# Costruzione di macchine

Modulo di:

Progettazione probabilistica e affidabilità

Marco Beghini

Lezione 7:

Basi di statistica

# Campione e Popolazione

Estrazione da una popolazione (virtualmente infinita) di un campione di numerosità: $n$

$$\{x_1, x_2, x_3, \ldots, x_n\}$$

Equivale a ripetere $n$ volte, in modo indipendente, un esperimento aleatorio caratterizzato da una V.A. $X$ con una distribuzione fissata (non necessariamente nota).

L'**inferenza statistica** rappresenta l'insieme dei metodi che permettono di ricavare informazioni sulla popolazione a partire dal campione: stima delle caratteristiche.

Inevitabilmente non sarà mai possibile ottenere la certezza della stima ma la teoria fornisce il modo più corretto di farlo e anche i livelli o intervalli di confidenza corrispondenti.

# Trattamento dei dati campionari

**Esempio 7.1**

Campione di $n$=600 valori estratti da una V.A. uniformemente distribuita nell'intervallo [$a$, $b$] con $a$=0, $b$=4.

$$\{x_1, x_2, x_3, \ldots, x_{600}\}$$

1) Divisione del dominio in $M$ sottointervalli (per esempio uguali)

$$M = \left[1 + 3.3 \log(n)\right] \qquad n = 600 \Rightarrow M = 10$$

$$\Delta = \frac{b-a}{M} \qquad I_k = \left[a + (k-1)\Delta,\ a + k\Delta\right];\ \text{con}\ \ k = 1, .., M$$

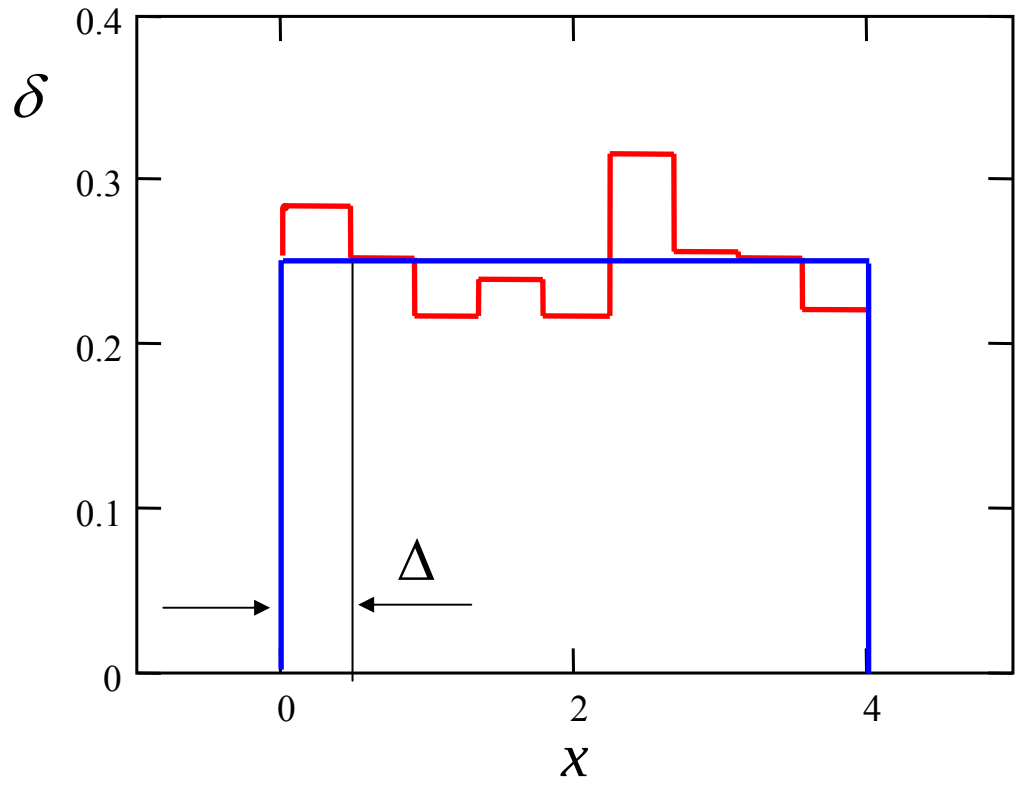
2) Numerosità di valori per sottointervallo

$$O_k = \#\{x_i \in I_k\}$$

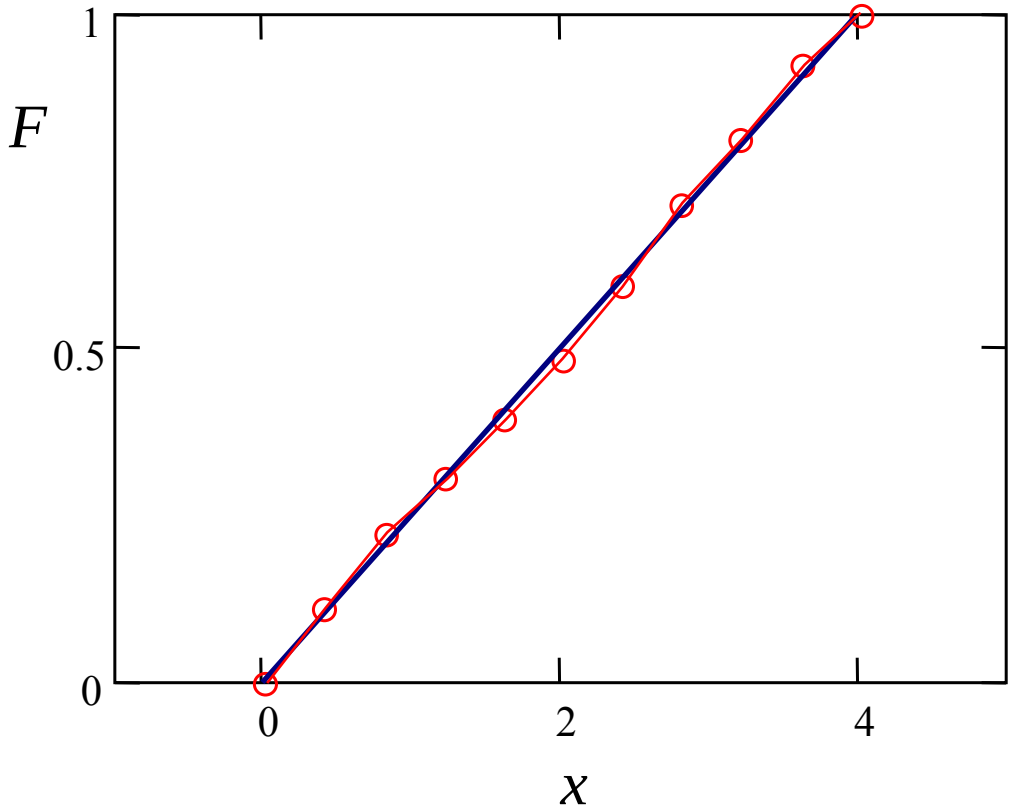
3) Densità di valori per sottointervallo

$$\delta_k = \frac{O_k}{n\Delta}$$

4) Istogramma della densità delle frequenze relative (confrontato con $f(x)$ )



5) Integrale dell'istogramma (confrontato con $F(x)$ )

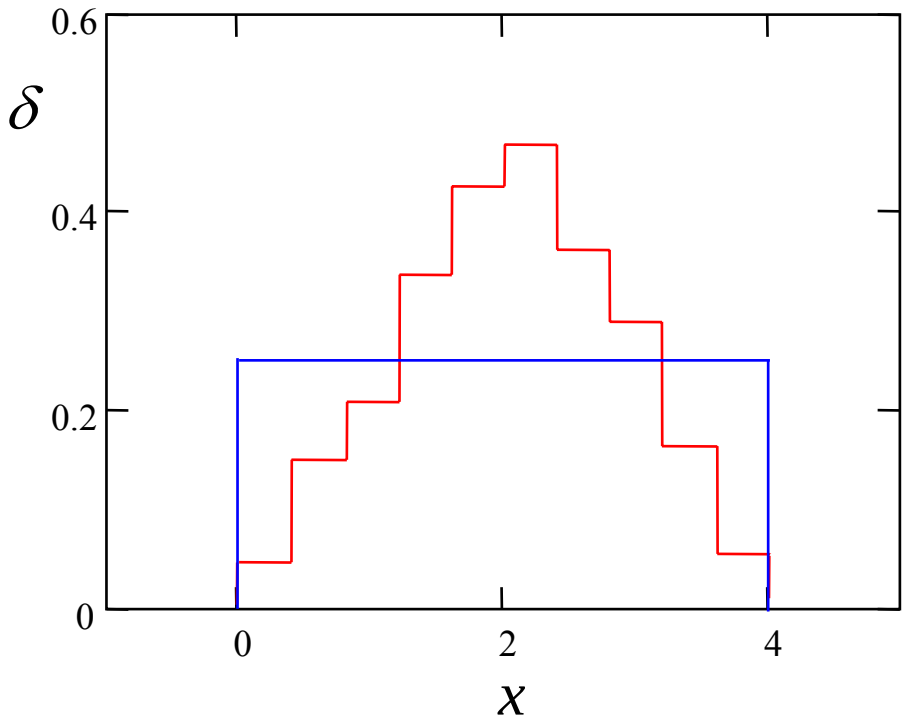


# Esperimenti numerici

## Esempio 7.2

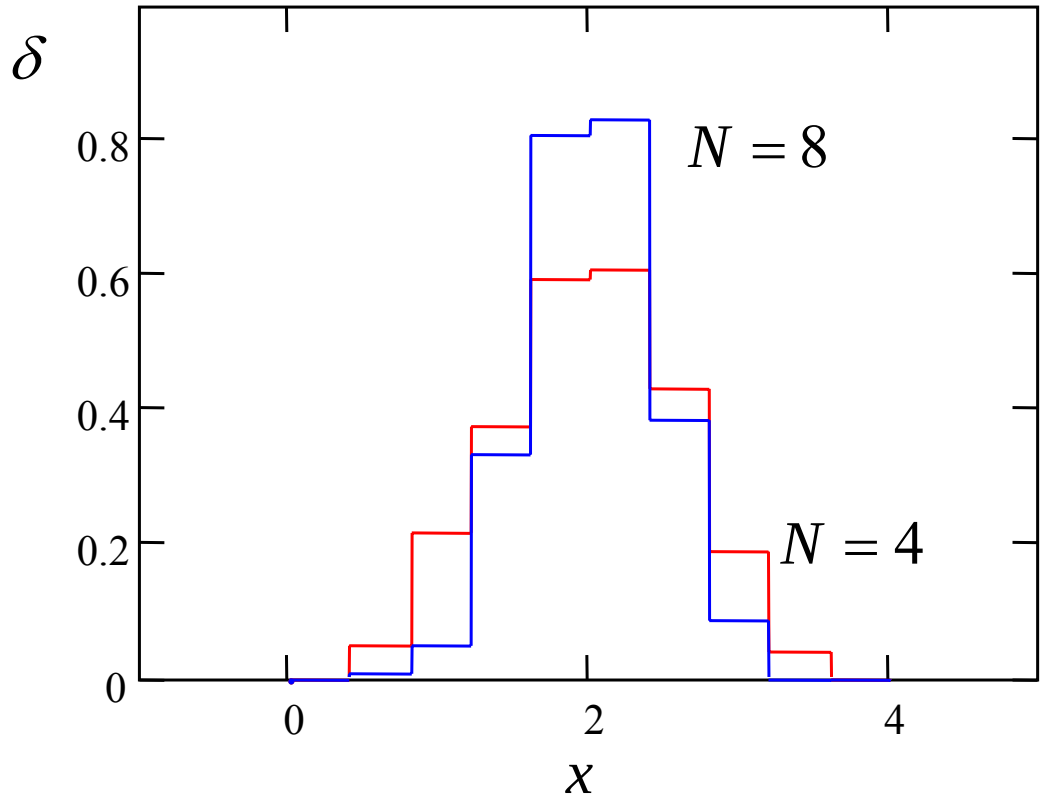
Due campioni, entrambi di $n$=600 valori estratti da $N$=2 V.A. indip. $X^{(1)}$, $X^{(2)}$, uniformemente distribuite nell'intervallo [$a$, $b$] ($a$=0, $b$=4), sono mediati per ottenere un campione di $X_m$=0.5($X^{(1)}$+$X^{(2)}$). Tracciare l'istogramma.

$$X_m = \frac{X^{(1)} + X^{(2)}}{2}$$



**Esempio 7.3**

Cosa succede aumentando il numero di elementi della media?




- La distribuzione si stringe
- La distribuzione tende a diventare campaniforme

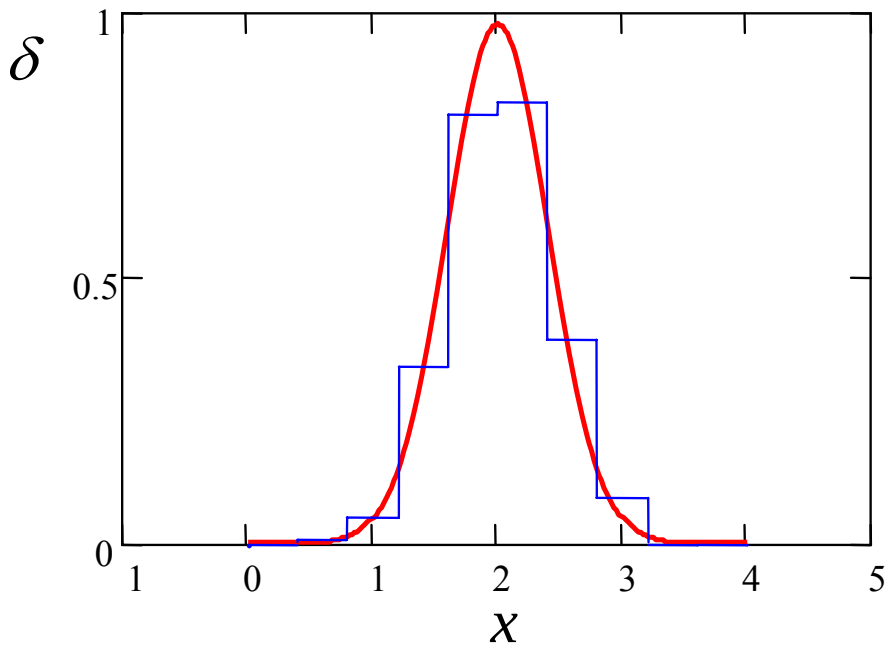
**Teorema (limite centrale)**

Date $N$ V.A. $X^{(i)}$ con la stessa distribuzione (continua, discreta e qualunque) con media $\mu$ e deviazione standard $\sigma$ la V.A. ottenuta come media:

$$X_m = \frac{1}{N}\sum_{i=1}^{N} X^{(i)}$$

quando $N$ è grande ha le seguenti caratteristiche (tende a):

$$X_m = N\left(x, \mu, \frac{\sigma}{\sqrt{N}}\right)$$



# Campione e Popolazione

**Problema**: prevedere (stimare) le caratteristiche della popolazione (distribuzione della V.A. di partenza) elaborando il campione di *n* elementi estratto casualmente:

$$\{x_1, x_2, \ldots, x_n\}$$

## Previsione della media $\mu$ (1/3)

Media campionaria:

$$m = \frac{\sum_{i=1}^{n} x_i}{n}$$

Si dimostra che:

La media campionaria è una stima corretta ed efficiente della media $\mu$

# Previsione della media $\mu$ (2/3)

Supponiamo di fare la media di $M$ campioni ognuno di numerosità $n$ di una V.A. $X$:

$$\left\{x_1^{(1)}, x_2^{(1)}, ..., x_n^{(1)}\right\} \qquad m_1 = \frac{1}{n}\sum_{i=1}^{n} x_i^{(1)}$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\left\{x_1^{(j)}, x_2^{(j)}, ..., x_n^{(j)}\right\} \qquad m_j = \frac{1}{n}\sum_{i=1}^{n} x_i^{(j)}$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\left\{x_1^{(M)}, x_2^{(M)}, ..., x_n^{(M)}\right\} \qquad m_M = \sum_{i=1}^{n} x_i^{(M)}$$

si ottiene un nuovo campione:

$$\left\{m_1, m_2, ..., m_M\right\}$$

di numerosità $M$ delle medie campionarie, ovvero un campione della V.A. media campionaria di $n$ elementi estratti a caso da $X$.

Se la V.A. $X$ di partenza è normale con: $X = N\left(x, \mu, \sigma_X\right)$

la V.A. media campionaria è:

$$N\left(m, \mu, \frac{\sigma_X}{\sqrt{n}}\right)$$

# Previsione della media $\mu$ (3/3)

Anche se la variabile $X$ di partenza non è normale **ma *n* è elevato,** la media campionaria è approssimabile con:

$$N\left(m, \mu, \frac{\sigma_X}{\sqrt{n}}\right)$$

Il modello gaussiano tende a rappresentare adeguatamente la distribuzione di ogni grandezza prodotta dall'azione combinata di molti piccoli effetti concomitanti che si sommano

Per questo motivo la distribuzione normale è il modello più adatto per modellare le incertezze di misura

**Esempio 7.4**

Un campionamento con $n = 8$ di una V.A. $X$ che ha deviazione standard (nota) $\sigma_x = 4$ ha fornito i seguenti valori:

6.37 12.6 15.7 9.1 14.8 12.9 8.2 8.8

cosa si può concludere sulla media $\mu$ di $X$ ?

1) Calcoliamo la media del campione: $m = 11.06$

2) Il valore ottenuto è stato estratto (a caso) dalla variabile media campionaria con numerosità $n$=8, per cui è una V.A. gaussiana che ha media (non nota) $\mu$ e deviazione standard:

$$\sigma_m = \frac{\sigma_X}{\sqrt{n}} = \frac{4}{\sqrt{8}} = \sqrt{2}$$

3) Possiamo valutare la probabilità che il valore effettivamente misurato di $m$ differisca da $\mu$ al più di una certa quantità usando i quantili della distribuzione normale, in particolare:

$$P\left(\left|m-\mu\right| < \sigma_m\right) = 0.6826$$

$$P\left(\left|m-\mu\right| < 2\sigma_m\right) = 0.9545$$

$$P\left(\left|m-\mu\right| < 3\sigma_m\right) = 0.9973$$

4) Ribaltando il ragionamento, possiamo affermare che, fissato un determinato livello di confidenza, la media **incognita** della popolazione si colloca nell'intervallo corrispondente:

| Confidenza | posizione | caso numerico |
|---|---|---|
| 68% | $m-\frac{\sigma_X}{\sqrt{n}}<\mu<m+\frac{\sigma_X}{\sqrt{n}}$ | $9.6<\mu<12.5$ |
| 95% | $m-2\frac{\sigma_X}{\sqrt{n}}<\mu<m+2\frac{\sigma_X}{\sqrt{n}}$ | $8.2<\mu<13.9$ |
| 99.7% | $m-3\frac{\sigma_X}{\sqrt{n}}<\mu<m+3\frac{\sigma_X}{\sqrt{n}}$ | $6.8<\mu<15.3$ |

**Nota**: per ridurre l'intervallo di indeterminazione, a parità di livello di confidenza, è necessario aumentare la numerosità del campione. Attenzione alla dipendenza: $\sqrt{n}$

**Esercizio 7.1**

Sapendo che la tensione di snervamento $x$ di un materiale ha deviazione standard $\sigma_x$=10MPa, stimare, con confidenza del 95%, il valor medio della tensione di snervamento avendo a disposizione le seguenti misure (valori in MPa):

220 187 197 196 200

Risposta: $191 < \mu < 209$

# Riassunto

La variabile aleatoria *m*, media campionaria su *n* elementi estratti da una popolazione con distribuzione gaussiana $N\left(x,\mu,\sigma_x\right)$ , produce la seguente V.A. *z* gaussiana standard:

$$z=\frac{m-\mu}{\sigma_x/\sqrt{n}};\qquad \text{con distribuzione: } N\left(z,0,1\right)$$

**Nota.** Se $X$ non è gaussiana, $z$ tende comunque alla stessa distribuzione limite quando *n* è grande;
per distribuzioni originarie campaniformi è sufficiente che: $n>(3\div4)$.

# Previsione della varianza (1/3)

Supponiamo di avere un campione di numerosità $n$ estratto da una V.A. gaussiana $X$:

$$\{x_1, x_2, \ldots, x_n\}$$

Conoscendo il valor medio $\mu$ di $X$, la miglior stima che possiamo fare della varianza sulla base del campione è la seguente:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

Se la media non è nota, la migliore approssimazione che può essere fatta consiste nell'uso della sua stima:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$$

Questa grandezza è però inadatta come stimatore perché:

- $m$ dipende dagli stessi dati per cui ci sono solo $n-1$ informazioni indipendenti sulle variazioni
- lo stimatore non è corretto (calcolando la media dei valori ottenuti da tanti campioni non si ottiene la varianza di $X$)

# Previsione della varianza (2/3)

Lo **stimatore corretto** della varianza, se non si conosce la media, è in effetti il seguente:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - m\right)^2$$

- $s^2$ è chiamata **varianza campionaria** (anche se sarebbe più corretto chiamarlo stimatore campionario della varianza della popolazione)
- $n-1$ sono i gradi di libertà ($g$)
- $s^2$ è uno specifico valore assunto di una V.A. che può essere definita come l'estrazione dal seguente processo:

$$\left\{x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)}\right\} \qquad m_1 = \frac{1}{n}\sum_{i=1}^{n} x_i^{(1)} \qquad s^2{}_1 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^{(1)} - m_1\right)^2$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\left\{x_1^{(j)}, x_2^{(j)}, \ldots, x_n^{(j)}\right\} \qquad m_j = \frac{1}{n}\sum_{i=1}^{n} x_i^{(j)} \qquad s^2{}_j = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^{(j)} - m_j\right)^2$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\left\{x_1^{(M)}, x_2^{(M)}, \ldots, x_n^{(M)}\right\} \qquad m_M = \frac{1}{n}\sum_{i=1}^{n} x_i^{(M)} \qquad s^2{}_M = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^{(M)} - m_M\right)^2$$

- Il valor medio degli $s^2_j$ tende effettivamente al valore esatto $\sigma_x{}^2$

# Previsione della varianza (3/3)

Consideriamo la seguente V.A. :

$$\chi^2 = (n-1)\frac{s^2}{\sigma_x^2}$$

è stato verificato che tale V.A. ha una distribuzione nota (detta distribuzione $\chi^2$ *chi-quadro*) con la seguente espressione:

$$f\left(\chi^2\right) = \frac{\left(\chi^2\right)^{\frac{g-2}{2}}}{2^{\frac{g}{2}} \cdot \Gamma\left(\frac{g}{2}\right)} \cdot e^{-\frac{\chi^2}{2}}$$

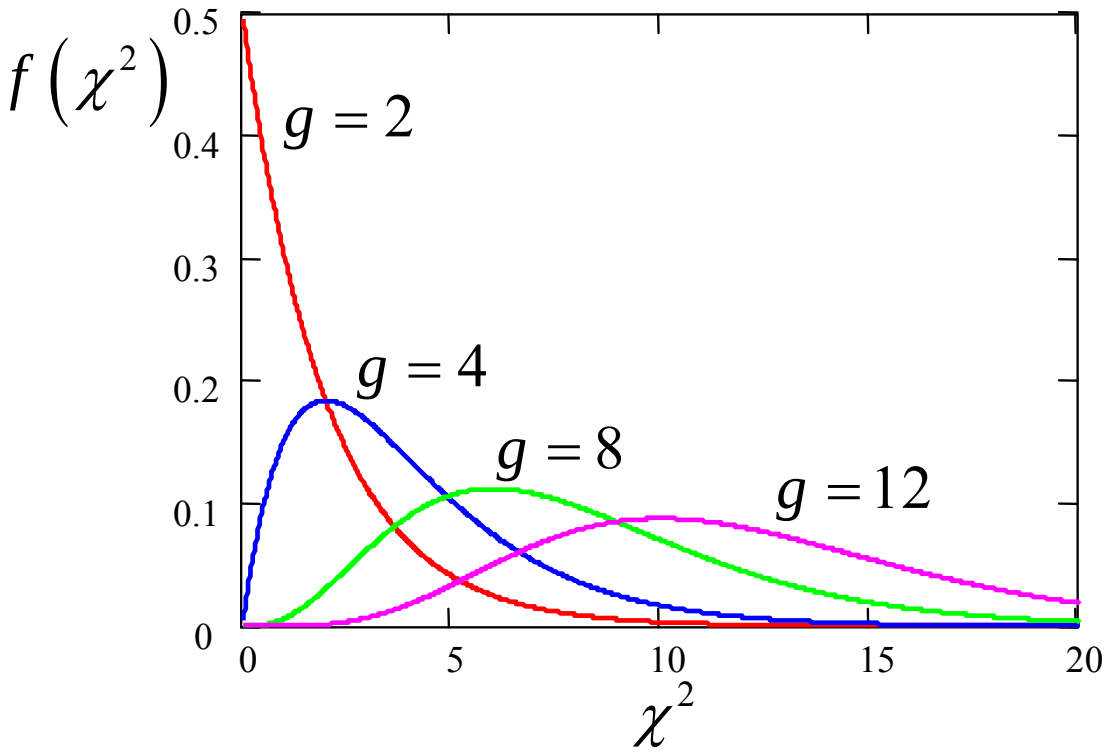
In cui $g = n-1$ sono i gradi di libertà.

La distribuzione *chi-quadro:*

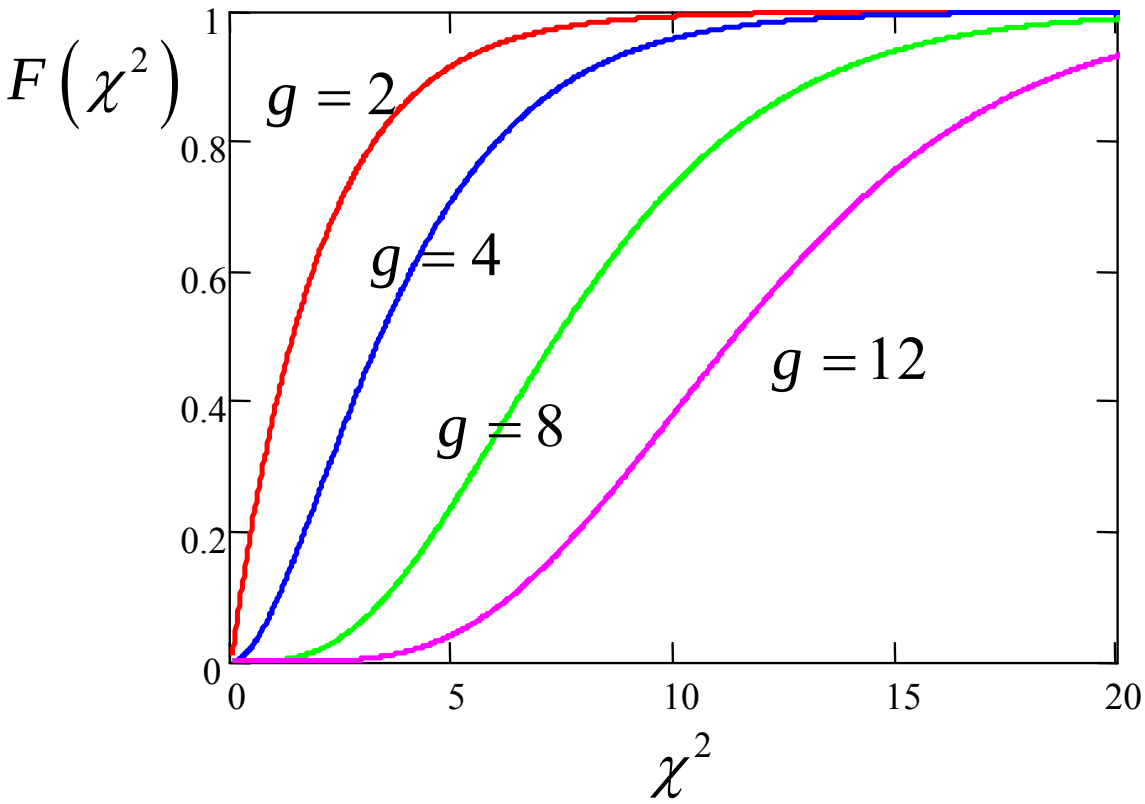
- è definita nei reali positivi
- è parametrizzata in $g$
- ha le seguenti caratteristiche (la varianza è riferita alla media):

$$E\left(\chi^2\right) = g; \qquad VAR\left(\chi^2\right) = 2g$$

# Distribuzione di densità del *chi quadro*



# Distribuzione cumulata del *chi quadro*



# Chi-quadro ridotto

Si introduce anche il *chi-quadro ridotto*:

$$\chi_r^2 = \frac{\chi^2}{g}$$

chiamato anche *chi-quadro per grado di libertà* con distribuzione:

$$f\left(\chi_r^2\right) = g\,\frac{\left(\chi_r^2\right)^{\frac{g-2}{2}}}{2^{\frac{g}{2}}\cdot\Gamma\left(\frac{g}{2}\right)}\cdot e^{-\frac{\chi_r^2}{2}}$$

Il *chi-quadro ridotto* fornisce la distribuzione della quantità:

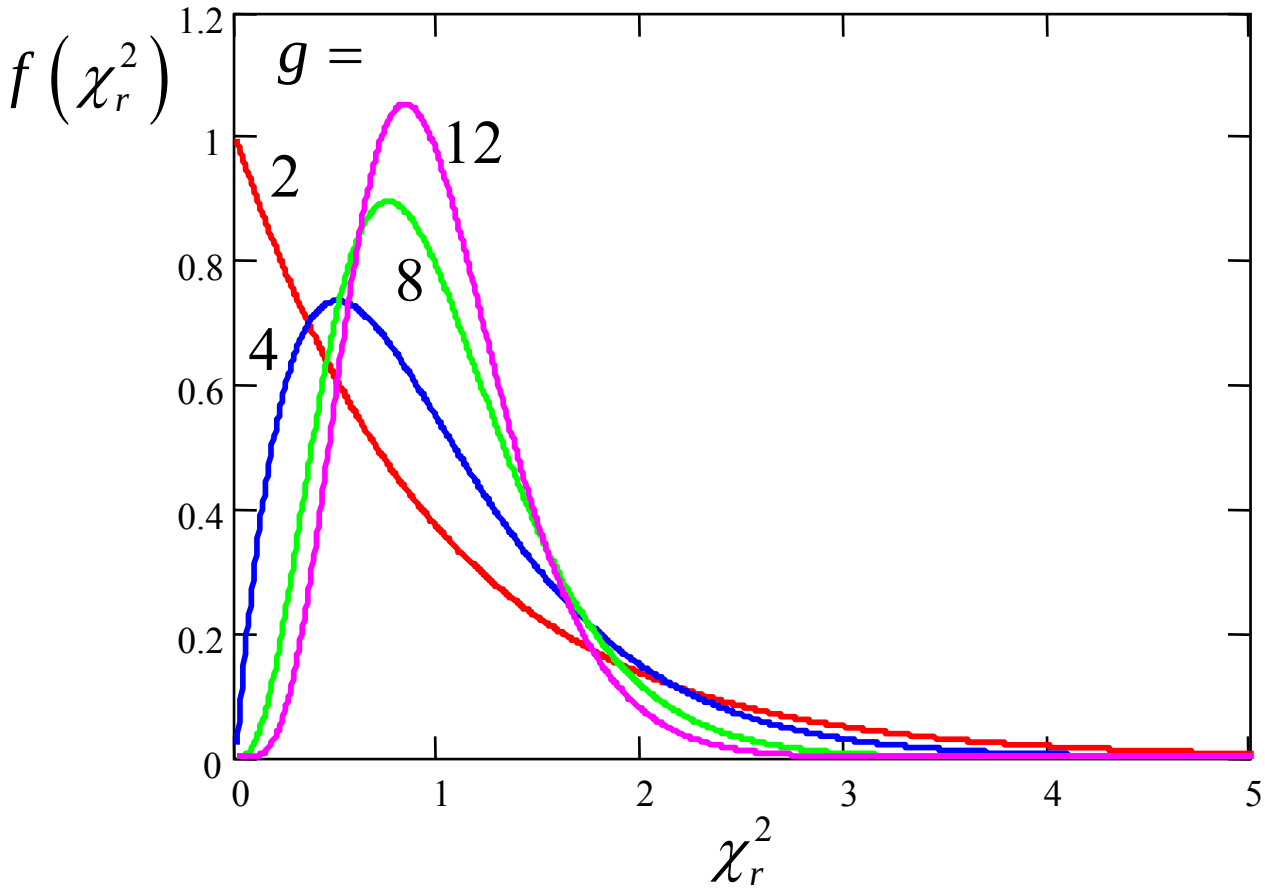
$$\chi_r^2 = \frac{s^2}{\sigma_x^2}$$

La distribuzione *chi-quadro ridotto*:

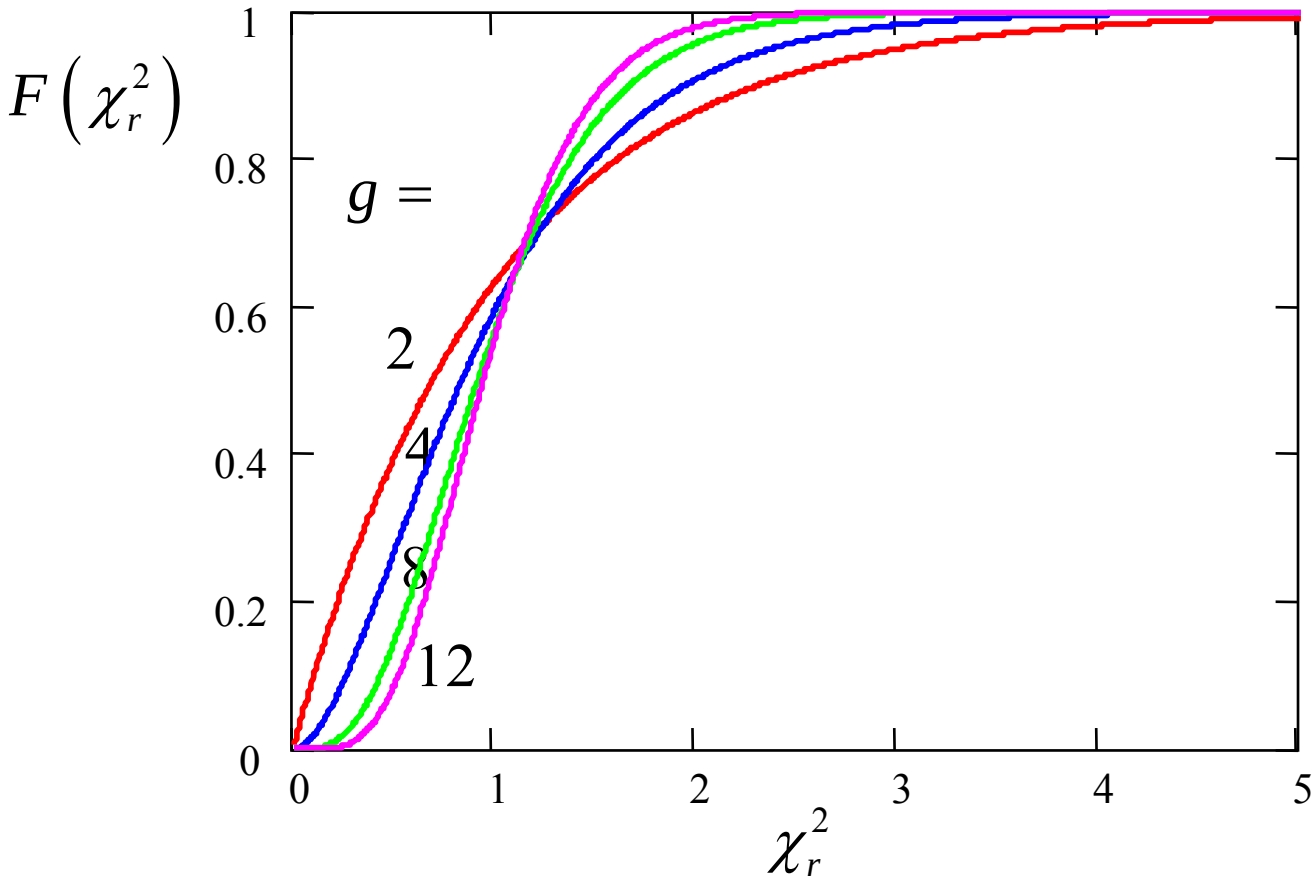
- è definita nei reali positivi
- è parametrizzata in $g$
- ha le seguenti caratteristiche (la varianza è riferita alla media):

$$E\left(\chi_r^2\right) = 1; \qquad VAR\left(\chi_r^2\right) = \frac{2}{g}$$

# Distribuzione di densità del *chi-quadro ridotto*

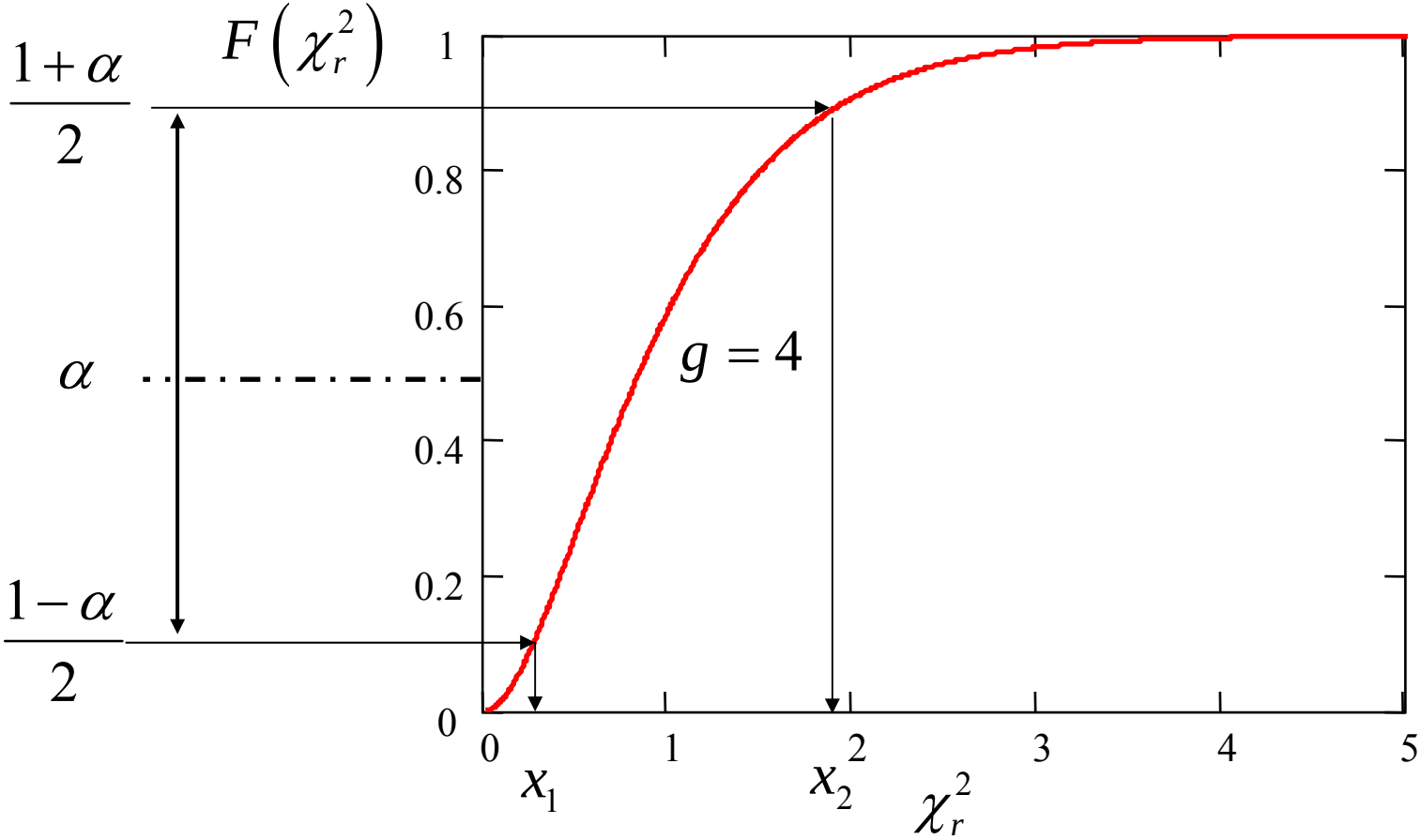


# Distribuzione cumulata del *chi-quadro ridotto*



# Intervalli di confidenza del *chi quadro ridotto*

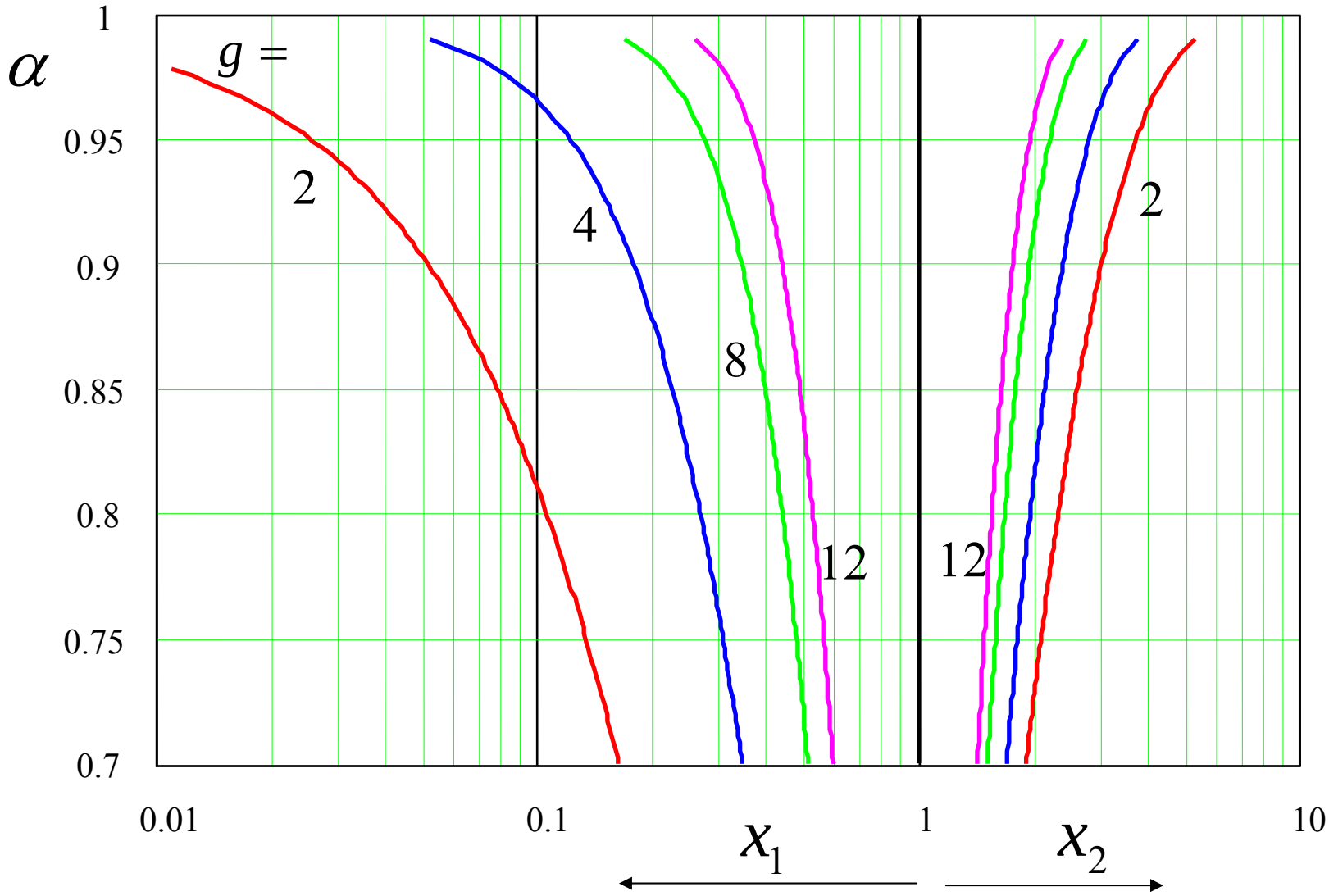
Determinare l'intervallo centrale di valori entro cui si colloca la frazione $\alpha$ dei dati di $\chi_r^2$




Esempio numerico con $g$=4 e $\alpha$=0.8 $\quad x_1 = 0.266; \quad x_2 = 1.945$

$$\chi_1^2 = x_1 g = 1.064; \quad \chi_2^2 = x_2 g = 7.779$$

# Intervalli di confidenza del *chi quadro ridotto*



**Esempio 7.5**

Le $n$=10 misure rappresentano l'usura in dischi di freno ($10^{-1}$mm)

8.05 9.1 6.2 7.5 10.3 5.0 8.53 6.8 9.6 11.46

Determinare gli scarti quadratici medi della popolazione rispettivamente con confidenza del 90% e 99%.

Dal campione: $m = 8.154; \quad s^2 = 3.232; \quad s = 1.798$

Confidenza del 90% $g = 9; \quad \alpha = 0.90$

$$x_1 = 0.369; \quad x_2 = 1.88$$

$$0.9 = P\left( x_1 < \frac{s^2}{\sigma_x^2} < x_2 \right) = P\left( \frac{s^2}{x_2} < \sigma_x^2 < \frac{s^2}{x_1} \right)$$

$$1.719 < \sigma_x^2 < 8.748$$

$$1.31 < \sigma_x < 2.96$$

Con confidenza del 99% : $g = 9; \quad \alpha = 0.99$

$$x_1 = 0.193; \quad x_2 = 2.621$$

$$0.99 = P\left( x_1 < \frac{s^2}{\sigma_x^2} < x_2 \right) = P\left( \frac{s^2}{x_2} < \sigma_x^2 < \frac{s^2}{x_1} \right)$$

$$1.233 < \sigma_x^2 < 16.77$$

$$1.11 < \sigma_x < 4.10$$

## Previsione della media con varianza non nota (1/2)

- Come nell'esempio precedente, tipicamente è noto il solo campione e quindi anche la dispersione della popolazione deve essere stimata

- Spesso quest'ultima può essere stimata con notevole dispersione

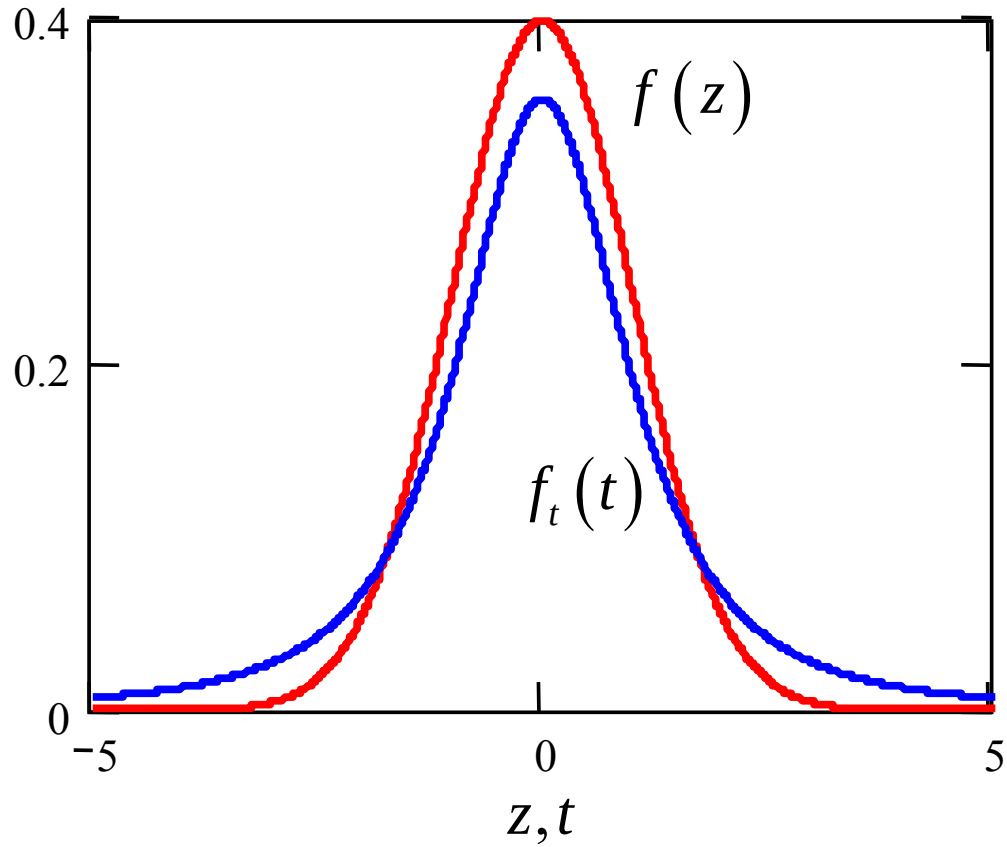
- La non conoscenza della $\sigma$ comporta un degrado del livello di confidenza con cui si può stimare la media della popolazione

- Possiamo comunque sostituire $\sigma$ con il valore $s$ stimato dal campione e quindi considerare la seguente VA:

$$t = \frac{m - \mu}{s / \sqrt{n}} \qquad \text{invece che} \qquad z = \frac{m - \mu}{\sigma / \sqrt{n}}$$

La V.A. $z$ è una Normale standard mentre la $t$ è una V.A. sempre con media nulla ma definita da una distribuzione con maggiore dispersione

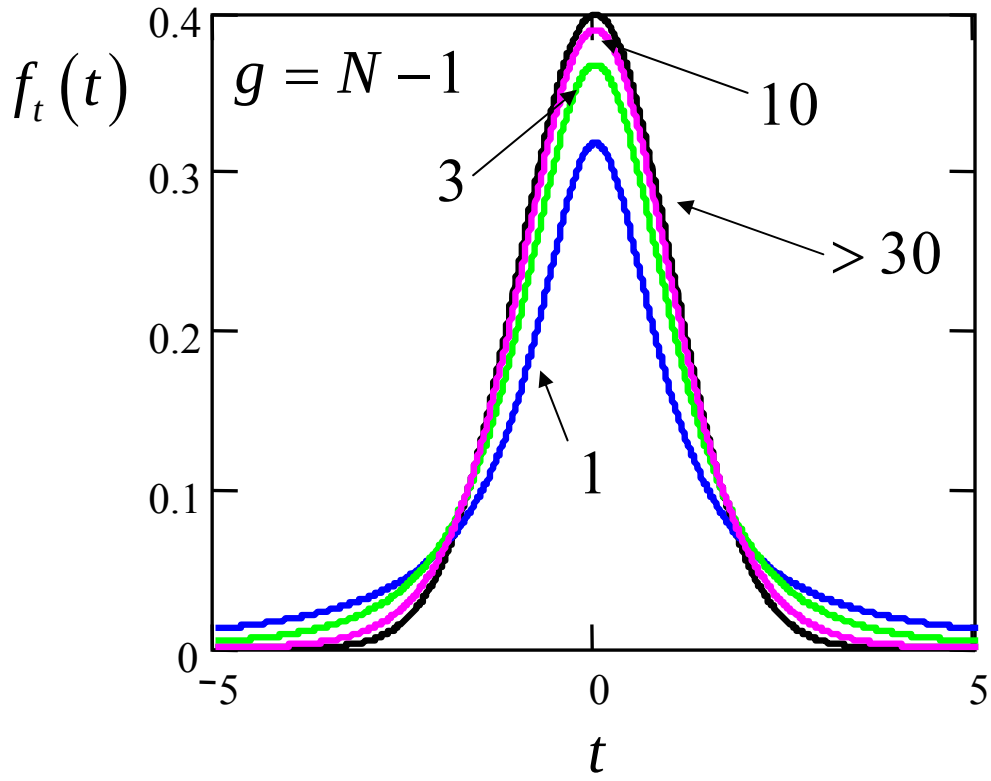
# Previsione della media con varianza non nota (2/2)



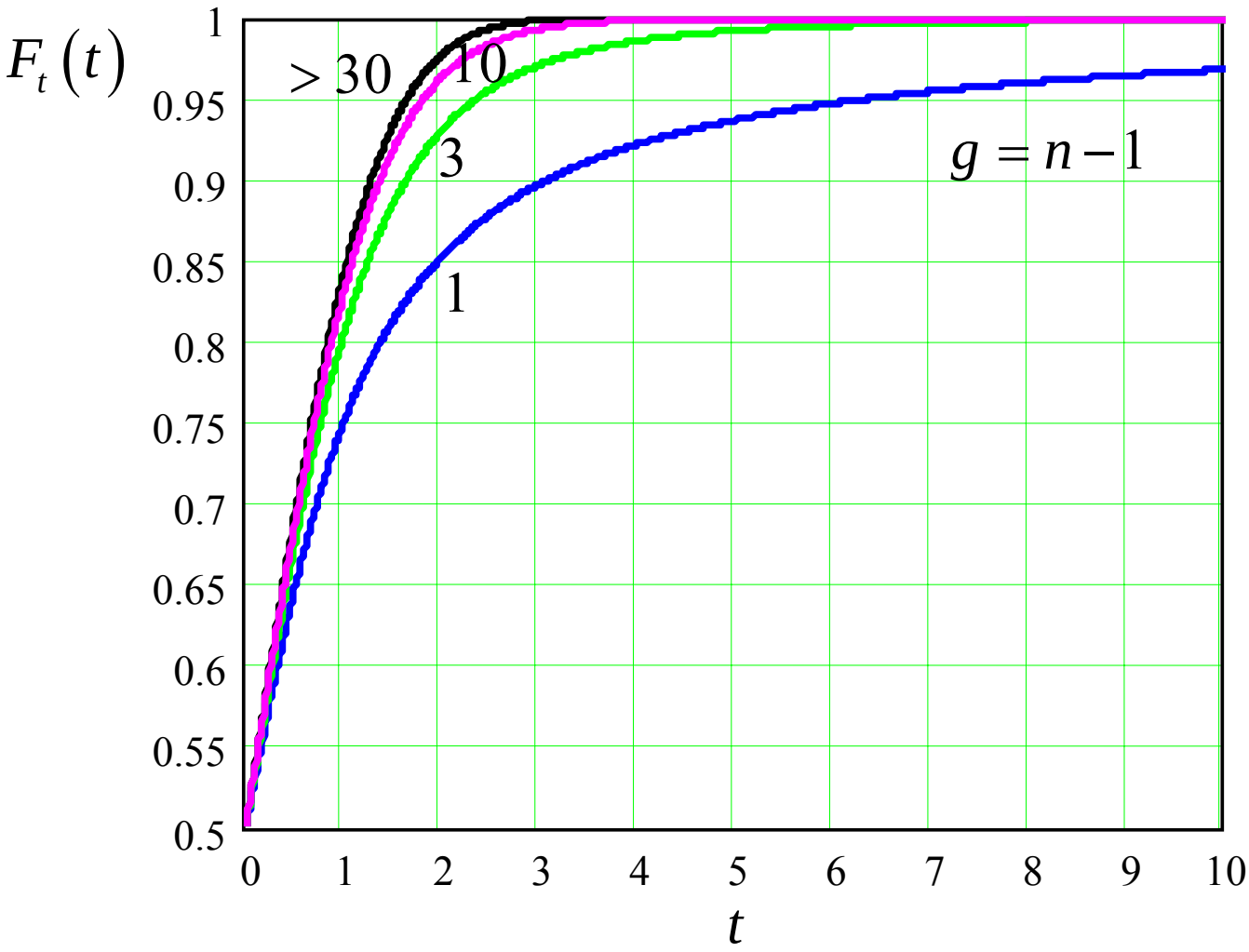

Possiamo prevedere che la distribuzione della V.A. $t$ sia vicina alla V.A. $z$ se la numerosità $n$ del campione è elevata

# Distribuzione *t* di Student

$$f_t(t) = \frac{1}{\pi g} \frac{\Gamma\left(\frac{g+1}{2}\right)}{\Gamma\left(\frac{g}{2}\right)} \cdot \left(1 + \frac{t^2}{g}\right)^{-\frac{g+1}{2}} \qquad g = n - 1$$



# Cumulata della *t* di Student



**Esempio 7.6**

Date le seguenti $n$=6 misure di rottura di barre nominalmente uguali (kN)

9.6 9.85 8.8 9.2 9.45 8.9

valutare con livelli di confidenza $\alpha$: 90% e 99% la media della resistenza.

Dal campione: $m = 9.3; \quad s = 0.409; \quad g = 5$

Dalla cumulata della $t$ di Student con $g$=5:

$$\alpha_1 = 0.90 \qquad \frac{1-\alpha_1}{2} = 0.95$$

$$F_t(t_1) = 0.95 \Rightarrow t_1 = 2.015$$

$$P\left(\frac{|\mu - m|}{s/\sqrt{n}} < t_1\right) = \alpha_1 \quad \Rightarrow \text{con confidenza del 90\%} \quad 8.96 < \mu < 9.64$$

$$\alpha_2 = 0.99 \qquad \frac{1-\alpha_2}{2} = 0.995$$

$$F_t(t_2) = 0.995 \Rightarrow t_2 = 4.03$$

$$P\left(\frac{|\mu - m|}{s/\sqrt{n}} < t_2\right) = \alpha_2 \quad \Rightarrow \text{con confidenza del 99\%} \quad 8.63 < \mu < 9.97$$

# Modelli da campioni

È possibile ricavare informazioni sulla densità di distribuzione della popolazione a partire da un campione?
Esempio: è gaussiana la distribuzione delle tensioni di snervamento?
Oppure: una distribuzione di tempi di guasto è di Weibull?

L'esame degli istogrammi è molto critica e richiede un numero enorme di dati ($n$ elevatissimo)

È opportuno lavorare con la cumulata che è più regolare in quanto integrale della distribuzione

Si considera un campione e si pongono le misure in ordine di valore crescente, a ognuno di tali misure è necessario associare un valore di probabilità cumulata, ovvero la probabilità che la misura sia minore o uguale di quella.

$$\{x_1, x_2, x_3, ...., x_n\} \qquad x_j \leq x_{j+1}; \qquad 0 < j < n$$

Quale probabilità cumulata possiamo attribuire al *j*-esimo termine della sequenza?

Primo tentativo:
Nel campione vi sono $j-0.5$ valori misurati non maggiori quindi possiamo assumere come stima della probabilità cumulata:

$$F\left(x_j\right) \cong F_j = \frac{j-0.5}{n}$$

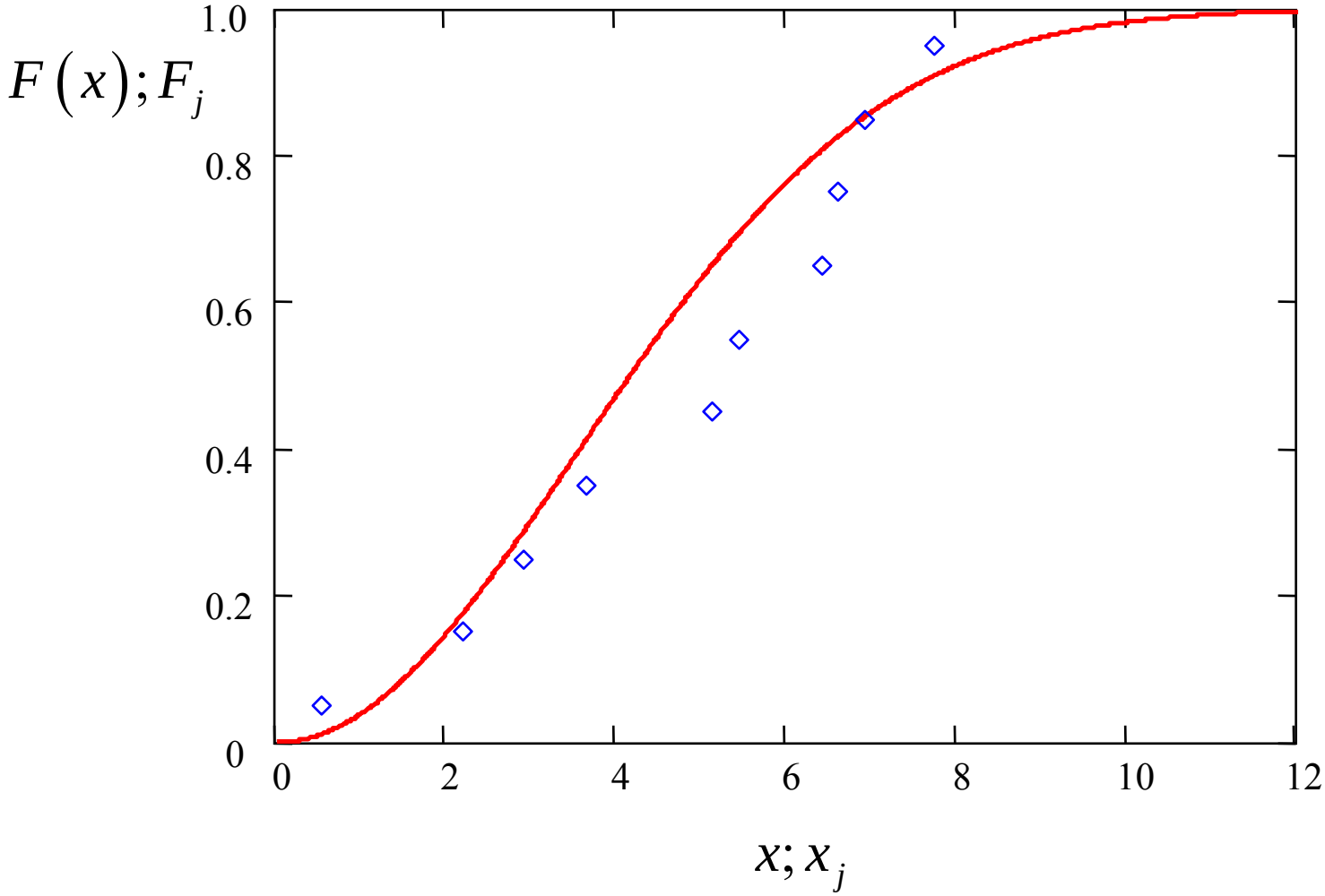
per esempio:

$$F\left(x_1\right) = \frac{0.5}{n}$$

Però, in un altro campione ordinato quasi sicuramente a valori corrispondenti di *x* corrisponderà un diverso numero di elementi che precedono

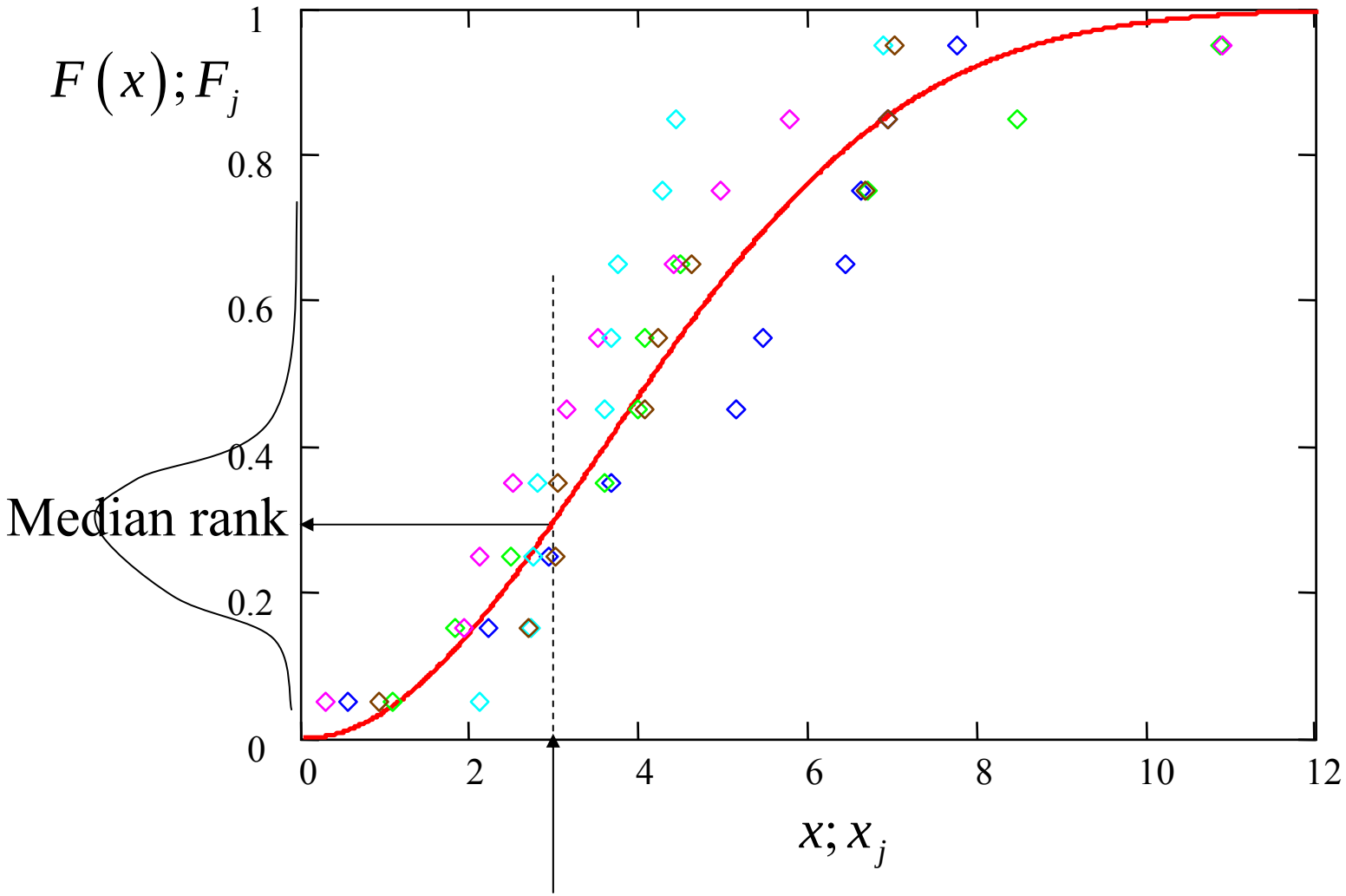
**Esempio 7.7**

Rappresentazione di un campione $n$=10 estratto da una V.A. di Weibull con: $\alpha = 5; \beta = 2$



**Esempio 7.8**

Rappresentazione di alcuni campioni tutti di numerosità $n$=10 estratti da una V.A. di Weibull con: $\alpha = 5; \beta = 2$



Si verifica che fissato un definito quantile $c$, il valore da considerare come $F_j$ per il corrispondente $c$ è dato dalla seguente relazione:

$$F_j^{(c)} = \frac{1}{1 + \frac{n-j+1}{j} \cdot \Phi\left(c, \nu_1, \nu_2\right)}$$

Con $\Phi\left(c, \nu_1, \nu_2\right)$ inversa della distribuzione cumulata di Fisher con gradi di libertà:

$$\nu_1 = 2\left(n-j+1\right); \quad \nu_2 = 2j$$

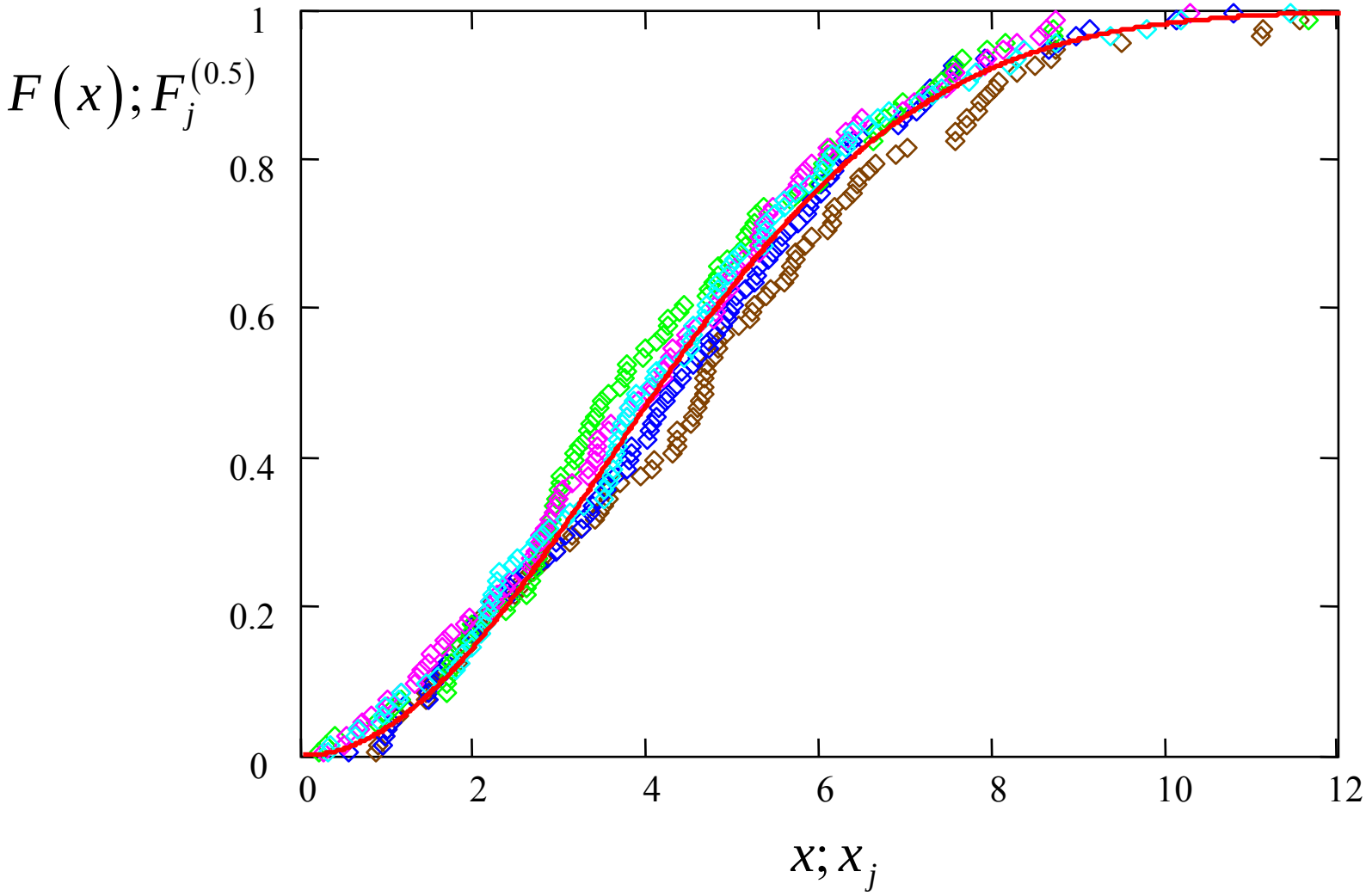
generalmente reperibile in forma tabulare nei manuali di statistica.

Spesso è sufficiente considerare il **rango mediano** (*median rank*) che corrisponde al quantile $c$=0.5. Si può allora usare la seguente espressione approssimata (ma molto accurata):

$$F_j^{(0.5)} \cong \frac{j-0.3}{n+0.4}$$

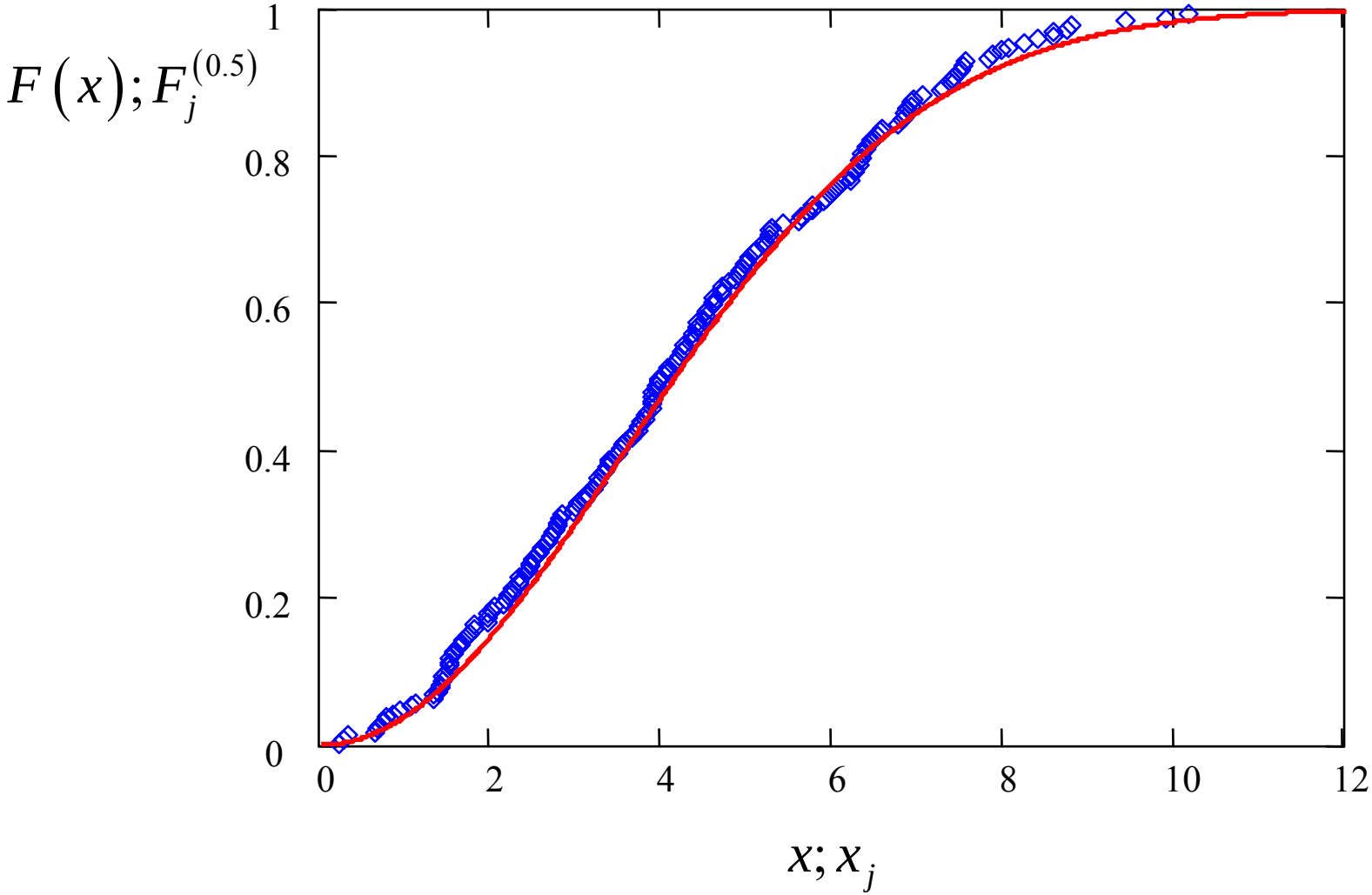
**Esempio 7.9**

Campioni di VA di Weibull con gli stessi parametri dell'esempio 7.8 con numerosità: $n$=100



**Esempio 7.10**

Campione con $n$=200



**Esempio 7.11**

Determinare dal seguente campione $n$=100 se la relativa tensione di snervamento può essere considerata gaussiana

| $j$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | …. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $x_j$ | 250 | 249 | 248 | 235 | 232 | 261 | 224 | 245 | 257 | 249 | 224 | 236 | 259 | 261 | ….. |

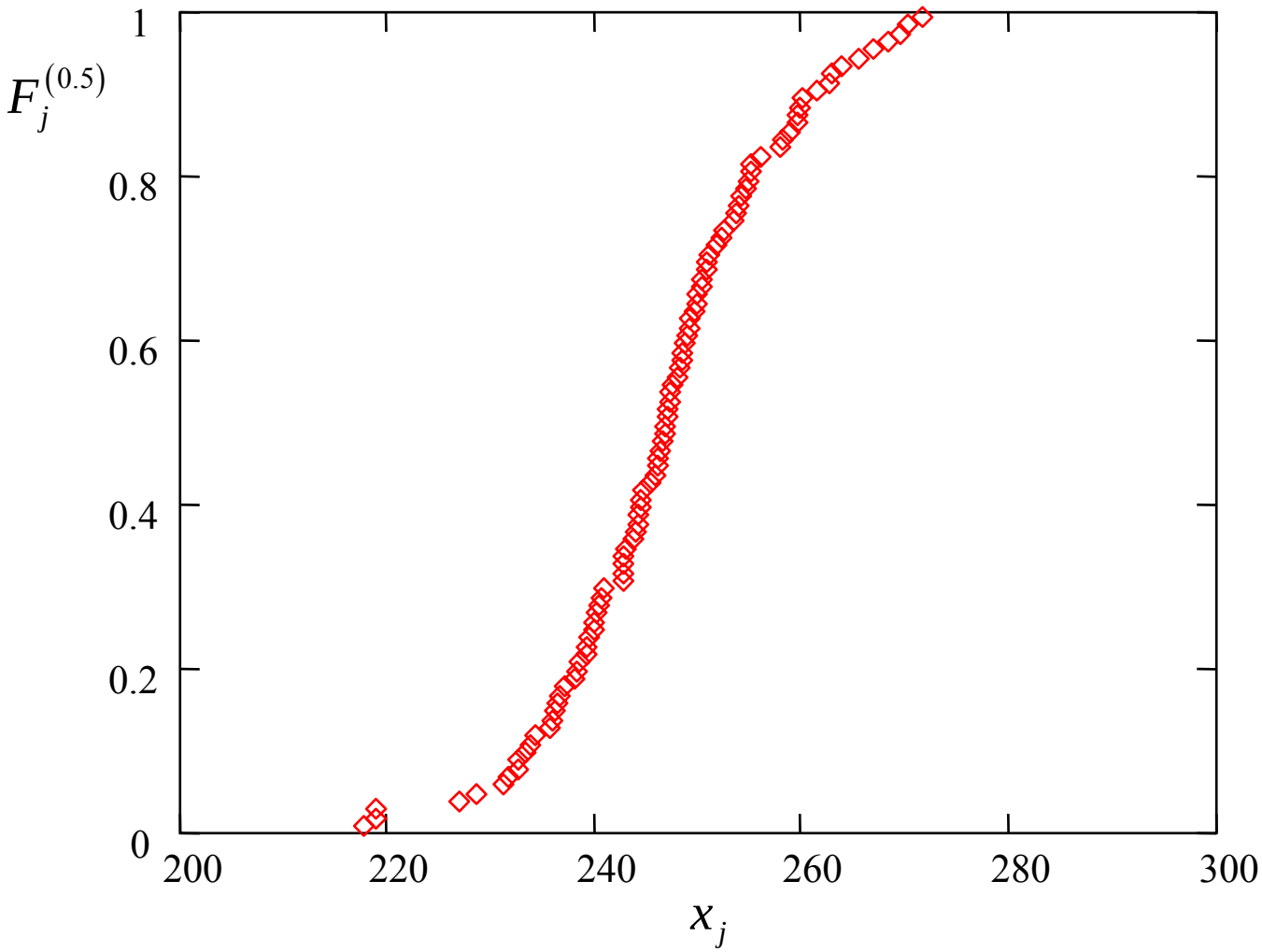
Caratteristiche della popolazione stimate dal campione

$$m = 247; \qquad s = 10.9$$

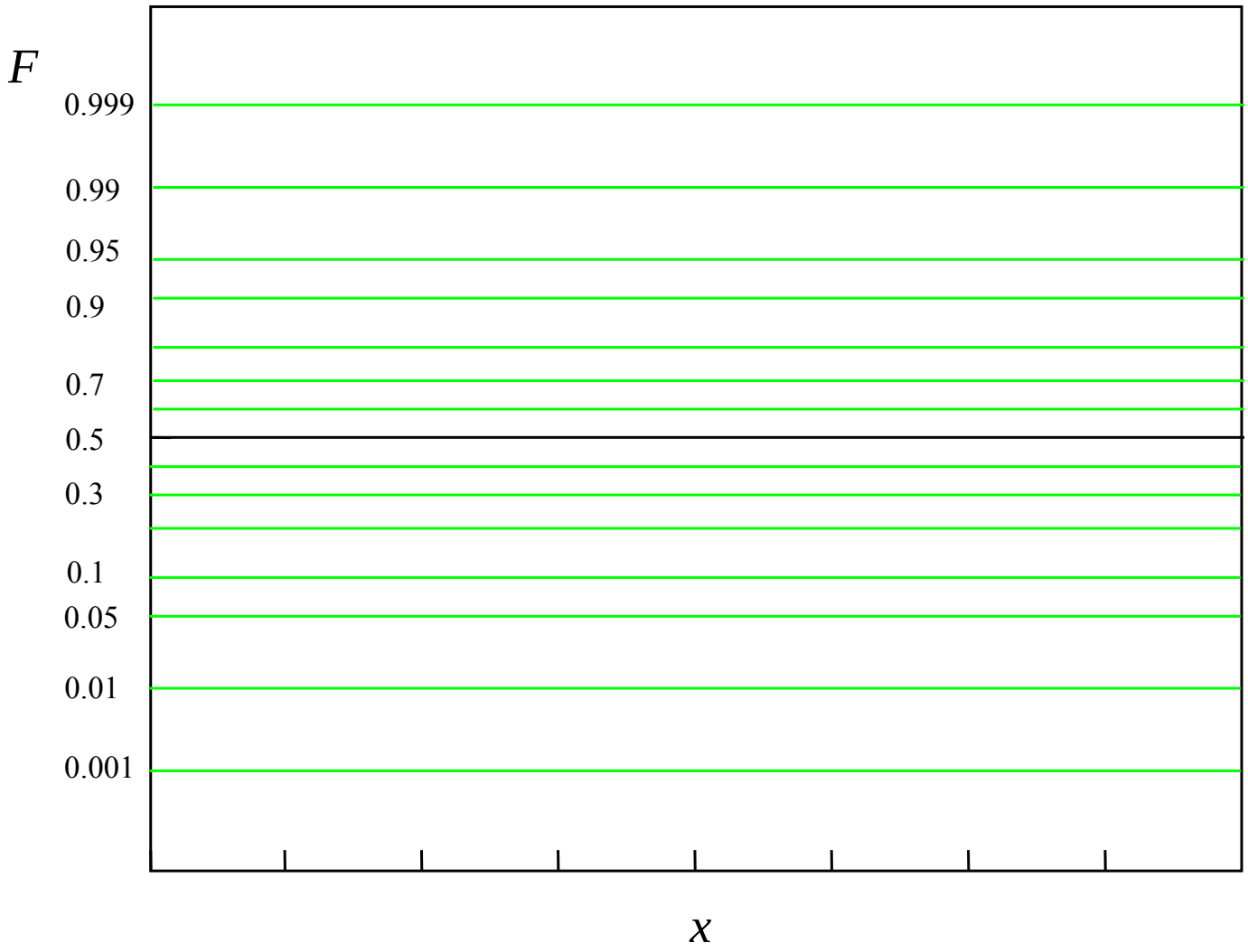
Ordinamento dei valori e calcolo del rango mediano:

| $F_j^{(0.5)}$ | 0.006 | 0.017 | 0.027 | 0.037 | 0.047 | …. | …. |
|---|---|---|---|---|---|---|---|
| $x_j$ | 218 | 219 | 219 | 227 | 228 | ….. | ….. |

# Cumulata campionaria



# Carta normale



# Campione su carta normale (scala delle ordinate distorta)