Το επαναληπτικό

Μη Παραμετρική Στατιστική

Παναγιώτης Παπασταμούλης
Αναπληρωτής Καθηγητής
Τμήμα Στατιστικής ΟΠΑ

papastamoulis@aueb.gr

2026

Περιεχόμενα

Άσκηση

Έστω $X_1, \ldots, X_n$ ανεξάρτητες και ισόνομες τ.μ. από κατανομή F. Να δειχτεί ότι

$$\mathrm{Cov}(\hat{F}_n(x), \hat{F}_n(y)) = \frac{1}{n}\left[F(\min(x, y)) - F(x)F(y)\right]$$

για δύο διακριτά σημεία x και y ($x \neq y$).

Πρόβλημα

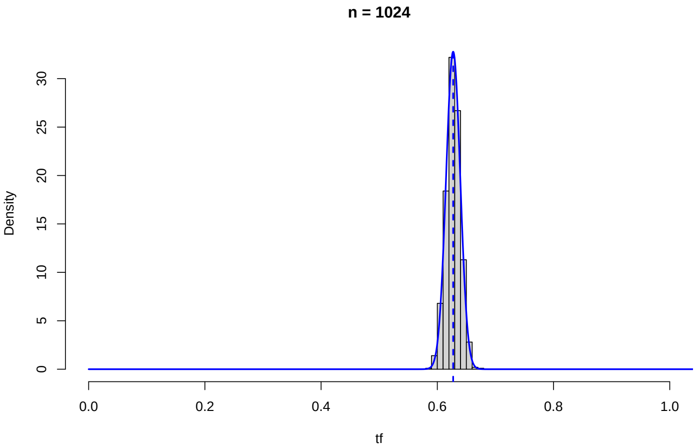
- Θεωρήστε τυχαίο δείγμα μεγέθους n απο εκθετική κατανομή $\mathcal{E}(1)$
- Ενδιαφερόμαστε για την εκτίμηση του συναρτησιακού $\sqrt{F(0.5)}$, όπου $F(x)$ η αθροιστική συνάρτηση κατανομής[1]
- Θεωρούμε τον εκτιμητή

$$\sqrt{\hat{F}_n(0.5)}$$

- Σκοπός: εύρεση ασυμπτωτικής κατανομής

[1] χωρίς να έχουμε γνώση της F

Μελέτη προσομοίωσης



Πρόβλημα

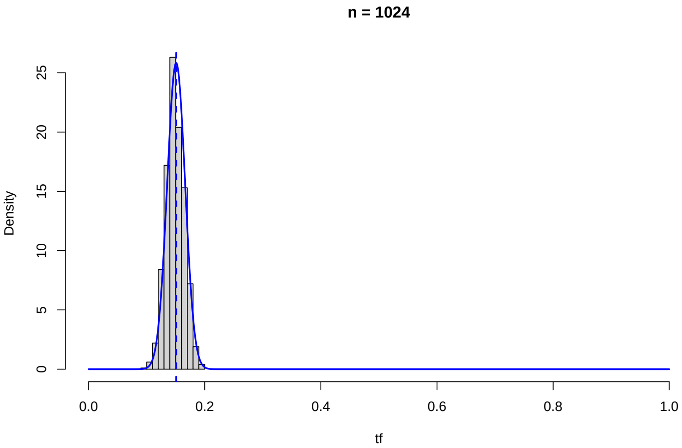
- Θεωρήστε τυχαίο δείγμα μεγέθους n απο κανονική κατανομή $\mathcal{N}(0,1)$
- Ενδιαφερόμαστε για την εκτίμηση του συναρτησιακού $\sqrt{F(-2)}$, όπου $F(x)$ η αθροιστική συνάρτηση κατανομής[2]
- Θεωρούμε τον εκτιμητή

$$\sqrt{\hat{F}_n(-2)}$$

- Σκοπός: εύρεση ασυμπτωτικής κατανομής

[2] χωρίς να έχουμε γνώση της F

Μελέτη προσομοίωσης

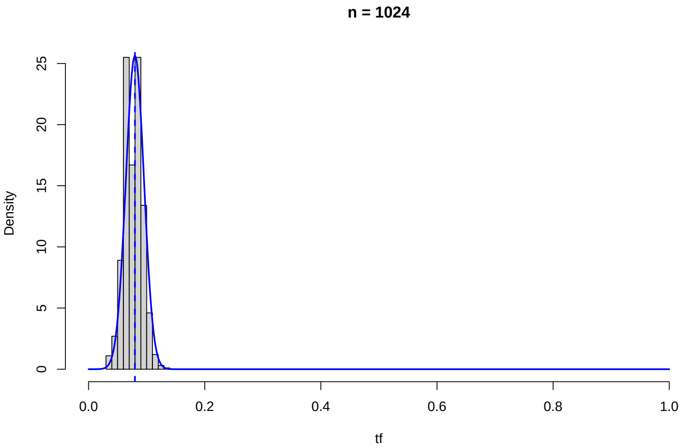


Πρόβλημα

- Θεωρήστε τυχαίο δείγμα μεγέθους n απο κανονική κατανομή t_1
- Ενδιαφερόμαστε για την εκτίμηση του συναρτησιακού $\sqrt{F(-50)}$, όπου $F(x)$ η αθροιστική συνάρτηση κατανομής[3]
- Θεωρούμε τον εκτιμητή
$$\sqrt{\hat{F}_n(-50)}$$
- Σκοπός: εύρεση ασυμπτωτικής κατανομής

[3] χωρίς να έχουμε γνώση της F

Μελέτη προσομοίωσης



Συνάρτηση επιρροής

Έστω $X_1, \ldots, X_n$ τυχαίο δείγμα n παρατηρήσεων από κατανομή F και $\hat{F}_n(x)$ η εμπειρική συνάρτηση κατανομής. Για δεδομένο x, να δείξετε ότι η ασυμπτωτική κατανομή του συναρτησιακού $\sqrt{\hat{F}_n(x)}$ είναι

$$\sqrt{\hat{F}_n(x)} \to \mathcal{N}\left(\sqrt{F(x)}, \frac{1-F(x)}{4n}\right)$$

Εισαγωγή

Ο έλεγχος του Kolmogorov (1933) για ένα δείγμα επεκτάθηκε από τον Smirnov (1939) για δύο δείγματα.

Στόχος: Σύγκριση δύο κατανομών βάσει ανεξάρτητων δειγμάτων.

- Δείγμα $X_1, \ldots, X_n \sim F_X$
- Δείγμα $Y_1, \ldots, Y_m \sim G_Y$
- Τα δείγματα είναι ανεξάρτητα

Υπόθεση Ελέγχου

Θέλουμε να ελέγξουμε:

$$H_0 : F_X(x) = G_Y(x), \quad \forall x \in \mathbb{R}$$

$$H_1 : F_X(x) \neq G_Y(x), \quad \text{για κάποιο } x \in \mathbb{R}$$

Ιδέα: Σύγκριση των εμπειρικών συναρτήσεων κατανομής.

Εμπειρικές Συναρτήσεις

Ορίζουμε:

- $\hat{F}_n(x)$: εμπειρική CDF του δείγματος X
- $\hat{G}_m(x)$: εμπειρική CDF του δείγματος Y

Η στατιστική συνάρτηση βασίζεται στη μέγιστη απόσταση μεταξύ τους.

Στατιστική Kolmogorov–Smirnov

Η στατιστική συνάρτηση ορίζεται ως:

$$D_{n,m} = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - \hat{G}_m(x)|$$

Ισοδύναμα:

$$D_{n,m} = \max_{1 \leq i \leq n+m} |\hat{F}_n(Z_i) - \hat{G}_m(Z_i)|$$

όπου $Z_1, \ldots, Z_{n+m}$ είναι το διατεταγμένο ενωμένο δείγμα.

Ανάλυση της Στατιστικής

Η στατιστική γράφεται ως:

$$D_{n,m} = \max\{D^+_{n,m}, D^-_{n,m}\}$$

όπου:

$$D^+_{n,m} = \sup_x \big(F_n(x) - G_m(x)\big)$$

$$D^-_{n,m} = \sup_x \big(G_m(x) - F_n(x)\big)$$

Θεώρημα

Έστω $X_1, X_2, \ldots, X_n$ ένα τυχαίο δείγμα από έναν πληθυσμό με άγνωστη αθροιστική συνάρτηση κατανομής $F_X(\cdot)$ και $Y_1, Y_2, \ldots, Y_n$ ένα τυχαίο δείγμα από έναν πληθυσμό με άγνωστη αθροιστική συνάρτηση κατανομής $G_Y(\cdot)$. Υποθέτουμε ότι τα δύο δείγματα είναι ανεξάρτητα. Έστω, επιπρόσθετα, $F_n(\cdot)$ και $G_n(\cdot)$ οι αντίστοιχες εμπειρικές αθροιστικές συναρτήσεις κατανομής αυτών των πληθυσμών. Η κατανομή, υπό τη μηδενική υπόθεση $H_0 : F_X(x) = G_Y(x), \forall x \in \mathbb{R}$, των στατιστικών συναρτήσεων $D^+_{n,n} = \sup\limits_x [F_n(x) - G_n(x)]$ και $D^-_{n,n} = \sup\limits_{x \in \mathbb{R}} (G_n(x) - F_n(x))$ δίνεται από τη σχέση:

$$P\left(D^+_{n,n} > d\right) = P\left(D^-_{n,n} > d\right) = \frac{\binom{2n}{[n(d+1)]}}{\binom{2n}{n}},$$

όπου με $[n(d+1)]$ συμβολίζουμε τον μεγαλύτερο ακέραιο που είναι μικρότερος ή ίσος από $n(d+1)$.

Θεώρημα

Έστω ότι $X_1, X_2, \ldots, X_n$*, ένα τυχαίο δείγμα από έναν πληθυσμό με άγνωστη αθροιστική συνάρτηση κατανομής* $F_X(\cdot)$ *και* $Y_1, Y_2, \ldots, Y_m$*, ένα τυχαίο δείγμα από έναν πληθυσμό με άγνωστη αθροιστική συνάρτηση κατανομής* $G_Y(\cdot)$*. Υποθέτουμε ότι τα δύο δείγματα είναι ανεξάρτητα. Έστω, επιπρόσθετα,* $F_n(\cdot)$ *και* $G_m(\cdot)$ *οι αντίστοιχες εμπειρικές αθροιστικές συναρτήσεις κατανομής αυτών. Όταν* n *και* m *είναι τέτοια, ώστε* $m, n \to \infty$*,* $\frac{m}{n} \to q$*, με* q *σταθερό θετικό αριθμό, τότε, υπό τη μηδενική υπόθεση* $H_0 : F_X(x) = G_Y(x), \forall x \in \mathbb{R}$*, ισχύει ότι:*

$$P\left(\sqrt{\frac{mn}{m+n}} D_{n,m} \leqslant d\right) \approx 1 - 2\sum_{j=1}^{\infty} e^{-2j^2 d^2}.$$

Άσκηση 1

Έστω δύο ανεξάρτητα τυχαία δείγματα

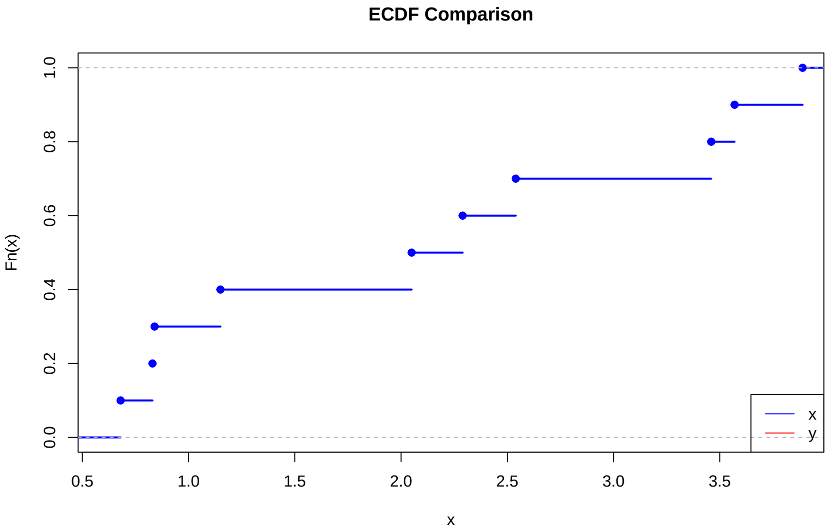
$$\boldsymbol{x} = (0.83, 3.57, 3.46, 2.05, 3.89, 2.54, 2.29, 1.15, 0.68, 0.84)$$

και

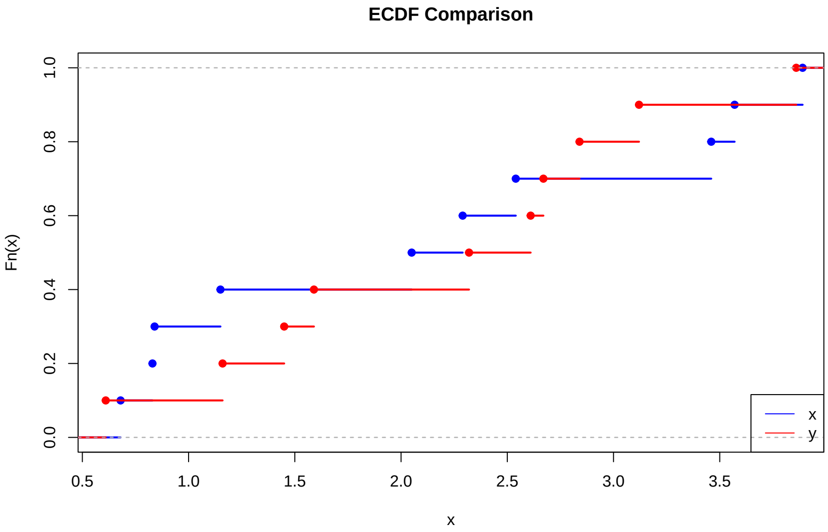
$$\boldsymbol{y} = (1.16, 1.45, 0.61, 2.67, 2.32, 2.84, 2.61, 1.59, 3.86, 3.12)$$

Να ελεγχθεί σε επίπεδο σημαντικότητας $\alpha = 5\%$ η υπόθεση ισότητας των κατανομών.

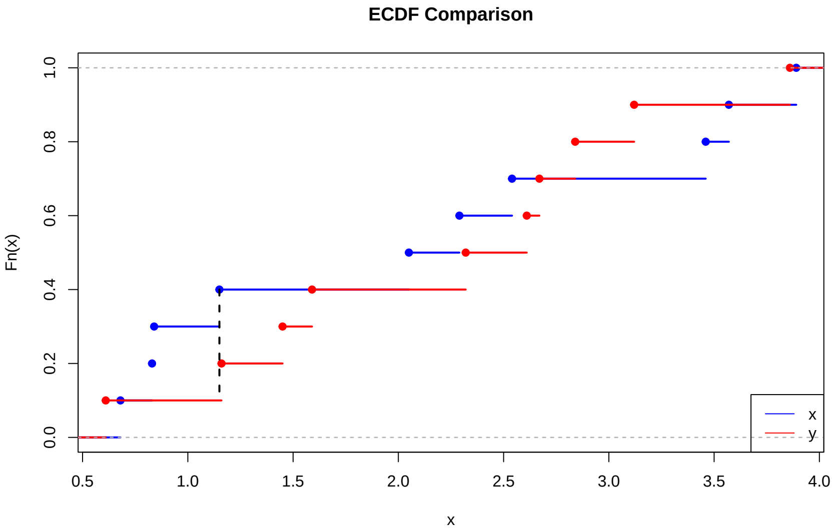
Εμπειρικές συναρτήσεις κατανομής και στατιστική συνάρτηση Kolmogorov-Smirnov



Εμπειρικές συναρτήσεις κατανομής και στατιστική συνάρτηση Kolmogorov-Smirnov



Εμπειρικές συναρτήσεις κατανομής και στατιστική συνάρτηση Kolmogorov-Smirnov



Αναλυτικά: $D_{n,m} = 0.3$

i	z_i	$\hat{F}_n(z_i)$	$\hat{G}_m(z_i)$	$\lvert\hat{F}_n(z_i) - \hat{G}_m(z_i)\rvert$
1	0.61	0.00	0.10	0.10
2	0.68	0.10	0.10	0.00
3	0.83	0.20	0.10	0.10
4	0.84	0.30	0.10	0.20
5	1.15	0.40	0.10	**0.30**
6	1.16	0.40	0.20	0.20
7	1.45	0.40	0.30	0.10
8	1.59	0.40	0.40	0.00
9	2.05	0.50	0.40	0.10
10	2.29	0.60	0.40	0.20
11	2.32	0.60	0.50	0.10
12	2.54	0.70	0.50	0.20
13	2.61	0.70	0.60	0.10
14	2.67	0.70	0.70	0.00
15	2.84	0.70	0.80	0.10
16	3.12	0.70	0.90	0.20
17	3.46	0.80	0.90	0.10
18	3.57	0.90	0.90	0.00
19	3.86	0.90	1.00	0.10
20	3.89	1.00	1.00	0.00

Με χρήση R

```
ks.test(x,y)

 Exact two-sample Kolmogorov-Smirnov test

data:  x and y
D = 0.3, p-value = 0.7869
alternative hypothesis: two-sided
```

Με χρήση R

```
ks.test(x,y)

 Exact two-sample Kolmogorov-Smirnov test

data:  x and y
D = 0.3, p-value = 0.7869
alternative hypothesis: two-sided
```

Δεν απορρίπτουμε την υπόθεση ισότητας των κατανομών.

Άσκηση 2

Έστω δύο ανεξάρτητα τυχαία δείγματα

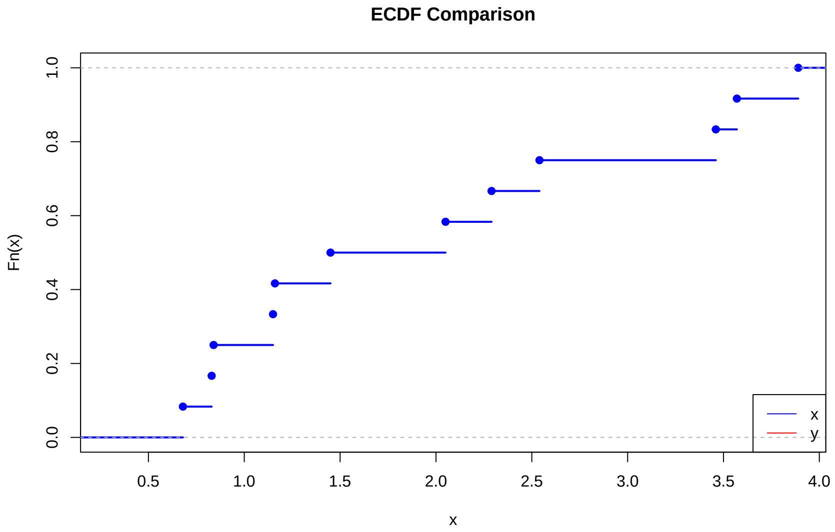
$$\boldsymbol{x} = (0.83, 3.57, 3.46, 2.05, 3.89, 2.54, 2.29, 1.15, 0.68, 0.84, 1.16, 1.45)$$

και

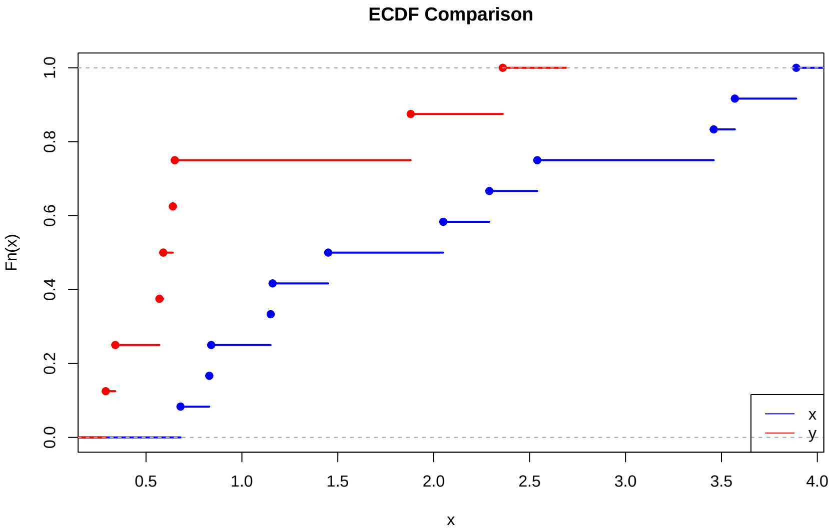
$$\boldsymbol{y} = (1.88, 0.65, 0.34, 0.59, 2.36, 0.64, 0.29, 0.57)$$

Να ελεγχθεί σε επίπεδο σημαντικότητας $\alpha = 5\%$ η υπόθεση ισότητας των κατανομών.

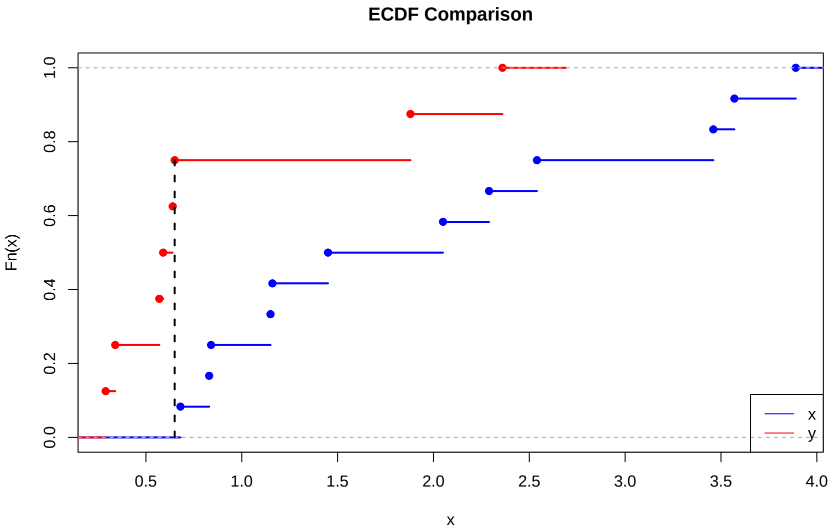
Εμπειρικές συναρτήσεις κατανομής και στατιστική συνάρτηση Kolmogorov-Smirnov



Εμπειρικές συναρτήσεις κατανομής και στατιστική συνάρτηση Kolmogorov-Smirnov



Εμπειρικές συναρτήσεις κατανομής και στατιστική συνάρτηση Kolmogorov-Smirnov



Αναλυτικά: $D_{n,m} = 0.75$

i	z_i	$\hat{F}_n(z_i)$	$\hat{G}_m(z_i)$	$\lvert\hat{F}_n(z_i) - \hat{G}_m(z_i)\rvert$
1	0.29	0.00	0.12	0.12
2	0.34	0.00	0.25	0.25
3	0.57	0.00	0.38	0.38
4	0.59	0.00	0.50	0.50
5	0.64	0.00	0.62	0.62
6	0.65	0.00	0.75	**0.75**
7	0.68	0.08	0.75	0.67
8	0.83	0.17	0.75	0.58
9	0.84	0.25	0.75	0.50
10	1.15	0.33	0.75	0.42
11	1.16	0.42	0.75	0.33
12	1.45	0.50	0.75	0.25
13	1.88	0.50	0.88	0.38
14	2.05	0.58	0.88	0.29
15	2.29	0.67	0.88	0.21
16	2.36	0.67	1.00	0.33
17	2.54	0.75	1.00	0.25
18	3.46	0.83	1.00	0.17
19	3.57	0.92	1.00	0.08
20	3.89	1.00	1.00	0.00

Με χρήση R

```
ks.test(x,y)

 Exact two-sample Kolmogorov-Smirnov test

data:  x and y
D = 0.75, p-value = 0.004763
alternative hypothesis: two-sided
```

Με χρήση R

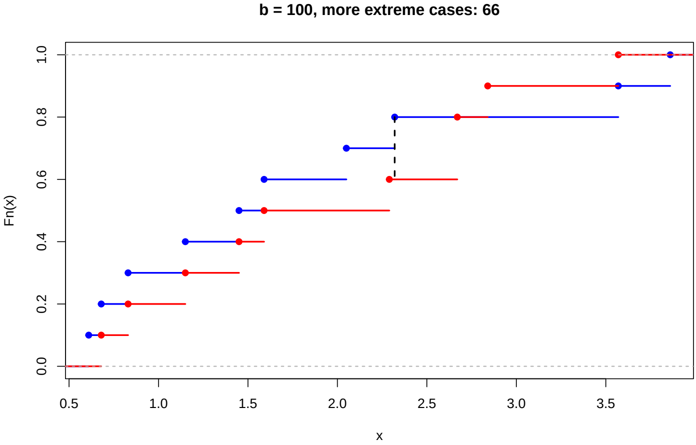
```
ks.test(x,y)

 Exact two-sample Kolmogorov-Smirnov test

data:  x and y
D = 0.75, p-value = 0.004763
alternative hypothesis: two-sided
```

Απορρίπτουμε την υπόθεση ισότητας των κατανομών.

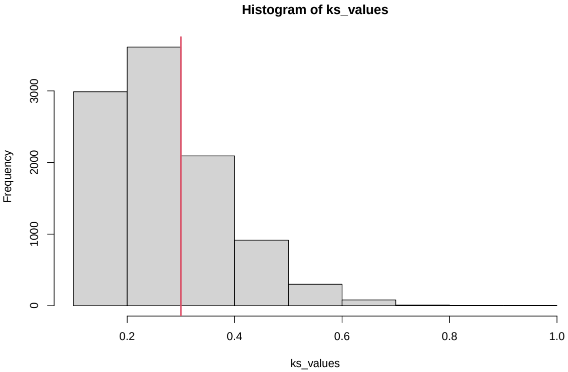
Bootstrap version

- Πώς θα μπορούσαμε να απαντήσουμε το παραπάνω μέ χρήση bootstrap;

Άσκηση 1: Bootstrap δείγματα (υπό την H_0)

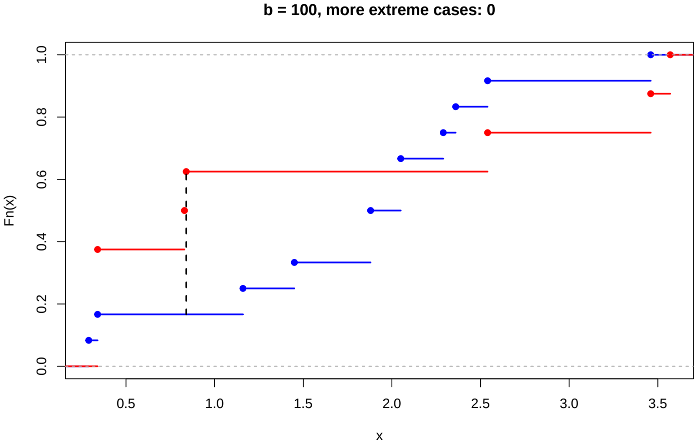


Bootstrap στα δεδομένα της Άσκησης 1

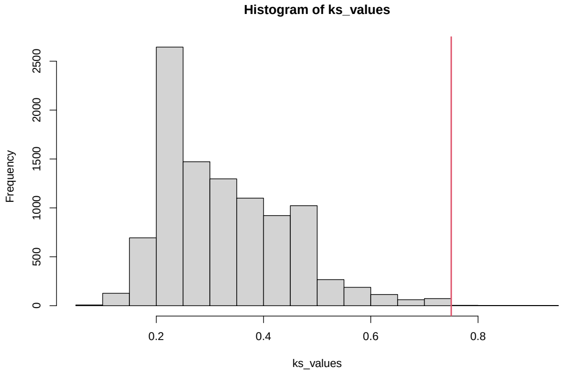


Εκτιμηθέν p-value $= 0.7012$

Άσκηση 2: Bootstrap δείγματα (υπό την H_0)



Bootstrap στα δεδομένα της Άσκησης 2



Εκτιμηθέν p-value $= 0.0037$