Descrizione sommaria degli argomenti svolti a lezione, anno accademico 2006-07

Lezione 1 (27/2, 2 ore). Introduzione al corso (lezioni ed esercitazioni: Franco Flandoli e David Barbato; testo: *S. Ross, Probabilità e Statistica per l'Ingegneria e le Scienze*, *Apogeo* 2002; materiale complementare e soprattutto compiti d'esame risolti: in rete alle pagine di Gubinelli e Flandoli). Si comincia dal Capitolo 3: oggetti del calcolo delle probabilità: universo o spazio degli eventi elementari, eventi come sottoinsiemi dell'universo (corrispondenza con le affermazioni o proposizioni) ed operazioni di unione, intersezione e complementare sugli eventi (corrispondenza con le operazioni di "o", "e" e "non" su proposizioni); probabilità (come funzione dagli eventi ai numeri di $[0,1]$) e sue regole. Probabilità condizionale (universo ridotto a B quando si sa che l'evento B si è verificato e nuova valutazione delle probabilità dei vari eventi); formula di fattorizzazione. Tutte le nozioni sono state esemplificate tramite l'esercizio 1.i del 4/6/02 (homepage Gubinelli).

Osservazioni sulle probabilità condizionate: si comportano come probabilità, ad esempio nel senso che $P(A^c|B) = 1 - P(A|B)$. Invece non c'è alcuna relazione tra $P(A|B)$ e $P(A|B^c)$.

Lezione 2 (28/2, 2 ore). Richiami su quanto visto nella lezione 1. Formula di Bayes: dimostrazione, interpretazione come mezzo per il calcolo della probabilità delle cause quando si osserva un determinato effetto (conoscendo le probabilità condizionate degli effetti sapendo le cause). In ingegneria serve ad esempio nella ricostruzine di segnali e nella ricerca delle cause di guasti. Esemplificazione tramite l'esercizio 1.ii del 4/6/02.

Elementi facili di calcolo combinatorico. Principio di enumerazione. Esempio 1: permutazioni (riordinamenti possibili di n oggetti diversi); loro cardinalità = $n!$ (dimostrazione tramite principio di enumerazione). Esempio 2: disposizioni di k elementi in n posti; loro cardinalità = $n(n-1)\bullet\bullet\bullet(n-k+1)$ (dimostrazione tramite principio di enumerazione). Riformulazione delle disposizioni: numero di stringhe ordinate di k elementi che si possono formare con n oggetti.

Esempio 3: numero di insiemi (non ordinati) di k elementi che si possono formare con n oggetti? $\binom{n}{k}$, coefficiente binomiale. Dimostrazione usando i due esempi precedenti (detti $D_{k,n}$ e $C_{k,n}$ i numeri degli esempi 2 e 3, vale $D_{k,n} = C_{k,n} \bullet k!$).

Spazi finiti ($\Omega = \{\omega_1, ..., \omega_N\}$, probabilità descritta dai numeri $\{p_1, ..., p_N\}$ dati da $p_i = P(\omega_i)$), spazi equiprobabili ($p_i = \frac{1}{N}$, $P(A) = \frac{\sharp A}{N}$, dove $\sharp A$ indica il numero di elementi in A).

Per casa: esercizio del Lotto: vengono estratti cinque numeri diversi dai primi 90 numeri; non si considera l'ordine; che probabilità c'è di effettuare un ambo (cioè di dichiarare, prima dell'estrazione, due numeri che poi risultano estratti)?

Lezione 3 (1/3, 1 ora). Interpretazione con albero della formula di fattorizzazione e del calcolo della causa più probabile (nell'uso della formula di Bayes). Esercizi 1.1 del 22/7/03 e 5 del 17/5/03.

Lezione 4 (6/3, 2 ore). Argomenti previsti. Definizione di indipendenza tra due eventi A e B, prima a partire dalla probabilità condizionale ($P(A) = P(A|B)$), poi nella forma simmetrica $P(A \cap B) = P(A)P(B)$. Esercizio 1.3 del 22/7/03.

Complementi di calcolo combinatorico. In quanti modi si possono mettere k crocette in n caselle ordinate? $\binom{n}{k}$. Se chiamiamo "successi" le crocette, "esperimenti" le caselle, $\binom{n}{k}$ è il numero di modi con cui si possono realizzare k successi in n esperimenti.

Probabilità di k successi in n prove (esperimenti) indipendenti, ciascuna con probabilità di successo p:

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

$k = 0, 1, ..., n$. Lo abbiamo dimostrato nel seguente modo. Abbiamo preso un universo S formato dalle stringhe di n "zeri" o "uni", dove "uno" sta per successo e "zero" per insuccesso. La probabilità di una stringa con k successi è $p^k(1-p)^{n-k}$ (gli esperimenti sono indipendenti, quindi la probabilità che simultaneamente succendano certe cose nei vari esperimenti è il prodotto delle probabilità). Posto A l'insieme in S di tutte le stringhe con k successi, dobbiamo calcolare $P(A)$. Ma ciascun elemento di A ha probabilità $p^k(1-p)^{n-k}$ e ci sono $\binom{n}{k}$ elementi in A. Pertanto $P(A)$, essendo la somma delle probabilità dei suoi elementi, vale $\binom{n}{k} p^k (1-p)^{n-k}$.

Inizio dello studio dei capitoli 4 e 5 del Ross. Variabili aleatorie (per ora discrete). Alcuni esempi di variabili aleatorie emerse nell'esercizio 1.3 del 22/7/03. Definizione di v.a. di Bernoulli di parametro p (X che può assumere solo i valori 0 e 1, con $P(X = 1) = p$, $P(X = 0) = 1 - p$). Definizione di v.a. binomiale di parametri n e p: somma di n v.a. di Bernoulli indipendenti di parametro p, ovvero numero di successi in n prove indipendenti con probabilità di successo p in ciascuna prova; vale $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$, $k = 0, 1, ..., n$.

Lezione 5 (7/3, 2 ore). Esercitazione. 1.3 del 4/6/02 sulla distribuzione binomiale. 1.a e 1.b del 30/5/06 su probabilità condizionali, fattorizzazione e distribuzione binomiale. 1.1 e 1.2 del 21/6/06 su coefficiente binomiale e probabilità elementare. 1 del 17/5/06 risolto con l'albero. 1.1 del 9/2/1006 risolto con l'albero. Di quest'ultimo esercizio si possono già fare anche i punti due e tre.

Lezione 6 (8/3, 1 ora). Variabili aleatorie discrete: loro descrizione tramite i valori e la massa (o distribuzione) di probabilità. Rappresentazione grafica della massa di probabilità. V.a. di Bernoulli, binomiale, verifica della somma pari ad uno, grafico.

Definizione di v.a. di Poisson, verifica della somma pari ad uno, e teorema degli eventi rari (dimostrato).

Lezione 7 (13/3, 2 ore). Media aritmetica di un campione; valor medio (o atteso) di una v.a. discreta; suo legame intuitivo con la media aritmetica; definizione di varianza; proprietà del valor medio (linearità; ; esempi di valor medio e varianza per v.a. di Bernoulli, binomiali e di Poisson.

Lezione 8 (14/3, 2 ore). Definizione dei momenti. Funzione generatrice dei momenti (FGM) di una v.a. Definizione $\phi_X(t) = E[e^{tX}]$. Proprietà principali: le derivate in zero della $\phi_X(t)$ danno il valor medio delle potenze di X: $\phi'(0) = E[X]$, $\phi''(0) = E[X^2]$; la FGM della somma di due v.a. indipendenti è uguale al prodotto delle rispettive FGM. Dimostrazione di queste proprietà.Calcolo della generatrice per le v.a. di Bernoulli, binomiale e Poisson e calcolo di media e varianza di queste variabili tramite la generatrice. Verifica che la generatrice di una $B(n,p)$ tende a quella di una $\mathsf{P}(\lambda)$ nel regime del teorema degli eventi rari.

Lezione 9 (15/3, 1 ora). Proprietà della varianza; calcolo di valor medio e varianza per v.a. di Bernoulli, binomiali e di Poisson, senza la generatrice.

Lezione 10 (20/3, 2 ore). Esercizi vari su calcoli di valori medi, varianza, generatrice per alcune variabili discrete. Funzione di distribuzione (ripartizione) cumulativa: definizione e rappresentazione grafica.

Lezione 11 (21/3, 1 ora). Esercizi sulla cumulativa ed sui valori medi. In particolare, es. 2 del 4/4/03 ed alcuni da altri compitini di Aprile.

Lezione 12 (22/3, 1 ora). V.a. geometriche, proprietà di mancanza di memoria, esempio

del lotto (i numeri ritardatari hanno la stessa probabilità degli altri). Nascono nel problema dell'istante del primo successo in una sequenza di esperimenti indipendenti: una v.a. geometrica rappresenta tale primo istante. Queste cose non sono in programma ma solo facoltative. Invece, sono in programma le seguenti definizioni e risultati (che si suggerisce di copiare sul formulario): X è geometrica di parametro $p \in (0,1)$ se assume i valori $k = 1, 2, ...$ con probabilità

$$P(X = k) = pq^{k-1}.$$

Media, varianza e funzione generatirce sono

$$\mu = \frac{1}{p}, \quad \sigma^2 = \frac{q}{p^2}, \quad \phi(t) = \frac{pe^t}{1 - qe^t}.$$

Lezione 13 (27/3, 2 ore). Esercizi sul calcolo di probabilità e valori medi per v.a. discrete, e di loro trasformazioni.

V.a. continue: esempi a livello intuitivo; densità di probabilità, legame tra suoi integrali e probabilità associate alla v.a.

Esempio delle v.a. uniformi: calcolo della costante di normalizzazione e di alcune probabilità.

Esempio delle v.a. esponenziali di parametro λ: calcolo della costante di normalizzazione e di alcune probabilità. Calcolo della funzione di ripartizione cumulativa e della finzione di affidabilità $P(X > t) = e^{-\lambda t}$.

Lezione 14 (28/3, 2 ore). Esercizi sul calcolo di probabilità e valori medi per v.a. discrete, e di loro trasformazioni.

V.a. continue: valor medio. Definizione, osservazioni sulle possibili divergenze, calcolo per le v.a. uniformi ed esponenziali. Interpretazione grafica del risultato $\mu = \frac{1}{\lambda}$.

Lezione 15 (29/3, 1 ora). Esercitazione (compitino).

Lezione 16 (3/4, 2 ore). V.a. continue: valore medio di una trasformazione, $E[g(X)] = \int_{-\infty}^{+\infty} g(x)f(x)dx$. Formula per la varianza di una v.a. continua. Varianza di una v.a. esponenziale: $\frac{1}{\lambda^2}$. Somiglianza tra v.a. geometriche ed esponenziali. Mancanza di memoria dell'esponenziale (dimostrazione).

Funzione generatrice dei momenti per v.a. continue. Esercizi per casa: verificare che la generatrice di una $Exp(\lambda)$ è $\phi(t) = \frac{\lambda}{\lambda - t}$, definita solo per $t < \lambda$. Usarla per calcolare media e varianza. Verificare che la somma di due v.a. esponenziali indipendenti non è esponenziale.

Densità gaussiana standard. Funzione generatrice della gaussiana standard: $\phi(t) = e^{\frac{t^2}{2}}$. Verifica che la media è zero, la varianza uno. Densità gaussiana generica, con parametri μ e σ^2. Senza dimostrazione, viene affermato che la generatrice è $\phi(t) = e^{t\mu + \frac{t^2\sigma^2}{2}}$ e che da essa si verifica che μ e σ^2 sono media e varianza. Grafico della densità gaussiana generica, differenze rispetto alla standard.

Definizione della funzione $\Phi(x)$, funzione di ripartizione cumulativa della gaussiana standard.

Lezione 17 (4/4, 2 ore). Se $X \sim N(\mu, \sigma^2)$ allora la sua funzione di ripartizione cumulativa $F(x)$ si può calcolare come

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right)$$

(dimostrazione). Vale $\Phi(-x) = 1 - \Phi(x)$, utile per ricondursi alle tavole quando $-x < 0$. Esercizi vari di calcolo di probabilità del tipo $P(X < \lambda)$ e $P(X > \lambda)$ per $X \sim N(\mu, \sigma^2)$.

Definizione di quantile q_α e di z_α ($= q_{1-\alpha}$). Primi esercizi sul calcolo di quantili. Enunciata la proprietà $q_\alpha = -q_{1-\alpha}$, utile per ricondursi alle tavole quando $\alpha < 1/2$.

Teorema: X, Y gaussiane indipendenti, α, β, γ numeri reali, implica $\alpha X + \beta Y + \gamma$ gaussiana. Dimostrazione usando le generatrici. Media e varianza di $\alpha X + \beta Y + \gamma$ si calcolano anche in modo elementare, solo la gaussianità non è facile.

Lezione 18 (12/4, 1 ora). Come passare da una gaussiana generica alla gaussiana standard. Come calcolare la funzione $\Phi(x)$ con x positivo e negativo usando la tabella; come calcolare i quantili usando la tabella, per α maggiore di un mezzo e minore di un mezzo. Esercizi: per $X \sim N(5;16)$ calcolare $P(X > 9)$ e trovare λ tale che $P(X > \lambda) = 0.8$, Per $X \sim N(1;1)$ calcolare $P(X > 3)$. Per $X \sim N(5;4)$ calcolare $P(|X - 5| > 1)$. Per casa: data $X \sim N(7;8)$ trovare λ tale che $P(X > \lambda) = 0.1$.

Lezione 19 (17/4, 2 ore). Statistica: studieremo due problemi principali, la costruzione di un modello e la verifica di un modello. Per modello intendiamo la distribuzione di probabilità di una grandezza aleatoria. Es.: il prezzo a metro quadro degli alloggi di una certa zona. Prima di tutto si deve decidere la classe di distribuzioni che vogliamo usare (gaussiane, Poisson, ecc.). Nella realtà applicativa, questo passo è molto difficile; un aiuto viene dal tracciare un istogramma a partire da alcuni dati sperimentali. Se ad esempio abbiamo i valori del prezzo a metro quadro relativi a 50 appartamenti, possiamo usare questi dati per tracciare un istogramma. A volte esso suggerisce (entro certi limiti) una classe da usare. Scelta la classe, resta il problema di stimarne i parametri.

Indichiamo con $X_1, ..., X_n$ un campione (i valori che otterremo con n esperimenti o osservazioni; scriveremo invece $x_1, ..., x_n$ per i valori numerici specifici ottenuti in n esperimenti già eseguiti). Le v.a. $X_1, ..., X_n$ sono (per definizione di campione) indipendenti e con la stessa distribuzione. Indichiamo con μ e σ^2 la media e varianza, comune, di queste variabili. Per stimare la media μ si usa $\overline{X} = \frac{X_1 + ... + X_n}{n}$. Vale $E[\overline{X}] = \mu$ (veirificato), che si interpreta dicendo che $\overline{X}$ è uno stimatore corretto o non distorto di μ. Vale inoltre $Var[\overline{X}] = \frac{\sigma^2}{n}$, che interpretata anche graficamente fa capire che, per n grande, i valori più probabili di $\overline{X}$ saranno vicini a μ.

Inoltre, se le $X_1, ..., X_n$ sono gaussiane, allora lo è anche $\overline{X}$, cioè $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. A partire da questo fatto si ottiene subito

$$P(|\overline{X} - \mu| \leq \delta) = \Phi\left(\frac{\delta}{\sigma}\sqrt{n}\right) - \Phi\left(-\frac{\delta}{\sigma}\sqrt{n}\right).$$

Chiamando $1 - \alpha$ il termine a destra e risolvendo in δ si trova il risultato fondamentale

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

Questo viene espresso sinteticamente come $\mu = \overline{X} \pm \frac{\sigma q_{1-\frac{\alpha}{2}}}{\sqrt{n}}$ con probabilità $1 - \alpha$.
L'intervallo $\overline{X} \pm \frac{\sigma q_{1-\frac{\alpha}{2}}}{\sqrt{n}}$ si dice intervallo di confidenza.

Lezione 20 (18/4, 2 ore). Dopo aver riassunto i risultati precedenti e svolto un esercizio, si osserva che la *precisione* $\delta = \frac{\sigma q_{1-\frac{\alpha}{2}}}{\sqrt{n}}$ dell'intervallo di confidenza peggiora, nel senso che δ cresce, se si richiede un rischio α inferiore, cioè si richiede un risultato più sicuro (per irsultato intendiamo la dichiarazione dell'intervallo). Inoltre, la precisione aumenta nel senso che δ decresce, se si prende n più grande. Però va quadruplicata la numerosità per dimezzare δ.

Viene enunciato il teorema limite centrale, intepretata la standardizzazione. Viene poi utilizzato per arrivare all'affermazione: Se X è qualsiasi con media e varianza μ, σ^2,

sappiamo che $\mu = \overline{X} \pm \frac{\sigma q_{1-\frac{\alpha}{2}}}{\sqrt{n}}$ vale con probabilità che tende a $1-\alpha$ per $n \to \infty$. Quindi la teoria degli intervalli di confidenza fin qui vista si applica, almeno approssimativamente, a tutte le distribuzioni (per numerosità grandi).

Lezione 21 (19/4, 1 ora). Riassunto della situazione, prima di passare ai test. Il problema della stima nasce quando si ha un problema descritto da una grandezza aleatoria di cui non si conosce la distribuzione di probabilità. Bisogna scegliere la classe (es. gaussiane, Poisson,...) e ricondursi al problema di stimare i parametri. Sappiamo stimare il valor medio μ, tramite $\overline{X}$; sappiamo che $\overline{X}$ è uno stimatore non distorto di μ. Se X è gaussiana $N(\mu, \sigma^2)$, sappiamo che $\mu = \overline{X} \pm \frac{\sigma q_{1-\frac{\alpha}{2}}}{\sqrt{n}}$ con probabilità $1-\alpha$. Se X è qualsiasi con media e varianza μ, σ^2, sappiamo che $\mu = \overline{X} \pm \frac{\sigma q_{1-\frac{\alpha}{2}}}{\sqrt{n}}$ vale con probabilità che tende a $1-\alpha$ per $n \to \infty$ (purtroppo l'approssimazione non è controllata).

Fatte queste premesse, osserviamo che per molte classi di distribuzioni alcuni parametri sono o si riconducono al valor medio: per la Bernoulli p, p è μ; così per le Poisson e naturalmente per le gaussiane. Ma per le geometriche ed esponenziali il parametro è il reciproco della media: $\lambda = \frac{1}{\mu}$ per $Exp(\lambda)$, ad esempio. Quindi stimiamo λ con $\frac{1}{\overline{X}}$. Va bene, ma va osservato che $\frac{1}{\overline{X}}$ non è uno stimatore non distorto.

Altri parametri richiedono altri stimatori. La varianza σ^2 si può stimare con $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$, che è non distorto. Purtroppo è basato su μ che di solito è incognito quanto σ^2. E' naturale introdurre la variante $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ basata solo sul campione sperimentale, ma si dimostra che $E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right] = \frac{n-1}{n}\sigma^2$, cioè è distorto (di poco). Per questo si preferisce lo stimatore

$$S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

che è stimatore non distorto di σ^2. La sua radice viene poi presa come stima di σ.

Infine, osserviamo che le formule del tipo $\mu = \overline{X} \pm \frac{\sigma q_{1-\frac{\alpha}{2}}}{\sqrt{n}}$ contengono un analogo problema: σ è incognito (di solito) quanto μ. Allora si può ad esempio dire che, se $X \sim N(\mu, \sigma^2)$ con probabilità $1-\alpha$ vale approssimativamente $\mu = \overline{X} \pm \frac{S q_{1-\frac{\alpha}{2}}}{\sqrt{n}}$.

Vale inoltre una versione esatta di questo risultato: se $X \sim N(\mu, \sigma^2)$ con probabilità $1-\alpha$ vale $\mu = \overline{X} \pm \frac{S t^{(n-1)}_{1-\frac{\alpha}{2}}}{\sqrt{n}}$, dove $t^{(n)}_{\beta}$ indica il quantile t di Student a n gradi di libertà (di livello β). Anche i quantili t di Student sono tabulati.

Lezione 22 (24/4, 2 ore). Teoria dei test (verifica dei modelli). Si parte da un modello ipotizzato A e da un campione sperimentale $x_1, ..., x_n$; si vuole sottoporre a giudizio A sulla base del campione. Stategia naturale: si crea il modello B associato al campione e si confrontano A e B. Se sono sufficientemente diversi, si rifiuta il modello A, altrimenti si dichiara che non c'è contraddizione tra il campione ed il modello A.

Operativamente, nel caso di un modello gaussiano di varianza nota σ^2, e media ipotizzata μ_0, si fissa un livello α (numero piccolo in $(0,1)$, es. 0.05), calcolano $\bar{x}$ e $q_{1-\frac{\alpha}{2}}$, si controlla se $|\bar{x} - \mu_0| > \frac{\sigma q_{1-\frac{\alpha}{2}}}{\sqrt{n}}$; se accade ciò, si rifiuta l'ipotesi μ_0, altrimenti non la si può rifiutare.

Canonicamente, si esegue il test precedente controllando se $|z| > q_{1-\frac{\alpha}{2}}$, dove $z = \frac{\bar{x}-\mu_0}{\sigma}\sqrt{n}$.

Se la varianza non è nota, e si calcola al suo posto S^2 dal campione, si controlla se

$|t| > t_{1-\frac{\alpha}{2}}^{(n-1)}$, dove $t = \frac{\bar{x}-\mu_0}{S}\sqrt{n}$ e dove $t_{1-\frac{\alpha}{2}}^{(n-1)}$ è il quantile t di Student di livello $1 - \frac{\alpha}{2}$ con $n - 1$ gradi di libertà.

Viene svolto l'esercizio 3, parti i e ii, del 4/6/02.

Lezione 23 (26/4, 1 ora). Complementi sulla teoria della stima. Supponiamo di aver trovato il risultato $\mu = \bar{x} \pm \frac{\sigma q_{1-\frac{\alpha}{2}}}{\sqrt{n}}$ a livello di confidenza $1 - \alpha$, per la media μ di una gaussiana X di varianza nota σ^2. Supponiamo che venga chiesto di calcolare la probabilità che la gaussiana X assuma valori $> \lambda$, per un certo λ dato. Se la media fosse μ, varrebbe

$$P(X > \lambda) = 1 - \Phi\left(\frac{\lambda - \mu}{\sigma}\right).$$

Siccome non conosciamo μ ma abbiamo per essa un intervallo di confidenza, possiamo dire che:

a livello di confidenza $1 - \alpha$, $P(X > \lambda)$ è compresa nell'intervallo di estremi $1 - \Phi\left(\frac{\lambda-\mu^-}{\sigma}\right)$ e $1 - \Phi\left(\frac{\lambda-\mu^+}{\sigma}\right)$, dove $\mu^- = \bar{x} - \frac{\sigma q_{1-\frac{\alpha}{2}}}{\sqrt{n}}$, $\mu^+ = \bar{x} + \frac{\sigma q_{1-\frac{\alpha}{2}}}{\sqrt{n}}$. Viene svolto l'es. 3 del 26/1/06.

Complementi sulla teoria dei test. Nella lezione precedente abbiamo introdotto la strategia dei testi così: avendo un'ipotesi, es. media μ_0 ed un campione sperimentale $x_1, ..., x_n$, dal campione si costruisce un modello, con l'intervallo di confidenza per μ, e si rifiuta l'ipotesi μ_0 se essa non sta nell'intervallo di confidenza. Prendiamo ora un altro punto di vista, più legato alla struttura logica $(A \Rightarrow B) \Leftrightarrow (nonB \Rightarrow nonA)$. Supponiamo al solito di avere l'ipotesi di media μ_0 ed un campione sperimentale $x_1, ..., x_n$. Se vale l'ipotesi (A) allora valgono certe conseguenze (B), a meno di una piccola probabilità. Basta allora controllare se il campione soddisfa le conseguenze: se $x_1, ..., x_n$ non soddisfa le conseguenze (*nonB*) allora concludiamo che non vale l'ipotesi di media μ_0 (*nonA*). Ci sono però tante possibili conseguenze dell'ipotesi di media μ_0. Il modo di valutare quale conseguenza sia più utile è il concetto di potenza di un test.

Lezione 24 (2/5, 2 ore). Come abbiamo detto, si può costruire un test seguendo il ragionamento: se vale l'ipotesi di media μ_0 allora valgono certe conseguenze, a meno di una piccola probabilità; basta allora controllare se il campione soddisfa o meno le conseguenze. Ci sono però tante possibili conseguenze dell'ipotesi di media μ_0. Una è quella già utilizzata, ovvero che $|\overline{X} - \mu_0| \leq \frac{\sigma q_{1-\frac{\alpha}{2}}}{\sqrt{n}}$ (salvo che con probabilità α). Ma ad esempio un'altra è $|X_1 - \mu_0| \leq \sigma q_{1-\frac{\alpha}{2}}$ (salvo che con probabilità α), cioè una condizione solo sul primo elemento del campione X_1. Il test basato sulla prima condizione ora descritta è quello delle lezioni precedenti: se il campione $x_1, ..., x_n$ soddisfa $|\bar{x} - \mu_0| > \frac{\sigma q_{1-\frac{\alpha}{2}}}{\sqrt{n}}$ (che canonicamente verifichiamo nella forma $|z| > q_{1-\frac{\alpha}{2}}$), rifiutiamo l'ipotesi. Possiamo però costurire un test sulla seconda conseguenza: dato il campione $x_1, ..., x_n$, se $|x_1 - \mu_0| > \sigma q_{1-\frac{\alpha}{2}}$, allora rifiutiamo l'ipotesi. Quale test è migliore?

Definizione di potenza. Calcolo della potenza per i due test e verifica che il migliore dei due è quello basato su $\bar{x}$.

Linguaggio della teoria dei test: ipotesi nulla, ipotesi alternativa, livello o significatività del test, errore di prima specie, errore di seconda specie, potenza. p-value.

Esercizio 3 del 19/9/06 e 4 del 17/5/06.

Lezione 25 (3/5, 1 ora). Abbiamo compeltato le basi di statistica su stime e test. Tra gli argomenti che dovremo ancora sviluppare ci sono: DOE, regressione, carte di controllo e soglie, altri esempi di stime e test.

DOE (Design Of Experiments, progettazione di esperimenti) è una teoria che si occupa di vari aspetti progettuali legati alla statistica. Uno dei più semplici è il seguente, che

illustriamo con un esempio. Le FFSS vogliono esaminare il traffico passeggeri Roma-Milano del venerdì pomeriggio per capire se è strutturalmente superiore al servizio offerto e ricalibrare il servizio. Descriviamo il numero di passeggeri in quella tratta in quel momento della settimana con una v.a. *N*. Dobbiamo decidere il tipo di distribuzione. *N* assume valori interi positivi; se assumesse valori mediamente bassi, sarebbe conveniente usare Poisson, binomiali, geometriche o altre distribuzioni discrete. Dal momento che assume valori molto alti, è anche ragionevole, se non addirittura molto migliore, usare distribuzioni continue. Scegliamo ora le gaussiane, per semplicità. Decidiamo quindi di descrivere *N* con una $N(\mu, \sigma^2)$. Come conoscere μ e σ^2 ? Attraverso osservazioni settimanali del traffico, poi usando la teoria della stima. In questa fase, di progettazione delle osservazioni ("esperimenti"), che ragionamenti ha senso fare? Ad esempio, si può decidere che precisione δ si vuole nella stima di μ e con quale livello $1 - \alpha$ di confidenza; scelti questi due valori, es. $\delta = 100$ passeggeri, $\alpha = 0.05$, ci chiediamo: quante osservazioni dobbiamo eseguire? Questo è un esempio semplice di DOE. La risposta (teorica) si ottiene invertendo la formula $\delta = \frac{\sigma q_{1-\frac{\alpha}{2}}}{\sqrt{n}}$.

Dal punto di vista pratico c'è il problema della mancata conoscenza di σ. Si può ad esempio procedere ad un primo insieme di osservazioni e stimare σ (mediante lo stimatore *S*) con tali campioni; se poi la numerosità richiesta eccede il numero di osservazioni già eseguite, si devono continuare le osservazioni per il necessario. Oppure si può decidere che un certo numero di altre linee e momenti della settimana è simile alla tratta sotto studio, raccogliere i dati di quelle linee in una sola settimana, ed usarli per stimare σ.

Un altro problema pratico è che in genere il valore di *n* che si trova appare piuttosto alto e d'altra parte le richieste su δ e α non sono così imperative ma possono essere parzialmente modificate. Quindi è naturale tabulare i valori di *n* al variare di δ e α (come vedremo meglio in seguito) per operare delle scelte.

Si suggerisce di collegarsi al sito www.istat.it, link "Tavole di dati" (sulla sinistra), link "La ricerca e lo sviluppo in Italia", aprire i dati (sulla destra in alto) e copiare su un foglio Excel la tavola 1.6. Provare ad affrontare il seguente problema: c'è differenza tra le spese per la ricerca universitaria negli anni 2002 e 2004?

Lezione 26 (8/5, 2 ore). Cenni sull'uso di Excel per analisi statistiche. Dopo aver copiato su un foglio Excel le due serie di dati descritte nella lezione precedente, si può calcolare media e deviazione stazionaria campionarie, $\bar{x}$ ed *S*, coi relativi comandi di Excel (scelta una casella, digitato =, si clicca su funzione, poi su funzioni statistiche, poi su media e si inseriscono i dati). Interessante è fare un istogramma. Non con l'opzione grafica di istogramma, che è una semplice funzione; ma attraverso il pacchetto Analisi Dati che si trova sotto Strumenti. Se non c'è, va caricato: da Strumenti si entra in Componenti aggiuntive e si prende Strumenti di Analisi. A quel punto è caricato e visibile sotto Strumenti. Con la funzione Istogramma di Analisi Dati si ottiene una tabella fatta dei punti di suddivisione e della numerosità di eventi in ciascun intervallo, poi la si visualizza con un grafico Istogramma.

Esercizio: realizzare un istogramma cumulativo: la spezzata costante a tratti che sale di $\frac{1}{n}$ (dove *n* è la numerosità del campione) in corrispondenza di ogni valore della *x* in cui c'è un dato sperimentale).

Circa l'esempio della lezione precedente, vengono esaminati visivamente i valori, si riconosce una modesta crescita ad occhio con alcune eccezioni, si calcolano media e deviazione di ciascun gruppo ($\bar{x}_1 = 239$, $\bar{x}_2 = 250$, $S_1 = 203$, $S_2 = 209$). Vengono visualizzati gli istogrammi di entrambe. Per capire se c'è stato un aumento nella spesa per R&S universitaria si descrivono due strategie.

La prima si basa sull'idea che dalla prima serie (2002) si ricava un modello, che per semplicità è stato preso come una gaussiana di media $\mu_0 = 239$ e deviazione $\sigma = 203$. Poi

si confronta il secondo campione col modello, eseguendo un test per la media. Si calcola $z = \frac{\bar{x}_2 - \mu_0}{\sigma}\sqrt{n}$ e si confronta col quantile. La scelta della significatività α è soggettiva; prendiamo il valore usuale 0.05. Vale $q_{0.975} = 1.96$, mentre *z* vale circa 0.23. Quindi il test osserva che non c'è contraddizione tra il campione e l'ipotesi, cioè non c'è motivo di ritenere che siano aumentate le spese per R&S universitarie.

Un secondo metodo consiste nell'introdurre la v.a. Y = incremento di spesa tra il 2002 e il 2004. I suoi valori sperimentali sono dati dalle differenze tra i valori 2004 e quelli 2002. La media campionaria ora è $\bar{x}_{incr} = 11$, che non è altro però che $\bar{x}_2 - \bar{x}_1$, quindi anche $\bar{x}_2 - \mu_0$. L'ipotesi nulla da sottoporre a test ora è che l'icremento sia una v.a. gaussiana con media $\mu_0^{incr} = 0$. Quindi il test consiste nel confrontare $z_{incr} = \frac{\bar{x}_{incr} - \mu_0^{incr}}{\sigma_{incr}}\sqrt{n}$ col quantile. Il numeratore $(\bar{x}_{incr} - \mu_0^{incr}) \bullet \sqrt{n}$ coincide col numeratore della *z* del metodo precedente: $(\bar{x}_2 - \mu_0) \bullet \sqrt{n}$. Invece σ_{incr} è un pò cambiata. Infatti essendo $Y = X_2 - X_1$, se ipotizziamo X_1 e X_2 indipendenti, vale $Var[Y] = Var[X_1] + Var[X_2]$, quindi approssimativamente $\sigma_{incr} \sim \sqrt{S_1^2 + S_2^2} \sim \sqrt{2}\,S_1$. Quindi z_{incr} è circa $\frac{1}{\sqrt{2}}$ più grande di *z*, ma questo non basta a modificare l'esito del test. In realtà le considerazioni fatte su $Var[Y]$ sono discutibili (l'indipendenza), descritte qui a titolo di esempio di ragionamento. Alternativamente, è meglio calcolare σ_{incr} dal campione degli incrementi (e magari usare il quantile *t* di Student per essere più precisi).

Esercizio per casa: nello stesso sito, prendere i dati della spesa per R&S delle imprese 2002 e vedere se c'è un legame (una "correlazione") con la spesa universitaria.

Lezione 27 (10/5, 1 ora). Vengono visualizzati i dati di spesa per R&S di università (*X*) e imprese (*Y*) 2002, in un grafico. Vengono discusse alcune impressioni intuitive: 1) i dati sono posizionati un po' a cono, non proprio nelle vicinanze di una retta ma meglio di più di una retta; 2) forse sembrano posizionati a parabola; 3) c'è un certo grado di clusterizzazione (raggruppamento): due cluster ben separati, più due punti isolati; 4) se si eliminassero Piemonte e Lombardia, i dati sarebbero molto più vicini ad una singola retta, non più parabolici, raggruppati in due cluster. I due dati di Piemonte e Lombardia vengono detti Outliers (punti anomali), o punti influenti (o che fanno leva) in quanto modificano radicalmente il risultato.

Viene calcolata la retta di regressione dei dati tutti insieme. Si presuppone che i dati seguano la legge $Y = aX + b + errore$, si stima *a* col comando Excel "Pendenza", *b* col comando Excel "Intercetta", ottenendo i valori $\hat{a} \sim 2$, $\hat{b} \sim -120$. Si visualizza questa retta di regressione sul grafico, e si immagina quale sarebbe stata senza Piemonte e Lombardia.

Si discutono alcuni difetti delle analisi appena fatte, in particolare il fatto che i dati sono fortemente disomogenei a causa delle grandi differenze di ampiezza delle regioni italiane. Ad esempio, il fatto che per certe regioni la spesa sia sensibilmente inferiore alla media nazionale (o viceversa) è più che altro dovuto alla piccolezza della regione, o per lo meno gioca anche tale fattore. Questo ha sicuramente falsato il confronto 2002-2004: le deviazioni standard erano enormi a causa della disomogeneità dei dati e, comparendo a denominatore di *z*, lo hanno reso così piccolo da essere troppo distante dal quantile. Analogamente, la clusterizzazione appena descritta forse è dovuta più alle dimensioni della regione che all'entità della spesa (cioè non è interpretabile come regioni "buone" o "cattive" in quanto a politiche universitarie, ma solo regioni grandi o piccole). Si capisce quindi che andrebbero normalizzati (standardizzati, uniformati) i dati usando ad esempio la popolazione regionale. Si decide di usare come nuove grandezze la spesa pro capite.

Esercizi per casa: 1) calcolare $\hat{a}$ e $\hat{b}$ senza Piemonte e Lombardia; visualizzare tutto su Excel; 2) cercare sul sito ISTAT i dati sulla popolazione regionale, calcolare la spesa pro capite e ripetere con essa le cose fatte sui dati di spesa universitaria 2002-2004 e spesa università-imprese 2002.

Lezione 28 (15/5, 2 ore). Regressione lineare: vedi brevi dispense in proposito, su aspetti teorici ed implementazione con Excel.

Lezione 29 (16/5, 2 ore). Completamento delle brevi dispense su regressione lineare, escluso per ora il calcolo degli stimatori per $r > 1$.

Richiamo sulla varianza e sue proprietà. Definizione di covarianza, sue proprietà (in parte simili a quelle della varianza); linearità nei due argomenti, covarianza nulla per v.a. indipendenti, e teorema: se $Y = aX + b + \varepsilon$ con X ed ε indipendenti, allora $Cov(X, Y) = aVar[X]$. Ne emerge che la covarianza misura il legame lineare tra variabili aleatorie. Però l'entità numerica dipende dall'unità di misura. Dato un campione, calcolata la covarianza empirica $\widehat{Cov}(X, Y) = \frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})$, che non sarà mai esattamente zero, come giudicare se è circa zero oppure no, a causa dell'unità di misura? Allora è utile una modifica della covarianza, ovvero il coefficiente di correlazione $\rho(X, Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$. Anch'esso è nullo per v.a. indipendenti, varia però tra -1 e 1, non dipende dall'unità di misura ($\rho(\lambda X, Y) = \rho(X, Y)$ per ogni $\lambda > 0$), e se vale un legame lineare esatto $Y = aX + b$ risulta $\rho(X, Y) = 1$ se la retta è inclinata positivamente ($a > 0$), $\rho(X, Y) = -1$ se la retta è inclinata negativamente ($a < 0$). L'impressione grafica è che una nuvola disordinata pressapoco circolare di punti ha correlazione molto vicina a zero, es. 0.1, una nuvola abbastanza allungata attorno ad una retta inclinata positivamente ha correlazione ad es. 0.9, una nuvola abbastanza allungata attorno ad una retta inclinata positivamente ha correlazione ad es. -0.9.

La correlazione può essere usata per tanti scopi. Uno è diretto: capire se certe grandezze sono legate oppure no. Ad esempio, le spese per R&S di università e imprese, normalizzate per numero di abitanti, sono correlate o no? A priori non è chiaro.

Nota: con Excel è facilissimo calcolare la correlazione (c'è il comando tra le funzioni).

Un secondo uso è per capire se, dato un campione $(x_1, y_1), ..., (x_n, y_n)$, è più ragionevole un modello lineare oppure uno nonlineare. Infatti si può trasformare la stringa delle x tramite una funzione f, ottenendo così le coppie $(f(x_1), y_1), ..., (f(x_n), y_n)$; si può poi calcolare la correlazione sia delle coppie originarie sia di quelle trasformate; se ad esempio capita che è maggiore (in valore assoluto) la correlazione delle coppie trasformate, vuol dire che è più fedele il modello nonlineare $Y = af(X) + b + \varepsilon$. Il preblema pratico di questi ragionamenti sulle trasformazioni è trovare una buona trasformazione, cosa per cui non ci sono regole, ma solo tentativi basati sulla visualizzaizone dei dati.

Un terzo uso dei covarianza e correlazione è nella teoria della regressione, come si vede nelle brevi dispense sulla regressione.

Lezione 30 (17/5, 12 ora). Esercizi su stime e test gaussiani e non gaussiani, test unilaterali, valore p.

Lezione 31 (22/5, 2 ore). Vengono consolidati i test unilaterali tramite un esercizio: descrizione delle motivazioni alla base della strategia del test, esecuzione del test, calcolo del valore p, calcolo della potenza (unilaterale).

Lezione 32 (23/5, 2 ore). Viene consolidato l'uso del TLC allo scopo di eseguire test (e stime) con calcoli gaussiani per variabili qualsiasi, in modo approssimato. Viene applicato al caso della Bernoulli e della binomiale.

Vengono consolidati alcuni calcoli con v.a. gaussiane: calcolo di probabilità e di quantili o soglie.

Lezione 33 (24/5, 12 ora). Esercitazione valida per l'esonero dalla prova scritta.

Per le regole della prova orale si rimanda alla pagina del docente.