Πανεπιστήμιο Αιγαίου

Ανάλυση Κατηγορικών Δεδομένων

Ενότητα 2: Πίνακες συνάφειας – Ανεξαρτησία – Έλεγχοι Ποσοστών

Στέλιος Ζήμερας

Τμήμα Μαθηματικών

Εισαγωγική Κατεύθυνση: Στατιστικής και Αναλογιστικών – Χρηματοοικονομικών Μαθηματικών

Σάμος, Ιούνιος 2015

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Πανεπιστήμιο Αιγαίου**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Πινάκες συνάφειας

- εξερεύνηση σχέσεων μεταξύ τυχαίων μεταβλητών. Είναι λογικό λοιπόν, στην ανάλυση των κατηγορικών δεδομένων να μας ενδιαφέρει η σχέση μεταξύ δύο ή περισσότερων κατηγορικών μεταβλητών.
- Έστω *X* και *Y* είναι δύο κατηγορικές μεταβλητές με την *X* να έχει *I* επίπεδα και την *Y* να έχει *J* επίπεδα. Αν κατηγοριοποιούμε ένα υποκείμενο με βάση αυτές τις δύο μεταβλητές τότε η διμεταβλητή που δημιουργείται έχει μία κατανομή. Την κατανομή αυτήν την αναπαριστούμε με ένα πίνακα συνάφειας

	Βαρύτητα συμπτωμάτων	
Φύλο	Χαμηλή	Υψηλή
Άνδρες	49	51
Γυναίκες	92	100

Πινάκες συνάφειας

- Ένας πίνακας συνάφειας ο οποίος περιγράφει τη σχέση μεταξύ δύο κατηγορικών μεταβλητών καλείται πίνακας δύο εισόδων (two-way table). Ένας πίνακας συνάφειας ο οποίος περιγράφει τη σχέση μεταξύ τριών κατηγορικών μεταβλητών καλείται πίνακας τριών εισόδων (three-way table) κ.τ.λ. Ένας πίνακας δύο εισόδων ο οποίος έχει *I* γραμμές και *J* στήλες καλείται *IxJ* πίνακας. Ο Πίνακας 2 είναι ένας 2x2 πίνακας.

Πινάκες συνάφειας

- Η από κοινού συνάρτηση πιθανότητας του συνδυασμού των ενδεχομένων *i* και *j* γράφεται:

$$\pi_{ij} = P(X = i, Y = j),\ i = 1,...,I,\ j = 1,...,J$$

όπου το $\sum_i \sum_j \pi_{ij} = 1$

Οι περιθώριες πιθανότητες των *X* και *Y* ορίζονται από τα περιθώρια αθροίσματα:

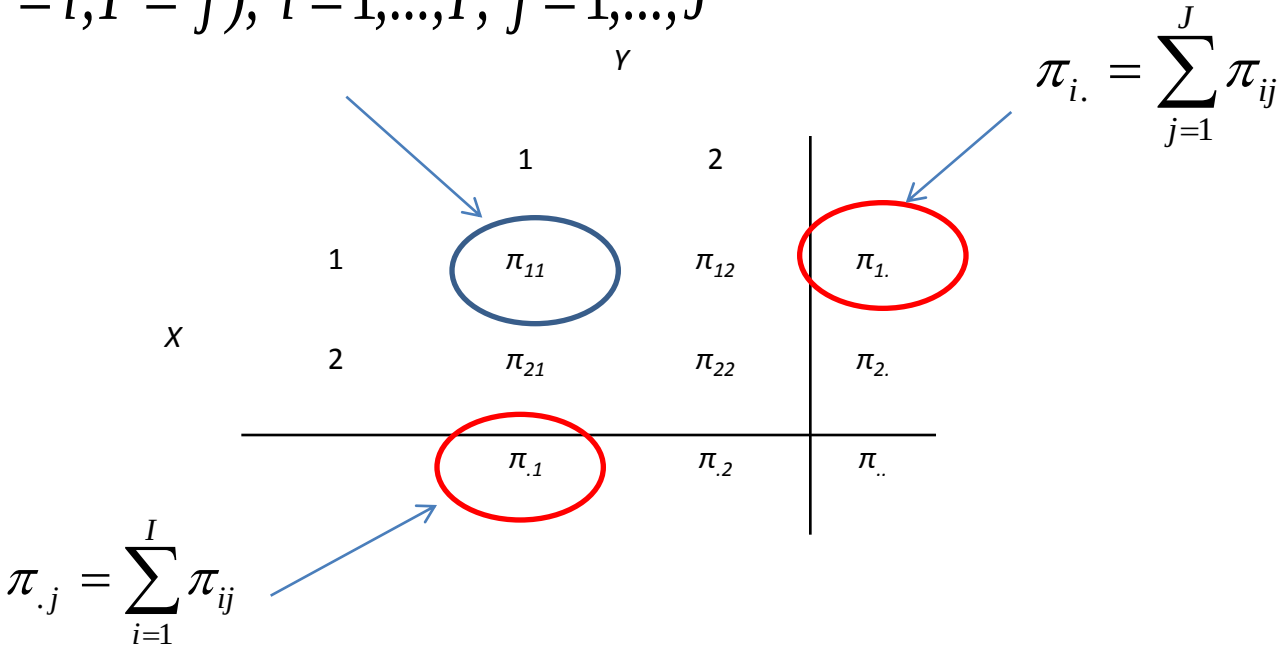
$$\pi_{i.} = \sum_{j=1}^{J} \pi_{ij} \qquad \sum_{i=1}^{I} \pi_{i.} = 1$$

$$\pi_{.j} = \sum_{i=1}^{I} \pi_{ij} \qquad \sum_{j=1}^{J} \pi_{.j} = 1$$

Πινάκες συνάφειας

Από κοινού και περιθώριες πιθανότητες

$$\pi_{ij} = P(X = i, Y = j),\ i = 1,\ldots,I,\ j = 1,\ldots,J$$



- Όμως δοθέντος ότι η *X* έχει την τιμή *i*, η *Y* έχει δεσμευμένη πιθανότητα

$$\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i.}} \qquad \sum_{j=1}^{J} \pi_{j|i} = 1$$

Ανεξαρτησία

Μία απλή σχέση (μη εξάρτησης) ανάμεσα στις *X* και *Y* είναι αυτή της ανεξαρτησίας για την οποία ισχύει:

$$\pi_{ij} = \pi_{i.}\pi_{.j} \Leftrightarrow \frac{\pi_{ij}}{\pi_{i.}} = \pi_{.j}$$

$$\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i.}}$$

$$\pi_{j|i} = \pi_{.j} \quad \forall i = 1,\ldots,I$$

Δηλαδή εάν υπάρχει ανεξαρτησία μεταξύ των μεταβλητών *X* και *Y* τότε οι δεσμευμένες πιθανότητες *P (Y = j\X = i)* είναι ίδιες για όλες τις τιμές της μεταβλητής *X* και βέβαια ίση με την *P (Y = j).*

Ανεξαρτησία

- $n_{..}$ = 292 υποκείμενα καταχωρήθηκαν σε έναν πίνακα συνάφειας 2x2 όσον αφορά το φύλο τους και τη βαρύτητα συμπτωμάτων της νόσου Alzheimer.
- $n_{1.}$ = 100 ήταν άνδρες και $n_{2.}$ = 192 γυναίκες.
- n_{12} = 51 άνδρες έδειξαν υψηλή βαρύτητα συμπτωμάτων.
- Η δειγματική εκτίμηση της από κοινού πιθανότητας ένα υποκείμενο να είναι άνδρας και να έχει υψηλή βαρύτητα συμπτωμάτων είναι p_{12} = 51/292 = 0.175.

	Βαρύτητα συμπτωμάτων		
Φύλο	Χαμηλή	Υψηλή	Σύνολο
Άνδρες	n_{11} = 49	n_{12} = 51	$n_{1.}$ = 100
Γυναίκες	n_{21} = 92	n_{22} = 100	$n_{2.}$ = 192
Σύνολο	$n_{.1}$ = 141	$n_{.2}$ = 151	$n_{..}$ = 292

$$p_{ij} = \frac{n_{ij}}{n..}$$

$$p_{12} = \frac{n_{12}}{n..}$$

Ανεξαρτησία

- Το ποσοστό της υψηλής βαρύτητας συμπτωμάτων δοθέντος ότι ένα υποκείμενο είναι άνδρας είναι 51/100 = 0.51.
- Το ποσοστό της χαμηλής βαρύτητας συμπτωμάτων δοθέντος ότι ένα υποκείμενο είναι άνδρας είναι 49/100 = 0.49.
- **Τα ποσοστά (0.51,0.49) αποτελούν τη δεσμευμένη κατανομή της βαρύτητας συμπτωμάτων δοθέντος ότι ένα υποκείμενο είναι άνδρας.**

	Βαρύτητα συμπτωμάτων		
Φύλο	Χαμηλή	Υψηλή	Σύνολο
Άνδρες	$n_{11} = 49$	$n_{12} = 51$	$n_{1.} = 100$
Γυναίκες	$n_{21} = 92$	$n_{22} = 100$	$n_{2.} = 192$
Σύνολο	$n_{.1} = 141$	$n_{.2} = 151$	$n_{..} = 292$

$$p_{j|i} = \frac{n_{ij}}{n_{i.}}$$

$$p_{2|1} = \frac{n_{12}}{n_{1.}}$$

Θεώρημα

Έστω τα X_i, i=1,2,..k έχουν πολυωνυμική κατανομή με αναμενόμενες συχνότητες np_i, i=1,2,..k. Έστω ότι Y_i, i=1,2,..k είναι ανεξάρτητες μεταβλητές που η κάθε μια ακολουθεί κατανομή Poisson με παραμετρους $\lambda_i=np_i$, i=1,2,..k. Δείξτε ότι η υπό-συνθήκη κατανομή των $Y=\sum_{i=1}^{n} Y_i = n$ είναι η πολυωνυμική κατανομή των X

Απόδειξη

Η από κοινού συνάρτηση πιθανότητας για τα n_i είναι το γινόμενο των πιθανοτήτων

$\lambda_i=np_i$

$$p(Y_1 = y_1, Y_2 = y_2, ..., Y_k = y_k) = \prod_{i=1}^{k} \left[\frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} \right]$$

Θεώρημα

Η από κοινού συνάρτηση πιθανότητας είναι το γινόμενο των πιθανοτήτων

$$p\left(Y_1 = y_1, Y_2 = y_2, ..., Y_k = y_k\right) = \left[\frac{\left(np_1\right)^{y_1} e^{-np_1}}{y_1!}\right]\left[\frac{\left(np_2\right)^{y_2} e^{-np_2}}{y_2!}\right] ..... \left[\frac{\left(np_k\right)^{y_k} e^{-np_k}}{y_k!}\right]$$

$$p\left(Y_1 = y_1, Y_2 = y_2, ..., Y_k = y_k\right) = \frac{n^{y_1+y_2+...+y_k} p_1^{y_1} p_2^{y_2} .... p_k^{y_k}}{y_1! y_2! ... y_k!} e^{-n}$$

(1)

Θεώρημα

Ισχύει ότι: $p_1 + p_2 + \ldots + p_k = 1$

Η υπό-συνθήκη κατανομή ισούται με:

$$p\left(Y_1 = y_1, Y_2 = y_2, \ldots\ldots, Y_k = y_k \mid \sum_{i=1}^{n} Y_i \right) = \frac{p\left(Y_1 = y_1, Y_2 = y_2, \ldots\ldots, Y_k = y_k\right) p\left(\sum_{i=1}^{n} Y_i \right)}{p\left(\sum_{i=1}^{n} Y_i \right)} \quad (2)$$

Από την (1) βλέπουμε ότι ο αριθμητής της υπό-συνθήκη πιθανότητας της (2) γράφεται:

$$p\left(Y_1 = y_1, Y_2 = y_2, \ldots\ldots, Y_k = y_k @ \sum_{i=1}^{n} Y_i \right) = \frac{n^n p_1^{y_1} p_2^{y_2} \ldots . p_k^{y_k}}{y_1! y_2! \ldots y_k!} e^{-n}$$

Θεώρημα

Γνωρίζουμε ότι $Y = \sum_{i=1}^{n} Y_i = n$ ακολουθεί κατανομή Poisson με παράμετρο $np_1 + np_2 + \ldots + np_\kappa = n$

Άρα ο παρονομαστής της (2) ισούται με:

$$p\left(\sum_{i=1}^{n} Y_i\right) = \frac{n^n}{n!} e^{-n}$$

Επομένως

$$p\left(Y_1 = y_1, Y_2 = y_2, \ldots\ldots, Y_k = y_k \mid \sum_{i=1}^{k} Y_i = n\right) = \frac{n!}{n_1! n_2! \ldots n_k!} p_1^{y_1} p_2^{y_2} \ldots . p_k^{y_k}$$

Πολυωνυμική Κατανομή

- Ωστόσο, μπορούμε να φτάσουμε στην πολυωνυμική κατανομή ξεκινώντας από την κατανομή Poisson. Αν έχουμε k ανεξάρτητες κατανομές Poisson $X_1, \ldots, X_k$ με παραμέτρους $m_1, \ldots, m_k$ αντίστοιχα τότε το άθροισμα τους είναι κατανομή Poisson με παράμετρο $m_1 + \ldots + m_k$. Αν δεσμεύσουμε πάνω στην τιμή του αθροίσματος $X_1 + \ldots + X_k = n$ τότε οι παρατηρήσεις X_i, $i = 1, \ldots, k$ δεν ακολουθούν κατανομή Poisson αλλά:

$$P\left(X_i = n_i, i = 1,\ldots,k \mid \sum_{j=1}^{k} X_j = n \right) = \frac{P\left(X_i = n_i, i = 1,\ldots,k \right)}{P\left(\sum_{j=1}^{k} X_j = n \right)}$$

$$= \frac{\prod_{i=1}^{k} \frac{e^{-m_i} m_i^{n_i}}{n_i!}}{\left(e^{-\sum_j m_j} \right) \frac{\left(\sum_j m_j \right)^n}{n!}} = \frac{n!}{\prod_i n_i!} \prod_i \left(\frac{m_i}{\sum_j m_j} \right)^{n_i}$$

Βοηθήματα

Γνωρίζουμε ότι $$p\left(Y_1 = y_1, Y_2 = y_2, ..., Y_k = y_k \mid \sum_{i=1}^{k} Y_i \right) = \frac{n!}{\prod_i n_i!} \prod_i \left(\frac{\lambda_i}{\sum_j \lambda_j} \right)^{y_i}$$

Γνωρίζουμε $\lambda_1 = np_1, \lambda_2 = np_2, ..., \lambda_k = np_k$

$$np_1 + np_2 + ... + np_k = n$$

Άρα

$$p\left(Y_1 = y_1, Y_2 = y_2, ..., Y_k = y_k \mid \sum_{i=1}^{k} Y_i \right) = \frac{n!}{\prod_i n_i!} \prod_i \left(\frac{\lambda_i}{\sum_j \lambda_j} \right)^{y_i}$$

np_i

$\lambda_1 + \lambda_2 + \ldots + \lambda_\kappa = n$

$$p\left(Y_1 = y_1, Y_2 = y_2, ..., Y_k = y_k \mid \sum_{i=1}^{k} Y_i \right) = \frac{n!}{n_1! n_2! \ldots n_k!} \prod_i \left(\frac{np_i}{n} \right)^{y_i}$$

$$p\left(Y_1 = y_1, Y_2 = y_2, ..., Y_k = y_k \mid \sum_{i=1}^{k} Y_i \right) = \frac{n!}{n_1! n_2! \ldots n_k!} \prod_i (p_i)^{y_i}$$

Άσκηση

Έχουμε 100 ερωτήσεις οι οποίες έχουν 4 πιθανές απαντήσεις όπου η μία είναι σωστή. Για κάθε μια ερώτηση, ο ερωτώμενος διαλέγει τυχαία μια απάντηση.

1. Προσδιορίστε την κατανομή των σωστών απαντήσεων
2. Να βρεθεί ο μέσος και η τυπική απόκλιση της κατανομής
3. Προσδιορίστε την κατανομή των (n_1, n_2, n_3, n_4)
4. Να βρεθούν: $E(n_i)$, $Var(n_i)$, $Cov(n_i n_j)$, $Cor(n_i n_j)$.

Απάντηση

- Εφόσον ο ερωτώμενος επιλέγει τυχαία μια απάντηση τότε η πιθανότητα να επιλέξει την σωστή είναι p=0.25. Αρα αν X ο αριθμός των σωστών απαντήσεων σε σύνολο n=100 ερωτήσεων τότε

$$X \sim Bin(100, 0.25)$$

$$E(X) = np = 100 * 0.25 = 25$$

$$\sigma(X) = \sqrt{Var(X)} = \sqrt{np(1-p)} = \sqrt{100(0.25)(1-0.25)} = \sqrt{18.75} = 4.33$$

- Εφόσον n μεγάλο τότε προσεγγιστικά θα ισχύει

$$X \sim N\left(np, \sqrt{np(1-p)}\right) = N(25, 4.33)$$

Κεντρικό Οριακό Θεώρημα

Απάντηση

- Επομένως

$$p(X \geq 50) = p\left(\frac{X-25}{4.33} \geq \frac{50-25}{4.33}\right) = p(z \geq 5.77) \cong 0.0000003$$

- Για κάθε μια από τις πιθανές απαντήσεις επιλέγεται μια πιθανότητα $n_1=n_2= n_3= n_4=0.25$. Αρα θα ισχύει:

$$(n_1, n_2 n_3, n_4) \sim Multi(100, 0.25, 0.25, 0.25, 0.25)$$

$$E(n_i) = np_i = np = 100 * 0.25 = 25$$

$$Var(n_i) = np_i(1 - p_i) = 100 * 0.25(1 - 0.25) = 18.75$$

$$Cov(n_i, n_j) = -np_i p_j = -100 * 0.25 * 0.25 = -6.25$$

$$Corr(n_i, n_j) = \frac{Cov(n_i, n_j)}{\sqrt{Var(n_i)}\sqrt{Var(n_j)}} = \frac{-6.25}{\sqrt{18.75 * 18.75}} = -\frac{1}{3}$$

Σύγκριση ποσοστών σε πινάκες συνάφειας

- Μεγάλα δείγματα
- Δείγμα: αποτελείται από δυο στατιστικά χαρακτηριστικά

Το μέγεθος n

Τον συνολικό αριθμό των χαρακτηριστικών του δείγματος που ικανοποιεί τις ιδιότητες των περιορισμών x

Σημειακή εκτίμηση

Το ποσοστό του πληθυσμού π εκτιμάται από το δειγματικό ποσοστό

$$\hat{\pi} = \frac{x}{n}$$

Όταν το δείγμα είναι μεγάλο η κατανομή του $\hat{\pi}$ είναι προσεγγιστικά η κανονική κατανομή με $\hat{\pi} \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$

Σύγκριση ποσοστών σε πίνακες συνάφειας

- 95% διάστημα εμπιστοσύνης ποσοστού

$$\hat{\pi} \pm z_{a/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

- Έλεγχος υποθέσεων

$H_0 : \pi=\pi_0$

$H_1 : \pi>\pi_0$

$H_1 : \pi<\pi_0$

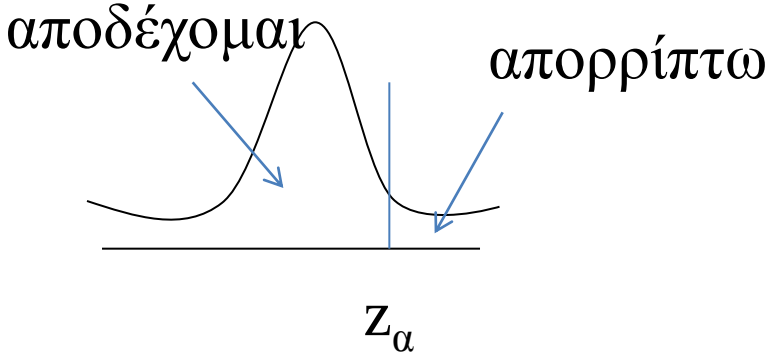
$H_1 : \pi\neq\pi_0$

$$z_* = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\hat{\pi}_0(1-\hat{\pi}_0)}{n}}} \sim N(0,1)$$

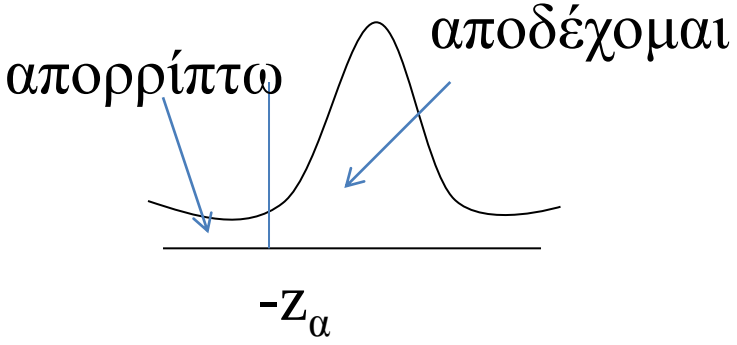
Σύγκριση ποσοστών σε πίνακες συνάφειας

Απόρριψη $H_1 : \pi > \pi_0$ αν $z_* > z_\alpha$

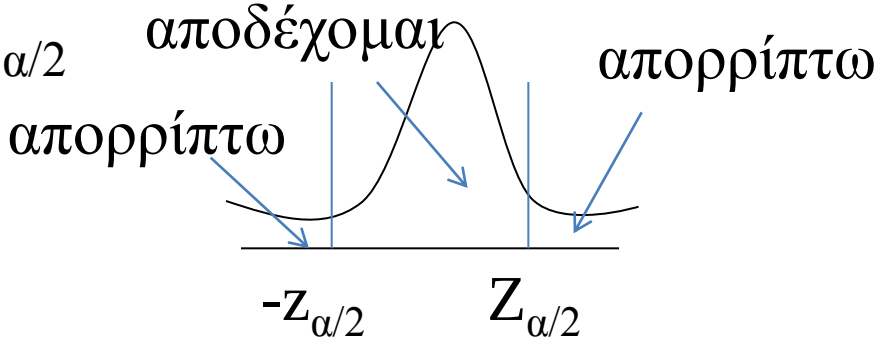
$\alpha = 0.05$ (5%)



Απόρριψη $H_1 : \pi < \pi_0$ αν $z_* < -z_\alpha$



Απόρριψη $H_1 : \pi \neq \pi_0$ αν $|z_*| \neq z_{\alpha/2}$



Άσκηση

- 30% των οδηγών αυτοκινήτων απέτυχε να μαντέψει αν τα αγωνιστικά αυτοκίνητα αγοράζονται από διαφορετικά άτομα σε σχέση με τα οικογενειακά.

Σε δείγμα 150 ατόμων, 60 απέτυχαν την δοκιμασία. Υπάρχει υπόνοια ότι το ποσοστό των αποτυχιών είναι σημαντικό ($\alpha=5\%$)

Απάντηση

Δεδομένα άσκησης

n=150, x=60

Άρα: $\hat{\pi} = \frac{x}{n} = \frac{60}{100} = 0.4$

Εκτίμηση του ποσοστού των ατόμων που απάντησαν λάθος

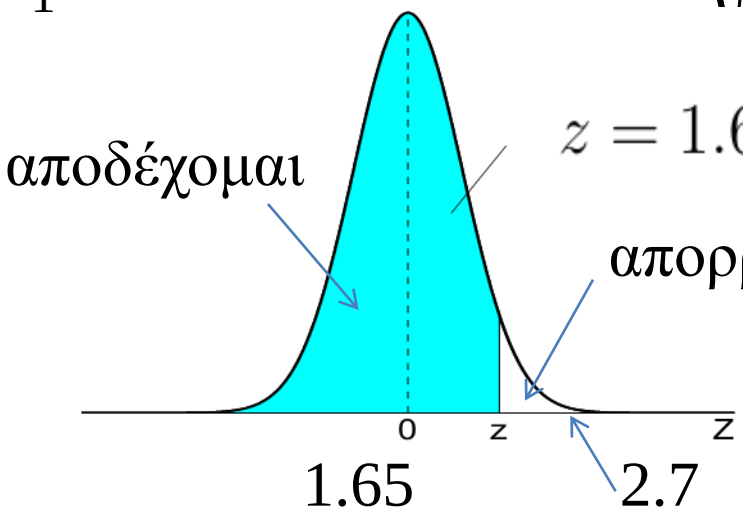
Έλεγχος υποθέσεων

$H_0 : \pi=0.3$

$H_1 : \pi>0.3$

$$z_* = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\hat{\pi}_0(1-\hat{\pi}_0)}{n}}} = \frac{0.4-0.3}{\sqrt{\frac{0.3*0.7}{150}}} = 2.7$$

$$z = 1.65 \Longleftrightarrow \Phi(z) = 0.95$$



2.7>1.65 άρα απορρίπτω την H_0

Άσκηση

- Το 1996 μέσα στα πλαίσια μεγάλης κοινωνικής μελέτης στην Αμερική τέθηκε το ερώτημα: αν οι γυναίκες που είναι έγκυος είναι σωστό να κάνουν έκτρωση στην περίπτωση που είναι παντρεμένες και δεν θέλουν άλλα παιδιά. 842 απάντησαν ναι και 982 απάντησαν όχι. Αν π είναι η πιθανότητα των ατόμων που απάντησαν θετικά να ελεγχθεί η υπόθεση H_0 : π=0.5, και να υπολογισθεί το 95% διάστημα εμπιστοσύνης της εκτίμησης του ποσοστού.

Απάντηση

Δεδομένα άσκησης

N=1824, x=842

Άρα: $\hat{\pi} = \frac{x}{n} = \frac{842}{1824} = 0.462$

Έλεγχος υποθέσεων

$H_0 : \pi = 0.5$

$H_1 : \pi < 0.5$

$$z_* = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\hat{\pi}_0(1-\hat{\pi}_0)}{n}}} = \frac{0.462 - 0.5}{\sqrt{\frac{0.5*0.5}{1824}}} = -3.26$$

$$p(z_* < -3.26) = 0.0006$$

95% διάστημα εμπιστοσύνης

$$\hat{\pi} \pm z_{a/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

$$\left[0.426 - 1.96\sqrt{\frac{0.426(1-0.426)}{1824}}, 0.426 + 1.96\sqrt{\frac{0.426(1-0.426)}{1824}}\right] = [0.439, 0.485]$$

Σύγκριση ποσοστών σε πινάκες συνάφειας 2Χ2

- Στον Πίνακα δίνεται δείγμα 419 γυναικών ταξινομημένο ως προς το αν πάσχουν από κατάθλιψη και αν είχαν κάποια τραυματική εμπειρία στη ζωή τους. Είναι το ποσοστό των γυναικών με κατάθλιψη το ίδιο για τις γυναίκες με τραυματική εμπειρία και χωρίς?

	Κατάθλιψη, *Y*		
Τραυματική εμπειρία, *X*	Όχι	Ναι	Σύνολο
Ναι	131	33	164
Όχι	251	4	255
Σύνολο	382	37	419

Σύγκριση Υ σταθερά με Χ=1 και Χ=2

Σύγκριση ποσοστών σε πινάκες συνάφειας 2Χ2

Η σύγκριση μπορεί να γίνει χρησιμοποιώντας τις δεσμευμένες πιθανότητες *Y* δοθέντος $X = 1$ και $X = 2$.

$$\pi_{y=2|x=1} - \pi_{y=2|x=2} = \frac{\pi_{12}}{\pi_{1.}} - \frac{\pi_{22}}{\pi_{2.}}$$

$$p_{j|i} = \frac{n_{ij}}{n_{i.}}$$

Η διαφορά παίρνει τιμές μεταξύ -1 και +1. Όταν η διαφορά είναι 0 τότε η απάντηση στην μεταβλητή *Y* δεν εξαρτάται από την τιμή $X = 1$ ή $X = 2$ οπότε λέμε ότι η Y είναι ανεξάρτητη της *X*. Οι πιθανότητες και συμβολίζονται και ως και αντίστοιχα.

Σύγκριση ποσοστών σε πινάκες συνάφειας 2Χ2

Η εκτίμηση της διαφοράς μεταξύ των δύο ποσοστών είναι:

$$p_{y=2|x=1} - p_{y=2|x=2} = p_{2|1} - p_{2|2} = \frac{33}{164} - \frac{4}{255} = 0.201 - 0.0156 = 0.185$$

	Κατάθλιψη, *Y*		
Τραυματική εμπειρία, *X*	Όχι	Ναι	Σύνολο
Ναι	131	33	164
Όχι	251	4	255
Σύνολο	382	37	419

Σύγκριση ποσοστών σε πινάκες συνάφειας 2Χ2

έλεγχο υποθέσεων

$H_0 : \pi_{2|1} = \pi_{2|2}$

$H_1 : \pi_{2|1} \neq \pi_{2|2}$

Κλασικός τρόπος εκτίμησης ποσοστών (όχι κατηγορικά δεδομένα)

$$p_1 - p_2 \sim N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} \right)$$

Γνωρίζουμε ότι τα ποσοστά $\hat{\pi}_{2|1}$ και $\hat{\pi}_{2|2}$ (εκτιμήσεις) ακολουθούν ασυμπτωτικά κανονική κατανομή.

$$\hat{\pi}_{2|1} - \hat{\pi}_{2|2} \sim N\left(\pi_{2|1} - \pi_{2|2}, \frac{\pi_{2|1}(1-\pi_{2|1})}{n_{1.}} + \frac{\pi_{2|2}(1-\pi_{2|2})}{n_{2.}} \right)$$

ΓΙΑΤΙ; $\hat{\pi}_{2|1} - \hat{\pi}_{2|2} \cong p_{1|1} - p_{1|2}$

τρόπος εκτίμησης ποσοστών (κατηγορικά δεδομένα)

Τα $\hat{\pi}_{j|i}$ είναι σημειακή εκτίμηση μέσα από το δείγμα αντιπροσωπεύοντας κατά 95% την πιθανότητα εκτίμησης του αντίστοιχου ποσοστού του πληθυσμού

Σύγκριση ποσοστών σε πινάκες συνάφειας 2Χ2

$H_0 : \pi_{2|1}=\pi_{2|2}$

κάτω από τη μηδενική υπόθεση ισχύει ότι:

$$\hat{\pi}_{2|1}-\hat{\pi}_{2|2} \sim N\left(\pi_{2|1}-\pi_{2|2}, \pi_{2|1}\left(1-\pi_{2|1}\right)\left(\frac{1}{n_{1.}}+\frac{1}{n_{2.}}\right)\right)$$

Κάτω από την H_0 η στατιστική συνάρτηση ελέγχου είναι

$$z_0 = \frac{p_{2|1}-p_{2|2}}{\sqrt{p_c\left(1-p_c\right)\left(\frac{1}{n_{1.}}+\frac{1}{n_{2.}}\right)}} \quad \text{με} \quad p_c = \frac{n_{1.}\cdot p_{2|1}+n_{2.}\cdot p_{2|2}}{n_{1.}+n_{2.}}$$

Κοινή συνάρτηση των $\pi_{2|1}=\pi_{2|2}$

95% διάστημα εμπιστοσύνης για την διαφορά

$$p_{2|1}-p_{2|2} \pm 1.96\sqrt{p_c\left(1-p_c\right)\left(\frac{1}{n_{1.}}+\frac{1}{n_{2.}}\right)}$$