Statistica Inferenziale

La verifica di ipotesi

Davide Barbieri

Inferenza statistica

- Inferenza: procedimento di induzione, dal particolare al generale.
- Stima di un parametro della popolazione partendo da un campione
- L'inferenza ha sempre un valore statistico, probabilistico e dipende dalla dimensione del campione.

Verifica di ipotesi

- Test effettuato per trarre conclusioni di carattere generale (su una popolazione) basandosi sui dati di uno o più campioni.
- Verificare, su base statistica, le probabilità associate alle ipotesi alternative di un esperimento.
- Es. verificare la validità (efficacia) di un farmaco.

Ipotesi

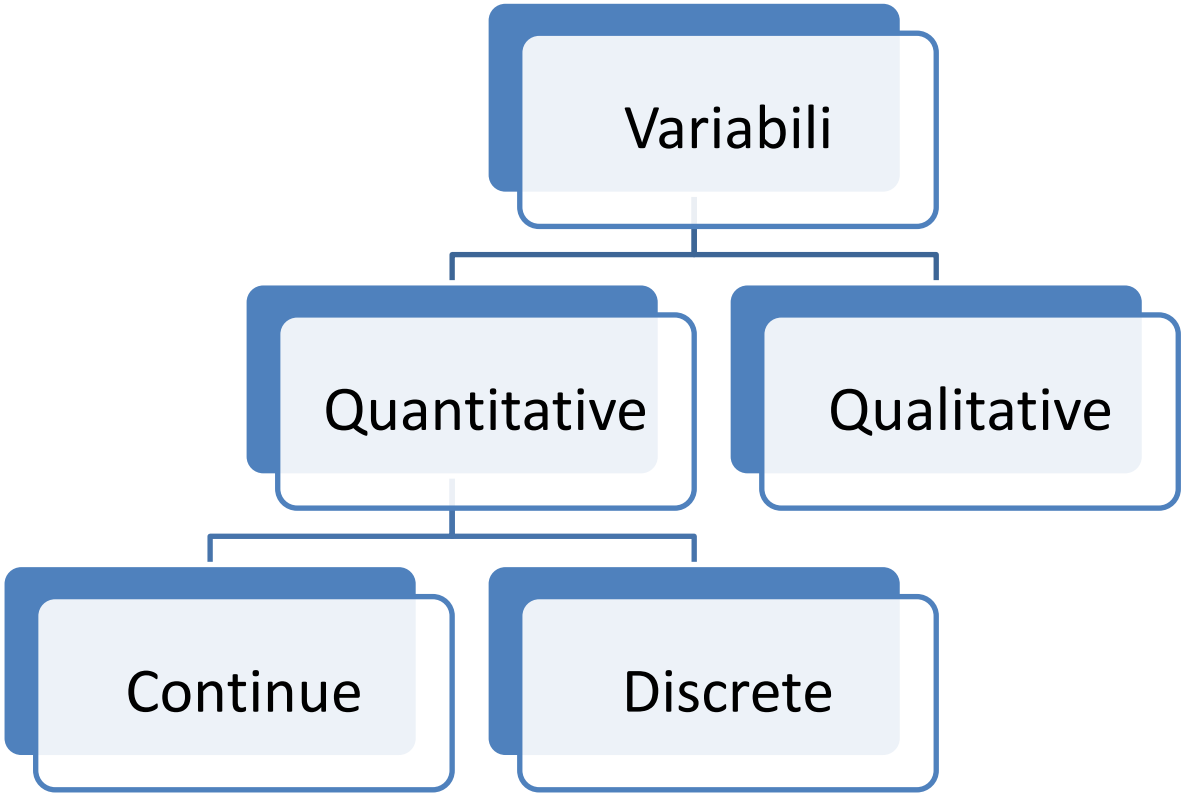
- Ipotesi nulla H_0: ipotesi su un parametro di popolazione considerata valida fino a prova contraria.
- Ipotesti alternativa H_1: valida se l'ipotesi nulla si dimostra (probabilmente) falsa.

Tipi di errore

	H_0 è vera	H_0 è falsa
Non respingo H_0	OK	Errore di tipo II β
Respingo H_0	Errore di tipo I α	OK

- $\alpha = P(H_0 \text{ respinta} \mid H_0 \text{ vera})$: significatività
- $\beta = P(H_0 \text{ non respinta} \mid H_0 \text{ falsa})$: potenza

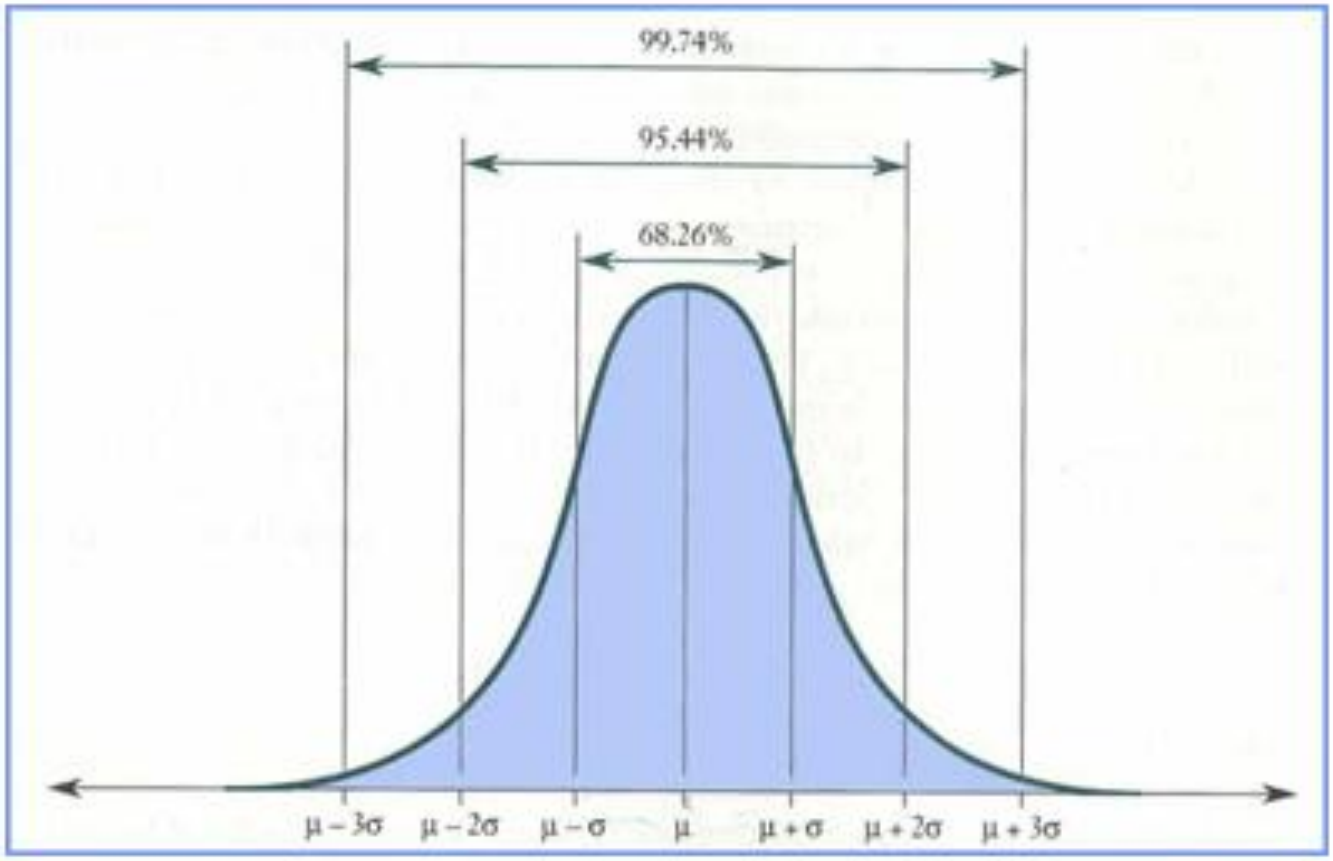
Variabili aleatorie



Verifica di ipotesi 1

Variabili quantitative continue

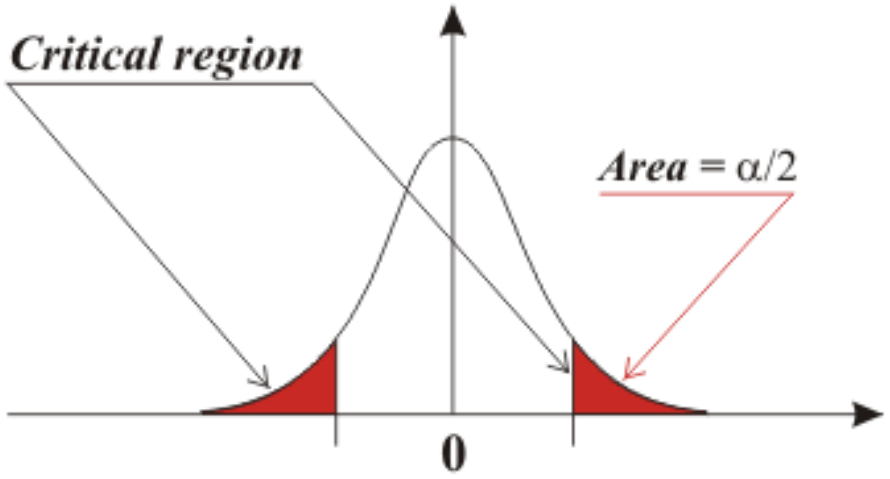
Distribuzione normale



Code del test

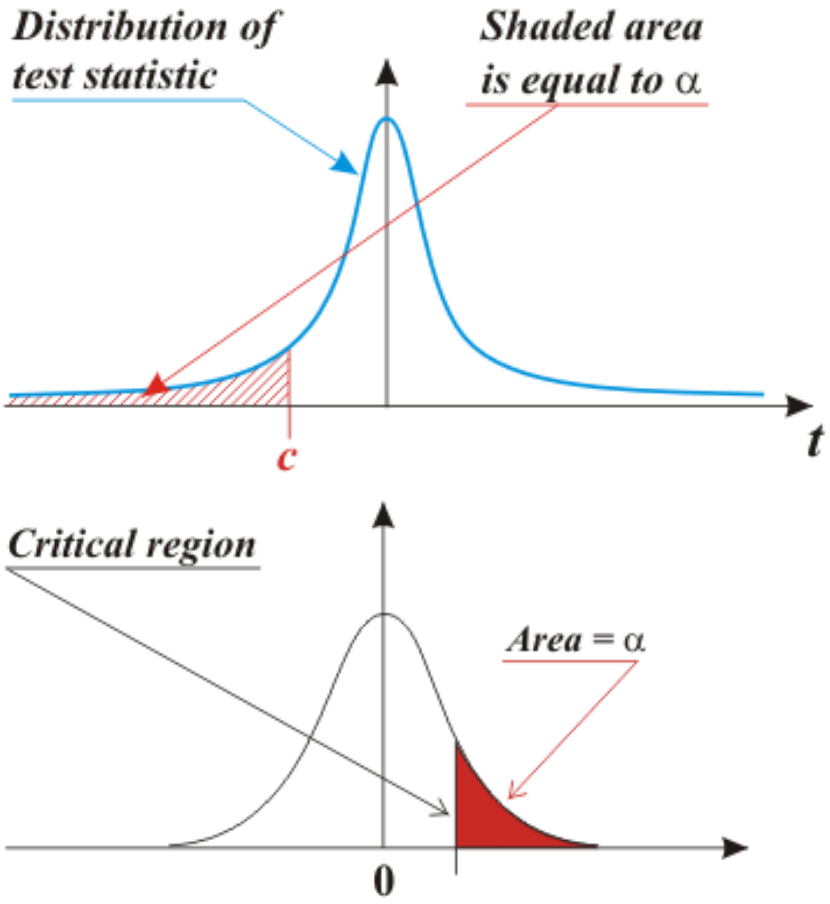
	Test a due code	Test a una coda sinistra	Test a una coda destra
Segno H_0	=	= oppure ≥	= oppure ≤
Segno H_1	≠	<	>
Area di rigetto	Entrambe le code	Coda sinistra	Coda destra

Test a due code



- p: probabilità di commettere errore di tipo I associata alla statistica del test
- Se $p < \alpha/2 \rightarrow$ respingo H_0 e convalido H_1
- Se $p > \alpha/2 \rightarrow$ convalido H_0 e respingo H_1

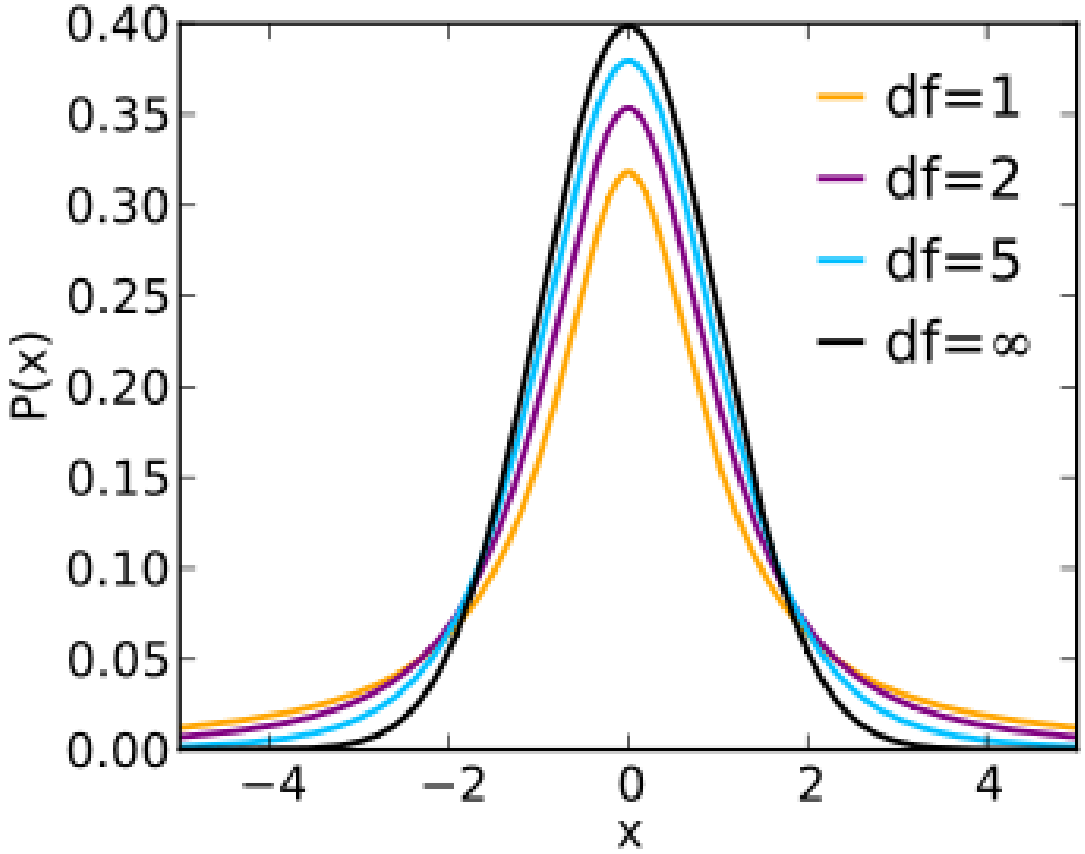
Test a una coda



t di Student

- μ e σ ignote.
- Verifica che la media di un campione corrisponda a quella di popolazione (valore atteso).
- Verifica che le medie di due campioni dipendenti non siano significativamente differenti.
- Verifica che le medie di due popolazioni siano uguali basandosi sulle medie di due campioni indipendenti.

Distribuzione di t



df=n-1 gradi di libertà

Procedura del test

1. Definire H_0 e H_1
2. Calcolare $t = (\underline{x}_1 - \underline{x}_2)/E_s$ ($E_s = s/\sqrt{n}$)
3. Identificare t_c corrispondente a df e α
4. Confrontare t e t_c
5. Accettare o rigettare H_0

Valori di t

df	Area nella Coda di Destra sotto la Curva di Distribuzione *t*					
	0.10	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**
1	3.078	6.314	12.706	31.821	63.656	318.289
2	1.886	2.920	4.303	6.965	9.925	22.328
3	1.638	2.353	3.182	4.541	5.841	10.214
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.894
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552

Es. 1: campioni dipendenti

Verificare l'efficacia di un programma di allenamento per sollevatori olimpici.

- Un solo campione di 5 atleti
- $t = (\underline{x}_1 - \underline{x}_2)/E_s$
- $df = n-1$ gradi di libertà
- $E_s = s/\sqrt{n}$ errore standard
- $s = \sqrt{((\sum(D_i-\underline{D})^2/(n-1))}$ dev. std campionaria
- $D_i = x_1-x_2$ scostamenti i=1..n
- $\underline{D} = \sum D_i/n$ scostamento medio

H_0: la forza resta uguale $\underline{x}_2=\underline{x}_1$			
H_1: la forza aumenta $\underline{x}_2>\underline{x}_1$			
n	**x_1**	**x_2**	**Di = x_2-x_1**
1	100	110	10
2	80	87	7
3	120	120	0
4	70	70	0
5	90	100	10
media	92	97,4	5,4
α=	0,05		
df=n-1	4		
$s = \sqrt{((\sum(D_i-\underline{D})^2/(n-1))}$	5,08		
Errore std $E_s = s/(\sqrt{n})$	2,27		
t_c=	2,13		
$t = (\underline{x}_2-\underline{x}_1)/E_s$	**2,38**		
rigetto H0			

Es. 2: campioni indipendenti

- H_0: Peso medio neonati maschi e femmine uguale $\mu_1 = \mu_2$
- H_1: peso medio maschi superiore $\mu_1 > \mu_2$
- $t = (\underline{x}_1 - \underline{x}_2)/E_s$
- $df = (n_1\text{-}1) + (n_2\text{-}1) = n_1 + n_2 - 2$
- $E_s = \sqrt{((Ds^2 / n_1) + (Ds^2/n_2))}$ errore standard
- $Ds^2=(\Sigma x^2_1 + \Sigma x^2_2)/((n_1\text{-}1)+(n_2\text{-}1))$ varianza media

H_0: peso medio uguale nei due sessi $\underline{x}_1=\underline{x}_2$				
H_1: peso medio maggiore nei maschi $\underline{x}_1>\underline{x}_2$				
n	**$\underline{x}_1$**	**$\underline{x}_2$**	**$\sum x_1^2$ (devianza)**	**$\sum x_2^2$**
100	3400	3150	33223680	24265113
Varianza media Ds^2 = $(\sum x_1^2+\sum x_2^2)/((n_1-1)+(n_2-1))$=				290347
Errore standard $E_s = \sqrt{((Ds^2/n_1)+(Ds^2/n_2))}$ = $\sqrt{(2Ds^2/n)}$ =				76,20
α =	0,05			
df=2n-2=	198			
t_c =	1,66			
t=	**3,28**			
Rigetto H0				

Verifica di ipotesi 2

Variabili categoriali

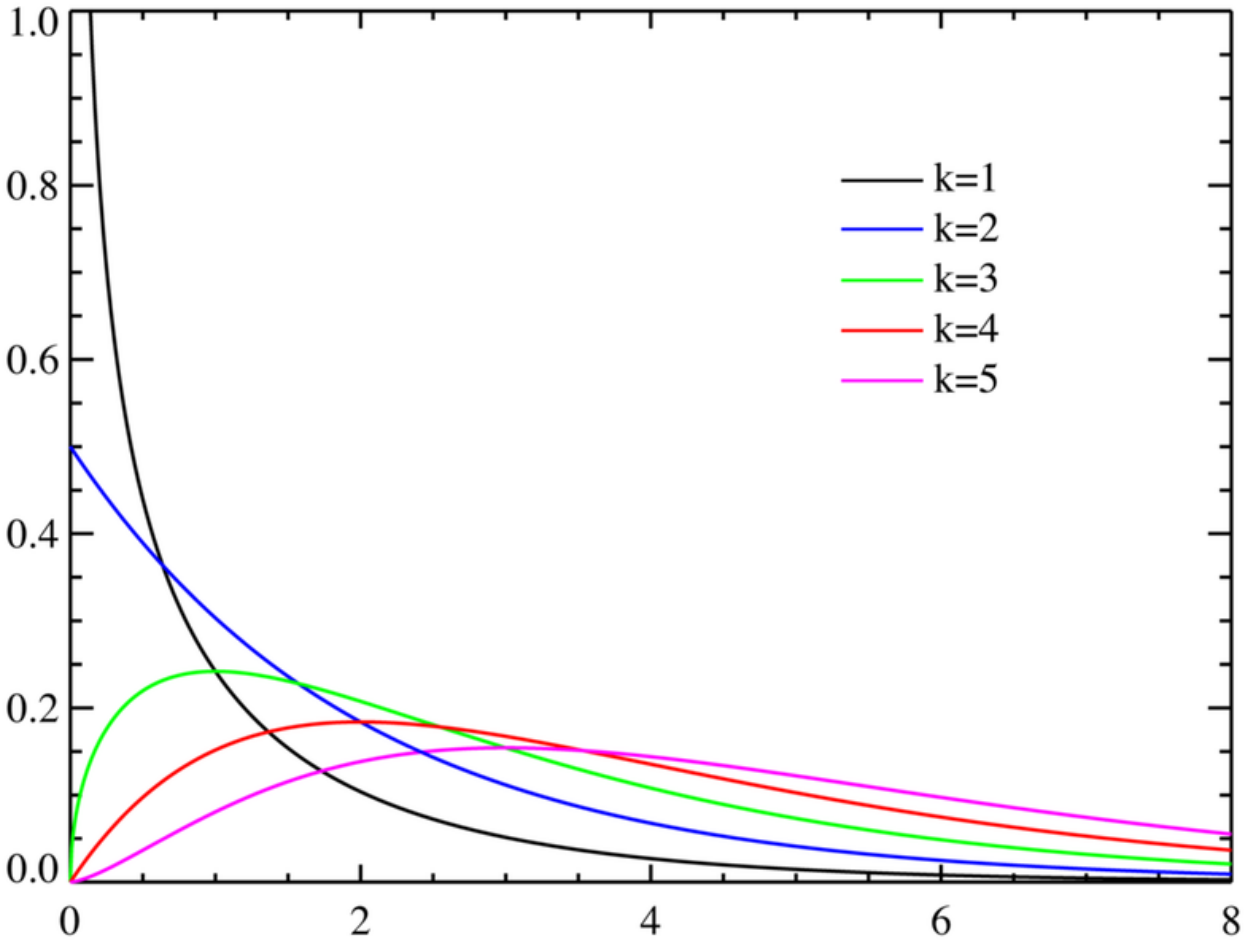
Statistica non parametrica

- Non è possibile fare ipotesi di normalità sulla distribuzione di provenienza del campione (per es. nel caso di campioni piccoli).
- Non si conoscono i parametri di popolazione (come media e varianza).
- Se il campione tende alla normalità, aumentando gli elementi, allora si possono usare test parametrici.

Chi quadro X^2

- Variabili categoriali
- Asimmetrica verso destra.
- Crescendo i gradi di libertà tende alla normalità
- $X^2 = \sum(f_o - f_e)^2/f_e$

X^2 densità di probabilità



Condizioni

- Dati indipendenti: nessun soggetto può apparire in più di una cella della tabella.
- Non più del 20 % delle frequenze attese nella tabella può essere < 5 (altrimenti si usa il test esatto di Fisher).
- Nessuna cella deve avere una frequenza attesa < 1 (altrimenti si usa il test esatto di Fisher).

Es: genere e pratica sportiva

	Non pratica sport	Pratica sport	Totale
M	14	43	57
F	11	32	43
totale	25	75	100

Esperimenti multinomiali

- Una sola variabile con k≥2 categorie.
- df = k-1
- n: dimensione del campione
- p=1/k probabilità che un elemento appartenga ad una delle categorie se H_0 è vera
- Frequenza attesa f_e = np

Es: pazienti psichiatrici

- H_0: frequenza schizofrenici nati in inverno uguale a quella dei nati nelle altre stagioni[1] P(In) = P(Pr) = P(Es) = P(Au)
- H_1: P(In) > P(Pr) e P(In) > P(Es) e P(In) > P(Au), cioè P(In) = max(P)

[1] In: inverno, Pr: primavera, Es: estate, Au: autunno

				f_o	f_e	$(f_o-f_e)^2/f_e$
n=	100		**Pr**	21	25	0,64
k=	4		**Es**	22	25	0,36
df=	3		**Au**	25	25	0
f_e= n*1/k	25		**In**	32	25	1,96
α=	0,05				**X^2 =**	**2,96**
X^2_c =	7,81					
Non rigetto H0						
df	**α=0,05**	**α=0,01**				
1	3,84	6,63				
2	5,99	9,21				
3	7,81	11,34				
4	9,48	13,27				
5	11,07	15,08				

Test di indipendenza

- Verificare l'indipendenza di due o più variabili categoriali.
- Tabelle di contingenza
- df = (r-1)(c-1) gradi di libertà
- H_0: $P(x_1x_2) = P(x_1)P(x_2)$ variabili indipendenti
- H_1: $P(x_1x_2) \neq P(x_1)P(x_2)$
- Valore atteso: E = (tot_riga*tot_colonna)/n

Es: programmi di allenamento

- Due campioni di atleti, che seguono due diversi programmi di allenamento, A e B.
- H_0: P(AV) = P(BV) e P(AS) = P(BS) il numero di vittorie V e sconfitte S è indipendente dal programma seguito
- H_1: P(AV) > P(BV) e P(AS) < P(BS) il numero di vittorie e sconfitte dipende dal programma di allenamento.

Osservati	A	B	Tot			
V_o	38	29	67			
S_o	7	17	24			
Tot	45	46	91	=n		
df=	1		(numero righe - 1) * (numero colonne -1)			
α=	0,05					
Attesi	A	B				
V_e	33,13	33,87				
S_e	11,87	11,87				
X^2=	5,63					
X^2_c=	3,84					
Rigetto H0						