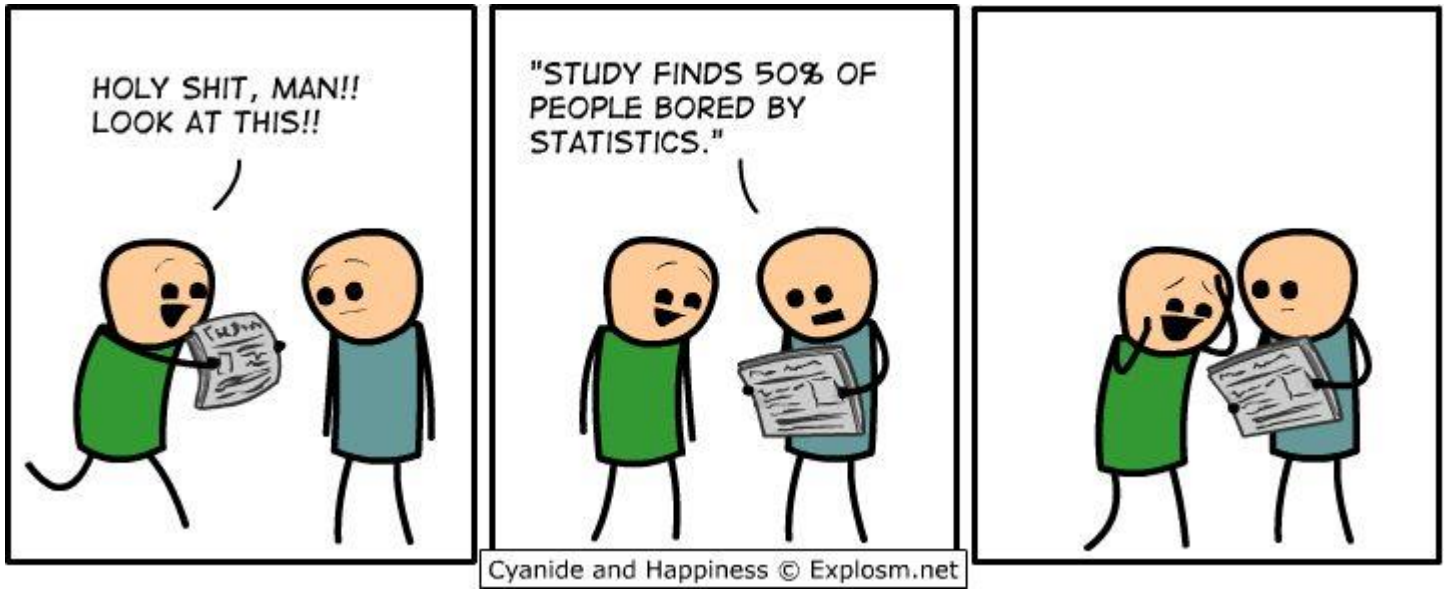
ΣΤΑΤΙΣΤΙΚΗ ΚΑΙ ΘΕΩΡΙΑ ΑΠΟΦΑΣΕΩΝ

ΔΙΑΛΕΞΗ: ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ

Διδάσκουσα: Ε. Γάκη, Επίκ. Καθηγήτρια


HOLY SHIT, MAN!!
LOOK AT THIS!!
"STUDY FINDS 50% OF
PEOPLE BORED BY
STATISTICS."
Cyanide and Happiness © Explosm.net

Περιεχόμενα Ενότητας

- Ανάλυση Διακύμανσης
  - Κατά έναν Παράγοντα
  - Κατά δύο Παράγοντες
    - Με μία παρατήρηση ανά κυψελίδα
    - Με m παρατηρήσεις ανά κυψελίδα

Περιεχόμενα Μαθήματος

- Ανάλυση Διακύμανσης
  - Κατά έναν Παράγοντα

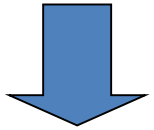
Ανάλυση Διακύμανσης

Η ανάλυση διακύμανσης (analysis of variance) είναι μια τεχνική που χρησιμεύει ως βάση για τη στατιστική ανάλυση μοντέλων που συνδέονται με τις **μεθόδους πειραματικών σχεδιασμών (experimental designs)**.

Ειδικότερα η ανάλυση διακύμανσης ασχολείται με τον προσδιορισμό των πηγών της μεταβλητότητας που παρατηρείται στα δειγματικά δεδομένα.

Ανάλυση Διακύμανσης

- Τα πειραματικά δεδομένα επηρεάζονται από ένα μεγάλο πλήθος πηγών μεταβλητότητας (sources of variation).
- Ένας καλός πειραματικός σχεδιασμός αποσκοπεί στο να εντοπίσει την κύρια πηγή μεταβλητότητας όπως επίσης και το ποσοστό της μεταβλητότητας που οφείλεται σε καθένα από τους διαφορετικούς παράγοντες που μας ενδιαφέρει να εξετάσουμε.
- Η υπόλοιπη διακύμανση των δεδομένων θεωρείται ότι οφείλεται σε τυχαίους παράγοντες και για το λόγο αυτό ονομάζεται λάθος (error).
- Ένας καλός σχεδιασμός ελαττώνει τη διακύμανση του λάθους όσο το δυνατό περισσότερο έτσι ώστε οι διαφορές της διακύμανσης που οφείλονται στους λόγους που μας ενδιαφέρουν να καθορισθούν όσο το δυνατόν ακριβέστερα.



Η ανάλυση διακύμανσης αποσκοπεί ακριβώς στο να καθορίσει, αφενός μεν όλες τις πηγές που συνεισφέρουν στη συνολική διακύμανση, αφετέρου δε το ποσοστό της διακύμανσης που μπορεί να αποδοθεί σε κάθε μια από τις πηγές αυτές.

Ανάλυση Διακύμανσης

- Η απλούστερη εφαρμογή της ανάλυσης διακύμανσης στοχεύει στον έλεγχο της υπόθεσης ότι οι μέσοι ορισμένων πληθυσμών δε διαφέρουν. Με αυτήν την έννοια η ανάλυση διακύμανσης είναι μια γενίκευση των ελέγχων για τη διαφορά δύο μέσων όπως αυτοί παρουσιάστηκαν στη Στατιστική Α (χρήση της κατανομής t του Student).

- Προϋποθέσεις:
  - **1η Υπόθεση:** Οι πληθυσμοί κατανέμονται κανονικά.
  - **2η Υπόθεση:** Οι πληθυσμοί έχουν ίσες διακυμάνσεις (ομοιογένεια διακύμανσης -homogeneity of variance).

Ανάλυση Διακύμανσης κατά Έναν Παράγοντα

Πλήρως Τυχαιοποιημένος Σχεδιασμός

Στον Πλήρως Τυχαιοποιημένο Σχεδιασμό (Completely Randomized Design) θέλουμε να ελέγξουμε τις διαφορές μεταξύ των μέσων k πληθυσμών όταν οι μονάδες τοποθετούνται τυχαία σε καθεμία από k ομάδες.

Κάθε ομάδα συνδέεται με ένα από τα k επίπεδα μεταχείρισης ή αγωγές (treatments or treatment levels).

Ανάλυση Διακύμανσης κατά Έναν Παράγοντα

Έχουμε k ανεξάρτητους πληθυσμούς και θέλουμε να ελέγξουμε την υπόθεση

H_0: $\mu_1 = \mu_2, \ldots, = \mu_\kappa$

Έναντι της εναλλακτικής

H_1: δύο τουλάχιστον μέσοι διαφέρουν.

Υποθέτουμε ότι:

- Οι πληθυσμοί κατανέμονται κανονικά.
- Οι πληθυσμοί έχουν ίσες διακυμάνσεις.

Ανάλυση Διακύμανσης κατά Έναν Παράγοντα

Έστω ότι παίρνουμε k τυχαία ανεξάρτητα δείγματα, μεγέθους $n_1, \ldots, n_k$ αντίστοιχα, ένα από κάθε επίπεδο μεταχείρισης/αγωγής (*οι k αγωγές, μπορεί να είναι k διαφορετικά λιπάσματα σε ένα γεωργικό πειραματικό σχεδιασμό, k μέθοδοι διδασκαλίας ενός γνωστικού αντικειμένου ή μιας δεξιότητας, k διαφορετικές μηχανές για την παραγωγή ενός προϊόντος κλπ.)*

Τότε θα έχουμε συνοπτικά ότι:

	Α Γ Ω Γ Ε Σ						
	1	2	. . .	i	. . .	k	
	Y_{11}	Y_{21}	. . .	Y_{i1}	. . .	Y_{k1}	
	Y_{12}	Y_{22}	. . .	Y_{i2}	. . .	Y_{k2}	
	.	.	.	.	.	.	
	.	.	.	.	.	.	
	.	.	.	.	.	.	
	Y_{1n}	Y_{2n}	. . .	Y_{in}	. . .	Y_{kn}	
Σύνολα	$Y_{1.}$	$Y_{2.}$	. . .	$Y_{i.}$	. . .	$Y_{k.}$	$Y_{..}$
Μέσοι	$\overline{Y}_{1.}$	$\overline{Y}_{2.}$	. . .	$\overline{Y}_{i.}$	. . .	$\overline{Y}_{k.}$	$\overline{Y}_{..}$

Ανάλυση Διακύμανσης κατά Έναν Παράγοντα

Η αντικατάσταση ενός δείκτη από τελεία υποδηλώνει άθροιση ως προς το δείκτη αυτό. Συγκεκριμένα:

$$Y_{i.} = \sum_{j=1}^{n_i} Y_{ij} \qquad Y.. = \sum_{i=1}^{k} Y_{i.} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}$$

$$\overline{Y}_{i.} = Y_{i.} / n_i \qquad \overline{Y}.. = Y../n = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n} Y_{ij}}{n}$$

$\bar{Y}_{..}$

Ανάλυση Διακύμανσης κατά Έναν Παράγοντα

Ένα μέτρο της συνολικής μεταβλητότητας (total variation) μεταξύ όλων των παρατηρήσεων μπορεί να υπολογισθεί αθροίζοντας τα τετράγωνα των αποκλίσεων κάθε παρατήρησης από τον μέσο

$$\bar{Y}_{..}$$

ο οποίος βασίζεται σε όλες τις παρατηρήσεις.

Ανάλυση Διακύμανσης κατά Έναν Παράγοντα

Με άλλα λόγια η συνολική μεταβλητότητα μπορεί να υπολογισθεί βάσει του τύπου:

Συνολική Μεταβλητότητα = $\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}..\right)^2$

όπου:

k = πλήθος αγωγών, μεταχειρίσεων, ομάδων, στηλών

n_i = πλήθος μονάδων που υφίστανται την i αγωγή

Y_{ij} = η j παρατήρηση που υφίσταται την i αγωγή

Η συνολική μεταβλητότητα ονομάζεται και Συνολικό Άθροισμα Τετραγώνων (total sum of squares) και συμβολίζεται με SST.

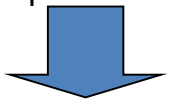
$\bar{Y}_{..}$

Ανάλυση Διακύμανσης κατά Έναν Παράγοντα

Γιατί υπάρχει διαφορά μεταξύ των παρατηρήσεων;

Υπάρχουν διαφορές τόσο μεταξύ των τιμών της ίδιας ομάδας όσο και μεταξύ των μέσων των διαφόρων ομάδων.

- Οι διαφορές μεταξύ των τιμών της ίδιας ομάδας οφείλονται στις ιδιαιτερότητες των μονάδων που την αποτελούν. Παρά το γεγονός ότι οι μονάδες της ίδιας ομάδας αντιμετωπίζονται με τον ίδιο τρόπο (υποβάλλονται στην ίδια αγωγή) οι αποδόσεις τους διαφέρουν.
- Οι διαφορές μεταξύ των μέσων των διαφόρων ομάδων οφείλονται εν μέρει στη διαφορετική αντιμετώπισή τους (διαφορετική αγωγή) και εν μέρει στις ιδιαιτερότητες των μονάδων που τις απαρτίζουν. Έτσι ακόμα και στην περίπτωση που οι ομάδες υποβληθούν στην ίδια αγωγή και πάλι θα υπάρχουν αποκλίσεις μεταξύ των μέσων τους.



Κατά συνέπεια αν η μηδενική υπόθεση ισχύει τότε οι διαφορές μεταξύ των μέσων θα οφείλονται αποκλειστικά στις ιδιαιτερότητες των μονάδων που τις αποτελούν. Στην περίπτωση που η μηδενική υπόθεση δεν ισχύει τότε οι διαφορές μεταξύ των μέσων θα είναι προφανώς μεγαλύτερες καθώς θα οφείλονται τόσο στη διαφορετική αγωγή όσο και στις ιδιαιτερότητες των μονάδων. Η διαπίστωση αυτή αποτελεί τη βάση για το στατιστικό έλεγχο των διαφορών των μέσων των ομάδων.

$\bar{Y}_{..}$

Ανάλυση Διακύμανσης κατά Έναν Παράγοντα

Η συνολική μεταβλητότητα μπορεί να αναλυθεί σε δύο συνιστώσες:

- η μία εκφράζει τη μεταβλητότητα μεταξύ των αγωγών (between treatments variation) δηλαδή τη μεταβλητότητα μεταξύ των ομάδων (στηλών)
- η άλλη εκφράζει την εντός των αγωγών μεταβλητότητα (within treatments variation) δηλαδή την εντός των ομάδων (στηλών) μεταβλητότητα.

Ανάλυση Διακύμανσης κατά Έναν Παράγοντα

Μεταξύ των αγωγών μεταβλητότητα (between treatments variation)

$$\sum_{i=1}^{k} n_i \left(\bar{Y}_{i.} - \bar{Y}.. \right)^2$$

όπου:

k	=	πλήθος αγωγών, μεταχειρίσεων, ομάδων, στηλών
n_i	=	πλήθος μονάδων που υφίστανται την i αγωγή
$\bar{Y}..$	=	ο μέσος της i αγωγής, μεταχείρισης, ομάδας, στήλης

Η μεταξύ των αγωγών μεταβλητότητα ονομάζεται και άθροισμα τετραγώνων μεταξύ αγωγών (treatment sum of squares) συμβολίζεται δε με SSTr.

Ανάλυση Διακύμανσης κατά Έναν Παράγοντα

Εντός των αγωγών μεταβλητότητα (within treatments variation)

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\overline{Y}_{i.}\right)^2$$

όπου:

k = πλήθος αγωγών, μεταχειρίσεων, ομάδων, στηλών.

n_i = πλήθος μονάδων που υφίστανται την i αγωγή

Y_{ij} = η j παρατήρηση που υφίσταται την i αγωγή

$\overline{Y}_{..}$ = ο μέσος της i αγωγής, μεταχείρισης, ομάδας, στήλης

Η εντός των αγωγών μεταβλητότητα ονομάζεται και άθροισμα τετραγώνων λαθών (error sum of squares) συμβολίζεται δε με SSE .

Ανάλυση Διακύμανσης κατά Έναν Παράγοντα

Αποδεικνύεται και θεωρητικά ότι ισχύει η σχέση SST = SSTr + SSE

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}..\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - \frac{Y_{..}^2}{n}$$

$$SSTr = \sum_{i=1}^{k} n_i \left(\overline{Y}_{i.} - \overline{Y}..\right)^2 = \sum_{i=1}^{k} \frac{Y_{i.}^2}{n_i} - \frac{Y_{..}^2}{n}$$

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{i.}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - \sum_{i=1}^{k} \frac{Y_{i.}^2}{n_i}$$

Ανάλυση Διακύμανσης κατά Έναν Παράγοντα

- Για να προσδιορίσουμε αν οι μέσοι των διαφόρων ομάδων που υφίστανται τις διαφορετικές αγωγές είναι όλοι ίσοι μπορούμε να συγκρίνουμε δύο διαφορετικές εκτιμήτριες της διακύμανσης του πληθυσμού. Η μία εκτιμήτρια βασίζεται στο μεταξύ των ομάδων άθροισμα τετραγώνων δηλαδή στο SSTr. Η άλλη εκτιμήτρια στηρίζεται στο άθροισμα τετραγώνων των λαθών δηλαδή στο SSE.

- Αν η μηδενική υπόθεση είναι αληθής οι εκτιμήσεις που θα προκύψουν από τις δύο αυτές εκτιμήτριες θα είναι περίπου ίσες. Στην περίπτωση όμως που η μηδενική υπόθεση δεν είναι αληθής η εκτίμηση που βασίζεται στο SSTr θα είναι μεγαλύτερη.

Ανάλυση Διακύμανσης κατά Έναν Παράγοντα

Όπως είναι γνωστό η διακύμανση εκτιμάται γενικά, από το πηλίκο του αθροίσματος των τετραγώνων των αποκλίσεων των παρατηρήσεων από τον αντίστοιχο μέσο τους δια των καταλλήλων βαθμών ελευθερίας.

Το πηλίκο αυτό ονομάζεται **μέσο τετραγωνικό σφάλμα (mean squared error).**

Ανάλυση Διακύμανσης κατά Έναν Παράγοντα

Η μεταξύ των αγωγών διακύμανση εκτιμάται από το

$$MSTr = \frac{SSTr}{k - 1}$$

Η εντός των αγωγών διακύμανση εκτιμάται από το

$$MSE = \frac{SSE}{n - k}$$

Τότε, αν Ho είναι σωστή, έχουμε ότι

$$F_o = \frac{MSTr}{MSE} \sim F_{k-1,n-k}$$

Απορρίπτουμε την Ho αν $F_o > F_{k-1,\, n-k,\, 1-\alpha}$

Ανάλυση Διακύμανσης κατά Έναν Παράγοντα

Πίνακας Ανάλυσης Διακύμανσης κατά Έναν Παράγοντα
(One-Way ANOVA Table)

Πηγή Μεταβλητότητας	Αθροίσματα Τετραγώνων S S	Βαθμοί Ελευθερίας D F	Μέσα Τετραγωνικά Σφάλματα M S	F (κάτω από την H_0)
Μεταξύ αγωγών (between treatments)	SSTr	k-1	MSTr=SSTr/k-1	F_o=MSTr/MSE
Εντός των αγωγών (σφάλμα) (within treatments)	SSE	n-k	MSE=SSE/n-k	
Σύνολο	SST	n-1		

Παράδειγμα

Ο Πίνακας περιέχει τις μηνιαίες πωλήσεις (σε χιλιάδες €) μιας εταιρείας κινητής τηλεφωνίας που έχει 14 καταστήματα σε τρεις διαφορετικές πόλεις.

α. Να δημιουργηθεί ο Πίνακας Ανάλυσης Διακύμανσης.
β. Να ελεγχθεί σε επίπεδο σημαντικότητας α=0.05 η υπόθεση ότι οι πωλήσεις δεν επηρεάζονται από την πόλη όπου βρίσκονται τα καταστήματα.

Καταστήματα Πόλης 1 (Y_{1i})	2	3	4				
Καταστήματα Πόλης 2 (Y_{2i})	4	5	5	6	4	5	6
Καταστήματα Πόλης 3 (Y_{3i})	2	3	3	4			

Παράδειγμα

Κατ. Πόλης. 1	Κατ. Πόλης. 2	Κατ. Πόλης. 3	
2	4	2	
3	5	3	
4	5	3	
	6	4	
	4		
	5		
	6		
Y_1 =9	Y_2 =35	Y_3 =12	Y =56
$\bar{Y}_{1.} = 3$	$\bar{Y}_{2.} = 5$	$\bar{Y}_{3.} = 3$	$\bar{Y}_{..} = 4$

n =14 (n_1=3, n_2=7, n_3=4), k=3, α=0.05

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^{\ 2} = 4+9+16+16+25+25+36+16+25+36+4+9+9+16 = 246$$

$$\frac{Y_{..}^2}{n} = \frac{56^2}{14} = \frac{3136}{14} = 224$$

$$\sum_{i=1}^{k}\frac{Y_{i.}^2}{n_i} = \frac{9^2}{3} + \frac{35^2}{7} + \frac{12^4}{4} = \frac{81}{3} + \frac{1225}{7} + \frac{144}{4} = 27+175+36 = 238$$

Παράδειγμα

$$SST = 246 - 224 \rightarrow SST = 22$$

$$SST_r = 238 - 224 \rightarrow SST_r = 14 \rightarrow MST_r = \frac{SST_r}{k-1} = \frac{14}{2} \rightarrow MST_r = 7$$

$$SSE = SST - SST_r \rightarrow SSE = 8 \rightarrow MSE = \frac{SSE}{n-k} = \frac{8}{11} \rightarrow MSE = 0.73$$

$$\Rightarrow F_0 = \frac{7}{0.73} \rightarrow F_0 \cong 9.59$$

Πίνακας Ανάλυσης Διακύμανσης_

ΠΗΓΗ ΜΕΤΑΒΛΗΤΟΤΗΤΑΣ	SS	DF	MS	F_0
Μεταξύ Αγωγών	14	k-1=2	7	9.59
Εντός Αγωγών	8	n-k=11	0.73	
ΣΥΝΟΛΟ	22	n-1=13		

Παράδειγμα

Έλεγχος Υποθέσεων:

H_0: Οι πωλήσεις δεν επηρεάζονται από την πόλη όπου βρίσκονται τα καταστήματα (μι=μ2=μ3)

H_1: Οι πωλήσεις επηρεάζονται από την πόλη όπου βρίσκονται τα καταστήματα (δύο τουλάχιστον μέσοι διαφέρουν

Κριτήριο Απόρριψης:

Απορρίπτουμε την H_0 αν $F_0 > F_{k-1,n-k,1-a}$

Στη συγκεκριμένη περίπτωση

$F_0 = 9,59$

$a = 0,05$

$F_{k-1, n-k, 1-a} = F_{2, 11, 0.95} = 3.982$

Συμπέρασμα:

Επειδή $F_0 = 9,59 > 3.982 = F_{2, 11, 0.95}$ απορρίπτουμε την H_0 και δεχόμαστε την H_1. Δηλαδή δεχόμαστε ότι η οι πωλήσεις επηρεάζονται από την πόλη όπου βρίσκονται τα καταστήματα

Παράδειγμα

Η αντίστοιχη λύση στο EXCEL

Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Column 1	3	9	3	1		
Column 2	7	35	5	0,666667		
Column 3	4	12	3	0,666667		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	14	2	7	9,625	0,003834	3,982298
Within Groups	8	11	0,727273			
Total	22	13				