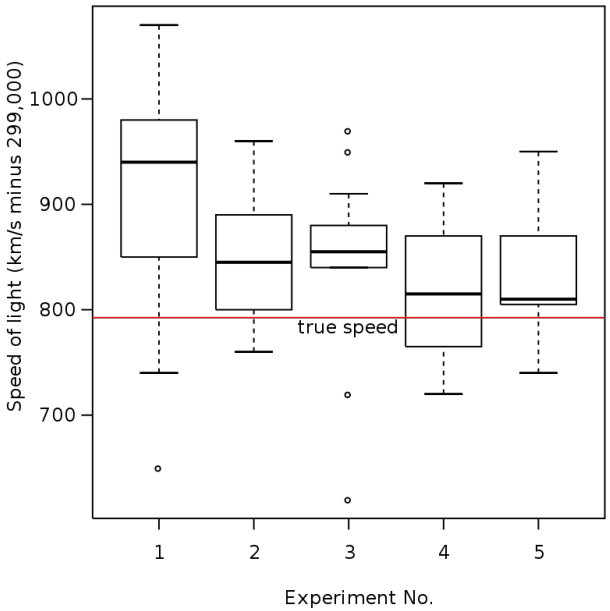


Βιοστατιστική

Περιγραφική Στατιστική

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

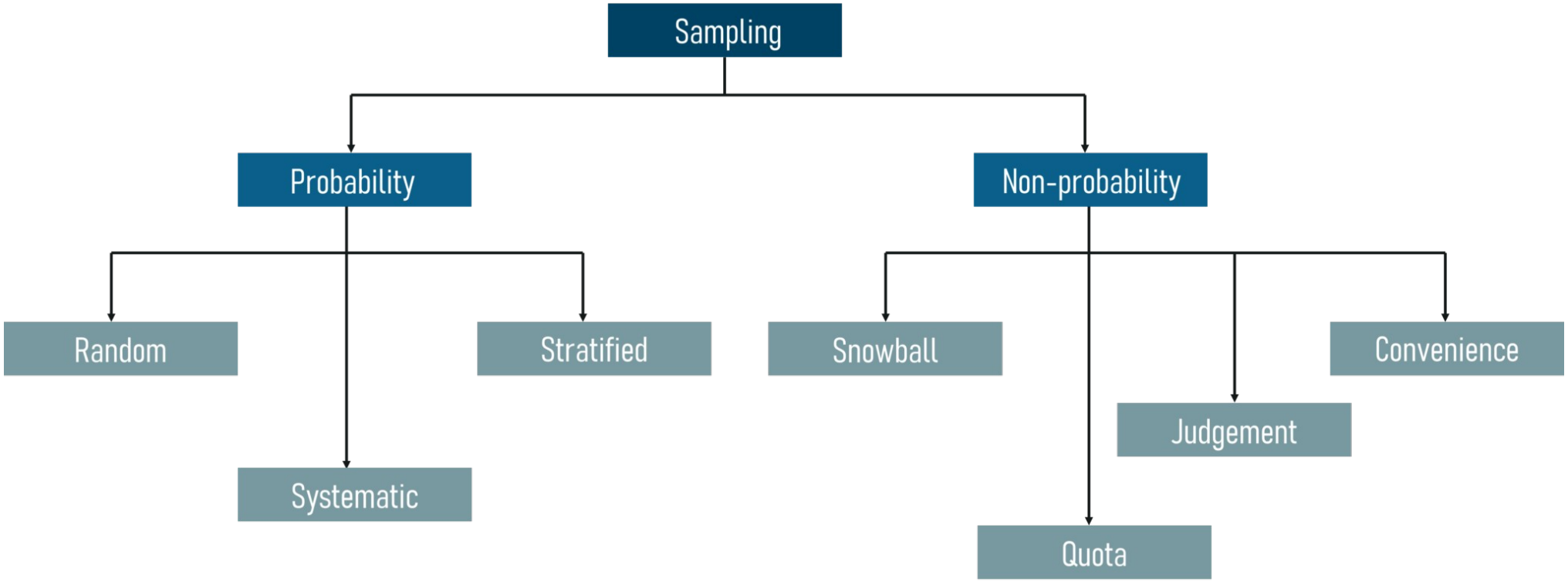
Πηγές υλικού

- 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)

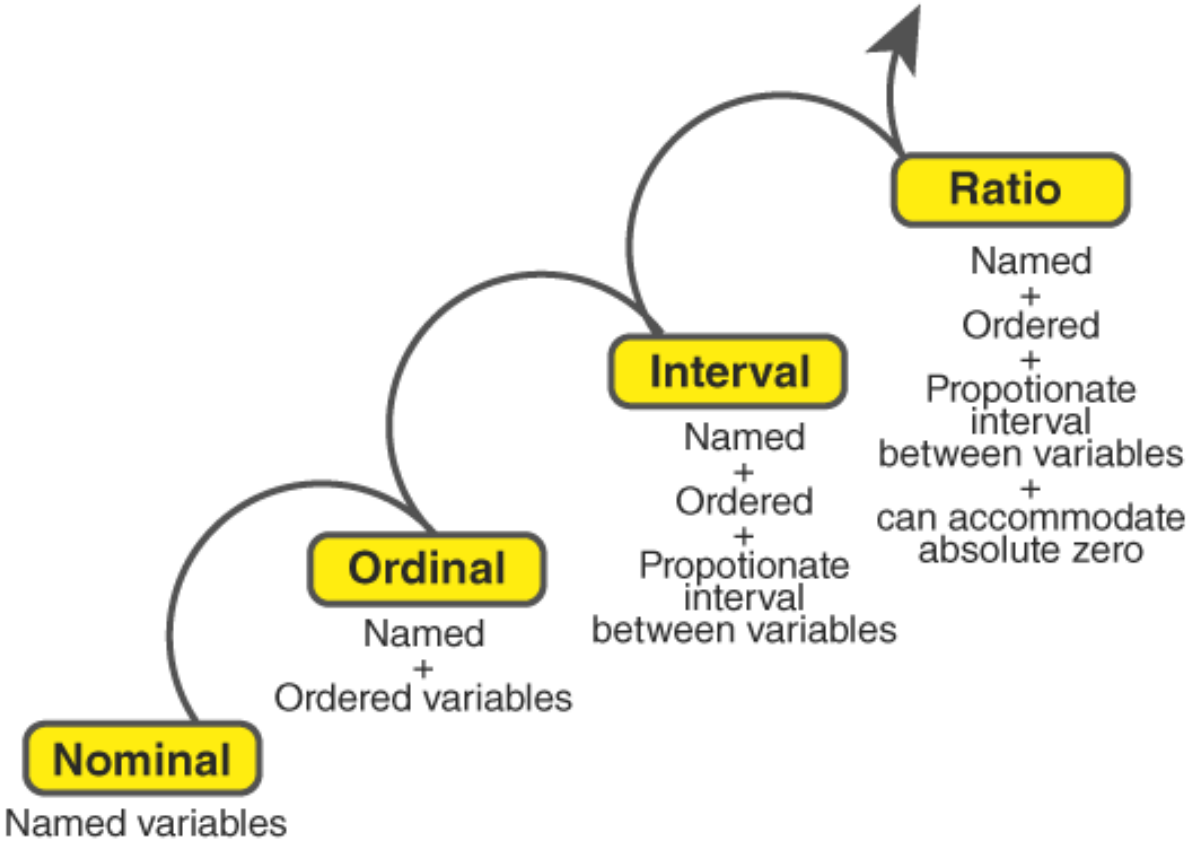
Η στατιστική

- 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

Τυχαία & μη τυχαία δειγματοληψία



LEVELS OF MEASUREMENT



Περιγραφική στατιστική

- Descriptive (or summary) statistics
  - *ελάχιστο (min), μέγιστο (max), μέσος (mean), διάμεσος (median), τυπική απόκλιση (standard deviation) και άλλα...*
- Ένας τρόπος για να συνοψίσουμε και να παρουσιάσουμε πληροφορίες σχετικά με ένα σύνολο δεδομένων
  - “Πρώτη αίσθηση” για ένα σύνολο δεδομένων
- Επιβεβαιώνουμε κάποια σαφή μοτίβα (πρότυπα) στα δεδομένα, αν υπάρχουν
- Εντοπισμός τυχόν παρατυπιών και προβλημάτων κατά τη δειγματοληψία
- Μας οδηγεί στην επιλογή του κατάλληλου στατιστικού μοντέλου

Μέτρα κεντρικής τάσης

- Περιγραφή ενός συνόλου δεδομένων με μία τιμή
  - Η πιο “τυπική” ή η πιο “κοινή” ή η πιο “μέση” τιμή
- Μέτρα κεντρικής τάσης
  - Τύπος (mode): η πιο κοινή τιμή
  - Διάμεσος (median): η κεντρική τιμή
  - Μέσος (mean, average)
- Ο τύπος, η διάμεσος ή ο μέσος όρος ενός δείγματος, τις περισσότερες φορές διαφέρουν από τα αντίστοιχα μέτρα του πληθυσμού

Παράμετροι και στατιστικά

- Μια **παράμετρος** είναι ιδιότητα του πληθυσμού
- Ένα **στατιστικό** είναι ιδιότητα του δείγματος
  - Παρέχει μια εκτίμηση μιας πληθυσμιακής παραμέτρου
  - Όσο το μέγεθος *n* του δείγματος πλησιάζει το μέγεθος *N* του πληθυσμού, τα στατιστικά τείνουν να μοιάζουν με τις παραμέτρους
- Συνήθως χρησιμοποιούμε ελληνικά γράμματα για την παράμετρο ενός πληθυσμού και ένα λατινικό γράμμα για το στατιστικό ενός δείγματος
  - π.χ, το ***μ*** προσδιορίζει τον μέσο του πληθυσμού, ενώ το ***M*** δηλώνει τη μέση τιμή του δείγματος

Παράμετροι και στατιστικά

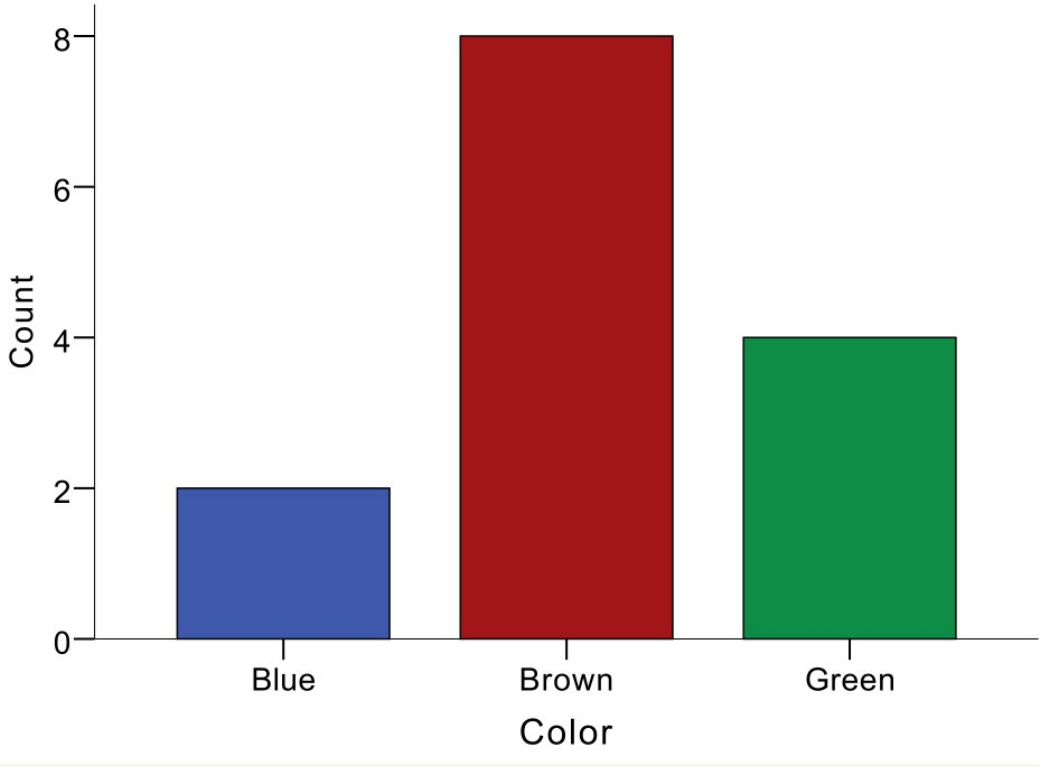
- Ένας άλλος συμβολισμός είναι η χρήση του “καπέλου” για το δείγμα:
  - Το μ προσδιορίζει τον μέσο του πληθυσμού
  - Το $\hat{\mu}$ προσδιορίζει τον μέσο του δείγματος

Τύπος

- Η πιο κοινή τιμή
  - Ο τύπος αυτού του συνόλου τιμών είναι 10:
    - 16 **10** 12 **10** 9 14 13
- Ο τύπος είναι η καλύτερη τιμή για να μαντέψουμε
  - Καλύτερα για διακριτά και όχι για συνεχή δεδομένα, ειδικά για κατηγορικά δεδομένα, όπου δεν υπάρχει ιεραρχική σχέση

Τύπος

- Καταγράψαμε σε 14 ανθρώπους το χρώμα των ματιών τους
- Ποιος είναι ο τύπος του δείγματος;
- Ποια είναι η καλύτερη “υπόθεση” για το χρώμα ενός τυχαίου ανθρώπου του δείγματος;
- Εφαρμογή στη γενετική...



Διάμεσος

- Η κεντρική τιμή σε ένα σύνολο αριθμών
  - Όταν οι αριθμοί είναι τοποθετημένοι με τη σειρά, η διάμεσος είναι η μεσαία τιμή: 12 20 24 **34** 35 80 83
  - Εάν το μέγεθος του δείγματος είναι ζυγός αριθμός, λαμβάνεται το μέσο σημείο μεταξύ των δύο κεντρικών τιμών ως διάμεσος: 5 6 **8** **9** 12 15, οπότε η διάμεσος είναι (8+9)/2 = 8,5
- Η διάμεσος δεν είναι ευαίσθητη σε ακραίες τιμές
  - Πλεονέκτημα: εξαλείφει την επίδραση των ακραίων τιμών
  - Μειονέκτημα: αγνοεί τις μη κεντρικές τιμές

Αριθμητικός μέσος

- Είναι το πιο ευρέως χρησιμοποιούμενο μέτρο κεντρικής τάσης
  - γνωστό και ως μέσος ή μέσος όρος
- Ο μέσος όρος του ακόλουθου συνόλου δεδομένων:
  10 16 10 12 9 14 είναι (16+10+12+10+9+14+13) / 7 = 12

- Για ένα σύνολο αριθμών xi με i=1,2,....n ισχύει:

$$M = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Ο μέσος αξιοποιεί όλες τις τιμές του δείγματος

Ερωτήσεις

- Ποιο είναι το καλύτερο μέτρο κεντρικής τάσης για καθένα από τα ακόλουθα;

    1) Βάρος 50 τυχαίων φοιτητών
    2) Εισόδημα οικογενειών στην Ελλάδα
    3) Έξοδα για στέγαση 100 φοιτητών, σε τρεις κατηγορίες:
        (a) χαμηλότερα από 200 ευρώ
        (b) μεταξύ 200 και 500 ευρώ
        (c) υψηλότερα από 500 ευρώ

Μέτρα διασποράς

- Συγκρίνετε τα δύο σύνολα δεδομένων:
  - D1: 12 13 13 14 15 14
  - D2: 5 9 12 15 20 20
    - Έχουν σχεδόν ίδιους μέσους και διάμεσους αλλά είναι πολύ διαφορετικά
- Οι αριθμοί στο D2 είναι πιο απλωμένοι
  - Έχουν μεγαλύτερη **διασπορά**
- Μέτρα διασποράς:
  - εύρος, τεταρτημόρια, διακύμανση, τυπική απόκλιση...

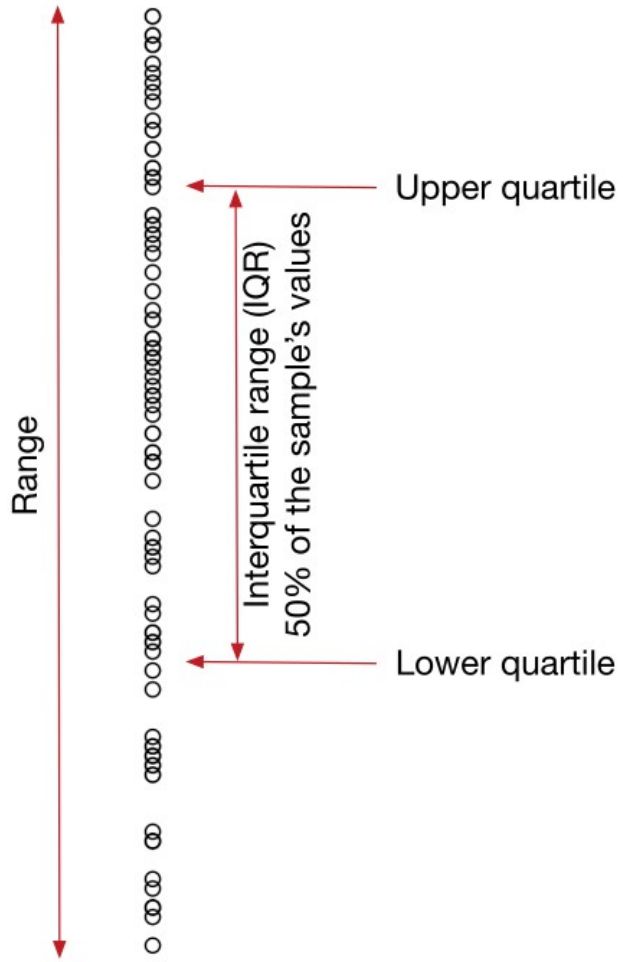
Εύρος

- Είναι η διαφορά ανάμεσα στη μέγιστη και την ελάχιστη τιμή:
  - D1: 12 13 13 14 15 14
  - D2: 5 9 12 15 20 20
- D1: 15 - 12 = 3
- D2: 20 - 5 = 15

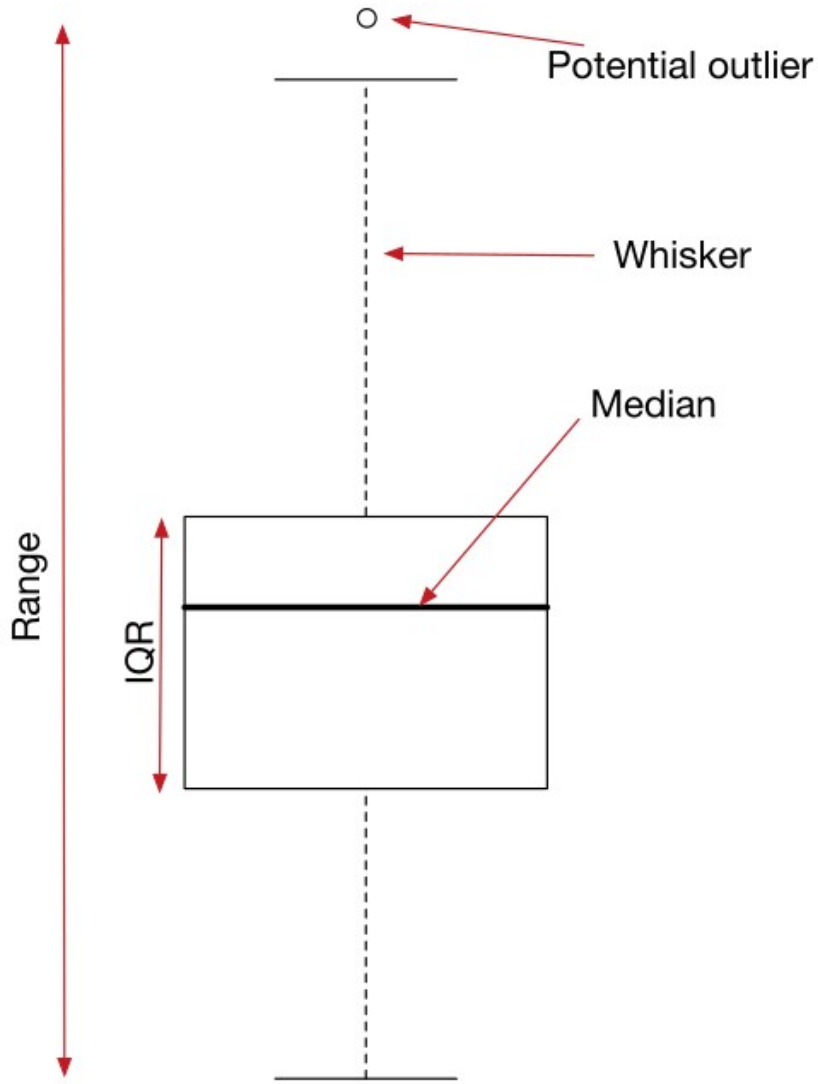
- Το εύρος είναι πολύ ευαίσθητο στις ακραίες τιμές

Τεταρτημόρια

- Τα τεταρτημόρια είναι οι **τρεις τιμές** που χωρίζουν ένα σύνολο από *n* ταξινομημένους ιεραρχικά αριθμούς σε τέσσερα ίσα υποσύνολα
  - Το πρώτο (κατώτερο) τεταρτημόριο διαχωρίζει το 25% των χαμηλότερων αριθμών
  - Το δεύτερο (μεσαίο) τεταρτημόριο είναι ο διάμεσος
  - Το τρίτο (ανώτερο) τεταρτημόριο διαχωρίζει ο 25% των υψηλότερων αριθμών



Boxplot



Διακύμανση (variance)

Variance

$$Var = \frac{\sum_{i=1}^{n} (x_i - M)^2}{n}$$

sum of squares

sample size

Τυπική απόκλιση (standard deviation)

Standard deviation (scaled to use the same units as the original data)

$$SD = \sqrt{Var} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - M)^2}{n}}$$

Είναι η τετραγωνική ρίζα της διακύμανσης

Η πιο διαδεδομένη μέτρηση διασποράς των τιμών

Παράδειγμα

Consider the following dataset that gives the weight of six 15-month old babies (in kilograms):

8 10 10 12 9 11

$$M = \frac{(8+10+10+12+9+11)}{6} = 10$$

$$Var = \frac{(8-10)^2+(10-10)^2+(10-10)^2+(12-10)^2+(9-10)^2+(11-10)^2}{6} = 1.667$$

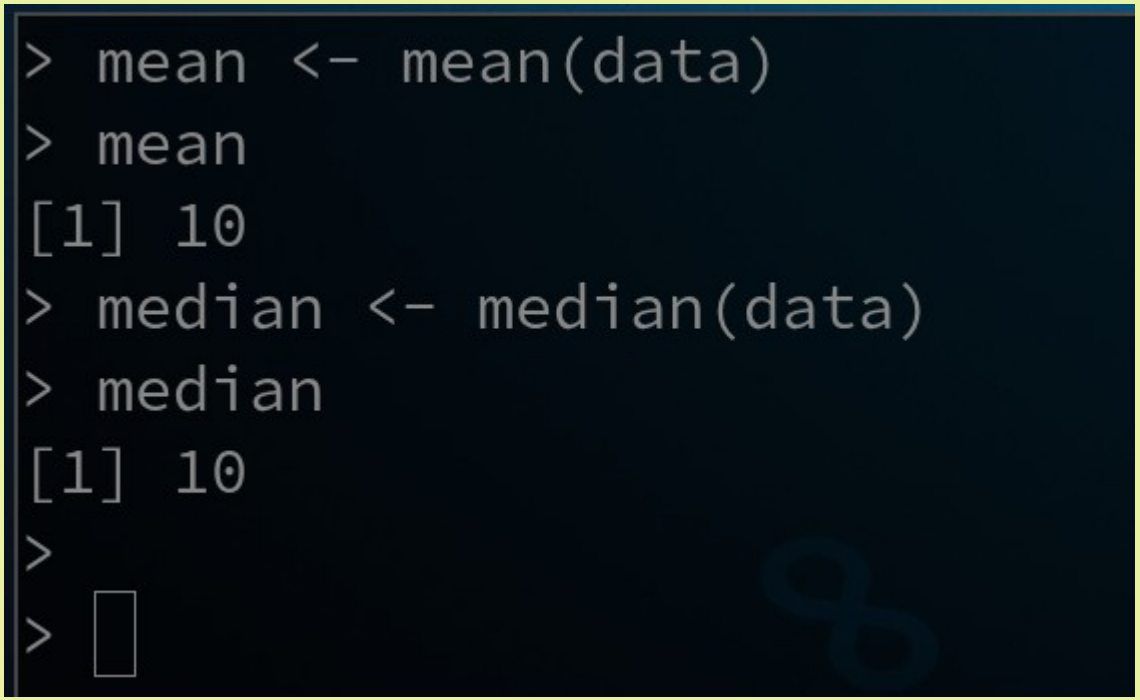
$$SD = \sqrt{1.667} = 1.29$$

Ας λύσουμε την άσκηση με R

- Δημιουργούμε ένα σετ δεδομένων με τις τιμές της άσκησης (βάρος βρεφών 15 μηνών)

```
> data <- c(8,10,10,12,9,11)
> data
[1]  8 10 10 12  9 11
>
```

Μέσος και διάμεσος



Διακύμανση και τυπική απόκλιση

```
> variance <- var(data)
> variance
[1] 2
> sd <- sd(data)
> sd
[1] 1.414214
```

Γιατί υπάρχει διαφορά στην τιμή που υπολογίσαμε με το χέρι και στην τιμή από το R;

Υπολειμματικές τιμές (residuals)

- Οι ακατέργαστες αποκλίσεις της κάθε τιμής από τον μέσο όρο λέγονται **υπολειμματικές τιμές**

$$x_i - M$$

For the following dataset (M = 10)

8 10 10 12 9 11

The residuals are as follows:

-2 0 0 2 -1 1

Το άθροισμα όλων των υπολειμματικών τιμών ενός δείγματος είναι πάντα **0**

Εκτιμητές πληθυσμιακών παραμέτρων

- Πώς εκτιμούμε τον μέσο όρο ενός πληθυσμού, τη διακύμανση ή την τυπική του απόκλιση από ένα δείγμα;
- Ερωτήματα:
  - Είναι ο μέσος όρος ενός δείγματος ένας καλός εκτιμητής του πληθυσμιακού μέσου;
  - Είναι η τυπική απόκλιση ενός δείγματος ένας καλός εκτιμητής της τυπικής απόκλισης του πληθυσμού;

Αποτελεσματικοί και αμερόληπτοι εκτιμητές

- Ένα καλό στατιστικό θα πρέπει να είναι ένας αποτελεσματικός και αμερόληπτος εκτιμητής της αντίστοιχης παραμέτρου του πληθυσμού
- Ένα **αποτελεσματικό** στατιστικό έχει μικρότερο σφάλμα
  - τείνει να είναι κοντά στην παράμετρο του πληθυσμού
  - παρουσιάζει μικρότερες διακυμάνσεις από δείγμα σε δείγμα
- Ένα **αμερόληπτο** στατιστικό δεν έχει bias (είναι αντικειμενικό)
  - Μακροπρόθεσμα (σταθερά), δεν υπερεκτιμά ούτε υποτιμά την πραγματική παράμετρο του πληθυσμού

Αμερόληπτοι και μεροληπτικοί εκτιμητές

- Τα στατιστικά των δειγμάτων κεντρικής τάσης όπως οι μέσοι και οι διάμεσοι είναι **αμερόληπτοι** εκτιμητές
  - Έτσι, χρησιμοποιούμε συχνά για το $\hat{\mu}$ για να προσεγγίσουμε τη μέση τιμή του πληθυσμού μ
- Όμως τα στατιστικά για τη διασπορά είναι **μεροληπτικά**
  - Τείνουν να υποτιμούν την πραγματική παράμετρο του πληθυσμού
  - Ένα μικρό δείγμα είναι απίθανο να συλλάβει τα άκρα ενός πληθυσμού

Αμερόληπτοι εκτιμητές διασποράς

The population variance is usually represented as σ^2 and its unbiased estimator is:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu})^2}{n-1}$$

$n-1$ degrees of freedom

The unbiased estimator of the population standard deviation σ is:

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

$n-1$ degrees of freedom

Βαθμοί ελευθερίας

- Βαθμοί ελευθερίας
  - ο αριθμός των παραμέτρων στον υπολογισμό ενός στταιστικού που μπορούν να ποικίλουν ελεύθερα
- Γιατί ***n-1***;
  - Όταν είναι γνωστός ο μέσος, απαιτούνται μόνο n-1 ανεξάρτητες παρατηρήσεις για να υπολογίσουμε τη διακύμανση (ή την τυπική απόκλιση)
  - Αυτό γίνεται αν θεωρήσουμε τις υπολειμματικές τιμές ενός δείγματος

$$x_n = \hat{\mu} - (x_1 + x_2 + \ldots + x_{n-1})$$

Περιγραφικά και συμπερασματικά στατιστικά

- Οι αμερόληπτες εκτιμήσεις της διακύμανσης ενός πληθυσμού (ή της τυπικής του απόκλισης) είναι γνωστά ως **συμπερασματική** (inferential) διακύμανση ή **συμπερασματική** τυπική απόκλιση
  - Λέγεται και “επαγωγική”
- Τα περιγραφικά στατιστικά απλά περιγράφουν το δείγμα
- Με τα συμπερασματικά στατιστικά, προσπαθούμε να συμπεράνουμε τις πληθυσμιακές παραμέτρους από ένα δείγμα

Άθροισμα τετραγώνων

- Γιατί στα στατιστικά της διασποράς χρησιμοποιούμε το άθροισμα τετραγώνων των υπολειμματικών τιμών...

$$\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

- και όχι το άθροισμα των απόλυτων τιμών τους;

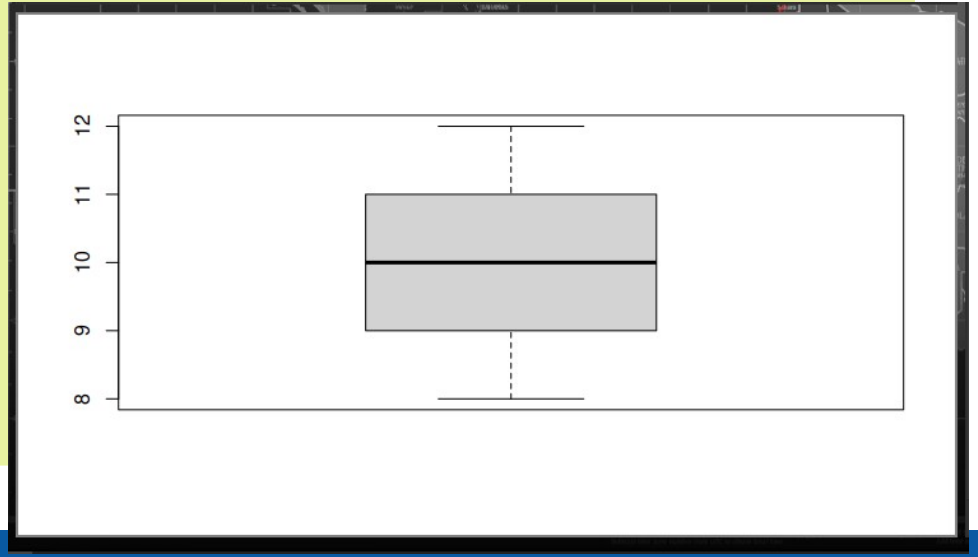
$$\sum_{i=1}^{n}|x_i - \hat{\mu}|$$

Άθροισμα τετραγώνων

- Οι υπολογισμοί με απόλυτες τιμές είναι επίπονοι αλλά δεν είναι αυτός ο κύριος λόγος...
- Το μέτρο κεντρικής τάσης που μειώνει το άθροισμα των απόλυτων τιμών των υπολειμματικών τιμών είναι η διάμεσος και όχι ο μέσος
  - Καθώς ο μέσος χρησιμοποιείται πιο συχνά στις αναλύσεις δεδομένων, οδηγούμαστε στο άθροισμα των τετραγώνων
  - Στις περιπτώσεις αναλύσεων με διάμεσους, μπορούμε να χρησιμοποιούμε το άθροισμα των απόλυτων τιμών

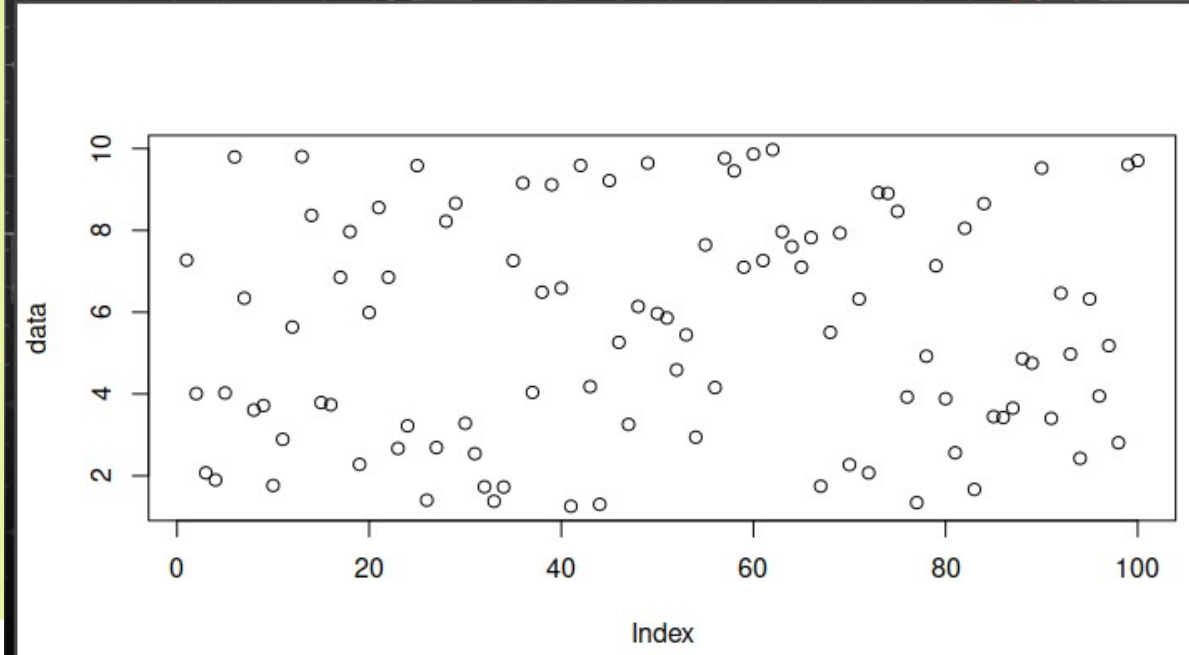
Summary

```
> summary(data)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   8.00    9.25   10.00   10.00   10.75   12.00
> boxplot(data)
```



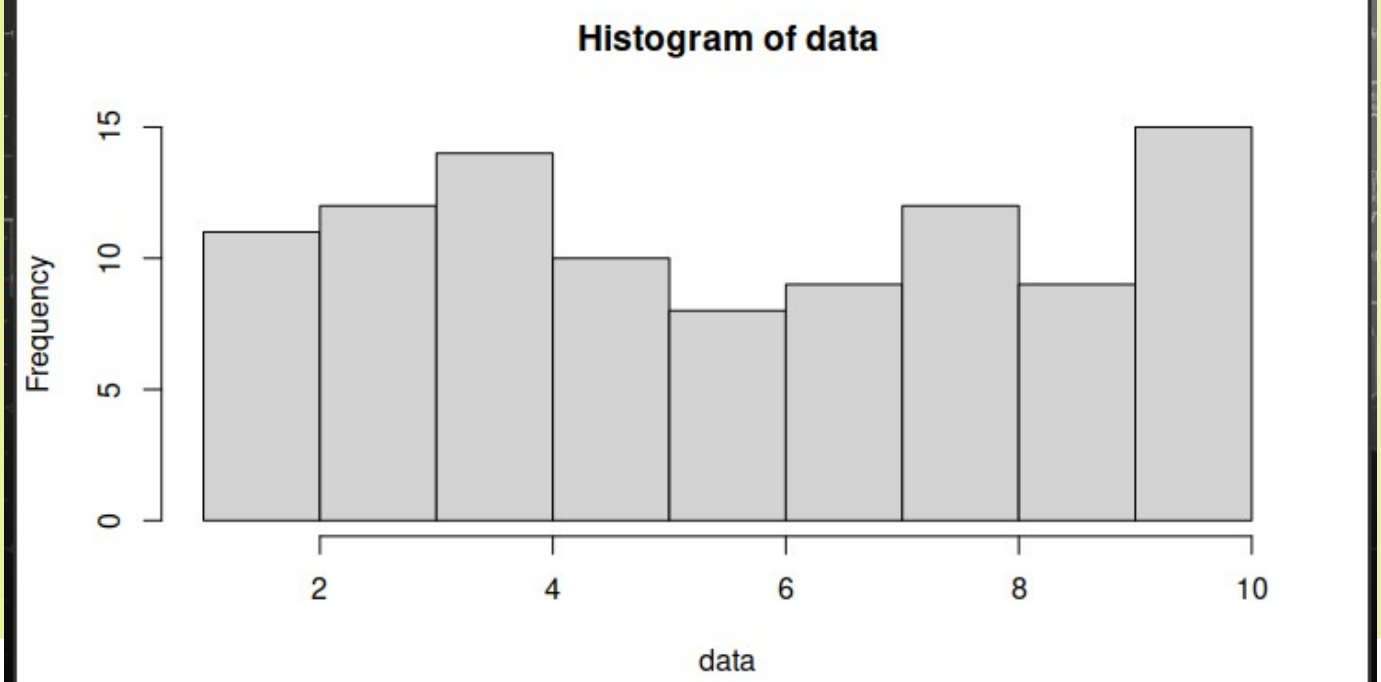
Νέα ποσοτικά δεδομένα

```
> data = runif(100,1,10)
> plot(data)
>
```



Νέα ποσοτικά δεδομένα

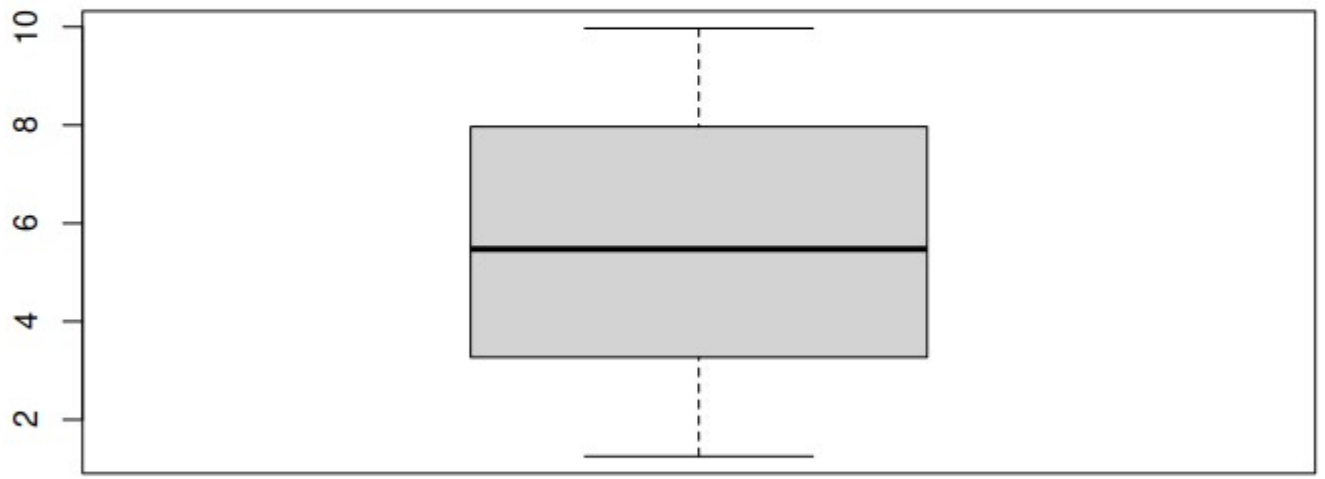
```
> hist(data)
>
```



Περιγραφική στατιστική

```
> mean <- mean(data)
mean
[1] 5.560556
> median <- median(data)
median
[1] 5.47571
> variance <- var(data)
variance
[1] 7.441719
> sd <- sd(data)
sd
[1] 2.727951
> summary(data)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.254   3.276   5.476   5.561   7.965   9.971
```

```
> boxplot(data)
>
```



Πληθυσμός και δείγμα

- Ορίζουμε πληθυσμό αριθμών από το 1 ως το 1000 και καλούμε ένα τυχαίο δείγμα 20 ατόμων

```
> population <- 1:1000
> sample <- sample(population, 20)
> sample
 [1] 892 295 984 565 463 875 953  56 487 598  65 368  72
 623 315 608 869 235 721
[20] 280
>
```

Ιδιότητες του δείγματος

- Μέγεθος και εύρος δείγματος

```
> length(sample)
[1] 20
> range(sample)
[1]  56 984
>
```

```
> IQR(sample)
[1] 466.75
```

Περιγραφικά στατιστικά του δείγματος

- Μέτρα κεντρικής τάσης

```
> mean(population)
[1] 500.5
> median(population)
[1] 500.5
> mean(sample)
[1] 516.2
> median(sample)
[1] 526
>
```

Περιγραφικά στατιστικά του δείγματος

- Μέτρα διασποράς (αμερόληπτα)

```
> var(sample)
[1] 91447.96
> sd(sample)
[1] 302.4036
>
```

- Μέτρα διασποράς (μεροληπτικά)

```
> var_biased <- sum((sample - mean(sample))^2)/length(sa
mple)
> var_biased
[1] 86875.56
> sd_biased <- sqrt(var_biased)
> sd_biased
[1] 294.7466
>
```

Υπολειμματικές τιμές

- Αποκλίσεις από τον μέσο όρο (άθροισμα)

```
> sum(population - mean(population))
[1] 0
> sum(sample - mean(sample))
[1] -9.094947e-13
```

- Αποκλίσεις από τον μέσο όρο (άθροισμα τετραγώνων)

```
> sum((sample - mean(sample))^2)
[1] 1737511
```

- Αποκλίσεις από τον μέσο όρο (άθροισμα απόλυτων τιμών)

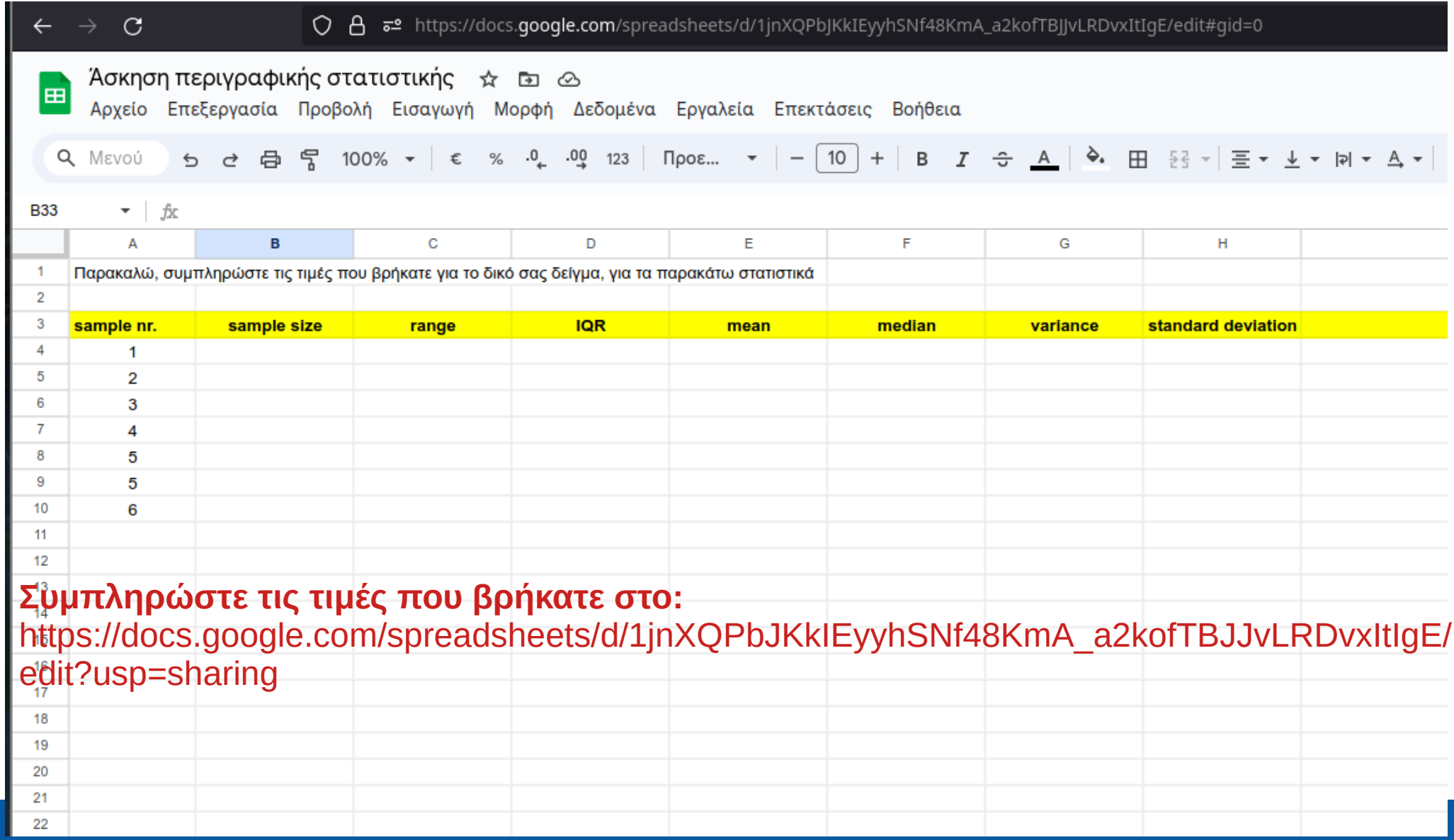
```
> sum(abs(sample - mean(sample)))
[1] 5052
```

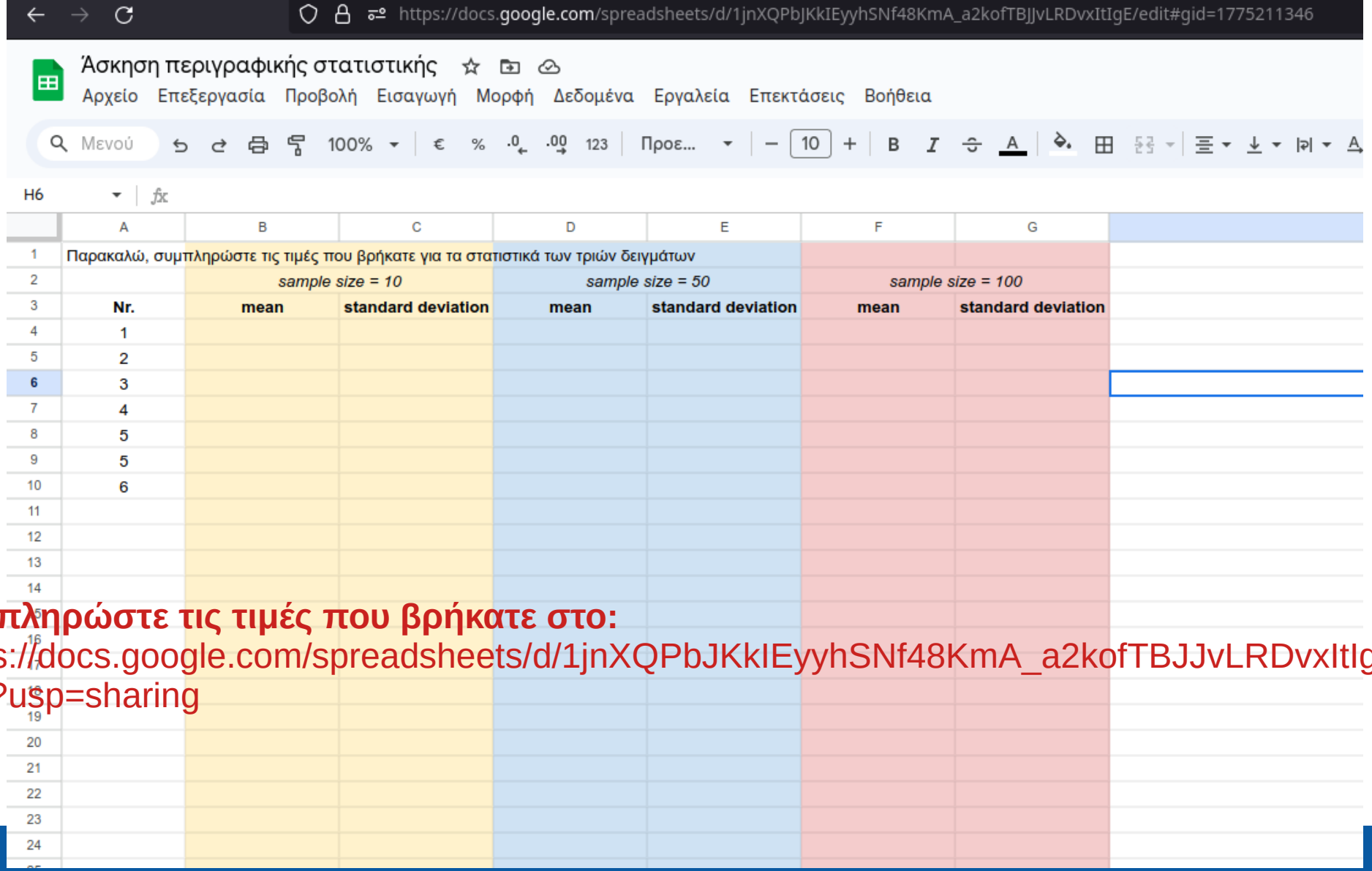
Homework

- **Άσκηση 1**
  - Δημιουργήστε έναν πληθυσμό αριθμών από το 1 ως το 2000 και ένα τυχαίο δείγμα μεγέθους 40
  - Υπολογίστε τα στατιστικά στο δείγμα:
    - εύρος
    - IQR
    - μέσο
    - διάμεσο
    - διακύμανση
    - τυπική απόκλιση

Homework

- **Άσκηση 2**
  - Για τον ίδιο πληθυσμό αριθμών από το 1 ως το 2000:
    - Δημιουργήστε τρία δείγματα μεγέθους 10, 50 και 100
    - Για κάθε ένα δείγμα υπολογίστε τον μέσο και τη διακύμανση


https://docs.google.com/spreadsheets/d/1jnXQPbJKkIEyyhSNf48KmA_a2kofTBJJvLRDvxItIgE/edit#gid=0
Άσκηση περιγραφικής στατιστικής
Αρχείο Επεξεργασία Προβολή Εισαγωγή Μορφή Δεδομένα Εργαλεία Επεκτάσεις Βοήθεια
Μενού
100%
Προε...
10
B33
Παρακαλώ, συμπληρώστε τις τιμές που βρήκατε για το δικό σας δείγμα, για τα παρακάτω στατιστικά
sample nr.
sample size
range
IQR
mean
median
variance
standard deviation
1
2
3
4
5
5
6
Συμπληρώστε τις τιμές που βρήκατε στο:
https://docs.google.com/spreadsheets/d/1jnXQPbJKkIEyyhSNf48KmA_a2kofTBJJvLRDvxItIgE/edit?usp=sharing



Συμπληρώστε τις τιμές που βρήκατε στο:
https://docs.google.com/spreadsheets/d/1jnXQPbJKkIEyyhSNf48KmA_a2kofTBJJvLRDvxItIgE/edit?usp=sharing

Thank you