STATISTICA

Nome media	Adatta a dati	Media semplice	Media ponderata	Casi particolari
ARITMETICA	Additivi	$M(x) = \frac{\sum x_i}{n}$	$M^P(x) = \frac{\sum x_i f_i}{\sum f_i}$	
ARMONICA	Inv. proporzionali	$M_a(x) = \frac{n}{\sum \frac{1}{x_i}}$	$M_a^P(x) = \frac{\sum f_i}{\sum \frac{f_i}{x_i}}$	$x_i = 0 \rightarrow IMP$
GEOMETRICA	Grandezze moltiplicative (tassi di crescita, …)	$M_g(x) = \sqrt[n]{\prod x_i}$ $M_g(x) = e^{\left(\frac{\sum ln(x_i)}{n}\right)}$	$M_g^P(x) = \sqrt[\sum f_i]{\prod x_i^{f_i}}$ $M_g^P(x) = e^{\left(\frac{\sum ln(x_i) f_i}{\sum f_i}\right)}$	$\begin{cases} x_i < 0 \rightarrow N/A \\ x_i = 0 \rightarrow M_g(x) = 0 \end{cases}$
QUADRATICA	Quadratici	$M_2(x) = \sqrt{\frac{\sum x_i^2}{n}}$	$M_2^P(x) = \sqrt{\frac{\sum x_i^2 f_i}{\sum f_i}}$	$\left.\begin{matrix} x_i < 0 \\ x_j > 0 \end{matrix}\right\} \rightarrow N/A$
DI POTENZA	Formula generale da cui derivare tutte le altre medie (anche quelle superiori alla quadratica)	$M_r(x) = \sqrt[r]{\frac{\sum x_i^r}{n}}$	$M_r^P(x) = \sqrt[r]{\frac{\sum x_i^r f_i}{\sum f_i}}$	$\begin{cases} r = -1 \rightarrow M_\alpha(x) \\ r = 0 \rightarrow M_g(x) \\ r = 1 \rightarrow M(x) \\ r = 2 \rightarrow M_2(x) \\ r = 3 \rightarrow M_3(x) \end{cases}$

CONDIZIONE DI INTERNALITÀ

$$x_{min} \leq M_\alpha(x) \leq M_g(x) \leq M(x) \leq M_2(x) \leq \cdots \leq x_{max}$$

COVARIANZA

È un indice che dà un'idea del tipo di relazione che esiste tra due grandezza x e y.

$$\begin{cases} cov < 0 & direttam.\ prop. \\ cov > 0 & inversam. prop. \end{cases}$$

Si calcola tramite due metodi:

$$cov(x,y) = M\big[\big(x - M(x)\big)\big(y - M(y)\big)\big]$$

$$\boldsymbol{cov(x,y) = M(xy) - M(x)M(y)}\ \textit{meno laboriosa}$$

MEDIANA

È il valore che occupa la **POSIZIONE CENTRALE** della distribuzione **ORDINATA**.

Se la distribuzione è di 5 elementi, la mediana sarà il TERZO elemento; se la distribuzione è di 6 elementi, la mediana sarà la COPPIA TERZO-QUARTO elemento; se hanno lo stesso valore, ovviamente non si avrà la coppia ma solo il singolo valore. *ATTENZIONE ai valori ponderati.*

QUANTILI

Sono i genitori della mediana e individuano quelle x che occupano **DETERMINATE POSIZIONI** all'interno di una distribuzione **ORDINATA**.

- **QUARTILI**: individuano le x che si trovano nei QUARTI (sono 4)
- **DECILI**: individuano le x che si trovano nei DECIMI (sono 10)
- **PERCENTILI**: individuano le x che si trovano nei CENTESIMI (sono 100)

x	y
1	20
4	35
12	25
16	40
Σ 33	120

$$Q_1 = 25\% = 30° \; posto = 4$$

$$Q_2 = 50\% = 60° \; posto = mediana = 12$$

$$D_6 = 60\% = 72° \; posto = 12$$

$$D_9 = 90\% = 108° \; posto = 16$$

$$P_{25} = 25\% = 30° \; posto = 4$$

$$P_{75} = 75\% = 90° \; posto = 16$$

MODA

È l'elemento della distribuzione con **MAGGIORE FREQUENZA**. Se c'è più di un elemento con la stessa frequenza, la moda non esiste e la distribuzione si dice ZERO MODALE.

INDICI DI VARIABILITÀ

VALORE	METODO DIRETTO		METODO INDIRETTO
	TIPO	FORMULA	FORMULA
SCARTO QUADRATICO MEDIO $E(x), \sigma(x), \sigma = \sqrt{V(x)}$	SEMPLICE	$\sqrt{\frac{\sum\left[(x_i - M(x))^2\right]}{n}}$	$\sqrt{M(x^2) - (M(x))^2}$
	PONDERATO	$\sqrt{\frac{\sum\left[(x_i - M(x))^2 f_i\right]}{\sum f_i}}$	
VARIANZA $V(x), \sigma^2(x), \sigma^2$	SEMPLICE	$\frac{\sum\left[(x_i - M(x))^2\right]}{n}$	$M(x^2) - (M(x))^2$
	PONDERATO	$\frac{\sum\left[(x_i - M(x))^2 f_i\right]}{\sum f_i}$	

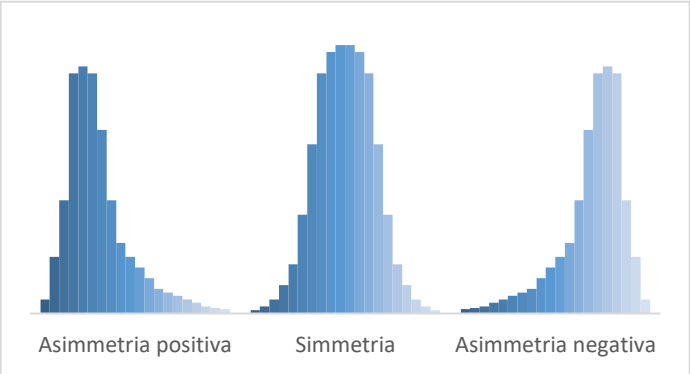
INDICI DI FORMA

Descrivono la "forma" della distribuzione, che può essere **SIMMETRICA** o **APPIATTITA (asimmetrica)**.

Indice di skewness di Pearson

$$sk = \frac{m - moda}{\sigma}$$

- $sk > 0$: asimmetria positiva, coda a destra
- $sk = 0$: simmetria
- $sk < 0$: asimmetria negativa, coda a sinistra



TEST DI INDIPENDENZA NELLE TABELLE A DOPPIA ENTRATA

Verifica se tra due distribuzioni esiste un collegamento (v. dipendenti) o se sono separate (v. indipendenti).

f — tab. freq. osservate

		MIGLIORAMENTO PERCEPITO: SÌ	NO	Σ
SOSTANZA	FARMACO	250	50	300
	PLACEBO	50	50	100
	Σ	300	100	400

f* — tab. freq. teoriche

		MIGLIORAMENTO PERCEPITO: SÌ	NO	Σ_{riga}
SOSTANZA	FARMACO	$\frac{300 \cdot 300}{400} = 225$	$\frac{300 \cdot 100}{400} = 75$	300
	PLACEBO	$\frac{100 \cdot 300}{400} = 75$	$\frac{100 \cdot 100}{400} = 25$	100
	$\Sigma_{colonna}$	300	100	400

Ogni cella della **f*** contiene il risultato dei dati della f secondo questa formula:

$$\boldsymbol{cella\ f^* = \frac{\Sigma_{riga} \cdot \Sigma_{colonna}}{\Sigma_{tot}}}$$

Ora si calcola il **CHI-QUADRATO CALCOLATO** con la formula:

$$\boldsymbol{\chi_C^2 = \sum \frac{(f - f^*)^2}{f^*}}$$

dove f e f^* corrispondono alla stessa cella nelle due rispettive tabelle.

Viene ora confrontato il χ_C^2 con il **CHI-QUADRATO TEORICO** $\boldsymbol{\chi_T^2}$, ottenuto dalla tabella dei valori teorici in base ai **GDL** $\boldsymbol{\nu}$ (Gradi Di Libertà) e al livello di **SIGNIFICATIVITÀ** $\boldsymbol{\alpha}$ (che è indicato dal testo dell'esercizio e solitamente vale 1% o 5%).

$$\boldsymbol{GDL = \nu = (r-1)(c-1) = (n°righe - 1)(n°colonne - 1)}$$

In base ai chi-quadrati calcolati otteniamo:

$$\begin{cases} \chi_C^2 < \chi_T^2 & variabili\ indipendenti \\ \chi_C^2 > \chi_T^2 & variabili\ dipendenti \end{cases}$$

V DI CRAMER

Serve a calcolare il valore di connessione tra due variabili non appena si è scoperto che sono effettivamente collegate.

$$\boldsymbol{V = \sqrt{\frac{\chi_C^2}{\Sigma_{tot} \cdot (min(r,c) - 1)}}} \;\; con \;\; 0 \leq V \leq 1 \;\; [conn.nulla \ldots conn.totale]$$

REGRESSIONE LINEARE

La retta di regressione lineare descrive l'andamento ideale della distribuzione.

$$y' = bx + a \rightarrow \begin{cases} b = \dfrac{cov(x,y)}{\sigma^2(x)} \\ a = M(y) - bM(x) \end{cases} = \begin{cases} na + b\sum x_i = \sum y_i \\ a\sum x_i + b\sum x_i^2 = \sum x_i y_i \end{cases}$$

COEFFICIENTE DI CORRELAZIONE LINEARE E DI DETERMINAZIONE (BONTÀ DI ACCOSTAMENTO)

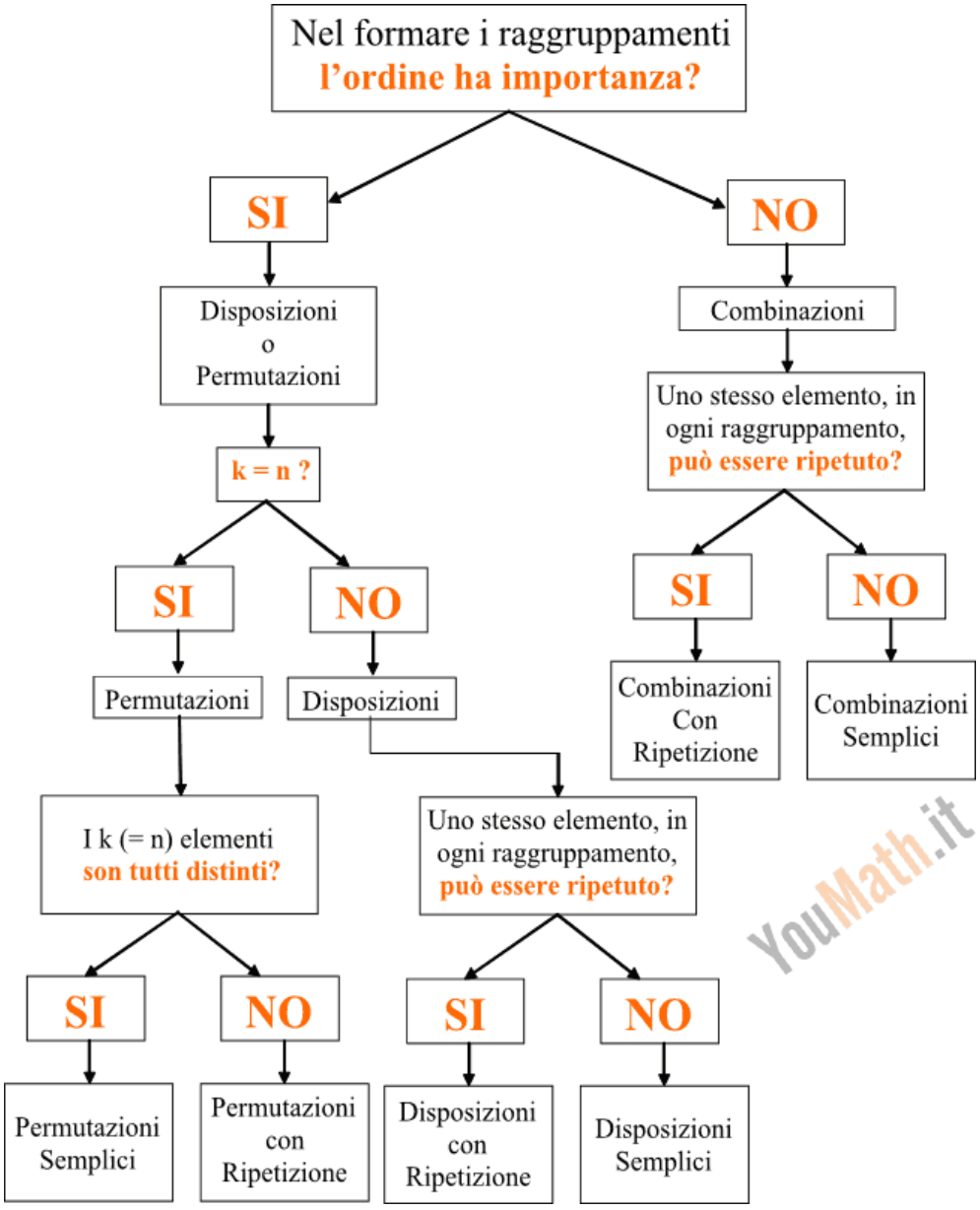
Dopo aver stabilito la retta di regressione, usiamo questi due coefficienti per ottenere informazioni riguardo la relazione tra i due fenomeni studiati.

$$coeff.corr.lineare: \; r = \frac{cov(x,y)}{\sigma(x) \cdot \sigma(y)} \;\; con \;\; -1 \leq r \leq 1 \;\; dove \begin{cases} r = -1 & relazione\ inversa \\ r = 0 & relazione\ lineare \\ r = 1 & relazione\ diretta \end{cases}$$

$$coeff.di\ determinazione: \; 0 \leq r^2 \leq 1 \;\; dove \begin{cases} r^2 = 0 & modello\ scarso \\ 0 < r^2 < 1 & modello\ buono \\ r^2 = 1 & modello\ perfetto \end{cases}$$

Calcolo combinatorio

$\begin{cases} n = n° \ elementi \\ s = n° \ spazi \ (= k) \end{cases}$	**DISPOSIZIONI**	**PERMUTAZIONI**	**COMBINAZIONI**
Distinzione per	ORDINE e DIVERSITÀ	solo ORDINE	solo DIVERSITÀ
Semplici	$D_{n,s} = \frac{n!}{(n-s)!}$	$P_n = n!$	$C_{n,s} = \binom{n}{s} = \frac{n!}{(n-s)!\,s!}$
Ripetizioni	$D'_{n,s} = n^s$	$P_n^{n_1,\dots,n_k} = \frac{n!}{n_1!\dots n_k!}$	$C'_{n,s} = \binom{n+s-1}{s} = \frac{(n+s-1)!}{(n-1)!\,s!}$



Probabilità

DFINIZIONE CLASSICA

$$P(x) = \frac{n^\circ\ casi\ favorevoli}{n^\circ\ casi\ possibili}$$

DEFINIZIONE ASSIOMATICA DI KOLMOGOV

$\begin{cases} 0 \le P(e) \le 1 \\ P(e_{IMPOS}) = 0 \\ P(e_{CERTO}) = 1 \\ P(e^c) = 1 - P(e) \end{cases}$	**COMPATIBILI**		**INCOMPATIBILI** (uno esclude l'altro)
$E_1 \cup E_2$	$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$		$P(E_1 \cup E_2) = P(E_1) + P(E_2)$
$E_1 \cap E_2$	**DIPENDENTI** (uno è responsabile dell'andamento dell'altro)	**INDIPENDENTI**	$P(E_1 \cap E_2) = 0$
	$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2 \mid E_1)$	$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2)$	

VARIABILI CASUALI (O ALEATORIE O STOCASTICHE)

$\begin{cases} n, p, m, \sigma, \alpha\ \ forniti \\ q = 1 - p \end{cases}$		**DESCRIZIONE**	**FORMULA**	**MOMENTI**		
			$P(x)$	$M(x)$	$V(x)$	$E(x)$
DISCRETE	**BINOMIALE**	Numero di volte in cui si verifica un evento su n prove	$\binom{n}{k} p^x q^{n-x}$	np	npq	$\sqrt{npq}$
	DI POSISSON (degli eventi rari)	Numero di volte in cui si verifica un evento di prob. infinitesima in un numero infinito di tentativi	$\frac{e^{-m} \cdot m^x}{x!}$	m	m	$\sqrt{m}$
	GEOMETRICA	Numero di insuccessi da sopportare prima di ottenere un successo	pq^x	$\frac{q}{p}$	$\frac{q}{p^2}$	$\frac{\sqrt{q}}{p}$
CONTINUE	**NORMALE** (gaussiana)	Descrive la curva della distribuzione classica (come si nota dai MOMENTI)	$\frac{e^{-\frac{1}{2}\left(\frac{m-x}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}}$	m	σ^2	σ
	STANDARDIZZATA	Semplificazione della NORMALE; si impostano i $M(x)$ e $V(x)$ come mostrato e si trasforma $x \to u$; poi si usano le probabilità delle u ottenute (che si trovano in una tabella)	$u = \frac{x - m}{\sigma}$	0	1	σ
	ESPONENZIALE NEGATIVA	Descrive i tassi di mortalità	$\alpha \cdot e^{-\alpha x}$	$\frac{1}{\alpha}$	$\frac{1}{\alpha^2}$	$\frac{1}{\alpha}$

INFERENZA

È una tecnica che viene utilizza per descrivere dati relativi ad una POPOLAZIONE conoscendo però solo un suo sottoinsieme di elementi (scelti il più casualmente possibile) detto CAMPIONE.

Abbiamo visto solo un metodo, la VERIFICA DI IPOTESI, dove inizialmente ci si prefissa due ipotesi:

- **IPOTESI NULLA H_0**: ovvero il risultato che vorremmo ottenere
- **IPOTESI ALTERNATIVA H_1**: ovvero il risultato opposto a quello aspettato

	H_0	H_1	CALCOLATO	TEORICO
MEDIA SINGOLA	$\mu = \mu_0$	$\mu \neq \mu_0$	$\mu_C = \frac{M - \mu_0}{\frac{\sigma}{\sqrt{n}}}$	$\mu_T = 50\% - \frac{\alpha}{2}$
MEDIA DOPPIA	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$\mu_C = \frac{m_1 - m_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_1^2}{n_2}}}$	$\mu_T = 50\% - \frac{\alpha}{2}$
μ_T da controllare sulla tabella della normale standardizzata				
DATI APPAIATI	$\Delta \leq 0$	$\Delta > 0$	$t_C = \frac{dm}{\sqrt{\frac{s^2 d}{n}}} = \frac{M(x-y)}{\sigma_d}\sqrt{n-1}$ $s^2 d = \sigma_d^2 \cdot \frac{n}{n-1}$	$t_T = \begin{cases} \nu: \ n-1 \ GDL \\ \alpha: \ dal \ testo \ \% \end{cases}$
t_T da controllare sulla tabella del χ_T^2				

INTERVALLO DI CONFIDENZA PER LA MEDIA DELLA POPOLAZIONE

$$P\left(m - \mu_T \cdot \frac{\sigma}{\sqrt{n}} m < \mu < m + \mu_T \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Codice R

Media di una tabella

```
> labels = c("A", "B", "C")
> vals1 = c(24, 27, 21)
> vals2 = c(18, 21, 25)
.. vals<n>
> table = data.frame(labels, vals1, vals2, .., vals<n>)
# rowMeans(<values_table[<rows>, <cols>]>)
> means = rowMeans(table[1:3,2:<n>])
> means = round(means, 2)
> voti=data.frame(labels, vals1, vals2, .., vals<n>, media)
# barplot(<values>, names.arg=<labels>)
> barplot(vals2, names.arg=labels)
# barplot(table$vals2, names.arg=table$labels)
# barplot(table[,3], names.arg=labels)
```

Simmetria

```
# γ = 0 -> distribuzione simmetrica
# γ < 0 -> distribuzione asimmetrica negativa
# γ > 0 -> distribuzione asimmetrica positiva
> gamma = function(x) {
        m3 = mean((x-mean(x))^3)
        skew = m3/(sd(x)^3)
        skew
}
# β = 3 -> distribuzione MESOCURTICA
# β < 3 -> distribuzione PLATICURTICA
# β > 3 -> distribuzione LEPTOCURTICA
> beta = function(x) {
        m4 = mean((x-mean(x))^4)
        curt = m4/(sd(x)^4)
        curt
}
# γ2 = 0 -> distribuzione MESOCURTICA
# γ2 < 0 -> distribuzione PLATICURTICA
# γ2 > 0 -> distribuzione LEPTOCURTICA
> gamma2 = function(x) {
        beta(x) - 3
}
> x = c(0, 1, 1, 2, 2, 3, 4, 5)
```

TABELLA A DOPPIA ENTRATA

```
> matrx = matrix(c(122, 203, 167, 118, 528, 178, 673, 212), nrow=4, byrow=TRUE)
> labelsRows = c("A", "B", "C", "D")
> labelsCols = c("X", "Y")
> dimnames(titanic) = list(labelsRows, labelsCols)
> matrx
    X   Y
A   122 203
B   167 118
C   528 178
D   673 212
> mosaicplot(matrx)
# test chi quadrato, correct=FALSE se matrice 2x2
> testchiq=chisq.test(matrx, correct=TRUE)
> testchiq
       Pearson's Chi-squared test
data: STAGE
X-squared = 190.4011, df = 3, p-value < 2.2e-16
# χ²_C = 190.4011 ≫ χ²_T = 11.35, con α = 1% e ν = 3 GDL, si rifiuta l'ipotesi nulla di indipendenza e
si conferma la connessione fra i fenomeni

# calcolo il V di Cramer
> chiq = testchiq$statistic
> chiq
X-squared
190.4011
# totale elementi
> N = sum(matrx)
# V = sqrt( χ²_C / (N·(min(<rows>,<cols>)−1)) )
> V=sqrt( chiq / (N * (2-1)) )
> V
X-squared
0.2941201
# esiste una discreta connessione fra i due fenomeni
```

REGRESSIONE LINEARE

```
> Xs = c(30, 50, 40, 85, 60, 80, 70)
> Ys = c(10, 18, 16, 30, 20, 28, 26)
> line = lm(Ys~Xs)
> plot(Xs, Ys)
> abline(line, col="blue")
> segments(Xs, fitted(line), Xs, Ys, lty=2)
> title(main="Regressione lineare fra X e Y")
> summary (rettastat)
Call:
lm(formula = voto ~ orestudio)

Residuals:
    1       2       3       4        5        6       7
-0.97167 0.08215 1.55524 -0.07365 -1.39093 -0.33711 1.13598


Coefficients:
       Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.55241 1.43653 0.385 0.716
orestudio 0.34731 0.02308 15.050 2.35e-05 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 1.159 on 5 degrees of freedom
Multiple R-squared: 0.9784, Adjusted R-squared: 0.9741
F-statistic: 226.5 on 1 and 5 DF, p-value: 2.346e-05

# a=0,55241; b=0,37431
# Y’ = 0.37431*X + 0.55241
# analisi dei residui
> plot(fitted(rettastat), residuals(rettastat))
> abline(0, 0)

# coeff. di correlazione lineare
> R = cor(Xs, Ys)
> R
[1] 0.9891421
# R molto vicino a 1, forte relazione lineare diretta fra le due variabili

# coeff. di determinazione (bontà di accostamento)
> R2 = R^2
> R2
[1] 0.978402
# R2 molto vicino a 1, modello teorico usato (retta) si adatta molto bene ai valori osservati
(tabella); R2 è presente anche nella terza parte dell’output della summary()
```

VARIABILE CASUALE DISCRETA BINOMIALE

```
> dbinom() # calcolo probabilità per ogni k
> pbinom() # probabilità cumulata da 0 a k (somma prob. da 0 a k)
> qbinom() # inversa della probabilità cumulata (ottiene k da una percentuale)
> rbinom() # simula la variabile casuale t volte
```

```
### PROB. DI PASSARE UNA MATERIA 70%, 5 STUDENTI
# vettore dei k
> k = c(0:5)
# calcolo delle probabilità da 0 a 5
# dbinom(k, n, p)
> passato = dbinom(k, 5, 0.7)
> probs
[1] 0.00243 0.02835 0.13230 0.30870 0.36015 0.16807
> barplot(passato, names.arg=k)
# > plot(k, passato, 'h') ; 'h' per un grafico a linee

### 30 DOMANDE A CROCETTA, 3 SCELTE PER DOMANDA
# calcolo probabilità cumulata da 0 a 10
# pbinom(k_max, n, p)
> pbinom(10, 30, 1/3)
[1] 0.5847596

# ottenimento della k che da percentuale 50%
# qbinom(percent, n, p)
> qbinom(0.5, 30, 1/3)
[1] 10

# simula t volte l'evento
# rbinom(t, n, p)
> rbinom(5, 30, 1/3)
[1] 7 10 13 11 7
```

VARIABILE CASUALE DISCRETA DI POISSON

```
> dpois() # calcolo probabilità per ogni k
> ppois() # probabilità cumulata da 0 a k (somma prob. da 0 a k)
> qpois() # inversa della probabilità cumulata (ottiene k da una percentuale)
> rpois() # simula la variabile casuale t volte e ne scrive i risultati
```

```
# calcolo delle probabilità con media 2%
# dpois(k, λ)
> dpois(k, 2)
[1] 0.13533528 0.27067057 0.27067057 0.18044704 0.09022352 0.03608941
```

VARIABILE CASUALE CONTINUA NORMALE

```
> dnorm() # calcolo probabilità per ogni k
> pnorm() # probabilità cumulata da 0 a k (somma prob. da 0 a k)
> qnorm() # inversa della probabilità cumulata (ottiene k da una percentuale)
> rnorm() # simula la variabile casuale t volte e ne scrive i risultati
```

```
# creo l'asse delle x per contenere tutti i valori necessari, quindi da 120 a 240 cm
> x = seq(120, 240, by=0.01)
# creo la distribuzione normale
# dnorm(x, m, σ)
> distr = dnorm(x, 168, 12)
> plot(x, distr, type="l", xlab="Altezza in cm", ylab="Densità di probabilità", col="red")

# lower.tail=TRUE  cumulativa con i valori a sinistra
# lower.tail=FALSE cumulativa con i valori a destra
# pnorm(x, mean=<m>, sd=<σ>, lower.tail=<TRUE,FALSE>)
> pnorm(50, mean=60, sd=20, lower.tail=TRUE)
[1] 0.3085375

# il valore, che a partire da sinistra, include la distribuzione alla percentuale data
# qnorm(%, m, sd)
> qnorm(0.70, 60, 20)
[1] 70.48801 #il 70% della distribuzione è incluso prima del valore ottenuto

# ottiene 100 valori reali distribuiti sulla normale con media 10 e scarto 3
> rnorm(100, 10, 3)
```

VERIFICA DI IPOTESI SINGOLA

```
# verifica sul campione camp con
# H0:mu=7 e H1:mu!=7 ("two.sided")
# alpha=1% (conf.level = 1 - alpha)
> camp = c(7, 6.5, 7.5, 6, 5, 4.5, 8, 7, 7, 6.5)
> t.test(camp, mu=7, alternative="two.sided", conf.level=0.99)
# alternative="<two.sided>,<greater>,<less>"
```

VERIFICA DI IPOTESI SU DUE MEDIE

```
# verifica sulle medie con
# H0:M(qi.f)=M(qi.m) e H1:M(qi.f)!=M(qi.m) ("var.equal=TRUE")
# alpha=1% (conf.level = 1 - alpha)
> m1 = c(105, 99, 115, 130, 80)
> m2 = c(103, 104, 115, 118, 106)
> t.test(m1, m2, var.equal=TRUE, conf.level=0.99)
```

VERIFICA DI IPOTESI SU DATI APPAIATI

```
# verifica sul campione camp con
# H0:before-after<=0 e H1:before-after>0 ("greater")
# alpha=5% (conf.level = 1 - alpha)
> before = c(80, 76, 92, 87, 79, 86, 98, 92, 90, 89, 85, 88)
> after  = c(76, 75, 86, 82, 78, 86, 95, 90, 88, 85, 84, 86)
> t.test(before, after, alternative="greater", conf.level=0.95)
# alternative="<two.sided>,<greater>,<less>"
```