κεφάλαιο 7

ΠΛΗΡΗΣ ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΣΤΟ SPSS

(COMPLETE DESCRIPTIVE STATISTICAL ANALYSIS IN SPSS)

Στα προηγούμενα κεφάλαια παρουσιάστηκαν αναλυτικά οι μέθοδοι περιγραφικής διερεύνησης των μεταβλητών, δηλαδή της κατανομής συχνότητας (frequency distributions), των μέτρων κεντρικής θέσης (measures of central location), των μέτρων διασποράς (measures of dispersion), των τιμών z (z-scores), της λοξότητας (skewness), της κύρτωσης (kurtosis). Οι μέθοδοι αυτές ανήκουν στην **περιγραφική στατιστική** (descriptive statistics) και έχουν ως στόχο τη διερεύνηση της "ταυτότητας" κάθε μεταβλητης, πριν αυτή χρησιμοποιηθεί σε κάποια **στατιστική ανάλυση**, που μπορεί να είναι **παραμετρική** (parametric) ή **μη παραμετρική** (non parametric).

Μια μεταβλητή για να αναλυθεί **παραμετρικά** (δηλαδή μέσω εκτίμησης πληθυσμιακών παραμέτρων από κατανομή γνωστης μορφής, μέσου και τυπικής απόκλισης), θα πρέπει να ισχύουν οι εξής βασικές προϋποθέσεις, που τίθενται συνήθως ως στατιστικές παραδοχές (statistical assumptions):

(i) η μεταβλητή να έχει μετρηθεί στην **διαστημική κλίμακα** (interval scale), δηλαδή να αποδίδει ποσοτικά δεδομένα (quantitative data),

(ii) να έχει **ομοιογένεια διασποράς** (homogeneity of variance) σε εξαρτημένα ή ανεξάρτητα δείγματά της (samples) ή ομάδες της (groups),

(iii) να μην περιέχει **ακραίες τιμές** (outliers), δηλαδή πολύ μικρές ή πολύ μεγάλες σε σχέση με τις υπόλοιπες τιμές της κατανομής και

(iv) να είναι **κανονική** (normal), δηλαδή η κατανομή της να είναι συμμετρική (symmetrical) και μεσόκυρτη (mesokurtic).

Αν η μεταβλητή έχει χοντρικά, έστω, τα χαρακτηριστικά αυτά, τότε μπορεί να εκπροσωπηθεί από τον μέσο (M) και την τυπική απόκλιση (s) και να χρησιμοποιηθεί σε: (α) σε συσχετίσεις (correlations) της με άλλες μεταβλητές ή (β) σε συγκρίσεις (comparisons) δειγμάτων της (samples).

Αν, όμως, η **μεταβλητή** ανήκει στην **διατακτική** (ordinal) ή στην **ονομαστική** (nominal) **κλίμακα** τα δεδομένα της είναι ποιοτικά (qualitative) και μπορεί να εκπροσωπηθεί από το διάμεσο (median) ή την κορυφή (mode), αντίστοιχα. Τότε, κατάλληλη στατιστική ανάλυση είναι η μη παραμετρική (non parametric), δηλαδή *ελεύθερη κατανομής* (distribution free). Το ίδιο ισχύει αν η μεταβλητή είναι μεν ποσοτική, αλλά αποκλίνει σαφώς από τις παραδοχές αυτές, εκτός αν μετασχηματισθεί (transformed), για να απαλλαγεί από ακραίες τιμές και τυχόν ασυμμετρία κατανομής.

Στο παρόν κεφάλαιο αναλύονται περιγραφικά μέσω της διαδικασίας **Frequencies** και **Explore** του **SPSS** 3 **διαφορετικές περιπτώσεις κατανομικής δομής** (distributional shape): (α) συμμετρική, (β) λοξή, (γ) μετασχηματισμένη και χωρίς λίγες πολύ ακραίες τιμές της. Στο πρώτο παράδειγμα γίνεται πλήρης ερμηνεία των στατιστικών, ενώ στα άλλα 2 παραδείγματα αντιμετωπίζεται η ασυμμετρία και η ύπαρξη ακραίων τιμών.

7.1 Μια συμμετρική κατανομή (κανονική)

Τα δεδομένα αφορούν τη μεταβλητή *κοιλιακοί* (αριθμός αναδιπλώσεων από την εδραία θέση σε 30″) δείγματος 339 ανδρών 18-26 ετών. Σκοπός της ανάλυσης είναι η επίδειξη της πλήρους διερεύνησης των χαρακτηριστικών κεντρικής θέσης, διασποράς και κατανομής μιας μεταβλητης.

1. Διαδικασία "Frequencies" (Συχνότητες)

```
SPSS: ANALYZE >> DESCRIPTIVE STATISTICS >>
FREQUENCIES
FREQUENCIES VARIABLES=ΚΟΙΛΙΑΚΟΙ /NTILES=4
 /STATISTICS=STDDEV VARIANCE RANGE MINIMUM
MAXIMUM SEMEAN MEAN MEDIAN MODE SKEWNESS
SESKEW KURTOSIS SEKURT
 /HISTOGRAM NORMAL /ORDER=VARIABLE.
```

N = πλήθος Valid (έγκυρων τιμών)
Missing = 0 (δεν ελλείπουν τιμές)
Mean (μέσος, M) = ΣX / N = 25.89 .26
Standard error of mean (τυπικό σφάλμα του μέσου)

$SEM = S / \sqrt{N} = 0.163$

Median = διάμεσος (Md) = 26
Mode (Mo) = κορυφή = 27
Standard Deviation (τυπική απόκλιση)

$s = \sqrt{\text{variance}} = 3.004$

Variance (διασπορά) $s^2 = \Sigma(X-M)^2/(N-1)$
Skewness (λοξότητα) Sk = 0.136
Std. error of Skewness (SEsk)
(τυπικό σφάλμα λοξότητας) = 0.132
Kurtosis (κύρτωση) Ku = 0.047
Std. error of kurtosis (SEku)
(τυπικό σφάλμα κύρτωσης) = 0.264
Range (εύρος) R = 17
Minimum (ελάχιστο) = 18
Maximum (μέγιστο) = 35

Πίνακας περιγραφικών στατιστικών

Statistics

ΚΟΙΛΙΑΚΟΙ		
N	Valid	339
	Missing	0
Mean		25,89
Std. Error of Mean		,163
Median		26,00
Mode		27
Std. Deviation		3,004
Variance		9,023
Skewness		,136
Std. Error of Skewness		,132
Kurtosis		,047
Std. Error of Kurtosis		,264
Range		17
Minimum		18
Maximum		35
Percentiles	25	24,00
	50	26,00
	75	28,00

Percentiles(εκατοστημόρια)
C25 = Q1 = 24 = 1ο τεταρτημόριο
C50 = Q2 = 26 = 2ο τεταρτημόριο
C75 = Q3 = 28 = 3ο τεταρτημόριο.

Λοξότητα και κύρτωση κοντά στο 0: η κατανομή είναι μάλλον κανονική.

Πίνακας κατανομής συχνότητας για την μεταβλητή "Κοιλιακοί".

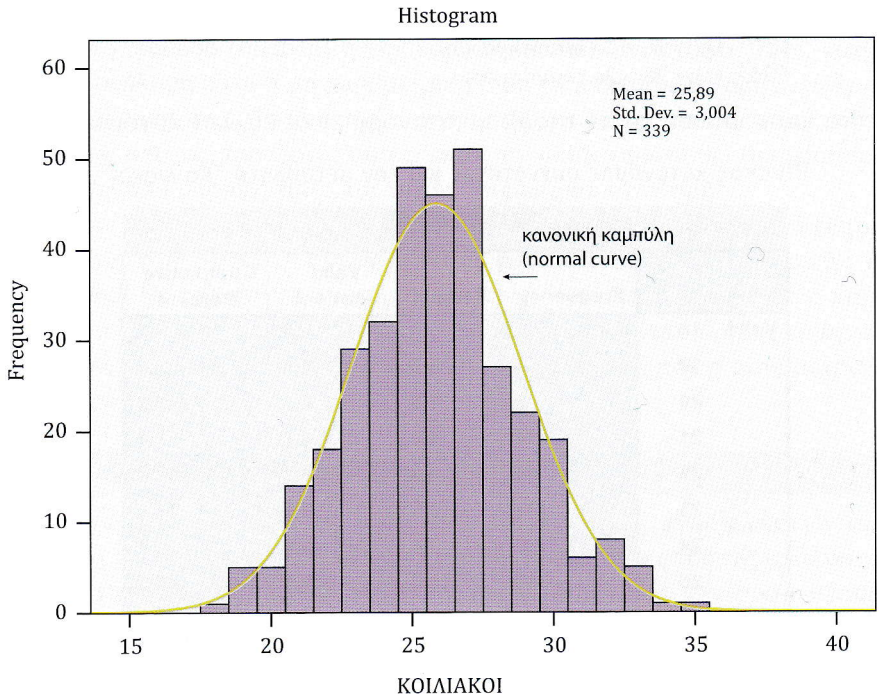
ΚΟΙΛΙΑΚΟΙ

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	18	1	,3	,3	,3
	19	5	1,5	1,5	1,8
	20	5	1,5	1,5	3,2
	21	14	4,1	4,1	7,4
	22	18	5,3	5,3	12,7
	23	29	8,6	8,6	21,2
	24	32	9,4	9,4	30,7
	25	49	14,5	14,5	45,1
	26	46	13,6	13,6	58,7
	27	51	15,0	15,0	73,7
	28	27	8,0	8,0	81,7
	29	22	6,5	6,5	88,2
	30	19	5,6	5,6	93,8
	31	6	1,8	1,8	95,6
	32	8	2,4	2,4	97,9
	33	5	1,5	1,5	99,4
	34	1	,3	,3	99,7
	35	1	,3	,3	100,0
	Total	339	100,0	100,0	

Frequency (συχνότητα) = αριθμός ατόμων με ίδια επίδοση.
Valid (έγκυρο) = έγκυρες διαφορετικές τιμές (βαθμίδες) της μεταβλητής: υπάρχουν **18 διαδοχικές τιμές X**: από 18 έως 35 κοιλιακούς / λεπτό.
Percent (εκατοστιαία) = σχετική συχνότητα, επί τοις % άτομα με ίδια επίδοση.
Cumulative Percent (αθροιστική %) = αθροιστική σχετική συχνότητα.

Ενδεικτικά παραδείγματα:

Ελάχιστη τιμή = 18 με συχνότητα f=1, Μέγιστη τιμή = 35 με συχνότητα f=1,
Επικρατούσα τιμή = 27 (έχει τη μέγιστη συχνότητα f=51),
Διάμεσος = 26 (ανήκει στην κλάση που περιλαμβάνει το 50% των τιμών),
Frequency (συχνότητα) f = 18 για την τιμή 22:
18 άτομα πέτυχαν επιδόσεις από 18 μέχρι 22 κοιλιακούς,
Cumulative percent (αθροιστική εκατοστιαία) c%f = 93.8:
το κατώτερο 93.8% της κατανομής έχει επίδοση από 18 έως 30 κοιλιακούς.



Histogram = **ιστόγραμμα κατανομής** συχνότητας της μεταβλητης "Κοιλιακοί".
Οριζόντιος άξονας: κλίμακα των τιμών (βαθμίδων) της μεταβλητης.
Κατακόρυφος άξονας (frequency): κλίμακα των συχνοτήτων (f).
Ράβδοι = στήλες με ύψος ανάλογο της αντίστοιχης συχνότητας (f).
Γραμμή (πάνω στο ιστόγραμμα): καμπύλη της κανονικής κατανομής.
Η κατανομή είναι συμμετρική και σχεδόν μεσόκυρτη. Αυτό επιβεβαιώνεται από την καλή προσαρμογή της στην καμπύλη της κανονικής κατανομής (γραμμή) , αλλά και από τα τυπικά πηλίκα (z):

Skewness / SE_{sk} = 0.136 / 0.132 = 1.030 < 2, άρα συμμετρία

Kurtosis / SE_{ku} = 0.047 / 0.264 = 0.1780 < 2, άρα μέση κύρτωση.

Η μεταβλητή αυτή πληροί την προϋπόθεση της κανονικότητας (normality) και μπορεί να αναλυθεί με όποια παραμετρική στατιστική μέθοδο, αφού μπορεί να εκπροσωπηθεί από τον μέσο (M) και την τυπική απόκλιση (s) σε εκτιμήσεις πληθυσμιακών παραμέτρων (μ, σ).

2. Διαδικασία Explore (Διερεύνηση)

SPSS: ANALYZE >> DESCRIPTIVE STATISTICS >> EXPLORE

```
EXAMINE VARIABLES=ΚΟΙΛΙΑΚΟΙ
 /PLOT BOXPLOT STEMLEAF HISTOGRAM NPPLOT
 /PERCENTILES(5,10,25,50,75,90,95) HAVERAGE
 /STATISTICS DESCRIPTIVES EXTREME /CINTERVAL 95.
```

M-Estimators				
	Huber's M-Estimator[a]	Tukey's Biweight[b]	Hampel's M-Estimator[c]	Andrews' Wave[d]
ΚΟΙΛΙΑΚΟΙ	25,88	25,87	25,87	25,87

a. The weighting constant is 1.339.

b. The weighting constant is 4.685.

c. The weighting constants are 1.700, 3.400, and 8.500

d. The weighting constant is 1.340*pi.

Ο πίνακας αυτός παρουσιάζει τις τιμές 4 **εκτιμητών-M** (m-estimators) της κεντρικής θέσης της κατανομής και είναι χρήσιμοι για ασύμμετρες κατανομές λόγω ακραίων τιμών που ακόμα και μετά από κατάλληλο μετασχηματισμό (π.χ. log ή square root) παραμένουν ασύμμετρες. Στις περιπτώσεις αυτές χρησιμοποιούνται "εύρωστα στατιστικά" (robust statistics), ανεπηρέαστα από τις ακραίες τιμές.

Στον υπολογισμό του μέσου (M) κάθε τιμή X συμμετέχει με ίδιο "βάρος" 1 (weight = 1) ανεξάρτητα από το σημείο της κατανομής όπου βρίσκεται (στο κέντρο ή στα άκρα). Στον "5% trimmed Mean" οι τιμές στο κεντρικό 90% έχουν βάρος 1 και αυτές στα δύο άκρα (5%) έχουν βάρος 0 (και έτσι αφαιρούνται στον υπολογισμό του μέσου).

Οι M-εκτιμητές υπολογίζουν το κέντρο της κατανομής (μέσος) με "βάρη" που μειώνονται σταδιακά από το 1 στο 0, όσο οι τιμές μετατοπίζονται από το κέντρο στα άκρα της κατανομής. Τα βάρη αυτά υπολογίζονται με βάση την τυπική απόσταση (standardized distance) κάθε τιμής X από το κέντρο της κατανομής.

Όταν η κατανομή είναι συμμετρική, οι 4 M-εκτιμητές δίνουν σχεδόν ίδια τιμή M.

Από τον πίνακα "Descriptives" βλέπουμε ότι ο (αριθμητικός) μέσος είναι M = 25.89. Οι εκτιμήσεις των 4 M-estimators είναι σχεδόν ίδιες (≈ 25.9) με τον μέσο, επειδή η κατανομή είναι συμμετρική.

Σε περιπτώσεις ασύμμετρων κατανομών (skewed distributions) κατάλληλος εκτιμητης-M είναι, εκτός του 5% trimmed mean, αυτός του Huber, που πετυχαίνει καλύτερη εκτίμηση του πληθυσμιακού μέσου (μ). Οι άλλοι 3 M-εκτιμητές σε ασύμμετρες κατανομές δίνουν καλύτερη εκτίμηση του πληθυσμιακού διάμεσου (δμ).

Ο πίνακας "Descriptives" παραθέτει τα περιγραφικά στατιστικά της μεταβλητής.

Mean (μέσος, M) = ΣX / N = 25.89.

Standard error (of mean) (τυπικό σφάλμα του μέσου) $SE_M = S\sqrt{N} = 0.163$.

95% Confidence Interval for Mean (διάστημα εμπιστοσύνης 95% για τον μέσο):

Mean ± SE_{mean} * $t_{(0.95)}$ = 25.89 ± 0.163 * 1.968 = 25.89 ± 0.321
lower bound (κατώτερο όριο) = 25.57, upper bound (ανώτερο όριο) = 26.21.

Η τιμή $t_{(0.95)}$ = 1.968 βρίσκεται από το παράρτημα Δ με δίπλευρο έλεγχο στο α=0.05 και βαθμούς ελευθερίας df = N – 1 = 339 – 1 = 338 ≈ 300.

Descriptives

			Statistic	Std. Error
ΚΟΙΛΙΑΚΟΙ	Mean		25,89	,163
	95% Confidence Interval for Mean	Lower Bound	25,57	
		Upper Bound	26,21	
	5% Trimmed Mean		25,85	
	Median		26,00	
	Variance		9,023	
	Std. Deviation		3,004	
	Minimum		18	
	Maximum		35	
	Range		17	
	Interquartile Range		4	
	Skewness		,136	,132
	Kurtosis		,047	,264

Η ερμηνεία του "95% CI" είναι η εξής: με πιθανότητα 95% ο πληθυσμιακός μέσος (μ) "πέφτει" στα όρια 25.57 – 26.21, με την πιθανότητα αυτή να είναι ελάχιστη στα δύο άκρα (όρια) και μέγιστη στο κέντρο της κατανομής, που αντιστοιχεί ο μέσος 25.89.

5% trimmed mean = (βελτιωμένος) μέσος της κατανομής = 25.85, που υπολογίστηκε αφού αφαιρέθηκαν 5% των τιμών από το κατώτερο και 5% των τιμών από το ανώτερο άκρο της (εύρωστο στατιστικό). Έτσι, ο μέσος αυτός είναι ανεπηρέαστος από τις ακραίες (πολύ μικρές και πολύ μεγάλες) τιμές της κατανομής.

Παρατηρούμε ότι ο μέσος "5% trimmed mean" = 25.85 είναι σχεδόν ίσος με τον μέσο M = 25.89 και αυτό οφείλεται στο γεγονός ότι η κατανομή είναι συμμετρική και ο υπολογισμός των 2 μέσων δεν επηρεάστηκε από τις λίγες και συμμετρικά κατανεμημένες ακραίες τιμές που παρατίθενται στον πίνακα "Extreme Values".

Median = διάμεσος (Md) = 26,

Variance (διασπορά) $s^2 = \Sigma(X - M)^2/(N - 1) = 9.023$,

Standard Deviation (τυπική απόκλιση) $s = \sqrt{\text{variance}} = \sqrt{9.023} = 3.004$,

Minimum (ελάχιστο) = 18, Maximum (μέγιστο) = 35,

Range (εύρος) R = 17 = 35 – 18,

Interquartile range (ενδοτεταρτημοριακό εύρος) = Qi = Q3 – Q1 = 28 – 24 = 4,

Skewness (λοξότητα, Sk) = 0.136,

Std. error of Skewness SE_{sk} (τυπικό σφάλμα λοξότητας) = 0.132,

Kurtosis (κύρτωση) Ku = 0.047,
Std. error of kurtosis SE_{ku} (τυπικό σφάλμα κύρτωσης) = 0.264.

Χρήσιμες Διαπιστώσεις:

Ο μέσος M = 25.89 είναι σχεδόν ίσος με το διάμεσο Md = 26 και αυτό δείχνει ότι η κατανομή τείνει να είναι συμμετρική.

Μια πιο σίγουρη εκτίμηση της λοξότητας και της κύρτωσης της κατανομής μπορεί να επιτευχθεί, αν πάρουμε τα πηλίκα των εκτιμήσεων αυτών προς τα τυπικά τους σφάλματα:

(λοξότητα) / (τυπικό σφάλμα λοξότητας) = 0.136 / 0.132 = 1.03 < 2,
z-τιμή στα όρια ±2 και δείχνει καλή συμμετρία,

(κύρτωση) / (τυπικό σφάλμα κύρτωσης) = 0.047 / 0.264 = 0.178 < 2,
z-τιμή στα όρια ±2 και δείχνει πολύ καλή μέση κύρτωση.

Η συνεκτίμηση των δύο αυτών πηλίκων μας οδηγεί στο συμπέρασμα ότι η κατανομή των 339 τιμών της μεταβλητης "κοιλιακοί" είναι κανονική, πράγμα που επιβεβαιώνεται και με το ιστόγραμμα.

Βήματα στην αξιολόγηση της κανονικότητας της κατανομής:

1) Ιστόγραμμα & Φυλλόγραμμα,

2) Θηκόγραμμα,

3) Λοξότητα & Κύρτωση,

4) Έλεγχος Shapiro–Wilk με σημαντικότητα στο 0.01 ή στο 0.001 ανάλογα με το μέγεθος δείγματος (N).

Extreme Values

			Case Number	Value
ΚΟΙΛΙΑΚΟΙ	Highest	1	279	35
		2	228	34
		3	128	33
		4	202	33
		5	227	33[a]
	Lowest	1	66	18
		2	77	19
		3	74	19
		4	63	19
		5	30	19[b]

a. Only a partial list of cases with the value 33 are shown in the table of upper extremes.

b. Only a partial list of cases with the value 19 are shown in the table of lower extremes.

Στον πίνακα "Extreme Values" οι 5 μεγαλύτερες (highest) και οι 5 μικρότερες (lowest) τιμές τής κατανομής.

Οι 5 μεγαλύτερες τιμές είναι 33, 33, 33, 34, 35 και αντιστοιχούν στα άτομα (case) 227, 202, 128, 228, 279.

Οι 5 μικρότερες τιμές είναι 18, 19, 19, 19, 19 και ανήκουν στα άτομα (case) 66, 77, 74, 63, 30.

Case number = περίπτωση (άτομο), Value = τιμή.

Ο πίνακας αυτός περιλαμβάνει πάντα τις 5 ανώτερες και τις 5 κατώτερες τιμές, χωρίς να αξιολογεί αν είναι απόμακρες (outliers) ή ακραίες (extreme).

Το SPSS αξιολογεί το *πόσο ακραία* είναι κάποια τιμή X με βάση τα εξής κριτήρια:

α) αν $X > Q3 + 1.5 * (Q3 - Q1)$ ή $X < Q1 - 1.5 * (Q3 - Q1)$
τότε η τιμή X ορίζεται ως απόμακρη (outlier)

β) αν $X > Q3 + 3 * (Q3 - Q1)$ ή $X < Q1 - 3 * (Q3 - Q1)$
τότε η τιμή X ορίζεται ως απόμακρη (outlier).

Στον πίνακα "Percentiles" βλέπουμε ότι Q3 = 28 & Q1 = 24, οπότε:

$$Q3 + 1.5 * (Q3 - Q1) = 28 + 1.5 * (28 - 24) = 32$$

$$Q1 - 1.5 * (Q3 - Q1) = 24 - 1.5 * (28 - 24) = 18$$

δηλαδή κάθε τιμή $X > 32$ ή $X < 18$ αποτελεί μια απόμακρη τιμή (outlier),

$$Q3 + 3 * (Q3 - Q1) = 28 + 3 * (28 - 24) = 40$$

$$Q1 - 3 * (Q3 - Q1) = 24 - 3 * (28 - 24) = 12$$

δηλαδή κάθε τιμή $X > 40$ ή $X < 12$ αποτελεί μια ακραία τιμή (extreme).

Υπάρχει μία μόνο απόμακρη (>32) τιμή, η X=35 (άτομο, case = 279).
Δεν υπάρχουν ακραίες τιμές (> 40 ή < 12) και αυτό φαίνεται και στο θηκόγραμμα (boxplot), όπου αποτυπώνεται μόνο η απόμακρη τιμή X = 35 (case = 279).

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average (Definition 1)	ΚΟΙΛΙΑΚΟΙ	21,00	22,00	24,00	26,00	28,00	30,00	31,00
Tukey's Hinges	ΚΟΙΛΙΑΚΟΙ			24,00	26,00	28,00		

Στον πίνακα "**Percentiles**" δίνονται τα **βασικά εκατοστημόρια**:

5ο εκατοστημόριο = C5 = 21,

10ο εκατοστημόριο = C10 = 22,

25ο εκατοστημόριο = C25 = Q1 (1ο τεταρτημόριο) = 24,

50ο εκατοστημόριο = C50 = Median = 26,

75ο εκατοστημόριο = C75 = Q3 (3ο τεταρτημόριο) = 28,

90ο εκατοστημόριο = C90 = 30,

95ο εκατοστημόριο = C95 = 31.

Επίσης, στον πίνακα αυτόν δίνονται οι εκτιμήσεις "Tukey Hinges" των 3 τεταρτημόριων, που για το συγκεκριμένο δείγμα 339 δεδομένων είναι ίδια με τα Q1, Q2, Q3. Ο υπολογισμός των Tukey Hinges είναι λίγο διαφορετικός από αυτόν των εκατοστημορίων: 1) ο διάμεσος (Md) χωρίζει την κατανομή σε 2 μισά (με 50% των τιμών στο κάθε ένα), 2) κάθε μισό χωρίζεται από τα hinges σε δύο ίσα μισά (με 25% των τιμών στο κάθε ένα).

Η αξιολόγηση των μέτρων κεντρικής θέσης, των μέτρων διασποράς και της μορφής της κατανομής (δηλαδή κατά πόσο η κατανομή είναι συμμετρική & μεσόκυρτη και επομένως κανονική) πρέπει να γίνεται συνδυαστικά με την αξιολόγηση του ιστογράμματος και άλλων σχετικών γραφημάτων της κατανομής.

Ο έλεγχος κανονικότητας της κατανομής μπορεί να γίνει και πιθανολογικά μέσω των ελέγχων Kolmogorov-Smirnov και Shapiro-Wilk (Shapiro & Wilk, 1965; Shapiro et. al. 1968).

Οι έλεγχοι αυτοί συγκρίνουν την κατανομή του δείγματος με την κανονική αλλά τείνουν να δίνουν κανονικότητα σε μικρά δείγματα (π.χ. $N < 30$) και μη κανονικότητα σε μεγάλα δείγματα (π.χ. $N > 100$). Για το λόγο αυτό η σημαντικότητά τους μπορεί να ελεγχθεί στο $\alpha = 0.01$ ή ακόμα και στο $\alpha = 0.001$ ή αν το δείγμα είναι πολύ μεγάλο, μπορεί η αξιολόγηση της κατανομής να γίνει με το ιστογράμμα και τα μέτρα λοξότητας και κύρτωσης.

Ο έλεγχος κανονικότητας Shapiro-Wilk είναι ισχυρότερος του Kolmogorov-Smirnov για δείγματα $N \leq 50$ και είναι γενικά προτιμητέος (Razali & Wah, 2011).

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
ΚΟΙΛΙΑΚΟΙ	,093	339	,000	,987	339	,004

a. Lilliefors Significance Correction

Με βάση τα αποτελέσματα των 2 αυτών ελέγχων θα έπρεπε να συμπεράνουμε ότι η κατανομή δεν είναι κανονική (συμμετρική και μεσόκυρτη), επειδή και οι δύο έλεγχοι έδειξαν ότι υπάρχει στατιστικώς σημαντική διαφορά μεταξύ της δειγματικής κατανομής και του υποδείγματος της κανονικής κατανομής (normal curve). Όμως, στο παράδειγμά μας έχουμε ένα αρκετά μεγάλο δείγμα ($N = 339$) και, έτσι, οι 2 αυτοί έλεγχοι μπορούν να αγνοηθούν και να βασίσουμε την αξιολόγηση της κανονικότητας της κατανομής σε άλλα σχετικά στατιστικά, όπως στο ιστόγραμμα (το οποίο είναι συμμετρικό) και στις τιμές της λοξότητας και της κύρτωσης σε σχέση με τα τυπικά τους σφάλματα, που όπως είδαμε ήδη συνιστούν κανονικότητα.

Στο γράφημα **Stem-and-Leaf** (στέλεχος-&-φύλλο) παρατίθενται όλες οι αρχικές τιμές και οι επαναλήψεις τους. Το σχήμα αυτό δείχνει τη συμμετρία της κατανομής και την ακραία (ανώτερη) τιμή. Για παράδειγμα, υπάρχουν 5 τιμές 20, 32 τιμές 24, 22 τιμές 29 και 1 τιμή 34. Από το σχήμα αυτό φαίνεται η κορυφή (27) καθότι έχει συχνότητα $f = 51$ (όπως φαίνεται και από το μέγιστο μήκος της αντί-

στοιχης στήλης). Το γράφημα αυτό έχει όλα τα στοιχεία του ιστογράμματος και επιπλέον δίνει όλες τις τιμές σε διαδοχικές ομάδες και τυχόν ακραίες τιμές (extremes) με βάση τα κριτήρια που αναπτύχθηκαν πιο πάνω.

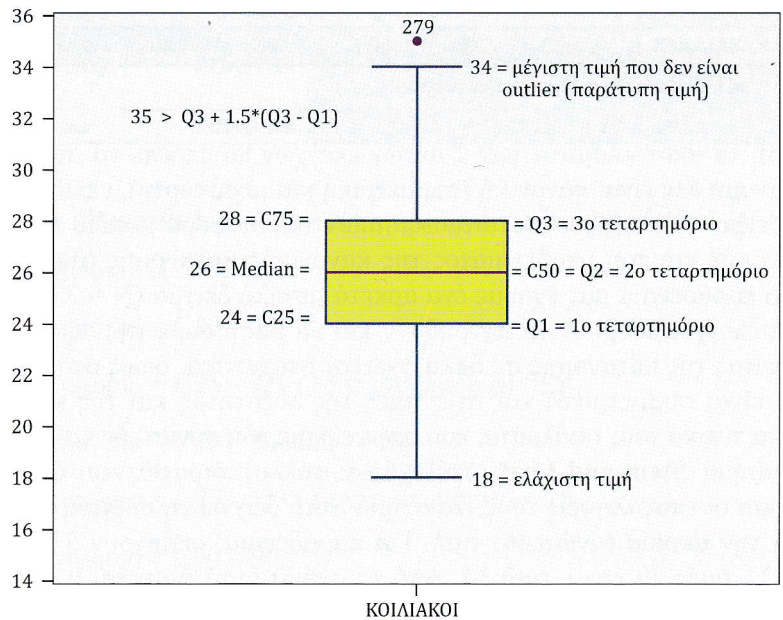
Από το γράφημα αυτό μπορούμε να εντοπίσουμε μερικά χρήσιμα στοιχεία:

```
ΚΟΙΛΙΑΚΟΙ Stem-and-Leaf Plot

 Frequency
          Stem &  Leaf
     1.00   18 .  0
     5.00   19 .  00000
     5.00   20 .  00000
    14.00   21 .  00000000000000
    18.00   22 .  000000000000000000
    29.00   23 .  00000000000000000000000000000
    32.00   24 .  00000000000000000000000000000000
    49.00   25 .  0000000000000000000000000000000000000000000000000
    46.00   26 .  0000000000000000000000000000000000000000000000
    51.00   27 .  000000000000000000000000000000000000000000000000000
    27.00   28 .  000000000000000000000000000
    22.00   29 .  00000000000000000000
    19.00   30 .  0000000000000000000
     6.00   31 .  000000
     8.00   32 .  00000000
     5.00   33 .  00000
     1.00   34 .  0
     1.00 Extremes    (>=35.0)

 Stem width:          1
 Each leaf:       1 case(s)
```

Θηκόγραμμα (boxplot) της κατανομής των 339 τιμών.



1) Υπάρχει μια απόμακρη (outlier) τιμή ίση με 35, που απέχει πάνω από ενάμιση μήκος θήκης από το 3ο τεταρτημόριο Q3 = 34, και ανήκει στο άτομο (case) 279:

35 > Q3 + 1.5 * (Q3 – Q1) = 28 + (1.5 * (28 – 24)) = 28 + (1.5 * 4) = 28 + 6 = 32,

2) Το 3ο τεταρτημόριο (Q3, 3st quartile) είναι η τιμή 28 και διαχωρίζει το πρώτο 75% των τιμών (C75) από το υπόλοιπο 25% των τιμών της κατανομής.

3) Ο διάμεσος (Median) είναι η τιμή 26 χωρίζει την κατανομή σε δύο ίσα τμήματα από 50% των τιμών σε κάθε ένα.

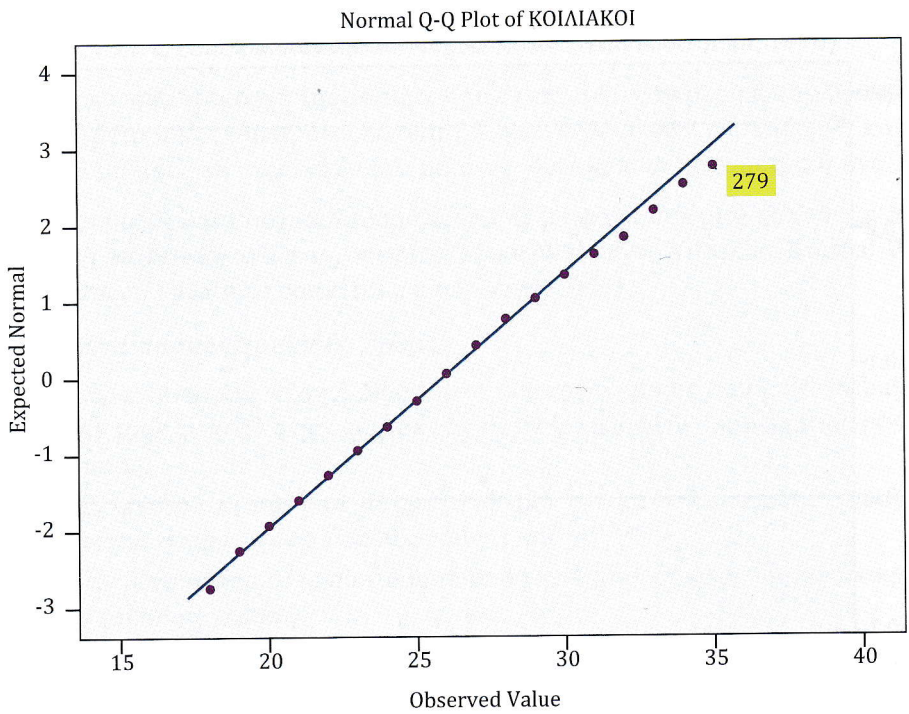
4) Το 1ο τεταρτημόριο (Q1, 1st quartile) είναι η τιμή 18 και διαχωρίζει το πρώτο 25% των τιμών (C25) από το υπόλοιπο 75% των τιμών της κατανομής.

5) Ο διάμεσος (26) χωρίζει τη θήκη (πλαίσιο) σε δύο σχεδόν ίσα μισά.

6) Υπάρχει ίση απόσταση από το διάμεσο (26) των 2 whisker (2 οριζόντιων γραμμών), δηλαδή της μεγαλύτερης (34) και της μικρότερης (18) τιμής που δεν είναι απόμακρη (outlier).

7) Με βάση τα στοιχεία αυτά συμπεραίνουμε ότι η κατανομή είναι συμμετρική.

Γράφημα "Normal Q-Q Plot" της κατανομής των 339 τιμών.



Quantile (q) είναι η τιμή X_i που ορίζει το πρώτο q% των τιμών της κατανομής: οι N τιμές X σε σειρά μεγέθους διαιρούν την κλίμακα X σε N + 1 μέρη. Έτσι, η αναλογία των τιμών που πέφτουν πριν την X_i είναι η i/(N + 1) και για κάθε q η $i = q\ (N + 1)$. Π.χ. σε N=50 τιμές η quantile q = 0.21 είναι η X που ορίζει το πρώτο 21% των τιμών: η i = q (N + 1) = 0.21 (50 + 1) = 10.71η, δηλαδή η X μεταξύ $10^{ης}$ και $11^{ης}$ σε σειρά τιμής.

Normal Q-Q Plot: *Quantile-Quantile Plot.* Τα quantiles των παρατηρήσεων υποτυπώνονται έναντι τω αντίστοιχων της κανονικής (normal expected).

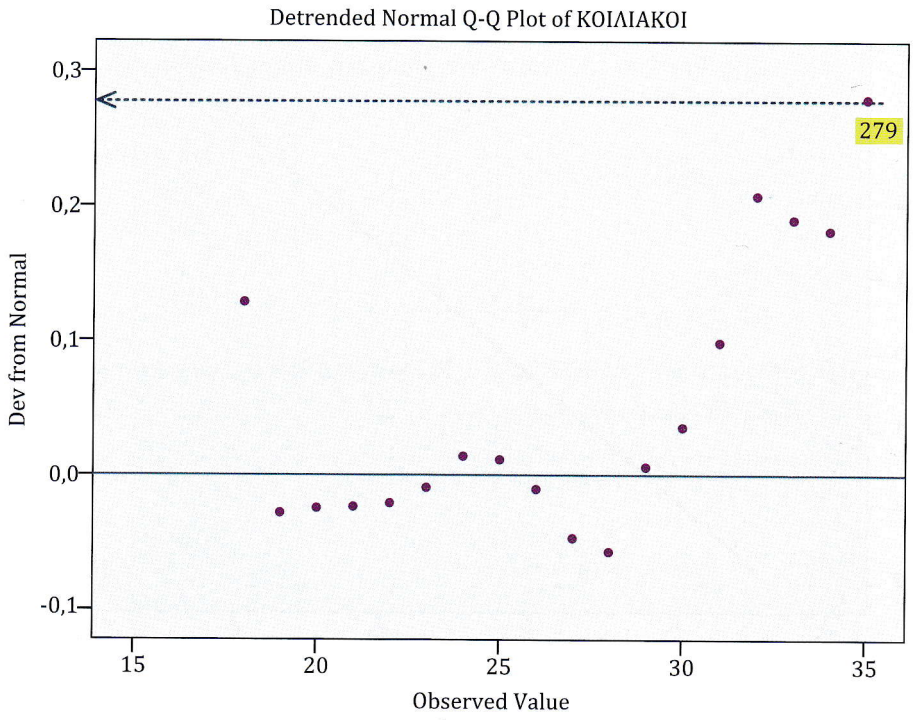
Το υποτύπωμα (plot) δίνει τα σημεία (ο) που αντιπροσωπεύουν την πραγματική κατανομή και μια ευθεία γραμμή που συμβολίζει την τέλεια κανονικότητα.

Στην κατανομή αυτή υπάρχουν 18 διαδοχικές ακέραιες τιμές X (observed value) και έτσι παρήχθησαν 18 αντίστοιχες quantiles που ως τυπικές τιμές z αντιπαραβάλλονται με τις τυπικές τιμές z των αντίστοιχων quantiles της κανονικής κατανομής.

Η ερμηνεία του γραφήματος αυτού είναι η εξής:

Όσο τα σημεία που προέκυψαν από τις τιμές (observed values) είναι κοντά ή και πάνω στην γραμμή, τόσο η κατανομή τείνει να είναι κανονική (normal). Το συγκεκριμένο υποτύπωμα δείχνει ότι με εξαίρεση την περίπτωση (case) 279 (που όπως είδαμε είναι η απόμακρη τιμή 35) τα σημεία πέφτουν πολύ κοντά στη γραμμή της κανονικότητας και επομένως η κατανομή είναι κανονική.

Γράφημα "Detrended Normal Q-Q Plot" της κατανομής των 339 τιμών.



Detrended Normal Q-Q Plot: Detrended *Quantile-Quantile Plot.* Οι τυποποιημένες (z) αποκλίσεις (Dev from Normal) των quantiles από την κανονική κατανομή υποτυπώνονται έναντι της μηδενικής απόκλισης (γραμμή).

Το γράφημα αυτό αποτελεί προέκταση του "Normal Q-Q Plot" και αξιολογεί σε τυπικές τιμές z πόσο αποκλίνουν από την κανονικότητα οι τιμές X της κατανομής.

Τα σημεία (**o**) του γραφήματος αυτού προκύπτουν ως εξής: κάθε quantile Xi μετασχηματίζεται σε z-τιμή (απόκλιση / τυπική απόκλιση) και αφαιρείται από την z-τιμή της αντίστοιχης προσδοκώμενης quantile της κανονικής (expected normal). Οι αποκλίσεις αυτές υποτυπώνονται έναντι των αντίστοιχων αρχικών (observed).

Η ερμηνεία του γραφήματος αυτού είναι η εξής: Τα σημεία που απέχουν πολύ από τη γραμμή κανονικότητας συνεισφέρουν περισσότερο στην όποια απόκλιση της συνολικής κατανομής από την κανονικότητα.

Στο παράδειγμα των 339 τιμών λίγες μόνο τιμές αποκλίνουν από την κανονικότητα με κύρια αυτή της περίπτωσης 279 με τιμή X = 35 που απέχει ≈ 0.28z από την κανονικότητα, ενώ οι άλλες απέχουν το πολύ μέχρι 0.2z και αφορούν όσες είναι > 30. Όμως, οι αποκλίσεις αυτές δεν επαρκούν για να αμφισβητηθεί η κανονικότητα της κατανομής. Έτσι, συνάγεται ότι η συγκεκριμένη κατανομή είναι αδρά κανονική.

Τι γίνεται σε περίπτωση μεγάλης ασυμμετρίας στην κατανομή;

Αν σε μια ανάλυση παρατηρηθεί **μεγάλη ασυμμετρία,** τότε έχουμε 3 επιλογές:

1) να μετασχηματισθούν οι τιμές X ώστε να επιτευχθεί η συμμετρία της κατανομής και να συνεχίσουμε με κάποια παραμετρική στατιστική ανάλυση, όπως π.χ. t-test, ANOVA, Pearson correlation, regression (Bland & Altman, 1996).

2) να διερευνηθεί η εκδοχή της αφαίρεσης από την ανάλυση μερικών πολύ ακραίων τιμών (μετά από επαρκή αιτιολόγηση), ώστε να βελτιωθεί περαιτέρω η κατανομική δομή προς τη συμμετρία και μετά να γίνει κάποια παραμετρική ανάλυση.

3) να εφαρμοσθεί στα αρχικά δεδομένα (χωρίς μετασχηματισμό) κάποια μη παραμετρική στατιστική ανάλυση, όπως π.χ. Mann-Whitney ή Wilcoxon, Kruskal-Wallis ή Friedman, Spearman correlation κ.ά (Siegel, 1956).

Στα επόμενα **παραδείγματα** θα δούμε:

(α) τα μέτρα κεντρικής θέσης, διασποράς και κατανομικής δομής (distributional shape) μιας μεταβλητής με αρκετή θετική λοξότητα (substantial positive skewness),

(β) πώς βελτιώνονται αυτά τα μέτρα μετά από τον κατάλληλο μετασχηματισμό των αρχικών τιμών (data transformation) και

(γ) πως βελτιώνονται τα μέτρα αυτά μετά από την αφαίρεση μερικών πολύ ακραίων τιμών (extreme values).

Σύμφωνα με κρατούσες απόψεις έμπειρων αναλυτών (π.χ. Tabachnick & Fidell, 2007, σελ. 86-89), τα αρχικά δεδομένα μιας μεταβλητής μπορούν να μετασχηματισθούν ως εξής σε σχέση με την κατεύθυνση και το βαθμό ασυμμετρίας της κατανομής:

Κατανομή με	Μετασχηματισμός
μέτρια θετική λοξότητα	SQRT(X)
έντονη θετική λοξότητα χωρίς μηδενικές τιμές	LOG10(X)
έντονη θετική λοξότητα με μηδενικές τιμές	LOG10(X+C)
μέτρια αρνητική λοξότητα	SQRT(D–X)
έντονη αρνητική λοξότητα	LOG10(D–X)

C = 1 – ελάχιστη τιμή, D = μέγιστη τιμή + 1.

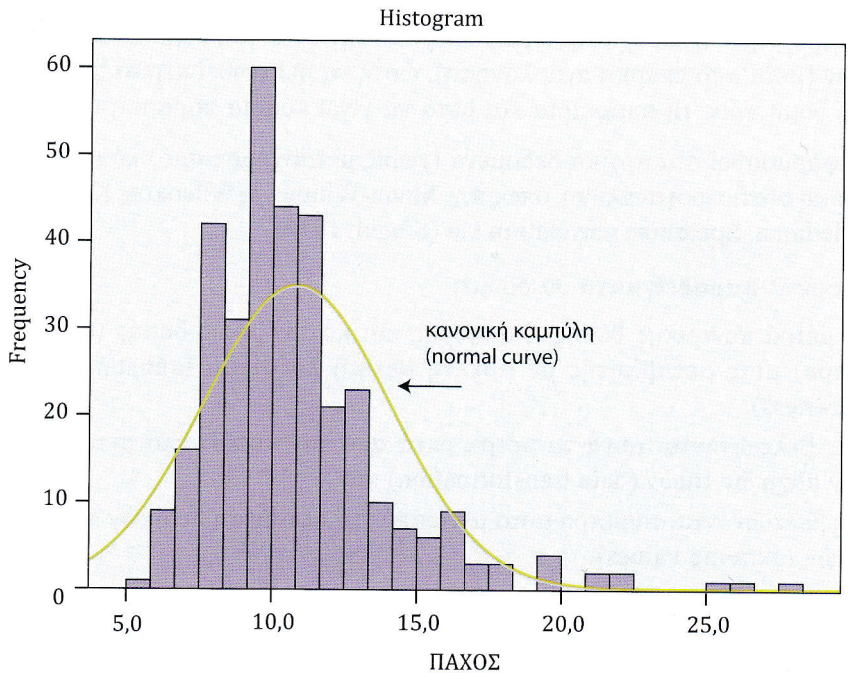
7.2 Μια ασύμμετρη κατανομή (λοξή θετικά)

Τα δεδομένα αφορούν τη μεταβλητή *δερματικό πάχος* (σε mm) δείγματος 339 ανδρών 18-26 ετών. Θα γίνει διερεύνηση των μέτρων κεντρικής θέσης, διασποράς και κατανομικής δομής.

Μια πρώτη εικόνα της κατανομικής δομής της μεταβλητής *πάχος* μας δίνει το **ιστόγραμμα** των 339 τιμών. Η κατανομή είναι έντονα λεπτόκυρτη (leptoku-rtic) και ασύμμετρη (λοξή) δεξιά (positively skewed).

Αυτό φαίνεται και από την απόκλιση του ιστογράμματος από την κανονική καμπύλη (γραμμή πάνω στο γράφημα).

Στο δεξί άκρο της κατανομής παρατηρούνται μερικές μεγάλες τιμές πάνω από το 20 και κάποιες λιγότερες πάνω από το 25.



Οι τιμές αυτές δίνουν τη θετική λοξότητα στην κατανομή και επηρεάζουν δραστικά προς τα πάνω την εκτίμηση του αριθμητικού μέσου.

Στον πίνακα "Descriptives" παρατίθενται τα περιγραφικά στατιστικά.
Το μέτρο 95% Confidence Interval for Mean εξηγήθηκε στη σελ. 75 και 122.

Descriptives

			Statistic	Std. Error
ΠΑΧΟΣ	Mean		10,818	,1757
	95% Confidence Interval for Mean	Lower Bound	10,472	
		Upper Bound	11,163	
	5% Trimmed Mean		10,512	
	Median		10,100	
	Variance		10,464	
	Std. Deviation		3,2348	
	Minimum		5,8	
	Maximum		28,0	
	Range		22,2	
	Interquartile Range		2,9	
	Skewness		1,856	,132
	Kurtosis		5,322	,264

Οι τιμές έχουν εύρος 22.2 με μέγιστο 28.0 και ελάχιστο 5.8.

Ο μέσος M = 10.818 διαφέρει αισθητά από τον μέσο "5% Trimmed" = 10.512 και από το διάμεσο Md = 10.100. Αυτό αποτελεί άλλη ένδειξη ότι η κατανομή είναι λοξή δεξιά (θετικά) με κάποιες ακραίες μεγάλες τιμές και επιβεβαιώνεται από τις τιμές λοξότητας και κύρτωσης:

Skewness = 1.856 που δείχνει έντονη θετική λοξότητα,

Kurtosis = 5.322 και δείχνει λεπτόκυρτη κατανομή.

και από τα πηλίκα:

$Sk / SE_{sk} = 1.856 / 0.132 = 14.06 >> 2$, άρα έντονα ασύμμετρη θετικά,

$Ku / SE_{ku} = 5.322 / 0.264 = 20.16 >> 2$, άρα έντονα λεπτόκυρτη.

Συνδυάζοντας τα 2 αυτά πηλίκα διαπιστώνουμε πιθανολογικά ότι η κατανομή των 339 τιμών πάχους δεν είναι κανονική. Για να δούμε ειδικότερα που οφείλεται αυτή η έντονη απόκλιση από την κανονικότητα εξετάζουμε και τα υπόλοιπα αποτελέσματα της διαδικασίας explore.

Στο γράφημα "Stem-and-Leaf" παρατίθενται οι αρχικές τιμές σε ομάδες.
Κύριο στοιχείο είναι η αναφορά στις 22 ανώτερες τιμές που είναι ≥ 16.2 και που το φυλλόγραμμα τις ομαδοποιεί μαζί δείχνοντας ουσιαστικά πώς θα φαινόταν η κατανομή χωρίς αυτές, δηλαδή με πιο συμμετρική μορφή.

Από τον πίνακα "Percentiles" διαπιστώνουμε ότι το ενδοτεταρτημοριακό εύρος

$$Qi = Q3 - Q1 = 11.8 - 8.9 = 2.9, \text{ οπότε}$$

```
ΠΑΧΟΣ Stem-and-Leaf Plot

Frequency
     Stem & Leaf
  1    5. 8
 13    6. 1333445567778
 37    7. 000133344444555556666677777777899999
 40    8. 000001122223333334445555667777789999999
 68    9.
000001112222223333333344444444444445555555666666677777778888999999
 49   10. 00000001111122222223333333344444445566777888999999
 50   11. 000011111112222333444444445555555556667788888888999
 23   12. 00122344455555666778899
 18   13. 000011123345667788
  8   14. 14448888
  8   15. 25556899
  2   16. 01
22 Extremes   (>=16.2)

Stem width:       1.0
Each leaf:        1 case(s)
```

$$Q3 + 1.5 * Qi = 11.8 + 1.5 * 2.9 = 16.15$$

--> όσες τιμές είναι > 16.15 είναι απόμακρες (outliers, o)

$$Q3 + 3 * Qi = 11.8 + 3 * 2.9 = 20.5$$

--> όσες τιμές είναι > 20.5 είναι ακραίες (extremes, *).

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average (Definition 1)	ΠΑΧΟΣ	7,050	7,600	8,900	10,100	11,800	14,850	16,900
Tukey's Hinges	ΠΑΧΟΣ			8,900	10,100	11,800		

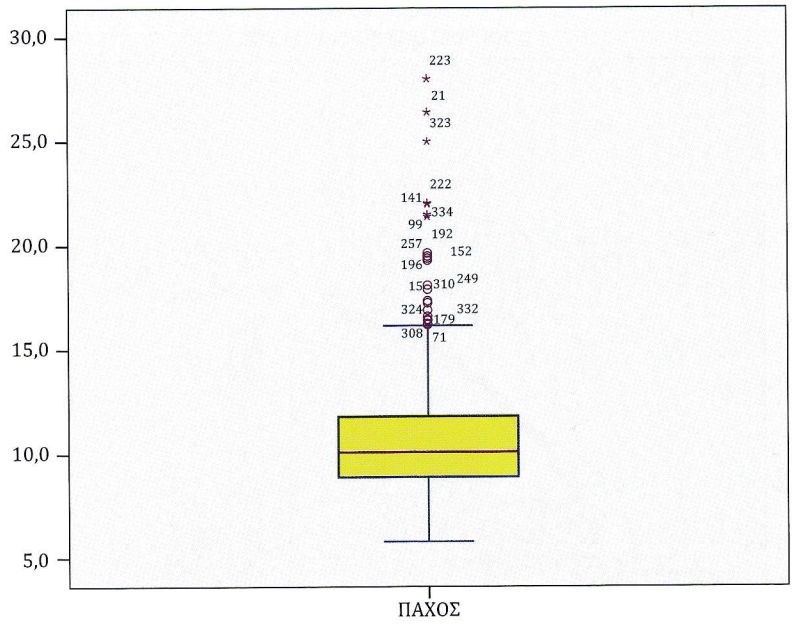
Στο **θηκόγραμμα** (boxplot) ο διάμεσος (μαύρη γραμμή) δεν χωρίζει ισόποσα τη θήκη, προφανώς λόγω των 22 πολύ μεγάλων τιμών, οπότε συμπεραίνουμε ότι η κατανομή δεν είναι κανονική, αλλά εντόνως λοξή δεξιά.

Όσες περιπτώσεις (cases) συμβολίζονται με **o** αφορούν **απόμακρες τιμές** (outliers), δηλαδή τιμές που απέχουν πάνω από 1.5 μήκος θήκης (Qi = 2.9) από το 1ο ή το 3ο τεταρτημόριο (εδώ έχουμε μόνο ανώτερες απόμακρες):

$$Q3 + 1.5 * Qi = 11.8 + 1.5 * 2.9 = 16.15 \text{ --> όσες είναι} > 16.15.$$

Όσες περιπτώσεις (cases) συμβολίζονται με * αφορούν **ακραίες τιμές** (extreme), δηλαδή τιμές που απέχουν πάνω από 3 μήκη θήκης (Qi = 2.9) από το 1ο ή το 3ο τεταρτημόριο (εδώ έχουμε μόνο ανώτερες ακραίες):

$$Q3 + 3 * Qi = 11.8 + 3 * 2.9 = 20.5 \text{ --> όσες είναι} > 20.5,$$



δηλαδή τα άτομα 223, 21, 323, 334, 222, 99, 141 με αντίστοιχες τις 7 τιμές 28, 26.4, 25, 22.1, 22, 21.5, 21.4. Στο κεφάλαιο 7.4 θα δούμε πόσο βελτιώνεται η διασπορά και η κατανομή μετά την αφαίρεσή τους.

Στο **γράφημα** "**Normal Q-Q Plot**" και "**Detrended Normal Q-Q Plot**" βλέπουμε πιο καθαρά τις 7 αυτές περιπτώσεις, αφού με διπλό "κλικ" πάνω στο γράφημα και δεξί "κλικ" πάνω σε κάθε μία από αυτές τις περιπτώσεις (cases) επιλέξαμε "show data values", ώστε να εμφανιστούν οι κωδικοί τους.

7.3 Μια μετασχηματισμένη κατανομή (Συμμετρία)

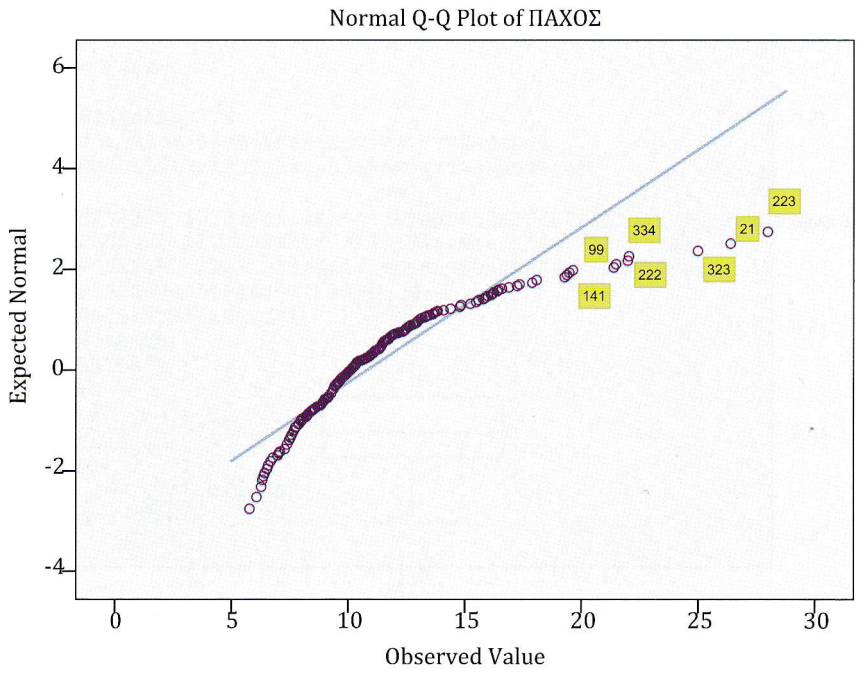
Όπως είδαμε στο κεφάλαιο 7.2, η κατανομή των 339 τιμών "Πάχους" έχει μεγάλη θετική ασυμμετρία (substantial positive skewness) και επομένως καταλληλότερος μετασχηματισμός τους είναι ο LOG10(X), ώστε να μειωθεί δραστικά η επίδραση των ακραίων τιμών και η ασυμμετρία της κατανομής.

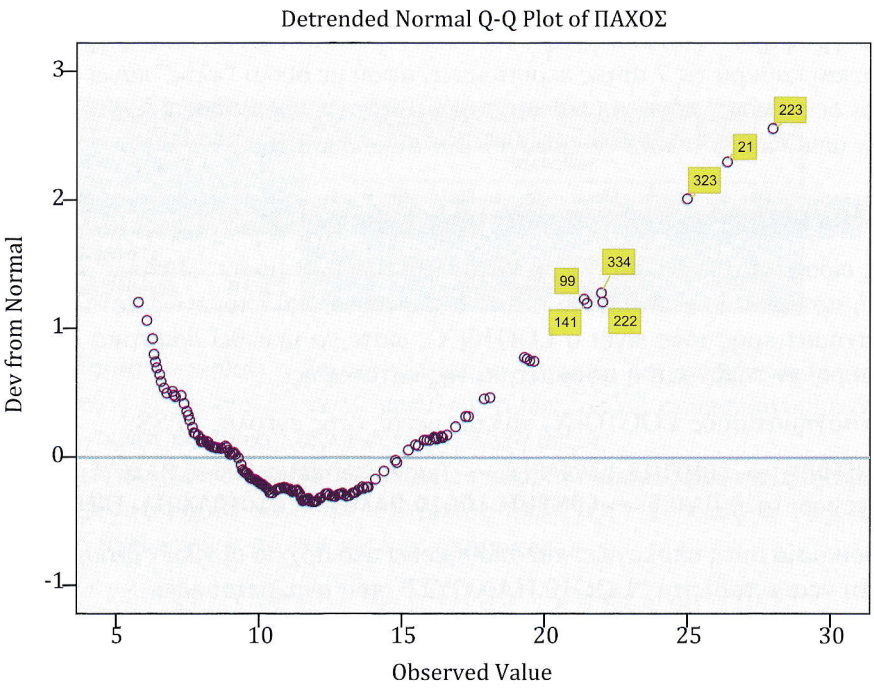
Ο μετασχηματισμός LOG1O(X) γίνεται με τις εξής εντολές SPSS:

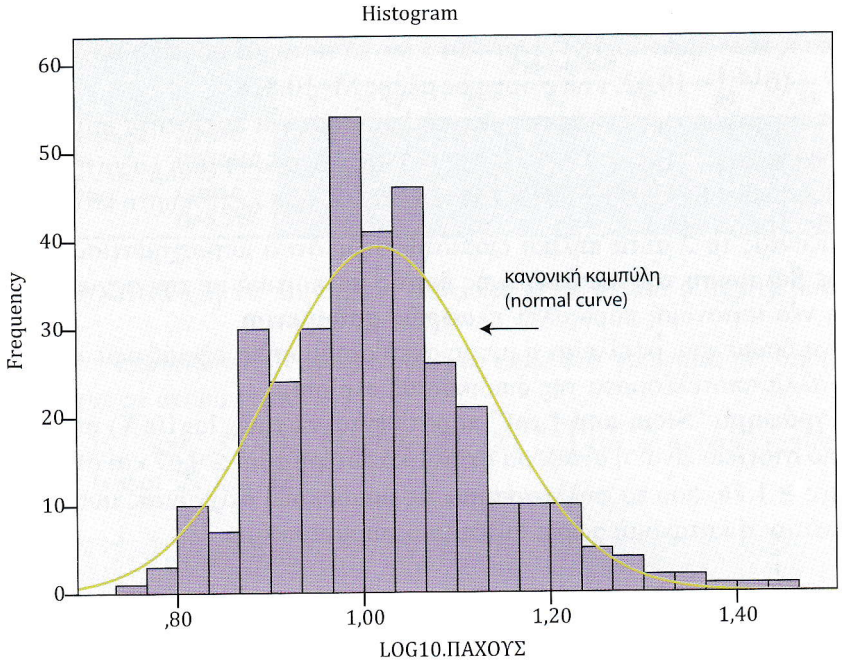
TRANSFORM >> COMPUTE VARIABLE >> Target Variable: LOG10.ΠΑΧΟΥΣ >> Numeric Expression: Lg10(ΠΑΧΟΣ) >> **COMPUTE LOG10.ΠΑΧΟΥΣ=LG10(ΠΑΧΟΣ). EXECUTE.**

Η διαδικασία αυτή υπολογίζει και αποθηκεύει στο αρχείο αρχικών δεδομένων (data file) τη νέα μεταβλητή "LOG10.ΠΑΧΟΥΣ", που αναλύεται εδώ.

Μια πρώτη εικόνα της κατανομής της μεταβλητής "LOG10.ΠΑΧΟΥΣ" μας δίνει το **ιστόγραμμα** των 339 μετασχηματισμένων τιμών. Η κατανομή είναι χοντρικά συμμετρική (με ελάχιστη θετική λοξότητα). Επισημαίνεται ότι σε πραγματικές έρευνες βελτιώσεις (μετασχηματισμοί) αυτού του βαθμού είναι επιθυμητές, καθότι επιτρέπουν την εφαρμογή παραμετρικής στατιστικής ανάλυσης (Bland & Altman, 1996).


Normal Q-Q Plot of ΠΑΧΟΣ
Expected Normal
Observed Value
6
4
2
0
-2
-4
0
5
10
15
20
25
30
223
21
334
99
323
222
141


Detrended Normal Q-Q Plot of ΠΑΧΟΣ
Dev from Normal
Observed Value
3
2
1
0
-1
5
10
15
20
25
30
223
21
323
334
99
141
222



Στο δεξί άκρο της κατανομής παρατηρούνται λίγες μεγάλες τιμές πάνω από το log10(X) = 1.4, που αντιστοιχεί στην αρχική τιμή $X = 10^{1.4} = 25$.

Οι τιμές αυτές δίνουν ελάχιστη θετική λοξότητα στην κατανομή και επανέρχονται στην αρχική κλίμακα μέτρησης (mm) ως $X = 10^{LOG(X)}$.

Στον πίνακα "Descriptives" παρατίθενται τα περιγραφικά στατιστικά.

Descriptives

			Statistic	Std. Error
LOG10.ΠΑΧΟΥΣ	Mean		1,0180	,00623
	95% Confidence Interval for Mean	Lower Bound	1,0058	
		Upper Bound	1,0303	
	5% Trimmed Mean		1,0131	
	Median		1,0043	
	Variance		,013	
	Std. Deviation		,11467	
	Minimum		,76	
	Maximum		1,45	
	Range		,68	
	Interquartile Range		,12	
	Skewness		,721	,132
	Kurtosis		1,019	,264

Το μέτρο 95% Confidence Interval for Mean εξηγήθηκε ήδη στα κεφ.4.6.2 & 7.1.2.

Οι τιμές log10(X) έχουν εύρος 0.68 με μέγιστο 1.45 και ελάχιστο 0.76.

Ο μέσος των τιμών log(X) είναι 1.018 και αντιστοιχεί σε μέσο αρχικών τιμών = $10^{\log(X)} = 10^{1.018} = 10.42$, ενώ ο αρχικός μέσος M=10.818.

Η κατανομή διατηρεί ελάχιστη θετική λοξότητα και λεπτότητα καθότι:

Skewness / SEsk = 0.721 / 0.132 = 5.46 > 2, άρα θετική λοξότητα,

Kurtosis / SEku = 1.019 / 0.264 = 3.86 > 2, άρα λεπτόκυρτη κατανομή.

Συνδυάζοντας τα 2 αυτά πηλίκα διαπιστώνουμε ότι ο μετασχηματισμός log10(X) επέφερε **βελτίωση της κατανομικής δομής** συγκριτικά με την αρχική (κεφ. 7.2), αλλά η νέα κατανομή παραμένει **ελαφρώς ασύμμετρη**.

Για να δούμε πού οφείλεται η μικρή αυτή ασυμμετρία εξετάζουμε στη συνέχεια και τα άλλα αποτελέσματα της διαδικασίας explore.

Στο **γράφημα "Stem-and-Leaf"** παρατίθενται οι τιμές log10(X) σε ομάδες.

Κύριο στοιχείο είναι η αναφορά στην 1 κατώτερη τιμή ≤ 0.67 και στις 13 ανώτερες τιμές ≥ 1.26, που το φυλλόγραμμα τις ομαδοποιεί δείχνοντας ουσιαστικά πώς θα φαινόταν η κατανομή χωρίς αυτές (πιο συμμετρική).

LOG10.ΠΑΧΟΥΣ Stem-and-Leaf Plot

```
Frequency
    Stem & Leaf
 1   Extremes   (=<.76)
 3    7. 899
13    8. 0001112223444
34    8. 566666666777778888888888889999999
37    9. 000000011111111122222233333333444444
71    9. 55555555556666666666666677777777777777777777888888888888888999999999999
64   10. 0000000000000000011111111111111122222233333333344444444444444444
49   10. 55555555556666666666666777777777777888899999999
28   11. 00000001111111112222333333 44
13   11. 5557777899999
12   12. 000001112233
 1   12. 5
13   Extremes  (>=1.26)

Stem width:     .10
Each leaf:    1 case(s)
```

Από τον πίνακα "Percentiles" διαπιστώνουμε ότι το ενδοτεταρτημοριακό εύρος (Qi) είναι Q3 – Q1 = 1.0719 – 0.9494 = 0.1225, οπότε η μία μικρότερη τιμή log10(X) είναι απόμακρη (outlier) ως < 0.7657

$$Q1 - 1.5 * Qi = 0.9494 - 1.5 * 0.1225 = 0.7657 \text{ --> } X = 10^{0.7657} = 5.83,$$

υπάρχουν άλλες 11 απόμακρες (outliers) > 1.2557

$$Q3 + 1.5 * Qi = 1.0719 + 1.5 * 0.1225 = 1.2557 \text{ --> } X = 10^{1.2557} = 18.02,$$

και 2 ακραίες (extremes, *) ως > 1.4394

$$Q3 + 3 * Qi = 1.0719 + 3 * 0.1225 = 1.4394 \text{ --> } X = 10^{1.4394} = 27.50.$$

Στο **θηκόγραμμα (boxplot)** διαπιστώνουμε ότι ο διάμεσος (μαύρη γραμμή) χωρίζει σχεδόν στη μέση τη θήκη, αλλά υπάρχουν 14 παράτυπες τιμές που έχουν ως

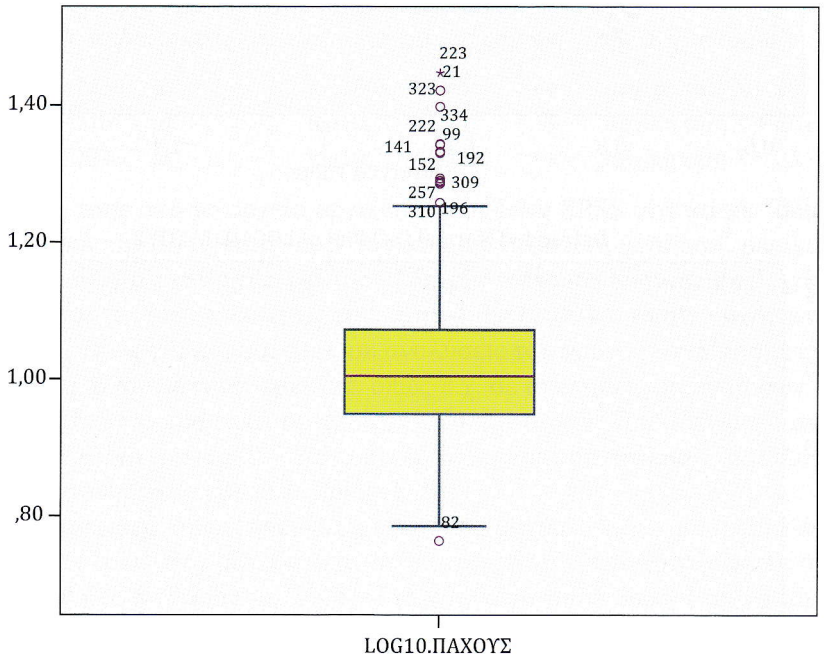
Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average (Definition 1)	LOG10. ΠΑΧΟΥΣ	,8482	,8808	,9494	1,0043	1,0719	1,1717	1,2279
Tukey's Hinges	LOG10. ΠΑΧΟΥΣ			,9494	1,0043	1,0719		

αποτέλεσμα η κατανομή να παραμένει ελαφρώς λοξή θετικά (slightly positively skewed).

Οι 14 αυτές περιπτώσεις (cases) είναι εμφανείς στο γράφημα ως απόμακρες (outliers, **o**) για τα άτομα (case)

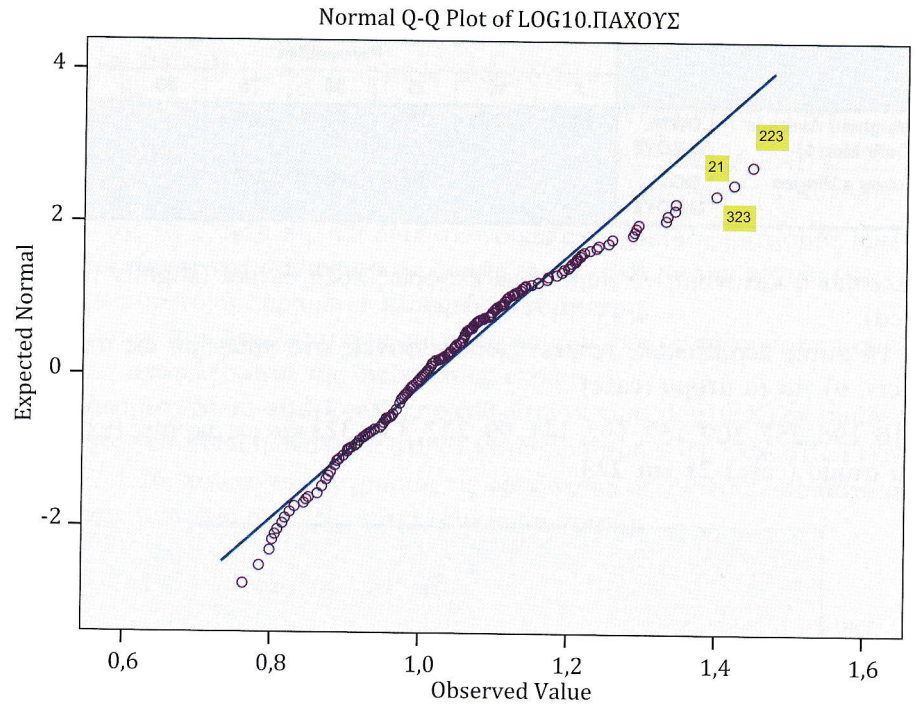
82, 310, 196, 257, 309, 152, 192, 141, 99, 222, 334, **323** και ως ακραίες (extreme, *) για τα άτομα (case) **21** και **223**.

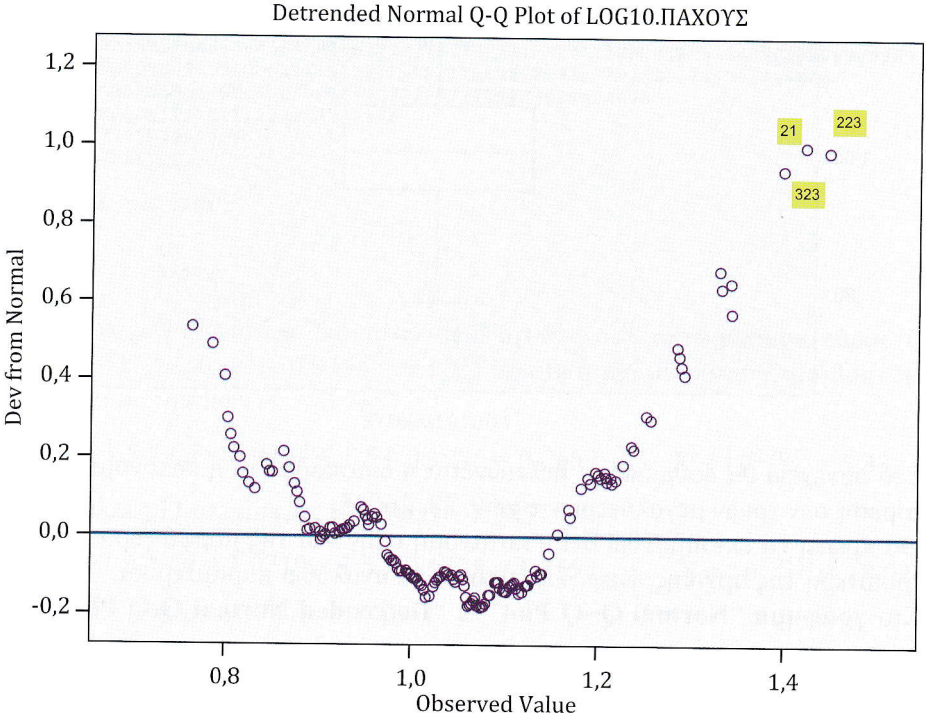


Στο συνέχεια θα δούμε πόσο βελτιώνεται η διασπορά και η κατανομή μετά την αφαίρεση των τριών μεγαλύτερων τιμών: 28, 26, 25.

Θα πρέπει να επισημανθεί ότι η κατανομή στην νέα της μορφή (δραστικά συμμετρικότερη της αρχικής, Κεφ. 7.2) μπορεί να αναλυθεί παραμετρικά.

Στο **γράφημα "Normal Q–Q Plot" & "Detrended Normal Q-Q Plot"** φαίνεται πιο καθαρά απόκλιση από την κανονικότητα σε 3 κυρίως περιπτώσεις (cases): 223 & 21 (ως ακραίες, *) και 323 (ως η μέγιστη απόμακρη, outlier). Αυτές αντιστοιχούν στις τιμές log10(X) = 1.45, 1.42, 1.40 και στις αρχικές τιμές X = 28, 26, 25, και ξεχωρίζουν σαφώς από τις άλλες 336 αρχικές τιμές.


Normal Q-Q Plot of LOG10.ΠΑΧΟΥΣ
Expected Normal
Observed Value
4
2
0
-2
0,6
0,8
1,0
1,2
1,4
1,6
223
21
323


Detrended Normal Q-Q Plot of LOG10.ΠΑΧΟΥΣ
Dev from Normal
Observed Value
1,2
1,0
0,8
0,6
0,4
0,2
0,0
-0,2
0,8
1,0
1,2
1,4
21
223
323

7.3.1 Η κατανομή των τιμών log10(X) χωρίς 3 ακραίες (N=336)

Στο κεφάλαιο 7.2 είδαμε ότι η κατανομή των 339 τιμών "Πάχους" έχει μεγάλη θετική ασυμμετρία (substantial positive skewness) και η ανάλυσή της παραμετρικά θα ήταν αδόκιμη.

Στο Κεφάλαιο 7.3 είδαμε ότι ο μετασχηματισμός των αρχικών τιμών X σε log10(X) έδωσε κατανομή σχεδόν συμμετρική, δηλαδή με μικρή θετική λοξότητα, που οφείλεται κυρίως στις 3 μεγαλύτερες τιμές 28, 26, 25, που συνιστούν μια μικρή συστάδα (cluster) ξέχωρη από τις άλλες 336 τιμές.

Όπως ήδη επισημάνθηκε, στην νέα αυτή μορφή της η κατανομή μπορεί να αναλυθεί παραμετρικά, καθότι η μικρή λοξότητα που εμφάνισε δεν αποτελεί πρόβλημα για τις περισσότερες παραμετρικές στατιστικές μεθόδους. Όμως, η ερμηνεία των στατιστικών αποτελεσμάτων απαιτεί το μετασχηματισμό των τιμών log10(X) στις αρχικές ως $X = 10^{LOG10(X)}$.

Στη συνέχεια θα δούμε αν η εν λόγω κατανομή μπορεί να βελτιωθεί μετά την **αφαίρεση των** 3 αυτών **πολύ ακραίων τιμών.**

Οι 3 αυτές περιπτώσεις αφορούν τις τιμές "πάχους" των εξής ατόμων:

άτομο 223ο	με	log10(X) = 1.45	→	αρχική τιμή X = 28,
άτομο 21ο	με	log10(X) = 1.42	→	αρχική τιμή X = 26 &
άτομο 323ο	με	log10(X) = 1.40	→	αρχική τιμή X = 25.

Η αφαίρεσή τους από το αρχείο αρχικών δεδομένων SPSS γίνεται με "delete".

Η πρακτική της αφαίρεσης, πριν την κύρια στατιστική ανάλυση, μερικών πολύ ακραίων τιμών, (influential cases) προκειμένου να έχουμε ουσιαστική βελτίωση της διασποράς και της κανονικότητας της κατανομής δεν είναι σπάνια στην επιστημονική έρευνα, όπως για παράδειγμα, σε αναλύσεις του χρόνου αντίδρασης (Ratcliff, 1993). Σε **μη πειραματικές έρευνες**, όπως π.χ. σε αναλύσεις οικονομικών και κοινωνικών δεικτών η πρακτική αυτή είναι λίγο πιο συχνή. Παράδειγμα η αφαίρεση τιμών με υπόλοιπο (residual) > 2.5 τυπικές αποκλίσεις σε ανάλυση της Ολυμπιακής επίδρασης στο εμπόριο (Rose & Spiegel, 2011).

Σε **πειραματικές**, όμως, **έρευνες** η αφαίρεση αρχικών τιμών μπορεί να έχει δραματικές επιπτώσεις στο μέγεθος και στην εγκυρότητα των στατιστικών αποτελεσμάτων (Bakker & Wicherts, 2014). Για τον λόγο αυτό θα πρέπει να γίνεται με φειδώ και μόνο μετά από συνεκτίμηση στοιχείων όπως το μέγεθος του δείγματος (N), η διασπορά, η θεωρητική βάση της έρευνας, η κύρια στατιστική ανάλυση, και το αν η ανάλυση είναι διερευνητική (exploratory) ή επιβεβαιωτική (confirmatory) κάποιας θεωρίας.

Οι 3 αυτές τιμές στην αρχική τους μορφή ήταν οι X = 28, 26, 25.
Η αφαίρεσή τους από το αρχείο δεδομένων SPSS έγινε ως εξής:

DATA >> SELECT CASES >> IF CONDITION IS SATISFIED: ΠΑΧΟΣ < 25.

Έτσι, φιλτραρίστηκαν απλά οι 3 αυτές τιμές χωρίς να αφαιρεθούν τελείως από το αρχείο και η ανάλυση "έτρεξε" χωρίς αυτές τις τιμές με N=336.

Από την Explore παρατίθεται ο πίνακας "Descriptives", το ιστόγραμμα (histogram), το φυλλόγραμμα (Stem-&-Leaf) και το θηκόγραμμα (boxplot).

Από τον πίνακα "Descriptives" διαπιστώνεται ότι τα μέτρα κεντρικής θέσης έχουν συγκλίνει μεταξύ τους (μέσος = 1.0144, μέσος "5% trimmed" = 1.0111, διάμεσος = 1.0043), πράγμα που δείχνει βελτίωση της συμμετρίας της κατανομής, όπως αυτό επιβεβαιώνεται και από τις τιμές λοξότητας (0.489) και κύρτωσης (0.334), αλλά και από τα πηλίκα τους με τα τυπικά τους σφάλματα (0.489 / 0.133 = 3.68, 0.334 / 0.265 = 1.26).

Descriptives

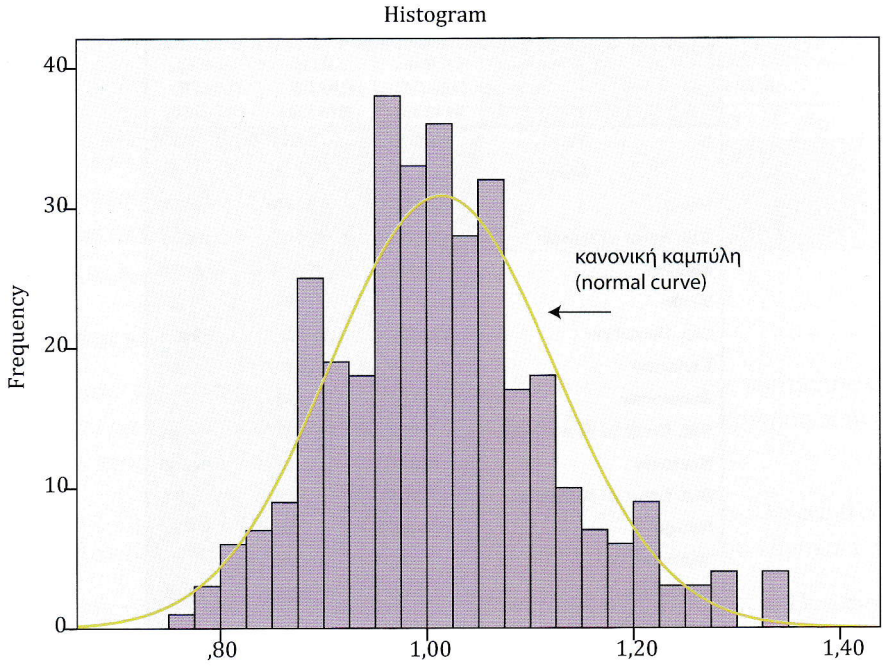
			Statistic	Std. Error
LOG10.ΠΑΧΟΣ. N336	Mean		1,0144	,00592
	95% Confidence Interval for Mean	Lower Bound	1,0027	
		Upper Bound	1,0260	
	5% Trimmed Mean		1,0111	
	Median		1,0043	
	Variance		,012	
	Std. Deviation		,10857	
	Minimum		,76	
	Maximum		1,34	
	Range		,58	
	Interquartile Range		,12	
	Skewness		,489	,133
	Kurtosis		,334	,265

Το μέτρο 95% Confidence Interval for Mean εξηγήθηκε στη σελ. 75 και 122.
Η κατανομή των 336 μετασχηματισμένων τιμών είναι εμφανώς συμμετρικότερη της αρχικής, αλλά και αυτής των 339 μετασχηματισμένων τιμών, που είχαν δείξει μια μικρή θετική λοξότητα, η οποία, όμως, έχει ελαττωθεί αισθητά στην παρούσα ανάλυση και φαίνεται από την καλή προσαρμογή του ιστογράμματος στην κανονική καμπύλη (γραμμή πάνω στο γράφημα).

Από το φυλλόγραμμα (Stem-&-Leaf) φαίνεται ότι η νέα κατανομή περιέχει 11 παράτυπες log10 τιμές: 1 ελάχιστη (≤ 0.76), 10 μέγιστες (≥ 1.26). Θα δούμε στη συνέχεια εντελώς "πειραματικά" ποια βελτίωση μπορεί να επιφέρει η ταυτόχρονη αφαίρεση και των 11 αυτών παράτυπων τιμών.

```
LOG10.ΠΑΧΟΣ (N=336, οι 3 μεγαλύτερες εκτός) Stem-and-Leaf Plot

Frequency
    Stem &  Leaf
 1 Extremes    (=<.76)
 3   7. 899
13   8. 0001112223444
34   8. 566666666777778888888888889999999
37   9. 000000011111111112222223333333444444
71   9. 555555555666666666666667777777777777777777888888888888888999999999999
64 10. 000000000000000001111111111111112222223333333334444444444444444
49 10. 555555555556666666666667777777777788888999999999
28 11. 00000001111111112222333333344
13 11. 5557777899999
12 12. 000001112233
 1 12. 5
10 Extremes    (>=1.26)
Stem width:      .10
Each leaf:       1 case(s)
```



7.3.2 Η κατανομή των τιμών log10(X) χωρίς 14 ακραίες (N=325)

Στην προηγούμενη ανάλυση (κεφ. 7.3.1) είδαμε ότι η κατανομή των 336 τιμών log10(X) (η αρχική χωρίς 3 ακραίες) είχε καλή προσέγγιση της κανονικότητας, όπως έδειξαν η λοξότητα, η κύρτωση και το ιστόγραμμα.

Για λόγους καθαρά επίδειξης της επίδρασης των πολύ ακραίων τιμών έγινε η ίδια ανάλυση, αλλά αυτή τη φορά αφού αφαιρέθηκαν και οι 11 ακραίες τιμές log10(X) που περιλαμβάνονται στο φυλλόγραμμα (Stem-&-Leaf).
Έτσι, αφαιρέθηκαν όλες (11) οι τιμές log10(X) που ήταν ≤ 0.76 ή ≥ 1.26.

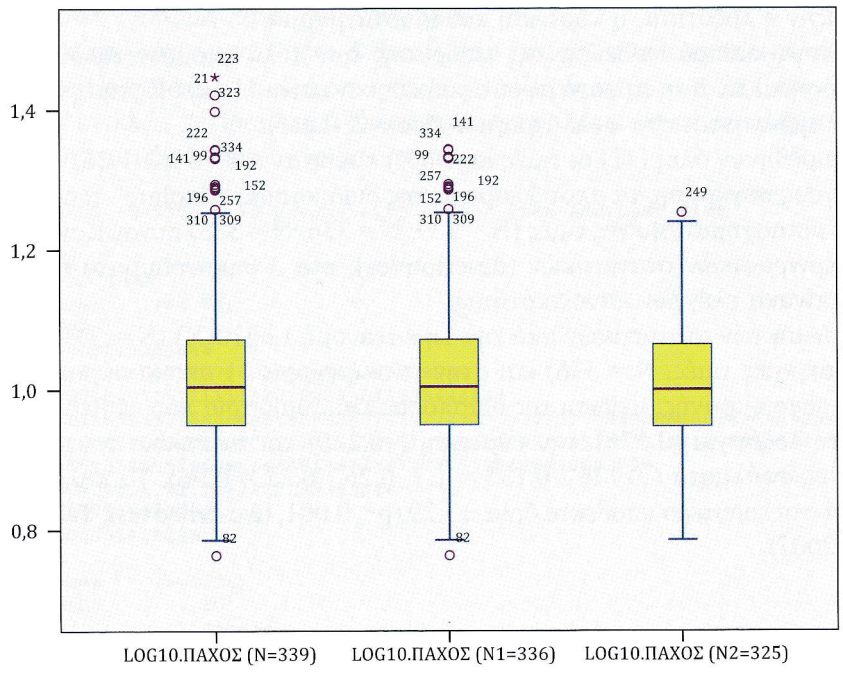
Για λόγους σύγκρισης τα αποτελέσματα της διαδικασίας "Explore" των 3 κατανομών με μετασχηματισμένες τιμές (N = 339, N = 336, N = 325) συνοψίζονται στον πίνακα περιγραφικών στατιστικών (descriptives), στα 3 θηκογράμματα (boxplot) και στον πίνακα ελέγχων κανονικότητας.

Η βελτίωση των στατιστικών από την 1η κατανομή Log10(X) (N = 339) στη 2η (χωρίς 3 ακραίες τιμές, N = 336) και στην τελική (χωρίς 11 ακόμα ακραίες τιμές, N = 325) είναι εμφανής: μείωση της διασποράς και συμμετρία που επιβεβαιώνεται και από τη λοξότητα (0.176), την κύρτωση (–0.260) και τα πηλίκα τους προς τα τυπικά τους σφάλματα (0.176 / 0.135 = 1.3, 0.26 / 0.27 = 0.96). Τα πηλίκα αυτά είναι μέσα στα ευρύτερα αποδεκτά όρια ±3.29 ($p < 0.001$, two-tailed test, Tabachnick & Fidell, 2007).

Statistics				
		LOG10. ΠΑΧΟΣ (N=339)	LOG10. ΠΑΧΟΣ (N1=336)	LOG10. ΠΑΧΟΣ (N2=325)
N	Valid	339	336	325
	Missing	0	3	14
Mean		1,0180	1,0144	1,0063
Std. Error of Mean		,00623	,00592	,00536
Median		1,0043	1,0043	1,0000
Mode		,97	,97	,97
Std. Deviation		,11467	,10857	,09666
Variance		,013	,012	,009
Skewness		,721	,489	,176
Std. Error of Skewness		,132	,133	,135
Kurtosis		1,019	,334	-,260
Std. Error of Kurtosis		,264	,265	,270
Range		,68	,58	,47
Minimum		,76	,76	,79
Maximum		1,45	1,34	1,25
Percentiles	25	,9494	,9494	,9432
	50	1,0043	1,0043	1,0000
	75	1,0719	1,0719	1,0654

Το μέτρο 95% Confidence Interval for Mean εξηγήθηκε στα κεφ. 4.6.2 & 7.1.2.

Η βελτίωση της συμμετρίας της κατανομής φαίνεται και από τους ελέγχους κανονικότητας, που στην 3η κατανομή δείχνουν μη σημαντική απόκλιση ($p = 0.086$ και $p = 0.045$, δηλαδή πιθανότητα σφάλματος τύπου I > 0.01).



Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
LOG10.ΠΑΧΟΣ (N=339)	,073	339	,000	,970	339	,000
LOG10.ΠΑΧΟΣ (N1=336)	,065	336	,002	,982	336	,000
LOG10.ΠΑΧΟΣ (N2=325)	,047	325	,086	,991	325	,045

a. Lilliefors Significance Correction

Προβλήματα και Ασκήσεις για Λύση στο SPSS

1. Εφάρμοσε τη διαδικασία ANALYZE >> Reports --> Case summaries --> Statistics για την αναφορά (report) και τον υπολογισμό των περιγραφικών στατιστικών των τιμών 3, 4, 7, 4, 4, 6, 8, 2, 1, 9.

2. Εφάρμοσε τη διαδικασία ANALYZE >> Descriptive Statistics --> Descriptives για τον υπολογισμό των περιγραφικών στατιστικών των 10 βαρών του Πίνακα 1.1.

3. Εφάρμοσε τη διαδικασία ANALYZE >> Descriptive Statistics --> Descriptives για τον υπολογισμό των περιγραφικών στατιστικών των 15 βαθμολογιών Μπάσκετ του Πίνακα 4.4.

4. Εφάρμοσε τη διαδικασία ANALYZE >> Descriptive Statistics --> Frequencies (Statistics, Charts) για τον υπολογισμό των περιγραφικών στατιστικών και του ιστογράμματος των τιμών 6, 1, 4, 2, 5, 2, 7, 4, 2, 3, 5, 5, 4, 3, 6, 3, 4, 4, 5, 7.

5. Εφάρμοσε τη διαδικασία ANALYZE >> Descriptive Statistics --> Frequencies (Statistics, Charts) για τον υπολογισμό των περιγραφικών στατιστικών και του ιστογράμματος των τιμών 1, 2, 2, 2, 3, 3, 3, 4, 4, 4, 4, 4, 5, 5, 5, 5, 6, 6, 7, 7.

6. Εφάρμοσε τη διαδικασία ANALYZE >> Explore (Statistics, Plots) για τον υπολογισμό των περιγραφικών στατιστικών (descriptive statistics) και την κατασκευή του φυλλογράμματος (stem-& leaf), του ιστογράμματος (histogram), των γραφημάτων και των ελέγχων κανονικότητας (normality plots) και του θηκογράμματος (box plot) των τιμών 6, 1, 4, 2, 5, 2, 7, 4, 2, 3, 5, 5, 4, 3, 6, 3, 4, 4, 5, 7, 6, 1, 4, 2, 5, 2, 7, 4, 2, 3, 5, 5, 4, 3, 6, 3, 4, 4, 5, 7, 6, 1, 4, 2, 5, 2, 7, 4, 2, 3, 5, 5, 4, 3, 6, 3, 4, 4, 5, 7.

7. Εφάρμοσε τη διαδικασία ANALYZE >> Explore (Statistics, Plots) για τον υπολογισμό των περιγραφικών στατιστικών (descriptive statistics) και την κατασκευή του φυλλογράμματος (stem-& leaf), του ιστογράμματος (histogram), των γραφημάτων και των ελέγχων κανονικότητας (normality plots) και του θηκογράμματος (box plot) των τιμών 1, 2, 2, 3, 4, 4, 4, 5, 5, 6, 6, 6, 8, 8, 9, 9, 9, 9, 9, 9, 9, 10, 12, 12, 12, 12, 12, 12, 12, 12, 12, 13, 14, 14, 14, 14, 14, 15, 15, 15, 15, 15, 15, 15, 16, 16, 16, 16, 16, 16, 17, 17, 17, 17, 17, 18, 18, 18, 18, 19, 19, 19, 20, 20, 20.

Χρήσιμες Επισημάνσεις για τις ακραίες τιμές (extreme values, outliers)

Η ύπαρξη ακραίων (πολύ μικρών ή πολύ μεγάλων) τιμών είναι σύνηθες φαινόμενο στα μη τυχαία δείγματα (non random samples). Άρα η εμφάνιση τους είναι πολύ συχνή στην εμπειρική έρευνα (empirical research), δηλαδή στην επιστημονική έρευνα που βασίζεται στην ανάλυση δείγματος. Στις τεχνολογικές επιστήμες, που στο δείγμα περιλαμβάνονται δεδομένα τεχνικής υφής, η εμφάνιση ακραίων τιμών μπορεί να είναι αποτέλεσμα δυσλειτουργίας (disfunction) ή θορύβου (noise) της μετρητικής διαδικασίας και αντιμετωπίζεται με φιλτράρισμα (filtering), προκειμένου να μειωθεί ή ακόμα και να απαλειφθεί η επίδρασή τους στα στατιστικά του δείγματος (π.χ.στην εμβιομηχανική της κίνησης, Winter, 2009, σελ.301-316). Σε άλλες, όμως, επιστήμες, όπως για παράδειγμα στην εκπαίδευση που στο δείγμα περιλαμβάνονται δεδομένα "ανθρώπινης συμπεριφοράς", η εμφάνιση ακραίων τιμών είναι συνήθως αποτέλεσμα ακραίων βιολογικών και ψυχολογικών λειτουργιών που δεν μπορούν (ούτε πρέπει) να φιλτραριστούν. Στις περιπτώσεις αυτές η αντιμετώπιση της ακραίας επίδρασής τους στα στατιστικά του δείγματος γίνεται συνήθως με τον κατάλληλο μετασχηματισμό των αρχικών τιμών (Bland & Altman, 1996). Εναλλακτική λύση στο πρόβλημα αυτό είναι η χρήση της κατάλληλης μη παραμετρικής μεθόδου ανάλυσης, όπως, για παράδειγμα ο έλεγχος Mann-Whitney αντί του κλασικού ελέγχου t-test για τη σύγκριση 2 ανεξάρτητων δειγμάτων (Bakker & Wicherts, 2014). Σε πιο σπάνιες περιπτώσεις και μετά από επαρκή μεθοδολογική αιτιολόγηση μερικές πολύ "δραστικές" ακραίες τιμές (highly influential cases) μπορεί να απαλειφθούν από το αρχικό δείγμα. Συστηματική έρευνα στατιστικής προσομείωσης (simulation study) δείχνει ότι η ακρίβεια και η ισχύς αναλύσεων όπως η συσχέτιση (correlation) και ο έλεγχος (t-test) βελτιώνονται σημαντικά και ουστατικά μετά την αφαίρεση κάποιων πολύ ακραίων τιμών (Osborne & Overbay, 2004). Τέλος η μελέτη ειδικά των ακραίων τιμών, που δεν είναι αποτέλεσμα σφάλματος του μηχανισμού που τις προκάλεσε, έχει και αυτή επιστημονική αξία, ακόμη και στο πλαίσιο της στατιστικής ερμηνείας των "θαυμάτων" στη ζωή (Kruskal, 1988, 929).