ΑΝΑΛΥΣΗ ΔΙΑΣΠΟΡΑΣ΄Η ΔΙΑΚΥΜΑΝΣΗΣ
(ANALYSIS OF VARIANCE – VARIANCE ANALYSIS – ANOVA

ANOVA

Αγλαΐα Καλαματιανού
ΣΗΜΕΙΩΣΕΙΣ ΔΙΔΑΣΚΑΛΙΑΣ ΓΙΑ ΤΟ ΜΑΘΗΜΑ "ΣΤΑΤΙΣΤΙΚΗ ΙΙ: ΕΙΣΑΓΩΓΗ ΣΤΙΣ ΜΕΘΟΔΟΥΣ ΔΙΜΕΤΑΒΛΗΤΗΣ ΚΑΙ ΠΟΛΥΜΕΤΑΒΛΗΤΗΣ ΑΝΑΛΥΣΗΣ"
Τμήματα Κοινωνιολογίας και Ψυχολογίας Παντείου Πανεπιστημίου
Ακαδ. Έτος 2018-2019

Πηγές:
Healey, J. F. (2012) Statistics: A tool for social research, Ninth edition, Wadsworth
Agresti , A & Finlay, B. (2009) Statistical Methods for the social sciences, Prentice Hall
Γεωπονικό Πανεπιστήμιο Αθηνών/Γιώργος Κ. Παπαδόπουλος, www.aua.gr/gpapadopoulos
https://libguides.library.kent.edu/SPSS/OneWayANOVA

Η ανάλυση διακύμανσης προτάθηκε από τον Sir Ronald A. Fisher το 1918. Ευρέως έγινε γνωστή μετά το 1925 όταν εκδόθηκε το κλασικό πλέον βιβλίο του R. A. Fisher, Statistical Methods for Research Workers, στο οποίο είχε συμπεριλάβει και την ανάλυση διακύμανσης. Η ανάλυση διακύμανσης «γεννήθηκε»/προέκυψε κατά την ενασχόληση του Fisher με δύσκολα προβλήματα στατιστικής συμπερασματολογίας που εμφανίζονται στον γεωργικό πειραματισμό (πολλές πηγές μεταβλητότητας και συχνά εμφανιζόμενες ετερογένειες, και μάλιστα, προς διάφορες κατευθύνσεις του πειραματικού αγρού π.χ. ως προς τη γονιμότητα, την κλίση και την υγρασία των εδαφών, τις προηγούμενες καλλιέργειες, κτλ.). Η προσέγγιση της λύσης τέτοιου είδους προβλημάτων που πρότεινε ο Fisher, βασίζεται στην τυχαιοποίηση και στην επανάληψη και ως μαθηματικό εργαλείο για την υποστήριξη αυτής της προσέγγισης πρότεινε την ανάλυση διακύμανσης. Γι' αυτό, όπως θα δούμε στη συνέχεια, στην ανάλυση διακύμανσης έχει επικρατήσει να χρησιμοποιείται ορολογία (και όχι μόνο) που χρησιμοποιείται στον γεωργικό πειραματισμό και γενικότερα στον πειραματισμό, παρότι δεν εφαρμόζεται μόνο στην ανάλυση πειραματικών δεδομένων.

Πηγή: Γεωπονικό Πανεπιστήμιο Αθηνών/Γιώργος Κ. Παπαδόπουλος (www.aua.gr/gpapadopoulos

ΓΕΝΙΚΑ :

Η ανάλυση της διασποράς (ANOVA) είναι μια ευρέως διαδεδομένη μέθοδος ελέγχου σημαντικότητας (test of significance), ή άλλως, ελέγχου υποθέσεων αναφορικά με την σύγκριση των μέσων τιμών τριών ή περισσοτέρων πληθυσμών (συχνά αναφέρονται και ως ομάδες).

Συνεπώς η ανάλυση της διασποράς (ANOVA) μπορεί να θεωρηθεί ως μια επέκταση των στατιστικών ελέγχων που αφορούν στη σύγκριση των μέσων τιμών δύο πληθυσμών (που μας είναι γνωστοί πχ. *t*-test, σύγκριση των μέσων φοιτητικών επιδόσεων σε δύο πανεπιστήμια ή μεταξύ ανδρών και γυναικών)

Παράδειγμα: (Healey, J. F. (2012) Statistics: A tool for social research, Ninth edition, Wadsworth, σελ. 242)

Είναι γνωστό ότι η θανατική ποινή έχει μια ηθική διάσταση και μπορεί να επηρεαστεί από το θρησκευτικό υπόβαθρο ενός ατόμου.

Ενδιαφέρει επομένως να εξεταστεί αν η υποστήριξη της θέσης περί θανατικής ποινής –εξαρτημένη μεταβλητή- (μετρούμενη σε μια κλίμακα διαστήματος/αναλογίας –ποσοτική κλίμακα) διαφέρει μεταξύ πληθυσμών με διαφορετικό θρήσκευμα δηλαδή μεταξύ πληθυσμών που ορίζονται από τις τιμές μιας (ανεξάρτητης) μεταβλητής (πχ. Προτεστάντες, Καθολικοί, Εβραίοι, άθρησκοι, κ.ο.κ)

Μια τέτοια διερεύνηση μπορεί να γίνει συγκρίνοντας τη μέση (τιμή) υποστήριξη μεταξύ πληθυσμών

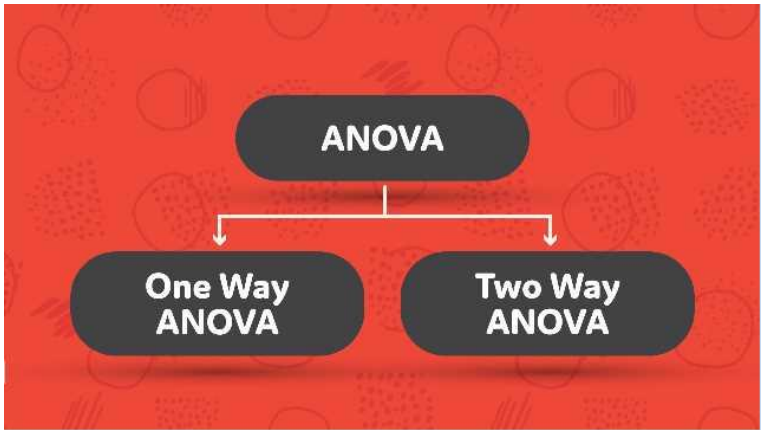
Τότε αν το ενδιαφέρον ήταν η σύγκριση μεταξύ δύο πληθυσμών (πχ. Προτεστάντες –άθρησκοι) τότε οι γνωστοί μας έλεγχοι θα μπορούσαν να εφαρμοστούν

Αν όμως το ενδιαφέρον είναι η σύγκριση μεταξύ τριών ή περισσοτέρων πληθυσμών (πχ. Προτεστάντες, Καθολικοί, Εβραίοι, άθρησκοι, κ.ο.κ) τότε μέθοδος της ανάλυσης της διασποράς (ANOVA) είναι κατάλληλη

Συμπερασματικά: Στην ANOVA (Analysis of Variance) συγκρίνουμε τους μέσους όρους (means) περισσότερων από δυο πληθυσμών (populations).

Ενα άλλο παράδειγμα: Σύγκριση της μέσης ετήσιας φαρμακευτικής δαπάνης των νοικοκυριών διαφορετικών περιοχών μιας χώρας)

Υπάρχουν δύο τύποι ANOVA



- Ο απλούστερος τύπος ANOVA ονομάζεται **one-way ANOVA** – **Ανάλυση της διασποράς κατά έναν παράγοντα** – μία ανεξάρτητη μεταβλητή οι τιμές τις οποίας ορίζουν τους πληθυσμούς. Διαφορετικά λέμε πως οι τιμές μιας μεταβλητής (κατηγορικής/ποιοτικής) ή ενός παράγοντα επηρεάζουν τις τιμές μιας ποσοτικής μεταβλητής

Εμείς θα ασχοληθούμε με αυτόν τον τύπο ανάλυσης της διασποράς (one-way ANOVA)

- **Two way ANOVA** – **Ανάλυση της διασποράς κατά δύο παράγοντες** – δύο ανεξάρτητες μεταβλητές οι τιμές των οποίων συνδυαστικά ορίζουν τους πληθυσμούς. Διαφορετικά λέμε πως οι τιμές δύο παραγόντων (δύο ανεξάρτητες κατηγορικές/ποιοτικές μεταβλητές) επηρεάζουν τις τιμές μιας ποσοτικής μεταβλητής

- **Πολυμεταβλητή ANOVA (***multivariate ANOVA***):** Πως πολλοί παράγοντες μαζί επηρεάζουν μια ποσοτική μεταβλητή

Τι ακριβώς ελέγχουμε με την One-Way ANOVA ;

1. Ελέγχουμε την υπόθεση ότι οι μέσες τιμές των πληθυσμών (ή ομάδων) είναι ίσες. Πιο συγκριμένα οι στατιστικές υποθέσεις διαμορφώνονται ως εξής:

Μηδενική υπόθεση **H_0:** $\mu_1 = \mu_2 = \ \ldots \ = \mu_k$

Εναλλακτική υπόθεση **H_1 :** $\mu_i \neq \mu_j$, *i, j* = 1,2,..,*k (τουλάχιστον ένα ζευγάρι διαφέρει).*

2. Αν η **H_0 απορριφθεί, τότε ίσως όλες οι μέσες τιμές διαφέρουν ή κάποιες από αυτές διαφέρουν ή ίσως μόνο μία διαφέρει.** Δεν μπορεί να μας πει ποιες ακριβώς ομάδες (πληθυσμοί) διαφέρουν. Για να δούμε που υπάρχουν διαφορές χρειαζόμαστε άλλες αναλύσεις (πχ. κατά ζεύγη ή post hoc tests)

3. Η στατιστική του ελέγχου είναι η *F-statistic* (*F-ratio*)

4. Η ύπαρξη έστω και μιας διαφοράς ερμηνεύεται ότι ο παράγοντας (κατηγορική μεταβλητή) επηρεάζει σημαντικά την ποσοτική μεταβλητή

Ως συνήθως ο ερευνητής επιθυμεί να απορρίψει την μηδενική υπόθεση

Ερώτηση: πως διαμορφώνονται οι υποθέσεις αναφορικά με το παράδειγμα υποστήριξη θανατικής ποινής και θρησκεύματος ?

Γιατί δεν χρησιμοποιούμε πολλά πχ. t-tests ώστε να ελέγξουμε συγκρίνουμε όλους τους δυνατούς συνδυασμούς μέσων τιμών και προτιμούμε την ANOVA?

1) Μια τέτοια διαδικασία πολλαπλών ελέγχων είναι αρκετά χρονοβόρα ακόμη και όταν το k είναι μικρό. Για παράδειγμα, αν θέλουμε να συγκρίνουμε k = 4 μέσες τιμές, πρέπει να κάνουμε 6 συνολικά ελέγχους, $\binom{4}{2}$=6. Αν k = 5 πρέπει να κάνουμε 10 ελέγχους, …. Αν k = 10, απαιτούνται 45 έλεγχοι! Όμως, αυτό το πρόβλημα θα μπορούσε να αντιμετωπισθεί εύκολα με χρήση κατάλληλου λογισμικού.

2) Το πραγματικό όμως πρόβλημα που δημιουργείται αφορά στην διόγκωση του σφάλματος

Π.χ. Σύγκριση 3 πληθυσμών/δειγμάτων (1, 2, 3)

t-tests: 1-2, 1-3, 2-3 με α=0.05 για το καθένα

Συνολική «εμπιστοσύνη» (πιθανότητα να μην έχει γίνει λάθος σε κανένα test) = $(0.95)^3$=0.857

Πιθανότητα να έχει γίνει λάθος σε ένα τουλάχιστο test = 1-0.857 = 0.143 > 0.05

Συμπέρασμα: Η πιθανότητα σφάλματος αυξάνεται πολύ αυξανομένου του αριθμού των ελέγχων (συγκρίσεων)

Π.χ. για 5 πληθυσμούς (10 συγκρίσεις) έχουμε πιθανότητα ενός τουλάχιστον σφάλματος

$1-(0.95)^{10}=0.40$ **(!!)**

Οι υποθέσεις της One-Way ANOVA (Απαιτήσεις αναφορικά με τα δεδομένα)

1. Η εξαρτημένη μεταβλητή είναι συνεχής μεταβλητή (μετράται σε μια κλίμακα διαστήματος ή αναλογίας) πχ. χρόνος απόκρισης σε ένα ερέθισμα, IQ score, ακαδημαϊκή επίδοση στη κλίμακα [0-100].

2. Η ανεξάρτητη μεταβλητή θα πρέπει να έχει τουλάχιστον τρείς κατηγορικές τιμές. Όταν η ανεξάρτητη έχει δύο τιμές τότε ο έλεγχος ANOVA συμπίπτει με ένα πχ. *t*-test

3. Οι παρατηρήσεις πρέπει να είναι ανεξάρτητες. Δηλαδή δεν πρέπει να υπάρχει σχέση μεταξύ των παρατηρήσεων από τον κάθε πληθυσμό ή μεταξύ των πληθυσμών. Για παράδειγμα τα μέλη του κάθε πληθυσμού θα πρέπει να είναι διαφορετικά (δεν μπορεί κάποιος να είναι μέλος σε δύο ή περισσότερους πληθυσμούς). Προφανώς η υπόθεση αυτή είναι θέμα σχεδιασμού της έρευνας και όχι κάτι που μπορεί να ελεγθεί στατιστικά. Αν η υπόθεση δεν ισχύει τότε κάποιο άλλο στατιστικό τεστ πρέπει να χρησιμοποιηθεί (πχ. για επαναλαμβανόμενες μετρήσεις) αντί της ANOVA.

4. Τυχαίο δείγμα δεδομένων από τον πληθυσμό

5. Δεν πρέπει να υπάρχουν ακραίες τιμές (outliers) στα δεδομένα αναφορικά προφανώς με την εξαρτημένη μεταβλητή.

6. Κανονικότητα (Normality) – Η εξαρτημένη μεταβλητή θα πρέπει να ακολουθεί την κανονική κατανομή για κάθε τιμή της ανεξάρτητης μεταβλητής.

7. Ομοιογένεια των διασπορών - Ισότητα διασπορών δηλαδή η διασπορά των δεδομένων στις διαφορετικές ομάδες θα πρέπει να είναι η ίδια

<u>Η ανάλυση διασποράς (ανάλυση της διακύμανσης) κατά έναν παράγοντα, ANOVA, στη πράξη</u>:

Η ανάλυση διασποράς (ανάλυση της διακύμανσης) κατά έναν παράγοντα, ANOVA, συγκρίνει τις μέσες τιμές μεταξύ περισσοτέρων των δύο ανεξάρτητων δειγμάτων προκειμένου να προσδιοριστεί εάν υπάρχει στατιστική απόδειξη ότι οι αντίστοιχοι πληθυσμοί (από τους οποίους προέρχονται τα δείγματα) έχουν διαφορετικές μέσες τιμές.

Η ανάλυση διασποράς κατά έναν παράγοντα είναι μια παραμετρική μέθοδος.

Οι μεταβλητές που χρησιμοποιούνται σε αυτή τη δοκιμασία είναι γνωστές ως:
- Εξαρτημένη μεταβλητή
- Ανεξάρτητη μεταβλητή (επίσης γνωστή ως μεταβλητή ομαδοποίησης ή παράγοντας)
Η ανεξάρτητη μεταβλητή διαιρεί τις στατιστικές μονάδες σε περισσότερους των δύο αμοιβαία αποκλειόμενους πληθυσμούς (επίπεδα ή ομάδες)

Χρήσεις της ANOVA

Η ANOVA χρησιμοποιείται συχνά για την ανάλυση δεδομένων από τους ακόλουθους τύπους μελετών:

Μελέτες πεδίου- δειγματοληπτικές έρευνες

Πειράματα

Οιονεί πειράματα (Quasi-experiments)

Η ANOVA χρησιμοποιείται συνήθως για να ελεγχθούν τα εξής:

Στατιστικές διαφορές μεταξύ των μέσων δύο ή περισσοτέρων πληθυσμών (ή ομάδων)

Στατιστικές διαφορές μεταξύ των μέσων για δύο ή περισσότερες παρεμβάσεις (πχ. θεραπείες, μεθόδους κά)

Στατιστικές διαφορές μεταξύ των μέσων δύο ή περισσότερων επιδόσεων, βαθμών κτλ.

Υπενθυμίζουμε ότι:

Τόσο η ANOVA όσο και ο έλεγχος ανεξάρτητων δειγμάτων (t-test) μπορούν να συγκρίνουν τις μέσες τιμές δύο ομάδες. Ωστόσο, μόνο η ANOVA μπορεί να συγκρίνει τα τις μέσες τιμές σε τρεις ή περισσότερες ομάδες.

Εάν η μεταβλητή ομαδοποίησης (ανεξάρτητη μεταβλητή) έχει μόνο δύο κατηγορίες (τιμές), τότε τα αποτελέσματα της ANOVA και του t-test θα είναι ισοδύναμα. Δηλαδή σε μια τέτοια περίπτωση αν εφαρμόσουμε και τις δύο μεθόδους τότε θα δούμε ότι $t^2 = F$.

Στη πράξη ακολουθούνται διάφοροι κανόνες κατά την εφαρμογή της ANOVA:

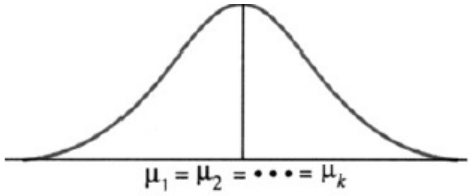
Κάθε δείγμα (ομάδα) πρέπει να έχει τουλάχιστον 6 στοιχεία (άτομα ή άλλες μονάδες)
Ιδανικά πρέπει να είναι περισσότερα Η συμπερασματολογία για τον πληθυσμό θα είναι πιο περιορισμένη όταν τα στοιχεία σε κάθε δείγμα είναι λίγα

Τα ισορροπημένα σχέδια (balanced designs) (δηλαδή ο ίδιος αριθμός ατόμων σε κάθε ομάδα) είναι ιδανικά. Τα ακραίως μη ισορροπημένα σχέδια αυξάνουν την πιθανότητα για παραβίαση των απαιτήσεων/υποθέσεων που με τη σειρά τους μειώνουν την εγκυρότητα της μεθόδου

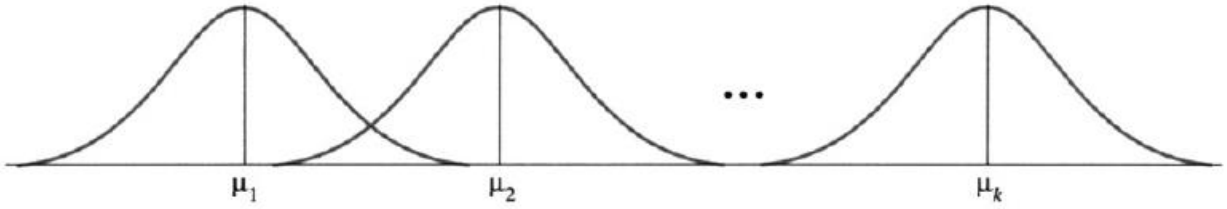
Ο έλεγχος ANOVA F-τεστ

Ας θεωρήσουμε ότι από καθέναν από k κανονικούς πληθυσμούς με κοινή διασπορά σ^2 και μέσες τιμές αντίστοιχα, μ_1 , μ_2 , … , μ_k , παίρνουμε ένα τυχαίο δείγμα μεγέθους αντίστοιχα, n_1 , n_2 , … , n_k για να κάνουμε, με βάση αυτά τα k δείγματα, τον έλεγχο της

Μηδενικής υπόθεσης **H_0:** $\mu_1 = \mu_2 = \ \ldots \ = \mu_k$



Εναλλακτική υπόθεση **H_1 :** $\mu_i \neq \mu_j$, *i, j* = 1,2,..,*k (τουλάχιστον ένα ζευγάρι διαφέρει)*



Η στατιστική του ελέγχου στην ANOVA συμβολίζεται με F και για τον υπολογισμό της τιμής της συνήθως χρησιμοποιείται ο παρακάτω πίνακας

Παράδειγμα: (Healey, J. F. (2012) Statistics: A tool for social research, Ninth edition, Wadsworth, σελ. 251)

Σε ένα μεγάλο πανεπιστήμιο οργανώθηκε ένα πείραμα με σκοπό διερευνηθεί επίδραση της μεθόδου διδασκαλίας σε ένα εισαγωγικό μάθημα στατιστικής στην επίδοση των φοιτητών. Οι φοιτητές ομαδοποιήθηκαν με τυχαίο τρόπο σε τρείς ομάδες και διδάχτηκαν με τρείς διαφορετικές μεθόδους Α, Β και Γ αντιστοίχως. Στο τέλος του εξαμήνου επιλέχθηκαν τυχαία δείγματα μεγέθους 9 από τις τελικές επιδόσεις κάθε τμήματος. Οι επιδόσεις μετρήθηκαν στην κλίμακα [0-100] και τα σχετικά δεδομένα δίνονται στον παρακάτω πίνακα. Το ερώτημα είναι αν υπάρχει σημαντική διαφορά στην επίδοση των φοιτητών εξαρτώμενη από τη μέθοδο διδασκαλίας;

Προφανώς πρέπει να ελέγξουμε την υπόθεση

H_0: $\mu_A = \mu_B = \mu_\Gamma$

Όπου μ_A , μ_B και μ_Γ *συμβολίζουν τις μέσες επιδόσεις στις ομάδες (πληθυσμούς) των φοιτητών που διδάσκονται με τις μεθόδους Α, Β και Γ αντιστοίχως*

Πίνακας 1: Τελικές επιδόσεις ανάλογα με τη μέθοδο διδασκαλίας

A		B		Γ	
x_i	x_i^2	x_i	x_i^2	x_i	x_i^2
55	3025	56	3136	50	2500
57	3249	60	3600	52	2704
60	3600	62	3844	60	3600
63	3969	67	4489	61	3721
72	5184	70	4900	63	3969
73	5329	71	5041	69	4761
79	6241	82	6724	71	5041
85	7225	88	7744	80	6400
92	8464	95	9025	82	6724
$\sum_{i=1}^{9} x_i =$	636		651		588
$\sum_{i=1}^{9} x_i^2 =$	46,286		48,503		39,420
$\bar{X}_A =$	70.67	$\bar{X}_B =$	72.33	$\bar{X}_\Gamma =$	65.33

$\bar{X}$= 1875/27 = 69.44

Μεταβλητότητα μεταξύ και μέσα στις ομάδες

Ερώτημα: Γιατί η ANOVA (ανάλυση της διασποράς) η οποία χρησιμοποιείται για τη σύγκριση των μέσων τιμών των πληθυσμών ονομάζεται ανάλυση διασποράς;

Διότι (όπως χαρακτηριστικά αναφέρουν οι Agresti & Finlay (2009, σελ. 371) στην ANOVA η στατιστική του ελέγχου συγκρίνει τις μέσες τιμές χρησιμοποιώντας δύο εκτιμήσεις για τη διασπορά, σ^2 , για κάθε ομάδα.

- **Η μία εκτίμηση αφορά στην μεταβλητότητα *μεταξύ* της μέσης τιμής κάθε δείγματος $\bar{X}_i$, $i = 1,2, \ldots, k$ και της μέσης τιμής του συνολικού δείγματος $\bar{X}$, γνωστή ως εκτίμηση της διασποράς μεταξύ ομάδων (between-groups estimate of variance) ή απλά μεταβλητότητα μεταξύ ομάδων (δειγμάτων)**

Για την περίπτωση του παραδείγματος η εκτίμηση αφορά στη μεταβλητότητα των $\bar{X}_A = 70.67$, $\bar{X}_B = 72.33$ και $\bar{X}_\Gamma =$ 65.33 από την μέση τιμή του συνολικού δείγματος $\bar{X}$= 1875/27 = 69.44

- **Η άλλη εκτίμηση αφορά στην μεταβλητότητα *μέσα σε κάθε ομάδα (δείγμα)* δηλαδή στην μεταβλητότητα των επιμέρους παρατηρήσεων, $x_{i\,,}\ i = 1,2, \ldots, n_k$ κάθε δείγματος από την μέση τιμή του αντίστοιχου δείγματος της μέσης τιμής κάθε δείγματος $\bar{X}_i$ $i = 1,2, \ldots, k$ γνωστή ως εκτίμηση της διασποράς μέσα στις ομάδες (Within-groups estimate of variance) ή απλά μεταβλητότητα εντός των ομάδων (δειγμάτων)**

Η στατιστική του ελέγχου ***F*** είναι ο λόγος των δύο εκτιμήσεων της διασποράς.

Για τον έλεγχο της μηδενικής υπόθεσης, **H_0:** $\mu_1 = \mu_2 = \ \ldots \ = \mu_k$, γενικά

Για το παράδειγμά μας της **H_0:** $\mu_A = \mu_B = \ \mu_\Gamma$

Η στατιστική του ελέγχου ***F*** είναι ο λόγος των δύο εκτιμήσεων της διασποράς δηλαδή, **εκτίμηση της διασποράς μεταξύ ομάδων και εκτίμηση της διασποράς μέσα στις ομάδες**

Η στατιστική του ελέγχου έχει τη μορφή

$$F = \frac{\textbf{εκτίμηση της διασποράς μεταξύ ομάδων}}{\textbf{εκτίμηση της διασποράς μέσα στις ομάδες}}$$

Και ονομάζεται η στατιστική F ανάλυσης της διασποράς ή συντομότερα, η στατιστική F της ANOVA.

Η εκτίμηση της διασποράς εντός των ομάδων είναι μια αμερόληπτη εκτίμηση της $\boldsymbol{\sigma^2}$ ανεξάρτητα από το εάν η H_0 είναι αληθινή.

Αντίθετα, η εκτίμηση της διασποράς μεταξύ των ομάδων είναι αμερόληπτη μόνο αν η H_0 είναι αληθινή και τότε έχει περίπου την ίδια τιμή με εκείνη της διασποράς εντός των ομάδων και συνεπώς η τιμή της *F* αναμένεται κοντά στο 1.0, πέραν του σφάλματος δειγματοληψίας.

Εάν η H_0 δεν είναι αληθινή, τότε η διασπορά μεταξύ των ομάδων τείνει να υπερεκτιμά την $\boldsymbol{\sigma^2}$ συνεπώς τείνει να είναι μεγαλύτερη από την εκτίμηση εντός των ομάδων και η τιμή της *F* τείνει να είναι μεγαλύτερη του 1.0, (και ακόμη μεγαλύτερη με μεγαλύτερα δείγματα).

Όταν η H_0 είναι αληθινή, η παραπάνω *F* στατιστική ακολουθεί την (δειγματική) κατανομή *F*. Η *p*-value βρίσκεται στην δεξιά ουρά της κατανομής. Όσο μεγαλύτερη είναι η τιμή της στατιστικής *F*, τόσο μικρότερη είναι η τιμή *p*-value.

Συνέχεια παραδείγματος σύγκριση μεθόδων διδασκαλίας

Τα αποτελέσματα που προκύπτουν από την εφαρμογή της ANOVA σε δεδομένα με τη χρήση κάποιου σχετικού λογισμικού συνήθως παρουσιάζονται σε έναν πίνακα γνωστός ως **Πίνακας ANOVA.**

Ο οποίος αυτός παρουσιάζει τις δύο εκτιμήσεις της διασποράς του πληθυσμού $\boldsymbol{\sigma^2}$ δηλαδή της διασποράς μεταξύ δειγμάτων και εντός δειγμάτων κάτω από τον τίτλο "mean square" και την τιμή της *F* στατιστικής που είναι ο λόγος των δύο διασπορών

Πίνακας 2. Για το παράδειγμά μας ο πίνακας ANOVA έχει την παρακάτω μορφή

	Sum of Squares	df	Mean Square	F
Μεταξύ ομάδων	SSB (= 240.82)	dfb (=2)	MSB (120.41)	MSB/MSW=
Εντός ομάδων (Error)	SSW (= 3776.51)	dfw (=24)	MSW (157.36)	120.41/157.36= 0.77
Total	SST (=4017.33)	dfT		

Σημείωση: ο παρακάτω συμβολισμός χρησιμοποιείται επίσης σε αποτελέσματα από σχετικό λογισμικό

SSB ή SSR (the regression sum of squares)

SSW ή SSE (the error sum of squares)

MSB ή MSR

MSW ή MSE

Υπολογισμός της διασποράς μεταξύ ομάδων και εντός ομάδων

ΒΑΣΙΚΗ ΠΑΡΑΤΗΡΗΣΗ:
ΤΟΣΟ Η ΕΚΤΙΜΗΣΗ ΤΗΣ ΔΙΑΣΠΟΡΑΣ ΜΕΤΑΞΥ ΟΜΑΔΩΝ ΟΣΟ ΕΝΤΟΣ ΟΜΑΔΩΝ ΑΠΟΤΕΛΟΥΝ ΕΝΑ ΜΕΤΡΟ ΜΕΤΑΒΛΗΤΟΤΗΤΑΣ ΔΙΑΙΡΟΥΜΕΝΟ ΜΕ ΕΝΑΝ ΟΡΟ ΠΟΥ ΑΝΤΙΣΤΟΙΧΕΙ ΣΕ ΒΑΘΜΟΥΣ ΕΛΕΥΘΕΡΙΑΣ

Χρήσιμος συμβολισμός – ορολογία

Η ποσότητα $SST = \sum_{i=1}^{n} x_i^2 = n\bar{X}^2$
ονομάζεται συνολικό άθροισμα τετραγώνων (total sum of squares)
και εύκολα μπορεί να δειχθεί ότι ισούται με το άθροισμα των τετραγώνων των αποκλίσεων των δεδομένων (συνολικό δείγμα) από την μέση τους τιμή δηλαδή ισούται με $\sum_{i=1}^{n}(x_i - \bar{X})^2$ **που αποτελεί τον αριθμητή στη σχέση υπολογισμού της διαποράς** n **δεδομένων.**

Για τα παράδειγμά μας και με βάση τα στοιχεία του πίνακα 1 έχουμε
$SST = \sum_{i=1}^{n} x_i^2 - n\bar{X}^2 = \sum_{i=1}^{27} x_i^2 - n\bar{X}^2 = (46{,}286 + 48{,}503 + 39{,}420) - 27(69.44)^2 = 4017.33$, που παρουσιάζεται και στον Πίνακα 2

Για να κατασκευάσουμε τις δύο επιμέρους εκτιμήσεις της διασποράς η συνολική διασπορά $\boldsymbol{SST}$ επιμερίζεται σε δύο μέρη που αντιστοιχούν στη διασπορά εντός των δειγμάτων που συμβολίζεται με $\boldsymbol{SSW}$, **sum of squares within** και στη διασπορά μεταξύ των δειγμάτων που συμβολίζεται με $\boldsymbol{SSB}$, **sum of squares between,** και ισχύει

$$\boldsymbol{SST} = \boldsymbol{SSB} + \boldsymbol{SSW}$$

Για τον υπολογισμό του $\boldsymbol{SSB}$ έχουμε

$$\boldsymbol{SSB} = \sum_{j=1}^{k} \boldsymbol{n_i}(\overline{\boldsymbol{X}}_j - \overline{\boldsymbol{X}})^2$$

όπου $j = 1, 2, \ldots, k$ ο αριθμός των δειγμάτων από αντίστοιχους πληθυσμούς με αντίστοιχα μεγέθη $\boldsymbol{n_i}$ και αριθμητικούς μέσους $\overline{\boldsymbol{X}}_j$ και $\overline{\boldsymbol{X}}$ ο αριθμητικός μέσος του συνολικού δείγματος, όπως προηγουμένως.

Για το παράδειγμά μας έχουμε (και με βάση τα στοιχεία από Πίνακα 1)
$\boldsymbol{SSB} = \sum_{j=1}^{3} \boldsymbol{n_i}(\overline{\boldsymbol{X}}_j - \overline{\boldsymbol{X}})^2 = (9)(70.67 \;\; 69.44)^2 + \;\; (9)(72.33 \;\; 69.44)^2 + \;\; (9)(65.33 \;\; 69.44)^2 =$
$13.62 + 75.17 + 152.03 = 240.82$

Από τα παραπάνω έχουμε

$$\boldsymbol{SSW} = \boldsymbol{SST} - \boldsymbol{SSB}$$

Και για το παράδειγμά μας υπολογίζουμε
$\boldsymbol{SSW} = \boldsymbol{SST} - \boldsymbol{SSB} = 4017.33 - 240.82 = 3776.51$

Οι τιμές των $\boldsymbol{SSW}$ και $\boldsymbol{SSB}$ εμφανίζονται και στον Πίνακα 2

Παρατήρηση με τα μέχρις εδώ αποτελέσματα:

Εάν η μηδενική υπόθεση είναι αληθινή, τότε δεν θα πρέπει να υπάρχει μεγάλη διακύμανση από ομάδα σε ομάδα (για το παράδειγμα, μεταξύ μεθόδων), και οι τιμές των *SSW* και *SSB* θα πρέπει να είναι περίπου ίσες.

Αν η μηδενική υπόθεση δεν είναι αληθινή, θα υπάρχουν μεγάλες διαφορές μεταξύ των ομάδων σε σχέση με τις διαφορές εντός των ομάδων και η τιμή του *SSB* θα πρέπει να είναι πολύ μεγαλύτερη από εκείνη του *SSW*.

Η τιμή του *SSB* θα αυξάνεται καθώς η διαφορά μεταξύ των μέσων τιμών των ομάδων αυξάνεται και ειδικότερα όταν δεν υπάρχουν μεγάλες διακυμάνσεις εντός των ομάδων (*SSW*). Όσο μεγαλύτερη είναι η τιμή του *SSB* συγκριτικά με εκείνη του *SSW*, τόσο πιο πιθανό είναι να απορρίψουμε τη μηδενική υπόθεση.

Οι βαθμοί ελευθερίας, df

Το επόμενο βήμα για τον υπολογισμό των εκτιμήσεων της διασποράς του πληθυσμού είναι να διαιρέσουμε τα παραπάνω αθροίσματα, $\boldsymbol{SSW}$ και $\boldsymbol{SSB}$ με τους αντίστοιχους βαθμούς ελευθερίας.

Έστω:

dfw, οι βαθμοί ελευθερίας που σχετίζονται με το ***SSW***, και
dfb, οι βαθμοί ελευθερίας που σχετίζονται με το ***SSB***

Υπολογίζονται ως εξής:

$$dfw = n - k$$
$$dfb = k\text{-}1$$

Όπου n είναι το μέγεθος του συνολικού δείγματος και k ο αριθμός των δειγμάτων (από τους αντίστοιχους πληθυσμούς ή ομάδες).

Για το παράδειγμά μας έχουμε
$dfw = n - k = 27 - 3 = 24$
$dfb = k\text{-}1 = 3\text{-}1 = 2$
που παρουσιάζονται και στον Πίνακα 2.

Οι πραγματικές εκτιμήσεις της διασποράς του πληθυσμού, ονομάζονται μέσες τετραγωνικές εκτιμήσεις (**mean square estimates**) και χρησιμοποιούνται τα ακρωνύμια **MSW** και **MSB** για να δηλώσουν αντιστοίχως τις εκτιμήσεις της διασποράς για εντός και μεταξύ των δειγμάτων.

Οι εκτιμήσεις αυτές υπολογίζονται ως εξής:

$$\boldsymbol{MSW = \frac{SSW}{dfw}}$$

$$\boldsymbol{MSB = \frac{SSB}{dfb}}$$

Για το παράδειγμά μας έχουμε

$$\boldsymbol{MSW = \frac{SSW}{dfw} = \frac{3776.51}{24}} = 157.36$$

$$\boldsymbol{MSB = \frac{SSB}{dfb} = \frac{240.82}{2}} = 120.41$$

Παρουσιάζονται και στον Πίνακα 2

Τότε η τιμή της *F* στατιστικής του ελέγχου υπολογίζεται ως εξής:

$$\boldsymbol{F = \frac{MSB}{MSW}}$$

Και για το παράδειγμά μας έχουμε

$\boldsymbol{F = \frac{MSB}{MSW} = \frac{120.41}{157.36}}$ =0.77

Έλεγχος: η παραπάνω *F* στατιστική ακολουθεί την *F* κατανομή με
$dfw = n - k$ και $dfb = k$-1 βαθμούς ελευθερίας

Για το παράδειγμά μας έχουμε
$dfw = n - k$ = 27 - 3 = 24
$dfb = k$-1 = 3-1 = 2
Και η κρίσιμη *F* τιμή για α=0.05 είναι *F (κρίσιμη)* =3.30

Συγκρίνοντας την κρίσιμη τιμή και την τιμή της στατιστικής *F* που υπολογίστηκε παραπάνω καταλαβαίνουμε ότι δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση ότι οι μέσες τιμές των πληθυσμών είναι ίσες και συμπεραίνουμε ότι παρατηρούμενες διαφορές μεταξύ των μέσων τιμών δειγμάτων οφείλονται σε σφάλματα δειγματοληψίας. Επομένως το τελικό συμπέρασμα είναι ότι η επίδοση των φοιτητών στο μάθημα της στατιστικής δεν διαφέρει σημαντικά λαμβάνοντας υπόψη τη μέθοδο διδασκαλίας.

ΑΣΚΗΣΕΙΣ

1. Ένα τυχαίο δείγμα 15 εθνών έχουν επιλεγεί από τρία επίπεδα ανάπτυξης. Τα "Χαμηλού εισοδήματος" έθνη είναι σε μεγάλο βαθμό γεωργικά και έχουν το χαμηλότερο επίπεδο ποιότητας ζωής. Τα έθνη με "υψηλό εισόδημα" είναι βιομηχανικά, και οι άνθρωποι πιο εύποροι και σύγχρονοι. Τα έθνη με "Μεσαίο εισόδημα" είναι μεταξύ των προαναφερόμενων δύο άκρων. Το ερώτημα είναι αν αυτά τα χαρακτηριστικά αντικατοπτρίζονται επίσης στις διαφορές στην αναμενόμενη διάρκεια ζωής (αριθμός ετών που ο μέσος πολίτης αναμένεται να ζήσει) μεταξύ των ατόμων των τριών κατηγοριών εθνών; (Τα στοιχεία για τα 15 έθνη αφορούν το έτος 2009) . Ο παρακάτω πίνακας παρέχει τα σχετικά δεδομένα:

Χαμηλό εισόδημα		Μεσαίο εισόδημα		Υψηλό εισόδημα	
Έθνος	Αν. διάρ. ζωής	Έθνος	Αν. διάρ. ζωής	Έθνος	Αν. διάρ. ζωής
Cambodia	61	China	73	Australia	81
Malawi	46	Indonesia	71	Canada	81
Nepal	64	Pakistan	66	Japan	83
Niger	53	South Korea	80	Russia	68
Sudan	58	Turkey	72	United Kingdom	79

Source: Population Reference Bureau. 2009 World Population Data Sheet. Available at http://www.prb.org/. Επίσης αναφέρεται από τον Healey, J. F. (2012) Statistics: A tool for social research, Ninth edition, Wadsworth

1α) τι είδους στατιστικό πρόβλημα προκύπτει και πως διατυπώνεται ?
1β) τι υποθέσεις πρέπει να ισχύουν?
1γ) Ποιο το συμπέρασμά σας αναφορικά με το ερώτημα της άσκησης ? (τιμή της στατιστικής F=17.69, ορίστε εσείς επίπεδο σημαντικότητας

2. Τίθεται το ερώτημα: Είναι οι σεξουαλικά ενεργοί έφηβοι καλύτερα ενημερωμένοι για το AIDS και άλλα πιθανά προβλήματα υγείας που σχετίζονται με το σεξ από τους έφηβους που είναι σεξουαλικά αδρανείς;
Ένα τεστ γενικών γνώσεων για το σεξ και την υγεία δόθηκε σε τυχαία δείγματα εφήβων που αντιστοίχως ήταν σεξουαλικά αδρανείς, σεξουαλικά ενεργοί αλλά μόνο με έναν μόνο σύντροφο και σεξουαλικά ενεργοί με περισσότερους του ενός συντρόφους. Υπάρχει κάποια σημαντική διαφορά στις μεταξύ βαθμών ενημέρωσης στους αντιστοίχους τρεις πληθυσμούς από επιλέχτηκαν τα δείγματα;). Ο παρακάτω πίνακας παρέχει τα σχετικά δεδομένα:

σεξουαλικά αδρανείς	σεξουαλικά ενεργοί με έναν μόνο σύντροφο	σεξουαλικά ενεργοί με πολλούς συντρόφους
10	11	12
12	11	12
8	6	10
10	5	4
8	15	3
5	10	15

Source: Healey, J. F. (2012) Statistics: A tool for social research, Ninth edition, Wadsworth

1α) τι είδους στατιστικό πρόβλημα προκύπτει και πως διατυπώνεται ?
1β) τι υποθέσεις πρέπει να ισχύουν?
1γ) Ποιο το συμπέρασμά σας αναφορικά με το ερώτημα της άσκησης ? (τιμή της στατιστικής F=0.08, ορίστε εσείς επίπεδο σημαντικότητας