Metodi statistici per l'analisi dei dati

Massimiliano Grosso
Dipartimento di Ingegneria Meccanica, Chimica e dei Materiali
E-mail: massimiliano.grosso@dimcm.unica.it
Web: http://people.unica.it/massimilianogrosso

1

Metodi Statistici per l'Analisi dei Dati

INTRODUZIONE

2

Motivazioni

Richiami di statistica – Esperimenti replicati

- Obbiettivo del corso:
  - **Pianificazione** degli esperimenti in maniera tale che i risultati della campagna sperimentali possano essere analizzati con **metodi statistici**, per giungere a conclusioni **oggettive** del processo in esame

Due fasi distinte:

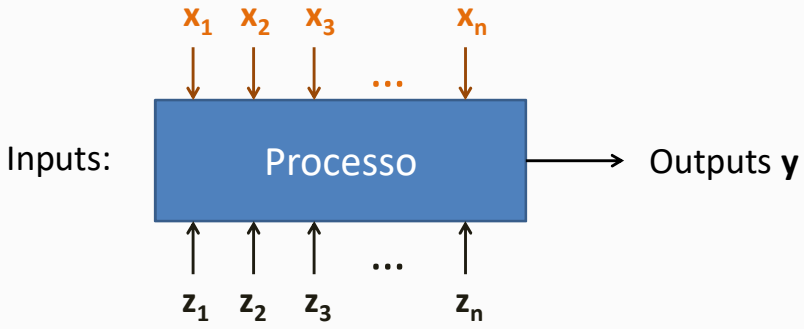
1. **Pianificazione** della campagna sperimentale (Design Of Experiments: **DOE**)
2. Analisi **statistica** dei risultati

3

Motivazioni

Richiami di statistica – Esperimenti replicati

Fattori **controllabili**:

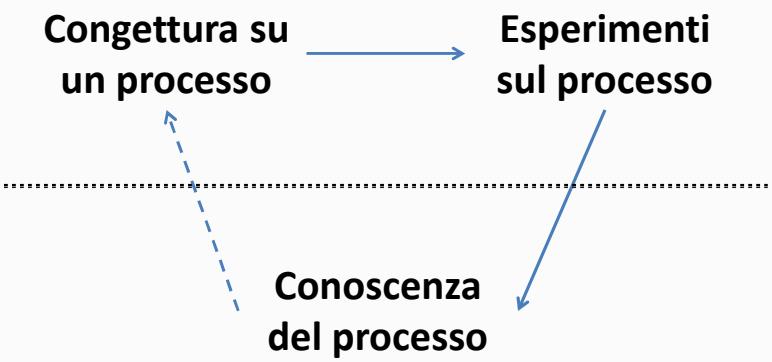


Fattori **incontrollabili**:

4

Motivazioni

- Lo studio di un processo è una procedura iterativa



5

Progettazione campagna sperimentale – Concetti di base

- I principi di base della progettazione della campagna sperimentale sono:

1. Replicazione
2. Randomizzazione
3. Blocking

6

Progettazione campagna sperimentale – Concetti di base

Richiami di statistica – Esperimenti replicati

- **Replicazione**

- Ripetere gli esperimenti nelle stesse condizioni più volte

1. Permette di ottenere una stima «**genuina**» dell'errore sperimentale
2. Permette una stima più **precisa** della variabile di output

- **Replica della misura sperimentale** è un concetto diverso (e più potente) della **misura ripetuta**
  - Nell'ultimo caso si può valutare al più la variabilità intrinseca del sistema di misura

7

Progettazione campagna sperimentale – Concetti di base

Richiami di statistica – Esperimenti replicati

- **Randomizzazione**

- **Ordine** con cui sono eseguite le misure sperimentali deve essere del tutto **casuale**
- Randomizzando l'ordine delle esperienze si possono *compensare* eventuali effetti di ulteriori fattori (non considerati nel modello) che possono essere presenti e che risultano essere soggetti a variazioni nel tempo

8

Progettazione campagna sperimentale – Concetti di base

- **Blocking**

- Tecnica di progettazione della campagna sperimentale usata per aumentare la precisione con cui sono effettuati i confronti tra i fattori di interesse.
- Il Blocking è usato per ridurre la variabilità relativa a fattori di disturbo
  - fattori che possono influenzare la risposta ma a cui non siamo interessati
- **Blocco – Definizione**
- Un insieme di condizioni sperimentali relativamente omogenee

9

Linee guida per una campagna sperimentale

1. Definizione del problema
2. Scelta dei fattori, livelli e intervalli
3. Selezione delle variabili da misurare
4. Pianificazione della campagna sperimentale
5. Esperimenti
6. Analisi statistica dei dati
7. Conclusioni

- Le linee guida riportate sono valide qualunque sia il livello di conoscenza del processo

10

Conoscenza del processo – Outline degli strumenti a disposizione

Richiami di statistica – Esperimenti replicati

Conoscenza del processo

Half fractional design	Screening preliminare delle variabili
Full factorial design	Valutazione delle variabili che influenzano il processo
Modelli empirici	Modelli lineari
Ottimizzazione	Response Surface Model
Modelli a principi primi	Regressione non lineare

Oggetto del corso

11

Metodi statistici per l'analisi dei dati

RICHIAMI DI STATISTICA – ESPERIMENTI REPLICATI

12

Introduzione alla sezione

- La discussione permetterà di rivedere diversi concetti di base di statistica
  - Variabili aleatorie
  - Distribuzioni di probabilità
  - Campioni aleatori
  - Distribuzioni di campionamento
  - Test delle ipotesi – Intervalli di fiducia
- Per il momento esperimenti effettuati **sempre nelle stesse condizioni**.
- N.B. Da non confondere **esperimenti replicati nelle stesse condizioni** con **misure ripetute**

13

Esempio introduttivo

- Si intende monitorare la qualità di una crema destinata ad uso alimentare.
- A tal proposito sono considerati 10 diversi campioni della crema e, per ciascuno di essi è misurata la viscosità
  - 10 misure sperimentali di viscosità riportate in tabella
- L'insieme di misure di viscosità è un **campione** sperimentale.

j	Controllo (cp)
1	70.00
2	70.52
3	73.00
4	72.00
5	71.44
6	71.00
7	72.88
8	71.60
9	71.84
10	72.60
$\bar{y}$	**71.69**

14

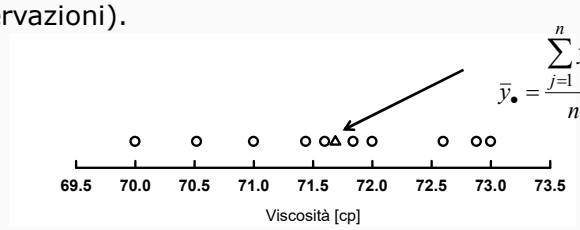
Concetti di statistica di base

- Le **prove sperimentali** (etichettate con il pedice j) differiscono tra loro per effetto delle fluttuazioni dovute all'**errore sperimentale**.
- La presenza dell'errore sperimentale implica che la singola misura sia l'esito di una **variabile aleatoria** (ovvero, non è possibile a priori la sua previsione).

15

Concetti di statistica di base – Descrizioni grafiche della variabilità

- **Diagramma per punti**
- Utile per campioni di piccole dimensioni (sino a 20 osservazioni).



- Il diagramma permette di riconoscere il **trend centrale** e la **dispersione** dei dati.

16

Concetti di statistica di base – Indici di posizione e dispersione del campione

Richiami di statistica – Esperimenti replicati

- Scalari per identificare il trend centrale:
- **Media aritmetica**

$$\bar{y}_\bullet = \frac{\sum_{j=1}^{n} y_j}{n} = 71.69$$

- **Mediana**: rappresenta il valore centrale che divide il campione in due parti uguali costituiti rispettivamente dai valori inferiori e superiori ad esso

17

Concetti di statistica di base – Indici di posizione e dispersione del campione

Richiami di statistica – Esperimenti replicati

- Misure della **dispersione** dei dati:
- **Varianza:**

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(y_j - \bar{y}_\bullet\right)^2$$

La somma dei quadrati è divisa per (*n*-1) anziché *n*

- **Deviazione standard**
- È la radice quadrata della varianza

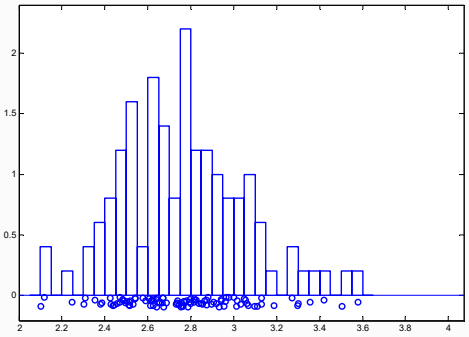
$$s = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}\left(y_j - \bar{y}_\bullet\right)^2}$$

- Utile perché ha le stesse dimensioni della variabile y

18

Concetti di statistica di base – Descrizioni grafiche della variabilità – Frequenze

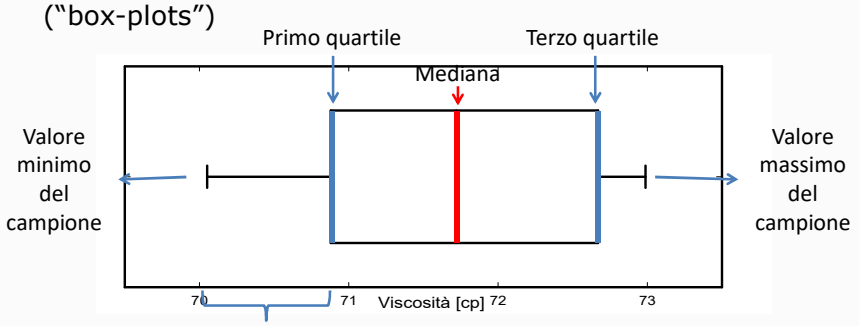
- In presenza di campioni di dimensioni maggiori è possibile riportare i dati negli istogrammi delle **frequenze assolute** (o **relative**) del campione di dati.
- L'istogramma è costruito dividendo l'asse orizzontale in intervalli (in genere di uguale lunghezza) e disegnando un rettangolo sul *j*-esimo intervallo la cui area sia proporzionale a n_j, numero di osservazioni che cadono nell'intervallo.



19

Concetti di statistica di base – Descrizioni grafiche della variabilità

- Rappresentazione dei campioni tramite "diagrammi a scatola" ("box-plots")



Il **25%** delle osservazioni cade in questo intervallo

Il **50 %** delle osservazioni cade in questo intervallo

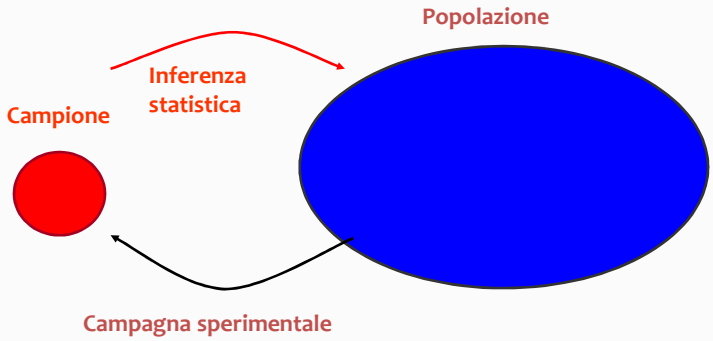
Il **75%** delle osservazioni cade in questo intervallo

20

Campioni e distribuzioni campionarie

Richiami di statistica – Esperimenti replicati

- L'obiettivo dell'inferenza statistica è trarre delle conclusioni su una popolazione a partire da un suo campione



Dal campione si intende ottenere informazioni sulla popolazione generatrice non nota

21

Caratterizzazione della Popolazione – Distribuzioni di probabilità

Richiami di statistica – Esperimenti replicati

- La struttura di probabilità di una variabile aleatoria (**VA**) Y è descritta dalla sua **funzione densità di probabilità** (probability density function: **pdf**) $f(y)$.
- Proprietà fondamentali della pdf di una VA:

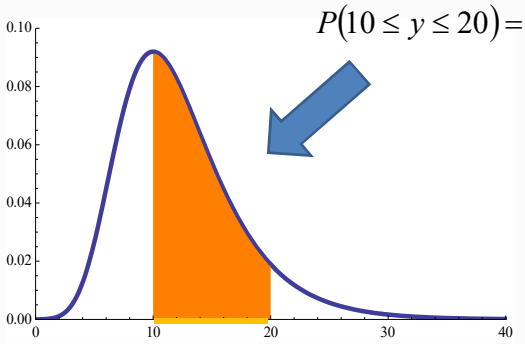
1. $P(a \le y \le b) = \int_a^b f(y)dy$
2. $\int_{-\infty}^{+\infty} f(y)dy = 1$
3. $f(y) \ge 0$

22

Caratterizzazione della Popolazione – Distribuzioni di probabilità

- Esempio di funzione densità di probabilità

$$P(10 \le y \le 20) = \int_{10}^{20} f(y)dy$$



23

Distribuzioni di probabilità – Scalari associati

- **Media** di una variabile aleatoria Y (anche definito **valore atteso**)

$$\mu = \int_{-\infty}^{+\infty} y\, f(y)dy = E[Y]$$

- **Definizione**
- L'operatore **Valore Atteso** E[X] restituisce il risultato medio che si osserverebbe per in presenza di infinite osservazioni della Variabile Aleatoria X

24

Caratterizzazione della Popolazione – Scalari associati ad una VA

- **Varianza** di una variabile aleatoria Y

$$\sigma^2 = V(Y) = \int_{-\infty}^{+\infty} (y-\mu)^2 f(y) dy = E\left[(Y-\mu)^2\right]$$

- Varianze piccole sono associate ad incertezze piccole.

25

Caratterizzazione della Popolazione – Scalari associati ad una VA

- Alcune proprietà di interesse delle VA. 1/2
  1. $E[c] = c$
  2. $E[Y] = \mu$
  3. $E[cY] = cE[y] = c\mu$
  4. $V[c] = 0$
  5. $V[Y] = \sigma^2$
  6. $V[cY] = c^2 V[Y] = c^2\sigma^2$

26

Caratterizzazione della Popolazione – Scalari associati ad una VA

Richiami di statistica – Esperimenti replicati

- Alcune proprietà di interesse delle VA. 2/2
- In presenza di più variabili aleatorie:

6. $E[Y_1 + Y_2] = E[Y_1] + E[Y_2] = \mu_1 + \mu_2$

7. $V[Y_1 + Y_2] = V[Y_1] + V[Y_2] + 2\,\mathrm{cov}(Y_1, Y_2)$

- Dove è definita la covarianza delle VA Y_1 e Y_2:

$$\mathrm{cov}(Y_1, Y_2) = E[(Y_1 - \mu_1)(Y_2 - \mu_2)]$$

27

Caratterizzazione della Popolazione – Scalari associati ad una VA

Richiami di statistica – Esperimenti replicati

- **Statistica – Definizione**:
- Una **statistica** è una **funzione** delle **osservazioni** di un **campione** che non contiene parametri **ignoti** della **popolazione** che ha generato il campione (es: media e varianza).
- Esempi di statistiche:
- **Media aritmetica** $\bar{Y} = \dfrac{\sum_{i=1}^{n} Y_i}{n}$
- **Varianza campionaria** $S^2 = \dfrac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1}$

28

Campioni e distribuzioni campionarie - Stimatori

Richiami di statistica – Esperimenti replicati

- **Stimatore – definizione:**
- Uno **stimatore** di un **parametro** ignoto è una statistica che mira a valutare il parametro stesso.
- La media aritmetica e la varianza campionaria sono esempi di stimatori **puntuali**.
- Lo stimatore puntuale del generico parametro θ è in genere indicato con il simbolo del cappuccio:

$$\hat{\theta}$$

- Esempio media aritmetica:

$$\bar{Y} = \sum Y_i / n = \hat{\mu}$$

- Un valore numerico puntuale calcolato da un campione di dati, prende il nome di **stima**.

29

Campioni e distribuzioni campionarie - Stimatori

Richiami di statistica – Esperimenti replicati

- Proprietà stimatori
- **Imparzialità**: Uno stimatore si dice **imparziale** (*unbiased*) se il suo valore atteso coincide con il valore vero del parametro

$$E\left[\hat{\theta}\right] = \theta$$

- NB sebbene il valore vero non sarà mai noto è possibile valutare il verificarsi della imparzialità.

- **Efficienza:** È una misura della **varianza dello stimatore**. Se dispongo di più stimatori devo scegliere quello con varianza minima ovvero quello con la massima efficienza.

30

Campioni e distribuzioni campionarie - Stimatori

Richiami di statistica – Esperimenti replicati

- Lo stimatore **media aritmetica** è **imparziale**:

$$E[\bar{Y}] = E\left[\frac{\sum_{i=1}^{n} Y_i}{n}\right] = \frac{1}{n} E\left[\sum_{i=1}^{n} Y_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[Y_i] =$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu$$

31

Campioni e distribuzioni campionarie - Stimatori

Richiami di statistica – Esperimenti replicati

- Lo stimatore media aritmetica è **efficiente**:

$$V[\bar{Y}] = V\left[\frac{\sum_{i=1}^{n} Y_i}{n}\right] = V\left[\sum_{i=1}^{n} \frac{1}{n} Y_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} V[Y_i] =$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

32

Campioni e distribuzioni campionarie - Stimatori

Richiami di statistica – Esperimenti replicati

- In maniera analoga si può dimostrare che la **varianza campionaria** S^2 è **imparziale**

$$E\left[S^2\right] = E\left[\frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}{n-1}\right] = \frac{1}{n-1}\sum_{i=1}^{n} E\left[\left(Y_i - \bar{Y}\right)^2\right] =$$

$$= \frac{1}{n-1} E\left[SS\right]$$

- dove SS è la **somma corretta dei quadrati** delle osservazioni y_i

$$SS = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$$

33

Campioni e distribuzioni campionarie - Stimatori

Richiami di statistica – Esperimenti replicati

- Dimostrazione imparzialità varianza – Continua

$$E\left[SS\right] = E\left[\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2\right] = E\left[\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2\right] =$$

$$= \sum_{i=1}^{n}\left(\mu^2 + \sigma^2\right) - n\left(\mu^2 + \sigma^2/n\right) = \left(n-1\right)\sigma^2$$

- da cui:

$$E\left[S^2\right] = \frac{1}{n-1} E\left[SS\right] = \sigma^2$$

34

Campioni e distribuzioni campionarie – Definizione gradi di libertà

Richiami di statistica – Esperimenti replicati

- Il **numero di gradi di libertà** di una **somma di quadrati** è data dal numero di elementi **indipendenti** presenti nella somma.
- Esempio: SS ha n-1 g.d.l.

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

- In SS **non tutti gli elementi sono indipendenti**.
- Il valore della media $\bar{y}$ deve essere tale da soddisfare il vincolo:

$$\sum_{i=1}^{n} (y_i - \bar{y}) = 0$$

35

Campioni e distribuzioni campionarie – Definizione gradi di libertà

Richiami di statistica – Esperimenti replicati

- **Risultato generale:**
- Se y è una variabile aleatoria di varianza σ^2 e una somma degli scarti quadratici ricavata da essa ha ν g.d.l., allora

$$E\left[\frac{SS}{\nu}\right] = \sigma^2$$

- **Proprietà importante** per le applicazioni successive

36

Caratterizzazione della Popolazione – Esempi di distribuzione

- **Distribuzione di tipo normale o Gaussiana**
- La densità di probabilità è data da:

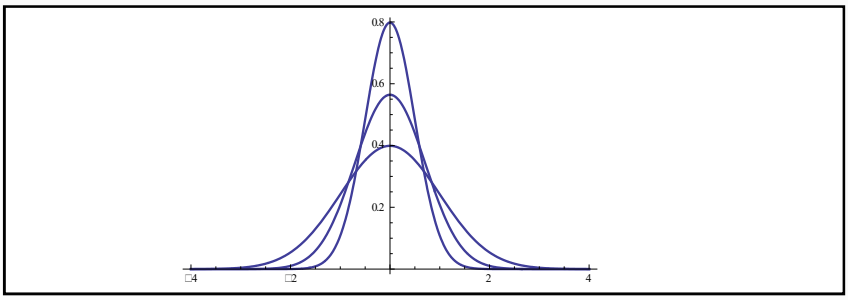
$$f(y)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right) \qquad -\infty < y < +\infty$$

- La funzione è definita lungo tutto l'asse reale (ovvero un qualunque numero reale può essere un esito di una VA di tipo normale)
- Il grafico di tale funzione è una curva a campana simmetrica rispetto a $y=\mu$
- La distribuzione dipende da due parametri, μ e σ^2.

37

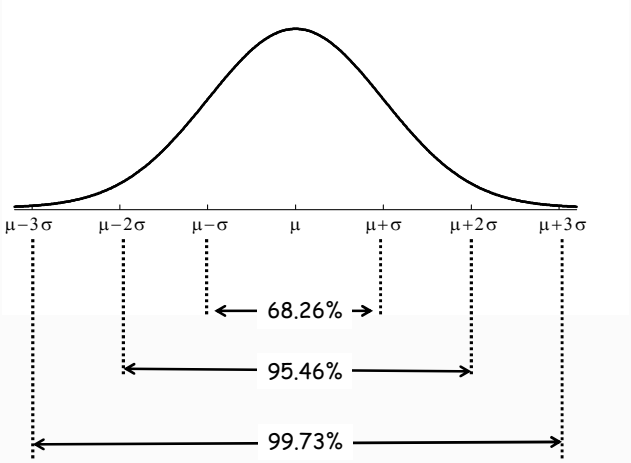
Distribuzione normale

In figura sono riportate tre gaussiane con egual media e varianza 0.25, 0.5, 1



38

Distribuzione normale



Aree sottese dalla distribuzione normale

N.B.

Questo è vero per ogni valore di μ e σ nel caso della Gaussiana!

39

Distribuzione normale di tipo standard – Definizione

- Data una variabile aleatoria Y (di tipo gaussiano) di media μ e varianza σ^2

$$Y \sim N(\mu, \sigma^2)$$

- Si consideri la seguente trasformazione lineare:

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

- È facile verificare che la nuova VA Z ha media 0 e varianza unitaria:

$$Z \sim N(0,1) \Rightarrow$$ **Gaussiana di tipo standard**

40

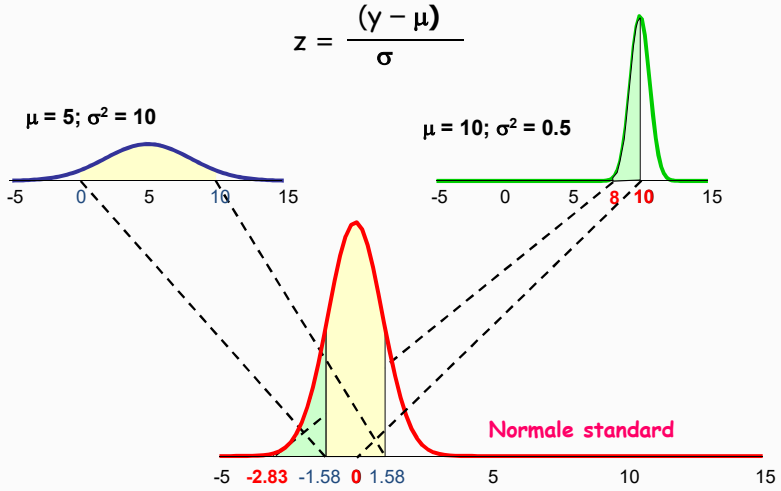
Funzioni di VA Gaussiane
Trasformazioni lineari

- Nota la funzione di distribuzione standard è possibile ricavare le proprietà di una qualsiasi distribuzione gaussiana
- In particolare, è possibile calcolare la probabilità che si verifichi un dato evento per un generico processo, con media e varianza note.
- Questo è possibile sapendo solo i valori della distribuzione di tipo standard.

41

Calcolo probabilità per una Gaussiana generica

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



42

Calcolo probabilità per una Gaussiana generica

- Esempio: calcolare quale è la probabilità che si verifichi un evento appartenente all'intervallo [0,5] per la variabile aleatoria di media 3 e deviazione standard 2:
- Si deve calcolare quale è la probabilità che la variabile aleatoria di tipo standard assuma un valore nell'intervallo corrispondente.

Calcolo probabilità per una Gaussiana generica

- Dobbiamo calcolare la probabilità:

$$P(0 < X < 5)$$

- Gli estremi dell'intervallo corrispondente per la distribuzione di tipo standard possono essere facilmente calcolati

$$z_1 = \frac{x_1 - \mu_X}{\sigma_X} = \frac{0-3}{2}$$
$$z_2 = \frac{x_2 - \mu_X}{\sigma_X} = \frac{5-3}{2} = 1$$

$$\Rightarrow \quad P(0 < X < 5) = P(-1.5 < Z < 1) = 0.8413 - 0.0668 = 77.4\%$$

Calcolo probabilità per una Gaussiana generica

- Esercizi
- Sia Y una variabile aleatoria di tipo normale, di media $\mu = 16$ e varianza $\sigma^2 = 25$
- Calcolare:
  - P(Y > 20)
  - P(20 < Y < 25)
  - P(Y < 10)
  - P(12 < Y < 24)

45

Teorema del limite centrale

- **Teorema del limite centrale**
- Sia $y_1, y_2, \ldots, y_n$ una successione di n VA indipendenti ed identicamente distribuite tali che $E[y_i]=\mu$ e $V(y_i)=\sigma^2$.
- Sia inoltre $x_n=y_1+y_2+\ldots+y_n$
- Allora:

$$Z_n = \frac{X_n - n\mu}{\sqrt{n\sigma^2}}$$

- tende ad una **VA Gaussiana di tipo standard** per n $\rightarrow\infty$

46

Teorema del limite centrale

- **VA Gaussiana è ideale per descrivere l'errore sperimentale**
- La VA di tipo normale è un valido modello matematico per descrivere le incertezze presenti nella misura sperimentale
  - È ragionevole assumere che le deviazioni dal valore vero provengano da diverse fonti indipendenti

47

Variabili Aleatorie derivate dalla gaussiana - Variabile χ^2

- Si considerino k VA di tipo Standard indipendenti z_1, z_2, …. z_k
- La variabile aleatoria scalare

$$X = Z_1^2 + Z_2^2 + \ldots + Z_k^2$$

- prende il nome di variabile aleatoria χ^2 a *k* gradi di libertà.
- Tale variabile aleatoria è caratterizzata completamente da un solo parametro, il numeri di gradi di libertà *k*.
- La pdf ha espressione:

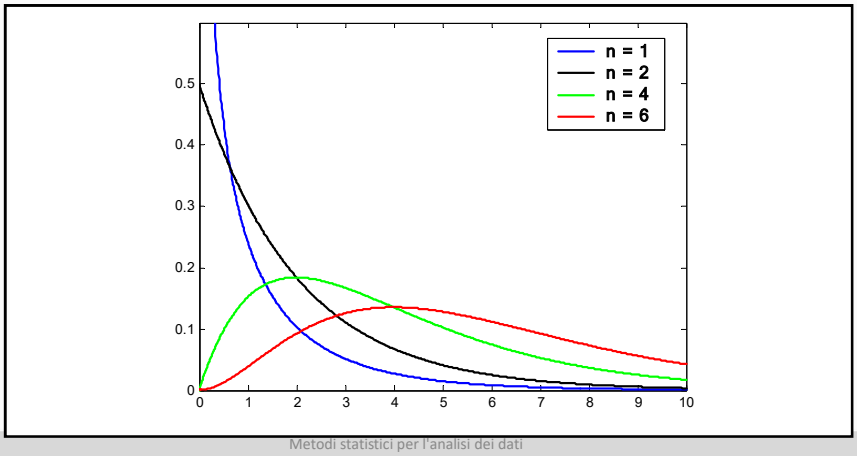
$$f(x) = \frac{1}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} \exp\left(-\frac{x}{2}\right) \qquad x > 0$$

48

Variabile aleatoria χ^2

Richiami di statistica – Esperimenti replicati

- Funzione densità di probabilità



49

Variabile aleatoria χ^2

Richiami di statistica – Esperimenti replicati

- Proprietà di una variabile aleatoria χ^2 a *k* gradi di libertà

$$\mu = k$$
$$\sigma^2 = 2k$$

- Il massimo si osserva in corrispondenza di $y = n-2$.
- Per $n \to \infty$ la distribuzione χ^2 tende ad una gaussiana.

50

Variabile aleatoria χ^2 – Esempio

- Esempio di VA che segue la distribuzione di tipo χ^2:
- Siano y_1, y_2, …, y_n un campione di dati generati da una VA di tipo Gaussiano $N(\mu,\sigma^2)$. Allora:

$$\frac{SS}{\sigma^2} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sigma^2} \sim \chi^2_{n-1}$$

- Da cui, con semplici passaggi, si può ricavare la seguente relazione per la stima S^2 della varianza:

$$S^2 = \frac{SS}{n-1} \Rightarrow (n-1)\frac{S^2}{\sigma^2} \approx \chi^2_{n-1}$$

51

VA derivate dalla gaussiana Distribuzione T-student

- Siano dati una variabile aleatoria Z Gaussiana di tipo standard (ovvero $Z \sim \mathcal{N}(0,1)$, ed una χ^2 ad r gradi di libertà
- La variabile aleatoria :

$$T_r = \frac{Z}{\sqrt{\frac{\chi^2_r}{r}}}$$

è una distribuzione *T di student ad r gradi di libertà*.

52

VA derivate dalla Gaussiana
Distribuzione T di student

- Espressione analitica della T di student

$$f_r(y) = \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{r\pi}\,\Gamma(r/2)} \frac{1}{\left(\frac{y^2}{r}+1\right)^{\frac{r+1}{2}}} \qquad -\infty < y < \infty$$

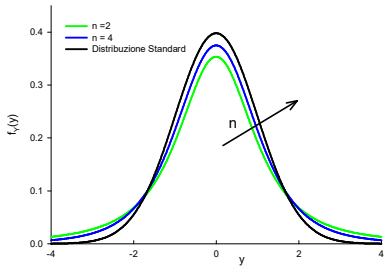
- Proprietà:
- Dipende da un solo parametro il numero intero *r*

Media: $\mu_{t,r} = 0$

Varianza: $\sigma^2_{t,r} = \frac{r}{r-2} \qquad (r > 2)$

VA derivate dalla Gaussiana
Distribuzione T di student

- In figura sono mostrate le funzioni densità per 1,3,6 gradi di libertà.



William Gosset "creatore" della T di student

- La T è simmetrica rispetto a y=0
- Per r →+∞ la T di student tende ad una gaussiana di tipo standard.

Variabile aleatoria di tipo t di student – Esempio

- Esempio di VA che segue la distribuzione di tipo t di student:
- Siano y_1, y_2, …, y_n un campione di dati generati da una VA di tipo Gaussiano $N(\mu,\sigma^2)$. Allora, la quantità:

$$t = \frac{\bar{y} - \mu}{\sqrt{S^2/n}}$$

- Segue una distribuzione di tipo t di student a (n-1) g.d.l.

55

VA derivate dalla Gaussiana La distribuzione F di Fisher

- Siano Y e W due VA di tipo χ^2 rispettivamente ad *u* e *v* gradi di libertà.
- Il rapporto

$$F_{u,v} = \frac{\chi_u^2 / u}{\chi_v^2 / v}$$

è una VA di tipo **F di Fisher** ad *(u,v)* gradi di libertà.

- La VA ha due parametri, *u* e *v*.

56

VA derivate dalla Gaussiana
La distribuzione F di Fisher

Richiami di statistica – Esperimenti replicati

- Espressione analitica della F di Fisher

$$f(y;u,v)=\frac{\Gamma\left(\frac{u+v}{2}\right)}{\Gamma\left(\frac{u}{2}\right)\Gamma\left(\frac{v}{2}\right)}\left(\frac{u}{v}\right)^{n/2}\frac{y^{\frac{u-2}{2}}}{\left(1+\left(\frac{u}{v}\right)y\right)^{\frac{u+v}{2}}} \qquad 0<y<\infty$$

Media: $\mu_F=\frac{v}{v-2}, \qquad (v>2)$

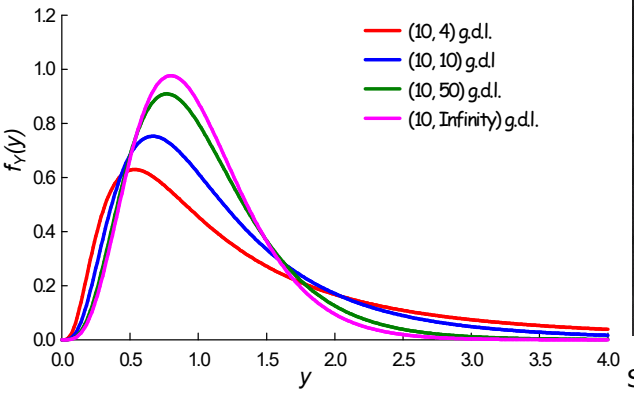
Varianza: $\sigma_F^2=\frac{2\,v^2\,(u+v\text{-}2)}{u(v\text{-}4)\,(v\text{-}2)^2} \qquad v>4$

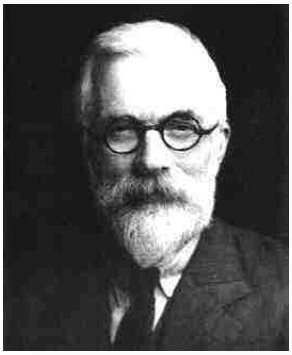
57

VA derivate dalla Gaussiana
La distribuzione F di Fisher

Richiami di statistica – Esperimenti replicati

- Grafici della F di Fisher al variare dei gradi di libertà





Sir Ronald Aylmer Fisher
1890 - 1962

58

Variabile aleatoria di tipo F di Fisher – Esempio

- Esempio di VA che segue la distribuzione di tipo F di Fisher:
- Siano:
  - $y_{1,1}$, $y_{1,2}$, …, $y_{1,n1}$ un campione di n_1 osservazioni provenienti da una data popolazione
  - $y_{2,1}$, $y_{2,2}$, …, $y_{2,n2}$ un campione di n_2 osservazioni provenienti da una altra popolazione
- Si suppone inoltre che la varianza σ^2 sia la stessa per entrambe le popolazioni. Allora:

$$\frac{S_1^2}{S_2^2} \approx F_{n1-1,n2-1}$$

- Dove S_1^2 e S_2^2 sono le due varianze campionarie calcolate per i due campioni

Analisi del campione di dati con strumenti statistici – Ulteriori sviluppi

- Nei prossimi lucidi si illustreranno delle tecniche utili per approfondire ulteriormente la conoscenza del campione di dati a disposizione:
  - **Test statistici delle ipotesi**
  - **Intervalli di fiducia**

Analisi del campione di dati con strumenti statistici – Ulteriori sviluppi

Richiami di statistica – Esperimenti replicati

- **Modello statistico per il campione di dati:**

$$y_j = \mu + \varepsilon_j \qquad \varepsilon_j \approx N(0, \sigma^2) \qquad j = 1,2,\ldots,n$$

y_i
j-esima osservazione sperimentale

μ
Media della risposta Costante

ε_j
Variabile aleatoria normale associata con la j-esima osservazione

61

Test delle ipotesi – Introduzione

Richiami di statistica – Esperimenti replicati

- Torniamo al campione sperimentale di prodotti alimentari dell'esempio introduttivo.
- Da pregressi studi sull'impianto si sa che nella linea produttiva non sono graditi materiali troppo viscosi (motivi: perdite di carico, costi elevati di esercizio etc.).
- Da pregresse analisi si è stabilito un **valore di soglia** per la viscosità:

$$\nu = 72.5$$

- al di sopra del quale risulta difficile la lavorazione del prodotto.

62

Test delle ipotesi - Definizione

- Un'**ipotesi statistica** è un'assunzione che noi facciamo sui parametri di una distribuzione o, equivalentemente, di un modello.
- L'ipotesi riflette qualche **congettura** sul problema in esame.
- Nel caso dell'esempio introduttivo, si vuole stabilire se
  - la viscosità della crema possa essere **almeno pari** al valore critico oppure
  - vi è una differenza **significativa** rispetto al valore ν=72.5.

Test statistici – Definizione del problema

- Un **test statistico** di un'ipotesi è una procedura in cui si conclude se è possibile *non rigettare l'ipotesi* (cioè non si può escludere che essa sia vera) oppure *rigettare l'ipotesi.*
  - Si usa un campione e si cerca di concludere se tale campione è compatibile o meno con l'ipotesi nulla di partenza.
- Nell'esempio preso in considerazione, si vuole testare se il campione sperimentale possa derivare da una variabile aleatoria di media ν = 72.5

Test delle Ipotesi - Ipotesi nulla

- Il test delle ipotesi richiede l'introduzione di una **ipotesi nulla H_0**:

$$H_0: \ \mu = \mu_0 = 72.5$$

- In alternativa è possibile che la viscosità sia **realmente** minore del valore di soglia. Questa ipotesi, in **contrasto con l'ipotesi nulla**, è l'**ipotesi alternativa H_1**:

$$H_1: \ \mu < \mu_0 = 72.5$$

- Tutti i test delle ipotesi statistici richiedono la formulazione di un'ipotesi nulla e di un'ipotesi alternativa
- L'ipotesi nulla e l'ipotesi alternativa sono **esaustive** e **mutuamente esclusive**.

Test statistici – Errori che si possono commettere nella procedura

- **Errore di tipo I (o errore α)**
- Probabilità di rigettare l'ipotesi nulla nonostante essa fosse vera

α=P(errore di tipo I)=P(rigetto H_0|H_0 è vera)

- è anche il **livello di significatività** del test.

- **Errore di tipo II (o errore β)**
- Probabilità di *non* rigettare l'ipotesi nulla nonostante essa fosse falsa

β=P(errore di tipo II)=P(non rigetto H_0|H_0 è falsa)

Test statistici – Sviluppo della procedura

- Parte della procedura consiste nel calcolo dell'insieme di valori che portano al rigetto di H_0.
- Tale insieme di valori prende il nome di **regione critica** o **regione di rigetto** del test.

67

Test statistici – Caso varianza σ^2 nota – Ricetta 1/4

- N.B. Tale eventualità non è solo di interesse didattico: l'incertezza presente nelle misure sperimentali può essere nota a priori, per esempio da pregresse misure.
- Per l'esempio si assume $\sigma^2=1$

1. Scegliere un **livello di significatività** α del test (in genere $\alpha=0.05$)
2. Calcolare il **valore critico** z_α tale che:

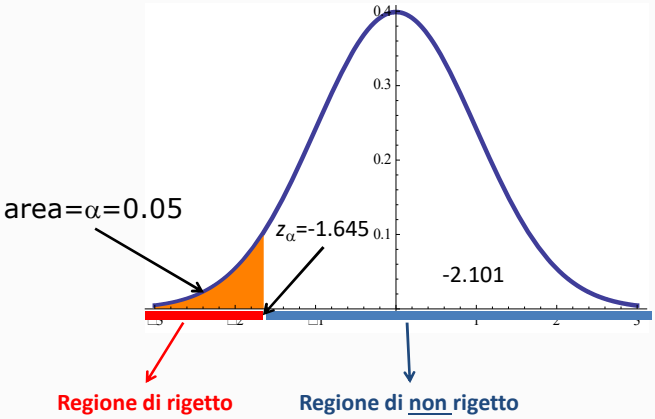
$$P(Z \leq z_\alpha) = \alpha$$

- Nel caso in esame, per $\alpha=0.05$ si può leggere dalle tabelle $z_\alpha=-1.64485$

68

Test statistici – Esempio: Caso varianza nota – Ricetta 2/4

- Distribuzione normale di tipo standard con l'evidenzia delle regioni critiche



69

Test statistici – Esempio: Caso varianza nota – Ricetta 3/4

- Calcolare

$$z_0 = \frac{\bar{y} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}$$

- Dove:
  - $\bar{y}$ è la media campionaria
  - σ^2 è la varianza dell'errore sperimentale
  - n è la dimensione del campione

70

Test statistici – Esempio: Caso varianza nota – Ricetta 3/4

- Si confronta il valore di z_0 osservato con il valore critico z_α

$$z_0 > z_a$$

- **non** rigettiamo l'ipotesi nulla H_0: non si hanno evidenze sperimentali tali da affermare che la media sia **significativamente minore del valore di riferimento**

$$z_0 < z_\alpha$$

- Si **rigetta** l'ipotesi nulla: la media è **significativamente minore** di μ_0.
- Il «rischio» di affermare la conclusione sbagliata è pari al livello di significatività α del test

71

Test delle Ipotesi sulla media - Teoria

- Caso varianza σ^2 nota
- Se l'ipotesi nulla

$$H_0 : \mu = \mu_0$$

- **fosse vera**, la variabile aleatoria media campionaria

$$\overline{Y} = \frac{\sum Y_i}{n}$$

- si comporterebbe come una distribuzione gaussiana di media μ_0 e varianza σ^2/n

$$\overline{Y} \approx N\left(\mu_0, \frac{\sigma^2}{n}\right)$$

72

Test delle Ipotesi sulla media - Teoria

Richiami di statistica – Esperimenti replicati

- Pertanto, se H_0 fosse vera, la variabile aleatoria

$$Z = \frac{\overline{Y} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}$$

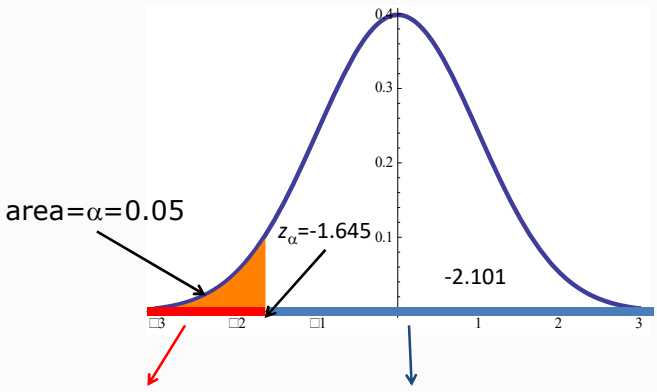
- sarebbe **una distribuzione normale di tipo standard** e il valore osservato z_0 sarebbe un esito che rispetta tale VA.

73

Test delle Ipotesi sulla media - Teoria

Richiami di statistica – Esperimenti replicati

- Al di sopra di z_α è poco plausibile che la variabile aleatoria Z assuma valori



74

Test delle ipotesi sulla media - Esempio

- Si consideri di nuovo l'esempio.
- Il test delle ipotesi è sul valore medio:

$$H_0: \ \mu = \mu_0$$
$$H_1: \ \mu < \mu_0$$

- Con un livello di significatività a = 5 %

75

Test delle ipotesi sulla media - Esempio

- Si valuta innanzitutto il valore z_α tale che $P(Z<z_\alpha)=\alpha=0.05$.

$$P(Z < z_\alpha) = \alpha \quad \Rightarrow \quad z_\alpha = -1.645$$

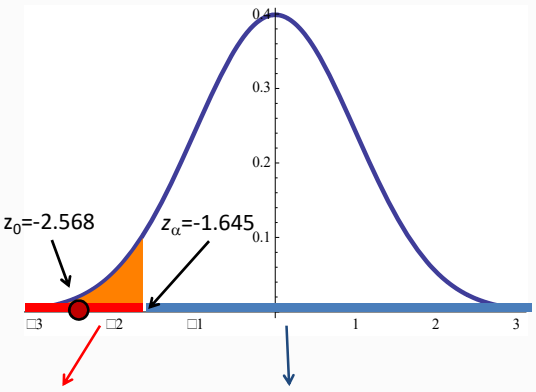
- **Se** l'ipotesi nulla **fosse vera**, il risultato

$$z_0 = \frac{\bar{y} - \mu_0}{\sigma}\sqrt{n} = \frac{71.69 - 72.5}{1}\sqrt{10} = -2.568$$

- **sarebbe** un valore osservato di una variabile aleatoria **normale di tipo standard**.

76

Test delle ipotesi sulla media - Esempio



Regione di rigetto

Regione di non rigetto

Il valore osservato z_0 rientra nella regione in cui la variabile aleatoria ha poche probabilità di cadere

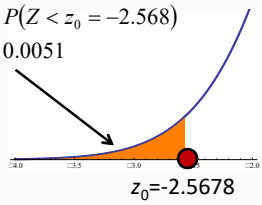
C'è un 5% di probabilità che il valore osservato appartenga alla VA supposta nell'ipotesi nulla H_0 e sia comunque rigettata

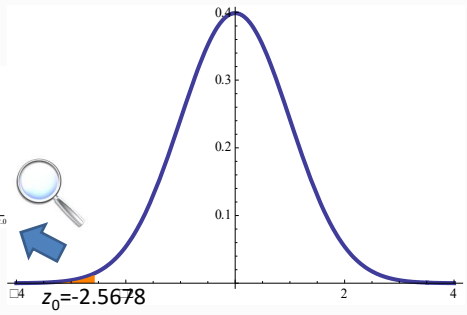
77

Test statistici – Uso del p-value

- Approccio **alternativo** a quello classico dell'individuazione delle zone di rigetto.
- Il **p-value** rappresenta la probabilità che la statistica test stimata assuma un valore almeno uguale al valore osservato della statistica nel caso in cui l'ipotesi nulla fosse vera.
- Nel caso dell'esempio:

$$p-value = P(Z < z_0 = -2.568) = 0.0051$$





78

Test statistici – Uso del p-value

- **Pro**
- Informazione più quantitativa
- **Contro:**
- Necessita di calcolatori con programmi specifici (o comunque competenze di programmazione avanzata)

79

Test delle Ipotesi - Ipotesi alternative 1/4

- Nel problema in esame si assume che il nostro campione di dati sperimentali sia caratterizzato da una variabile aleatoria che abbia una funzione densità di probabilità che coinvolge un parametro ignoto θ e si assume l'ipotesi nulla che

  $H_0 : \ \theta = \theta_0$

- L'ipotesi alternativa era:

  $H_1 : \ \theta < \theta_0$

- Ma non è l'unica alternativa che possiamo considerare.

80

Test delle Ipotesi - Ipotesi alternative 2/4

- In altri casi la natura può suggerire altri tipi di alternative:

$$H_1: \ \theta > \theta_0$$

- Oppure

$$H_1: \ \theta \neq \theta_0$$

- Le prime 2 alternative si chiamano **one-sided**. L'ultima **two-sided**

81

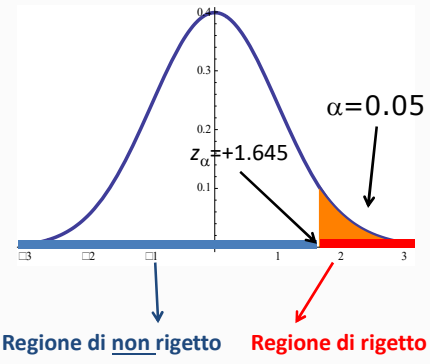
Test delle Ipotesi – Ipotesi alternative 3/4

- Nel caso di ipotesi alternativa

$$H_1: \ \theta > \theta_0$$

- Si deve determinare il valore critico z_α tale che tutti i valori superiori ad esso abbiano una probabilità di verificarsi pari a α
- Dobbiamo escludere i valori per cui la distribuzione gaussiana standard assume valori tali che

$$P(Z > z_\alpha) = \alpha$$



82

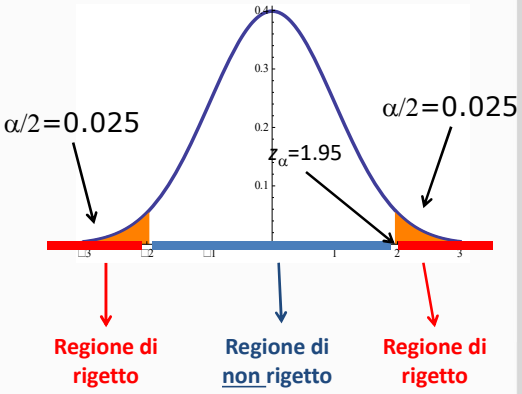
Test delle Ipotesi – Ipotesi alternative 4/4

- Nel caso di ipotesi alternativa

$$H_1: \quad \theta \neq \theta_0$$

- Ricordiamo che è una ipotesi alternativa «two-sided»
- Si deve determinare il valore critico z_α tale che

$$P(|Z| > z_\alpha) = \alpha$$



83

Test delle ipotesi sulla media - Varianza ignota

- Nel caso in cui non fosse nota la varianza s2 non è possibile sfruttare la statistica per determinare i valori critici dei test statistici

$$z = \sqrt{n}\frac{\bar{Y} - \mu_0}{\sigma}$$

- È possibile ricorrere alla stima S^2 della varianza campionaria

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

- Se l'ipotesi nulla fosse vera, allora la variabile aleatoria

$$t = \frac{\bar{Y} - \mu_0}{\sqrt{S^2/n}}$$

- Sarebbe una distribuzione t di student ad (n-1) gdl.

84

Test delle ipotesi sulla media – Varianza ignota

- Ricetta
- Fissare un livello di significatività del test (es: α= 5%)
- Calcolare il valore t_a per cui:

$$P(t \le t_\alpha) = \alpha$$

- dove t è la distribuzione di student ad r=n-1 gradi di libertà.
- Calcolare S^2:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

- Calcolare

$$t_0 = \sqrt{n}\,\frac{\bar{y} - \mu_0}{\sqrt{S^2}}$$

  - $t_0 < t_a$: rigettare H_0
  - $t_0 > t_a$: non rigettare H_0.

85

Test delle ipotesi sulla media – Varianza ignota – Esercizio

- Ritorniamo al campione in esame
- Si fissa un livello di significatività α=0.05 per il test
- Dalle tabelle si determina il valore t_α:

$$P(t \le t_\alpha) = 0.05 \Rightarrow t_{\alpha,19} = -1.833$$

- Si calcola il valore stimato per la varianza:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = 0.983$$

- Da cui è possibile calcolare la statistica t_0:

$$t_0 = \sqrt{n}\,\frac{\bar{y} - \mu_0}{\sqrt{S^2}} = -2.589$$

86

Test delle ipotesi sulla media – Varianza ignota – Esercizio

- Quindi

$$t_0 < t_{\alpha,19}$$

- Si **rigetta** l'ipotesi nulla.
- Alternativamente, è possibile calcolare il **p-value**

$$P(t_r \le t_0 = -2.59) = 0.0146$$

- Da notare come il p-value sia più elevato rispetto a quello stimato nel caso della varianza nota
  - La mancanza di informazioni sul processo si riflette in delle conclusioni più incerte.

87

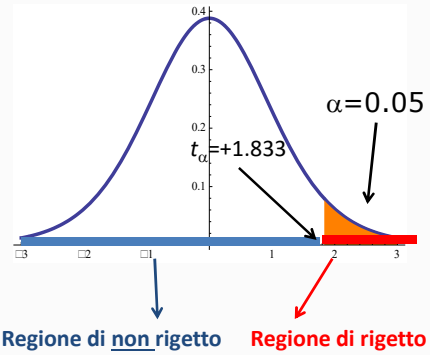
Test delle ipotesi sulla media – Altre ipotesi alternative

- Nel caso di ipotesi alternativa

$$H_1: \ \mu > \mu_0$$

- Si deve determinare il valore critico t_α tale che tutti i valori superiori ad esso abbiano una probabilità di verificarsi pari a α
- Dobbiamo escludere i valori per cui la t di student assuma valori tali che

$$P(t_r > t_\alpha) = \alpha \quad (r = 9\,gdl)$$



88

Test delle ipotesi sulla media – Altre ipotesi alternative

Richiami di statistica – Esperimenti replicati

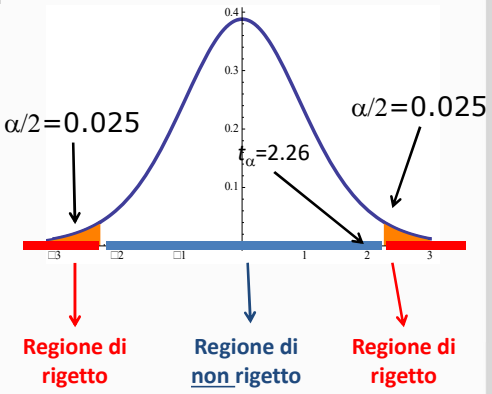
- Nel caso di ipotesi alternativa «two sided»

$$H_1: \ \mu \neq \mu_0$$

- Si deve determinare il valore critico z_α per cui

$$P(|T| > t_{r,\alpha}) = \alpha$$

$$P(t_r > t_\alpha) = \alpha \quad (r = 9\, gdl)$$



89

Intervalli di fiducia - Introduzione

Richiami di statistica – Esperimenti replicati

- Nell'esaminare un campione di dati sperimentali, si può essere interessati ad un'informazione più qualitativa di una semplice stima puntuale di parametri.
- Ad esempio, si può essere interessati a determinare un **intervallo** di valori in cui è molto probabile cada il valore vero del parametro.
- Tale tipo di inferenza prende il nome di **inferenza di intervallo** e il risultato della procedura è un **intervallo di fiducia** (anche denominato **intervallo di confidenza**)
- Per esempio, si può essere interessati ad un intervallo di fiducia per la media μ della viscosità.

90

Intervalli di fiducia Procedura

- Si suppone che θ sia il parametro incognito da stimare
- Si sceglie una probabilità γ vicina a 1 (in genere γ=0.95). Tale probabilità prende il nome di **livello di fiducia**.
- In seguito si determinano due quantità L e U tali che

$$P(L \le \theta \le U) = \gamma$$

- L'intervallo di estremi *L* e *U* prende il nome di **intervallo di fiducia** e si indica con il simbolo:

$$CONF\{L \le \theta \le U\}$$

Intervalli di fiducia della Media – Caso varianza non nota.

Determinazione intervallo di fiducia:

1. Scegliere un livello di fiducia $\gamma=1-\alpha$
2. Ricavare (per esempio da tabelle) il valore $t_{\alpha/2}$ tale che:

$$P(-t_{\alpha/2} \le T_r \le t_{\alpha/2}) = \gamma = 1-\alpha$$

essendo T_r la T di student a r=n-1 gdl

3. Calcolare media e varianza del campione dei dati sperimentali.
3. L'intervallo di fiducia per la media sarà:

$$CONF\left\{\bar{y} - t_{\alpha/2}\sqrt{\frac{S^2}{n}} \le \mu \le \bar{y} + t_{\alpha/2}\sqrt{\frac{S^2}{n}}\right\}$$

Intervalli di fiducia della Media nel caso di varianza non nota.

Richiami di statistica – Esperimenti replicati

- La variabile aleatoria:

$$Z = \sqrt{n}\frac{\overline{Y} - \mu}{\sigma}$$

- È una variabile normale di tipo standard
- Si può ulteriormente dimostrare che la variabile aleatoria:

$$W = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 = (n-1)\frac{S^2}{\sigma^2} \approx \chi^2_{n-1}$$

- È una variabile aleatoria χ^2 a *n-1* gradi di libertà

93

Intervalli di fiducia della Media nel caso di varianza non nota.

Richiami di statistica – Esperimenti replicati

- In conclusione la variabile aleatoria:

$$T = \frac{Z}{\sqrt{W/(n-1)}} = \sqrt{n}\frac{\dfrac{\overline{Y} - \mu}{\sigma}}{\sqrt{\dfrac{\sum\left(Y_i - \overline{Y}\right)^2}{\sigma^2}\Big/(n-1)}} = \frac{\overline{Y} - \mu}{\sqrt{\dfrac{S^2}{n}}}$$

- È una variabile aleatoria di tipo t di student ad n-1 gradi di libertà

94

Intervalli di fiducia della Media nel caso di varianza non nota.

- Dalla definizione di probabilità è possibile ricavare la relazione:

$$P\left(-t_{\alpha/2} \le T_r \le t_{\alpha/2}\right) = P\left(-t_{\alpha/2} \le \frac{\bar{y}-\mu}{\sqrt{\frac{S^2}{n_1}}} \le t_{\alpha/2}\right) = \gamma$$

- da cui con qualche passaggio è possibile ricavare l'intervallo di fiducia desiderato:

$$P\left(\bar{y} - t_{\alpha/2} S\sqrt{\frac{1}{n}} \le \mu \le \bar{y} + t_{\alpha/2} S\sqrt{\frac{1}{n}}\right) = \gamma \quad CONF\left\{\underbrace{\bar{y} - t_{\alpha/2} S\sqrt{\frac{1}{n}}}_{L} \le \underbrace{\mu}_{\theta} \le \underbrace{\bar{y} + t_{\alpha/2} S\sqrt{\frac{1}{n}}}_{U}\right\}$$

95

Intervalli di fiducia della Media – Esercizio

Determinazione intervallo di fiducia:

1. Si sceglie un livello di fiducia γ=95%
2. Ricavare il valore $t_{\alpha/2}$ tale che:

$$P\left(-t_{\alpha/2} \le T_r \le t_{\alpha/2}\right) = 95\% \Rightarrow t_{\alpha/2} = 2.262$$

essendo T_r la T di student a r=9 gdl

3. Calcolare media e varianza del campione dei dati sperimentali.

$$\bar{y} = 71.69, \qquad S^2 = 0.9834$$

3. L'intervallo di fiducia per la media sarà:

$$CONF\left\{71.69 - 2.262\sqrt{\frac{0.9834}{10}} \le \mu \le 71.69 + 2.262\sqrt{\frac{0.9834}{10}}\right\} = \left\{70.98 \le \mu \le 72.4\right\}$$

96

Intervalli di fiducia della Media – Esercizio

- Da notare che nell'intervallo di fiducia calcolato non ricade il valore 72.5, confermando che tale valore è molto improbabile per la media della popolazione.
- In generale, si deve ricordare che, per le proprietà di simmetria della t di student:

$$P(T_r \le -t_{\alpha/2}) = P(T_r \ge +t_{\alpha/2})$$

- Il valore di $t_{\alpha/2}$ può essere calcolato anche dalla relazione:

$$P(T_r \le t_{\alpha/2}) = \frac{1}{2}(1+\gamma)$$

97

Diagramma in scala probabilistica

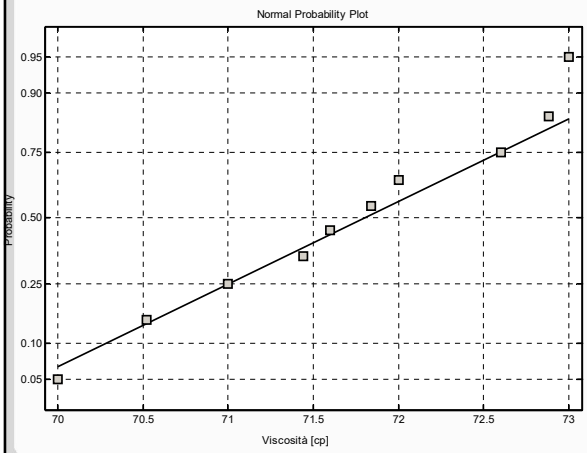
- Da notare che il modello statistico preso in considerazione parte dall'assunzione che i dati sperimentali seguano una distribuzione di tipo Gaussiano.
- Tale assunzione può essere verificata costruendo un **diagramma in scala probabilistica**.
- La procedura è abbastanza semplice e consiste in un'analisi di tipo **grafico**.
- Per costruire il diagramma si deve:
  - ordinare i dati dal più piccolo al più grande
  - le osservazioni così ordinate sono rappresentate rispetto la loro frequenza cumulativa osservata
  - la scala in ordinata non è lineare ma è tale che, se i dati rispettassero una dispersione di tipo **Gaussiano**, essi si disporrebbero approssimativamente **lungo una retta**

98

Diagramma in scala probabilistica

Richiami di statistica – Esperimenti replicati

- Esempio dati crema



- In linea di principio, è possibile implementare il metodo a mano, ma risulta molto pesante.
- La maggior parte dei software di uso comune supportano la rappresentazione su carta probabilistica.

99

Conclusioni – Concetti importanti

Richiami di statistica – Esperimenti replicati

- Esperimento come esito di una variabile aleatoria
  - VA di tipo Gaussiano
- Campagna sperimentale esito di una variabile aleatoria
  - VA di tipo student (o, in casi fortunati, di tipo Gaussiano)
- Con gli strumenti della statistica è possibile inferire conclusioni rigorose sul processo.
- Sono stati introdotti i concetti (verranno ampiamente ripresi nel seguito):
  - Test statistici
  - Intervalli di fiducia
- Diagrammi in scala probabilistica

100