Ανάλυση Διασποράς

Έστω ότι μας δίνονται δείγματα που προέρχονται από άγνωστους πληθυσμούς. Πόσο διαφέρουν οι μέσες τιμές τους; Με άλλα λόγια: πόσο πιθανό είναι να προέρχονται από πληθυσμούς με την ίδια μέση τιμή;
Για να απαντήσουμε στα παραπάνω ερωτήματα, διατυπώνουμε τη μηδενική υπόθεση H_0 ότι όλα τα δείγματα προέρχονται από πληθυσμούς *με την ίδια μέση τιμή*.
Έλεγχος μηδενικής υπόθεσης: ο λόγος F που ορίσαμε για τον έλεγχο της ισότητας δύο διασπορών μπορεί να χρησιμεύσει για τον έλεγχο της παραπάνω μηδενικής υπόθεσης.
Η διαδικασία ονομάζεται **Ανάλυση Διασποράς** (ANalysis Of VAriance – ANOVA) και περιλαμβάνει τα ακόλουθα στάδια:

- Βρίσκουμε τη διασπορά των μέσων τιμών των δειγμάτων για να εκτιμήσουμε συνακόλουθα τη διασπορά του πληθυσμού (βλ. κατανομή δειγματοληψίας και ΚΟΘ): αυτό το ονομάζουμε **διακύμανση κατά παράγοντες** (between-groups variance)
- Εκτιμούμε τη διασπορά του πληθυσμού με βάση τις διασπορές στο εσωτερικό κάθε δείγματος: αυτό το ονομάζουμε **διακύμανση σφάλματος** (within groups variance, error variance)
- Συγκρίνουμε την πρώτη διασπορά με τη δεύτερη: αν η πρώτη είναι μικρή σε σχέση με τη δεύτερη, επαληθεύεται η μηδενική υπόθεση αλλιώς διαψεύδεται.

Νόημα / Φυσική Σημασία / Εφαρμογή: αυτά εφαρμόζονται όταν μελετάμε την επίδραση διαφορετικών παραγόντων στη συμπεριφορά ενός πληθυσμού. Έτσι, με κάθε διαφορετικό δείγμα παρατηρούμε την (πιθανή) επίδραση ενός ξεχωριστού παράγοντα (αν έχουμε σχεδιάσει κατάλληλα το πείραμα) και η κατανομή κάθε δείγματος περιέχει:
α) την επίδραση του αντίστοιχου παράγοντα
β) τις τυχαίες αποκλίσεις, μετρητικά σφάλματα κλπ.

Ουσιαστικά, αυτό που θέλουμε, είναι να ξεχωρίσουμε αυτές τις δύο διαφορετικές επιδράσεις και να δούμε αν υπάρχουν ή όχι σημαντικές επιδράσεις των υπό μελέτη παραγόντων στα δείγματα του πληθυσμού.

Πληροφοριακά, οι παράγοντες αναφέρονται και ως “treatments” (επεξεργασία ή κατεργασία). Η ορολογία αυτή προέρχεται από έρευνες στον τομέα της γεωπονίας όπου εφαρμόζονταν διαφορετικά “treatments” (με λιπάσματα, φυτοφάρμακα κλπ) σε διάφορα είδη φυτών.

<u>Παράδειγμα 1</u>: Ισομεγέθη δείγματα
Παρατηρούμε 4 ομάδες παιδιών που παίζουν με παιγνίδια διαφορετικού χρώματος και μετράμε πόσο χρόνο αφιερώνουν σε κάθε είδος παιγνιδιού.
Εναλλακτικά, η ίδια ακριβώς διαδικασία θα μπορούσε να εφαρμοστεί σε προβλήματα όπως:

- Μελέτη χημικής διεργασίας σε διαφορετικά μοντέλα αντιδραστήρα, με διαφορετικούς καταλύτες (μέτρηση χρόνου για ολοκλήρωση αντίδρασης) για να επιλεγεί ο αποτελεσματικότερος καταλύτης.
- Καταγραφή ανάπτυξης ομάδων φυτών σε διάφορα εδάφη με διαφορετικό λίπασμα για να επιλεγεί το καταλληλότερο

κλπ.

Στον επόμενο πίνακα καταγράφονται οι χρονικές διάρκειες παιγνιδιού και στις τελευταίες γραμμές, ο αριθμός παιδιών κάθε ομάδας, οι μέσες τιμές και οι διασπορές (ανά χρώμα παιγνιδιού).

#1	#2	#3	#4
κόκκινο	κίτρινο	πράσινο	μπλε
..... 10 τιμές ανά ομάδα			
n_1=10	n_2 = 10	n_3 = 10	n_4 = 10
$<x_1>$=3.4	$<x_2>$ = 5.0	$<x_3>$ = 2.4	$<x_4>$ = 2.5
s_1^2=4.5	s_2^2=5.6	s_3^2=1.2	s_4^2=1.8

Συνολικά
N = 40
$<x>_{ολικό}$ = 3.3
$s^2_{ολικό}$ = 4.1

Εδώ, οι παράγοντες είναι οι #1, #2, #3, #4 (π.χ. κόκκινο, κίτρινο, πράσινο, μπλε).
Η διακύμανση κατά παράγοντες επομένως είναι αυτή που υπολογίζουμε από τις μέσες τιμές κάθε ομάδας, δηλαδή η διακύμανση των αριθμών 4.5, 5.6, 1.2 και 1.8.
Άρα, η διακύμανση κατά παράγοντες είναι χαρακτηριστικό της *κατανομής δειγματοληψίας* και σχετίζεται με την κατανομή του πληθυσμού όπως ορίζει το *Κεντρικό Οριακό Θεώρημα*, δηλαδή:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Αλλά:

- δε γνωρίζουμε τη διασπορά της διακύμανσης της κατανομής δειγματοληψίας του πληθυσμού όλων των δυνατών παραγόντων, οπότε βασιζόμαστε σε μια εκτιμήτρια αυτής, το $s_{\bar{x}}$ που θα υπολογίσουμε από τους 4 αριθμούς, βάσει της σχέσης όπου διαιρούμε με το $n_{\bar{x}} - 1$ = 4-1 = 3.
  Άρα, αντικαθιστούμε το $\sigma_{\bar{x}}$ με το $s_{\bar{x}}$.
- Σύμφωνα με τη μεθοδολογία της ανάλυσης διασποράς θα προβούμε σε *διαφορετικές εκτιμήσεις της διασποράς του πληθυσμού* τις οποίες θα συγκρίνουμε μεταξύ τους, άρα αντικαθιστούμε επίσης, το σ με το s.
- Στον παρονομαστή υπεισέρχεται το μέγεθος δείγματος, άρα εδώ n = 10.

Επομένως:

$$s^2 = n\, s^2_{\bar{x}}$$

Εδώ βρίσκουμε

$$\mu_{\bar{x}} = 3.425$$

Επομένως οι διαφορές από τη μέση τιμή και τα τετράγωνά τους είναι
0,07500 0,00562
1,67500 2,80563
-0,92500 0,85563
-0,82500 0,68063
Αθροίζουμε τα τετράγωνα και διαιρούμε διά 3 για να βρούμε 1.45 και πολλαπλασιάζουμε επί μέγεθος δείγματος n = 10 για να βρούμε
s^2 = 10 (1.45) = 14.5
Αυτή είναι η διασπορά κατά παράγοντες.

Η διασπορά του σφάλματος για κάθε παράγοντα εκτιμάται αν πάρουμε τα τετράγωνα των τυπικών αποκλίσεων (διασπορές) για κάθε ομάδα, οπότε έστω ότι βρίσκουμε
s_1^2 = 4.5, s_2^2 = 5.6, s_3^2 = 1.2, s_4^2 = 1.8
Τώρα, για να βρούμε από αυτές μια δεύτερη εκτίμηση της διασποράς όλου του πληθυσμού, μπορούμε να πάρουμε το μέσο όρο των παραπάνω τιμών για να βρούμε s = 3.28

Τώρα, έχουμε *δύο εκτιμήσεις* για τη διασπορά του πληθυσμού:

- μία που προέρχεται από τις *πιθανές* διαφοροποιήσεις που προκαλούν οι διαφορετικοί παράγοντες κατά ομάδα
- και μία δεύτερη που μπορεί να προέρχεται από τυχαίους παράγοντες που είναι πέρα από τον έλεγχό μας και είναι κοινοί σε κάθε ομάδα – στο βαθμό που μπορούμε να εξασφαλίσουμε κάτι τέτοιο από το σχεδιασμό του πειράματος.

Υπάρχουν δύο περιπτώσεις:

- Αν οι παράγοντες που εξετάζουμε δεν προκαλούν καμία σημαντική διαφοροποίηση, τότε οι μέσες τιμές των ομάδων θα πρέπει να συγκλίνουν και η διασπορά των μέσων τιμών θα είναι μικρή. Επομένως, η πρώτη εκτίμηση της διασποράς περιμένω να είναι μικρότερη ή το πολύ περίπου ίση με τη διασπορά που παρατηρείται σε κάθε ομάδα λόγω τυχαίων παραγόντων.
- Αν οι παράγοντες που εξετάζουμε προκαλούν σημαντική διαφοροποίηση, τότε περιμένω μια διασπορά κατά παράγοντα, μεγαλύτερη από αυτή ή το πολύ περίπου ίση με αυτή λόγω τυχαίων παραγόντων.

Πώς μπορώ να εκτιμήσω το "περίπου ίση";
Έχω μια κατάσταση ανάλογη με αυτή όπου θέλουμε να δούμε πόσο σημαντική είναι η διαφορά των διασπορών δύο διαφορετικών δειγμάτων.
Επομένως μπορώ να χρησιμοποιήσω το λόγο F που θα οριστεί ως εξής:

F = διασπορά κατά παράγοντα : διασπορά λόγω σφάλματος

και εν προκειμένω βρίσκω F = 14.5 / 3.28 = 4.42

Για να χρησιμοποιήσουμε το κριτήριο F χρειαζόμαστε και τους βαθμούς ελευθερίας. Για τη διασπορά κατά παράγοντα, έχω 4 ομάδες άρα $df_1 = 4-1 = 3$. Για τη διασπορά λόγω σφάλματος έχουμε 10 μετρήσεις ανά ομάδα και 4 ομάδες, οπότε $df_2 = 4 \text{ X } (10-1) = 4 \text{ X } 9 = 36$.

Από πίνακες, για επίπεδο σημαντικότητας 1% βρίσκουμε ότι η κρίσιμη περιοχή είναι:
F >= 4.38.
Συμπεραίνουμε ότι ο λόγος F για το συγκεκριμένο παράδειγμα είναι (οριακά) μέσα στην κρίσιμη περιοχή, άρα η παρατηρούμενη διασπορά λόγω διαφορετικών παραγόντων είναι σημαντική.
Αν η μηδενική υπόθεση ήταν ότι οι παράγοντες δεν ευθύνονται για τη διασπορά, τότε αυτή πρέπει να απορριφθεί.

<u>Παράδειγμα 2</u>: Ανισομεγέθη δείγματα
Η λογική της μεθοδολογίας είναι παρόμοια, με τη διαφορά ότι οι μέσες τιμές και διασπορές πρέπει να *σταθμιστούν με βάση το μέγεθος των δειγμάτων*.
Οι υπολογισμοί γίνονται πιο εύκολα αν χρησιμοποιήσουμε αθροίσματα τετραγώνων αντί για τον ορισμό των διασπορών, όπως αποδεικνύεται κατωτέρω:

$$\sum_i (x_i - \bar{x})^2 = \sum_i x_i^2 - 2\sum_i x_i \bar{x} + \bar{x}^2 = \sum_i x_i^2 - 2N\bar{x}\bar{x} + \bar{x}^2 = \sum_i x_i^2 - N\bar{x}^2 =$$

$$= \sum_i x_i^2 - N\left(\frac{\sum_i x_i}{N}\right)^2 = \sum_i x_i^2 - \frac{\left(\sum_i x_i\right)^2}{N}$$

Έστω η επόμενη παραλλαγή του παραδείγματος με τις 4 ομάδες:

	#1		#2		#3		#4	
	x	x^2	x	x^2	x	x^2	x	x^2
	..... n_i τιμές ανά ομάδα i							
Sums	25	143	35	253	41	257	24	146
	n_1=5		n_2 = 6		n_3 = 7		n_4 = 4	
Συνολικά αθροίσματα:								
N = 22	$(\Sigma x)_{ολικό}$ = 125				$(\Sigma x_T^2)_{ολικό}$ = 799			

Ισχύει η ταυτότητα

Ολικό Διπλό Άθροισμα Τετραγώνων =
n φορές το (Άθροισμα Τετραγώνων Διασπορών κατά Παράγοντα)
+ Διπλό Άθροισμα Τετραγώνων Τυχαίων Διασπορών Ομάδων

γνωστή ως **ταυτότητα του Αθροίσματος Τετραγώνων για Ανάλυση Διασποράς** – (ANOVA sum of squares identity) οπότε, βρίσκουμε τα δύο που είναι πιο εύκολα και μέσω της ταυτότητας προσδιορίζουμε και το τρίτο.

Για ίσο μέγεθος δειγμάτων η ταυτότητα γράφεται:

$$\sum_{i=1}^{g}\sum_{j=1}^{n}\left(y_{ij}-\bar{y}_{..}\right)^2 = n\sum_{i=1}^{g}\left(\bar{y}_{i\cdot}-\bar{y}_{..}\right)^2 + \sum_{i=1}^{g}\sum_{j=1}^{n}\left(y_{ij}-\bar{y}_{i\cdot}\right)^2$$

όπου η τελεία συμβολίζει άθροισμα ως προς το δείκτη τον οποίο αντικαθιστά και η μπάρα παριστάνει διαίρεση αυτού του αθροίσματος με το αντίστοιχο πλήθος.
και συμβολικά:

$$SS_T = SS_t + SS_E$$

όπου T = total, t = treatment (αναφέρεται στους παράγοντες που εξετάζουμε) και E = error (αναφέρεται στους τυχαίους παράγοντες).

Για άνισα δείγματα, οι παραπάνω ποσότητες τροποποιούνται ως εξής:

$$SS_T = \sum_{i=1}^{g}\sum_{j=1}^{n} y_{ij}^2 - \frac{y_{..}^2}{N} \quad , \quad SS_t = \sum_{i=1}^{g}\frac{y_{i\cdot}^2}{n} - \frac{y_{..}^2}{N} \quad \text{και} \quad SS_E = SS_T - SS_t$$

όπου g σημαίνει “groups”, ομάδες.
Βάσει των παραπάνω αθροισμάτων, οι *εκτιμήτριες* των διασπορών ορίζονται ως εξής:
Κατά παράγοντες:

$MS_t = SS_t / (g-1)$

τείνει στο $\sigma^2 + n\Sigma_{i=1}^{g}\tau_i^2/(g-1)$, όπου τ_i οι διασπορές λόγω παραγόντων

Ως προς σφάλμα

$MS_E = SS_E / g(n-1)$

τείνει στο σ^2

Οπότε η μηδενική υπόθεση γράφεται
H_0: $\tau_1 = \tau_2 = ... \tau_g = 0$
και η εναλλακτική είναι τουλάχιστον ένα τ να διαφέρει από το μηδέν.
Το στατιστικό ελέγχου είναι ο λόγος $F_0 = MS_t / MS_E$
Κριτήριο απόρριψης: $F_0 > F_{\alpha,\ g-1,\ g(n-1)}$ όπου οι δείκτες είναι: α, βαθμός εμπιστοσύνης, g-1 βαθμοί ελευθερίας βάσει αριθμού ομάδων g και g(n-1) οι βαθμοί ελευθερίας βάσει συνολικού αριθμού παρατηρήσεων (θεωρήσαμε βαθμούς ελευθερίας κατά ομάδες, n-1 και πολλαπλασιάσαμε επί τον αριθμό ομάδων).

Τότε, για το συγκεκριμένο παράδειγμα, βρίσκουμε

$$\sum_{i=1}^{g}\sum_{j=1}^{n} y_{ij}^2 = 799 \quad , \quad y_{..}^2 = 125^2 = 15625 \quad ,$$

$SS_T = 799 - (125)^2/22 = 88.8$

Παρόμοια,
$SS_t = (25)^2/5 + (35)^2/6 + (41)^2/7 + (24)^2/4 - (125)^2/22 = 125+204.2 +240.1+144-710.2 = 3.1$
$SS_E = SS_T - SS_t = 88.8 - 3.1 = 85.7$

Συνηθίζεται να συμπληρώνουμε τον παρακάτω πίνακα Ανάλυσης Διασπορών Μονοπαραγοντικού Πειράματος.

Πηγή	**Άθροισμα**	**df**	**Μέσο Τετράγωνο (εκτίμηση διασποράς)**	**F_0**
Σύνολο	88.8	21	-----	
Παράγοντες	3.1	3	1.03	
Σφάλμα	85.7	18	4.76	

Οπότε το στατιστικό ελέγχου είναι F = 1.03 / 4.76 = 0.22 που είναι μικρότερο από τις τιμές της κατανομής του F τόσο για 1% όσο και για 5% (3.16 και 5.09, αντίστοιχα), άρα οι διαφορές οφείλονται σε τυχαίους και όχι στους εξεταζόμενους παράγοντες.

Προϋποθέσεις εφαρμογής ανάλυσης διασποράς

- Όπως και για τον έλεγχο με χρήση κατανομής του t, απαιτείται
  - κανονική κατανομή πληθυσμού
  - ίσες διασπορές δειγμάτων

  Στην πράξη, θέλουμε ιστογράμματα περίπου συμμετρικά και με παρόμοιο πλάτος.
- όταν έχουμε πάνω από δύο δείγματα.

Ερώτηση: Γιατί δεν εφαρμόζουμε το t (ή z) αλλά την ανάλυση διασπορών για περισσότερα από δύο δείγματα;
Η απάντηση έγκειται στο ότι οι μεταξύ τους συνδυασμοί είναι g (g – 1) / 2, δηλαδή αυξάνονται περίπου με το τετράγωνο του αριθμού δειγμάτων, g. Αν εφαρμόσουμε τον έλεγχο του t για κάθε ζεύγος, έχουμε μεγαλύτερη πιθανότητα για σφάλμα τύπου I, δηλαδή να επιβεβαιωθεί η εναλλακτική υπόθεση ενώ στην πραγματικότητα πρόκειται για τυχαίο αποτέλεσμα.
Π.χ. για επίπεδο σημαντικότητας 5% και 4 ομάδες, από διωνυμική κατανομή βρίσκουμε 17.1% πιθανότητα ότι τουλάχιστον ένα ζεύγος θα δώσει επαλήθευση της εναλλακτικής υπόθεσης κατά τύχη.