DEFINIZIONE TIPOLOGIA DI STUDIO

indagine GLOBALE o PARZIALE o CAMPIONARIA

studi TRASVERSALI o LONGITUDINALI } 1 UNITÀ STATISTICA

" OCCASIONALI o RIPETUTI } + UNITÀ STATISTICHE

" OSSERVAZIONALI o SPERIMENTALI } MODALITÀ DI INTERVENTO

" DESCRITTIVI o ANALITICI } FINALITÀ DI RICERCA

DEFINIZIONE DELLE VARIABILI

QUALITATIVE CATEGORIALI
- ORDINABILI
- NOMINALI

QUANTITATIVE NUMERICHE
- CONTINUE QUALSIASI
- DISCRETE NATURALI

TIPOLOGIE GRAFICI

DIAGRAMMA A BARRE
DIAGRAMMA A TORTA
DIAGRAMMA DI PARETO
} VARIABILI QUALITATIVE O QUANTITATIVE DISCRETE O PER VARIABILI MISURATE CON SCALA NOMINALE O ORDINALE

ISTOGRAMMA
POLIGONO
POLIGONO CUMULATIVO
} VARIABILI QUANTITATIVE CONTINUE

ANALISI DESCRITTIVA

tecniche:	studio:
• FREQUENZE →	FORMA della DISTRIBUZIONI
• VALORI MEDI →	TENDENZA CENTRALE ...
• MISURE DI VARIABILITÀ →	VARIABILITÀ ...

VARIABILE QUALITATIVA NOMINALE → MODA

esempio tipo di dissesto
SCALA NOMINALE { I, FC, IC ... }
TIPO DI GRAFICO: DIAGRAMMA A BARRE
forma DISTRIBUZIONE:
- TENDENZIALMENTE UNIFORME
- TENDENZIALMENTE NON UNIFORME
  - UNIMODALE
  - PLURIMODALE

VARIABILE QUALITATIVA ORDINATA → MODA

esempio grado di urbanizzazione in 50 comuni
SCALA ORDINALE { BASSO, MEDIO/BASSO, MEDIO ... }
TIPO DI GRAFICO: DIAGRAMMA A BARRE
forma DISTRIBUZIONE:
- TENDENZIALMENTE UNIFORME → QUANTILI
- TENDENZIALMENTE NON UNIFORME
  - UNIMODALE TENDENZIALMENTE { NEGATIVA → MEDIANA; SIMMETRICA → MEDIANA; POSITIVA → MEDIANA }
  - PLURIMODALE → QUANTILI { FORMA A J; FORMA A U }

VARIABILE QUANTITATIVA DISCRETA → MODA

esempio n° difetti in 50 barre di acciaio
SCALA NUMERICA: { 0, 1, 2, 3, ... }
TIPO DI GRAFICO: { DIAGRAMMA A BARRE; DIAGRAMMA A BASTONCINI; DOT PLOT numerosità ≤ 30 }
forma DISTRIBUZIONE:
- TENDENZIALMENTE UNIFORME → QUANTILI
- TENDENZIALMENTE NON UNIFORME
  - UNIMODALE TENDENZIALMENTE { NEGATIVA → MEDIANA; SIMMETRICA → MEDIA; POSITIVA → MEDIANA }
  - PLURIMODALE → QUANTILI { FORMA A J; FORMA A U }

VARIABILE QUANTITATIVA CONTINUA → MODA

esempio resistenza a compressione in MPa
SCALA NUMERICA CONTINUA
TIPO DI GRAFICO: { ISTOGRAMMA; POLIGONO }
forma DISTRIBUZIONE:
- TENDENZIALMENTE UNIFORME → QUANTILI
- TENDENZIALMENTE NON UNIFORME
  - UNIMODALE TENDENZIALMENTE { NEGATIVA → MEDIANA; SIMMETRICA → MEDIA; POSITIVA → MEDIANA }
  - PLURIMODALE → QUANTILI { FORMA A J; FORMA A U }

VALORI MEDI

INTENSITÀ MEDIA: MEDIA
POSIZIONE CENTRALE: MEDIANA $\frac{n+1}{2}$ (SE PARI: posizioni $\frac{n}{2}$ e $\frac{n+1}{2}$, E MEDIA DEI DUE VALORI!!!)
POSIZIONI CENTRALI E NON: QUANTILI { $1° \to (n+1)/4$; $2° \to (n+1)/2$; $3° \to 3(n+1)/4$ }
MODALITÀ PIÙ FREQUENTI: MODA

MISURE DI VARIABILITÀ

- RANGE
- RANGE INTERQUARTILE
- VARIANZA
- DEVIAZIONE STANDARD
- COEFFICIENTE DI VARIAZIONE
- FREQUENZA RELATIVA DELLA MODA

per variabili QUANTITATIVE:
- CONFRONTO TRA MEDIA E MEDIANA
- BOXPLOT

FORMA	MISURA VARIABILITÀ
– UNIFORMI:	RANGE, IQR
– SIMMETRICHE:	D.S., RANGE, IQR
– ASIMMETRICHE:	RANGE, IQR
– PLURIMODALI:	RANGE, IQR

VARIANZA: $\frac{\sum_{j=1}^{n}(x_j-\bar{x})^2}{n}$

$\sigma^2 = \frac{1}{N}\sum_{j=1}^{N}(x_j-\bar{x})^2$ → $S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j-\bar{x})^2$

DEVIAZIONE STANDARD: $\sigma = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(x_j-\bar{x})^2}$ → $S = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x_j-\bar{x})^2}$

↳ distanza media dei valori dalla loro media

$S = \sqrt{\frac{\sum_{i=1}^{k}(x_i-\bar{x})^2 \times n_i}{n-1}}$ NON suddivisa in classi

$S = \sqrt{\frac{\sum_{i=1}^{k}(x_i-\bar{x}) \times n_i}{n-1}}$

COEFFICIENTE di VARIAZIONE: indipendente dall'ordine di grandezza della variabile

esempio trovare quale di due prodotti con valute di differente presenta un prezzo maggiormente variabile:

A) media = 50 euro, dev. stand. = 5 euro → $CV_A\% = \left(\frac{5}{50}\right)\times 100 = 10\%$

B) media = 100 $, dev. stand = 5 $ → $CV_B\% = \left(\frac{5}{100}\right)\times 100 = 5\%$

PREZZO B MENO VARIABILE dell'A

$$CV\% = \left(\frac{S}{\bar{x}}\right)\times 100$$

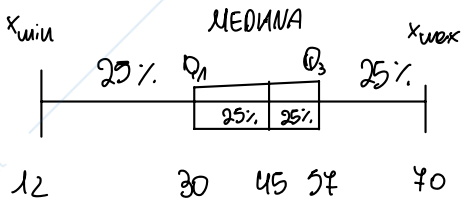
SCALA NOMINALE → freq. moda

ORDINALE → freq. moda, range, rang

NUMERICA → freq. moda, range, range inter deviazione standard

BOXPLOT
- x_{min} (minimo)
- PRIMO QUARTILE (Q_1)
- MEDIANA
- TERZO QUARTILE (Q_3)
- x_{max} (massimo)

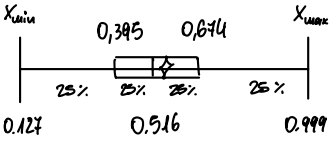
esempio
- $x_{min} = 12$
- PRIMO = 30
- MED = 45
- TERZO = 57
- $x_{max} = 70$



ESERCITAZIONE es 3:

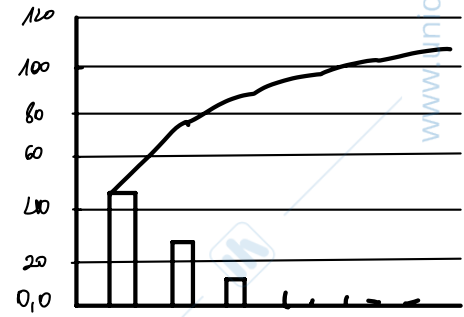
CLASSE IV_0-10

$x_{min} = 0.127$
PRIMO Q. = 0.395
MEDIANA = 0.516
TERZO Q. = 0.674
$x_{max} = 0.999$
MEDIA = 0.543 ✧



es 5:

60:69	freq. ASS	freq rel %	freq % cum.
CIVILE	87	$\frac{87}{203}\times 100 = 42.9$	42.9
ISTRUZIONE	58	$\frac{28}{203}\times 100 = 28.6$	71.5
SANITA'	21	$\frac{21}{203}\times 100 = 10.3$	81.9
RELIGIOSO	11	5.4	87.2
MILITARE	10	4.9	92.2
NON IDENT.	10	4.9	97.1
MOB e TRASP.	5	2.5	99.6
TECNOLOGICO	1	2.5	100.0
TOTALE	203	100	



FORMA	VALORE MEDIO	MISURA VARIABILITA
UNIFORME	QUANTILI	RANGE, RANGE IQR
SIMMETRICHE	MEDIA	D.S, RANGE, IQR
ASIMMETRICHE	MODA, MEDIANA, QUANTILI	RANGE, IQR
PLURIMODALI	QUANTILI	RANGE, IQR

tra 0 e 1 compresi
0 → evento impossibile
1 → evento certo

LA VARIABILITÀ, INCERTEZZA E PROBABILITÀ

PROVA, EVENTO, PROBABILITÀ

3 APPROCCI: IMPOSTAZIONE CLASSICA, FREQUENTISTA, SOGGETTIVISTA

esperimenti con:	CON:	CON:
1) incertezza risultato	1)	1)
2) ripetibili nelle stesse condizioni	2)	eventi unici senza esperienza consolidata
3) con risultati equiprobabili		

E = EVENTO

$f_n(E)$ = numero di volte che E si è verificato

Allora $Pr(E)$ = PROBABILITÀ

VARIABILI CASUALI

MISURE	PARAMETRI DI POPOLAZIONE	STATISTICHE CAMPIONARIE
MEDIA	μ	$\bar{x}$
VARIANZA	σ^2	S^2
DEVIAZIONE STANDARD	σ	S

MODELLI PER VARIABILI CASUALI CONTINUE

se la variabile può assumere tutti i valori compresi in un intervallo reale → $f(x) = Pr(x \in A)$

DISTRIBUZIONE NORMALE O GAUSSIANA

- diversi fenomeni continui sembrano seguire una distribuzione normale;
- può approssimare numerose distribuzioni discrete;
- teorema del limite centrale.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(1/2)[(x-\mu)/\sigma]^2}$$

e = 2.718
π = 3.14
μ = media della pop.
σ = dev. stand.
x = valori assunti dalla variabile X

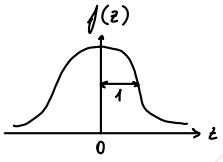
diventa

DISTRIBUZIONE NORMALE STANDARDIZZATA

consente di ottenere una variabile casuale normale standardizzata con valor atteso pari a 0 e deviazione standard pari a 1

La standardizzazione è data dalla trasformazione di X in Z:

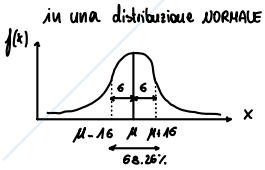
$$z = \frac{X - \mu}{\sigma}$$



+ TAVOLE DISTRIBUZIONE NORMALE STANDARDIZZATA

- riga: parte intera di z
- colonna: seconda cifra decimale di z
- incrocio: PROBABILITÀ di trovare un valore ≤ a z

in una distribuzione NORMALE:



$\mu \pm 1\sigma$ include circa il 68.26% dei valori
$\mu \pm 2\sigma$ include circa il 95.4% dei valori
$\mu \pm 3\sigma$ include circa il 99.7% dei valori

ne discendono REGOLE EMPIRICHE: (n≥30)

$\bar{x} \pm DS$ ... 68% ...
$\bar{x} \pm 2DS$ ... 95.4% ...
$\bar{x} \pm 3DS$ ... 99.7% ...

MODELLI PER VARIABILI CASUALI DISCRETE

se l'insieme dei possibili valori (o supporto) è un insieme discreto di numeri reali → $p(x) = Pr(X = x)$

- di Bernoulli
- Binomiale
- Geometrica
- di Poisson

MODELLO di BERNOULLI

Una prova può dar luogo a due risultati:
successo di E (1) e INSUCCESSO di E (0) (x = 0,1)

$Pr(x=1) = p$
$Pr(x=0) = (1-p)$ } PROBABILITÀ

$X \sim Ber(p)$

RISULTATI DICOTOMICI (binario)

MODELLO BINOMIALE

esperimento casuale ripetuto n volte nelle medesime condizioni, in cui in ogni singola prova può dar luogo a due possibili risultati su un dato evento: evento verificato (successo) ed evento non verificato (insuccesso)

(x = 0,1, ..., n)

$Pr(E) = p$ dove p costante

$X \sim Biu(n, p)$

MODELLO di POISSON

in un certo ambito circoscritto, un evento (E) successo può verificarsi più volte e in modo indipendente

x = 0,1,2, ... → supporto discreto infinito numerabile

$X \sim Po(\lambda)$

$\lambda = E(x)$ = n° medio di successi nell'ambito considerato (media)

MODELLO GEOMETRICO

esperimento ipotetico: successione di sottoprove bernoulliane indipendenti.
ogni sottoprova: SUCCESSO (E) e INSUCCESSO (E)

x = 1, 2, ... → supporto discreto infinito numerabile

$X \sim Geo(p)$

$Pr(E) = p$
$E(x) = 1/p$ = n° medio di sottoprove fino al successo

MODELLI PER VARIABILI CASUALI CONTINUE

- Esponenziale
- di Weibull
- GEV

Variabili casuali connesse alla Normale:
- Lognormale
- Chi-quadrato
- di Fisher
- di Student

DISTRIBUZIONE ESPONENZIALE

introdotta per fenomeni di cui interessa la DURATA di vita, di resistenza...

$\lambda (x \sim E(\lambda))$

valore medio $1/\lambda$ è la DURATA media del fenomeno riferimento agli esperimenti di Poisson

$S(x) = 1 - F(x) = Pr(X > x)$

DISTRIBUZIONE di WEIBULL

Esponenziale generalizzata con $\lambda > 0$ e $\beta > 0$
introdotta per fenomeni di cui interessa la DURATA di vita, di resistenza...

$\beta = 1$, funzione di rischio costante (Esp.)
$\beta > 1$, " " " è funzione crescente del tempo
$\beta < 1$, " " " è funzione decrescente del tempo

DISTRIBUZIONE GENERALIZZATA DEI VALORI ESTREMI

usata come approssimazione per modellare i massimi o i minimi di lunghe (finite) sequenze di variabili aleatorie

DISTRIBUZIONE LOGNORMALE

$Y \sim N(\mu, \sigma)$ dove N NORMALE

$Y = \ln X$ v.c. Lognormale X

Viene usata al posto della NORMALE quando il fenomeno di interesse non può assumere valori negativi, ma è UNIMODALE e con forma approssimativamente campanulare e simmetrica

INTRODUZIONE ALLE TECNICA DI INFERENZA STATISTICA

GENERALIZZARE LE INFORMAZIONI TRATTE DAL CAMPIONE ALLA TOTALITÀ DELLA POPOLAZIONE

- SELEZIONARE CAMPIONE
- VALUTARE GRADO DI ATTENDIBILITÀ
- INFERENZA STATISTICA
  - ESTENDERE RISULTATI ALLA POPOLAZIONE
  - CONTROLLARE GRADO DI INCERTEZZA

MISURE INTRODOTTE AI FINI DELL'ANALISI DESCRITTIVA DIVENTANO OGGETTO DI INFERENZA

STIMA

PUNTUALE (340 gr): QUANTITÀ INCOGNITA AD UN UNICO VALORE. È UNA SINGOLA STATISTICA. MEDIA CAMPIONARIA È STIMATORE PUNTUALE DELLA MEDIA DELLA POPOLAZIONE μ

PER INTERVALLO (365 gr, 375 gr): QUANTITÀ INCOGNITA AD UN INSIEME DI VALORI. COSTRUENDO UN INTERVALLO ENTRO CUI CI POTREBBE ESSERE IL VALORE VERO DELLA QUANTITÀ INCOGNITA

LUNGHEZZA = $2z_{\alpha/2}(\sigma/\sqrt{n})$
$z_{\alpha/2}$: VALORE CRITICO
α: probabilità che il parametro si trovi fuori dall'intervallo
(m) intervallo di confidenza

INTERVALLI DI CONFIDENZA PER LA MEDIA:

1) VARIANZA NOTA → LIVELLO DI CONFIDENZA (90%, 95%, ...)
$Pr\left(\bar{X} - z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} \le \bar{X} + z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}\right) = ...$

2) VARIANZA NON NOTA
$\left(\bar{X} - t_{n-1;\alpha/2} \times \frac{S}{\sqrt{n}}, \bar{X} + t_{n-1;\alpha/2} \times \frac{S}{\sqrt{n}}\right)$

3) POPOLAZIONE NON NORMALE
$\left(\bar{X} - z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} \le \bar{X} + z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}\right)$ $\left(\bar{X} - t_{n-1;\alpha/2} \times \frac{S}{\sqrt{n}}, \bar{X} + t_{n-1;\alpha/2} \times \frac{S}{\sqrt{n}}\right)$

TEOREMA DEL LIMITE CENTRALE

PER LA MAGGIOR PARTE DELLE POPOLAZIONI, LA DISTRIBUZIONE DELLA MEDIA CAMPIONARIA È APPROSSIMATIVAMENTE NORMALE, PURCHÈ SI CONSIDERINO CAMPIONI DI ALMENO 30 OSSERVAZIONI

SE DISTRIBUZIONE POPOLAZIONE È ABBASTANZA SIMMETRICA, LA DISTRIBUZIONE DELLA MEDIA → DISTRIBUZIONE NORMALE PURCHÈ CI SIANO ALMENO 5-15 OSSERVAZIONI

SE POPOLAZIONE HA DISTRIBUZIONE NORMALE LA MEDIA CAMPIONARIA È DISTRIBUITA SECONDO LA LEGGE NORMALE, INDIPENDENTEMENTE DA AMPIEZZA

$$\bar{X} \sim N(\mu, \sigma/\sqrt{n})$$

$$z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

$$Pr\left(\bar{X} - z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

VARIANZA σ NON NOTA: $t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$

$$\bar{X} - t_{n-1/\alpha/2} \times \frac{S}{\sqrt{n}} \le \mu \le \bar{X} - t_{n-1/\alpha/2} \times \frac{S}{\sqrt{n}}$$

INTERVALLO DI CONFIDENZA PER LA PROPORZIONE:

$$p - z_{\alpha/2} \times \sqrt{p(1-p)/n} \le \pi \le p + z_{\alpha/2} \times \sqrt{p(1-p)/n}$$

VERIFICA DI IPOTESI: TEST SU UN CAMPIONE

PROCEDURA INFERENZIALE, SI FA UN' IPOTESI SU QUANTITÀ INCOGNITA DELLA POPOLAZIONE (PARAMETRO) E DECIDERE SECONDO CAMPIONE CASUALE (STATISTICA CAMPIONARIA) SE ESSA È ACCETTABILE O NO.

TEST DI IPOTESI → PROCESSO CONVENTIVO O MENO L'IPOTESI DI RICERCA

PRESENZA DI
- DIFFERENZA PARAMETRO DA VALORE INTERESSE
- DIFFERENZE TRA PARAMETRI DI GRUPPI DIVERSI
- RELAZIONI TRA VARIABILI

VERIFICA DI IPOTESI PARAMETRICA COMPOSTA DA

- IPOTESI NULLA H_0: ASSENZA DIFFERENZE DEL PARAMETRO DA VALORE / ASSENZA DIFFERENZE TRA PARAMETRI O ASSENZA DI RELAZIONI TRA VARIABILI → ATTEGGIAMENTO CONSERVATIVO
- IPOTESI ALTERNATIVA H_1: NUOVA INTUIZIONE DEL RICERCATORE

1) FASI TEST STATISTICO: TEST Z PER LA MEDIA (σ NOTO)

1° PASSO: FORMULAZIONE IPOTESI

$H_0: \mu = 368$ gr
$H_1: \mu \neq 368$ gr
TEST DI IPOTESI BIDIREZIONALE O A DUE CODE

$H_0: \mu = 368$ gr
$H_1: \mu < 368$ gr
TEST DI IPOTESI UNIDIREZIONALE O A UNA CODA

2° PASSO: SCELTA DI UNA STATISTICA TEST

STATISTICA CAMPIONARIA LA CUI DISTRIBUZIONE CAMPIONARIA DEVE ESSERE NOTA E COMPLETAMENTE SPECIFICATA SOTTO L'IPOTESI NULLA (μ e σ).
SE X È NORMALE POSSIAMO APPLICARE IL TEOREMA DEL LIMITE CENTRALE:

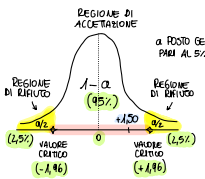
$$\bar{X} \sim N(\mu, \sigma/\sqrt{n}) \qquad z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

3° PASSO: REGIONE DI ACCETTAZIONE E DI RIFIUTO

IPOTIZZIAMO VERA L'IPOTESI NULLA
AD ESEMPIO:

$\bar{X} \sim N(368, 15/\sqrt{25})$ e $z = \frac{\bar{X} - 368}{15/\sqrt{25}}$

SE LA MEDIA CAMPIONARIA DEL CAMPIONE OSSERVATO CADE IN UNA DELLE DUE ZONE DI RIFIUTO POSSIAMO: (TRA -1,96 e +1,96)
- SI È VERIFICATO UN VALORE MOLTO RADO SOTTO L'IPOTESI NULLA, MA NON RIFIUTIAMO L'IPOTESI NULLA
- ESSENDOSI VERIFICATO UN VALORE CHE È MOLTO RADO SOTTO L'IPOTESI NULLA, RIFIUTIAMO L'IPOTESI NULLA



I VALORI DELLA STATISTICA (OVVERO LE MEDIE CAMPIONARIE) CHE CADONO NELLA REGIONE DI ACCETTAZIONE PORTANO ALL'ACCETTAZIONE DELL'IPOTESI NULLA

I VALORI DELLA STATISTICA (OVVERO LE MEDIE CAMPIONARIE) CHE CADONO NELLA REGIONE DI RIFIUTO PORTANO AL RIFIUTO DELL'IPOTESI NULLA

PERTANTO PER $\alpha = 5\%$
RIFIUTARE H_0 SE $z < -1,96$ O $z > +1,96$
OPPURE SE $\bar{X} < \mu - 1,96 \times \sigma/\sqrt{n}$ O $\bar{X} > \mu + 1,96 \times \sigma/\sqrt{n}$

NON RIFIUTARE H_0 ALTRIMENTI

NELL'ESEMPIO LA MEDIA CAMPIONARIA RISULTA 372,5:

$$z = \frac{372,5 - 368}{15/\sqrt{25}} = +1,50$$

CHE RISULTA ESSERE ALL'INTERNO DELLA REGIONE DI ACCETTAZIONE.
IL TEST PORTA A NON RIFIUTARE L'IPOTESI NULLA H_0
LA DIFFERENZA TRA LA MEDIA CAMPIONARIA OSSERVATA (372,5) E IL VALORE IPOTIZZATO DEL LA MEDIA DELLA POPOLAZIONE (368) NON È STATISTICAMENTE SIGNIFICATIVA.
IL PROCESSO È SOTTO CONTROLLO

UN ALTRO MODO PER EVIDENZIARE IL TEST È QUELLO DI CALCOLARE IL **P-VALUE**:
- SE IL P-VALUE È $\geq \alpha$, IPOTESI NULLA NON È RIFIUTATA
- SE IL P-VALUE È $< \alpha$, IPOTESI NULLA È RIFIUTATA

ESEMPIO: SE I VALORI CRITICI = -1,50 e +1,50
ALLORA LA PROBABILITÀ CHE Z ASSUMA UN VALORE MINORE DI +1,50 È DI 0.9332; PER CUI LA PROBABILITÀ CHE ASSUMA UN VALORE SUPERIORE A +1,50 È: 1 - 0.9332 = 0.0668. ALLO STESSO MODO LA PROBABILITÀ CHE ASSUMA UN VALORE MINORE A -1,50 È 0.0668. PERTANTO IL P-VALUE ASSUME UN VALORE DI: 0.0668 + 0.0668 = 0.1336
QUINDI 0,1336 > α (0.05) → NON RIFIUTIAMO H_0

SOTTO CONTROLLO DI CHI CONDUCE IL TEST
- ERRORE DEL 1 TIPO: SI RIFIUTA L'IPOTESI NULLA QUANDO È VERA
- ERRORE DEL 2 TIPO: SI ACCETTA L'IPOTESI NULLA QUANDO È FALSA

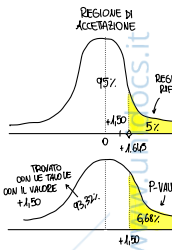
AUMENTARE DIMENSIONE DEL TEST
- $1 - \alpha$ CONFIDENZA DEL TEST
- $1 - \beta$ POTENZA DEL TEST

I TEST AD UNA CODA UNIDIREZIONALI

IPOTESI NULLA $H_0: \mu = 368$
IPOTESI ALTERNATIVA: $H_1: \mu > 368$

FISSANDO $\alpha = 0.05$ (5%)
LA REGIONE CRITICA È CONTENUTA NELLA CODA DESTRA DELLA DISTRIBUZIONE NORMALE STANDARDIZZATA E CORRISPONDE AL 5%
IL VALORE CRITICO LASCIA ALLA SUA SINISTRA UNA MASSA DI PROBABILITÀ PARI A 0.95 (95%)
QUINDI, DALLE TAVOLE IL VALORE CRITICO È 1,645 (GUARDO IL VALORE CHE PIÙ SI AVVICINA A 0.9500, IN QUESTO CASO È LA MEDIA TRA DUE VALORI E PRENDO I CORRISPETTIVI VALORI DELLA COLONNA E RIGA = 1,6; 0,04 E 0,05 → 0.045
z = +1,50, CADE ALL'INTERNO DELLA REGIONE DI ACCETTAZIONE E QUINDI L'IPOTESI NULLA È ACCETTATA

SECONDO IL P-VALUE, DOBBIAMO CALCOLARE LA PROBABILITÀ CHE Z SIA > A +1,50.
P-VALUE = 1 − 0.9332 = 0.0668 (6,68%)
0.0668 > 0.05 → L'IPOTESI NULLA NON PUÒ ESSERE RIFIUTATA



2) TEST DI IPOTESI T PER LA MEDIA (σ NON NOTO)

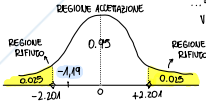
$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$ → SE NORMALE O NON T-STUDENT n-1 GRADI DI LIBERTÀ

ESEMPIO:
$H_0: \mu = 120$
$H_1: \mu \neq 120$
$\alpha = 0.05$ (5%)
$n = 12$
$\bar{X} = 112.85$ m/s
$S = 20.80$ m/s

$t = \frac{112.85 - 120}{20.80/\sqrt{12}} = -1,19$

t-student con 12 − 1 = 11 GRADI DI LIBERTÀ

TEST A DUE CODE QUINDI DETERMINO I VALORI CRITICI CON LE TAVOLE DELLA DISTRIBUZIONE T DI STUDENT
GRADO DI LIBERTÀ = 11
AREA CODA DI DESTRA = ...
... = $\alpha/2 = \frac{0.05}{2} = 0.025$
VALORE CRITICO = 2.2010



IL VALORE CADE ALL'INTERNO DELLA ZONA DI ACCETTAZIONE, QUINDI L'IPOTESI NULLA NON DEVE ESSERE RIFIUTATA

3) TEST DI IPOTESI Z PER LA PROPORZIONE

VERIFICARE IPOTESI SU π

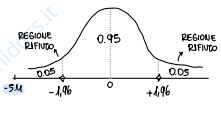
PROPORZIONE CAMPIONARIA X/n
SE NUMERO SUCCESSI X E DI INSUCCESSI (n-X) SONO ENTRAMBI MAGGIORI DI 5, LA DISTRIBUZIONE DELLA PROPORZIONE DI SUCCESSI PUÒ ESSERE APPROSSIMATA DALLA DISTRIBUZIONE NORMALE E QUINDI:

$Z = \frac{p - \pi}{\sqrt{\pi(1-\pi)/n}}$

$H_0: \pi = 0.5$
$H_1: \pi \neq 0.5$

PROPORZIONE CAMPIONARIA: $369/894 = 0.41$

$\alpha = 0.05$
REGIONE ACCETTAZIONE = 0.95
$\frac{0.41 - 0.5}{0.5(1-0.5)/894} = -5.4$ → CADE NELLA REGIONE DI RIFIUTO QUINDI RIFIUTIAMO L'IPOTESI NULLA
VALORE CRITICO CON TABELLE: 0.9750 → 1,96



4) LEGAME TRA INTERVALLI DI CONFIDENZA E VERIFICA DI IPOTESI

$n = 25$
$\bar{X} = 372,5$ grammi
$\sigma = 15$ grammi
LIVELLO DI CONFIDENZA = 95%
$\alpha = 0.05$ (5%)
VALORE CRITICO = 1,96

AVREMO QUINDI IL SEGUENTE INTERVALLO DI CONFIDENZA:
$\bar{X} \pm 1,96 \frac{15}{\sqrt{25}} = (366,6; 378,4)$

AVENDO IPOTIZZATO NELL'IPOTESI NULLA 368 ED ESSENDO CHE $366,6 \leq 368 \leq 378,4$ NON RIFIUTIAMO L'IPOTESI NULLA

VERIFICA DI IPOTESI: TEST SU DUE CAMPIONI

CONFRONTO TRA MEDIE DI DUE POPOLAZIONI INDIPENDENTI

(LA SELEZIONE DI UN CAMPIONE NON HA EFFETTO SULLA SECONDA)

CONSIDERIAMO **DUE** POPOLAZIONI E SUPPONIAMO DI ESTRARRE UN CAMPIONE DI AMPIEZZA n_1 DALLA 1ª POP. E UN CAMPIONE n_2 DALLA 2ª POP., ENTRAMBI OSSERVATI IN RELAZIONE AD UNA **VARIABILE CASUALE QUANTITATIVA** X

μ_1 e μ_2 LE MEDIE

σ_1 e σ_2 LE DEV. STANDARD

POPOLAZIONI CON DISTRIBUZIONE NORMALE ORVERO ($n_1 \geq 30$, $n_2 \geq 30$) NON NORMALE MA CON AMPIEZZE ABBASTANZA GRANDI PER POTER APPLICARE IL TEOREMA DEL LIMITE CENTRALE

VOGLIAMO VERIFICARE L'IPOTESI NULLA CHE LE MEDIE DELLE DUE POPOLAZIONI SIANO UGUALI TRA LORO:

$H_0: \mu_1 = \mu_2 \quad (\mu_1 - \mu_2 = 0)$

CONTRO UN IPOTESI ALTERNATIVA BIDIREZIONALE O UNIDIREZIONALE:

$H_1: \mu_1 \neq \mu_2 \quad (\mu_1 - \mu_2 \neq 0)$, $H_1: \mu_1 > \mu_2 \quad (\mu_1 - \mu_2 > 0)$

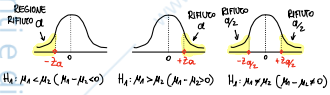
$H_1: \mu_1 < \mu_2 \quad (\mu_1 - \mu_2 < 0)$

LA STATISTICA TEST È UNA FUNZIONE DELLA DIFFERENZA TRA LE **MEDIE CAMPIONARIE** DEI DUE CAMPIONI: $\overline{X}_1 - \overline{X}_2$

PER IL CONFRONTO TRA LE MEDIE DI 2 POP. INDIPENDENTI POSSIAMO:

1) TEST Z SULLA DIFFERENZA DI DUE MEDIE, VARIANZE σ_1 E σ_2 NOTE;
2) TEST t SULLA DIFFERENZA DI DUE MEDIE, VARIANZE OMOGENEE ($\sigma_1 = \sigma_2 = \sigma$) NON NOTE;
3) TEST t SULLA DIFFERENZA DI DUE MEDIE, VARIANZE NON OMOGENEE E NON NOTE.

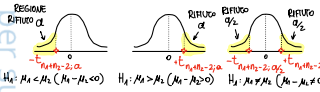
1) $Z = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$; NORMALE $N(0,1)$; α (5% o 1%)



$H_1: \mu_1 < \mu_2 \ (\mu_1 - \mu_2 < 0)$ $H_1: \mu_1 > \mu_2 \ (\mu_1 - \mu_2 > 0)$ $H_1: \mu_1 \neq \mu_2 \ (\mu_1 - \mu_2 \neq 0)$

2) $t = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ DOVE $S_p^2 = \dfrac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1-1)+(n_2-1)}$; t DI STUDENT (n_1+n_2-2) GRADI DI LIBERTÀ ; α (5% o 1%)

(VARIANZA COMBINATA)



$H_1: \mu_1 < \mu_2 \ (\mu_1 - \mu_2 < 0)$ $H_1: \mu_1 > \mu_2 \ (\mu_1 - \mu_2 > 0)$ $H_1: \mu_1 \neq \mu_2 \ (\mu_1 - \mu_2 \neq 0)$

INTERVALLO DI CONFIDENZA PER LA DIFFERENZA TRA LE MEDIE DI DUE POPOLAZIONI INDIPENDENTI $(\mu_1 - \mu_2)$:

$(\overline{X}_1 - \overline{X}_2) \pm t_{n_1+n_2-2;\,\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ (VALORE CRITICO)

AREA ACCETTAZIONE: $1-\alpha/2$

CONFRONTO TRA MEDIE DI DUE POPOLAZIONI NON INDIPENDENTI

(CAMPIONI APPAIATI / CAMPIONI RIPETUTE)

L'ANALISI NON CONSIDERA LE DUE MEDIE DEI SINGOLI CAMPIONI MA LA **MEDIA DELLE DIFFERENZE TRA I DUE CAMPIONI**

VALORI	1 CAMPIONE	2 CAMPIONE	DIFFERENZA
1	X_{11}	X_{21}	$D_1 = X_{11} - X_{21}$
...	...	...	...
j	X_{1j}	X_{2j}	$D_j = X_{1j} - X_{2j}$
...	...	...	...
n	X_{1n}	X_{2n}	$D_n = X_{1n} - X_{2n}$

SI VUOLE VERIFICARE L'IPOTESI NULLA CHE LA MEDIA DELLE DIFFERENZE TRA LE DUE POPOLAZIONI UGUALE A 0:

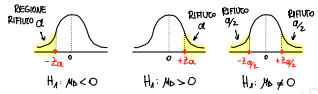
$H_0: \mu_D = 0$

CONTRO UN IPOTESI ALTERNATIVA BIDIREZIONALE O UNIDIREZIONALE:

$H_1: \mu_D \neq 0$ $H_1: \mu_D < 0$ $H_1: \mu_D > 0$

TEST Z SULLA MEDIA DELLE DIFFERENZE, VARIANZA DELLE DIFFERENZE σ_D NOTA

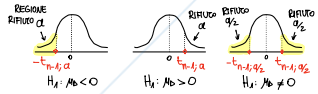
$Z = \dfrac{\overline{D} - \mu_D}{\frac{\sigma_D}{\sqrt{n}}}$ DOVE $\overline{D} = \dfrac{\sum_{j=1}^{n} D_j}{n}$; NORMALE $N(0,1)$; α (5% o 1%)



$H_1: \mu_D < 0$ $H_1: \mu_D > 0$ $H_1: \mu_D \neq 0$

TEST t SULLA MEDIA DELLE DIFFERENZE, VARIANZA DELLE DIFFERENZE σ_D NON NOTA

$t = \dfrac{\overline{D} - \mu_D}{\frac{S_D}{\sqrt{n}}}$ DOVE $\overline{D} = \dfrac{\sum_{j=1}^{n} D_j}{n}$; $S_D = \sqrt{\dfrac{\sum_{j=1}^{n}(D_j - \overline{D})^2}{n-1}}$; t DI STUDENT $n-1$ GRADI DI LIBERTÀ ; α (5% o 1%)



$H_1: \mu_D < 0$ $H_1: \mu_D > 0$ $H_1: \mu_D \neq 0$

INTERVALLO DI CONFIDENZA PER LA MEDIA DELLE DIFFERENZE PER DUE POPOLAZIONI NON INDIPENDENTI (μ_D):

$\overline{D} \pm t_{n-1;\,\alpha/2}\dfrac{S_D}{\sqrt{n}}$

AREA ACCETTAZIONE: $1-\alpha/2$

CONFRONTO TRA LE PROPORZIONI DI DUE POPOLAZIONI

π_1 e π_2 PROPORZIONI DEI SUCCESSI NELLA POPOLAZIONE

p_1 e p_2 PROPORZIONI CAMPIONARIE DEI SUCCESSI (frequenze relative)

IPOTESI NULLA $H_0: \pi_1 = \pi_2$

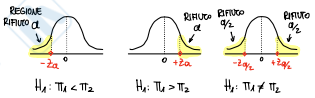
IPOTESI ALTERNATIVA $H_1: \pi_1 \neq \pi_2$; $H_1: \pi_1 < \pi_2$; $H_1: \pi_1 > \pi_2$

TEST Z SULLA DIFFERENZA TRA DUE PROPORZIONI

DOVE X_1 E X_2 SONO IL NUMERO DI SUCCESSI NEI DUE CAMPIONI

$Z = \dfrac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ DOVE $p_1 = \dfrac{X_1}{n_1}$ $p_2 = \dfrac{X_2}{n_2}$ $\bar{p} = \dfrac{X_1 + X_2}{n_1 + n_2}$

NORMALE $N(0,1)$; α (5% o 1%)



$H_1: \pi_1 < \pi_2$ $H_1: \pi_1 > \pi_2$ $H_1: \pi_1 \neq \pi_2$

INTERVALLO DI CONFIDENZA PER LA DIFFERENZA TRA DUE PROPORZIONI

$\pi_1 - \pi_2$: $(p_1 - p_2) \pm z_{\alpha/2}\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}$

AREA ACCETTAZIONE: $1-\alpha/2$

CONFRONTO TRA VARIANZE DI DUE POPOLAZIONI (INDIPENDENTI)

IPOTESI NULLA $H_0: \sigma_1^2 = \sigma_2^2$

IPOTESI ALTERNATIVA $H_1: \sigma_1^2 < \sigma_2^2$; $H_1: \sigma_1^2 > \sigma_2^2$; $H_1: \sigma_1^2 \neq \sigma_2^2$

TEST F SULLA DIFFERENZA TRA DUE VARIANZE CAMPIONARIE:

$F = S_1^2 / S_2^2$ (NUMERATORE / DENOMINATORE)

DISTRIBUZIONE F DI FISCHER CON (n_1-1) E (n_2-1) GRADI DI LIBERTÀ

α (5% o 1%)

VALORE CRITICO SUPERIORE F_U SI TROVA CON LE TAVOLE DI FISCHER;

MENTRE VALORE CRITICO INFERIORE F_L CON $F_L = 1/F_U^*$

DOVE F_U^* È IL VALORE CRITICO SUPERIORE CON GRADI DI LIBERTÀ INVERTITI: (n_2-1) E (n_1-1) (NUMERATORE / DENOMINATORE)

ANALISI DELLA VARIANZA (ANOVA) A UNA VIA

(GRUPPI DEFINITI SULLA BASE DI 1 SINGOLO FATTORE)

CONSENTE DI CONFRONTARE MEDIE DI PIÙ DI 2 GRUPPI → (POPOLAZIONI) INDIVIDUATI SULLA BASE DI 1 O PIÙ FATTORI DI INTERESSE E SONO FORMATI SECONDO I LIVELLI (MODALITÀ)

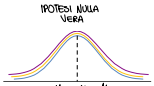
ANCHE SE SI PARLA DI ANALISI DELLA VARIANZA IN REALTÀ L'OGGETTO DI INTERESSE SONO LE DIFFERENZE TRA MEDIE NEI DIVERSI GRUPPI E TRAMITE L'ANALISI DELLA VARIABILITÀ

CONSIDERIAMO UN FATTORE DI INTERESSE A CARATTERIZZATO DA C LIVELLI (TRATTAMENTI) E UNA VARIABILE CAUSALE QUANTITATIVA X (RISPOSTA O RISULTATO)

I C GRUPPI RAPPRESENTANO POPOLAZIONI CON DISTRIBUZIONE NORMALE, CON TUTTI LA STESSA VARIANZA E LE OSSERVAZIONI ESTRATTE CASUALMENTE

GRUPPO 1:	LIVELLO A_1	$X_{11}, X_{21}, \ldots X_{1n1}$
GRUPPO 2:	LIVELLO A_2	$X_{12}, X_{22}, \ldots X_{2n2}$
...	...	...
GRUPPO C:	LIVELLO A_C	$X_{1c}, X_{2c}, \ldots X_{cnc}$

IPOTESI NULLA H_0: $\mu_1 = \mu_2 = \ldots = \mu_c$



IPOTESI ALTERNATIVA H_1: NON TUTTE LE μ_j SONO UGUALI (CON $j = 1,2,\ldots,c$)

PER VERIFICARE L'IPOTESI NULLA, LA VARIABILITÀ TOTALE (SST) (SOMMA DEI QUADRATI TOTALE) VIENE SCOMPOSTA IN DUE COMPONENTI:

- VARIABILITÀ TRA I GRUPPI (SSA)
- VARIABILITÀ ENTRO I GRUPPI (SSW)

$SST = SSA + SSW$

G.L. $n-1$ | G.L. $c-1$ | G.L. $n-c$

VARIABILITÀ TOTALE: $SST = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(X_{ij} - \overline{\overline{X}})^2$

DOVE $\overline{\overline{X}} = \dfrac{\sum_{j=1}^{c}\sum_{i=1}^{n_j} X_{ij}}{n}$ = MEDIA COMPLESSIVA, $n = \sum_{j=1}^{c} n_j$

VARIABILITÀ TRA GRUPPI: $SSA = \sum_{j=1}^{c} n_j(\overline{X}_j - \overline{\overline{X}})^2$

DOVE $\overline{X}_j = \dfrac{\sum_{i=1}^{n_j} X_{ij}}{n_j}$ = MEDIA CAMPIONARIA DEL j-ESIMO CAMPIONE

VARIABILITÀ ENTRO GRUPPI: $SSW = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(X_{ij} - \overline{X}_j)^2$

$n = \sum_{j=1}^{c} n_j \qquad T_j = \sum_{i=1}^{n_j} X_{ij} \qquad T = \sum_{j=1}^{c} T_j \qquad C = \dfrac{T^2}{n}$

$SST = \sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(X_{ij} - \overline{\overline{X}}\right)^2 \qquad SST = \sum_{j=1}^{c}\sum_{i=1}^{n_j} X_{ij}^2 - C$

$SSA = \sum_{j=1}^{c} n_j\left(\overline{X}_j - \overline{\overline{X}}\right)^2 \qquad SSA = \sum_{j=1}^{c}\dfrac{T_j^2}{n_j} - C$

$SSW = \sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(X_{ij} - \overline{X}_j\right)^2 \qquad SSW = SST - SSA$

DIVIDENDO CIASCUNA SOMMA DEI QUADRATI PER I RISPETTIVI GRADI DI LIBERTÀ, SI OTTENGONO TRE VARIANZE, O MEDIE DEI QUADRATI: MST, MSA, MSW.

$MST = \dfrac{SST}{n-1} \qquad MSA = \dfrac{SSA}{c-1} \qquad MSW = \dfrac{SSW}{n-c}$

INOLTRE VIENE RIPORTATA LA STATISTICA TEST F PER VERIFICARE L'IPOTESI NULLA:

$H_0 = \mu_1 = \mu_2 = \ldots = \mu_c$

H_1 = NON TUTTE LE μ_j SONO UGUALI TRA LORO

$F = \dfrac{SSA/(c-1)}{SSW/(n-c)} = \dfrac{MSA}{MSW}$

SE L'IPOTESI NULLA È **VERA** IL VALORE DOVREBBE ESSERE CIRCA 1

SE L'IPOTESI NULLA È **FALSA** IL VALORE DOVREBBE ESSERE MOLTO SUPERIORE ALL'UNITÀ

LA STATISTICA F HA DISTRIBUZIONE F DI FISHER CON $(c-1)$ GRADI DI LIBERTÀ AL NUMERATORE E $(n-c)$ GRADI DI LIBERTÀ AL DENOMINATORE "TABELLE DI FISHER" → F_U

REGIONE DI RIFIUTO: α ← $\alpha \simeq F_U$

REGIONE ACCETTAZIONE: $1-\alpha$

SE $F \geq F_U$ (CALCOLATO ≥ TAVOLE) L'IPOTESI NULLA VIENE RIFIUTATA

NELL'ANOVA CI SONO ALMENO 2 MEDIE SIGNIFICATIVAMENTE DIVERSE, PER TROVARLE USIAMO DEI METODI CHIAMATI CONFRONTI MULTIPLI → PROCEDURA DI TUKEY-CRAMER $CK(c-1)/2$

RANGE CRITICO $= Q_\alpha\sqrt{\frac{MSW}{2}\left(\frac{1}{n_j} + \frac{1}{n_{j'}}\right)}$

DOVE Q_α È IL VALORE CRITICO CON: c GRADI AL NUMERATORE E n-c GRADI AL DENOMINATORE

DUE MEDIE SONO SEGNALATE COME SIGNIFICATIVAMENTE DIVERSE TRA LORO SE LA DIFFERENZA IN VALORE ASSOLUTO DI TALI MEDIE È MAGGIORE DEL RANGE CRITICO

ES. LA FORNITORI

$k(k-1)/2 = 6$ COPPIE DI MEDIE

NUM = 4 DEN = 16

TABELLE DI FISHER → 4.05

RANGE CRITICO $= 4.05\sqrt{\frac{6.084}{2}\left(\frac{1}{5} + \frac{1}{5}\right)} = 4.4$

CONFRONTI	DIFF. MEDIE	RANGE CRITICO
1 VS 2	4.44	4.322
1 VS 3	...	4.322
1 VS 4	...	4.322
2 VS 3	...	4.322
2 VS 4	...	4.322
3 VS 4	...	4.322

4.44 ≥

INTERVALLO DI CONFIDENZA AL $(1-\alpha)\%$:

$\overline{X}_j - \overline{X}_{j'} \pm Q_\alpha\sqrt{\frac{MSW}{2}\left(\frac{1}{n_j} + \frac{1}{n_{j'}}\right)}$

DUE MEDIE SONO SEGNALATE SIGNIFICATIVAMENTE DIVERSE TRA LORO SE L'INTERVALLO DI CONFIDENZA NON INCLUDE IL VALORE 0

ES. (0.269, 9.211)

ANALISI DELLA VARIANZA (ANOVA) A DUE VIE

IPOTESI:
- EFFETTO PRINCIPALE DEL FATTORE A
- EFFETTO PRINCIPALE DEL FATTORE B
- INTERAZIONE TRA A E B

Origine della variabilità	Somma dei quadrati	Gradi di libertà	Media dei Quadrati	F_0
A trattamenti	SS_A	a-1	$MS_A = SS_A/(a-1)$	MS_A/MS_E
B trattamenti	SS_B	b-1	$MS_B = SS_B/(b-1)$	MS_B/MS_E
Interazione	SS_{AB}	(a-1)(b-1)	$MS_{AB} = SS_{AB}/(a-1)(b-1)$	MS_{AB}/MS_E
Errore	SS_E	ab(n-1)	$MS_E = SS_E/(ab(n-1))$	
Totale	SS_T	abn-1		

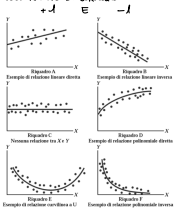
- GRAFICO DI PROBABILITÀ NORMALE DEI RESIDUI
- GRAFICO DEI RESIDUI RISPETTO AI VALORI STIMATI
- GRAFICO DEI RESIDUI NELL'ORDINE TEMPORALE DELLA RACCOLTA DEI DATI
- GRAFICO DEI RESIDUI RISPETTO AD ALTRE VARIABILI

CONNESSIONE E CORRELAZIONE

SE VI È ASSENZA DI RELAZIONI TRA DUE O PIÙ VARIABILI
→ INDIPENDENTI (QUANDO IL VARIARE DI UNA NON MODIFICA L'ALTRA)

MISURE DI CONNESSIONE SE VARIABILI CATEGORIALI (tipo ASS E REL)
MISURE DI CORRELAZIONE SE VARIABILI NUMERICHE (LEGAME STATISTICO)
↓
PER DUE VARIABILI QUANTITATIVE SI PUÒ USARE UN DIAGRAMMA DI DISPERSIONE → SERVE A RILEVARE EVENTUALI TENDENZE

COEFFICIENTE DI CORRELAZIONE TRA +1 E −1



COEFFICIENTE DI CORRELAZIONE:

$$r = \frac{cov(X,Y)}{S_X S_Y} \quad \frac{\text{COVARIANZA}}{\text{PRODOTTO D.S. CAR.}}$$

$$COV(X,Y) = \frac{\sum_{j=1}^{n}(x_j-\bar{x})(y_j-\bar{y})}{n-1}$$

$$\bar{x} = \frac{\text{SOMMA VALORI X}}{\text{NUMERO DELLE X}} \quad \bar{y} = \frac{\text{SOMMA VALORI Y}}{\text{NUMERO DELLE Y}}$$

OGNI n VALORE DI X $(x_j - \bar{x})$

$$S_X = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x_j-\bar{x})^2} \quad S_Y = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(y_j-\bar{y})^2}$$

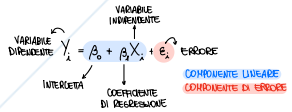
MEDIA DEI VALORI TROVATI

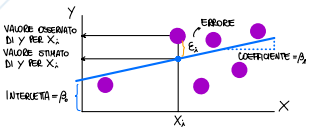
CORRELAZIONE E REGRESSIONE

STUDIO DELLA ASSOCIAZIONE TRA VARIABILI QUANTITATIVE

MODELLO ATTRAVERSO CUI PREVEDERE I VALORI DI UNA VARIABILE DIPENDENTE O RISPOSTA A PARTIRE DAI VALORI DI UNA O PIÙ VARIABILI INDIPENDENTI O ESPLICATIVE

- LINEARE SEMPLICE
  - LINEARE O CURVILINEA
  - RELAZIONE FORTE, DEBOLE O NESSUNA RELAZIONE





STIMA O VALORE PREDETTO DI Y PER L'UNITÀ i-ESIMA: $\hat{Y}_i = b_0 + b_1 X_i$ (VALORE OSSERVATO DELLA X PER L'UNITÀ i-ESIMA)

b_0: STIMA DELL'INTERCETTA; b_1: STIMA DEL COEFFICIENTE DI REGRESSIONE

IL SEGNO DI b_1 INDICA SE LA RELAZIONE LINEARE È POSITIVA O NEGATIVA

METODO DEI MINIMI QUADRATI
- DETERMINARE b_0 E b_1

$$min\sum(Y_i-\hat{Y}_i)^2 = min\sum(Y_i-(b_0+b_1X_i))^2$$

$$b_1 = \frac{SQXY}{SQX}; \quad b_0 = \bar{Y} - b_1\bar{X}$$

$$\bar{Y} = \frac{\sum_{i=1}^{n}Y_i}{n}; \quad \bar{X} = \frac{\sum_{i=1}^{n}X_i}{n}$$

$$SQXY = \sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y}) = \sum_{i=1}^{n}X_iY_i - \frac{\left(\sum_{i=1}^{n}X_i\right)\left(\sum_{i=1}^{n}Y_i\right)}{n}$$

$$SQX = \sum_{i=1}^{n}(X_i-\bar{X})^2 = \sum_{i=1}^{n}X_i^2 - \frac{\left(\sum_{i=1}^{n}X_i\right)^2}{n}$$

INTERPOLAZIONE ED ESTRAPOLAZINE

PREVISIONE DI Y CORRISPONDE AD UN VALORE DI X INTERNO ALL'INTERVALLO

PREVISIONE DI Y CORRISPONDE AD UN VALORE DI X CHE NON CADE NELL'INTERVALLO

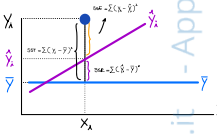
VARIABILITÀ TOTALE DELLA Y IN DUE COMPONENTI

$$SST = SSR + SSE$$

SST: SOMMA TOTALE DEI QUADRATI
$SST = \sum(Y_i - \bar{Y})^2$
MISURA LA DISPERSIONE DEI VALORI DELLA Y ATTORNO ALLA PROPRIA MEDIA

SSR: SOMMA DEI QUADRATI DELLA REGRESSIONE
$SSR = \sum(\hat{Y}_i - \bar{Y})^2$
MISURA LA VARIAZIONE DELLA Y ATTRIBUIBILE ALLA VARIABILE X

SSE: SOMMA DEI QUADRATI DEGLI ERRORI
$SSE = \sum(Y_i - \hat{Y}_i)^2$
MISURA LA VARIAZIONE DELLA Y NON ATTRIBUIBILE ALLA X



COEFFICIENTE DI DETERMINAZIONE

$$r^2 = \frac{SSR}{SST} \quad 0 \le r^2 \le 1$$

↓ NESSUNA RELAZIONE TRA X E Y

→ CORRELAZIONE LINEARE DIRETTA (+1) O INVERSA (−1) PERFETTA

ERRORE STANDARD DELLA STIMA

$$S_{YX} = \sqrt{\frac{SSE}{n-2}} = \frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{n-2}$$

MINORE S_{YX} → RELAZIONE FORTE
MAGGIORE S_{YX} → RELAZIONE DEBOLE

ANALISI DEI RESIDUI e_i

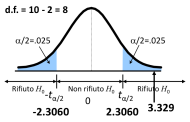
$e_i = Y_i - \hat{Y}_i$ RESIDUO

IPOTESI DI LINEARITA' (retta orizzontale)
DISTRIBUZIONE NORMALE DEGLI ERRORI
OMOSCHEDASTICITA' (varianza)
INDIPENDENZA DAGLI ERRORI (indipendenza)

INFERENZA NEL MODELLO DI REGRESSIONE

$H_0: \beta_0 = 0$ (NESSUNA RELAZIONE LINEARE)
$H_1: \beta_0 \neq 0$ (ESISTENZA RELAZIONE LINEARE)

$$t_{STAT} = \frac{b_1 - \beta_1}{S_{b_1}} = \frac{0.10977 - 0}{0.032973} = 3.32928$$



$$F_{STAT} = \frac{MSR}{MSE}$$

$$MSR = \frac{SSR}{1}$$

$$MSE = \frac{SSE}{n-2}$$

RIFIUTO H_0 SE $F > F_U$
CON F_U VALORE CRITICO CHE LASCIA A DESTRA PROBABILITA' PARI AD α

$b_1 \pm t_{\alpha/2} S_{b_1}$ G.L. = n-2

APPLICAZIONE CORRETTA DELL'ANALISI DI REGRESSIONE

Sintetizziamo i passi per una corretta applicazione dell'analisi di regressione:

1. iniziare l'analisi con un'attenta osservazione del diagramma di dispersione, per cogliere l'eventuale relazione tra X e Y;
2. dopo la stima del modello e prima di passare a impiegarne i risultati, verificare se le assunzioni alla base del modello sono soddisfatte:
   a) grafico dei residui rispetto alla variabile X per verificare l'assunzione di omoschedasticità delle varianze;
   b) istogramma dei residui e normal probability plot, per verificare l'assunzione di normalità degli errori;
   c) grafico dei residui rispetto all'ordine temporale di raccolta dei dati, per verificare l'assunzione di indipendenza degli errori.
3. se le ipotesi alla base del modello non risultano violate, procedere alle inferenze sul modello (test di ipotesi sui coefficienti, intervalli di confidenza per la risposta media e la previsione).