Εφαρμοσμένη Στατιστική

Δημήτριος Μπάγκαβος

Τμήμα Μαθηματικών και Εφαρμοσμένων Μαθηματικών
Πανεπιστήμιο Κρήτης

16 Φεβρουαρίου 2018

Εισαγωγή.

- Έστω ότι $X_1, X_2, \ldots, X_n$ είναι ένα τυχαίο δείγμα παρατηρήσεων μιας τυχαίας μεταβλητής X.
- Μας ενδιαφέρει να ελέγξουμε κατά πόσον οι τιμές αυτές αποτελούν ένδειξη ότι η τυχαία μεταβλητή X ακολουθεί μία συγκεκριμένη κατανομή, π.χ. μία ομοιόμορφη κατανομή στο διάστημα μεταξύ 0 και 1 ή μεταξύ 0 και 10, ή μία κανονική κατανομή με μέση τιμή 20 και τυπική απόκλιση 2.7.
- Εναλλακτικά, θα μπορούσαμε να ενδιαφερόμαστε να εξετάσουμε ένα πολύ γενικότερο ερώτημα όπως: «είναι εύλογο να υποθέσουμε ότι τα δεδομένα αυτά έχουν προέλθει από μια κανονική κατανομή με άγνωστη μέση τιμή και διασπορά;»
- Παρόλο που ο έλεγχος δείγματος μπορεί να αφορά οποιαδήποτε κατανομή, στην πράξη πιο σύνηθες είναι ο έλεγχος για κανονική κατανομή.
- Στη συνέχεια θα δούμε γραφικούς ελέγχους κανονικότητας καθώς και τις βασικές αρχές ελέγχου στατιστικών υποθέσεων οι οποίες χρησιμοποιούνται:
  - Για έλεγχο ότι η τ.μ. X προέρχεται από συγκεκριμένη κατανομή,
  - Γενικότερα, για έλεγχο οποιασδήποτε στατιστικής υπόθεσης θέλουμε να επαληθεύσουμε μέσω των διαθέσιμων δεδομένων.

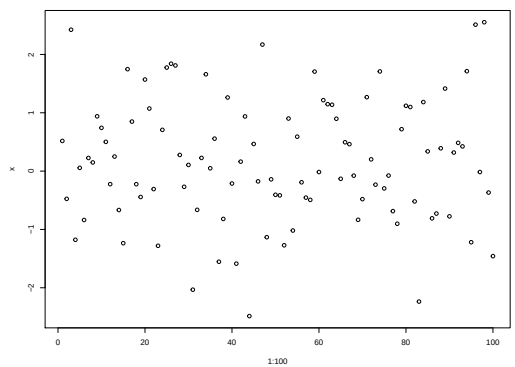
Γραφικές μέθοδοι επαλήθευσης τυχαιότητας.

- Πρακτικά, πριν από κάθε έλεγχο έχει νόημα να εξετάσουμε αν το διαθέσιμο δείγμα $X_1, X_2, \ldots, X_n$ είναι τυχαίο (αν είναι προϊόν κάποιας συστηματικής διαδικασίας δεν έχει νόημα να κοιτάμε αν ταιριάζει σε κάποια κατανομή).
- Ο έλεγχος τυχαιότητας γίνεται με μια απλή γραφική παράσταση της μορφής $(i, X_i), i = 1, 2, \ldots, n$.
- Αν στην γραφική παράσταση δεν διακρίνεται κάποιο σχήμα (κάποια πατέντα, κάποια συστηματική τάση των δεδομένων) τότε μπορούμε να υποθέσουμε ότι το δείγμα μας είναι τυχαίο οπότε έχει νόημα να ψάξουμε να βρούμε κάποια κατανομή.

Για παράδειγμα, για ένα δείγμα από την τυπική κανονική 100 παρατηρήσεων:

```
x<-rnorm(100)
plot(1:100, x)
```

η γραφική παράσταση (δίπλα εικόνα) δείχνει ότι η κατανομή των σημείων στο χώρο του γραφήματος είναι εντελώς τυχαία.



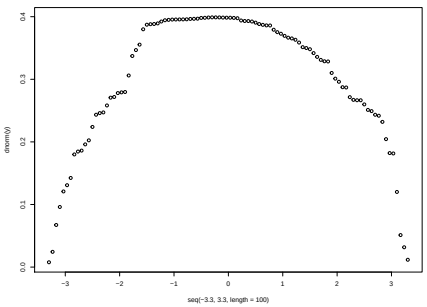
- Ο συγκεκριμένος τρόπος είναι ένας απλός – πρακτικός – τρόπος επαλήθευσης τυχαιότητας.

Γραφική μέθοδος επαλήθευσης συγκεκριμένης κατανομής.

- Ένας πρώτος έλεγχος για έλεγχο προέλευσης δείγματος $X_1, X_2, \ldots, X_n$ από οποιαδήποτε κατανομή είναι να δούμε αν η γραφική παράσταση $(i, X_{(i)}), i = 1, 2, \ldots, n$ ταιριάζει στο σχήμα της κατανομής.
- Οι παρατηρήσεις $X_{(i)}, i = 1, 2, \ldots, n$ είναι το διατεταγμένο κατ' αύξουσα σειρά δείγμα $X_1, X_2, \ldots, X_n$.

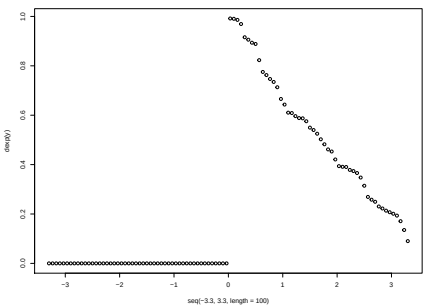
Για παράδειγμα, για ένα δείγμα από την τυπική κανονική 100 παρατηρήσεων, η γραφική

```
x<-rnorm(100)
y<-sort(x) #sort sample in ascending order
plot(seq(-3.3, 3.3, length=100), dnorm(y))
```



παράσταση (δίπλα) δείχνει ότι η κατανομή των σημείων μοιάζει με αυτή της κανονικής.
Αν χρησιμοποιήσουμε λάθος κατανομή, π.χ.

```
x<-rnorm(100)
y<-sort(x) #sort sample in ascending order
plot(seq(-3.3, 3.3, length=100), dexp(y))
```



το αποτέλεσμα δεν θα ταιριάζει σε κάποιο γνωστό σχήμα κατανομής.

- Δοκιμάστε τα παραπάνω με μέγεθος δείγματος 1000. Τι παρατηρείτε;

Γραφικοί έλεγχοι q-q plot και p-p plot.

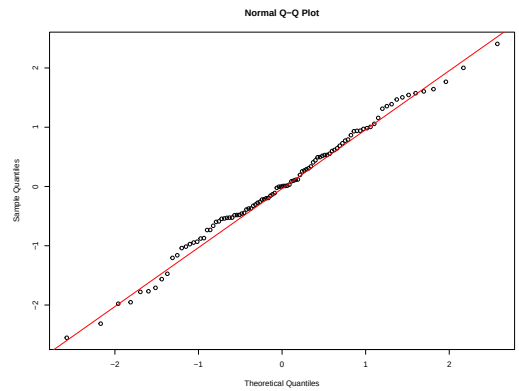
- Ο προηγούμενος τρόπος παρόλο που χρησιμοποιείται συχνά στην πράξη (γιατί είναι πολύ εύκολο να υλοποιηθεί):
  - Υποθέτει ότι ξέρουμε τις παραμέτρους της κατανομής που ελέγχουμε ή τις έχουμε εκτιμήσει (π.χ. με μέγιστη πιθανοφάνεια). Είναι ισοδύναμος τρόπος με το ιστόγραμμα, μόνο που για το ιστόγραμμα δεν χρειάζεται να ξέρουμε τις παραμέτρους της κατανομής.
  - Δεν έχει καμία επιστημονική βάση - δεν στηρίζεται σε κάποια αποδεδειγμένη μέθοδο (θεώρημα ή ιδιότητα): μπορεί απλά να τύχει να ταιριάζει το δείγμα με την συγκεκριμένη κατανομή.
- Μία μέθοδος που προσφέρει σχετική ασφάλεια για έλεγχο δείγματος είναι το λεγόμενο q-q (quantile quantile) plot το οποίο απεικονίζει τα ποσοστημόρια (quantiles) του δείγματος (υπολογισμένα εμπειρικά) σε σχέση με τα ποσοστημόρια (υπολογισμένα θεωρητικά) της επιλεγμένης κατανομής, σαν σημεία στο επίπεδο (x, y).
- Ίδιας λογικής είναι και το p-p (probability probability) plot το οποίο απεικονίζει την εμπειρική α.σ.κ. του δείγματος σε σχέση με την θεωρητική α.σ.κ.

Κατασκευή q-q plot.

- Για να κατασκευάσουμε το q-q plot για δεδομένο δείγμα $X_1, X_2, \ldots, X_n$
  1. Υπολογίζουμε τα εμπειρικά ποσοστιαία σημεία (σελίδα 39 διαφάνειες πρώτης εβδομάδας) - τα σημεία αυτά είναι οι τεταγμένες (y) του γραφήματος.
  2. Υπολογίζουμε τα ποσοστιαία σημεία της κατανομής που θεωρούμε ότι προέρχεται το δείγμα. Προκύπτουν από τη λύση της $F(\xi_p; \theta) = p$ ως προς ξ_p - αυτά τα σημεία είναι οι τετμημένες (x) του γραφήματος.
  3. Το ζητούμενο γράφημα προκύπτει από τα ζεύγη $(x = \xi_p, y)$. Το δύσκολο κομμάτι είναι η λύση της $F(\xi_p; \theta) = p$ η οποία απαιτεί αντιστροφή της F κάτι το οποίο δεν είναι πάντα εφικτό αναλυτικά.

Για δείγμα 100 παρατηρήσεων από $N(0, 1)$,

```
x<-rnorm(100)
qqnorm(y,  main = "Normal Q-Q Plot",  xlab = "
    Theoretical Quantiles",
          ylab = "Sample Quantiles",  plot.it =
    TRUE)
qqline(x, distribution = function(p) qnorm(p),
    prob = c(0.1, 0.6), col = 2)
```



Για σύγκριση δείγματος με άλλες κατανομές προτιμάται η χρήση της `qqPlot` του πακέτου `car`

Κατασκευή p-p plot.

- Για να κατασκευάσουμε το p-p plot για δεδομένο δείγμα $X_1, X_2, \ldots, X_n$ σχηματίζουμε τα ζεύγη $(\hat{F}(x), F(x;\theta))$ όπου

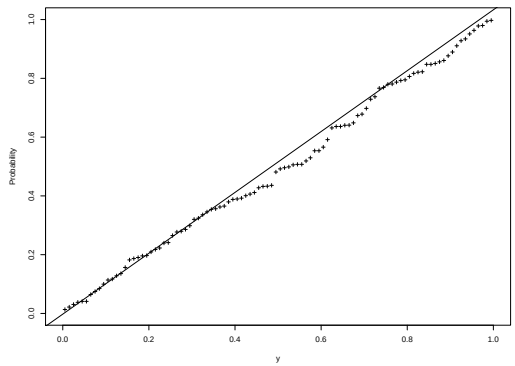
$$\hat{F}(x) = n^{-1} \sum_{i=1}^{n} 1_{\{X_i \leq x\}} \tag{1}$$

  είναι η εμπειρική α.σ.κ (παρατηρείστε ότι δεν χρειάζεται να κάνουμε καμία υπόθεση για να την υπολογίσουμε)
- και $F(x;\theta)$ είναι η υποτιθέμενη θεωρητική α.σ.κ.

Για δείγμα 100 παρατηρήσεων από $N(0,1)$,

```
ppplot.das(rnorm(100), pdist = pnorm, xlab = NULL,
     ylab = "Probability", line = TRUE,
                    lwd = 2, pch = 3, cex = 0.7)
```

όπου η εντολή `ppplot.das` είναι διαθέσιμη από το πακέτο `StatDA`



- Από το γράφημα βλέπουμε ότι η σχηματιζόμενη γραφική παράσταση είναι σχεδόν (πολύ κοντά) ευθεία το οποίο είναι ένδειξη ότι το δείγμα προέρχεται από την κατανομή που προσδιορίσαμε.
- Και πάλι όμως, τόσο το q-q plot όσο και το p-p plot δεν είναι αποδείξεις ότι το δείγμα προέρχεται από την συγκεκριμένη κατανομή.

Βασικές αρχές ελέγχου στατιστικών υποθέσεων.

- Ο λόγος είναι ότι απλά μπορεί να τύχει το συγκεκριμένο δείγμα να ικανοποιεί τη συνθήκη που ελέγχουμε (δηλ. μπορεί να τύχει να φαίνονται ίσα ενώ να μην είναι πραγματικά).
- Για να είμαστε σε θέση να ελέγξουμε υποθέσεις για την άγνωστη κατανομή πιθανότητας μιας τ.μ. με βάση ένα δείγμα χρειαζόμαστε μιας διαδικασία που θα αναφέρεται σε όλα ποσοστιαία σημεία / πιθανότητες και όχι μόνο σε αυτά που διαλέξαμε για τις γρ. παραστάσεις των q-q plot και p-p plot.
- Με άλλα λόγια, χρειαζόμαστε διαδικασίες σχεδιασμένες να απαντήσουν στο ερώτημα: αποτελούν οι παρατηρήσεις μας ένα δείγμα από κάποια συγκεκριμένη κατανομή;
- Αυτό απαντάται χρησιμοποιώντας κάποια ικανή και αναγκαία συνθήκη (η οποία απορρέει από κάποια χαρακτηριστική ιδιότητα της κατανομής)
- Το δείγμα χρησιμοποιείται για την επιβεβαίωση της ικανής και αναγκαίας συνθήκης.
- Η ικανή και αναγκαία συνθήκη συνήθως μεταφράζεται σε κάποια στατιστική συνάρτηση: τη συνάρτηση ελέγχου της υπόθεσης μας.

Βασικές αρχές ελέγχου στατιστικών υποθέσεων.

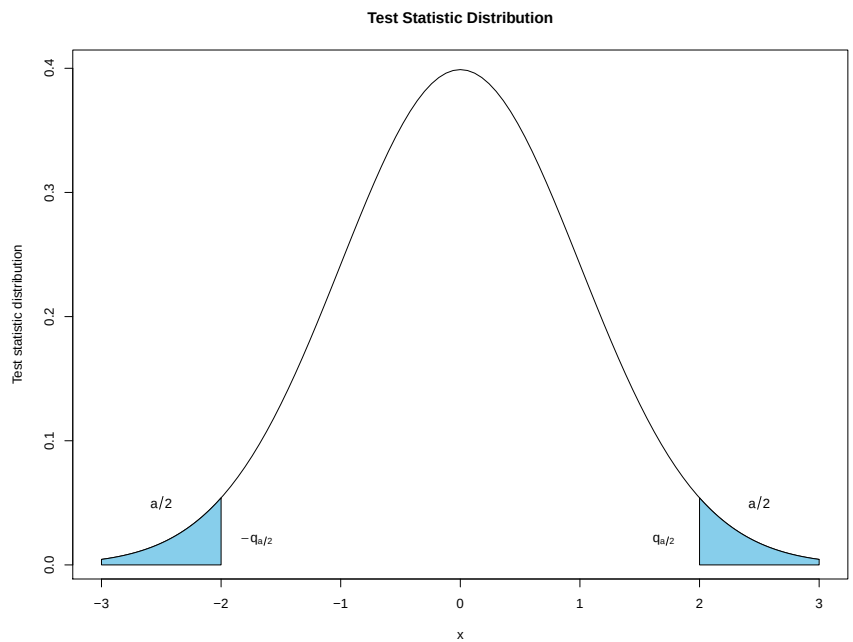
- Αυτές οι διαδικασίες είναι μέρος γενικότερου κομματιού της στατιστικής που λέγεται έλεγχος υποθέσεων και έχουν τα εξής χαρακτηριστικά:
  1. Σχηματίζουμε τη μηδενική υπόθεση, (H_0), η οποία εκφράζει την εικασία που θέλουμε να ελέγξουμε: π.χ. για να ελέγξουμε ότι το δείγμα μας προέρχεται από την τυπική κανονική κατανομή θα είχαμε $H_0 : F(x) = F(x; 0, 1)$ όπου $F(x; 0, 1)$ η α.σ.κ της $N(0, 1)$.
  2. Σχηματίζουμε την εναλλακτική υπόθεση, (H_1) η οποία εκφράζει την περίπτωση που η H_0 δεν ισχύει.
     - Η H_1 μπορεί να περιέχει όλες της εναλλακτικές καταστάσεις της H_0, π.χ.: $H_1 : F(x) \neq F(x; 0, 1)$
     - η μόνο κάποιες συγκεκριμένες, π.χ. $H_1 : F(x) = F(x; 1, 2)$
  3. Σχηματίζουμε μία στατιστική συνάρτηση, βάση κάποιας χαρακτηριστικής ιδιότητας της υπόθεσης που θέλουμε να ελέγξουμε.
  4. Όταν ισχύει η H_0, θα επαληθεύεται η ιδιότητα και βάση αυτού μπορούμε να συμπεράνουμε με πιο πολύ σιγουριά (και πάλι όχι 100%) αν ισχύει η υπόθεση μας.
- Είναι σημαντικό να τονίσουμε εδώ ότι και πάλι δεν θα έχουμε 100% σιγουριά στο αποτέλεσμα. Οι στατιστικοί έλεγχου υποθέσεων βασίζονται στο λεγόμενο επίπεδο σημαντικότητας βάση του οποίου καθορίζεται η αποδοχή ή απόρριψη της υπόθεσης που ελέγχουμε.

Επίπεδο σημαντικότητας.

- Για να έχει νόημα ο στατιστικός έλεγχος υποθέσεων που διεξάγουμε θα πρέπει το στατιστικό τεστ, όταν επαναλαμβάνεται με τυχαία δείγματα κάτω από τις ίδιες συνθήκες να αποτυπώνει την πραγματικότητα (απόρριψη ή αποδοχή της H_0, αναλόγως με την πραγματικότητα) ένα `μεγάλο ποσοστό φορών'.
- Αυτό το ποσοστό φορών είναι το επίπεδο σημαντικότητας και καθορίζεται από την πιθανότητα η στατιστική συνάρτηση ελέγχου του τεστ να πάρει τιμές έτσι ώστε το τεστ να επαληθεύεται πολλές φόρες.

Στη γενική περίπτωση, αν η κατανομή της στατιστικής συνάρτησης $T(\mathbf{X})$ – υπό την H_0 – έχει ποσοστιαία σημεία $-q_{\frac{\alpha}{2}}, q_{\frac{\alpha}{2}}$ (δηλαδή στο διάστημα $(-q_{\frac{\alpha}{2}}, q_{\frac{\alpha}{2}})$ περικλείεται το $100(1-\alpha)\%$ των τιμών της $T(\mathbf{X})$) τότε ισχύει

$$1-\alpha = P(-q_{\frac{\alpha}{2}} < T(\mathbf{X}) < q_{\frac{\alpha}{2}})$$



Κρίσιμη περιοχή.

- Όπως φάνηκε και από το σχήμα, οι τιμές $-q_{\frac{\alpha}{2}}, q_{\frac{\alpha}{2}}$ καθορίζουν την κρίσιμη περιοχή του τεστ (οι περιοχές με το μπλε χρώμα).
- Όταν η στατιστική συνάρτηση παίρνει τιμές εκεί, τότε η H_0 απορρίπτεται.
- Αντίθετα, το διάστημα $(-q_{\frac{\alpha}{2}}, q_{\frac{\alpha}{2}})$ καθορίζει την περιοχή αποδοχής - όταν η $T(\mathbf{X})$ πάρει τιμές σε αυτό το διάστημα τότε η H_0 γίνεται αποδεκτή.
- Από τα παραπάνω φαίνεται ότι το α είναι το ποσοστό των απορρίψεων της H_0 (δεδομένου ότι αυτή είναι αληθής) που είμαστε διατεθειμένοι να ανεχθούμε.
- Η στατιστική συνάρτηση $T(\mathbf{X})$ είναι τ.μ. με γνωστή κατανομή όταν η H_0 είναι αληθής.
- Οι H_0, H_1 διατυπώνονται με βάση τις στατιστικές παραμέτρους του μοντέλου που υιοθετείται για την ανάλυση των δεδομένων και μπορεί να είναι απλές, σύνθετες, μονόπλευρες η δίπλευρες
- Η μορφή της H_1 καθορίζει της θέση της κρίσιμης περιοχής στον άξονα των x.

Κρίσιμη περιοχή.

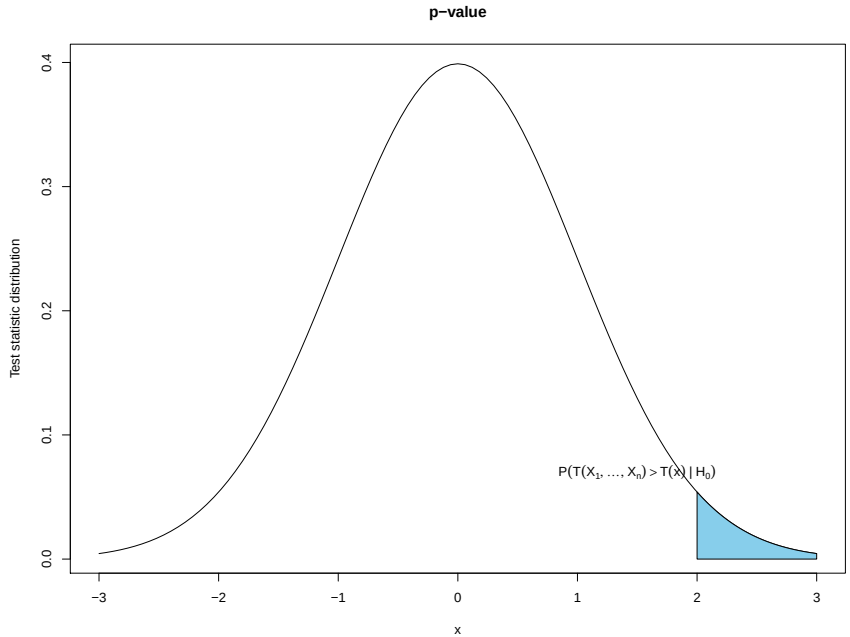
- Αν είναι δεξιόπλευρη (π.χ. $H_0 : \theta = \theta_0, H_1 : \theta > \theta_0$) τότε η κρίσιμη περιοχή είναι της μορφής $[q_\alpha, +\infty)$.
- Αν είναι αριστερόπλευρη (π.χ. $H_0 : \theta = \theta_0, H_1 : \theta < \theta_0$) τότε η κρίσιμη περιοχή είναι της μορφής $(-\infty, -q_\alpha]$.
- Αν είναι διπλευρη (π.χ. $H_0 : \theta = \theta_0, H_1 : \theta \neq \theta_0$) τότε η κρίσιμη περιοχή είναι της μορφής $(-\infty, -q_{\frac{\alpha}{2}}] \cup [q_{\frac{\alpha}{2}}, +\infty)$.
- Τα ποσοστιαία σημεία $q_\alpha, q_{\alpha/2}$ κ.λ.π. υπολογίζονται από την κατανομή της $T(\mathbf{X})$ όταν αληθεύει η H_0
- Δεν υπάρχει θεωρία που να μας εξασφαλίζει 100% αν απορρίπτουμε σωστά η λάθος μια στατιστική υπόθεση.
- Όταν απορρίπτουμε λανθασμένα την H_0 τότε το σφάλμα που κάνουμε έχει πιθανότητα μικρότερη ή ίση του α το οποίο το καθορίζουμε εμείς και έχει μικρή τιμή (1%, 5%, 10%).
- Για να κάνουμε σωστά έναν έλεγχο τα βασικά βήματα είναι να καθορίσουμε σωστά την περιοχή απόρριψης / αποδοχής και να υπολογίσουμε σωστά την τιμή της στατιστικής συνάρτησης.

Τύποι σφαλμάτων κατά τον στατιστικών υποθέσεων.

- Στον στατιστικό έλεγχο υποθέσεων μπορεί να συμβούν τα εξής δύο ειδών σφάλματα:
  1. Σφάλμα τύπου I: Απόρριψη της μηδενικής υπόθεσης ενώ αυτή είναι σωστή:
  $$\alpha = P(\text{ απόρριψης της } H_0 | H_0 \text{ αληθής })$$
  2. Σφάλμα τύπου II: Αποδοχή της μηδενικής υπόθεσης H_0 ενώ αυτή είναι λανθασμένη:
  $$\beta = P(\text{ αποδοχή της } H_0 | H_0 \text{ είναι λάθος})$$
- Ισχύς του στατιστικού τεστ είναι η πιθανότητα απόρριψης της H_0 ενώ αυτή είναι λάθος,
$$\gamma = 1 - \beta = P(\text{ απόρριψη της } H_0 | H_0 \text{ είναι λάθος})$$
- Το επίπεδο σημαντικότητας ενός ελέγχου είναι η πιθανότητα να παρατηρηθεί τιμή του στατιστικού μεγαλύτερη από αυτή που έδωσε το δείγμα των παρατηρήσεων
$$\begin{aligned} p - \text{value} &= P(T(\mathbf{X}) \text{ πιο ακραία από την } T(x) | H_0) \\ &= P(T(\mathbf{X}) > T(x) | H_0) = 1 - F_{T|H_0}(x). \end{aligned}$$

Απόρριψη και αποδοχή στατιστικών υποθέσεων.

Η στατιστική συνάρτηση είναι τυχαία μεταβλητή και ως τυχαία μεταβλητή έχει κάποια κατανομή. Η πιθανότητα του $p-$value είναι η πιθανότητα η στατιστική συνάρτηση $T(X_1,\ldots,X_n)$ να λάβει μια ακραία τιμή, όταν η H_0 είναι αληθινή.



- Διαισθητικά, αν το $p-$value είναι «κοντά» στο 0 τότε συμπεραίνουμε ότι είναι «απίθανο», δεδομένης της H_0, να εμφανιστεί το συγκεκριμένο δείγμα $X_1, X_2, \ldots, X_n$, και όπως είναι φυσικό φτάνουμε στο συμπέρασμα ότι μάλλον δεν πρέπει να ισχύει η H_0.
- Για «μικρό» λοιπόν $p-$value είναι λογικό να απορρίπτουμε την H_0.

Ορισμός

Η τιμή $p-$value είναι το μικρότερο επίπεδο σημαντικότητας, α, στο οποίο η αρχική υπόθεση μπορεί να απορριφθεί.

Απόρριψη και αποδοχή στατιστικών υποθέσεων.

Προσοχή:

Θα πρέπει να τονισθεί ότι η p – value δεν είναι η πιθανότητα ότι η H_0 είναι σωστή.

- Στην κλασική στατιστική θεωρία δεν υπάρχει τρόπος να προσδιορίσουμε την πιθανότητα να είναι σωστή η μηδενική υπόθεση.
- Αυτό γιατί σε κάθε έλεγχο υπόθεσης, η H_0 θα είναι πάντα ή σωστή ή λάθος.

Ερμηνεία του p – value

Αυτό που παρέχει η τιμή p – value είναι η πιθανότητα να παρατηρηθεί τιμή / δείγμα πιο ακραίο σε σχέση με αυτό που έχουμε διαθέσιμο

Έλεγχος κανονικότητας με το τεστ του Kolmogorov

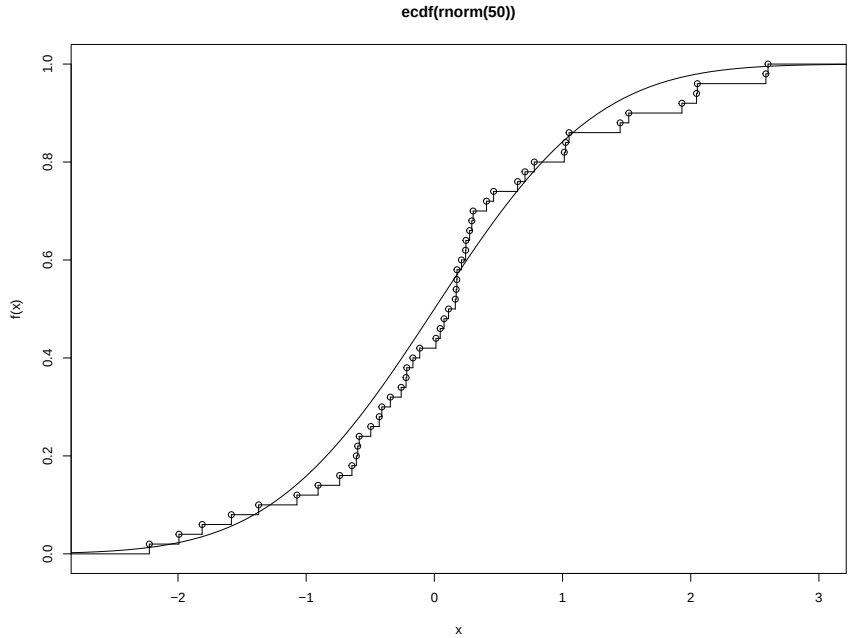
- Το τεστ του Kolmogorov βασίζεται στη σύγκριση της αθροιστικής συνάρτησης κατανομής $F(x)$ με την αντίστοιχη εμπειρική συνάρτηση κατανομής που ορίστηκε στην (1).
- Εάν η διαφορά των 2 συναρτήσεων είναι μεγάλη, αυτό είναι ένδειξη για να απορρίψουμε την μηδενική υπόθεση και να συμπεράνουμε ότι η πραγματική, αλλά άγνωστη συνάρτηση κατανομής $F(x)$ δεν ορίζεται από την μηδενική υπόθεση.
- Εννοείται πως εφόσον μας ενδιαφέρει να ελέγξουμε αν το δείγμα προέρχεται από κανονική κατανομή, ως $F(x) = \Phi(x)$ θεωρούμε την α.σ.κ. της $N(0, 1)$.
- Το ερώτημα που προκύπτει για να ορίσουμε πλήρως το τεστ είναι πως θα καθορίσουμε αν η απόσταση είναι μεγάλη η μικρή.
- Αυτό που βολεύει είναι το κατώτερο άνω φράγμα (sup) οπότε ο ορισμός του τεστ είναι:

$$D_n = \sup_x |F(x) - \hat{F}(x)| \tag{2}$$

- Σημαντική παρατήρηση: το συγκεκριμένο τεστ υποθέτει ότι καμία παράμετρος της κατανομής δεν εκτιμάται από τα δεδομένα, αντίθετα τις θεωρεί δεδομένες.

Έλεγχος κανονικότητας με το τεστ του Kolmogorov

- Το τεστ βασίζεται στη μέγιστη διαφορά μεταξύ θεωρητικής και εμπειρικής α.σ.κ υπολογισμένη σε κάθε σημείο x του πεδίου ορισμού της κατανομής που ελέγχουμε.
- Προφανώς, μεγάλη διαφορά των 2 κατανομών σημαίνει απόκλιση από την κατανομή που ελέγχουμε.



- Ο έλεγχος $H_0 : F(x) = \Phi(x)$ έναντι της $H_1 : F(x) \neq \Phi(x)$, για δείγμα $X_1, \ldots, X_n$ στην R γίνεται με την:

```
as<-ks.test(rnorm(50), "pnorm")
as
```

```
One-sample Kolmogorov-Smirnov test

data:  rnorm(50)
D = 0.14407, p-value = 0.2275
alternative hypothesis: two-sided
```

Ερμηνεία: Η τιμή D που επιστρέφει η R είναι η τιμή της στατιστικής συνάρτησης (2) για το συγκεκριμένο δείγμα.
Βάσει της τιμής **p-value** πρέπει να απορρίψουμε η να δεχθούμε την H_0;

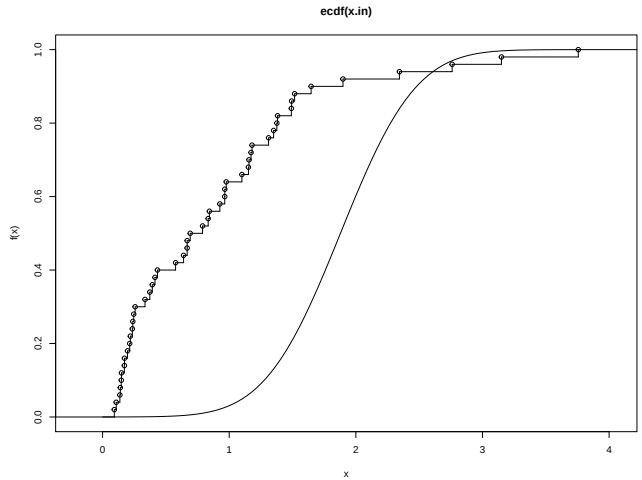
Γενικότερη χρήση του Kolmogorov τεστ.

- Τι γίνεται όταν το δείγμα δεν είναι όντως από την κατανομή που ελέγχουμε;
- Π.χ. για δείγμα 50 παρατηρήσεων από την εκθετική κατανομή,

```
x.in <- rexp(50)
as<-ks.test(x.in, "pnorm")
as
```

```
One-sample Kolmogorov-Smirnov test

data:  x.in
D = 0.53684, p-value = 5.396e-14
alternative hypothesis: two-sided
```



- Εδώ η τιμή `p-value` είναι ουσιαστικά 0, το οποίο δείχνει ότι το δείγμα όντως δεν προέρχεται από την κανονική.
- Το ίδιο φαίνεται και από τη γρ. παράσταση όπου ξεκάθαρα η απόσταση της θεωρητικής α.σ.κ της $N(0,1)$ και της $\hat{F}(x)$ (βλέπε σχέση (1)) είναι πολύ μεγάλη.

Άσκηση: Χρησιμοποιείστε το τεστ του Kolmogorov για να ελέγξετε την υπόθεση (H_0) ότι τα δεδομένα προέρχονται από την εκθετική κατανομή

Γενικότερη χρήση του Kolmogorov τεστ.

- Εξ΄ ορισμού, το τεστ του Kolmogorov μπορεί να χρησιμοποιηθεί για έλεγχο προέλευσης δείγματος από οποιαδήποτε κατανομή αρκεί αυτή να είναι συνεχής.
- Παρατηρείστε ότι προκειμένου να εφαρμόσουμε το τεστ δεν χρειάζεται να υποθέσουμε κάτι άλλο για τα διαθέσιμα δεδομένα,
- Αυτό σημαίνει ότι το τεστ είναι πολύ γενικό, άρα και ευρέως διαδεδομένο στη χρήση του.
- Τι γίνεται όμως όταν οι παράμετροι της κατανομής του δείγματος είναι άγνωστοι και πρέπει να εκτιμηθούν;
- Σε αυτή την περίπτωση, ναι μεν η τιμή του στατιστικού D_n εξακολουθεί να ισχύει, αλλά επειδή έχει αλλάξει η κατανομή του, η `p-value` που υπολογίζουμε δεν έχει καμία πρακτική αξία (δεν μπορεί να χρησιμοποιηθεί για έλεγχο).
- Το επόμενο τεστ (το τεστ του Lilliefors) απευθύνεται σε περιπτώσεις όπου οι παράμετροι της κατανομής που ελέγχουμε εκτιμώνται από τα δεδομένα.

Το τεστ του Lilliefors.

- Έστω δείγμα $X_1, \ldots, X_n$ από άγνωστη κατανομή $F(x)$.
- Για να ελέγξουμε αν το δείγμα προέρχεται από την κανονική κατανομή,
  - εκτιμάμε τη μέση τιμή και διακύμανση της κατανομής του δείγματος από τις σχέσεις

$$\bar{X} = n^{-1} \sum_{i=1}^{n} X_i, \; s = \left\{ (n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \right\}^{\frac{1}{2}}$$

  - Υπολογίζουμε τις 'κανονικοποιημένες' τιμές του δείγματος

$$Z_i = \frac{X_i - \bar{X}}{s}, i = 1, 2, \ldots, n$$

  - Βάση των $Z_i, i = 1, 2, \ldots, n$, υπολογίζουμε την εμπειρική α.σ.κ από την (1), έστω $\hat{F}(x)$ και μετά την

$$L = \sup_x |\Phi(x) - \hat{F}(x)| \tag{3}$$

  που είναι και η στατιστική συνάρτηση του τεστ.

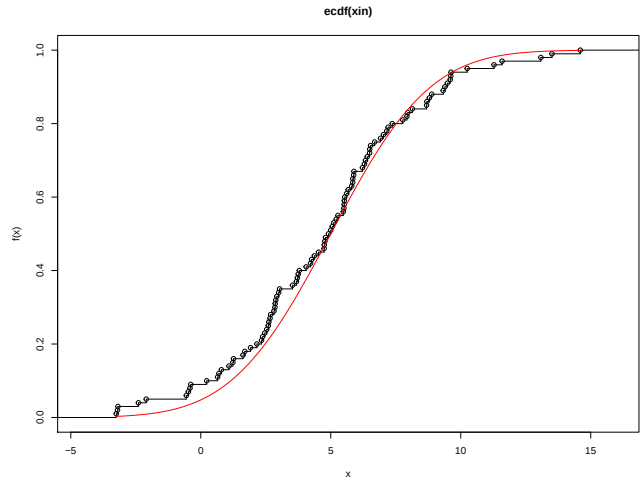
Το τεστ του Lilliefors.

- Για τον υπολογισμό του τεστ υπάρχει η έτοιμη συνάρτηση `lillie.test()` του πακέτου **nortest**.

```
xin<-rnorm(100, mean = 5, sd = 3)
tst<-lillie.test(xin)

Lilliefors (Kolmogorov-Smirnov)
    normality test

data:  xin
D = 0.055925, p-value = 0.6185
```



- Όπως και στην περίπτωση του Kolmogorov τεστ, έτσι και εδώ η μηδενική υπόθεση είναι ότι τα δεδομένα έρχονται από την υποτιθέμενη κατανομή (για το παρόν τεστ αυτή είναι η κανονική) και η εναλλακτική ότι τα δεδομένα δεν προέρχονται από την κανονική.
- Όπως φαίνεται από την **p-value** αλλά και από την γραφική παράσταση της κατανομής των δεδομένων η H_0 γίνεται αποδεκτή.
- Προσοχή: Σε αυτό το τεστ δεν υπάρχει η επιλογή να προσδιορίσουμε άλλη κατανομή ($\neq \Phi(x)$) για την εναλλακτική υπόθεση.

Το τεστ των Shapiro-Wilk.

- Με δεδομένο δείγμα $X_1, \ldots, X_n$ από άγνωστη κατανομή $F(x)$, ένα ακόμα τεστ για έλεγχο κανονικότητας δείγματος είναι το τεστ των Shapiro-Wilk με στατιστική συνάρτηση

$$W = \frac{\left\{\sum_{i=1}^{n} a_i X_{(i)}\right\}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{4}$$

- όπου $X_{(i)}$ είναι η i-στή διατεταγμένη παρατήρηση, δηλ. η i-στή μικρότερη παρατήρηση του δείγματος.

$$(a_1, \ldots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}}$$

όπου $(m_1, \ldots, m_n)$ είναι οι αναμενόμενες τιμές των διατεταγμένων παρατηρήσεων και V είναι ο πίνακας συνδιακύμανσης τους.

Η εντολή `shapiro.test()` κάνει τον πιο πάνω υπολογισμό αυτόματα και επίσης υπολογίζει την αντίστοιχη `p-value`.

```
> shapiro.test(rnorm(100, mean = 5, sd = 3))

    Shapiro-Wilk normality test

data:  rnorm(100, mean = 5, sd = 3)
W = 0.98294, p-value = 0.2236
```

Το τεστ των Anderson-Darling.

- Τα 2 πρώτα τεστ που είδαμε βασίζονται στην απόσταση της δειγματικής κατανομής των δεδομένων και της κατανομής που υποθέτουμε στη μηδενική υπόθεση.
- Συνολικά το εμβαδόν της απόστασης μεταξύ των 2 συναρτήσεων είναι:

$$\int_{-\infty}^{+\infty} (\hat{F}(x) - F(x))^2 \, dF(x)$$

όπου η κάθε παρατήρηση είναι ζυγισμένη με την αντίστοιχη συνάρτηση πυκνότητας της.

- Μπορούμε να ελέγξουμε ακόμα περισσότερο την επιρροή κάθε παρατήρησης στην παραπάνω συνάρτηση ορίζοντας

$$A = n \int_{-\infty}^{+\infty} \frac{(\hat{F}(x) - F(x))^2}{F(x)(1 - F(x))} \, dF(x)$$

δηλαδή το κάθε σημείο έχει βάρος $f(x)(F(x)(1 - F(x))^{-1})$ ώστε παρατηρήσεις στα άκρα να έχουν πιο μεγάλο βάρος.

- Μία πιο χρήσιμη στην πράξη έκδοση του στατιστικού είναι η

$$A^2 = -n - \sum_{i=1}^{n} \frac{2i - 1}{n} \left\{ \ln(F(X_{(i)})) + \ln(1 - F(X_{(n+1-i)})) \right\}$$

Το τεστ των Anderson-Darling.

Όπως και για τα προηγούμενα τεστ, η R έχει έτοιμη συνάρτηση, την `ad.test()` του πακέτου `nortest` που αυτοματοποιεί τους υπολογισμούς:

```
> ad.test(rnorm(100, mean = 5, sd = 3))

        Anderson-Darling normality test

data:  rnorm(100, mean = 5, sd = 3)
A = 0.22351, p-value = 0.8208
```

Παρατήρηση 1.

Η κατανομή της στατιστικής συνάρτησης επηρεάζεται από την κατανομή που ελέγχουμε.

Παρατήρηση 2.

Το τεστ των Anderson-Darling θεωρείται πιο ισχυρό σε σχέση με τα προηγούμενα για την ανίχνευση της πραγματικότητας.

Αν ελέγξουμε π.χ. για εκθετική κατανομή, βλέπουμε ότι τεστ ανιχνεύει (βάση του αποτελέσματος) ότι η προέλευση του δείγματος δεν είναι η κανονική.

```
> ad.test(rexp(100))

        Anderson-Darling normality test

data:  rexp(100)
A = 5.0986, p-value = 1.086e-12
```