ΕΜ272: APPLIED STATISTICS

LECTURE 1

(10/02)

- Introduce my self

- Παρουσίαση Μαθήματος. Τι είναι το Μάθημα;

↳ Εφαρμοσμένη Στατιστική (επιθ. προσδιορ: εφαρμοσμένη)

⇒ Εφαρμογές σε ρεαλιστικά προβλήματα από μια πληθώρα
Διαφορετικών Πεδίων Επιστημών

* Χρήση Υπολογιστικών Στατιστικών Πακέτων (MATLAB, SPSS, MINITAB, R ...)

Εργαστήριο: MATLAB

* Μάθημα: Αρχικά 4 ώρες θεωρία → Μετά 4 ώρες θεωρία + 2 lab
↳ Μετά 2 ώρες θεωρία + 4 lab.

* Εξέταση: Εργαστηριακές Ασκήσεις (~3) με έκθεση
+ Τελική Εξέταση (ή Τελικό Project)

* Φοιτητές: Applied Math, Math, CSD

Q → Πόσοι; , ,

- Προαπαιτούμενα: Πιθανότητες, Στατιστική.

Q: Τι έχετε κάνει;

⇒ Πιθανότητες: χώροι πιθανότητας, ανεξαρτησία, τύπος του Bayes,
Διασπορά, κατανομές (διωνυμική, Bernoulli, Poisson)

⇒ Νόμος των μεγάλων αριθμών

⇒ Στατιστική : Παραμετρικά στατιστικά μοντέλα
Στατιστικά Δείγματα
Παραμετρικοί χώροι
Διαστήματα Εμπιστοσύνης
Προβλήματα Ελέγχου Κανονικών Πληθυσμών

- Θ. Από που μπορεί να προέρχονται τα φυσικά προβλήματα;

⇒ Φυσικές επιστήμες (Φυσική, Χημεία, Βιολογία,)

⇒ Ιατρική

⇒ Οικονομικές επιστήμες

⇒ κοινωνικές

→ Παιδαγωγικά θέματα

⇒ - - - -

Προσοχή: Η εφαρμοσμένη στατιστική σημαίνει πάντα την ανάλυση δεδομένων

Applied Statistics ≡ Στατιστική Ανάλυση Δεδομένων

Απλό παράδειγμα: Κατανομή Ύψους. Έστω δεδομένα ύψους N. Βρείτε την κατανομή πιθανότητας.

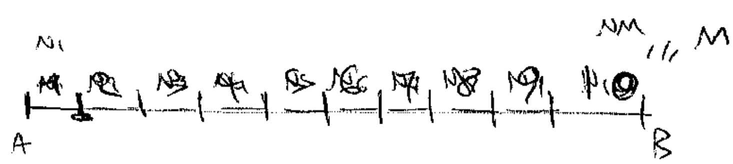
↳ Μεταβλητή x: ύψος

↳ Πιθανότητα $P(x)$

Q: Έστω δείγμα N. Πώς βρίσκουμε την πιθανότητα $P(x = y)$

$$P(x) = \frac{N(y-h < x < y)}{N}$$

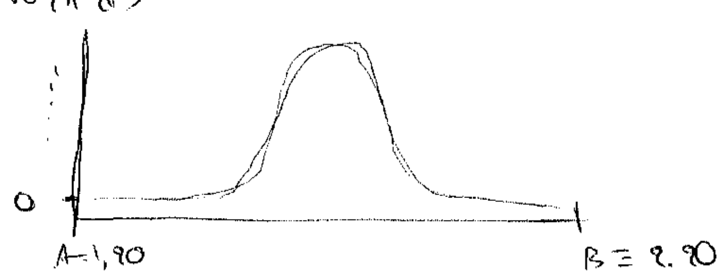
- Ορίζουμε όρια A, B, διάστημα h



$M = \frac{B-A}{h}$. e.g. $A = 1.20$, $B = 2.20$

$P_1 = P(A+h) = \frac{N_1}{N}$ . . $\sum_{i=1}^{M} P_i = \frac{N}{N} = 1$

- κατανομή πιθανότητας



$P_1 = P_0 \pm \epsilon_1$

- Χρειάζεται: Ανάλυση του προβλήματος

- Q: Επίδραση N. Πόσο είναι το σφάλμα; $\epsilon_1 = \epsilon_1(N) = ?$

π.χ. Αν $N = 1000$. Τι μπορούμε να πούμε για $N = 10.000.000$

@ : Στην πραγματικότητα η κατανομή ύψους $^{\text{συνήθως}}$ δεν είναι κανονική! Είναι περίπου κανονική.

@ : Είναι το πρόβλημα καλά ορισμένο;

Όχι! Γιατί? Δεν διευκρινίζει

EM 272 : APPLIED STATISTICS

LECTURE 2

(19/02)

- Print → share syllabus
- Introduce for the rest
- Συζήτηση ύλης.

Εισαγωγικά (Ενότητα 1)

* Ορολογία : Δεδομένα (data)

  Δειγματοληψία (sampling) : συλλογή δεδομένων

  Περιγραφική στατιστική (descriptive statistics) : περιγραφή δεδομένων

  Στατιστική συμπερασματολογία (statistical inference) : απόκτηση συμπερασμάτων

↳ Δεδομένα συλλέγονται από πληθυσμό (population)

  Δείγμα (sample) : υποσύνολο του πληθυσμού

  Τυχαία Μεταβλητή (random variable) : οποιοδήποτε ~~[illegible]~~ χαρακτηριστικό του οποίου η τιμή αλλάζει μέσα στο στοιχεία του πληθυσμού.

  (τ.μ.)

↳ ποιοτική (qualitative) : είναι κατηγορίες

  ποσοτική (quantitative) : παίρνει αριθμ. τιμές

  συνεχής (continuous) : τιμές σε ένα συνεχές διάστημα

  διακριτή (discrete) : τιμές σε ένα σύνολο διακεκριμένων τιμών

- Παραδείγματα : Δείγμα : πληθυσμός   τ.μ. : ύψος, φύλο (ποιοτική), βάρος, ηλικία (ποσοτικές)

  : » 18 ετών   : » » » , X

  [illegible] / [illegible] τ.μ. ⇒

* Περιγραφική Στατιστική

(Frequency)
- Συχνότητα, έστω δείγμα $x_1, x_2, \dots x_n$ με πιθανές τιμές (διακεκριμένες)

f a_i, $a_1, a_2, \dots a_m$

f_i : πόσες φορές εμφανίζεται η τιμή a_i στο δείγμα

- σχετική συχνότητα (relative frequency)

|||

Ποσοστό (percent) $p_i = \frac{f_i}{n}$

- Για κατηγορίες που μπορούν να μπουν σε διάταξη ορίζεται και η

Αθροιστική συχνότητα $\tilde{F}_i$ $\tilde{F}_i = \sum_{j=1}^{i} f_j$ $j < i$

Αθροιστικό ποσοστό $\tilde{P}_i = \frac{\tilde{F}_i}{n}$ $a_j < a_i$

- Παρουσίαση : πίνακα συχνοτήτων (frequency table), ραβδόγραμμα (bar chart), κυκλικό διάγραμμα (pie chart)

Παράδειγμα : Βάρος πληθυσμού

Έλλειμμα προϋπολογισμού / μήνα

1% 2% 3% 1% 4% 5% 6% 3% 8% 2% 10% 15%

$a_1 = 1\%$ $f_1 = \frac{2}{12} = 1/6$

$a_2 = 2\%$ $f_2 = \frac{2}{12} = 1/6$

$a_3 = 3\%$ $f_3 = 1/6$ $F_3 = 1/6 + 1/6 + 1/6 = 3/6 = 1/2$

: ποσοστό που το χρέος είναι έως 3% είναι 1/2

Πίνακας συχνοτήτων

a	f	F
1%	1/6	1/6
2%	1/6	2/6
3%	1/6	3/6
.	.	.

(Αντι)
Παράδειγμα : Δείγμα 100 φοιτητές . Ερώτηση: Μέσο μεταφοράς προς το πανεπιστήμιο.

→ Μπορούμε να σχηματίσουμε πίνακα - γραφήματα f_i, p_i αλλά ΟΧΙ $\tilde{F}_i$, $\tilde{P}_i$ γιατί δεν υπάρχει σχέση διατάξεως μεταξύ των τιμών της τ.μ. (μέσο μεταφοράς)

- Ομαδοποίηση Δεδομένων

Χωρισμός σε ομάδες (groups) → κλάσεις διαστημάτων

→ Υπολογισμός συχνότητας για κάθε ομάδα.

(π.χ. κατανομή ύψους)

- Πιθανότητες - Ορισμοί :

↳ Πιθανότητα να ισχύει το A $P(A)$ $P(A) \geq 0$, $\sum_A P(A) = 1$

↳ » » » (από κοινού) και B $P(A,B)$

↳ » » » αν ισχύει το B $P(A|B) = \dfrac{P(A,B)}{P(B)}$

$P(A,B) = \left(P(A|B)P(B) = P(B|A)P(A)\right)$

$P(A) = \sum_B P(A,B)$

- Ανεξάρτητες $P(A,B) = P(A)P(B)$

- Κανόνας του Bayes $P(B|A) = \dfrac{P(A|B)P(B)}{P(A)}$

↳ συνδυασμοί $\binom{n}{r} = \frac{n!}{(n-r)!\,r!}$

- Μέτρα Στατιστικών Δεδομένων

↳ Μέτρα Χώρου (Statistics of location

⇒ Μέσος όρος (Mean, Average)

Αριθμητικός → $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Γεωμετρικός $GM_x = \sqrt[n]{\prod_{i=1}^{n} x_i}$

Q: Πότε χρησιμοποιείται ο γεωμετρικός μέσος όρος;

A: (wikipedia): Όταν χρειάζεται ο υπολογισμός "average factors" (μέσων ποσοστών)

π.χ. Χρηματιστήριο με αύξηση 10% στο 1ο χρόνο, 20% στο 2ο και πτώση 15% στο 3ο.

Τότε $GM_x = \frac{1,1 \cdot 1,2 \cdot 0,85}{3} = 1,0391$: μέση αύξηση κάθε έτος.

Αρμονικός (Harmonic) $\frac{1}{H_x} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}$

$\Rightarrow H_x = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$

Q: Πότε χρησιμοποιείται ο αρμονικός μέσος όρος;

A (wikipedia): π.χ. σε ένα ταξίδι έχουμε στη μισή διαδρομή ταχύτητα 40 k/h και στην άλλη μισή 60 k/h → μέση ταχύτητα = H_x

$= \frac{2}{\frac{1}{40}+\frac{1}{60}} = 48$ k/h

⇒ Δειγματική Διάμεσος (Median)

: κεντρική τιμή όταν διατάξουμε τα δεδομένα κατ' αύξουσα σειρά

n : περιττός $\tilde{x}$: τιμή στη θέση $x_{(n+1)/2}$

n : άρτιος $\tilde{x} = \dfrac{x_{(n/2)} + x_{(n/2+1)}}{2}$

⇒ Percentile. (Δειγματικά ποσοστιαία) : Ποσοστό σημείων που βρίσκονται πριν από μια τιμή αναφοράς.

Διάταξη κατ' αύξουσα σειρά

P-th percentile is $\dfrac{(n+1)}{100/p}$

η τιμή της x στη θέση

⇒ Μέτρα Μεταβλητότητας (Statistics of dispersion)

: Διαφορετικά δείγματα μπορεί να έχουν το ίδιο μέτρο αλλά διαφορετική διασπορά

⇒ Εύρος (Range) $(= X_{max} - X_{min})$

⇒ Διακύμανση ή Διασπορά (Variance)

πληθυσμό $\sigma^2 = \dfrac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$

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

(Q : Ποια η διαφορά ?

⇒ Τυπική απόκλιση (standard deviation)

$s = \sqrt{s^2}$, $\sigma = \sqrt{\sigma^2}$

Συντελεστής τυπικής απόκλισης $V = \dfrac{\sigma \times 100}{\bar{x}} = \dfrac{s \times 100}{\bar{x}}$

Τυπικό
⇒ Σφάλμα (Standard error) $s_{\bar{x}} = \frac{s}{\sqrt{n}}$, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

⇒ Ενδοτεταρτομοριακό εύρος IR (Interquartile Range)

$$IR = \underbrace{Q_3}_{\text{75-εκατοστιαίο σημείο}} - \underbrace{Q_1}_{\text{25-εκατοστιαίο σημείο}}$$

p-εκατοστιαίο σημείο όταν p% του δείγματος είναι μικρότερες από αυτή τη παρατήρηση.

⇒ Σύνοψη 5 αριθμών (5 number summary): Διάμεσος, Q_1, Q_3
X_{min}, X_{max}

⇒ Στατιστική ροπών (Moment Statistic)

r-τη κεντρική ροπή $\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^r$

1st ⇒ $\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x}) = 0$

2nd ⇒ $= s^2$

3rd, 4th used για υπολογισμό ασυμμετρίας (skewness)
και κύρτωσης

— Λύσεις μαθήματος → έχει ανοιχτεί ?

5'

LECTURE 3 (22/02)

Είδαμε : Βασικές έννοιες
Ορολογία
Περιγραφική στατιστική

— Λίστα Γραμματεία, Μιχάλης
ΕΜ 272

Κατανομές (Distributions)

* Διακριτές (Discrete)

↳ Διωνυμική (Binomial) : Έστω δείγμα με 2 δυνατότητες, τιμές A, B
$P(A)=p$
$P(B)=q$

Πιθανότητα $P(X=x) = \binom{n}{x} p^x q^{n-x}$

Παράδειγμα : σύνολο φρούτων (μήλο (A), μπανάνα (B)),

1 φρούτο $\{A, B\}$
$\{p, q\}$
2 φρούτα $\{AA, AB, BB\}$

πιθανότητα $\{p^2, 2pq, q^2\}$ $(p+q)^n$

$$P(X=x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

μέσος $\mu = np$ $\sigma = \sqrt{npq}$ $\sigma^2 = npq$

↳ Poisson : καλή προσέγγιση της Διωνυμικής όταν το ένα γεγονός είναι σπάνιο ($\pi.\chi.$ $p<0.1$) και το μέγεθος του δείγματος μεγάλο ($n \geq 55$)

X : Διωνυμική με μεγάλο n και μικρό p $\lambda = np$

$$P(X=x) = \frac{e^{-\lambda}\lambda^x}{x!}$$ $\lambda = \mu = \sigma^2$

¿ έλεγξέ το!

Παράδειγμα : (Birthday Paradox)

Έστω n άνθρωποι. Υπάρχουν $\binom{n}{2}$ ζευγάρια. Ορίζουμε ως επιτυχία αν υπάρχει ζευγάρι το οποίο έχει κοινή ημερομηνία γέννησης

Πιθανότητα $1/365 = P(B|A)$

Αριθμός επιτυχιών $\mu = np = \binom{n}{2}/365 = \frac{n(n-1)}{2}\frac{1}{365} = \frac{n(n-1)}{730} = \lambda$

Πιθανότητα ότι δεν υπάρχουν 2 άτομα με ίδια ημερομηνία γέννησης

$$P(X=0) = e^{-\lambda}\frac{\lambda^0}{0!} = e^{-\lambda} = \exp\left(\frac{-n(n-1)}{730}\right)$$

n ώστε η πιθανότητα να είναι $< 0,5$: $\exp\left(\frac{-n(n-1)}{730}\right) \le 1/2$

$$\Rightarrow \exp\left\{\frac{n(n-1)}{730}\right\} \ge 2, \quad n(n-1) \ge 730\ln(2) \Rightarrow \underline{\underline{n \geq 23}}$$

Q : Ποια είναι η πιθανότητα αν $n=70$, $n=500$?

Q: Ποια είναι η πιθανότητα κάποιος να έχει την ίδια ημερομηνία με τη δική μας?

* Συνεχής

↳ Uniform (ομοιόμορφη)

$$f(x) = \begin{cases} \frac{1}{\beta - \alpha}, & \text{if } \alpha < x < \beta \\ 0, & \text{otherwise} \end{cases}$$

$$\mu = \frac{a+b}{2} \qquad \sigma^2 = \frac{(b-a)^2}{12}$$

↳ Κανονική (normal)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$: συμμετρική γύρω από το μέσο

$\mu \pm \sigma$ περιέχει 68.3% σημ.
$\pm 2\sigma$ " 95.5% "
$\pm 3\sigma$ " 99.7% "

→ Γάμμα : χρόνος έως ότου n γεγονότα πραγματοποιηθούν

$$f(x) = \frac{\lambda e^{-\lambda x}(\lambda x)^{n-1}}{\Gamma(n)} \quad \text{for } x \geq 0$$

$$\Gamma(t) = \int_0^\infty e^{-y} y^{n-1} dy = (n-1)!$$ για ακέραιες τιμές του n

↳ Εκθετική : χρόνος έως ότου ένα γεγονός συμβεί
(συνεχές ανάλογο της Γεωμετρικής κατανομής)

$f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ $\quad \mu = 1/\lambda \quad , \quad \sigma^2 = 1/\lambda^2$

$F(x) = 1 - e^{-\lambda x}$

$P(X > t+s \mid X > t) = P(X > s)$ $s, t \geq 0$: Δεν έχει μνήμη

↳ Chi-square : τετράγωνο σφάλματος στο n-διάστατο χώρο
χ^2

Αν Y_i είναι κανονικές κατανομές

$\sum_{i=1}^{n} Y_i^2$ ακολουθεί χ^2 κατανομή

↳ t ή student κατανομή

$\bar{X}-\mu$ ενός δείγματος είναι κανονική κατανομή

$\frac{\bar{X}-\mu}{\sigma}$ » » $\mu=0,\ \sigma=1$

αλλά $\frac{\bar{Y}_i-\mu}{S_{\bar{Y}_i}}$: πιο "πλατιά" από την κανονική κατανομή.

τυπικό σφάλμα

Εκτίμηση - Διαστήματα εμπιστοσύνης

Η κατανομή μιας τ.μ. χαρακτηρίζεται από μ : μέση τιμή
σ^2 : διασπορά

Εκτίμηση μιας παραμέτρου : σημειακή εκτίμηση (point estimate)
» διαστήματος τιμών $(1-\alpha)\%$ confidence interval :

α : (στάθμη) επίπεδο σημαντικότητας (significance level)

π.χ. Εκτίμηση μέσης τιμής μ. Χρειάζεται να γνωρίζουμε $\bar{x}$, σ^2, μορφή της κατανομής

κανονική Διαστημα εμπιστοσύνης : $\bar{x} \pm (\text{κρίσιμη τιμή}) \times S_{\bar{x}}$
τυπικό σφάλμα

$$\bar{X} \pm z_{1-\alpha/2} \cdot \frac{S}{\sqrt{n}}$$

— Εμπεριέχει την αληθινή τιμή της παραμέτρου με πιθανότητα $(1-\alpha)\%$

- Μικρά δείγματα ($n<30$): δεν μπορεί να οριστεί παραμετρικά μη κανονικής κατανομής
→ χρειάζεται μη-παραμετρική μέθοδο

- Διαφορά μέσων τιμών.

για κανονικές ($n_1, n_2 \geq 30$)

$$(\bar{X}_1 - \bar{X}_2) \pm z_{1-\alpha/2} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

ΕΜ 272 : ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

LECTURE 4 (26/02/2010)

- Λίστα : em 272
  email : majordomo@tem.uoc.gr
  subject : subscribe em272
- Lab : Τρίτη ή Πέμπτη
  15:00-18:00.
  Παρασκευή : 13:00-15:00 Η-205
  Δευτέρα: Επόμενο φορά Lab
  Ερ-σ: 5/15 : δεν έχουν κάνει καθόλου MATLAB!

Είδαμε : κατανομές

Διαστήματα εμπιστοσύνης

Confidence Intervals.

Ορίζεται

$$\bar{X} \pm (\text{κρίσιμη τιμή}) \times S_{\bar{X}}$$

$$\pm z_{1-a/2} \frac{s}{\sqrt{n}} \quad , \quad \text{κανονική} \quad , \quad n > 30$$

$$\pm t_{n-1,1-a/2} \frac{s}{\sqrt{n}} \quad , \quad \text{»} \quad , \quad n < 30$$

$z_{1-a/2}$: κρίσιμη τιμή της τυπικής κανονικής κατανομής

$t_{n-1,1-a/2}$: » » » t - » » με n-1 βαθμούς ελευθερίας.

: μη κανονική , $n < 30$: μη-παραμετρικό

(δεν μπορεί να οριστεί με βάση κάποια γνωστή κατανομή)

→ μη παραμετρική μέθοδο.

$L_1 = \bar{X} - \ldots$

$L_2 = \bar{X} + \ldots$

$$P(L_1 \leq \mu \leq L_2) = 1 - a$$

π.χ. $a = 0{,}05$ $\quad z_{1-a/2} = 1{,}96$ $\quad P(\bar{X} - 1.96 \cdot \sigma_{\bar{x}} \leq \mu \leq \bar{X} + 1.96\sigma_{\bar{x}}) = 0{,}95$

Παράδειγμα: $n = 10$ 95% εμπιστοσύνη $(\alpha = 0.05)$

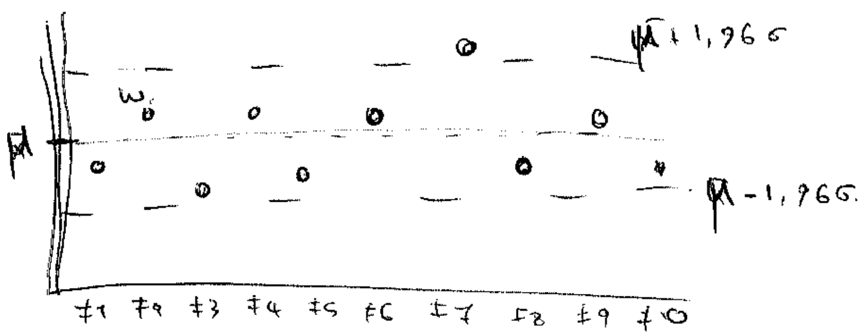
$L_1 = \bar{x} - 2.3\, s_{\bar{x}}$

$L_2 = \bar{x} + 2.3\, s_{\bar{x}}$

Για να ελαττώσουμε το εύρος του διαστήματος εμπιστοσύνης το τυπικό σφάλμα της μέσης τιμής $(s_{\bar{x}})$ πρέπει να ελαττωθεί.

Πως: $\downarrow \sigma$ ή $\uparrow n$

Διάστημα εμπιστοσύνης



: είναι ποσοστό των μετρήσεων είναι μέσα στο $\mu \pm 1{,}96\sigma$ διάστημα.

$z_{1-\alpha/2}\, \sigma$: tolerance (ανεκτικότητας) ε

Συνήθως γράφετε ως ποσοστό της μετρήσιμης τιμής $\varepsilon = \dfrac{100\, z_{1-\alpha/2}\, \sigma}{\omega_i}$

Προσοχή: μ: άγνωστος

Central Limit Theorem (Κεντρικό Οριακό Θεώρημα) CLT

Κατανομή Μέσων Τιμών

$X_1, X_2 \ldots X_n$: τυχαίο δείγμα από οποιαδήποτε κατανομή με μέση τιμή μ και διασπορά σ^2

n μεγαλώνει

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$ προσεγγίζει την κανονική $N(0,1)$ κατανομή

$$\Rightarrow \frac{n\bar{X} - n\mu}{\sqrt{n}\,\sigma} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\,\sigma} \simeq N(0,1)$$

Q: Τι συμβαίνει όταν $n \to \infty$? Δέλτα function !

$$\bar{X} \simeq N\left(\mu, \frac{\sigma^2}{n}\right)$$

- Νόμος των Μεγάλων Αριθμών (Law of Large Numbers) LLN.

↳ Ασυμπτωτικά αποτελέσματα της μέσης τιμής

$n \to \infty \qquad \bar{X} - \mu \to 0$

CLT

$n \to \infty \qquad \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \simeq N(0,1)$

EMΦ72: APPLIED STATISTICS

LECTURE 4

Λίστα: em972
majordomo subscribe em972

Έλεγχος Υποθέσεων

- Διαστήματα εμπιστοσύνης
- Central limit Theorem +

π.χ.: ερώτημα αν διαφέρουν οι μέσες τιμές των τ.μ. X_1 και X_2

Έλεγχος στατιστικής υπόθεσης

(α) ↳ ξεκινάμε με μία μηδενική υπόθεση (null hypothesis) H_0: συνήθως θέλουμε να απορρίψουμε για να δεχτούμε την εναλλακτική υπόθεση (alternative hypothesis) H_1

π.χ. $H_0: \mu_1 = \mu_2$ $\quad H_1: \mu_1 \neq \mu_2$

(β) ↳ επιλέγουμε κατάλληλη στατιστική ελέγχου (test statistic) q που ακολουθεί γνωστή κατανομή ((π.χ. $q = z \sim N(0,1)$, $q = t \sim t_{n_1+n_2-2}$ ≡ ταυτότητα)

Ορίζουμε ένα σύνολο τιμών της q που είναι απίθανο (σε στάθμη σημαντικότητας α) να πάρει η q αν ισχύει η H_0 το οποίο ονομάζεται απορριπτική περιοχή (rejection region) R. Αυτό το σύνολο αντιστοιχεί στις ουρές της κατανομής, π.χ. για $H_0: \mu_1 = \mu_2$ $\quad q = z \quad R = \{|z| > z_{1-\alpha/2}\}$

(γ) ↳ Υπολογίζουμε την τιμή της q στο δείγμα $\hat{q}$ (δειγματική στατιστική ελέγχου) και εξετάζουμε αν ανήκει στο R για να απορρίψουμε την H_0.

p-τιμή (p-value): χαμηλότερη στάθμη σημαντικότητας α για την οποία μπορούμε να απορρίψουμε την H_0 με βάση το δείγμα

π.χ. $q = z$ $\quad p = P(|z| > \tilde{z})$: $\tilde{z}$: τιμή της στατιστικής q από το δείγμα που είναι ίδιο με την κρίσιμη τιμή $z_{1-p/2}$.

↳ (δ) Υπάρχει πλήρης συμφωνία του αποτελέσματος από το διάστημα εμπιστοσύνης και του στατιστικό έλεγχο στην ίδια στάθμη σημαντικότητας α.

π.χ. αν απορρίψουμε την $H_0: \mu = \mu_0$ για $\alpha = 0.05$ τότε το 95% διάστημα εμπιστοσύνης της μ δεν περιέχει την μ_0.
Το ίδιο ισχύει για την σύγκριση του $\bar{x}_1$ και $\bar{x}_2$.

Ο έλεγχος μπορεί να είναι μονόπλευρος (one-sided test),

π.χ. $H_0: \mu_1 \leq \mu_2 \quad H_1: \mu_1 > \mu_2$, τότε ορίζεται μονόπλευρα και η απορριπτική περιοχή, π.χ. $R = \{z > z_{1-\alpha}\}$.

- Type I / Type II Errors

Απόφαση	True State	
	Ho True	Ho False
Απόρριψη Ho	Type I error α	Correct $1-b$
Μη >> Ho	Correct $1-\alpha$	Type II error b

Type I error: όταν η μηδενική υπόθεση απορρίπτεται ενώ είναι true (False Alarm Rate)
Type II error: >> >> >> δεν >> >> >> false (Miss Rate)

P(Type I error) $= \alpha$ ⟶ απόρριψη της Ho ενώ ισχύει
P(Type II error) $= b$ ↳ αποδοχή της Ho ενώ δεν ισχύει

– Δύναμη (power) pw: $1-\beta$

<u>Ερώτηση</u>: Υποθέτουμε ότι υπάρχει μια διαφορά. Θα είμαστε ικανοί να την καταλάβουμε (βρούμε);

Power είναι η πιθανότητα να ανιχνεύσουμε μια πραγματική διαφορά (σε δεδομένο p-διάστημα εμπιστοσύνης) δεδομένων των παραμέτρων του προβλήματος.

Η power εξαρτάται από:

1) Μέγεθος δείγματος n : $\uparrow n$ $\uparrow pw (\equiv 1-\beta)$

2) Μέγεθος διαφοράς : $\mu_1 - \mu_2 \downarrow$, $pw \downarrow$

3) α :

EM 272: Applied Statistics

LECTURE 5 (01/03/10) (→ Διαφάνειες)

Είδαμε: Διαστήματα εμπιστοσύνης

Έλεγχος υποθέσεων: Βασικές Αρχές

Σφάλματα τύπου I, II

I: Απόρριψη H_0 εάν είναι true

II: Αποδοχή της H_0 εάν είναι false

Έλεγχος Διαφορών Μεταξύ 2 Δειγμάτων

Έστω δύο δείγματα: $x_1, x_2, \ldots x_{n_x}$, $y_1, y_2, \ldots y_{n_y}$

t-test: Έλεγχος αν η μέση τιμή δύο συνόλων δεδομένων (δειγμάτων) είναι διαφορετική

$$t = \frac{\text{Διαφορά μέσων τιμών}}{\text{Διαφορά διασποράς μέσων τιμών}} = \frac{\bar{x} - \bar{y}}{SE(\bar{x} - \bar{y})}$$

$$SE(\bar{x} - \bar{y}) = \sqrt{\frac{S_{\bar{x}}^2}{n_x} + \frac{S_{\bar{y}}^2}{n_y}}$$

Αν $n_x = n_y = n$: $S^2 = \frac{S_x^2 + S_y^2}{2}$, $SE(\bar{x} - \bar{y}) = \sqrt{\frac{S^2}{n} + \frac{S^2}{n}}$

- Βαθμοί ελευθερίας στο t-test: $dF = n_1 + n_2 - 2 = n_x + n_y - 2$

⟹

- Υπολογίζουμε p-value χρησιμοποιώντας την t - κατανομή

Η p-value είναι η πιθανότητα να απορρίψουμε τη μηδενική υπόθεση (Ho) ενώ αυτή είναι αληθινή (σφάλμα τύπου Ι

→ Δεν είναι κατάλληλη μέθοδος για περισσότερα από 2 δείγματα.

Παράδειγμα : Μέντιουμ. : Διάβασμα τραπουλόχαρτων!
Μαντική Ικανότητα

Έλεγχος - Έχουμε 25 κάρτες τις αναποδογυρίζουμε και ρωτάμε το μέντιουμ να μας πει τι "χρώμα" είναι

H_0 : το πρόσωπο δεν είναι καλός μέντιουμ!

1 κάρτα

C : # των χαρτιών που θα προβλέψει σωστά.

$p = 1/4$

Q : Πότε απορρίπτουμε την H_0; → Αν $C = 25$! 24, 23 17?

C : κρίσιμη - Απορριπτική περιοχή → συνδέεται με το σφάλμα τύπου I.

$C = 25$ πιθανότητα (σφάλμα τύπου I) $= P$ (reject H_0 | H_0 is valid) $=$

$$P(X \geq 25 \mid p = 1/4) = \left(\frac{1}{4}\right)^{25} \simeq 10^{-15}$$

: πιθανότητα τυχαίας πρόβλεψης 25 χαρτιών!

$c = 10$. $P(X \geq 10 \mid p = 1/4) = \simeq 0{,}07$?

Διαδικασία :

α) H_0

β) T statistic → μέτρημα χαρτιών, επίπεδο α.

γ) Υπολογίζουμε το q.

π.χ. $\alpha = 1\%$

$$P(X \geq c \mid p = 1/4) \leq 0{,}01 \quad \rightarrow \quad c = 19 : \text{το ελαχιστοποιεί}$$

Παράδειγμα R. Fisher (1935) J. Neyman & Pearson, E. " } Ανάπτυξη του Ελέγχου Υποθέσεων

Μια κυρία υποστηρίζει ότι έχει τη δυνατότητα να προσδιορίσει το τρόπο παρασκευής και το είδος τσαγιού από τη γεύση

α) Ho : η κυρία δεν έχει αυτή τη δυνατότητα

b) Test statistic : # επιτυχιών σε 8 προσπάθειες

c) Distribution-κατανομή : Διωνυμική (Yes/No)

Κρίσιμη περιοχή βασίζεται σε κριτήριο <5% : 8 επιτυχίες

8 επιτυχίες > 98% εμπιστοσύνη

Αποτέλεσμα : Η Ho απορρίφθηκε!

Παράδειγμα: Δύο τρόποι κατασκευής ενός εξαρτήματος δίνουν την ακόλουθη μέση διάρκεια ζωής (συνεχής ώρες λειτουργίας)

$\mu_A = 1100\,h$ $\mu_B = 1300\,h$. κανονική κατανομή με γνωστή απόκλιση $\sigma = 270\,h$

Ένας νέος κατασκευαστής ισχυρίζεται ότι ένα καινούργιο παρόμοιο εξάρτημα έχει $\mu = \mu_B$ και χαμηλότερο κόστος παραγωγής

Τεστ: 12 εξαρτήματα ελέγχονται και δίνουν $\bar{x} = 1260\,h$

Ερώτηση: Είναι $\mu = \mu_B$;

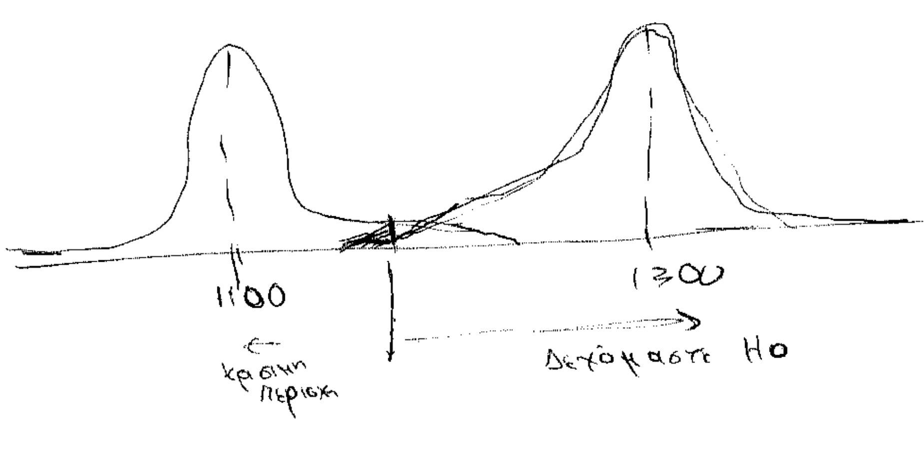
α) $H_0 : \mu = \mu_B = 1300$

$H_1 : \mu = \mu_A = 1100$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{12}} = \frac{270}{\sqrt{12}} = 77,94$$

β) Στατιστική ελέγχου: Μέτρηση νέων εξαρτημάτων

Επιλογή α. κανονική κατανομή



p - τιμή = 0,05

γ)

↳ 2 τρόποι ελέγχου αν η στατιστική ελέγχου ανήκει στην απορριπτική περιοχή.

1) Υπολογίζουμε την απορριπτική περιοχή (για την κατανομή Β)

π.χ. $\alpha = 0,05$ (5%)

$P(Z \leq 0,05) = -1,64 \Rightarrow \bar{X}_a = \mu_B - 1,64 \cdot \sigma_{\bar{x}}$

$= 1300 - 1,64 \times 77,94 = 1172.2$

→ Άρα εφόσον $\bar{X} = 1260 > \bar{X}_a$: Δεχόμαστε την Ho

2) Υπολογίζουμε την πιθανότητα να έχουμε απόκλιση από τη στατιστική τιμή της Ho

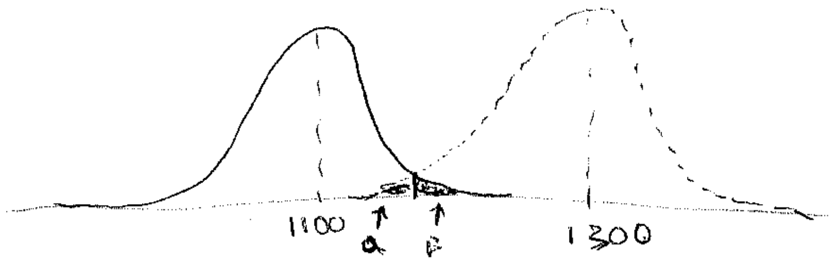
$P = P(T \geq t_n(X))$ ή $p = P(T \leq t_n(X))$

Αυτή χρησιμοποιείται από τα statistic software.

$$P = P\left(Z \leq \frac{(\bar{X} - \mu_B)}{\sigma_{\bar{X}}}\right) = P\left(Z \leq \frac{(1260 - 1300)}{77,94}\right) = 0,304$$

Εφόσον $P > 0,05$ δεν έχουμε καμία απόδειξη απόρριψης της Ho!

Σφάλμα τύπου ΙΙ (αποδοχή της Ho ενώ είναι False)

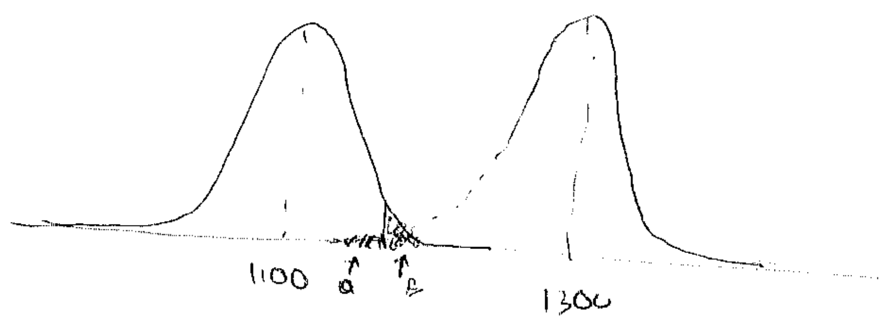


$\bar{X}_a = 1200$: όριο απόφασης

$$\alpha = \beta = P\left(z \le \frac{1200-1300}{77,94} \right) = N_{0,1}(-1,283) = 0,10$$

$$\beta = P\left(z \le \frac{1200-1100}{77,94} \right) = \text{»} = 0.10$$

Όταν ελαττώνουμε το α → ↑ β



π.χ $\alpha = 0,05$

$$\beta = P\left(z \ge \frac{(x_a - \mu_a)}{\sigma_x} \right) = P\left(z \ge \frac{1178,8-1100}{77,94} \right) = 0,177\,!$$

εναλλακτικές H_1

$\alpha = 0,05$

$H_1: \mu = \mu_A$	$z = (\mu_A - \bar{x}_{0,05})/\sigma_x$	β	1 - β
1100	0,93	0,18	0,81
1178,2	0,00	0,50	0,50
1200	-0,36	0,64	0,36
1300	-1,64	0,95	0,05

– Για να αυξήσουμε το $(1-\beta)$ μπορούμε να αυξήσουμε το n

$\rightarrow \downarrow \sigma_{\bar{x}}$

π.χ. $n = 24$, $\alpha = 0,05$

$\bar{X}_A = \mu_B - 1,64 \times \sigma_{\bar{x}} = 1300 - 1,64 \cdot 55,11$
$= \underline{1}209,6$

μ_A	$z = (\mu_A - \bar{X}_{0,05}) / \sigma_{\bar{x}}$	β	$1-\beta$
1100	1,99	0,02	0,98
1150	1,08	0,14	0,86
1200	0,17	0,43	0,57
1300	-1,64	0,95	0,05

EM 272 : Applied Statistics

LECTURE 6 (08/03/10)

Είδαμε: Έλεγχος στατιστικών υποθέσεων

Βασικά παραδείγματα

Διαστήματα εμπιστοσύνης e.g. $P(\mu \in [95,105]) = \delta = 1-\alpha$

Εργαστήριο - Lab : Παρασκευή 13:00-15:00 Η205.

Πέμπτη 15:00-17:00 Η205

ή Τετάρτη 13:00-16:00 Η203 ← 3η : 2 x 1.5 per group

Common Test Statistics

e.g. One-sample z-test

$$z_{1-ay1} \equiv z_0 = \frac{\bar{x} - \mu_0}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

or $\mu_0 = \bar{x} - (\text{κρίσιμη τιμή})\, s_{\bar{x}} = \bar{x} - z_{1-\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$

$H_0 : \mu = \mu_0$

ή $\bar{x} = \mu_0 \pm z_{1-\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$

$p = P(T \geq t_n(x))$ ή $P(T \leq t_n(x))$

$$P = P\left(|z| \leq \frac{\bar{x} - \mu_0}{\left(\frac{\sigma}{\sqrt{n}}\right)}\right) = P(|z| \leq z_0)$$

If $P(|z| \leq z_0) \geq \alpha$ accept H_0

$< \alpha$ reject H_0

↳ Test statistics → Παραμετρικά

Μη παραμετρικά.

↳ Περίληψη διαφορετικών test.

Στατιστικός έλεγχος 2 συνόλων δεδομένων (δειγμάτων)

π.χ. A, B μ_A, μ_B sets

$H_0: \mu_A = \mu_B$ (ή $\mu_A - \mu_B = 0$)

$H_1: \mu_A \neq \mu_B$: two-sided test

ή $H_0: \mu_A \leq \mu_B$, $H_1: \mu_A > \mu_B$

ή $H_0: \mu_A \geq \mu_B$, $H_1: \mu_A < \mu_B$ } One-sided test

π.χ. sets $\underline{X} = \{x_1, x_2, \ldots x_n\}$ $\underline{Y} = \{y_1, y_2, \ldots y_n\}$

N_{μ_A, σ_A} N_{μ_B, σ_B}

παράμετροι: μ_A, μ_B (or $\mu_A + \underline{\Delta}$), σ_A, σ_B

Αν ορίσουμε $\underline{\underline{D}} = \{(x_1 - y_1), (x_2 - y_2), \ldots (x_n - y_n)\}$, σ_D

Έχουμε ως παραμέτρους μόνο Δ, σ_D

Βασική Ιδέα: ανεξάρτητα δείγματα ("independent samples")

↳ τα δεδομένα έχουν επιλεχθεί με τέτοιο τρόπο ώστε όλοι οι παράγοντες που τα επηρεάζουν, εκτός αυτών που μας ενδιαφέρουν να λαμβάνονται υπόψιν με τον ίδιο τρόπο

Παράδειγμα: Ανεξάρτητα Δείγματα

(a) ⇒ Drug, Placebo : Δράση ενός φαρμάκου. Δύο δείγματα,

στο ένα χορηγείται φάρμακο στο άλλο placebo. κατόπιν μετρίεται η επίδραση στην ασθένεια.

- Αν υπάρχουν και άλλοι παράμετροι που επηρεάζουν, όπως φύλο, ηλικία τότε για να είναι τα 2 δείγματα ανεξάρτητα πρέπει να επιλεχθούν ώστε να είναι πραγματικά τυχαία όσον αφορά φύλο και ηλικία

(b) ⇒ Πρόγραμμα TV : Δείγμα → Άνδρες, Γυναίκες

- Πρέπει τα δείγματα να είναι τυχαία όσο αφορά άλλες ιδιότητες όπως : ηλικία, μορφωτικό επίπεδο, εισόδημα, κλπ.

Ανάλυση κατά ζευγάρια (paired samples)

(a) : σύνολο ασθενών που τη μια φορά λαμβάνουν placebo και την άλλη φάρμακο.

(b) ζευγάρια "συζύγων". Ο/Η σύζυγος μπορεί (συνήθως) να έχουν παρόμοια δεδομένα όσο αφορά την ηλικία, παιδεία, συνήθειες. Προσοχή : κάλλιστα αυτές οι παραδοχές μπορεί να είναι λάθος

Παράδειγμα : (Two-tailed Test) : Αξιολόγηση Δασκάλων.

Σε ένα σχολείο οι μαθητές κατανέμονται τυχαία σε 2 καθηγητές Μαθηματικών. κ. Α κ. Β

Μετά την κατανομή κ. Α : 30 μαθητές

κ. Β : 25 »

Στο τέλος του χρόνου όλοι οι μαθητές παίρνουν το ίδιο διαγώνισμα. Οι μαθητές του κ. Α έχουν μέσο όρο βαθμού 78 και τυπική απόκλιση 10

Οι » » κ Β έχουν μέσο όρο 85 και τυπική απόκλιση 15

Έλεγχος της υπόθεσης ότι οι κ. Α και κ Β είναι το ίδιο καλοί καθηγητές. Επίπεδο σημαντικότητας $a = 0,1$

α) Μηδενική Υπόθεση

$H_0: \mu_1 - \mu_2 = 0$

$H_1: \mu_1 - \mu_2 \neq 0$

Two sided test : Η H_0 θα απορριφθεί αν $\mu_1 - \mu_2 > \varepsilon$

ή $\mu_1 - \mu_2 < \varepsilon$

b) Στατιστική Ελέγχου : οι βαθμοί

t-κατανομή (διαφορετικές αποκλίσεις - διασπορές)

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{10^2}{30} + \frac{15^2}{25}} = \sqrt{3,33+9} = 3,51$$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} = \frac{\left(\frac{10^2}{30} + \frac{15^2}{25}\right)^2}{\frac{\left(\frac{10^2}{30}\right)^2}{29} + \frac{\left(\frac{15^2}{25}\right)^2}{24}}$$

$$\Rightarrow df = \frac{(3,33 + 9)^2}{\frac{3,33^2}{29} + \frac{9^2}{24}} = \frac{152,03}{0,382 + 3,375}$$

$$= \frac{152,03}{3,757} = 40,47$$

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{SE} = \frac{(78 - 85) - 0}{3,51} = \frac{-7}{3,51} = -1,99$$

$\Rightarrow$ p-value : πιθανότητα ένα t-score με 40 βαθμούς ελευθερίας (degrees of freedom) να είναι $< -1,99$ ή $> 1,99$

πιθανότητα να παρατηρήσουμε μια τιμή της στατιστικής ελέγχου ίση με την t-ελέγχου

t - κατανομή

$$P(t < -1,99) = 0,027$$

$$P(t > +1,99) = 0,027$$

$\Rightarrow$ p-value $= 0,054$

<u>Αποτέλεσμα</u> (φόσσων p-value $< \alpha = 0,10$ δεν μπορούμε να δεχτούμε την Ho!

Common Test Statistics

In the table below, the symbols used are defined at the bottom of the table. Many other tests can be found in the literature (other articles).

Name	Formula	Assumptions or notes
One-sample z-test	$z = \frac{\overline{x} - \mu_0}{(\sigma/\sqrt{n})}$	(Normal population **or** $n > 30$) **and** σ known. (z is the distance from the mean in relation to the standard deviation of the mean). For non-normal distributions it is possible to calculate a minimum proportion of a population that falls within k standard deviations for any k (see: *Chebyshev's inequality*).
Two-sample z-test	$z = \frac{(\overline{x}_1 - \overline{x}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	Normal population **and** independent observations **and** σ_1 and σ_2 are known
One-sample t-test	$t = \frac{\overline{x} - \mu_0}{(s/\sqrt{n})}$, $df = n - 1$	(Normal population **or** $n > 30$) **and** σ unknown. For non-normal populations, n should be large enough to ensure both that dist. of mean is close to normal and that s is a good estimate of σ.
Paired t-test	$t = \frac{\overline{d} - d_0}{(s_d/\sqrt{n})}$, $df = n - 1$	(Normal population of differences **or** $n > 30$) **and** σ unknown
Two-sample pooled t-test, equal variances*	$t = \frac{(\overline{x}_1 - \overline{x}_2) - d_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$, $df = n_1 + n_2 - 2$ [5]	(Normal populations **or** $n_1 + n_2 > 40$) **and** independent observations **and** $\sigma_1 = \sigma_2$ **and** σ_1 and σ_2 unknown
Two-sample unpooled t-test, unequal variances*	$t = \frac{(\overline{x}_1 - \overline{x}_2) - d_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$,	(Normal populations **or** $n_1 + n_2 > 40$) **and** independent observations **and** $\sigma_1 \neq \sigma_2$ **and** σ_1 and σ_2 unknown

	$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$ [6]	
One-proportion z-test	$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	$n \cdot p_0 > 10$ and $n(1 - p_0) > 10$ and it is a SRS (Simple Random Sample).
Two-proportion z-test, pooled	$z = \frac{(\hat{p}_1 - \hat{p}_2) - d_p}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$ $\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$	$n_1p_1 > 5$ and $n_1(1 - p_1) > 5$ and $n_2p_2 > 5$ and $n_2(1 - p_2) > 5$ and independent observations
Two-proportion z-test, unpooled	$z = \frac{(\hat{p}_1 - \hat{p}_2) - d_p}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$	$n_1p_1 > 5$ and $n_1(1 - p_1) > 5$ and $n_2p_2 > 5$ and $n_2(1 - p_2) > 5$ and independent observations
One-sample chi-square test	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$	One of the following • All expected counts are at least 5 • All expected counts are > 1 and no more that 20% of expected counts are less than 5
*Two-sample F test for equality of variances	$F = \frac{s_1^2}{s_2^2}$	Arrange so $s_1^2 \geq s_2^2$ and reject H_0 for $F > F(\alpha/2, n_1 - 1, n_2 -$ [7]

Definitions of symbols:

- α, the probability of Type I error (rejecting a null hypothesis when it is in fact true)
- n = sample size
- n_1 = sample 1 size
- n_2 = sample 2 size
- $\overline{x}$ = sample mean
- μ_0 = hypothesized population mean
- μ_1 = population 1 mean
- μ_2 = population 2 mean
- σ = population standard deviation
- σ^2 = population variance
- s = sample standard deviation
- s^2 = sample variance
- s_1 = sample 1 standard deviation
- s_2 = sample 2 standard deviation

- t = t statistic
- df = degrees of freedom
- $\overline{d}$ = sample mean of differences
- d_0 = hypothesized population mean difference
- s_d = standard deviation of differences

- $\hat{p}$ = x/n = sample proportion, unless specified otherwise
- p_0 = hypothesized population proportion
- p_1 = proportion 1
- p_2 = proportion 2
- d_p = hypothesized difference in proportion
- $\min\{n_1, n_2\}$ = minimum of n_1 and n_2
- $x_1 = n_1 p_1$
- $x_2 = n_2 p_2$
- χ^2 = Chi-squared statistic
- F = F statistic

In general, the subscript 0 indicates a value taken from the null hypothesis, H_0, which should be used as much as possible in constructing its test statistic.

EM279 : Applied Statistics

LECTURE 7 (18/03/10)

Είδαμε: Έλεγχος Στατιστικών Υποθέσεων

Παραδείγματα

Lab : Παρασκευή 13:00 - 15:00 Η205

Τετάρτη 13:00 - 15:00 Η205 → Free ok

ή 13:00 - 16:00 Η203

CH. 3

Ανάλυση Διακύμανσης (ή Διασποράς) - Analysis of Variance.

Είδαμε ότι ο έλεγχος 2 στατιστικών δειγμάτων γίνεται με τους συνήθεις στατιστικούς ελέγχους.

Τι γίνεται σε περίπτωση που θέλουμε έλεγχο > 2 δειγμάτων;

π.χ. k ανεξάρτητους πληθυσμούς μ_j, σ_j $\quad j = 1, 2, \dots k$.

→ τυχαίο δείγμα $\bar{x}_j$, s_j

	1	2	...	k
Πληθυσμός				
Μέση τιμή	μ_1	μ_2	...	μ_k
Τυπική Απόκλιση	σ_1	σ_2	...	σ_k
Δείγμα.				
Μέση τιμή	$\bar{x}_1$	$\bar{x}_2$	...	$\bar{x}_k$
Τυπική Απόκλιση	s_1	s_2	...	s_k
Μέγεθος Δείγματος	n_1	n_2	...	n_k.

π.χ. k διαφορετικά πειράματα (ή θεραπείες) ενός προβλήματος.

Πλήρως τυχαιοποιημένα σχέδια (Completely Randomized Designs)

Έστω $H_0: \mu_1 = \mu_2 \dots \mu_k$

H_1: Δύο τουλάχιστον από τους μ's διαφέρουν μεταξύ τους.

Πιθανή Λύση: Πολλαπλά t-tests.

Προβλήματα:

(A) Χρειαζόμαστε $\binom{k}{2}$ t-tests. π.χ. $k=5$ $\binom{5}{2} = \frac{5!}{2!\,3!} = \frac{1\cdot2\cdot3\cdot4\cdot5}{1\cdot2\cdot1\cdot2\cdot3} = \underline{10}$

$\frac{k!}{2!\,(k-2)!}$

$\binom{n}{r}$ n αντικείμενα ανά r $\Rightarrow \equiv {}_nC_r = \frac{n!}{r!\,(n-r)!}$

$\binom{10}{2} = \frac{10!}{2!\,8!} = 45 \dots$

Άρα πολλά t-tests!

(B) Μπορεί να οδηγηθούμε σε λάθος συμπεράσματα.

Η πιθανότητα σφάλματος τύπου I (Ho reject / Ho valid) αυξάνεται!

Γιατί;

π.χ. $\alpha = 0,05$ για 2 ελέγχους $H_0^1: \mu_1 = \mu_3$
$H_0^2: \mu_1 = \mu_k$

$P(H_0^1, H_0^2) = P(H_0^1)\cdot P(H_0^2) = (1-\alpha)(1-\alpha) = 0,95^2 = 0,9025$

Άρα συνολικό σφάλμα τύπου I $= 1 - 0,9025 = 0,0975$.

Γενικά » » » $I = 1-(1-\alpha)^m$ m: # συγκρίσεων

Ανάλυση Διακύμανσης ως προς έναν Παράγοντα

↳ υποθέτουμε ότι υπάρχει ένας μόνο παράγοντας ως προς τον οποίο οι πληθυσμοί διαφοροποιούνται μεταξύ τους.

Βασική παραδοχή (όπως και στα t-test) : $\sigma_1^2 = \sigma_2^2 = \dots \sigma_k^2 = \sigma^2$

ίδια διακύμανση

– Εκτίμηση κοινής διακύμανσης σ^2 μέσω της

$$S_w^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + \dots (n_k-1)s_k^2}{\underbrace{n_1+n_2+\dots+n_k}_{N} - k}$$

= διασπορά στο εσωτερικό των ομάδων (within groups variability)

Γενικά η συνολική μεταβλητότητα k πληθυσμών αναλύεται σε

↳ 1) διασπορά κάθε πληθυσμού γύρω από τη μέση τιμή του πληθυσμού

↳ 2) » » » » » συνολική μέση τιμή

$$S_B^2 = \frac{n_1(\bar{X}_1-\bar{X})^2 + n_2(\bar{X}_2-\bar{X})^2 + \dots + n_k(\bar{X}_k-\bar{X})^2}{k-1}$$

(between groups variability)

όπου $\bar{X} = \dfrac{n_1\bar{X}_1 + n_2\bar{X}_2 + \dots + n_k\bar{X}_k}{n_1+n_2+\dots+n_k} = \dfrac{n_1\bar{X}_1 + n_2\bar{X}_2 + \dots + n_k\bar{X}_k}{N}$

Εάν ισχύει η $H_0: \mu_1 = \mu_2 = \dots \mu_k$ πρέπει $S_B^2 = S_w^2$ ή

$F = \dfrac{S_B^2}{S_w^2} = 1$. Αν υπάρχουν διαφορές $F \neq 1$ (>1)

Παρένθεση: κατανομή F (οικογένεια κατανομών)

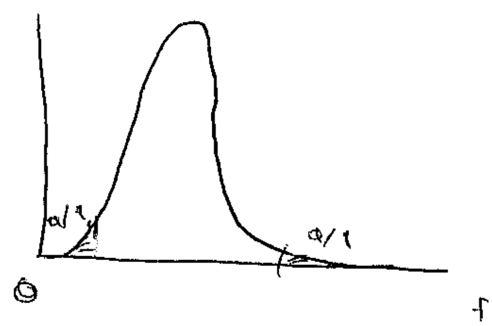
Αφορά ελέγχους διακυμάνσεων 2 κανονικά κατανεμημένων πληθυσμών

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$: ακολουθεί την κατανομή

F με n_1-1, n_2-1 βαθμούς ελευθερίας

π.χ. $H_0 : \sigma_1^2 = \sigma_2^2$

$H_1 : \sigma_1^2 \neq \sigma_2^2$



Αν η τιμή της $F = \frac{s_1^2}{s_2^2}$ που υπολογίζεται από τα διαθέσιμα δεδομένα είναι $>$ $F_{n_1-1,\, n_2-1,\, \alpha/2}$ που ορίζει μια περιοχή εμβαδού $\alpha/2$ στο δεξιό άκρο της αντίστοιχης κατανομής τότε η H_0 απορρίπτεται! (προσοχή: $s_1^2 > s_2^2$)

Παράδειγμα: Δύο αναλγητικά σκευάσματα χορηγούνται σε ασθενείς

$n_1 = 13$, $n_2 = 13$ → Διακυμάνσεις $s_1^2 = 64$, $s_2^2 = 16$

$\alpha = 0{,}05$. Να ελεγχθεί αν $\sigma_1^2 > \sigma_2^2$

$H_0 : \sigma_1^2 = \sigma_2^2$

$H_1 : \sigma_1^2 > \sigma_2^2$.

→ $F_{12,12,0{,}05} = 2{,}69 \equiv F_{crit.}$

$$F = \frac{s_1^2}{s_2^2} = \frac{64}{16} = 4 > F_{crit.} \rightarrow$$

→ απορρίπτουμε H_0 → άρα δεχόμαστε ότι $\sigma_1^2 > \sigma_2^2$.

Άρα, εφόσον ελέγχουμε διακυμάνσεις χρησιμοποιούμε την κατανομή F. $\Rightarrow$

$\Rightarrow$ $F = \dfrac{S_B^2}{S_\omega^2}$ ακολουθεί κατανομή F με $k-1, N-k$ β.ε.

$F_{k-1, N-k, \alpha} \equiv F_{crit}$

<u>Διαδικασία</u>: Υπολογίζουμε F. Αν $F > F_{crit}$, $H_0 \Rightarrow$ <u>reject</u>

Αν $F \leq F_{crit}$, H_0: accept, δηλαδή $\rightarrow$

$\rightarrow$ $\mu = \mu_1 = \mu_2 = \ldots = \mu_k$

$\sigma = \sigma_1 = \sigma_2 = \ldots = \sigma_k$.

ANOVA : συμβολισμός

Έχουμε k πληθυσμούς μ_j, σ_j $\quad j=1,2,\ldots k$.

και τ.δ. $\bar{X}_j$, S_j , X_{ij} : $i=1,2,\ldots n_j$
$j=1,2,\ldots k$

Συνολική διακύμανση των πληθυσμών :

$$S_T^2 = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\bar{X})^2}{N-1} \qquad (1)$$

Οι δύο συνιστώσες της συνολικής διασποράς είναι :

$$- S_W^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2+\cdots+(n_k-1)S_k^2}{N-k} = \frac{\sum_{j=1}^{k}(n_j-1)S_j^2}{N-k}$$

$$\left[S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right]$$

$$= \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\bar{X}_j)^2}{N-k}. \qquad (2)$$

$$- S_B^2 = \frac{n_1(\bar{X}_1-\bar{X})^2+n_2(\bar{X}_2-\bar{X})^2+\cdots+n_k(\bar{X}_k-\bar{X})^2}{k-1} = \frac{\sum_{j=1}^{k}n_j(\bar{X}_j-\bar{X})^2}{k-1} \qquad (3)$$

Οι αριθμητές στις παραπάνω σχέσεις έχουν μια συγκεκριμένη ονομασία.

⟶

⇒ (α) της s_T^2 : συνολικό άθροισμα τετραγώνων (Total Sum of Squares) TSS

$$TSS = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2$$

⇒ (β) της s_W^2 : άθροισμα τετραγώνων στο εσωτερικό των ομάδων (Sum of Squares Within-groups) SSW

$$SSW = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$$

⇒ (γ) της s_B^2 : άθροισμα τετραγώνων μεταξύ των ομάδων (Sum of squares between-groups) SSB

$$SSB = \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x})^2$$

$TSS = SSB + SSW$ → Απόδειξη ;

Αντίστοιχα :

$s_B^2 = \dfrac{SSB}{k-1}$: μέσο τετράγωνο μεταξύ των ομάδων

$s_W^2 = \dfrac{SSW}{N-k}$: » » στο εσωτερικό » »

LECTURE 8

ANOVA Πίνακας Είδαμε:

Πηγή Διασποράς	Άθροισμα Τετραγώνων	Βαθμοί Ελευθερίας	Μέσο Τετράγωνο	F
Μεταξύ των Ομάδων	SSB	$k-1$	$s_B^2 = \frac{SSB}{k-1}$	$\frac{s_B^2}{s_W^2}$
Στο εσωτερικό των ομάδων	SSW	$N-k$	$s_W^2 = \frac{SSW}{N-k}$	
Σύνολο	SST	$N-1$		

Οι υπολογισμοί είναι αρκετά πολύπλοκοι και χρονοβόροι →

→ Χρήση SOFTWARE.

Προσοχή - Απαιτείται: Έλεγχος προϋποθέσεων της μεθόδου

Ορθή ανάγνωση των αποτελεσμάτων

Ώστε
⇒ να έχουμε Ασφαλή συμπεράσματα.

$$SSB = \sum_{j=1}^{k} n_j (\bar{X}_j - \bar{X})^2$$

$$SSW = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2$$

$$SST = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2$$

ή TSS

Παράδειγμα: Επιδημιολογική Μελέτη

Μετρήθηκε ο αριθμός των λευκών αιμοσφαιρίων σε 3 ομάδες υγιών ατόμων με διαφορετικές συνήθειες καπνίσματος.

Όλοι ήταν άνδρες ηλικίας 25-34 ετών.

1η ομάδα: 42 μη καπνιστές $n_1 = 42$

$\bar{X}_1 = 8071,4$, $s_1 = 1771,6$

2η ομάδα: 26 καπνιστές $n_2 = 26$

με < 20 τσιγάρα/ημέρα $\bar{X}_2 = 8680,8$, $s_2 = 1750,2$

3η ομάδα: 32 καπνιστές $n_3 = 32$

> 20 τσιγάρα/ημέρα $\bar{X}_3 = 10809,4$, $s_3 = 1795,7$

Έλεγχος: αν οι πληθυσμιακές μέσες τιμές των λευκών αιμοσφαιρίων των τριών ομάδων διαφέρουν σημαντικά μεταξύ τους.

Αρχικά:

$$- \bar{X} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + n_3\bar{X}_3}{n_1+n_2+n_3} = \frac{42(8071,4) + 26(8680,8) + 32(10809,4)}{\underbrace{(42+26+32)}_{=100}} = 9106.$$

Κατόπιν υπολογίζουμε τις διάφορες συνιστώσες της διασποράς

$$- SSB = n_1(\bar{X}_1-\bar{X})^2 + n_2(\bar{X}_2-\bar{X})^2 + n_3(\bar{X}_3-\bar{X})^2 =$$

$$= 42(8071,4-9106)^2 + 26(8680,8-9106)^2 + 32(10809,7-9106)^2$$

$$= 142\,503\,114.$$

$- SSW = (n_1-1)s_1^2 + (n_2-1)s_2^2 + (n_3-1)s_3^2 =$

$= (42-1)(1771,6^2) + (26-1)(1750,2)^2 + (32-1)(1795,4)^2 =$

$= 305188596$

και

$$s_B^2 = \frac{SSB}{k-1} = \frac{142503114}{3-1} = 71251557$$

$$s_W^2 = \frac{SSW}{N-k} = \frac{\cancel{565193286}\ 305188596}{100-3} = 3146273,5 \quad \underline{\text{OK}} \quad \cancel{5826738,9}$$

Υποθέσεις:

$H_0: \mu_1 = \mu_2 = \mu_3$

H_A: Δύο τουλάχιστον από τις μ_1, μ_2, μ_3 διαφέρουν μεταξύ τους
$\mu_1 \neq \mu_2$ ή $\mu_2 \neq \mu_3$ ή $\mu_1 \neq \mu_3$

- Επίπεδο σημαντικότητας ελέγχου $\alpha = 0,05$

- Τιμή κριτηρίου ελέγχου $F = \dfrac{s_B^2}{s_W^2} = \dfrac{71251557}{\cancel{5826738,9}\ 3146273,5} = \cancel{12,228}\ 22,646$

Κρίσιμη τιμή της κατανομής $F_{crit} = F_{2,97,0,05} \simeq 3,11$

- Εφόσον $F > F_{crit}$: H_0 : rejected

↳ Ο μέσος αριθμός λευκών αιμοσφαιρίων διαφοροποιείται σε δύο τουλάχιστον από τους τρεις πληθυσμούς.

ANOVA: Πολλαπλές συγκρίσεις

Είδαμε ότι η $H_0: \mu_1 = \mu_2 = \dots = \mu_k$ έχει ως εναλλακτική υπόθεση

H_A: 2 τουλάχιστον από τις μ είναι διαφορετικές μεταξύ τους. ⇒

⇒ Δεν μας βοηθά στον προσδιορισμό των μέσων τιμών που διαφέρουν μεταξύ τους. ⇒

⇒ Απαιτούνται έλεγχοι πολλαπλών συγκρίσεων.

(Multiple Comparison Tests)

Πολλά διαφορετικά Test

A) Έλεγχος του Fisher (Least Significant Difference LSD) (1935)

: Διευρυμένη εκδοχή του t-test.

Πρώτα κάνουμε το τυπικό F-test. Όταν έχει απορριφθεί η H_0 για επίπεδο σημαντικότητας α. η ελάχιστη σημαντική διαφορά δύο οποιονδήποτε μέσων τιμών μ_i και μ_j είναι:

$$LSD = t_{N-k,\alpha/2} \sqrt{s_w^2 \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

. Αν $|\bar{x}_i - \bar{x}_j| \geq LSD$

⇒ μ_i, μ_j : διαφέρουν σημαντικά μεταξύ τους

η 100(1-α)% Δ.Ε.

$$\left(\bar{x}_i - \bar{x}_j\right) \pm t_{N-k,\alpha/2} \sqrt{s_w^2\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

↳ Αρκετά "ελαστικός" έλεγχος. Δεν διασφαλίζει πάντα ότι το συνολικό σφάλμα τύπου Ι είναι α.!

Β) Έλεγχος του Tukey

: Χρησιμοποιεί την τυποποιημένη κατά Student κατανομή εύρους (studentized range distribution) και υπολογίζεται από:

$$\frac{\bar{X}_{max} - \bar{X}_{min}}{\sqrt{\frac{S_W^2}{n}}}$$

$\bar{X}_{max} - \bar{X}_{min}$: εύρος των δειγματικών μέσων τιμών κ ομάδων.

⇒ Για $n_i = n_j = n$

Έλεγχος με την κρίσιμη τιμή

$$W = q_{\alpha, k, N-k} \sqrt{\frac{S_W^2}{n}}$$

τιμή της studentized range distribution

Αν $|\bar{X}_i - \bar{X}_j| \geq q_{\alpha, k, N-k} \sqrt{\frac{S_W^2}{n}}$ ⇒

⇒ Οι μ_i, μ_j διαφέρουν σημαντικά μεταξύ τους.

Αν $n_i \neq n_j$

$$W_{ij} = q_{\alpha, k, N-k} \sqrt{\frac{S_W^2}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

Εφαρμοσμένη Στατιστική (Applied Statistics) 9.1

(12/06)

LECTURE 9

(Πώς τέλεσαι το πρόβλημα + 10%)

Είδαμε:

k : ανεξάρτητοι πληθυσμοί

- ANOVA με 1 μεταβλητή

$$\left.\begin{array}{l}\bar{x}_j,\ \mu_j\\ s_j,\ \sigma_j\end{array}\right\}\ j=1,2,\dots k$$

↳ πίνακας ANOVA

Πηγή Διασποράς	Άθροισμα Τετραγώνων	Βαθμοί Ελευθερίας	Μέσο Τετράγωνο	F
Μεταξύ των ομάδων	$SSB=\sum_{j=1}^{k} n_j(\bar{x}_j-\bar{x})^2$	$k-1$	$S_B^2=\frac{SSB}{k-1}$	$\frac{S_B^2}{S_W^2}$
Εσωτερικό των Ομάδων	$SSW=\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j)^2$	$N-k$	$S_W^2=\frac{SSW}{N-k}$	
	$SST=\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij}-\bar{x})^2$	$N-1$		

- $H_0: \mu_1=\mu_2=\dots\mu_k$

$H_A: \neq H_0$.

: Δεν μας βοηθά στον προσδιορισμό των μέσων τιμών που διαφέρουν μεταξύ τους. Απαιτούνται ⇒

⇒ Πολλαπλά Test

Διαφορετικές Μέθοδοι:

A) Test Fisher (Ελάχιστη Σημαντική Διαφορά)

$$LSD = t_{N-k,\,a/2}\sqrt{s_W^2\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

$|\bar{x}_i-\bar{x}_j| \geq LSD$ ⇒

μ_i, μ_j διαφέρουν σημαντικά μεταξύ τους

Δεν μας δίνει πάντα ότι το συνολικό σφάλμα τύπου I είναι a!

B) Test Tukey

↳ Studentized Range Distribution

$$\frac{\bar{x}_{max}-\bar{x}_{min}}{\sqrt{s_W^2/n}}$$

Έλεγχος με $\omega = q_{a,k,N-k}\sqrt{\frac{s_W^2}{n}}$ $\quad |\bar{x}_i-\bar{x}_j| \geq \omega$

Γ) Έλεγχος του Bonferroni (Bonferroni Correction)

: Απλά t-test με τροποποιημένο επίπεδο σημαντικότητας α.

Είδαμε ότι σε κ πληθυσμούς $\binom{k}{2} = \frac{k(k-1)}{2}$ διαφορετικά t-test οδηγούν σε αύξηση του συνολικού σφάλματος τύπου I. ⇒

⇒ Τροποποιημένο α^* → $\alpha^* = \dfrac{\alpha}{\binom{k}{2}}$

α^* : επίπεδο σημαντικότητας σε κάθε επιμέρους έλεγχο ώστε το συνολικό σφάλμα τύπου I να είναι α.

π.χ. $k = 3$, $\alpha = 0{,}05$ $\binom{3}{2} = 3$ $\alpha^* = \dfrac{0{,}05}{3} = 0{,}016$

⇒ Άρα $H_0: \mu_i = \mu_j$

→ υπολογίζουμε την $t_{ij} = \dfrac{(\bar{x}_i - \bar{x}_j)}{\sqrt{s_w^2\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$

→ κρίσιμη τιμή της κατανομής $t_{N-k,\alpha^*/2}$

Δ.Ε. $(\bar{x}_i - \bar{x}_j) \pm t_{N-k,\alpha^*/2}\sqrt{s_w^2\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$

Γενικά:

* Για μικρό αριθμό πληθυσμών (π.χ. $k < 5$) ο έλεγχος Bonferroni, είναι πολύ χρήσιμος.

* Για μεγαλύτερο αριθμό πληθυσμών το $\alpha^* = \dfrac{\alpha}{\binom{k}{2}}$ γίνεται πολύ μικρό και επομένως το σφάλμα τύπου I κάθε επιμέρους ελέγχου γίνεται πολύ μικρό και το σφάλμα τύπου II πολύ μεγάλο! Τότε

$\Rightarrow$ Χρησιμοποιείται περισσότερο το test Tukey.

* Προσοχή: Υπάρχουν ακόμη πολλά διαφορετικά εναλλακτικά test πολλαπλών συγκρίσεων.

π.χ. Duncan, Scheffé, Newman, Sidak, Dunnet κ.α.

Βασικό κριτήριο επιλογής: το είδος των συγκρίσεων που θα γίνουν και το πείραμα που ακολουθεί η μελέτη.

Applied Statistics

ANOVA - Γραμμικό Υπόδειγμα

Group (ομάδες)

	1	2	3	...	k	
	X_{11}	X_{12}	X_{13}	...	X_{1k}	
	X_{21}	X_{22}	X_{23}	...	X_{2k}	
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
	$X_{n_1 1}$	$X_{n_2 2}$	$X_{n_3 3}$	...	$X_{n_k k}$	
Άθροισμα	$T_{.1}$	$T_{.2}$	$T_{.3}$	...	$T_{.k}$	$T_{..}$
Μέση τιμή	$\bar{X}_{.1}$	$\bar{X}_{.2}$	$\bar{X}_{.3}$	...	$\bar{X}_{.k}$	$\bar{X}_{..}$

Όπου
$$\bar{X}_{.j} = \frac{T_{.j}}{n_j}$$

$$T_{..} = \sum_{j=1}^{k} T_{.j} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} X_{ij}$$

$$\bar{X}_{..} = \frac{T_{..}}{N}, \quad N = \sum_{j=1}^{k} n_j$$

$X_{ij} = \mu_j + \varepsilon_{ij}$, $\varepsilon_{ij} \equiv X_{ij} - \mu_j$: error (σφάλμα)

$\mu = \dfrac{\sum_{j=1}^{k} \mu_j}{k}$: συνολική μέση τιμή όλων των πληθυσμών

$\mu_j = \mu + \tau_j$ $\quad \tau_j \equiv \mu_j - \mu$: group effect

$X_{ij} = \mu + \tau_j + \varepsilon_{ij}$

Οι προϋποθέσεις της ANOVA (ανάλυσης διακύμανσης) επαναδιατυπώνονται ως εξής:

α) k δεδομένα: τυχαία και ανεξάρτητα δείγματα από τους αντίστοιχους πληθυσμούς με $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 = \sigma^2$

β) τ_j: άγνωστες με $\sum_{j=1}^{k} \tau_j = 0$

$$\left(\tau_j = \mu_j - \mu \Rightarrow \sum_{j=1}^{k} \tau_j = \sum_{j=1}^{k} (\mu_j - \mu) = \sum_{j=1}^{k} \mu_j - k \frac{\sum_{j=1}^{k} \mu_j}{k} = 0 \right)$$

γ) $\varepsilon_{ij} = x_{ij} - \mu_j \Rightarrow$

ε_{ij}: ανεξάρτητα, κανονική κατανομή $\left(\sum_{j=1}^{k} \varepsilon_{ij} = \sum_{j=1}^{k} (x_{ij} - \mu_j) \right)$ με μέση τιμή 0 και διακύμανση ίση με τη διακύμανση των x_{ij}, εφόσον x_{ij} και ε_{ij} διαφέρουν κατά μια σταθερά

$\Rightarrow$ Οι υποθέσεις της ανάλυσης $H_0: \mu_1 = \mu_2 = \dots = \mu_k$

H_A: Δύο τουλάχιστον από τις μ_j διαφέρουν μεταξύ τους

γίνονται:

$H_0: \tau_j = 0, \quad j = 1, 2, \dots k$

H_A: Ένα τουλάχιστον $\tau_j \neq 0$.

$$x_{ij} = \mu + \tau_j + \varepsilon_{ij} \Rightarrow x_{ij} - \mu = \underbrace{\tau_j}_{\mu_j - \mu} + \underbrace{\varepsilon_{ij}}_{x_{ij} - \mu_j}$$

$$\Rightarrow x_{ij} - \mu = (\mu_j - \mu) + (x_{ij} - \mu_j), \quad \begin{matrix} i = 1, 2, \dots n_j \\ j = 1, 2, \dots k \end{matrix} \qquad (1)$$

Προσοχή: ο πληθυσμιακός μέσος μ εκτιμάται από το δειγματικό $\bar{x}$, η διαφορά $x_{ij}-\mu$ εκτιμάται από την ποσότητα $x_{ij}-\bar{x}_{..}$. Επίσης η διαφορά $(\mu_j-\mu)$ εκτιμάται από τη $(\bar{x}_{.j}-\bar{x}_{..})$ και η $(x_{ij}-\mu_j)$ από την $(x_{ij}-\bar{x}_{.j})$

Άρα η (1) γίνεται

$$x_{ij}-\bar{x}_{..} = (\bar{x}_{.j}-\bar{x}_{..}) + (x_{ij}-\bar{x}_{.j}) \qquad (2)$$

Παίρνοντας τα αθροίσματα των τετραγώνων έχουμε

NOTE: TO BE PROVEN BY STUDENTS

[ok by Γωγώ!]

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_{..})^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left[(\bar{x}_{.j}-\bar{x}_{..})+(x_{ij}-\bar{x}_{.j})\right]^2$$

$$= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{x}_{.j}-\bar{x}_{..})^2 + \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_{.j})^2 + \underbrace{2\sum_{j=1}^{k}\left[(\bar{x}_{.j}-\bar{x}_{..})\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_{.j})\right]}_{A} \qquad (3)$$

Αλλά εφόσον $\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_{.j})=0 \Rightarrow A = 0$

για $j=1,2 \ldots k$.

και $\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\bar{x}_{.j}-\bar{x}_{..})^2 = \sum_{j=1}^{k} n(\bar{x}_{.j}-\bar{x}_{..})^2$

Άρα (3) $\Rightarrow$

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_{..})^2 = \sum_{j=1}^{k} n_j(\bar{x}_{.j}-\bar{x}_{..})^2 + \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_{.j})^2 \qquad (4)$$

Τελικά η (4) είναι η γνωστή μας ισότητα.

$$SST = SSB + SSW$$

Note : SSW → συνδέεται με την $(X_{ij} - \bar{X}_{.j})$ ή την $(X_{ij} - \mu_j)$

δηλαδή τα ε_{ij} → άρα αποτελεί ένα μέτρο της διασποράς των σφαλμάτων.

(error sum of squares)

και S_W^2 → (error mean square)

Υπολογιστικά : $SST = \sum_{j=1}^{k} \sum_{i=1}^{n_j} X_{ij}^2 - \frac{T_{..}^2}{N}$

$$SSB = \sum_{j=1}^{k} \frac{T_{.j}^2}{n_j} - \frac{T_{..}^2}{N}$$

$$SSW = \sum_{j=1}^{k} \sum_{i=1}^{n_j} X_{ij}^2 - \sum_{j=1}^{k} \frac{T_{.j}^2}{n_j}.$$

APPLIED STATISTICS

LECTURE 10 (19/04/10)

Ασκήσεις 1. Απορίες ? — Τι είναι PDF, CDF
Παρατηρήσεις: Αναφορά!

Είδαμε: ANOVA με 1 μεταβλητή - Γραμμικό Υπόδειγμα

- Γιατί κάνουμε ANOVA;
- Τι είναι ;

* Στην ANOVA με 1 μεταβλητή σκοπός είναι η σύγκριση μέσων τιμών μιας τυχαίας μεταβλητής σε κ πληθυσμούς (ομάδες)

Παραδοχή: Όλοι οι άλλοι παράγοντες οι οποίοι μπορούν να επιδράσουν στη μεταβλητότητα των τιμών στο εσωτερικό των ομάδων, είναι ελεγχόμενοι.

Όταν η παραπάνω παραδοχή δεν ισχύει μπορούμε να ελέγξουμε την επίδραση ενός δεύτερου παράγοντα, με τη σχεδίαση τυχαιοποιημένων μπλοκ (randomized block design)

(Fisher 1990)

Γενικεύσεις:

- Block: οι παρατηρήσεις χωρίζονται σε n block με τέτοιο τρόπο ώστε μέσα σε κάθε block οι μονάδες να είναι όσο το δυνατόν περισσότερο ομοιογενείς

- Ο αριθμός των παραγόντων που ελέγχονται > 1

Παράδειγμα: Σε μια ιατρική μελέτη μελετάται η σχέση της ποσότητας χοληστερόλης στο αίμα σε γυναίκες αναφορικά με την ηλικία και τις καπνιστικές συνήθειες

Παράγοντας B ηλικία (π.χ ≥ ή < 50 ετών)

Παράγοντας A κάπνισμα
- μη καπνίστριες
- καπνίστριες (< 10 τσιγάρα/μέρα)
- καπνίστριες (> 10 »)

π.χ. $n = 150$: n_1 (ηλικία < 40) = 50, n_2 = 50, n_3 = 50

$m_1 = 50$
$m_2 = 50$
$m_3 = 50$

Παράγοντας A	Παράγοντας B: < 40	40 < ≤ 60	> 60
μη καπνίστριες			
καπνίστριες < 10 τσ/μέρα			
καπνίστριες > 10 τσ/μέρα			

ANOVA Με 2 Μεταβλητές Παράγοντες (Factors)

(Generalized Randomized Block Design)

Παράγοντας A	Παράγοντας B: Group 1	2	...	k	Άθροισμα	Μέση Τιμή
1	X_{111}	X_{121}	...	X_{1k1}	$T_{1..}$	$\bar{X}_{1..}$
	X	$X_{12?}$				
⋮	⋮	⋮		⋮		
	$X_{1..}$	X_{12n}	...	X_{1kn}		
2	X_{211}	X_{221}	...	X_{2k1}	$T_{2..}$	$\bar{X}_{2..}$
⋮	⋮	⋮		⋮		
	X_{21n}	X_{22n}	...	X_{2kn}		
m	X_{m11}	X_{m21}	...	X_{mk1}	$T_{m..}$	$\bar{X}_{m..}$
	⋮	⋮		⋮		
	X_{m1n}	X_{m2n}	...	X_{mkn}		
Άθροισμα	$T_{.1.}$	$T_{.2.}$		$T_{.k.}$	$T_{...}$	
Μέση Τιμή	$\bar{X}_{.1.}$	$\bar{X}_{.2.}$		$\bar{X}_{.k.}$		$\bar{X}_{...}$

→ 2 μεταβλητές με αριθμό κατηγοριών m και k αντίστοιχα.

- Two-way Factorial Design.

* <u>Υπολογίζουμε</u>

$$T_{ij} = \sum_{l=1}^{n} x_{ijl} \quad , \qquad \bar{x}_{ij} = \frac{T_{ij}}{n} \qquad \begin{matrix} i = 1,2,\dots m \\ j = 1,2,\dots k \end{matrix}$$

$$x_{ijl} = \mu + a_i + \beta_j + (a\beta)_{ij} + \varepsilon_{ijl} \qquad (1)$$

μ : συνολική μέση τιμή (grand mean)

a_i = επίδραση του επιπέδου i του παράγοντα A στη τιμή x_{ijl}

$(\bar{x}_{i..} - \bar{x}_{...})$ (2a)

β_j : » » j » » B » » x_{ijl}

$(\bar{x}_{.j.} - \bar{x}_{...})$ (2b)

$(a\beta)_{ij}$: αλληλεπίδραση των επιπέδων i, j » » x_{ijl}

$$= [\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x}_{...}] = [\bar{x}_{ij.} - (\bar{x}_{i..} - \cancel{\bar{x}_{...}}) - (\bar{x}_{.j.} - \bar{x}_{...}) - \cancel{\bar{x}_{...}}] \quad (2c)$$

ε_{ij} : σφάλμα κατά την εκτίμηση των x_{ijk}, δηλαδή η επίδραση όλων των άλλων εξωγενών παραγόντων οι οποίοι δεν λαμβάνονται υπόψιν.

$$\varepsilon_{ij} = x_{ijl} - \mu_j = (x_{ijl} - \bar{x}_{ij}) \qquad (2d)$$

Έχουμε (1) $\xrightarrow{2a,\,2b,\,2c,\,2d}$

$$(X_{ijl}-\mu)=(\bar{X}_{i\cdot\cdot}-\bar{X}_{\cdot\cdot\cdot})+(\bar{X}_{\cdot j\cdot}-\bar{X}_{\cdot\cdot\cdot})$$

$$+(\bar{X}_{ij\cdot}-\bar{X}_{i\cdot\cdot}-\bar{X}_{\cdot j\cdot}+\bar{X}_{\cdot\cdot\cdot})+(X_{ijl}-\bar{X}_{ij\cdot}) \Rightarrow$$

$$\Rightarrow \sum_{i=1}^{m}\sum_{j=1}^{k}\sum_{l=1}^{n}(X_{ijl}-\bar{X}_{\cdot\cdot\cdot})^2=\sum_{i=1}^{m}\sum_{j=1}^{k}\sum_{l=1}^{n}(\bar{X}_{i\cdot\cdot}-\bar{X}_{\cdot\cdot\cdot})^2+\sum_{i=1}^{m}\sum_{j=1}^{k}\sum_{l=1}^{n}(\bar{X}_{\cdot j\cdot}-\bar{X}_{\cdot\cdot\cdot})^2$$

$$+\sum_{i=1}^{m}\sum_{j=1}^{k}\sum_{l=1}^{n}(\bar{X}_{ij\cdot}-\bar{X}_{i\cdot\cdot}-\bar{X}_{\cdot j\cdot}+\bar{X}_{\cdot\cdot\cdot})^2+\sum_{i=1}^{m}\sum_{j=1}^{k}\sum_{l=1}^{n}(X_{ijl}-\bar{X}_{ij\cdot})^2$$

$$=\underbrace{nk\sum_{i=1}^{m}(\bar{X}_{i\cdot\cdot}-\bar{X}_{\cdot\cdot\cdot})^2}_{SSA}+\underbrace{nm\sum_{j=1}^{k}(\bar{X}_{\cdot j\cdot}-\bar{X}_{\cdot\cdot\cdot})^2}_{SSB}$$

$$+\underbrace{n\sum_{i=1}^{m}\sum_{j=1}^{k}(\bar{X}_{ij\cdot}-\bar{X}_{i\cdot\cdot}-\bar{X}_{\cdot j\cdot}+\bar{X}_{\cdot\cdot\cdot})^2}_{SSAB}+\underbrace{\sum_{i=1}^{m}\sum_{j=1}^{k}\sum_{l=1}^{n}(X_{ijl}-\bar{X}_{ij\cdot})^2}_{SSE} \quad (3)$$

$$TSS=\underbrace{SSA}^{\text{Παράγοντας A}}+\underbrace{SSB}^{\text{Παράγοντας B}}+\underbrace{SSAB}^{\text{Αλληλεπίδραση AB}}+\underbrace{SSE}^{\text{Σφάλμα}}$$

Βαθμοί Ελευθ. $(mkn-1)$ $(m-1)$ $(k-1)$ $(m-1)(k-1)$ $mk(n-1)$

Μέσα Τετράγωνα

$\Rightarrow$ $MSA=\dfrac{SSA}{(m-1)}$, $MSB=\dfrac{SSB}{k-1}$, $MSAB=\dfrac{SSAB}{(m-1)(k-1)}$ $MSE=\dfrac{SSE}{mk(n-1)}$

Έλεγχος Υποθέσεων.

α. ~~Παράγοντας~~ Επίδραση A

$H_0: \alpha_i = 0 \quad , \quad i = 1, 2, \ldots m$

H_A: τουλάχιστο ένα $\alpha_i \neq 0$

↳ ελέγχεται η επίδραση του A επί των τιμών x_{ijk}

$$F = \frac{MSA}{MSE}$$

b. ~~Παράγοντας~~ Επίδραση B

$H_0: \beta_j = 0 \quad , \quad j = 1, 2, \ldots k$

H_A: τουλάχιστον ένα $\beta_j \neq 0$

$$F = \frac{MSB}{MSE}$$

γ. Έλεγχος ~~επιδράσεων~~ αλληλεπιδράσεων A, B

$H_0: (\alpha\beta)_{ij} = 0 \quad , \quad i = 1, 2, \ldots m \quad , \quad j = 1, 2, \ldots k$

H_A: τουλάχιστον ένα $(\alpha\beta)_{ij} \neq 0$

$$F = \frac{MSAB}{MSE}$$

Πίνακας ANOVA (Τυχαιοποιημένου σχεδίου, 2 παράγοντες)

Πηγή Διασποράς	Άθροισμα Τετραγώνων	Βαθμοί ελευθερίας	Μέσο τετράγωνο	F
Παράγοντας A	SSA	$m-1$	$MSA = \frac{SSA}{(m-1)}$	$\frac{MSA}{MSE}$
Παράγοντας B	SSB	$k-1$	$MSB = \frac{SSB}{(k-1)}$	$\frac{MSB}{MSE}$
Αλληλεπίδραση AB	SSAB	$(m-1)(k-1)$	$MSAB = \frac{SSAB}{(m-1)(k-1)}$	$\frac{MSAB}{MSE}$
Σφάλμα	SSE	$mk(n-1)$	$MSE = \frac{SSE}{mk(n-1)}$	
Σύνολο	TSS	$mkn-1$		

APPLIED STATISTICS

LECTURE 11 (26/05/10)

ANOVA – 2 παράγοντες

Παράδειγμα:

Σε μία ιατρική μελέτη ελέγχουμε τη σχέση χοληστερόλης στο αίμα σε γυναίκες αναφορικά με την ηλικία και τις καπνιστικές συνήθειες

36 γυναίκες

- Παράγοντας A : κάπνισμα
  - μη καπνίστριες
  - καπνίστριες I (< 10 τσιγάρα/μέρα)
  - » II (> 10 »)
- Παράγοντας B : ηλικία
  - ηλικία I (< 40 ετών)
  - » II (40< και <60 ετών) 40-60
  - » III (> 60 ετών)

Παράγοντας A		Παράγοντας B ηλικία I	ηλικία II	ηλικία III	Άθροισμα	Μέση τιμή
μη καπνίστριες		208	210	240		
		226	201	230		
		200	190	220	2520	210
		180	195	210		
	Άθροισμα	814	806	900		
	Μέση τιμή	203,5	201,5	225		
καπνίστριες I		215	222	245		
		207	232	247	2610	217,5
		190 180	212	232		
			207	221		
	Άθροισμα	792	873	945		
	Μέση τιμή	198	218,25	236,25		
καπνίστριες II		209	215	260	2720	226,67
		227	203	259		
		210	235	233		
		200	230	250		
	Άθροισμα	840	883	997		
	Μέση τιμή	210	220,75	249,25		
Άθροισμα		2446	2562	2842	7850	

Υπολογίζουμε όλες τις ποσότητες που χρειαζόμαστε:

$$TSS = \sum_{i=1}^{m}\sum_{j=1}^{k}\sum_{l=1}^{n} x_{ijl}^2 - \underbrace{\frac{\left(\sum_{i=1}^{m}\sum_{j=1}^{k}\sum_{l=1}^{n} x_{ijl}\right)^2}{m\cdot k\cdot n}}_{C}$$

Έχουμε: $m = 3$, $k = 3$, $n = 4$

$$C = \frac{7850^2}{36} = 1711736,111$$

$$TSS = (208^2 + 226^2 + 200^2 + \dots 250^2) - 1711736,111 =$$

$$= \underbrace{528366 + 572174 + 621010}_{1721550} - 1711736,111 = 9813,889$$

$$SSA = \frac{\sum_{i=1}^{m} T_{i\cdot}^2}{k\cdot n} - C = \frac{2590^2 + 2610^2 + 2720^2}{12} - C = 1672,222$$

$$SSB = \frac{\sum_{j=1}^{k} T_{\cdot j\cdot}^2}{m\cdot n} - C = \frac{2446^2 + 2562^2 + 2842^2}{12} - C = 6907,55$$

$$SSAB = \frac{\sum_{i=1}^{m}\sum_{j=1}^{k} T_{ij\cdot}^2}{n} - \frac{\sum_{i=1}^{m} T_{i\cdot}^2}{k\, n} - \frac{\sum_{j=1}^{k} T_{\cdot j\cdot}^2}{m\cdot n} + C$$

$$= \frac{814^2 + 806^2 + 900^2 + 792^2 + 873^2 + 945^2 + 840^2 + 883^2 + 997^2}{4}$$

$$- \frac{2590^2 + 2610^2 + 2720^2}{12} - \frac{2446^2 + 2562^2 + 2842^2}{12} + 1711736,111$$

$SSAB = 1720.987 - 1713{,}408{,}333 - 1718643{,}667 + 1711736{,}111$

$= 671{,}111$

$$SSE = \sum_{i=1}^{m}\sum_{j=1}^{k}\sum_{b=1}^{n} x_{ijb}^2 - \frac{\sum_{i=1}^{m}\sum_{j=1}^{k} T_{ij.}^2}{n}$$

$= 1725550 - 1720.987 = 4563$

Βαθμοί ελευθερίας:

↳ συνολικό άθροισμα $m \cdot k \cdot n - 1 = 3 \cdot 3 \cdot 4 - 1 = 35$

Παράγοντα A $m-1 = 3-1 = 2$

» B $k-1 = 2$

» Αλληλεπίδραση $(m-1)(k-1) = 4$

» Σφαλμάτων $mk(n-1) = 3 \cdot 3 \cdot 3 = 9$

Μέσα τετράγωνα:

$$MSA = \frac{SSA}{m-1} = 836{,}111$$

$$MSB = \frac{SSB}{k-1} = 3453{,}775$$

$$MSAB = \frac{SSAB}{(m-1)(k-1)} = 167{,}78$$

$$MSE = \frac{SSE}{mk(n-1)} = 507$$

Έλεγχοι

Note : $x_{ijk} = \mu + a_i + \beta_j + (ab)_{ij} + \epsilon_{ijk}$
Επίπεδο σημαντικότητας 90% , $a = 0{,}10$

A) Έλεγχος παράγοντα A (γονιμοποίησης)

$H_0 : \mu_A^1 = \mu_A^2 = \mu_A^3$

ότι : $a_1 = a_2 = a_3 = 0$

H_A : $a_1 \neq 0$ ή $a_2 \neq 0$ ή $a_3 \neq 0$

$F = \frac{MSA}{MSE} = \frac{836{,}111}{507} = 1{,}649$ $\quad a = 0{,}05 = 5\%$ $\quad F_{2,9,a} = F_{2,9,0{,}05} = 4{,}26$)

$F < F_{2,9,01}$: Ho: Accepted

B) Έλεγχος παράγοντα B (ηλικία)

$H_0 : B_1 = B_2 = B_3 = 0$

H_A : $B_1 \neq 0$ ή $B_2 \neq 0$ ή $B_3 \neq 0$

$F = \frac{MSB}{MSE} = \frac{3453{,}475}{507} = 6{,}81$ $\quad F_{2,9,a}$

$F > F_{2,9,a}$: Ho : Απορρίπτεται!

C) Έλεγχος αλληλεπιδράσεων A, B

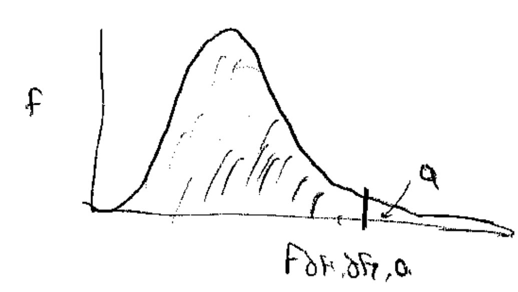
$H_0 : (ab)_{11} = (ab)_{12} = (ab)_{13} = \ldots (ab)_{33} = 0$

A_A : τουλάχιστον ένα $(ab)_{ij} \neq 0$

$F = \frac{MSAB}{MSE} = 0{,}331$ $\quad F_{4,9,01} = F_{4,9,0.05} = 3{,}63$

$F < F_{4,9,a}$

Note



APPLIED STATISTICS

LECTURE 12 (03/05/10)

Είδαμε: ANOVA - 2 παράγοντες

- Σχέσεις αθροισμάτων → Πίνακα (;)

$$SST = SSA + SSB + SSAB + SSE$$

$SST = \sum_{i=1}^{m} \sum_{j=1}^{k} \sum_{l=1}^{n} (x_{ijl} - \bar{x}_{...})^2$

$SSA = nk \sum_{i=1}^{m} (\bar{x}_{i..} - \bar{x}_{...})^2$

$SSB = nm \sum_{j=1}^{k} (\bar{x}_{.j.} - \bar{x}_{...})^2$

$SSAB = n \sum_{i=1}^{m} \sum_{j=1}^{k} (\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x}_{...})^2$

$SSE = \sum_{i=1}^{m} \sum_{j=1}^{k} \sum_{l=1}^{n} (x_{ijl} - \bar{x}_{ij.})^2$

- Παράδειγμα: Ιατρική μελέτη.

Πίνακα (;)

- Εργαστήριο MATLAB (μή ορθογώνιος πίνακας, $n_j \neq n$) ?

- Ασκήσεις 1

Notes: καλύτερα γραφήματα.

Kurtosis.

$$g_2 = \frac{m_4}{m_2^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} = \frac{E(X-\mu)^4}{\sigma^4}$$

$$\text{or } = \frac{m_4}{m_2^2} - 3 = \quad " \quad - 3$$

MATLAB:

k = kurtosis (X) : X : vector → 1 number

X : matrix → kurtosis for each column of X.

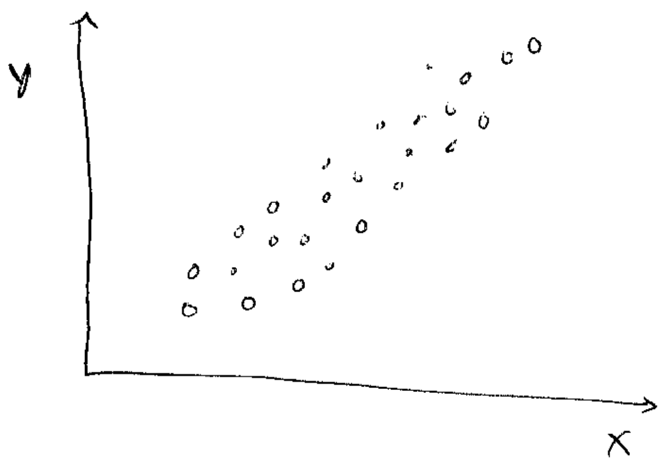
Θεωρητική τιμή: Uniform $k = -1.2$

Normal $k = 0$

ΣΥΣΧΕΤΙΣΗ

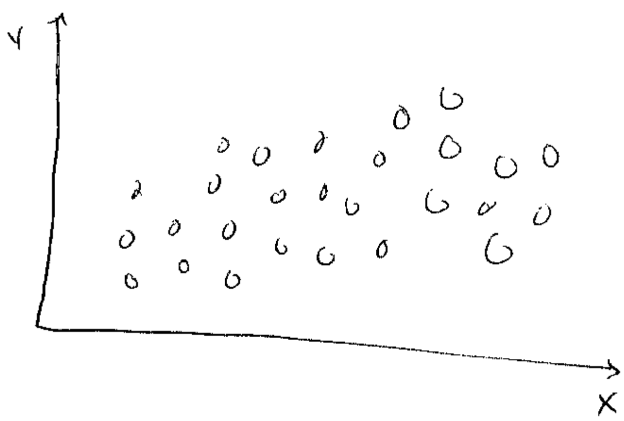
- Σχέσεις μεταξύ τυχαίων μεταβλητών.

π.χ. Γραμμική εξάρτηση μιας τ.μ. από μια ή περισσότερες μεταβλητές. X, Y



Διαγράμματα Διασποράς (Scatter Plots)

← Ισχυρή γραμμική σχέση μεταξύ δύο συνεχών τ.μ.



← Ασθενής γραμμική σχέση

* Συντελεστής Συσχέτισης

Δύο τ.μ. X, Y με διασπορές $\sigma_X^2 = Var[X]$, $\sigma_Y^2 = Var[Y]$

και συνδιασπορά $\sigma_{XY} = Cov[X,Y] = E[(X-\mu_X)(Y-\mu_Y)]$

Note: Συνεχή χώρο.

$\mu_n = \int_{-\infty}^{\infty} (x-c)^n f(x) dx$ E: expected value.

$\mu'_n = E(X^n) = \int_{-\infty}^{\infty} x^n dF(x)$, $\left(E(X) = \sum x_i p(x_i) \right.$

F: cumulative distribution function $F(x) = P(X \le x) = \int^{x} f(t)dt$ $= \int_{-\infty}^{\infty} x f(x) dx$

- Correlation Coefficient.

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$: ανεξάρτητος μονάδων

τιμές στο διάστημα $[-1, 1]$

$\rho \to 1$: ισχυρή θετική συσχέτιση

$\rho \to -1$: » αρνητική »

$\rho \to 0$: γραμμική ανεξαρτησία των X, Y

Ποιοτικός έλεγχος : scatter diagram

* Εκτίμηση του συντελεστή συσχέτισης

(συντελεστής συσχέτισης Pearson) r

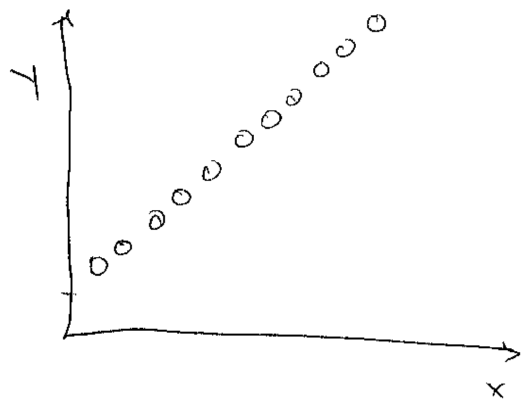
Έστω n ζευγάρια παρατηρήσεων των X, Y

$$\{(x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)\}$$

$$r = \frac{S_{XY}}{S_X S_Y} \quad , \quad S_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \right)$$

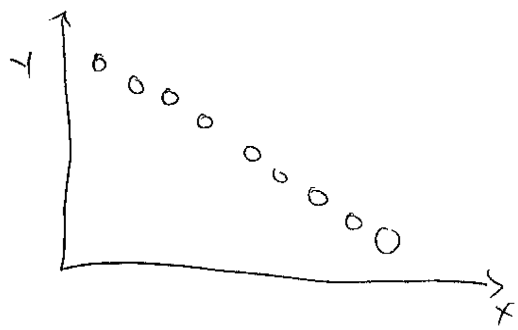
$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

Παραδείγματα : scatter plots

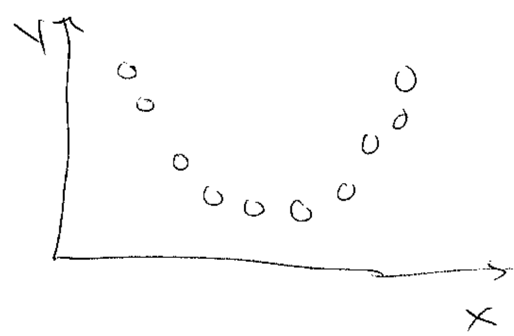


: πλήρης θετική συσχέτιση $r=1$

$Y = Y_0 + \alpha X$



πλήρης αρνητική συσχέτιση $r=-1$



: Ανυπαρξία γραμμικής σχέσης $r=0$

Έλεγχος υποθέσεων : παρόμοιος με τους ελέγχους υποθέσεων που έχουμε αναφέρει 12.5

π.χ.

$H_0: \rho = 0$ (ελέγχει αν τα X, Y είναι γραμμικά ανεξάρτητα)

Το τυπικό σφάλμα του δειγματικού συντελεστή r εκτιμάται από την ποσότητα

$$\hat{se}(r) = \sqrt{\frac{1-r^2}{n-2}}$$

Έλεγχος με το κριτήριο: (για $H_0: \rho = \rho_0$)

$$t = \frac{r - \rho_0}{\sqrt{\frac{(1-r^2)}{(n-2)}}} = r\sqrt{\frac{n-2}{1-r^2}} \qquad \left(r = \frac{S_{xy}}{S_x S_y} \right)$$

Όταν τα ζεύγη (x_i, y_i) λαμβάνονται τυχαία από τον αντίστοιχο πληθυσμό των X, Y (και η X, Y ακολουθούν κανονική κατανομή) η παραπάνω ποσότητα ακολουθεί κατανομή t με $n-2$ βαθμούς ελευθερίας

Προσοχή:

Για $H_0: \rho = \rho_0 \neq 0$ ακολουθείται διαφορετική διαδικασία ελέγχου

Μετασχηματίζοντας την τιμή του δειγματικού συντελεστή συσχέτισης ως

$$z_r = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right)$$ αποδεικνύεται

[Fisher (1921) Metron, 1, 3-21
Μετασχηματισμός z του Fisher]

ότι η δειγματοληπτική κατανομή της ποσότητας z_r είναι κανονική με μέση τιμή

$$z_\rho = \frac{1}{2} \ln\left(\frac{1+\rho}{1-\rho}\right)$$ και τυπική απόκλιση $\frac{1}{\sqrt{n-3}}$

Άρα αν

$H_0: \rho = \rho_0 \neq 0$

$H_A: \rho \neq \rho_0$

Χρησιμοποιείται η $Z = \dfrac{z_r - z_\rho}{\frac{1}{\sqrt{n-3}}}$ η οποία ακολουθεί τυπική κανονική κατανομή.

APPLIED STATISTICS (10/05/10)

LECTURE 13

ΑΝΑΛΥΣΗ ΠΑΛΙΝΔΡΟΜΗΣΗΣ (Regression Analysis)

: Περιγράφει τη μεταβλητότητα μιας τ.μ. Y από τιμές άλλων τ.μ. $X_1 X_2$ κλπ.

Στόχος: εύρεση μοντέλου που περιγράφει την εξάρτηση της Y (εξαρτημένη μεταβλητή) από μια άλλη μεταβλητή X (ανεξάρτητη μεταβλητή)

Απλή Γραμμική Παλινδρόμηση (Simple Linear Regression)

: Η ανεξάρτητη τ.μ. είναι 1 και η εξάρτηση θεωρείται γραμμική.

Υποθέσεις:

1. Η μέση τιμή της τ.μ. Y για κάθε x της X, $E[Y|X=x]$ είναι γραμμική συνάρτηση της x

$$\mu_{Y/x} = E[Y|X=x] = \alpha + \beta x$$

2. Η διασπορά της Y είναι σταθερή $Var[Y|X=x] = \sigma^2_{Y/x} = \sigma^2$

3. Η κατανομή της Y ως προς τη X είναι κανονική

$$Y|X=x \sim N(\alpha + \beta x, \sigma^2)$$

Επιτρέπει παραμετρικό έλεγχο και εκτίμηση των α, β

Γενικά : $Y_i = \alpha + \beta X_i + \varepsilon_i$

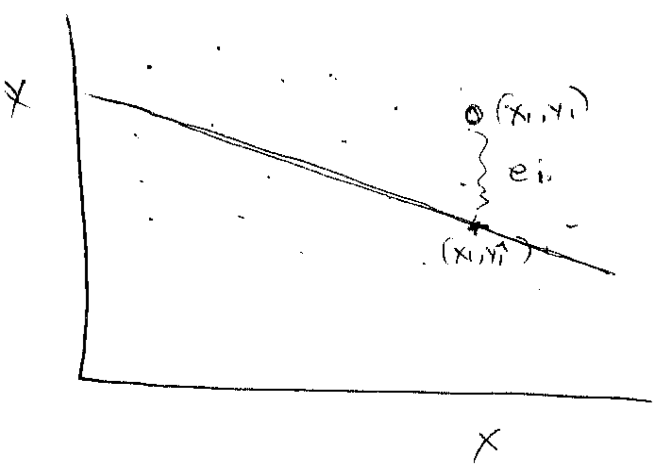
ε_i : σφάλμα παλινδρόμησης (regression error)

$Var[\varepsilon_i] = \sigma_e^2 = \sigma^2$ $\qquad \varepsilon_i = Y_i - \mu_{Y_i/X_i}$

Εκτίμηση Παραμέτρων Απλής Γραμμικής Παλινδρόμησης

Έστω ζεύγη $\{(X_1, Y_1), (X_2, Y_2) \ldots (X_n, Y_n)\}$

Μέθοδος Ελαχίστων Τετραγώνων (Method of Least Squares)



Ελαχιστοποίηση των

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

$$\hat{\beta} = \frac{S_{XY}}{S_X^2} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} \quad (\text{διότι } \bar{Y} = \hat{\alpha} + \hat{\beta}\bar{X})$$

→ Πληθυσμιακή ευθεία της παλινδρόμησης $\mu_{Y/X} = \alpha + \beta X$

Επαγωγική Μέθοδος Εκτίμησης:

Τυπικές αποκλίσεις κατανομών $se(\hat{B}) = \frac{\sigma_{Y/X}}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$

Δειγματικές αποκλίσεις: $se(\hat{\alpha}) = \sigma_{Y/X}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$

$s_{Y/X} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{n-2}}$ Τυπική απόκλιση της Παλινδρόμησης (standard deviation from regression)

$\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$: άθροισμα των τετραγώνων των σφαλμάτων

$e_i \equiv Y_i - \hat{Y}_i$: υπόλοιπο (residual) - σφάλμα ελαχίστων τετραγώνων

or $s^2 \equiv se^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$: $n-2$ γιατί από τους βαθμούς ελευθερίας n του δείγματος αφαιρούμε δύο για τις δύο παραμέτρους που έχουν ήδη εκτιμηθεί!

↳ Βασιζόμενοι στην κλίση της ευθείας $\hat{B}$ μπορούμε να ελέγξουμε την τιμή της κλίσης B της Παλινδρομικής ευθείας

Δηλαδή να ελέγξουμε την υπόθεση

$H_0: B = B_0$

$H_A: B \neq B_0$

Έλεγχος με τη βοήθεια της ποσότητας

$$t = \frac{\hat{\beta} - \beta_0}{se(\hat{\beta})}$$: κατανομή t με n-2 dof.

ή $t = \dfrac{\hat{a} - a_0}{se(a)}$

π.χ. $\beta_0 = 0$ $\mu_{Y|X} = a + 0X = a$. (η Y δεν εξαρτάται γραμμικά από την X)

Διαστήματα εμπιστοσύνης για τις εκτιμώμενες παραμέτρους: (a, β)

$$\left.\begin{array}{l} \hat{a} \pm t_{n-2,1-a/2} \times se(\hat{a}) \\ \hat{\beta} \pm t_{n-2,1-a/2} \times se(\hat{\beta}) \end{array}\right\}$$ a% Διάστημα Εμπιστοσύνης (ΔΕ) των συντελεστών a, β της ευθείας της παλινδρόμησης

<u>Εκτίμηση - πρόβλεψη εξαρτημένης μεταβλητής.</u>

Σημειακή εκτίμηση $\hat{Y}_0$ της μέσης τιμής της Y για κάθε τιμή X_0 της X. Το $(1-a)\%$ Δ.Ε είναι

$$\hat{y} \pm t_{n-2,a/2}\, se(\hat{y}) \equiv (a + \beta x_0) \pm t_{n-2,1-a/2}\, S_{Y/X} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

τυπικό σφάλμα της $\hat{Y}$

$$se(\hat{Y}) = S_{Y/X} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Όρια της πρόβλεψης της Y για μια τιμή X_0 :
(limits of prediction)

↳ $\hat{Y}$: προβλεπόμενη τιμή νέου ατόμου-δεδομένου που προστίθεται στον πληθυσμό.

Πρέπει να συνυπολογίσουμε μια επιπλέον μεταβλητότητα λόγω της διασποράς των τιμών της Y γύρω από τη μέση τιμή του υπο-πληθυσμού στον οποίο ανήκουν.

$$\Rightarrow se(\hat{Y}) = \sqrt{S^2_{Y/X} + se(\bar{Y})^2} = S_{Y/X}\sqrt{1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}$$

<u>Αξιολόγηση Ευθείας Παλινδρόμησης - Συντελεστής Προσδιορισμού</u>

$\Rightarrow R^2 = r^2$ (coefficient of determination) $\left[r = \frac{S_{XY}}{S_X S_Y} \right]$

$r \in [-1,1] \Rightarrow R^2 \in [0,1]$

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{\text{Άθροισμα τετραγώνων από τη Γραμμική παλινδρόμηση}}{\text{Συνολικό Άθροισμα τετραγώνων}}$$

$$r^2 = \frac{\left[\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})\right]^2}{\left[\sum_{i=1}^{n}(X_i-\bar{X})^2\right]\left[\sum_{i=1}^{n}(Y_i-\bar{Y})^2\right]}$$

$\underline{\underline{Q}}$: Απόδειξη ότι $R^2 = r^2$; (($R^2 \Rightarrow \hat{Y}_i - \bar{Y} = (\hat{a} + \hat{B}X_i - \hat{a} - \hat{B}\bar{X})$)) ....

Επίσης καθώς $\sum_{i=1}^{n}(Y_i-\bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2 + \sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$

Έχουμε

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2} = 1 - \frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}$$

Διορθωμένος συντελεστής προσδιορισμού (adjusted coefficient of determination)

$$\bar{R}^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2/n-2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2/n-1} = 1 - \frac{S^2_{Y/X}}{S_Y^2}$$

<u>Έλεγχος υπόθεσης:</u>

$H_0: B = 0$

$H_A: B \neq 0$

με τη βοήθεια του λόγου $F = \frac{MSR}{MSE}$

MSR: Μέσο τετράγωνο της παλινδρόμησης $= \frac{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}{k=(2-1)}$

βαθμοί ελευθερίας γραμμικής παλινδρόμησης

MSE: Μέσο τετράγωνο των σφαλμάτων $= \frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{n-2}$

F κατανομή με $k, n-2$ dof.

Μετασχηματισμοί των Μεταβλητών

Συχνά η σχέση τ.μ. X, Y δεν είναι γραμμική.

↳ Με τη βοήθεια κατάλληλου μετασχηματισμού μπορεί να μετατραπεί σε γραμμική.

π.χ. $Y_t = Y_0 e^{rt} \rightarrow$ τ.μ. $Y, t \rightarrow$

$$\Rightarrow \ln Y_t = \ln(Y_0 e^{rt}) = \ln Y_0 + rt$$

$$y' = \alpha' + \beta t$$

π.χ. $y = \alpha + \beta x^2 \rightarrow x' \rightarrow \sqrt{x}$ $\Rightarrow y = \alpha + \beta x'$

$\sqrt{y} = \sqrt{\alpha x^2} \Rightarrow y' = \alpha' x$

– Γενικά Χρειάζονται Μη Γραμμικά Μοντέλα (Non linear Models)

– Τα παραπάνω ονομάζονται και Γενικευμένα Γραμμικά Μοντέλα (Generalized Linear Models)

π.χ. $\log y = b_1 + b_2 w$

ή $\log\left(\frac{P}{1-P}\right) = b_1 + b_2 w \Rightarrow P = (1-P)e^{b_1+b_2w} = e^{b_1+b_2w} - Pe^{b_1+b_2w} \Rightarrow$

$$\Rightarrow P(1+e^{b_1+b_2w}) = e^{b_1+b_2w} \Rightarrow P = \frac{e^{b_1+b_2w}}{1+e^{b_1+b_2w}} = \frac{1}{1+e^{-b_1-b_2w}}$$

APPLIED STATISTICS

LECTURE 14

Πολλαπλή Γραμμική Παλινδρόμηση (Multiple Linear Regression)

Η τ.μ. Y εξαρτάται γραμμικά από τις $X_1, X_2, \ldots X_k$

$$y_i = \underset{(\beta_0)}{\alpha} + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i \qquad i = 1, 2, \ldots n$$

Υποθέσεις: όπως και στην απλή γραμμική παλινδρόμηση

π.χ.: X_1: ηλικία

X_2: ύψος

X_3: βάρος

Y: ενεργειακή πρόσληψη ενηλίκων $= \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

Σε μορφή πινάκων

$$\underline{\underline{Y}} = \underline{\underline{X}}\,\underline{\beta} + \underline{\varepsilon}$$

$\underline{X}$: dimension $n \times (k+1)$

Y: » $n \times 1$

β: » $(k+1) \times 1$

$\underline{\varepsilon}$: » $n \times 1$

↳ $E[\underline{\varepsilon}] = 0$

$V(\underline{\varepsilon}) = \sigma^2 \underline{I}$

Λύση ελαχίστων τετραγώνων

$$\hat{\beta} = (\underline{\underline{X}}^T \underline{X})^{-1} X^T Y \qquad \left(\text{minimizes error } \underline{\varepsilon}^T \varepsilon\right)$$

$\hat{\varepsilon}_i = Y_i - \hat{Y}_i \quad i = 1, \ldots n$: Υπόλοιπα (residuals)

Εξίσωση παλινδρόμησης $\hat{Y} = \underset{\hat{\beta}_0}{\hat{\alpha}} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots \hat{\beta}_k X_k$

Προσαρμογή της εξίσωσης της πολλαπλής Παλινδρόμησης.

Έχουμε $\sum_{i=1}^{n}(Y_i-\bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2 + \sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$

Συντελεστής πολλαπλού προσδιορισμού

$$R^2_{y.1,2,\dots,k} = \frac{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2} = 1 - \frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}$$

Μερική συσχέτιση

Συντελεστής Μερικής συσχέτισης (Partial correlation coefficient) : συσχέτιση μεταξύ της Y και της X_i όταν οι επιδράσεις των υπολοίπων X έχουν απομακρυνθεί

Υπολογίζεται ως εξής:

1. Εξίσωση Πολλαπλής Παλινδρόμησης μεταξύ της Y και των $X_1 \dots X_k$

2. " " " " X_i και των $X_1,\dots X_{i-1}, X_{i+1},\dots X_k$

3. Δειγματικός συντελεστής συσχέτισης μεταξύ των υπολοίπων των (1) και (2). Οι (1), (2) απομακρύνουν τις γραμμικές επιδράσεις των ανεξάρτητων μεταβλητών $X_1,\dots X_{i-1}, X_{i+1} \dots X_k$ στις X_i και Y

$$r_{Y1.2} = \frac{r_{Y1} - r_{Y2}\,r_{12}}{\sqrt{(1-r_{Y2}^2)(1-r_{12}^2)}} \quad : Y - X_1$$

2 X (X_1, X_2)

$$r_{Y2.1} = \frac{r_{Y2} - r_{Y1}\,r_{12}}{\dots} \quad : Y - X_2.$$

→ Για περισσότερες από 2 ανεξ. μ. απαιτείται Software.

Έλεγχοι Υποθέσεων :

$H_0 : \beta_i = \beta_{i0} \quad , \; i = 1, 2 .. k$

$H_A : \beta_i \neq \beta_{i0}$ (υποθέτουμε ότι οι τιμές των υπολοίπων ανεξ. μ. ~~Xq~~ ≠ X_i παραμένουν σταθερές

Κρίσιμη ποσότητα :

$$t = \frac{\hat{\beta}_i - \beta_{i0}}{se(\hat{\beta}_i)}$$

$$t = \frac{\hat{\alpha} - \alpha_0}{se(\hat{\alpha})}$$

Χρήση κατηγορικών Μεταβλητών

Κωδικοποίηση κατηγορικών μεταβλητών (ψευδομεταβλητές)

π.χ. Δίτιμες Μεταβλητές → (0, 1)

Παράδειγμα: Ενεργειακή Πρόσληψη

X_2 : Φύλο $\begin{cases} 0 & \text{γυναίκες} \\ 1 & \text{άνδρες} \end{cases}$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4$$

↳ ... $Se(\hat{\beta}_i)$ $i = 0, 1, 2, 3, 4$.

Εισαγωγή Όρων Αλληλεπίδρασης - Μη Γραμμικών Όρων

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots$ (Γραμμικοί Όροι)

$+ \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \dots$ (Όροι Αλληλεπίδρασης

$+ \beta_{11} X_1^2 + \beta_{22} X_2^2 + \dots$ (Όροι 2ης τάξης)

$+ \beta_{111} X_1^3 + \beta_{222} X_2^3 + \dots$ (» 3ης »)

⋮ ⋮

APPLIED STATISTICS

LECTURE 14-15 (17/05/10)

ΛΟΓΑΡΙΘΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ (Logistic Regression)

Είδαμε την γραμμική παλινδρόμηση

$$\mu_{Y/x_1,x_2..x_k} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. \beta_k X_k$$

εκτιμάται με τη βοήθεια γραμμικού υποδείγματος

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_k X_k + \varepsilon$$

Διτιμη εξαρτημένη μεταβλητή (π.χ. $Y = 0$ ή 1)
αποτυχία επιτυχία

$\Rightarrow \mu_Y = P = P(Y=1)$: πιθανότητα P της επιτυχίας

$P = \alpha + \beta X$: Δεν είναι κατάλληλο $\rightarrow$ $\alpha + \beta X \in \mathbb{R}$

$P \in [0,1]$

Λύση

$\Rightarrow \dfrac{P}{1-P} \xrightarrow{\text{λογάριθμος}} \ln\left\{\dfrac{P}{1-P}\right\} = \alpha + \beta X$

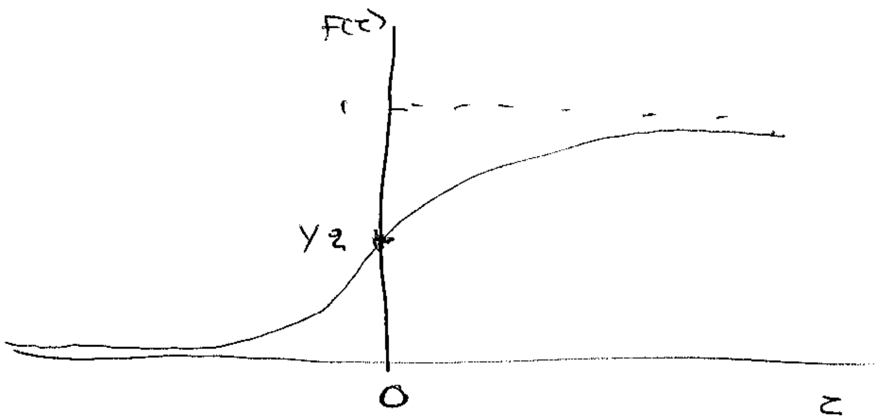
$\hookrightarrow \in [0, +\infty]$ $\qquad \hookrightarrow \in [-\infty, +\infty]$

$$\text{logit}(P) \equiv \ln\left\{\frac{P}{1-P}\right\} = \alpha + \beta X$$

$$\Rightarrow \frac{p}{1-p} = e^{a+\beta x} = e^{z} \Rightarrow p = e^{z} - p e^{z} \Rightarrow$$

$$\Rightarrow p = \frac{e^{z}}{1+e^{z}} = \frac{1}{1+e^{-z}} = f(z)$$

: εκτίμηση πιθανότητας επιτυχίας p της δίτιμης τ.μ. Y.



$F(z) \in [0,1]$

$z \in [-\infty,+\infty]$

Γενική μορφή $\ln\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 X_1 + \dots \beta_k X_k$

$$\Rightarrow \frac{p}{1-p} = e^{\beta_0+\beta_1 X_1+\dots\beta_k X_k}.$$

- κατηγορία: Generalized Linear Models
  Log-linear models

Κάθε ένας από τους β_i εκφράζει τη μεταβολή του $\ln\left(\frac{p}{1-p}\right)$ (σχετικής πιθανότητας) για μια μονάδα αύξησης της αντίστοιχης τ.μ. X_i

π.χ. Δίτιμη τ.μ. φύλο

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X = \beta_0 + \beta_1 (\text{φύλο})$$

$X = 0$: γυναίκες
$= 1$: άνδρες

$$\frac{p_A}{1-p_A} = e^{\beta_0+\beta_1}$$

$$\frac{p_\Gamma}{1-p_\Gamma} = e^{\beta_0}$$

$$\psi = \frac{p_A/(1-p_A)}{p_\Gamma/(1-p_\Gamma)} = e^{\beta_1}$$

Λόγος σχετικών πιθανοτήτων της επιτυχίας

π.χ. πιθανότητα τροχαίου (Accident Analysis and Prevention 32, 815-995 (2000) Chliaoutakis et al.

198 οδηγούς

X_1 : 0 : γυναίκες
1 : άνδρες

$$\ln\left(\frac{p}{1-p}\right) = 0,488 - 0,662 X_1 \quad \rightarrow \quad \psi = e^{-0,662} = 0,52$$

⇒ Σχετική πιθανότητα τροχαίου στους άνδρες είναι 0,52 φορές μικρότερη από την αντίστοιχη των γυναικών

Εκτίμηση συντελεστών με τη χρήση μιας

Συνάρτησης Πιθανοφάνειας (Likehood function)

$$L(B_0, B_1, \dots B_k) = \prod_{i=1}^{n} p_i^{y_i} (1-p_i)^{1-y_i}$$

y_i είναι η τιμή (0, ή 1) της Y για την παρατήρηση $i = 1, 2, \dots n$

p_i : πιθανότητα επιτυχίας της Y " "

$$p_i = \frac{1}{1 + e^{-(B_0 + B_1 X_1 + \dots B_k X_k)}}, \quad i = 1, 2, \dots n$$

L : πιθανότητα να προκύψουν οι τιμές Y_i, $i = 1, 2, \dots n$

(Υπενθυμίζουμε $P(Y=y) = p^y (1-p)^{1-y}$ διωνυμική μεταβλητή)

πιθανότητα να προκύψει μια απ' τις 2 τιμές

- Αποδεικνύεται ότι ένας εκτιμώμενος συντελεστής $\hat{B}_1$ (για αρκετά μεγάλο δείγμα) ακολουθεί κ.κ με μέση τιμή B_1 και τυπικό σφάλμα $se(\hat{B}_1)$ $\quad \dfrac{\hat{B}_1 - B_1}{se(\hat{B}_1)}$: τυπική κανονική κατανομή (τ.κ.κ)

Έλεγχος υποθέσεων.

$H_0 : B_1 = 0$ (ή $e^{B_1} = 1$)

$H_A : B_1 \neq 0$

Με τη βοήθεια του κριτηρίου $z = \frac{\hat{B_i}}{se(\hat{B_i})}$

: κριτήριο του Wald

$\Rightarrow$ εφόσον ισχύει η H_0 το z ακολουθεί z κ. κ.

Μεγάλα δείγματα.

– Μειονέκτημα : Όταν η $\hat{B_i}$ είναι πολύ μεγάλη $se(\hat{B_i})$ είναι αναλογικά πολύ μικρότερη $\rightarrow$

$\rightarrow$ Likelihood ratio statistic (λόγος μέγιστων τιμών συνάρτησης πιθανοφάνειας)

$$-2\ln\left(\frac{\hat{L_i}}{\hat{L_q}}\right) = -2\ln \hat{L_i} - (-2\ln \hat{L_q})$$

$\hat{L_i}$: Max της L όταν ο B_i <u>δεν</u> περιλαμβάνεται στο υπόδειγμα (δηλ. η αντίστοιχη X_i)

$\hat{L_q}$: » » L » » ο B_i περιλαμβάνεται » »
(» »)

=

Μεγιστοποίηση της L $\rightarrow$ για συντελεστές $B_0, B_1 .. B_k$.

Μηδενίζουμε τις μερικές παραγώγους ως προς $B_0, B_1 .. B_k$ της $\ln L$

$L, \ln L$: μεγιστοποιούνται για τις ίδιες τιμές των παραμέτρων $\rightarrow$

για εξισώσεις με για αγνώστους $\rightarrow$ (μεγάλο φορτίο)

Διαστήματα Εμπιστοσύνης

100 $(1-\alpha)\%$ Δ.Ε. για τον B_1

$$\hat{B}_1 \pm z_{\alpha/2}\, se(\hat{B}_1)$$

α

$$e^{[\hat{B}_1 \pm z_{\alpha/2}\, se(\hat{B}_1)]}$$

π.χ. πιθανότητα τροχαίου ατυχήματος

$$\ln\left[\frac{\hat{P}}{1-\hat{P}}\right] = 0{,}488 - 0{,}662 X_1$$

$$se(\hat{B}_1) = 0{,}309$$

$H_0 : B_1 = 0$.

$$\frac{\hat{B}_1}{se(\hat{B}_1)} = -2{,}192 \rightarrow$$

< κρίσιμη τιμής τ.κ.κ. που αντιστοιχεί σε $\alpha = 0{,}05$

η H_0: απορρίπτεται!

→ Άρα H_A: $B_1 \neq 0$.

95% Δ.Ε $\left(e^{[\hat{B}_1 - z_{\alpha/2}\, se(\hat{B}_1)]},\ e^{[\hat{B}_1 + z_{\alpha/2}\, se(\hat{B}_1)]}\right) = (0{,}285, 0{,}933)$