

6. ΣΥΜΠΕΡΑΣΜΑΤΟΛΟΓΙΑ ΣΤΗ ΓΡΑΜΜΙΚΗ ΣΥΣΧΕΤΙΣΗ

Είναι φυσικό ότι ο δειγματικός συντελεστής R , ως μια στατιστική συνάρτηση, είναι μιά τυχαία μεταβλητή. Όπως είπαμε ήδη μπορεί να χρησιμοποιηθεί σαν εκτιμητή του ρ . Για να μπορέσουμε να κάνουμε στατιστική συμπερασματολογία για το ρ χρειάζεται να ξέρουμε την από κοινού κατανομή των X και Y . Ιδιαίτερο ενδιαφέρον παρουσιάζει η περίπτωση που το (X,Y) ακολουθεί την διμεταβλητή κανονική κατανομή.

6.1 Η Διμεταβλητή Κανονική Κατανομή (The Bivariate Normal Distribution)

Ορισμός: Το άνυσμα (X,Y) ακολουθεί τη διμεταβλητή (ή διδιάστατη κανονική κατανομή) αν η από κοινού συνάρτηση πυκνότητας πιθανότητας $f(x,y)$ έχει τη μορφή

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[\left(\frac{x-\mu_x}{\sigma_x} \right)^2 - 2\rho \left[\frac{x-\mu_x}{\sigma_x} \right] \left[\frac{y-\mu_y}{\sigma_y} \right] + \left[\frac{y-\mu_y}{\sigma_y} \right]^2 \right] \right\}$$

όπου μ_x , μ_y , σ_x^2 , σ_y^2 και ρ είναι οι παράμετροι της κατανομής. Μπορούν να αποδειχθούν οι παρακάτω ιδιότητες της διμεταβλητής κανονικής.

Αν (X,Y) δ.μετ. $N(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$ τότε

$$I_1) X \sim N(\mu_x, \sigma_x^2)$$

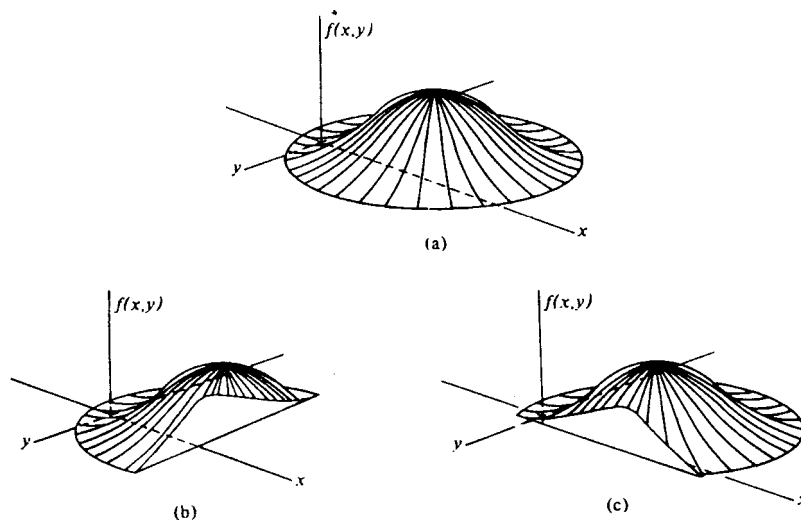
$$I_2) Y \sim N(\mu_y, \sigma_y^2)$$

$$I_3) Y | (X=x) \sim N \left(\mu_y + \rho \frac{\sigma_y}{\sigma_x} (x - \mu_x), \sigma_y^2 (1 - \rho^2) \right)$$

$$I_4) X | (Y=y) \sim N \left(\mu_x + \rho \frac{\sigma_x}{\sigma_y} (y - \mu_y), \sigma_x^2 (1 - \rho^2) \right)$$

Οι ιδιότητες αυτές είναι σημαντικές διότι, εάν θέσουμε $\mu_y - \rho \frac{\sigma_y}{\sigma_x} \mu_x = \alpha$ και $\rho \frac{\sigma_y}{\sigma_x} = \beta$ από την I_3 θα έχουμε, $Y | (X=x) \sim N(\alpha + \beta x, \sigma_y^2(1-\rho^2))$ δηλαδή η $Y|x$ ακολουθεί κανονική κατανομή που η μέση τιμή της είναι γραμμική συνάρτηση του x και η διασπορά της είναι ανεξάρτητη από το x . Αυτό συνεπάγεται ότι αν έχουμε ένα διμεταβλητό τυχαίο δείγμα $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ από μία διμεταβλητή κανονική κατανομή όλη η στατιστική συμπερασματολογία που έχουμε αναπτύξει για τη γραμμική παλινδρόμηση με σταθερό και ελεγχόμενο X θα ισχύει και για την περίπτωση που το X είναι τυχαία μεταβλητή.

Στο σχήμα που ακολουθεί φαίνεται η γραφική παράσταση μιας διμεταβλητής κανονικής κατανομής όπως και των αντιστοίχων περιθωρίων κατανομών της.



Η Διδιάστατη Κανονική Κατανομή

Παράδειγμα: Εστω X και Y τα βάρη ενός αδελφού και μίας αδελφής, αντίστοιχα, σε μία οικογένεια. Ας υποθέσουμε ότι το (X,Y) ακολουθεί τη διμεταβλητή κανονική κατανομή με παραμέτρους $\mu_x=69.5\text{kg}^*$, $\mu_y=64.5\text{kg}^*$, $\sigma_x=2.5\text{kg}^*$, $\sigma_y=2.4\text{kg}^*$ και $\rho=0.6$.

- α) Να βρεθεί η ευθεία γραμμικής παλινδρόμησης του Y επί του X.
 β) Να βρεθεί η ευθεία γραμμικής παλινδρόμησης του X πάνω στο Y.
 γ) Να βρεθεί η δεσμευμένη διασπορά του Y|x και X|y.
 δ) Να βρεθεί η δεσμευμένη κατανομή των βαρών των αδελφών
 (γυναικών) που έχουν αδελφούς των οποίων το βάρος είναι 68
 kgr*.

Λύση:

α) Έχουμε ότι

$$E(Y|x) = \alpha + \beta x$$

όπου

$$\alpha = \mu_y - \rho \frac{\sigma_y}{\sigma_x} \mu_x = 24.468$$

$$\beta = \rho \frac{\sigma_y}{\sigma_x} = 0.576$$

Επομένως, η ευθεία γραμμικής παλινδρόμησης του Y|x είναι

$$E(Y | x) = 24.468 + 0.576 x.$$

β) Εδώ έχουμε

$$E(X | y) = \alpha^* + \beta^* y$$

όπου

$$\alpha^* = \mu_x - \rho \frac{\sigma_x}{\sigma_y} \mu_y = 29.1875$$

$$\beta^* = \rho \frac{\sigma_x}{\sigma_y} = 0.625$$

Επομένως

$$E(X|y) = 29.1875 + 0.625 y$$

γ)

$$V(Y|x) = \sigma_y^2 (1 - \rho^2) = 3.686$$

$$V(X|y) = \sigma_x^2 (1 - \rho^2) = 4.00$$

δ) Επειδή $(X, Y) \sim$ διμ. κανονική κατανομή θα έχουμε

$$Y | (X = x) \sim N\left(\mu_x + \rho \frac{\sigma_y}{\sigma_x}(x - \mu_x), \sigma_y^2(1 - \rho^2)\right)$$

Επομένως

$$Y | (X = 68) \sim N(63.64, 3.686)$$

6.2 Συμπερασματολογία για το ρ

Όταν και το X και το Y είναι τυχαίες μεταβλητές (όπως για παράδειγμα στην περίπτωση που το (X, Y) ακολουθεί τη διμεταβλητή κανονική κατανομή) δεν σημαίνει υποχρεωτικά ότι μπορούμε να θεωρήσουμε τη συγκεκριμένη τιμή της τυχαίας μεταβλητής σαν "αιτία" για την αντίστοιχη τιμή της άλλης τυχαίας μεταβλητής. (Όταν οι δύο τυχαίες μεταβλητές έχουν κάποια γραμμική σχέση). Με άλλα λόγια, όταν δύο τυχαίες μεταβλητές είναι γραμμικά συνδεδεμένες αυτό δεν συνεπάγεται μια σχέση "αιτίας"- "αποτελέσματος".

Για παράδειγμα, αν X είναι το ύψος του πατέρα σε μία οικογένεια και Y είναι το ύψος ενός γιού, δεν είναι παράλογο να μιλήσουμε για σχέση αιτίας - αποτελέσματος (ψηλοί πατέρες έχουν, συνήθως, ψηλούς γιούς). Όταν όμως X είναι το ύψος ενός αδελφού και το Y το ύψος μίας αδελφής σε μία οικογένεια και υπάρχει γραμμική σχέση ανάμεσα στο X και το Y δεν πρέπει να θεωρήσουμε ότι μία συγκεκριμένη τιμή του X είναι αιτία για μία τιμή του Y . X και Y μπορεί να έχουν κάποια συσχέτιση αυτό όμως δεν σημαίνει ότι η μία είναι αιτία της άλλης. Και οι δύο σχετίζονται στην πραγματικότητα με κάποιο άλλο παράγοντα (στη συγκεκριμένη περίπτωση τους ίδιους γονείς).

Στην περίπτωση που μία σχέση αιτίας - αποτελέσματος δεν έχει έννοια χρησιμοποιούμε το συντελεστή συσχέτισης ρ σαν μέτρο του βαθμού γραμμικής εξάρτησης των X και Y (αντί να χρησιμοποιήσουμε τη γραμμική παλινδρόμηση της μίας στην άλλη).

Συγκεκριμένα, ξέρουμε ήδη ότι αν X και Y είναι ανεξάρτητες τυχαίες μεταβλητές τότε είναι και ασυσχέτιστες ($\rho=0$). Το αντίστροφο **δεν συμβαίνει πάντα**. Αν δηλαδή, $\rho=0$ αυτό δεν σημαίνει υποχρεωτικά ότι X και Y είναι ανεξάρτητες.

Παρόλα αυτά αποδεικνύεται εύκολα ότι αν (X,Y) ακολουθούν τη διμεταβλητή κανονική κατανομή τότε $\rho=0 \Rightarrow X,Y$ ανεξάρτητες. (Μάλιστα, μπορεί να αποδειχθεί ότι η διμεταβλητή κανονική κατανομή είναι η μόνη κατανομή που πληροί την ιδιότητα αυτή).

Η ιδιότητα αυτή βοηθά στον έλεγχο της υπόθεσης της ανεξαρτησίας δύο τυχαίων μεταβλητών όταν αυτές ακολουθούν τη διμεταβλητή κανονική κατανομή. Αντί δηλαδή να ελέγξουμε την υπόθεση της ανεξαρτησίας, μπορούμε να ελέγξουμε την υπόθεση $\rho=0$.

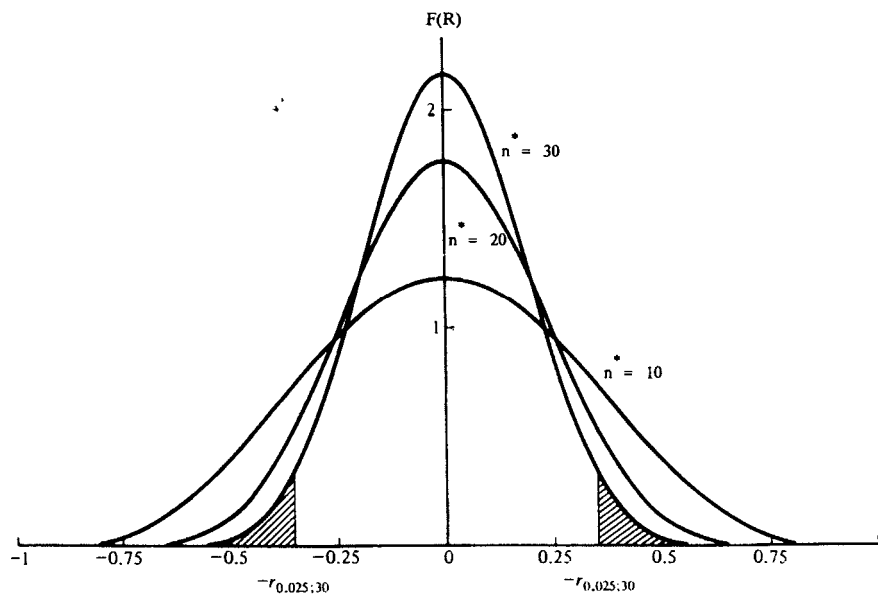
6.3 Έλεγχοι Υποθέσεων για το $\rho = 0$

Αν για διμεταβλητό κανονικό πληθυσμό θέλουμε να ελέγξουμε την υπόθεση $H_0: \rho=0$ (δηλαδή την υπόθεση ότι X και Y είναι ανεξάρτητες) χρειαζόμαστε την κατανομή της εκτιμήτριας R του ρ . Αποδεικνύεται ότι, κάτω από την μηδενική υπόθεση, η συνάρτηση πυκνότητας πιθανότητας του R είναι

$$f(R) = \left\{ \frac{\Gamma\left[\frac{n^*+1}{2}\right]}{\sqrt{\pi} \Gamma\left[\frac{n^*}{2}\right]} \right\} [1 - R^2]^{(n^*-2)/2}$$

όπου n είναι το μέγεθος του δείγματος, $n^*=n-2$ οι βαθμοί ελευθερίας και $\Gamma(\alpha)$ είναι η συνάρτηση γάμμα.

Η συνάρτηση $f(R)$ είναι συμμετρική γύρω από το $r=0$. Το σχήμα 9 που ακολουθεί δίνει τις γραφικές παραστάσεις της $f(R)$ για $n=12, 22$ και 32 ($n^*=10, 20, 30$).



Γραφικές παραστάσεις της $f(R)$ για $n = 12, 22, 32$

Ο Πίνακας που ακολουθεί δίνει το $100(1-\alpha)$ εκατοστιαίο σημείο $r_{n-2, 1-\alpha/2}$ της κατανομής.

n-2	Άνω κριτικές τιμές του r				
	0.900	0.950	0.975	0.990	0.995
1	0.951	0.988	0.997	1.000	1.000
2	0.800	0.900	0.950	0.980	0.990
3	0.687	0.805	0.878	0.934	0.959
4	0.608	0.729	0.811	0.882	0.917
5	0.551	0.669	0.755	0.833	0.875
6	0.507	0.621	0.707	0.789	0.834
7	0.472	0.582	0.666	0.750	0.798
8	0.443	0.549	0.632	0.715	0.765
9	0.419	0.521	0.602	0.685	0.735
10	0.398	0.497	0.576	0.658	0.708
11	0.380	0.476	0.553	0.634	0.684
12	0.365	0.457	0.532	0.612	0.661
13	0.351	0.441	0.514	0.592	0.641
14	0.338	0.426	0.497	0.574	0.623
15	0.327	0.412	0.482	0.558	0.606
16	0.317	0.400	0.468	0.542	0.590
17	0.308	0.389	0.456	0.529	0.575
18	0.299	0.378	0.444	0.515	0.561
19	0.291	0.369	0.433	0.503	0.549
20	0.284	0.360	0.423	0.492	0.537
21	0.277	0.352	0.413	0.482	0.526
22	0.271	0.344	0.404	0.472	0.515
23	0.265	0.337	0.396	0.462	0.505
24	0.260	0.330	0.388	0.453	0.496
25	0.255	0.323	0.381	0.445	0.487
26	0.250	0.317	0.374	0.437	0.479
27	0.245	0.311	0.367	0.430	0.471
28	0.241	0.306	0.361	0.423	0.463
29	0.237	0.301	0.355	0.416	0.456
30	0.233	0.296	0.349	0.409	0.449
40	0.202	0.257	0.304	0.358	0.393
60	0.165	0.211	0.250	0.295	0.325
120	0.117	0.150	0.178	0.210	0.232

Ανάλογα με την εναλλακτική υπόθεση, σε επίπεδο σημαντικότητας α , έχουμε

i) αν $H_1 : \rho > 0$, απορρίπτουμε την H_0 αν $R > r_{n-2, 1-\alpha}$

ii) αν $H_1 : \rho < 0$, απορρίπτουμε την H_0 αν $R < -r_{n-2, 1-\alpha}$

iii) αν $H_1 : \rho \neq 0$, απορρίπτουμε την H_0 αν $R > r_{n-2, 1-\alpha/2}$

ή αν $R < -r_{n-2, 1-\alpha/2}$

Στην περίπτωση που οι πίνακες της στατιστικής συνάρτησης $R_{n-2, 1-\alpha}$ δεν είναι διαθέσιμοι χρησιμοποιούμε την ακόλουθη σχέση της τυχαίας μεταβλητής R με την κατανομή t .

Θεώρημα: Αν $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ είναι ένα τυχαίο δείγμα από τη διμεταβλητή κανονική κατανομή τότε

$$T = R \sqrt{\frac{n-2}{1-R^2}} \sim t_{n-2}$$

Ο έλεγχος επομένως της στατιστικής υπόθεσης $H_0 : \rho = 0$ γίνεται με τη συνήθη μέθοδο, ανάλογα με τη εναλλακτική υπόθεση, σε επίπεδο σημαντικότητας α με την χρήση των πινάκων της κατανομής t .

Σημείωση: Η σχέση της κατανομής του R με την κατανομή t έχει μεγαλύτερη σημασία από την απλή χρήση της μιας ως εναλλακτικής στατιστικής συνάρτησης της άλλης στους ελέγχους υποθέσεων που προκύπτει από το προηγούμενο θεώρημα. Αυτό γιατί, από την προηγούμενη σχέση, αν εκφράσουμε την T μέσω των S_{xx} , S_{yy} και S_{xy} θα έχουμε

$$\begin{aligned} T &= R \sqrt{\frac{n-2}{1-R^2}} \\ &= \sqrt{\frac{(n-2)S_{xy}^2}{S_{yy}S_{xx} - S_{xy}^2}} \end{aligned}$$

$$\begin{aligned}
&= \frac{S_{xy}}{S_{xx}} \sqrt{\frac{S_{xx}^2}{[S_{yy}S_{xx} - S_{xy}^2]/(n-2)}} \\
&= \frac{S_{xy}}{S_{xx}} \sqrt{\frac{(n-2)S_{xx}}{nS_{Y|x}^2}} \\
&= \frac{S_{xy}}{S_{xx}} \sqrt{\frac{(n-2)S_{xx}}{(n-2)S_{Y|x}^{*2}}} \\
&= \frac{\hat{a} \sqrt{S_{xx}}}{\sqrt{S_{Y|x}^{*2}}}
\end{aligned}$$

Η συνάρτηση αυτή όμως συμπίπτει με τη στατιστική συνάρτηση

$$T = \frac{\hat{\beta} - \beta_0}{\sqrt{S_{Y|x}^{*2}/S_{xx}}}$$

που χρησιμοποιήσαμε για τον έλεγχο υποθέσεων για το β της γραμμικής παλινδρόμησης για $\beta_0 = 0$, με X τυχαία μεταβλητή.

Βλέπουμε δηλαδή, ότι ο έλεγχος της υπόθεσης για μηδενική συσχέτιση στη διμεταβλητή κανονική κατανομή (ή ισοδύναμα, ο έλεγχος ανεξαρτησίας) είναι ισοδύναμος με την υπόθεση $H_0: \beta=0$ στην γραμμική παλινδρόμηση για προκαθορισμένες τιμές του x .

Η σχέση αυτή μας επιτρέπει να ελέγχουμε υποθέσεις για έλλειψη σχέσης γραμμικής παλινδρόμησης ($\beta=0$) χωρίς να ανησυχούμε για το κατά πόσο η ανεξάρτητη μεταβλητή είναι ελεγχόμενη ή τυχαία.

6.4 Διαστήματα εμπιστοσύνης για $\rho \neq 0$

Αν $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ είναι ένα τυχαίο δείγμα από ένα πληθυσμό που ακολουθεί τη διμεταβλητή κανονική κατανομή και $\rho \neq 0$ υπάρχει δυνατότητα να κατασκευάσουμε διαστήματα εμπιστοσύνης για το ρ . Για το λόγο αυτό χρησιμοποιούμε το R που

είναι μια σημειακή εκτιμήτρια του ρ . Επειδή η ακριβής κατανομή του R όταν $\rho \neq 0$ είναι εξαιρετικά πολύπλοκη αναγκάζομαστε να κατασκευάσουμε διαστήματα εμπιστοσύνης κατά προσέγγιση χρησιμοποιώντας ένα μετασχηματισμό που είναι γνωστός σαν μετασχηματισμός του Fisher. Έτσι, κατασκευάζουμε αρχικά διάστημα εμπιστοσύνης για μια συνάρτηση του ρ . Η συνάρτηση αυτή (μετασχηματισμός του Fisher) είναι η

$$\mu_{Z'} = \frac{1}{2} \ln \frac{1+\rho}{1-\rho} = \tanh^{-1} \rho$$

(Η επιλογή του μετασχηματισμού αυτού οφείλεται στο γεγονός ότι, όπως θα δούμε στη συνέχεια, υπάρχει αμερόληπτη εκτιμήτρια του μ_Z , με γνωστή κατανομή).

Σύμφωνα με την μέθοδο αυτή κατασκευάζουμε πρώτα ένα διάστημα εμπιστοσύνης για το μ_Z , και μετά μετασχηματίζουμε το διάστημα αυτό σε ένα διάστημα εμπιστοσύνης για το ρ .

Μια αμερόληπτη σημειακή εκτιμήτρια για το μ_Z είναι, προφανώς, η

$$Z' = \frac{1}{2} \ln \frac{1+R}{1-R} = \tanh^{-1} R$$

Μπορεί να αποδειχθεί ότι, ασυμπτωτικά,

$$Z' \sim N[\mu_{Z'}, 1/(n-3)]$$

(Στην πράξη, η προσέγγιση αυτή είναι καλή όταν $n \geq 25$). Επομένως, ασυμπτωτικά

$$\frac{Z' - \mu_{Z'}}{1/\sqrt{n-3}} \sim N(0,1)$$

και σαν ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για την παράμετρο

$$\mu_{Z'} = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}$$

μπορούμε να χρησιμοποιήσουμε το διάστημα που έχει ακραία σημεία τα

$$\frac{1}{2} \ln \frac{1+R}{1-R} \pm Z_{1-\alpha/2} / \sqrt{n-3}$$

Χρησιμοποιώντας τον αντίστροφο του μετασχηματισμού του Fisher εφαρμοζόμενο στα άκρα του διαστήματος αυτού παίρνουμε τα άκρα διαστήματος εμπιστοσύνης για το ρ . Για το μετασχηματισμό και τον αντίστροφο μετασχηματισμό υπάρχουν πίνακες που διευκολύνουν τους υπολογισμούς. Οι πίνακες αυτοί δίνονται στο παράρτημα.

Παράδειγμα: Ένα τυχαίο δείγμα 28 παρατηρήσεων $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, $n = 28$, από μια διμεταβλητή κατανομή έδωσε συντελεστή συσχέτισης (δειγματικό) 0.46 (δηλαδή $r = 0.46$). Να κατασκευασθεί ένα 95% διάστημα εμπιστοσύνης για το συντελεστή συσχέτισης ρ του πληθυσμού (X, Y) .

Λύση: Υποθέτουμε ότι $(X, Y) \sim$ διμεταβλητή κανονική κατανομή. Θα κατασκευάσουμε πρώτα ένα διάστημα εμπιστοσύνης για το

$$\mu_{Z'} = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}.$$

Έχουμε, $Z_{0.975} = 1.96$. Το 95% διάστημα εμπιστοσύνης για το $\mu_{Z'}$ είναι το

$$\frac{1}{2} \ln \frac{1+0.46}{1-0.46} \pm 1.96 / \sqrt{28-3}$$

και χρησιμοποιώντας τους πίνακες ή κάνοντας τις πράξεις, βρίσκουμε $0.4973 \pm 1.96/5 \Rightarrow (0.1053, 0.8893)$.

Χρησιμοποιώντας τον αντίστροφο μετασχηματισμό του Fisher, ή τους πίνακες, βρίσκουμε ότι το αντίστοιχο διάστημα εμπιστοσύνης για το ρ θα έχει άκρα, κατά προσέγγιση, 0.105 και 0.711. Επομένως, ένα κατά προσέγγιση 95% διάστημα εμπιστοσύνης για το ρ στην περίπτωση αυτή είναι το (0.105, 0.711).

Σημείωση: Εκτός από τον μετασχηματισμό του Fisher υπάρχει και μία γραφική μέθοδος για τον καθορισμό διαστημάτων εμπιστοσύνης για το ρ όταν τα τυχαία δείγματα προέρχονται από μία διμεταβλητή κανονική κατανομή. Το πλεονέκτημα της γραφικής μεθόδου συνίσταται στο ότι με αυτήν δεν απαιτούνται αριθμητικού υπολογισμοί. Το μειονέκτημα της μεθόδου είναι ότι για κάθε επίπεδο σημαντικότητας α απαιτείται διαφορετικό γράφημα, ότι τα γραφήματα

μπορούν να χρησιμοποιηθούν μόνο σε δείγματα μικρού μεγέθους, και ότι έχουν ακρίβεια, το πολύ, δύο δεκαδικών ψηφίων. Παραδείγματα τέτοιων γραφικών πινάκων, για διαφορετικά α δίνονται στο παράτημα.

Παράδειγμα: Έστω ότι έχουμε ένα τυχαίο δείγμα 10 παρατηρήσεων με δειγματικό συντελεστή συσχέτισης $r = 0.4$ από ένα πληθυσμό που ακολουθεί διμεταβλητή κανονική κατανομή. Να κατασκευασθεί ένα 95% διάστημα εμπιστοσύνης για το ρ .

Λύση: Χρησιμοποιώντας το γράφημα του παραρτήματος που αντιστοιχεί σε $\alpha = 0.95$ και θεωρώντας την κατακόρυφη γραμμή που ενώνει τα σημεία 0.4 εντοπίζουμε τα σημεία στα οποία η γραμμή αυτή τέμνει τις δύο καμπύλες που αντιστοιχούν σε $n=10$. Οι οριζόντιες συντεταγμένες των δύο αυτών σημείων τομής είναι τα 95% άκρα του ζητούμενου διαστήματος εμπιστοσύνης. Στο παράδειγμα μας τα σημεία αυτά είναι, περίπου, το -0.30 και +0.78.

6.5 Έλεγχοι Υποθέσεων για $\rho \neq 0$

Ο μετασχηματισμός του Fisher χρησιμοποιείται και για ελέγχους υποθέσεων για τιμές του $\rho \neq 0$. Στην περίπτωση αυτή χρησιμοποιείται η στατιστική συνάρτηση

$$Z = \frac{Z' - \mu_{Z'}}{1/\sqrt{n-3}}$$

που όπως είδαμε ασυμπτωτικά είναι $N(0, 1)$.