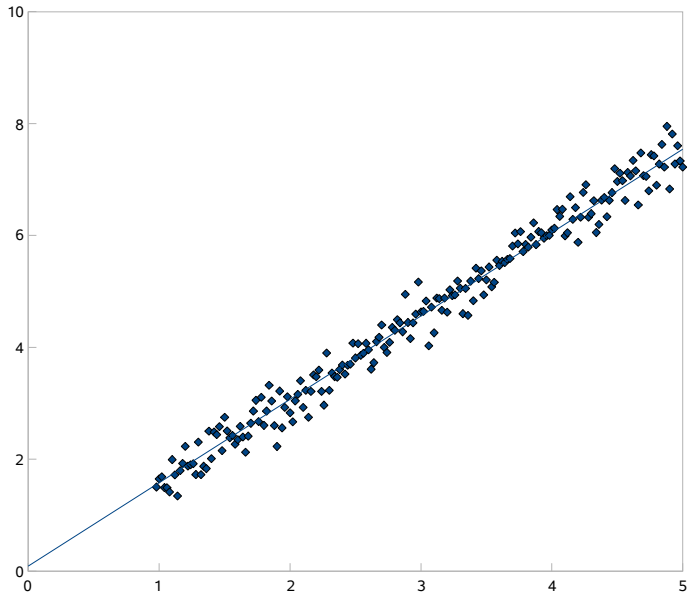
Συσχέτιση μεταξύ δύο συνόλων δεδομένων

Διαγράμματα διασποράς (scattergrams)

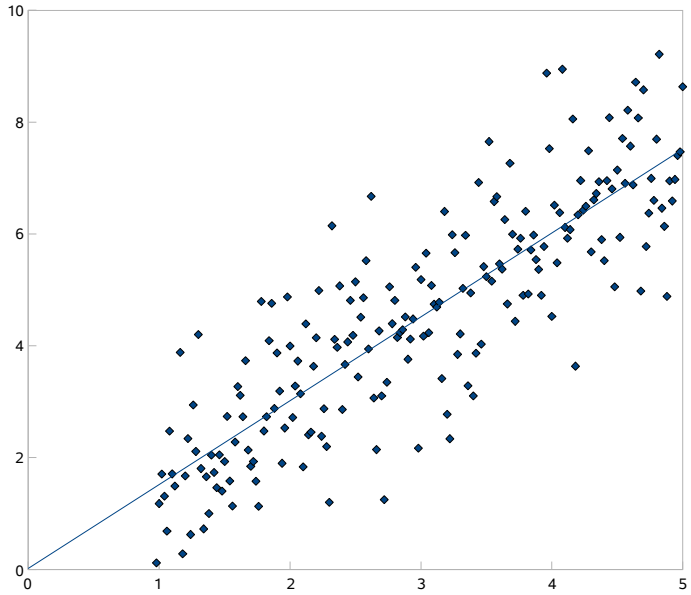
Η οπτική απεικόνιση δύο συνόλων δεδομένων μπορεί να αποκαλύψει με παραστατικό τρόπο πιθανές τάσεις και μεταξύ τους συσχετίσεις, οι οποίες μπορεί να έχουν

- *γραμμική*
- *μη γραμμική* μορφή

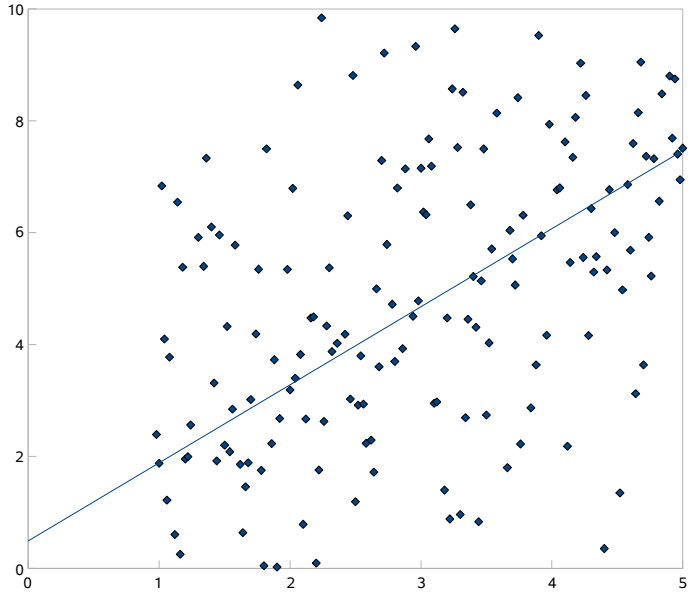
Όσο πιο *ισχυρή* η συσχέτιση τόσο μεγαλύτερη η σαφήνεια με την οποία αναδεικνύεται ως ευθεία ή καμπύλη, τόσο πιο κοντά σε αυτή τη γραμμή τείνουν να βρίσκονται τα σημεία.



Σχήμα 1 Σύνολα δεδομένων με ισχυρή θετική γραμμική συσχέτιση.



Σχήμα 2 Σύνολα δεδομένων με ασθενή θετική γραμμική συσχέτιση



Σχήμα 3 Σύνολα δεδομένων που είναι αμφίβολο αν συσχετίζονται, είτε γραμμικά είτε μη γραμμικά.

Γραμμική παλινδρόμηση και συντελεστής συσχέτισης

Όταν δύο σύνολα μετρήσεων συσχετίζονται γραμμικά, μπορούμε να εκφράσουμε ποσοτικά το πόσο ισχυρή είναι η συσχέτισή τους ορίζοντας το λεγόμενο ***συντελεστή γραμμικής συσχέτισης***. Το στατιστικό του δείγματος συμβολίζεται ως r ενώ ο συντελεστής συσχέτισης του πληθυσμού παριστάνεται με ρ.

Έστω τα σύνολα παρατηρήσεων για τις τυχαίες μεταβλητές x και y. Έχουμε υπολογίσει αντίστοιχα τις μέσες τιμές $\bar{x}$, $\bar{y}$ και τις τυπικές αποκλίσεις s_x, s_y. Για να τις φέρουμε σε ένα κοινό σύστημα αναφοράς, ορίζουμε τις ανηγμένες μεταβλητές $\tilde{x}=\frac{x-\bar{x}}{s_x}$ και $\tilde{y}=\frac{y-\bar{y}}{s_y}$ που λένε πόσο απέχει κάθε μεταβλητή από τη μέση τιμή της μετρημένη σε μονάδες τυπικής απόκλισης – και αντίστοιχα θα κάναμε με τα μ_x, σ_x και μ_y, σ_y για τον πληθυσμό αντί για το δείγμα. Όταν τα x και y συσχετίζονται θετικά ή αρνητικά, τό ίδιο θα κάνουν και οι ανηγμένες μεταβλητές γιατί απλώς μετατοπίσαμε την αρχή των αξόνων και αλλάξαμε μονάδα μέτρησης. Αλλά το ότι οι ανηγμένες μεταβλητές έχουν κοινή αρχή είναι πλεονέκτημα γιατί π.χ. αν συσχετίζονται θετικά, τότε αν η μία είναι θετική, μάλλον θα είναι και η άλλη και παρόμοια αν η μία είναι αρνητική τότε μάλλον θα είναι και η άλλη, οπότε το γινόμενό τους θα πρέπει να είναι πιο συχνά θετικό. Παρόμοια, αν συσχετίζονται αρνητικά, το γινόμενό τους θα πρέπει να είναι πιο συχνά αρνητικό.

Τότε, το γινόμενο $\tilde{x}\,\tilde{y}$ μπορεί να βοηθήσει στο να ορίσουμε ένα μέτρο της συσχέτισης μεταξύ των δύο συνόλων δεδομένων. Η πιο απλή σκέψη είναι να πάρουμε τη μέση τιμή ορίζοντας

$$\rho=\frac{1}{N}\sum_{i=1}^{N}\tilde{x}_i\,\tilde{y}_i$$

για τον πληθυσμό (εφόσον είναι πεπερασμένος). Ένας άλλος τρόπος για να καταλήξουμε στον παραπάνω ορισμό είναι ο εξής: τα σύνολα **x** και **y** των δεδομένων μας μπορούμε να τα δούμε ως ανύσματα ενός αφηρημένου χώρου N διαστάσεων. Αν ορίσουμε τα αντίστοιχα μοναδιαία ανύσματα $\hat{\boldsymbol{x}}$ και $\hat{\boldsymbol{y}}$ διαιρώντας το καθένα με το μήκος του, τότε τα δεδομένα είναι τόσο πιο συσχετισμένα όσο πιο παράλληλα είναι τα δύο ανύσματα. Ένα μέτρο αυτής της σχέσης είναι το εσωτερικό γινόμενό τους. Αν ορίσουμε τα ανύσματα **x** και **y** όχι με βάση τα αρχικά δεδομένα αλλά με τα μετατοπισμένα κατά $\bar{x}$ και $\bar{y}$, το οποίο δεν αλλάζει τίποτα επί της ουσίας, τότε το εσωτερικό γινόμενο των αντίστοιχων μοναδιαίων ανυσμάτων γράφεται ως

$$\hat{\boldsymbol{x}}\cdot\hat{\boldsymbol{y}}=\frac{\left(x_1-\bar{x}\,,x_2-\bar{x}\,,\ldots\right)}{\sqrt{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots}}\cdot\frac{\left(y_1-\bar{y}\,,y_2-\bar{y}\,,\ldots\right)}{\sqrt{(y_1-\bar{y})^2+(y_2-\bar{y})^2+\ldots}}=$$
$$\frac{\left((x_1-\bar{x})(y_1-\bar{y})+(x_2-\bar{x})(y_2-\bar{y})+\ldots\right)}{\sqrt{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\ldots}\sqrt{(y_1-\bar{y})^2+(y_2-\bar{y})^2+\ldots}}$$

Αν πολλαπλασιάσουμε και διαιρέσουμε με N τότε στον παρονομαστή θα προκύψει το γινόμενο των τυπικών αποκλίσεων, οπότε παίρνουμε τον προηγούμενο ορισμό για τον πληθυσμό. Αντίστοιχα, για να πάρουμε τα s_x, s_y στον παρονομαστή αντί των διασπορών του συνολικού πληθυσμού, θα πολλαπλασιάσουμε και διαιρέσουμε με n-1, καταλήγοντας στη σχέση:

$$r=\frac{1}{n-1}\sum_{i=1}^{n}\tilde{x}_i\,\tilde{y}_i$$

που είναι το στατιστικό του συντελεστή συσχέτισης για πεπερασμένο δείγμα.

Από την παραπάνω εξαγωγή με τη μορφή εσωτερικού γινομένου προκύπτει και μια ενδιαφέρουσα ιδιότητα: επειδή το εσωτερικό γινόμενο είναι το μήκος της προβολής του ενός ανύσματος πάνω στο άλλο και χρησιμοποιήσαμε μοναδιαία ανύσματα συνεπάγεται ότι ο συντελεστής συσχέτισης θα είναι πάντα μεταξύ -1 και 1:

- $|\rho|$, $|r| <= 1$.
- Τιμή = 1 σημαίνει απόλυτη συσχέτιση
- Τιμή =-1 σημαίνει πλήρη *αντισυσχέτιση*
- Τιμή = 0 σημαίνει απόλυτη έλλειψη συσχέτισης

Από τον ορισμό του συντελεστή συσχέτισης μπορούμε να βρούμε ένα γραμμικό μοντέλο που περιγράφει τη συσχέτιση των δύο συνόλων μετρήσεων με ακρίβεια που περιγράφεται από την τιμή του συντελεστή. Πράγματι, αν θεωρήσουμε το μοντέλο

y = ax + b

τότε βρίσκουμε ότι

$$y-\bar{y}=ax+b-a\,\bar{x}-b=a\,(x-\bar{x})$$

και διαιρώντας κατά μέλη με s_y βρίσκουμε

$$\tilde{y}=\frac{y-\bar{y}}{s_y}=a\,\frac{x-\bar{x}}{s_x}\,\frac{s_x}{s_y}=a\,\frac{s_x}{s_y}\,\tilde{x}$$

Από τον ορισμό του συντελεστή συσχέτισης για δείγμα

$$r=\frac{1}{n-1}\sum_{i=1}^{n}\tilde{x}_i\,\tilde{y}_i=\frac{1}{n-1}\sum_{i=1}^{n}\tilde{x}_i\,a\,\frac{s_x}{s_y}\tilde{x}_i=a\,\frac{s_x}{s_y}\,\frac{1}{n-1}\sum_{i=1}^{n}\tilde{x}_i^{\,2}=a\,\frac{s_x}{s_y}\,\frac{n}{n-1}\,\bar{\tilde{x}}^2$$

Για μεγάλο δείγμα που προσεγγίζει τον πληθυσμό, ο λόγος n/(n-1) τείνει στη μονάδα. Επίσης, το μέσο τετράγωνο της ανηγμένης μεταβλητής μια σταθερή ποσότητα και, υποθέτοντας κανονική κατανομή του πληθυσμού, μπορεί να δειχθεί ότι τείνει στη μονάδα όταν το δείγμα είναι αρκετά μεγάλο. Τότε,

$$r=a\,\frac{s_x}{s_y}=\frac{\tilde{y}}{\tilde{x}}$$

από όπου εύκολα βρίσκουμε ότι

$$y=r\,\frac{s_y}{s_x}\,x+\bar{y}-r\,\frac{s_y}{s_x}\,\bar{x}$$

δηλαδή, για τους συντελεστές του γραμμικού υποδείγματος ισχύει για μεν την κλίση

$$a = r\frac{s_y}{s_x}$$

για δε, την αποτέμνουσα

$$b = \bar{y} - a\,\bar{x}$$

όπως και θα περιμέναμε.

Αν η κοινή κατανομή των μεταβλητών δεν είναι η διδιάσταση κανονική κατανομή, τότε ισχύει γενικότερα

$$a = \frac{r}{\langle \tilde{x}^2 \rangle}\frac{s_y}{s_x}$$

Γραμμική παλινδρόμηση – ελάχιστα τετράγωνα – πρόβλεψη

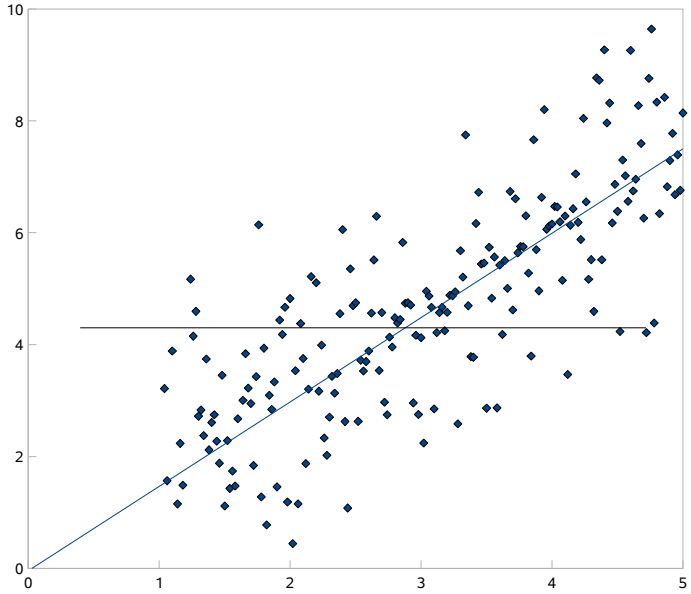
Είδαμε πώς μπορούμε να εκτιμήσουμε ποσοτικά τη γραμμική συσχέτιση (αν υπάρχει) μεταξύ δύο συνόλων δεδομένων όταν έχουν 1 με 1 αντιστοιχία. Από το συντελεστή συσχέτισης μπορέσαμε να εξάγουμε και την εξίσωση μιας ευθείας που αποτελεί το ακριβέστερο γραμμικό μοντέλο των δεδομένων. Αυτός δεν είναι ο μόνος τρόπος για να καταστρώσουμε ένα γραμμικό μοντέλο για τα δεδομένα μας. Το μοντέλο θέλουμε να είναι τέτοιο ώστε να έχει όσο το δυνατό μικρότερη απόκλιση από τις παρατηρήσεις που κάναμε. Με αυτό το σκεπτικό μπορούμε να βρούμε τους συντελεστές του μέσα από μια διαδικασία ελαχιστοποίησης μιας συνάρτησης σφάλματος.

Αν το μοντέλο που θέλουμε να προσαρμόσουμε στα δεδομένα παρασταθεί ως Y=aX+b και τα δεδομένα μας αποτελούνται από ζεύγη (x_i, y_i), τότε για κάθε παρατήρηση i το σφάλμα του μοντέλου θα είναι $Y-y_i = aX+b-y_i$. Μπορούμε να ορίσουμε μια συνάρτηση για το συνολικό σφάλμα ως το άθροισμα των τετραγώνων των σφαλμάτων: $F = \sum_{i=1}^{n}(Y-y_i)^2 = \sum_{i=1}^{n}(aX_i+b-y_i)^2$ (κανονικά, είναι η ρίζα της παραπάνω ποσότητας αλλά αυτό δεν αλλάζει την ουσία για το συγκεκριμένο θέμα). Δε μπορούμε να χρησιμοποιήσουμε τις ίδιες τις διαφορές (δηλαδή την παράσταση $\left|\sum_{i=1}^{n}(Y-y_i)\right|$) γιατί αυτές δε θα έχουν το ίδιο πρόσημο και θα αλληλοαναιρούνται δίνοντας ψευδή εντύπωση για το σφάλμα. Θα μπορούσαμε να πάρουμε το άθροισμα των απόλυτων τιμών των διαφορών, $\sum_{i=1}^{n}|Y-y_i|$ το οποίο είναι σωστό, μόνο που τα τετράγωνα είναι πιο εύκολα στο μαθηματικό χειρισμό γιατί δεν παρουσιάζουν ασυνέχεια. Σκοπός μας είναι να ελαχιστοποιήσουμε το συνολικό τετραγωνικό σφάλμα F βρίσκοντας τις καλύτερες παραμέτρους του μοντέλου, a και b. Όπως γνωρίζουμε, αυτές μπορούν να βρεθούν ελαχιστοποιώντας τη συνάρτηση F ως προς a, b, δηλαδή λύνοντας το σύστημα των εξισώσεων $\frac{\partial F}{\partial a} = \frac{\partial F}{\partial b} = 0$

Προκύπτει ότι η ευθεία ελαχίστων τετραγώνων συμπίπτει με αυτή που προκύπτει από το συντελεστή συσχέτισης. Αλλά πρέπει να καταλάβουμε ότι η παλινδρόμηση είναι πιο γενική διαδικασία γιατί η ανεξάρτητη μεταβλητή μπορεί να είναι τυχαία ή επιλεγμένη βάσει κάποιου προτύπου ενώ ο συντελεστής συσχέτισης αφορά μόνο τυχαίες μετρήσεις.

Υπάρχουν δύο ευθείες παλινδρόμησης, η $\tilde{y} = r\,\tilde{x}$ και η $\tilde{x} = r\,\tilde{y}$ (δηλαδή η $\tilde{y} = r^{-1}\tilde{x}$) που εν γένει είναι διαφορετικές αλλά προφανώς όσο ο συντελεστής συσχέτισης πλησιάζει το 1 ή -1 τόσο πιο πολύ τείνουν να ταυτιστούν.

Αν δεν υπάρχει καμία συσχέτιση, η ευθεία παλινδρόμησης βρίσκουμε αμέσως ότι είναι μία οριζόντια ευθεία $y = \bar{y}$



Σχήμα 4 Η οριζόντια γραμμή ενδέχεται να δίνει μηδενικό άθροισμα σφαλμάτων σε σχέση με τα δεδομένα (τα θετικά σφάλματα αναιρούντα από τα αρνητικά) ενώ είναι κάκιστη αναπαράσταση αυτών. Έτσι, το άθροισμα των σφαλμάτων δεν είναι η καλύτερη εκτίμηση του συνολικού σφάλματος για προσαρμογή του μοντέλου βάσει αυτής.

Μη γραμμική συσχέτιση

Μπορεί να έχουμε μικρό r αλλά να υπάρχει ισχυρή μη γραμμική συσχέτιση. Τη μη γραμμική συσχέτιση μπορούμε να διαπιστώσουμε

- μετασχηματίζοντας μία δοκιμαστική μη γραμμική σχέση που πιθανολογούμε ότι είναι το κατάλληλο υπόδειγμα, σε γραμμική και κάνοντας ***γραμμική παλινδρόμηση*** (=προσαρμογή γραμμικού μοντέλου). Π.χ. αν η σχέση που ανταποκρίνεται στα δεδομένα είναι της μορφής $Y = aX^n$ υπάρχουν δύο τρόποι για να μετατρέψουμε τη σχέση σε γραμμική: α) ορίζουμε τη μεταβλητή $Z = X^n$ οπότε έχουμε το μοντέλο $Y = aZ$ β) παίρνουμε λογαρίθμους $\log Y = \log a + n \log X$ και έχουμε το μοντέλο $Y' = A + nX'$, όπου $X' = \log X$, $Y' = \log Y$, $A = \log a$.
- κάνοντας ***μη γραμμική παλινδρόμηση***, προσαρμόζοντας το πιθανολογούμενο μοντέλο με ελάχιστα τετράγωνα

Μπορεί να οριστεί και ***μη γραμμικός συντελεστής συσχέτισης*** ως εξής:

$$r^2 = \frac{\sum (y_{est} - \bar{y})^2}{\sum (y - \bar{y})^2}$$

όπου y_{est} η ***καμπύλη παλινδρόμησης*** (το μοντέλο) που προσαρμόζουμε στα δεδομένα.

Πρακτικός υπολογισμός του r

Ο γραμμικός συντελεστής συσχέτισης μπορεί να γραφεί και ως εξής, διευκολύνοντας τον υπολογισμό του στην πράξη, π.χ. με το Excel ή το OpenOfficeCalc:

$$r = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{\left[n \sum x^2 - (\sum x)^2\right]\left[n \sum y^2 - (\sum y)^2\right]}}$$

Συσχέτιση =/= Αιτιακή σχέση!

Επειδή, βάσει της συσχέτισης δύο μεταβλητών x και y μπορούμε να χρησιμοποιήσουμε τη μία για να προβλέψουμε μια πιθανή τιμή της άλλης, λέμε ότι η y "εξηγείται" από τη x, και πιο συγκεκριμένα, το r^2 μας δίνει το ποσοστό της διασποράς του y που "εξηγείται" από το x.

Αν και έχει καθιερωθεί, αυτή η ορολογία δεν είναι ακριβής. Δεν είναι απαραίτητο να υπάρχει αιτιακή σχέση μεταξύ των x και y. Το τετράγωνο του συντελεστή συσχέτισης λέγεται και *συνδιακύμανση* των x και y.

Αξιοπιστία συντελεστή συσχέτισης (έλεγχος υποθέσεων)

Όπως και με τα άλλα στατιστικά, μπορούμε να ορίσουμε το *επίπεδο σημαντικότητας* του συντελεστή συσχέτισης.

Ακόμη και σε τυχαίο δείγμαμπορούμε να βρούμε, ασθενείς έστω, συσχετίσεις. Για να ποσοτικοποιήσουμε την εμπιστοσύνη του συντελεστή συσχέτισης, πρέπει να βρούμε τι πιθανότητα υπάρχει να προκύψει η τιμή του κατά τύχη.

Μπορούμε να υπολογίσουμε μια θεωρητική κατανομή δειγματοληψίας για το σκοπό αυτό, δηλαδή:

- υποθέτουμε τυχαία δείγματα (x, y)
- τα x, y σχηματίζουν πληθυσμό με κανονική κατανομή
- τα x, y είναι ασυσχέτιστα ($\rho = 0$)

Τότε, μπορούμε να υπολογίσουμε την πιθανότητα να λάβουμε συγκεκριμένες τιμές του r για τυχαία δείγματα με διάφορα μεγέθη ή βαθμούς ελευθερίας.

Αυτή η πιθανότητα δίνεται σε πίνακες που περιέχουν *κρίσιμες τιμές του r*. Στην αριστερή στήλη δίνονται οι βαθμοί ελευθερίας df = n-1 και οι επόμενες στήλες αντιστοιχούν σε πιθανότητες 5%, 2.5%, 1% και 0.5% να λάβουμε τιμή του συντελεστή ίση ή μεγαλύτερη από αυτή που αντιστοιχεί σε συγκεκριμένο df. Αυτές οι τιμές προκύπτουν με τη βοήθεια της κατανομής t.

Για να μπορούμε να δώσουμε συντελεστή συσχέτισης με συγκεκριμένη εμπιστοσύνη πρέπει τα δείγματα να επιλέγονται τυχαία.

Η εμπιστοσύνη μας λέει όχι μόνο τη σημαντικότητα ενός συγκεκριμένου συντελεστή αλλά δίνει και ανάλογη πληροφορία για ομάδες συντελεστών: έστω ότι αναζητούμε συσχετίσεις μεταξύ 12 μεταβλητών (π.χ. δημογραφική έρευνα που περιλαμβάνει εισόδημα, ηλικία, χρόνια εκπαίδευσης κλπ) με επίπεδο εμπιστοσύνης 5%. Οι συνδυασμοί τους είναι 66. Αν βρούμε 3 σημαντικές συσχετίσεις, αυτό δεν είναι σημαντικό αποτέλεσμα γιατί αναμένεται το 5% των συσχετίσεων, εν προκειμένω 3.3, να βρεθούν σημαντικές κατά τύχη.

Εμπιστοσύνη πρόβλεψης βάσει γραμμικού υποδείγματος

Όπως και για το συντελεστή συσχέτισης, έτσι και για τις προβλέψεις βάσει γραμμικής παλινδρόμησης, υπάρχει πιθανότητα να προκύψει κατά τύχη. Αντιμετώπιση:

* Μπορούν να οριστούν διαστήματα εμπιστοσύνης για τις προβλέψεις.

* Διασταύρωση: χρήση εξίσωσης παλινδρόμησης βάσει ενός δείγματος για προβλέψεις με βάση ένα άλλο δείγμα και σύγκριση προβλέψεων με το άλλο δείγμα (εκτιμώμενο y(x) με πραγματικό y).

Το γενικό κριτήριο ελέγχου, του χ2

Ορισμός και χρήση του χ^2

Τα κριτήρια ελέγχου (z, t, F) που έχουμε χρησιμοποιήσει ως τώρα έχουν περιορισμούς ως προς τα δείγματα (ίσες διασποράς, μέγεθος) και την κανονικότητα του πληθυσμού. Για να κάνουμε ελέγχους εκεί όπου οι προϋποθέσεις αυτές παραβιάζονται, εφαρμόζουμε το κριτήριο του χ^2.

Το χ^2 μας λέει τι πιθανότητα έχουμε να πάρουμε το συγκεκριμένο δείγμα από πληθυσμό με συγκεκριμένη κατανομή. Το χ^2 χρησιμεύει σε περιπτώσεις όπου οι άλλοι έλεγχοι δεν είναι εύκολο να εφαρμοστούν, π.χ. όταν τα δεδομένα είναι κατηγοριοποιημένα σε μη ποσοτικές κατηγορίες, ή όταν τα δεδομένα δεν έχουν κανονική κατανομή ή ίσες διακυμάνσεις.

Ορισμός:

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{(f_i - F_i)^2}{F_i} \right]$$

όπου F η προβλεπόμενη και f η παρατηρούμενη συχνότητα για μια κατηγορία. Εύκολα βλέπουμε ότι πρόκειται για άθροισμα τετραγώνων των ποσοστιαίων διαφορών συχνοτήτων, σταθμισμένων με τις θεωρητικές συχνότητες.

Παράδειγμα:

Προβλέπουμε (π.χ. με βάση στοιχεία του DNA) ότι ο τοπικός πληθυσμός θα έχει 40% μελαχροινούς, 40% ξανθούς και 20% κοκκινομάλληδες. Αν πάρουμε ένα δείγμα από 50 άτομα, περιμένουμε να βρούμε 20, 20 και 10, αντίστοιχα – *αυτό είναι η θεωρία μας ή η μηδενική υπόθεση*.

Από τυχαίο δείγμα βρίσκουμε 30, 15 και 5. Τότε υπολογίζουμε

f	F	f-F	$(f-F)^2$	$(f-F)^2/F$
30	20	10	100	5
15	20	-5	25	1.25
5	10	-5	25	2.50

Άρα, $\chi^2 = 8.75$

Έχουμε τρεις κατηγορίες, άρα 2 βαθμούς ελευθερίας.

Ανατρέχουμε σε πίνακες κρίσιμων σημείων της κατανομής του χ^2 που έχουν την ίδια λογική όπως και οι άλλοι που έχουμε δει ως τώρα και βλέπουμε για df = 2 Βλέπουμε ότι υπάρχει 2.5% πιθανότητα να προκύψει τιμή πάνω από 7.377 κατά τύχη, άρα η παρατηρούμενη διαφορά είναι σημαντική και η αρχική θεωρία απορρίπτεται.

Άλλο παράδειγμα.

Διατυπώνεται η θεωρία ότι το Σαββατοκύριακο τα τροχαία αυξάνονται. Δεν έχουμε τη δυνατότητα για σαφή αριθμητική πρόβλεψη άρα διατυπώνουμε μηδενική υπόθεση βάσει ομοιόμορφης κατανομής = κάθε μέρα ίδια πιθανότητα ατυχημάτων και συγκρίνουμε αντίστοιχη πρόβλεψη με συγκεκριμένο δείγμα. Π.χ. αν είχαμε 70 παρατηρήσεις, σύμφωνα με τη μηδενική υπόθεση θα είχαμε 10 τροχαία την ημέρα.

Προσοχή!

- Όπως είναι διατυπωμένο το κριτήριο, δεν έχει σημασία η σειρά των κατηγοριών, άρα αν τους αλλάξουμε τη σειρά, πάλι την ίδια τιμή θα πάρουμε
- Πρέπει να κυττάζουμε πάλι τα δεδομένα για να δούμε αν όντως επιβεβαιώνουν την εναλλακτική υπόθεση εκτός από το να διαψεύδουν τη μηδενική. Π.χ. δοκιμάζουμε ζιζανιοκτόνο για συγκεκριμένα φυτά και μετράμε τον πληθυσμό πριν και μετά. Η μηδενική υπόθεση θα ήταν ότι δεν υπήρχε μεταβολή στους σχετικούς πληθυσμούς. Κάθε σημαντική μεταβολή θα τη διαψεύσει, *αλλά μπορεί να μην είναι η επιθυμητή* (δηλαδή, το ζιζανιοκτόνο να σκοτώνει λάθος φυτά!)