Ζητήματα οργάνωσης & διαχείρισης δεδομένων από έρευνες με ερωτηματολόγιο

Διονύσιος Γ. Φραγκόπουλος,
Ηλεκτρολόγος Μηχανικός και Μηχανικός Η/Υ, PhD, MSc

Θεσσαλονίκη, Παρασκευή 8 Μαΐου 2020

Μεταπτυχιακό πρόγραμμα «Επικοινωνία»,
Τμήμα Δημοσιογραφίας & ΜΜΕ
Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

** επικοινωνία: dfragos@ee.auth.gr*

1

Στόχοι της παρουσίασης

Κατανόηση των:

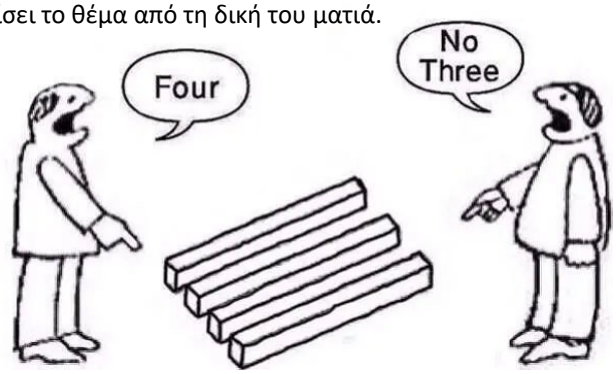
- **Δομή μίας έρευνας - το ερωτηματολόγιο ως εργαλείο μέτρησης**
  Γενικό πλαίσιο μίας έρευνας και χρήσης του ερωτηματολογίου ως εργαλείο μέτρησης με καταγραφή μεταβλητών και των σχέσεών τους
- **Παλινδρόμηση, συσχέτιση**
  Βασικές αρχές περιγραφής σχέσεων μεταβλητών και πρόβλεψης τιμών
- **Διαχείριση ελλειπουσών τιμών**
  Τύποι ελλειπουσών τιμών και τρόπων διαχείρισής τους με παράθεση των διαθέσιμων τεχνικών πρόβλεψης/απόδοσης αυτών των τιμών
- **Σχεδιασμός πολύπλοκης δομής έρευνας ερωτηματολογίου**
  Ταυτόχρονη εξέταση διαφορετικών αντικειμένων σε μία έρευνα.

2

2

Δομή μίας έρευνας
Στάση του ερευνητή

Ο ερευνητής οφείλει να είναι μεθοδικός, προσεκτικός και να αναφέρει πως αντιλαμβάνεται τον κόσμο για να είναι σε θέση η ερευνητική κοινότητα να προσεγγίσει το θέμα από τη δική του ματιά.



Everything we hear is an opinion, not a fact. Everything we see is a perspective, not the truth. *Marcus Aurelius**

Ο Marcus Aurelius Antoninus Augustus (121-180 μ.Χ.) υπήρξε στωικός φιλόσοφος ενώ η φράση αυτή ανήκει μάλλον στην φιλοσοφική οπτική του σκεπτικισμού

3

3

Δομή μίας έρευνας
Μεθοδολογία συνυφασμένη με το «παράδειγμα»

- **Ποιοτική έρευνα (qualitative paradigm)**
  - Συναντάται σε πολλές διαφορετικές μορφές κυρίως στις θεωρητικές επιστήμες ως ερμηνευτική (interpretivism), κονστρουκτιβισμός (constructivism), νατουραλισμός (naturalism) κ.α.
  - Υπάρχουν περισσότερες από μία πλευρές στην προσέγγιση της πραγματικότητας, ο ερευνητής μπορεί να είναι είτε μόνο παρατηρητής ή να συμμετέχει σε αυτό που καλείται να μελετήσει *(οντολογική και επιστημολογική διάσταση)*

- **Ποσοτική έρευνα (quantitative paradigm)**
  - Προέρχεται από τη φιλοσοφική στάση του θετικισμού
  - Υπάρχει μόνο μία αλήθεια για την πραγματικότητα, μια αντικειμενική πραγματικότητα *(οντολογική διάσταση)*
  - Ο ερευνητής είναι ανεξάρτητος από την πραγματικότητα που καλείται να μελετήσει, άρα δεν την επηρεάζει ούτε επηρεάζεται από αυτή *(επιστημολογική διάσταση)*

4

4

Παράδειγμα σχεδιασμού έρευνας ερωτηματολογίου
Σκοπός

- **Αφορά τα Ακαδημαϊκά και Ερευνητικά Ιδρύματα της Ελλάδας**
  - 36 Ακαδημαϊκά Ιδρύματα στη χώρα (2017)
  - 15 Ερευνητικοί φορείς
- **Εξέταση παραγόντων επιτυχίας σχεδιασμού και υλοποίησης πληροφοριακών συστημάτων**
  - Αφορά όσους είχαν συμμετάσχει στο σχεδιασμό και την υλοποίηση πληροφοριακών συστημάτων σε ρόλο Εποπτείας ή Διοίκησης έργου
- **Εξέταση αποδοχής χρήσης της τεχνολογίας**
  - Αποδοχή χρήσης διοικητικών πληροφοριακών συστημάτων
  - Αποδοχή χρήσης εκπαιδευτικών πληροφοριακών συστημάτων
- **Μέσο διεξαγωγής της έρευνας**
  - Αποστολή προσκλήσεων συμμετοχής, μέσω ηλεκτρονικού ταχυδρομείου, σε έρευνα με χρήση ηλεκτρονικής πλατφόρμας (limesurvey)

5

5

Παράδειγμα σχεδιασμού έρευνας ερωτηματολογίου
Το πρόβλημα

- **Πώς αποτυπώνεται η σημαντικότητα 205 παραγόντων;**
  - Οι παράγοντες πρέπει να αξιολογηθούν με κλίμακα σημαντικότητας πέντε σημείων
  - Πρέπει να γίνει διερευνητική ανάλυση παραγόντων
  - Το κοινό που απευθύνεται το ερωτηματολόγιο διαθέτει ελάχιστο ελεύθερο χρόνο και δεν πρέπει να υπάρξει σφάλμα μη συμμετοχής στην έρευνα
  - Το κοινό που πρέπει να απαντήσει στο ερωτηματολόγιο οφείλει να απαντήσει και το προκαταρκτικό μέρος της έρευνας (δημογραφικά και ειδικά χαρακτηριστικά), άρα σύνολο 205 + 8 ερωτήσεις (**μέσος εκτιμώμενος χρόνος συμπλήρωσης 30 λεπτά**).
  - Ακόμη και η ποιοτική ομαδοποίηση των αρχικών παραγόντων οδηγεί σε τελική λίστα 44, άρα 44 +8 ερωτήσεων **(μέσος εκτιμώμενος χρόνος συμπλήρωσης 7 λεπτά)**.
  - Η μερική αποτύπωση κάποιων μόνο παραγόντων (π.χ. 10) σε μία μόνο ομάδα δεν επιτρέπει την ενιαία ανάλυση και εξαγωγή συμπερασμάτων (π.χ. συσχετίσεις ή SEM)
- **Πώς αποτυπώνεται παράλληλα η αποδοχή χρήσης της τεχνολογίας;**
  - Πλήρη αποτύπωση των μη άμεσα μετρήσιμων υποκείμενων παραγόντων

6

6

Εργαλείο της έρευνας
Το ερωτηματολόγιο ως εργαλείο μέτρησης

- **Χαρακτηριστικά ερωτηματολογίων**
  - Ιδιοσυμπληρούμενο ή αυτό-συμπληρούμενο (self-administered)
  - Εμμέσως συμπληρούμενο (standardized interviews)
  - Πολλαπλοί τρόποι διάθεσης
    - Ταχυδρομική αποστολή (έντυπο)
    - Ηλεκτρονικό ταχυδρομείο
    - Με μορφή συνέντευξης (έμμεσα)
    - Μέσω ηλεκτρονικής πλατφόρμας

- **Ερωτήσεις (μεταβλητές)**
  - Ανοιχτού τύπου (open type questions)
    - Περισσότερος χρόνος συμπλήρωσης
    - Ελευθερία έκφρασης
    - Απαιτείται προσοχή για την ανάλυση/ερμηνεία
  - Κλειστού τύπου (close type questions)
    - Προκαθορισμένες απαντήσεις εντός πλαισίου
    - Κωδικοποιούνται σε μορφή κλίμακας (άρα αριθμητικά ερμηνεύσιμη)

7

7

Εργαλείο της έρευνας
Κοινοί τύποι λεκτικών και αριθμητικών κλιμάκων

- **Διχοτομικές κλίμακες**
  - Ερωτήσεις που καθοδηγούν σε απάντηση δύο διακριτών επιλογών (Ναι/Όχι)

  - Χρησιμοποιείτε κινητό τηλέφωνο; *Ναι* ☐ *Όχι* ☐

- **Κλίμακες απλής/πολλαπλής επιλογής**
  - Επιλογή μεταξύ διακριτών απαντήσεων (περισσότερων των δύο)

  - Επιλέξτε τις ηλεκτρονικές συσκευές που χρησιμοποιείτε:
    Κινητό τηλέφωνο ☐
    Φορητός υπολογιστής ☐

- **Κλίμακες (τύπου) Likert**
  - Διατύπωση μίας καταφατικής πρότασης και αποτύπωση βαθμού συμφωνίας του ερωτώμενου με αυτή ή πχ βαθμού σημαντικότητας

  - Πόσο συχνά χρησιμοποιείτε ηλεκτρονική συσκευή:
    Καθόλου ☐ *Λίγο* ☐ *Αρκετά* ☐ *Πολύ* ☐ *Πάρα πολύ* ☐

8

8

Άμεσα μετρήσιμες μεταβλητές
Τύποι μεταβλητών όπως προκύπτουν από λεκτικές ή αριθμητικές κλίμακες

- **Κατηγορικές ή διακριτές μεταβλητές (categorical)**
  - **Ονομαστικές ή διακριτές μεταβλητές (nominal)** είναι αυτές που διαθέτουν τις ιδιότητες αμοιβαίου αποκλεισμού και της πληρότητας περιεχομένου (π.χ. περιοχή διαμονής, φύλο)
  - **Τακτικές ή διατεταγμένες μεταβλητές (ordinal)** είναι αυτές που διαθέτουν τη βασική ιδιότητα της ταξινόμησης σε μία βαθμωτή κλίμακα (π.χ. σημαντικότητα παράγοντα, βαθμός συμφωνίας με διατύπωση).

- **Συνεχείς μεταβλητές (continuous)**
  - **Μεταβλητές διαστήματος (interval)** στις οποίες μας ενδιαφέρει η πραγματική απόσταση που χωρίζει τις τιμές που μετρούνται (π.χ. βαθμολογία σε μάθημα)
  - **Αναλογικές μεταβλητές (ratio)** στις οποίες μας ενδιαφέρει τόσο η πραγματική απόσταση που χωρίζει τις τιμές που μετρούνται αλλά διαθέτουν επίσης και ένα πραγματικό μηδενικό σημείο ως σημείο αναφοράς (π.χ. εισόδημα).

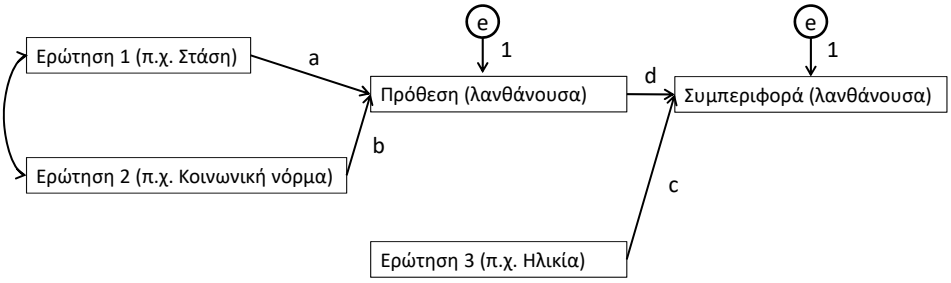
9

9

Μη άμεσα μετρήσιμες μεταβλητές και σχέσεις
Τύποι μεταβλητών

- **Παρατηρήσιμες μεταβλητές (observed variables) <u>ίσως</u> ως δείκτες**
  Κάθε ερώτηση του ερωτηματολογίου που αφορά μία μέτρηση και μπορούν να παρατηρηθούν άμεσα

- **Λανθάνουσες μεταβλητές (latent variables) ως δείκτες**
  Η αποτύπωση μη άμεσα μετρήσιμων διαστάσεων (αφηρημένων εννοιών ή υποκείμενων παραγόντων, π.χ. συμπεριφορά, προτίμηση, πρόθεση)

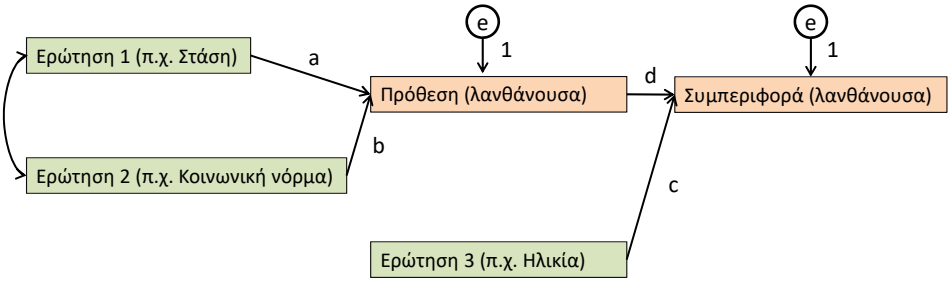


10

10

Μη άμεσα μετρήσιμες μεταβλητές και σχέσεις
Τύποι μεταβλητών

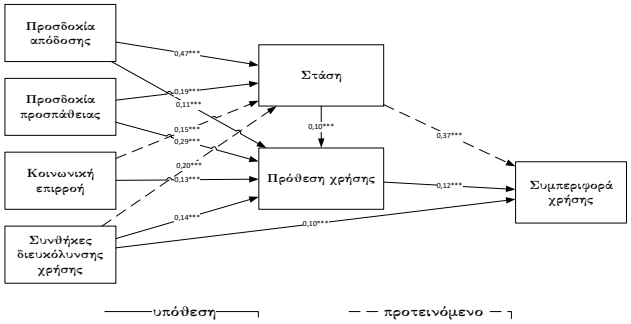
- **Ανεξάρτητες μεταβλητές (independent variables)**
  Οι άμεσα μετρήσιμες μεταβλητές που κατά την ανάλυση θεωρούνται δεδομένες (το αίτιο)
- **Εξαρτημένες μεταβλητές (dependent variables)**
  Οι μεταβλητές που εξαρτώνται ή προκύπτουν από άλλη/άλλες (το αποτέλεσμα)



11

11

Μη άμεσα μετρήσιμες μεταβλητές και σχέσεις
Παράδειγμα – Αποδοχή χρήσης της τεχνολογίας (UTAUT)



Προσδοκία απόδοσης: Αν θεωρεί ο χρήστης ότι η τεχνολογία βελτιώνει την απόδοσή του σε εργασίες
Προσδοκία προσπάθειας: Αν θεωρεί ο χρήστης ότι η τεχνολογία τον βοηθά να φέρει σε πέρας εργασίες με μικρότερη προσπάθεια
Κοινωνική επιρροή: Αν το κοινωνικό περιβάλλον του χρήστη έχει υιοθετήσει και πιθανώς επιδιώκει τη χρήση της τεχνολογίας
Συνθήκες διευκόλυνσης χρήσης: Αν ο χρήστης διαθέτει τις συνθήκες/εργαλεία που θα του επιτρέψουν τη χρήση της τεχνολογίας
Στάση: Η στάση του του χρήστη απέναντι στη χρήση της τεχνολογίας
Πρόθεση χρήσης: Η πρόθεση του χρήστη να χρησιμοποιήσει την τεχνολογία
Συμπεριφορά: Αν ο χρήστης χρησιμοποιεί την τεχνολογία

Πηγή: Dwivedi, Yogesh K., Nripendra P. Rana, Anand Jeyaraj, Marc Clement, and Michael D. Williams. 2017. "Re-examining the Unified Theory of Acceptance and Use of Technology (UTAUT): Towards a Revised Theoretical Model." Review of Information Systems Frontiers. doi: 10.1007/s10796-017-9774-y

12

12

Αξιοπιστία δεδομένων
Λανθάνουσες μεταβλητές - Εσωτερική συνάφεια ή ενδοσυνάφεια

- **Alpha (> 0,70 αποδεκτή, > 0,80 ικανοποιητική, 0,90 πάρα πολύ καλή)**
  - Από το 1951 που προτάθηκε ο συγκεκριμένος δείκτης από τον Cronbach χρησιμοποιείται τυφλά από την πλειοψηφία των ερευνητών.
- **Omega (> 0,70 αποδεκτή, > 0,80 ικανοποιητική, 0,90 πάρα πολύ καλή)**
  - Σε πρόσφατη βιβλιογραφία προτείνεται αντί του Alpha η χρήση του Omega, καθώς αποφεύγονται προβλήματα πόλωσης (bias) και άλλοι συγκεκριμένοι στατιστικοί περιορισμοί.
- **AVE (Average Variance Extracted) (> 0,50)**
  - Συντελεστής ένδειξης της αξιοπιστίας σύγκλισης (convergence validity)
  - Δείχνει κατά πόσο η διακύμανση των τιμών της λανθάνουσας μεταβλητής περιγράφεται ικανοποιητικά από τους δείκτες της σε σχέση με το σφάλμα μέτρησης
- **Φορτίσεις από την εφαρμογή επιβεβαιωτικής ανάλυσης**
  - Είναι η πιο ασφαλής τεχνική καθώς λαμβάνει υπόψη πολλές στατιστικές παραμέτρους.

13

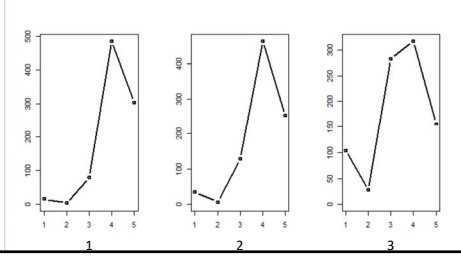
13

Αξιοπιστία δεδομένων
Λανθάνουσες μεταβλητές - Εσωτερική συνάφεια ή ενδοσυνάφεια

- **Στάση: Η στάση του του χρήστη απέναντι στη χρήση της τεχνολογίας (παράδειγμα)**

1. Η χρησιμοποίηση πληροφοριακών συστημάτων εκπαίδευσης είναι καλή πρακτική.
2. Τα πληροφοριακά συστήματα εκπαίδευσης κάνουν την διαδικασία της διδασκαλίας πιο ενδιαφέρουσα.
3. Προτιμώ να χρησιμοποιώ τα πληροφοριακά συστήματα εκπαίδευσης στην διαδικασία της διδασκαλίας από τις κλασικές πρακτικές.

$\text{Alpha} = 0{,}808, \quad \text{Omega} = 0{,}751, \quad \text{AVE} = 0{,}599$



14

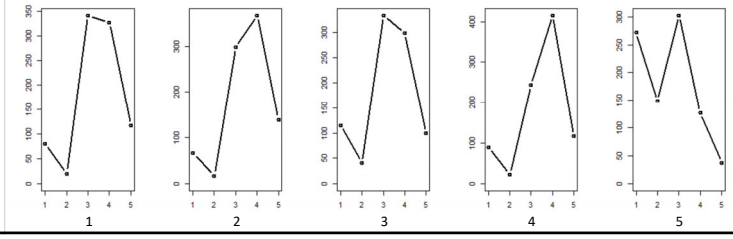
14

Αξιοπιστία δεδομένων
Λανθάνουσες μεταβλητές - Εσωτερική συνάφεια ή ενδοσυνάφεια

- **Κοινωνική επιρροή: Αν το κοινωνικό περιβάλλον του χρήστη έχει υιοθετήσει και πιθανώς επιδιώκει τη χρήση της τεχνολογίας (παράδειγμα)**

1. Τα άτομα που με γνωρίζουν καλά, θεωρούν ότι οφείλω να χρησιμοποιώ πληροφοριακά συστήματα εκπαίδευσης.
2. Τα άτομα των οποίων τη γνώμη εκτιμώ ιδιαίτερα, θεωρούν ότι πρέπει να χρησιμοποιώ πληροφοριακά συστήματα εκπαίδευσης.
3. Οι φοιτητές ζητούν από εμένα, στα μαθήματά μου, τη χρήση πληροφοριακών συστημάτων εκπαίδευσης.
4. Γενικά, το πανεπιστήμιό μου έχει υποστηρίξει ιδιαίτερα τη χρήση πληροφοριακών συστημάτων εκπαίδευσης.
5. Το πανεπιστήμιό μου με υποχρεώνει να χρησιμοποιώ πληροφοριακά συστήματα εκπαίδευσης.

$\text{Alpha} = 0{,}740, \quad \text{Omega} = 0{,}722, \quad \text{AVE} = 0{,}424 \quad | \quad \text{Alpha} = 0{,}795, \quad \text{Omega} = 0{,}795, \quad \text{AVE} = 0{,}662$



15

15

Σχέση μεταξύ μεταβλητών
Μοντέλα δομικών εξισώσεων (Structural Equation Modeling)

- **Μία ομάδα στατιστικών εργαλείων για την ανάλυση δεδομένων**
  Εξετάζει ταυτόχρονα σχέσεις μεταξύ εξαρτημένων ή ανεξάρτητων μεταβλητών ως επέκταση της παλινδρόμησης και της παραγοντικής ανάλυσης, με βασικές τεχνικές τις:

  - **Διερευνητική ανάλυση - Exploratory Factor Analysis – EFA**

  Ανάλυση παραγόντων (factor analysis) - Διερεύνηση των μετρήσεων ώστε να γίνει ομαδοποίηση των μετρήσιμων μεταβλητών, ώστε τελικά να εντοπιστούν πιθανές λανθάνουσες που μπορούν εξηγηθούν σύμφωνα με τα αποτελέσματα της ποιοτικής έρευνας

  - **Επιβεβαιωτική ανάλυση - Confirmatory Factor Analysis – CFA**

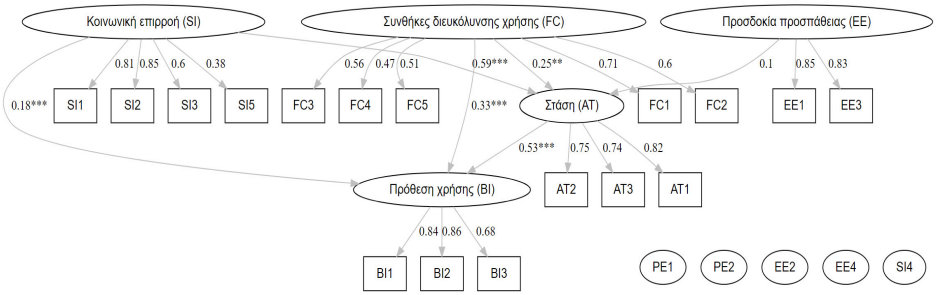
  Επιβεβαίωση ύπαρξης μεταβλητών και σχέσεων μεταξύ τους όπως αυτές περιγράφονται πιθανώς σε περιγραφικό μοντέλο προηγούμενης έρευνας/βιβλιογραφίας

16

16

Σχέση μεταξύ μεταβλητών
Μοντέλα δομικών εξισώσεων (Structural Equation Modeling)

- **Παράδειγμα επιβεβαιωτικής ανάλυσης - Αποδοχή χρήσης της τεχνολογίας (UTAUT)**



Διάγραμμα τυποποιημένων συντελεστών συσχέτισης (Beta)

Πηγή: Φραγκόπουλος Δ. 2018. "Εισαγωγή και ολοκλήρωση ICT συστημάτων στις εκπαιδευτικές, οργανωσιακές και επικοινωνιακές δομές της Ανώτατης εκπαίδευσης" Διδακτορική διατριβή, http://ikee.lib.auth.gr/record/303595?ln=el

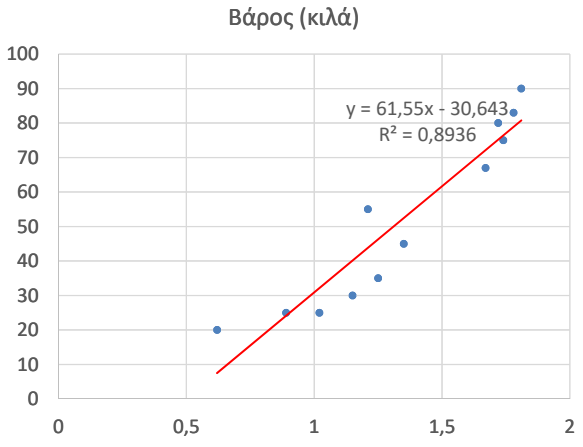
17

17

Περιγραφή σχέσης μεταβλητών
Συνεχείς μεταβλητές

- **Έστω ότι έχουμε αποτυπώσει τις τιμές του βάρους σε σχέση με το ύψος ενός ατόμου σε δύο ερωτήσεις του ερωτηματολογίου μας**

Υψος (μ)	Βάρος (κιλά)
1,72	80
1,78	83
1,67	67
1,81	90
1,74	75
1,21	55
1,35	45
1,25	35
1,15	30
1,02	25
0,89	25
0,62	20



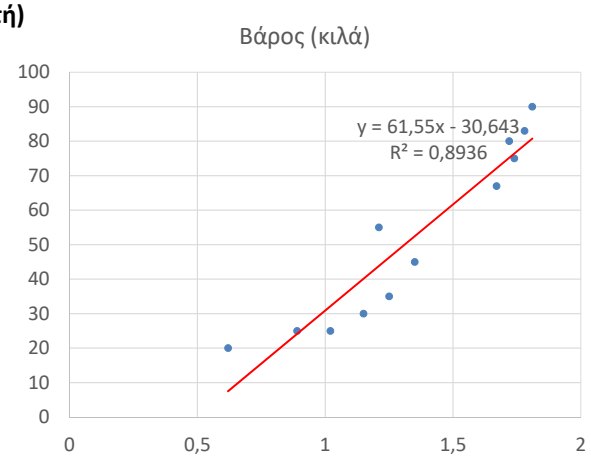
18

18

Πρόβλεψη και συνοχή τιμών
Συνεχείς μεταβλητές

- **Θέλουμε να προβλέψουμε τις πιθανές τιμές του βάρους (εξαρτημένη μεταβλητή) για ύψη ατόμων που δεν έχουμε λάβει απαντήσεις (ανεξάρτητη μεταβλητή)**

Υψος (μ)	Βάρος (κιλά)
1,72	80
1,78	83
1,67	67
1,81	90
1,74	75
1,21	55
1,35	45
1,25	35
1,15	30
1,02	25
0,89	25
0,62	20



19

19

Πρόβλεψη και συνοχή τιμών
Συνεχείς μεταβλητές - γραμμική παλινδρόμηση

- **Γραμμική παλινδρόμηση (linear regression)**
  Είναι μία προσέγγιση μοντελοποίησης της σχέσης μεταξύ μίας εξαρτημένης μεταβλητής Y με μία η περισσότερες ανεξάρτητες (x_1, x_2, ..) ή αλλιώς μια προσαρμογή γραμμής στη διασπορά των τιμών

  Y = b*x +a

  Στο παράδειγμά μας:

  Y= 61,55 * x – 30,643

  με R^2 = 0,8936 που σημαίνει ότι 89,36% των εξαρτημένων τιμών μπορούν να περιγραφούν από το μοντέλο μας (***προσαρμογή – fit***)

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-30,6435	9,418787	-3,25344	0,00867	-51,6299	-9,65712
Υψος (μ)	61,54978	6,715169	9,165783	3,51E-06	46,58745	76,5121

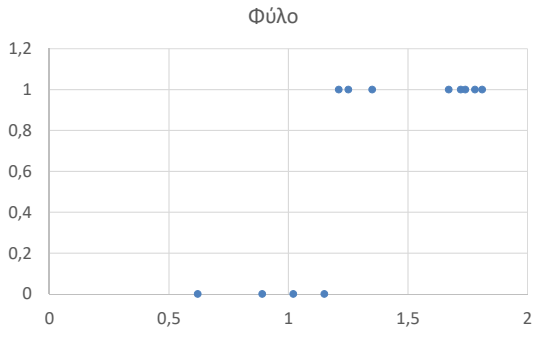
20

20

Πρόβλεψη και συνοχή τιμών
Διχοτομικές μεταβλητές

- **Γραμμική παλινδρόμηση (linear regression)**

Ύψος (μ)	Φύλο	Κωδ. Φύλου
1,72	Άντρας	1
1,78	Άντρας	1
1,67	Άντρας	1
1,81	Άντρας	1
1,74	Άντρας	1
1,21	Άντρας	1
1,35	Άντρας	1
1,25	Άντρας	1
1,15	Γυναίκα	0
1,02	Γυναίκα	0
0,89	Γυναίκα	0
0,62	Γυναίκα	0



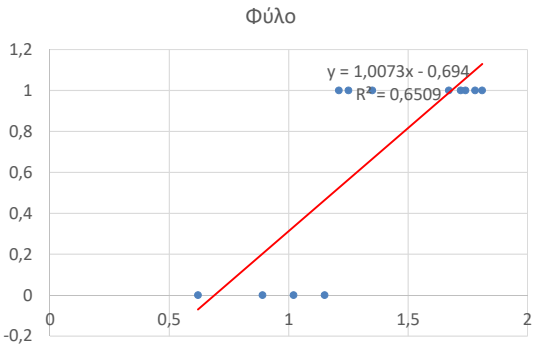
21

21

Πρόβλεψη και συνοχή τιμών
Διχοτομικές μεταβλητές

- **Γραμμική παλινδρόμηση (linear regression)** (για y>=0,5 -> 1, για y<0,5 ->0)
  Μπορεί να μας δώσει αρνητικές τιμές ή μεγαλύτερες του 1 που δεν έχουν κανένα νόημα στην κωδικοποίηση του φύλου.

Ύψος (μ)	Φύλο	Κωδ. Φύλου
1,72	Άντρας	1
1,78	Άντρας	1
1,67	Άντρας	1
1,81	Άντρας	1
1,74	Άντρας	1
1,21	Άντρας	1
1,35	Άντρας	1
1,25	Άντρας	1
1,15	Γυναίκα	0
1,02	Γυναίκα	0
0,89	Γυναίκα	0
0,62	Γυναίκα	0



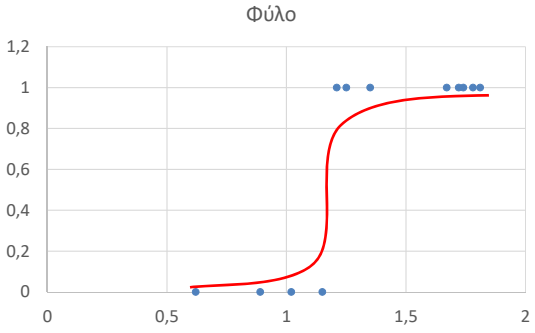
22

22

Πρόβλεψη και συνοχή τιμών
Διχοτομικές τιμές – δυαδική λογιστική παλινδρόμηση

- **Δυαδική λογιστική παλινδρόμηση (binomial logistic regression)**

Ύψος (μ)	Φύλο	Κωδ. Φύλου
1,72	Άντρας	1
1,78	Άντρας	1
1,67	Άντρας	1
1,81	Άντρας	1
1,74	Άντρας	1
1,21	Άντρας	1
1,35	Άντρας	1
1,25	Άντρας	1
1,15	Γυναίκα	0
1,02	Γυναίκα	0
0,89	Γυναίκα	0
0,62	Γυναίκα	0

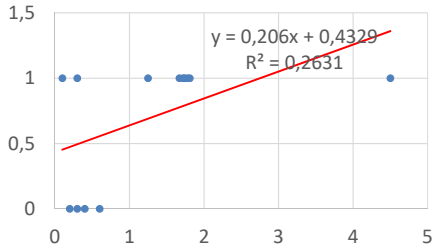


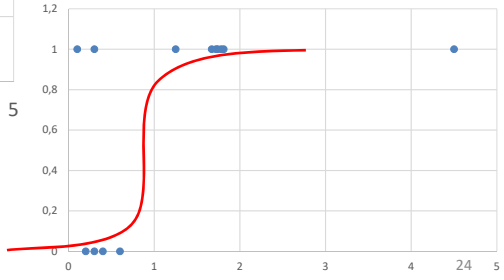
23

23

Πρόβλεψη και συνοχή τιμών
Διχοτομικές μεταβλητές – δυαδική λογιστική παλινδρόμηση

- **Δυαδική λογιστική παλινδρόμηση (binomial logistic regression)**
  - Η λογιστική παλινδρόμηση δεν αποτυγχάνει στον εντοπισμό της σωστής τιμής της μεταβλητής y πάνω από την οποία το φύλο είναι Άντρας.
  - Η λογιστική παλινδρόμηση είναι ιδανική για ταξινόμηση (classification), στην περίπτωσή μας μεταξύ Άντρα/Γυναίκας



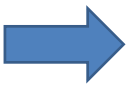


24

Πρόβλεψη και συνοχή τιμών
Τεχνική μετατροπής κατηγορικών μεταβλητών σε διχοτομικές

- **Χρήση ψευδομεταβλητών (dummy variables)**
  Οι κατηγορικές μεταβλητές μετατρέπονται σε διχοτομικές ώστε να μπορεί να εφαρμοστεί γραμμική ή δυαδική λογιστική παλινδρόμηση

Ύψος (μ)	Επάγγελμα
1,72	Φοιτητής
1,78	Καθηγητής
1,67	Τεχνικός
1,81	Τεχνικός
1,74	Καθηγητής
0,3	Καθηγητής
0,1	Φοιτητής
1,25	Καθηγητής
0,6	Τεχνικός



Ύψος (μ)	Επάγγελμα	Επάγγελμα Τεχνικός	Επάγγελμα Φοιτητής	Επάγγελμα Καθηγητής
1,72	Φοιτητής	0	1	0
1,78	Καθηγητής	0	0	1
1,67	Τεχνικός	1	0	0
1,81	Τεχνικός	1	0	0
1,74	Καθηγητής	0	0	1
0,3	Καθηγητής	0	0	1
0,1	Φοιτητής	0	1	0
1,25	Καθηγητής	0	0	1
0,6	Τεχνικός	1	0	0

25

25

Πρόβλεψη και συνοχή τιμών
Μη συνεχείς εξαρτημένες μεταβλητές – λογιστική παλινδρόμηση

- **Δυαδική λογιστική παλινδρόμηση (binomial logistic regression)**
  Χρησιμοποιείται όταν η εξαρτημένη μεταβλητή είναι δίτιμη ή δυαδική ή διχοτομική (binary) ή διμερής (π.χ. Ναι/Όχι)

- **Τακτική λογιστική παλινδρόμηση (ordinal logistic regression)**
  Χρησιμοποιείται όταν η εξαρτημένη μεταβλητή είναι τακτική (ordinal) (π.χ. μια ερώτηση της κλίμακας διαφωνώ καθόλου, λίγο, μέτρια, αρκετά, πολύ)

- **Ονομαστική λογιστική παλινδρόμηση**
  Χρησιμοποιείται όταν η εξαρτημένη μεταβλητή είναι ονομαστική (nominal) ή πολυωνυμική (polynomial) ή πολυχοτομική (polychotomus) ή κατηγορική αδιαβάθμητη (non-ordered categorical) ή πολυμερής μεταβλητή απόκρισης

26

26

Πρόβλεψη και συνοχή τιμών
Σύγκριση γραμμής με λογιστική παλινδρόμηση

Χαρακτηριστικό σύγκρισης	Γραμμική παλινδρόμηση	Λογιστική παλινδρόμηση
Γραμμική σχέση μεταξύ ανεξάρτητης και εξαρτημένης μεταβλητής	προαπαιτούμενο	αδιάφορο
Ανεξάρτητες μεταβλητές	Μπορούν να συσχετίζονται	Δεν πρέπει να υπάρχει συσχέτιση (**Μη ύπαρξη πολυσυγγραμμικότητας - multicollinearity**).
Εξαρτημένη μεταβλητή	Μόνο συνεχείς μεταβλητές	Μεταβλητές με διακριτά χαρακτηριστικά
Έννοια του συντελεστή παλινδρόμησης (Coefficient)	Αν όλες οι μεταβλητές παραμένουν σταθερές, η μοναδιαία αύξηση μίας μεταβλητής οδηγεί σε μοναδιαία αύξηση/μείωση της εξαρτημένης μεταβλητής κατά Χ.	Διαφορετική ανάλογα με τον τύπο λογιστικής παλινδρόμησης (δυαδική, κτλ.)
Τεχνική μείωσης σφάλματος	Χρήση μεθόδου ελαχίστων τετραγώνων (ordinary least squares)	Χρήση μέγιστης πιθανοφάνειας (maximum likelihood)

27

27

Σχέση μεταξύ μεταβλητών
Βασικές τεχνικές ελέγχου σχέσεων μεταξύ μεταβλητών

- **Συσχέτιση μεταβλητών (correlation coefficient)**
  Είναι δείκτης ελέγχου της σχέσης των τιμών δύο μεταβλητών (-1 < r < +1)

- **Συνδιακύμανση μεταβλητών (covariance coefficient)**
  Αν θέλουμε να έχουμε εικόνα του πόσο μεταβάλλεται μία μεταβλητή αν κάποια άλλη με την οποία σχετίζεται μεταβληθεί.

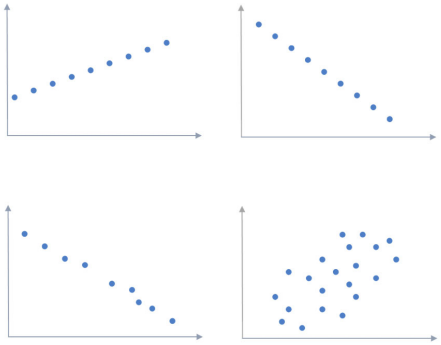
28

28

Σχέση μεταξύ μεταβλητών
Συσχέτιση μεταβλητών (correlation coefficient)

Μεταβλητή 1	Μεταβλητή 2	Μέθοδος
συνεχής	συνεχής	Pearson
συνεχής	τακτική	Polyserial
τακτική	τακτική	Polychoric
διχοτομική	διχοτομική	Tetrachoric
συνεχής ή τακτική	συνεχής ή τακτική	Spearman*
διχοτομική ή τακτική	διχοτομική ή τακτική	Kendal*

* Ως μη παραμετρικές μέθοδοι υπολογισμού



Γραμμική συσχέτιση

29

29

Σχέση μεταξύ μεταβλητών
Συσχέτιση μεταβλητών (correlation coefficient)

	Total	JABA	JEAB	TBA	BAP	TPR	AVB
Total	1						
JABA	.892[a]	1					
JEAB	.631[a]	.302[a]	1				
TBA	.716[a]	.428[a]	.592[a]	1			
BAP	.733[a]	.666[a]	.306[a]	.696[a]	1		
TPR	.492[a]	.210	.338[a]	.534[a]	.324[b]	1	
AVB	.459[a]	.239[b]	.360[a]	.410[a]	.193	.577[a]	1

[a] Correlation is significant at the 0.01 level (2-tailed)

[b] Correlation is significant at the 0.05 level (2-tailed)

** Ο πίνακας συσχέτισης βοηθά να εντοπιστεί το φαινόμενο της **πολυγραμμικότητας (multicollinearity)**, δηλαδή της υψηλής συσχέτισης (συνήθως >0,90) μεταξύ δύο μεταβλητών. Άλλος δείκτης του ίδιου φαινομένου είναι ο VIF (variance inflation factor) όταν είναι μεγαλύτερος του 10.*

Πηγή: https://www.researchgate.net/publication/275652409_Research_Rankings_of_Behavior_Analytic_Graduate_Training_Programs_and_Their_Faculty

30

30

Διαχείριση ελλειπουσών τιμών

- **Τρεις τύποι ελλειπουσών τιμών (missing values)**
  - Εντελώς τυχαία (Missing Completely at Random - MCAR)
  - Μερικώς τυχαία, λόγω κάποιας συνθήκης γνωστής από πριν (Missing At Random - MAR)
  - Με μη τυχαίο τρόπο, λόγω δομικού προβλήματος του ερωτηματολογίου (Missing Not At Random – MNAR)

Ύψος (μ)	Φύλο
1,72	**N/A**
1,78	Γυναίκα
1,67	Γυναίκα
1,81	**N/A**
1,74	**N/A**
0,3	Άντρας
0,1	Άντρας
1,25	Άντρας
0,6	**N/A**

Οφείλουμε να ελέγχουμε ειδικά την περίπτωση της εμφάνισης ελλειπουσών τιμών λόγω δομικού προβλήματος γιατί τότε υπάρχει πόλωση (bias) που πρέπει να λάβουμε υπόψη μας.

Αυτό που επιδιώκουμε είναι να διαθέτουμε μόνο εντελώς τυχαία δημιουργούμενες ελλείπουσες τιμές

Little, R.J.A., Rubin, D.B., (1987) Statistical Analysis with Missing Data. Wiley

31

31

Διαχείριση ελλειπουσών τιμών

- **Ελλείπουσες τιμές με μερικώς τυχαίο τρόπο (Missing At Random - MAR)**
  Η μεταβλητή στην οποία οφείλονται οι ελλείπουσες τιμές μας είναι γνωστή εκ των προτέρων

Ύψος (μ)	Φύλο
1,72	**N/A**
1,78	**N/A**
1,70	**N/A**
1,81	**N/A**
1,69	Γυναίκα
0,3	Άντρας
0,1	Γυναίκα
1,25	Γυναίκα
0,6	Γυναίκα

Έστω ότι έχουμε βάλει μία ερώτηση ελέγχου κατά την οποία δεν καταγράφεται το Φύλο για ύψος μεγαλύτερο από 1,70 μέτρα

Άλλο παράδειγμα μπορεί να είναι ότι ξέρουμε ότι δε θα λάβουμε απαντήσεις από συγκεκριμένες γεωγραφικές περιοχές για μία ερώτησή μας γιατί υπάρχει προκατάληψη για αυτό το θέμα σε αυτές τις συγκεκριμένες περιοχές

32

32

Διαχείριση ελλειπουσών τιμών

- **Ελλείπουσες τιμές με μη τυχαίο τρόπο, λόγω δομικού προβλήματος (Missing Not At Random – MNAR)**

Ύψος (μ)	Φύλο
1,72	**N/A**
1,78	**N/A**
1,70	**N/A**
1,81	**N/A**
1,69	Γυναίκα
0,3	Γυναίκα
0,1	Γυναίκα
1,25	Γυναίκα
0,6	Γυναίκα

Έστω ότι έχουμε συμπεριλάβει μία ερώτηση από σφάλμα για την αποτύπωση του φύλου που η διατύπωσή της προσβάλει τους Άντρες, οπότε δεν έχουν συμπληρώσει το φύλο τους

33

33

Διαχείριση ελλειπουσών τιμών

- **Παλιά (“old in Texas”) αντιμετώπιση ελλειπουσών τιμών**
  - Διαγράφονται όλες οι εγγραφές στις οποίες έστω μία τιμή ερώτησης από όσες συνολικά έχει απαντήσει ο συμμετέχων είναι ελλείπουσα (διαγραφή κατά λίστα - list-wise deletion). Πρέπει να έχουμε MCAR αλλιώς η διαγραφή θα δημιουργήσει πόλωση (bias).

23	21	22	10	-	24	56	100

- **Μερική αντιμετώπιση ελλειπουσών τιμών**
  - Υπολογίζονται οι στατιστικοί δείκτες σύμφωνα με τα ζεύγη των υπαρχουσών ζευγών τιμών, ουσιαστικά παραβλέποντας τις ελλείπουσες τιμές (Διαγραφή κατά ζεύγη - pairwise deletion). Μπορεί να δημιουργήσει, ως πρόβλημα, μη θετικούς πίνακες συσχέτισης (not positive definite).

23	21	22	10	-	24	56	100

- **Παραγωγή πλήρους δείγματος (συνόλου τιμών)**
  - Δημιουργούνται δείγματα με τα στατιστικά χαρακτηριστικά του αρχικού ελλιπούς δείγματος ή αλλιώς την ομοιογένεια του (congeniality), θέτοντας στη θέση των ελλειπουσών τιμών κάποιες που εμφανίζονται στατιστικά πιο πιθανό να είναι αληθείς (π.χ. μέθοδος παλινδρόμησης, μέθοδος EM, απόδοση πολλαπλών ελλιπών στοιχείων - multiple imputation - MI)

Marsh, H. W. (1998). Pairwise deletion for missing data in structural equation models: Nonpositive definite matrices, parameter estimates, goodness of fit, and adjusted sample sizes. Structural Equation Modeling: A Multidisciplinary Journal, 5, 22-36.
Buuren, S.V. and Groothuis-Oudshoorn, K., 2010. mice: Multivariate imputation by chained equations in R. Journal of statistical software, pp.1-68.

34

Διαχείριση ελλειπουσών τιμών
Παραγωγή πλήρους δείγματος

- **Τεχνική αντικατάστασης με μέσο όρο, διάμεσο, επικρατούσα τιμή**
  - Αντικαθίσταται κάθε ελλείπουσα τιμή με την τιμή του μέσου όρου όλων των τιμών της μεταβλητής (ή το διάμεσο, ή την επικρατούσα τιμή).
  - Συνήθως ενισχύεται το τυπικό σφάλμα και δε λαμβάνονται υπόψη όλα τα στατιστικά χαρακτηριστικά του δείγματος/δεδομένων
- **Τεχνική με χρήση ΕΜ ή παλινδρόμηση**
  - Υπολογίζεται η πιο πιθανή τιμή της ελλείπουσας λαμβάνοντας υπόψη τις υπόλοιπες διαθέσιμες τιμές της συγκεκριμένης μεταβλητής για άλλους συμμετέχοντες σε συνάρτηση με τις τιμές των υπόλοιπων μεταβλητών **(predictors)** που δε διαθέτουν ελλείπουσες τιμές.
  - Στις συγκεκριμένες τεχνικές δε λαμβάνεται υπόψη η συνολική αβεβαιότητα στην επιλογή της πιθανής μεταβλητής δημιουργώντας πόλωση (bias)
- **Τεχνική της πολλαπλής απόδοσης συνόλων, MCMC και mice**
  - Γίνεται υπολογισμός πολλαπλών πιθανών τιμών ως μέρος μίας διαδικασίας κατά την οποία συγκεκριμένα χαρακτηριστικά της προηγούμενης απόδοσης ενός πλήρους συνόλου λαμβάνονται υπόψη στον υπολογισμό του επόμενου με σκοπό το μικρότερο τυπικό σφάλμα στο πλαίσιο της εγγενούς αβεβαιότητας του δείγματος/δεδομένων.

Η τεχνική FIML (Full information maximum likelihood), διαθέσιμη στο AMOS και την R, δεν παράγει νέες τιμές αλλά επιτρέπει τον υπολογισμό π.χ. μοντέλων δομικών εξισώσεων με μεγαλύτερη ακρίβεια 35

35

Διαχείριση ελλειπουσών τιμών
Παραγωγή πλήρους δείγματος – Πολλαπλή απόδοση τιμών

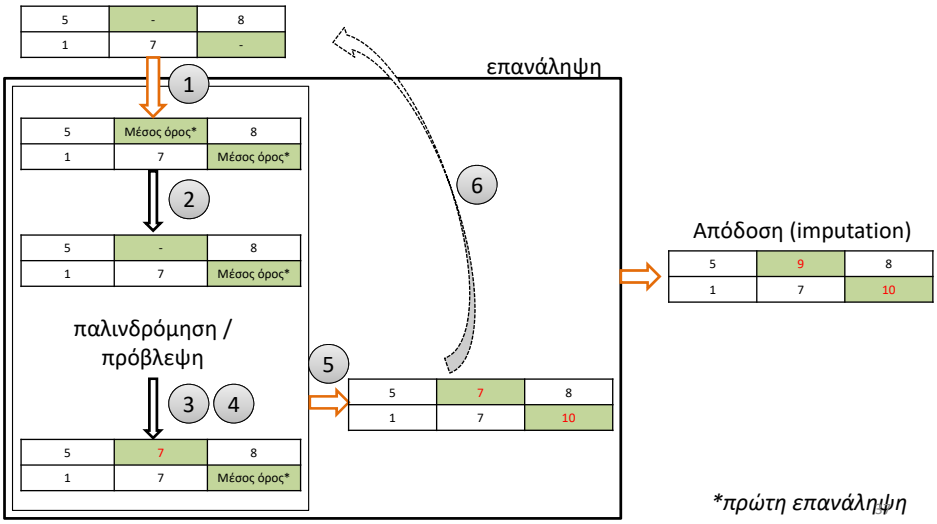
- **Βήματα υπολογισμού πολλαπλών συνόλων**
  1. Αντικαθίσταται κάθε ελλείπουσα τιμή με την τιμή του π.χ. του μέσου όρου όλων των τιμών της κάθε μεταβλητής ώστε να δημιουργηθεί ένα σύνολο αναφοράς
  2. Μία ελλείπουσα τιμή εμφανίζεται πάλι στο σύνολο αναφοράς ακριβώς όπως υπήρχε στο αρχικό ελλιπές σύνολο
  3. Οι διαθέσιμες πλέον τιμές, δηλαδή τόσο αυτές που είχαν μετρηθεί αλλά και αυτές που αντικαταστάθηκαν με το μέσο όρο της κάθε μεταβλητής χρησιμοποιούνται ώστε να προβλεφθεί η πιο πιθανή τιμή της ελλείπουσας (με χρήση γραμμικής ή λογιστικής παλινδρόμησης ανάλογα με τον τύπο της μεταβλητής που αντιστοιχεί). Κάθε ανεξάρτητη (ή και η εξαρτημένη) μεταβλητή που χρησιμοποιείται για την πρόβλεψη ονομάζεται **μεταβλητή πρόβλεψης (predictor)**
  4. Η τιμή που προβλέφθηκε στο προηγούμενο βήμα τοποθετείται στη θέση της ελλείπουσας τιμής και χρησιμοποιείται όπως και όλες οι υπόλοιπες τιμές του συνόλου για την πρόβλεψη των υπόλοιπων αρχικά ελλειπουσών τιμών.
  5. Τα βήματα 2-4 επαναλαμβάνονται για όλες τις αρχικά ελλείπουσες τιμές μέχρι να δημιουργηθούν πιθανές τιμές για όλες
  6. Τα βήματα 2-5 περιγράφουν μία πλήρη πρόβλεψη και επαναλαμβάνονται ως **επανάληψη (iteration)** ώστε σε κάθε νέα επανάληψη να δημιουργηθεί ένα νέο σύνολο λαμβάνοντας υπόψη τις τιμές που προβλέφθηκαν στην προηγούμενη.

36

36

Διαχείριση ελλειπουσών τιμών
Παραγωγή πλήρους δείγματος – Πολλαπλή απόδοση τιμών

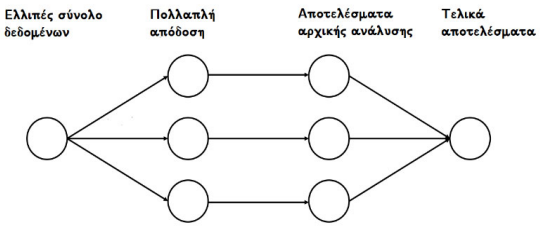
- **Βήματα υπολογισμού πολλαπλών συνόλων**



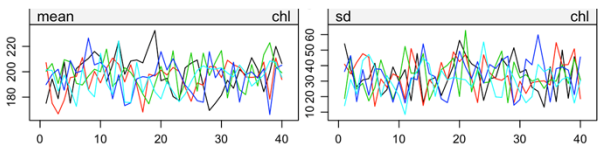
37

Διαχείριση ελλειπουσών τιμών
Παραγωγή πλήρους δείγματος – Πολλαπλή απόδοση τιμών

- **Λογικό διάγραμμα διαδικασίας**



- **Σύγκλιση τιμών**
  - Ελέγχεται η σύγκλιση του μέσου όρου (mean) και της τυπικής απόκλισης (standard deviation) ώστε να μην εμφανίζονται ακραίες τιμές



38

38

Παράδειγμα σχεδιασμού έρευνας ερωτηματολογίου
Το πρόβλημα

- **Πώς αποτυπώνεται η σημαντικότητα 205 παραγόντων;**
  - Οι παράγοντες πρέπει να αξιολογηθούν με κλίμακα σημαντικότητας πέντε σημείων
  - Πρέπει να γίνει διερευνητική ανάλυση παραγόντων
  - Το κοινό που απευθύνεται το ερωτηματολόγιο διαθέτει ελάχιστο ελεύθερο χρόνο και δεν πρέπει να υπάρξει σφάλμα μη συμμετοχής στην έρευνα
  - Το κοινό που πρέπει να απαντήσει στο ερωτηματολόγιο οφείλει να απαντήσει και το προκαταρκτικό μέρος της έρευνας (δημογραφικά και ειδικά χαρακτηριστικά), άρα σύνολο 205 + 8 ερωτήσεις (**μέσος εκτιμώμενος χρόνος συμπλήρωσης 30 λεπτά**).
  - Ακόμη και η ποιοτική ομαδοποίηση των αρχικών παραγόντων οδηγεί σε τελική λίστα 44, άρα 44 +8 ερωτήσεων **(μέσος εκτιμώμενος χρόνος συμπλήρωσης 7 λεπτά)**.
  - Η μερική αποτύπωση κάποιων μόνο παραγόντων (π.χ. 10) σε μία μόνο ομάδα δεν επιτρέπει την ενιαία ανάλυση και εξαγωγή συμπερασμάτων (π.χ. συσχετίσεις ή SEM)

- **Πώς αποτυπώνεται παράλληλα η αποδοχή χρήσης της τεχνολογίας;**
  - Πλήρη αποτύπωση των μη άμεσα μετρήσιμων υποκείμενων παραγόντων

39

39

Η λύση στο «άλυτο» πρόβλημα

- **Η δημιουργία δομικών ελλειπουσών τιμών με ελεγχόμενο τρόπο**
  - Η συμπλήρωση μίας ομάδας παραγόντων με σχεδιασμένο και δομημένο τρόπο ώστε να αποτυπώνονται σωστά τα στατιστικά χαρακτηριστικά του δείγματος

- **Σχεδιασμός Δομικών Ελλιπών Στοιχείων με χρήση τριπλής φόρμας**

Δημιουργία τριών ομάδων ερωτήσεων και προβολή των δύο από αυτές μόνο σε κάθε συμμετέχοντα, μαζί με την κοινή Χ ομάδα (Δημογραφικά και Ειδικά χαρακτηριστικά) σε τρεις φόρμες

φόρμα συμπλήρωσης	Ομάδες ερωτήσεων			
	X	A	B	C
1	1	1	1	0
2	1	1	0	1
3	1	0	1	1

Αποτέλεσμα 33% κενά δεδομένα (ελλείπουσες τιμές) σε κάθε ερώτηση για κάθε συμμετέχοντα, αλλά εξέταση περισσότερων ερωτήσεων

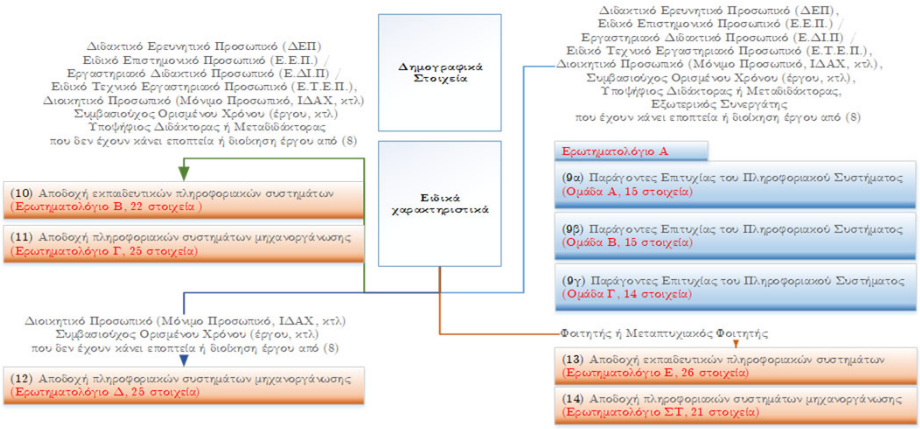
συνδυασμός	διάταξη ομάδων
1	XAB
2	XCA
3	XBC
4	AXB
5	CXA
6	BXC

Πηγή: Graham, J. W., B. J. Taylor, A. E. Olchowski, and P. E. Cumsille. 2006. "Planned missing data designs in psychological research." Review of. Psychol Methods 11 (4):323-43. doi: 10.1037/1082-989x.11.4.323.

40

40

Τελική δομή της έρευνας ερωτηματολογίου



Πηγή: Φραγκόπουλος Δ., 2018, "Εισαγωγή και ολοκλήρωση ICT συστημάτων στις εκπαιδευτικές, οργανωσιακές και επικοινωνιακές δομές της Ανώτατης εκπαίδευσης" Διδακτορική διατριβή, http://ikee.lib.auth.gr/record/303595?ln=el

41

41

Διεξαγωγή της έρευνας του παραδείγματος

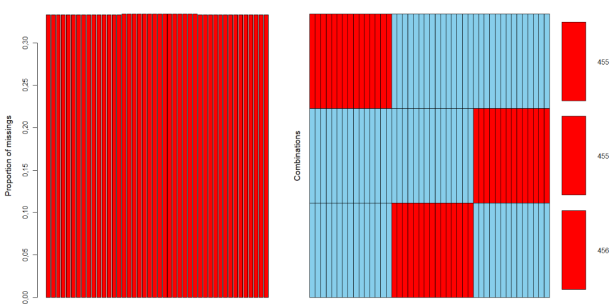
- **36 Ακαδημαϊκά Ιδρύματα στην Ελλάδα (23 ΑΕΙ και 13 ΤΕΙ) και 15 ερευνητικοί φορείς**
- **Τεχνική συλλογής δεδομένων (Φεβρουάριος - Απρίλιος 2018):**
  - 71.151 **(προσωπικές)** ηλεκτρονικές προσκλήσεις συμμετοχής σε έρευνα ερωτηματολογίου από τις οποίες εκτιμάται ότι παραδόθηκαν επιτυχώς περίπου 20.000, σε όλες τις ομάδες χρηστών των ιδρυμάτων και φορέων
  - 8.838 συνδέσεις στην πλατφόρμα, 8.054 μοναδικές επισκέψεις.
  - επισκέψεις έγιναν από 48 διαφορετικές χώρες
  - 48 διαφορετικές περιοχές της Ελλάδας και 138 περιοχές άλλων 47 χωρών
  - 79,41% επέλεξαν σταθερό υπολογιστή ή laptop για τη διαδικασία της συμπλήρωσης, 18,30% κινητό τηλέφωνο και 1,83% tablet
  - 5.922 πλήρεις απαντήσεις συνολικά

42

42

Η λύση στο «άλυτο» Πρόβλημα
Το δεύτερο πρόβλημα

- **Μοτίβο ελλειπουσών τιμών**
  - Το ερωτηματολόγιο Α συμπληρώθηκε από 1369 άτομα
  - 44 παράγοντες επιτυχίας σε τρεις ομάδες (15, 15 και 14 στοιχεία)
  - Αλλά… δε μπορεί να εφαρμοστεί κάποια τεχνική SEM εφόσον υπάρχουν ελλείπουσες τιμές, αν γίνει διαγραφή των εγγραφών με ελλείπουσες τιμές θα διαθέτουμε 0 εγγραφές για ανάλυση



Πηγή: Φραγκόπουλος Δ., 2018, "Εισαγωγή και ολοκλήρωση ICT συστημάτων στις εκπαιδευτικές, οργανωσιακές και επικοινωνιακές δομές της Ανώτατης εκπαίδευσης" Διδακτορική διατριβή, http://ikee.lib.auth.gr/record/303595?ln=el

43

Εφαρμογή σε περιβάλλον SPSS
Η λύση στο δεύτερο πρόβλημα

- **Παράδειγμα με μικρό αριθμό ελλειπουσών τιμών**
  - Little's MCAR test
  - Listwise, pairwise, EM και Regression Missing Value Analysis
  - Multiple Imputation - Analyze Patterns
    - Minimum percentage missing -> 0
  - Multiple Imputation – Impute Missing Data Values
  - Convergence Graphs

- **Παράδειγμα με χρήση τριπλής φόρμας και κοινή φόρμα ελέγχου**
  (2->X, 3->A, 3->B, 3-> C με A,B,C ~33% ελλείπουσες τιμές)
  - Little's MCAR test
  - Multiple Imputation - Analyze Patterns
    - Minimum percentage missing -> 0
  - Multiple Imputation – Impute Missing Data Values
  - Convergence Graphs

44

Άλλες τεχνικές για διαχείριση μικρών δειγμάτων

- **Απόδοση συνόλων με συνοπτικό διαμοιρασμό (parcel summary imputation)**
  - Φτιάχνονται επιπλέον πλήρεις μεταβλητές που προστίθενται ως Predictors για να μπορεί να υπολογιστεί επιτυχώς η απόδοση τιμών σε μικρά δείγματα που έχουν μετρηθεί πολλές μεταβλητές

- **Παθητική απόδοση συνόλων (passive multiple imputation)**
  - Η παθητική απόδοση τιμών εφαρμόζει συνοπτικό διαμοιρασμό απόδοσης σε πολλαπλές αποδόσεις, όπου για να λάβει χώρα η επόμενη απόδοση υπολογίζονται ξανά οι επιπλέον τιμές των επιπλέον μεταβλητών από την προηγούμενη.

45

45

Ευχαριστώ

Διονύσιος Φραγκόπουλος Επικοινωνία: dfragos@ee.auth.gr

We should be suspicious of any dataset (large or small) which appears perfect.

— David J. Hand

Imputing one value for a missing datum cannot be correct in general, because we don't know what value to impute with certainty (if we did, it wouldn't be missing).

— Donald B. Rubin

Van Buuren, S., 2018. Flexible imputation of missing data. CRC press.
https://stefvanbuuren.name/fimd/

46

46