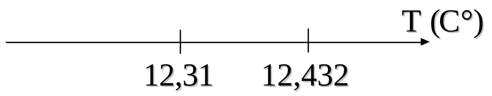
Variabile casuale

- *X* è una variabile casuale (v.c.)
  - In statistica i termini "aleatorio", "casuale", "stocastico" sono aggettivi che si associano agli eventi ottenuti come risultati di una prova
  - Una **variabile casuale** (o **variabile aleatoria** o **variabile stocastica**) può essere pensata come il risultato numerico di un esperimento quando questo non è prevedibile con certezza (ossia è non deterministico)
    - Ad esempio, il risultato del lancio di un dado a sei facce è una variabile casuale con possibili valori gli interi da 1 a 6
      - Variabile casuale a (6) valori discreti
    - La temperatura ambientale misurata ogni 22 gennaio alle 12:00 a Cassino è una v.c. di media 12,31°
      - Variabile casuale continua

Distribuzione di probabilità

- *Ad una variabile casuale X si associa la legge di probabilità, che associa ad ogni sottoinsieme dell'insieme dei possibili valori di X la probabilità **P** che la v.c. X assuma valore in esso.*
- *X* è una variabile casuale (v.c.) continua
  - Il dominio di X può essere limitato o illimitato
  - Poichè X è continua, $P(X = x)=0$
  - Serve una ***Probability density function*** (PDF)



Distribuzioni continue

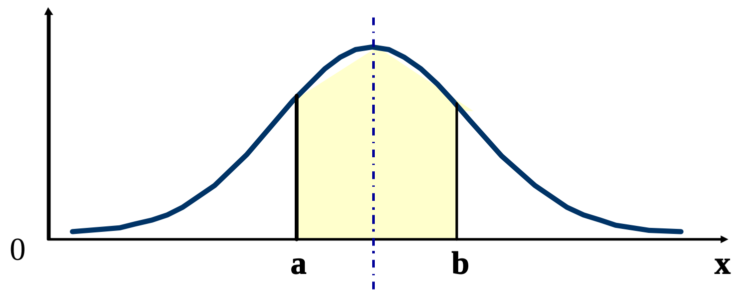
- X è una variabile casuale (v.c.) continua
  - ***Probability density function*** (PDF)
    - È una function $\boldsymbol{f(x)}$ con 3 proprietà:
      1. $f(x) \geq 0$ for all real values x
         - La densità di probabilità ovviamente è positiva
      2. For any fixed a and b with $a \leq b$, the probability that X will fall between a and b is the area under $f(x)$ between a and b

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

Continuous Distributions

Proprietà 2:

Shaded Area is the Probability That X is Between a and b



$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

Distribuzioni continue

- X è una variabile casuale (v.c.) continua
  - ***Probability density function*** (PDF)
    - È una function ***f*(*x*)** con 3 proprietà:
      1. $f(x) \geq 0$ for all real values x
         - La densità di probabilità ovviamente è positiva
      2. For any fixed a and b with $a \leq b$, the probability that X will fall between a and b is the area under $f(x)$ between a and b:

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

      3. The total area under f(x) is 1:

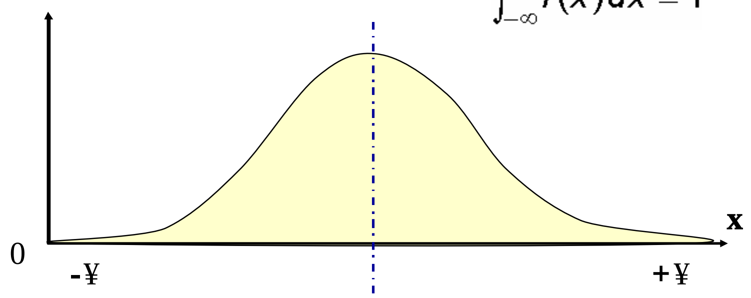
$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

Continuous Distributions

Proprietà 3:

tutta l'area sottesa rappresenta tutta la probabilità e quindi è uguale al 100%, cioè 1

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$



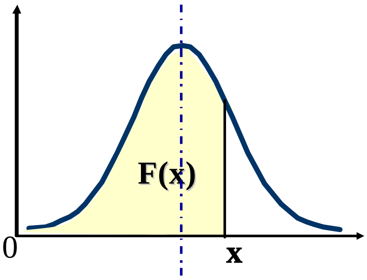
Continuous Distributions (cont'd.)

- *Cumulative distribution function* (CDF) - probability that the v.c. will be ≤ a fixed value *x*:

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)\,dt$$

- Properties of continuous CDFs
  - $0 \le F(x) \le 1$ for all x
  - As $x \to -\infty$, $F(x) \to 0$
  - As $x \to +\infty$, $F(x) \to 1$
  - $F(x)$ is nondecreasing in x
  - $F(x)$ is a continuous function with slope equal to the PDF:

$$f(x) = F'(x)$$

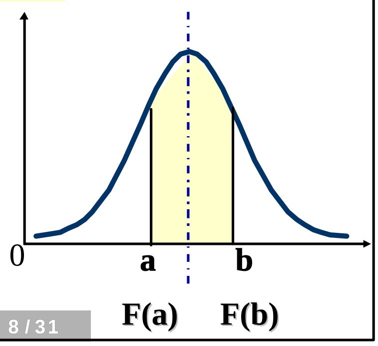


Continuous Distributions (cont'd.)

- *Cumulative distribution function* (CDF) - probability that the v.c. will be ≤ a fixed value *x*:

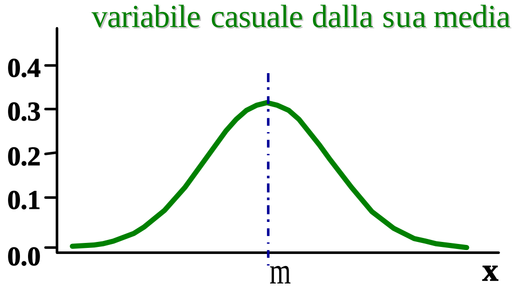
$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)\,dt$$

$$P(a \le X \le b) = \int_{a}^{b} f(x)\,dx = F(b)-F(a)$$



Expected Values, Variances, and Standard Deviations

- Valor atteso or mean of X is $\mu = \mu_X = E(X) = \int_{-\infty}^{+\infty} x\, f(x)\, dx$
  - Media pesata, tramite il peso f(x), della variabile casuale
- Variance of X is $\sigma^2 = \sigma_X^2 = Var(X) = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)\, dx$
  - Media pesata, tramite f(x), della distanza al quadrato della variabile casuale dalla sua media

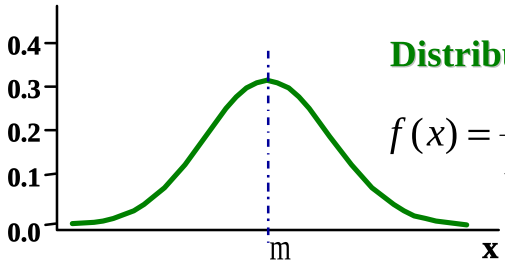


Expected Values, Variances, and Standard Deviations

- Valor atteso or mean of X is $\mu = \mu_X = E(X) = \int_{-\infty}^{+\infty} x\, f(x)\, dx$
- Variance of X is

$$\sigma^2 = \sigma_X^2 = Var(X) = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)\, dx$$

- Standard deviation of X is $\sigma = \sigma_X = +\sqrt{Var(X)}$



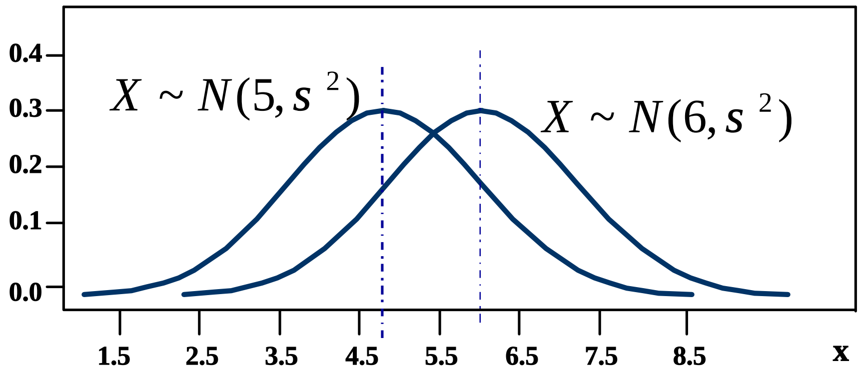
Distribuzione normale

$$f(x) = \frac{1}{\sqrt{2ps^2}} e^{-(x-m)^2/2s^2}$$

$$X \sim N(m, s^2)$$

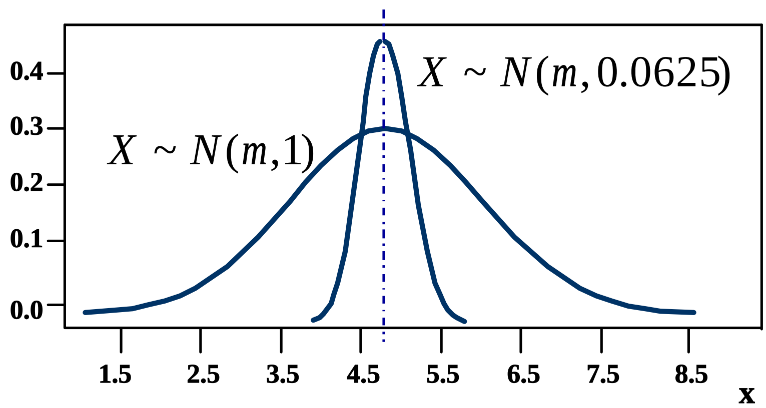
Expected Values, Variances, and Standard Deviations

Effects of m on the Probability Density Function of a Normal Random Variable



Expected Values, Variances, and Standard Deviations

Effects of s^2 on the Probability Density Function of a Normal Random Variable



Expected Values, Variances, and Standard Deviations

The Standard Normal Distribution

Let *Z* be a normal random variable with mean *0* and variance *1*; that is

$$Z \sim N(0,1)$$

We say that *Z* follows the standard normal distribution. Denote the cumulative distribution function as *F(z)*, and *a* and *b* as two numbers with *a* < *b*, then

$$P(a < Z < b) = F(b) - F(a)$$

Expected Values, Variances, and Standard Deviations

The Standard Normal Distribution

se

$$X \sim N(m, s^2)$$

the random variable Z = (X - m)/s has a standard normal distribution: Z ~ N(0, 1)

Sampling (campionamento)

- *Il campionamento sta alla base della **inferenza** statistica, la quale si divide in due grandi capitoli: la **stima** e la verifica (o **test**) **d'ipotesi***
- STIMA
  - estimate or infer something about a *population* or *process* based on only a *sample* from it
    - Think of a R.V. with a distribution governing the population
    - *Random sample* (**campione**) is a set of *independent and identically distributed* (IID) observations $X_1, X_2, \ldots, X_n$
    - In simulation, sampling is making some runs of the model and collecting the output data
    - Don't know *parameters* of population (or distribution) and want to estimate them or infer something about them based on the sample

La stima campionaria

- **Population parameter**
  - **Population mean** $m = E(X)$
  - **Population variance** s^2
  - **Population proportion**
- **Parameter – need to know whole population**
- **Fixed (but unknown)**

- **Sample estimate**
  - **Sample mean**
  - **Sample variance**
  - **Sample proportion**
- ***Sample statistic***
  - **can be computed from a sample**
  - **Varies from one sample to another – is a RV itself, and has a distribution, called the *sampling distribution***

Sampling Distributions

- Have a statistic, like sample mean or sample variance
  - Its value will vary from one sample to the next
- The sample mean

Media campionaria $\overline{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

- The sample variance

Varianza campionaria $s^2 = \dfrac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$

Sampling Distributions

- Some sampling-distribution results
  - Sample mean $\overline{X}$

    $E(\overline{X}) = \mu, \quad Var(\overline{X}) = \sigma^2/n$

    If $X \sim N(\mu, \sigma^2)$ then $\overline{X} \sim N(\mu, \sigma^2/n)$

$\overline{X}$ è a sua volta una v.c., anch'essa distribuita normalmente, se la popolazione X è normale

$\overline{X} \sim N(\mu, \sigma/\sqrt{n})$ for large n

Regardless of distribution of X,

  - Sample variance s^2

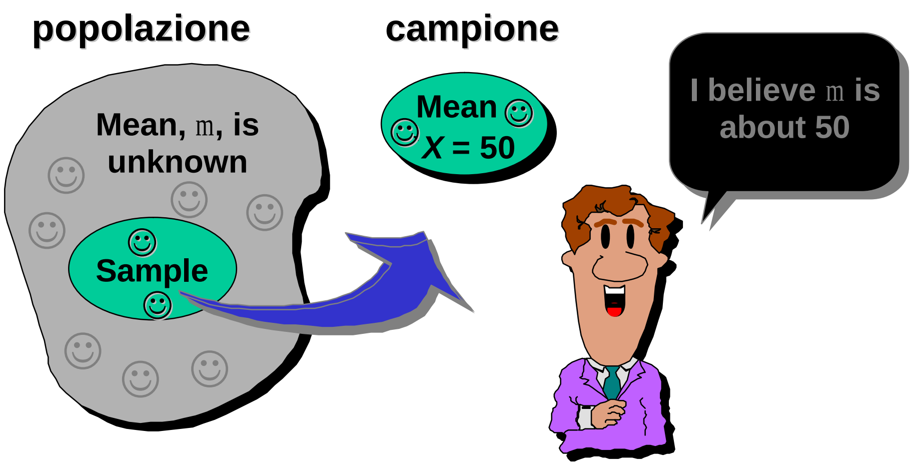
    $E(s^2) = s^2$

ma anche se X non è normale, per i grandi numeri è distribuita normalmente

La stima campionaria puntuale

- A sample statistic that estimates (in some sense) a population parameter
- Properties
  - **Unbiased**: *E*(estimate) = parameter
  - **Efficient**: *Var*(estimate) is lowest among competing point estimators
  - **Consistent**: *Var*(estimate) decreases (usually to 0) as the sample size *n* increases

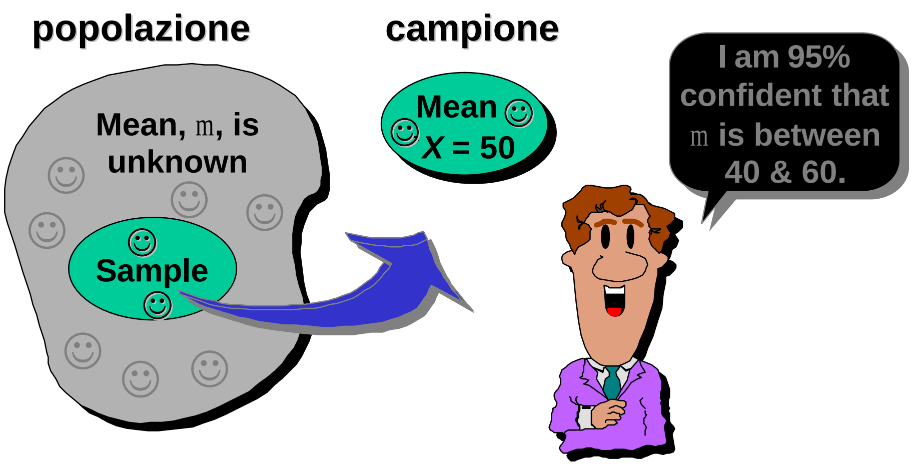
Stima puntuale



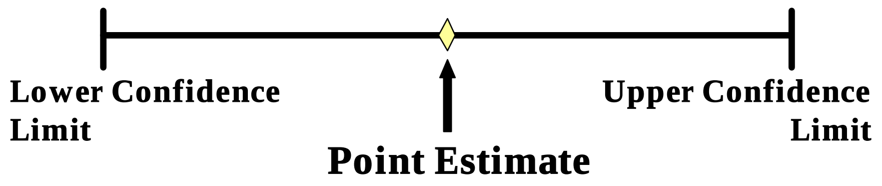
Stima puntuale

Estimate Population Parameter...		with Sample Statistic
Mean	m	$\bar{x}$
Variance	s^2	s^2
Differences	m_1 - m_2	$\bar{x}_1 - \bar{x}_2$

Stima per intervalli

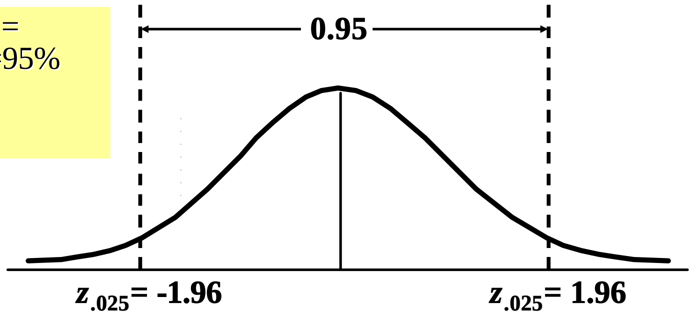


La stima: Confidence Intervals



confidence level =
=100 (1 – a)% =95%

a% =5%



La stima: Confidence Intervals

- A point estimator is just a single number, with some uncertainty or variability associated with it
- *Confidence interval* quantifies the likely imprecision in a point estimator
  - An interval that contains (*covers*) the unknown population parameter with specified (high) probability 1 – a
  - Called a 100 (1 – a)% confidence interval for the parameter
- Confidence interval for the population mean m:

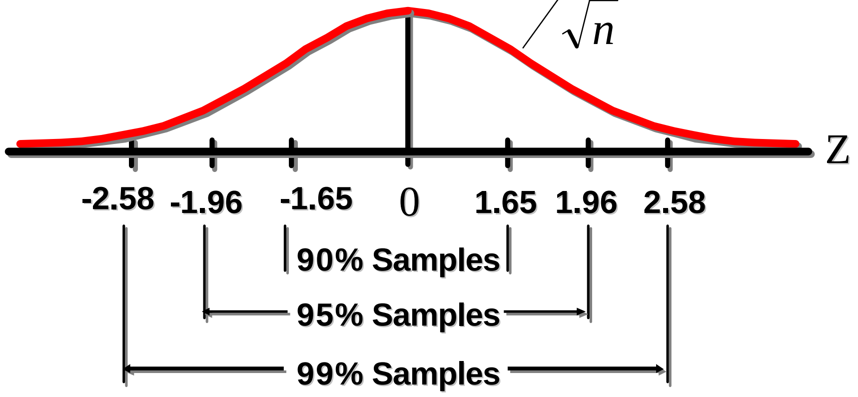
Confidence Interval

- Alcuni risultati
  - **Media campionaria** $\overline{X}$
    - Valore atteso $E(\overline{X}) = \mu,$
    - Varianza $Var(\overline{X}) = \sigma^2/n$
    - Se $X \sim N(\mu,\sigma)$ then $\overline{X} \sim N(\mu,\sigma/\sqrt{n})$
    - Indipedentemente dalla distribuzione di X
      $\overline{X} \sim N(\mu,\sigma/\sqrt{n})$ for large n
  - Varianza campionaria s^2
    - $E(s^2) = s^2$

Confidence Interval (s nota, X normale)

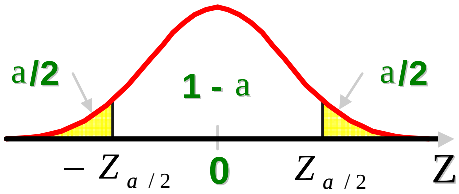
se $X \sim N(m,\ s)$ Þ $\overline{X} \sim N(\mu,\sigma/\sqrt{n})$

Þ $\dfrac{\overline{X} - m}{s/\sqrt{n}} = Z$ ~*N(0,1)*



Confidence Interval
(σ nota, X normale)

$$-Z_{\alpha/2} \le \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \le Z_{\alpha/2}$$



$$\overline{X} - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

esempio

$n = 25$
$\overline{X} = 50$
$\sigma = 10$

95% confidence interval for μ?

$$50 - 1.96 \cdot \frac{10}{\sqrt{25}} \le \mu \le 50 + 1.96 \cdot \frac{10}{\sqrt{25}}$$

$$46.08 \le \mu \le 53.92$$

Confidence Interval for Mean
(σ non nota)

- Varianza campionaria s^2
  - $E(s^2) = \sigma^2$

$$\frac{\overline{X} - \mu}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \approx \frac{\overline{X} - \mu}{s / \sqrt{n}}$$

  - La variabile $T = \dfrac{\overline{X} - \mu}{s / \sqrt{n}}$

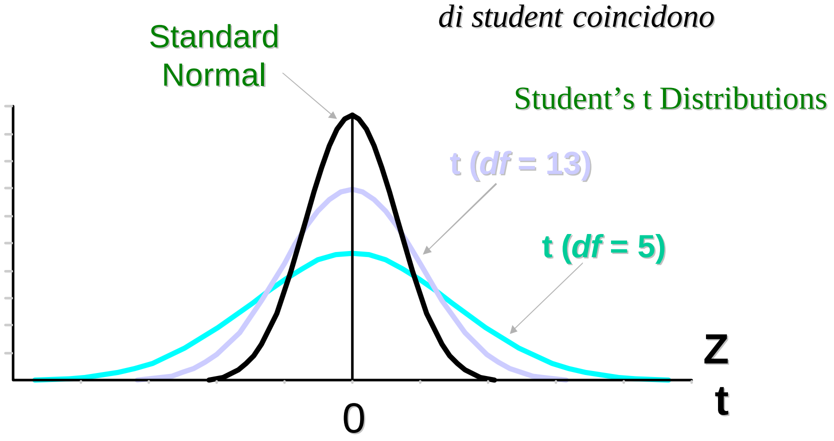
segue una distribuzione $\mathbf{t_{n-1}}$ di Student con **n-1** g.d.l.

Intervallo di confidenza al 100(1-α)%

$$\overline{X} - t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}, \overline{X} + t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}$$

Confidence Interval for Mean
(s non nota)

*Per n elevato (n>25) la normale standardizzata e la **t** di student coincidono*



Confidence Interval for Mean
(s non nota)

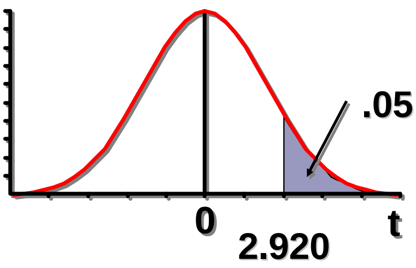
Student's t Table

v	$t_{.10}$	$t_{.05}$	$t_{.025}$
1	3.078	6.314	12.706
2	1.886	2.920	4.303
3	1.638	2.353	3.182

a / 2

t values

Assume:
***n* = 3**
***df* = *n* - 1 = 2**
a **= .10**
a**/2 =.05**



Confidence Interval for Mean

Possibili casi

s **nota**
X~N(μ,σ²)
$\forall\, n$

s **nota**
X qualunque
$n > 25$

$\bar{X} \sim N(\mu, \sigma/\sqrt{n})$

Intervallo di confidenza costruito tramite **Z**

s **non nota**
X~N(μ,σ²)
$n \leq 25$

Intervallo di confidenza costruito tramite **T**

X non normale
$n \leq 25$

?