

R - Biostatistics in Biosciences

Dr. Timokratis Karamitros
Researcher

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

tkaram@pasteur.gr



Αρχές Πειραματικού Σχεδιασμού

Ο θεμελιώδης στόχος του πειραματικού σχεδιασμού είναι:

Να εξαχθούν δεδομένα από ένα σύστημα έτσι ώστε η διακύμανση στα δεδομένα να μπορεί να αποδοθεί απερίφραστα στη συγκεκριμένη διαδικασία που διερευνούμε.

Αυτό επιτυγχάνεται μεγιστοποιώντας τη **Στατιστική Ισχύ (Statistical Power)**.

Στατιστική Ισχύς (Statistical Power)

Η **Στατιστική Ισχύς** είναι η πιθανότητα μια μελέτη να ανιχνεύσει μια **επίδραση (effect)** όταν αυτή η επίδραση είναι πραγματικά παρούσα.

Επηρεάζεται από:

Το μέγεθος της επίδρασης: Οι μεγάλες επιδράσεις ανιχνεύονται ευκολότερα.

Το μέγεθος του δείγματος: Τα μεγάλα δείγματα προσφέρουν μεγαλύτερη ευαισθησία στον έλεγχο.

Η μεταβλητότητα των δεδομένων: Όσο λιγότερο μεταβλητά είναι τα δεδομένα, τόσο μικρότερες επιδράσεις μπορούμε να ανιχνεύσουμε.

Για να μεγιστοποιήσουμε την ισχύ, πρέπει να κάνουμε δύο πράγματα:

Χρήση μέγιστου εφικτού μεγέθους δείγματος (Μεγιστοποίηση n).

Λήψη μέτρων για την ελαχιστοποίηση της μεταβλητότητας στα δεδομένα (Μείωση του σφάλματος).

Βασική Ορολογία Πειραματικού Σχεδιασμού

Πειραματική Μονάδα (Experimental Unit): Η φυσική οντότητα στην οποία εφαρμόζεται η θεραπεία (π.χ. ένα μεμονωμένο φυτό, ένα ζώο).

Θεραπεία (Treatment): Οποιαδήποτε χειραγώγηση εφαρμόζεται στις πειραματικές μονάδες (π.χ. ένα συγκεκριμένο διατροφικό συμπλήρωμα).

Ομάδα Θεραπείας (Treatment Group): Μια ομάδα πειραματικών μονάδων που λαμβάνουν την ίδια θεραπεία.

Ομάδα Ελέγχου (Control Group): Συμπληρωματική ομάδα που δεν λαμβάνει καμία θεραπεία ή λαμβάνει μια τυπική θεραπεία (πχ placebo).

Πειραματικός Παράγοντας (Experimental Factor): Μια συλλογή σχετιζόμενων θεραπειών και ελέγχων (π.χ. "Διατροφικό Συμπλήρωμα").

Επίπεδα του Παράγοντα (Levels of the Factor): Οι διαφορετικές θεραπείες / έλεγχοι που αποτελούν τον παράγοντα (π.χ. τα τέσσερα διαφορετικά συμπληρώματα συν ο έλεγχος "κανένα").

Αντιγραφή/Επανάληψη (Replication)

Η Επανάληψη είναι η διαδικασία ανάθεσης πολλών πειραματικών μονάδων στην ίδια θεραπεία (ή συνδυασμό θεραπειών - treatments).

Σημασία: Η Επανάληψη αυξάνει το μέγεθος του δείγματος, το οποίο με τη σειρά του **αυξάνει τη στατιστική ισχύ** του ελέγχου, επιτρέποντάς μας να ανιχνεύσουμε τις επιδράσεις, ειδικά στη βιολογία όπου το υλικό είναι εγγενώς μεταβλητό.

Βασική Ιδέα: Αυξημένη Επανάληψη = περισσότερη στατιστική ισχύς.

Timokratis Karamitros
Bioinformatics and Applied Genomics Unit

Ανεξαρτησία και Ψευδοαναπαραγωγή (Pseudoreplication)

Η Αρχή της Ανεξαρτησίας

Οι περισσότεροι στατιστικοί έλεγχοι (όπως το t-test και η ANOVA) προϋποθέτουν ότι τα δεδομένα είναι **ανεξάρτητα**.

Ανεξαρτησία: Σημαίνει ότι η τιμή μιας μέτρησης από ένα αντικείμενο **δεν επηρεάζεται** από τις τιμές άλλων αντικειμένων.

Συνήθεις Πηγές Μη-Ανεξαρτησίας στη Βιολογία:

Γενετική: Π.χ., ποντίκια από την ίδια γέννα είναι πιο πιθανό να είναι παρόμοια μεταξύ τους.

Γεωγραφία: Π.χ., δείγματα από κοντινές τοποθεσίες έχουν παρόμοιο μικροκλίμα ή τύπο εδάφους.

Δειγματοληψία εντός Βιολογικών "Μονάδων": Π.χ., τα φύλλα στο ίδιο δέντρο είναι πιο παρόμοια μεταξύ τους από ό,τι τα φύλλα από διαφορετικά δέντρα.

Πειραματικές Διατάξεις Εργαστηρίου: Π.χ., φυτά που καλλιεργούνται στην ίδια γλάστρα ή ψάρια που κρατούνται στο ίδιο ενυδρείο επηρεάζονται από τις ίδιες συνθήκες.

Timokratis Karamitros
Bioinformatics and Applied Genomics Unit

Ψευδοαναπαραγωγή (Pseudoreplication)

Η **Μη-Ανεξαρτησία** οδηγεί στην **Ψευδοαναπαραγωγή (Pseudoreplication)**, η οποία είναι μια τεχνητή αύξηση του μεγέθους του δείγματος που προκαλείται από τη χρήση μη-ανεξάρτητων δεδομένων.

Παράδειγμα: Μέτρηση Πετάλων Φυτών

Ενδιαφερόμαστε για το αν τα φυτά παράγουν διαφορετικό αριθμό πετάλων σε δύο διαφορετικούς τύπους εδάφους (Α και Β). Έχουμε 3 φυτά σε κάθε έδαφος, και κάθε φυτό παράγει 4 άνθη

Σωστός Έλεγχος (Ανεξάρτητα Δεδομένα)

Πειραματική Μονάδα: Το Φυτό.

Δειγματοληψία: Μέτρηση ενός άνθους ανά φυτό.

Σύνολο Ανεξάρτητων Δειγμάτων: $n = 3$ σε κάθε έδαφος.

Αποτέλεσμα: $p = 0.29$ (Μη σημαντικό).

Συμπέρασμα: Το πρώτο τεστ είναι **σωστό**. Τα άνθη εντός του ίδιου φυτού **δεν είναι ανεξάρτητα** (πιθανόν για γενετικούς λόγους), άρα το **φαινόμενο της σημαντικότητας** στο δεύτερο τεστ είναι **ψευδές** (spurious).

Λάθος Έλεγχος (Ψευδοαναπαραγωγή)

Πειραματική Μονάδα (Λάθος): Το Άνθος.

Δειγματοληψία: Μέτρηση όλων των ανθέων (4) ανά φυτό.

Σύνολο Φαινομενικών Δειγμάτων: $n = 12$ σε κάθε έδαφος.

Αποτέλεσμα: $p = 0.009$ (Φαινομενικά σημαντικό).

Αν είχαμε μόνο **ένα φυτό** σε κάθε έδαφος, και μετρούσαμε το ύψος του **20 φορές**, αυτές οι 20 επαναλαμβανόμενες μετρήσεις **δεν** παρέχουν καμία επιπλέον πληροφορία για την επίδραση του **τύπου εδάφους**. Παρέχουν μόνο πληροφορίες για τη μεταβλητότητα της τεχνικής μέτρησής μας. Η ανεξάρτητη μονάδα παραμένει το **ένα φυτό** ανά έδαφος.

Διακρίνοντας την Ψευδοαναπαραγωγή

Παράδειγμα: Αφθονία Ζωοπλαγκτόν

Στόχος: Να εξεταστεί η επίδραση των **ψαριών** στην αφθονία του **ζωοπλαγκτόν**.

Πείραμα (Λάθος Σχεδιασμός): Λαμβάνονται πολλαπλά δείγματα από μία λίμνη με ψάρια και από μία λίμνη χωρίς ψάρια.

Λανθασμένο Συμπέρασμα: Αποδίδουμε τη διαφορά στην αφθονία του ζωοπλαγκτόν στην **παρουσία ψαριών**.

Το Σφάλμα: Δεν έχουμε μετρήσει την επίδραση των ψαριών, αλλά τη **διαφορά μεταξύ των δύο συγκεκριμένων λιμνών**. Οι δύο λίμνες μπορεί να διαφέρουν σε πολλούς άλλους παράγοντες (π.χ. διατροφή, κλίμα, pH), όχι μόνο στην παρουσία ψαριών.

Σωστός Σχεδιασμός: Για να ελέγξουμε την επίδραση των ψαριών, οι **ανεξάρτητες αντιγραφές** πρέπει να είναι **ολόκληρες λίμνες**. Χρειαζόμαστε δεδομένα από **αρκετές λίμνες με ψάρια** και **αρκετές λίμνες χωρίς ψάρια**.

Πειραματική Μονάδα (Σωστή): Η **Λίμνη** (με ή χωρίς ψάρια).

Πειραματική Μονάδα (Λάθος): Το **Δείγμα** εντός της λίμνης.

Διαφορά μεταξύ Επαναλαμβανόμενων Μετρήσεων και Πραγματικής Αντιγραφής

Η Αξία των Πολλών Δειγμάτων εντός της Μονάδας

Είναι **αληθές** ότι είναι καλύτερο να λαμβάνουμε πολλά δείγματα (samples) ή μετρήσεις από κάθε μονάδα/τοποθεσία (π.χ. 20 δείγματα ζωοπλαγκτόν από μία λίμνη) παρά μόνο ένα.

Όφελος: Η λήψη πολλών μετρήσεων από το ίδιο αντικείμενο προστατεύει από την πιθανότητα τα αποτελέσματα να επηρεαστούν από ένα μόνο δείγμα. Δίνει μια **πιο ακριβή εικόνα** της συγκεκριμένης μονάδας.

Προσοχή: Αυτό είναι σημαντικό για την ακρίβεια της μέτρησης της μονάδας, αλλά **δεν είναι** το ίδιο με το να έχουμε μετρήσεις από περισσότερα **ανεξάρτητα αντικείμενα** (π.χ. λίμνες).

Ποια είναι η βιολογική μονάδα αντιγραφής που είναι σχετική με την επίδραση που προσπαθούμε να ελέγξουμε;

Ερώτημα Ελέγχου

Διαφορά Πυκνότητας Πλαγκτόν
μεταξύ Δύο Συγκεκριμένων Λιμνών
(Α και Β)

Επίδραση των Ψαριών στην
Πυκνότητα Πλαγκτόν

Σωστή Μονάδα Αντιγραφής

Δείγματα εντός της λίμνης

Ολόκληρη η Λίμνη (με ή χωρίς
ψάρια)

Σχεδιασμός

Λήψη 10 δειγμάτων από τη Λίμνη Α
και 10 από τη Λίμνη Β και σύγκρισή
τους.

Λήψη δειγμάτων από **πολλές λίμνες**
με ψάρια και από **πολλές λίμνες**
χωρίς ψάρια.

Η **Πραγματική Αντιγραφή (True Replication)** αυξάνει τη **στατιστική ισχύ** του ελέγχου για την
ανίχνευση διαφορών μεταξύ **πειραματικών μονάδων** ως προς τον παράγοντα που μας ενδιαφέρει
(π.χ. η παρουσία/απουσία ψαριών).

Εν ολίγοις, οι **πολλαπλές μετρήσεις** εντός μιας μονάδας αυξάνουν την **ακρίβεια** της μέτρησης, αλλά
η **αντιγραφή σε πολλαπλές ανεξάρτητες μονάδες** αυξάνει τη **στατιστική ισχύ** για τη γενίκευση
συμπερασμάτων σχετικά με την επίδραση ενός παράγοντα.

Πειράματα με Συγχυτικούς Παράγοντες και Θόρυβο

Μορφές Ανεπιθύμητης Διακύμανσης

Μορφή Διακύμανσης

Συγχυτική Διακύμανση (Confounding Variation)

Θόρυβος (Noise)

Περιγραφή

Προκύπτει όταν μία ή
περισσότερες άλλες πηγές
διακύμανσης λειτουργούν
παράλληλα με τον παράγοντα
που διερευνάται.

Περιγράφει τη διακύμανση που
είναι **άσχετη** με τον παράγοντα
που διερευνάται.

Πρόβλημα

Καθιστά δύσκολη ή αδύνατη την
απερίφραστη απόδοση των
παρατηρούμενων επιδράσεων σε μία μόνο
αιτία.

Αυξάνει τη συνολική μεταβλητότητα των
αποτελεσμάτων, καθιστώντας
δυσκολότερη τη **στατιστική ανίχνευση** της
επίδρασης του παράγοντα.

Συγχυτικοί Παράγοντες (Confounding)

Σε Μελέτες Παρατήρησης: Όταν πολλοί παράγοντες αλλάζουν συστηματικά μαζί.

Παράδειγμα: Η μέτρηση των ρυθμών ανάπτυξης φυτών σε διαφορετικά υψόμετρα. Το υψόμετρο συνδέεται συστηματικά με τη θερμοκρασία, την ακτινοβολία UV, την κατακρήμνιση και την ταχύτητα του ανέμου. Οποιαδήποτε σχέση με την ανάπτυξη δεν μπορεί να αποδοθεί σε έναν μόνο παράγοντα.

Σε Πειραματικές Μελέτες (Procedural Confounding): Όταν η εφαρμογή μιας θεραπείας παράγει από μόνη της άλλες ανεπιθύμητες επιδράσεις.

Timokratis Karamitros
Bioinformatics and Applied Genomics Unit

Θόρυβος (Noise)

Ο θόρυβος δημιουργείται από τις ίδιες διαδικασίες που δημιουργούν τη σύγχυση, αλλά συμβαίνει ακόμα και όταν οι συγχυτικοί παράγοντες δεν **ευθυγραμμίζονται** με τις θεραπείες.

Διαφορά με τη Σύγχυση:

Σύγχυση: Όταν οι υπο-ομάδες φυτών αναπτύσσονται **κυρίως σε διαφορετικά υψόμετρα** (η διακύμανση λόγω υψομέτρου συγχέεται με τη διακύμανση λόγω υπο-είδους).

Θόρυβος: Όταν τα δείγματα των υπο-ομάδων είναι **διασκορπισμένα σε πολύ διαφορετικά υψόμετρα** (η διακύμανση λόγω υψομέτρου προσθέτει ανεπιθύμητη διακύμανση — **θόρυβο** — στις εκτιμήσεις του ρυθμού ανάπτυξης).

Timokratis Karamitros
Bioinformatics and Applied Genomics Unit

Διαχείριση Συγχυτικών Παραγόντων και Θορύβου

Τέσσερις βασικές τεχνικές χρησιμοποιούνται για την αντιμετώπιση αυτών των ανεπιθύμητων επιδράσεων:

Τυχαία κατανομή (randomisation)

Ομαδοποίηση σε block (blocking)

Πειραματικός έλεγχος

Πρόσθετες θεραπείες

Timokratis Karamitros
Bioinformatics and Applied Genomics Unit

Τυχαία Κατανομή (Randomisation)

Η τυχαία κατανομή είναι θεμελιώδης αρχή του πειραματικού σχεδιασμού και χρησιμεύει ως «ασφάλεια» έναντι απρόβλεπτων συγχυτικών παραγόντων.

Αρχή: Κάθε πειραματική μονάδα πρέπει να επιλέγεται ή να κατανέμεται σε μια συγκεκριμένη θεραπεία "τυχαία".

Παράδειγμα: Τυχαία επιλογή ασθενών που θα λάβουν φάρμακο έναντι εικονικού φαρμάκου, ή τυχαία τοποθέτηση πειραματικών αγροτεμαχίων σε ένα χωράφι.

Οφέλη: Η τυχαία κατανομή προστατεύει από διάφορες πιθανές μεροληψίες και συγχυτικούς παράγοντες, συμπεριλαμβανομένων των ακούσιων μεροληψιών που εισάγονται κατά την εγκατάσταση του πειράματος.

Σχέση με το Θόρυβο: Η τυχαία κατανομή **δεν μειώνει** τον θόρυβο. Στην πραγματικότητα, εάν η τυχαία κατανομή αφαιρέσει αποτελεσματικά τη σύγχυση, μπορεί να **φαίνεται** ότι αυξάνει τη διακύμανση, αλλά αυτό είναι ένα απαραίτητο τίμημα για τη σωστή ερμηνεία των επιδράσεων της θεραπείας.

Timokratis Karamitros
Bioinformatics and Applied Genomics Unit

Ομαδοποίηση σε blocks (Blocking)

Το Blocking είναι μια τεχνική οργάνωσης του πειραματικού υλικού που αντιμετωπίζει τη δυνητική **συγχυτική διακύμανση** και τον **θόρυβο** που προκαλείται από την εγγενή ετερογένεια του βιολογικού υλικού. Θεωρείται η σημαντικότερη έννοια πειραματικού σχεδιασμού μετά την αντιγραφή.

Ομαδοποίηση σε Μπλοκ: Ομαδοποιούμε τα υπό μελέτη αντικείμενα σε **μπλοκ** έτσι ώστε:

Η διακύμανση **εντός** των μπλοκ να είναι **μικρή**.

Η διακύμανση **μεταξύ** των μπλοκ να είναι **μεγάλη** (αναμενόμενη).

Αντιπροσώπευση Θεραπειών: Κάθε θεραπεία πρέπει να εφαρμόζεται **τουλάχιστον μία φορά** μέσα σε κάθε μπλοκ.

Παράδειγμα: Διαφορετικές Δίαιτες σε Ποντίκια

Σε ένα πείραμα με τρεις δίαιτες (I, II, III) και πέντε γέννες ποντικών (A...E), οι οποίες αναμένεται να έχουν διαφορετικές αποκρίσεις:

Μπλοκ: Οι **γέννες** (A, B, C, D, E).

Σχεδιασμός: Από κάθε γέννα, επιλέγουμε τρία ποντίκια (A1, A2, A3) και αναθέτουμε το καθένα τυχαία σε μία από τις τρεις δίαιτες (I, II, III).

Θεραπεία	Γέννα A (Μπλοκ 1)	Γέννα B (Μπλοκ 2)	...	Γέννα E (Μπλοκ 5)
I	A1	B1	...	E1
II	A2	B2	...	E2
III	A3	B3	...	E3

Αυτός ο σχεδιασμός **εγγυάται** ότι αυξάνει τη στατιστική ισχύ του πειράματος για την ανίχνευση των επιδράσεων των θεραπειών, επειδή η διακύμανση λόγω της γέννας (inter-litter variation) μπορεί να ελεγχθεί και να αφαιρεθεί κατά την ανάλυση.

Σχέση με άλλους Ελέγχους

Στην περίπτωση μόνο δύο θεραπειών, το blocking είναι ισοδύναμο με τον **t-test ζευγαρωμένων δειγμάτων (paired-sample t-test)**.

Σε περιπτώσεις με περισσότερα από δύο μπλοκ, η ανάλυση γίνεται συνήθως με **ANOVA** (Blocked Experimental Design ANOVA).

Η τυχαία κατανομή εξακολουθεί να είναι σημαντική:

Τα ποντίκια επιλέγονται **τυχαία** από κάθε γέννα για να κατανεμηθούν στις θεραπείες.

Οι γέννες θεωρούνται **τυχαίες** (δεν επιλέχθηκαν σκόπιμα για να διαφέρουν, αλλά αναμένεται να υπάρχει κάποια διαφορά).

Timokratis Karanikolas
Bioinformatics and Applied Genomics Unit

Πειραματικός Έλεγχος (Experimental Control)

Έλεγχος Πρωτοκόλλου και Ακρίβεια Μετρήσεων

Η πρώτη μορφή ελέγχου αφορά τη βελτίωση του πρωτοκόλλου της μελέτης και την ακρίβεια των μετρήσεων.

Ακρίβεια: Πρέπει να εξετάζουμε την ακρίβεια των μετρήσεων σε σχέση με το μέγεθος των επιδράσεων που μας ενδιαφέρουν.

Παράδειγμα: Η μέτρηση του ύψους ενός σπορόφυτου πρέπει να γίνεται με τυπική προσέγγιση (standard approach) και με την κατάλληλη ακρίβεια (π.χ. στο πλησιέστερο χιλιοστό, όχι εκατοστό), αλλά δεν έχει νόημα να μετράμε το χρονικό διάστημα ανάπτυξης σε χιλιοστά του δευτερολέπτου.

Timokratis Karanikolas
Bioinformatics and Applied Genomics Unit

Έλεγχος των Συνθηκών

Η δεύτερη μορφή ελέγχου περιλαμβάνει τη χειραγώγηση του πειραματικού περιβάλλοντος ώστε όλες οι πειραματικές μονάδες να βιώνουν τις ίδιες συνθήκες.

Ομοιομορφία Συνθηκών: Στην απλούστερη περίπτωση, αυτό σημαίνει τον έλεγχο εξωτερικών παραγόντων (π.χ. θερμοκρασία) έτσι ώστε όλες τα treatments να λαμβάνουν χώρα σε **πανομοιότυπες** συνθήκες.

Σταθερότητα: Δεν είναι πάντα απαραίτητο οι συνθήκες να είναι απόλυτα σταθερές, αλλά αρκεί όποια διακύμανση και αν συμβαίνει να είναι **ίδια για όλες τις θεραπείες**.

Αντιμετώπιση Διαδικαστικής Σύγχυσης (Procedural Confounding)

Πιο σύνθετα προβλήματα προκύπτουν όταν η ανεπιθύμητη διακύμανση παράγεται ως **παράγωγο της θεραπείας** που εφαρμόζουμε (procedural confounding).

Παράδειγμα (Αποσύνθεση Φυλλικής Στρώσης):

Θεραπεία: Μεταβολή της ποσότητας φυλλικής στρώσης που τοποθετείται στο έδαφος για να μελετήσουμε την επίδραση στην μικροβιακή κοινότητα.

Συγχυτικός Παράγοντας: Η φυσική παρουσία της στρώσης φύλλων επηρεάζει το **μικροκλίμα** στην επιφάνεια του εδάφους (π.χ. ξηρότητα).

Λύση Ελέγχου: Δημιουργία μιας **τεχνητής μη-αποσυντιθέμενης στρώσης** (π.χ. artificial litter) η οποία αναμειγνύεται με την πραγματική στρώση. Με αυτόν τον τρόπο, κάθε αγροτεμάχιο έχει **σταθερό φυσικό όγκο** φυλλικής στρώσης (έλεγχος μικροκλίματος), αλλά **διαφορετική ποσότητα** αποσυντιθέμενου υλικού (η θεραπεία).

Paired-Sample t-test (Έλεγχος t Ζευγαρωμένων Δειγμάτων)

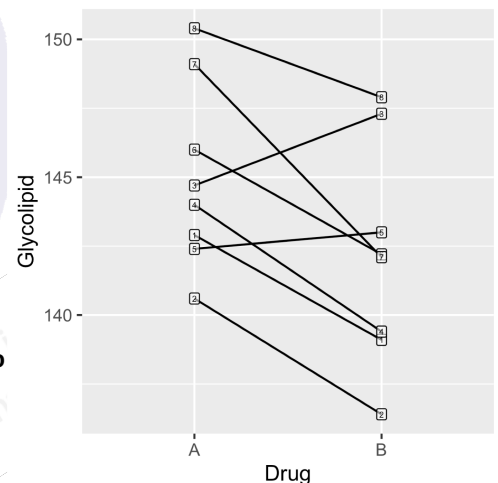
Ο έλεγχος t ζευγαρωμένων δειγμάτων (Paired-sample t-test) είναι η κατάλληλη μέθοδος ανάλυσης για δεδομένα που προέρχονται από **σχεδιασμό με ομαδοποίηση σε Μπλοκ** ή από **επαναλαμβανόμενες μετρήσεις** στο ίδιο υποκείμενο, καθώς λαμβάνει υπόψη τη μη-ανεξαρτησία.

Πότε Χρησιμοποιούμε το Paired-Sample t-test;

Χρησιμοποιούμε τον έλεγχο t ζευγαρωμένων δειγμάτων όταν τα δεδομένα σχηματίζουν **ζεύγη μη-ανεξάρτητων παρατηρήσεων**.

Ορισμός Paired-Sample Data: Η πρώτη τιμή στο Δείγμα A συνδέεται με την πρώτη τιμή στο Δείγμα B, η δεύτερη τιμή στο A με τη δεύτερη τιμή στο B, κ.ο.κ.

Κοινό Παράδειγμα: Μέτρηση συγκέντρωσης γλυκολιπιδίων στο **ίδιο άτομο** πριν και μετά τη θεραπεία, ή υπό δύο διαφορετικές συνθήκες (π.χ. **φάρμακο A vs φάρμακο B**).



Γιατί ο Paired-Sample σχεδιασμός είναι χρήσιμος;

Στη βιολογία, υπάρχει συχνά μεγάλη διακύμανση **μεταξύ των ατόμων - μονάδων** (among-unit variation) (π.χ. μεταξύ ασθενών). Αυτή η διακύμανση μπορεί να **επισκιάσει** (obscure) την επίδραση της θεραπείας που μας ενδιαφέρει.

Μείωση Διακύμανσης: Ο σχεδιασμός ζευγαρωμένων δειγμάτων λειτουργεί ως **έλεγχος (control) για τη διακύμανση μεταξύ των ατόμων - μονάδων** (π.χ. διαφορές λόγω ηλικίας, φύλου, βάρους των ασθενών). Εστιάζοντας στη διαφορά **εντός** του κάθε ατόμου, αυξάνουμε τη **στατιστική ισχύ** του ελέγχου.

Παράδειγμα Φαρμάκων Γλυκολιπιδίων:

Non-Paired Design (Λάθος): 8 ασθενείς χωρίζονται τυχαία σε 2 ομάδες (n=4 έκαστη) και λαμβάνουν διαφορετικό φάρμακο (A ή B). Η μεγάλη διακύμανση στα επίπεδα γλυκολιπιδίων μεταξύ των ασθενών θα οδηγούσε σε **μεγάλο τυπικό σφάλμα** στη διαφορά των μέσων, καθιστώντας πιθανότατα τη διαφορά **μη-σημαντική**.

Paired Design (Σωστό): Και οι 8 ασθενείς λαμβάνουν και τα δύο φάρμακα.

Όφελος: Παρόλο που τα επίπεδα γλυκολιπιδίων διαφέρουν πολύ μεταξύ των ασθενών, το Paired Design μας επιτρέπει να **αφαιρέσουμε** (factor out) τη διακύμανση μεταξύ των ατόμων και να επικεντρωθούμε στις **διαφορές μεταξύ των δύο θεραπειών** εντός του ίδιου ατόμου.

Πώς Πραγματοποιείται το Paired-Sample t-test;

Ο έλεγχος t ζευγαρωμένων δειγμάτων **δεν είναι ένας νέος τύπος ελέγχου**. Είναι απλώς μια εφαρμογή του **one-sample t-test** (έλεγχος t ενός δείγματος) στη διαφορά των ζευγαριών.

Η «τεχνική» είναι να εργαστούμε απευθείας με τις **διαφορές** μεταξύ των ζευγαριών τιμών:

Υπολογισμός των Διαφορών: Υπολογίζουμε τη διαφορά (D) για κάθε ζεύγος παρατηρήσεων (π.χ. $D = \text{Drug B} - \text{Drug A}$) για κάθε ασθενή.

Υπολογισμός του Μέσου της Διαφοράς: Βρίσκουμε τον μέσο όρο όλων αυτών των διαφορών.

Έλεγχος Σημαντικότητας: Ελέγχουμε αν ο μέσος όρος των διαφορών είναι **στατιστικά διαφορετικός από το μηδέν**.

$H_0: \text{mean_D} = 0$ (Δεν υπάρχει διαφορά μεταξύ των φαρμάκων)

Εάν τα δύο φάρμακα είχαν παρόμοια επίδραση, θα περιμέναμε ο μέσος όρος των διαφορών (mean_D) να είναι περίπου μηδέν.

(προ-)Υποθέσεις του Paired-Sample t-test

Οι υποθέσεις του paired-sample t-test είναι ίδιες με αυτές του one-sample t-test, αλλά εφαρμόζονται στο **δείγμα των διαφορών** και όχι στα αρχικά δεδομένα:

-- **Κλίμακα Μέτρησης:** Οι διαφορές πρέπει να μετρώνται σε κλίμακα **διαστήματος** ή **αναλογίας** (πράγμα που εξασφαλίζεται αν τα αρχικά δεδομένα είναι σε αυτές τις κλίμακες).

-- **Κανονικότητα:** Το **δείγμα των διαφορών** πρέπει να προέρχεται από μια κανονική κατανομή.

Σημαντικό: Δεν υπάρχει απαίτηση τα **αρχικά δεδομένα** να προέρχονται από κανονική κατανομή. Ακόμη και αν τα αρχικά δεδομένα δεν είναι κανονικά, οι διαφορές μεταξύ των ζευγαριών είναι συχνά αποδεκτά κανονικές.

Paired-sample t-test στην R

Η R προσφέρει τη δυνατότητα paired-sample t-test, ώστε να μη χρειάζεται να υπολογίζουμε χειροκίνητα τις διαφορές. Υπολογίζει αυτόματα τις διαφορές μεταξύ των ζευγών και εκτελεί ένα one-sample t-test πάνω σε αυτές. Στο παράδειγμα με τα δεδομένα glycolipid θα δούμε πρώτα την παραδοσιακή και έπειτα τη σύντομη μέθοδο της R.

```
glycolipid<-read.csv("GLYCOLIPID.CSV")  
glimpse(glycolipid)
```

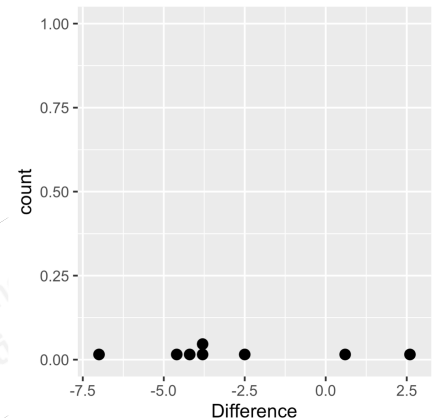
Υπολογισμός των Διαφορών

Αρχικά υπολογίζουμε τις διαφορές ανά ασθενή και εκτελούμε one-sample t-test στις διαφορές.

```
glycolipid_diffs <-  
  glycolipid %>%  
  group_by(Patient) %>%  
  summarise(Difference = diff(Glycolipid))
```

```
glycolipid_diffs
```

```
ggplot(glycolipid_diffs, aes(x = Difference)) +  
  geom_dotplot() + theme_grey(base_size = 22)
```



Εκτέλεση one-sample t-test στις διαφορές

```
t.test(glycolipid_diffs$Difference)
```

Ερμηνεία αποτελεσμάτων t-test

Το output του t-test περιλαμβάνει το είδος του τεστ, τα δεδομένα και τα βασικά στατιστικά μεγέθη.

$t = -2.6209$, $df = 7$, $p\text{-value} = 0.03436$

Το p-value δείχνει ότι η μέση ενδο-ατομική διαφορά είναι στατιστικά σημαντική στο επίπεδο $p < 0.05$.

Reporting:

«Individual patients had significantly lower serum glycolipid concentrations when treated with Drug B than when treated with Drug A ($t = 2.62$, $d.f. = 7$, $p < 0.05$).»

Στο paired-sample t-test:

Το τεστ βασίζεται στο δείγμα των διαφορών και όχι στα αρχικά ζεύγη μετρήσεων

Οι βαθμοί ελευθερίας ισούνται με τον αριθμό των ζευγών μείον ένα

Το συμπέρασμα αφορά αλλαγές εντός των ίδιων ατόμων και όχι διαφορές μεταξύ ομάδων

Η διεπαφή Pair() στην R

Από την έκδοση 4.5, η R προσφέρει τη διεπαφή Pair() για paired t-tests. Η διεπαφή αυτή καθιστά ρητό ποιές δύο μεταβλητές σχηματίζουν ζεύγος.

Για να τη χρησιμοποιήσουμε, τα δεδομένα πρέπει να είναι σε wide format. (Χρησιμοποιούμε τη συνάρτηση pivot_wider από το πακέτο tidyr), και διαλέγουμε τις μεταβλητές που χρειαζόμαστε)

```
library(tidyr)
glycolipid_wide <- glycolipid %>%
  select(Patient, Drug, Glycolipid) %>%
  pivot_wider(
    names_from = Drug,
    values_from = Glycolipid
  )

glycolipid_wide
```

Timokratis Karamitros
Bioinformatics and Applied Genomics Unit

Paired t-test με Pair()

Αφού τα δεδομένα είναι σε wide format, εκτελούμε το paired t-test με τη σύνταξη Pair().

```
t.test(Pair(B, A) ~ 1, data = glycolipid_wide)
```

Συμπέρασμα

Η διεπαφή Pair() δεν αλλάζει το στατιστικό αποτέλεσμα, αλλά βελτιώνει τη σαφήνεια του κώδικα.

Καθιστά ξεκάθαρο ποια δεδομένα συγκρίνονται και μειώνει την ανάγκη για ενδιάμεσους υπολογισμούς.

Timokratis Karamitros
Bioinformatics and Applied Genomics Unit

Assumptions and diagnostics

Στις περισσότερες περιπτώσεις, έχουμε ήδη στο μυαλό μας μια στατιστική ανάλυση όταν συλλέγουμε δεδομένα. Αν και είναι δελεαστικό να προχωρήσουμε απευθείας στην ανάλυση, αυτό σπάνια είναι καλή ιδέα.

Η αρχική διερεύνηση των δεδομένων μας βοηθά να κατανοήσουμε τι πραγματικά «λένε» και να ελέγξουμε αν οι υποθέσεις της ανάλυσης που σκοπεύουμε να χρησιμοποιήσουμε είναι ρεαλιστικές.

Η αρχική οπτική επιθεώρηση είναι κρίσιμο βήμα καθώς μας επιτρέπει να εντοπίσουμε πιθανά προβλήματα πριν την ανάλυση

Γιατί οι υποθέσεις είναι σημαντικές?

Η παλινδρόμηση και η ANOVA είναι ειδικές περιπτώσεις του γενικού γραμμικού μοντέλου. Πρόκειται για παραμετρικά μοντέλα, τα οποία βασίζονται σε συγκεκριμένες υποθέσεις σχετικά με τα δεδομένα. Αν αυτές οι υποθέσεις παραβιάζονται έντονα, τα αποτελέσματα και τα p -values δεν είναι αξιόπιστα.

Τα παραμετρικά μοντέλα απαιτούν έλεγχο υποθέσεων

Χωρίς έλεγχο, τα συμπεράσματα μπορεί να είναι παραπλανητικά

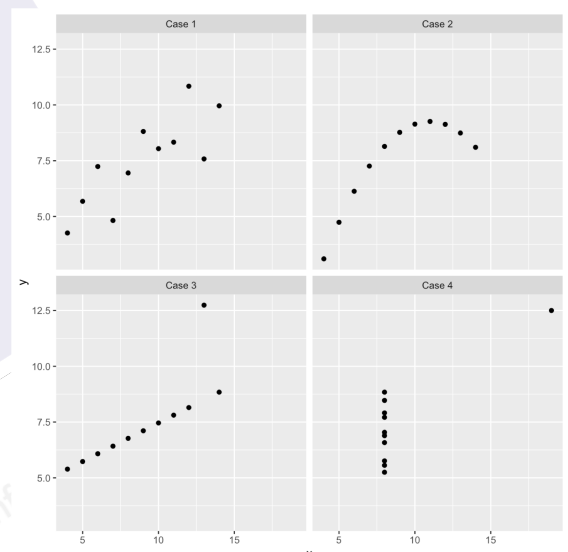
Κατανόηση των δεδομένων

Πριν από οποιαδήποτε ανάλυση, πρέπει να εξετάζουμε τα δεδομένα γραφικά.

Η εξερεύνηση των δεδομένων προηγείται της ανάλυσης
Τα γραφήματα αποκαλύπτουν δομές που δεν φαίνονται από αριθμούς

Το παράδειγμα του Anscombe

Τα τέσσερα διαγράμματα διασποράς του Anscombe έχουν σχεδόν ίδιους μέσους όρους, διακυμάνσεις, κλίσεις και τομές. Παρ' όλα αυτά, η σχέση μεταξύ x και y είναι εντελώς διαφορετική σε κάθε περίπτωση.
Ίδιες στατιστικές περιγραφές, διαφορετικές σχέσεις
Η παλινδρόμηση από μόνη της δεν αρκεί



The «Anscombe's quartet»

Τι δείχνουν τα τέσσερα παραδείγματα

<https://www.youtube.com/watch?v=Kd--Q-aTwpM>

Κάθε γράφημα αφηγείται διαφορετική «ιστορία» για τη σχέση των μεταβλητών, παρότι το γραμμικό μοντέλο δίνει τα ίδια αποτελέσματα.

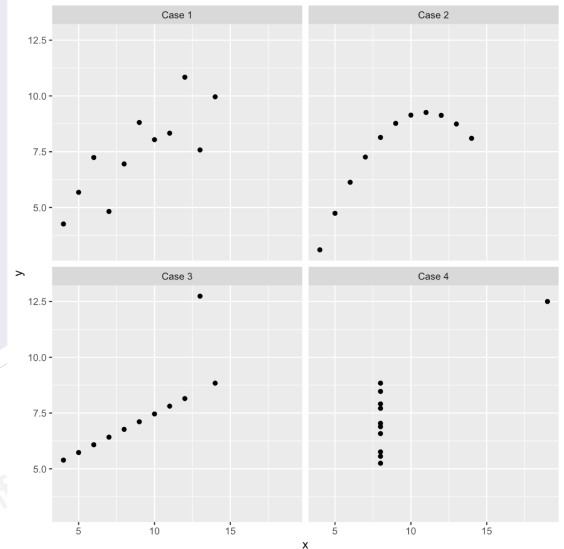
Περίπτωση 1: γραμμική σχέση, κανονικές κατανομές

Περίπτωση 2: μη γραμμική αλλά τέλεια σχέση

Περίπτωση 3: ένας outlier καταστρέφει τη σχέση

Περίπτωση 4: ένας outlier δημιουργεί ψευδή συσχέτιση

Αν δεν εξετάσουμε τα δεδομένα, υπάρχει σοβαρός κίνδυνος να εφαρμόσουμε λάθος στατιστικό μοντέλο. Τα παραδείγματα του Anscombe είναι ακραία, αλλά παρόμοια προβλήματα εμφανίζονται συχνά και σε πραγματικά δεδομένα.



Διαγνωστικά Γραφήματα για την Παλινδρόμηση στο R

Για να ελέγξουμε την καταλληλότητα και την εγκυρότητα ενός μοντέλου γραμμικής παλινδρόμησης (Linear Regression), είναι απαραίτητο να δημιουργήσουμε **διαγνωστικά γραφήματα**. Αυτά μας επιτρέπουν να αξιολογήσουμε εάν πληρούνται οι βασικές υποθέσεις του μοντέλου, ειδικά όσον αφορά τα **κατάλοιπα (residuals)**.

Θα χρησιμοποιήσουμε το παράδειγμα απελευθέρωσης γλυκόζης από το κεφάλαιο της Απλής Παλινδρόμησης.

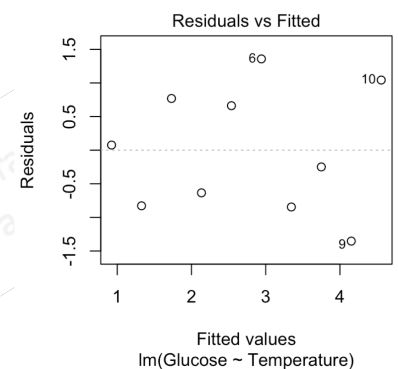
```
vicia_germ <- read.csv("GLUCOSE.CSV")
```

```
# Προσαρμογή του μοντέλου γραμμικής παλινδρόμησης:  
# Επίδραση της Θερμοκρασίας (Temperature) στην απελευθέρωση  
Γλυκόζης (Glucose)
```

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

Για να δημιουργήσετε το γράφημα **Residuals vs Fitted Values**:

```
plot(vicia_model, which = 1, add.smooth = FALSE)
```



Στη συνάρτηση `plot(vicia_model, which = 1, add.smooth = FALSE)`

- Το πρώτο όρισμα, `vicia_model`, είναι το όνομα του προσαρμοσμένου αντικειμένου μοντέλου (`lm`)
- Το δεύτερο όρισμα, `which = 1`, ελέγχει την έξοδο της συνάρτησης `plot` και την οδηγεί να παράγει το γράφημα **Κατάλοιπα vs Προσαρμοσμένες Τιμές** (`Residuals vs Fitted Values`).
- Το όρισμα `add.smooth = FALSE` λέει στη συνάρτηση να μην προσθέσει μια γραμμή «loess smooth». Αυτή η γραμμή υποτίθεται ότι βοηθά στην ανάδειξη προτύπων, αλλά τείνει να **υπερ-προσαρμόζει** (`overfit`) τη σχέση και οδηγεί τους χρήστες να βλέπουν προβλήματα που δεν υπάρχουν. Για αυτόν τον λόγο, είναι συχνά προτιμότερο να καταστέλλεται.

Έλεγχος της Υπόθεσης Γραμμικότητας

Το γράφημα `Residuals vs Fitted Values` μας επιτρέπει να αξιολογήσουμε την **υπόθεση γραμμικότητας**:

Τι ψάχνουμε:

Οποιαδήποτε ένδειξη συστηματικής τάσης (π.χ. καμπύλη U ή σχήμα καμπούρας) στη διασπορά των σημείων

Συμπέρασμα: Δεδομένου ότι δεν υπάρχει προφανές πρότυπο (τα σημεία είναι τυχαία κατανεμημένα γύρω από το 0), μπορούμε να **αποδεχτούμε την υπόθεση της γραμμικότητας**.

Έλεγχος της Υπόθεσης Κανονικότητας

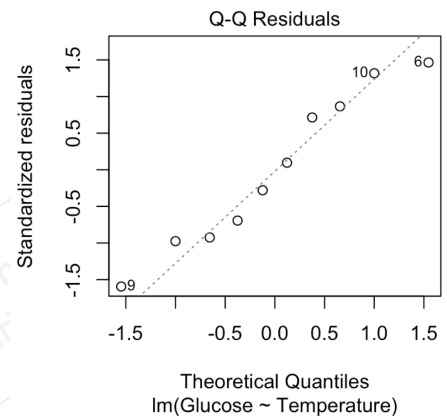
Χρησιμοποιούμε τη συνάρτηση `plot` για να δημιουργήσουμε το διαγνωστικό γράφημα κανονικής πιθανότητας (Normal Probability Plot) ρυθμίζοντας το `which = 2`:

```
plot(vicia_model, which = 2)
```

Αυτό παράγει το γράφημα **Normal Q-Q (Quantile-Quantile)**.

Τι ψάχνουμε: Στο Normal Q-Q Plot, ψάχνουμε να δούμε αν τα σημεία δεδομένων βρίσκονται κοντά στην **διαγώνια γραμμή αναφοράς 45°**.

Εάν τα σημεία ακολουθούν τη γραμμή, τότε τα κατάλοιπα κατανέμονται κανονικά.



Έλεγχος της Υπόθεσης Σταθερής Διακύμανσης (Ομοιοσκεδαστικότητα)

Για να αξιολογήσουμε την υπόθεση της **σταθερής διακύμανσης (constant variance)** ή **ομοιοσκεδαστικότητας (homoscedasticity)**, χρησιμοποιούμε το διαγνωστικό γράφημα **Scale-Location** (`which = 3`).

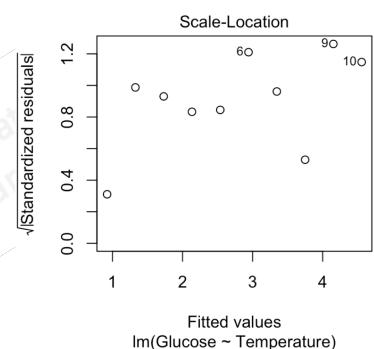
```
plot(vicia_model, add.smooth = FALSE, which = 3)
```

`which = 3`: Καθορίζει το γράφημα Scale-Location. Αυτό το γράφημα απεικονίζει τις **τετραγωνικές ρίζες των τυποποιημένων καταλοίπων** ($\sqrt{|\text{Standardized residuals}|}$) έναντι των **προσαρμοσμένων τιμών (Fitted Values)**.

Τι Ψάχνουμε στο Scale-Location Plot

Για να ισχύει η υπόθεση της ομοιοσκεδαστικότητας:

- Τα σημεία πρέπει να είναι **τυχαία κατανεμημένα** και να σχηματίζουν ένα **οριζόντιο «σύννεφο»**.
- Η διασπορά των σημείων (το εύρος τους) να είναι **παρόμοια** για όλες τις τιμές των προσαρμοσμένων τιμών. Αν η διασπορά των σημείων αυξάνεται ή μειώνεται συστηματικά καθώς αυξάνονται οι προσαρμοσμένες τιμές, τότε έχουμε **ετεροσκεδαστικότητα (heteroscedasticity)** και η υπόθεση παραβιάζεται.



Διαγνωστικά Γραφήματα για one-way ANOVA

Η έννοια των **διαγνωστικών γραφημάτων** (diagnostic plots) ισχύει για οποιοδήποτε Γενικό Γραμμικό Μοντέλο (General Linear Model), συμπεριλαμβανομένης της ανάλυσης **One-way ANOVA**.

Θα χρησιμοποιήσουμε το παράδειγμα της πυκνότητας των corncrake (ορτυκιών) από το αντίστοιχο κεφάλαιο.

```
# Προσαρμογή του μοντέλου ANOVA:  
# Επίδραση του Συμπληρώματος (Supplement) στην Αύξηση Βάρους (WeightGain)  
  
corncrake_model <- lm(WeightGain ~ Supplement, data = corn_crake)
```

Normal Probability Plot (which = 2)

Το γράφημα κανονικής πιθανότητας χρησιμοποιείται για την αναγνώριση αποκλίσεων από την **υπόθεση της κανονικότητας**. Αυτό το γράφημα μας επιτρέπει να ελέγξουμε αν τα κατάλοιπα (η διακύμανση κάθε σημείου από τον μέσο όρο της ομάδας του) είναι πιθανό να έχουν αντληθεί από **κανονική κατανομή**.

Δημιουργούμε το γράφημα **Normal Q-Q** για το παράδειγμά μας ρυθμίζοντας το `which = 2`:

Όπως και στην παλινδρόμηση, αναμένουμε τα σημεία να ακολουθούν τη γραμμή αναφοράς.

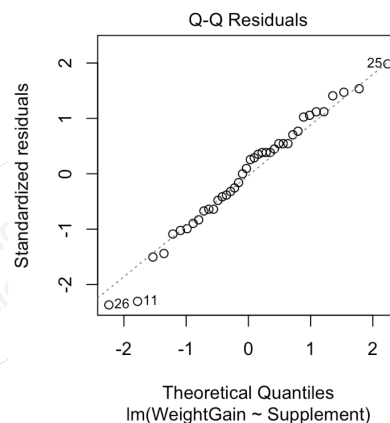
```
plot(corncrake_model, which = 2)
```

Όπως παρατηρήσατε, η ανάλυση του **Normal Q-Q Plot** (`which = 2`) δείχνει ότι:

Τα σημεία **δεν αποκλίνουν συστηματικά** από τη γραμμή αναφοράς.

Μερικές μικρές αποκλίσεις στο κατώτερο άκρο είναι **συνήθεις** και δεν αποτελούν λόγο ανησυχίας.

Συμπέρασμα: Η **υπόθεση της κανονικότητας** των καταλοίπων ικανοποιείται.



Έλεγχος Σταθερής Διακύμανσης (Scale-Location Plot, which = 3)

Το γράφημα **Scale-Location** χρησιμοποιείται για την αξιολόγηση της **υπόθεσης της σταθερής διακύμανσης** (constant variance) ή **ομοιοσκεδαστικότητας** (homoscedasticity).

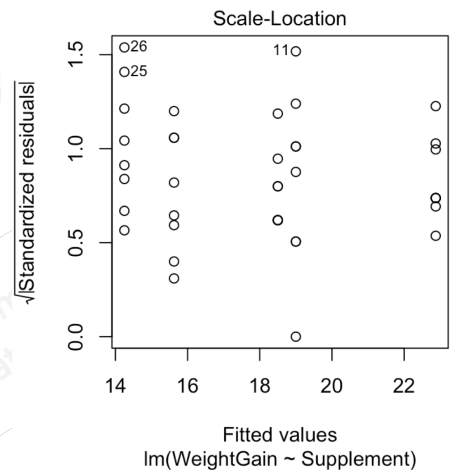
Μας επιτρέπει να δούμε εάν η **μεταβλητότητα των καταλοίπων** είναι περίπου **σταθερή εντός κάθε ομάδας** (Supplement):

```
plot(corncrake_model, which = 3, add.smooth = FALSE)
```

Τι Παρατηρούμε στο Scale-Location Plot:

Το γράφημα απεικονίζει τη **ρίζα της απόλυτης τιμής των τυποποιημένων καταλοίπων** έναντι των **προσαρμοσμένων τιμών** (οι οποίες είναι οι μέσοι όροι των ομάδων Supplement).

- Ελέγχουμε εάν η **διασπορά των σημείων** γύρω από τον οριζόντιο άξονα είναι περίπου **ίδια** για όλες τις ομάδες.
- Εάν οι ομάδες με υψηλότερο μέσο όρο έχουν μεγαλύτερη διασπορά καταλοίπων (ή το αντίστροφο), τότε παραβιάζεται η ομοιοσκεδαστικότητα.
- Εάν τα σημεία σχηματίζουν ένα οριζόντιο "σύννεφο" με σταθερό πλάτος, η υπόθεση ικανοποιείται.



Το βασικό προειδοποιητικό σημάδι που αναζητούμε στο γράφημα Scale-Location είναι ένα **συστηματικό πρότυπο** στη διασπορά των καταλοίπων.

Τι αναζητούμε: Να διαπιστώσουμε αν το **μέγεθος** (magnitude) των καταλοίπων τείνει να **αυξάνεται ή να μειώνεται** καθώς αυξάνονται οι προσαρμοσμένες τιμές (Fitted Values / Μέσοι Όροι Ομάδων).

Σημασία: Εάν εμφανιστεί ένα τέτοιο πρότυπο (π.χ., σχήμα χωνιού), υποδηλώνει ότι η **διακύμανση αλλάζει με τον μέσο όρο** (Heteroscedasticity / Ετεροσκεδαστικότητα).

Συμπέρασμα για το Παράδειγμα (corncrake):

Δεδομένου ότι **δεν υπάρχει τέτοιο πρότυπο** στο γράφημα Scale-Location για το μοντέλο ANOVA του corncrake, η **υπόθεση της σταθερής διακύμανσης (constant variance) ή ομοιοσκεδαστικότητας (homoscedasticity) ικανοποιείται**.

Κατηγορικά Δεδομένα (Categorical Data)

Timokratis Karamitros
Bioinformatics and Applied Genomics Unit

Κατηγορικά Δεδομένα (Categorical Data)

Τι είναι και πώς διαφέρουν;

Στη Βιολογία συχνά μετράμε **ποσοτικές μεταβλητές** (μήκος, βάρος), όπου υπολογίζουμε μέσους όρους.

Υπάρχουν όμως περιπτώσεις όπου η "μέτρηση" είναι **κατηγορική** (π.χ. φύλο, χρώμα, είδος).

Περιορισμός: Στα κατηγορικά δεδομένα **δεν** μπορούμε να υπολογίσουμε μέσο όρο (π.χ. δεν υπάρχει "μέσο χρώμα" ή "μέσο φύλο").

Τι αναλύουμε: Δουλεύουμε με τις **συχνότητες εμφάνισης** (counts/frequencies) των κατηγοριών.

Timokratis Karamitros
Bioinformatics and Applied Genomics Unit

Η Κατανομή Χ² (Chi-square)

Βασικά Χαρακτηριστικά

Χρησιμοποιείται για τον έλεγχο υποθέσεων σε κατηγορικά δεδομένα.

Παράμετροι: Ορίζεται από μία μόνο παράμετρο, τους **βαθμούς ελευθερίας (degrees of freedom - df)**.

Τιμές: Παίρνει μόνο **θετικές τιμές** (από 0 έως + ∞).

Σχήμα: Είναι ασύμμετρη κατανομή (δεξιά στρεβλή - skewed to the right).

Προσομοίωση της Κατανομής στην R

Προσομοίωση της Κατανομής στην R Πώς προκύπτει το στατιστικό Χ²;

Αν έχουμε μια αναμενόμενη τιμή (E) και μια παρατηρούμενη τιμή (O), η ποσότητα κατανομή Χ².

$$\frac{(O - E)^2}{E}$$

ακολουθεί την

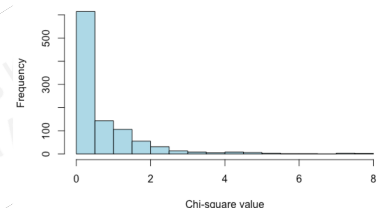
Κώδικας R για προσομοίωση:

```
# Προσομοίωση 1000 δειγματοληψιών με αναμενόμενη συχνότητα 25%
expected <- 250
observed <- rbinom(1000, size = 1000, prob = 0.25)

# Υπολογισμός του στατιστικού Χ2 για κάθε δείγμα
chi_sq_stats <- ((observed - expected)^2) / expected

# Οπτικοποίηση της κατανομής των αποτελεσμάτων
hist(chi_sq_stats, breaks = 20, main = "Distribution of Chi-square
statistic", xlab = "Chi-square value", col = "lightblue")
```

Distribution of Chi-square statistic



Τύποι Ελέγχων χ^2

Υπάρχουν δύο κύριοι τύποι ελέγχου ανάλογα με το ερώτημά μας:

Έλεγχος Καλής Προσαρμογής (Goodness of Fit):

Έχουμε **μία** κατηγορική μεταβλητή.

Ελέγχουμε αν οι παρατηρούμενες συχνότητες ταιριάζουν με μια θεωρητική προσδοκία (π.χ. αναλογία φύλου 1:1).

Έλεγχος Πίνακα Συνάφειας (Contingency Table Test):

Έχουμε **δύο** κατηγορικές μεταβλητές.

Ελέγχουμε αν υπάρχει **συσχέτιση** μεταξύ τους (π.χ. αν το χρώμα ματιών σχετίζεται με το φύλο).

Παράδειγμα - Goodness of Fit

Έλεγχος αναλογίας φύλου στους φοιτητές

Υπόθεση: Η αναλογία ανδρών/γυναικών είναι 1:1.

Δεδομένα: Σε 164 φοιτητές, παρατηρούμε:

Άνδρες: 74 (Αναμενόμενο: 82)

Γυναίκες: 90 (Αναμενόμενο: 82)

Ερώτημα: Η διαφορά οφείλεται στην τύχη (sampling variation) ή υπάρχει πραγματική απόκλιση;

Παράδειγμα - Contingency Table

Συσχέτιση Φύλου και Χρώματος Ματιών

Δημιουργούμε έναν πίνακα με δύο ταξινομήσεις:

	Μπλε Μάτια	Καφέ Μάτια	Σύνολο
Άνδρες	44	20	64
Γυναίκες	58	42	100

Στόχος: Να δούμε αν η πιθανότητα να έχει κάποιος καφέ μάτια διαφέρει μεταξύ ανδρών και γυναικών.

Προϋποθέσεις του χ^2 Test

Για να είναι έγκυρο το τεστ, πρέπει να ισχύουν τα εξής:

Ανεξαρτησία: Οι μετρήσεις πρέπει να είναι ανεξάρτητες (κάθε αντικείμενο ανήκει σε μία μόνο κατηγορία).

Μέγεθος Δεδομένων: Οι **αναμενόμενες** συχνότητες (expected counts) πρέπει να είναι ≥ 5 .

Μόνο Συχνότητες: Το τεστ γίνεται **πάντα** σε πραγματικούς αριθμούς (counts).

Ποτέ σε ποσοστά (%) ή μέσους όρους.

Μόνο σε ακέραιους αριθμούς παρατηρήσεων.

Έλεγχος Καλής Προσαρμογής (χ^2 Goodness of Fit)

Πότε τον χρησιμοποιούμε;

Όταν έχουμε **μία** κατηγορική μεταβλητή.

Όταν θέλουμε να συγκρίνουμε τις παρατηρούμενες συχνότητες με **προκαθορισμένες αναμενόμενες τιμές**.

Παραδείγματα:

Αναλογία Φύλου: Είναι η αναλογία αρσενικών-θηλυκών 1:1 ή υπάρχει μεροληψία;

Πειράματα Επιλογής: Αν προσφέρουμε σε σκίουρους 4 είδη τροφής, τις επιλέγουν εξίσου (αναλογία 1:1:1:1);

Γενετική Μέντελ: Ταιριάζουν οι απόγονοι στην αναμενόμενη φαινοτυπική αναλογία 3:1;

Πώς λειτουργεί ο έλεγχος;

Τα βήματα της ανάλυσης

Μηδενική Υπόθεση (H_0): Δεν υπάρχει διαφορά μεταξύ παρατηρούμενων και αναμενόμενων συχνοτήτων (π.χ. "Η αναλογία είναι 1:1").

Υπολογισμός Αναμενόμενων Τιμών (E): Πόσες παρατηρήσεις θα περιμέναμε σε κάθε κατηγορία αν ίσχυε η H_0 ;

Υπολογισμός Στατιστικού χ^2 : Μια τιμή που συνοψίζει τη διαφορά παρατηρούμενου-αναμενόμενου.

Βαθμοί Ελευθερίας (df): $n - 1$ (όπου n ο αριθμός των κατηγοριών).

p-value: Η πιθανότητα να δούμε τέτοια (ή μεγαλύτερη) απόκλιση λόγω τύχης.

Προϋποθέσεις και Περιορισμοί

Ανεξαρτησία: Κάθε παρατήρηση είναι μοναδική και ανεξάρτητη.

Μέγεθος Δείγματος: Οι **αναμενόμενες** τιμές πρέπει να είναι > 5 .

Προσοχή: Αν οι αναμενόμενες τιμές είναι πολύ μικρές, το p-value δεν είναι αξιόπιστο.

Δεδομένα: Χρησιμοποιούμε πάντα τους **αρχικούς αριθμούς (counts)** και ποτέ ποσοστά ή μέσους όρους.

Εφαρμογή στην R (Παράδειγμα *Silene dioica*)

Θέλουμε να ελέγξουμε αν η αναλογία αρσενικών/θηλυκών φυτών είναι 1:1

```
# 1. Εισαγωγή παρατηρούμενων συχνοτήτων (105 αρσενικά, 87 θηλυκά)
observed_freqs <- c(105, 87)

# 2. Ορισμός αναμενόμενων αναλογιών (0.5 για το καθένα)
n_cat <- length(observed_freqs)
expected_probs <- rep(1, n_cat) / n_cat

# 3. Εκτέλεση του ελέγχου Chi-square
chisq.test(observed_freqs, p = expected_probs)
```

X-squared = 1.6875, df = 1, p-value = 0.1939

Τι σημαίνει αυτό;

- **p-value = 0.1939:** Εφόσον το $p > 0.05$, **δεν** απορρίπτουμε τη μηδενική υπόθεση.
- **Συμπέρασμα:** Η αναλογία φύλου στο δείγμα μας δεν διαφέρει σημαντικά από την αναμενόμενη 1:1.
- Οι διαφορές που παρατηρούμε οφείλονται πιθανότατα σε τυχαία δειγματοληψία.

Πίνακες Συνάφειας (Contingency Tables)

Πότε χρησιμοποιούμε αυτόν τον έλεγχο;

Όταν μελετάμε **δύο ή περισσότερες** κατηγορικές μεταβλητές ταυτόχρονα.

Όταν θέλουμε να εξετάσουμε αν υπάρχει **συσχέτιση (association)** μεταξύ των μεταβλητών.

Μηδενική Υπόθεση (H0): Οι δύο μεταβλητές είναι **ανεξάρτητες** (η μία δεν επηρεάζει την άλλη).

Παραδείγματα:

Φύλο και Χρώμα Ματιών: Διαφέρει η αναλογία καφέ ματιών μεταξύ ανδρών και γυναικών;

Βιότοπος και Χρώμα Πασχαλίτσας: Είναι πιο συχνές οι μαύρες πασχαλίτσες στις βιομηχανικές περιοχές σε σχέση με τις αγροτικές;

Κατανόηση της Συσχέτισης

Συσχέτιση σημαίνει ότι οι αναλογίες μιας μεταβλητής αλλάζουν ανάλογα με την κατηγορία της άλλης μεταβλητής.

	Μπλε Μάτια	Καφέ Μάτια
Άνδρες	22	10
Γυναίκες	29	21

Αυτός ο πίνακας ονομάζεται **"two-way contingency table"** επειδή συνοψίζει δύο μεταβλητές. Αν είχαμε και τρίτη (π.χ. ηλικία), θα ήταν "three-way".

Οπτικοποίηση Συσχέτισης

Δεν μας ενδιαφέρουν οι απόλυτοι αριθμοί, αλλά οι **αναλογίες**.

Πίνακας Χωρίς Συσχέτιση (Ανεξαρτησία)

	C1	C2
R1	10	20
R2	40	80

Το R1 είναι πάντα το 1/4 του R2, ανεξάρτητα από το αν είμαστε στο C1 ή στο C2.

Πίνακας Με Συσχέτιση

	C1	C2
R1	10	80
R2	40	20

Η αναλογία του R1 αλλάζει δραματικά ανάλογα με το αν είμαστε στο C1 ή στο C2.

Παράδειγμα - Χρώμα Πασχαλίτσας (*Adalia bipunctata*)

Ερευνούμε αν το περιβάλλον (Αγροτικό/Βιομηχανικό) επηρεάζει το χρώμα (Μαύρο/Κόκκινο).



Πίνακας Παρατηρούμενων Συχνοτήτων:

	Black	Red	Totals
Rural	30	70	100
Industrial	115	85	200
Totals	145	155	300

$$E = \frac{(\text{Σύνολο Γραμμής}) \times (\text{Σύνολο Στήλης})}{\text{Γενικό Σύνολο}}$$

Προϋποθέσεις του ελέγχου χ^2

Οι προϋποθέσεις παραμένουν οι ίδιες με τον έλεγχο καλής προσαρμογής:

Ανεξάρτητες Παρατηρήσεις: Τα δεδομένα δεν πρέπει να προέρχονται από επαναλαμβανόμενες μετρήσεις στα ίδια άτομα (π.χ. πριν και μετά από μια παρέμβαση).

Αναμενόμενες Τιμές: Πρέπει να είναι > 5 σε κάθε κελί.

Αν έχουμε πολύ μικρές αναμενόμενες τιμές, το τεστ δεν είναι αξιόπιστο και ίσως χρειαστούμε εναλλακτικές (όπως το *Fisher's Exact Test*).

Εκτέλεση στην R

```
# Δημιουργία των στηλών (Habitat και Colour)
habitat <- c(rep("Rural", 100), rep("Industrial", 200))
colour <- c(rep("Black", 30), rep("Red", 70),      # Δεδομένα Rural
            rep("Black", 115), rep("Red", 85))    # Δεδομένα Industrial

lady_bird_df1 <- data.frame(Habitat = habitat, Colour = colour)

head(lady_bird_df1)

# Μετατροπή σε contingency table
lady_bird_table <- xtabs(~ Habitat + Colour, data = lady_bird_df1)

# Εμφάνιση του πίνακα
print(lady_bird_table)

# Εκτέλεση του ελέγχου
chisq.test(lady_bird_table)
```



Ερμηνεία Αποτελεσμάτων `chisq.test()`

Pearson's Chi-squared test with Yates' continuity correction

data: lady_bird_table

X-squared = 19.103, df = 1, p-value = 1.239e-05

X-squared (19.103): Είναι η τιμή του στατιστικού χ^2 . Όσο μεγαλύτερη είναι αυτή η τιμή, τόσο μεγαλύτερη είναι η απόκλιση από αυτό που περιμέναμε αν οι μεταβλητές ήταν ανεξάρτητες.

df = 1 (Degrees of Freedom): Οι βαθμοί ελευθερίας. Για πίνακες συναφείας υπολογίζονται ως: $(\text{rows} - 1) \times (\text{cols} - 1)$. Εδώ: $(2-1) \times (2-1) = 1$.

p-value (1.239e-05): Το e-05 σημαίνει 10^{-5} άρα ότι μετακινούμε την υποδιαστολή 5 θέσεις αριστερά: 0.00001239

Όταν οι αναμενόμενες τιμές είναι < 5

Όπως είδαμε στις προϋποθέσεις, το χ^2 test βασίζεται στη μαθηματική προσέγγιση ότι οι αναμενόμενες τιμές (E) είναι αρκετά μεγάλες.

Αν έστω και ένα κελί έχει αναμενόμενη τιμή μικρότερη από 5, η R θα εκτελέσει τον έλεγχο αλλά θα εμφανίσει μια **προειδοποίηση (Warning)**: Warning: Chi-squared approximation may be incorrect

```
# Δεδομένα από ένα πολύ μικρό πείραμα
small_data <- matrix(c(3, 1, 2, 5), nrow = 2)
chisq.test(small_data)
```

X-squared = 0.73661, df = 1, p-value = 0.3907

Warning message:

In `chisq.test(small_data)` : Chi-squared approximation may be incorrect

Η λύση - Fisher's Exact Test

Όταν έχουμε μικρά δείγματα (πίνακες 2×2 με αναμενόμενες τιμές < 5), η ενδεδειγμένη λύση είναι το **Fisher's Exact Test**.

- Δεν βασίζεται σε προσεγγίσεις κατανομών.
- Υπολογίζει την **ακριβή πιθανότητα** εμφάνισης των δεδομένων μας.

```
# Χρήση του Fisher's Exact Test αντί για  
Chi-square  
fisher.test(small_data)  
  
data: small_data  
p-value = 0.2424  
alternative hypothesis: true odds ratio is  
not equal to 1  
95 percent confidence interval:  
 0.285564 468.452065  
sample estimates:  
odds ratio  
 6.069191
```

Εναλλακτική - Προσομοίωση Monte Carlo

Μια άλλη λύση που προσφέρει η R μέσα στην ίδια τη συνάρτηση `chisq.test()` είναι ο υπολογισμός του p-value μέσω **προσομοίωσης (Monte Carlo simulation)**. Αυτό είναι χρήσιμο αν έχετε μεγαλύτερους πίνακες (π.χ. 3×3) όπου το Fisher's test μπορεί να είναι υπολογιστικά βαρύ.

```
# Εκτέλεση Chi-square με υπολογισμό p-value μέσω προσομοίωσης  
chisq.test(small_data, simulate.p.value = TRUE, B = 2000)
```

Η παράμετρος $B = 2000$ ορίζει πόσες τυχαίες επαναλήψεις (replicates) θα κάνει η R για να βρει το p-value. Με αυτή τη μέθοδο, η προειδοποίηση "approximation may be incorrect" εξαφανίζεται.

Κατάσταση

Μεγάλο δείγμα (Όλα τα $E > 5$)

Μικρό δείγμα ($E < 5$) σε πίνακες 2×2

Μικρό δείγμα σε μεγάλους πίνακες

Προτεινόμενο Τεστ

`chisq.test()`

`fisher.test()`

`chisq.test(..., simulate.p.value = TRUE)`

Μη Παραμετρικές Δοκιμασίες (Non-parametric tests)

Timokratis Karamitros
Bioinformatics and Applied Genomics Unit

Τι είναι οι Μη Παραμετρικές Δοκιμασίες;

Βασίζονται σε μαθηματικά μοντέλα που υποθέτουν συγκεκριμένες κατανομές στον πληθυσμό (π.χ. Κανονική κατανομή).

Κάνουν πολύ πιο "ασθενείς" υποθέσεις. Δεν απαιτούν τα δεδομένα να προέρχονται από μια συγκεκριμένη κατανομή.

Ήδη γνωστές μη παραμετρικές μέθοδοι:

Bootstrap και Permutation tests.

Έλεγχοι χ^2 (Καλής προσαρμογής και Πίνακες συνάφειας).

Timokratis Karamitros
Bioinformatics and Applied Genomics Unit

Πλεονεκτήματα και Μειονεκτήματα

Πλεονεκτήματα (Γιατί να τις χρησιμοποιήσω;)

Ευελιξία: Κατάλληλες όταν παραβιάζονται οι προϋποθέσεις των παραμετρικών τεστ.

Τύπος Δεδομένων: Μπορούν να αναλύσουν **τακτικά (ordinal)** δεδομένα ή δεδομένα με **κατατάξεις (ranks)**.

Παράδειγμα: Βαθμολόγηση ελκυστικότητας φωτογραφιών, σειρά επιθετικότητας σε ζώα.

Μειονεκτήματα (Γιατί να τις αποφύγω αν μπορώ;)

Χαμηλότερη Ισχύς (Power): Αν τα δεδομένα είναι κατάλληλα για παραμετρικό τεστ, το μη παραμετρικό είναι λιγότερο πιθανό να βρει μια πραγματική διαφορά.

Λιγότερο Πληροφοριακές: Είναι πιο δύσκολο να καταλάβουμε ποιες ακριβώς διαφορές οδηγούν στο σημαντικό αποτέλεσμα (δυσκολία στα post-hoc tests).

Περιορισμένη πολυπλοκότητα: Δεν υπάρχουν μη παραμετρικά αντίστοιχα για όλα τα σύνθετα μοντέλα ANOVA.

Πώς λειτουργούν; (Η Αρχή της Κατάταξης)

Οι μη παραμετρικές δοκιμασίες δεν χρησιμοποιούν τις αρχικές τιμές, αλλά τη **σειρά κατάταξής** τους (ranks).

Sample A:	1.3	2.8	4.1	3.2
Sample B:	7.2	3.0	4.2	6.2

Βάζουμε όλα τα δεδομένα μαζί σε αύξουσα σειρά.
Αντικαθιστούμε κάθε τιμή με τη σειρά της (1, 2, 3...).

Sample A:	1.3	2.8	3.2	4.1
Sample B:	3.0	4.2	6.2	7.2

Κατάταξη (Ranks):

Δείγμα A: 1, 2, 4, 5 (Άθροισμα = 12)

Δείγμα B: 3, 6, 7, 8 (Άθροισμα = 24)

Συμπέρασμα: Το μεγάλο άθροισμα στο Δείγμα B υποδηλώνει ότι οι τιμές του τείνουν να είναι μεγαλύτερες από του Δείγματος A.

Αντιστοιχία Παραμετρικών και Μη Παραμετρικών Tests

Όταν οι προϋποθέσεις της κανονικότητας δεν ισχύουν, επιλέγουμε το αντίστοιχο μη παραμετρικό τεστ:

Παραμετρικό Τεστ	Μη Παραμετρικό Αντίστοιχο	Χρήση
One-sample / Paired t-test	Wilcoxon signed-rank test	Σύγκριση δείγματος με τιμή ή ζευγαρωτά δεδομένα
Two-sample t-test	Mann-Whitney U-test	Σύγκριση δύο ανεξάρτητων δειγμάτων
One-way ANOVA	Kruskal-Wallis test	Σύγκριση τριών ή περισσότερων δειγμάτων
Pearson Correlation	Spearman's Rank Correlation	Συσχέτιση δύο μεταβλητών

Μη Παραμετρικοί Έλεγχοι στην R

Η R διαθέτει ενσωματωμένες συναρτήσεις για όλους τους παραπάνω ελέγχους.

```
# Για Wilcoxon και Mann-Whitney (ανάλογα τις παραμέτρους)
wilcox.test(variable1, variable2)
```

```
# Για Kruskal-Wallis
kruskal.test(y ~ group, data = my_data)
```

```
# Για Spearman Correlation
cor.test(x, y, method = "spearman")
```

Wilcoxon signed-rank test

Είναι το μη παραμετρικό αντίστοιχο του **one-sample t-test** και του **paired-sample t-test**.

Χρησιμοποιείται για να συγκρίνουμε τη **διάμεσο (median)** ενός δείγματος με μια προκαθορισμένη τιμή ή για να συγκρίνουμε **ζευγαρωμένες παρατηρήσεις**.

Προϋποθέσεις:

Η μεταβλητή πρέπει να είναι τουλάχιστον **τακτική (ordinal)**, διαστήματος (interval) ή λόγου (ratio).

Δεν εφαρμόζεται σε ονομαστικά (nominal) δεδομένα.

Η κατανομή του πληθυσμού πρέπει να είναι **προσεγγιστικά συμμετρική**. Το τεστ μπορεί να αποτύχει αν η κατανομή παρουσιάζει έντονη δεξιά ή αριστερή στρέβλωση.

Εφαρμογή σε Ζευγαρωτά Δείγματα (Paired Design)

Όπως και στο paired t-test, η διαδικασία επικεντρώνεται στις **διαφορές** μεταξύ των ζευγαριών:

Υπολογίζονται οι διαφορές για κάθε ζεύγος παρατηρήσεων.

Ελέγχεται αν αυτές οι διαφορές διαφέρουν σημαντικά από το μηδέν.

Σημαντική Σημείωση: Συχνά, η κατανομή των διαφορών μπορεί να είναι κανονική ακόμα και αν τα αρχικά δείγματα δεν είναι. Αν οι διαφορές είναι κανονικές, προτιμάμε το t-test. Αν **δεν** είναι κανονικές, χρησιμοποιούμε το Wilcoxon.

Παράδειγμα - Άμυνα Φυτών και Κάμπιες

Δομή Δεδομένων και Οπτικοποίηση

Τα δεδομένα μας αποτελούνται από 40 γραμμές και 3 μεταβλητές:

ID: Μοναδικός κωδικός για κάθε ζεύγος φύλλων (1-20).

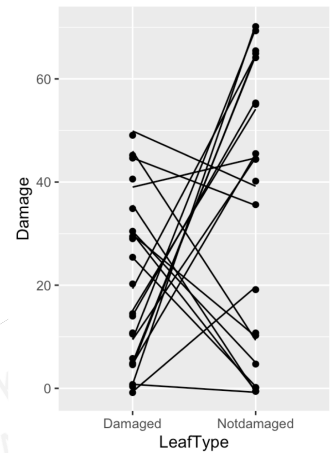
LeafType: Ο τύπος του φύλλου ("Damaged" vs "Notdamaged").

Damage: Το ποσοστό (%) κατανάλωσης του φύλλου.

```
leaves <- LEAF_DAMAGE.CSV (read.csv...)
```

Οπτικοποίηση με "Jitter"

Χρησιμοποιούμε τη μέθοδο **jitter** για να αποφύγουμε την επικάλυψη των σημείων που έχουν την ίδια τιμή.



Παρατήρηση: Υπάρχει μεγάλη παραλλακτικότητα (κάποιες κάμπιες τρώνε περισσότερο γενικά), αλλά διακρίνεται μια τάση προτίμησης προς τα μη κατεστραμμένα φύλλα.

```
jitter <- position_jitter(width = 0, height = 1)

ggplot(leaves, aes(x = LeafType, y = Damage, group = ID)) +
  geom_point(position = jitter) +
  geom_line(position = jitter)
```

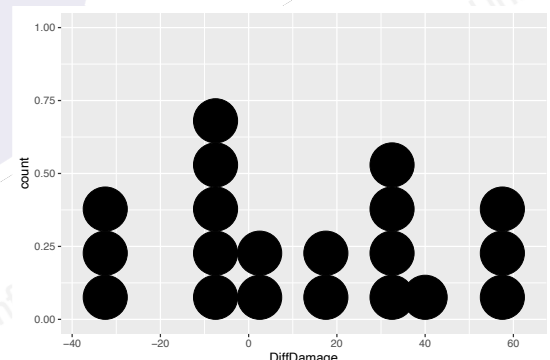
Έλεγχος Κανονικότητας των Διαφορών

Πριν επιλέξουμε το τεστ, εξετάζουμε την κατανομή των **διαφορών** εντός κάθε κάμπιας.

```
# Βήμα 1 -- Υπολογισμός των διαφορών
leaves_diff <- leaves %>%
  group_by(ID) %>%
  summarise(DiffDamage = diff(Damage))

# Βήμα 2 -- Δημιουργία dotplot
ggplot(leaves_diff, aes(x = DiffDamage)) +
  geom_dotplot(binwidth = 10)
```

Συμπέρασμα: Η κατανομή δεν φαίνεται κανονική, αλλά είναι προσεγγιστικά **συμμετρική**. Επομένως, η προϋπόθεση συμμετρίας του Wilcoxon ικανοποιείται.



Προετοιμάζοντας τα δεδομένα (Long to Wide)

Για να χρησιμοποιήσουμε τη συνάρτηση `Pair()`, πρέπει οι δύο μετρήσεις κάθε ζεύγους να βρίσκονται στην ίδια γραμμή. Μετατρέπουμε το dataframe από **long** μορφή σε **wide**.

```
library(tidyr)

# Μετατροπή σε wide format
leaves_wide <- leaves %>%
  pivot_wider(
    names_from = LeafType,
    values_from = Damage
  )
```

Η `pivot_wider()` παίρνει τις τιμές του `LeafType` και τις κάνει ονόματα στηλών, τοποθετώντας τις μετρήσεις δίπλα-δίπλα.

Εκτέλεση και Ερμηνεία του Wilcoxon Test

```
wilcox.test(Pair(Notdamaged, Damaged) ~ 1, data = leaves_wide)
```

Ανάλυση Εξόδου:

V = 134.5: Είναι το στατιστικό του ελέγχου (συχνά αναφέρεται ως **W** στις αναφορές).

p-value = 0.1156: Η τιμή είναι μεγαλύτερη από 0.05

Προειδοποιήσεις (Warnings): Η R μας ενημερώνει ότι υπάρχουν "ties" (ίδιες τιμές) ή "zeroes" (μηδενικές διαφορές), κάτι που καθιστά το p-value προσεγγιστικό.

Mann-Whitney U-test

Είναι η μη παραμετρική εναλλακτική του **independent two-sample t-test**.
Χρησιμοποιείται για τη σύγκριση της **διαμέσου (median)** δύο ανεξάρτητων δειγμάτων.

Προϋποθέσεις:

Η μεταβλητή είναι **τακτική (ordinal)**, διαστήματος (interval) ή λόγου (ratio).

Οι παρατηρήσεις μεταξύ των δύο ομάδων είναι **ανεξάρτητες**.

Οι κατανομές των δύο ομάδων έχουν **παρόμοιο σχήμα** (π.χ. δεν συγκρίνουμε μια κατανομή με έντονη δεξιά στρέβλωση με μια κατανομή με έντονη αριστερή στρέβλωση).

Μηδενική Υπόθεση (H₀): Οι δύο πληθυσμοί έχουν την ίδια διάμεσο. Ένα σημαντικό p-value σημαίνει ότι οι διάμεσοι (όχι οι μέσοι όροι!) διαφέρουν σημαντικά.

Συμπέρασμα: "Όταν δόθηκε η επιλογή, οι προνύμφες του *Operophtera brumata* δεν κατανάλωσαν σημαντικά μεγαλύτερες ποσότητες φύλλων από περιοχές του δέντρου που δεν είχαν υποστεί ζημιά, σε σύγκριση με εκείνες που είχαν υποστεί (Wilcoxon matched-pairs test: W=134.5, n = 20, p-value = 0.1156"

Παράδειγμα - Αναζήτηση τροφής από μυρμήγκια

Θα εξετάσουμε τον ρυθμό συλλογής τροφής (βιομάζα λείας ανά μυρμήγκι ανά μισή ώρα) σε δύο διαφορετικά είδη δέντρων: **Βελανιδιά (Oak)** και **Σφένδαμο (Sycamore)**.

Τα δεδομένα μας (ants):

•**Food:** Ο ρυθμός συλλογής τροφής.

•**Tree:** Το είδος του δέντρου.

```
ANTS1.CSV  
glimpse(ants)
```

Εκτέλεση στην R

Στην R, το Mann-Whitney U-test εκτελείται με τη συνάρτηση `wilcox.test()`. Παρόλο που το όνομα παραπέμπει στον Wilcoxon, πρόκειται για την ίδια μαθηματική διαδικασία.

```
# Εκτέλεση του ελέγχου με χρήση φόρμουλας  
wilcox.test(Food ~ Tree, data = ants)
```

Food ~ Tree: Η μεταβλητή με τις τιμές (Food) αριστερά από το ~ και η μεταβλητή ομαδοποίησης (Tree) δεξιά.

Ερμηνεία Αποτελεσμάτων

έξοδος της R θα μοιάζει με την εξής:

data: Food by Tree

W = 291, p-value = 0.2924

alternative hypothesis: true location shift is not equal to 0

Ανάλυση:

p-value = 0.2924: Εφόσον το $p > 0.05$, δεν απορρίπτουμε τη μηδενική υπόθεση.

Στατιστικό W: Αν και η R το ονομάζει **W**, στην αναφορά μας θα το καταγράψουμε ως **U** (αφού το ονομάζουμε Mann-Whitney).

Συμπέρασμα:

"Οι ρυθμοί συλλογής βιομάζας λείας δεν διέφεραν σημαντικά μεταξύ των μυρμηγκιών που συλλέγουν τροφή σε βελανιδιές και εκείνων σε σφενδάμους (Mann-Whitney U-test: $U = 291$, $p = 0.2924$)."

Kruskal-Wallis

Τι είναι;

Χρησιμοποιείται για τη σύγκριση της κεντρικής τάσης (διαμέσου) μεταξύ **περισσότερων από δύο** ανεξάρτητων ομάδων.

Είναι η μη παραμετρική εκδοχή της **one-way ANOVA**.

Προϋποθέσεις:

Η μεταβλητή είναι **τακτική (ordinal)**, διαστήματος ή λόγου.

Οι παρατηρήσεις είναι **ανεξάρτητες** μεταξύ των ομάδων.

Η κατανομή της μεταβλητής σε κάθε ομάδα είναι **παρόμοια** (ειδικά ως προς τη στρέβλωση).

Μηδενική Υπόθεση (H0): Όλες οι ομάδες προέρχονται από πληθυσμούς με την ίδια διάμεσο. Ένα σημαντικό p-value υποδηλώνει ότι τουλάχιστον μία ομάδα διαφέρει από τις υπόλοιπες.

Παράδειγμα - Μάθηση σε σουπιές

Το Πείραμα: Ερευνήθηκε πώς ο χρόνος έκθεσης σε μια κατάσταση επηρεάζει τη μάθηση στις σουπιές. Οι σουπιές επιτίθενται σε γαρίδες, αλλά αν οι γαρίδες είναι μέσα σε γυάλινο σωλήνα, αποτυγχάνουν και τελικά σταματούν τις επιθέσεις.

Δείγμα: 50 σουπιές χωρισμένες σε 5 ομάδες των 10.

Μεταβλητή Time (Έκθεση): 2, 4, 8, 15 και 30 λεπτά.

Μεταβλητή Strikes (Αποτέλεσμα): Μετά από 24 ώρες, μετρήθηκε ο αριθμός των επιθέσεων που έκαναν οι ίδιες σουπιές στις γαρίδες.

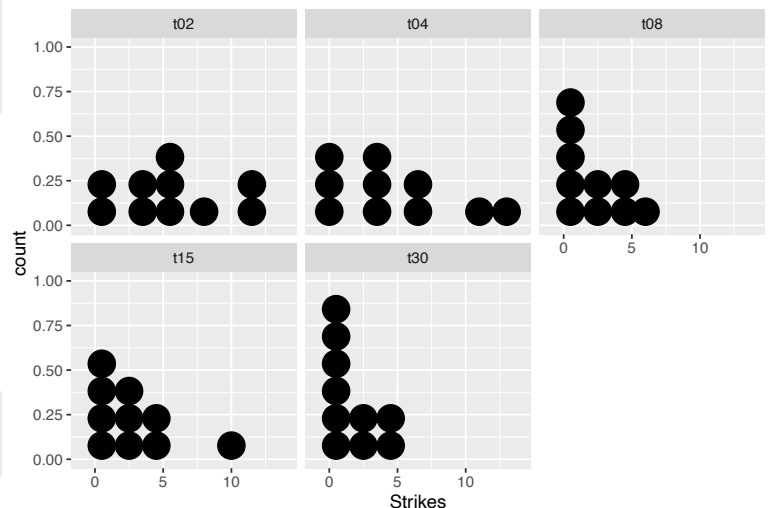
Οπτικοποίηση Δεδομένων

Πριν την ανάλυση, εξετάζουμε την κατανομή των επιθέσεων (Strikes) για κάθε χρονική ομάδα με dot plots.

CUTTLEFISH.CSV'

```
# Προβολή δεδομένων  
glimpse(cuttlefish)  
  
# Δημιουργία dot plots ανά ομάδα  
ggplot(cuttlefish, aes(x = Strikes)) +  
  geom_dotplot(binwidth = 2) +  
  facet_wrap(~ Time)
```

Παρατήρηση: Τα δεδομένα παρουσιάζουν στρέβλωση και περιλαμβάνουν πολλούς μικρούς ακέραιους αριθμούς με ίδιες τιμές. Αυτό καθιστά δύσκολη τη χρήση μετασχηματισμών και ιδανική τη χρήση του Kruskal-Wallis.



Εκτέλεση στην R

Η συνάρτηση `kruskal.test` χρησιμοποιεί τη γνωστή σύνταξη φόρμουλας (εξαρτημένη μεταβλητή ~ ομάδα).

```
kruskal.test(Strikes ~ Time, data = cuttlefish)
```

Ανάλυση Εξόδου:

data: Strikes by Time

Kruskal-Wallis chi-squared = 8.5788, df = 4, p-value = 0.07253

- **Kruskal-Wallis chi-squared = 8.5788:** Το στατιστικό του ελέγχου (συχνά αναφέρεται ως **H**).
- **df = 4:** Οι βαθμοί ελευθερίας (αριθμός ομάδων - 1).
- **p-value = 0.07253.** Εφόσον $p > 0.05$, δεν υπάρχει σημαντική διαφορά μεταξύ των ομάδων.

Συσχέτιση Spearman (ρ) Spearman's rank correlation

Αυτή είναι η λύση μας όταν θέλουμε να δούμε αν δύο μεταβλητές σχετίζονται, αλλά τα δεδομένα μας δεν είναι "ιδανικά".

Τι είναι;

Το μη παραμετρικό αντίστοιχο της συσχέτισης Pearson.
Ελέγχει την ύπαρξη **συσχέτισης** μεταξύ δύο αριθμητικών μεταβλητών.

Πλεονεκτήματα:

Δεν απαιτεί οι μεταβλητές να ακολουθούν την κανονική κατανομή.
Μπορεί να χρησιμοποιηθεί με **τακτικά (ordinal)** δεδομένα.
Είναι ανθεκτική σε κατανομές με μεγάλη στρέβλωση (skewed).

Βασικές Προϋποθέσεις:

Οι μεταβλητές είναι τουλάχιστον τακτικής κλίμακας.
Η σχέση μεταξύ τους είναι **μονοτονική**: Δηλαδή, καθώς η μία αυξάνεται, η άλλη τείνει είτε να αυξάνεται σταθερά είτε να μειώνεται σταθερά (χωρίς να σχηματίζει "λόφους" ή "κοιλιάδες").

Παράδειγμα - Αναπαραγωγή Μαυροπετεινού (Black Grouse)

Το Βιολογικό Ερώτημα: Στα "leks" (περιοχές όπου τα αρσενικά συγκεντρώνονται για επίδειξη), λίγα πουλιά καταφέρνουν πολλά ζευγαρώματα. Ένα αρσενικό που πέτυχε φέτος, θα τα πάει εξίσου καλά και την επόμενη χρονιά;

Τα Δεδομένα (grouse):

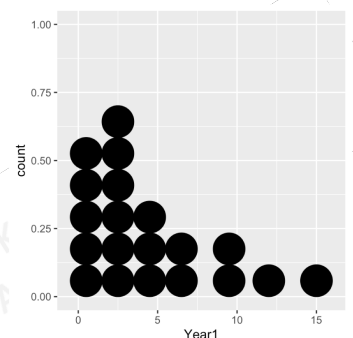
Year1: Αριθμός ζευγαρωμάτων το 1ο έτος.

Year2: Αριθμός ζευγαρωμάτων το 2ο έτος.

'GROUSE.CSV'

```
glimpse(grouse)
```

```
# Οπτικοποίηση κατανομής (dotplots)
ggplot(grouse, aes(x = Year1)) + geom_dotplot(binwidth = 2)
ggplot(grouse, aes(x = Year2)) + geom_dotplot(binwidth = 2)
```

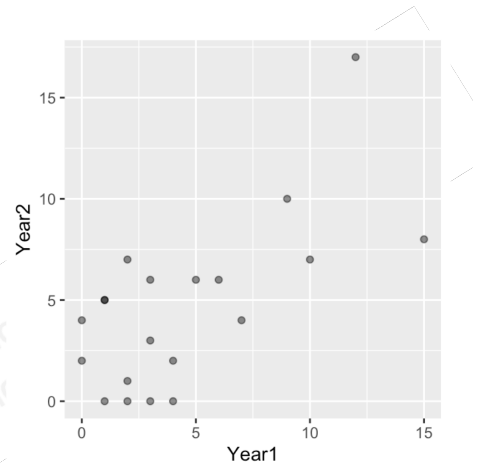


Παρατήρηση: Τα δεδομένα έχουν έντονη δεξιά στρέβλωση (οι περισσότεροι έχουν λίγα ζευγαρώματα, ελάχιστοι έχουν πολλά). Αυτό καθιστά τη συσχέτιση Spearman πιο ασφαλή από την Pearson

Οπτικοποίηση της Συσχέτισης

Πριν το τεστ, δημιουργούμε ένα scatter plot. Επειδή τα δεδομένα είναι ακέραιοι, χρησιμοποιούμε $\alpha = 0.5$ για να βλέπουμε πού συμπίπτουν τα σημεία.

```
ggplot(grouse, aes(x = Year1, y =  
Year2)) +  
  geom_point(alpha = 0.5) +  
  theme_minimal()
```



Γιατί Correlation και όχι Regression;

Δεν υπάρχει σχέση αιτιότητας (το έτος 1 δεν "προβλέπει" το έτος 2).

Δεν θέλουμε να προβλέψουμε τιμές, αλλά να δούμε αν υπάρχει **σύνδεση**.

Ο ερευνητής δεν ελέγχει καμία από τις δύο μεταβλητές.

Εκτέλεση στην R

Χρησιμοποιούμε τη συνάρτηση `cor.test()`, ορίζοντας ως μέθοδο την "spearman".

```
cor.test(~ Year1 + Year2, method = "spearman", data = grouse)
```

Ανάλυση Εξόδου:

- **S = 592.12**: Το στατιστικό του ελέγχου.
- **p-value = 0.01112**: Εφόσον $p < 0.05$, η συσχέτιση είναι στατιστικά σημαντική.
- **rho = 0.55**: Ο συντελεστής συσχέτισης. Δείχνει μια μέτρια, θετική σχέση.
- **Warning**: Το μήνυμα για τα "ties" (ίδιες τιμές) είναι αναμενόμενο και δεν επηρεάζει την εγκυρότητα της απόφασής μας.

Στη συσχέτιση Spearman, αναφέρουμε τον συντελεστή ρ , το μέγεθος του δείγματος n και το p-value. Δεν χρειάζονται βαθμοί ελευθερίας ή το στατιστικό S.

"Υπάρχει θετική συσχέτιση μεταξύ του αριθμού των ζευγαρωμάτων που πέτυχε ένα αρσενικό σε ένα lek και του αριθμού που πέτυχε το ίδιο αρσενικό στο επόμενο lek (Spearman's $\rho = 0.55$, $n = 20$, $p < 0.05$)."

R - Biostatistics in Biosciences

Dr. Timokratis Karamitros
Researcher

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

tkaram@pasteur.gr

