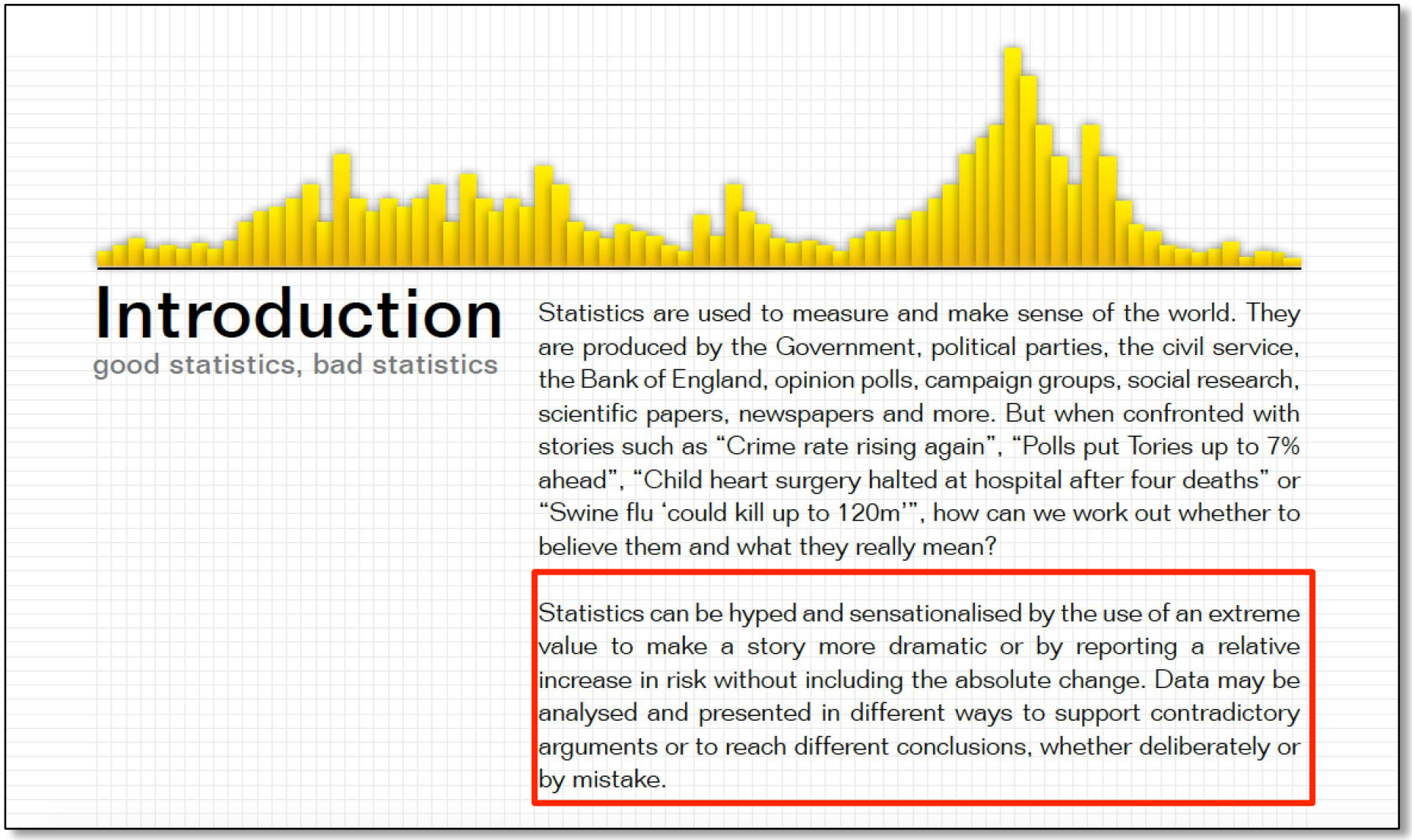
Pillola 4 («scatola degli attrezzi»)

Statistica

Si ringrazia il collega Prof. Lorenzo Magnea (dispense A.A. 2019-20) per gli spunti

A proposito di statistica



Natura e rischi dei dati statistici

Qual era la domanda?

Statistics are the product of conscious choices: **what to count** and **how to count it**. The results borrow from mathematics an air of precision and certainty. But choosing what to count, and how, comes down to human judgement about the best way to get the answer to the question, whether it is by designing and carrying out an experiment, a survey, a poll, a clinical trial, an observational study or a census.

"How big is the gap between the earnings of men and women? According to the Office for National Statistics (ONS), it is 12.8%. But the Government Equalities Office (GEO) says it is 23%. And the Equality and Human Rights Commission (EHRC) says it's 17.1%[1].

The differences in these figures arise from the different methods used to produce them: the ONS includes only full-time employees, excluding overtime and part-time workers. The GEO includes part-time workers because it says more women than men work part-time and it is wrong to exclude them. The EHRC figure uses the ONS data but compares the mean salaries not the median. It justifies this by saying that men are over-represented at one extreme of the earnings range, and women at the other.

Three figures – all of them right – but asking what is being compared and how it was calculated tells us why there is a difference." **Nigel Hawkes**

- Bisogna chiedersi con precisione che cosa sia stato misurato, e come
- Aspetti in parte secondari possono cambiare significativamente il risultato
- Le differenze possono essere volute, e sfruttate politicamente

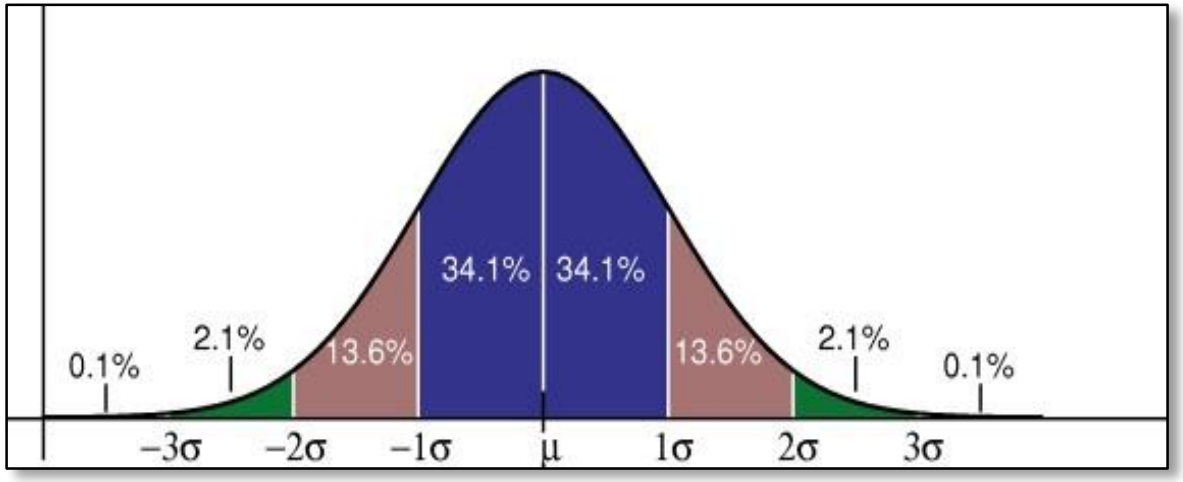
Campioni rappresentativi

Literary Digest carried out a survey before the 1936 US Presidential Election. It mailed out millions of ballot papers and got two million back; a huge sample, most of which backed the Republican candidate Alf Landon. But the addresses to which they had been sent came from a directory of car owners and from the telephone directory: a biased sample, since in 1936 only the better-off owned cars or had telephones. Franklin D Roosevelt, the Democrat, won the election in a landslide.

E’ necessario scegliere un campione rappresentativo della popolazione
… a meno che non si intenda manipolare le opinioni …

Un pollo a testa

- La statistica studia fenomeni che hanno una componente ***casuale***
- La prima informazione interessante è la ***media*** dei risultati
  - ✓ Questa informazione è spesso ***insufficiente***
- La seconda informazione interessante è la ***deviazione standard***
  - ✓ Tipicamente circa ***due terzi*** dei risultati distano dalla media meno di una deviazione standard
- In molti casi serve conoscere tutta la ***distribuzione*** dei dati
- Per un fenomeno casuale, se si hanno molti dati, la distribuzione ha la forma standard della curva ***gaussiana***. Ma ***non sempre***!



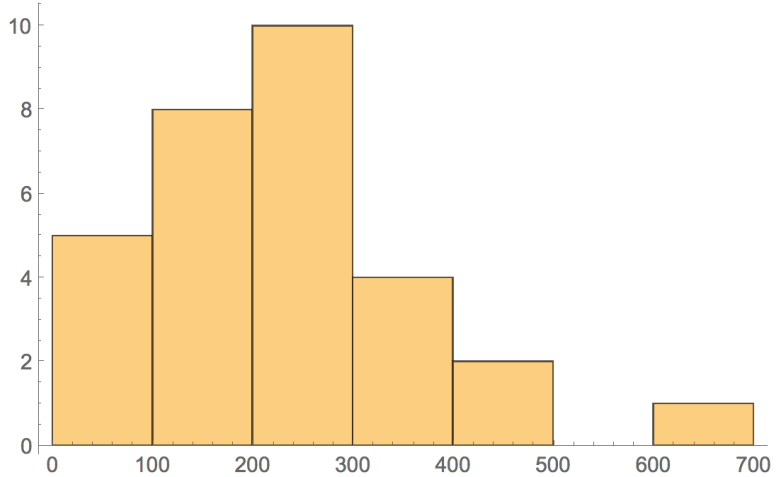
Un caso concreto

- Misuriamo quanti pedoni attraversano un certo incrocio ogni giorno per 30 giorni, per decidere la posizione di un semaforo, o di un'edicola
- Mostriamo i risultati nella forma di un istogramma
- Ci sono molte informazioni nella forma dell'istogramma non catturate dalla media e dalla deviazione standard

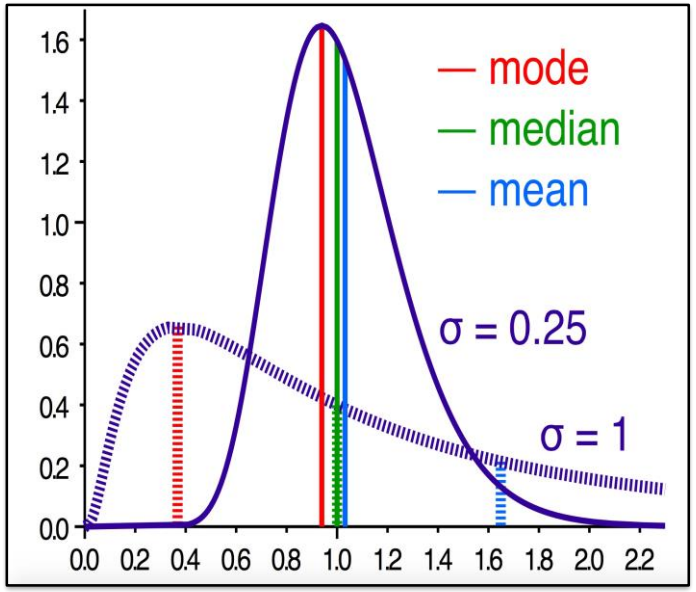
- Media: 228

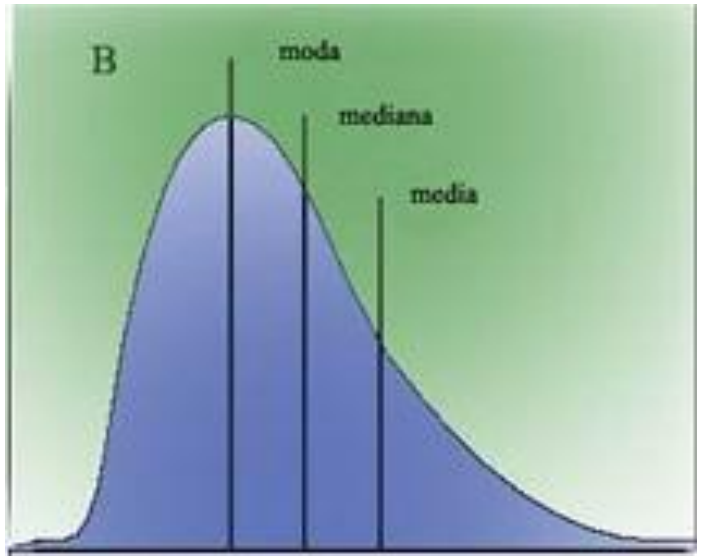
```
In[71]:= dati = {54, 26, 98, 23, 39, 144, 145, 187, 123, 199, 101,
          132, 188, 266, 255, 244, 227, 214, 276, 231, 287, 234,
          245, 388, 333, 344, 323, 406, 447, 675};

In[72]:= Histogram[dati, 8]
```

Out[72]=



Media, Mediana e Moda



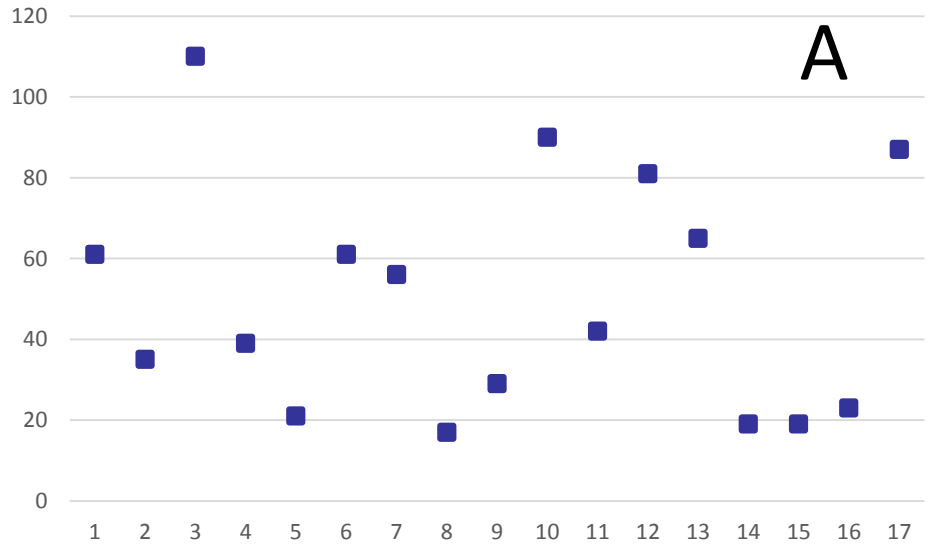


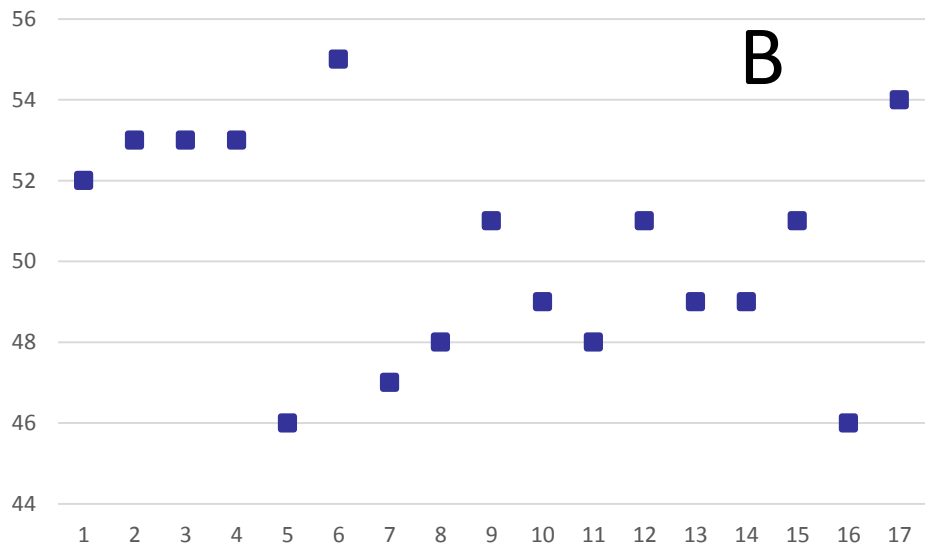
- Tre diversi modi di caratterizzare il "valore centrale" di una distribuzione
- La **media** è il rapporto tra i dati numerici e il numero dei dati
- La **moda** è il valore che si presenta con maggior frequenza
- La **mediana** è il valore centrale fra i dati numerici

Un esempio

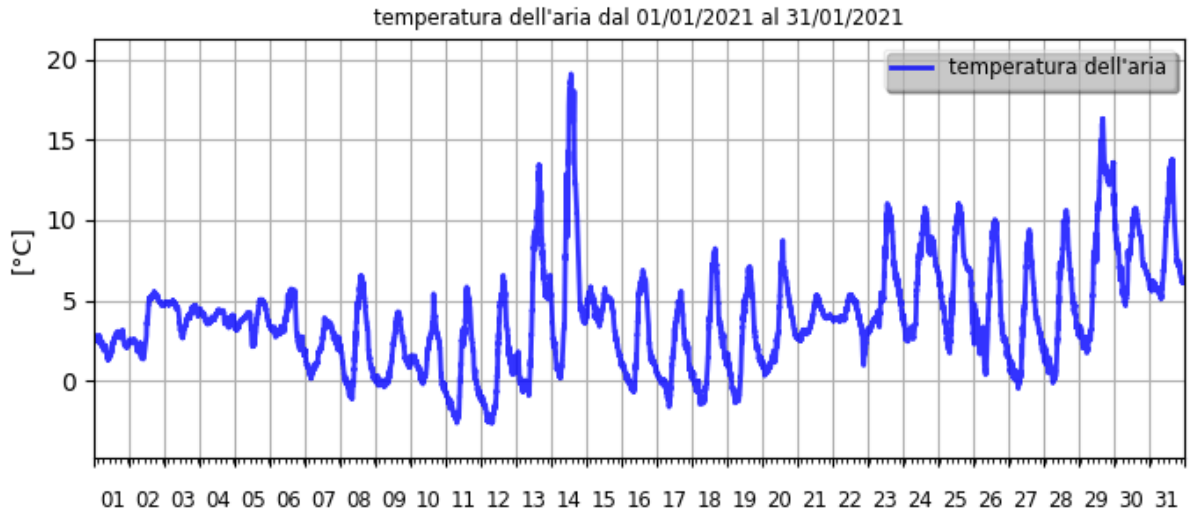
																		media	moda	mediana	dev. stand.
A	61	35	110	39	21	61	56	17	29	90	42	81	65	19	19	23	87	50.3	61	42	29.0
B	52	53	53	53	46	55	47	48	51	49	48	51	49	49	51	46	54	50.3	53	51	2.8

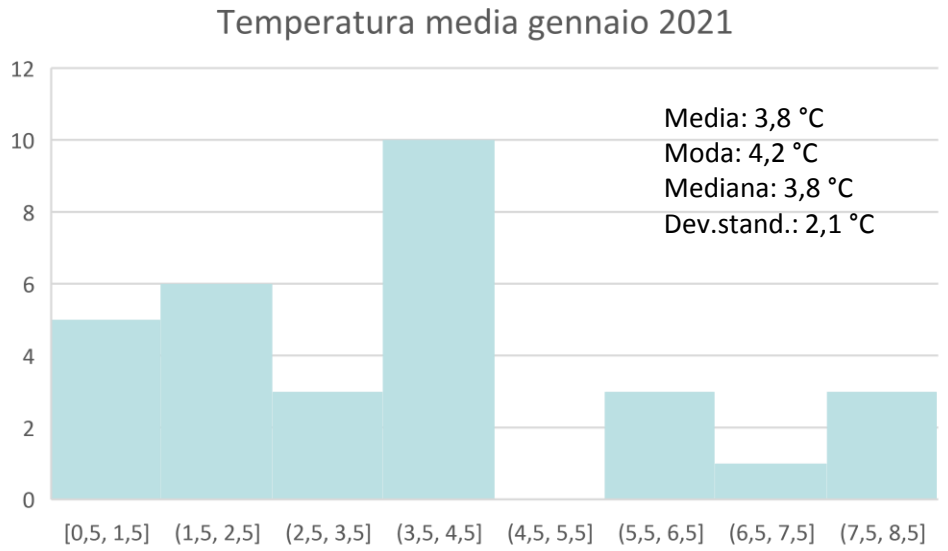
- L'insieme delle misure A ha media 50,3 e si distribuisce tra i valori estremi 17 e 110
- Anche l'insieme delle misure B ha media 50,3 ma si distribuisce tra i valori estremi 46 e 55
- Nei due casi, la media è identica, mentre variano gli altri parametri statistici
- Il parametro **deviazione standard** è quello che informa sullo «sparpagliamento» dei dati attorno al valore medio





Un secondo esempio





Media, Mediana e Moda

The way statistical data is summarised or presented can lead to wrong conclusions being drawn even if the statistics are correct. The results of studies are commonly captured by a single figure, but this figure might not represent everything that the study has found. Common pitfalls to be aware of are: there is more than one type of average, extreme values might not be very likely, and big (and small) numbers are difficult to comprehend without the context.

'Average' in news reports is often used to refer to the mean (that is, the sum of the listed values divided by the number of values in the list). But the mean doesn't always give us the most useful information: we might need the median (the middle point) or the mode (the most common value). If we wanted to know the typical salary earned in the UK the most representative value would be the median salary. This is because a few thousand people earning millions of pounds a year will affect the mean more than the multitudes of people earning tens of thousands of pounds a year.

For example, if you had a room with ten teachers all earning between £20,000 - £30,000, with a mean salary of £24,900 and a median salary of £25,000 and then someone who earns a million pounds walked into the room, the mean would increase to £114,000 but the median would hardly change. By using the median this distortion is reduced, providing a more representative average salary.

"You might want to know the average number of children in each household in a particular place. You work out that the mean is 2.3 by adding up all the children in the area and dividing by the number of households. But this number wouldn't tell you that the most common household setup has no children (that is, the mode is 0). If you ordered the households according to how many children they had and the household in the middle had just one child, the median would be 1." **Harriet Ball**

Averages are useful ways of encapsulating data, but when averages are used consider what sort of average it is and whether it is representative of what we are trying to find out.

La scelta tra media, mediana e moda dipende dalla domanda che si pone

Significatività statistica

3. How sure are we?

When evaluating data statisticians have to decide whether the results they are seeing are 'statistically significant'. Even if a result is statistically significant it doesn't mean it is practically significant or of importance to society. Confidence intervals give the scale of potential uncertainties in counting, measuring or observing data.

What does 'statistically significant' mean?

"To be honest, it's a tricky idea. It can tell us if the difference between a drug and a placebo or between the life expectancies of two groups of people, for example, could be just down to chance.

If a result is statistically significant we can be fairly confident that something real is going on. It means that a difference as large as the one observed is unlikely to have occurred by chance alone.

Statisticians use standard levels of 'unlikely'. Commonly they use significant at the 5% level (sometimes written as p=0.05). In this case a difference is said to be 'significant' because it has a less than 1 in 20 probability of occurring if all that is going on is chance." David Spiegelhalter

Occorre stabilire caso per caso un criterio di decisione (e poi verificare)

Significatività esistenziale

- Il 18 Ottobre 1995 l'"UK Committee for the Safety of Medicines" comunica che un certo tipo di pillola contraccettiva "**raddoppia il rischio di trombosi venosa**".
- La notizia viene riportata con enfasi dai media ("La pillola di tipo X causa la trombosi venosa").
  - Si verifica un calo misurabile nell'utilizzo di contraccettivi orali.
  - Nel 1996 nel Regno Unito nascono 26000 bambini più che nel 1995, e si effettuano 13600 aborti in più.
- La probabilità che una donna in età fertile che prende contraccettivi orali soffra di una trombosi venosa è circa 1/7000. Si tratta di una condizione quasi sempre curabile.
- Una stima non ufficiale suggerisce che il numero di morti per trombosi venose in assenza del comunicato sarebbe stato dell'ordine di: uno.

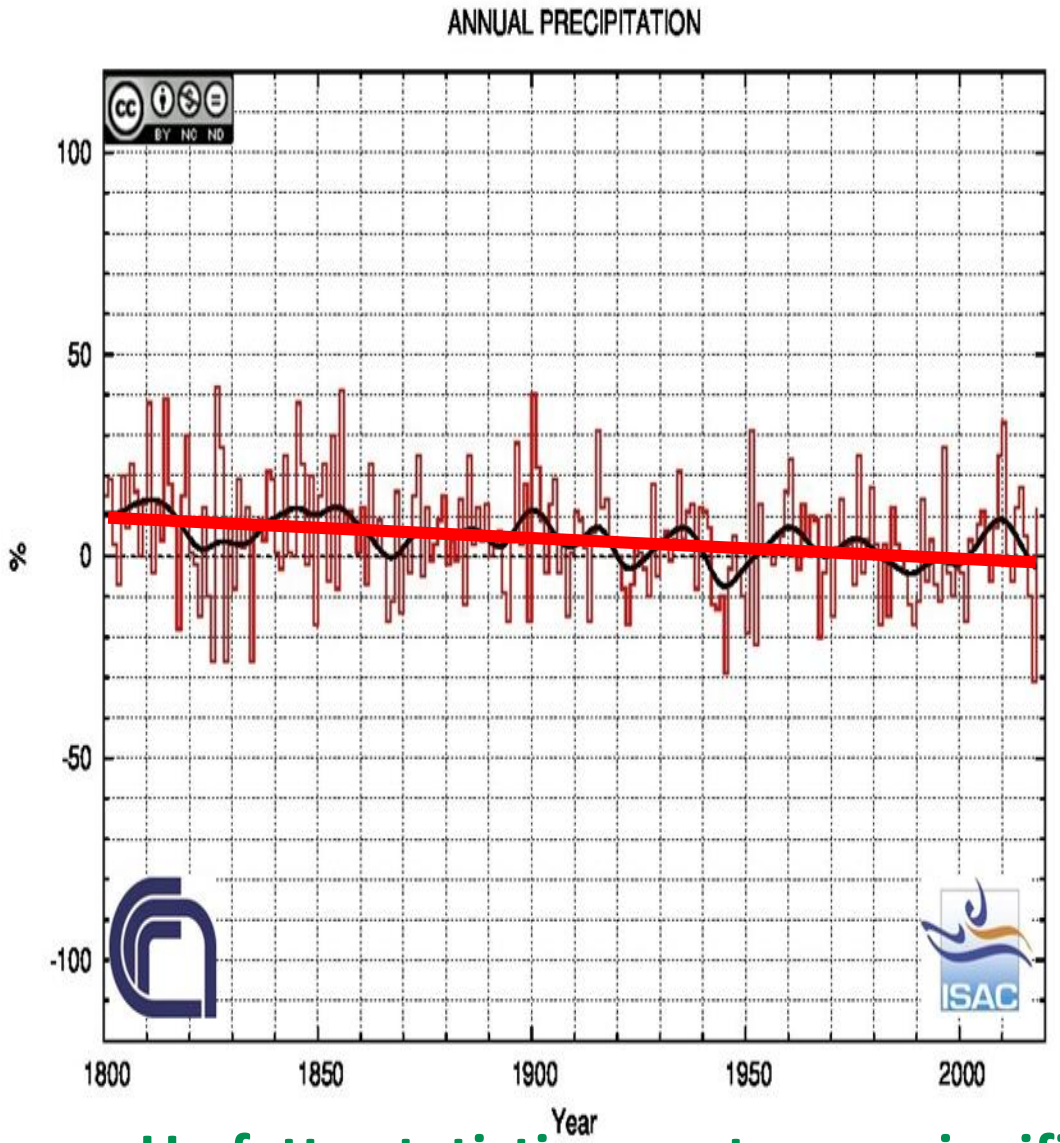
Un fatto statisticamente significativo non è necessariamente importante

Significatività esistenziale

- Uno studio della City University of New York è stato citato dai media riportando che "**il rischio di morte accidentale per bambini accuditi a casa è sette volte superiore a quello degli asili nido**".
- Licenziare subito la baby sitter?
- Se si guardano i numeri, sono piccolissimi: 1,6 decessi per 100000 bambini (a casa), e 0,23 decessi per 100000 bambini all'asilo nido.
- Due numeri così piccoli sono probabilmente entrambi contenuti negli intervalli di errore (sistematico e statistico).
- I numeri osservati sono una frazione minuscola dei decessi dovuti ad altri tipi di incidenti (circa l'uno per cento, per esempio una dozzina su 1100 negli Stati Uniti nel 2010).
- Se portare il bambino al nido comportasse un lungo tragitto in macchina, occorrerebbe tenere conto della molto maggiore pericolosità del traffico (circa 80 decessi negli USA nel 2010). Ecc…

Un fatto statisticamente significativo non è necessariamente importante

Significatività statistica



- Il grafico a sinistra mostra la variazione percentuale delle precipitazioni annuali in Italia dal 1800 a oggi
- Il range di variazione oscilla tra -30% e 40%
- La retta di trend rivela un calo del 10%
- Il trend NON risulta statisticamente significativo
- Tuttavia le precipitazioni si sono ridotte davvero del 10%, in media, sia pure con forti oscillazioni interannuali
- Questo calo è riscontrato nelle portate dei fiumi: ho meno acqua disponibile!

Un fatto statisticamente non significativo non è necessariamente non importante

Intervalli di (s)fiducia

As many as 5 million Americans infected with H1N1

Bank says inflation may reach 4%

Temperature 'could rise by 11 degrees', says study

Council job cuts 'could be as high as 100,000'

Fuel costs could rise by 40%, MPs told

Global unemployment could rise by 50m

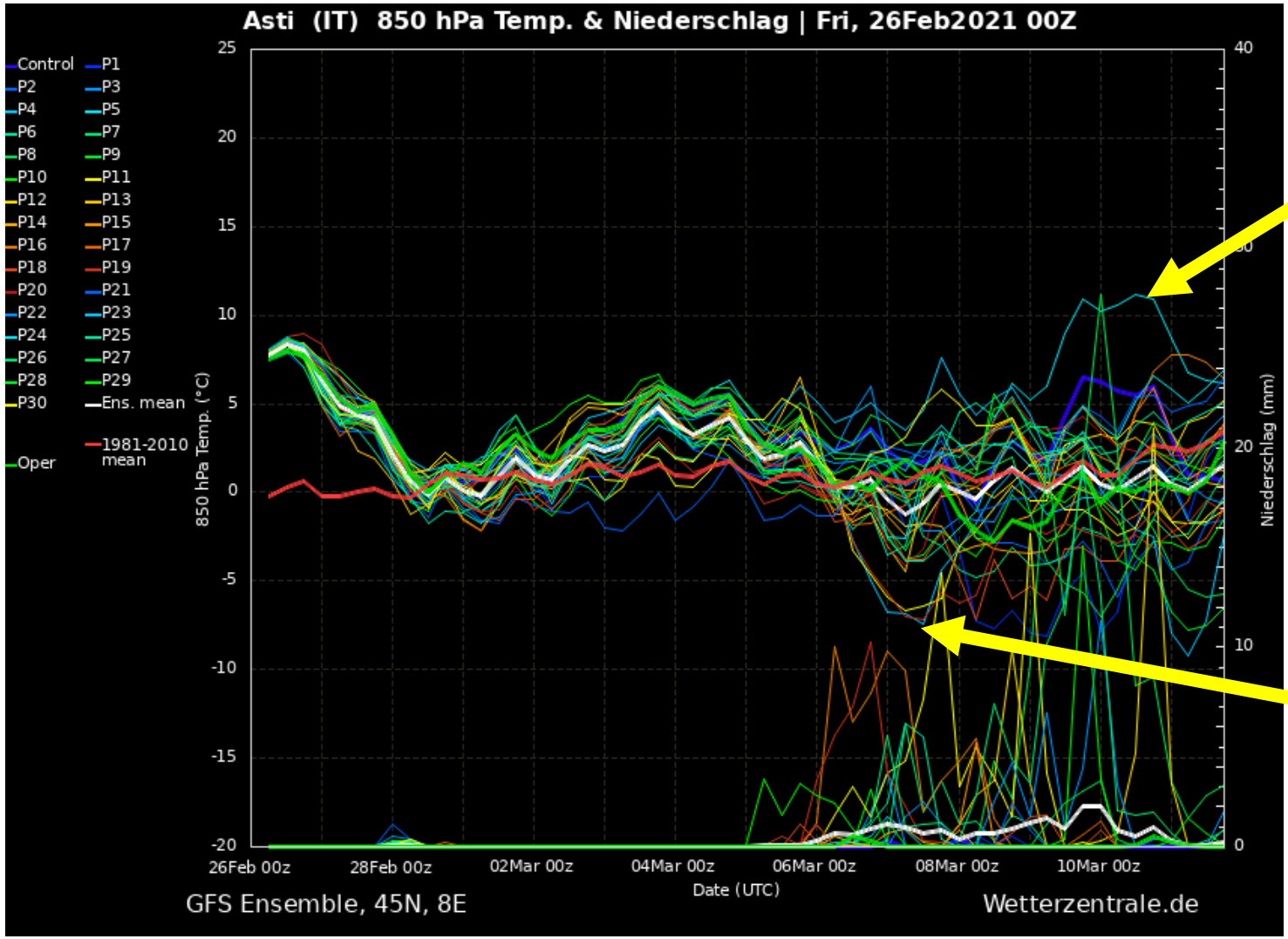
Most projections are given as a range, because we can't know precisely what is going to happen in the future. When the emphasis is on the most **extreme value** – "as many as 5 million Americans infected with H1N1", "Council job cuts 'could be as high as 100,000'" – it sounds alarming until you consider the likelihood (the probability) of what's being reported actually happening.

"'Could be as high as', 'may reach' ... phrases such as these hint at mischief at work, depending as they do on the most extreme possibilities rather than the most likely. Headlines like these might more accurately read: 'could be, but most likely won't'.

'Temperature could rise by 11 degrees C'[2]. This was one result from a model of climate sensitivity to rising CO_2 levels. But the model was run 2,000 times and this outcome was generated only once. The most common result was that temperature would rise by 3°C . Whilst 11°C is possible it's not the most likely result, but it was still widely reported. When reading such stories it is important to consider is it likely, given what else we know?" **Michael Blastland**

Fornire solo massimi e minimi valori di variabili statistiche è una manipolazione

Intervalli di (s)fiducia



Il più caldo

Nota: immagini selezionate a titolo esemplificativo e non collegate ai valori mostrati

Il più freddo

Selezionare i casi estremi per fare notizia è manipolazione dell'informazione

Percentuali di cosa

4. Percentages and risk; knowing the absolute and relative changes

"Bacon butties give you colon cancer", "passive smoking causes dementia", "knife killings at new high" – how should we react to such stories? We need to work out how many people are really affected and what our individual risks are. To understand the importance of any increase or decrease we need to know both the absolute and relative change.

Take the story about bacon sandwiches giving you colon cancer. It reported a 20% rise in the risk of colon cancer from eating red or processed meat. This may sound alarming, but what does it mean? It depends on how large your risk of getting colon cancer is to start with.

A person has around a 5% chance of getting colon cancer during their lifetime – the 'absolute' risk. If you eat a bacon sandwich every day you increase your risk of getting colon cancer by 20% – the 'relative' risk increase. So what it means is that your lifetime risk of getting colon cancer is now 6%, an increase of 1% (that is, 20% of 5% = 1%).

"Why do reports prefer to talk about relative percentage risks without mentioning the absolute risk? The suspicion must be that this allows the use of 'bigger numbers': 20% is big enough to be a scare, the absolute change 1% or even 1 person in every 100, is less disturbing." **Michael Blastland**

Se vengono forniti incrementi è importante rapportarli ai dati assoluti

Percentuali di cosa

"Recent newspaper articles reported that '9 in 10 people carry a gene that increases chance of high blood pressure'. The actual study had found a genetic variant in 1 in 10 people that **reduced** the risk of high blood pressure – this would not have received much press coverage, but by shifting the attention to the group who did not have this reduced risk the story became 'news'." **David Spiegelhalter**

Representing something as a proportion does not tell us what the absolute numbers are. A Whiskas advert told us that "8 out of 10 cats prefer Whiskas", but not how many cats had been asked... This was then changed to "8 out of 10 owners that expressed a preference said their cats preferred Whiskas". A more likely situation but we're still not told how many owners were asked and how many of these expressed an opinion.

Or, for example, the actual number of violent crimes might not have changed in a few years, but if the number of other types of crime has gone down then violent crimes as a proportion of the total numbers of crimes will increase.

"Before reacting to a percentage you have to think what it is really telling you and to do that you need to put it in context. Take, for example, the statistic that 99% of deaths in the first four weeks of life occur in developing countries. Although that sounds horrifying, around 90 per cent of all births take place in developing countries. And so the chances of a baby dying in its first four weeks are 'only' 11 times greater there – bad enough, in all conscience." **Christina Pagel**

Se vengono fornite percentuali è importante capire di quale quantità

Riassumendo

- Le indagini statistiche sono spesso soggette a ipotesi
- Occorre sapere con precisione che cosa è stato misurato
- Spesso serve conoscere bene la distribuzione della probabilità
- La definizione di significatività statistica ha elementi soggettivi
- Statisticamente significativo non significa rilevante
- Le statistiche possono essere manipolate selezionando campioni non rappresentativi, o fornendo dati parziali