Teoria e tecniche dei test

Lezione 2 – 2013/14

ALCUNE NOZIONI STATITICHE DI BASE

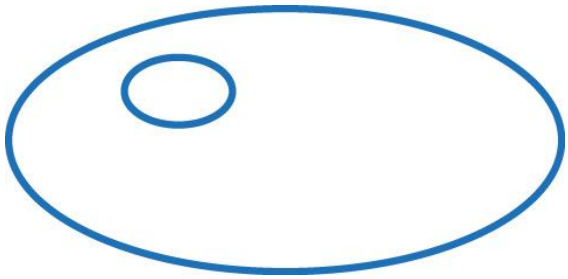
Concetti di base

- Campione e popolazione
- La distribuzione delle variabili
- Indicatori di tendenza centrale
- Indicatori di dispersione
- La distribuzione normale
- Indicatori della forma della distribuzione
- La correlazione

Campione e popolazione (1)

- La **popolazione** è l'insieme di individui o oggetti che si vogliono studiare. Questi individui o oggetti vengono denominati **unità statistiche**.
- Una **variabile** è una caratteristica di ogni appartenente alla popolazione.
- Un **campione** è una parte di popolazione.
- L'**errore di campionamento** è la differenza tra una caratteristica misurata sull'intera popolazione e la stessa riscontrata in un campione di quella popolazione.
- Il **grado di variabilità** è una misura di come gli elementi della popolazione differiscono gli uni dagli altri in riferimento alla variabile sotto studio.

La popolazione e un campione



Campione e popolazione (2)

- Un **parametro** è un valore numerico che descrive una caratteristica della popolazione.
- Una **statistica** è un valore numerico che descrive una caratteristica di un campione.
- La **dimensione della popolazione** è il numero delle unità statistiche della popolazione. Il relativo simbolo è *N*.
- La **dimensione del campione** si indica con *n*.

Campione e popolazione (3)

- I **dati qualitativi** descrivono una caratteristica particolare di un'osservazione campionaria. Nella maggior parte dei casi non sono numerici.
- I dati creati assegnando codifiche numeriche alle diverse categorie, senza che tali numeri abbiano un reale significato, sono chiamati **dati nominali**.
- I dati che sono creati assegnando numeri alle categorie dove l'ordine di assegnazione ha un significato sono chiamati **dati ordinali**.
- Le **scale di Likert** sono utilizzate per raccogliere informazioni su atteggiamenti e opinioni incluso il grado di consenso di una affermazione, frequenza di uso, importanza di un argomento, qualità e gradimento.

Campione e popolazione (4)

- I dati che sono intrinsecamente numerici sono chiamati **dati quantitativi**.
- I **dati discreti** possono assumere solo determinati valori. Questi valori sono spesso numeri interi o comunque non decimali.
- I **dati continui** possono assumere un infinito numero di valori possibili entro un intervallo di valori della scala numerica. Tali valori sono molto spesso il risultato di misurazioni.
- Gli **strumenti della statistica descrittiva** permettono di sintetizzare i dati.
- Una **inferenza** è una deduzione o una conclusione.

La distribuzione delle variabili (1)

- Una **tabella di frequenza** o **distribuzione di frequenza** registra ogni categoria, valore, o classe di valori che una variabile potrebbe avere e il corrispondente numero di volte che ognuna di esse ricorre nei dati. La frequenza della *i*-ma classe è indicata con f_i.
- La **frequenza relativa** di una classificazione consiste nel numero di volte in cui una osservazione si ritrova all'interno della classificazione stessa, rappresentata come una porzione del numero totale di osservazioni. La frequenza relativa può essere espressa come una frazione, decimale o percentuale.

La distribuzione delle variabili (2)

- La **frequenza relativa cumulata** di una classe è la somma della frequenza relativa di quella classe con quelle di tutte le classi precedenti. Rappresenta una porzione del numero totale delle osservazioni e può essere espressa come una frazione, un numero decimale o una percentuale.

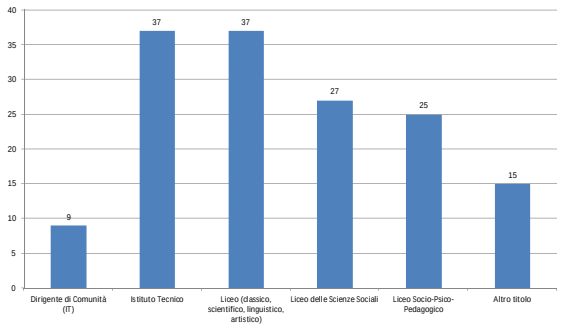
Esempio di distribuzione di frequenza (dai dati raccolti in aula S1)

Tipo di Diploma	frequenza	frequenza %	frequenza % cumulata
Dirigente di Comunità (IT)	9	6,0%	6,0%
Istituto Tecnico	37	24,7%	30,7%
Liceo (classico, scientifico, linguistico, artistico)	37	24,7%	55,3%
Liceo delle Scienze Sociali	27	18,0%	73,3%
Liceo Socio-Psico-Pedagogico	25	16,7%	90,0%
Altro titolo	15	10,0%	100,0%
Totale	**150**		

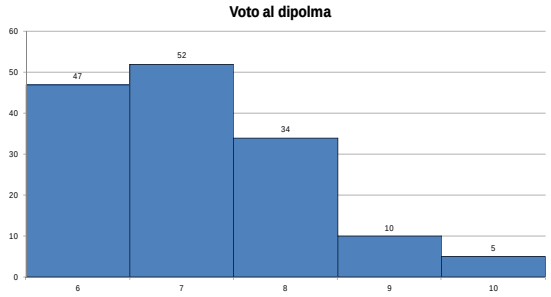
La distribuzione delle variabili (3)

- Il **diagramma a barre** rappresenta la frequenza o la frequenza relativa di una tabella di frequenza sotto forma di un rettangolo o barra o colonna.
- Un **istogramma** è molto simile a un diagramma a barre ma la scala di misura dell'asse delle ascisse deve tener conto che i dati sono intrinsecamente ordinati.

Esempio di diagramma a barre (dai dati raccolti in aula S1)



Esempio di istogramma (dai dati raccolti in aula S1)



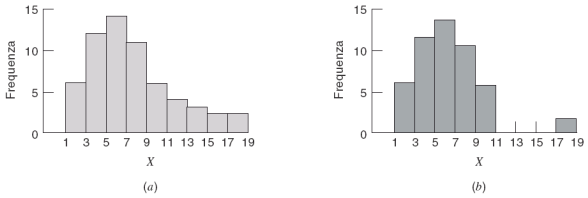
Indici riassuntivi numerici

- La **tendenza centrale** o **posizione** di un insieme di dati indica dove, numericamente, i dati sono posizionati o concentrati.
- La **forma** di un insieme di dati descrive come i dati si distribuiscono intorno ai valori centrali relativamente alla simmetria o asimmetria.
- La **variabilità** di un insieme di dati descrive come i dati sono disposti intorno ai valori della tendenza centrale.

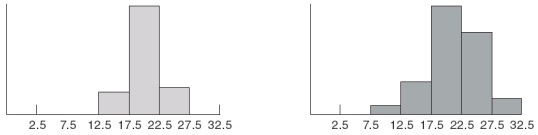
Distribuzione delle variabili (4)

- Quando i dati sono distribuiti uniformemente su entrambi i lati del picco la distribuzione è **simmetrica**.
- Quando i dati non sono distribuiti uniformemente su entrambi i lati del picco la distribuzione è **asimmetrica**.

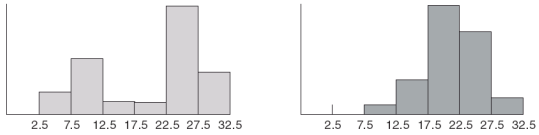
Istogrammi con (*a*) dati asimmetrici, (*b*) valori estremi



Istogrammi con diversa variabilità che mostrano la dispersione dei dati



Istogrammi con diversa variabilità che mostrano diverse regolarizzazioni



Indicatori di tendenza centrale (1)

- Una **statistica** è un descrittore numerico calcolato dai dati campionari ed è usato per descrivere il campione. Le statistiche, di norma, si rappresentano con lettere romane.
- Un **parametro** è un descrittore numerico usato per descrivere la popolazione. I parametri, di norma, si rappresentano con lettere greche.

Indicatori di tendenza centrale (2)

- La **media aritmetica** sintetizza la posizione (tendenza centrale) della distribuzione d'un insieme di dati. Si trova sommando tutti i valori dei dati e dividendo per il numero totale delle osservazioni.
- La **media della popolazione** è indicata dalla lettera greca μ (mu).
- La **mediana** è il valore dell'osservazione centrale d'una distribuzione ordinata di dati.
- La **moda** è il valore con la frequenza più alta nel campione.

Indicatori di dispersione

- Il **campo di variazione** (range), ***R***, è la differenza fra l'osservazione maggiore e quella minore del campione.

$$R = Max - Min$$

- La **varianza campionaria**, s^2, è la media dei quadrati degli scarti tra ciascun valore e la media campionaria. Lo **scarto quadratico medio *s*** è la radice quadrata positiva della varianza.
- Lo **scarto quadratico medio** e la **varianza della popolazione** si indicano rispettivamente con σ (sigma) e σ^2.

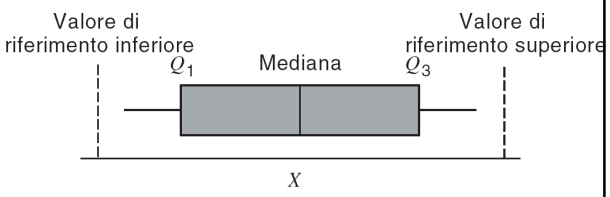
Indicatori di posizione (1)

- Il **valore standardizzato** misura di quanti "scarti quadratici medi" un valore dista dalla media.
- Un **valore anomalo** (outlier) è un valore che ha una probabilità molto bassa di verificarsi.
- Il ***p* -esimo percentile** di un insieme di dati è il valore per cui una percentuale pari a *p* delle osservazioni è inferiore o uguale a esso.
- Il **rank di percentile** di un valore è la percentuale dei dati del campione pari o al di sotto del valore di interesse.

Indicatori di posizione (2)

- Il **primo quartile Q_1** è un valore tale che il 25% dei dati è inferiore o uguale a esso.
- Il **terzo quartile Q_3** è un valore tale che il 75% dei dati è inferiore o uguale a esso.
- Un **grafico a scatola** è una rappresentazione grafica che utilizza le statistiche di sintesi per rappresentare la distribuzione di un insieme di dati.
- Lo **scarto interquartile** (**SIQ**) è la differenza tra il terzo e il primo quartile, $Q_3 - Q_1$.
- Il **valore di riferimento inferiore** di un grafico a scatola è posizionato a $Q_1 - 1.5$ (SIQ) e quello superiore a $Q_3 + 1.5$ (SIQ).

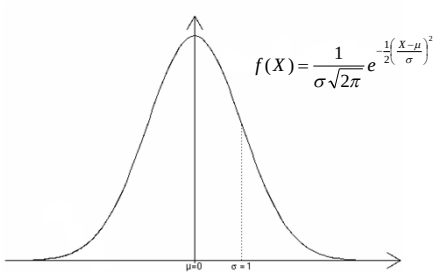
Grafico a scatola e baffi



La distribuzione normale (1)

- La **distribuzione normale** o gaussiana è la più importante distribuzione di probabilità di *variabili continue.*
- È caratterizzata da due parametri: **media** (μ) e **varianza** (σ^2)
- Assume valori compresi tra -∞ e +∞
- Descrive una curva di tipo simmetrico **a campana**, che raggiunge il punto più alto (massimo valore di y) in corrispondenza del valore μ
- media, mediana e moda coincidono
- presenta due punti di flesso in corrispondenza di μ+σ e μ-σ; cioè i punti in cui la curva da convessa diventa concava si trovano in corrispondenza a ±1 deviazione standard dalla media

La Distribuzione Normale

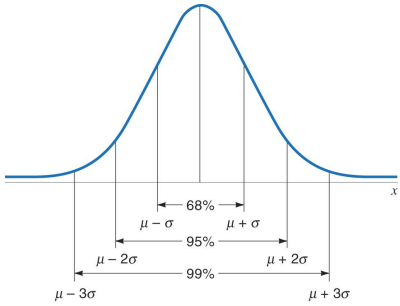


26

La distribuzione normale (2)

- La **regola empirica** dice che per una distribuzione simmetrica "a campana":
  - Circa il 68% delle osservazioni si trovano entro ± uno scarto quadratico medio della media.
  - Circa il 95% delle osservazioni si trovano entro ± due scarti quadratici medi della media.
  - Quasi tutte (più di 99%) le osservazioni si trovano entro ± tre scarti quadratici medi della media.

La regola empirica



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

La Distribuzione Normale (3)

- La **distribuzione normale standardizzata** si ottiene con la trasformazione lineare dei punti grezzi in punti *z*

$$z = \frac{x - \bar{X}}{s_X}$$

Dove *x* è il valore del soggetto, $\bar{X}$ è la media del campione, e s_X è la deviazione standard del campione

29

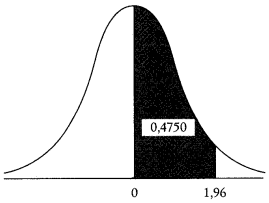
La distribuzione normale (4)

- Nella distribuzione normale standardizzata le probabilità corrispondenti alle superfici racchiuse dalla curva normale possono essere calcolate.
- Queste probabilità sono state tabulate e vengono riportate in apposite **tabelle**.

30

La distribuzione normale (4)

Ad esempio, grazie alle tabelle è possibile conoscere l'area compresa tra le z = 0 e z = 1,96



31

Indicatori della forma della distribuzione

- La normalità della distribuzione è fondamento di molte analisi.
- L'**asimmetria** indica quanto una curva si allontana da una distribuzione normale in termini di spostamento a sinistra o a destra
- La **curtosi** indica quanto una curva si allontana da una distribuzione normale in termini di maggiore appiattimento o maggiore allungamento
- Con valori degli indici di asimmetria e curtosi **compresi tra -1 e 1** la distribuzione può considerarsi normale

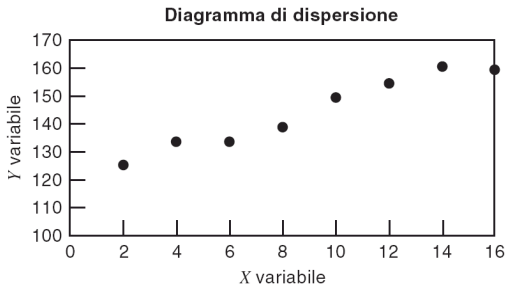
32

La correlazione

- Il **coefficiente di correlazione** è una misura della forza di una relazione lineare tra due variabili X e Y. Una correlazione di -1 corrisponde ad una perfetta relazione negativa e una correlazione di +1 corrisponde ad un perfetta relazione positiva.
- Il coefficiente di correlazione di **Pearson**, a partire dai punti z, si calcola nel seguente modo:

$$r_{xy} = \frac{\sum z_x z_y}{N}$$

La rappresentazione grafica dei dati di una correlazione



La regressione lineare semplice (1)

- La relazione tra le variabili *X* e *Y*, ovvero il **modello di regressione lineare semplice**, può essere descritta dall'equazione di una retta:

$$\hat{y} = b_0 + b_1 x$$

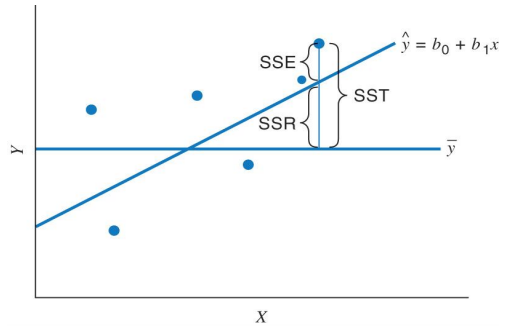
Dove $\hat{y}$ è il valore di *y* predetto in base all'equazione, *x* è il valore del soggetto nella variabile X, b_0 e b_1 sono stime dei parametri che indicano rispettivamente il punto in cui la retta incontra l'asse delle y e l'inclinazione della retta stessa.

- La tecnica per trovare l'equazione della retta che minimizza la somma totale dei quadrati delle deviazioni tra dati osservati e punti sulla retta si dice **metodo dei minimi quadrati**.

La regressione lineare semplice (2)

- La distanza tra il valore predetto di *Y*, $\hat{y}$ e il valore osservato, *y*, è detta **deviazione** o **errore di previsione**.
- **L'errore standard della stima, $s_{y|x}$**, è una misura di quanto i dati osservati variano attorno alla retta di regressione.

Componenti della variazione dei valori di *y*



Tre tipi di relazioni:

Perfettamente negativa, nessuna relazione e perfettamente positiva