UNIVERSITA' DEGLI STUDI "MAGNA GRÆCIA"DI CATANZARO

Polo Universitario di Vibo Valentia

Corso di Laurea in Infermieristica

Statistica Medica

Dr. Agostino Scardamaglio

Università Magna Graecia di Catanzaro
Corso di laurea in Infermieristica

Programma di Statistica Medica

Dr. Agostino Scardamaglio

1. Introduzione alla statistica medica
1.1 Finalità della statistica
1.2 Termini e definizioni
1.3 Dati e concetto di funzione

2. Metodi di rilevazione e presentazione dei dati
2.1 Ricerca osservazionale e sperimentale
2.2 Il disegno d'indagine
2.3 Strumenti per la ricerca scientifica: griglia d'osservazione, questionario e intervista
2.4 Tabelle
2.5 Grafici

3. Distribuzioni di frequenza
3.1 Frequenze assolute, relative, cumulate
3.2 Classi di frequenza
3.3 Contingenze
3.4 Misure di connessione e Chi quadro

4. Variabilità della distribuzione dei dati
4.1 Misure di tendenza centrale: media, mediana e moda
4.2 Misure di dispersione: range, range interquartile e deviazione standard
4.3 Misure di posizione: valore di z, i 5 valori di sintesi, il diagramma box-plot
4.4 Analisi della varianza, Errori sistematici e casuali
4.5 Relazioni tra variabili: il concetto di correlazione

5. Teoria della probabilità
5.1 Le varie definizioni del concetto di probabilità
5.2 Probabilità totale e composta, legge dei grandi numeri
5.3 Elementi di calcolo combinatorio
5.4 Teorema di Bayes e tests diagnostici

6. Dalla probabilità all'inferenza
6.1 Distribuzioni di probabilità, gradi di libertà
6.2 Distribuzione normale e di t di Student
6.3 Inferenza, statistiche, parametri, stimatori e stime
6.4 Stima della media di una popolazione
6.5 Teoria dei campioni e tipi di campionamento

7. Il Sistema d'ipotesi
7.1 Tests parametrici e non parametrici
7.2 Teorema del limite centrale
7.3 Requisiti di un test statistico
7.4 Intervalli di confidenza
7.5 Verifica delle ipotesi
7.6 Tipi di errore

Perché bisogna conoscere la STATISTICA?

1. **LA MEDICINA STA DIVENTANDO QUANTITATIVA**
2. **LA STATISTICA PERVADE LA LETTERATURA MEDICA**
3. **E' INDISPENSABILE PER PROGRAMMARE, ESEGUIRE, INTERPRETARE GLI STUDI E LE RICERCHE IN AMBITO BIOMEDICO.**

UMG Dubium sapientiae initium
Lezione I

Metodi di analisi statistica

Statistica Descrittiva: si occupa della presentazione, organizzazione e sintesi dei dati: tabelle, grafici, indici di sintesi

Statistica Inferenziale: ci permette di generalizzare i risultati ottenuti dai dati raccolti da un piccolo campione per una popolazione più ampia:

- Stima di parametri
- Test di ipotesi

Finalità della Statistica

Descrivere i dati

condensare anche un gran numero di dati rilevati in pochi valori riassuntivi, capaci di indicare importanti proprietà della popolazione oggetto di indagine

Esplorare le relazioni

definire e descrivere le relazioni tra le variabili rilevate

Fare previsioni

utilizzare i dati raccolti per prevedere i valori che ci si aspetta di trovare nella popolazione oggetto di indagine in particolari condizioni

UMG Dubium sapientiae initium
Lezione I

Classificare

descrivere ed analizzare gruppi definiti sulla base di caratteristiche comuni misurate dalle variabili rilevate.

Valutare ipotesi

stabilire quanto è verosimile che esista una relazione tra le variabili.

Generare ipotesi

grazie alle 5 fasi precedentemente descritte le variabili divengono meglio comprensibili, ed è possibile che questo porti a proporre nuove idee a proposito della popolazione indagata.

Tutti gli obiettivi che abbiamo elencato sono sistemi differenti per affrontare lo stesso problema:

LA VARIABILITA'

I metodi statistici chiamano in causa osservazioni che variano da campione a campione e portano un certo grado di incertezza in ogni analisi.

L'obiettivo primario di pressoché tutti i metodi statistici è quello di comprendere il comportamento di un fenomeno tenendo conto degli effetti di questa variabilità.

UMG Dubium sapientiae initium
Lezione I

Variabilità biologica

- diversi fattori contribuiscono a differenziare i soggetti in studio (ad es. aspetto esteriore, attività fisiologiche, fattori genetici)

- esiste una variabilità anche in uno stesso individuo dovuta al tempo o ad altri fattori (metabolici, emozionali, etc.)

Variabilità della misura

- la variabilità dei dati dipende anche dall'operazione di misura
- gli errori di misura possono essere legati all'operatore, alla strumentazione o alla tecnica impiegata
- gli errori di misura possono riguardare le rilevazioni riferite a soggetti diversi o la ripetizione di una misura su uno stesso soggetto
- l'entità dell'errore aumenta quando una misura viene ricavata indirettamente da altre misure

UMG Dubium sapientiae initium
Lezione I

Termini e definizioni

Censo

la collezione di dati da ogni elemento della popolazione

Campione

un sottoinsieme degli elementi della popolazione

Popolazione

la completa collezione degli elementi (soggetti, misure, campioni chimici, ecc.) oggetto di studio. La collezione è completa nel senso che include assolutamente TUTTI gli elementi d'interesse.

Termini e definizioni

Parametro

un numero che descrive una caratteristica della **POPOLAZIONE**

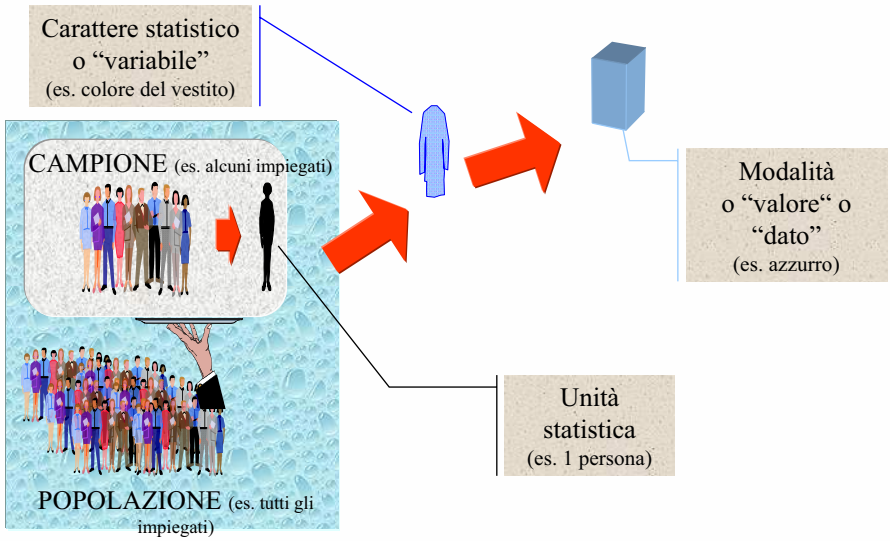
Statistica

misura numerica che descrive una qualche caratteristica del **CAMPIONE**

Popolazione ⟺ **Parametro**

Campione ⟺ **Statistica**

UMG Dubium sapientiae initium
Lezione I



Tipologia dei dati

Quantitativi

numeri che rappresentano conteggi o misure (es. peso, altezza, ecc.)

Qualitativi (o categorici)

Caratteri che possono essere classificati in diverse categorie distinte da caratteristiche non numeriche (es.sesso, titolo di studio, ecc.)

UMG Dubium sapientiae initium
Statistica medica

Quantitativi

Discreti

sono definiti quando il numero di possibili valori che la variabile può assumere è una quantità "finita" ovvero enumerabile. Es. numero di uova deposte da una gallina = 0, 1, 2, 3, …

Continui

i dati presentano un'infinità di possibili valori che corrispondono ad una qualche scala continua che copre un certo intervallo senza interruzioni o salti. Es. peso di neonato alla nascita (3.500 grammi)

Dati continui …

SCALA INTERVALLARE:

Non esiste uno "zero" reale: il punto di origine è arbitrario
Es: temperatura in °C, tempo

Ha senso fare differenze, ma NON rapporti (40 °C non è il doppio di 20°C!)

SCALA RAPPORTO:

Esiste uno zero reale: il punto di origine è definito

E' possibile fare rapporti e moltiplicazioni tra valori. Es: peso, altezza, temperatura in Kelvin, ematocrito, ecc.

UMG Dubium sapientiae initium
Statistica medica

Qualitativi

Nominali

sono caratterizzati da dati costituiti da nomi, etichette, categorie. NON possono essere disposti in un qualche ordine logico (crescente o decrescente)

Es. Sesso: Maschio-Femmina Stato:Vivo-Morto

Ordinali

sono dati che possono essere disposti secondo un ordine definito, ma le differenze tra i valori non possono essere determinate o sono prive di senso

Es. Titolo di studio, Giovane-Adulto-Anziano ecc.

Il dato statistico

Il dato statistico può esprimere l'intensità oppure la frequenza con cui si manifesta un fenomeno o carattere.

La frequenza indica il numero di volte in cui si è manifestato il carattere studiato.

L'intensità indica il valore, la misura, la quantizzazione del fenomeno oggetto di studio.

UMG Dubium sapientiae initium
Statistica medica

Dato grezzo
D. non disposto secondo un particolare ordinamento ma secondo il succedersi cronologico dell'acquisizione.

Dato ordinato
Serie (per variabili qualitative). Valori disposti in ordine di grandezza cresente o decrescente. Es.: elenco di persone in ordine alfabetico.

Seriazione (per variabili quantitative). Valori disposti secondo l'incremento o il decremento numerico. Es.: farmaci ordinati secondo l'incremento o il decremento del costo.

Fenomeno
E' l'evento oggetto dell'indagine statistica. Viene anche definito carattere. I fenomeni analizzabili statisticamente sono caratterizzati da risultati incerti.

Variabile
Carattere quantitativo espresso mediante numero su scala, intervalli o rapporti.
Può essere intesa come funzione statistica che associa l'insieme rappresentato delle unità statistiche (varabile indipendente x) all'insieme dei rispettivi valori di frequenza delle modalità (varabile dipendente f(x)).

Statistica medica

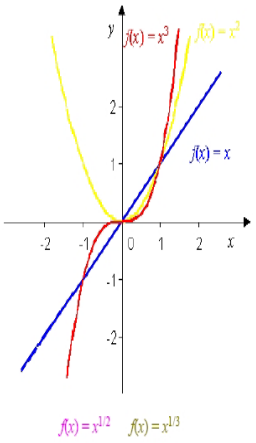
Il concetto di funzione

E' importante sia per la matematica che per la statistica: infatti cercare le cause, le implicazioni, le conseguenze e l'utilita' di una funzione, significa mostrare il legame che esiste fra cose diverse.
Fra tutte le definizioni di funzioni si è preferita la definizione di Dirichlet:
Si definisce funzione y della variabile x un legame fra due variabili, una detta variabile indipendente x e l'altra detta variabile dipendente y, tali che abbiano senso le operazioni da effettuare sulla x per ottenere i valori della y e per ogni valore della x corrisponda un solo valore della y

$$y= f(x)$$

Il concetto grafico di funzione

Se attribuiamo ad x il significato di ascissa di un punto del piano cartesiano 0xy ed al corrispondente valore di y il significato di ordinata dello stesso punto, allora ad ogni coppia (x, f(x)) di valori corrispondenti possiamo associare un punto P del piano aventi tali valori per coordinate. L'espressione grafica del succedersi giustapposti di questi infiniti punti e costituita da una linea denominata diagramma o grafico della funzione



metodi per la rilevazione dei dati

- osservazione
- esperimento

UMG Dubium sapientiae initium
Lezione II

La ricerca osservazionale

Ricerca descrittiva

ha lo scopo di descrivere in modo sistematico una particolare situazione o evento per spiegare o prevedere in che modo la situazione o l'evento possano presentarsi nel futuro o essere modificati

Ricerca correlazionale

studia la relazione tra variabili

La ricerca sperimentale

Si serve dell'esperimento che consiste nella modificazione deliberata di alcune variabili in una data situazione, allo scopo di alterarne la natura in modo controllato e di verificare la relazione di causalità eventualmente esistente fra due o più variabili.

UMG Dubium sapientiae initium
Lezione II

Classificazione della ricerca in base al tempo

- **Ricerca retrospettiva**
  prende in esame dati già raccolti
- **Ricerca prospettica**
  prende in esame dati raccolti nel presente
- **Ricerca trasversale**
  i dati vengono raccolti una sola volta, senza follow up
- **Ricerca longitudinale**
  i dati vengono raccolti in momenti differenti su una coorte di soggetti seguiti nel tempo

Disegno d'indagine

1. Definizione degli obiettivi
2. Definizione dell'universo e scelta della lista
3. Scelta del periodo di riferimento
4. Definizione del piano di campionamento
5. Scelta delle variabili da rilevare
6. Definizione dell'unità di analisi e di rilevazione
7. Scelta della tecnica di rilevazione
8. Formulazione del questionario e pretest

UMG Dubium sapientiae initium
Lezione II

Strumenti per la ricerca scientifica

- la griglia di osservazione
- il questionario
- l'intervista

Griglia di osservazione

1.formulazione di una o più domande per ogni obiettivo conoscitivo in relazione alla complessità del fenomeno ipotizzato

2.formulazione delle domande di controllo per verificare l'attendibilità delle risposte

3.assegnazione di un ordine di successione:

raggruppando le domande relative allo stesso argomento

rispettando il processo cognitivo dell'intervistato

(es. *disposizione a imbuto* o *funnel sequence*: dal generale allo specifico) segnalando il passaggio da un argomento all'altro (*effetto alone*)

4.predisposizione di un'introduzione che illustri la committenza, gli scopi, il tema e gli argomenti della ricerca

UMG Dubium sapientiae initium
Lezione II

Il questionario

A) **ADDESTRAMENTO DEGLI INTERVISTATORI** *(breafing)* allo scopo di:

- presentare sinteticamente gli obiettivi della ricerca
- chiarire la struttura e i contenuti del questionario
- illustrare le tecniche di impiego
- istruire all'approccio con l'intervistato
- stabilire i luoghi e i tempi di somministrazione e di consegna
- definire i criteri di selezione del campione

B) **CODIFICA** attribuzione di un codice a ciascuna delle possibili risposte per ogni domanda, che consenta l'elaborazione statistica o l'analisi del contenuto

Tipologia d'indagine

STANDARDIZZATA: le domande sono tutte precedentemente codificate e somministrate senza chiedere ulteriori chiarimenti, secondo un ordine rigido (→ analisi quantitativa e comparativa)

NON-STANDARDIZZATA: l'intervistatore si limita a seguire una scaletta (non necessariamente scritta) adattando la formulazione delle domande ai contesti in cui si inscrive il colloquio (→ scoperta)

SEMI-STANDARDIZZATA: l'intervistatore pone le domande secondo una forma e una sequenza precedentemente stabilita, tuttavia è libero di chiedere chiarimenti ed approfondimenti della risposta

UMG Dubium sapientiae initium
Lezione II

Tipi di domande

CHIUSE

il tipo e il numero delle risposte previste dal ricercatore sono le uniche tra le quali l'intervistato può scegliere

APERTE

consentono all'intervistato di esprimere anche opinioni che aggiungono dettagli o che si discostano, anche se di poco, da quelle previste

LIBERE

non prevedono alcun vincolo per la risposta se non la coerenza e un ragionevole limite di tempo

Tipi di somministrazione del questionario

CON INTERVISTATORE

tecnica di rilevazione, che si esprime nella forma dell'interazione verbale e che permette di giungere a informazioni specifiche, relative ad un fenomeno le cui cause non sempre sono chiare ai soggetti che nel fenomeno sono immersi

AUTOSOMMINISTRATO

Il questionario viene consegnato al rispondente previa comunicazione delle istruzioni per la sua compilazione

UMG Dubium sapientiae initium
Lezione II

Tabelle di frequenza

elencano le classi o categorie di valori, insieme alle frequenze assolute (conteggi degli elementi) entro ciascuna categoria, e frequenze relative

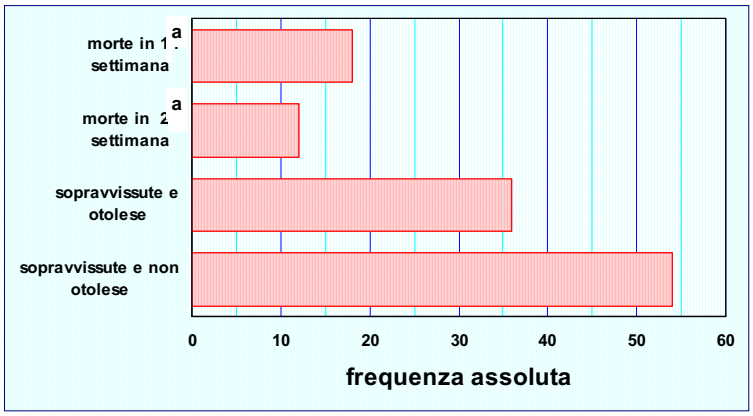
estremi di classe		valore centrale	frequenze
apparenti	reali		n
44.25-45.75	44.3-45.7	**45.0**	2
45.75-47.25	45.8-47.2	**46.5**	5
47.25-48.75	47.3-48.7	**48.0**	7
48.75-50.25	48.8-50.2	**49.5**	14
50.25-51.75	50.3-51.7	**51.0**	16
51.75-53.25	51.8-53.2	**52.5**	9
53.25-54.75	53.3-54.7	**54.0**	5
54.75-56.25	54.8-56.2	**55.5**	1
56.25-57.75	56.3-57.7	**57.0**	1

UMG Dubium sapientiae initium
Lezione II

Rappresentazioni grafiche

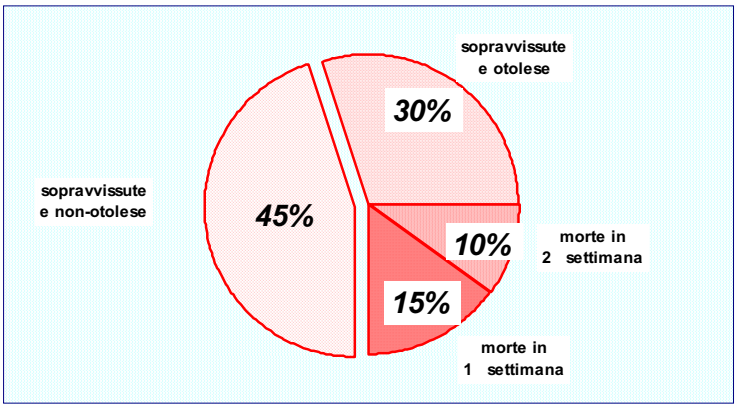
elencano le classi o categorie di valori, insieme alle frequenze assolute (conteggi degli elementi) entro ciascuna categoria, e frequenze relative

Grafico a barre



UMG Dubium sapientiae initium
Lezione II

Grafico a torta

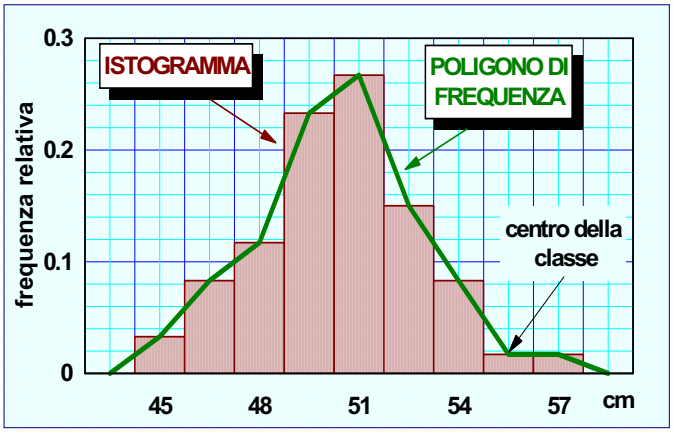


Istogrammi e poligoni di frequenza

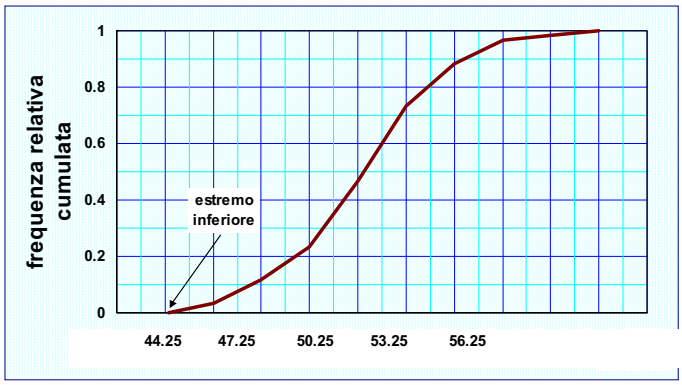
Negli istogrammi e nei poligoni di frequenza le **frequenze** sono **proporzionali all'area** (delimitata dalla spezzata che li costituisce e inclusa tra due valori reali sull'asse orizzontale), e non all'altezza della figura. Ovviamente, quando le classi hanno tutte la stessa ampiezza, l'area è proporzionale anche all'altezza. I valori riportati sull'asse verticale indicano la densità di frequenza per una prefissata ampiezza di classe

UMG Dubium sapientiae initium
Lezione II

Istogramma



Ogiva di Galton



Le informazioni raccolte per essere "trattate" da un computer devono essere organizzate in strutture chiamate comunemente

Data Base o File Dati.

Le informazioni vengono, comunemente, organizzate per riga, cioè su ogni riga, consecutivamente, vengono elencati i dati relativi ad un soggetto.

UMG Dubium sapientiae initium
Lezione III

N.	NOME	SESSO	ETA'	ALTEZZA	PESO	PAS	GLIC.
1	Rossi Amerigo	M	32	172	64	140	190
2	Bianchi Paolo	M	47	170	80	148	180
3	ValenziAlberica	F	45	168	51	125	150
4	Alinori Alfonso	M	27	183	85	130	170
5							
6							

Concetto di tabella

Forma nella quale si compendia la rilevazione statistica (matrice dei dati).
La **matrice dei dati** presenta tante righe quante sono le unità statistiche osservate e tante colonne quante sono le variabili statistiche considerate (con l'aggiunta di una colonna ed una riga per le intestazioni.
In essa si rappresentano il valore delle variabili delle **n colonne** associate alle **n righe** delle unità statistiche osservate o del loro raggruppamento in classi di modalità.

UMG Dubium sapientiae initium
Lezione III

Elementi di una tabella

Le righe. Contengono il valore delle variabili associate (funzionali) alle singole unità statistiche ovvero al loro raggruppamento in classi di modalità (etichette delle categorie).
La prima riga descrive nella prima cella il carattere della singola unità statistica ovvero dei suoi raggruppamenti in classi di modalità. Nelle celle successive viene descritta la variabile associata all'unità o la classe di modalità.

Le colonne. Contengono i valori delle variabili associate (funzionali) alle singole unità statistiche ovvero al loro raggruppamento in classi di modalità.

Notazione statistica

- *sommatoria*: $\sum_{i=1}^{n} x_i$

L'espressione si legge: sommatoria, per i che va da 1 a n, delle i-^{me} osservazioni della variabile x.

- *pedice* (i, j o altro) associato ad una variabile:
- x_i significa i-^{ma} osservazione della variabile x.

UMG Dubium sapientiae initium
Lezione III

Notazione statistica

Esempio: la variabile x assume i seguenti 5 valori:

i	x
1	3
2	8
3	4
4	2
5	10.7

quindi n = 5. Se i = 2, allora $x_i = x_2 = 8$.

$$\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5 = 27.7$$

Quindi somma dei 5 valori di x.

Sintesi tabellare dei caratteri statistici:

Se abbiamo n dati relativi ad un indagine condotta su n individui ad ogni modalita x_i del carattere X andiamo ad associare il numero di volte in cui la modalità si manifesta n_i

n= numero delle unità statistiche rilevate
X=carattere oggetto di studio
k=num totale dei diversi valori assunti dal carattere X (modalità)
x_i=modalità i-esima del carattere X $\quad i$=1,….,k
n_i=frequenze assolute

Lezione III

Carattere

E' l'oggetto dello studio che si osserva, si misura e si rileva attraverso le unità statistiche. Esempi: il carattere sesso ha le modalità M e F.
Il carattere tempo di spostamento si manifesta attraverso infinite modalità (n. reali positivi).
Il carattere n. di figli ha come modalità i numeri interi positivi e lo zero.

Tipi di carattere

C. Qualitativi (Mutabili)
Sono espressi in forma nominale. Si dividono in:
Sconnessi. Non ordinabili sec. criteri oggettivi (colore dei capelli, gelato preferito).
Ordinabili. Seguono una progressione (graduatorie)
Ciclici. Sono ordinabili in modo ciclico con inizio arbitrario.

C. Quantitativi (variabili)
Sono espressi mediante numeri su scala, intervallo o rapporto. Si dividono in:
Discreti. Le modalità sono n. interi positivi.
Continui. Le modalità sono n. reali.

UMG Dubium sapientiae initium
Lezione III

Modalità

Denominazione delle varie manifestazioni del fenomeno nell'ambito della variabilità.
Il concetto è strettamente connesso a quello di classi di frequenza in quanto le n unità statistiche dei dati grezzi vengono raggruppate in classi di f. secondo le k modalità del carattere.
L'individuazione e la classificazione delle modalità relativa a caratteri qualitativi e quantitativi discreti e in genere agevole. Si deve ricorrere talvolta a qualche artificio per distinguere le modalità di caratteri quantitativi continui.
Le modalità devono essere esaustive e non sovrapposte.
Per Maria il carattere peso in Kg assume modalità 55 mentre il carattere colore degli occhi assume modalità verde.

Frequenze

I dati numerici ordinati (es. fatture ordinate secondo l'importo) costituiscono una seriazione. La differenza tra il numero più grande e quello più piccolo di una seriazione si chiama **campo di variazione** o **range**.

Quest'ultimo può essere diviso in un certo numero di **classi** di ampiezza diversa.

La conta del numero di dati che cadono all'interno di ciascuna classe, costituisce la **frequenza**.

UMG Dubium sapientiae initium
Lezione III

Aspetti della frequenza

Frequenza assoluta: N. intero che rappresenta il numero di unità statistiche sulle quali è stata osservata la medesima modalità.
Frequenza relativa: Rapporto tra la frequenza assoluta e il numero totale delle osservazioni
Frequenza percentuale: Frequenza relativa moltiplicato 100.
Frequenza cumulata (assoluta, relativa, percentuale): Somme di tutte le frequenze che si susseguono via via dalla prima all'ultima classe.

La distribuzione di frequenza

E' una delle rappresentazioni statistiche fondamentali. Si costruisce raggruppando in classi le n unità statistiche secondo le k modalità del carattere osservato. In pratica:
1) si individuano i numeri maggiore e minore tra i dati grezzi tra i quali è contenuto il campo di variazione (range);
2) si divide il campo di variazione per un numero opportuno di classi;
3) si contano il numero di dati che cadono all'interno di ciascuna classe (frequenza assoluta)

UMG Dubium sapientiae initium
Lezione III

Sudivisione in classi di frequenza

- Il numero di classi deve essere equilibrato (circa la radice quadrata del n. di osservazioni).
- Le classi devono avere la stessa ampiezza.
- Le classi devono in genere essere limitate in un intervallo caratterizzato da un limite sup. ed inf.

Si devono il più possibile evitare classi aperte. L'ampiezza delle classi o modulo (differenza tra limite superiore ed inferiore) deve essere equilibrato.

Limiti superiori di una classe

Sono i valori più grandi che possono effettivamente appartenere alla classe

Limiti superiori

Classe	Frequenze
0 - 2	20
3 - 5	14
6 - 8	15
9 - 11	2
12 - 14	1

UMG Dubium sapientiae initium
Lezione III

Limiti inferiori di una classe

Sono i valori più piccoli che possono effettivamente appartenere alla classe

Limiti inferiori

Classe	Frequenze
0 - 2	20
3 - 5	14
6 - 8	15
9 - 11	2
12 - 14	1

Il **n. delle classi** deve essere equilibrato (circa la radice quadrata del n. delle osservazioni).

$N.Cls = \sqrt{n}$ N.Cls= n. classi n= n. osservazioni

L'ampiezza delle classi deve essere significativa, cioè, equiampia.

Le classi devono essere continue (devono essere considerati tutti i valori nel campo di variazione della variabile) e contigue (non ci devono essere sovrapposizioni tra classi o discontinuità.

La prima e/o l'ultima classe possono essere aperte, cioè possono essere definite in modo che non sia specificato uno degli estremi (quello inferiore per la prima classe e quello superiore per l'ultima).

Lezione III

FREQUENZE ASSOLUTE

FREQUENZE RELATIVE

FREQUENZE PERCENTUALI

FREQUENZE CUMULATE $N_i = \sum_{j=1}^{i} n_j$

MODALITA'

x_i	n_i	f_i	p_i	N_i
x_1	n_1	$n_1/n=f_1$	f_1*100	n_1
x_2	n_2	$n_2/n=f_2$	f_1*100	$n_{1+}n_2$
x_3	n_3	$n_2/n=f_2$	f_1*100	$n_{1+}n_{2+}n_3=n$
	n	1	100	

Analogamente alle Ni possono essere costruite anche le F_i e le P_i

Tabelle di frequenza a doppia entrata

Se X ha h modalità e Y ha K modalità le frequenza marginali vengono così calcolate:

X/Y	y_1	…	y_j	…	y_k	**Tot.**
x_1	n_{11}	…	n_{1j}	…	n_{1k}	$n_{1\cdot}$
…	…	…	…	…		
x_i	n_{i1}	…	n_{ij}	…	n_{ik}	$n_{i\cdot}$
…	…	…	…	…		
x_h	n_{h1}	…	n_{hj}	…	n_{hk}	$n_{h\cdot}$
Tot.	**n.1**	**…**	**n.j**	**…**	**n.k**	**n..**

X/Y	y_1	…	y_j	…	y_k	**Tot.**
x_1	f_{11}	…	f_{1j}	…	f_{1k}	$f_{1\cdot}$
…	…	…	…	…		
x_i	f_{i1}	…	f_{ij}	…	f_{ik}	$f_{i\cdot}$
…	…	…	…	…		
x_h	f_{h1}	…	f_{hj}	…	f_{hk}	$f_{h\cdot}$
Tot.	**f.1**	**…**	**f.j**	**…**	**f.k**	**1**

Distribuzioni marginali

Distribuzioni condizionate

Tab.di frequenza assoluta

$$n_{i.} = \sum_{j=1}^{k} nij \qquad n_{i.} = \sum_{i=1}^{h} nij$$

Tab.di frequenza relativa

$$f_{i\cdot} = \sum_{j=1}^{k} f_{ij} \qquad f_{i\cdot} = \sum_{i=1}^{h} f_{ij}$$

Lezione III

Tabelle doppie e misure di connessione

Per analizzare le relazioni o connessioni tra due caratteri statistici si utilizzano le tabelle doppie o tabelle a doppia entrata o tabelle d'incrocio (dall'inglese cross tabulation).
In generale una tabella mette in relazione le frequenze congiunte di due caratteri, ad esempio X e Y; Vengono generalmente indicate con **nij** la frequenza assoluta congiunta relativa alla i-esima modalità di riga e j-esima di colonna.

Tabelle doppie: 3 tipi di frequenze relative

Oltre alle frequenze assolute è possibile calcolare tre tipi di frequenze relative espresse in %:

- frequenze relative rispetto al totale delle unità statistiche;
- frequenze relative condizionate rispetto al totale di riga (detto marginale di riga);
- frequenze relative condizionate rispetto al totale di colonna (marginale di colonna).

UMG Dubium sapientiae initium
Lezione III

Tabelle doppie e concetto di funzione

Dati due caratteri ci si chiede se la conoscenza delle modalità di un carattere consenta di fare delle ipotesi sulle modalità del secondo carattere e il tipo di relazione che intercorre tra i due. Considerazioni:

- In fisica vi è una relazione MATEMATICA ben precisa tra lo spazio percorso ed il tempo impiegato nella caduta di un sasso sottoposto alla forza di gravità;
- in matematica finanziaria vi è una relazione matematica tra costo totale di un prodotto e l'importo dell'IVA;
- in statistica conoscendo la statura di una persona possiamo fare delle ipotesi più o meno sicure sul suo peso

La statistica studia quindi quelle relazioni che risultano più sfumate ed incerte

Indipendenza e connessione delle variabili

In statistica due caratteri si dicono **indipendenti** se la conoscenza delle modalità di uno dei due caratteri non ci permette di fare ipotesi sulle modalità del secondo.
Molto spesso due caratteri sono logicamente indipendenti e quindi ci aspettiamo che siano anche statisticamente indipendenti.

Ad esempio nella seguente tabella si nota che le righe presentano frequenze in proporzione.

UMG Dubium sapientiae initium
Lezione III

Tabella delle frequenze attese o teoriche

Se vi è indipendenza statistica allora la frequenza assoluta di ogni cella è uguale al prodotto del marginale di riga per quello di colonna diviso per il totale generale: es. la frequenza 8=20*24/60.

X \ Y	y1	y2	y3	Tot
x1	1	4	5	**10**
x2	2	8	10	**20**
x3	3	12	15	**30**
Tot.	**6**	**24**	**30**	**60**

In generale se indichiamo con ni. i marginali di riga e con n.j i marginali di colonna, in caso di indipendenza statistica dovrà verificarsi che
nij= (ni. x n.j)/ N

Tabella delle frequenze: considerazioni generali

1) Il totale di riga è uguale al totale di colonna;
2) La somma dei valori delle distribuzioni condizionate è uguale al totale generale di tabella.

X \ Y	y1	y2	y3	Tot
x1	Σ dei valori condizionati = 60			**10**
x2				**20**
x3				**30**
Tot.	**6**	**24**	**30**	**60**

Il file STAT1.xls mostra un esempio di calcolo del test su foglio excel.

UMG Dubium sapientiae initium
Lezione III

Tabella delle frequenze assolute

In generale una tabella che deriva da una osservazione di un fenomeno non presenta esattamente la situazione di indipendenza statistica, come ad esempio quella riportata all'inizio di questa lezione, anche nel caso di caratteri logicamente indipendenti (vi sono sempre fluttuazioni casuali ed errori).
Infatti le frequenze assolute sono diverse dalle frequenze attese o frequenze teoriche in caso di indipendenza.
Di seguito vengono riportate esempi di:

1) Tabelle di frequenza assolute
2) Tabelle di frequenza teoriche
3) Tabelle di contingenza

	a	b	c	n1
	d	e	f	n2
	g	h	i	n3
Tot	m1	m2	m3	N

Tab.di frequenza assoluta

	n1*m1/N	n1*m2/N	n1*m3/N	n1
	n2*m1/N	n2*m2/N	n2*m3/N	n2
	n3*m1/N	n3*m2/N	n3*m3/N	n3
Tot	m1	m2	m3	N

Tab.delle frequenze teoriche

UMG Dubium sapientiae initium
Lezione III

	a	b	c	n1
	d	e	f	n2
	g	h	i	n3
Tot	m1	m2	m3	N

Tab.di frequenza assoluta

	l	m	n	n1
	o	p	q	n2
	r	s	t	n3
Tot	m1	m2	m3	N

Tab.di frequenza teorica

	a-l	b-m	c-n	n1
	d-o	e-p	f-q	n2
	g-r	h-s	i-t	n3
Tot	m1	m2	m3	N

Tab.di contingenza

Misure di connessione e chi quadro

Per misurare la maggiore o minore dipendenza di due caratteri (mutabili) si utilizza un particolare indice che è chiamato indice χ2 (chi quadrato) calcolabile sia con il ricorso alle frequenze teoriche che alle frequenze osservate. Si propone il secondo metodo in quanto più semplice.

$$\chi^2 = n\left[\sum_{i=1}^{h}\sum_{j=1}^{k}\frac{n_{ij}^{\ 2}}{n_i.n._j}\right] \qquad \chi^2 = \frac{\chi^2}{\chi_{\max}^{\ 2}} = \frac{\chi^2}{n\min[h-1,k-1]}$$

Il file STAT1.xls mostra un esempio di calcolo del test su foglio excel.
Il Chi quadro viene calcolato da qualsiasi programma di elaborazione statistica.

Misure di tendenza centrale

Valori che si posizionano nel "mezzo" della distribuzione

∑	**indica la sommatoria di un insieme di valori**
x	**valore della variabile di interesse**
n	**dimensione del campione considerato**
N	**dimensione della Popolazione**

UMG Dubium sapientiae initium
Lezione IV

Media aritmetica

dato un campione di n elementi {x1, x2, ... xn} cioè un campione di dimensione (o numerosità) n, tratto da un universo rappresentato dalla variabile x, la media aritmetica è definita dall'espressione:

$$\bar{x} = \frac{\sum x}{n}$$

Facendo riferimento alle serie e seriazioni si può ottenere la media anche con la seguente espressione:

$$\bar{x} = \frac{\sum x\, f(x)}{\sum f(x)}$$

Dove *f(x)* rappresenta la frequenza assoluta o relativa della classe *x*.

Esempio di media aritmetica

Pulsazioni sotto sforzo (ciclette):

130 140 135 140 150 180 120 120 170 130 134 121 154 169 170 136 158 167 130 133 154 129 166 142

Media = 3478/24 = 144.92

UMG Dubium sapientiae initium
Lezione IV

Mediana

Si consideri un campione di valori di VES (*velocità di eritrosedimentazione*, mm/ora) misurati in 7 pazienti

{8, 5, 7, 6, 35, 5, 4}

In questo caso, la media (= 10 mm/ora) non è un valore tipico della distribuzione: soltanto un valore su 7 è superiore alla media! Conviene usare come indice del centro la mediana, definita come quel valore che divide a metà la distribuzione, sicché l'insieme dei valori è per metà minore e per metà maggiore della mediana.

Mediana

Quindi la **mediana** è il valore che occupa la posizione centrale dei dati una volta che questi siano stati ordinati in modo crescente.

Non è influenzata da valori estremi. Numero dispari di elementi:

$$Valore\ in\ posizione\ = \frac{n+1}{2}$$

Numero pari di elementi:

la media dei valori che occupano le posizioni (n/2) ed [(n/2)+1] nell'insieme ordinato dei numeri.

UMG Dubium sapientiae initium
Lezione IV

Esempio per un numero dispari di elementi:

6.72	**3.46**	**3.60**	**6.44**	**26.70**
3.46	**3.60**	**6.44**	**6.72**	**26.70**

(numero DISPARI di valori)

centro esato → **MEDIANA= 6.44**

Esempio per un numero pari di elementi:

6.72 3.46 3.60 6.44

3.46 3.60 6.44 6.72

(numero PARI di valori)

non c'è un centro esatto

$$\frac{3.60 + 6.44}{2}$$

MEDIANA= 5.02

Lezione IV

Moda

- E' il valore più frequente
  - Bimodale
  - Multimodale
  - No Moda
- E' l'unica misura di tendenza centrale che può essere usata con dati di tipo nominale

Esempio moda

a. 5 5 5 3 1 5 1 4 3 5	⇐	Moda = 5
b. 1 2 2 2 3 4 5 6 6 6 7 9	⇐	Bimodale - 2 e 6
c. 1 2 3 6 7 8 9 10	⇐	No Moda

Lezione IV

Si dice **media geometrica** l'antilogaritmo della media aritmetica dei logaritmi:

$$\overline{x}_g = \text{antilog}_{10}\left(\sum_{i=1}^{n} \log_{10}(x_i) \Big/ n\right)$$

Dalla definizione di logaritmo si ricava che la media geometrica di n valori si può calcolare come radice n-*esima* del loro prodotto:

$$\overline{x}_g = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

$\text{antilog}_{10}(2.398)=250.034$

dove la differenza è dovuta ad errori di arrotondamento.

Logaritmi e fenomeni biomedici

Una delle leggi fondamentali della fisiologia afferma che la risposta eccitatoria di un organismo ad uno stimolo è proporzionale al logaritmo dello stimolo:

Legge di Weber-Fechner: Risposta $\propto$ log(stimolo)

Tale legge è valida anche in altri ambiti, quali la farmacologia (l'effetto di un principio attivo è proporzionale non alla sua dose ma al logaritmo della dose), la microbiologia, l'enzimologia e l'immunologia.

UMG Dubium sapientiae initium
Lezione IV

Distribuzione dei dati

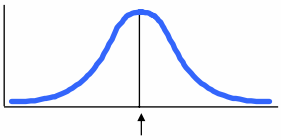
Simmetrica

I dati sono distribuiti in modo simmetrico se la parte sinistra e destra dell'istogramma sono pressoché speculari

Asimmetrica

Se la distribuzione non è simmetrica, e si estende di più in una direzione

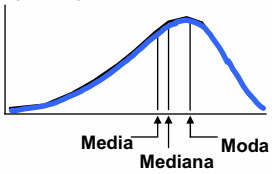
Distribuzione dei dati



Moda (Mo) = Media (Ma) = Mediana (Me)

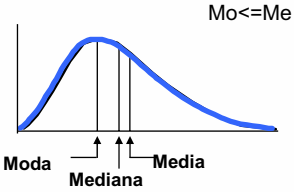
SIMMETRICA

Ma<=Me <=Mo



ASIMMETRICA A SINISTRA
(negativa)

Mo<=Me <=Ma



ASIMMETRICA A DESTRA
(positiva)

Lezione IV

Misure di variazione (dispersione)

La media (o la mediana), di per sè, non dà informazioni sulla **dispersione** dei valori di un insieme di dati.

Esempio:

Gli insiemi di valori di VES

{A}: { 8, 5, 7, 6, 35, 5, 4}

{B}: { 11, 8, 10, 9, 17, 8, 7}

hanno la stessa media (= 10), ma in {A} i valori sono più dispersi che in {B}:

in {A} i valori sono inclusi tra 4 e 35

in {B} i valori sono inclusi tra 7 e 17

Range

$X_{\max} - X_{\min}$ valore più alto – valore più basso

il **range** di {A} è $R_A = 35 - 4 = 31$
il **range** di {B} è $R_B = 17 - 7 = 10$

Il **range** è il più ***intuitivo*** fra gli indici di dispersione, ha però l'inconveniente di basarsi solo sui due valori estremi, nei quali più evidentemente si manifesta la variabilità di campionamento e l'errore di misura.

UMG Dubium sapientiae initium
Lezione IV

Varianza

Misura la dispersione dei valori intorno alla media, ovvero definisce lo “scarto quadratico medio”

Simbologia $\begin{cases} s^2 \\ \sigma^2 \end{cases}$

Varianza in simboli

$$\sigma^2 = \frac{\sum (x-\mu)^2}{\mathrm{N}}$$ **Varianza della popolazione**

$$s^2 = \frac{\sum (x-\bar{x})^2}{n-1}$$ **Varianza campionaria**

UMG Dubium sapientiae initium
Lezione IV

Deviazione standard

E' la RADICE QUADRATA della Varianza
Ha la stessa unità di misura della media

$$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}$$ **DevStandard popolazione**

$$\sigma = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$$ **DevStandard campionaria**

Coefficiente di variazione

Il **coefficiente di variazione** non ha dimensione: è un indice di variabilità relativa, utilizzabile per confrontare la dispersione di variabili con differenti unità di misura.

$$CV\% = 100 \times \frac{s}{\overline{x}}$$

Lezione IV

Nell'***esempio*** dei due insiemi di valori di VES si ha:

{A}: $D = 8^2 + 5^2 + \ldots 4^2 - (8 + 5 + \ldots 4)^2 / 7 = 1440 - 700 = 740$

$s^2 = 740 / 6 = 123.33$

$s = = 11.1$ $\pm s = (-1.1, 21.1)$

$CV\% = 100(11.1/10) = 111\%$

{B}: $D = 11^2 + 8^2 + \ldots 7^2 - (11 + 8 + \ldots 7)^2 / 7 = 768 - 700 = 68$

$s^2 = 68 / 6 = 11.33$

$s = = 3.4$ $\pm s = (6.6, 13.4)$

$CV\% = 100(3.4/10) = 34\%$

Misure di posizione

Valore di Z (z score)

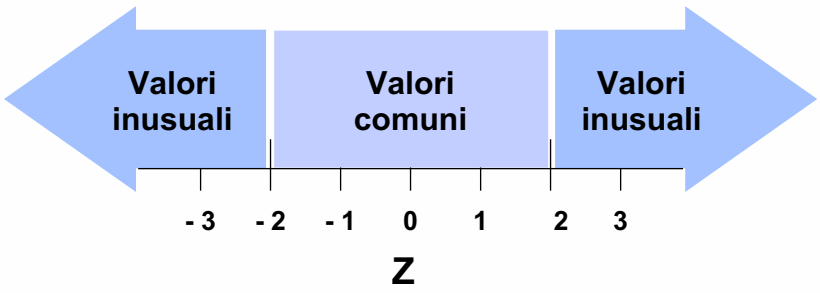
Quanto un dato valore x si discosta dalla media, misurato in unità di deviazioni standard

$$z = \frac{x - \mu}{\sigma}$$ popolazione

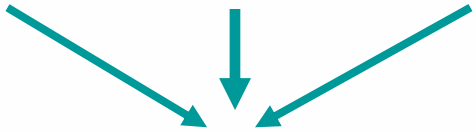
$$z = \frac{x - \bar{x}}{s}$$ campione

UMG Dubium sapientiae initium
Lezione IV

Interpretazione di z



Quartili, Decili, Percentili



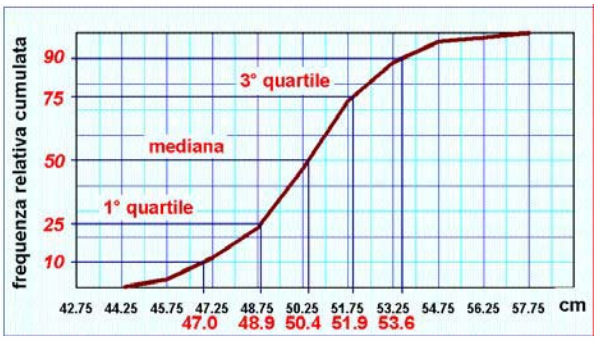
Frattili

Una distribuzione può essere descritta per mezzo dei suoi **frattili**. Si dice frattile p-*esimo* di una distribuzione quel valore x_p tale che la frequenza relativa cumulata $F(x_p) = p$

Lezione IV

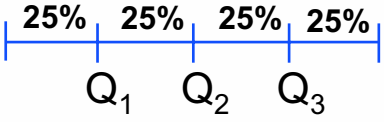
Frattili

Nei **grafici cumulati**, i valori riportati sull'asse verticale indicano la **frequenza** delle rilevazioni con **valore pari o minore** dei valori in corrispondenza sull'asse orizzontale.



Quartili

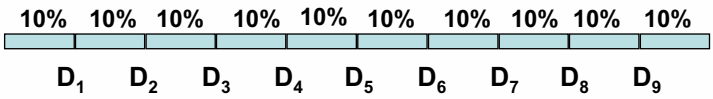
Q_1, Q_2, Q_3
dividono la distribuzione in quattro porzioni ad uguale numerosità



Lezione IV

Decili

D1, D2, D3, D4, D5, D6, D7, D8, D9
dividono la distribuzione in quattro porzioni ad ugual numerosità



I 5 valori di sintesi e l'analisi della varianza

Per farsi un'dea delle principali caratteristiche della distribuzione di un carattere quantitativo X si effettua l'analisi della varianza (ANOVA) partendo dai cinque indici sopra indicati:

$X_{\min}$	1) la modalità più piccola
Q_1	2) il primo quartile
Me	3) la mediana
Q_3	4) il terzo quartile
$X_{\max}$	5) la modalità più grande

UMG Dubium sapientiae initium
Lezione IV

I 5 valori di sintesi

Sulla base di questi valori si possono ricavare informazioni sulla **posizione della distribuzione** attraverso:

Me	la mediana
$MiQ = \frac{Q_1 + Q_3}{2}$	la media interquartile
$MR = \frac{x_{\max} + x_{\min}}{2}$	il midrange

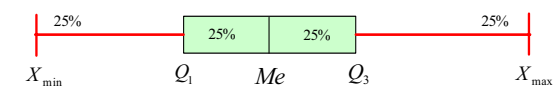
e sulla **variabilità** attraverso:

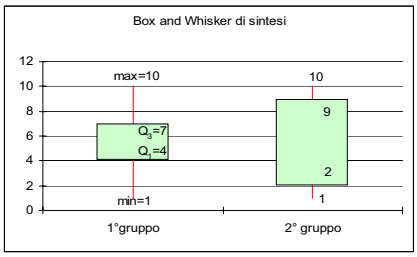
$Q_3 - Q_1$	range interquartile
$X_{\max} - X_{\min}$	range

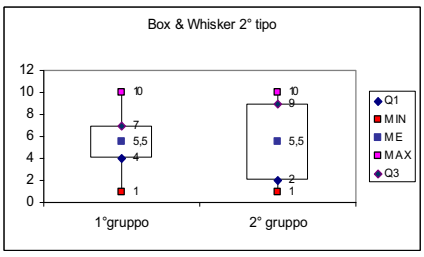
Il "diagramma a scatola e baffi" o "Box-plot" o "Box and whisker"

E' una rappresentazione grafica che utilizza i 5 numeri di sintesi per il confronto tra due o più collettivi. I quartili vengono rappresentati rispetto ad un conveniente asse, che diviene una scatola tra Q1 e Q3 ed all'interno della quale viene disegnata una linea verticale in corrispondenza della mediana.
Alle due estremità del grafico si pongono gli estremi della distribuzione, cioè il valore più piccolo a sinistra (contiene il 25% delle osservazioni) e quello più grande a destra (contiene il 25% delle osservazioni). I due valori vengono uniti alla scatola attraverso una linea (baffo).
La scatola contiene il 50% delle osservazioni centrali della distribuzione (il 25% a destra e il 25% a sinistra della linea mediana.

Lezione IV







Lo studio della varianza è proposto nel file di esercitazione STAT2.xls

Errori sistematici (δ)

Gli **errori sistematici** si manifestano nella tendenza deterministica di un dato metodo a **sovrastimare** (o **sottostimare**) il vero valore θ . Pertanto, l'universo delle misure che si possono virtualmente ottenere quando con tale metodo si misura ha media μ che differisce dal valore (δ = μ - θ).

Gli errori sistematici hanno cause ben determinate, inerenti o al **metodo** (**es.:** scarsa selettività del reagente usato per la titolazione di un certo soluto), o alle **condizioni di esecuzione** del procedimento analitico (**es.:** strumento non calibrato correttamente

Una misura è tanto più accurata quanto minore è l'entità dell'errore sistematico (δ) da cui è affetta.

UMG Dubium sapientiae initium
Lezione IV

Errori casuali (ε)

Misurazioni dello stesso valore , ripetute in uno stesso procedimento analitico, e in condizioni il più possibile simili, portano spesso a misure differenti: non è possibile ripetere la misurazione in modo del tutto identico.

La somma di tutte le **piccole e imprevedibili** variazioni nell'esecuzione delle varie operazioni analitiche fa sì che le misure fluttuino attorno a un valore μ, che si scosta più o meno dal valore , a seconda dell'entità dell'errore sistematico. Tali **fluttuazioni** attorno a μ (ε = x - μ) sono dette **errori casuali**.

Una misura è tanto più precisa quanto minore è l'entità dell'errore casuale (ε) da cui è affetta.

In conclusione

Per riassumere, l'errore totale di una misura esente da errori grossolani può essere espresso come **somma** di una componente **sistematica** e di una componente **casuale**.

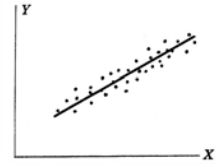
errore totale		errore sistematico		errore casuale
$(x-\vartheta)$	=	$(\mu-\vartheta)$	+	$(x-\mu)$
η	=	δ	+	ε
attendibilità		accuratezza		Precisione

Lezione IV

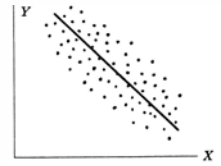
Relazione tra variabili

Spesso si vuole trovare la relazione che lega due o più variabili (es. la pressione di un gas dipende da temperatura e volume).

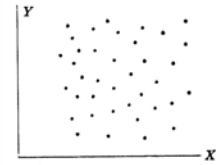
Date due variabili X e Y costruiamo un diagramma di dispersione con i loro valori.
Se tutti i punti giacciono più o meno su una retta, la correlazione è detta lineare e la relazione fra le variabili sarà retta da un'equazione lineare



Se Y cresce al crescere di X la correlazione è *positiva o diretta*:



Se Y decresce al crescere di X, la correlazione è detta negativa o inversa:



Se non c'è relazione fra le variabili diciamo che sono *incorrelate*:

UMG Dubium sapientiae initium
Lezione IV

Correlazione

Per sapere se esiste un 'legame' tra due caratteri quantitativi, e cioè se uno di essi esercita un'influenza sull'altro, ad esempio il peso delle persone e la loro altezza, si utilizzano gli indici di **correlazione**, i quali danno anche una misura di questo 'legame'. Quando la dipendenza tra due variabili è lineare si parla di correlazione lineare.
L'indice usato è detto indice di correlazione di Bravais-Person.

$$r = \frac{\sum (x - M_x)(y - M_y)}{n\sigma_x\sigma_y} \qquad -1 \le r \le 1$$

x, y sono le serie dei dati
Mx, My sono le medie aritmetiche rispettivamente di x e y
n è il numero totale dei dati
SIGMAx e **SIGMAy** rispettivamente la dev. standard delle x e delle y.

per r= 1 si ha il massimo di correlazione diretta
per r= -1 si ha il massimo di correlazione inversa
per r= 0 non si ha correlazione

UMG Dubium sapientiae initium
Lezione IV

Correlazione diretta e inversa

La correlazione si dice **diretta** se ai valori crescenti di una variabile corrispondono valori pure crescenti dell'altra variabile, ad esempio reddito e consumi, altezza e peso.
La correlazione si dice **inversa** se ai valori crescenti di una variabile corrispondono valori decrescenti dell'altra variabile, ad esempio altitudine e pressione atmosferica

Esempio nel file STAT2.Xls

Concetto di probabilità

La probabilità rappresenta il modello interpretativo per la valutazione di fenomeni deterministici e/o sperimentali corrispondenti alle modalità tipolologiche casuale e/o aleatoria.

Un esperimento, ad esempio, può dar luogo ad un risultato, fra un certo numero di risultati possibili, di esito ignoto o non determinabile a priori in modo univoco.

UMG Dubium sapientiae initium
Lezione V

Concetto di probabilità

Esempio:

1) Osservare con quale accelerazione cade una mela equivale alla valutazione di un fenomeno deterministico di tipi casuale.
2) Osservare “testa” nel lancio di una moneta equivale alla valutazione di un fenomeno sperimentale di tipo aleatorio aleatorio

Alcune definizioni

Esperimento aleatorio: lancio di uma moneta di dadi, ecc.

Spazio degli eventi Ω (o spazio campione): l'insieme di tutti i possibili esiti del nostro esperimento

Evento (casuale-aleatorio): possibile esito dell'esperimento nel caso di un evento di tipo aleatorio

UMG Dubium sapientiae initium
Lezione V

Definizioni del concetto di probabilità

Definizione empirica: E' il valore costante intorno al quale tende a stabilizzarsi la frequenza relativa di un evento al crescere del numero delle prove di un dato esperimento.

Definizione classica: La probabilità che un evento accada è data dal rapporto tra il numero dei casi favorevoli e il numero di casi possibili (Laplace 1749-1827).

$$P(E) = \frac{n.casi.favorevoli}{n.casi.possibili}$$

$$p = \frac{r}{k}$$

p= probabilità che un evento accada P(E)
r= casi favorevoli
k= casi possibili
q= probabilità che un evento non accada
E= Eventi possibili

Definizione Classica (*continuazione*)

La probabilità che un evento si verifichi è indicata con: $0 \leq p \leq 1$

La probabilità che un evento non si manifesti è indicata con: $q = 1 - p$

La probabilità che un evento è certo è indicata con: $p + q = 1$

La probabilità che si verifichi l'evento impossibile è indicata con: $p + q = 0$

Presupposti:

Gli eventi possibili sono tutti tra loro mutuamente esclusivi. Potrà accadere uno solo degli eventi. Vengono considerati solo gli eventi utili allo scopo prefissato. Tutti gli eventi sono equiprobabili e la probabilità teorica è conoscibile a priori.

UMG Dubium sapientiae initium
Lezione 5

Definizione Classica (esempi)

P(faccia di un dato)= 1/6

P(una carta di un mazzo di 40 carte)= 1/40

P(faccia di una moneta)= 1/2

Es.: la probabilità di ottenere un numero superiore o uguale a 5 lanciando un dado. Poiché gli eventi utili sono due (5 e 6) la probabilità sarà P(5,6)= 2/6 = 1/3 = 0,33
La probabilità di estrarre un asso da un mazzo di 40 carte: poiché ci sono 4 assi all'interno del mazzo, P(asso)= 4/40 = 1/10 = 0,1 .

Definizione frequentista

La probabilità di un evento è stabilita dal rapporto tra la frequenza con cui questo evento è comparso e il numero di prove effettuate.
Se indichiamo con fn(E) la frequenza relativa con cui l'evento E si è verificato in una serie di n prove effettuate tutte nelle stesse condizioni, allora:

$$P(E) = \lim_{n \to \infty} fn(E)$$

Due eventi A e B si dicono **incompatibili** se non possono verificarsi contemporaneamente.
Due eventi A e B si dicono **necessari** (o colletivamente esaustivi) se almeno uno di loro si verifica certamente.

UMG Dubium sapientiae initium
Lezione V

Considerazioni

La situazione di equiprobabilità è valida solo in certe particolari condizioni, spesso fittizie come quelle dei giochi d'azzardo, mentre, nella realtà, gli eventi non presentano questa caratteristica.
Ad esempio, avere capelli neri è equiprobabile ad avere capelli rossi? Se ciò fosse vero dovremmo quotidianamente osservare lo stesso numero di persone con capelli neri o rossi mentre l'esperienza quotidiana ci dice che questo non è vero.
La probabilità di avere capelli rossi è sicuramente inferiore (o non uguale) a quella di avere capelli neri.
Per calcolare esattamente la probabilità dei due eventi dovremmo conoscere esattamente, in modo univoco e definitivo, quali sono i fattori che producono l'apparire del fenomeno "capelli rossi" o di quello "capelli neri".

Probabilità totale

Se l'evento A e l'evento B si escludono a vicenda (incompatibili), la probabilità di ottenere A o B P(A o B) è uguale alla somma della probabilità di A più la probabilità di B in simboli: $P(AoB) = P(A) + P(B)$
In un mazzo di 52 carte sia che P(asso o re) = P(asso) + P(re) = 1/13 + 1/13 = 2/13
Tale principio può essere esteso anche a due eventi che non si escludono a vicenda.

$$P(AoB) = P(A) + P(B) - P(AeB)$$

dove P(A e B) rappresenta la probabilità di ottenere, uno dopo, l'altro sia A che B.
P(donne o piche) = P(donna) + P(picche) - P(donne e picche) = 4/52 + 13/52 - 1/52 = 16/52 = 4/13

UMG Dubium sapientiae initium
Lezione V

Probabilità composta

La probabilità che due eventi A e B accadono l'uno dopo l'altro è detta probabilità composta.
Se l'esito di un esperimento (evento A) è indipendente dall'esito del precedente esperimento (evento B), gli eventi A e B sono detti indipendenti.
La probabilità composta di due **eventi indipendenti** A e B è data da: P(A e B) = P(a) x P(b)
Ad es. la probabilità di ottenere croce in due lanci successivi si ottiene moltiplicando le probabilità di avere croce nei singoli lanci:
P(croce e croce)= P(croce)xP(croce) = 1/2x1/2 = 1/4

Probabilità composta condizionata

Quando il manifestarsi di un evento B influisce in qualche modo sul manifestarsi dell'evento A, si dice che gli eventi consecutivi A e B sono **dipendenti**. In questo caso la probabilità che accada A se B si è verificato è diversa dalla probabilità che accada A se B non si è verificato.
La probabilità di A, posto che B si sia verificato, indicata col simbolo P(A|B), si chiama probabilità condizionata (dal fatto che B sia accaduto). Il simbolo "|" denota **condizione**.
La probabilità composta di due eventi consecutivi dipendenti A e B è allora la probabilità di ottenere uno di questi eventi moltiplicata la probabilità condizionata di ottenere l'altro, posto che il primo si sia verificato. In simboli:

$$P(A \text{ e } B) = P(A) \times P(B|A)$$

UMG Dubium sapientiae initium
Lezione V

Legge dei grandi numeri

Per $N_{tot} \rightarrow \infty$ la frequenza tende alla probabilità (a priori).

In effetti il rapporto di probabilità è il limite teorico raggiungibile solo con un numero infinito di prove.

Elementi di calcolo combinatorio

n fattoriale: viene indicato con il simbolo n! (n fattoriale) che sintetizza una serie di prodotti. In simboli:

$$n! = n(n-1)\cdot(n-2)\cdot\ldots\cdot(1)$$

Es.: 7! = 7(7-1)(7-2)(7-3)(7-4)(7-5)(7-6)(1) = 7x6x5x4x3x2x1 = 5040

Per convenzione 0! = 1

Tab. fattoriali

0! = 1
1! = 1
2! = 2
3! = 6
4! = 24
5! = 120
6! = 720
7! = 5040

UMG Dubium sapientiae initium
Lezione V

Combinazioni

Le combinazioni di n oggetti diversi presi r alla volta sono i gruppi di r elementi che si possono formare con gli n elementi di partenza in modo che ciascun gruppo sia diverso dagli altri almeno per un elemento.
Il numero di combinazioni di n oggetti diversi presi r alla volta C(n,r)

$$C(n,r) = \frac{n(n-1)\cdot\ldots\cdot(n-r+1)}{r!} = \frac{n!}{r!(n-r)!}$$

Qualche esempio

Calcoliamo il numero di combinazioni di un tris di assi (in un mazzo ce ne sono 4): C(4,3) = (4x3x2x1) / (3x2x1)x(4 - 3) = 24 / 6 = 4

Calcolo alternativo: $$C(4,3) = \frac{4 \cdot 3 \cdot 2}{1 \cdot 2 \cdot 3} = \frac{24}{6} = 4$$

Calcoliamo quante cinquine si possono formare con i 90 numeri del lotto:

$$C(90,5) = \frac{90 \cdot 89 \cdot 88 \cdot 87 \cdot 86}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5} = 43949268$$

UMG Dubium sapientiae initium
Lezione V

Teorema di Bayes

Dividendo l'equazione della probabilità composta condizionata per P(A):

$$P(B/A) = \frac{P(A/B) \cdot P(B)}{P(A)}$$

per k eventi reciprocamente incompatibili e collettivamente esaustivi, e per B1, B2, ..., B3 eventi mutuamente esclusivi, si ottiene il Teorema di Bayes:

$$P(B_i/A) = \frac{P(A/B_i) \cdot P(B_i)}{P(A/B_1)P(B_1) + (PA/B_2)P(B_2) + \ldots + P(A/B_K)P(B_K)}$$

Dove:
P(Bi) è la probabilità a priori che è attribuita alla popolazione Bi prima che siano conosciuti i dati,
P(A/Bi) rappresenta la probabilità aggiuntiva dopo che è stata misurata la probabilità di A.

Proviamo adesso a pensare agli eventi Ei come le cause che determinano l'evento A. Allora, se si è verificato A, con quale probabilità la causa è Ei? In altre parole si vuole conoscere la probabilità P(Ei|A):

$$P(E_i \mid A) = \frac{P(E_i) * P(A \mid E_i)}{\sum_{i=1}^{n} P(E_i) * P(A \mid E_i)}$$

P(Ei) = Probabilità a priori (non dipendono dal esito A)

P(A| Ei) = Verosimiglianza (con quale probabilità Ei determina A)

P(Ei|A) = Probabilità a posteriori (verificatosi A, con quale probabilità Ei si verifica)

UMG Dubium sapientiae initium
Lezione V

Teorema di Bayes: esempio

Un laboratorio ha messo a punto un alcool-test in base al quale il 2% delle persone controllate dalla polizia è risultato essere in stato d'ebbrezza. In base all'esperienza si è constatato, inoltre, che il 95% dei casi di alcool-test ha dato esito positivo in caso di reale ebbrezza, mentre nel 96% dei casi, ha dato esito negativo in caso di persone sobrie. Quale è la probabilità che una persona sia realmente ebbra, in caso di esito positivo del test?

E= evento "ubriaco" NE = evento "non ubriaco" A = evento "test positivo" B = evento "test negativo"

P(E) = 0.02 P(NE) = 1-P(E) = 0.98

P(A|E) = 0.95 P(B|E) = 1-P(A|E) = 0.05

P(B|NE) = 0.96 P(A|NE)= 1-P(B|NE) = 0.04

Risulterà: $$P(E \mid A) = \frac{P(A \mid E) * P(E)}{P(A \mid E) * P(E) + P(A \mid NE) * P(NE)}$$

ovvero: $$P(E \mid A) = \frac{0.95 * 0.02}{0.95 * 0.02 + 0.04 * 0.98} \cong 0.33$$

Non molto buono! Se aumentassi P(B|NE) = 0.99?

$$P(E \mid A) = \frac{0.95 * 0.02}{0.95 * 0.02 + 0.01 * 0.98} \cong 0.66$$

Decisamente meglio!

Distribuzioni di probabilità

E' noto che una variabile statistica può assumere diverse modalità e, alla luce della teoria della probabilità, ciascuna modalità ha una certa probabilità di manifestarsi.

In una distribuzione, come ad ogni modalità si può associare la sua frequenza allo stesso modo si può associare la sua probabilità.

A seconda che si manifestino modalità discrete o continue si hanno distribuzioni di probabilità discrete o continue.

UMG Dubium sapientiae initium
Lezione VI

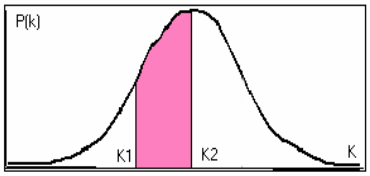
Distribuzioni di probabilità

Quando la variabile K assume una serie continua di valori (es. peso dei neonati) è chiamata variabile casuale continua e P(K) è la funzione di densità di probabilità. La rappresentazione grafica di una distribuzione di probabilità continua è una curva cioè una funzione continua la cui equazione è:

Lezione VI

$$Y = P(K).$$

Distribuzioni di probabilità



L'area compresa tra la curva e l'asse delle ascisse è uguale a 1 e l'area sotto la curva compresa tra le perpendicolari all'asse Y, K=K1 e K=K2 (ombreggiata) rappresenta la probabilità che K assuma valori compresi tra K1 e K2.
P[K1<K<K2].

Lezione VI

Distribuzioni di probabilità

La statistica è un'estensione del calcolo delle probabilità con l'introduzione di nuove variabili (variate):

- la probabilità viene fatta passare da un numero razionale ... ad un numero reale;

- può essere infinitesima anche se poi si darà significato sempre alla probabilità finita tramite integrazioni;

- si suppongono valide tutte le leggi delle probabilità già stabilite;

- non si può più definire la probabilità come rapporto fra casi favorevoli e casi possibili

Distribuzioni di probabilità

Nello studio della popolazione individuate le variabili e la loro relazione funzionale, si possono ottenere in un piano cartesiano poligoni o curve di distribuzione a seconda che si tratti di dati discreti o continui. Il fenomeno non sempre è subito interpretabile come una serie di punti che raffigurino una retta, un poligono o una curva. Più spesso le coordinate costituiscono una "nube di punti" che, a seconda della loro densità, posso rappresentare una delle raffigurazioni innanzi citata.

Il tipo di grafico che meglio interpreta tale situazione è quello "dispersione" che generalmente consente anche la raffigurazione della curva di tendenza (interpolante).

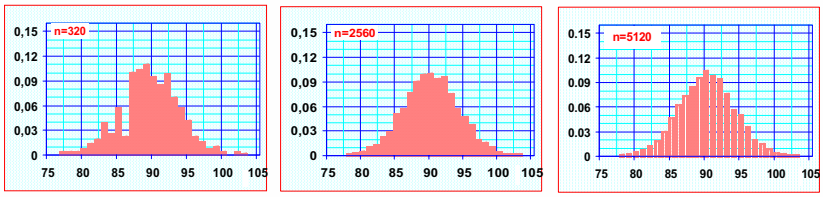
UMG Dubium sapientiae initium
Lezione VI

Distribuzioni di probabilità

In ogni caso, individuata la curva di frequenza (distribuzione di frequenza teorica) che meglio rappresenta il fenomeno, possiamo ipotizzare che le nostre variabili, all'aumentare della dimensione del campione (per n tendente all'infinito), si avvicinano sempre più ad essa. In tal modo si possono usare le relazioni già studiate nella statistica descrittiva alla luce della teoria campionaria e probabilistica dei tests statistici.

Distribuzioni degli errori di misura

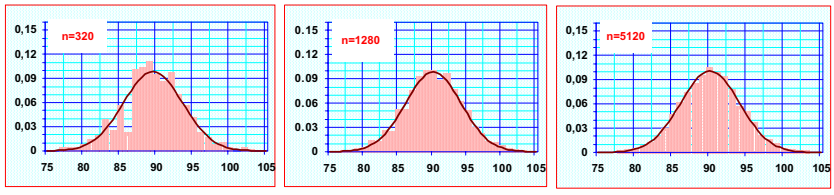
Si supponga di eseguire, in condizioni assai simili e con lo stesso metodo analitico, un **gran numero** di titolazioni di una soluzione di glucosio avente concentrazione θ=90 mg/dl, e di riportare in grafico le **frequenze relative** dei valori ottenuti (x) con le prime 20, 40, ... 5120 misure.



UMG Dubium sapientiae initium
Lezione VI

Forma della distribuzione

All'aumentare del numero di misure, i valori tendono ad accentrarsi attorno alla loro media e l'istogramma assume una forma ***a campana*** sempre più regolare, che può essere approssimata con una funzione reale nota come **funzione di gauss** o **funzione normale**.



Comportamento degli errori

Gli errori casuali di misura ($\varepsilon = x - \mu$), considerati nel loro complesso, mostrano un comportamento tipico che può essere così descritto:

- Gli ***errori piccoli*** sono più frequenti di quelli ***grandi***;
- Gli errori di ***segno negativo*** tendono a manifestarsi con la stessa frequenza di quelli con segno positivo;
- All'aumentare del numero delle misure si ha che ~2/3 dei valori tendono ad essere inclusi nell'intervallo **media ± 1 deviazione standard**
- Il **95%** ~ dei valori tende ad essere incluso nell'intervallo **media ± 2 deviazioni standard**

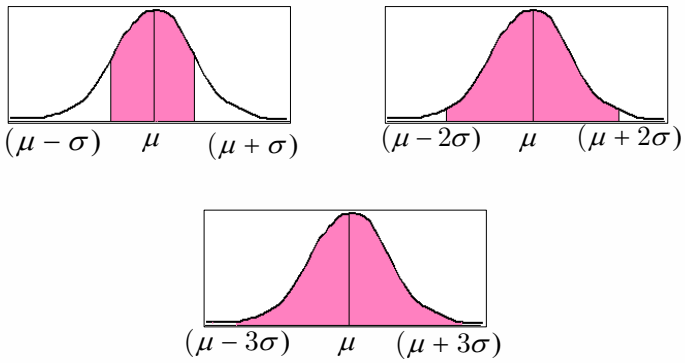
UMG Dubium sapientiae initium
Lezione VI

Distribuzioni di probabilità

Teorema di Chebyshev: per ciascun gruppo di dati (popolazione o campione) e qualsiasi costante k maggiore di 1, la proporzione dei dati che deve giacere nell'intervallo di k deviazioni standard a destra e a sinistra della loro media è almeno di $1 - \frac{1}{k^2}$ deviazioni standard.

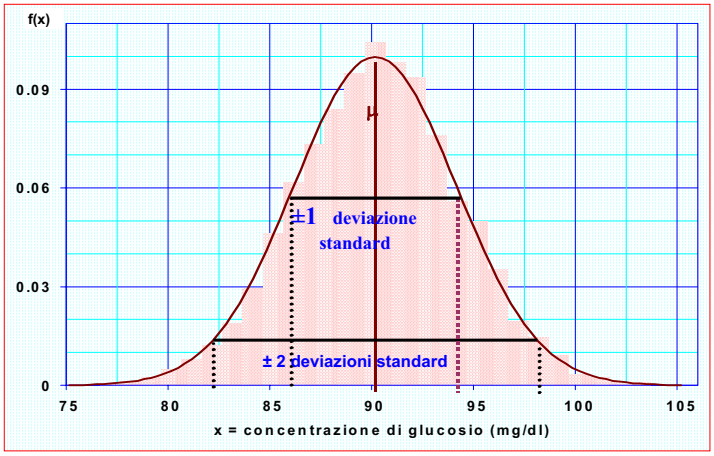
- il 68,27% dei casi è compreso tra $(\mu - \sigma)$ e $(\mu + \sigma)$
- il 95,45% dei casi è compreso tra $(\mu - 2\sigma)$ e $(\mu + 2\sigma)$
- il 99,73% dei casi è compreso tra $(\mu - 3\sigma)$ e $(\mu + 3\sigma)$

Distribuzioni di probabilità

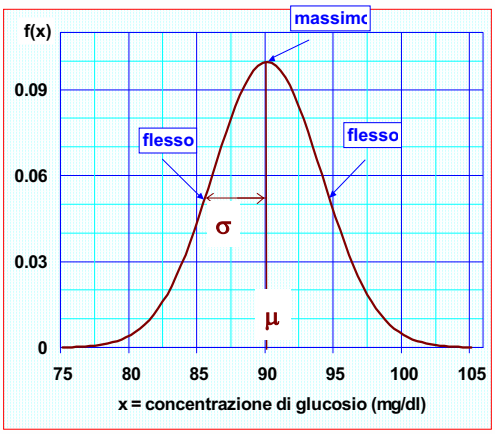


UMG Dubium sapientiae initium
Lezione VI

La funzione di Gauss



La funzione di Gauss



$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

dove:

σ è la deviazione standard della totalità delle misure;
μ è la media della totalità delle misure;
e = *base dei logaritmi naturali* (*e* = 2.71828...).
π è il rapporto tra circonferenza e diametro π = 3.14159...);

Lezione VI

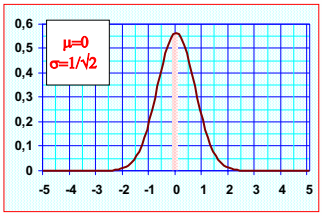
La funzione di Gauss Standard

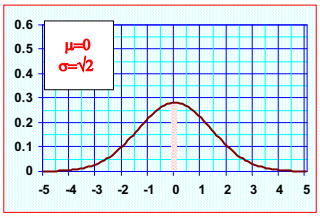
Si può trasformare una generica funzione gaussiana f(x) con media μ e varianza σ^2, in una **funzione gaussiana standard** ϕ(z) con media 0 varianza 1, se si pone :

$$z = \frac{(x-\mu)}{\sigma}$$

… z è detta “deviata gaussiana standard”

Forma della gaussiana

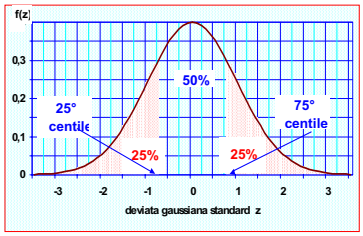


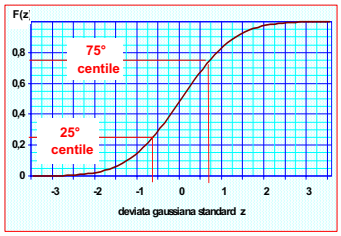


Tutte le gaussiane hanno la stessa identica forma, benché quelle con deviazione standard maggiore siano più larghe e più basse di quelle con deviazione standard minore.

UMG Dubium sapientiae initium
Lezione VI

Distribuzioni di probabilità





☜L'area sottesa alla funzione di Gauss, da -∞ ad un dato valore z=z*, indica la *frequenza relativa* dei valori $z \leq z^*$.

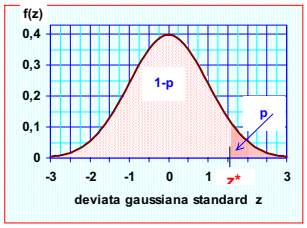
z* = coniugato di z speculare rispetto all'asse reale

☜Tale area è data dall'integrale di f(z) *definito* tra -∞ e z*:

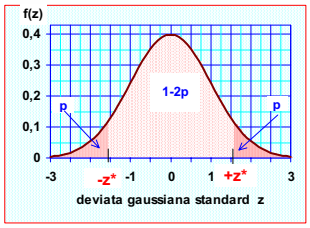
$$F(z^*) = \int_{-\infty}^{z^*} \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2} z^2\right) dz$$

F(z*) rappresenta la *distribuzione cumulativa* di f(z).

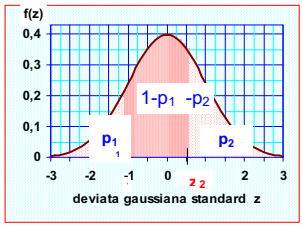
Distribuzioni di probabilità



☝ Detto p (0<p<1) il valore dell'area a destra di +z*, l'area a sinistra di +z* vale (1-p).



☝ L'area a sinistra di -z* è uguale all'area a destra di +z*. Detto p (0<p<1) il valore di tale area, l'area esterna a z* vale 2p, e l'area interna vale (1-2p).

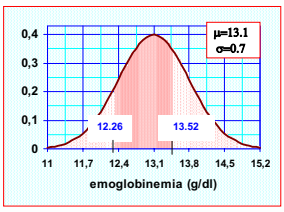


☝L'area compresa tra due valori $z_1^* < z_2^*$ si ricava per differenza $(1-p_1-p_2)$, dove p_1 è il valore dell'area a sinistra di z_1^*, e p_2 quello dell'area a destra di z_2^*.

UMG Dubium sapientiae initium
Lezione VI

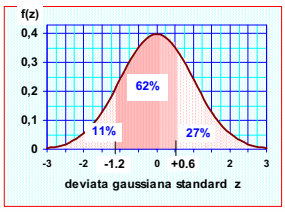
Applicazione della gaussiana standard

In una popolazione di ragazze di età inclusa tra i 18 e i 25 anni, la concentrazione di emoglobina nel sangue (x) **approssima** la **distribuzione gaussiana** con media μ=13.1 g/dl e deviazione standard σ=0.7 g/dl. In base a queste sole informazioni possiamo calcolare, ***ad esempio***, quante ragazze hanno emoglobinemia inclusa tra 12.26 e 13.52 g/dl. Infatti:



$$z_1 = \frac{(12.26 - 13.10)}{0.7} = -1.2$$

$$z_2 = \frac{(13.52 - 13.10)}{0.7} = +0.6$$



Distribuzione dell'emoglobina in una popolazione di ragazze di età compresa tra i 18 e i 25 anni.Nell'11% delle ragazze i valori di Hb ☝ sono minori di 12.26 g/dl, e nel 27% sono maggiori di 13.52 g/dl. Quindi il 62% delle ragazze ha valori di Hb compresi tra 12.26 e 13.52 g/dl.

I gradi di libertà (GdL)

	A1	A2	A3
B1	60	53	12
B2	53	23	16
B3	55	48	20

Dovendo suddividere i 168 eventi contenuti nella categoria A1 nelle 3 celle corrispondenti alle categorie di B, noi abbiamo libertà disporre quanti eventi vogliamo in 2 sole celle, la terza è costretta a contenere gli eventi restanti.

Lo stesso ragionamento viene fatto per a2, a3 e per ciascuno dei valori di B.

Nelle tabelle vi sono delle celle (per convenzione le ultime) che non possono contenere qualsiasi numero ma solo quanto resta per poter sommare al totale gli eventi di quella categoria.

$$GdL = (r - 1) \times (c - 1)$$

UMG Dubium sapientiae initium
Lezione VI

Il parametro t

La media della distribuzione campionaria coincide con la media della popolazione, mentre lo scarto quadratico medio vale:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$

Per N>30 la distribuzione campionaria della media è approssimabile alla distribuzione normale.

Una buona stima dell'errore standard vero è l'errore standard del campione

$$s_m = \sqrt{\frac{s^2}{N}} = \frac{s}{\sqrt{N}}$$

Il parametro t

Utilizzando l'errore standard campionario il parametro z viene modificato:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Sostituendo la stima del parametro della varianza della popolazione con quello della varianza campionaria si ottiene i parametro t consistente nel rapporto:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

UMG Dubium sapientiae initium
Lezione VI

Il parametro t

Caratteristica importante del parametro t è che non è distribuito normalmente. La sua distribuzione sarà più dispersa di quella di z.

Essa è stata calcolata dal matematico inglese Gosset, che la pubblicò sotto lo pseudonimo di Student.

Si tratta di una famiglia di distribuzioni, a seconda del numero di gradi di libertà, che vale:

GdL = N-1

dove N è il numero di osservazioni del campione.

Il parametro t

I valori della famiglia di distribuzioni t sono tabulati.

Per campioni molto grandi, il valore di s_m oscilla poco intorno al suo valore medio, che è σ_m.

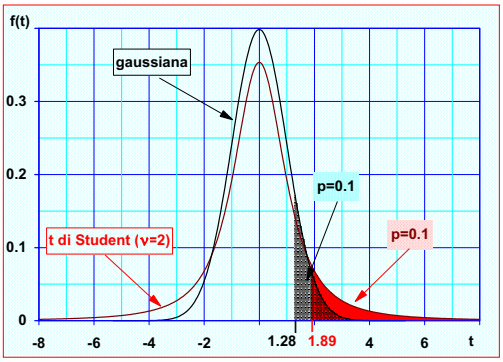
Quindi per valori molto grandi la distribuzione t si avvicina molto a quella di z, ed arriva a coincidere per infiniti gradi di libertà.

Per piccoli campioni (N<30) le differenze sono notevoli, data l'oscillazione casuale di s_m intorno a σ_m.

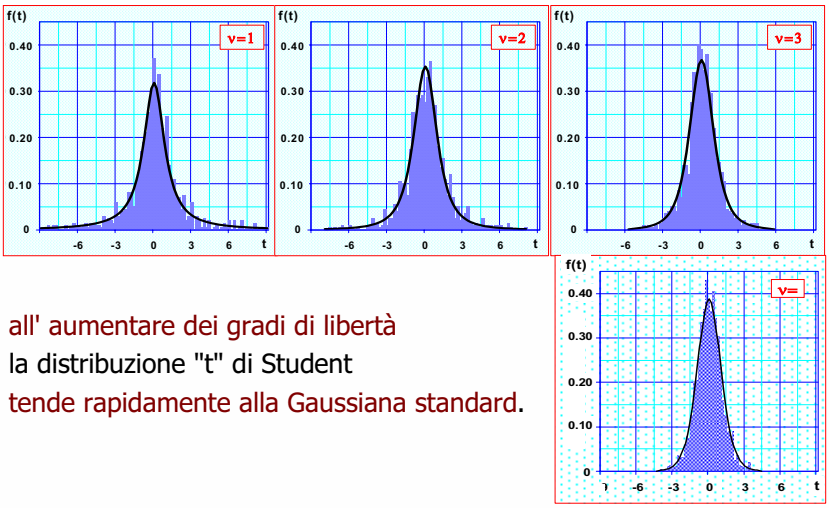
UMG Dubium sapientiae initium
Lezione VI

La distribuzione t di Student

Questa ha ***code più alte, fianchi più stretti e varianza maggiore*** rispetto alla Gaussiana standard:



$$\frac{\bar{x} - \mu}{s / \sqrt{n}}$$ ~ t di Student (con ν=n-1 g.d.l.)



all' aumentare dei gradi di libertà
la distribuzione "t" di Student
tende rapidamente alla Gaussiana standard.

Lezione VI

La distribuzione t di Student

A causa della sua forma, la distribuzione "t" di Student ha **percentili con valore assoluto tanto più elevato** rispetto a quello dei corrispondenti percentili della Gaussiana **quanto minore è il numero di gradi di libertà.**

Ad esempio, il 90° percentile della gaussiana standard è 1.282, mentre i corrispondenti percentili delle "t" di Student con 1, 2, 3 e 9 g.d.l. sono rispettivamente 3.078, 1.886, 1.638 e 1.383.

Inferenza

L'INFERENZA si propone di trovare, per via induttiva, conclusioni (informazioni) sulle caratteristiche della popolazione attraverso un campione estratto da essa.

Una STATISTICA è la funzione che riferita alla variabili x1, x2, …, xn genera una variabile casuale. S= f(x1,x2,…,xn) applicata ai valori x1,x2, …, xn del campione assume un valore numerico.

Un PARAMETRO di una popolazione (media, varianza) è la stima dedotta dalle osservazioni effettuate su di un campione casuale.

Una VARIABILE CASUALE è il risultato dell'associazione di una probabilità agli eventi (normalmente espresse dai numeri).

UMG Dubium sapientiae initium
Lezione VI

Inferenza

Una STIMA rappresenta l'utilizzazione di parametri campionari per trarre conclusioni sui parametri di tutta la popolazione (ad es. si ricavano i parametri μ e σ da quelli campionari). Si effettua così la stima dei parametri di tutta la popolazione.

Si può stimare un parametro di una popolazione in un punto (stima puntuale) o in un intervallo (stima intervallare).

Uno STIMATORE CORRETTO è quel parametro campionario, ad es la media, che risulta essere uguale al corrispondente parametro della popolazione. Se questo parametro non è uguale ci troviamo di fronte ad uno STIMATORE DISTORTO.

Inferenza in simboli

$N = popolazione\_osservata$

$f_a(n) = distribuzione\_popolazione\_N$

$n = campione\_casuale\_semplice$

estratto da N (n_1,n_2, …, $n_{n)}$ con distribuzione $f_a(n)$

$a = a(f) = parametro\_ignoto$

$\hat{a} = \hat{a}(n_{1,\ldots,} n_n)$

stima di a, $\hat{a}$ è una variabile casuale

$E = a - \hat{a}$

tale differenza rappresenta l'incertazza sull'errore di stima

Lezione VI

Le stime

Una STIMA PUNTUALE consiste in un solo specifico valore: quello che meglio di ogni altro può servire per stimare un parametro della popolazione sotto osservazione.

Una STIMA INTERVALLARE individua un campo di valori e la probabilità (livello di confidenza) che l'intervallo contenga il parametro incognito della popolazione. E' la più utilizzata ed è preferibile a quella puntuale perché indica con sicurezza l'intervallo entro il quale una stima può essere corretta al livello di confidenza prescelto.

Le conclusioni raggiunte sulla base di parametri campionari vengono poi estese a tutta la popolazione con una precisione di valutazione decisa preventivamente.

Stima della media di una popolazione

Quando si vuole stimare la media di una popolazione il parametro campionario da utilizzare è la media campionaria μx i cui limiti di confidenza sono:

$$\mu x \pm zc \frac{\sigma}{\sqrt{Nc}} \cdot \sqrt{\frac{Np - Nc}{Np - 1}}$$

N_c=Numerosità campionaria
N_p=Numerosità popolazione
z_c= funzione normale campionaria

Per n=>30 per calcolare i limiti di confidenza si utilizza la stima campionaria della che produce una stima corretta della Dev-standard della popolazione.

Per Nc<30 essa non approssima in modo sufficiente σ e si deve così utilizzare la teoria dei piccoli campioni (distribuzione t di Student).

UMG Dubium sapientiae initium
Lezione VI

La teoria dei campioni

• Il **campione** è quella parte limitata di popolazione che viene presa in esame;
• La **numerosità o ampiezza** del campione è determinata dal numero di elementi che lo compongono;
• Le **modalità di estrazione** del campione (**campionamento)** possono seguire uno schema probabilistico, quando ogni elemento della popolazione ha una probabilità nota di essere estratto, non probabilistico (a quote, a convenienza) altrimenti;
• Il **campionamento casuale o random** è uno dei principali metodi per ottenere campioni probabilistici. Tale tipo di campionamento segue generalmente due regole: con ripetizione e senza ripetizione.

Tipi di campionamento probabilistico

Possiamo distinguere vari tipi di campione, a seconda del metodo utilizzato per produrlo:

- ***casuale semplice:*** tutte le unità della popolazione di riferimento hanno la stessa probabilità di essere incluse nel campione;
- ***stratificato:*** la popolazione di riferimento viene stratificata e il campione si ottiene da successivi campionamenti sui vari strati;
- ***a grappoli:*** Quando la popolazione di riferimento è naturalmente suddivisa in gruppi di unità spazialmente contigue.

UMG Dubium sapientiae initium
Lezione VI

Distribuzioni campionarie in popolazioni normali

Quando la popolazione da cui si estrae il campione è distribuita normalmente è possibile aggiungere altre proprietà secondo se ci troviamo nel caso di:

a) **popolazioni con σ nota**
b) **popolazioni con σ ignota**

Popolazioni con σ nota

La distribuzione campionaria delle medie è anch'essa normale con:

$$\mu_{\bar{x}} = \mu \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

• la media delle medie campionarie èuguale a quella della popolazione;
• la deviazione standard delle medie campionarie è inferiore alla deviazione standard della popolazione

UMG Dubium sapientiae initium
Lezione VI

Popolazioni con σ nota (continuazione)

Se le medie campionarie sono distribuite normalmente con media μ e deviazione standard allora è distribuita come una variabile normale standardizzata: $\frac{\sigma}{\sqrt{n}}$

$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

attraverso quest'ultima relazione è possibile determinare la probabilità che un campione, di numerosità *n*, con media, provenga da una popolazione con mediaμse è nota la deviazione standard (σ) della popolazione.

Popolazioni con σ ignota

Quando la varianza della popolazione è ignota la dobbiamo stimare a partire dalla varianza campionaria utilizzano la relazione:

$$\hat{\sigma}^2 = \frac{n}{n-1} s^2$$

Da questa relazione possiamo poi ricavare la deviazione standard della distribuzione campionaria delle medie.

UMG Dubium sapientiae initium
Lezione VI

Tabelle di distribuzione

Sono delle tabelle dove si possono rilevare i valori di un particolare probabilità di distribuzione (riportato in colonna) per un particolare grado di libertà (riportato in riga).

Vengono di seguito riportate rispettivamente le tabelle di distribuzione:

1) Normale standard
2) T Student
3) Chi quadro

z*	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414
0.1	.46017	.45620	.45224	.44828	.44433	.44038	.43644	.43251	.42858	.42465
0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
1.0	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08691	.08534	.08379	.08226
1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
2.0	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831

UMG Dubium sapientiae initium

Lezione VI

z*	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
2-4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
3.0	.00135	.00097	.00069	.00048	.00034	.00023	.00016	.00011	.00007	.00005
4.0	.00003	.00002	.00001	.00001	.00001	.00000	.00000	.00000	.00000	.00000

ν	.7500	.8000	.8500	.9000	.9500	.9750	.9900	.9950	.9990	.9995
1	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.22	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850

UMG Dubium sapientiae initium
Lezione VI

ν	.7500	.8000	.8500	.9000	.9500	.9750	.9900	.9950	.9990	.9995
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617	3.160	3.373
250	0.675	0.843	1.039	1.285	1.651	1.969	2.341	2.596	3.123	3.330
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
INF	0.675	0.842	1.036	1.282	1.645	1.960	2.327	2.576	3.091	3.291

	probabilità											
v	.005	.010	.025	.050	.100	.250	.750	.900	.950	.975	.990	.995
1	0.00	0.00	0.00	0.00	0.02	0.10	1.32	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	0.58	2.77	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	1.21	4.11	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	1.92	5.39	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	2.67	6.63	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	3.45	7.84	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	4.25	9.04	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	5.07	10.22	13.36	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	4.17	5.90	11.39	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	6.74	12.55	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	7.58	13.70	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	8.44	14.85	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	9.30	15.98	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	10.17	17.12	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	11.04	18.25	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	11.91	19.37	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	12.79	20.49	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	13.68	21.60	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	14.56	22.72	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	15.45	23.83	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	16.34	24.93	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	17.24	26.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	18.14	27.14	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	19.04	28.24	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	19.94	29.34	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	20.84	30.43	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	21.75	31.53	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	22.66	32.62	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	23.57	33.71	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	24.48	34.80	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	33.66	45.62	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	42.94	56.33	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	52.29	66.98	74.40	79.08	83.30	88.38	91.95
90	59.20	61.75	65.65	69.13	73.29	80.62	98.65	107.6	113.1	118.1	124.1	128.3
120	83.85	86.92	91.57	95.70	100.6	109.2	130.1	140.2	146.6	152.2	159.0	163.6

Il Sistema d'ipotesi

I problemi di scelta tra due (o più) ipotesi, in statistica vengono chiamati problemi di verifica d'ipotesi. Le ipotesi sono generalmente chiamate: ipotesi nulla H0 e ipotesi alternativa H1.

Lo strumento utilizzato per affrontare problemi di verifica d'ipotesi viene chiamato **test statistico**. Quest'ultimo rappresenta il mezzo utile per verificare quanto i dati a disposizione siano o meno a favore delle mie ipotesi.

A livello teorico alcuni test sono più adatti di altri in certe condizioni per il loro comportamento asintotico. I tests si dividono in **parametrici** e **non parametrici**.

UMG Dubium sapientiae initium
Lezione VII

I tests parametrici

Assumono che i nostri dati si distribuiscano con delle distribuzioni note (es:. Gaussina). Il test t, l'analisi della varianza, la correlazione, la regressione, insieme con gli altri test di statistica multivariata sono parte dei metodi di inferenza detti "classici" o "parametrici". Prima della loro applicazione, è fondamentale che vengano verificati e soddisfatti alcuni assunti che riguardano la popolazione d'origine:

1) **Indipendenza dei gruppi campionari**: le osservazioni di ogni gruppo dovrebbero essere formate per estrazione casuale da una popolazione, in cui ogni soggetto abbia la stessa probabilità di essere incluso in qualsiasi gruppo. In questo modo, i fattori aleatori o non controllati dovrebbero essere casualmente distribuiti e non generare distorsioni od errori sistematici.

I tests parametrici (continuazione)

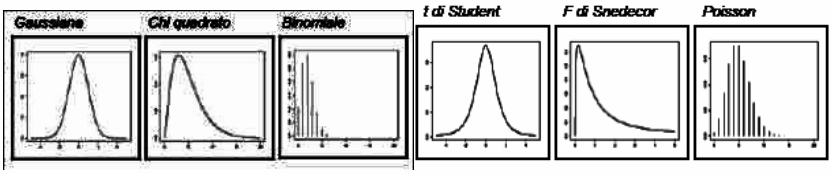
2) **Normalità delle distribuzioni**: da essa deriva la relazione tra popolazione e campioni, secondo il teorema del limite centrale:se, da una popolazione con media μ e varianza σ2 normalmente distribuita, si estraggono casualmente alcuni campioni di campioni di dimensione n, le loro medie si distribuiranno normalmente con media generale μ e varianza della media σ2/n la non normalità è indice serio di estrazione non casuale.

3) **Omoscedasticità o omogeneità delle varianze**: se sono formati per estrazione casuale dalla medesima popolazione, i vari gruppi devono avere varianze eguali

UMG Dubium sapientiae initium
Lezione VII

I tests parametrici (continuazione)

Le distribuzioni, ovvero le curve di frequenza teorica, rappresentano il comportamento delle nostre variabili all'aumentare della dimensione del campione (per n che tende all'infinito). Le distribuzioni più comuni sono:



I tests non parametrici

Test non parametrici: non viene fatta nessuna assunzione sul tipo di distribuzione dei dati originali.

Hanno le seguenti caratteristiche:

• non dipendono dalla forma di distribuzione della popolazione

• non prevedono il calcolo della media, bensì della mediana come misura della tendenza centrale

• permettono inferenze anche su dati qualitativi o di rango.

UMG Dubium sapientiae initium
Lezione VII

I tests non parametrici (continuazione)

TEST U di MANN-WHITNEY
Confronto di dati ordinali. Il test U è adatto al confronto di due serie di dati ordinali. Ad esempio due serie di punteggi assegnati in un test.

TEST KOLMOGOROV-SMIRNOV
Confronto dati quantitativi. Si suddivide l'intervallo di variazione in classi di frequenza di uguale ampiezza. Ad ogni classe si attribuiscono le frequenze cumulative del primo e del secondo campione.

TEST di WILCOXON o test dei ranghi per confronto di gruppi in campioni appaiati, analogo del test t.

TEST di SPEARMAN o di correlazione dei ranghi. E' l'analogo non parametrico del coefficiente di correlazione di Pearson.

Il teorema del limite centrale

Teorema del limite centrale richiede l'assunzione di indipendenza.

Se una variata X ha una ha una distribuzione f(X), la media di un campione X(n) su n osservazioni tende ad essere distribuita normalmente al crescere di n. Cioè per n sufficientemente grande la media campionaria si distribuisce asintoticamente come una normale.

$$\bar{x}_{n\to\infty} \to N(\mu, \frac{\sigma^2}{n})$$

UMG Dubium sapientiae initium
Lezione VII

Requisiti di un test statistico (parametrico)

1) Deve risultare nota la funzione di distribuzione della V.C. descritta dal test sotto l'ipotesi H0 in modo da poter fissare a priori la zona di rifiuto H0.
2) Il test deve essere non distorto nel senso che per qualsiasi valore di n e del parametro delta di non centralità (cioè per qualsiasi ipotesi alternativa H1) si deve sempre avere 1-β.
3) Il test deve essere consistente cioè all'aumentare dell'ampiezza n del campione il valore β deve tendere a zero. Deve, cioè, essere minima la probabilità di accettare H0 quando è falsa (potenza del test).
4) La performance del test è importante nel calcolo della dimensione del campione. Se il test è poco affidabile (bassa sensibilità e/o specificità), la numerosità del campione dovrà essere alta.

Applicazione del Sistema d'ipotesi

1) Considero la mia variabile di interesse

2) Ipotizzo una ragionevole distribuzione asintotica per la mia variabile nella intera popolazione

3) Formulo un corretto sistema di ipotesi

4) Utilizzo un appropriato test statistico che grazie alla distribuzione ipotizzata precedentemente e alla assunzione di indipendenza avrà una certa distribuzione asintotica

5) Confronto il valore del test con la distribuzione sotto l'ipotesi H0.

UMG Dubium sapientiae initium
Lezione VII

Intervalli di confidenza

Siano μ(PC) e σ(PC) la media e la deviazione standard della stima del generico parametro campionario PC. Se la distribuzione del PC approssima la distribuzione normale (ragionevolmente per n=>30) possiamo presumere che il relativo parametro cada con probabilità del:

68,27% nell'intervallo PC+-σ(PC)

95,45% " " PC+-2σ(PC)

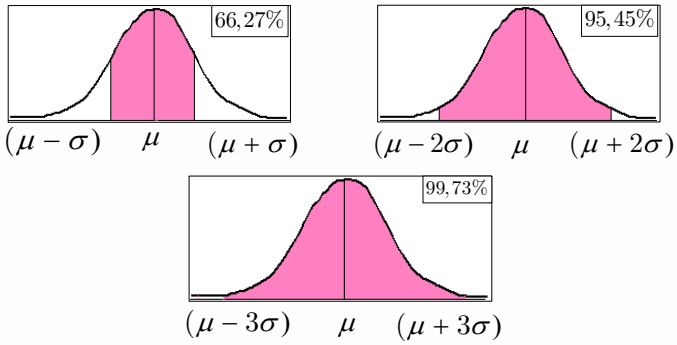
99,73% " " PC+-3σ(PC)

Sono molto utilizzati anche i seguenti intervalli di confidenza:

95% " " PC+-1,96σ(PC)

99% " " PC+-2,58σ(PC)

Intervalli di confidenza (continuazione)



Tali intervalli vengono chiamati **intervalli di confidenza** e i relativi estremi definiti **limiti di confidenza o limiti fiduciali** al 68,27%, al 95,45%, al 99,73%

Lezione VII

Verifica delle ipotesi (1)

Definito un sistema d'ipotesi, il test statistico mi permette di accettare o rifiutare l'ipotesi nulla. In realtà l'accettare l'ipotesi nulla significa non avere elementi sufficienti per rifiutarla.

Fisso α che è la probabilità di rifiutare l'ipotesi nulla (H0) quando è vera. La probabilità α deve essere divisa in due α/2 per scostamenti a dx e per scostamenti a sx.

I valori Tx forniti dal test consentono di verificare se i dati sono a favore o meno dell'ipotesi nulla. I valori nella code portano ad un rifiuto di H0. Valori centrali portano ad accettare H0.

Verifica delle ipotesi (2)

Generalmente si sceglie H0 come ipotesi che si vuole rifiutare. Se rifiuto H0 conosco l'errore α che sto commettendo.

Se non la rifiuto l'errore che sto commettendo è più difficile da determinare.

I tests di significatività (tests delle ipotesi, regole di decisione) permettono di decidere se accettare o rifiutare l'ipotesi H0 allorquando i risultati degli esperimenti sul campione differiscono significativamente dai risultati attesi.

UMG Dubium sapientiae initium
Lezione VII

Verifica delle ipotesi (3)

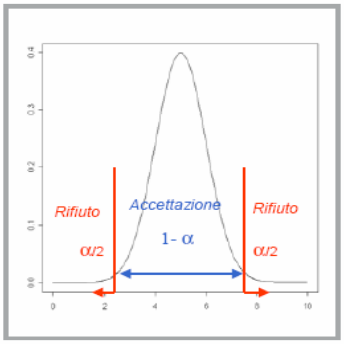
Dopo aver formulato l'ipotesi e prima di estrarre il campione, si indica con α la probabilità massima con cui si accetta di incorrere nell'errore di I tipo.Tale probabilità è detta livello di significatività del test.

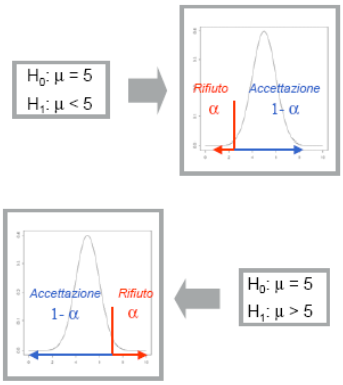
La percentuale 1-α rappresenta l'intervallo di confidenza. Tra gli α più usati vi sono quelli dello 0,05 e 0,01 ossia, 5% e 1%.

Con β si indica la probabilità di commettere un errore di II tipo. La quantità 1-β si chiama potenza del test. Essa rappresenta la probabilità di rifiutare l'ipotesi H0 quando è vera H1.

Regione di rifiuto e di Regione di rifiuto e di accettazione

Abbiamo visto che i valori di Tx ci servono per verificare se i nostri dati sono a favore o meno dell'ipotesi nulla





- Valori nelle code portano ad un rifiuto di H0
- Valori centrali portano ad accettare H0

Lezione VII

Regione di rifiuto e di Regione di rifiuto e di accettazione

Le regioni di accettazione e di rifiuto dipendono però dal tipo di ipotesi scelte:

- Nel caso di ipotesi: H0: μ = 5 H1: μ ≠ 5 La regione di rifiuto sarà bilaterale
- Nel caso di ipotesi: H0: μ = 5 H1: μ < 5 o μ > 5

La regione di rifiuto dipenderà dall'ipotesi alternativa.

Tipi di errore

Prendere una decisione prevede correre dei rischi. Il rischio che corriamo è di prendere una decisione sbagliata.
Quanti tipi di errori posso fare e come faccio a minimizzarli ?
Fisso il livello dell'errore del primo tipo.
Minimizzo il livello dell'altro, motivo per cui le due ipotesi, H0 e H1, non sono simmetriche.

Errori del I tipo: ho α probabilità di rifiutare H0 quando è vera
Errori del II tipo: ho β probabilità di accettare H0 quando è falsa

UMG Dubium sapientiae initium
Lezione VII

Tipi di errore (1)

Generalmente si sceglie come H0 l'ipotesi che si vuole rifiutare:

• Se rifiuto H0 conosco l'errore che sto commettendo: α

• Se non la rifiuto l'errore che commetto è più difficile da determinare: β

Non e' proponibile andare a controllare dove cadono i valori osservati sulla distribuzione sotto H0.

Il test statistico (TX) ci deve restituire un valore numerico attraverso il quale siamo in grado di prendere una decisione.

Se il test ha valori piccoli allora i dati sembrano soddisfare H0, se ha valori grandi, in valore assoluto, allora i dati sembrano non soddisfare H0.

Conclusioni sul sistema d'ipotesi

Il sistema d'ipotesi è composto dalle seguenti 3 fasi:

1) Test di formulazione dell'ipotesi
2) Test del criterio di decisione
3) Test di verifica del rischio d'errore

UMG Dubium sapientiae initium
Lezione VII

Test di formulazione dell'ipotesi

Chiamo **ipotesi**, o **ipotesi zero**, o **ipotesi nulla** (H_0), l'ipotesi per la quale resta definita la distribuzione di campionamento. Chiamo **ipotesi alternativa**, o **altra ipotesi** (H_1) l'insieme delle altre possibili ipotesi.

Test del criterio di decisione

	SE È VERA H_0	SE È VERA H_1
... e in base al campione decido che è vera H_0	**decisione giusta** **protezione**: $(1-\alpha)$	**decisione sbagliata** **errore** di tipo II: β
... e in base al campione decido che è vera H_1	**decisione sbagliata** **errore** di tipo I: α	**decisione giusta** **potenza**: $(1-\beta)$

UMG Dubium sapientiae initium
Lezione VII

Test di verifica del rischio d'errore

Protezione $(1-\alpha)$**:**
probabilità di accettare H_0 quando è vera H_0

Potenza del test $(1-\beta)$**:**
probabilità di rifiutare H_0 quando è vera una specifica H_1

Rischio di errore di tipo I (α)**:**
probabilità di rifiutare H_0 quando è vera H_0

Rischio di errore di tipo II (β)**:**
probabilità di accettare H_0 quando è vera una specifica H_1