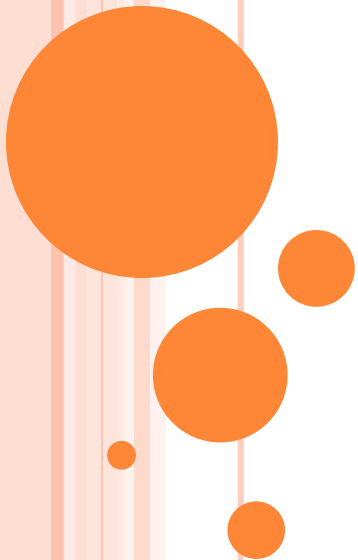


Πανεπιστήμιο Πειραιώς
Σχολή Χρηματοοικονομικής και Στατιστικής
Τμήμα Χρηματοοικονομικής και Τραπεζικής Διοικητικής

Πρόγραμμα Μεταπτυχιακών Σπουδών στη
Χρηματοοικονομική και Τραπεζική

ΠΡΟΠΑΡΑΣΚΕΥΑΣΤΙΚΟ ΜΑΘΗΜΑ ΣΤΗ ΣΤΑΤΙΣΤΙΚΗ

Μέρος 1ο



1.1 DATA SOURCES – DATA COLLECTION

ο Πρωτογενή Δεδομένα - Primary Data:

Collected from the original source

- Collected through:
 - ο Στατιστικά Πειράματα - Statistical experiments (controlled experiments)
 - ο Survey methods (μέθοδοι επισκόπησης)
 - ο Παρατήρηση - Observation (participant / non-participant where the observer is involved in the data generation)
- Ερωτηματολόγια - Questionnaires:
 - ο Μικρότερο κόστος
 - ο Πρέπει να είναι πολύ καλά σχεδιασμένα

1.1 DATA SOURCES – DATA COLLECTION (CONTINUED)

- **Δευτερογενή Δεδομένα - Secondary Data:**
Collected from other than the primary sources
 - Statistical Agencies
 - Research Institutions
 - Data providers (Bloomberg, Thomson/Reuters, etc.)

1.1 DATA SOURCES – DATA COLLECTION (CONTINUED)

- **Πληθυσμός - Population:** Every observation possibly available
- **Πλαίσιο Δειγματοληψίας - Sampling Frame:** Relevant information for all members of a population (παράδειγμα: τηλεφωνικός κατάλογος)
- **Απογραφή - Census:** Complete enumeration of a population's characteristics / data
 - Usually not cost effective

1.1 DATA SOURCES – DATA COLLECTION (CONTINUED)

- **Δείγμα - Sample:**

Selected in order to represent the population with respect to one or many characteristics of interest

- Cost effective

- **Απλό Τυχαίο Δείγμα - Simple Random Sample:**

Randomly choosing among the members of a population

- **Μη-Τυχαίο Δείγμα - Non-Random Sample:**

Sample that maintains certain characteristics of the population

- Stratified (Στρωματοποιημένο) Sample: Use of different strata (segments) of the population, maintaining their level of contribution to the total population
- **Cluster (συστάδα):** A specific group of a population that is totally representative with respect to (the value) one or many characteristics
- **Πολυσταδιακή Δειγματοληψία - Multistage Sampling:**
Selection of appropriate clusters at different stages of a study
 - In order to cut down sampling costs

1.2 PRESENTATION OF DATA

- Με πίνακα - Tabular form (table of data):
 - One way classification
 - In classes (κλάσεις) or ranges
 - Frequency (Συχνότητα) of a class: Number of observations falling into one class
 - Table of frequencies -> frequency distribution
 - Class characteristics:
 - mid-point (μέσο)
 - lower limit (κάτω όριο)
 - upper limit (άνω όριο)
 - Προσοχή: το άνω όριο μιας κλάσης μπορεί να είναι ίδιο με το κάτω όριο της επόμενης όταν έχουμε συνεχή δεδομένα (δηλαδή όταν τα δεδομένα μπορούν να έχουν οποιαδήποτε τιμή σε ένα διάστημα).
 - More classes -> more informative but more complicated
 - Less classes -> more simple but some information may be lost
 - Ο αριθμός των κλάσεων εξαρτάται από το μέγεθος του δείγματος
 - Τα διαστήματα που αντιστοιχούν σε κάθε κλάση μπορεί να μην έχουν ίδιο εύρος
 - Open ended intervals: when there is no upper or lower limit

1.2 PRESENTATION OF DATA (CONTINUED)

- Relative frequency (Σχετική Συχνότητα) of a class: The **percentage** of the observations that fall into one class
- Relative frequency distribution (κατανομή): A table with the relative frequencies of the classes.
- Cumulative frequency distribution (αθροιστική κατανομή συχνοτήτων): We extend the left limit of each class in order to be equal to the left limit of the first class. We then re-calculate the frequency distribution for the new classes.
 - Αντιπροσωπεύει τον **αριθμό** των παρατηρήσεων που έχουν τιμή ίση ή μικρότερη του άνω ορίου της κάθε κλάσης.
- Αθροιστική σχετική κατανομή συχνοτήτων - Cumulative relative frequency distribution:
 - Αντιπροσωπεύει το **ποσοστό** των παρατηρήσεων που έχουν τιμή ίση ή μικρότερη του άνω ορίου της κάθε κλάσης.
- **Attention:** Cumulative distributions require ordered data (διατεταγμένα δεδομένα).

1.2 PRESENTATION OF DATA (CONTINUED)

- Two way classification tables:
 - Multiple columns (one for each characteristic)
- When time is irrelevant or all data are sampled at the same time/ period, we have cross-sectional data (διαστρωματικά δεδομένα)
- When data represent evolution over a time horizon, we have time series data (χρονοσειρές)
- When we have cross sectional data that evolve over time, we have panel (two-dimensional) data

1.3 MEASURES OF CENTRAL TENDENCY

- Arithmetic Mean (“Αριθμητικός Μέσος” ή απλά “Μέσος”):

- Population (Πληθυσμιακός Μέσος)

$$\text{Ungrouped: } \mu = \frac{1}{N} \sum_{i=1}^N X_i \quad \text{Grouped: } \mu = \frac{\sum_{i=1}^N f_i X_i}{\sum_{i=1}^N f_i}$$

- Sample (Δειγματικός Μέσος)

$$\text{Ungrouped: } \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{Grouped: } \bar{X} = \frac{\sum_{i=1}^n f_i X_i}{\sum_{i=1}^n f_i}$$

- Arithmetic Mean is an “average” value, f_i : συχνότητες

1.3 MEASURES OF CENTRAL TENDENCY (CONTINUED)

- Median (Διάμεσος):
 - The value below which lie half of the rest observations, when the number of observations is odd, or
 - The average between the two middle values, if the number of observations is even.
 - It is not sensitive to outliers but it does not use much of the information
- Mode (Κορυφή): The observation, or the mid point of the class with higher frequency than the ones of the adjacent observations (or classes)
 - Attention: There might be many modes!!!
- Επικρατούσα τιμή: Η τιμή με τη μεγαλύτερη συχνότητα εμφάνισης (αν υπάρχει)

1.3 MEASURES OF CENTRAL TENDENCY (CONTINUED)

- Geometric Mean (Γεωμετρικός Μέσος): its logarithm is the mean of the logarithms of the observations.
 - very useful in financial data or when averaging percentages or rates

1.4 MEASURES OF DISPERSION

- Range (Εύρος): Higher minus Lower value
- Quantiles (Ποσοστημόρια):
 - Quartiles (τεταρτημόρια): separate sorted data into 4 segments (Q_1 , Q_2 , Q_3)
 - Deciles (δεκατημόρια): separate sorted data into 10 segments

1.4 MEASURES OF DISPERSION (CONTINUED)

- Interquartile range (Ενδοτεταρτημοριακό Εύρος):

$$IQR = Q_3 - Q_1$$

- Is a measure of dispersion around the median
 - Is not affected by extreme observations (outliers)
- Quartile deviation: average deviation of the quartiles from the median.

$$QD = IQR/2$$

- Coefficient of Quartile Deviation:

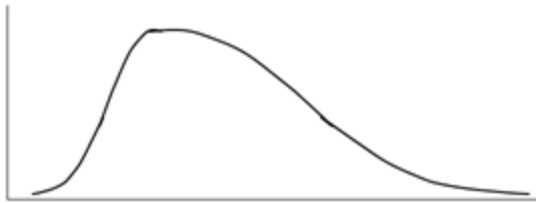
$$CQD = (Q_3 - Q_1)/(Q_3 + Q_1)$$

1.4 MEASURES OF DISPERSION (CONTINUED)

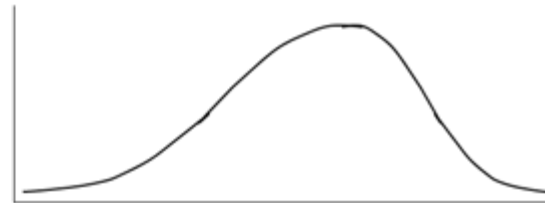
- Μέση (απόλυτη) απόκλιση - Average (absolute) deviation: Average distance of the observations from the mean
- Διασπορά - Variance: Average squared distance of the observations from the mean
 - More sensitive to observation which are far from the mean
- Standard deviation (s.d): the square root of variance
 - Maintains the measurement unit of the observation
- Coefficient of Variation: σ/μ , relative s.d.

1.5 SHAPE OF FREQUENCY DISTRIBUTIONS

- Coefficient of Skewness (Συντελεστής Ασυμμετρίας):
 - Μέτρο ασυμετρίας της κατανομής συχνοτήτων
 - Ευαίσθητο σε παρατηρήσεις που είναι σχετικά μακριά από το μέσο



Right-skewed (skew > 0)

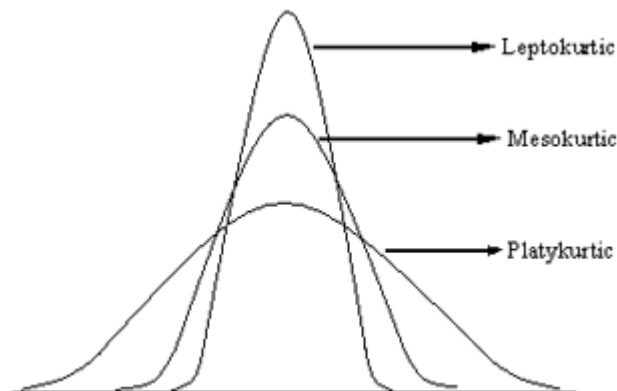


Left-skewed (skew < 0)

- Coefficient of Kurtosis (Συντελεστής Κύρτωσης):
 - Χαρακτηρίζει πόσο μυτερή είναι η κατανομή συχνοτήτων
 - Very sensitive to large deviations from the mean
 - Kurt <3 : Πλατύκυρτη (platykurtic distribution)
 - Kurt >3 : Λεπτόκυρτη (leptokurtic distribution)
 - Kurt $=3$: mesokurtic distribution (as is the normal distribution)

1.5 SHAPE OF FREQUENCY DISTRIBUTIONS (CONTINUED)

- Coefficient of Kurtosis (Συντελεστής Κύρτωσης):
 - Χαρακτηρίζει πόσο μυτερή είναι η κατανομή συχνοτήτων
 - Very sensitive to large deviations from the mean
 - $Kurt < 3$: Πλατύκυρτη (platykurtic distribution)
 - $Kurt > 3$: Λεπτόκυρτη (leptokurtic distribution)
 - $Kurt = 3$: mesokurtic distribution (as is the normal distribution)



ΣΥΜΒΟΛΙΣΜΟΙ

$X_i, i = 1, 2, 3, \dots, n$	Παρατηρήσεις
X_{max}	Μέγιστη παρατήρηση
X_{min}	Ελάχιστη παρατήρηση
n	Πλήθος των παρατηρήσεων
Δ	Εύρος της τάξης (αναφερόμαστε σε τάξεις ίσου εύρους)
m_i	Κεντρική τιμή της i τάξης
K	Πλήθος των τάξεων
f_M	Συχνότητα της τάξης της διαμέσου
f_{Q_i}	Συχνότητα της τάξης του i τεταρτημορίου
F_{M-1}	Αθροιστική συχνότητα της προηγούμενης τάξης από αυτή της διαμέσου
F_{Q_i-1}	Αθροιστική συχνότητα της προηγούμενης τάξης από αυτή του i τεταρτημορίου
L_M	Κατώτερο όριο της τάξης της διαμέσου
L_{Q_i}	Κατώτερο όριο της τάξης του i τεταρτημορίου
L_{T_e}	Κατώτερο όριο της τάξης της επικρατούσας τιμής
Δ_1	Διαφορά μεταξύ της συχνότητας της τάξης της επικρατούσας τιμής και της συχνότητας της προηγούμενης τάξης
Δ_2	Διαφορά μεταξύ της συχνότητας της τάξης της επικρατούσας τιμής και της συχνότητας της επόμενης τάξης

ΤΥΠΟΛΟΓΙΟ ΠΕΡΙΓΡΑΦΙΚΗΣ ΣΤΑΤΙΣΤΙΚΗΣ

Αριθμητικός Μέσος

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

$$\bar{X} = \frac{\sum_{i=1}^k f_i m_i}{\sum_{i=1}^k f_i}$$

Διάμεσος

- Αν n περιττός

M : Η τιμή της παρατήρησης στη

θέση $\frac{n+1}{2}$ δηλαδή η $\left(X_{\frac{n+1}{2}} \right)$

- Αν n άρτιος

$$M = \frac{1}{2} \left(X_{\frac{n}{2}+1} + X_{\frac{n}{2}} \right)$$

$$M = L_M + \delta \frac{\left(\frac{n}{2} - F_{M-1} \right)}{f_M}$$

Επικρατούσα Τιμή

T_0 : η τιμή με τη μεγαλύτερη
συχνότητα εμφάνισης

$$T_0 = L_{T_0} + \delta \frac{\Delta_1}{\Delta_1 + \Delta_2}$$

ΤΥΠΟΛΟΓΙΟ ΠΕΡΙΓΡΑΦΙΚΗΣ ΣΤΑΤΙΣΤΙΚΗΣ

Εύρος

$$R = X_{\max} - X_{\min}$$

Διακύμανση

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

$$S^2 = \frac{\sum_{i=1}^k f_i (m_i - \bar{X})^2}{\sum_{i=1}^k f_i - 1}$$

Τυπική Απόκλιση

$$S = +\sqrt{S^2}$$

$$S = +\sqrt{S^2}$$

Συντελεστές Ασυμμετρίας

$$\beta_3 = \frac{\frac{\sum_{i=1}^n (X_i - \bar{X})^3}{n}}{S^3}$$

$$\beta_3 = \frac{\frac{\sum_{i=1}^k f_i (m_i - \bar{X})^3}{\sum_{i=1}^k f_i}}{S^3}$$

Συντελεστής Κύρτωσης

$$\beta_4 = \frac{\frac{\sum_{i=1}^n (X_i - \bar{X})^4}{n}}{S^4}$$

$$\beta_4 = \frac{\frac{\sum_{i=1}^k f_i (m_i - \bar{X})^4}{\sum_{i=1}^k f_i}}{S^4}$$

ΑΣΚΗΣΗ 1

Δίνεται σε αύξουσα τάξης μεγέθους οι μετρήσεις του ύψους ενός δείγματος 34 φοιτητών:

1.55	1.59	1.60	1.60	1.61	1.62	1.63	1.63
1.63	1.64	1.65	1.66	1.67	1.67	1.67	1.69
1.70	1.70	1.70	1.70	1.71	1.72	1.75	1.75
1.78	1.78	1.78	1.81	1.82	1.85	1.85	1.90
1.92	1.98						

A μέρος

Ζητείται:

- α) Να ομαδοποιηθούν τα δεδομένα σε μια κατανομή συχνοτήτων με $k=5$ κλάσεις
- β) Με βάση το παραπάνω πίνακα συχνοτήτων να κατασκευάσετε το ιστόγραμμα του δείγματος σας
- γ) Να υπολογιστούν η διάμεσος και το ενδοτεταρτημοριακό εύρος από τα μη ταξινομημένα δεδομένα
- δ) Να υπολογιστούν ο αριθμητικός μέσος και η τυπική απόκλιση από τη κατανομή συχνοτήτων
- ε) Να υπολογιστούν ο συντελεστής κύρτωσης και ο συντελεστής ασυμμετρίας του δείγματος και να χαρακτηριστεί ανάλογα με τις τιμές που θα βρείτε

ΑΣΚΗΣΗ 1

Β μέρος

Παράλληλα με τις παραπάνω μετρήσεις ύψους για το συγκεκριμένο δείγμα φοιτητών έγινε συλλογή δεδομένων για το βάρος τους (οι τιμές δίνονται σε αντίστοιχη σειρά με τα ύψη):

52	67	75	58	95	56	88	65
52	81	76	73	86	67	60	75
79	69	76	79	66	61	61	87
89	59	74	74	82	65	96	94
101	113						

Με βάση και αυτές τις παρατηρήσεις:

α) Να κατασκευαστεί πίνακας συχνοτήτων των δύο μεταβλητών με 5 κλάσεις για το ύψος και 7 κλάσεις για το βάρος

β) Να κατασκευάσετε το scatter plot των μεταβλητών Ύψους – Βάρους. Τι μπορείτε να πείτε για τον τρόπο που συνδέονται τα δύο αυτά μεγέθη;

γ) Να υπολογίσετε το συντελεστή συσχέτισης (ρ) Ύψους – Βάρους.