Εφαρμοσμένη Στατιστική με Έμφαση στις Επιστήμες Υγείας
Κωδικός Βιβλίου στον Εύδοξο: 59397001
Συγγραφείς: Α. Σαχλάς & Σ. Μπερσίμης

Βιοστατιστική
Κωδικός Βιβλίου στον Εύδοξο: 41236
Συγγραφείς: Δ. ΤΡΙΧΟΠΟΥΛΟΣ, Α. ΤΖΩΝΟΥ, Κ. ΚΑΤΣΟΥΓΙΑΝΝΗ

ΣΤΑΤΙΣΤΙΚΗ :

Παράσταση – Περιγραφή δεδομένων

Σύγκριση δεδομένων – Εξαγωγή συμπερασμάτων

Σχέση αιτίου - αιτιατού

Με τις στατιστικές μεθόδους επιδιώκεται αφενός η συνοπτική αλλά εμπεριστατωμένη παρουσίαση των ευρημάτων μιας μελέτης (**περιγραφική στατιστική**) και αφετέρου η συναγωγή συμπερασμάτων που βασίζονται στα ευρήματα αυτά (**συμπερασματολογική στατιστική / επαγωγική στατιστική**)

Πιθανότητα (P, Probability) είναι μέτρο του πόσο αναμενόμενο να συμβεί ένα γεγονός ή μια θέση (ισχυρισμός) να είναι αληθής. Οι πιθανότητες παίρνουν τιμές μεταξύ 0 (δεν θα συμβεί) και 1 (θα συμβεί). Όσο μεγαλύτερη η πιθανότητα ενός γεγονότος, τόσο βέβαιοι είμαστε ότι θα συμβεί.

Ως **μεταβλητή** θεωρούμε κάθε χαρακτηριστικό το οποίο μπορεί να μεταβληθεί ή να διαφοροποιηθεί κατά μήκος του χρόνου, από τόπο σε τόπο, από άτομο σε άτομο ή από ομάδα σε ομάδα (π.χ. ηλικία, ύψος, εισόδημα, συγκέντρωση χοληστερόλης, αρτηριακή πίεση, ρυθμός γεννητικότητας κτλ)

- **Ποιοτική** ονομάζεται η μεταβλητή που περιγράφει κάποιο ποιοτικό χαρακτηριστικό ενός ατόμου ή μιας ομάδας.
- **Ποσοτική** ονομάζεται η μεταβλητή που μπορεί να μετρηθεί με τη συνήθη έννοια του όρου
  - Συνεχής
  - Ασυνεχής

- Ως **ανεξάρτητη** χαρακτηρίζεται μια μεταβλητή όταν επηρεάζει μια άλλη μεταβλητή.
- Ως **εξαρτημένη** χαρακτηρίζεται μια μεταβλητή όταν επηρεάζεται από μια άλλη μεταβλητή.

ΠΕΡΙΓΡΑΦΗ ΔΕΔΟΜΕΝΩΝ

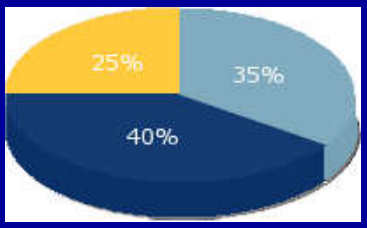
Γραφικές μέθοδοι

- **Ραβδογράμματα - Ιστογράμματα (Συχνότητα)**
- **Κυκλικά διαγράμματα**
- **Διαγράμματα πλαισίου**

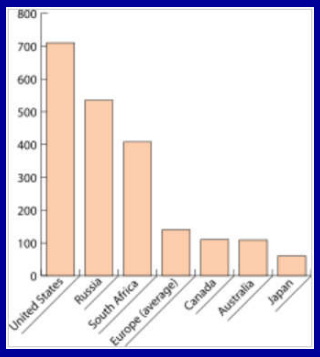
Αριθμητικοί στατιστικοί δείκτες ή μέτρα *(statistics)*

- **Κεντρικής τάσης**
  - **Μέση τιμή *(mean)***
  - **Διάμεσος *(median)***
  - **Επικρατούσα τιμή *(mode)***
- **Διασποράς**
  - **Εκατοστημόρια ή ποσοστιαία σημεία *(percentiles)***
  - **Διακύμανση ή Διασπορά *(Variance)***
  - **Τυπική απόκλιση *(standard deviation)***

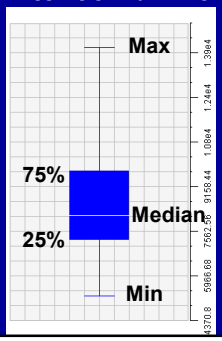
Διάγραμμα πίτας (Pie chart)



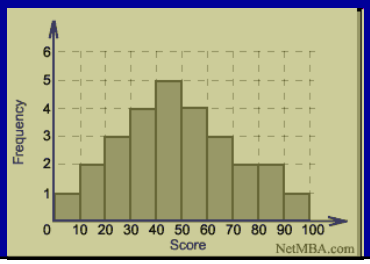
Ραβδόγραμμα (bar chart)



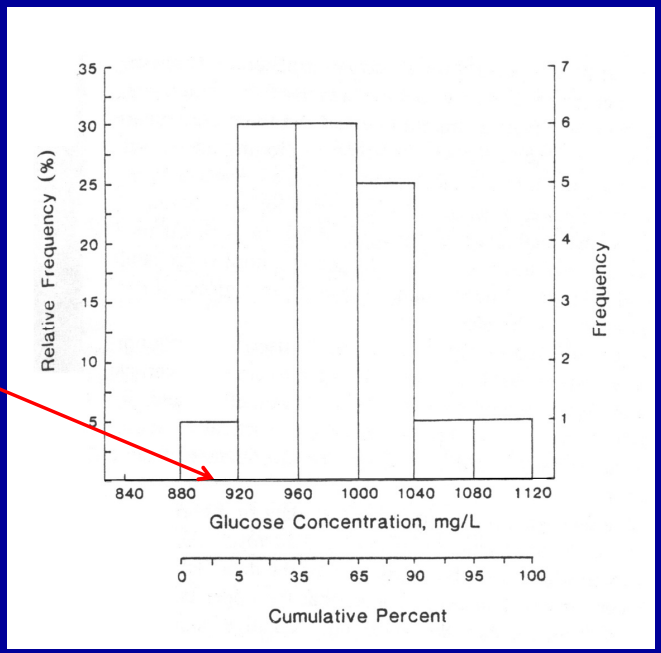
Διάγραμμα πλαισίου ή Θηκόγραμμα (Box plot)

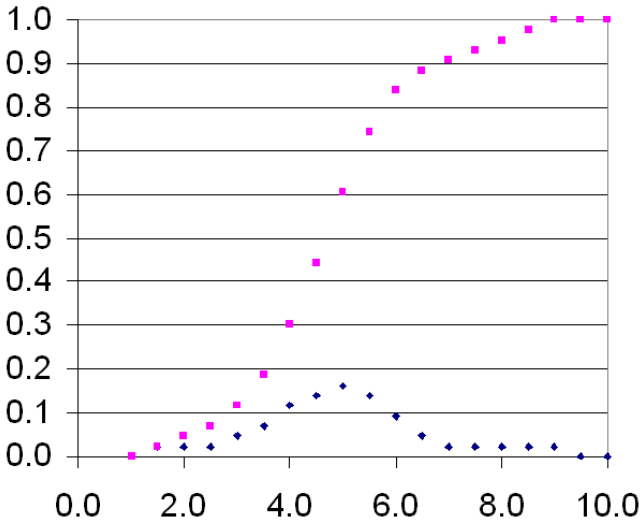


Ιστόγραμμα (Histogram)



Κλάσεις (bins)





https://plot.ly/create/histogram/

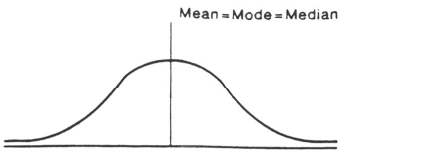
Μέση τιμή :
(mean ή average)

$$\overline{x} = \frac{1}{N} \sum^{N} x_i$$

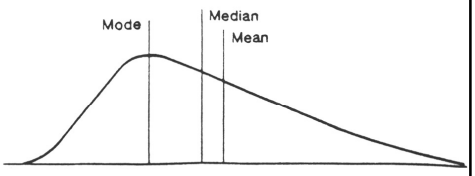
Διάμεσος :
(median)

- **Ταξινόμηση των μετρήσεων κατά μέγεθος**
- **Επιλογή της τιμής στο μέσον των μετρήσεων**

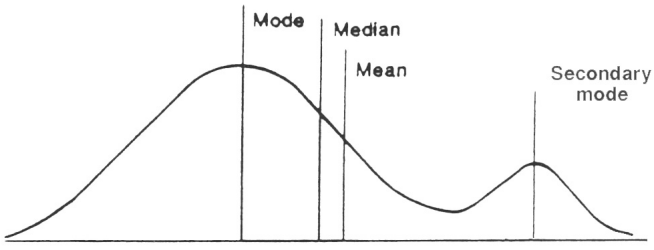
Επικρατούσα τιμή : Η συχνότερη τιμή
(mode)



Normal (gaussian) distribution, symmetric about



Nonnormal distribution



Bimodal distribution

Εκατοστημόρια ή ποσοστιαία σημεία *(percentiles)*

- **Διατάσσουμε τα δεδομένα κατά τάξη μεγέθους**
- **Το p-εκατοστημόριο είναι η τιμή που έχει p% των μετρήσεων μικρότερες από αυτήν**

Τεταρτημόρια *(quartiles)*

Ειδικά εκατοστημόρια

- **$Q_1 \rightarrow 25\%$**
- **$Q_2 \rightarrow 50\%$**
- **$Q_3 \rightarrow 75\%$**

1,2,3,6,10,12
Θέση της διαμέσου (median) (n+1)/2 = (6+1)/2 = 3.5 → Q2 = (3+6)/2 = 4.5
Θέση του 25% (Q1) = .25(n+1) = 1.75 → Q1 = (1+2)/2 = 1.5
Θέση του 75% (Q3) = .75(n+1) = 5.25 → Q3 = (10+12)/2 = 11
Έκταση (Range) = 12 -1 = 11
Μέση τιμή (Mean) = (1+2+3+6+10+12)/6 = 5.66
Κυρίαρχη τιμή (Mode) = καμία

ΜΕΤΡΑ ΤΗΣ ΔΙΑΣΠΟΡΑΣ

Έκταση ή εύρος *(range)* : $x_{max} - x_{min}$

Διακύμανση

– Διασπορά *(variance)* :

$$s^2 = \frac{\sum_1^N (x_i - \bar{x})^2}{N-1} = \frac{1}{N-1}\left[\sum_1^N x_i^2 - \frac{\left(\sum_1^N x_i\right)^2}{N}\right]$$

Πληθυσμού «σ^2» Δείγματος «s^2»

Τυπική (ή σταθερή) απόκλιση *(standard deviation)* :

$$s = \sqrt{\frac{\sum_1^N (x_i - \bar{x})^2}{N-1}}$$

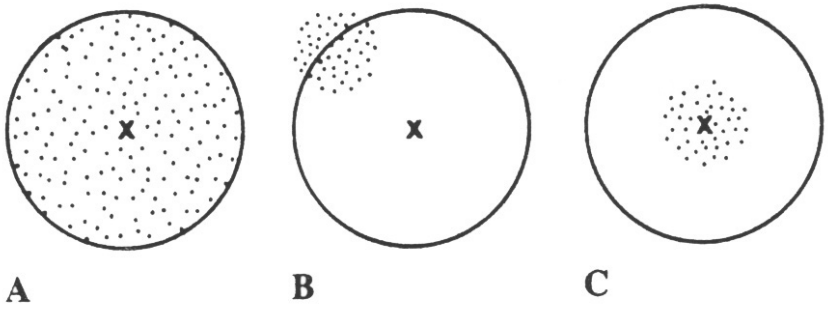
Συντελεστής διακύμανσης *(coefficient of variance)* :

ή

$$\%CV = \frac{100\% s}{\bar{x}}$$

Συντελεστής μεταβλητότητας *(coefficient of variation)*

Accurate but imprecise | inaccurate but precise | Accurate and precise



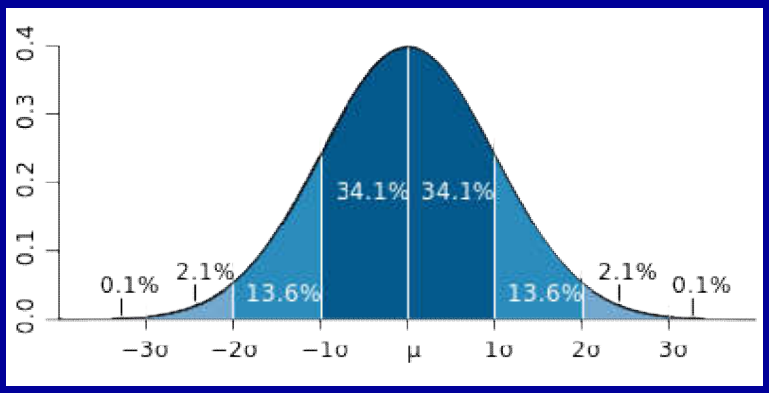
Precision: ορθότητα ή επαναληψιμότητα

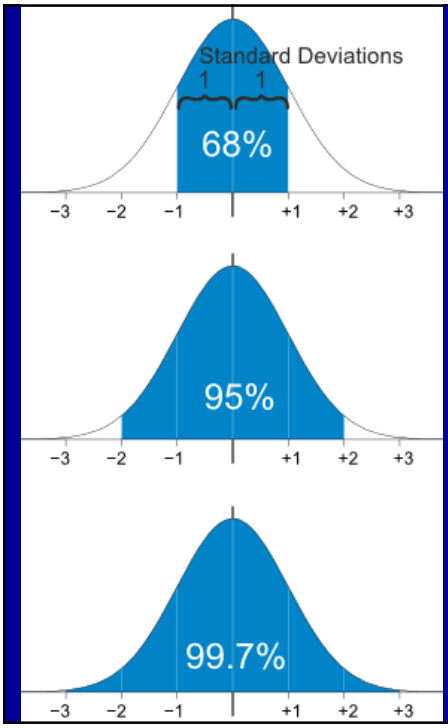
Accuracy: ακρίβεια

Κανονική κατανομή

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$N(\mu,\sigma^2)$





$N(0,1)$

$$f(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}}$$

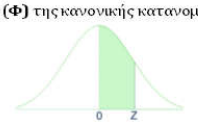
Τυποποιημένη μεταβλητή

$$Z = \frac{X-\mu}{\sigma}$$

http://davidmlane.com/hyperstat/z_table.html

http://onlinestatbook.com/2/normal_distribution/standard_normal.html

Αθροιστικές πιθανότητες (**Φ**) της κανονικής κατανομής για θετικές τιμές της Z



Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817

$\Phi(-\infty < Z \le z) = 0.5 + \Phi(0 \le Z \le z)$

Υπολογισμός πιθανοτήτων στην Κανονική κατανομή

Αν τα επίπεδα της χοληστερόλης ενός πληθυσμού σε mg/dl ακολουθούν την κατανομή N(210, 900), ποια είναι η πιθανότητα ένα άτομο που επιλέγεται τυχαία από αυτόν τον πληθυσμό να έχει επίπεδο χοληστερόλης:

Α) Μεταξύ 180 και 210 mg/dl ;

Β) Μεγαλύτερο από 225 mg/dl;

Γ) Μικρότερο από 150 mg/dl;

Δ) Μεταξύ 195 και 225 mg/dl;

$$z = \frac{X - \mu}{\sigma}$$

μ = 210mg/dl, σ = 30mg/dl

Τυποποιούμε πρώτα τα άκρα των διαστημάτων της μεταβλητής για να μπορούμε να ανατρέξουμε στον πίνακα της κανονικής κατανομής

Α) $$P(180 \le X \le 210) = P\left(\frac{180-210}{30} \le \frac{X-210}{30} \le \frac{210-210}{30}\right) = P(-1 \le z \le 0)$$

$$= P(z \le 0) - P(z \le -1) = 0,5 - 0,1587 = 0,3413$$

Β) $$P(X \ge 225) = P\left(\frac{X-210}{30} \ge \frac{225-210}{30}\right) = P(z \ge 0,5) = 1 - P(z \le 0,5) = 1 - 0,6915 = 0,3085$$

Γ) $$P(X \le 150) = P\left(\frac{X-210}{30} \le \frac{150-210}{30}\right) = P(z \le -2) = 0{,}0228$$

Δ) $$P(195 \le X \le 225) = P\left(\frac{195-210}{30} \le \frac{X-210}{30} \le \frac{225-210}{30}\right) = P(-0{,}5 \le z \le 0{,}5)$$

$$= P(z \le 0{,}5) - P(z \le -0{,}5) = 0{,}6915 - 0{,}3085 = 0{,}3830$$

Κεντρικό οριακό θεώρημα (central limit theorem)

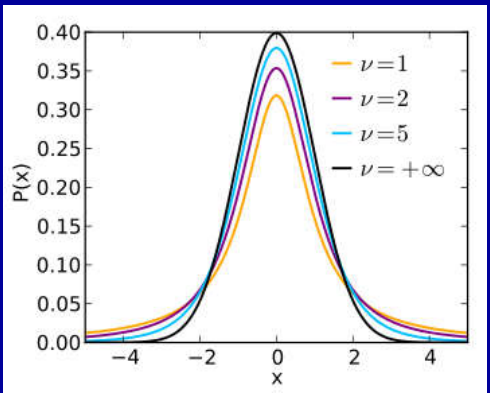
Για αρκούντως μεγάλα δείγματα από έναν πληθυσμό, οι μέσες τιμές ακολουθούν περίπου την κανονική κατανομή, ανεξάρτητα από το είδος της κατανομής του πληθυσμού. Όσο μεγαλύτερα τα δείγματα τόσο καλύτερα προσεγγίζεται η κανονική κατανομή.

Έστω $X_1, X_2, \ldots, X_n$ ανεξάρτητες μεταβλητές και $S_n = X_1 + X_2 + \ldots + X_n$

Για μεγάλα «n», η τ.μ. $$Z = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$ **ακολουθεί**

την κανονική κατανομή $N(0,1)$

Κατανομή t του Student



$$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

Η προσπάθεια της επαγωγικής στατιστικής είναι μελετώντας δείγματα να συνάγει συμπεράσματα τα οποία να γενικεύονται στον πληθυσμό.

Τι είναι πληθυσμός;

Τι είναι δείγμα;

ΕΚΤΙΜΗΤΙΚΗ

- **Εκτίμηση σε σημείο**

  Δίνουμε τους αριθμητικούς δείκτες του δείγματος ως προσεγγιστική (αβέβαιη) εκτίμηση αυτών του πληθυσμού

  εκτιμώμενο τυπικό σφάλμα της μέσης τιμής : $s_{\bar{x}} = \frac{s}{\sqrt{N}}$

  τυπικό σφάλμα : $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$

  τυπικό σφάλμα της διακύμανσης : $\sigma^2 \sqrt{\frac{2}{N-1}}$

Εκτίμηση σε διαστήματα εμπιστοσύνης ή αξιοπιστίας (*confidence intervals*)

Δίνουμε ένα διάστημα μέσα στο οποίο αναμένεται με συγκεκριμένη πιθανότητα να εμπίπτει μια παράμετρος του πληθυσμού

$$\mu = \bar{x} \pm t \cdot s_{\bar{x}}$$

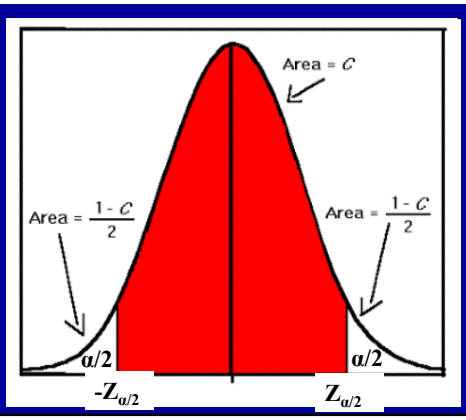
Στάθμη σημαντικότητας (σ.σ): 1-a→100(1-a)% δ.ε.

• Διάστημα εμπιστοσύνης για τη μέση τιμή (μεγάλα δείγματα)

Η μεταβλητή "Z" ακολουθεί την κανονική κατανομή $Z_{\alpha/2} = \dfrac{\bar{X}_{\alpha/2} - \mu}{\sigma/\sqrt{n}}$

Αναζητούμε την πιθανότητα για κάποιο "z" από πίνακες της τυποποιημένης κανονικής κατανομής

$$P\left[\bar{X}_n - Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) \le \mu \le \bar{X}_n + Z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right] = 1-\alpha$$



Ο δείκτης σωματικής μάζας X(kgr/m²) ενός ατόμου υπολογίζεται αν διαιρέσουμε το βάρος του με το τετράγωνο του ύψους του. Για άνδρες ηλικίας 30-40 ετών είναι γνωστό ότι $X \sim N(\mu,\sigma^2)$. Να προσδιορισθεί το 95% δ.ε. για την σωματική μάζα μ των ανδρών εάν από τυχαίο δείγμα 49 ανδρών από αυτόν τον πληθυσμό προέκυψε $\bar{X} = 25$ και $s^2=9$.

Δ.ε. 95% → $Z_{\alpha/2}=Z_{0.025}$ –(πίνακες) → 1.96

$$\bar{X}_n - 1.96\left(\frac{3}{\sqrt{49}}\right) \le \mu \le \bar{X}_n + 1.96\left(\frac{3}{\sqrt{49}}\right)$$

• Διάστημα εμπιστοσύνης για τη μέση τιμή (μικρά δείγματα)

Η μεταβλητή *"t"* ακολουθεί την κατανομή *"t"*

$$t_{n-1} = \frac{\bar{Y}_{a/2} - \mu}{s / \sqrt{n}}$$

Αναζητούμε την τιμή *"t"* από πίνακες για *"v"* β.ε και a

$$\bar{X}_n - t_{n-1}\left(\frac{s}{\sqrt{n}}\right) \le \mu \le \bar{X}_n + t_{n-1}\left(\frac{s}{\sqrt{n}}\right)$$

Από δείγμα 15 υγιών γυναικών ηλικίας 25-40 ετών, υπολογίσθηκε για την αμυλάση του ορού ότι $\bar{X} = 96$ μονάδες/100ml και $s=35$ μονάδες/100ml. Να υπολογισθεί το 90% διάστημα εμπιστοσύνης για την αληθή τιμή της μέσης τιμής μ της αμυλάσης στον πληθυσμό των υγιών γυναικών στις ίδιες ηλικίες.

Δ.ε. 90% → $t_{14,0.05}$ –(πίνακες) → 1.76

$$\bar{X}_n - 1.76\left(\frac{35}{\sqrt{15}}\right) \le \mu \le \bar{X}_n + 1.76\left(\frac{35}{\sqrt{15}}\right)$$

ΒΑΘΜΟΙ ΕΛΕΥΘΕΡΙΑΣ

Έστω ότι έχετε 7 καπέλα και τον περιορισμό ότι θέλετε να φοράτε κάθε ημέρα της εβδομάδας ένα διαφορετικό καπέλο.

1η ημέρα: Οποιοδήποτε από τα καπέλα
2η ημέρα: Διαλέγουμε 1 από τα 6 εναπομείναντα
3η ημέρα: Διαλέγουμε 1 από τα 5 εναπομείναντα
4η ημέρα:
7η ημέρα: Θα φορέσετε υποχρεωτικά το ένα που απέμεινε

Άρα έχετε ελευθερία επιλογής μόνο μέχρι την 6η ημέρα (7-1).

Οι βαθμοί ελευθερίας συχνά ορίζονται ως ο αριθμός των παρατηρήσεων (πληροφορίες) στα δεδομένα που είναι ελεύθερα να μεταβληθούν κατά τον υπολογισμό στατιστικών παραμέτρων που έχουν σχέση με μεταβλητότητα

Βαθμοί ελευθερίας (degrees of freedom) β.ε. είναι ο αριθμός των τιμών στον τελικό υπολογισμό ενός στατιστικού που είναι ελεύθερες να μεταβληθούν.

Ή αλλιώς, ο αριθμός των ανεξάρτητων πληροφοριών που συμμετέχουν στην εκτίμηση μιας παραμέτρου .

Γενικά οι β.ε. στην εκτίμηση μιας παραμέτρου είναι ίσοι με τον αριθμό των ανεξάρτητων μετρήσεων που συμμετέχουν στον υπολογισμό της μείον τον αριθμό των παραμέτρων που χρησιμοποιούνται σε προηγούμενα βήματα για τον υπολογισμό της παραμέτρου αυτής (π.χ. η διακύμανση έχει N − 1 β.ε., διότι υπολογίζεται από N τυχαίες μετρήσεις μείον τη μοναδική παράμετρο που προσδιορίσθηκε σε προηγούμενο βήμα, την μέση τιμή, και η οποία αξιοποίησε ήδη τις N τιμές).

- **Διάστημα εμπιστοσύνης για τη διασπορά (μεγάλα και μικρά δείγματα)**

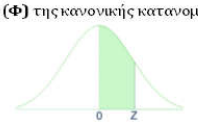
  Ακολουθεί την κατανομή *"X^2"*

  Αναζητούμε τις τιμές *"X^2"* από πίνακες για $P(X^2 > X^2_{v;a}) = a$

$$\frac{(n-1)\cdot s_n^2}{X^2_{(n-1),a/2}} < \sigma^2 < \frac{(n-1)\cdot s_n^2}{X^2_{(n-1),(1-a/2)}}$$

$$X^2 = \frac{(n-1)\cdot s^2}{\sigma^2}$$

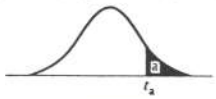
Αθροιστικές πιθανότητες **(Φ)** της κανονικής κατανομής για θετικές τιμές της Z



Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817

$$\Phi(-\infty < z \leq Z) = 0.5 + \Phi(0 \leq z \leq Z)$$

Τιμών $t_{\nu,a}$ της t_ν-κατανομής ώστε $P(t_\nu > t_{\nu,a}) = a$



B.ε	a = .10	a = .05	a = .025	a = .010	a = .005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
inf.	1.282	1.645	1.960	2.326	2.576

• **Διάστημα εμπιστοσύνης για τη διαφορά μέσων δύο πληθυσμών (μεγάλα ανεξάρτητα δείγματα)**

- **Ακολουθεί την κανονική κατανομή**
- **Αναζητούμε τις τιμές "Z" από πίνακες της τυποποιημένης κανονικής κατανομής**

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$\mu_1 - \mu_2 = \bar{X} - \bar{Y} \pm Z_{a/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

• **Διάστημα εμπιστοσύνης για τη διαφορά μέσων δύο κανονικών πληθυσμών με κοινό "σ²" (μικρά ανεξάρτητα δείγματα)**

- **Ακολουθεί την κατανομή *"t"* με $v_1 + v_2$ β.ε.**
- **Αναζητούμε την τιμή *"t"* από πίνακες για $v_1 + v_2$ και a/2**

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{s \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\mu_1 - \mu_2 = \bar{x} - \bar{y} \pm t \cdot s \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Δηλαδή χρειαζόμαστε το τυπικό σφάλμα του μέσου για το συγκεκριμένο πρόβλημα

$$s_{\hat{x}} = \frac{s}{\sqrt{N}} \qquad s_{\hat{x}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad s_{\hat{d}} = \sqrt{\frac{s_d^2}{n}}$$

•**Διάστημα εμπιστοσύνης για τη διαφορά μέσων εξαρτημένων δειγμάτων**

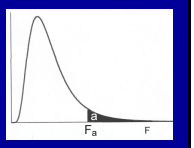
• Διάστημα εμπιστοσύνης για τη διαφορά "μέσων" ζευγαρωτών δειγμάτων

- Διαφορά *"δ"* μεταξύ των μέσων τιμών των δύο δειγμάτων
- Ακολουθεί την κατανομή *"t"* με $v = n-1$ β.ε.
- Αναζητούμε την τιμή *"t"* από πίνακες για *v* και a/2

• Διάστημα εμπιστοσύνης για τον λόγο των "διασπορών" δύο κανονικών πληθυσμών

- Ακολουθεί την κατανομή *"F"* με v_1, v_2 β.ε.
- Αναζητούμε την τιμή *"F"* από πίνακες για v_1, v_2 και a/2

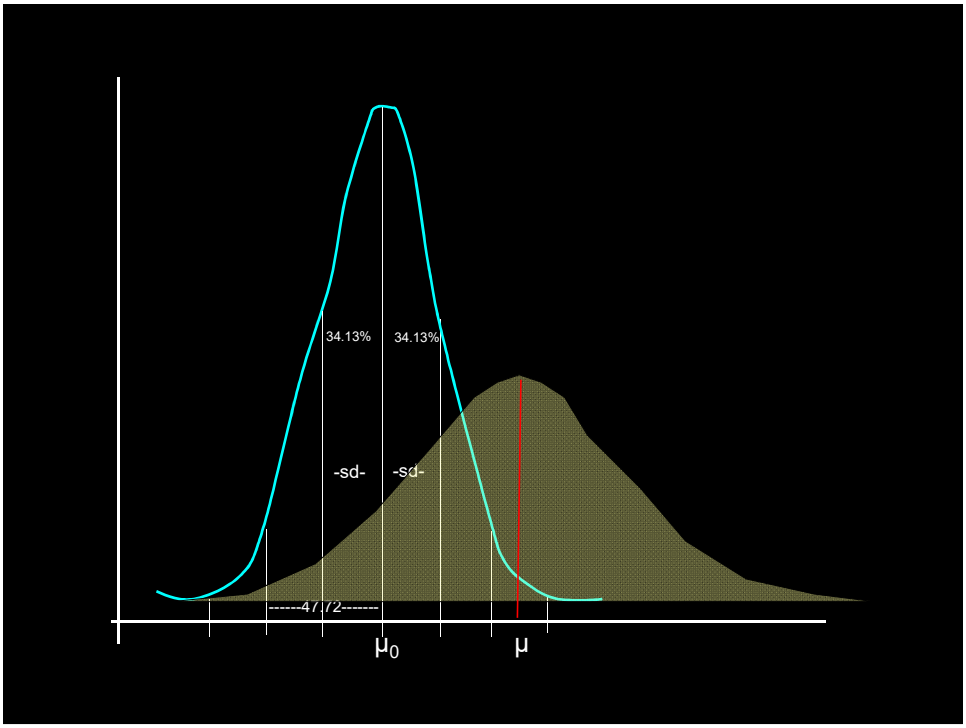
Πίνακας της κατανομής F για $P(F > F_{(\kappa-1),(n-\kappa)}) = a$



v_1: αριθμητής

v_2: παρονομαστής

$n-\kappa$ \ $\kappa-1$	1	2	3	4	5	6	7	8	9
1	161.40	199.50	215.70	224.60	230.20	234.00	236.80	238.90	240.50
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39



ΔΟΚΙΜΑΣΙΑ ΥΠΟΘΕΣΕΩΝ (ΕΛΕΓΧΟΣ ΣΗΜΑΝΤΙΚΟΤΗΤΑΣ)
Significance testing

Συγκρίνοντας δείγματα από δύο πληθυσμούς διαπιστώνουμε πάντοτε διαφορές μεταξύ των μέσων $\bar{x}$ αλλά και μεταξύ των διασπορών τους s^2.

Απηχούν πραγματικές διαφορές μεταξύ των πληθυσμών;

Μηδενική υπόθεση (null) $H_0 : \mu_1 = \mu_2$ ή $\sigma_1^2 = \sigma_2^2$

Εναλλακτική υπόθεση (alternative) $H_1 : \mu_1 \neq \mu_2$ ή $\sigma_1^2 \neq \sigma_2^2$

Σφάλμα 1ου είδους (α) : Πιθανότητα εσφαλμένης απόρριψης της H_0

Σφάλμα 2ου είδους (β) : Πιθανότητα εσφαλμένης απόρριψης της H_1

- Έλεγχος για τη μέση τιμή “μ” μεγάλου δείγματος, σ.σ. “*α*”

$$z = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} = \frac{\textit{δειγματικός μέσος - υποτιθέμενος μέσος}}{\textit{τυπικό σφάλμα του δειγματικού μέσου}}$$

Κ.Ο.Θ →Το «z» ακολουθεί την κανονική κατανομή

Για κάθε “*α*” αντιστοιχεί ένα κρίσιμο “z_α”

Η H_0 απορρίπτεται όταν $z > z_\alpha$

- Έλεγχος για διαφορά δύο μέσων τιμών (μεγάλα δείγματα), σ.σ. “*α*”

$$z = \frac{(\bar{y}_1 - \bar{y}_2) - \delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

ακολουθεί την κανονική κατανομή

Για κάθε “*α*” αντιστοιχεί ένα κρίσιμο “z_α”

Η H_0 απορρίπτεται όταν $z > z_\alpha$

Έστω ότι είσαστε δικαστής και πρέπει να κηρύξετε τον κατηγορούμενο "αθώο" (H_0) ή "ένοχο" (H_1).

H_0: αθώος (τα πειστήρια ενοχής δεν αρκούν για καταδίκη)
H_1: ένοχος

Προκειμένου να απορρίψετε την υπόθεση της αθωότητας (H_0) χρειάζεστε ικανοποιητικά ενοχοποιητικά πειστήρια.

• Εάν απαιτήσετε λίγα πειστήρια για την καταδίκη, τότε θα αυξάνονταν το ποσοστό των αθώων που θα καταδικάζονταν (σφάλμα τύπου Ι). Την ίδια στιγμή θα αυξάνονταν το ποσοστό των ενόχων που θα καταδικάζονταν (ορθή απόρριψη της μηδενικής υπόθεσης).
• Εάν απαιτήσετε πολλά πειστήρια, τότε θα αυξάνονταν το ποσοστό των αθώων που θα κηρύσσονταν αθώοι (ορθή αποδοχή της μηδενικής υπόθεσης). Την ίδια στιγμή θα αυξάνονταν το ποσοστό των ενόχων που θα κηρύσσονταν αθώοι (σφάλμα τύπου ΙΙ)

Η πιθανότητα απόρριψης της H_0, όταν είναι όντως λανθασμένη, ονομάζεται **ισχύς (power)** της δοκιμασίας. Pr (απόρριψη H_0 | H_1 ορθή)
Η ισχύς χρησιμοποιείται συχνά για τον υπολογισμό του απαιτούμενου μεγέθους του δείγματος.

$H_0 : \mu = \mu_0$

Εάν θεωρήσουμε ότι ισχύει η H_0 τότε το δείγμα με μέση τιμή **μ** προέρχεται από την ίδια κατανομή που προέρχεται και η μέση τιμή αναφοράς μ_0. Άρα αναμένεται η **μ** να βρίσκεται μέσα στο διάστημα εμπιστοσύνης για το μ_0 σε κάποια σ.σ (έστω α=0.05).

Τυποποιώντας την απόσταση δύο μέσων τιμών ($\mu - \mu_0$) διαιρούμε με το τυπικό σφάλμα και όχι με την τυπική απόκλιση, όπως θα ήταν εάν τυποποιούσαμε απλώς την τυχαία μεταβλητή. Εάν η διαφορά είναι >1.96 φορές το τυπικό σφάλμα του μέσου, τότε η μ βρίσκεται εκτός του δ.ε και έχει πιθανότητα εμφάνισης < 0.05. Στην περίπτωση αυτή μπορεί να θεωρηθεί ότι μια τέτοια μέση τιμή δεν μπορεί να αντιπροσωπεύει την ίδια κατανομή με αυτήν της μ_0. Άρα απορρίπτουμε την H_0.

- **Έλεγχος για τη μέση τιμή "μ" μικρού δείγματος,** σ.σ. "*α*"

$$t = \frac{\bar{y} - \mu_0}{s / \sqrt{n}}$$

ακολουθεί την κατανομή *"t"*

s από τις τιμές y, β.ε.= n - 1

- **Έλεγχος μέσων τιμών μ_1, μ_2 δύο δειγμάτων**

Όταν $n_1 \neq n_2$ → unpaired t-test (διαφορετικά υποκείμενα)

Δείγματα από ανεξάρτητους πληθυσμούς
Διακυμάνσεις όχι σημαντικά διαφορετικές
} Unpaired

Π.χ. Σύγκριση τιμών γλυκόζης από ασθενείς δύο διαφορετικών νοσοκομείων

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - \delta}{s \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

ακολουθεί την κατανομή *"t"*

β.ε. = $n_1 + n_2 - 2$

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

=Σταθμισμένη SD

t-test_immunoglobulin.xls

Έλεγχος σημαντικότητας για την διασπορά, σ.σ. "α"

- **Σύγκριση δειγματικής διασποράς s^2 με θεωρητική σ^2**

$$X^2 = \frac{(n-1)\cdot s^2}{\sigma^2}$$ ακολουθεί την κατανομή *"X^2"*

$$H_0 : s^2 = \sigma^2$$

$$H_1 : s^2 \neq \sigma^2$$

Για κάθε "*α*" και *"ν=n-1"* αντιστοιχεί ένα κρίσιμο $X^2_{\alpha/2;\nu}$

Η H_0 απορρίπτεται όταν $X^2 > X^2_{\alpha/2;\nu}$

- **Σύγκριση των διασπορών δύο πληθυσμών, σ.σ. "α"**

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$$

$$H_1 : \frac{\sigma_1^2}{\sigma_2^2} > 1$$

$$F = \frac{s_1^2}{s_2^2}$$ ακολουθεί την κατανομή *"F"*

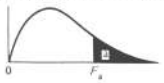
Για κάθε "*α*" και *"β.ε. ν1,ν2"* αντιστοιχεί ένα κρίσιμο $F_{\nu1,\nu2;\alpha}$

Η H_0 απορρίπτεται όταν $F \geq F_{\nu1,\nu2;\alpha}$ για $s_1^2 > s_2^2$

Folate.xls

ΠΙΝΑΚΑΣ IV

Των τιμών $F_{\nu_1,\nu_2,\alpha}$ της F-κατανομής για τις οποίες $P(F > F_{\nu_1,\nu_2,\alpha}) = \alpha$



Βαθμοί Ελευθερίας ($\alpha = .05$)

αριθμητής v_1, $s_1 > s_2$

v_2 \ v_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Όταν $n_1 = n_2$ → κατά ζεύγη (paired) t-test (ίδια υποκείμενα)

Διακυμάνσεις όχι σημαντικά διαφορετικές

$$t = \frac{(\bar{x} - \bar{y}) - \delta_0}{s_d / \sqrt{n}}$$

ακολουθεί την κατανομή *"t"*

s_d: τυπική απόκλιση διαφορών "δ", β.ε. = n -1

$$\delta = \sum_1^n \frac{x_i - y_i}{n} = \sum_1^n \frac{x_i}{n} - \sum_1^n \frac{y_i}{n} = \bar{x} - \bar{y}$$

n : αριθμός ζευγών

PO2.xls

- Έλεγχος σημαντικότητας για το «***p***» της διωνυμικής κατανομής (μεγάλα δείγματα)

  Έστω «***x***» επιτυχίες σε «***n***» δοκιμές. Συγκρίνουμε την αναλογία ***x/n*** των επιτυχιών με μια δοθείσα πιθανότητα p_0 με την βοήθεια της τυποποιημένης μεταβλητής ***Z***.

$$Z = \frac{\frac{x}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

6_17.xls

ΑΝΙΣΑ ΔΕΙΓΜΑΤΑ, ΑΝΙΣΕΣ ΔΙΣΑΠΟΡΕΣ

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}} \qquad s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

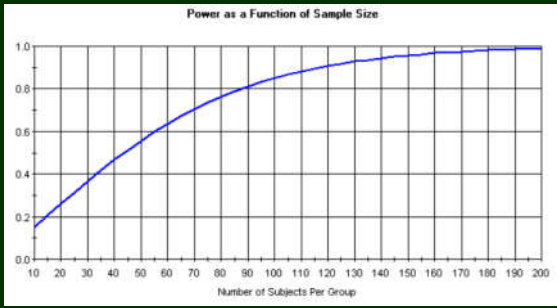
$$\text{d.f.} = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}.$$

Power analysis

Στατιστική ισχύς (Statistical Power) : P = 1 – β, P τουλάχιστον 0.80

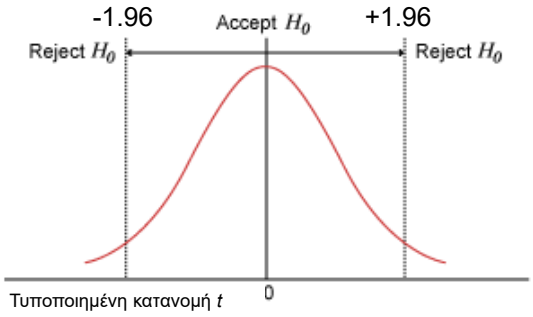
Σφάλμα 2ου είδους (β) : Πιθανότητα εσφαλμένης απόρριψης της H_1

Ποιο μέγεθος δείγματος θα μου δώσει την απαιτούμενη ακρίβεια;

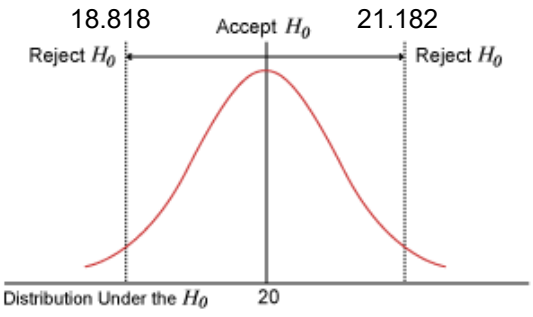


Η στατιστική ισχύς εξαρτάται από το μέγεθος της διαφοράς που μελετάμε και από το μέγεθος του δείγματος.

Εάν η στατιστική ισχύς είναι υψηλή, η πιθανότητα να κάνουμε σφάλμα 2ου είδους (να συμπεράνουμε ότι δεν υπάρχει διαφορά ενώ υπάρχει) είναι μικρή.

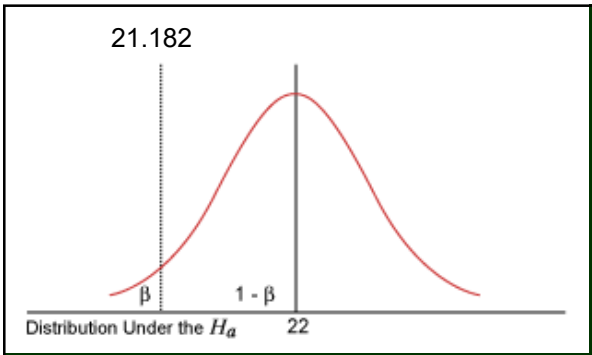


n = 44
Πληθυσμ. μέσος μ_0 = 20
Δειγματικός μέσος μ = 22
Τυπική απόκλιση s = 4
Τυπ.σφάλμα se = 0.603
z = 3.317
z_{cr} = 1.96



Διάστημα εμπιστοσύνης εάν ισχύει η H_0 :

Για z = +/-1.96 οι αντίστοιχες τιμές της x σε υποθετική κατανομή με κέντρο το μ_0 = 20 θα είναι = 20 +/- 0.603(1.96) = 21.182 and 20 – 0.603(1.96) = 18.818.



Πόση είναι η πιθανότητα να είναι μικρότερη από -21.182 στην εναλλακτική κατανομή; Αυτή είναι η πιθανότητα β να αποδεχθούμε την Ho ενώ είναι λάθος.
Ποια τιμή z αντιστοιχεί στην τιμή 21.182 στην εναλλακτική κατανομή;

$$z = \frac{21.182 - 22}{0.603} = -1.356 \rightarrow P = 0.4115 + 0.5 = 0.9115$$

ΑΝΑΛΥΣΗ ΤΗΣ ΔΙΑΚΥΜΑΝΣΗΣ (ANOVA)

Σύγκριση περισσοτέρων των δύο ανεξαρτήτων δειγμάτων από κανονικούς πληθυσμούς με «ίδια» διακύμανση

Συνολική εκτίμηση για το αν οι μέσοι όλων των δειγμάτων είναι μεταξύ τους ίσοι ή αν τουλάχιστον ένας από αυτούς διαφέρει.

Ελέγχουμε αν υπάρχει σημαντική διαφορά μεταξύ της διασποράς των μέσων τιμών και της συνολικής διασποράς των δειγμάτων.

$$s_\alpha^2 = \sum_1^\kappa n_i (\bar{x}_i - \bar{x})^2 \qquad s_\upsilon^2 = \sum_1^\kappa (n_i - 1) s_i^2$$

$$F = \frac{\text{τετράγωνα μεταξύ δειγμάτων}}{\text{τετράγωνα εντός δειγμάτων}} = \frac{s_\alpha^2/(\kappa - 1)}{s_\upsilon^2/(n - \kappa)}$$

$n = n_1 + n_2 + \ldots + n_\kappa$

κ: πλήθος δειγμάτων

Hb_ANOVA.xls

H_0: $\bar{x}_1 = \bar{x}_2 = \bar{x}_3 = \cdots = \bar{x}_\kappa$

H_1: Τουλάχιστον ένας από τους δειγματικούς μέσους δεν είναι ίσος με κάποιον άλλον

ΠΙΝΑΚΑΣ ΑΝΑΛΥΣΗΣ ΔΙΑΣΠΟΡΑΣ

$VR > F_{(\kappa-1),(n-\kappa),\alpha}$

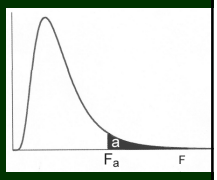
Μεταβολή	Αθρ. Τετραγ.	Β.ε.	Μέση μεταβλητ.	F
Μεταξύ δειγμάτων	s_α^2	κ-1	$s_\alpha^2/(\kappa-1)$	$\dfrac{s_\alpha^2/(\kappa-1)}{s_\upsilon^2/(n-\kappa)}$
Εντός δειγμάτων	s_υ^2	n-κ	$s_\upsilon^2/(n-\kappa)$	

Προϋπόθεση : διασπορές των δειγμάτων ίσες.

09__3_Example.xls

Η H_0 απορρίπτεται εάν $F > F_{(\kappa-1),(n-\kappa)}$

Πίνακας της κατανομής F για $P(F > F_{(\kappa-1),(n-\kappa)}) = a$



n-κ \ κ-1	1	2	3	4	5	6	7	8	9
1	161.40	199.50	215.70	224.60	230.20	234.00	236.80	238.90	240.50
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39

ΕΛΕΓΧΟΣ ΚΑΝΟΝΙΚΟΤΗΤΑΣ

Προϋπόθεση όλων των «παραμετρικών δοκιμασιών» που αναφέρθηκαν είναι η «κανονικότητα» των δεδομένων. Πώς ελέγχουμε την κανονικότητα;

- Γράφημα συσχέτισης των ποσοστιαίων σημείων (quantiles) των δεδομένων με αυτά της κανονικής κατανομής (**QQ-plot**).

  Ευθεία γραμμή → κανονικότητα

- Έλεγχος **λοξότητας (skewness)** και **κύρτωσης (kurtosis)**

$$\text{Συντ. λοξότητας} = \frac{\sum_{1}^{N}(x_i - \bar{x})^3}{(N-1)s^3}$$

μετρά την ασυμμετρία της κατανομής

>0→ η δεξιά ουρά μεγαλύτερη από την αριστερή

$$\text{Συντ. κύρτωσης} = \frac{\sum_{1}^{N}(x_i - \bar{x})^4}{(N-1)s^4}$$

μετρά το πόσο πιο απότομη ή πιο επίπεδη είναι η κατανομή σε σχέση με την κανονική.

>0→ απότομη στο κέντρο με μακρές ουρές, <0→ επίπεδη στο κέντρο με μικρές ουρές

Αποδεχόμαστε ότι τα δεδομένα ακολουθούν την κανονική κατανομή, αν οι συντελεστές ασυμμετρίας και κύρτωσης είναι στο διάστημα (-1,1).

P-value

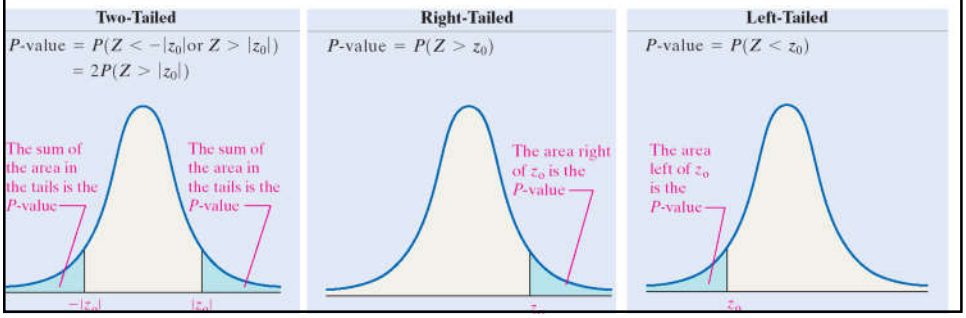
Στις δοκιμασίες σημαντικότητας η ***p*-value** είναι η πιθανότητα του να πάρει το στατιστικό *z* ή *t* τουλάχιστον τόσο ακραίες τιμές όσο αυτή που παρατηρήθηκε, θεωρώντας ότι ισχύει η μηδενική υπόθεση. Ο μελετητής απορρίπτει συνήθως την μηδενική υπόθεση όταν η *p*-value βρίσκεται να είναι μικρότερη από την στάθμη σημαντικότητας ***α*** που επέλεξε, συνήθως **0.05** ή **0.01**. Μια τόσο μικρή *p*-value δείχνει ότι το παρατηρηθέν αποτέλεσμα θα ήταν πολύ απίθανο να συμβεί εάν ισχύει η μηδενική υπόθεση (δηλαδή ότι η παρατηρηθείσα σχέση είναι πολύ απίθανο να είναι αποτέλεσμα καθαρής τύχης.)

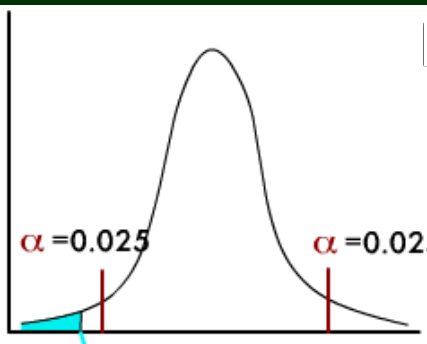
Υπολογισμός της *p*-value

Υπολογίζουμε το στατιστικό $z = \dfrac{\bar{x} - \mu}{s / \sqrt{n}}$

Ανατρέχουμε στους πίνακες του **z** (διαφάνειες) και βρίσκουμε την πιθανότητα **Φ** στην οποία αντιστοιχεί.

***p*-value = 2×(0.5 – Φ)** για αμφίπλευρη δοκιμασία σημαντικότητας.





s = 5.67, n = 100

H_o: $\mu = 35$

H_a: $\mu \neq 35$

Significance Level: $\alpha = 0.05$

Sample Data: $\bar{x} = 33.6$

test statistic: $z = -2.47$

P-value = 0.0068 times 2 (for a 2-sided test) = 0.0136

Conclusion: since the *P*-value of 0.0136 is less than the significance level of $\alpha = 0.05$, we reject the null hypothesis.

$P(-\infty \leq z \leq -2.47) = 2x[0.5 - P(-2.47 \leq z \leq 0)] = 2x(0.5 - 0.49324) = 2x0.0068 = 0.0136 < 0.05$

ΠΟΙΟΤΙΚΑ ΔΕΔΟΜΕΝΑ

Κατηγορίας (nominal) - Διάταξης (ordinal)

Ποσοστά - Σχετικές Συχνότητες - Αναλογίες

- $$p = \frac{\text{Πλήθος εμφανίσεων κάποιου φαινομένου}}{\text{Σύνολο παρατηρήσεων}}$$

- **Σε μεγάλα δείγματα ακολουθεί κανονική κατανομή**

  Δηλ. όταν *np(1-p)* $\geq$ *10*

ΣΥΓΚΡΙΣΗ ΑΝΑΛΟΓΙΑΣ ΕΝΟΣ ΔΕΙΓΜΑΤΟΣ ΜΕ ΓΝΩΣΤΗ ΑΝΑΛΟΓΙΑ

(Η τ.μ. παίρνει δύο τιμές)

p = x/n
H_0: p = p_0
H_1: p ≠ p_0

το p προκύπτει από τον ίδιο πληθυσμό που αναφέρεται το p_0

$$z = \frac{p - p_0}{\sqrt{p_0(1-p_0)/n}}$$

Διάστημα εμπιστοσύνης

$$= \mathbf{p} \pm \mathbf{z}_{\mathbf{a}/2}\sqrt{\frac{\mathbf{p}(1-\mathbf{p})}{\mathbf{n}}}$$

ΣΥΓΚΡΙΣΗ ΑΝΑΛΟΓΙΩΝ ΑΠΟ ΔΥΟ ΠΛΗΘΥΣΜΟΥΣ

$H_0: (p_1 - p_2) = D_0$, $p_1 = x_1/n_1$, $p_2 = x_2/n_2$

$H_1: (p_1 - p_2) \neq D_0$

Περίπτωση Α: $D_0 = 0 \rightarrow H_0: p_1 = p_2$

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$p = \frac{x_1 + x_2}{n_1 + n_2}$$

Περίπτωση Β: $D_0 \neq 0 \rightarrow H_0: p_1 - p_2 = D_0$

$$z = \frac{(p_1 - p_2) - D_0}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

ΣΥΓΚΡΙΣΗ ΑΝΑΛΟΓΙΩΝ ΔΥΟ ΔΕΙΓΜΑΤΩΝ ΚΑΤΑ ΖΕΥΓΗ
Ανάλυση "χ^2" (Μη παραμετρικός έλεγχος)

- Η τ.μ. παίρνει δύο τιμές (Κατηγορίες 1, 2)
  - Πίνακες 2x2

Ομάδα	Κατηγορία 1	Κατηγορία 2	Σύνολο
1 2	π_{11} π_{21}	π_{12} π_{22}	$\pi_{11}+\pi_{12}=N_1$ $\pi_{21}+\pi_{22}=N_2$
Σύνολο	$\pi_{11}+\pi_{21}$	$\pi_{12}+\pi_{22}$	$N=N_1+N_2$

- Αναμενόμενες τιμές κατ. 1 ομάδας 1, 2

$\theta_{11} = (\pi_{11}+\pi_{12})*(\pi_{11}+\pi_{21})/N$, $\theta_{21} = (\pi_{21}+\pi_{22})*(\pi_{11}+\pi_{21})/N$

- Αναμενόμενες τιμές κατ. 2 ομάδας 1, 2

$\theta_{12} = (\pi_{11}+\pi_{12})*(\pi_{12}+\pi_{22})/N$, $\theta_{22} = (\pi_{21}+\pi_{22})*(\pi_{12}+\pi_{22})/N$

Εξετάζουμε την τιμή της μεταβλητής :

$$X^2 = \frac{(\pi_{11}-\theta_{11})^2}{\theta_{11}} + \frac{(\pi_{12}-\theta_{12})^2}{\theta_{12}} + \frac{(\pi_{21}-\theta_{21})^2}{\theta_{21}} + \frac{(\pi_{22}-\theta_{22})^2}{\theta_{22}}$$

ή $$X^2 = \frac{\pi_{11}^2}{\theta_{11}} + \frac{\pi_{12}^2}{\theta_{12}} + \frac{\pi_{21}^2}{\theta_{21}} + \frac{\pi_{22}^2}{\theta_{22}} - N$$

η οποία ακολουθεί κατανομή "χ^2" με β.ε.=1,

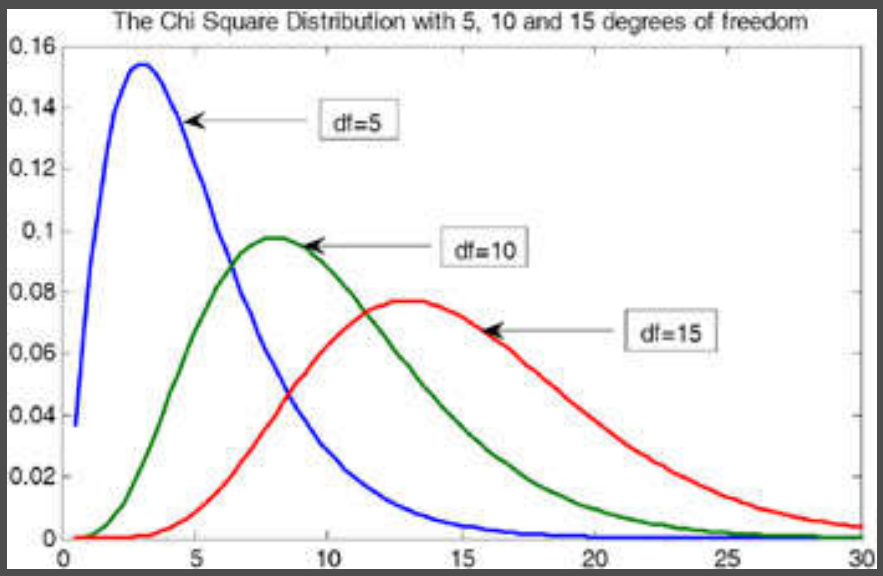
όταν οι αναμενόμενες αναλογίες δεν είναι < 5

Η χ^2 εξετάζει αμφίπλευρα, αλλά στον πίνακα αναζητούμε για σ.σ=0.05 και όχι 0.05/2.

x_square_schmerzlind.xls x_square_operation.xls

Για 2 ομάδες και 2 κατηγορίες μπορεί επίσης να χρησιμοποιηθεί ο τύπος:

$$X^2 = \frac{(\pi_{11}\cdot\pi_{22}-\pi_{12}\cdot\pi_{21})^2(\pi_{11}+\pi_{12}+\pi_{21}\cdot\pi_{22})}{(\pi_{11}+\pi_{12})\cdot(\pi_{21}+\pi_{22})\cdot(\pi_{12}+\pi_{22})\cdot(\pi_{12}+\pi_{21})}$$



ΣΥΚΡΙΣΗ ΑΝΑΛΟΓΙΩΝ S ΔΕΙΓΜΑΤΩΝ ΜΕ K ΚΑΤΗΓΟΡΙΕΣ
Ανάλυση "χ^2"

Πίνακες Συνάφειας (contingency tables)

	Κατηγορία "j"				
Ομάδα "i"	1	2	.	k	Σύνολο γραμμών
1	π_{11}	π_{12}	.	π_{1k}	$n_{1k}=\Sigma n_{1j}$
2	π_{21}	π_{22}	.	π_{2k}	$n_{2k}=\Sigma n_{2j}$
.	.	.	.	.	.
s	π_{s1}	π_{s2}		π_{sk}	$n_{sk}=\Sigma n_{sj}$
Σύνολο στηλών	$n_{s1}=\Sigma n_{i1}$	$n_{s2}=\Sigma n_{i2}$	.	$n_{sk}=\Sigma n_{ik}$	$N=\Sigma\Sigma n_{ij}$

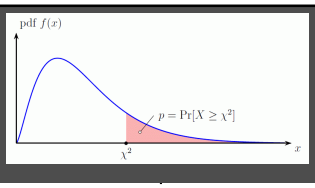
$$\theta_{ij} = \frac{(\sum_{i=1}^{s} n_{ij}) \cdot (\sum_{j=1}^{k} n_{ij})}{N} \qquad X^2 = \sum_{i=1}^{s}\sum_{j=1}^{k} \frac{(\pi_{ij} - \theta_{ij})^2}{\theta_{ij}}$$

ή απλούστερα

$$X^2 = \sum_{i=1}^{s}\sum_{j=1}^{k} \left[\frac{\pi_{ij}^2}{\theta_{ij}} - N \right]$$

β.ε. = (*s*-1)×(*k*-1)

Fisher Exact Test : Εναλλακτική για την χ2 για μικρούς αριθμούς



Κρίσιμες τιμές του χ^2

df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319

ΜΗ ΠΑΡΑΜΕΤΡΙΚΕΣ ΔΟΚΙΜΑΣΙΕΣ

Δεν προαπαιτούν να ακολουθούν οι πληθυσμοί κάποια κατανομή.

Ενδιαφέρει μόνο η τάξη και όχι η τιμή της τυχαίας μεταβλητής.

- **Δοκιμασία Kolmogorov –Smirnov**
  (Καλής προσαρμογής, Ομογένειας)
- **Κριτήριο των Ροών ή Wald-Wolfowitz**
  (Δοκιμασία τυχαιότητας)
- **Δοκιμασία προσήμου (sign test)**
- **Αθροίσματα τάξεων (rank sum test)**
  - **Δοκιμασία Wilcoxon**
  - **Δοκιμασία Mann-Whithey**
  - **Δοκιμασία Kruskal-Wallis**

ΔΟΚΙΜΑΣΙΑ ΟΜΟΓΕΝΕΙΑΣ (Kolmogorov – Smirnov)

- **Ελέγχει αν δύο δείγματα προέρχονται από την ίδια κατανομή.**

H_0= Τα δύο δείγματα προέρχονται από την ίδια κατανομή

H_1= Τα δύο δείγματα προέρχονται διαφορικές κατανομές

- **Χρησιμοποιεί τον δείκτη D = max $|F_1(x) - F_2(x)|$**

  $F_1(x)$ και $F_2(x)$ είναι οι αθροιστικές συχνότητες των δύο δειγμάτων

- **Η μηδενική υπόθεση απορρίπτεται, όταν το $D \geq D_{a,n1,n2}$ από τον πίνακα VII.**

Για m,n > 15

α=0.05 $$D_{a,m,n} = 1.36\sqrt{\frac{m+n}{m \cdot n}}$$ α=0.01 $$D_{a,m,n} = 1.63\sqrt{\frac{m+n}{m \cdot n}}$$

10__2_Example.xls

ΔΟΚΙΜΑΣΙΑ ΟΜΟΓΕΝΕΙΑΣ (Kolmogorov – Smirnov)

- Για ένα δείγμα που ελέγχεται εάν είναι κανονικό:

H_0= Το δείγμα προέρχονται από κανονικό πληθυσμό

H_1= Το δείγμα δεν προέρχονται από κανονικό πληθυσμό

- Χρησιμοποιεί τον δείκτη $D = \max |F_1(x) - F_N(x)|$

  $F_1(x)$ και $F_N(x)$ είναι οι αθροιστικές συχνότητες του δείγματος και της κανονικής κατανομής

- Η μηδενική υπόθεση απορρίπτεται, όταν το $D \geq D_{a,n}$ από τον πίνακα VI.

Για n > 100

α=0.05 $D_{a,n} = \frac{1.36}{\sqrt{n}}$ α=0.01 $D_{a,n} = \frac{1.63}{\sqrt{n}}$

ΔΟΚΙΜΑΣΙΑ ΤΩΝ ΡΟΩΝ ή WALD-WOLFOWITZ

(Δοκιμασία τυχαιότητας)

Εξετάζει αν μια ακολουθία παρατηρήσεων είναι τυχαία ή όχι. Π.χ.

ΕΕ ΑΑ ΕΕ ΑΑΑ Ε ΑΑΑ ΕΕΕΕ ΑΑΑ

Κάθε διαδοχή ομοίων συμβόλων λέγεται «ροή».

Το πλήθος των συμβόλων μιας ροής λέγεται «μήκος ροής».

Στο παράδειγμα έχουμε 8 ροές

n_1=9 Ε, n_2=11 Α

Εάν $n_1, n_2 \geq 10$ τότε το πλήθος των ροών U ακολουθεί την

$N(\mu_u, \sigma^2)$ με $\mu_u = \frac{2n_1n_2}{n_1+n_2} + 1$ και $\sigma_u^2 = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}$

$$Z = \frac{U - \mu_u}{\sigma_u}$$

Alignment.xls

ΔΟΚΙΜΑΣΙΑ ΠΡΟΣΗΜΟΥ (Sign test)

- **Ανάλογη προς τη δοκιμασία *"t"***
- **Χρησιμοποιεί τον διάμεσο (median)**
- **Ζευγαρωτές παρατηρήσεις**
- **Πλήθος μη μηδενικών διαφορών > 5**

H_0: $M_1 = M_2$

Η H_0 απορρίπτεται, όταν το πλήθος των αρνητικών *(N_-)* και θετικών *(N_+)* διαφορών είναι "άνισο". Αυτό συμβαίνει όταν το $N_m = \min(N_-, N_+) \leq$ της μικρότερης τιμής του διαστήματος εμπιστοσύνης που προβλέπεται από πίνακες της διωνυμικής κατανομής για το πλήθος των μη μηδενικών διαφορών και για συγκεκριμένο "a".

Sign_test.xls

Όρια εμπιστοσύνης για το N_p (διωνυμική κατανομή), p=0.05; *N*=0 έως 99

N	0	1	2	3	4	5	6	7	8	9
0	-	-	-	-	-	-	0-6	0-7	0-8	1-8
10	1-9	1-10	2-10	2-11	2-12	3-12	3-13	4-13	4-14	4-15
20	5-15	5-16	5-17	6-17	6-18	7-18	7-19	7-20	8-20	8-21
30	9-21	9-22	9-23	10-23	10-24	11-24	11-25	12-25	12-26	12-27
40	13-27	13-28	14-28	14-29	15-29	15-30	15-31	16-31	16-32	17-32
50	17-33	18-33	18-34	18-35	19-35	19-36	20-36	20-37	21-37	21-38
60	21-39	22-39	22-40	23-40	23-41	24-41	21-42	25-42	25-43	25-44
70	26-44	26-45	27-45	27-46	28-46	28-47	28-48	29-48	29-49	30-49
80	30-50	31-50	31-51	32-51	32-52	32-53	33-53	33-54	34-54	34-55
90	35-55	35-56	36-56	36-57	37-57	37-58	37-59	38-59	38-60	39-60

***N*: πλήθος μη μηδενικών διαφορών**

Εναλλακτικά υπολογίζουμε το $Z = \dfrac{N_+ - N/2}{\sqrt{N}/2}$

ΔΟΚΙΜΑΣΙΑ ΑΘΡΟΙΣΜΑΤΩΝ ΤΑΞΕΩΝ ή ΘΕΣΕΩΝ ή ΒΑΘΜΩΝ (rank sums) ΚΑΤΑ WILCOXON

- Ζευγαρωτές παρατηρήσεις (N_1=N_2)
- H_0: $M_1 = M_2$
- Κατατάσσουμε τις απόλυτες τιμές των διαφορών κατά αύξουσα σειρά.
- Η σειρά εμφάνισης των διαφορών αποτελεί την τάξη τους.
- Γράφουμε δίπλα σε κάθε διαφορά την τάξη της και το πρόσημό της.
- Αθροίζουμε τις τάξεις των θετικών διαφορών (T_+)
- Αθροίζουμε τις τάξεις των αρνητικών διαφορών (T_-)

Στο βιβλίο T → W

- Η μηδενική υπόθεση απορρίπτεται όταν:

  T = min {T_+,T_-} ≤ T_c. (T_c από τον πίνακα XI, n=πλήθος μη μηδενικών παρατηρήσεων)

-

- Για πλ. διαφ. N > 25 η T ακολουθεί κανονική κατανομή.

ή εάν N(N+1)/2 > 20

$$Z = \frac{T - \mu_T}{\sigma_T}$$

$$\mu_T = \frac{N(N+1)}{4} \qquad \sigma_T = \sqrt{\frac{N(N+1)(2N+1)}{24}}$$

Wilc_Sign_Rank_Sum.xls

10_12_Example.xls

ΠΙΝΑΚΑΣ XI

Πίνακας κρίσιμων τιμών του T στο test **Wilcoxon** για ζευγαρωτές παρατηρήσεις

Μονόπλευρο test a	Δίπλευρο test a	n = 5	n = 6	n = 7	n = 8	n = 9	n = 10	n = 11	n = 12	n = 13	n = 14	n = 15	n = 16
.05	.10	1	2	4	6	8	11	14	17	21	26	30	36
.025	.05		1	2	4	6	8	11	14	17	21	25	30
.01	.02			0	2	3	5	7	10	13	16	20	24
.005	.01				0	2	3	5	7	10	13	16	19

		n = 17	n = 18	n = 19	n = 20	n = 21	n = 22	n = 23	n = 24	n = 25	n = 26	n = 27	n = 28
.05	.10	41	47	54	60	68	75	83	92	101	110	120	130
.025	.05	35	40	46	52	59	66	73	81	90	98	107	117
.01	.02	28	33	38	43	49	56	62	69	77	85	93	102
.005	.01	23	28	32	37	43	49	55	61	68	76	84	92

		n = 29	n = 30	n = 31	n = 32	n = 33	n = 34	n = 35	n = 36	n = 37	n = 38	n = 39
.05	.10	141	152	163	175	188	201	214	228	242	256	271
.025	.05	127	137	148	159	171	183	195	208	222	235	250
.01	.02	111	120	130	141	151	162	174	186	198	211	224
.005	.01	100	109	118	128	138	149	160	171	183	195	208

		n = 40	n = 41	n = 42	n = 43	n = 44	n = 45	n = 46	n = 47	n = 48	n = 49	n = 50
.05	.10	287	303	319	336	353	371	389	408	427	446	466
.025	.05	264	279	295	311	327	344	361	379	397	415	434
.01	.02	238	252	267	281	297	313	329	345	362	380	398
.005	.01	221	234	248	262	277	292	307	323	339	356	373

Όρια ασφαλείας του δείκτη για τη δοκιμασία προσήμου αθροιστικών τάξεων κατά Wilcoxon

	a=0.05	a=0.025
Πλήθος διαφορών≠0	**$T_{0.95}$**	**$T_{0.975}$**
5	15	-
6	17	21
7	22	24
8	26	30
9	29	35
10	35	39
11	40	46
12	44	52
13	49	57
14	55	63
15	60	70
16	66	76
17	71	83
18	77	91
19	84	98
20	90	106

Για n > 20

$$T_p = z_p \sqrt{\frac{n(n+1)(2n+1)}{6}}$$

z_p : το p εκατοστιαίο σημείο της κανονικής κατανομής

ΔΟΚΙΜΑΣΙΑ ΑΘΡΟΙΣΜΑΤΩΝ ΤΑΞΕΩΝ (rank sums) ΚΑΤΑ MANN-WHITNEY (U-test)

- Δείγματα διαφορετικού μεγέθους N_1,N_2
- Διατάσσουμε τις παρατηρήσεις των δύο δειγμάτων σε αύξουσα σειρά, σαν να ανήκαν στον ίδιο πληθυσμό.
- Σημειώνουμε την τάξη κάθε παρατήρησης.
- Υπολογίζουμε τα αθροίσματα T_1 και T_2 των τάξεων για κάθε δείγμα.
- $T = \min \{T_1, T_2\}$
- Εξετάζουμε, αν το T εμπίπτει στα όρια αποδοχής που βρίσκουμε σε πίνακες για N_1, N_2 και συγκεκριμένη σ.σ.

Στον πίνακα που ακολουθεί το N_1 είναι το N του δείγματος με το μικρότερο T.

Acceptance region for the rank sum T (Mann-Whitney-Wilcoxon 2 sample test), $p = 0.05$

N_1 / N_2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
2	—	—	—	—	—	—	—	36-52	45-63	55-75	66-88	79-101	92-116	106-132	121-149
3	—	—	—	—	15-30	22-38	29-48	38-58	47-70	58-82	69-96	82-110	95-126	110-142	125-160
4	—	—	—	10-26	16-34	23-43	31-53	40-64	49-77	60-90	72-104	85-119	99-135	114-152	130-170
5	—	—	6-21	11-29	17-38	24-48	33-58	42-70	52-83	63-97	75-112	89-127	103-144	118-162	134-181
6	—	—	7-23	12-32	18-42	26-52	34-64	44-76	55-89	64-104	79-119	92-136	107-153	122-172	139-191
7	—	—	7-26	13-35	20-45	27-57	36-69	46-82	57-96	69-111	82-127	96-144	111-162	127-181	144-201
8	—	3-19	8-28	14-38	21-49	29-61	38-74	49-87	60-102	72-118	85-135	100-152	115-171	131-191	149-211
9	—	3-21	8-31	14-42	22-53	31-65	40-79	51-93	62-109	75-125	89-142	104-160	119-180	136-200	154-221
10	—	3-23	9-33	15-45	23-57	32-70	42-84	53-99	65-115	78-132	92-150	107-169	124-188	141-209	159-231
11	—	3-25	9-36	16-48	24-61	34-74	44-89	55-105	68-121	81-139	96-157	111-177	128-197	145-219	164-241
12	—	4-26	10-38	17-51	26-64	35-79	46-94	58-110	71-127	84-146	99-165	115-185	132-206	150-228	169-251
13	—	4-28	10-41	18-54	27-68	37-83	48-99	60-116	73-134	88-152	103-172	119-193	136-215	155-237	174-261
14	—	4-30	11-43	19-57	28-72	38-88	50-104	62-122	76-140	91-159	106-180	123-201	141-223	160-246	179-271
15	—	4-32	11-46	20-60	29-76	40-92	52-109	65-127	79-146	94-166	110-187	127-209	145-232	164-256	184-281

Εάν N_1, $N_2 > 20$

$U = T - m(m+1)/2$, $\mu_T = mn/2$, $\sigma^2_T = mn(m+n+1)/12$, $Z = \dfrac{U - \mu_T}{\sigma_T}$

ΔΟΚΙΜΑΣΙΑ KRUSKAL - WALLIS

- Μη παραμετρικό ανάλογο της ANOVA
- H_0: Τα δείγματα είναι ομογενή (οι μέσες τάξεις των κ δειγμάτων δεν διαφέρουν σημαντικά μεταξύ τους.)
- Η τ.μ. πρέπει να έχει συνεχή κατανομή και να είναι διατάξιμη.
- Χρησιμοποιούμε τις τάξεις των παρατηρήσεων και τις υποβάλουμε σε "ανάλυση των διακυμάνσεων".
- Υπολογίζουμε τον δείκτη *"H"* ο οποίος ακολουθεί την κατανομή "χ^2"

$$H = \frac{12}{n(n+1)} \cdot \sum_{j=1}^{\kappa} \frac{R_j^2}{n_j} - 3(n+1)$$

κ: πλήθος δειγμάτων, n_j: πλήθος του δείγματος "j", $n=\Sigma n_j$

R_j: άθροισμα των τάξεων στο δείγμα "j"

Για κ=3 και n_j μέχρι 5 ανατρέχουμε στον πίνακα XII.

Στην περίπτωση πολλαπλότητας «*μ*» των τιμών διορθώνουμε την τιμή του Η διαιρώντας με τον αριθμό C, που λαμβάνει υπόψη το άθροισμα των κύβων των βαθμών πολλαπλότητας καθώς και το άθροισμά των:

$$H = \frac{\frac{12}{n(n+1)} \cdot \sum_{j=1}^{\kappa} \frac{R_j^2}{n_j} - 3(n+1)}{C}$$

$$C = 1 - \frac{\left(\mu_1^3 + \mu_2^3 + \ldots + \mu_\rho^3\right) - \left(\mu_1 + \mu_2 + \ldots + \mu_\rho\right)}{n^3 - n}$$

Table The Kruskal-Wallis test

Critical region : H $\geq$ tabulated value

K = 3					K = 4						K = 5						
Sample Sizes			α = 0.05	α = 0.01	Sample sizes				α = 0.05	α = 0.01	Sample sizes					α = 0.05	α = 0.01
2	2	2	-	-	2	2	1	1	-	-	2	2	1	1	1	-	-
3	2	1	-	-	2	2	2	1	5.679	-	2	2	2	1	1	6.750	-
3	2	2	4.714	-	2	2	2	2	6.167	6.667	2	2	2	2	1	7.133	7.533
3	3	1	5.143	-	3	1	1	1	-	-	2	2	2	2	2	7.418	8.291
3	3	2	5.361	-	3	2	1	1	-	-	3	1	1	1	1	-	-
3	3	3	5.600	7.200	3	2	2	1	5.833	-	3	2	1	1	1	6.583	-
4	2	1	-	-	3	2	2	2	6.333	7.133	3	2	2	1	1	6.800	7.600
4	2	2	5.333	-	3	3	1	1	6.333	-	3	2	2	2	1	7.309	8.127
4	3	1	5.208	-	3	3	2	1	6.244	7.200	3	2	2	2	2	7.682	8.682
4	3	2	5.444	6.444	3	3	2	2	6.527	7.636	3	3	1	1	1	7.111	-
4	3	3	5.791	6.745	3	3	3	1	6.600	7.400	3	3	2	1	1	7.200	8.073
4	4	1	4.967	6.667	3	3	3	2	6.727	8.015	3	3	2	2	1	7.591	8.576
4	4	2	5.455	7.036	3	3	3	3	7.000	8.538	3	3	2	2	2	7.910	9.115
4	4	3	5.598	7.144	4	1	1	1	-	-	3	3	3	1	1	7.576	8.424
4	4	4	5.692	7.654	4	2	1	1	5.833	-	3	3	3	2	1	7.769	9.051
5	2	1	5.000	-	4	2	2	1	6.133	7.000	3	3	3	2	2	8.044	9.505
5	2	2	5.160	6.533	4	2	2	2	6.545	7.391	3	3	3	3	1	8.000	9.451
5	3	1	4.960	-	4	3	1	1	6.178	7.067	3	3	3	3	2	8.200	9.876
5	3	2	5.251	6.909	4	3	2	1	6.309	7.455	3	3	3	3	3	8.333	10.20
5	3	3	5.648	7.079	4	3	2	2	6.621	7.871							
5	4	1	4.985	6.955	4	3	3	1	6.545	7.758							
5	4	2	5.273	7.205	4	3	3	2	6.795	8.333							
5	4	3	5.656	7.445	4	3	3	3	6.984	8.659							
5	4	4	5.657	7.760	4	4	1	1	5.945	7.909							
5	5	1	5.127	7.309	4	4	2	1	6.386	7.909							
5	5	2	5.338	7.338	4	4	2	2	6.731	8.346							
5	5	3	5.705	7.578	4	4	3	1	6.635	8.231							
5	5	4	5.666	7.823	4	4	3	2	6.874	8.621							
5	5	5	5.780	8.000	4	4	3	3	7.038	8.876							

Παραμετρικές δοκιμασίες	**Προϋποθέσεις παραμετρικών δοκιμασιών**	**Μη παραμετρικές εναλλακτικές**
Δύοανεξάρτητα δείγματα Student's t test	1) Τα δεδομένα και των δύο δειγμάτων έχουν συλλεγεί τυχαία 2) Τα δεδομένα και των δύο δειγμάτων προέρχονται από πληθυσμούς με κανονική κατανομή 3) Οι διακυμάνσεις τους είναι ίσες	Mann-Whitney U test
Κατά ζεύγη Student's t test	1) Η διαφορές (d_i) πρέπει να προέρχονται από πληθυσμό διαφορών με κανονική κατανομή	Wilcoxon signed rank
ANOVA	1) Τα δεδομένα όλων των δειγμάτων έχουν συλλεγεί τυχαία 2) Τα δεδομένα όλων των δειγμάτων προέρχονται από πληθυσμούς με κανονική κατανομή 3) Οι διακυμάνσεις τους είναι ίσες	Kruskal-Wallis H test
Συντελεστής συσχέτισης του Pearson	1) Τα δεδομένα Y για κάθε X πρέπει να έχουν συλλεγεί τυχαία από κανονική κατανομή των τιμών Y. 2) Τα δεδομένα X για κάθε Y πρέπει να έχουν συλλεγεί τυχαία από κανονική κατανομή των τιμών X.	Συντελεστής συσχέτισης τάξεως του Spearman

ΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ - ΣΥΣΧΕΤΙΣΗ
(Simple Linear Regression - Correlation)

Εύρεση μιας μαθηματικής ευθείας που εξηγεί τα δεδομένα

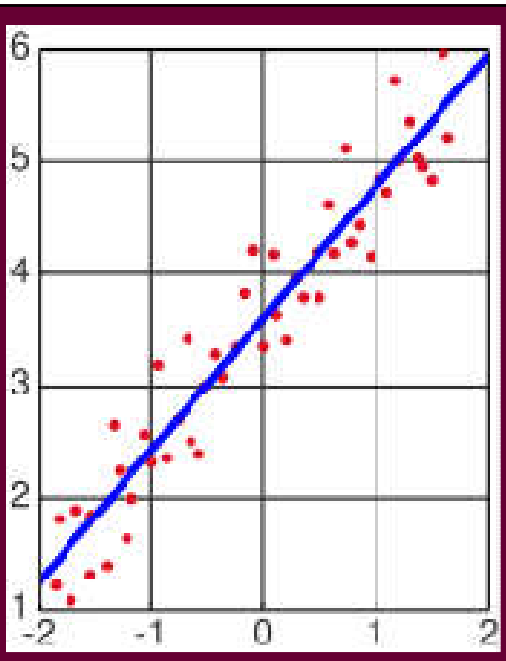
$$y = \alpha + \beta x + e$$

α : τεταγμένη επί την αρχή (intercept) β : κλίση (slope)

e : τυχαίο σφάλμα
x : ελεγχόμενη (predictor)
y : απόκριση (response)

Μέθοδος ελαχίστων τετραγώνων

$$y_i = a + bx_i + e_i \qquad \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$



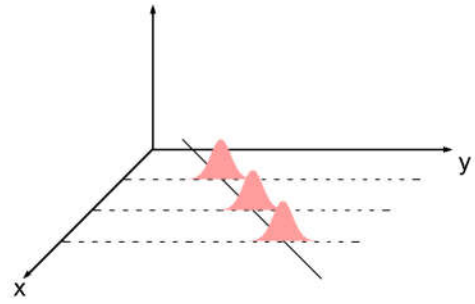
$$b = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$a = \bar{y} - b\bar{x}$$

Προϋποθέσεις για τη χρήση της απλής γραμμικής παλινδρόμησης

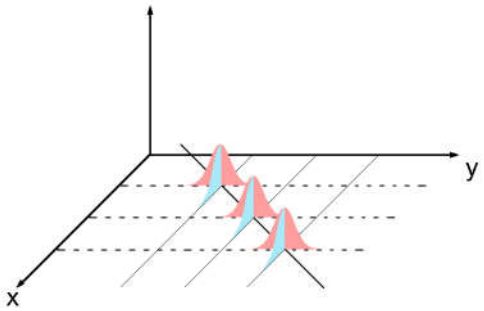
- Τα τυχαία σφάλματα στην "x" είναι αμελητέα
- Για κάθε τιμή της "x" υπάρχει μια κανονική κατανομή τιμών της "y".
- Η κατανομή του "y" για κάθε τιμή του "x" έχει την ίδια διακύμανση

Linear_Regr.xls



Απλή γραμμική παλινδρόμηση

Παλινδρόμηση Deming



Τυπικό σφάλμα της εκτίμησης
(standard error of the estimation)

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$$

Τυπικό σφάλμα για το "b"

δ.ε. $\beta = b \pm t \cdot s_b$, β. ε. n-2, a

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

Τυπικό σφάλμα για το "a"

δ.ε. $\alpha = a \pm t \cdot s_a$, β. ε. n-2, a

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

ΠΟΛΛΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ
(Multiple Linear Regression)

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e$$

Έστω δείγμα μεγέθους n

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_k x_{1k} + e_1$$
$$y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_k x_{2k} + e_2$$
$$\cdot$$
$$y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots + \beta_k x_{nk} + e_n$$

$$Y = X \cdot \beta + e$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots x_{1k} \\ 1 & x_{21} & x_{22} & \dots x_{2k} \\ \dots & & & \\ 1 & x_{n1} & x_{n2} & \dots x_{nk} \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_n \end{bmatrix} \qquad e = \begin{bmatrix} e_1 \\ e_2 \\ \dots \\ e_n \end{bmatrix}$$

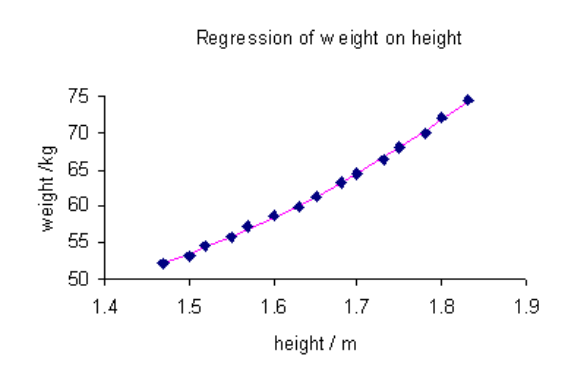
$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \dots \\ \hat{\beta}_n \end{bmatrix}$$

$$\hat{\beta} = \left(X^{'}X\right)^{-1} X^{'}Y$$

Απώλεια βάρους Y[gr]	Χρόνος έκθεσης X_1[h]	Σχετική υγρασία X_2
4.3	4	0.2
5.5	5	0.2
6.8	6	0.2
8.0	7	0.2
4.0	4	0.3
5.2	5	0.3
6.6	6	0.3
7.5	7	0.3
2.0	4	0.4
4.0	5	0.4
5.7	6	0.4
6.5	7	0.4

$$\hat{y} = 5.49 + 1.32x_1 - 8x_2$$

ΜΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ
(Non Linear Regression)



$$y = \alpha + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k + e$$

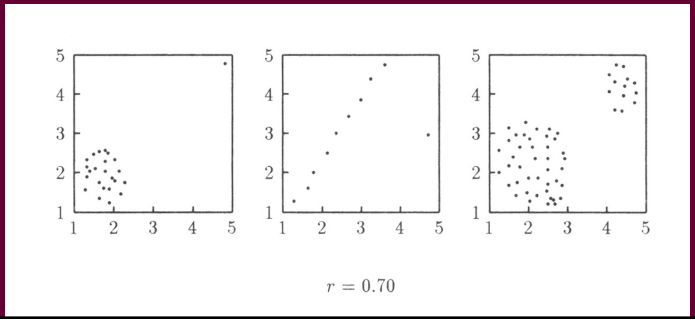
ΣΥΣΧΕΤΙΣΗ (Correlation)

Μέθοδος για την μέτρηση του βαθμού συμμεταβλητότητας των μεταβλητών.

Συντελεστής συσχέτισης Pearson

(Correlation coefficient)

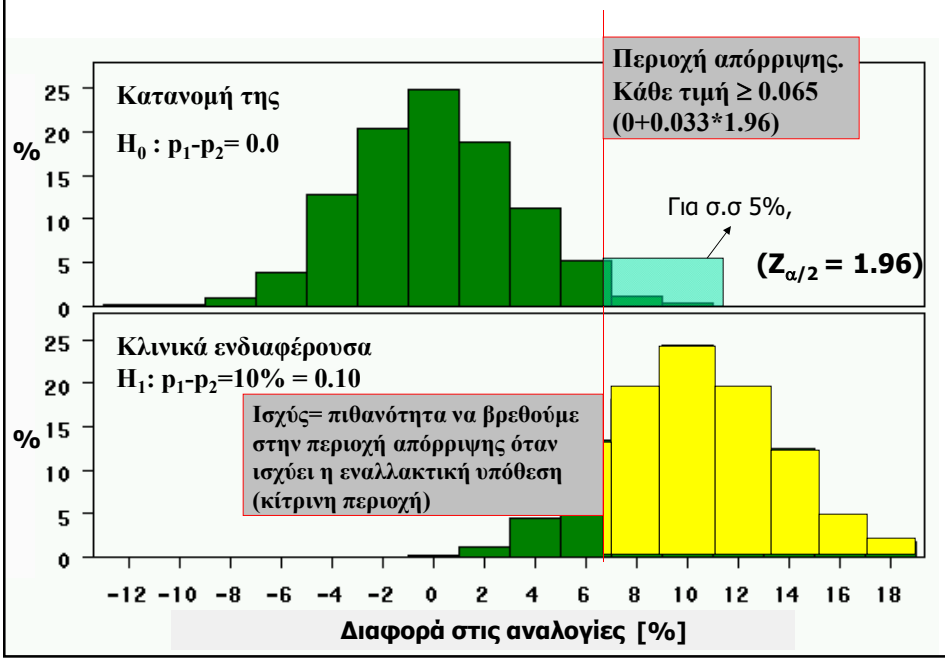
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

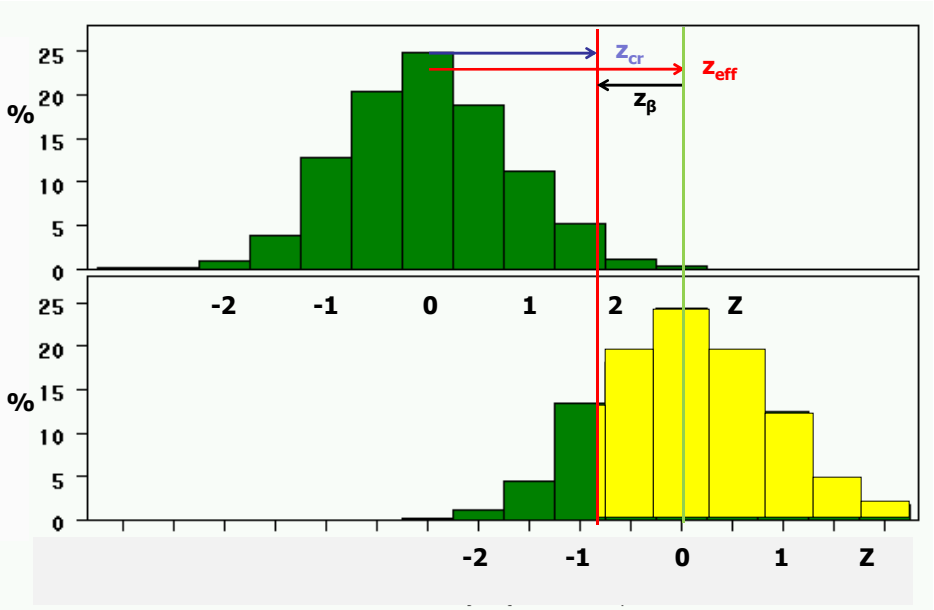


Εισαγωγή στο μέγεθος δείγματος και στον υπολογισμός της ισχύος μιας δοκιμασίας

Ποια είναι η πιθανότητα (Ισχύς) να απορρίψουμε την μηδενική υπόθεση όταν ισχύει η εναλλακτική;

(Ποια είναι η πιθανότητα να ανιχνεύσουμε μια πραγματική διαφορά;)





$$\gamma = P(Z > \frac{0.065}{0.033} - \frac{0.10}{0.033}) = P(Z > z_{cr} - z_{eff})$$

$$\gamma = P(Z > -1.06) = 85\%$$

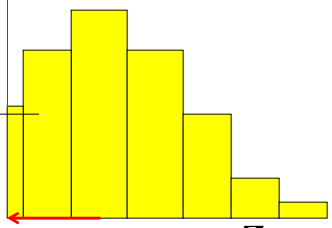
Περιοχή απόρριψης. Κάθε τιμή ≥ 0.065 (0+0.033*1.96)

z_{cr} : όριο απόρριψης της H_0

$$z_{eff} = \frac{\text{διαφορά}}{SE}$$

Ισχύς= πιθανότητα να βρεθούμε στην περιοχή απόρριψης όταν ισχύει η εναλλακτική υπόθεση (κίτρινη περιοχή)

Ισχύς:



z_{cr} - z_{eff} → z_β

Z

Παράγοντες που επηρεάζουν την Ισχύ

1. Μέγεθος της διαφοράς ↑
2. Τυπική απόκλιση ↓
3. Μεγάλο μέγεθος δείγματος ↑
4. Απαιτούμενη σ.σ ↓

Παραδείγματα

- Παράδειγμα : Θέλετε να υπολογίσετε πόση θα ήταν η ισχύς του να βρείτε διαφορά ΔIQ = 3.0 μεταξύ δύο ομάδων : 30 άνδρες γιατρούς και 30 γυναίκες. Εάν η αναμενόμενη τυπική απόκλιση είναι περίπου10, τότε το τυπικό σφάλμα της διαφοράς θα είναι περίπου:

$$SE = \sqrt{\frac{10^2}{30} + \frac{10^2}{30}} = 2.57 \qquad z_{eff} = \frac{\text{διαφορά}}{SE} = \frac{3.0}{2.57} = 1.167$$

$$z_\beta = z_{cr} - z_{eff} = 1.96 - 1.167 = 0.793$$

γ ή P = 0.5 - 0.2852 = 0.2148

Παραδείγματα

- Πόσο θα έπρεπε να είναι το μέγεθος του κάθε δείγματος για να επιτευχθεί ισχύς 80% (αντιστοιχεί σε z_β=0.84)

$$\mathbf{z}_\beta = \mathbf{z}_{\mathbf{cr}} - \mathbf{z}_{\mathbf{eff}} = 1.96 - \frac{3}{\sqrt{2 \cdot 10^2 / \mathbf{n}}} = -0.84 \Rightarrow \mathbf{n} = 174$$

$$\mathrm{n} = \frac{2\sigma^2 (Z_\beta + Z_{\alpha/2})^2}{\text{διαφορά}^2} = \frac{2 \cdot 10^2 (0.84 + 1.96)^2}{3^2} = 174$$

Απαιτούμενο μέγεθος δείγματος για σύγκριση δύο αναλογιών

Παράδειγμα: Προτίθεμαι να διεξαγάγω μια μελέτη για να ελέγξω εάν ο καρκίνος του παγκρέατος συνδέεται με την κατανάλωση καφέ. Εάν απαιτήσω ισχύ 80% για την ανίχνευση διαφοράς 0.1(10%) στα ποσοστά των καταναλωτών καφέ και του δείγματος ελέγχου, από πόσα άτομα θα πρέπει να αποτελείται η κάθε ομάδα; Περίπου το μισό του πληθυσμού πίνει καφέ. (Εάν η κατανάλωση καφέ και ο καρκίνος του παγκρέατος συνδέονται, θα αναμέναμε μεγαλύτερο ποσοστό καρκίνου μεταξύ των καταναλωτών καφέ από ό,τι στο δείγμα ελέγχου)

Απαιτούμενο μέγεθος δείγματος για σύγκριση δύο αναλογιών

Για τον υπολογισμό του z_β χρειάζεται το SE για την διαφορά των δύο ποσοστών.

$$SE = \sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

Με $p = 0.5$ και $n_1 = n_2 = n \rightarrow SE = 0.5\sqrt{(2/n)}$

$$\mathbf{z}_\beta = \mathbf{z}_{cr} - \mathbf{z}_{eff} = 1.96 - \frac{0.10}{0.5\sqrt{2/\mathbf{n}}} = -0.84 \Rightarrow \mathbf{n} = 392$$