Lezioni di Metodologia Sperimentale

Cesare Bini
Dipartimento di Fisica
Sapienza Università di Roma

Programma delle lezioni

- Perché occorre fare un'analisi statistica dei dati sperimentali in Medicina ?
- Il caso delle variabili continue:
  - come presentare e riassumere i dati
  - perché la distribuzione di Gauss ?
  - quando entra in gioco la distribuzione t-Student ?
- Il caso delle variabili discrete:
  - necessità di un modello statistico
  - i fenomeni di Bernoulli
  - i fenomeni di Poisson
  - il limite gaussiano
- Intermezzo semi-filosofico: che cos'è la probabilità ?
- I test di ipotesi
  - Che significato dobbiamo attribuire a P ?
  - Alcuni esempi di base: Gauss, t-Student e χ^2
  - Alcuni test usati in medicina

Perché occorre fare un'analisi statistica dei dati sperimentali in Medicina ?

- Cosa fate voi in Medicina ?
  - Set di dati sperimentali (*campione*)
  - *Ipotesi* (anche diverse ipotesi) da "rigettare" o "corroborare";
  - Numero magico *P* calcolato sulla base dei dati e dell'ipotesi: se $P<0.05$ l'ipotesi è "rigettata", se $P>0.05$ l'ipotesi è "corroborata".
- Alcune considerazioni immediate di cui tenere sempre conto (e su cui torneremo):
  (1) Importanza della scelta e delle caratteristiche del campione (unbiased sample, blind o double-blind analysis, control "placebo" samples, repeatibility conditions,…)
  (2) "rigettare" o "corroborare" ma non "verificare": "Absence of evidence is not evidence of absence"
  (3) *P* è una probabilità: ➔ il risultato che si ottiene non è deterministico ma probabilistico;

Due esempi di GedankenExperiment

- Vogliamo stabilire se un certo trattamento medico (farmacologico o chirurgico o di qualsiasi tipo) abbia o meno un effetto positivo su una certa patologia.
- Due campioni di pazienti. Sottoponiamo il primo campione al trattamento (*campione trattato*) e non sottoponiamo il secondo campione(*campione di controllo* o *placebo*) allo stesso trattamento.
- Caratteristiche dei due campioni:
  - equivalenza, età, sesso, condizioni generali dei pazienti;
  - inconsapevolezza del fatto di essere o meno trattati (placebo controlled experiment);
  - i medici non sono a conoscenza di quali pazienti sono stati trattati (*double-blind analysis*)
- Facciamo dunque l'ipotesi che queste condizioni di base siano ben soddisfatte. Al termine di un periodo di tempo fissato a priori andiamo a vedere l'esito del trattamento confrontando i due campioni.

Esempio (1) - I

- Criterio di "guarigione" univoco (non esistono "mezze guarigioni").
- Conteggio, in ciascun campione, di "guariti" e non "guariti".
- Definisco i simboli rilevanti:
  - N_t = numero di pazienti del campione trattato;
  - N_c = numero di pazienti del campione di controllo;
  - n_t = numero di pazienti guariti del campione trattato;
  - n_c = numero di pazienti guariti del campione di controllo;
- Deve essere:
  - $0 \le n_t \le N_t$
  - $0 \le n_c \le N_c$
- E' inoltre naturale definire altre due grandezze: le *frazioni di pazienti guariti*:
  - $$f_t = \frac{n_t}{N_t}$$
  - $$f_c = \frac{n_c}{N_c}$$

Esempio (1) - II

- f_t ed f_c sono $0 \leq f_t \leq 1$, $0 \leq f_c \leq 1$, esprimibili anche come %. Sono numeri reali, non interi. Sono grandezze "comode".
- La risposta del mio test sarà:
  - se $f_t > f_c$, il trattamento ha un effetto positivo
  - se $f_t = f_c$, il trattamento non ha effetto
  - se $f_t < f_c$, il trattamento è addirittura nocivo!
- MA:
  - che significa >, =, < in questo contesto ?
  - e.g.: trovo f_t=23.2% ed f_c=19.4%, cosa posso dire ? Sarà un f_t = 23.2% ± qualcosa e f_c=19.4% ± qualcos'altro.
- Quindi devo trovare un criterio per stabilire quando una cosa sia "*significativamente*" maggiore, minore o uguale ad un'altra. E' qui la motivazione dell'analisi statistica !

Esempio (2) - I

- L'esito del trattamento non è guarigione o no ma la variazione di una certa grandezza *X* che segnala lo stato della patologia.
- *X* misurata per tutti i pazienti di ambedue i campioni prima e dopo il trattamento.
- Definisco i simboli rilevanti: chiamo X^t_{ini} e X^t_{fin} i valori iniziali e finali di X per il campione trattato; X^c_{ini} e X^c_{fin} quelli del campione di controllo. E' comodo costruire dei "*vettori*" di dati (tabelle):

$$X^t_{ini}(k), k = 1, N_t \qquad X^c_{ini}(k), k = 1, N_c$$

$$X^t_{fin}(k), k = 1, N_t \qquad X^c_{fin}(k), k = 1, N_c$$

- Devo confrontare non numeri, ma insiemi di numeri: devo imparare a trattare e riassumere un insieme di numeri.
- Di nuovo:
  - se " X^t_{fin} " >, =, < " X^t_{ini} " diversa conclusione
  - ma deve essere "*significativamente*" >, =, <
  - Inoltre, in questo caso a che serve il campione di controllo ?

Proviamo a ripetere un esperimento

- Torniamo all'esempio (1). Prendo un terzo campione, "equivalente" ai primi due, e lo tratto esattamente come il campione "trattato". Che mi aspetto ?
  - $N_t = 100$, $n_t = 24$
  - $N_r = 100$, $n_r = 24$? $= 0$? $=26$? $= 100$?
- Se dovessi "scommettere" su un valore, su quale scommetterei ?
- Se immaginassi di prendere tanti campioni, tutti "equivalenti", il valore che otterrei di n_t non sarebbe sempre lo stesso, ma sarebbe un valore che fluttua secondo una certa legge.

➔ La statistica insegna a trattare queste fluttuazioni e permette dunque di fare anche previsioni.

Che succede se prendo un campione più numeroso ?

- Sempre nel quadro dell'esempio (1):
  - $N_t = 10$; $n_t = 4$; $f_t = 40\%$
  - $N_c = 10$; $n_c = 1$, $f_c = 10\%$

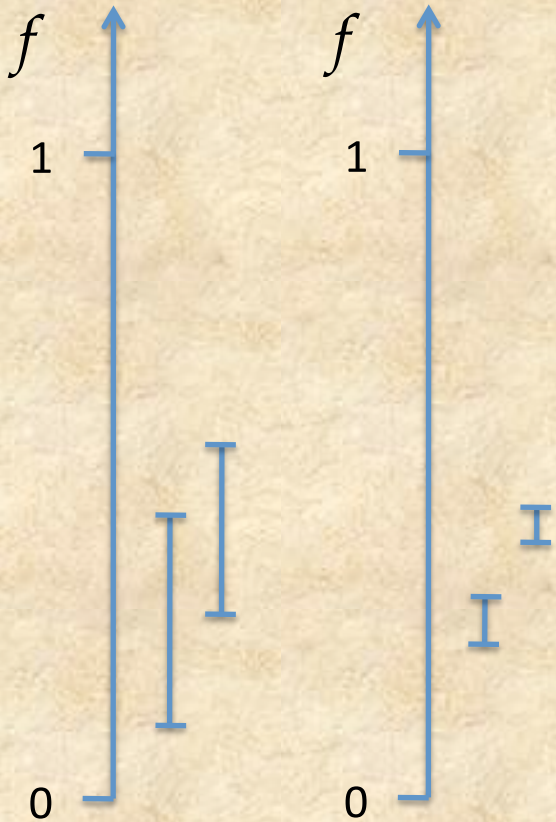
  cosa concludo ? Voi annuncereste la grande scoperta ?

Se invece:
  - $N_t = 100$; $n_t = 40$; $f_t = 40\%$
  - $N_c = 100$; $n_c = 10$; $f_c = 10\%$

  vi sentireste più sicuri ?

➔sembra che al crescere della dimensione del campione si guadagni in capacità di discriminare.
(il ± qualcosa diminuisce ?)

Significatività



Nel primo caso:

f_t=(32±8)%

f_c=(23±14)%

Nel secondo caso:

f_t=(32±2)%

f_c=(23±3)%

Stessi valori centrali ma incertezze ridotte.

Conclusioni:

Primo caso "Absence of evidence"

Secondo caso "Significant evidence"

Ricapitolando:

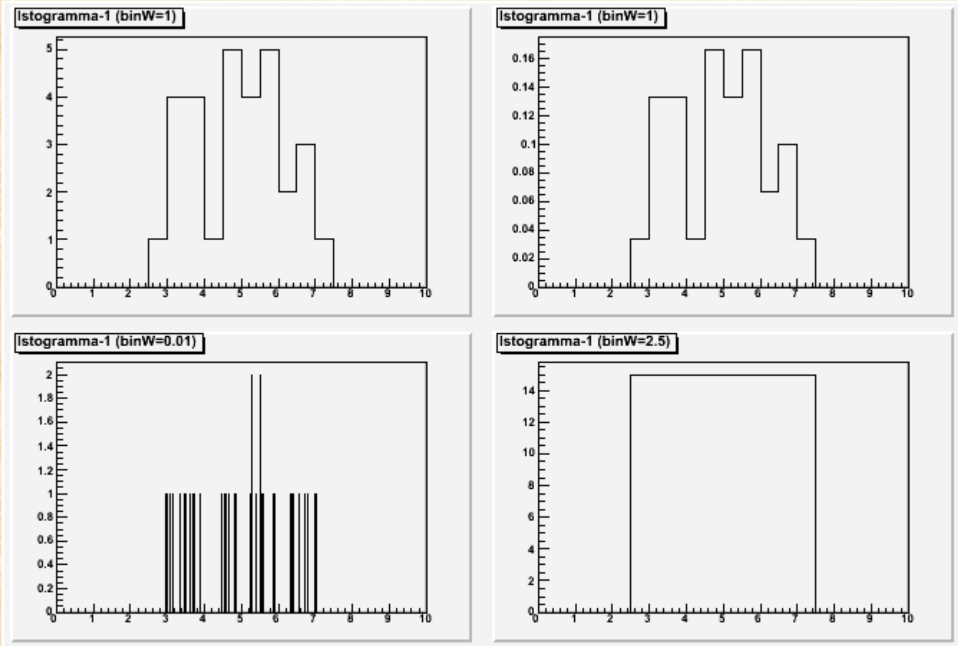
- Abbiamo imparato che:
  - Qualsiasi test clinico si basa su un campione;
  - Il risultato di una misura su un campione è caratterizzato da fluttuazioni;
- Cosa chiediamo all'analisi statistica dei dati ?
  - un modello per trattare le fluttuazioni
  - un metodo per presentare i dati e riassumerli
  - un metodo per confrontare numeri, relativi a campioni diversi: significatività.
- Prossime lezioni:
  - Esempio – (2): impariamo a trattare i dati
  - Esempio – (1): costruiamo un modello statistico.

Il caso delle variabili continue

- Prendiamo in esame l'Esempio – (2): ciascun campione ci fornisce 2 gruppi di dati. Vediamo come trattare ciascun campione.
- Si tratta di variabili continue: X non è un conteggio, ma un "numero reale" caratterizzato da delle unità di misura (ml^{-1}, g/cm^3, …);
- Abbiamo un "vantaggio" un campione di misure ripetute "equivalenti".
  - : come presento i dati ?
  - : come li riassumo ?

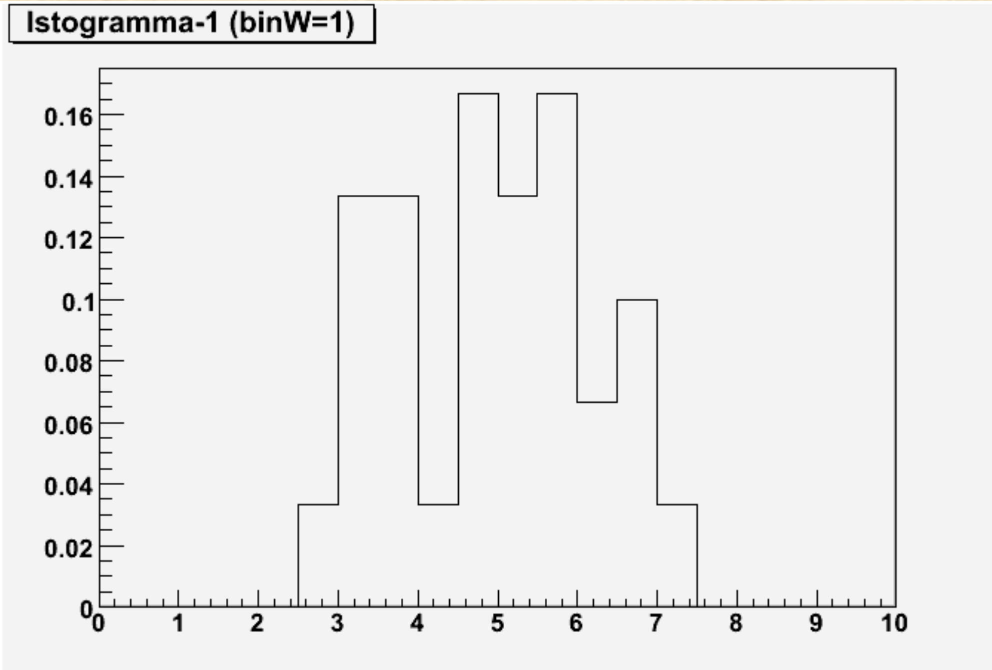
Tabella/sequenza di numeri

3.16	5.52	4.56
3.88	5.57	6.39
4.85	4.45	5.53
3.72	3.34	5.52
4.66	2.97	6.79
6.33	6.70	5.28
5.25	6.55	3.60
3.70	5.40	5.28
3.47	4.82	5.89
7.01	4.81	3.07



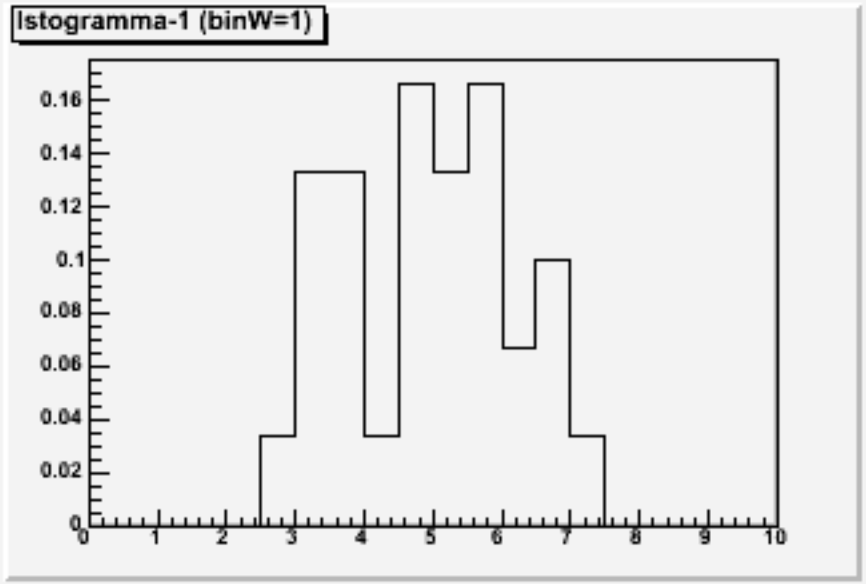
Faccio un *istogramma* delle misure
- sto perdendo informazioni (non sempre);
- scelgo il "binning" (3 esempi, 2 "estremi");
- N. eventi o frequenze: n o n/N ?

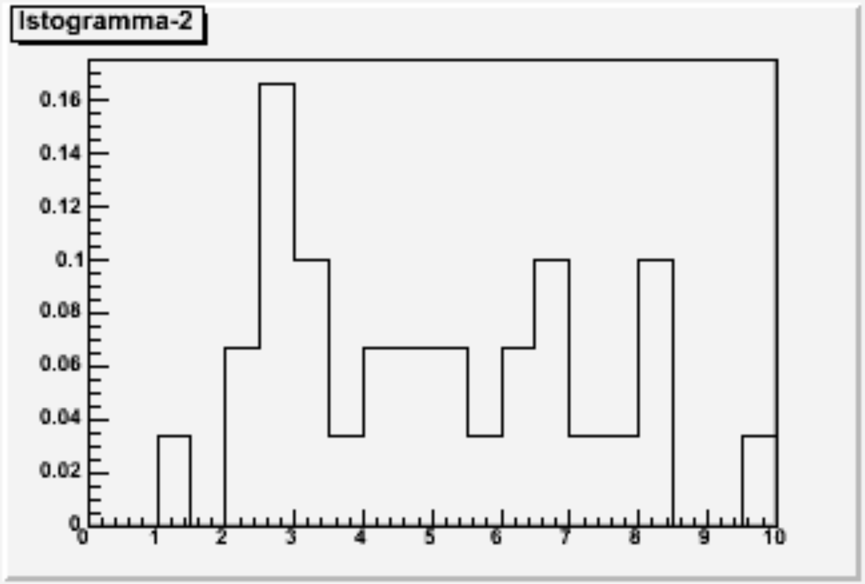
Caratteri di un istogramma



- Forma
- Valore centrale
- Dispersione

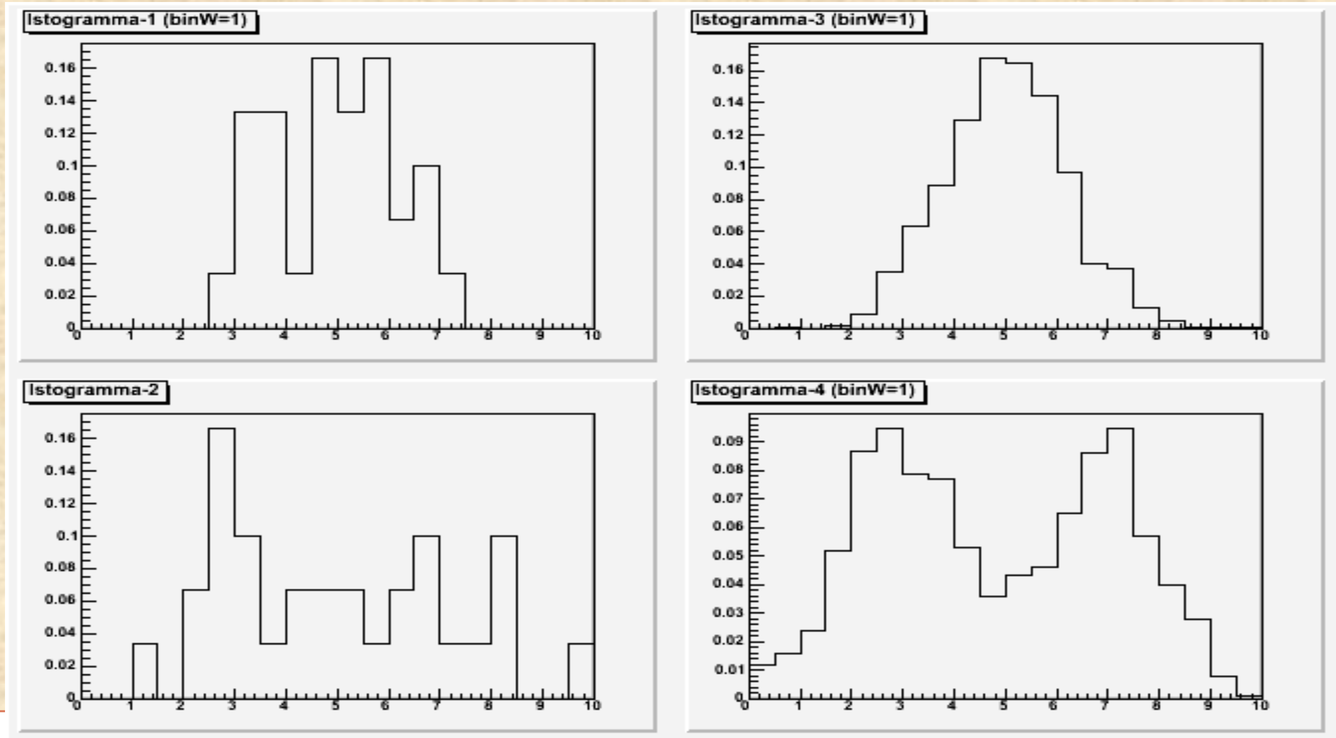
Forma dell'istogramma - I





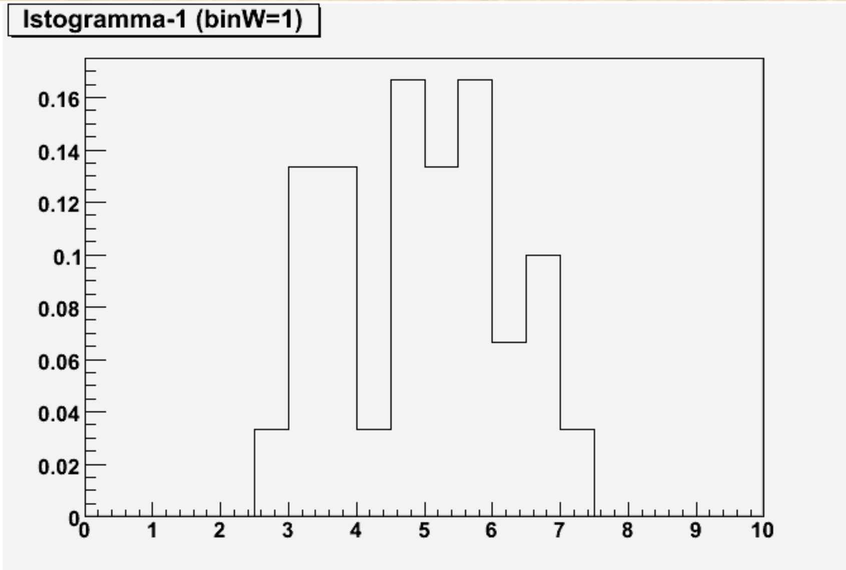
- Istogramma-1: misure semplicemente "sparpagliate" – non si osservano strutture;
- Istogramma-2: sembrano dividersi in due zone – capire se le diverse zone sono legate a diversi caratteri del campione: campanello c'allarme per lo sperimentatore

Forma dell'istogramma - II



- Aumento la "statistica" (la dimensione del campione da 30 a 1000 valori):
  - NON si stringe o si allarga l'istogramma;
  - semplicemente diventa "meglio definito";
  - I sospetti di "struttura" acquistano più solidità !

Valore centrale

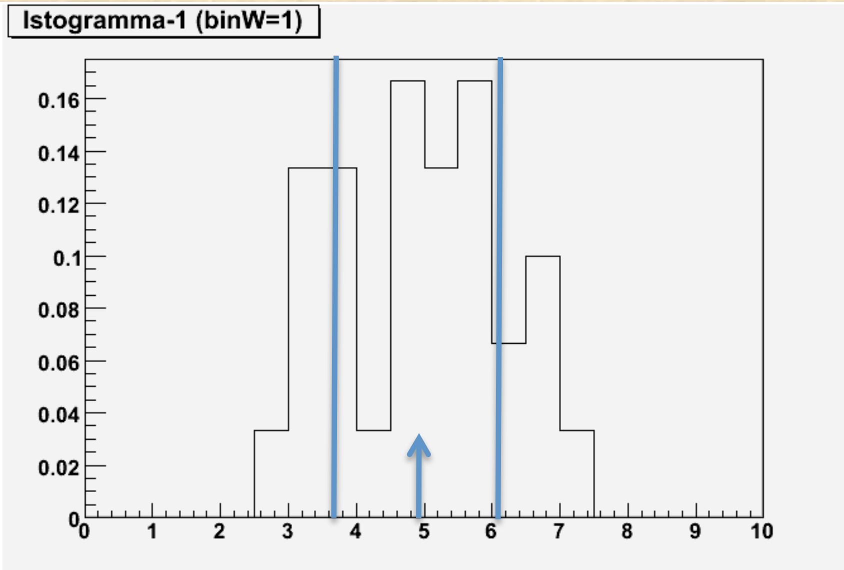


(Max+Min)/2 =	4.99
Moda =	4.75 o 5.75
Mediana =	5.06
Media =	4.93

- (Max + Min)/2
- Moda = "valore più frequente";
- Mediana = quel valore che divide l'istogramma in parti uguali;
- Media aritmetica:

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_N}{N} = \frac{\sum_{k=1}^{N} X_K}{N}$$

Dispersione - I



(Max-Min)/2 =	2.02
$s(X)$ =	1.18
$s^2(X)$ =	1.40

- (Max - Min)/2
- Deviazione Standard Campionaria

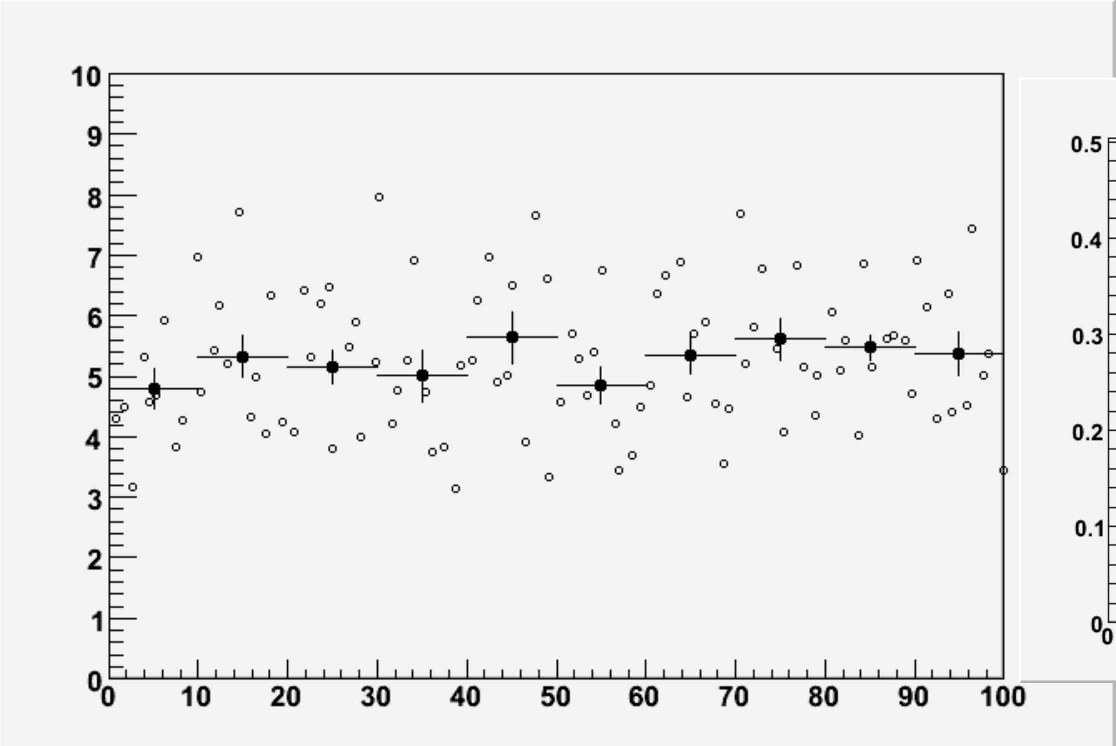
$$s(X)=\sqrt{\frac{\sum_{k=1}^{N}(X_k-\overline{X})^2}{N-1}}$$

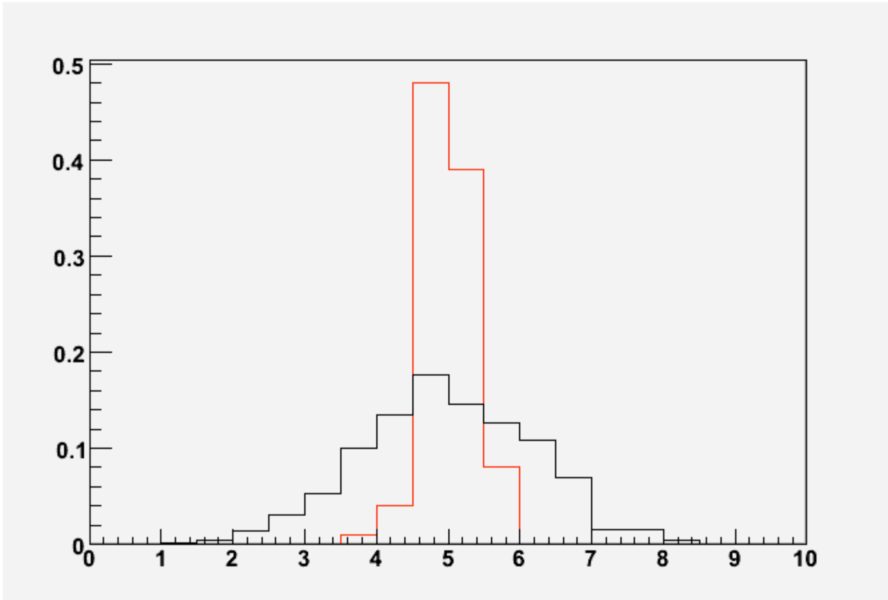
- Si chiama Varianza campionaria $s^2(X)$.

Consideriamo ora l'intervallo: $\overline{X} \pm s(X) = 4.93 \pm 1.18 = 3.75 \leftrightarrow 6.11$
che significato ha ? Comprende circa il 70% dei valori ottenuti.
Se facessi $\overline{X} \pm 3s(X)$ comprenderei il 100% dei valori ottenuti.

Dispersione - II

"Dumping" delle fluttuazioni:
La media aritmetica ha il "potere" di ridurre le fluttuazioni !





Campione di 100 misure che raggruppo in 10 sottocampione da 10 misure ciascuno.
Faccio la media per ogni sottocampione. La dispersione si riduce di $\approx 3 \approx \sqrt{10}$ volte.

Dispersione - III

Due domande ora. Dato il mio campione di 30 misure dalle quali ho calcolato $\overline{X}$ e $s(X)$:

(1) Se effettuo una 31-esima misura, dove mi aspetto che cada ?

(2) Se ripeto una sequenza di altre 30 misure, dove mi aspetto cada la media delle nuove 30 misure ?

Risposta:

- (1) nel ≈70% dei casi dentro l'intervallo $\overline{X} \pm s(X)$
- (1) nel 100% dei casi dentro $\overline{X} \pm 3s(X)$
- (2) nel ≈70% dei casi dentro l'intervallo $\overline{X} \pm \frac{s(X)}{\sqrt{N}}$
- (2) nel 100% dei casi dentro $\overline{X} \pm 3\frac{s(X)}{\sqrt{N}}$

Ricapitolando

Qual è dunque la migliore stima di *X* che posso dare sulla base del mio campione ?

$$\overline{X} \pm \frac{s(X)}{\sqrt{N}}$$

Il significato è: se ripeto una sequenza di *N* misure e ne faccio la media aritmetica, con una probabilità del 70% troverò una media dentro quell'intervallo.

Al crescere di *N*, migliora la mia conoscenza e la mia capacità di discriminare tra ipotesi:

$N \rightarrow N\times 10 \;\Rightarrow\; incertezza \rightarrow incertezza\ /\ 3$

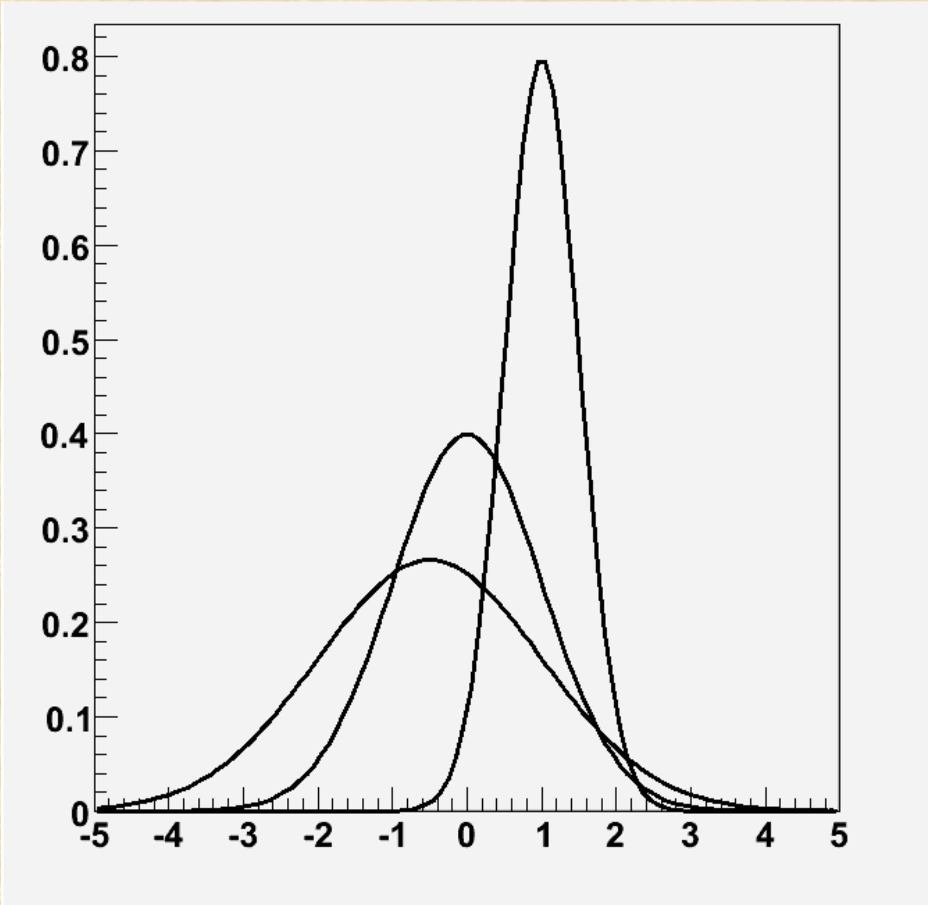
$N \rightarrow N\times 100 \;\Rightarrow\; incertezza \rightarrow incertezza\ /\ 10$

La distribuzione di Gauss - I

- Perché in genere gli istogrammi hanno forme "a campana" ?
- Teorema del Limite Centrale: "la somma di tante variabili causali di qualunque forma, tende a distribuirsi secondo la distribuzione di Gauss"

$$G(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- dipendente da due parametri: μ e σ
  - μ corrisponde alla media aritmetica
  - σ corrisponde alla deviazione standard campionaria



La distribuzione di Gauss - II

TABELLA della gaussiana.

-Gaussiana Standardizzata

$m=(x-\mu)/\sigma$

$(\mu=0, \sigma=1)$

-F(z) = Prob(m<z)

-F(-z) = 1 – F(z)

-Percentili / Quartili

Esempio: la distribuzione dei pesi delle bambine di 10 anni è descritta da una distribuzione di Gauss con μ=36.2 kg e σ=5.0 kg. Mia figlia pesa 44.1 kg: a che percentile si trova ?

(1) z = (44.1-36.2)/5.0 = 1.58

(2) F(z) = 0.94295

(3) ➔ 94-esimo percentile

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0,1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0,2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0,3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0,4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0,5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0,6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0,7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0,8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0,9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1,0	0.84134	0.84375	0.84614	0.84850	0.85083	0.85.314	0.85543	0.85769	0.85993	0.86214
1,1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1,2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1,3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1,4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92786	0.92922	0.93056	0.93189
1,5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1,6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1,7	0.95543	0.95637	0.95.728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1,8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1,9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2,0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2,1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2,2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2,3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2,4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2,5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2,6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2,7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2,8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2,9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3,0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99897	0.99900
3,1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3,2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3,3	0.99952	0199953	0.99957	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3,4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976

La distribuzione t di Student - I

Se conosco μ e σ di una distribuzione di Gauss, conosco in maniera completa il comportamento della variabile X. Ma in molti casi, non conosco direttamente μ e σ, al massimo posso valutare media e deviazione standard campionaria dal campione di dimensione N e assumerli come "stime" di μ e σ.
Si vede allora che:

se $\dfrac{X-\mu}{\sigma}$ è distribuita secondo la distribuzione di Gauss;

allora $\dfrac{X-\overline{X}}{s(X)/\sqrt{N}}$ è distribuita secondo la distribuzione t-Student

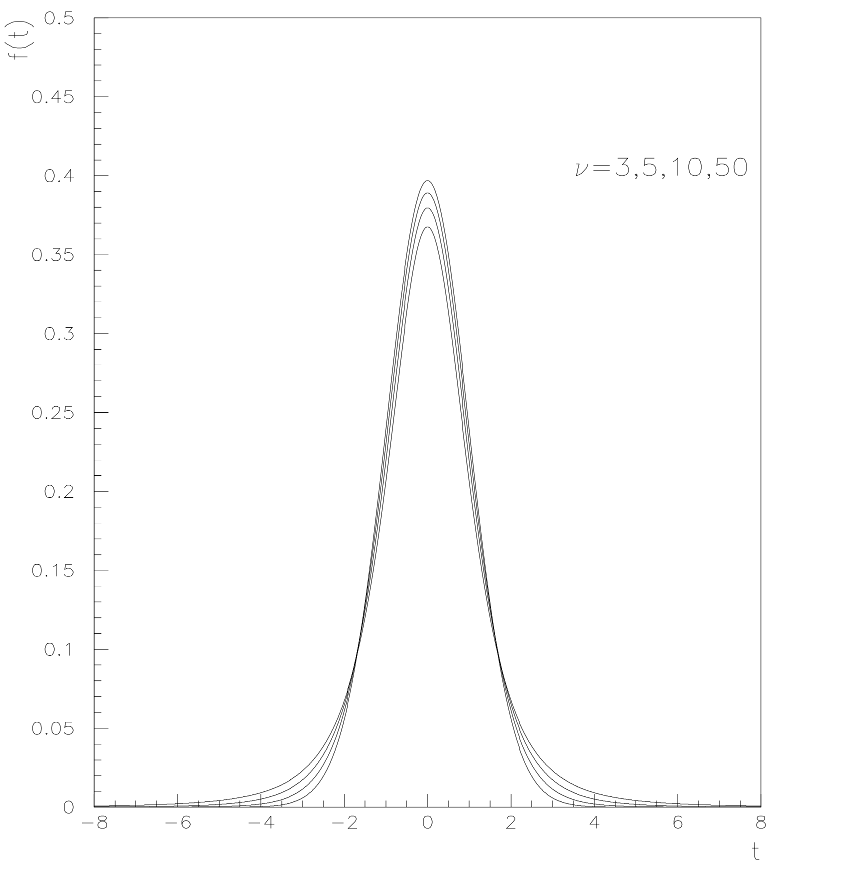
Quindi: la t-Student è una "correzione" alla Gaussiana quando sono in presenza di un campione. Per N grandi t-Student → Gaussiana.

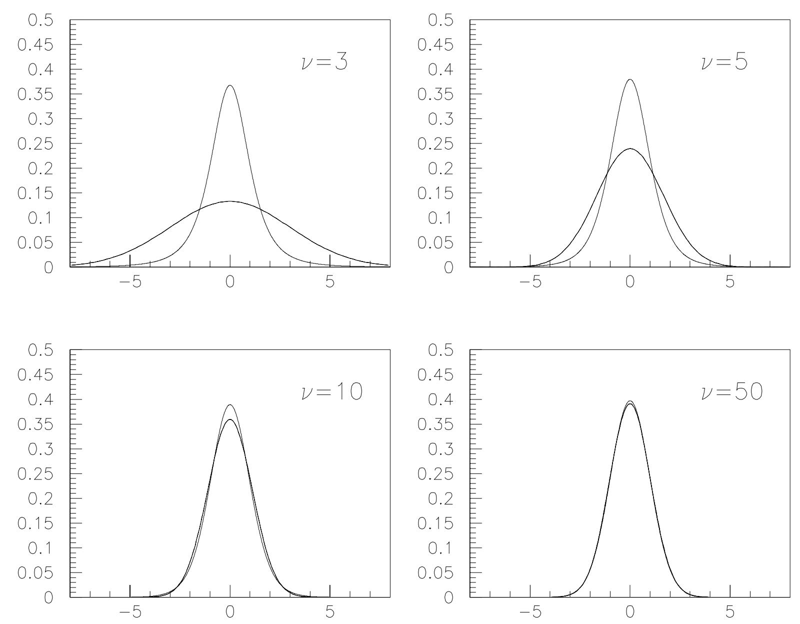
Il numero N di misure del mio campione si chiama
"***Numero di gradi di libertà***"

La distribuzione t di Student - II

E' una Gaussiana con "code più estese"

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\,\Gamma(\nu/2)}\left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$





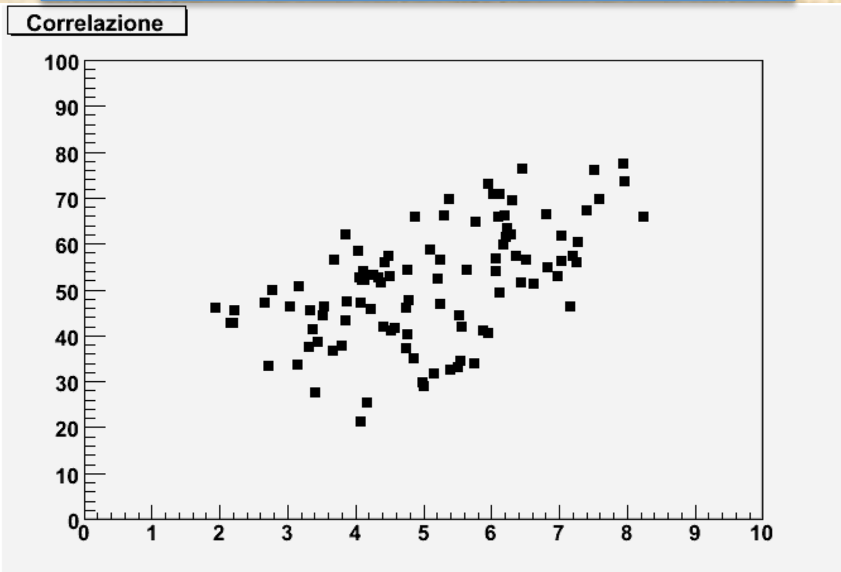
Tabella/sequenza d*i coppie* di numeri

Aggiungiamo un'informazione: per esempio l'età del paziente

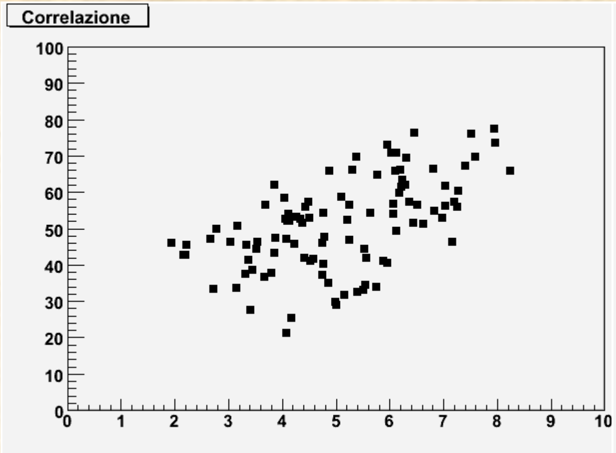
- Ho una sequenza di coppie di numeri;
- Come tratto questo caso graficamente ?
- A cosa sono interessato ?

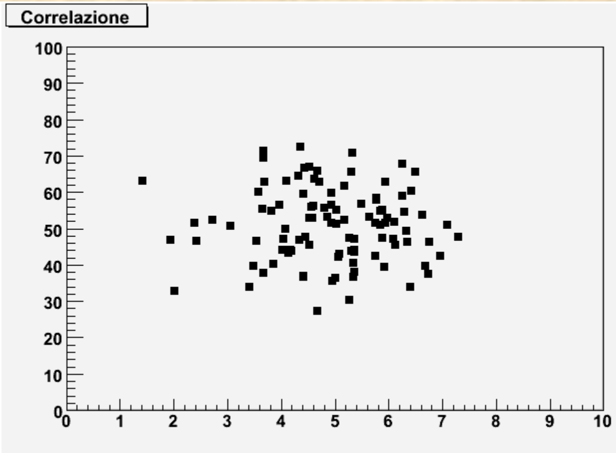
2.68, 42	5.20, 43	4.40, 53
2.96, 32	5.06, 56	4.34, 68
6.77, 45	4.89, 60	6.97, 53
7.00, 57	0.61, 52	4.46, 37
3.98, 46	4.86, 39	6.97, 67
2.10, 49	4.83, 81	5.58, 46
6.01, 62	4.45, 45	5.25, 51
5.47, 81	5.52, 68	4.78, 56
5.83, 44	3.05, 37	6.15, 37
9.26, 61	4.35, 32	1.91, 38

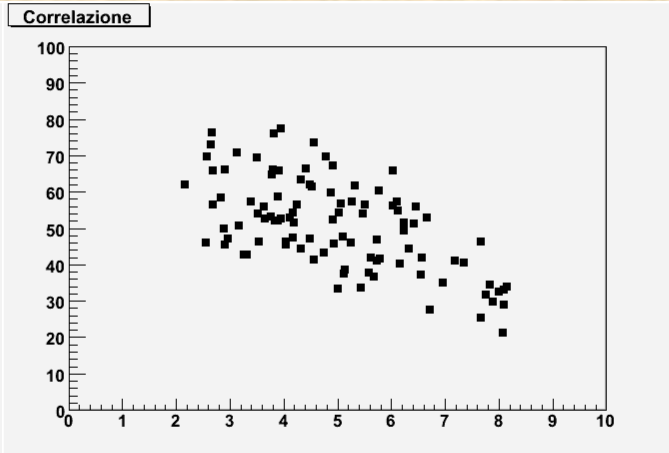
Grafico di correlazione – Scatter Plot



Covarianza e coefficiente di correlazione - I



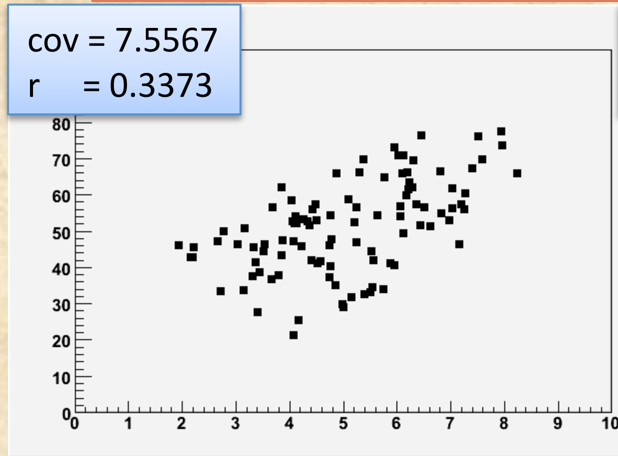


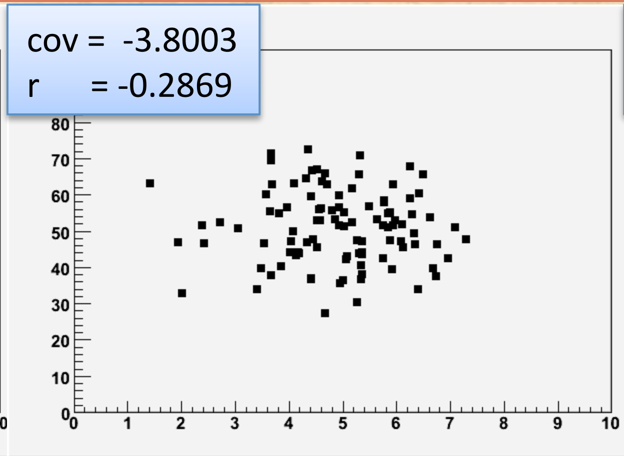


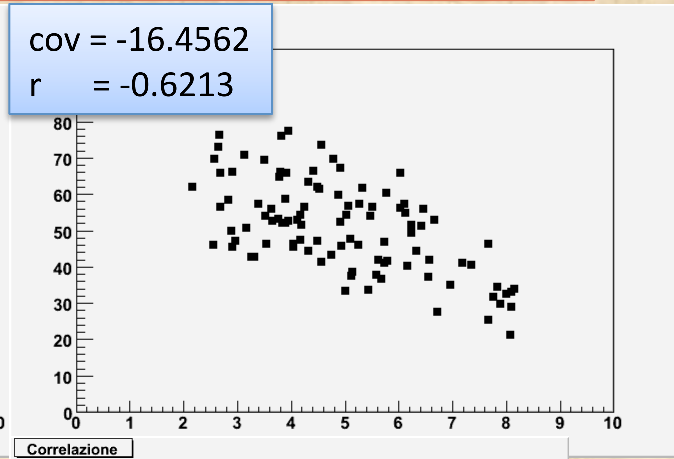
Occorre trovare un metodo per "quantificare" il grado di correlazione
Chiamiamo x_k gli N valori della grandezza e y_k gli N valori delle età.
(1) Covarianza campionaria **cov(*x*,*y*)**
(2) Coefficiente di correlazione campionario ***r*(*x*,*y*)**

$$\mathrm{cov}(x,y) = \frac{\sum_{k=1}^{N}(x_k - \bar{x})(y_k - \bar{y})}{N-1} \qquad r(x,y) = \frac{\sum_{k=1}^{N}(x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{N}(x_k - \bar{x})^2 \sum_{k=1}^{N}(y_k - \bar{y})^2}}$$

Covarianza e coefficiente di correlazione - II

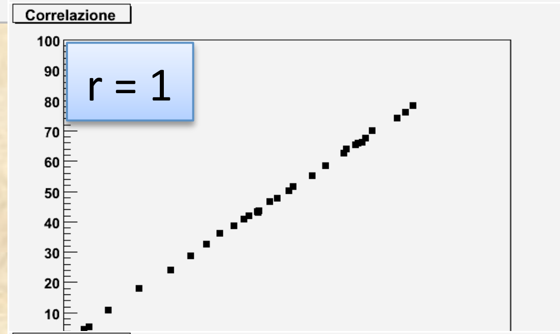


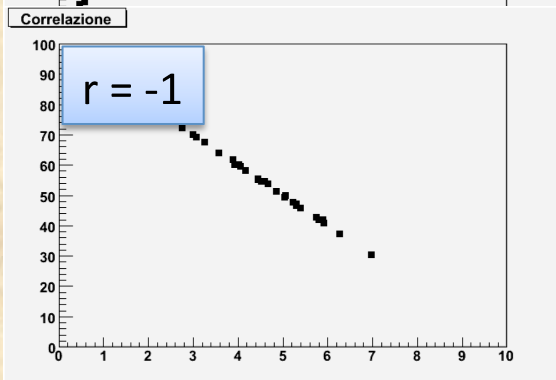




Si dimostra facilmente che **-1 < *r* < 1**
***r* = -1** le variabili sono totalmente anti-correlate
***r* = 0** le variabili sono scorrelate (o non correlate)
***r* = 1** le variabili sono totalmente correlate

Il coefficiente di correlazione è una variabile "comoda".
Nel seguito vedremo come si fa a fare il
Test di Correlazione





Esercizi - I

- Dopo anni di esperienza è nota che la distribuzione della concentrazione di rame nel sangue umano è ben descritta da una distribuzione di Gauss di parametri:

  $\mu = 3.2 \times 10^{-5}\ \mathrm{cm}^{-3}$;

  $\sigma = 0.2 \times 10^{-5}\ \mathrm{cm}^{-3}$.

  All'ultimo esame del sangue trovo:

  $X = 3.5 \times 10^{-5}\ \mathrm{cm}^{-3}$.

  Devo preoccuparmi?

  Se invece trovo

  $X = 1.9 \times 10^{-5}\ \mathrm{cm}^{-3}$?

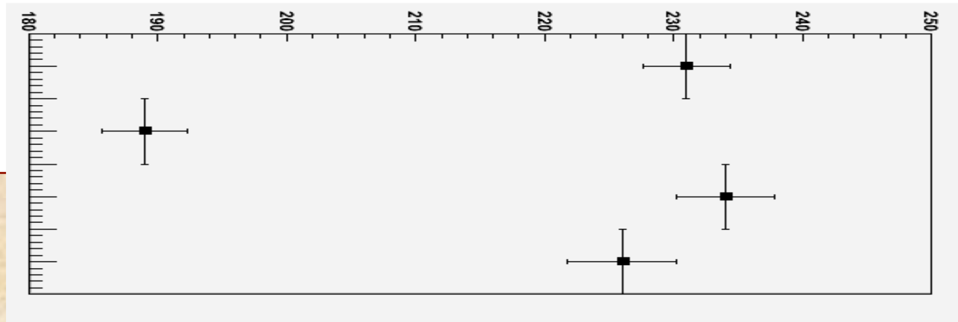
- Un medico mi spiega che l'intervallo di accettabilità del valore del colesterolo tra 150 e 220 mg/dl corrisponde ad un intervallo di probabilità del 90% calcolato su una popolazione gaussiana. Determinare μ e σ di tale distribuzione.

Esercizi - II

- Esempio –(2) della prima lezione:
  - ($N_t = 100$, $N_c = 100$)

$$\overline{X}^{t}_{ini} = 231 \qquad s(X^{t}_{ini}) = 34$$
$$\overline{X}^{t}_{fin} = 189 \qquad s(X^{t}_{fin}) = 33$$
$$\overline{X}^{c}_{ini} = 234 \qquad s(X^{c}_{ini}) = 38$$
$$\overline{X}^{c}_{fin} = 226 \qquad s(X^{c}_{fin}) = 42$$

  - Cosa concludiamo ?



Il caso delle variabili discrete

- Torniamo all'esempio-(1): conto n_t guarigioni su N_t pazienti trattati. Cosa posso dire della "**frazione di pazienti guariti**" ? o della "**probabilità che un paziente guarisca**" ?
- Faccio semplicemente $f_t = n_t \,/\, N_t$?
- Ma devo anche dire $f_t \pm$ quanto? per potermi confrontare per esempio con f_c
- (uso il termine probabilità in modo "intuitivo": probabilità ≈ frazione)
- Devo avere un ***modello statistico*** che mi permetta di conoscere come mi aspetto le distribuzioni di n_t e di n_c.

Il processo di Bernoulli - I

- Considero un fenomeno casuale che può avere luogo secondo due modalità:
  - successo   prob(successo) = $\boldsymbol{p}$
  - insuccesso   prob(insuccesso) = $\boldsymbol{1-p}$
- Facciamo $\boldsymbol{N}$ prove: quanti successi mi aspetto ?
- Esempi:
  - test o croce: $\boldsymbol{p}$=1/2, $\boldsymbol{N}$=numero di lanci
  - lancio del dado: $\boldsymbol{p}$(i)=1/6, $\boldsymbol{N}$=numero di lanci
- Cosa ci aspettiamo ?

Il processo di Bernoulli - II

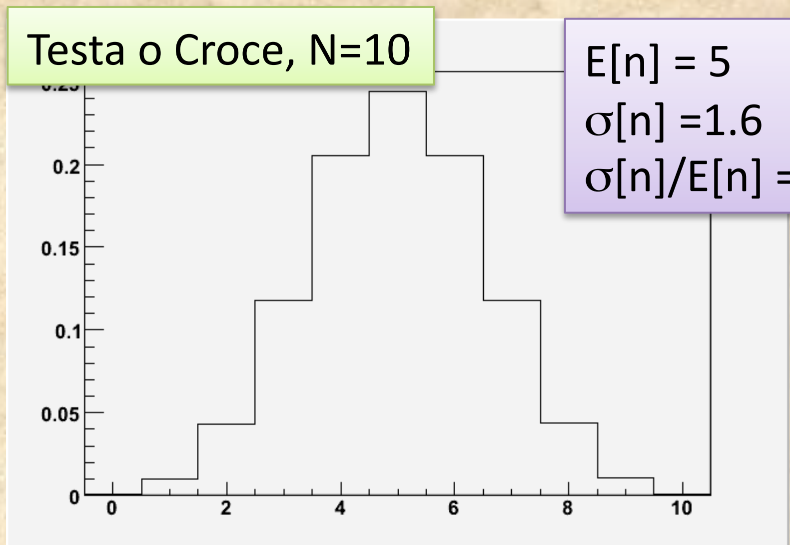
- Distribuzione binomiale: probabilità ***p(n)*** di ottenere ***n*** successi in ***N*** prove se la probabilità del successo è ***p***.

$$p(n) = \binom{N}{n} p^n (1-p)^{N-n}$$

- Applichiamola al caso di testa o croce:
  - ***p***=0.5 ***N***=10
  - ***p***=0.5 ***N***=100
- E al caso del lancio del dado (***p***(1))
  - ***p***=1/6 ***N***=10
  - ***p***=1/6 ***N***=100
- In ogni caso:
  - "Valor medio della distribuzione" ***E[n] = Np***
  - "Deviazione standard della distribuzione" ***σ[n] = Np(1-p)***
- Precisazione: ***E[n]*** vs. $\overline{n}$; ***σ[n]*** vs. ***s(n)***

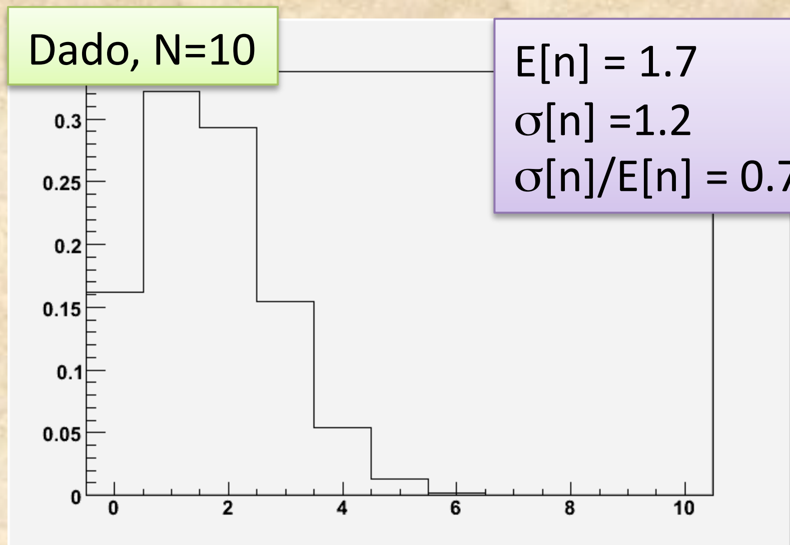
Il processo di Bernoulli - III

Testa o Croce, N=10



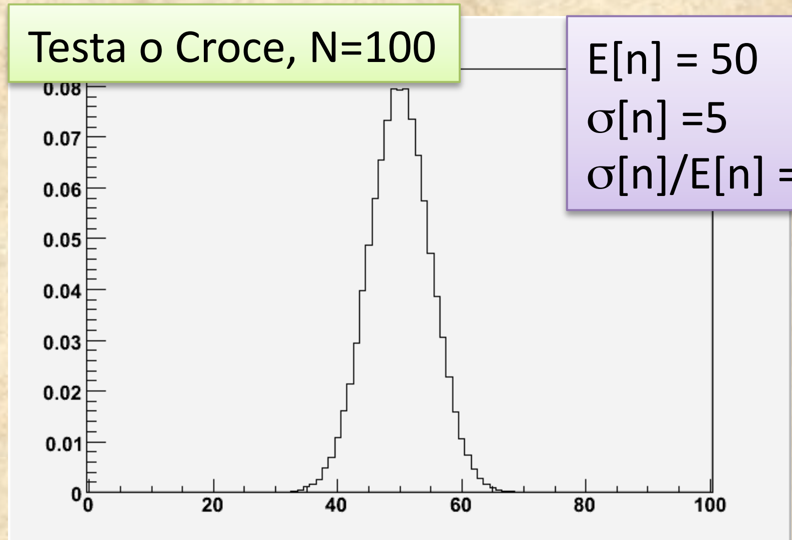
E[n] = 5
σ[n] =1.6
σ[n]/E[n] = 0.32

Dado, N=10



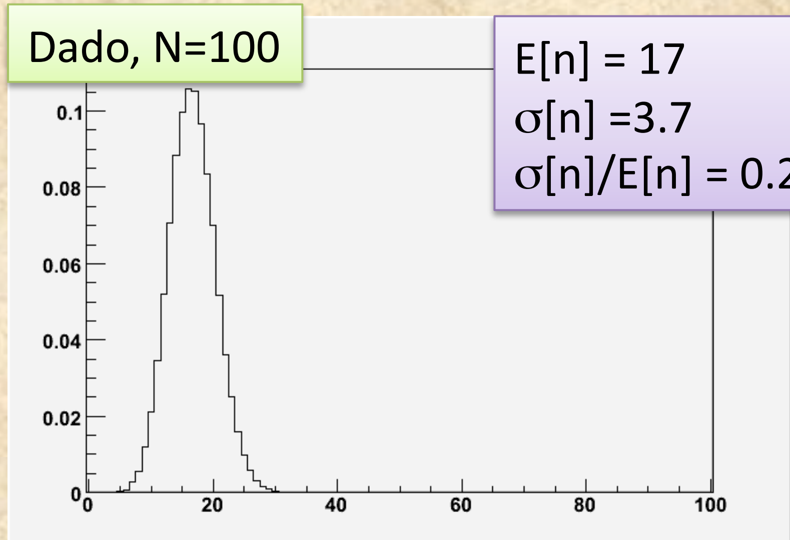
E[n] = 1.7
σ[n] =1.2
σ[n]/E[n] = 0.70

Testa o Croce, N=100



E[n] = 50
σ[n] =5
σ[n]/E[n] = 0.1

Dado, N=100



E[n] = 17
σ[n] =3.7
σ[n]/E[n] = 0.22

Il processo di Bernoulli - IV

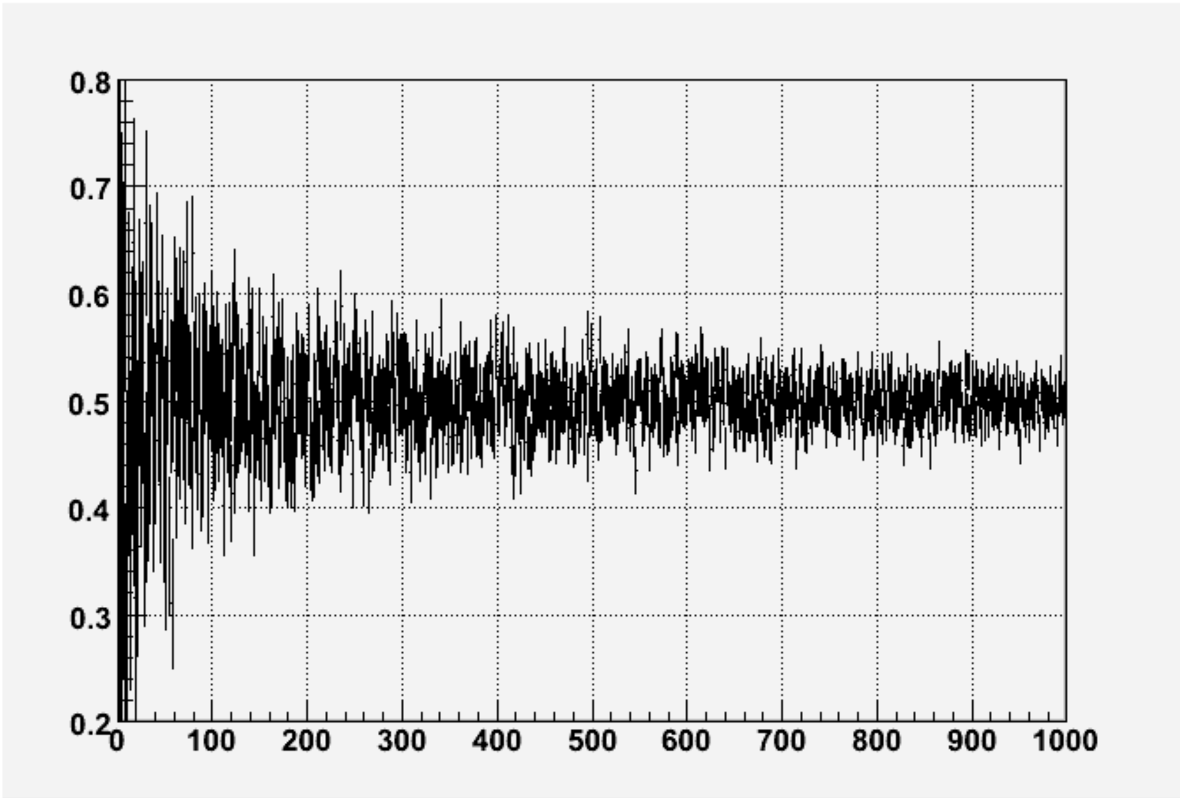
- Osservazioni:
  - Al crescere di *N*, crescono il valor medio e la deviazione standard, ma il rapporto σ/E diminuisce (restringimento relativo)
  - Al crescere di *N*, la distribuzione tende ad assomigliare ad una gaussiana
- Applico il modello binomiale per stimare l'incertezza sulla mia frazione di guarigioni *f*:
  - identifico *f* con *p* del successo;
  - identifico *n* con *E[n]*;
  - $E[n] = n = Np \;\Rightarrow\; f = p = n/N$
  - Valuto l'incertezza $\sigma[f] = \sigma[n]/N$
  - da cui:

$$f = \frac{n}{N} \pm \sqrt{\frac{n}{N^2}\left(1 - \frac{n}{N}\right)}$$

  - Al crescere di N la mia conoscenza su *f* migliora "come 1/√N"

Il processo di Bernoulli - V

Stima della probabilità di fare T (o C) al crescere del numero di lanci.



Il processo di Bernoulli – un esempio

Supponiamo allora che ho fatto un esperimento su un campione di N_t = 118 pazienti, e ho contato n_t=24 guarigioni. Qual è la migliore stima della probabilità di guarigione con la sua incertezza ?

Applico la formula vista in precedenza:

$$f = \frac{n}{N} \pm \sqrt{\frac{n}{N^2}\left(1 - \frac{n}{N}\right)} = 0.2034 \pm 0.0037 = (20.3 \pm 3.7)\%$$

Che significato ha questo risultato ? Avendo visto che per N grandi la binomiale → gaussiana,

$\mu \pm 1\sigma$ → $(20.3 \pm 3.7)\%$ → 68% probabilità;

$\mu \pm 2\sigma$ → $(20.3 \pm 7.4)\%$ → 95% probabilità;

$\mu \pm 3\sigma$ → $(20 \pm 11)\%$ → 99.7% probabilità.

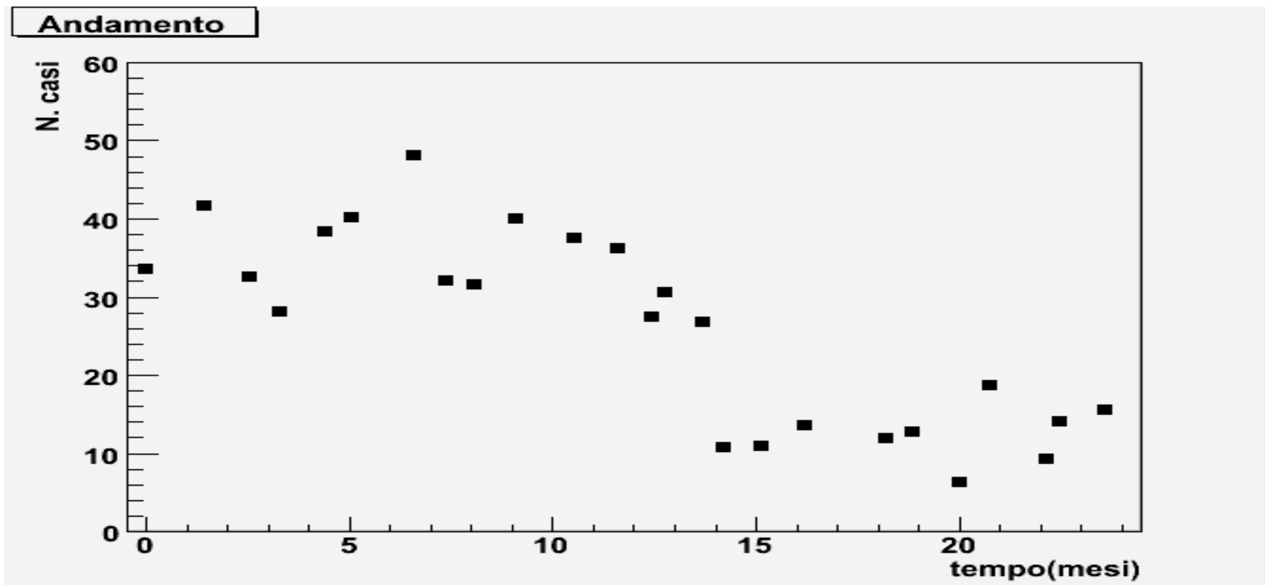
CAVEAT -1: solo per N "grandi" queste probabilità sono valide rigorosamente !

CAVEAT-2: tutto valido trascurando effetti "sistematici"…

Il processo di Poisson - I

- Non tutti i fenomeni di conteggio sono di tipo binomiale:
  - incidenza di una malattia su un intera popolazione;
  - contagi per unità di tempo
- Esempio (variante sull'Esempio-(1)):
  - ad un certo tempo ho intrapreso a trattare la popolazione con un certo farmaco: prima si avevano 34 casi di malattie al mese, dopo alcuni mesi ne misuro in media 13.

    Il trattamento ha avuto effetto ?



Il processo di Posson - II

- Il processo di Poisson è il modello statistico che descrive il conteggio di "eventi casuali".
- Stiamo contando eventi nel tempo. Dividiamo l'asse dei tempi in intervalli δt tali che:
  - in ogni intervallo $p(2) << p(1) << p(0)$
  - in ogni intervallo $p(1) = \alpha\, \delta t$
  - p(1) in un intervallo NON dipende da p(1) nell'intervallo precedente o successivo
  - ➔ Eventi del tutto casuali, non periodici, non "a bunch"
- Se valgono queste condizioni, il numero *n* di conteggi nell'intervallo generico Δt segue la statistica di Poisson.

Il processo di Posson - III

- n = variabile poissoniana:
  - $0 \leq n \leq \infty$;
  - detto λ il "conteggio medio" si ha la distribuzione di Poisson:

$$p(n) = \frac{\lambda^n e^{-\lambda}}{n!}$$

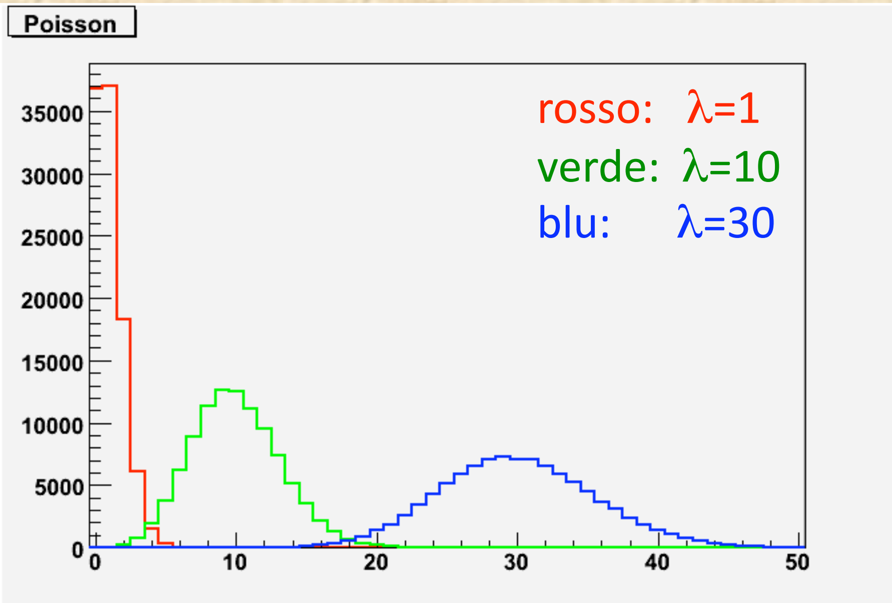
  - $E[n] = \lambda$
  - $\sigma[n] = \sqrt{\lambda}$
- Dunque se conto n, quanto sono incerto su n ?
  - identifico n con λ e dico $\lambda = n \pm \sqrt{n}$
  - passo al "rate" facendo:

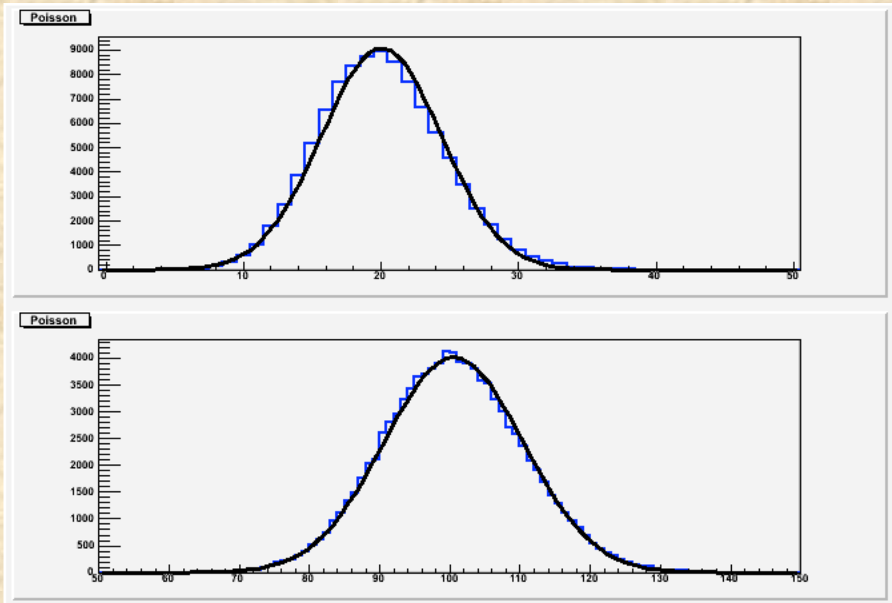
$$r = \frac{n}{\Delta t} \pm \frac{\sqrt{n}}{\Delta t}$$

  - quindi anche in questo caso al crescere di n, la mia conoscenza su λ migliora come $1/\sqrt{n}$

Il processo di Posson - IV

Come è fatta la distribuzione di Poisson:
- asimmetrica per $\lambda < 20 \div 30$
- a campana e tendente ad una distribuzione di Gauss per λ più grandi.





Il processo di Posson – un esempio

- Supponiamo di aver diagnosticato nello scorso mese n=56 nuovi casi della malattia K in tutta la nazione. Quanto vale il “rate” di quella malattia nell’ultimo mese ? E quanto sono incerto su di esso ?
- Applichiamo la formula del “rate”, esprimendolo in “nuovi casi/giorno”

$$r = \frac{n}{\Delta t} \pm \frac{\sqrt{n}}{\Delta t} = (1.87 \pm 0.25) d^{-1}$$

- Significato probabilistico: stesse considerazioni di prima per il limite gaussiano.
- CAVEAT: tutto valido solo nel limite in cui le ipotesi del processo poissoniano sono valide ! (applichiamo al caso del battito cardiaco…)

Ricapitolando.

Abbiamo esaminato due casi:

- frazione di "guarigioni";
- "rate" di malattie

Nel primo caso abbiamo considerato il modello binomiale:

$$f = \frac{n}{N} \pm \sqrt{\frac{n}{N^2}\left(1 - \frac{n}{N}\right)}$$

Nel secondo caso il modello di Poisson

$$r = \frac{n}{\Delta t} \pm \frac{\sqrt{n}}{\Delta t}$$

Caratterstica comune, al crescere di N o di n:

(1) la mia incertezza relativa diminuisce;

(2) ottengo intervalli di probabilità gaussiani.

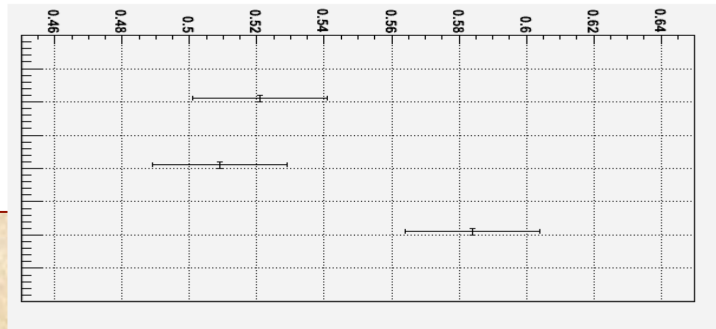
Esercizi - I

- L'incidenza alla nascita della sindrome genetica X è dello 0.12% per età della madre inferiore ai 30 anni e del 0.28% per età superiore a 30 anni. La signora Y ha 10 figli di cui 3 avuti prima dei 30 anni e 7 dopo i 30 anni. Quant'è la probabilità che nessuno sia affetto da sindrome X?

- In Italia la probabilità di superare i 70 anni di vita è pari al 78.2%. Quant'è la probabilità che almeno 4 dei miei 5 figli superino i 70 anni ?
- La radioattività ambientale standard, misurata con un certo contatore è caratterizzata da un valor medio $r=1.8 \times 10^{-3}$ conteggi al secondo. Metto quel contatore a casa mia e lo lascio contare per un giorno intero. Se ottengo N=404, devo preoccuparmi ?

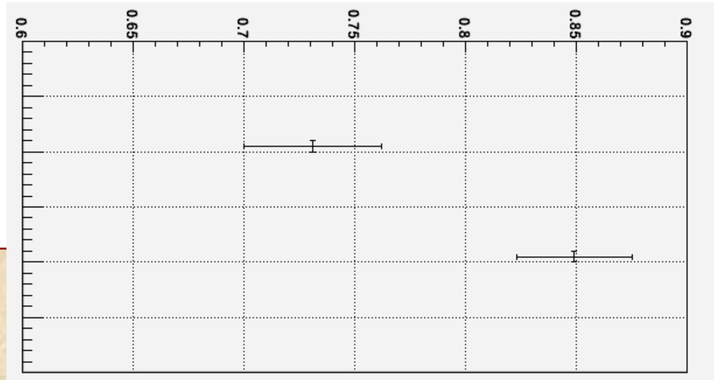
Esercizi - II

- Esempio –(1) della prima lezione (basato su un lavoro pubblicato (*Am Heart J 2006;151:934-42*)). N = numero di pazienti del campione; n = numero di pazienti che dopo un tempo fissato hanno “complicazioni cardiache”
  - Gruppo 1: N = 604, n=315;
  - Gruppo 2: N = 597, n=304;
  - Gruppo 3: N = 601, n=351.
- Il lavoro si conclude con due affermazioni:
  - (1): Gruppo 1 = Gruppo 2
  - (2): Gruppo 3 > Gruppo 1
- Calcoliamo le frazioni di “complicazioni cardiache”:
  - $f_1 = (52.1 \pm 2.0)\%$
  - $f_2 = (50.9 \pm 2.0)\%$
  - $f_3 = (58.4 \pm 2.0)\%$



Esercizi - III

- Il lavoro era in risposta ad un precedente studio (*South Med J 1988;81:826 -9.*) molto simile che considerava solo i gruppi 1 e 2:
  - Gruppo 1: N = 192, n=27;
  - Gruppo 2: N = 201, n=44;
- Lavoro che si concludeva con l'affermazione:
  - Gruppo 1 < Gruppo 2
- Calcoliamo le frazioni nei due gruppi:
  - $f_1 = (73.1 \pm 3.1)\%$
  - $f_2 = (84.9 \pm 2.6)\%$



Alcune considerazioni

Nel caso di fenomeni di conteggio, abbiamo visto che utilizzando i modelli di Bernoulli e Poisson è possibile fare delle affermazioni "probabilistiche" sulle "*frazioni*" e sui "*rate*", sotto forma di:

Valore centrale ± barra di incertezza.

(1) Si tratta di affermazioni probabilistiche;

(2) Il contenuto di probabilità è rigorosamente corretto solo nel limite di alti N o n;

(3) Le affermazioni valgono solo nel limite di validità dei modelli;

(4) Stiamo assumendo che i campioni siano stati scelti in modo perfettamente coerente. Se questo non è, mi aspetto che le incertezze si allarghino.

Nell'ultima lezione vedremo come confrontare in modo quantitativo diverse frazioni o diversi rate.

Ricapitolando.

Abbiamo visto che la conclusione di una ricerca in Medicina si traduce in un "Numero Magico" **p**, che per qualche motivo cambia il risultato a seconda se è > o < di 0.05.

Abbiamo visto che in generale dunque le conclusioni sono di natura probabilistica e non deterministica.

➔ si impone lo studio della probabilità anche a livello concettuale per dare significato alle nostre conclusioni.

Finora abbiamo fatto uso di una nozione di probabilità intuitiva. Abbiamo introdotto la nozione di *intervallo di confidenza*, nella cui definizione la probabilità è essenzialmente l'espressione quantitativa di quanto crediamo che il valor vero della misura cada nell'intervallo che abbiamo fornito (68%, 95%, 99.7%...).

Sappiamo inoltre che la probabilità è in qualche modo legata alla "frequenza" con cui un fenomeno si presenta.

Intermezzo semi-filosofico: che cos'è la Probabilità ?

Primo dibattito: la definizione.

Due classi di definizioni che corrispondono a due diversi approcci:

(1) Probabilità ***ontologica***: la probabilità è una proprietà delle cose, è "scritta nella natura", è legata all'oggetto che intendo conoscere;

(2) Probabilità ***epistemica***: la probabilità è legata al soggetto che conosce, rappresenta lo stato della mia conoscenza.

La cosiddetta definizione assiomatica (Kolmogorov), valida indipendentemente dall'interpretazione, fonda la matematica della probabilità.

Metodi di calcolo - I

- **Definizione "combinatoria"**:

  N_p = Numero di casi possibili (equiprobabili)

  N_f = Numero di casi favorevoli

  ➔ $\mathbf{p = N_f / N_p}$

  - Esempio: lancio di due monetine, estrazione di una carta da un mazzo…;
  - Problema di principio nella definizione;
  - Non tutti i problemi di probabilità sono riconducibili a questo schema.
- **Definizione "frequentista":**
  - su n prove ho misurato una frequenza del successo pari a f_n:
  - ➔ $\mathbf{p = \lim f_n}$ **(per $n \rightarrow \infty$)**
  - Esempio: imparare dall'esperienza
  - Funziona: ***Principio di stabilità della frequenza*** o ***Legge empirica del caso***.
  - Non tutti i problemi di probabilità sono riconducibili a questo schema

Metodi di calcolo - II

- **“Degree of belief”**: ***quanto io (soggetto) credo al verificarsi di un dato evento sulla base delle conoscenze che ho in questo momento.***
- Cosa facciamo quando usiamo la probabilità ? Selezioniamo cose vere, cose false, e cose sul cui contenuto di verità siamo incerti. Ma le cose incerte non sono completamente ignote, diciamo che c’è una “scala” naturale che ha per estremi da una parte la totale certezza della verità e dall’altra la totale certezza della falsità. La probabilità “quantifica” questa scala. Appare naturale introdurre una scala tra 0 e 1 o anche, nel linguaggio comune tra 0 e 100%.
- Formalizzabile mediante la **“Scommessa Coerente”**: punto una quantità X di denaro e accetto di ottenerne in caso di successo una quantità Y: ➔ attribuisco all’evento una probabilità:

$$p = \frac{X}{Y}$$

- p si chiama anche “Speranza Matematica”.

Il linguaggio della probabilità - I

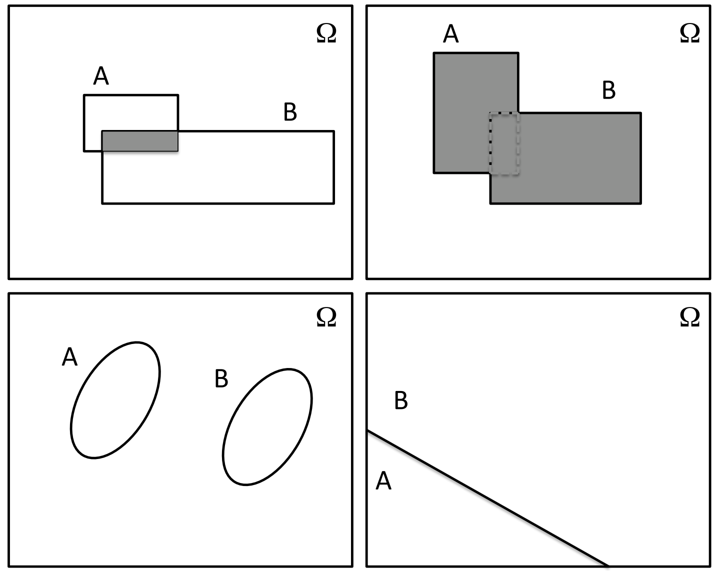
- Evento A nello spazio degli eventi Ω
- Eventi Composti:
  - Somma logica (OR): $A \cup B$
  - Prodotto logico (AND): $A \cap B$
- Eventi incompatibili:

$$A \cap B = \emptyset$$

- Eventi opposti:

$$A \cap B = \emptyset$$

$$A \cup B = \Omega$$



Il linguaggio della probabilità - II

- P(A) = probabilità dell'evento A

  Assioma della positività: $0 \leq P(A) \leq 1$

  Assioma della certezza: $P(\Omega) = 1$ $\quad P(\emptyset) = 0$

  Assioma dell'unione: $P(A \cup B) = P(A) + P(B)$

  se A e B sono incompatibili.

- A e B compatibili, si ha il Teorema della Probabiltà Totale:

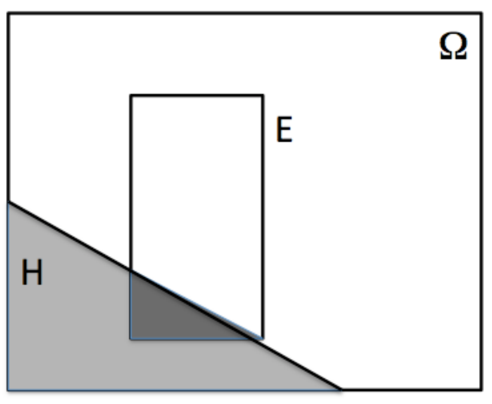
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- P(A/H) = probabilità dell'evento A data la condizione H; probabilità condizionata

  Teorema delle Probabiltà composte:

$$P(E \cap H) = P(E/H)P(H)$$

$$P(E \cap H) = P(H/E)P(E)$$



Esercizi - I

- Un’urna contiene 5 palline numerate da 1 a 5. Estraendole una ad una senza reintrodurle ogni volta, quant’è la probabilità di ottenere la sequenza 1-2-3-4-5 ? E se ogni volta reintroduco la pallina estratta quanto vale la probabilità di ottenere la stessa sequenza ? Infine quanto cambiano le 2 probabilità se anziché cercare la sequenza 1-2-3-4-5 cerco la sequenza 2-5-4-1-3 ?

- Quant’è la probabiltà che estraendo da un mazzo di 40 carte ottenga un 7 o una casta di denari ? E che ottenga invece il 7 di denari ?

Il Teorema di Bayes - I

- *Supponiamo che il fenomeno che sto studiando possa aver luogo secondo N modalità diverse (gli N eventi* $\mathbf{A_i}$*). Supponiamo inoltre di aver osservato l'evento* **B** *e di essere interessati a sapere quale o quali tra le N modalità* $\mathbf{A_i}$ *siano plausibilmente le "cause" di* **B***. Detto in termini più espliciti, se l'effetto* **B** *può essere dovuto a N cause diverse* $\mathbf{A_i}$*, voglio, dall'osservazione dell'effetto, stabilire la probabilità di ciascuna causa.*
- *Questo processo si chiama INFERENZA statistica: dall'osservazione di un effetto voglio dire qualcosa circa le sue possibili cause*

Il Teorema di Bayes - II

- Dunque, le A_k sono le N possibili cause e B è l'effetto. La probabilità della k-esima causa dato l'effetto B è:

$$P(A_k / B) = \frac{P(B / A_k) P(A_k)}{\sum_{i=1}^{N} P(B / A_i) P(A_i)}$$

- P(B/A_k) sono le "likelihood", che mi dicono quanto è verosimile che la causa A_k dia l'effetto B; dipendono dal mio esperimento.
- P(A_k) sono le "prior", è tutto quello che so prima del mio esperimento, sono le "probabilità a priori".

Esercizi - II

- Un test del virus HIV è caratterizzato dalle seguenti prestazioni:

  p(+/infetto)=99.0% ,

  p(-/non infetto)=99.3% .

  Calcolare quant'è la probabilità che facendo un test e risultando positivo una persona sia effettivamente infetta in 2 casi:

  (1) il test è fatto su tutta la popolazione nazionale italiana (per la quale il ministero della salute stima una frazione di infetti dello 0.2% circa);

  (2) il test è fatto solo su un campione "a rischio" in cui ci si aspetta che circa la metà delle persone testate sia infetta.

- Dai dati dell'esercizio precedente (caso(1)): Quanto cambia la probabilità di essere infetto se ripeto 3 volte il test e per tre volte la persona risulta positiva ?

Inferenza Statistica

- Dall'osservazione sul mio campione (intrinsecamente limitato) passo ad un intervallo di probabilità sul "valore vero" di una grandezza.
  - Inferenza frequentista: costruisco un intervallo di confidenza di probablità α (e.g. α=95%) tale che se ripeto N volte la misura su N campioni indipendenti, nel 95% dei casi l'intervallo costruito "contiene" il valore vero.
  - Inferenza bayesiana: attraverso il teorema di Bayes, considero il valore vero come una variabile causale e calcolo la sua distribuzione di probabilità da cui ricavo l'intervallo di confidenza voluto.
- Quindi: intervallo di confidenza del 95% significa che io, sulla base delle mie conoscenze, dico che il valor vero è contenuto la dentro al 95%, la probabilità che sia fuori è del 5%. Se ripetesse l'esperimento 100 volte, 5 volte l'intervallo da me costruito non conterrebbe il valor vero.

Test di Ipotesi - I

- Torniamo agli esempi della prima lezione:
  - Campione trattato (n_t / N_t oppure $X_{ini}^t(k)$, $X_{fin}^t(k)$);
  - Campione placebo (n_c / N_c oppure $X_{ini}^c(k)$, $X_{fin}^c(k)$);
  - Ipotesi da testare.
- Dalla combinazione di questi numeri e dalle caratteristiche dell'ipotesi sotto test, voglio ricavare il valore magico P (p-value) sulla base del quale trarre la mia conclusione.
- Test di Ipotesi (o di significatività):
  - costruisco una statistica campionaria ζ funzione del campione e dell'ipotesi di distribuzione di probabilità nota $p(\zeta)$;
  - calcolo il valore di ζ sulla base del mio campione ζ_{exp};
  - calcolo $p(\zeta>\zeta_{exp})$

Test di Ipotesi - II

(1) Confronto tra le medie di due campioni

- Caso gaussiano;
- caso t-Student;

(2) Test di correlazione tra due grandezze

(3) Confronto tra "frazioni" di guarigioni

- Caso gaussiano;
- Relative Risk (RR) o Hazard Ratio (HR);
- Odds Ratio (OR);
- Test "esatto" di Fisher.

(4) F-test, confronto tra varianze.

Confronto tra le medie di due campioni - I

Supponiamo di aver analizzato ciascuno dei 4 campioni della variabile X e di avere ottenuto intervalli di confidenza al 68%. Ci riferiamo per semplicità ad uno dei due campioni (omettiamo t o c)

$$\overline{X}_{ini} \pm \frac{s(X_{ini})}{\sqrt{N}} \qquad \overline{X}_{fin} \pm \frac{s(X_{fin})}{\sqrt{N}}$$

L'ipotesi che voglio sottoporre a test è che non ci sia stata variazione (null hypothesys).

Costruisco la statistica campionaria:

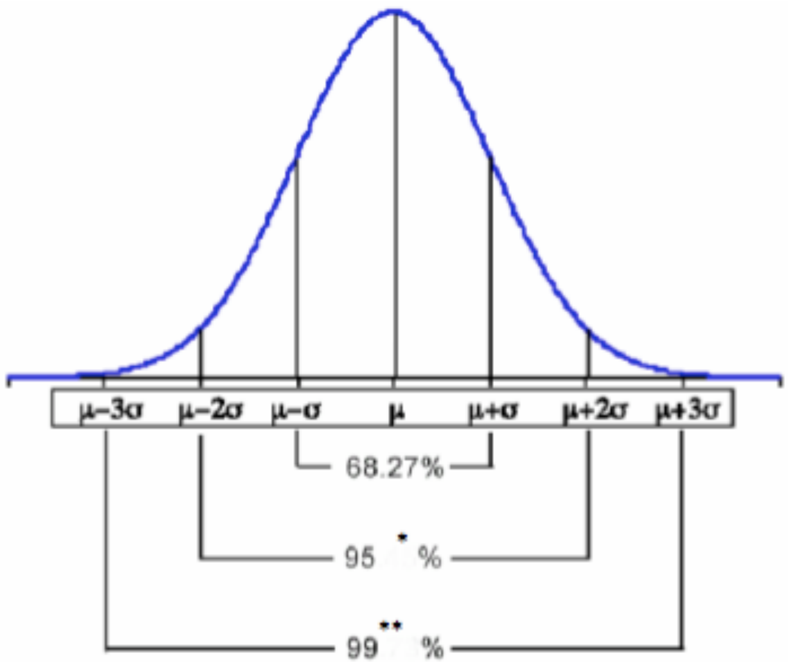
$$\zeta = \frac{\overline{X}_{fin} - \overline{X}_{ini}}{\sqrt{\frac{s^2\left(X_{fin}\right)}{N} + \frac{s^2\left(X_{ini}\right)}{N}}}$$

numeratore = differenza tra le medie;

denominatore = combinazione delle due incertezze.

Confronto tra le medie di due campioni - II

Caso (1): N "grande", siamo nel limite gaussiano. Se l'ipotesi è corretta, ζ è una variabile *gaussiana standardizzata*. Dunque la p(ζ) è



Confronto tra le medie di due campioni - III

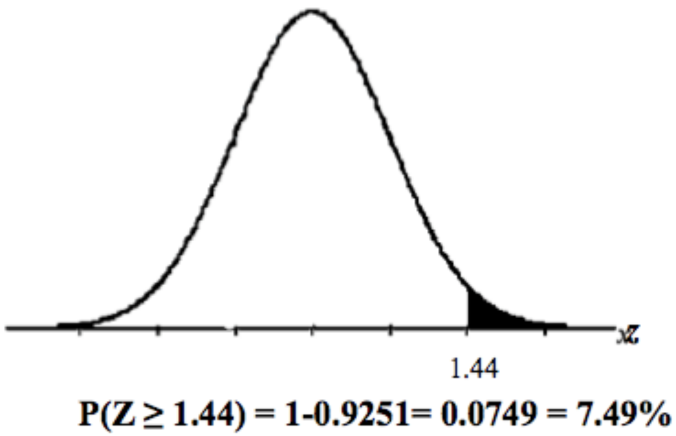
Due esempi:

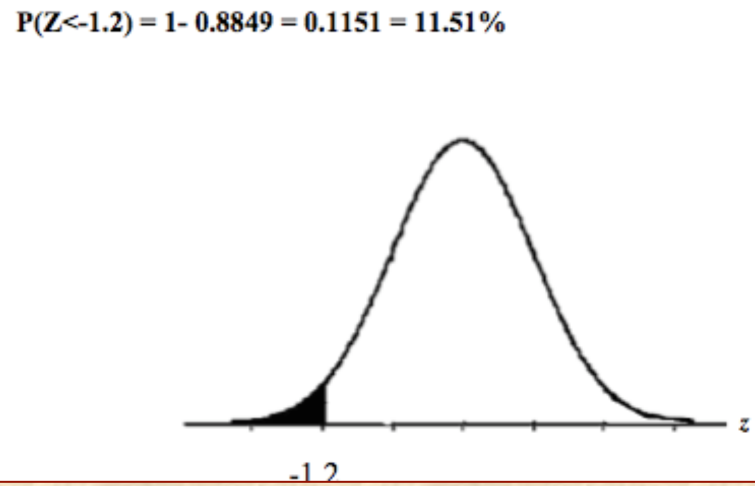
Nel primo ottengo ζ_{exp} = 1.44. Accetto l'ipotesi ?

Calcolo $p(\zeta>\zeta_{exp})$ e ottengo (usando le tabelle) 7.5%. Se avevo fissato il limite al 5% devo dire che l'ipotesi nulla è accettata.

Nel secondo ottengo ζ_{exp} = -1.2. Accetto l'ipotesi ?

Lo stesso procedimento mi da: $p(\zeta>\zeta_{exp})$ = 88.49%. Di nuovo, devo accettare l'ipotesi (sono nella coda a sinistra)





Confronto tra le medie di due campioni - IV

Caso (2): N non tanto "grande", non siamo nel limite gaussiano. Assumiamo però che $s(X_{ini})=s(X_{fin})=s$, per esempio faccio la media di $s(X_{ini})$ e $s(X_{fin})$.

ζ diventa:

$$\zeta = \frac{\overline{X}_{fin} - \overline{X}_{ini}}{s\sqrt{\frac{2}{N}}}$$

che è distribuita come una t-Student con $2N-2$ gradi di libertà.

(ugualmente se i due campioni avessero un diverso numero di pazienti N_1 ed N_2 sarebbe comunque una t-Student

$$\zeta = \frac{\overline{X}_{fin} - \overline{X}_{ini}}{s\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

con N_1+N_2-2 gradi di libertà)

Confronto tra le medie di due campioni - V

Test t-Student.

Calcolo il numero di gradi di libertà e il valore ζ_{exp}

Fisso il livello di confidenza (e.g. 95%)

Vedo dove si colloca ζ_{exp}

Esempio:

ngl=6;

ζ_{exp}=-3.31;

Conclusione ?

Ipotesi rigettata.

N.B. Il caso ngl=∞ corrisponde al limite gaussiano.

	α=68.3%	α=90%	α=95%	α=99%	α=99.73%
N-1	t	t	t	t	t
2	1.84	6.31	12.7	63.7	236
3	1.32	2.96	4.30	9.93	19.2
4	1.20	2.35	3.18	5.84	9.22
5	1.15	2.13	2.78	4.60	6.62
6	1.11	2.02	2.57	4.03	5.51
8	1.08	1.90	2.37	3.50	4.53
10	1.06	1.83	2.26	3.25	4.09
13	1.05	1.78	2.18	3.05	3.76
20	1.03	1.73	2.09	2.86	3.45
30	1.02	1.70	2.05	2.76	3.28
50	1.01	1.68	2.01	2.68	3.16
100	1.00	1.66	1.98	2.63	3.08
∞	1.00	1.65	1.96	2.58	3.00

Test di correlazione tra due grandezze

Supponiamo di aver effettuato una sequenza di *N* coppie di misure delle due grandezze x_1 e x_2 e di aver determinato per via sperimentale il valore di r (coefficiente di correlazione campionario).

Consideriamo la seguente ipotesi: le due grandezze non sono correlate (null hypothesys).

Si dimostra che la grandezza:

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

è distribuita come una *t* di Student con *N*-2 gradi di libertà.

Considero la tabella dei limiti su r al 95% per N tra 3 e 102:

N-2	1	5	10	15	20	25	30	50	100
r	0.97	0.75	0.58	0.48	0.42	0.38	0.35	0.27	0.19

Esempio. Trovo r = 0.58: cosa concludo ? Se N>12 c'è correlazione, se < no.

Confronto tra "frazioni": la tabellina 2×2

Test di ipotesi sull'esperimento dei pazienti guariti e non.
Null Hypothesys: i 2 campioni sono "uguali"
➔il trattamento non ha nessun effetto.
La Tabellina 2×2 è un modo utile per mostrare il risultato.

	"guariti"	"non guariti"	pazienti
campione trattato	n_t	N_t-n_t	N_t
campione di controllo	n_c	N_c-n_c	N_c
totale	n=n_t+n_c	N-n	N

Dalla Tabellina si ricavano direttamente le frazioni f_t ed f_c.
Facciamo prima l'ipotesi che la distribuzione binomiale sia nel limite gaussiano (*N* elevati)

Caso gaussiano.

Nel limite gaussiano costruisco una ζ gaussiana:

$$\zeta = \frac{f_t - f_c}{\sqrt{\dfrac{f_t(1-f_t)}{N_t} + \dfrac{f_c(1-f_c)}{N_c}}}$$

Calculato ζ_{exp}, farò le stesse considerazioni che abbiamo visto nel caso delle medie:

Se $-1.96<\zeta_{exp}<1.96$ ➔ null hypothesys accettata

Se $\zeta_{exp}>1.96$ OR $\zeta_{exp}<-1.96$ ➔ null hypothesys rigettata

➔ ho fatto una scoperta !

Relative Risk (RR) - I

	"guariti"	"non guariti"	pazienti
campione trattato	n_t	N_t-n_t	N_t
campione di controllo	n_c	N_c-n_c	N_c
totale	$n=n_t+n_c$	N-n	N

Calcoliamo il doppio rapporto RR detto "rischio relativo":

$$RR = \frac{n_t}{N_t} \bigg/ \frac{n_c}{N_c} = \frac{f_t}{f_c}$$

Mi aspetto che valori vicini a 1 implicano assenza di effetto
Devo costruire un intervallo di confidenza al 95%: se contiene 1 vuol dire che al 95% non c'è alcun effetto.

Relative Risk (RR) - II

Quindi la mia ζ è in questo caso RR. Nel caso di alti valori di N, f(RR) è ben approssimata da una gaussiana.
Quindi considero una gaussiana di parametri:

$$\mu = RR$$

$$\sigma = \sqrt{\frac{RR}{f_c}\left(\frac{1-f_f}{N_t} + RR\frac{1-f_c}{N_c}\right)}$$

se il valore 1, è contenuto entro l'intervallo μ ± 2σ
➔Ipotesi accettata
In alcuni casi si usa anzichè RR il suo logaritmo. La distribuzione è sempre gaussiana con parametri:

$$\mu = \ln(RR)$$

$$\sigma = \sqrt{\frac{1-f_t}{N_t f_t} + \frac{1-f_c}{N_c f_c}}$$

Un esempio.

- **Findings: 10 096 patients were allocated to tranexamic acid and 10 115 to placebo, of whom 10 060 and 10 067, respectively, were analysed. All-cause mortality was signifi cantly reduced with tranexamic acid (1463 [14.5%] tranexamic acid group vs 1613 [16.0%] placebo group; relative risk 0.91, 95% CI 0.85–0.97; p=0.0035). The risk of death due to bleeding was signifi cantly reduced (489 [4.9%] vs 574 [5.7%]; relative risk 0.85, 95% CI 0.76–0.96; p=0.0077).**

Odds Ratio (OR)

	"guariti"	"non guariti"	pazienti
campione trattato	n_t	N_t-n_t	N_t
campione di controllo	n_c	N_c-n_c	N_c
totale	n=n_t+n_c	N-n	N

Calcoliamo il doppio rapporto OR detto "rischio relativo":

$$OR = \frac{\dfrac{n_t}{N_t - n_t}}{\dfrac{n_c}{N_c - n_c}}$$

Anche in questo caso mi aspetto valori vicini a 1 in caso di null hypothesys.

Devo costruire un intervallo di confidenza al 95%: se contiene 1 vuol dire che al 95% non c'è alcun effetto.

Test esatto di Fisher

	"guariti"	"non guariti"	pazienti
campione trattato	n_t	N_t-n_t	N_t
campione di controllo	n_c	N_c-n_c	N_c
totale	$n=n_t+n_c$	N-n	N

Fissando N_t, N_c ed n (cioè gli elementi "marginali" della tabellina) ho tante possibili tabelline, al variare, per esempio, di n_t. Fisher nel 1930 ha calcolato la probabilità che una tale tabellina sia consistente con la null hypothesys:

$$p = \frac{N_t!N_c!n!(N-n)!}{N!n_t!n_c!(N_t-n_t)!(N_c-n_c)!}$$

Si tratta di un risultato che vale anche per piccoli numeri !

Esempio

Ho a disposizione solo 16 casi della malattia che sto studiando. Ne scelgo 8 da sottoporre al trattamento e lascio gli altri 8 come campione di controllo. Al termine ho 1 decesso tra gli 8 trattati e 4 nel campione di controllo. Posso dire che il trattamento riduce la probabilità di decesso ?

	"deceduti"	"sopravvissuti"	pazienti
campione trattato	*1*	*7*	*8*
campione di controllo	*4*	*4*	*8*
totale	*5*	*11*	*16*

Posso costruire 6 tabelline.
Per ciascuna calcolo la P:

P(0-5) = 0.013
P(1-4) = 0.128
P(2-3) = 0.359
P(3-2) = 0.359
P(4-1) = 0.128
P(5-0) = 0.013

Cosa concludo ? Adottando una tecnica "simil-gaussiana" posso chiedermi: quant'è la probabilità di ottenere 1 o meno di 1 decesso ? Si ha:
$P(\leq 1) = P(0) + P(1) = 0.013+0.128 = 0.141$, il 14.1% ampiamente sopra il fatidico 0.05 ! Non posso dire che il trattamento funziona.

F-test

Vogliamo confrontare non le medie ma le ***deviazioni standard campionarie*** di due campioni. Per esempio ci chiediamo se a seguito di un trattamento è aumentata la variabilità di un certo indicatore di patologia o, più semplicemente se due metodi per misurare la stessa cosa hanno lo stesso grado di variabilità.

→ Campione (1): N_1 pazienti, dev.st. = $s_1(X)$

→ Campione (2): N_2 pazienti, dev.st. = $s_2(X)$

Se X è gaussiana si dimostra che posso definire una statistica campionaria ζ:

$$\zeta = \frac{s_1^2}{s_2^2}$$

caratterizzata da una f(ζ) nota: la cosiddetta variabile F.

Dalle tabelle della variabile F definiamo un intervallo di confidenza al 95% per testare la *null hypotheys*: i due campioni hanno la stessa dev.standard.

Cosa manca:

Abbiamo imparato a trattare il confronto tra due campioni: i test di ipotesi visti permettono:

- di stabilire se la null hypothesys è all'interno di un intervallo del 95% C.L;
- di calcolare il P-value.

Cosa succede se anziché avere 2 campioni (t e c) ne ho 3 o più ?

→ analisi di regressione (fit);

→ test del χ^2;

→ analisi della varianza (ANOVA)