Dispensa formule - Riassunto Statistica

Statistica / Statistics (Università Commerciale Luigi Bocconi)

Capitolo 1

Popolazione: insieme completo di tutte le unità di oggetto di studio. La dimensione N può essere infinita.

Campione: sottoinsieme delle unità osservate nella popolazione. La dimensione è *n* .

Parametro: caratteristica specifica della popolazione.

Statistica: caratteristica specifica del campione.

Statistica descrittiva: comprende metodici grafici e numerici, utilizzati per sintetizzare ed elaborare dati per trasformarli in informazioni.

Statistica inferenziale: fornisce le basi per le previsioni e le stime che consentono di trasformare le informazioni in conoscenza.

Capitolo 2

Unità Statistica: oggetto dell'osservazione di ogni fenomeno individuale che costituisce il fenomeno collettivo. L'insieme delle unità statistiche costituisce il collettivo o popolazione statistica.

Carattere: aspetto rilevato in corrispondenza di ogni unità statistica.

Modalità: differenti forme secondo cui si manifesta il carattere. Sono le categorie o i valori che ciascun carattere presenta in corrispondenza di ogni unità statistica.

Dati Qualitativi: appartengono a gruppi o categorie. Sono divisi in:

- Nominali: la codifica numerica è scelta per pura convenienza; Scala Nominale. Ex/ sesso, cittadinanza, orientamento
- Ordinali: indicano un ordine gerarchico degli elementi; Scala Ordinale. Ex/ voto scolastico, livello di soddisfazione

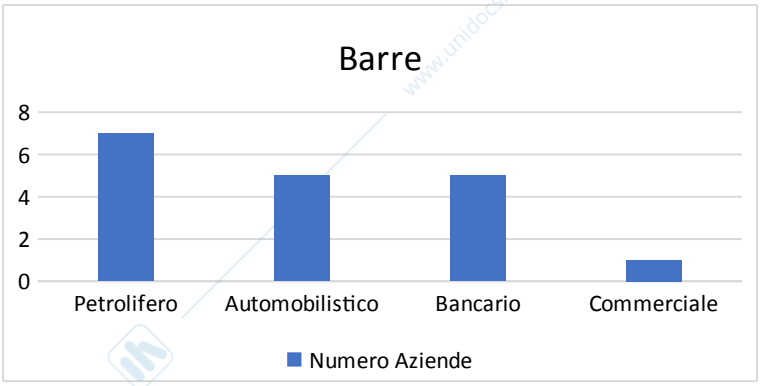
Dati Quantitativi: assumono valori numerici. Sono divisi in:

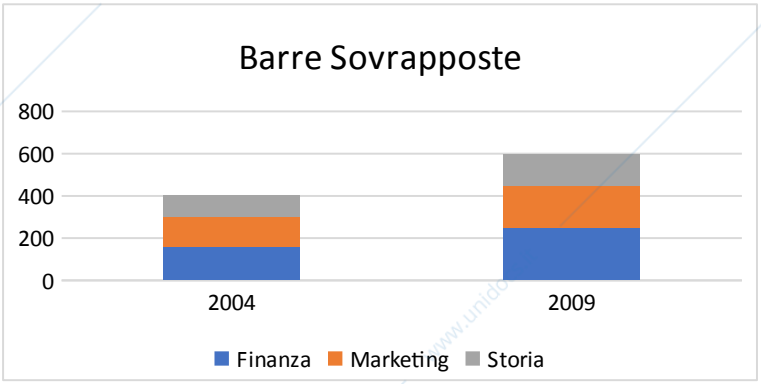
- Discreti: hanno un numero finito di valori e sono generati da un conteggio; Scala Non Rapporto/Intervalli. Ex/ studenti iscritti, numero azioni, volumi di vendita
- Continui: possono assumere un qualunque valore all'interno di un determinato intervallo e sono generati da una misurazione; Scala Rapporto. Ex/ altezza, peso, temperatura, distanza

Distribuzione di Frequenza: tabella per organizzare i dati.

Modalità o Classe (ex/ Settore attività)	Frequenza (ex/ Numero di aziende)
Petrolifero	7
Automobilistico	5
Bancario	5
Commerciale	1

Diagramma a barre: qualitativo ordinale/nominale.





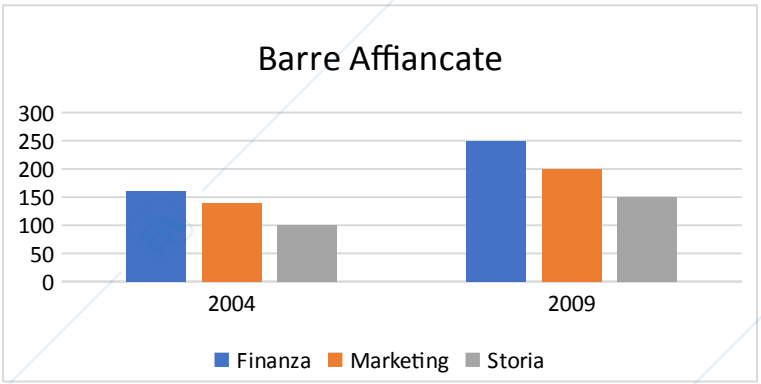


Diagramma a Torta: qualitativo ordinale/nominale.

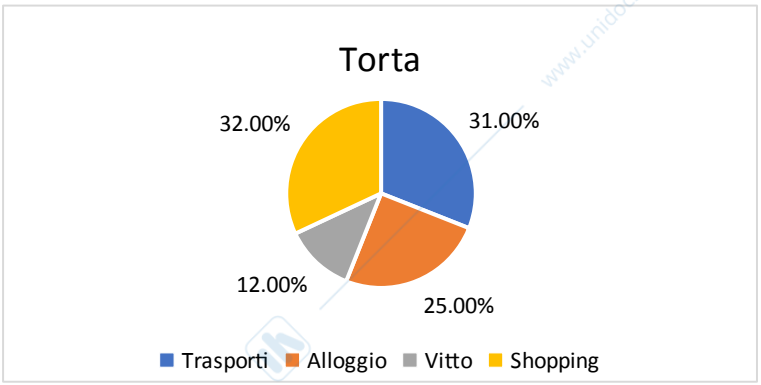
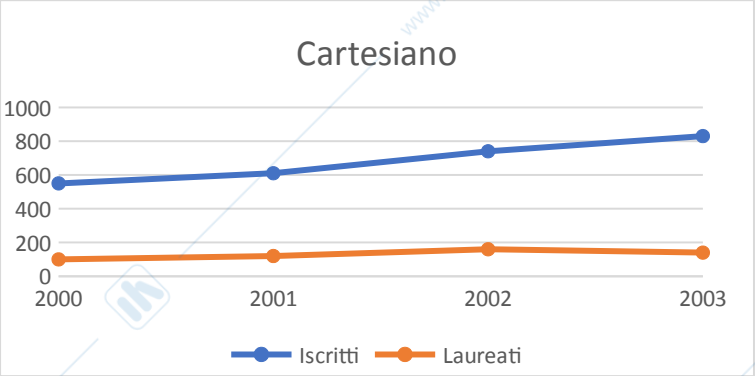
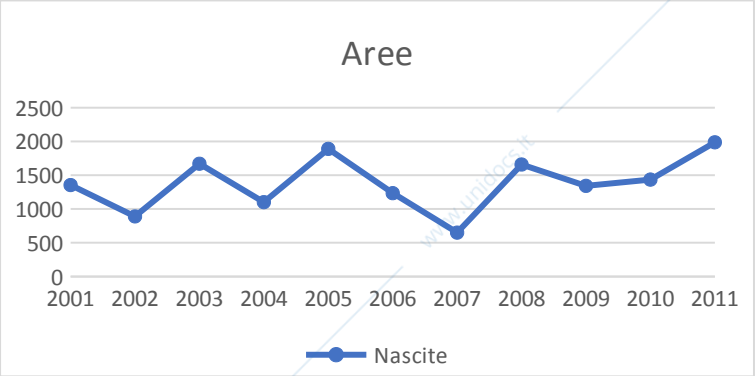


Diagramma di Pareto: rappresenta le frequenze delle cause di difettosità. Serve per separare le poche cause rilevanti dalle numerose insignificanti.

Cartesiano: rappresenta una serie storica in istanti di tempo diversi. Quantitativo discreto.



Aree: quantitativo continuo.



Ampiezza dell'intervallo: $w = \frac{(Val.MAX - Val.min)}{N° classi}$

Distribuzione frequenze relative: ottenuta dividendo ciascuna frequenza per il numero complessivo di osservazioni.

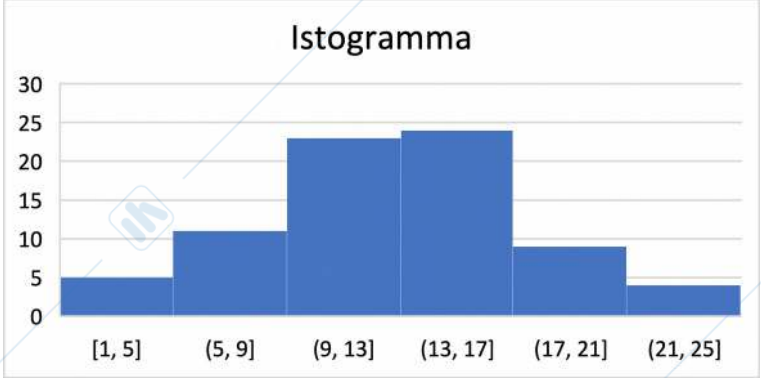
Distribuzione frequenze cumulate: contiene il numero totale di osservazioni con valori minori del limite superiore di ciascuna classe.

Distribuzione frequenze relative cumulate: si cumulano le frequenze relative.

Uso del cellulare (minuti)	Frequenza n_i	Frequenza cumulata N_i	Frequenza relativa f_i	Frequenza relativa cumulata F_i
220;230	5	5	0,0454	0,0454
230;240	8	13	0,0727	0,1182
240;250	13	26	0,1182	0,2364
250;260	22	48	0,2	0,4364
260;270	32	80	0,2909	0,7273
270;280	13	93	0,1182	0,8455
280;290	10	103	0,0909	0,9364
290;300	7	110	0,0636	1
	110		1	

Densità di frequenza: $c_i = \frac{f_i}{ampiezza\ classe}$

Istogramma: quantitativo continuo. Se le classi hanno tutte la stessa ampiezza, l'altezza di ciascun rettangolo è proporzionale al numero di osservazioni della classe; altrimenti sarà uguale alla densità di frequenza.



Simmetria: un istogramma è simmetrico se le osservazioni sono bilanciate o distribuite intorno al centro.

Asimmetria: una distribuzione è asimmetrica o obliqua se le osservazioni non sono distribuite in modo simmetrico rispetto al valore centrale.

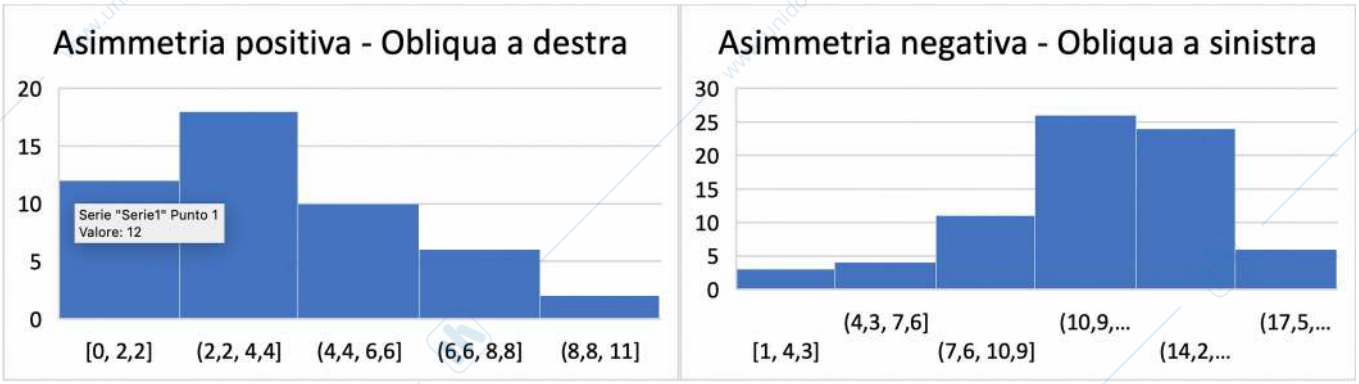


Diagramma di Dispersione: quantitativi discreti. Rappresenta l'osservazione congiunta di due variabili. Evidenzia:

- Eventuale relazione tra le due variabili
- Presenza di eventuali valori anomali (outlier)

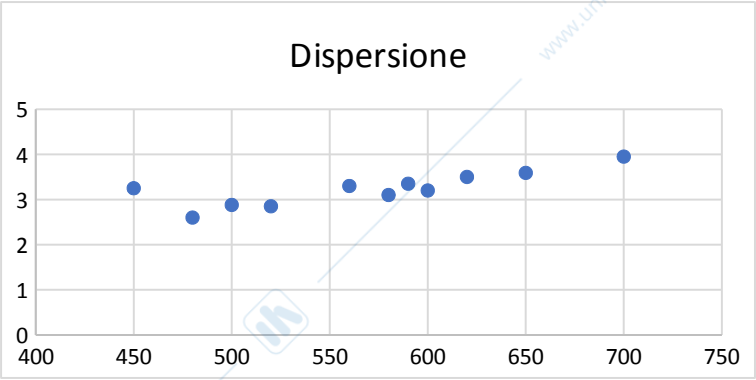


Tabella a doppia entrata: elenca la frequenza delle osservazioni per ogni combinazione di classi di misura di due variabili. È indicata come *r x c*

Tabella di contingenza: entrambe le variabili sono qualitative.

Zona	Attrezzi	Legname	Vernici	Altri	**Totale**
Nord	100	50	50	50	**250**
Sud	50	95	45	60	**250**

Totale	**150**	**145**	**95**	**110**	**500**

Capitolo 3

Media: somma dei valori di tutte le osservazioni divisa per il numero di osservazioni.

- Dati non raggruppati: $\acute{x}=\frac{\sum x_i}{n}$
- Dati raggruppati (classi): $\acute{x}=\frac{\sum x_i \cdot f_i}{\sum f_i}$

Mediana: osservazione centrale di un insieme di osservazioni.

- Dati non raggruppati:
  - *n* dispari: $Me=(N+1)\cdot 0{,}5$
  - *n* pari: $Me=\frac{N\cdot 0{,}5+(N+1)\cdot 0{,}5}{2}$
- Dati raggruppati (classi): $Me=L_{inf}+(p-F_{inf})\cdot(\frac{\Delta_i}{f_i})$

Moda: modalità che si presenta il maggior numero di volte (dati qualitativi).

Outlier: osservazioni eccezionalmente elevate/basse che tendono a far aumentare/diminuire la media rispetto alla mediana, determinando asimmetria positiva/negativa.

- Outlier inferiori: $T_1=Q_1-1{,}5\cdot D.I.$
  - $T_1<min$ → NO outlier inferiori
  - $T_1>min$ → SI outlier inferiori
- Outlier superiori: $T_2=Q_3+1{,}5\cdot D.I.$
  - $T_2>MAX$ → NO outlier superiori
  - $T_2<MAX$ → SI outlier superiori

Campo di variazione (range): differenza tra il masimo e il minimo dei valori osservati.

$$gamma=x_{MAX}-x_{min}$$

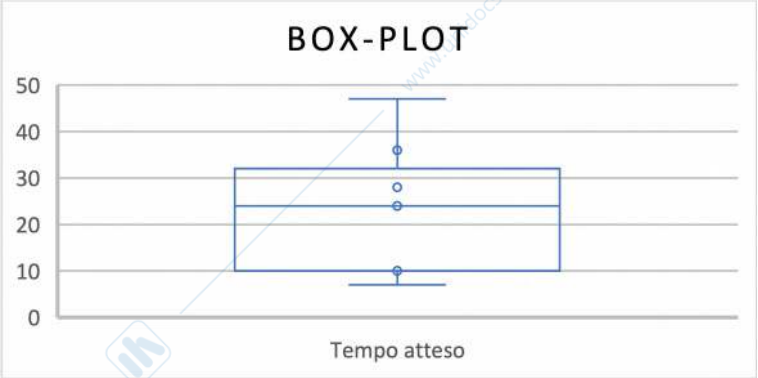
Quartili: misure di tendenza non centrale ottenute a partire dalle frequenze cumulate delle osservazioni.

- Dati non raggruppati:
  - $Q_1=(N+1)\cdot 0{,}25$
  - $Q_2=(N+1)\cdot 0{,}5$
  - $Q_3=(N+1)\cdot 0{,}75$
  - $P_{80}=(N+1)\cdot 0{,}8$
- Dati raggruppati (classi): $Me=L_{inf}+(p-F_{inf})\cdot(\frac{\Delta_i}{f_i})$

Differenza Interquartile: misura la variabilità del 50% centrale dei dati.

$$D.I.=Q_3-Q_1$$

Box Plot: esprime graficamente i cinque numeri di sintesi.



Varianza: misura quanto i valori osservati si discostino quadraticamente rispetto alla media aritmetica.

	Popolazione	Campione
Implicita	$\sigma^2=\frac{\sum(x_i-\acute{x})^2}{N}$	$s^2=\frac{n}{n-1}\left(\frac{\sum(x_i-\acute{x})^2}{n}\right)$
Esplicita	$\sigma^2=\frac{\sum x_i^2}{N}-\acute{x}^2$	$s^2=\frac{n}{n-1}\left(\frac{\sum x_i^2}{n}-\acute{x}^2\right)$
Dati raggruppati	$\sigma^2=\frac{\sum(x_i-\acute{x})^2\cdot f_i}{N}$	$s^2=\frac{n}{n-1}\left(\frac{\sum(x_i-\acute{x})^2\cdot f_i}{n}\right)$

Deviazione standard (s.q.m.): esprime la dispersione dei dati intorno ad un indice di posizione, quale può essere, ad esempio, la media aritmetica.

Popolazione	Campione
$\sigma=\sqrt{\sigma^2}$	$s=\sqrt{s^2}$

Coefficiente di Variazione: permette di valutare la dispersione dei valori attorno alla media.

Popolazione	Campione
$CV=\frac{\sigma}{[\acute{x}]}$	$CV=\frac{s}{[\acute{x}]}$

Covarianza: misura la relazione lineare tra due variabili. Un valore positivo indica una relazione diretta o positiva, un valore negativo indica una relazione inversa o negativa.

	Popolazione	Campione
Implicita	$Cov(X,Y)=\sigma_{xy}=\frac{\sum(x_i-\acute{x})\cdot(y_i-\acute{y})}{N}$	$Cov(X,Y)=s_{xy}=\frac{n}{n-1}\left(\frac{\sum(x_i-\acute{x})\cdot(y_i-\acute{y})}{n}\right)$
Esplicita	$Cov(X,Y)=\sigma_{xy}=\frac{\sum x_i\cdot y_i-\frac{\sum x_i\cdot y_i}{N}}{N}$	$Cov(X,Y)=s_{xy}=\frac{n}{n-1}\left(\frac{\sum x_i\cdot y_i}{n}-\acute{x}\cdot\acute{y}\right)$

Coefficiente di Correlazione Lineare: esprime una relazione di linearità tra due variabili. Varia tra +1 e -1.

- $0<r<+1$: relazione lineare positivia
- $-1<r<0$: relazione lineare negativa
- $r=0$: nessuna relazione lineare

Popolazione	Campione
$\rho=\frac{Cov(X,Y)}{\sigma_x\cdot\sigma_y}$	$r=\frac{Cov(X,Y)}{s_x\cdot s_y}$

Retta di Regressione: mostra l'effetto sulla variabile dipendente *Y* di un cambiamento della variabile indipendente *X*.

$$Y=b_0+b_1\cdot X$$

Pendenza: variazione di *Y* per ogni variazione unitaria di *X*.

$$b_1 = \frac{Cov(X, Y)}{\sigma_x^2}$$

Origine: $b_0 = \acute{y} - b_1 \cdot \acute{x}$

Devianza: fornisce un grado di dispersione di una certa variabile dal proprio valore mediano.

$$D(X) = \sum (x_i - \acute{x})^2$$

Capitolo 4

Esperimento aleatorio: processo che porta a due o più risultati senza poter prevedere quale si realizzerà.

Eventi elementari: possibili risultati di un esperimento casuale.

Spazio campionario: insieme degli eventi elementari.

Evento (E): un qualsiasi sottoinsieme di eventi elementari di uno spazio campionario. Un evento si verifica quando il risultato dell'esperimento casuale è uno degli eventi che lo costituiscono.

Intersezione: insieme di tutti gli eventi elementari di *S* che appartengono sia ad *A* che a *B*.

$$A=\{1,3,5\} \qquad B=\{2,3,4\} \qquad A\cap B=\{3\}$$

Unione: insieme di tutti gli eventi complementari di *S* che appartengono ad almeno uno dei due eventi.

$$A=\{1,3,5\} \qquad B=\{2,3,4\} \qquad A\cup B=\{1,2,3,4,5\}$$

Evento complementare: insieme degli eventi elementari appartenenti ad *S* ma non ad *A*.

$$S=\{1,2,3,4,5,6\} \qquad A=\{1,3,5\} \qquad \acute{A}=\{2,4,6\}$$

Definizione classica: la probabilità di un evento è la proporzione di volte che l'evento si verifica.

$$P(A)=\frac{N_A}{N}$$

Combinazioni: $c_k^n=\binom{n}{k}=\frac{n!}{k!(n-k)!}$

Regola evento complementare: $P(\acute{A})=1-P(A)$

Regola additiva: $P(A\cup B)=P(A)+P(B)-P(A\cap B)$

Probabilità condizionata: sapendo che l'evento *B* si è verificato, la probabilità condizionata dell'evento *A*:

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

Regola moltiplicativa: $P(A\cap B)=P(A|B)\cdot P(B)$

Indipendenza statistica: due eventi sono statisticamente indipendenti se e solo se:

$$P(A\cap B)=P(A)\cdot P(B)$$

Capitolo 5

Variabile aleatoria: assume valori numerici in corrispondenza ai risultati di un esperimento aleatorio.

Variabile aleatoria discreta: assume al più un insieme numerabile di valori.

Variabile aleatoria continua: assume un qualunque valore in un intervallo.

Funzione di probabilità *P(x)*: esprime la probabilità che *X* assuma il valore *x*, come funzione di *x*, per una variabile aleatoria discreta *X*.

$$P(x)=P(X=x) \qquad \text{per ogni valore di } x$$

Funzione di ripartizione: esprime la probabilità che *X* non superi il valore x_0, come funzione di x_0, per una variabile aleatoria *X*.

$$F(x_0)=P(X\le x_0) \qquad -\infty<x_0<+\infty$$

Valore atteso: $E(X)=\mu=\sum_x xP(x)$

Varianza: $\sigma^2 = Var(X) = E\left[(X-\mu)^2\right] = \sum_x (x-\mu)^2 P(x)$

Distribuzione Binomiale: definisce la distribuzione di probabilità di *n* prove ripetute e indipendenti, quando i risultati possibili di ciascuna prova possono essere soltanto due: Successo *p*, Insuccesso *(1-p)*

Funzione Probabilità	Media	Varianza
$P(x)=\binom{n}{x}\cdot p^x\cdot(1-p)^{n-x}$	$\mu=n\cdot p$	$\sigma^2=n\cdot p\cdot(1-p)$

p: successo %

$(1-p)$: insuccesso %

x: n° successi

$(n-x)$: n° insuccessi

Distribuzione Normale: è una distribuzione di probabilità continua spesso usata come prima approssimazione per descrivere variabili casuali a valori reali che tendono a concentrarsi attorno a un singolo valor medio.

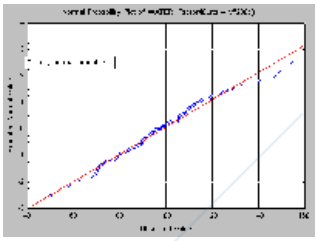
	Media	Varianza
$X \quad N(\mu,\sigma^2)$	μ	σ^2

Calcolo probabilità in un intervallo: $P(a<X<b)=F(b)-F(a)$

Standardizzazione: per dimostrare che $P_x(X)$ è effettivamente una funzione di densità di probabilità, si ricorre innanzi tutto alla standardizzazione della variabile casuale, cioè alla trasformazione tale per cui risulta:

$$Z=\frac{X-\mu}{\sigma}$$

Normal Probability Plot: metodo per verificare l'ipotesi di normalità e stabilire se il modello normale possa fornire buone approssimazioni per la distribuzione effettiva. Se i dati seguono la distribuzione normale il Plot sarà una linea retta.



Approssimazione Binomiale a Normale: quando il numero *n* delle prove è grande, il calcolo con la distribuzione binomiale è molto lungo. Per facilitare il calcolo, conviene approssimarla con la distribuzione normale.

$$Z=\frac{X-np}{\sqrt{np(1-p)}}$$

La regola pratica da seguire per capire se si può approssimare con la normale consiste nel verificare se valgono entrambe le seguenti condizioni:

$np\geq 5 \quad np(1-p)\geq 5$

Capitolo 8

Stimatore: considerato un parametro della popolazione, è una variabile aleatoria funzione delle variabili campionarie, i cui valori forniscono approssimazioni per il parametro non noto.

Stima: ogni singolo valore ricavato grazie allo stimatore.

Stimatore puntuale: considerato un parametro della popolazione (come la media), lo stimatore puntuale è una funzione delle variabili campionarie che determina un unico valore.

Stima puntuale: valore indicato dallo stimatore puntuale (ad esempio la media campionaria $\acute{X}$ è uno stimatore puntuale della media μ della popolazione e il valore che assume in corrispondenza di una particolare realizzazione campionaria viene chiamata stima puntuale $\acute{x}$).

Stimatore non distorto: uno stimatore puntuale $\hat{\theta}$ viene definito non distorto (o corretto) per il parametro della popolazione θ se il suo valore atteso coincide con il parametro stesso.

$$E(\hat{\theta})=\theta$$

Distorsione: sia $\hat{\theta}$ uno stimatore di θ, la distorsione è definita come la differenza tra la sua media e θ:

$$D(\hat{\theta})=E(\hat{\theta})-\theta$$

Stimatore asintoticamente non distorto: la differenza tra il valore atteso dello stimatore puntuale e il parametro da stimare diminuisce al crescere dell'ampiezza del campione. All'aumentare dell'ampiezza del campione, la distorsione diventa sempre più piccola.

Stimatore efficiente: su più stimatori non distorti per lo stesso parametro, è quello con la varianza più piccola.

Stimatore per intervallo: determina gli estremi di un intervallo di valori che verosimilmente contiene il parametro da stimare.

Stima per intervallo: stima determinata dallo stimatore per intervallo.

Capitolo 9

Intervalli di confidenza per la media:

			MARGINE D'ERRORE= VAR. CAS. X ERR. STD.			
PARAMETRO (θ)		STIMA (θ_0 ¿)	VARIABILE CASUALE	ERRORE STANDARD	NOTE	
Media	μ	$\acute{x}$	$z_{1-\frac{\alpha}{2}}$	$\sqrt{\frac{\sigma^2}{n}}$	σ^2 nota	
			$z_{1-\frac{\alpha}{2}}$	$\sqrt{\frac{s_c^2}{n}}$	σ^2 ignota – Grandi Campioni (n>30)	$s_c^2=s^2\cdot\left(\frac{n}{n-1}\right)$
			$_{n-1}t_{\frac{\alpha}{2}}$	$\sqrt{\frac{s_c^2}{n}}$	σ^2 ignota – Piccoli Campioni (n<30)	$s_c^2=s^2\cdot\left(\frac{n}{n-1}\right)$

Intervalli di confidenza per la proporzione:

PARAMETRO (θ)		STIMA (θ_0 ¿	MARGINE D'ERRORE= VAR. CAS. X ERR. STD.		NOTE
			VARIABILE CASUALE	ERRORE STANDARD	
Proporzione	π	p	$z_{1-\frac{\alpha}{2}}$	$\sqrt{\frac{p(1-p)}{n}}$	Grandi campioni (n>30)

Intervalli di confidenza per la differenza tra medie:

PARAMETRO (θ)		STIMA (θ_0 ¿	MARGINE D'ERRORE= VAR. CAS. X ERR. STD.		NOTE	
			VARIABILE CASUALE	ERRORE STANDARD		
Differenza tra Medie	$\mu_1-\mu_2$	$\acute{x}_1-\acute{x}_2$	$z_{1-\frac{\alpha}{2}}$	$\sqrt{\left(\frac{\sigma_1^2}{n_1}\right)+\left(\frac{\sigma_2^2}{n_2}\right)}$	σ^2 nota	
			$z_{1-\frac{\alpha}{2}}$	$\sqrt{\left(\frac{s_{c1}^2}{n_1}\right)+\left(\frac{s_{c2}^2}{n_2}\right)}$	σ^2 ignota – Grandi Campioni (n>30)	$s_c^2=s^2\cdot\left(\frac{n}{n-1}\right)$
			$_{n_1+n_2-2}t_{\frac{\alpha}{2}}$	$\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$	σ^2 ignota – Piccoli Campioni (n<30)	$s_p^2=\frac{(n_1-1)s_{c1}^2+(n_2-1)s_{c2}^2}{n_1+n_2-2}$

Intervalli di confidenza per la differenza tra proporzioni:

PARAMETRO (θ)		STIMA (θ_0 ¿	MARGINE D'ERRORE= VAR. CAS. X ERR. STD.		NOTE
			VARIABILE CASUALE	ERRORE STANDARD	
Differenza tra Proporzioni	$\pi_1-\pi_2$	p_1-p_2	$z_{1-\frac{\alpha}{2}}$	$\sqrt{\frac{p_1(1-p_1)}{n_1}\cdot\frac{p_2(1-p_2)}{n_2}}$	Grandi campioni (n>30)

Ampiezza campionaria per la media: $n=\frac{\left(z_{\frac{\alpha}{2}}\right)^2\sigma^2}{(ME)^2}$

Ampiezza campionaria per la proporzione: $n=\frac{\left(z_{\frac{\alpha}{2}}\right)^2 0{,}25}{(ME)^2}$

Capitolo 10

Ipotesi nulla H_o: ipotesi che viene considerata vera a meno di ottenere prove evidenti della validità del suo contrario.

Ipotesi alternativa H_1: ipotesi contro la quale viene verificata l'ipotesi nulla e che viene considerata vera se l'ipotesi nulla è considerata falsa.

Ipotesi semplice: ipotesi che specifica un singolo valore per il parametro della popolazione considerato.

Ipotesi composta: ipotesi che specifica uno o più intervalli di valori per il parametro della popolazione considerato.

Ipotesi alternativa unilaterale: ipotesi alternativa che considera tutti i possibili valori del parametro della popolazione a destra oppure a sinistra rispetto a quelli specificati dall'ipotesi nulla.

Ipotesi alternativa bilaterale: ipotesi alternativa che considera tutti i possibili valori del parametro della popolazione diversi dal valore specificato dall'ipotesi nulla semplice.

Regola di decisione: in base alla sua formulazione si rifiuta o non si rifiuta l'ipotesi nulla sulla base dell'evidenza campionaria.

Errore di primo tipo: errore commesso quando si rifiuta un'ipotesi nulla vera.

Errore di secondo tipo: errore commesso quando non si rifiuta un'ipotesi nulla falsa.

Livello di significatività: probabilità di rifiutare un'ipotesi nulla quando è vera. A volte viene espresso in termini percentuali $(100 \cdot \alpha\,\%)$

Potenza: probabilità di rifiutare un'ipotesi nulla quando è falsa.

Verifica d'ipotesi:

PARAMETRO (Θ)		Ipotesi	Bilaterale	Unilaterale SX	Unilaterale DX	STIMA (θ_0)	VARIABILE CASUALE	ERRORE STANDARD	NOTE
Media	μ	H_0 (Nulla) H_1 (Alternativa) Reg. Acc. H_0 Reg. Rif. H_0	$\mu = \mu_0$ $\mu \neq \mu_0$ $-z_{1-\alpha/2} < Test < +z_{1-\alpha/2}$ $Test < -z_{1-\alpha/2}$; $Test > +z_{1-\alpha/2}$	$\mu \geq \mu_0$ $\mu < \mu_0$ $Test > -z_{1-\alpha}$ $Test < -z_{1-\alpha}$	$\mu \leq \mu_0$ $\mu > \mu_0$ $Test < +z_{1-\alpha}$ $Test > +z_{1-\alpha}$	x_m	$z_{1-\alpha/2}$ $z_{1-\alpha}$	$\sqrt{(\sigma^2/n)}$ $\sqrt{(s_c^2/n)}$	con σ^2 Nota con σ^2 Ignota ; Grandi Campioni (n>30 ; n>120) dove $s_c^2 = s^2 * [(n/(n-1)]$
		H_0 (Nulla) H_1 (Alternativa) Reg. Acc. H_0 Reg. Rif. H_0	$\mu = \mu_0$ $\mu \neq \mu_0$ $-t_{n-1,\alpha/2} < Test < +t_{n-1,\alpha/2}$ $Test < -t_{n-1,\alpha/2}$; $Test > +t_{n-1,\alpha/2}$	$\mu \geq \mu_0$ $\mu < \mu_0$ $Test > -t_{n-1,\alpha}$ $Test < -t_{n-1,\alpha}$	$\mu \leq \mu_0$ $\mu > \mu_0$ $Test < +t_{n-1,\alpha}$ $Test > +t_{n-1,\alpha}$	x_m	$t_{n-1,\alpha/2}$ $t_{n-1,\alpha}$	$\sqrt{(s_c^2/n)}$	con σ^2 Ignota ; Piccoli Campioni (n≤30 ; n≤120) dove $s_c^2 = s^2 * [(n/(n-1)]$
Proporzione	π	H_0 (Nulla) H_1 (Alternativa) Reg. Acc. H_0 Reg. Rif. H_0	$\pi = \pi_0$ $\pi \neq \pi_0$ $-z_{1-\alpha/2} < Test < +z_{1-\alpha/2}$ $Test < -z_{1-\alpha/2}$; $Test > +z_{1-\alpha/2}$	$\pi \geq \pi_0$ $\pi < \pi_0$ $Test > -z_{1-\alpha}$ $Test < -z_{1-\alpha}$	$\pi \leq \pi_0$ $\pi > \pi_0$ $Test < +z_{1-\alpha}$ $Test > +z_{1-\alpha}$	p	$z_{1-\alpha/2}$ $z_{1-\alpha}$	$\sqrt{[p^*(1-p)/n]}$	Grandi Campioni (n>30 ; n>120)
Differenza tra Medie	$\mu_1 - \mu_2$	H_0 (Nulla) H_1 (Alternativa) Reg. Acc. H_0 Reg. Rif. H_0	$\mu_1 - \mu_2 = 0$ $\mu_1 - \mu_2 \neq 0$ $-z_{1-\alpha/2} < Test < +z_{1-\alpha/2}$ $Test < -z_{1-\alpha/2}$; $Test > +z_{1-\alpha/2}$	$\mu_1 - \mu_2 \geq 0$ $\mu_1 - \mu_2 < 0$ $Test > -z_{1-\alpha}$ $Test < -z_{1-\alpha}$	$\mu_1 - \mu_2 \leq 0$ $\mu_1 - \mu_2 > 0$ $Test < +z_{1-\alpha}$ $Test > +z_{1-\alpha}$	$x_{m1} - x_{m2}$	$z_{1-\alpha/2}$ $z_{1-\alpha}$	$\sqrt{[(\sigma_1^2/n_1) + (\sigma_2^2/n_2)]}$ $\sqrt{[(s_{c1}^2/n_1) + (s_{c2}^2/n_2)]}$	con σ^2 Nota con σ^2 Ignota ; Grandi Campioni (n>30 ; n>120) dove $s_c^2 = s^2 * [(n/(n-1)]$
		H_0 (Nulla) H_1 (Alternativa) Reg. Acc. H_0 Reg. Rif. H_0	$\mu_1 - \mu_2 = 0$ $\mu_1 - \mu_2 \neq 0$ $-t_{n1+n2-2,\alpha/2} < Test < +t_{n1+n2-2,\alpha/2}$ $Test < -t_{n1+n2-2,\alpha/2}$; $Test > +t_{n1+n2-2,\alpha/2}$	$\mu_1 - \mu_2 \geq 0$ $\mu_1 - \mu_2 < 0$ $Test > -t_{n1+n2-2,\alpha}$ $Test < -t_{n1+n2-2,\alpha}$	$\mu_1 - \mu_2 \leq 0$ $\mu_1 - \mu_2 > 0$ $Test < +t_{n1+n2-2,\alpha}$ $Test > +t_{n1+n2-2,\alpha}$	$x_{m1} - x_{m2}$	$t_{n1+n2-2,\alpha/2}$ $t_{n1+n2-2,\alpha}$	$\sqrt{[s_p^2(1/n_1 + 1/n_2)]}$	con σ^2 Ignota ; Piccoli Campioni (n≤30 ; n≤120) dove $s_p^2 = [(n_1-1)s_{c1}^2 + (n_2-1)s_{c2}^2] / (n_1+n_2-2)$

P-Value: probabilità che rappresenta il livello di significatività al quale l'ipotesi nulla può essere rifiutata.

Potenza di un test:

1. È tanto maggiore quanto più la vera media è distante dalla media ipotizzata nell'ipotesi nulla μ_o
2. Più basso è il livello di significatività di α, più è bassa la potenza.
3. Più è grande la varianza della popolazione, più è bassa la potenza del test.
4. Più è grande l'ampiezza campionaria, più è potente il test

Capitolo 13

Test Chi-Quadrato: estratto un campione casuale di *n* osservazioni, chiamate $O_1, O_2, \ldots, O_K$, se l'ipotesi nulla identifica con $p_1, p_2, \ldots, p_K$ le probabilità di appartenere a ogni categoria, la frequenza attesa di ogni categoria sotto H_0 è:

$$E_i = n\, p_i \quad i = 1, 2, \ldots, K$$

Se l'ipotesi nulla è vera e il campione abbastanza grande (frequenze attese ≥ 5), si costruisce la statistica test:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

che risulta seguire una distribuzione chi-quadrato con *(K – 1)* gradi di libertà.

Test sulla bontà di adattamento: si rifiuta H_0 se

$$\sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} > \chi^2_{K-1,\alpha}$$

Test Chi-Quadrato per le tabelle di contingenza: la statistica test:

$$\chi^2 = \sum_{i=1}^{r} \sum_{i=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

segue una distribuzione chi-quadrato con *(r – 1) (c – 1)* gradi di libertà, se non più del 20% delle frequenze attese E_{ij} è minore di 5.

Test di indipendenza per tabelle di contingenza: se l'ipotesi nulla è:

H_0: non ci sono associazioni tra le due caratteristiche della popolazione

allora la stima del numero atteso di osservazioni in ciascuna cella, sotto H_0, è:

$$E_{ij} = \frac{R_i C_j}{n} \qquad R_i C_j \text{ totali di riga e colonna}$$

Un test di indipendenza per tabelle di contingenza è che si rifiuta H_0 se:

$$\sum_{i=1}^{r} \sum_{i=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} > \chi^2_{(r-1)(c-1),\alpha}$$

Capitolo 12

Test sull'assenza di correlazione: sia *r* il coefficiente di correlazione lineare campionario, l'ipotesi nulla è:

$H_0: \rho = 0$

1) $H_1: \rho > 0$ si rifiuta H_0 se: $\frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} > t_{n-2,\alpha}$

2) $H_1: \rho < 0$ si rifiuta H_0 se: $\frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} \leftarrow t_{n-2,\alpha}$

3) $H_1: \rho \neq 0$ si rifiuta H_0 se: $\frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} \leftarrow t_{n-2,\alpha}$ o $\frac{r\sqrt{(n-2)}}{\sqrt{(1-r^2)}} > t_{n-2,\alpha}$

4) $t_{n-2,\frac{\alpha}{2}} = 2$ si rifiuta H_0 se: $|r| > \frac{2}{\sqrt{n}}$

Modello di regressione lineare semplice: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

β_0 = origine = $\acute{y} - b_1 \cdot \acute{x}$

β_1 = pendenza = $\frac{Cov(X,Y)}{\sigma_x^2}$

ε_i = errore aleatorio = $y_i - \hat{y}_i$

Metodo dei minimi quadrati per la stima dei coefficienti: permette di determinare gli stimatori dei coefficienti β_0 e β_1 minimizzando la somma dei quadrati degli errori ε_i:

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Stimatore di b_1: $r\frac{S_Y}{S_X}$

Stimatore di b_0: $\acute{y} - b_1 \cdot \acute{x}$

Ipotesi standard per il modello di regressione lineare:

1. Le *Y* sono funzioni lineari delle *X* e di una componente aleatoria di errore: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
2. Le x_i sono costanti o realizzazioni di una variabile aleatoria *X*, non correlata con le componenti aleatorie di errore ε_i.
3. Gli errori aleatori ε_i sono variabili aleatorie con media 0 e varianza σ^2 costante per ogni *i*.
   $E(\varepsilon_i) = 0 \qquad E(\varepsilon_i^2) = \sigma^2$
4. Gli errori aleatori ε_i non sono correlati tra di loro e quindi: $E(\varepsilon_i \varepsilon_j) = 0 \; per \; ogni \; i \neq j$

Analisi della varianza: la devianza totale di un modello di regressione (*SST*) può essere scomposta in una componente spiegata dal modello (*SSR*) e in una componente non spiegata o residua (*SSE*):

$$SST = SSR + SSE$$

$$SST = \sum_{i=1}^{n} (y_i - \acute{y})^2$$

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \acute{y})^2 = b_1^2 \sum_{i=1}^{n} (x_i - \acute{x})^2$$

$$SSE = \sum_{i=1}^{n} ¿¿¿$$

Coefficiente di determinazione R^2: $R^2=\frac{SSR}{SST}=1-\frac{SSE}{SST}$ $0<R^2<1$ (valore elevato→miglior modello)

Coefficiente di correlazione: per il modello di regressione lineare semplice, il coefficiente di determinazione R^2 coincide con il quadrato del coefficiente di correlazione:

$$R^2=r^2$$

Stima della varianza del modello: uno stimatore è la varianza dell'errore o varianza dei residui:

$$\hat{\sigma}^2=s_e^2=\frac{\sum_{i=1}^{n} e_i^2}{n-2}=\frac{SSE}{n-2}$$

Varianza dello stimatore dei minimi quadrati per β_1:

Varianza nota	Varianza ignota
$\sigma_{b_1}^2=\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\acute{x})^2}=\frac{\sigma^2}{(n-1)s_x^2}$	$s_{b_1}^2=\frac{s_e^2}{\sum_{i=1}^{n}(x_i-\acute{x})^2}=\frac{s_e^2}{(n-1)s_x^2}$

Basi per l'inferenza su β_1: se valgono le ipotesi standard per il modello di regressione e gli errori sono distribuiti normalmente, la variabile aleatoria:

$$T=\frac{b_1-\beta_1}{s_{b1}}$$

è distribuita secondo una variabile *t* di Student con *(n – 2)* gradi di libertà.

Test di ipotesi su β_1: se gli errori del modello sono distribuiti normalmente e valgono le ipotesi standard:

1) $H_0:\beta_1=\beta_1^{¿}$ oppure $H_0:\beta_1\leq\beta_1^{¿}$

$H_1:\beta_1>\beta_1^{¿}$

si rifiuta H_0 se $\frac{b_1-\beta_1^{¿}}{s_{b1}}>t_{n-2,\alpha}$

2) $H_0:\beta_1=\beta_1^{¿}$ oppure $H_0:\beta_1\geq\beta_1^{¿}$

$H_1:\beta_1<\beta_1^{¿}$

si rifiuta H_0 se $\frac{b_1-\beta_1^{¿}}{s_{b1}}<t_{n-2,\alpha}$

3) $H_0:\beta_1=\beta_1^{¿}$

$H_1:\beta_1\neq\beta_1^{¿}$

si rifiuta H_0 se $\frac{b_1-\beta_1^{¿}}{s_{b1}}>t_{n-2,\alpha}$ oppure $\frac{b_1-\beta_1^{¿}}{s_{b1}}<t_{n-2,\alpha}$

Intervallo di confidenza per β_1: se gli errori del modello sono distribuiti normalmente e valgono le ipotesi standard, un intervallo di confidenza a livello $100(1-\alpha)\%$ è dato da:

$$b_1-t_{n-2,\alpha}s_{b1}<\beta_1<b_1+t_{n-2,\alpha}s_{b1}$$

Test F su β_1:

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

si rifiuta H_0 se $F \geq F_{1, n-2, \alpha}$