Εκτίμηση αθροιστικής συνάρτησης κατανομής

Μη Παραμετρική Στατιστική

Παναγιώτης Παπασταμούλης
Αναπληρωτής Καθηγητής
Τμήμα Στατιστικής ΟΠΑ

papastamoulis@aueb.gr

2026

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ
SCHOOL OF INFORMATION SCIENCES & TECHNOLOGY

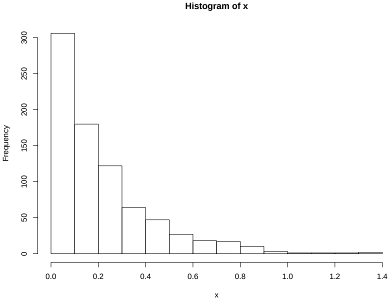
ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
DEPARTMENT OF STATISTICS

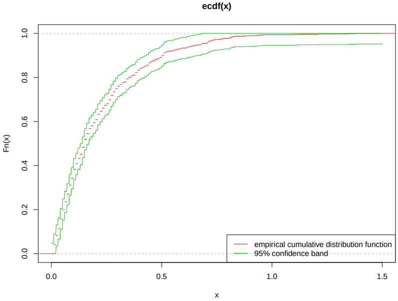
Περιεχόμενα: Εκτίμηση αθροιστικής συνάρτησης κατανομής και συναρτησιακών

1. Εκτίμηση αθροιστικής συνάρτησης κατανομής
   - Εμπειρική συνάρτηση κατανομής
   - Ζώνη εμπιστοσύνης DKW
   - Σημειακά διαστήματα εμπιστοσύνης για την $F(x)$
2. Έλεγχος καλής προσαρμογής Kolmogorov-Smirnov
3. Εκτίμηση συναρτησιακών της συνάρτησης κατανομής
   - Στατιστικά συναρτησιακά και εκτιμητές αντικατάστασης
   - Συνάρτηση επιρροής και μη παραμετρική μέθοδος Δέλτα

Nerve data

- Cox and Lewis (1966)
- $n = 799$ χρόνοι αναμονής μεταξύ διαδοχικών συσπάσεων νευρικής ίνας





Ιστόγραμμα δεδομένων

Εμπειρική συνάρτηση κατανομής + 95% ζώνη εμπιστοσύνης

Εκτίμηση αθροιστικής συνάρτησης κατανομής

Έστω $X_1, \ldots, X_n$ τυχαίο δείγμα από κάποια (άγνωστη) συνάρτηση κατανομής F (διακριτή ή συνεχής).

Ορισμός (Εμπειρική συνάρτηση κατανομής)

Η εμπειρική συνάρτηση κατανομής ορίζεται ως

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}\{X_i \leqslant x\}. \tag{1}$$

Στον παραπάνω ορισμό:

$$\mathrm{I}\{X_i \leqslant x\} = \begin{cases} 1, & \text{αν } X_i \leqslant x \\ 0, & \text{διαφορετικά.} \end{cases}$$

Παράδειγμα ($n = 10$)

- Δεδομένα:

$$(-0.90, 0.18, 1.59, -1.13, -0.08, 0.13, 0.71, -0.24, 1.98, -0.14)$$

- Διατεταγμένο δείγμα

$$(-1.13, -0.90, -0.24, -0.14, -0.08, 0.13, 0.18, 0.71, 1.59, 1.98)$$

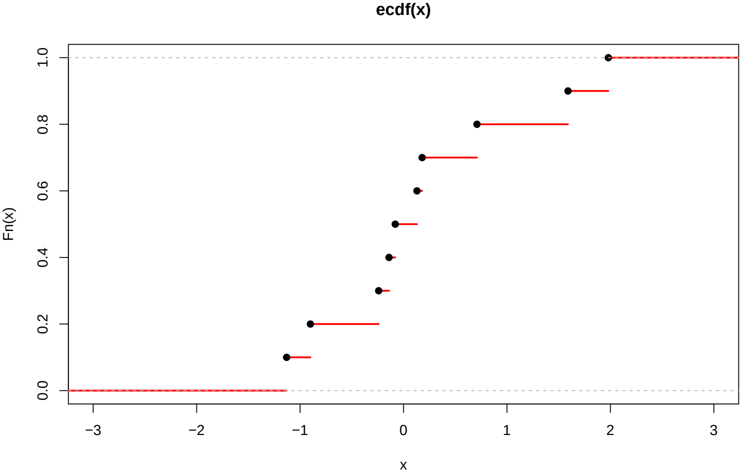
- Υπολογισμός $\widehat{F}_n(x)$

x	$\widehat{F}_n(x)$
$(-\infty, -1.13)$	0
$[-1.13, -0.90)$	$1/10$
$[-0.90, -0.24)$	$2/10$
$[-0.24, -0.14)$	$3/10$
$[-0.14, -0.08)$	$4/10$
$[-0.08, 0.13)$	$5/10$
$[0.13, 0.18)$	$6/10$
$[0.18, 0.71)$	$7/10$
$[0.71, 1.59)$	$8/10$
$[1.59, 1.98)$	$9/10$
$[1.98, \infty)$	1

- Παράδειγμα: $\widehat{F}_n(1.64) = 0.9$ (πραγματική τιμή: $F(1.64) \approx 0.95$)

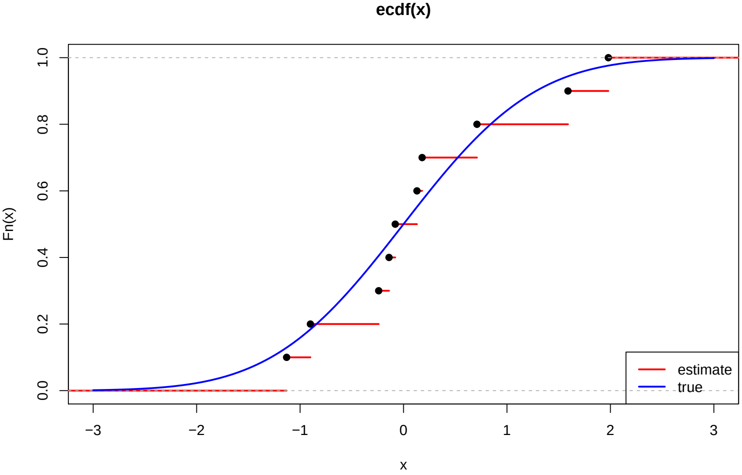
Παράδειγμα ($n = 10$)

Δεδομένα: $(-0.90, 0.18, 1.59, -1.13, -0.08, 0.13, 0.71, -0.24, 1.98, -0.14)$

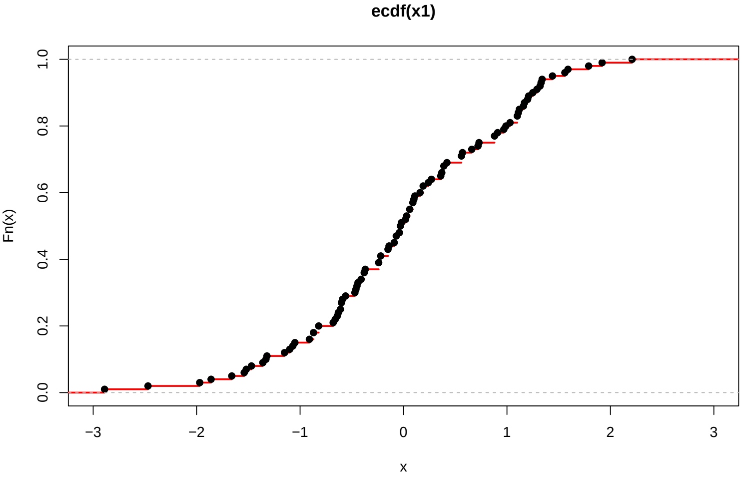


Παράδειγμα ($n = 10$)

Δεδομένα: $(-0.90, 0.18, 1.59, -1.13, -0.08, 0.13, 0.71, -0.24, 1.98, -0.14)$

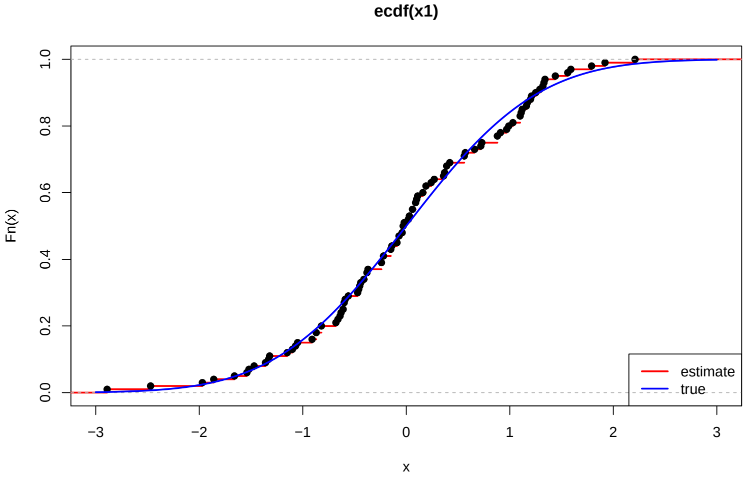


Παράδειγμα ($n = 100$)



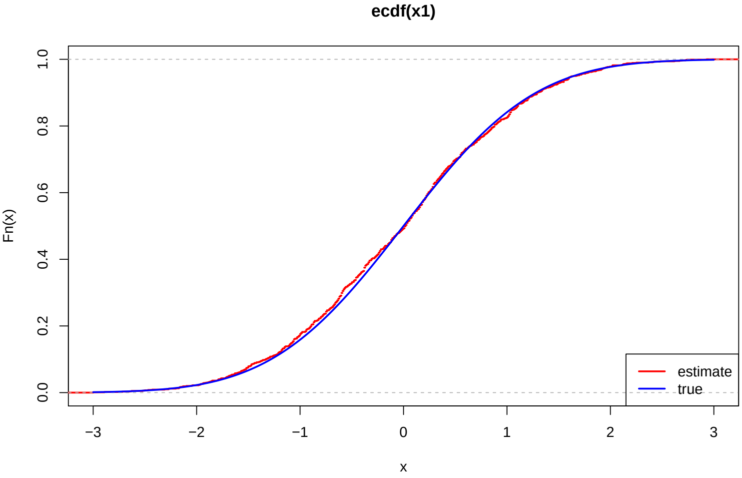
Παράδειγμα: $\widehat{F}_n(1.64) = 0.97$ (πραγματική τιμή: $F(1.64) \approx 0.95$)

Παράδειγμα ($n = 100$)



Παράδειγμα: $\widehat{F}_n(1.64) = 0.97$ (πραγματική τιμή: $F(1.64) \approx 0.95$)

Παράδειγμα ($n = 1000$)



Παράδειγμα: $\widehat{F}_n(1.64) = 0.95$ (πραγματική τιμή: $F(1.64) \approx 0.95$)

Ιδιότητες Εμπειρικής Συνάρτησης Κατανομής

1. **Μέση τιμή και διασπορά**

$$\mathrm{E}\widehat{F}_n(x) = F(x) \quad \text{και} \quad \mathrm{Var}\widehat{F}_n(x) = \frac{F(x)(1-F(x))}{n}$$

   για κάθε $x \in \mathbb{R}$.

2. **Θεμελιώδες θεώρημα Glivenko-Cantelli.** Για κάθε συνάρτηση κατανομής F ισχύει ότι:

$$\sup_{x\in\mathbb{R}} \left|\widehat{F}_n(x) - F(x)\right| \stackrel{a.s.}{\to} 0$$

3. **Ανισότητα DKW (Dvoretzky-Kiefer-Wolfowitz).** Για κάθε $\varepsilon > 0$:

$$\mathrm{P}\left(\sup_{x\in\mathbb{R}} \left|\widehat{F}_n(x) - F(x)\right| > \varepsilon\right) \leqslant 2e^{-2n\varepsilon^2}$$

   (για όλα τα n).

Απόδειξη για το (1)

- Για δοθέν $x \in \mathbb{R}$, έστω η τυχαία μεταβλητή

$$Y = n\widehat{F}_n(x) = \sum_{i=1}^{n} \mathrm{I}\{X_i \leqslant x\} = \text{«πόσα } X \text{ από τα } n \text{ είναι} \leqslant x\text{»}.$$

Απόδειξη για το (1)

- Για δοθέν $x \in \mathbb{R}$, έστω η τυχαία μεταβλητή

$$Y = n\widehat{F}_n(x) = \sum_{i=1}^{n} \mathrm{I}\{X_i \leqslant x\} = \text{«πόσα } X \text{ από τα } n \text{ είναι} \leqslant x\text{»}.$$

- Ξεκάθαρα: $Y \sim \mathcal{B}(n, F(x))$.

Απόδειξη για το (1)

- Για δοθέν $x \in \mathbb{R}$, έστω η τυχαία μεταβλητή

$$Y = n\widehat{F}_n(x) = \sum_{i=1}^{n} \mathrm{I}\{X_i \leqslant x\} = \text{«πόσα } X \text{ από τα } n \text{ είναι} \leqslant x\text{»}.$$

- Ξεκάθαρα: $Y \sim \mathcal{B}(n, F(x))$.
- Συνεπώς για τη μέση τιμή

$$\begin{aligned} \mathrm{E}(Y) &= nF(x), \quad \forall x \in \mathbb{R} \Leftrightarrow \\ \mathrm{E}(n\widehat{F}_n(x)) &= nF(x), \quad \forall x \in \mathbb{R} \Leftrightarrow \\ \mathrm{E}(\widehat{F}_n(x)) &= F(x), \quad \forall x \in \mathbb{R} \end{aligned}$$

Απόδειξη για το (1)

- Για δοθέν $x \in \mathbb{R}$, έστω η τυχαία μεταβλητή

$$Y = n\widehat{F}_n(x) = \sum_{i=1}^{n} \mathrm{I}\{X_i \leqslant x\} = \text{«πόσα } X \text{ από τα } n \text{ είναι } \leqslant x\text{»}.$$

- Ξεκάθαρα: $Y \sim \mathcal{B}(n, F(x))$.
- Συνεπώς για τη μέση τιμή

$$\begin{aligned} \mathrm{E}(Y) &= nF(x), \quad \forall x \in \mathbb{R} \Leftrightarrow \\ \mathrm{E}(n\widehat{F}_n(x)) &= nF(x), \quad \forall x \in \mathbb{R} \Leftrightarrow \\ \mathrm{E}(\widehat{F}_n(x)) &= F(x), \quad \forall x \in \mathbb{R} \end{aligned}$$

- Διασπορά

$$\begin{aligned} \mathrm{Var}(Y) &= nF(x)(1 - F(x)), \quad \forall x \in \mathbb{R} \Leftrightarrow \\ n^2\mathrm{Var}(\widehat{F}_n(x)) &= nF(x)(1 - F(x)), \quad \forall x \in \mathbb{R} \Leftrightarrow \\ \mathrm{Var}(\widehat{F}_n(x)) &= \frac{F(x)(1 - F(x))}{n}, \quad \forall x \in \mathbb{R}. \end{aligned}$$

Συνέπεια του (3): Μία ζώνη εμπιστοσύνης για την F

Μέσω της ανισότητας DKW μπορούμε να κατασκευάσουμε μία **ζώνη εμπιστοσύνης** για τη συνάρτηση κατανομής (F).

- Έστω $\alpha \in (0,1)$.
- Ορίζουμε

$$L_n(x) = \max\left\{\widehat{F}_n(x) - \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}, 0\right\}$$

$$U_n(x) = \min\left\{\widehat{F}_n(x) + \sqrt{\frac{1}{2n}\log\frac{2}{\alpha}}, 1\right\}$$

Λήμμα ($100(1-\alpha)\%$ DKW ζώνη εμπιστοσύνης για την F)

Γιά κάθε F και για κάθε n

$$\mathrm{P}(L_n(x) \leqslant F(x) \leqslant U_n(x) \text{ για κάθε } x) \geqslant 1-\alpha.$$

Απόδειξη

Από την ανισότητα DKW για $\varepsilon_n^2 = \frac{1}{2n}\log\frac{2}{\alpha}$ προκύπτει ότι

$$
\begin{aligned}
\mathrm{P}\left(\sup_{x\in\mathbb{R}}\left|\widehat{F}_n(x)-F(x)\right|>\varepsilon_n\right) &\leqslant 2e^{-2n\varepsilon_n^2} \Leftrightarrow \\
\mathrm{P}\left(\left|\widehat{F}_n(x)-F(x)\right|>\varepsilon_n\right) &\leqslant 2e^{-2n\frac{1}{2n}\log\frac{2}{\alpha}}, \quad \forall x\in\mathbb{R} \Leftrightarrow \\
\mathrm{P}\left(\left|\widehat{F}_n(x)-F(x)\right|>\varepsilon_n\right) &\leqslant \alpha, \quad \forall x\in\mathbb{R} \Leftrightarrow \\
\mathrm{P}\left(\left|\widehat{F}_n(x)-F(x)\right|\leqslant\varepsilon_n\right) &\geqslant 1-\alpha, \quad \forall x\in\mathbb{R} \Leftrightarrow \\
\mathrm{P}\left(-\varepsilon_n\leqslant\widehat{F}_n(x)-F(x)\leqslant\varepsilon_n\right) &\geqslant 1-\alpha, \quad \forall x\in\mathbb{R} \Leftrightarrow \\
\mathrm{P}\left(-\widehat{F}_n(x)-\varepsilon_n\leqslant -F(x)\leqslant -\widehat{F}_n(x)+\varepsilon_n\right) &\geqslant 1-\alpha, \quad \forall x\in\mathbb{R} \Leftrightarrow \\
\mathrm{P}\left(\widehat{F}_n(x)-\varepsilon_n\leqslant F(x)\leqslant \widehat{F}_n(x)+\varepsilon_n\right) &\geqslant 1-\alpha, \quad \forall x\in\mathbb{R}
\end{aligned}
$$

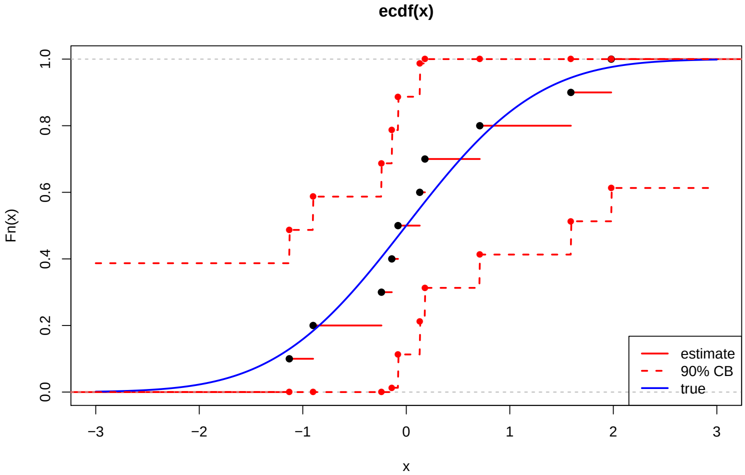
η απόδειξη ολοκληρώνεται θέτοντας $L(x)=\max\{\widehat{F}_n(x)-\varepsilon_n, 0\}$ και $U(x)=\min\{\widehat{F}_n(x)+\varepsilon_n, 1\}$.

Παράδειγμα: 90% ζώνη εμπιστοσύνης ($n = 10$)

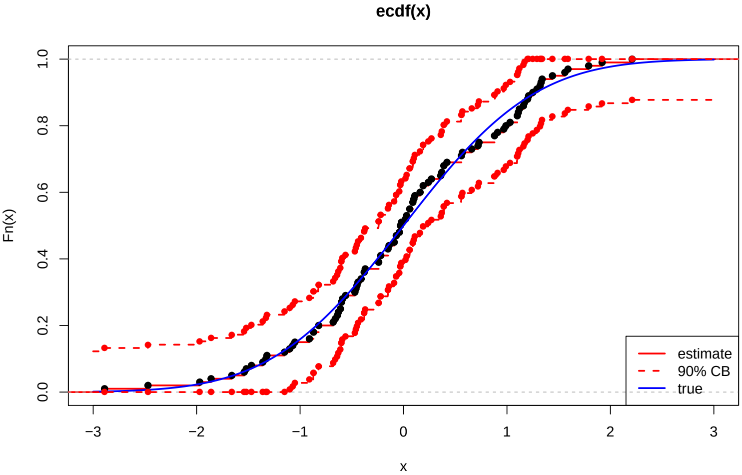
- $n = 10$
- $\alpha = 0.1$
- $\varepsilon_n = \sqrt{\frac{1}{2n} \log \frac{2}{\alpha}} \approx 0.387$
- $L_n(x) = \max\{\widehat{F}_n(x) - 0.387, 0\}$
- $U_n(x) = \min\{\widehat{F}_n(x) + 0.387, 1\}$

x	$\widehat{F}_n(x)$	$L_n(x)$	$U_n(x)$
$(-\infty, -1.13)$	0	0	0.387
$[-1.13, -0.90)$	0.1	0	0.487
$[-0.90, -0.24)$	0.2	0	0.587
$[-0.24, -0.14)$	0.3	0	0.687
$[-0.14, -0.08)$	0.4	0.013	0.787
$[-0.08, 0.13)$	0.5	0.113	0.887
$[0.13, 0.18)$	0.6	0.213	0.987
$[0.18, 0.71)$	0.7	0.313	1
$[0.71, 1.59)$	0.8	0.413	1
$[1.59, 1.98)$	0.9	0.513	1
$[1.98, \infty)$	1	0.613	1

Παράδειγμα: 90% ζώνη εμπιστοσύνης ($n = 10$)



Παράδειγμα: 90% ζώνη εμπιστοσύνης ($n = 100$)



Διαστήματα εμπιστοσύνης για την $F(x)$ σε δοθέν x

- Θέλουμε να καθορίσουμε ℓ_n, u_n τέτοια ώστε

$$\mathrm{P}(\ell_n \leqslant F(x) \leqslant u_n) \geqslant 1 - \alpha$$

- Όμως

$$Y = n\widehat{F}_n(x) \sim \mathcal{B}(n, F(x))$$

- Οπότε το πρόβλημα μπορεί να αναδιατυπωθεί ως εξής:
  - Παρατηρούμε $Y \sim \mathcal{B}(n, p)$
  - $p = F(x) \in [0, 1]$ (άγνωστη παράμετρος)
  - και θέλουμε να βρούμε διάστημα εμπιστοσύνης για το p.
- Μερικές τεχνικές
  1. Clopper-Pearson: Ακριβές Δ.Ε. μέσω της διωνυμικής κατανομής
  2. Wald: ασυμπτωτικό Δ.Ε
  3. Wilson: ασυμπτωτικό Δ.Ε.
  4. μετασχηματισμός arcsin: ασυμπτωτικό Δ.Ε. με σταθεροποιήση (ασυμπτωτικής) διασποράς

Ασυμπτωτικό ΔΕ (Wald)

- Έστω $\widehat{p}_n = Y/n$. Από Κεντρικό Οριακό Θεώρημα

$$\sqrt{n}\frac{\widehat{p}_n - p}{\sqrt{p(1-p)}} \xrightarrow{d} \mathcal{N}(0,1)$$

Ασυμπτωτικό ΔΕ (Wald)

- Έστω $\widehat{p}_n = Y/n$. Από Κεντρικό Οριακό Θεώρημα

$$\sqrt{n}\frac{\widehat{p}_n - p}{\sqrt{p(1-p)}} \xrightarrow{d} \mathcal{N}(0,1)$$

- (Ασθενής) Νόμος Μεγάλων Αριθμών

$$p_n \xrightarrow{P} p$$

Ασυμπτωτικό ΔΕ (Wald)

- Έστω $\widehat{p}_n = Y/n$. Από Κεντρικό Οριακό Θεώρημα

$$\sqrt{n}\frac{\widehat{p}_n - p}{\sqrt{p(1-p)}} \xrightarrow{d} \mathcal{N}(0,1)$$

- (Ασθενής) Νόμος Μεγάλων Αριθμών

$$p_n \xrightarrow{P} p$$

- Από όπου προκύπτει **(γιατί;)**

$$\sqrt{n}\frac{\widehat{p}_n - p}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}} \xrightarrow{d} \mathcal{N}(0,1)$$

Ασυμπτωτικό ΔΕ (Wald)

- Έστω $\widehat{p}_n = Y/n$. Από Κεντρικό Οριακό Θεώρημα

$$\sqrt{n}\frac{\widehat{p}_n - p}{\sqrt{p(1-p)}} \xrightarrow{d} \mathcal{N}(0,1)$$

- (Ασθενής) Νόμος Μεγάλων Αριθμών

$$p_n \xrightarrow{P} p$$

- Από όπου προκύπτει **(γιατί ;)**

$$\sqrt{n}\frac{\widehat{p}_n - p}{\sqrt{\widehat{p}_n(1-\widehat{p}_n)}} \xrightarrow{d} \mathcal{N}(0,1)$$

- Λύνοντας ως προς p προκύπτει ότι το τυχαίο διάστημα

$$\left[\widehat{p}_n - z_{\alpha/2}\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}, \widehat{p}_n + z_{\alpha/2}\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}\right]$$

είναι το $100(1-\alpha)\%$ Α.Δ.Ε ίσων ουρών για το $p = F(x)$.

Παράδειγμα ($n = 10$)

- Δεδομένα:

$$(-0.90, 0.18, 1.59, -1.13, -0.08, 0.13, 0.71, -0.24, 1.98, -0.14)$$

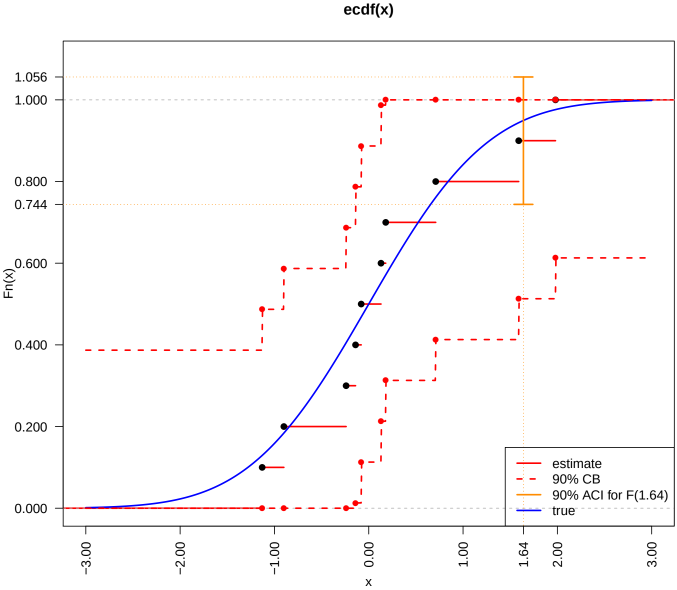
- Να υπολογιστεί το 90% ΑΔΕ ίσων ουρών του Wald για την $F(1.64)$.

- Υπολογισμός $\widehat{F}_n(x)$

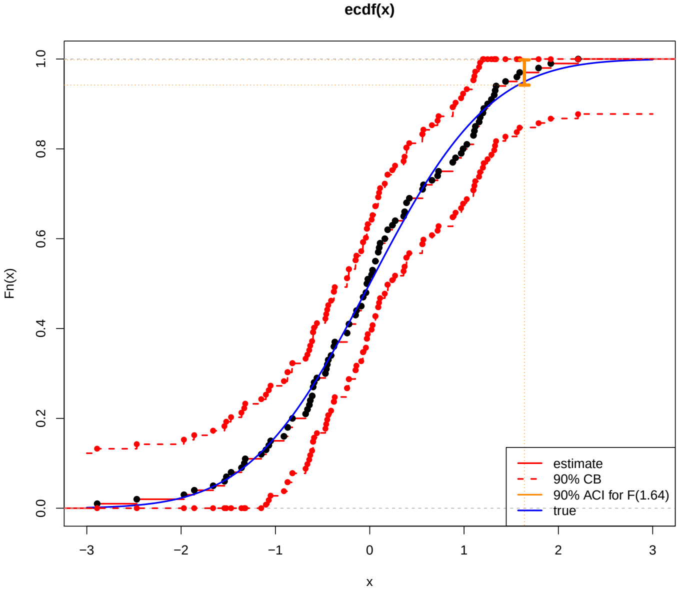
x	$\widehat{F}_n(x)$
$(-\infty, -1.13)$	0
$[-1.13, -0.90)$	$1/10$
$[-0.90, -0.24)$	$2/10$
$[-0.24, -0.14)$	$3/10$
$[-0.14, -0.08)$	$4/10$
$[-0.08, 0.13)$	$5/10$
$[0.13, 0.18)$	$6/10$
$[0.18, 0.71)$	$7/10$
$[0.71, 1.59)$	$8/10$
$[1.59, 1.98)$	$9/10$
$[1.98, \infty)$	1

- Είναι $\widehat{p}_n = \widehat{F}_n(1.64) = 0.9$, $\alpha = 0.1$, $z_{0.05} \approx 1.64$ και $n = 10$
- 90% ΑΔΕ για $F(1.64)$: $[0.744, 1.056]$
- Παρατηρήστε ότι τα άκρα του ΑΔΕ μπορεί να υπερβαίνουν τα όρια του παραμετρικού χώρου.

Παράδειγμα ($n = 10$)



Παράδειγμα ($n = 100$): $[0.942, 0.998]$



Έλεγχος καλής προσαρμογής Kolmogorov-Smirnov

Ενδιαφέρει ο έλεγχος της

$$H_0 : F(x) = F_0(x), \forall x \in \mathbb{R}, \quad \text{vs} \quad H_1 : \exists x : F(x) \neq F_0(x).$$

- Ορίζουμε ως στατιστική συνάρτηση Kolmogorov-Smirnov την

$$D_n = \sup_{x \in \mathbb{R}} \left| \widehat{F}_n(x) - F(x) \right|.$$

- «Μεγάλες» τιμές της $D_n \Rightarrow$ απόρριψη H_0
- Ασυμπτωτικά,

$$\mathrm{P}(\sqrt{n} D_n > k) \xrightarrow{H_0} 2 \sum_{j=1}^{\infty} (-1)^{j+1} e^{-2j^2k^2}$$

- (Ασυμπτωτικό) p-value:

$$2 \sum_{j=1}^{\infty} (-1)^{j+1} e^{-2j^2k^2},$$

όπου k η πραγματοποίηση της $\sqrt{n} D_n$ στο δείγμα.

Η βασική ιδέα για την απόδειξη

Λήμμα

Η κατανομή της τυχαίας μεταβλητής $\sup_{x\in\mathbb{R}} \left|\widehat{F}_n(x) - F(x)\right|$ *δεν εξαρτάται από την* F *(αν* F *συνεχής).*

- Αν $X_i \sim F$, τότε $U_i := F(X_i) \sim \mathcal{U}(0,1)$
- Πράγματι, για $0 \leqslant t \leqslant 1$:

$$
\begin{aligned}
F_U(t) &:= \mathrm{P}(U_i \leqslant t) \\
&= \mathrm{P}(F(X_i) \leqslant t) \\
&= \mathrm{P}(X_i \leqslant F^{-1}(t)) \\
&= F\left(F^{-1}(t)\right) \\
&= t.
\end{aligned}
$$

Η βασική ιδέα για την απόδειξη

Συνεπώς

$$
\begin{aligned}
\sup_{x\in\mathbb{R}} \left|\widehat{F}_n(x) - F(x)\right| &= \sup_{t\in[0,1]} \left|\widehat{F}_n(F^{-1}(t)) - F(F^{-1}(t))\right| \\
&= \sup_{t\in[0,1]} \left|\frac{1}{n}\sum_{i=1}^{n} \mathrm{I}\{X_i \leqslant F^{-1}(t)\} - t)\right| \\
&= \sup_{t\in[0,1]} \left|\frac{1}{n}\sum_{i=1}^{n} \mathrm{I}\{F(X_i) \leqslant t\} - t)\right| \\
&= \sup_{t\in[0,1]} \left|\frac{1}{n}\sum_{i=1}^{n} \mathrm{I}\{U_i \leqslant t\} - t)\right| \\
&= \sup_{t\in[0,1]} \left|\widehat{G}_n(t) - t)\right|,
\end{aligned}
$$

όπου $\widehat{G}_n(\cdot)$ η εμπειρική συνάρτηση κατανομής των $U_i = F(X_i)$, $i = 1, \ldots, n$.

Τι σημαίνει αυτό για το τεστ

Επειδή η κατανομή του D_n δεν εξαρτάται από το F:

- Οι κρίσιμες τιμές είναι καθολικές.
- Μπορούν να υπολογιστούν και να πινακοποιηθούν μία φορά για πάντα.
- Το τεστ ισχύει για **κάθε συνεχή κατανομή**.

Αυτό είναι το βασικό χαρακτηριστικό που καθιστά το τεστ Kolmogorov–Smirnov απαλλαγμένο από παραμέτρους (ελεύθερο κατανομής).

Σύνδεση με το Οριακό Αποτέλεσμα

Καθώς $n \to \infty$,

$$\sqrt{n}\, D_n \;\to\; \sup_{0 \le u \le 1} |B(u)|,$$

όπου $B(u)$ είναι μια «γέφυρα Brown» (Brownian bridge)[1].
Αυτό οδηγεί στην **κατανομή Kolmogorov**, η οποία παρέχει τις ασυμπτωτικές κρίσιμες τιμές του τεστ.

[1] ειδική περίπτωση Gaussian process

Παράδειγμα ($n = 10$)

- $(-0.90, 0.18, 1.59, -1.13, -0.08, 0.13, 0.71, -0.24, 1.98, -0.14)$
- Να ελεγχθεί η υπόθεση ($\alpha = 5\%$)

$$H_0 : F(x) = F_0(x), \forall x \in \mathbb{R}, \quad \text{VS} \quad H_1 : \exists x : F(x) \neq F_0(x),$$

όπου $F_0(x)$ η συνάρτηση κατανομής της $\mathcal{N}(0, 1)$.

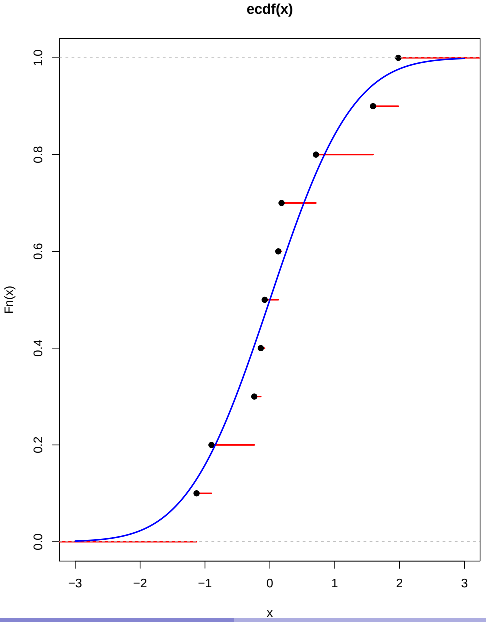
- Υπολογισμός $\widehat{F}_n(x)$

x	$\widehat{F}_n(x)$
$(-\infty, -1.13)$	0
$[-1.13, -0.90)$	1/10
$[-0.90, -0.24)$	2/10
$[-0.24, -0.14)$	3/10
$[-0.14, -0.08)$	4/10
$[-0.08, 0.13)$	5/10
$[0.13, 0.18)$	6/10
$[0.18, 0.71)$	7/10
$[0.71, 1.59)$	8/10
$[1.59, 1.98)$	9/10
$[1.98, \infty)$	1

- Συνάρτηση κατανομής της $\mathcal{N}(0, 1)$: $F_0(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{t^2}{2}\right\} \mathrm{d}t.$

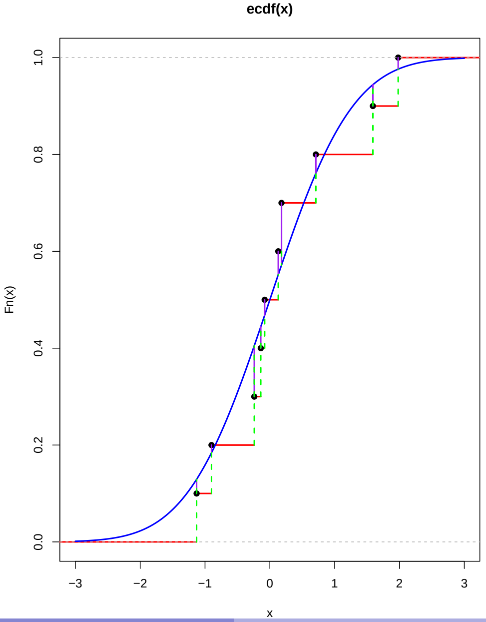
Παράδειγμα (συνέχεια)

Αρκεί το στατιστικό να υπολογιστεί ως max επί των $x = X_i$ και των $x = X_i - \varepsilon$ για μικρό ε.



Παράδειγμα (συνέχεια)

Αρκεί το στατιστικό να υπολογιστεί ως max επί των $x = X_i$ και των $x = X_i - \varepsilon$ για μικρό ε.



Παράδειγμα (συνέχεια)

- Έστω
  - $\widehat{F}_n(x_0^-) = \lim_{x \to x_0^-} \widehat{F}_n(x)$
  - $D^-(x) = |\widehat{F}_n(x^-) - F_0(x)|$
  - $D^+(x) = |\widehat{F}_n(x) - F_0(x)|$
- Για κάθε x στο παρατηρηθέν δείγμα υπολογίζουμε $D^-(x)$ και $D^+(x)$
- Η στατιστική συνάρτηση Kolmogorov-Smirnov ισούται με

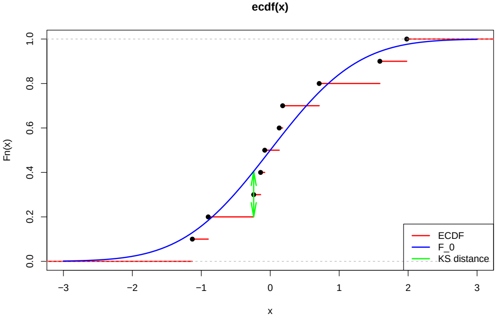
$$D_n = \max_x \{D^-(x), D^+(x)\}.$$

Παράδειγμα (συνέχεια)

	x	$\widehat{F}_n(x^-)$	$\widehat{F}_n(x)$	$F_0(x)$	D^+	D^-
1	-1.130	0.000	0.100	0.129	0.029	0.129
2	-0.900	0.100	0.200	0.184	0.016	0.084
3	-0.240	0.200	0.300	0.405	0.105	0.205
4	-0.140	0.300	0.400	0.444	0.044	0.144
5	-0.080	0.400	0.500	0.468	0.032	0.068
6	0.130	0.500	0.600	0.552	0.048	0.052
7	0.180	0.600	0.700	0.571	0.129	0.029
8	0.710	0.700	0.800	0.761	0.039	0.061
9	1.590	0.800	0.900	0.944	0.044	0.144
10	1.980	0.900	1.000	0.976	0.024	0.076

- Συνεπώς $D_n = 0.205$.
- Η στήλη $F_0(x)$ υπολογίστηκε μέσω της `pnorm()` στην R.

Παράδειγμα (συνέχεια)



- Είναι $k = \sqrt{n}D_n = \sqrt{10} \times 0.205 \approx 0.648$
- (Ασυμπτωτικό) p-value: $2\sum_{j=1}^{\infty}(-1)^{j+1}e^{-2j^2k^2} \approx 0.794 > 0.05$
- Αποτυγχάνουμε να απορρίψουμε την H_0 σε επίπεδο σημαντικότητας 5%.

Ορισμός

Έστω X μια τυχαία μεταβλητή από έναν πληθυσμό με αθροιστική συνάρτηση κατανομής $F(\cdot)$. Ονομάζεται **στατιστικό συναρτησιακό** μια απεικόνιση T, η οποία απεικονίζει την αθροιστική συνάρτηση F σε έναν πραγματικό αριθμό ή διάνυσμα.

Εκτίμηση συναρτησιακών της συνάρτησης κατανομής

- Ενδιαφερόμαστε για την εκτίμηση παραμέτρων που γράφονται ως συναρτήσεις της συνάρτησης κατανομής F, δηλαδή: $T(F)$ όπου T συνάρτηση της F.
    - Μέσος: $\mu = \mathrm{E}(X_1)$ γράφεται ως

$$\mu = \int x \mathrm{d}F = \begin{cases} \int x f(x) \mathrm{d}x, & \text{αν } X_1 \text{ συνεχής} \\ \sum x f(x), & \text{αν } X_1 \text{ διακριτή} \end{cases}$$

      όπου στην συνεχή περίπτωση: $f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x}$ ενώ στην διακριτή περίπτωση $f(x) = \mathrm{P}(X_1 = x) = F(x) - F(x^-)$.
    - (κάτω) α-ποσοστιαίο σημείο: $q_\alpha = F^{-1}(\alpha) = \inf\{x : F(x) \geqslant \alpha\}$.
- Εκτός από τη σημειακή εκτίμηση παραμέτρων, ενδιαφέρει και η συμπερασματολογία
    - Εκτίμηση διασποράς/τυπικών σφαλματών του εκτιμητή
    - Κατασκευή διαστημάτων εμπιστοσύνης για την παράμετρο.

Εκτιμητής αντικατάστασης (plug-in estimator)

Ορισμός (plug-in estimator)

Ο εκτιμητής αντικατάστασης (plug-in estimator) $\widehat{\theta}_n$ *της* $\theta = T(F)$ *ορίζεται ως*

$$\widehat{\theta}_n := T(\widehat{F}_n).$$

Εκτιμητής αντικατάστασης (plug-in estimator)

Ορισμός (plug-in estimator)

Ο εκτιμητής αντικατάστασης (plug-in estimator) $\widehat{\theta}_n$ *της* $\theta = T(F)$ *ορίζεται ως*

$$\widehat{\theta}_n := T(\widehat{F}_n).$$

- Ο εκτιμητής αντικατάστασης προκύπτει απλά αντικαθιστώντας την F με την $\widehat{F}_n$ στη συνάρτηση T.

Εκτιμητής αντικατάστασης (plug-in estimator)

Ορισμός (plug-in estimator)

Ο εκτιμητής αντικατάστασης (plug-in estimator) $\widehat{\theta}_n$ *της* $\theta = T(F)$ *ορίζεται ως*

$$\widehat{\theta}_n := T(\widehat{F}_n).$$

- Ο εκτιμητής αντικατάστασης προκύπτει απλά αντικαθιστώντας την F με την $\widehat{F}_n$ στη συνάρτηση T.
- Σε πολλές περιπτώσεις παίρνουμε ως εκτιμητή το δειγματικό «εμπειρικό» ανάλογο της ποσότητας που μας ενδιαφέρει.

Εκτιμητής αντικατάστασης (plug-in estimator)

Ορισμός (plug-in estimator)

Ο εκτιμητής αντικατάστασης (plug-in estimator) $\widehat{\theta}_n$ της $\theta = T(F)$ ορίζεται ως

$$\widehat{\theta}_n := T(\widehat{F}_n).$$

- Ο εκτιμητής αντικατάστασης προκύπτει απλά αντικαθιστώντας την F με την $\widehat{F}_n$ στη συνάρτηση T.
- Σε πολλές περιπτώσεις παίρνουμε ως εκτιμητή το δειγματικό «εμπειρικό» ανάλογο της ποσότητας που μας ενδιαφέρει.
- Προσοχή:
  - Η πραγματική (F) είναι συνάρτηση κατανομής η οποία, γενικά, μπορεί να είναι συνεχής/διακριτή.
  - Η εμπειρική ($\widehat{F}_n$) είναι διακριτή: θέτει μάζα $1/n$ σε κάθε μία από τις n τιμές του δείγματος.

Γραμμικά συναρτησιακά

Ορισμός (Γραμμικά συναρτησιακά)

Ένα συναρτησιακό της μορφής $T(F) = \int \alpha(x) \mathrm{d}F(x)$ *καλείται γραμμικό συναρτησιακό.*

Γραμμικά συναρτησιακά

Ορισμός (Γραμμικά συναρτησιακά)

Ένα συναρτησιακό της μορφής $T(F) = \int \alpha(x)\mathrm{d}F(x)$ *καλείται γραμμικό συναρτησιακό.*

- Υπενθύμιση:

$$\int \alpha(x)\mathrm{d}F(x) = \begin{cases} \int \alpha(x) f(x)\mathrm{d}x, & \text{(συνεχής περίπτωση)} \\ \sum_j x_j f(x_j), & \text{(διακριτή περίπτωση)} \end{cases}$$

Γραμμικά συναρτησιακά

Ορισμός (Γραμμικά συναρτησιακά)

Ένα συναρτησιακό της μορφής $T(F) = \int \alpha(x)\mathrm{d}F(x)$ καλείται γραμμικό συναρτησιακό.

- Υπενθύμιση:

$$\int \alpha(x)\mathrm{d}F(x) = \begin{cases} \int \alpha(x)f(x)\mathrm{d}x, & \text{(συνεχής περίπτωση)} \\ \sum_j x_j f(x_j), & \text{(διακριτή περίπτωση)} \end{cases}$$

- Επειδή $\widehat{F}_n$ διακριτή, ο **εκτιμητής αντικατάστασης για ένα γραμμικό συναρτησιακό** $T(F) = \int \alpha(x)\mathrm{d}F(x)$ είναι

$$T(\widehat{F}_n) = \int \alpha(x)\mathrm{d}\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \alpha(X_i).$$

Παράδειγμα Ι: μέση τιμή

- Έστω $\mu = T(F) = \int x \mathrm{d}F$.

Παράδειγμα Ι: μέση τιμή

- Έστω $\mu = T(F) = \int x \mathrm{d}F$.
- Είναι γραμμικό συναρτησιακό της F με $\alpha(x) = x$.

Παράδειγμα Ι: μέση τιμή

- Έστω $\mu = T(F) = \int x \mathrm{d}F$.
- Είναι γραμμικό συναρτησιακό της F με $\alpha(x) = x$.
- Εκτιμητής αντικατάστασης:

$$\widehat{\mu} = \int x \mathrm{d}\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \alpha(X_i) = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n.$$

Παράδειγμα I: μέση τιμή

- Έστω $\mu = T(F) = \int x \mathrm{d}F$.
- Είναι γραμμικό συναρτησιακό της F με $\alpha(x) = x$.
- Εκτιμητής αντικατάστασης:

$$\widehat{\mu} = \int x \mathrm{d}\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \alpha(X_i) = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n.$$

- Είναι $\mathrm{Var}(\widehat{\mu}) = \dfrac{\sigma^2}{n}$

Παράδειγμα Ι: μέση τιμή

- Έστω $\mu = T(F) = \int x \mathrm{d}F$.
- Είναι γραμμικό συναρτησιακό της F με $\alpha(x) = x$.
- Εκτιμητής αντικατάστασης:

$$\widehat{\mu} = \int x \mathrm{d}\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \alpha(X_i) = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n.$$

- Είναι $\mathrm{Var}(\widehat{\mu}) = \dfrac{\sigma^2}{n}$
- Τυπικό σφάλμα: $\mathrm{se} = \sqrt{\mathrm{Var}(\widehat{\mu})} = \dfrac{\sigma}{\sqrt{n}}$.

Παράδειγμα I: μέση τιμή

- Έστω $\mu = T(F) = \int x \mathrm{d}F$.
- Είναι γραμμικό συναρτησιακό της F με $\alpha(x) = x$.
- Εκτιμητής αντικατάστασης:

$$\widehat{\mu} = \int x \mathrm{d}\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \alpha(X_i) = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n.$$

- Είναι $\mathrm{Var}(\widehat{\mu}) = \dfrac{\sigma^2}{n}$
- Τυπικό σφάλμα: $\mathrm{se} = \sqrt{\mathrm{Var}(\widehat{\mu})} = \dfrac{\sigma}{\sqrt{n}}$.
- Έστω $\widehat{\sigma}$ εκτιμητής του σ.

Παράδειγμα I: μέση τιμή

- Έστω $\mu = T(F) = \int x \mathrm{d}F$.
- Είναι γραμμικό συναρτησιακό της F με $\alpha(x) = x$.
- Εκτιμητής αντικατάστασης:

$$\widehat{\mu} = \int x \mathrm{d}\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \alpha(X_i) = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n.$$

- Είναι $\mathrm{Var}(\widehat{\mu}) = \dfrac{\sigma^2}{n}$
- Τυπικό σφάλμα: $\mathrm{se} = \sqrt{\mathrm{Var}(\widehat{\mu})} = \dfrac{\sigma}{\sqrt{n}}$.
- Έστω $\widehat{\sigma}$ εκτιμητής του σ.
- Plug-in εκτιμητής του τυπικού σφάλματος: $\widehat{\mathrm{se}} = \dfrac{\widehat{\sigma}}{\sqrt{n}}$.

Παράδειγμα Ι: μέση τιμή

- Έστω $\mu = T(F) = \int x \mathrm{d}F$.
- Είναι γραμμικό συναρτησιακό της F με $\alpha(x) = x$.
- Εκτιμητής αντικατάστασης:

$$\widehat{\mu} = \int x \mathrm{d}\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \alpha(X_i) = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n.$$

- Είναι $\mathrm{Var}(\widehat{\mu}) = \dfrac{\sigma^2}{n}$
- Τυπικό σφάλμα: $\mathrm{se} = \sqrt{\mathrm{Var}(\widehat{\mu})} = \dfrac{\sigma}{\sqrt{n}}$.
- Έστω $\widehat{\sigma}$ εκτιμητής του σ.
- Plug-in εκτιμητής του τυπικού σφάλματος: $\widehat{\mathrm{se}} = \dfrac{\widehat{\sigma}}{\sqrt{n}}$.
- $100(1-\alpha)\%$ διάστημα εμπιστοσύνης (ασυμπτωτικό) για το $\mu = T(F)$

$$\left[\overline{X}_n - z_{\alpha/2}\frac{\widehat{\sigma}}{\sqrt{n}}, \overline{X}_n + z_{\alpha/2}\frac{\widehat{\sigma}}{\sqrt{n}}\right].$$

Παράδειγμα ΙΙ: διασπορά

- Έστω $\sigma^2 = T(F) = \int (x-\mu)^2 \mathrm{d}F$.

Παράδειγμα ΙΙ: διασπορά

- Έστω $\sigma^2 = T(F) = \int (x-\mu)^2 \mathrm{d}F$.
- Επειδή $\mu = \int x \mathrm{d}F(x)$, ισχύει ότι $\sigma^2 = \int x^2 \mathrm{d}F(x) - \left(\int x \mathrm{d}F(x)\right)^2$.

Παράδειγμα II: διασπορά

- Έστω $\sigma^2 = T(F) = \int (x-\mu)^2 \mathrm{d}F$.
- Επειδή $\mu = \int x \mathrm{d}F(x)$, ισχύει ότι $\sigma^2 = \int x^2 \mathrm{d}F(x) - \left(\int x \mathrm{d}F(x)\right)^2$.
- Εκτιμητής αντικατάστασης:

$$
\begin{aligned}
\widehat{\sigma}^2 &= \int x^2 \mathrm{d}\widehat{F}_n(x) - \left(\int x \mathrm{d}\widehat{F}_n(x)\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} (X_i - \widehat{\mu})^2.
\end{aligned}
$$

Παράδειγμα ΙΙ: διασπορά

- Έστω $\sigma^2 = T(F) = \int (x-\mu)^2 \mathrm{d}F$.
- Επειδή $\mu = \int x \mathrm{d}F(x)$, ισχύει ότι $\sigma^2 = \int x^2 \mathrm{d}F(x) - \left(\int x \mathrm{d}F(x)\right)^2$.
- Εκτιμητής αντικατάστασης:

$$\begin{aligned}
\widehat{\sigma}^2 &= \int x^2 \mathrm{d}\widehat{F}_n(x) - \left(\int x \mathrm{d}\widehat{F}_n(x)\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu})^2.
\end{aligned}$$

- Παρατηρήστε ότι $\widehat{\sigma}^2 \neq S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

Παράδειγμα III: λοξότητα

- Έστω μ και σ^2 η μέση τιμή και διασπορά της τυχαίας μεταβλητής X.

Παράδειγμα ΙΙΙ: λοξότητα

- Έστω μ και σ^2 η μέση τιμή και διασπορά της τυχαίας μεταβλητής X.
- Η λοξότητα ορίζεται ως

$$\kappa = \frac{\mathrm{E}(X-\mu)^3}{\sigma^3} = \frac{\int (x-\mu)^3 \mathrm{d}F(x)}{\left\{\int (x-\mu)^2 \mathrm{d}F(x)\right\}^{3/2}}$$

και μετράει την ασυμμετρία της κατανομής της X.

Παράδειγμα III: λοξότητα

- Έστω μ και σ^2 η μέση τιμή και διασπορά της τυχαίας μεταβλητής X.
- Η λοξότητα ορίζεται ως

$$\kappa = \frac{\mathrm{E}(X-\mu)^3}{\sigma^3} = \frac{\int (x-\mu)^3 \mathrm{d}F(x)}{\left\{\int (x-\mu)^2 \mathrm{d}F(x)\right\}^{3/2}}$$

  και μετράει την ασυμμετρία της κατανομής της X.
- Έχουμε ότι $\widehat{\mu} = \overline{X}_n$ και $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \widehat{\mu})^2$.

Παράδειγμα III: λοξότητα

- Έστω μ και σ^2 η μέση τιμή και διασπορά της τυχαίας μεταβλητής X.
- Η λοξότητα ορίζεται ως

$$\kappa = \frac{\mathrm{E}(X-\mu)^3}{\sigma^3} = \frac{\int (x-\mu)^3 \mathrm{d}F(x)}{\left\{\int (x-\mu)^2 \mathrm{d}F(x)\right\}^{3/2}}$$

  και μετράει την ασυμμετρία της κατανομής της X.
- Έχουμε ότι $\widehat{\mu} = \overline{X}_n$ και $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu})^2$.
- Εκτιμητής αντικατάστασης του κ

$$\widehat{\kappa} = \frac{\int (x-\mu)^3 \mathrm{d}\widehat{F}_n(x)}{\left\{\int (x-\mu)^2 \mathrm{d}\widehat{F}_n(x)\right\}^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu})^3}{\widehat{\sigma}^3}.$$

Εφαρμογή

Έστω τυχαίο δείγμα μεγέθους $n = 10$:

$$(-0.90, 0.18, 1.59, -1.13, -0.08, 0.13, 0.71, -0.24, 1.98, -0.14).$$

Να υπολογιστούν οι εκτιμητές αντικατάστασης για τη μέση τιμή, διασπορά και λοξότητα της κατανομής των δεδομένων και ένα 95% Α.Δ.Ε για τη μέση τιμή.

Εφαρμογή

Έστω τυχαίο δείγμα μεγέθους $n = 10$:

$$(-0.90, 0.18, 1.59, -1.13, -0.08, 0.13, 0.71, -0.24, 1.98, -0.14).$$

Να υπολογιστούν οι εκτιμητές αντικατάστασης για τη μέση τιμή, διασπορά και λοξότητα της κατανομής των δεδομένων και ένα 95% Α.Δ.Ε για τη μέση τιμή.

- $\widehat{\mu} = 0.21$

Εφαρμογή

Έστω τυχαίο δείγμα μεγέθους $n = 10$:

$$(-0.90, 0.18, 1.59, -1.13, -0.08, 0.13, 0.71, -0.24, 1.98, -0.14).$$

Να υπολογιστούν οι εκτιμητές αντικατάστασης για τη μέση τιμή, διασπορά και λοξότητα της κατανομής των δεδομένων και ένα 95% Α.Δ.Ε για τη μέση τιμή.

- $\widehat{\mu} = 0.21$
- $\widehat{\sigma}^2 \approx 0.87$ (παρατηρήστε ότι αυτό είναι διαφορετικό του S_n^2)

Εφαρμογή

Έστω τυχαίο δείγμα μεγέθους $n = 10$:

$$(-0.90, 0.18, 1.59, -1.13, -0.08, 0.13, 0.71, -0.24, 1.98, -0.14).$$

Να υπολογιστούν οι εκτιμητές αντικατάστασης για τη μέση τιμή, διασπορά και λοξότητα της κατανομής των δεδομένων και ένα 95% Α.Δ.Ε για τη μέση τιμή.

- $\widehat{\mu} = 0.21$
- $\widehat{\sigma}^2 \approx 0.87$ (παρατηρήστε ότι αυτό είναι διαφορετικό του S_n^2)
- $\widehat{\kappa} \approx 0.54$

  Παρατήρηση: η παραπάνω εκτίμηση της λοξότητας υπολογίζεται στην `R` μέσω της εντολής **`skewness(x, type = 1)`** του πακέτου `'e1071'`.

Εφαρμογή

Έστω τυχαίο δείγμα μεγέθους $n = 10$:

$$(-0.90, 0.18, 1.59, -1.13, -0.08, 0.13, 0.71, -0.24, 1.98, -0.14).$$

Να υπολογιστούν οι εκτιμητές αντικατάστασης για τη μέση τιμή, διασπορά και λοξότητα της κατανομής των δεδομένων και ένα 95% Α.Δ.Ε για τη μέση τιμή.

- $\widehat{\mu} = 0.21$
- $\widehat{\sigma}^2 \approx 0.87$ (παρατηρήστε ότι αυτό είναι διαφορετικό του S_n^2)
- $\widehat{\kappa} \approx 0.54$

  Παρατήρηση: η παραπάνω εκτίμηση της λοξότητας υπολογίζεται στην `R` μέσω της εντολής `skewness(x, type = 1)` του πακέτου `'e1071'`.
- 95% Α.Δ.Ε για το μ: $[-0.37, 0.79]$

  Παρατήρηση: η εκτίμηση του τυπικού σφάλματος του δειγματικού μέσου υπολογίστηκε ως:

$$\widehat{\text{se}} = \frac{\widehat{\sigma}}{\sqrt{n}} \approx \frac{\sqrt{0.87}}{\sqrt{10}} \approx 0.31.$$

Παράδειγμα IV: συντελεστής συσχέτισης

- Έστω $Z = (X, Y)$ διδιάστατη τυχαία μετάβλητη με συνάρτηση κατανομής $F(x, y)$.

Παράδειγμα IV: συντελεστής συσχέτισης

- Έστω $Z = (X, Y)$ διδιάστατη τυχαία μεταβλητή με συνάρτηση κατανομής $F(x, y)$.
- Συντελεστής συσχέτισης $\rho = T(F) = \frac{\mathrm{E}(X-\mu_X)(Y-\mu_Y)}{\sigma_X \sigma_Y} = \frac{\mathrm{E}(XY)-\mu_X \mu_Y}{\sigma_X \sigma_Y}$

Παράδειγμα IV: συντελεστής συσχέτισης

- Έστω $Z = (X, Y)$ διδιάστατη τυχαία μετάβλητη με συνάρτηση κατανομής $F(x, y)$.
- Συντελεστής συσχέτισης $\rho = T(F) = \frac{\mathrm{E}(X-\mu_X)(Y-\mu_Y)}{\sigma_X \sigma_Y} = \frac{\mathrm{E}(XY)-\mu_X \mu_Y}{\sigma_X \sigma_Y}$
- Προφανώς $T(F) = g(T_1(F), T_2(F), T_3(F), T_4(F), T_5(F))$, όπου
  $T_1(F) = \int x \mathrm{d}F(z) \quad T_2(F) = \int y \mathrm{d}F(z) \quad T_3(F) = \int xy \mathrm{d}F(z)$
  $T_4(F) = \int x^2 \mathrm{d}F(z) \quad T_5(F) = \int y^2 \mathrm{d}F(z)$
  και
  $$g(t_1, t_2, t_3, t_4, t_5) = \frac{t_3 - t_1 t_2}{\sqrt{\left(t_4 - t_1^2\right)\left(t_5 - t_2^2\right)}}.$$

Παράδειγμα IV: συντελεστής συσχέτισης

- Έστω $Z = (X, Y)$ διδιάστατη τυχαία μετάβλητη με συνάρτηση κατανομής $F(x, y)$.
- Συντελεστής συσχέτισης $\rho = T(F) = \frac{\mathrm{E}(X-\mu_X)(Y-\mu_Y)}{\sigma_X \sigma_Y} = \frac{\mathrm{E}(XY)-\mu_X\mu_Y}{\sigma_X \sigma_Y}$
- Προφανώς $T(F) = g(T_1(F), T_2(F), T_3(F), T_4(F), T_5(F))$, όπου

  $T_1(F) = \int x \mathrm{d}F(z) \quad T_2(F) = \int y \mathrm{d}F(z) \quad T_3(F) = \int xy \mathrm{d}F(z)$

  $T_4(F) = \int x^2 \mathrm{d}F(z) \quad T_5(F) = \int y^2 \mathrm{d}F(z)$

  και

$$g(t_1, t_2, t_3, t_4, t_5) = \frac{t_3 - t_1 t_2}{\sqrt{\left(t_4 - t_1^2\right)\left(t_5 - t_2^2\right)}}.$$

- Αντικαθιστώντας την F με $\widehat{F}_n$ στις $T_1(F), T_2(F), T_3(F), T_4(F), T_5(F)$ προκύπτει ότι ο εκτιμητής αντικατάστασης είναι

$$\begin{aligned}\widehat{\rho} &= g(T_1(\widehat{F}_n), T_2(\widehat{F}_n), T_3(\widehat{F}_n), T_4(\widehat{F}_n), T_5(\widehat{F}_n)) \\ &= \frac{\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sqrt{\sum_{i=1}^n \left(X_i - \overline{X}_n\right)^2}\sqrt{\sum_{i=1}^n \left(Y_i - \overline{Y}_n\right)^2}}\end{aligned}$$

  δηλαδή ο *δειγματικός συντελεστής συσχέτισης*.

Παράδειγμα V: ποσοστιαία σημεία

- Υπενθύμιση ορισμού: (κάτω) p-ποσοστιαίο σημείο
  - Για F γνησίως αύξουσα $q_p = F^{-1}(p)$
  - δηλαδή το q_p είναι το μοναδικό μέλος του συνόλου $\{x : F(x) = p\}$
  - Γενικός ορισμός μέσω γενικευμένης αντίστροφης συνάρτησης: $q_p = F^{-1}(p) = \inf\{x : F(x) \geqslant p\}$, $p \in (0, 1)$.

Παράδειγμα V: ποσοστιαία σημεία

- Υπενθύμιση ορισμού: (κάτω) p-ποσοστιαίο σημείο
  - Για F γνησίως αύξουσα $q_p = F^{-1}(p)$
  - δηλαδή το q_p είναι το μοναδικό μέλος του συνόλου $\{x : F(x) = p\}$
  - Γενικός ορισμός μέσω γενικευμένης αντίστροφης συνάρτησης: $q_p = F^{-1}(p) = \inf\{x : F(x) \geqslant p\}$, $p \in (0, 1)$.
- Εκτιμητής αντικατάστασης

$$
\begin{aligned}
\widehat{q}_p &= \widehat{F}_n^{-1}(p) \\
&= \inf\{x : \widehat{F}_n(x) \geqslant p\} \\
&= X_{[j]}
\end{aligned}
$$

όπου

- $(X_{[1]}, \ldots, X_{[n]})$ είναι το διατεταγμένο δείγμα
- j ο δείκτης για τον οποίον $\dfrac{j-1}{n} < p \leqslant \dfrac{j}{n}$

Παράδειγμα V: ποσοστιαία σημεία

- Υπενθύμιση ορισμού: (κάτω) p-ποσοστιαίο σημείο
  - Για F γνησίως αύξουσα $q_p = F^{-1}(p)$
  - δηλαδή το q_p είναι το μοναδικό μέλος του συνόλου $\{x : F(x) = p\}$
  - Γενικός ορισμός μέσω γενικευμένης αντίστροφης συνάρτησης: $q_p = F^{-1}(p) = \inf\{x : F(x) \geqslant p\}$, $p \in (0,1)$.
- Εκτιμητής αντικατάστασης

$$\begin{aligned} \widehat{q}_p &= \widehat{F}_n^{-1}(p) \\ &= \inf\{x : \widehat{F}_n(x) \geqslant p\} \\ &= X_{[j]} \end{aligned}$$

όπου
  - $(X_{[1]}, \ldots, X_{[n]})$ είναι το διατεταγμένο δείγμα
  - j ο δείκτης για τον οποίον $\dfrac{j-1}{n} < p \leqslant \dfrac{j}{n}$
- Το $\widehat{q}_p$ είναι το *δειγματικό (κάτω) p-ποσοστιαίο σημείο.*

Παράδειγμα V: ποσοστιαία σημεία

- Παράδειγμα: ο εκτιμητής του $q_{0.5}$ (διάμεσος)

$$\begin{aligned}\widehat{q}_{0.5} &= \widehat{F}_n^{-1}(0.5)\\ &= \inf\{x : \widehat{F}_n(x) \geqslant 0.5\}\\ &= \begin{cases} X_{[n/2]}, & \text{αν } n \text{ άρτιος}\\ X_{[(n+1)/2]}, & \text{αν } n \text{ περιττός}\end{cases}\end{aligned}$$

Παράδειγμα V: ποσοστιαία σημεία

- Παράδειγμα: ο εκτιμητής του $q_{0.5}$ (διάμεσος)

$$\begin{aligned}\widehat{q}_{0.5} &= \widehat{F}_n^{-1}(0.5)\\ &= \inf\{x : \widehat{F}_n(x) \geqslant 0.5\}\\ &= \begin{cases} X_{[n/2]}, & \text{αν } n \text{ άρτιος} \\ X_{[(n+1)/2]}, & \text{αν } n \text{ περιττός} \end{cases}\end{aligned}$$

- Παράδειγμα: εκτιμητής του $q_{0.9}$

$$\begin{aligned}\widehat{q}_{0.9} &= \widehat{F}_n^{-1}(0.9)\\ &= \inf\{x : \widehat{F}_n(x) \geqslant 0.9\}\\ &= X_{[j]}\end{aligned}$$

όπου j ο δείκτης για τον οποίον: $\frac{j-1}{n} < 0.9 \leqslant \frac{j}{n}$. Πχ:
  - Αν $n = 10$: $\widehat{q}_{0.9} = X_{[9]}$ διότι: $\frac{9-1}{10} = 0.8 < 0.9 \leqslant \frac{9}{10} = 0.9$.
  - Αν $n = 11$: $\widehat{q}_{0.9} = X_{[10]}$ διότι: $\frac{10-1}{11} \approx 0.82 < 0.9 \leqslant \frac{10}{11} \approx 0.91$.

Παράδειγμα V: ποσοστιαία σημεία

- Παράδειγμα: ο εκτιμητής του $q_{0.5}$ (διάμεσος)

$$
\begin{aligned}
\widehat{q}_{0.5} &= \widehat{F}_n^{-1}(0.5) \\
&= \inf\{x : \widehat{F}_n(x) \geqslant 0.5\} \\
&= \begin{cases} X_{[n/2]}, & \text{αν } n \text{ άρτιος} \\ X_{[(n+1)/2]}, & \text{αν } n \text{ περιττός} \end{cases}
\end{aligned}
$$

- Παράδειγμα: εκτιμητής του $q_{0.9}$

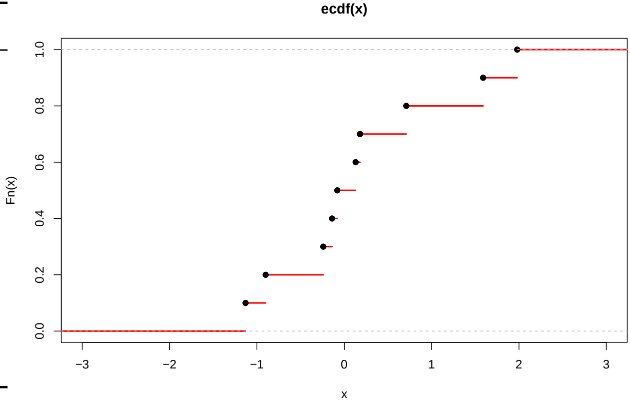
$$
\begin{aligned}
\widehat{q}_{0.9} &= \widehat{F}_n^{-1}(0.9) \\
&= \inf\{x : \widehat{F}_n(x) \geqslant 0.9\} \\
&= X_{[j]}
\end{aligned}
$$

όπου j ο δείκτης για τον οποίον: $\frac{j-1}{n} < 0.9 \leqslant \frac{j}{n}$. Πχ:
  - Αν $n = 10$: $\widehat{q}_{0.9} = X_{[9]}$ διότι: $\frac{9-1}{10} = 0.8 < 0.9 \leqslant \frac{9}{10} = 0.9$.
  - Αν $n = 11$: $\widehat{q}_{0.9} = X_{[10]}$ διότι: $\frac{10-1}{11} \approx 0.82 < 0.9 \leqslant \frac{10}{11} \approx 0.91$.
- Παρατήρηση: στην R `quantile(..., type = 1)`.

Παράδειγμα V: ποσοστιαία σημεία

Δεδομένα: $(-0.90, 0.18, 1.59, -1.13, -0.08, 0.13, 0.71, -0.24, 1.98, -0.14)$

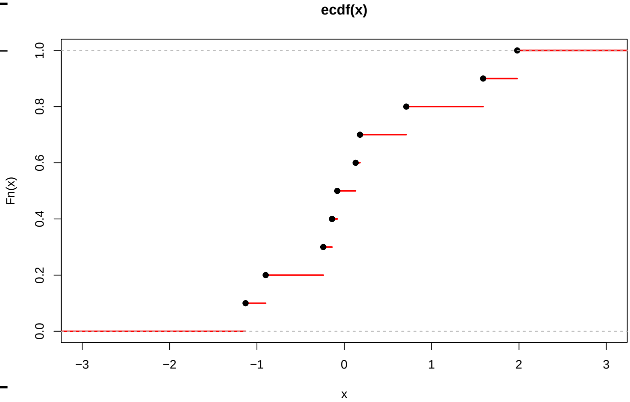
x	$\widehat{F}_n(x)$
$(-\infty, -1.13)$	0
$[-1.13, -0.90)$	$1/10$
$[-0.90, -0.24)$	$2/10$
$[-0.24, -0.14)$	$3/10$
$[-0.14, -0.08)$	$4/10$
$[-0.08, 0.13)$	$5/10$
$[0.13, 0.18)$	$6/10$
$[0.18, 0.71)$	$7/10$
$[0.71, 1.59)$	$8/10$
$[1.59, 1.98)$	$9/10$
$[1.98, \infty)$	1



Παράδειγμα V: ποσοστιαία σημεία

Δεδομένα: $(-0.90, 0.18, 1.59, -1.13, -0.08, 0.13, 0.71, -0.24, 1.98, -0.14)$

x	$\widehat{F}_n(x)$
$(-\infty, -1.13)$	0
$[-1.13, -0.90)$	1/10
$[-0.90, -0.24)$	2/10
$[-0.24, -0.14)$	3/10
$[-0.14, -0.08)$	4/10
$[-0.08, 0.13)$	5/10
$[0.13, 0.18)$	6/10
$[0.18, 0.71)$	7/10
$[0.71, 1.59)$	8/10
$[1.59, 1.98)$	9/10
$[1.98, \infty)$	1

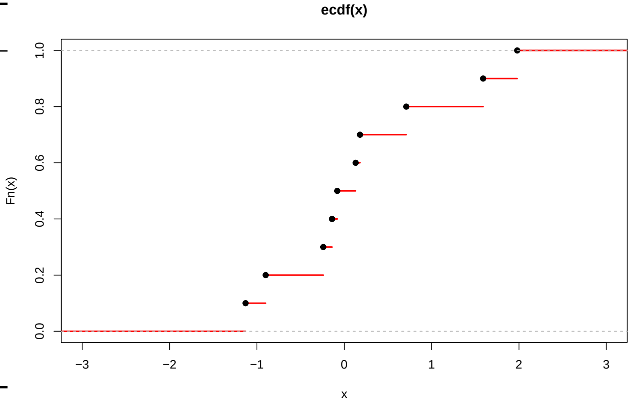


- $\widehat{q}_{0.5} = -0.08$

Παράδειγμα V: ποσοστιαία σημεία

$$\text{Δεδομένα}:\ (-0.90, 0.18, 1.59, -1.13, -0.08, 0.13, 0.71, -0.24, 1.98, -0.14)$$

x	$\widehat{F}_n(x)$
$(-\infty, -1.13)$	0
$[-1.13, -0.90)$	$1/10$
$[-0.90, -0.24)$	$2/10$
$[-0.24, -0.14)$	$3/10$
$[-0.14, -0.08)$	$4/10$
$[-0.08, 0.13)$	$5/10$
$[0.13, 0.18)$	$6/10$
$[0.18, 0.71)$	$7/10$
$[0.71, 1.59)$	$8/10$
$[1.59, 1.98)$	$9/10$
$[1.98, \infty)$	1

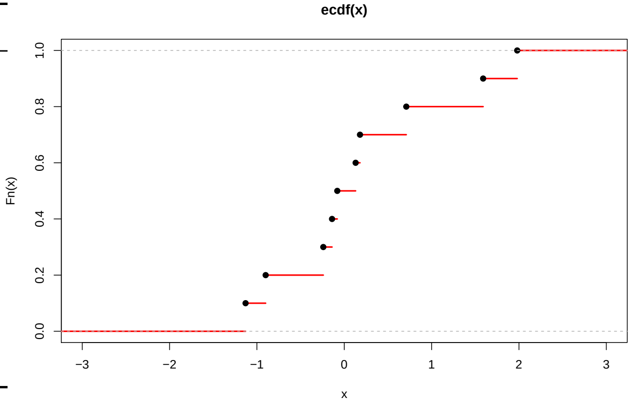


- $\widehat{q}_{0.5} = -0.08$
- $\widehat{q}_{0.9} = 1.59$

Παράδειγμα V: ποσοστιαία σημεία

Δεδομένα: $(-0.90, 0.18, 1.59, -1.13, -0.08, 0.13, 0.71, -0.24, 1.98, -0.14)$

x	$\widehat{F}_n(x)$
$(-\infty, -1.13)$	0
$[-1.13, -0.90)$	1/10
$[-0.90, -0.24)$	2/10
$[-0.24, -0.14)$	3/10
$[-0.14, -0.08)$	4/10
$[-0.08, 0.13)$	5/10
$[0.13, 0.18)$	6/10
$[0.18, 0.71)$	7/10
$[0.71, 1.59)$	8/10
$[1.59, 1.98)$	9/10
$[1.98, \infty)$	1



- $\widehat{q}_{0.5} = -0.08$
- $\widehat{q}_{0.9} = 1.59$
- $\widehat{q}_{0.86} = 1.59$

Εφαρμογή: Διαστήματα εμπιστοσύνης για ποσοστιαία σημεία

- Ακριβή: μέσω της διωνυμικής κατανομής
- Ασυμπτωτικά: μέθοδος Δέλτα

Διωνυμικά Διαστήματα εμπιστοσύνης για ποσοστιαία σημεία

- Σκοπός: κατασκευή $100(1-\alpha)\%$ Δ.Ε για το q_p

Διωνυμικά Διαστήματα εμπιστοσύνης για ποσοστιαία σημεία

- Σκοπός: κατασκευή $100(1-\alpha)\%$ Δ.Ε για το q_p
- Αρκεί να επιλέξουμε $r < s$ τέτοια ώστε:

$$\mathrm{P}\left(X_{[r]} < q_p \leqslant X_{[s]}\right) \geqslant 1-\alpha$$

Διωνυμικά Διαστήματα εμπιστοσύνης για ποσοστιαία σημεία

- Σκοπός: κατασκευή $100(1-\alpha)\%$ Δ.Ε για το q_p
- Αρκεί να επιλέξουμε $r < s$ τέτοια ώστε:

$$\mathrm{P}\left(X_{[r]} < q_p \leqslant X_{[s]}\right) \geqslant 1-\alpha$$

- Παρατηρούμε ότι

$$\mathrm{P}\left(X_{[r]} < q_p \leqslant X_{[s]}\right) = 1 - \left\{\mathrm{P}\left(X_{[r]} \geqslant q_p\right) + \mathrm{P}\left(X_{[s]} < q_p\right)\right\}$$

Διωνυμικά Διαστήματα εμπιστοσύνης για ποσοστιαία σημεία

- Σκοπός: κατασκευή $100(1-\alpha)\%$ Δ.Ε για το q_p
- Αρκεί να επιλέξουμε $r < s$ τέτοια ώστε:

$$\mathrm{P}\left(X_{[r]} < q_p \leqslant X_{[s]}\right) \geqslant 1-\alpha$$

- Παρατηρούμε ότι

$$\mathrm{P}\left(X_{[r]} < q_p \leqslant X_{[s]}\right) = 1 - \left\{\mathrm{P}\left(X_{[r]} \geqslant q_p\right) + \mathrm{P}\left(X_{[s]} < q_p\right)\right\}$$

- Αρκεί το άθροισμα στις αγκύλες να είναι $\leqslant \alpha$:

$$\mathrm{P}\left(X_{[r]} \geqslant q_p\right) \leqslant \frac{\alpha}{2}$$
$$\mathrm{P}\left(X_{[s]} < q_p\right) \leqslant \frac{\alpha}{2}$$

Διωνυμικά Διαστήματα εμπιστοσύνης για ποσοστιαία σημεία

- $\{X_{[s]} < q_p\}$: τουλάχιστον s παρατηρήσεις είναι μικρότερες του q_p
- $\{X_{[r]} \geqslant q_p\}$: το πολύ $r-1$ παρατηρήσεις είναι μικρότερες του q_p
- Ορίζω $Y = \sum_{i=1}^{n} \mathrm{I}\{X_i < q_p\}$
- Για συνεχή F: $\mathrm{P}(X_1 < q_p) = \mathrm{P}(X_1 \leqslant q_p) = p$
- Επειδή X_i ανεξάρτητες και ισόνομες: $Y \sim \mathcal{B}(n, p)$
- Οπότε

$$\mathrm{P}\left(X_{[s]} \leqslant q_p\right) = \mathrm{P}(Y \geqslant s) = \sum_{i=s}^{n} \binom{n}{i} p^i (1-p)^{n-i}$$

- Ομοίως:

$$\begin{aligned}\mathrm{P}\left(X_{[r]} \geqslant q_p\right) &= 1 - \mathrm{P}\left(X_{[r]} < q_p\right) \\ &= 1 - \mathrm{P}(Y \geqslant r) = \sum_{i=0}^{r-1} \binom{n}{i} p^i (1-p)^{n-i}\end{aligned}$$

Διωνυμικά Διαστήματα εμπιστοσύνης για ποσοστιαία σημεία

Συνοψίζοντας, δείξαμε την

Πρόταση

Το τυχαίο διάστημα

$$\left(X_{[r]}, X_{[s]}\right]$$

όπου

- $r = \max\{k = 0, 1, \ldots, n : F_{n,p}(k-1) \leqslant \alpha/2\}$
- $s = \min\{k = 1, \ldots, n+1 : F_{n,p}(k-1) \geqslant 1 - \alpha/2\}$
- $F_{n,p}(\cdot)$ *είναι η αθροιστική συνάρτηση κατανομής της* $\mathcal{B}(n, p)$

είναι ένα $100(1-\alpha)\%$ *διάστημα εμπιστοσύνης για το* q_p, $p \in (0, 1)$.

Σχόλια

Διωνυμικά Διαστήματα εμπιστοσύνης για ποσοστιαία σημεία

Συνοψίζοντας, δείξαμε την

Πρόταση

Το τυχαίο διάστημα

$$\left(X_{[r]}, X_{[s]}\right]$$

όπου

- $r = \max\{k = 0, 1, \ldots, n : F_{n,p}(k-1) \leqslant \alpha/2\}$
- $s = \min\{k = 1, \ldots, n+1 : F_{n,p}(k-1) \geqslant 1 - \alpha/2\}$
- $F_{n,p}(\cdot)$ *είναι η αθροιστική συνάρτηση κατανομής της* $\mathcal{B}(n, p)$

είναι ένα $100(1-\alpha)\%$ *διάστημα εμπιστοσύνης για το* q_p, $p \in (0, 1)$.

Σχόλια

- Στην πρόταση πρέπει να ορίσουμε $X_{[0]} = -\infty$ και $X_{[n+1]} = \infty$.
- Το προηγούμενο Δ.Ε είναι συντηρητικό.

Εφαρμογή

Έστω $(-0.90, 0.18, 1.59, -1.13, -0.08, 0.13, 0.71, -0.24, 1.98, -0.14)$. Να υπολογιστεί το 95% Διωνυμικό διάστημα εμπιστοσύνης για τη διάμεσο.

Εφαρμογή

Έστω $(-0.90, 0.18, 1.59, -1.13, -0.08, 0.13, 0.71, -0.24, 1.98, -0.14)$. Να υπολογιστεί το 95% Διωνυμικό διάστημα εμπιστοσύνης για τη διάμεσο.

- Διάμεσος: $q_{0.5}$ (και σημειακά: $\widehat{q}_{0.5} = -0.08$)

Εφαρμογή

Έστω $(-0.90, 0.18, 1.59, -1.13, -0.08, 0.13, 0.71, -0.24, 1.98, -0.14)$. Να υπολογιστεί το 95% Διωνυμικό διάστημα εμπιστοσύνης για τη διάμεσο.

- Διάμεσος: $q_{0.5}$ (και σημειακά: $\widehat{q}_{0.5} = -0.08$)
- $F_{n,p} = F_{10,0.5}$: συνάρτηση κατανομής της $\mathcal{B}(10, 0.5)$

Εφαρμογή

Έστω $(-0.90, 0.18, 1.59, -1.13, -0.08, 0.13, 0.71, -0.24, 1.98, -0.14)$. Να υπολογιστεί το 95% Διωνυμικό διάστημα εμπιστοσύνης για τη διάμεσο.

- Διάμεσος: $q_{0.5}$ (και σημειακά: $\widehat{q}_{0.5} = -0.08$)
- $F_{n,p} = F_{10,0.5}$: συνάρτηση κατανομής της $\mathcal{B}(10, 0.5)$
- Υπολογισμός r και s

k	$F_{10,0.5}(k-1)$	$x_{[k]}$
0	0.000	$-\infty$
1	0.001	-1.13
2	0.011	-0.90
3	0.055	-0.24
4	0.172	-0.14
5	0.377	-0.08
6	0.623	0.13
7	0.828	0.18
8	0.945	0.71
9	0.989	1.59
10	0.999	1.98
11	1.000	∞

$\alpha = 0.05$

$r = \max\{k = 0, 1, \ldots, 10 : F_{10,0.5}(k-1) \leqslant 0.025\}$

$= 2 \Rightarrow$

$x_{[r]} = x_{[2]} = -0.90$

$s = \min\{k = 1, \ldots, 11 : F_{10,0.5}(k-1) \geqslant 0.975\}$

$= 9 \Rightarrow$

$x_{[s]} = x_{[9]} = 1.59$

Εφαρμογή

Έστω $(-0.90, 0.18, 1.59, -1.13, -0.08, 0.13, 0.71, -0.24, 1.98, -0.14)$. Να υπολογιστεί το 95% Διωνυμικό διάστημα εμπιστοσύνης για τη διάμεσο.

- Διάμεσος: $q_{0.5}$ (και σημειακά: $\widehat{q}_{0.5} = -0.08$)
- $F_{n,p} = F_{10,0.5}$: συνάρτηση κατανομής της $\mathcal{B}(10, 0.5)$
- Υπολογισμός r και s

k	$F_{10,0.5}(k-1)$	$x_{[k]}$
0	0.000	$-\infty$
1	0.001	-1.13
2	0.011	-0.90
3	0.055	-0.24
4	0.172	-0.14
5	0.377	-0.08
6	0.623	0.13
7	0.828	0.18
8	0.945	0.71
9	0.989	1.59
10	0.999	1.98
11	1.000	∞

$\alpha = 0.05$

$r = \max\{k = 0, 1, \ldots, 10 : F_{10,0.5}(k-1) \leqslant 0.025\}$
$= 2 \Rightarrow$

$x_{[r]} = x_{[2]} = -0.90$

$s = \min\{k = 1, \ldots, 11 : F_{10,0.5}(k-1) \geqslant 0.975\}$
$= 9 \Rightarrow$

$x_{[s]} = x_{[9]} = 1.59$

- 95% Δ.Ε για τη διάμεσο: $\left(x_{[2]}, x_{[9]}\right] = (-0.90, 1.59]$

Εφαρμογή (εντολές R)

```
> x <- c(-0.90, 0.18, 1.59, -1.13, -0.08, 0.13, 0.71,
+        -0.24, 1.98, -0.14)
> n <- length(x)
> p <- 0.5
> cbind(0:(n+1), pbinom(q=(-1):n, size = n, prob = p)
+       c(-Inf, sort(x), Inf))
```

Κάποιοι προβληματισμοί

- Γνωρίζουμε ότι $\widehat{F}_n$ είναι «κοντά» στην F (για κάθε F).

Κάποιοι προβληματισμοί

- Γνωρίζουμε ότι $\widehat{F}_n$ είναι «κοντά» στην F (για κάθε F).
- Είναι αυτό επαρκές για να έχουμε $T(\widehat{F}_n)$ κοντά στην $T(F)$;

Κάποιοι προβληματισμοί

- Γνωρίζουμε ότι $\widehat{F}_n$ είναι «κοντά» στην F (για κάθε F).
- Είναι αυτό επαρκές για να έχουμε $T(\widehat{F}_n)$ κοντά στην $T(F)$;
- Όχι απαραίτητα

Κάποιοι προβληματισμοί

- Γνωρίζουμε ότι $\widehat{F}_n$ είναι «κοντά» στην F (για κάθε F).
- Είναι αυτό επαρκές για να έχουμε $T(\widehat{F}_n)$ κοντά στην $T(F)$;
- Όχι απαραίτητα
- Αντιπαράδειγμα:
  - έστω F παραγωγίσιμη με συνάρτηση πυκνότητας $f(x)$.
  - έστω $T(F) = F'(x)|_{x=x_0} = f(x_0)$
  - Η $\widehat{F}_n$ δεν είναι παραγωγίσιμη στα $x_0 = x_1, \ldots, x_n$, ενώ για $x_0 \neq x_1, \ldots, x_n$ έχουμε $\widehat{F}'_n(x_0) = 0$.

Κάποιοι προβληματισμοί

- Γνωρίζουμε ότι $\widehat{F}_n$ είναι «κοντά» στην F (για κάθε F).
- Είναι αυτό επαρκές για να έχουμε $T(\widehat{F}_n)$ κοντά στην $T(F)$;
- Όχι απαραίτητα
- Αντιπαράδειγμα:
  - έστω F παραγωγίσιμη με συνάρτηση πυκνότητας $f(x)$.
  - έστω $T(F) = F'(x)|_{x=x_0} = f(x_0)$
  - Η $\widehat{F}_n$ δεν είναι παραγωγίσιμη στα $x_0 = x_1, \ldots, x_n$, ενώ για $x_0 \neq x_1, \ldots, x_n$ έχουμε $\widehat{F}'_n(x_0) = 0$.
- Η συνέπεια και η ασυμπτωτική κανονικότητα εκτιμητών συναρτησιακών της F απαιτούν επί πλέον ιδιότητες για την $T(F)$ όπως η συνέχεια και η διαφορισιμότητα συναρτησιακών.

Ορισμός

Έστω T ένα συναρτησιακό, F είναι συνάρτηση κατανομής στο $\mathbb{R}$ και δ_x είναι η εκφυλισμένη κατανομή στο x. Τότε η **συνάρτηση επιρροής** (Influence Function) ορίζεται ως,

$$\mathrm{IF}_F(x) = \lim_{\varepsilon \to 0} \frac{T((1-\varepsilon)F + \varepsilon\delta_x) - T(F)}{\varepsilon},$$

αν το όριο υπάρχει για κάθε $x \in \mathbb{R}$.

Ορισμός

Έστω T ένα συναρτησιακό, F είναι συνάρτηση κατανομής στο $\mathbb{R}$ και δ_x είναι η εκφυλισμένη κατανομή στο x. Τότε η **συνάρτηση επιρροής** (Influence Function) ορίζεται ως,

$$\mathrm{IF}_F(x) = \lim_{\varepsilon \to 0} \frac{T((1-\varepsilon)F + \varepsilon\delta_x) - T(F)}{\varepsilon},$$

αν το όριο υπάρχει για κάθε $x \in \mathbb{R}$.

Αν αλλοιώσω ελαφρώς τα δεδομένα σε ένα μόνο σημείο, πόσο αλλάζει ο εκτιμητής μου;

Υπενθυμίζεται ότι $\delta_x(y)$ είναι ένα σημείο μάζας στο $x \in \mathbb{R}$, αν

$$\delta_x(y) = \begin{cases} 1 & , \quad y \geqslant x, \\ 0 & , \quad y < x. \end{cases}$$

Ουσιαστικά, η συνάρτηση επιρροής είναι η παράγωγος του συναρτησιακού $T(F)$ στην κατεύθυνση του δ_x. Δηλαδή, αν θεωρήσουμε $F_\varepsilon = (1-\varepsilon)F + \varepsilon\delta_x$, τότε

$$\mathrm{IF}_F(x) = \frac{d}{d\varepsilon} T(F_\varepsilon)\bigg|_{\varepsilon=0}, \tag{2}$$

υπό την προϋπόθεση ότι υπάρχει η παράγωγος.

Influence Function: Μέσος vs Διάμεσος

Μέσος

Για $T(F) = \mathbb{E}[X]$,

$$\mathrm{IF}(x)_F = x - \mathbb{E}[X]$$

- Μη φραγμένο influence function
- Ένα outlier μπορεί να επηρεάσει αυθαίρετα τον εκτιμητή

Διάμεσος

Για διάμεσο m,

$$\mathrm{IF}(x)_F = \frac{1}{2f(m)}\,\mathrm{sign}(x - m)$$

- Φραγμένο influence function
- Ανθεκτική σε ακραίες τιμές

Ορισμός

Έστω $X_1, X_2, \ldots, X_n$, είναι ένα τυχαίο δείγμα από έναν πληθυσμό με α.σ.κ. $F(\cdot)$. Η **εμπειρική συνάρτηση επιρροής** συμβολίζεται με $\widehat{\mathrm{IF}}(x)$ και ορίζεται ως εξής: $\widehat{\mathrm{IF}}(x) = \mathrm{IF}_{F_n}(x)$, δηλαδή,

$$\widehat{\mathrm{IF}}(x) = \lim_{\varepsilon \to 0} \frac{T((1-\varepsilon)F_n + \varepsilon\delta_x) - T(F_n)}{\varepsilon}.$$

Παράδειγμα

Μέσος

Για $T(F) = \mathbb{E}[X]$,

$$\widehat{\mathrm{IF}}(x) = x - \bar{x}.$$

- Μη φραγμένη ως συνάρτηση του x

Διάμεσος

Έστω $\hat{m}$ η εμπειρική διάμεσος και $\hat{f}(\hat{m})$ εκτίμηση της πυκνότητας στο $\hat{m}$.

$$\widehat{\mathrm{IF}}(x) = \frac{1}{2\,\hat{f}(\hat{m})}\,\mathrm{sign}(x - \hat{m}).$$

- Φραγμένη ως συνάρτηση του x

Αξιολόγηση της Empirical IF στις Παρατηρήσεις

Η εμπειρική συνάρτηση επιρροής είναι συνάρτηση του x.

Στην πράξη μας ενδιαφέρει η επίδραση κάθε παρατήρησης:

$$\widehat{\mathrm{IF}}(x_1), \widehat{\mathrm{IF}}(x_2), \ldots, \widehat{\mathrm{IF}}(x_n)$$

Μέσος

$$\widehat{\mathrm{IF}}(x_i) = x_i - \bar{x}, \quad i = 1, \ldots, n$$

Διάμεσος

$$\widehat{\mathrm{IF}}(x_i) = \frac{1}{2\,\hat{f}(\hat{m})} \operatorname{sign}(x_i - \hat{m}), \quad i = 1, \ldots, n$$

Σημείωση: Αυτά τα $\widehat{\mathrm{IF}}(x_i)$ χρησιμοποιούνται για τον υπολογισμό της ασυμπτωτικής διασποράς (μεταξύ άλλων).

Θεώρημα

Έστω $T(F) = \int a(x)dF(x)$ *είναι ένα γραμμικό συναρτησιακό, τότε*

1. $\mathrm{IF}_F(x) = a(x) - T(F)$ *και* $\widehat{\mathrm{IF}}(x) = a(x) - T(F_n)$.
2. *Για οποιαδήποτε συνάρτηση κατανομής* G,
$$T(G) = T(F) + \int \mathrm{IF}_F(x)dG(x). \tag{3}$$
3. $\int \mathrm{IF}_F(x)dF(x) = 0$.
4. *Έστω* $\tau^2 = \int \mathrm{IF}_F^2(x)dF(x)$. *Τότε,* $\tau^2 = \int (a(x) - T(F))^2 dF(x)$ *και, αν* $\tau^2 < \infty$,
$$\sqrt{n}(T(F) - T(F_n)) \to \mathcal{N}(0, \tau^2). \tag{4}$$
5. *Έστω* $\widehat{\tau}^2 = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mathrm{IF}}^2(X_i) = \frac{1}{n}\sum_{i=1}^{n}(a(X_i) - T(F_n))^2$. *Τότε,* $\widehat{\tau}^2 \xrightarrow{P} \tau^2$ *και* $\widehat{\mathrm{se}}/\mathrm{se} \xrightarrow{P} 1$, *όπου* $\widehat{\mathrm{se}} = \widehat{\tau}/\sqrt{n}$ *και* $\mathrm{se} = \sqrt{\mathrm{Var}(T(F_n))}$.
6. $$\frac{\sqrt{n}(T(F) - T(F_n))}{\widehat{\tau}} \to \mathcal{N}(0, 1). \tag{5}$$

Παρατηρήσεις

- Η συνάρτηση επιρροής συμπεριφέρεται όπως η συνάρτηση σκορ στην παραμετρική στατιστική συμπερασματολογία:
$$\int \mathrm{IF}_F(x)dF(x) = 0 \text{ και } \mathrm{Var}(T(F_n)) = \int \mathrm{IF}_F^2(x)dF(x)/n.$$
- Από τη σχέση (5) του Θεωρήματος προκύπτει ότι η στατιστική συνάρτηση $\dfrac{T(F) - T(F_n)}{\widehat{\mathrm{se}}}$ ακολουθεί (προσεγγιστικά) την κατανομή $\mathcal{N}(0,1)$. Το αποτέλεσμα αυτό είναι γνωστό στη βιβλιογραφία ως **μη παραμετρική μέθοδος Δέλτα**.
- Από την κανονική προσέγγιση προκύπτει το ασυμπτωτικό διάστημα εμπιστοσύνης για το γραμμικό συναρτησιακό $T(F)$.

Διάστημα Εμπιστοσύνης με τη Μη Παραμετρική Μέθοδο Δέλτα

Θεώρημα

Έστω $X_1, X_2, \ldots, X_n$ *είναι ένα τυχαίο δείγμα από έναν πληθυσμό με αθροιστική συνάρτηση κατανομής* $F(\cdot)$ *και* $T(F)$ *ένα γραμμικό συναρτησιακό. Υποθέτουμε ότι:*

$$\tau^2 = \int (a(x) - T(F))^2 dF(x) < \infty.$$

Τότε ένα $100(1-a)\%$ *ασυμπτωτικό διάστημα εμπιστοσύνης για το γραμμικό συναρτησιακό* $T(F)$ *είναι το:*

$$\left(T(F_n) - z_{a/2} \frac{\widehat{\tau}}{\sqrt{n}}, T(F_n) + z_{a/2} \frac{\widehat{\tau}}{\sqrt{n}} \right) \tag{6}$$

ή ισοδύναμα το

$$\left(T(F_n) - z_{a/2} \widehat{\mathrm{se}}, T(F_n) + z_{a/2} \widehat{\mathrm{se}} \right)$$

όπου $\widehat{\mathrm{se}} = \dfrac{\widehat{\tau}}{\sqrt{n}}$, *με* $\widehat{\tau}^2 = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} (a(X_i) - T(F_n))^2.$

Ασυμπτωτικό ΔΕ για τη Μέση Τιμή με τη Μη Παραμετρική Μέθοδο Δέλτα

Παράδειγμα

Έστω $X_1, X_2, \ldots, X_n$ είναι ένα τυχαίο δείγμα από έναν πληθυσμό με αθροιστική συνάρτηση κατανομής $F(\cdot)$. Έστω ότι

$$\tau^2 = \int (a(x) - T(F))^2 dF(x) = \int (x-\theta)^2 dF(x) < \infty.$$

Να δείξετε ότι το τυχαίο διάστημα

$$(\bar{X} - z_{a/2}\widehat{\mathrm{se}}, \bar{X} + z_{a/2}\widehat{\mathrm{se}}),$$

όπου $\widehat{\mathrm{se}} = \dfrac{\widehat{\tau}}{\sqrt{n}}$ είναι ένα $100(1-a)\%$, κατά σημείο, ασυμπτωτικό διάστημα εμπιστοσύνης για την πληθυσμιακή μέση τιμή θ.

Ασυμπτωτικό ΔΕ για την $F(\beta)$ με τη Μη Παραμετρική Μέθοδο Δέλτα

Παράδειγμα

Έστω $X_1, X_2, \ldots, X_n$ είναι ένα τυχαίο δείγμα από έναν πληθυσμό με αθροιστική συνάρτηση κατανομής $F(\cdot)$. Έστω ότι

$$\hat{\tau}^2 = \frac{1}{n}\sum_{i=1}^{n}\{\delta_{X_i}(\beta) - \hat{F}_n(\beta)\}^2$$

όπου $\hat{F}_n(\cdot)$ η εμπειρική συνάρτηση κατανομής. Να δείξετε ότι το τυχαίο διάστημα

$$(\hat{F}_n(\beta) - z_{a/2}\widehat{\mathrm{se}}, \hat{F}_n(\beta) + z_{a/2}\widehat{\mathrm{se}}),$$

όπου $\widehat{\mathrm{se}} = \widehat{\tau}/\sqrt{n}$ είναι ένα $100(1-a)\%$ ασυμπτωτικό διάστημα εμπιστοσύνης για την $F(\beta)$.

Να δείξετε ότι το παραπάνω συμπίπτει με το Ασυμπτωτικό ΔΕ Wald (σελίδα 17)