ΜΗ ΠΑΡΑΜΕΤΡΙΚΗ ΣΤΑΤΙΣΤΙΚΗ - ΚΑΤΗΓΟΡΙΚΑ ΔΕΔΟΜΕΝΑ
(Non parametric statistics - Categorical data analysis)
(Distributions Free methods)
(Κεφ.7 Εισαγωγή στη στατιστική - Διδ. Σημειώσεις Μπατσίδη, Ζωγράφου)

Μάθημα 1ο **01/03/2017**

(Εκτός εξεταστέας ύλης)

Κλίμακες μέτρησης

ποιοτικές
- Ονοματική ή ονομαστική: μόνο κατηγορίες, π.χ. αριθμοί αυτοκινήτων
- Διατάξιμη: Συγκριτική σχέση, π.χ. βαθμοί αξιωματικών

ποσοτικές
- Διαστηματική: Τιμή στη διαφορά 2 μετρήσεων (το μηδέν δεν είναι σταθερό) π.χ. θερμοκρασί
- Αναλογική ή λόγου: Το σταθερό μηδέν, π.χ. βάρη, ύψη

Πολυωνυμική κατανομή (Multinomial)

Μια δοκιμή με $E_1, \dots, E_k$ δυνατά αποτελέσματα και $p_1 = P(E_1), \dots, p_k = P(E_k)$, $\sum_{i=1}^{k} p_i = 1$.
n ανεξάρτητες δοκιμές και $x_1, \dots, x_k$ εμφανίσεις, αντίστοιχα, $\sum_{i=1}^{k} x_i = n$.
$(X_1, X_2, \dots, X_{k-1}) \sim M(n, p_1, p_2, \dots, p_{k-1})$.

$$P_{X_1,\dots,X_{k-1}}(x_1, \dots, x_{k-1}) = \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k}, \quad \sum_{i=1}^{k} p_i = 1, \quad \sum_{i=1}^{k} x_i = n.$$

ΙΔΙΟΤΗΤΕΣ:

1) $X_i \sim B(n, p_i)$ $\qquad Var(X_i) = n \cdot p_i (1 - p_i)$
2) $(X_i, X_j) \sim M(n, p_i, p_j)$ $\qquad Cov(X_i, X_j) = n \cdot p_i p_j$

$$\sqrt{n}\left(\frac{X_1}{n} - p_1, \dots, \frac{X_k}{n} - p_k\right) \xrightarrow[n \to \infty]{} N_k(0, \Sigma)$$

Εκτίμηση με τη μέθοδο μέγιστης πιθανοφάνειας (ΕΜΠ)

Εστω $X_1, \ldots, X_n$ τ.δ. από πληθυσμό με σ.π. $P(x,\theta)$ ή σ.π.π. $f(x,\theta)$

Η συνάρτηση $L(\theta|x) = \prod_{i=1}^{n} P(x_i,\theta)$ ή $L(\theta|x) = \prod_{i=1}^{n} f(x_i,\theta)$ ονομάζεται συνάρτηση πιθανοφάνειας (likelihood function).

Ο εκτιμητής $\hat{\theta} = \theta(X_1, \ldots, X_n)$ του θ, λέγεται εκτιμητής μέγιστης πιθανοφάνειας (ΕΜΠ) του θ αν $L(\hat{\theta}|x) = \max_{\theta \in \Theta} L(\theta|x)$, όπου Θ ο παραμετρικός χώρος του θ.

ΠΑΡΑΤΗΡΗΣΕΙΣ:

α) Ισοδύναμα αν $\ln L(\hat{\theta}|x)$

β) κανένα, ένα ή πολλά πεπερασμένα.

ΠΑΡΑΔΕΙΓΜΑ 1: Εστω $X_1, \ldots, X_n$ τ.δ. από $E\kappa\theta(1/\theta)$. ΕΜΠ;

ΛΥΣΗ

$$L(\theta|x) = \prod_{i=1}^{n} \frac{1}{\theta} e^{-\frac{x_i}{\theta}} = \frac{1}{\theta^n} e^{-\frac{\sum^n x_i}{\theta}} \text{ και}$$

$$\ln L(\theta|x) = -n\ln\theta - \frac{\sum^n x_i}{\theta},$$

$$\frac{\partial \ln L(\theta|x)}{\partial \theta} = -\frac{n}{\theta} + \frac{\sum x_i}{\theta^2} \quad (=0 \rightarrow)$$

$$\hat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{X} \quad \text{με} \quad \left.\frac{\partial^2 \ln L(\theta|x)}{\partial \theta^2}\right|_{\theta=\hat{\theta}} = \frac{n}{\theta^2} - 2\frac{\sum x_i}{\hat{\theta}^3} = \frac{n^3}{(\sum x_i)^2} - 2\frac{n^3}{(\sum x_i)^2} = -\frac{n^3}{(\sum x_i)^2} < 0$$

άρα maximum

ΠΑΡΑΔΕΙΓΜΑ 2: Έστω τ.δ. $X_1, \ldots, X_n$ από Poisson(θ). ΕΜΠ;

ΛΥΣΗ:

$$L(\theta|x) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \frac{e^{-n\theta}\theta^{\sum x_i}}{\prod_{i=1}^{n} x_i!}$$

$$\ln L(\theta|x) = -n\theta + (\sum x_i)\ln\theta - \sum_{i=1}^{n} \ln x_i!$$

$$\frac{\partial \ln L(\theta|x)}{\partial \theta} = -n + \frac{\sum x_i}{\theta} \quad (=0 \Rightarrow)$$

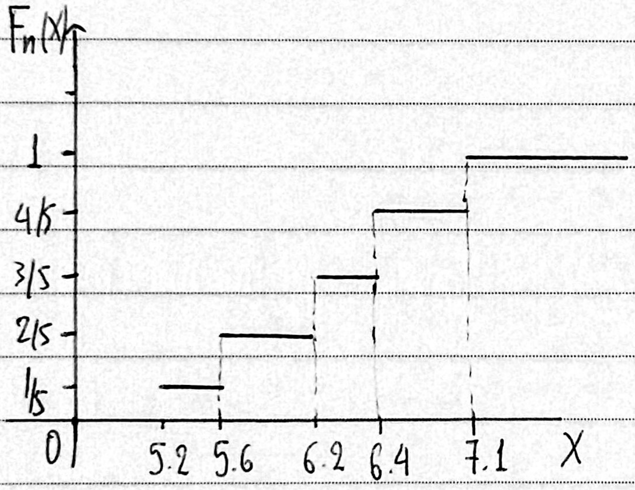
$$\hat{\theta} = \frac{\sum x_i}{n} = \bar{x}$$

$$\left.\frac{\partial^2 \ln L(\theta|x)}{\partial \theta^2}\right|_{\theta=\hat{\theta}} = -\frac{n^2}{\sum x_i} < 0$$, άρα maximum.

Εμπειρική Αθροιστική Συνάρτηση Κατανομής (ε.α.σ.κ.)

ΟΡΙΣΜΟΣ: Έστω $X_1, \dots, X_n$ τ.δ. από πληθυσμό με α.σ.κ. $F(x)$. Η ε.α.σ.κ. $F_n(x)$ ισούται με το μέρος (ποσοστό) των X_i's που είναι μικρότερη ή ίση του x, $-\infty < x < +\infty$. Δηλαδή

$$F_n(x) = \frac{\text{αριθμός των } X_i\text{'s} \leq x}{n} = \frac{\sum_{i=1}^{n} I_{(-\infty, x]}(x_i)}{n}, \quad x \in \mathbb{R}$$ όπου $I_c(x_i) = \begin{cases} 1, & x_i \leq c \\ 0, & \text{αλλιώς} \end{cases}$



Χρόνοι 5 μαθητών λυκείου
6.2, 5.6, 7.1, 6.4, 5.2

ΘΕΩΡΗΜΑ 1: Έστω $X_1, \dots, X_n$ τ.δ. από πληθυσμό με α.σ.κ. $F(x)$. Για σταθερό x, η τ.μ. $n * F_n(x) \sim B(n, F(x))$, όπου $F_n(x)$ η ε.α.σ.κ.

ΑΠΟΔΕΙΞΗ: Οι δυνατές τιμές της ε.α.σ.κ. ανήκουν στο σύνολο $\left\{0, \frac{1}{n}, \dots, \frac{n-1}{n}, 1\right\}$

Για σταθερό x, η τιμή της ε.α.σ.κ. είναι ίση με $\frac{k}{n}$, $k = 0, 1, \dots, n$, αν k το πλήθος δειγματικών τιμών είναι $\leq x$. Άρα, ο υπολογισμός της $P\left(F_n(x) = \frac{k}{n}\right)$ είναι η πιθανότητα να έχουμε k επιτυχίες σε n δοκιμές με $p = P(E) = P(X_i \leq x) = F(x)$

Άρα, $P\left(F_n(x)=\frac{k}{n}\right)=\binom{n}{k}[F(x)]^k[1-F(x)]^{n-k}$ ή

$$P\left(n\cdot F_n(x)=k\right)=\binom{n}{k}[F(x)]^k[1-F(x)]^{n-k}$$ ή

$$n\cdot F_n(x)\sim B(n, F(x)).$$

$E[n\cdot F_n(x)]=n\cdot F(x)\longrightarrow E[F_n(x)]=F(x)$ ← αμερόληπτος εκτιμητής.

$$Var[n\cdot F_n(x)]=n[F(x)][1-F(x)]\longrightarrow Var[F_n(x)]=\frac{F(x)\cdot[1-F(x)]}{n}\xrightarrow[n\to\infty]{}0$$

$$F_n(x)\xrightarrow[n\to\infty]{\text{ασυμ.}}N\left(F(x),\ \frac{F(x)[1-F(x)]}{n}\right)$$