Esercitazione 8 maggio 2014

Esercizio 2 dal tema d'esame del 13.01.2014 (parte II).
L'età media di n gruppo di 10 studenti che hanno appena conseguito la laurea triennale è di 22 anni.
a) Costruire un intervallo di confidenza al 95% per l'età media della popolazione dei neo-laureati della triennale, supponendo che l'età sia distribuita normalmente con varianza nota $\sigma^2 = 45$.
b) Determinare la numerosità del campione necessaria perché l'errore massimo nella stima della media sia di 0.7, assumendo sempre un intervallo di confidenza del 95%.

Soluzione.
Sappiamo che la media campionaria è uno stimatore non distorto e consistente della media, e che con le ipotesi di gaussianità introdotte, è una variabile gaussiana con media (incognita) μ e varianza $45/n$, con n=10 nel nostro caso.
a) L'intervallo di confidenza del 95% per la media si ricava allora dal fatto che la media campionaria standardizzata è una normale standard, per cui

$$P\left(\frac{|\bar{X}_n - \mu|}{\sqrt{\sigma^2/n}} \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

ove $z_{\alpha/2}$ è il valore tale che $P(Z \geq z_{\alpha/2}) = \alpha/2$ e Z è una variabile normale standard.

Cioè, noti n e σ^2, per ogni valore di α possiamo trovare $z_{\alpha/2}$ tale che la probabilità che un campione $x_1, \ldots, x_n$ abbia media campionaria $\bar{x}$ tale che

(1) $$\bar{x} - z_{\frac{\alpha}{2}}\sqrt{\sigma^2/n} \leq \mu \leq \bar{x} + z_{\frac{\alpha}{2}}\sqrt{\sigma^2/n}$$

è 1- α. La quantità $|\bar{X}_n - \mu|$ è l'errore di stima e, pertanto, la quantità $z_{\frac{\alpha}{2}}\sqrt{\sigma^2/n}$ è l'errore *massimo* di stima al livello 1- α, e la lunghezza dell'intervallo di confidenza è, ovviamente, $2\, z_{\alpha/2}\sqrt{(\sigma^2/n)}$.
Per i dati dell'esercizio si ha, applicando la formula (1), che l'intervallo di confidenza del 95% per l'età media è $(22 - 1.96\sqrt{(45/10)},\ 22 + 1.96\sqrt{(45/10)}) = (17.84, 26.15)$.

b) Cerchiamo n tale che

$$P(|\bar{X}_n - \mu| \leq 0.7) = 0.95$$

Per quanto svolto al punto a), basta imporre $z_{\frac{\alpha}{2}}\sqrt{\sigma^2/n} \leq 0.7$, cioè $1.96\sqrt{(45/n)} \leq 0.7$, che, risolto, fornisce $n \geq 358$.

Esercizio 12.16 p. 342 del libro di testo
Soluzione.
a) La media campionaria, pari a 0.45, è la stima della proporzione p di favorevoli al candidato A nell'intera popolazione. Il numero di favorevoli nel campione è 300 x 0.45 = 135, quella dei contrari 300 x 0.55 = 165: siccome entrambi questi valori sono > 30, sono soddisfatte le condizioni per l'uso della formula dell'intervallo di confidenza asintotico per la proporzione:

(2) $$\bar{x} - z_{\frac{\alpha}{2}}\sqrt{\bar{x}(1-\bar{x})/n} \leq p \leq \bar{x} + z_{\frac{\alpha}{2}}\sqrt{\bar{x}(1-\bar{x})/n}$$

che è del tutto simile alla formula (1), con la varianza stimata da $\bar{x}(1-\bar{x})/n$ perché la varianza della media campionaria di una popolazione bernoulliana deriva dalla varianza della Binomiale.

Con i dati dell'esercizio: $\bar{x}$ = 0.45, *n* = 300, $z_{\alpha/2}$ = 1.96 l'intervallo cercato è (0.394, 0.506), la cui lunghezza (risposta al punto b)) è 0.112.

Non svolgete il punto c).

d) Osserviamo intanto che l'intervallo di confidenza trovato al punto a) contiene, seppur di poco, valori superiori a 0.50.
Per rispondere non servono altri calcoli: se alziamo il livello di confidenza dell'intervallo a 99%, il corrispondente quantile $z_{0.005}$ è maggiore del quantile $z_{0.025}$ usato per l'intervallo al 95%, e quindi, dalla formula (2), l'intervallo certamente si allarga. Siccome già quello del 95% superava il valore 0.5, a maggior ragione adesso. Pertanto, il vero valore dei favorevoli al candidato A nella popolazione potrebbe anche essere > 50%. Tuttavia, non dobbiamo stupirci se, ad elezioni compiute e scrutinio ultimato, il candidato A risulti aver raccolto il 43% dei favori, perché anche 0.43 è un valore nell'intervallo di confidenza al livello del 95% (e anche del 99%).

Domanda aggiuntiva. Supponiamo ora di aver rilevato una percentuale di favorevoli ad A pari, ancora, al 45% ma in un campione di 1000 individui. Con riferimento all'intervallo di confidenza del 95% di cui al punto a), il nuovo intervallo di confidenza al 95% risulta più breve: da (2) si vede, infatti, che al crescere di *n* il termine $\sqrt{\bar{x}(1-\bar{x})/n}$ diminuisce. Il nuovo intervallo è (0.409, 0.490) e quindi, sulla base di un campione di 1000 intervistati, possiamo affermare con la confidenza del 99%, che il candidato A non otterrà la maggioranza assoluta alle elezioni.

Esercizio (da completarsi venerdì 16).

In una città si raccolgono informazioni sul consumo annuo di energia elettrica per unità abitativa. Nel caso di unità abitative di metratura confrontabile, la varianza del consumo di energia elettrica è un indicatore dei livelli di efficienza energetica, di interesse per l'impresa erogatrice così come per l'Amministrazione Locale.
In un campione di 101 unità abitative confrontabili si è osservata una deviazione standard campionaria del consumo di energia elettrica in un certo anno pari a 173.8 kWh.
a) Costruire un intervallo di confidenza al 95% per la varianza del consumo annuo di energia elettrica per unità abitative confrontabili a quelle del campione.
b) Fornire una limitazione superiore alla varianza, al livello del 95%.
Soluzione.
a) Un intervallo di confidenza al 95% per la varianza è della forma

(3) $$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right)$$

ove

$$S^2 = \frac{1}{n-1}\sum_i^n (X_i - \bar{X}_n)^2$$

è lo stimatore non distorto e consistente della varianza della popolazione.
Al denominatore nella formula (3) ci sono i quantili della distribuzione chi quadrato ad *n*-1 gradi di libertà: si veda la Figura 1.

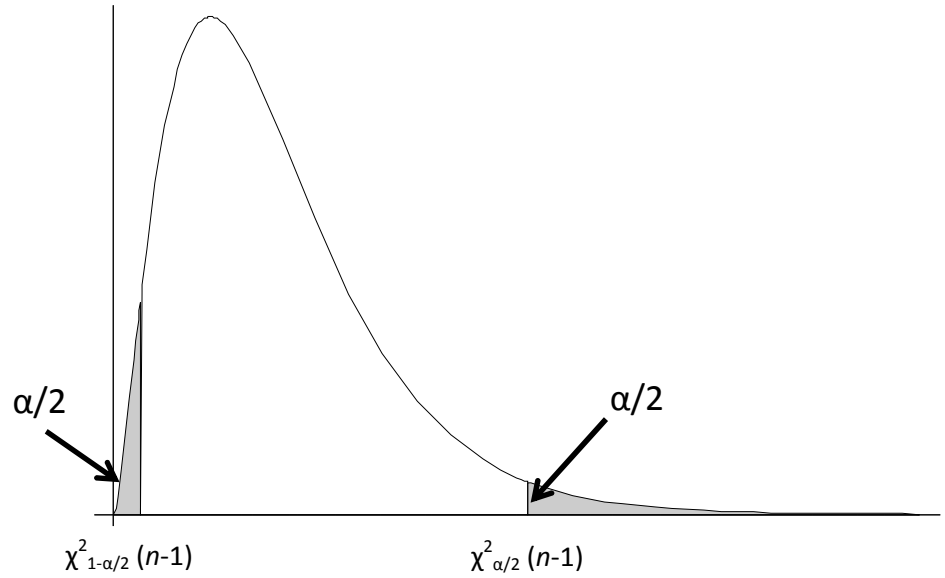

Figura 1 – Esempio di una densità chi quadrato e dei quantili usati nella formula (3).

Per i dati del problema: $\chi^2_{\alpha/2}(n-1)$ = 129.5613 mentre $\chi^2_{1-\alpha/2}(n-1)$ = 74.2219 e quindi l'intervallo cercato è (100 x 173.8^2 / 129.5316 , 100 x 173.8^2 / 74.2219) = (23314.4, 40697.5). Facendo la radice quadrata dei due estremi abbiamo l'intervallo di confidenza per la deviazione standard: (152.7, 201.7) kWh.

Svolgeremo il punto b) venerdì 16.

Esercizio 1 dal tema d'esame del 10.06.2013 (II)
Sia *p* la proporzione di italiani inclini a votare il partito XYZ alle prossime elezioni.
Da un'indagine effettuata tramite interviste la percentuale campionaria di favorevoli al partito è risultata essere del 48%. Quanto deve valere la dimensione campionaria *n* perché l'errore di stima di *p* sia inferiore a 0.05 con un livello di confidenza del 99%?
Soluzione
Usiamo ancora la formula (2). Ragionando come nel primo esercizio, l'errore massimo di stima (in valore assoluto) è

$$z_{\frac{\alpha}{2}}\sqrt{\bar{x}(1-\bar{x})/n}$$

(posso sbagliare per eccesso o per difetto), e quindi basta imporre 2.57583 x √(0.48x0.52/*n*) ≤ 0.05, da cui *n* ≥ 663.2 e quindi *n* ≥ 664. Quindi, se il risultato del 48% di favorevoli è stato ottenuto su un campione di almeno 664 individui, possiamo dire che l'intervallo di confidenza al 99% per *p* è contenuto nell'intervallo (0.48 - 0.05, 0.48 + 0.05) = (0.43, 0.53). Se il campione intervistato è di dimensione inferiore, l'incertezza, data dalla lunghezza dell'intervallo di confidenza, risulta maggiore.
Complemento 1. Possiamo chiederci che dimensione deve avere il campione per assicurarsi, al livello del 99%, che il partito XYZ non raggiunga la maggioranza assoluta. Significa determinare il valore di *n* per cui l'errore massimo sia pari a 0.02 (= 0.50 – 0.48). Con gli stessi calcoli si ottiene *n* > 4024.1, quindi *n* ≥ 4025. Si vede allora come ridurre l'errore apparentemente di poco richieda un grosso aumento della dimensione campionaria.

Complemento 2. Determiniamo la dimensione campionaria in anticipo, per avere un errore di stima non superiore a 0.05, qualunque sia l'esito dell'indagine.
Osserviamo che $p(1-p) \le \frac{1}{4}$ perché' la funzione $p(1-p)$ per p tra 0 ed 1 disegna una parabola con concavità rivolta verso il basso e vertice in $p = \frac{1}{2}$. Allora, qualunque sia la media campionaria $\bar{x}$ si ha

$$\sqrt{\bar{x}(1-\bar{x})/n} \le \sqrt{1/4n}$$

e pertanto, basta richiedere $z_{\alpha/2}\sqrt{(1/4n)} \le 0.05$ da cui, risolvendo, $n \ge 663.5$, cioè $n \ge 664$. Il risultato è uguale a quello di prima perché la media campionaria 0.48 è molto vicina a 0.50. Questo risultato, però, vale per qualunque valore della media campionaria si ottenga poi, a fine indagine.

Esercizio 2 dal tema d'esame del 1.06.2010
Di 10 individui sono stati annotati il sesso e la pressione arteriosa sistolica (la "massima") ottenendo:

Individuo	1	2	3	4	5	6	7	8	9	10
Sesso	F	M	F	F	F	M	F	F	M	M
Pressione	127	116	101	110	121	119	112	102	130	125

Si vuole testare l'ipotesi che la pressione media sia la stessa negli uomini e nelle donne.
1. Quali ipotesi bisogna fare?
(*2. Testare l'ipotesi di omoschedasticità. -> non svolto.*)
3. Dire se l'ipotesi di uguaglianza delle medie si può rifiutare al livello del 5%.
Soluzione.
1. Trattandosi di una dimensione campionaria molto bassa, possiamo solo supporre che i due campioni, gruppo dei maschi e gruppo delle femmine, siano indipendenti, che la pressione della popolazione dei maschi sia gaussiana con media μ_M e varianza σ^2_M incognite; che la pressione della popolazione delle femmine sia gaussiana con media μ_F e varianza σ^2_F incognite e che le due varianze siano uguali. Sotto queste ipotesi si può condurre un test

$$H_0 : \mu_M = \mu_F \text{ contro l'alternativa } H_1 : \mu_M \neq \mu_F .$$

Sotto l'ipotesi nulla di uguaglianza delle medie, mi aspetto che presi due campioni, uno di uomini e uno di donne, la differenza delle medie dei due campioni sia "vicina" a 0. Il valore prefissato di $\alpha = 5\%$ conduce a dare un significato al termine "vicina". Infatti, sotto le ipotesi prima elencate la statistica

$$T = \frac{\bar{X}_M - \bar{X}_F}{\sqrt{S_P^2\left(\frac{1}{n_M} + \frac{1}{n_F}\right)}}$$

ove

$$S_P^2 = \frac{(n_M - 1)S_M^2 + (n_F - 1)S_F^2}{n_M + n_F - 2}$$

ha distribuzione t di Student a $(n_M + n_F - 2)$ gradi di libertà.

Calcoliamo allora la media campionaria e la varianza campionaria della pressione sia per gli uomini che per le donne ed utilizziamo il fatto di conoscere la distribuzione esatta della statistica *T* per stabilire la probabilità di "osservare i risultati ottenuti dai due campioni". Si ha:

$$\bar{x}_M = 122.5$$
$$\bar{x}_F = 112.167$$
$$s^2_M = 39.0$$
$$s^2_F = 106.167$$
$$s^2_P = 80.9792$$

Quindi, dai nostri campioni si ricava che *T* = (122.5-112.167)/(√80.9792) = 1.14826. Questo valore è abbastanza vicino a 0? In Figura 2 è rappresentata la distribuzione di *T* assieme con la regione di rifiuto del test, cioè la regione tale che quando *T* sconfina lì dentro, non posso più dire che *T* è abbastanza vicina a 0. Infatti, se *T* cade nell'area grigia vuol dire che il risultato campionario cade in un intervallo di valori che ha probabilità molto bassa, < 0.05, di essere osservato se davvero la pressione media degli uomini è uguale alla pressone media delle donne.

Nel nostro caso si vede che il valore campionario di *T* è davvero abbastanza vicino a 0, al livello α del 5%, infatti rimane dentro l'area bianca.
In altre parole, usando lo stesso ragionamento dell'esercizio 2 dal tema d'esame del 18.9.2013 svolto alla scorsa esercitazione, se il valore campionario di *T* cade nell'area bianca allora la probabilità di osservare un valore campionario della differenza standardizzata delle medie nei due gruppi uguale o superiore a 1.14826 è abbastanza alto, superiore a 0.025 (= α/2). Non c'è quindi abbastanza evidenza nei dati per rifiutare l'ipotesi nulla di uguale media della pressione per uomini e donne.

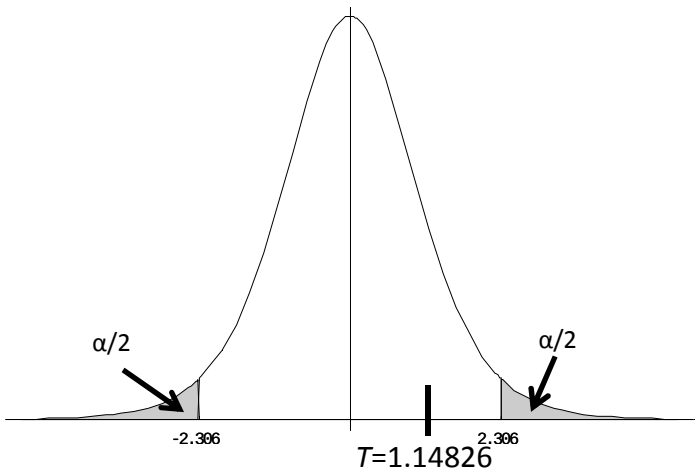

Figura 2 – Densità della distribuzione *t* di Student a 8 gradi di libertà: l'area grigia indica la regione di rifiuto del test al livello α, *T* = 1.1482 è il valore campionario della statistica test.