

# ΚΕΦΑΛΑΙΟ 1: ΑΝΑΛΥΣΗ ΜΙΑΣ ΜΕΤΑΒΛΗΤΗΣ

## ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

-9-

Εισαγωγή Η επιλογή της σωστής στατιστικής τεχνικής για την περιγραφή μιας μεταβλητής εξαρτάται αποκλειστικά από το χαρακτήρα της, τη διάκρισή της δηλαδή σε ποιοτική ή ποσοτική. Συνήθως, τα πρωτογενή δεδομένα που διαθέτουμε σε μία μελέτη είναι τεράστια σε μέγεθος και έχουν ανατεράχαστη μορφή με αποτέλεσμα να μην μπορούμε να διακρίνουμε τις "πληροφορίες" που περιέχουν. Οι μέθοδοι της περιγραφικής Στατιστικής αποτελούν το επιστημονικό εργαλείο για τη σύνοψη, ταξινόμηση και παρουσίασή της σε εύληπτη μορφή.

Συνήθως χρησιμοποιούνται τρεις διαφορετικές τεχνικές:

- ① οι πίνακες συχνοτήτων
- ② οι γραφικές παραστάσεις
- ③ τα στατιστικά μέτρα

Ας δούμε ένα παράδειγμα.

### Παράδειγμα 1 (Εφαρμογή στο SPSS)

Έρευνα στην οποία λαμβάνεται τυχαίο δείγμα 50 μαθητών από τους νομούς Μιλιών, Δράμας, Κοζάνης, Πέλλας, Καβάλας. Καταγράφεται το ύψος τους σε cm, το βάρος τους σε kg, η ηλικία τους σε μήνες και ο χρόνος που έμειναν στο αγώνισμα των 1000m σε δευτερόλεπτα.

Μελέτη της μεταβλητής νομός. Ποιοτική μεταβλητή που αναπαριστά τον νομό καταγωγής των παιδιών που πήραν μέρος στην έρευνα.

Διαδικασία Frequencies\*: Μπορούμε να πάρουμε έναν πίνακα συχνότητων για την μεταβλητή ποσός. Αν θέλουμε να κατασκευάσουμε κάποιο γράφημα ενεργοποιούμε την επιλογή Charts.

Με την επιλογή Format καθορίζουμε τον τρόπο εμφάνισης των κατηγοριών του πίνακα συχνότητων. Ο πίνακας συχνότητων μπορεί να ταξινομηθεί ως προς τη σειρά εμφάνισης των διαφορετικών κατηγοριών στα δεδομένα ή προς τη συχνότητα εμφάνισής τους. Αυτό μπορεί να γίνει είτε κατά αύξουσα ή κατά φθίνουσα σειρά.

Κάνουμε κυκλικό διάγραμμα (pie-chart) και ραβδόγραμμα (το ιστόγραμμα δεν έχει νόημα για τις ποιοτικές μεταβλητές).

\* Analyze → Descriptives Statistics

(Charts → Pie)

(Charts → Bar)

→ Frequencies

Statistics (Περιγραφικά Μέτρα)

• Bar charts

• Pie charts

Διαδικασία Descriptives: Βοηθά στον υπολογισμό

διάφορων στατιστικών μέτρων μία ή περισσότερων ποσοτικών... (Σελ. 11)

Συμπλήρωμα:

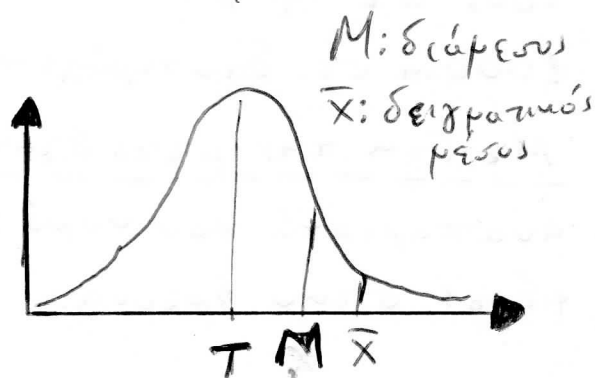
Κατανομή δεδομένων από τύπος συμμετρική. Η τιμή ή τον αριθμητικό μέσο, συμπίπτει με διάμεσο και επικρατούσα τιμή.

$$\bar{X} = M = T \text{ και } M = \frac{Q_1 + Q_3}{2}$$

Θετική ασυμμετρία

$$\bar{X} > M > T, \quad M < \frac{Q_1 + Q_3}{2}$$

T: επικρατούσα τιμή



μεταβλητών.

-11-

Υάνουμε: Descriptive Statistics → Descriptives

Ιpsos, βάρος, ηλικία

Για να κατασκευάσουμε έναν πίνακα συχνοτήτων ή και ποσοτικής μεταβλητής θα πρέπει να συνδυάσουμε την εντολή "Recode" με τη διαδικασία "Frequencies". Ας δούμε τι μπορούμε να κάνουμε με τη μεταβλητή ipso.

Transform → Recode into Different Variables

ορίζουμε τις κλάσεις του υπό κατασκευή πίνακα συχνοτήτων.

Ανάγνωση Output - Frequencies

Στην πρώτη στήλη (Frequency) του πίνακα συχνοτήτων καταγράφεται ο αριθμός των παιδιών που υπήρχαν στο δείγμα (50 περιπτώσεις) από τον κάθε νομό. Το επί της % ποσοστό των σχολικών μεθόδων του δείγματος για κάθε νομό εμφανίζεται στη στήλη της σχετικής συχνότητας (Percent). Αφού δεν υπάρχουν missing values (ελλείπουσες τιμές) οι στήλες Percent και Valid Percent ταυτίζονται. Οι αθροιστικές συχνότητες εμφανίζονται στην τελευταία στήλη του πίνακα.

Ραβδογράμματα Χρησιμοποιούνται κυρίως για τη διαγραμματική απεικόνιση των απόλυτων και των σχετικών συχνοτήτων ή και ποσοτικής μεταβλητής.

Κυκλικά διαγράμματα Εύκολος τρόπος παρουσίασης σχετικών συχνοτήτων κυρίως ή και ποσοτικής μεταβλητής ή και ενός ποσοτικού μεθόδου που απαρτίζεται από επί μέρους διακριτά πράγματα.

Ανάγνωση Output - Descriptives

Skewness (Ασυμμετρία) Μέτρο της συρρεπίας ή της ασυρρεπίας ή και κατανομής

Ο συντελεστής ασυμμετρίας συνήθως συμβολίζεται με:

$$B_1 = \frac{1}{n} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s} \right)^3$$

Αν  $B_1 < 0$  έχουμε αρνητική ή αριστερή ασυμμετρία

Αν  $B_1 > 0$  έχουμε θετική ή δεξιά ασυμμετρία



Κύρτωση (Kurtosis) Μέτρο διαπλάτυνσης της κατανομής

Δίνει πληροφορίες γύρω από την αιχμηρότητα της κατανομής

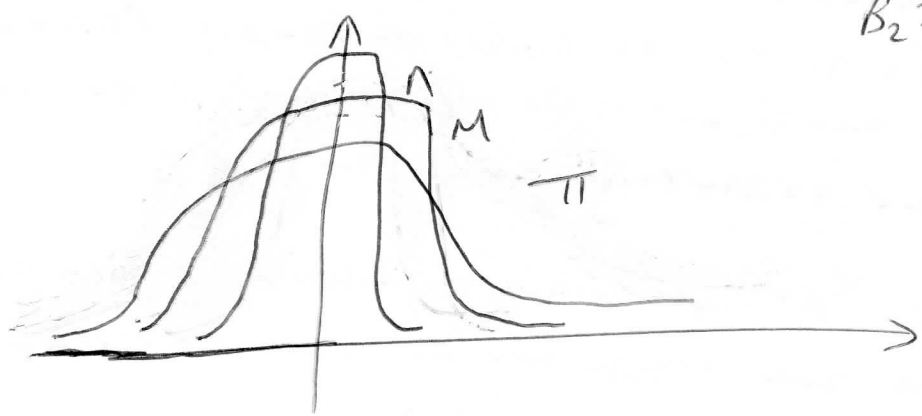
Μέτρο  
κύρτωσης

$$B_2 = \frac{1}{n} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s} \right)^4$$

$B_2 = 3$  (μεσόκυρτη)

$B_2 < 3$  (πλατύκυρτη)

$B_2 > 3$  (στενόκυρτη)



Στα διάφορα πακέτα χρησιμοποιούνται οι ακόλουθοι τύποι:

$$SKEW = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s} \right)^3$$

$$KURT = \left\{ \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s} \right)^4 \right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Stem-and-leaf-plot: Διάγραμμα ρίσχου και φύλλων ή κορμού-και-φύλλων ή φυλλόγραμμα (αποδίδεται στον John Tukey)

Η γραφική προσέγγιση του φυλλογραφήματος ενδείκνυται κατά κύριο λόγο για την επεξεργασία μεγάλων πλήθους παρατηρήσεων (περίπου 150).

π.χ. Έστω ότι έχουμε τις ηλικίες

27 34 34 43 21 38 46 38 22 35

Τις διατάσσουμε: 21 22 27

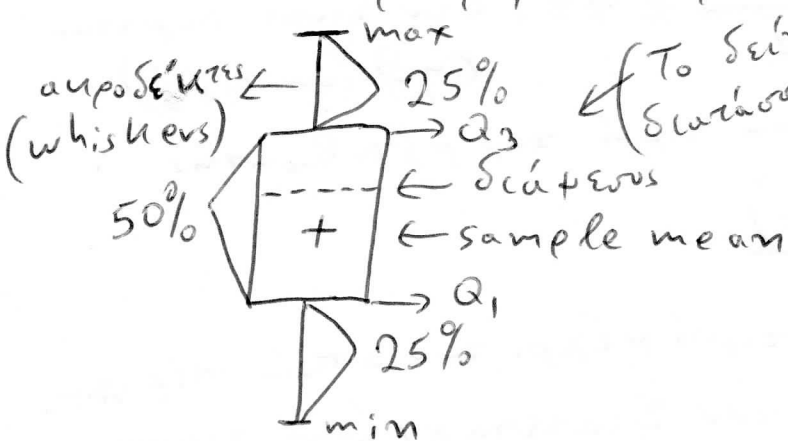
34 34 35 38 38  
43 46

| Κορμός-Stem | Φύλλο-Leaf |
|-------------|------------|
| 2           | 1 2 7      |
| 3           | 4 4 5 8 8  |
| 4           | 3 6        |

π.χ. Κορμός-Stem φύλλο-Leaf  
2 1  
4 6

Οι παρατηρήσεις εντός ορθογώνιου παραλληλογράμμου χαρακτηρίζονται ως παράτυπα σημεία (outliers) ή σαν ακραίες τιμές (extreme values).

Boxplot (Θηνόγραμμα): Μέθοδος γραφικής παράστασης των βασικών μέτρων θέσης στο εύρος των τιμών των υπό μελέτη χαρακτηριστικών. Δίνει τη δυνατότητα διερεύνησης σχετινά με την ύπαρξη ή μη ακραίων παρατηρήσεων (outliers). Δίνει τις τιμές των δύο τεταρτημορίων, της διαμέσου, της ελάχιστης και μέγιστης τιμής των δεδομένων. Απεικονίζεται επίσης και το ενδοτεταρτημοριακό εύρος  $Q_3 - Q_1$ .



Αυροδέκτες Ενίων την ελάχιστη και τη μέγιστη τιμή της μεταβλητής με την άνω και την κάτω πλευρά του ορθογώνιου.

Η θέση της διαμέσου δίνει το είδος της ασυμμετρίας.

-14-

- Εάν η κατανομή είναι συμμετρική η διαμέσος (median) θα βρίσκεται στο μέσο και θα συντίθεται με την διευκρινιστική μέση τιμή  $\bar{X}$ ,
- εάν βρίσκεται προς το  $Q_1$  λέμε ότι η κατανομή έχει θετική ασυμμετρία
- εάν βρίσκεται προς το  $Q_3$  λέμε ότι η κατανομή έχει αρνητική ασυμμετρία.

Διάμεσος (Median) χωρίζει το σύνολο των διατεταγμένων δεδομένων σε δύο ισοπαγήδη υποσύνολα, 50% των παρατηρήσεων έχουν τιμές μικρότερες ή ίσες της διαμέσου.

Θέση διαμέσου (αφού διατάξουμε τις τιμές από τη μικρότερη στην μεγαλύτερη)  $= \frac{n+1}{2}$

Αν  $n=7$  (περιττός) π.χ. 60 65 67 (68) 70 78 80  
 $\frac{n+1}{2} = 4$  Median = 68

Αν  $n=8$  (άρτιος) π.χ. 62 65 67 70 78 79 81 82  
Θέση διαμέσου  $= \frac{n+1}{2} = 4.5$  <sup>εδώ</sup> <sub>διάμεσος</sub>

Ευρβατική θεωρούμε  $\frac{70+78}{2} = 74$  (δεν είναι παρατηρήσιμη τιμή)  
Είναι όμως δυνατή θεωρητική τιμή της μεταβλητής.

Τεταρτηρόρια (Quantiles)

$Q_1$ : πρώτο τεταρτηρόριο, η τιμή μέχρι την οποία περιλαμβάνεται το  $\frac{1}{4}$  (25%) των διατεταγμένων τιμών του δείγματος

$Q_3$ : τρίτο τεταρτηρόριο, το  $\frac{3}{4}$  (το 75%) των διατεταγμένων

τιμών του δείγματος

$Q_2$ : Δεύτερο τεταρτημόριο  $\equiv$  Συρπύπτει με τη Διάμεσο

$IQR = Q_3 - Q_1 = \text{ενδοτεταρτημοριακό εύρος}$   
(interquartile range)

$Q = \frac{1}{2} IQR = \frac{1}{2} (Q_3 - Q_1)$  (ημι-ενδοτεταρτημοριακό εύρος)  
(semi-interquartile range)

Πιθανοθεωρητικά γραφήματα  $P-P$ ,  $Q-Q$

Γραφήματα  $P-P$ ,  $Q-Q$ : Με τα γραφήματα αυτά προσπαθούμε να διαπιστώσουμε πόσο κοντά σε κάποια κατανομή των προσδιορίζουμε είναι τα δεδομένα που επεξεργαστήραστε. Συνήθως ενδιαφερόμαστε για την κανονική κατανομή και ριζώρε για διαγράμματα διασποράς της παρατηρούμενης κατανομής ως προς την κανονική κατανομή. Όσο πιο εύκολα φαίνεται ότι τα σημεία είναι συγκεντρωμένα γύρω από μία ευθεία (συνήθως είναι η διχοτόμος της χανίια των  $x-y$ ) τόσο περισσότερο ενισχύεται η υπόθεση ότι τα δεδομένα ακολουθούν τη μεταβλητή που έχει υποτεθεί.

Γράφημα  $Q-Q$  Γράφημα τύπου Quantile-Quantile παρουσιάζει τα εκατοστημόρια (ποσοστιαία σημεία) της παρατηρούμενης ως προς την αναμενόμενη κανονική κατανομή. Τα εκατοστημόρια της υποτιθέμενης κανονικής κατανομής υπολογίζονται με διάφορους αλγόριθμους π.χ. Tukey και τοποθετούνται στον άξονα  $y$  ενώ τα αντίστοιχα παρατηρούμενα στον άξονα  $x$ . Επιπλέον το SPSS δίνει και μία χ.π. που δείχνει τις ατομικές αποστάσεις μεταξύ παρατηρούμενων και επιμώμενων εκατοστημορίων.

-16-

Τα σφάλματα αυτά καταγράφονται γύρω από μία ορισμένη γραμμή που αντιστοιχεί συνήθως 0.

Analysis Output

Skewness =  $\frac{\sum z^3}{n}$ ,  $z$ : the distance of each score from the mean

SEskew =  $\sqrt{\frac{6}{n}}$  for large samples

Kurtosis =  $\left( \frac{\sum z^4}{n} \right) - 3$

$z_{skew} = \frac{\sum z^3/n}{\sqrt{6/n}}$

SEkurt =  $\sqrt{\frac{24}{n}}$

$z_{kurt} = \frac{\text{kurtosis}}{\sqrt{\frac{24}{n}}}$

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

$-1.96 \leq z_{kurt} \leq +1.96$   
 $\Rightarrow 95\% \text{ confidence}$

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

that the distribution is normal

$z_{kurt} > 0$  is normal (leptokurtic)  
more peaked than normal

$z_{kurt} < 0$  (platykurtic)

Analysis Output

Frequencies for Ipsos

Mean = 145.680 Median = 144,5

C.V. = coefficient of variation = συντελεστής μεταβλητότητας =  $\frac{S}{\bar{x}} \times 100$

=  $\frac{7.098}{145.680} = 0.04872 = 4,87\%$  μικρότερο από το 20%

δείγμα ομοιογενές

$\frac{0.422}{0.337} = \frac{\text{skew}}{\text{SEskew}}$

ο λόγος της 203% της = 1.25 ενώ [-2,2] και της

αυριότητας = 1.27 είναι ενώ το [-2,2]. Περιμένουμε λοιπόν  
" -0.341/0.662

η κατανομή της ipsos να είναι κατά την κανονική. Το ιστόγραμμα δεν δίνει μεχάρες απομαρτίσεις από την κανονικότητα.

Recode της μεταβλητής ipsos

Transform → Recode → Old and New Values

New = ipsos - R

Analyze → Frequencies

Charts → Histograms  
with normal curves

Format → Ascending counts

Κατασκευή πολυγώνων συχνότητας για τη μεταβλητή ipsos

Δίνει με πιο ενδιαφέρον τρόπο από το ιστόγραμμα τη μορφή των δεδομένων. Χρησιμοποιείται μόνο για τη διαγραμματική απεικόνιση συνεχών δεδομένων.

Επιανακατασκευάζει το ύψος δίνοντας νέα τρόπο μέτρο της υψότητας στην οποία ανήκει η αρχική τιμή.

|                                                             |             |           |       |                  |
|-------------------------------------------------------------|-------------|-----------|-------|------------------|
| Σχεδιάζεται με βάση το ιστόγραμμα με αφετηρία της κατανομής | ipsos - R → | 135 - 140 | 137,5 | (μέτρα υψότητας) |
|                                                             |             | 140 - 145 | 142,5 |                  |
|                                                             |             | 145 - 150 | 147,5 |                  |
|                                                             |             | 150 - 155 | 152,5 |                  |
|                                                             |             | 155 - 160 | 157,5 |                  |

Το πολύγωνο συχνότητας ενδυναμώνεται κυρίως για την παρόχρηση σύγκρισης δύο ή περισσότερων κατανομών.

Graphs → Line → Simple  
(Summaries of group of cases)

Refine  
Category Axis → ipsos - R (Line represents  
N of cases)

Περισσότερη περιγραφική ανάλυση προσφέρει η διαδυσκία Explore για περισσότερες της μιας μεταβλητών

- Prescriptives (εύρος, μέση τιμή, διάμεσος, ωπική απόκλιση, διακύμανση) ipsos, baros, ηλικία

## M-estimators

Estimators which obtained from the minimization of

$$\sum_{i=1}^n \rho(x_i, \theta), \text{ where } \rho \text{ is a function with certain properties. The solutions } \hat{\theta} = \arg \min_{\theta} \left( \sum_{i=1}^n \rho(x_i, \theta) \right)$$

are called M-estimators. Maximum likelihood estimators are thus a special case of M-estimators.

M-estimators are solutions  $\theta$  which minimize  $\sum_{i=1}^n \rho(x_i, \theta)$ . Often is simpler to differentiate with respect to  $\theta$  and solve for the root of the derivative. Then the M-estimator is said to be of  $\psi$ -type. Otherwise is of  $\rho$ -type

Examples Mean  $(X_1, \dots, X_n) \text{ iid } \sim F$

$$\rho(x, \theta) = \frac{(x - \theta)^2}{2} \Rightarrow \hat{\theta} = \bar{X} \Rightarrow \text{mean M-estimator of } \rho\text{-type}$$

Outliers Πέντε μεγαλύτερες  
Πέντε μικρότερες τιμές

Percentiles  $5^{00}, 10^{00}, 25^{00}, 50^{00}, 75^{00}, 90^{00}$  και  $95^{00}$  είναι ορόσημα απείων

Ανάγνωση του Output για την περιθώριση χρόνου με τη διαδικασία Explore

Mean = 361.84 } Median < Mean  $\Rightarrow$  η κατανομή της περιθώρισης χρόνου έχει  
Median = 361.5 }  
= skew/se } **ΘΕΤΙΚΗ** λοξότητα (ασυμμετρία)  
 $Z_{skew} = 0.388$  } λοξότητα εντός  $[-2, 2]$   
= curt/se } **0.6419** ("κοντά" σε κανονική κατανομή)  
 $Z_{curt} = 0.6419$

Οι τιμές των λοβόμια, υέρτων  $< 0 \Rightarrow$  κατανομή με ουρά -29-  
προς τα αριστερά (μικρές τιμές)

5% trimmed Mean = αποκορμένος μέσος όρος (με αποκοπή  
είναι 362,03 και η Mean = 361,84 5% των  
ακραίων τιμών)

Μικρή διαφορά συνεπώς η ύπαρξη ακραίων  
και παρά των σημείων δεν επηρεάζει σημαντικά  
τις τιμές του μέσου.

Το Οημόγραμμα, το φυλλογράφημα και το ιστόγραμμα δείχνει  
ότι δεν απέχουμε "πολύ" από την κανονική κατανομή

Η διαφέρως στο Οημόγραμμα είναι περίπου στο μέσο

IQR = 33 δεν φαίνεται να υπάρχει μεγάλη διασπορά των  
μετρήσεων.

$$C.V. = \frac{24,19678}{361,84} = 0.0668 \approx 7\%$$

$$\begin{aligned} Q_3 &= 378 \\ Q_1 &= 345 \end{aligned} \quad \left\{ \begin{aligned} Q_3 - Q_1 &= 33 \end{aligned} \right.$$

M-estimators Το SPSS προσφέρει 4 ανθεκτικούς (robust)  
κ.ρ.π. της κεντρικής τάσης. Προτιμώνται όταν υπάρχουν  
outliers γιατί για τον υπολογισμό τους οι απορριπτόμενες  
τιμές αναβιβάζονται με μικρότερα βάρη από ότι οι κεντρικές  
τιμές.

Σε ασύμμετρες κατανομές ο M-estimator Huber's καδώς  
επίσης και ο αποκορμένος μέσος όρος τείνουν προς τον  
πληθυσμιακό μέσο όρο ενώ οι υπολοίποι M-επιτηρητές  
είναι περὶ την πληθυσμιακή διαφέρως. Στην περίπτωση  
μας οι M-επιτηρητές είναι κοντά στην Mean γιατί η  
αποτελεί ένδειξη συμμετρικής κατανομής.

## Ta Tests for Normality

Kolmogorov-Smirnov και Shapiro-Wilk

και τα Q-Q plots επιβεβαιώνουν τις ενδείξεις κανονικότητας της κατανομής των μετρήσεων χρόνου.

### Kolmogorov-Smirnov Test

Η εμπειρική συνάρτηση κατανομής  $F_n(x)$  for  $n$  iid observations

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I_{X_i \leq x} = S(x)$$

(ποσοστό παρατηρήσεων του δείγματος οι οποίες είναι το πολύ ίσες με την τιμή  $x \forall x \in (-\infty, \infty)$ )

Ο έλεγχος K-S συγκρίνει την αθροιστική σ.κ.  $F(x)$  με την αντίστοιχη εμπειρική σ.κ.  $S(x) \equiv F_n(x)$  βασισμένη σε ένα δείγμα  $n$  παρατηρήσεων.

(βλέπε Ξενοδόχου 422-430)

$$S(x) = F_n(x) = \frac{\# \text{ τιμών δείγματος} \leq x}{n}$$

$$\text{Ελέγχουμε } H_0: F(x) = F_0(x) \quad \forall x \in (-\infty, \infty)$$

$$H_1: F(x) \neq F_0(x)$$

$X_1, \dots, X_n$  τ.δ. μεγέθους  $n$  από κάποιο άγνωστο

κατανομή με σ.κ.  $F(x)$ . Έστω  $F_0(x)$  μία σαφής καθορισμένη κατανομή. Η μεδιστη ματανομή απόσταση

μεταξύ  $S(x), F_0(x)$

$$T = D_n = D = \sup_x |F_n(x) - F_0(x)|$$

$$\text{Μεγαλύτερες } T \text{ συνιστούν } \leq \sup_x |F_0(x) - F_n(x)|$$

$$\text{στη απωφ. της } H_0 \leq \sup_x |F_0(x) - S(x)|$$

Η ακριβής κατανομή της  $T$  προκύπτει από Kolmogorov

$T_{\alpha}(1-\alpha)$  ποσοστιαία κριθείς της υπάρχουν σε

πίνακες. Η απόδειξη της κανονικότητας επιβεβαιώνεται από το  $p$ -value.

# Shapiro-Wilk Test for Normality

-21-

W-test for goodness of fit ( $K-S$ , Lilliefors,  $\chi^2_{K-1}$ )

π.δ.  $X_1, \dots, X_n$  iid  $\sim F_X(x)$ ,  $x \in \mathbb{R}$ , σ.κ.

Ελέγχουμε:

$H_0: F_X(x) \sim F$  με άγνωστη μέση τιμή και άγνωστη διασπορά,  $F$  κανονική

$H_1: F_X(x) \sim$  σ.κ. ρία ρη-κανονική

Στατιστική συνάρτηση ελέγχου

$$T_3 = W = \frac{\left[ \sum_{i=1}^K a_i (X^{(n+1-i)} - X^{(i)}) \right]^2}{D},$$

$$D = \sum_{i=1}^n (X_i - \bar{X})^2, \quad X^{(i)}: \text{η } i\text{-οστή παρατήρηση}$$

διατεταγμένου δείγματος  
κάτα αύξουσα τάξη  
μεγέθους,  $K \approx \frac{n}{2}$  ακέραια

$a_i, i=1, \dots, K$

σταθεροί συντελεστές που δίνονται από κατάλληλους πίνακες  
για δοθέν  $n$  και  $K \approx \frac{n}{2}$ .

(βλέπε Ξεναλάκη: Μη-Παραμετρική Στατιστική, σελ. 500-501)

