# Γραμμική συσχέτιση τυχαίων μεταβλητών

## Ανεξάρτητες τυχαίες μεταβλητές (independent random variables)

Δύο τυχαίες μεταβλητές **X**, **Y** λέγονται ανεξάρτητες αν η μία δεν επηρεάζει την πιθανότητα πραγματοποίησης της άλλης. Ο ορισμός είναι ο ακόλουθος

$$P(\mathrm{X}=a,\mathrm{Y}=b)=P(\mathrm{X}=a)\cdot P(\mathrm{Y}=b),$$

για κάθε δύο αριθμούς α,b.

Ας θεωρήσουμε την ρήψη 2 νομισμάτων και τις μεταβλητές **X**, **Y** οι οποίες ορίζονται ως εξής

| | Κ | Γ |
|---|---|---|
| $\mathrm{X}(\omega)$ | 0 | 1 |
| $\mathrm{Y}(\omega)$ | 0 | 1 |

Οι **X**, **Y** είναι ανεξάρτητες εφόσον η ρήψη του κάθε νομίσματος δεν επηρεάζει την ρήψη του άλλου. Αυτό μπορούμε να το πετύχουμε ρίχνοντας τα νομίσματα σε διαφορετικά μέρη του δωματίου.

$$P(\mathrm{X}=0,\mathrm{Y}=1)=\frac{1}{4}$$

$$P(\mathrm{X}=0)\cdot P(\mathrm{Y}=1)=\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{4}$$

Μια σημαντική ιδιότητα των ανεξάρτητων τυχαίων μεταβλητών είναι η ακόλουθη

$$\mathbf{E}(\mathrm{X}\cdot\mathrm{Y}) = \mathbf{E}\mathrm{X}\cdot\mathbf{E}\mathrm{Y}$$

Η μέση τιμή των **X**, **Y** στο προηγούμενο παράδειγμα είναι 1/2 και συνεπώς η μέση τιμή του γινομένου τους θα είναι

$$\mathbf{E}(\mathrm{X}\cdot\mathrm{Y}) = 1/4.$$

Ποια είναι η μέση τιμή του γινομένου των **X**, **Y** κατά την ρήψη 2 τίμιων ζαριών, όπου **X** είναι η ένδειξη του πρώτου και **Y** η ένδειξη του δεύτερου ζαριού;

## Συνδιακύμανση (covariance)

Για δύο τυχαίες μεταβλητές **X**, **Y** ορίζουμε την συνδιακύμανση τους **cov(X,Y)** ως εξής:

$$\mathrm{cov}(\mathrm{X},\mathrm{Y}) = \mathbf{E}(\mathrm{X}-\mu_{\mathrm{X}})\cdot(\mathrm{Y}-\mu_{\mathrm{Y}}),$$

$\mu_{\mathrm{X}} = \mathbf{E}\mathrm{X}$ και $\mu_{\mathrm{Y}} = \mathbf{E}\mathrm{Y}.$

Κάνοντας πράξεις παίρνουμε

$$\mathrm{cov}(\mathrm{X},\mathrm{Y}) = \mathbf{E}(\mathrm{X}\cdot\mathrm{Y}) - \mathbf{E}\mathrm{X}\cdot\mathbf{E}\mathrm{Y}.$$

(*) Να θυμίσουμε ότι ο τύπος της διακύμανσης είναι

$$\mathrm{Var}(\mathrm{X}) = \mathbf{E}\mathrm{X}^2 - (\mathbf{E}\mathrm{X})^2.$$

Η συνδιακύμανση είναι ας πούμε μια ποσότητα που εκφράζει κατά πόσο «συσχετίζονται» οι δύο τυχαίες μεταβλητές και μπορεί να πάρει τόσο θετικές όσο και αρνητικές τιμές.

Αν οι **X**, **Y** είναι ανεξάρτητες τότε η συνδιακύμανση τους είναι μηδέν,

$$\mathrm{cov}(\mathrm{X},\mathrm{Y}) = \mathbf{E}\mathrm{X}\cdot\mathbf{E}\mathrm{Y} - \mathbf{E}\mathrm{X}\cdot\mathbf{E}\mathrm{Y} = 0$$

**Παράδειγμα.** Ας θεωρήσουμε το πείραμα της ρήψης 2 τίμιων νομισμάτων. Θα θεωρήσουμε τις τυχαίες μεταβλητές **X,Y,Z** τις οποίες ορίζουμε στο παρακάτω πινακάκι

| | ΚΚ | ΚΓ | ΓΚ | ΓΓ |
|---|---|---|---|---|
| X | 1 | 1 | 0 | 0 |
| Y | 1 | 0 | 1 | 0 |
| Z | 2 | 1 | 1 | 0 |

Οι **X,Y** είναι ανεξάρτητες γιατί η **X** αφορά το αποτέλεσμα του «πρώτου» νομίσματος ενώ η **Y** το αποτέλεσμα του «δεύτερου».

$$\left.\begin{aligned}\mathbf{E}X = \mathbf{E}Y = 1/2 \\ \mathbf{E}(X\cdot Y) = \mathbf{E}X\cdot\mathbf{E}Y\end{aligned}\right\} \Rightarrow \operatorname{cov}(X,Y) = 0$$

Οι **X,Z** όμως δεν είναι ανεξάρτητες, αν η **Z** έχει την τιμή 2, τότε η **X** έχει αναγκαστικά την τιμή 1. Για να υπολογίσουμε την **cov(X,Z)** κάνουμε πράξεις,

$$\left.\begin{aligned}\mathbf{E}(X\cdot Z) &= 2\cdot P(X\cdot Z=2)+1\cdot P(X\cdot Z=1)+0\cdot P(X\cdot Z=0)\\ &= 2\cdot\frac{1}{4}+1\cdot\frac{1}{4}=\frac{3}{4}\\ \mathbf{E}Z &= 2\cdot P(Z=2)+1\cdot P(Z=1)+0\cdot P(Z=0)=1\\ \mathbf{E}X &= 1/2\end{aligned}\right\} \Rightarrow \operatorname{cov}(X,Z) = \frac{1}{4}$$

## Συντελεστής (γραμμικής) συσχέτισης (correlation coefficient)

Δεν είναι όμως η συνδιακύμανση η ποσότητα που χρησιμοποιούμε για να εκφράσουμε κατά πόσο συσχετίζονται δύο τυχαίες μεταβλητές **X,Y** αλλά ο συντελεστής συσχέτισης **ρ(X,Y)** που ορίζεται ως εξής,

$$\rho(\mathrm{X},\mathrm{Y}) = \frac{\mathrm{cov}(\mathrm{X},\mathrm{Y})}{\sigma(\mathrm{X})\cdot\sigma(\mathrm{Y})}$$

με την προϋπόθεση οι τυπικές αποκλίσεις **σ(X)** και **σ(Y)** να μην είναι μηδέν.

Ο συντελεστής συσχέτισης παίρνει τιμές από -1 έως 1, δηλαδή ισχύει

$$-1 \le \rho(\mathrm{X},\mathrm{Y}) \le 1$$

$$\begin{aligned}
\rho(\mathrm{X},\mathrm{Y}) = 0 &\rightarrow \mathrm{X},\mathrm{Y} \quad \text{ασυσχέτιστες} \\
\rho(\mathrm{X},\mathrm{Y}) > 0 &\rightarrow \mathrm{X},\mathrm{Y} \quad \text{θετικά συσχετισμένες} \\
\rho(\mathrm{X},\mathrm{Y}) < 0 &\rightarrow \mathrm{X},\mathrm{Y} \quad \text{αρνητικά συσχετισμένες}
\end{aligned}$$

Στην πράξη όταν δύο μεταβλητές είναι θετικά συσχετισμένες και η μία από τις δύο είναι μεγάλη, αυτό ενισχύει την πιθανότητα και η άλλη να είναι μεγάλη. Όταν είναι αρνητικά συσχετισμένες και η μία από τις δύο είναι μεγάλη, τότε αυτό ενισχύει την πιθανότητα η άλλη να είναι μικρή. Όταν είναι ασυσχέτιστες και γνωρίζουμε κάτι για την **X** δεν μπορούμε να πούμε κάτι για την **Y**.

**X,Y** ανεξάρτητες συνεπάγει **X,Y** ασυσχέτιστες (το αντίστροφο δεν ισχύει)

Όσο πιο κοντά είμαστε στο 1 ή -1 τόσο μεγαλώνει και η σιγουριά μας όταν πιθανολογούμε για την τιμή της **Y** όταν γνωρίζουμε την τιμή της **X**.

Στην ακραία περίπτωση όπου **ρ(X,Y)=1** ή **-1,** τότε οι **X,Y** συνδέονται με μία σχέση της μορφής **Y=αX+β,** δηλαδή αν γνωρίζουμε την τιμή της **X** γνωρίζουμε με πιθανότητα **1** την τιμή της **Y**.

**Παράδειγμα.** Ας πάμε πάλι στο πείραμα της ρήψης 2 τίμιων νομισμάτων και τις τυχαίες μεταβλητές **X,Y,Z** τις οποίες ορίσαμε πριν

| | ΚΚ | ΚΓ | ΓΚ | ΓΓ |
|---|---|---|---|---|
| X | 1 | 1 | 0 | 0 |
| Y | 1 | 0 | 1 | 0 |
| Z | 2 | 1 | 1 | 0 |

Η συνδιακύμανση των **X,Y** είναι όπως είδαμε μηδέν και συνεπώς ο συντελεστής συσχέτισης θα είναι και αυτός μηδέν,

$$\mathrm{cov}(\mathrm{X},\mathrm{Y})=0 \Rightarrow \rho(\mathrm{X},\mathrm{Y})=\frac{0}{\sigma(\mathrm{X})\cdot\sigma(\mathrm{Y})}=0$$

Οι τυπικές αποκλίσεις των **X,Y** είναι 1/2.

Ο συντελεστής συσχέτισης των **X,Z** θα είναι

$$\mathrm{cov}(\mathrm{X},\mathrm{Z})=\frac{1}{4} \Rightarrow \rho(\mathrm{X},\mathrm{Z})=\frac{1/4}{\sigma(\mathrm{X})\cdot\sigma(\mathrm{Z})}=\frac{1/4}{(1/2)\cdot(1/\sqrt{2})}=\frac{\sqrt{2}}{2}$$

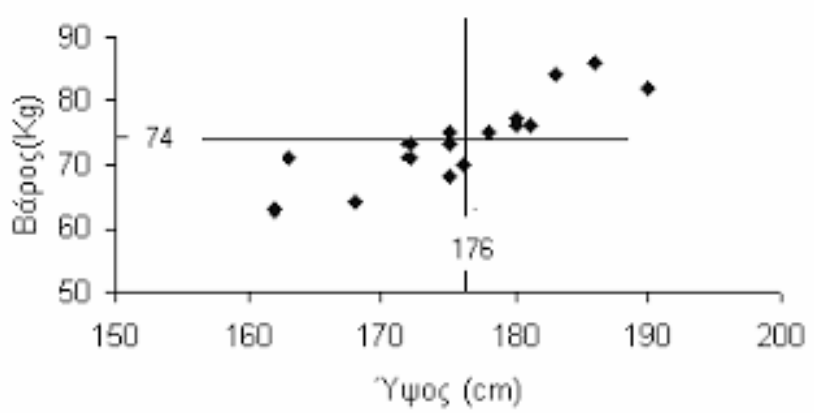
Αναμενόμενο να έχουμε θετική συσχέτιση. Για «μεγάλες» τιμές της **X** αναμένουμε «μεγάλες» τιμές της **Z**. Θεωρήστε την μεταβλητή **Z'** η οποία μετράει τα «Γράμματα». Τώρα ποια είναι η εκτίμηση σας για το πρόσημο της **ρ(X,Z')**;

# Διάγραμμα διασποράς (scatter plot)

Ας υποθέσουμε ότι εκτελούμε ένα πείραμα πολλές φορές και παίρνουμε ζεύγη τιμών (x,y) δύο τυχαίων μεταβλητών **X,Y**, οπότε προκύπτει ένα πλήθος **v** δειγμάτων

$$(x_1, y_1), (x_2, y_2), \ldots, (x_v, y_v).$$

Φανταστείτε για παράδειγμα ότι μετράτε το βάρος και το ύψος των Ελλήνων. Τα δείγματα αυτά που είναι ζευγάρια αριθμών μπορούμε να τα αποτυπώσουμε στο επίπεδο και να πάρουμε το διάγραμμα διασποράς, όπως φαίνεται στο διπλανό σχήμα. Σε κάθε σημείο αντιστοιχεί και ένα δείγμα.



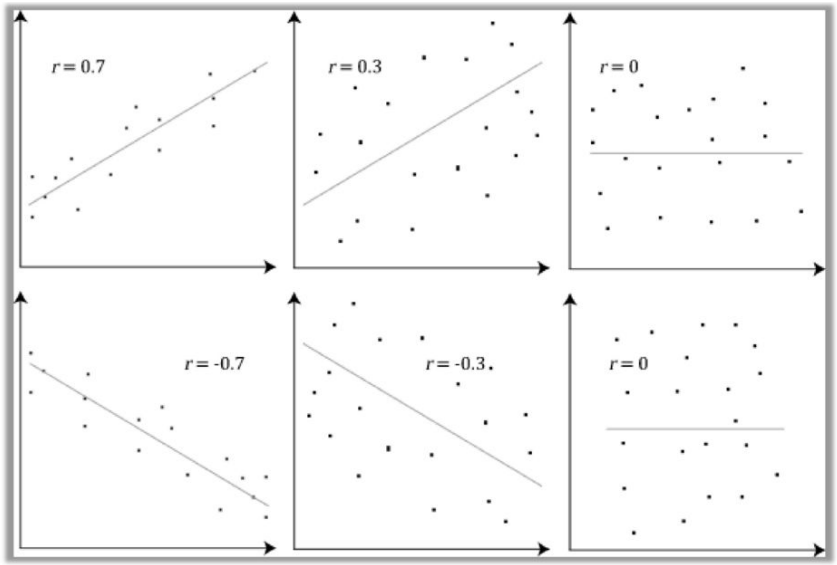

Όσο μεγαλύτερο είναι το **v** τόσο περισσότερα πράγματα μας λέει το διάγραμμα διασποράς για την κατανομή του ζεύγους **(X,Y)**.

Μπορούμε από το διάγραμμα διασποράς να βγάλουμε συμπεράσματα για την συσχέτιση των **X,Y**; Η απάντηση είναι συνήθως ναι, αρκεί να έχουμε αρκετά μεγάλο πλήθος δειγμάτων.

Ο συντελεστής συσχέτισης μας λέει κατά πόσον οι δύο μεταβλητές σχετίζονται μέσω μιας σχέσης της μορφής

$$\mathrm{Y} = \alpha \cdot \mathrm{X} + \beta$$

και όπως γνωρίζουμε δύο μεταβλητές που συνδέονται με μια τέτοια σχέση σχηματίζουν μία ευθεία γραμμή στο επίπεδο.



Μπορούμε να εκτιμήσουμε τον συντελεστή συσχέτισης δύο μεταβλητών **X,Y** από τα δείγματα που έχουμε συλλέξει αντικαθιστώντας στον τύπο

$$\rho(\mathrm{X},\mathrm{Y}) = \frac{\mathrm{cov}(\mathrm{X},\mathrm{Y})}{\sigma(\mathrm{X})\cdot\sigma(\mathrm{Y})}$$

την δειγματική συνδιακύμανση **s(X,Y)** και τις δειγματικές τυπικές αποκλίσεις **s(X)**, **s(Y)**, οπότε θα πάρουμε τον τύπο του δειγματικού συντελεστή συσχέτισης **r(X,Y)**,

$$r(\mathrm{X},\mathrm{Y}) = \frac{s(\mathrm{X},\mathrm{Y})}{s(\mathrm{X})\cdot s(\mathrm{Y})}.$$

Οι τύποι είναι οι ακόλουθοι,

$$s(X,Y) = \frac{1}{\nu-1}\left[(x_1-\overline{x})(y_1-\overline{y}) + (x_2-\overline{x})(y_2-\overline{y}) + \ldots + (x_\nu-\overline{x})(y_\nu-\overline{y})\right]$$

$$s(\mathrm{X}) = \sqrt{\frac{1}{\nu-1}\left[(x_1-\overline{x})^2 + (x_2-\overline{x})^2 + \ldots + (x_\nu-\overline{x})^2\right]}$$

$$s(\mathrm{Y}) = \sqrt{\frac{1}{\nu-1}\left[(y_1-\overline{y})^2 + (y_2-\overline{y})^2 + \ldots + (y_\nu-\overline{y})^2\right]}$$

και μετά από πράξεις και απλοποιήσεις παίρνουμε τους τύπους

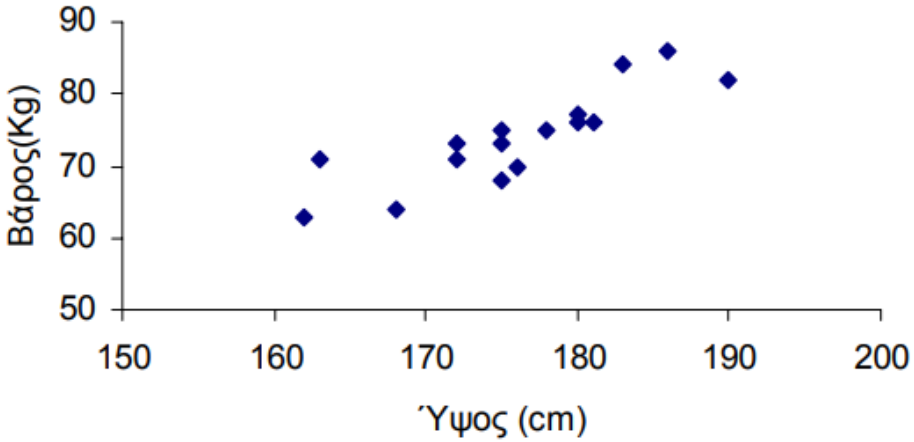
$$r(\mathrm{X},\mathrm{Y}) = \frac{\sum_{i=1}^{\nu}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{\nu}(x_i-\overline{x})^2}\cdot\sqrt{\sum_{i=1}^{\nu}(y_i-\overline{y})^2}} = \frac{\sum_{i=1}^{\nu}x_i y_i - \nu\overline{x}\cdot\overline{y}}{\sqrt{\sum_{i=1}^{\nu}x_i^2 - \nu\overline{x}^2}\cdot\sqrt{\sum_{i=1}^{\nu}y_i^2 - \nu\overline{y}^2}}$$

Για παράδειγμα θα πάρουμε τα διπλανά δείγματα από 16 εργάτες μίας βιομηχανίας.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Ύψος (cm) | 183 | 162 | 172 | 181 | 180 | 168 | 176 | 180 |
| Βάρος (Kg) | 84 | 63 | 71 | 76 | 77 | 64 | 70 | 76 |

| | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|
| Ύψος (cm) | 190 | 175 | 178 | 175 | 186 | 172 | 175 | 163 |
| Βάρος (Kg) | 82 | 68 | 75 | 73 | 86 | 73 | 75 | 71 |

Το διάγραμμα διασποράς φαίνεται στο διπλανό σχήμα και εκτιμούμε αμέσως μία θετική συσχέτιση μεταξύ του ύψους και του βάρους, όπως άλλωστε ήταν και αναμενόμενο.




Μπορείτε να κάνετε τις πράξεις και να τον υπολογίσετε ή να χρησιμοποιήσετε ένα πρόγραμμα επεξεργασίας δεδομένων, όπως είναι το PSPP ή το SPSS.

Πολλές φορές όταν δεν έχουμε έναν υπολογιστή ίσως χρειαστεί να κάνουμε πράξεις φτιάχνοντας ένα τέτοιο πινακάκι

| $x_i$ | $y_i$ | $x_i-\overline{x}$ | $y_i-\overline{y}$ | $(x_i-\overline{x})^2$ | $(y_i-\overline{y})^2$ | $(x_i-\overline{x})(y_i-\overline{y})$ |
|---|---|---|---|---|---|---|
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $\sum x_i$ | $\sum y_i$ | | | $\sum (x_i-\overline{x})^2$ | $\sum (y_i-\overline{y})^2$ | $\sum (x_i-\overline{x})(y_i-\overline{y})$ |

και εφαρμόζοντας τον τύπο

$$r(\mathrm{X},\mathrm{Y})=\frac{\sum_{i=1}^{\nu}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{\nu}(x_i-\overline{x})^2}\cdot\sqrt{\sum_{i=1}^{\nu}(y_i-\overline{y})^2}}.$$

Ή μπορούμε να φτιάξουμε το πινακάκι

| $x_i$ | $y_i$ | $x_i^2$ | $y_i^2$ | $x_iy_i$ |
|---|---|---|---|---|
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $\sum x_i$ | $\sum y_i$ | $\sum x_i^2$ | $\sum y_i^2$ | $\sum x_iy_i$ |

και να εφαρμόσουμε τον τύπο

$$r(\mathrm{X},\mathrm{Y})=\frac{\sum_{i=1}^{\nu}x_iy_i-\nu\overline{x}\cdot\overline{y}}{\sqrt{\sum_{i=1}^{\nu}x_i^2-\nu\overline{x}^2}\cdot\sqrt{\sum_{i=1}^{\nu}y_i^2-\nu\overline{y}^2}}$$

**Παράδειγμα 1.**

| $x_i$ | $y_i$ | $x_i - \bar{x}$ | $y_i - \bar{y}$ | $(x_i - \bar{x})^2$ | $(y_i - \bar{y})^2$ | $(x_i - \bar{x}) \cdot (y_i - \bar{y})$ |
|---|---|---|---|---|---|---|
| 1 | 4 | -2 | 4 | 4 | 16 | -8 |
| 2 | 2 | -1 | 2 | 1 | 4 | -2 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | -2 | 1 | -2 | 1 | 4 | -2 |
| 5 | -4 | 2 | -4 | 4 | 16 | -8 |
| $\sum x_i = 15$ | $\sum y_i = 0$ | | | $\sum (x_i - \bar{x})^2 = 10$ | $\sum (y_i - \bar{y})^2 = 40$ | $\sum (x_i - \bar{x}) \cdot (y_i - \bar{y}) = -20$ |

$\bar{x} = 3,\ \bar{y} = 0$

$$r = \frac{\sum_{i=1}^{5} (\chi_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{5} (\chi_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{5} (y_i - \bar{y})^2}} = \frac{-20}{\sqrt{10} \cdot \sqrt{40}} = -1$$

**Παράδειγμα 2.**

| $x_i$ | $y_i$ | $x_i^2$ | $y_i^2$ | $x_i \cdot y_i$ |
|---|---|---|---|---|
| 1 | -2 | 1 | 4 | -2 |
| 3 | 0 | 9 | 0 | 0 |
| 5 | 1 | 25 | 1 | 5 |
| 7 | 3 | 49 | 9 | 21 |
| 9 | 5 | 81 | 25 | 45 |
| 10 | 6 | 100 | 36 | 60 |
| 12 | 8 | 144 | 64 | 96 |
| 13 | 10 | 169 | 100 | 130 |
| $\sum x_i = 60$ | $\sum y_i = 31$ | $\sum x_i^2 = 578$ | $\sum y_i^2 = 239$ | $\sum x_i \cdot y_i = 355$ |

$$\bar{x} = 7{,}5 \quad \kappa\alpha\iota \quad \bar{y} = 3{,}9$$

$$r = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\sum_{i=1}^{\nu} x_i y_i - \nu \cdot \bar{\chi} \cdot \bar{y}}{\sqrt{\sum_{i=1}^{\nu} \chi_i^2 - \nu \cdot \bar{\chi}^2} \cdot \sqrt{\sum_{i=1}^{\nu} y_i^2 - \nu \cdot \bar{y}^2}} = \frac{355 - 8 \cdot 7{,}5 \cdot 3{,}9}{\sqrt{578 - 8 \cdot 7{,}5^2} \cdot \sqrt{239 - 8 \cdot 3{,}9^2}} = 0{,}99$$