**Statistica I, Ing. Gestionale, a.a. 2009/10**
**Registro delle lezioni**

**Lezione 1** (2/3). Introduzione al corso; materiale e comunicazioni alla pag. di F. Flandoli, http://www2.ing.unipi.it/~a008484/dispStatisticaGestionaleTriennale.html. Si osserva che ci sono tre corsi a carattere statistico nel percorso di Gestionale a Pisa, ovvero Statistica I alla triennale, Statistica II alla magistrale, Statistica Applicata (Prof. Lanzetta) tra i facoltativi della triennale, corsi coordinati ta loro.

La statistica si può grosso modo dividere in Statistica descrittiva da un lato, e Statistica basata sul Calcolo delle Probabilità dall'altro. Il capitolo 2 del Ross è dedicato al primo argomento. Si consiglia la lettura del paragrafo 2.2 in relazione ai temi trattati in questa lezione. Gli altri capitoli puntano alla Statistica basata sul Calcolo delle Probabilità, sempre aiutandosi comunque con le idee pratiche e grafiche della statistica descrittiva.

Si suggerisce un po' di uso del programma di calcolo `R` per capire più in conceto alcuni argomenti del corso.

Esercizio. i) Scaricare `R` da rete (basta cercare R stat con Google). ii) rintracciare le temperature di alcune città italiane e scriverle nel comando

`T<-c(.,.....,.)`

iii) eseguire i comandi `mean(T)`, `sd(T)`, `hist(T)`, `hist(T,k)` per qualche k, `hist(T,freq=TRUE)`.

Vengono descritte le formule per la media $\overline{x}$ e la deviazione standard $S$ di un campione $x_1, ..., x_n$

$$\overline{x} = \frac{x_1 + ... + x_n}{n}, \quad S^2 = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})^2 .$$

e la variante

$$\frac{1}{n} \sum_{i=1}^{n} |x_i - \overline{x}|$$

(che potremmo chiamare "media delle distanze dalla media"). Entrambi sono indicatori che quantificano l'*incertezza* (la dispersione, la casualità, l'aleatorietà, la variabilità, dei dati ecc.).

Viene spiegata la differenza tra istogramma assoluto (quello di `hist(T)`) e relativo (quello di `hist(T,freq=FALSE)`).

Viene evidenziato l'aspetto contingente degli istogrammi, dipendente dai dati specifici o altri dettagli come il numero di dati o il numero di classi.

Emerge come naturale l'idea che dietro agli istogrammi coi loro dettagli accidentali ci sia una curva più regolare che descrive il fenomeno fisico. Questo introduce il concetto di densità di probabilità (pdf) (definizione 4.2.2). Viene introdotta la densità gaussiana standard (par. 5.5). La specifica `freq=FALSE` negli istogrammi serve ad ottenere area uno sotto l'istogramma, in modo che sia confrontabile con la densità, es. gaussiana.

Viene descritto (a titolo facoltativo) il calcolo con cui si verifica che $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ è una pdf:

$$\left( \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx \right)^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} e^{-\frac{y^2}{2}} dx dy = \int_0^{2\pi} d\theta \int_0^{\infty} e^{-\frac{\rho^2}{2}} \rho d\rho$$
$$= 2\pi \int_0^{\infty} e^{-t} dt = 2\pi.$$

**Lezione 2** (3/3). Cenno di studio di funzione per riconoscere il grafico di $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. Comandi di `R` per ottenenerlo:

```
X=-4+(1:8000)/1000
Y=dnorm(X,0,1)
plot(X,Y)
```

Gaussiane qualsiasi, di parametri $\mu$ e $\sigma$: grafico, somiglianze e differenze rispetto al caso standard, in particolare il ruolo di $\sigma$ (esempi grafici), verifica della proprietà di area uno (col cambio di variabile $y = \frac{x-\mu}{\sigma}$). La trasformazione $y = \frac{x-\mu}{\sigma}$: significato grafico, standardizzazione.

Come scegliere $\mu$ e $\sigma$ a partire da un campione sperimentale $x_1, ..., x_n$? L'idea è che $\mu$ e $\sigma$ (che immaginiamo essere i parametri veri di un modello) sono approssimati dai numeri empirici (sperimentali) $\overline{x}$ ed $S$. Riconoscimento che $\sigma$ misura la dispersione del grafico gaussiano.

**Lezione 3** (4/3). Richiamo su quanto visto fino ad ora: dato un campione sperimentale $x_1, ..., x_n$, da esso possiamo calcolare un istogramma, $\overline{x}$ ed $S$, e con questi costruire una gaussiana.

A cosa può servire? Immaginiamo ad esempio di vendere un certo prodotto e voler tenere una riserva per accontentare tutti o quasi i clienti. Quanto deve essere ampia la riserva? Se disponiamo di un campione delle richieste passate della clientela, relative a quel prodotto, possiamo estrapolare una densità (per ora gaussiana) nel modo detto sopra, e calcolare il numero $\lambda$ alla cui destra sta un'area di ampiezza da noi decisa a priori, es. 5%. La probabilità che le richieste superino $\lambda$ è il 5%. Nel 95% dei casi le richeiste saranno inferiori

a $\lambda$. Questo è il valore che terremo nella riserva, se abbiamo deciso a priori di correre un rischio del 5%. Vedremo meglio più avanti che tale valore si chiama quantile di ordine 0.95 e si calcola col comando `qnorm(0.95,m,s)`.

Si noti per inciso che disponendo di un campione di numerisotà 100, avrebbe senso evitare le gaussiane e prendere come $\lambda$ un numero alla cui destra stanno gli ultimi 5 valori più grandi del campione. Fatte le dovute proporzioni, si può usare lo stesso metodo con campioni di numerosità diversa, ma se la numerosità è troppo bassa diventano troppo pochi i valori alla destra di $\lambda$ ed il metodo perde di significato o affidabilità.

Si inizia lo studio del capitolo 3, sui fondamenti del calcolo delle probabilità. Dopo le definizioni di universo, eventi, eventi elementari e probabilità, e l'illustrazione delle varie regole (esemplificate intuitivamente col concetto di massa), si torna alle pdf, prendendo come esempio di universo tutto $\mathbb{R}$, come eventi gli intervalli o insiemi più generici e come probabilità di un insieme $A$ l'area sottesa dalla pdf su $A$, ovvero $\int_A f(x)\,dx$. Le regole delle probabilità sono verificate. L'evento "le richieste superano $\lambda$" è un esempio.

Viene poi descritto il concetto di probabilità uniforme su uno spazio finito, esemplificandolo col problema del lancio di due dadi. Si confrontano i due esempi (finito e continuo): nel caso finito i singoli eventi elementari hanno probabilità non nulla e che genera la probabilità di ogni altro evento, per somma; nel caso continuo ogni punto ha probabilità nulla, quindi non è significativo.

Si conclude con un esempio speciale di spazio finito. Si considerano $n$ persone, ciascuna che può scegliere una possibilità A o B, codificate con 1 o 0. Ciascuna persona sceglie 1 con probabilità $p$, uguale per tutte le persone. Gli eventi elementari di $S$ sono le stringhe di 0 e 1. Se una stringa ha $k$ numeri uguali a 1, la sua probabilità è

$$p^k (1-p)^{n-k}.$$

Capiremo meglio questa formula del prodotto e trarremo le conseguenze di questo calcolo.

**Lezione 4** (9/3). Probabilità condizionale, interpretazione intuitiva e grafica. Indipendenza tra eventi, legame con la probabilità condizionale ($A$ e $B$ sono indipendenti se e solo se $P(A|B) = P(A)$, se e solo se $P(B|A) = P(B)$, quando $P(B) > 0$ e $P(A) > 0$). Differenza rispetto al concetto di eventi disgiunti. Spiegazione della formula $p^k (1-p)^{n-k}$, tramite l'indipendenza.

Formula di fattorizzazione.

Esercizio: vendite di vino a F (50%) e G (50%); le vendite a F riguardano vino B per i 3/4; le vendite a G riguardano vino B per 1/4. Che probabilità c'è di vendere vino B? (Soluzione con la formula delle probabilità totali).

Calcolo combinatorio: principio di enumerazione, disposizioni, permutazioni e combinazioni.

**Lezione 5** (10/3). Formula di Bayes.

Esercizio. Se una fotocopiatrice può fare cattive fotocopie sia per surriscaldamento (S) sia per dispersione di toner (D), se tra i due problemi S accade il 70% delle volte e D il 30%, se quando accade S le fotocopie vengono male il 20% mentre se accade D l'80% delle volte, quando osserviamo una fotocopia venuta male, è più probabile che sia a causa di S o D?

Esercizio. Se una banca ha 100 correntisti, indipendenti, ciascuno che si presenta durante un giorno con probabilità $p$, che probabilità c'è che si presentino 20 correntisti? Ricordando che $p^{20}(1-p)^{100-20}$ è la probabilità di una ben precisa ventina di correntisti, e che il numero di ventine è $\binom{100}{20}$, la probabilità richiesta è $\binom{100}{20}p^{20}(1-p)^{100-20}$.

**Lezione 6** (11/3). Concetto di variabile aleatoria, esposto attraverso esempi (es. tempo di vita, quantità venduta in futuro, costo futuro di un bene). V.a. discrete e continue. Rappresentazione grafica di v.a. discrete, tramite lista dei valori e delle loro probabilità. La successione $(p_k)$ viene detta funzione massa di probabilità. Devono essere numeri $\geq 0$ a somma 1. Grafico della massa.

Esempi: Bernoulli e binomiale (verifica di somma 1 usando il binomio di Newton). Simbolo $X \sim B(n,p)$ (Bernoulli: $B(1,p)$). Vedremo un altro legame tra i due. Si rammentano le formule

$$\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k!} = \frac{n!}{k!(n-k)!}$$

ed i casi particolari $\binom{n}{0} = 1$, $\binom{n}{n} = 1$, $\binom{n}{1} = n$, $\binom{n}{n-1} = n$, usati per esemplificare $p_0, p_1, p_{n-1}, p_n$ della binomiale. Diversi grafici della binomiale, a seconda del valore di $p$. Si rammenta che il numero $p_k = \binom{n}{k}p^k(1-p)^{n-k}$ è la probabilità che in una sequenza lunga $n$ di 0 e 1 ci siano $k$ valori 1, supponendo indipendenti i valori e probabilità $p$ di avere 1.

V.a. continue, funzione densità di probabilità (pdf), $P(X \in I) = \int_I f(x)\,dx$.

Esempi: uniforme su $[a,b]$, esponenziale di parametro $\lambda$. In entrambi i casi abbiamo calcolato una certa costante imponendo area 1.

Funzione di ripartizione (detta anche funzione di distribuzione cumulativa, cumulative distribution function, cdf): $F(t) = P(X \leq t)$. Ha senso

sia per le discrete sia per le continue, cioè unifica un po' le teorie. Nel caso continuo si lega alla pdf $f$ in due modi:

$$F(t) = \int_0^t f(x)\,dx$$

e

$$F'(t) = f(t)$$

nei punti in cui $f$ è continua, e quindi $F$ è derivabile.

**Lezione 7** (16/3). Definizione di valor medio per v.a. discrete. Esempio $B(1,p)$, esempio di variabile che vale $-1, 0, 1$ con ugual probabilità ed altri esempi. Interpretazione grafica (media pesata dei valori, baricentro, centro di massa, comunque sull'asse delle ascisse). Esempio: $B(n,p)$ con calcoli espliciti; cenno:

$$\begin{aligned} E[X] &= \sum k \binom{n}{k} p^k (1-p)^{n-k} = \sum \frac{n!}{(k-1)!\,(n-k)!} p^k (1-p)^{n-k} \\ &= n \sum \frac{(n-1)!}{(k-1)!\,(n-k)!} p^k (1-p)^{n-k} \\ &= np \sum \frac{(n-1)!}{(k-1)!\,(n-k)!} p^{k-1} (1-p)^{n-k} = np\,(p+(1-p))^{n-1} = np. \end{aligned}$$

Richiamo sul concetto di media aritmetica $\overline{x} = \frac{x_1+\ldots+x_n}{n}$ di un campione sperimentale. Legame tra i due concetti:

$$\begin{aligned} \frac{x_1+\ldots+x_n}{n} &= \frac{(0+\ldots+0)+\ldots+(k+\ldots+k)+\ldots}{n} \\ &= \frac{n_0}{n}\cdot 0 + \ldots + \frac{n_k}{n}\cdot k + \ldots \end{aligned}$$

dove $n_k$ è il numero di elementi $x_i$ uguali a $k$. Se ipotizziamo che le frequenze empiriche $\frac{n_k}{n}$ (quelle degli istogrammi) siano prossime alle probabilità teoriche $p_k$ (quelle della funzione massa di probabilità), allora la somma precedente è circa uguale a

$$p_0 \cdot 0 + \ldots + p_k \cdot k + \ldots = E[X].$$

Definizione di v.a. indipendenti:

$$P(X_1 \in I_1, \ldots, X_n \in I_n) = \prod_{i=1}^{n} P(X_i \in I_i)$$

per ogni sequenza di insiemi $I_1, ..., I_n$. Se sono v.a. discrete, è sufficiente prendere come insiemi $I_1, ..., I_n$ i punti:

$$P(X_1 = a_1, ..., X_n = a_n) = \prod_{i=1}^{n} P(X_i = a_i)$$

per ogni sequenza di punti $a_1, ..., a_n$.

Teorema 1: se $X_1, ..., X_n$ sono indipendeti $B(1, p)$, allora $S = X_1 + ... + X_n$ è una $B(n, p)$. Dimostrazione: si osserva che l'evento $(S = k)$ è l'unione degli eventi che specificano esattamente quali $k$ delle $X_i$ sono uguali a 1. Ciascuno di tali eventi, ad es.

$$(X_1 = 1, ..., X_k = 1, X_{k+1} = 0, ..., X_n = 0)$$

per l'indipendenza ha probabilità data dal prodotto, ad es.

$$P(...) = \prod_{i=1}^{n} P(X_i = a_i) = p^k (1-p)^{n-k}$$

essendoci $k$ valori $a_i$ uguali ad 1. Il numero degli eventi di questo tipo è $\binom{n}{k}$. Quindi la probabilità complessiva è la somma di $\binom{n}{k}$ valori di probabilità, tutti però uguali a $p^k (1-p)^{n-k}$, quindi $\binom{n}{k} p^k (1-p)^{n-k}$.

Teorema 2: linearità del valor medio. Osservazione: non richiede l'ipotesi di indipendenza.

Teorema 3: se $X, Y$ sono indipendenti, allora $E[XY] = E[X] E[Y]$. Osservazione: il viceversa non è vero: da questa singola condizione non si riesce a dedurre la sequenza di condizioni che definisce l'indipendenza.

Esempio: dimostrazione che per $X \sim B(n, p)$ vale $E[X] = np$, usando i teoremi 1 e 2.

**Lezione 8** (17/3). Valor medio per v.a. continue. Esempi: uniforme, esponenziale, gaussiana. Interpretazione grafica. Definizioni di mediana e moda, osservazioni sul legame con la media.

Valor medio di trasformazioni: esempio di $E[X^2]$ e $E[X^3]$ in un caso discreto.

**Lezione 9** (18/3). Valor medio di una trasformazione:

$$E[\varphi(X)] = \sum \varphi(a_k) p_k, \qquad E[\varphi(X)] = \int_{-\infty}^{\infty} \varphi(x) f(x) dx$$

nei dua casi, continuo e discreto. Casi particolari rilevanti: momenti di ordine $k$, $E\left[X^k\right]$, e varianza

$$Var\left[X\right] = E\left[\left(X-\mu\right)^2\right]$$

dove $\mu = E\left[X\right]$. E' lo scarto quadratico medio. Media di una costante: uguale alla costante. Usando questo e la linearità si trova la formula $Var\left[X\right] = E\left[X^2\right] - \mu^2$. Note: $E\left[X^2\right] \geq \mu^2$ sempre; $E\left[X^2\right] = \mu^2$ solo se $X$ è costante (quindi uguale alla sua media).

Calcolo della varianza per la Bernoulli.

Definizione di covarianza; generalizza la varianza; definizione di v.a. scorrelate; teorema: indipendenti implica scorrelate (equivalente al teorema secondo cui indipendenti implica $E\left[XY\right] = E\left[X\right]E\left[Y\right]$). Formula generale per la varianza della somma $X+Y$; caso particolare quando le v.a. sono indipendenti. Utilità nella gestione di un portfolio (diversificazione degli investimenti tra titoli indipendenti).

Calcolo della varianza per la binomiale. Implicazioni per esempi con grande numero di utenti indipendenti: piccolo scarto rispetto alla media.

**Lezione 10** (23/3). Esercizio 1 del compito di MMS del 6/6/2007, parti i), ii), iii), iv), v). Per completare gli ultimi dettagli di alcune domande serve conoscere la varianza di esponenziali e gaussiane, la generatrice delle esponenziali, il calcolo di probabilità gaussiane con le tavole, che vedremo.

Funzione generatrice dei momenti, come si calcolano i momenti, Teorema sulla generatrice della somma di v.a. indipendenti (con dimostrazione), esempi di Bernoulli, binomiale, gaussiana.

**Lezione 11** (24/3). Teorema: $X, Y$ gaussiane indipendenti, $a, b, c$ numeri reali, implica $aX + bY + c$ gaussiana (con dimostrazione). Calcolo della sua media e varianza. Standardizzazione, in generale e per le gaussiane; dalle gaussiane qualsiasi alle canoniche.

**Lezione 12** (25/3). Funzione di ripartizione, grafico tipico, uso per calcolare probabilità; operazione inversa: quantili; trasformazione nel caso gaussiano, dalla $F$ alla $\Phi$ canonica; calcolo di probabilità e quantili tramite le tavole; esercizi. Formule del tipo $\mu \pm \sigma q_\alpha$ per le soglie (es. scorte di magazzino), assegnato il rischio.

Cenno alle V.a. di Poisson, media e varianza (per ora senza dimostrazione e senza generatrice). Tabella di media, varianza e generatrice per Bernoulli, binomiali, Poisson, esponenziali e gaussiane (per esercizio la varianza di queste due ultime a partire dalla generatrice; generatrice delle esponenziali per esercizio).

Ancora da svolgere, su Poisson ed esponenziali: teorema degli eventi rari, mancanza di memoria, teorema sul minimo.

**Esercizi** suggeriti sul programma svolto fino ad ora:
23/4/2008: primi 5 punti
secondo compitino 08: primi 3 punti
4/6/2008: punti 1,2,3,4,6
15/7/2008: punti 1,2,3,4
16/9/2008: punti 1,2,3,4,5,6,7,8,9
27/1/2009: punti 1,2,3,4,5,6,8
17/2/2009: punti 1,2,3,4,5,6,7,8
4/5/2009: punti 1,3,4,5,8
12/6/2009: punti 1,2,3,4,5,6,7
2/7/2009: punti 1,2,3,6,7,8,9
17/09/2009: punti 1,2,3 (provare),8,9,10.

**Lezione 13*** (8/4). Il testo delle lezioni contrassegnate con * si può trovare alla pagina **http://users.dma.unipi.it/barsanti/statistica_2010/**.

Introduzione alla statistica, campione, media campionaria e sue proprieta. Teorema limite centrale con esempi.

**Lezione 14*** (13/4). Approssimazione gaussiana della v.a. binomiale con esempio, distribuzione della media campionaria. Esempio che fa uso della distribuzione della media campionaria. Varianza campionaria e suo valore atteso.

**Lezione 15*** (14/4). Variabile chi quadro e sua applicazione per la determinazione della legge dalla varianza campionaria.

**Lezione 16*** (15/4). Esercizio sull'uso della variabile chi quadro. Campionamento aleatorio e sue proprieta'. Trattamento statistico delle risposte binarie in un sondaggio. Esercizi vari di statistica.

**Lezione 17*** (20/4). Stima puntuale, metodo della massima verosimiglianza, applicazione alla stima dei parametri della distribuzione esponenziale e bernoulliana. Altri esempi di applicazione della massima verosimiglianza per la stima dei parametri della distribuzione di Poisson, della gaussiana e della distribuzione uniforme. Stima di intervalli, concetto di intervallo di confidenza ed esempio di calcolo per la media di una distribuzione gaussiana.

**Lezione 18*** (21/4). Intervalli di confidenza per la media unilaterali e bileterali (con sigma nota) al variare della confidenza e della numerosita' campionaria.

**Lezione 19*** (22/4). Utilizzo della t di Student nel calcolo degli intervalli di confidenza quando anche la sigma e' stimata dai dati. Intervalli di

confidenza per la sigma con l'uso della distribuzione del chi quadro. Intervalli di confidenza per la differenza fra due medie sia nel caso di sigma nota che nel caso di sigma stimata dai dati.

**Lezione 20*** (27/4). Intervalli di confidenza per la stima di una probabilita'. Esercizio di ripasso sulla variabile poissoniana. Introduzione alla verifica delle ipotesi. Ipotesi nulla, ipotesi alternativa, regione critica, errori di prima e seconda specie, significativita'.

Come **esercizi relativi alle lezioni 13-20** si suggerisce di esaminare attentamente gli esercizi svolti a lezione (vedi pdf pag. http://users.dma.unipi.it/barsanti/statistica_2
più i seguenti:

23/4/2008, punto n. 6;
4/08, ultimi due punti;
4/6/2008, punti n. 5, 7, 9;
24/6/08, punti n. 3, 4, 6;
15/7/08, punti n. 5, 6, più punti n. 1, 2 della pagina seguente;
16/9/08, punti n. 10, 11;
17/2/2009, punti n. 7, 9;
4/5/2009, punto n. 7;
12/6/2009, punto n. 10;
2/7/2009, punti n. 3, 5, 10;
17/9/2009, punti n. 5, 7.

**Lezione 21** (28/4). L'algoritmo del test (bilaterale) per la media di una gaussiana con varianza nota viene esemplificato col seguente esercizio. Un certo sistema di servizio (si pensi ad esempio agli sportelli di una banca) è ben dimensionato se ci sono in media 100 richieste al giorno (se sono di più bisogna aumentarlo, se sono di meno si stanno sprecando risorse). Forse il mercato è cambiato e le richieste non sono più 100 in media. Si registra un campione per 9 giorni:

$$98, 112, 103, 96, 108, 115, 102, 99, 109.$$

Al 95%, il servizio è ben dimensionato? Si supponga, sulla base di esperienze passate, che sia $\sigma = 4$. Sol:

$$z = \frac{\overline{x} - \mu_0}{\sigma}\sqrt{n} = \frac{104.2 - 100}{4}\sqrt{9} = 3.15$$

maggiore (in valore assoluto) di $q_{1-\frac{\alpha}{2}} = z_{\frac{\alpha}{2}} = 1.96$ (vale $\alpha = 0.05$, $\frac{\alpha}{2} = 0.025$, $1 - \frac{\alpha}{2} = 0.975$, $q_{0.975} = 1.96$). Il sistema non è ben dimensionato.

Nella risoluzione si è interpretato graficamente il test, tramite una gaussiana e le sue code. Si è insistito sul rischio, del 5%, necessariamente presente. E' l'area totale delle code. Se non si specifica tale rischio a priori, non ha senso confrontare un campione con una media ipotizzata $\mu_0$. La media empirica $\overline{x}$ è sempre diversa da $\mu_0$, per la casualità del campione e non perche' $\mu_0$ sia falsa. Quindi il punto è capire se $\overline{x}$ dista da $\mu_0$ in modo eccessivo oppure no. Il grado si anomalia della distanza di $\overline{x}$ da $\mu_0$ è dato dalle code, individuate da $\alpha$.

**Lezione 22** (29/4). Viene risolto un altro esercizio sul test (esercizio 10 del 27/1/2009), commentando di nuovo il significato, la ragionevolezza, di ciò che fa il test, in termini grafici. Se vale l'ipotesi $H_0$ (che in tale esercizio è "il numero medio di forme è 50", o più concisamente "$\mu_0 = 50$"), la v.a. $\overline{X}$ è una $N\left(\mu_0, \frac{\sigma^2}{n}\right)$. I suoi valori cadono molto raramente nelle due piccole code di area totale 0.05. Quindi, se il valore empirico $\overline{x}$ cade proprio in tali code, siamo indotti a concludere che era sbagliata l'ipotesi (perche' sotto di essa una tale cosa non doveva accadere). Quindi il test si può eseguire controllando se $\overline{x}$ cade nelle code della gaussiana $N\left(\mu_0, \frac{\sigma^2}{n}\right)$. Standardizzando, lo si può eseguire controllando se $z = \frac{\overline{x}-\mu_0}{\sigma}\sqrt{n}$ cade nelle code della gaussiana canonica, cioè controllando se

$$|z| > q_{1-\frac{\alpha}{2}}.$$

Vengono poi mostrate alcune varianti (alcune inquietanti) dell'idea precedente, tutte giuste: invece di prendere le due code simmetriche, si considera come anomala solo una coda di ampiezza 0.05; oppure una fascia centrale di ampiezza 0.05. Questa osservazione molto critica si capirà meglio più avanti. Non c'è comunque nulla di sbagliato. E' l'idea generale dei test, con cui si possono inventare tantissimi test validi: se vale l'ipotesi $H_0$, una certa grandezza da noi scelta ($\overline{X}$ o altre), assume valori in certe zone molto raramente, per cui se l'analoga grandezza empirica lo fa, rifiutiamo l'ipotesi.

Si sottolinea che un test o rifiuta $H_0$, oppure non è in grado di farlo. Non ha senso dire che il test conferma $H_0$. Anche questo sarà sempre più chiaro.

Concetto di $p$-value (valore $p$). E' la probabilità che l'indicatore da noi usato assuma valori più estremi di quello empirico. Si raffigura la densità $N\left(\mu_0, \frac{\sigma^2}{n}\right)$, si disegna il valore di $\overline{x}$, da esso si trovano le due code corrispondenti: il $p$-value è l'area di tali code.

Negli esempi precedenti:

$$p = P\left(\left|\frac{\overline{X} - \mu_0}{\sigma}\sqrt{n}\right| > \left|\frac{\overline{x} - \mu_0}{\sigma}\sqrt{n}\right|\right).$$

Dopo alcuni calcoli si trova

$$p = 2 - 2\Phi\left(\left|\frac{\overline{x} - \mu_0}{\sigma}\sqrt{n}\right|\right).$$

I software calcolano il $p$-value, non chiedono $\alpha$. Se il $p$-value viene piccolo, significa che l'indicatore assume raramente valori più estremi di quello empirico. Quindi l'ipotesi non è ragionevole.

Se si sceglie $\alpha$ in anticipo, e si calcola il $p$-value invece di eseguire il test, si può dedurre il risultato del test per quel valore di $\alpha$? Sì. se $\alpha > p$, si rifiuta $H_0$, altrimenti no. Questo viene spiegato graficamente, con la solita densità $N\left(\mu_0, \frac{\sigma^2}{n}\right)$ e le sue code.

Infine, viene esaminato un esercizio un po' diverso.

*Esercizio.* Un servizio ferroviario viene cancellato se il numero medio di passeggeri è inferiore a 20. I passeggeri che viaggiano ad una certa ora tra Firenze e Pisa vengono contati, per 10 giorni. I valori sono

$$17, 13, 9, 23, 14, 12, 16, 11, 14, 21.$$

A livello di significatività 95%, c'è ragione di cancellare quel servizio serale?

*Soluzione.* Si può eseguire il solito test, ma questo non è aderente al problema. Meglio ragionare dall'inizio. Se la media è 20 (caso limite), $\overline{X}$ è una $N\left(\mu_0, \frac{\sigma^2}{n}\right)$. Se i valori di $\overline{X}$ cadono nella coda destra, non c'è nulla di male: può significare che il numero medio di passeggeri è anche maggiore di 20, meglio ancora. Il problema sussiste solo se $\overline{X}$ cade nella coda sinistra, cioè assume valori troppo piccoli. Allora è solo la coda sinistra che rappresenta la situazione anomala. Prendiamo quindi una coda sinistra di ampiezza 0.05, senza coda destra. Il test consiste nel vedere se $\overline{X}$ cade in essa, ovvero, dopo standardizzazione, se

$$\frac{\overline{x} - \mu_0}{\sigma}\sqrt{n} < -q_{1-\alpha}.$$

Questo test *unilaterale* è più "intelligente" di quello solito *bilaterale*, vista la particolare domanda dell'esercizio. Questo è un esempio in cui si vede

l'utilità delle varianti dette sopra. Capiremo meglio tutto questo col concetto di potenza di un test.

**Lezione 23** (4/5). Approfondimenti sui test statistici. Vengono formulati due test, uno basato sulla media $\overline{X}$ (si rifiuta quando cade nelle code di una $N\left(\mu_0, \frac{\sigma^2}{n}\right)$), l'altro basato semplicemente su $X$ stessa, un solo esperimento (si rifiuta quando cade nelle code di una $N(\mu_0, \sigma^2)$). Intuitivamente, come mai pensiamo che il primo sia migliore? Se le code hanno area $\alpha$ (es. 0.05) in entrambi i casi, in entrambi abbiamo la stessa probabilità $\alpha$ di errore di prima specie: la probabilità che il test rifiuti $H_0$ quando era vera, è $\alpha$. Cosa li distingue?

Viene introdotto il concetto di errore di seconda specie (accettare $H_0$ quando è falsa). Viene calcolata la sua probabilità, detta $\beta$. Ci si accorge che è una funzione della media vera $\mu$. Attraverso alcuni esercizi preliminari si trova

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

Come funzione della media vera standardizzata, più precisamente di

$$d = \left|\frac{\mu - \mu_0}{\sigma}\sqrt{n}\right|$$

la funzione $\beta$ è detta curva caratteristica operativa (curva OC).

La potenza del test è, per definizione, $1 - \beta(\mu)$, cioè la probabilità di rifiutare $H_0$ quando è falsa.

I due test visti all'inizio differiscono in potenza. Lo si spiega attraverso l'inyerpretazione grafica di $\beta$, che risulta molto più alto per il secondo test.

Infine, viene confrontato il caso unilaterale col bilaterale, sempre in termini di $\beta$ o di potenza. Quando si cerca di scoprire che il valor medio è unilateralmente diverso da $\mu_0$, il test unilaterale è più potente.

**Lezione 24** (5/5). Esercizio (esempio) 8.36 del libro. Svolto sia col test bilaterale sia unilaterale. Spiegazione grafica dettagliata del fatto che la potenza è maggiore per il test unilaterale, che quindi è il migliore tra quelli a noi noti (eppure non ha rifiutato l'ipotesi). Calcolo del valore $p$. Per tentare di far meglio, l'azienda produttrice dell'esercizio, nel tentativo di arrivare a rifiutare l'ipotesi (cosa che desidera), potrebbe aumentare la numerosità campionaria. Oppure prendere un $\alpha$ meno stringente, ma dal calcolo del $p$-value si vede che gli unici valori di $\alpha$ per cui si rifiuterebbe l'ipotesi non sono proponibili.

**Lezione 25** (6/5). Esercizio: calcolare la potenza del test unilaterale destro.

Sol. Se il test ha come ipotesi alternativa $H_1$ = "la media è $> \mu_0$ ", la probabilità $\beta(\mu)$ dell'errore di seconda specie se la media vera è una certa $\mu > \mu_0$ è (dopo alcuni calcoli con ausilio del grafico)

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

Quindi la potenza è $1 - \beta(\mu)$.

Come incide la numerosità sulla potenza? A livello analitico: $n$ più grande implica $\frac{\mu_0-\mu}{\sigma}\sqrt{n}$ più negativo, quindi $\beta(\mu)$ più piccolo. Quindi potenza maggiore. Si vede anche a livello grafico, mostrando che al crescere di $n$ le gaussiane centrate in $\mu_0$ e $\mu$ diventano più concentrate, e quindi più separate, ecc.

Design Of Experiments (DOE). Si tratta di un'ampia teoria che guida alla progettazione razionale di esperimenti. Qui ci occupiamo solo di un dettaglio, però importantissimo: la scelta della numerosità $n$ del campione da esaminare. Abbiamo visto che $n$ influisce sulla potenza. Quindi la domanda è: fissata la significatività $\alpha$, scelto il tipo di test (unilaterale, bilaterale, ecc.), quale $n$ bisogna scegliere per avere una certa potenza? O più precisamente, per avere una certa potenza di osservare una certa differenza $\mu - \mu_0$?

Caso unilaterale destro: fissati $\alpha$, $\mu$ e $\beta(\mu)$, es. $\beta(\mu) = 0.05$, si risolve l'equazione

$$\Phi\left(\frac{\mu_0 - \mu}{\sigma}\sqrt{n} + q_{1-\alpha}\right) = 0.05$$

ovvero

$$\frac{\mu_0 - \mu}{\sigma}\sqrt{n} + q_{1-\alpha} = q_{0.05}$$

da cui si trova $n$. Nota: la soluzione non è un intero, quindi si prende l'intero immediatamente superiore

$$n \geq \left(\frac{(q_{0.1} - q_{1-\alpha})\,\sigma}{\mu_0 - \mu}\right)^2.$$

Vanno bene ovviamente tutti quelli ancora superiori, ma quello è il più piccolo che va bene.

A livello pratico, ci sono molte scelte da fare, prima di calcolare $n$. Una delle più critiche è $\mu$. Un'idea possibile, anche se vaga, è che $\mu$ sia il primo

valore critico diverso da $\mu_0$, cioè il primo valore che, se realizzato, provoca delle conseguenze rilevanti, e che quindi deve essere rilevato dal test.

Quando si eseguono test nelle aziende? Ad esempio quando si fa monitoraggio, ad esempio con le *carte di controllo*. Vengono spiegate sommariamente le carte di controllo, le bande, il campionamento a tempi regolari, l'allarme quando si esce dalle bande, l'analogia con l'eseguire un test ad ogni istante di controllo. Vengono calcolate le bande: $\mu_0 \pm \frac{\sigma q_{1-\frac{\alpha}{2}}}{\sqrt{n}}$, come per l'intervallo di confidenza (ma lo scopo è un test, non la stima della media). Operativamente, pensando ad un esempio concreto, come si realizza una carta? Vanno scelti $\alpha$ ed $n$, fissati una volta per tutte. Per sceglerli si deve aver chiaro il significato di ogni elemento della teoria del test. $\alpha$ è la probabilità di usciere dalle bande quando invece la media è rimasta $\mu_0$. In molti casi questo non è così grave: quando accade basta rifare il campionamento, per controllare meglio. $n$ va scelto per avere una certa potenza, relativamente ad un certo $\mu$. Si deve sapere, ad es. dagli esperti delle cose prodotte, quali deviazioni da $\mu_0$ rendono inservibili o pericolose le cose prodotte. $\mu$ corrisponde a tali valori critici. A quel punto va scelto $\beta(\mu)$. E' la probabilità di non accorgersi di un cambiamento di $\mu_0$, quando questo è avvenuto. Questo sì che può essere periciloso: vendere cose fuori norma senza saperlo. Allora $\beta(\mu)$ va preso molto piccolo, e trovato $n$ in corrispondenza.

**Lezione 26** (11/5). p.187 e seguenti: richiamo sulla definizione di v.a. chi quadro, grafico delle densità al variare di $n$, media $n$ e deviazione standard proporzionale a $\sqrt{n}$ (concentrazione relativa). Numeri (simili a i quantili) $\chi^2_{\alpha,n}$, tavole relative, interpretazione geometrica.

p. 220: ricordando la definizione di $S^2$ e la proprietà dimostrata $E[S^2] = \sigma^2$, vale il teorema:

$$\frac{S^2}{\sigma^2}(n-1)$$

ha legge $\chi^2_{n-1}$. La dimostrazione è difficile ma un'idea si ha nel seguente modo. Introduciamo

$$S^2_\mu := \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2.$$

Esercizio per casa: mostrare che $E\left[S^2_\mu\right] = \sigma^2$.

Si vede subito che

$$\frac{S^2_\mu}{\sigma^2} n = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2$$

cioè $\frac{S_\mu^2}{\sigma^2}n$ è somma di $n$ quadrati di gaussiane canoniche indipendenti. Quindi $\frac{S_\mu^2}{\sigma^2}n$ è $\chi_n^2$, per definizione. Non è così strano quindi che $\frac{S^2}{\sigma^2}(n-1)$ possa essere una chi quadro; meno facile da intuire è che abbia solo $n-1$ gradi di libertà.

p. 323 e seguenti: test chi quadro per la varianza, caso unilaterale (destro). Esempio 8.5.1, anche proprio eseguendo il test a livello $\alpha = 0.05$, tramite le tavole. L'idea intuitiva del test è la seguente: ci di chiede se $S^2$ sia significativamente più grande di $\sigma_0^2$ (l'ipotesi nulla è che la varianza sia $\sigma_0^2$, oppure $\leq \sigma_0^2$); se lo è, rifiutiamo l'ipotesi nulla. Il problema è confrontare $S^2$ con $\sigma_0^2$ (non basta la semplice disuguaglianza $S^2 > \sigma_0^2$): di quanto deve essere $S^2$ maggiore di $\sigma_0^2$ per rifiutare $\sigma_0^2$? Tutto è relativo ai parametri del problema (es. $n$) ed al rischio $\alpha$ che vogliamo correre. Allora, come nel caso del test per la media, il metodo matematico consiste nell'introdurre una grandezza universale (es. la standardizzazione) associata alla grandezza statistica in oggetto. Per la media, si introdiceva la standardizzazione $Z = \frac{\overline{X}-\mu_0}{\sigma}\sqrt{n}$. Per la varianza, si introduce $\frac{S^2}{\sigma_0^2}(n-1)$. Il problema se $S^2$ sia significativamente maggiore di $\sigma_0^2$ si traduce nel problema se $\frac{S^2}{\sigma_0^2}(n-1)$ sia significativamente grande. Sappiamo che $\frac{S^2}{\sigma_0^2}(n-1)$ è una $\chi_{n-1}^2$, quando l'ipotesi nulla è vera. Quindi $\frac{S^2}{\sigma_0^2}(n-1)$ supera $\chi_{\alpha,n-1}^2$ solo con probabilità $\alpha$. Preso $\alpha$ piccolo, se per il campione sperimentale vale $\frac{S^2}{\sigma_0^2}(n-1) > \chi_{\alpha,n-1}^2$, giudichiamo questo incompatibile con l'ipotesi nulla ($\frac{S^2}{\sigma_0^2}(n-1)$ è troppo grande). Quindi rifiutiamo l'ipotesi.

Viene anche calcolato il valore $p$: la probabilità che una $\chi_{n-1}^2$ superi il valore sperimentale di $\frac{S^2}{\sigma_0^2}(n-1)$. Bisogna usare le tavole della chi quadro al contrario, cosa spesso poco agevole per via dei pochi valori.

**Lezione 27** (12/5). Problema: filato prodotto correttamente: $\mu_0 = 0.2$ mm, $\sigma_0 = 0.02$ mm. Spessore da evitare: 0.3 mm, o 0.1 mm. Creare carte di controllo.

Osservazione 1: si tratta di un processo ad alta precisione, cioè con $\sigma_0$ molto piccola. Un campione ha probabilità piccolissima di superare 0.3 mm per caso (servono 5 sigma, quasi impossibile).

Osservazione 2: ha quindi senso tenere sotto controllo la media, invece che il singolo esemplare. Infatti il singolo esemplare è improbabile che superi 0.3 mm per caso, se la media resta quella. Il pericolo non è nella causalità, ma in un peggioramento sistematico della media.

Osservazione 3: oppure il pericolo è in un peggioramento della varianza:

se fosse ad es. $\sigma = 0.05$ mm., basta arrivare a $2\sigma$ per raggiungere 0.3 mm per caso in un esemplare. Questo sarebbe frequente.

Conclusione: vanno tenute sotto controllo media e varianza.

Intanto si crea una carta di controllo per la media, in cui $UCL = \mu_0 + \frac{\sigma_0 q_{1-\frac{\alpha}{2}}}{\sqrt{n}}$ e $LCL = \mu_0 - \frac{\sigma_0 q_{1-\frac{\alpha}{2}}}{\sqrt{n}}$. Vanno scelti $\alpha$ ed $n$. Come già detto in una lezione precedente, $\alpha$ forse può essere scelto non troppo severo, es. 0.05, mentre $n$ va scelto con lo scopo di avere una certa potenza. Va identificato un valore $\mu$ che non si vuole raggiungere, va scelta la relativa potenza (qui bisogna essere severi, trattandosi della probabilità di accettare del filato troppo spesso o sottile), e calcolato $n$. Lo faremo nella prossima lezione.

Poi si deve creare una carta di controllo per la varianza. Si può impostare così: l'indicatore è positivo, c'è un'unico limite, $UCL$. Un modo è prendere come indicatore $\frac{S^2}{\sigma_0^2}(n-1)$, e $UCL = \chi^2_{\alpha,n}$. Di nuovo vanno scelti $\alpha$ ed $n$, con ragionamenti analoghi.

**Lezione 28** (13/5). Esercizi di preparazione al compito. Il compito conterrà alcuni esercizi semplici di calcolo ed altri più articolati, da presentare con una discussione più ampia.

Vengono svolti quattro esercisi semplici di calcolo:

*Esercizio 1*: se $X, Y$ sono Poisson di parametro 5, calcolare $E\left[X\left(e^Y - Y^2\right)\right]$.

*Esercizio 2*: se $X$ è una $N(3,4)$, trovare $\lambda$ tale che $P(X > \lambda) = 0.8$.

*Esercizio 3*: se $X$ è una $N(-5,9)$, calcolare $P(X > 0)$.

*Esercizio 4*: se $X$ è una $B(2, 0.7)$, calcolare $E\left[\frac{X-1}{X+1}\right]$.

Viene poi svolto un esercizio più strutturato, sulla falsariga del problema della volta precedente. la prosuzione corretta di un filato deve avere spessore medio 0.2 mm con deviazione standard 0.02 mm. Se lo spessore di un esemplare raggiunge 0.3 mm, si nota nei tessuti; se raggiunge 0.1 mm, il filo si può strappare.

*Domanda 1*: calcolare la probabilità che un esemplare superi 0.3 mm.

*Domanda 2*: a parità di $\sigma$, che spessore medio dovremmo avere affinche' un esemplare su 1000 superi 0.3 mm?

*Domanda 3*: alla luce del risultato appena ottenuto, costruire una carta di controllo per la media, esaminando criticamente le scelte da effettuare in fase di DOE.

**Lezione 29** (18/5). Esercizio. Nei mesi

$$T = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12$$

sono state registrate le vendite di un prodotto, pari a

$$X = 18, 2, 30, 1, 1, 12, 12, 7, 9, 17, 13, 3.$$

Per risparmiare calcoli, nel seguito si usino i seguenti risultati intermedi:

$$\sum x_i = 125, \sum x_i^2 = 2115, \sum t_i = 78, \sum t_i^2 = 650, \sum x_i t_i = 763.$$

Domanda 1. Tracciare un istogramma con intevalli $(0, 5]$, $(5, 10]$, $(10, 15]$, ecc. e sovrapporre la gaussiana stimata.

Domanda 2. Calcolare la quantità da tenere in magazzino sufficiente al 90%.

Sol. Dobbiamo trovare il numero $q$ tale che $P(X \leq q) = 0.9$. Vale $q = \mu + \sigma q_{0.9} \approx 10.416 + 8.6 \cdot 1.28 = 21.424$.

Domanda 3. Quanti mesi di osservazione servirebbero per stimare la media con un errore pari all'unità, al 95%? Che precisione si ottiene in tre mesi?

Domanda 4. Sappiamo che nei mesi sopra considerati è peggiorata la crisi economica. Questo ha influito sulle vendite?

Circa la domanda 4 viene data la formula per la covarianza empirica

$$\widehat{Cov} = \frac{1}{n-1} \sum (x_i - \overline{x})(y_i - \overline{y})$$

discusso il fatto che $\widehat{Cov} > 0$ corrisponde a dati che crescono o calano di pari passo, mentre $\widehat{Cov} < 0$ corrisponde a dati che si muovono in controtendenza; viene poi evidenziato il difetto che $\widehat{Cov}$ dipende dall'ordine di grandezza dei numeri in gioco e delle loro fluttuazioni, quindi si introduce il coefficiente di correlazione

$$r = \frac{\widehat{Cov}}{S_X S_Y}$$

che ha le stesse proprietà sul segno ma è universale e varia tra -1 ed 1 (si trova tutto al par. 2.6). Nell'esempio si trova $r = -0.145$, che indica sostanzialmente assenza di legame tra tempo e vendite.

*Esercizio per casa*: ripetere le stesse cose per i dati

$$T = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12$$
$$X = 1, 13, 11, 23, 10, 24, 8, 28, 3, 1, 9, 10.$$

**Lezione 30** (19/5). Retta di regressione (regressione lineare semplice, non multipla), calcolo dei coefficienti col metodo dei minimi quadrati, par. 9.2:

$$B = r\frac{S_Y}{S_X}, \qquad A = \overline{y} - B\overline{x}.$$

Calcolo e rappresentazione grafica nel caso dell'esercizio precedente.

Si torna all'istogramma dell'esercizio precedente, dove la densità gaussiana non sembra una buona approssimazione. Si pensa di provare con le densità esponenziali. Esse dipendono da un solo parametro $\lambda$. Come ricavarlo (stimarlo) dai dati? Prima idea:

$$\mu = \frac{1}{\lambda}$$

quindi

$$\lambda \approx \frac{1}{\overline{x}}.$$

Potenziale contraddizione: anche $\sigma = \frac{1}{\lambda}$, quindi anche $\lambda \approx \frac{1}{S}$ dovrebbe essere una ragionevole stima, ma se $S$ ed $\overline{x}$ differiscono molto, forse il modello esponenziale non è ragionevole.

Note sugli esercizi assegnati. Ci sono alcuni errori nella risoluzione degli esercizi dei compiti di MMS suggeriti. Ecco la lista: 27/01/09 es. 3, errore nella varianza; 16/09/2008 es. 10, errore nel calcolo di $\Phi$; 17/02/2009 es. 9, manca un quadrato nel calcolo di $n$; 15/07/2008, è sbagliato il valore di $q_{0.975}$.

**Lezione 31** (20/5). Torniamo ai dati dell'esercizio della lezione 29.

Domanda 1. Trovare la densità esponenziale associata ai dati e disegnarla.

Sol. Possono esserci varie densità esponenziali, a seconda di come stimiamo $\lambda$. Usiamo prima di tutto $\lambda \approx \frac{1}{\overline{x}} = \frac{1}{10.416} = 0.096$, che poi disegneremo sovrapposta all'istogramma. Questa stima è anche supportata dal metodo della massima verosimiglianza. Un'altra stima sarebbe $\lambda \approx \frac{1}{S} = \frac{1}{8.6} = 0.116$. Più avanti vedremo un altro metodo di stima più adattato al problema della coda destra.

Domanda 2. Calcolare la quantità da tenere in magazzino sufficiente al 90%.

Sol. Dobbiamo trovare il numero $q$ tale che $P(X \leq q) = 0.9$. Se $X \sim Exp(\lambda)$, $F_X(t) = 1 - e^{-\lambda t}$, quindi dobbiamo trovare $q$ tale che $1 - e^{-\lambda q} = 0.9$, dove $\lambda$ è uno di quelli stimati sopra. Vale $e^{-\lambda q} = 0.1$, $\lambda q = \log 10$, $q = \frac{\log 10}{\lambda} =$

$\frac{\log 10}{0.096} = 23.985$. E' maggiore di 21.424, quindi forse più realistico alla luce del valore 30 che si è osservato nel campione.

Domanda 3. Che precisione ha, approssimativamente, la stima di $\frac{1}{\lambda}$ data da $\overline{x}$, al 95%? Come si dovrebbe ragionare se è richiesta la precisione della stima di $\lambda$, ovvero se è richiesta una valutazione dell'errore $\left|\lambda - \frac{1}{\overline{x}}\right|$?

Nota: si usano gli intervalli di confidenza $\frac{1}{\lambda} = \overline{x} \pm \frac{\sigma q_{0.975}}{\sqrt{n}}$, che però sono approssimati per due ragioni. La prima è che non si conosce la vera $\sigma$; questo sarebbe rimediabile prendendo $\frac{St_{0.975}^{(n-1)}}{\sqrt{n}}$. La seconda è che le v.a. di partenza non sono gaussiane, quindi $\overline{X}_n$ non è $N\left(\mu, \frac{\sigma^2}{n}\right)$ e quindi gli estremi dell'intervallo di confidenza non andrebbero calcolati tramite i quantili gaussiani. Ma per il TLC, $\overline{X}_n$ è approssimativamente $N\left(\mu, \frac{\sigma^2}{n}\right)$, quindi le varie conclusioni sono circa corrette.

Nota: l'errore relativo, al 95%, cioè $\left|\frac{\overline{x}-\mu}{\mu}\right|$, è circa 0.5, molto alto.

Nota: vale $\left|\lambda - \frac{1}{\overline{x}}\right| = \frac{|\overline{x}-\mu|}{\overline{x}\mu}$, quindi si possono riportare le informazioni su $|\overline{x} - \mu|$ a quelle per $\left|\lambda - \frac{1}{\overline{x}}\right|$. Facendo i conti si vede che l'errore relativo $\frac{\left|\lambda - \frac{1}{\overline{x}}\right|}{\lambda}$ è di nuovo circa 0.5.

*Esercizio per casa* (Domanda 4). Trovare $n$ in modo da avere errore *relativo* inferiore a 2/10.

Domanda 5 (rinviata). Se non si disponeva del campione della lezione 29, che difficoltà c'era a rispondere alla domanda 4? Se avessimo ipotizzato, da analisi di mercato, che le vendite avrebbero avuto una media tra 5 e 15 ed una deviazione tra 5 e 15?

Problema: come decidere razionalmente se è meglio la gaussiana o l'esponenziale? mmaginiamo di inventare un test, in cui l'ipotesi nulla $H_0$ sia del tipo: la densità è la gaussiana $N\left(10.416, 8.6^2\right)$; oppure, la densità è esponenziale di parametro 0.096. Magari il test rifiuta una delle due e non l'altra, permettendoci di scegliere. Anche se il test non discriminasse, possiamo scegliere la densità che conquista il miglior valore $p$.

Relativamente alla partizione con cui facciamo l'istogramma (o un'altra, fissata), si calcolino le frequenze teoriche $p_i$ e quelle sperimentali $\widehat{p}_i$. Intendiamo: $p_i$ è la probabilità (ad es. nell'ipotesi gaussiana) di cadere in quell'intervallo della partizione; se tale intervallo ha estremi $[a_i, b_i]$ vale

$$p_i = F\left(b_i\right) - F\left(a_i\right)$$

che, nel caso gaussiano diventa

$$\Phi\left(\frac{b_i-\mu}{\sigma}\right)-\Phi\left(\frac{a_i-\mu}{\sigma}\right)$$

mentre nel caso esponenziale diventa

$$1-e^{-\lambda b_i}-\left(1-e^{-\lambda a_i}\right).$$

Invece, la frequenza sperimentale $\widehat{p}_i$ è quella del campione: ad es. $\frac{4}{12}$ per l'intervallo $(0,5]$, ecc.

Sia $k$ il numero di intervallini. Se prendiamo $\sum_{i=1}^{k}\left|\widehat{p}_i-p_i\right|^2$, stiamo valutando quanto distano le due densità. Però così diamo poco peso alle code (assai importanti per le applicazioni), in quanto nelle code i valori $\widehat{p}_i$ e $p_i$ sono più piccoli e quindi $\left|\widehat{p}_i-p_i\right|^2$ è troppo piccolo, indipendentemente dal fatto che la densità sia una buona approssimazione dell'istogramma. Prendendo la somma degli errori relativi

$$\sum_{i=1}^{k}\left|\frac{\widehat{p}_i-p_i}{p_i}\right|^2$$

si cura questo difetto, ma si ottiene un'espressione molto instabile, molto sensibile a piccole variazioni, in quanto i denominatori $p_i^2$ sono troppo piccoli, proprio sulle code. Un miglior equilibrio è l'espressione

$$T_0=\sum_{i=1}^{k}\frac{\left|\widehat{p}_i-p_i\right|^2}{p_i}.$$

Esercizio per casa: calcolare $T_0$ per la gaussiana e per l'esponenziale, confrontandoli.

Nota: la migliore è quella col $T_0$ più piccolo (frequenze più vicine). Si può vedere che questo corrisponde al miglior $p$-value e ad una maggior possibilità di non venir rifiutati dal test. Attenzione: la grandezza $T$ che si userà nel cosidetto test chi quadro è uguale a $nT_0$

$$T=nT_0.$$

**Lezione 32** (25/5).

Esercizio 1. Data $X\sim N(5,4)$, calcolare $P(X>3)$.

Esercizio 2. Se $X$ rappresenta il quadagno giornaliero, quale guadagno è garantito al 95%? Ovvero, qual'è il "guadagno minimo" che ci si aspetta al 95%?

Esercizio 3. Ed il "guadagno massimo", al $98\%$?

Esercizio 4. Con che probabilità il guadagno non supererà 8?

Esercizio 5. Se $X$ ed $Y$ sono come sopra ma indipendenti, guadagni relativi a due diversi giorni, con che probabilità il guadagno totale dei due giorni sarà maggiore di 6?

Esercizio 6. Considerando 25 giorni lavorativi in un mese ed i guadagni indipendenti ed ugualmente distribuiti, ma non più $N(5,4)$ bensì esponenziali di media 5, calcolare approssimativamente la probabilità di guadagnare più di 100.

Soluzioni (cenni) relative ad alcuni esercizi.

Esercizio 2. Cerchiamo il valore $g$ tale che $P(X \geq g) = 0.95$ (il valore minimo, superato il 95% delle volte). Dopo aver fatto un disegno ed un confronto grafico con la gaussiana canonica, si capisce (o in altro modo) che

$$g = 5 - 2 \cdot q_{0.95} = ...$$

Esercizio 5. Il guadagno totale dei due giorni è $X+Y$, che è $N(5+5, 4+4) = N(10, 8)$. Quindi bisogna calcolare

$$P(X + Y > 6) = 1 - \Phi\left(\frac{6-10}{\sqrt{8}}\right) = ...$$

Esercizio 6. Detti $X_1, ..., X_{25}$ i guadagni giornalieri, vale

$$P(X_1 + ... + X_{25} > 100) = P\left(\frac{X_1 + ... + X_{25} - 25 \cdot \mu}{\sqrt{25}\sigma} > \frac{100 - 25 \cdot \mu}{\sqrt{25}\sigma}\right)$$

dove prendiamo $\mu = E[X_i] = 5$, $\sigma^2 = Var[X_i] = 5^2$. Per il TLC la v.a. $\frac{X_1+...+X_{25}-25\cdot\mu}{\sqrt{25}\sigma}$ è approssimativamente una $Z \sim N(0,1)$, quindi

$$P(X_1 + ... + X_{25} > 100) \approx P\left(Z > \frac{100 - 25 \cdot 5}{\sqrt{25} \cdot 5}\right) = 1 - \Phi\left(\frac{100 - 25 \cdot 5}{\sqrt{25} \cdot 5}\right) = ...$$

---

Esercizio. Due laboratori esaminano la percentuale di difettosità di certi pezzi prodotti. Il LabA trova $p_A = 0.045$, mentre il LabB trova $p_B = 0.091$.

Il committente vuole capire questa discrepanza ed esamina lui stesso 12 esemplari, trovandone uno difettoso. Quale laboratorio gli sembra migliore? Può rifiutare alcune delle loro dichiarazioni?

Sol. $\widehat{p} = \frac{1}{12} = 0.083$. Quanto è distante (relativamente ai parametri del problema, es. 12 esemplari)? Dette $X_1, ..., X_{12}$ delle v.a. che valgono 1 se il corrispondente pezzo è difettoso, con probabilità $p$, zero altrimenti, posto $S = X_1 + ... + X_{12}$ = numero di pezzi difettosi, $S$ è $B(12, p)$, quindi potremmo tentare di usare le binomiali per ragionare sul problema. Questo aprirebbe la strada ai test binomiali, test esatti, che però non trattiamo.

Possiamo invece, in virtù del TLC (ma 12 è un po' bassa), pensare che $\frac{S}{n}$ è circa gaussiana, quindi applichiamo approssimativamente il test gaussiano per la media. Esaminando la dichiarazione del LabA, l'ipotesi nulla è $\mu_0$ (o se si vuole $p_0$) $= 0.045$, e la media empirica è proprio

$$\frac{S}{n} = \widehat{p} = \frac{1}{12} = 0.083.$$

Quindi calcoliamo

$$z = \frac{0.083 - 0.045}{\sigma}\sqrt{12}.$$

C'è il problema di $\sigma$. Prendiamo la $\sigma$ del modello ipotizzato, che essendo una $B(1, p_0)$ porta a $\sigma^2 = p_0(1 - p_0)$, quindi

$$\sigma = \sqrt{0.045 \cdot 0.955} = 0.0497.$$

Pertanto $z = 0.635$. Questo va confrontato con un quantile preso a priori, es. $q_{0.975} = 1.96$. Il test non è significativo: non possiamo rifiutare la dichiarazione di LabA (a maggior ragione quella di LabB che è più vicina al risultato sperimentale).

---

Altro approccio. Problema dell'adattamento (corrispondenza tra proporzioni empiriche e densità teorica prefissata) e test chi quadro: si confronta

$$T = n \sum_{i=1}^{k} \frac{|\widehat{p}_i - p_i|^2}{p_i}$$

(che è approssimativamente $\chi^2_{k-1}$, per $n$ grande) col quantile $\chi^2_{\alpha,k-1}$.

**Lezione 33** (26/5). Esercizio. Calcolare $T$ per entrambi i laboratori i casi e confrontarli. Eseguire il test chi-quadro al 95%.

Sol. $\widehat{p}_1 = \frac{11}{12} = 0.917$, $\widehat{p}_2 = \frac{1}{12} = 0.083$,

$$p_1^A = \Phi\left(\frac{25 - 10.416}{8.6}\right) = \Phi\left(1.696\right) = 0.955$$
$$p_2^A = 1 - 0.955 = 0.045.$$

$$p_1^B = 1 - e^{-0.096 \cdot 25} = 0.909$$
$$p_2^B = 1 - 0.909 = 0.091$$

da cui

$$T^A = 12 \cdot \frac{|0.917 - 0.955|^2}{0.955} + 12 \cdot \frac{|0.083 - 0.045|^2}{0.045} = 0.403$$

$$T^B = 12 \cdot \frac{|0.917 - 0.909|^2}{0.909} + 12 \cdot \frac{|0.083 - 0.091|^2}{0.091} = 0.009$$

quindi è migliore l'approssimazione esponenziale della coda. Ma $\chi^2_{0.05,1} = 3.841$, non superato da entrambi. Non possiamo rifiutare nessuna delle due ipotesi.

Nota: il test gaussiano della lezione precedente è identico a quello appena svolto: i quadrati dei due numeri $z = 0.635$ e $q_{0.975} = 1.96$ su cui si basava il test gaussiano, sono esattamente pari ai due numeri $T^A = 0.403$ e $\chi^2_{0.05,1} = 3.841$ su cui si basa il test chi-quadro.

Esercizio. Dividere i dati della lezione 29 nelle due classi $(-\infty, 25]$, $(25, \infty)$. Calcolare frequenze empiriche e frequenze teoriche sia gaussiane sia esponenziali, calcolare $T$ in entrambi i casi e confrontarli. Eseguire il test chi-quadro al 95%.

**Lezione 34** (27/5). Esercizio 7 del capitolo 8. Attenzione: la domanda b richiede l'uso della potenza (vedi DOE, ricerca della numerosità per avere una potenza data).

Esercizio. Si prendano i dati dell'esempio 8.4.4. Il programma di prevenzione illustrato in quell'esempio ha avuto effetto?

Sol. Come illustrato nell'esempio, si traduce il problema in uno che sappiamo studiare: si calcolano le differenze e si considera come nuova variabile di studio la differenza tra il valore nuovo (dopo l'innovazione) e quello vecchio (prima dell'innovazione). L'ipotesi nulla è (come in tutti i test fatti fino

ad ora) che l'innovazione non abbia avuto effetto, cioè le cose stiano come prima, cioè $\mu_0 = 0$ ($\mu_0$ è la media delle differenze tra dopo e prima). Dette $w_1, ..., w_{10}$ le dieci differenze negli stabilimenti, calcoliamo

$$\frac{\overline{w} - \mu_0}{\sigma}\sqrt{n}$$

prendendo come $\sigma$ la deviazione $S_W$ dei dati $w_1, ..., w_{10}$ stessi. Vale $\frac{\overline{w}-\mu_0}{\sigma}\sqrt{n} = \frac{-2.15-0}{3}\sqrt{10}$. Eseguiamo un test unilaterale del tipo $\frac{\overline{w}-\mu_0}{\sigma}\sqrt{n} < -q_{1-\alpha} = -164$ (al 95%). Risulta significativo, l'innovazione ha avuto effetto.

Si chiede poi di calcolare il $p$-dei-dati, che essendo al probabilità di avere valori più estremi di quello sperimentale, è dato da

$$P\left(Z < -\frac{2.15}{3}\sqrt{10}\right) = \Phi\left(-\frac{2.15}{3}\sqrt{10}\right) = ...$$

Infine, viene richiesta la potenza di questo test di accorgersi di un calo di 3 unità. Disegnandosi l'intervallo in cui viene accettata l'ipotesi nulla $(\mu_0 - \delta, \infty)$, $\delta = \frac{\sigma q_{1-\alpha}}{\sqrt{n}}$ (è un test unilaterale) e la gaussiana di $\overline{W}$, che è una $N\left(\mu, \frac{\sigma^2}{n}\right) = N\left(-3, \frac{3^2}{10}\right)$, si deve calcolare $\beta(\mu) = \beta(-3)$, la probabilità dell'errore di seconda specie, cioè la probabilità secondo la gaussiana $N\left(-3, \frac{3^2}{10}\right)$ di cadere nell'intervallo di accettazione $(\mu_0 - \delta, \infty)$, probabilità che vale

$$\begin{aligned} 1 - F(\mu_0 - \delta) &= 1 - \Phi\left(\frac{\mu_0 - \delta - \mu}{\frac{\sigma}{\sqrt{n}}}\right) = 1 - \Phi\left(\frac{\mu_0 - \mu}{\sigma}\sqrt{n} - q_{1-\alpha}\right) \\ &= 1 - \Phi\left(\frac{3}{3}\sqrt{10} - 1.64\right) \end{aligned}$$

da cui la potenza vale $\Phi\left(\frac{3}{3}\sqrt{10} - 1.64\right)$.