

# Εισαγωγή στη Στατιστική Μοντελοποίηση

# Τι είναι Μοντέλο

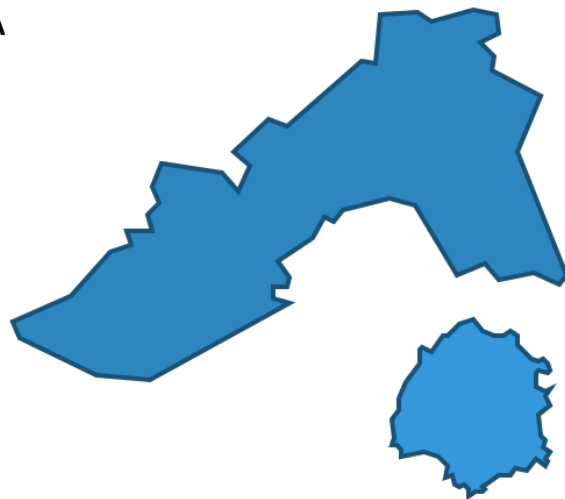
Μοντέλο = απλοποιημένη αναπαράσταση της πραγματικότητας.

Χρησιμεύει για:

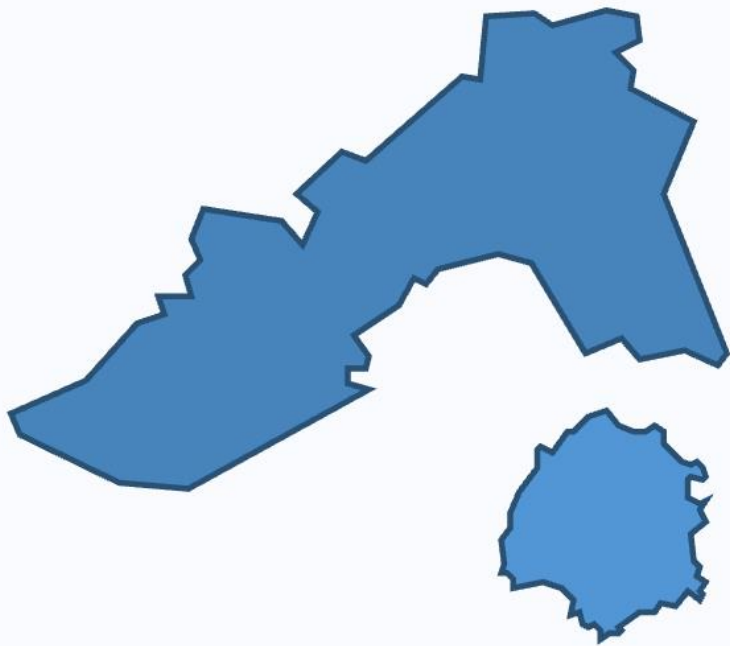
- Εξήγηση
- Πρόβλεψη
- Βελτίωση συστημάτων

ΔΕΔΟΜΕΝΑ = ΜΟΝΤΕΛΟ + ΣΦΑΛΜΑ

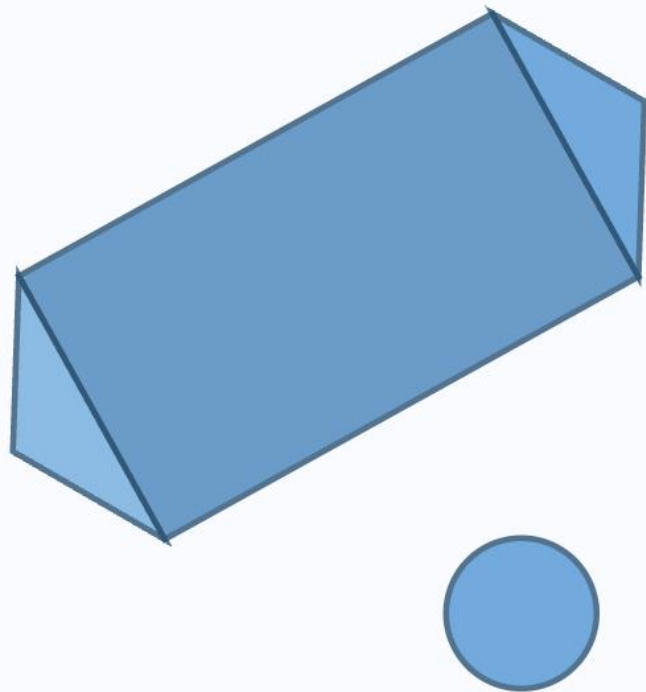
Παράδειγμα: Εκτίμηση εμβαδού Νομού Καβάλας  
με γεωμετρικά σχήματα



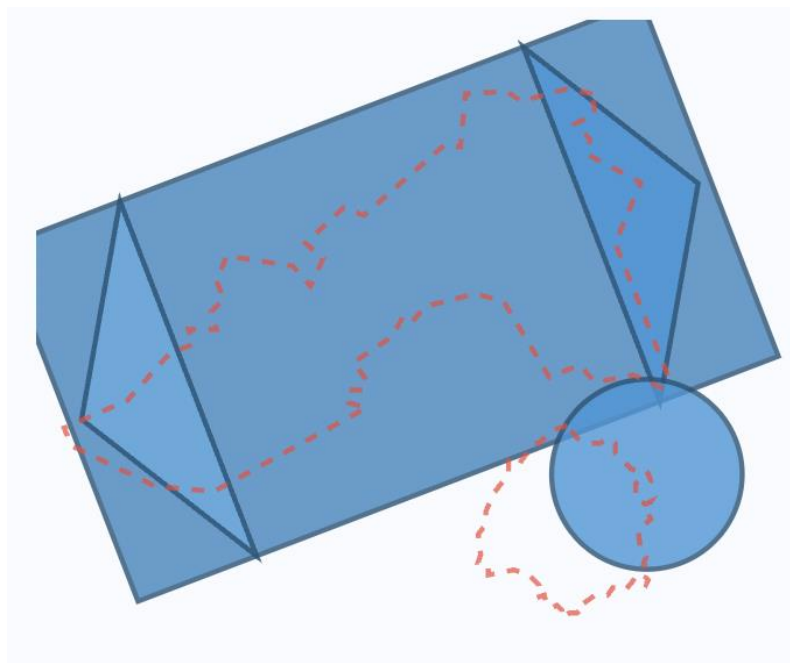
## Αρχικός Χάρτης



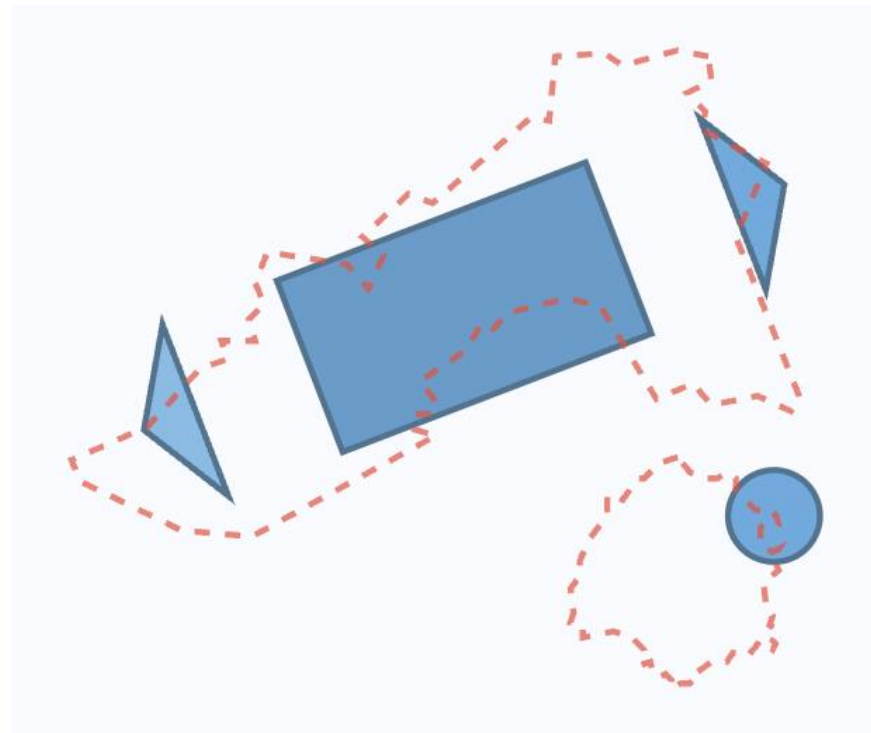
## Γεωμετρική Προσέγγιση



- Πραγματικό περίγραμμα → πολύπλοκο
- Μοντέλο:  $\text{Εμβαδόν Ν. Καβάλας} = \text{Εμβαδόν Γεωμετρικών Σχημάτων} + \text{Άλλα Πράγματα}$
- Σφάλμα: περιοχές που μένουν εκτός ή μέσα → Υπερ- ή υπο-εκτίμηση εμβαδού



Υπερεκτίμηση (Model Overfitting)



Υποεκτίμηση (Model Underfitting)

**Εμβαδόν Ν. Καβάλας = Εμβαδόν Γεωμετρικών Σχημάτων + Άλλα Πράγματα**

# Στατιστικά Μοντέλα

- ✓ Τα χρησιμοποιούμε για να συνοψίσουμε κατανομές.
- ✓ Τα χρησιμοποιούμε για να κάνουμε προβλέψεις σχετικά με το ποια θα μπορούσε να είναι η τιμή της επόμενης παρατήρησης που θα προστεθεί σε μια κατανομή.
- ✓ Τα χρησιμοποιούμε, επίσης, για να εξηγήσουμε τη μεταβλητότητα σε μία μεταβλητή με μια άλλη.

## ΈΝΑ ΑΠΛΟ ΜΟΝΤΕΛΟ

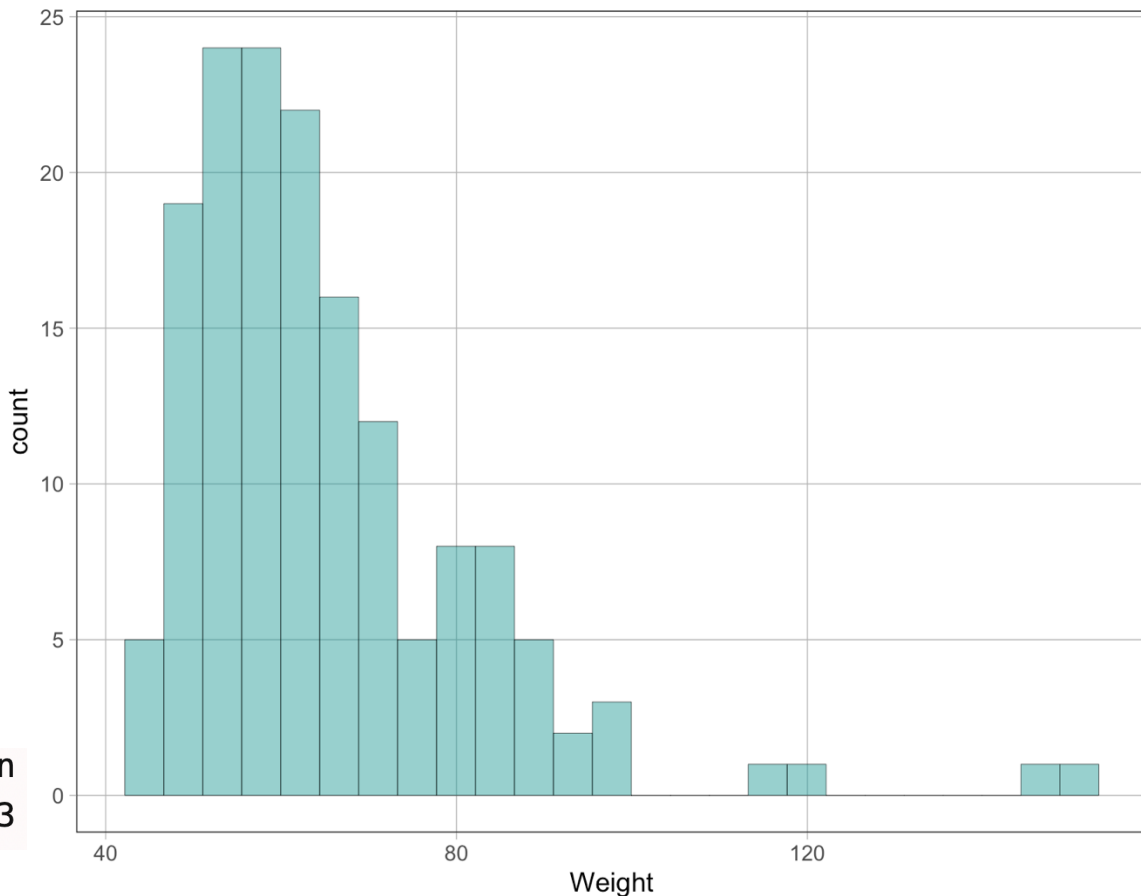
Αν έπρεπε να διαλέξετε μία μόνο τιμή για να αντιπροσωπεύσετε ολόκληρη την κατανομή μιας μεταβλητής, ποια θα ήταν αυτή;

Σκεπτόμενοι με διαφορετικό τρόπο: αν θέλατε να προβλέψετε ποια θα ήταν η τιμή της επόμενης τυχαία επιλεγμένης παρατήρησης, ποια τιμή θα ήταν η καλύτερη πρόβλεψή σας;

# Μοντελοποίηση Κατανομής με έναν μόνο Αριθμό

Ποια τιμή θα επιλέγατε για να αναπαραστήσετε ή να μοντελοποιήσετε την κατανομή του βάρους φοιτητών (Weight) και γιατί;

Αν θέλατε να προβλέψετε την τιμή του βάρους του επόμενου τυχαία επιλεγμένου φοιτητή, ποια τιμή θα ήταν η πρόβλεψή σας;

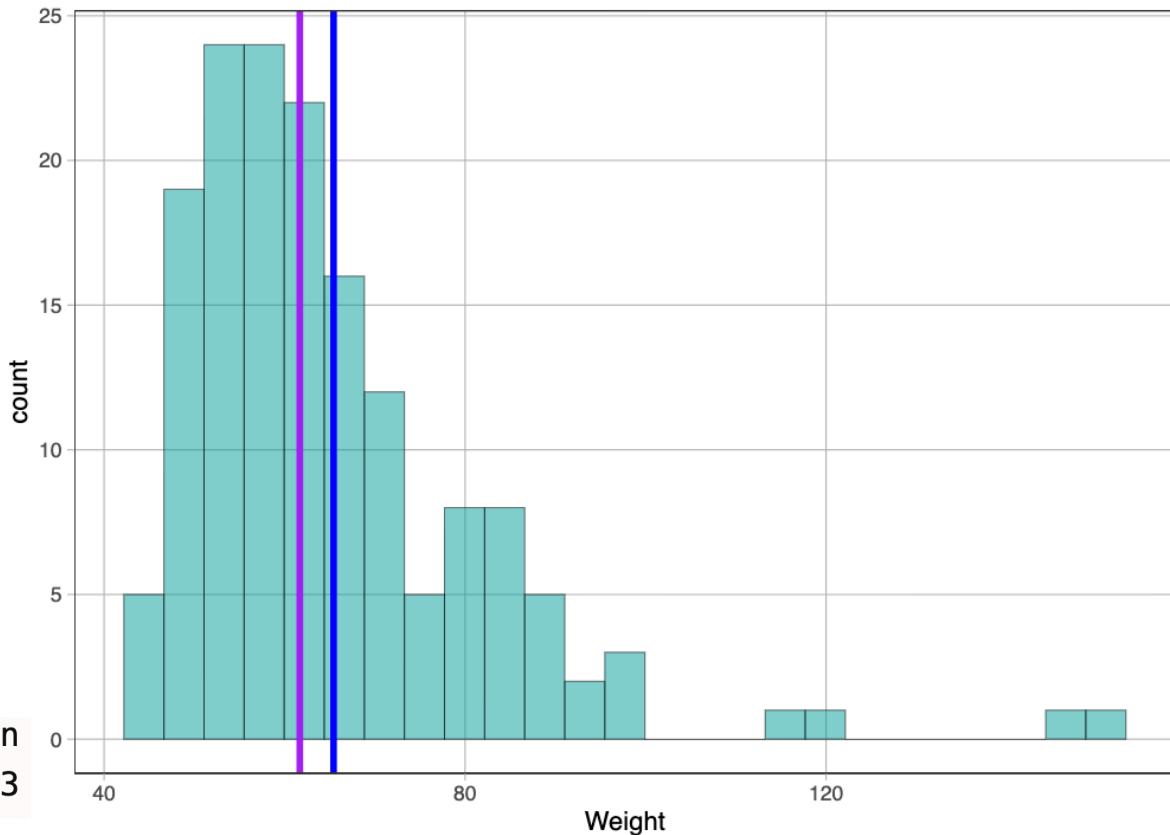


| min   | Q1    | median | Q3    | max    | mean  |
|-------|-------|--------|-------|--------|-------|
| 43.09 | 54.43 | 61.69  | 71.67 | 149.69 | 65.43 |

# Μοντελοποίηση Κατανομής με έναν μόνο Αριθμό

Ποια τιμή θα επιλέγατε για να αναπαραστήσετε ή να μοντελοποιήσετε την κατανομή του βάρους φοιτητών (Weight) και γιατί;

Αν θέλατε να προβλέψετε την τιμή του βάρους του επόμενου τυχαία επιλεγμένου φοιτητή, ποια τιμή θα ήταν η πρόβλεψή σας;



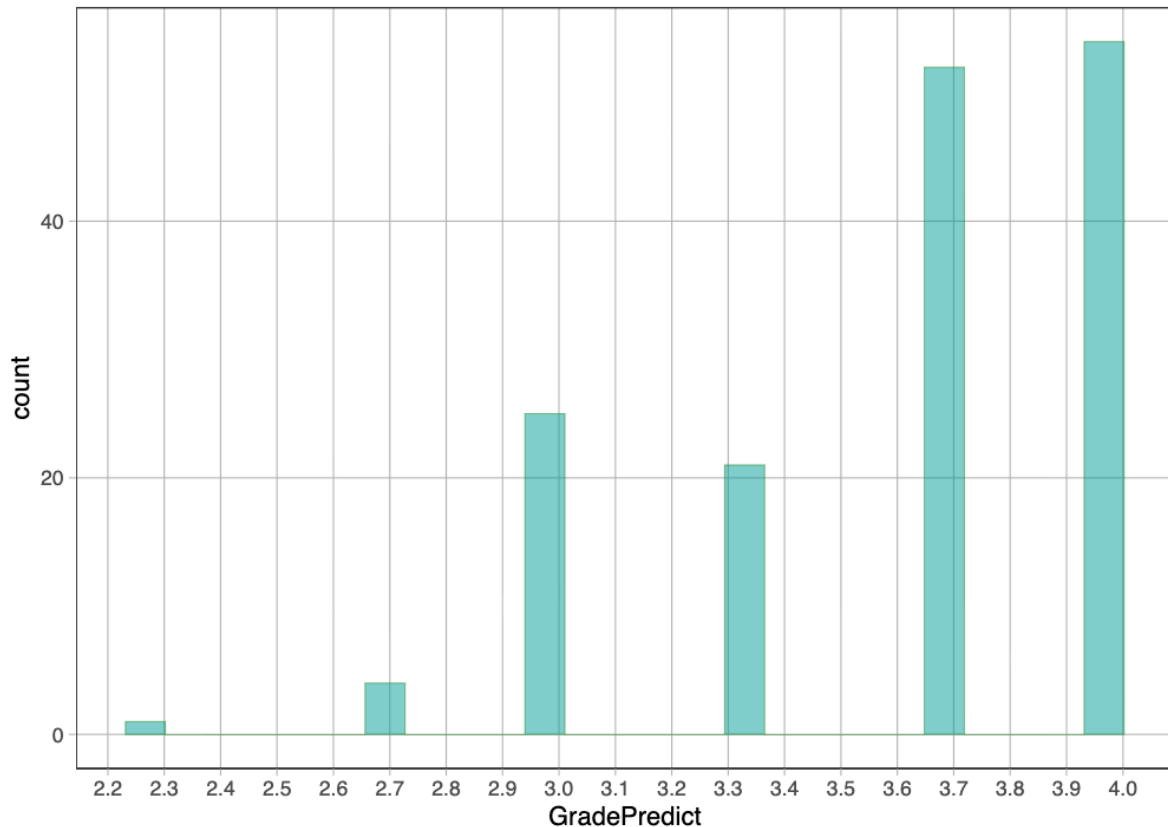
|       |       |        |       |        |       |
|-------|-------|--------|-------|--------|-------|
| min   | Q1    | median | Q3    | max    | mean  |
| 43.09 | 54.43 | 61.69  | 71.67 | 149.69 | 65.43 |

# Μοντελοποίηση Κατανομής με έναν μόνο Αριθμό

Ποια τιμή θα επιλέγατε για να αναπαραστήσετε ή να μοντελοποιήσετε την κατανομή της βαθμολογίας φοιτητή (GradePredict) και γιατί;

Αν θέλατε να προβλέψετε την τιμή της βαθμολογίας του επόμενου τυχαία επιλεγμένου φοιτητή, ποια τιμή θα ήταν η πρόβλεψή σας;

| min | Q1  | median | Q3 | max | mean     |
|-----|-----|--------|----|-----|----------|
| 2.3 | 3.3 | 3.7    | 4  | 4   | 3.603822 |

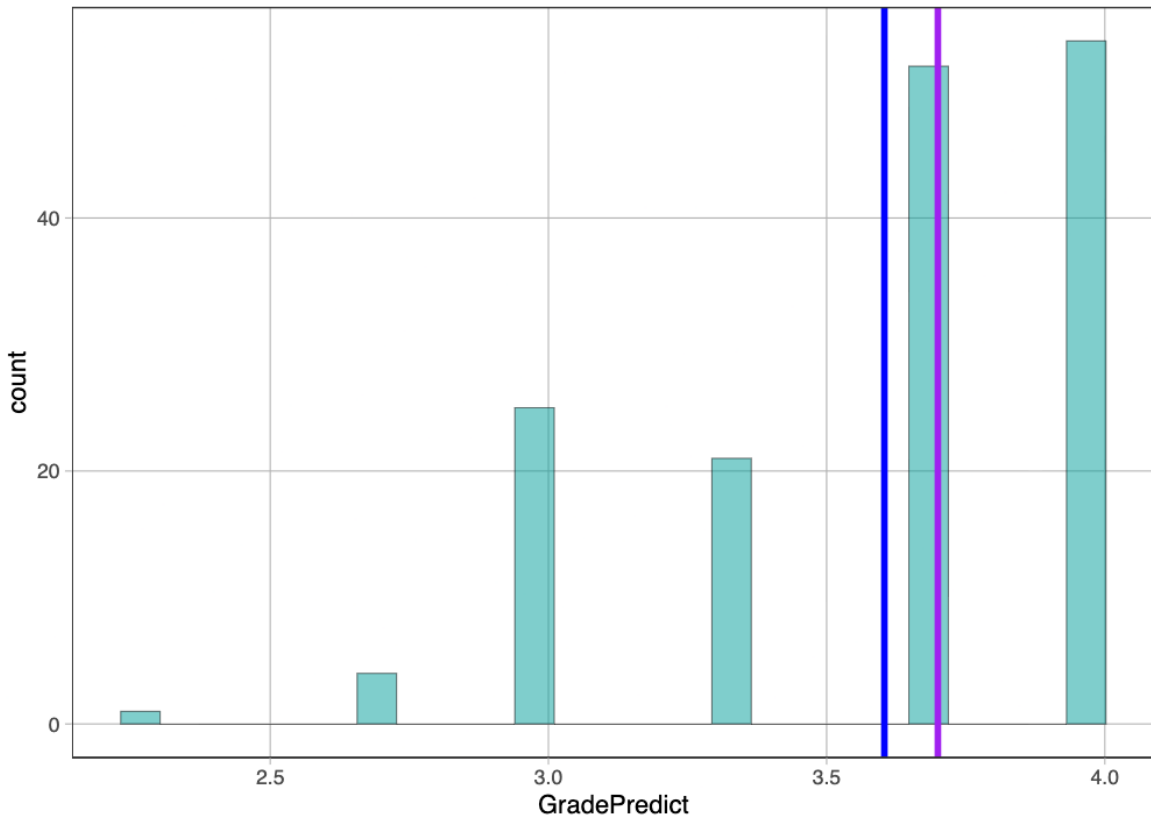


# Μοντελοποίηση Κατανομής με έναν μόνο Αριθμό

Ποια τιμή θα επιλέγατε για να αναπαραστήσετε ή να μοντελοποιήσετε την κατανομή της βαθμολογίας φοιτητή (GradePredict) και γιατί;

Αν θέλατε να προβλέψετε την τιμή της βαθμολογίας του επόμενου τυχαία επιλεγμένου φοιτητή, ποια τιμή θα ήταν η πρόβλεψή σας;

| min | Q1  | median | Q3 | max | mean     |
|-----|-----|--------|----|-----|----------|
| 2.3 | 3.3 | 3.7    | 4  | 4   | 3.603822 |

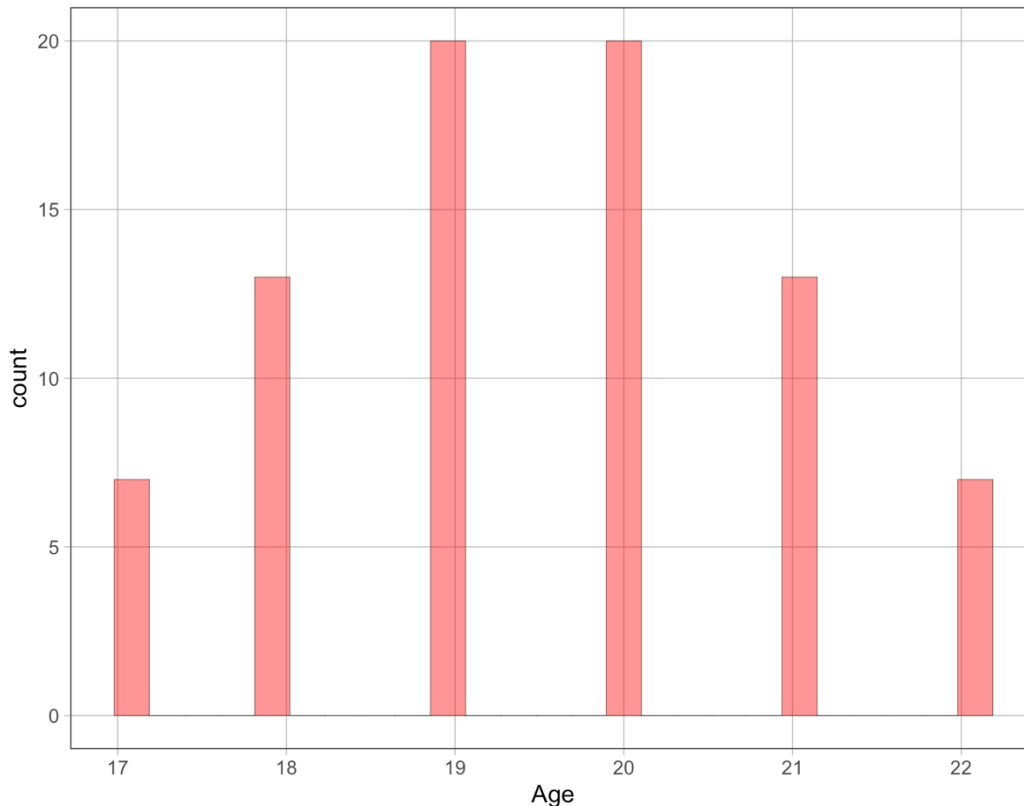


# Μοντελοποίηση Κατανομής με έναν μόνο Αριθμό

Ποια τιμή θα επιλέγατε για να αναπαραστήσετε ή να μοντελοποιήσετε την κατανομή της ηλικίας φοιτητή (Age) και γιατί;

Αν θέλατε να προβλέψετε την ηλικία του επόμενου τυχαία επιλεγμένου φοιτητή, ποια τιμή θα ήταν η πρόβλεψή σας;

|     |       |        |       |     |      |
|-----|-------|--------|-------|-----|------|
| min | Q1    | median | Q3    | max | mean |
| 17  | 18.75 | 19.5   | 20.25 | 22  | 19.5 |

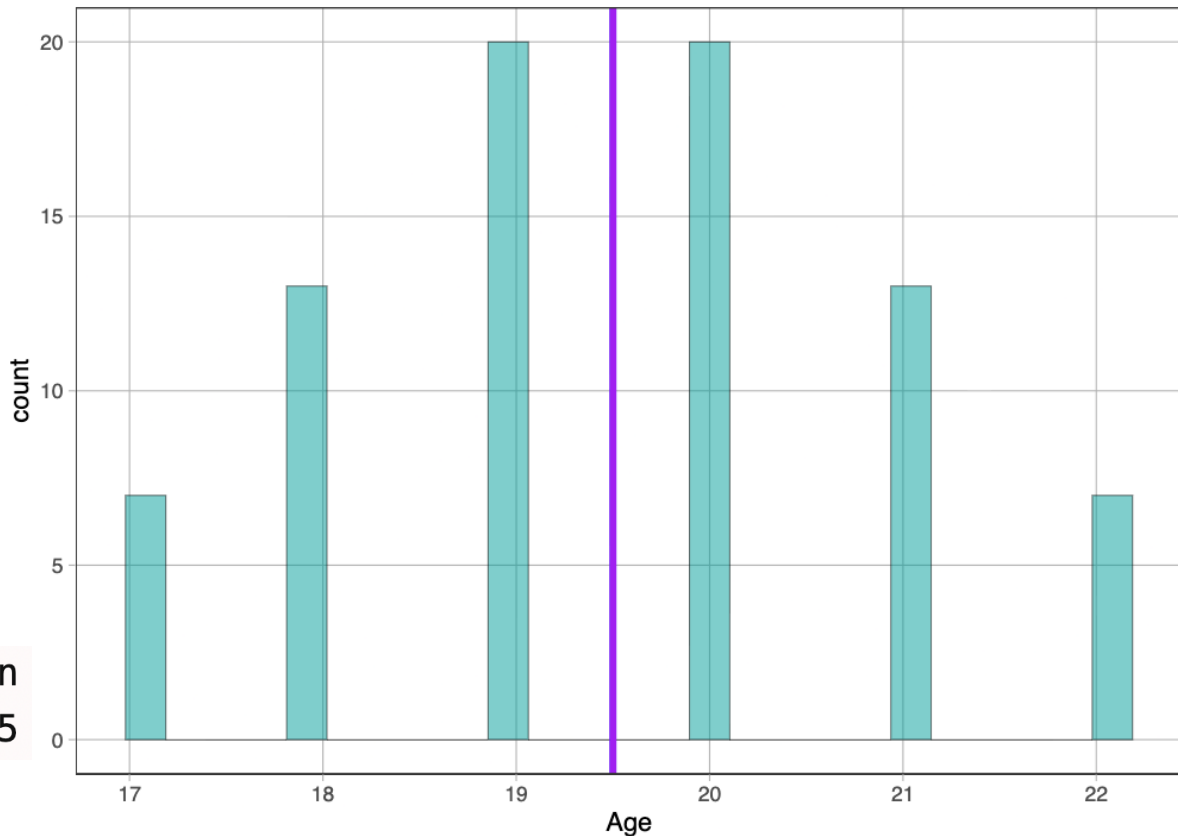


# Μοντελοποίηση Κατανομής με έναν μόνο Αριθμό

Ποια τιμή θα επιλέγατε για να αναπαραστήσετε ή να μοντελοποιήσετε την κατανομή της ηλικίας φοιτητή (Age) και γιατί;

Αν θέλατε να προβλέψετε την ηλικία του επόμενου τυχαία επιλεγμένου φοιτητή, ποια τιμή θα ήταν η πρόβλεψή σας;

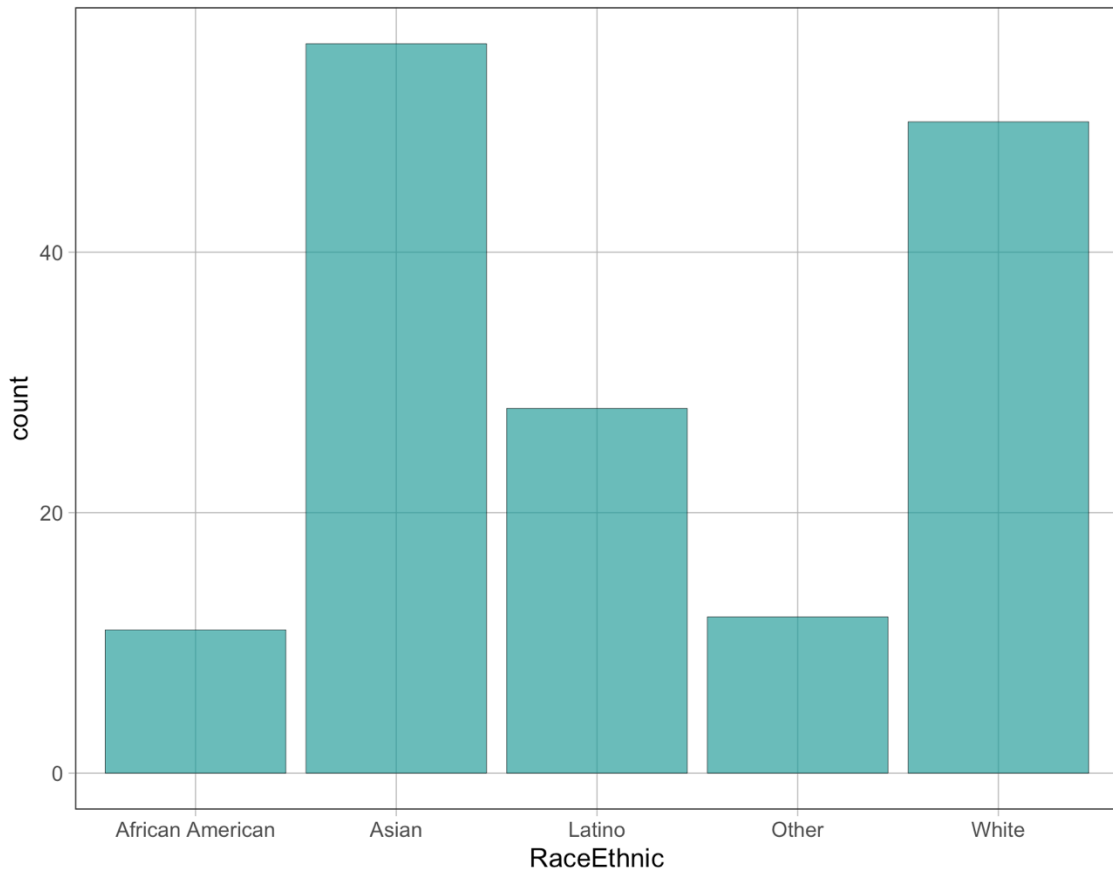
| min | Q1    | median | Q3    | max | mean |
|-----|-------|--------|-------|-----|------|
| 17  | 18.75 | 19.5   | 20.25 | 22  | 19.5 |



# Μοντελοποίηση Κατανομής με έναν μόνο Αριθμό

Ποια τιμή θα επιλέγατε για να αναπαραστήσετε ή να μοντελοποιήσετε την κατανομή της φυλετικής προέλευσης (RaceEthnic) φοιτητή και γιατί;

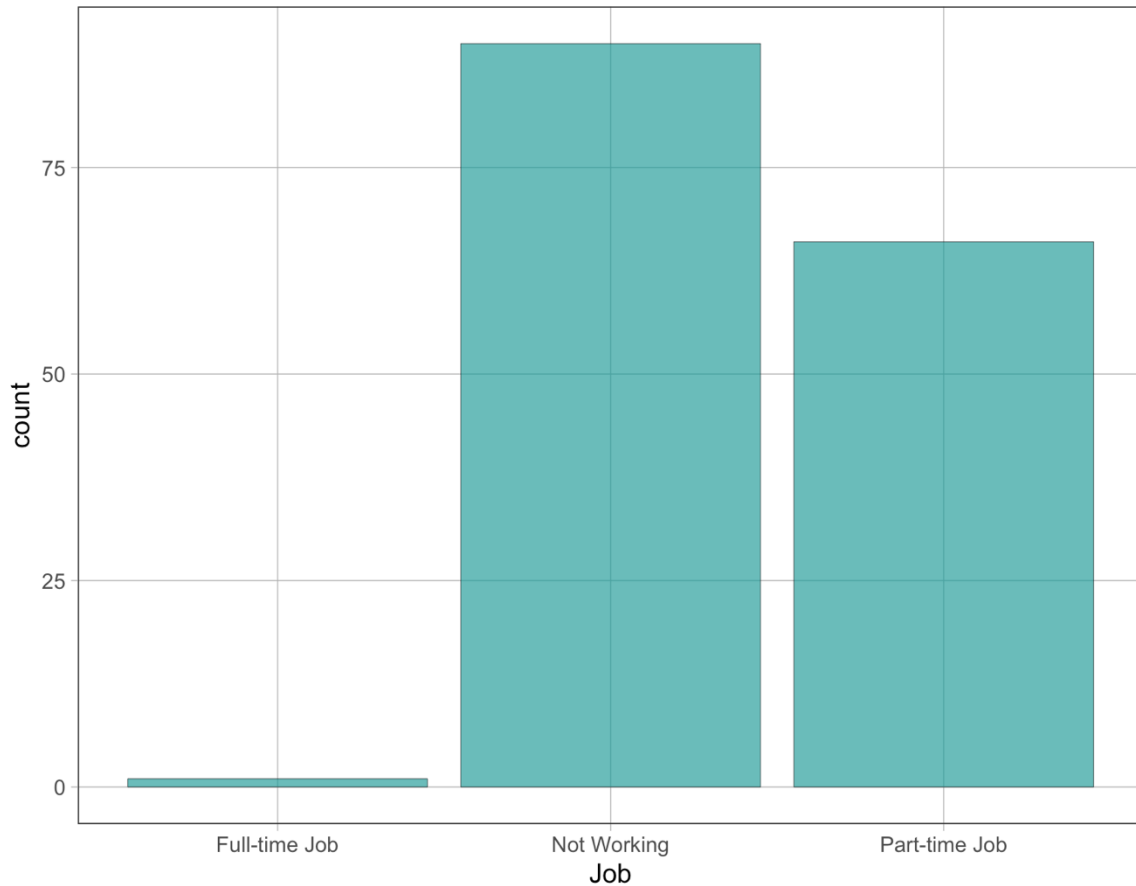
Αν θέλατε να προβλέψετε τη φυλετική προέλευση του επόμενου τυχαία επιλεγμένου φοιτητή, ποια τιμή θα ήταν η πρόβλεψή σας;



# Μοντελοποίηση Κατανομής με έναν μόνο Αριθμό

Ποια τιμή θα επιλέγατε να αναπαραστήσετε ή να μοντελοποιήσετε την κατανομή του είδους απασχόλησης φοιτητή (Job) και γιατί;

Αν θέλατε να προβλέψετε το είδος απασχόλησης του επόμενου τυχαία επιλεγμένου φοιτητή, ποια τιμή θα ήταν η πρόβλεψή σας;

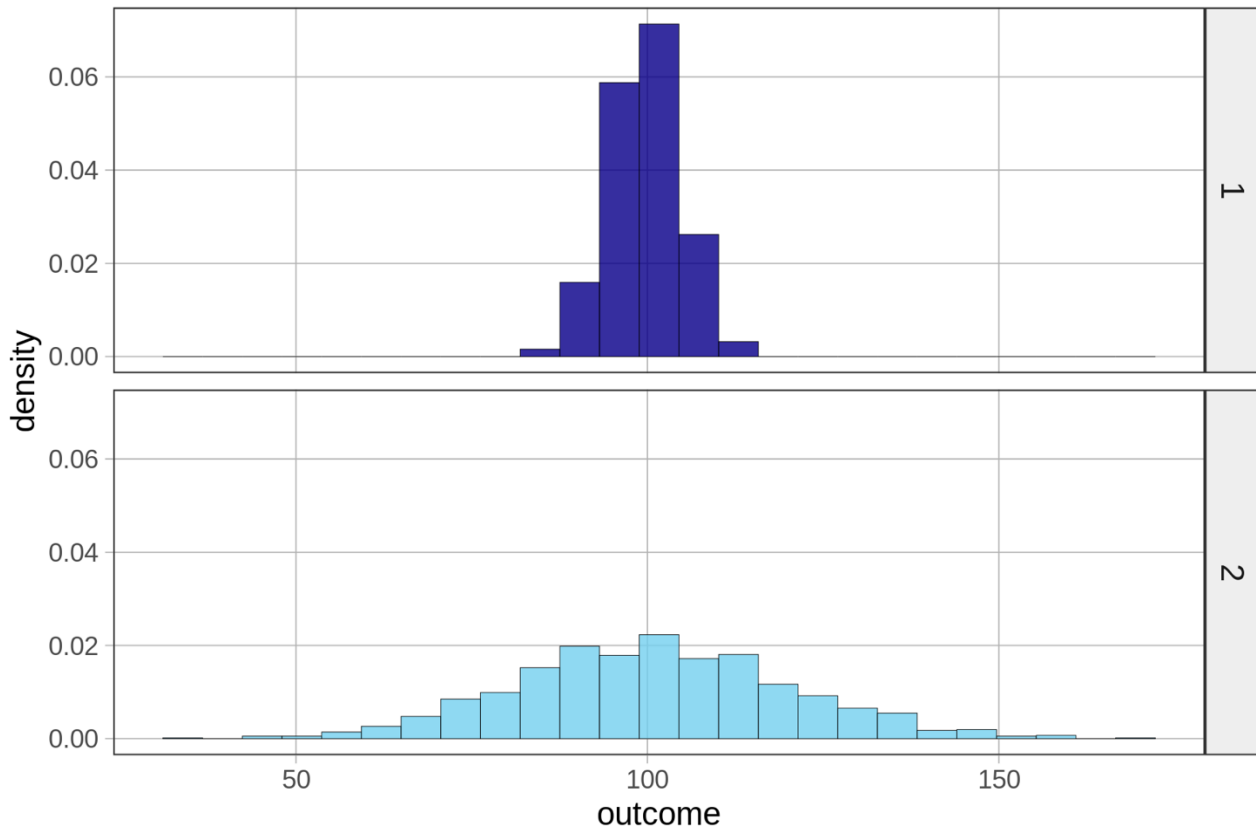


# Μοντέλο και Σφάλμα

Ποια τιμή θα επιλέγατε ως το καλύτερο μοντέλο για την κατανομή 1 (με το σκούρο μπλε) και γιατί;

Ποια τιμή θα επιλέγατε ως το καλύτερο μοντέλο για την κατανομή 2 (με το γαλάζιο) και γιατί;

Σε ποια κατανομή (1 ή 2) προσαρμόζεται το μοντέλο καλύτερα στα δεδομένα;  
Γιατί;

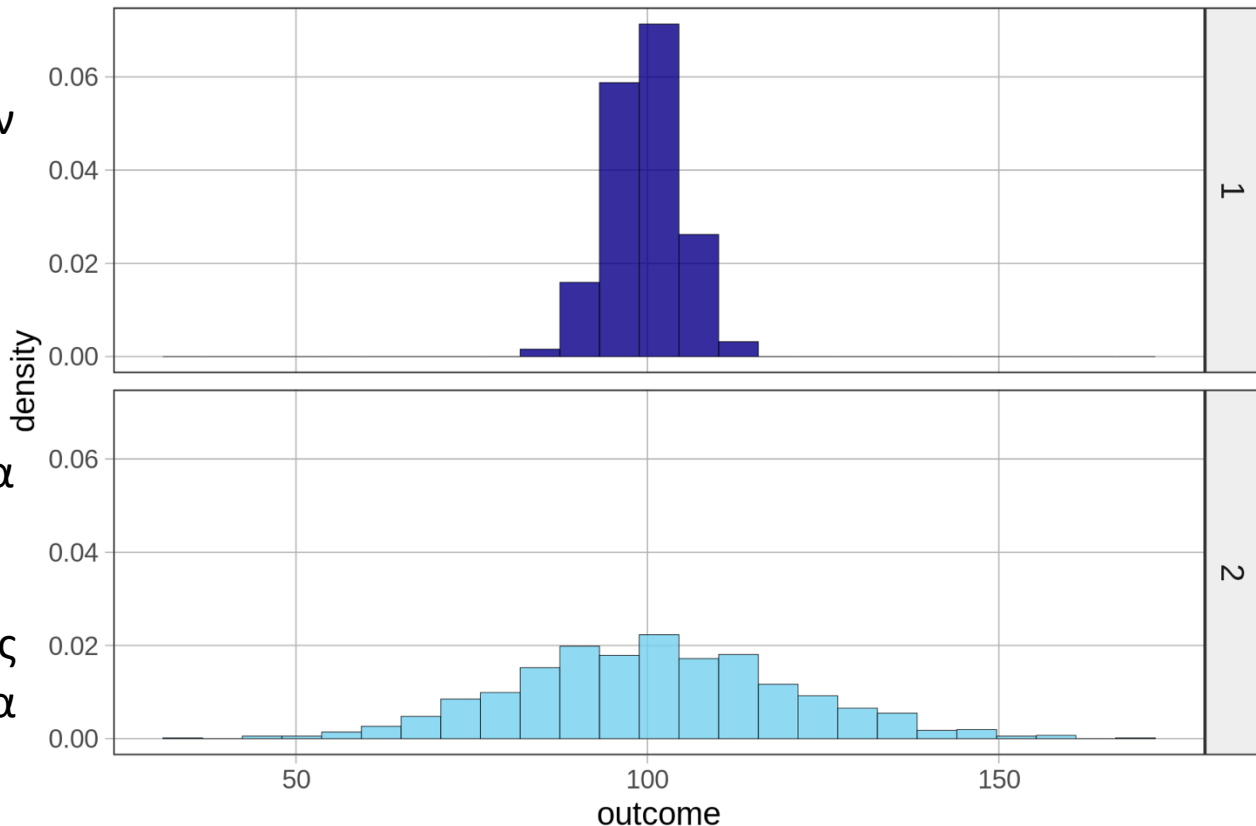


# Μοντέλο και Σφάλμα

- Το σφάλμα από το μοντέλο μπορεί να θεωρηθεί ως η απόκλιση των παρατηρούμενων τιμών από αυτήν την τιμή πρόβλεψης.

- **Στόχος:** ελαχιστοποίηση του σφάλματος

- Ένα μοντέλο μιας τιμής για μια κατανομή με μικρότερη διασπορά φαίνεται να έχει λιγότερο σφάλμα, και επομένως καλύτερη προσαρμογή, από ένα μοντέλο μιας τιμής για μια κατανομή με μεγαλύτερη διασπορά.



# Μοντελοποίηση Κατανομής με έναν μόνο Αριθμό

## Ποσοτική Μεταβλητή

Μέσος Όρος ή Διάμεσος

## Ποιοτική Μεταβλητή

Επικρατούσα Τιμή

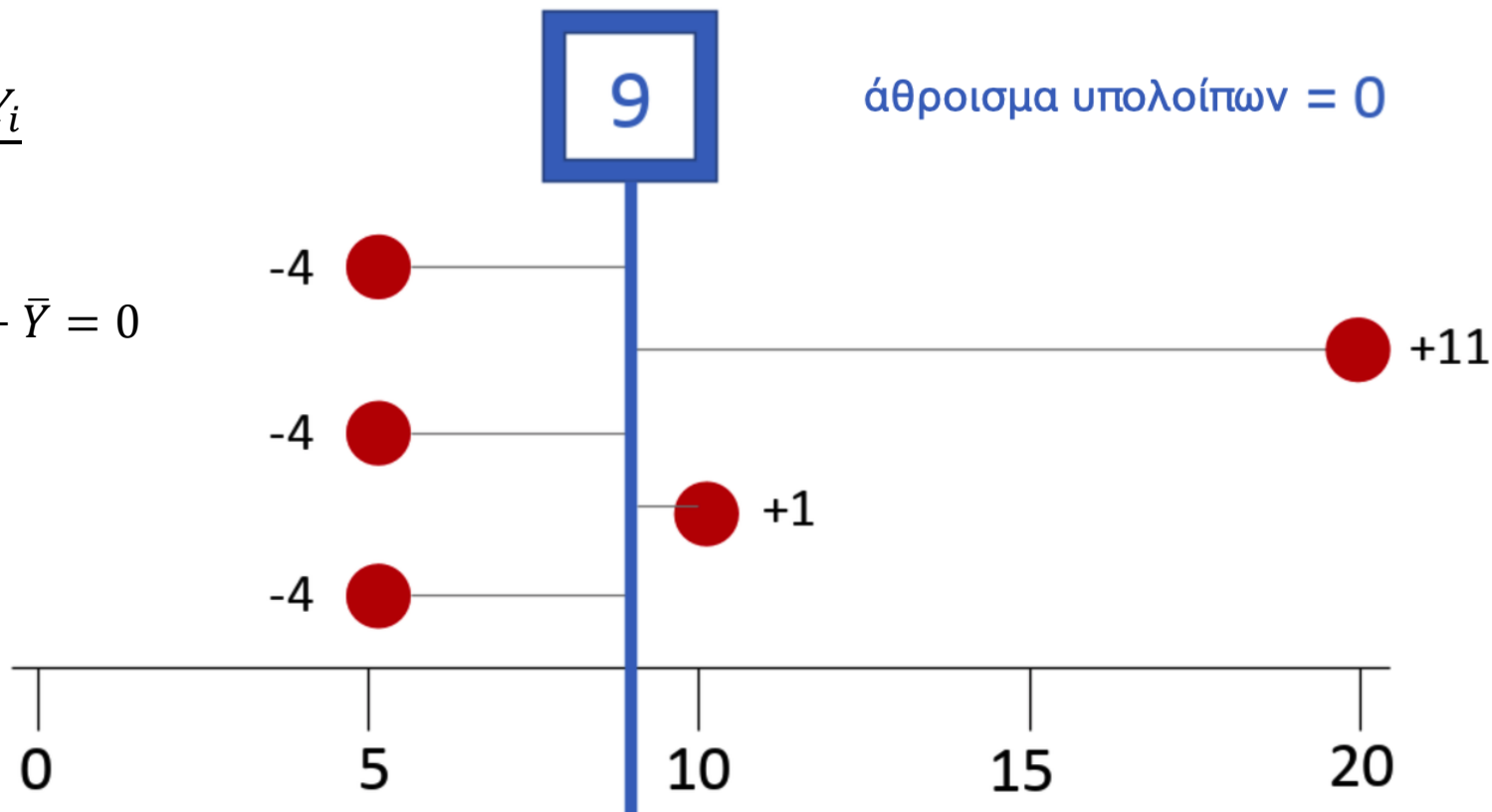
| Χαρακτηριστικό                 | Διάμεσος              | Μέσος όρος            |
|--------------------------------|-----------------------|-----------------------|
| Υπολογισμός                    | Μεσαία θέση           | Άθροισμα / n          |
| Ανθεκτικότητα σε ακραίες τιμές | Ναι                   | Όχι                   |
| Συμμετρικές κατανομές          | Διάμεσος = μέσος όρος | Διάμεσος = μέσος όρος |
| Ασύμμετρες κατανομές           | Καλύτερη              | Παραπλανητικός        |

**Παράδειγμα:** 5, 5, 5, 10, 20 → Διάμεσος = 5, Μέσος όρος = 9

# Ο Μέσος Όρος ως Σημείο Ισορροπίας

$$\bar{Y} = \frac{\sum Y_i}{n}$$

$$\sum Y_i - \bar{Y} = 0$$



# Το «Κενό» Μοντέλο

$$Y_i = b_0 + e_i$$

όπου  $b_0 = \bar{Y}$

(ο εκτιμημένος μέσος όρος)

- Καμία ανεξάρτητη μεταβλητή
- Πρόβλεψη ίδια για όλους:  $b_0$
- Χρησιμεύει ως μοντέλο αναφοράς

```
lm(Thumb ~ NULL, data = Fingers)
```

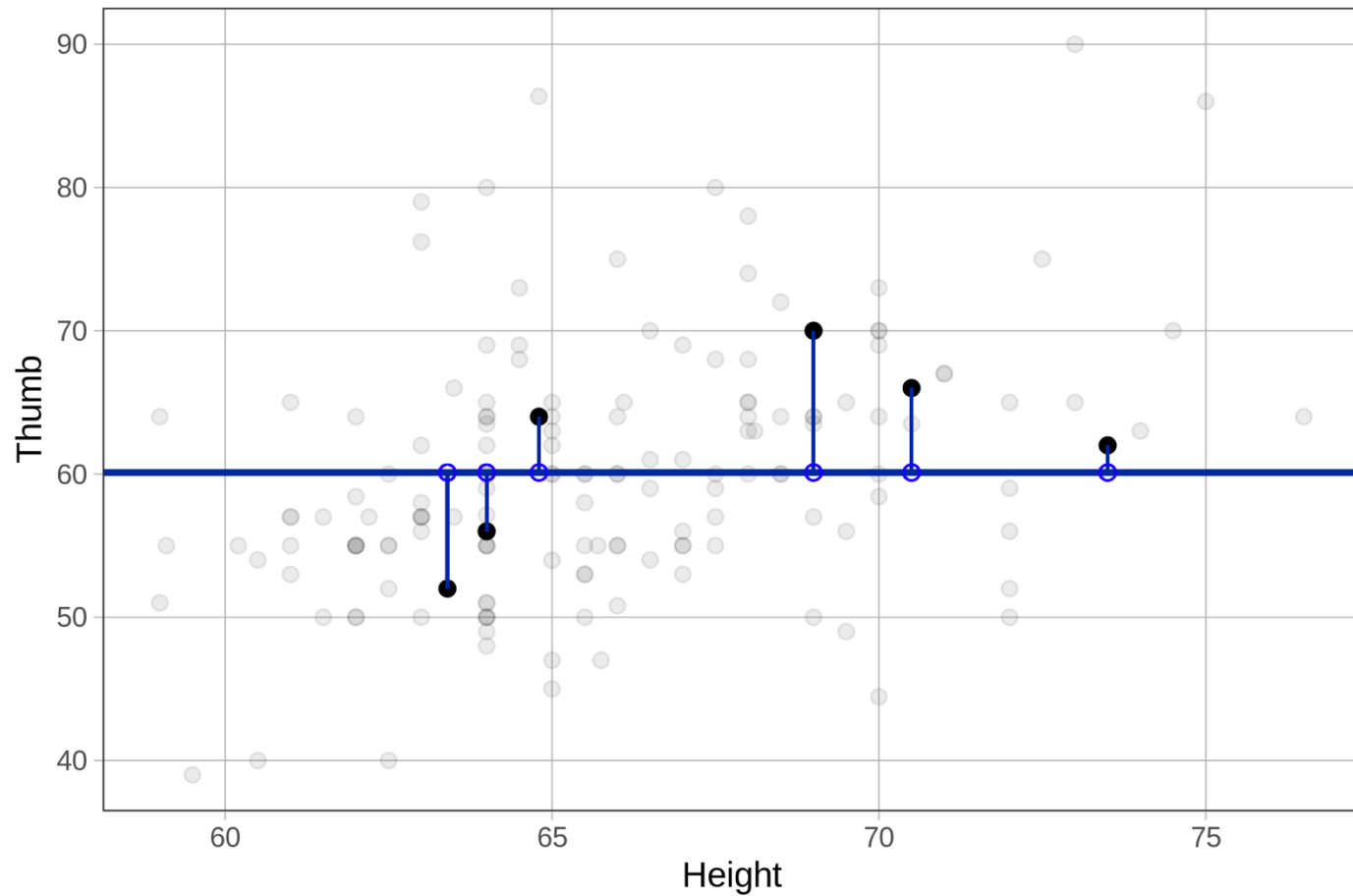
Call:

```
lm(formula = Thumb ~ NULL, data  
= Fingers)
```

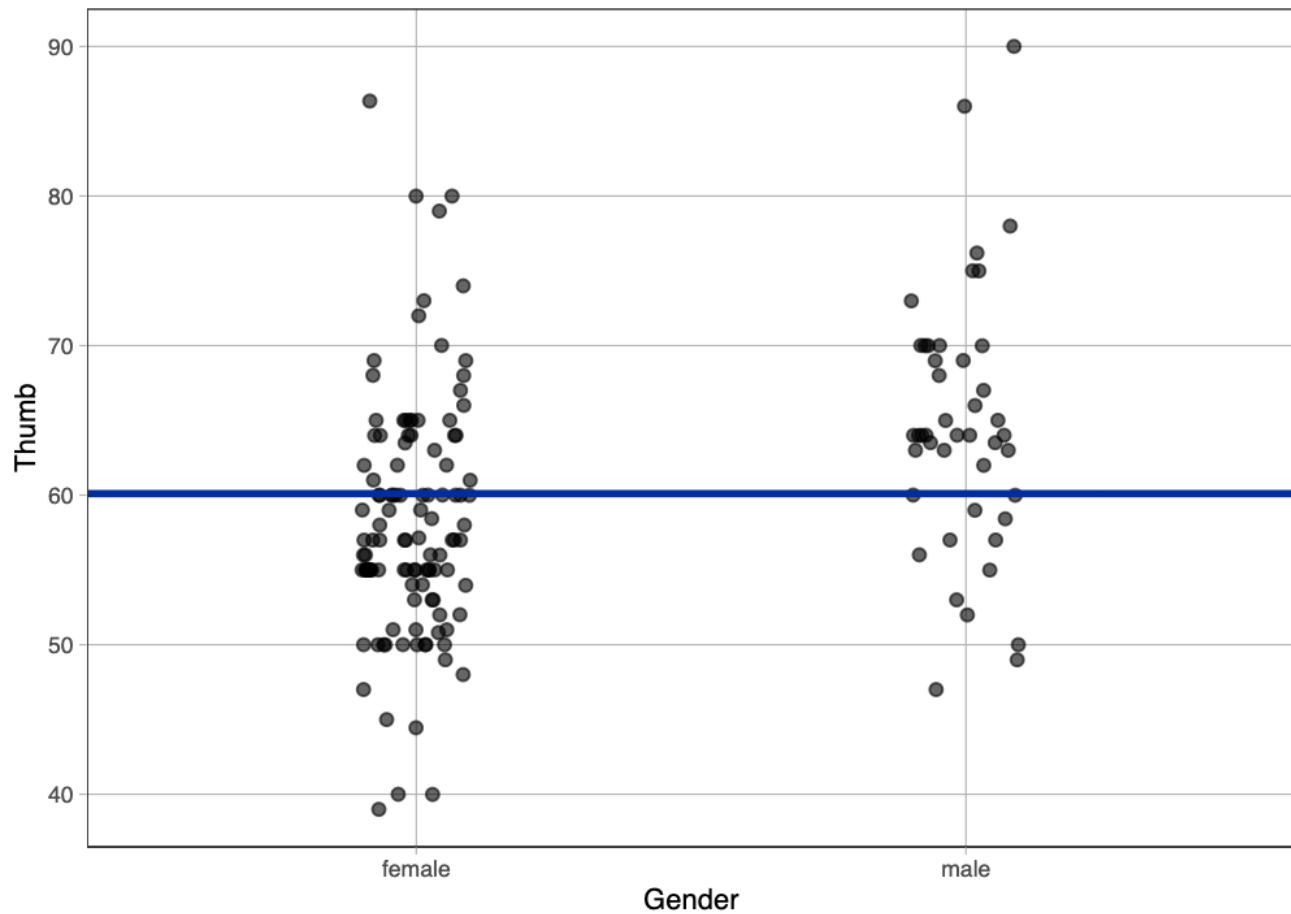
Coefficients:

```
(Intercept)  
60.1
```

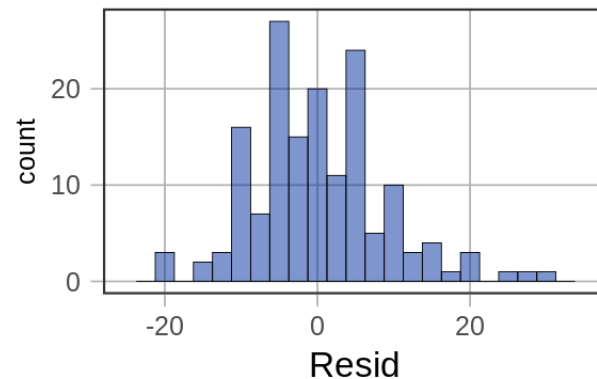
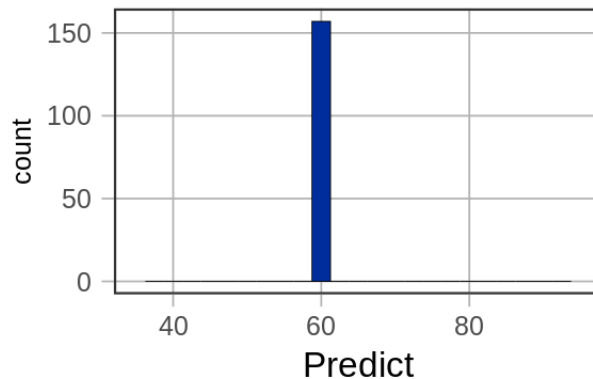
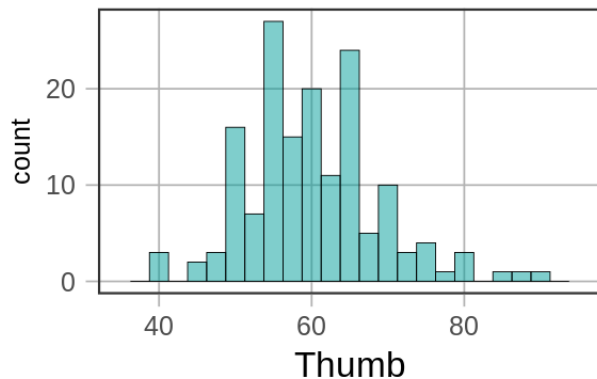
# Σφάλματα Κενού Μοντέλου



# Σφάλματα Κενού Μοντέλου



# Η Κατανομή των Υπολοίπων



- Εξετάστε τις παραπάνω κατανομές.
- Πώς μοιάζει η κατανομή των δεδομένων του μήκους αντίχειρα;
- Πώς θα σχολιάζατε την κατανομή της Predict (προβλέψεις του μοντέλου);
- Πώς θα σχολιάζατε την κατανομή της Resid (υπόλοιπα ή σφάλματα);
- Τι κοινό έχουν και σε τι διαφέρουν η κατανομή των δεδομένων του μήκους αντίχειρα και η κατανομή των σφαλμάτων;

# ΔΕΔΟΜΕΝΑ = ΜΟΝΤΕΛΟ + ΣΦΑΛΜΑ

Μήκος αντίχειρα = Μέσος Όρος + Σφάλμα

$$Y_i = \bar{Y} + e_i$$

$$Y_i = b_0 + e_i$$

Γενικό Γραμμικό Μοντέλο

$$\underbrace{Y_i}_{\text{Μήκος αντίχειρα}} = \underbrace{b_0}_{\text{Πρόβλεψη}} + \underbrace{e_i}_{\text{Υπόλοιπο}}$$

Το Μοντέλο στη ΔΠΔ

$$Y_i = \beta_0 + \epsilon_i$$

|                                                                                   | Δεδομένα                                                                                                     | Πληθυσμός                                                                                                                         |
|-----------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------|
|                                                                                   | Μοντέλο που κατασκευάζεται<br>βάσει δεδομένων                                                                | Μοντέλο που προσπαθούμε να<br>εκτιμήσουμε                                                                                         |
|                                                                                   | (εκτιμημένο)                                                                                                 | (άγνωστο)                                                                                                                         |
| Λεκτική εξίσωση                                                                   | Μήκος αντίχειρα ατόμου $i$ =<br>μέσος όρος δείγματος + σφάλμα                                                | Μήκος αντίχειρας ατόμου $i$ =<br>μέσος όρος πληθυσμού +<br>σφάλμα                                                                 |
| Σημειογραφία για<br>μοντέλο μιας<br>παραμέτρου που<br>χρησιμοποιεί το<br>μέσο όρο | $Y_i = \bar{Y} + e_i$                                                                                        | $Y_i = \mu + \varepsilon_i$                                                                                                       |
|                                                                                   | • $Y_i$ είναι ο αντίχειρας του<br>ατόμου $i$                                                                 | • $Y_i$ είναι ο αντίχειρας του<br>ατόμου $i$                                                                                      |
|                                                                                   | • $\bar{Y}$ είναι ο μέσος όρος του<br>δείγματος                                                              | • $\mu$ είναι ο μέσος όρος του<br>πληθυσμού (άγνωστος)                                                                            |
|                                                                                   | • $e_i$ είναι η διαφορά μεταξύ του<br>μήκους αντίχειρα του ατόμου $i$<br>και του μέσου όρου του<br>δείγματος | • $\varepsilon_i$ είναι η διαφορά μεταξύ του<br>μήκους αντίχειρα του ατόμου $i$<br>και του μέσου όρου του<br>πληθυσμού (άγνωστος) |

Μοντέλο μίας παραμέτρου

$$Y_i = b_0 + e_i$$

$$Y_i = \beta_0 + \varepsilon_i$$