



**Marica Franzini**

**Laboratorio di Geomatica - DICAr**

**Università di Pavia**

**email: [marica.franzini@unipv.it](mailto:marica.franzini@unipv.it)**



# Elementi di statistica

## Perché è importante la statistica?

---

L'interesse subentra quando occorre effettuare misure di qualità ed affidabili.

Un topografo chiede alla statistica di essere guidato:

1. nel comprendere la bontà e l'affidabilità delle misure fatte;
2. nel confronto di misure prese in momenti diversi e nel comprendere la causa di eventuali differenze (es. frana).

## Precisione ed accuratezza

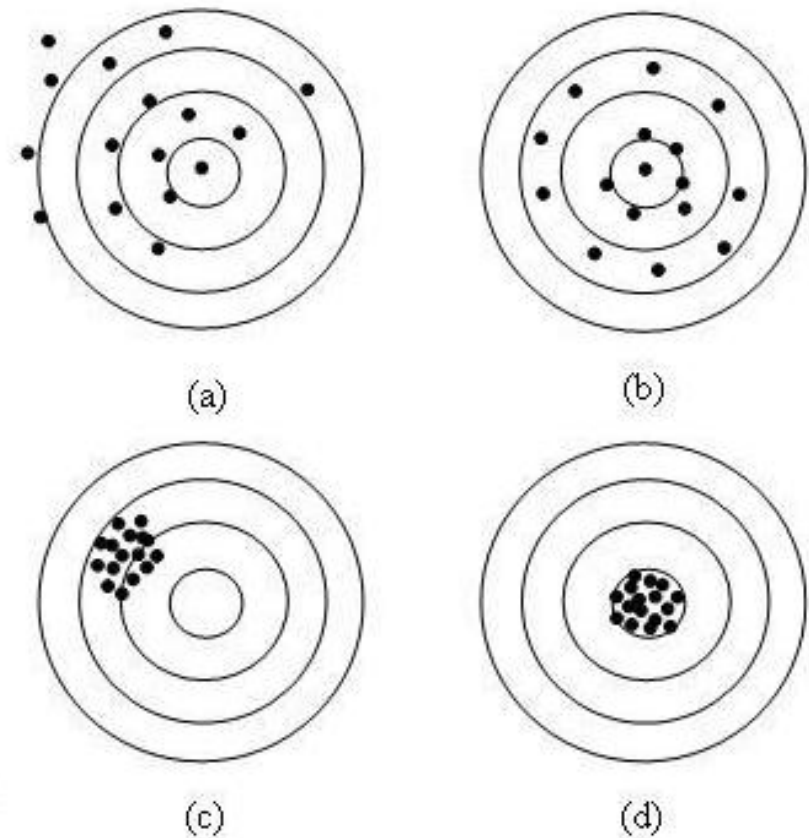
La qualità di una misura si esprime generalmente mediante due termini:

- ✓ precisione - descrive la concentrazione di misure ripetute
- ✓ accuratezza - descrive la distanza fra le misure ed il valore vero

Esempio del bersaglio:

- tiratore poco preciso e poco accurato
- tiratore accurato ma poco preciso
- tiratore preciso ma poco accurato
- tiratore preciso ed accurato

Spesso queste definizioni vengono invertite per cui è sempre buona norma prestare molta attenzione al contesto in cui vengono usate.



# Classificazione degli errori

---

Gli errori vengono comunemente suddivisi in tre categorie:

1. errori grossolani (o blunder)
2. errori sistematici (o bias)
3. errori accidentali (o random error)

## Errori grossolani

---

Sono causati da fattori esterni alla misura vera e propria.

Un esempio di errore grossolano può essere una svista dell'operatore nel leggere l'altezza strumentale o nel collimare un dettaglio sulla facciata di un edificio.

Tali errori non devono essere trattati con metodi statistici ma individuati ed eliminati; la ripetizione di misure salvaguarda in parte da questa fonte di errore.

## Errori sistematici

---

Originano dalla mancata o non corretta considerazione di alcuni aspetti dei fenomeni fisici coinvolti nelle misure.

Se effettuiamo delle misure con un distanziometro laser, senza considerare la costante del prisma, commettiamo degli errori sistematici su tutte le misure.

Questo esempio fa capire come difficilmente gli errori sistematici siano individuabili mediante ripetizione in quanto presenti costantemente ad ogni nuova misura.

Esistono tuttavia delle procedure e degli accorgimenti che permettono di ridurre od eliminare tale sorgente di errore. Un nuovo set di misure effettuato ad esempio da un altro operatore potrebbe mettere in luce un errore commesso in precedenza e correggerne le conseguenze.

## Errori accidentali

---

Sono responsabili delle piccole fluttuazioni che la ripetizione di misure di precisione evidenzia.

Sono dovuti a un complesso di ragioni: piccole imperfezioni degli strumenti, variazioni minime delle condizioni ambientali, etc.

## Variabile aleatoria continua

---

In teoria della probabilità, una variabile aleatoria (o variabile casuale) può essere pensata come il risultato numerico di un esperimento.

In particolare, si definisce variabile casuale continua una variabile casuale che può assumere tutti i valori compresi in un intervallo .

Il lancio di un dado o di una moneta è un esempio di variabile casuale discreta.

In topografia le misure di angoli e distanze sono variabili casuali continue. Nel GPS le misure di pseudo-range effettuate continuamente dal ricevitore e la determinazione delle coordinate 3D sono un esempio di variabile aleatoria continua.



## Distribuzione di probabilità continue

---

La Funzione Densità di Probabilità (FDP) continua definisce analiticamente come si distribuiscono i valori assunti da una variabile aleatoria continua.

Se continuo a misurare le coordinate di un vertice e riporto i risultati in un grafico qual è l'aspetto della figura che ottengo?

Una distribuzione normale (o curva di Gauss); essa è la distribuzione continua più utilizzata in statistica.

Quando si dispone di un'espressione matematica adatta alla rappresentazione di un fenomeno continuo, siamo in grado di calcolare la probabilità che la variabile aleatoria assuma valori compresi in intervalli.

## La curva di Gauss

---

E' caratterizzata dai parametri  $\mu$  e  $\sigma$ .

La **media**  $\mu$  rappresenta il punto di simmetria e anche il punto in cui assume il valore massimo.

La **deviazione standard**  $\sigma$  rappresenta la larghezza della curva:  $\sigma$  piccolo significa curva piccata, *stretta* (indicata anche come STD - Standard Deviation).

La sua espressione analitica è la seguente:

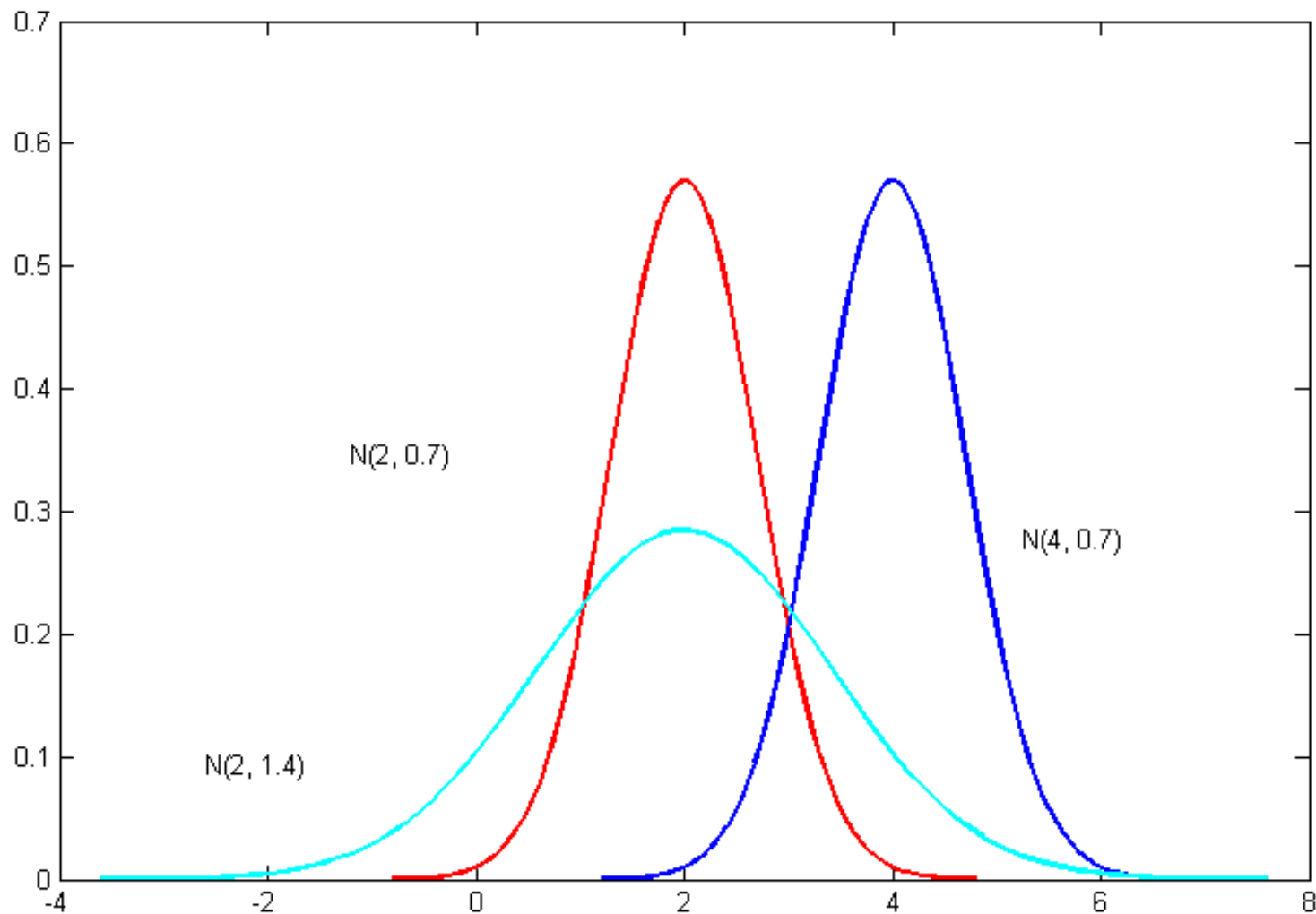
$$f_N(X; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

dove:

- ✓ x sono i valori assunti dalla variabile aleatoria

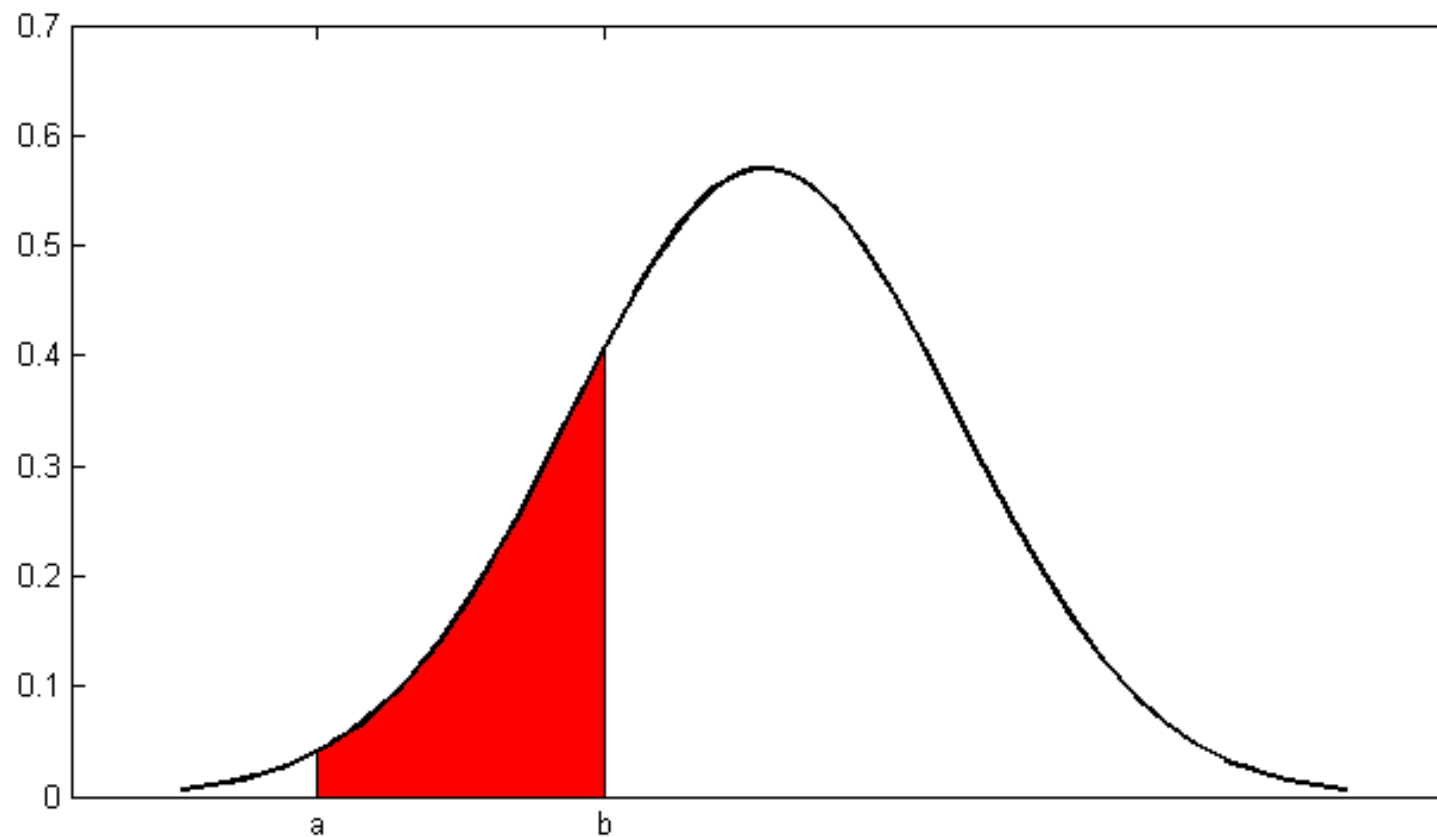
## La curva di Gauss - 2

---



## Interpretazione della curva di Gauss

---



$P([a,b]) = \text{area sottesa}$

La probabilità che la variabile aleatoria assuma valori compresi tra  $a$  e  $b$  è data dall'area sottesa alla curva.

Ovviamente regioni ed intervalli differenti hanno probabilità differenti.

## Stima di media e deviazione standard

---

Fenomeno aleatorio  $X$ .

Disponibilità di  $n$  misure ripetute di  $X$ :  $X_i$  con  $i = 1, 2, \dots, n$

La loro media si stima con la formula

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

e la deviazione standard si stima

$$\sigma = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^n (X_i - \mu)^2}$$

Un'altra figura spesso utilizzata è la varianza:  $\sigma^2 = \frac{1}{n-1} \cdot \sum_{i=1}^n (X_i - \mu)^2$

## Fenomeni aleatori n-dimensionali

---

Quanto indicato fino a questo momento può essere esteso nel caso di variabili ad n-dimensioni. Esempio: la determinazione di un punto con GPS è un fenomeno aleatorio tridimensionale. Generalizzando:

- ✓ vettore  $n$ -dimensionale delle medie

$$\mu = (\mu_1, \mu_2, \dots, \mu_n)^t$$

- ✓ matrice di varianza-covarianza

$$C_{XX} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \dots & \sigma_2^2 & \dots & \sigma_{2n} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \sigma_n^2 \end{bmatrix}$$

Sulla diagonale: varianze delle singole componenti.

Fuori dalla diagonale, ad esempio  $\sigma_{12}$ : covarianza fra componente 1 e 2.

# Esempio baseline GPS

Osservazioni ? X

**GPS**

Data / ora: 09/10/2011 10:21:20      Da: PP-01      A: PP-02       Attivo

---

Errore centr.:  m      Err. quota:  m      Quota Target:  m

---

Vettore baseline:      DX  m      DY  m      DZ  m

Usa matrice covarianza       Usa valori assoluti e relativi stimati

M0:       Assoluti:  m

Qbox:                   Relativi:  ppm

## Intervalli n-sigma - 1

---

Qual è la giusta interpretazione da dare al significato di sigma?

Un esempio:

	Lettura cerchio orizzontale su B	Lettura cerchio orizzontale su C	Angolo interno
1	210,5832	340,9302	130,3470
2	210,5832	340,9326	130,3494
3	210,5814	340,9326	130,3512
4	210,5877	341,0295	130,4418
5	210,6835	341,0273	130,3438
6	210,5127	340,8909	130,3782
			130,3686
			0,0380



## Intervalli n-sigma - 2

---

Nel caso precedente, per l'angolo interno, avevamo:

✓  $\mu$ : 130,3686

✓  $\sigma$ : 0,0380

quali sono i valori che può assumere la variabile casuale?

Spesso si sentono affermazioni del tipo:

il valore è  $130,3686 \pm 0,0380$  ossia l'angolo sarà compreso al massimo nell'intervallo di valori [130,3306 - 130,4066].

E' corretta questa affermazione? NO!

## Intervalli n-sigma - 3

Avevamo visto che la probabilità che la variabile aleatoria assuma valori compresi tra a e b è data dall'area sottesa alla curva.

Per la curva di Gauss si ha che:

$$P([\mu - \sigma, \mu + \sigma]) = 0.683$$

$$P([\mu - 2\sigma, \mu + 2\sigma]) = 0.955$$

$$P([\mu - 3\sigma, \mu + 3\sigma]) = 0.997$$

Interpretazione:

facendo 100 misure, in media  
68 cadono nell'intervallo 1-sigma;  
95 cadono nell'intervallo  
2-sigma, ecc..

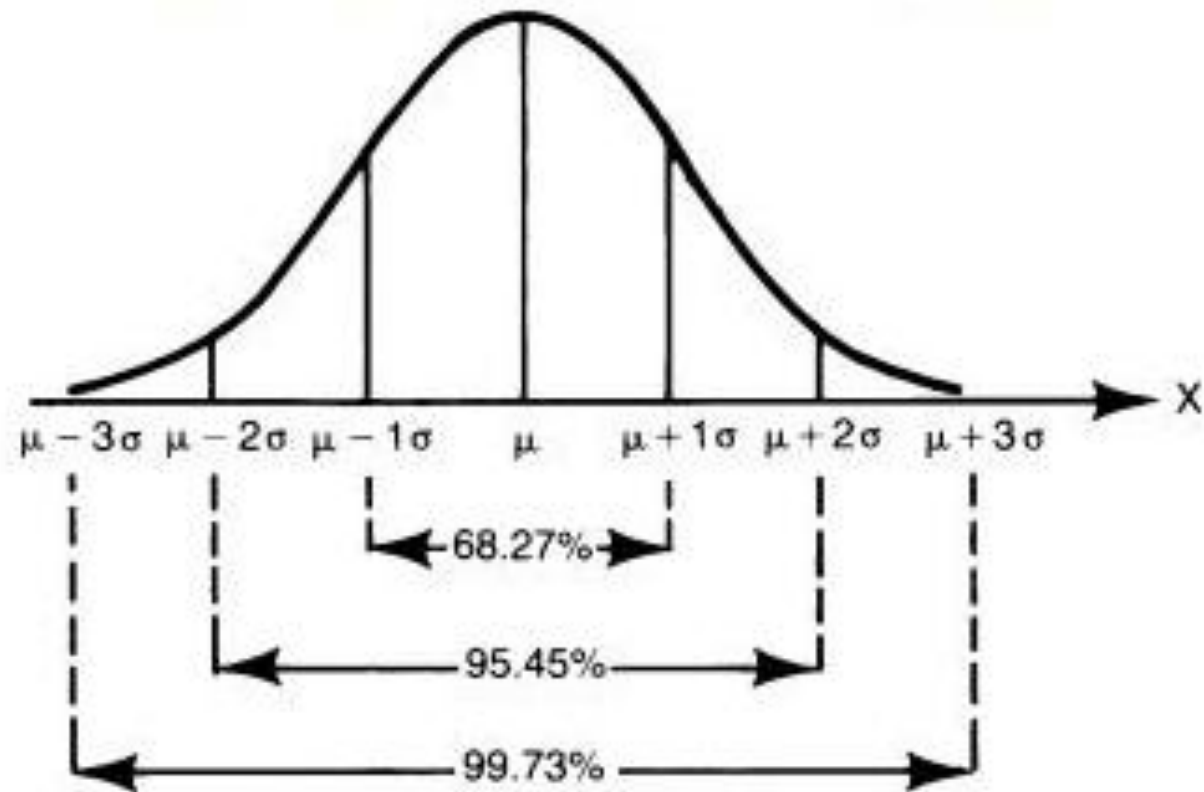


Figure 2

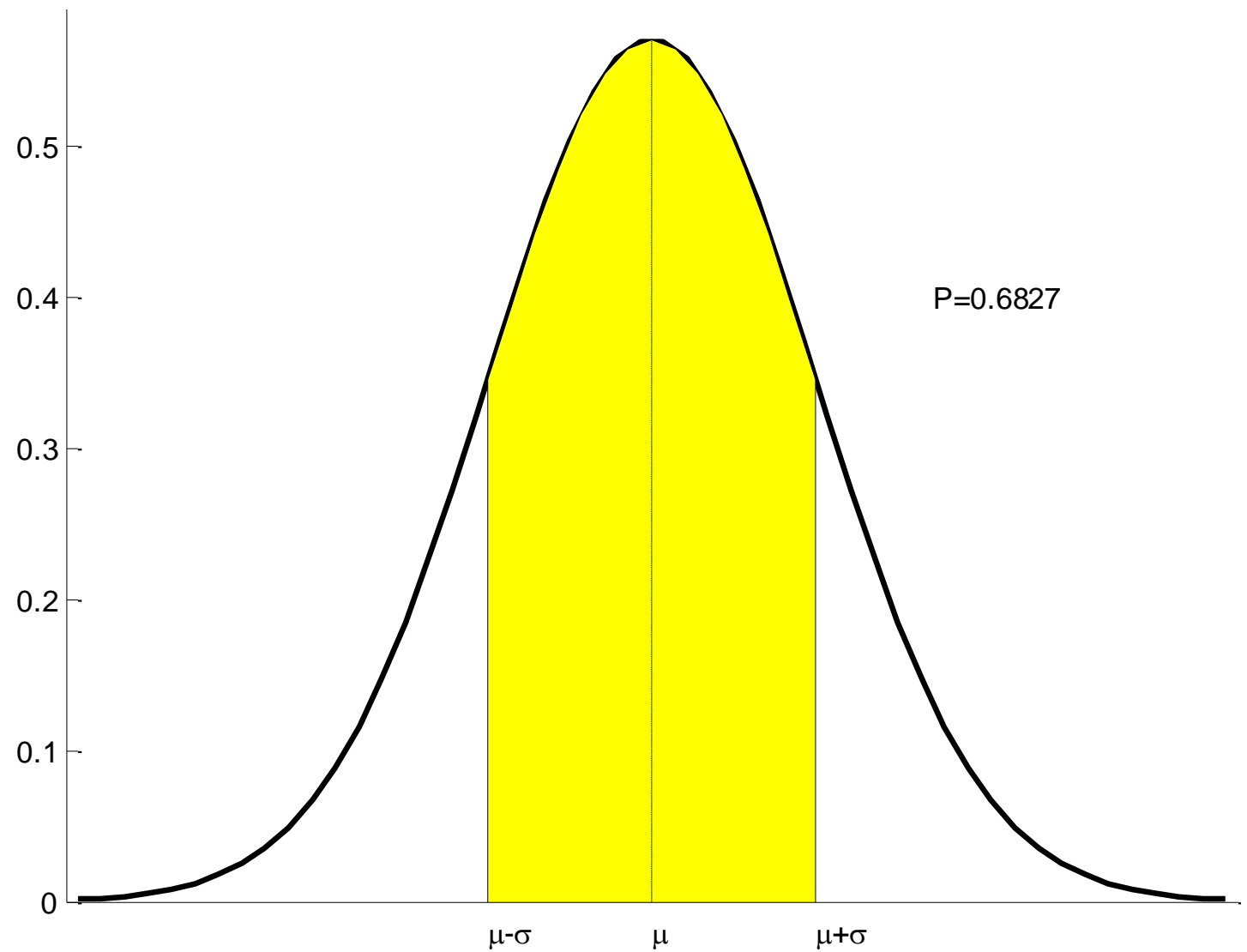
99.73%	99%	95.45%	95%	90%	80%
3.00	2.58	2.00	1.96	1.645	1.28

## Intervalli notevoli per la curva di Gauss

---

Intervallo 1-sigma,  
avente probabilità

$$P = 0.6827$$

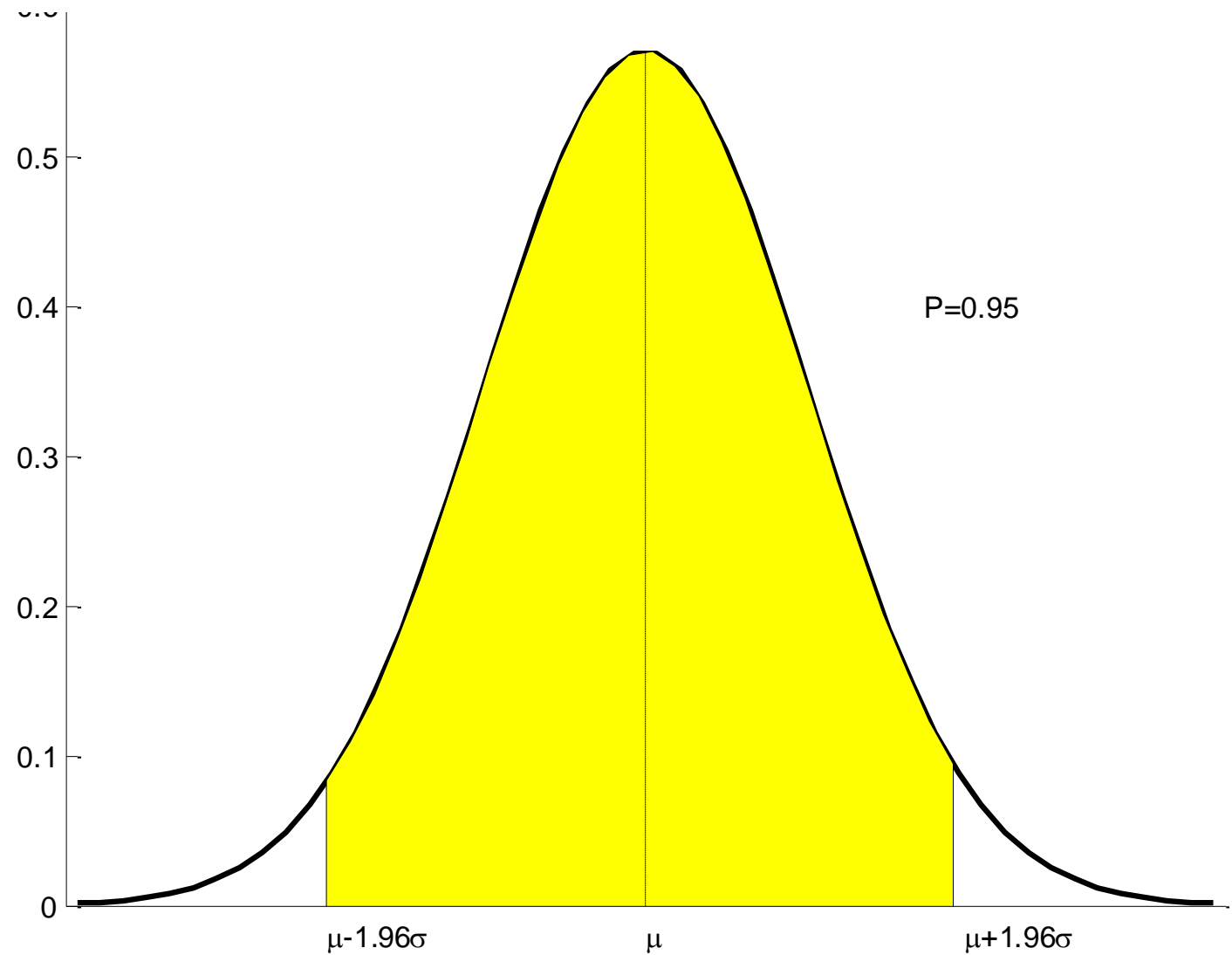


## Intervalli notevoli per la curva di Gauss - 2

---

Intervallo 1.96-sigma,  
avente probabilità

$$P = 0.95$$

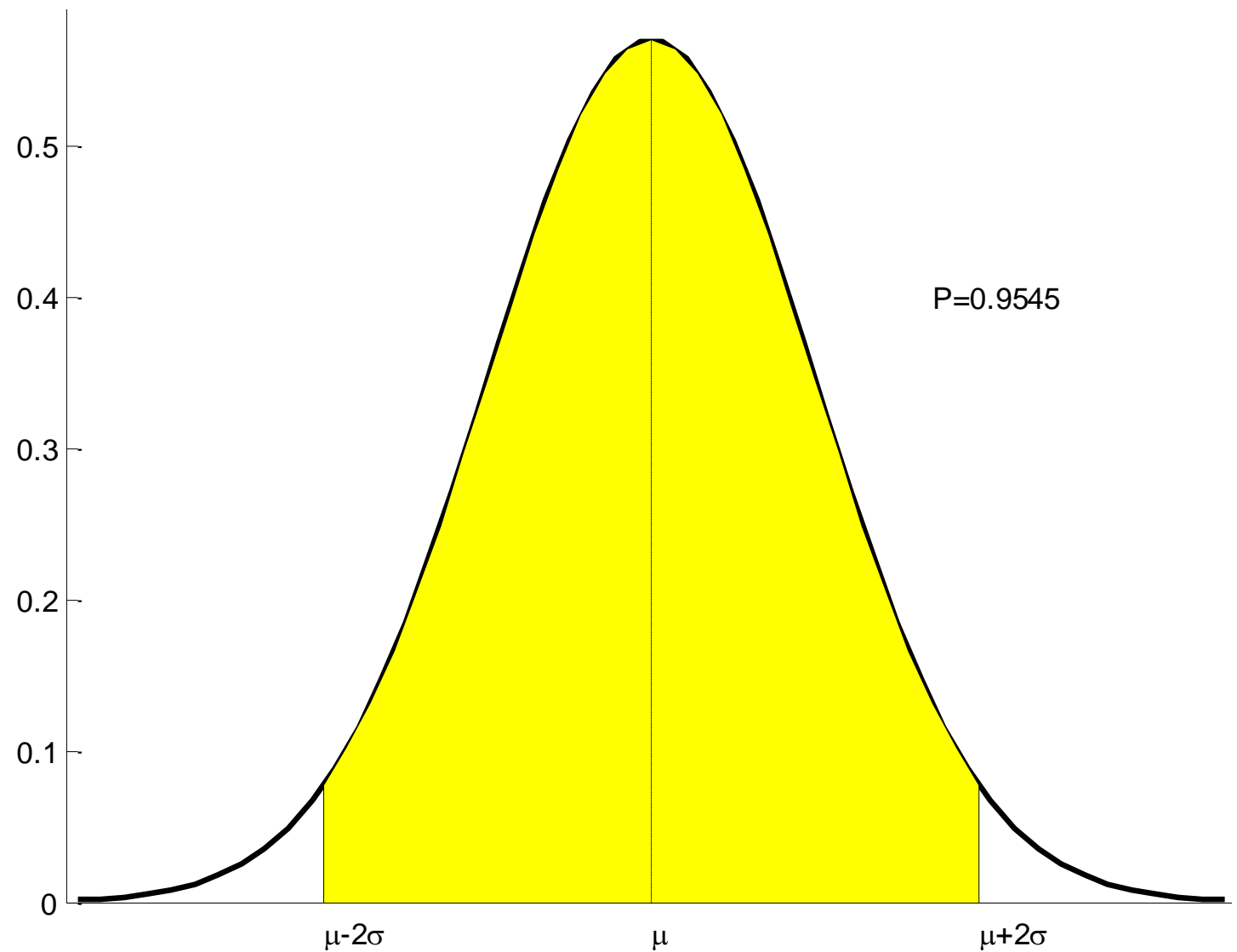


## Intervalli notevoli per la curva di Gauss - 3

---

Intervallo 2-sigma,  
avente probabilità

$$P = 0.9545$$

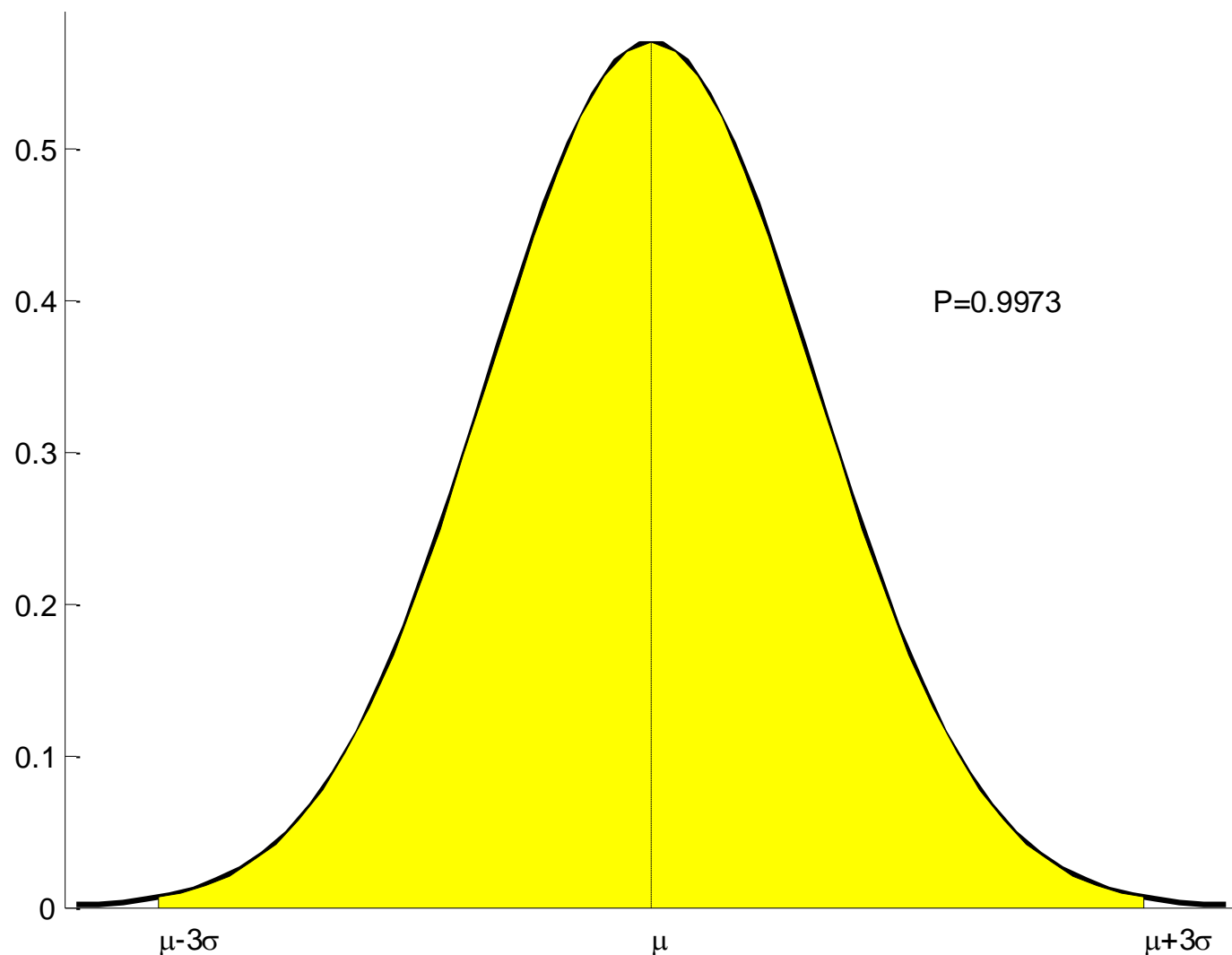


## Intervalli notevoli per la curva di Gauss - 4

---

Intervallo 3-sigma,  
avente probabilità

$$P = 0.9973$$



## Esempio: monitoraggio di una frana

---

Si vuole monitorare una frana misurando ripetutamente, ad intervalli regolari, la distanza tra il punto A, interno alla frana, ed il punto B, localizzato su un terreno stabile.

La prima serie ripetuta di misure, effettuate il primo giorno, mostra:

- ✓  $\mu = 121.780 \text{ m}$
- ✓  $\sigma = 0.012 \text{ m} = 1.2 \text{ cm}$

Una settimana dopo, con lo stesso strumento, viene effettuata una nuova misura:

- ✓  $d = 121.800 \text{ m}$

La frana si sta muovendo? Occorre evacuare il paese sotto la frana?

## Esempio: monitoraggio di una frana - 2

---

Se la frana non si è mossa, la nuova misura effettuata è frutto della variabilità figlia degli errori accidentali di misura.

Supponiamo di analizzare l'accaduto in termini di livello di significatività al 95%. Sotto questa premessa ci aspettiamo che le misure di distanza cadano in un intervallo  $2\sigma$  pari a:

$$[121.780 - (2 * 0.012), 121.780 + (2 * 0.012)] = [121.756, 121.804]$$

La nuova misura è all'interno dell'intervallo?

Sì, statisticamente non c'è evidenza di movimento.

La nuova misura è al di fuori?

Sì, è compatibile con un possibile movimento.



## Vantaggi delle osservazioni ripetute

---

Perché occorre effettuare numerose ripetizioni?

1. Per aver la possibilità di individuare errori grossolani
2. Per avere una migliore stima del valore vero della variabile casuale: la media di misure ripetute è più vicina al valore vero delle singole misure.
3. Per avere la possibilità di stimare la dispersione delle osservazioni

## Riassumendo

---

Due terminologie per la qualità: precisione ed accuratezza

Tre tipologie di errori: grossolani, sistematici ed accidentali

Due figure statistiche: media e deviazione standard

Esiste un legame tra tutte queste grandezze? SI!

## L'esempio del bersaglio

---

Riutilizzando l'esempio del bersaglio, possiamo considerare la serie di tiri effettuati come una variabile casuale e determinarne il comportamento attraverso una curva di Gauss.

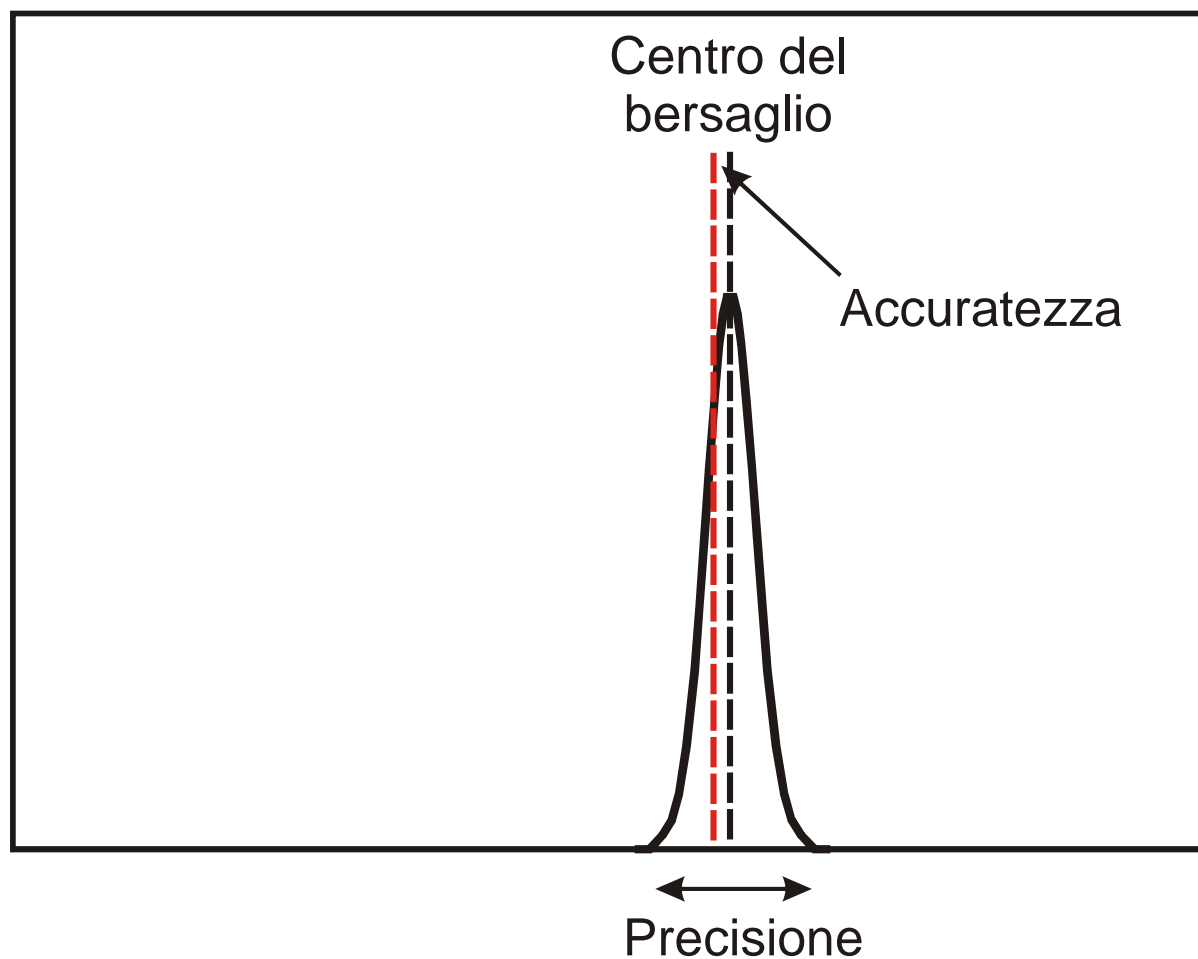
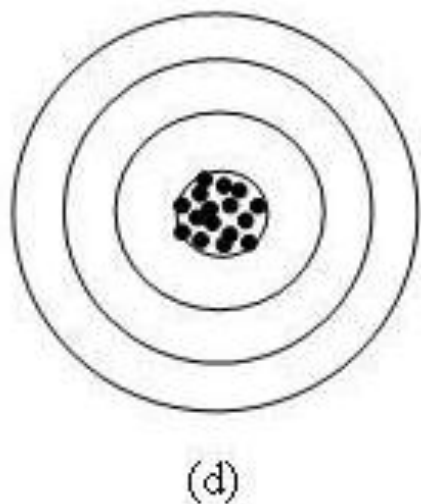
In questo caso, l'ipotetico valore vero che stiamo ricercando è il centro del bersaglio.

Il tiratore effettuerà una serie di spari e, dall'esito di queste operazioni ripetute, si cercherà di trarre alcune conclusioni sulla sua abilità.

La **media** dei tiri fornirà un'indicazione sull'**accuratezza** del tiratore mentre la **deviazione standard** sul suo livello di **precisione**.

## L'esempio del bersaglio - caso d

Caso ideale: il tiratore è molto abile e mostra grande accuratezza e precisione. La media dei tiri è molto vicina al valore vero (il centro del bersaglio) e la deviazione standard contenuta mostra come la dispersione dei tiri sia limitata.



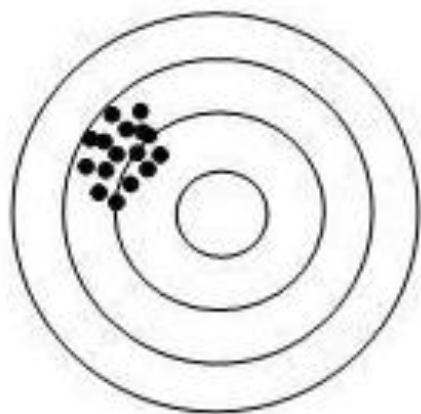
## L'esempio del bersaglio - caso c

Tiratore molto preciso ma poco accurato.

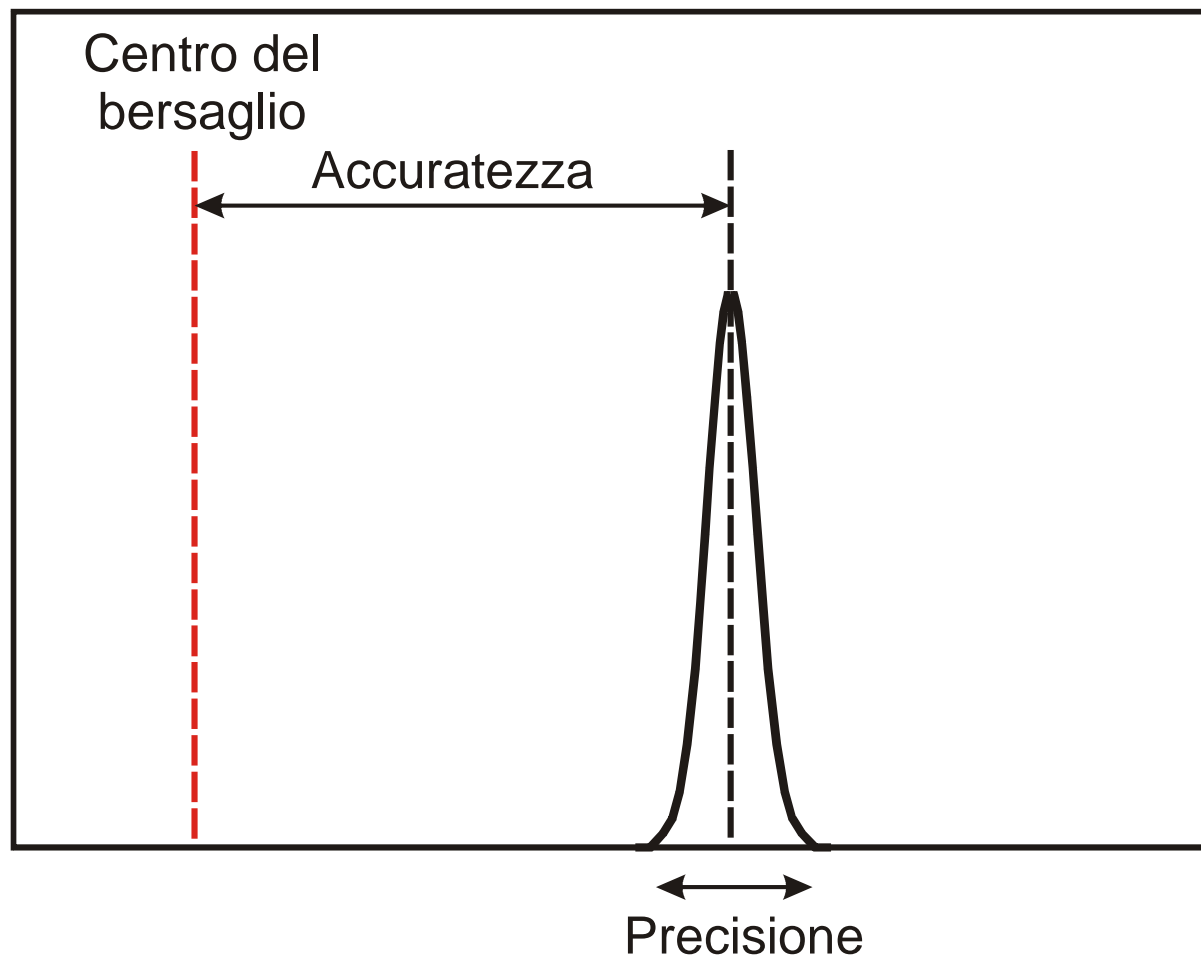
La media dei tiri è lontana dal valore vero (il centro del bersaglio) ma la deviazione standard è contenuta.

Qual è l'origine di tale fenomeno?

Tipicamente la presenza di un **errore sistematico** (minino srettificato, occhio sifulo, etc.)



(c)



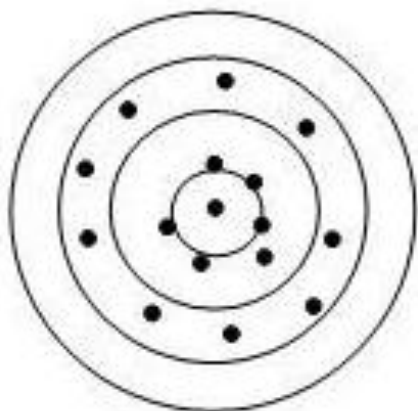
## L'esempio del bersaglio - caso b

Tiratore molto accurato ma poco preciso.

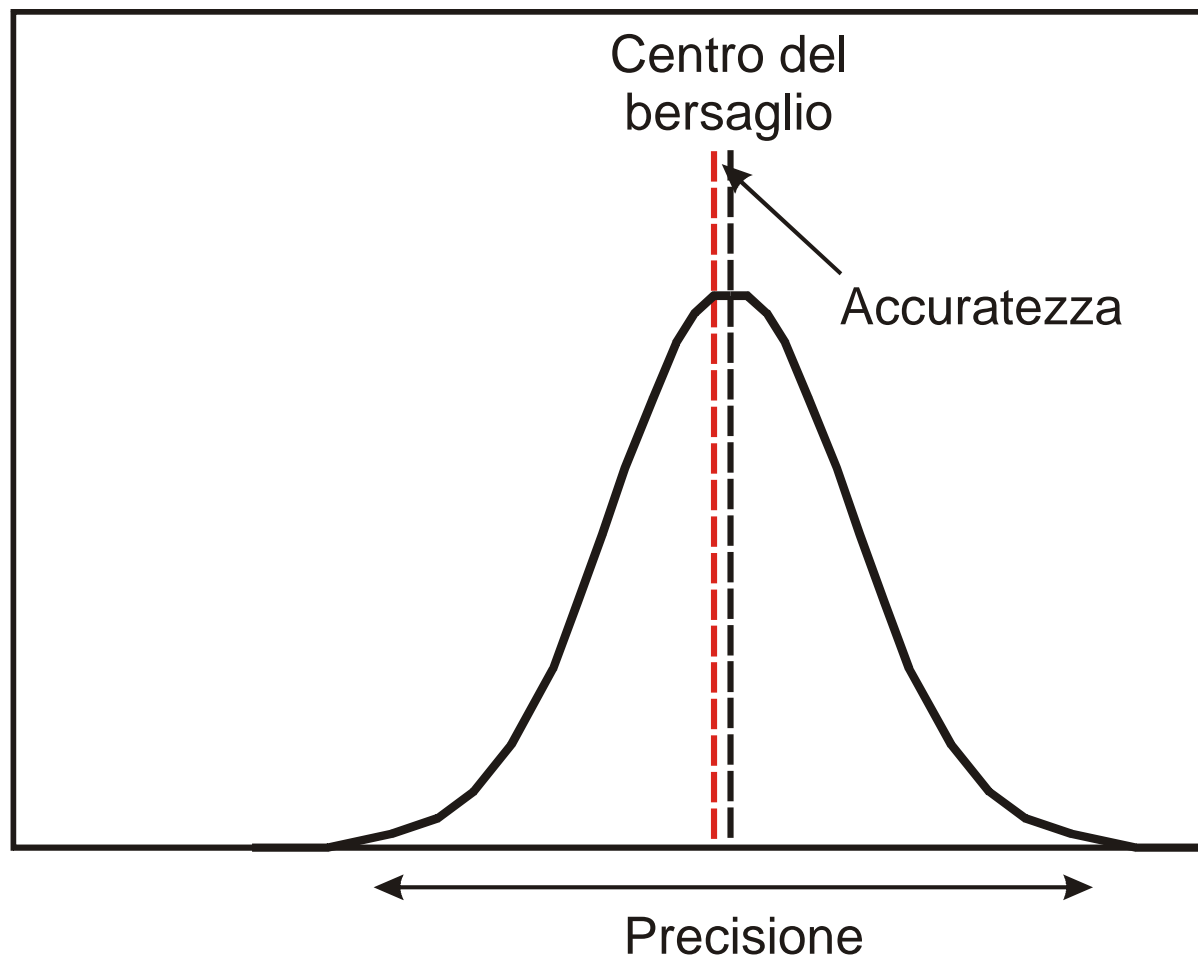
La media dei tiri è molto vicina al valore vero (il centro del bersaglio) ma la deviazione standard è grande sintomo di poca precisione.

Qual è l'origine di tale fenomeno?

Tipicamente la presenza di **errori accidentali** (condizioni ambientali diverse, variazioni nell'impugnatura dell'arma, etc.)



(b)

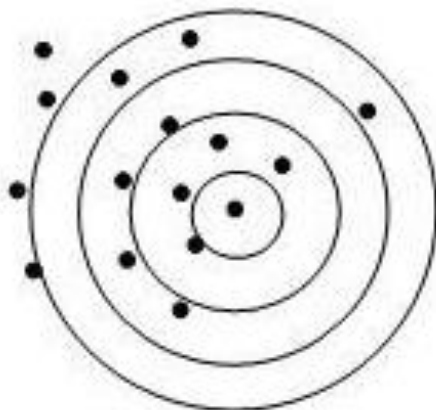


## L'esempio del bersaglio - caso a

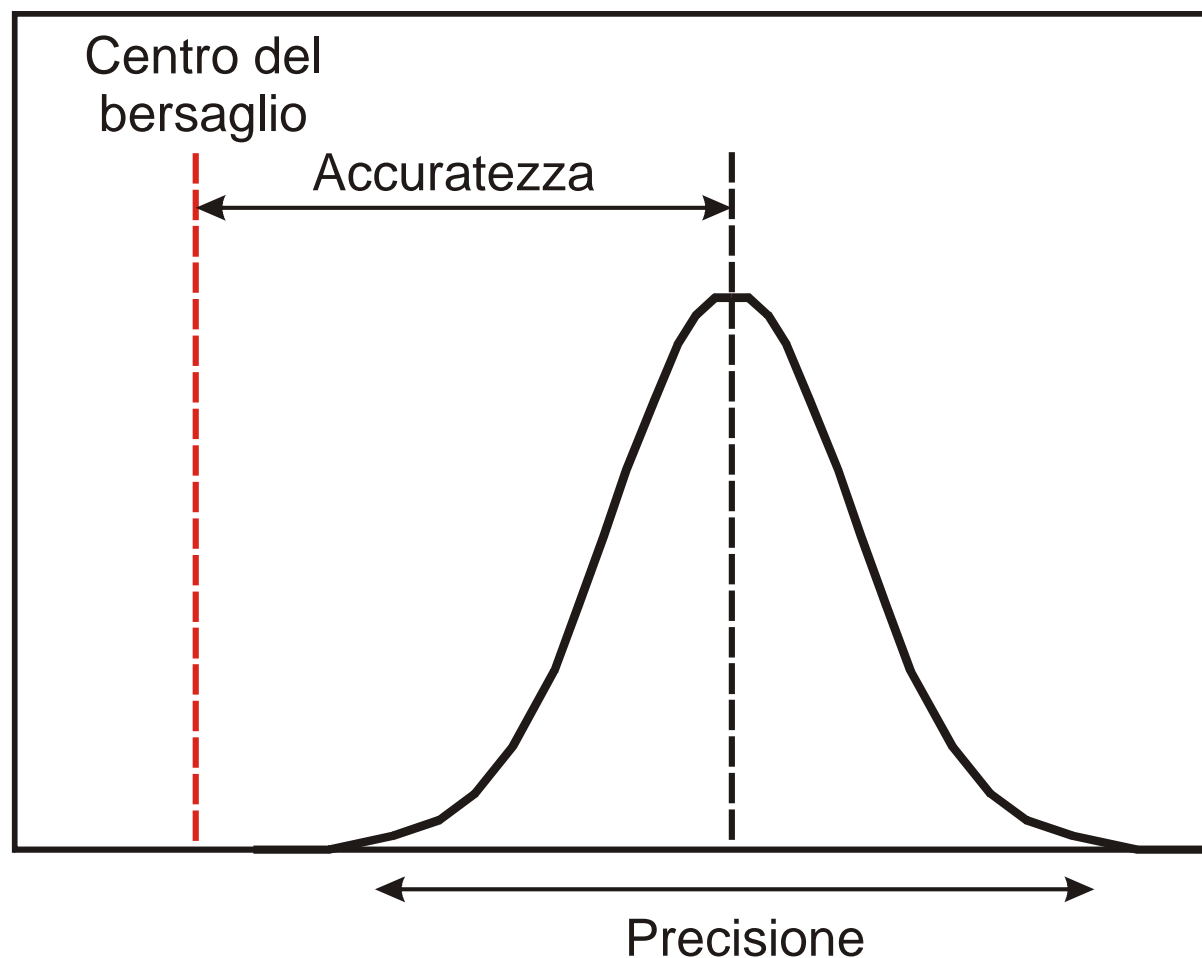
Tiratore poco accurato e preciso.

La media dei tiri è lontana dal valore vero (il centro del bersaglio) e anche la deviazione standard è grande.

Qual è l'origine di tale fenomeno? L'insieme di **errori sistematici ed accidentali**.



(a)



## Riassumendo - 2

---

Abbiamo visto come le figure statistiche caratterizzanti al curva di Gauss possono essere interpretate in funzione della presenza di errori sistematici ed accidentali; l'analisi permette inoltre di esprimere un giudizio in termini di accuratezza e precisione.



## Un esempio reale

---

Quando si effettua un rilievo topografico si misurano delle quantità come angoli e distanze. Per non incorrere in errori grossolani si effettuano misure ripetute → questo permette inoltre di analizzare la precisione delle misure effettuate tramite la deviazione standard ( $\sigma$ ).

	Lettura cerchio orizzontale su B	Lettura cerchio orizzontale su C	Angolo interno
1	210,5832	340,9302	130,3470
2	210,5832	340,9326	130,3494
3	210,5814	340,9326	130,3512
4	210,5877	341,0295	130,4418
5	210,6835	341,0273	130,3438
6	210,5127	340,8909	130,3782
			130,3686
			0,0380

La media ( $\mu$ ) non fornisce tuttavia nessuna informazione sull'eventuale presenza di errori sistematici ma solo una stima del valore vero.

## Un esempio reale - 2

---

Durante una validazione è però possibile effettuare un'analisi completa. Validare delle misure significa confrontarle con dei valori ottenuti da una tecnica più precisa e considerati, per questo motivo, "veri" .

Alcuni esempi:

- validare una livellazione GPS su alcuni capisaldi rilevati tramite livellazione geometrica;
- validare un rilievo differenziale su vertici precedentemente rilevati con un rilievo statico;
- validare una triangolazione aerea collimando sulle immagini punti rilevati con metodi topografici.

## Un esempio reale - 2

---

Validazione dell'accuratezza delle immagini Google Earth su Pavia

Scarti determinati come differenze tra le coordinate derivate da collimazioni su Google Earth e quelle ottenute per via topografica.

# 60 punti	Scarti	
	E [m]	N [m]
<b>Media</b>	15,926	-1,162
<b>STD</b>	0,691	1,108
<b>EQM</b>	15,941	1,606

Media: errore sistematico di circa 16 metri in direzione Est-Ovest

STD: errori accidentali dell'ordine del metro

EQM: ??

Domanda: la media sulla coordinate Nord è un errore sistematico? Test statistici!

## Errore quadratico medio

---

Abbiamo visto che la media può essere utilizzata per stanare gli errori sistematici e che la deviazione standard fornisce un'indicazione sugli errori accidentali.

All'operatore (topografo) piacerebbe aver una stima compatta delle due fonti di errore → errore quadratico medio

$$eqm = \sqrt{media^2 + std^2} = \sqrt{\mu^2 + \sigma^2}$$

L'eqm rappresenta una stima complessiva della qualità delle misura analizzate.

**ATTENZIONE:** in letteratura è possibile reperire diversi nomi. Spesso l'eqm è indicato come RMSE (Root Mean Squared Error) mentre la std con RMS (Root Mean Error) - bisogna sempre leggere attentamente il significato che hanno nel contesto in cui sono state inserite!