Μη - Παραμετρική Στατιστική

Φροντιστήριο #5

Μη-παραμετρικοί έλεγχοι υποθέσεων για το πρόβλημα κλίμακας (scale problem) για δύο ανεξάρτητα δείγματα

Εισαγωγή

Θεωρούμε δύο ανεξάρτητα δείγματα από δύο πληθυσμούς. Σύμφωνα με την μηδενική υπόθεση H_0, υποθέτουμε ότι οι δύο πληθυσμοί σχηματίζονται από την ίδια κατανομή.

Μας ενδιαφέρει να ανιχνεύσουμε τυχόν διαφορές των δύο πληθυσμών ως προς τις διακυμάνσεις τους.

Το κλασσικό παραμετρικό test για τον έλεγχο της ισότητας των διακυμάνσεων είναι το F-test. Σε αυτό το test, θεωρούμε ότι οι δύο πληθυσμοί κατανέμονται σύμφωνα με την κανονική κατανομή και η σ.σ. ελέγχου είναι:

$$F_{m-1,n-1} = \frac{\sum_{i=1}^{m}(X_i-\bar{X})^2}{m-1} \Bigg/ \frac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{n-1}$$

η οποία, κάτω από την $H_0: \frac{\sigma_X^2}{\sigma_Y^2}=1$ (όπου σ_X^2, σ_Y^2 είναι οι διακυμάνσεις των πληθυσμών X και Y) ακολουθεί την κατανομή F με $m-1$ και $n-1$ β.ε. Το F-test, γενικά δεν είναι "στιβαρό" (robust) για να αντιμετωπίσει αποκλίσεις από την υπόθεση της κανονικότητας. Συνεπώς, αν υπάρχουν σημαντικές αμφιβολίες για την ισχύ της υπόθεσης της κανονικότητας για τον έλεγχο αυτό, μη-παραμετρικοί ανάλογοι έλεγχοι είναι κατάλληλοι για να εφαρμοστούν.

Το F-test, επιπλέον, δεν απαιτεί κάποια υπόθεση που να αφορά τη θέση των δύο κανονικών πληθυσμών. Τα "μεγέθη" των δύο δειγματικών διακυμάνσεων συγκρίνονται και κάθε μία από τις δειγματικές διακυμάνσεις υπολογίζεται ως ένα μέτρο απόκλισης γύρω από το κάθε δειγματικό μέσο.

Για το F-test η σχέση μεταξύ των δύο (κανονικών) πληθυσμών μπορεί να γραφεί:

$$F_{Y-\mu_Y}(x) = F_{X-\mu_X}\left(\frac{\sigma_X}{\sigma_Y}x\right) = F_{X-\mu_X}(\theta x) \ \forall x \quad (1)$$

και για κάποιο $\theta > 0$, όπου $\theta = \frac{\sigma_X}{\sigma_Y}$ και

$$F_{\frac{(X-\mu_X)}{\sigma_X}}(x) = \Phi(x).$$

Έτσι, η μηδενική υπόθεση H_0 που καλούμαστε να ελέγξουμε είναι: $H_0: \theta = 1$. Δηλαδή, μπορούμε να πούμε ότι κάτω από την H_0, οι κατανομές των τ.μ. $X-\mu_X$ και $Y-\mu_Y$ διαφέρουν μόνο κατά έναν παράγοντα κλίμακας θ (scale factor θ) για οποιεσδήποτε τιμές των μ_X και μ_Y. Συνεπώς,

$$E(X-\mu_X) = \theta E(Y-\mu_Y) \text{ και}$$
$$var(X) = \theta^2 var(Y)$$

Αν θεωρήσουμε ως μέτρα θέσης τις διαμέσους των δύο πληθυσμών X και Y, M_X και M_Y, αντίστοιχα, τότε μπορούμε να γράψουμε τη σχέση (1) ως εξής:

$$F_{Y-M_Y}(x) = F_{X-M_X}(\theta x) \quad \forall x \text{ και για κάποιο } \theta>0. \qquad (2)$$

Θεωρώντας ότι η σχέση (2) αντιστοιχεί στο αντίστοιχο της σχέσης (1), μη-παραμετρικό test, που βασίζεται στις διαμέσους M_X και M_Y των πληθυσμών X και Y (και όχι στις μέσες τιμές τους μ_X και μ_Y).

Το μη-παραμετρικό test που θα σχεδιάσουμε θα συνενώνει τα δύο δείγματα από τους πληθυσμούς X και Y σε ένα **ενιαίο** διατεταγμένο δείγμα κατασκευάζοντας μία κατάλληλη ελεγχοσυνάρτηση τάξεων μεγέθους γραμμικής μορφής όπως αυτή που είδαμε στο Φρ. #4 (linear rank test statistic).

Αν M_X και M_Y είναι γνωστά, τότε θέτοντας για το κάθε δείγμα: $X_i' = X_i - M_X$ και

$$Y_j' = Y_j - M_Y, \quad i=1,2,\dots,m \text{ και } j=1,2,\dots,n$$

οι πληθυσμοί X′ και Y′ θα έχουν μηδενικές διαμέσους, και οι τ.μ. X′ και Y′ στο ενιαίο διατεταγμένο δείγμα θα μπορούν να αναδείξουν τις διαφορές στις διακυμάνσεις χωρίς να "επηρεάζονται" από τις αντίστοιχες διαφορές στις διαμέσους τους. Τότε, μπορούμε να γράψουμε:

$$F_{Y'}(x) = F_{X'}(\theta x) \ \forall x \text{ και για κάποιο } \theta > 0 \qquad (2)'$$

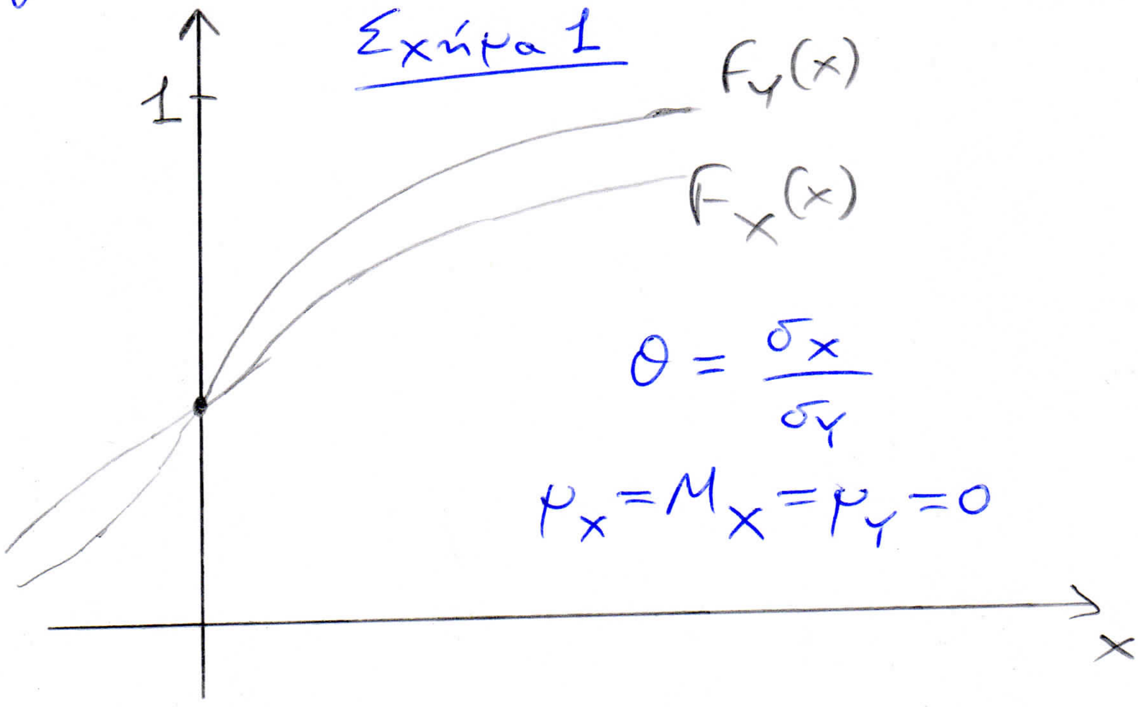
Αν επιπλέον υποθέσουμε $M_X = M_Y = M$, τότε μπορούμε να γράψουμε ως μηδενική υπόθεση στον έλεγχο την:

$$F_Y(x) = F_X(\theta x) \ \forall x \text{ και για κάποιο } \theta > 0, \ \theta \neq 1$$

δηλαδή ότι η σ.κ. του πληθυσμού Y είναι η ίδια με αυτή του πληθυσμού X εκτός από μία παράμετρο θ είτε "συμπιεσμένη" είτε όχι ($\theta > 1$ ή $\theta < 1$).

Στο Σχήμα 1, η σχέση $F_Y(x) = F_X(\theta x)$ απεικονίζεται για την $F_Y(x) = \Phi(x)$ και $\theta > 1$.

Σχήμα 1



Αρκετά tests που βασίζονται στους βαθμούς των παρατηρήσεων στο ενιαίο διατεταγμένο δείγμα έχουν προταθεί στη Βιβλιογραφία για το πρόβλημα κλίμακας (scale problem).

Αν αυτά τα tests επιθυμούμε να είναι χρήσιμα για να ανιχνεύσουν διαφορές μεταξύ των δύο πληθυσμών στις διακυμάνσεις τους θα πρέπει να υποθέσουμε ότι είτε οι διάμεσοι (ή οι μέσες τιμές) των δύο πληθυσμών είναι ίσες (και άγνωστες) ή ότι οι δειγματικές παρατηρήσεις μπορούν να προσαρμοστούν έτσι ώστε να έχουν τα ίδιο μέτρο θέσης για παράδειγμα, αφαιρώντας την κατάλληλη παράμετρο θέσης από τις μετρήσεις του ενός δείγματος. Πως γίνεται αυτό; Μπορεί ενδεχομένως να επιλεγεί ένα κατάλληλο σύνολο "βαρών" για την γραμμική ελεγχοσυνάρτηση τάξεων μεγέθους ώστε να απεικονίζεται η διακύμανση των παρατηρήσεων γύρω από μία κοινή κεντρική τιμή (μέση τιμή ή διάμεσος).

Freund-Ansari-Bradley-David-Barton tests

Στο ενιαίο διατεταγμένο δείγμα των $N=m+n$ μετρήσεων αν δεν έχουμε ισοβαθμίες ο μέσος βαθμός (average rank) είναι η μέση τιμή των πρώτων N ακέραιων αριθμών $\frac{N+1}{2}$.

Η απόκλιση κάθε βαθμού της i-οστής διατεταγμένης παρατήρησης από το μέσο βαθμό είναι:

$$i-\frac{(N+1)}{2}$$ και η "ποσότητα" αυτής της διαφοράς είναι μία ένδειξη σχετικής διακύμανσης των τιμών.

Αν χρησιμοποιήσουμε τις απόλυτες τιμές αυτών των αποκλίσεων έτσι ώστε να αποδώσουμε ίσα "βάρη" στις θετικές και στις αρνητικές αποκλίσεις, η γραμμική ελεγχοσυνάρτηση τάξεων μεγέθους (linear rank statistic) είναι:

$$A_N=\sum_{i=1}^{N}\left|i-\frac{(N+1)}{2}\right|Z_i,$$ όπου Z_i είναι οι τ.μ. που έχουμε ήδη ορίσει ως:

$$Z_i=\begin{cases}0, & \text{αν η i-οστή παρατήρηση ανήκει στο δείγμα των Y.}\\ 1, & \text{αν η i-οστή παρατήρηση ανήκει στο δείγμα των X.}\end{cases}$$

Διάφορες εναλλακτικές μορφές της παραπάνω σ.σ. ελέγχου, έχουν προταθεί κατά καιρούς από τους Freund and Ansari (1957), Ansari and Bradley (1960) και David and Barton (1958). Όλοι οι έλεγχοι που έχουν προταθεί είναι ισοδύναμοι.

- Η ελεγχοσυνάρτηση Freund-Ansari-Bradley έχει τη μορφή:

$$F_N=\frac{m(N+1)}{2}-A_N,$$

όπου A_N όπως ορίστηκε παραπάνω.

Η ακριβής κατανομή της F_N έχει προσδιοριστεί και για τον έλεγχο της $H_0: F_{Y-M}(x) = F_{X-M}(\theta x)$ $\forall x$ και για κάποιο $\theta > 0$, $\theta \neq 1$, υποθέτοντας ότι οι δύο πληθυσμοί έχουν την ίδια διάμεσο μπορούν να προσδιοριστούν οι p-τιμές και οι κρίσιμες περιοχές (αφού προσδιοριστούν τα κρίσιμα σημεία από κατάλληλους πίνακες τιμών).

$$H_0: F_{Y-M}(x) = F_{X-M}(\theta x), \quad \theta > 0, \ \theta \neq 1, \ \forall x$$

Εναλλακτική	Κρίσιμη περιοχή	P-value
$H_1: \theta > 1$	$F_N \leq k_1$	$P(F_N \leq f \mid H_0)$
$H_1: \theta < 1$	$F_N \geq k_2$	$P(F_N \geq f \mid H_0)$
$H_1: \theta \neq 1$ (3)	$F_N \leq k_3$ ή $F_N \geq k_4$	

Κάτω από την H_0 η κατανομή της F_N είναι συμμετρική γύρω από τη μέση τιμή αν N είναι άρτιος.

Τα k_1, k_2, k_3, k_4 προσδιορίζονται από κατάλληλους πίνακες, για την F_N.

Η ισχύς μίας από τις δύο ανισότητες στην εναλλακτική μορφή (3) καθορίζει τον υπολογισμό του p-value. Πίνακες για την F_N για $N \leq 20$, είναι διαθέσιμοι.

Για μεγάλα μεγέθη δείγματος, η κανονική προσέγγιση της κατανομής της F_N μπορεί να χρησιμοποιηθεί.

Αποδεικνύεται ότι:

αν N άρτιος	αν N περιττός
$E(F_N) = \frac{m(N+2)}{4}$	$E(F_N) = \frac{m(N+1)^2}{4N}$
$var(F_N) = \frac{mn(N^2-4)}{48(N-1)}$	$var(F_N) = \frac{mn(N+1)(N^2+3)}{48N^2}$

• Ένα εναλλακτικό ισοδύναμο test προτάθηκε από τους David and Barton (1958).

Η ελεγχοσυνάρτηση των David and Barton είναι πιο πολύπλοκη και έχει τη μορφή:

$$B_N = \sum_{i=1}^{[(N+1)/2]} \left(\left[\frac{N+2}{2}\right] - i\right) Z_i + \sum_{i=[(N+1)/2]+1}^{N} \left(i - \left[\frac{N+1}{2}\right]\right) Z_i$$

όπου $[x]$ είναι το ακέραιο μέρος του x. Για N άρτιο, οι B_N και F_N είναι ισοδύναμες σ.σ. ελέγχου (με ίσες μέσες τιμές και διασπορές). Για N περιττό η ισοδυναμία δεν ισχύει, και η ακριβής σχέση μεταξύ B_N και F_N είναι:

$$B_N + F_N = m\left[\frac{(N+2)}{2}\right].$$

Η ακριβής κατανομή της B_N κάτω από την H_0, δίνε- -9-
ται από κατάλληλους πίνακες τιμών (π.χ. για $n=m$,
με $m, n \leq 8$, David and Barton (1958)). Για την ασυμπτω-
τική κανονική προσέγγιση, αποδεικνύεται ότι:

$$E(B_N) = \frac{m(N+2)}{4}, \quad var(B_N) = \frac{mn(N^2-4)}{48(N-1)},$$

αν N άρτιος

και

$$E(B_N) = \frac{m(N^2-1)}{4N}, \quad var(B_N) = \frac{mn(N+1)(N^2+3)}{48N^2}$$

αν N περιττός.

Σύνοψη και Παράδειγμα

Υποθέσεις (για τον έλεγχο Ansari-Bradley)

α) Τα δεδομένα προέρχονται από 2 τ.δ. από δύο πληθυσμούς 1 και 2, X και Y.

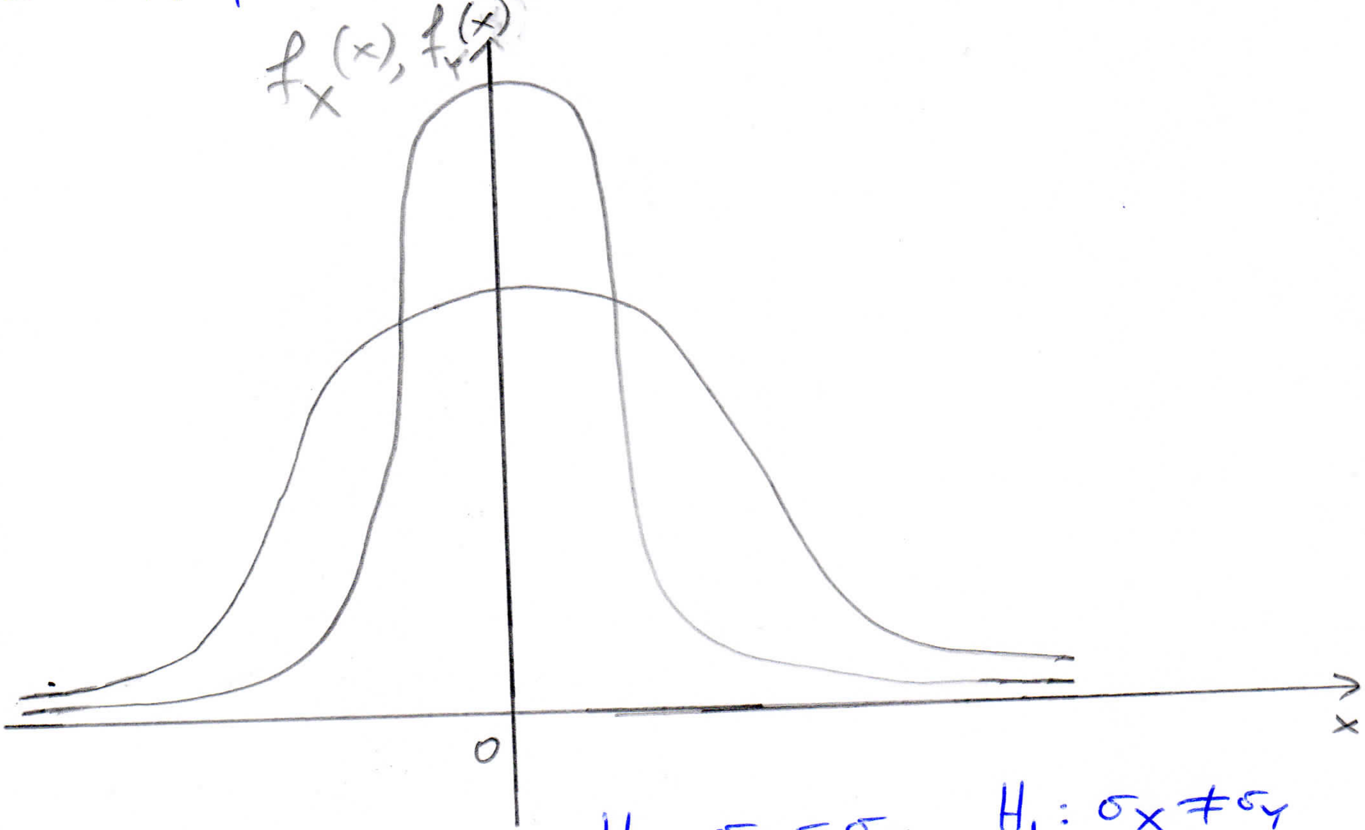
β) Οι κατανομές των πληθυσμών είναι συνεχείς.

γ) Τα δύο δείγματα είναι ανεξάρτητα.

δ) Τα δεδομένα μετρώνται τουλάχιστον σε διατάξιμη κλίμακα (ordinal scale).

ε) Οι δύο πληθυσμοί υποθέτουμε ότι έχουν ίσες διαμέσους και έχουν πιθανώς διαφορές ως προς τη διασπορά (ή την τυπική απόκλιση).

Οι κατανομές των πληθυσμών, έχουν π.χ. τη μορφή του σχήματος που ακολουθεί:



Ελέγχουμε: $H_0: \sigma_X = \sigma_Y$ $\quad H_1: \sigma_X \neq \sigma_Y$

ή $H_0: \sigma_X \leq \sigma_Y$ $\quad H_1: \sigma_X > \sigma_Y$

ή $H_0: \sigma_X \geq \sigma_Y$ $\quad H_1: \sigma_X < \sigma_Y$

Ελεγχοσυνάρτηση AB

- Τοποθετούμε το σύνολο των δεδομένων $N = n + m$, από την μικρότερη τιμή στην μεγαλύτερη τιμή.
- Η μικρότερη και η μεγαλύτερη τιμή στο ενιαίο δείγμα λαμβάνουν το βαθμό 1.
- Η αμέσως επόμενη μικρότερη και η αμέσως επόμενη μεγαλύτερη τιμή στο ενιαίο δείγμα λαμβάνουν το βαθμό 2 κ.ο.κ.
- Αν N άρτιος οι βαθμοί εισάγονται ως εξής:

  $1, 2, 3, \ldots, N/2, N/2, \ldots, 3, 2, 1.$

• Αν N περιττός οι βαθμοί εισάγονται ως εξής:

$$1,2,3,\ldots,\frac{(N-1)}{2},\frac{(N+1)}{2},\frac{(N-1)}{2},\ldots,3,2,1$$

• $AB=\sum_{i=1}^{m} R_i(X)$ δηλ. AB είναι το άθροισμα των βαθμών στις τιμές της X.

Απόφαση

(α) $\cancel{H}_0$ αν $AB < x_{1-\alpha/2,m,n}$ ή $AB \geq x_{\alpha/2,m,n}$

(β) $\cancel{H}_0$ αν $AB < x_{1-\alpha,m,n}$

(γ) $\cancel{H}_0$ αν $AB \geq x_{\alpha,m,n}$

Οι τιμές x (κρίσιμες τιμές δίνονται από κατάλληλους πίνακες τιμών για το AB-test).

Παράδειγμα

Έχουμε δεδομένα από έναν καρδιακό δείκτη (cardiac index) σε δύο groups ασθενών.

Group 1: ασθενείς με κανονική προσθετική λειτουργία βαλβίδας (normal prosthetic valve function).

Group 2: ασθενείς με μη-κανονική προσθετική λειτουργία βαλβίδας (abnormal prosthetic valve function).

Θέλουμε να ξέρουμε αν η διασπορά (ή αντίστοιχα η τυπική απόκλιση) ως προς τον καρδιακό δείκτη (variable of interest) στους 2 πληθυσμούς είναι ή όχι στατιστικά σημαντικά διαφορετική με βάση τα δύο δείγματα, σε εσσ $\alpha=5\%$.

Group 1 (X) 3.84 2.6 1.19 2.00

Group 2 (Y) 3.97 2.5 2.7 3.36 2.30

$H_o: \sigma_X = \sigma_Y$ $\qquad$ $H_1: \sigma_X \neq \sigma_Y$

ή $\sigma_X^2 = \sigma_Y^2$ $\qquad$ $\sigma_X^2 \neq \sigma_Y^2$

Συνδυάζοντας τα δύο δείγματα, έχουμε: $(N=9)$

	1.19	2	2.3	2.5	2.6	2.7	3.36	3.84	3.97
Group	x	x	y	y	x	y	y	x	y
Βαθμός	1	2	3	4	5	4	3	2	1

$AB = 1+2+5+2 = 10$

$m=4, \; n=5$

$\cancel{H_o}$ αν $AB < x_{1-\alpha/2,4,5} = 8$

ή αν $AB \geq x_{\alpha/2,4,5} = 16$

Η απόφαση είναι μη-απόρριψη της Ho. Άρα υπάρχουν ισχυρές ενδείξεις μη-απόρριψης της Ho συνεπώς οι δύο πληθυσμοί φαίνεται ότι δεν έχουν στατιστικά σημαντική διαφορά ως προς τις διακυμάνσεις ή τις τυπικές αποκλίσεις τους.

Με την R, το test πραγματοποιείται με την εντολή ansari.test (p-value = 0.7937 για αμφίπλευρο έλεγχο). Ο κώδικας σε R και τα αποτελέσματα επισυνάπτονται στην επόμενη σελίδα.

```
 x=c(3.84,2.6,1.19,2)
> y=c(3.97,2.5,2.9,3.36,2.3)
> ansari.test(x,y,alternative="two.sided",exact=TRUE)

        Ansari-Bradley test

data:  x and y
AB = 10, p-value = 0.7937
alternative hypothesis: true ratio of scales is not equal to 1
```