Registro di MMS
parte di Statistica
a.a. 2007/08

Lezione 1 (26/2/08, 1 ora). Eventi, eventi elementari (esiti), universo S (spazio degli esiti). Formalizzazione matematica: S è un insieme "ambiente", gli eventi sono i sottoinsiemi, gli esiti sono gli elementi di S.

Esempio 1. Un'azienda intraprende un'azione, che potrà condurre ad un successo economico oppure no, e potrà produrre dei danni ambientali oppure no. S ha quattro eventi elementari a, b, c, d: a = successo economico con danno ambientale, b = successo senza danno, c = insucesso con danno, d = insucesso senza danno. L'evento $A = \{a, b\}$ corrisponde a "successo economico", l'evento $B = \{a, c\}$ corrisponde a "danno ambientale", per fare due esempi.

Esempi 2 e 3: lancio di un dado, lancio di due dadi.

Principali operazioni su eventi: $\cup$, $\cap$, complementare.

Lezione 2 (27/2/08, 2 ore). La famiglia di tutti gli eventi è un'algebra rispetto alle operazioni di $\cup$ e $\cap$: commutatività, associatività, distributività; $\emptyset$ è elemento neutro per $\cup$, S per $\cap$. Regole di De Morgan.

Probabilità. Interpretazione intuitiva tramite frequenze empiriche relative (N esperimenti indipendenti effettuati nelle stesse condizioni, ciascuno con possibili esiti alternativi $e_1, ..., e_k$, N_i = numero di esperimenti con esito e_i = frequenza empirica assoluta, $f_i = \frac{N_i}{N}$ = frequenza empirica relativa; la probabilità dell'esito e_i viene immaginata come una idealizzazione di f_i, o come il valore limite di f_i per $N \to \infty$, anche se non ha senso chiedersi se tale limite esista in quanto si parla di esperimenti reali).

Regole formali: $0 \leq P(A) \leq 1$, $P(\emptyset) = 0$, $P(S) = 1$, additività su insiemi disgiunti, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ per insiemi qualsiasi, $P(A^c) = 1 - P(A)$. Esercizio per casa: rispondere alla domanda dell'esempio 3.4.1 del testo (Ross).

Nota: nelle teorie booleane, ci sono solo i valori 0 ed 1, mentre in probabilità si usano tutti i valori tra 0 ed 1: 0 per eventi impossibili (più precisamente, che hanno frequenza nulla di avvenimento), 1 per eventi certi, valori intermedi per tutte le altre situazioni.

Spazi equiprobabili, $P(A) = \frac{n_A}{n}$, dove n è la cardinalità di S, n_A la cardinalità di A. Principio di enumerazione. Esempio 3.5.1 del libro.

Probabilità condizionale. Noto un evento B, ogni A avrà una nuova probabilità $P^{new}(A)$. Il nuovo universo è B, quindi $P^{new}(B) = 1$. Gli esiti in B^c sono impossibili, quindi solo $A \cap B$ sopravvive, quindi è ragionevole assumere che $P^{new}(A)$ sia proporzionale a $P(A \cap B)$. per rispettare la condizione $P^{new}(B) = 1$ dobbiamo prendere $P^{new}(A) = \frac{P(A \cap B)}{P(B)}$.

Lezione 3 (28/2/08, 2 ore). Definizione di probabilità condizionata: se $P(B) > 0$, chiamiamo probabilità di A sapendo B (o condizionata a B) il numero

$$P(A|B) := \frac{P(A \cap B)}{P(B)}.$$

Si può verificare che soddisfa, rispetto ad A, le regole di una probabilità, con universo B: $0 \leq P(A|B) \leq 1$, $P(\emptyset) = 0$, $P(B) = 1$, se $A_1 \cap A_2 = \emptyset$ allora

$$P(A_1 \cup A_2|B) = P(A_1|B) + P(A_2|B).$$

Invece non soddisfa le regole delle probabilità come funzione di B, con A fissato (ad es. $P(A|\emptyset)$ non ha senso, invece che valere 0).

Esempi 3.6.1 e 3.6.3. Il secodo illustra il calcolo di $P(A \cap B)$ usando $P(A|B)$, attraverso la formula

$$P(A \cap B) = P(A|B)\,P(B).$$

Formula di fattorizzazione (anche dimostrazione) e interpretazione grafica con albero. Esempio 3.7.1.

Formula di Bayes (anche dimostrazione), sia nella versione

$$P(B|A) = \frac{P(A|B)\,P(B)}{P(A)}$$

sia in quella con una partizione $B_1, B_2, ...$

$$P(B_k|A) = \frac{P(A|B_k)\,P(B_k)}{P(A|B_1)\,P(B_1) + P(A|B_2)\,P(B_2) + ...}.$$

Interpretazione con cause/effetti: $B_1, B_2, ...$ sono diverse cause possibili, A è un effetto che può verificarsi. Le probabilità "naturali" sono le $P(A|B_k)$, cioè le probabilità che la causa B_k provochi l'effetto A. Bayes permette di calcolare la probabilità delle diverse cause, quando si osserva un effetto A, ad esempio per conoscere la causa più probabile. Esempio 3.7.7.

Esercizi per casa dal capitolo 3: esercizi 1, 2, 3, 4, 5 e 25, 26, 27, 29.

Lezioni 4-5-6 (4-5-6/3/08, 5 ore): vedere homepage Dr. Alessio Rovai.

Lezione 7 (11/3/08, 1 ora). V.a. discrete introdotte attraverso esempi (frammenti dai paragrafi 4.1, 4.2, 4.4 e 5.1). Loro massa probabilità e valor medio. V.a. di Bernoulli, esemplificate con l'esito del controllo della qualità di una moto. V.a. binomiali $B(n,p)$, esemplificate col numero di moto difettose in un gruppo di 10 moto soggette a controllo della qualità (in generale, numero di successi in n esperimenti indipendenti in ciascuno dei quali c'è probabilità p di successo). Grafico della massa di probabilità nei due esempi ed interpretazione nel valor medio per la binomiale.

Lezione 8 (12/3/08, 2 ore). Elementi semplici di calcolo combinatorico (dal paragr. 3.5): principio di enumerazione $n!$ come numero di permutazioni; $n(n-1)\cdots(n-k+1)$ come numero di modi in cui si possono mettere i numeri da 1 a k in un casellario di n caselle; coefficiente binomiale $\binom{n}{k}$ come numero di modi in cui si possono mettere k asterischi indistinguibili in un casellario di n caselle.

Giustificazione della formula delle probabilità $\binom{n}{k}p^k(1-p)^{n-k}$ delle v.a. binomiali (come nel seguito della def. 5.1.2) e rappresentazione delle binomiali come somma di Bernoulli indipendenti (formula 5.1.4).

Lezione 9 (13/3/08, 2 ore). V.a. binomiali: aspetti applicativi. Si consiglia di costruire un foglio Excel. Richiamo sull'interpretazione e la rappresentazione come somma di Bernoulli indipendenti. Calcolo di probabilità di code, tipo $P(X>1)$, interpretazione grafica. Calcolo di soglie: fissato un rischio α, calcolo della soglia λ tale che $P(X>\lambda)\geq 1-\alpha$, e problemi analoghi.

Lezione 10 (18/3/08, 1 ora). Definizione generale di valor medio (valore atteso) $E[X]$ per v.a. discrete, giustificazione della definizione tramite frequenze relative e media aritmetica campionaria come sul libro (paragr. 4.4). Formula per $E[g(X)]$ (proposizione 4.5.1) alcuni esempi. [Non l'esempio 4.4.3 del libro]. Varianza e deviazione standard (solo definizione come al paragr. 4.6 ed anticipazione dell'interpretazione grafica di σ).

Lezione 11 (19/3/08, 2 ore). Proprietà del valor medio: linearità, dimostrata solo nel caso del corollario 4.5.2, più enunciato generale come nel paragr. 4.5.1; teorema 4.7.3 su $E[XY]=E[X]E[Y]$ per v.a. indipendenti (senza dim.) e su $Var[X+Y]=Var[X]+Var[Y]$ per v.a. indipendenti (con dim.); formula 4.6.2. [Ancora non la covarianza]. Esempi: media e varianza di v.a. di Bernoulli e binomiali. Proprietà di concentrazione delle binomiali ad alta numerosità. Es. per casa: $Var[aX]=a^2Var[X]$, $Var[X-Y]=Var[X]+Var[Y]$ per v.a. indipendenti.

V.a. di Poisson (paragr. 5.2): definizione, teorema degli eventi rari (approssimazione tramite binomiali) ed interpretazione come numero di successi in situazioni in cui il numero totale di esperimenti è indeterminato e molto alto, e si conosce il numero medio λ di successi.

Esercizi per casa: si possono in generale provare a svolgere gli esercizi numero 2 dei compiti d'esame fino al 2006 incluso, e l'esercizio numero 1 dei compiti del 2007, escludendo però le domande più difficili e quelle che coinvolgono gaussiane ed esponenziali.

Lezione 13 (27/3/08, 2 ore). Esercizio: data la v.a. discreta X che vale 2 con prob. 1/4, 0 con prob. 1/2, -2 con prob. 1/4, calcolare media, varianza, $E\left[e^{tX}\right]$, $E\left[5X - 2X^2 + 3\right]$. Se X ed Y sono indipendenti con la stessa legge della X precedente, calcolare $E\left[2XY + X^2 - XY^2\right]$, $Var\left[3X - 5Y\right]$. Esercizio: $Var\left[aX\right] = a^2 Var\left[X\right]$, $Var\left[X + b\right] = Var\left[X\right]$. Esercizio per casa (un po' più difficile): nelle ipotesi dell'esercizio precedente, calcolare $E\left[\frac{X^2}{1+Y^2}\right]$, $P\left(X \leq 1\right)$, $P\left(X^2 \leq 1\right)$.

Concetto di v.a. continua, densità di probabilità, probabilità calcolate a partire dalla densità ($P\left(a \leq X \leq b\right) = \int_a^b f\left(x\right) dx$ e casi simili), funzione di ripartizione (o distribuzione) cumulativa $F\left(t\right) = P\left(X \leq t\right)$ (data da $\int_{-\infty}^t f\left(x\right) dx$ per le variabili continue), sua monotonia e limiti all'infinito, grafico costante a tratti nel caso di v.a. discrete. Tutto questo è descritto al paragrafo 4.2.

Lezioni 14-15-16 (1-2-3/4/08, 5 ore): vedere homepage Dr. Alessio Rovai.

Lezione 17 (8/4/08, 1 ora). Definizione di v.a. esponenziale (par. 5.6), $P\left(X > t\right) = e^{-\lambda t}$ (funzione di affidabilità o di sopravvivenza), formule per $F(t)$, $E\left[X\right]$, $Var\left[X\right]$. Interpretazione grafica, confronto grafico tra due casi, $\lambda = 2$ e $\lambda = 0.5$. Nota: $\sigma = \mu$, elevata aleatorietà, in contrasto con binomiali e Poisson.

Lezione 18 (9/4/08, 2 ore). Proprietà di assenza di memoria delle v.a. esponenziali: enunciato, dimostrazione, interpretazione applicativa (mancanza di logoramento). Sistema (con tempo di vita X) composto da N sottosistemi (con tempi di vita $X_1, ..., X_N$); in serie: $X = \min\left(X_1, ..., X_N\right)$; in parallelo: $X = \max\left(X_1, ..., X_N\right)$; con sostituzione: $X = X_1 + ... + X_N$. Teorema sul minimo di v.a. esponenziali indipendenti (con dimostrazione). Nei casi in parallelo e con sostituzione non esplicitiamo la densità ma osserviamo che sarà concentrata, invece che esponenziale; ad es. osserviamo per $X = X_1 + ... + X_N$ che vale $E\left[X\right] = N \cdot \frac{1}{\lambda}$, $Var\left[X\right] = N \cdot \frac{1}{\lambda^2}$, quindi

$\sigma_X = \sqrt{N} \cdot \frac{1}{\lambda}$, cioè si ritrova un andamento simile a binomiali e Poisson.

$F' = f$ nei punti di continuità di f.

Funzione generatrice dei momenti $\phi(t)$. Proprietà $\phi'(t) = E[X]$, $\phi''(t) = E[X^2]$. Calcolo di $\phi(t)$ per le v.a. esponenziali. Per casa: da qui calcolare $E[X]$ e $Var[X]$ per le v.a. esponenziali.

Lezione 19 (10/4/08, 2 ore). Vari esempi di funzione generatrice, per Bernoulli, binomiale, Poisson. Ci si appoggia sulla Proposizione 4.8.1, dimostrata. Calcolo rigoroso di $E[X]$ e $Var[X]$ per le v.a. di Poisson.

V.a. Gaussiane. Densità gaussiana $N(\mu, \sigma^2)$ dipendente dai parametri μ e σ^2; suo grafico, interpretazione dei parametri e dipendenza da σ; funzione generatrice (calcolata), da cui si trova che μ e σ^2 sono $E[X]$ e $Var[X]$. Definizione di $\Phi(t)$, funzione di ripartizione cumulativa standard.

Lezione 20 (16/4/08, 2 ore). Proposizione 5.5.1 sulle gaussiane e sua conseguenza:

$$F_{\mu,\sigma}(t) = \Phi\left(\frac{t-\mu}{\sigma}\right)$$

dove $F_{\mu,\sigma}(t)$ è la funzione di ripartizione cumulativa di una $N(\mu, \sigma^2)$. Uso delle tavole per calcolare $\Phi(t)$ con $t \in [0, 3.5]$, formula $\Phi(-t) = 1 - \Phi(t)$ per i valori negativi, varie interpretazioni geometriche dei calcoli di questo tipo (anche calcolo di $P(X \geq t)$, $P(a \leq X \leq b)$ per gaussiane generiche).

Circa la Proposizione 5.5.1, si insiste sul fatto che la dimostrazione serve soprattutto per dimostrare la gaussianità, in quanto il valore esplicito di media e varianza è ovvio da calcoli ben più facili. Si insiste anche sul concetto di standardizzazione, usato per dimostrare $F_{\mu,\sigma}(t) = \Phi\left(\frac{t-\mu}{\sigma}\right)$: data una v.a. qualsiasi X avente media μ e varianza σ^2, la v.a. $\frac{X-\mu}{\sigma}$ si chiama standardizzata di X, in quanto ha media nulla e varianza unitaria; se X è gaussiana, anche $\frac{X-\mu}{\sigma}$ lo è.

Risoluzione di due esercizi: $P(|Y-3| > 1)$ se $Y \sim N(3, 2)$; $P(XY < 0)$ se X è indipendente da Y e vale ± 1 con ugual probabilità.

Lezione 21 (17/4/08, 2 ore). Le combinazioni affini $aX + bY + c$ di gaussiane indipendenti sono gaussiane (con dimostrazione, prima dell'es. 5.5.4).

Esercizio (si veda anche il paragrafo 6.2): date $X_1, ..., X_n$ indipendenti, con la stessa distribuzione, valor medio μ e varianza σ^2, dire quello che si può delle v.a. $S_n = X_1 + ... + X_n$ e $\overline{X}_n = \frac{S_n}{n}$; svolgere l'esercizio sia in generale sia quando le X_k sono gaussiane. Infine, scrivere la standardizzata di S_n (risultato: $\frac{S_n - n\mu}{\sqrt{n}\sigma}$).

Osservazione: quando le X_k sono gaussiane, la v.a. $\frac{S_n - n\mu}{\sqrt{n}\sigma}$ è una gaussiana standard $N(0,1)$.

Teorema Limite Centrale (teorema 6.3.1): in generale, la v.a. $\frac{S_n - n\mu}{\sqrt{n}\sigma}$ *converge* ad una gaussiana standard $N(0,1)$ per $n \to \infty$, nel senso che

$$P\left(a \leq \frac{S_n - n\mu}{\sqrt{n}\sigma} \leq b\right) \underset{n\to\infty}{\rightarrow} P(a \leq Z \leq b) = \Phi(b) - \Phi(a)$$

dove $Z \sim N(0,1)$. Esempi ed esercizi.

Può essere utile riassumere il programma sul libro svolto fino ad ora, programma richiesto in vista del primo compitino. Il capitolo 3 è fondamentale, specialmente per quanto riguarda la probabilità condizionale, l'indipendenza, la formula di fattorizzazione e la formula di Bayes, mentre il calcolo combinatorico serve solo funzionalmente agli sviluppi sulle binomiali necessari nel seguito.

Del capitolo 4 abbiamo visto le definizioni fondamentali di variabile aleatoria discreta e continua, con le relative densità di probabilità, la funzione di ripartizione cumulativa, la definizione di valor medio e varianza e le loro proprietà di calcolo, la funzione generatrice dei momenti e le sue proprietà, come operare su trasformazioni di variabili. Invece non abbiamo svolto il paragrafo 4.3, la parte del 4.7 riguardante la covarianza, ed il 4.9.

Del capitolo 5 abbiamo imparato a lavorare sulle v.a. di Bernoulli, binomiali e di Poisson ed abbiamo visto i legami tra esse. Abbiamo poi studiato a fondo le v.a. esponenziali e gaussiane, in tutti i loro aspetti. Abbiamo invece per ora tralasciato ipergeometriche, uniformi e dal paragrafo 5.6.1 incluso in poi. L'unico argomento sulle gaussiane *non* ancora svolto sono i *quantili* q_α e z_α.

Infine, abbiamo completato lo studio delle gaussiane studiando il Teorema Limite Centrale del paragrafo 6.3, esemplificato sia graficamente sia con esercizi. Anche il calcolo del paragrafo 6.2 è stato svolto a titolo di esercizio.

Lezione 22 (22/4/08, 1 ora). Esercizi sulle gaussiane e teorema limite centrale.

Lezione 23 (23/4/08, 2 ore). Compitino.

Lezione 24 (24/4/08, 2 ore). Concetto di campione sperimentale e media campionaria (par. 6.2), intese come v.a. (nell'ottica pre-sperimentale, di progettazione degli esperimenti - DOE = Design Of Experiments). Proprietà della media campionaria $\overline{X}$. Elenco di ragioni (proprietà) per cui riteniamo che $\overline{X}$ sia una buona approssimazione (*stima*) di μ. Nel caso particolare $X_i \sim N(\mu, \sigma^2)$, vale $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, quindi possiamo calcolare $P\left(\mu - \delta \leq \overline{X} \leq \mu + \delta\right)$, quantificando l'affermazione "$\overline{X}$ è vicino a μ con elevata probabilità". Tale probabilità è anche uguale a $P\left(\overline{X} - \delta \leq \mu \leq \overline{X} + \delta\right)$: se μ è incognito, questo quantifica con che probabilità abbiamo l'informazione quantitativa $\overline{X} - \delta \leq \mu \leq \overline{X} + \delta$ su μ (intervallo di confidenza).

Di solito è più interessante assegnare queste probabilità e calcolare il relativo valore di δ (=errore dell'approssimazione). Bisogna quindi saper risolvere problemi inversi relativi alle gaussiane.

Il problema inverso più elementare è: data una gaussiana canonica $Z \sim N(0,1)$ ed una probabilità $\alpha \in (0,1)$, trovare x tale che $P(Z \leq x) = \alpha$. In altri termini, dobbiamo risolvere l'equazione $\Phi(x) = \alpha$. La soluzione si chiama *quantile di ordine* α, denotato q_α. Questo problema viene interpretato graficamente e illustrato con l'uso (inverso) delle tavole. Ovviamente il problema $P(Z \leq x) = 1 - \alpha$ ha come soluzione $q_{1-\alpha}$; siccome questo problema è molto comune (ovvero che venga assegnata l'area α della coda destra), la soluzione x viene anche indicata col simbolo z_α. Vale quindi, per definizione, $q_{1-\alpha} = z_\alpha$.

Un problema inverso più generale è: data una gaussiana generica $X \sim N(\mu, \sigma^2)$ ed una probabilità $\alpha \in (0,1)$, trovare x tale che $P(X \leq x) = \alpha$. La soluzione è

$$x = \mu + \sigma q_\alpha.$$

Intuitivamente questo si può capire confrontando i grafici delle due gaussiane X e Z: differiscono per la traslazione in μ e per il cambio di unità di misura σ sull'asse delle ascisse. Rigorosamente basta riscrivere l'equazione $P(X \leq x) = \alpha$ nella forma $\Phi\left(\frac{x-\mu}{\sigma}\right) = \alpha$, così da ricondurci al problema precedente, per cui vale $\frac{x-\mu}{\sigma} = q_\alpha$, da cui si trova $x = \mu + \sigma q_\alpha$.

Lezione 25 (29/4/08, 1 ora). Consolidamento sui quantili. Quantili notevoli: $q_{0.975} = z_{0.025} = 1.96$, $q_{0.95} = z_{0.05} = 1.645$, e forse qualche altro.

Problema inverso a due code: data una gaussiana canonica $Z \sim N(0,1)$ ed una probabilità $\alpha \in (0,1)$, distribuiamo α equamente sulle due code (si interpreti graficamente), cercando il valore x per cui $P(-x \leq Z \leq x) = 1-\alpha$. La soluzione è $x = q_{1-\alpha/2}$. Ad esempio, per $\alpha = 0.05$, si trova $x = q_{0.975} = 1.96$ (mentre per il problema $P(Z \leq x) = 1-\alpha$ si trovava $x = q_{0.95} = 1.645$).

Intervalli di confidenza (par. 7.3). Significato ed importanza della dichiarazione "al 95%" (o simili).

Lezione 26 (30/4/08, 2 ore). Da par. 7.3 abbiamo appreso che

$$P\left(-\frac{\sigma q_{1-\alpha/2}}{\sqrt{n}} \leq \overline{X}-\mu \leq \frac{\sigma q_{1-\alpha/2}}{\sqrt{n}}\right)=1-\alpha.$$

Ad es., al 95%, possiamo dire che vale $-\frac{\sigma 1.96}{\sqrt{n}} \leq \overline{X}-\mu \leq \frac{\sigma 1.96}{\sqrt{n}}$, ovvero $\overline{X}-\frac{\sigma 1.96}{\sqrt{n}} \leq \mu \leq \overline{X}+\frac{\sigma 1.96}{\sqrt{n}}$, che scriviamo nella forma

$$\mu=\overline{X} \pm \frac{\sigma 1.96}{\sqrt{n}}.$$

Iniziamo la risoluzione di un esercizio realistico e complesso. Un commerciante vuole introdurre un nuovo prodotto e si chiede quanti esemplari del prodotto deve tenere a disposizione giornalmente. Chiamiamo N il numero (aleatorio) di esemplari che gli vengono richiesti giornalmente, ed indichiamo con m il numero che dovrebbe tenere a disposizione (l'incognita del problema).

Se ad esempio sapesse che N è una $N(5,4)$, m sarebbe la soluzione del problema $P(N \leq m)=1-\alpha$, con α fissato dal commerciante in base al rischio che vuole correre di non accontentare tutta la clientela. La probabilità $\alpha \in (0,1)$ misura tale rischio. la soluzione sappiamo essere $m=5+2q_{1-\alpha}$.

Ci sono due ordini di problemi in questa traccia di risoluzione. Il primo è che la gaussianità di N è un'ipotesi molto discutibile: N assume valori interi positivi, una gaussiana assume valori reali. Il secondo è che, pur accettando per semplicità l'ipotesi di gaussianità di N, non conosciamo la sua media e la sua varianza (i valori 5 e 4 li avevamo inventati per illustrare la strategia di risoluzione). Bisogna allora pianificare una campagna di indagini per stimare μ e σ^2.

Supponiamo di effettuare n osservazioni sperimentali (registrare il numero di richieste in n giorni). Avremo il campione $N_1, ..., N_n$ da cui calcoleremo $\overline{N}$, come approssimazione di μ, $S^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(N_i-\overline{N}\right)^2$ come approssimazione di σ^2.

Una parentesi, che riprenderemo: ora che sappiamo analizzare le v.a., cosa ci spinge a ritenere che S^2 sia una buona approssimazione di σ^2? Dimostreremo ad esempio che è uno stimatore corretto (non distorto): $E[S^2]=\sigma^2$. Ad es., se avessimo considerato l'espressione apparentemente più semplice $\frac{1}{n}\sum_{i=1}^{n}\left(N_i-\overline{N}\right)^2$, la sua media sarebbe stata $\frac{n-1}{n}\sigma^2$, vicina a σ^2 per n grande, ma un po' diversa per n piccolo.

Tornando al nostro problema, sappiamo che quando avremo effettuato gli esperimenti avremo un'iformazione del tipo: al 95% vale

$$\mu = \overline{N} \pm \frac{\sigma 1.96}{\sqrt{n}}.$$

Cosa useremo per calcolare m, il numero $m = \overline{N} + \sigma q_{1-\alpha}$, oppure $m = \overline{N} + \frac{\sigma 1.96}{\sqrt{n}} + \sigma q_{1-\alpha}$, o cos'altro? Se $\frac{\sigma 1.96}{\sqrt{n}}$ è grande, la differenza è sostanziale. Nasce qui come da altri ragionamenti l'ovvio desiderio di avere un errore $\frac{\sigma 1.96}{\sqrt{n}}$ piccolo. Possiamo agire sulla numerosità n degli esperimenti. Quindi, se prefissiamo che l'errore deve essere inferiore ad un certo valore δ, dobbiamo risolvere la disequazione (nell'incognita n) $\frac{\sigma 1.96}{\sqrt{n}} \leq \delta$, da cui $n \geq \left(\frac{\sigma 1.96}{\delta}\right)^2$.

Lezione 27 (7/5/08, 1 ora). Esercizio sulla stima al 95% della media μ incognita di una $N(\mu, 4)$, dato un campione $x_1, ..., x_{25}$ con media $\overline{x} = 275$.

Esercizio sulla progettazione di esperimenti al fine stimare la media μ incognita di una $N(\mu, 4)$, al 95%, con precisione pari a 1. Bisogna trovare il più piccolo valore della numerosità n che soddisfa questi due requisiti.

Problema nella pratica: se non si conosce μ non si conosce nemmeno σ. A volte si hanno informazioni pregresse su situazioni simili e si usa il σ di quelle, anche se potrebbe non essere il valore giusto. In fase di analisi di dati $x_1, ..., x_n$ si approssima σ con la deviazione sperimentale S. A causa di questa approssimazione, la formula $\delta = \frac{S q_{1-\alpha/2}}{\sqrt{n}}$ non è più esattamente vera. Diventa invece esattamente vera una sua variante, $\delta = \frac{S t^{(n-1)}_{1-\alpha/2}}{\sqrt{n}}$, dove $t^{(k)}_{\beta}$ è il quantile della distribuzione t di Student a k gradi di libertà, di ordine β (analogo al quantile gaussiano q_β, sono che dipende dal parametro aggiuntivo k).

Lezione 28-29 (8-13/5/08, 3 ore): vedere homepage Dr. Alessio Rovai.

Lezione 30 (14/5/08, 2 ore). Riassunto di statistica: i) stima dei parametri incogniti, ii) verifica di parametri dichiarati. Stima: i) puntuale, ii) intervallare. Stima puntuale: i) statistica descrittiva, ii) complementi teorici (stimatori corretti ecc.). $\overline{X}$ è un buon stimatore di μ. S^2 è un buon stimatore di σ^2? Teorema: è corretto (dim. come nel par. 6.4). Definizione di v.a. Chi quadro (par. 5.8.1), grafico, quantili, media e varianza. Teorema 6.5.1, senza dimostrazione. Conseguenza: $Var\left[S^2\right] = \frac{\sigma^4 C}{n-1}$ (C costante opportuna), per cui anche S^2 si stringe attorno alla sua media σ^2, come $\overline{X}$ si stringe attorno a μ ($Var\left[\overline{X}\right] = \frac{\sigma^2}{n}$).

Ricordiamo il teorema sulla stima intervallare in ipotesi gaussiane: $\mu = \overline{X} \pm \frac{\sigma q_{1-\frac{\alpha}{2}}}{\sqrt{n}}$ con probabilità $1 - \alpha$. Cosa si può dire se X non è gaussiana, ma soddisfa le ipotesi del TLC? Il TLC si può riformulare come $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \to N(0,1)$, quindi

$$P\left(-q_{1-\frac{\alpha}{2}} \leq \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq q_{1-\frac{\alpha}{2}}\right) \to 1-\alpha$$

da cui, con gli stessi calcoli del par. 7.3, troviamo: $\mu = \overline{X} \pm \frac{\sigma q_{1-\frac{\alpha}{2}}}{\sqrt{n}}$ con probabilità *approssimativamente* $1-\alpha$.

Esempio: $X \sim B(1,p)$, $\mu = p$, $\sigma^2 = p(1-p)$, quindi posto $\widehat{p} = \overline{X}$ (proporzione sperimentale di successi), vale

$$p = \widehat{p} \pm \frac{\sqrt{p(1-p)}q_{1-\frac{\alpha}{2}}}{\sqrt{n}}$$

con probabilità approssimativamente $1-\alpha$.

Viene affrontato il problema: σ non si conosce. A posteriori, cioè dopo gli esperimenti, si conosce S, oppure $\widehat{p}$ e quindi $\sqrt{\widehat{p}(1-\widehat{p})}$ nel caso delle proporzioni, quindi si possono usare questi valori come approssimazioni di σ. Ricordiamo inoltre che, nel caso gaussiano, $\frac{St_{1-\frac{\alpha}{2},n-1}}{\sqrt{n}}$ è una formula esatta per l'errore. Oppure, nel caso delle proporzioni, si può risolvere la disequazione

$$p \leq \widehat{p} + \frac{\sqrt{p(1-p)}q_{1-\frac{\alpha}{2}}}{\sqrt{n}}$$

e l'altra simile, trovando un intervallo esplicito ed esatto per p.

Lezione 31 (15/5/08, 2 ore). Riassunto, commenti ed esempi sulle formule degli intervalli di confidenza sia nel caso gaussiano che generale. Collegamento col problema di calcolare soglie (verrà ripreso).

Errore assoluto ed errore relativo, ricerca di n nei due casi.

Viene in particolare esemplificata la dipendenza di n da certe scelte, ragionando sui compromessi tipici nelle applicazioni per tenere n ad un livello ragionevole.

A priori (in fase di DOE), i) si possono fare degli esperimenti preliminari ottenendo un valore S che stima grossolanamente σ; ii) si possono cercare valori noti in letteratura in situazioni simili; iii) in alcuni casi si sa a priori che il parametro incognito appartiene ad un certo insieme e si può prendere il caso peggiore: ad esempio siccome $p \in [0,1]$, vale sempre $\sqrt{p(1-p)} \leq 0.5$.

Lezione 32 (20/5/08, 1 ora). Carte di controllo (Cap. 13) e test statistici (Cap. 8). Viene introdotta l'idea tramite le carte di controllo della produzione industriale di un certo prodotto, la cui caratteristica sia descritta da una v.a. X, il cui valor medio μ_0 deve assumere un certo valore se la produzione è corretta. Periodicamente si esegue un controllo su un campione di numerosità n, si calcola la media campinaria $\overline{x}$, la si disegna sulla carta e si controlla se è uscita dalla banda delimitata da UCL ed LCL. Questi limiti sono calcolati, nell'ipotesi $X \sim N(\mu_0, \sigma^2)$ con σ^2 nota, usando la proprietà $\overline{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$, quindi tagliando due code simmetriche di ampiezza $\frac{\alpha}{2}$ dalla distribuzione di $\overline{X}$. Si trova $UCL = \mu_0 + \frac{\sigma}{\sqrt{n}} q_{1-\frac{\alpha}{2}}$ e $LCL = \mu_0 - \frac{\sigma}{\sqrt{n}} q_{1-\frac{\alpha}{2}}$. Il controllo appena detto equivale a controllare se $|z| > q_{1-\frac{\alpha}{2}}$ dove $z = \frac{\overline{x}-\mu_0}{\sigma}\sqrt{n}$: questo è il modo usuale di eseguire un test.

Lezione 33 (21/5/08, 2 ore). La strategia di test viene capita in un secondo modo, facendo riferimento agli intervalli di confidenza. Se abbiamo un campione da una gaussiana di media μ incognita, sappiamo che $\mu = \overline{X} \pm \frac{\sigma q_{1-\frac{\alpha}{2}}}{\sqrt{n}}$ con probabilità $1-\alpha$. Allora, se viene ipotizzato che la media sia μ_0 e vogliamo controllare se il campione è coerente con questa ipotesi (cioè se il campione rifiuta l'ipotesi o non permette di rifiutarla), basta vedere se μ_0 appartiene all'intervallo di confidenza. Questo conduce alla solita condizione $|z| < q_{1-\frac{\alpha}{2}}$. Se $|z| < q_{1-\frac{\alpha}{2}}$, non rifiutiamo l'ipotesi, altrimenti la rifiutiamo.

In analogia con la teoria degli intervalli di confidenza, bisognerebbe prefissare α. Qui però è uso comune (es. dei software) non fissare α a priori ma cercare il *valore* p di demarcazione tra gli α che avrebbero condotto a rifiutare l'ipotesi μ_0 e gli altri. Dopo aver discusso intuitivamente il problema si arriva a capire che il test avrebbe rifiutato l'ipotesi per tutti gli $\alpha > p$. Inoltre p si calcola risolvendo l'equazione

$$|z| = q_{1-\frac{p}{2}}$$

da cui si trova $p = 2 - 2\Phi(|z|)$.

Viene poi ampliato il linguaggio relativo alla teoria dei test (ipotesi nulla e ipotesi alternativa). Viene discusso il caso unilaterale, in cui l'ipotesi alternativa è del tipo "media maggiore di μ_0". In questo caso, rifacendoci all'intuizione basata sulla carta di controllo, si disegna una carta con solo la linea UCL, che dovendo tagliare fuori un'area pari ad α vale $UCL = \mu_0 + \frac{\sigma}{\sqrt{n}} q_{1-\alpha}$. Il test corrispondente si esegue controllando se $z > q_{1-\alpha}$ (in questo caso si rifiuta l'ipotesi e si afferma che "la media maggiore di μ_0". Per esercizio, si cerchi di capire se il valore p del caso bilaterale, p^B, è maggiore o minore di p^U, il valore p del caso unilaterale.

Lezione 34 (22/5/08, 2 ore). Test per la media: fermo restando in tutti i casi che l'ipotesi nulla è "la media è μ_0", abbiamo visto: i) il caso bilaterale, in cui $\mathcal{H}_1$ è "la media è diversa da μ_0", il test è significativo (rifiuta l'ipotesi nulla) se $|z| > q_{1-\frac{\alpha}{2}}$, il valore p è $p^B = 2 - 2\Phi(|z|)$; i) il caso unilaterale con $\mathcal{H}_1 =$ "la media è maggiore di μ_0", il test è significativo se $z > q_{1-\alpha}$, il valore p è $p^{U^+} = 1 - \Phi(z)$; i) il caso unilaterale con $\mathcal{H}_1 =$ "la media è minore di μ_0", il test è significativo se $z < -q_{1-\alpha}$, il valore p è $p^{U^-} = 1 - \Phi(-z)$. Questi valori di p ed i test sono stati ricavati o motivati, anche su base intuitiva.

Se $z > 0$, $p^B = 2p^{U^+}$, quindi $p^{U^+} < p^B$. Questo mostra che il test unilaterale rifiuta l'ipotesi più facilmente di quello bilaterale.

DOE: p dipende da z, quindi si può calcolare solo a posteriori. Per ora non abbiamo grandezze che permettano di decidere la numerosità in fase di DOE. Unica cosa: le precedenti considerazioni su p possono portare a scegliere tra diversi test.

Errore di I e II specie. Si calcola la probabilità dell'errore di I specie e si trova che vale α. Quindi α non è la probabilità di arrivare ad una conclusione sbagliata eseguendo un test (come invece era per gli intervalli di confidenza) ma è solo la probabilità di rifiutare quando invece l'ipotesi nulla è vera. Si calcola poi la probabilità β dell'errore di II specie osservando che essa dipende da $\mu \neq \mu_0$ e trovando che vale

$$\beta(\mu) = \Phi\left(\frac{\mu_0 - \mu}{\sigma}\sqrt{n} + q_{1-\frac{\alpha}{2}}\right) - \Phi\left(\frac{\mu_0 - \mu}{\sigma}\sqrt{n} - q_{1-\frac{\alpha}{2}}\right).$$

Si definisce potenza del test la funzione $1 - \beta(\mu)$. Essa è la probabilità di rifiutare l'ipotesi nulla quando la media vera è $\mu \neq \mu_0$. La potenza non dipende dai dati sperimentali e dipende da n. Quindi può essere usata in fase di DOE. Ci si può chiedere: fissata la significatività α, data l'ipotesi "la media è μ_0", nota σ^2, che numerosità serve per avere una certa potenza (es. 90%) di accorgersi di un incremento della media pari ad un valore δ (cioè avere $1 - \beta(\mu_0 + \delta) = 90\%$)?

Si riprende, per concludere un problema già discusso in lezioni passate: detto X il numero (aleatorio) di richieste giornaliere di un prodotto, calcolare $\lambda =$ numero di pezzi da tenere a disposizione per soddisfare la clientela il 95% delle volte (in media). Se X è una v.a. discreta, appunto il numero di pezzi, si deve trovare il più piccolo intero λ tale che $P(X > \lambda) \leq 0.05$. Se X fosse stata una v.a. continua, si poteva risolvere l'equazione $P(X > \lambda) = 0.05$, trovando un numero reale λ (e prendendo il primo intero superiore se vogliamo un numero intero). La ricerca di λ si può eseguire con un software per varie distribuzioni (Poisson ecc.), oppure con le tavole dei quantili nel caso gaussiano.

Il problema nasce se i parametri della distribuzione di X non sono noti e vogliamo stimarli. Esemplifichiamo supponendo che X sia una Poisson di media μ incognita, oppure una gaussiana di media μ incognita e (per semplicità) varianza σ^2 nota. Da un campione otteniamo una stima $\overline{x}$ di μ. Possiamo, un po' grossolanamente, far finta che μ valga $\overline{x}$ ed usare questo valore per trovare la soglia λ, che chiameremo nel seguito $\overline{\lambda}$. Per trovarla, nel caso Poisson si deve usare Excel, nel caso gaussiano vale

$$\overline{\lambda} = \overline{x} + \sigma q_{0.95}.$$

Però sappiamo anche che $\mu = \overline{x} \pm \delta$ a livello di confidenza prefissato $1 - \alpha$ (approssimativamente per una Poisson, esattamente per una gaussiana), dove δ è dato da una certa espressione. Possiamo allora effettuare il calcolo di λ nei due casi estremi, in cui prendiamo come media μ i due valori $\overline{x}+\delta$ e $\overline{x}-\delta$. Chiamiamo $\lambda^- < \lambda^+$ queste due soglie estreme, che tra l'altro comprendono la soglia $\overline{\lambda}$. Possiamo affermare che, a livello di confidenza $1-\alpha$, la soglia vera λ sta nell'intervallo tra λ^- e λ^+. Inoltre, se vogliamo effettuare una scelta di λ molto cautelativa, possiamo prendere proprio λ^+ (il peggior valore di λ compatibile con l'intervallo di confidenza per la media).

Riassunto del programma per il secondo compitino. Il riassunto descritto qui non si intende esaustivo, solo il registro riporta tutti i dettagli.

I temi principali sono quelli che si trovano usualmente negli esercizi numero 3 dei compiti del 2005/06 e precedenti (salvo eccezioni di alcuni compiti strutturati diversamente), e degli esercizi numero 2 dei compiti del 2006/07 (di nuovo salvo eccezioni di alcuni compiti strutturati diversamente).

Includono comunque la capacità di calcolare probabilità e soglie (quantili) gaussiani, che a volte si trovano negli esercizi su probabilità e valori medi; si vedano anche le lezioni 20, 24, 25. Analogamente, essendo la matematica spesso tutta collegata, è necessario saper usare anche argomenti precedenti al primo compitino, come le proprietà generali dei valori medi e di tutte le v.a. studiate, la formula di fattorizzazione, il TLC e così via.

Più specificamente di statistica, si deve saper calcolare l'intervallo di confidenza per la media per v.a. gaussiane e qualsiasi, e saper ragionare sulla formula dell'errore (precisione) sia come se si fosse in fase di DOE sia in fase post-sperimentale. Saper ragionare significa sapersi muovere di fronte alla difficoltà della mancata conoscenza di σ, o più classicamente saper valutare la numerosità che serve per ottenere certe prestazioni (un certo errore assoluto o relativo, con una certa confidenza). Nel caso gaussiano è bene conoscere anche la formula con la t di Student. Molto importante: la stima poi si collega di solito al problema successivo di calcolare grandezze di interesse per le v.a. esaminate, come certe soglie (vedere ad es. l'ultima lezione).

Circa la teoria dei test bisogna aver capito le varie nozioni ed idee, saper eseguire i test bilaterali e unilaterali per la media, calcolare il valore p, calcolare la potenza (bilaterale), ed aver capito almeno potenzialmente come si potrebbe decidere la numerosità (in fase di DOE) per avere una potenza preassegnata.