Μάθημα 2ο 08/03/17

Ενας πληθυσμός - ένα δείγμα

χ^2 τεστ καλής προσαρμογής

Εστω $X_1, \dots, X_n$ τ.δ. από πληθυσμό με ασκ. $F(x)$.
Ενδιαφέρει ο έλεγχος $H_0: F(x) = F_0(x)$ έναντι $H_a: F(x) \neq F_0(x)$.
Διαιρούμε τον δειγματικό χώρο σε k κατηγορίες, $i = 1, \dots, k$,
και έστω n_i οι παρατηρούμενες συχνότητες των δεδομένων με $\sum_{i=1}^{k} n_i = n$

Υπολογίζουμε την πιθανότητα η τ.μ. X να πάρει τιμές στις k κατηγορίες και έστω
$p_i = P(X \in i^{\text{οστή}} \text{ κατηγορία} \mid F_0(x))$

Εστω $e_i = np_i$ η αναμενόμενη συχνότητα $[(n_1, \dots, n_k) \sim M(n, p_1, \dots, p_k)$ ή $n_i \sim B(n, p_i)]$

Karl-Pearson

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i} \underset{\text{ασυμπ.}}{\sim} \chi^2_{k-1}$$

γιατί $\sum_{i=1}^{k} n_i = n = \sum_{i=1}^{k} n p_i$

η προσέγγιση είναι καλύτερη αν $e_i \geq 1$
αλλιώς πρέπει να συμπτυχθούν οι κατηγορίες

παράμετροι s για εκτίμηση, τότε $k-s-1$.

ΘΕΩΡΗΜΑ (Ζωγράφος-σελ 15): Αν $(n_1, \ldots, n_k) \sim M(n, p_1, \ldots, p_k)$ τότε η σ.σ.

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_{i_0})^2}{np_{i_0}} \underset{\text{ασυμ.}}{\sim} \left(\xrightarrow[n\to\infty]{d}\right) \chi^2_{k-1}$$ όταν η $H_0: p_i = p_{i_0}$, p_{i_0} γνωστό, είναι αληθής

όπου, $0 < p_{i_0} < 1$, $i = 1, \ldots, k$, $\sum_{i=1}^{k} p_{i_0} = 1$

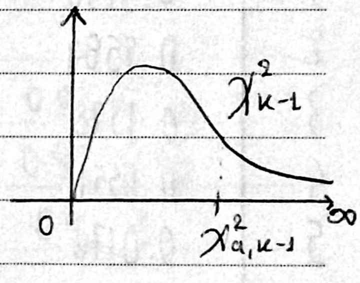
ΠΑΡΑΔΕΙΓΜΑ 1. (7.1)

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i} \left(= \sum_{i=1}^{k} \frac{n_i^2}{e_i} - n \right) \sim \chi^2_{k-1} \boxed{(k-s-1)}$$

$n = 90$ $\quad H_0: p_i = \frac{1}{6}$, $i = 1, \ldots, 6$ έναντι $H_a: p_i \neq \frac{1}{6}$, $e_i = n \cdot p_i = 90 \cdot \frac{1}{6} = 15$

πλευρά	1	2	3	4	5	6
συχνότητα (n_i)	13	17	16	18	12	14
αναμενόμενη συχνότητα (e_i)	15	15	15	15	15	15

$$\chi^2 = \sum_{i=1}^{6} \frac{(n_i - e_i)^2}{e_i} = \frac{(13-15)^2}{15} + \ldots + \frac{(14-15)^2}{15} = 1.87$$



$\chi^2_{0.05,\,6-1} = \chi^2_{0.05,\,5} = 11.07$
$\uparrow$
a

$1.87 < 11.07$ δεν απορ. H_0.

ΠΑΡΑΔΕΙΓΜΑ 2 (7.2), **ΠΑΡΑΔΕΙΓΜΑ 3 (1.5, 1.6 - Μπατσίδης)**

ΠΑΡΑΔΕΙΓΜΑ 4 (ανάλογο του 2)
$n = 340$
Τα δεδομένα έχουν ως εξής:

μεταβλήθηκε 3 φορές, 45 μέρες

Αριθμός ημερήσιων μεταβολών (X)	0	1	2	3	4	5	6	≥7
συχνότητα (n_i)	62	123	84	45	17	5	3	1

X = αριθμός μεταβολών της μετοχής κατά τη διάρκεια μιας μέρας.

$H_0: F(x) = F_0(x) = P(\lambda) = P(X=x) = \frac{e^{-x}\lambda^x}{x!}$, $x = 0, 1, \ldots$

$H_a: F(x) \neq P(\lambda)$

$\hat{\lambda} = \bar{x} = \frac{\sum_1^6 x_i f_i}{n} = 1.6.$

Άρα, $F(x) = F_0(x) = P(\lambda = 1.6)$

X	$P(X=x_i) = p_i$	n_i	$e_i = n \cdot p_i$ (n = 340)	i
0	0.2019	62	68.696	1
1	0.3230	123	109.820	2
2	0.2584	84	87.856	3
3	0.1378	45	46.852	4
4	0.0551	17	18.734	5
5	0.0176	5	5.984	6
6	0.0047	3 } 4	1.5982 } 1.972	7
≥7	0.0011	1	1 > 0.374	8
		340		

↳ πάμε στις γειτονικές κατηγορίες και τις ενώνουμε

$$\chi^2 = \sum_{i=1}^{7} \frac{(n_i - e_i)^2}{e_i} = \frac{(62 - 68.696)^2}{68.696} + \ldots + \frac{(4 - 1.972)^2}{1.972} = 4.874.$$

Επειδή $4.874 < 11.07 = \chi^2_{0.05,5}$ $\quad (k - s - 1 = 7 - 1 - 1)$

δεν απορρίπτω H_0.

Τεστ Kolmogorov-Smirnov

Έστω $X_1, \dots, X_n$ τ.δ. από πληθυσμό με α.σ.κ. $F(x)$.

Ενδιαφέρει: $H_0: F(x) = F_0(x)$ έναντι $H_a: F(x) \neq F_0(x)$

Ο έλεγχος γίνεται με το στατιστικό: $D_n = \sup_x |F_n(x) - F_0(x)|$, όπου, $F_n(x) = \frac{(\text{αρ. παρατηρ. } X_i \leq x)}{n} = \frac{1}{n}\sum_i^n I(X_i \leq x)$

είναι η ε.α.σ.κ.

Ισοδύναμα, $D_n = \max_x |F_n(x) - F_0(x)|$ με κρίσιμη περιοχή $D_n \geq D_{n,a}$

ΠΑΡΑΔΕΙΓΜΑ 1 (7.3)

0.79, 0.15, 0.43, 0.01, 0.93, 0.15, 0.27, 0.19, 0.32, 0.96 || $U(0,1)$, $n=10$.

$H_0: F(x) = F_0(x) \equiv U(0,1)$ έναντι $H_a: F(x) \neq U(0,1)$

$f(x) = 1,\ 0<x<1$ και $F(x) = \begin{cases} 0 & , x<0 \\ x & , 0<x<1 \\ 1 & , x \geq 1 \end{cases}$ αφού $F(x) = \int_0^x 1\,dx$

X_i	0.01	0.15 (2)	0.19	0.27	0.32	0.43	0.79	0.93	0.96
$F_0(x_i)$	0.01	0.15	0.19	0.27	0.32	0.43	0.79	0.93	0.96
$F_n(x_i)$	1/10	3/10	4/10	0.5	0.6	0.7	0.8	0.9	1
$\lvert F_n(x_i) - F_0(x_i)\rvert$	0.09				0.28				

$D_n = \max_{x_i} |F_n(x_i) - F_0(x_i)| = 0.28 < D_{n,a} = D_{10,0.05} = 0.409$, δεν απορρίπτω H_0.

ΠΑΡΑΔΕΙΓΜΑ 2:

33.3, 31.4, 31.0, 33.4, 33.5, 37.0, 36.2, 34.9, 33.7, 34.4

$n=10$, $N(\mu=32, \sigma^2=3.24)$, $\sigma=1.8$.

$X \sim N(\mu, \sigma^2) \Rightarrow Z = \frac{X-\mu}{\sigma} \sim N(0,1)$

$H_0: F(x) = F_0(x) = N(\mu=32, \sigma^2=3.24)$ έναντι $H_a: F(x) \neq N(\mu=32, \sigma^2=3.24)$.

π.χ. $F_0(31.0) = P(X \leq 31.0) = P\left(Z = \frac{X-\mu}{\sigma} \leq \frac{31-32}{1.8}\right) = P(Z \leq -0.55) = 0.2912$ κ.λ.π.

X_i	31.0	31.4	33.3	33.4	33.5	33.7	34.4	34.9	36.2	37.0
$F_0(x_i)$	0.2912	0.3707	0.7642	0.7823	0.7967	0.8264	0.9082	0.9463	0.9901	0.9973
$F_n(x_i)$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$\lvert F_n(x_i)-F_0(x_i)\rvert$	0.1912	0.1767	0.4642	0.3823	0.2967	0.2264	0.2082	0.1463	0.0901	0.0027

Επειδή $\underset{\substack{\uparrow \\ 10}}{D_n} = 0.4642 > D_{n,a} (= D_{10,0.05}) = 0.409$, άρα απορρίπτω την H_0

δηλαδή τα δεδομένα δεν προέρχονται από $N(\mu=32, \sigma^2=3.24)$.