Μη-Παραμετρική Στατιστική

Φροντιστήριο #3

Εισαγωγή στο γενικό πρόβλημα για δύο δείγματα

Θα ασχοληθούμε με δεδομένα από δύο ανεξάρτητα τυχαία δείγματα τα οποία μπορούν να σχηματιστούν ανεξάρτητα από καθέναν από δύο πληθυσμούς. Τα δείγματα είναι ανεξάρτητα υπό την έννοια ότι κάθε μέτρηση στο ένα δείγμα είναι ανεξάρτητη κάθε μέτρησης στο δεύτερο δείγμα. Θεωρούμε δύο πληθυσμούς X και Y με συναρτήσεις κατανομών, αντίστοιχα, F_X και F_Y. Έστω ένα δείγμα μεγέθους m από τον πληθυσμό X και ένα άλλο δείγμα μεγέθους n που σχηματίζεται ανεξάρτητα από τον πληθυσμό Y.

Έχουμε $X_1, X_2, \dots, X_m$

$Y_1, Y_2, \dots, Y_n$

Συνήθως, η υπόθεση που μας ενδιαφέρει να εξετάσουμε είναι: $H_0: F_Y(x) = F_X(x)$ $\forall x$ δηλαδή αν τα δύο δείγματα σχηματίζονται από τον ίδιο πληθυσμό.

Αν για παράδειγμα, θεωρήσουμε κάποια παραμετρική υπόθεση όπως π.χ. ότι οι πληθυσμοί **κατανέμονται** κανονικά, τότε το t-test για την ισότητα των μέσων τιμών ή το F-test για την ισότητα των διακυμάνσεων

είναι οι πιο ισχυροί έλεγχοι. Όμως αυτά και άλλα μαθηματικά tests μπορεί να είναι αρκετά ευαίσθητα σε αποκλίσεις από την υπόθεση της κανονικότητας ή από άλλες σημαντικές υποθέσεις. Αν υπάρχουν υποψίες για πιθανές αποκλίσεις ή μη-επαρκής πληροφορία για να ελεγχθούν οι προϋποθέσεις της κανονικότητας μπορεί να υιοθετηθεί μια μη-παραμετρική διαδικασία ελέγχου.

Πρακτικά, χρησιμοποιούνται και υιοθετώνται κάποιες άλλες υποθέσεις για τη μορφή των υπό-εξέταση πληθυσμών. Μία πρώτη τέτοια υπόθεση μας είναι το μοντέλο μετατόπισης (shift model or location model). Κάτω από μία τέτοια υπόθεση, θεωρούμε ότι οι πληθυσμοί X και Y είναι ίδιοι, εκτός πιθανώς από μία μετατόπιση κατά μία ποσότητα θ, πιθανώς άγνωστη, δηλ. $F_Y(x) = P(Y \le x) = P(X \le x-\theta)$
$= F_X(x-\theta)$ $\forall x$ και $\theta \neq 0$. Αυτό σημαίνει ότι $X+\theta$ και Y έχουν την ίδια κατανομή ή ότι η X κατανέμεται όπως η $Y-\theta$. Ο πληθυσμός Y είναι ο ίδιος με τον X αν $\theta = 0$, "μετατοπίζεται" προς τα δεξιά αν $\theta > 0$ και "μετατοπίζεται" προς τα αριστερά αν $\theta < 0$. Κάτω από αυτήν την υπόθεση, οι πληθυσμοί έχουν το ίδιο σχήμα, την ίδια διασπορά και το ποσό της μετατόπισης θ θα πρέπει να είναι ίσο με τη

διαφορά των μέσων τιμών $\mu_Y - \mu_X$ (ή τη διαφορά των διαμέσων $M_Y - M_X$).

Άλλη υπόθεση που μπορεί να τεθεί για τη μορφή του πληθυσμού είναι το μοντέλο κλίμακας (scale model). Σύμφωνα με την υπόθεση αυτή, οι πληθυσμοί X και Y είναι ίδιοι εκτός (πιθανώς) από μία ποσότητα θ, $\theta > 0$, $\theta \neq 1$. Μπορούμε, τότε, να γράψουμε:

$$F_Y(x) = P(Y \leq x) = P(X \leq \theta) = F_X(\theta x) \quad \forall x$$ και $\theta > 0$, $\theta \neq 1$. Αυτό σημαίνει ότι οι X/θ και Y έχουν την ίδια κατανομή $\forall \theta > 0$ ή ότι η X κατανέμεται ως θY. Επίσης $\text{var}(X) = \theta^2 \text{var}(Y)$ και $E(X) = \theta E(Y)$.

Η υπόθεση του σύνθετου μοντέλου μεταπόπισης-κλίμακας (location-scale model) ενσωματώνει ιδιότητες και από τα δύο μοντέλα.

Ο έλεγχος Kolmogorov-Smirnov για δύο δείγματα

Στην περίπτωση των δύο δειγμάτων, χρησιμοποιούμε ένα στατιστικό έλεγχου το οποίο για τα δύο δείγματα συγκρίνει τις εμπειρικές συναρτήσεις κατανομών τους.

Έστω οι διατεταγμένες στατιστικές μεγέθους m και η αντίστοιχα, από δύο πληθυσμούς με συνεχείς συναρτήσεις κατανομών F_X και F_Y που δίνονται όπως παρακάτω: $X_{(1)}, X_{(2)}, \ldots, X_{(m)}$ και

$Y_{(1)}, Y_{(2)}, \dots, Y_{(n)}$. Οι αντίστοιχες εμπειρικές συναρτήσεις κατανομής $S_m(x)$ και $S_n(x)$ ορίζονται ως εξής:

$$S_m(x) = \begin{cases} 0, & \text{αν } x < X_{(1)} \\ k/m, & \text{αν } X_{(k)} \le x < X_{(k+1)}, \text{ για } k = 1, 2, \dots, m-1 \\ 1, & \text{αν } x \ge X_{(m)} \end{cases}$$

και

$$S_n(x) = \begin{cases} 0, & \text{αν } x < Y_{(1)} \\ k/n, & \text{αν } Y_{(k)} \le x < Y_{(k+1)}, \text{ για } k = 1, 2, \dots, n-1 \\ 1, & \text{αν } x \ge Y_{(n)} \end{cases}$$

Μπορούμε να πούμε ότι $S_m(x)$ και $S_n(x)$ είναι τα ποσοστά των παρατηρήσεων των X και Y που δεν ξεπερνούν την τιμή x.

Αν ισχύει η $H_0: F_Y(x) = F_X(x)$ $\forall x$ τότε οι πληθυσμιακές κατανομές ταυτίζονται και κατ' ουσίαν θεωρούμε ότι έχουμε δύο δείγματα από τον "ίδιο" πληθυσμό. Πολύ λογικά οι εμπειρικές συναρτήσεις κατανομής των X και Y μπορούν να θεωρηθούν οι καταλληλότεροι εκτιμητές για τις πληθυσμιακές συναρτήσεις κατανομών των X και Y.

Χρειαζόμαστε έναν ορισμό εγγύτητας (closeness definition) για να ορίσουμε το κριτήριο Kolmogorov-Smirnov για τα δύο δείγματα. Το κριτήριο συμβολίζεται ως $D_{m,n}$ και είναι η μέγιστη απόλυτη διαφορά μεταξύ των δύο εμπειρικών συναρτήσεων κατανομών:

$$D_{m,n} = \max_x \left| S_m(x) - S_n(x) \right|$$

Η παραπάνω σ.σ. εξέχου είναι κατάλληλη για το

εναλλακτικό δίπλευρο έλεγχο:

$$H_A: F_Y(x) \neq F_X(x) \text{ για κάποιο } x \quad (\text{Έλεγχος } A)$$

και η κρίσιμη περιοχή καθορίζεται από την ανισότητα:

$$D_{m,n} \geq c_\alpha \quad \text{όπου} \quad P(D_{m,n} \geq c_\alpha \mid H_0) \leq \alpha$$

Η p-τιμή είναι $P(D_{m,n} \geq D_0 \mid H_0)$ όπου D_0 είναι η παρατηρούμενη τιμή του $D_{m,n}$ στατιστικού που υπολογίζουμε. Η ακριβής κατανομή του στατιστικού $D_{m,n}$ κάτω από την H_0 δίνεται από κατάλληλους πίνακες (βλέπε π.χ. Πίνακα 21, βιβλίο Ξεκαλάκη).

Για τη μονόπλευρη εναλλακτική υπόθεση:

$$H_1: F_X(x) > F_Y(x) \text{ για τουλάχιστον ένα } x \quad (\text{Έλεγχος } B)$$

χρησιμοποιούμε το στατιστικό:

$$D^+_{m,n} = \max_x \left[S_{m,X}(x) - S_{n,Y}(x) \right]$$

ομοίως για τη μονόπλευρη εναλλακτική υπόθεση:

$$H_1: F_X(x) < F_Y(x) \text{ για τουλάχιστον ένα } x. \quad (\text{Έλεγχος } \Gamma)$$

χρησιμοποιούμε το στατιστικό:

$$D^-_{m,n} = \max_x \left[S_{n,Y}(x) - S_{m,X}(x) \right]$$

Σχόλια Το test K-S για δύο δείγματα είναι εύκολο στην εφαρμογή, χρησιμοποιώντας την ακριβή κατανομή για κάθε τιμή των m και n στο εύρος των τιμών των πινάκων και χρησιμοποιώντας κατάλληλη ασυμπτωτική κατανομή για μεγαλύ-

τερα δειγματικά μεγέθη (βλέπε πίνακα 21, βιβλίο Ξενιαδάκη). Ο πίνακας 20 του βιβλίου της Ξενιαδάκη, χρησιμοποιείται για την περίπτωση n=m.

Ο έλεγχος Β είναι κατάλληλος για προβλήματα όπου έχει ενδιαφέρον να ελέγξουμε αν τα Χ είναι μετατοπισμένα προς τα αριστερά των Υ (δηλαδή αν τα Χ τείνουν να είναι μικρότερα από τα Υ). Αντίστοιχα, ο έλεγχος Γ είναι κατάλληλος είναι κατάλληλος για τα προβλήματα στα οποία επιθυμούμε να ελέγξουμε αν τα Χ είναι μετατοπισμένα προς τα δεξιά των Υ (δηλαδή αν τα Χ τείνουν να είναι μεγαλύτερα από τα Υ).

Ακριβής κατανομή για την ελεγχοσυνάρτηση $D_{m,n}$.

Κάτω από την H_0, η ακριβής κατανομή της ελεγχοσυνάρτησης $D_{m,n}$ έχει υπολογιστεί. Πιο συγκεκριμένα, για μικρά μεγέθη δειγμάτων (π.χ. μικρότερα από 25 ή αν $m+n \leq 16$ ή/και $2 \leq m \leq n \leq 12$) υπάρχουν κατάλληλοι πίνακες τιμών. Το "σώμα" αυτών των πινάκων δίνει τις τιμές $mn D_{m,n}$ που είναι σημαντικές σε διάφορα επίπεδα στατιστικής σημαντικότητας. Γνωρίζοντας τις τιμές των $m, n, mn D_{m,n}$ και αναλόγως με το αν έχουμε μονόπλευρο ή αμφίπλευρο έλεγχο βρίσκουμε τα κρίσιμα σημεία του ελέγχου. Για παράδειγμα, για $m=6, n=8$ απορρίπτουμε την H_0 σε επίπεδο $\alpha=1\%$ όταν $mn D_{m,n} \geq 38$, στον μονόπλευρο έλεγχο.

Μεγάλα δείγματα: Αν είτε m είτε n είναι μεγαλύτερα από 25, τότε π.χ. για τον αμφίπλευρο έλεγχο, πάλι

υπάρχουν καταλληλοι πίνακες. Υπολογίζουμε από τα δεδομένα μας την τιμή του στατιστικού $D_{m,n}$ και συγκρίνουμε την τιμή αυτή με τη κρίσιμη τιμή από τον πίνακα για δοθείσες τιμές των m και n και για συγκεκριμένη τιμή του επιπέδου στατιστικής σημαντικότητας α. Αν $D_{m,n} > C_\alpha \sqrt{\frac{m+n}{mn}}$ τότε η H_0 απορρίπτεται σε επίπεδο α για τον δίπλευρο έλεγχο. Για παράδειγμα, για $m=55, n=60$ και για $\alpha=0.05$ για τον αμφίπλευρο έλεγχο, ο πίνακας δίνει την τιμή με την οποία η $D_{m,n}$ θα πρέπει να είναι ίση ή μεγαλύτερη ώστε να απορριφθεί η H_0. Για αυτά τα δεδομένα, η H_0 απορρίπτεται σε επίπεδο $\alpha=5\%$ αν $D_{m,n} > \underset{(\alpha=5\%)}{1.36} \sqrt{\frac{m+n}{mn}}$

$$= 1.36 \sqrt{\frac{55+60}{(55)(60)}} = 0.254$$

Για τον μονόπλευρο έλεγχο, όταν m και n είναι μεγάλα, τότε $D_{m,n} = \max_x [S_m(x) - S_n(x)]$

Ο Goodman (1954) απέδειξε ότι η ο.σ.

$$\chi^2 = 4 D_{m,n}^2 \frac{mn}{m+n}$$ προσεγγίζεται από την κατανομή χ^2 με $df=2$ αν τα δειγματικά μεγέθη αυξάνουν $(n,m \to \infty)$. Σε μία τέτοια περίπτωση χρησιμοποιούμε καταλλήλους πίνακες της κατανομής χ^2. Η προσέγγιση με την χ^2 μπορεί να χρησιμοποιηθεί και με μικρά δείγματα αλλά αυτό συνήθως οδηγεί σε έναν συντηρητικό έλεγχο. Δηλαδή

το λάθος στη χρήση της X^2-προσέγγισης με μικρά -8-
δείγματα οδηγεί σχεδόν πάντα τον έλεγχο προς
τη σωστή κατεύθυνση (π.χ. σωστή απόρριψη της H_0).
Σε κάθε περίπτωση προτιμάται η ακριβής κατανομή
για το στατιστικό $D_{m,n}$ σε μικρά δείγματα.

Ας δούμε ένα παράδειγμα.

Παράδειγμα Ερευνητής θέλει να ελέγξει αν δύο τάξεις παιδιών συμπεριφέρονται στατιστικά ίδια ως προς μία ψυχολογική μέθοδο που τους εφαρμόζεται. Ο ερευνητής ελέγχει την υπόθεση:

H_0: δεν υπάρχει στατιστικά σημαντική διαφορά στα ποσοστά λαθών που κάνουν οι δύο τάξεις παιδιών, έναντι της

H_1: τα παιδιά της μεγαλύτερης τάξης έχουν μικρότερα (λιγότερα) ποσοστά λαθών από τα παιδιά της μικρότερης τάξης.

Δίνεται ο ακόλουθος πίνακας με τα ποσοστά των συνολικών λαθών στις δύο τάξεις.

Τάξη με παιδιά 11 ετών	Τάξη με παιδιά 7 ετών
35.2	39.1
39.2	41.2
40.9	45.2
38.1	46.2
34.4	48.4
29.1	48.7
41.8	55.0
24.3	40.6
32.4	52.1
	47.2

Επιλέγουμε $\alpha = 0.01, m = 9, n = 10$

Εφαρμόζουμε το one-sided K-S two sample test.

Τα δεδομένα μας ταξινομούνται σε κλάσεις διαστημάτων ώστε να υπολογιστούν οι δύο εμπειρικές συναρτήσεις κατανομών $S_m(x), S_n(x)$.

	$S_m(x)$	$S_n(x)$	$S_m(x) - S_n(x)$
24-27	1/9	0/10	0.111
28-31	2/9	0/10	0.222
32-35	5/9	0/10	0.556
36-39	7/9	1/10	0.678
40-43	9/9	3/10	0.7
44-47	9/9	6/10	0.4
48-51	9/9	8/10	0.2
52-55	9/9	10/10	0

Παρατηρούμε ότι η μέγιστη διαφορά παρατηρείται για τιμή $D_{m,n} = 0.7$. Άρα $mn D_{m,n} = 9 \cdot 10 \cdot 0.7 = 63$.

Για $\alpha = 0.01$ η κρίσιμη τιμή από τον πίνακα είναι: $c_{0.01} = 61$. Επειδή $mn D_{m,n} > c_{0.01}$ υπάρχουν ισχυρές ενδείξεις απόρριψης της H_0.

Συμπεραίνουμε ότι τα παιδιά 11 ετών φαίνεται ότι μάθουν λιγότερα λάθη από τα παιδιά των 7 ετών όταν και στις δύο ομάδες παιδιών εφαρμοστεί η ψυχολογική μέθοδος.

Για την εκτέλεση του ελέγχου Kolmogorov-Smirnov για τα δύο δείγματα χρησιμοποιούμε τη συνάρτηση ks.test, στην R. Χρήσιμα ορίσματα της συνάρτησης είναι τα ακόλουθα:

alternative = { "g" / "l" } για τον καθορισμό του μονόπλευρου ελέγχου

και exact = { TRUE / FALSE } για τον καθορισμό της μεθόδου υπολογισμού της p-τιμής (p-value).

Παρατήρηση: Στον έλεγχο K-S για δύο δείγματα υποθέτουμε ότι οι πληθυσμοί X και Y είναι ίδιοι εκτός ίσως από μία πιθανή διαφοροποίηση στις θέσεις τους που εμφράζεται από μία μετατόπιση θ (δηλ. ότι $X+\theta$ και Y έχουν την ίδια κατανομή ή ότι X και $Y-\theta$ έχουν την ίδια κατανομή). Άρα, εναλλακτικά οι έλεγχοι υποθέσεων:

$H_0: F_X(x) = F_Y(x)$ $\forall x$ έναντι:

$H_1: F_X(x) \neq F_Y(x)$ για ένα τουλάχιστον x

$H_1: F_X(x) \geq F_Y(x)$* ισχύει η γνήσια ανισότητα για ένα τουλαχιστόν x

$H_1: F_X(x) \leq F_Y(x)$

*Αν $F_X(x) \geq F_Y(x)$ με αυστηρή ανισότητα για κάποιο x τότε λέμε ότι η τ.μ. Y είναι στοχαστικά μεγαλύτερη από την τ.μ. X. Συμβολισμός: $\boxed{Y \geq^{st} X.}$

Μπορούν να γραφούν ως εξής: (για το μοντέλο θέσης)

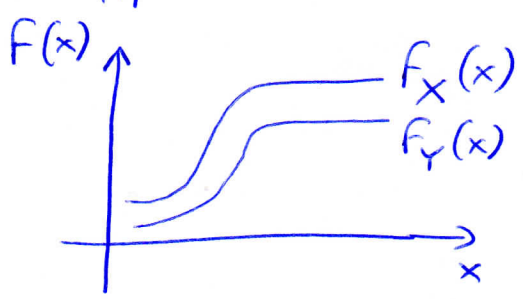
$H_0: \theta = 0$ έναντι:

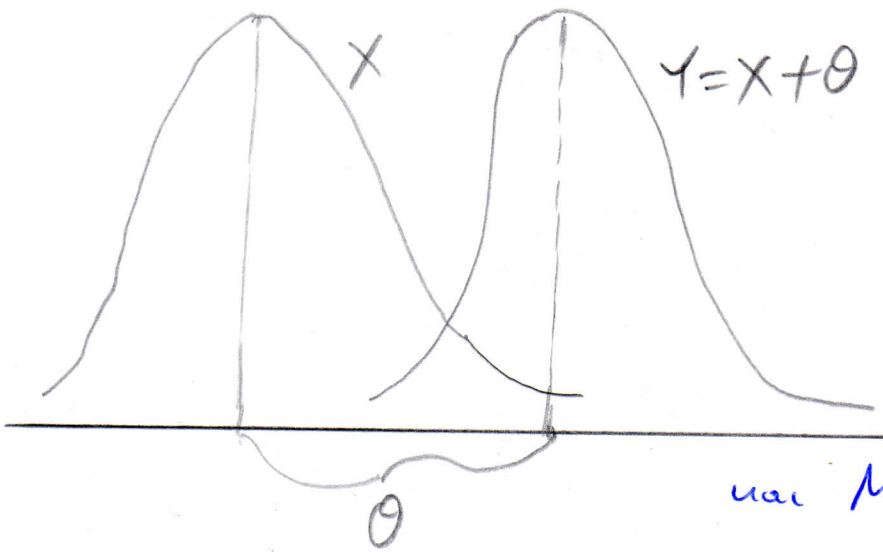
$H_1: \theta \neq 0$

$H_1: \theta > 0$

$H_1: \theta < 0$

Σχήμα για (*)





Οι X και Y έχουν ίδια μορφή ίδια διασπορά ενώ η θ συνήθως είναι: $\theta = \mu_Y - \mu_X$ ή $\theta = M_Y - M_X$ όπου μ_X, μ_Y: μέσες τιμές και M_X, M_Y διάμεσοι X, Y.

Η p-value του ελέγχου είναι ίση με: (για τον αμφίπλευρο έλεγχο)

$\cdot P(D_{n,m} > d_{n,m})$ όπου $d_{n,m}$ είναι η παρατηρούμενη τιμή της $D_{n,m}$

Είδαμε ότι η ακριβής κατανομή της $D_{n,m}$ είναι δύσκολο να υπολογιστεί και έχουν κατασκευαστεί ειδικοί πίνακες με τα ποσοστιαία σημεία της $D_{n,m}$. Μπορεί ωστόσο, να αποδειχθεί ότι:

$$\cdot \lim_{n,m\to\infty} P\left(\sqrt{\frac{nm}{n+m}}\, D_{n,m} \le d\right) = 1 - 2\sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2d^2}$$

άρα μία προσεγγιστική τιμή της p-value του ελέγχου για μεγάλα δείγματα είναι η:

$$p\text{-value} = P(D_{n,m} > d_{n,m}) = 1 - P(D_{n,m} \le d_{n,m})$$

$$\simeq 2\sum_{i=1}^{\infty} (-1)^{i-1} e^{-2\frac{nm}{n+m} i^2 d_{n,m}^2}$$

Στο παράδειγμά μας έχουμε:

$$H_0: F_X(x) = F_Y(x) \quad \forall x$$

vs

$$H_1: F_X(x) \geq F_Y(x) \text{ για ένα τουλάχιστον } x.$$

(F_X: παιδιά μεγαλύτερης τάξης (11 ετών); F_Y: παιδιά μικρότερης τάξης (7 ετών))

$$D^+_{n,m} = \max_x [S_n(x) - S_m(x)]$$

με $D^+_{n,m} = 0.7$

και $p\text{-τιμή} = P(D^+_{n,m} \geq 0.7) = 0.009637$

Για τα συνήθη επίπεδα σ.σ. η H_0 απορρίπτεται άρα υπάρχουν ισχυρές ενδείξεις υπέρ της H_1. Η υλοποίηση του ελέγχου στην R δίνεται στην επόμενη σελίδα μαζί με την υλοποίηση των γραφημάτων για τις εμπειρικές σ.κ. για τα **δύο** δείγματα, των παιδιών μεγαλύτερης τάξης 11 ετών και των παιδιών μικρότερης τάξης 7 ετών.

Για την παρασκευή των γραφημάτων των δύο εμπειρικών σ.κ. για τα δύο δείγματα χρησιμοποιούμε την library(ggplot2).

Θέτουμε cdf1 και cdf2 τις δύο ε.σ.κ. για τα δύο δείγματα.

Ο τίτλος του γραφήματος δίνεται στο όρισμα της εντολής plot, main="ECDF"

Με την εντολή lines σωδιάζουμε τα δύο γραφήματα.

Τα υπόλοιπα γραφήματα καθορίζουν την μορφή και το χρώμα των γραμμών που θα έχουν τα δύο γραφήματα. (col.points, lty).

Με την εντολή legend βάζουμε στη θέση (30, 0.8) την επεξηγηματική "ετικέτα" με τα ονόματα "sample 1" και "sample 2".

```
library(ggplot2)
sample1=c(35.2,39.2,40.9,38.1,34.4,29.1,41.8,24.3,32.4)
sample2=c(39.1,41.2,45.2,46.2,48.4,48.7,55,40.6,52.1,47.2)
# create ECDF of data
cdf1 = ecdf(sample1)
cdf2 = ecdf(sample2)
plot(cdf1, verticals=TRUE, do.points=FALSE, col.points=2,
col.hor=2, col.vert=2, lty=1, main="ECDFS")
lines(cdf2, verticals=TRUE, do.points=FALSE, col.points=3,
col.hor=3, col.vert=3,lty=2)
leg.names=c("sample1","sample2")
legend(30, 0.8, leg.names, lty=1:2, col=2:3)
#KS_test_two_samples
ks.test(sample1,sample2, exact=TRUE)
ks.test(sample1,sample2, alternative="g")
```

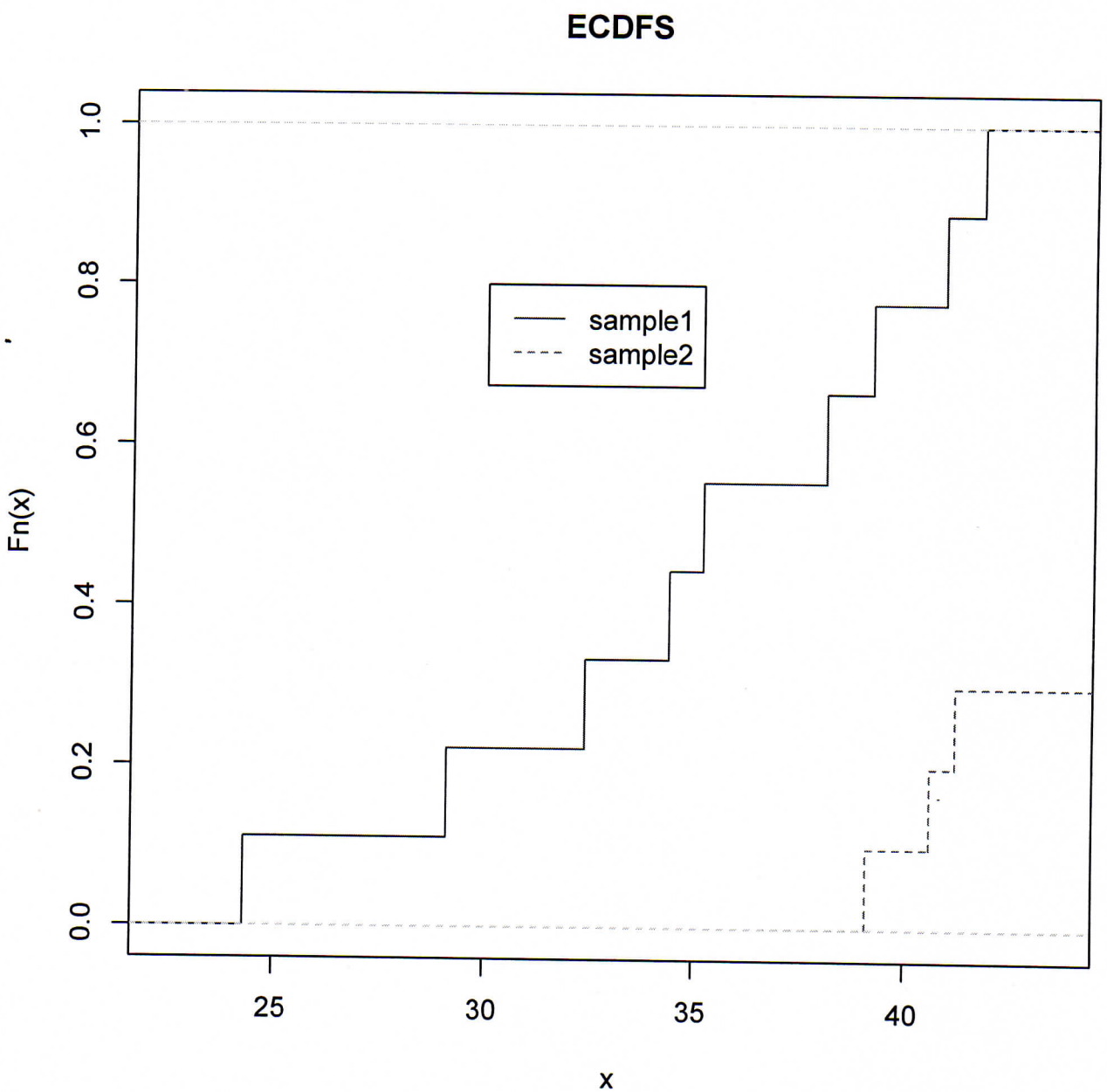
```
 Two-sample Kolmogorov-Smirnov test

data:  sample1 and sample2
D = 0.7, p-value = 0.007036
alternative hypothesis: two-sided

> ks.test(sample1,sample2, alternative="g")

        Two-sample Kolmogorov-Smirnov test

data:  sample1 and sample2
D^+ = 0.7, p-value = 0.009637
alternative hypothesis: the CDF of x lies above that of y
```


ECDFS
sample1
sample2
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
25
30
35
40
x