



SAPIENZA
UNIVERSITÀ DI ROMA

INTRODUZIONE AL DOE come strumento di sviluppo prodotto

Francesca Campana
Parte 2
Concetti di base



SAPIENZA
UNIVERSITÀ DI ROMA

CONCETTI STATISTICI DI PARTENZA

- DESCRIPTORI DI UNA VARIABILE RANDOM
- GRAFICI UTILI
- DISTRIBUZIONI CAMPIONARIE
- INFERENZA
- SCELTA DELLA NUMEROSITA' DEL CAMPIONE



Concetti statistici di partenza

Statistica descrittiva

Popolazione

Campione

Riguarda insiemi di dati.

I campioni si considerano estratti “casualmente” come candidati rappresentativi di una popolazione.

Il comportamento del campione è di tipo probabilistico.

Le caratteristiche del campione sono dette **statistiche**:

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

Media e Varianza dei campioni esprimono posizione e dispersione del campione.

Concetti statistici di partenza

Statistica descrittiva

Descrittori del campione:

Per la posizione:

- Media e Mediana

Per la forma:

- Range = max – min
- Varianza o Deviazione Standard
- Quartili o Percentili
- L'intervallo dell'interquartile (IQR)

Mediana e Quartili sono utili nella descrizione di campioni con

“outlier” = dati fuori scala o asimmetrie

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19

grammi	grammi
-8	-20
-7	-7
-6	-6
-5	-5
-5	-5
-4	-4
-3	-3
-3	-3
-2	-2
-1	-1
1	1
1	1
2	2
4	4
5	5
5	5
7	7
7	7
9	9

media

-0.15789

-0.78947

mediana

-1

-1

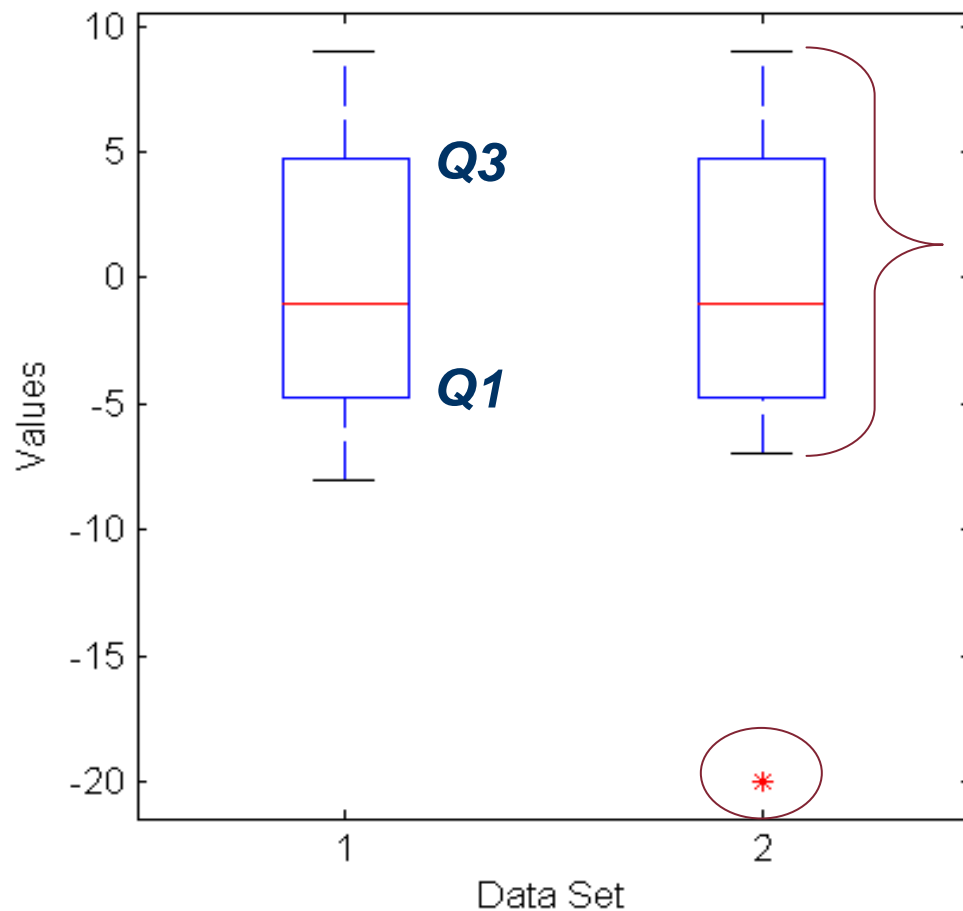
new_media

0.277778

03/05/2012

Concetti statistici di partenza

Statistica descrittiva / BoxPlot



La rappresentazione tramite boxplot sintetizza questi dati

limite per outlier

abbiamo un outlier se il suo valore

è: < 1.5 Primo Quartile

> 1.5 Terzo Quartile

Q0=Min	-8
Q1	-4.5
Q2=Mediana	-1
Q3	4.5
Q4=Max	9

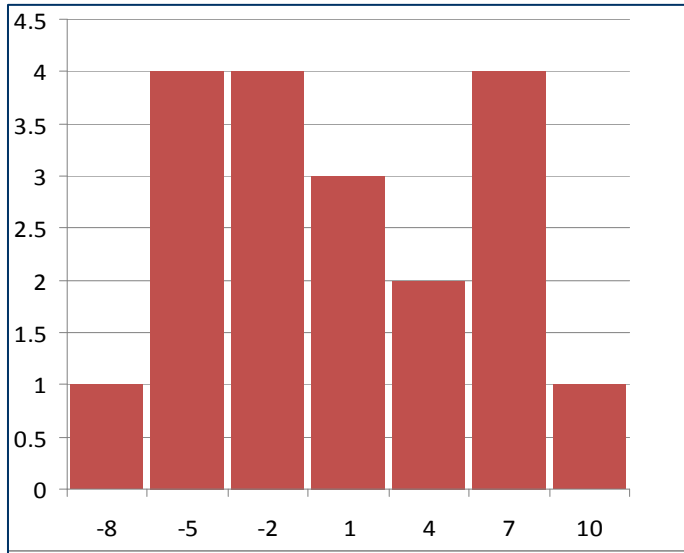
03/05/2012



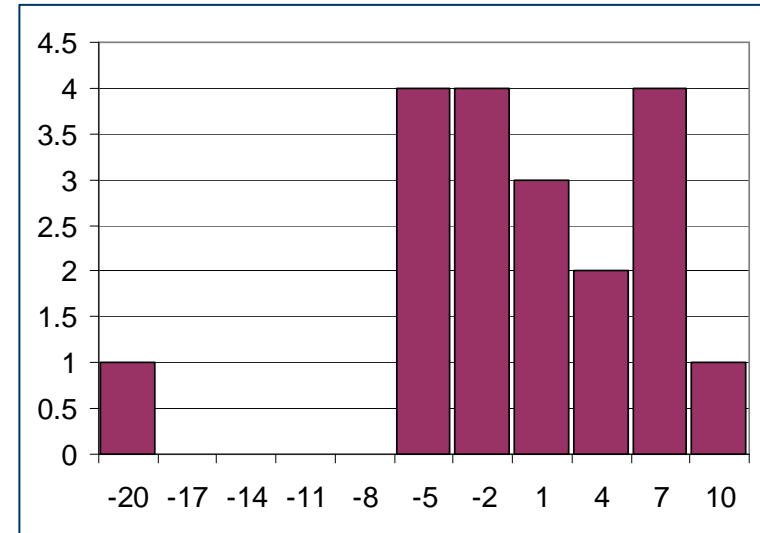
SAPIENZA
UNIVERSITÀ DI ROMA

Concetti statistici di partenza

Statistica descrittiva / ISTOGRAMMA



range	freq
-20	1
-17	0
-14	0
-11	0
-8	0
-5	4
-2	4
1	3
4	2
7	4
10	1



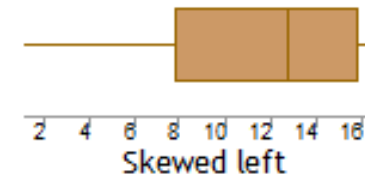
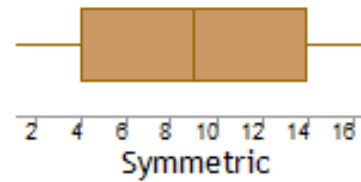
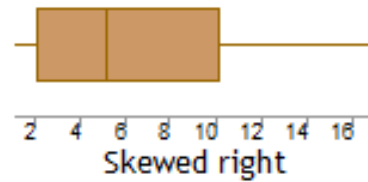
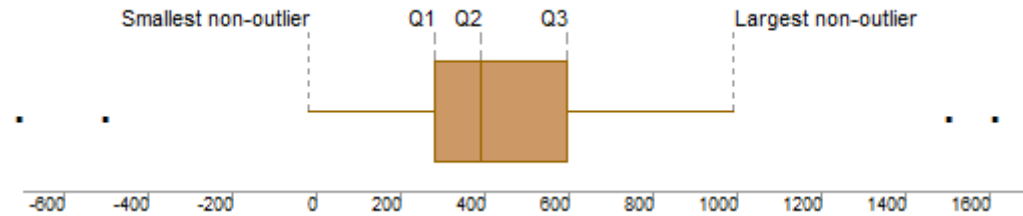
range	freq
-8	1
-5	4
-2	4
1	3
4	2
7	4
10	1

Dati n campioni il numero ottimale di bin per rendere significativo il grafico è \sqrt{n}



Concetti statistici di partenza

Statistica descrittiva / BOXPLOT



Sia il boxplot che l'istogramma delle frequenze mettono in luce la forma del campione (ed eventuali asimmetrie)

03/05/2012



SAPIENZA
UNIVERSITÀ DI ROMA

Concetti statistici di partenza

Statistica descrittiva / Diagrammi di frequenza cumulativa

La frequenza cumulativa consente di stimare la % di elementi che rispettano un dato valore all'interno della distribuzione campionaria



range	freq	cumulativa	c. relativa
-8	1	1	5.26
-5	4	5	26.32
-2	4	9	47.37
1	3	12	63.16
4	2	14	73.68
7	4	18	94.74
10	1	19	100.00



Concetti statistici di partenza

Statistica descrittiva



Sotto opportune ipotesi il campione (e la relativa distribuzione campionaria) riflette gli andamenti della popolazione.

Le funzioni di distribuzione statistica delle popolazioni sono definite dai cosiddetti **parametri**:

$$\mu = E(y) = \begin{cases} \int_{-\infty}^{+\infty} yf(y)dy \\ \sum_{i \rightarrow \infty} y_i p(y_i) \end{cases}$$

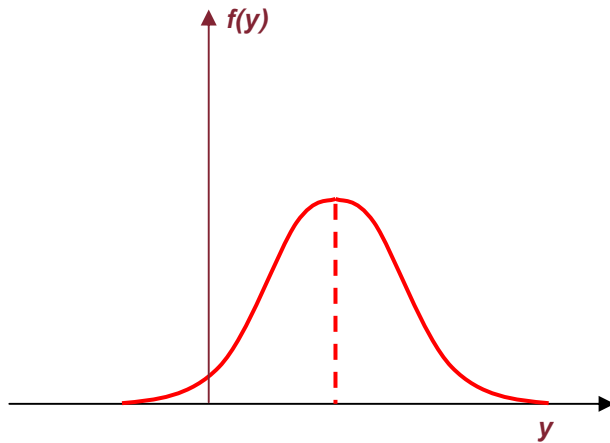
$$\sigma^2 = V(y) = \begin{cases} \int_{-\infty}^{+\infty} (y - \mu)^2 f(y)dy \\ \sum_{i \rightarrow \infty} (y_i - \mu)^2 p(y_i) \end{cases}$$

Concetti statistici di partenza

Statistica descrittiva

Una delle più importanti è la **distribuzione normale**.

Il suo andamento descrive l'errore sperimentale.



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

$$P(\mu \pm 1\sigma) = 68\%$$

$$P(\mu \pm 2\sigma) = 95\%$$

$$P(\mu \pm 3\sigma) = 99,7\%$$

la **distribuzione normale**

standard trasforma la y in una curva normale a media nulla:

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

Z serve a calcolare via tabella l'integrale della distribuzione, quindi la $P(y)$

Concetti statistici di partenza

Statistica descrittiva

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

Z serve a calcolare via tabella l'integrale della distribuzione, quindi la P(y)

Una produzione di sacchetti di carta è caratterizzata da una resistenza distribuita normalmente secondo i parametri:

$$\mu = 275 \text{ kPa} \quad \sigma = 13 \text{ kPa}$$

Si richiede una resistenza di almeno 241 kPa quindi si cerca:

$$P(y \geq 241) = 1 - P(y \leq 241)$$

$$Z = (241 - 275) / 13 = -2.61 \text{ da tabella si trova:}$$

$$P(Z) = 0.995$$

Concetti statistici di partenza

Statistica descrittiva

Una produzione di alberini registra per i diametri i seguenti parametri:

$$\mu=0.2508 \text{ in } \sigma=0.0005 \text{ in}$$

Le specifiche di progetto sono: 0.25 ± 0.0015 in

Quanti alberini cadono nella specifica?

$$Z_{\text{inf}} = -4.6$$

$$Z_{\text{sup}} = 1.4$$

$$P(Z_{\text{sup}}) - P(Z_{\text{inf}}) = 0.9192 - 0.000 = 0.9192$$

91.92% di pezzi soddisfacenti

Concetti statistici di partenza

Statistica descrittiva

Esistono altre funzioni di densità utili in campo ingegneristico:

- **Distribuzione binomiale** per valutare il numero di elementi difettosi nella popolazione (y variabile discreta: “successo”/“insuccesso”):

$$p(Y) = \binom{n}{y} \lambda^y (1-\lambda)^{(n-y)} \quad \lambda \in (0,1) \quad \begin{array}{l} \mu = n\lambda \\ \sigma^2 = n\lambda(1-\lambda) \end{array}$$

- **Distribuzione di Poisson** per valutare elementi difettosi in una unità di prodotto:

$$p(y) = \frac{e^{-\lambda} \lambda^y}{y!} \quad \lambda > 0 \quad \begin{array}{l} \mu = \lambda \\ \sigma^2 = \lambda \end{array}$$

- **Distribuzione esponenziale** per l'affidabilità con tasso di avaria costante:

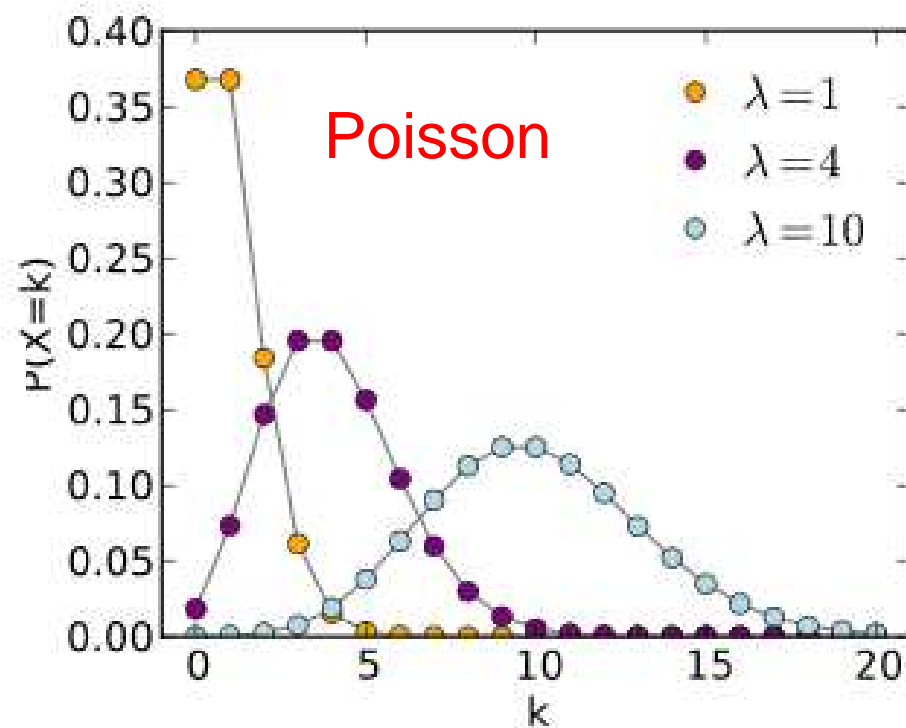
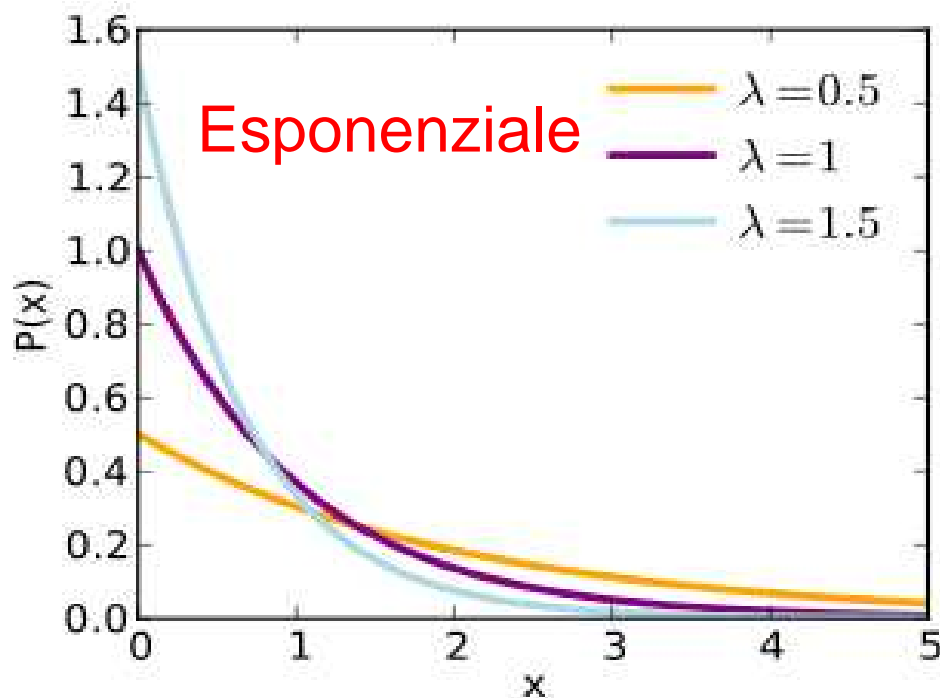
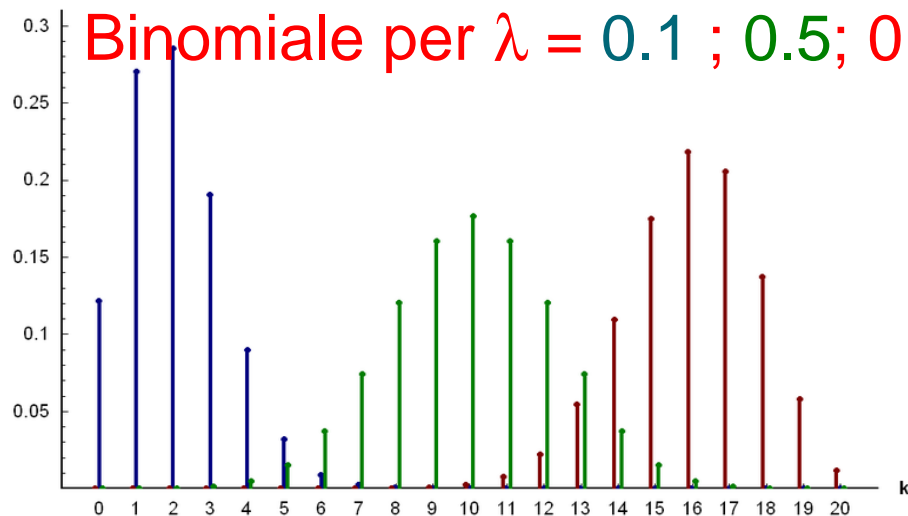
$$p(y) = \lambda e^{-\lambda x} \quad \begin{array}{l} \mu = 1/\lambda \\ \sigma^2 = 1/\lambda^2 \end{array}$$



Concetti statistici di partenza

Statistica descrittiva

Binomiale per $\lambda = 0.1 ; 0.5; 0.9$



03/05/2012



SAPIENZA
UNIVERSITÀ DI ROMA

Concetti statistici di partenza

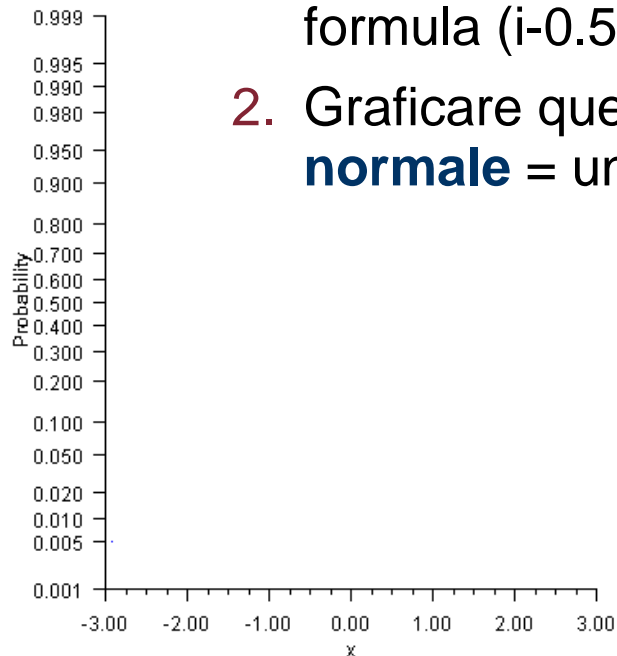
Statistica descrittiva/ PROBABILITY PLOT

Come si comprende il comportamento di un campione?

- Probability plot
- Inferenza statistica

Il Probability Plot diagramma i campioni rispetto alla supposta distribuzione teorica.

1. Ordinare le y_i e calcolare le frequenze cumulate secondo la formula $(i-0.5)/n$
2. Graficare questi valori e le corrispondenti y_i su **una carta normale** = una carta con l'ordinata in scala normale



Se i punti sono su una retta
l'ipotesi è accettabile!

03/05/2012

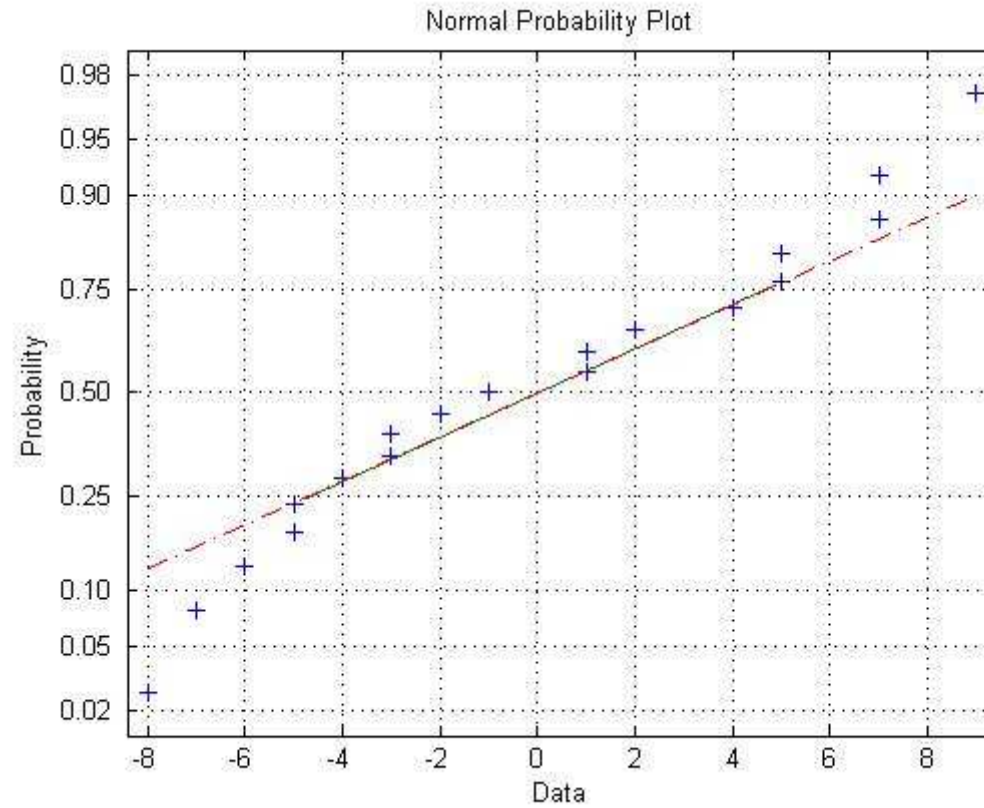


SAPIENZA
UNIVERSITÀ DI ROMA

Concetti statistici di partenza

Statistica descrittiva/ PROBABILITY PLOT

	grammi	$(j-0.5)/19$
1	-8	0.026
2	-7	0.079
3	-6	0.132
4	-5	0.184
5	-5	0.237
6	-4	0.289
7	-3	0.342
8	-3	0.395
9	-2	0.447
10	-1	0.500
11	1	0.553
12	1	0.605
13	2	0.658
14	4	0.711
15	5	0.763
16	5	0.816
17	7	0.868
18	7	0.921
19	9	0.974



03/05/2012

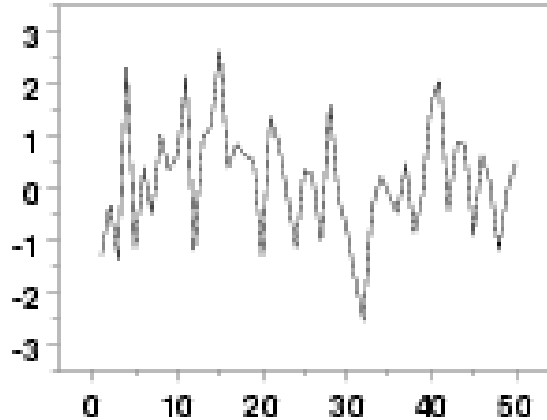


SAPIENZA
UNIVERSITÀ DI ROMA

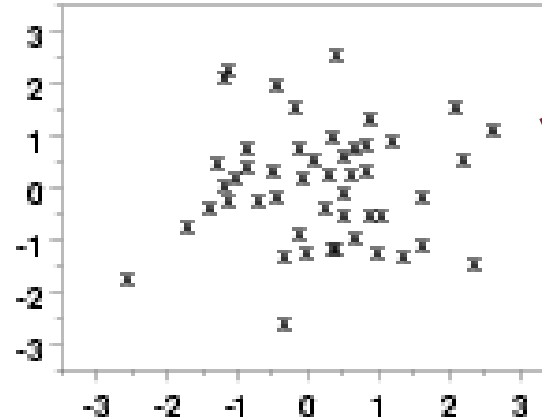
Concetti statistici di partenza

Statistica descrittiva/ Interpretazione grafici

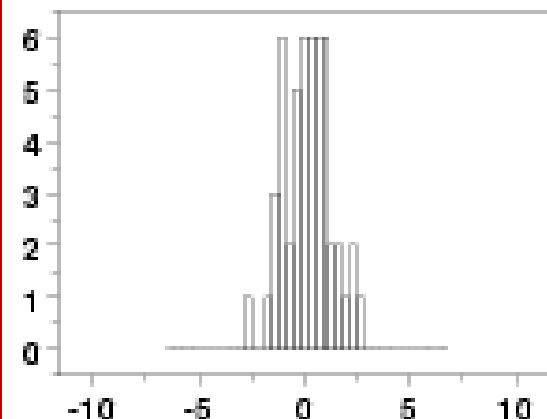
Normal Random Numbers: 4-Plot



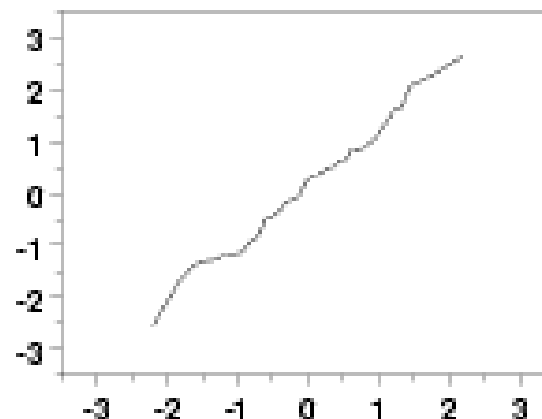
RUNSEQUENCE PLOT Y



LAG PLOT Y



HISTOGRAM Y



NORMAL PROBABILITY PLOT Y

N.B. il **lag plot** studia la correlazione tra dati sperimentali nel tempo.

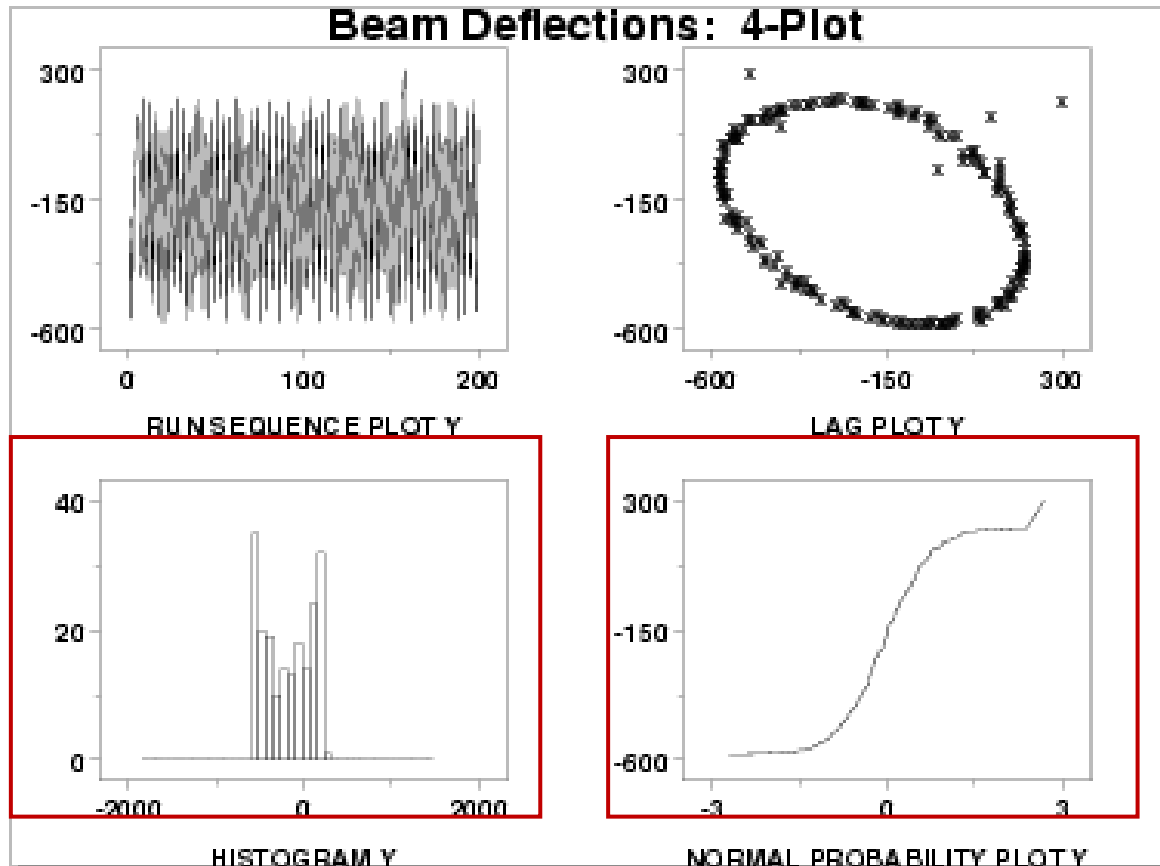
Esso grafica Y_i in funzione di Y_{i-1} , se c'è ciclicità i dati non sono a comportamento casuale.

Dalla forma dell'istogramma delle frequenze e dalla rettilinearità della parte centrale del probability plot l'ipotesi di campione di tipo "normale" (o gaussiano) sembra confermata



Concetti statistici di partenza

Statistica descrittiva/ Interpretazione grafici



Un probability plot ad S è sintomatico di un comportamento uniforme ad istogramma multimodale.

Il lag plot conferma l'ipotesi.

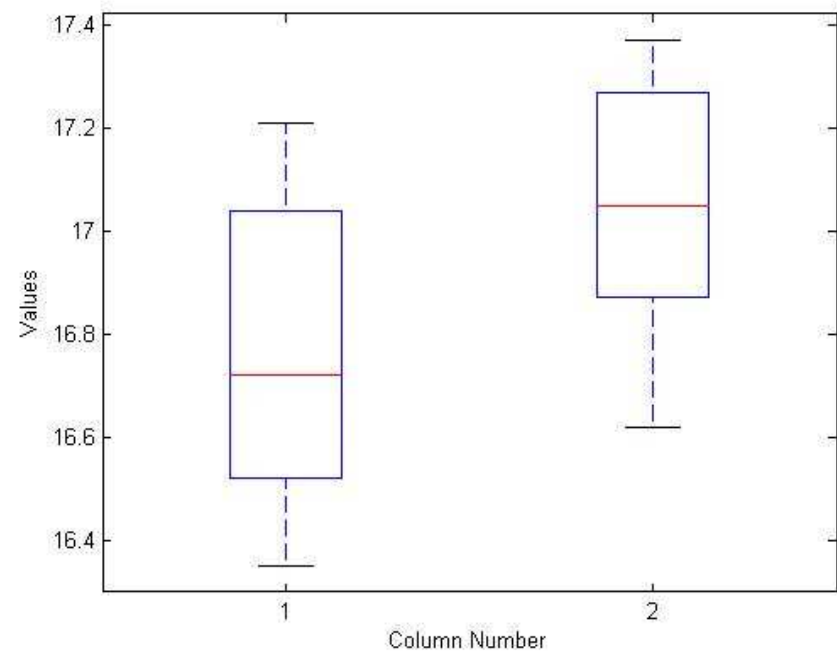
Concetti statistici di partenza

Esempio di riepilogo

Resistenza di due diverse composizioni di calcestruzzo
Testate con 10 replicazioni per composizione

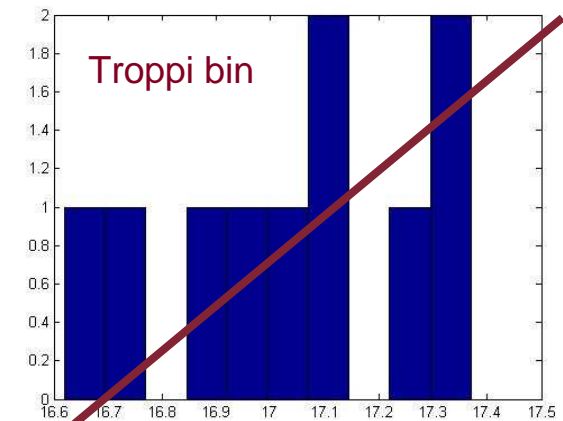
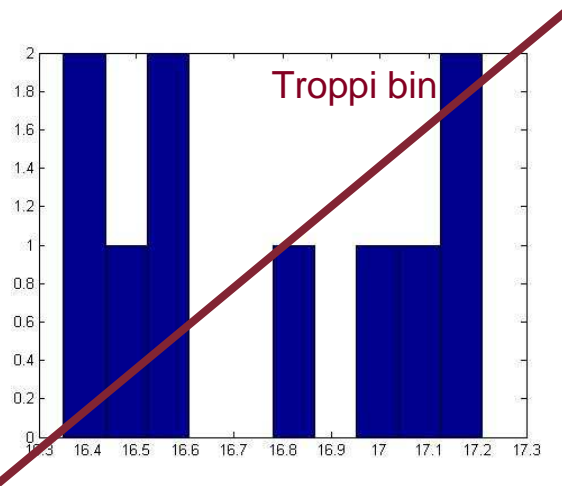
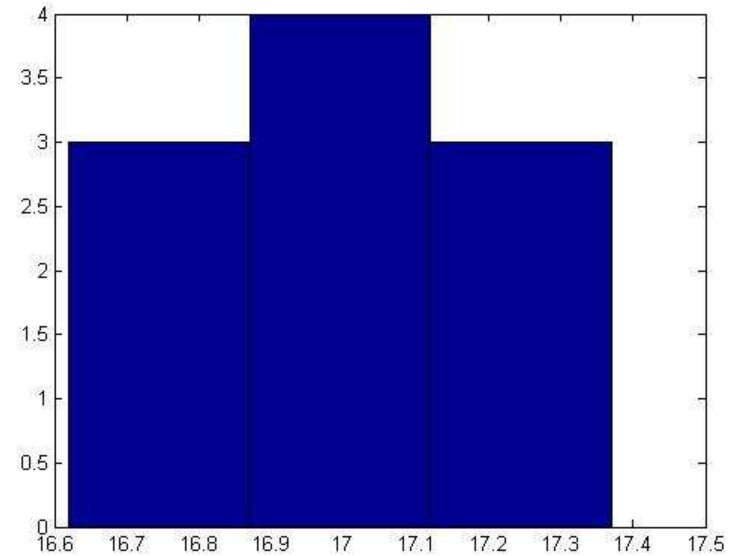
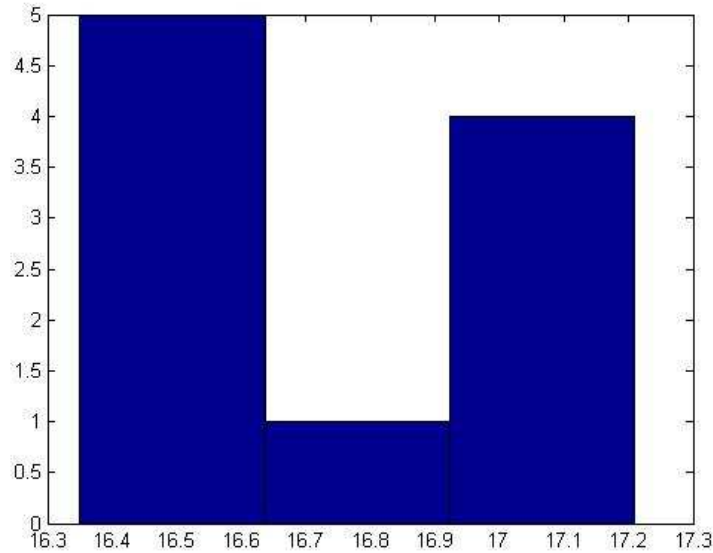
	modified mortar kgf/cm ²	Unmodified kgf/cm ²
1	16.85	16.62
2	16.4	16.75
3	17.21	17.37
4	16.35	17.12
5	16.52	16.98
6	17.04	16.87
7	16.96	17.34
8	17.15	17.02
9	16.59	17.08
10	16.57	17.27

	kgf/cm ²	kgf/cm ²
media	16.76	17.04
var	0.10	0.06
stad.dev	0.32	0.25
Q0	16.35	16.62
Q1	16.53	16.90
Q2	16.72	17.05
Q3	17.02	17.23
Q4	17.21	17.37



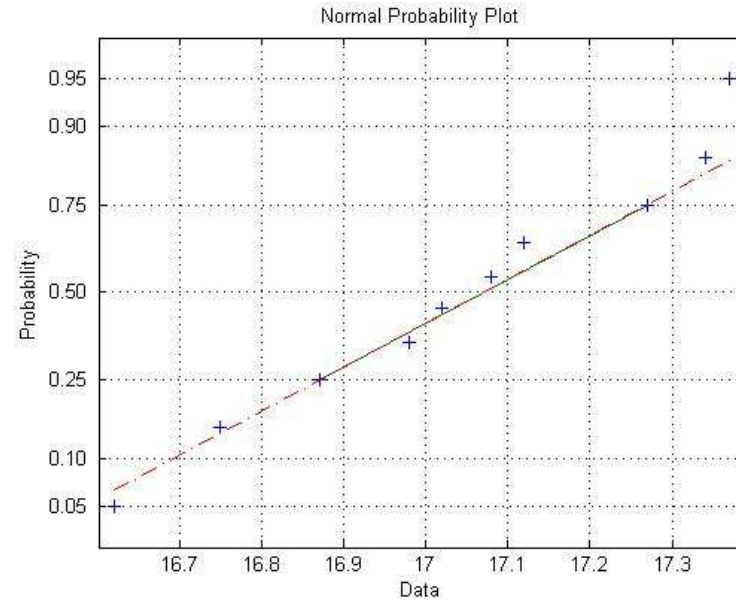
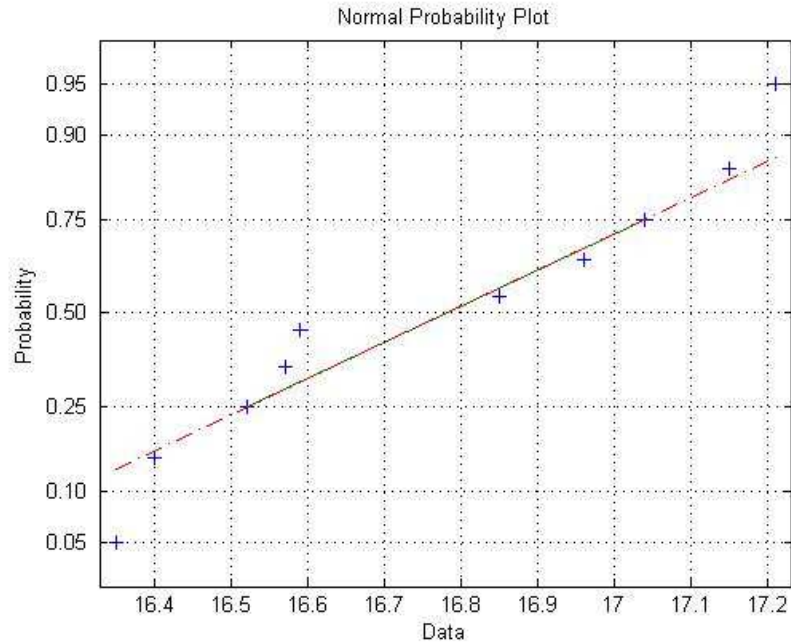
Concetti statistici di partenza

Esempio di riepilogo



Concetti statistici di partenza

Esempio di riepilogo



	kgf/cm ²	kgf/cm ²
media	16.76	17.04
var	0.10	0.06



Concetti statistici di partenza

Inferenza statistica

	modified mortar	Unmodified
1	16.85	16.62
2	16.4	16.75
3	17.21	17.37
4	16.35	17.12
5	16.52	16.98
6	17.04	16.87
7	16.96	17.34
8	17.15	17.02
9	16.59	17.08
10	16.57	17.27

Con quale percentuale di errore possiamo affermare che le due miscele sono uguali? Diverse?

$$y_{1j} = \mu_1 + \varepsilon_{1j} \quad j = 1,10$$

$$y_{2j} = \mu_2 + \varepsilon_{2j} \quad j = 1,10$$



Occorre definire:

- una statistica di test
- un ipotesi di errore

Concetti statistici di partenza

Inferenza statistica

Scelta della statistica del test:

- confrontare le medie assumendo varianze uguali
- inferenza sulle media di una distribuzione normale, varianza incognita
- inferenza sulla varianza

Inferire = confrontare i dati con una ipotesi

Ipotesi nulla H_0

verificate attraverso i valori di α e β

Ipotesi alternativa H_1

$\alpha = P(\text{rifiutare } H_0 \text{ sebbene sia vera}) \rightarrow \text{rischio del produttore}$

$\beta = P(\text{accettare } H_1 \text{ sebbene sia falsa}) \rightarrow \text{rischio dell'acquirente}$

Concetti statistici di partenza

Table 2-3 Tests on Means with Variance Known

Hypothesis	Test Statistic	Criteria for Rejection
$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	$Z_0 = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}$	$ Z_0 > Z_{\alpha/2}$
$H_0: \mu = \mu_0$ $H_1: \mu < \mu_0$		$Z_0 < -Z_\alpha$
$H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$		$Z_0 > Z_\alpha$
$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$	$Z_0 = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	$ Z_0 > Z_{\alpha/2}$
$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$		$Z_0 < -Z_\alpha$
$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$		$Z_0 > Z_\alpha$

Concetti statistici di partenza

Table 2-4 Tests on Means of Normal Distributions, Variance Unknown

Hypothesis	Test Statistic	Criteria for Rejection
$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	$t_0 = \frac{\bar{y} - \mu_0}{S/\sqrt{n}}$	$ t_0 > t_{\alpha/2, n-1}$
$H_0: \mu = \mu_0$ $H_1: \mu < \mu_0$		$t_0 < -t_{\alpha, n-1}$
$H_0: \mu = \mu_0$ $H_1: \mu > \mu_0$		$t_0 > t_{\alpha, n-1}$
if $\sigma_1^2 = \sigma_2^2$		
$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$	$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	$ t_0 > t_{\alpha/2, v}$
$v = n_1 + n_2 - 2$		
if $\sigma_1^2 \neq \sigma_2^2$		
$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 < \mu_2$	$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$	$t_0 < -t_{\alpha, v}$
$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 > \mu_2$	$v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}}$	$t_0 > t_{\alpha, v}$



Concetti statistici di partenza

Table 2-7 Tests on Variances of Normal Distributions

Hypothesis	Test Statistic	Criteria for Rejection
$H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 \neq \sigma_0^2$	$\chi_0^2 = \frac{(n-1)S^2}{\sigma_0^2}$	$\chi_0^2 > \chi_{\alpha/2, n-1}^2$ or $\chi_0^2 < \chi_{1-\alpha/2, n-1}^2$
$H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 < \sigma_0^2$		$\chi_0^2 < \chi_{1-\alpha, n-1}^2$
$H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 > \sigma_0^2$		$\chi_0^2 > \chi_{\alpha, n-1}^2$
$H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 \neq \sigma_2^2$	$F_0 = \frac{S_1^2}{S_2^2}$	$F_0 > F_{\alpha/2, n_1-1, n_2-1}$ or $F_0 < F_{1-\alpha/2, n_1-1, n_2-1}$
$H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 < \sigma_2^2$	$F_0 = \frac{S_2^2}{S_1^2}$	$F_0 > F_{\alpha, n_2-1, n_1-1}$
$H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 > \sigma_2^2$	$F_0 = \frac{S_1^2}{S_2^2}$	$F_0 > F_{\alpha, n_1-1, n_2-1}$

Concetti statistici di partenza

Inferenza statistica/Numero di campioni

Grazie all'inferenza ed al teorema del limite centrale possiamo valutare il numero di campioni sufficiente per una indagine.



Curve Operative

Concetti statistici di partenza

Statistica descrittiva/ Numero di campioni

Se il campione è scelto “casualmente” e la statistica rappresenta il parametro della popolazione allora:

- per n che tende all'infinito media ed $E(y)$ coincidono
- la media di S^2 è la varianza della popolazione ed è il minimo valore possibile tra tutti i possibili campioni

$$\begin{aligned} E(S^2) &= E\left(\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}\right) = \frac{1}{n-1} E\left(\sum_{i=1}^n (y_i - \bar{y})^2\right) \\ &= \frac{1}{n-1} E\left(\sum_{i=1}^n y_i^2 - n\bar{y}^2\right) = \frac{1}{n-1} (n-1)\sigma^2 \end{aligned}$$

Concetti statistici di partenza

Statistica descrittiva/ Sulla scelta del campione

Teorema del limite centrale

per $i = 1, n$ $x_i \sim N(\mu, \sigma)$

se $y = x_1 + \dots + x_n$

$$z_{n \rightarrow \infty} = \frac{y - n\mu}{\sqrt{n\sigma^2}}$$

Se n variabili sono normali anche la loro somma lo è.

- La somma di errori di misura è ancora una distribuzione normale
- Se ogni x_i rappresenta un campione al limite di n abbiamo descritto la popolazione → quanto vale n nella pratica?

$n = 3$ o 4 ? $n = 15$? $n > 40$?

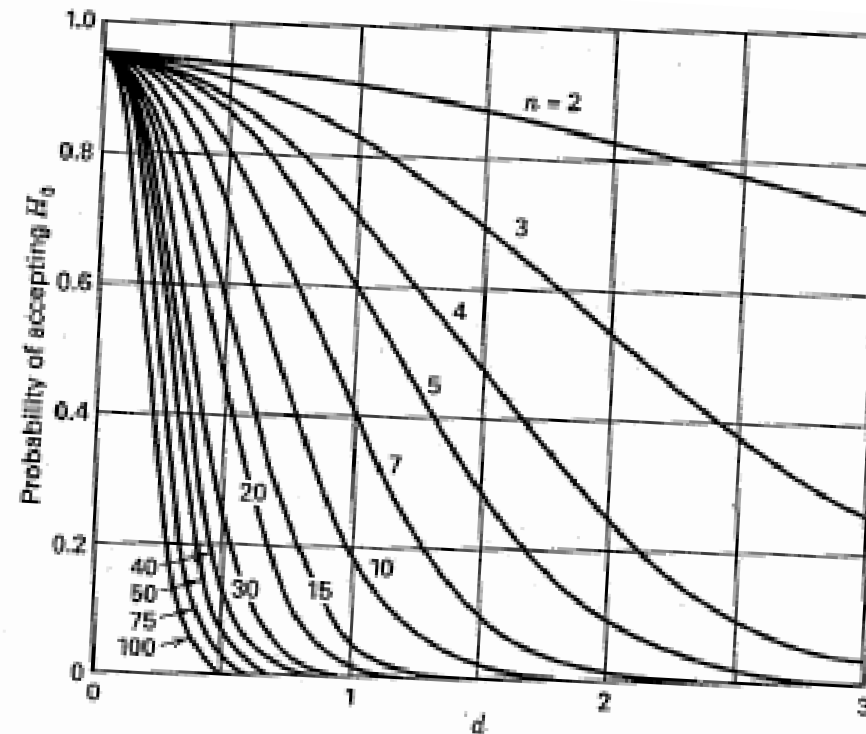
Asimmetria crescente

SPERIENZA
UNIVERSITÀ DI ROMA

Concetti statistici di partenza

Statistica descrittiva/Sulla scelta del campione

α	β	$\Delta=0.5\sigma$	$\Delta=\sigma$	$\Delta=1.5\sigma$
.01	.01	98	25	11
.01	.05	73	18	8
.01	.10	61	15	7
.01	.20	47	12	6
.01	.50	27	7	3
.05	.01	75	19	9
.05	.05	53	13	6
.05	.10	43	11	5
.05	.20	33	8	4
.05	.50	16	4	3
.10	.01	65	16	8
.10	.05	45	11	5
.10	.10	35	9	4
.10	.20	25	7	3
.10	.50	11	3	3
.20	.01	53	14	6
.20	.05	35	9	4
.20	.10	27	7	3
.20	.20	19	5	3
.20	.50	7	3	3



■ FIGURE 2.12 Operating characteristic curves for the two-sided t -test with $\alpha = 0.05$. (Reproduced with permission from "Operating Characteristics for the Common Statistical Tests of Significance," C. L. Ferris, F. E. Grubbs, and C. L. Weaver, *Annals of Mathematical Statistics*, June 1946.)

Concetti statistici di partenza

Statistica descrittiva/Sulla scelta del campione

- La scelta del numero di campioni influisce sulla precisione dell'analisi attraverso l'ascissa della OC = scarto tra risposta attesa e media.
- Nelle carte di controllo n e ΔT di campionamento giocano un ruolo combinato estremamente importante.