x = 75$ Kg $\quad s = 12$ Kg $\quad n = 16$ $\quad 1-\alpha = 95\%$ $\quad t_{15,\,\alpha/2} = 2.131$

Equation to be used: $95\%\ C.I. = x \pm t_{\alpha/2} \cdot s/\sqrt{n} = x \pm t_{\alpha/2} \cdot S.E.$

1st step: computing standard error
$S.E. = s/\sqrt{n} = 12/\sqrt{16} = 12/4 = 3$ Kg

2nd step: computing confidence interval

$95\%\ C.I. = x \pm t_{15,\,\alpha/2} \cdot S.E. = 75 \pm 2.131 \cdot 3 =$ [81.39 Kg ; 68.61 Kg]

The interval ranging from 68.61 Kg (lower limit) to 81.39 Kg (upper limit) has 95% probability to comprise true population mean.

Confidence interval

$\theta = \mu$

confidence level = 0.95

$$\overline{x} - 1.96 * \sigma / \sqrt{n} < \mu < \overline{x} + 1.96 * \sigma / \sqrt{n}$$

for whatever confidence level = $1-\alpha$

$$\overline{x} - Z_{\alpha/2} * \sigma / \sqrt{n} < \mu < \overline{x} + Z_{\alpha/2} * \sigma / \sqrt{n}$$

for whatever parameter θ

$$\hat{\theta} - Z_{\alpha/2} * S.E.(\hat{\theta}) < \theta < \hat{\theta} + Z_{\alpha/2} * S.E.(\hat{\theta})$$

3rd problem: Computing the confidence interval of a population PROPORTION

Problem: Which is the 95% confidence interval of the proportion (prevalence) of asthma in a population, if the relative frequency of asthma in a sample of 225 subjects is 4.89% ?

Data: p = 0,05 n = 225 1-α= 95% $z_{\alpha/2}$ = 1.96 C.I. = ?

Equation to be used: 95% C.I. = $p \pm z_{\alpha/2} \cdot \sqrt{p(1-p)/n} = p \pm z_{\alpha/2} \cdot S.E.$

1st step: computing standard error

S.E. = $\sqrt{p(1-p)/n} = \sqrt{0.0489(1-0.0489)/225} = \sqrt{0.0489*0.9511/225} = 0.0144 = 1.44\ \%$

2nd step: computing confidence interval

95% C.I. = $p \pm z_{\alpha/2} \cdot S.E.$ =

Upper limit = 4.89 + 1.96*1.44 = 7.71%

Lower limit = 4.89 – 1.96*1.44 = 2.07%

The interval ranging from 2.07% (lower limit) to 7.71% (upper limit) has 95% probability to comprise the true asthma prevalence in that population.

CONFIDENCE INTERVAL AT A (1-α)

CONFIDENCE LEVEL OF A PROPORTION

If $np \geq 10$ and $n(1-p) \geq 10 \Rightarrow \hat{\pi} = p \sim N(\pi, \pi(1-\pi)/n)$

We use $p(1-p)/n$ to estimate $\pi(1-\pi)/n$

$$p - Z_{\alpha/2} * \sqrt{p(1-p)/n} < \pi < p + Z_{\alpha/2} * \sqrt{p(1-p)/n}$$

for $1-\alpha = 95\%$

$$p - 1.96 * \sqrt{p(1-p)/n} < \pi < p + 1.96 * \sqrt{p(1-p)/n}$$

4th problem: Using the Confidence Interval to plan sample size.

A study aims at computing the prevalence (probability) of asthma in a population. Preliminary data from the current literature suggest that asthma prevalence could be about 6%. Which is the sample size necessary to estimate asthma prevalence with a width of the 95% confidence interval not greater than 2% ?

Data: $p = 0.06$ $1-\alpha = 95\%$ $z_{\alpha/2} = 1.96$ CI width $\leq 2\%$ $n = ?$

$(p + z_{\alpha/2} \cdot \sqrt{p(1-p)/n}) - (p - z_{\alpha/2} \cdot \sqrt{p(1-p)/n}) \leq \delta$

$2 z_{\alpha/2} \cdot \sqrt{p(1-p)/n} \leq \delta$ dividing by $2 z_{\alpha/2}$

$\sqrt{p(1-p)/n} \leq \delta /(2 z_{\alpha/2})$ squaring the equation

$p(1-p)/n \leq \delta^2 /(2 z_{\alpha/2})^2$ multiplying by n and dividing by the 2nd member

$p(1-p) (2 z_{\alpha/2})^2/ \delta^2 \leq n$

$n \geq 0.06*0.94 * (2 * 1.96)^2/ 0.02^2$ $n \geq 0.0564 * (3.92)^2/ 0.0004$

$n \geq 0.0564 * 15.36 / 0.0004$ **$n \geq 2166.58$** **$n \geq 2167$**