

ΕΛΕΓΧΟΣ ΚΑΝΟΝΙΚΟΤΗΤΑΣ ΔΕΔΟΜΕΝΩΝ

## Έλεγχος των δεδομένων για εκλείπουσες τιμές (missing values)

- Κάθε αρχείο με δεδομένα θα έχει ορισμένες (λίγες ή περισσότερες) εκλείπουσες τιμές...
- Το πρόβλημα είναι σημαντικό μόνο εφόσον οι τιμές αυτές δεν κατανέμονται μέσα στην κατανομή τυχαία!
- Ο έλεγχος για τον τρόπο με τον οποίο κατανέμονται οι εκλείπουσες τιμές μέσα στο αρχείο γίνεται στο SPSS με το εργαλείο Missing Value Analysis:
  - Analyze → Missing Value Analysis → Τοποθέτηση των μεταβλητών που μας ενδιαφέρει να ελεγχθούν στο κατάλληλο παράθυρο (Quantitative ή Categorical Variables) → Patterns → Cases with missing values, sorted by missing value patterns
  - Στον πίνακα που θα προκύψει (βλ. παρακάτω) κοιτάζουμε το ποσοστό των εκλειπουσών τιμών (δεν θέλουμε να ξεπερνάει το 1%) καθώς και το πού βρίσκονται οι συγκεκριμένες τιμές.
- Στην περίπτωση που έχουμε πολλές εκλείπουσες τιμές και μας δημιουργούν πρόβλημα (π.χ., στην περίπτωση της χρήσης της εντολής compute) μπορούμε να συμπληρώσουμε τις τιμές αυτές (βλ. επόμενη διαφάνεια)...

### Πίνακας Σχημάτων Εκλειπουσών Τιμών από το SPSS

Missing Patterns (cases with missing values)													
Case	# Missing	% Missing	Missing and Extreme Value Patterns <sup>a</sup>										
			keyb1test2	keyb1test3	keyb1test4	vrinterfacetest1	vrinterfacetest2	vrinterfacetest3	vrinterfacetest5	gender	vrinterfacetest4	keyb1test5	keyb1test1
5	1	9,1											
10	1	9,1											
14	1	9,1											
19	11	100,0											

- indicates an extreme low value, while + indicates an extreme high value. The range used is (Q1 - 1.5\*IQR, Q3 + 1.5\*IQR).

a. Cases and variables are sorted on missing patterns.

### Πώς συμπληρώνουμε τις εκλείπουσες τιμές;

- Το SPSS έχει διάφορους τρόπους να αντιμετωπίζει τις εκλείπουσες τιμές:
- Για παράδειγμα, μπορούμε να χρησιμοποιήσουμε την επιλογή **Replace Missing Values** στο μενού **Transform**.
- Ωστόσο, οι επιλογές (inserting mean, median or linear interpolation) δεν είναι τόσο αξιόπιστες μέθοδοι εκτίμησης των εκλειπουσών τιμών.
- Η καλύτερη επιλογή είναι να χρησιμοποιήσουμε τις μεθόδους που υπάρχουν στο **Missing Value Analysis**.
- Επιλέγουμε: **Analyze → Missing Value Analysis**
- Επιλέγουμε τις ποσοτικές και τις κατηγορικές μεταβλητές και τις τοποθετούμε στα κατάλληλα πλαίσια (επιλέγουμε τόσο τις μεταβλητές που έχουν εκλείπουσες τιμές και θέλουμε να τις συμπληρώσουμε όσο και εκείνες που θα χρησιμοποιήσουμε στη συνέχεια για τον υπολογισμό των τιμών των μεταβλητών με εκλείπουσες τιμές)
- Κλικ στο **Patterns** → επιλέγουμε **‘Tabulated cases grouped by missing value patterns’** και **‘Cases with Missing Values, sorted by missing value patterns’**
- Πατάμε **Continue** → **OK**
- **Σημειώνουμε το ποσοστό των εκλειπουσών τιμών για τις μεταβλητές που θα αντικαταστήσουμε.**  
*Προσέξτε ότι συνιστάται το ποσοστό των εκλειπουσών τιμών να μην ξεπερνάει το 1% σε μια μεταβλητή. Εφόσον αυτή η προϋπόθεση ισχύει, προχωράμε στο επόμενο στάδιο...*

### Πώς συμπληρώνουμε τις εκλείπουσες τιμές (συνέχεια);

- Επιστρέφουμε στο κουτί διαλόγου του **Missing Value Analysis**.
- Τοποθετούμε τη μεταβλητή με τις εκλείπουσες τιμές ως **quantitative variable**
- Από τα **Estimation methods**, επιλέγουμε **EM**
- Κάνουμε κλικ στο κουμπί **Variables**
- Κάνουμε κλικ στο κουμπί **Select variables**
- Επιλέγουμε τις μεταβλητές με εκλείπουσες τιμές και τις μετακινούμε στο κουτί **Predicted variables**
- Επιλέγουμε όλες τις μεταβλητές που θα χρησιμοποιηθούν για τον υπολογισμό των τιμών που θα αντικαταστήσουν τις εκλείπουσες τιμές και τις μετακινούμε στο κουτί **Predictor variables**
- **Continue**
- Κλικ στο κουμπί **EM** και τσεκάρουμε την επιλογή **Save Completed Data**
- Κλικ στο **File** και δίνουμε ένα όνομα για το καινούριο αρχείο δεδομένων, το οποίο θα περιέχει τις υπολογισμένες τιμές.
- Κλικ **Save**
- Κλικ **Continue**
- Ανοίγουμε το καινούριο αρχείο με τα δεδομένα που μόλις δημιουργήσαμε

Όπως καταλαβαίνετε, πρόκειται ουσιαστικά για μια παλινδρόμηση: υπολογίζουμε τις τιμές που θα αντικαταστήσουν τις εκλείπουσες τιμές βάσει των τιμών σε άλλες μεταβλητές (αλλά και στην ίδια τη μεταβλητή με τις εκλείπουσες τιμές).

## Κατασκευή του ιστογράμματος της μεταβλητής στο SPSS

Επιλέγουμε:

**[Graphs → Legacy Dialogs → Histogram...]**

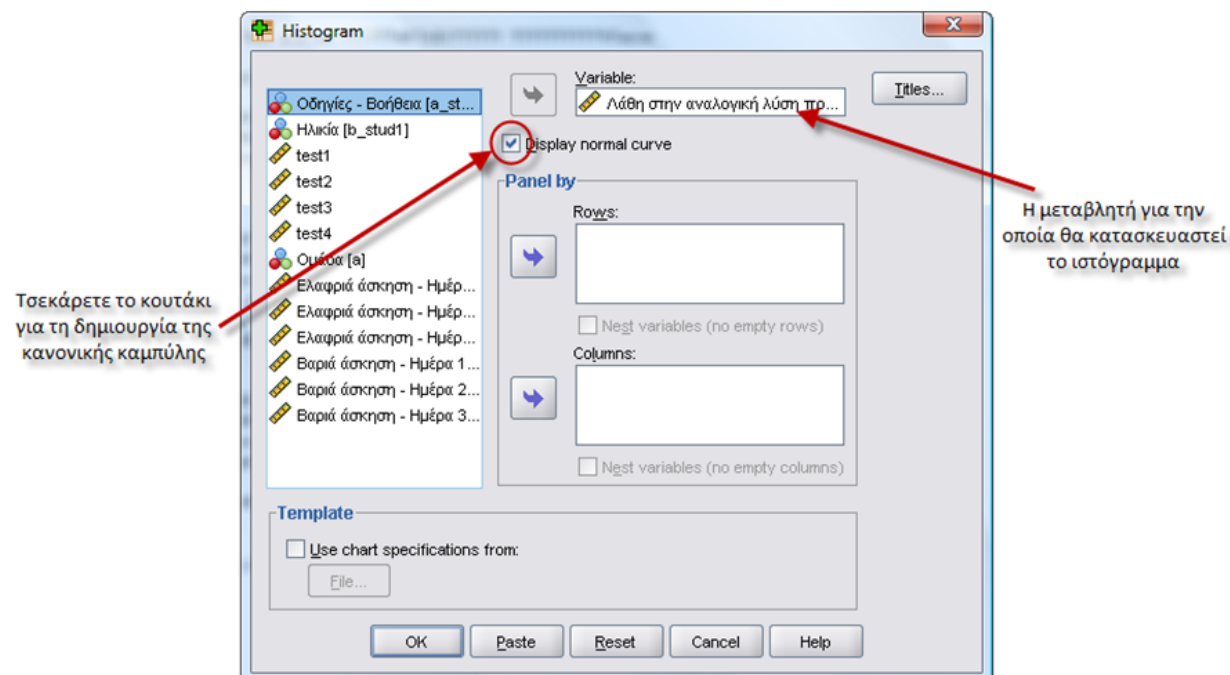
Στο παράθυρο που ανοίγει (διπλανό Σχήμα) θα πρέπει

(α) να μετακινήσετε στο πλαίσιο **[Variable:]** τη μεταβλητή για την οποία θα κατασκευαστεί το ιστόγραμμα, και

(β) να τσεκάρετε το κουτάκι στην επιλογή για τη δημιουργία της κανονικής καμπύλης πάνω από το ιστόγραμμα **[Display normal curve]**.

Βεβαίως, δεν υπάρχουν ασφαλή κριτήρια για το βαθμό απόκλισης του ιστογράμματος από την κανονική καμπύλη...

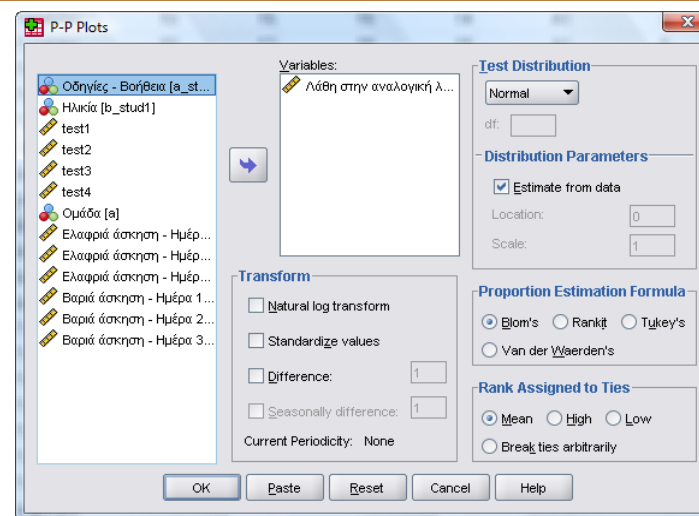
### Το παράθυρο Histogram



## Κατασκευή P-P plots (Proportion-Proportion)

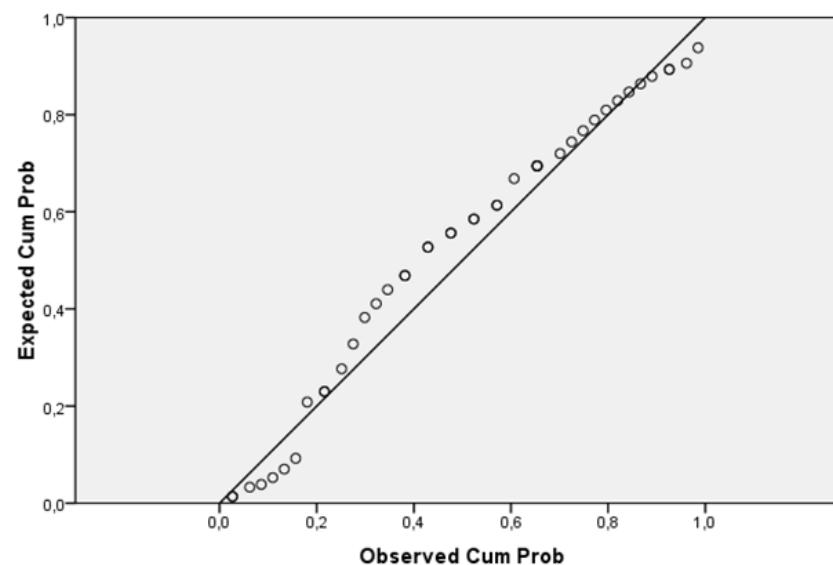
- Πρόκειται για ένα διάγραμμα όπου η παρατηρηθείσα αθροιστική σχετική συχνότητα σχεδιάζεται απέναντι στην αναμενόμενη αθροιστική σχετική συχνότητα έτσι όπως θα έδειχναν αν η κατανομή ήταν κανονική.
- Επιλέγουμε  
**[Analyze → Descriptive Statistics → P-P Plots...]**  
 και στο παράθυρο που ανοίγει (βλ. διπλανό Σχήμα) μεταφέρουμε στο πλαίσιο **[Variables:]** τη μεταβλητή για την οποία θα κατασκευαστεί το διάγραμμα.
- Κάνοντας κλικ στο κουμπί **[OK]** θα δημιουργηθεί ένα διάγραμμα όπως αυτό του διπλανού Σχήματος.
- Όλα τα σημεία θα πρέπει να βρίσκονται πάνω στη διαγώνιο εφόσον η μεταβλητή είναι κανονικά κατανεμημένη.
- Κατασκευάζεται εύκολα αλλά δεν υπάρχουν κοινά αποδεκτά κριτήρια για τον καθορισμό της απόστασης που μπορεί να έχουν τα σημεία από τη διαγώνιο ώστε να θεωρηθεί κανονική η κατανομή...

## Το παράθυρο P-P Plots



## Ένα παράδειγμα P-P plot

Normal P-P Plot of Λάθη στην αναλογική λύση προβλημάτων



## Συμμετρία και κύρτωση (skew &amp; kurtosis)

- Επιλέγουμε **[Analyze → Descriptive Statistics → Descriptives...]** και στο παράθυρο που ανοίγει μεταφέρουμε στο πλαίσιο **[Variables:]** τη μεταβλητή για την οποία θα υπολογιστούν οι συγκεκριμένοι δείκτες.
- Κάνουμε κλικ στο κουμπί **[Options...]** και στο παράθυρο διαλόγου που ανοίγει επιλέγουμε το **[Kurtosis]** και το **[Skewness]**.
- Πατάμε **[Continue]**, μετά **[OK]** και στο παράθυρο Viewer θα πάρουμε ένα πίνακα με όλους τους δείκτες που έχουμε ζητήσει.
- Και οι δύο τιμές πρέπει να είναι μηδενικές για κανονική κατανομή. Ένας απλός κανόνας για να δεχτούμε ότι μια μεταβλητή είναι κανονική είναι να μην ξεφεύγουν οι τιμές της συμμετρίας και της κύρτωσης από το +2 ως το -2.
- Άλλοι είναι πιο «αυστηροί» και συνιστούν το εύρος από το +1 ως το -1... Αυτός ο κανόνας εφαρμόζεται στις περιπτώσεις μεγάλων δειγμάτων ( $N > 300$ ).

Αν το δείγμα είναι μικρό ( $N < 100$ )...

... υπολογίστε τις τυπικές τιμές για τη συμμετρία και την κύρτωση βάσει του τύπου που ακολουθεί και απορρίψτε ως μη κανονικές τις μεταβλητές εκείνες που έχουν τυπική τιμή μεγαλύτερη από το 1,96:

$$Z_{\text{skew}} = \text{Skew} / \text{SE}_{\text{skew}}$$

$$Z_{\text{kurtosis}} = \text{Kurtosis} / \text{SE}_{\text{kurtosis}}$$

Αν το δείγμα ήταν μέτριου μεγέθους ( $100 < N < 300$ ), τότε υπολογίζουμε και πάλι τις τυπικές τιμές για τη συμμετρία και την κύρτωση και απορρίπτουμε ως μη κανονικές εκείνες τις μεταβλητές που έχουν έστω μία από αυτές τις τυπικές τιμές μεγαλύτερη από 3,29.

## Ένα παράδειγμα πίνακα περιγραφικών από το SPSS

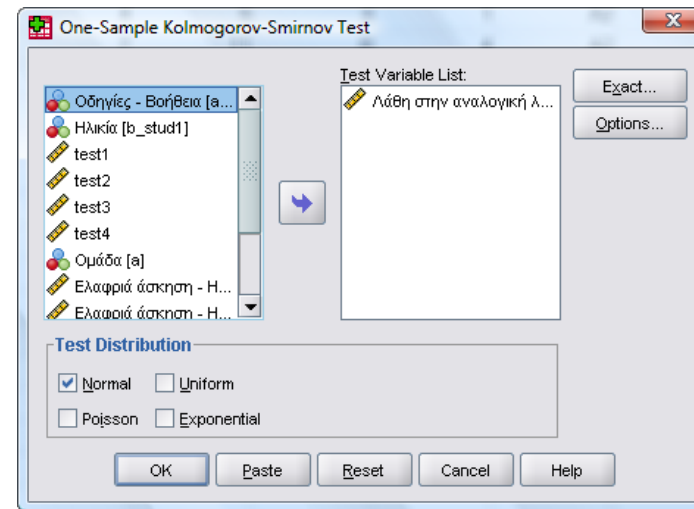
Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Χρόνος ολοκλήρωσης έργου στο πληκτρολόγιο - δεύτερη προσπάθεια	18	8	18	13,44	2,833	-,511	,536	-,737	1,038
Valid N (listwise)	18								

## Τυπικά τεστ κανονικότητας

- Το γνωστότερο είναι το Kolmogorov-Smirnov (ένα δεύτερο που υπολογίζει το SPSS είναι το Shapiro-Wilk).
- Επιλέγουμε **[Analyze → Nonparametric Tests → 1-Sample K-S...]** και μεταφέρουμε στο πλαίσιο διαλόγου **[Test Variable List:]** τη μεταβλητή για την οποία θα πραγματοποιηθεί το τεστ.
- Στο παράδειγμά μας το αποτέλεσμα του τεστ δεν είναι στατιστικά σημαντικό ( $p=.555$ ), επομένως θα πρέπει να δεχτούμε ότι η μεταβλητή που ελέγχθηκε δεν αποκλίνει από την κανονικότητα.
- Το πρόβλημα με τα τεστ αυτά είναι ότι όσο μεγαλώνει το δείγμα ( $N > 300$ ) τόσο αυξάνει η πιθανότητα να απορριφθεί μια μεταβλητή, η οποία αποκλίνει ελάχιστα από την κανονικότητα.
- Τα τεστ αυτά είναι πολύ «συντηρητικά» και μπορεί να απορρίψουν ως μη κανονική μια μεταβλητή που απέχει ελάχιστα από την κανονικότητα.

## Το παράθυρο του Kolmogorov-Smirnov στο SPSS



## Πίνακας με αποτελέσματα Kolmogorov-Smirnov

One-Sample Kolmogorov-Smirnov Test

		Λάθη στην αναλογική λύση προβλημάτων
N		42
Normal Parameters <sup>a</sup>	Mean	30,07
	Std. Deviation	13,617
Most Extreme Differences	Absolute	,122
	Positive	,074
	Negative	-,122
Kolmogorov-Smirnov Z		,793
Asymp. Sig. (2-tailed)		,555

a. Test distribution is Normal.