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

statistica descrittiva

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

Concetti statistici di partenza

Statistica descrittiva

La Statistica descrittiva 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:

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

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

Elaborare le due colonne come se fossero due scenari possibili di esperimento

Caso A) colonna di sinistra dati senza outlier (rappresentativi)

Caso B) colonna di destra dati con un outlier (quindi dati meno attendibili)

Includendo tutti i dati la media del caso B differisce molto dal caso A – se si esclude l'outlier da -0.789 passa a 0.278

Come si verifica la presenza di outlier? Analizzando i quartili

	Delta_peso (g)	Delta_peso (g)
1	-8	-20
2	-7	-7
3	-6	-6
4	-5	-5
5	-5	-5
6	-4	-4
7	-3	-3
8	-3	-3
9	-2	-2
10	-1	-1
11	1	1
12	1	1
13	2	2
14	4	4
15	5	5
16	5	5
17	7	7
18	7	7
19	9	9
media	-0.15789	-0.78947
mediana	-1	-1
new_media		0.277778

Concetti statistici di partenza

Statistica descrittiva

Come prima cosa occorre ordinare i campione in senso crescente (a prescendire dall'ordine di sperimentazione)

I quartili si ottengono dividendo l'insieme dei campioni in quattro set uguali

La mediana è il valore del campione centrale (o la media dei due valori centrali)

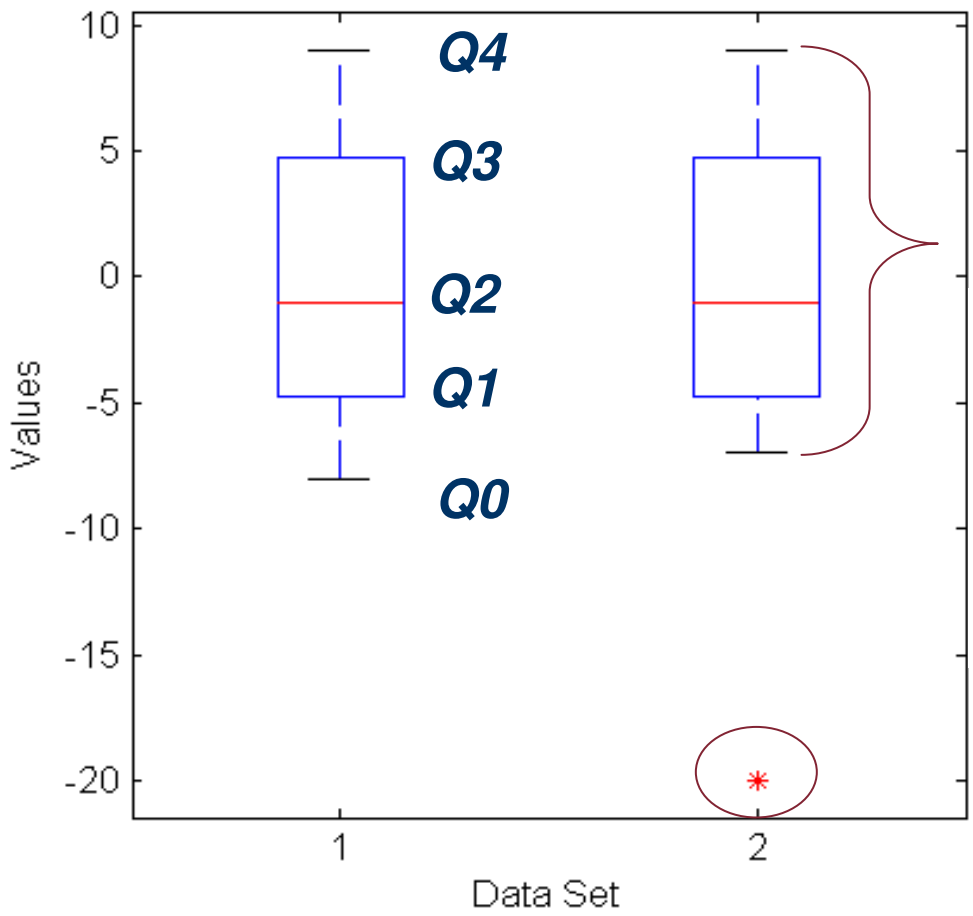
Il primo quartile è il valore limite tra il primo set inferiore e il secondo: Q1=-4.5,

Il terzo quartile è il valore limite tra il terzo set e il quarto: Q1=4.5

	Delta_peso (g)	Delta_peso (g)
1	-8	-20
2	-7	-7
3	-6	-6
4	-5	-5
5	-5	-5
6	-4	-4
7	-3	-3
8	-3	-3
9	-2	-2
10	-1	-1
11	1	1
12	1	1
13	2	2
14	4	4
15	5	5
16	5	5
17	7	7
18	7	7
19	9	9
media	-0.15789	-0.78947
mediana	-1	-1
new_media		0.277778

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

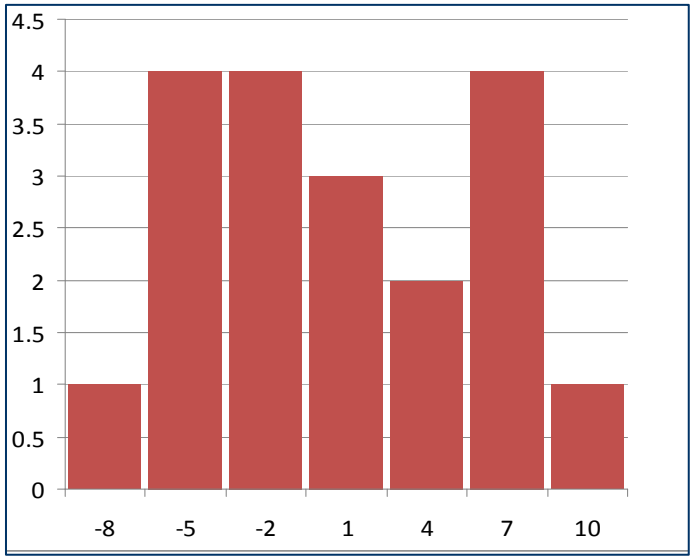
> 1.5 Terzo Quartile

Trovati escludendo il valore/I valori incriminato

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

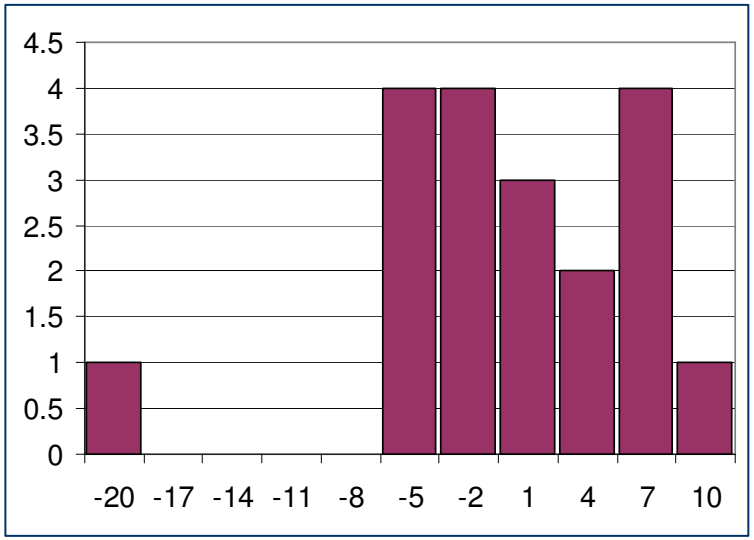
Concetti statistici di partenza

Statistica descrittiva / ISTOGRAMMA



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

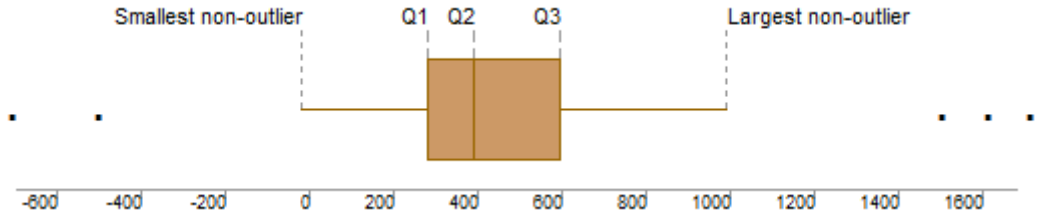
range	freq
-20	1
-17	0
-14	0
-11	0
-8	0
-5	4
-2	4
1	3
4	2
7	4
10	1

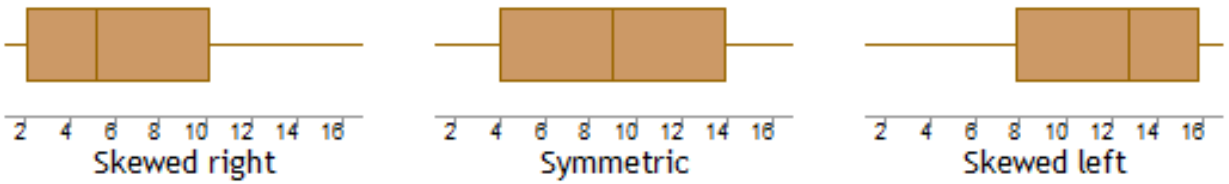


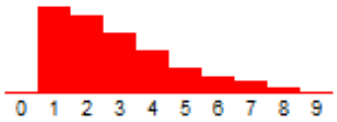
Dati *n* campioni il numero ottimale di bin per rendere significativo il grafico è sqrt(n)

Concetti statistici di partenza

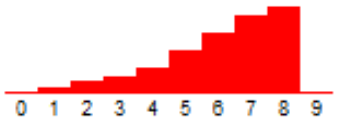
Statistica descrittiva / BOXPLOT







Skewed right



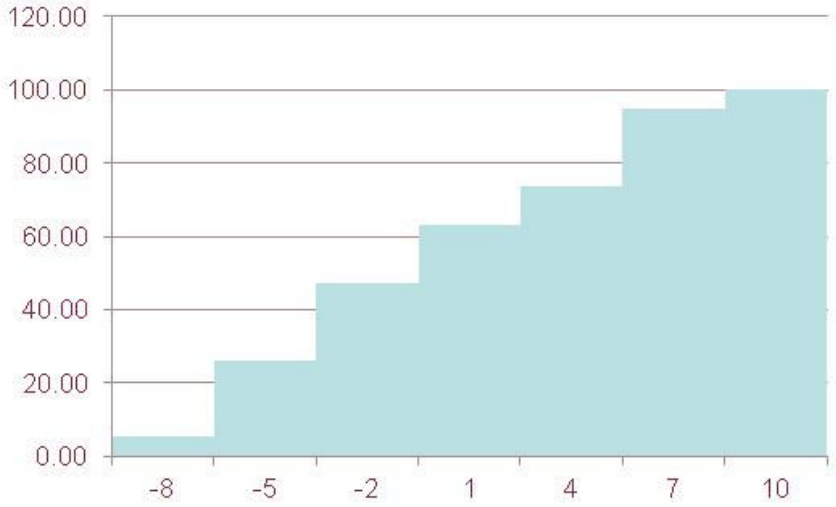
Skewed left

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

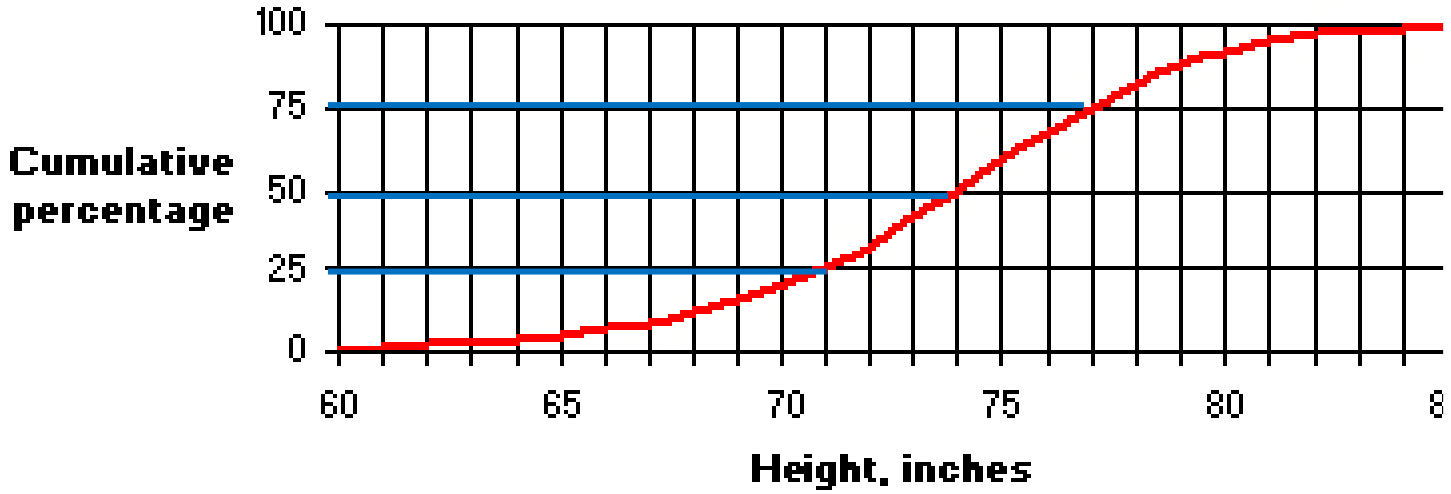
Concetti statistici di partenza

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:

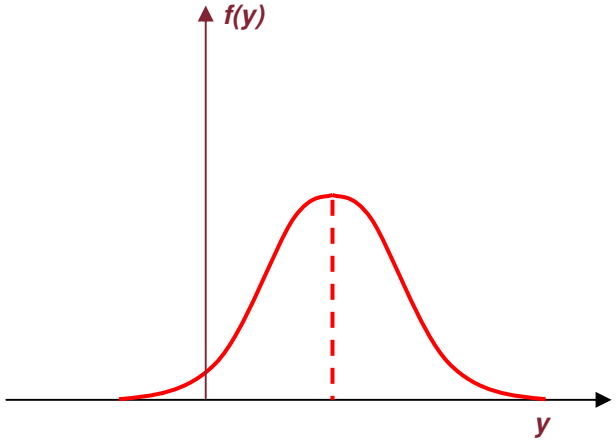
$$\mu = E(y) = \begin{cases} \int_{-\infty}^{+\infty} y f(y)\,dy \\ \sum_{i \to \infty} y_i p(y_i) \end{cases} \qquad \sigma^2 = V(y) = \begin{cases} \int_{-\infty}^{+\infty} (y-\mu)^2 f(y)\,dy \\ \sum_{i \to \infty} (y_i-\mu)^2 p(y_i) \end{cases}$$

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:

μ=275 kPa $\sigma = 13$ kPa

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

P(y > = 241) = 1 – P(y<=241)

Z=(241-275)/13= - 2.61 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:

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

Le specifiche di progetto sono: 0.25 ± 0.0015 in

Quanti alberini cadono nella specifica?

Zinf= - 4.6

Zsup=1.4

P(Zsup)-P(Zinf) = 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"):

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

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

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

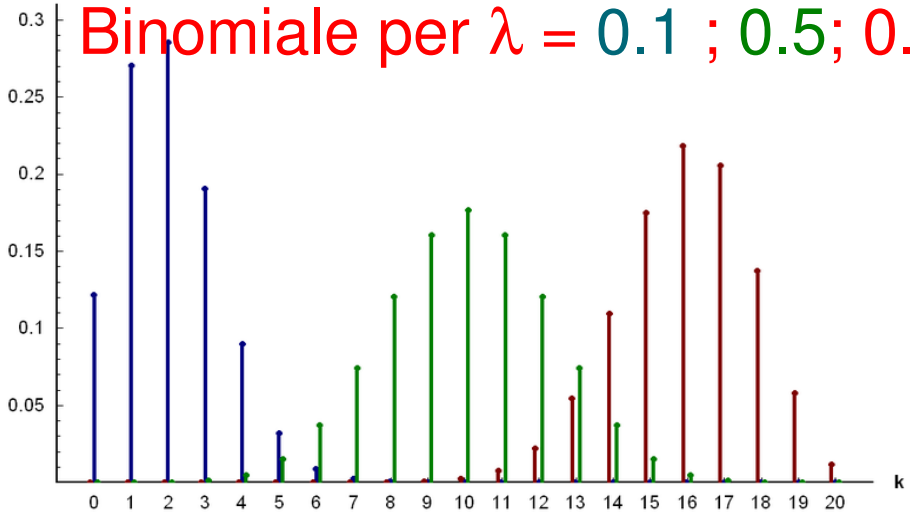
- **Distribuzione esponenziale** per l'affidabilità con tasso di avaria costante:

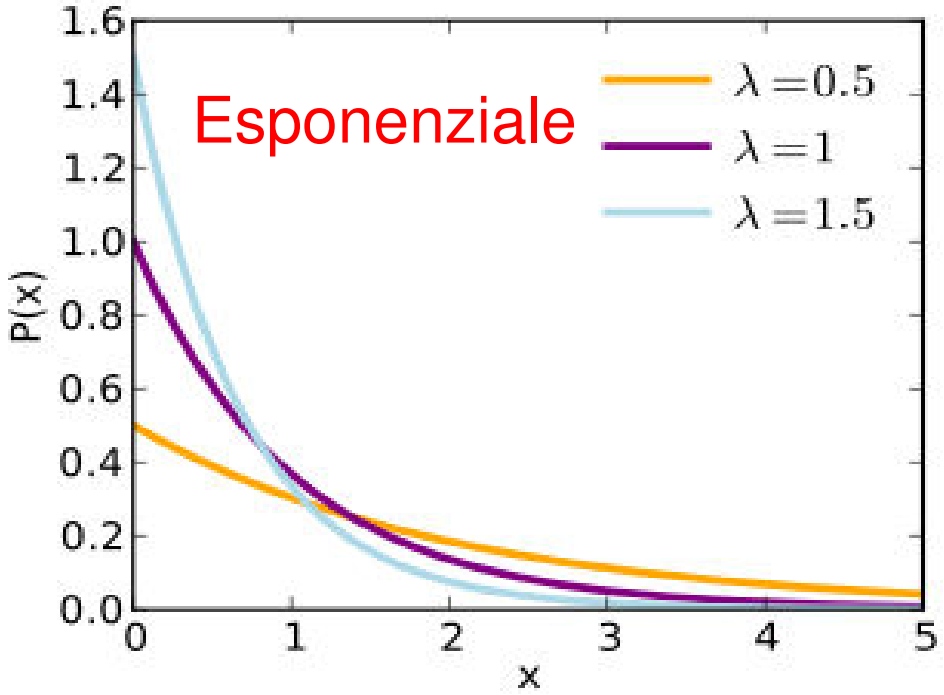
$$p(y) = \lambda e^{-\lambda x} \qquad \begin{matrix} \mu = 1/\lambda \\ \sigma^2 = 1/\lambda^2 \end{matrix}$$

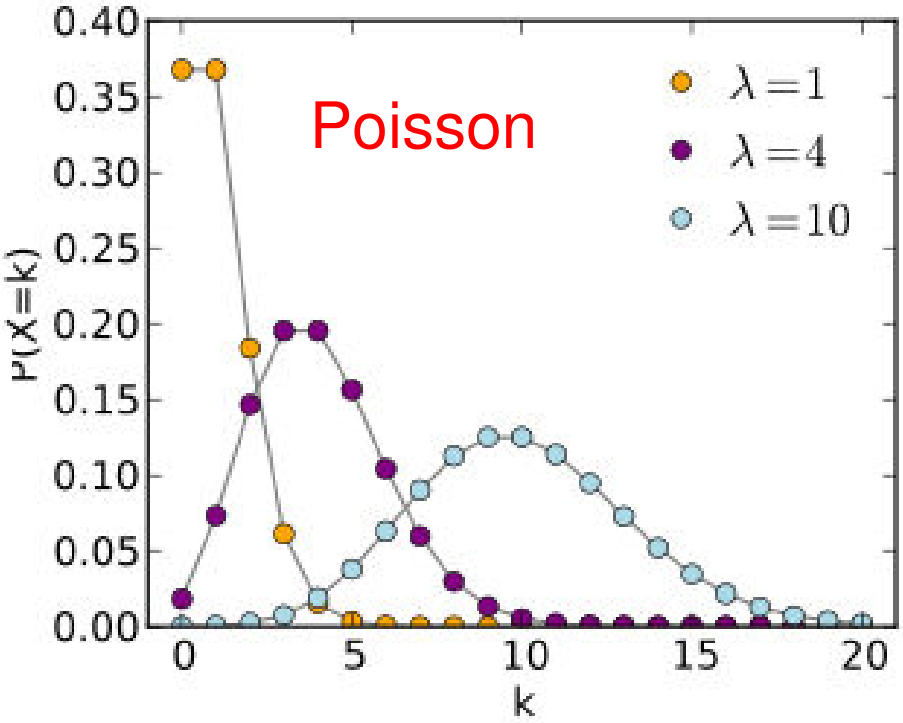
Concetti statistici di partenza

Statistica descrittiva

Binomiale per λ = 0.1 ; 0.5; 0.9







Concetti statistici di partenza

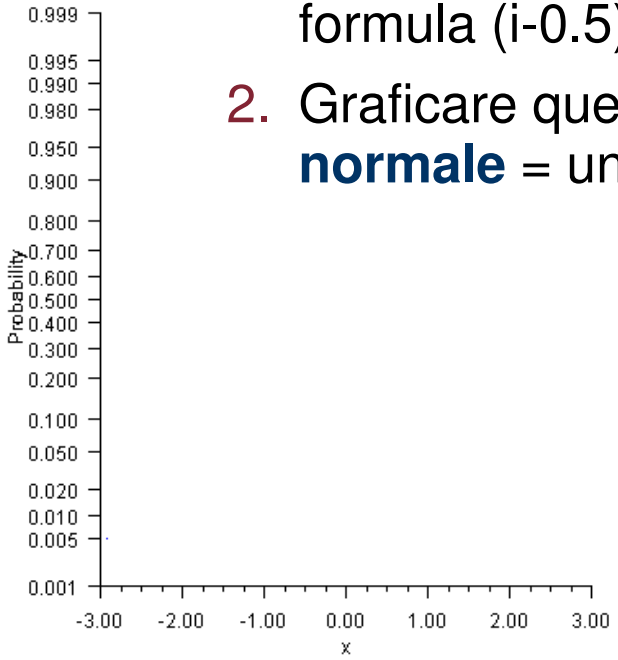
Statistica descrittiva/ PROBABILITY PLOT

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

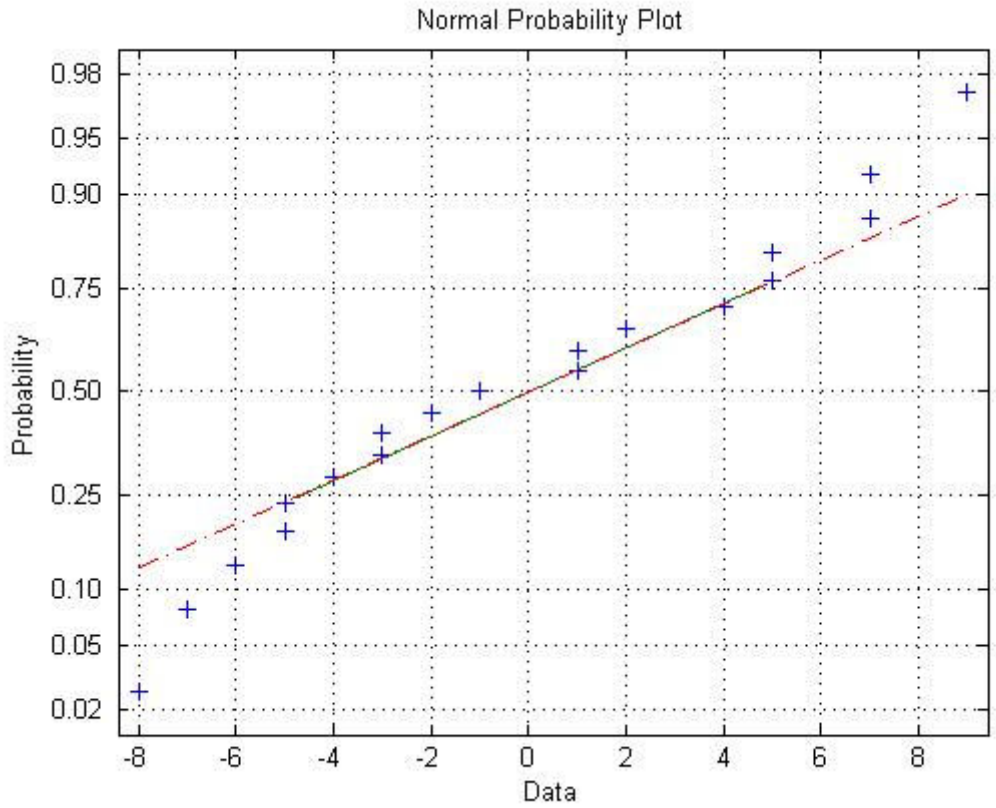


Se i punti sono su una retta l'ipotesi è accettabile!

Concetti statistici di partenza

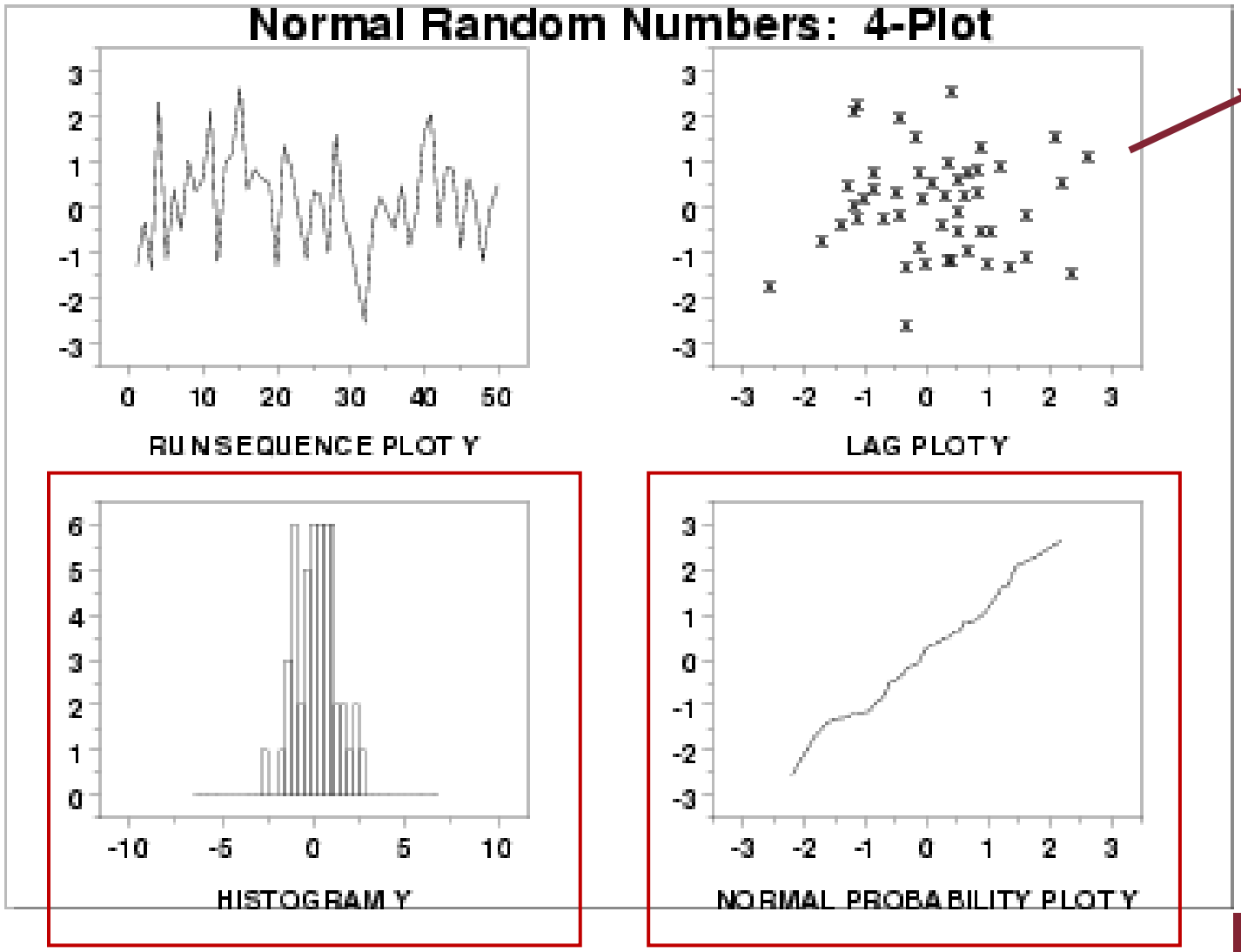
Statistica descrittiva/ PROBABILITY PLOT

	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



Concetti statistici di partenza

Statistica descrittiva/ Interpretazione grafici



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

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

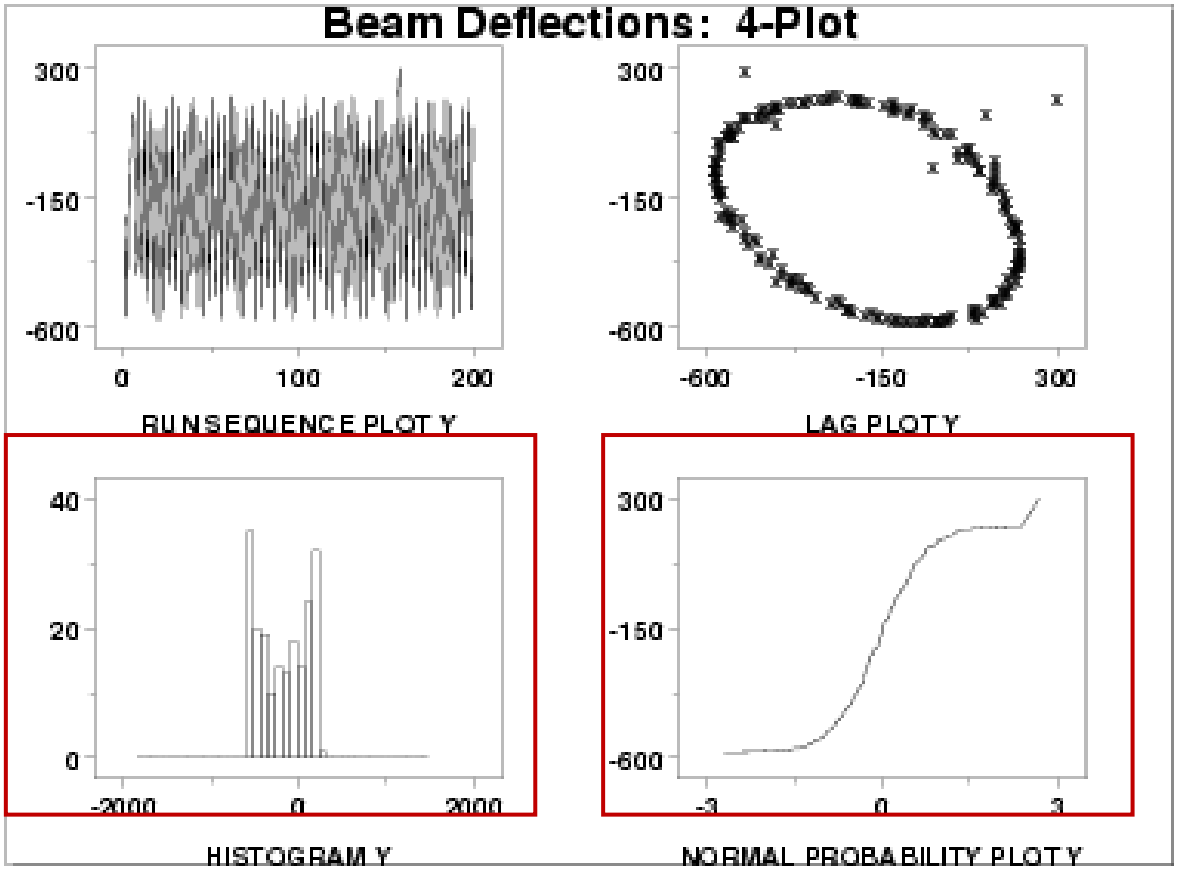
Dalla forma dell'istogramma delle frequenze

e dalla rettilineità della parte centrale del probability plot

l'ipotesi di campione di tipo "normale" (o gausssiano) 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.

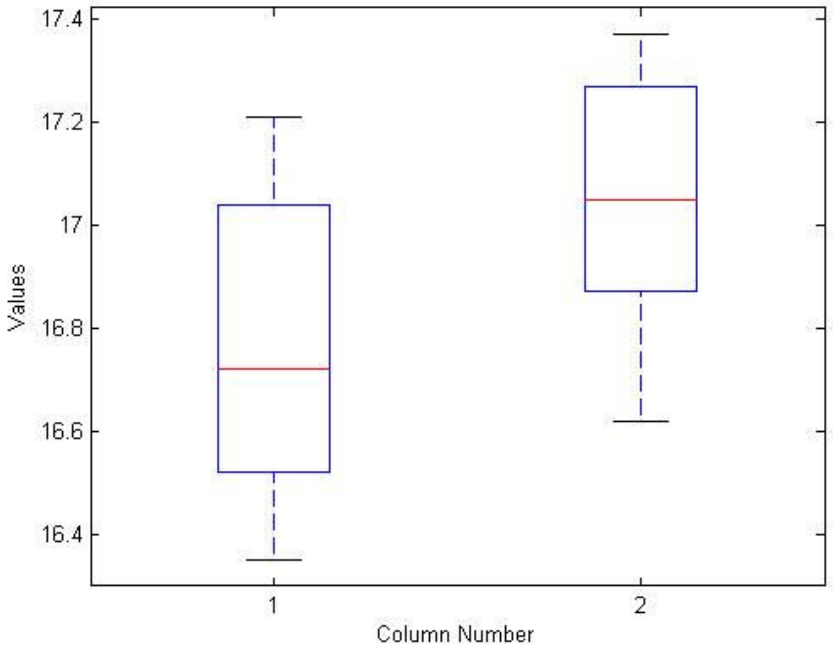
Concetti statistici di partenza

Esempio di riepilogo

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

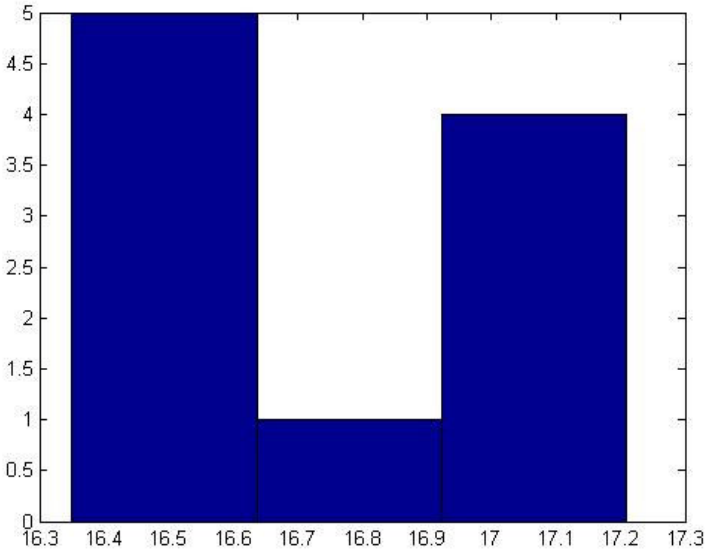
	modified mortar kgf/cm^2	Unmodified kgf/cm^2
1	16.85	16.62
2	16.4	16.75
3	17.21	17.37
4	16.35	17.12
5	16.52	16.98
6	17.04	16.87
7	16.96	17.34
8	17.15	17.02
9	16.59	17.08
10	16.57	17.27

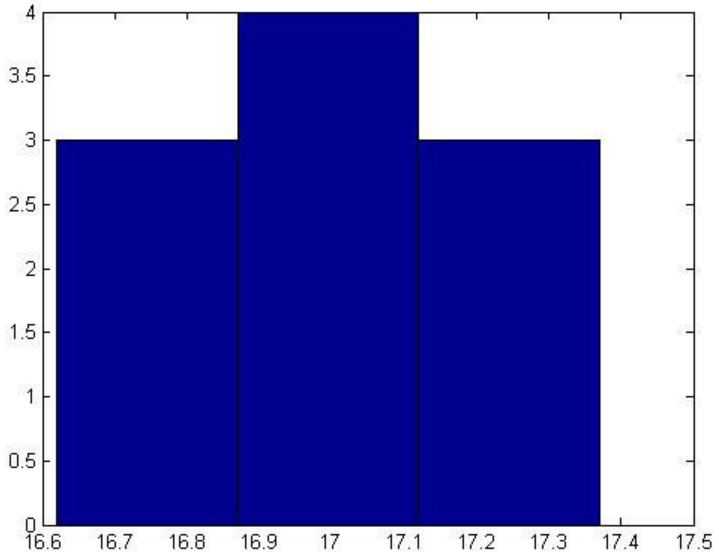
	kgf/cm^2	kgf/cm^2
media	16.76	17.04
var	0.10	0.06
stad.dev	0.32	0.25
Q0	16.35	16.62
Q1	16.53	16.90
Q2	16.72	17.05
Q3	17.02	17.23
Q4	17.21	17.37

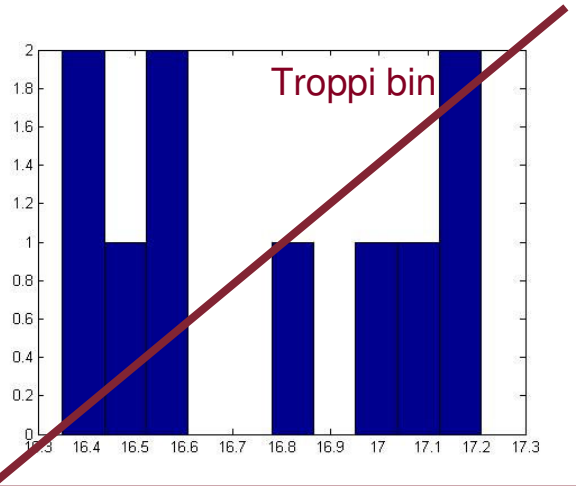


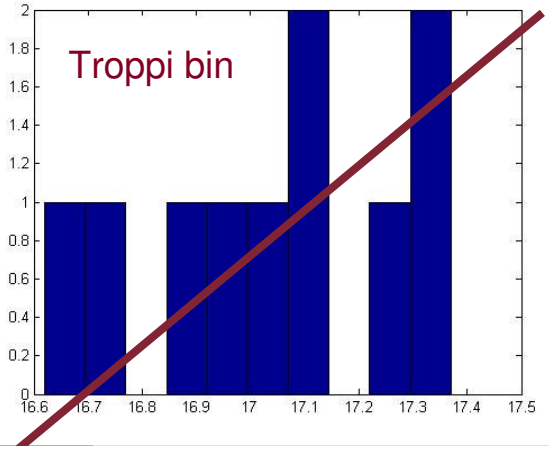
Concetti statistici di partenza

Esempio di riepilogo



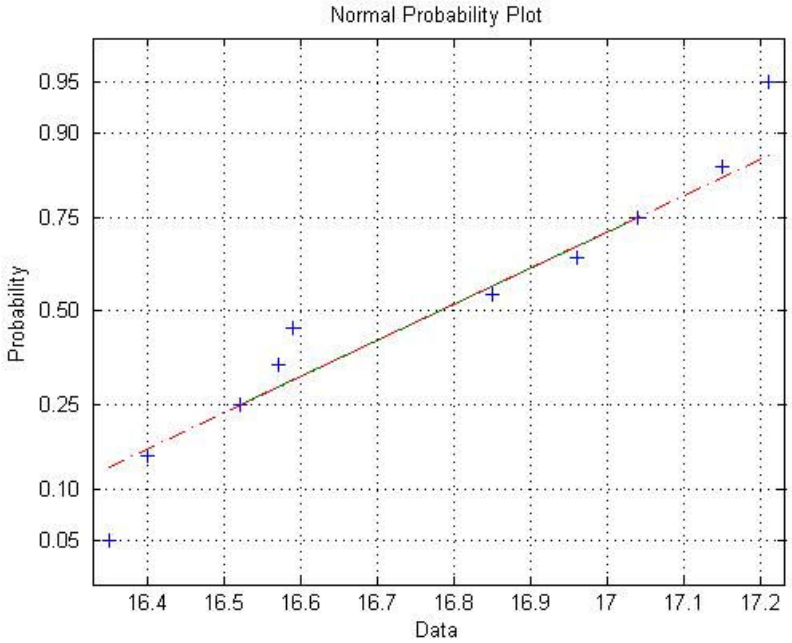


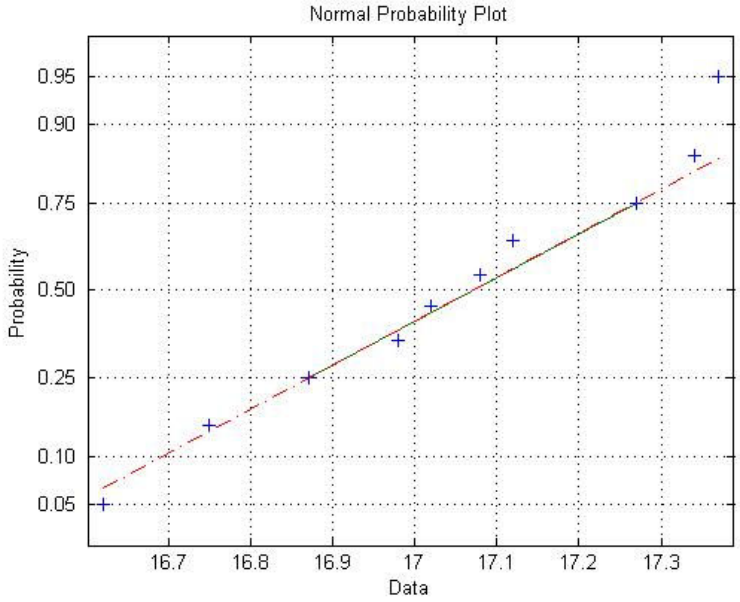




Concetti statistici di partenza

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?

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

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

Occorre definire:

- una statistica di test
- un ipotesi di errore

Concetti statistici di partenza

Inferenza statistica

Inferire = confrontare i dati con una ipotesi

Scelta della statistica del test:

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

Ipotesi nulla H0

Ipotesi alternativa H1

verificate attraverso i valori di α e β

α = P(rifiutare H0 sebbene sia vera) → rischio del produttore

β = P(accettare H1 sebbene sia falsa) → rischio dell'acquirente

Concetti statistici di partenza

Table 2-3 Tests on Means with Variance Known

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

Concetti statistici di partenza

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

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

SAPIENZA
UNIVERSITÀ DI ROMA

Concetti statistici di partenza

Table 2-7 Tests on Variances of Normal Distributions

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

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

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

$$= \frac{1}{n-1}E\left(\sum_{i=1}^{n} y_i^{\,2} - n\bar{y}^2\right) = \frac{1}{n-1}(n-1)\sigma^2$$

Concetti statistici di partenza

Statistica descrittiva/ Sulla scelta del campione

Teorema del limite centrale

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

$$se \qquad y = x_1 + \ldots + x_n$$

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

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

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

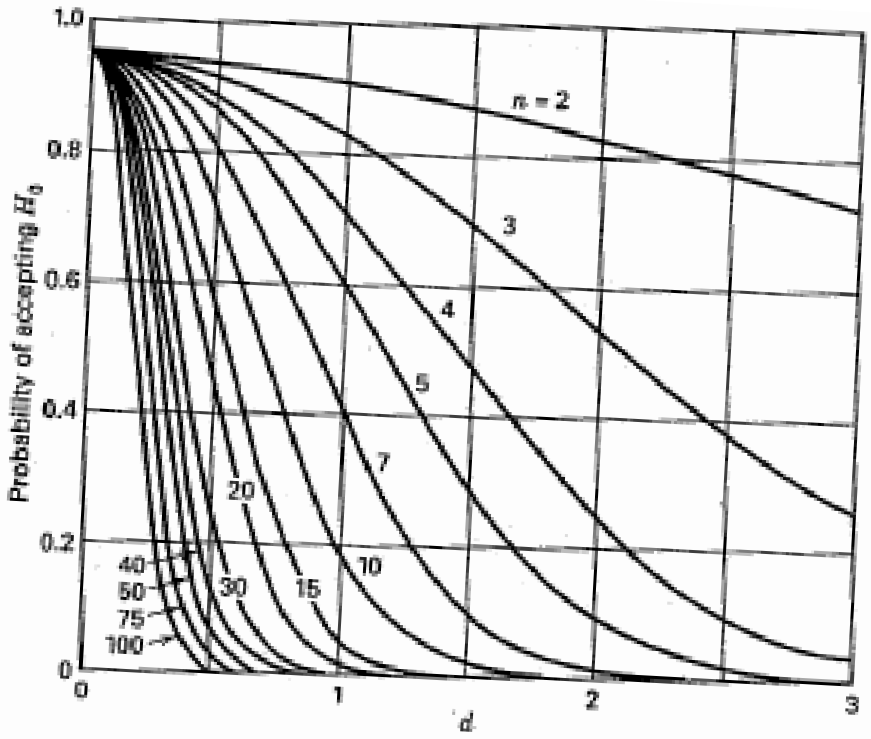
n= 3 o 4 ? n = 15 ? n > 40?

Asimmetria crescente

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.