Πανεπιστήμιο Αιγαίου

Ανάλυση Κατηγορικών Δεδομένων

Ενότητα 1: Εισαγωγή

Στέλιος Ζήμερας

Τμήμα Μαθηματικών

Εισαγωγική Κατεύθυνση: Στατιστικής και Αναλογιστικών – Χρηματοοικονομικών Μαθηματικών

Σάμος, Δεκέμβριος 2014

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

Εισαγωγή

- Αριθμητικά δεδομένα αντιστοιχούν σε πραγματοποιήσεις τυχαίων μεταβλητών των οποίων οι δυνατές τιμές **παίρνονται από ένα υποσύνολο των ακεραίων ή των πραγματικών αριθμών**.
- Ωστόσο εκτός από αυτού του τύπου τα δεδομένα μπορούμε να ξεχωρίσουμε και άλλα τα οποία προέρχονται από πραγματοποιήσεις τυχαίων μεταβλητών των οποίων **οι τιμές λαμβάνονται από ένα σύνολο που δεν αναπαριστάται απαραίτητα από αριθμούς**. Τέτοιες τυχαίες μεταβλητές για παράδειγμα, θα μπορούσαν να είναι απαντήσεις που έχουνε δώσει οι ερωτώμενοι σε ερωτήσεις κάποιου ερωτηματολογίου και υποδεικνύουν την θέση των ερωτώμενων για κάποιο εξεταζόμενο χαρακτηριστικό.

Εισαγωγή

- Τέτοιες τυχαίες μεταβλητές ονομάζονται κατηγορικές μεταβλητές και τα αντίστοιχα δεδομένα ονομάζονται κατηγορικά. Οι μετρήσεις που προκύπτουν μέσα από τέτοιες διαδικασίες καλούνται **κατηγορικές** ή **ποιοτικές**.

Εισαγωγή

- Ονοματικές (nominal), όπου η κάθε μέτρηση αντιπροσωπεύει την κατηγορία στην οποία ανήκει ο ερωτώμενος. Οι αριθμοί που θα μπορούσαν να χρησιμοποιηθούν για την αναπαράσταση των τιμών λειτουργούν σαν ετικέτες που απλά περιγράφουν τις διαφορετικές κατηγορίες. Για παράδειγμα: οικονομικές- κοινωνικές ομάδες, θρήσκευμα

Εισαγωγή

- Διατεταγμένες (ordinal), όπου οι μετρήσεις δείχνουν τη σειρά και την διάταξη των στοιχείων ή ομάδων. Οποιοσδήποτε αριθμός μπορεί να χρησιμοποιηθεί για να περιγράψει τις μετρήσεις αυτές. Θα πρέπει όμως οι αριθμοί αυτοί να διατηρούν την διάταξη των διαφορετικών κατηγοριών. Το χαρακτηριστικό θερμοκρασία θα μπορούσε να έχει τέσσερις κατηγορίες: «πολύ ζεστό», «ζεστό», «κρύο» και «πολύ κρύο» οι οποίες θα μπορούσαν να αναπαρασταθούν από τους αριθμούς 4, 3, 2 και 1 αντίστοιχα.
- **Στις περιπτώσεις ωστόσο τιμών ονομαστικών κλιμάκων αυτό δεν έχει κανένα νόημα.**

Κατανομές

Διακριτές κατανομές πιθανότητας

Έστω Χ μια διακριτή τ.μ. και υποθέτουμε ότι οι τιμές που μπορεί να πάρει είναι x_1, x_2,..... διαταγμένες σε αύξουσα σειρά.

Οι πιθανότητες να πάρει η μεταβλητή Χ τις τιμές αυτές είναι:

$$P(X=x_n)=f(x_n),\ n=1,2,\ldots$$

με σ.π.π

$f(x_n)\geq 0$

$\Sigma f(x_n)=1$

Διωνυμική Κατανομή

- Αν έχουμε ανεξάρτητες ταυτοτικές δοκιμές οι οποίες μετρούν πόσο συχνά εμφανίζεται ένα ενδεχόμενο τότε η κατάλληλη κατανομή είναι η διωνυμική. Τυπικά, το αποτέλεσμα που μας ενδιαφέρει ονομάζεται «επιτυχία». Αν η πιθανότητα επιτυχίας είναι π σε μια σειρά από N ανεξάρτητες ταυτοτικές δοκιμές, τότε ο συνολικός αριθμός επιτυχιών X έχει τη διωνυμική κατανομή με παραμέτρους N και π. Τότε γράφουμε:

$$X \sim B(N, \pi)$$

Διωνυμική Κατανομή

$$P(X=n)=\binom{N}{n}\pi^{n}(1-\pi)^{N-n}$$

$$E(X)=N\pi \qquad Var(X)=N\pi(1-\pi)$$

- Η διακριτή συνάρτηση καλείται διωνυμική κατανομή επειδή για x=0,1,2,… δίνει τους όρους του αναπτύγματος του διωνύμου:

$$(q+p)^{n}=q^{n}+\binom{n}{1}pq^{n-1}+\binom{n}{2}p^{2}q^{n-2}+....+p^{2}=\sum_{x=0}^{n}\binom{n}{x}p^{x}q^{n-x}$$

Διωνυμική Κατανομή

Παράδειγμα

Η πιθανότητα να φέρουμε 2 φορές κεφάλι σε 6 ρίψεις είναι:

$$p(x=2)=\binom{6}{2}\left(\frac{1}{2}\right)^{2}\left(\frac{1}{2}\right)^{6-2}=\frac{6!}{4!2!}\left(\frac{1}{2}\right)^{2}\left(\frac{1}{2}\right)^{4}=\frac{15}{64}$$

Κατανομή Poisson

- Φαινόμενα τα οποία έχουν πολύ μικρή πιθανότητα εμφάνισης αλλά είναι διαθέσιμος ένας μεγάλος αριθμός δοκιμών, μπορούν να μοντελοποιηθούν από την κατανομή Poisson. Για παράδειγμα, ο αριθμός των αυτοκτονιών σε ένα χρόνο. Η πιθανότητα κάποιος να αυτοκτονήσει είναι πολύ μικρή αλλά σε ένα μεγάλο πληθυσμό ένας σχετικά σημαντικός αριθμός ατόμων τελικά θα το κάνουν.
- Το Poisson μοντέλο δειγματοληψίας για τα κατηγορικά δεδομένα θεωρεί ότι στην συλλογή των δεδομένων το συνολικό μέγεθος του δείγματος *n* είναι τυχαία μεταβλητή.

Κατανομή Poisson

- Το Poisson μοντέλο υποθέτει ότι ο αριθμός των υποκειμένων σε κάθε κελί ενός πίνακα συνάφειας είναι ανεξάρτητες Poisson τυχαίες μεταβλητές. Αν συμβολίσουμε με X_i τη τυχαία μεταβλητή που μετρά τον αριθμό των υποκειμένων στο κελί *i* τότε:

$$P(X_i = n_i) = \frac{e^{-m_i} m_i^{n_i}}{n_i!} \qquad i = 1,2,...,k$$

$$E(X_i) = m_i \qquad Var(X_i) = m_i$$

Κατανομή Poisson

- Αν στην δυωνυμική κατανομή το $n \to \infty$ και η πιθανότητα p είναι μικρή τότε q=1-p $\to 1$ και

$$m = np$$

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

$$P\left(X_1 = n_1, X_2 = n_2, \ldots, X_k = n_k \mid m\right) = \prod_{i=1}^{k} \frac{e^{-m_i} m_i^{n_i}}{n_i !}$$

όπου *m* είναι το διάνυσμα των παραμέτρων των Poisson μεταβλητών. Γίνεται φανερό πως το ολικό μέγεθος δείγματος ακολουθεί Poisson κατανομή με παράμετρο $n = \sum n_i$ ακολουθεί Poisson κατανομή με παράμετρο $\sum m_i$

Κατανομή Poisson

Προσέγγιση κατανομής Poisson από διωνυμική

Δυωνυμική κατανομή:

$$P(X=n)=\binom{N}{n}\pi^{n}(1-\pi)^{N-n}$$

Θέτω λ=Nπ με π=λ/N. Επομένως

$$p(X=n)=\binom{N}{n}\left(\frac{\lambda}{N}\right)^{n}\left(1-\frac{\lambda}{N}\right)^{N-n}=\frac{\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right).....\left(1-\frac{n-1}{N}\right)}{n!}\lambda^{n}\left(1-\frac{\lambda}{N}\right)^{N-n}\Rightarrow$$

1

$$p(X=n)\to\frac{\lambda^{n}e^{-\lambda}}{n!}$$

$$\lim_{n\to\infty}\left(1-\frac{k}{n}\right)^{n}=e^{k}$$

$e^{-\lambda}$

Πολυωνυμική Κατανομή

- Κάθε ενδεχόμενο έχει πιθανότητα εμφάνισης π_j, $j = 1,..., k$. Οπωσδήποτε, $\pi_1 + ... + \pi_k = 1$. Συμβολίζουμε με X_j τον αριθμό των φορών που εμφανίστηκε το ενδεχόμενο j στις η δοκιμές. Τότε γράφουμε:

$$\left(X_1, X_{2,}, ..., X_k\right) \sim Mult\left(n, \pi_1, \pi_2, ..., \pi_\kappa\right)$$

και ισχύει:

$$P\left(X_1 = n_1, X_2 = n_2, ..., X_k = n_k\right) = \frac{n!}{n_1!...n_k!}\pi_1^{n_1}...\pi_k^{n_k}$$

Πολυωνυμική Κατανομή

Εύκολα βλέπουμε ότι:

$$E\left(X_j\right) = n\pi_j \qquad Var\left(X_j\right) = n\pi_j\left(1-\pi_j\right) \qquad Cov\left(X_i, X_j\right) = -n\pi_i\pi_j$$

Θα μπορούσε κάποιος να αναπαραστήσει τον πίνακα συνδιακύμανσης των X_i ως:

$$n\left(diag\left(\pi\right) - \pi\pi'\right)$$

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:

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

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

Πολυωνυμική Κατανομή

- Αυτή η κατανομή είναι πολυωνυμική με παραμέτρους

$$\left(n, \frac{m_1}{\sum_j m_j}, \ldots, \frac{m_k}{\sum_j m_j} \right)$$

όπου οι πιθανότητες π_i σε κάθε κελί είναι: $\frac{m_i}{\sum_j m_j}$

- Η ερμηνεία αυτού του αποτελέσματος είναι η ακόλουθη: αν έχουμε *k* κελιά και ο αριθμός των υποκειμένων θεωρείται ότι παράγεται από ανεξάρτητες κατανομές Poisson τότε δεσμεύοντας στο συνολικό μέγεθος του δείγματος παίρνουμε ότι ο αριθμός των υποκειμένων ακολουθεί την πολυωνυμική κατανομή. Ο τρόπος δειγματοληψίας λέγεται πολυωνυμική δειγματοληψία.

Δυαδικά Δεδομένα (Binary Data)

- Δυαδικά δεδομένα προέρχονται από τυχαίες μεταβλητές που παίρνουν δύο δυνατές τιμές: ναι/όχι, σωστό/λάθος, άντρας /γυναίκα. Ας υποθέσουμε ότι έχουμε *p* μεταβλητές / ερωτήσεις και *n* ερωτώμενους. Τα δεδομένα για κάθε υποκείμενο όταν $n = 5$ και $p = 9$ θα είναι για παράδειγμα της μορφής:

011010110 110011011 111000110 001100110 010101011

Δυαδικά Δεδομένα (Binary Data)

- Ένας τρόπος περιγραφής των δεδομένων αυτών είναι διαμέσου ενός πίνακα δεδομένων (data matrix) τον οποίο, προς το παρόν, μπορούμε να ονομάσουμε *X*. Κάθε γραμμή αυτού του πίνακα περιέχει τα δεδομένα κάθε υποκειμένου και αποκαλείται σχήμα απάντησης (response pattern). Υπάρχουν συνολικά 2^p δυνατά σχήματα απάντησης. Αν το *n* είναι πολύ μεγαλύτερο του 2^p πολλά από τα σχήματα απάντησης θα εμφανίζονται στο δείγμα περισσότερες από μια φορά. Αυτό μας επιτρέπει να παρουσιάσουμε τον πίνακα σαν ένα πίνακα συχνοτήτων. Ο πίνακας *X'X* διαστάσεων *(p* x *p)* περιέχει τις εξής πληροφορίες για τα δεδομένα:

Δυαδικά Δεδομένα (Binary Data)

i. Τα διαγώνια στοιχεία του πίνακα είναι το άθροισμα της κάθε στήλης του πίνακα *X,* δηλαδή ο συνολικός αριθμός των θετικών απαντήσεων για το υποκείμενο *i.*

ii. Το *(i, j)* στοιχείο του πίνακα *X'X* δίνει τον αριθμό των ερωτώμενων που απάντησαν θετικά στις ερωτήσεις *i* και *j.*

Δυαδικά Δεδομένα (Binary Data)

- Έστω ότι ο *X* πίνακας περιλαμβάνει τα δεδομένα έξι υποκειμένων, κάθε ένα εκ των οποίων απαντούν θετικά ή αρνητικά σε τρεις ερωτήσεις. Ένας τέτοιος πίνακας *X* μπορεί να είναι της μορφής:

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix}$$

Δυαδικά Δεδομένα (Binary Data)

- Ο *X′X* πίνακας είναι: $\begin{pmatrix} 3 & 2 & 0 \\ 2 & 5 & 1 \\ 0 & 1 & 1 \end{pmatrix}$

από τον οποίο καταλαβαίνουμε ότι στην πρώτη ερώτηση απάντησαν θετικά τρία υποκείμενα, ενώ στις επόμενες δύο ερωτήσεις απάντησαν θετικά πέντε και ένα υποκείμενα αντίστοιχα. Δύο υποκείμενα απάντησαν θετικά στις ερωτήσεις ένα και δύο, ενώ ένα υποκείμενο απάντησε θετικά στις ερωτήσεις δύο και τρία. Τέλος, κανένα υποκείμενο δεν απάντησε θετικά στις ερωτήσεις ένα και τρία.

Δεδομένα με Περισσότερες από Δύο Κατηγορίες (Polytomous Data)

- Ας υποθέσουμε ότι υπάρχουν *p* ερωτήσεις / μεταβλητές και $c_1, c_2, \ldots, c_p$ δηλώνουν τον αριθμό των δυνατών απαντήσεων (κατηγοριών) για κάθε μεταβλητή. Τα δεδομένα αυτά ονομάζονται πολυωνυμικά δεδομένα. Ένας τρόπος αναπαράστασης των δεδομένων αυτών είναι όπως στην περίπτωση των δυαδικών δεδομένων με τη βοήθεια αριθμών που αναπαριστούν κάθε κατηγορία. Δυστυχώς όμως μια τέτοια αναπαράσταση δε θα έδινε καμιά ερμηνεία στον πίνακα *X′X*. Προτιμότερη είναι η παρακάτω αναπαράσταση.

Δεδομένα με Περισσότερες από Δύο Κατηγορίες (Polytomous Data)

- Έστω ένα διάνυσμα από δυαδικές μεταβλητές που να υποδηλώνει ποια απάντηση (κατηγορία) έχει επιλέξει ο κάθε ερωτώμενος, π.χ. *i'* = (0,0,1,0,0) δείχνει ότι ο ερωτώμενος έχει επιλέξει την απάντηση (κατηγορία) 3 για την ερώτηση (μεταβλητή) *i.* Η διάσταση του πίνακα *X* εδώ είναι

$$n \times \sum_{i=1}^{p} c_i$$

Δεδομένα με Περισσότερες από Δύο Κατηγορίες (Polytomous Data)

Ας υποθέσουμε ότι πέντε υποκείμενα απαντούν σε δύο ερωτήσεις εκ των οποίων η πρώτη έχει δύο δυνατές απαντήσεις (κατηγορίες) και η δεύτερη τρεις. Τα δεδομένα π.χ. είναι:

21 21 12 21 13

Τότε ο πίνακας *X* είναι ένας πίνακας με πέντε γραμμές (αριθμός υποκειμένων) και πέντε στήλες εκ των οποίων οι δύο πρώτες αντιστοιχούν στις δύο δυνατές απαντήσεις της πρώτης ερώτησης ενώ οι τρεις τελευταίες στις τρεις δυνατές απαντήσεις της δεύτερης ερώτησης.

Δεδομένα με Περισσότερες από Δύο Κατηγορίες (Polytomous Data)

$$X = \left(\begin{array}{cc|ccc} 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{array}\right)$$

- Παρατηρείστε ότι ο πίνακας *X* μπορεί να αναπαρασταθεί ως $X = (X_1, X_2)$ όπου ο πίνακας X_1 είναι διάστασης 5x2 και ο πίνακας X_2 είναι διάστασης 5 x 3. Ο πίνακας $X'X$ δίνεται από

Δεδομένα με Περισσότερες από Δύο Κατηγορίες (Polytomous Data)

$$\begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix} = \left(\begin{array}{cc|ccc} 2 & 0 & 0 & 1 & 1 \\ 0 & 3 & 3 & 0 & 0 \\ \hline 0 & 3 & 3 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{array}\right)$$

- Παρατηρείστε ότι οι πίνακες και έχουν την ίδια ερμηνεία όπως και στην περίπτωση των δυαδικών δεδομένων. Δηλαδή, την πρώτη απάντηση της ερώτησης ένα την επέλεξαν δύο υποκείμενα ενώ τη δεύτερη απάντηση της ερώτησης ένα την επέλεξαν τρία υποκείμενα. Με τον ίδιο τρόπο φαίνεται πως τρία υποκείμενα επέλεξαν την πρώτη απάντηση της δεύτερης ερώτησης ενώ από ένα υποκείμενο επέλεξαν τις απαντήσεις δύο

Δεδομένα με Περισσότερες από Δύο Κατηγορίες (Polytomous Data)

- και τρία της δεύτερης ερώτησης αντίστοιχα. Ο πίνακας έχει ιδιαίτερο ενδιαφέρον και δίνεται στον Πίνακα

		X_2			
		Απάντ. 1	Απάντ. 2	Απάντ. 3	Σύνολο
X_1	Απάντ. 1	0	1	1	2
	Απάντ. 2	3	0	0	3
	Σύνολο	3	1	1	5

- Ο Πίνακας 1 είναι όπως και οι παραπάνω πίνακες ένας πίνακας συχνοτήτων και δείχνει τον αριθμό των υποκειμένων που επιλέγουν μια απάντηση από την πρώτη ερώτηση και μια απάντηση από τη δεύτερη ερώτηση. Το στοιχείο (2,1) για παράδειγμα μας δείχνει ότι τρία υποκείμενα επέλεξαν την απάντηση δύο και την απάντηση ένα στις ερωτήσεις ένα και δύο αντίστοιχα.

Δεδομένα με Περισσότερες από Δύο Κατηγορίες (Polytomous Data)

- Τέτοιοι πίνακες αναπαριστούν τη σχέση μεταξύ των δύο ερωτήσεων μιας και εάν υπάρχει εξάρτηση μεταξύ των δύο μεταβλητών θα αναμένουμε μεγαλύτερο αριθμό υποκειμένων να επιλέγει συγκεκριμένο συνδυασμό απαντήσεων από τις δύο ερωτήσεις. Πίνακες αυτού του τύπου είναι ιδιαίτερου ενδιαφέροντος και εξετάζονται στη συνέχεια.