Analisi statistica di dati biomedici
Analysis of biological signals

III Parte – Verifica delle ipotesi (b)

Agostino Accardo (accardo@units.it)

Master in Ingegneria Clinica – LM in Neuroscienze
2013-2014 e segg.

Test sulla differenza tra proporzioni o frequenze

Differenze tra proporzioni (caso binomiale)

IPOTESI: 2 gruppi di osservazioni INDIPENDENTI
Un carattere con 2 sole possibilità (valori), p.es. 'miglioramento'/'non miglioramento', con probabilità ***p*** e ***q=1-p***.
(avremo $p_1, q_1 \; e \; n_1$ e $p_2, q_2 \; e \; n_2$ nei due gruppi)

Per valori di n_1, n_2 grandi (>80-100), una buona stima di p_1 e p_2 è data dalle frequenze relative f_1 *e* f_2 , inoltre la distribuzione **binomiale** si può approssimare con una Normale.
Poiché i gruppi sono indipendenti, allora la differenza tra i gruppi avrà media pari alla differenza delle medie e Varianza pari alla somma delle varianze, quindi

H_0: i 2 campioni provengono dalla stessa popolazione
devono quindi avere lo stesso valore di ***p***, stimabile con

$$\hat{\boldsymbol{p}} = \frac{\boldsymbol{n_1 f_1 + n_2 f_2}}{\boldsymbol{n_1 + n_2}}$$

e $\hat{\boldsymbol{q}}$ = 1 - $\hat{\boldsymbol{p}}$

Utilizzeremo:

$$\boldsymbol{Z = \frac{f_1 - f_2}{\sqrt{p \cdot q \cdot (1/n_1 + 1/n_2)}}}$$

Che si potrà confrontare con Z_α *o* $Z_{\alpha/2}$.

Per *n* piccoli al posto di $(\boldsymbol{f_1 - f_2})$ si userà la correzione (continuity correction): $|\boldsymbol{f_1 - f_2}| - \frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$

altrimenti il test sarebbe troppo ottimista.

Esempio:

25 pazienti:

12=n1, ricevettero un trattamento e 9 dissero di aver ricevuto benefici

13=n2, ricevettero un placebo e 4 dissero di aver ricevuto benefici

Frequenze osservate:

f1= 9/12=0.75, f2= 4/13=0.3077 => f1 - f2=0.4423

Stimiamo $\widehat{\boldsymbol{p}} = \frac{\boldsymbol{n_1 f_1 + n_2 f_2}}{\boldsymbol{n_1 + n_2}}$ = (9+4)/(12+13) = 0.52 e lo standard error

$\sqrt{\boldsymbol{p} \cdot \boldsymbol{q} \cdot (^1/_{n_1} + {}^1/_{n_2})}$ = 0.20

Z= 0.4423/0.20 = 2.21 corrispondente ad un p-value di 0.027 che è < 5%
($Z_{0.05}$ = 1.64)

QUINDI esiste evidenza di una differenza significativa tra i 2 trattamenti

Differenze tra proporzioni (caso analogo al binomiale ma su osservazioni NON indipendenti - paired)

p.es. due osservazioni sullo **stesso** individuo (p.es effetto di due farmaci)

In questo caso l'errore standard della differenza non è più basato sulla sola varianza di ciascuna proporzione, ma deve tener conto dei risultati 'correlati'. Si dividono le osservazioni (Ti) in 4 gruppi a seconda che la caratteristica sia presente o meno in ciascun membro della coppia (es*: presenza di un sintomo prima di un trattamento (T_1) e dopo (T_2)):*

T_1	T_2	n° coppie
Si	Si	a
Si	No	b
No	Si	c
No	No	d
		n

Nota: Considero '**b**' e '**c**' perché sono le sole situazioni che cambiano!

In questo caso si valuta:

$$Z = \frac{b - c}{\sqrt{b + c}}$$

e si confronta con Z_α.

Per *n* piccoli (correzione):

$$Z = \frac{|b - c| - 1}{\sqrt{b + c}}$$

Differenze tra frequenze (Tabelle di frequenza o di contingenza)

CASO GENERALE: TABELLE r x c

	CONSUMO CAFFEINA (mg/day)				
STATO CIVILE	0	1-150	151-300	>300	TOTALE
Coniugato	652	1537	598	242	3029
Separato/Divorziato/Vedovo	36	46	38	21	141
Single	218	327	106	67	718
TOTALE	906	1910	742	330	3888

H_0: le 2 variabili (stato civile/ consumo caffeina) sono scorrelate nella popolazione da cui è stato estratto il campione.

Dalla tabella delle frequenze osservate, si ricava la tabelle di quelle attese (teoriche), basandosi sul mantenimento delle distribuzioni marginali, le quali sono esenti da interdipendenza tra le variabili.
Le frequenze attese sono così ricavabili:

									Totale
	$T_1 \cdot \frac{\sum A}{T}$	$T_2 \cdot \frac{\sum A}{T}$	$T_3 \cdot \frac{\sum A}{T}$	$T_4 \cdot \frac{\sum A}{T}$	…	…	…	…	$\sum A$
	$T_1 \cdot \frac{\sum B}{T}$	…	…	…	…	…	…	…	$\sum B$
	…	$T_2 \cdot \frac{\sum C}{T}$	…	…	…	…	…	…	$\sum C$
	…	…	…	$T_4 \cdot \frac{\sum D}{T}$	…	…	…	…	$\sum D$
Totale	T_1	T_2	T_3	T_4	…	…	…	…	T

Infine si calcola:

$$\boldsymbol{\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}}$$

dove le $\boldsymbol{O_{ij}}$ sono le frequenze osservate e le $\boldsymbol{E_{ij}}$ sono le frequenze attese.
Tanto maggiore sarà questo valore, tanto più i valori osservati sono diversi da quelli attesi.
Si confronta quindi la $\boldsymbol{\chi^2}$ con $\chi^2_{\alpha,\nu}$ con $\nu = (c-1)(r-1)$ (gradi di libertà)

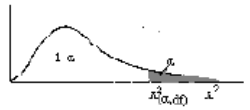
Nell'esempio $\boldsymbol{\chi^2 = 51.61}$ *e* $\boldsymbol{\chi^2_{0.01,6} = 22.46}$ $\Rightarrow$ *si rifiuta* H_0*, portando alla conclusione che ESISTE un legame significativo tra le due variabili!*

NOTE:

- è importante ricordare che se si trova un legame tra le variabili (H_0 falsa) questo NON INDICA necessariamente che esiste una RELAZIONE CAUSALE tra esse!
- In generale ci sono altri fattori che influenzano entrambe le variabili e provocano l'associazione trovata.
- L'ampiezza di $\boldsymbol{\chi^2}$ non indica la forza del legame tra le variabili, ma piuttosto la forza dell'evidenza che l'ipotesi nulla è falsa.
- Il $\boldsymbol{\chi^2}$ si può applicare solo se l'80% delle celle nella tabella delle frequenze attese è >5 e se ciascuna frequenza attesa è >1, altrimenti altri metodi (per tabelle piccole)

TABELLA χ^2

VALORI CRITICI DELLA DISTRIBUZIONE χ^2 (con gdl da 1 a 30)



Gradi di libertà	*Area della coda superiore*											
	.995	*.99*	*.975*	*.95*	*.90*	*.75*	*.25*	*.10*	*.05*	*.025*	*.01*	*.005*
1			0.001	0.004	0.016	0.102	1.323	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	0.575	2.773	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	1.213	4.108	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	1.923	5.385	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	2.675	6.626	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	3.455	7.841	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	4.255	9.037	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	5.071	10.219	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	5.899	11.389	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	6.737	12.549	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	7.584	13.701	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	8.438	14.845	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	9.299	15.984	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	10.165	17.117	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	11.037	18.245	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	11.912	19.369	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	12.792	20.489	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	13.675	21.605	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	14.562	22.718	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	15.452	23.828	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	16.344	24.935	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.042	17.240	26.039	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	18.137	27.141	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	19.037	28.241	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	19.939	29.339	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	20.843	30.435	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	21.749	31.528	36.741	40.113	43.194	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	22.657	32.620	37.916	41.337	44.461	48.278	50.993
29	13.121	14.257	16.047	17.708	19.768	23.567	33.711	39.087	42.557	45.722	49.5B8	52.336
30	13.787	14.954	16.791	18.493	20.599	24.478	34.800	40.256	43.773	46.979	50.892	53.672

CASO PARTICOLARE: TABELLE 2X2

Tabella delle Osservazioni

N= a+b+c+d

	C_1	C_2	totale
R_1	a	b	a+b
R_2	c	d	c+d
totale	a+c	b+d	a+b+c+d

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

da confrontare con $\chi^2_{\alpha,1}$

Per piccoli campioni, si usa la correzione di Yates:

$$\chi^2 = \frac{N(|ad-bc| - \frac{N}{2})^2}{(a+b)(a+c)(b+d)(c+d)}$$

Se più di un elemento della tabella dei valori attesi è <5 si userà il **test esatto di Fisher**.

Test esatto di Fisher

Questo test è adatto anche nel caso in cui si hanno a disposizione dati NON NORMALI. È basato sul calcolo diretto della probabilità che venga 'estratta' proprio quella tabella.
Si calcola:

$$P = \frac{(a+b)!\,(a+c)!\,(b+d)!\,(c+d)!}{N!\,a!\,b!\,c!\,d!}$$

per ciascuna possibile differente tabella che produca gli stessi totali:

Esempio:

	V_1	V_2	totale
Z_1	1	5	6
Z_2	8	2	10
totale	9	7	16

0 6	1 5	2 4	3 3	4 2	5 1	6 0
9 1	8 2	7 3	6 4	5 5	4 6	3 7
P_1	P_2	P_3	P_4	P_5	P_6	P_7

P1=0.00087, P2=0.0236, P3=0.157, P4=0.327, P5=0.33, P6=0.11, P7=0.01
Si sommano quindi tutti i P_i delle distribuzioni che cadono fino a quella osservata (nell'esempio sino alla SECONDA situazione $P_1 + P_2$), si raddoppia il valore (per avere 2 code) e si confronta direttamente con la probabilità di rifiutare H_0, ovvero con l'α prefissato
Siccome $(P_1+P_2)*2 = 0.049$ è $< \alpha = 0.05$, allora rifiuteremo l'ipotesi nulla e diremo che esiste una relazione tra V e Z.

Tabella 2 x 2, osservazioni dipendenti

Non si usa il χ^2 ma si confronta Z= $\frac{|b-c|-1}{\sqrt{b+c}}$ con Z_α o $Z_{\alpha/2}$

Oppure si fa il quadrato, Z^2, e si confronta con $\chi^2_{\alpha,1}$ = **test di Mc Nemar**

TEST A PIU' CAMPIONI

ANOVA (Analisi della Varianza) (parametrico)

H_0: i campioni provengono dalla medesima popolazione (stesse MEDIA e VARIANZA)

Anziché esaminare la differenza tra le medie si analizza la differenza tra le varianze (altrimenti si utilizza t-Student per i confronti a coppie i campioni, applicando opportune procedure che tengano conto che la probabilità dell'errore di I tipo cresce col numero di confronti => test di Bonferroni, di Tukey, di Scheffè, di Dunnet….).

Requisiti: i Campioni, tra loro indipendenti, provengono da popolazioni Normali (per testare Normalità→Normal Plot) con varianze omogenee (per testarlo => test Bartlett)

La variabilità dei dati è dovuta
- sia dal fatto che i soggetti appartengono a gruppi (o trattamenti) diversi, VARIANZA TRA GRUPPI, var(t)
- sia ad una variabilità individuale tra i soggetti anche di uno stesso gruppo/trattamento (che è la parte dovuta a errori di misura, diversità individuali, fattori non controllabili, ecc.), VARIANZA ENTRO I GRUPPI, var(E)

La varianza totale sarà quindi $var_{tot}=var(t)+var(E)$

Se i campioni provengono tutti dalla medesima popolazione (o da popolazioni 'indistinguibili') allora

$var(t) \sim 0$ e $var(E) \sim$ la varianza del fenomeno

altrimenti $var(t) \gg var(E)$ (tanto maggiore, quanto maggiori saranno le differenze tra i gruppi)

Si userà allora il **test di Fisher** per vedere quanto le varianze siano diverse tra loro:

$$F = \frac{var(t)}{var(E)}$$

Si confronta F con F_{α,n_t,n_E} con n_t e n_E = gradi di libertà del numeratore e denominatore.
Se F risulta essere minore, allora l'ipotesi nulla è vera, ovvero tutti i campioni provengono dalla medesima popolazione.

OSS: Se $var(t) < var(E)$, allora evidentemente H_0 è VERA!

La **varianza entro i gruppi** vale (la devianza media complessiva):

$$var(E) = \frac{\sum_{k=1}^{j} \sum_{i=1}^{n_k} (x_{i_k} - \overline{x_k})^2}{\sum_{k=1}^{j} (n_k - 1)}$$

j è il numero di gruppi; n_k è il numero di elementi del gruppo k-esimo;

$\sum_{k=1}^{j} (n_k - 1)$ è il numero di gradi di libertà ENTRO i gruppi;

$\sum_{i=1}^{n_k} (x_{i_k} - \overline{x_k})^2$ è la devianza nel k-esimo gruppo.

Questa espressione rappresenta una **stima della varianza della popolazione** (di cui i gruppi sono i campioni estratti), MIGLIORE di ciascuna varianza ottenibile separatamente in ciascun gruppo (in quanto tiene conto di un numero maggiore di osservazioni rispetto ciascun gruppo).

La **varianza tra i gruppi** (che si avvicina a zero quanto più le medie sono simili tra loro) sarà:

$$var(t) = \frac{\sum_{k=1}^{j} (\overline{x_k} - \bar{x})^2 \cdot n_k}{j - 1}$$

$\bar{x}$ è la media totale su tutte le osservazioni;

$\sum_{k=1}^{j} (\overline{x_k} - \bar{x})^2$ è la devianza di ciascun gruppo dalla media totale.

NOTA: solitamente è più comodo calcolare var_{tot} e var(t) per poi ricavare var(E) per semplice differenza, con

$$var_{tot} = \sum_{k=1}^{j} \sum_{i=1}^{n_k} {x_{i_k}}^2 \; - \; \frac{(\sum_{k=1}^{j} n_k * \overline{x_k})^2}{\sum_{k=1}^{j} n_k}$$

Si applicherà quindi il test di Fisher:

$$F = \frac{var(t)}{var(E)} < F_{\alpha, n_t, n_E}$$

con $n_t = j - 1; \qquad n_E = \sum_{k=1}^{j}(n_k - 1) = \sum_{k=1}^{j} n_k - j$

Esempio: Tasso colesterolo in 3 gruppi H0: no diff significative tra le medie

Professionisti (A) n_A=12; $\overline{x_A}$= 285; ${\sigma_A}^2$ =3140

Impiegati (B) n_B=14; $\overline{x_B}$= 224; ${\sigma_B}^2$ =1380

Agricoltori (C) n_C=10; $\overline{x_C}$= 195; ${\sigma_C}^2$ =666

$var(E) = 1772$ $var(t) = 23768$ F=13.4 $n_t = 2$ n_E = 33 $F_{0.05, n_t, n_E}$ =3.31

$\Rightarrow$ Rifiuto H0: Almeno uno dei 3 gruppi è significativamente distinto dagli altri due

Nota: con 2 gruppi il test di Fisher dà gli stessi risultati del t-Student

Test di Bartlett: necessario per valutare se le varianze sono tra loro omogenee e quindi nel caso in cui l'ANOVA dia valida l'H0.

H_0: le varianze sono stime indipendenti di varianza di una popolazione e le differenze sono dovute al caso.

Si valuta:

$$\frac{A}{B} = \frac{2.3026 \cdot \left[(n-k)\log_{10}\bar{s}^2 - \sum_{i=1}^{k}(n_i - 1)\log_{10} s_i^2\right]}{1 + \frac{1}{3(k-1)} \cdot \left[\sum_{i=1}^{k}\frac{1}{n_i - 1} - \frac{1}{n-k}\right]}$$

k è il numero di gruppi; n_i è il numero di campioni nell'i-simo gruppo; $n = \sum_{i=1}^{k} n_i$; s_i^2 è la varianza nell'i-simo gruppo; $\bar{s}^2$ è la varianza complessiva.

Il rapporto si distribuisce come una χ^2 con k-1 gradi di libertà.

Se $\chi^2_{A/B} < \chi^2{}_{\alpha,k-1}$, allora l'ipotesi nulla va accettata e le varianze sono OMOGENEE

Test di Kruskal-Wallis (non Parametrico)

se le ipotesi richieste per l'analisi della varianza non sono soddisfatte
ma almeno si hanno: a) Indipendenza tra i campioni; b) Numerabilità;
allora si può utilizzare questo test che rappresenta il caso generale del test di Mann-Whitney

H_0: i gruppi appartengono alla stessa popolazione (le differenze tra le sommatorie dei ranghi sono attribuibili solo al caso)

Per valutare il test, si prendono le N osservazioni tutte insieme (tutti i gruppi), si valutano i ranghi e quindi la statistica:

$$H = \frac{12 \cdot \sum_{i=1}^{k} n_i (\bar{R}_i - \bar{R})^2}{N(N+1)}$$

Con:

n_i numero di osservazioni nell'i-esimo gruppo;
R_i sommatoria dei Ranghi dell'i-esimo gruppo;
$\bar{R}_i$ rango medio dell'i-esimo gruppo;
$\bar{R}$ media dei Ranghi;
k numero di gruppi.

H cresce col crescere della variazioni fra i gruppi e si distribuisce come una $\chi^2_{\alpha,k-1}$ (ad una sola coda perché H può solo crescere)

Se $H > \chi^2_{\alpha,k-1}$, si respingerà l'ipotesi nulla => esiste una differenza significativa tra i gruppi

VERIFICA DELLE IPOTESI (sintesi)

1 CAMPIONE:

Test sulla media: Popolaz con distrib Normale e nota σ => Z – test

Popolaz con distrib Normale ma ignota σ => t-Student

Ignota distribuzione => Sign / Wilcoxon signed rank sum test

Test sulla frequenza: Tabelle con sufficiente numerosità => χ^2

Tabelle con bassa numerosità => Kolmogorov

2 CAMPIONI:

Test su differenza di medie: distrib Normale e nota σ => Z – test

distrib Normale ma ignota σ => t-Student + Fisher (se H0 è vera)

Ignota distribuzione => Wilcoxon-Mann-Whitney

Test sulle diff di frequenza (tabelle 2 x 2): test Z, Z modificato, χ^2 + Yates, test esatto di Fisher, test di McNemar

3 O PIU' CAMPIONI:

Test su differenze di varianze: Popolaz con distr Normale => ANOVA + Bartlett

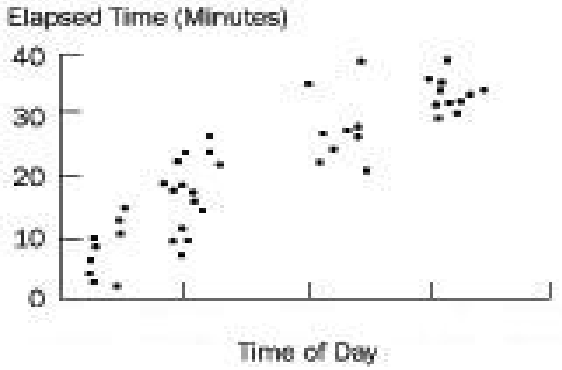
Ignota distribuzione => Kruskal – Wallis

Tabelle con 2 variabili (tabelle r x c) => χ^2

RELAZIONI TRA FENOMENI (VARIABILI)

RELAZIONE TRA 2 VARIABILI

Retta di Regressione (relazione lineare): si può utilizzare per predire Y, data una qualsiasi X



Si parte da uno SCATTER DIAGRAM

Assunti:

- i valori di Y (variabile DIPENDENTE) devono essere distribuiti come una Normale per ciascun valore di x
- la varianza di Y deve essere identica per ciascuna x, ovvero deve essere verificata l'OMOSCEDASTICITA'
- la relazione deve essere lineare.

NON è necessario che entrambe le variabili siano aleatorie (casuali), né che X sia Normale!

Per verificare i 3 assunti, si calcola la relazione: $Y = a + bX$

a e *b* opportuni per minimizzare le distanze verticali (RESIDUI): $\sum_{i=1}^{n}(y_i - Y(x_i))^2$, dove y_i sono i valori osservati e $Y(x_i)$ sono quelli teorici; si ricava:

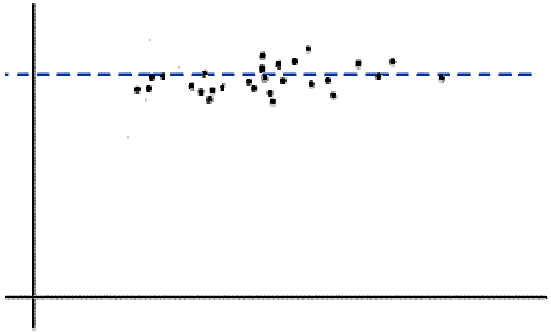
$$\boldsymbol{b = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{cov(x,y)}{var(x)}} \qquad \boldsymbol{a = \bar{y} - b\bar{x}}$$

Un discorso simile si può effettuare scambiando le 2 variabili considerando la X come variabile dipendente e la Y indipendente
Si dovrà quindi calcolare:

$$X = a' + b'Y$$

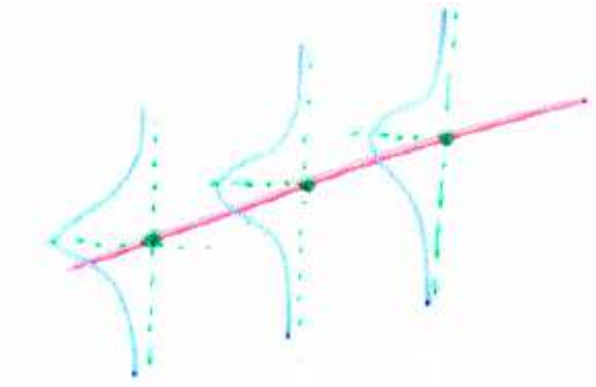
Distanze calcolate in orizzontale anziché in verticale
Se gli assunti sono veri, i residui devono essere distribuiti Normalmente e questo si può testare con il Normal Plot

Residui =>



Nota: *b* e *b'* possono essere utilizzati come indici di concordanza, indicando quanto cresce in media una variabile al crescere unitario dell'altra. Esse rappresentano anche asimmetria nel rapporto tra variabili (=> coeff. correlazione)

Nell'ipotesi che la distribuzione di Y in corrispondenza ad ogni x_i sia normale:



nell'ipotesi che la varianza sia uguale per tutti i punti (OMOSCEDASTICITA') si possono valutare le deviazioni standard della pendenza (*b*) e dell'intercetta (*a*):

$$\sigma_a = \sigma \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x-\bar{x})^2}} \qquad \sigma_b = \frac{\sigma}{\sqrt{\sum(x-\bar{x})^2}}$$

Per stimare σ ($\hat{\sigma} = errore\ standard\ della\ stima$) utilizzo S_{yx}:

$$\hat{\sigma} = S_{yx} = \sqrt{\frac{\sum(y_i - Y(x_i))^2}{n-2}}$$

Per testare la bontà dei valori della pendenza e dell'intercetta trovati si userà il t-test
Nel caso della **pendenza** (b) si avrà:

$$t = \frac{b - \beta_0}{\sigma_b}$$

da confrontare con $t_{\alpha,n-2}$. β_0 è il valore da testare *(Es. H_0: non esiste relazione tra qual è la probabilità che un campione con particolari y in x prefissate, dia una pendenza $b \geq \beta_0$)*.
Se $\beta_0 = 0$, si testa l'ipotesi che non vi sia alcun legame tra x e y.
L'intervallo di confidenza della pendenza sarà quindi:

$$b \pm t_{\alpha,n-2} \cdot \sigma_b$$

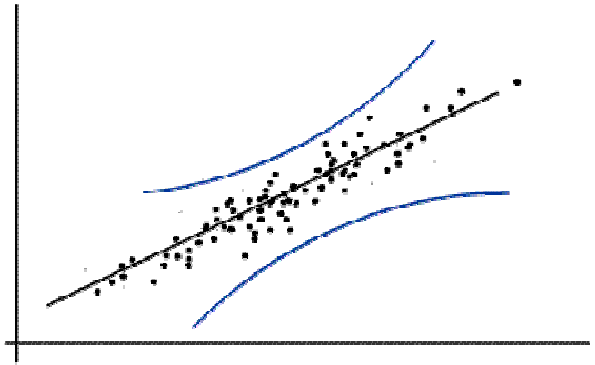
Per l’**intercetta** (a) si usa:

$$t = \frac{a - \alpha_0}{\sigma_a}$$

con $t_{\alpha,n-2}$ e α_0 un valore da testare, come per la pendenza.
L’intervallo di confidenza dell’intercetta sarà:

$$a \pm t_{\alpha,n-2} \cdot \sigma_a$$

Limiti di confidenza rispetto la retta si ottengono combinando i possibili valori (a intervalli) di a e b



INTERPRETAZIONI E LIMITI
affinchè i risultati siano significativi
- le osservazioni devono essere indipendenti (Es: 1 sola misura per ogni individuo)
- non si deve usare la relazione oltre il campo delle x da cui si è partiti (no estrapolazioni)
- data x, si può predire Y, ma non viceversa
- gli intervalli di confidenza per b indicano l’incertezza nella forza della relazione tra y e x
- la retta di regressione indica quanto della variabilità di y può essere spiegata (in modo lineare) da x e quanta variabilità resta non spiegata (quota parte dovuta a rumore)

Coefficiente di correlazione (lineare)

Per uniformare le informazioni delle 2 rette di regressione, ovvero non considerare più una variabile dipendente e una indipendente, ma entrambe aleatorie, si utilizza il coefficiente di correlazione lineare di Bravais-Pearson:

$$r = \pm\sqrt{b' \cdot b} = \frac{cod(x,y)}{dev(x)dev(y)} = \frac{cov(x,y)}{var(x)var(y)} =$$

$$= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \cdot (y_i - \bar{y})^2}} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)\sigma_x\sigma_y} =$$

$$= \frac{\sum x_i y_i - \sum x_i \sum y_i / n}{\sqrt{\left[\sum x_i^2 - (\sum x_i)^2/n\right]\left[\sum y_i^2 - (\sum y_i)^2/n\right]}}$$

$$-1 \leq r \leq 1 \quad \text{(adimensionale)}$$

È una misura simmetrica che dà informazione sull'interdipendenza tra le variabili, ovvero una misura della dispersione dei dati rispetto ad un andamento lineare.
Se r=0, non c'è correlazione, più $r \to 1\ (-1)$ maggiore è la correlazione.

Per vedere se è significativamente distante da 0, si valuta:

$$t = \frac{r}{\sqrt{(1-r^2)/n-2}}$$

e si confronta con $t_{\alpha,n-2}$.

Assunti:
- per calcolare l'intervallo di confidenza è necessario che sia la x che la y provengano da distribuzioni normali (W-test per valutarlo)
- le osservazioni devono essere indipendenti (1 sola osservazione per ciascun individuo, NON ripetute!), altrimenti **l'ANALISI NON E' VALIDA**!

Una volta trovato che r è significativamente vicino a 1 (o -1) **non si può** direttamente **dire** se x dipende da y o viceversa o addirittura che x e y dipendano da un terzo fattore

Esiste anche il **COEFFICIENTE DI DETERMINAZIONE**: $r^2 = b \cdot b'$ che esprime la variabilità di Y, attraverso la variabilità di X.

$1 - r^2$ esprime la porzione di varianza di Y che dipende da fattori diversi da X.

Esistono anche correlazioni basate sui Ranghi (NON parametriche), con ipotesi iniziale di sola indipendenza, non di Normalità.
- Coefficiente di Spearman=r_s, si calcola come r ma sui ranghi delle x e delle y
- Coefficiente di Kendall=r_t

APPROCCIO MATRICIALE

regressione lineare $\boldsymbol{Y = XB + \varepsilon}$

$$Y = \begin{vmatrix} y_1 \\ \vdots \\ y_n \end{vmatrix}, X = \begin{vmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{vmatrix}, \mathrm{B} = \begin{vmatrix} \beta_0 \\ \beta_1 \end{vmatrix}, \varepsilon = \begin{vmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{vmatrix}$$

←con valor medio nullo e varianza = σ

↑ la colonna con "1" indica che l'intercetta è inclusa in questa matrice

Ai minimi quadrati avremo (noti che siano X e Y):

$$\widehat{\mathrm{B}} = (X^T \cdot X)^{-1} \cdot X^T \cdot Y$$

$$E(Y) = X\mathrm{B}$$

$$\varepsilon\varepsilon^T = \begin{vmatrix} \varepsilon_1^2 & \varepsilon_1\varepsilon_2 & \dots \\ \varepsilon_2\varepsilon_1 & \varepsilon_2^2 & \dots \\ \vdots & \vdots & \varepsilon_n^2 \end{vmatrix} \Longrightarrow E(\varepsilon\varepsilon^T) = \begin{vmatrix} \varepsilon_1^2 & \dots & 0 \\ 0 & \varepsilon_2^2 & 0 \\ 0 & \dots & \varepsilon_n^2 \end{vmatrix}$$

perché le ε sono indipendenti => $E(\varepsilon_i\varepsilon_j) = 0$ per $i \neq j$

correlazione parziale e multipla

$$R = \begin{vmatrix} 1 & r_{12} & \cdots & r_{1k} \\ r_{21} & 1 & \cdots & r_{2k} \\ \vdots & \vdots & \cdots & \cdots \\ r_{k1} & \cdots & \cdots & 1 \end{vmatrix}$$

regressione multipla lineare: $\boldsymbol{Y = XB + \varepsilon}$

$$X = \begin{bmatrix} x_0 & \cdots & x_{k1} \\ \vdots & \ddots & \vdots \\ x_0 & \cdots & x_{kn} \end{bmatrix}, \mathrm{B} = \begin{vmatrix} \beta_0 \\ \vdots \\ \beta_n \end{vmatrix}$$

$$X \cdot X^T = \begin{vmatrix} n & \sum x_{1i} & \sum x_{2i} & \cdots & \sum x_{ki} \\ \sum x_{1i} & \sum x_{1i}^2 & & \vdots \ \vdots \ \vdots & \\ \vdots & & & \vdots \ \vdots \ \vdots & \\ \sum x_{ki} & \vdots \ \cdots & \cdots & \cdots & \sum x_{kn}^2 \end{vmatrix} \Longrightarrow \hat{\beta} = (X^T \cdot X)^{-1} X^T Y$$

Come scegliere il test piu' adatto per la verifica di ipotesi

Scala di misura	*Due gruppi con sogg. diversi*	*>2 gruppi con sogg. diversi*	*Prima e dopo con gli stessi sogg.*	*Piu' tempi negli stessi sogg.*	*Associazione fra 2 variabili*
Intervallare ("normale")	*T-test di Student*	***ANOVA***	*Paired t test*	***ANOVA per misure ripetute***	*Correlazione di Pearson e regressione lineare*
Nominale	*Chi quadro*	*Chi quadro*	*Test di McNemar*	*Test Q di Cochrane*	*Coefficiente di contingenza (Test K di Kendall)*
Ordinale	*Test per la somma dei ranghi di Mann-Whitney*	*Test di Kruskal-Wallis*	*Test di Wilcoxon*	*Test di Friedman*	*Correlazione dei ranghi (test di Spearman)*

I primi 23 numeri fattoriali

1! =	1
2! =	2
3! =	6
4! =	24
5! =	120
6! =	720
7! =	5.040
8! =	40.320
9! =	362.880
10! =	3.628.800
11! =	39.916.800
12! =	479.001.600
13! =	6.227.020.800
14! =	87.178.291.200
15! =	1.307.674.368.000
16! =	20.922.789.888.000
17! =	355.687.428.096.000
18! =	6.402.373.705.728.000
19! =	121.645.100.408.832.000
20! =	2.432.902.008.176.640.000
21! =	51.090.942.171.709.440.000
22! =	1.124.000.727.777.607.680.000
23! =	25.852.016.738.884.976.640.000