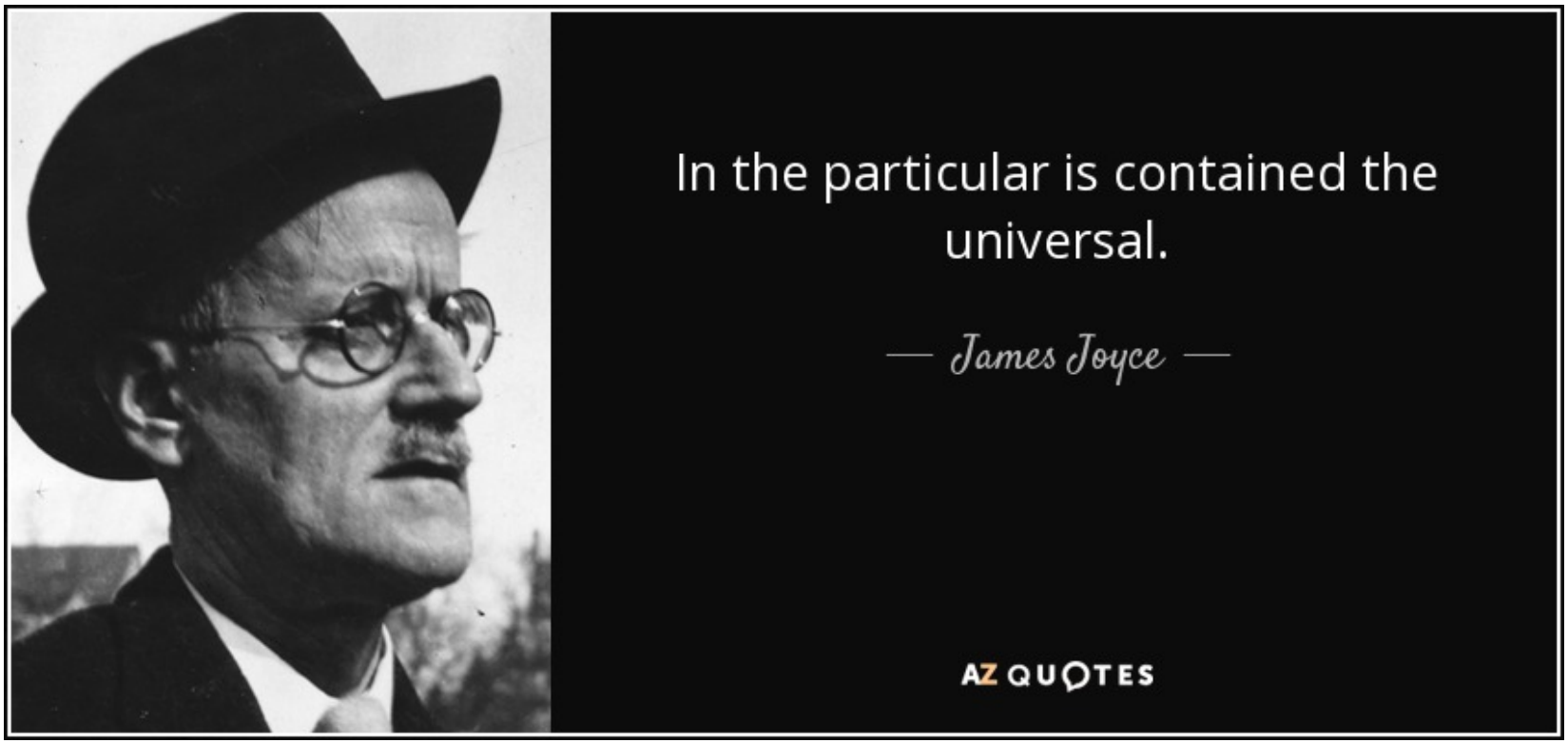
# 704SM Biostatistica

Massimo Borelli

ottobre 2016

UNIVERSITÀ DEGLI STUDI DI TRIESTE
Dipartimento di Scienze della Vita

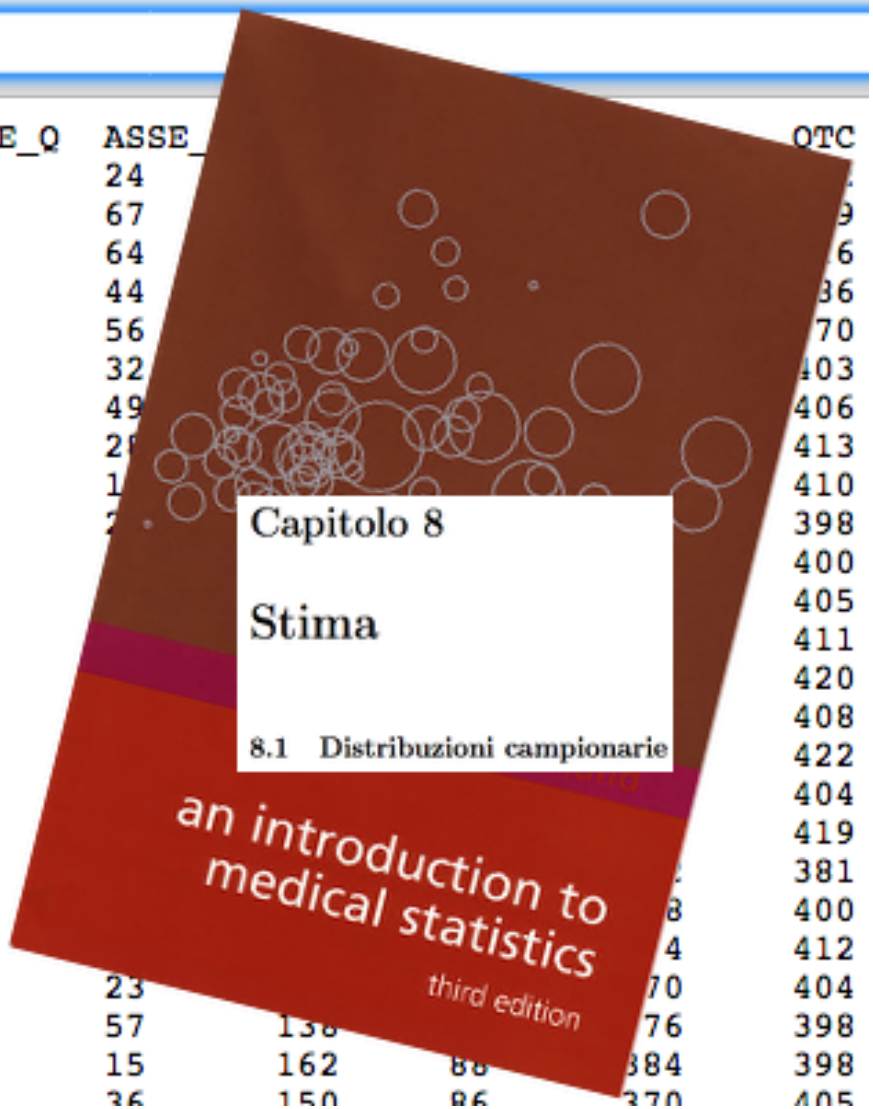
SOCIETA' DEI MATEMATICI
E NATURALISTI DI MODENA
www.socnatmatmo.unimore.it


In the particular is contained the universal.
— James Joyce —
AZ QUOTES

# le stime

www.dmi.units.it/~borelli/dataset/heart.txt

| CODPAZ | SESSO | ETA | FC | ASSE_P | ASSE_Q | ASSE_ | | | QTC |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 46 | 83 | 28 | 1 | 24 | | | |
| 100 | 2 | 20 | 75 | 77 | 66 | 67 | | | 9 |
| 1000 | 2 | 40 | 82 | 36 | 82 | 64 | | | 6 |
| 1001 | 2 | 19 | 66 | 66 | 69 | 44 | | | 36 |
| 1002 | 1 | 9 | 88 | 62 | 86 | 56 | | | 70 |
| 1003 | 2 | 48 | 62 | 35 | 44 | 32 | | | 403 |
| 1004 | 1 | 48 | 72 | 63 | -12 | 49 | | | 406 |
| 1005 | 2 | 20 | 79 | 36 | 57 | 2 | | | 413 |
| 1006 | 1 | 59 | 70 | 19 | -11 | 1 | | | 410 |
| 1007 | 2 | 34 | 80 | 55 | 64 | 2 | | | 398 |
| 1008 | 2 | 58 | 58 | 53 | -19 | | | | 400 |
| 101 | 2 | 48 | 86 | 43 | 13 | | | | 405 |
| 1010 | 2 | 35 | 68 | 2 | 12 | | | | 411 |
| 1011 | 1 | 55 | 73 | 65 | 63 | | | | 420 |
| 1012 | 1 | 43 | 52 | 61 | 52 | | | | 408 |
| 1013 | 2 | 42 | 56 | 77 | 64 | | | | 422 |
| 1014 | 2 | 27 | 66 | 56 | 75 | | | | 404 |
| 1017 | 2 | 50 | 66 | 33 | 13 | | | | 419 |
| 1018 | 1 | 42 | 64 | 22 | 35 | | | | 381 |
| 1019 | 1 | 31 | 60 | 72 | 57 | | | | 400 |
| 102 | 2 | 47 | 81 | 43 | 60 | | | | 412 |
| 1020 | 1 | 61 | 79 | 62 | 59 | 23 | | | 404 |
| 1021 | 2 | 41 | 71 | 38 | 72 | 57 | 138 | | 398 |
| 1022 | 2 | 57 | 66 | 42 | 51 | 15 | 162 | | 398 |
| 1023 | 1 | 58 | 79 | 69 | 15 | 36 | 150 | | 405 |



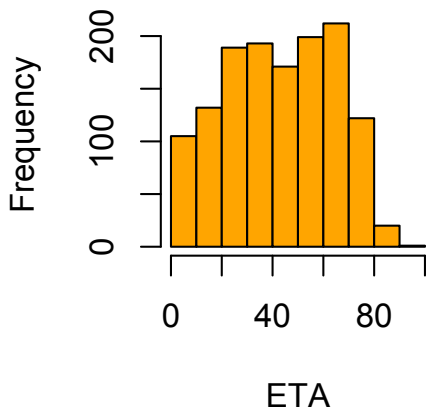
# la spinosa questione dello standard error

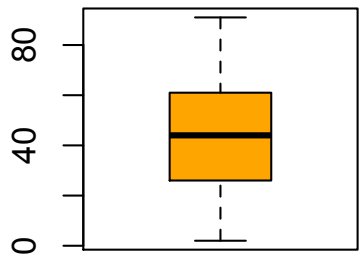
## domande

- Ma che differenza c'è tra deviazione standard ed errore standard?
- Quali 'errori' ci fa fare l'errore standard?
- Quando usare l'uno o l'altro?
- Quale è il 'pregio' che ha l'errore standard?

# che differenza c'è tra SD e SE?

```r
www = "http://www.dmi.units.it/borelli/dataset/heart.txt"
heart = read.table(www, header = TRUE)
attach(heart)
mean(ETA)        # 43.08
sd(ETA)          # 21.35
length(ETA)      # 1344
hist(ETA)        # non e' distribuita normalmente
```



# che differenza c'è tra SD e SE?

```
> sample(ETA, 4)
[1] 58  9  4 73
>
> mediacampionaria = mean(sample(ETA, 4))
> mediacampionaria
[1] 51.25
```

# che differenza c'è tra SD e SE?

**domanda**

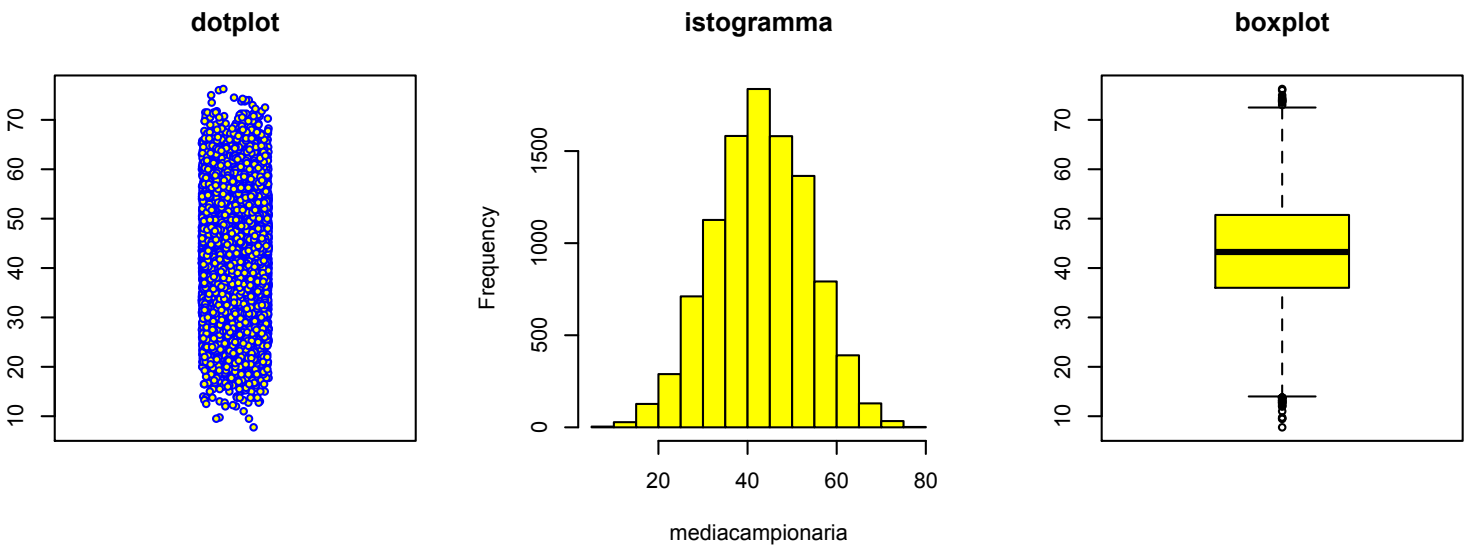
Ma cosa succede se invece di calcolare una sola mediacampionaria ne calcolo decine e decine di migliaia? E in particolare:

- quanto varrà la media?
- quanto varrà la deviazione standard?

# che differenza c'è tra SD e SE?

```
quantevolte = 99999
mediacampionaria = numeric(quantevolte)

for(i in 1: quantevolte)
    mediacampionaria[i] = mean(sample(ETA, 4))
```



# che differenza c'è tra SD e SE?

```
> mean(mediacampionaria)
[1] 43.04459
> mean(ETA)
[1] 43.07961
```

Jakob Bernoulli

●●●○○ I WIND 22:56 93%

it.m.wikipedia.org

## Legge debole dei grandi numeri

Se, data una successione di variabili casuali $X_1, X_2, \ldots, X_n, \ldots$ aventi la stessa media $\mu$, la stessa varianza finita e indipendenti, si considera la media campionaria

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

la legge (debole) dei grandi numeri afferma che per ogni $\varepsilon > 0$:

$$\lim_{n \to \infty} \mathrm{P}\left(\left|\bar{X}_n - \mu\right| < \varepsilon\right) = 1.$$

ossia la media campionaria converge in probabilità alla media comune delle $X_i$.

Con maggior

# che differenza c'è tra SD e SE?

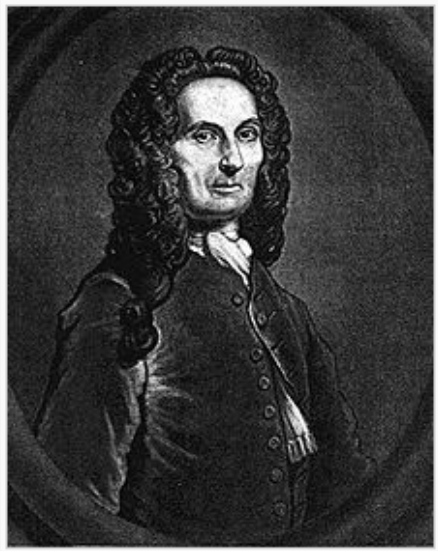
**risposta: dal particolare all'universale**

Quindi se io ho dei dati numerici tratti da un **campione** e calcolo la **media** (campionaria) ottengo una stima della **media vera** (ignota) di quel carattere di tutta la **popolazione**?

- sì

# che differenza c'è tra SD e SE?

```
> sd(mediacampionaria)
[1] 10.62414
> sd(ETA)
[1] 21.35423
```



Abraham de Moivre

●●●○○ I WIND 22:57 92%

it.m.wikipedia.org

## Teoremi centrali del limite

I **teoremi centrali del limite** sono una famiglia di teoremi di convergenza debole nell'ambito della teoria della probabilità.

Una delle formulazioni più note del teorema è la seguente:

Sia $X_j$ una delle $n$ variabili aleatorie indipendenti e identicamente distribuite, e siano $E[X_j] = \mu$ e $Var[X_j] = \sigma^2 \quad \forall j$, con $0 < \sigma^2 < \infty$ (ovvero $X_j \in \mathcal{L}^2$).

Posto $Y_n = \dfrac{\sum_{j=1}^{n} X_j - n\mu}{}$ allora

# che differenza c'è tra SD e SE?

**attenzione**

Quindi se io ho dei dati numerici tratti da un **campione** e calcolo la dispersione (**deviazione standard**) della mediacampionaria non ottengo una stima della dispersione di quel carattere di tutta la **popolazione**?

- no!

# la spinosa questione dello standard error

**risposta**

- Ma che differenza c'è tra deviazione standard ed errore standard?

```
> sd(mediacampionaria)
[1] 10.62414
> sd(ETA)
[1] 21.35423
```

$$s.e. = \frac{\sigma}{\sqrt{n}}$$

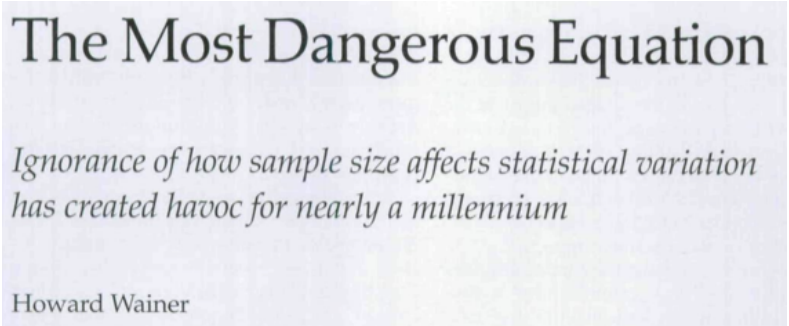
# la spinosa questione dello standard error

## The Most Dangerous Equation

*Ignorance of how sample size affects statistical variation has created havoc for nearly a millennium*

Howard Wainer

# la spinosa questione dello standard error


The Most Dangerous Equation

*Ignorance of how sample size affects statistical variation has created havoc for nearly a millennium*

Howard Wainer

Lanciamo una monetina per 10 volte. È plausibile che possano uscire meno di 4 teste?

Lanciamo una monetina per 10mila volte. È plausibile che possano uscire meno di 4mila teste?

# la spinosa questione dello standard error

## Una simulazione

Un milione di ripetizioni di quattro esperimenti:

- 10 lanci di una monetina
- 100 lanci di una monetina
- 1000 lanci di una monetina
- 10000 lanci di una monetina

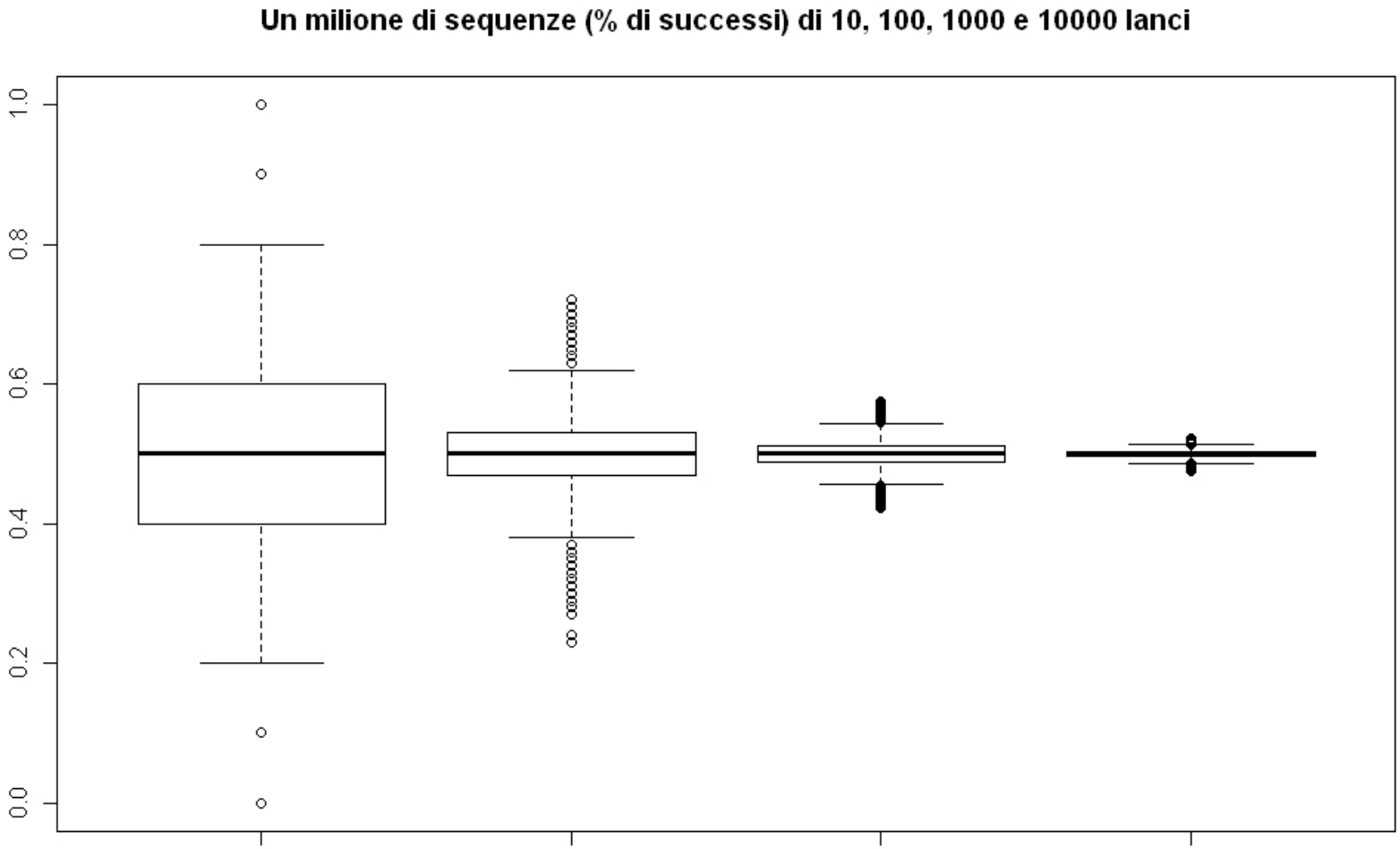
$$\begin{pmatrix} \textit{testa}\ \mathbf{0} & \textit{croce}\ \mathbf{1} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

# la spinosa questione dello standard error

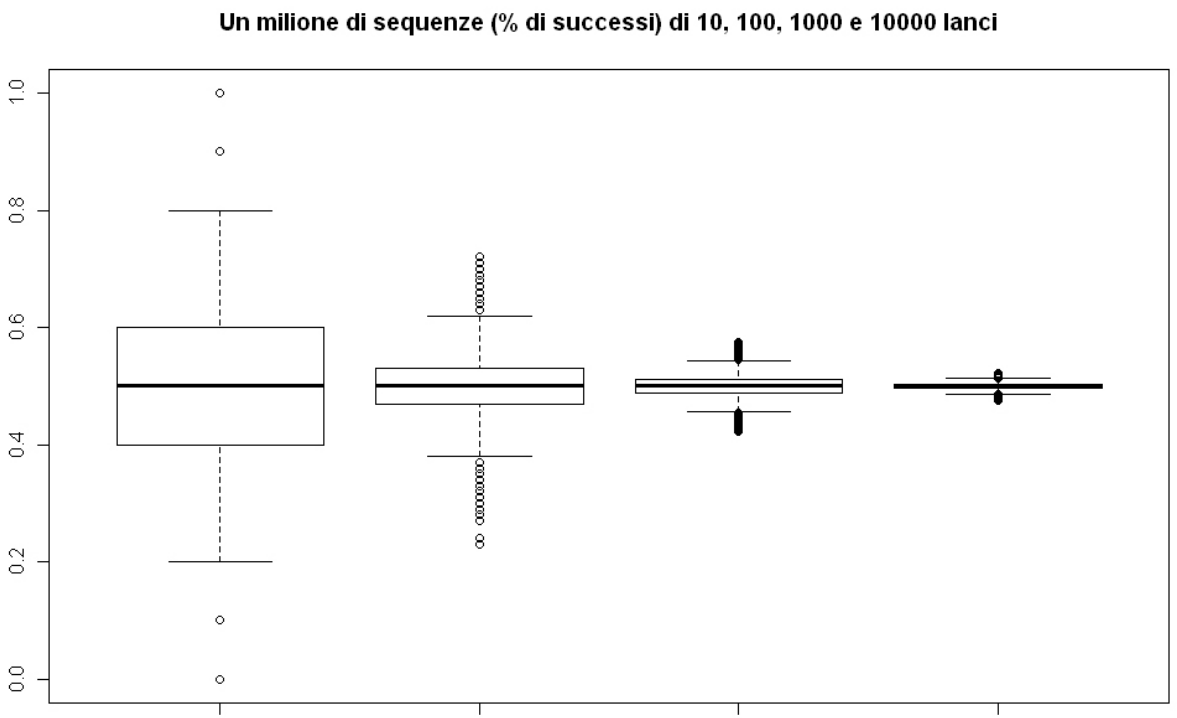
**domande**

- Quali 'errori' ci fa fare l'errore standard?

# la spinosa questione dello standard error



# la spinosa questione dello standard error




## cosa ci dice questo grafico?

- nei 'piccoli campioni' ci può essere molta variabilità rispetto ai 'grandi campioni'
- i 'grandi campioni' aumentano la **affidabilità** del parametro stimato (indice di centralità: media / mediana)

# The most dangerous hospital or the most dangerous equation?

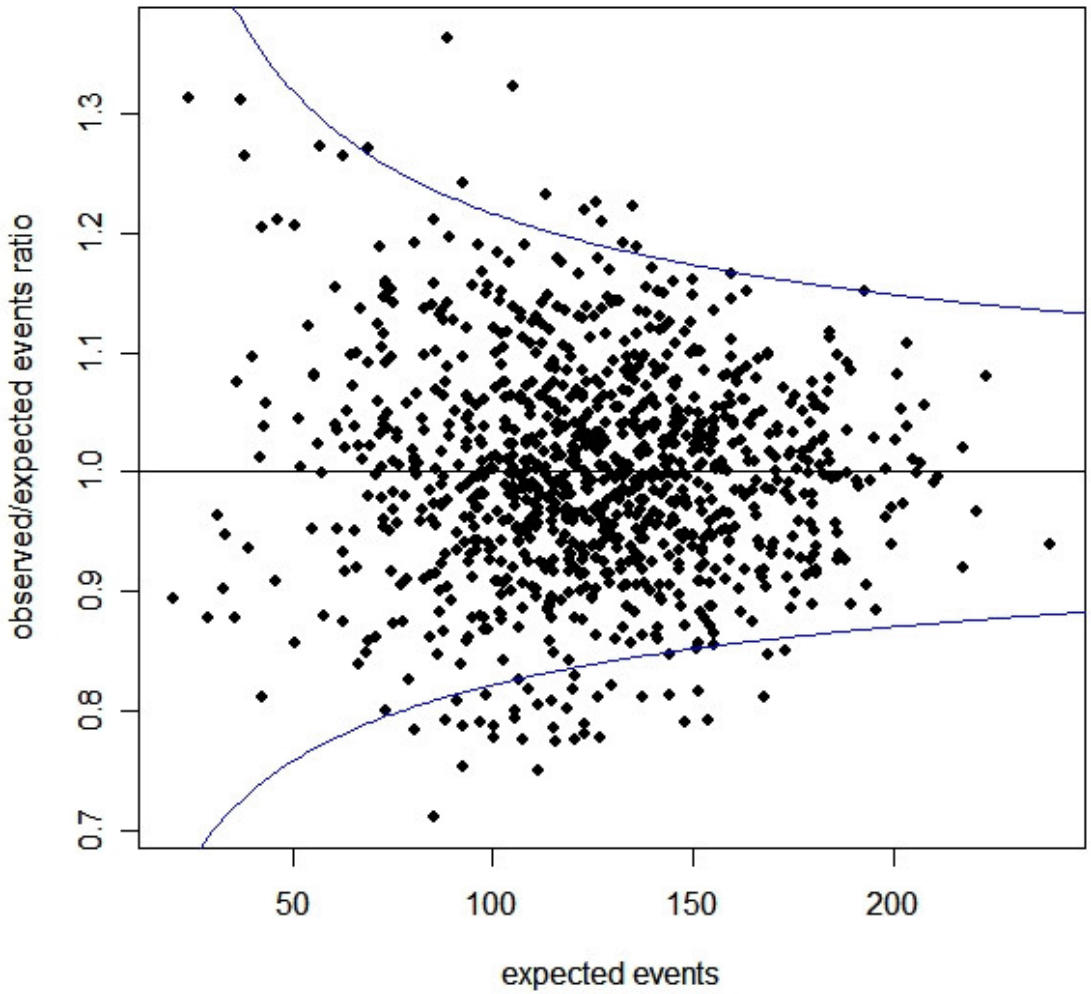
Yu-Kang Tu*[1,2] and Mark S Gilthorpe[1]

## Mortality in NHS hospitals

According to an article in the Daily Telegraph [1] (accessed online on 25/04/2007), the George Elliot Hospital (the only hospital run by the George Elliot Hospital NHS Trust) may have been the most dangerous hospital in England during the 2005/6 financial year. This is because its Hospital Standardised Mortality Ratio (HSMR) was 1.43, i.e. the number of patient deaths in this hospital was 43% higher than expected. In contrast, the hospital run by the Royal Free Hampstead Trust may have bee safest, since its HSMR was only 0.74, i.e. the numb patient deaths in this Trust was 26% lower than exp The source of information in the Daily Telegraph wa vided by an organization called Dr Foster Intelli which recently published a report entitled "How h is your hospital" [2], in which the performance of Trusts was assessed against several indicators, su post-operative mortality and emergency readmissio

**The most dangerous hospital or the most dangerous equation?**

Yu-Kang Tu*[1,2] and Mark S Gilthorpe[1]



R.it Cronaca

# In Friuli c'è il paese d'Italia dove si legge di più

*E' Pertegada. Ma in tutta la regione i lettori abbondano: 500 mila su un milione di persone*

di TOMMASO CERNO

PERTEGADA
www.pertegada.it/ ▾ Translate this page
PERTEGADA, frazione del comune di LATISANA- PROVINCIA DI UDINE (Italia), conta oggi circa 2350 abitanti, occupati nell'agricoltura (note sono alcune ...

# la spinosa questione dello standard error

**domande**

- Quando usare l'uno o l'altro?

# Quando usare l'uno o l'altro

Martin Bland:

Capita frequentemente, inoltre, di fare confusione tra "errore standard" e "deviazione standard": è comprensibile, dal momento che l'errore standard è una deviazione standard (della distribuzione della media campionaria); in questo contesto i due termini vengono spesso usati in modo interscambiabile. Faremo d'ora in avanti riferimento alla seguente convenzione: useremo "errore standard" quando vorremo misurare la precisione di una stima; useremo invece "deviazione standard" in riferimento alla variabilità di un campione, di una popolazione o di una distribuzione. Se dunque volessimo quantificare la bontà della stima della media delle

# la spinosa questione dello standard error

**domande**

- Quale è il 'pregio' che ha l'errore standard?

# il 'pregio' che ha l'errore standard

| esperimento<br>N | media<br>$\mu$ | deviazione st.<br>$\sigma$ | radice quadrata<br>$\sqrt{N}$ | errore st.<br>$\sigma/\sqrt{N}$ |
|---|---|---|---|---|
| 10 lanci | 5.0 | 1.6 | 3.2 | 0.50 |
| 100 lanci | 50.0 | 5.0 | 10 | 0.50 |
| 1000 lanci | 499.9 | 15.8 | 31.6 | 0.50 |
| 10000 lanci | 5000.3 | 50.1 | 100 | 0.50 |