Μεθοδολογία Κοινωνικών Επιστήμων

Ανάλυση ποσοτικών δεδομένων Ι

ΠΜΣ Πολιτικής Επιστήμης και Κοινωνιολογίας

ΕΚΠΑ/Τμήμα Πολιτικής Επιστήμης & Δημόσιας Διοίκησης

- Εξάμηνο: Χειμερινό 2024-25
- Μάνος Τσατσάνης (Επίκουρος Καθηγητής)
- etsats@pspa.uoa.gr

Βασικές προσεγγίσεις στην ποσοτική ανάλυση

Περιγραφική στατιστική (descriptive statistics)

- Στατιστική ανάλυση, με σκοπό την περιγραφή και την περίληψη δειγματοληπτικών δεδομένων.

Συμπερασματική ή επαγωγική στατιστική (inferential statistics)

- **Στατιστική, η οποία χρησιμοποιείται προκειμένου να**
  - εξαχθούν συμπεράσματα και γενικεύσεις από το δείγμα για τον πληθυσμό έρευνας
  - ελεγχθούν ερευνητικές υποθέσεις

Μονομεταβλητή ανάλυση
(ανάλυση μίας μεταβλητής κάθε φορά)

- Πίνακες συχνοτήτων
  - Αριθμός ανθρώπων ή περιπτώσεων σε κάθε κατηγορία
  - Συχνά εκφράζεται και ως ποσοστό του δείγματος
  - Τα δεδομένα των μεταβλητών διαστήματος/αναλογίας χρειάζεται να ομαδοποιηθούν
- Διαγράμματα
  - Μπορεί να είναι ραβδογράμματα, θηκογράμματα ή κυκλικά διαγράμματα (μεταβλητές διάταξης και ονομαστικές μεταβλητές)
  - Ιστογράμματα (μεταβλητές διαστήματος/αναλογίας)

Κατανομή συχνότητας

Πίνακες Συχνοτήτων (Frequency tables)

Πολιτική αυτοτοποθέτηση

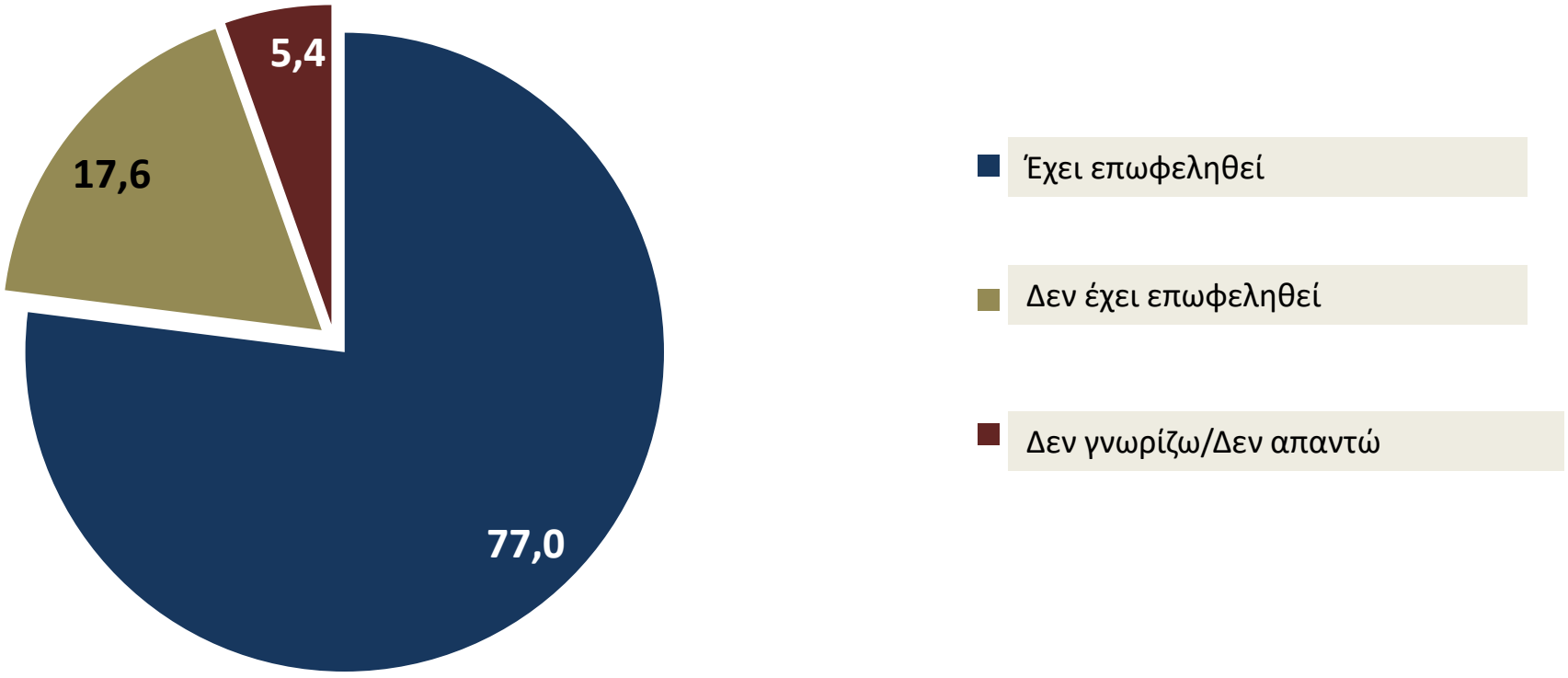
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Αριστεροί	211	9,5	9,5	9,5
	Κεντροαριστεροί	378	17,1	17,1	26,6
	Κεντρώοι	666	30,1	30,1	56,7
	Κεντροδεξιοί	307	13,9	13,9	70,6
	Δεξιοί	366	16,6	16,6	87,1
	Τίποτα από τα παραπάνω δεν τους εκφράζει	216	9,8	9,8	96,9
	ΔΓ/ΔΑ	69	3,1	3,1	100,0
	Total	2213	100,0	100,0	

Το κράτος πρέπει να διαχωριστεί πλήρως από την εκκλησία

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Συμφωνώ	1078	48,7	48,7	48,7
	Ούτε ούτε (αυθ)	274	12,4	12,4	61,1
	Διαφωνώ	760	34,4	34,4	95,4
	ΔΓ/ΔΑ	101	4,6	4,6	100,0
	Total	2213	100,0	100,0	

- Τρόπος παρουσίασης του αριθμού εμφανίσεων κάθε κατηγορίας μιας μεταβλητής σε ένα δείγμα.
- Ονομαστικές και διατακτικές μεταβλητές.

Παράδειγμα: "Λαμβάνοντας τα πάντα υπόψη σας, θα λέγατε ότι η Ελλάδα έχει γενικά επωφεληθεί ή όχι από το γεγονός ότι είναι μέλος της Ευρωπαϊκής Ένωσης;" (%)



Tsirbas, Y., & Sotiropoulos, D.A. (2015) 'What Do Greek Political Elites Think about Europe and the Crisis? An Exploratory Analysis' *L.S.E. Hellenic Observatory Papers on Greece and Southeast Europe, GreeSE paper No. 93*

Μέτρα Κεντρικής Τάσης

- Αριθμητικός μέσος (Mean)
- Διάμεσος (Median)
- Επικρατούσα τιμή (Mode)
- Αριθμητικός μέσος (μέσος όρος): το άθροισμα όλων των τιμών μιας ομάδας, διαιρεμένο με τον αριθμό αυτών των τιμών.
  - $\overline{X}$ ή Μ
  - Πολύ ευαίσθητος σε ακραίες τιμές.
  - Αναλογικές μεταβλητές (και ίσων διαστημάτων).

Διάμεσος

- Κατάλληλο και για διατακτικές μεταβλητές.
- Το μεσαίο σημείο από ένα σετ τιμών. Εκατέρωθέν του βρίσκεται το 50% των τιμών.
- Αν είναι ζυγός ο αριθμός των τιμών τότε είναι ο μέσος όρος των δύο μεσαίων τιμών.
- Προτιμάται όταν υπάρχουν ακραίες τιμές γιατί δεν επηρεάζεται από αυτές.

Επικρατούσα τιμή

- Επικρατούσα τιμή (mode): Η τιμή που εμφανίζεται συχνότερα.
- Το πιο γενικό και λιγότερο ακριβές, αλλά είναι σημαντικό για την κατανόηση των χαρακτηριστικών ειδικών ομάδων τιμών.
- Συχνότερο λάθος: η αναφορά του αριθμού εμφανίσεων και όχι της τιμής που αντιπροσωπεύει.
- **Bimodal:** κατανομή με δύο επικρατούσες τιμές.
- Μπορεί να υπολογιστεί για κάθε μεταβλητή, καθώς οι διατακτικές μεταβλητές και οι μεταβλητές ίσων διαστημάτων είναι και ονομαστικές.

Πότε χρησιμοποιούμε τι;

- Εξαρτάται από το τι θέλουμε να περιγράψουμε.
- Ποιοτικές, ονομαστικές, μεταβλητές μπορούν να περιγραφούν μόνο με το mode. Οι διατακτικές με mode και διάμεσο.
- Μ.Ο. και διάμεσος χρησιμοποιούνται για ποσοτικές μεταβλητές. Ο μέσος όρος είναι ο πιο συνήθης τρόπος να συνοψίσουμε την κεντρική τάση μίας μεταβλητής

Μέτρα Διασποράς

- Προσέγγιση του κατά πόσο τα δεδομένα είναι συγκεντρωμένα γύρω από μια κεντρική τιμή ή διάσπαρτα.
- Αποτελούν και ένδειξη του κατά πόσο τα μέτρα κεντρικής τάσης περιγράφουν ικανοποιητικά ένα δείγμα ή έναν πληθυσμό.
- Αν τα δεδομένα είναι πολύ διάσπαρτα, τότε είναι απαραίτητη η αναφορά των μέτρων διασποράς.

Μέτρα Διασποράς

– Π.χ. αν οι φοιτητές/τριες ενός συγκεκριμένου έτους φοίτησης έχουν μέσο όρο βαθμολογίας 7,5 στα μαθήματα που έχουν περάσει, αυτό μπορεί να σημαίνει ότι έχουν όλοι μέσο όρο 7,5 σε όλα τα μαθήματα (μηδενική διασπορά), αλλά μπορεί να σημαίνει ότι υπάρχουν πολλοί με Μ.Ο. 5 και πολλοί με Μ.Ο. 10 (μεγάλη διασπορά).

											Μέσος όρος	Διασπορά
7,5	7,5	7,5	7,5	7,5	7,5	7,5	7,5	7,5	7,5	7,5	7,5	Μηδενική
10	5	10	9	6.5	5	7	8	7	5	10	7,5	Μεγάλη

Διακύμανση/Διασπορά

<u>Διακύμανση/διασπορά (Variance)</u>: Προσθέτουμε τα τετράγωνα των διαφορών από τον μ.ό. και διαιρούμε με τον αριθμό των περιπτώσεων. Σημ. στην περίπτωση που αναζητούμε τη διασπορά σε ένα δείγμα και όχι σε έναν πληθυσμό, διαιρούμε με n-1 (αριθμό περιπτώσεων μειωμένο κατά 1).

Μέσος όρος δείγματος = 7			
Φοιτ	Μ.Ο.	Διαφορά από μ.ό.	Τετράγωνο διαφοράς
1	5	-2	4
2	6	-1	1
3	6	-1	1
4	7	0	0
5	8	1	1
6	10	3	9
Άθροισμα τετραγώνων			16
Διακύμανση			16/5=3,2

-Επειδή πήραμε τα τετράγωνα των διαφορών, αυτό σημαίνει ότι μέσω της διακύμανσης δώσαμε μεγαλύτερο βάρος στις μεγάλες διαφορές από τον μ.ο.

-Η μονάδα μέτρησης είναι διαφορετική από την αρχική.

Τυπική απόκλιση

Για να λύσουμε το παραπάνω πρόβλημα υπολογίζουμε την **Τυπική Απόκλιση** (standard deviation, s), ως την τετραγωνική ρίζα της διακύμανσης.

$$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$$

Στο παραπάνω παράδειγμα, η τυπική απόκλιση από τον μέσο στο δείγμα είναι 1,79 (τετραγωνική ρίζα του 3,2)

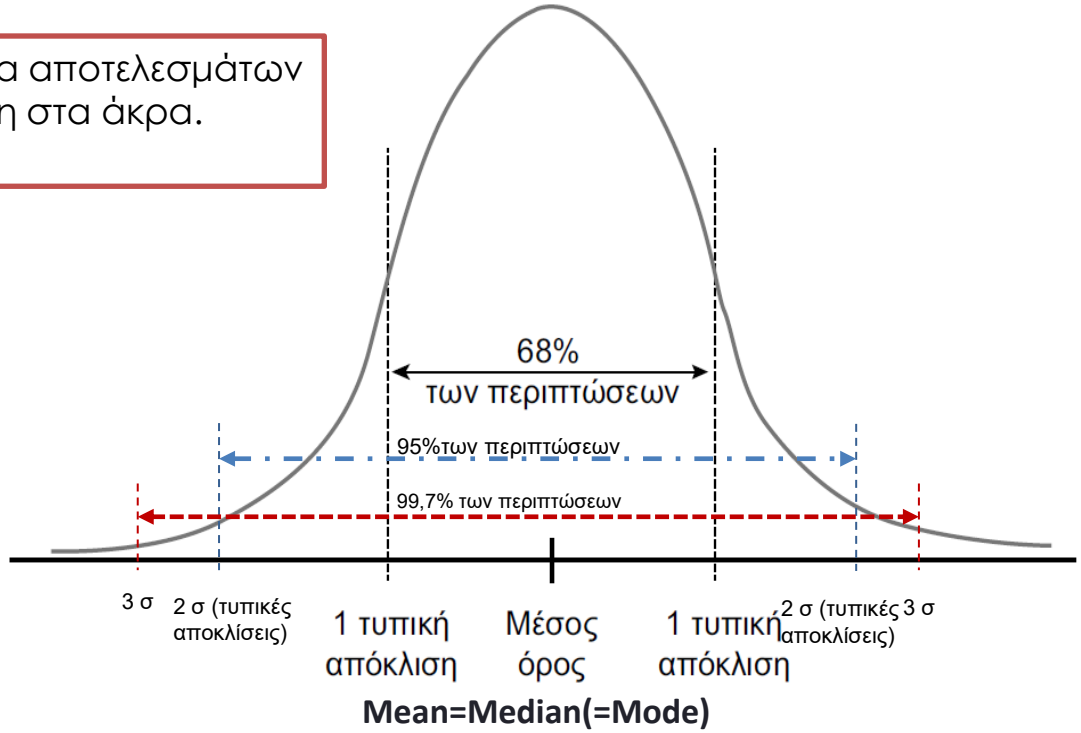
Η τυπική απόκλιση μετριέται στην ίδια μονάδα μέτρησης όπως και τα δεδομένα μας και αυτό είναι πλεονέκτημα.

Κανονική κατανομή

- Μια κατανομή, η οποία έχει συμμετρική μορφή και ομοιάζει με καμπάνα
- Σε μια κανονική κατανομή το 68% των περιπτώσεων θα ενέπιπτε στο διάστημα μεταξύ μιας τυπικής απόκλισης άνω του μέσου όρου (το κέντρο της κατανομής) και μιας τυπικής απόκλισης κάτω του μέσου όρου. Και περίπου το 95% στο διάστημα +-2 τυπικών αποκλίσεων.
- Τα βασικά στατιστικά είναι παραμετρικά, βασίζονται δηλαδή σε παραδοχές που έχουν να κάνουν με παραμέτρους της κατανομής του πληθυσμού έρευνας. Η βασικότερη των παραδοχών αυτών αφορά την κανονική κατανομή.
- Η μέση τιμή μεγάλου αριθμού ανεξάρτητων παρατηρήσεων πλησιάζει τη μέση τιμή του πραγματικού πληθυσμού (Κεντρικό Οριακό Θεώρημα - Central limit theorem)

Κανονική κατανομή (επίσης: Γκαουσιανή)

Μεγαλύτερη συχνότητα αποτελεσμάτων στο μέσο και μικρότερη στα άκρα. Συμμετρία.



Γράφημα 15.1 Κανονική κατανομή

Adler & Clark, 2018, σ. 569

Διαστήματα Εμπιστοσύνης

ΜΕΓΕΘΟΣ ΔΕΙΓΜΑΤΟΣ	ΑΠΟΤΕΛΕΣΜΑ (π.χ. % υποστήριξης θανατικής ποινής)	ΔΕΙΓΜΑΤΟΛΗΠΤΙΙΚΟ ΣΦΑΛΜΑ (+/-)	ΔΙΑΣΤΗΜΑ ΕΜΠΙΣΤΟΣΥΝΗΣ	ΕΥΡΟΣ ΤΙΜΩΝ ΜΕΣΑ ΣΤΟ ΟΠΟΙΟ ΕΙΜΑΣΤΕ ΚΑΤΑ 95% ΣΙΓΟΥΡΟΙ ΌΤΙ ΒΡΙΣΚΟΝΤΑΙ ΤΑ ΑΠΟΤΕΛΕΣΜΑΤΑ ΣΤΟΝ ΠΡΑΓΜΑΤΙΚΟ ΠΛΗΘΥΣΜΟ
100	52%	10%	20%	42-62
204	52%	7%	14%	45-59
400	52%	5%	10%	47-57
1100	52%	3%	6%	49-55
10.000	52%	1%	2%	51-53

Είμαστε κατά 95% σίγουροι ότι το αποτέλεσμα στον πραγματικό πληθυσμό θα είναι εντός του εύρους τιμών που ορίζεται από το διάστημα εμπιστοσύνης.

Υπολογισμός 95% διαστήματος εμπιστοσύνης για ποσοστά:

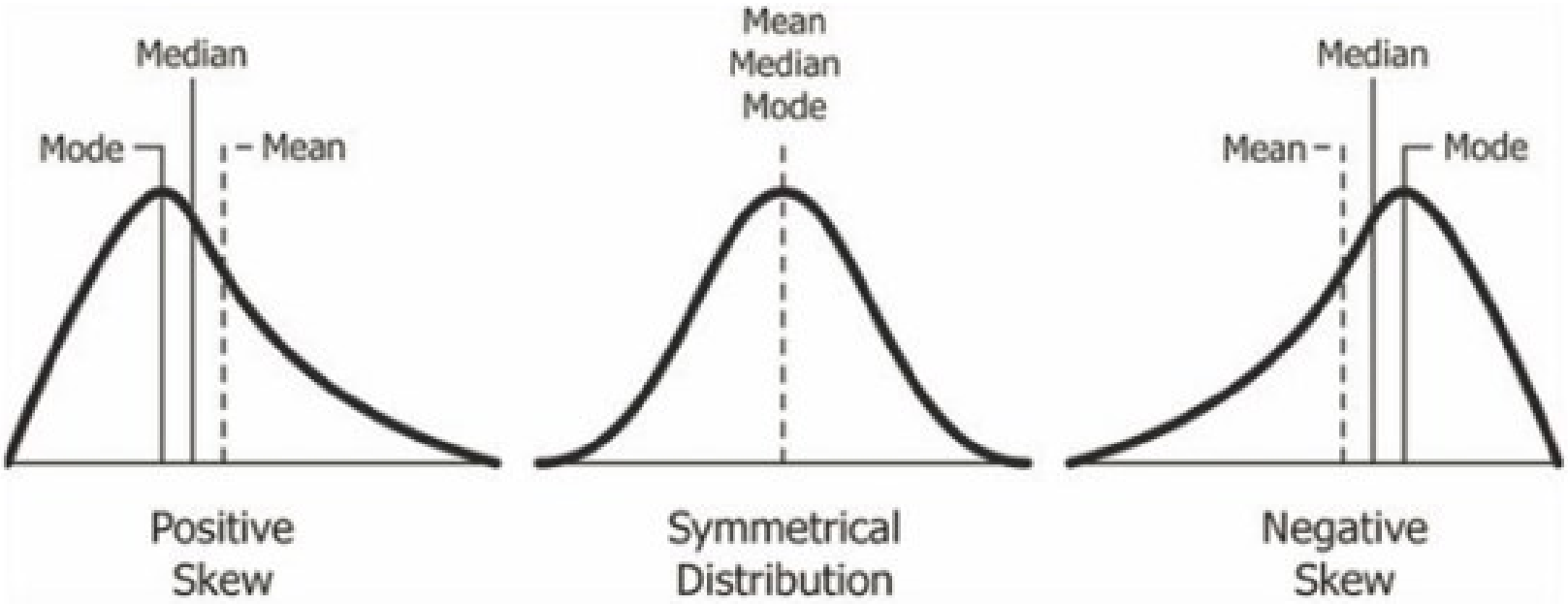
Άνω όριο CI = $\hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, κάτω όριο CI = $\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Υπολογισμός 95% διαστήματος εμπιστοσύνης για μέσους όρους:

Άνω όριο CI = $\bar{x} + 1.96\frac{s}{\sqrt{n}}$, κάτω όριο CI = $\bar{x} - 1.96\frac{s}{\sqrt{n}}$

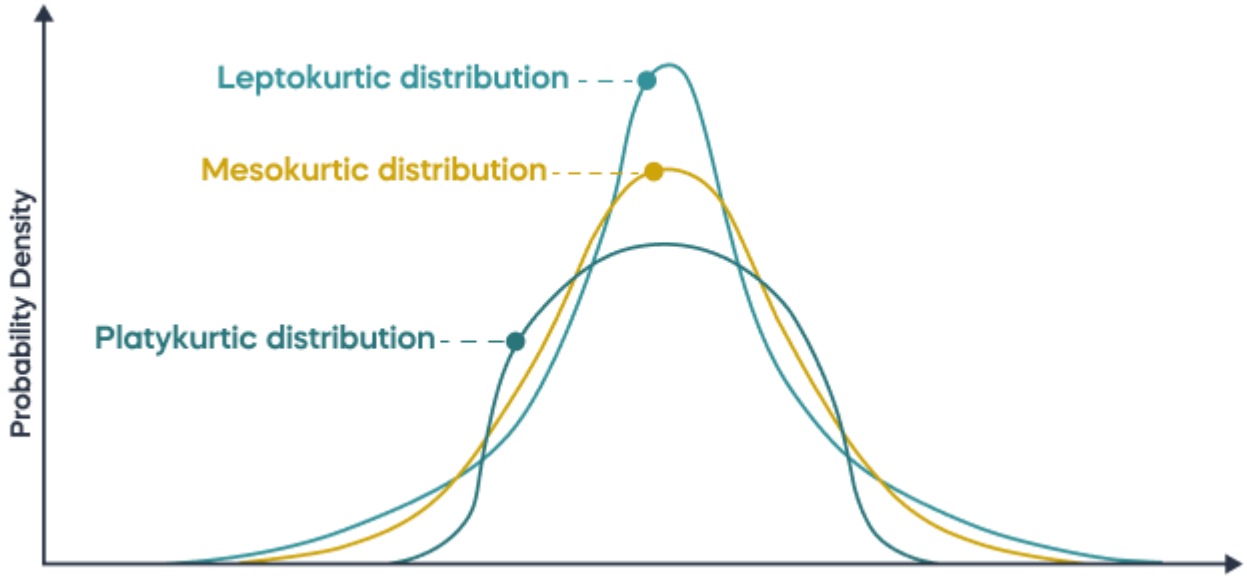
Skewness (στρέβλωση/ασυμμετρία)

Στρέβλωση= η έλλειψη συμμετρίας μιας κατανομής, οριζόντια (η μια «άκρη» είναι μακρύτερη από την άλλη).



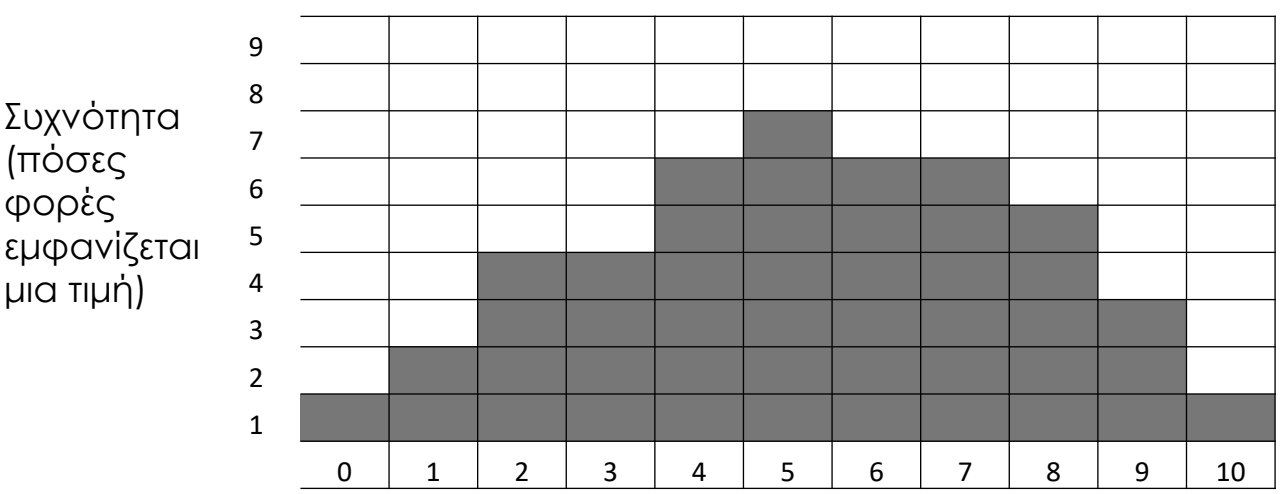
Kurtosis (κυρτότητα)

Κύρτωση= η έλλειψη συμμετρίας μιας κατανομής, κάθετα. Παρόλο που διάμεσος=μέσος, όταν έχουμε πολύ λεπτόκυρτες κατανομές έχουμε περισσότερες ακραίες τιμές σε σχέση με την κανονική κατανομή (μεσόκυρτη), όταν έχουμε πλατύκυρτες, παρατηρούμε λιγότερες ακραίες τιμές σε σχέση με την κανονική κατανομή



Ιστόγραμμα…

Ο ευκολότερος τρόπος να ελέγξουμε αν μια κατανομή είναι κανονική, είναι «με το μάτι», δηλαδή κατασκευάζοντας ένα γράφημα. Υπάρχουν, προφανώς, και στατιστικοί τρόποι.
Ιστόγραμμα συχνότητας (στο ιστόγραμμα πυκνότητας, κάθε μπάρα αντιπροσωπεύει το ποσοστό της συγκεκριμένης κατηγορίας).



Βαθμός εμπιστοσύνης στις τηλεοπτικές ειδήσεις, όπου 0 σημαίνει καμία εμπιστοσύνη και 10 απόλυτη εμπιστοσύνη (υποθετικό παράδειγμα).

Μπορείτε να υπολογίσετε:
-Το μέγεθος του δείγματος
-Τον μέσο όρο
-Τη διάμεσο
-Το ποσοστό όσων έχουν βαθμό εμπιστοσύνης 8 και άνω;

Διμεταβλητή ανάλυση
(ανάλυση δύο μεταβλητών κάθε φορά)

- Διερευνά τις σχέσεις μεταξύ δύο μεταβλητών
- Αναζητά τη σύμπτωση της διακύμανσης μιας μεταβλητής με αυτή μιας άλλης και τη συσχέτιση
- Δε μπορεί να θεμελιώσει αιτιώδη σύνδεση
- Πίνακες συνάφειας
  - Συνδέει τις συχνότητες δύο μεταβλητών
  - Βοηθά στον εντοπισμό οποιωνδήποτε μορφών διασύνδεσης

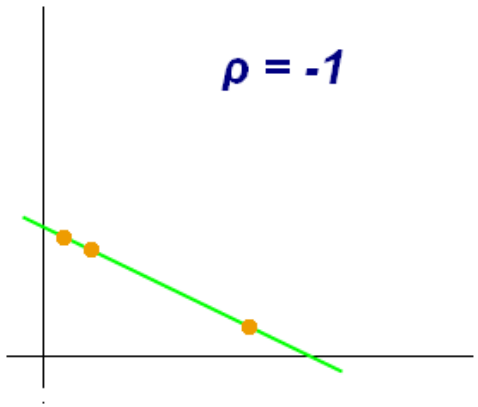
Ο συντελεστής συσχέτισης Pearson's *r* : *η σχέση μεταξύ δύο μεταβλητών διαστήματος/αναλογίας*

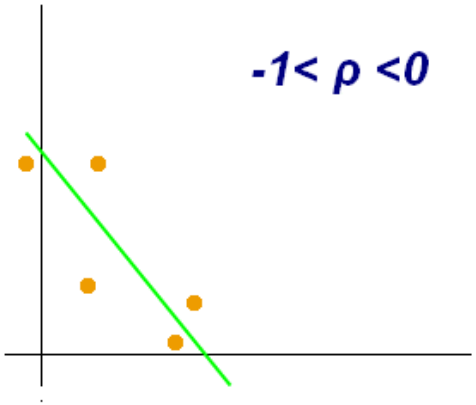
- Ο συντελεστής δείχνει την ισχύ και την κατεύθυνση της σχέσης
- Βρίσκεται μεταξύ-1 (τέλεια αρνητική συσχέτιση) και +1 (τέλεια θετική συσχέτιση)

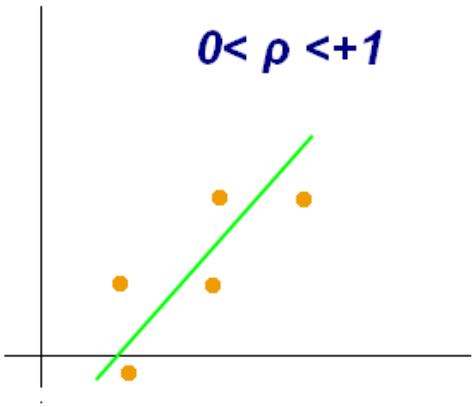
$$r = \frac{cov_{xy}}{s_x s_y} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

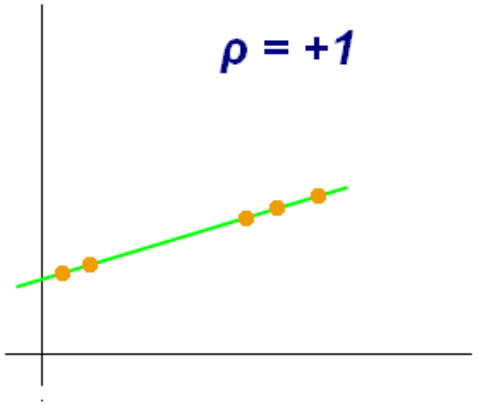
- Συντελεστής προσδιορισμού **r^2**
  - Υπολογίζεται αν υψώσουμε την τιμή του r στο τετράγωνο
  - Εκφράζει το μέρος της διακύμανσης της μίας μεταβλητής που οφείλεται στην άλλη

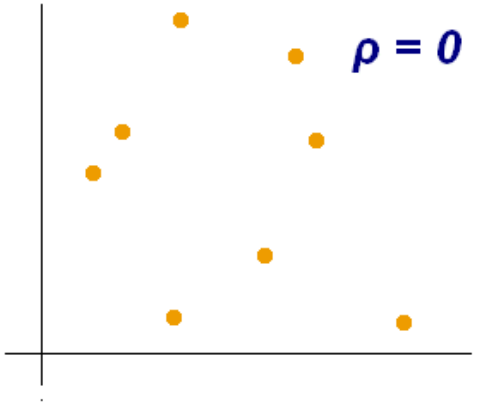
Ο συντελεστής συσχέτισης Pearson's *r* : *η σχέση μεταξύ δύο μεταβλητών διαστήματος/αναλογίας*











Συντελεστές συσχέτισης μεταξύ μεταβλητών άλλου τύπου

- Ο συντελεστής **Spearman's *rho*** αφορά στη σχέση μεταξύ δύο μεταβλητών διάταξης ή μίας μεταβλητής διάταξης και μίας μεταβλητής διαστήματος/αναλογίας (τιμές από -1 μέχρι +1)
- Ο συντελεστής **point biserial** (r_{pb}): Μπορεί να χρησιμοποιηθεί όταν η μία μεταβλητή είναι διχοτομική και η άλλη συνεχής *(τιμές από -1 μέχρι +1)*
- Ο συντελεστής **Cramer's *V*:** αφορά στη σχέση μεταξύ δύο ονομαστικών μεταβλητών ή μιας ονομαστικής μεταβλητής και μιας μεταβλητής διάταξης *(τιμές μεταξύ* 0 και 1)
- Ο συντελεστής **φ (phi):** αφορά στη σχέση μεταξύ δύο διχοτομικών μεταβλητών *(τιμές μεταξύ* 0 και 1)

Έλεγχος υποθέσεων: Διαδικασία ελέγχου στατιστικής σημαντικότητας

1. **Διατυπώστε** μια μηδενική υπόθεση- που ορίζει ότι οι δύο υπό εξέταση μεταβλητές δεν έχουν καμία σχέση μεταξύ τους στον πληθυσμό από όπου επιλέχθηκε το δείγμα
2. **Καθορίστε** το αποδεκτό για εσάς επίπεδο στατιστικής σημαντικότητας
3. **Χρησιμοποιήστε** μια στατιστική δοκιμασία (π.χ. X^2, t-test, F-test)
4. **Αν επιτευχθεί ένα αποδεκτό επίπεδο**

   -**απορρίψτε** τη μηδενική υπόθεση

   Αν δεν επιτευχθεί αποδεκτό επίπεδο

   -**αποδεχτείτε την**

Στατιστική Σημαντικότητα

- Επειδή είναι αδύνατο να εξαλείψουμε όλες τις πιθανές πηγές σφάλματος (παρόλο που οφείλουμε να προσπαθήσουμε), πρέπει να ορίσουμε ένα επίπεδο εμπιστοσύνης (ή πιθανότητας σφάλματος).
- Π.χ. επίπεδο σημαντικότητας $p<0.05$ σημαίνει ότι

  -υπάρχει 5% πιθανότητα τα αποτελέσματά μας να οφείλονται στην τύχη ή, αντίθετα,

  -ότι είμαστε 95% σίγουροι ότι τα αποτελέσματά μας ισχύουν στον πληθυσμό έρευνάς μας και οφείλονται σε επίδραση που περιγράφεται στην υπόθεσή μας (επίπεδο εμπιστοσύνης)

- Τελικά: Στατιστική σημαντικότητα είναι το επίπεδο ρίσκου που αναλαμβάνει ο ερευνητής να απορρίψει τη μηδενική υπόθεση όταν στην πραγματικότητα αυτή ισχύει.

Έλεγχος υποθέσεων ή πότε κάτι είναι «σημαντικό» (στατιστικά);

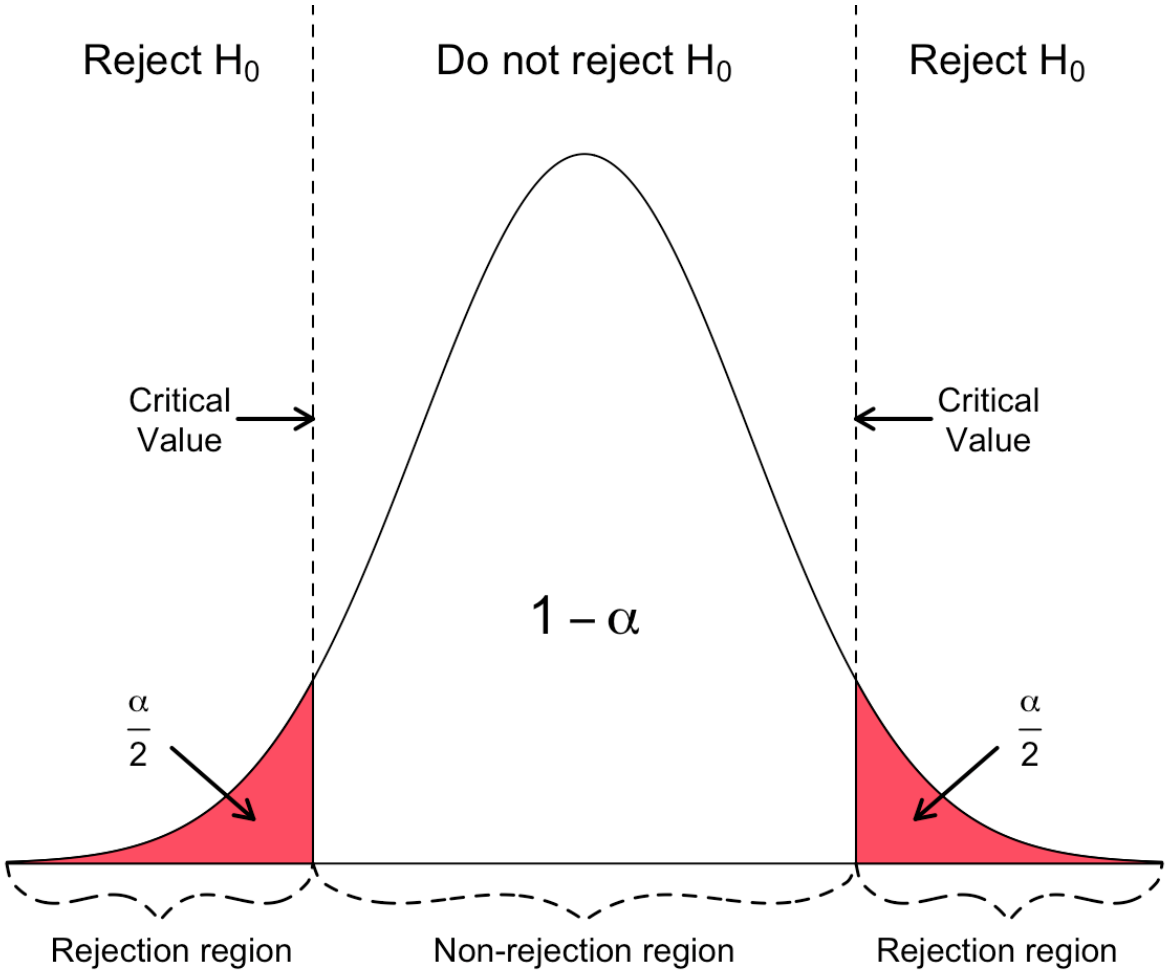
Ψευδώς θετικό

Ψευδώς αρνητικό

Όταν απορρίπτουμε μια μηδενική υπόθεση που ισχύει.

Όταν δεχόμαστε μια μηδενική υπόθεση που δεν ισχύει.

Έλεγχος υποθέσεων ή πότε κάτι είναι «σημαντικό» (στατιστικά);



Υπολογισμός t value για συντελεστή συσχέτισης Pearson's r

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Ο έλεγχος χ^2

Ο έλεγχος **x^2** εφαρμόζεται σε πίνακες συνάφειας μεταξύ ονομαστικών/κατηγορικών μεταβλητών. Μας επιτρέπει να καθορίσουμε το βαθμό εμπιστοσύνης μας στην ύπαρξη μιας σχέσης μεταξύ δύο μεταβλητών στον πληθυσμό. Περιλαμβάνει τον υπολογισμό μιας **αναμενόμενης συχνότητας** ή τιμής για κάθε κελί του πίνακα- δηλαδή μιας συχνότητας που να οφείλεται καθαρά στην τύχη. Η τιμή του x^2 προσδιορίζεται με τον υπολογισμό της διαφοράς μεταξύ της πραγματικής και της αναμενόμενης τιμής για κάθε κελί και μετά το άθροισμα αυτών των διαφορών.

$$\chi^2 = \sum \frac{(παρατηρούμενη\ τιμη - αναμενόμενη\ τιμη)^2}{αναμενόμενη\ τιμη}$$

Το αν θα επιτευχθεί στατιστική σημαντικότητα από μια τιμή x^2 εξαρτάται όχι μόνο από το μέγεθός της, αλλά και από τον αριθμό των κατηγοριών των δύο αναλυόμενων μεταβλητών. Αυτό το τελευταίο καθορίζεται από τους λεγόμενους '**βαθμούς ελευθερίας**' που συνδέονται με τον πίνακα.

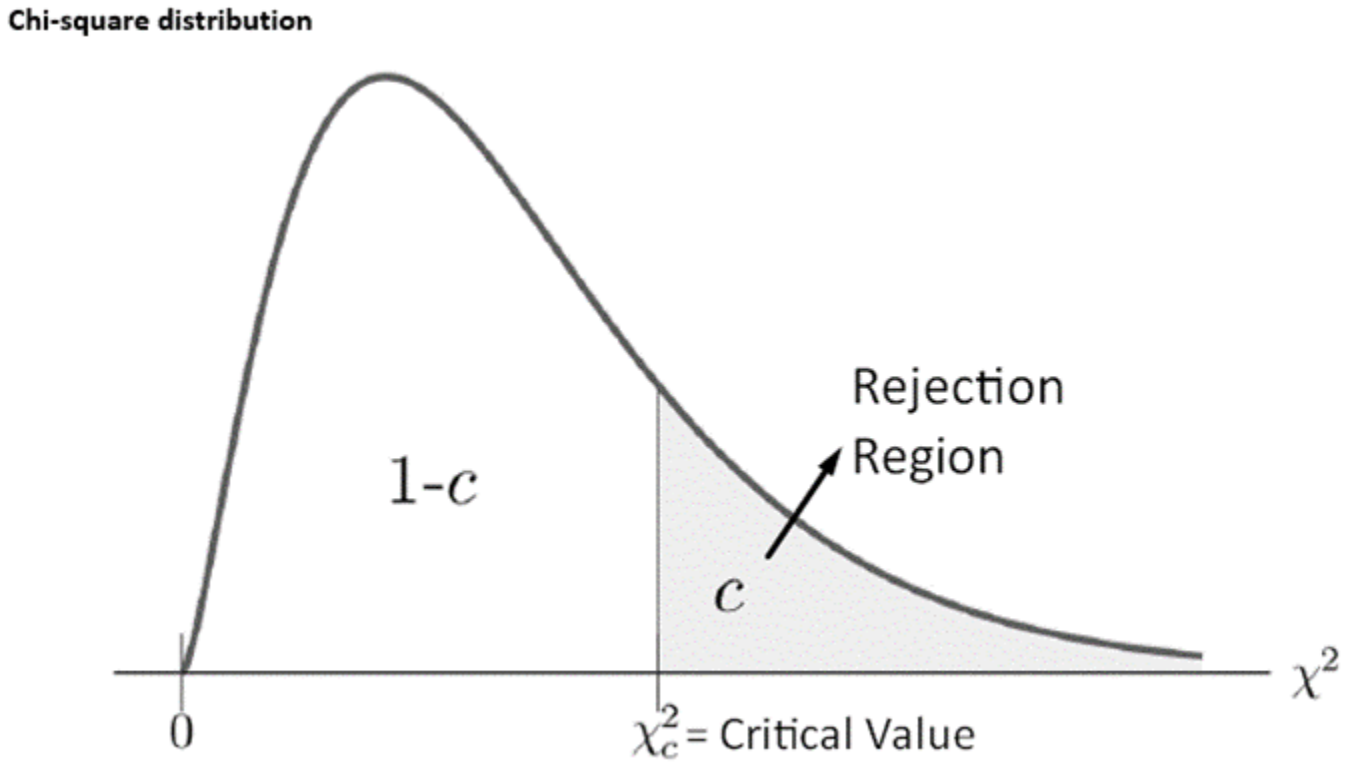
Βαθμοί ελευθερίας = (αριθμός γραμμών -1) x (αριθμός στηλών-1)

Ο έλεγχος χ^2

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of x^2								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36	27.69
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68	29.14
15	5.229	7.261	8.547	11.037	14.339	18.25	22.31	25.00	30.58
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30	32.00
17	6.408	8.672	10.085	12.792	16.338	20.49	24.77	27.59	33.41
18	7.015	9.390	10.865	13.675	17.338	21.60	25.99	28.87	34.80
19	7.633	10.117	11.651	14.562	18.338	22.72	27.20	30.14	36.19
20	8.260	10.851	12.443	15.452	19.337	23.83	28.41	31.41	37.57
22	9.542	12.338	14.041	17.240	21.337	26.04	30.81	33.92	40.29
24	10.856	13.848	15.659	19.037	23.337	28.24	33.20	36.42	42.98
26	12.198	15.379	17.292	20.843	25.336	30.43	35.56	38.89	45.64
28	13.565	16.928	18.939	22.657	27.336	32.62	37.92	41.34	48.28
30	14.953	18.493	20.599	24.478	29.336	34.80	40.26	43.77	50.89
40	22.164	26.509	29.051	33.660	39.335	45.62	51.80	55.76	63.69
50	27.707	34.764	37.689	42.942	49.335	56.33	63.17	67.50	76.15
60	37.485	43.188	46.459	52.294	59.335	66.98	74.40	79.08	88.38

Έλεγχος υποθέσεων ή πότε κάτι είναι «σημαντικό» (στατιστικά);



Έλεγχος υπόθεσης με ονομαστικές μεταβλητές V-Dem (δεδομένα για το έτος 2018)

Μηδενική υπόθεση (ανεξαρτησίας):

H_0 : Ο τύπος του πολιτικού καθεστώτος δεν σχετίζεται με τον τύπο του συστήματος διακυβέρνησης

Εντολές *R*

```
#κάνουμε εγκατάσταση του πακέτου "gmodels" (για να χρησιμοποιήσουμε αργότερα την εντολή "CrossTable"
install.packages("gmodels")
install.packages("smplot2") #για την εντολή "sm_statCorr()"
install.packages("rcompanion") #για την εντολή "cramerV"
library(gmodels) #ενεργοποιούμε το πακέτο
library(tidyverse) #ενεργοποιούμε το πακέτο
library(rcompanion) #ενεργοποιούμε το πακέτο
library(smplot2) #ενεργοποιούμε το πακέτο

#δημιουργούμε διχοτομική μεταβλητή "parl" με βάση τη μεταβλητή v2ex_legconhog του V-Dem (βλ. codebook σ. 160)
Vdem14$parl[Vdem14$v2ex_legconhog==1] <- "Parliamentary"
Vdem14$parl[Vdem14$v2ex_legconhog==0] <- "Presidential"
```

Έλεγχος υπόθεσης με ονομαστικές μεταβλητές V-Dem (δεδομένα για το έτος 2018)

```
#φτιάχνουμε dataframe με δεδομένα του 2018 (πρέπει ήδη να υπάρχει και από προηγούμενο παράδειγμα)
y2018 <- subset(Vdem14, year==2018)


#δημιουργία ονομαστικής μεταβλητής "libdemf" για τον τύπο καθεστώτος (πρέπει ήδη να υπάρχει και από προηγούμενο παράδειγμα)
Vdem14$libdemf[Vdem14$libdem<34] <-"Autocracies"
Vdem14$libdemf[Vdem14$libdem>33 & Vdem14$libdem <=66] <- "Hybrid"
Vdem14$libdemf[Vdem14$libdem>66] <- "Democracies"

#έλεγχος χ²
#δημιουργούμε τον πίνακα συνάφειας
contigencytable <- table(y2018$libdemf, y2018$parl)
#χρησιμοποιούμε το Pearson's chi-square test
CrossTable(contigencytable, chisq = TRUE, expected = TRUE, format = "SPSS")
cramerV(contigencytable) #υπολογίζουμε το Cramer's V
```

```
   Cell Contents
|-------------------------|
|                   Count |
|         Expected Values |
| Chi-square contribution |
|             Row Percent |
|          Column Percent |
|           Total Percent |
|-------------------------|

Total Observations in Table:  179

             |
             | Parliamentary  |  Presidential  |     Row Total  |
-------------|----------------|----------------|----------------|
 Autocracies |             27 |             52 |             79 |
             |         38.838 |         40.162 |                |
             |          3.608 |          3.489 |                |
             |         34.177% |         65.823% |         44.134% |
             |         30.682% |         57.143% |                |
             |         15.084% |         29.050% |                |
-------------|----------------|----------------|----------------|
 Democracies |             30 |              9 |             39 |
             |         19.173 |         19.827 |                |
             |          6.114 |          5.912 |                |
             |         76.923% |         23.077% |         21.788% |
             |         34.091% |          9.890% |                |
             |         16.760% |          5.028% |                |
-------------|----------------|----------------|----------------|
      Hybrid |             31 |             30 |             61 |
             |         29.989 |         31.011 |                |
             |          0.034 |          0.033 |                |
             |         50.820% |         49.180% |         34.078% |
             |         35.227% |         32.967% |                |
             |         17.318% |         16.760% |                |
-------------|----------------|----------------|----------------|
Column Total |             88 |             91 |            179 |
             |         49.162% |         50.838% |                |
-------------|----------------|----------------|----------------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  19.19059     d.f. =  2     p =  6.804818e-05
```

Έλεγχος υπόθεσης με ονομαστικές μεταβλητές V-Dem (δεδομένα για το έτος 2018)

Output της εντολής CrossTable()

Μπορούμε να απορρίψουμε τη μηδενική υπόθεση H_0

Έλεγχος υπόθεσης με συνεχείς μεταβλητές V-Dem (δεδομένα για το έτος 2018)

```
#explore variables for subset of dataframe
hist(Vdem14$e_gdppc[Vdem14$year==2018]) #gdp per capita (βλ. σ. 396 codebook)
hist(Vdem14$libdem[Vdem14$year==2018])

#create scatterplot
plot(x=y2018$e_gdppc, y=y2018$libdem)

#Δείκτης Pearson's
rcor.test(y2018$e_gdppc, y=y2018$libdem, alternative="two.sided",
          method="pearson", conf.level = 0.95)

#Scatterplot using ggplot with regression fit line and 95% confidence intervals


ggplot(subset(Vdem14, Vdem14$year==2018), aes(x=e_gdppc, y=libdem)) +
          geom_point(color="blue", size=2, shape=1) +
          geom_smooth(method=lm) +
          sm_statCorr() +
          ylab("Liberal Democracy (0-100)") +
          xlab("GDP per capita")
```