Corso Integrato di Statistica Informatica e Analisi dei Dati Sperimentali

A.A 2009-2010

Esercitazione E

Scopo dell'esercitazione
Applicazioni del teorema del limite centrale. Rappresentazione delle incertezze di misura. Calcolo dell'intervallo di confidenza.

1 Riepilogo della teoria

1.1 Teorema del limite centrale

Sia X é una variabile aleatoria con valore atteso μ e deviazione standard σ. La $\bar{X}$ la variabile aleatoria $\bar{X}$=medie effettuate su campioni di numerositá N é ancora una variabile aleatoria e come tale ha una sua distribuzione di probabilitá $p(\bar{X})$ caratterizzata da un valore atteso $\mu_{\bar{x}}$ e da una varianza $\sigma^2_{\bar{x}}$.
Il teorema del limite centrale afferma che:

- il valore atteso della distribuzione delle media campionarie é eguale a quello della variabile X, ovvero: $\mu_{\bar{x}} = \mu$;
- per N sufficientemente grande la distribuzione delle medie campionarie approssima una distribuzione Gaussiana;
- la deviazione standard della distribuzione delle medie é:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$

1.2 Incertezza di misura

La deviazione standard della distribuzione di probabilitá di una variabile aleatoria rappresenta l'**incertezza** (o **errore standard**) sul valore della grandezza osservata. Nel riportare il valore di una grandezza, risultato di un osservazione diretta o indiretta, va sempre riportata l'incertezza standard. Nel riportare il valore di una grandezza, le cifre significative indicano la precisione con si conosce o si é misurato il valore. Il numero di cifre significative da usare nel riportare l'incertezza è sempre 2. Quindi se conosciamo $x = 12.346789$ cm e la sua incertezza é: $\sigma = 0.03427852634$ cm, si riporta il valore: $x = 12.347 \pm 0.0342$ cm. Se $x = 1245.346789$ m con $\sigma = 35.34278526$ m, si riporta: $x = 1245 \pm 35$ m.
Nell'arrotondare le cifre significative ricordarsi che: se la prima cifra trascurata é minore di 5 allora il numero si tronca, se la prima cifra trascurata é maggiore o uguale a 5 bisogna aggiungere 1 all'ultima cifra utile. Ad esempio se $x = 0.12563$ diventa $x = 0.1256$ se troncato alla 4^a decimale ma $x = 013$ se troncato alla seconda decimale.
Per riportare un valore osservato con la sua incertezza si possono utilizzare diverse convenzioni del tutto equivalenti:

1. La convenzione standard é di riportare la grandezza con il suo errore standard:
   X = (12.345 ± 0.026) mm.
   Per numeri molto piccoli o molto grandi é comodo usare la notazione esponenziale o usare multipli e sottomultipli ella grandezza. quindi X = (0.00023 ± 0.00003) mm diventa: X = $(2.3 \pm 3) \cdot 10^{-4}$ mm oppure X = $(0.23 \pm 0.3) \cdot 10^{-4}$ μm.
2. L'incertezza si puó riportare tra parentesi dopo il valore della grandezza (suggerito dalle norme ISO). Ad esempio: X=12.345(26) mm é equivalente a scrivere: X=(12.345 ± 26) mm.
3. Si puó utilizzare il valore relativo ($\epsilon = \sigma_x/x$) o o l'errore relativo percentuale ($\epsilon\% = 100\epsilon$). Quindi per X = (12.345 ± 0.026)mm, essendo $\epsilon = 0.026/12.345 \sim 0.002 = 0.2\%$, si puó scrivere:
   X = 12.35 mm $\pm$ 2% = 12.35 mm $\pm$ 0.002
   oppure:
   X = 12.35 (1 $\pm$ 0.2%) mm = 12.35 (1 $\pm$ 0.002) mm.

1.3 Calcolo dell'incertezza in alcuni casi notevoli

Per una variabile che segue la distribuzione binomiale l'incertezza sul numero di successi osservato N_s su N_T prove é: $\sigma = \sqrt{N_s(1 - f_s)}$ con $f_s = N_s/N_T$.

Per una variabile che segue la distribuzione di Poisson l'incertezza é: $\sigma = \sqrt{N}$ dove N é il numero di eventi osservato. Se la misura viene eseguite M volte e $\bar{N}$ é il numero medio di eventi osservato, l'incertezza su $\bar{N}$ é:

$$\sigma_{\bar{N}} = \frac{\sqrt{\bar{N}}}{\sqrt{M}}$$

Per una distribuzione di frequenze sperimentali (istogramma) l'incertezza standard sulla frequenze assoluta N_i della classe i-esima é:

$$\sigma_i = \sqrt{N_i(1-f_i)} \sim \sqrt{N_i}$$

dove $f_i = N_i/N_T$ con N_T é il numero totale di dati utilizzato per l'istogramma.
L'errore sulle frequenze relative della classe i-esima é:

$$\sigma_{f_i} = \frac{\sigma_i}{N_T} = \frac{\sqrt{f_i(1-f_i)}}{\sqrt{N_T}} \sim \sqrt{\frac{f_i}{N_T}}$$

Dove N_T é il numero totale di dati utilizzato per la costruzione dell'istogramma.
Per una variabile che segue la distribuzione di Normale con valore atteso μ e varianza σ^2 l'incertezza é: σ.
L'incertezza sul valore medio ottenuto da un campione di N misure é:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$

Se non si conosce la deviazione standard si usa:

$$s^2 = \frac{1}{N-1}\sum_{i=1} N(x-\bar{x})^2$$

ovvero la deviazione standard campionaria e quindi:

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{N}}$$

Se il valore medio e la deviazione standard di una grandezza X sono calcolati utilizzando una tabella di frequenze con N_c classi la deviazione standard della media é:

$$\sigma_{\bar{x}} \sim \sqrt{\frac{\sigma^2}{N_c} + \frac{\Delta_c^2}{12}}$$

dove Δ_c é l'ampiezza della singola classe. Se $\Delta_c \ll \sigma$ si ha:

$$\sigma_{\bar{x}} \sim \frac{\sigma}{\sqrt{N_c}}$$

1.4 Intervallo di confidenza

Sia X una variabile aleatoria con valore atteso μ e deviazione standard σ. Indichiamo F(x) la funzione di distribuzione di probabilitá di X e p(x) la sua funzione densitá di probabilitá.
L'intervallo di confidenza è un intervallo intorno al valore atteso μ, la probabilitá di osservare un valore della variabile X all'interno di questo intervallo é detta **Confidenza** C: $P(x_{min} \leq X \leq x_{max}) = C$ ovvero la probabilitá di osservare un valore della X esterno all'intervallo é $\alpha = 1 - C$:

$$P(X \leq x_{min} \cup X \geq x_{max}) = \alpha$$

α é detto anche livello di significativitá dell'intervallo. L'intervallo di confidenza é definito dando i suoi valori estremi:

$$[x_{min}, x_{max}] = [\mu - dx_{min}, \mu - dx_{max}]$$

A priori la distribuzione di probabilitá della variabile X non esimmetrica, per cui non é detto che dx_{min} e dx_{max} siano eguali.
Il valore vero μ di una distribuzione di solito non é noto. Supponiamo peró di aver osservato il valore x, la probabilitá che x sia all'interno dell'intervallo di confidenza é:

$$P(\mu - dx_{min} \leq x \leq \mu + dx_{max}) = C$$

invertendo le disuguaglianze:

$$P(x - dx_{min} \leq \mu \leq x + dx_{max}) = C$$

ovvero possiamo assumere, con probabilitá C, l'intervallo $[x_{min}, x_{max}]$ attorno al valore osservato racchiuda il valore vero. Il che equivale a dire che il rischio di sbagliare affermando che il valore vero é esterno all'intervallo dato é $\alpha = 1 - C$

1.5 Calcolo degli intervalli di confidenza

Per una variabile X che segue una distribuzione normale con valore atteso μ e varianza σ^2 la funzione EXCEL: **INV.NORM**(α, μ, σ) calcola il valore x tale che $P(x \leq \alpha)$ ovvero, indicando con $F(x)$ la distribuzione di probabilitá della variabile x, si ha:

$$F(x) = \int_{-\infty}^{x} p(x)dx = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}}\, e^{\frac{-(x'-\mu)^2}{2\sigma^2}} dx' = \alpha$$

Quindi

$$x = F^{-1}(\alpha) = \text{INV.NORM}(\alpha, \mu, \sigma)$$

Indicando con C la confidenza richiesta e ponendo: $\alpha = (1-C)/2$, i limiti dell'intervallo di confidenza $[x_{min}, x_{max}]$ si calcolano usando:

$$\begin{aligned} x_{min} &= \text{INV.NORM}(\alpha, \bar{\text{x}}, \sigma) \\ x_{max} &= \text{INV.NORM}(1-\alpha, \bar{\text{x}}, \sigma) \end{aligned}$$

dove $\bar{x}$ é la migliore stima del valore vero che posso ottenere dai dati.
Nota: α é la proabilitá di osservare un valore minore del limite inferiore dell'intervallo, questa, per la simmetria della funzione di distribuzione, é eguale alla rpobabilitá di osservare un valore maggiore del limite superiore del'intervallo, quindi 2α é la probabilitá di osservare un valore esterno all'intervallo dato.
Per una variabile Z che segue una distribuzione normale standard (con valore atteso $\mu = 0$ e varianza $\sigma^2 = 1$) la funzione EXCEL: **INV.NORM.ST**(α) calcola il valore z tale che $P(z \leq \alpha)$ ovvero:

$$F(z) = \int_{-\infty}^{z} p(z)dz = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}\, e^{\frac{-z^2}{2}} dz = \alpha$$

quindi

$$z = F^{-1}(\alpha)$$

Possiamo quindi calcolare l'intervallo di confidenza tra $z_{min} = F^{-1}(\alpha/2)$ e $z_{max} = F^{-1}(1-\alpha/2)$ che definisce la regione in cui la variabile z é contenuta con confidenza $C = 1-\alpha$.
Notiamo che per una variabile normale standard l'intervallo di confidenza é simmetrico rispetto all'origine:

$$F^{-1}(\alpha/2) = -F^{-1}(1-\alpha/2)$$

Possiamo quindi calcolare un unico valore limite per la z:

$$z_{lim} = |F^{-1}(\alpha/2)|$$

per definire l'intervallo di confidenza: $[-z_{lim}, z_{lim}]$.
Quindi se C é la confidenza richiesta, poniamo: $\alpha = (1-C)/2$ e calcoliamo

$$z_{lim} = \text{INV.NORM.ST}(1-\alpha) = \text{abs}\,(\text{INV.NORM.ST}(1-\alpha))$$

La funzione Gnuplot: `invnorm`$(1-\alpha)$ é equivalente alla funzione EXCEL `INV.NORM.ST` .
L'uso della variabile standardizzata Z era particolarmente utile per l'uso di valori tabulati della distribuzione normale Standard. Se X é una variabile che segue una distribuzione normale con valore atteso μ e varianza σ^2 si puó costruire una variabile Z:

$$Z = \frac{X-\mu}{\sigma}$$

che segue una distribuzione normale standard. Possiamo quindi calcolare i limiti dell'intervallo di confidenza assumendo una distribuzione normale standard e poi invertire l'equazione precedente e calcolare:

$$\begin{aligned} x_{min} &= \mu - \sigma\, z_{lim} \\ x_{max} &= \mu + \sigma\, z_{lim} \end{aligned} \tag{1}$$

Non conoscendo il valore vero μ si usa la sua migliore stima ovvero la media campionaria dei dati $\bar{x}$.

1.6 Intervallo di confidenza per la media compionaria

Se la media campionaria é calcolata su un set di dati dei quali si conosce la σ si possono usare le formule viste sopra per il calcolo dell'intervallo di confidenza usando

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$

dove N é la numerositá del campione. Se la varianza della distribuzione é nota l'intervallo di confidenza si calcola utilizzando la funzione inversa della distribuzione Normale o della distribuzione normale standard come visto sopra. Se la σ dei dati non é nota ma deve essere stimata dal campione dei dati la variabile

$$t_\lambda = \frac{\bar{x} - \mu}{s_{\bar{x}}}$$

segue una distribuzione di probabilitá nota come t-student, che approssima bene una distribuzione normale per N elevato (maggiore di 10-20). La distribuzione della t dipende dal parametro λ, ovvero il numero di gradi di libertá del campione. In questo caso $\lambda = N - 1$. Questa é una distribuzione simmetrica con valore atteso $\bar{t} = 0$ e deviazione standard che dipende dal numero di gradi di libertá: $\sigma_t = \frac{\lambda}{\lambda - 2}$
La funzione EXCEL: **INV.T**(α, λ) calcola il valore t_{lim} per il quale la probabilitá di osservare un valore di t al di fuori dell'intervallo $[-t_{lim}, t_{lim}]$ é appunto α.
Quindi se C é la confidenza richiesta e $\alpha = 1 - C$, gli estremi dell'intervallo di confidenza sono $[-t_{lim}, t_{lim}]$ calcolati con:

$$t_{lim} = \text{INV.T}(\alpha, \lambda)$$

Questa puó essere invertita:

$$\begin{aligned} x_{min} &= \bar{x} - s_{\bar{x}}\, t_{lim} \\ x_{max} &= \bar{x} + s_{\bar{x}}\, t_{lim} \end{aligned} \tag{2}$$

Quando la deviazione standard é nota (σ) e il valore osservato risulta da una media di N misure é possibile usare la funzione CONFIDENZA implementata in EXCEL: se $\bar{x}$ e s sono rispettivamente la media e la deviazione standard campionarie calcolate su un insieme di N dati,

$$\Delta = CONFIDENZA(\alpha, s, N)$$

é il valore per cui all'intervallo: $[\bar{x} - \Delta, \bar{x} + \Delta]$ é associata la confidenza $C = 1 - \alpha$ ovvero la probabilitá che l'intervallo dato contenga il valore vero é

$$P(\bar{x} - \Delta \leq \mu \leq \bar{x} + \Delta) = \int_{\bar{x}-\Delta}^{\bar{x}+\Delta} p(x)d(x) = C = 1 - \alpha$$

mentre la probabilitá che l'intervallo dato non lo contenga é appunto α

2 Esercizi

2.1 Esercizio E1

Riportare in un documento elettronico i valori delle variabili con l'incertezza associata.

X	σ	u.m.	X	σ	u.m.
0.000541272	0.00000234	g	8.807696	1.698535	mg
6.963879	0.345670	km	13.05671	$4.2883{\cdot}10^{-2}$	V
23.54947	0.0345	A	$9.28836{\cdot}10^{-2}$	4.29E-007	mF
156.9618	6.9348765	mm	696.4838	15.77133	m^2
656.3443	0.23445	N	0.00084373	0.0001254	cm^3

2.2 Esercizio E2

Il file `Cifre_sgnificative.xls` contiene una tabella con misure ripetute di diverse grandezze. Per ognuna di esse calcolare media ed errore standard della media. Riportare in una tabella le medie con gli errori standard.

2.3 Esercizio E3

Calcolare gli intervalli di confidenza al 90%, 95%, 99% per le seguenti variabili il cui valore osservato é X_{oss} assumendo che la distribuzione delle variabili sia normale con deviazione standard nota σ:

X	σ	X	σ
54.1272	0.0023	$8.802{\cdot}10^{-5}$	$1.6{\cdot}10^{-7}$
6.96	0.03	13.05	$4.3{\cdot}10^{-2}$
23.5	0.05	$9.285{\cdot}10^{-2}$	4.E-003
156.9	1.5	696	15
656.3	0.2	0.00084	0.00012

Calcolare gli intervalli di confidenza al 95% per le seguenti variabili sapendo che il valore medio osservato é $\bar{X}_{oss}$, risultato di N misure. Si assuma la distribuzione delle variabili X normale con deviazione standard nota σ:

X_{oss}	σ	N	X_{oss}	σ	N
54.1272	0.0023	7	$8.802{\cdot}10^{-5}$	$1.6{\cdot}10^{-7}$	15
6.96	0.03	5	13.05	$4.3{\cdot}10^{-2}$	10
23.5	0.05	9	$9.285{\cdot}10^{-2}$	4.0E-3	12
156.9	1.5	4	696	15	11
656.3	0.2	12	0.00084	0.00012	20

2.4 Esercizio E4

Utilizzare gli applet di almeno due dei seguenti siti per calcolare i propri tempi di reazione:
http://www.topendsports.com/testing/reaction-timer.htm
http://faculty.washington.edu/chudler/java/reacttime.html
http://www.gamemakers.de/reflextester/
http://www.maniacworld.com/Test_your_reflexes.htm
http://www.javascriptkit.com/script/cut62.shtml
http://faculty.washington.edu/chudler/java/reacttime.html
Scrivere una breve relazione indicando i risultati ottenuti, il proprio tempo di reazione medio con l'errore, l'intervallo di confidenza al 95%.
Confrontare i risultati medi con l'istogramma riportato sul sito:
http://www.topendsports.com/testing/reactionquiz.shtml?view_results

2.5 Esercizio E5

Riferendosi all'esercizio A.2 dell'esercitazione B. Il pesi misurati di due campioni di frutti sono registrati nel file `frutti.dat` (esercitazione B): in prima colonna i dati per il campione A e in seconda colonna per il campione B. Per entrambi i frutti calcolare e riportare nella relazione:

- il peso medio con l'incertezza standard (errore) per i due frutti,
- gli intervalli di confidenza al 90%;
- la probabilitá che un frutto A o B pesi meno dei 2/3 del valore medio.
- la probabilitá che un frutto A o B pesi più di una volta e mezza il valore medio.

2.6 Esercizio E6

Riferendosi ai dati **A.4-5** stabilire valore attesi, errore e intervalli di confidenza al 95% per i tempi di apnea (file **tempi_di_apnea.txt**) e per il peso degli insetti (file **insetti.txt**).

2.7 Esercizio E7

In riferimento all'esperimento C2 dell'esercitazione C (In un secchio sono mescolati alcuni fagioli neri insieme a molti bianchi. Pescare una manciata di fagioli e contare il numero di fagioli neri per 30 volte. Il numero di fagioli neri pescato è una variabile aleatoria che segue una distribuzione di Poisson il cui parametro λ è il numero medio di fagioli neri nel pugno).

- Utilizzare i dati per calcolare l'errore standard σ_λ sul valore medio λ.
- Confrontare l'istogramma sperimentale con la distribuzione teorica ottenuto utilizzando come parametro della distribuzione di Poisson $\lambda' = \lambda - \sigma_\lambda$

- Confrontare l'istogramma sperimentale con la distribuzione teorica ottenuto utilizzando come parametro della distribuzione di Poisson $\lambda' = \lambda - \sqrt{\lambda}$
- Commentare i risultati in una breve relazione.