ΒΟΗΘΗΤΙΚΕΣ ΣΗΜΕΙΩΣΕΙΣ ΓΙΑ SPSS

ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΜΕ ΧΡΗΣΗ Η/Υ

Κωνσταντίνος Ζαφειρόπουλος

Τμήμα Διεθνών και Ευρωπαϊκών Σπουδών

Άδειες Χρήσης

Χρηματοδότηση

Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα. Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Πανεπιστήμιο Μακεδονίας**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.

Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

ΜΕΤΑΒΛΗΤΕΣ

Ορισμός: μία ερώτηση (σε ένα ερωτηματολόγιο) που δέχεται μοναδική απάντηση

Τα είδη των μεταβλητών καθορίζονται από τις τιμές τους:

1. ποιοτικές (με κατηγορία)
   - ονομαστικές (nominal) π.χ. φύλο, κόμμα, εθνικότητα
   - διάταξης (ordinal) πχ επίπεδο εκπαίδευσης (πρωτοβάθμια, δευτεροβάθμια, τριτοβάθμια)
2. ποσοτικές
   - διαστήματος (interval) π.χ. θερμοκρασία (το μηδέν δεν δηλώνει ανυπαρξία του φαινομένου της θερμότητας και δεν υπάρχει η έννοια του πολλαπλάσιου ή τουλάχιστον αξιόπιστη ένδειξη- η ίδια θερμοκρασία έχει άλλη ένδειξη σε βαθμούς κελσίου και άλλη σε Φαρενάιτ)
   - αναλογίας (ratio) π.χ. βάρος σε κιλά (το μηδέν δηλώνει ανυπαρξία του φαινομένου και υπάρχει η έννοια του πολλαπλάσιου)

ΣΗΜΕΙΩΣΗ: Όλες οι ποσοτικές μεταβλητές μπορούν να γίνουν κατηγορίες και συνεπώς ποιοτικές μεταβλητές. Και οι ποιοτικές μεταβλητές μπορούν να έχουν αριθμήτικές τιμές, πχ 1 για της πρώτη κατηγορία, 2 για την δεύτερη, κλπ, αλλά οι τιμές δεν εκφράζουν ποσότητα.

ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ (Descriptive Statistics)

- ✓ μέσος όρος (mean): είναι το άθροισμα των τιμών διά το πλήθος τους (διαιρούμε με όσες τιμές προσφέρουν πληροφορία – βαθμοί ελευθερίας)

$$\overline{x} = \frac{\sum x_i}{n}$$

- ✓ διασπορά (variance): μετράμε αποστάσεις από τον μέσο όρο

$$s^2 = \frac{\sum (x_i - \overline{x})^2}{n-1}$$

- ✓ τυπική απόκλιση (standard deviation): προσέγγιση-εκτίμηση της μέσης απόστασης των παρατηρήσεων από τον μέσο όρο τους

$$s = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}}$$

ΕΠΑΓΩΓΙΚΗ ΣΤΑΤΙΣΤΙΚΗ Ή ΣΤΑΤΙΣΤΙΚΗ ΣΥΜΠΕΡΑΣΜΑΤΟΛΟΓΙΑ (Statistical Inference)

Συμπεράσματα για όλο τον πληθυσμό γνωρίζοντας μόνο το δείγμα (n)

1. δείγμα (στατιστικά: $\overline{x}$, s^2, s)
2. εκτίμηση
3. περιθώριο σφάλματος

Δείγμα		**Πληθυσμός**
μπορεί να είναι πολλά		είναι μόνο ένα
στατιστικά	→ εκτίμηση →	Παράμετροι
από ένα στατιστικό προσπαθούμε να βρούμε την παράμετρο:		
$\overline{x}$	→	μ (μέση τιμή)
s^2	→	σ^2
s	→	σ

ΠΕΡΙΓΡΑΦΙΚΑ ΓΙΑ ΜΙΑ ΠΟΣΟΤΙΚΗ ΜΕΤΑΒΛΗΤΗ

Εντολή: Analyze → Descriptive Statistics → Descriptives

ΣΥΧΝΟΤΗΤΕΣ ΓΙΑ ΠΟΙΟΤΙΚΕΣ ΜΕΤΑΒΛΗΤΕΣ

Εντολή: Analyze → Descriptive Statistics → Frequencies

Variables: συνήθως μία ποιοτική μεταβλητή

ΠΙΝΑΚΑΣ ΣΥΝΑΦΕΙΑΣ Ή ΔΙΠΛΗΣ ΕΙΣΟΔΟΥ

Διασταύρωση δύο κατηγορικών (συνήθως ποιοτικών) μεταβλητών.

Εντολή: Analyze → Descriptive Statistics → Crosstabs

ΣΥΓΚΡΙΣΗ ΜΕΣΩΝ ΟΡΩΝ

Υπολογισμός μέσων όρων μιας ποσοτικής μεταβλητής ανά κατηγορία μιας ποιοτικής μεταβλητής.

Εντολή: Analyze → Compare means → Means

Dependent list: ποσοτική μεταβλητή
Independentlist: ποιοτική μεταβλητή

ΕΠΑΝΑΚΩΔΙΚΟΠΟΙΗΣΗ ΔΕΔΟΜΕΝΩΝ

Εντολή: Transform→ Recode into different variables

Επιλέγουμε την μεταβλητή και την περνάμε στο διπλανό πλαίσιο.
Στο Output Variable ορίζουμε νέο όνομα και ετικέτα (name & label)για τη μεταβλητή και πατάμε το Change.
Στην επιλογή Old and new values επανακωδικοποιούμε τις τιμές

π.χ. Old value → Value: 3
New value → Value: 2
Add
Old → New: 3 → 2

Χρησιμοποιούμε την επιλογή Range για να μετατρέψουμε μια συνεχή μεταβλητή σε κατηγορική. π.χ. για να χωρίσουμε τις τιμές στις εξής τρεις κατηγορίες α) ≤ 25, β) 25-50 και γ) ≥ 50 επιλέγουμε:

- ⊙ Range Lowest through 25
- ⊙ Range 25 through 50
- ⊙ Range 50 through highest

ΑΥΤΟΜΑΤΗ ΕΠΑΝΑΚΩΔΙΚΟΠΟΙΗΣΗ

Εντολή: Transform → Automatic recode

New name: _____
Add new name
Recode starting from

- lowest value
- highest value

ΕΠΙΛΟΓΗ ΠΕΡΙΠΤΩΣΕΩΝ

Εντολή: Data → Select Cases

Ενεργοποιούμε το If condition is satisfied και επιλέγουμε το If για να ορίσουμε τη συνθήκη.

Επεξήγηση συμβόλων:

&	and
|	or
~	not
~=	not equal (όχι ίσον)
**	ύψωση σε δύναμη

ΠΡΟΣΟΧΗ! Όταν τελειώσουμε τη διαδικασία που μας ενδιαφέρει, πάντα διαγράφουμε τη συνθήκη και δουλεύουμε με όλες τις περιπτώσεις για να αποφύγουμε σφάλματα.

ΕΛΕΓΧΟΙ ΥΠΟΘΕΣΕΩΝ

Ερώτημα: Αρχείο Emploee data.sav Να δημιουργηθεί ο πίνακας συνάφειας με μεταβλητές το φύλο και τη θέση εργασίας. Σύμφωνα με τον πίνακα, οι άντρες ή οι γυναίκες βρίσκονται ψηλότερα στην ιεραρχία; Παίζει ρόλο το φύλο στη θέση εργασίας που κατέχει κάποιος εργαζόμενος; (Ή αλλιώς: Επηρεάζει το φύλο τη θέση εργασίας; **Είναι το φύλο και η θέση εργασίας εξαρτημένα ή ανεξάρτητα;)**

Εντολή: Analyze → Descriptive Statistics → Crosstabs

Rows: gender
Columns: employment category

Αν πατήσουμε OK ως εδώ έχουμε τον πίνακα συνάφειας (και έχουμε ολοκληρώσει το πρώτο σκέλος του ερωτήματος). Αλλά…

Στην επιλογή Cells στην υποκατηγορία Percentages τσεκάρουμε το ☑ Row για να πάρουμε τα ποσοστά που χρειαζόμαστε για να απαντήσουμε αν οι άντρες ή οι γυναίκες βρίσκονται υψηλότερα στην ιεραρχία.

Στο Output…

Gender * Employment Category Crosstabulation

			Employment Category			
			Clerical	Custodial	Manager	Total
Gender	Female	Count	206	0	10	216
		% within Gender	95,4%	,0%	4,6%	100,0%
	Male	Count	157	27	74	258
		% within Gender	60,9%	10,5%	28,7%	100,0%
Total		Count	363	27	84	474
		% within Gender	76,6%	5,7%	17,7%	100,0%

Απάντηση: Οι άντρες βρίσκονται υψηλότερα στην ιεραρχία από τις γυναίκες (με βάση το ποσοστό % within gender). [στην απάντηση δίνουμε και τα ποσοστά που αναγράφονται στον πίνακα]

Για να απαντήσουμε στο τελευταίο ερώτημα θα χρησιμοποιήσουμε επαγωγική στατιστική. Η διαδικασία που θα ακολουθήσουμε λέγεται Έλεγχος Υποθέσεων.

H_0: Είναι η μηδενική υπόθεση, η κύρια υπόθεσή μας. Είναι πάντα η υπόθεση της μη διαφοροποίησης ή της ισότητας.
H_1: Είναι η εναλλακτική υπόθεση.

Στην περίπτωσή μας:

H_0: Το φύλο και η θέση εργασίας είναι ανεξάρτητα.

H_1: Το φύλο και η θέση εργασίας είναι εξαρτημένα.

Ένας πρώτος τρόπος να δούμε αν οι μεταβλητές μας είναι ανεξάρτητες ή όχι είναι με βάση τις αναμενόμενες συχνότητες. Αναμενόμενες συχνότητες είναι οι συχνότητες που θα είχε ο πίνακας αν οι μεταβλητές ήταν ανεξάρτητες. Για να υπολογίσει το πρόγραμμα τις αναμενόμενες συχνότητες, στην επιλογή Cells στην υποκατηγορία Counts επιλέγουμε και το ☑ Expected. Εάν οι μεταβλητές είναι ανεξάρτητες, οι παρατηρούμενες και οι αναμενόμενες συχνότητες πρέπει να είναι πολύ κοντά, να έχουν μικρή απόσταση.

Gender * Employment Category Crosstabulation

			Employment Category			
			Clerical	Custodial	Manager	Total
Gender	Female	Count	206	0	10	216
		Expected Count	165,4	12,3	38,3	216,0
		% within Gender	95,4%	,0%	4,6%	100,0%
	Male	Count	157	27	74	258
		Expected Count	197,6	14,7	45,7	258,0
		% within Gender	60,9%	10,5%	28,7%	100,0%
Total		Count	363	27	84	474
		Expected Count	363,0	27,0	84,0	474,0
		% within Gender	76,6%	5,7%	17,7%	100,0%

ΣΗΜΕΙΩΣΗ: Στον πίνακα «Crosstabulation» του Output

Count (εννοείται observed): είναι η παρατηρούμενη συχνότητα

Expected count: είναι η αναμενόμενη συχνότητα

% within gender: είναι η δεσμευμένη πιθανότητα ή πιθανότητα υπό συνθήκη

ΔΟΚΙΜΑΣΙΑ X^2

Προκειμένου να πραγματοποιηθεί ο έλεγχος υποθέσεων, υπολογίζουμε το X^2. Το **X^2** είναι ένα στατιστικό που παράγεται λαμβάνοντας υπόψη τις αποστάσεις των παρατηρούμενων από τις αναμενόμενες συχνότητες και αφορά το σύνολο των κελιών του πίνακα. Οι όροι που χρησιμοποιούνται για τον υπολογισμό του είναι τετραγωνικοί (δηλαδή ≥ 0). Όσο πιο μεγάλο είναι το X^2, τόσο πιο σίγουροι είμαστε για την εξάρτηση.

Για να υπολογίσουμε το X^2 στο SPSS, στην επιλογή Statistics επιλέγουμε το ☑ Chi-square.

Στο Output παίρνουμε τον πίνακα «Chi-Square Tests».

- X^2: είναι ο αριθμός στο πρώτο κελί (Pearson Chi-square / Value)
- df: είναι οι βαθμοί ελευθερίας (degrees of freedom)
- asymptotic significance: είναι το επίπεδο ή η στάθμη σημαντικότητας και συμβολίζεται με το p.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	79,277[a]	2	,000
Likelihood Ratio	95,463	2	,000
N of Valid Cases	474		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 12,30.

Γενικά,

- αν $p > 0,05$ δεν μπορούμε να απορρίψουμε την H_0
- αν $p < 0,05$ απορρίπτουμε την H_0 και δεχόμαστε την H_1

Στην περίπτωσή μας:

$p = 0,000$ (δεν είναι ακριβώς μηδέν αλλά πολύ μικρό),δηλαδή $p < 0,05$ και επομένως απορρίπτουμε την H_0 και δεχόμαστε την H_1

Απάντηση: Το φύλο και η θέση εργασίας είναι εξαρτημένα.

ΠΡΟΣΟΧΗ! Ο πίνακας «Chi-Square Tests» έχει μια υποσημείωση. Το ποσοστό που δίνεται στην παρένθεση της υποσημείωσης δεν πρέπει να είναι πάνω από 20% για να είναι αξιόπιστος ο έλεγχος X^2. Ο περιορισμός αυτός δεν ελέγχεται όταν ο πίνακας συνάφειας είναι 2^x2.

T-Test ΕΛΕΓΧΟΣ ΥΠΟΘΕΣΕΩΝ

Στην ουσία είναι έλεγχος μέσων τιμών.

Προϋποθέσεις:

1. Δύο μεταβλητές: μία ποσοτική και μία ποιοτική με μόνο δύο κατηγορίες (αναλύουμε την ποσοτική μεταβλητή με βάση την ποιοτική, η οποία απλά ορίζει ομάδες).
2. Κανονικότητα της ποσοτικής μεταβλητής για κάθε τιμή της ποιοτικής μεταβλητής (κανονική κατανομή).

Ερώτημα: Να ελεγχθεί αν υπάρχει στατιστικά σημαντική διαφορά ανάμεσα στους μισθούς των λευκών και αυτών που ανήκουν σε μειονότητα.

ΕΛΕΓΧΟΣ ΚΑΝΟΝΙΚΟΤΗΤΑΣ

H_0: Η μεταβλητή ακολουθεί την κανονική κατανομή.
H_1: Η μεταβλητή δεν ακολουθεί την κανονική κατανομή.

Εντολή: Analyze → Descriptive Statistics → Explore

Dependent list: current salary (ποσοτική μεταβλητή)
Factor list: minority classification (ποιοτική μεταβλητή)
Στην επιλογή Plots επιλέγουμε το ☑Normality plots with tests

Στο Output μας ενδιαφέρει ο πίνακας «Tests of normality».

Tests of Normality

	Minority Classification	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Current Salary	No	,195	370	,000	,802	370	,000
	Yes	,244	104	,000	,634	104	,000

a. Lilliefors Significance Correction

- κοιτάζουμε το Kolmogorov-Smirnov Test για N > 50
- κοιτάζουμε το Shapiro-Wilk για N ≤ 50
- το N φαίνεται από τη στήλη df
- και στα δύο αυτά τεστ και για τις δύο κατηγορίες της ποιοτικής μεταβλητής μας ενδιαφέρει το Significance
  - αν $p > 0,05$ δεν μπορούμε να απορρίψουμε την H_0
  - αν $p < 0,05$ απορρίπτουμε την H_0 και δεχόμαστε την H_1

Εάν υπάρχει κανονικότητα συνεχίζουμε το T-Test. Αν δεν υπάρχει κανονικότητα κάνουμε μη παραμετρικούς ελέγχους (Non-Parametric Statistics) εναλλακτικά στο T-Test.

Στην περίπτωσή μας:

- No: p = 0,000 στο Kolmogorov-Smirnov Test, δηλαδή p < 0,05 άρα απορρίπτουμε την H_0 και δεχόμαστε την H_1 (δεν υπάρχει κανονικότητα)
- Yes: p = 0,000 στο Kolmogorov-Smirnov Test, δηλαδή p < 0,05 άρα απορρίπτουμε την H_0 και δεχόμαστε την H_1 (δεν υπάρχει κανονικότητα)

Καταχρηστικά θα κάνουμε T-Test σαν να είχαμε κανονικότητα και στη συνέχεια θα δούμε και την σωστή επιλογή, τους μη παραμετρικούς ελέγχους.

T-Test έλεγχος υποθέσεων (συνέχεια)

$H_0: \mu_1 = \mu_2$
$H_1: \mu_1 \neq \mu_2$

όπου μ: μέση τιμή πληθυσμού
μ_1: μέση τιμή μισθού λευκών
μ_2: μέση τιμή μισθού μειονοτικών

Εντολή: Analyze → Compare means → Independent-Samples T-Test

Test variable: current salary (ποσοτική μεταβλητή)
Grouping variable: minority classification (ποιοτική μεταβλητή)
Define groups → Δηλώνουμε τους αριθμούς ανάλογα με την κωδικοποίηση στο Data View. Στην περίπτωσή μας:

- Group 1: 0
- Group 2: 1

Στο Output παίρνουμε δύο πίνακες:

1. Ο πίνακας «Group Statistics» μας δίνει το σύνολο των περιπτώσεων (N), το μέσο όρο (mean) και την τυπική απόκλιση (standard deviation) και για τις δύο κατηγορίες της ποιοτικής μεταβλητής.

Group Statistics

	Minority Classification	N	Mean	Std. Deviation	Std. Error Mean
Current Salary	No	370	$36,023.31	$18,044.096	$938.068
	Yes	104	$28,713.94	$11,421.638	$1,119.984

2. Στον πίνακα «Independent Samples Test» μας ενδιαφέρουν τα εξής:
   - ✓ Levene's Test for Equality of Variances: είναι τεστ για την ισότητα των διασπορών.
     - ➢ αν p > 0,05 οι διασπορές είναι ίσες
     - ➢ αν p < 0,05 οι διασπορές είναι άνισες

Όταν οι διασπορές είναι ίσες, στην ανάγνωση του υπόλοιπου πίνακα (t-test for equality of means) κοιτάμε την πρώτη γραμμή (equal variances assumed). Όταν είναι άνισες, κοιτάμε τη δεύτερη γραμμή (equal variances not assumed).

Στην περίπτωσή μας: p = 0,000, δηλαδή p < 0,05 άρα οι διασπορές είναι άνισες και επομένως από εδώ και πέρα θα κοιτάμε μόνο τη δεύτερη γραμμή.

✓ T-Test for Equality of Means

- δηλώνουμε πόσο είναι το t (t = 5,003)
- Significance (2-tailed): δηλώνουμε πόσο είναι το p (p = 0,000) και με βάση αυτό τον αριθμό δεχόμαστε ή απορρίπτουμε την H_0.
  - αν p > 0,05 δεν μπορούμε να απορρίψουμε την H_0
  - αν p < 0,05 απορρίπτουμε την H_0 και δεχόμαστε την H_1

  Στην περίπτωσή μας p < 0,05 άρα απορρίπτουμε την H_0 ($\mu_1 = \mu_2$) και δεχόμαστε την H_1 ($\mu_1 \neq \mu_2$).

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Current Salary	Equal variances assumed	28,487	,000	3,915	472	,000	$7,309.369	$1,867.111	$3,640.491	$10,978.246
	Equal variances not assumed			5,003	262,188	,000	$7,309.369	$1,460.936	$4,432.707	$10,186.030

Απάντηση: Υπάρχει στατιστικά σημαντική διαφορά του μισθού των λευκών και των μειονοτικών.

ΜΗ ΠΑΡΑΜΕΤΡΙΚΟΙ ΕΛΕΓΧΟΙ

Στους μη παραμετρικούς έλεγχους (Non-Parametric Statistics) δεν υπάρχει η προϋπόθεση της κανονικότητας. Ελέγχουν την ομοιογένεια (και όχι την ισότητα των μέσων τιμών) και χρησιμοποιούν βαθμούς διατακτικότητας (ranks).

H_0: Υπάρχει ομοιογένεια στους μισθούς λευκών – μειονοτικών.
H_1: Δεν υπάρχει ομοιογένεια στους μισθούς λευκών – μειονοτικών.

Εντολή: Analyze → Nonparametric tests → Independent Samples

Καρτέλα Fields:
- Test fields: current salary (ποσοτική μεταβλητή)
- Groups: minority classification (ποιοτική μεταβλητή)

Καρτέλα Settings:
- επιλέγουμε το ⊙Customize Tests και στη συνέχεια το ☑Mann-Whitney U (2 samples)

Run

Στο Output…

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Current Salary is the same across categories of Minority Classification.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Με βάση το Significance δεχόμαστε ή απορρίπτουμε την H_0

- αν $p > 0,05$ δεν μπορούμε να απορρίψουμε την H_0
- αν $p < 0,05$ απορρίπτουμε την H_0 και δεχόμαστε την H_1

Στην περίπτωσή μας:
$p = 0,000$ δηλαδή $p < 0,05$ άρα απορρίπτουμε την H_0 και δεχόμαστε την H_1

Απάντηση: Δεν υπάρχει ομοιογένεια στους μισθούς λευκών – μειονοτικών. Άρα υπάρχει στατιστικά σημαντική διαφορά του μισθού των λευκών και των μειονοτικών.

ΖΕΥΓΑΡΩΤΟ T-Test (Paired-Samples T-Test)

Προϋποθέσεις:

1. Δύο ποσοτικές μεταβλητές (που πρέπει να αναφέρονται στο ίδιο πράγμα)
2. Κανονικότητα

Ερώτημα: Να ελεγχθεί αν υπάρχει στατιστικά σημαντική διαφορά ανάμεσα στον αρχικό και τον τωρινό μισθό.

Έλεγχος κανονικότητας

H_0: Η μεταβλητή ακολουθεί την κανονική κατανομή.
H_1: Η μεταβλητή δεν ακολουθεί την κανονική κατανομή.

Εντολή: Analyze → Descriptive Statistics → Explore

Dependent list: current salary & beginning salary (βάζουμε και τις δύο ποσοτικές μεταβλητές)
Factor list: δεν βάζουμε τίποτα γιατί δεν έχουμε ποιοτική μεταβλητή (προσοχή!)
Στην επιλογή Plots επιλέγουμε το ☑Normality plots with tests

Στο Output μας ενδιαφέρει ο πίνακας «Tests of normality».

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	Df	Sig.	Statistic	df	Sig.
Current Salary	,208	474	,000	,771	474	,000
Beginning Salary	,252	474	,000	,715	474	,000

a. Lilliefors Significance Correction

- κοιτάζουμε το Kolmogorov-Smirnov Test για N > 50
- κοιτάζουμε το Shapiro-Wilk για N ≤ 50
- το N φαίνεται από τη στήλη df
- και στα δύο αυτά τεστ και για τις δύο μεταβλητές μας ενδιαφέρει το Significance
  - αν p > 0,05 δεν μπορούμε να απορρίψουμε την H_0
  - αν p < 0,05 απορρίπτουμε την H_0 και δεχόμαστε την H_1

Εάν υπάρχει κανονικότητα συνεχίζουμε το ζευγαρωτό T-Test. Αν δεν υπάρχει κανονικότητα κάνουμε μη παραμετρικούς ελέγχους (Non-Parametric Statistics) εναλλακτικά στο ζευγαρωτό T-Test.

Στην περίπτωσή μας:

- current salary: p = 0,000 στο Kolmogorov-Smirnov test, δηλαδή p < 0,05 άρα απορρίπτουμε την H_0 και δεχόμαστε την H_1(δεν υπάρχει κανονικότητα)

- beginning salary: p = 0,000 στο Kolmogorov-Smirnov test, δηλαδή p < 0,05 άρα απορρίπτουμε την H_0 και δεχόμαστε την H_1 (δεν υπάρχει κανονικότητα)

Καταχρηστικά θα προχωρήσουμε στο ζευγαρωτό T-Test σαν να είχαμε κανονικότητα για χάρη του παραδείγματος.

Ζευγαρωτό T-Test (συνέχεια)

H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 \neq \mu_2$

όπου μ_1: μέσος αρχικός μισθός
μ_2: μέσος τωρινός μισθός

Εντολή: Analyze → Compare means → Paired-Samples T-Test

Paired Variables:

Pair	Variable 1	Variable 2
1	current salary	beginning salary
2		

Στο Output παίρνουμε τρεις πίνακες:

1. με βάση τον πίνακα «Paired Samples Statistics» δίνουμε τα περιγραφικά (μέσος όρος, σύνολο περιπτώσεων, τυπική απόκλιση)

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Current Salary	$34,419.57	474	$17,075.661	$784.311
	Beginning Salary	$17,016.09	474	$7,870.638	$361.510

2. στον πίνακα «Paired Samples Correlations» μας ενδιαφέρει το correlation, ο συντελεστής συσχέτισης. Παίρνει τιμές από -1 μέχρι 1 (- αρνητική συσχέτιση, + θετική συσχέτιση). Θέλουμε να έχει υψηλή θετική τιμή (πάνω από 0,4) αλλιώς δεν κάνουμε T-Test.

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Current Salary & Beginning Salary	474	,880	,000

3. στον πίνακα «Paired Samples Test» δηλώνουμε το t (t = 35,036) και το df (df = 473) και με βάση το Sig. (2-tailed) δεχόμαστε ή απορρίπτουμε την H_0.

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Current Salary - Beginning Salary	$17,403.481	$10,814.620	$496.732	$16,427.407	$18,379.555	35,036	473	,000

Στην περίπτωσή μας: p = 0,000 δηλαδή p < 0,05 άρα απορρίπτουμε την H_0 ($\mu_1 = \mu_2$) και αποδεχόμαστε την H_1 ($\mu_1 \neq \mu_2$)

Απάντηση: Υπάρχει στατιστικά σημαντική διαφορά ανάμεσα στον αρχικό και τον τωρινό μισθό.

Ο εναλλακτικός μη παραμετρικός έλεγχος στο ζευγαρωτό T-Test είναι ο παρακάτω.

ΜΗ ΠΑΡΑΜΕΤΡΙΚΟΙ ΕΛΕΓΧΟΙ – ΕΛΕΓΧΟΣ ΤΟΥ WILCOXON

Ερώτημα: Να ελεγχθεί αν υπάρχει στατιστικά σημαντική διαφορά ανάμεσα στον αρχικό και τον τωρινό μισθό.

H_0: Υπάρχει ομοιογένεια ανάμεσα στον αρχικό και τον τωρινό μισθό.
H_1: Δεν υπάρχει ομοιογένεια ανάμεσα στον αρχικό και τον τωρινό μισθό.

Εντολή: Analyze→ Nonparametric Tests→ Related Samples

Στην καρτέλα Fields στο Test Fields περνάμε τα current & beginning salary
Στην καρτέλα Settings επιλέγουμε το ⊙Customize Tests και στη συνέχεια το ☑Wilcoxon matched-pair signed-rank (2 samples)
Run

ΣτοOutput…

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The median of differences between Current Salary and Beginning Salary equals 0.	Related-Samples Wilcoxon Signed Ranks Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Με βάση το Significance δεχόμαστε ή απορρίπτουμε την H_0

- αν $p > 0,05$ δεν μπορούμε να απορρίψουμε την H_0
- αν $p < 0,05$ απορρίπτουμε την H_0 και δεχόμαστε την H_1

Στην περίπτωσή μας:
$p = 0,000$ δηλαδή $p < 0,05$ άρα απορρίπτουμε την H_0 και δεχόμαστε την H_1

Απάντηση: Δεν υπάρχει ομοιογένεια ανάμεσα στον αρχικό και τον τωρινό μισθό και επομένως υπάρχει στατιστικά σημαντική διαφορά ανάμεσα στον αρχικό και τον τωρινό μισθό.

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ Ή ΔΙΑΣΠΟΡΑΣ (ANOVA (analysis of variance)) ΜΕ ΕΝΑΝ ΠΑΡΑΓΟΝΤΑ

Προϋποθέσεις:

1. Μία ποσοτική και μία ποιοτική μεταβλητή με πάνω από δύο κατηγορίες (αν οι κατηγορίες ήταν δύο θα κάναμε T-Test).
2. Η ποσοτική μεταβλητή πρέπει να ακολουθεί κανονική κατανομή (κανονικότητα) για κάθε τιμή της ποιοτικής.

Θέλουμε το within (εσωτερική διασπορά) να είναι μικρό και το between (εξωτερική διασπορά – απόσταση) να είναι μεγάλο.

F = s^2 between / s^2 within

- μικρό F οι ομάδες δεν διαφέρουν
- μεγάλο F οι ομάδες διαφέρουν

Ερώτημα: Να ελεγχθεί αν υπάρχει στατιστικά σημαντική διαφορά του (τωρινού) μισθού ανάμεσα στις τρεις ομάδες εργαζομένων.

Έλεγχος κανονικότητας

H_0: Η μεταβλητή ακολουθεί την κανονική κατανομή.
H_1: Η μεταβλητή δεν ακολουθεί την κανονική κατανομή.

Εντολή: Analyze → Descriptive Statistics → Explore

Dependent list: current salary (ποσοτική μεταβλητή)
Factor list: employment category (ποιοτική μεταβλητή)
Στην επιλογή Plots επιλέγουμε το ☑Normality plots with tests

Στο Output μας ενδιαφέρει ο πίνακας «Tests of normality».

Tests of Normality

	Employment Category	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
Current Salary	Clerical	,107	363	,000	,882	363	,000
	Custodial	,276	27	,000	,818	27	,000
	Manager	,109	84	,016	,929	84	,000

a. Lilliefors Significance Correction

- κοιτάζουμε το Kolmogorov-Smirnov Test για N > 50
- κοιτάζουμε το Shapiro-Wilk για N ≤ 50
- το N φαίνεται από τη στήλη df
- και στα δύο αυτά τεστ και για τις τρεις κατηγορίες της ποιοτικής μεταβλητής μας ενδιαφέρει το Significance (θα πρέπει p > 0,05 και για τις τρεις κατηγορίες

για να υπάρχει κανονικότητα – αν έστω και ένα από τα τρία p είναι < 0,05 δεν υπάρχει κανονικότητα)

- αν p > 0,05 δεν μπορούμε να απορρίψουμε την H_0
- αν p < 0,05 απορρίπτουμε την H_0 και δεχόμαστε την H_1

Εάν υπάρχει κανονικότητα συνεχίζουμε την ανάλυση διακύμανσης. Αν δεν υπάρχει κανονικότητα κάνουμε μη παραμετρικούς ελέγχους (Non-Parametric Statistics) εναλλακτικά στην ανάλυση διακύμανσης.

Στην περίπτωσή μας,

- clerical: p = 0,000 στο Kolmogorov-Smirnov Test, δηλαδή p < 0,05 άρα απορρίπτουμε την H_0 και δεχόμαστε την H_1 (δεν υπάρχει κανονικότητα)
- custodial: p = 0,000 στο Shapiro-Wilk, δηλαδή p < 0,05 άρα απορρίπτουμε την H_0 και δεχόμαστε την H_1 (δεν υπάρχει κανονικότητα)
- manager: p = 0,016 στο Kolmogorov-Smirnov Test, δηλαδή p < 0,05 άρα απορρίπτουμε την H_0 και δεχόμαστε την H_1 (δεν υπάρχει κανονικότητα)

Καταχρηστικά θα προχωρήσουμε στο ANOVA.

Ανάλυση διακύμανσης (συνέχεια)

H_0: Δεν υπάρχει διαφορά στους μέσους όρους → μ1=μ2=μ3
H_1: Υπάρχει διαφορά στους μέσους όρους

Εντολή: Analyze → Compare Means → One-Way ANOVA

Dependent List: current salary (ποσοτική μεταβλητή)
Factor: employment category (ποιοτική μεταβλητή)
Στο Post Hoc επιλέγουμε το ☑Scheffe
Στο Options στην υποκατηγορία Statistics επιλέγουμε το ☑Descriptive

Στο Output παίρνουμε τους εξής πίνακες:

1. Ο πίνακας «Descriptives» μας δίνει το σύνολο των περιπτώσεων (N), το μέσο όρο (mean) και την τυπική απόκλιση (standard deviation) και για τις τρεις κατηγορίες της ποιοτικής μεταβλητής.

Descriptives

Current Salary

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
Clerical	363	$27,838.54	$7,567.995	$397.217	$27,057.40	$28,619.68	$15,750	$80,000
Custodial	27	$30,938.89	$2,114.616	$406.958	$30,102.37	$31,775.40	$24,300	$35,250
Manager	84	$63,977.80	$18,244.776	$1,990.668	$60,018.44	$67,937.16	$34,410	$135,000
Total	474	$34,419.57	$17,075.661	$784.311	$32,878.40	$35,960.73	$15,750	$135,000

2. Στον πίνακα «ANOVA» δηλώνουμε:

ANOVA

Current Salary

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	8,944E10	2	4,472E10	434,481	,000
Within Groups	4,848E10	471	1,029E8		
Total	1,379E11	473			

- το F [στην περίπτωσή μας π.χ. F(2,471) = 434,481 όπου 2 και 471 είναι οι βαθμοί ελευθερίας (ή df) των between groups και within groups]
- το p με βάση το Significance
  - αν $p > 0,05$ δεν μπορούμε να απορρίψουμε την H_0
  - αν $p < 0,05$ απορρίπτουμε την H_0 και δεχόμαστε την H_1

  Αν δεχτούμε την H_1 πρέπει να κάνουμε και έλεγχο Post Hoc (σε ομάδες ανά δύο) για να δούμε πού οφείλεται η διαφορά / ποιες ομάδες διαφέρουν μεταξύ τους. [Στην περίπτωσή μας $p = 0,000$ δηλαδή $p < 0,05$ άρα απορρίπτουμε την H_0 και δεχόμαστε την H_1]

3. Στον πίνακα «Multiple comparisons» με βάση το Significance δηλώνουμε τα p για τα ζεύγη των κατηγοριών.

Multiple Comparisons

Current Salary
Scheffe

(I) Employment Category	(J) Employment Category	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Clerical	Custodial	$-3,100.349	$2,023.760	,310	$-8,069.80	$1,869.10
	Manager	$-36,139.258*	$1,228.352	,000	$-39,155.54	$-33,122.98
Custodial	Clerical	$3,100.349	$2,023.760	,310	$-1,869.10	$8,069.80
	Manager	$-33,038.909*	$2,244.409	,000	$-38,550.17	$-27,527.65
Manager	Clerical	$36,139.258*	$1,228.352	,000	$33,122.98	$39,155.54
	Custodial	$33,038.909*	$2,244.409	,000	$27,527.65	$38,550.17

*. The mean difference is significant at the 0.05 level.

- αν $p > 0,05$ δεν μπορούμε να πούμε ότι υπάρχει στατιστικά σημαντική διαφορά
- αν $p < 0,05$ υπάρχει στατιστικά σημαντική διαφορά

Στην περίπτωσή μας,

- clerical – custodial: $p = 0,310$ δηλαδή $p > 0,05$ άρα δεν υπάρχει στατιστικά σημαντική διαφορά
- clerical – manager: $p = 0,000$ δηλαδή $p < 0,05$ άρα υπάρχει στατιστικά σημαντική διαφορά
- custodial – manager: $p = 0,000$ δηλαδή $p < 0,05$ άρα υπάρχει στατιστικά σημαντική διαφορά

Απάντηση: Υπάρχει στατιστικά σημαντική διαφορά του (τωρινού) μισθού ανάμεσα στις τρεις ομάδες εργαζομένων και οφείλεται στη διαφορά που έχει ο μισθός των manager από τις άλλες δύο κατηγορίες.

ΜΗ ΠΑΡΑΜΕΤΡΙΚΟΙ ΕΛΕΓΧΟΙ – ΕΛΕΓΧΟΣ KRUSKAL-WALLIS

Εφόσον στην πραγματικότητα δεν υπάρχει κανονικότητα, ακολουθούμε την παρακάτω διαδικασία.

H_0: Υπάρχει ομοιογένεια ως προς το μισθό ανάμεσα στις τρεις ομάδες εργαζομένων.
H_1: Δεν υπάρχει ομοιογένεια ως προς το μισθό ανάμεσα στις τρεις ομάδες εργαζομένων.

Εντολή: Analyze→ Nonparametric Tests→ Independent Samples

Στην καρτέλα Fields στο Test Fields περνάμε το current salary και στο Groups περνάμε το employment category.
Στην καρτέλα Settings επιλέγουμε το ⊙Customize Tests και στη συνέχεια το ☑Kruskal-Wallis 1-way ANOVA (k samples)
Run

ΣτοOutput…

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Current Salary is the same across categories of Employment Category.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Με βάση το Significance δεχόμαστε ή απορρίπτουμε την H_0

- αν $p > 0,05$ δεν μπορούμε να απορρίψουμε την H_0
- αν $p < 0,05$ απορρίπτουμε την H_0 και δεχόμαστε την H_1

Στην περίπτωσή μας:
$p = 0,000$ δηλαδή $p < 0,05$ άρα απορρίπτουμε την H_0 και δεχόμαστε την H_1

Απάντηση: Δεν υπάρχει ομοιογένεια ως προς το μισθό ανάμεσα στις τρεις ομάδες εργαζομένων και επομένως υπάρχει στατιστικά σημαντική διαφορά του (τωρινού) μισθού ανάμεσα στις τρεις ομάδες εργαζομένων.

ΠΡΟΣΟΧΗ! Για να δούμε ποιες ομάδες διαφοροποιούνται από ποιες πρέπει να κάνουμε ελέγχους ανά δύο (δηλαδή τρεις ελέγχους Mann-Whitney) αλλά μόνο αφού πρώτα έχουμε επιλέξει περιπτώσεις (Select Cases).

Έλεγχοι ανά δύο (Mann-Whitney)

1. Data → Select cases → If condition is satisfied → If → jobcat = 1 | jobcat = 2
   Analyze → Nonparametric tests → Independent Samples
   Καρτέλα Fields:
   - Test fields: current salary (ποσοτική μεταβλητή)
   - Groups: minority classification (ποιοτική μεταβλητή)

   Καρτέλα Settings:
   - επιλέγουμε το ⊙Customize Tests και στη συνέχεια το ☑Mann-Whitney U (2 samples)

   Run
   Στο Output…

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Current Salary is the same across categories of Employment Category.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

p = 0,000 δηλαδή p < 0,05 άρα απορρίπτουμε την H_0 και δεχόμαστε την H_1 (δεν υπάρχει ομοιογένεια ανάμεσα στην πρώτη και τη δεύτερη ομάδα)

2. Data → Select cases → If condition is satisfied → If → jobcat = 1 | jobcat = 3
   Τα υπόλοιπα ακριβώς ίδια με πριν…
   Στο Output…

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Current Salary is the same across categories of Employment Category.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

p = 0,000 δηλαδή p < 0,05 άρα απορρίπτουμε την H_0 και δεχόμαστε την H_1 (δεν υπάρχει ομοιογένεια ανάμεσα στην πρώτη και την τρίτη ομάδα)

3. Data → Select cases → If condition is satisfied → If → jobcat = 2 | jobcat = 3
   Τα υπόλοιπα ακριβώς ίδια με πριν…
   Στο Output…

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Current Salary is the same across categories of Employment Category.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

p = 0,000 δηλαδή p < 0,05 άρα απορρίπτουμε την H_0 και δεχόμαστε την H_1 (δεν υπάρχει ομοιογένεια ανάμεσα στην δεύτερη και την τρίτη ομάδα)

Τελική απάντηση: Υπάρχει στατιστικά σημαντική διαφορά του (τωρινού) μισθού ανάμεσα και στις τρεις ομάδες εργαζομένων (όλες διαφέρουν από όλες).

ΣΥΝΤΕΛΕΣΤΕΣ ΣΥΣΧΕΤΙΣΗΣ (correlation coefficients)

Συσχέτιση δε σημαίνει σχέση ή αιτιότητα, αλλά εμφάνιση ταυτόχρονη ή με την ίδια ή αντίθετη φορά.

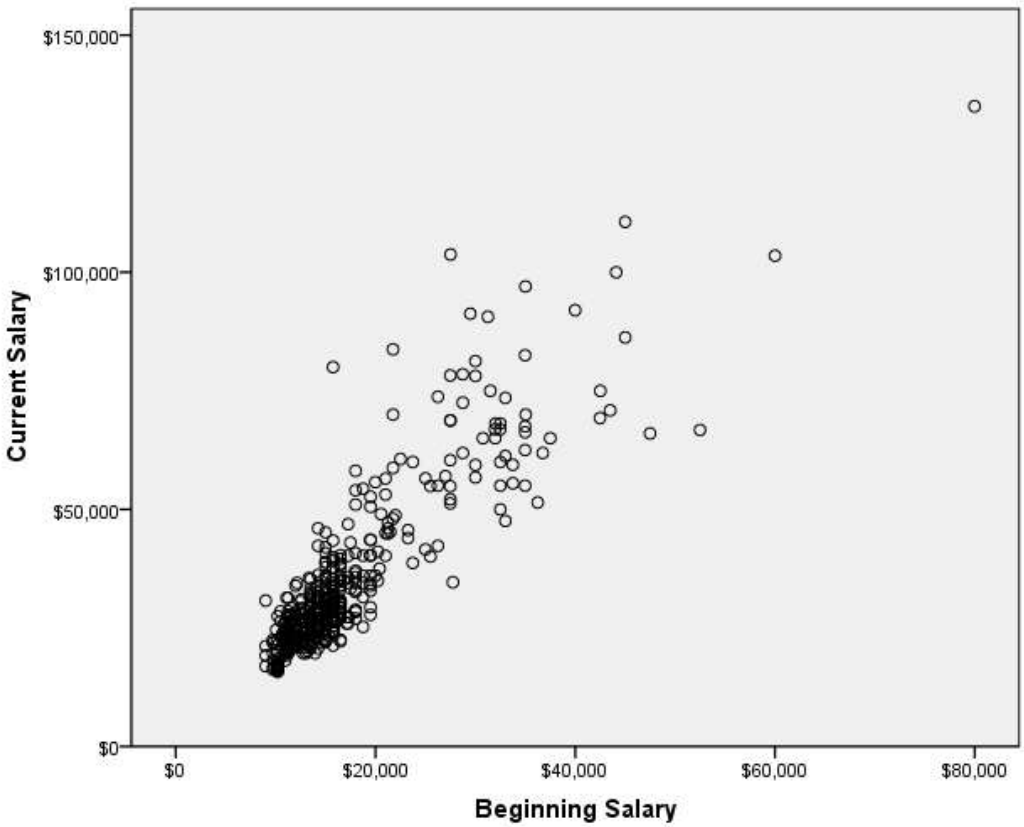
- θετική συσχέτιση: μεγαλώνει το ένα – μεγαλώνει και το άλλο, μικραίνει το ένα – μικραίνει και το άλλο (π.χ.)
- αρνητική συσχέτιση: μεγαλώνει το ένα – μικραίνει το άλλο

Θα δημιουργήσουμε ένα διάγραμμα διασποράς (scatterplot). Είναι γραφική παράσταση που συνδέει δύο ποσοτικές μεταβλητές (η μία μπαίνει στον άξονα Χ και η άλλη στον Υ).

Εντολή: Graphs→ Chart Builder

Στην καρτέλα Gallery επιλέγουμε το Scatter/Dot και σέρνουμε την πρώτη εικόνα (simple scatter) στο παραπάνω πλαίσιο.
X-axis: beginning salary (ανεξάρτητη μεταβλητή)*
Y-axis: current salary (εξαρτημένη μεταβλητή)*
* Ο συντελεστής συσχέτισης είναι συμμετρικός (δεν έχει σημασία ποια μεταβλητή μπαίνει στο x και ποια στο y), αλλά για περαιτέρω ανάλυση θα τις ξεχωρίζουμε σε εξαρτημένη (current salary) και σε ανεξάρτητη (beginning salary).

Στο Output…



Στο γράφημα βλέπουμε ότι υπάρχει γραμμικότητα (υπάρχει η τάση όλες οι τελείες να είναι πάνω στην ίδια γραμμή). Όσο πιο μεγάλη διασπορά έχει το νέφος, τόσο πιο πολύ απομακρυνόμαστε από τη γραμμικότητα.

Οι συντελεστές συσχέτισης παίρνουν τιμές από -1 έως 1. Το -1 δείχνει τέλεια αρνητική συσχέτιση (αρνητική κλίση στο γράφημα), ενώ το 1 δείχνει τέλεια θετική συσχέτιση (θετική κλίση στο γράφημα). Το 0 (τυχαία κατανομή ή καμπύλη) δείχνει ότι δεν υπάρχει συσχέτιση.

Στο δικό μας γράφημα βλέπουμε ότι όσοι είχαν χαμηλό αρχικό μισθό, εξακολουθούν να έχουν χαμηλό σημερινό μισθό και όσοι είχαν υψηλό αρχικό μισθό, εξακολουθούν να έχουν υψηλόσημερινό μισθό.

ΣΥΝΤΕΛΕΣΤΗΣ ΣΥΣΧΕΤΙΣΗΣ ΤΟΥ PEARSON

Ο συντελεστής του Pearson είναι συντελεστής γραμμικής συσχέτισης, συμβολίζεται με r και παίρνει τιμές από -1 έως 1 ($-1 \leq r \leq 1$). Χρησιμοποιείται σε ποσοτικές μεταβλητές.

- ✓ -1 έως -0,5 θεωρούμε ότι είναι υψηλός αρνητικός συντελεστής συσχέτισης
- ✓ -0,5 έως -0,2: θεωρούμε ότι είναι χαμηλός αρνητικός συντελεστής συσχέτισης
- ✓ -0,2 έως 0,2: θεωρούμε ότι ο συντελεστής συσχέτισης είναι μηδενικός
- ✓ 0,2 έως 0,5: θεωρούμε ότι είναι χαμηλός θετικός συντελεστής συσχέτισης
- ✓ 0,5 έως 1: θεωρούμε ότι είναι υψηλός θετικός συντελεστής συσχέτισης

* οι χαμηλοί συντελεστές ισχύος εκφράζουν τάση και οι υψηλοί βεβαιότητα

H_0: Δεν υπάρχει συσχέτιση (ο συντελεστής συσχέτισης είναι μηδενικός).
H_1: Υπάρχει συσχέτιση (ο συντελεστής συσχέτισης δεν είναι μηδενικός πληθυσμό).

ΠΡΟΣΟΧΗ! Δεν δηλώνουμε τις υποθέσεις στις συσχετίσεις (είναι για δική μας ευκολία).

Ερώτημα: Να βρεθούν οι συντελεστές συσχέτισης ανάμεσα στον αρχικό και τον τωρινό μισθό.

Εντολή: Analyze → Correlate → Bivariate

Variables: current salary & beginning salary (βάζουμε όλες τις μεταβλητές)
Στο πλαίσιο Correlation Coefficients επιλέγουμε συντελεστές συσχέτισης (στην περίπτωσή μας το ☑Pearson).

Στο Output παίρνουμε τον πίνακα Correlations ή πίνακα συσχετίσεως.
ΠΡΟΣΟΧΗ! Πάντα κοιτάζουμε μόνο το κάτω τρίγωνο που σχηματίζει η διαγώνιος με τις μονάδες.

- μας ενδιαφέρει το Pearson Correlation και δηλώνουμε το r
  στην περίπτωσή μας, r = 0,88 άρα έχουμε υψηλό θετικό συντελεστή συσχέτισης
- με βάση το Significance ελέγχουμε αν επαληθεύεται ή όχι το H_0
  - αν $p > 0,05$ δεν μπορούμε να απορρίψουμε την H_0
  - αν $p < 0,05$ απορρίπτουμε την H_0 και δεχόμαστε την H_1

  Μέσα στα κελιά συχνά εμφανίζονται ένα ή δύο αστεράκια. Ένα αστεράκι * δηλώνει ότι η συσχέτιση είναι στατιστικά σημαντική σε επίπεδο 0,05. Δύο αστεράκια ** δηλώνουν ότι η συσχέτιση είναι στατιστικά σημαντική σε επίπεδο 0,01.

Correlations

		Current Salary	Beginning Salary
Current Salary	Pearson Correlation	1	,880**
	Sig. (2-tailed)		,000
	N	474	474
Beginning Salary	Pearson Correlation	,880**	1
	Sig. (2-tailed)	,000	
	N	474	474

**. Correlation is significant at the 0.01 level (2-tailed).

ΠΡΟΣΟΧΗ! Αυτό που εμείς πρέπει να δηλώνουμε είναι ότι η συσχέτιση (r = 0,88) είναι υψηλή θετική και στατιστικά σημαντική σε επίπεδο σημαντικότητας σε επίπεδο 0,01.

Ερώτημα: Να υπολογιστούν και να σχολιαστούν οι συντελεστές συσχέτισης ανάμεσα στις εξής μεταβλητές: αρχικός μισθός, σημερινός μισθός, προϋπηρεσία, χρόνος υπηρεσίας στην εργασία, επίπεδο εκπαίδευσης και αν κάποιος ανήκει σε μειονότητα (ψευδομεταβλητή-dummy variable).

Στο Output, με βάση τον πίνακα Correlations, για όλα τα κελιά στο κάτω τρίγωνο δηλώνουμε τα εξής:

Correlations

		Beginning Salary	Current Salary	Previous Experience (months)	Months since Hire	Educational Level (years)	Minority Classification
Beginning Salary	Pearson Correlation	1	,880**	,045	-,020	,633**	-,158**
	Sig. (2-tailed)		,000	,327	,668	,000	,001
	N	474	474	474	474	474	474
Current Salary	Pearson Correlation	,880**	1	-,097*	,084	,661**	-,177**
	Sig. (2-tailed)	,000		,034	,067	,000	,000
	N	474	474	474	474	474	474
Previous Experience (months)	Pearson Correlation	,045	-,097*	1	,003	-,252**	,145**
	Sig. (2-tailed)	,327	,034		,948	,000	,002
	N	474	474	474	474	474	474
Months since Hire	Pearson Correlation	-,020	,084	,003	1	,047	,050
	Sig. (2-tailed)	,668	,067	,948		,303	,282
	N	474	474	474	474	474	474
Educational Level (years)	Pearson Correlation	,633**	,661**	-,252**	,047	1	-,133**
	Sig. (2-tailed)	,000	,000	,000	,303		,004
	N	474	474	474	474	474	474
Minority Classification	Pearson Correlation	-,158**	-,177**	,145**	,050	-,133**	1
	Sig. (2-tailed)	,001	,000	,002	,282	,004	
	N	474	474	474	474	474	474

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

1. Ανάμεσα στον τωρινό και τον αρχικό μισθό ο συντελεστής συσχέτισης (r = 0,880**) είναι υψηλός θετικός και στατιστικά σημαντικός σε επίπεδο σημαντικότητας 0,01.
2. Ανάμεσα στην προϋπηρεσία και τον αρχικό μισθό ο συντελεστής συσχέτισης (r = 0,045) είναι χαμηλός θετικός και μη στατιστικά σημαντικός.
3. Ανάμεσα στην προϋπηρεσία και τον τωρινό μισθό ο συντελεστής συσχέτισης (r = -0,097*) είναι μηδενικός και στατιστικά σημαντικός σε επίπεδο σημαντικότητας 0,05.
4. Ανάμεσα στον χρόνο υπηρεσίας στην εργασίακαι τον αρχικό μισθό ο συντελεστής συσχέτισης (r = -0,020) είναι μηδενικός και μη στατιστικά σημαντικός.
5. Ανάμεσα στον χρόνο υπηρεσίας στην εργασίακαι τον τωρινό μισθό ο συντελεστής συσχέτισης (r = 0,084) είναι μηδενικός και μη στατιστικά σημαντικός.
6. Ανάμεσα στον χρόνο υπηρεσίας στην εργασία και την προϋπηρεσία ο συντελεστής συσχέτισης (r = 0,003) είναι μηδενικός και μη στατιστικά σημαντικός.
7. Ανάμεσα στο επίπεδο εκπαίδευσης και τον αρχικό μισθό ο συντελεστής συσχέτισης (r = 0,633**) είναι υψηλός θετικός και στατιστικά σημαντικός σε επίπεδο σημαντικότητας 0,01.

8. Ανάμεσα στο επίπεδο εκπαίδευσης και τον τωρινό μισθό ο συντελεστής συσχέτισης (r = 0,661**) είναι υψηλός θετικός και στατιστικά σημαντικός σε επίπεδο σημαντικότητας 0,01.
9. Ανάμεσα στο επίπεδο εκπαίδευσης και την προϋπηρεσία ο συντελεστής συσχέτισης (r = -0,252**) είναι χαμηλός αρνητικός και στατιστικά σημαντικός σε επίπεδο σημαντικότητας 0,01.
10. Ανάμεσα στο επίπεδο εκπαίδευσης και τον χρόνο υπηρεσίας στην εργασίαο συντελεστής συσχέτισης (r = 0,047) είναι μηδενικός και μη στατιστικά σημαντικός.
11. Ανάμεσα στο αν κάποιος ανήκει σε μειονότητα και τον αρχικό μισθό ο συντελεστής συσχέτισης (r = -0,158**) είναι μηδενικός και στατιστικά σημαντικός σε επίπεδο σημαντικότητας 0,01.
12. Ανάμεσα στο αν κάποιος ανήκει σε μειονότητα και τον τωρινό μισθό ο συντελεστής συσχέτισης (r = -0,177**) είναι μηδενικός και στατιστικά σημαντικός σε επίπεδο σημαντικότητας 0,01.
13. Ανάμεσα στο αν κάποιος ανήκει σε μειονότητα και την προϋπηρεσία ο συντελεστής συσχέτισης (r = 0,145**) είναι μηδενικός και στατιστικά σημαντικός σε επίπεδο σημαντικότητας 0,01.
14. Ανάμεσα στο αν κάποιος ανήκει σε μειονότητα και τον χρόνο υπηρεσίας στην εργασίαο συντελεστής συσχέτισης (r = 0,050) είναι μηδενικός και μη στατιστικά σημαντικός.
15. Ανάμεσα στο αν κάποιος ανήκει σε μειονότητα και το επίπεδο εκπαίδευσης ο συντελεστής συσχέτισης (r = -0,133**) είναι μηδενικός και στατιστικά σημαντικός σε επίπεδο σημαντικότητας 0,01.

ΣΥΝΤΕΛΕΣΤΗΣ ΤΟΥ SPEARMAN

Μη παραμετρικός συντελεστής συσχέτισης. Υπολογίζεται στους διατακτικούς αριθμούς του (rank) και συμβολίζεται με Ps. Αν δεν είναι και οι δύο μεταβλητές ποσοτικές, χρησιμοποιούμε το συντελεστή του Spearman (συσχετίσεις σε διατακτικές μεταβλητές, αλλά και σε ποσοτικές μεταβλητές ή μία και μία).

Ερώτημα: Να βρεθούν οι συντελεστές συσχέτισης ανάμεσα στον αρχικό και τον τωρινό μισθό.

Εντολή: Analyze → Correlate → Bivariate

> Variables: current salary & beginning salary (βάζουμε όλες τις μεταβλητές)
> Στο πλαίσιο Correlation Coefficients επιλέγουμε συντελεστές συσχέτισης (στην περίπτωσή μας το ☑Spearman).

Η ανάλυση είναι ίδια με το συντελεστή του Pearson.

ΠΑΛΙΝΔΡΟΜΗΣΗ (Regression)

Είναι ένα μοντέλο πρόβλεψης. Προβλέψεις για παρατηρήσεις έξω από το μοντέλο δεν είναι ακριβείς, μπορούμε όμως να κάνουμε προβλέψεις για τα ενδιάμεσα σημεία.

Y εξαρτημένη μεταβλητή, X ανεξάτητη. Σημείωση: αν έχουμε πολλές ανεξάρτητες μεταβλητές δεν έχουμε απλή, αλλά πολλαπλή παλινδρόμηση.

Ευθεία ελαχίστων τετραγώνων (least square) ή γραμμή παλινδρόμησης ή γραμμική παλινδρόμηση (linear regression) είναι μία και μοναδική γραμμή, που ελαχιστοποιεί το άθροισμα των τετραγώνων των αποστάσεων των σημείων από την ευθεία

Το σημείο με συντεταγμένες $\bar{x}$, $\bar{y}$ ή αλλιώς κέντρο βάρους (centroid) του νέφους.

$\hat{y}$: εκτιμώμενη τιμή του y ή εκτίμηση ή πρόβλεψη (όσο πιο κοντά είναι στο y, τόσο το καλύτερο)
e: είναι το σφάλμα ή υπόλοιπο (residual) ή κατάλοιπο, δηλαδή η διαφορά της προβλεπόμενης από την παρατηρούμενη τιμή ($e = y - \hat{y}$). Ο μέσος όρος των σφαλμάτων είναι μηδέν.

$y - \bar{y} = (y - \hat{y}) + (\hat{y} - \bar{y})$
οι πραγματικές τιμές των y και οι προβλεπόμενες τιμές των y έχουν τον ίδιο μέσο όρο
↓
$(y - \bar{y})^2 = (y - \hat{y})^2 + (\hat{y} - \bar{y})^2$
υψώνουμε στο τετράγωνο για να παραπέμπει σε διασπορά
↓

$$\sum (y - \bar{y})^2 = \sum (y - \hat{y})^2 + \sum (\hat{y} - \bar{y})^2$$

$\sum (y - \hat{y})^2$ή $\sum e^2$ θέλουμε να είναι μηδέν

συντελεστής προσδιορισμού ή $R^2 = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2}$: παίρνει τιμές από 0 έως 1, θέλουμε να είναι κοντά στο 1, δηλαδή τα σφάλματα να είναι μηδενικά (πάνω από 0,4 θεωρείται μεγάλο, αλλά πάντοτε ανάλογα με την επιστήμη…). Το R^2 εκφράζει πόσο % από τη διασπορά της Y ερμηνεύεται από τη διασπορά του μοντέλου με ανεξάρτητη μεταβλητή την X (ή τις X).

F: αν η παλινδρόμηση είναι καλή, θα πρέπει να μας δίνει ένα μεγάλο F **(το significance είναι μικρότερο από 0,05)**

Ερώτημα: Να υπολογιστεί το μοντέλο παλινδρόμησης με εξαρτημένη μεταβλητή τον σημερινό μισθό και ανεξάρτητη τον αρχικό μισθό. Να γίνει έλεγχος καλής προσαρμογής (goodness of fit) του μοντέλου.

Εντολή: Analyze → Regression → Linear

Dependent: current salary
Independent: beginning salary

Στο Output…

- Στον πίνακα Model Summary δηλώνουμε το R^2 = 0,774 (η ερμηνεία: η X ερμηνεύει την Y κατά 77,4% ή η διασπορά της Y ερμηνεύεται κατά 77,4% από τη διασπορά του μοντέλου). Επιπλέον πληροφορίες: το R^2 είναι ο συντελεστής του Pearson στο τετράγωνο, μόνο όταν έχουμε μία ανεξάρτητη μεταβλητή, ενώ Adjusted R^2 είναι το διορθωμένο R^2 για όλο τον πληθυσμό.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,880[a]	,775	,774	$8,115.356

a. Predictors: (Constant), Beginning Salary

- Στον πίνακα ANOVA μας ενδιαφέρει το F και το significance. Αν το Sig. < 0,05 τότε το F είναι μεγάλο και άρα η γραμμική παλινδρόμηση είναι στατιστικά σημαντική.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1,068E11	1	1,068E11	1622,118	,000[a]
	Residual	3,109E10	472	6,586E7		
	Total	1,379E11	473			

a. Predictors: (Constant), Beginning Salary

b. Dependent Variable: Current Salary

- Το R^2 και το F είναι οι δείκτες καλής προσαρμογής. Όταν R^2 είναι υψηλό και το Sig. ANOVA < 0,05 σημαίνει ότι το μοντέλο μας έχει καλή προσαρμογή.

- Στον πίνακα Coefficients μας ενδιαφέρει η τιμή B, η οποία μας δίνει την εξίσωση της παλινδρόμησης:
  - $y = b_0 + b_1x$
    εξαρτημένη = ConstantB + συντελεστής ανεξάρτητης B * ανεξάρτητη
    Το b_1 εκφράζει τη μεταβολή της Y για μια μονάδα αύξησης της X.
    Στην περίπτωσή μας, y = 1928,206 + 1,909 * x
  - Beta (β) είναι ο συντελεστής της X, αν στο μοντέλο μας δεν χρησιμοποιήσουμε τις αυθεντικές τιμές, αλλά τις τυποποιημένες (standardized). Τα ελέγχουμε σε μοντέλα με πολλές ανεξάρτητες μεταβλητές και χρησιμεύουν στο να ταξινομήσουμε τις ανεξάρτητες μεταβλητές ως προς την ερμηνευτική τους ικανότητα. Δεν λαμβάνουμε υπόψη τα πρόσημα, αλλά τις απόλυτες τιμές.
  - Τα t και significance συνιστούν τον έλεγχο σημαντικότητας της ανεξάρτητης μεταβλητής. Ελέγχει αν το B της ανεξάρτητης μεταβλητής μπορεί ποτέ να είναι μηδέν. Αν το p=Sig. < 0.05 η x είναι στατιστικά σημαντική.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1928,206	888,680		2,170	,031
	Beginning Salary	1,909	,047	,880	40,276	,000

a. Dependent Variable: Current Salary

Ερώτημα 2: Να υπολογιστεί το μοντέλο παλινδρόμησης με εξαρτημένη μεταβλητή τον σημερινό μισθό και ανεξάρτητες τον αρχικό μισθό, τους μήνες εργασίας στην εταιρία και την προηγούμενη προϋπηρεσία. Να γίνει έλεγχος καλής προσαρμογής (goodness of fit) του μοντέλου. Να γίνει ανάλυση των Beta και ανάλυση σημαντικότητας των συντελεστών των ανεξάρτητων μεταβλητών.

Εντολή: Analyze → Regression → Linear

Dependent: current salary
Independent: beginning salary, months since hire, previous experience

Στο Output…

- Στον πίνακα Model Summary δηλώνουμε ότι το R^2 = 0,804, που σημαίνει ότι η διασπορά της Y (εξαρτημένη) ερμηνεύεται κατά 80,4% από τη διασπορά του μοντέλου.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,897[a]	,804	,803	$7,586.187

a. Predictors: (Constant), Previous Experience (months), Months since Hire, Beginning Salary

- Στον πίνακα ANOVA δηλώνουμε ότι το F = 642,151 και το p = 0,000. Επειδή p < 0,05, αυτό σημαίνει ότι το F είναι μεγάλο και άρα η γραμμική παλινδρόμηση είναι στατιστικά σημαντική.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1,109E11	3	3,696E10	642,151	,000[a]
	Residual	2,705E10	470	5,755E7		
	Total	1,379E11	473			

a. Predictors: (Constant), Previous Experience (months), Months since Hire, Beginning Salary

b. Dependent Variable: Current Salary

- Το R^2 και το F είναι οι δείκτες καλής προσαρμογής. Το γεγονός ότι το R^2 είναι υψηλό και το p < 0,05 σημαίνει ότι το μοντέλο μας έχει καλή προσαρμογή.

- Στον πίνακα Coefficients αρχικά μας ενδιαφέρει η τιμή B, η οποία μας δίνει την εξίσωση της παλινδρόμησης: $y = b_0 + b_1x$ | $y = b_0 + b_1x_1 + b_2x_2 + b_3x_3$. Στην περίπτωσή μας: $y = -10266,629 + (1,927 * x_1) + (173,203 * x_2) + (-22,509 * x_3)$. Το επόμενο που μας ενδιαφέρει είναι τα Beta, που χρησιμεύουν στο να ταξινομήσουμε τις ανεξάρτητες μεταβλητές ως προς την ερμηνευτική τους ικανότητα. Έτσι, πρώτος σε ερμηνευτική ικανότητα είναι ο αρχικός μισθός (Beta = 0,888), δεύτερη η προηγούμενη προϋπηρεσία (Beta = -0,138) και τρίτη οι μήνες εργασίας στην εταιρεία (Beta = 0,102). Τέλος, μας ενδιαφέρουν τα t

και το Sig, τα οποία συνιστούν τον έλεγχο σημαντικότητας της ανεξάρτητης μεταβλητής.

- t_1 = 43,435 και p_1 = 0,000 < 0,05 άρα ο συντελεστής της X_1 (beginning salary) είναι στατιστικά σημαντικός
- t_2 = 4,995 και p_2 = 0,000 < 0,05 άρα ο συντελεστής της X_2 (months since hire) είναι στατιστικά σημαντικός
- t_3 = -6,742 και p_3 = 0,000 < 0,05 άρα ο συντελεστής της X_3 (previous experience) είναι στατιστικά σημαντικός

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-10266,629	2959,838		-3,469	,001
	Beginning Salary	1,927	,044	,888	43,435	,000
	Months since Hire	173,203	34,677	,102	4,995	,000
	Previous Experience (months)	-22,509	3,339	-,138	-6,742	,000

a. Dependent Variable: Current Salary

Ερώτημα 3: Ekloges. sav. Να υπολογιστούν οι συντελεστές συσχέτισης ανάμεσα στα ποσοστά των κομμάτων του 2004 και του 2007.

Εντολή: Analyze → Correlate → Bivariate

Στο Variable(s) βάζουμε πρώτα όλα τα ποσοστά του 2004 και μετά όλα τα ποσοστά του 2007. Πατάμε Paste και στο παράθυρο Syntax πέρνουμε γραμμένη την εντολή. Αμέσως μετά το τελευταίο 2004 και πριν το πρώτο 2007 γράφουμε with. Αφού αφήσουμε τον κέρσορα κάπου πάνω στην εντολή, στο μενού πατάμε το πράσινο τριγωνικό κουμπί και στο Output παίρνουμε τον πίνακα Correlations.

Correlations

		p_nd2007	p_pasok2007	p_kke2007	p_syn2007	p_laos2007	p_allo2007
p_nd2004	Pearson Correlation	,931**	-,466**	-,419**	-,441**	,185	-,284*
	Sig. (2-tailed)	,000	,000	,001	,001	,172	,034
	N	56	56	56	56	56	56
p_pasok2004	Pearson Correlation	-,523**	,919**	-,327*	-,144	-,582**	-,354**
	Sig. (2-tailed)	,000	,000	,014	,290	,000	,008
	N	56	56	56	56	56	56
p_kke2004	Pearson Correlation	-,455**	-,400**	,962**	,421**	,127	,449**
	Sig. (2-tailed)	,000	,002	,000	,001	,352	,001
	N	56	56	56	56	56	56
p_syn2004	Pearson Correlation	-,511**	-,297*	,560**	,870**	,248	,685**

	Sig. (2-tailed)	,000	,026	,000	,000	,066	,000
	N	56	56	56	56	56	56
p_laos2004	Pearson Correlation	,019	-,520**	,138	,313*	,938**	,627**
	Sig. (2-tailed)	,889	,000	,309	,019	,000	,000
	N	56	56	56	56	56	56
p_allo2004	Pearson Correlation	-,381**	-,459**	,516**	,746**	,583**	,835**
	Sig. (2-tailed)	,004	,000	,000	,000	,000	,000
	N	56	56	56	56	56	56

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Θα συζητήσουμε τα στοιχεία της διαγωνίου του πίνακα Correlations. Ένας μεγάλος συντελεστής σημαίνει ότι η ΝΔ εκεί που είχε υψηλά ποσοστά το 2004 εξακολουθεί να έχει υψηλά ποσοστά το 2007 κ.ο.κ. – διατηρείται η ισορροπία (εργαλείο γεωγραφίας – συντελεστής χωρικής συσχέτισης).

Συμπεράσματα: τη μεγαλύτερη σταθερότητα έχει το ΚΚΕ, το ΛΑΟΣ, η ΝΔ και λιγότερο το ΠΑΣΟΚ και ο ΣΥΝΑΣΠΙΣΜΟΣ.

Ερώτημα 4: Να δημιουργηθεί το μοντέλο παλινδρόμησης με εξαρτημένη μεταβλητή τη ΝΔ 2007 και ανεξάρτητη τη ΝΔ 2004.

Εντολή: Analyze → Regression → Linear

Dependent: ΝΔ 2007
Independent: ΝΔ 2004

- Στον πίνακα Model Summmary το $R^2 = 0,866$, που σημαίνει ότι η διασπορά της y (εξαρτημένη – ΝΔ 2007) ερμηνεύεται κατά 86,6% από τη διασπορά του μοντέλου (ανεξάρτητη – ΝΔ 2004).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,931[a]	,866	,864	1,73470

a. Predictors: (Constant), p_nd2004

- Με βάση τον πίνακα ANOVA, F = 349,024 και p = 0,000. Επειδή p < 0,05, αυτό σημαίνει ότι το F είναι μεγάλο και άρα η γραμμική παλινδρόμηση είναι στατιστικά σημαντική. Το R^2 και το F είναι οι δείκτες καλής προσαρμογής. Το γεγονός ότι το R^2 είναι υψηλό και το p < 0,05 σημαίνει ότι το μοντέλο μας έχει καλή προσαρμογή.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1050,281	1	1050,281	349,024	,000[a]
	Residual	162,496	54	3,009		
	Total	1212,778	55			

a. Predictors: (Constant), p_nd2004

b. Dependent Variable: p_nd2007

- Στον πίνακα Coefficients αφενός μας ενδιαφέρει η τιμή B, η οποία μας δίνει την εξίσωση της παλινδρόμησης: ΝΔ2007 = -0,772 + 0,954 ΝΔ2004 (το b_1 εκφράζει τη μεταβολή της Y για μια μονάδα αύξησης της X) και αφετέρου το t = 18,682 και το p = 0,000, τα οποία συνιστούν τον έλεγχο σημαντικότητας του συντελεστή της ανεξάρτητης μεταβλητής. Επειδή p < 0,05 η X (ανεξάρτητη – ΝΔ 2004) είναι στατιστικά σημαντική.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-,772	2,412		-,320	,750
	p_nd2004	,954	,051	,931	18,682	,000

a. Dependent Variable: p_nd2007

Ερώτημα 5: Να δημιουργηθεί ένα μοντέλο παλινδρόμησης με εξαρτημένη μεταβλητή τη ΝΔ 2007 και ανεξάρτητες όλα τα κόμματα του 2004.

Εντολή: Analyze → Regression → Linear

Dependent: ΝΔ 2007
Independent: ΝΔ 2004, ΠΑΣΟΚ 2004, ΚΚΕ 2004, ΣΥΡΙΖΑ 2004, ΛΑΟΣ 2004, ΑΛΛΟ 2004

- Tolerance είναι ένας δείκτης που δείχνει πότε έχουμε πρόβλημα πολυσυγγραμικότητας (multicolinearity) – η μεγάλη συσχέτιση μεταξύ των ανεξάρτητων μεταβλητών δημιουργεί πρόβλημα. Υπάρχει ένα όριο… Λύσεις: μη συμπερίληψη όλων των μεταβλητών (οικονομία) ή factor analysis. Μια μεταβλητή δεν μπαίνει στο μοντέλο (ΠΑΣΟΚ 2004).

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	p_allo2004, p_nd2004, p_kke2004, p_laos2004, p_syn2004[a]	.	Enter

a. Tolerance = ,000 limits reached.

b. Dependent Variable: p_nd2007

- Στον πίνακα Model Summmary το R^2 = 0,911, που σημαίνει ότι η διασπορά της Υ (εξαρτημένη – ΝΔ 2007) ερμηνεύεται κατά 91,1% από τη διασπορά του μοντέλου.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,954[a]	,911	,902	1,47074

a. Predictors: (Constant), p_allo2004, p_nd2004, p_kke2004, p_laos2004, p_syn2004

- Με βάση τον πίνακα ANOVA, F = 102,135 και p = 0,000. Επειδή p < 0,05, αυτό σημαίνει ότι το F είναι μεγάλο και άρα η γραμμική παλινδρόμηση είναι στατιστικά σημαντική. Το R^2 και το F είναι οι δείκτες καλής προσαρμογής. Το γεγονός ότι το R^2 είναι υψηλό και το p < 0,05 σημαίνει ότι το μοντέλο μας έχει καλή προσαρμογή.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1104,624	5	220,925	102,135	,000[a]
	Residual	108,154	50	2,163		
	Total	1212,778	55			

a. Predictors: (Constant), p_allo2004, p_nd2004, p_kke2004, p_laos2004, p_syn2004

b. Dependent Variable: p_nd2007

- Στον πίνακα Coefficients αφενός μας ενδιαφέρουν: η τιμές B, που μας δίνουν την εξίσωση της παλινδρόμησης: ΝΔ 2007 = -0,073 + 0,997 ΝΔ 2004 – 0,047 ΚΚΕ 2004 + 0,760 ΣΥΝΑΣΠΙΣΜΟΣ 2004 – 0,469 ΛΑΟΣ 2004 – 1,575 ΑΛΛΟ 2004 (κάθε b δηλώνει τη μεταβολή που επέρχεται στην Y για κάθε μονάδα αύξησης της αντίστοιχης X)
  Το επόμενο που μας ενδιαφέρει είναι τα Beta, που χρησιμεύουν στο να ταξινομήσουμε τις ανεξάρτητες μεταβλητές ως προς την ερμηνευτική τους ικανότητα. Έτσι, πρώτο σε ερμηνευτική ικανότητα είναι η ΝΔ 2004 (Beta = 0,972), δεύτερο το ΑΛΛΟ 2004 (Beta = -0,207), τρίτο Ο ΣΥΝΑΣΠΙΣΜΟΣ 2004 (Beta = 0,162), τέταρτο ο ΛΑΟΣ 2004 (Beta = -0,095) και πέμπτο το ΚΚΕ 2004 (Beta = -0,025). Τέλος, μας ενδιαφέρουν τα t και το Sig, τα οποία συνιστούν τον έλεγχο σημαντικότητας της ανεξάρτητης μεταβλητής.
  - t_1 = 18,138 και p_1 = 0,000 < 0,05 άρα ο συντελεστής της X_1 (ΝΔ 2004) είναι στατιστικά σημαντικός
  - t_2 = -0,497 και p_2 = 0,621 > 0,05 άρα ο συντελεστής της X_2 (ΚΚΕ 2004) είναι μη στατιστικά σημαντικός
  - t_3 = 2,303 και p_3 = 0,025 < 0,05 άρα ο συντελεστής της X_3 (ΣΥΝΑΣΠΙΣΜΟΣ 2004) είναι στατιστικά σημαντικός
  - t_4 = -1,842 και p_4 = 0,071 > 0,05 άρα ο συντελεστής της X_4 (ΛΑΟΣ 2004) είναι μη στατιστικά σημαντικός
  - t_5 = -2,932 και p_5 = 0,005 < 0,05 άρα ο συνετελστής της X_5 (ΑΛΛΟ 2004) είναι στατιστικά σημαντικός

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-,073	2,946		-,025	,980
	p_nd2004	,997	,055	,972	18,138	,000
	p_kke2004	-,047	,095	-,025	-,497	,621
	p_syn2004	,760	,330	,162	2,303	,025
	p_laos2004	-,469	,255	-,095	-1,842	,071
	p_allo2004	-1,575	,537	-,207	-2,932	,005

a. Dependent Variable: p_nd2007

ΣΗΜΕΙΩΣΗ: Για να λύσουμε το πρόβλημα της πολυσυγγραμικότητας μπορούμε να επαναλάβουμε τη διαδικασία χωρίς το ΠΑΣΟΚ και χωρίς τις μεταβλητές που δεν είναι στατιστικά σημαντικές (ΚΚΕ, ΛΑΟΣ). Αλλά θα προτιμήσουμε μια διαφορετική μέθοδο (stepwise).

Η μέθοδος stepwise

Ερώτημα: Να ξαναγίνει το αρχικό μοντέλο χρησιμοποιώντας τη μέθοδο stepwise.

Εντολή: Analyze → Regression → Linear

Dependent: ΝΔ 2007
Independent: ΝΔ 2004, ΠΑΣΟΚ 2004, ΚΚΕ 2004, ΣΥΡΙΖΑ 2004, ΛΑΟΣ 2004, ΑΛΛΟ 2004
Method: stepwise

* Σε όλους τους πίνακες κοιτάμε πάντα το τελευταίο βήμα (στην περίπτωσή μας τη γραμμή 4).

- Στον πίνακα Model Summmary το $R^2 = 0,910$, που σημαίνει ότι η διασπορά της y (εξαρτημένη – ΝΔ 2007) ερμηνεύεται κατά 91% από τη διασπορά του μοντέλου.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,931[a]	,866	,864	1,73470
2	,945[b]	,893	,889	1,56111
3	,949[c]	,901	,896	1,51713
4	,954[d]	,910	,903	1,45984

a. Predictors: (Constant), p_nd2004

b. Predictors: (Constant), p_nd2004, p_laos2004

c. Predictors: (Constant), p_nd2004, p_laos2004, p_allo2004

d. Predictors: (Constant), p_nd2004, p_laos2004, p_allo2004, p_syn2004

- Με βάση τον πίνακα ANOVA, F = 129,518 και p = 0,000. Επειδή p < 0,05, αυτό σημαίνει ότι το F είναι μεγάλο και άρα η γραμμική παλινδρόμηση είναι

στατιστικά σημαντική. Το R^2 και το F είναι οι δείκτες καλής προσαρμογής. Το γεγονός ότι το R^2 είναι υψηλό και το $p < 0,05$ σημαίνει ότι το μοντέλο μας έχει καλή προσαρμογή.

ANOVA[e]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1050,281	1	1050,281	349,024	,000[a]
	Residual	162,496	54	3,009		
	Total	1212,778	55			
2	Regression	1083,613	2	541,807	222,319	,000[b]
	Residual	129,165	53	2,437		
	Total	1212,778	55			
3	Regression	1093,090	3	364,363	158,302	,000[c]
	Residual	119,688	52	2,302		
	Total	1212,778	55			
4	Regression	1104,089	4	276,022	129,518	,000[d]
	Residual	108,688	51	2,131		
	Total	1212,778	55			

a. Predictors: (Constant), p_nd2004

b. Predictors: (Constant), p_nd2004, p_laos2004

c. Predictors: (Constant), p_nd2004, p_laos2004, p_allo2004

d. Predictors: (Constant), p_nd2004, p_laos2004, p_allo2004, p_syn2004

e. Dependent Variable: p_nd2007

- Στον πίνακα Coefficients αφενός μας ενδιαφέρουν: η τιμές B, που μας δίνουν την εξίσωση της παλινδρόμησης: ΝΔ 2007 = -0,474 + 1,003 ΝΔ 2004 – 0,464 ΛΑΟΣ 2004 – 1,622 ΑΛΛΟΣ 2004 + 0,736 ΣΥΝΑΣΠΙΣΜΟΣ 2004 (κάθε b δηλώνει τη μεταβολή που επέρχεται στην y για κάθε μονάδα αύξησης της αντίστοιχης x)
  Το επόμενο που μας ενδιαφέρει είναι τα Beta, που χρησιμεύουν στο να ταξινομήσουμε τις ανεξάρτητες μεταβλητές ως προς την ερμηνευτική τους ικανότητα. Έτσι, πρώτο σε ερμηνευτική ικανότητα είναι η ΝΔ 2004 (Beta = 0,979), δεύτερο το ΑΛΛΟ 2004 (Beta = -0,213), τρίτο Ο ΣΥΝΑΣΠΙΣΜΟΣ 2004 (Beta = 0,156) και τέταρτο ο ΛΑΟΣ 2004 (Beta = -0,094). Τέλος, μας ενδιαφέρουν τα t και το Sig, τα οποία συνιστούν τον έλεγχο σημαντικότητας της ανεξάρτητης μεταβλητής.
  - t_1 = 18,989 και p_1 = 0,000 < 0,05 άρα ο συντελεστής της X_1 (ΝΔ 2004) είναι στατιστικά σημαντικός
  - t_2 = -1,838 και p_2 = 0,072 > 0,05 άρα ο συντελεστής της X_2 (ΛΑΟΣ 2004) είναι μη στατιστικά σημαντικός
  - t_3 = -3,089 και p_3 = 0,003 < 0,05 άρα ο συντελεστής της X_3 (ΑΛΛΟΣ 2004) είναι στατιστικά σημαντικός
  - t_4 = 2,272 και p_4 = 0,027 < 0,05 άρα ο συντελεστής της X_4 (ΛΑΟΣ 2004) είναι στατιστικά σημαντικός

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-,772	2,412		-,320	,750
	p_nd2004	,954	,051	,931	18,682	,000
2	(Constant)	-,940	2,171		-,433	,667
	p_nd2004	,988	,047	,964	21,082	,000
	p_laos2004	-,834	,226	-,169	-3,698	,001
3	(Constant)	2,274	2,638		,862	,393
	p_nd2004	,950	,049	,927	19,320	,000
	p_laos2004	-,554	,259	-,112	-2,138	,037
	p_allo2004	-,819	,404	-,108	-2,029	,048
4	(Constant)	-,474	2,812		-,169	,867
	p_nd2004	1,003	,053	,979	18,989	,000
	p_laos2004	-,464	,252	-,094	-1,838	,072
	p_allo2004	-1,622	,525	-,213	-3,089	,003
	p_syn2004	,736	,324	,156	2,272	,027

a. Dependent Variable: p_nd2007

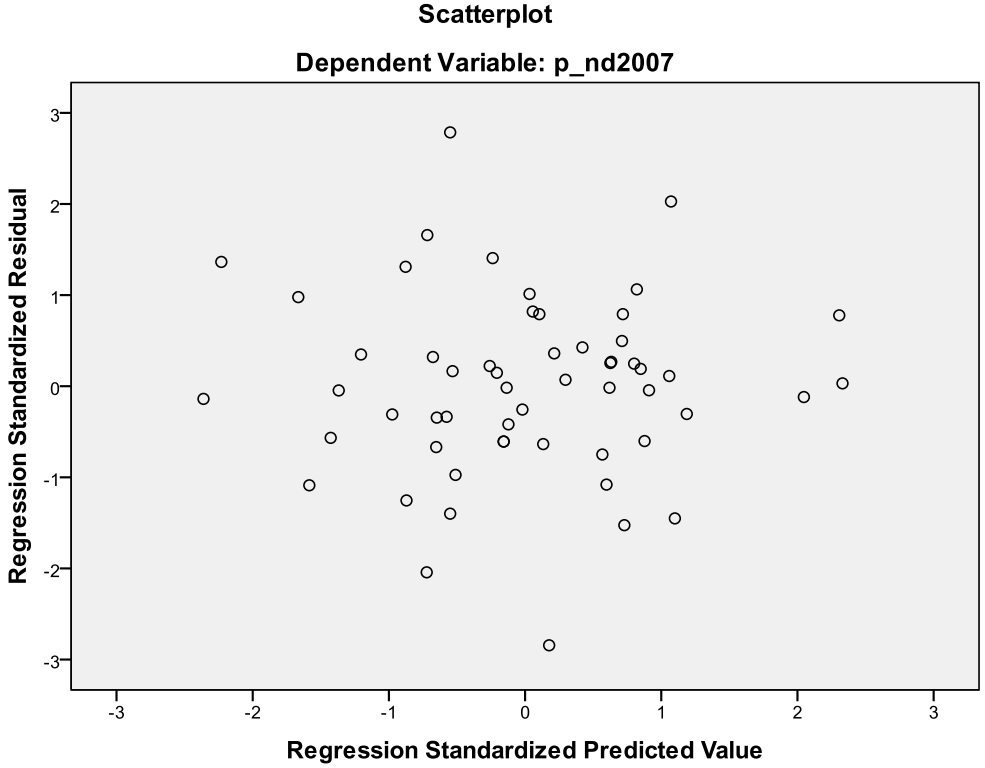
<u>Προϋποθέσεις παλινδρόμησης</u>:

1. **γραμμικότητα (linearity): ο μέσος της y για τα διάφορα επίπεδα της x είναι γραμμική συνάρτηση της X**
2. **ομοσκεδαστικότητα (homoscedasticity): ίση διασπορά των σφαλμάτων**
3. **ανεξαρτησία (independence): οι τιμές της Y που αντιστοιχούν στα διάφορα επίπεδα της X είναι ανεξάρτητες μεταξύ τους**
4. **κανονικότητα (normality): η κατανομή της Y για όλα τα επίπεδα της X είναι κανονική**

<u>Εντολή</u>: Analyze → Regression → Linear

Independent: ΝΔ 2007
Dependent: ΝΔ 2004
Στο κουμπί <u>Save</u> και στην υποκατηγορία Predicted values και στην υποκατηγορία Residuals επιλέγουμε τα ⊙ Unstandardized (2 νέες στήλες στα δεδομένα…)
Στην επιλογή <u>Plots</u>: Y = ZRESID και X = ZPRED (για να ελέγξουμε την γραμμικότητα και την ομοσκεδαστικότητα)

Με βάση το Scatterplot ελέγχουμε τη γραμμικότητα και την ομοσκεδαστικότητα. Αν το σχήμα είναι ασαφές, έχουμε γραμμικότητα. Αν παντού έχει το ίδιο πάχος (ή πλάτος), έχουμε ομοσκεδαστικότητα. Στο δικό μας σχήμα έχουμε και γραμμικότητα και ομοσκεδαστικότητα.



Ακραίες τιμές: είτε ως outliers είτε ως influential points (για τιμές πάνω από 3 ή κάτω από -3).

Επίσης, ελέγχουμε την κανονικότητα στα Residuals.

Εντολή: Analyze → Descriptive statistics → Explore

Dependent list: Residuals
Στο Plots επιλέγουμε το ☑ Normality plots with tests

Στον πίνακα Tests of Normality δηλώνουμε το p = 0,200 στο Kolmogorov-Smirnov test, δηλαδή p > 0,05 άρα έχουμε κανονικότητα.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Unstandardized Residual	,090	56	,200*	,985	56	,712

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.