Εφαρμογές Στατιστικών Μεθόδων σε Επιχειρηματικά Προβλήματα

Φροντιστήριο #1

Συσχέτιση δύο μεταβλητών

Θέλουμε να μελετήσουμε τη σχέση μεταξύ δύο μεταβλητών (ποσοτήτων, χαρακτηριστικών) X και Y. Ένας πρώτος απλός τρόπος για να διαπιστωθεί αν υπάρχει συσχέτιση μεταξύ των δύο μεταβλητών είναι η κατασκευή ενός διαγράμματος διασποράς (scatter plot). Το διάγραμμα διασποράς καθορίζει (προσδιορίζει) **τη πιθανή** σχέση μεταξύ των μεταβλητών X και Y. Διάφορες μορφές διαγραμμάτων διασποράς εμφανίζονται στα Σχήματα 8.1-8.3 του βιβλίου (Ι. Χαλκιάς).

Μία ποσοτική μέτρηση της έντασης της συσχέτισης μεταξύ δύο μεταβλητών X και Y μπορεί να γίνει με τη χρήση του συντελεστή συσχέτισης $r=r(X,Y)$. Ας υποθέσουμε ότι για τιμές $x_1, x_2, \dots, x_n$ που δίνουμε στη μεταβλητή X καταγράφουμε τις αντίστοιχες τιμές $y_1, y_2, \dots, y_n$ που λαμβάνει η μεταβλητή Y. Τότε, για τις μετρήσεις $x_1, x_2, \dots, x_n$ της X και $y_1, y_2, \dots, y_n$ της Y ο (δειγματικός) συντελεστής συσχέτισης $r=r(X,Y)$ ορίζεται ως εξής:

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

όπου $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ (αριθμητικός μέσος των μετρήσεων (τιμών) της X)

και $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ (ο αριθμητικός μέσος των μετρήσεων (τιμών) της Y)

Όμως,

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^{n}(x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y})$$

$$= \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i\bar{y} - \sum_{i=1}^{n}\bar{x}y_i + \sum_{i=1}^{n}\bar{x}\bar{y}$$

$$= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} - \cancel{\bar{x}n\bar{y}} + \cancel{n\bar{x}\bar{y}}$$

$$= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}$$

και

$$\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2)$$

$$= \sum_{i=1}^{n} x_i^2 - 2\sum_{i=1}^{n} x_i\bar{x} + \sum_{i=1}^{n}\bar{x}^2$$

$$= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - 2\bar{x}n\bar{x} + n\bar{x}^2$$

$$= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$= \sum_{i=1}^{n} x_i^2 - n\frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n^2} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

$$= \frac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

Ομοίως

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}y_i^2-n\bar{y}^2=\frac{n\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2}{n}$$

Μπορούμε να γράφουμε:

$$r=r(X,Y)=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sqrt{\left(n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2\right)\left(n\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2\right)}}$$

Αποδεικνύεται ότι: $-1\leq r(X,Y)\leq +1$.

- Τιμές του r που τείνουν στην τιμή -1, υποδηλώνουν έντονη αρνητική γραμμική συσχέτιση μεταξύ των X και Y.
- Τιμές του r που τείνουν στην τιμή $+1$, υποδηλώνουν έντονη θετική γραμμική συσχέτιση μεταξύ των X και Y.
- Τιμές του r που τείνουν στην τιμή 0, υποδηλώνουν μη-ύπαρξη γραμμικής συσχέτισης μεταξύ των X και Y.

– Ο συντελεστής γραμμικής συσχέτισης r καλείται δειγματικός συντελεστής γραμμικής συσχέτισης διότι υπολογίζεται από (το δείγμα) τις μετρήσεις των μεταβλητών X και Y.

– Δεν εκφράζεται σε κάποια μονάδα μέτρησης, είναι ένας καθαρός αριθμός.

Έλεγχος στατιστικής σημαντικότητας του συντελεστή γραμμικής συσχέτισης. -4-

Θέλουμε να ελέγξουμε αν η γραμμική συσχέτιση μεταξύ των X και Y είναι στατιστικά σημαντική. Βασιζόμενοι στο δείγμα των μετρήσεων των μεταβλητών X και Y θέλουμε να διερευνήσουμε αν ο συντελεστής γραμμικής συσχέτισης του πληθυσμού των μετρήσεων των μεταβλητών X και Y, ρ είναι διάφορος του μηδενός.

Οι προς έλεγχο υποθέσεις είναι:

$H_0: \rho = 0$ (Δεν υπάρχει γραμμική συσχέτιση μεταξύ των X και Y)

vs

$H_1: \rho \neq 0$ (Υπάρχει στατιστικά σημαντική γραμμική συσχέτιση μεταξύ των X και Y)

Ο έλεγχος υποθέσεων πραγματοποιείται με το κριτήριο t. Χρησιμοποιούμε τη ελεγχοσυνάρτηση:

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

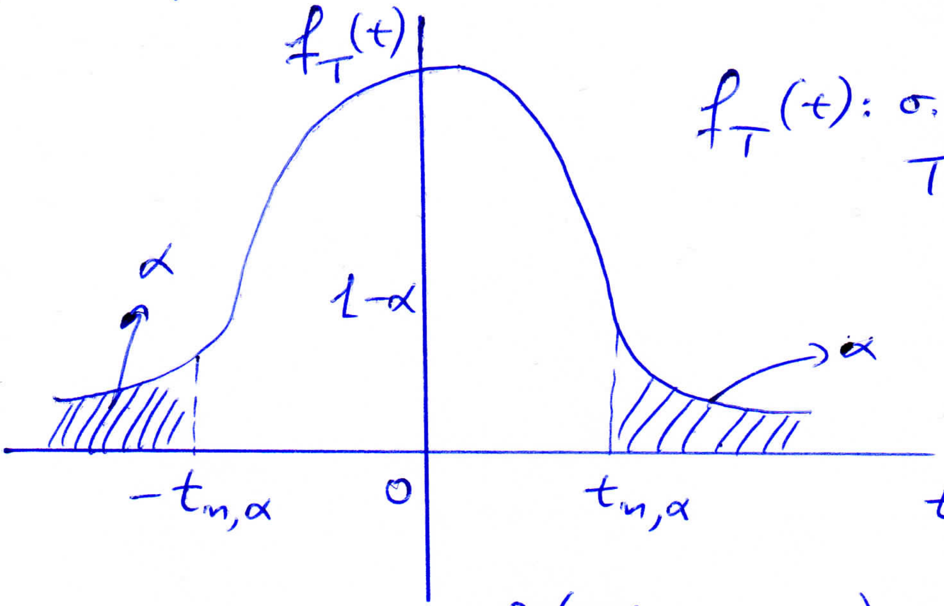
Όταν ισχύει η H_0, (κάτω από την H_0)

$$T \sim t_{n-2}$$

Απαραίτητη υπόθεση: Οι μεταβλητές X και Y να κατανέμονται τουλάχιστον οριακά σύμφωνα με την κανονική κατανομή. Το κριτήριο t δεν είναι ιδιαίτερα ευαίσθητο στην υπόθεση της κανονικότητας.

ΥΠΕΝΘΥΜΙΣΗ: Κατανομή t_n-student

n: βαθμοί ελευθερίας. Η γ.π. της σ.π. της $T \sim t_n$ έχει κωδωνοειδή μορφή, είναι συμμετρική ως προς τον κατακόρυφο άξονα στο O και έχει πιο παχιές ουρές από την $N(0,1)$ (είναι πιο πεπλατυσμένη). Όσο το n μεγαλώνει προσεγγίζεται πολύ καλά από την $N(0,1)$



$f_T(t)$: σ.π. της τ.μ. $T \sim t_n$.

Αν $T \sim t_n$ τότε $t_{n,\alpha}: P(T > t_{n,\alpha}) = \alpha$

Επίσης, λόγω συμμετρίας $t_{n,1-\alpha} = -t_{n,\alpha}$ ⊠

Απόφαση: Απορρίπτουμε την H_0 σε εσσ α αν

$$|T| > t_{n-2,\alpha/2}$$

Προσοχή!! Μία χαμηλή τιμή του r δεν σημαίνει ότι πάντα η σχέση μεταξύ των X και Y είναι ασθενής. Οι μεταβλητές X και Y μπορεί να συσχετίζονται έντονα αλλά η σχέση να είναι καμπυλόγραμμη.

Παραδείγματα

<u>#1</u> Σε 10 παιδιά ηλικίας 10-15 ετών μετρήθηκε το βάρος Y σε Kg και το ύψος X σε cm.

X	130	148	154	145	162	158	152	144	138	164
Y	34	42	45	41	46	49	43	46	36	50

Να εξετασθεί αν το βάρος είναι ή όχι ανεξάρτητο του ύψους.

<u>Λύση</u>

Μετά από υπολογισμούς έχουμε: $\bar{x}=149,5$

$\bar{y}=43,2$ $\quad \sum_{i=1}^{10}(x_i-\bar{x})^2=10,7^2$, $\quad \sum_{i=1}^{10}(y_i-\bar{y})^2=5,18^2$

$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=49,88$

Ελέγχουμε τις υποθέσεις: $H_0: \rho=0$ vs $H_1: \rho\neq 0$

όπου ρ είναι ο συντελεστής συσχέτισης για τον πληθυσμό των μετρήσεων των μεταβλητών X και Y. Υποθέτουμε ότι οι μεταβλητές X και Y κατανέμονται (έστω οριακά) κανονικά.

Χρησιμοποιούμε τη σ.σ. ελέγχου:

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

όπου n: μέγεθος δείγματος μετρήσεων (εδώ $n=10$)

και r ο δειγματικός συντελεστής συσχέτισης των X και Y. Από τα δεδομένα, έχουμε:

$$r=\frac{49,88}{\sqrt{10,7^2 \cdot 5,18^2}}=0,9$$

Συνεπώς $T=\dfrac{0,9\sqrt{8}}{\sqrt{1-0,9^2}}\simeq 5,84$

Η H_0 απορρίπτεται σε ε.σ.σ. $\alpha=5\%$, αν

$|T|>t_{n-2,\alpha/2}=t_{8,0.025}=2,306$ (από πίνακες της t_8)

Συμπέρασμα: η H_0 απορρίπτεται σε ε.σ.σ. $\alpha=5\%$ άρα το βάρος δεν φαίνεται να είναι ανεξάρτητο του ύψους. Δηλαδή τα δεδομένα παρέχουν ισχυρές ενδείξεις ότι το βάρος συσχετίζεται στατιστικά σημαντικά με το ύψος. $\boxtimes$

#2 Ένα τ.δ. $n=6$ ζευγών μετρήσεων των μεταβλητών X και Y που προέρχεται από δύο κανονικούς πληθυσμούς δίνει συντελεστή γραμμικής συσχέτισης $r=0.874$. Υπάρχει ή όχι στατιστικά σημαντική θετική συσχέτιση μεταξύ των μεταβλητών X και Y στον πληθυσμό σε εσσ $\alpha=0.01$;

Λύση Έχουμε $n=6$, $r=0.874$

Ελέγχουμε σε ε.σ.σ. $\alpha=0.01=1\%$ την

$H_0: \rho=0$
vs
$H_1: \rho>0$

Χρησιμοποιούμε τη σ.σ. ελέγχου:

$$T=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \overset{H_0}{\sim} t_{n-2}=t_4$$

Όμως
$$T=\frac{0.874\cdot 2}{\sqrt{1-0,874^2}}=\frac{1,748}{0,48}\simeq 3,64$$

Απόφαση: η H_0 απορρίπτεται σε εσσ $\alpha=1\%$ αν $T>t_{n-2,\alpha}=t_{4,0.01}=3,747$

Η ανισότητα δεν ισχύει.

Συνεπώς η H_0 δεν απορρίπτεται. Άρα μια μη-στατιστικά σημαντική θετική συσχέτιση φαίνεται ότι υπάρχει μεταξύ των μεταβλητών X και Y.

#3 Οι μηνιαίες πωλήσεις μπύρας (Y) σε εκ. βαρέλια και η μέση μηνιαία θερμοκρασία (X) σε βαθμούς κελσίου καταγράφηκαν από το Διευθυντή πωλήσεων ενός μεγάλου εργοστασίου για το χρονικό διάστημα Δεκ-Σεπτεμ. και τα παρακάτω δεδομένα συγκεντρώθηκαν:

$$\sum x_i = 115,\quad \sum y_i = 32{,}7,\quad \sum x_i^2 = 1585,\quad \sum y_i^2 = 109{,}07$$

$$\sum x_i y_i = 395{,}4$$

Να ελεγχθεί στατιστικά αν ισχύει ο ισχυρισμός ότι καθώς αυξάνει η θερμοκρασία αυξάνει και η κατανάλωση μπύρας σε επίπεδο σημαντικότητας $\alpha = 0.05$.

Λύση Θα ελέγξουμε τις υποθέσεις:

$H_0: \rho = 0$ (δεν υπάρχει γραμμική συσχέτιση μεταξύ της θερμοκρασίας και της κατανάλωσης μπύρας)

$H_1: \rho > 0$ (υπάρχει θετική γραμμική συσχέτιση μεταξύ της θερμοκρασίας και της κατανάλωσης μπύρας).

Έχουμε $n = 10$. Από τα παραπάνω δεδομένα, υπολογίζουμε το συντελεστή γραμμικής συσχέτισης r του δείγματος.

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

$$= \frac{10\cdot(395,4) - 115\cdot(32,27)}{\sqrt{10(1585) - 115^2}\sqrt{10\cdot 109,07 - 32,7^2}} = 0,816$$

Απορρίπτεται η H_0 αν $T = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} > t_{n-2,\alpha}$ για ε.σ.σ. α

Όμως $T = \dfrac{0,816\sqrt{8}}{\sqrt{1-0,816^2}} = 3,9924$

Από τους πίνακες της t_8

$t_{8,0.05} = 1,86$

Παρατηρούμε ότι $T > t_{8,0.05}$ άρα τα δεδομένα παρέχουν ισχυρές ενδείξεις ότι οι δύο μεταβλητές X και Y συσχετίζονται θετικά σε εσσ $\alpha = 5\%$.
Άρα, ισχύει ο ισχυρισμός σε $\alpha = 5\%$ ότι καθώς αυξάνει η θερμοκρασία αυξάνει και η κατανάλωση μπύρας.