Ανάλυση διακύμανσης (Μονοδιάστατη)

One-Way ANOVA

Ανάλυση διακύμανσης

- Η μονοδιάστατη ανάλυση διακύμανσης εξετάζει εάν δύο ή περισσότεροι ανεξάρτητοι πληθυσμοί έχουν τον ίδιο η διαφορετικό μέσο όρο.
- Στην περίπτωση που έχουμε μόνο δύο ανεξάρτητους πληθυσμούς μπορούμε να διερευνήσουμε το ίδιο ερώτημα με το ανάλογο t-test
- Γιατί δεν μπορούμε να κάνουμε πολλαπλά t-test??

Ανάλυση διακύμανσης

- Η εξαρτημένη μεταβλητή (*dependent ή response* variable) είναι η μεταβλητή που συγκρίνουμε
- Η ανεξάρτητη μεταβλητή (*factor ή grouping* variable) είναι η μεταβλητή που ομαδοποιεί τα δείγματα στους διαφορετικούς πληθυσμούς που συγρκίνουμε
  - Η ανεξάρτητη μεταβλητή έχει k επίπεδα (όπου k≥2)
- Η ανάλυση διακύμανσης αναφέρεται ως μονοδιάστατη (*one way ANOVA*) γιατί έχουμε μία μόνο ανεξάρτητη μεταβλητή.

Προϋποθέσεις μονοδιάστατης ανάλυσης διακύμανσης

- Τα δεδομένα είναι αντιπροσωπευτικά
- Ομοιογένεια διακυμάνσεων
- Ακολουθούν την κανονική κατανομή

Μηδενική υπόθεση

- Η *μηδενική υπόθεση* στη μονοδιάστατη ανάλυση διακύμανσης είναι ότι όλοι οι μέσοι όροι είναι ίσοι

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$$

- Η *εναλλακτική υπόθεση* είναι ότι τουλάχιστον ένας μέσος όρος διαφέρει σημαντικά.

Παράδειγμα

- Έχει παρατηρηθεί ότι οι μαθητές στην αίθουσα διδασκαλίας μπορεί να διακριθούν ανάλογα με το που κάθονται (είτε στις μπροστά θέσεις, είτε στις μεσαίες, είτε στις πίσω)
- Έχει παρατηρηθεί ότι οι μαθητές των τελευταίων εδράνων είναι συνήθως πιο ανήσυχοι, κάνουν περισσότερο βαβούρα, παίζουν με τα κινητά τους, κ.ο.κ.
- Το ερώτημα λοιπόν είναι **εάν το που κάθεται ο μαθητής επηρεάζει τις επιδόσεις του**

Ανάλυση διακύμανσης

Η ανάλυση διακύμανσης δεν εξετάζει εάν ο ένας μέσος όρος είναι μικρότερος από τον άλλο, αλλά εάν είναι ίσοι ή διαφέρουν.

Παράδειγμα συνέχεια

- Έστω ότι πήραμε τυχαίο δείγμα από μαθητές ανάλογα με το που κάθονται
- Καταγράψαμε τις επιδόσεις τους σε μια δοκιμασία κατανόησης του μαθήματος και έχουμε σκορ:
  - Πρώτα: 82, 83, 97, 93, 55, 67, 53
  - Μέσαια: 83, 78, 68, 61, 77, 54, 69, 51, 63
  - Τελευταία: 38, 59, 55, 66, 45, 52, 52, 61

Περιγραφικά στατιστικά δείγματος

	Πρώτα	Μεσαία	Τελευταία
Μέγεθος δείγματος			
Μέσος όρος			
Τυπική απόκλιση			
Διακύμανση			

Περιγραφικά στατιστικά δείγματος

	Πρώτα	Μεσαία	Τελευταία
Μέγεθος δείγματος	7	9	8
Μέσος όρος	75.71	67.11	53.50
Τυπική απόκλιση	17.63	10.95	8.96
Διακύμανση	310.90	119.86	80.29

Διακύμανση

- Είναι όλες οι τιμές ίδιες?
  - Όχι άρα υπάρχει διακύμανση στα δεδομένα
  - Η διακύμανση μπορεί να μετρηθεί για το σύνολο των δεδομένων
  - Ως SS(Total) αναφερόμαστε στο άθροισμα των τετραγώνων των αποκλίσεων από το μέσο όρο
  - Το άθροισμα αυτό είναι ο αριθμητής στον υπολογισμό της διακύμανσης

Διακύμανση

- Είναι οι τιμές του κάθε πληθυσμού ίδιες?
  - Όχι υπάρχει διακύμανση των παρατηρήσεων και σε κάθε επιμέρους πληθυσμό
  - Ως SS(Within) αναφερόμαστε στο μέρος του αθροίσματος των τετραγώνων που οφείλεται ακριβώς στη διακύμανση μεταξύ των ατόμων του κάθε πλυθησμού

Διακύμανση

- Είναι η διακύμανση των επιμέρους πληθυσμών ίδια με τη συνολική διακύμανση?
  - Υπάρχει μέρος της διακύμανσης που παρατηρείται **εντός** του κάθε επιμέρους πληθυσμού και υπάρχει επιπλέον διακύμανση **μεταξύ** των πληθυσμών
  - Με τον όρο SS(Between) αναφερόμαστε στο μέρος του αθροίσματος των τετραγώνων που οφείλεται ακριβώς στη διακύμανση μεταξύ των πληθυσμών

Ανάλυση διακύμανσης

- Άρα υπάρχουν δύο πηγές διακύμανσης
  - Η διακύμανση εντός του κάθε πληθυσμού, SS(W), ή αλλιώς η φυσική διακύμανση των ατόμων ή η διακύμανση που οφείλεται στο τρόπο εκτίμησης κ.ο.κ. (άρα διακύμανση που δεν οφείλεται στον παράγοντα ενδιαφέροντος)
  - Η διακύμανση μεταξύ των πληθυσμών, SS(B), που είναι η διακύμανση που οφείλεται στη δράση του παράγοντα που μελετάμε (της ανεξάρτητης μεταβλητής)

Ανάλυση διακύμανσης

- Ο βασικός πίνακας της ανάλυσης διακύμανσης

Πηγή	SS	df	MS	F	p
Εντός					
Μεταξύ					
Συνολικά					

Συνολικά

- Ο μέσος όρος όλων των δεδομένων
- Ο μέσος όρος αυτός υπολογίζεται εάν αγνοήσουμε τον παράγοντα ενδιαφέροντος

- Στο παράδειγμά μας είναι 65.08

Διακύμανση μεταξύ των πληθυσμών

- SS(Between)
  - Η διακύμανση μεταξύ του μέσου όρου του κάθε επιμέρους πληθυσμού και του συνολικού μέσου όρου
  - Η διαφορές αυτές είναι σταθμισμένες με το μέγεθος δείγματος κάθε πληθυσμού

$$SS\left(B\right)=\sum_{i=1}^{k}n_i\left(\overline{x}_i-\overline{\overline{x}}\right)^2$$

Διακύμανση μεταξύ των πληθυσμών (παράδειγμα)

SS(B)=1902

$$SS\left(B\right)=7\left(75.71-65.08\right)^{2}+9\left(67.11-65.08\right)^{2}+8\left(53.50-65.08\right)^{2}$$

Διακύμανση εντός των πληθυσμών

- SS(Within)
  - Είναι το άθροισμα των σταθμισμένων διακυμάνσεων κάθε επιμέρους πληθυσμού
  - Η στάθμιση γίνεται με τους βαθμούς ελευθερίας (df)
  - Οι βαθμοί ελευθερίας (df) για κάθε επιμέρους πληθυσμό είναι το μέγεθος δείγματος (κάθε πληθυσμού) μείον ένα

Διακύμανση εντός των πληθυσμών

$$SS(W) = \sum_{i=1}^{k} df_i s_i^2$$

$$SS(W) = df_1 s_1^2 + df_2 s_2^2 + \cdots + df_k s_k^2$$

- Στο παράδειγμά μας SS(Within) 3386

Ανάλυση διακύμανσης

- Ο βασικός πίνακας της ανάλυσης διακύμανσης διαμορφώνεται ως εξής

Πηγή	SS	df	MS	F	p
Μεταξύ	1902				
Εντός	3386				
Συνολικά	5288				

Βαθμοί ελευθερίας (df)

- Ένας βαθμός ελευθερίας αντιστοιχεί σε κάθε τιμή που μπορεί να ποικίλει μέχρις ότου οι υπόλοιπες τιμές γίνουν υποχρεωτικές
- Συχνά οι βαθμοί ελευθερίας υπολογίζονται ως το μέγεθος δείγματος μείον ένα

Βαθμοί ελευθερίας

- Οι βαθμοί ελευθερίας για μεταξύ των πληθυσμών είναι ο αριθμός των ανεξάρτητων πληθυσμών που συγκρίνουμε μείον ένα
  - Στο παράδειγμα έχουμε 3 πληθυσμούς άρα df(B) = 2
- Οι βαθμοί ελευθερίας εντός των πληθυσμών είναι το άθροισμα των βαθμών ελευθερίας των επιμέρους πληθυσμών (όπου για κάθε πληθυσμός είναι μέγεθος δείγματος μείον 1)
  - df(W) = 6 + 8 + 7 = 21
- Οι βαθμοί ελευθερίας συνολικά είναι το συνολικό μέγεθος δείγματος μείον ένα
  - df(Total) = 24 – 1 = 23

Ανάλυση διακύμανσης

- Ο βασικός πίνακας της ανάλυσης διακύμανσης διαμορφώνεται ως εξής

Πηγή	SS	df	MS	F	p
Μεταξύ	1902	2			
Εντός	3386	21			
Συνολικά	5288	23			

Σταθμισμένο άθροισμα τετραγώνων

- MS = SS / df

Παράδειγμά μας

- MS(B) = 1902 / 2 = 951.0
- MS(W) = 3386 / 21 = 161.2
- MS(T) = 5288 / 23 = 229.9
  - MS(Total) **δεν** είναι το άθροισμα του MS(Between) και του MS(Within).

Ανάλυση διακύμανσης

- Ο βασικός πίνακας της ανάλυσης διακύμανσης διαμορφώνεται ως εξής

Πηγή	SS	df	MS	F	p
Μεταξύ	1902	2	951.0		
Εντός	3386	21	161.2		
Συνολικά	5288	23	229.9		

Ανάλυση διακύμανσης

- Η ανάλυση διακύμανσης στηρίζεται στον επιμερισμός της διακύμανσης εντός του κάθε πληθυσμού και μεταξύ των πληθυσμών
- F test statistic (αντίστοιχο του t)
  - Το F είναι ο λόγος των δύο σταθμισμένων αθροισμάτων τετραγώνων
  - F = MS(B) / MS(W)
- Για το παράδειγμά μας, F = 951.0 / 161.2 = 5.9

Ανάλυση διακύμανσης

- Ο βασικός πίνακας της ανάλυσης διακύμανσης διαμορφώνεται ως εξής

Πηγή	SS	df	MS	F	p
Μεταξύ	1902	2	951.0	5.9	
Εντός	3386	21	161.2		
Συνολικά	5288	23	229.9		

Ανάλυση διακύμανσης

- Το ερώτημα λοιπόν είναι **πόσο πιθανό να παρατηρηθεί αυτή η τιμή F εάν οι μέσοι όροι ήταν ίσοι**?
- Για το λόγο αυτό χρησιμοποιούμε την **κατανομή F** με βαθμούς ελευθερίας τόσο για το εντός όσο και για το μεταξύ (άρα στο παράδειγμά μας df (Between) 2 & df(Within) 21
- $P(F_{2,21} > 5.9) = 0.009$

Ανάλυση διακύμανσης

- Ο βασικός πίνακας της ανάλυσης διακύμανσης διαμορφώνεται ως εξής

Πηγή	SS	df	MS	F	p
Μεταξύ	1902	2	951.0	5.9	0.009
Εντός	3386	21	161.2		
Συνολικά	5288	23	229.9		

Συμπέρασμα

- Αφού $p = 0.009$, (άρα $p<0.05$), μπορούμε να απορρίψουμε τη μηδενική υπόθεση.
- Άρα δεν μπορούμε να ισχυριστούμε ότι όλοι οι μέσοι όροι είναι ίσοι

Συμπέρασμα

- Τα στοιχεία επαρκούν για τον ισχυρισμό ότι υπάρχει διαφορά στο μέσο όρο των επιδόσεων των φοιτητών αναλόγως με το που κάθονται στην αίθουσα
- Το αποτέλεσμα της ANOVA δε μας λέει ποια είναι η διαφορά (μόνο ότι υπάρχει)
- Για να περιγράψουμε τη διαφορά χρειαζόμαστε άλλα tests