ΣΤΑΤΙΣΤΙΚΗ (ΜΥ0202)

Μ.Ν. Ντυκέν,

Πανεπιστήμιο Θεσσαλίας
Τ.Μ.Χ.Π.Π.Α.

Βόλος, 2018-2019

ΔΙΑΛΕΞΗ 10

Συμπερασματική Στατιστική

Εκτίμηση κατά διάστημα εμπιστοσύνης

ΕΙΣΑΓΩΓΗ ΣΤΗΝ ΣΥΜΠΕΡΑΣΜΑΤΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

1. Ένας από τους σημαντικούς σκοπούς της Στατιστικής είναι η <u>εκτίμηση των παραμέτρων</u> που χαρακτηρίζουν την κατανομή ενός <u>πληθυσμού</u> σχετικά με την μεταβλητή που εξετάζουμε.
2. Σε πολλές περιπτώσεις, οι τιμές των παραμέτρων (μέση τιμή, τυπική απόκλιση, αναλογία) **για τον πληθυσμό** είναι **<u>άγνωστες</u>**.
3. Χρησιμοποιώντας ένα <u>αντιπροσωπευτικό δείγμα</u>, μπορούμε να **<u>εκτιμήσουμε</u>** τις τιμές των παραμέτρων.
4. Δεδομένου ότι η εκτίμηση βασίζεται σε υπό σύνολο του πληθυσμού, **δεν αποτελεί ακριβές υπολογισμός**. Υπάρχει επομένως **<u>σφάλμα εκτίμησης</u>**.
5. Όμως υπάρχουν μεθόδους με βάση τις οποίες μπορούμε να ελέγξουμε <u>σε ποιο βαθμό το σφάλμα περιορίζεται</u> προκειμένου να κάνουμε **<u>γενικεύσεις για όλο τον πληθυσμό</u>**.

Η συμπερασματική Στατιστική αποτελείται από εκείνες τις μεθόδους που μας επιτρέπουν να <u>διατυπώσουμε συμπεράσματα για όλο τον πληθυσμό</u> με βάση όμως τα δεδομένα που προέρχονται από το δείγμα.

Η ΛΟΓΙΚΗ ΤΗΣ ΣΥΜΠΕΡΑΣΜΑΤΙΚΗΣ ΣΤΑΤΙΣΤΙΚΗΣ

ΠΛΗΘΥΣΜΟΣ

Χαρακτηριστικά του πληθυσμού:
Γνωστά;

ΝΑΙ

Τυχαία μεταβλητή Χ:
όλες οι τιμές γνωστές

μ

σ

ΌΧΙ

Δείγμα
Δειγματικές Τιμές
Χ1, Χ2,…,Χn

$\bar{X}$

s

Σημειακοί Εκτιμητές

Μια τιμή ανά παράμετρο

Συμπερασματική Στατιστική:

Σημειακοί εκτιμητές – Εκτίμηση κατά διάστημα

Έλεγχος Υποθέσεων

Γραμμική Παλινδρόμηση

ΣΗΜΕΙΑΚΟΙ ΕΚΤΙΜΗΤΕΣ

- Η δειγματική μέση τιμή $\bar{X}$ (η οποία υπολογίζεται με βάση τα δεδομένα του δείγματος) αποτελεί '**μια**' εκτίμηση της μέσης τιμής $\boldsymbol{\mu}$ του πληθυσμού.
- Η δειγματική διακύμανση s^2 (βάσει των δεδομένων του δείγματος) αποτελεί '**μια**' εκτίμηση της διακύμανσης $\boldsymbol{\sigma}^2$ του πληθυσμού.

➔ Οι δειγματικές τιμές $\bar{X}$ και s^2 διαφέρουν από δείγμα σε δείγμα.

- Κατά συνέπεια, υπάρχει σχετική αβεβαιότητα όσον αφορά την αξιοπιστία της δειγματικής εκτίμησης.

- Όμως από τις δειγματικές τιμές, μπορούμε να βρούμε διαστήματα μέσα στα οποία βρίσκεται με συγκεκριμένο ποσοστό βεβαιότητας (π.χ. 95%) η μέση τιμή $\boldsymbol{\mu}$, η τυπική απόκλιση $\boldsymbol{\sigma}$ του πληθυσμού όπως και άλλοι παράμετροι (αναλογία: ποσοστό ατόμων) .

ΔΙΑΣΤΗΜΑ ΕΜΠΙΣΤΟΣΥΝΗΣ

- Οποιοδήποτε διάστημα επεκτείνεται ομοιόμορφα δεξιά και αριστερά από την παράμετρο που εξετάζουμε (μέση τιμή **μ**, τυπική απόκλιση **σ,** αναλογία **p,** κ.ά.) ονομάζεται **διάστημα εμπιστοσύνης (Δ.Ε.) και εκφράζεται ως εξής για την μέση τιμή:**

$$\boldsymbol{\mu - \beta \leq \mu \leq \mu + \beta}$$

- Ο ορισμός του β εξαρτάται από μια σειρά παραμέτρων.

ΔΙΑΣΤΗΜΑ ΕΜΠΙΣΤΟΣΥΝΗΣ

- Η εύρεση του διαστήματος εμπιστοσύνης μιας παραμέτρου (όπως η μέση τιμή)
  a) γίνεται με τη βοήθεια **<u>ενός γνωστού στατιστικού</u>** που **<u>περιέχει την εκτιμώμενη (από το δείγμα) παραμέτρου</u>**. Ο γνωστός αυτός στατιστικός ακολουθεί συστηματικά μια γνωστή κατανομή (όπως π.χ. η Κανονική Κατανομή, η Διωνυμική).
  b) Βασίζεται στο **επίπεδο εμπιστοσύνης (1-α)%** που επιλεγεί ο ίδιος ο ερευνητής **(π.χ. 95%).** Το επίπεδο εμπιστοσύνης αντανακλά το **<u>ποσοστό των παρατηρήσεων</u>** της μεταβλητής Χ **<u>που ανήκουν στο διάστημα εμπιστοσύνης.</u>**

     Κατά συνέπεια, το ποσοστό των τιμών της μεταβλητής Χ που βρίσκονται **<u>εκτός του διαστήματος εμπιστοσύνης</u>**, συμβολίζεται με **α** και ονομάζεται **επίπεδο σημαντικότητας (σφάλμα, ρίσκο).**

     Επιλέγοντας 95% επίπεδο εμπιστοσύνης (βεβαιότητα) δεχόμαστε 5% σφάλμα στην εκτίμηση της παραμέτρου.

ΔΙΑΣΤΗΜΑ ΕΜΠΙΣΤΟΣΥΝΗΣ ΤΗΣ ΜΕΣΗΣ ΤΙΜΗΣ

- Το **ακριβές** (1-α)% Δ.Ε. της μέσης τιμής είναι:

$$\mu \pm z_a \times \frac{\sigma}{\sqrt{n}}.\sqrt{d}$$

Όπου:

✓ Για **επίπεδο εμπιστοσύνης** = 95%, υπολογίζουμε το 95% Δ.Ε.,

✓ Το **επίπεδο σημαντικότητας** : α = 5% (0,05)

➔ $\mathbf{z_a = 1{,}96}$ (σύμφωνα με το Πίνακα της Κανονικής Κατανομής).

✓ n = μέγεθος δείγματος (n > 30)

✓ $\frac{\sigma}{\sqrt{n}}.\sqrt{d}$ = ακριβές τυπικό σφάλμα με $\boldsymbol{d = 1 - \frac{n}{N}}$

✓ $u = z_a.\frac{\sigma}{\sqrt{n}}.\sqrt{d}$= **ΑΚΡΙΒΕΣ Σφάλμα Δειγματοληψίας**

ΒΑΣΙΚΕΣ ΠΡΟΥΠΟΘΕΣΕΙΣ

- Η μεταβλητή X σε επίπεδο πληθυσμού χαρακτηρίζεται από την μέση τιμή: **μ** και την τυπική απόκλιση: **σ**

- Αν το δείγμα είναι αρκετά μεγάλο και αντιπροσωπευτικό, τότε η εκτίμηση της μέσης τιμής με βάση τα δεδομένα του δείγματος βασίζεται στην Κανονική Κατανομή:

$$\overline{X} \to N(\mu, \sigma^2)$$

- Όταν η μέση τιμή **μ** και η τυπική απόκλιση **σ** του πληθυσμού δεν είναι γνωστές (πιο πιθανό σενάριο),

  τότε για την εύρεση του διαστήματος εμπιστοσύνης:

  (α) χρησιμοποιούμε τις **σημειακές εκτιμήσεις**: $\overline{X}$ και s του δείγματος.

  (β) υπολογίζουμε συστηματικά το (1-α)% Δ.Ε. της μέσης τιμής

ΜΕΓΕΘΟΣ ΔΕΙΓΜΑΤΟΣ ΚΑΙ Δ.Ε. ΤΗΣ ΜΕΣΗΣ ΤΙΜΗΣ

(1-α)% Δ.Ε. = $\mu \pm z_a \times \frac{\sigma}{\sqrt{n}} . \sqrt{d}$ και $d \approx 1 - \frac{n}{N}$

Μέγεθος δείγματος: n > 30

Χρήση της ***Κανονικής* Κατανομής**

α = 5% ➔ (1-α) = 95%

➔ $z_{\alpha} = 1{,}96$

α = 1% ➔ (1-α) = 99%

➔ $z_{\alpha} = 2{,}576$

Μέγεθος δείγματος: n ≤ 30

Χρήση της **Κατανομής *Student***

α = 5% ➔ (1-α) = 95%

α = 1% ➔ (1-α) = 99%

➔ z_{α} = t(n-1; a/2)

Βλέπε πίνακα Student

Πως χρησιμοποιούμε τον Πίνακα Student?

Κατανομή Student για (n-1) και α/2

n-1	α/2 = 0.10	α/2= 0,05	α/2 = 0,025	α/2 = 0,01	α/2 = 0,005
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	$\boxed{2.093}$	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	$\boxed{2.797}$
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
∞	1.282	1.645	1.960	2.326	2.576

Για n =20
και
σφάλμα α = 5%

➔ α/2 = 2,5% = 0.025

➔ t(**n-1**, 0.025) =
t(19, 0.025) = 2.093

Για n =25
και
σφάλμα α = 1%

➔ α/2 = 0.5% = 0.005

➔ t(**n-1**, 0.005) =
t(24, 0.025) = 2.797

Πως χρησιμοποιούμε τον Πίνακα Student?

Κατανομή Student για (n-1) και α/2

n-1	α/2 = 0.10	α/2= 0,05	α/2 = 0,025	α/2 = 0,01	α/2 = 0,005
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
∞	1.282	1.645	1.960	2.326	2.576

Για n > 30 (∞)

όπως π.χ. n = 31

Παρατηρούμε ότι:

Αν α/2 = 0.025 δηλαδή α = 0.05 (5%),

t(**n-1**, 0.025) = t(30, 0.025) = t(∞, 0.025) = **1.96**

Αν α/2 = 0.005 δηλαδή α = 0.01 (1%),

t(**n-1**, 0.005) = t(30, 0.005) = t(∞, 0.005) = **2.576**

Παραδείγματα εκτίμησης κατά διάστημα **της μέσης τιμής.**

Παράδειγμα 1:

Μια δειγματοληπτική έρευνα σε 200 νοικοκυριά ενός Δήμου όπου συνολικά κατοικούν 2000 νοικοκυριά έδωσε τα ακόλουθα αποτελέσματα: κατά μέσο όρο, το μηνιαίο εισόδημα ανέρχεται σε 1100€ με διασπορά = 547600.

1. Δώστε (α) το 95% διάστημα εμπιστοσύνης και (β) το 99% διάστημα εμπιστοσύνης για το μέσο εισόδημα. Συμπέρασμα.

2. Η ίδια έρευνα επαναλήφθηκε ένα χρόνο μετά την 1η έρευνα. Τα αποτελέσματα της νέας έρευνας έδειξαν ότι, το μέσο εισόδημα = 1050 με διασπορά = 608400. Δώστε το 95% και 99% Δ.Ε. για το μέσο εισόδημα. Συμπέρασμα.

Παράδειγμα 1:

Μια δειγματοληπτική έρευνα σε 200 νοικοκυριά ενός Δήμου όπου συνολικά κατοικούν 2000 νοικοκυριά έδωσε τα ακόλουθα αποτελέσματα: κατά μέσο όρο, το μηνιαίο εισόδημα ανέρχεται σε 1100€ με διασπορά = 547600.

1. Δώστε (α) το 95% διάστημα εμπιστοσύνης και (β) το 99% διάστημα εμπιστοσύνης για το μέσο εισόδημα. Συμπέρασμα.

2. Η ίδια έρευνα επαναλήφθηκε ένα χρόνο μετά την 1η έρευνα. Τα αποτελέσματα της νέας έρευνας έδειξαν ότι, το μέσο εισόδημα = 1050 με διασπορά = 608400. Δώστε το 95% και 99% Δ.Ε. για το μέσο εισόδημα. Συμπέρασμα.

Τι γνωρίζουμε;

Από τα δεδομένα, έχουμε n= 200 και N= 2000 ➔ n/N = 0,1 ➔ d= 1- 0,1 = 0,9

Από το δείγμα, έχουμε $\bar{X} = 1000 \, , \sigma^2 = 547600 \Rightarrow \sigma = 740$

1.α 95% Δ.Ε. : $\mu \pm z_a \dfrac{\sigma}{\sqrt{n}} \sqrt{d} = 1100 \pm 1{,}96 \times \dfrac{740}{\sqrt{200}} \sqrt{0{,}9} = 1100 \pm 97{,}30$

n = 200 > 30 ➔ z_α = 1,96

Το εύρος του Δ.Ε. = 2 x 97,30 = 194,6 ➔ Το εύρος σε σχέση με τη μέση τιμή = 194,6 / 1100 ≈ 18% της μέσης τιμής (περιορισμένο εύρος), το μέσο εισόδημα κυμαίνεται μεταξύ: **1002,7 ≤ μ ≤ 1197,3**

1.β 99% Δ.Ε. : $\mu \pm z_a \frac{\sigma}{\sqrt{n}} \sqrt{d} = 1100 \pm 2{,}576 \times \frac{740}{\sqrt{200}} \sqrt{0{,}9} = 1100 \pm 127{,}87$

n = 200 > 30 ➔ z_α = 2,576 (βλέπε πίνακα)

Το εύρος του Δ.Ε. = 2 x 127,87 = 255,74 ➔ Το εύρος σε σχέση με τη μέση τιμή = 255,74 / 1100 ≈ 23% της μέσης τιμής (και εδώ έχουμε περιορισμένο εύρος), μπορούμε να στηρίξουμε ότι το <u>μέσο εισόδημα των κατοίκων του Δήμου</u> κυμαίνεται μεταξύ: 972,13 ≤ μ ≤ 1227,87.

2. Ένα χρόνο μετά, έχουμε: $\overline{X} = 1050 \,,\, \sigma^2 = 608400 \Rightarrow \sigma = 780$

α/ 95% Δ.Ε. : $\mu \pm z_a \frac{\sigma}{\sqrt{n}} \sqrt{d} = 1050 \pm 1{,}96 \times \frac{780}{\sqrt{200}} \sqrt{0{,}9} = 1050 \pm 102{,}56$

Το εύρος του Δ.Ε. = 2 x 102,56 = 205,12 ➔ Το εύρος σε σχέση με τη μέση τιμή = 205,12 / 1050 ≈ 20% της μέσης τιμής (περιορισμένο εύρος), <u>μέσο εισόδημα των κατοίκων του Δήμου</u> κυμαίνεται μεταξύ: **947,54 ≤ μ ≤ 1152,56.**

β/ Για το 99% Δ.Ε., το μέσο εισόδημα κυμαίνεται μεταξύ : **915,21 ≤ μ ≤ 1184,79** και το εύρος του αντιστοιχεί περίπου στο 26% της μέσης τιμής.

Παράδειγμα 2:

Μια δειγματοληπτική έρευνα σε 50 καταστήματα της χώρας έδειξε ότι η μέση τιμή ενός προϊόντος ανέρχεται σε 1,25€ με διασπορά = 0,25.

Δώστε (α) το 95% διάστημα εμπιστοσύνης και (β) το 98% διάστημα εμπιστοσύνης για τη μέση τιμή του προϊόντος. Συμπέρασμα.

Παράδειγμα 2:

Μια δειγματοληπτική έρευνα σε 50 καταστήματα της χώρας έδειξε ότι η μέση τιμή ενός προϊόντος ανέρχεται σε 1,25€ με διασπορά = 0,25.

Δώστε (α) το 95% διάστημα εμπιστοσύνης και (β) το 98% διάστημα εμπιστοσύνης για τη μέση τιμή του προϊόντος. Συμπέρασμα.

Τι γνωρίζουμε;

N άγνωστο όμως πολύ μεγάλο (καταστήματα της χώρας): n/N →∞ ➔ d =1

Από το δείγμα έχουμε : $\overline{X} = 1{,}25,\ \sigma^2 = 0{,}25 \Rightarrow \sigma = 0{,}5$

α/ 95% Δ.Ε. : $\mu \pm z_a \dfrac{\sigma}{\sqrt{n}} = 1{,}25 \pm 1{,}96 \times \dfrac{0{,}05}{\sqrt{50}} = 1{,}25 \pm 0{,}13$

n = 50 > 30 ➔ z_α = 1,96

Το εύρος του Δ.Ε. = 0,26 είναι περιορισμένο➔ Με 5% σφάλματος, μπορούμε να πούμε ότι τα καταστήματα εφαρμόζουν κοινή πολιτική ως προς τη τιμή του προϊόντος.

β/ 98% Δ.Ε.: $\mu \pm z_a \dfrac{\sigma}{\sqrt{n}} = 1{,}25 \pm 2{,}326 \times \dfrac{0{,}05}{\sqrt{50}} = 1{,}25 \pm 0{,}21.$

Είναι φανερό ότι και με 2% σφάλμα, το εύρος δεν είναι τόσο μικρό.

Διάστημα εμπιστοσύνης της αναλογίας.

ΔΙΑΣΤΗΜΑ ΕΜΠΙΣΤΟΣΥΝΗΣ ΤΗΣ ΑΝΑΛΟΓΙΑΣ: p

- Σε αρκετές περιπτώσεις, τα μεγέθη που εξετάζουμε εκφράζονται σε ποσοστά όπου p = ποσοστό ατόμων με συγκεκριμένο χαρακτήρα (ή απλή αναλογία).
- Όπως και για την μέση τιμή, μπορούμε να υπολογίσουμε τα όρια ακρίβειας του ποσοστού, δηλαδή να εκτιμήσουμε παρόμοιο διάστημα εμπιστοσύνης.
- Παράδειγμα: Μια έρευνα σε ένα δείγμα 200 ατόμων έδωσε τα ακόλουθα αποτελέσματα σχετικά με το κάπνισμα.

Καπνιστές	Αριθμός ατόμων	Αναλογία
Ναι	120	0.6
Όχι	80	0.4

- Το ποσοστό ατόμων που καπνίζουν = 60% ($p = 0.6$) και το αντίστοιχο για τους μη καπνιστές είναι 40% ($q = 1-p = 0.4$).
- Το ζήτημα είναι το ακόλουθο: σε ποιο βαθμό μπορούμε να γενικεύσουμε το αποτέλεσμα αυτό στο συνολικό πληθυσμό; Μπορούμε, βάσει των αποτελεσμάτων της έρευνας να θεωρήσουμε ότι, το 60% του πληθυσμού είναι καπνιστές;

ΔΙΑΣΤΗΜΑ ΕΜΠΙΣΤΟΣΥΝΗΣ ΤΗΣ ΑΝΑΛΟΓΙΑΣ: p

- Για να γενικεύσουμε με σημαντικό βαθμό βεβαιότητας, το αποτέλεσμα της έρευνας, θα πρέπει να βρούμε το Δ.Ε.
- Έστω :
  N = μέγεθος πληθυσμού
  n = μέγεθος του δείγματος
  p = αναλογία ατόμων (με βάση το δείγμα) που έχουν το χαρακτηριστικό και q = 1 - p
- Γνωρίζουμε επίσης ότι η αναλογία p ακολουθεί κατανομή Bernouilli όπου:
  $\mu = p$
  $\sigma^2 = p.q = p.(1-p) \Rightarrow \sigma = \sqrt{p.q}$
- Επομένως το (1-α)% Δ.Ε. για το ποσοστό δίνεται από:

$$p \pm z_a \times \frac{\sigma}{\sqrt{n}}.\sqrt{d} = p \pm z_a \times \sqrt{\frac{pq}{n}}.\sqrt{d} = p \pm z_a \times \sqrt{\frac{p(1-p)}{n}}.\sqrt{d}$$

$d = 1 - \frac{n}{N}$ Και $\sqrt{\frac{pq}{n}} =$ τυπικό σφάλμα του ποσοστού p

Προσοχή
Αν n μικρό, τότε παίρνουμε t(n-1, a/2)
Βλέπε πίνακα

Παραδείγματα εκτίμησης κατά διάστημα της αναλογίας.

Παράδειγμα 3:

Μια δειγματοληπτική έρευνα σε 25 φοιτητές της Θεσσαλίας έδειξε ότι:
(α) το ενοίκιο που πληρώνουν, ανέρχεται κατά μέσο όρο στα 210€ με τυπική απόκλιση = 65€.
(β) 20 από αυτούς δεν είναι καθόλου ικανοποιημένοι με το ενοίκιο που πληρώνουν.

1. Δώστε το 95% διάστημα εμπιστοσύνης για το μέσο ενοίκιο. Συμπέρασμα

2. Δώστε το 95% Δ.Ε. για το ποσοστό των μη ικανοποιημένων φοιτητών.

Παράδειγμα 3:

Μια δειγματοληπτική έρευνα σε 25 φοιτητές της Θεσσαλίας έδειξε ότι:
(α) το ενοίκιο που πληρώνουν, ανέρχεται κατά μέσο όρο στα 210€ με τυπική απόκλιση = 65€.
(β) 20 από αυτούς δεν είναι καθόλου ικανοποιημένοι με το ενοίκιο που πληρώνουν.

1. Δώστε το 95% διάστημα εμπιστοσύνης για το μέσο ενοίκιο. Συμπέρασμα

2. Δώστε το 95% Δ.Ε. για το ποσοστό των μη ικανοποιημένων φοιτητών.

Τι γνωρίζουμε;

Ν άγνωστο όμως μεγάλο (Φοιτητές της Θεσσαλίας): n/N →∞ → d =1

n = 25 < 30, θα πρέπει να χρησιμοποιούμε τον συντελεστή t(n-1; α/2).
Για 95% Δ.Ε., α = 5% (0,05) → α/2 = 0,025
Από τον πίνακα του Student, βρίσκουμε την τιμή του t(24; 0,025)= 2,064 (γραμμή 24 και στήλη α/2 =0.025)

1. 95% Δ.Ε. : $\mu \pm t(n-1; a/2)\dfrac{\sigma}{\sqrt{n}} = 210 \pm 2{,}064 \times \dfrac{65}{\sqrt{25}} = 210 \pm 26{,}83$

   Το μέσο ενοίκιο κυμαίνεται μεταξύ: **183,17 ≤ μ ≤ 236,83.** Το εύρος ξεπερνά τα 50€ (26% της μέσης τιμής), μπορεί να θεωρηθεί σχετικά μεγάλο για το προϋπολογισμό ενός φοιτητή.

2. Για να βρούμε το 95% Δ.Ε. για το ποσοστό των μη ικανοποιημένων φοιτητών, χρησιμοποιούμε το ποσοστό (αναλογία) που προέκυψε από το δείγμα.

<u>Γνωρίζουμε</u> από την έρευνα σε δείγμα 25 φοιτητών, ότι 20 στους 25 δεν είναι ικανοποιημένοι. Επομένως η αναλογία είναι p =0,8 και κατά συνέπεια q= 0,2.

n = 25 < 30, θα πρέπει να χρησιμοποιούμε τον συντελεστή t(n-1; α/2) = 2,064.

95% Δ.Ε. για την αναλογία δίνεται από: $p \pm t(n-1; a/2) \times \sqrt{\frac{pq}{n}}.\sqrt{d}$

Όπως αναφέρθηκε στο 1ο ερώτημα, n = 25 ➔ d=1

95% Δ.Ε. : $p \pm t(n-1; a/2) \times \sqrt{\frac{pq}{n}} = 0{,}8 \pm 2{,}064 \times \sqrt{\frac{0{,}8 \times 0{,}2}{25}} = 0{,}8 \pm 0{,}17 \Rightarrow 0{,}63 \leq p \leq 0{,}97$

Το διάστημα είναι ιδιαίτερα μεγάλο. Η εκτίμηση της αναλογίας δεν είναι στατιστικά ικανοποιητική και αυτό οφείλεται στο πολύ μικρό μέγεθος του δείγματος που δεν επιτρέπει αξιόπιστο συμπέρασμα.

Διάστημα εμπιστοσύνης για:

(α) τη διαφορά των μέσων τιμών δύο δειγμάτων,

(β) τη διαφορά των αναλογίων σε δύο δείγματα.

ΔΙΑΣΤΗΜΑ ΕΜΠΙΣΤΟΣΥΝΗΣ ΓΙΑ ΤΗ ΔΙΑΦΟΡΑ ΤΩΝ ΜΕΣΩΝ ΤΙΜΩΝ 2 ΔΕΙΓΜΑΤΩΝ

	1ο δείγμα	2ο δείγμα
Μέγεθος δείγματος	n_1	n_2
Μέση τιμή	$\overline{X}_1$	$\overline{X}_2$
Διασπορά	s_1^2	s_2^2

1η περίπτωση: τα δύο δείγματα είναι **μεγάλα ($n_1 \geq 30$ & $n_2 \geq 30$)**

Το Δ.Ε. για τη διαφορά των μέσων τιμών = $(\overline{X}_1 - \overline{X}_2) \pm z_a \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

2η περίπτωση: τα δύο δείγματα είναι **μικρά (ή τουλάχιστον ένα από τα δύο)**

Το Δ.Ε. για τη διαφορά των μέσων τιμών = $(\overline{X}_1 - \overline{X}_2) \pm t(v, a/2).s.\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

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

ΔΙΑΣΤΗΜΑ ΕΜΠΙΣΤΟΣΥΝΗΣ ΓΙΑ ΤΗ ΔΙΑΦΟΡΑ ΤΩΝ ΑΝΑΛΟΓΙΩΝ ΣΕ 2 ΔΕΙΓΜΑΤΑ

	1ο δείγμα	2ο δείγμα
Μέγεθος δείγματος	n_1	n_2
Ποσοστό	p_1 (q_1)	p_2 (q_2)

1η περίπτωση: τα δύο δείγματα είναι **μεγάλα ($n_1 \geq 30$ & $n_2 \geq 30$)**

Το Δ.Ε. για τη διαφορά των ποσοστών = $$(p_1 - p_2) \pm z_a \cdot \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

2η περίπτωση: τα δύο δείγματα είναι **μικρά (ή τουλάχιστον ένα από τα δύο)**

Το Δ.Ε. για τη διαφορά των ποσοστών = $$(p_1 - p_2) \pm t(\nu, a/2) \cdot \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

όπου $v = (n_1 + n_2 - 2)$

Παράδειγμα 4:

Μια δειγματοληπτική έρευνα σε 50 άνδρες και 50 γυναίκες έδειξε ότι,
(α) η μέση ηλικία των ανδρών = 52 με διασπορά 16 ενώ η μέση ηλικία των γυναικών = 55 με διασπορά = 9
(β) το 60% των ανδρών καπνίζουν έναντι του 50% για τις γυναίκες

1. Δώστε το 95% διάστημα εμπιστοσύνης για τη διαφορά των μέσων ηλικιών. <u>Συμπέρασμα.</u>

2. Δώστε το 95% διάστημα εμπιστοσύνης για τη διαφορά των ποσοστών των καπνιστών. <u>Συμπέρασμα</u>.

Παράδειγμα 4:

Μια δειγματοληπτική έρευνα σε 50 άνδρες και 50 γυναίκες έδειξε ότι,
(α) η μέση ηλικία των ανδρών = 52 με διασπορά 16 ενώ η μέση ηλικία των γυναικών = 55 με διασπορά = 9
(β) το 60% των ανδρών καπνίζουν έναντι του 50% για τις γυναίκες

1. Δώστε το 95% διάστημα εμπιστοσύνης για τη διαφορά των μέσων ηλικιών. Συμπέρασμα.
2. Δώστε το 95% διάστημα εμπιστοσύνης για τη διαφορά των ποσοστών των καπνιστών. Συμπέρασμα.

	Άνδρες	Γυναίκες
Μέγεθος δείγματος	50	50
Μέση ηλικία	52	55
Διασπορά	16	9

Τα δύο δείγματα > 30 ➔

$$95\%\,\Delta.\mathrm{E.}:(\bar{X}_1-\bar{X}_2)\pm z_a.\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

$$95\%\,\Delta.\mathrm{E.}:(52-55)\pm 1.96.\sqrt{\frac{16}{50}+\frac{9}{50}}=-3\pm 1{,}39\Rightarrow -4{,}39\leq \bar{X}_1-\bar{X}_2\leq -1{,}61$$

Εφόσον το Δ.Ε. **δεν περιλαμβάνει την τιμή 0** και η διαφορά είναι συστηματικά αρνητική, μπορούμε να πούμε ότι, υπάρχει πραγματική διαφορά μεταξύ των ανδρών και των γυναικών ως προς την ηλικία τους.

Για το 95% Δ.Ε. της διαφοράς των ποσοστών καπνιστών και δεδομένου ότι τα δύο δείγματα > 30, έχουμε τα ακόλουθα:

	Άνδρες	Γυναίκες
Μέγεθος δείγματος	50	50
p (% καπνιστών)	0,6	0,5
q(% μη καπνιστών)	0,4	0,5

$$95\%\ \Delta.\mathrm{E}.: (p_1 - p_2) \pm z_a \cdot \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

$$95\%\,\Delta.\mathrm{E}.: (0{,}6 - 0{,}5) \pm 1.96.\sqrt{\frac{0{,}6 \times 0{,}4}{50} + \frac{0{,}5 \times 0{,}5}{50}} = 0{,}10 \pm 0{,}19 \Rightarrow -0{,}09 \le p_1 - p_2 \le +0{,}29$$

Το Δ.Ε. περιλαμβάνει την τιμή 0, δηλαδή η διαφορά μπορεί να είναι και αρνητική και θετική! Αυτό δεν γίνεται και κατά συνέπεια, η διαφορά δεν είναι στατιστικά σημαντική.

Μια δειγματοληπτική έρευνα σε 69 φοιτητές του ΤΜΧΠΠΑ και του ΤΠΜ του Πανεπιστημίου Θεσσαλίας έδωσε τα ακόλουθα αποτελέσματα, σχετικά με την πρόθεσή τους να φύγουν στο εξωτερικό:

Τμήμα	Θέλουν να φύγουν	Δεν θέλουν να φύγουν	Σύνολο
ΤΜΧΠΠΑ	22	18	40
ΤΠΜ	21	8	29
Σύνολο	43	26	69

1. Δώστε το 95% και 99% Δ.Ε. για το ποσοστό φοιτητών που σκέφτονται να φύγουν στο εξωτερικό. Συμπέρασμα.
2. Δώστε το 95% Δ.Ε. για τη διαφορά μεταξύ των δύο τμημάτων του ποσοστού των φοιτητών που σκέφτονται να φύγουν; Είναι σημαντική η διαφορά;

Για τον υπολογισμό των Δ.Ε., θα πρέπει να υπολογίσουμε τα σχετικά ποσοστά των φοιτητών που θέλουν να φύγουν (p) και των υπόλοιπων (q=1-p):

Τμήμα	Θέλουν να φύγουν	Δεν θέλουν να φύγουν	Σύνολο
ΤΜΧΠΠΑ	55,0	45,0	100,0
ΤΠΜ	72,4	27,6	100,0
Σύνολο	62,3	37,7	100,0

(1.) 95% Δ.Ε. για το ποσοστό φοιτητών (δηλαδή το σύνολο) που σκέφτονται να φύγουν στο εξωτερικό.

Το 62,3% των 69 φοιτητών που εξετάζουμε θέλουν να φύγουν: p = 62,3 & q= 37,7.

n = 69 > 30 ➔ z_a = 1,96 για 95% Δ.Ε. και z_a = 2,576 για 99% Δ.Ε.

Θεωρούμε ότι, n/N → 0 ➔ d → 1

➔

$$95\%\ \Delta.\mathrm{E.} = p \pm z_a \cdot \sqrt{\frac{pq}{n}} = 62{,}3 \pm 1{,}96\sqrt{\frac{62{,}3 \times 37{,}7}{69}} = 62{,}3 \pm 11{,}4$$

50,9 < p < 73,7

$$99\%\ \Delta.\mathrm{E.} = p \pm z_a \cdot \sqrt{\frac{pq}{n}} = 62{,}3 \pm 2{,}576\sqrt{\frac{62{,}3 \times 37{,}7}{69}} = 62{,}3 \pm 15{,}0$$

47,3 < p < 77,3

Συμπέρασμα: Τα Δ.Ε. αναδεικνύουν με 95% βεβαιότητας ότι τουλάχιστον το 50% των φοιτητών θέλουν να φύγουν ή με 99% βεβαιότητας, τουλάχιστον 47%!

(2.) 95% Δ.Ε. για τη διαφορά των ποσοστών μεταξύ των 2 τμημάτων.

Τι γνωρίζουμε;

Τμήμα	**n**	**p**	**q**
ΤΜΧΠΠΑ	40	55,0	45,0
ΤΠΜ	29 < 30	72,4	27,6

Για ένα από τα δύο δείγματα, n < 30, επομένως το 95% της διαφοράς των ποσοστών δίνεται από:

$$95\%\ \Delta.\text{E.} = (p_1 - p_2) \pm t(v, a/2) \times \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

Όπου $v = n_1 + n_2 - 2 = 40 + 29 - 2 = 67$. Ο πίνακας δεν μας δίνει $v = 67$ ➔ κατά προσέγγιση, παίρνουμε $v = 60$ που είναι πιο κοντά από το $v = 120$.

Για 95% Δ.Ε., $t(60, 0{,}025) = 2{,}000$

$$95\%\ \Delta.\text{E.} = (72{,}4 - 55{,}0) \pm 2 \times \sqrt{\frac{72{,}4 \times 27{,}6}{29} + \frac{55{,}0 \times 45{,}0}{40}} = 17{,}4 \pm 2 \times \sqrt{68{,}9 + 61{,}9} = 17{,}4 \pm 22{,}9$$

➔ **- $-5{,}5 \leq p1 - p2 \leq 40{,}3$** ➔ Το Δ.Ε. είναι μεγάλο και **<u>περιλαμβάνει το 0</u>**!!!

➔ Τα δύο ποσοστά παρουσιάζουν διαφορά η οποία είναι απόλυτα τυχαία και όχι αντιπροσωπευτική. <u>Δεν μπορούμε να πούμε ότι, υπάρχει στατιστικά σημαντική διαφορά</u>.

Σχέση μεταξύ σφάλματος εκτίμησης και μεγέθους δείγματος

Σφάλμα εκτίμησης - Μέγεθος δείγματος

Η μέση τιμή όπως και η αναλογία που «υπολογίζουμε» με βάση τα δεδομένα του δείγματος αποτελούν **<u>εκτιμήσεις</u>** για το πληθυσμό και περιλαμβάνουν σφάλμα δειγματοληψίας.

[A] $$\mu \pm \underbrace{z_a \times \frac{\sigma}{\sqrt{n}}.\sqrt{d}}_{U}$$

[B] $$p \pm \underbrace{z_a \times \sqrt{\frac{pq}{n}}.\sqrt{d}}_{U}$$

U = σφάλμα δειγματοληψίας = σφάλμα εκτίμησης και υπολογίζεται με τα δεδομένα του δείγματος.

Σφάλμα εκτίμησης - Μέγεθος δείγματος

[A] $U = z_a \frac{\sigma}{\sqrt{n}} \sqrt{d}$ **: 1η περίπτωση βασισμένη στη μέση τιμή και διασπορά μιας επιλεγμένης μεταβλητής**

Σε περίπτωση που n/N → 0 ➔ $U = z_a \frac{\sigma}{\sqrt{n}} \Rightarrow n = z_a^2 \frac{\sigma^2}{U^2}$

Αν θέλουμε <u>μικρό σφάλμα δειγματοληψίας</u>, θα πρέπει να επιλέγουμε σχετικά μεγάλο μέγεθος δείγματος, ειδικά όταν γνωρίζουμε ότι ο πληθυσμός που μελετάμε παρουσιάζει σημαντική διασπορά ή / και ανισότητες.

Σφάλμα εκτίμησης - Μέγεθος δείγματος

[Β] $U = z_a \times \sqrt{\frac{pq}{n}}.\sqrt{d}$ **: 2η Περίπτωση βασισμένη σε μια αναλόγια του πληθυσμού.**

Σε περίπτωση που n/N → 0 ➔ $U = z_a \times \sqrt{\frac{pq}{n}} \Rightarrow n = z_a^2 \frac{pq}{U^2}$

όπου p = ποσοστό του πληθυσμού που έχει το χαρακτηριστικό που εξετάζουμε.

Αν p γνωστό (π.χ. Γνωρίζουμε ότι, το ποσοστό ανεργίας στην Ελλάδα είναι της τάξης του 28% σύμφωνα με τα δεδομένα της Eurostat), τότε μπορούμε να υπολογίσουμε τον «κατάλληλο» μέγεθος δείγματος για

(α) ένα επιλεγμένο επίπεδο εμπιστοσύνης **(1-α)%**

και ταυτόχρονα

(β) ένα συγκεκριμένο και επιλεγμένο σφάλμα εκτίμησης **U%**

Ο ερευνητής επιλέγει

Σφάλμα εκτίμησης - Μέγεθος δείγματος

Παράδειγμα:

Ορισμός του κατάλληλου ***Μεγέθους δείγματος*** *για μια έρευνα που θα αφορούσε τον κίνδυνο ανεργίας για τον Οικονομικά ενεργό πληθυσμό της Ελλάδας, δεδομένου ότι, το ποσοστό ανεργίας = 28%.*

Διαδικασία:

(α) επιλογή του επιπέδου εμπιστοσύνης: (1-α)% = 95% ➔ $\boldsymbol{z_a = 1{,}96}$

(β) επιλογή του σφάλματος εκτίμησης : ***U = ??? : εξετάζουμε διάφορα εναλλακτικά επίπεδα***

$$U = 1{,}96 \times \sqrt{\frac{0{,}28 \times 0{,}72}{n}} \Rightarrow n = 1{,}96^2 \frac{0{,}28 \times 0{,}72}{U^2} = \frac{0{,}7745}{U^2}$$

U		n	Στρογγυλοποίηση
10%	0,1	77,45	77
5%	0,05	309,8	310
1%	0,01	7744,7	7745

Σφάλμα εκτίμησης - Μέγεθος δείγματος

(γ) Σύγκριση με ***εναλλακτικά επίπεδα*** *τόσο για το* ***επίπεδο εμπιστοσύνης*** *όσο και για το* ***σφάλμα εκτίμησης***

(1-a)% = 99%

➔ $z_a = 2{,}58$

$$U = 2{,}58 \times \sqrt{\frac{0{,}28 \times 0{,}72}{n}} \Rightarrow n = 2{,}58^2 \frac{0{,}28 \times 0{,}72}{U^2} = \frac{1{,}3419}{U^2}$$

U		(1-α)% = 95% $z_a = 1{,}96$		(1-α)% = 99% $z_a = 2{,}58$	
		n	Στρογγυλοποίηση	n	Στρογγυλοποίηση
10%	0,1	77,45	77	134,2	134
5%	0,05	309,8	310	536,8	537
1%	0,01	7744,7	7745	13419,3	13419

(δ) ***Απόφαση:*** εξαρτάται σε σημαντικό βαθμό από το ***<u>κόστος εφαρμογής της έρευνας πεδίου</u>***. Π.χ. αν το κόστος ανά ερωτηματολόγιο = 5€ και ο προϋπολογισμός μας δεν ξεπέρνα τα 1600€, τότε δεν μπορούμε να επιλέξουμε πάνω από 320 ερωτηματολόγια ➔ n = 310

Σφάλμα εκτίμησης - Μέγεθος δείγματος

Τι γίνεται αν το ποσοστό p δεν είναι γνωστό;

Σε αυτή την περίπτωση, επιλέγουμε συστηματικά p = 50% (0,5) διότι η αναλογία αυτή οδηγεί στο **μέγιστο σφάλμα εκτίμησης**.

p	q	p.q
0,5	0,5	**0,25**
0,4	0,6	0,24
0,3	0,7	0,21
0,2	0,8	0,16
0,1	0,9	0,09