Στατιστικοί έλεγχοι για την επιλογή μοντέλου και την εγκυρότητα ενός προβλεπτικού αλγορίθμου

Ομιλητής: Α. Μπατσίδης

Τμήμα Μαθηματικών, Πανεπιστήμιο Ιωαννίνων, 45110 Ιωάννινα
abatsidis@uoi.gr

26-4-2017

Η ομιλία χωρίζεται σε δύο μέρη.

- Το πρώτο μέρος έχει ως θέμα τους στατιστικούς ελέγχους για την επιλογή μοντέλου. Έχει περισσότερο ερευνητικό χαρακτήρα. Τα ερευνητικά αποτελέσματα που θα παρουσιαστούν βασίζονται στην εργασία:
  M.D. Jiménez-Gamero, A. Batsidis, M.V. Alba-Fernández (2016). Fourier methods for model selection, Ann. Inst. Statis. Math., 68, 105-133.
- Το δεύτερο μέρος επικεντρώνεται στο πρόβλημα της αξιολόγησης της εγκυρότητας ενός προβλεπτικού αλγορίθμου. Έχει περισσότερο εκπαιδευτικό, πρακτικό χαρακτήρα. Βασίζεται σε υπό διαμόρφωση εργασία με τον συνάδελφο Πολυχρόνη Οικονόμου.

ΔΙΑΡΘΩΣΗ ΤΗΣ ΟΜΙΛΙΑΣ

Στα αρχικά στάδια της ανάπτυξής της η Στατιστική περιοριζόταν στη συλλογή, καταγραφή, οργάνωση και συνοπτική παρουσίαση, μέσω πινάκων, γραφημάτων και υπολογισμών απλών δεικτών (μέσοι όροι, ποσοστά κ.λ.π.).

Αρχικά λοιπόν η Στατιστική είχε εμπειρικό (περιγραφικό) χαρακτήρα (Περιγραφική Στατιστική).

Από τις αρχές του 20ου αιώνα όμως, με την ανάπτυξη και τη μαθηματική θεμελίωση της Θεωρίας Πιθανοτήτων, η Στατιστική, εκτός από την αρχική περιγραφική της μορφή (που εξακολουθεί να διατηρεί), άρχισε παράλληλα να προσλαμβάνει αυστηρή και μαθηματική μορφή (Μαθηματική-Επαγωγική Στατιστική) έχοντας ως κύριο στόχο την ανάπτυξη μεθοδολογιών για να είναι δυνατή η επέκταση των συμπερασμάτων στο όλο (πληθυσμό) από τη μελέτη και ανάλυση ενός μέρους (δείγμα).

Πληθυσμός: είναι το σύνολο των οντοτήτων ή συμβάντων των οποίων ένα ή περισσότερα χαρακτηριστικά (τυχαίες μεταβλητές) ενδιαφερόμαστε να μελετήσουμε.

Ο όρος τυχαία μεταβλητή χρησιμοποιείται για να περιγράψει μια ιδιότητα ή ένα χαρακτηριστικό γνώρισμα των μελών ενός πληθυσμού το οποίο μεταβάλεται από μέλος σε μέλος του πληθυσμού.

Η έννοια της τυχαίας μεταβλητής γενικεύεται στις p διαστάσεις ($p-$διάστατη τυχαία μεταβλητή ή τυχαίο διάνυσμα). Μια p-διάστατη τυχαία μεταβλητή αποτελείται από p συνήθεις τυχαίες μεταβλητές οι οποίες αποτελούν τις συντεταγμένες του τυχαίου διανύσματος.

Σε πρακτικά προβλήματα το μέγεθος του υπό εξέταση πληθυσμού είναι τόσο μεγάλο ώστε είναι αδύνατο να μελετηθεί όλος ο πληθυσμός. Έτσι εκλέγουμε ένα υποσύνολο του πληθυσμού (δείγμα) το οποίο μελετούμε και αναλύουμε. Τυχαίο δείγμα είναι το δείγμα που εκλέγεται κατά τέτοιο τρόπο ώστε όλα τα μέλη του πληθυσμού να έχουν την ίδια δυνατότητα να συμπεριληφθούν σε αυτό.

Το δείγμα συνίσταται από μετρήσεις, που γίνονται σε μέλη ενός πληθυσμού και αφορούν χαρακτηριστικά τους. Οι μετρήσεις αυτές θεωρούνται τιμές τυχαίων μεταβλητών, έχουν δηλαδή προκύψει από ένα πείραμα τύχης σύμφωνα με κάποια κατανομή πιθανότητας (το μοντέλο που τις περιγράφει).

Άλλες φορές είναι γνωστή η συναρτησιακή έκφραση της κατανομής αλλά είναι άγνωστοι οι παράμετροι που υπεισέρχονται στην έκφρασή της (Παραμετρική Στατιστική), ενώ άλλες φορές είναι εντελώς άγνωστη η μορφή της συνάρτησης κατανομής (Μη Παραμετρική Στατιστική).

Γίνεται αντιληπτό ότι καθοριστικό ρόλο στην εξαγωγή ορθών συμπερασμάτων στα πλαίσια της Παραμετρικής Στατιστικής διαδραματίζει η ορθή επιλογή μοντέλου.

Το πρώτο μέρος της ομιλίας θα επικεντρωθεί στην επιλογή μοντέλου μέσω του ελέγχου στατιστικών υποθέσεων.

Ο Έλεγχος Στατιστικών Υποθέσεων αποτελεί κλάδο της Στατιστικής Συμπερασματολογίας.

Μια στατιστική υπόθεση είναι μια υπόθεση που δύναται να εξεταστεί χρησιμοποιώντας κάποιες παρατηρήσεις, έστω $X_1, ..., X_n$, που μπορούν να θεωρούνται τιμές τυχαίων μεταβλητών. Με άλλα λόγια οποιαδήποτε υπόθεση αναφέρεται στη συμπεριφορά τυχαίων μεταβλητών για τις οποίες μπορούμε να έχουμε παρατηρήσεις είναι μια στατιστική υπόθεση.

Η υπόθεση η οποία ορίζεται με την ελπίδα να απορριφθεί ονομάζεται μηδενική υπόθεση (null hypothesis, H_0). Για να την ελέγξουμε χρειαζόμαστε μια εναλλακτική υπόθεση (alternative hypothesis, H_a ή H_1).

Ο έλεγχος στατιστικών υποθέσεων δίνει τον τρόπο για να αποφασίσουμε για την ορθότητα ή μη της H_0 έναντι της H_a. Η όλη διαδικασία για το σκοπό αυτό λέγεται στατιστικό τεστ.

Βήματα κατασκευής στατιστικού τεστ

- Εύρεση μιας στατιστικής συνάρτησης, δηλαδή μιας συνάρτησης των $X_1, ..., X_n$, με γνωστή κατανομή όταν αληθεύει η H_0.
- Προσδιορισμός της περιοχής απόρριψης της H_0 (κρίσιμη περιοχή).
- Δύο τύποι σφαλμάτων:
  Σφάλμα Τύπου I: Ο στατιστικός απορρίπτει την H_0, ενώ αληθεύει η H_0.
  Σφάλμα Τύπου II: Ο στατιστικός αποδέχεται την H_0, ενώ αληθεύει η H_a.
- Οι αντίστοιχες πιθανότητες συμβολίζονται με α και β. Η πιθανότητα Σφάλματος Τύπου I είναι γνωστή και ως επίπεδο σημαντικότητας.
- Το ιδανικό θα ήταν να μπορούσε κανείς να βρει διαδικασία ελέγχου που θα ελαχιστοποιεί τις ποσότητες α και β ταυτόχρονα (ανέφικτο).
- Κατασκευή κρίσιμων περιοχών με προκαθορισμένο α και την προσδοκία να μεγιστοποιείται το $\gamma = 1 - \beta$ (ισχύς).

Έστω $X_1, X_2, \ldots, X_n$ ένα τυχαίο δείγμα από έναν άγνωστο πληθυσμό F και δύο παραμετρικές οικογένειες $\mathcal{F}$ και $\mathcal{G}$, με παραμέτρους $\theta \in \Theta \subseteq \mathbb{R}^k$ και $\gamma \in \Gamma \subseteq \mathbb{R}^r$, που μπορεί να είναι είτε μη εμφωλευμένες, είτε επικαλυπτόμενες είτε εμφωλευμένη η μία στην άλλη.

- Δύο παραμετρικές οικογένειες κατανομών $\mathcal{F}$ και $\mathcal{G}$ λέμε ότι είναι μη εμφωλευμένες ή ξεχωριστές (nonnested ή separate) αν

$$\mathcal{F} \cap \mathcal{G} = \emptyset.$$

- Δύο παραμετρικές οικογένειες κατανομών $\mathcal{F}$ και $\mathcal{G}$ λέμε ότι είναι επικαλυπτόμενες (overlapping) αν

$$\mathcal{F} \cap \mathcal{G} \neq \emptyset, \mathcal{F} \nsubseteq \mathcal{G} \text{ και } \mathcal{G} \nsubseteq \mathcal{F}.$$

- Η παραμετρική οικογένεια κατανομών $\mathcal{G}$ λέμε ότι είναι εμφωλευμένη στην $\mathcal{F}$ αν

$$\mathcal{G} \subset \mathcal{F}.$$

Το πρόβλημα επιλογής του κατάλληλου μοντέλου μέσω του ελέγχου υποθέσεων ανάγεται στον έλεγχο της υπόθεσης αν τα μοντέλα έχουν τον ίδιο βαθμό εγγύτητας στον αληθινό πληθυσμό, έναντι της υπόθεσης ότι το ένα είναι πλησιέστερα στον αληθινό πληθυσμό από ότι το άλλο. Επομένως ανάγεται στον έλεγχο της

$$H_0 : \ D(F, \mathcal{F}) = D(F, \mathcal{G})$$

έναντι μιας εκ των εναλλακτικών

$$H_{1\mathcal{F}} : \ D(F, \mathcal{F}) < D(F, \mathcal{G}) \quad \text{ή} \quad H_{1\mathcal{G}} : \ D(F, \mathcal{F}) > D(F, \mathcal{G})$$

όπου D ένα μέτρο της εγγύτητας ή απόκλισης (divergence) των δύο πληθυσμών.

Στη στατιστική βιβλιογραφία έχουν εμφανιστεί πλήθος μέτρων εγγύτητας, απόκλισης δύο πληθυσμών και έχουν αξιοποιηθεί σε στατιστικούς ελέγχους υποθέσεων.

Μια εκτενής ανασκόπηση της βιβλιογραφίας στο πλαίσιο αυτό δίνεται στη μονογραφία του Pardo (2006).

Στο σημείο αυτό θα ήταν παράλειψη να μην αναφερθεί ότι ο Τομέας Στατιστικής του Τμήματος Μαθηματικών του Πανεπιστημίου Ιωαννίνων έχει μια διαχρονική παράδοση σε τέτοια θέματα καθώς μέλη του (εν ενεργεία και διατελέσαντα, Papaioannou, Ferentinos, Zografos, Tsairidis, Micheas, Avlogiaris) έχουν συμβολή, άλλος μεγαλύτερη και άλλος μικρότερη, στην περιοχή αυτής της Στατιστικής Επιστήμης.

Ειδική μνεία πρέπει να γίνει στην επί πολλά έτη ερευνητική συνεργασία του Κ. Ζωγράφου με την ερευνητική ομάδα του Leadro Pardo (M.L. Menendez^{+}, Julio Angel Pardo, Domingo Morales, Nirian Martin).

Η πιο ευρέως χρησιμοποιούμενη μέθοδος ελέγχου της

$$H_0 : \ D(F, \mathcal{F}) = D(F, \mathcal{G})$$

έναντι μιας εκ των εναλλακτικών

$$H_{1\mathcal{F}} : \ D(F, \mathcal{F}) < D(F, \mathcal{G}) \quad \text{ή} \quad H_{1\mathcal{G}} : \ D(F, \mathcal{F}) > D(F, \mathcal{G})$$

είναι αυτή του Vuong (1989) όπου D το πλέον διαδεδομένο μέτρο απόκλισης που παρουσιάστηκε από τους Kullback and Leibler (1951) και διεξοδικά μελετήθηκε από τον Kullback (1959).
Έστω $f(\cdot)$ και $g(\cdot)$ δύο συναρτήσεις πυκνότητας πιθανότητας. Τότε το (KL) μέτρο απόκλισης της g από την f:

$$D_X^{KL} = \int f(x) ln \left(\frac{f(x)}{g(x)} \right) dx$$

Οι έλεγχοι του Vuong (1989) εφαρμόζονται όταν πληρούνται οι συνθήκες A1-A7 της εργασίας του και βασίζονται στη στατιστική συνάρτηση:

$$\frac{1}{n}\sum_{i=1}^{n}\log\frac{f\left(X_i;\hat{\theta}_n\right)}{g\left(X_i;\hat{\gamma}_n\right)}$$

όπου f και g οι συναρτήσεις πυκνότητας πιθανότητας των $\mathcal{F}$ και $\mathcal{G}$, αντίστοιχα, ενώ

$$\sum_{i=1}^{n}\log f\left(X_i;\hat{\theta}_n\right)=\sup_{\theta\in\Theta}\sum_{i=1}^{n}\log f\left(X_i;\theta\right)$$

και

$$\sum_{i=1}^{n}\log g\left(X_i;\hat{\gamma}_n\right)=\sup_{\gamma\in\Gamma}\sum_{i=1}^{n}\log g\left(X_i;\gamma\right),$$

αντίστοιχα.

Περιπτώσεις μη εφαρμογής της μεθόδου του Vuong (1989).

- Δεν εφαρμόζεται η μεθοδολογία του Vuong (1989) όταν η πιθανοφάνεια ενός μοντέλου δε μπορεί να υπολογιστεί. Για παράδειγμα για stable (ευσταθείς) κατανομές.
- Ο εκτιμητής μέγιστης πιθανοφάνειας της μιας οικογένειας κατανομών δεν συγκλίνει υπό την άλλη (βλέπε Cauchy vs Normal).
- Το πεδίο ορισμού ενός εκ των δύο ανταγωνιστικών μοντέλων εξάρτάται από την άγνωστη παράμετρο, (βλέπε αποκομμένη (truncated) Laplace κατανομή με παραμέτρους θέσης και κλίμακας) κ.α.

Κοινό χαρακτηριστικό των παραδειγμάτων:

η χαρακτηριστική συν/ση ($E[exp(it^{'}X)]$) δίνεται σε απλή κλειστή μορφή.

- Η χαρακτηριστική συνάρτηση χαρακτηρίζει πλήρως μια κατανομή.
- Η χαρακτηριστική συνάρτηση υπάρχει πάντοτε, σε αντίθεση με τη ροπογεννήτρια συνάρτηση ($E[exp(t^{'}X)]$).
- Cramer (1946): πρωτοπαρουσιάζεται ο ορισμός της εμπειρικής χαρακτηριστικής συνάρτησης

$$c_n(t) = \frac{1}{n}\sum_{j=1}^{n} \exp(\mathrm{i}t'X_j).$$

- Press (1972): εκτίμηση παραμέτρων ευσταθών κατανομών.
- Heathcote (1972): έλεγχο καλής προσαρμογής.
- Από τότε πλήθος ερευνητικών εργασιών αξιοποιούν τη χαρακτηριστική συνάρτηση και την εμπειρική χαρακτηριστική συνάρτηση σε ποικίλα ερευνητικά θέματα.

Ενδεικτικές εργασίες:

- Εκτιμητική, Έλεγχοι υποθέσεων: Heathcote (1977), Feuerverger and Mureika (1977), Koutrouvelis (1980, 1981, 1982), Feuerverger and McDunnough (1981a, 1981b), Epps and Singleton (1986), Csorgo and Heathcote (1987), Alba-Fernandez et al. (2001, 2008), Henze et al. (2003, 2014).
- Έλεγχοι καλής προσαρμογής: Koutrouvelis and Kellermeier (1981), Csorgo (1986), Baringhaus and Henze (1988), Epps and Pulley (1983, 1986), Meintanis and Koutrouvelis (1990, 1991), Epps (1995), Fan (1997),Gurtler and Henze (2000), Henze and Wagner (1997), Huskova and Meintanis (2007,2009,2010,2012), Matsui and Takemura (2005), Jimenez-Gamero (2014), Jimenez-Gamero and Kim (2015), Koutrouvelis (2016)
- Έλεγχοι αλλαγής σημείου: Huskova and Meintanis (2006a,b) Hlavka and Huskova (2012).
- Εργασίες ανασκόπησης: Csorgo (1984), Huskova and Meintanis (2008 a, b), Meintanis (2016).

Ως μέτρο απόστασης δύο d-διάστατων πληθυσμών, με χαρακτηριστικές συναρτήσεις $c_1(t)$ και $c_2(t)$, $t \in \mathbb{R}^d$, θεωρούμε το:

$$D^2(c_1, c_2) = \int_{\mathbb{R}^d} |c_1(t) - c_2(t)|^2 w(t) dt,$$

όπου για $z = a + \mathrm{i}b$, $|z|^2 = a^2 + b^2$, $w(t) > 0$, με $w(t) = w(-t)$, $\forall t \in \mathbb{R}^d$.

- Κεντρική ιδέα: επιλογή του μοντέλου που ελαχιστοποιεί την απόσταση μεταξύ ενός εκτιμητή της πληθυσμιακής χαρακτηριστικής συνάρτησης και ενός εκτιμητή της χαρακτηριστικής συνάρτησης του μοντέλου.
- Η πληθυσμιακή χ.σ. εκτιμάται από την ε.χ.σ. Η εκτίμηση της χ.σ. του μοντέλου προκύπτει με αντικατάσταση των άγνωστων παραμέτρων από κατάλληλους εκτιμητές.
- Minimum integrated squared error (ISE) εκτιμητές (Heathcote, 1977).

Ορισμός (Minimum) integrated squared error (ISE) εκτιμητής

Έστω $X_1, X_2, \ldots, X_n$ τ.δ. από έναν πληθυσμό με χαρ/κη συν/ση $c(t)$ και $\mathcal{F} = \{c_F(t;\theta) = u_F(t;\theta) + \mathrm{i}\, v_F(t;\theta);\ \theta \in \Theta \subseteq \mathbb{R}^k\}$ μια οικογένεια κατανομών με ταυτοποιήσιμα (identifiable) στοιχεία δηλαδή $c(t;\theta_1) \neq c(t;\theta_2)$, υπό την έννοια ότι $\sup_{t\in[0,1]^d} |c(t;\theta_1) - c(t;\theta_2)| > 0$, όταν $\theta_1 \neq \theta_2$. Είναι

$$\begin{aligned} \hat{\theta}_n &= \arg\min_{\theta\in\Theta} D^2\left(c_n(t), c_F(t;\theta)\right) \\ &= \arg\min_{\theta\in\Theta} \int |c_n(t) - c_F(t;\theta)|^2 w(t) dt \end{aligned}$$

Heathcote (1977) και Csörgő (1981): ασυμπτωτικές ιδιότητες όταν το μοντέλο είναι σωστά προσδιορισμένο ($c(t) \in \mathcal{F}$).

Ορισμός

Έστω ότι $c(t) \notin \mathcal{F}$. Θα είναι $\theta_* \in \Theta$ η τιμή της παραμέτρου θ που είναι τέτοια ώστε $c_F(t;\theta_*)$ να είναι το στοιχείο της οικογένειας $\mathcal{F}$ πλησιέστερα στην $c(t)$, δηλαδή

$$\begin{aligned} \theta_*(w) = \theta_* &= \arg\min_{\theta\in\Theta} D^2(c(t), c_F(t;\theta)) \\ &= arg\min_{\theta\in\Theta} \int |c(t) - c_F(t;\theta)|^2 w(t)dt. \end{aligned}$$

- Αν $c(t) \in \mathcal{F}$, δηλαδή αν $c(t) = c_F(t;\theta)$, για κάποιο $\theta \in \Theta$, τότε $\theta_* = \theta$.
- Η τιμή θ_* μπορεί είτε να μην υπάρχει είτε αν υπάρχει να μην είναι μοναδική.

ΥΠΟΘΕΣΗ 1: $D^2(c(t), c_F(t;\theta)) = \int |c(t) - c_F(t;\theta)|^2 w(t)dt$ λαμβάνει μοναδικό ελάχιστο στο $\theta_* \in \Theta$.

Η Υπόθεση 1 είναι συνηθισμένη σε παρόμοιες εργασίες όπου η εκτίμηση των παραμέτρων προκύπτει από την ελαχιστοποίηση ενός μέτρου εγγύτητας μεταξύ του πληθυσμού και της οικογένειας κατανομών (βλέπε White (1982 Assumption A3(b)),Broniatowski and Keziou (2009, Assumption C.1).

Θεώρημα 1

Αν ισχύει η Υπόθεση 1 τότε

$$\hat{\theta}_n \xrightarrow{a.s.} \theta_*.$$

Αυτό σημαίνει ότι οι τιμές του $\hat{\theta}_n$ προσεγγίζουν την τιμή του θ_*, με την έννοια ότι τιμές για τις οποίες η $\hat{\theta}_n$ δε συγκλίνει στην θ_* έχουν πιθανότητα 0.

$$P(lim_{n\to\infty}\hat{\theta}_n = \theta_*) = 1.$$

ΥΠΟΘΕΣΗ 2: Οι συναρτήσεις $u_F(t;\theta)$ και $v_F(t;\theta)$ είναι δύο φορές συνεχώς διαφορίσιμες (twice continuously differentiable) ως προς θ, για όλα τα θ σε μια ανοικτή περιοχή (neighborhood) $\Theta_1 \subset \Theta$ του θ_*. Επιπρόσθετα, η πρώτη και δεύτερη παράγωγος ως προς θ των $u_F(t;\theta)$ και $v_F(t;\theta)$ είναι ομοιόμορφα ($\forall\theta \in \Theta_1$) φραγμένες από W-ολοκληρώσιμές συναρτήσεις.

Η Υπόθεση 2 εξασφαλίζει ότι

$$D^2\left(c_n(t), c_F(t;\theta)\right) = \int |c_n(t) - c_F(t;\theta)|^2 w(t)dt,$$

είναι δύο φορές συνεχώς διαφορίσιμη ως προς θ, $\forall\theta \in \Theta_1$ και μπορεί να την παραγωγίσουμε υπό το ολοκλήρωμα.

Θεώρημα 2

Αν ισχύουν οι Υποθέσεις 1 και 2 και $\theta_* \in int\Theta$, τότε

$$\sqrt{n}(\hat{\theta}_n - \theta_*) \xrightarrow{\mathcal{L}} N_k(0, \Sigma),$$

όπου $\Sigma = D_2(\theta_*)^{-1} A(\theta_*) D_2(\theta_*)^{-1}$ και $A(\theta) = E\{h(X;\theta)h(X;\theta)'\}$, με $h(x;\theta) = (h_1(x;\theta), \ldots, h_k(x;\theta))'$,

$$\begin{aligned} h_j(x;\theta) &= \int \{\cos(t'x) - u_F(t;\theta)\} \frac{\partial}{\partial \theta_j} u_F(t;\theta) w(t) dt \\ &+ \int \{\sin(t'x) - v_F(t;\theta)\} \frac{\partial}{\partial \theta_j} v_F(t;\theta) w(t) dt, 1 \leq j \leq k, \end{aligned}$$

$$D_2(\theta) = \frac{1}{2} \frac{\partial^2}{\partial \theta \partial \theta'} D^2(c(t); c_F(t;\theta))$$

Έστω $X_1, X_2, \ldots, X_n$ τ.δ. από έναν πληθυσμό με χ.σ. $c(t)$, και δυο οικογένειες κατανομων

$$\mathcal{F} = \{c_F(t;\theta) = u_F(t;\theta) + iv_F(t;\theta);\ \theta \in \Theta \subseteq \mathbb{R}^k\},$$

$$\mathcal{G} = \{c_G(t;\gamma) = u_G(t;\gamma) + iv_G(t;\gamma);\ \gamma \in \Gamma \subseteq \mathbb{R}^r\}.$$

$$H_0 :\ \mu_{FG}(\theta_*, \gamma_*) = 0$$

$$H_{1\mathcal{F}} :\ \mu_{FG}(\theta_*, \gamma_*) < 0 \quad \text{ή} \quad H_{1\mathcal{G}} :\ \mu_{FG}(\theta_*, \gamma_*) > 0,$$

όπου

$$\begin{aligned} \mu_{FG}(\theta_*, \gamma_*) &= D^2(c(t), \mathcal{F}) - D^2(c(t), \mathcal{G}) \\ &= D^2(c(t), c_F(t;\theta_*)) - D^2(c(t), c_G(t;\gamma_*)). \end{aligned}$$

Θεώρημα 3

Υποθέτουμε ότι $u_F(t;\theta)$ και $v_F(t;\theta)$ είναι συνεχείς ως προς θ για κάθε t, $u_G(t;\gamma)$ και $v_G(t;\gamma)$ είναι συνεχείς ως προς γ για κάθε t. Αν ισχύει η Υπόθεση 1 για τις $\mathcal{F}$, $\mathcal{G}$, $\theta_* \in int\Theta$, $\gamma_* \in int\Gamma$:

$$T(\hat{\theta}_n, \hat{\gamma}_n) = D^2(c_n(t), c_F(t;\hat{\theta}_n)) - D^2(c_n(t), c_G(t;\hat{\gamma}_n)) \xrightarrow{a.s.} \mu_{FG}(\theta_*, \gamma_*).$$

$$T(\hat{\theta}_n, \hat{\gamma}_n) = \frac{1}{n}\sum_{j=1}^{n} \xi(X_i, \hat{\theta}_n, \hat{\gamma}_n), \tag{1}$$

όπου

$$\begin{aligned}\xi(x,\theta,\gamma) &= \int \{u_G(t;\gamma) - u_F(t;\theta)\}\{2\cos(t'x) - u_G(t;\gamma) - u_F(t;\theta)\}w(t)dt \\ &+ \int \{v_G(t;\gamma) - v_F(t;\theta)\}\{2\sin(t'x) - v_G(t;\gamma) - v_F(t;\theta)\}w(t)dt.\end{aligned}$$

Θεώρημα 4

Έστω ότι οι οικογένειες κατανομών $\mathcal{F}$ και $\mathcal{G}$ ικανοποιούν τις υποθέσεις του Θεωρήματος 2.
(α) Αν $c_F(t;\theta_*) = c_G(t;\gamma_*)$, τότε

$$nT(\hat{\theta}_n, \hat{\gamma}_n) \xrightarrow{\mathcal{L}} T_1 = \sum_{j=1}^{k+r} \lambda_j \chi^2_{1j},$$

όπου $\chi^2_{11}, \chi^2_{12}, ...,$ είναι ανεξάρτητες χ^2_1 και $\{\lambda_j\}$ είναι το σύνολο των ιδιοτιμών ενός συγκεκριμένου πίνακα.
(β) Αν $c_F(t;\theta_*) \neq c_G(t;\gamma_*)$, τότε

$$\sqrt{n}\left\{T(\hat{\theta}_n, \hat{\gamma}_n) - \mu_{FG}(\theta_*, \gamma_*)\right\} \xrightarrow{\mathcal{L}} N(0, \sigma^2_{FG}(\theta_*, \gamma_*)),$$

με $\sigma^2_{FG}(\theta, \gamma) = var\{\xi(X, \theta, \gamma)\} > 0.$

Επομένως, είναι σημαντικό να γνωρίζουμε αν $c_F(t;\theta_*) = c_G(t;\gamma_*)$ ή όχι.

Θεώρημα 5

$$\sigma^2_{FG}(\theta,\gamma) = 0 \Longleftrightarrow c_F(t;\theta) = c_G(t;\gamma).$$

Επομένως ο έλεγχος της $c_F(t;\theta_*) = c_G(t;\gamma_*)$ έναντι της $c_F(t;\theta_*) \neq c_G(t;\gamma_*)$ είναι ισοδύναμος με τον έλεγχο της

$$\begin{aligned} H_{0\sigma}: &\quad \sigma^2_{FG}(\theta_*,\gamma_*) = 0, \\ H_{1\sigma}: &\quad \sigma^2_{FG}(\theta_*,\gamma_*) > 0. \end{aligned}$$

$$\hat{\sigma}^2_{FG}(\hat{\theta}_n,\hat{\gamma}_n) = \frac{1}{n}\sum_{j=1}^{n}\xi^2(X_j,\hat{\theta}_n,\hat{\gamma}_n) - \left\{\frac{1}{n}\sum_{j=1}^{n}\xi(X_j,\hat{\theta}_n,\hat{\gamma}_n)\right\}^2.$$

Θεώρημα 6

(α) Αν $\mathcal{F}$ και $\mathcal{G}$ ικανοποιούν την Υπόθεση 1, $\theta_* \in int\Theta$, $\gamma_* \in int\Gamma$, $u_F(t;\theta)$ και $v_F(t;\theta)$ συνεχείς συναρτήσεις του θ για κάθε t, $u_G(t;\gamma)$ και $v_G(t;\gamma)$ συνεχείς συναρτήσεις του γ για κάθε t, τότε

$$\hat{\sigma}^2_{FG}(\hat{\theta}_n, \hat{\gamma}_n) \xrightarrow{a.s.} \sigma^2_{FG}(\theta_*, \gamma_*).$$

(β) Αν οι οικογένειες κατανομών $\mathcal{F}$ και $\mathcal{G}$ ικανοποιούν τις υποθέσεις του Θεωρήματος 2 και $\sigma^2_{FG}(\theta_*, \gamma_*) = 0$, τότε

$$T_\sigma = 0.25n\hat{\sigma}^2_{FG}(\hat{\theta}_n, \hat{\gamma}_n) \xrightarrow{\mathcal{L}} T_2 = \sum_{j=1}^{k+r} \lambda^\sigma_j \chi^2_{1j},$$

όπου $\chi^2_{11}, \chi^2_{12}, ...,$ είναι ανεξάρτητες χ^2_1 και $\{\lambda^\sigma_j\}$ είναι το σύνολο των ιδιοτιμών συγκεκριμένου πίνακα.

Σε αυτήν την περίπτωση πάντοτε ισχύει ότι $c_F(t;\theta_*) \neq c_G(t;\gamma_*)$.

Θεώρημα 4: Αν $c_F(t;\theta_*) \neq c_G(t;\gamma_*)$, τότε υπό την $H_0 : \mu_{FG}(\theta_*, \gamma_*) = 0$

$$\sqrt{n}T(\hat{\theta}_n, \hat{\gamma}_n) \xrightarrow{\mathcal{L}} N(0, \sigma^2_{FG}(\theta_*, \gamma_*)).$$

Θεώρημα 6:

$$\hat{\sigma}^2_{FG}(\hat{\theta}_n, \hat{\gamma}_n) \xrightarrow{a.s.} \sigma^2_{FG}(\theta_*, \gamma_*).$$

$$D = \sqrt{n}T(\hat{\theta}_n, \hat{\gamma}_n)/\hat{\sigma}_{FG}(\hat{\theta}_n, \hat{\gamma}_n). \tag{2}$$

Κανόνας απόφασης με επίπεδο σημαντικότητας $\alpha \in (0, 1)$:

- αν $D < -Z_{1-\alpha/2}$ τότε επιλέγουμε το μοντέλο $\mathcal{F}$.
- αν $D > Z_{1-\alpha/2}$ τότε επιλέγουμε το μοντέλο $\mathcal{G}$.
- αν $|D| \leq Z_{1-\alpha/2}$ τότε δεν υπάρχουν ενδείξεις για να προτιμήσουμε κάποιο από τα δύο μοντέλα $\mathcal{F}$ και $\mathcal{G}$,

όπου $\Phi(Z_{1-\alpha/2}) = 1 - \alpha/2$, με Φ την αθροιστική συνάρτηση κατανομής της $N(0, 1)$.

Σε αυτήν την περίπτωση δεν ξέρουμε αν ισχύει ή όχι ότι $c_F(t;\theta_*) = c_G(t;\gamma_*)$.

Πρώτο βήμα:Έλεγχος της $H_{0\sigma} : \sigma^2_{FG}(\theta_*, \gamma_*) = 0$ με επίπεδο σημαντικότητας α_1.

- Θεώρημα 6: Υπό την $H_{0\sigma}$, $T_\sigma \xrightarrow{\mathcal{L}} T_2 = \sum_{j=1}^{k+r} \lambda_j^\sigma \chi^2_{1j}$.
- $\{\lambda_j^\sigma\}$ είναι το σύνολο των ιδιοτιμών συγκεκριμένου πίνακα (εξαρτάται από τις άγνωστες τιμές θ_*, γ_*).
- $\{\hat{\lambda}_j^\sigma\}$ το σύνολο των ιδιοτιμών του πίνακα με αντικατάσταση των θ_*, γ_* από $\hat{\theta}_n, \hat{\gamma}_n$.
- Τότε $\sup_x \left| P_{0\sigma}(T_\sigma \le x) - P_*(\hat{T}_\sigma \le x) \right| \xrightarrow{a.s.} 0$, όπου $\hat{T}_\sigma = \sum_{j=1}^{k+r} \hat{\lambda}_j^\sigma \chi^2_{1j}$ και P_* η δεσμευμένη πιθανότητα δοθέντος των δεδομένων.
- Αν $T_\sigma < \hat{t}_{\sigma,1-\alpha_1}$, τότε ισοδύναμα μοντέλα, όπου $\hat{t}_{\sigma,1-\alpha_1}$ τέτοιο ώστε $P_*\left(\hat{T}_\sigma \le \hat{t}_{\sigma,1-\alpha_1}\right) = 1 - \alpha_1$.
- Νέο πρόβλημα: η κατανομή γραμμικού συνδυασμού χ^2 είναι άγνωστη γενικά. Τρόπος επίλυσης: προσομοίωση της δεσμευμένης κατανομής του $\hat{T}_\sigma$ δοθέντος των δεδομένων.

Αν στο προηγούμενο βήμα $T_\sigma \geq \hat{t}_{\sigma,1-\alpha_1}$ τότε με επίπεδο σημαντικότητας α_1 δεν απορρίπτεται η υπόθεση $H_{0\sigma} : \sigma^2_{FG}(\theta_*, \gamma_*) = 0$, δηλαδή η υπόθεση ότι $c_F(t;\theta_*) = c_G(t;\gamma_*)$, προχωρούμε χρησιμοποιώντας το Θεώρημα 4 στο επόμενο βήμα.

Δεύτερο βήμα (αν $T_\sigma \geq \hat{t}_{\sigma,1-\alpha_1}$): Έλεγχος $H_0 : \mu_{FG}(\theta_*, \gamma_*) = 0$.

- Αν $T_\sigma \geq \hat{t}_{\sigma,1-\alpha_1}$ και $D < -Z_{1-\alpha_2/2}$ τότε $\mathcal{F}$.
- Αν $T_\sigma \geq \hat{t}_{\sigma,1-\alpha_1}$ και $D > Z_{1-\alpha_2/2}$ τότε $\mathcal{G}$.

Επίπεδο σημαντικότητας $\alpha \leq \max\{\alpha_1, \alpha_2\}$.

Έστω τώρα ότι η οικογένεια κατανομών $\mathcal{G}$ είναι εμφωλευμένη στην $\mathcal{F}$ δηλαδή $\mathcal{G} \subset \mathcal{F}$. Η οικογένεια κατανομών $\mathcal{G}$ δεν μπορεί να δώσει ποτέ καλύτερη προσαρμογή από την $\mathcal{F}$.

- Αποδεικνύεται ότι η H_0 σε αυτήν την περίπτωση είναι ισοδύναμη με τον έλεγχο της $c_F(t;\theta_*) = c_G(t;\gamma_*)$
- Κανόνας Απόφασης (Θεώρημα 6, πρώτο βήμα προηγούμενου ελέγχου): Αν $T_\sigma \geq \hat{t}_{\sigma,1-\alpha}$ τότε επιλέγουμε $\mathcal{F}$, ενώ αν $T_\sigma < \hat{t}_{\sigma,1-\alpha}$ τότε ισοδύναμα μοντέλα.

- Στόχος: να προταθεί μια μεθοδολογία για κάποιες περιπτώσεις (εδώ παρουσιάζονται δύο) που δεν εφαρμόζεται η μεθοδολογία του Vuong (1989) και να αξιολογηθεί η απόδοσή της μέσω μελέτης προσομοίωσης.
- Να συγκριθεί με την κλασική μεθοδολογία για περιπτώσεις που και οι δύο μπορούν να εφαρμοστούν (εδώ παρουσιάζεται μία).
- Σε κάθε περίπτωση 10000 δείγματα μεγέθους n ($n = 50, 100, 200, 300$) από τις $\mathcal{F}$ και $\mathcal{G}$ για διάφορες τιμές των παραμέτρων θ και γ δημιουργήθηκαν.
- Εφαρμόζονται οι μεθοδολογίες και καταγράφεται κάθε φορά το ποσοστό κάθε απόφασης.
- Επιλογή $w(t)$: επιδρά τόσο στην ισχύ της μεθόδου όσο και στον υπολογιστικό χρόνο. Επιλέξαμε συναρτήσεις τέτοιες ώστε να είναι πιο εύκολοι οι υπολογισμοί.

Η αποκομμένη αριστερά οικογένεια κατανομών στο σημείο $\theta \in \Theta \subseteq \mathbb{R}$ έχει συνάρτηση πυκνότητας πιθανότητας

$$f(x;\theta) = \frac{l(x)}{1 - L(\theta)}, \quad x > \theta,$$

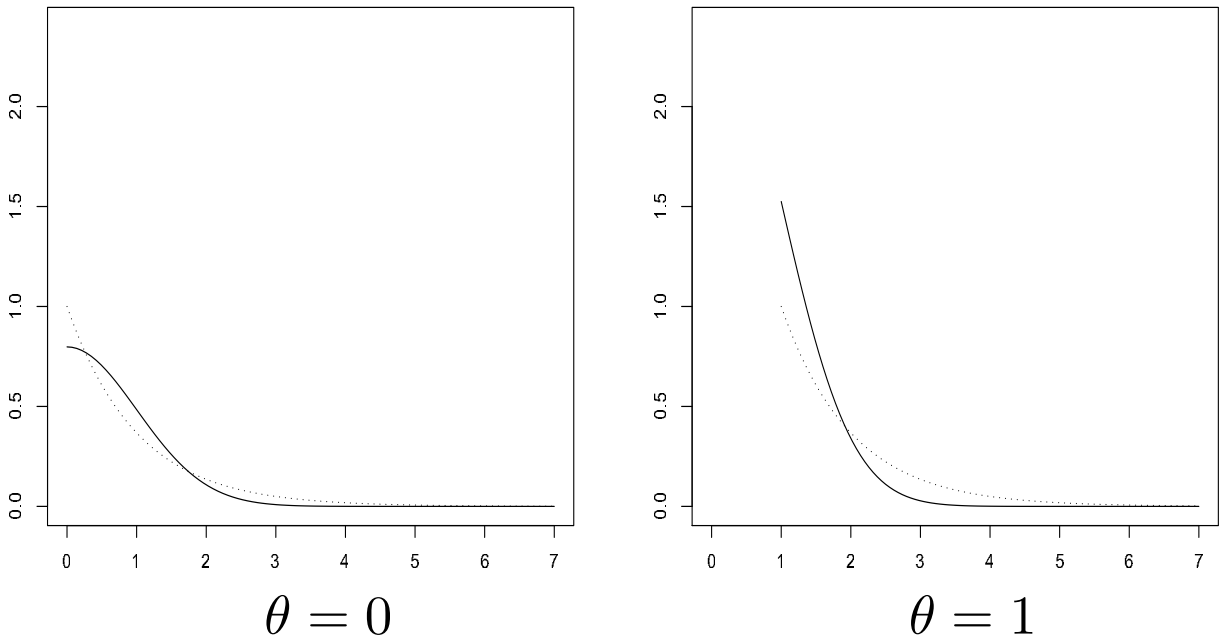
l μια πλήρως ορισμένη συνάρτηση πυκνότητας πιθανότητας, και L η αθροιστική συνάρτηση κατανομής της l.

Ειδικές περιπτώσεις: $\Theta = \mathbb{R}$, $\mathcal{F}$ με αρχική την $N(0,1)$ και $\mathcal{G}$ η Laplace με μέση τιμή 0 και διακύμανση 2.

$$l_1(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right), x \in R$$

και

$$l_2(x) = \frac{1}{2} \exp\left(- \mid x \mid\right), x \in R.$$



Σχήμα 1. Συναρτήσεις πυκνότητας πιθανότητας της αποκομμένης Laplace (διακεκκομένη) και κανονικής κατανομή (μαύρη γραμμή).

Ποσοστά επιλογής παραλλαγής Vuong σε 10000 προσομοιωμένα δείγματα. F =κανονική, G =Laplace.

		w η συν/ση π.π. της $N(0,1)$					
Αληθής		$\theta = 0$			$\theta = 1$		
	n	F	G	FG	F	G	FG
F	50	43.09	0.00	55.91	81.82	0.00	18.18
	100	63.26	0.00	36.74	97.08	0.00	2.92
	200	88.08	0.00	11.92	99.96	0.00	0.04
	300	96.52	0.00	3.48	100.00	0.00	0.00
G	50	0.70	12.26	87.04	0.06	37.73	62.21
	100	0.10	26.76	73.14	0.00	71.12	28.88
	200	0.00	54.30	45.70	0.00	95.65	4.35
	300	0.00	73.74	26.26	0.00	99.47	0.53

FG: τα δυο μοντέλα ισοδύναμα.

- $\mathcal{F}$ η οικογένεια των κανονικών κατανομών με μέση τιμή 0 και διακύμανση $\theta \in \Theta = (0, \infty)$, με συνάρτηση πυκνότητας πιθανότητας

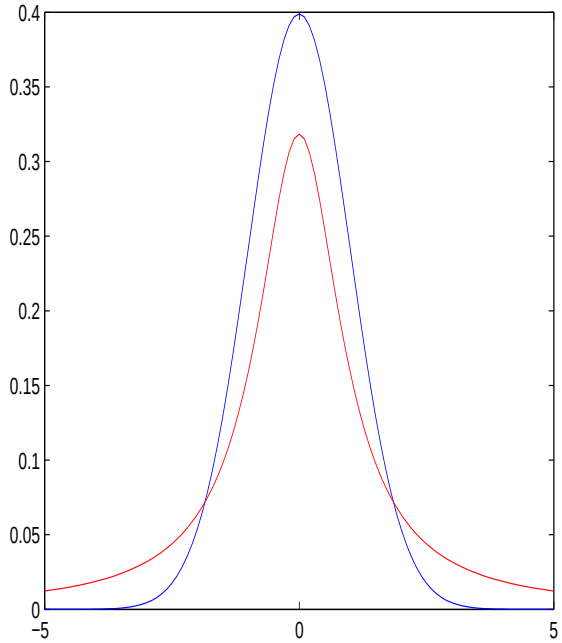
$$f(x;\theta) = \frac{1}{\sqrt{2\pi\theta}} \exp\left(-\frac{x^2}{2\theta}\right), x \in R.$$

- $\mathcal{G}$ είναι η οικογένεια των Cauchy κατανομών με παράμετρο θέσης 0 και κλίμακας $\gamma \in \Gamma = (0, \infty)$, με συνάρτηση πυκνότητας πιθανότητας

$$g(x;\gamma) = \frac{1}{\pi\gamma}\left[1 + \left(\frac{x}{\gamma}\right)^2\right]^{-1}, x \in R.$$

- $\hat{\theta}_{ML} = \frac{1}{n}\sum_i X_i^2$ δεν συγκλίνει όταν ισχύει η $\mathcal{G}$, που αποτελεί απαραίτητη συνθήκη για την εφαρμογή των κλασικών μεθοδολογιών.

- $w(t) = \exp(-|t|)$.



Σχήμα 2. Συναρτήσεις πυκνότητας πιθανότητας της κανονικής κατανομής με $\theta = 1$ (μπλε) και της Cauchy με $\gamma = 1$ (κόκκινο).

Ποσοστά επιλογής της παραλλαγής Vuong σε 10000 προσομοιωμένα δείγματα. F =κανονική $\theta = 1$, G =Cauchy $\gamma = 1$.

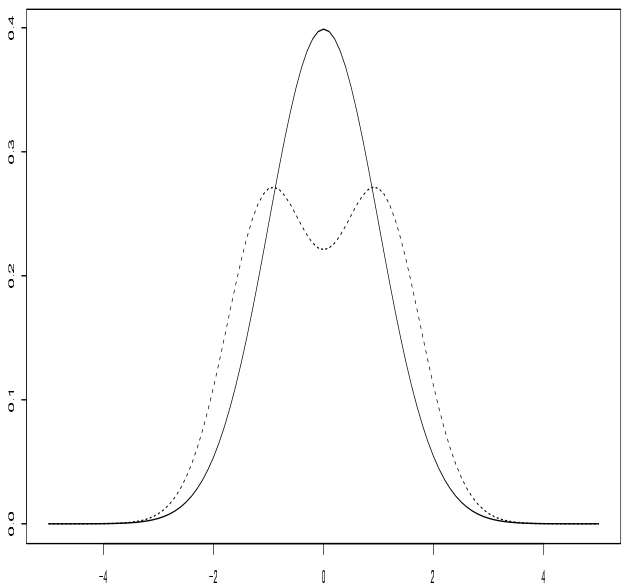
	Δειγματοληψία=F			Δειγματοληψία=G		
n	F	G	FG	F	G	FG
50	72.28	0.00	27.72	0.06	33.97	65.97
100	93.70	0.00	6.30	0.00	64.59	35.41
200	99.78	0.00	0.22	0.00	92.94	7.06

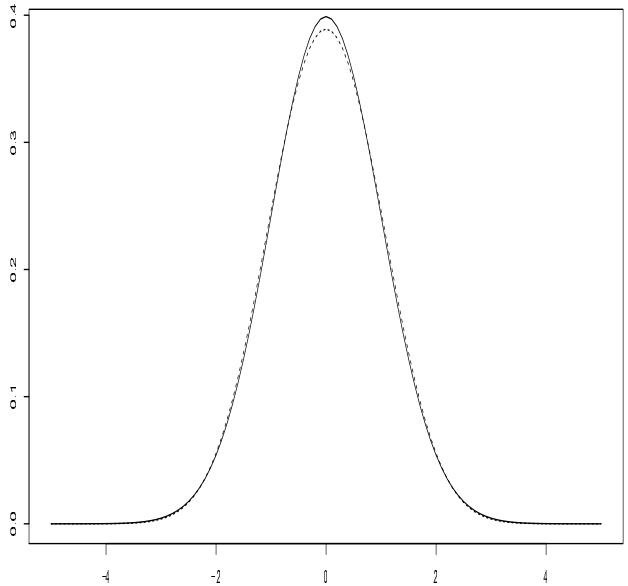
- $\mathcal{F}$ η οικογένεια των κανονικών κατανομών με μέση τιμή 0 και διακύμανση $\theta \in \Theta = (0, \infty)$.

- $\mathcal{G}$ είναι η οικογένεια $0.5N(a, \gamma) + 0.5N(-a, \gamma)$, με a γνωστή τιμή και $\gamma \in \Gamma = (0, \infty)$.

- Κατανομή μίξης: είναι η κατανομή μιας τυχαίας μεταβλητής που προκύπτει από μια συλλογή τυχαίων μεταβλητών ως εξής: αρχικά μια τυχαία μεταβλητή επιλέγεται τυχαία σύμφωνα με δοθείσες πιθανότητες και έπειτα καταγράφεται η τιμή της τυχαίας μεταβλητής που έχει εκλεγεί.

- Η μίξη είναι κάτι τελείως διαφορετικό από το γραμμικό συνδυασμό τυχαίων μεταβλητών. Έτσι ο γραμμικός συνδυασμός κανονικών κατανομών με διαφορετικές μέσες τιμές εξακολουθεί να ακολουθεί κανονική κατανομή, ενώ η μίξη θα έχει δύο κορυφές όταν οι μέσες τιμές είναι μακριά και επομένως δεν είναι κανονική κατανομή.





1ο γράφημα: συν/ση πυκνότητας πιθανότητας της τυπικής κανονικής (μαύρη γραμμή) και της πλησιέστερης μίξης (διακεκομμένη) για $a = 1$.
2ο γράφημα: συν/ση πυκνότητας πιθανότητας της τυπικής κανονικής (μαύρη γραμμή) και της πλησιέστερης μίξης (διακεκομμένη) για $a = 0.5$.

Ποσοστά επιλογής σε 10000 δείγματα ($n = 100$). $F = N(0,1)$, G =μίξη.

		$a=1$			$a=0.5$		
		F	G	FG	F	G	FG
(α)	ML	62.55	0.00	37.45	0.30	6.95	92.75
	$CF1$	67.40	0.00	32.60	0.20	5.00	94.80
	$CF2$	72.05	0.00	27.95	0.50	1.30	98.20
	$CF3$	69.80	0.00	30.20	0.80	1.20	98.00
(β)	ML	3.20	1.25	95.55	0.15	13.60	86.25
	$CF1$	1.90	0.05	98.05	7.35	0.05	92.60
	$CF2$	1.30	2.10	96.60	4.00	0.15	95.85
	$CF3$	1.15	3.25	95.60	3.70	0.20	96.10
(γ)	ML	0.05	14.45	85.50	0.20	13.30	86.50
	$CF1$	0.00	20.20	79.80	0.10	8.55	91.35
	$CF2$	0.20	14.50	85.30	0.30	2.70	97.00
	$CF3$	0.25	13.50	86.25	0.65	2.35	97.00

Δειγματοληψία: (α) $N(0,1)$, (β) $N(0,2)$, (γ) $0.5N(1,1) + 0.5N(-1,1)$.

$w(t)$ η συνάρτηση πυκνότητας πιθανότητας της $CF1 : N(0,1)$ $CF2 : N(0,4)$, $CF3 : N(0,9)$.

- Επέκταση σε περισσότερα των δύο μοντέλα μπορεί να επιτευχθεί με τεχνικές πολλαπλών συγκρίσεων (βλέπε Shimodaira 1998).
- Καθώς ο υπολογισμός των ISE εκτιμητών μπορεί να είναι χρονοβόρος, μπορούν να χρησιμοποιηθούν άλλοι συνεπείς εκτιμητές (αλλάζουν κάποια ασυμπτωτικά αποτελέσματα).
- Διακριτά δεδομένα: Έλεγχος καλής προσαρμογής, επιλογής μοντέλου και ξεχωριστών οικογενειών: προσέγγιση με πιθανογεννήτρια αντί χαρακτηριστική συνάρτηση
  M.D. Jiménez-Gamero and A. Batsidis (2017). Minimum distance estimators for count data based on the probability generating function with applications. Metrika, In Press.
- Επέκταση σε εξαρτημένα δεδομένα. Επέκταση σε μοντέλα παλινδρόμησης, ημιπαραμετρικά.

- Οι αλγόριθμοι πρόβλεψης χρησιμοποιούνται συχνά στην πράξη και διαδραματίζουν καθοριστικό ρόλο στη λήψη αποφάσεων για το μέλλον. Συνηθέστερα το αποτέλεσμα ενός προβλεπτικού αλγορίθμου είναι μια μεταβλητή.
- Ο έλεγχος της εγκυρότητας ενός αλγορίθμου περιλαμβάνει μεταξύ άλλων τη σύγκριση των προβλεπόμενων τιμών με τις πραγματικές που μας γίνονται γνωστές.
- Παρότι αυτή η σύγκριση είναι σημαντική στα περισσότερα βιβλία και συγγράμματα δε δίνεται ξεκάθαρη απάντηση στο ερώτημα πως διενεργείται στατιστικά και πολλοί φοιτήτες ή/και χρήστες της Στατιστικής χρησιμοποιούν λανθασμένες μεθοδολογίες (π.χ. $t-$ τεστ εξαρτημένων δειγμάτων, Kolmogorov-Smirnov έλεγχος ισότητας δύο πληθυσμών, Wilcoxon signed rank test).
- Μέσω ενός πραγματικού συνόλου δεδομένων θα παρουσιαστεί η μέθοδος που μπορεί να χρησιμοποιηθεί και ειδικότερα θα τονιστούν οι προυποθέσεις εφαρμογής της.

- Ο Οργανισμός Οικονομικής Συνεργασίας και Ανάπτυξης (Ο.Ο.Σ.Α.) αποτελείται από 35 μέλη-κράτη. Δύο φορές το χρόνο εκδίδει το O-ECD Economic Outlook στο οποίο μεταξύ άλλων παρουσιάζονται προβλέψεις για πληθώρα οικονομικών δεικτών των μελών κρατών.
- Στο παράδειγμα αυτό θα αξιολογήσουμε την εγκυρότητα του προβλεπτικού αλγορίθμου του ΟΟΣΑ ως προς το ρυθμό ανάπτυξης του ΑΕΠ (σε ποσοστό) των μελών κρατών του για το 2015
- Η σύγκριση των αληθινών τιμών θα γίνει με τις προβλέψεις του Νοεμβρίου 2014 (Economic Outlook No 96) και του Νοεμβρίου 2013 (Economic Outlook No 94).

Πίνακας: Ρυθμός ανάπτυξης του ΑΕΠ σε ποσοστό (%) για κάθε μέλος το 2015 με τις αντίστοιχες προβλέψεις του 2013 και 2014 του ΟΟΣΑ.

Χώρα	Προβλέψεις 2013 για 2015	Προβλέψεις 2014 για 2015	Πραγματικό 2015	Χώρα	Προβλέψεις για 2015	Προβλέψεις 2014 για 2015	Πραγματικό 2015
Australia	3.053	2.475	2.431	Korea	3.991	3.756	2.612
Austria	2.219	0.895	0.822	Latvia	*	3.216	2.738
Belgium	1.532	1.355	1.501	Luxembourg	2.325	2.220	3.536
Canada	2.598	2.551	1.078	Mexico	4.174	3.900	2.470
Chile	4.916	3.236	2.318	Netherlands	0.867	1.439	1.951
Czech Republic	2.275	2.347	4.536	New Zealand	2.851	2.971	2.978
Denmark	1.898	1.361	1.605	Norway	3.145	1.765	1.611
Estonia	3.951	2.409	1.538	Poland	3.328	2.993	3.941
Finland	1.940	0.933	0.210	Portugal	1.113	1.319	1.596
France	1.571	0.764	1.209	Slovak Republic	2.922	2.835	3.831
Germany	1.962	1.079	1.486	Slovenia	0.629	1.362	2.317
Greece	1.791	2.252	-0.312	Spain	0.982	1.742	3.205
Hungary	1.653	2.118	3.148	Sweden	3.038	2.817	3.865
Iceland	2.768	3.287	4.170	Switzerland	2.700	1.543	0.782
Ireland	2.163	3.300	26.286	Turkey	4.142	3.170	3.972
Israel	3.455	3.059	2.514	United Kingdom	2.489	2.685	2.222
Italy	1.425	0.220	0.615	United States	3.377	3.068	2.596
Japan	0.962	0.839	0.571				

Δεδομένου ότι η Λετονία επικύρωσε την αίτηση ένταξη της το 2016, δεν υπήρχε πρόβλεψη για το 2015 στο 2013 (για αυτό το *).

- Κεντρική ιδέα: ένας έγκυρος προβλεπτικός αλγόριθμος δημιουργεί προβλέψεις που μπορούν να θεωρηθούν αναμενόμενες τιμές των πραγματικών τιμών. Αυτό έχει ως συνέπεια να μην πέφτουν ακριβώς στη διαγώνιο των 45 μοιρών αλλά να βρίσκονται διάσπαρτα τυχαία γύρω από αυτή. Όσο πιο κοντινά είναι σε αυτήν τη διαγώνιο τόσο πιο έγκυρος ο αλγόριθμος.
- Επομένως μπορεί να υιοθετήσουμε το μοντέλο της απλής γραμμικής παλινδρόμησης

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, ..., n$$

και να ελέγξουμε την υπόθεση

$$H_0 : \beta_0 = 0 \ \& \ \beta_1 = 1 \quad \text{έναντι της} \quad H_1 : \text{δεν ισχύει η } H_0.$$

- Αυτή η υπόθεση, υπό κάποιες υποθέσεις, μπορεί να ελεγχθεί με το F τεστ για τον έλεγχο γραμμικών υποθέσεων.

Υλοποίηση:

I. Υπολόγισε το $SSE_0 = \sum_{i=1}^{n}(y_i - x_i)^2$.

II. Προσάρμοσε το μοντέλο $y_i = \beta_0 + \beta_1 y_i + \epsilon_i$, $i = 1, ..., n$. Έστω $\hat{\beta}_0$ και $\hat{\beta}_1$, οι εκτιμητές ελαχίστων τετραγώνων. Υπολόγισε το

$$SSE_1 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

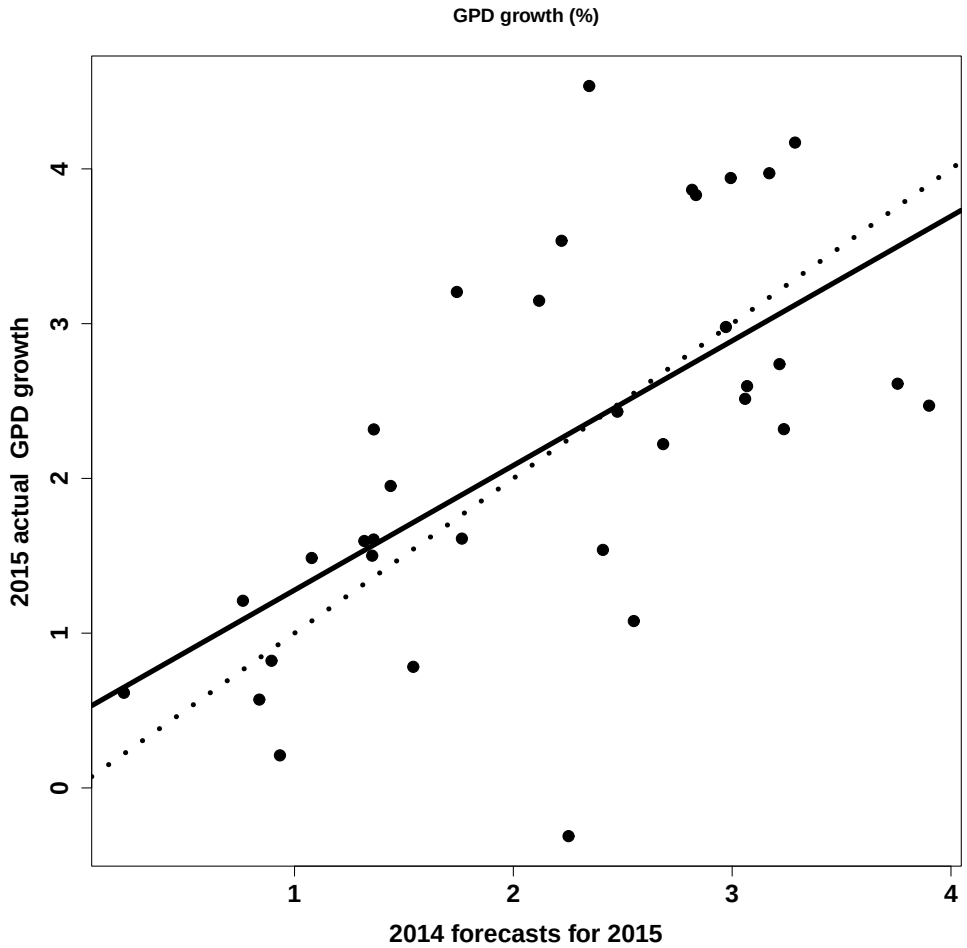
III. Υπολόγισε το

$$F_{obs} = \frac{(SSE_0 - SSE_1)/2}{SSE_1/(n-2)}.$$

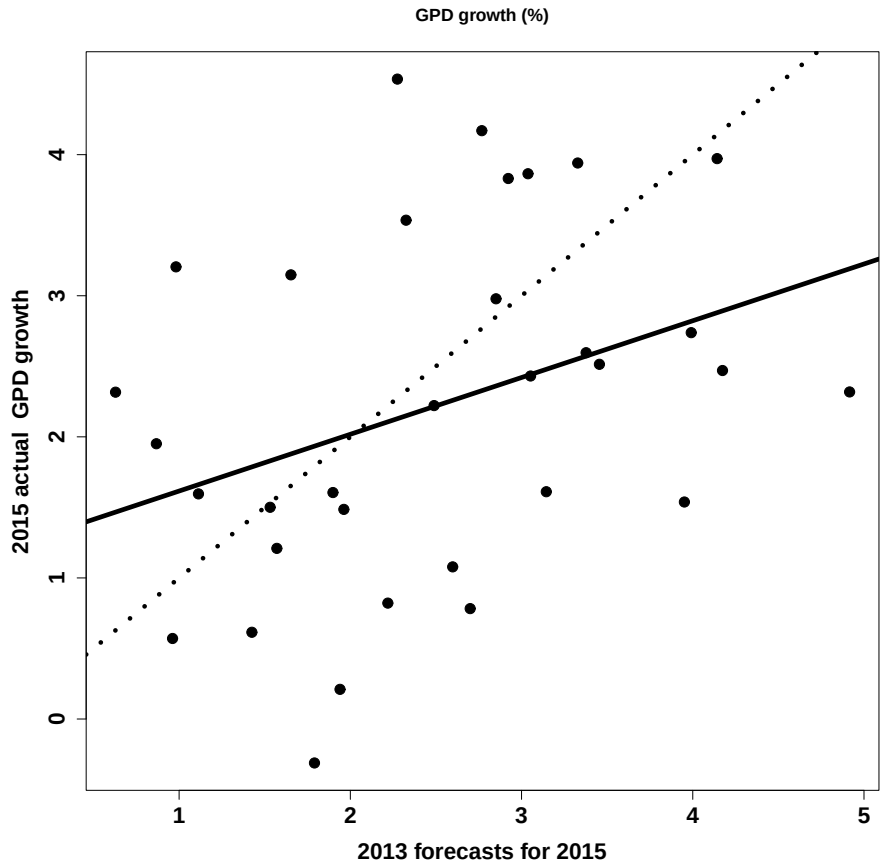
IV. Απορρίπτεται η H_0 αν $F_{obs} \geq F_{2,n-2,\alpha}$ ή ισοδύναμα αν η p-τιμή$=1 - F_{2,n-2}(F_{obs})$ είναι μικρότερη του α, όπου $F_{2,n-2}$ είναι η αθροιστική συνάρτηση κατανομής της F κατανομής με παραμέτρους 2 και $n-2$.

Προυποθέσεις:

- **Υπόθεση 1.** Οι τιμές της μεταβλητής του άξονα $X-$ είναι γνωστές χωρίς σφάλματα.
- **Υπόθεση 2.** Τα σφάλματα $\epsilon_i, i = 1, ..., n$ ασυσχέτιστα, από κανονική κατανομή με μέση τιμή 0 και σταθερή διακύμανση σ^2.
- Οι υποθέσεις θεωρούνται από κάποιους συγγραφείς περιοριστικές (βλέπε για παράδειγμα Mitchell, 1997; Mayer et al., 1994; Harrison, 1990).
- Η Υπόθεση 1 δεν παραβιάζεται αν ως X τιμές θεωρηθούν οι προβλέψεις.
- Προσομοίωση: το F-τεστ είναι ευαίσθητο σε αποκλίσεις από τη σταθερή διακύμανση, ενώ η ευαισθησία αυτή μειώνεται όσο αυξάνει το μέγεθος του δείγματος.
- Προσομοίωση: η απόκλιση από την κανονικότητα δε δημιουργεί πρόβλημα στην περίπτωση συμμετρικών δεδομένων, ενώ σε κάθε άλλη περίπτωση υπάρχει πρόβλημα.



F=0.642, p-τιμή=0.533, $\hat{y} = 0.4729 + 0.8054x$



F=5.583 p-τιμή=0.008, $\hat{y} = 1.2134 + 0.4024x$. Σε αυτήν την περίπτωση ο αλγόριθμος δεν μπορεί να θεωρηθεί έγκυρος.

Κύρια Βιβλιογραφία 1ου μέρους:

M. Broniatowski, A. Keziou (2009) Parametric estimation and testing through divergences and the duality technique, JMVA, 100, 16–36.

D.R. Cox (1961) Tests of separate families of hypothesis, in: Proceedings of the forth Berkeley symposium in mathematical statistics and probability, 105–123.

D.R. Cox (1962) Further results on tests of separate families of hypothesis, Journal of the Royal Statistical Society B, 24, 406–424.

S. Csörgő (1981) The Empirical Characteristic Process When Parameters Are Estimated. In *Contributions to Probability*, pp. 708–723. Academic Press.

Granger, C.W.J., King, M.L., White, H. (1995). Comments on testing economic theories and the use of model selection criteria. J. Econometrics 67, 173-187.

C.R. Heathcote (1977) The integrated squared error estimation of parameters, Biometrika 64, 255–64.

M.D. Jiménez-Gamero, A. Batsidis, V. Alba-Fernández (2016) Fourrier methods for model selection, Ann. Inst. Statis. Math., 68, 105-133.

H. Shimodaira (1998) An application of multiple comparison techniques to model selection, Ann. Inst. Statist. Math. 50, 1–13.

H. White (1982) Regularity conditions for Cox's test of non-nested hypothesis, Journal of Econometrics 19, 301–315.

Q.H. Vuong (1989) Likelihood ratio tests for model selection and non-nested hypotheses, Econometrica 57, 257–306.

Κύρια Βιβλιογραφία 2ου μέρους:

Mayer, D. G., Stuart, M. A. and Swain, A. J. (1994). Regression of real-world data on model output: an appropriate overall test of validity. Agric. Systems, 45, 93-104.

Reynolds M. R., Burkhart H. E. and Daniels R. F. (1981) Procedures for statistical validation of stochastic simulation models. Forest Science 27, 349–364.

Mitchell P.L. (1997) Misuse of regression for empirical validation of models, Agr. Syst. 54, 313–326.

Harrison, S. R. (1990). Regression of a model on real-system output: an invalid test of model validity. Agric. Systems, 35, 183-90.

Thornton, P. K. and Hansen, J.W. (1996). A Note on Regressing Real-world Data on Model Output, Agricultural Sysrems 50 (1996) 411-414.