Βιοστατιστική

Εισαγωγή στη Στατιστική για Βιολόγους

Αριστοτέλης Παπαγεωργίου, Τμ. ΜΒΓ ΔΠΘ, apapage@mbg.duth.gr

Πηγές υλικού

- Διαφάνειες και ασκήσεις του Theophanis Tsandilas (National Institute for Research in Digital Science and Technology, INRIA, Γαλλία)
- Διαφάνειες και ασκήσεις του Σπύρου Γαλατσίδα (Τμ. Δασολογίας & Διαχείρισης Περιβάλλοντος & Φυσικών Πόρων, ΔΠΘ)
-

Early beginnings

Photo: Matthias Kabel

450 BC Hippias of Elis uses the average value of the length of a king's reign (the mean) to work out the date of the first Olympic Games, some 300 years before his time.

400 BC In the Indian epic the *Mahabharata*, King Rtuparna estimates the number of fruit and leaves (2095 fruit and 50000000 leaves) on two great branches of a vibhitaka tree by counting the number on a single twig, then multiplying by the number of twigs. The estimate is found to be very close to the actual number. This is the first recorded example of sampling – "but this knowledge is kept secret", says the account.

500 400 300 200 100

431 BC Attackers besieging Plataea in the Peloponnesian war calculate the height of the wall by counting the number of bricks. The count was repeated several times by different soldiers. The most frequent value (the mode) was taken to be the most likely. Multiplying it by the height of one brick allowed them to calculate the length of the ladders needed to scale the walls.

Mathematical foundations

iStock/Thinkstock

1560 Gerolamo Cardano calculates probabilities of different dice throws for gamblers.

1560

1580

1654 Pascal and Fermat correspond about dividing stakes in gambling games and together create the mathematical theory of probability.

1640

1761 The Rev. Thomas Bayes proves Bayes' theorem – the cornerstone of conditional probability and the testing of beliefs and hypotheses.

1713 Jacob Bernoulli's *Ars conjectandi* derives the law of large numbers – the more often you repeat an experiment, the more accurately you can predict the result.

1749 Gottfried Achenwall coins the word "statistics" (in German, *Statistik*); he means the information you need to run a nation state.

1808 Gauss, with contributions from Laplace, derives the normal distribution – the bell-shaped curve fundamental to the study of variation and error.

1720

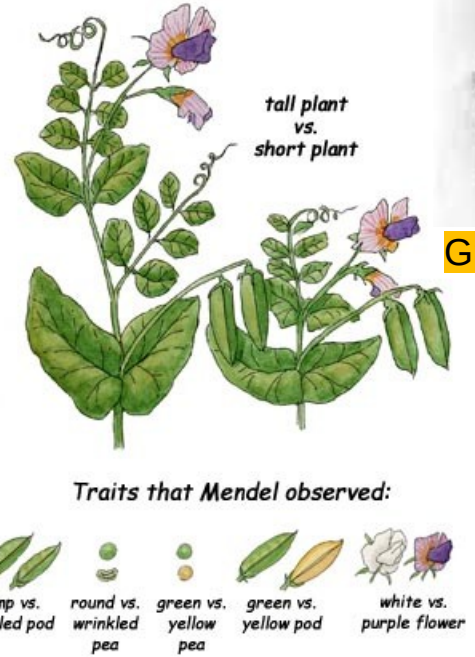
1740

1800

Gregor Mendel

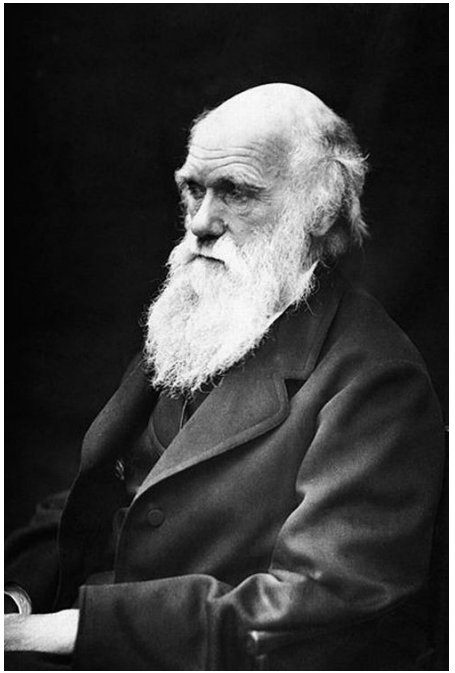
- Το πρώτο βιολογικό πείραμα
- Από τα αποτελέσματα (αναλογίες παρατηρήσεων) κατέληξε στη θεωρία
- Ανακάλυψε τη σωματιδιακή και δυαδική φύση της κληρονομικότητας
- Δεν έκανε στατιστική ανάλυση για έλεγχο των αναλογιών 1:3
- Είχε δίκιο!

Gregor Mendel (1822-1884)



Charles Darwin

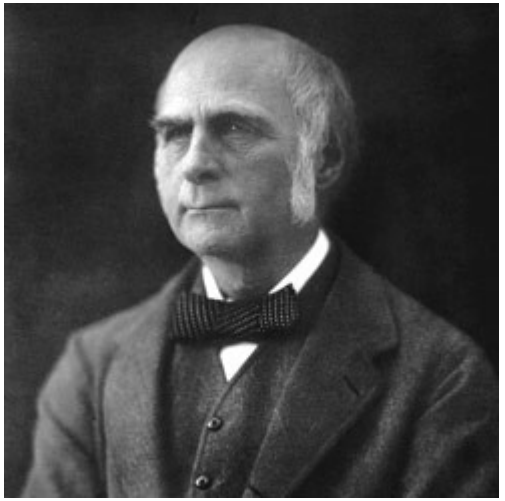
- Περιέγραψε τον μηχανισμό της φυσικής επιλογής στην εξέλιξη
- Θεμελίωσε μια δυναμική εξελικτική θεωρία
- Δεν χρησιμοποίησε καθόλου μαθηματικά
- Δεν εξήγησε τον μηχανισμό της κληρονόμησης



Charles Darwin (1809–1882)

Francis Galton

- Στατιστικός, ψυχολόγος, κοινωνιολόγος, ανθρωπολόγος, εξερευνητής, γενετιστής
- Ιδρυτής της “ευγονικής” και του “κοινωνικού δαρβινισμού”
- Μελέτησε τις βιομετρικές διαφορές στους ανθρώπινους πληθυσμούς
- Πρότεινε εναλλακτική θεωρία κληρονομικότητας από αυτή του Mendel (βιομετρική σχολή)
- Ανακάλυψε τις έννοιες “συσχέτιση” και “παλινδρόμηση”



Francis Galton (1822–1911)

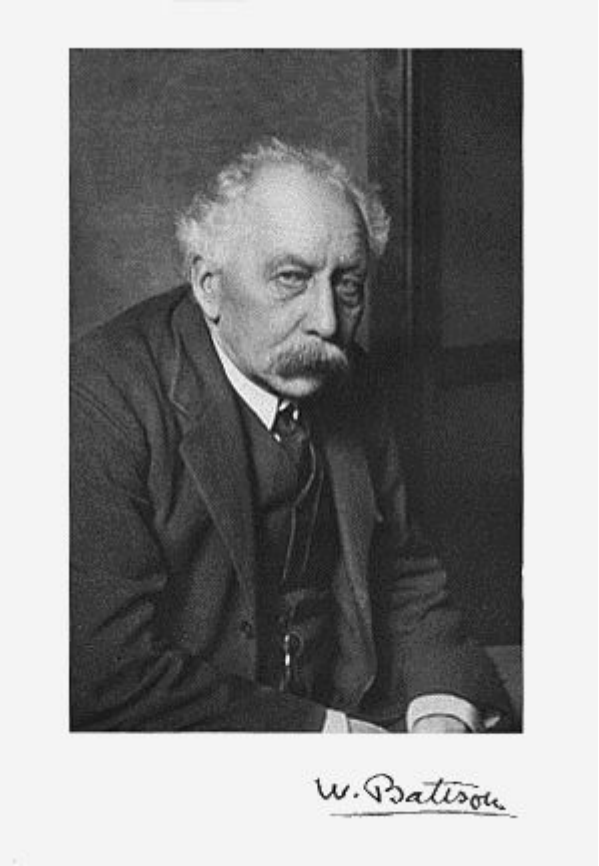
Karl Pearson

- Υπερασπιστής (ιδρυτής) της βιομετρικής σχολής και σημαντικός οπαδός της ευγονικής
- Ανέπτυξε σημαντικές στατιστικές μεθόδους (τεστ χ^2, τυπική απόκλιση)
- Καθιέρωσε τους συντελεστές “συσχέτισης” και “παλινδρόμησης”
- Παρά την ευγονική, ταυτόχρονα συνετέλεσε στην πρόοδο της επιστήμης και στον κοινωνικό της ρόλο

Karl Pearson (1857-1936)

William Bateson

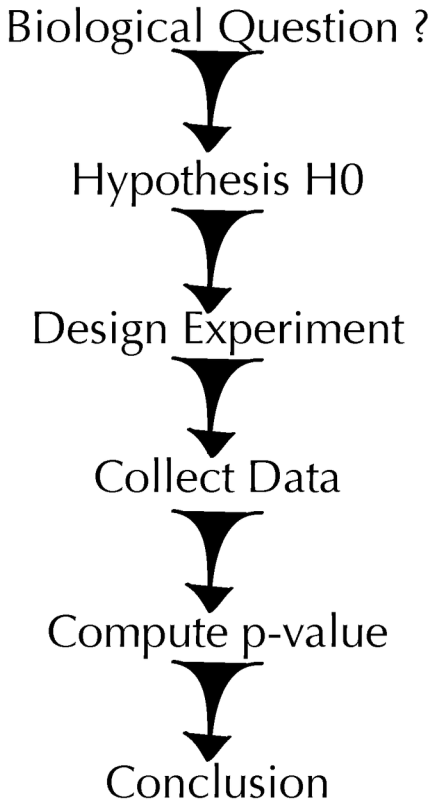
- Υπερασπιστής της Μενδελικής σχολής, ιδρυτής της γενετικής επιστήμης
  - Μαζί με τους Charles Davenport και Wilhelm Johannsen
- Τη δεκαετία του 1930 οι απόψεις των βιομετρικών είχαν πλέον απορριφτεί και η μενδελική κληρονόμηση κυριαρχεί

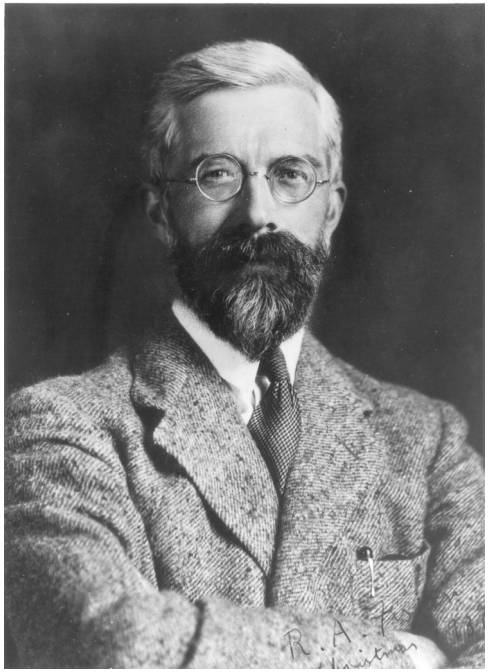


William Bateson (1861-1926)

Έλεγχος υποθέσεων και ANOVA

- Γραμμική άλγεβρα, **μεικτά μοντέλα** (γενετική βελτίωση ζώων και φυτών)
- Έλεγχος υποθέσεων
- Ανάλυση διακύμανσης (ANOVA)
- Fishers Exact Test
- *Statistical Methods for Research Workers (1925)*
- *The Genetical Theory of Natural Selection (1930)*

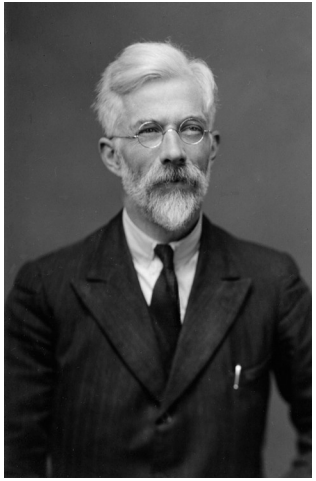




R.A. Fisher (1890-1962)

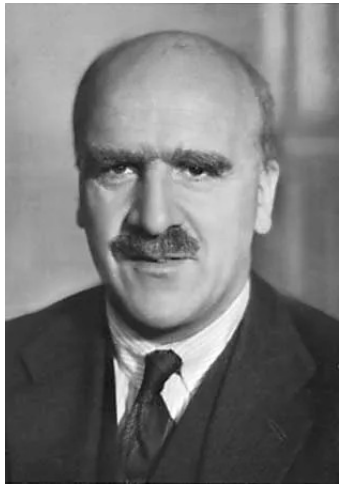
Η “νέα σύνθεση”

- Ερμηνεία της εξελικτικής θεωρίας του Δαρβίνου μέσα από τη στατιστική των μενδελικών
- Η βιολογία και η γενετική είναι ποσοτικές επιστήμες



R.A. Fisher (1890-1962)

R.A. Write (1890-1962)

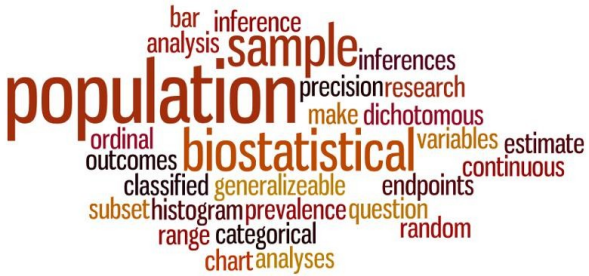


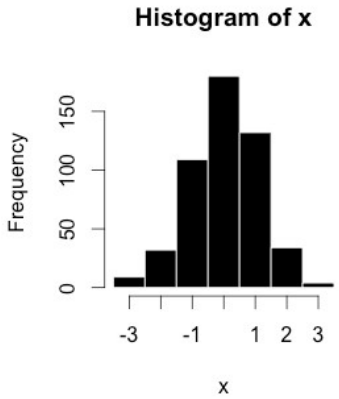
J.B.S. Haldane (1892-1964)

Η στατιστική

- Ο κλάδος της επιστήμης που ασχολείται με:
  - Το σχεδιασμό της συλλογής δεδομένων
  - Την οργάνωση, επεξεργασία και παρουσίαση δεδομένων και αποτελεσμάτων επεξεργασίας
  - Την ανάλυση των δεδομένων, τη διατύπωση συμπερασμάτων και τη λήψη αποφάσεων

Βιοστατιστική

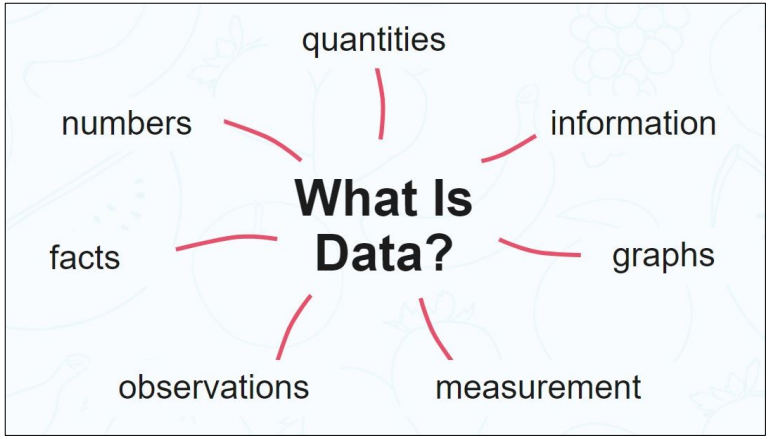




- Η εφαρμογή των κανόνων και των εργαλείων (μεθόδων) της στατιστικής για την επεξεργασία ερωτημάτων βιολογικού χαρακτήρα
  - Υγεία - ιατρική
  - Αγροτική παραγωγή
  - Οικολογία - βιοποικιλότητα

Αριθμοί και δεδομένα

- Οι αριθμοί είναι αφηρημένα σύμβολα που χρησιμοποιούμε για μετρήσεις και παρατηρήσεις
- Τα **δεδομένα** (data) αντιπροσωπεύουν πραγματικές οντότητες του φυσικού κόσμου
  - Τα δεδομένα μπορεί να είναι και αριθμοί

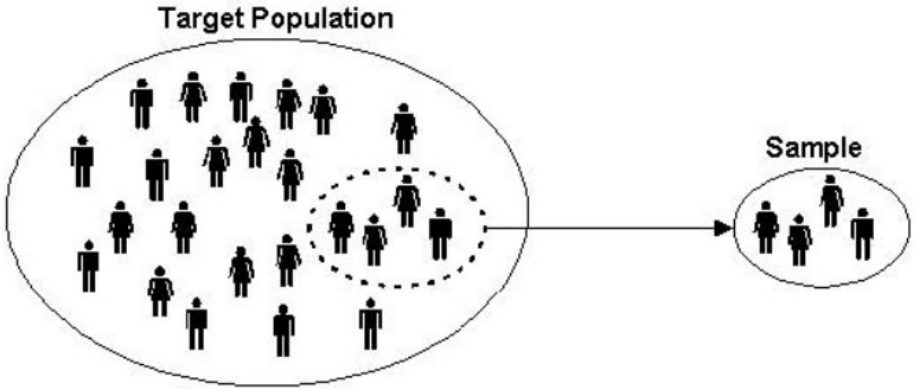


Αριθμοί και “περιεχόμενο”

- Μια σειρά αριθμών από μόνη της δεν έχει νόημα
  - π.χ. 8 10 8 12 14 13 12 13
- Προσθέτοντας **περιεχόμενο** στους αριθμούς, αυτοί γίνονται δεδομένα γιατί αποκτούν νόημα, π.χ.
  - Οι ηλικίες 8 παιδιών
  - Οι βαθμοί 8 μαθητών / μαθητριών
  - Ο αριθμός των σφαλμάτων σε 100 πειράματα που κατέγραψαν 8 ερευνητές/ερευνήτριες
  - Το μήκος 8 διαφορετικών φύλλων ενός φυτού

Δειγματοληψία

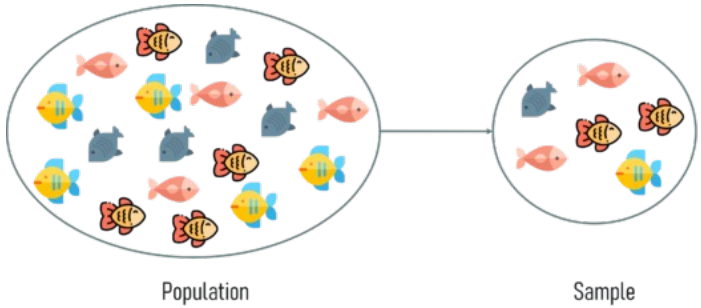
- Το περιεχόμενο (νόημα) αφορά τη διαδικασία παραγωγής των δεδομένων
- **Πληθυσμός**: το σύνολο των μονάδων της μελέτης
- **Δείγμα**: ένα υποσύνολο του πληθυσμού για το οποίο συγκεντρώνονται δεδομένα και γίνονται αναλύσεις



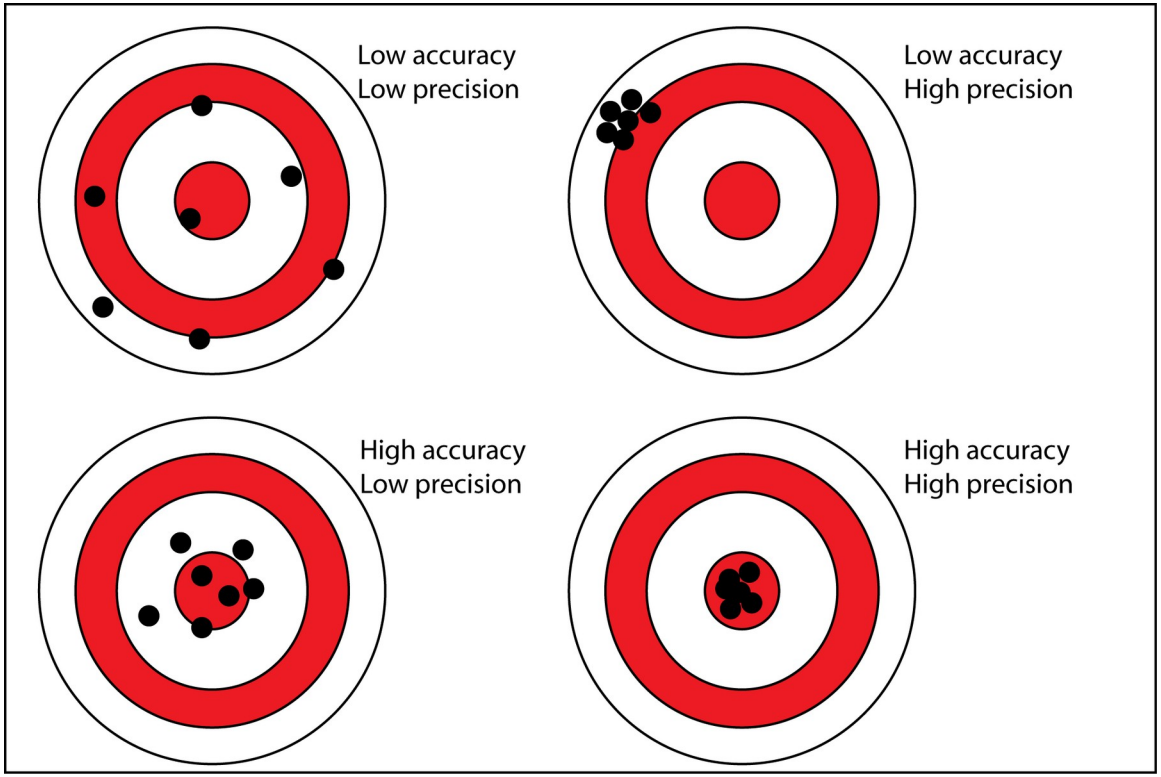
- Η δειγματοληψία είναι κεντρική έννοια για την κατανόηση της στατιστικής και θεμελιώδες συστατικό των περισσότερων στατιστικών αναλύσεων

Στατιστική

- Το δείγμα, σαν υποσύνολο του πληθυσμού, δεν είναι απαραίτητο να είναι αντιπροσωπευτικό
- Η πληροφορία του δείγματος είναι κάποιες φορές αβέβαιη
- **Ο ρόλος της στατιστικής είναι η αντιμετώπιση (διόρθωση) αυτής της αβεβαιότητας**

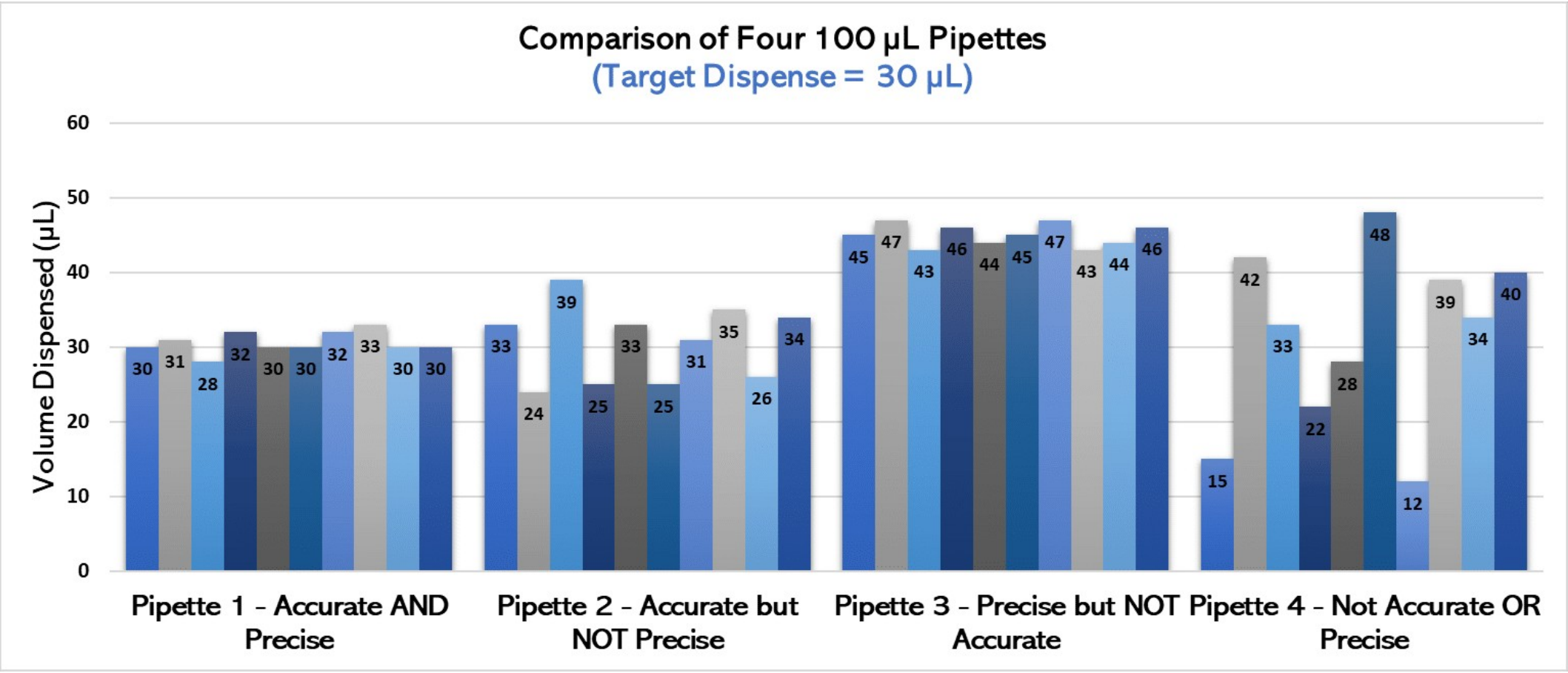


Ακρίβεια και αξιοπιστία των μετρήσεων



- **Ακρίβεια** (accuracy) είναι η εγγύτητα των μετρήσεων προς την αληθινή τιμή του υπό μελέτη χαρακτηριστικού
- **Αξιοπιστία** (precision) ή επαναληψιμότητα (repeatability) είναι η εγγύτητα μεταξύ ανεξάρτητων μετρήσεων του ίδιου χαρακτηριστικού, όταν μετριέται κάτω από τις ίδιες προϋποθέσεις και συνθήκες.

Ακρίβεια και αξιοπιστία μιας ογκομετρικής πιπέτας



Βελτίωση ακρίβειας και αξιοπιστίας

- Πώς μπορούμε να βελτιώσουμε την ακρίβεια και την αξιοπιστία στο προηγούμενο παράδειγμα;
- Πώς μπορούμε να βελτιώσουμε την ακρίβεια και την αξιοπιστία σε μια δειγματοληψία;

Πιθανότητες

- Η αβεβαιότητα συνήθως ποσοτικοποιείται και αποδίδεται ως **πιθανότητα**
- Η πιθανότητα να συμβεί ένα γεγονός μπορεί να γραφτεί ως *P(x)* ή *Pr(x)*
  - Παίρνει τιμές από 0 ως 1
- Μια λογική (αλλά όχι μοναδική) ερμηνεία μιας πιθανότητας:
  - η σχετική συχνότητα με την οποία ένα γεγονός X εμφανίζεται σε βάθος χρόνου

Ένα παράδειγμα: ρίψη νομίσματος

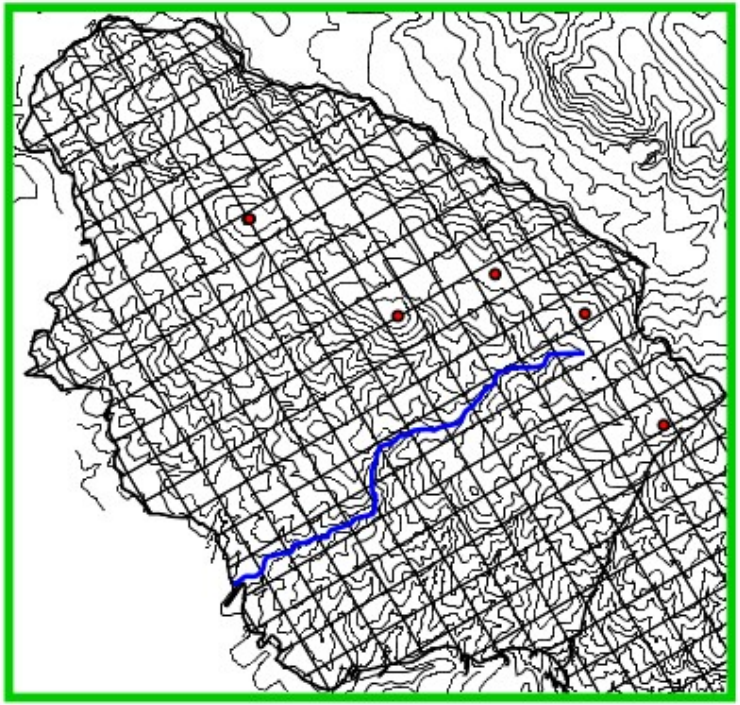
- Θεωρούμε ένα πείραμα ρίψης με ένα “σωστό” νόμισμα
  - Μακροπρόθεσμα (π.χ. μετά από 1.000.000 δοκιμές), θα έχουμε έναν (περίπου) ίσο αριθμό των δύο γεγονότων
    - Κορώνα ή γράμματα
  - Pr(Head) = Pr(Tail) = 0,5
- Πριν συνεχίσουμε πρέπει να δούμε την έννοια του “πληθυσμού”

Στατιστικός πληθυσμός (population)

- Είναι ένα σαφώς καθορισμένο σύνολο ή ολότητα για το οποίο ενδιαφερόμαστε και από το οποίο συγκεντρώνουμε δεδομένα
  - Οι κάτοικοι της Αλεξανδρούπολης σε μια έρευνα γνώμης
  - Οι ασθενείς μιας χώρας για μια συγκεκριμένη ασθένεια, τους οποίους παρακολουθούμε και καταγράφουμε την πρόοδο
  - Οι πελεκάνοι μιας λίμνης, για τους οποίους καταγράφουμε φύλο, ηλικία και σωματικές διαστάσεις
  - Τα δέντρα ενός είδους σε μια περιοχή, από τα οποία μαζεύουμε ιστό για γενετικές αναλύσεις
  - Οι καταχωρήσεις σε μια βάση δεδομένων (cases)

Στατιστικές μονάδες & μέλη του πληθυσμού (statistical units, cases)

- άτομα (βιολογικά όντα)
- αντικείμενα
- επιφάνειες
- σημεία
- γεγονότα



Πληθυσμός και διαδικασία δειγματοληψίας

- Στο παράδειγμα του νομίσματος (γεγονός), ο πληθυσμός ισοδυναμεί με ένα απείρως μεγάλο και υποθετικό δείγμα
- Μια καλά σχεδιασμένη έρευνα, θα χρησιμοποιεί μια διαδικασία δειγματοληψίας που σχετίζεται με τους στόχους της έρευνας αυτής
- Η διαδικασία δειγματοληψίας είναι συνήθως ατελής, π.χ. λόγω μεροληψίας στο δείγμα
  - Μια καλά σχεδιασμένη μελέτη θα ελαχιστοποιήσει τον αντίκτυπο τέτοιων προβλημάτων

Ένα παράδειγμα: ρίψη νομίσματος

- Μια ερευνητική ομάδα στοχεύει να εκτιμήσει την πιθανότητα του γεγονότος “κορώνα” *Pr(Heads)* σε μια σειρά νομισμάτων (π.χ. 10)
  - Ποιος είναι ο στατιστικός πληθυσμός;
  - Ποιες είναι οι πιθανές πηγές μεροληψίας σε ένα δείγμα;
  - Περιγράψτε μια διαδικασία δειγματοληψίας που θα ελαχιστοποιεί την επίδραση τέτοιων μεροληψιών

Μέγεθος δείγματος

- Το **μέγεθος του δείγματος *n*** είναι ο αριθμός των παρατηρήσεων σε ένα δείγμα
- Όσο μεγαλύτερο είναι το μέγεθος *N* ενός πληθυσμού,τόσο λιγότερες πληροφορίες (αναλογικά) περιέχει ένα δείγμα μεγέθους *n*
- Η ικανότητα γενίκευσης των συμπερασμάτων που παίρνουμε από το δείγμα για τον πληθυσμό, εξαρτάται από την επάρκεια της διαδικασίας δειγματοληψίας σε σχέση με τους στόχους της έρευνας

Παράδειγμα

- Είναι αρκετό ένα δείγμα 100 υγειών ανθρώπων (εθελοντών) που ρωτάμε σε δρόμο της Θεσσαλονίκης αν είναι διαθέσιμοι;
- Εξαρτάται από το ερώτημα της έρευνας:
  - μπορεί να είναι επαρκές εάν ο στόχος είναι να αξιολογηθεί η επίδραση της καφεΐνης στον χρόνο αντίδρασης
    - Πρέπει να τους κεράσουμε έναν καφέ!
  - είναι ανεπαρκές εάν ο στόχος είναι να αξιολογηθούν οι παρενέργειες μιας νέας φαρμακευτικής ουσίας
  - είναι ανεπαρκές για να αξιολογήσει το μέσο εισόδημα της ελληνικής οικογένειας

Ανεξάρτητες και εξαρτημένες παρατηρήσεις

- Οι παρατηρήσεις σε ένα δείγμα μπορούν να θεωρηθούν ως **ανεξάρτητες** αν οι πληροφορίες για κάθε παρατήρηση δεν παρέχουν πληροφορίες για άλλες παρατηρήσεις
- Δύο παρατηρήσεις είναι **εξαρτημένες** (ή συσχετισμένες) αν είναι κατά κάποιο τρόπο συνδεδεμένες
  - Οι διατροφικές συνήθειες των μελών της ίδιας οικογένειας
  - Επαναλαμβανόμενα τεστ μνήμης που έγιναν από το ίδιο άτομο
  - Το μήκος του φύλλου μέσα στο ίδιο φυτό
- Σε ποια κατηγορία ανήκει το παράδειγμα της ρίψης του νομίσματος;
  - Και κάτω από ποιες προϋποθέσεις;

Στατιστική μοντελοποίηση

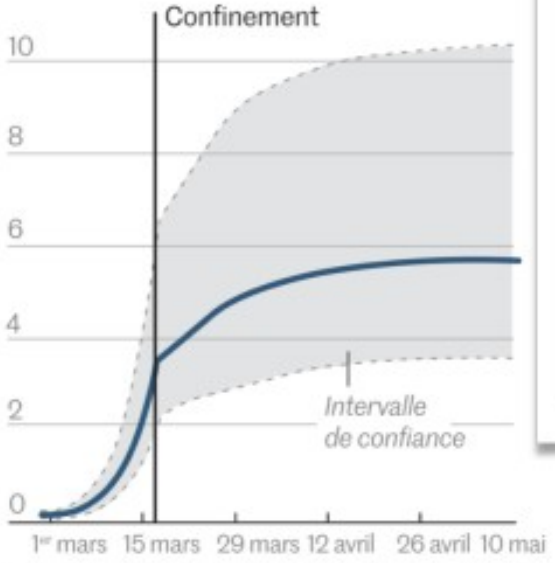
- Τα δείγματα σχετίζονται κατά κάποιο τρόπο με τον πληθυσμό από τον οποίο προέρχονται
- Ο στόχος της στατιστικής μοντελοποίησης είναι να καταλάβουμε τη διαδικασία που δημιούργησε τα παρατηρούμενα δεδομένα και να προβλέψουμε νέες παρατηρήσεις
- Μια ειδική περίπτωση στατιστικής μοντελοποίησης είναι ο **έλεγχος υποθέσεων**. Παραδείγματα υποθέσεων:
  - Οι άνδρες είναι πιο ψηλοί από τις γυναίκες
  - Η πρόσβαση στην τριτοβάθμια εκπαίδευση επηρεάζει θετικά το εισόδημα
  - Η ανάγνωση από χαρτί οδηγεί σε καλύτερη απομνημόνευση από την ανάγνωση από tablet.

Why learning statistics?

5,7 % de la population seraient infectés par le Covid-19 au 11 mai

Part de la population qui **pourrait être infectée** par le Covid-19 à la date du 11 mai en France métropolitaine, en %

Dans la population totale



Le Monde, 21/4/2020

Comments by readers

Mickaël R. 14/05/2020 - 03H29

9,9 % (marge de 6,6 à 15,7 %) des habitants d'Ile-de-France auraient été contaminés au 11 mai et 9,1 % (marge 6,0 à 14,6 %) "
A ce niveau ce ne sont plus des marges, mais des abymes.
La marge comptent pour plus d'un tiers du pourcentage calculé.

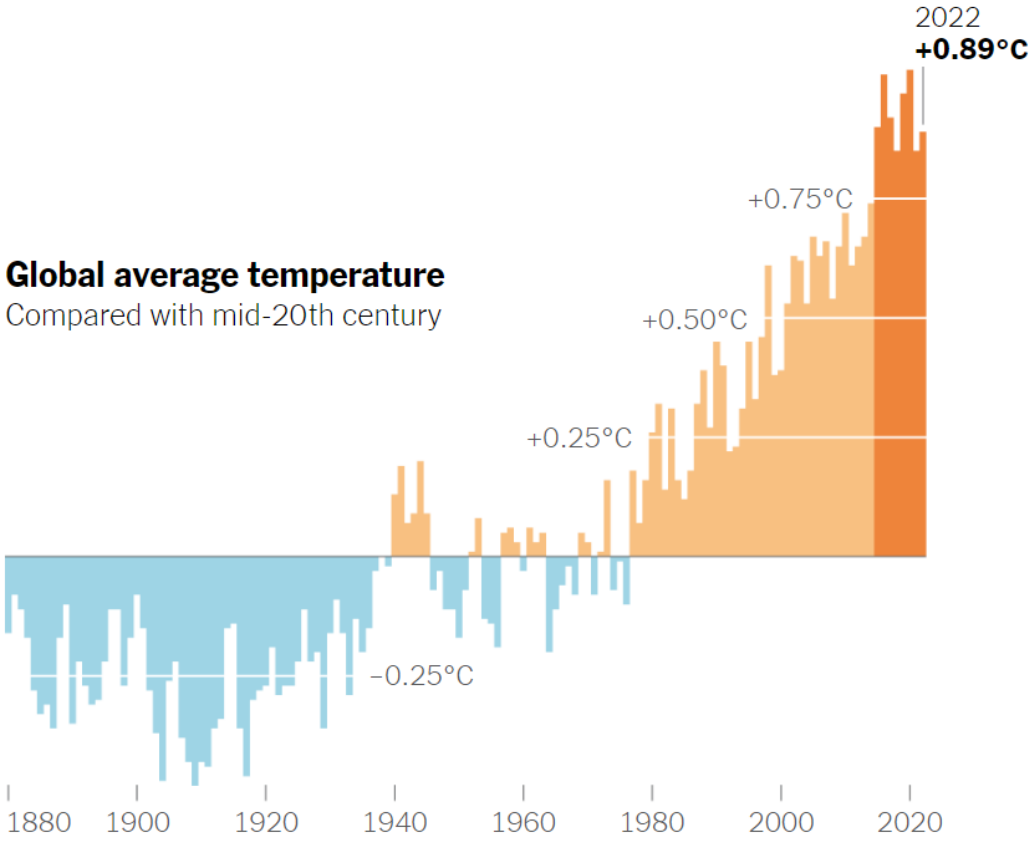
Masquer la réponse | Répondre | Signaler un abus

Eric B 14/05/2020 - 11H40

Au moins ces incertitudes sont évaluées proprement. Avez-vous des données qui permettraient de réduire cette incertitude? J'imagine que non...

Γιατί είναι χρήσιμη η στατιστική;

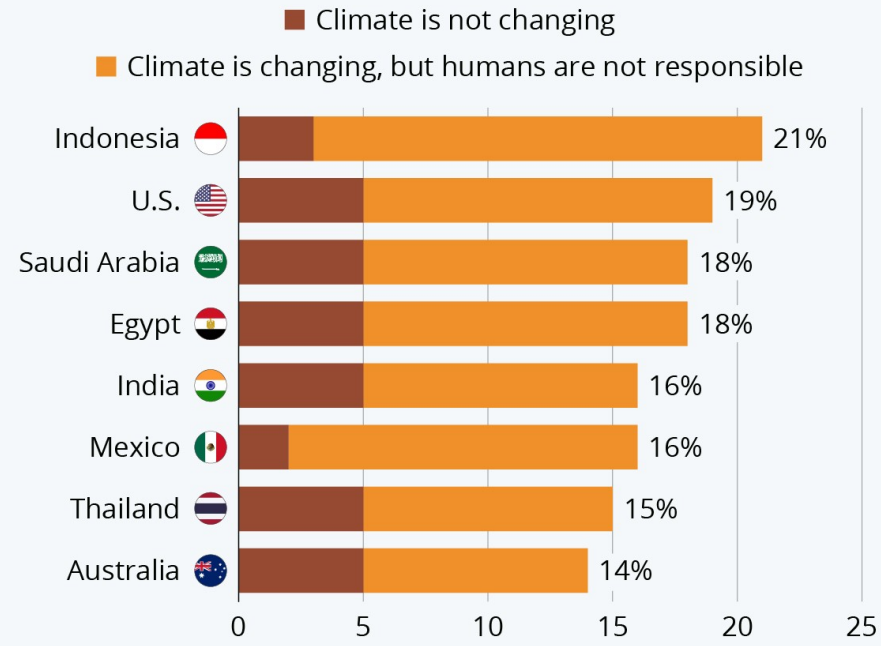
- Πολλοί ισχυρισμοί και πεποιθήσεις αλλά και προκαταλήψεις και στερεότυπα βασίζονται σε άτυπα συμπεράσματα στατιστικής
  - Συχνά βασίζονται σε ανεπαρκή ή μεροληπτική δειγματοληψία
  - Συχνά με βάση ελλιπή ή λανθασμένα μοντέλα
  - Συχνά αποτυγχάνουν να διακρίνουν μεταξύ συσχέτισης και αιτιότητας
- Πολύ συχνά, σημαντικές αποφάσεις που λαμβάνονται στην πολιτική βασίζονται σε αληθείς (ή ψευδείς) στατιστικές αποδείξεις



Source: NASA Goddard Institute for Space Studies

Where Climate Change Deniers Live

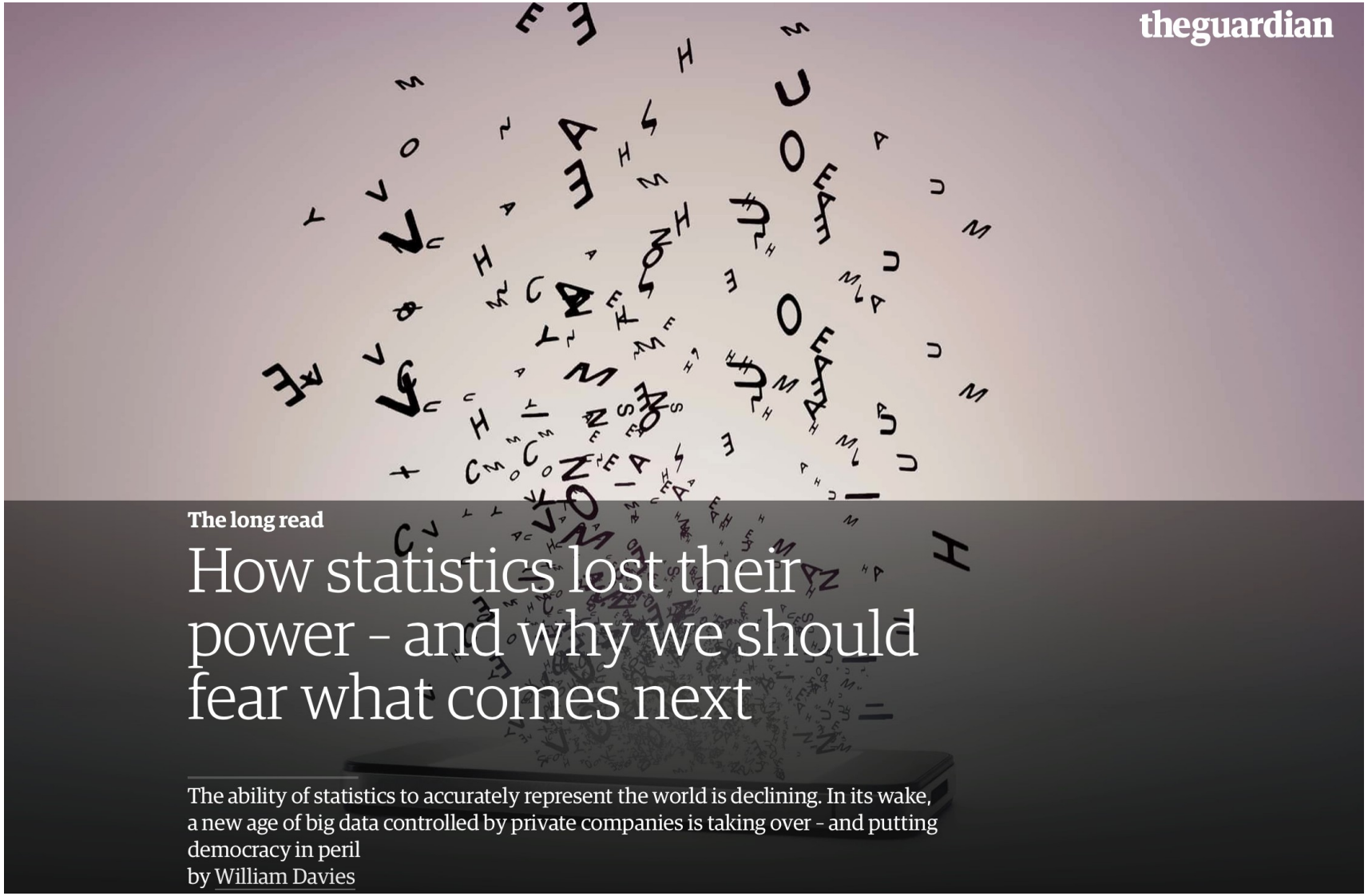
Countries with highest share saying they think climate change is not real or humans are not responsible



26,000 people in 25 countries surveyed July 30-Aug 24, 2020
Source: YouGov

statista

theguardian



The long read

How statistics lost their power – and why we should fear what comes next

The ability of statistics to accurately represent the world is declining. In its wake, a new age of big data controlled by private companies is taking over – and putting democracy in peril

by William Davies

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For A

News & Comment > News > 2018 > September > Article

NATURE | NEWS

Over half of psychology studies fail reproducibility test

Largest replication study to date casts doubt on many published positive results.

Monya Baker

27 August 2015

Rights & Permissions

Don't trust everything you read in the psychology literature. In fact, two thirds of it should probably be distrusted.

In the biggest project of its kind, Brian Nosek, a social psychologist and head of the Center for Open Science in Charlottesville, Virginia, and 269 co-authors repeated work reported in 98 original papers from three psychology journals, to see if they independently came up with the same results.

Brian Nosek's team set out to replicate scores of

Γιατί είναι χρήσιμη η στατιστική;

- Η στατιστική είναι ένα θεμελιώδες εργαλείο έρευνας για πολλούς επιστημονικούς κλάδους
  - ...αλλά ακόμη και στην έρευνα, πολύ συχνά γίνεται κακή χρήση

Open access, freely available online

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on *p*-values. Research findings are defined here as any relationship reaching

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance, the post-study probability that it is true is the positive predictive value, PPV.

Η εργασία με τις περισσότερες αναφορές παγκοσμίως

Σκοπός του μαθήματος

- Η κατανόηση βασικών εννοιών της στατιστικής και η αντίληψη της στατιστικής προσέγγισης για τα βιολογικά δεδομένα
- Κατανόηση στατιστικών μεθόδων πρακτικού ενδιαφέροντος στις βιολογικές επιστήμες
- Απόκτηση εμπειρίας πάνω σε πραγματικά βιολογικά δεδομένα
- Εξοικείωση με τη R

Γιατί να μάθω R;

- Το R είναι μια γλώσσα προγραμματισμού και λογισμικό ανοιχτού κώδικα, αλλά και περιβάλλον για στατιστικούς υπολογισμούς και γραφήματα
- Είναι διαθέσιμο για τις περισσότερες πλατφόρμες υπολογιστών (Windows, Mac OS, Linux)
- Υπάρχει ένα ευρύ φάσμα στατιστικών πακέτων γραμμένων σε R. Καλύπτουν σχεδόν όλα όσα μπορεί να χρειαστείτε για τις στατιστικές σας αναλύσεις.
- Το R χρησιμοποιείται ευρέως από την ερευνητική κοινότητα

Μειονεκτήματα της R

- Απαιτητική εκμάθηση και η χρήση CLI για πολλούς, σε σύγκριση με εμπορικό στατιστικό λογισμικό με γραφικό περιβάλλον, όπως SPSS, Statistica και JMP
- Το R υποστηρίζει αρκετούς σχετικούς τύπους δεδομένων (λίστες, διανύσματα, μήτρες, πλαίσια) και είναι εύκολο να μπερδευτεί κανείς
- Η δημιουργία ενός καλού γραφήματος μπορεί να είναι αρκετά επίπονη, π.χ. υπολογίζοντας παραμέτρους
- Όμως για όλα τα παραπάνω, με την πρόοδο της χρήσης του R, οι δυσκολίες ξεπερνιούνται

[Home]

Download

CRAN

R Project

About R
Logo
Contributors
What's New?
Reporting Bugs
Development Site
Conferences
Search

R Foundation

Foundation
Board
Members
Donors
Donate

Help With R

Getting Help

Getting Help with R

Helping Yourself

Before asking others for help, it's generally a good idea for you to try to help yourself. R includes extensive facilities for accessing documentation and searching for help. There are also specialized search engines for accessing information about R on the internet, and general internet search engines can also prove useful (see below).

R Help: `help()` and `?`

The `help()` function and `?` help operator in R provide access to the documentation pages for R functions, data sets, and other objects, both for packages in the standard R distribution and for contributed packages. To access documentation for the standard `lm` (linear model) function, for example, enter the command `help(lm)` or `help("lm")`, or `?lm` or `?"lm"` (i.e., the quotes are optional).

To access help for a function in a package that's *not* currently loaded, specify in addition the name of the package: For example, to obtain documentation for the `rlm()` (robust linear model) function in the **MASS** package, `help(rlm, package="MASS")`.

Standard names in R consist of upper- and lower-case letters, numerals (`0-9`), underscores (`_`), and periods (`.`), and must begin with a letter or a period. To obtain help for an object with a *non-standard* name (such as the help operator `?`), the name must be quoted: for example, `help('?')` or `?"?"`.

You may also use the `help()` function to access information about a package in your library — for example, `help(package="MASS")` — which displays an index of available help pages for the package along with some other information.

Help pages for functions usually include a section with executable examples illustrating how the functions work. You can execute these examples in the current R session via the `example()` command; e.g., `example(lm)`

CRAN
Mirrors
What's new?
Search
CRAN Team

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Task Views
Other

Documentation
Manuals
FAQs
Contributed

Donations
Donate

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux (Debian, Fedora/Redhat, Ubuntu)
- Download R for macOS
- Download R for Windows

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2023-10-31, Eye Holes) R-4.3.2.tar.gz, read what's new in the latest version.
- Sources of R alpha and beta releases (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are available here. Please read about new features and bug fixes before filing corresponding feature requests or bug reports.
- Source code of older versions of R is available here.
- Contributed extension packages

Questions About R

- If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

Supporting CRAN

- CRAN operations, most importantly hosting, checking, distributing, and archiving of R add-on packages for various platforms, crucially rely on technical, emotional, and financial support by the R community.

  Please consider making financial contributions to the R Foundation for Statistical Computing.

What are R and CRAN?

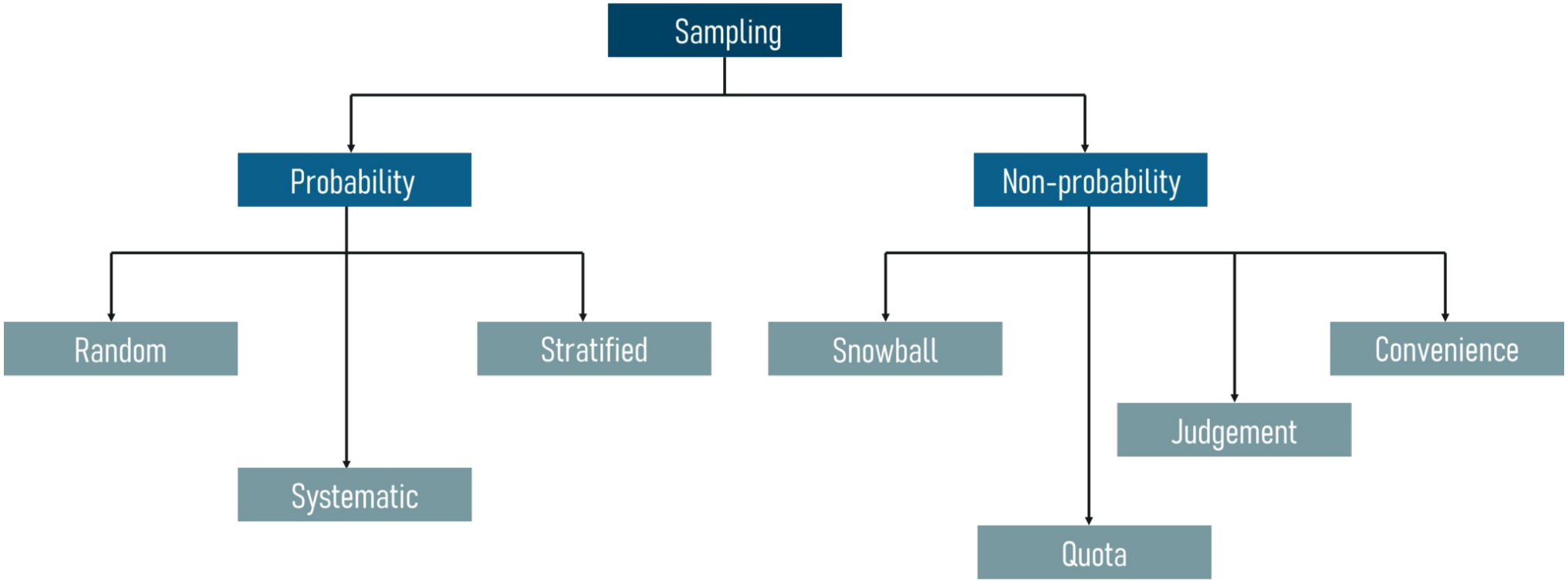
R is 'GNU S', a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques: linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering, etc. Please consult the R project homepage for further information.

CRAN is a network of ftp and web servers around the world that store identical, up-to-date, versions of code and documentation for R. Please use the CRAN mirror nearest to you to minimize network load.

Κατεβάστε και εγκαταστήστε

R
R Studio

Τυχαία & μη τυχαία δειγματοληψία



Απλή τυχαία δειγματοληψία

- Κάθε μέλος του πληθυσμού έχει την ίδια πιθανότητα να επιλεγεί στο δείγμα
- Η “τυχαιότητα” πρέπει να διασφαλίζεται αυστηρά, αλλιώς έχουμε “snowball sampling”
- *Παράδειγμα: έρευνα κοινής γνώμης σε 1000 ανθρώπους στην Αλεξανδρούπολη*

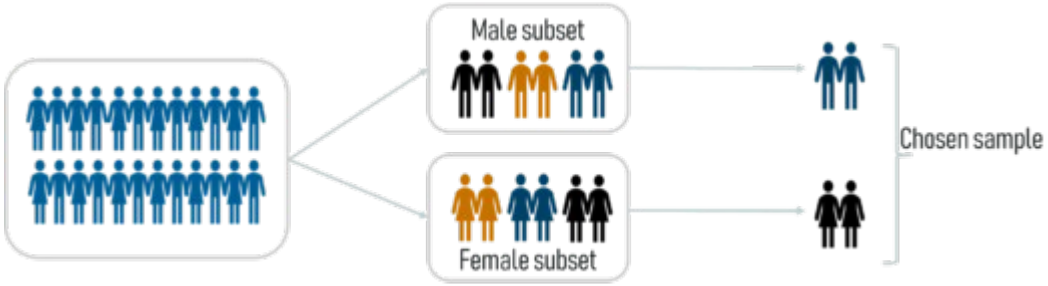
Η τυχαιότητα απαιτεί μεγαλύτερα δείγματα για να είναι αντιπροσωπευτική η δειγματοληψία, γι’ αυτό οδηγούμαστε σε στρωματωμένη, συστηματική, κλπ.

Συστηματική δειγματοληψία

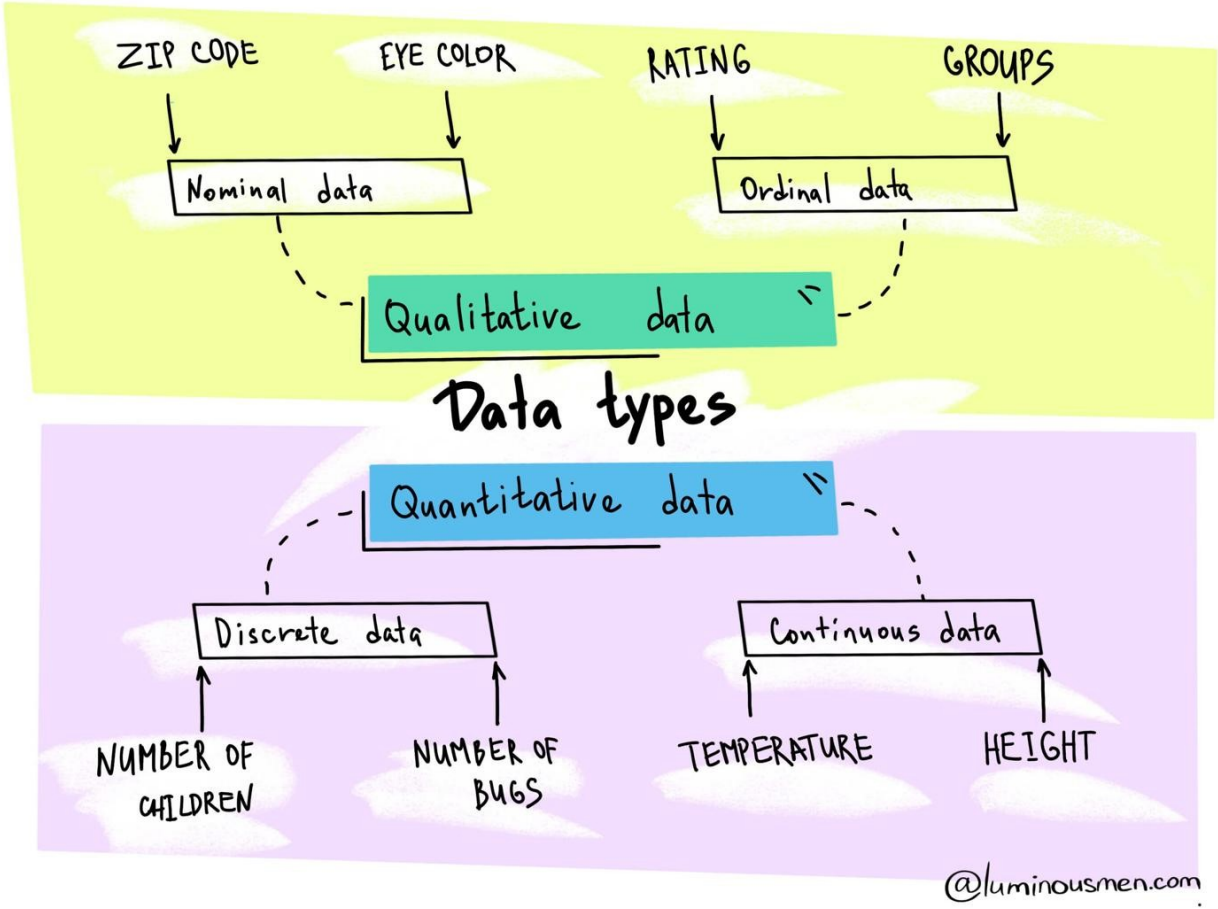
- Κάθε **νιοστό** (n^{th}) μέλος του πληθυσμού επιλέγεται στο δείγμα
- Πρέπει να υπάρχει γνώση παραμέτρων του πληθυσμού, π.χ. μέγεθος, πυκνότητα
- *Παράδειγμα: συλλογή δειγμάτων από 100 δέντρα σε ένα δάσος για αναλύσεις DNA*

Στρωματωμένη δειγματοληψία

- Ένα **στρώμα** (stratum) είναι ένα υποσύνολο του πληθυσμού που χαρακτηρίζεται από ένα κοινό στοιχείο (π.χ. φύλλο), που είναι σημαντικό για την έρευνά μας
- Γίνεται τυχαία δειγματοληψία μέσα σε κάθε στρώμα, έτσι ώστε το συνολικό δείγμα να είναι αντιπροσωπευτικό ως προς την εκπροσώπηση των στρωμάτων
- *Παράδειγμα: το βάρος και το ύψος μιας ομάδας ανθρώπων*

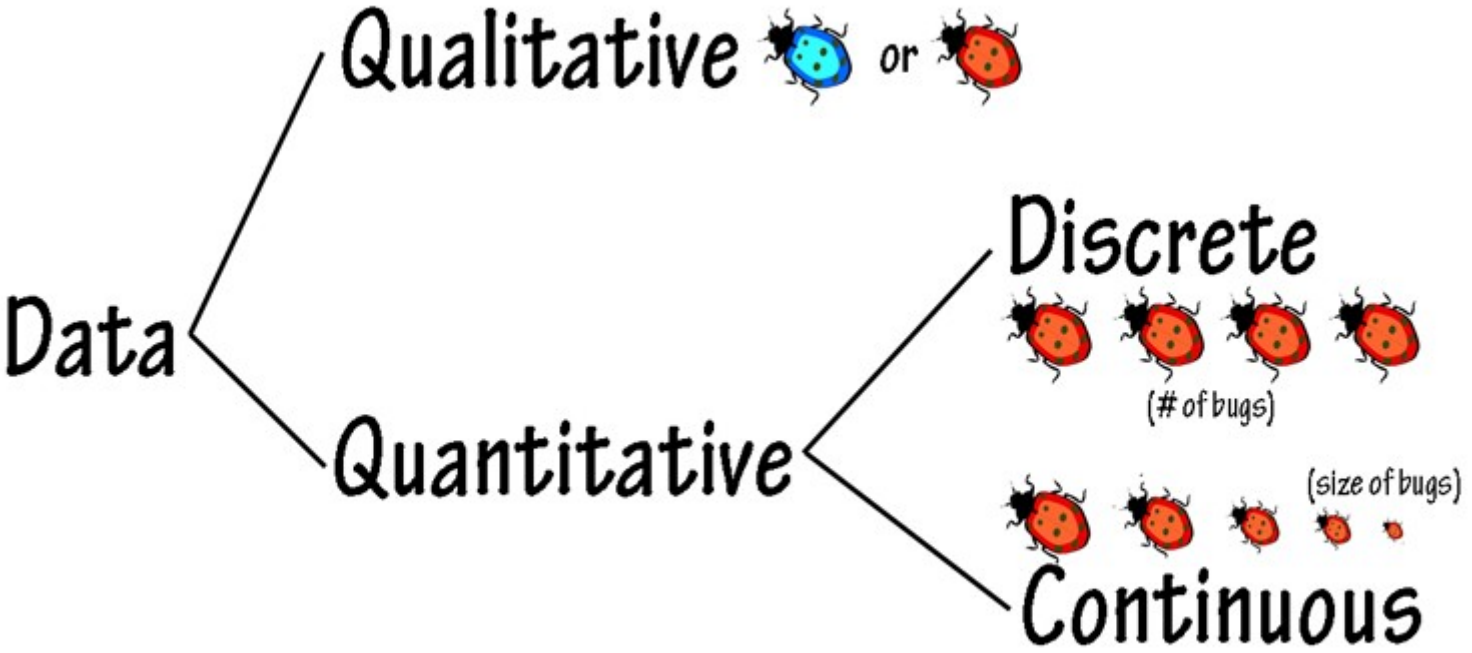


Τύποι δεδομένων



Σε ποια κατηγορία ανήκουν τα **βιολογικά** δεδομένα;

Τύποι δεδομένων

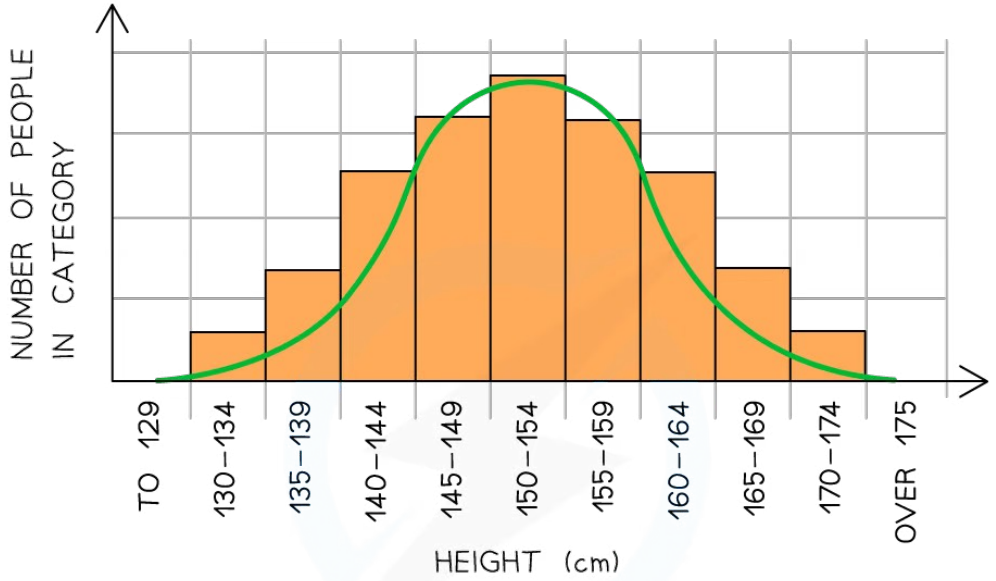


Ποσοτικά δεδομένα

- **Διακριτά** ποσοτικά δεδομένα
- Προκύπτουν από απαρίθμηση
- Παραδείγματα:
  - Διωνυμικά δεδομένα που λαμβάνουν μόνο δύο δυνατές τιμές (0,1)
  - Συχνότητες ή τιμές καταμέτρησης

Συνεχή ποσοτικά δεδομένα

- Προκύπτουν από μετρήσεις
- Κάθε μέτρηση μπαίνει σε ένα διάστημα τιμών
  - Και όχι σε ένα συγκεκριμένο σημείο



FEATURES OF CONTINUOUS VARIATION:

- NO DISTINCT CLASSES OR CATEGORIES EXIST
- CHARACTERISTICS CAN BE MEASURED AND FALL WITHIN A RANGE BETWEEN TWO EXTREMES

Κλίμακες μέτρησης

- **Ονομαστικές**
  - Μπορούν να αναπαρασταθούν και με αριθμούς αλλά δεν υπάρχει ιεράρχηση
- Χρώμα, επάγγελμα

- **Τακτικές**
  - Διατηρούν την πληροφορία σχετικά με το σχετικό (όχι απόλυτο) μέγεθος αυτού που μετριέται
- Ηλικιακή ομάδα (παιδί, έφηβοι, ενήλικες)

Categorical

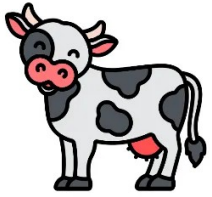
Nominal

Pen

Pencil

Paper

Dog



Cow

Cat

Ordinal

Excellent

Good

Bad

Fantastic

Okay

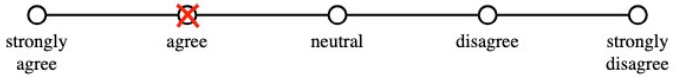
Don’t like

B1. Which of the following video-communication tools do you currently use or have you used in the past? *

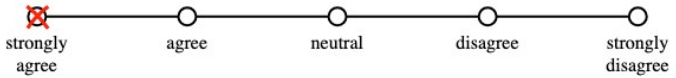
Une seule réponse possible par ligne.

	never	rarely	occasionally	frequently
Skype	◯	◯	◯	◯
Zoom	◯	◯	◯	◯
Jitsi	◯	◯	◯	◯
Google Meet	◯	◯	◯	◯
Blackboard	◯	◯	◯	◯
Discord	◯	◯	◯	◯
Teams	◯	◯	◯	◯

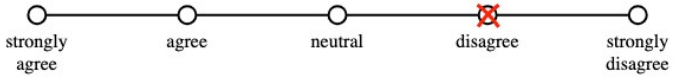
1. The website has a user friendly interface.



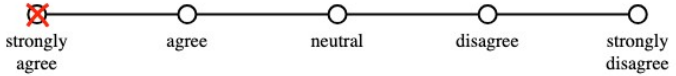
2. The website is easy to navigate.



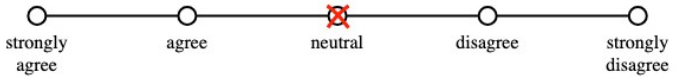
3. The website's pages generally have good images.



4. The website allows users to upload pictures easily.



5. The website has a pleasing color scheme.



Τι τύπου μετρήσεις έχουμε από τις απαντήσεις της ερώτησης αυτής;

Συχνότητες γονοτύπων και αλληλομόρφων

- Διακριτά δεδομένα είναι οι συχνότητες γονοτύπων και αλληλομόρφων σε ένα γονίδιο σε έναν πληθυσμό
- Τι τύπου δεδομένα είναι τα γενετικά δεδομένα;

```
> genotype = c("AA","Aa","aa","Aa","AA","Aa","AA","aa","Aa","Aa","aa","Aa","aa","Aa","AA","Aa","aa","Aa","Aa")
> genotype
 [1] "AA" "Aa" "aa" "Aa" "AA" "Aa" "AA" "aa" "Aa" "Aa" "aa" "Aa" "aa" "Aa" "AA"
[16] "Aa" "aa" "Aa" "Aa"
> table(genotype)
genotype
aa Aa AA
 5 10  4
> _
```

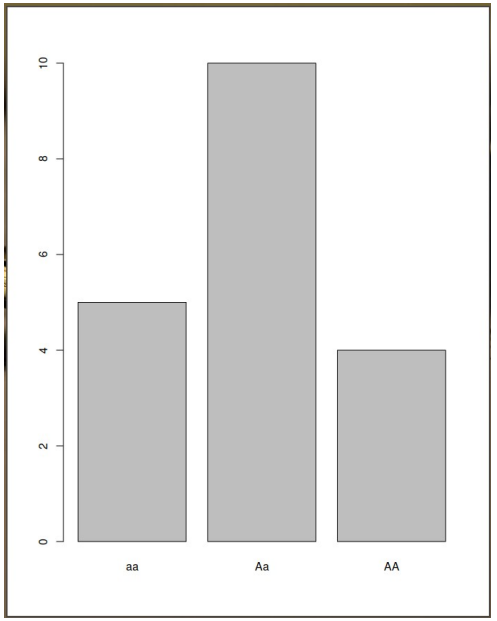
Συχνότητες γονοτύπων και αλληλομόρφων

- Στο R, οι κατηγορικές μεταβλητές λέγονται factors

```
> genotypeF = factor(genotype)
> levels(genotypeF)
[1] "aa" "Aa" "AA"
> table(genotypeF)
genotypeF
aa Aa AA
 5 10  4
> _
```

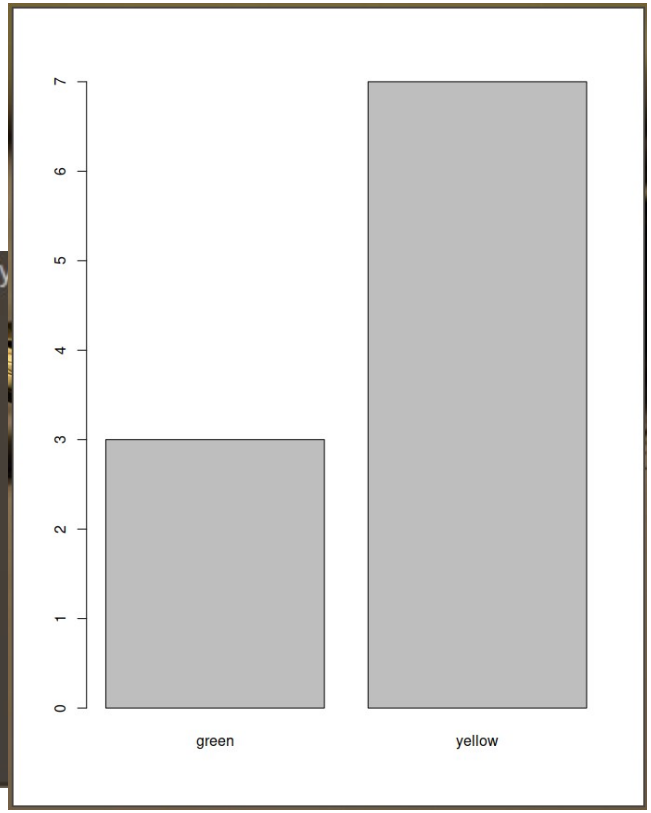
- Γράφημα συχνοτήτων γονοτύπων (ιστόγραμμα)

```
> plot(genotypeF)
```



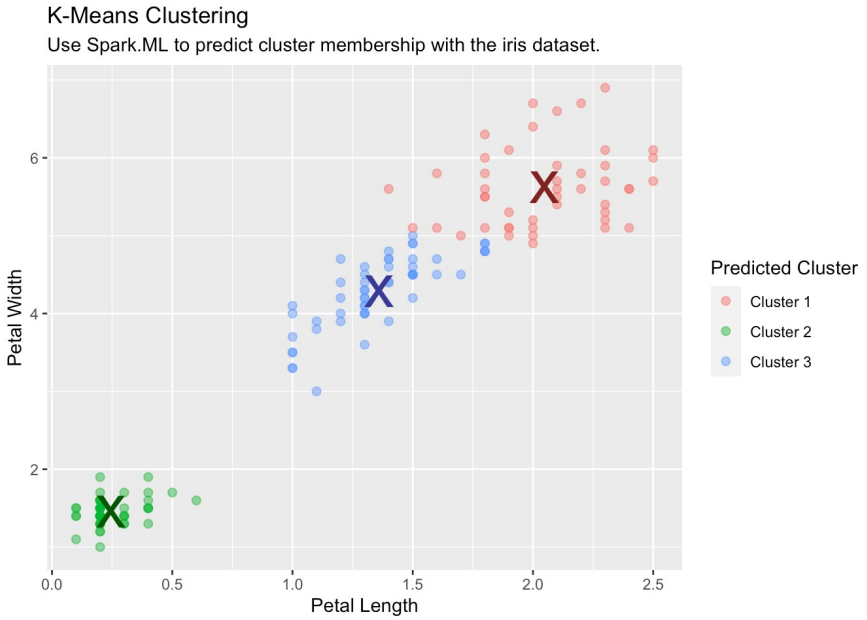
Συχνότητες φαινοτύπων

```
> phenotype = c("green","yellow","yellow","yellow","yellow","yellow","yellow","y
llow","green","green")
> phenotype
 [1] "green"  "yellow" "yellow" "yellow" "yellow" "yellow" "yellow" "yellow"
 [9] "green"  "green"
> phenotypeF = factor(phenotype)
> levels(phenotypeF)
[1] "green"  "yellow"
> table(phenotypeF)
phenotypeF
 green yellow
     3      7
> plot(phenotypeF)
> _
```



Κατάταξη σε κατηγορίες

- Ιεραρχικά (ranking)
  - π.χ. κλινική εικόνα σε μια συγκεκριμένη ασθένεια
- Κατάταξη σε συστάδες/ομάδες (clusters/groups)
  - π.χ. υπαγωγή σε γεωγραφικές ομάδες
  - π.χ. υπαγωγή σε προβλεπόμενες βιολογικές ομάδες



Κλίμακες μέτρησης

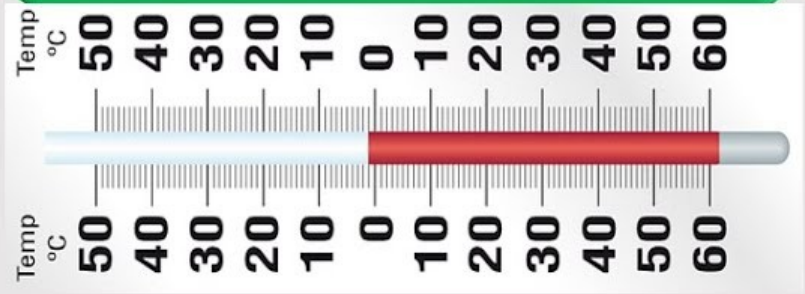
- **Διάστημα**
  - Διατηρεί συνεχείς, γραμμικές σχέσεις μεταξύ αυτού που μετριέται
  - Οι αποστάσεις μεταξύ των επόμενων σημείων της κλίμακας είναι ίσες
  - Η παρουσία του μηδενός είναι ασαφής
  - Θερμοκρασία σε βαθμούς Κελσίου
- Θερμοκρασία σε βαθμούς Κελσίου

- **Αναλογία**
  - Διάστημα που έχει «αληθινό» μηδέν
- Θερμοκρασία σε βαθμούς Kelvin, βάρος σε κιλά

	Interval	Ratio
Well-defined **intervals**	YES	YES
The **zero point** indicates the **absence of a quantity**	NO	YES
Difference measured in terms of **distance**	YES	YES
Difference measured in terms of **ratios**	NO	YES

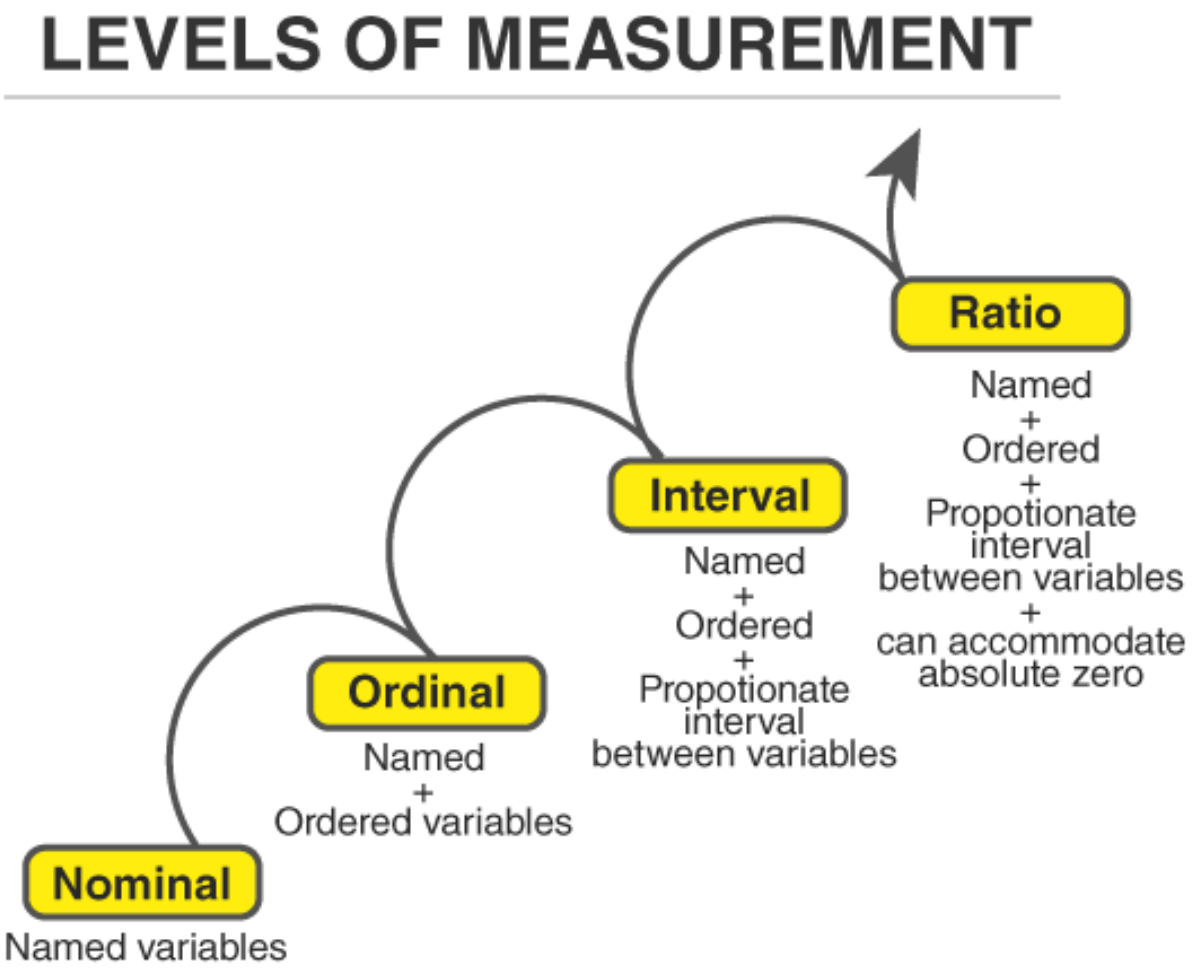
Difference Between Interval and Ratio Scale

Interval Scale



Ratio Scale

Σύνοψη


LEVELS OF MEASUREMENT
Nominal
Named variables
Ordinal
Named
+
Ordered variables
Interval
Named
+
Ordered
+
Propotionate interval between variables
Ratio
Named
+
Ordered
+
Propotionate interval between variables
+
can accommodate absolute zero

Κλίμακες μέτρησης

- Οι κλίμακες περιορίζουν τις μαθηματικές πράξεις που επιτρέπονται σε δεδομένα συγκεκριμένου τύπου:
  - Τα ονομαστικά δεδομένα περιορίζονται σε πράξεις όπως η καταμέτρηση
  - Τα τακτικά δεδομένα περιορίζονται σε πράξεις όπως η κατάταξη
  - Τα δεδομένα διαστήματος επιτρέπουν επίσης την πρόσθεση και την αφαίρεση (αλλά όχι τον πολλαπλασιασμό ή τη διαίρεση)
  - Οι κλίμακες αναλογίας επιτρέπουν το πλήρες εύρος της αριθμητικής λειτουργίας και επιτρέπουν αναλογίες μεταξύ αριθμών (10/5 = 2 σημαίνει ότι το 10 είναι δύο φορές μεγαλύτερο από το πέντε)

Thank you