

costruzione della stima intervallare. Ma è anche una situazione che difficilmente si incontra in pratica. Un caso più realistico è quello in cui entrambi μ e σ^2 sono ignoti. Consideriamo allora ancora il caso di popolazione normale ma con anche σ^2 **ignota**. Questa volta l'informazione ausiliaria a priori è che sia *noto* oppure sia *realisticamente ipotizzabile* che il fenomeno di interesse è interpretabile dalla v.c. $X \sim N(\mu, \sigma^2)$ con *entrambi* i parametri ignoti.

18.3 Studentizzazione e variabile casuale T di Student

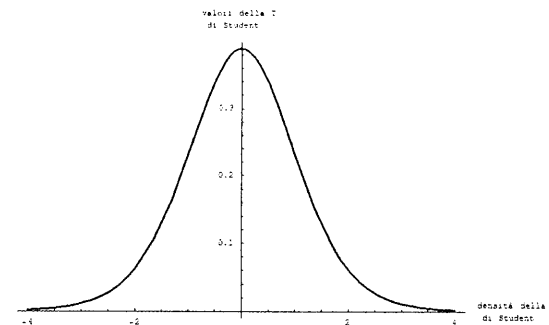
Siamo ancora sotto la condizione iniziale: $X \sim N(\mu, \sigma^2)$. Allora il teorema $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ vale ancora. Ora però non conosciamo il valore σ^2 e così non possiamo più standardizzare e usare le tavole della Z . Che facciamo di fronte a un parametro ignoto? Lo stimiamo! La stima per σ^2 è la varianza campionaria corretta con i gradi di libertà (chi non se lo ricorda, vada a rileggersi il paragrafo 17.10): $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$. Se nella standardizzazione sostituiamo la stima s^2 all'ignota σ^2 si effettua una diversa trasformazione che chiamiamo **studentizzazione**.

$$\text{Studentizzazione di } \bar{X}: \frac{\bar{X} - \mu}{\sqrt{S^2/n}}$$

Notate che per essere del tutto corretti dal punto di vista teorico, abbiamo scritto la studentizzazione usando le maiuscole anziché le minuscole. Dovrebbe essere chiaro a questo punto che questo significa usare gli *stimatori* cioè tenere sotto controllo *tutti* i possibili campioni di ampiezza n che ci potrebbero capitare. Se non è ancora del tutto chiaro, consiglio di ripartire dal Capitolo 16.

Un altro teorema di teoria delle probabilità ci assicura che lo stimatore media campionaria studentizzata (anziché standardizzata) non è più una $Z \sim N(0,1)$ ma è un'altra v.c. chiamata **T di Student**.

La T di Student è una v.c. *diversa* dalla $Z \sim N(0,1)$ ma la sua funzione di densità è molto simile: ha sempre forma **campanulare**, è centrata sullo 0 ma ha varianza più grande di 1. Rispetto alla Z ha, come si dice, le code *più pesanti* cioè le code della campana sono un po' più lontane dall'asse delle ascisse:



La v.c. T di Student ha un solo parametro detto **gradi di libertà**. Per di popolazione normale con entrambi i parametri ignoti, studentizzando lo stimatore media campionaria si ottiene una T di Student con $n - 1$ gradi di libertà, pari cioè a quelli che usiamo per *correggere* la varianza campionaria s^2 che mettiamo al denominatore:

$$\frac{\bar{X} - \mu}{\sqrt{S^2/n}} = T_{n-1}$$

Anche per la v.c. T di Student esistono le **tavole** per individuarne probabilità e valori (quando non abbiamo sottomano un computer). Le tavole della T di Student sono pubblicate su qualunque manuale di Statistica, si trovano facilmente su Internet e sono anche riportate alla fine di questo capitolo.

18.4 Intervallo di confidenza per la media μ con popolazione normale e σ^2 ignota

La metodologia di costruzione è analoga a quella con σ^2 nota. La differenza consiste nell'utilizzare la T di Student con $(n - 1)$ gradi di libertà anziché la $Z \sim N(0,1)$:

1. si estrae un campione bernoulliano di ampiezza n e ci si procura i dati (campionari): $(x_1 \cdots x_i \cdots x_n)$;
2. si calcolano le stime puntuali per *entrambi* i parametri ignoti:

$$\text{la media campionaria: } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \text{ per } \mu,$$

$$\text{la varianza campionaria (corretta): } s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \text{ per } \sigma^2;$$

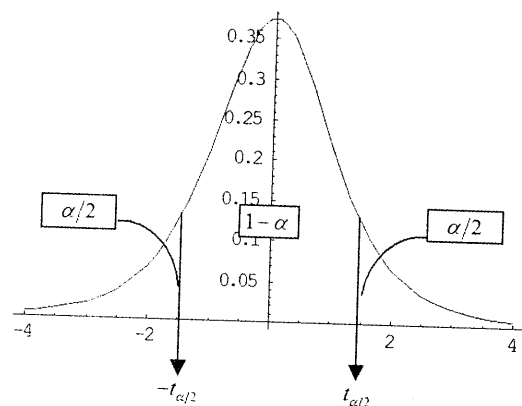
3. si sceglie il livello di confidenza $(1 - \alpha)$ da cui si ottiene la probabilità di sbagliare α e le probabilità $\alpha/2$ delle code.

4. si studentizza lo stimatore media campionaria $\bar{X}$ e si ottiene la v.c. T di Student con $(n - 1)$ gradi di libertà: $\frac{\bar{X} - \mu}{\sqrt{S^2/n}} = T_{n-1}$.

Sulla T (anziché sulla Z) si applica una teoria in tutto simile (passaggi e ragionamenti) a quella che abbiamo visto al passo 4. del paragrafo precedente. Quello che cambia è che salteranno fuori dei **T-score** (anziché degli Z-score):

$$P\left(-t_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sqrt{S^2/n}} \leq t_{\alpha/2}\right) = 1 - \alpha$$

Graficamente:



Il T-score positivo $t_{\alpha/2}$ si trova sulle tavole della T di Student. Il suo simmetrico $-t_{\alpha/2}$ si ottiene cambiando semplicemente il segno.

Ora invertiamo la (doppia) disuguaglianza all'interno delle parentesi tonde e riscriviamo la probabilità in modo che l'intervallo sia *centrato* sul parametro μ che vogliamo stimare:

$$P\left(\bar{X} - t_{\alpha/2}\sqrt{\frac{S^2}{n}} \leq \mu \leq \bar{X} + t_{\alpha/2}\sqrt{\frac{S^2}{n}}\right) = 1 - \alpha$$

5. Sostituendo i dati campionari si ottiene finalmente:

IC per μ a l.c. $(1 - \alpha)$ con popolazione normale e σ^2 ignota:

$$\left[\bar{x} - t_{\alpha/2}\sqrt{\frac{s^2}{n}}, \bar{x} + t_{\alpha/2}\sqrt{\frac{s^2}{n}} \right]$$

ESEMPIO: UNA NUOVA DIDATTICA

Una docente universitaria sta sperimentando una nuova metodologia didattica. Per valutare gli effetti della sua nuova metodologia vuole una stima (intervallare) della votazione media riportata agli esami scritti dagli studenti sottoposti alla nuova didattica. Seleziona casualmente 10 studenti e ne rileva il voto all'esame scritto (in 30-esimi):

Studente	1	2	3	4	5	6	7	8	9	10
Voto/30	26	21	26	19	23	23	30	19	28	21

La docente assume che la votazione agli esami scritti per tutti i suoi studenti segua realisticamente una distribuzione normale. Per la stima intervallare sceglie il 99% di confidenza.

Chiamiamo (come al solito) X la v.c. che interpreta il fenomeno di interesse "voto all'esame scritto" sulla U di studenti sottoposti alla nuova didattica. Siamo nel caso di popolazione normale e σ^2 ignota. La votazione media μ è il parametro che si vuole stimare. Il sottoinsieme di 10 studenti selezionati casualmente rappresenta un campione bernoulliano di ampiezza $n=10$; la tabella fornisce i dati campionari:

$$(x_1, x_2, \dots, x_{10}) = (26, 21, 26, 19, \dots, 21)$$

Calcoliamo le stime (puntuali) per entrambi i parametri ignoti:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{10} (26 + 21 + 26 + \dots + 28 + 21) = 23.6 \quad \text{per } \mu$$

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

$$= \frac{1}{9} \left[(26 - 23.6)^2 + (21 - 23.6)^2 + \dots + (21 - 23.6)^2 \right] = 14.267$$

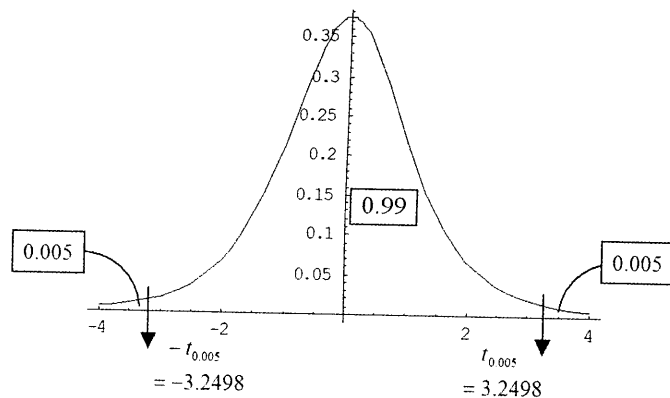
Studentizziamo:

$$\frac{\bar{X} - \mu}{\sqrt{S^2/n}} = T \quad \text{con } (n-1) = 9 \text{ gradi di libertà}$$

Il l.c. è scelto pari a 99%; calcoliamo allora la probabilità $(1 - \alpha)$ e le probabilità $\alpha/2$ delle code:

$$1 - \alpha = 0.99 \quad \alpha = 0.01 \quad \alpha/2 = 0.005$$

Sulle tavole della T di Student, alla riga 9 gradi di libertà e colonna 0.005 (oppure col computer) troviamo il T-score $t_{\alpha/2} = t_{0.005} = 3.2498$ che lascia a destra una probabilità pari a 0.005. Graficamente:



Infine calcoliamo l'IC per μ :

$$\left[\bar{x} - t_{\alpha/2} \sqrt{\frac{s^2}{n}}, \bar{x} + t_{\alpha/2} \sqrt{\frac{s^2}{n}} \right]$$

$$= \left[23.6 - 3.2498 \times \sqrt{\frac{14.267}{10}}, 23.6 + 3.2498 \times \sqrt{\frac{14.267}{10}} \right]$$

$$= [19.72, 27.48]$$

Ci dice che la docente può *confidare* al 99% che l'intervallo di voti compresi fra 19.72 e 27.48 *contenga* la votazione media all'esame scritto per i suoi studenti sottoposti alla nuova didattica.

Provate a controllare, per esercizio, che questo IC è un po' più *largo* di quello che si otterrebbe usando le tavole della Z (sbagliato!) anziché quelle della T (corretto!).

18.5 Intervalli di confidenza (approssimati) per grandi campioni per la media e per la %

Ora ci mettiamo nel caso in cui non si sa *nulla* circa il fenomeno in U : non si hanno informazioni ausiliarie a priori, oppure non si ritiene realistica l'ipotesi che la popolazione sia normale oppure ancora si sa che la popolazione *non* è normale. Consideriamo cioè tutte le situazioni non previste nei paragrafi precedenti.

Nonostante la v.c. normale funzioni bene per molti fenomeni (l'abbiamo visto nel Capitolo 15) nelle applicazioni molti altri casi rimangono fuori. Pensiamo ad esempio al fenomeno "reddito netto mensile da lavoro dipendente in Italia" oppure "risparmio annuo delle famiglie italiane" oppure "voto finale dei laureati

italiani nel 2006-07". Sono realisticamente modellabili con una curva a campana e simmetrica? Sarebbe come dire che in Italia i lavoratori dipendenti percepiscono in massa uno stipendio più o meno *medio* e quelli che guadagnano tantissimo sono tanti quanto quelli vicino o al di sotto della soglia di povertà; sarebbe come assumere che le famiglie più ricche risparmino come le famiglie più povere o che i bravi che si laureano con la lode sono tanti quanto quelli che si laureano con il minimo dei voti (...invece i dati ci dicono che sono di più i bravi, anzi le brave). Siamo poi certamente in un caso di popolazione *non* normale quando vogliamo stimare una percentuale o una frequenza relativa di un fenomeno dicotomico. Sappiamo infatti che in questo caso la v.c. adeguata è la binomiale (...e chi non lo ricorda rileggi i paragrafi 15.5 e 17.13).

Costruire un IC richiede comunque, rispetto alla stima puntuale, delle informazioni *in più*. Se non abbiamo informazioni ausiliarie a priori su X , cioè se non siamo nel caso di popolazione normale, dobbiamo allora avere *molti dati* cioè essere nel caso di **grandi campioni**. *Solo* se il campione è, come si dice, *sufficientemente grande* possiamo infatti appellarci a un teorema di teoria delle probabilità fondamentale (centrale!) nell'inferenza statistica. Questo teorema si chiama (appunto) **teorema centrale del limite** e lo indicheremo brevemente con TCL. Noi non abbiamo abbastanza Matematica per dimostrare il TCL e neanche per enunciarlo rigorosamente. Ci limitiamo ad intuire *come e quando* utilizzarlo per risolvere il nostro problema di costruzione degli IC.

Qualunque sia la distribuzione del fenomeno X in U , se l'ampiezza campionaria n tende all'infinito, in formule $n \rightarrow \infty$, allora gli stimatori media campionaria $\bar{X}$ (stimatore per μ) e frequenza relativa campionaria $\hat{P}$ (stimatore per p o anche per la percentuale $100 \cdot p$) sono normali. Questo è il risultato teorico. Nella pratica i campioni possono essere grandi ma (ovviamente) non sono infiniti. Allora il risultato teorico si può usare così: quando n è **sufficientemente grande**, gli stimatori $\bar{X}$ e $\hat{P}$ sono **approssimativamente** normali (tecnicamente si dice: **asintoticamente normali**). In formule indicheremo la distribuzione asintotica con il simbolo $\approx$ (per distinguerla dalla distribuzione esatta $\sim$):

Per il TCL, se n è **sufficientemente grande** vale:

$$\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right) \text{ e } \hat{P} \approx N\left(p, \frac{p(1-p)}{n}\right)$$

Siccome con n sufficientemente grande, grazie al TCL abbiamo ritrovato la Normale, allora per *grandi campioni* possiamo usare la metodologia degli IC basata sulla $Z \sim N(0,1)$. Attenzione: qui la normalità è *approssimata* e conseguentemente si tratterà di **IC approssimati per grandi campioni** con un *effettivo* IC **approssimativamente** pari all' $(1 - \alpha)$ scelto. Gli IC approssimati per grandi campioni si usano per la media μ quando non si può assumere la normalità della popolazione