Μεθοδολογία των επιστημών του Ανθρώπου : Στατιστική

Εργαστήριο 7 (συν):

1. Να εξηγηθεί με ποιο τρόπο οι παρακάτω παράγοντες επιδρούν στο εύρος ενός διαστήματος εμπιστοσύνης.
   a. Αύξηση του μεγέθους του δείγματος
   b. Αύξηση της μεταβλητότητας του δείγματος.
   c. Αύξηση του συντελεστή εμπιστοσύνης
2. Να χρησιμοποιηθεί το αρχείο gssft.sav για να γίνει έλεγχος της υπόθεσης ότι στους εργαζόμενους με πλήρη απασχόληση η τιμή του μέσου χρόνου εργασίας εβδομαδιαία είναι 40 ώρες(μεταβλητή hrs1).
   a. Ποιο είναι το 95% διάστημα εμπιστοσύνης για το μέσο χρόνο εβδομαδιαίας απασχόλησης σε εργαζόμενους με πλήρη απασχόληση.

1ος τρόπος: Από το μενού {Analyze => Descriptive Statistics => Explore} εισάγουμε τη μεταβλητή hrs1 στο πλαίσιο {Dependent List}. Επίσης αν θέλουμε μόνο πίνακα αποτελεσμάτων επιλέγουμε {Statistics} στον τομέα {Display}. Μετά το «κλικ» στο OK του παραθύρου παίρνουμε στο output τους εξής πίνακες:

Πίνακας 1.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
hrs1 Number of Hours Worked Last Week	741	99,2%	6	0,8%	747	100,0%

Στο πίνακα φαίνεται να υπάρχουν 6 άτομα, από τα 747 του δείγματος, χωρίς τιμή στη μεταβλητή hrs1

Πίνακας 2.

Descriptives

			Statistic	Std. Error
hrs1 Number of Hours Worked Last Week	Mean		46,29	,414
	95% Confidence Interval for Mean	Lower Bound	45,48	
		Upper Bound	47,10	
	5% Trimmed Mean		45,84	
	Median		42,00	
	Variance		126,992	
	Std. Deviation		11,269	
	Minimum		5	
	Maximum		89	
	Range		84	
	Interquartile Range		10	
	Skewness		,731	,090
	Kurtosis		1,735	,179

Το 95% διάστημα εμπιστοσύνης για τη μέση τιμή των ωρών εβδομαδιαίας εργασίας είναι το [45,48 47,10]

2ος τρόπος: Όταν κάνουμε το έλεγχο της μηδενικής υπόθεσης για μια συγκεκριμένη τιμή πχ 40 τότε στο «εργαστήριο_6_απαντήσεις» δείξαμε ένα άλλο απλό τρόπο να κατασκευάσουμε το ζητούμενο διάστημα εμπιστοσύνης προσθέτοντας την τιμή 40 στο διάστημα εμπιστοσύνη που μας δίνει η διαδικασία "one sample T-test". ΑΑτο συμβαίνει επειδή η εν λόγω διαδικασία μα δίνει δίαστημα εμπιστοσύνης για τη διαφορά **μ-μ₀** που στην προκείμενη περίπτωση είναι η διαφορά **μ-40**.

b. Βασιζόμενοι στο 95% διάστημα εμπιστοσύνης μπορείτε να απορρίψετε την μηδενική υπόθεση ότι ο πληθυσμιακός μέσος είναι 43 ώρες την εβδομάδα

   Για να απαντηθεί το ερώτημα τοποθετώ την τιμή 43 σε σχέση με το 95% διάστημα εμπιστοσύνης [45,48 47,10] που κατασκευάσαμε στο ερώτημα a. Επειδή η τιμή 43 βρίσκεται έξω από το 95% διάστημα και οδηγούμαστε στην απόρριψη της μηδενικής υπόθεσης (σε επίπεδο σημαντικότητας (100-95=5%). Και επειδή η τιμή 43 βρίσκεται στην περιοχή μικρότερων τιμών (43<45,48) συμπεραίνουμε ότι η μέση τιμή του πληθυσμού είναι σημαντικά μεγαλύτερη από 43.

3. Ποια από τις παρακάτω προτάσεις δεν είναι αληθής και ποια όχι
   a. Με το 95% διάστημα εμπιστοσύνης για τη μέση τιμή ενός πληθυσμού μπορεί να γίνει έλεγχος της μηδενικής υπόθεσης για την για οποιαδήποτε τιμή μ_o του πληθυσμού σε επίπεδο σημαντικότητας α=0,05 ή 5%.

      Αλήθεια
   b. Το διάστημα εμπιστοσύνης για τη διαφορά ανάμεσα τη μέση τιμή του πληθυσμού (μ) στη τιμή της μηδενικής υπόθεσης (μ_o) για την οποία κάνουμε έλεγχο δίνεται τόσο από την επιλογή "One sample T-test" όσο και από την επιλογή "Explore".

      Δεν αληθεύει. Μπορεί να γίνει μόνο στο "One samle T-test"

c. Το βασικό πλεονέκτημα του διαστήματος εμπιστοσύνης έναντι του κλασσικού ελέγχου υπόθεσης είναι ότι μας προμηθεύει με την ακριβή πιθανότητα (τιμή p ή p-value) εσφαλμένης απόρριψης της Ho.

   Δεν αληθεύει. Αυτό ισχύει στον έλεγχο υποθέσεων.

d. Για οποιαδήποτε μηδενική υπόθεση ($\mu=\mu_o$ για τη μέση τιμή ενός πληθυσμού ή $\mu_1 = \mu_2$ για τη σύγκριση των μέσων δύο πληθυσμών) αν η τιμή 0 βρίσκεται στο διάστημα εμπιστοσύνης που κατασκευάζεται δεν απορρίπτεται η Ho.

   Αληθεύει.

e. Το 80% διάστημα εμπιστοσύνης είναι πάντα μεγαλύτερου μήκους από το 95% διάστημα εμπιστοσύνης που δημιουργείται από τα ίδια δεδομένα.

f. Από τη διαδικασία του «one sample t-test» του SPSS για να δημιουργήσω ένα διάστημα εμπιστοσύνης για τη μέση τιμή του πληθυσμού μιας μεταβλητής (πχ educ) θα πρέπει να βάλω ως τιμή της μηδενικής υπόθεσης (Value) την τιμή 0.

g. Από την επιλογή «Explore» του SPSS μπορώ να δημιουργήσω ένα διάστημα εμπιστοσύνης για τη διαφορά ως προς το μέσο μορφωτικό επίπεδο (educ) μεταξύ ανδρών και γυναικών (sex).

4. Να χρησιμοποιηθούν τα δεδομένα του αρχείου gss.sav για να ελεγχθεί με τη βοήθεια ενός 90% διαστήματος εμπιστοσύνης:

   a. Πως διαφέρουν άνδρες και γυναίκες μεταξύ του ως προς τα έτη εκπαίδευσης (educ); Ποια είναι η διαφορά τους στο δείγμα; Ποια είναι η εκτιμώμενη διαφορά στον πληθυσμό;

      Το διάστημα για τη διαφορά των μέσων τιμών για τις άγνωστες μέσες πληθυσμιακές τιμές ανδρών και γυναικών μπορεί να κατασκευαστεί ευθέως από την επιλογή “Independent samples t-test”. Πράγματι το δύο φύλα συνθέτουν δύο ανεξάρτητες ομάδες στην περίπτωση των ερευνών όπου ένα τυχαίο δείγμα συμμετεχόντων απαντά σε ένα ερωτηματολόγιο που συμπεριλαμβάνει και δημογραφικά χαρακτηριστικά.

      Από το μενού Analyze => Compare Means => Independent Samples T Test εισάγουμε στο πλαίσιο Grouping Variable τη μεταβλητή sex και δίνουμε στο Define Groups τις τιμές 1 και 2 που αντιστοιχούν στα δύο φύλα. Στο πλαίσιο Test Variable(s) εισάγουμε από τη λίστα τη μεταβλητή tvhours.(ακριβώς την ίδια διαδικασία περιγράφουμε στο αρχείο

«εργαστήριο_7_απαντήσεις». Η μοναδική διαφορά είναι ότι εδώ από την επιλογή {Options} ορίζουμε αντί 95 το 90 στο πλαίσιο {Confidence Interval Percentage} Οι παρακάτω πίνακες δίνουν το αποτέλεσμα του ελέγχου αλλά και το ζητούμενο διάστημα εμπιστοσύνης.

Πίνακας 1. Group Statistics

	sex Respondent's Sex	N	Mean	Std. Deviation	Std. Error Mean
educ Highest Year of School Completed	1 Male	639	13,19	3,349	,132
	2 Female	857	12,92	2,849	,097

Πίνακας 2.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									90% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
educ Highest Year of School Completed	Equal variances assumed	20,705	,000	1,652	1494	,099	,265	,161	,001	,529
	Equal variances not assumed			1,613	1242,5	,107	,265	,164	-,005	,536

Το 90% διάστημα εμπιστοσύνης για τη διαφορά των μέσων τιμών των ωρών εβδομαδιαίας εργασίας μεταξύ ανδρών και γυναικών είναι το [-0,005 0,536]

Η ισότητα των διακυμάνσεων ελέγχεται από την τιμή p του Ελέγχου Levene. Εδώ επειδή p<0,001 απορρίπτεται η H_o:$\sigma_1^2=\sigma_2^2$

Αφού η ισότητα των διακυμάνσεων των δύο πληθυσμών απορρίπτεται (p<0,001), η τιμή t που θα χρησιμοποιηθεί για την κατασκευή του διαστήματος προκύπτει από την εναλλακτική κατανομή της κάτω γραμμής {equal variances not assumed}. Επειδή το 0, δηλαδή η τιμή της μηδενικής υπόθεσης (H_o: $\mu_1=\mu_2$ ή **H_o: $\mu_1-\mu_2=0$**) για την διαφορά $\mu_1-\mu_2$, βρίσκεται εντός του 90% διαστήματος [-0,005 0,536] δεν απορρίπτεται η μηδενική υπόθεση σε επίπεδο 0,1 ή 10%. Δηλαδή τα δύο φύλα δεν διαφέρουν σημαντικά ως προς το μορφωτικό του επίπεδο. Η διαφορά μέσων τιμών μεταξύ των δυο δειγμάτων είναι 0,265 (στήλη Mean Difference του πίνακα 2) υπέρ των ανδρών όπως φαίνεται από τον Πίνακα 1. Η εκτιμώμενη διαφορά στον πληθυσμό αν ζητάμε σημειακή εκτίμηση (δηλαδή μια μοναδική τιμή) είναι 0,265 δηλαδή η ίδια με εκείνη που παρατηρήθηκε στα 2 δείγματα. Αν ζητάμε εκτίμηση με 90% διάστημα εμπιστοσύνης είναι από -0,005 (δηλαδή μικρή υπεροχή των γυναικών) έως 0,536 (υπεροχή των ανδρών κατά ½ περίπου έτος σπουδών)

b. Κατά πόσο άνδρες και γυναίκες ξοδεύουν τον ίδιο χρόνο στην εργασία τους (hrs1). Ποια είναι η διαφορά τους στο δείγμα; Ποια είναι η εκτιμώμενη διαφορά στον πληθυσμό.

c. Πως διαφέρουν άνδρες και γυναίκες μεταξύ του ως προς το ατομικό τους εισόδημα (rincomdol). Ποια είναι η διαφορά τους στο δείγμα; Ποια είναι η εκτιμώμενη διαφορά στον πληθυσμό.

d. Να κατασκευαστεί ξανά ένα 80% διάστημα εμπιστοσύνης για να ελέγξετε τις διαφορές στις 3 παραπάνω περιπτώσεις. Τα συμπεράσματά σας είναι τώρα διαφορετικά; Γιατί;

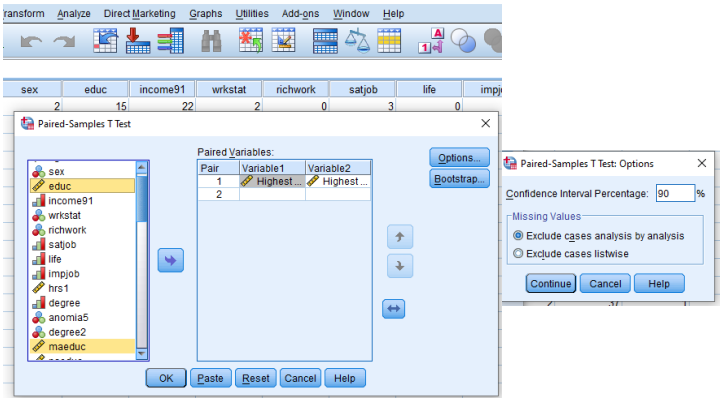
5. Να χρησιμοποιηθούν τα δεδομένα του αρχείου gss.sav για να ελεγχθεί με τη βοήθεια ενός 90% διαστήματος εμπιστοσύνης πως διαφέρουν ως προς τα έτη εκπαίδευσης οι σημερινές γυναίκες από τις μητέρες τους:

Στην περίπτωση αυτή τα δείγματα τιμών εκπαίδευσης των γυναικών και των μητέρων τους είναι σχετιζόμενα αφού η εκπαίδευση μιας μητέρας είναι ζευγάρι με την τιμή της εκπαίδευσης της κόρης της.

Ο έλεγχος της υπόθεσης για την διαφορά των μέσων τιμών (H_o: $\mu_1=\mu_2$ ή **H_o: $\mu_1-\mu_2=0$**) αυτή τη φορά θα διεξαχθεί από την επιλογή {Analyze => Compare Means => Paired Samples T- Test}.

Πριν ξεκινήσουμε τη διαδικασία ελέγχου πρέπει πρώτα να επιλέξουμε με την επιλογή {Data => Select Cases} το δείγμα των γυναικών από τα δεδομένα του αρχείου μας (τη χρήση του select θα την βρείτε σε προηγούμενα εργαστήρια αλλα και σε videos που υπάρχουν στην e-class)

Στη συνέχεια κάνουμε τις κατάλληλες επιλογές των δύο μεταβλητών (educ και maeduc) και συντελεστή εμπιστοσύνης στην εντολή {Analyze => Compare Means => Paired Samples T- Test} όπως φαίνεται παρακάτω και κάνουμε κλικ στο {OK}:



Το αποτέλεσμα της παραπάνω διαδικασίας δίνεται στους πίνακες:

Πίνακας 1.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	maeduc Highest Year of School Completed, Mother	10,94	735	3,385	,125
	educ Highest Year of School Completed	13,26	735	2,744	,101

Πίνακας 2.

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	maeduc Highest Year of School Completed, Mother & educ Highest Year of School Completed	735	,474	,000

Πίνακας 3.

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	90% Confidence Interval of the Difference Lower	Upper			
Pair 1	maeduc Highest Year of School Completed, Mother - educ Highest Year of School Completed	-2,317	3,191	,118	-2,511	-2,123	-19,686	734	,000

Το 90% διάστημα εμπιστοσύνης για τη διαφορά των μέσων τιμών εκπαίδευσης ανάμεσα στις σημερινές γυναίκες και τις μητέρες τους είναι το [-2,511 -2,123]

Το συμπέρασμα το ελέγχου προκύπτει από το 90% διάστημα εμπιστοσύνης [-2,511 -2,123] που δίνεται στον πίνακα 3. Αφού η τιμή 0 είναι έξω από το διάστημα απορρίπτεται η H_o. Η διαφορά $\mu_{\text{μητέρων}}$-$\mu_{\text{γυναικών}}$ είναι σημαντικά μικρότερη του 0 αφού όλο το διάστημα που την εκτιμά ανήκει στις αρνητικές τιμές. Συμπερασματικά η μέση εκπαίδευση των σύγχρονων γυναικών είναι σημαντικά υψηλότερη (κατά 2,317 χρόνια) από την μέση εκπαίδευση των μητέρων τους σε επίπεδο σημαντικότητας 0,10 ή 10%.

Φυσικά το ίδιο αποτέλεσμα προκύπτει από τον ίδιο πίνακα 3 χρησιμοποιώντας την λογική του ελέγχου υποθέσεων. Πράγματι, στην στήλη {Sig.(2-tailed)} έχουμε την τιμή $p < 0,001$ η οποία προκύπτει από τον έλεγχο της υπόθεσης με την κατανομή t, και φυσικά απορρίπτει ισχυρά την H_o. Η τιμή p θα μπορούσε να προσδιοριστεί ακριβέστερα ζητώντας περισσότερα δεκαδικά ψηφία (όσα επιτρέπει το λογισμικό) στον πίνακα 3.