

R - Programming in Biosciences

Dr. Timokratis Karamitros
Researcher

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

tkaram@pasteur.gr



Στατιστικές Συγκρίσεις:
Εισαγωγή στη Στατιστική Ελέγχου
(Test Statistic)

Διενέργεια Συγκρίσεων

Η επιστημονική διερεύνηση συχνά απαιτεί να εκτιμήσουμε αν **δύο στατιστικοί πληθυσμοί** διαφέρουν σε κάποια ποσοτική μεταβλητή (π.χ., μέση δραστηριότητα ενζύμου, μέση βιομάζα)

Συνεχίζουμε με το παράδειγμα των μωβ και πράσινων φυτών, αλλά αυτή τη φορά εξετάζουμε τη **διαφορά στο μέγεθος (βιομάζα)**:

- **Ερώτημα & Υπόθεση:** Υποθέτουμε ότι τα μωβ φυτά είναι **πιο εύρωστα** (fitter) και, κατά συνέπεια, **μεγαλύτερα** από τα πράσινα.
- **Στατιστικοί Πληθυσμοί:** Ορίζουμε δύο ξεχωριστούς πληθυσμούς: τον **Πληθυσμό Μωβ Φυτών** και τον **Πληθυσμό Πράσινων Φυτών**.
- **Μεταβλητή:** Χρησιμοποιούμε τη **Μάζα Ξηρού Βάρους (Dry Weight Biomass)**, μια αριθμητική μεταβλητή κλίμακας αναλογίας, που βρίσκεται στη στήλη Weight του αρχείου MORPH_DATA.CSV.
- **Παράμετροι Πληθυσμού:** Επειδή μας ενδιαφέρει αν τα μωβ φυτά είναι **γενικά** μεγαλύτερα, οι παράμετροι ενδιαφέροντος είναι οι **Μέσοι Όροι Ξηρού Βάρους** κάθε μορφής.

Εξέταση των Δεδομένων

Υπολογίζουμε τις **σημειακές εκτιμήσεις** του μέσου ξηρού βάρους, της τυπικής απόκλισης (sd) και του μεγέθους δείγματος (n) για κάθε μορφή:

```
# using morph data...
morph_data %>%
  # ...group the data by colour morph category
  group_by(Colour) %>%
  # ... calculate the mean, sd and sample size of weight in each category
  summarise(mean = mean(Weight),
             sd = sd(Weight),
             samp_size = n())
```

Colour	mean	sd	samp_size
<chr>	<dbl>	<dbl>	<int>
1 Green	638.	50.5	155
2 Purple	648.	38.2	95

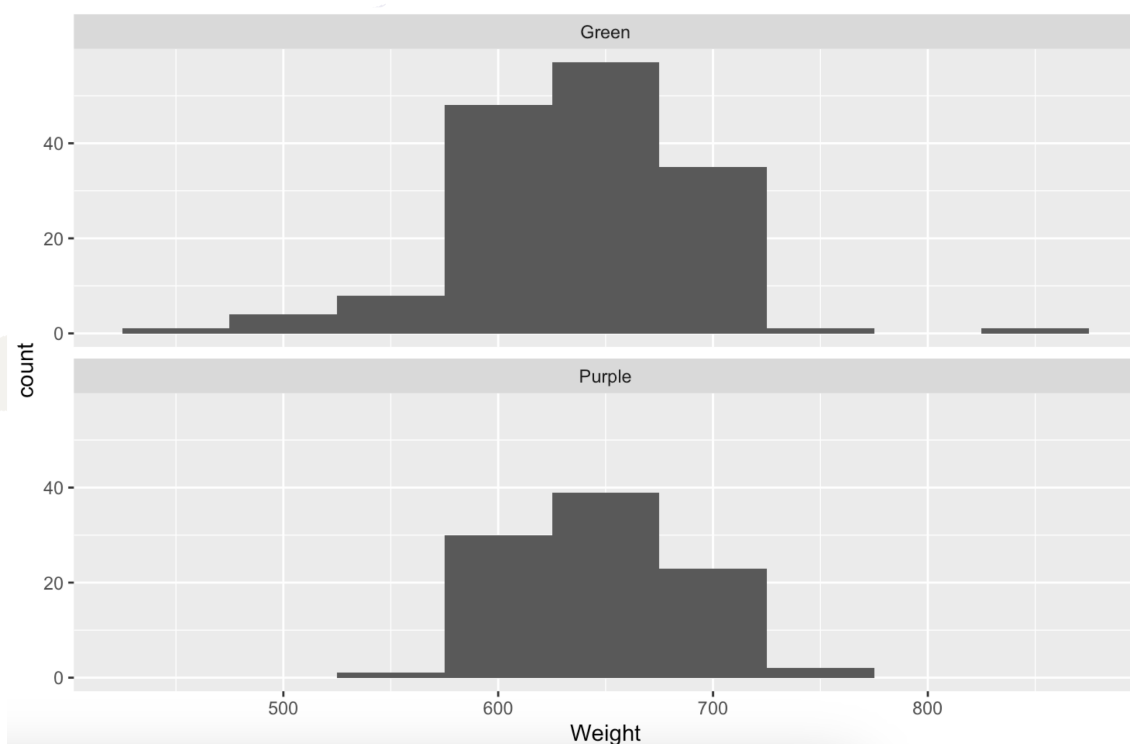
Σημειακή Εκτίμηση: Ο μέσος όρος του μωβ μορφότυπου (648) είναι όντως **μεγαλύτερος** από αυτόν του πράσινου (638).

Αβεβαιότητα: Οι εκτιμήσεις sd δείχνουν παρόμοια μεταβλητότητα.

Αν και παρατηρούμε μια διαφορά στους μέσους όρους των δειγμάτων, δεν μπορούμε να συμπεράνουμε τίποτα ακόμα. Πρέπει να ελέγξουμε αν αυτή η διαφορά οφείλεται απλώς σε **τυχαία διακύμανση δειγματοληψίας**.

Οπτική Εξέταση: Τα ιστογράμματα των βαρών δείχνουν ότι, παρόλο που η μωβ μορφή τείνει να έχει υψηλότερα βάρη, υπάρχει **σημαντική επικάλυψη** μεταξύ των δύο κατανομών.

```
ggplot(morph_data, aes(x = Weight)) +  
  geom_histogram(binwidth = 50) +  
  facet_wrap(~Colour, ncol = 1) # <- separate histogram for each colour morph!
```



Η Στατιστική Ελέγχου (Test Statistic)

Για να αξιολογήσουμε αν η διαφορά μεταξύ των δύο πληθυσμιακών μέσων είναι **στατιστικά σημαντική**, ακολουθούμε τα επόμενα βήματα:

1. **Μηδενική Υπόθεση (H_0):** Υποθέτουμε ότι **δεν υπάρχει διαφορά** μεταξύ των πληθυσμιακών μέσων, δηλαδή ότι και τα δύο δείγματα προέρχονται από έναν πληθυσμό με τον ίδιο μέσο όρο.
2. **Μηδενική Κατανομή:** Χρησιμοποιούμε τα πραγματικά δείγματα για να βρούμε τι θα συνέβαινε αν λαμβάναμε επανειλημμένα νέα δείγματα υπό την υπόθεση «καμίας διαφοράς». Αυτό συνοψίζεται στη **Μηδενική Κατανομή** μιας κατάλληλης **Στατιστικής Ελέγχου** (← **Νέα Έννοια**).
3. **p-value:** Υπολογίζουμε την πιθανότητα να παρατηρήσουμε μια διαφορά **ίση ή πιο ακραία** από αυτή που βρέθηκε στο δείγμα, αν ίσχυε η H_0 .
4. **Συμπέρασμα:** Εάν το p-value είναι αρκετά απίθανο (πχ <0.05), συμπεραίνουμε ότι το αποτέλεσμα είναι **στατιστικά σημαντικό**.

Η Στατιστική Ελέγχου (Test Statistic)

Όταν συγκρίνουμε περισσότερες από μία παραμέτρους, χρειαζόμαστε μια **ενιαία αριθμητική σύνοψη** της «επίδρασης ενδιαφέροντος». Αυτός ο αριθμός ονομάζεται **Στατιστική Ελέγχου**. Σκοπός της είναι να συνοψίσει την επίδραση σε έναν μόνο αριθμό για τη διενέργεια στατιστικού ελέγχου.

Έλεγχος Μεταθέσεων (Permutation Tests)

Ο Έλεγχος Μεταθέσεων είναι μια μέθοδος που χρησιμοποιείται για να καταδειχθεί η λογική της H_0 .

Λογική: Η H_0 υποθέτει ότι οι ετικέτες ('Purple' και 'Green') είναι **άνευ σημασίας** και θα μπορούσαν να είχαν ανατεθεί τυχαία.

Διαδικασία Μετάθεσης:

Δημιουργούμε τεχνητά δείγματα **αναθέτοντας τυχαία** τις ετικέτες 'Purple' και 'Green' στις **τιμές βάρους**, διατηρώντας το αρχικό **μέγεθος δείγματος** κάθε ομάδας

```
table(plant_morphs)
  Green Purple
    155     95
```

Επαναλαμβάνουμε τη διαδικασία 1000 - 10000 φορές.

Για κάθε τεχνητό δείγμα, υπολογίζουμε τη **Στατιστική Ελέγχου**, η οποία στην περίπτωση μας είναι η **διαφορά μεταξύ των μέσων όρων** του μωβ και του πράσινου φαινότυπου.

Διενέργεια Ελέγχου Μεταθέσεων

Κάθε τυχαία μετάθεση αποδίδει μια τιμή από τη **Μηδενική Κατανομή** της διαφοράς μεταξύ των μέσων, υπό την υπόθεση ότι η διαφορά στον πληθυσμό είναι **μηδέν**.

• Παραδείγματα Διαφοράς Μέσων:

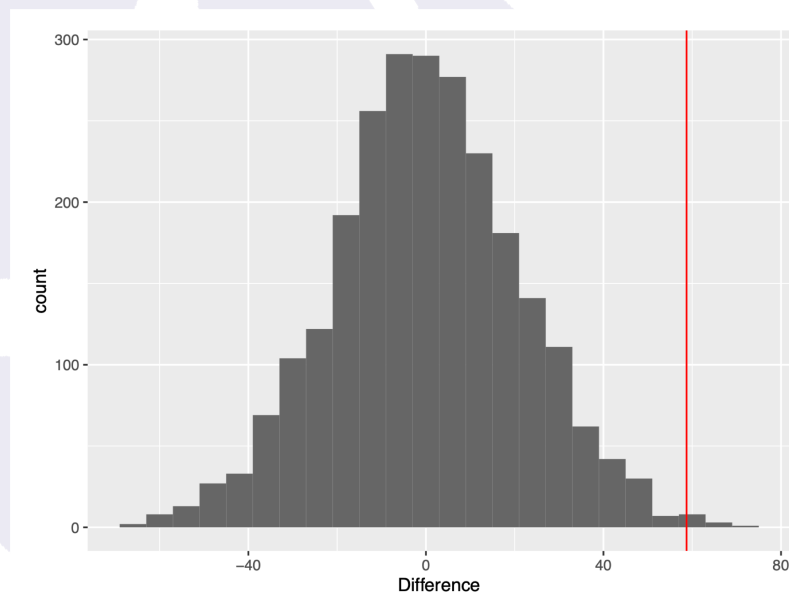
- 1η μετάθεση: 4.07
- 2η μετάθεση: -19.97

• Μηδενική Κατανομή: Η κατανομή όλων αυτών των διαφορών είναι η προσέγγιση της Μηδενικής Κατανομής.

- **Χαρακτηριστικό:** Η κατανομή αυτή **κεντράρεται στο μηδέν**, όπως αναμένεται αν οι ετικέτες είναι τυχαίες.

• Στατιστική Ελέγχου (Παρατηρούμενη): Η κόκκινη γραμμή δείχνει τη **διαφορά** μεταξύ των μέσων όρων (βάρους) του **πραγματικού δείγματος** (π.χ. AveragePurple – AverageGreen = 767 – 708 = 59).

Η παρατηρούμενη διαφορά βρίσκεται στην άκρη της δεξιάς ουράς της Μηδενικής Κατανομής. Αυτό υποδηλώνει ότι είναι **απίθανο** η παρατηρούμενη διαφορά να οφείλεται απλώς σε τυχαία διακύμανση, αν οι πληθυσμιακοί μέσοι όροι ήταν ίσοι.



Υπολογισμός του p-value

Χρησιμοποιούμε τη Μηδενική Κατανομή για να ποσοτικοποιήσουμε την πιθανότητα να βρεθεί η παρατηρούμενη διαφορά υπό τη H_0 .

Αποτέλεσμα (Υποθετικό): Μόνο 10 από τις 2500 μεταθέσεις έχουν διαφορά ίση ή πιο θετική από την παρατηρούμενη.

p-value: $p = 10/2500 = 0.004$

Ερμηνεία και Συμπέρασμα

Ερμηνεία: Το $p=0.004$ είναι η πιθανότητα να παρατηρηθεί η διαφορά (59) ή μια μεγαλύτερη διαφορά, αν δεν υπήρχε καμία διαφορά μεταξύ των βαρών των πληθυσμών (μωβ και πράσινων φυτών).

Συμπέρασμα: Δεδομένου ότι $p=0.004$ είναι **μικρότερο** από το συμβατικό επίπεδο σημαντικότητας $p < 0.05$, καταλήγουμε ότι η διαφορά στους μέσους όρους ξηρού βάρους μεταξύ των μωβ και πράσινων μορφών είναι **στατιστικά σημαντική**.

Περίληψη Αναφοράς:

Η μέση μάζα ξηρού βάρους των μωβ φυτών ($n=95$) ήταν **σημαντικά μεγαλύτερη** από αυτή των πράσινων φυτών ($n=155$) (μονόπλευρος έλεγχος μεταθέσεων, $p < 0.05$).

Η **Στατιστική Ελέγχου**, ένας ενιαίος αριθμός που ποσοτικοποιεί την «επίδραση» (π.χ., τη διαφορά μεταξύ των μέσων)

Στη συνέχεια, συγκρίνουμε την παρατηρούμενη Στατιστική Ελέγχου με τη Μηδενική Κατανομή για να υπολογίσουμε το p -value.

Ο Ρόλος της Στατιστικής Σημαντικότητας

Η στατιστική ανάλυση, μέσω της χρήσης p -values, μπορεί να μας βοηθήσει να διακρίνουμε εάν οι παρατηρούμενες διαφορές στα δεδομένα μας είναι **γνήσιες (genuine differences)** ή οφείλονται απλώς σε **τυχαία διακύμανση δειγματοληψίας (random variation)**.

Τι Μπορεί να μας Πει: Μας δίνει ένα μέτρο της **πιθανότητας** το αποτέλεσμα μας να είναι τυχαίο, **υποθέτοντας ότι δεν υπάρχει καμία επίδραση (H_0)**. Εάν αυτή η πιθανότητα είναι πολύ χαμηλή ($p < 0.05$), λέμε ότι η διαφορά είναι **Στατιστικά Σημαντική**.

Τι Δεν Μπορεί να μας Πει: Η στατιστική σημαντικότητα **δεν μπορεί** να μας πει εάν αυτή η διαφορά είναι **βιολογικά ενδιαφέρουσα, σημαντική, ή μεγάλης κλίμακας** (βιολογική σημασία).

Αυτή η αξιολόγηση —εάν μια στατιστικά σημαντική διαφορά έχει **πρακτική ή βιολογική σημασία**— είναι αποκλειστική **ευθύνη του επιστήμονα** που διεξάγει την έρευνα.

Λίγα Λόγια για τη Μηδενική Υπόθεση (H_0)

Μηδενική Υπόθεση (H_0): Είναι μια δήλωση για το τι θα περιμέναμε να δούμε αν δεν υπήρχε καμία επίδραση του παράγοντα που εξετάζουμε στη μεταβλητή που μετράμε.

Παράδειγμα: Στο παράδειγμα των φυτών, η H_0 ήταν: **Δεν υπάρχει διαφορά στη μέση βιομάζα των μωβ και πράσινων φυτών.**

Λειτουργία: Όλοι οι στατιστικοί έλεγχοι λειτουργούν αρχικά προσδιορίζοντας μια H_0 και στη συνέχεια αξιολογώντας αν τα παρατηρούμενα δεδομένα **αποκλίνουν** από αυτήν με τρόπο που είναι **ανεξάρτητος από τη δειγματοληπτική διακύμανση**.

Εναλλακτική Υπόθεση (H_1 ή Test Hypothesis): Είναι η δήλωση της επίδρασης που αναμένουμε (ή ελπίζουμε!) να δούμε.

Παράδειγμα: Τα μωβ και πράσινα φυτά **διαφέρουν στο μέσο μέγεθός τους**.

Αποδοχή, Απόρριψη και p-value

p-value: Ο στατιστικός έλεγχος δίνει μια **πιθανότητα** που μας λέει πόσο πιθανό είναι να είχαμε λάβει το παρατηρούμενο αποτέλεσμα, ή ένα πιο ακραίο, **αν η H_0 ήταν πραγματικά αληθής**.

Εύρος p-value	Απόφαση	Ερμηνεία
Μικρό ($p < \text{threshold}$, π.χ. <0.05)	Απορρίπτουμε τη Μηδενική Υπόθεση H_0	Το αποτέλεσμα είναι απίθανο να οφείλεται σε δειγματοληπτική διακύμανση. Λέμε ότι υπάρχει στατιστικά σημαντική διαφορά.
Μεγάλο ($p \geq \text{threshold}$, π.χ. ≥ 0.05)	Δεν απορρίπτουμε τη Μηδενική Υπόθεση H_0	Το αποτέλεσμα είναι πιθανό να έχει προκύψει από δειγματοληπτική διακύμανση.

Σημαντική Διάκριση: Η **απόρριψη** της H_0 **δεν** σημαίνει ότι «αποδεικνύεται» ότι η Εναλλακτική Υπόθεση H_1 είναι αληθής.

Το να **μην απορρίπτουμε** τη H_0 **δεν** είναι το ίδιο με το να **αποδεχόμαστε** ότι η H_0 είναι αληθής. Ένας λόγος γι' αυτό μπορεί να είναι ότι, αν το δείγμα είναι μικρό, μπορεί να υπάρχει μια επίδραση, αλλά ο έλεγχος δεν ήταν αρκετά ισχυρός για να την ανιχνεύσει (έλλειψη επαρκών δεδομένων).

Δεν είναι απαραίτητο να αναφέρεται η H_0 κατά την επιστημονική επικοινωνία

Η έμφαση πρέπει να δίνεται στις **ουσιαστικές βιολογικές υποθέσεις** (δηλαδή, στην Εναλλακτική Υπόθεση, H_1)

Το Κατώφλι Σημαντικότητας $p < 0.05$

Σύμβαση: Συμβατικά, ένα αποτέλεσμα θεωρείται **στατιστικά σημαντικό** αν $p < 0.05$. Αυτό το κατώφλι ονομάζεται **επίπεδο σημαντικότητας** (significance level).

Μια πιθανότητα 0.05 αντιστοιχεί σε **1 στις 20 περιπτώσεις** (ή 5 στις 100). Αυτό σημαίνει ότι:

Αν **δεν** υπήρχε **πραγματική επίδραση** του παράγοντα που διερευνάται, θα περιμέναμε να πάρουμε ένα αποτέλεσμα **“σημαντικό”** στο $p=0.05$ περίπου **5 φορές στα 100 δείγματα**.

Αυτό δείχνει ότι ένα «σημαντικό» αποτέλεσμα δεν είναι κάτι εξαιρετικά σπάνιο.

Προσοχή στα p-values!

Παρερμηνείες:

- **Ασυνέπεια Κατωφλίου:** Είναι παράλογο να απορρίπτουμε μια ιδέα αν $p=0.045$, αλλά να είμαστε απρόθυμοι να την υιοθετήσουμε αν $p=0.055$, καθώς η ουσιαστική διαφορά είναι αμελητέα.
- **Εξάρτηση:** Η ακριβής τιμή του p-value επηρεάζεται από:
 - Το **μέγεθος της αληθινής επίδρασης** που μελετάται.
 - Την **ποσότητα των δεδομένων** που αναλύονται.
 - Την **καταλληλότητα** του στατιστικού μοντέλου για τα δεδομένα.
Προσοχή: Είναι εύκολο να προκύψει ένα πολύ μικρό p-value ακόμα και όταν η επίδραση είναι αδύναμη ή ανύπαρκτη, αν το μοντέλο είναι ακατάλληλο.
- Δεν είναι μέτρο απόδειξης/μεγέθους:
 - **Δεν είναι σωστό** να εξισώνεται το μέγεθος του p-value με τη **βαρύτητα των αποδείξεων** ή να ερμηνεύεται ως δήλωση για το **πόσο μεγάλη είναι η επίδραση**.

Παρουσίαση των p-values

Αν και η R εμφανίζει τα p-values με πολλά δεκαδικά ψηφία, κατά την αναφορά των αποτελεσμάτων χρησιμοποιούνται συνήθως τέσσερις κατηγορίες:

Κατηγορία	Εύρος	Περιγραφή
$p > 0.05$	$p > 0.05$	Μη στατιστικά σημαντικό (NS)
$p < 0.05$	$0.01 < p < 0.05$	Σημαντικό *
$p < 0.01$	$0.001 < p < 0.01$	Πολύ σημαντικό **
$p < 0.001$	$p < 0.001$	Εξαιρετικά σημαντικό ***

Αν και οι τέσσερις κατηγορίες είναι συνήθως επαρκείς, η αναφορά της **πραγματικής πιθανότητας** μπορεί να είναι χρήσιμη. Για παράδειγμα, το να αναφέρεται $p = 0.06$ αντί για $p > 0.05$ είναι κατατοπιστικό, καθώς η τιμή είναι πολύ κοντά στο κατώφλι, υποδεικνύοντας πιθανή παρουσία μιας επίδρασης.

Βιολογική vs. Στατιστική Σημαντικότητα

Αυτό είναι ένα ζωτικής σημασίας σημείο: **Δεν πρέπει να συγχέουμε τη στατιστική σημαντικότητα με τη βιολογική σημασία.**

Στατιστική Σημαντικότητα: Ένα αποτέλεσμα μπορεί να είναι **στατιστικά ιδιαίτερα σημαντικό** (π.χ., $p < 0.001$) αλλά **βιολογικά ασήμαντο** (trivial).

Παράγοντες Σημαντικότητας

Η **σημαντικότητα ενός αποτελέσματος** εξαρτάται από έναν συνδυασμό τριών παραγόντων:

- Το **μέγεθος της αληθινής επίδρασης** στον πληθυσμό.
- Η **μεταβλητότητα των δεδομένων**.
- Το **μέγεθος του δείγματος** (n).

Ακόμη και μια μικροσκοπική επίδραση μπορεί να είναι στατιστικά σημαντική εάν το **μέγεθος του δείγματος είναι πολύ μεγάλο**.

Timokratis Karamitros
Bioinformatics and Applied Genomics Unit

ΠΑΡΑΜΕΤΡΙΚΟΙ ΕΛΕΓΧΟΙ

Timokratis Karamitros
Bioinformatics and Applied Genomics Unit

Κανονική Κατανομή (Normal Distribution)

Κανονική Κατανομή (ή Κατανομή Gauss),

Εφαρμογή και Τύπος Μεταβλητής

Η Κανονική Κατανομή είναι κατάλληλη για **αριθμητικές μεταβλητές** που μετρώνται σε κλίμακα **διαστήματος (interval)** ή **αναλογίας (ratio)**.

Αν και η μεταβλητή πρέπει να είναι **συνεχής (continuous)**, η Κανονική Κατανομή μπορεί να παρέχει μια αξιοπρεπή προσέγγιση για ορισμένους τύπους **διακριτών αριθμητικών δεδομένων**.

Περιγραφή της Κατανομής

Η Κανονική Κατανομή περιγράφεται πλήρως από δύο μόνο παραμέτρους:

Τον **Μέσο (Mean)**, ο οποίος αντιπροσωπεύει την κεντρική της τάση.

Την **Τυπική Απόκλιση (Standard Deviation - S.D.)**, η οποία αντιπροσωπεύει τη διασπορά της.

Αν γνωρίζουμε αυτές τις δύο ποσότητες για μια συγκεκριμένη Κανονική Κατανομή, γνωρίζουμε **όλα όσα πρέπει να γνωρίζουμε** για αυτήν την κατανομή.

Ο Κανόνας των 4 S.D. (Εύρος 95%)

Εάν μια μεταβλητή ακολουθεί Κανονική Κατανομή, τότε περίπου το **95% των τιμών της** θα πέσει μέσα σε ένα διάστημα που έχει εύρος **4 Τυπικές Αποκλίσεις**:

• **Άνω Όριο:** Μέσος + 2 x Τυπική Απόκλιση

• **Κάτω Όριο:** Μέσος - 2 x Τυπική Απόκλιση

Μαθηματικές Ιδιότητες

Όταν **προσθέτουμε ή αφαιρούμε** δύο κανονικά κατανεμημένες μεταβλητές για να δημιουργήσουμε μια νέα μεταβλητή, η προκύπτουσα μεταβλητή θα είναι επίσης **Κανονικά Κατανεμημένη**.

Ομοίως, εάν **πολλαπλασιάσουμε** μια κανονικά κατανεμημένη μεταβλητή με έναν αριθμό, η προκύπτουσα μεταβλητή θα παραμείνει **Κανονικά Κατανεμημένη**.

Τυπικό Σφάλμα του Μέσου (Standard Error of the Mean - SE)

Το **Τυπικό Σφάλμα του Μέσου (SE)** είναι μια εκτίμηση του πόσο ο μέσος όρος του δείγματος μπορεί να αποκλίνει από τον πραγματικό πληθυσμιακό μέσο όρο.

s^2 είναι η **διασπορά** (variance) του δείγματος (όπου s είναι η τυπική απόκλιση του δείγματος).
 n είναι το **μέγεθος** του δείγματος (sample size).

Τότε, το Τυπικό Σφάλμα του Μέσου (SE):

$$SE = \frac{s}{\sqrt{n}} = \sqrt{\frac{s^2}{n}}$$

Η Κατανομή t (The t Distribution)

Η κατανομή t προκύπτει κάθε φορά που:

- Λαμβάνουμε ένα δείγμα από μια **κανονικά κατανεμημένη μεταβλητή**.
- Εκτιμούμε τον **πληθυσμιακό μέσο όρο** από αυτό το δείγμα.
- Και στη συνέχεια, **διαιρούμε τον μέσο όρο με το τυπικό του σφάλμα (SE)**.

Η **Κατανομή Δειγματοληψίας** αυτής της νέας ποσότητας ακολουθεί την **Κατανομή t του Student**.

Επειδή η κατανομή t περιλαμβάνει την **ανακλιμάκωση ενός μέσου με το τυπικό του σφάλμα (SE)**, το σχήμα της εξαρτάται μόνο από το **μέγεθος του δείγματος (sample size)**. Αυτό μας επιτρέπει να κατασκευάσουμε απλούς **στατιστικούς ελέγχους** για την αξιολόγηση διαφορών μεταξύ μέσων όρων, τα λεγόμενα **t-tests**.

One-Sample t-test

Ο έλεγχος t ενός δείγματος είναι ένας απλός στατιστικός έλεγχος που χρησιμοποιείται όταν θέλουμε να **συγκρίνουμε τον πληθυσμιακό μέσο όρο ενός αριθμητικού δείγματος με μια συγκεκριμένη, προκαθορισμένη τιμή**. Ο έλεγχος αξιολογεί αν ο πληθυσμιακός μέσος όρος είναι πιθανό να διαφέρει από αυτήν την τιμή.

Χρησιμοποιούμε τον έλεγχο t ενός δείγματος για να ελέγξουμε αν ο μέσος όρος των δεδομένων μας **διαφέρει στατιστικά σημαντικά** από μια τιμή που μας ενδιαφέρει (πχ μια αναμενόμενη τιμή, ένα θεωρητικό όριο, ή μια προηγούμενη μέτρηση).

Ο έλεγχος t ενός δείγματος είναι ουσιαστικά ένας έλεγχος της **Μηδενικής Υπόθεσης** ότι ο πληθυσμιακός μέσος όρος ισούται με την τιμή που μας ενδιαφέρει.

Περίπτωση

Θεωρητικό Μοντέλο

Ρύθμιση/Νομοθεσία

Ερώτημα Ελέγχου

Ένα θεωρητικό μοντέλο πρόβλεψε ότι ένα ζώο θα πρέπει να εγκαταλείπει μια περιοχή αναζήτησης τροφής μετά από **10 λεπτά**. Έχουμε δεδομένα από 25 ζώα.

Παρακολουθούμε τη ρύπανση της θάλασσας και θέλουμε να ελέγξουμε αν η μέση πυκνότητα των κολοβακτηριδίων υπερβαίνει το **νομοθετημένο όριο**.

Στατιστικός Έλεγχος

Μπορούμε να ελέγξουμε αν ο μέσος χρόνος αναζήτησης τροφής **διαφέρει σημαντικά** από την πρόβλεψη των 10 λεπτών.

Ο έλεγχος t ενός δείγματος μπορεί να ελέγξει αν η μέση τιμή για ολόκληρη την παραλία **υπερβαίνει αυτό το όριο**.

(προ-) Υποθέσεις του Ελέγχου t Ενός Δείγματος

Για να είναι έγκυρος ο έλεγχος t ενός δείγματος, πρέπει να πληρούνται αρκετές υποθέσεις, οι οποίες παρατίθενται κατά **φθίνουσα σειρά σημαντικότητας**:

1. Ανεξαρτησία (Independence)

Η υπόθεση της ανεξαρτησίας είναι η **πιο σημαντική**:

Ορισμός: Κάθε παρατήρηση στο δείγμα **δεν πρέπει να «εξαρτάται»** από τις άλλες.

Σημασία: Εάν τα δεδομένα δεν είναι ανεξάρτητα, τα p -values που παράγονται από τον έλεγχο t θα είναι **αναξιόπιστα**.

Κίνδυνος: Συγκεκριμένα, τα p -values τείνουν να είναι **πολύ μικρά** όταν παραβιάζεται η υπόθεση της μη-ανεξαρτησίας. Αυτό οδηγεί στον κίνδυνο του **λανθασμένου συμπεράσματος** ότι μια διαφορά είναι στατιστικά σημαντική, ενώ στην πραγματικότητα δεν είναι.

2. Κλίμακα Μέτρησης (Measurement Scale)

Απαιτήση: Η μεταβλητή που αναλύεται πρέπει να μετράται σε κλίμακα **διαστήματος (interval)** ή **αναλογίας (ratio)**, δηλαδή πρέπει να είναι μια **αριθμητική μεταβλητή**.

Λογική: Δεν είναι λογική η εφαρμογή ενός ελέγχου t σε μεταβλητή που δεν μετράται σε αυτές τις κλίμακες.

(προ-) Υποθέσεις του Ελέγχου t Ενός Δείγματος

3. Κανονικότητα (Normality)

Απαίτηση: Ο έλεγχος t ενός δείγματος θα παράγει εντελώς αξιόπιστα p-values μόνο όταν η μεταβλητή είναι **Κανονικά Κατανεμημένη** στον πληθυσμό.

Σημαντικότητα: Αυτή η υπόθεση είναι **λιγότερο σημαντική** από ό,τι πιστεύουν πολλοί:

Ο έλεγχος t είναι **ανθεκτικός** (robust) σε ήπιες αποκλίσεις από την κανονικότητα όταν το **μέγεθος του δείγματος είναι μικρό**.

Όταν το **μέγεθος του δείγματος είναι μεγάλο**, η υπόθεση της κανονικότητας **σχεδόν δεν έχει σημασία**.

Timokratis Karamitros
Bioinformatics and Applied Genomics Unit

One-Sample t-test in R

Το Νέο Ερώτημα:

Παλαιότερη Αναφορά: Μια αναφορά πριν από 20 χρόνια βρήκε ότι το μέσο βάρος των μωβ μορφών ήταν **710 γραμμάρια** (g).

Ερώτημα Ελέγχου: Θέλουμε να αξιολογήσουμε εάν το μέσο βάρος των μωβ φυτών στον **σημερινό πληθυσμό** διαφέρει από αυτήν την παλαιότερη τιμή, υποθέτοντας ότι μπορεί να έχουν προσαρμοστεί στις τοπικές συνθήκες.

Για αυτό το παράδειγμα, χρειαζόμαστε **μόνο τα δεδομένα της μωβ μορφής**.

```
# get just the purple morphs  
purple_morphs <- filter(morph_data, Colour == "Purple")
```

Timokratis Karamitros
Bioinformatics and Applied Genomics Unit

Οπτικοποίηση Δεδομένων και Έλεγχος Υποθέσεων

Υπολογισμός Συνοπτικών Στατιστικών

```
purple_morphs %>%  
  summarise(mean = mean(Weight),  
            sd = sd(Weight),  
            samp_size = n())
```

Μέση Τιμή: Ο μέσος όρος ξηρού βάρους για τα μωβ φυτά είναι περίπου 647.58.

Τυπική Απόκλιση: Η διασπορά στο δείγμα είναι 38.24

Μέγεθος Δείγματος: Το δείγμα μας περιέχει 94 παρατηρήσεις.

Timokratis Karamitros
Bioinformatics and Applied Genomics Unit

Έλεγχος Υποθέσεων

Ανεξαρτησία

Πρέπει να υποθέσουμε ότι η μέτρηση του βάρους κάθε φυτού είναι **ανεξάρτητη** από τις άλλες. Αυτό εξαρτάται από τον σχεδιασμό του πειράματος (π.χ., τα φυτά δεν προέρχονται από την ίδια μητρική γραμμή, ούτε καλλιεργήθηκαν στην ίδια γλάστρα). Υποθέτουμε ότι οι παρατηρήσεις είναι ανεξάρτητες. **(Υποτίθεται)**

Κλίμακα Μέτρησης

Η μεταβλητή Weight (ξηρό βάρος) είναι μια **αριθμητική μεταβλητή** σε κλίμακα **αναλογίας**. **(Ισχύει)**

Κανονικότητα

Οπτικός Έλεγχος (Ιστόγραμμα): Οπτικοποιώντας το ιστόγραμμα των βαρών για τα 95 μωβ φυτά, μπορούμε να εκτιμήσουμε αν η κατανομή τους προσεγγίζει την καμπύλη κανονικής κατανομής (καμπύλη καμπάνας).

Δεδομένου ότι το μέγεθος του δείγματος ($n=95$) είναι **αρκετά μεγάλο** ($n > 30$), ο έλεγχος t είναι **ανθεκτικός** (robust) σε ήπιες αποκλίσεις από την κανονικότητα **(Ανθεκτικό)**

Timokratis Karamitros
Bioinformatics and Applied Genomics Unit

Διεξαγωγή του t-test Ενός Δείγματος

Χρησιμοποιούμε τη συνάρτηση `t.test()` για να συγκρίνουμε τον μέσο όρο του δείγματος με την παλαιότερη αναμενόμενη τιμή (710g).

Μηδενική Υπόθεση (H0): Ο μέσος πληθυσμιακός όρος είναι ίσος με 710g.

Εναλλακτική Υπόθεση (H1): Ο μέσος πληθυσμιακός όρος είναι διαφορετικός από 710g (Αυτόματος αμφίπλευρος έλεγχος - two-tailed test).

```
t.test(purple_morphsWeight, mu = 710)
```

`purple_morphsWeight`: Το αριθμητικό διάνυσμα των βαρών που αναλύουμε.

`mu = 710`: Η τιμή του **πληθυσμιακού μέσου όρου** (μ) που ελέγχουμε υπό τη Μηδενική Υπόθεση.

Το αποτέλεσμα του ελέγχου θα μας δώσει τη **Στατιστική Ελέγχου t**, τους **Βαθμούς Ελευθερίας (df)** και το **p-value** που απαιτείται για να βγάλουμε το συμπέρασμά μας.

Ερμηνεία του αποτελέσματος

One Sample t-test

```
data: purple_morphsWeight
t = -15.912, df = 94, p-value < 2.2e-16
alternative hypothesis: true mean is not
equal to 710
95 percent confidence interval:
 639.7785 655.3587
sample estimates:
mean of x
 647.5686
```

Στοιχείο

Τιμή

Επεξήγηση

t

-15.912

Η **Στατιστική Ελέγχου** (t-statistic). Αυτός είναι ο αριθμός που συγκρίνεται με την Κατανομή t (Student's t-distribution).

df

94

Οι **Βαθμοί Ελευθερίας** (degrees of freedom, $df = n - 1 = 95 - 1 = 94$). Αυτή η τιμή απαιτείται για τον προσδιορισμό του σχήματος της Κατανομής t και τον υπολογισμό του p-value.

p-value

< 2.2e-16

Η πιθανότητα να παρατηρηθεί το αποτέλεσμα μας (ή πιο ακραίο), αν ίσχυε η H_0 .

Στατιστική Σημαντικότητα

Το $p\text{-value} < 2.2e-16$ είναι **πολύ μικρότερο** από το συμβατικό όριο σημαντικότητας 0.05.

•**Συμπέρασμα:** Υπάρχει **στατιστικά σημαντική διαφορά** μεταξύ της μέσης μάζας ξηρού βάρους και της αναμενόμενης τιμής των 710g.

•**Αναφορά:** Δεδομένου ότι το p-value είναι μικρότερο από 0.001 το αποτέλεσμα αναφέρεται ως **$p < 0.001$ (***)**.

Alternative hypothesis: true mean is not equal to 710: Επιβεβαιώνει την **Εναλλακτική Υπόθεση** (H_A) και, εμμέσως, απορρρίπτει τη **Μηδενική Υπόθεση**.

95 percent confidence interval: Δείχνει το **95% διάστημα εμπιστοσύνης** για τη διαφορά μεταξύ των μέσων (το εύρος των πιθανών τιμών του «αληθινού» μέσου).

Συνολική Ερμηνεία: Τα σημερινά μωβ φυτά έχουν μέση μάζα ξηρού βάρους που είναι **στατιστικά σημαντικά διαφορετική** από τα 710g που αναφέρθηκαν πριν από 20 χρόνια

Έλεγχος t Δύο Δειγμάτων (Two-Sample t-test)

Ο έλεγχος t δύο δειγμάτων είναι η **παραμετρική** εκδοχή της διαδικασίας μετάθεσης (permutation procedure) για τη σύγκριση δύο ομάδων: Σύγκριση των μέσων όρων μιας **αριθμητικής** μεταβλητής από **δύο ανεξάρτητους** πληθυσμούς.

Σκοπός του ελέγχου t δύο δειγμάτων είναι να αξιολογήσει αν ο μέσος όρος διαφέρει στατιστικά σημαντικά μεταξύ των δύο πληθυσμών. Η **Μηδενική Υπόθεση H_0** είναι ότι **δεν υπάρχει διαφορά** μεταξύ των δύο πληθυσμιακών μέσων.

Ο έλεγχος t δύο δειγμάτων ελέγχει αν η παρατηρούμενη διαφορά στους μέσους όρους των δειγμάτων είναι μεγαλύτερη από αυτή που θα αναμενόταν από την τυχαία διακύμανση, αν ίσχυε η H_0 .

Timokratis Karamitros
Bioinformatics and Applied Genomics Unit

(Προ)- Υποθέσεις του Ελέγχου t Δύο Δειγμάτων

Ουσιαστικά οι ίδιες με αυτές του ελέγχου t ενός δείγματος

Ανεξαρτησία (Independence) Κάθε παρατήρηση σε κάθε ένα από τα δύο δείγματα πρέπει να είναι **ανεξάρτητη** από τις άλλες.

Όπως και στον έλεγχο t ενός δείγματος, εάν τα δεδομένα δεν είναι ανεξάρτητα, τα p-values που παράγονται θα είναι **αναξιόπιστα** και συνήθως **υπερβολικά μικρά**. Αυτό αυξάνει τον κίνδυνο να καταλήξουμε ψευδώς σε συμπέρασμα στατιστικής σημαντικότητας.

Κλίμακα Μέτρησης (Measurement Scale)

Η μεταβλητή που συγκρίνεται πρέπει να είναι **αριθμητική**. Δεν έχει νόημα να εφαρμόζεται ο έλεγχος t σε μια κατηγορική μεταβλητή.

Κανονικότητα (Normality) Η μεταβλητή πρέπει να είναι **κανονικά κατανοημένη** σε κάθε έναν από τους **δύο πληθυσμούς** από τους οποίους ελήφθησαν τα δείγματα.

Ωστόσο, ο έλεγχος t δύο δειγμάτων είναι **αρκετά ανθεκτικός** (robust) σε ήπιες αποκλίσεις από την κανονικότητα. Όταν τα **μεγέθη των δειγμάτων είναι μεγάλα**, η υπόθεση αυτή έχει **μικρή σημασία**.

Timokratis Karamitros
Bioinformatics and Applied Genomics Unit

Αξιολόγηση των Υποθέσεων του Ελέγχου t Δύο Δειγμάτων

Ανεξαρτησία & Κλίμακα Μέτρησης

Ανεξαρτησία: Η εκτίμηση της ανεξαρτησίας απαιτεί την κριτική εξέταση του **τρόπου συλλογής των δεδομένων**. Δεν μπορούμε να τις ελέγξουμε με στατιστική συνάρτηση, αλλά πρέπει να σκεφτούμε αν η τιμή κάθε παρατήρησης επηρεάζεται από κάποια άλλη.

Κλίμακα Μέτρησης: Αυτό αξιολογείται άμεσα από τον **τύπο της μεταβλητής** (πρέπει να είναι αριθμητική κλίμακας διαστήματος ή αναλογίας).

Κανονικότητα: Πρέπει να εξετάσουμε την κατανομή σε **κάθε ομάδα ξεχωριστά**, όχι την κατανομή του συνδυασμένου δείγματος.

Μέθοδος: Μπορούμε να οπτικοποιήσουμε την κατανομή κάθε ομάδας χρησιμοποιώντας ένα **ζεύγος ιστογραμμάτων ή διαγράμματα σημείων (dot plots)**.

Κριτήριο:

Εάν και τα δύο δείγματα φαίνονται **περίπου κανονικά** (π.χ., περίπου σχήματος καμπάνας), τότε μπορούμε να προχωρήσουμε στον έλεγχο t.

Εάν παρατηρηθεί **απόκλιση από την κανονικότητα**, πρέπει να λάβουμε υπόψη τα **μεγέθη των δειγμάτων (n)**.

Μεγάλα Δείγματα: Ο έλεγχος t είναι **ανθεκτικός (robust)** σε ήπιες αποκλίσεις όταν τα μεγέθη των δειγμάτων είναι **μεγάλα** (π.χ., $n > 100$ σε κάθε ομάδα). Σε αυτή την περίπτωση, μπορούμε συνήθως να προχωρήσουμε.

```
morph_data %>%  
  # group the data by plant morph  
  group_by(Colour) %>%  
  # calculate the mean, standard deviation and sample size  
  summarise(mean = mean(Weight),  
             sd = sd(Weight),  
             samp_size = n())
```

Διεξαγωγή Ελέγχου t Δύο Δειγμάτων

Έλεγχος Κανονικότητας και Μέγεθος Δείγματος

Όπως γνωρίζουμε, τα δείγματα είναι:

Πράσινα φυτά: ???

Μωβ φυτά: ???

Αν και το μέγεθος δείγματος των πράσινων είναι μεγάλο (>100), το $n=77$ για τα μωβ είναι επαρκές ώστε ο έλεγχος t να είναι **αρκετά ανθεκτικός** (robust) σε ήπιες αποκλίσεις από την κανονικότητα.

Ελέγχουμε την κατανομή κάθε ομάδας ξεχωριστά:

```
ggplot(morph_data, aes(x = Weight)) +  
  geom_histogram(binwidth = 50) +  
  facet_wrap(~Colour, ncol = 1)
```

Αξιολόγηση: Δεν υπάρχει κάτι το ιδιαίτερα «μη-κανονικό» (non-normal) στις κατανομές των δύο δειγμάτων—φαίνονται περίπου κανονικές. Επομένως, είναι λογικό να υποθέσουμε ότι προήλθαν και τα δύο από κανονικά κατανεμημένους πληθυσμούς και να προχωρήσουμε.

Διεξαγωγή του Ελέγχου t Δύο Δειγμάτων με τη Συνάρτηση t.test

Δεδομένα: Το morph_data περιέχει τη μεταβλητή μέτρησης (Weight) και τη μεταβλητή ομάδας (Colour).

```
t.test(Weight ~ Colour, morph_data)
```

Ορίσματα:

Φόρμουλα (Formula): Weight ~ Colour

Περιλαμβάνει το σύμβολο ~ (**tilde**), το οποίο υποδηλώνει μια φόρμουλα.

Αριστερά της ~ (Weight): Η αριθμητική μεταβλητή της οποίας τους μέσους όρους θέλουμε να συγκρίνουμε.

Δεξιά της ~ (Colour): Η **μεταβλητή-δείκτης** (indicator variable) που καθορίζει σε ποια από τις δύο ομάδες ανήκει κάθε παρατήρηση.

Δεδομένα (Data): morph_data

Το όνομα του data frame που περιέχει τις μεταβλητές που αναφέρονται στη φόρμουλα.

Ο έλεγχος αυτός θα υπολογίσει αν η **διαφορά μεταξύ του μέσου βάρους** των μωβ και των πράσινων φυτών είναι στατιστικά σημαντική.

Ερμηνεία των Αποτελεσμάτων του Έλεγχου t Δύο Δειγμάτων

Welch Two Sample t-test: Αυτό επιβεβαιώνει ότι χρησιμοποιήθηκε η έκδοση Welch του ελέγχου t δύο δειγμάτων. Ο έλεγχος Welch είναι πιο ανθεκτικός (robust) επειδή λαμβάνει υπόψη την πιθανότητα άνισων διακυμάνσεων μεταξύ των δύο ομάδων.

data: Weight by Colour: Υπενθυμίζει ότι συγκρίναμε τους μέσους όρους της μεταβλητής Weight, όπου η συμμετοχή σε ομάδα καθορίστηκε από τη μεταβλητή Colour.
Στατιστική Ελέγχου και p-value

Στοιχείο	Τιμή	Επεξήγηση
t	-1.6839	Η Στατιστική Ελέγχου t. Το πρόσημο (αρνητικό) υποδηλώνει ότι ο μέσος όρος της δεύτερης ομάδας (Green, λόγω της αλφαβητικής σειράς) είναι μεγαλύτερος από την πρώτη (Purple), ή, πιο γενικά, ποια ομάδα αφαιρέθηκε από ποια.
d.f.	236.94	Οι Βαθμοί Ελευθερίας. Λόγω της διόρθωσης Welch (για άνισες διακυμάνσεις), οι βαθμοί ελευθερίας μπορεί να είναι ένας δεκαδικός αριθμός.
p-value	0.09351	Η πιθανότητα να παρατηρηθεί το αποτέλεσμα μας, αν ισχύει η H0.

Συμπέρασμα Σημαντικότητας

ο p-value είναι 0.09351. Συγκρίνοντας το με το συμβατικό όριο $p < 0.05$:
 $0.09351 > 0.05$

Δεν υπάρχει **στατιστικά σημαντική διαφορά** στη μέση μάζα ξηρού βάρους μεταξύ των δύο μορφών φυτών στο επίπεδο 5%.

alternative hypothesis: true difference in means is not equal to 0: Επιβεβαιώνει ότι η Εναλλακτική Υπόθεση H_A είναι ότι η αληθινή διαφορά στους μέσους όρους δεν είναι μηδέν (αμφίπλευρος έλεγχος).

95 percent confidence interval: Δείχνει το **95% διάστημα εμπιστοσύνης** για την **αληθινή διαφορά** μεταξύ των μέσων. Εάν το διάστημα αυτό περιέχει το μηδέν (0), όπως συμβαίνει όταν $p > 0.05$, αυτό επιβεβαιώνει ότι δεν απορρίπτουμε τη H₀.

sample means: Ανακεφαλαιώνει τους μέσους όρους του δείγματος για κάθε ομάδα

Έλεγχοι Συσχέτισης (Correlation Tests)

Η **συσχέτιση (correlation)** είναι ένα στατιστικό μέτρο **σύνδεσης (association)** μεταξύ δύο μεταβλητών. Σύνδεση είναι κάθε σχέση μεταξύ των μεταβλητών που τις καθιστά εξαρτημένες με κάποιο τρόπο: γνωρίζοντας την τιμή της μίας μεταβλητής, λαμβάνουμε πληροφορίες για τις πιθανές τιμές της άλλης.

Συσχέτιση vs. Σύνδεση: Αν και οι όροι χρησιμοποιούνται συχνά εναλλακτικά, η συσχέτιση έχει μια πιο στενή έννοια. Η ανάλυση συσχέτισης ποσοτικοποιεί τον βαθμό στον οποίο μια σύνδεση τείνει προς ένα συγκεκριμένο πρότυπο μέσω ενός μέτρου που ονομάζεται **συντελεστής συσχέτισης (correlation coefficient)**.

Timokratis Karamitros
Bioinformatics and Applied Genomics Unit

Συντελεστής Συσχέτισης: Βασικές Ιδιότητες

Ο συντελεστής συσχέτισης (όπως ο Συντελεστής Συσχέτισης **Pearson's r**) μετρά τον βαθμό στον οποίο δύο μεταβλητές τείνουν προς μια **ευθεία γραμμική σχέση**.

Εάν **δεν υπάρχει σχέση** μεταξύ των μεταβλητών, ο συντελεστής συσχέτισης θα είναι **μηδέν (0)**. Όσο πιο κοντά στο 0 είναι η τιμή, τόσο πιο αδύναμη είναι η σχέση.

Μια **τέλεια συσχέτιση** θα είναι είτε **-1** είτε **+1**, ανάλογα με την κατεύθυνση.

Η τιμή του συντελεστή συσχέτισης υποδηλώνει την **κατεύθυνση** (θετική/αρνητική) και την **ισχύ** της σύνδεσης.

Δεν λέει τίποτα για την **κλίση** της σχέσης (δηλαδή, πόσο απότομα αλλάζει η μία μεταβλητή σε σχέση με την άλλη). Ένας συντελεστής συσχέτισης από μόνος του **δεν μας λέει** αν μια σύνδεση είναι πιθανό να είναι «πραγματική» (δηλαδή, διαφορετική από το μηδέν στον πληθυσμό).

Πρέπει να χρησιμοποιήσουμε έναν **έλεγχο στατιστικής σημαντικότητας** για να αξιολογήσουμε αν μια συσχέτιση είναι πιθανό να **διαφέρει από το μηδέν**.

Timokratis Karamitros
Bioinformatics and Applied Genomics Unit

Ο Συντελεστής Συσχέτισης Pearson (r) και ο Έλεγχος Σημαντικότητας

Μέτρο **γραμμικής σύνδεσης** (linear association) μεταξύ **αριθμητικών** μεταβλητών.

«**Pearson's Correlation Test**»: Υπάρχει ένας τυπικός παραμετρικός έλεγχος που σχετίζεται με τον συντελεστή συσχέτισης Pearson. Θα τον ονομάζουμε «**Έλεγχο Συσχέτισης Pearson**».

Μηδενική Υπόθεση H_0 : Η H_0 αυτού του ελέγχου είναι ότι **δεν υπάρχει γραμμική συσχέτιση** στον πληθυσμό (δηλαδή, ο πληθυσμιακός συντελεστής συσχέτισης $r = 0$).

Timokratis Karamitros
Bioinformatics and Applied Genomics Unit

Ο Έλεγχος Συσχέτισης Pearson (Pearson's Correlation Test)

Καθορισμός μιας **Μηδενικής Υπόθεσης H_0**

Υπολογισμός μιας κατάλληλης **Στατιστικής Ελέγχου** (t-statistic)

Προσδιορισμός της **Μηδενικής Κατανομής** (Null Distribution)

Υπολογισμός ενός **p-value** από τον παρατηρούμενο συντελεστή

Timokratis Karamitros
Bioinformatics and Applied Genomics Unit

(Προ-) Υποθέσεις του Έλεγχου Συσχέτισης Pearson

Κλίμακα Μέτρησης: Και οι δύο μεταβλητές μετρούνται σε κλίμακα **διαστήματος (interval)** ή **αναλογίας (ratio)** (πρέπει να είναι αριθμητικές).

Κανονικότητα: Οι δύο μεταβλητές είναι **Κανονικά Κατανεμημένες** στον πληθυσμό.

Γραμμικότητα: Η σχέση μεταξύ των μεταβλητών είναι **γραμμική**.

*** Η υπόθεση της **Γραμμικότητας** είναι η μόνη που πρέπει να ισχύει για να είναι ο **Συντελεστής Pearson (r)** ένα έγκυρο μέτρο σύνδεσης. Ωστόσο, οι υποθέσεις 1 και 2 πρέπει να πληρούνται για να είναι κατάλληλος ο **σχετικός στατιστικός έλεγχος**.

Νέο Παράδειγμα: Alzheimer's Disease (AD)

Ερευνητές ενδιαφέρονται να χρησιμοποιήσουν έναν βιοδείκτη πλάσματος, το **ολικό tau (t-Tau)**, για να αξιολογήσουν την εξέλιξη της νόσου Alzheimer.

Μεταβλητές:

t-Tau: Επίπεδα πρωτεΐνης t-Tau στο πλάσμα (αριθμητική μεταβλητή).

MMSE Score: Βαθμολογία από το Mini-Mental State Examination (αριθμητική μεταβλητή, < 26 σημαίνει γνωστική εξασθένηση).

Σκοπός: Να αξιολογηθεί η **σύνδεση** μεταξύ των επιπέδων t-Tau και του βαθμού γνωστικής εξασθένησης (MMSE score) σε 40 ασθενείς με AD.

Οπτικοποίηση Δεδομένων και Έλεγχος Υποθέσεων

Έλεγχος Υποθέσεων:

Κλίμακα Μέτρησης: Τόσο το t-Tau όσο και το MMSE score είναι **αριθμητικές μεταβλητές**

Γραμμικότητα/Κανονικότητα: Αυτά πρέπει να ελεγχθούν **οπτικά** πριν από τη διεξαγωγή του ελέγχου.

```
alzheimers<- read.csv(file = "ALZHEIMERS.CSV")  
  
glimpse(alzheimers)  
  
ggplot(alzheimers, aes(x = MMSE, y = tTau)) +  
  geom_point()
```

Παρατήρηση: Υπάρχει μια **αρνητική σύνδεση** μεταξύ της γνωστικής απόδοσης (MMSE score) και των επιπέδων t-Tau. Αυτό σημαίνει ότι όσο αυξάνεται το t-Tau, τόσο μειώνεται το MMSE score (και αντίστροφα).

Σχήμα: Η σχέση φαίνεται να ακολουθεί ένα πρότυπο «**ευθείας γραμμής**» (straight line pattern).

Συμπέρασμα: Ο συντελεστής συσχέτισης Pearson είναι ένα **εύλογο μέτρο σύνδεσης** για αυτά τα δεδομένα. Η Υπόθεση #3 **ισχύει**.

Κανονικότητα?

```
ggplot(alzheimers, aes(x = MMSE)) +  
  geom_dotplot(binwidth = 2)  
  
ggplot(alzheimers, aes(x = tTau)) +  
  geom_dotplot(binwidth = 2)
```

Διεξαγωγή του Έλεγχου Συσχέτισης Pearson στην R

Χρησιμοποιούμε τη συνάρτηση `cor.test` στο R για να υπολογίσουμε τόσο τον συντελεστή συσχέτισης όσο και τη στατιστική σημαντικότητά του:

```
cor.test(~ MMSE + tTau, method = "pearson", data = alzheimers)
```

Φόρμουλα: `~ MMSE + tTau`

Η συνάρτηση `cor.test` αναμένει τη χρήση της σύνταξης φόρμουλας του R.

Οι δύο μεταβλητές που αναλύονται (MMSE και tTau) εμφανίζονται **δεξιά του ~** (tilde) και διαχωρίζονται με το σύμβολο +

Μέθοδος: `method = "pearson"`

Αυτό το όρισμα καθορίζει ποιος τύπος συντελεστή συσχέτισης θα υπολογιστεί.

Δεδομένα: `data = alzheimers`

Η χρήση της σύνταξης `~ X + Y` στο `cor.test`

Έμφαση στη Μη Κατευθυντικότητα: Αυτή η σύμβαση υπάρχει για να τονίσει την **απουσία «κατευθυντικότητας» (directionality)** στις συσχετίσεις.

Συνάρτηση	Σύνταξη Φόρμουλας	Ερμηνεία
t.test	<code>Y ~ X</code>	Η μεταβλητή απόκρισης (Y) εξαρτάται από τη μεταβλητή ομάδας (X). Υποδηλώνει κατεύθυνση ή εξάρτηση (π.χ., σύγκριση μέσου Y στις ομάδες X).
cor.test	<code>~ X + Y</code>	Συγκρίνει τις μεταβλητές X και Y. Και οι δύο μεταβλητές εμφανίζονται δεξιά του ~ διαχωρισμένες με +

Ερμηνεία των Αποτελεσμάτων του Έλεγχου Συσχέτισης Pearson

Στοιχείο Εξόδου	Τιμή	Επεξήγηση
t	-8.0191	Η Στατιστική Ελέγχου t. Αυτός ο t-στατιστικός αριθμός χρησιμοποιείται για τον έλεγχο της συσχέτισης και δεν σχετίζεται με τη σύγκριση μέσων .
df	38	Οι Βαθμοί Ελευθερίας (degrees of freedom). Υπολογίζεται ως $n - 2$, όπου $n=40$ είναι το μέγεθος του δείγματος. Το R χρησιμοποιεί αυτή την τιμή και την t-στατιστική για τον υπολογισμό του p-value.
p-value	1.075e-09	Η πιθανότητα να παρατηρηθεί αυτή η συσχέτιση (ή πιο ακραία), αν ίσχυε η H_0 ($r=0$).
cor (Pearson)	-0.8033816	Ο Συντελεστής Συσχέτισης Pearson (r), ο οποίος εκτιμά την ισχύ και την κατεύθυνση της σχέσης.

Στατιστική Σημαντικότητα & Συντελεστής Συσχέτισης

Το p-value είναι 1.075×10^{-9} (δηλαδή ~ 0.000000001075).

Καθώς $p < 0.05$ (ή $p < 0.001$), συμπεραίνουμε ότι υπάρχει μια **στατιστικά σημαντική** συσχέτιση μεταξύ των επιπέδων t-Tau και του MMSE score. Η H_0 μηδενικής συσχέτισης απορρίπτεται.

Ο Συντελεστής Συσχέτισης Pearson είναι $r \sim -0.80$.

Αρνητική Κατεύθυνση: Το αρνητικό πρόσημο επιβεβαιώνει τη **αρνητική σύνδεση** που παρατηρήθηκε στο scatter plot: καθώς τα επίπεδα t-Tau **αυξάνονται**, το MMSE score (γνωστική απόδοση) **μειώνεται**.

Ισχύς: Η τιμή -0.80 υποδηλώνει μια **μέτρια προς ισχυρή** αρνητική σύνδεση μεταξύ των δύο μετρήσεων.

Συνοπτικά: Στατιστικά σημαντική αρνητική γραμμική συσχέτιση μεταξύ των επιπέδων t-Tau και του MMSE score ($r = -0.80$, $t = -8.02$, $df=38$, $p < 0.001$). Αυτό υποδηλώνει ότι οι ασθενείς με υψηλότερα επίπεδα του βιοδείκτη t-Tau στο πλάσμα τείνουν να έχουν χαμηλότερη γνωστική απόδοση.

Συσχετίσεις και Παλινδρόμηση

Μεγάλο μέρος της βιολογίας ασχολείται με τις σχέσεις μεταξύ αριθμητικών μεταβλητών. Στο προηγούμενο κεφάλαιο, μάθαμε για τη συσχέτιση (correlation), η οποία ποσοτικοποιεί την ισχύ και την κατεύθυνση μιας σύνδεσης, αλλά δεν μας λέει τίποτα για τη μορφή της σχέσης ούτε μας επιτρέπει να κάνουμε προβλέψεις.

Αντίθετα με τη συσχέτιση, μια ανάλυση παλινδρόμησης (regression analysis) μας επιτρέπει να κάνουμε ακριβείς δηλώσεις για το πώς μια **εξαρτημένη μεταβλητή (y)** εξαρτάται από τις τιμές μιας άλλης **(ανεξάρτητη μεταβλητή, x)**.

Γραφικά, αυτό το αξιολογούμε χρησιμοποιώντας ένα scatter plot (διάγραμμα διασποράς).

Συσχέτιση και Παλινδρόμηση (Relationships and Regression)

Στη βιολογία, ένα μεγάλο μέρος της ανάλυσης αφορά τη μελέτη σχέσεων μεταξύ **αριθμητικών μεταβλητών** (π.χ., βάρος σε σχέση με το μήκος ψαριών, ή η ποικιλότητα ειδών σε σχέση με το pH του εδάφους).

Τεχνική	Σκοπός	Τι μας λέει
Συσχέτιση	Ποσοτικοποιεί την ισχύ και την κατεύθυνση της σύνδεσης.	$r \sim 0$ (αδύναμη) έως $r = 1$ (ισχυρή). Δεν περιγράφει τη μορφή της σχέσης, ούτε επιτρέπει προβλέψεις.
Παλινδρόμηση	Επιτρέπει την εκτίμηση του πώς μία μεταβλητή εξαρτάται από τις τιμές μιας άλλης.	Επιτρέπει ακριβείς δηλώσεις για τη μορφή της σχέσης και τη χρήση της για πρόβλεψη .

Απλή Γραμμική Παλινδρόμηση (Simple Linear Regression)

Μια ανάλυση γραμμικής παλινδρόμησης (linear regression analysis) μας επιτρέπει:

- Να αξιολογήσουμε τη στατιστική σημαντικότητα της σχέσης.
- Να περιγράψουμε τη σχέση μαθηματικά — δηλαδή να βρούμε την εξίσωση της γραμμής καλύτερης προσαρμογής.

Η απλή γραμμική παλινδρόμηση (Simple linear regression) μας επιτρέπει να προβλέψουμε πώς μια **εξαρτημένη μεταβλητή (y)** ανταποκρίνεται ή εξαρτάται από μια άλλη (**ανεξάρτητη μεταβλητή, x**), υποθέτοντας μια σχέση ευθείας γραμμής.

Το «απλή» (simple) σημαίνει ότι το μοντέλο λαμβάνει υπόψη μόνο μία ανεξάρτητη μεταβλητή.

Το «γραμμική» (linear) σημαίνει ότι η σχέση ακολουθεί μια ευθεία γραμμή.

Συνήθως συμβολίζουμε την **εξαρτημένη μεταβλητή** ως «y» και την **ανεξάρτητη μεταβλητή** ως «x». Στο γράφημα, η y βρίσκεται στον y-άξονα και η x στον x-άξονα.

Σε ένα πειραματικό πλαίσιο, η χειραγωγούμενη μεταβλητή είναι η **ανεξάρτητη μεταβλητή** και το μετρούμενο αποτέλεσμα είναι η **εξαρτημένη μεταβλητή**.

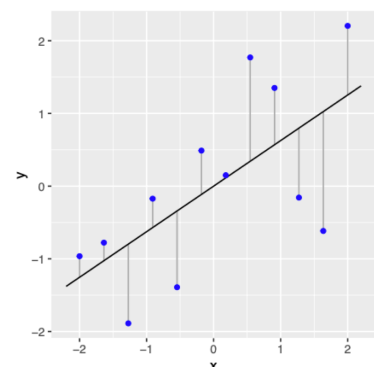
Πώς Λειτουργεί η Απλή Γραμμική Παλινδρόμηση

Εύρεση της Γραμμής Καλύτερης Προσαρμογής (Best Fit Line)

Όταν σχεδιάζουμε μια ευθεία γραμμή μέσα από τα σημεία δεδομένων, οι κάθετες αποστάσεις μεταξύ της γραμμής και κάθε σημείου ονομάζονται **υπόλοιπα (residuals)**.

Τα υπόλοιπα αντιπροσωπεύουν τη «υπόλοιπη» διακύμανση αφού η γραμμή έχει προσαρμοστεί στα δεδομένα.

Η παλινδρόμηση λειτουργεί βρίσκοντας τη γραμμή που **ελαχιστοποιεί** το μέγεθος των υπολοίπων.



Ερμηνεία μιας Παλινδρόμησης

Η εξίσωση για μια σχέση ευθείας γραμμής είναι: $y = a + b \cdot x$, όπου:

y είναι η **εξαρτημένη μεταβλητή**.

x είναι η **ανεξάρτητη μεταβλητή**.

a είναι η τομή (intercept, η τιμή του y όταν $x=0$).

b είναι η κλίση (slope) της γραμμής.

Η κλίση (b) μας λέει την ποσότητα με την οποία **αλλάζει η εξαρτημένη μεταβλητή (y)** για μια αλλαγή μιας μονάδας στην **ανεξάρτητη μεταβλητή (x)**.

Timokratis Karavitros
Bioinformatics and Applied Genomics Unit

Συσχέτιση ή Παλινδρόμηση;

Η επιλογή μεταξύ των δύο εξαρτάται από τον τρόπο συλλογής των δεδομένων και τον στόχο της ανάλυσης.

Προέλευση Δεδομένων:

--Πειραματική Μελέτη (manipulated variable): Χρησιμοποιούμε **παλινδρόμηση**, καθώς περιγράφει την αιτιώδη σχέση (η χειραγωγούμενη μεταβλητή είναι η ανεξάρτητη). Η συσχέτιση είναι γενικά ακατάλληλη.

--Μελέτη Παρατήρησης (observational study): Και οι δύο μέθοδοι μπορεί να είναι κατάλληλες.

Στόχος της Ανάλυσης:

--Αν ο στόχος είναι να βρούμε τη **μορφή** της σχέσης, την **εξίσωση** ή να κάνουμε **προβλέψεις** για την εξαρτημένη μεταβλητή, πρέπει να χρησιμοποιήσουμε **παλινδρόμηση**.

--Αν θέλουμε απλώς να γνωρίζουμε την **ισχύ** και την **κατεύθυνση** της σύνδεσης, αρκεί η **συσχέτιση**.

Timokratis Karavitros
Bioinformatics and Applied Genomics Unit

ΑΠΛΗ ΠΑΛΙΝΔΡΟΜΗΣΗ ΣΤΗΝ R

Ένας φυσιολόγος φυτών μελετά τη σχέση μεταξύ της δραστηριότητας του ενζύμου αμυλάση και της θερμοκρασίας στους βλασθήρες του φασολιού (*Vicia faba*). Διεξάγει ένα πείραμα για να βρει τη σχέση μεταξύ της απελευθέρωσης γλυκόζης (από τη διάσπαση του αμύλου από την αμυλάση) και της θερμοκρασίας (στο εύρος 2 - 20C).

Θέλουμε να δούμε αν υπάρχει μια στατιστικά σημαντική σχέση μεταξύ της θερμοκρασίας και της απελευθέρωσης γλυκόζης.



```
vicia_germ<- read.csv(file = "GLUCOSE.CSV")
```

```
glimpse(vicia_germ)
```

Το σύνολο δεδομένων περιέχει δύο **αριθμητικές** μεταβλητές:

Temperature: Η **ανεξάρτητη μεταβλητή** (experimental treatment).

Glucose: Η **εξαρτημένη μεταβλητή** (μετρούμενο αποτέλεσμα).

Οπτικοποίηση των Δεδομένων

Οπτικοποιούμε τα δεδομένα χρησιμοποιώντας ένα απλό διάγραμμα διασποράς

```
ggplot(vicia_germ, aes(x = Temperature, y = Glucose)) +  
  geom_point()
```

Το διάγραμμα διασποράς υποδηλώνει ότι η σχέση μεταξύ της **ανεξάρτητης μεταβλητής** (x) και της **εξαρτημένης μεταβλητής** (y) είναι **γραμμική**, οπότε η προσαρμογή ενός γραμμικού μοντέλου έχει νόημα.

Προσαρμογή Μοντέλου και Έλεγχοι Σημαντικότητας

Βήμα 1: Προσαρμογή του Μοντέλου (Model Fitting)

Υπολογίζουμε τη γραμμή καλύτερης προσαρμογής. Χρησιμοποιούμε τη συνάρτηση **lm (linear model)**.

```
vicia_model <- lm(Glucose ~ Temperature, data = vicia_germ)
```

Το πρώτο όρισμα είναι η **φόρμουλα** (Glucose ~ Temperature).

Η **εξαρτημένη μεταβλητή** (Glucose) πρέπει να βρίσκεται **αριστερά** του ~ (**tilde**)

Η **ανεξάρτητη μεταβλητή** (Temperature) πρέπει να βρίσκεται **δεξιά** του ~ (**tilde**)

Το δεύτερο όρισμα είναι το όνομα του data frame (vicia_germ).

Αναθέσαμε το αποτέλεσμα της προσαρμογής στο αντικείμενο vicia_model.

```
print(vicia_model)
```

Αυτό εκτυπώνει μια γρήγορη περίληψη του μοντέλου και πληροφορίες για τους **συντελεστές** (coefficients) της προσαρμοσμένης γραμμής: την τομή (Intercept) και την κλίση (Temperature), η οποία ονομάζεται με το όνομα της **ανεξάρτητης μεταβλητής**.

Βήμα 2: Έλεγχος Στατιστικής Σημαντικότητας

Συνήθως θέλουμε να προσδιορίσουμε αν η κλίση του μοντέλου διαφέρει σημαντικά από το μηδέν (δηλαδή, αν υπάρχει πραγματική σχέση).

Ο απαιτούμενος έλεγχος F test γίνεται χρησιμοποιώντας τη συνάρτηση anova:

```
anova(vicia_model)
```

Η έξοδος της anova είναι ένας Πίνακας Ανάλυσης Διακύμανσης (Analysis of Variance Table).

Ο έλεγχος F βασίζεται πάντα σε έναν λόγο δύο διακυμάνσεων (Analysis of Variance).

Ανάλυση εξόδου

Η έξοδος του ανοιχτού δίνει τον F-test, ο οποίος είναι ουσιαστικά ένας έλεγχος για το αν η κλίση του μοντέλου είναι σημαντικά διαφορετική από το μηδέν (δηλαδή, αν υπάρχει σημαντική σχέση μεταξύ της ανεξάρτητης και της εξαρτημένης μεταβλητής)

Ο πίνακας συνοψίζει τους υπολογισμούς του F-test:

Df (degrees of freedom / Βαθμοί Ελευθερίας):

Temperature: Df = 1. Αυτός είναι ο βαθμός ελευθερίας που σχετίζεται με την ανεξάρτητη μεταβλητή (Temperature), δηλαδή τη διακύμανση που εξηγείται από το μοντέλο.

Residuals: Df = 8. Αυτοί είναι οι βαθμοί ελευθερίας που σχετίζονται με την υπολειπόμενη διακύμανση (Error df), που υπολογίζεται ως $n - 2$ ($10 - 2 = 8$, όπου $n=10$ είναι το μέγεθος του δείγματος).

Sum Sq (Sum of Squares / Άθροισμα Τετραγώνων): το εξηγούμενο άθροισμα τετραγώνων (ESS), η διακύμανση στην εξαρτημένη μεταβλητή (Glucose) που εξηγείται από την ανεξάρτητη μεταβλητή (Temperature)

Residuals: το υπολειπόμενο άθροισμα τετραγώνων (RSS), η διακύμανση που δεν εξηγείται από το μοντέλο.

Mean Sq (Mean Square / Μέσο Τετράγωνο): Υπολογίζεται ως Sum Sq / Df. Το Mean Sq είναι μια εκτίμηση της διακύμανσης.

F value (F-statistic / F-στατιστική) Αυτός είναι ο λόγος διακύμανσης (F-ratio), που υπολογίζεται ως Mean Sq(Temperature) / Mean Sq(Residuals). Η F-στατιστική ποσοτικοποιεί πόση διακύμανση εξηγείται όταν συμπεριλαμβάνουμε τον όρο της κλίσης του μοντέλου. Μεγαλύτερες τιμές υποδεικνύουν ισχυρότερη σχέση μεταξύ της ανεξάρτητης και της εξαρτημένης μεταβλητής.

Pr(>F) (p-value / τιμή-p):

Αυτή είναι η πιθανότητα το παρατηρούμενο αποτέλεσμα (ή πιο ακραίο) να προέκυψε τυχαία λόγω δειγματοληψίας, αν δεν υπήρχε πραγματική σχέση (H_0 : κλίση = 0).

Συμπέρασμα Στατιστικής Σημαντικότητας

Τιμή p (0.005657): Η τιμή p είναι μικρότερη από 0.05 ($p < 0.05$).

Απόφαση: Απορρίπτουμε τη μηδενική υπόθεση.

Συμπέρασμα: Υπάρχει μια στατιστικά σημαντική σχέση μεταξύ της Temperature (ανεξάρτητη μεταβλητή) και της Glucose (εξαρτημένη μεταβλητή).

Σχεδιασμός της Προσαρμοσμένης Γραμμής και των Δεδομένων

```
pred_data <- data.frame(Temperature = seq(2, 20, length.out = 25))
```

Υπολογισμός των Προβλέψεων

Χρησιμοποιούμε τη συνάρτηση `predict` για να υπολογίσουμε τις προβλεπόμενες τιμές:

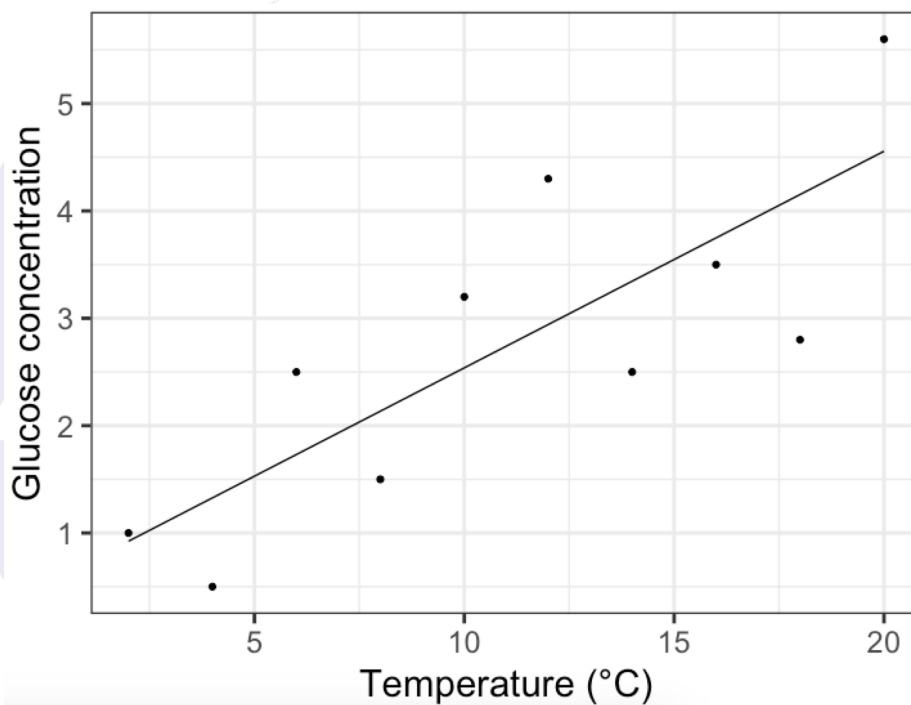
```
predict(vicia_model, pred_data)
```

Για να χρησιμοποιήσουμε αυτές τις προβλέψεις στο `ggplot2`, τις προσθέτουμε στο `data frame` `pred_data` χρησιμοποιώντας τη συνάρτηση `mutate`:

```
pred_data <- mutate(pred_data, Glucose = predict(vicia_model,  
pred_data))
```

Σχεδιάζουμε τις προβλέψεις μαζί με τα αρχικά δεδομένα:

```
ggplot(pred_data, aes(x = Temperature, y = Glucose)) +  
geom_line() + geom_point(data = vicia_germ) +  
xlab("Temperature (°C)") +  
ylab("Glucose concentration") +  
theme_bw(base_size = 22)
```



Ανάλυση Διακύμανσης Μίας Κατεύθυνσης (One-way ANOVA)

Η ANOVA χρησιμοποιείται για να συγκρίνει τις διαφορές μεταξύ των μέσων όρων μιας αριθμητικής **εξαρτημένης μεταβλητής** σε **περισσότερες από δύο ομάδες** που ορίζονται από μία **κατηγορική ανεξάρτητη μεταβλητή**.

Γιατί όχι πολλαπλά t-tests;

Η διεξαγωγή πολλών t-tests αυξάνει τον κίνδυνο εύρεσης ψευδώς θετικών αποτελεσμάτων (false positives). Η ANOVA αντιμετωπίζει αυτό το πρόβλημα με έναν μόνο, συνολικό έλεγχο.

Πώς Λειτουργεί η ANOVA (Ανάλυση Διακύμανσης)

Η ANOVA συγκρίνει τους μέσους όρους αναλύοντας και συγκρίνοντας διαφορετικές πηγές διακύμανσης στα δεδομένα. Χωρίζει τη **Συνολική Διακύμανση (Total Variation)** σε δύο κύρια μέρη:

Πηγή Διακύμανσης	Περιγραφή	Ισοδύναμο στην Παλινδρόμηση
Διακύμανση μεταξύ των Ομάδων (Between-group)	Η διακύμανση που εξηγείται από τις διαφορές μεταξύ των μέσων όρων των ομάδων (από την ανεξάρτητη μεταβλητή).	Εξηγούμενη Διακύμανση
Διακύμανση εντός των Ομάδων (Within-group)	Η διακύμανση λόγω σφάλματος ή άλλων ανεξήγητων παραγόντων, εντός κάθε ομάδας.	Διακύμανση Σφάλματος (Residual)

Ο Υπολογισμός του F-test

Η ANOVA χρησιμοποιεί τα **Αθροίσματα Τετραγώνων (Sum of Squares, SS)** και τους **Βαθμούς Ελευθερίας (df)** για να υπολογίσει τα **Μέσα Τετράγωνα (Mean Squares, MS)**, τα οποία είναι εκτιμήσεις της διακύμανσης.

Υπολογισμός MS: $MS = SS / df$.

Υπολογισμός F-ratio: Συγκρίνει την εξηγούμενη διακύμανση με τη διακύμανση σφάλματος.

$$F = \frac{\text{Treatment MS}}{\text{Residual MS}}$$

Μεγάλο F: Υποδηλώνει ότι η διακύμανση μεταξύ των ομάδων είναι μεγάλη σε σχέση με τη διακύμανση σφάλματος, άρα οι διαφορές μεταξύ των μέσων είναι **στατιστικά σημαντικές** ($p\text{-value} < \alpha$).

Συνολικός Έλεγχος: Ο F-test μας λέει μόνο αν **κάποιοι** από τους μέσους όρους διαφέρουν, όχι **ποιοι** συγκεκριμένοι μέσοι όροι διαφέρουν μεταξύ τους.

Εφαρμογή και Προϋποθέσεις

Αιτιότητα

Η ANOVA δεν συνεπάγεται αιτιότητα. Μόνο ένας **πειραματικός σχεδιασμός** με χειραγώγηση της **ανεξάρτητης μεταβλητής** και τυχαιοποίηση επιτρέπει συμπεράσματα αιτιότητας.

Ίσες Επαναλήψεις

Δεν είναι απαραίτητες. Η One-way ANOVA λειτουργεί καλά ακόμα και με άνισα μεγέθη δειγμάτων.

Σύγκριση με t-test

Η ANOVA μπορεί να χρησιμοποιηθεί και για σύγκριση δύο ομάδων. Σε αυτή την περίπτωση, δίνει το ίδιο p-value με το t-test (υπό την προϋπόθεση ίσων διακυμάνσεων).

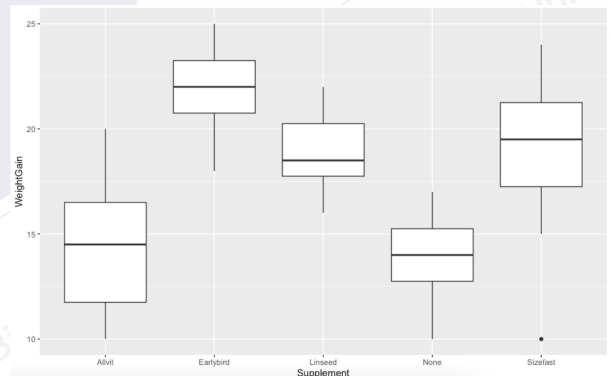
Ανάλυση Διακύμανσης Μίας Κατεύθυνσης (One-way ANOVA) στην R

Ο στόχος είναι να εφαρμόσουμε One-way ANOVA στο R χρησιμοποιώντας το παράδειγμα των νεοσσών (corncrake) όπου συγκρίνεται η **αύξηση βάρους** (WeightGain, εξαρτημένη μεταβλητή) μεταξύ 5 διαφορετικών **συμπληρωμάτων** (Supplement, ανεξάρτητη κατηγορική μεταβλητή) και μιας ομάδας ελέγχου.

Οπτικοποίηση των Δεδομένων

Πριν την προσαρμογή του μοντέλου, οπτικοποιούμε τα δεδομένα με ένα boxplot για να δούμε τη διασπορά και τους μέσους όρους κάθε ομάδας:

```
ggplot(corn_crake, aes(x = Supplement, y = WeightGain)) +  
geom_boxplot()
```



Ανάλυση Διακύμανσης Μίας Κατεύθυνσης (One-way ANOVA) στην R

Προσαρμογή του Μοντέλου (Model Fitting)

Για να προσαρμόσουμε το μοντέλο ANOVA στο R, χρησιμοποιούμε τη συνάρτηση lm (linear model), όπως και στη γραμμική παλινδρόμηση:

```
corncrake_model <- lm(WeightGain ~ Supplement, data = corn_crake)
```

Φόρμουλα: WeightGain (αριθμητική **εξαρτημένη μεταβλητή**) βρίσκεται αριστερά της ~.
Supplement (κατηγορική **ανεξάρτητη μεταβλητή**) βρίσκεται δεξιά.

Έλεγχος Σημαντικότητας (Significance Test)

Για να λάβουμε την τιμή p και να καθορίσουμε αν υπάρχει στατιστική υποστήριξη για διαφορές μεταξύ των μέσων όρων των ομάδων, χρησιμοποιούμε τη συνάρτηση anova:

```
anova(corncrake_model)
```

Πίνακας Αποτελεσμάτων ANOVA

Στήλη	Περιγραφή
Df	Βαθμοί Ελευθερίας (degrees of freedom)
Sum Sq	Άθροισμα Τετραγώνων (sum of squares)
Mean Sq	Μέσο Τετράγωνο (mean square)
F value	F-ratio (Λόγος Διακύμανσης), η F-στατιστική
Pr(>F)	Τιμή p

Timokratis Karamitros
Bioinformatics and Applied Genomics Unit

Ερμηνεία F-ratio και Τιμής p:

Ο F-ratio είναι η στατιστική ελέγχου. Μετρά πόσο μεγάλες είναι οι διαφορές μεταξύ των μέσων όρων των θεραπειών σε σχέση με το σφάλμα.

Μηδενική Υπόθεση (H0): Όλοι οι μέσοι όροι είναι ίσοι (κανένα αποτέλεσμα θεραπείας).

Αποτέλεσμα: Για τα δεδομένα των νεοσσών, $F = 9.7$ και $p = 2.06e-05$ *** ($p < 0.001$).

Συμπέρασμα: Το αποτέλεσμα είναι **στατιστικά σημαντικό**. Υπάρχει ένδειξη ότι υπάρχουν διαφορές στην αύξηση βάρους μεταξύ **τουλάχιστον κάποιων** από τις θεραπείες συμπληρώματος.

Προσοχή: Αυτός είναι ένας **συνολικός έλεγχος σημαντικότητας (global test)**. Δεν μας λέει **ποιες** συγκεκριμένες ομάδες διαφέρουν μεταξύ τους (π.χ. ποιο συμπλήρωμα είναι καλύτερο από το control). Για αυτό, απαιτούνται **post hoc multiple comparison tests** (θα παρουσιαστούν αργότερα).

Timokratis Karamitros
Bioinformatics and Applied Genomics Unit

Παρουσίαση Αποτελεσμάτων

Εμφανίζουμε τους μέσους όρους κάθε ομάδας μαζί με ένα μέτρο αβεβαιότητας, όπως το **Τυπικό Σφάλμα του Μέσου (Standard Error of the Mean, SE)**.

Υπολογισμός Μέσου και SE (dplyr):

$$SE = \frac{\text{Τυπική Απόκλιση του Δείγματος}}{\sqrt{\text{Μέγεθος Δείγματος}}}$$

```
corncrake_stats <- corn_crake %>%  
group_by(Supplement) %>%  
summarise(Mean = mean(WeightGain), SE = sd(WeightGain)/sqrt(n()))
```

Timokrits Karaitros
Bioinformatics and Applied Genomics Unit

Σχεδιασμός με Σημεία και Ράβδους Σφάλματος (geom_point και geom_errorbar):

Χρησιμοποιούμε το `corncrake_stats` (summary data) για το γράφημα.

```
ggplot(data = corncrake_stats, aes(x = Supplement, y = Mean, ymin = Mean -  
SE, ymax = Mean + SE)) +  
geom_point(colour = "blue", size = 3) +  
geom_errorbar(width = 0.1, colour = "blue") +  
coord_flip() + theme_bw()
```

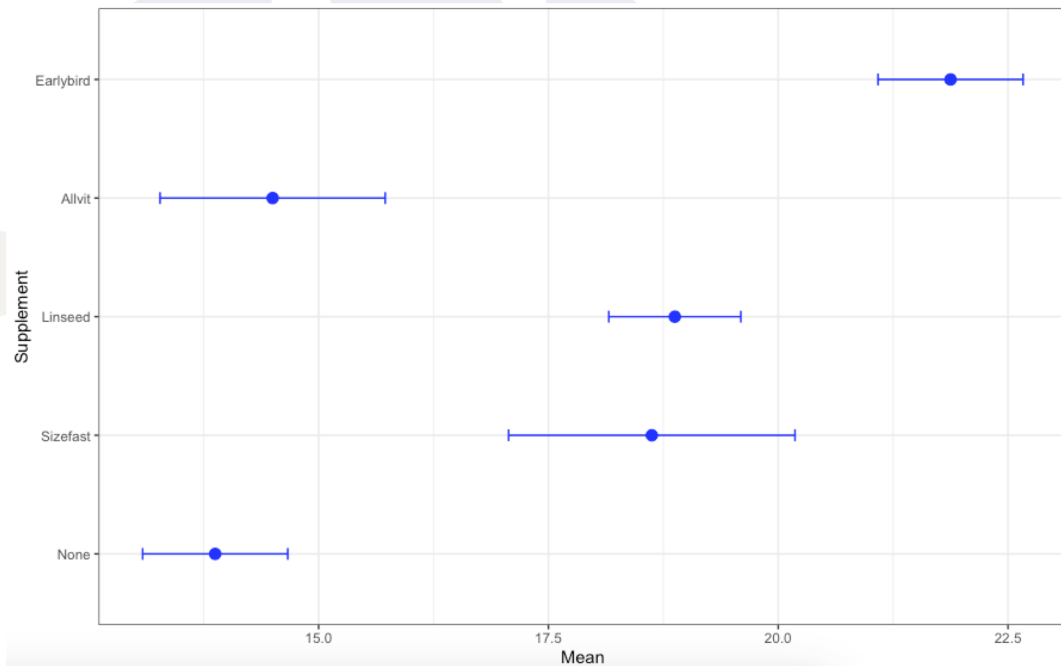
Διόρθωση Σειράς: Για να εμφανίζεται η ομάδα ελέγχου (None) πρώτη, μετατρέπουμε την **ανεξάρτητη μεταβλητή** (Supplement) σε **παράγοντα (factor)** και ορίζουμε τη σειρά (levels).

```
corncrake_stats <- corncrake_stats %>%  
mutate(Supplement = factor(Supplement, levels = c("None", "Sizefast",  
"Linseed", "Allvit", "Earlybird")))
```

```
(re - run)
```

Timokrits Karaitros
Bioinformatics and Applied Genomics Unit

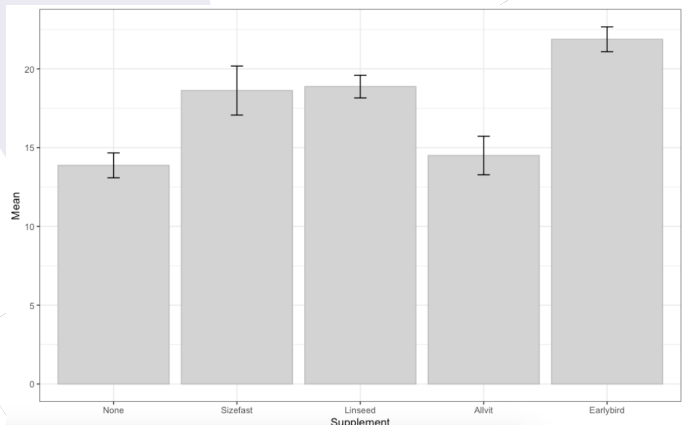
Σχεδιασμός με Σημεία και Ράβδους Σφάλματος (geom_point και geom_errorbar):



Εναλλακτικός Σχεδιασμός (Bar Plot):

Αντικαθιστούμε το `geom_point` με το `geom_col` για να δημιουργήσουμε ένα ραβδόγραμμα.

```
ggplot(data = corncrake_stats, aes(x = Supplement, y = Mean, ymin = Mean - SE, ymax = Mean + SE)) +  
  geom_col(fill = "lightgrey", colour = "grey") +  
  geom_errorbar(width = 0.1, colour = "black") +  
  theme_bw()
```



Έλεγχοι Πολλαπλών Συγκρίσεων (Multiple Comparison Tests)

Ο **συνολικός έλεγχος σημαντικότητας (global test)** της ANOVA μάς είπε μόνο ότι υπάρχει σημαντική επίδραση του συμπληρώματος στην **εξαρτημένη μεταβλητή** (WeightGain), αλλά όχι **ποιες** συγκεκριμένες ομάδες διαφέρουν. Οι **Post hoc Έλεγχοι Πολλαπλών Συγκρίσεων** χρησιμοποιούνται για να αξιολογήσουν αυτές τις διαφορές.

1. Εισαγωγή στους Post hoc Ελέγχους

- Post hoc:** Σημαίνει ότι οι συγκρίσεις γίνονται χωρίς προηγούμενη συγκεκριμένη υπόθεση, εξετάζοντας όλα τα πιθανά ζεύγη μέσων.

- Πολλαπλές Συγκρίσεις:** Αναφέρονται στο γεγονός ότι γίνονται πολλές ζευγαρωτές συγκρίσεις ταυτόχρονα, ενώ ελέγχεται το συνολικό ποσοστό ψευδώς θετικών αποτελεσμάτων.

2. Ο Έλεγχος Tukey's HSD

Ο πιο ευρέως χρησιμοποιούμενος έλεγχος είναι ο Tukey's Honestly Significant Difference (HSD) επειδή είναι **συντηρητικός**: έχει χαμηλό ποσοστό ψευδώς θετικών αποτελεσμάτων (false positive).

Εφαρμογή Tukey HSD στο R

Μετατροπή του lm σε aov αντικείμενο:

Χρησιμοποιούμε τη συνάρτηση aov στο προσαρμοσμένο μοντέλο lm.

```
corncrake_aov <- aov(corncrake_model)
```

Εκτέλεση του TukeyHSD: Εφαρμόζουμε τη συνάρτηση TukeyHSD στο αντικείμενο aov.

```
TukeyHSD(corncrake_aov, ordered = TRUE)
```

Το όρισμα ordered = TRUE ταξινομεί τους μέσους όρους της **ανεξάρτητης μεταβλητής** (Supplement) από τον μικρότερο προς τον μεγαλύτερο.

Ερμηνεία

Έξοδος TukeyHSD:

Ο πίνακας εξόδου περιέχει όλες τις ζευγαρωτές συγκρίσεις και τις ακόλουθες στήλες:

diff: Η διαφορά μεταξύ των μέσων όρων του ζεύγους.

lwr, upr: Το κάτω και άνω όριο του 95% διαστήματος εμπιστοσύνης για τη διαφορά των μέσων.

p adj: Η **προσαρμοσμένη τιμή p (adjusted p-value)** για τη σύγκριση, η οποία λαμβάνει υπόψη τις πολλαπλές συγκρίσεις.

Ερμηνεία:

Αν $p_{adj} < 0.05$ (ή το διάστημα εμπιστοσύνης **δεν περιλαμβάνει το 0**), η διαφορά είναι **στατιστικά σημαντική**.

Παράδειγμα: Η σύγκριση Earlybird - None έχει $p_{adj} = 0.0000599$ ($p < 0.05$), άρα η διαφορά είναι σημαντική.

Εφαρμογή Tukey HSD στο R με το Πακέτο agricolae

Αυτόματος Υπολογισμός Ομάδων με agricolae

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

```
HSD.test(corncrake_aov, "Supplement", console=TRUE)
```

Το πρώτο όρισμα είναι το aov αντικείμενο.

Το δεύτερο όρισμα ("Supplement") είναι το όνομα της **ανεξάρτητης μεταβλητής** (παράγοντα).

Το `console=TRUE` διασφαλίζει ότι τα αποτελέσματα εκτυπώνονται στην οθόνη.

Συμπέρασμα Letters Notation:

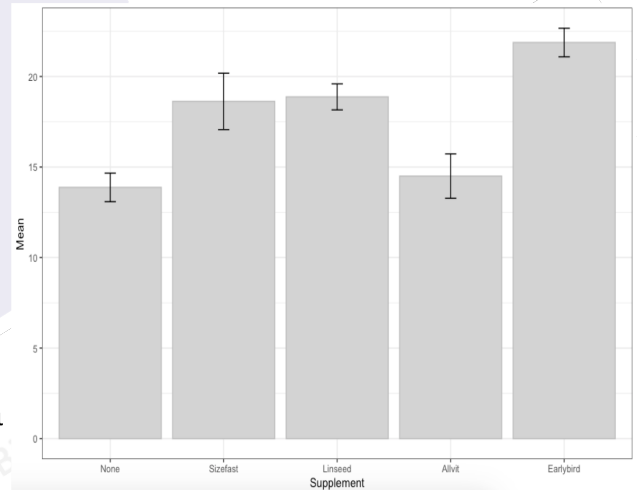
Η έξοδος του HSD.test περιλαμβάνει έναν πίνακα στο τέλος που παρουσιάζει την **ταυτότητα των ομάδων με γράμματα (letters notation)**, τα ονόματα των θεραπειών και τους μέσους όρους τους.

Η έξοδος του HSD.test περιλαμβάνει έναν πίνακα στο τέλος που παρουσιάζει την **ταυτότητα των ομάδων με γράμματα (letters notation)**, τα ονόματα των θεραπειών και τους μέσους όρους τους.

WeightGain groups		
Earlybird	21.875	a
Linseed	18.875	a
Sizefast	18.625	ab
Allvit	14.500	bc
None	13.875	c

Θεραπείες που μοιράζονται ένα κοινό γράμμα (π.χ. Sizefast και Linseed μοιράζονται το a **δεν διαφέρουν σημαντικά** μεταξύ τους (Tukey test, $p > 0.05$)).

Θεραπείες που δεν μοιράζονται κανένα γράμμα (π.χ. Earlybird (a) και Allvit (bc), None (c)) **διαφέρουν σημαντικά** μεταξύ τους ($p < 0.05$)).



Παρουσίαση των Αποτελεσμάτων

Κατά την παρουσίαση των αποτελεσμάτων σε μια έκθεση, είναι απαραίτητο να συνδυάσετε τα ευρήματα του **συνολικού ελέγχου ANOVA** με τις λεπτομέρειες του ελέγχου **Tukey HSD**.

Πρέπει να παρέχεται μια επεξήγηση (π.χ. στη λεζάντα του πίνακα ή του σχήματος) που να δηλώνει:

Ποιος έλεγχος χρησιμοποιήθηκε (Tukey test).

Τι σημαίνουν οι κωδικοί γραμμάτων ("Means followed by the same letter did not differ significantly").

Το κρίσιμο επίπεδο σημαντικότητας ($p > 0.05$)

R - Programming in Biosciences

Dr. Timokratis Karamitros
Researcher

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

tkaram@pasteur.gr

