R – Biostatistics in Biosciences

Dr. Timokratis Karamitros
Senior Researcher

Head, Bioinformatics and Applied Genomics Unit
Hellenic Pasteur Institute, Athens, GR

tkaram@pasteur.gr

HELLENIC PASTEUR INSITUTE 100 YEARS

Why in R?

Learning a new topic becomes easier if it is motivated by interesting and engaging applied problems!

We will explore each new topic with a relevant problem from biology
each new topic with a relevant problem from biology

We then try to reach the solution intuitively before discussing the related statistical methods

Timokratis Karamitros
Bioinformatics and Applied Genomics Unit

Γιατί γενικότερα?

Για να κάνεις επιστήμη, ή για να κατανοήσεις την επιστήμη που κάνουν άλλοι, χρειάζεται να κατανοήσεις τις αρχές του πειραματικού σχεδιασμού, της συλλογής δεδομένων, της στατιστικής ανάλυσης και της επιστημονικής παρουσίασης

πρέπει να ξέρεις πώς να παίρνεις λογικές αποφάσεις σχετικά με τον σχεδιασμό μελετών και τη συλλογή δεδομένων και στη συνέχεια να τα ερμηνεύεις σωστά

στόχος αυτού του μαθήματος είναι να σου δώσει μια καλή πρακτική κατανόηση των βασικών τεχνικών στατιστικής που χρησιμοποιούνται ευρέως στη βιολογία, να αποκτήσεις ένα σύνολο δεξιοτήτων που μπορείς να εφαρμόσεις σε εργαστηριακές ασκήσεις, μαθήματα πεδίου και σε ερευνητικές εργασίες

Ζούμε σε έναν κόσμο που κατακλύζεται ολοένα και περισσότερο από δεδομένα, και οι άνθρωποι που γνωρίζουν πώς να τα ερμηνεύουν έχουν μεγάλη ζήτηση!!!

Προαπαιτούμενα προγραμματισμού

Έχεις χρησιμοποιήσει R και RStudio στο παρελθόν? έχεις ξαναγράψει κώδικα R και ξέρεις πώς να τον «τρέχεις» ??

Η R είναι μια γλώσσα προγραμματισμού σχεδιασμένη πρωτίστως για την υποστήριξη της ανάλυσης δεδομένων. Τη χρησιμοποιούμε για να πραγματοποιήσουμε όλες τις στατιστικές αναλύσεις

Η R είναι μια πλήρης γλώσσα προγραμματισμού, που σημαίνει ότι αλληλεπιδρούμε μαζί της γράφοντας κώδικα. Ουσιαστικά, η R περιμένει εντολές με τη μορφή κειμένου (κώδικας R). Ο χρήστης μπορεί να πληκτρολογήσει αυτές τις εντολές ή να τις στείλει μέσω άλλου βοηθητικού προγράμματος.

Χρήση πακέτων

Τα πακέτα της R επεκτείνουν τη βασική λειτουργικότητα της γλώσσας, ώστε να μπορούμε να κάνουμε περισσότερα με αυτήν. Συγκεντρώνουν κώδικα R, δεδομένα και τεκμηρίωση σε μια τυποποιημένη μορφή, που καθιστά εύκολη τη χρήση και την κοινή τους διάθεση.

Το πακέτο **readr** για την εισαγωγή δεδομένων στη R
Το πακέτο **dplyr** για τον χειρισμό δεδομένων
Το πακέτο **ggplot2** για τη δημιουργία γραφημάτων

```
install.packages("dplyr")
library("dplyr")
```


1. Εργασία με Δεδομένα

Όλα τα παραδείγματα χρησιμοποιούν σύνολα δεδομένων που είναι αποθηκευμένα σε αρχεία CSV.

Χρησιμοποιείτε τη **readcsv** από το πακέτο *readr* για να εισάγετε αυτά τα δεδομένα στη R.

Όταν χρησιμοποιούμε την **readcsv**, καταλήγουμε με ένα *tibble*: την έκδοση του tidyverse για τα data frames.

Τα *tibbles* και τα *data frames* είναι αντικείμενα που μοιάζουν με πίνακες, με γραμμές και στήλες. Συγκεντρώνουν διαφορετικές μεταβλητές, αποθηκεύοντας καθεμία ως μια στήλη με όνομα.

τα δεδομένα μέσα σε αυτά τα αντικείμενα είναι οργανωμένα σύμφωνα με τις αρχές των *tidy data*.

Τα tidy data έχουν μια συγκεκριμένη δομή που τα κάνει εύκολα στον χειρισμό, τη μοντελοποίηση και την οπτικοποίηση.

Ένα *tidy* σύνολο δεδομένων είναι εκείνο στο οποίο κάθε μεταβλητή βρίσκεται μόνο σε μία στήλη και κάθε γραμμή περιέχει μία μοναδική παρατήρηση.

Διαχείριση δεδομένων με dplyr

Το dplyr καθιστά εύκολο τον χειρισμό «ορθογώνιων δεδομένων» που βρίσκονται σε tibbles και data frames.

select: λαμβάνει ένα υποσύνολο μεταβλητών,
mutate: δημιουργεί νέες μεταβλητές,
filter: λαμβάνει ένα υποσύνολο γραμμών,
arrange: αναδιατάσσει τις γραμμές,
summarise: υπολογίζει συνοπτικές πληροφορίες για ομάδες.
rename και **groupby**

Οπτικοποίηση με ggplot2

Μια οπτικοποίηση δημιουργείται ορίζοντας ένα ή περισσότερα *layers* (στρώματα), όπου κάθε στρώμα συνδέεται με κάποια δεδομένα και ένα σύνολο κανόνων για το πώς αυτά θα εμφανιστούν.

Θα πρέπει να μπορείτε να:
οπτικοποιείτε κατανομές μιας μεταβλητής χρησιμοποιώντας dot plots και ιστογράμματα (για αριθμητικές μεταβλητές) και bar plots (για κατηγορικές μεταβλητές),
οπτικοποιείτε συσχετίσεις μεταξύ ζευγών μεταβλητών με scatter plots (αριθμητική vs αριθμητική), bar plots (κατηγορική vs κατηγορική) και box plots (αριθμητική vs κατηγορική).

Επιστημονική Μέθοδος και Διαψευσιμότητα (falsificationism)

1.1 Στάδια στην Επιστημονική Διαδικασία

Παρατηρήσεις (Observations)
Παρατήρηση: Πληροφορίες ή εντυπώσεις για γεγονότα ή αντικείμενα. Οι ερωτήσεις δεν προέρχονται από καθαρή αφηρημένη σκέψη, αλλά από την παρατήρηση του φυσικού κόσμου (άμεσες παρατηρήσεις, μοτίβα σε δεδομένα, συσσωρευμένο έργο άλλων).
Ερωτήσεις (Questions)
Ερώτηση: Τι είναι αυτό που θέλουμε να μάθουμε. Είναι σημαντικό η ερώτηση να είναι αρκετά εστιασμένη ώστε να μπορεί να απαντηθεί σε μια συγκεκριμένη μελέτη.
Υποθέσεις (Hypotheses)
Υπόθεση: Μια εξήγηση που προτείνεται για να δικαιολογήσει τα παρατηρούμενα γεγονότα.
Το κριτικό χαρακτηριστικό στη βιολογία είναι ότι η υπόθεση πρέπει να προσδιορίζει κάποια βιολογική διαδικασία ή μηχανισμό. Μία ερώτηση συχνά δημιουργεί περισσότερες από μία υποθέσεις.
Προβλέψεις (Predictions)
Πρόβλεψη: Τι αναμένουμε να δούμε αν η υπόθεση είναι αληθής. Οι υποθέσεις μπορεί να μην είναι άμεσα ελέγξιμες. Χρειαζόμαστε προβλέψεις για να καθορίσουμε τι δεδομένα να συλλέξουμε και τι μοτίβο να αναμένουμε.

Ιδανική Πρόβλεψη: Αυτή που είναι **μοναδική** για τη συγκεκριμένη υπόθεση, έτσι ώστε αν η πρόβλεψη επαληθευτεί, μόνο μία υπόθεση να μπορεί να είναι υπεύθυνη.

Ένα Παράδειγμα: Οι Γαμβάροι (Gammarus)

Στάδιο	**Παράδειγμα**
Παρατήρηση	Οι γαμβάροι (Gammarus) βρίσκονται σχεδόν αποκλειστικά **κάτω από πέτρες** σε ένα ρέμα.
Ερώτηση	Γιατί ο *Gammarus* περνάει τον περισσότερο χρόνο του κάτω από πέτρες;
Υποθέσεις	1. Χρειάζονται **καταφύγιο** από το ρεύμα. 2. Η **τροφή** (φυλλόπτωση) συσσωρεύεται κάτω από τις πέτρες. 3. Προστατεύονται από **οπτικά κυνηγούς θηρευτές** (π.χ., ψάρια).
Προβλέψεις	**Καταφύγιο:** Μεγαλύτερη αναλογία *Gammarus* σε ανοιχτό χώρο σε **ρέματα με αργή ροή**. **Τροφή:** *Gammarus* **δεν** θα συσσωρευτεί κάτω από πέτρες από τις οποίες έχει αφαιρεθεί όλη η φυλλόπτωση. **Θήρευση:** *Gammarus* θα συσσωρεύεται περισσότερο κάτω από πέτρες σε ρέματα **με παρουσία ψαριών**.

Έλεγχος Υποθέσεων: Διαψευσιμότητα

Το σημαντικό στην επιστημονική εξέταση είναι ότι οι υποθέσεις **είτε απορρίπτονται είτε δεν απορρίπτονται**, αλλά **σπάνια αποδεικνύονται** (εκτός από ασήμαντες περιπτώσεις).

- **Γιατί δεν αποδεικνύονται;** Επειδή δεν μπορούμε να είμαστε σίγουροι ότι έχουμε σκεφτεί **όλες τις πιθανές εξηγήσεις**. Η εύρεση αποδεικτικών στοιχείων που υποστηρίζουν μια πρόβλεψη **δεν** εγγυάται ότι η υπόθεση είναι η **μόνη αληθινή**.
- **Η Δύναμη της Απόρριψης:** Αντίθετα, αν βρούμε στοιχεία που **αντιφάσκουν άμεσα** με την πρόβλεψη, μπορούμε να είμαστε **σίγουροι** ότι η υπόθεση **δεν μπορεί να είναι αληθής**.
  Παράδειγμα: Η υπόθεση ότι **όλα** τα κωνοφόρα είναι αειθαλή **διαψεύδεται** οριστικά με την εύρεση του πρώτου **λάριου** (φύλλοβόλο κωνοφόρο).
- **Διαψευσιμότητα (Falsificationism):** Ο ιδεατός κύκλος **διατύπωσης υποθέσεων** και στη συνέχεια αναζήτησης αποδεικτικών στοιχείων ικανών να τις **διαψεύσουν**, είναι το μοντέλο που πρότεινε ο φιλόσοφος **Karl Popper**.

Βεβαιότητα και Επιστημονική Πρόοδος

1. Η Έννοια της «Αποδοχής» Υποθέσεων

Στην πράξη, οι επιστήμονες **αποδέχονται** συνεχώς υποθέσεις, παρόλο που η λογική της διαψευσιμότητας υποδηλώνει ότι μόνο η απόρριψη (rejection) είναι οριστική.

- **Αποδοχή μέσω Αντίστασης:** Μια υπόθεση γίνεται πιο «αποδεκτή» και θεωρείται ως η **«καλύτερη εκτίμηση της αλήθειας»** όταν επαναλαμβανόμενες προσπάθειες να βρεθούν **αντιπαραδείγματα (counter-evidence)** αποτυγχάνουν.
- **Προσωρινή Αλήθεια:** Αυτή η αποδοχή είναι πάντα **προσωρινή**. Υπάρχει η επιφύλαξη ότι μπορεί να εμφανιστεί μια **καλύτερη ιδέα** (υπόθεση) στο μέλλον.

2. Οι Φιλοσοφίες του Πραγματικού Επιστημονικού Έργου

Οι Φιλοσοφίες του Πραγματικού Επιστημονικού Έργου

Α. Η Θεωρία του Thomas Kuhn: Η «Αλλαγή Παραδείγματος» (Paradigm Shift)

Ο Kuhn υποστήριξε ότι η επιστήμη προχωρά μέσα από:

- **Περιόδους Στασιμότητας (Κανονική Επιστήμη):** Οι επιστήμονες εργάζονται μέσα σε ένα **αποδεκτό παράδειγμα (paradigm)** – ένα σύνολο κοινών πεποιθήσεων και μεθόδων.
- **Συσσώρευση Ασυμβατότητας:** Αρχίζουν να συσσωρεύονται στοιχεία που **δεν υποστηρίζουν** πλήρως το υπάρχον παράδειγμα.
- **Επανάσταση:** Τελικά, η ασυμβατότητα οδηγεί σε μια **επιστημονική επανάσταση** που **απορρίπτει ολόκληρο** το παλιό παράδειγμα και προτείνει μια **νέα άποψη**.
- **Προέλευση:** Η φράση **«αλλαγή παραδείγματος» (paradigm shift)** προέρχεται από αυτήν τη φιλοσοφία.

Οι Φιλοσοφίες του Πραγματικού Επιστημονικού Έργου

Β. Η Πρόταση του Imre Lakatos: Τα «Ερευνητικά Προγράμματα» (Research Programmes)

Ο Lakatos προσπάθησε να συμβιβάσει τις απόψεις του Kuhn, προτείνοντας ότι οι επιστημονικές ιδέες είναι οργανωμένες σε **«ερευνητικά προγράμματα»**:

- **Πυρήνας (Core Ideas):** Υπάρχουν **βασικές ιδέες** μέσα στο πρόγραμμα που **δεν αμφισβητούνται** εύκολα.
- **Προστατευτική Ζώνη (Auxiliary Hypotheses):** Υπάρχουν **περιφερειακές ιδέες** που συνεχώς **αμφισβητούνται** και **προσαρμόζονται** μέσω της διαψευσιμότητας.
- **Πρόοδος:** Αυτή η διαδικασία επιτρέπει σε κάθε ερευνητικό πρόγραμμα να **προοδεύει** ως σύνολο.

Timokratis Karamitros Bioinformatics and Applied Genomics Unit

Πρακτική Εφαρμογή

Όταν σχεδιάζεται μια μελέτη για ένα **συγκεκριμένο πρόβλημα**, ο **βασικός κύκλος της διαψευσιμότητας** παραμένει μια **πολύ καλή προσέγγιση**.

- **Οδηγός, Όχι Κανόνας:** Η μέθοδος της διαψευσιμότητας είναι ένα **πλαίσιο (approach)** για τη διεξαγωγή αυστηρής και παραγωγικής επιστημονικής έρευνας, **δεν είναι ένα αυστηρό σύνολο κανόνων**.
- **Προϋπόθεση:** Η κατανόηση αυτής της «κανονικής» διαδικασίας είναι απαραίτητη προϋπόθεση για να μπορέσει κανείς να **«σπάσει τους κανόνες»** με εποικοδομητικό τρόπο.

2. Παρατηρήσεις και Δεδομένα στην Επιστήμη

Η Σημασία της Συλλογής Δεδομένων

Η συλλογή δεδομένων, όσο **τετριμμένη** κι αν φαίνεται, είναι καθοριστικής σημασίας.

- **Δεδομένα και Θεωρία:** Τα δεδομένα που συλλέγουμε **εγείρουν ερωτήματα** και **ιδέες** για ένα σύστημα, και ταυτόχρονα μας επιτρέπουν να **ελέγξουμε** αυτές τις ιδέες.
- **Καθοριστικοί Παράγοντες:** Τα χαρακτηριστικά των δεδομένων μας (π.χ., **ποιότητα, ποσότητα, ανάλυση**) καθορίζουν:
  - Τους **τύπους ανάλυσης** που μπορούμε να πραγματοποιήσουμε.
  - Την **εμπιστοσύνη** που μπορούμε να έχουμε στα συμπεράσματα που εξάγονται.
- **Βασική Αρχή: Καμία** αναλυτική πολυπλοκότητα **δεν μπορεί να αντλήσει πληροφορίες που δεν υπάρχουν** ήδη στα δεδομένα.

Τύποι Δεδομένων και Μεταβλητές

Υπάρχουν πολλά **διαφορετικά είδη δεδομένων**, τα οποία οργανώνονται σε **στατιστικές μεταβλητές**.

Α. Είδη Δεδομένων:
Εκτός από τα απλά δεδομένα όπως **ποσότητες** (**'πόσο'**) ή **ταξινομήσεις** (**'τι είδος'**), υπάρχουν πιο πολύπλοκοι τύποι, όπως:

Χωρικοί χάρτες κατανομής ειδών.
Ακολουθίες DNA.
Δίκτυα (π.χ., διατροφικές σχέσεις [food webs] ή ρυθμιστικά δίκτυα γονιδίων).

Β. Στατιστικές Μεταβλητές (Variables):
Ορισμός: Ο όρος «μεταβλητή» (variable) χρησιμοποιείται χαλαρά από τους στατιστικολόγους για να σημαίνει **οποιοδήποτε χαρακτηριστικό** μετράται ή ελέγχεται πειραματικά σε διαφορετικά αντικείμενα.
Μορφές: Οι μεταβλητές **δεν είναι απαραίτητα αριθμητικές** (όπως το μέγεθος ενός κυττάρου), αλλά μπορεί να είναι και **μη αριθμητικές** (όπως το στάδιο ανάπτυξης).

Γ. Το Σύνολο Δεδομένων (Data Set):
Ορισμός: Ένα **σύνολο** σχετικών μεταβλητών αναφέρεται ως **σύνολο δεδομένων (data set)**.
Παράδειγμα: Σε ένα σύνολο δεδομένων για τη χωρική κατανομή ειδών, οι μεταβλητές μπορεί να περιλαμβάνουν:
Θέση **x** (συντεταγμένη).
Θέση **y** (συντεταγμένη).
Παρουσία/απουσία ενός είδους (ταξινόμηση).
Περιβαλλοντικοί παράγοντες που μετρήθηκαν.

Τύποι Μεταβλητών (Variables)

Η κατανόηση του **είδους των δεδομένων** που συλλέγουμε είναι κρίσιμη, καθώς καθορίζει τις **στατιστικές αναλύσεις** που μπορούν να εφαρμοστούν. Οι μεταβλητές ενός συνόλου δεδομένων (data set) ταξινομούνται συνήθως σε **Αριθμητικές** (Numeric) και **Κατηγορικές** (Categorical).

Βασική Ταξινόμηση

Τύπος Μεταβλητής	**Περιγραφή**	**Ερώτηση**	**Περαιτέρω Ταξινόμηση**
Αριθμητική	Περιγράφει μια **μετρήσιμη ποσότητα** με έναν αριθμό.	Πόσο; / Πόσα;	**Διαστήματος (Interval)** vs. **Λόγου (Ratio)**
Κατηγορική	Περιγράφει ένα **χαρακτηριστικό** ή μια **ιδιότητα** μιας παρατήρησης.	Τι τύπος; / Ποια κατηγορία;	**Ονομαστική (Nominal)** vs. **Διατακτική (Ordinal)**

Κατηγορικές Μεταβλητές (Categorical Variables)

Οι κατηγορικές μεταβλητές ταξινομούνται ανάλογα με το αν υπάρχει μια **φυσική σειρά** μεταξύ των κατηγοριών.

Ονομαστικές Μεταβλητές (Nominal)
Προκύπτουν όταν οι παρατηρήσεις καταγράφονται ως κατηγορίες που **δεν έχουν φυσική διάταξη** ή σειρά μεταξύ τους.
Χαρακτηριστικό: Δεν υπάρχει καμία φυσική, λογική ή ποσοτική σχέση σειράς μεταξύ των κατηγοριών.
Παραδείγματα:

- **Οικογενειακή κατάσταση:** Άγαμος, Παντρεμένος, Χήρος, Διαζευγμένος.
- **Φύλο (Sex):** Άρρεν, Θήλυ.
- **Μορφή χρώματος (Colour morph):** Κόκκινο, Κίτρινο, Μαύρο.

Διατακτικές Μεταβλητές (Ordinal)
Συμβαίνουν όταν οι παρατηρήσεις μπορούν να καταταχθούν σε μια **σημαντική (meaningful) σειρά**, αλλά η ακριβής **«απόσταση»** μεταξύ των στοιχείων **δεν είναι σταθερή** ή γνωστή

Παράδειγμα Συμπεριφοράς (Κλίμακα Aggressiveness):

Μπορούμε να πούμε ότι το σκορ **3** (ξεκινάει επίθεση) είναι **πιο επιθετικό** από το σκορ **2** (επιθετική επίδειξη), αλλά **δεν** μπορούμε να πούμε ότι το **2** είναι **διπλάσια** επιθετικό από το **1** (αγνοεί).

Συμπεριφορά	**Σκορ**
Ξεκινάει επίθεση (initiates attack)	3
Επιθετική επίδειξη (aggressive display)	2
Αγνοεί (ignores)	1
Υποχωρεί (retreats)	0

Παράδειγμα Κατάταξης (Rank Orderings):

Η **σειρά τερματισμού** σε έναν αγώνα (1η, 2η, 3η, κ.λπ.) είναι διατακτική μεταβλητή.

Μας λέει ότι ο 1ος είναι καλύτερος από τον 2ο.

Δεν μας λέει αν η διαφορά χρόνου μεταξύ του 1ου και του 2ου ήταν **ίδια** με τη διαφορά μεταξύ του 2ου και του 3ου.

Αριθμητικές Μεταβλητές (Numeric Variables)

Οι αριθμητικές μεταβλητές μετρούν ποσότητες και ταξινομούνται ανάλογα με την ύπαρξη ή μη ενός **πραγματικού μηδέν**

2.2.3 Κλίμακα Λόγου (Ratio Scale)

Οι μεταβλητές κλίμακας λόγου έχουν **πραγματικό μηδέν** (true zero) και μια **γνωστή και συνεπή μαθηματική σχέση** μεταξύ οποιωνδήποτε σημείων της κλίμακας μέτρησης.

Πραγματικό Μηδέν: Το μηδέν αντιπροσωπεύει την **παντελή απουσία** του μετρούμενου χαρακτηριστικού.

Σημαντικοί Λόγοι: Έχει νόημα να μιλάμε για **λόγους (ratios)**. Για παράδειγμα, μπορούμε να πούμε ότι τα 60 K (Κέλβιν) είναι **διπλάσια ζεστά** από τα 30 K, επειδή το 0 K (το απόλυτο μηδέν) αντιπροσωπεύει μηδενική θερμική ενέργεια.

Άλλα Παραδείγματα:

- **Μήκος** (το μηδέν σημαίνει μηδενικό μήκος).
- **Βάρος**.
- **Αριθμός αντικειμένων**.

Κλίμακα Διαστήματος (Interval Scale)

Οι μεταβλητές κλίμακας διαστήματος έχουν **συνεπή αριθμητική κλίμακα**, αλλά ξεκινούν από ένα **αυθαίρετο σημείο**.

Αυθαίρετο Μηδέν: Το μηδέν της κλίμακας είναι **τεχνητό** και δεν αντιπροσωπεύει την απουσία του χαρακτηριστικού.

Σημαντικές Διαφορές: Οι **διαφορές** μεταξύ των τιμών είναι **σημαντικές**. Π.χ., η διαφορά μεταξύ 60^0C και 70^0C είναι ίδια με τη διαφορά μεταξύ -20 και -10^0C.

Μη Σημαντικοί Λόγοι: Οι λόγοι **δεν** έχουν νόημα. Δεν μπορούμε να πούμε ότι οι 60^0C είναι **διπλάσια ζεστοί** από τους 30^0C . Αυτό γίνεται εμφανές αν σκεφτούμε ότι η ίδια θερμοκρασία μπορεί να μετρηθεί στην κλίμακα Φαρενάιτ (όπου το 0^0C είναι 32^0F).

Γιατί Είναι Σημαντική η Διάκριση; (Κλίμακα Λόγου vs. Κλίμακα Διαστήματος)

Η διάκριση μεταξύ των κλιμάκων **Λόγου (Ratio)** και **Διαστήματος (Interval)** είναι ζωτικής σημασίας διότι **επηρεάζει τις μαθηματικές πράξεις** που μπορούμε να εφαρμόσουμε σε μια αριθμητική μεταβλητή, ώστε να έχουμε ένα **ουσιαστικό (meaningful)** αποτέλεσμα.

1. Πράξεις στην Κλίμακα Διαστήματος (Interval Scale)

- **Επιτρεπόμενες Πράξεις:** Μπορούμε να κάνουμε **πρόσθεση** και **αφαίρεση**.
  - Οι **διαφορές** και οι **αποστάσεις** μεταξύ των τιμών είναι **ουσιαστικές**.
  - *Παράδειγμα:* Μπορούμε να πούμε ότι η διαφορά μεταξύ 30C και 20C είναι 10C.
- **Μη Επιτρεπόμενες Πράξεις: Δεν** μπορούμε να κάνουμε **πολλαπλασιασμό** ή **διαίρεση** (υπολογισμό λόγων) και να καταλήξουμε σε ένα ουσιαστικό αποτέλεσμα.
  Λόγος: Αυτό συμβαίνει επειδή το μηδέν είναι **αυθαίρετο** (δεν σημαίνει πλήρη απουσία του μετρούμενου χαρακτηριστικού).

2. Πράξεις στην Κλίμακα Λόγου (Ratio Scale)

Επιτρεπόμενες Πράξεις: Μπορούμε να κάνουμε **πρόσθεση, αφαίρεση, πολλαπλασιασμό** και **διαίρεση**.

Οι **διαφορές, οι αποστάσεις** και οι **λόγοι** είναι **όλες ουσιαστικές**.

Παράδειγμα: Μπορούμε να πούμε ότι τα 4 κιλά είναι **διπλάσια** σε βάρος από τα 2 κιλά, επειδή το 0 κιλά σημαίνει μηδενικό βάρος.

Ποιος Τύπος Δεδομένων Είναι ο Καλύτερος;

Όλοι οι τύποι δεδομένων είναι **χρήσιμοι**, αλλά η **καταλληλότητα** τους εξαρτάται από τις **στατιστικές αναλύσεις** που απαιτούνται και, κυρίως, από την **ερευνητική ερώτηση**.

1. Η Ιεραρχία των Δεδομένων

Γενικά, οι μεταβλητές **λόγου (ratio)** είναι οι **καλύτερα προσαρμοσμένες** για στατιστική ανάλυση, καθώς επιτρέπουν το ευρύτερο φάσμα μαθηματικών πράξεων. Ωστόσο, στην πράξη:

- **Βιολογικά Συστήματα:** Συχνά, είναι δύσκολο ή πρακτικά αδύνατο να αναπαρασταθούν βιολογικά συστήματα μόνο με δεδομένα λόγου.
- **Πόροι:** Η συλλογή δεδομένων λόγου υψηλής ποιότητας μπορεί να απαιτεί **πολύ περισσότερους πόρους** (χρόνο, χρήμα) από όσους είναι διαθέσιμοι.
- **Ερευνητικό Ερώτημα:** Το κεντρικό ερώτημα πρέπει να είναι: **«Τι είδους δεδομένα χρειαζόμαστε για να απαντήσουμε στην ερώτησή μας;»**

Εάν είναι σαφές εκ των προτέρων ότι, για παράδειγμα, τα **διατακτικά (ordinal)** δεδομένα (κλίμακα κατάταξης) δεν είναι αρκετά λεπτομερή για την απάντηση, τότε πρέπει είτε να βρεθεί ένας **καλύτερος τρόπος συλλογής** των δεδομένων είτε να **εγκαταλειφθεί** αυτή η προσέγγιση.

Μετατροπές και Απώλεια Πληροφορίας

Μετατροπή	**Δυνατή;**	**Αποτέλεσμα**
Λόγου (Ratio) →Διαστήματος (Interval)	**ΝΑΙ**	Π.χ., από Κ →C
Διαστήματος (Interval) → Λόγου (Ratio)	**ΟΧΙ**	Δεν μπορεί να προστεθεί «πραγματικό μηδέν»
Διαστήματος (Interval) →Διατακτική (Ordinal)	**ΝΑΙ**	Π.χ., από θερμοκρασία →«Ζεστό/Κρύο»
Διατακτική (Ordinal) → Διαστήματος (Interval)	**ΟΧΙ**	Δεν μπορεί να προστεθεί ακριβής απόσταση

Συνέπεια: Οι μετατροπές από μια **πιο λεπτομερή** σε μια **λιγότερο λεπτομερή** κλίμακα (π.χ., από Λόγου σε Διαστήματος, ή Διαστήματος σε Διατακτική) είναι δυνατές, αλλά **αποφεύγονται** στην πράξη, διότι αναπόφευκτα οδηγούν σε **απώλεια πληροφορίας**.

Ακρίβεια και Επαναληψιμότητα (Accuracy and Precision)

Στην καθημερινή γλώσσα, οι όροι **ακρίβεια** και **επαναληψιμότητα** (ή ακριβολογία) χρησιμοποιούνται σχεδόν ως συνώνυμα. Ωστόσο, στην **επιστημονική έρευνα** έχουν **διακριτές έννοιες**

- **Ακρίβεια (Accuracy):** Αναφέρεται στο πόσο **κοντά** βρίσκεται μια μέτρηση στην **πραγματική (αληθινή) τιμή** αυτού που προσπαθούμε να μετρήσουμε.
- **Επαναληψιμότητα / Ακριβολογία (Precision):** Αναφέρεται στο πόσο **επαναλαμβανόμενη (συνεπής)** είναι μια μέτρηση, **ανεξάρτητα** από το αν βρίσκεται κοντά στην πραγματική τιμή.

Παράδειγμα Ζύγισης (Ισορροπία)

Σενάριο	Χαρακτηριστικό	Επεξήγηση
Παλιός, κακοσυντηρημένος ζυγός (μέτρηση ±0.1g)	**Χαμηλή Επαναληψιμότητα**	Διαφορετικά αποτελέσματα κάθε φορά. Κάποια μπορεί να είναι κοντά στην αληθινή τιμή (τυχαία ακριβή), αλλά τα αποτελέσματα είναι ασυνεπή.
Νέος ηλεκτρονικός ζυγός (μέτρηση ±0.01g) με **λάθος μηδενισμό** (0.2g απόκλιση)	**Υψηλή Επαναληψιμότητα, Χαμηλή Ακρίβεια**	Επαναλαμβανόμενη ζύγιση δίνει **πανομοιότυπα** αποτελέσματα, αλλά όλα είναι **λανθασμένα** (απέχουν 0.2g από την αληθινή τιμή).

Αναλογία Σκοποβολής (Target Analogy)

Συνδυασμός	Περιγραφή
Υψηλή Ακρίβεια, Υψηλή Επαναληψιμότητα	Όλες οι βολές είναι **πολύ κοντά η μία στην άλλη** (επαναληψιμότητα) **και** συγκεντρωμένες στο **κέντρο** του στόχου (ακρίβεια).
Χαμηλή Ακρίβεια, Υψηλή Επαναληψιμότητα	Όλες οι βολές είναι **πολύ κοντά η μία στην άλλη** (επαναληψιμότητα) αλλά **μακριά** από το κέντρο του στόχου (χαμηλή ακρίβεια).
Υψηλή Ακρίβεια, Χαμηλή Επαναληψιμότητα	Οι βολές είναι **διασκορπισμένες**, αλλά ο **μέσος όρος** τους (το κέντρο της διασποράς) βρίσκεται **κοντά** στο κέντρο του στόχου.
Χαμηλή Ακρίβεια, Χαμηλή Επαναληψιμότητα	Οι βολές είναι **διασκορπισμένες** και **μακριά** από το κέντρο του στόχου.

Συνέπειες Ακρίβειας και Επαναληψιμότητας

Η κατανόηση της **ακρίβειας (accuracy)** και της **επαναληψιμότητας (precision)** των δεδομένων είναι θεμελιώδης. Όπως φαίνεται στο διάγραμμα του στόχου, ο **ιδανικός στόχος** είναι η υψηλή ακρίβεια και η υψηλή επαναληψιμότητα (πάνω αριστερά).

- **Προτεραιότητα στην Ακρίβεια:** Όταν η υψηλή επαναληψιμότητα δεν είναι εφικτή, είναι συνήθως προτιμότερο να έχουμε μετρήσεις για τις οποίες είμαστε **λογικά βέβαιοι για την ακρίβειά τους** (κάτω αριστερά), παρά πιο επαναλαμβανόμενες μετρήσεις των οποίων η ακρίβεια είναι αμφισβητήσιμη (πάνω δεξιά).
- **Μέσος Όρος:** Ο υπολογισμός του μέσου όρου των τιμών στην περίπτωση της **χαμηλής επαναληψιμότητας αλλά υψηλής ακρίβειας** (κάτω αριστερά) θα δώσει μια τιμή αρκετά κοντά στο κέντρο (την αληθινή τιμή). Αντίθετα, ο μέσος όρος στην περίπτωση της **υψηλής επαναληψιμότητας αλλά χαμηλής ακρίβειας** (πάνω δεξιά) **δεν** βελτιώνει καθόλου την ακρίβεια.

Υπονοούμενη Επαναληψιμότητα – Σημαντικά Ψηφία

Όταν δηλώνουμε ένα αποτέλεσμα, κάνουμε μια **έμμεση δήλωση** σχετικά με την **επαναληψιμότητα** της μέτρησης μέσω του αριθμού των **σημαντικών ψηφίων** (significant figures) που χρησιμοποιούμε.

- **Παράδειγμα:** Ένα αποτέλεσμα 12.375mm υποδηλώνει μεγαλύτερη επαναληψιμότητα από ένα αποτέλεσμα 12.4mm.
- **Σωστή Αναφορά:** Μια τιμή 12.4 που μετρήθηκε με την ίδια επαναληψιμότητα με το 12.735 θα έπρεπε να γραφτεί σωστά ως **12.400**.
- **Πρακτικός Κανόνας:** Όταν αναφέρετε αποτελέσματα, χρησιμοποιήστε γενικά τον **ίδιο αριθμό σημαντικών ψηφίων** με τα αρχικά δεδομένα.

Διακριτά Δεδομένα (Discrete Data)

Για **διακριτά δεδομένα** (όπως ο αριθμός των αυγών σε μια φωλιά), η επαναληψιμότητα της μέτρησης δεν αποτελεί πρόβλημα (π.χ., χρησιμοποιούμε **4**, όχι 4.0).
Μέσος Όρος: Επιτρέπεται η αναφορά του **μέσου μεγέθους** ως **4.3 αυγά**, καθώς το 4 υπονοεί 4.0.
Μεγάλοι Αριθμοί: Όταν οι αριθμοί είναι **μεγάλοι**, η επαναληψιμότητα επανέρχεται στο προσκήνιο.

300 000 μυρμήγκια: Υπονοεί επαναληψιμότητα +/- 50 000.
320 987 μυρμήγκια: Υποδηλώνει μια απίθανα ακριβή μέτρηση.

Σφάλμα, Μεροληψία και Προκατάληψη (Error, Bias and Prejudice)

Το **σφάλμα** είναι σχεδόν πάντα παρόν στα βιολογικά δεδομένα, αλλά η χειρότερη μορφή σφάλματος είναι η **μεροληψία (bias)**.

Α. Μεροληψία (Bias)

- **Ορισμός:** Είναι μια **συστηματική έλλειψη ακρίβειας**. Τα δεδομένα όχι μόνο είναι ανακριβή, αλλά **όλα αποκλίνουν από την αληθινή τιμή προς την ίδια κατεύθυνση**

- **Σημασία:** Υπάρχει σημαντική στατιστική διάκριση μεταξύ μετρήσεων που διαφέρουν από την αληθινή τιμή **τυχαία** και εκείνων που διαφέρουν **συστηματικά**.

- **Συνέπεια:** Οι μετρήσεις με κάποια έλλειψη επαναληψιμότητας (όπως η περίπτωση Γ, κάτω αριστερά) μπορούν να δώσουν μια **λογική εκτίμηση** της αληθινής τιμής αν ληφθεί ο μέσος όρος τους. Η μεροληψία **δεν** διορθώνεται με τη λήψη μέσου όρου.

Β. Τρόποι Αποφυγής της Μεροληψίας

- Η αποφυγή της μεροληψίας στη συλλογή δεδομένων είναι μία από τις πιο σημαντικές δεξιότητες στον σχεδιασμό επιστημονικών ερευνών. Ορισμένες μορφές μεροληψίας είναι:
- **Μη Τυχαία Δειγματοληψία:** Πολλές τεχνικές δειγματοληψίας είναι **επιλεκτικές**.
  - *Παράδειγμα:* Η παγίδευση αρθροπόδων ευνοεί τα **πολύ δραστήρια είδη**.
  - *Παράδειγμα:* Η μελέτη αντιδράσεων διαφυγής στο εργαστήριο μπορεί να είναι μεροληπτική αν η διαδικασία σύλληψης επέλεξε οργανισμούς με **χειρότερες** αντιδράσεις διαφυγής.
- **Εγκλιματισμός Βιολογικού Υλικού:** Οργανισμοί που διατηρούνται σε εργαστήριο για καιρό ή για πολλές γενιές μπορεί να **αλλάξουν τα χαρακτηριστικά τους** (π.χ., μέσω φυσικής επιλογής) και να δώσουν **μεροληπτική εντύπωση** της συμπεριφοράς τους στη φύση.
- **Παρέμβαση από τη Διαδικασία Έρευνας:** Η ίδια η διαδικασία μέτρησης συχνά **παραμορφώνει** το μετρούμενο χαρακτηριστικό.
  - *Παράδειγμα:* Μέτρηση αδρεναλίνης σε μικρό θηλαστικό **επηρεάζοντας** τα επίπεδα αδρεναλίνης κατά τη διαδικασία.
  - *Παράδειγμα:* Παγίδες εδάφους με αιθανόλη **προσελκύουν** είδη εντόμων που συνήθως τρέφονται με σάπια φρούτα.
- **Μεροληψία Ερευνητή (Investigator Bias):** Οι μετρήσεις μπορούν να επηρεαστούν από **συνειδητή ή ασυνείδητη προκατάληψη** του ερευνητή.
  - *Παράδειγμα:* **Στρογγυλοποίηση προς τα πάνω** τιμών σε δείγματα που περιμένετε να είναι μεγάλα ή **σκόπιμη επανάληψη** μιας τυχαίας δειγματοληψίας (quadrat) όταν δεν προσγειώνεται σε «τυπικό» βιότοπο.

3: Μάθηση από τα Δεδομένα (Μέρος Α)

Στατιστική Συχνότητας (Frequentist Statistics)

Πληθυσμοί (Populations)

Ενώ στη βιολογία ο πληθυσμός αναφέρεται σε μια ομάδα ατόμων του ίδιου είδους που αλληλεπιδρούν, στη στατιστική ο όρος έχει μια **πολύ πιο αφηρημένη έννοια**.

- **Ορισμός:** Ένας **στατιστικός πληθυσμός** είναι οποιαδήποτε ομάδα αντικειμένων ή στοιχείων που μοιράζονται **συγκεκριμένα χαρακτηριστικά ή ιδιότητες**.
- **Αντιμετώπιση στη Στατιστική Συχνότητας:** Θεωρούμε τους πληθυσμούς ως **σταθερές και άγνωστες οντότητες**.
- **Στόχος:** Ο στόχος της στατιστικής είναι να **μάθει κάτι** για αυτούς τους πληθυσμούς αναλύοντας δεδομένα που συλλέχθηκαν από αυτούς.
- **Ορισμός:** Ο **ερευνητής** είναι αυτός που **ορίζει** τον πληθυσμό, και το «κάτι που θέλουμε να μάθουμε» μπορεί να είναι οποιοδήποτε μετρήσιμο χαρακτηριστικό.
- **Παραδείγματα:**
  - Οι αναγνώστες ενός βιβλίου (φοιτητές βιολογίας, παρόμοιο εκπαιδευτικό υπόβαθρο).
  - Οι διαφορετικές περιοχές πευκοδασους στην Πελοπόνησο (παρόμοια οικολογικά χαρακτηριστικά).
  - Ένας βιολογικός πληθυσμός φυτών ή ζώων (π.χ., οι κάστορες στη Σκωτία).

Μάθηση για τους Πληθυσμούς: Τα 8 Βήματα

Ανεξάρτητα από τον τύπο του πληθυσμού ή την ερώτηση, η διαδικασία εκμάθησης από τα δεδομένα ακολουθεί θεμελιώδη κοινά βήματα.

Βήμα 1: Διευκρίνιση Ερωτημάτων, Υποθέσεων και Προβλέψεων

Ποτέ δεν πρέπει να ξεκινάμε τη συλλογή δεδομένων αν δεν έχουμε ορίσει σαφώς τις **ερωτήσεις, τις υποθέσεις** και τις **προβλέψεις** μας. Η συλλογή δεδομένων χωρίς σαφή επιστημονικό στόχο οδηγεί σε σπατάλη χρόνου και ενέργειας.

Βήμα 2: Απόφαση για τους Σχετικούς Πληθυσμούς

Πρέπει να αποφασίσουμε ποιος πληθυσμός (ή ποιοι πληθυσμοί) πρέπει να μελετηθεί.

Εργαστηριακά Πειράματα: Σε πειράματα (π.χ., προσθήκη θρεπτικών ουσιών σε λιβάδια), ο στατιστικός πληθυσμός **ορίζεται από τον πειραματικό σχεδιασμό**.

Π.χ., «Ο πληθυσμός των οικοσυστημάτων στα οποία προστέθηκαν λιπάσματα» – αυτός ο πληθυσμός **δεν υπήρχε** πριν γίνει το πείραμα.

Βασικό Σημείο: Οι στατιστικοί πληθυσμοί είναι έννοιες που **ορίζει ο ερευνητής**. Μπορεί να είναι **«πραγματικοί»** (π.χ., φοιτητές) ή **μη πραγματικοί** (π.χ., μη-πραγματοποιημένες πειραματικές εκτάσεις).

Βήμα 3: Απόφαση για τις Μεταβλητές προς Μελέτη

Πρέπει να αποφασίσουμε ποια χαρακτηριστικά (μεταβλητές) των στοιχείων του πληθυσμού θα μετρηθούν.

Παραδείγματα: Τυποποιημένη μέτρηση **πολιτικής στάσης**, μάζα **αποθηκευμένου άνθρακα** ανά μονάδα επιφάνειας, **σωματική μάζα** ατόμων.

Προσοχή: Ενώ η σωματική μάζα είναι σαφής, η μέτρηση μιας αφηρημένης έννοιας όπως η «πολιτική στάση» απαιτεί προσεκτική σκέψη (π.χ., μέτρηση και συνδυασμός πολλαπλών παραγόντων).

Μάθηση για τους Πληθυσμούς: Τα 8 Βήματα

Βήμα 4: Απόφαση για τις Σχετικές Παραμέτρους Πληθυσμού
Μια **παράμετρος πληθυσμού (population parameter)** είναι μια **αριθμητική ποσότητα** που περιγράφει τη μεταβλητή που θέλουμε να μελετήσουμε στον πληθυσμό (είναι ιδιότητα της μεταβλητής στον πληθυσμό, όχι των δεδομένων).
Παραδείγματα Παραμέτρων:

- **Μέσος Όρος Πληθυσμού (Population Mean):** Χρησιμοποιείται για να απαντήσουμε σε ερωτήσεις όπως «πόσο υπάρχει κατά μέσο όρο».
- **Διακυμάνσεις Πληθυσμού (Population Variances):** Χρησιμοποιούνται στη στατιστική γενετική για να κατανείμουν τη μεταβλητότητα.
- **Συντελεστής Συσχέτισης (Correlation Coefficient):** Χρησιμοποιείται για να κατανοήσουμε πώς **σχετίζονται** δύο πτυχές του πληθυσμού.

Βήμα 5: Συλλογή Αντιπροσωπευτικού Δείγματος
Επειδή δεν μπορούμε να μετρήσουμε κάθε μέλος του πληθυσμού, πρέπει να εργαστούμε με ένα **δείγμα (sample)** – ένα υποσύνολο του πληθυσμού.
Αντιπροσωπευτικό Δείγμα: Ένα δείγμα που έχει επιλεγεί έτσι ώστε να **αντικατοπτρίζει τα χαρακτηριστικά** ολόκληρου του πληθυσμού.
Σημασία: Η εξαγωγή χρήσιμων συμπερασμάτων για τον πληθυσμό είναι **πολύ δύσκολη** από ένα μη αντιπροσωπευτικό δείγμα.
Προσοχή: Περισσότερα δεδομένα **δεν** σημαίνουν καλύτερα δεδομένα, αν τα δεδομένα **δεν είναι αντιπροσωπευτικά** (π.χ., δειγματοληψία μόνο ηλικιωμένων ατόμων).

Μάθηση για τους Πληθυσμούς: Τα 8 Βήματα

Βήμα 6: Εκτίμηση της Παραμέτρου Πληθυσμού
Από το αντιπροσωπευτικό δείγμα, υπολογίζουμε μια **σημειακή εκτίμηση (point estimate)** της παραμέτρου πληθυσμού.
Σημειακή Εκτίμηση: Ένας αριθμός που αντιπροσωπεύει την **«καλύτερη εικασία»** μας για την άγνωστη, αληθινή τιμή της παραμέτρου (π.χ., ο μέσος όρος του δείγματος).
Παράδειγμα: Αν ενδιαφερόμαστε για τον μέσο όρο πληθυσμού, η προφανής σημειακή εκτίμηση είναι ο μέσος όρος του δείγματος.
Βήμα 7: Εκτίμηση Αβεβαιότητας
Μια σημειακή εκτίμηση **μόνη της είναι άχρηστη**, επειδή προέρχεται από ένα **περιορισμένο δείγμα**.
Αβεβαιότητα: Πρέπει πάντα να **ποσοτικοποιούμε** την αβεβαιότητα που σχετίζεται με την εκτίμησή μας. Αυτό το βήμα είναι ένα από τα πιο δύσκολα στην κατανόηση της στατιστικής συχνότητας και θα καλυφθεί αργότερα.
Βήμα 8: Απάντηση στην Ερώτηση!
Τέλος, χρησιμοποιώντας τις σημειακές εκτιμήσεις και τα μέτρα αβεβαιότητας, απαντάμε στην ερώτηση.
Γενικό Ερώτημα: Το μοτίβο που βλέπω είναι **«πραγματικό»** ή είναι απλώς αποτέλεσμα **τυχαιότητας**;
Στατιστικός Έλεγχος: Ο στατιστικός έλεγχος (όπως ο έλεγχος της μηδενικής υπόθεσης) συνδυάζει τις εκτιμήσεις και τα μέτρα αβεβαιότητας για να απαντήσει σε αυτό το ερώτημα.

Ένα Απλό Παράδειγμα (Βήματα 1-6)

Ένα φυτό έχει **δύο μορφές:** μωβ και πράσινη (πολυμορφισμός). Σε έναν γειτονικό πληθυσμό, η συχνότητα της μωβ μορφής είναι **25%** (ισορροπία Hardy-Weinberg). Υποθέτουμε ότι η μωβ μορφή έχει πλεονέκτημα στον νέο πληθυσμό

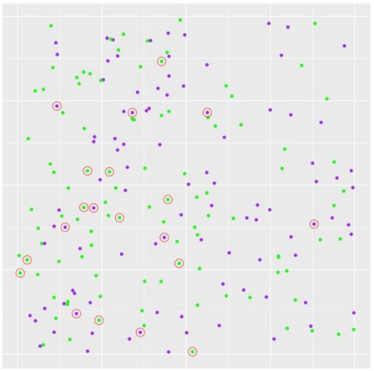
Βήμα	**Εφαρμογή στο Παράδειγμα**	**Αποτέλεσμα**
1. Ερώτημα/Πρόβλεψη	**Υπόθεση:** Η μωβ μορφή έχει πλεονέκτημα. **Πρόβλεψη:** Η συχνότητα της μωβ μορφής στον νέο πληθυσμό θα είναι **μεγαλύτερη από 25%**.	Πρόβλεψη > 25%
2. Πληθυσμός	Ο πληθυσμός όλων των φυτών στο εξεταζόμενο τοπίο.	Όλα τα φυτά
3. Μεταβλητή	Το **χρώμα** του φυτού (Μωβ ή Πράσινο).	Κατηγορική, Ονομαστική
4. Παράμετρος	Η **συχνότητα** (ποσοστό ή αναλογία) της μωβ μορφής στον **ευρύτερο πληθυσμό**.	Συχνότητα Μωβ
5. Δείγμα	Συλλέγουμε ένα **τυχαίο δείγμα 20 φυτών** από το τοπίο.	n=20
6. Εκτίμηση	Βρίσκουμε 12 πράσινα και 8 μωβ φυτά. **Σημειακή Εκτίμηση:** 8/20 = 40%.	Εκτίμηση = 40%

Και Τώρα Τι;

Η σημειακή μας εκτίμηση **40%** είναι **μεγαλύτερη** από το αναμενόμενο **25%**. Ωστόσο, επειδή το δείγμα μας είναι μικρό, αυτό το αποτέλεσμα **μπορεί να οφείλεται στην τύχη**. Το επόμενο βήμα είναι η εκτίμηση της **αβεβαιότητας** στην εκτίμηση του 40% (Βήμα 7: Σφάλμα Δειγματοληψίας), η οποία θα μας επιτρέψει να αποφασίσουμε αν η παρατηρούμενη διαφορά είναι «πραγματική».

Σφάλμα Δειγματοληψίας (Sampling Error)

Το Σφάλμα Δειγματοληψίας είναι η **φυσική μεταβλητότητα (διακύμανση)** στις εκτιμήσεις που προκύπτουν από δείγματα, ακόμα κι όταν η διαδικασία δειγματοληψίας είναι άψογη.



1. Παρατηρούμενη Διακύμανση και Ορισμός
Συνεχίζοντας το παράδειγμα του πολυμορφισμού των φυτών, παρατηρήσαμε ότι:
1ο Δείγμα (n=20): 8 μωβ φυτά → **40%** συχνότητα μωβ μορφής.
2ο Δείγμα (n=20): 9 μωβ φυτά → **45%** συχνότητα μωβ μορφής.

Η **διαφορά** μεταξύ αυτών των εκτιμήσεων (από 40% σε 45%) προέκυψε παρόλο που:
Ο **πληθυσμός** παρέμεινε **αμετάβλητος**.
Το σχέδιο δειγματοληψίας ήταν **πλήρως αξιόπιστο** και μη μεροληπτικό.

Αυτή η διακύμανση, η οποία οφείλεται αποκλειστικά στη **τυχαιότητα (chance variation)** της δειγματοληψίας, ονομάζεται **Σφάλμα Δειγματοληψίας (Sampling Error)** ή **Διακύμανση Δειγματοληψίας (Sampling Variation)**.

Σφάλμα Δειγματοληψίας (Sampling Error)

- **Έννοια του «Σφάλματος»:** Στη στατιστική, η λέξη «σφάλμα» (**error**) δεν σημαίνει **λάθος** ή **πρόβλημα** στον σχεδιασμό. Αναφέρεται στη **διακύμανση** ή τον **«θόρυβο» (noise)** στη διαδικασία.
- **Η Αναγκαιότητα της Στατιστικής:** Το Σφάλμα Δειγματοληψίας είναι ο **κύριος λόγος** για τον οποίο χρειαζόμαστε τη στατιστική. Αν αγνοήσουμε αυτόν τον «θόρυβο» που υπάρχει σε κάθε εκτίμηση, κινδυνεύουμε να εξάγουμε **λανθασμένα συμπεράσματα**.

2. Από την Εκτίμηση στην Κατανομή

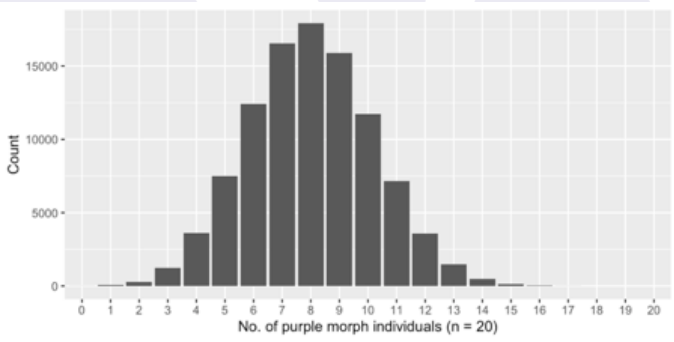
Είναι σημαντικό να συνειδητοποιήσουμε ότι το Σφάλμα Δειγματοληψίας **δεν** είναι ιδιότητα ενός **συγκεκριμένου δείγματος**, αλλά συνέπεια:

- Της φύσης της **μεταβλητής** που μελετάμε.
- Της **μεθόδου δειγματοληψίας** που χρησιμοποιείται.

Για να κατανοήσουμε το μέγεθος και τη φύση αυτού του σφάλματος, πρέπει να εξετάσουμε την έννοια της **Κατανομής Δειγματοληψίας (Sampling Distribution)**.

4.2 Κατανομές Δειγματοληψίας (Sampling Distributions)

Η **Κατανομή Δειγματοληψίας** είναι μια κεντρική έννοια στη Στατιστική Συχνότητας και είναι το εργαλείο που χρησιμοποιείται για την **ποσοτικοποίηση του Σφάλματος Δειγματοληψίας**.



Ορισμός και Δημιουργία

- **Διαδικασία:** Για να διερευνήσουμε το Σφάλμα Δειγματοληψίας, μπορούμε να προσομοιώσουμε τη λήψη **χιλιάδων ανεξάρτητων δειγμάτων** (στην περίπτωση αυτή, 100 000) του **ιδίου μεγέθους** (n=20) από τον **ίδιο πληθυσμό**.
- **Αποτέλεσμα:** Η **Κατανομή Δειγματοληψίας** είναι μια γραφική παράσταση (π.χ., ραβδόγραμμα) που συνοψίζει το **εύρος των αποτελεσμάτων** (π.χ., τον αριθμό των μωβ μορφών) που αναμένεται να παρατηρηθούν όταν επαναλαμβάνουμε την ίδια διαδικασία δειγματοληψίας ξανά και ξανά.
- **Σημασία:** Οποιαδήποτε διακύμανση παρατηρείται σε αυτή την κατανομή οφείλεται αποκλειστικά στο **Σφάλμα Δειγματοληψίας**, καθώς η παράμετρος του πληθυσμού παραμένει σταθερή.

Ανάλυση της Κατανομής Δειγματοληψίας

- Για το συγκεκριμένο παράδειγμα (n=20, συχνότητα μωβ μορφής στον πληθυσμό = 40%), η προσομοίωση έδειξε:
- **Πιο Συνηθισμένο Αποτέλεσμα (Mode):** Ο πιο συχνός αριθμός μωβ μορφών είναι **οκτώ (8)**, ο οποίος αντιστοιχεί ακριβώς στην **αληθινή τιμή** της παραμέτρου του πληθυσμού 8/20 = 40%).
- **Μέση Τιμή:** Η αληθινή τιμή της παραμέτρου (40%) τείνει να είναι η **πιο κοινή εκτίμηση** που αναμένεται να παρατηρηθεί υπό επαναλαμβανόμενη δειγματοληψία.
- **Περιορισμός στην Πραγματική Ζωή:** Στον πραγματικό κόσμο, οι επιστήμονες εργάζονται με **μόνο ένα δείγμα**. Η γνώση της Κατανομής Δειγματοληψίας είναι το κλειδί για να χρησιμοποιήσουμε αυτό το ένα δείγμα για να εκτιμήσουμε την άγνωστη παράμετρο.

Χρησιμότητα και Εύρος (Spread)

- Η Κατανομή Δειγματοληψίας είναι ζωτικής σημασίας για τη Στατιστική Συχνότητας διότι:
- **Ποσοτικοποιεί την Αβεβαιότητα:** Επιτρέπει τη μέτρηση της **αβεβαιότητας** (Σφάλμα Δειγματοληψίας) που περιέχεται στην εκτίμησή μας.
- **Εύρος Αποτελεσμάτων:** Η κατανομή δείχνει το **εύρος των αναμενόμενων αποτελεσμάτων**. Στο παράδειγμα, το εύρος των μωβ μορφών είναι περίπου 2 έως 15, που αντιστοιχεί σε εκτιμώμενες συχνότητες **10% έως 75%**.
  - **Σημασία:** Αυτό το ευρύ εύρος αποδεικνύει ότι, όταν λαμβάνουμε μόνο 20 άτομα, το **Σφάλμα Δειγματοληψίας** για την εκτίμηση συχνότητας μπορεί να είναι **αρκετά μεγάλο**.

Εξάρτηση της Κατανομής

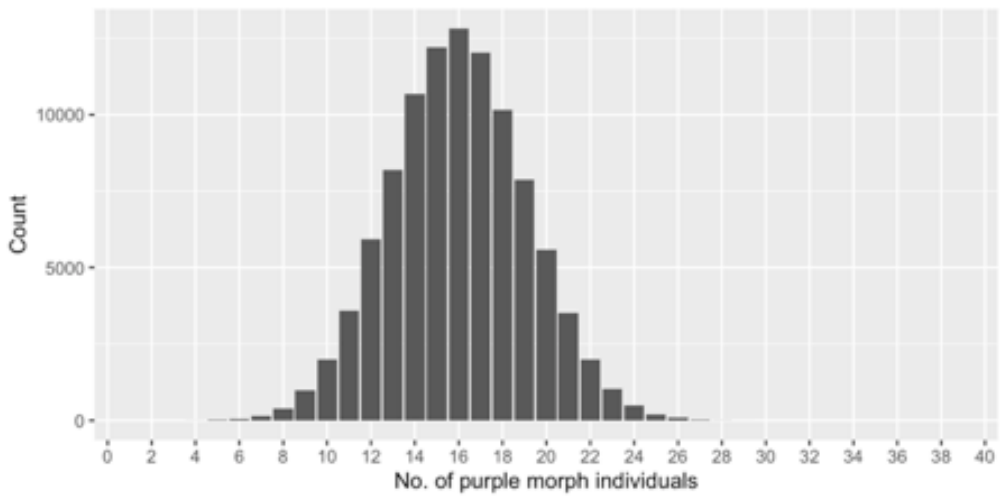
Η Κατανομή Δειγματοληψίας **δεν είναι σταθερή**. Αλλάζει εάν αλλάξει οποιοσδήποτε από τους δύο ακόλουθους παράγοντες:

- **Το Μέγεθος του Δείγματος (n):** Εάν το n άλλαζε από 20 σε 100, η κατανομή θα ήταν διαφορετική.
- **Η Παράμετρος του Πληθυσμού:** Εάν η αληθινή συχνότητα μωβ φυτών στον πληθυσμό ήταν 60% αντί για 40%, η κατανομή θα ήταν διαφορετική.
- Αυτό επιβεβαιώνει ότι το Σφάλμα Δειγματοληψίας είναι συνέπεια της **φύσης της μεταβλητής** και της **μεθόδου δειγματοληψίας** που χρησιμοποιείται.
- **Προοπτική:** Δεν χρειάζεται να υπολογίζουμε εμείς οι ίδιοι όλες αυτές τις λεπτομέρειες. Οι στατιστικολόγοι έχουν ήδη κάνει τη σκληρή δουλειά για να κατασκευάσουν τις Κατανομές Δειγματοληψίας για μια πληθώρα προβλημάτων, επιτρέποντάς μας να τις χρησιμοποιήσουμε για να ποσοτικοποιήσουμε την αβεβαιότητα και να απαντήσουμε στα επιστημονικά μας ερωτήματα.

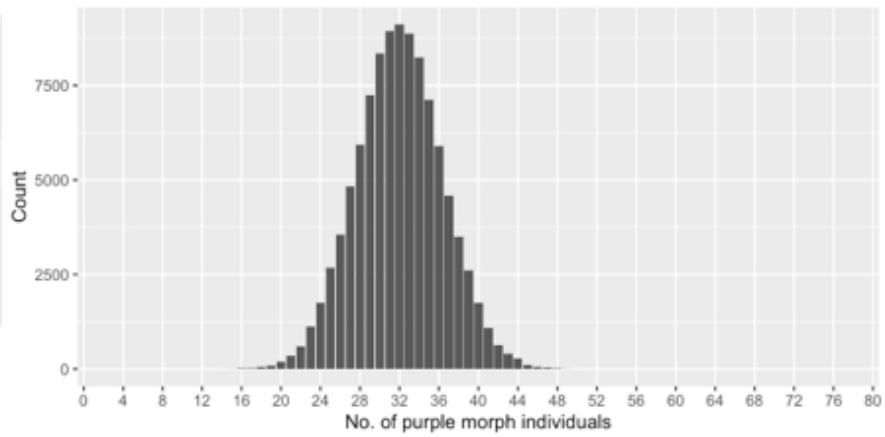
Η Επίδραση του Μεγέθους Δείγματος (n)

- Το μέγεθος του δείγματος (n) είναι ένας από τους πιο σημαντικούς παράγοντες σε ένα σχέδιο δειγματοληψίας, καθώς καθορίζει άμεσα το Σφάλμα Δειγματοληψίας (ή Διακύμανση Δειγματοληψίας). 1. Σύγκριση Κατανομών Δειγματοληψίας
- Για να κατανοήσουμε πώς επηρεάζει το n, το πείραμα προσομοίωσης επαναλαμβάνεται με δύο μεγαλύτερα μεγέθη δείγματος (n=40 και n=80) από τον ίδιο πληθυσμό (με αληθινή συχνότητα μωβ μορφής 40%).

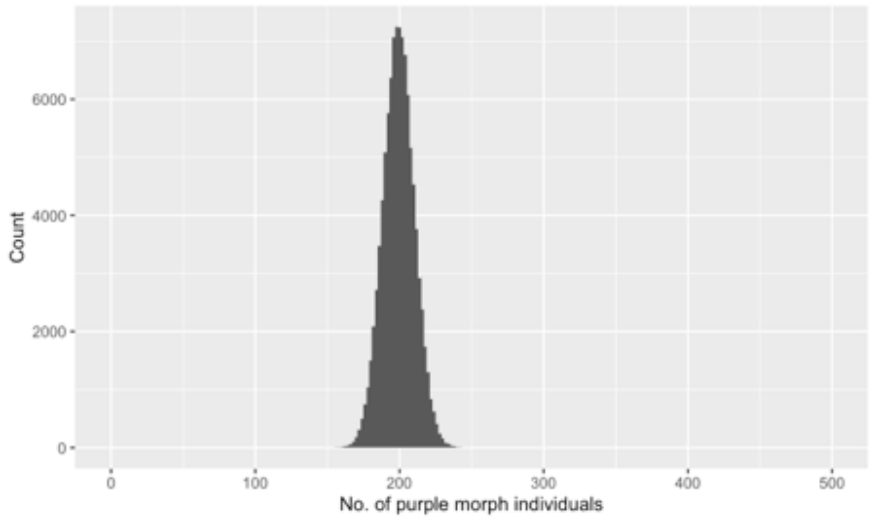
Μέγεθος Δείγματος (n)	Εύρος Αποτελεσμάτων (Αριθμός Μωβ)	Εκτιμώμενο Εύρος Συχνότητας (%)	Εύρος Σφάλματος Δειγματοληψίας
20	2 έως 15	10% – 75%	Πολύ Μεγάλο
40	6 έως 26	15% – 65%	Μεγάλο
80	16 έως 48	20% – 60%	Μικρότερο



Distribution of number of purple morphs sampled (n = 40)



Distribution of number of purple morphs sampled (n = 80)



Distribution of number of purple morphs sampled (n = 500)

Συμπέρασμα: Με την **αύξηση του μεγέθους του δείγματος** (n), το εύρος των αποτελεσμάτων (το **άπλωμα** της κατανομής) **μειώνεται**.

Αυτό σημαίνει ότι **μειώνουμε το Σφάλμα Δειγματοληψίας**.

Αυτό είναι διαισθητικά λογικό: η σύνθεση ενός **μεγάλου δείγματος** πρέπει να **προσεγγίζει στενότερα** τη σύνθεση του πραγματικού πληθυσμού σε σχέση με ένα μικρό δείγμα.

Απαιτήσεις για Ακρίβεια

Για να εκτιμηθεί με ακρίβεια μια συχνότητα, απαιτούνται συχνά **πολύ μεγάλα δείγματα**.

- **Δείγμα n=500:** Όταν το μέγεθος του δείγματος αυξάνεται σε n=500, το εύρος των αποτελεσμάτων περιορίζεται σε 160 έως 240 μωβ μορφές.
- **Εύρος Συχνότητας n=500:** Αυτό αντιστοιχεί σε εκτιμώμενες συχνότητες **32% έως 48%**.

Παρόλο που η βελτίωση είναι μεγάλη σε σύγκριση με τα μικρότερα δείγματα, ακόμα και με 500 άτομα υπάρχει **αρκετή αβεβαιότητα** (το εύρος είναι 16 ποσοστιαίες μονάδες).

Το Βασικό Συμπέρασμα:

- Χρειαζόμαστε **πολλά δεδομένα** για να μειώσουμε σημαντικά το **Σφάλμα Δειγματοληψίας** και να έχουμε μια ακριβή εκτίμηση της παραμέτρου του πληθυσμού.

Το Τυπικό Σφάλμα (Standard Error – SE)

Για να κάνουμε αυστηρές συγκρίσεις και να ποσοτικοποιήσουμε τη διασπορά της *Κατανομής Δειγματοληψίας*, χρειαζόμαστε ένα ποσοτικό μέτρο της μεταβλητότητας του σφάλματος δειγματοληψίας. Αυτό το μέτρο ονομάζεται **Τυπικό Σφάλμα (Standard Error – SE)**.

Ορισμός:

Το **Τυπικό Σφάλμα (SE)** μιας εκτίμησης είναι η **Τυπική Απόκλιση (Standard Deviation)** της **Κατανομής Δειγματοληψίας** της εκτίμησης.

Περιγραφή:

Το κεντρικό σημείο είναι ότι το SE είναι μέτρο του εύρους ή της διασποράς μιας κατανομής, συγκεκριμένα της *Κατανομής Δειγματοληψίας*.

Συντομογραφίες:

Χρησιμοποιούνται συχνά οι συντομογραφίες **SE** ή **s.e.**

Υπολογισμός του Τυπικού Σφάλματος με Προσομοίωση στην R

Για να κατανοήσουμε το SE, μπορούμε να το υπολογίσουμε με προσομοίωση στο R. Στόχος: υπολογισμός του αναμενόμενου SE για την εκτίμηση της συχνότητας της μωβ μορφής όταν:

- πλήθος δειγμάτων: 100000
- μέγεθος δείγματος: n=80
- αληθινή συχνότητα στον πληθυσμό: 40%

Καθορισμός μεταβλητών

```
# number of independent samples
nsamples <- 100000

# sample size to use for each sample
samplesize <- 80

# probability a plant will be purple
purpleprob <- 0.4
```

Εκτέλεση προσομοίωσης:

```
# repeat the sampling/estimation procedure many times
rawsamples <- rbinom(n = nsamples, size = samplesize, prob = purpleprob)
# convert results to %
percentsamples <- 100 * rawsamples / samplesize
```

Υπολογισμός του Τυπικού Σφάλματος με Προσομοίωση στην R

Η μεταβλητή percentsamples περιέχει 10^5 εκτιμήσεις συχνότητας (ως ποσοστά) που αποτελούν την Κατανομή Δειγματοληψίας. Οι πρώτες 50 τιμές έχουν ως εξής:

```
## [1] 33.75 35.00 42.50 41.25 46.25 35.00 42.50 45.00 43.75 40.00 41.25 52.50
## [13] 43.75 38.75 38.75 47.50 42.50 40.00 43.75 42.50 36.25 42.50 43.75 43.75
## [25] 35.00 42.50 35.00 37.50 41.25 40.00 45.00 40.00 41.25 38.75 46.25 45.00
## [37] 31.25 40.00 47.50 42.50 37.50 41.25 31.25 45.00 32.50 36.25 40.00 37.50
## [49] 40.00 45.00
```

Υπολογισμός του SE:

```
sd(percentsamples)
## [1] 5.494663
```

Αποτέλεσμα:

Το Τυπικό Σφάλμα (SE) είναι περίπου **5.5 ποσοστιαίες μονάδες**.

Χρησιμότητα του Τυπικού Σφάλματος

Το SE παρέχει ένα μέσο ποσοτικοποίησης της αβεβαιότητας μιας εκτίμησης.

Ποσοτικοποίηση μεταβλητότητας

Το SE ≈ 5.5 ποσοστιαίες μονάδες δείχνει πόση διασπορά αναμένεται στις εκτιμήσεις της συχνότητας της μωβ μορφής όταν δειγματοληπτούμε n=80 άτομα.

Εύρος εκτιμήσεων

Για "καλά συμπεριφερόμενες" κατανομές, περίπου το **95%** των εκτιμήσεων αναμένεται να βρίσκεται εντός **± 2 SE** (δηλαδή σε ένα εύρος περίπου 4 SE συνολικά).

Αυτό το εύρος μάς βοηθά να κατανοήσουμε πόσο αξιόπιστη ή ασταθής είναι μια σημειακή εκτίμηση.

Συμπέρασμα:

Το Τυπικό Σφάλμα είναι ένα κρίσιμο μέτρο αβεβαιότητας και πρέπει πάντα να συνοδεύει οποιαδήποτε σημειακή εκτίμηση

Ο Στόχος της Στατιστικής Συχνότητας

Ο σκοπός της μελέτης των **Κατανομών Δειγματοληψίας** και του **Τυπικού Σφάλματος (SE)** είναι να οικοδομηθεί η εννοιολογική βάση για τους στατιστικούς ελέγχους.
Πριν χρησιμοποιήσουμε τη Στατιστική Συχνότητας, πρέπει να κατανοήσουμε:

- **Πώς συμπεριφέρονται οι σημειακές εκτιμήσεις** υπό επαναλαμβανόμενη δειγματοληψία (δηλαδή, τις **Κατανομές Δειγματοληψίας**).

- **Πώς σχετίζονται** το **Σφάλμα Δειγματοληψίας** και το **Τυπικό Σφάλμα** με αυτές τις κατανομές.

Αυτές οι ιδέες είναι απαραίτητες γιατί οι **στατιστικοί έλεγχοι** (όπως ο υπολογισμός των **p-values**) βασίζονται στο να γνωρίζουμε **πόσο πιθανό** είναι το αποτέλεσμά μας να έχει προκύψει τυχαία, λαμβάνοντας υπόψη τη φυσική μεταβλητότητα που ορίζεται από το **SE**.

Στατιστική Σημαντικότητα και p-values

Η Στατιστική Συχνότητας απαντά στο εξής ερώτημα: **Τι θα συνέβαινε αν επαναλαμβάναμε** τη διαδικασία συλλογής δεδομένων πολλές φορές, **υποθέτοντας ότι ο πληθυσμός παραμένει ο ίδιος** κάθε φορά; Αυτή η ιδέα μας οδηγεί στη δημιουργία **Κατανομών Δειγματοληψίας**, οι οποίες αποτελούν τη βάση για τον υπολογισμό των p-values και της στατιστικής σημαντικότητας.

Εκτίμηση μιας Κατανομής Δειγματοληψίας

Ο τελικός μας στόχος είναι να διαπιστώσουμε αν η συχνότητα της μωβ μορφής στον νέο πληθυσμό είναι **πιθανό να είναι μεγαλύτερη από 25%**. Για να το κάνουμε αυτό, χρειαζόμαστε την Κατανομή Δειγματοληψίας της εκτίμησης της συχνότητας.

Επειδή στην πραγματικότητα έχουμε πρόσβαση μόνο σε **ένα δείγμα**, η λύση είναι να χρησιμοποιήσουμε αυτό το μοναδικό δείγμα για να **προσεγγίσουμε** τον πληθυσμό και στη συνέχεια να βρούμε την αναμενόμενη Κατανομή Δειγματοληψίας.

Επισκόπηση της Μεθόδου Bootstrapping (Επαναδειγματοληψία)

Η μέθοδος **bootstrapping** είναι ένας τρόπος χρήσης ενός δείγματος για την προσέγγιση του πληθυσμού.

Η Ιδέα: Υποθέτουμε ότι το δείγμα είναι ο αληθινός πληθυσμός. Στη συνέχεια, αντλούμε νέα δείγματα από αυτόν τον υποτιθέμενο πληθυσμό για να δημιουργήσουμε μια Κατανομή Δειγματοληψίας.

Αναλογία με την Κάλπη (The Hat Analogy)

Ας φανταστούμε ότι έχουμε γράψει το χρώμα κάθε φυτού του πραγματικού μας δείγματος (n=20) σε ένα ξεχωριστό κομμάτι χαρτί και τα έχουμε βάλει σε μια κάλπη (καπέλο).

Βήμα	Διαδικασία Bootstrapping
1.	Διαλέγουμε ένα κομμάτι χαρτί **τυχαία**.
2.	Καταγράφουμε την τιμή του (μωβ ή πράσινο) και το **τοποθετούμε πίσω** στην κάλπη. (Αυτό ονομάζεται **«δειγματοληψία με αντικατάσταση»**).
3.	Ανακινούμε την κάλπη.
4.	Επαναλαμβάνουμε τη διαδικασία μέχρι να έχουμε ένα νέο τεχνητό δείγμα, **ίδιου μεγέθους** με το αρχικό (π.χ., n=20). Αυτό είναι ένα **«bootstrapped δείγμα»**.
5.	Για κάθε bootstrapped δείγμα, υπολογίζουμε την ποσότητα που μας ενδιαφέρει (την αναλογία μωβ φυτών).
6.	Επαναλαμβάνουμε τα Βήματα 1-5 περίπου **10 000 φορές** για να δημιουργήσουμε την **Κατανομή Bootstrapping**.

Σημασία: Παρόλο που φαίνεται σαν «απάτη», αυτή η διαδικασία **προσεγγίζει** την πραγματική Κατανομή Δειγματοληψίας της συχνότητας της μωβ μορφής. Η μέθοδος αναπτύχθηκε από τον στατιστικολόγο **Bradley Efron**.

Φόρτωμα και κατανόηση των δεδομένων

Ας υποθέσουμε ότι έχουμε πραγματοποιήσει δειγματοληψία 250 φυτών από τον νέο πληθυσμό. Τα δεδομένα βρίσκονται σε ένα αρχείο CSV με όνομα **MORPHDATA.CSV**, και έχουν φορτωθεί σε ένα tibble με το όνομα `morphdata`

π.χ.:

```
morphdata <- read.csv(file = "MORPHDATA.CSV")
```

Το dataset έχει 250 γραμμές και δύο μεταβλητές:

- **Colour**: κατηγορική (Green / Purple)
- **Weight**: αριθμητική (το βάρος κάθε φυτού)

Εκκίνηση του bootstrap

Θέλουμε να δημιουργήσουμε μια προσεγγιστική **κατανομή δειγματοληψίας** για τη συχνότητα της μωβ μορφής. Το μόνο που μας χρειάζεται από το dataset είναι η μεταβλητή Colour, οπότε την εξάγουμε:

```
plant_morphs <- morphdata$Colour


# what is the set of values 'plant_morphs' can take?
unique(plant_morphs)

# show the first 50 values

head(plant_morphs, 50)
```

Υπολογίζουμε το μέγεθος δείγματος και τη σημειακή εκτίμηση (ποσοστό μωβ φυτών):

```
samp_size <- length(plant_morphs)
samp_size
mean_point_est <- 100 * sum(plant_morphs == "Purple") / samp_size
```

Η εκτίμηση είναι **38 % μωβ φυτά**.

Παραγωγή bootstrap δειγμάτων

Θα δημιουργήσουμε **10.000** bootstrap δείγματα:

```
n_samp <- 10000
```

Το βασικό εργαλείο είναι η συνάρτηση `sample()`, η οποία επιτρέπει δειγματοληψία *με επανάθεση*:

```
samp <- sample(plant_morphs, replace = TRUE)
first_bs_freq <- 100 * sum(samp == "Purple") / samp_size
```

Αυτό μάς δίνει *μία* bootstrap εκτίμηση συχνότητας. Αλλά χρειαζόμαστε χιλιάδες. Χρησιμοποιούμε την `replicate()` για να επαναλάβουμε αυτόματα τη διαδικασία:

```
boot_out <- replicate(n_samp, {
samp <- sample(plant_morphs, replace = TRUE)
100 * sum(samp == "Purple") / samp_size
})
```

Τώρα το `bootout` περιέχει **10.000 εκτιμήσεις** της συχνότητας μωβ φυτών.

```
head(boot_out, 30) %>% round(1)
```

Κατανόηση της bootstrap κατανομής

Οι τιμές στο `bootout` αντιπροσωπεύουν τι συχνότητες μωβ φυτών θα βλέπαμε αν μπορούσαμε να επαναλάβουμε τη δειγματοληψία άπειρες φορές. Είναι δηλαδή μια **bootstrapped κατανομή δειγματοληψίας**.
Ο μέσος της κατανομής είναι:

```
mean(boot_out) %>% round(2)     # ≈ 37.97 %
```

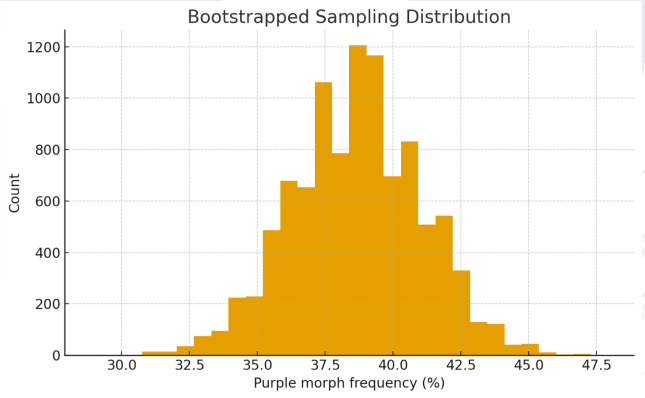
όπως και η αρχική σημειακή εκτίμηση — κάτι αναμενόμενο, αφού επαναδειγματοληπτούμε τα ίδια δεδομένα.
Το πιο χρήσιμο μέγεθος είναι το **τυπικό σφάλμα (SE)**:

```
sd(boot_out) %>% round(2)     # ≈ 3.08
```

Το SE μάς δείχνει πόσο ακριβής είναι η εκτίμησή μας.
Μικρό SE σημαίνει ότι η σημειακή εκτίμηση είναι σταθερή και αξιόπιστη.

Πώς αναφέρουμε τα αποτελέσματα

Όταν αναφέρουμε μια σημειακή εκτίμηση πρέπει πάντα να συνοδεύεται από το τυπικό σφάλμα της:
Η συχνότητα της μωβ μορφής (n = 250) ήταν 37.97 % (s.e. ± 3.08).
Το μέγεθος δείγματος αναφέρεται επίσης, γιατί είναι κρίσιμο για την ερμηνεία.



Στατιστική Σημαντικότητα και *p-value*

Το κεντρικό ερώτημα είναι: **Είναι η συχνότητα της μωβ μορφής μεγαλύτερη από 25% στον νέο πληθυσμό μελέτης;** Η απάντηση σε αυτή την ερώτηση δεν μπορεί να είναι οριστική από ένα δείγμα, αλλά πρέπει να είναι μια **πιθανολογική εκτίμηση**.

Διεξαγωγή της Αξιολόγησης

Για να καταλήξουμε στην πιθανολογική εκτίμηση, κάνουμε δύο βασικές υποθέσεις:
1. Υπόθεση Μηδενικής Διαφοράς: Υποθέτουμε ότι η αληθινή τιμή της συχνότητας της μωβ μορφής στον νέο πληθυσμό είναι **25%**. Αυτή είναι η **Μηδενική Υπόθεση (H_0)** — η υπόθεση **«καμίας επίδρασης»** ή **«καμίας διαφοράς»**.
2. Υπόθεση Σταθερής Κατανομής Δειγματοληψίας: Υποθέτουμε ότι το **αναμενόμενο σχήμα** της Κατανομής Δειγματοληψίας θα παρέμενε το ίδιο ακόμα και αν η H_0 ήταν αληθής.

Αν η συχνότητα του πληθυσμού είναι πράγματι 25%, πώς θα έμοιαζε η αντίστοιχη Κατανομή Δειγματοληψίας;
Αυτή η κατανομή ονομάζεται **Μηδενική Κατανομή (Null Distribution)**.

Κατασκευή της Μηδενικής Κατανομής

Χρησιμοποιώντας τη μέθοδο Bootstrapping, η Μηδενική Κατανομή (null dist) κατασκευάζεται μετατοπίζοντας την bootstrapped κατανομή, ώστε ο μέσος της να βρίσκεται στο 25%

```
null_dist <- boot_out - mean(boot_out) + 25
```


Οπτική Αξιολόγηση και *p-value*

Στη συνέχεια, συγκρίνουμε την **παρατηρούμενη εκτίμηση** (π.χ., 40% στο παράδειγμά μας) με τη **Μηδενική Κατανομή**.

Αξιολόγηση: Επειδή η παρατηρούμενη συχνότητα βρίσκεται στην **άκρη της «ουράς»** της Μηδενικής Κατανομής, συμπεραίνουμε ότι είναι **μάλλον απίθανο** να έχει προκύψει από τυχαία διακύμανση αν η συχνότητα του πληθυσμού ήταν πραγματικά 25%.

Ποσοτικοποίηση του p-value: Για μια πιο ακριβή δήλωση, ποσοτικοποιούμε πόσο συχνά οι τιμές της Μηδενικής Κατανομής ήταν **μεγαλύτερες** από την παρατηρούμενη εκτίμηση (meanpointest):

```
p_value <- sum(null_dist > mean_point_est) / n_samp
p_value
```

```
## [1] 0
```

Αυτός ο αριθμός ονομάζεται **p-value**.

Ερμηνεία του *p-value*

Το p-value είναι η **πιθανότητα** να αποκτήσουμε ένα αποτέλεσμα **ίσο ή πιο ακραίο** από αυτό που παρατηρήθηκε, **υποθέτοντας ότι η Μηδενική Υπόθεση (H_0) είναι αληθής**.

- **Χαμηλό p-value:** Δεδομένου ότι η H_0 είναι η υπόθεση «καμίας επίδρασης», ένα χαμηλό p-value ερμηνεύεται ως **απόδειξη ότι υπάρχει επίδραση**.
- **Βιολογική Ερμηνεία:** Στο παράδειγμά μας, το *p* υποδηλώνει ότι είναι **μάλλον απίθανο** να παρατηρηθεί η συχνότητά μας αν η αληθινή συχνότητα ήταν 25%. Επομένως, τα δεδομένα υποστηρίζουν την πρόβλεψη ότι η συχνότητα της μωβ μορφής είναι **μεγαλύτερη από 25%** στον νέο πληθυσμό.

Επίπεδο Σημαντικότητας (Significance Level)

Για να αποφασίσουμε πότε ένα p-value είναι αρκετά μικρό, εφαρμόζουμε ένα **κατώφλι** που ονομάζεται **Επίπεδο Σημαντικότητας**.

- **Συνήθης Σύμβαση:** Πολύ συχνά χρησιμοποιείται το κατώφλι του $p < 0.05$ (ή 5%).
- **Στατιστικά Σημαντικό:** Αν το p-value είναι μικρότερο από το επιλεγμένο επίπεδο σημαντικότητας (π.χ., $0 < 0.05$), το αποτέλεσμα λέγεται ότι είναι **στατιστικά σημαντικό**.
- **Προσοχή:** Το όριο του 5% είναι απλώς μια **σύμβαση** και **δεν** πρέπει να συγχέεται με τη **βιολογική σημασία**.

Συμπεράσματα

Η διαδικασία που εφαρμόσαμε ονομάζεται **Έλεγχος Σημαντικότητας (Significance Test)** και αποτελεί τη βάση όλων των ελέγχων στη Στατιστική Συχνότητας.
Η αλυσίδα συλλογισμών είναι:

1. **Υποθέτουμε** ότι δεν υπάρχει **«επίδραση»** (H_0).
2. **Κατασκευάζουμε** τη **Μηδενική Κατανομή** (null distribution), η οποία δείχνει τι θα συνέβαινε με συχνή δειγματοληψία υπό την H_0.
3. **Συγκρίνουμε** την εκτίμηση του δείγματός μας με τη Μηδενική Κατανομή για να βρούμε το **p-value** (τη συχνότητα με την οποία θα παρατηρούνταν το αποτέλεσμά μας ή ένα πιο ακραίο αποτέλεσμα υπό την H0).

Αν και χρησιμοποιήσαμε το **bootstrapping** για να κάνουμε αυτή τη διαδικασία «πραγματική», στο εξής θα χρησιμοποιούμε έτοιμα στατιστικά εργαλεία. Το κλειδί είναι να μπορούμε να **αναγνωρίζουμε τη Μηδενική Υπόθεση** και να **ερμηνεύουμε το σχετικό p-value**.