One Way ANOVA
(Analysis of Variance)

Ζιντζαράς Ηλίας, M.Sc., Ph.D.

Καθηγητής Βιομαθηματικών-Βιομετρίας
Εργαστήριο Βιομαθηματικών
Τμήμα Ιατρικής
Πανεπιστήμιο Θεσσαλίας

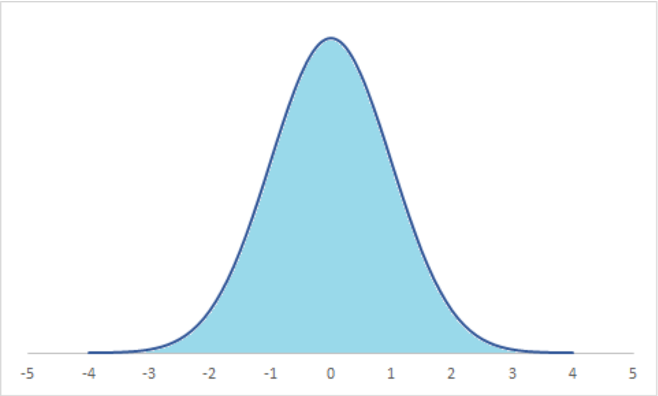
Institute for Clinical Research and Health Policy Studies
Tufts University School of Medicine
Boston, MA, USA

Θεόδωρος Μπρότσης, MSc, PhD Candidate
Ακαδημαϊκός Υπότροφος
(http://biomath.med.uth.gr)
Πανεπιστήμιο Θεσσαλίας
Email: tmprotsis@uth.gr

Γιατί ANOVA;

- Μέχρι τώρα συγκρίναμε δύο πληθυσμούς
  - T-Test για ανεξάρτητα δείγματα (T-Test for independent samples)
  - T-Test για εξαρτημένα δείγματα (Paired samples T-Test)
- Το να περιοριστούμε σε δύο πληθυσμούς αποτελεί φυσικά περιορισμό
- Τι γίνεται λοιπόν αν θελήσουμε να συγκρίνουμε τις μέσες τιμές περισσότερων των δύο πληθυσμών;
- Αν θελήσουμε να συγκρίνουμε πληθυσμούς με τον καθένα να αποτελείται από διάφορα επίπεδα ή υπο-ομάδες;
- Για το λόγο αυτό θα κάνουμε χρήση της ANOVA· ANalysis Of VAriance

Έστω πως θέλουμε να συγκρίνουμε τις μέσες τιμές ΤΡΙΩΝ δειγμάτων για να δούμε αν υπάρχουν διαφορές μεταξύ τους

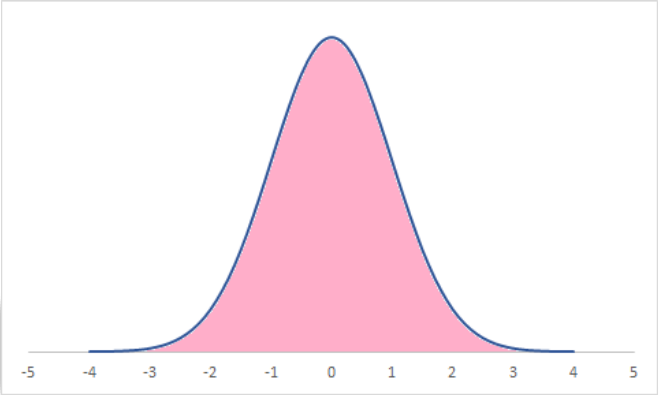


μ_1

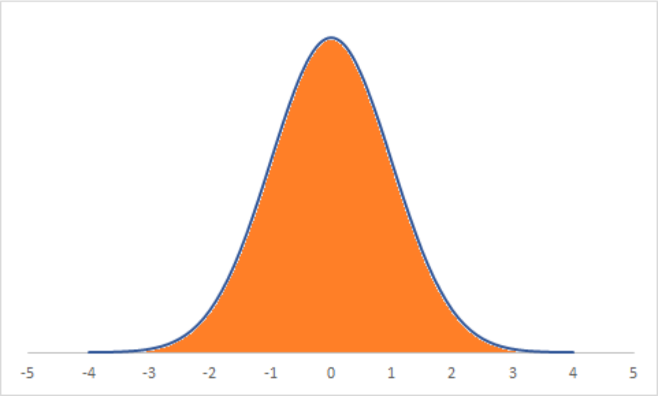
Ή είναι μία μέση τιμή τόσο μακριά από τις άλλες δύο ώστε να μην ανήκει στον ίδιο πληθυσμό;

Αυτό που ρωτάμε είναι:

Προέρχονται όλες αυτές οι μέσες τιμές από τον ίδιο πληθυσμό;

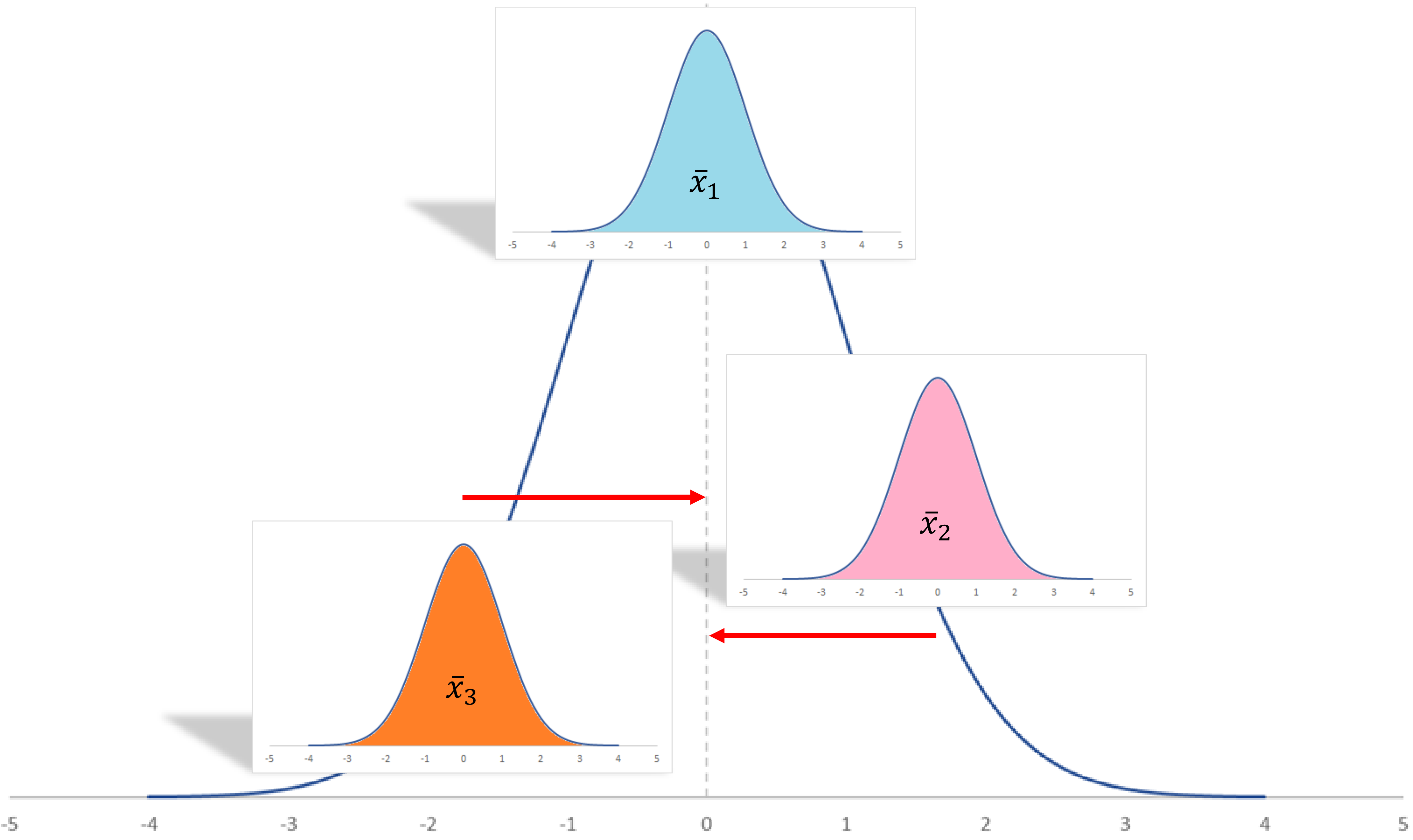


μ_2



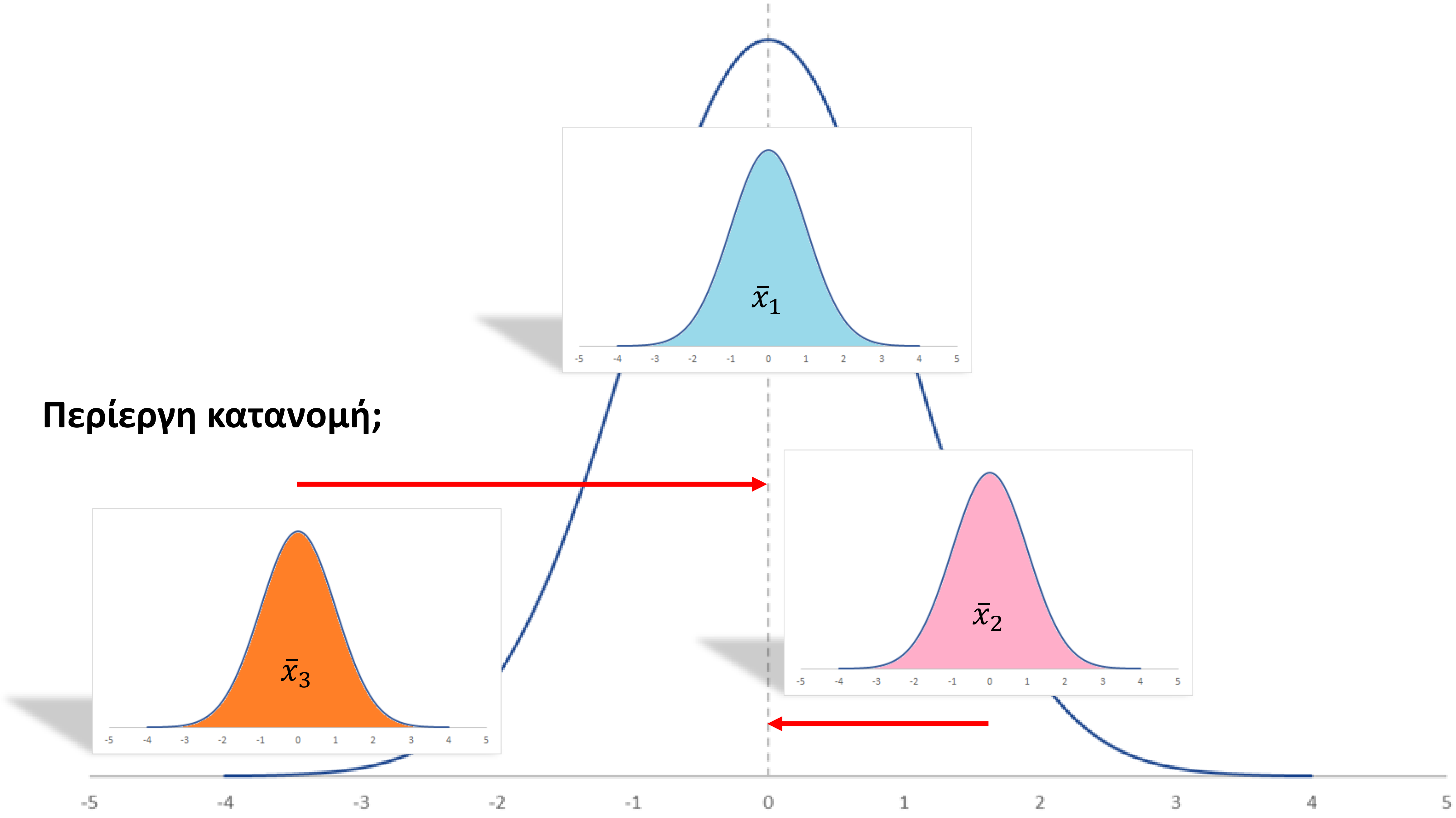
μ_3

Ή είναι όλες τόσο μακριά μεταξύ τους ώστε να μην προέρχονται από τον ίδιο πληθυσμό;

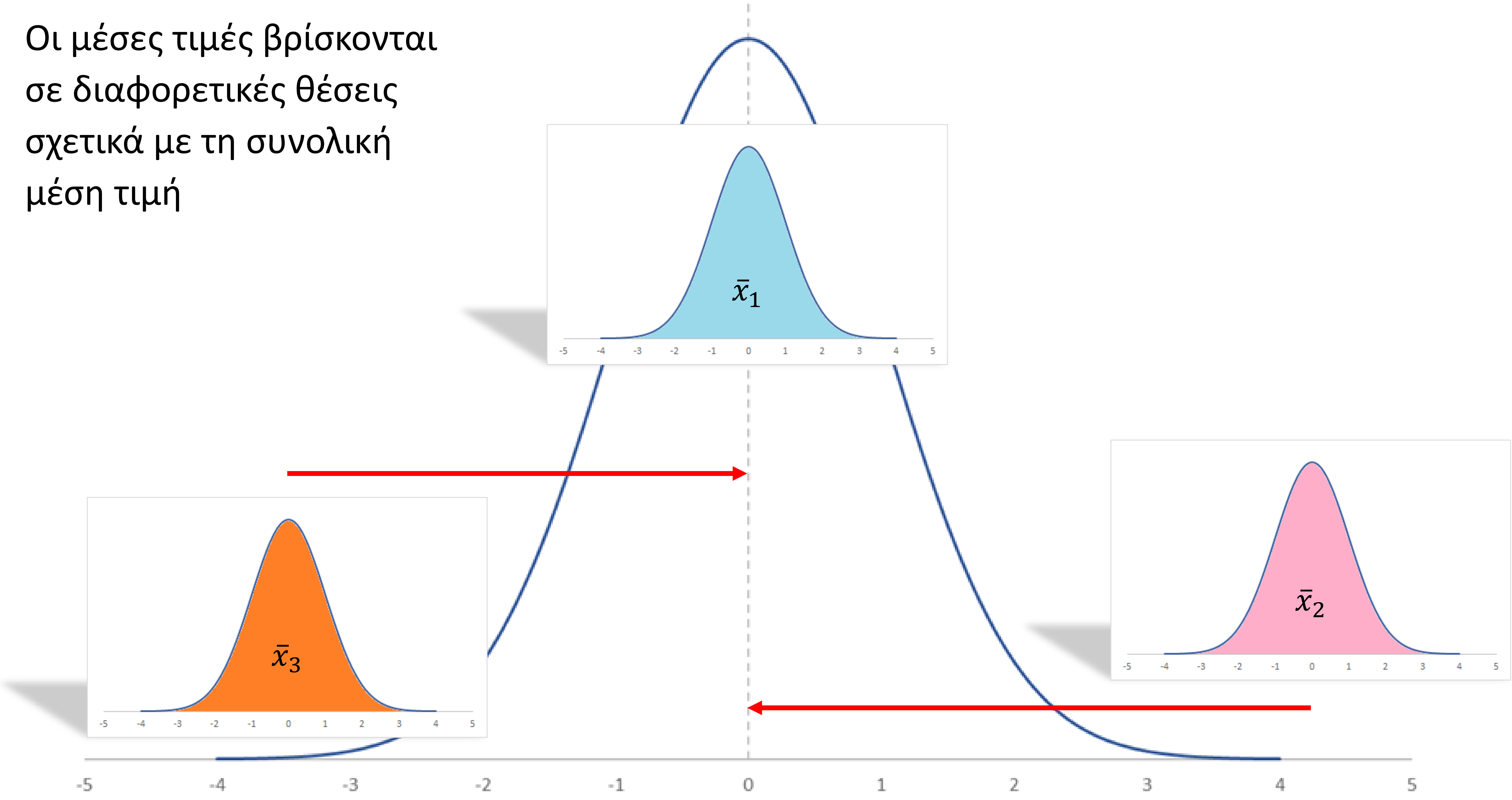


$\bar{x}_1$
$\bar{x}_2$
$\bar{x}_3$
-5 -4 -3 -2 -1 0 1 2 3 4 5

Περίεργη κατανομή;

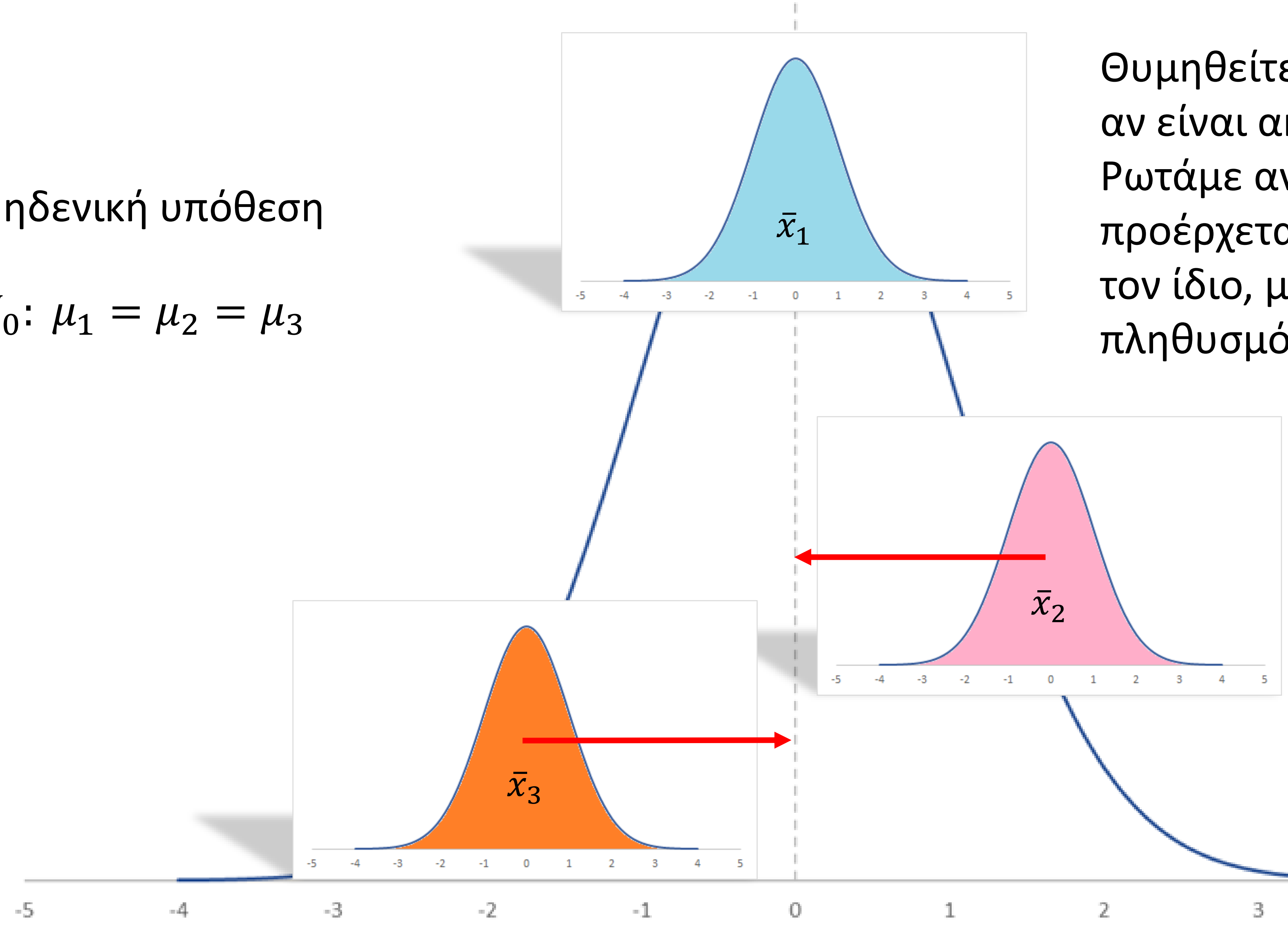


Οι μέσες τιμές βρίσκονται σε διαφορετικές θέσεις σχετικά με τη συνολική μέση τιμή



Μηδενική υπόθεση

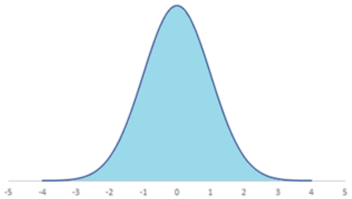
$$H_0\colon \mu_1 = \mu_2 = \mu_3$$



Θυμηθείτε, πως δεν ρωτάμε αν είναι ακριβώς ίδιες. Ρωτάμε αν η κάθε μέση τιμή προέρχεται πιθανότατα από τον ίδιο, μεγαλύτερο πληθυσμό.

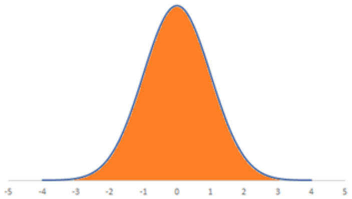
Variability AMONG / BETWEEN the sample means.

Πολλαπλά t-tests



$\bar{x}_1$

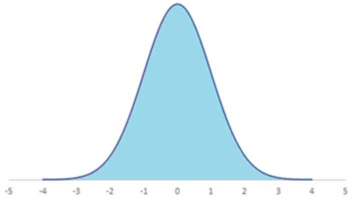
$\bar{x}_2$



$\bar{x}_3$

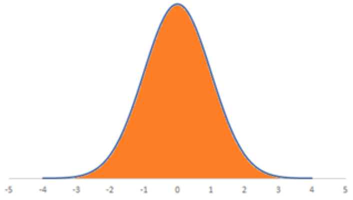
$H_0\colon\ \bar{x}_1 = \bar{x}_2; a = .05$

$H_0\colon\ \bar{x}_1 = \bar{x}_3; a = .05$



$\bar{x}_1$

$\bar{x}_2$



$\bar{x}_3$

$H_0\colon\ \bar{x}_2 = \bar{x}_3; a = .05$

Σύγκριση ανά ζεύγη σημαίνει τρία t-tests, ΟΛΑ με $\boldsymbol{a = .05}$ σφάλμα τύπου I με $\mathbf{95\%}$ εμπιστοσύνη

Το σφάλμα όμως συνδυάζεται με τους τρείς ελέγχους και έτσι γίνεται:

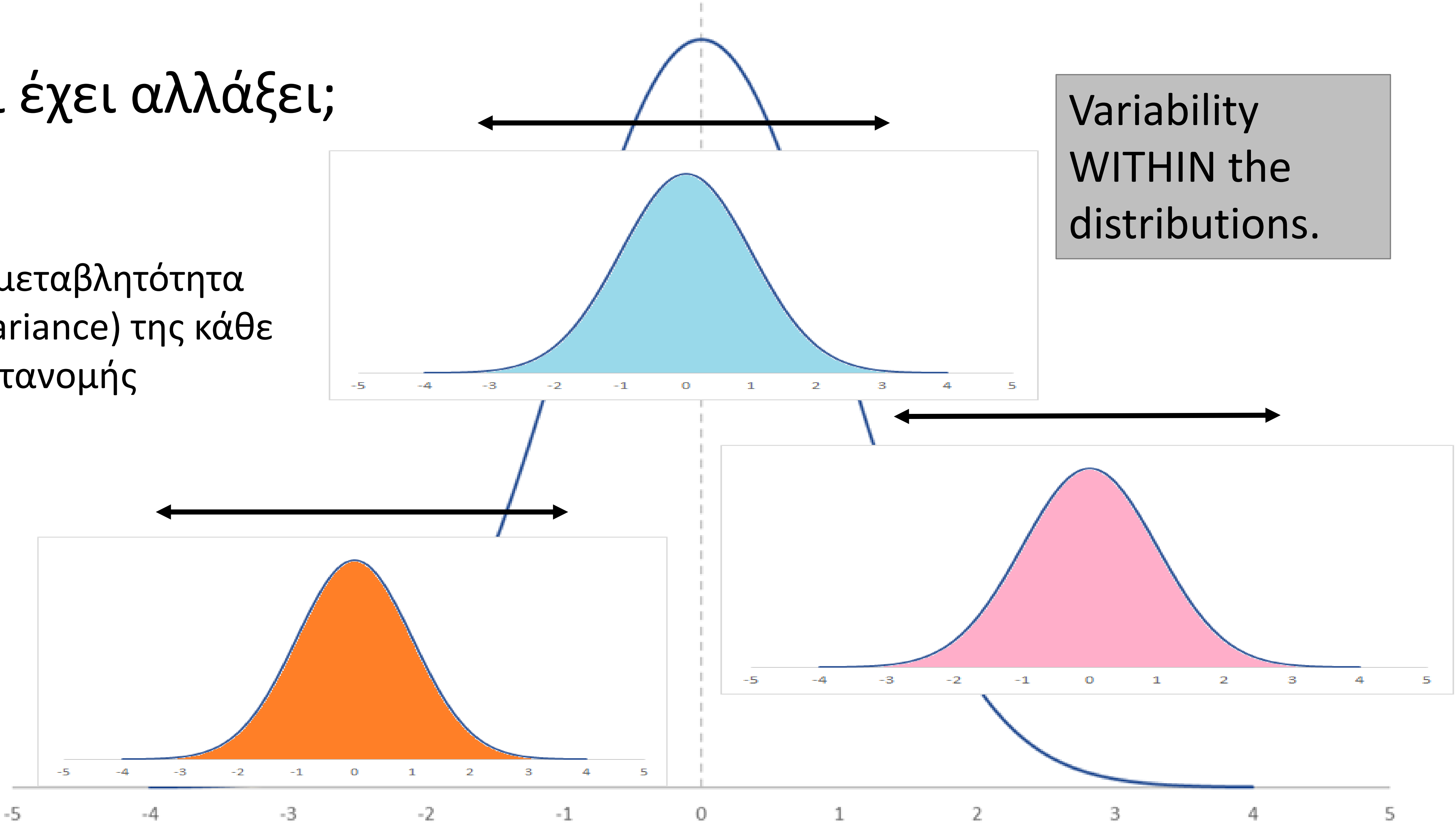
$$\mathbf{(1-.05)(1-.05)(1-.05) = 0.857}$$

$$\boldsymbol{a = 1 - .857 = .143}$$

Σφάλμα τύπου I: παρουσιάζεται όταν ο ερευνητής απορρίπτει εσφαλμένα τη μηδενική υπόθεση

Τι έχει αλλάξει;

Η μεταβλητότητα (variance) της κάθε κατανομής

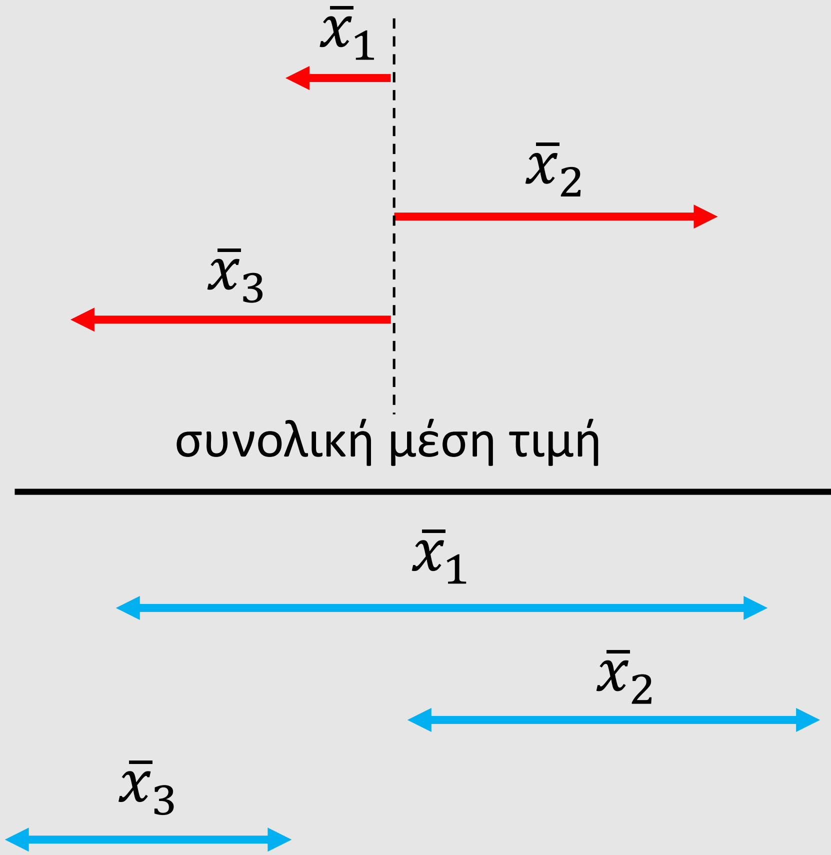


Variability WITHIN the distributions.

ANOVA: Analysis of Variance είναι μία αναλογία μεταβλητότητας

Variability AMONG / BETWEEN the sample means.

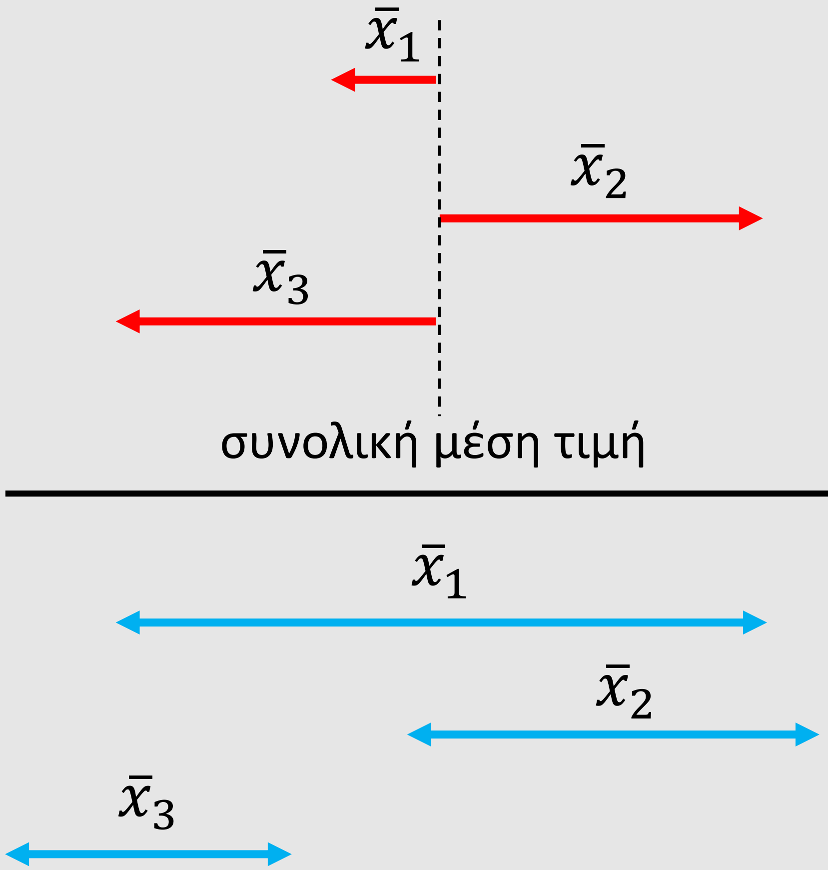
Variability WITHIN the distributions.



Απόσταση από τη συνολική μέση τιμή

“εντός έκταση”

ANOVA: Analysis of Variance είναι μία αναλογία μεταβλητότητας



Απόσταση από τη
συνολική μέση τιμή

“εντός έκταση”

$$= \frac{Variance\ Between}{Variance\ Within}$$

ANOVA: Analysis of Variance είναι μία αναλογία μεταβλητότητας

$$\left.\frac{Variance\ Between}{Variance\ Within}\right\}\ \text{Συστατικά συνολικής διακύμανσης}\ (Total\ Variance\ Components)$$

$$Variance\ Between + Variance\ Within = Total\ Variance$$

Διαμερισμός – διαχωρισμός συνολικής διακύμανσης στα συστατικά της

Αν η μεταβλητότητα ΜΕΤΑΞΥ (BETWEEN) των μέσων τιμών (απόσταση από τη συνολική μέση τιμή) στον αριθμητή είναι σχετικά μεγάλη σε σχέση με την μεταβλητότητα ΕΝΤΟΣ (WITHIN) των δειγμάτων (ενδογενή διακύμανση) στον παρονομαστή, η αναλογία είναι πολύ μεγαλύτερη από 1. Τότε, πιθανότητα, τα δείγματα ΝΑ ΜΗΝ προέρχονται από ένα κοινό πληθυσμό˙ ΑΠΟΡΡΙΨΗ ΤΗΣ ΜΗΔΕΝΙΚΗΣ ΥΠΟΘΕΣΗΣ ότι οι μέσες τιμές είναι ίδιες

ANOVA: Analysis of Variance είναι μία αναλογία μεταβλητότητας

$$\frac{\text{ΜΕΓΑΛΗ}}{\text{μικρή}} = \text{Απόρριψη της } H_0$$

Τουλάχιστον μία μέση τιμή είναι ακραία και κάθε κατανομή στενή· διαχωρίζονται μεταξύ τους

$$\frac{Variance\ Between}{Variance\ Within} \qquad \frac{\text{ΠΑΡΟΜΟΙΑ}}{\text{ΠΑΡΟΜΟΙΑ}} = \text{αποτυχία απόρριψης της } H_0$$

Οι μέσες τιμές είναι σχετικά κοντά στη συνολική μέση τιμή και/ή οι κατανομές επικαλύπτονται λίγο· Δύσκολα να διαχωριστούν

$$\frac{\text{μικρή}}{\text{ΜΕΓΑΛΗ}} = \text{αποτυχία απόρριψης της } H_0$$

Οι μέσες τιμές είναι σχετικά κοντά στη συνολική μέση τιμή και/ή οι κατανομές επικαλύπτονται αρκετά

ANOVA: Analysis of Variance είναι μία αναλογία μεταβλητότητας

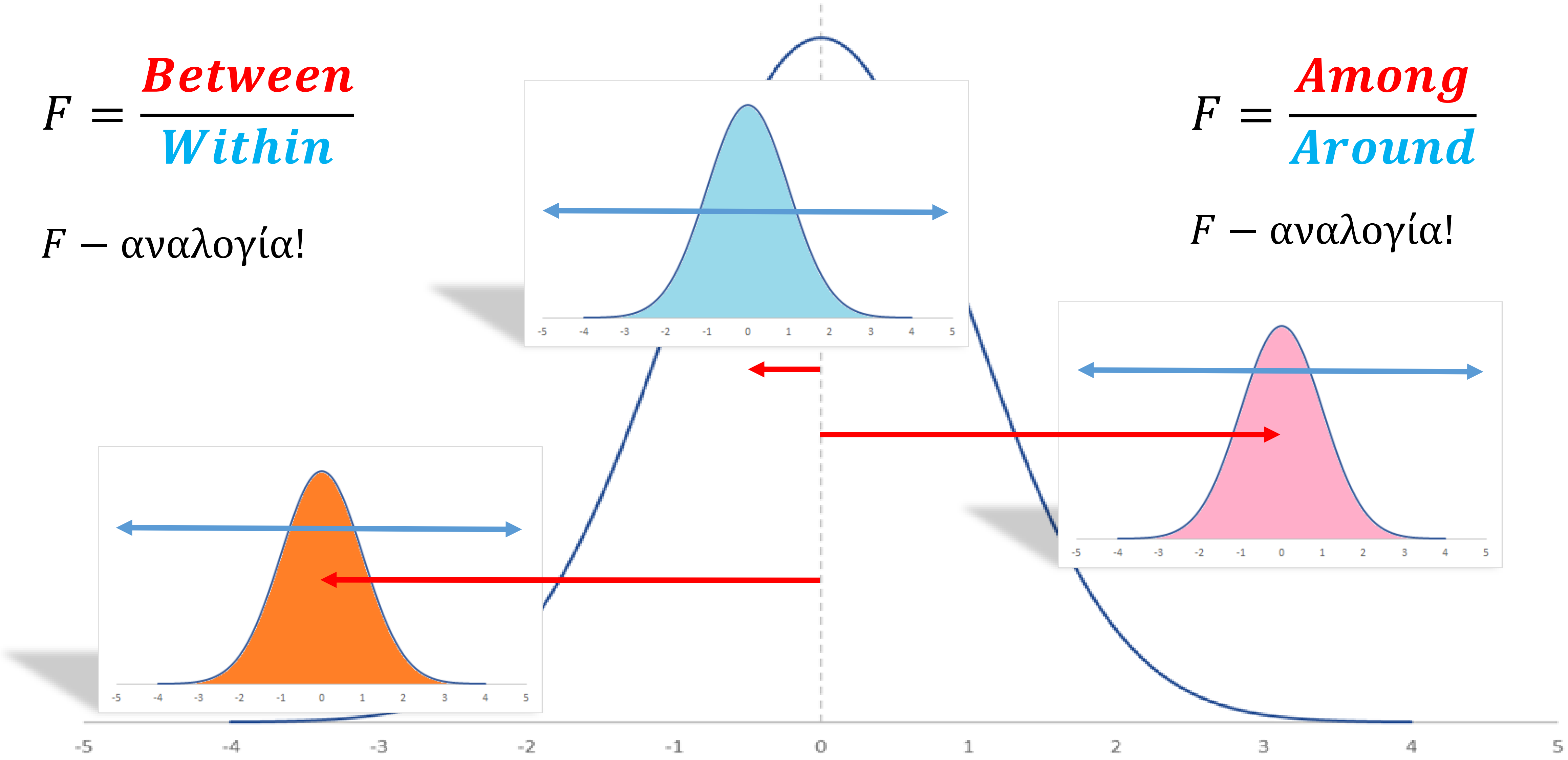
$$\textit{Variance Between} + \textit{Variance Within} = \textit{Total Variance}$$

$$F = \frac{\textbf{Between}}{\textbf{Within}}$$

$F -$ αναλογία!

$$F = \frac{\textbf{Among}}{\textbf{Around}}$$

$F -$ αναλογία!



Γιατί ANOVA;

- Μέχρι τώρα συγκρίναμε δύο πληθυσμούς
  - T-Test για ανεξάρτητα δείγματα (T-Test for independent samples)
  - T-Test για εξαρτημένα δείγματα (Paired samples T-Test)
- Το να περιοριστούμε σε δύο πληθυσμούς αποτελεί φυσικά περιορισμό
- Αν λοιπόν θελήσουμε να συγκρίνουμε τις μέσες τιμές περισσότερων των δύο πληθυσμών;
- Αν θελήσουμε να συγκρίνουμε πληθυσμούς με τον καθένα να αποτελείται από διάφορα επίπεδα ή υπο-ομάδες;
- Τότε θα κάναμε χρήση της ANOVA˙ ANalysis Of VAriance

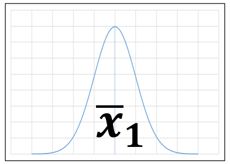
Σε μία μελέτη, έχει καταγραφεί το βάρος του ήπατος (εκφρασμένο ως ποσοστό του βάρους του σώματος) ποντικιών που ανήκουν σε k=4 ομάδες που τράφηκαν με 4 δίαιτες. Θέλουμε να ερευνήσουμε αν υπάρχουν συστηματικές διαφορές μεταξύ των 4 ομάδων.

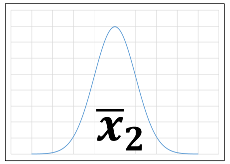
Ένας παράγοντας "δίαιτες"

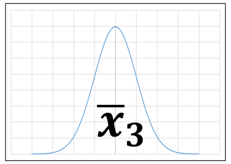
Columns / Groups

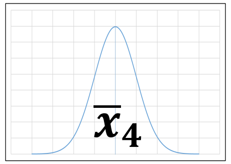
	a	b	c	d	
	3.42	3.17	3.34	3.64	
	3.96	3.63	3.72	3.93	
	3.87	3.38	3.81	3.77	
	4.19	3.47	3.66	4.18	
	3.58	3.39	3.55	4.21	
	3.76	3.41	3.51	3.88	
Μέση τιμή	**3.80**	**3.41**	**3.60**	**3.94**	**3.69 (total)**

Πηγές μεταβλητότητας









	a	b	c	d
	3.42	3.17	3.34	3.64
	3.96	3.63	3.72	3.93
	3.87	3.38	3.81	3.77
	4.19	3.47	3.66	4.18
	3.58	3.39	3.55	4.21
	3.76	3.41	3.51	3.88
Μέση τιμή	$\bar{x}_1 = 3.80$	$\bar{x}_2 = 3.41$	$\bar{x}_3 = 3.60$	$\bar{x}_4 = 3.94$

Συνολική μέση τιμή:

Η μέση τιμή όλων των βαρών του ήπατος

$$\bar{\bar{x}} = 3.69$$

- Ο έλεγχος για τις διαφορές μεταξύ των ομάδων βασίζεται στον εντοπισμό όλων των πηγών που διαμορφώνουν την μεταβλητότητα των δεδομένων (δηλ. τι κάνει τα 24 νούμερα να είναι διαφορετικά).
- Οπότε, η συνολική μεταβλητότητα (ή διακύμανση) αναλύεται στις επιμέρους πηγές διακύμανσης που την συνθέτουν (analysis of variance, ANOVA)

Πηγές μεταβλητότητας

	a	b	c	d	
	3.42	3.17	3.34	3.64	
	3.96	3.63	3.72	3.93	
	3.87	3.38	3.81	3.77	
	4.19	3.47	3.66	4.18	
	3.58	3.39	3.55	4.21	
	3.76	3.41	3.51	3.88	
Μέση τιμή	**3.80**	**3.41**	**3.60**	**3.94**	**3.69 (total)**

- Προφανώς, μία πηγή διακύμανσης είναι η επίδραση των **4 διαιτών**
- Μία άλλη πηγή διακύμανσης είναι η ενδογενή διακύμανση (within) μέσα στην κάθε ομάδα αφού το κάθε ποντίκι αντιδρά διαφορετικά στην ίδια δίαιτα˙ αυτή η μεταβλητότητα δεν μπορεί να ελεγχθεί και συνεπώς θεωρείται ως **τυχαίο σφάλμα** ή **τυχαία διακύμανση** (error)

Διακύμανση και άθροισμα τετραγώνων

- Δεδομένου ότι η ANOVA είναι εξ ορισμού η «ανάλυση της διακύμανσης», θα πρέπει να εξετάσουμε τη διακύμανση ως έννοια
- Ως διακύμανση ορίζουμε τη μέση τετραγωνική απόκλιση ενός σημείου από την μέση τιμή της κατανομής
  - Παίρνουμε την απόσταση κάθε σημείου από τη μέση τιμή, τετραγωνίζουμε, αθροίζουμε και βρίσκουμε τη μέση τιμή
- Αν βγάλουμε το «βρίσκουμε τη μέση τιμή» θα έχουμε το άθροισμα των τετραγώνων (sum of squares)
- Επομένως το sum of squares είναι η διακύμανση χωρίς την εύρεση της μέσης τιμής των τετραγωνικών αποκλίσεων

Διακύμανση και άθροισμα τετραγώνων

Διακύμανση δείγματος

$$s^2 = \frac{\sum(x-\mu)^2}{n-1}$$

$$s^2 = \frac{\sum(x-\mu)^2}{\cancel{n-1}}$$

Άθροισμα τετραγώνων
Sum of squares

$$SS = \sum(x-\mu)^2$$

Διαχωρισμός αθροίσματος τετραγώνων

SST
(total / overall)

SSC
(column / between / treatment)

SSE
(within / error)

Υπολογισμός του SST (sum of squares total)

	A	B	C	D	E
1		a	b	c	d
2		3.42	3.17	3.34	3.64
3		3.96	3.63	3.72	3.93
4		3.87	3.38	3.81	3.77
5		4.19	3.47	3.66	4.18
6		3.58	3.39	3.55	4.21
7		3.76	3.41	3.51	3.88
8					
9	**Μέση τιμή**	3.7966667	3.40833	3.59833	3.935

=AVERAGE(B2:B7)

Συνολική μέση τιμή:
Η μέση τιμή όλων των τιμών είναι $\bar{\bar{x}} = 3.69$

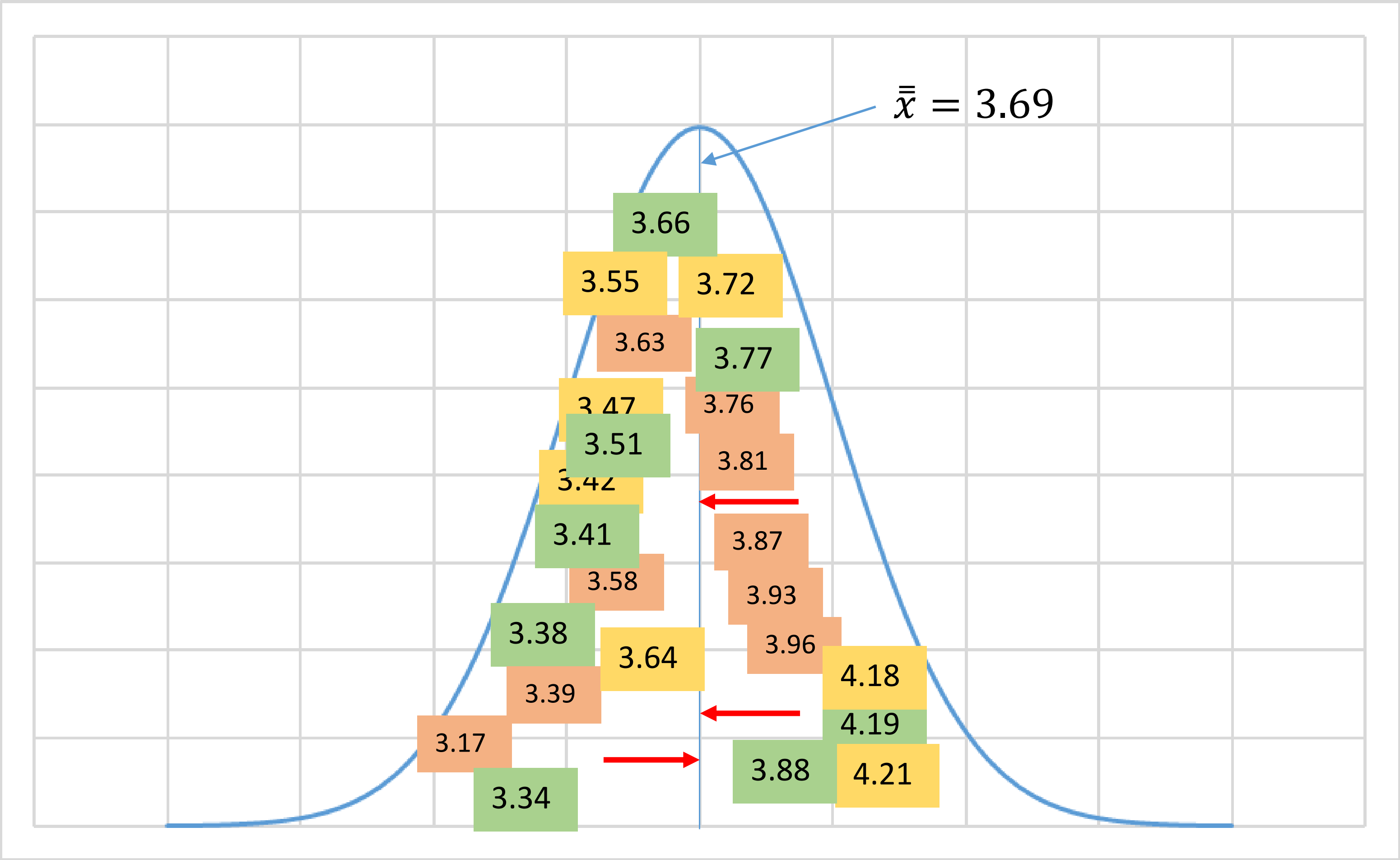
Υπολογισμός του συνολικού SST

1. Βρίσκουμε τις διαφορές ανάμεσα σε κάθε τιμή και την συνολική μέση τιμή
2. Υψώνουμε στη δύναμη του 2 τις διαφορές
3. Αθροίζουμε

=VAR.S(B2:E7)*23

1.835196

SST
(total / overall)



Υπολογισμός του συνολικού SST

1. Βρίσκουμε τις διαφορές ανάμεσα σε κάθε τιμή και την συνολική μέση τιμή
2. Υψώνουμε στη δύναμη του 2 τις διαφορές
3. Αθροίζουμε

Η τετραγωνισμένη απόσταση είναι κυριολεκτικά ΤΕΤΡΑΓΩΝΟ

Υπολογισμός του Between groups SSC

	A	B	C	D	E
1		a	b	c	d
2		3.42	3.17	3.34	3.64
3		3.96	3.63	3.72	3.93
4		3.87	3.38	3.81	3.77
5		4.19	3.47	3.66	4.18
6		3.58	3.39	3.55	4.21
7		3.76	3.41	3.51	3.88
8					
9	**Μέση τιμή**	3.7966667	3.40833	3.59833	3.935

=AVERAGE(B2:B7)

Συνολική μέση τιμή:
Η μέση τιμή όλων των τιμών είναι $\bar{\bar{x}} = 3.69$

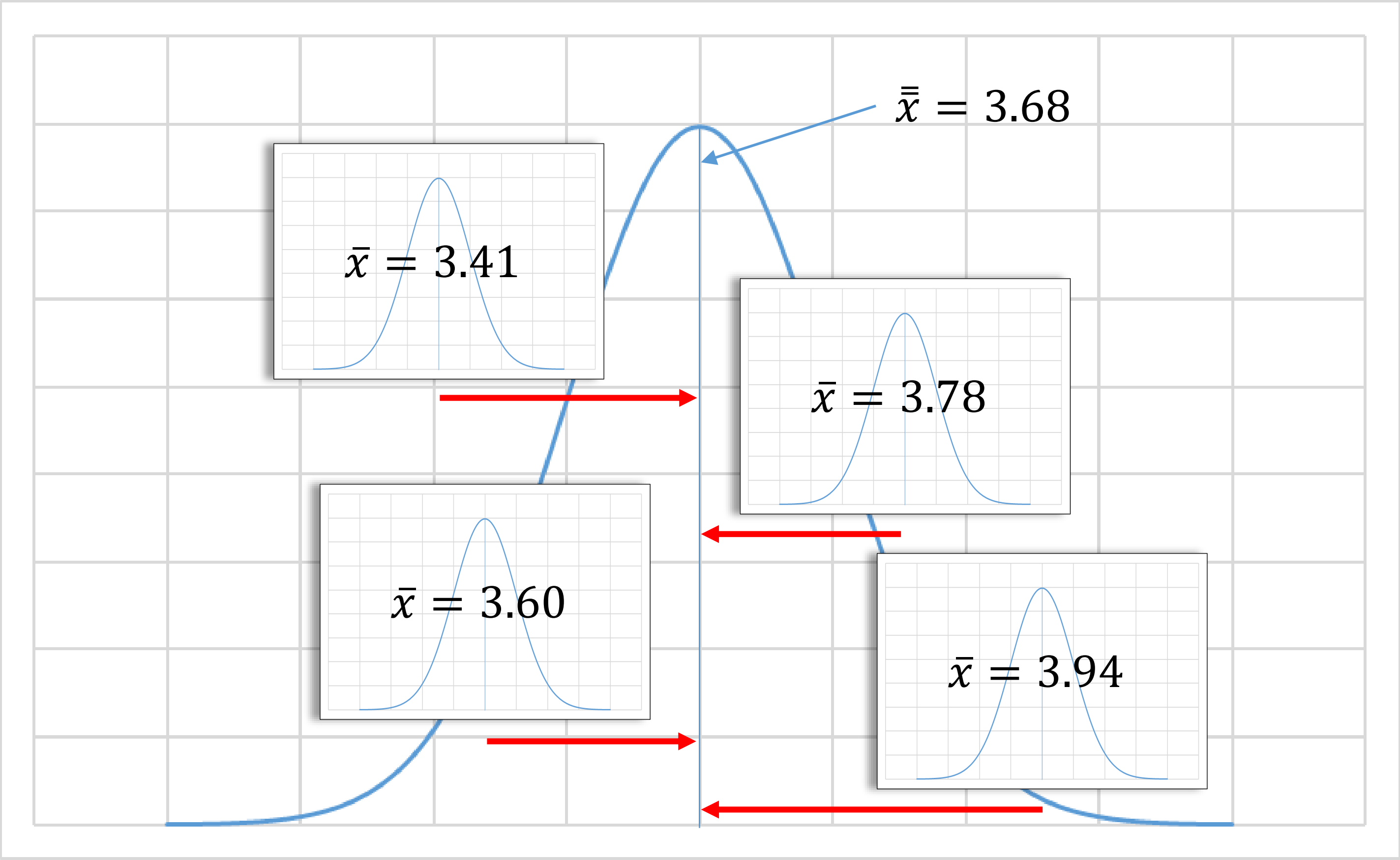
Υπολογισμός SSC between groups

1. Βρίσκουμε τις διαφορές ανάμεσα στη μέση τιμή των στηλών και την συνολική μέση τιμή
2. Υψώνουμε στη δύναμη του 2 τις διαφορές
3. Αθροίζουμε
4. Πολλαπλασιάζουμε με τον αριθμό των παρατηρήσεων της ομάδας

=VAR.S(B9:E9)*3*6

0.954145833

SSC
(column / between / treatment)
sum of squares



Υπολογισμός SSC
between groups

1. Βρίσκουμε τις διαφορές ανάμεσα στη μέση τιμή των στηλών και την συνολική μέση τιμή
2. Υψώνουμε στη δύναμη του 2 τις διαφορές
3. Αθροίζουμε
4. Πολλαπλασιάζου με με τον αριθμό των παρατηρήσεων της ομάδας

Υπολογισμός του error SSE

	A	B	C	D	E
1		a	b	c	d
2		3.42	3.17	3.34	3.64
3		3.96	3.63	3.72	3.93
4		3.87	3.38	3.81	3.77
5		4.19	3.47	3.66	4.18
6		3.58	3.39	3.55	4.21
7		3.76	3.41	3.51	3.88
8					
9	**Μέση τιμή**	3.7966667	3.40833	3.59833	3.935

=AVERAGE(B2:B7)

0.88105

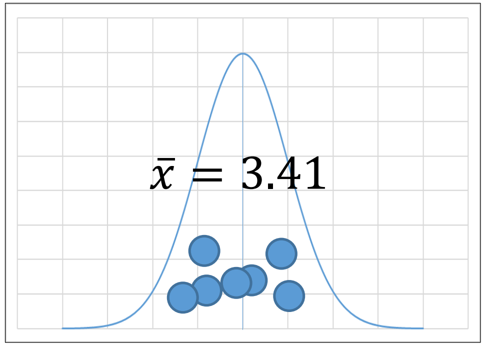
=VAR.S(B2:B7)*5+VAR.S(C2:C7)*5+VAR.S(D2:D7)*5+VAR.S(E2:E7)*5

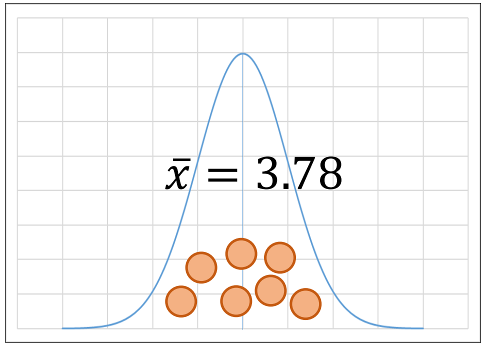
$$\cancel{\bar{\bar{x}} = 3.69}$$

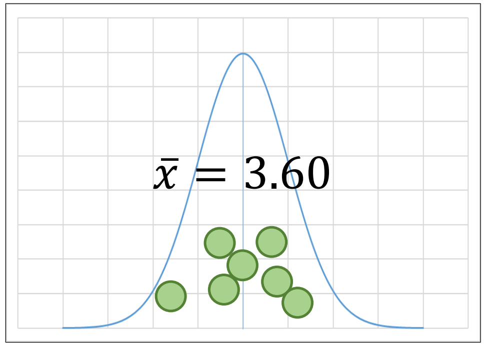
Υπολογισμός του Error SSE

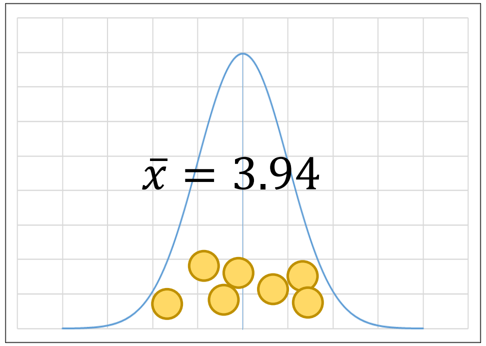
1. Βρίσκουμε τις διαφορές ανάμεσα σε κάθε τιμή και τη μέση τιμή της στήλης
2. Υψώνουμε στη δύναμη του 2 τις διαφορές
3. Αθροίζουμε
4. Προσθέτουμε το 3. για όλες τις στήλες

SSE
(within / error)
sum of squares









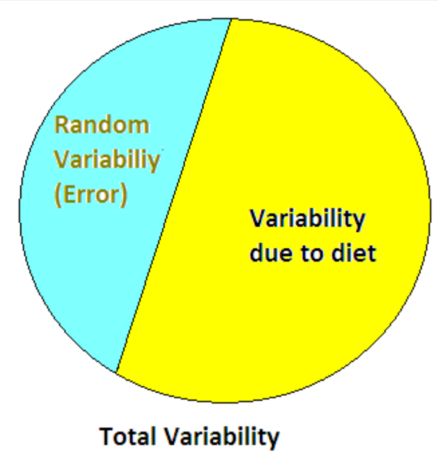
Υπολογισμός του Error SSE

1. Βρίσκουμε τις διαφορές ανάμεσα σε κάθε τιμή και τη μέση τιμή της στήλης
2. Υψώνουμε στη δύναμη του 2 τις διαφορές
3. Αθροίζουμε
4. Προσθέτουμε το 3. για όλες τις στήλες

One Way Anova

Η ANONA παρουσιάζεται με τη μορφή του παρακάτω πίνακα:

Source of variation (Πηγή μεταβλητότητας)	**df** (Βαθμοί ελευθερίας)	**SS** (Sum of squares)
Between groups (Μεταξύ των ομάδων)	**4-1=3 (C-1)**	**0.954**
Within groups (Εντός των ομάδων) **(error/random)**	**24-4=20 (N-C)**	**0.881**
Total	**24-1=23 (N-1)**	**1.83**



ANOVA

weight

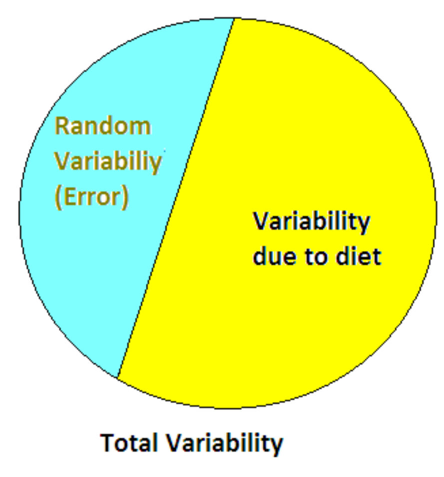
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.954	3	.318	7.220	.002
Within Groups	.881	20	.044		
Total	1.835	23			

N = συνολικές παρατηρήσεις, C = #columns/groups/treatments/diets

One Way Anova

Μετά υπολογίζεται, η **μέση** διακύμανση (Mean Square, MS), δηλ. MS=SS/df, της κάθε πηγής διακύμανσης

Source of variation	**df**	**SS**	**MS=SS/df** (μέσα τετράγωνα)
Between groups	**3**	**0.954**	**0.318**
Within groups (error/random)	**20**	**0.881**	**0.044=s^2**
Total	**23**	**1.83**	



ANOVA

weight

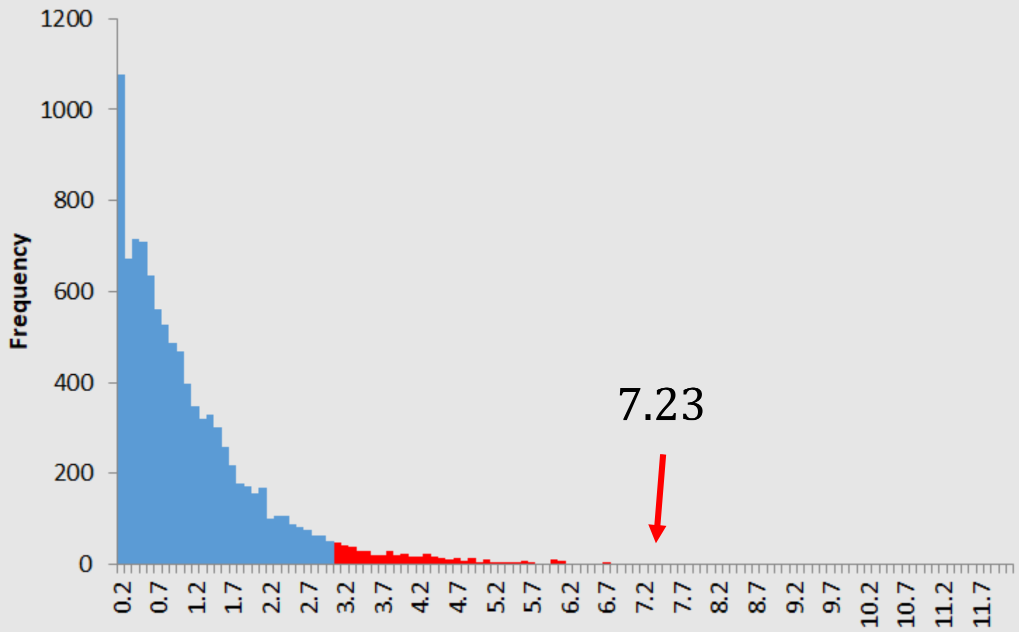
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.954	3	.318	7.220	.002
Within Groups	.881	20	.044		
Total	1.835	23			

$F - test$

- Κατόπιν, η μέση διακύμανση μεταξύ των 4 ομάδων συγκρίνεται με την μέση τυχαία διακύμανση (δηλ. με την ενδογενή διακύμανση της κάθε ομάδας) χρησιμοποιώντας το $\boldsymbol{F - Test}$:
- $F = \dfrac{\text{διακύμανση μεταξύ των ομάδων}}{\text{τυχαία διακύμανση}} = \dfrac{Between\ group\ MS}{Error\ MS} = \dfrac{0.318}{0.044} = 7.23$
- Αν η μέση διακύμανση μεταξύ των 4 ομάδων είναι μεγαλύτερη από την μέση τυχαία διακύμανση, δηλ. αν ο αριθμητής είναι μεγαλύτερος από τον παρονομαστή, τότε οι διαφορές μεταξύ των 4 ομάδων δεν είναι τυχαίες (είναι πραγματικές)
- Σε αυτή την περίπτωση, η τιμή του $F - test$ γίνεται πολύ μεγαλύτερη του 1

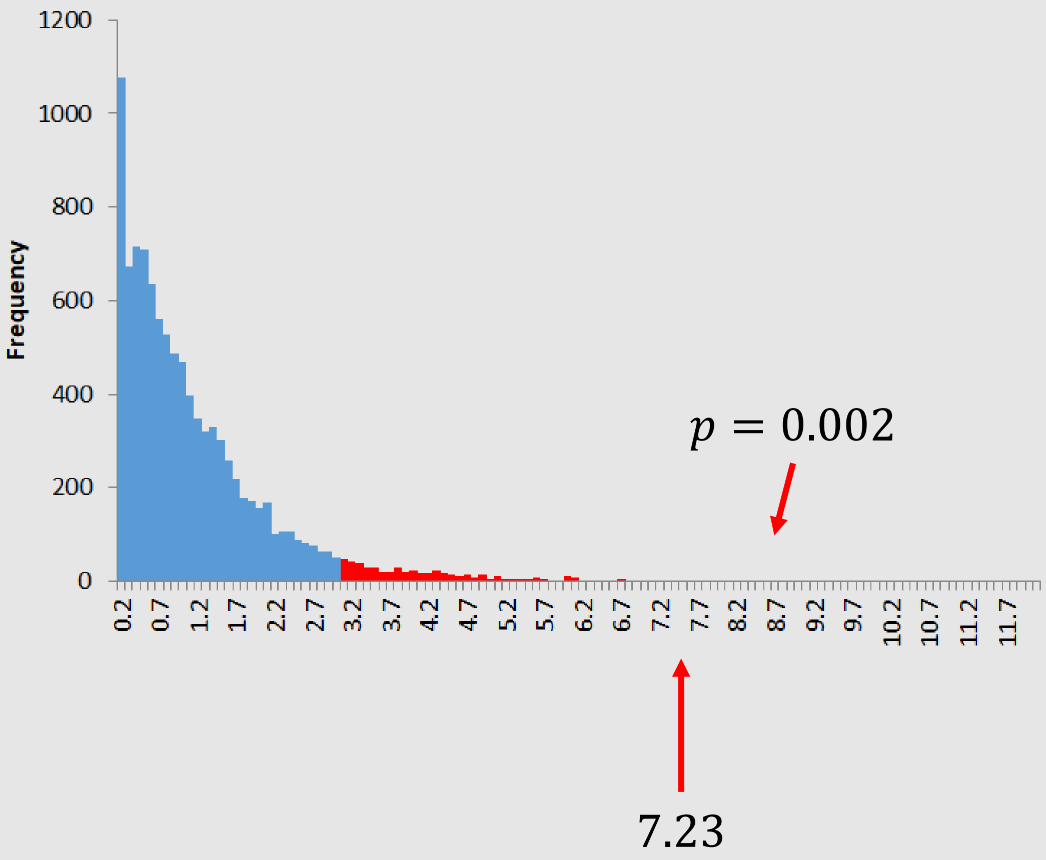
Σημαντικότητα

- Η σημαντικότητα της τιμής $F = 7.23$ προσδιορίζεται με παρόμοιο τρόπο με το $t - test$
- Προσομοιώνουμε τυχαία 10000 φορές την μελέτη και υπολογίζουμε κάθε φορά το F-test. Τότε τα 10000 $F - tests$ σχηματίζουν την $F -$ κατανομή
- Μετά βρίσκουμε το ποσοστό των $F - tests$ που είναι μεγαλύτερα από το $F = 7.23$



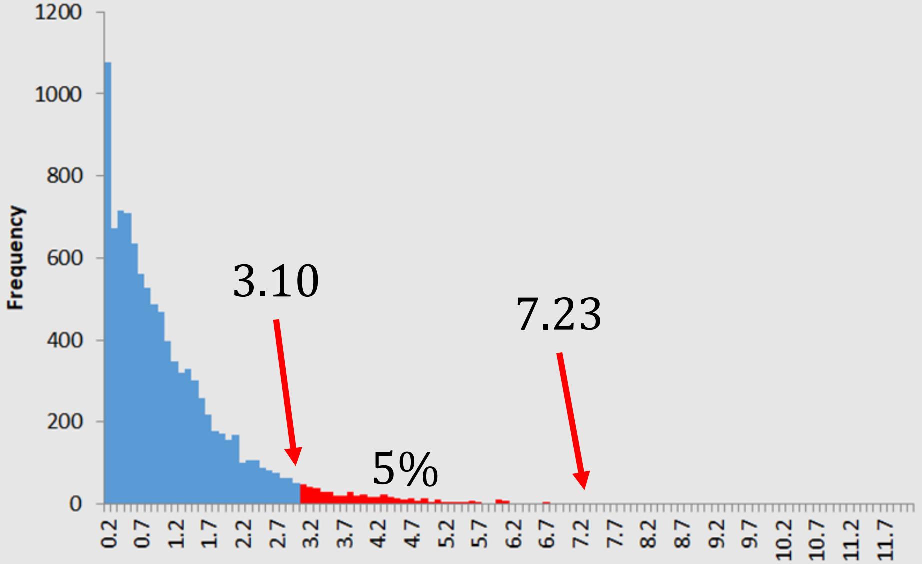
Σημαντικότητα

- Το ποσοστό των F-tests που είναι μεγαλύτερα από το $F = 7.23$ είναι $P = 0.002$
- Συνεπώς, οι 4 δίαιτες διαφέρουν μεταξύ τους σημαντικά (με μία μικρή πιθανότητα λάθους $P < 0.05$ ή $P = 0.002$)

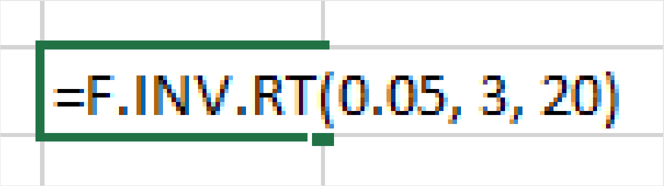


$F - test$

- Εναλλακτικά, η τιμή $F = 7.23$ συγκρίνεται με 5% σημείο της F-κατανομής με 3 και 20 df που είναι 3.1 (πίνακας F-κατανομής στην επόμενη διαφάνεια).
- Επειδή το $F = 7.23$ είναι μεγαλύτερο από το 3.1 συμπεραίνουμε ότι υπάρχει ένδειξη $(P < 0.05)$ ότι οι ομάδες (δίαιτες) διαφέρουν μεταξύ τους



Εύρεση του 5% σημείου της F κατανομής για 3 και 20 df στο Excel



Πίνακας f − κατανομής

Degrees of freedom in denominator	Degrees of freedom in numerator												
	1	2	3	4	5	6	7	8	9	10	20	40	∞
1	161.40	199.50	215.70	224.60	230.20	234.00	236.80	238.90	240.50	241.90	248.00	251.10	254.30
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.45	19.47	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.66	8.59	8.50
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.80	5.72	5.60
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.56	4.46	4.40
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.87	3.77	3.70
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.44	3.34	3.20
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.15	3.04	2.90
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	2.94	2.83	2.70
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.77	2.66	2.50
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.65	2.53	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.54	2.43	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.46	2.34	2.20
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.39	2.27	2.10
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.33	2.20	2.10
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.28	2.15	2.00
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.23	2.10	2.00
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.19	2.06	1.90
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.16	2.03	1.90
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.12	1.99	1.80
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	1.93	1.79	1.60
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.84	1.69	1.50
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.66	1.50	1.30
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.57	1.39	1.00

One Way ANOVA

Post Hoc Test

Διαδικασίες πολλαπλών συγκρίσεων

- Η ANOVA μας λέει μόνο αν οι μέσες τιμές των πληθυσμών είναι ίσες (αν πιθανότητα προέρχονται από τον ίδιο πληθυσμό)
- Αν το F-Test είναι σημαντικό, δεν γνωρίζουμε που ακριβώς βρίσκονται οι διαφορές
- Είναι απαραίτητο λοιπόν να κάνουμε επιμέρους συγκρίσεις μεταξύ των πληθυσμών:
  - A, B, C τα ζεύγη θα είναι AB, AC και BC ή $C(3,2) = \frac{3!}{(3-2)!2!} = \frac{1\times2\times3}{1\times(1\times2)} = \frac{6}{2} = 3$ ζεύγη
  - A, B, C, D τα ζεύγη θα είναι AB, AC, AD, BC, BD και CD ή $C(4,2) = 6$ ζεύγη
- Υπάρχουν πολλές διαδικασίες πολλαπλών ελέγχων: Fisher’s, LSD, Tukey HSD, Bonferroni

Διαδικασίες πολλαπλών συγκρίσεων

	Ομάδες / Δίαιτες			
	a	b	c	d
Μέση τιμή	3.80	3.41	3.60	3.94
Τυπική απόκλιση	0.27	0.15	0.17	0.22
Μέγεθος	6	6	6	6

ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	0.954145833	3	0.318048611	7.219763035	0.001805552	3.098391212
Within Groups	0.88105	20	0.0440525			
Total	1.835195833	23				

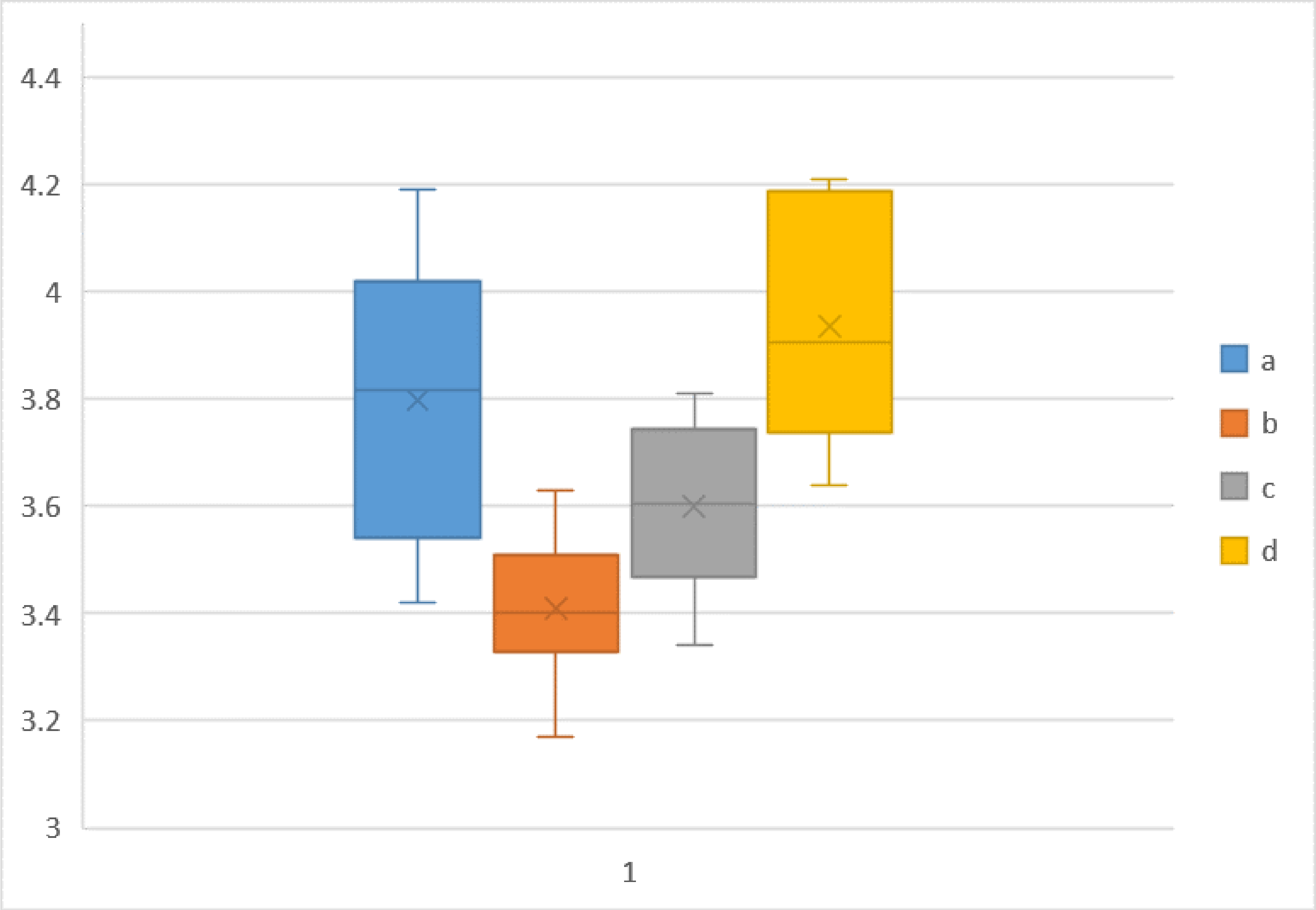
Πίνακας διαφορών

	d	a	c	b
d	0	0.14	0.34	0.53
a	-0.14	0	0.2	0.39
c	-0.34	-0.2	0	0.19
b	-0.53	-0.39	-0.19	0

Έξι έλεγχοι ζευγών

	Μέση τιμή #1	Μέση τιμή #2	Διαφορά
a vs b	3.80	3.41	0.39
a vs c	3.80	3.60	0.20
a vs d	3.80	3.94	-0.14
b vs c	3.41	3.60	-0.19
b vs d	3.41	3.94	-0.53
c vs d	3.60	3.94	-0.34

Ποια από αυτά τα ζεύγη συγκρίσεων περιέχει στατιστικά σημαντική διαφορά;


4.4
4.2
4
3.8
3.6
3.4
3.2
3
1
a
b
c
d

Έλεγχος Bonferroni

ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	0.954145833	3	0.318048611	7.219763035	0.001805552	3.098391212
Within Groups	0.88105	20	0.0440525			
Total	1.835195833	23				

$$H_0: \mu_i = \mu_j$$

$$H_\alpha: \mu_i \neq \mu_j$$

$$t = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$

$$SE = \sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

a vs b

$$t = \frac{3.80 - 3.41}{\sqrt{0.044\left(\frac{1}{6} + \frac{1}{6}\right)}}$$

$$t = \frac{0.39}{\sqrt{0.044(0.33)}}$$

$$t = \frac{0.39}{\sqrt{0.015}}$$

$$t = \frac{0.39}{0.1225}$$

$$t = 3.1837$$

$$20\ degrees\ of\ freedom$$

=T.DIST.2T(3.1837, 20)=0.004667

a vs b

$t = 3.1837$ =T.DIST.2T(3.1837, 20)=0.004667

- Η τιμή του $t - test$ ($t = 3.1837$) είναι μεγαλύτερη από το 5% σημείο της $t -$ κατανομής για 20 df (οι df του error) που είναι 2.09
- **Συνεπώς, υπάρχει σημαντική διαφορά μεταξύ των ομάδων a και b ($p < 0.05$ ή πιο συγκεκριμένα $p = 0.004$)**

95% διάστημα εμπιστοσύνης για μέση τιμή των διαφορών

$$\left((\bar{x}_a - \bar{x}_b) - t \times SE, D + t \times SE\right)$$

$$\left((3.80 - 3.41) - 2.09 \times 0.12, (3.80 - 3.41) + 2.09 \times 0.12\right)$$

$$(0.39 - 0.251, 0.39 + 0.251)$$

$$(0.139, 0.641)$$

Οπότε, με 95% βεβαιότητα η ομάδα a έχει μεγαλύτερο βάρος μεταξύ $\mathbf{0.14}$ και $\mathbf{0.64}$ από ότι η ομάδα b. Επειδή το 0 **δεν** συμπεριλαμβάνεται μέσα στο 95% CI, σημαίνει ότι η διαφορά είναι σημαντική

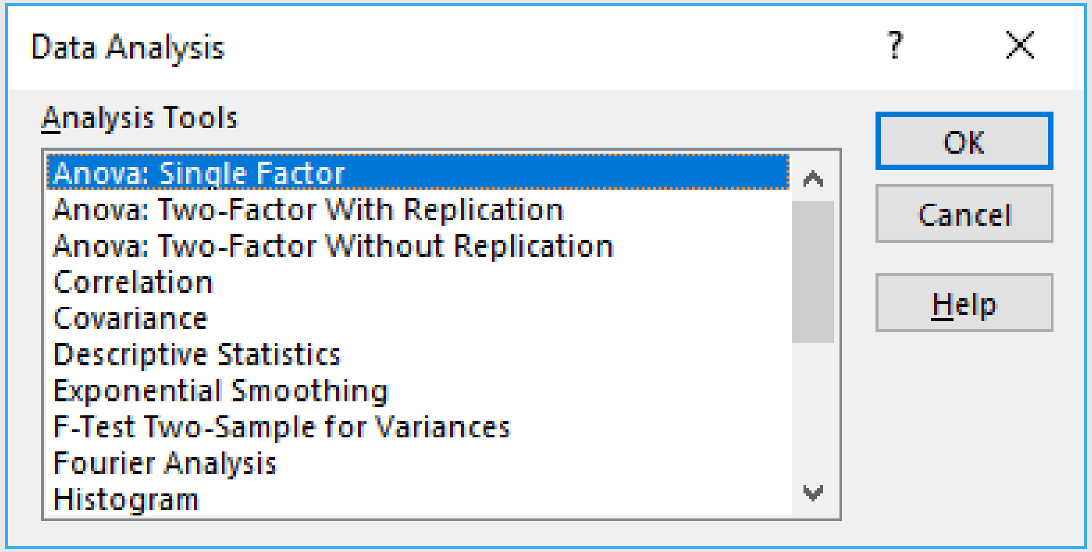
	t distribution		
	Percentage points of the t distribution		
	p-value		
degrees of freedom	0.05	0.01	0.001
	two tails	two tails	two tails
1	12.71	63.66	636.62
2	4.30	9.92	31.60
3	3.18	5.84	12.92
4	2.78	4.60	8.61
5	2.57	4.03	6.87
6	2.45	3.71	5.96
7	2.36	3.50	5.41
8	2.31	3.36	5.04
9	2.26	3.25	4.78
10	2.23	3.17	4.59
20	2.09	2.85	3.85

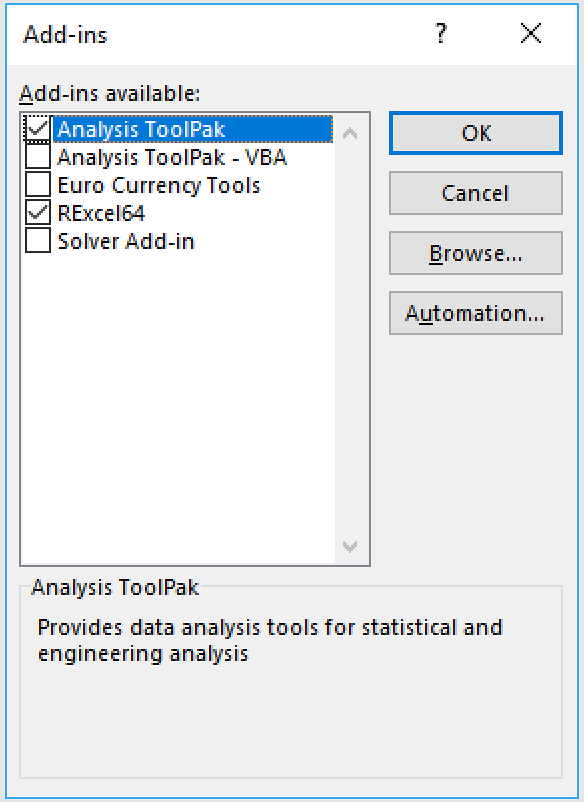
Διόρθωση Bonferroni

- Όλες οι πιθανές συγκρίσεις μεταξύ των ομάδων δεν είναι ανεξάρτητες και υπάρχει πάντα πιθανότητα να βρούμε ένα εσφαλμένο σημαντικό αποτέλεσμα
- Για τον λόγο αυτό όταν γίνονται πολλαπλές συγκρίσεις (k) μεταξύ ομάδων η στάθμη σημαντικότητας (P) πρέπει να διορθώνεται σε $P' = k * P$
- Συνεπώς, αν εκτελέσουμε 6 συγκρίσεις μεταξύ ομάδων, η σύγκριση μεταξύ της δίαιτας a και b θα είναι σημαντική σε $P = 6 * 0.004667 = 0.028$

Excel Data Analysis

File -> Options -> Add Ins -> Go … -> Analysis ToolPak

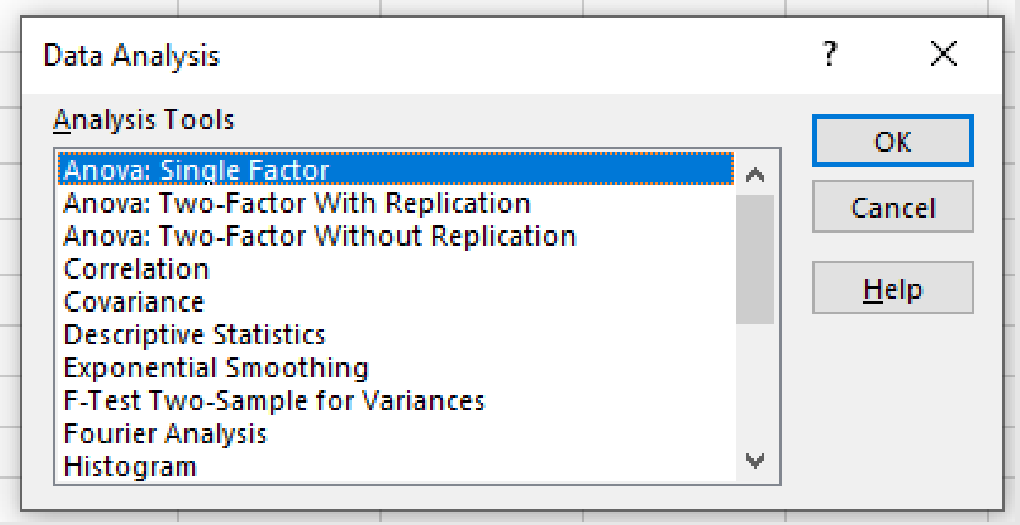




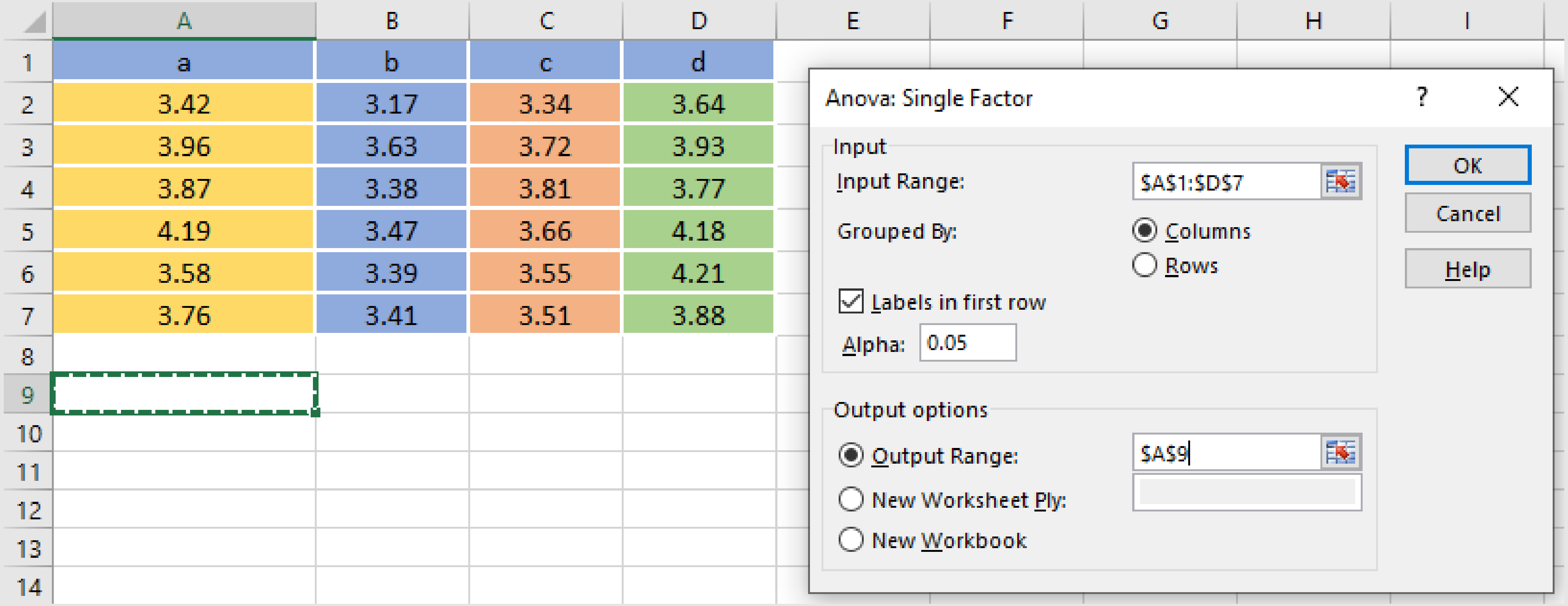
Excel Data Analysis

Data -> Data Analysis

a	b	c	d
3.42	3.17	3.34	3.64
3.96	3.63	3.72	3.93
3.87	3.38	3.81	3.77
4.19	3.47	3.66	4.18
3.58	3.39	3.55	4.21
3.76	3.41	3.51	3.88



Excel Data Analysis



	A	B	C	D
1	a	b	c	d
2	3.42	3.17	3.34	3.64
3	3.96	3.63	3.72	3.93
4	3.87	3.38	3.81	3.77
5	4.19	3.47	3.66	4.18
6	3.58	3.39	3.55	4.21
7	3.76	3.41	3.51	3.88

Excel Data Analysis

9	Anova: Single Factor						
10							
11	SUMMARY						
12	*Groups*	*Count*	*Sum*	*Average*	*Variance*		
13	a	6	22.78	3.79666667	0.07538667		
14	b	6	20.45	3.40833333	0.02217667		
15	c	6	21.59	3.59833333	0.02805667		
16	d	6	23.61	3.935	0.05059		
17							
18							
19	ANOVA						
20	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
21	Between Groups	0.95414583	3	0.31804861	7.21976304	0.00180555	3.09839121
22	Within Groups	0.88105	20	0.0440525			
23							
24	Total	1.83519583	23				