Διάλεξη 2

Διατεταγμένα Στατιστικά

Έστω $X_1, X_2, \dots, X_n$ ένα τ.δ. από F.

$X_{(1)} = \min\{X_1, \dots, X_n\} \rightarrow$ δειγματικό ελάχιστο

$X_{(2)} =$ 2η κατά ↑ σειρά παρατ.

$\vdots$

$X_{(k)} =$ την k κατά ↑ σειρά παρατ.

$\vdots$

$X_{(n)} = \max\{X_1, \dots, X_n\} \rightarrow$ δειγματικό μέγιστο

Καλούνται διατεταγμένα στατιστικά/παρατηρήσεις

$X_{(k)}$ ή $X_{k:n} \rightarrow k$ διατετ. παρατήρηση

Με τα διατ. στατ. ορίζουμε και άλλες ποσότητες ενδιαφέροντος.

$R_n = X_{(n)} - X_{(1)}$ δειγματικό εύρος

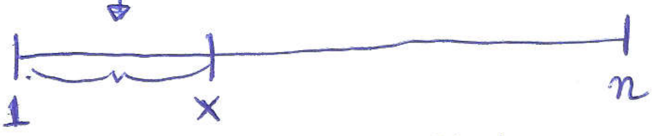
$$M_n = \begin{cases} X_{\left(\frac{n+1}{2}\right)}, & n \text{ περιττός} \\ \dfrac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n+1}{2}\right)}}{2}, & n \text{ άρτιος} \end{cases}$$ δειγματική διάμεσος

Επίσης δειγματικά ποσοστημόρια (υπάρχουν πολλοί ορισμοί).

$Q_a \rightarrow a$ δειγματικό ποσοστημόριο

στατιστικό που αφήνει $\sim 100 \cdot a\,\%$ παρατηρ. αριστερά της

$\{1, 2, \dots, n\}$ και $100 \cdot (1-a)\,\%$ παρατ. δεξιά της



αναζητούμε $x : \dfrac{x-1}{n-1} = a \Rightarrow x = 1 + a \cdot (n-1)$

$= a n + 1 - a$

π.χ. X_1, X_2, X_3, X_4 και θέλουμε $Q_{0.25}$

ας πάρουμε πρώτα $a \cdot n + 1 - a = 0.25 * 4 + 0.75 = 1.75$

$X_{(1.75)} \Rightarrow Q_{0.25} = \frac{1}{4} \cdot X_{(1)} + \frac{3}{4} X_{(2)}$

Γενικά αν $u = a \cdot n + 1 - a$, τότε

$$\boxed{Q_a = (1-w)\, X_{([u])} + w \cdot X_{([u]+1)}, \quad \text{όπου } W = x - [x]}$$

Θα μπορούσε να οριστεί μέσω γενικευμένης αντίστροφης, θα το δούμε αργότερα.

Ιδιαίτερο ενδιαφέρον έχουν οι κατανομές $U_{(i)}$, $1 \le i \le n$ που προέρχονται απο τ.δ. $U_1, \dots, U_n$ απο $U(0,1)$.

<u>Πρόταση</u>

Αν $\{U_i\}_{1 \le i \le n}$ τ.δ. απο $U(0,1)$, τότε

$$U_{(i)} \sim \mathrm{Beta}(i, n+1-i), \quad 1 \le i \le n.$$

<u>Απόδ</u>. Έστω $0 < p < 1$.

Θέτουμε $X_i = \mathbb{1}_{\{U_i \le p\}}$, $1 \le i \le n$.

Προφανώς $X_i \sim Be(p)$, αφού $X_i \in \{0,1\}$ και $P(X_i = 1) = P(U_i \le p) = p$

Παρατηρούμε ότι

$F_{U_{(i)}}(p) = P(U_{(i)} \le p)$ και αν $S_n = \sum_{i=1}^n X_i = \#$ επιτυχιών σε n δοκ. $Be(p)$

$$\{U_{(i)} \le p\} = \{S_n \ge i\}$$

Άρα $F_{U_{(i)}}(p) = P_p(S_n \ge i) = \sum_{k=i}^n P(S_n = k) = \sum_{k=i}^n \binom{n}{k} p^k (1-p)^{n-k}$

$$\Rightarrow f_{U_{(i)}}(p) = \sum_{k=i}^n k \binom{n}{k} p^{k-1} (1-p)^{n-k} - \sum_{k=i}^{n-1} (n-k) \binom{n}{k} p^k (1-p)^{n-k-1}$$

$(k=n \to$ μηδεν.$)$

Όμως

$$k \cdot \binom{n}{k} = k \frac{n!}{k!(n-k)!} = \frac{n!}{(k-1)!(n-k)!} = n \cdot \frac{(n-1)!}{(k-1)!\,[n-1-(k-1)]!}$$

$$= n \cdot \binom{n-1}{k-1} \Rightarrow (n-k)\binom{n}{k} = (n-k)\binom{n}{n-k} = \boxed{n \cdot \binom{n-1}{k}}$$

$$\Rightarrow f_{U_{(i)}}(p) = n\binom{n-1}{i-1} p^{i-1}(1-p)^{n-i}$$
$$+\ n\Bigg[\underbrace{\sum_{k=i+1}^{n}\binom{n-1}{k-1}p^{k-1}(1-p)^{n-k} - \sum_{k=i}^{n-1}\binom{n-1}{k}p^{k}(1-p)^{n-k-1}}_{=0}\Bigg]$$

$$\Rightarrow f_{U_{(i)}}(p) = n\binom{n-1}{i-1}p^{i-1}(1-p)^{n-i}$$
$$\Rightarrow U_{(i)} \sim Beta\,(i, n+1-i)$$

Υπενθύμιση

Αν $X \sim Beta\,(a, b)$, τότε $f_X(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$, $0 < x < 1$,

όπου $B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$, $a, b > 0$ [να γίνει η απόδειξη]

και $\Gamma(a) = \int_0^{+\infty} x^{a-1}e^{-x}dx$, $a>0$, $\Gamma(a+1) = a\Gamma(a)$, $\Gamma(n) = (n-1)!$

Πρόταση

Αν $X_1, \dots, X_n$ τ.δ. από απόλυτα συνεχή κατανομή F, τότε

$$F_{X_{(i)}}(x) = F_{U_{(i)}}(F(x)) \quad \text{και} \quad f_{X_{(i)}}(x) = n\binom{n-1}{i-1}f(x)F^{i-1}(x)(1-F(x))^{n-i}$$

όπου $U_{(i)}$ η i-διατ. παρατήρηση από η ομοιομ. $U_1, \dots, U_n$

Απόδ.

όπως πρίν, αντικαθιστώντας όπου $p = F(x)$.

$$F_{X_{(i)}}(x) = F_{U_{(i)}}(F(x)) \Rightarrow f_{X_{(i)}}(x) = f_{U_{(i)}}(F(x)) \cdot f(x).$$

Μια άλλη απόδειξη μπορεί να δωθεί

παρατηρώντας ότι αν $X \sim F$ και X συνεχής τ.μ,

τότε $F(X) \sim Unif(0,1)$ [η απόδειξη θα γίνει αργότερα στο μάθημα]

και $[F(X)]_{(i)} \stackrel{d}{=} F(X_{(i)})$ [η = ισχύει για $F \nearrow$]

Κατασκευή ακριβούς $(1-a)$ − Δ.Ε. για το p.

Δυϊκότητα Δ.Ε. & Ελέγχων υποθέσεων (έχει δειχθεί για αμφίπλευρα) επέκταση και για μονόπλευρους ελέγχους. Παράδειγμα:

$H_0: \mu=\mu_0$ vs $H_1: \mu > \mu_0$ τ.δ. $N(\mu, 1)$.

ελεγχοσυνάρτηση $\overline{X}_n$, κρίσιμη περιοχή: $\overline{X}_n \geq c_a$

με $P_{\mu_0}[\overline{X}_n \geq c_a] = a$.

Ισοδύναμα με p-value: $P_{\mu_0}[\overline{X}_n \geq \bar{x}] < a$ απορρίπτουμε

($\bar{x}$: τιμή στο δείγμα)

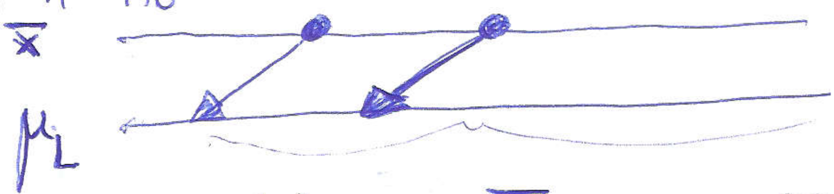
Ποιά είναι η μικρότερη τιμή του μ που θα γίνει δεκτή η H_0?

Λύνουμε $P_{\mu_0}[\overline{X}_n \geq \bar{x}] = a$ ως προς μ_0.

$$\frac{\overline{X}_n - \mu_0}{1/\sqrt{n}} \overset{\mu_0}{\sim} N(0,1) \Rightarrow P_{\mu_0}\left(\overset{N(0,1)}{\tilde{Z}} \geq \frac{\bar{x}-\mu_0}{1/\sqrt{n}}\right) = a$$

$$\Rightarrow \sqrt{n}(\bar{x}-\mu_0) = Z_a \Rightarrow \mu_0 = \bar{x} - Z_a \cdot \frac{1}{\sqrt{n}}$$

$$\underbrace{\text{p-value} \geq a}_{\text{δεκτή η } H_0} \Longleftrightarrow \mu_0 \geq \mu_L(\bar{x}) \equiv \bar{x} - Z_a \cdot \frac{1}{\sqrt{n}}.$$



$$\bar{x} \longrightarrow \mu_L(\bar{x}) \Rightarrow \overline{X}_n \longrightarrow \mu_L(\overline{X}_n) \equiv \mu_L$$

$(\mu_L, +\infty)$ αποτελεί $(1-a)\cdot 100\%$ − Δ.Ε. για το μ (μονόπλευρο).

$$\forall \mu \in \mathbb{R}, \quad P_\mu[\mu_L < \mu] = 1-a \quad \text{(ακριβές)}.$$

Ανάλογα βέβαια, $H_0: \mu = \mu_0$ vs $H_1: \mu < \mu_0$

με $\mu_U = \overline{X}_n + Z_a \cdot \frac{1}{\sqrt{n}}$ αποτελεί $(1-a)\cdot 100$ − ΔΕ για το μ.

Αν ~~[illegible]~~ συμπλέξουμε ελέγχους $a/2$ και $a/2$.

$\mu_L = \overline{X}_n - Z_{a/2} \cdot \frac{1}{\sqrt{n}}$ και $\mu_u = \overline{X}_n + Z_{a/2} \cdot \frac{1}{\sqrt{n}} \Rightarrow$

το (μ_L, μ_u) είναι $(1-a) \cdot 100\%$ Δ.Ε.

Πράγματι $P_\mu(\underbrace{\mu_L < \mu}_{A} \underbrace{< \mu_u}_{B}) = P(AB) = P(A) + P(B) - P(A \cup B)$

$= (1 - \frac{a}{2}) + (1 - \frac{a}{2}) - 1 = 1 - a$, $\underline{\forall \mu \in \mathbb{R}}$.

$\left[P(A \cup B) = P_\mu(\text{"}\mu_L \leq \mu\text{" ή "}\mu < \mu_u\text{"}) = 1 - \underbrace{P(\text{"}\mu_L \geq \mu\text{" \& "}\mu_u < \mu\text{"})}_{0} \right]$ (αφού $\mu_L \leq \mu_u$)

<u>Συμπέρασμα</u>

Ένα $(1-a) \cdot 100\%$ Δ.Ε. μπορεί να κατασκευαστεί από 2 ελέγχους μονόπλευρους με μέγεθος $\frac{a}{2}$ προσδιορίζοντας

$$\left.\begin{array}{ll} \mu_L : & P_{\mu_L}(\overline{X}_n \geq \overline{x}) = \frac{a}{2} \\ \mu_u : & P_{\mu_u}(\overline{X}_n \leq \overline{x}) = \frac{a}{2} \end{array}\right\} \Rightarrow (\mu_L, \mu_u) \text{ είναι } 100(1-a)\% \text{ Δ.Ε. για το } \mu.$$

<u>Για το p της $Be(p)$</u>

$H_0: p = p_0$ vs $H_1: p > p_0$

ελεγχοσυνάρτηση S_n, κρίσιμη περιοχή $S_n \geq s$

Αναζητούμε p_L : $P_{p_L}(S_n \geq \underset{\substack{\downarrow \\ \text{τιμή στο δείγμα}}}{i}) = \frac{a}{2}$, $\underline{i = 1, 2, \ldots, n-1}$

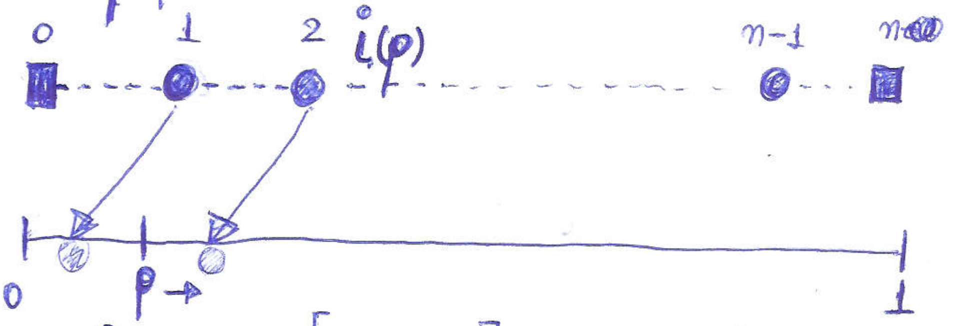
αντίστοιχα p_u : $P_{p_u}(S_n \leq i) = \frac{a}{2}$, $\underline{i = 1, \ldots, n-1}$

<u>Παρατήρηση</u>

$P_p(S_n \geq i) = P(U_{(i)} \leq p)$ και άρα $\nearrow$ ως προς p.

πετυχαίνουμε λύση και μάλιστα μοναδική.

Πώς διαμορφώνεται το Δ.Ε ?



Προσδιορισμός πιθανότητας με p_L : $P_p\left[p_L < p\right] = 1 - P_p\left[p_L \geq p\right] = 1 - P_p\left[S_n \geq i(p)\right]$

Όμως $P_p\left[S_n \geq i(p)\right] \leq P_{p_L(i(p))}\left[S_n \geq i(p)\right] = \frac{a}{2}$

ενδεχομένη διαφυγή μάζας.

$\Longrightarrow$

$$P_p\left[p_L < p\right] \geq 1 - \frac{a}{2} \quad , \forall p \in (0,1)$$

Παρόμοια $P_p\left[p < p_u\right] \geq 1 - \frac{a}{2} \quad , \forall p \in (0,1)$ $\Rightarrow$

$$P_p\left[p_L < p < p_u\right] \geq \left(1 - \frac{a}{2}\right) + \left(1 - \frac{a}{2}\right) - 1 = \underline{1 - a}$$

Τελικά $\inf_{0<p<1} P_p\left[p_L < p < p_u\right] = 1 - a$

<u>Ορολογία</u>

Ένα Δ.Ε. καλείται ακριβές, όταν

$$\inf_{\theta \in \Theta} P_\theta\left[\theta \in I\right] = 1 - a$$

<u>Σχόλ</u> : όμως μπορεί να μην είναι ακριβές για "πολλές" τιμές του θ.

Πόρισμα

Το "ακριβές" $(1-a)$- Δ.Ε. των Clopper-Pearson δίνεται για $i = 1, 2, \ldots, n-1$:

$$I^{CP}_{1-a}(i) = \left(b_{i, n+1-i; 1-\frac{a}{2}},\ b_{i+1, n-i; \frac{a}{2}}\right)$$

~~Αποδ~~ $i=0$: $I^{CP}_{1-a}(0) = (0,\ 1-\sqrt[n]{a})$

Αποδ $i=n$: $I^{CP}_{1-a}(n) = (\sqrt[n]{a},\ 1)$

Για $1 \le i \le n-1$, $P_p(P_L < p) \geq 1-\frac{a}{2}$ & προσδ. από τη σχέση

$$P_p(S_n \geq i) = P_p(U_{(i)} \leq p) = \frac{a}{2} \quad \& \quad U_{(i)} \sim Beta(i, n+1-i)$$

$P_L(i)$

$\Rightarrow \quad P_L(i) = b_{i, n+1-i; 1-\frac{a}{2}} \leftarrow (1-\frac{a}{2})$- άνω ποσοστημόριο

Επίσης το $P_u(i)$:

$$P_p(S_n \le i) = 1 - P_p(S_n \geq i+1) = 1 - P(U_{(i+1)} \le p) = \frac{a}{2}$$

$$\Rightarrow P(U_{(i+1)}(p)) = 1-\frac{a}{2} \quad \Rightarrow P_u(i) = b_{i+1, n-i; \frac{a}{2}}$$

Ασκήσεις

(1) Variance Stabilizing transformation για Δ.Ε. στο p και σύγκριση στην R με το παράδειγμα που δώθηκε

(2) Για $n = 10, 20, 100$ παράγεται δείγμα με $M = 10.000$ (monte-carlo) και εκτιμάτε τις πιθανότητες κάλυψης για 95% - Δ.Ε., για $p = \frac{1}{10}, \frac{3}{10}, \frac{1}{2}$ $(0.1, 0.3, 0.5)$

Ερώτηση: Μπορούμε να το κάνουμε για όλα τα p-μαζί? Ένα γράφημα

(3) (i) Ν.δ.ο. αν $X \sim G(a, \theta)$ και $Y \sim G(b, \theta)$, με $X \perp\!\!\!\perp Y$, τότε $Z = \frac{X}{X+Y} \sim B(a, b)$

(ii) Ν.δ.ο. αν $a = n_1$ και $b = n_2$, $n_1, n_2 \in \mathbb{N}$, τότε

τότε $W = \frac{Y}{X} \sim F_{2n_2, 2n_1}$

(iii) Ν.δ.ο. ότι τα Δ.Ε. δίνονται από ...