ΑΡΧΙΜΙΔΗΣ ΙΙΙ

Ενίσχυση Ερευνητικών Ομάδων στο ΤΕΙ Πάτρας

ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ Ι
Περιγραφική Στατιστική, Παραμετρικές και μη Παραμετρικές Διαδικασίες

ΓΙΩΡΓΟΣ ΒΛΑΧΟΠΟΥΛΟΣ

28/05/2015

1

1. Εισαγωγή

Τομείς Στατιστικής

1. **Περιγραφική Στατιστική**
2. **Επαγωγική Στατιστική**

- **Περιγραφική Στατιστική**

  Ασχολείται με την ποσοτική περιγραφή των φυσικών, κοινωνικών και λοιπών φαινομένων.

- **Επαγωγική Στατιστική**

  Ασχολείται με τη γενίκευση των συμπερασμάτων που προκύπτουν από τις περιγραφικές στατιστικές διαδικασίες και τις προβλέψεις για όλο το σύνολο.

2

2. Βασικές έννοιες

❖ **Πληθυσμός**

Το σύνολο των τιμών που εκφράζει ποσοτικά το υπό μελέτη χαρακτηριστικό.

Παραδείγματα

- ❖ Το σύνολο των ασθενών ενός νοσοκομείου με υψηλές τιμές αρτηριακής πίεσης
- ❖ Το σύνολο των επιπέδων σακχάρου στους ανθρώπους

3

2. Βασικές έννοιες

❖ **Συλλογή δεδομένων**

Μπορεί να γίνει με:

- ➢ **Απογραφή**

  Συγκέντρωση στοιχείων από όλο τον πληθυσμό
- ➢ **Δειγματοληψία**

  Συγκέντρωση στοιχείων από ένα τμήμα του πληθυσμού

4

2. Βασικές έννοιες

❖ **Τυχαία Μεταβλητή**

Εκφράζει το χαρακτηριστικό του πληθυσμού που μπορεί να προσδιοριστεί ή να πάρει κάποια τιμή μετρηθεί.

A. Ποιοτικές μεταβλητές (qualitative ή categorical ή string variables)

Δεν λαμβάνουν αριθμητικές τιμές, αλλά περιγράφονται οι κατηγορίες στις οποίες ταξινομούνται οι παρατηρήσεις.

Παραδείγματα : το φύλο, χρώμα μαλλιών κλπ

B. Ποσοτικές μεταβλητές (quantitative ή numerical variables)

Αυτές που είναι μετρήσιμες

Παραδείγματα : τιμή αρτηριακής πίεσης, βάρος, ηλικία κλπ

5

2. Βασικές έννοιες

❖ **Ποσοτικές Μεταβλητές**

Ασυνεχείς (ή διακριτές) μεταβλητές (discrete variables)

Μπορούν να λάβουν ορισμένες αριθμητικές τιμές.

Παραδείγματα :Πλήθος κυττάρων που εκφράζονται με συγκεκριμένο τρόπο, αριθμός ασθενών που παρουσίασαν παρενέργειες από την χορήγηση ενός σκευάσματος κλπ.

B. Συνεχείς μεταβλητές (continuous variable)

Μπορούν να λάβουν (θεωρητικά) οποιαδήποτε πραγματική τιμή εντός ενός επιλεγμένου διαστήματος.

Παραδείγματα : τιμή αρτηριακής πίεσης, βάρος, ηλικία κλπ

Παρατήρηση : Οι συνεχείς μεταβλητές μετρούνται με καθορισμένη εκ των προτέρων ακρίβεια οπότε «κατά βάθος» είναι Ασυνεχείς, π.χ. το βάρος ενός ασθενούς μπορεί θεωρητικά να πάρει μια οποιαδήποτε τιμή αλλά δε θα ήταν δυνατό να μετρηθεί βάρος 75,2564646456489 Kg

6

2. Βασικές έννοιες

Κλίμακες Μέτρησης

1. **Ονομαστική κλίμακα** (nominal scale) → Κάθε αντικείμενο κατατάσσεται σε μια μόνο κατηγορία. Μεταξύ των κατηγοριών δεν ορίζεται ούτε διάταξη ούτε απόσταση (π.χ Φύλο, χρώμα μαλλιών κλπ.)

2. **Διατεταγμένη κλίμακα** (ordinal scale) → Οι κατηγορίες ταξινομούνται με μια σειρά που δηλώνει μια διάταξη, μια ιεραρχία. Υπάρχει η ιδιότητα της διάταξης αλλά όχι της απόστασης (π.χ κατάσταση ασθενούς με τιμές: Ελαφριά, Σοβαρή Κρίσιμη)

3. **Κλίμακα διαστήματος** (interval scale) → Οι κατηγορίες ταξινομούνται με μια σειρά που δηλώνει μια διάταξη, μια ιεραρχία. Υπάρχει η ιδιότητα της διάταξης αλλά όχι της απόστασης. Δεν ορίζεται η αναλογικότητα και το μηδέν (π.χ. αξιολόγηση υπηρεσίας στην κλίμακα 1-5)

4. **Αναλογική κλίμακα** (ratio scale) → Χαρακτηρίζεται από τις ιδιότητες της διάταξης, της απόστασης και του μηδενικού στοιχείου (π.χ περιεκτικότητα αλκοόλ στο αίμα)

7

2. Βασικές έννοιες

❖ **(Τυχαία) Μεταβλητή** (συνέχεια)

D. Αναλογική κλίμακα μέτρησης

Χαρακτηρίζεται από τις ιδιότητες της διάταξης, της απόστασης και του μηδενικού στοιχείου.

Μηδενικό στοιχείο : Για ένα αντικείμενο μέτρησης που δεν έχει καθόλου την ιδιότητα που εξετάζεται τότε σε αυτό αντιστοιχίζεται η τιμή μηδέν.

Στην αναλογική κλίμακα είναι επιτρεπτές όλες οι μαθηματικές πράξεις του συνόλου των πραγματικών αριθμών.

Παράδειγμα αποτελεί η κλίμακα με τιμές 0 έως 100 που μετρά την τιμή χρέωσης παρεχόμενων υπηρεσιών υγείας.

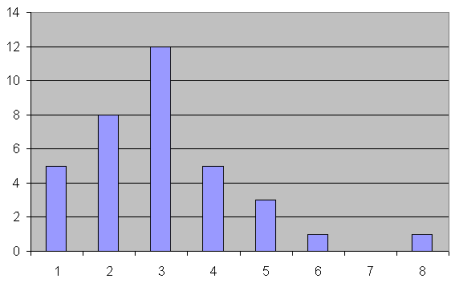
8

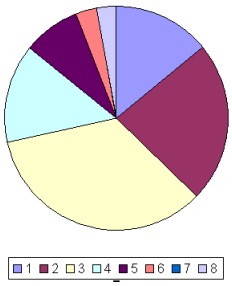
3. Περιγραφική στατιστική ανάλυση

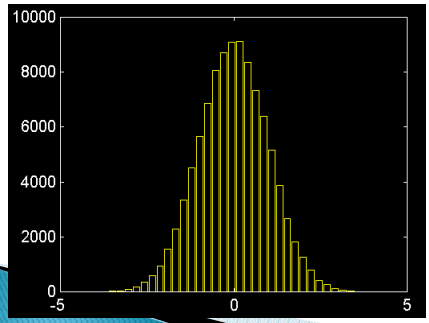
- **Συχνότητα (εμφάνισης)**
  Εκφράζει πόσες φορές εμφανίζεται η τιμή μιας μεταβλητής στα δεδομένα.
  - **(Απόλυτη) συχνότητα**
    Ισούται με το πλήθος εμφάνισης της τιμής της μεταβλητής στα δεδομένα
  - **Σχετική συχνότητα**
    Ισούται με την (απόλυτη) συχνότητα διαιρεμένη με το μέγεθος του δείγματος.
- **Πίνακας συχνοτήτων**
  Παρουσιάζει τις συχνότητες όλων των τιμών μιας μεταβλητής στα δεδομένα.
  Παράδειγμα : Ο πίνακας συχνοτήτων της μεταβλητής «φύλο» παρουσιάζει το πλήθος των ανδρών και των γυναικών στο δείγμα.
- **Απεικόνιση συχνοτήτων**
  - Με ραβδογράμματα → Διαγράμματα που παρουσιάζουν τις συχνότητες ως ξεχωριστές ράβδους μεταβλητού ύψους (συνήθως για ποιοτικές μεταβλητές)
  - Με διαγράμματα πίττας
  - Με ιστογράμματα → Χρησιμοποιείται για ποσοτικές μεταβλητές. Οι στήλες εμφανίζονται κατά σειρά μεγέθους των τιμών και δεν χωρίζονται με κενά.

9

3. Περιγραφική στατιστική ανάλυση







10

3. Περιγραφική στατιστική ανάλυση

- **Κατανομή συχνοτήτων (distribution function)**
  Αποτελεί το σύνολο όλων των πιθανών τιμών που μπορεί να πάρει μια τυχαία μεταβλητή μαζί με τις αντίστοιχες συχνότητες εμφάνισης κάθε τιμής.
- **Αθροιστική κατανομή συχνοτήτων (cumulative distribution function)**
  Αναπαριστά την πιθανότητα η τυχαία μεταβλητή να λαμβάνει τιμές μικρότερες ή ίση με συγκεκριμένη τιμή x.
- **Χαρακτηριστικές τιμές των κατανομών συχνοτήτων**
  - Μέτρα τάσης συγκέντρωσης τιμών (measures of central tendency)
    - Μέση τιμή (average)
    - Διάμεσος (median)
    - Επικρατούσα τιμή (most frequent value)

11

3. Περιγραφική στατιστική ανάλυση

- **Χαρακτηριστικές τιμές των κατανομών συχνοτήτων** (συνέχεια)
  - Μέτρα διασποράς (measures of spread)
    - Εύρος (range)
    - Ενδοτεταρτομοριακό εύρος (interquartile range)
    - Διασπορά (variance)
    - Τυπική απόκλιση (standard deviation)
    - Κεντρικές Ροπές (central moments)
    - Συντελεστής μεταβλητότητας (dispersion factor)
  - Μέτρα σχήματος (measures of shape)
    - Συντελεστής ασυμμετρίας (skewness)
    - Συντελεστής κύρτωσης (kurtosis)

12

3. Περιγραφική στατιστική ανάλυση

❖ **Μέση τιμή**

- Πληθυσμιακή μέση τιμή

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

- Δειγματική μέση τιμή

$$\bar{x} = \mathrm{E}(x) = \frac{\sum_{i=1}^{n} x_i}{n}$$

Συμμετρική κατανομή : οι τιμές της μεταβλητής διατάσσονται συμμετρικά γύρω από τη μέση τιμή.

13

3. Περιγραφική στατιστική ανάλυση

❖ **Διάμεσος**

Είναι η τιμή που χωρίζει το δείγμα σε δύο ισοπλήθη μέρη όταν έχουν διαταχθεί οι τιμές κατ' αύξουσα σειρά.

- Άρα οι μισές τιμές είναι μεγαλύτερες από τη διάμεσο και οι άλλες μισές τιμές μικρότερες.
- Αν το πλήθος των αριθμών είναι άρτιος, ως διάμεσος λαμβάνεται η μέση τιμή των δύο μεσαίων τιμών.

Παράδειγμα: Η διάμεσος των τιμών {6,6,8,8,10,10,10,17} είναι 9.

❖ **Επικρατούσα τιμή**

Είναι η συχνότερα εμφανιζόμενη τιμή στα δεδομένα.

- Σε ένα σύνολο τιμών είναι δυνατόν να υπάρξουν περισσότερες της μιας επικρατούσες τιμές.

Παράδειγμα: Η επικρατούσα τιμή των {6,6,8,8,10,10,10,17} είναι το 10.

14

3. Περιγραφική στατιστική ανάλυση

- **Επίδραση των απομονωμένων τιμών (outliers) στα μέτρα τάσης συγκέντρωσης τιμών**
  - Οι απομονωμένες τιμές (outliers) υπάρχουν σχεδόν σε όλα τα δεδομένα του πραγματικού κόσμου.
  - Η (δειγματική) μέση τιμή είναι ευαίσθητη στην παρουσία των απομονωμένων τιμών. Η ύπαρξη μιας τέτοιας τιμής μπορεί να μετακινήσει τη μέση τιμή πολύ μακριά από το μέσο των υπόλοιπων δεδομένων.
  - Αντίθετα, η διάμεσος και η επικρατούσα τιμή είναι ανθεκτικές στην παρουσία των απομονωμένων τιμών. Η παρουσία τέτοιων τιμών μεταβάλλει ελάχιστα τη διάμεσο.

15

3. Περιγραφική στατιστική ανάλυση

- **Σχέση Μέση τιμής – Διάμεσου – Επικρατούσας Τιμής**
  - Γενικά τα τρία χαρακτηριστικά δεν συμπίπτουν.
  - Συμπίπτουν μόνο όταν η κατανομή είναι συμμετρική και έχει μόνο μία κορυφή.
  - Αν η κατανομή δεν είναι συμμετρική και έχει μόνο μία κορυφή, τότε η μέση τιμή επηρεάζεται από τις ουρές της κατανομής και απομακρύνεται από το κέντρο της κατανομής προς τις ουρές, ενώ η διάμεσος βρίσκεται μεταξύ επικρατούσας τιμής (αντιστοιχεί στην κορυφή) και μέσης τιμής.

16

3. Περιγραφική στατιστική ανάλυση

❖ **Εύρος**

Είναι η διαφορά μεταξύ της μέγιστης μείον την ελάχιστη τιμή των δεδομένων.

❖ **Ενδοτεταρτομοριακό εύρος**

Είναι η απόσταση του 75ου από το 25ο ποσοστιαίο σημείο.

❖ **Διασπορά (ή διακύμανση)**

Είναι η μέση τιμή των τετραγώνων της απόστασης κάθε τιμής από τη μέση τιμή.

➢ Πληθυσμιακή διασπορά

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

➢ Δειγματική διασπορά

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

17

3. Περιγραφική στατιστική ανάλυση

❖ **Διασπορά** (συνέχεια)

✓ Όσο μεγαλύτερη είναι η διασπορά τόσο περισσότερο διασκορπισμένες είναι οι τιμές.

✓ Αν η διασπορά είναι μηδενική, τότε όλες οι παρατηρήσεις έχουν την ίδια τιμή.

❖ **Τυπική απόκλιση**

Είναι η τετραγωνική ρίζα της διασποράς.

➢ Πληθυσμιακή τυπική απόκλιση

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

➢ Δειγματική τυπική απόκλιση

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

18

3. Περιγραφική στατιστική ανάλυση

❖ **Τυπική απόκλιση** (συνέχεια)

- ✓ Όσο μεγαλύτερη είναι η τυπική απόκλιση τόσο περισσότερο διασκορπισμένες είναι οι τιμές.
- ✓ Αν η τυπική απόκλιση είναι μηδενική, τότε όλες οι παρατηρήσεις έχουν την ίδια τιμή.

Παρατήρηση :

Η τυπική απόκλιση (όπως και η διασπορά) δίνει μεγαλύτερη βαρύτητα σε εκείνες τις τιμές που απέχουν περισσότερο από τη μέση τιμή και μικρότερη σε εκείνες που βρίσκονται κοντά στη μέση τιμή.

❖ **Κεντρικές Ροπές**

Η k-τάξης κεντρική ροπή υπολογίζεται από τη σχέση : $m_k = E[(x-\mu)^k]$

❖ **Συντελεστής μεταβλητότητας**

Είναι η τυπική απόκλιση ως ποσοστό της μέσης τιμής.

19

3. Περιγραφική στατιστική ανάλυση

➢ **Επίδραση των απομονωμένων τιμών (outliers) στα μέτρα διασποράς**

- Το εύρος επηρεάζεται σημαντικά από τις απομονωμένες τιμές.
- Η τυπική απόκλιση, η διασπορά και οι κεντρικές ροπές επηρεάζονται σημαντικά από τις απομονωμένες τιμές. Μια τέτοια τιμή μπορεί να αυξήσει τα εν λόγω μέτρα κατά πολύ.
- Αντίθετα το ενδοτεταρτομοριακό εύρος δεν επηρεάζεται από την ύπαρξη απομονωμένων τιμών.

20

3. Περιγραφική στατιστική ανάλυση

- **Σημασία μέσης τιμής και τυπικής απόκλισης**
  - Με τη βοήθεια της τυπικής απόκλισης υπολογίζεται το ποσοστό των τιμών των δεδομένων που συγκεντρώνονται σε συγκεκριμένες αποστάσεις γύρω από τη μέση τιμή. Οι αποστάσεις αυτές μετρώνται σε πολλαπλάσια της τυπικής απόκλισης.
  - Ανισότητα Chebyshev

    Ανεξάρτητα από τη μορφή της κατανομής των δεδομένων ισχύουν τα ακόλουθα :
    - ✓ Τουλάχιστον το 75% των παρατηρήσεων βρίσκονται στο διάστημα

      $$\bar{x} \pm 2,5s$$
    - ✓ Τουλάχιστον το 88,8% των παρατηρήσεων βρίσκονται στο διάστημα

      $$\bar{x} \pm 3,5s$$

21

3. Περιγραφική στατιστική ανάλυση

- **Σημασία μέσης τιμής και τυπικής απόκλισης** (συνέχεια)
  - Η μέση τιμή αποτελεί το σημείο αναφοράς με το οποίο συγκρίνονται οι υπόλοιπες παρατηρήσεις. Η θέση κάθε παρατήρησης υπολογίζεται από το **τυπικό αποτέλεσμα** (z score) που ορίζεται ως

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

    Το τυπικό αποτέλεσμα δείχνει σε μονάδες τυπικής απόκλισης πόσο διαφέρει μια παρατήρηση πάνω ή κάτω από τη μέση τιμή.

    Παράδειγμα :
    - ➢ Αν το τυπικό αποτέλεσμα μιας τιμής είναι 2, αυτό σημαίνει ότι είναι μεγαλύτερη κατά 2 τυπικές αποκλίσεις από τη μέση τιμή.
    - ➢ Αν το τυπικό αποτέλεσμα μιας τιμής είναι -1, αυτό σημαίνει ότι είναι μικρότερη κατά 1 τυπική απόκλιση από τη μέση τιμή.

22

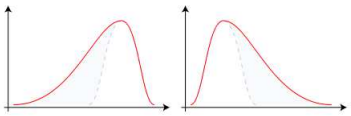
3. Περιγραφική στατιστική ανάλυση

❖ **Συντελεστής ασυμμετρίας (λοξότητα)**

➢ Δηλώνει το βαθμό ασυμμετρίας μιας κατανομής

$$skewness = \frac{\mu_3}{\sigma^3} = \frac{E\left[\left(x - E[x]\right)^3\right]}{\sigma^3}$$

- skewness=0 : Κανονική κατανομή
- skewness>0 : Κατανομή με ουρά προς τα δεξιά
- skewness<0 : Κατανομή με ουρά προς τα αριστερά
- |skewness|>1 : Κατανομή που διαφέρει σημαντικά από την Κανονική.



23

3. Περιγραφική στατιστική ανάλυση

❖ **Συντελεστής κύρτωσης**

➢ Μετρά το βαθμό συγκέντρωσης των τιμών της μεταβλητής γύρω από τη μέση τιμή (πόσο επιρρεπής είναι η κατανομή στην ύπαρξη outliers)

$$kurtosis = \frac{\mu_4}{\sigma^4} - 3 = \frac{E\left[\left(x - E[x]\right)^4\right]}{\sigma^4} - 3$$

- kurtosis = 0 : Κανονική κατανομή (μεσόκυρτη καμπύλη)
- kurtosis > 0 : Κατανομή στην οποία οι παρατηρήσεις συγκεντρώνονται περισσότερο γύρω από τη μέση τιμή σε σχέση με την κανονική κατανομή (λεπτόκυρτη καμπύλη)
- kurtosis < 0 : Κατανομή στην οποία οι παρατηρήσεις συγκεντρώνονται λιγότερο γύρω από τη μέση τιμή σε σχέση με την κανονική (πλατύκυρτη καμπύλη)

24

3. Περιγραφική στατιστική ανάλυση

- **Συναρτήσεις στο Matlab**

  Συντελεστής ασυμμετρίας : skewness(x)

  Συντελεστής κύρτωσης : kurtosis(x)

  Παρατήρηση 1: Εάν το όρισμα x είναι πίνακας, τότε οι συναρτήσεις δίνουν το ζητούμενο αποτέλεσμα για κάθε μία από τις στήλες του x.

  Παρατήρηση 2: Η συνάρτηση kurtosis(x) δεν αφαιρεί το 3 από τον υπολογισμό της κύρτωσης (άρα η κανονική κατανομή έχει κύρτωση=3)

  Παρατήρηση 3 : Οι δύο συντελεστές επηρεάζονται από την ύπαρξη απομονωμένων τιμών (outliers).

25

3. Περιγραφική στατιστική ανάλυση

- **<u>Το διάγραμμα boxplot (ή θηκόγραμμα)</u>**

  Βοηθά στη γραφική απεικόνιση της κατανομής των δεδομένων.

  Παρουσιάζει :

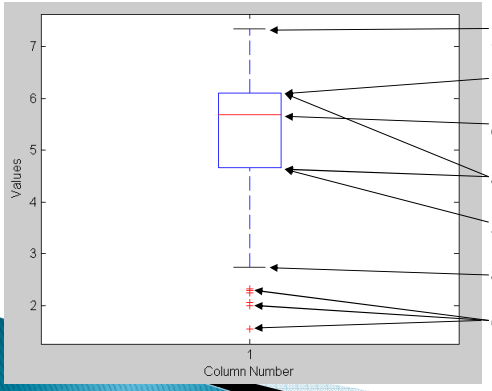
  - το 25ο τεταρτημόριο
  - το 75ο τεταρτημόριο
  - τη διάμεσο
  - το ενδοτεταρτομοριακό εύρος (R)
  - τις απομονωμένες τιμές (outliers)
  - τη μέγιστη τιμή (δεν συμπεριλαμβάνονται οι απομονωμένες τιμές)
  - την ελάχιστη τιμή (δεν συμπεριλαμβάνονται οι απομονωμένες τιμές)

Απομονωμένες τιμές (outliers) ονομάζονται αυτές που απέχουν >1,5R πάνω από το 75ο τεταρτημόριο ή κάτω από το 25ο τεταρτημόριο.
Ακραίες τιμές ονομάζονται αυτές που απέχουν >3 R πάνω από το 75ο τεταρτημόριο ή κάτω από το 25ο τεταρτημόριο.

26

3. Περιγραφική στατιστική ανάλυση

- **<u>Το διάγραμμα boxplot</u>** (συνέχεια)



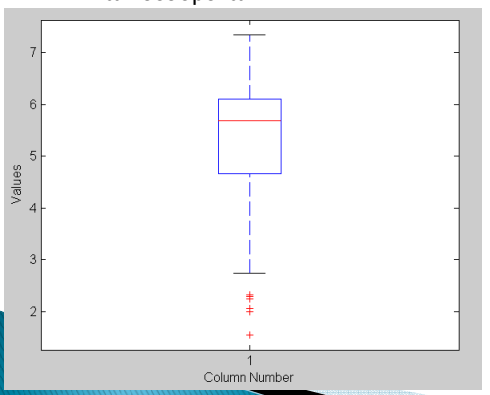
27

3. Περιγραφική στατιστική ανάλυση

- **<u>Το διάγραμμα boxplot</u>** (συνέχεια)

Πληροφορίες που εξάγονται από το διάγραμμα boxplot :

- Η θέση της διαμέσου δείχνει που βρίσκεται μια κεντρική τιμή των δεδομένων
- Το ύψος του κουτιού δίνει μια πρώτη οπτική εκτίμηση της μεταβλητότητας των δεδομένων



- Αν η γραμμή της διαμέσου δε βρίσκεται στο κέντρο του κουτιού, η κατανομή δεν είναι συμμετρική.
- Αν η διάμεσος είναι πιο κοντά στο πάνω άκρο τότε η κατανομή έχει ουρά προς τα αρνητικά.
- Αν η διάμεσος είναι πιο κοντά στο κάτω άκρο τότε η κατανομή έχει ουρά προς τα θετικά.

28

Έλεγχος υποθέσεων για τη μέση τιμή δύο δειγμάτων

- Στη συνέχεια περιγράφουμε τη διαδικασία ελέγχου της διαφοράς των (πληθυσμιακών) μέσων τιμών μ_1 και μ_2 δύο **ανεξάρτητων κανονικών πληθυσμών** χρησιμοποιώντας δύο τυχαία δείγματα, ένα από κάθε πληθυσμό.
- Κατά τον έλεγχο προσδιορίζουμε το διάστημα εμπιστοσύνης της διαφοράς των δύο μέσων τιμών.
- Διακρίνουμε τρεις περιπτώσεις :
  1. Οι διασπορές των δύο πληθυσμών είναι γνωστές.
  2. Οι διασπορές των δύο πληθυσμών είναι άγνωστες αλλά ίσες.
  3. Οι διασπορές των δύο πληθυσμών είναι άγνωστες αλλά διαφορετικές (άνισες).

29

Έλεγχος υποθέσεων για τη μέση τιμή δύο δειγμάτων

Πρώτη περίπτωση : Οι διασπορές των δύο πληθυσμών σ_1 και σ_2 είναι γνωστές

Βήμα 1 : Θέτουμε τη μηδενική υπόθεση H_0: $\mu_1=\mu_2$

Βήμα 2 : Θέτουμε την εναλλακτική υπόθεση H_1: $\mu_1\neq\mu_2$, ($\acute{\eta}$ H_1: $\mu_1>\mu_2$ $\acute{\eta}$ H_1: $\mu_1<\mu_2$).

Βήμα 3 : Θέτουμε $$q=\frac{\left(\overline{x_1}-\overline{x_2}\right)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$

(που ακολουθεί την τυποποιημένη κανονική κατανομή).

Βήμα 4 : Επιλέγουμε το επίπεδο σημαντικότητας a και υπολογίζουμε το αντίστοιχο σημείο $q_{a/2}$ ($\acute{\eta}$ q_a $\acute{\eta}$ $-q_a$).

Βήμα 5 : Υπολογίζουμε την τιμή του q για τα δείγματα.

30

Έλεγχος υποθέσεων για τη μέση τιμή δύο δειγμάτων

Δεύτερη περίπτωση : Οι διασπορές των δύο πληθυσμών είναι άγνωστες αλλά ίσες

Βήμα 1 : Θέτουμε τη μηδενική υπόθεση H_0: $\mu_1=\mu_2$.

Βήμα 2 : Θέτουμε την εναλλακτική υπόθεση H_1: $\mu_1\neq\mu_2$, (ή H_1: $\mu_1>\mu_2$ ή H_1: $\mu_1<\mu_2$).

Βήμα 3 : Θέτουμε $$t=\frac{\left(\overline{x_1}-\overline{x_2}\right)}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

(που ακολουθεί την κατανομή του Student t με n_1+n_2-2 βαθμούς ελευθερίας) όπου s είναι η δειγματική τυπική απόκλιση:

$$s=\sqrt{\frac{\sum_{i=1}^{n_1}\left(x_{1i}-\overline{x_1}\right)^2+\sum_{i=1}^{n_2}\left(x_{2i}-\overline{x_2}\right)^2}{n_1+n_2-2}}$$

31

Έλεγχος υποθέσεων για τη μέση τιμή δύο δειγμάτων

Δεύτερη περίπτωση : Οι διασπορές των δύο πληθυσμών είναι άγνωστες αλλά ίσες (συν.)

Βήμα 4 : Επιλέγουμε το επίπεδο σημαντικότητας a και υπολογίζουμε το αντίστοιχο σημείο $t_{a/2,n1+n2-2}$ (ή $t_{a,n1+n2-2}$ ή $-t_{a,n1+n2-2}$).

Βήμα 5 : Υπολογίζουμε την τιμή t για τα δείγματα.

Βήμα 6 : Απορρίπτουμε τη μηδενική υπόθεση υπέρ της

1. H_1 : $\mu_1\neq\mu_2$ εάν $|t|>t_{a/2,n1+n2-2}$ (αμφίπλευρος έλεγχος)
2. H_1 : $\mu_1>\mu_2$ εάν $t > t_{a,n1+n2-2}$ (μονόπλευρος έλεγχος προς τα δεξιά)
3. H_1 : $\mu_1<\mu_2$ εάν $t<-t_{a,n1+n2-2}$ (μονόπλευρος έλεγχος προς τα αριστερά)

32

Έλεγχος υποθέσεων για τη μέση τιμή δύο δειγμάτων

Τρίτη περίπτωση : Οι διασπορές των δύο πληθυσμών είναι άγνωστες αλλά διαφορετικές (συν.)

όπου s_1 και s_2 είναι οι δειγματικές τυπικές αποκλίσεις των δύο δειγμάτων :

$$s_1^2 = \frac{\sum_{i=1}^{n_1}\left(x_{1i} - \overline{x_1}\right)^2}{n_1 - 1} \qquad s_2^2 = \frac{\sum_{i=1}^{n_2}\left(x_{2i} - \overline{x_2}\right)^2}{n_2 - 1}$$

Βήμα 4 : Επιλέγουμε το επίπεδο σημαντικότητας a και υπολογίζουμε το αντίστοιχο σημείο $t_{a/2,\nu}$ (ή $t_{a,\nu}$ ή $-t_{a,\nu}$).

Βήμα 5 : Υπολογίζουμε την τιμή του t για τα δείγματα.

Βήμα 6 : Απορρίπτουμε τη μηδενική υπόθεση υπέρ της

1. $H_1 : \mu_1 \neq \mu_2$ εάν $|t| > t_{a/2,\nu}$ (αμφίπλευρος έλεγχος)
2. $H_1 : \mu_1 > \mu_2$ εάν $t > t_{a,\nu}$ (μονόπλευρος έλεγχος προς τα δεξιά)
3. $H_1 : \mu_1 < \mu_2$ εάν $t < -t_{a,\nu}$ (μονόπλευρος έλεγχος προς τα αριστερά)

33

Έλεγχος υποθέσεων για τη μέση τιμή δύο δειγμάτων

Τρίτη περίπτωση : Οι διασπορές των δύο πληθυσμών είναι άγνωστες αλλά διαφορετικές

Βήμα 1 : Θέτουμε τη μηδενική υπόθεση H_0: $\mu_1 = \mu_2$.

Βήμα 2 : Θέτουμε την εναλλακτική υπόθεση H_1: $\mu_1 \neq \mu_2$, (ή H_1: $\mu_1 > \mu_2$ ή H_1: $\mu_1 < \mu_2$).

Βήμα 3 : Θέτουμε $t = \dfrac{\left(\overline{x_1} - \overline{x_2}\right)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$

που ακολουθεί την κατανομή του Student t με ν βαθμούς ελευθερίας :

$$\nu = \frac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \dfrac{\left(\dfrac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

34

One-way ANOVA

- Πολλές φορές χρειάζεται να συγκρίνουμε τις μέσες τιμές **περισσότερων των δύο πληθυσμών**.
- Για παράδειγμα, επιθυμούμε να συγκρίνουμε τα αποτελέσματα τεσσάρων διαφορετικών θεραπειών που μειώνουν τα επίπεδα της χοληστερόλης συγκρίνοντας τις μέσες τιμές των τεσσάρων θεραπειών.
- Με άλλα λόγια, επιθυμούμε να εφαρμόσουμε έναν **έλεγχο μηδενικής υπόθεσης ότι k, k>2, ανεξάρτητοι πληθυσμοί έχουν ίσες μέσες τιμές** λαμβάνοντας μία ομάδα παρατηρήσεων (δείγμα) από κάθε πληθυσμό.
- Η μέθοδος που χρησιμοποιούμε για το σκοπό αυτό είναι η **Ανάλυση Διασποράς (ANOVA - ANalysis Of Variance).**
- Η μέθοδος εξετάζει τη μεταβλητότητα των παρατηρήσεων εντός των ομάδων και τη μεταβλητότητα των δειγματικών μέσων τιμών και καταλήγει σε συμπεράσματα για την ισότητα των πληθυσμιακών μέσων τιμών.

35

One-way ANOVA

- Εάν οι δειγματικές μέσες τιμές διαφέρουν περισσότερο από ό,τι αναμένεται με βάση τη μεταβλητότητα των παρατηρήσεων εντός των ομάδων, το συμπέρασμα είναι ότι οι πληθυσμιακές μέσες τιμές δεν είναι ίσες.
- Η **One-way ANOVA** είναι η περίπτωση της ANOVA που χρησιμοποιεί τις τιμές μιας μεταβλητής για το διαχωρισμό των ομάδων.
- Η μεταβλητή που χρησιμοποιείται για το διαχωρισμό των ομάδων ονομάζεται **παράγοντας (factor)**.

36

2. One-way ANOVA

Γιατί η ANOVA εξετάζει τη διασπορά;

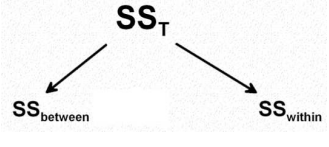
- Γιατί η ANOVA εξετάζει τη διασπορά ενώ ενδιαφερόμαστε για την εφαρμογή ελέγχων υποθέσεων για τη μέση τιμή;
- Τα συμπεράσματα για την πληθυσμιακή μέση τιμή βασίζονται στην εξέταση της διασποράς των δειγματικών μέσων τιμών.
- Με την ANOVA συγκρίνουμε την παρατηρούμενη διασπορά των δειγματικών μέσων τιμών ως προς την αναμενόμενη διασπορά με βάση τη **μηδενική υπόθεση ότι: «Όλες οι k πληθυσμιακές μέσες τιμές είναι ίσες.»**
- Εάν η διασπορά των δειγματικών μέσων τιμών διαφέρει από ό,τι αναμένεται με βάση τη μηδενική υπόθεση, έχουμε μία ένδειξη ότι αυτή η διαφορά οφείλεται στο γεγονός ότι μερικές (τουλάχιστον δύο) από τις ομάδες δεν έχουν την ίδια πληθυσμιακή μέση τιμή.

37

2. One-way ANOVA

Πώς λειτουργεί η ANOVA;

- Η ANOVA πρώτα εξετάζει πόσο διαφέρουν οι παρατηρήσεις εντός των ομάδων. Ως αποτέλεσμα προκύπτει η αναμενόμενη (με βάση τη μηδενική υπόθεση) διασπορά των δειγματικών μέσων τιμών των ομάδων.
- Στη συνέχεια, εξετάζει πόσο διαφέρουν μεταξύ τους οι δειγματικές μέσες τιμές των ομάδων.
- Εάν οι δειγματικές μέσες τιμές διαφέρουν μεταξύ τους περισσότερο από ό,τι αναμένεται, η μηδενική υπόθεση απορρίπτεται.



$$SS_T = \sum_{j=1}^{p}\sum_{i=1}^{n_j}\left(X_{ij}-\bar{X}\right)^2 = \sum_{j=1}^{p}\sum_{i=1}^{n_j}\left[\left(X_{ij}-\bar{X}_j\right)+\left(\bar{X}_j-\bar{X}\right)\right]^2$$

$$SS_{W/in} = \sum_{j=1}^{p}\sum_{i=1}^{n_j}\left(X_{ij}-\bar{X}_j\right)^2$$

$$SS_{Betw} = \sum_{j=1}^{p} n_j\left(\bar{X}_j-\bar{X}\right)^2$$

38

One-way ANOVA

Μεταβλητότητα εντός των ομάδων (within groups)

- Η εντός των ομάδων μεταβλητότητα δείχνει πόσο διαφέρουν οι παρατηρήσεις εντός των k ομάδων.
- Μία από τις παραδοχές της μεθόδου ANOVA είναι ότι όλες οι ομάδες προέρχονται από πληθυσμούς με ίσες πληθυσμιακές διασπορές.
- Αυτή η υπόθεση οδηγεί στην εκτίμηση της μέσης διασποράς κάθε ομάδας και στη συνέχεια στην εκτίμηση μιας μέσης τιμής των τιμών αυτών, η οποία αποτελεί τη **μεταβλητότητα εντός των ομάδων**.

39

One-way ANOVA

Μεταβλητότητα μεταξύ ομάδων (between groups)

- Καθεμία από τις ομάδες έχει μία δειγματική μέση τιμή (δηλαδή, έχουμε k δειγματικές μέσες τιμές).
- Υπολογίζουμε την τυπική απόκλιση των δειγματικών μέσων τιμών.
- Με βάση τη μηδενική υπόθεση ότι «όλες οι ομάδες προέρχονται από πληθυσμούς που έχουν ίσες πληθυσμιακές μέσες τιμές», η τυπική απόκλιση των δειγματικών μέσων τιμών μας δείχνει πώς ποικίλουν οι δειγματικές μέσες τιμές του ίδιου πληθυσμού.
- Η τυπική απόκλιση των δειγματικών μέσων τιμών αποτελεί μια εκτίμηση του τυπικού σφάλματος της μέσης τιμής.
- Το τετράγωνο της τυπικής απόκλισης είναι η **εκτίμηση της μεταβλητότητας μεταξύ των ομάδων**.

40

One-way ANOVA

Ανάλυση μεταβλητότητας

- Η μεταβλητότητα μεταξύ ομάδων είναι το αποτέλεσμα :
  - της μεταβλητότητας των παρατηρήσεων εντός των ομάδων και
  - της μεταβλητότητας των μέσων τιμών των πληθυσμών.
- Η μεταβλητότητα εντός των ομάδων δεν εξαρτάται από το εάν η μηδενική υπόθεση είναι αληθής.
- Η μεταβλητότητα μεταξύ ομάδων αποτελεί εκτίμηση της μεταβλητότητας εντός των ομάδων μόνο όταν η μηδενική υπόθεση είναι αληθής.
- Εάν η μηδενική υπόθεση δεν είναι αληθής, τότε η μεταβλητότητα μεταξύ των ομάδων διαφέρει σημαντικά από τη μεταβλητότητα εντός των ομάδων.

41

One-way ANOVA

Απόφαση

- Η απόφαση για τη μηδενική υπόθεση βασίζεται στη σύγκριση της μεταβλητότητας μεταξύ των ομάδων και της μεταβλητότητας εντός των ομάδων.
- Ο λόγος

$$F = \frac{\text{Μεταβλητοτητα μεταξυ ομαδων}}{\text{Μεταβλητοτητα εντος ομαδων}}$$

ακολουθεί την F-κατανομή :

$$f(x) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)}\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} \frac{x^{\frac{\nu_1-2}{2}}}{\left[1+\left(\frac{\nu_1}{\nu_2}\right)x\right]^{\frac{\nu_1+\nu_2}{2}}}$$

όπου ν_1 και ν_2 είναι οι βαθμοί ελευθερίας του αριθμητή και του παρονομαστή.

42

Διαγράμματα κανονικότητας και χ^2

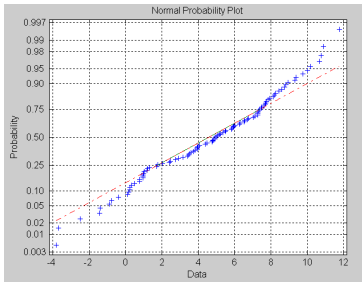
- **Τα διαγράμματα κανονικότητας (normal probability plots)**
  - ✓ Το **διάγραμμα κανονικότητας** είναι ένα χρήσιμο γράφημα για να εκτιμάται εάν τα δεδομένα προέρχονται από κανονική κατανομή.
  - ✓ Όπως έχουμε δει, πολλές στατιστικές διαδικασίες στηρίζονται στην παραδοχή ότι η κατανομή δεδομένων είναι κανονική.
  - ✓ Το διάγραμμα κανονικότητας **μπορεί να παρέχει κάποια διαβεβαίωση ότι η παραδοχή της κανονικότητας δεν παραβιάζεται ή να παρέχει έγκαιρη προειδοποίηση τυχόν απόκλισης** από την παραδοχή της κανονικότητας.
  - ✓ Εντολή Matlab : **normplot(x)**
    Παράδειγμα (κανονικότητας) :

```
x = normrnd(5,4,100,1);
normplot(x)
```

43

Διαγράμματα κανονικότητας και χ^2

- **Τα διαγράμματα κανονικότητας** (συν.)



  - ✓ Το **διάγραμμα κανονικότητας** περιλαμβάνει τρία γραφικά στοιχεία:
    - Τα "+" δείχνουν την εμπειρική πιθανότητα ως προς την τιμή των δεδομένων κάθε σημείου στο δείγμα.
    - Η συμπαγής γραμμή συνδέει το 25ο και το 75ο εκατοστημόριο των δεδομένων και αναπαριστά μία γραμμική παλινδρόμηση (μη ευαίσθητη στα ακραία σημεία του δείγματος).
    - Η διακεκομμένη γραμμή επεκτείνει τη συμπαγή γραμμή στα άκρα του δείγματος.

44

Έλεγχοι υποθέσεων κανονικότητας

- **Το Kolmogorov-Smirnov τεστ για κατανομή αναφοράς**
  - Το τεστ ελέγχει (στην περίπτωση αμφίπλευρου ελέγχου) :
    - τη **μηδενική υπόθεση ότι «το υπό εξέταση δείγμα προέρχεται από συγκεκριμένη κατανομή αναφοράς»** (δηλαδή, η συνάρτηση κατανομής του δείγματος συμπίπτει με τη συνάρτηση της κατανομής αναφοράς) έναντι
    - της εναλλακτικής υπόθεσης ότι «το δείγμα δεν προέρχεται από τη συγκεκριμένη κατανομή αναφοράς» (δηλαδή, η συνάρτηση κατανομής του δείγματος διαφέρει από τη συνάρτηση της κατανομής σε τουλάχιστον ένα σημείο).
  - Το τεστ συγκρίνει την εμπειρική αθροιστική κατανομή του δείγματος με την αθροιστική κατανομή αναφοράς. Με άλλα λόγια, το Kolmogorov-Smirnov τεστ συγκρίνει την αναλογία τιμών που είναι μικρότερες από x με την αντίστοιχη αναμενόμενη αναλογία τιμών με βάση την κατανομή αναφοράς.

45

Έλεγχοι υποθέσεων χ^2

Το χ-τετράγωνο τεστ για ανεξαρτησία μεταβλητών

Βήμα 1 : Θέτουμε τη **μηδενική** υπόθεση H_0: "**Δεν υπάρχει εξάρτηση μεταξύ των υπό μελέτη μεταβλητών**"

Βήμα 2 : Η εναλλακτική υπόθεση H_1 είναι: "Υπάρχει εξάρτηση μεταξύ των υπό μελέτη μεταβλητών".

Βήμα 3 : Εκτιμούμε τις θεωρητικά αναμενόμενες συχνότητες E_i σύμφωνα με τη μηδενική υπόθεση.

Βήμα 4 : Θέτουμε $x^2 = \sum_{i=1}^{r \cdot c} \frac{(O_i - E_i)^2}{E_i}$ (το οποίο ακολουθεί τη χ-τετράγωνο κατανομή με df=(c-1)(r-1) βαθμούς ελευθερίας), όπου O_i είναι οι παρατηρηθείσες συχνότητες.

Βήμα 5a : Υπολογίζουμε την τιμή του x^2 και (από τη χ-τετράγωνο cdf υπολογίζουμε) την πιθανότητα P που αντιστοιχεί στο x^2.

Βήμα 5b : Επιλέγουμε το επίπεδο σημαντικότητας a (π.χ. 5%).

Βήμα 6 : Απορρίπτουμε τη μηδενική υπόθεση H_0 υπέρ της εναλλακτικής Η[illegible] P>1-a.

46

Έλεγχοι υποθέσεων χ^2

Το χ-τετράγωνο τεστ για έλεγχο ταιριάσματος με κατανομή

Βήμα 1 : Θέτουμε τη **μηδενική υπόθεση** H_0: "**Τα δεδομένα του δείγματος ΑΚΟΛΟΥΘΟΥΝ μία θεωρητική κατανομή με γνωστές αναμενόμενες συχνότητες E_i των n κατηγοριών της μεταβλητής**."

Βήμα 2 : Η εναλλακτική υπόθεση H_1 είναι: "Τα δεδομένα του δείγματος δεν ακολουθούν τη θεωρητική κατανομή με γνωστές αναμενόμενες συχνότητες E_i των n κατηγοριών της μεταβλητής."

Βήμα 3 : Θέτουμε $x^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$ (που ακολουθεί τη χ-τετράγωνο κατανομή με df=n-1 βαθμούς ελευθερίας), όπου O_i είναι οι παρατηρηθείσες συχνότητες.

Βήμα 4a : Υπολογίζουμε την τιμή του x^2 και (από τη χ-τετράγωνο cdf υπολογίζουμε) την πιθανότητα P που αντιστοιχεί στο x^2.

Βήμα 4b : Επιλέγουμε το επίπεδο σημαντικότητας a (π.χ. 5%).

Βήμα 5 : Απορρίπτουμε την H_0 υπέρ της H_1 εάν P>1-a.

47

Έλεγχοι υποθέσεων χ^2

Τι πρέπει να προσέχουμε όταν εφαρμόζουμε τα x^2 τεστ;

- Το χ-τετράγωνο τεστ παρέχει **αναξιόπιστα αποτελέσματα** όταν
  - Η ελάχιστη θεωρητικά αναμενόμενη τιμή είναι μικρότερη από 1.
  - Περισσότερες από το 20% των θεωρητικά αναμενόμενων συχνοτήτων είναι μικρότερες από 5.
  - Έχουμε λιγότερες από 20 παρατηρήσεις.
  - Στις περιπτώσεις των 2x2 πινάκων συνάφειας με 20-40 παρατηρήσεις, έχουμε τουλάχιστον μία από τις θεωρητικά αναμενόμενες τιμές μικρότερη από 5.
  - ❖ Όταν ισχύει κάτι από τα παραπάνω, **δεν εφαρμόζουμε το χ-τετράγωνο τεστ**.
- Τονίζεται επίσης ότι τα x^2 τεστ **εφαρμόζονται στις αρχικές συχνότητες** και όχι σε λόγους ή ποσοστά που προκύπτουν από τις αρχικές συχνότητες.

48

Έλεγχος υποθέσεων για κανονικότητα και χ^2

- Ο έλεγχος x^2 χρησιμοποιείται όταν οι μεταβλητές είναι **ποιοτικές**.
- Όταν επιθυμούμε να προσδιορίσουμε εάν υπάρχει **συσχέτιση** μεταξύ δύο ποιοτικών μεταβλητών ή όταν επιθυμούμε να εξετάσουμε εάν ένα σύνολο δεδομένων **προέρχεται από μια καθορισμένη κατανομή**, τότε εφαρμόζουμε το x^2 έλεγχο.
- Σε καθεμία από τις δύο περιπτώσεις, υπάρχουν δύο κατηγορίες πληροφορίας που χρειαζόμαστε :
  1. την πραγματική συχνότητα κάθε «κελιού» του πίνακα συνάφειας (πραγματική ή παρατηρηθείσα συχνότητα) και
  2. την αναμενόμενη συχνότητα κάθε «κελιού», η οποία προέρχεται είτε από τη θεωρία ή μέσω μιας γνωστής σχέσης.

49

Τι πρέπει να προσέχουμε όταν εφαρμόζουμε τα x^2 τεστ;

- Το χ-τετράγωνο τεστ παρέχει **αναξιόπιστα αποτελέσματα** όταν:
  - Η ελάχιστη θεωρητικά αναμενόμενη τιμή είναι μικρότερη από 1.
  - Περισσότερες από το 20% των θεωρητικά αναμενόμενων συχνοτήτων είναι μικρότερες από 5.
  - Έχουμε λιγότερες από 20 παρατηρήσεις.
  - Στις περιπτώσεις των 2x2 πινάκων συνάφειας με 20-40 παρατηρήσεις, έχουμε τουλάχιστον μία από τις θεωρητικά αναμενόμενες τιμές μικρότερη από 5.
  - ❖ Όταν ισχύει κάτι από τα παραπάνω, **δεν εφαρμόζουμε το χ-τετράγωνο τεστ**.
- Τονίζουμε επίσης το ότι τα x^2 τεστ **εφαρμόζονται στις αρχικές συχνότητες** και όχι σε λόγους ή ποσοστά που προκύπτουν από τις αρχικές συχνότητες.

50

Η διόρθωση τουYates

- Η διόρθωση αυτή **εφαρμόζεται στους 2x2 πίνακες συνάφειας** όταν τουλάχιστον ένα κελί του πίνακα έχει αναμενόμενη συχνότητα μικρότερη από 5.
- Η διόρθωση του Yates υπολογίζεται εύκολα και τροποποιεί τον τύπο του x-τετράγωνο ως εξής :

$$x^2 = \sum_{i=1}^{r \cdot c} \frac{(O_i - E_i - 0.5)^2}{E_i}$$

- Η διόρθωση μειώνει την τιμή του x-τετράγωνο και έτσι αυξάνει την p-τιμή του.
- Η διόρθωση εμποδίζει την υπερεκτίμηση της στατιστικής σημαντικότητας για μικρού πλήθους δεδομένα.
- Η διόρθωση δεν προκαλεί παρά μικρή διαφορά όταν οι μετρήσεις είναι πολλές.

51

2. Μη παραμετρικοί έλεγχοι υποθέσεων

- Οι έλεγχοι υποθέσεων διακρίνονται σε :
  - ✓ **Παραμετρικοί** έλεγχοι
  - ✓ **Μη παραμετρικοί** έλεγχοι
- Τα z-τεστ και t-τεστ που έχουμε δει μέχρι τώρα βασίζονται στην παραδοχή ότι οι παρατηρήσεις:
  - ακολουθούν την κανονική κατανομή ή
  - τουλάχιστον, προσεγγιστικά ακολουθούν την κανονική κατανομή ή
  - ακολουθούν την κανονική κατανομή μετά από κάποιο μετασχηματισμό (π.χ. λογαριθμικό μετασχηματισμό)
- Τα z-τεστ και t-τεστ είναι **παραμετρικοί έλεγχοι.** Αυτοί είναι οι έλεγχοι των οποίων η εφαρμογή βασίζεται στην ύπαρξη παραμέτρων κατανομών (π.χ. N(0,1), t(df), F(df))

52

Μη παραμετρικοί έλεγχοι υποθέσεων

- Αντίθετα, **οι μη παραμετρικοί έλεγχοι** (**non-parametric tests** ή **distribution-free tests**) είναι οι έλεγχοι των οποίων η εφαρμογή δεν απαιτεί υποθέσεις για τις παραμέτρους κατανομών.
- Οι μη παραμετρικές διαδικασίες είναι **λιγότερο ισχυρές** από τους ελέγχους που έχουν σχεδιαστεί με βάση συγκεκριμένη κατανομή.
- Είναι καλύτερα να χρησιμοποιούμε ένα ισχυρότερο εργαλείο όταν οι παραδοχές του ικανοποιούνται, έστω και προσεγγιστικά, από το να χρησιμοποιούμε ένα λιγότερο ισχυρό εργαλείο με λιγότερες παραδοχές.
- Για παράδειγμα, όταν μια μη κανονική κατανομή μπορεί να μετασχηματιστεί σε κανονική με λογαριθμικό μετασχηματισμό, προτιμούμε να χρησιμοποιούμε τις μετασχηματισμένες παρατηρήσεις με μία παραμετρική μέθοδο από το να χρησιμοποιούμε τις αρχικές παρατηρήσεις με μία μη παραμετρική μέθοδο.

53

Μη παραμετρικοί έλεγχοι υποθέσεων

- Οι **μη παραμετρικοί έλεγχοι** δεν χρησιμοποιούν τις πραγματικές τιμές των παρατηρήσεων αλλά **το πλήθος των παρατηρήσεων** (π.χ. Sign test) **ή την τάξη (θέση) κάθε παρατήρησης** στο σύνολο όλων των δεδομένων (π.χ.Wilcoxon signed rank test, Mann-Whitney test, Wilcoxon rank sum test, Kruskal-Wallis test).
- Οι μη παραμετρικοί έλεγχοι εφαρμόζονται **σε ποσοτικά χαρακτηριστικά** στις περιπτώσεις όπου :
  - η κατανομή είναι προφανώς μη κανονική ή
  - η κατανομή είναι άγνωστη ή
  - το πλήθος των παρατηρήσεων είναι μικρό

ή γενικά **όταν είναι αδύνατο να εφαρμοστεί παραμετρικός έλεγχος**.

54

Μη παραμετρικοί έλεγχοι υποθέσεων

- Οι μη παραμετρικοί έλεγχοι μπορούν να εφαρμοστούν σε κάθε κατανομή. Ωστόσο, δίνουν καλύτερα αποτελέσματα:
  - όταν οι συγκρινόμενες ομάδες έχουν παρόμοιες κατανομές
  - όταν η κατανομή του υπό μελέτη χαρακτηριστικού είναι συνεχής.
- Η εφαρμογή των μη-παραμετρικών ελέγχων είναι ευκολότερη και απλούστερη από την εφαρμογή των αντίστοιχων παραμετρικών ελέγχων. Ωστόσο, ο υπολογισμός του "ορίου αξιοπιστίας" μιας διαφοράς που προκύπτει από τους μη παραμετρικούς ελέγχους είναι πολύ δύσκολος.

55

Μη παραμετρικοί έλεγχοι υποθέσεων

- Αντίστοιχοι μη παραμετρικοί έλεγχοι των κατά ζεύγη t-τεστ:
  - ✓ Το **Τεστ Προσήμου** ή **Προσημικός Έλεγχος (Sign test)**
  - ✓ Το **Βαθμολογικό Προσημικό Τεστ Wilcoxon (Wilcoxon signed-rank τεστ)**

  Εφαρμόζονται όταν η κατανομή είναι μη κανονική.
- Αντίστοιχοι μη παραμετρικοί έλεγχοι του t-τεστ δύο ανεξάρτητων δειγμάτων:
  - ✓ Το **Mann-Whitney τεστ**
  - ✓ Το **Βαθμολογικό Τεστ Wilcoxon (Wilcoxon Rank Sum τεστ)**

  Εφαρμόζονται όταν οι κατανομές είναι μη κανονικές αλλά ίδιες.
- Αντίστοιχος μη παραμετρικός έλεγχος της One-way ANOVA :
  - ✓ Το **Kruskal-Wallis τεστ**

  Εφαρμόζεται όταν οι κατανομές των δύο ή περισσότερων δειγμάτων είναι μη κανονικές αλλά ίδιες.

56

2. Μη παραμετρικοί έλεγχοι υποθέσεων

Χαρακτηριστικά	Έλεγχος	
	Πρώτη επιλογή	Δεύτερη επιλογή
Ποιοτικά	Έλεγχος χ-τετράγωνο	-
Ποσοτικά με κανονική κατανομή	t-τεστ	Mann-Whitney τεστ ή Wilcoxon rank sum τεστ
	Κατά ζεύγη t-τεστ	Sign test ή Wilcoxon signed-rank τεστ
	One-way ANOVA	Kruskal-Wallis τεστ
Ποσοτικά με μη κανονική ή άγνωστη κατανομή	Mann-Whitney τεστ ή Wilcoxon rank sum τεστ	Έλεγχος χ-τετράγωνο
	Sign test ή Wilcoxon signed-rank τεστ	
	Kruskal-Wallis τεστ	

57

2. Μη παραμετρικοί έλεγχοι υποθέσεων

- **Το Wilcoxon Rank Sum τεστ**
  - ✓ Εφαρμόζεται όταν :
    - θέλουμε να συγκρίνουμε (δηλαδή να διαπιστώσουμε εάν υπάρχει οποιαδήποτε διαφορά μεταξύ τους) δύο **ανεξάρτητες** ομάδες παρατηρήσεων
    - οι δύο ομάδες έχουν **μη κανονικές** κατανομές ή οι κατανομές τους είναι **άγνωστες** (δηλαδή το t-τεστ δεν μπορεί να εφαρμοστεί).
  - ✓ Υποθέτουμε ότι οι δύο ομάδες παρατηρήσεων προέρχονται από συνεχείς κατανομές οποιασδήποτε μορφής που είναι **ίδιες** με την εξαίρεση πιθανόν μιας ολίσθησης.

58

Μη παραμετρικοί έλεγχοι υποθέσεων

- **Το Wilcoxon Rank Sum τεστ**
  1. Θέτουμε τη **μηδενική υπόθεση H_0: ΔΕΝ ΥΠΑΡΧΕΙ καμία διαφορά μεταξύ των δύο ομάδων** (προέρχονται από κατανομές με ίσες διαμέσους)**.**
  2. Θέτουμε την **εναλλακτική υπόθεση H_1 : Υπάρχει διαφορά μεταξύ των δύο ομάδων** (δεν προέρχονται από κατανομές με ίσες διαμέσους)**.**
  3. **Ταξινομούμε** όλες τις παρατηρήσεις (και των δύο ομάδων n_1 και n_2).
     - ✓ Τις ταξινομούμε από τη μικρότερη προς τη μεγαλύτερη (ή το αντίθετο).
     - ✓ Η τάξη κάθε παρατήρησης είναι η θέση της στην ταξινομημένη λίστα, ξεκινώντας από το 1 για τη μικρότερη παρατήρηση.
     - ✓ Όλες οι ίσες τιμές ταξινομούνται με τη μέση τιμή των τάξεων που καταλαμβάνουν.

59

2. Μη παραμετρικοί έλεγχοι υποθέσεων

- **Το Wilcoxon Rank Sum Τεστ** (συν.)
  4. Υπολογίζουμε **το άθροισμα των τάξεων** κάθε ομάδας.
  5. Η Wilcoxon στατιστική W είναι **το μικρότερο** από τα δύο αθροίσματα.
  6. Επιλέγουμε το επίπεδο σημαντικότητας a.
  7. **Υπολογίζουμε την p-τιμή** (από τους πίνακες Wilcoxon) που αντιστοιχεί στη W στατιστική ή αλλιώς την κρίσιμη τιμή T_c.
  8. Απορρίπτουμε τη μηδενική υπόθεση υπέρ της εναλλακτικής εάν $p<a$ ή αλλιώς εάν $W<T_c$.

Παρατηρήσεις

α. Για n_1, n_2 >10, η στατιστική ακολουθεί προσεγγιστικά κανονική κατανομή με μέση τιμή $\frac{n_1 \cdot (n_1 + n_2 + 1)}{2}$ και τυπική απόκλιση $\sqrt{\frac{n_1 \cdot n_2 \cdot (n_1 + n_2 + 1)}{12}}$

60

2. Μη παραμετρικοί έλεγχοι υποθέσεων

- **Το Mann-Whitney Τεστ** (συν.)
  1. Θέτουμε τη **μηδενική υπόθεση H_0: ΔΕΝ υπάρχει καμία διαφορά μεταξύ των δύο ομάδων.**
  2. Θέτουμε την **εναλλακτική υπόθεση H_1 : Υπάρχει διαφορά μεταξύ των δύο ομάδων.**
  3. Για κάθε παρατήρηση του πρώτου δείγματος, **καταμετρούμε το πλήθος** των παρατηρήσεων του **δεύτερου** δείγματος **που είναι μικρότερες από αυτή**.
     - ✓ Προσθέτουμε το 1/2 για κάθε παρατήρηση του δευτέρου δείγματος που είναι ίση με την (υπό μελέτη) παρατήρηση του πρώτου δείγματος.
  4. Για κάθε παρατήρηση του δεύτερου δείγματος, **καταμετρούμε το πλήθος** των παρατηρήσεων του πρώτου δείγματος **που είναι μικρότερες από αυτή.**
     - ✓ Προσθέτουμε το 1/2 για κάθε παρατήρηση του πρώτου δείγματος που είναι ίση με την (υπό μελέτη) παρατήρηση του δεύτερου δείγματος.

61

Μη παραμετρικοί έλεγχοι υποθέσεων

- **Το Mann-Whitney Τεστ** (συν.)
  5. **Υπολογίζουμε το άθροισμα των καταμετρήσεων** για κάθε ομάδα.
  6. Η Mann-Whitney στατιστική U είναι το **μικρότερο από τα δύο αθροίσματα**.
  7. Επιλέγουμε το επίπεδο σημαντικότητας a.
  8. **Υπολογίζουμε την p-τιμή** (από τους Mann-Whitney πίνακες) που αντιστοιχεί στην U στατιστική.
  9. Απορρίπτουμε τη μηδενική υπόθεση υπέρ της εναλλακτικής εάν p<a.

62

Μη παραμετρικοί έλεγχοι υποθέσεων

- **Το Τεστ Πρόσημου (Sign Test)**
  - ✓ Εφαρμόζεται όταν :
    - συγκρίνουμε δύο **μη ανεξάρτητες** ομάδες παρατηρήσεων
    - οι **διαφορές** των δύο ομάδων προέρχονται από **οποιαδήποτε συνεχή κατανομή**.
  - ✓ Οι ομάδες παρατηρήσεων πρέπει να έχουν το **ίδιο πλήθος παρατηρήσεων**.
  - ✓ Βασίζεται στο **πρόσημο των διαφορών**.

63

Μη παραμετρικοί έλεγχοι υποθέσεων

- **Το Τεστ Πρόσημου** (συν.)
  1. Θέτουμε τη **μηδενική υπόθεση H_0: Οι διαφορές μεταξύ των ζευγών των αντίστοιχων δειγμάτων των δύο ομάδων προέρχονται από κατανομή με μηδενική διάμεσο.**
  2. Θέτουμε την **εναλλακτική υπόθεση H_1 :** Οι διαφορές μεταξύ των ζευγών των αντίστοιχων δειγμάτων των δύο ομάδων **ΔΕΝ** προέρχονται από κατανομή με μηδενική διάμεσο.
  3. Σύμφωνα με τη μηδενική υπόθεση, το πλήθος των αρνητικών διαφορών ισούται με το πλήθος των θετικών διαφορών (ζεύγη με μηδενικές διαφορές αγνοούνται).
  4. Εάν έχουμε δείγματα με πλήθος n, το πλήθος των θετικών διαφορών ακολουθεί τη διωνυμική κατανομή B(n,1/2).

  - ❖ **Σημείωση :** Η υπόθεση μηδενικής διαμέσου για τη διαφορά δεν είναι ισοδύναμη με την υπόθεση ίσων διαμέσων για τα δύο δείγματα. Το Τεστ Πρόσημου ελέγχει την πρώτη υπόθεση, όχι τη δεύτερη.

64

Μη παραμετρικοί έλεγχοι υποθέσεων

- **Το Τεστ Πρόσημου** (συν.)
  5. Η στατιστική είναι το πλήθος των παρατηρήσεων με θετική διαφορά.
  6. Επιλέγουμε το επίπεδο σημαντικότητας a.
  7. **Βρίσκουμε την p-τιμή** που αντιστοιχεί στη στατιστική χρησιμοποιώντας τη διωνυμική κατανομή.
  8. Απορρίπτουμε τη μηδενική υπόθεση υπέρ της εναλλακτικής εάν p<a.

65

2. Μη παραμετρικοί έλεγχοι υποθέσεων

- **Το Kruskal-Wallis Τεστ**
  - ✓ Εφαρμόζεται όταν :
    - συγκρίνουμε **δύο ή περισσότερες ανεξάρτητες** ομάδες παρατηρήσεων
    - οι δύο ή περισσότερες ομάδες παρατηρήσεων έχουν **μη κανονικές** κατανομές ή οι κατανομές τους είναι **άγνωστες** (δηλαδή όταν η one-way ANOVA δεν μπορεί να εφαρμοστεί).
  - ✓ Υποθέτουμε ότι οι ομάδες παρατηρήσεων προέρχονται από συνεχείς (οποιεσδήποτε) κατανομές που είναι **παρόμοιες** (παρόμοια σχήματα) με την εξαίρεση πιθανόν κάποιας ολίσθησης.

66

2. Μη παραμετρικοί έλεγχοι υποθέσεων

- **Το Kruskal-Wallis Τεστ** (συν.)
  1. Θέτουμε τη **μηδενική υπόθεση H_0: ΔΕΝ υπάρχει καμία διαφορά μεταξύ των ομάδων των παρατηρήσεων** (προέρχονται από τις ίδιες κατανομές)**.**
  2. Θέτουμε την **εναλλακτική υπόθεση H_1 : Υπάρχει διαφορά μεταξύ των ομάδων των παρατηρήσεων** (τουλάχιστον μία δεν προέρχεται από την ίδια κατανομή με τις υπόλοιπες)**.**
  3. **Ταξινομούμε** όλες τις παρατηρήσεις (από όλες τις ομάδες μαζί).
     - ✓ Τις κατατάσσουμε από τη μικρότερη προς τη μεγαλύτερη.
     - ✓ Η τάξη κάθε παρατήρησης είναι η θέση της σε αυτή την ταξινομημένη λίστα, ξεκινώντας από την τάξη 1 για τη μικρότερη παρατήρηση.
     - ✓ Όλες οι ίσες τιμές λαμβάνουν τη μέση τιμή των τάξεων που καταλαμβάνουν.

67

Μη παραμετρικοί έλεγχοι υποθέσεων

- **Το Kruskal-Wallis Τεστ** (συν.)
  4. **Υπολογίζουμε το άθροισμα των τετραγώνων (στηλών και σφάλματος) με βάση τις τάξεις** (δηλαδή One-way ANOVA με βάση τις τάξεις).
  5. Η Kruskal-Wallis στατιστική Η είναι :

$$H = df \frac{SS_{columns}}{SS_{error}}$$

  6. Όταν τα μεγέθη των δειγμάτων είναι μεγάλα (τουλάχιστον 5 παρατηρήσεις ανά δείγμα) και όλοι οι n πληθυσμοί έχουν την ίδια συνεχή κατανομή, η Η ακολουθεί προσεγγιστικά τη χ-τετράγωνο κατανομή με n-1 βαθμούς ελευθερίας.
  7. Βρίσκουμε **την p-τιμή** που αντιστοιχεί στην Η στατιστική.
  8. Καθορίζουμε το επίπεδο σημαντικότητας a.
  9. **Απορρίπτουμε τη μηδενική υπόθεση υπέρ της εναλλακτικής εάν p<a**.

68

Βιβλιογραφία

- 1) Γ. Βλαχόπουλος, Κ. Κουτσογιάννης, "Βιοστατιστική. Εφαρμογή με το SPSS και το R- Project", Εκδόσεις Αλγόριθμος, Πάτρα 2012
- 2) Ι. Αποστολάκης, Α. Καστανιά, Χρ. Πιερράκου, "Στατιστική επεξεργασια Δεδομενων στην Υγεία", Εκδόσεις Παπαζήση, Αθήνα 2003
- 3) W. Daniel, " Biostatistics, a foundation for analysis in the health sciences", Willey Series in Probability and Statistics, 2005
- 4) G.Van Belle, L. Fisher, P.Heagerty, T. Lumley, "Biostatistics, A methodology for the health sciences", Willey Series in Probability and Statistics, 2004
- 5) Παρουσιάσεις «Βιοστατιστική», Δ. Λιναράτος, πρόγραμμα μεταπτυχιακών σπουδών «ΤΕΧΝΟΛΟΓΙΕΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΣΤΗΝ ΙΑΤΡΙΚΗ ΚΑΙ ΤΗ ΒΙΟΛΟΓΙΑ»

69