

ΣΤΑΤΙΣΤΙΚΟΙ ΕΛΕΓΧΟΙ

- One-Sample t-Test
- Independent Samples t-Test
- Dependent Samples t-Test
- ANOVA
- Linear regression

Σύγκριση μέσου όρου πληθυσμού με τιμή ελέγχου

One-Sample t -Test

Μια σύντομη αναδρομή

Στα τέλη του 19^{ου} αιώνα μια μεγάλη αλλαγή για την επιστήμη ζυμώνονταν στην ζυθοποιία Guinness.

Ο William Gosset ανακάλυψε μια νέα μέθοδο για να εκτιμήσει τη δοκιμασία γεύσης της μπύρας από ένα μικρό δείγμα δοκιμαστών.

Το αποτέλεσμα ήταν μια στατιστική μέθοδος που άλλαξε την επιστήμη – και ίσως την μπύρα



- Το 1908, ο Gosset δημοσίευσε τη μέθοδό του στο περιοδικό Biometrika με το ψευδώνυμο «μαθητής» ('student'). Και για αυτό συχνά η μέθοδος λέγεται και 'student's t test'.

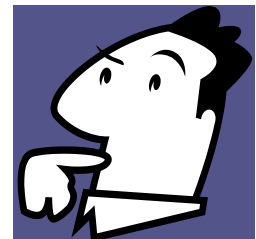


Πως δουλεύει το *one-sample t-test* για ΈΝΑ σύνολο?

- Ελέγχει εάν ο μέσος όρος του πληθυσμού από τον οποίο πήραμε δείγμα διαφέρει από μία τιμή ελέγχου.
 - Ποια είναι η τιμή ελέγχου;
 -
- Η τιμή ελέγχου συνήθως καθορίζεται από την πρότερη γνώση και από το ερώτημα που καλούμαστε να απαντήσουμε.

Παράδειγμα

- Θεωρητικά, τα τεστ ευφυΐας (που μετρούν το IQ) κατασκευάζονται έτσι ώστε ο μέσος όρος μεταξύ των ενηλίκων να είναι **100** μονάδες.
- Έστω ότι παίρνουμε ένα δείγμα έξι προπτυχιακών φοιτητών του πανεπιστημίου μας ($N = 6$), και ελέγχουμε εάν ο μέσος όρος του IQ των φοιτητών στο πανεπιστήμιο είναι υψηλότερος από 100.
- Με απλά λόγια, είναι οι φοιτητές μας εξυπνότεροι από το μέσο άνθρωπο;



- Έστω ότι οι τιμές IQ των έξι φοιτητών είναι

Φοιτητής 1: 110
Φοιτητής 2: 118
Φοιτητής 3: 110
Φοιτητής 4: 122
Φοιτητής 5: 110
Φοιτητής 6: 150



Ερευνητικό ερώτημα

Κατά μέσο όρο, ο πληθυσμός των προπτυχιακών φοιτητών του Πανεπιστημίου μας έχουν υψηλότερο δείκτη ευφυΐας από το μέσο άνθρωπο ($IQ = 100$);

- Πρώτα πρέπει να υπολογίσουμε το μέσο δείκτη ευφυΐας του δείγματος μας

$$\overline{X} = \frac{\sum X}{n} = \frac{110 + 118 + 110 + 122 + 110 + 150}{6} = 120$$

- Προφανώς αυτοί οι έξι έχουν μέσο όρο IQ υψηλότερο από 100, αλλά τι ισχύει στον πληθυσμό;

- Πόσο πιθανό είναι αυτό το αποτέλεσμα;
- Θα ίσχυε εάν είχαμε πάρει 6 διαφορετικούς φοιτητές;
- Τι θα συνέβαινε εάν παίρναμε πολλά σύνολα από 6 φοιτητές;

Το one-sample t -test θα μας απαντήσει αυτά τα ερωτήματα

Θα μας πει εάν το αποτέλεσμα είναι ‘σημαντικό’, ή με άλλα λόγια εάν είναι πιθανό να επαναληφθεί σε άλλο δείγμα του αυτού πληθυσμού.

Υπολογίζουμε τη διακύμανση των τιμών του δείγματος

IQ	Μέσος όρος	$X - \bar{X}$	$(X - \bar{X})^2$
110	120	-10	100
118	120	- 2	4
110	120	-10	100
122	120	2	4
110	120	-10	100
150	120	30	900

$$s_x^2 = \frac{\sum (X - \bar{X})^2}{n} = 201.\bar{3}$$

Για να βρούμε πόσο σημαντικά απέχει ο μέσος όρος από την τιμή ελέγχου χρειαζόμαστε ένα μέτρο να μετρήσουμε αυτή την απόσταση

$$\bar{X} - \mu$$

*Αλλά πρέπει να σταθμίσουμε αυτή την
απόσταση με τη διακύμανση του φαινομένου
που μελετάμε
Αυτό το σταθμισμένο μέτρο είναι η ποσότητα t*

$$t = \frac{\bar{X} - \mu}{\sqrt{\frac{s_x^2}{n-1}}}$$

Υπολογίζοντας το t στο παράδειγμα

$$t = \frac{\bar{X} - \mu}{\sqrt{\frac{s_x^2}{n-1}}} = \frac{120 - 100}{\sqrt{\frac{201.333}{6-1}}} \approx 3.152$$

Έχουμε μετρήσει 6 άτομα άρα **οι βαθμοί ελευθερίας** είναι $df = n - 1 = 5$

Οι βαθμοί ελευθερίας είναι σημαντικοί για τον υπολογισμό των πιθανοτήτων στο επόμενο βήμα

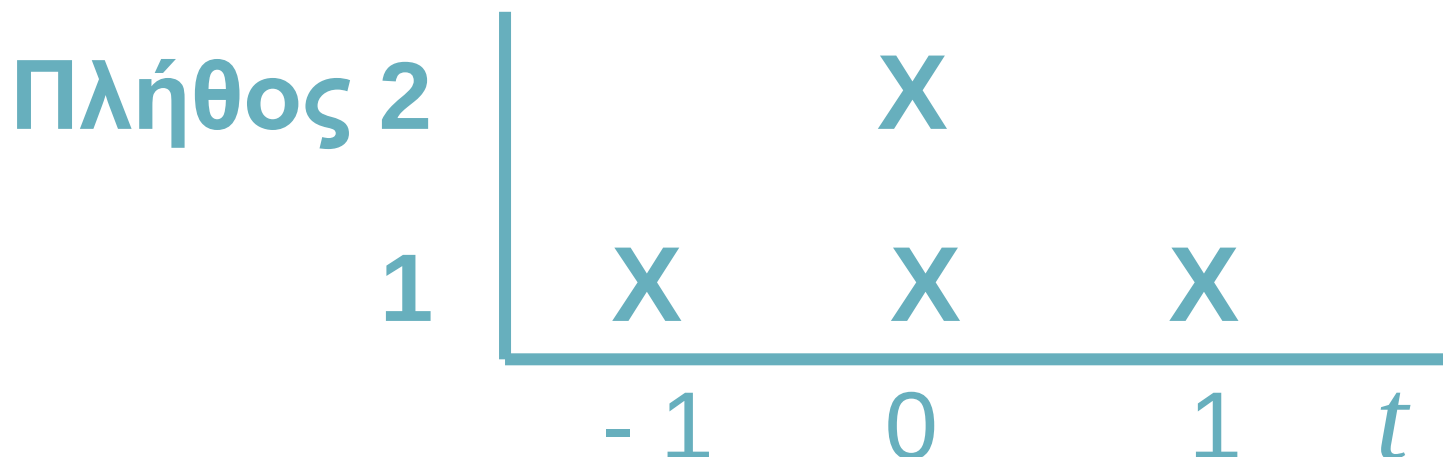
Και τώρα για κάτι διαφορετικό

Τώρα ας κάνουμε ένα διάλειμμα από τους τύπους και ας σκεφθούμε **την ευρύτερη εικόνα**.

Και αυτό είναι το δύσκολο μέρος.

Η κατανομή t

- Όπως οι καταγραφές, έτσι και η ποσότητα ' t ' ακολουθεί μια κατανομή.
- Εάν όντως η μέση τιμή IQ του πληθυσμού είναι ακριβώς 100, και εμείς πάρουμε πολλαπλά δείγματα 6 ατόμων, και υπολογίσουμε την τιμή t τότε τι περιμένουμε;
- Για παράδειγμα έστω 4 επαναλήψεις



Η κατανομή t

- Η πιο συχνή τιμή του ' t ' θα είναι 0. Γιατί; Γιατί ο αριθμητής στον τύπο υπολογισμού του t υπολογίζει τη διαφορά του μέσου όρου του δείγματος από την τιμή ελέγχου.
- Και εάν ο μέσος όρος είναι όντως ίσως με την τιμή ελέγχου, τότε τα περισσότερα δείγματα θα έχουν αυτόν το μέσο όρο.

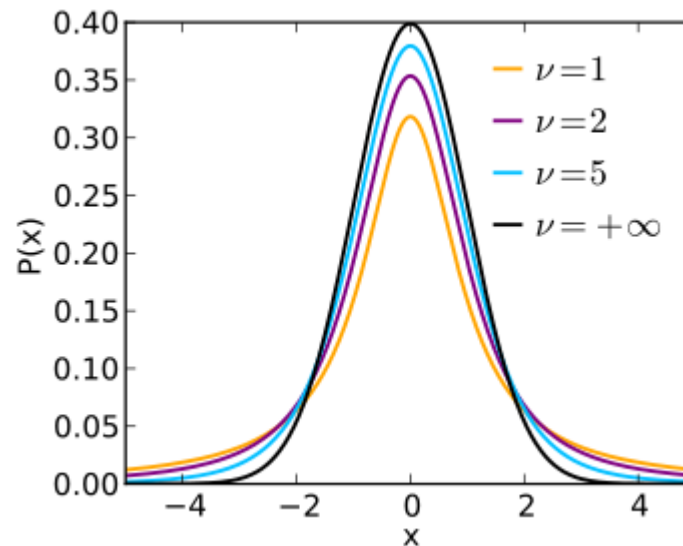
$$t = \frac{100 - 100}{\sqrt{\frac{s_x^2}{n - 1}}} = 0$$

Η κατανομή t

- Φυσικά πολλά δείγματα n ατόμων δε θα έχουν μέσο όρο ίσο με τον μέσο όρο του πληθυσμού
- Κάποια θα έχουν μεγαλύτερο μέσο όρο ($t > 0$) και κάποια μικρότερο ($t < 0$).
- Η συχνότητα με την οποία βρίσκουμε τιμές t μεγαλύτερες από 0 (ή μικρότερες από 0) είναι ακριβώς αυτό που εκτιμάει η κατανομή t .

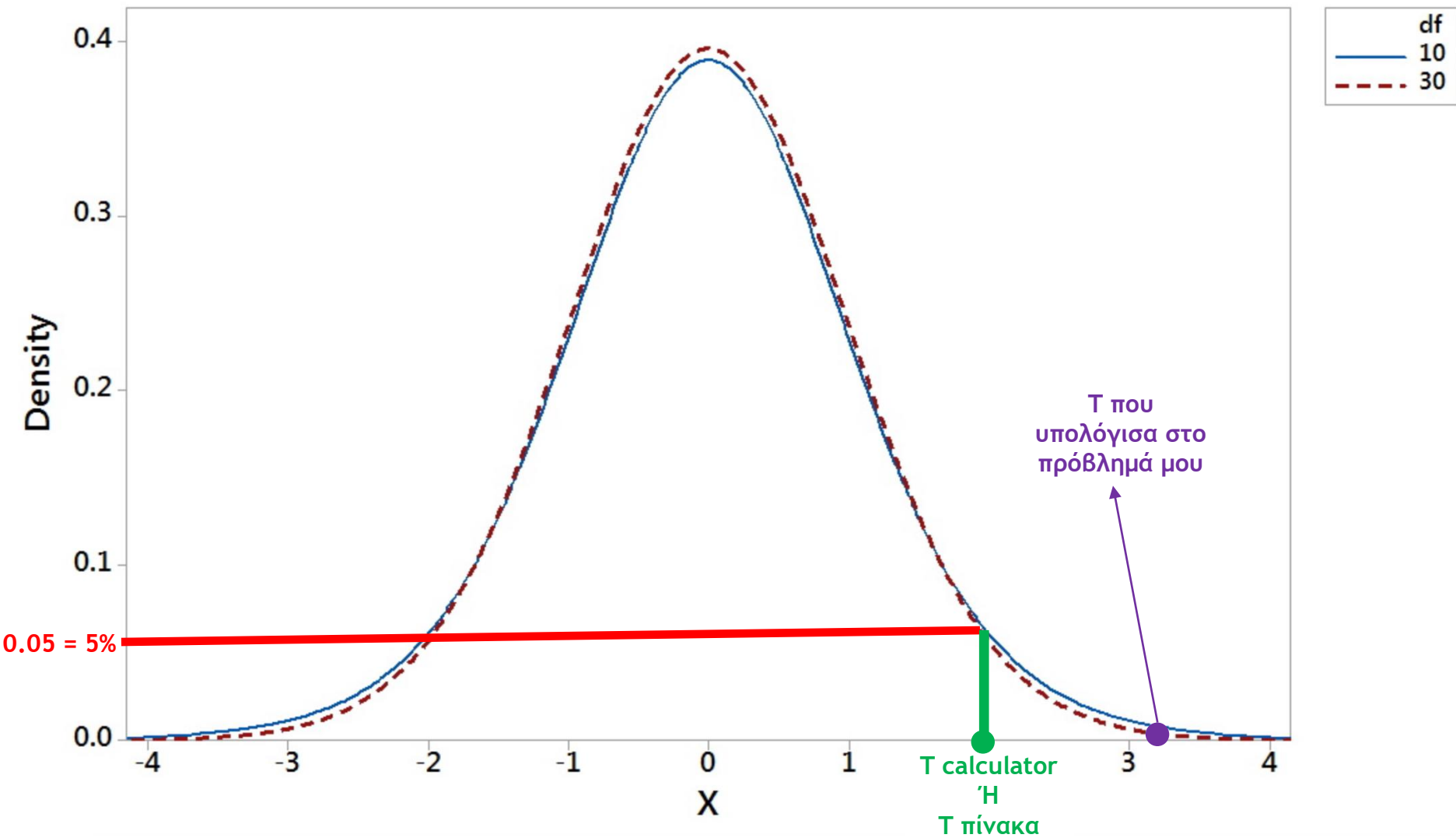
Η κατανομή t

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi} \Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad \Gamma(t) = \int_0^{\infty} x^{t-1} e^{-x} dx$$



Distribution Plot

T



Και τώρα προσοχή αυτό είναι το κρίσιμο μέρος

- Στο παράδειγμά μας το t είναι 3,152. Με άλλα λόγια είναι μεγαλύτερο από 0.
- Άρα δε φαίνεται το δείγμα μας να προήλθε από πληθυσμό με μέσο IQ 100.
- Εάν είχε έρθει θα περιμέναμε τιμή t κοντά στο 0.
- *Αλλά εάν είχε έρθει από τέτοιο πληθυσμό πόσο πιθανό θα ήταν να είχε καταγράψει τέτοια τιμή;*

Και τώρα προσοχή αυτό είναι το κρίσιμο μέρος

- Πως μπορούμε να υπολογίσουμε αυτή την πιθανότητα;
- Η κατανομή t μας εξυπηρετεί γιατί ακριβώς υπολογίζει πόσο συχνά θα παρατηρηθεί μια συγκεκριμένη τιμή t σε έναν πληθυσμό δειγμάτων με **μέση τιμή $t = 0$** .

Η κατανομή t

- Η κατανομή t που θα πρέπει να χρησιμοποιήσουμε εξαρτάται από τους βαθμούς ελευθερίας ($df = 5$).
- Γιατί;
 - Εάν μετρήσουμε μόνο λίγα άτομα, υπάρχει πιθανότητα να πετύχουμε ανάμεσα σε αυτά κάποια ακραία τιμή που θα επηρεάσει έντονα το δειγματικό μέσο όρο και άρα θα προκύψει μεγάλη τιμή της μετρικής t .
 - Γενικά, **μικρά δείγματα είναι λιγότερο αντιπροσωπευτικά του πληθυσμού.**

Και επίσης κρίσιμο

- Πότε θεωρούμε μια διαφορά **σημαντική**;
- Η πιο απλά πότε μπορούμε να απαντήσουμε το ερευνητικό μας ερώτημα με ναι ή όχι;

Ερευνητικό ερώτημα

Κατά μέσο όρο, ο πληθυσμός των προπτυχιακών φοιτητών του Πανεπιστημίου μας έχουν υψηλότερο δείκτη ευφυΐας από το μέσο άνθρωπο ($IQ = 100$);

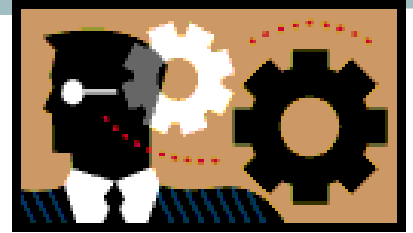
Μονόπλευρος ή αμφίπλευρος έλεγχος

- Το ερώτημα μας είναι διατυπωμένο ως **μονόπλευρος** έλεγχος, γιατί ρωτάει εάν ο μέσος όρος είναι **μεγαλύτερος** από την τιμή ελέγχου.
- Εάν το ερώτημα ήταν αν **διαφέρει** ο μέσος όρος από την τιμή ελέγχου (δηλαδή είναι είτε μικρότερος είτε μεγαλύτερος) τότε θα μιλήσουμε για τον **αμφίπλευρο έλεγχο** που είναι και η πιο συνηθισμένη περίπτωση
- Στην περίπτωση του μονόπλευρου ελέγχου για 5 βαθμούς ελευθερίας η κρίσιμη τιμή του t είναι **2,02**.

Απάντηση στο παράδειγμα

Στο παράδειγμά μας $t = 3,152$
άρα μεγαλύτερο από την κρίσιμη τιμή 2,02.
Άρα η διαφορά είναι σημαντική και μπορούμε να
απαντήσουμε το ερευνητικό ερώτημα θετικά.

Ή με πιο τεχνικούς όρους μπορούμε να
απορρίψουμε τη μηδενική υπόθεση ότι το
δείγμα μας προέρχεται από πληθυσμό με μέσο
όρο 100.



Μηδενική υπόθεση

- Στη στατιστική δοκιμασία που περιγράψαμε υπάρχει μια μηδενική και μια εναλλακτική υπόθεση.
- Η μηδενική υπόθεση ήταν ότι ο μέσος όρος του πληθυσμού είναι ίσος ή μικρότερος από την τιμή ελέγχου.
- Η εναλλακτική υπόθεση ήταν ότι ο μέσος όρος του πληθυσμού είναι μεγαλύτερος από την τιμή ελέγχου.

Τι μας είπε η κατανομή t

- Η κατανομή t μας είπε ότι στο 95% των δειγμάτων (μεγέθους 6 ατόμων) που θα πάρουμε από πληθυσμό με μέση ευφυΐα 100 η τιμή t θα είναι μικρότερη από 2,02
- Και μόνο σπάνια ($p < 0,05$) θα παρατηρούνται τιμές t μεγαλύτερες του 2,02

Άρα τι σημαίνει ότι εμείς είχαμε $t=3.152$

- Άρα με πιθανότητα σφάλματος μικρότερη από **5%** ($p < 0.05$) μπορούμε να ισχυριστούμε ότι το δείγμα μας δεν προήλθε από πληθυσμό με μέση ευφυΐα 100.
- Με βάση το ερώτημα μπορούμε να ισχυριστούμε ότι προήλθε από πληθυσμό με μέση ευφυΐα μεγαλύτερη από 100.
- Πόσο συχνά θα είναι εσφαλμένος αυτός ο ισχυρισμός;
- Αυτό είναι το επίπεδο στατιστικής σημαντικότητας

Πως το διατυπώνουμε αυτό επιστημονικά;



Ο μέσος όρος του δείκτη ευφυΐας των φοιτητών του πανεπιστημίου μας είναι σημαντικά μεγαλύτερος από τον αναμενόμενο μέσο όρο (τιμή ελέγχου 100) σύμφωνα με τη μονόπλευρη στατιστική δοκιμασία t για έναν πληθυσμό ($t(5)=3,152$ $p=0,012<0,05$).

Στατιστική σημαντικότητα

- Σύμφωνα με την επιστημονική παράδοση του εικοστού αιώνα, τιμές p οι οποίες είναι χαμηλότερες από 0,05 θεωρούνται σημαντικές.

Παράδειγμα

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Intelligence Quotient	6	120.0000	15.54349	6.34560

One-Sample Test

	Test Value = 100					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Intelligence Quotient	3.152	5	.025	20.00000	3.6881	36.3119

Παράδειγμα

GraphPad QuickCalcs: t test | ΝΕΚ δημοσιογράφων εκβ...

graphpad.com/quickcalcs/ttest1.cfm

Cart SEARCH Sign In

GraphPad Software

Scientific Software | Data Analysis Resource Center | Company | Support | How to Buy

QuickCalcs

1. Select category 2. Choose calculator 3. Enter data 4. View results

t test calculator

A *t* test compares the means of two groups. For example, compare whether systolic blood pressure differs between a control and treated group, between men and women, or any other two groups. Don't confuse *t* tests with correlation and regression. The *t* test compares one variable (perhaps blood pressure) between two groups. Use correlation and regression to see how two variables (perhaps blood pressure and heart rate) vary together. Also don't confuse *t* tests with ANOVA. The *t* tests (and related nonparametric tests) compare exactly two groups. ANOVA (and related nonparametric tests) compare three or more groups. Finally, don't confuse a *t* test with analyses of a contingency table (Fishers or chi-square test). Use a *t* test to compare a continuous variable (e.g., blood pressure, weight or enzyme activity). Use a contingency table to compare a categorical variable (e.g., pass vs. fail, viable vs. not viable).

1. Choose data entry format

- ☒ Enter up to 50 rows.
- ☐ Enter or paste up to 2000 rows.
- ☐ Enter mean, SEM and N.
- ☐ Enter mean, SD and N.

Caution: Changing format will erase your data.

2. Enter data

[Help me arrange the data.](#)

Label:

	Group 1	Group 2
1		
2		
3		
4		
5		
6		
7		
8		
9		

3. Choose a test

- ☒ Unpaired *t* test.
- ☐ Welch's unpaired *t* test (used rarely).
- ☐ Paired *t* test.

[Help me decide.](#)

4. View the results

GraphPad Prism
Organize, analyze and graph and present your scientific data.
[MORE >](#)

InStat
With InStat® you can analyze data in a few minutes.
[MORE >](#)

StatMate
StatMate® calculates sample size and power.
[MORE >](#)

 Questions about curve fitting or statistics?
[More >](#)

Στατιστική συμπερασματολογία

- Καταλήξαμε σε ένα συμπέρασμα για τον πληθυσμό των φοιτητών στηριζόμενοι σε ένα μικρό δείγμα.

Τελειώσαμε;

- ΟΧΙ
- Κάθε στατιστική δοκιμασία ισχύει μόνο κάτω από κάποιες προϋποθέσεις.

Προϋποθέσεις εφαρμογής του one sample t-test

- Αντιπροσωπευτικό δείγμα (αν το δείγμα είναι στρεβλό, κάθε συμπέρασμα είναι αμφίβολο)
- **Κανονική κατανομή** (η κατανομή t προκύπτει για επαναλαμβανόμενη δειγματοληψία σε πληθυσμό που ακολουθεί την κανονική κατανομή)

Σύγκριση δύο πληθυσμών

Independent samples t-
test

Τι κάνει το independent samples t test

- Συγκρίνει τους μέσους όρους δύο ανεξάρτητων ομάδων προκειμένου να καθορίσει αν οι μέσοι όροι των πληθυσμών από όπου προήρθαν τα δείγματα είναι σημαντικά διαφορετικοί

Σύγκριση μέσων όρων δύο
πληθυσμών

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{S_x^2}{N_x} + \frac{S_y^2}{N_y}}}$$

Η «δίαιτα» της σοκολάτας

Λίγη σοκολάτα μερικές φορές την εβδομάδα βοηθά στη διατήρηση μιας λεπτής σιλουέτας

ΔΗΜΟΣΙΕΥΣΗ: 27/03/2012, 13:47



Οι αμερικανοί επιστήμονες συστήνουν την κατανάλωση μαύρης σοκολάτας με μέτρο

[Tweet](#) 4
 [Share](#) 20
 [Like](#) 108
 [email](#)
[εκτύπωση](#)

Αθήνα

Όσοι τρώνε συχνά σοκολάτα, τείνουν να είναι πιο αδύνατοι, σύμφωνα με μία νέα αμερικανική επιστημονική έρευνα, η οποία διαπίστωσε ότι οι άνθρωποι που καταναλώνουν σοκολάτα μερικές φορές κάθε εβδομάδα, είναι κατά μέσο όρο πιο λεπτοί σε σχέση με όσους τρώνε μόνο κατά περίπτωση.

Οι ερευνητές, με επικεφαλής την δρ Μπεατρίς Γκολόμπ του **Πανεπιστημίου της Καλιφόρνια-Σαν Ντιέγκο**, που δημοσίευσαν τη σχετική μελέτη στο περιοδικό του Αμερικανικού Ιατρικού Συλλόγου Archives of Internal Medicine, μελέτησαν τη διατροφή

Συνάχι;
Βουλωμένη μύτη;
Καταρροή;

(ανακουφίσου)

- Ιατρική – Βιολογία
- Φυσική – Διάστημα
- Τεχνολογία – Πλανήτης
- Μαθηματικά – Πληροφορική
- Ψυχολογία – Κοινωνιολογία
- Ερευνα στην Ελλάδα

ΔΙΑΒΑΣΜΕΝΑ | ΣΧΟΛΙΑΣΜΕΝΑ

Ιατρική – Βιολογία- Τατουάζ...δοντιού εναντίον λοιμώξεων 03/04/2012, 17:02
[Σχολιάστε](#)

Ιατρική – Βιολογία- Φάλαινες εξπέρ στις μπουρμπουλήθρες! 03/04/2012, 13:02
[Σχόλια \(1\)](#)

Ιατρική – Βιολογία- Το «πλεκτό» του εγκεφάλου 03/04/2012, 18:03
[Σχολιάστε](#)

Ιατρική – Βιολογία- Ήταν οι δεινόσαυροι υδρόβιοι; 13:58
[Σχολιάστε](#)

Ένα παράδειγμα

Έστω δύο πληθυσμοί, αυτοί που έφαγαν σοκολάτα και αυτοί που δε έφαγαν

Το πρώτο δείγμα αποτελείται από 9 άτομα με μέσο όρο βάρους 50.3 κιλά και τυπική απόκλιση 4.5 κιλά

Το δεύτερο δείγμα αποτελείται από 12 άτομα με μέσο όρο βάρους 60.6 κιλά και τυπική απόκλιση 9.6 κιλά

Είναι αληθινή η διαφορά αυτή?

Διαφέρουν οι δύο μέσοι όροι

$$t = \frac{50,3 - 60,6}{\sqrt{\frac{4,5^2}{9} + \frac{5,6^2}{12}}} = \frac{-10,3}{\sqrt{\frac{20.25}{9} + \frac{31.36}{12}}}$$

Διαφέρουν οι δύο μέσοι όροι

$$t = \frac{-10,3}{\sqrt{2,25 + 2,61}} = \frac{-10,3}{2,2} = -4,67$$

Βαθμοί ελευθερίας

$$df = N_x + N_y - 2 = 9 + 12 - 2 = 19$$

Και αφού $t = -4,67$

$$P = 0,00016$$

Προϋποθέσεις

- Αντιπροσωπευτικότητα δείγματος
- Κανονικότητα κατανομής
- Ομοιογένεια διακυμάνσεων

Αν δεν ισχύουν οι προϋποθέσεις

- Εξηγήσαμε τι κάνουμε όταν δεν έχουμε ομοιογένεια διακυμάνσεων.
- Τι γίνεται αν δεν έχουμε κανονική κατανομή;
- Μία λύση ο μετασχηματισμός των δεδομένων
- Δεύτερη λύση η μη παραμετρική στατιστική

Μη παραμετρικός έλεγχος για τη σύγκριση διαμέσων

- Mann Whitney U test
- Συγκρίνει τους διάμεσους δύο πληθυσμών (και όχι τους μέσους όρους).

Μη-παραμετρική στατιστική

- Δε στηρίζεται στις τιμές των μετρήσεων αλλά στην τάξη τους.

Online application (Internet)

- <http://www.socscistatistics.com/tests/studentttest/Default2.aspx>
- <http://graphpad.com/quickcalcs/ttest1.cfm>
- ... και άλλα.

Έλεγχος t για εξαρτημένα δείγματα

Paired samples t-test

Τι ξέρουμε ήδη;

- Να **συγκρίνουμε το μέσο όρο των δύο πληθυσμών**, εάν αυτοί είναι **ανεξάρτητοι**.
- (Independent samples t-test)

Διαφορά με τα προηγούμενα

- Τι γίνεται αν οι καταγραφές δεν είναι ανεξάρτητες;

Παράδειγμα

- Έστω ότι μετρήσαμε τις επιδόσεις πέντε αθλητών πριν και μετά από προπόνηση ενός μηνός.
- Ο λόγος για να δούμε αν το συγκεκριμένο πρόγραμμα προπόνησης επηρέασε τις επιδόσεις των αθλητών.

	$X_1 :$ Πριν	$X_2 :$ Μετά
Aθλ1:	1	2
Aθλ2:	2	4.5
Aθλ3:	2	3
Aθλ4:	2	4.5
Aθλ5:	3	6

Τι βλέπουμε εδώ;

- Οι καταγραφές δεν είναι ανεξάρτητες.
- Κάθε γραμμή του πίνακα αντιστοιχεί **στον ίδιο αθλητή**. Άρα εξαρτημένα δείγματα
- Π.χ. ο *αθλητής 1* είχε επίδοση **1** πριν την προπόνηση και **2** μετά από αυτή.

	$X_1 :$ Πριν	$X_2 :$ Μετά	$(X_1 - \bar{X}_1)^2$	$(X_2 - \bar{X}_2)^2$
Αθλ1:	1	2	1	4
Αθλ2:	2	4.5	0	0.25
Αθλ3:	2	3	0	1
Αθλ4:	2	4.5	0	0.25
Αθλ5:	3	6	1	4
Μέσος όρος	2	4		
Τυπική απόκλιση	$s_x^2 = \frac{\sum (X - \bar{X})^2}{n} =$		0.4	1.9

One sample t-test

- Μπορούμε να συγκρίνουμε τον μέσο όρο των διαφορών με **τιμή ελέγχου το μηδέν** (0).
- Αν ο μέσος όρος είναι σημαντικά διαφορετικός σημαίνει ότι η προπόνηση επηρέασε τις επιδόσεις των αθλητών. Αλλιώς δεν είχε αποτέλεσμα.

Τι εξετάζουμε με το paired samples t-test

- **Δε συγκρίνουμε εάν ο μέσος όρος συνολικά αυξήθηκε ή μειώθηκε, αλλά εάν οι επιδόσεις του κάθε αθλητή άλλαξαν**
π.χ. θα μπορούσε να αυξηθεί ο μέσος όρος γιατί κάποιος αθλητής βελτίωσε πολύ την επίδοσή του ενώ όλοι οι άλλοι χειροτέρευσαν λίγο.
- **Και όπως κάθε στατιστικός έλεγχος, δεν επιχειρεί να βγάλει συμπέρασμα για το δείγμα αλλά για τον πληθυσμό.**
Με άλλα λόγια, εάν είχαμε πάρει 5 διαφορετικούς αθλητές πόσο πιθανό θα ήταν να είχαν επίσης βελτίωση στις επιδόσεις τους;

Πως δουλεύει;

	X_1 : Πριν	X_2 : Μετά	$X_1 - X_2$
Aθλ1:	1	2	-1
Aθλ2:	2	4.5	-2.5
Aθλ3:	2	3	-1
Aθλ4:	2	4.5	-2.5
Aθλ5:	3	6	-3
Μέσος όρος			-2
Τυπική απόκλιση			0,935

$$t = \frac{\bar{X} - \mu}{\sqrt{\frac{s_x^2}{n-1}}}$$

Πως δουλεύει;

$$t = \frac{\bar{X} - \mu}{\sqrt{\frac{s_x^2}{n}}} = \frac{-2 - 0}{\sqrt{\frac{0,935^2}{5}}} = -4.783$$

Πως δουλεύει;

$t = -4,783$, $df=4$, οπότε $p=0.009$

Άρα με **$p=0.009$** μπορούμε να ισχυριστούμε ότι η προπόνηση επηρέασε **σημαντικά** τις επιδόσεις των αθλητών.

Προϋποθέσεις εφαρμογής

- Αντιπροσωπευτικότητα δείγματος
 - Κανονικότητα κατανομής **των διαφορών**
-
- Απαιτείται να υπάρχει εξάρτηση μεταξύ των καταγραφών και οι καταγραφές να αφορούν την ίδια ποσοτική μεταβλητή.

Πότε το χρησιμοποιούμε;

- Η δοκιμασία t για **εξαρτημένα** δείγματα εκτιμά πόσο αξιόπιστα παρατηρούνται οι **διαφορές** ως συνέπεια κάποιου χειρισμού.

Ανάλυση διακύμανσης (Μονοδιάστατη)

One-Way ANOVA

A series of horizontal lines in teal and white, extending from the left edge of the slide and ending on the right, positioned below the 'One-Way ANOVA' text.

Ανάλυση διακύμανσης

- Η μονοδιάστατη ανάλυση διακύμανσης εξετάζει εάν δύο ή περισσότεροι ανεξάρτητοι πληθυσμοί έχουν τον ίδιο ή διαφορετικό μέσο όρο.
- Υπενθυμίζω ότι 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 συγκρίνουμε
 - Η ανεξάρτητη μεταβλητή έχει k επίπεδα (όπου $k \geq 2$)
- 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
- Τα υπολείμματα ακολουθούν την κανονική κατανομή

Μηδενική υπόθεση

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

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

- 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

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
 - Μέσσια: 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	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

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

Διακύμανση

- Είναι όλες οι τιμές ίδιες?
 - Όχι άρα υπάρχει διακύμανση στα δεδομένα
 - Η διακύμανση μπορεί να μετρηθεί για το σύνολο των δεδομένων
 - Ως $SS(\text{Total})$ 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 $SS(\text{Within})$ αναφερόμαστε στο μέρος του αθροίσματος των τετραγώνων που οφείλεται ακριβώς στη διακύμανση μεταξύ των ατόμων του κάθε πληθυσμού

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 $SS(\text{Between})$ 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 αν 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

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

Διακύμανση μεταξύ των πληθυσμών (παράδειγμα)

$$SS(B)=1902$$

$$SS(B) = 7(75.71 - 65.08)^2 + 9(67.11 - 65.08)^2 + 8(53.50 - 65.08)^2$$

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

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

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

- Στο παράδειγμά μας $SS(\text{Within}) = 3386$

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

- $MS(B) = 1902 / 2 = 951.0$
- $MS(W) = 3386 / 21 = 161.2$
- $MS(T) = 5288 / 23 = 229.9$
 - Το $MS(Total)$ 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		

SPSS

ANOVA

VAR00001

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1901,516	2	950,758	5,896	,009
Within Groups	3386,317	21	161,253		
Total	5287,833	23			

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

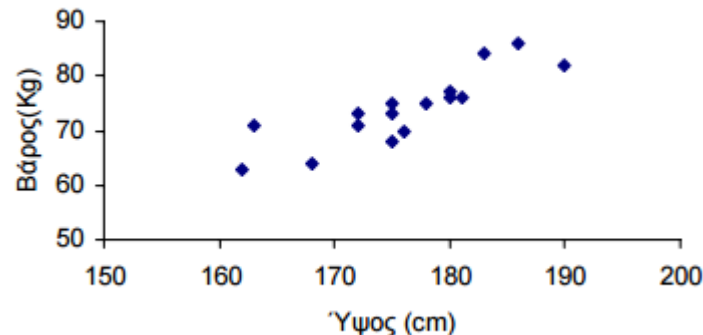
Συμπέρασμα

- 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 τα post hoc tests

ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ (LINEAR REGRESSION)

- **Πρόβλημα:** η σχέση δύο ή περισσότερων τυχαίων μεταβλητών.

Ας πούμε η σχέση μεταξύ ύψους και βάρους των φοιτητών



1. Ερευνάμε αν υπάρχει μια σχέση μεταξύ δύο μεταβλητών
2. Αν **ναι**, καθορίζουμε αυτή τη σχέση.

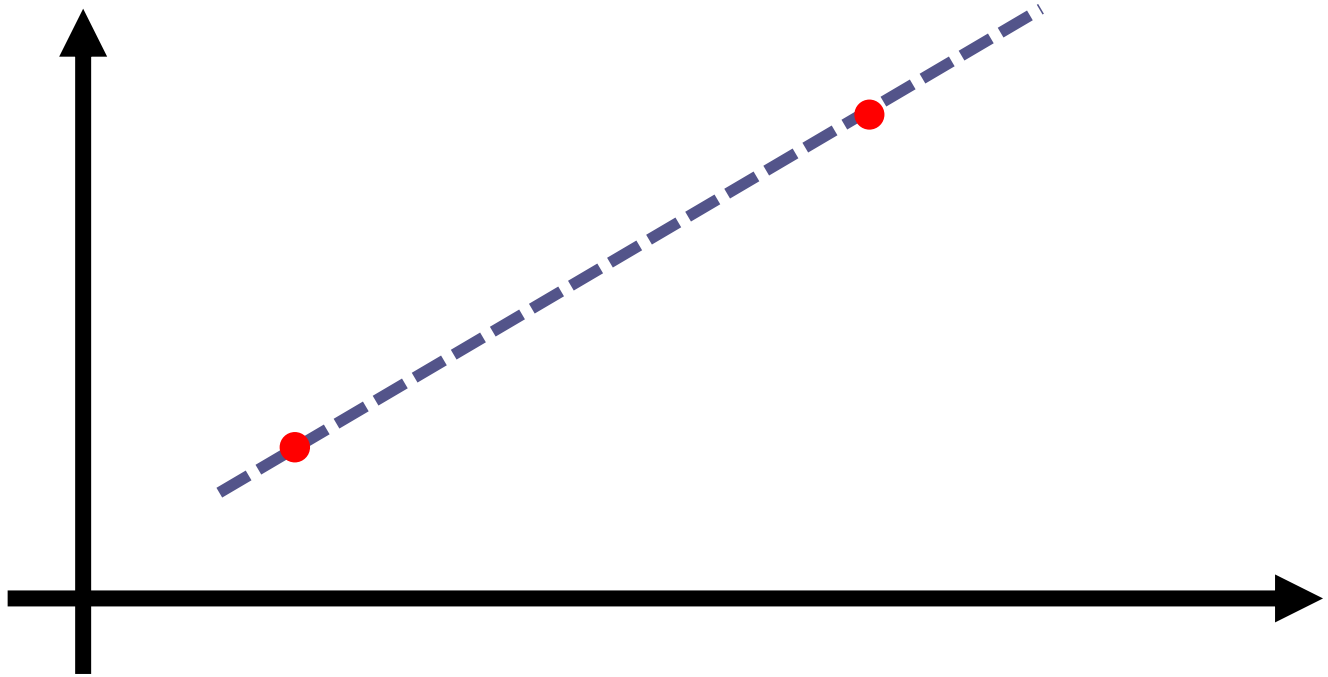
- Κατασκευή της μαθηματικής εξίσωσης που περιγράφει τη σχέση

Η απλούστερη μορφή τέτοιας σχέσης είναι η ευθεία

$$Y = aX + b$$

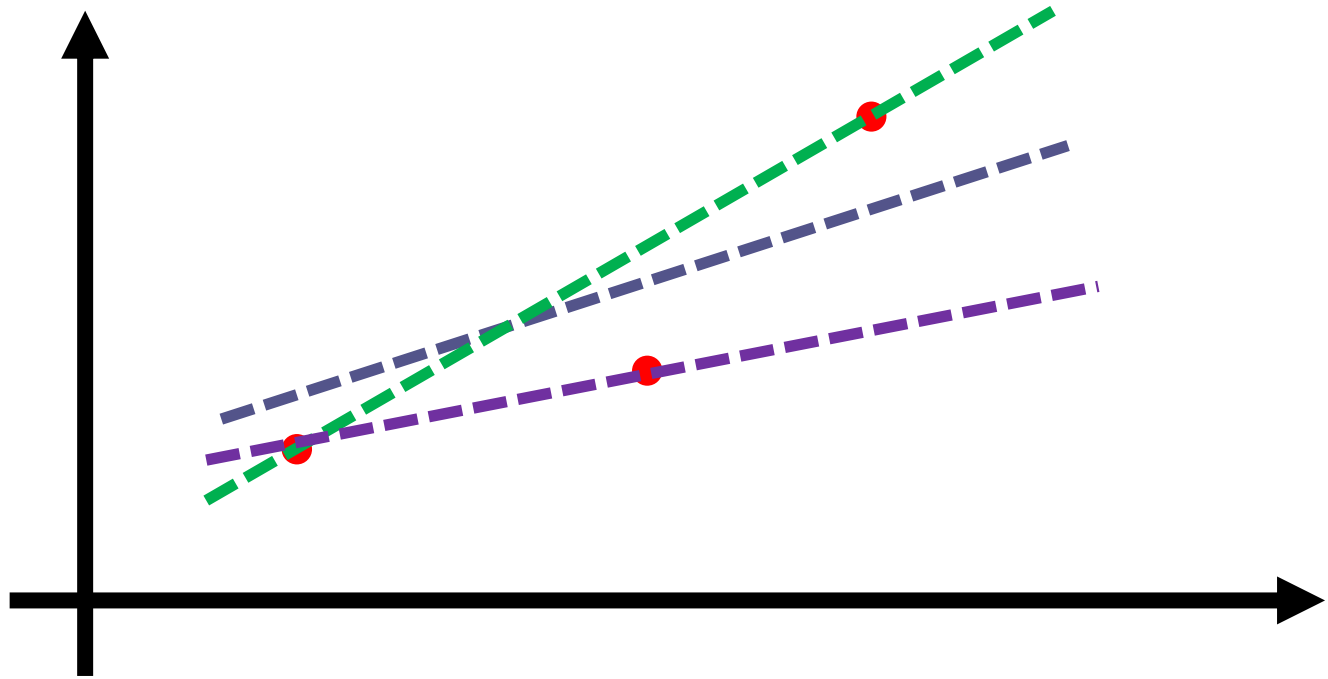
Πρέπει να βρούμε τα a και b

- Για να σκεφτούμε ...



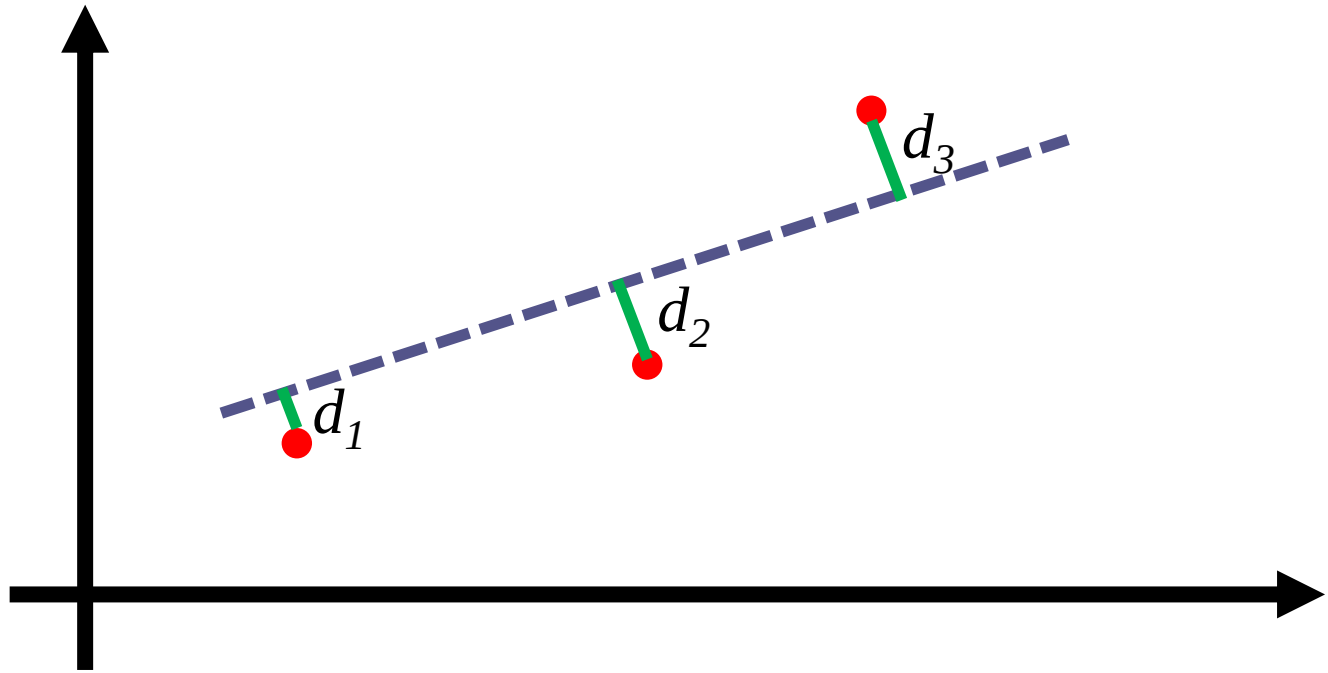
Δυο σημεία = μια και μοναδική ευθεία

- Για να σκεφτούμε ...



Τρία σημεία = ποιά ευθεία???

- Για να σκεφτούμε ...



Θέλω $d_1 + d_2 + d_3$ να είναι **ΕΛΑΧΙΣΤΟ**

***ΕΥΚΛΕΙΔΙΑ ΑΠΟΣΤΑΣΗ ΔΥΟ ΣΗΜΕΙΩΝ
ΤΟΥ ΕΠΙΠΕΔΟΥ***

$$A(x_1, y_1) \quad B(x_2, y_2)$$

$$d_{AB} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Οπότε γενικά ...

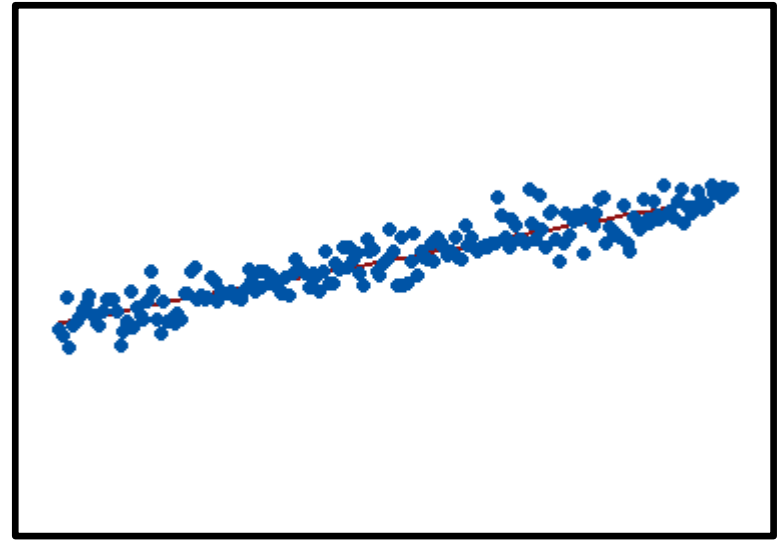
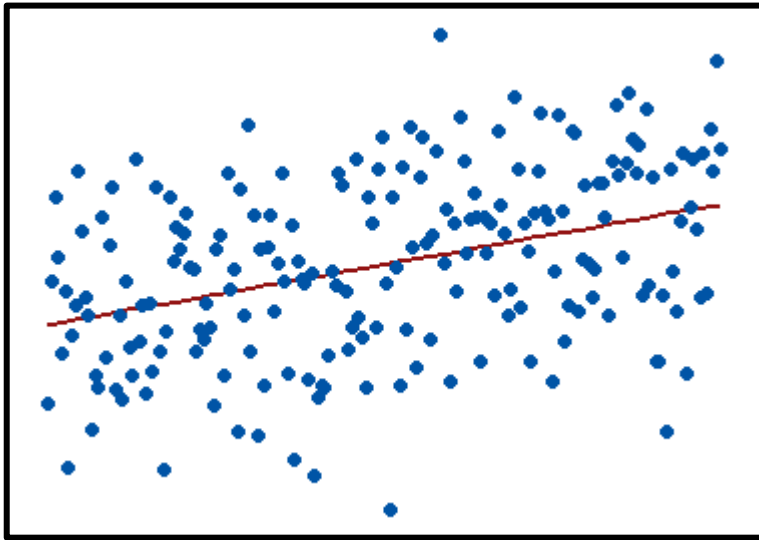
$$a = \frac{N \sum xy - (\sum x)(\sum y)}{N \sum x^2 - (\sum x)^2}$$

$$b = \frac{\sum y}{N} - a \frac{\sum x}{N}$$

Είναι «καλή» αυτή η ευθεία;;;
Συντελεστής προσδιορισμού r^2
= το ποσοστό της συνολικής
μεταβολής του Y που οφείλεται
στη μεταβολή του X .

$$r^2 = \left[\frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2} \sqrt{\sum (Y - \bar{Y})^2}} \right]^2$$

Είναι «καλή» αυτή η ευθεία;;;



Όσο πλησιέστερα είναι στο **1**, τόσο **ισχυρότερη** είναι η γραμμική συσχέτιση των Y και X

Για τιμές κοντά στο **0** έπεται ότι **δεν υπάρχει** καλή γραμμική συσχέτιση.