open source

Last updated Maggio 3, 2023

Test di Correlazione e Chi Quadro
Lezione 7

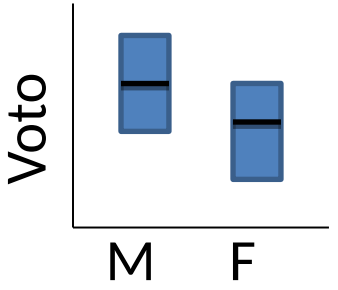
G. Bacaro

Statistica
CdL in Scienze e Tecnologie per l'Ambiente e la Natura
I anno, II semestre

Principali analisi statistiche

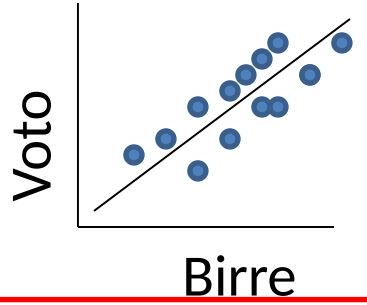
1. Confronto fra medie (2 o piú campioni)

Variabile quantitativa in funzione di una categorica (es. voto più alto M vs. F)



2. Correlazione e regressione

Relazione fra due variabile quantitative
(es. il voto medio dipende dal consumo di birre?)



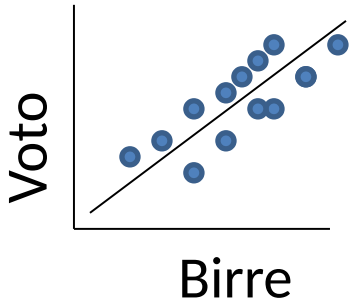
3. Analisi di tabelle di contigenza

Conteggi con due o più variabili categoriche
(es. essere astemi dipende dal genere?)

	Birre	
	SÌ	NO
M		
F		

Correlazione

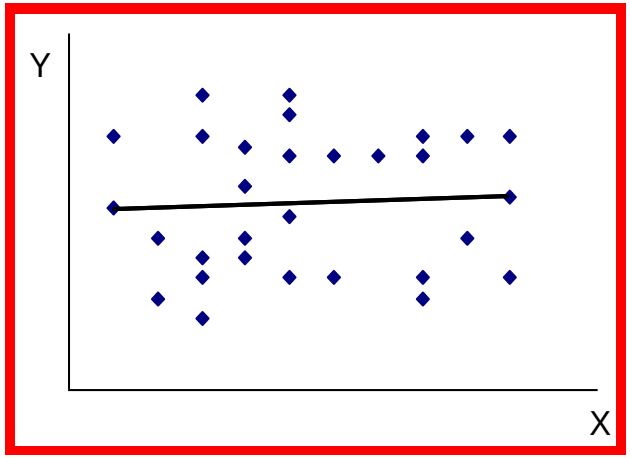
Legame - Associazione - Accordo – Relazione tra variabili

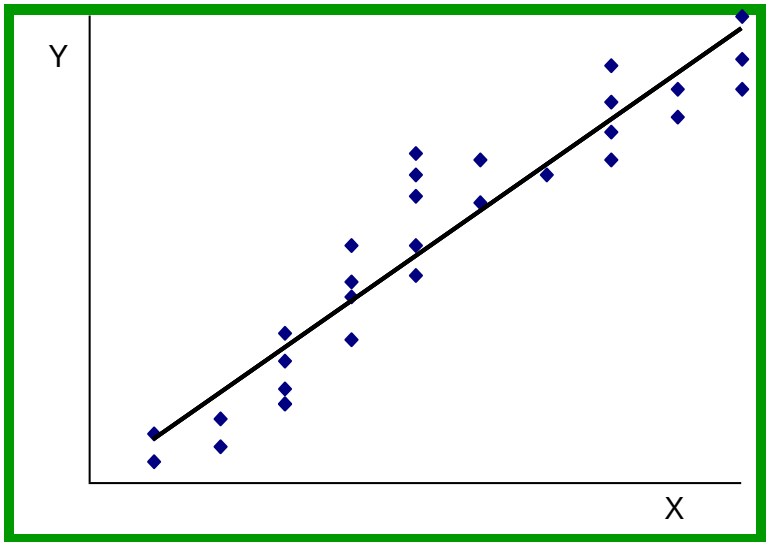


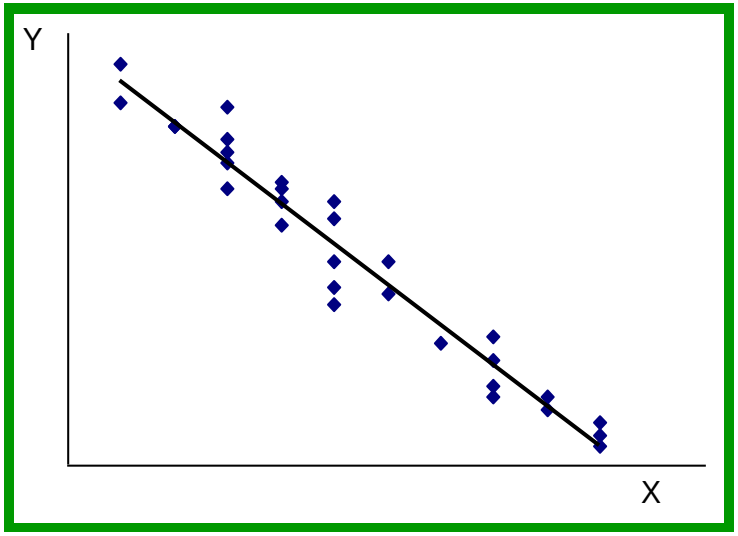
- valutare il grado di reciproca influenza tra due variabili;

- valutare il grado di associazione di due variabili che sono influenzate entrambe da una causa esterna.

La relazione esistente tra due variabili può essere analizzata graficamente ponendo i dati osservati in un diagramma a dispersione :







Il Coefficiente di Correlazione

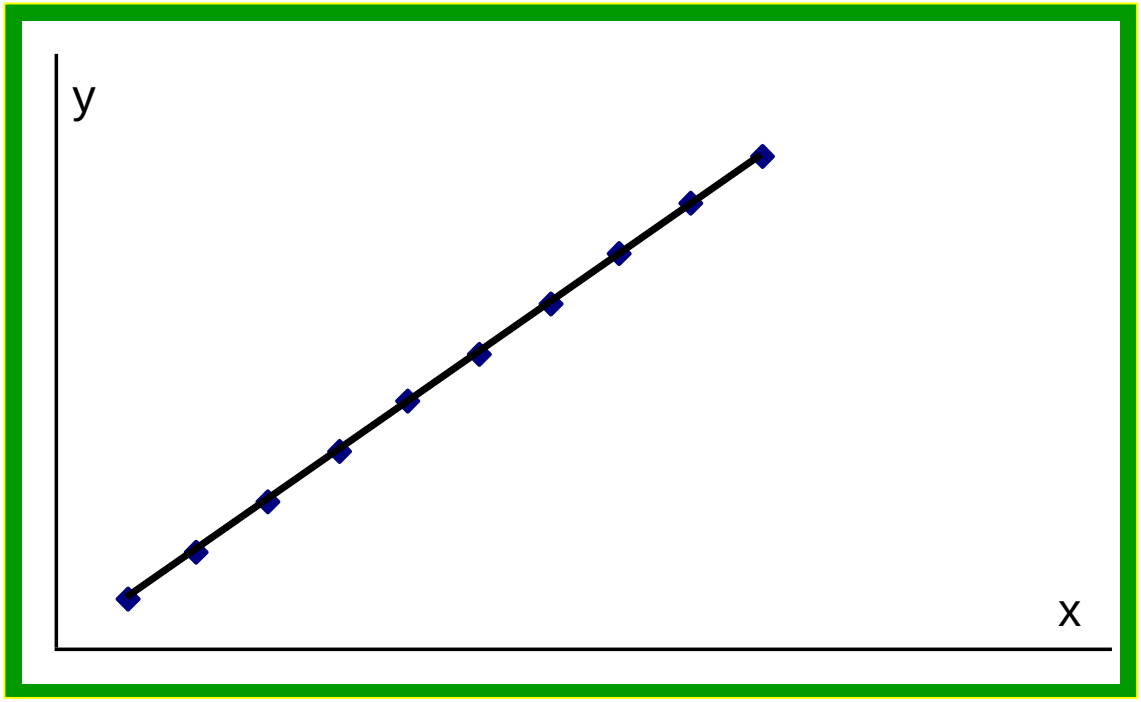
La **misura** della forza della **associazione** tra le due variabili è data dal **coefficiente di correlazione di Pearson**:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

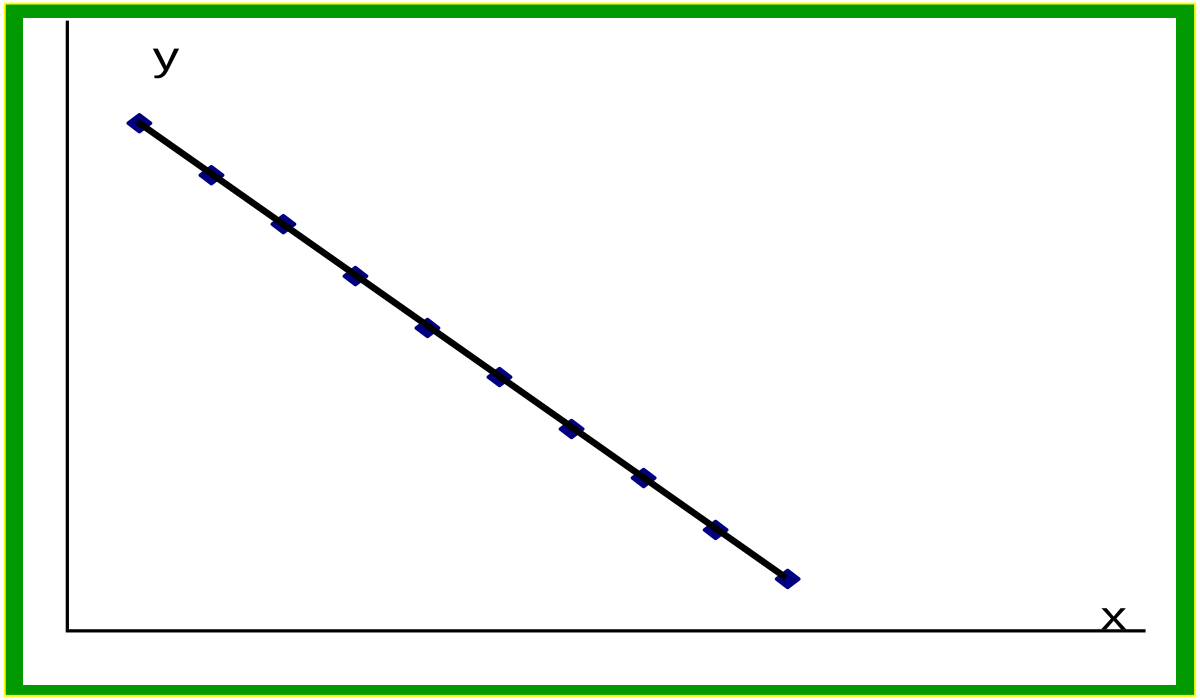
Con **$-1 < r < +1$**

La correlazione studia **<u>l'associazione lineare</u>** esistente tra due variabili.

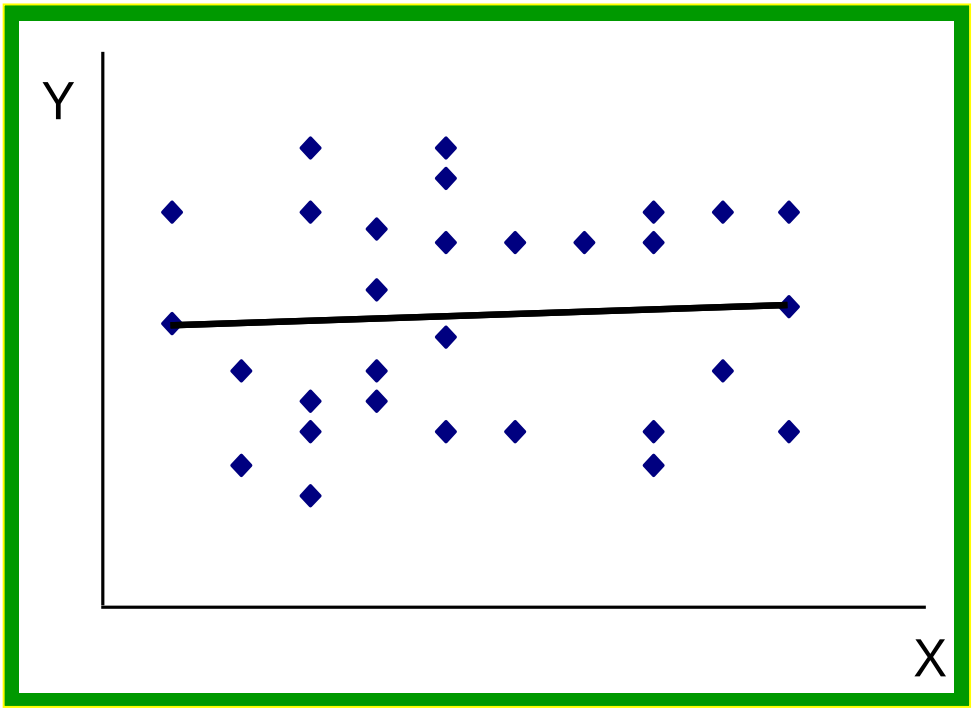
r = +1 : massima correlazione con **proporzionalità diretta** tra le due variabili, al crescere della X cresce anche la Y



r = -1 : massima correlazione con **proporzionalità inversa** tra le due variabili, al crescere della X decresce la Y (e viceversa).



r = 0 : vuol dire che **non esiste correlazione** tra le due variabili.

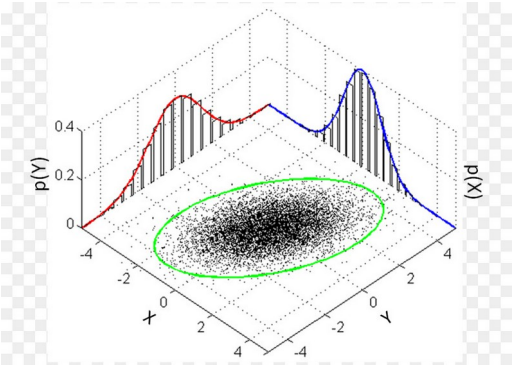


IL TEST DI VERIFICA DI IPOTESI

Il valore di r è comunque una stima campionaria del coefficiente di correlazione r della popolazione.
E' possibile eseguire un test di verifica relativa alla significatività del nostro r campionario.
Tale test verifica anche l'indipendenza delle due variabili se si assume che queste seguano una distribuzione normale bivariata.

ASSUNZIONI

+ La distribuzione di X e Y congiunte è una distribuzione normale bivariata.



IPOTESI

$$\begin{cases} H_0: \rho = 0 \\ H_1: \rho \neq 0 \end{cases}$$

STATISTICA TEST

$$T = r\sqrt{\frac{n-2}{1-r^2}}$$

DISTRIBUZIONE DELLA STATISTICA TEST

La statistica test ha una distribuzione t-Student con n-2 gradi di libertà.

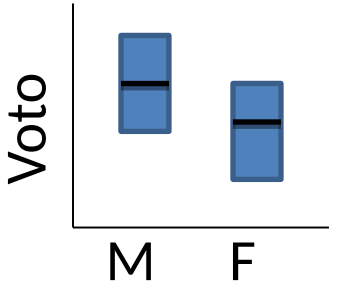
REGOLA DI DECISIONE

Conoscendo la distribuzione della statistica test, i suoi gradi di libertà e il livello di significatività ($\alpha = 0{,}05$), individuerò il valore tabulato con cui confrontare il valore calcolato.
Se $|t_{calc}| > |t_{tab}|$ allora rifiuto H_0.

Principali analisi statistiche

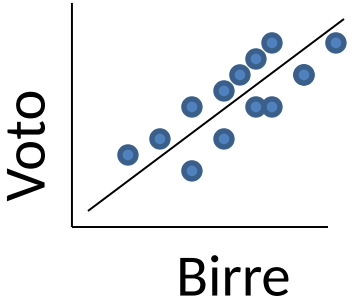
1. Confronto fra medie (2 o piú campioni)

Variabile quantitativa in funzione di una categorica (es. voto più alto M vs. F)



2. Correlazione e regressione

Relazione fra due variabile quantitative
(es. il voto medio dipende dal consumo di birre?)



3. Analisi di tabelle di contigenza

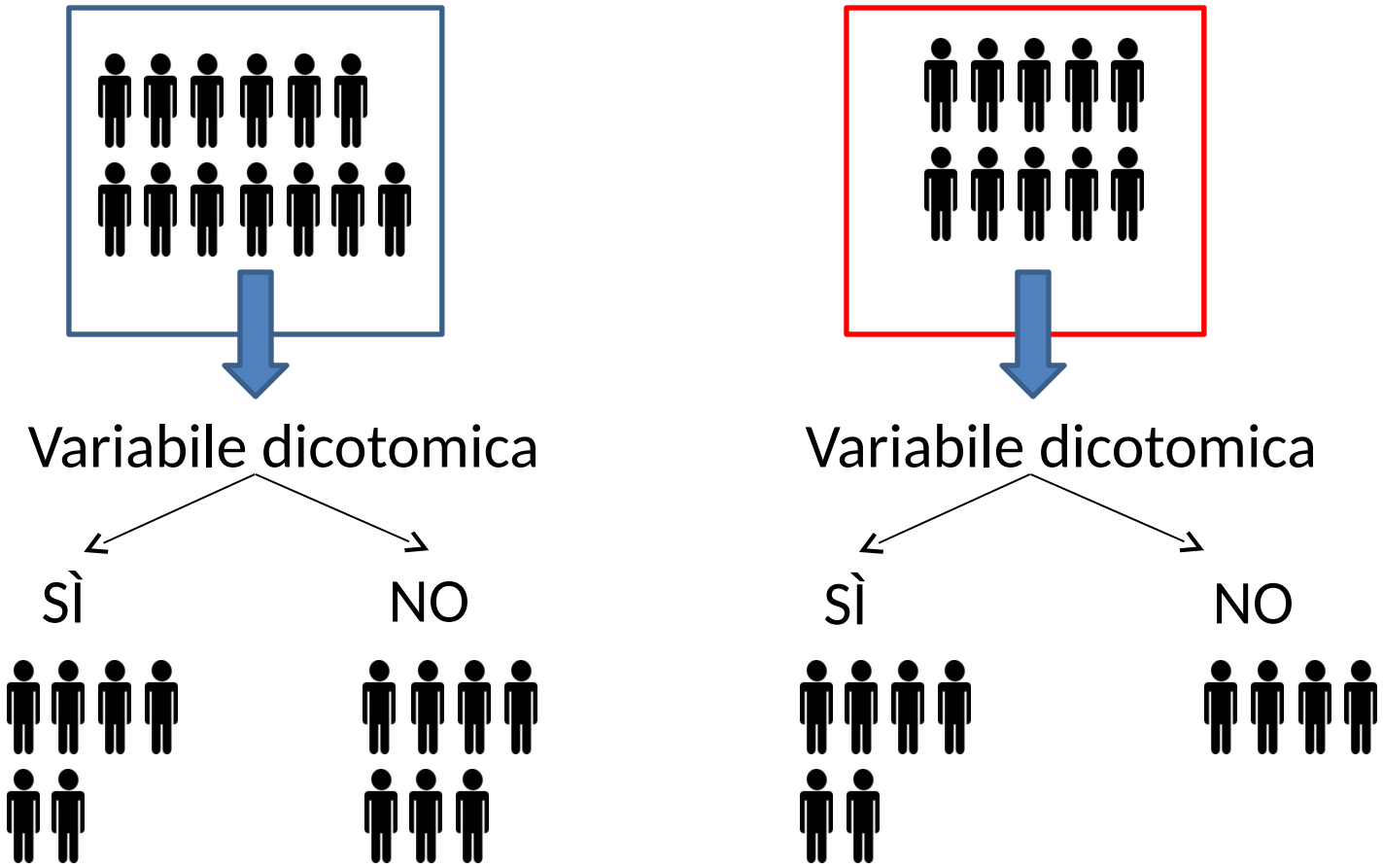
Conteggi con due o più variabili categoriche
(es. essere astemi dipende dal genere?)

	Birre	
	SÌ	NO
M		
F		

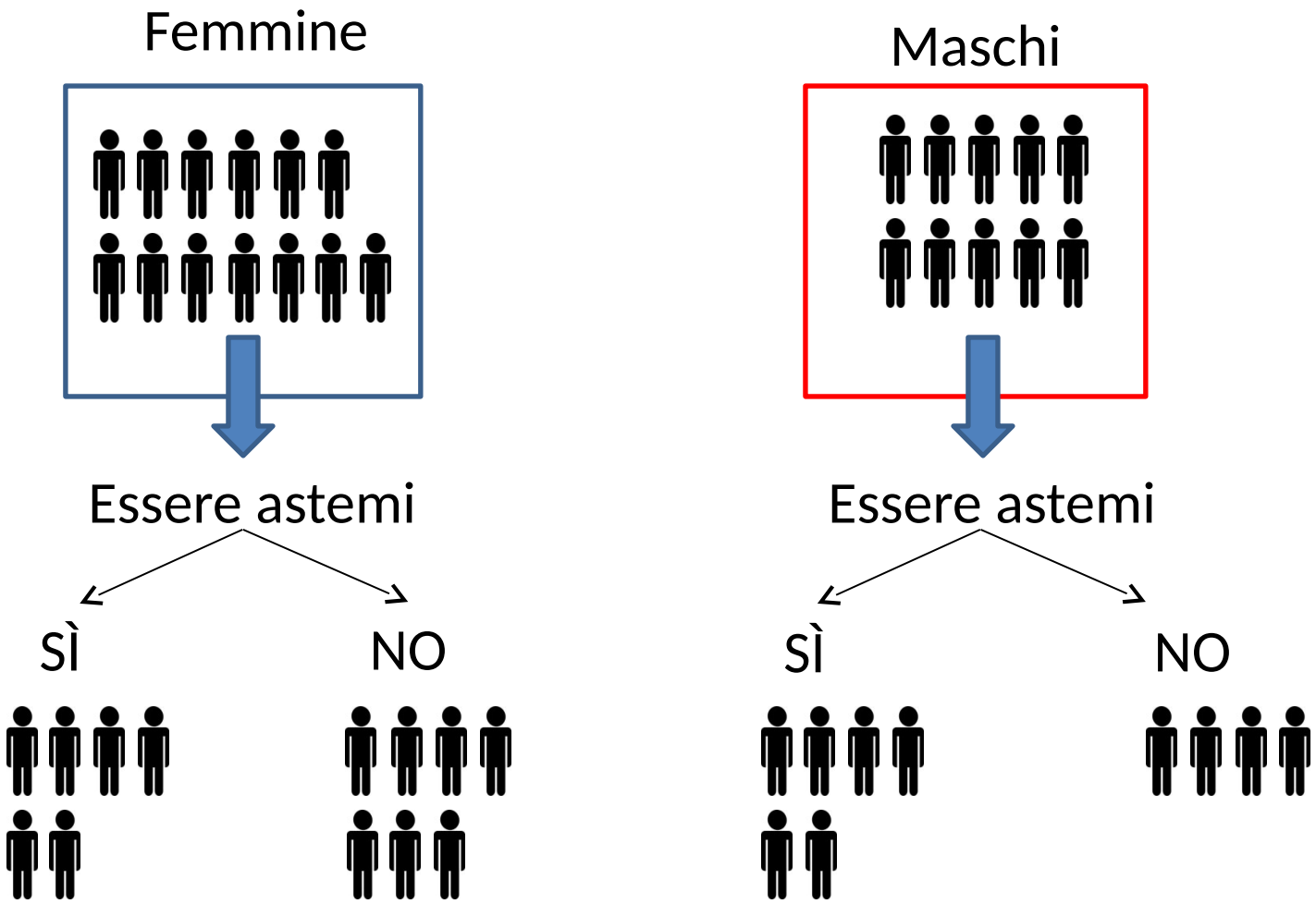
Test fra due proporzioni

Differenza fra due proporzioni

Confronto fra due gruppi indipendenti



Differenza fra due proporzioni



È diversa la proporzione di astemi fra M e F?

Il test del χ^2: Organizzare i dati

Organizzare i dati: la tabella di contingenza 2 x 2

	Gruppo 1	Gruppo 2	Totale riga
Successo	X_1	X_2	**X**
Insucesso	n_1-X_1	n_2-X_2	**n-X**
Totale colonna	**n_1**	**n_2**	**n**

Totale successi

Totale insuccessi

Totale individui Gruppo 1

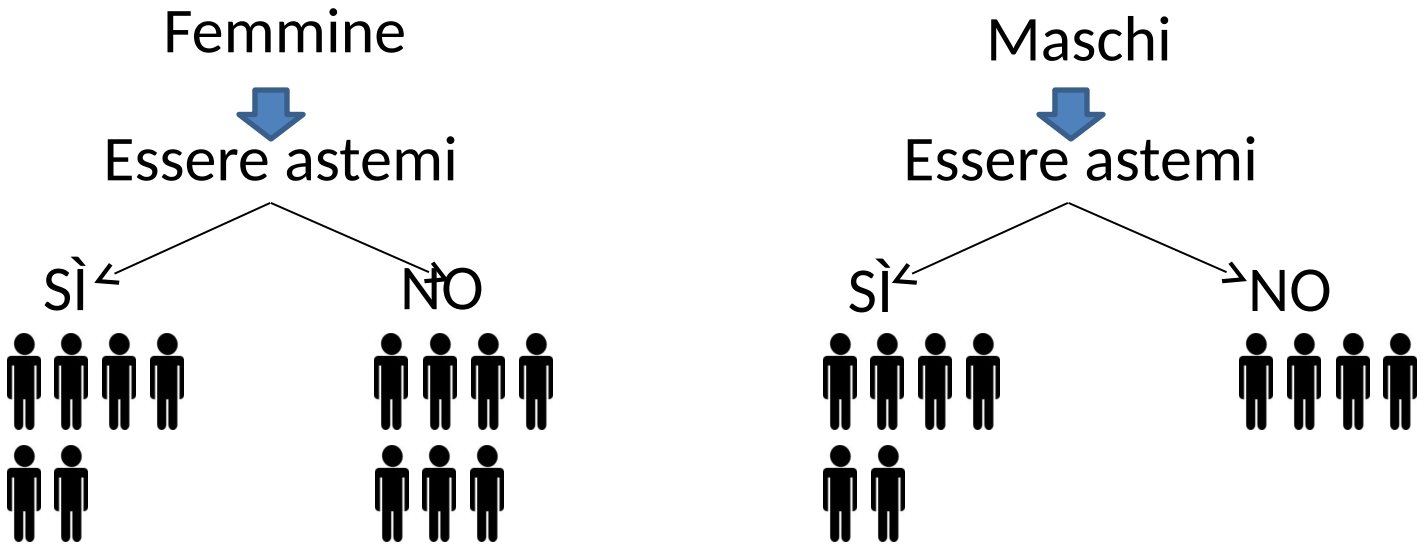
Totale individui Gruppo 2

Totale individui

Frequenze relative

	Gruppo 1	Gruppo 2	Totale riga
Successo	X_1/n_1	X_2/n_2	**(X)/n**
Insucesso	$(n_1-X_1)/n_1$	$(n_2-X_2)/n_2$	**(n-X)/n**
Totale colonna	**100%**	**100%**	**100%**

Il test del χ^2: Organizzare i dati



Astemi?	Femmine	Maschi	Totale riga
SÌ	6	6	**12**
NO	7	4	**11**
Totale colonna	**13**	**10**	**n=23**

Astemi?	Femmine	Maschi	Totale riga
SÌ	46.1%	60%	**52%**
NO	53.9%	40%	**48%**
Totale colonna	**100%**	**100%**	**100%**

Il test del χ^2: Calcolare le frequenze

Il test

Ipotesi:
Ho: le due proporzioni sono uguali
Ha: le due proporzioni sono diverse

$$\chi^2 = \sum_{\substack{Tutte\\ le\\ celle}} \frac{(|f_o - f_a| - 0.5)^2}{f_a}$$

Correzione di Yates

Astemi?	Femmine	Maschi	Tot riga
SÌ	6	6	**12**
NO	7	4	**11**
Tot colonna	**13**	**10**	**n=23**

Frequenze OSSERVATE (f_o)

Frequenze ATTESE (f_a)?

Il test del χ^2: Calcolare le frequenze

Calcolo delle frequenze attese: Frequenze che si avrebbero se Ho fosse vera

$$f_{attese} = \frac{\text{tot colonna} * \text{tot riga}}{n}$$

F astemi: 13*12/23=6.78

Astemi?	F	M	Tot riga
SÌ	**6**	6	**12**
NO	7	4	11
Tot colonna	**13**	10	**n=23**

Frequenze OSSERVATE (f_o)

	F	M
Astemi	6.78	5.21
Non astemi	6.21	4.78

Frequenze ATTESE (f_a)

Il test del χ^2: Eseguire il test

Frequenze OSSERVATE (f_o)

	F	M
Astemi	6	6
Non astemi	7	4

Frequenze ATTESE (f_a)

	F	M
Astemi	6.78	5.21
Non astemi	6.21	4.78

$$\chi^2 = \sum_{\substack{Tutte\\ le\\ celle}} \frac{(|f_o - f_a| - 0.5)^2}{f_a}$$

g.d.l.=(n righe-1)*(n colonne-1)

Se χ^2 calcolato > χ^2 critico rifiuto H0

χ^2 calcolato =0.434

Cosa concludiamo?

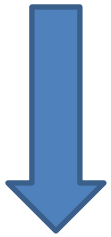
Tavola distribuzione CHI-QUADRATO

Gradi di libertà	*Livello di Probabilita' a*									
	1.00	*0.99*	*0.95*	*0.90*	*0.25*	*0.10*	*0.05*	*0.025*	*0.01*	*0.005*
1				0.02	1.32	2.71	3.84	5.02	6.64	7.88
2	0.01	0.02	0.10	0.21	2.77	4.61	5.99	7.38	9.21	10.60
3	0.07	0.12	0.35	0.58	4.11	6.25	7.82	9.35	11.35	12.84
4	0.21	0.30	0.71	1.06	5.39	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	1.15	1.61	6.63	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.64	2.20	7.84	10.65	12.59	14.45	16.81	18.55
7	0.99	1.24	2.17	2.83	9.04	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.73	3.49	10.22	13.36	15.51	17.54	20.09	21.96
9	1.74	2.09	3.33	4.17	11.39	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.94	4.87	12.55	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	4.58	5.58	13.70	17.28	19.68	21.92	24.73	26.76
12	3.07	3.57	5.23	6.30	14.85	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.89	7.04	15.98	19.81	22.36	24.74	27.69	29.82
14	4.08	4.66	6.57	7.79	17.12	21.06	23.69	26.12	29.14	31.32
15	4.60	5.23	7.26	8.55	18.25	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	7.96	9.31	19.37	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	8.67	10.09	20.49	24.77	27.59	30.19	33.41	35.72
18	6.27	7.02	9.39	10.87	21.61	25.99	28.87	31.53	34.81	37.16

Estensione del test del χ^2 a più di due gruppi

Il test del χ^2: estensione a c gruppi

Il test del χ^2 può essere utilizzato per testare anche se una proporzione è diversa fra più di due gruppi



Ipotesi:

H0: $\text{Proporzione}_1 = \text{Proporzione}_2 = \ldots = \text{Proporzione}_c$

Ha: Non tutte le proporzioni sono uguali (almeno due diverse)

$$\chi^2 = \sum_{\substack{Tutte \\ le \\ celle}} \frac{(f_o - f_a)^2}{f_a}$$

No correzione di Yates

Il test del χ^2: estensione a c gruppi

Sopravvissuto?	Dose 1	Dose 2	Dose 3	Tot riga
SÌ	15	12	5	**32**
NO	5	8	5	**18**
Tot colonna	**20**	**20**	**10**	**n=50**

Stesso procedimento per calcolare le frequenze attese!

Sopravvissuto?	Dose 1	Dose 2	Dose 3
SÌ	32*20/50	32*20/50	32*10/50
NO	18*20/50	18*20/50	18*10/50

Sopravvissuto?	Dose 1	Dose 2	Dose 3
SÌ	12.8	…	
NO	7.2		

Il test del χ^2: estensione a c gruppi

$$\chi^2_{calcolato} = \sum_{\substack{Tutte \\ le \\ celle}} \frac{(f_o - f_a)^2}{f_a}$$

g.d.l.=(n righe-1)*(n colonne-1)

Se χ^2 calcolato > χ^2 critico rifiuto H0

Tavola distribuzione CHI-QUADRATO

Gradi di libertà	*Livello di Probabilita' a*									
	1.00	*0.99*	*0.95*	*0.90*	*0.25*	*0.10*	*0.05*	*0.025*	*0.01*	*0.005*
1				0.02	1.32	2.71	3.84	5.02	6.64	7.88
2	0.01	0.02	0.10	0.21	2.77	4.61	5.99	7.38	9.21	10.60
3	0.07	0.12	0.35	0.58	4.11	6.25	7.82	9.35	11.35	12.84
4	0.21	0.30	0.71	1.06	5.39	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	1.15	1.61	6.63	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.64	2.20	7.84	10.65	12.59	14.45	16.81	18.55
7	0.99	1.24	2.17	2.83	9.04	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.73	3.49	10.22	13.36	15.51	17.54	20.09	21.96
9	1.74	2.09	3.33	4.17	11.39	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.94	4.87	12.55	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	4.58	5.58	13.70	17.28	19.68	21.92	24.73	26.76
12	3.07	3.57	5.23	6.30	14.85	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.89	7.04	15.98	19.81	22.36	24.74	27.69	29.82
14	4.08	4.66	6.57	7.79	17.12	21.06	23.69	26.12	29.14	31.32
15	4.60	5.23	7.26	8.55	18.25	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	7.96	9.31	19.37	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	8.67	10.09	20.49	24.77	27.59	30.19	33.41	35.72
18	6.27	7.02	9.39	10.87	21.61	25.99	28.87	31.53	34.81	37.16

Attenzione al calcolo dei g.d.l.!

Il test di indipedenza fra due variabili categoriche

Il test del χ^2: Il test di indipedenza

1. Variabili risposta (SÌ/NO)~ Variabile categorica 2 gruppi

2. Variabili risposta (SÌ/NO)~ Variabile con c gruppi

3. 2 Variabili categoriche con c gruppi

Il test del χ^2: Il test di indipendenza





	Quercia	Carpino	Salice	Tot riga
Lucanus				
Osmoderma				
Cerambix				
Tot colonna				

4 carpini con Cerambix

Il test del χ^2: estensione a c gruppi

Stesso procedimento per calcolare le frequenze attese!

$$\chi^2_{calcolato} = \sum_{\substack{Tutte \\ le \\ celle}} \frac{(f_o - f_a)^2}{f_a}$$

g.d.l.=(n righe-1)*(n colonne-1)

Se χ^2 calcolato > χ^2 critico rifiuto H0

Tavola distribuzione CHI-QUADRATO

Gradi di libertà	*Livello di Probabilita' a*									
	1.00	*0.99*	*0.95*	*0.90*	*0.25*	*0.10*	*0.05*	*0.025*	*0.01*	*0.005*
1				0.02	1.32	2.71	3.84	5.02	6.64	7.88
2	0.01	0.02	0.10	0.21	2.77	4.61	5.99	7.38	9.21	10.60
3	0.07	0.12	0.35	0.58	4.11	6.25	7.82	9.35	11.35	12.84
4	0.21	0.30	0.71	1.06	5.39	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	1.15	1.61	6.63	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.64	2.20	7.84	10.65	12.59	14.45	16.81	18.55
7	0.99	1.24	2.17	2.83	9.04	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.73	3.49	10.22	13.36	15.51	17.54	20.09	21.96
9	1.74	2.09	3.33	4.17	11.39	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.94	4.87	12.55	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	4.58	5.58	13.70	17.28	19.68	21.92	24.73	26.76
12	3.07	3.57	5.23	6.30	14.85	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.89	7.04	15.98	19.81	22.36	24.74	27.69	29.82
14	4.08	4.66	6.57	7.79	17.12	21.06	23.69	26.12	29.14	31.32
15	4.60	5.23	7.26	8.55	18.25	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	7.96	9.31	19.37	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	8.67	10.09	20.49	24.77	27.59	30.19	33.41	35.72
18	6.27	7.02	9.39	10.87	21.61	25.99	28.87	31.53	34.81	37.16

Attenzione al calcolo dei g.d.l.!

Il test del χ^2: Il test di indipendenza

Le ipotesi del test sono diverse!

H0: le due variabili categoriche sono indipendenti
(non vi è relazione)

Ha: le due variabili categoriche non sono indipendenti (una dipende dall'altra)

Nell'esempio precedente?

H0:?

Ha:?

Il test del χ^2





	Quercia	Carpino	Salice	Tot riga
Lucanus				
Osmoderma				
Cerambix				
Tot colonna				

Posso testare se il Lucanus ha una preferenza?

Il test del χ^2

	Quercia	Carpino	Salice
Lucanus			

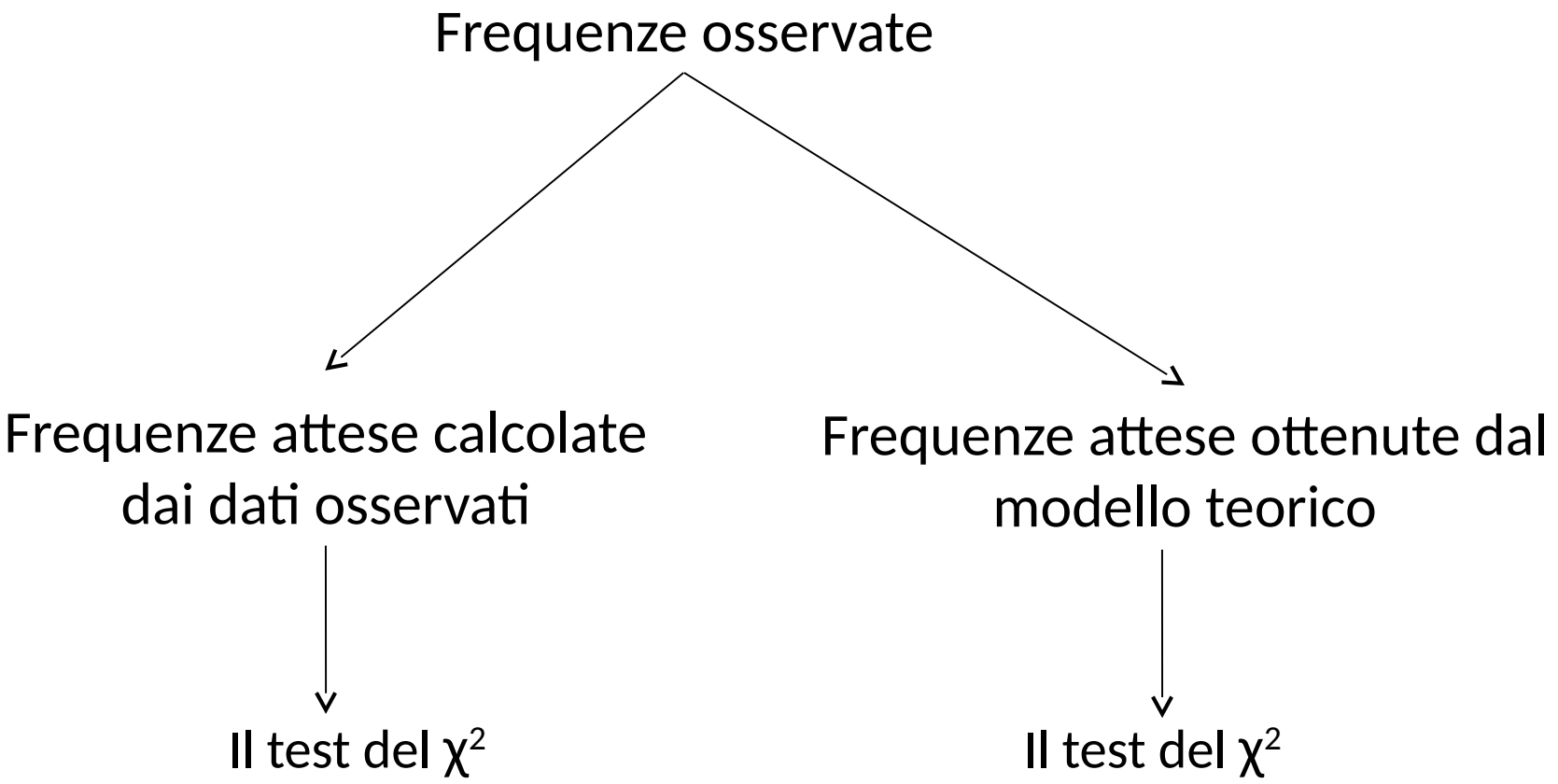
Idee?

Che informazione ci manca?

Il test del χ^2 per testare un modello teorico

Il test del χ^2 per testare un modello teorico

Le frequenze attese possono derivare da un modello!



Il test del χ^2 per testare un modello teorico

Ad es. ho un modello che indica che il 10% delle femmine e il 5% dei maschi di capriolo sviluppa una certa patologia entro i 3 anni

Patologia?	M	F	Tot
SÌ	4	14	18
NO	102	105	207

Le frequenze attese?

Patologia?	M	F
SÌ		
NO		

Il test del χ^2 per testare un modello teorico

Attenzione alle ipotesi del test!

Ho: il modello spiega i dati

Ha: il modello NON mi spiega i dati (i dati deviano dalle predizione del modello teorico)

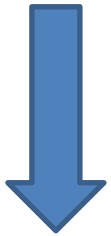
Vogliamo χ^2 calcolato > χ^2 critico?

Il test del χ^2: Limitazioni

1. Il test non funziona bene se le frequenza attese sono basse

Diversi suggerimenti…

Nessuna frequenza attesa dovrebbe essere<5



Test esatto di Fisher può essere utilizzato in questi casi

Il test del χ^2: Limitazioni

2. Il test lavora solo con frequenze (conteggi reali) e non con proporzioni (%)

40% femmine astemie (devo conoscere n!)
50% maschi astemi (devo conoscere n!)

3. Il test assume indipendenza delle frequenze (attenzione ai doppi conteggi!)

Dati quantitativi (medie)

2 gruppi → t test → Appaiato / Non appaiato

>2 gruppi → ANOVA

Dati con proporzioni

2 gruppi

>2 gruppi

→ Il test del χ^2