Ανάλυση Δεδομένων με χρήση του Στατιστικού Πακέτου R

Δημήτρης Φουσκάκης,
Καθηγητής,
Τομέας Μαθηματικών,
Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών,
Εθνικό Μετσόβιο Πολυτεχνείο.

Περιεχόμενα

Εισαγωγικά

- Όπως είδαμε και στην εισαγωγή, στην Στατιστική συνήθως ενδιαφερόμαστε να εκτιμήσουμε ένα άγνωστο μέγεθος, το οποίο καλείται παράμετρος και το οποίο συνήθως συνοψίζει κατά κάποιον τρόπο τις τιμές της υπό μελέτης μεταβλητής στον πληθυσμό, π.χ. τη μέση της τιμή. Η εκτίμησή μας γίνεται με την βοήθεια κατάλληλα επιλεγμένων δειγματοσυναρτήσεων, συναρτήσεων δηλαδή του δείγματος, οι οποίες καλούνται (σημειακές) εκτιμήτριες. Ο τρόπος επιλογής εκτιμητριών γίνεται είτε (α) με βάση την λογική, π.χ. αν θέλουμε να εκτιμήσουμε την μέση τιμή στον πληθυσμό μας ακούγεται λογικό να χρησιμοποιήσουμε ως εκτιμήτρια την μέση τιμή του δείγματος (plug in principle), είτε (β) με βάση διάφορες ιδιότητες, π.χ. η εκτιμήτρια μας θέλουμε να έχει μέση τιμή ίση με την ποσότητα όπου εκτιμά (αμεροληψία) είτε (γ) με βάση κάποιο κριτήριο κατασκευής (π.χ. εκτιμήτριες μέγιστης πιθανοφάνειας).
- Να υπενθυμίσουμε εδώ ότι οι εκτιμήτριες ως δειγματοσυναρτήσεις είναι τυχαίες μεταβλητές και άρα η ίδια εκτιμήτρια συνήθως παίρνει άλλη τιμή όταν παρατηρούμε άλλα δεδομένα. Συνήθως θέλουμε η εκτιμήτρια μας να έχει μικρή μεταβλητότητα (δηλαδή διασπορά), έτσι ώστε το τυπικό σφάλμα εκτίμησης (η τυπική απόκλιση της εκτιμήτριας) να είναι μικρό, δηλαδή οι τιμές της εκτιμήτριας μας να μην μεταβάλλονται πολύ από δείγμα σε δείγμα.

Εκτιμήτριες Μέγιστης Πιθανοφάνειας

- Το πιο διαδεδομένο κριτήριο κατασκευής εκτιμητριών είναι αυτό της μέγιστης πιθανοφάνειας (maximum likelihood). Πιθανοφάνεια καλείται η από κοινού σ.π.π. ή σ.μ.π. του τυχαίου δείγματος $X_1,...,X_n$. Αν το χαρακτηριστικό που μας ενδιαφέρει προέρχεται από ένα πληθυσμό με σ.π.π ή σ.μ.π. $f(x;\boldsymbol{\theta})$ όπου $\boldsymbol{\theta}$ είναι το διάνυσμα των αγνώστων παραμέτρων, τότε επειδή το τυχαίο δείγμα αποτελείται από ανεξάρτητες και ισόνομες τ.μ. η πιθανοφάνεια θα είναι ίση με

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i;\boldsymbol{\theta})$$

Εκτιμήτριες Μέγιστης Πιθανοφάνειας

- Η παραπάνω συνάρτηση είναι μια συνάρτηση του **θ** και μας δίνει την πιθανότητα το δείγμα μας να προέρχεται από την υποτιθέμενη κατανομή με παράμετρο **θ**. Συνεπώς μπορούμε να διαλέξουμε το **θ** έτσι ώστε να μεγιστοποιείται αυτή η πιθανότητα. Οι εκτιμήτριες που παίρνουμε με τον τρόπο αυτόν καλούνται Εκτιμήτριες Μέγιστης Πιθανοφάνειας (Ε.Μ.Π.) και τις συμβολίζουμε με $\hat{\boldsymbol{\theta}}$.

- Αρκετά συχνά δουλεύουμε με την λογαριθμική πιθανοφάνεια για λόγους ευκολίας υπολογισμών

$$l(\boldsymbol{\theta}) = \log(\mathrm{L}(\boldsymbol{\theta})) = \sum_{i=1}^{n} \log\left(\mathrm{f}(\mathrm{x}_i;\boldsymbol{\theta})\right)$$

Εκτιμήτριες Μέγιστης Πιθανοφάνειας

- Για την μεγιστοποίηση της παραπάνω συνάρτησης συνήθως δουλεύουμε αναλυτικά, παίρνουμε τις μερικές παραγώγους ως προς κάθε συνιστώσα του **θ,** τις εξισώνουμε με το μηδέν και λύνουμε το σύστημα που προκύπτει. Εν συνεχεία με την βοήθεια των δευτέρων μερικών παραγώγων ελέγχουμε αν το σημείο είναι πράγματι σημείο μεγίστου.
- Με την βοήθεια της R μπορούμε επίσης να δούμε το γράφημα της πιθανοφάνειας και να βρούμε με την βοήθειά του το μέγιστο.

Εκτιμήτριες Μέγιστης Πιθανοφάνειας

- Παράδειγμα 1: Έστω ότι ο αριθμός Χ των πετρελαιοφόρων που φθάνουν κάθε μέρα σε ένα λιμάνι είναι τ.μ. με κατανομή Poisson(λ), όπου λ άγνωστο. Τα παρακάτω δεδομένα αποτελούν τις παρατηρήσεις από τυχαίο δείγμα 20 ημερών

  9 4 5 5 7 13 8 3 6 5 4 5 10 5 5 4 3 3 4 7

Εκτιμήτριες Μέγιστης Πιθανοφάνειας

- Τότε

$$l(\lambda) = \log(\mathrm{L}(\lambda)) = \sum_{i=1}^{n} \log\left(e^{-\lambda}\lambda^{x_i} / x_i!\right) = -n\lambda + \sum_{i=1}^{n} x_i \log\lambda - \log\prod_{i=1}^{n} x_i!$$

- Εύκολα προκύπτει ότι η Ε.Μ.Π. του λ τότε είναι $\hat{\lambda} = \bar{X}.$
- Για τις συγκεκριμένες παρατηρήσεις η τιμή της εκτιμήτριάς μας είναι

$$\hat{\lambda} = \bar{x} = 5.75.$$

Εκτιμήτριες Μέγιστης Πιθανοφάνειας

- Η επόμενη συνάρτηση στην R υπολογίζει την λογαριθμική Poisson πιθανοφάνεια για συγκεκριμένο δείγμα και για 10000 διαφορετικές τιμές του λ.

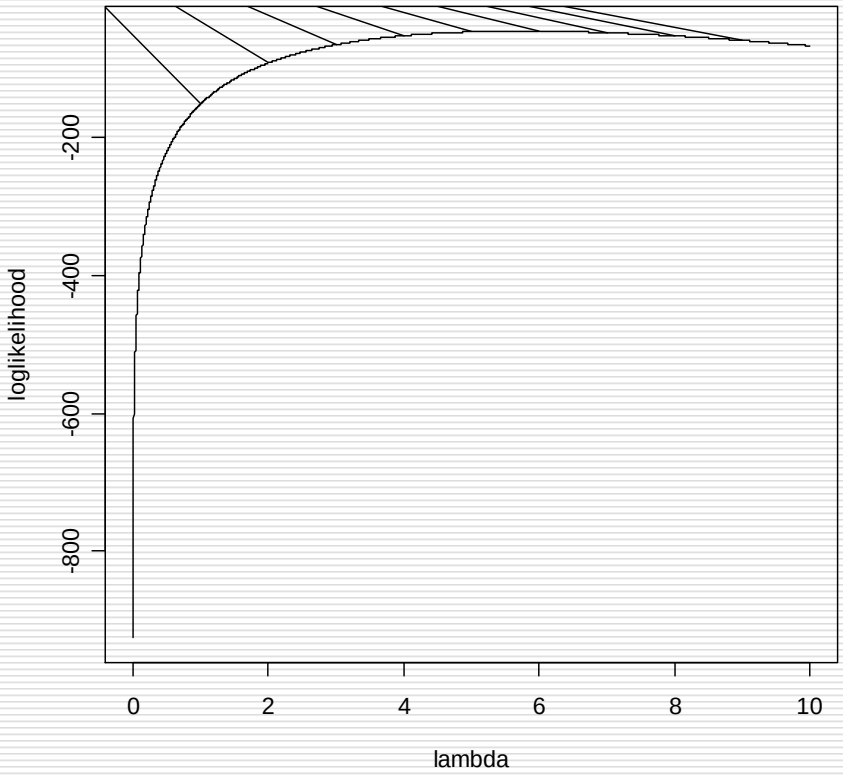
```
> lambda<-seq(0.001, 10, length=10000)
> poisson_loglikelihood<-function(data, lambda){
  results<-rep(NA,10000)
  for(i in 1:10000){
    results[i]<-sum(dpois(data,lambda[i],log=T))
  }
  return(results)
  }
```

Εκτιμήτριες Μέγιστης Πιθανοφάνειας

- Μπορούμε λοιπόν να κάνουμε ένα γράφημα της εν λόγω συνάρτησης και να δούμε εμπειρικά που μεγιστοποιείται.

```
> x<-c(9,4,5,5,7,13,8,3,6,5,4,5,10,5,5,4,3,3,4,7)
> results<-poisson_loglikelihood(x, lambda)
> plot(lambda, results, xlab="lambda", ylab="loglikelihood",
type="l")
```

Εκτιμήτριες Μέγιστης Πιθανοφάνειας



Εκτιμήτριες Μέγιστης Πιθανοφάνειας

- Για να βρούμε εμπειρικά το μέγιστο γράφουμε την εντολή

```
> lambda[order(results)[10000]]
[1] 5.75
```

- Η εντολή order(results)[10000] μας επιστρέφει τη θέση που υπάρχει η μεγαλύτερη *l*(λ), και άρα η εντολή lambda[order(results)[10000]] μας επιστρέφει πράγματι το λ που μεγιστοποιεί την *l*(λ).
- Η παραπάνω τιμή δεν σημαίνει κατ' ανάγκη ότι είναι το μέγιστο, καθώς έχουμε πάρει μόνο ένα αριθμό σημείων λ. Στην περίπτωσή μας είναι πράγματι το μέγιστο αφού

```
> mean(x)
[1] 5.75
```

Εκτιμήτριες Μέγιστης Πιθανοφάνειας

- Παράδειγμα 2: Ο χρόνος Χ (σε λεπτά) αναμονής στην ουρά για να εξυπηρετηθείτε από το ταμείο μιας τράπεζας έστω ότι είναι τ.μ. ομοιόμορφα κατανεμημένη στο διάστημα (0,θ), με θ άγνωστη παράμετρο. Έστω ότι έχουμε τις παρακάτω παρατηρήσεις 12 χρόνων αναμονής

5 0 2 10 6 4 3 8 9 7 8 4

Εκτιμήτριες Μέγιστης Πιθανοφάνειας

- Η σ.π.π. της ομοιόμορφης κατανομής στο διάστημα (0, θ) είναι:

$$f(x;\theta)=\theta^{-1},\ \ x\in(0,\theta),\ \ \theta>0$$

- Η πιθανοφάνεια τότε είναι

$$L(\theta)=\prod_{i=1}^{n}\theta^{-1}I_{(0,\theta)}(x_i)\quad \text{όπου}$$

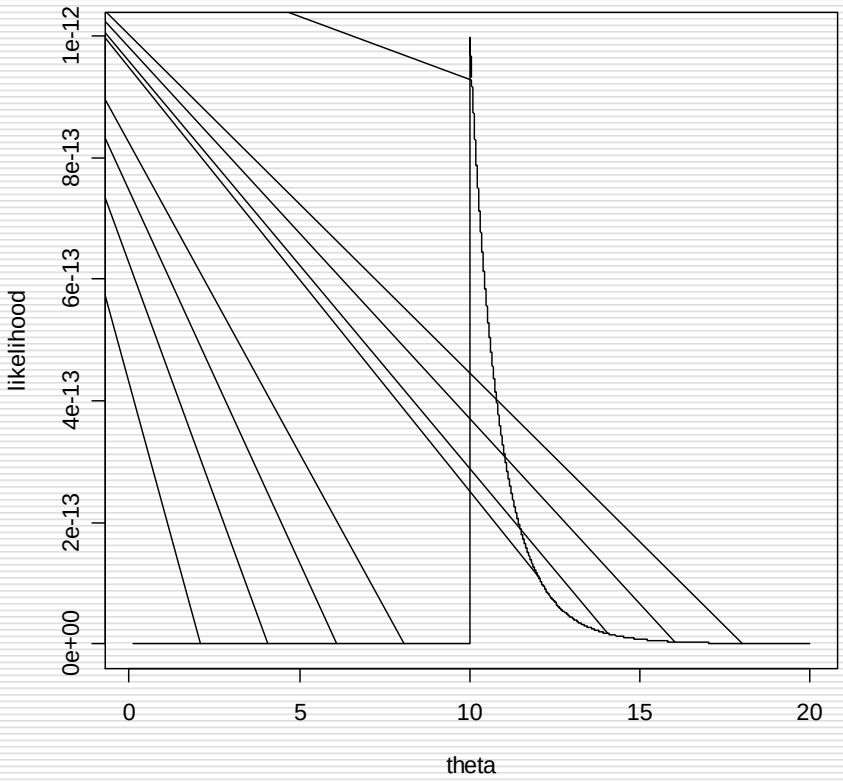
$$I_{(0,\theta)}(x_i)=\begin{cases}1,\ x\in(0,\theta),\\ 0,\ x\notin(0,\theta).\end{cases}\qquad \text{η δείκτρια συνάρτηση του συνόλου } (0,\theta)$$

Εκτιμήτριες Μέγιστης Πιθανοφάνειας

- Ο αναλυτικός τρόπος με την παράγωγο δεν μπορεί να δουλέψει στο εν λόγω παράδειγμα οπότε θα ζητήσουμε την βοήθεια της R για να βρούμε το μέγιστο.

```
> theta<-seq(0.1, 20, length=10000)
> uniform_likelihood<-function(data, theta){
    results<-rep(NA,10000)
    for(i in 1:10000){
        results[i]<-prod(dunif(data,0, theta[i]))
    }
    return(results)
    }
> y<-c(5, 0, 2, 10, 6, 4, 3, 8, 9, 7, 8, 4)
> results<-uniform_likelihood(y, theta)
> plot(theta, results, xlab="theta", ylab="likelihood",
    type="l")
```

Εκτιμήτριες Μέγιστης Πιθανοφάνειας



Εκτιμήτριες Μέγιστης Πιθανοφάνειας

- Η πιθανοφάνεια όπως περιμέναμε είναι 0 όταν $\theta < y_i$, για τουλάχιστον ένα i, ενώ όταν $\theta \geq y_i$ (για όλα τα i = 1,..,n) είναι μια φθίνουσα συνάρτηση του θ. Θέλουμε λοιπόν την μικρότερη τιμή του θ που να ικανοποιεί όμως την ανισότητα $\theta \geq y_i$ (για όλα τα i = 1,..,n), και άρα το μέγιστο (Ε.Μ.Π.) είναι το $\hat{\theta} = \max_{i=1,..n} \{y_i\}$.

Διαστήματα Εμπιστοσύνης

- Οι (σημειακές) εκτιμήσεις δεν μας δίνουν κάποια πληροφορία σχετικά με την ακρίβεια ή το σφάλμα εκτίμησης.
- Είναι λοιπόν χρήσιμο να προσδιορίσουμε, μέσω των εκτιμητριών και των τυπικών τους σφαλμάτων, ένα διάστημα το οποίο θα περιέχει την άγνωστη τιμή της παραμέτρου με καθορισμένη πιθανότητα, έστω γ.
- Σκοπός μας δηλαδή είναι να βρούμε δυο ποσότητες u και v (u<v) έτσι ώστε $P(u \leq \theta \leq v)=\gamma=1-\alpha$.
- Το [u,v] καλείται διάστημα εμπιστοσύνης (Δ.Ε.) με συντελεστή εμπιστοσύνης (σ.ε.) $\gamma=1-\alpha$.
- Το διάστημα εμπιστοσύνης του θ προσδιορίζεται με βάση την κατανομή της εκτιμήτριας του θ από το τυχαίο δείγμα, συνεπώς οι τιμές u και v είναι τυχαίες μεταβλητές. Αυτό σημαίνει ότι από διαφορετικό δείγμα ίδιου μεγέθους ενδέχεται να προκύψουν διαφορετικά Δ.Ε. για το θ.

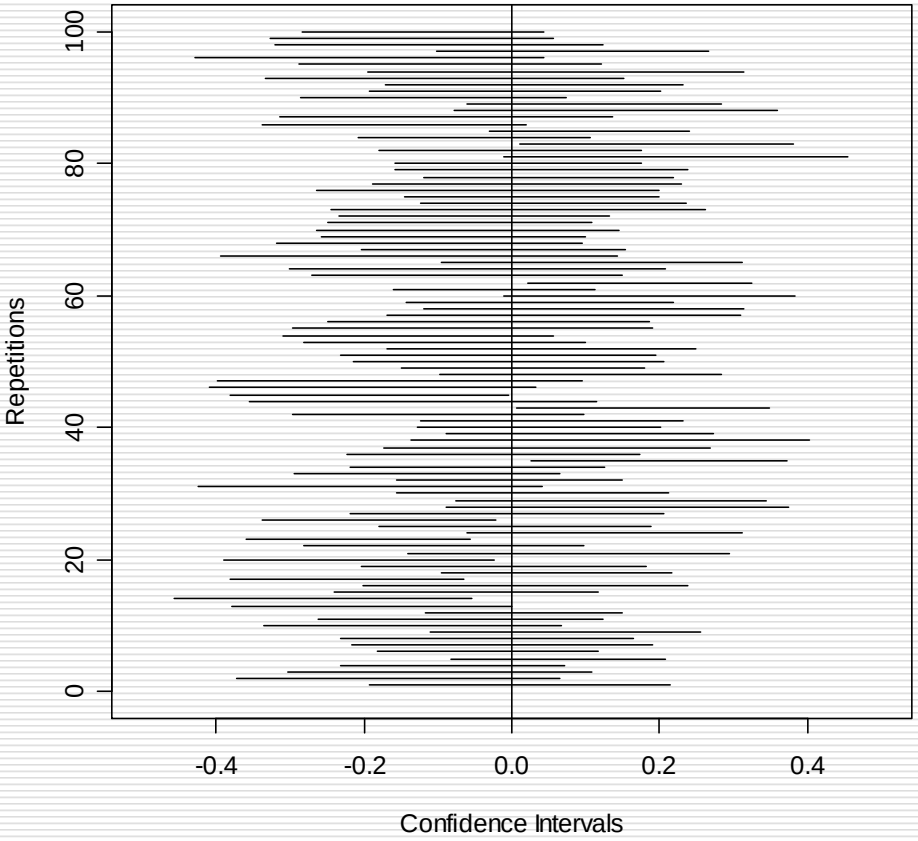
Διαστήματα Εμπιστοσύνης

- Το εύρος του Δ.Ε. εξαρτάται από το τυπικό σφάλμα της εκτιμήτριας του θ και τον συντελεστή εμπιστοσύνης. Όσο μεγαλύτερο είναι το τυπικό σφάλμα της εκτιμήτριας του θ τόσο μεγαλύτερο εύρος έχει το Δ.Ε. Επίσης όσο μεγαλύτερο συντελεστή εμπιστοσύνης έχουμε τόσο μεγαλύτερο εύρος έχει το Δ.Ε.
- Το τυπικό σφάλμα των εκτιμητριών είναι αντιστρόφως ανάλογο του μεγέθους του δείγματος n. Συνεπώς όσο το n αυξάνει τόσο το εύρος του Δ.Ε. θα μειώνεται.

Διαστήματα Εμπιστοσύνης

- Η πραγματική ερμηνεία ενός Δ.Ε. με σ.ε. γ είναι η ακόλουθη. Σε μια σειρά κατασκευών διαστημάτων εμπιστοσύνης μιας παραμέτρου, με ανεξάρτητα δείγματα του αυτού μεγέθους, ένα ποσοστό 100 γ% των διαστημάτων αυτών "αναμένεται" να περιέχουν την αληθή τιμή της παραμέτρου θ.

Διαστήματα Εμπιστοσύνης



Διαστήματα Εμπιστοσύνης

- Παράδειγμα: Έστω $X_1,\ldots,X_n$ τυχαίο δείγμα από πληθυσμό με μέση τιμή μ (άγνωστη) και διασπορά σ^2 (γνωστή). Μια λογική εκτιμήτρια για το μ είναι ο δειγματικός μέσος $\bar{X}$.
- Για μεγάλο n από το Κ.Ο.Θ. ξέρουμε ότι

$$\bar{X} \dot{\sim} N(\mu, \sigma^2/n).$$

- Συνεπώς το τυπικό σφάλμα της εκτιμήτριάς μας είναι $\sigma/\sqrt{n}$.

Διαστήματα Εμπιστοσύνης

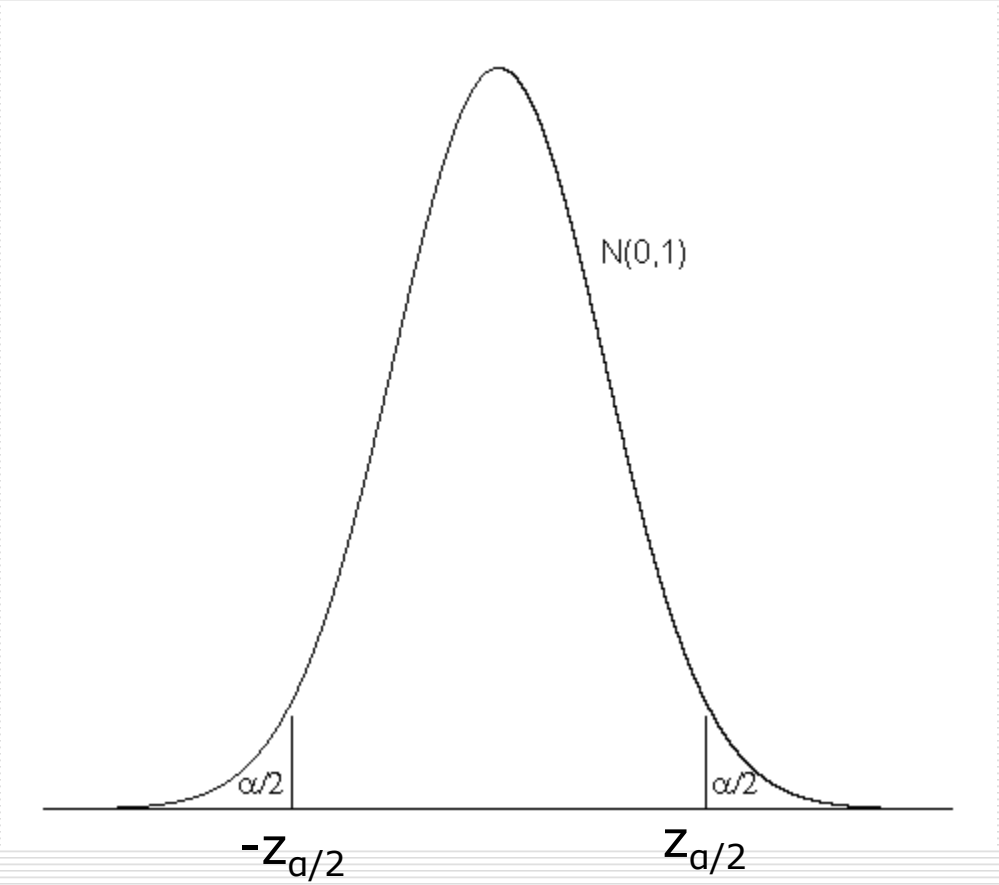
- Τότε όμως η τ.μ.

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

και συνεπώς αν $z_{a/2}$ το σημείο εκείνο της τυποποιημένης Κανονικής κατανομής για το οποίο $P(Z>z_{a/2})=a/2$ έχουμε

$$P(-z_{a/2}<Z<z_{a/2})=\gamma=1-a.$$

Διαστήματα Εμπιστοσύνης



Διαστήματα Εμπιστοσύνης

- Λύνοντας την ανισότητα $-z_{\alpha/2}<Z<z_{\alpha/2}$ ως προς μ προκύπτει το ακόλουθο γ% Δ.Ε. για το μ

$$\left(\bar{X}-z_{\alpha/2}\sigma/\sqrt{n},\ \bar{X}+z_{\alpha/2}\sigma/\sqrt{n}\right).$$

- Η πιθανότητα

$$P\left(\bar{X}-z_{\alpha/2}\sigma/\sqrt{n}<\mu<\bar{X}+z_{\alpha/2}\sigma/\sqrt{n}\right)=\gamma=1-\alpha$$

δεν εκφράζει την πιθανότητα το μ να πάρει τιμές στο εν λόγω διάστημα, διότι το μ είναι μια σταθερά (αν και άγνωστη), αλλά είναι η πιθανότητα το εν λόγω διάστημα να περιέχει την πραγματική τιμή του μ.

Διαστήματα Εμπιστοσύνης

- Αν $\bar{x}$ είναι η τιμή του $\bar{X}$ σε συγκεκριμένο δείγμα τότε θα εκτιμήσουμε το μ με το διάστημα

$$\left(\bar{x} - z_{\alpha/2}\sigma/\sqrt{n},\ \bar{x} + z_{\alpha/2}\sigma/\sqrt{n}\right)$$

το οποίο με πιθανότητα γ περιέχει το άγνωστο μ.

Έλεγχοι Υποθέσεων

- Στατιστική υπόθεση ονομάζεται κάθε υπόθεση που αφορά στην κατανομή μιας τυχαίας μεταβλητής Χ. Συνήθως μια υπόθεση αφορά στην άγνωστη παράμετρο, έστω θ της κατανομής της τυχαίας μεταβλητής Χ.
  - Παράδειγμα 1. Έστω Χ η διάρκεια ζωής συγκεκριμένων λαμπτήρων φθορίου. Ενδιαφερόμαστε να ελέγξουμε αν η μέση τιμή της τυχαίας μεταβλητής Χ, έστω μ, είναι 2000h, δηλαδή να ελέγξουμε αν ισχύει η μηδενική υπόθεση H_0: μ=2000 με εναλλακτική υπόθεση την H_1: μ≠2000. Η απόφασή μας για το αν θα απορρίψουμε ή όχι την μηδενική υπόθεση θα γίνει με βάση τα δεδομένα μας.
  - Λογική Ελέγχου Υποθέσεων: Αθώος (H_0 σωστή) μέχρι αποδείξεως του εναντίου (H_0 λανθασμένη). Άρα ή θα έχουμε αρκετές ενδείξεις από τα δεδομένα για να απορρίψουμε την μηδενική υπόθεση ή δεν θα έχουμε αρκετές ενδείξεις και δεν μπορούμε να την απορρίψουμε.

Έλεγχοι Υποθέσεων

- Ο έλεγχος $H_0 : \theta = \theta_0$

  $H_1 : \theta \neq \theta_0$

  καλείται αμφίπλευρος.

- Οι έλεγχοι

  $H_0 : \theta = \theta_0$ $\quad$ $H_0 : \theta = \theta_0$

  $H_1 : \theta > \theta_0$ $\quad$ $H_1 : \theta < \theta_0$

  καλούνται μονόπλευροι.

Έλεγχοι Υποθέσεων

- Για τον έλεγχο μιας (στατιστικής) υπόθεσης υπολογίζουμε το στατιστικό ελέγχου (test statistic). Συνήθως το στατιστικό ελέγχου είναι της μορφής

$$\frac{(\text{Εκτιμήτρια του } \theta) - (\text{Τιμή του } \theta \text{ με βάση την } H_0)}{\text{Τυπικό Σφάλμα Εκτιμήτριας του } \theta}$$

- Παρατηρήστε ότι το στατιστικό ελέγχου είναι μια δειγματοσυνάρτηση. Αυτό σημαίνει ότι από διαφορετικό δείγμα ίδιου μεγέθους ενδέχεται να προκύψουν διαφορετικές τιμές για την παραπάνω παράσταση.

Έλεγχοι Υποθέσεων

- Εν συνεχεία χωρίζουμε τον παραμετρικό χώρο σε περιοχή αποδοχής (οι τιμές του στατιστικού ελέγχου για τις οποίες δεν απορρίπτουμε την H_o) και σε κρίσιμη περιοχή (οι τιμές του στατιστικού ελέγχου για τις οποίες απορρίπτουμε την H_o). Ο εν λόγω διαχωρισμός του παραμετρικού χώρου εξαρτάται πέραν της τιμής του στατιστικού ελέγχου και από μια πιθανότητα α την οποία καλούμε επίπεδο σημαντικότητας του ελέγχου και ισούται με την πιθανότητα σφάλματος τύπου Ι.

Έλεγχοι Υποθέσεων

Απόφαση	Πραγματικότητα	
	H_0 Αληθής	H_1 Αληθής
Δεν απορρίπτω H_0	Σωστή Απόφαση	Σφάλμα Τύπου ΙΙ (πιθανότητα β)
Απορρίπτω H_0	Σφάλμα Τύπου Ι (πιθανότητα α)	Σωστή Απόφαση

- α = P(σφάλμα τύπου Ι) = P(απορρίπτω H_o | H_0 σωστή)
- β = P(σφάλμα τύπου ΙΙ) = P(δεν απορρίπτω H_o | H_1 σωστή)

Έλεγχοι Υποθέσεων

- Επιλέγουμε την κρίσιμη περιοχή έτσι ώστε να ελαχιστοποιούνται οι πιθανότητες των 2 ειδών σφαλμάτων.
- Κάτι τέτοιο δεν είναι πάντοτε εφικτό, οπότε στην πράξη κρατάμε το α σταθερό (π.χ. α=0.05) και ελαχιστοποιούμε το β.
- Η πιθανότητα P(απορρίπτω Ho | H_1 σωστή) = 1-β καλείται ισχύς του ελέγχου.

Έλεγχοι Υποθέσεων

1. Ορίζουμε τη μηδενική και την εναλλακτική υπόθεση με βάση το ερευνητικό μας ερώτημα.
2. Υπολογίζουμε το στατιστικό ελέγχου για τα δεδομένα μας.
3. Ορίζουμε την κρίσιμη περιοχή με βάση την προκαθορισμένη πιθανότητα σφάλματος τύπου Ι α.
4. Αν η τιμή του στατιστικού μας ελέγχου ανήκει στην κρίσιμη περιοχή απορρίπτουμε την μηδενική υπόθεση, αλλιώς δεν έχουμε με βάση τα δεδομένα σοβαρές ενδείξεις για να την απορρίψουμε.

Έλεγχοι Υποθέσεων

- Στο παράδειγμα 1 θέλουμε να ελέγξουμε αν η μέση διάρκεια ζωής συγκεκριμένων λαμπτήρων φθορίου είναι 2000h με εναλλακτική ότι δεν είναι. Έχουμε δηλαδή, σε ε.σ. έστω α, τον εξής έλεγχο:

$$H_0 : \mu = 2000$$

$$H_1 : \mu \neq 2000.$$

- Από τυχαίο δείγμα 50 τέτοιων λαμπτήρων έστω ότι η δειγματική μέση τιμή προέκυψε 1700h. Το ερώτημα λοιπόν είναι αν η διαφορά αυτή μεταξύ της τιμής του δείγματος και της υποτιθέμενης τιμής των 2000h μας δίνει σοβαρές ενδείξεις εναντίον της H_0 ή αν πιστεύουμε ότι προήλθε στην τύχη λόγω του συγκεκριμένου τυχαίου δείγματος που επιλέξαμε.
- Προφανώς όσο μεγαλύτερο είναι το τυπικό σφάλμα της εκτιμήτριας που χρησιμοποιούμε τόσο πιο εύκολα πιστεύουμε ότι η παρατηρούμενη αυτή διαφορά μπορεί να προέκυψε στην τύχη.

Έλεγχοι Υποθέσεων

- Υπολογίζουμε λοιπόν το στατιστικό ελέγχου

$$Z = \frac{\bar{X} - 2000}{SE(\bar{X})} = \frac{\bar{X} - 2000}{\sigma / \sqrt{n}}.$$

- Αν υποθέσουμε ότι γνωρίζουμε την τυπική απόκλιση του πληθυσμού, έστω σ=700h, τότε το παραπάνω στατιστικό για το συγκεκριμένο δείγμα παίρνει την τιμή

$$z = \frac{\bar{x} - 2000}{\sigma / \sqrt{n}} = \frac{1700 - 2000}{700 / \sqrt{50}} = -3.03.$$

Έλεγχοι Υποθέσεων

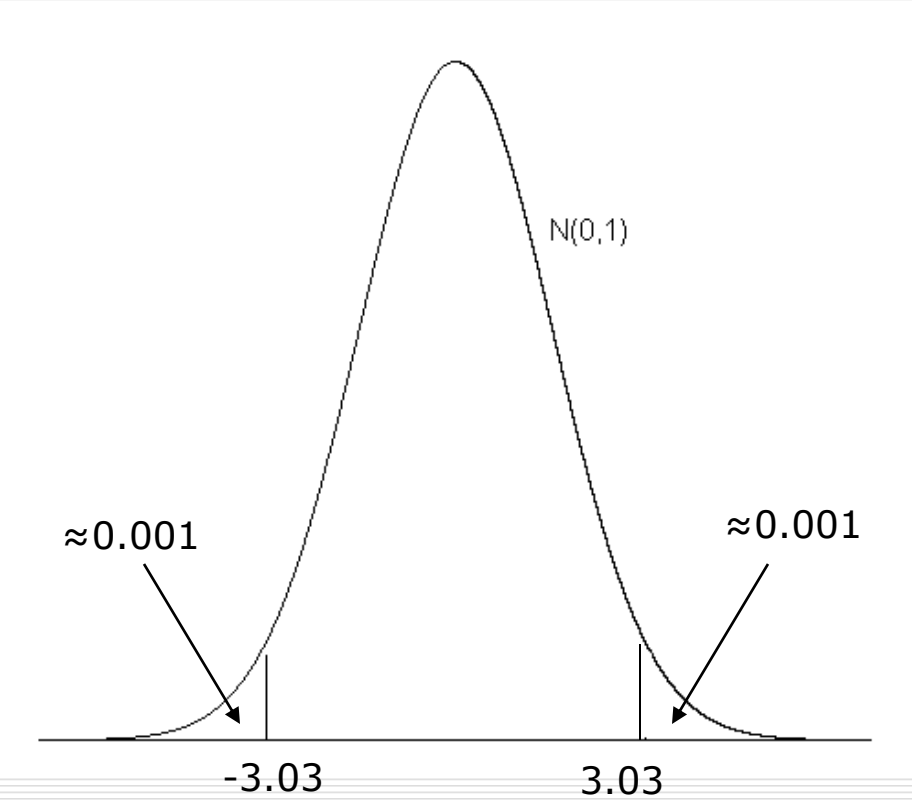
- Το ερώτημα λοιπόν είναι ποια είναι η πιθανότητα αν είχαμε κάποιο άλλο δείγμα να παρατηρούσαμε μια τιμή τόσο "ακραία" ή και ακόμα περισσότερο από το z=-3.03. Επειδή ο έλεγχος μας είναι αμφίπλευρος ως ακραία θεωρούμε οποιαδήποτε τιμή που είναι >3.03 ή <-3.03. Άρα αρκεί να βρούμε την περιοχή (πιθανότητα) κάτω από την σ.π.π. της τ.μ. Z δεξιά του 3.03 και αριστερά του -3.03 και να τις προσθέσουμε.

- Από το ΚΟΘ ξέρουμε ότι κάτω από την μηδενική υπόθεση

$$Z = \frac{\bar{X} - 2000}{\sigma / \sqrt{n}} \simeq N(0,1).$$

Έλεγχοι Υποθέσεων

- Από τους πίνακες της Κανονικής κατανομής βρίσκουμε ότι



Έλεγχοι Υποθέσεων

- Η πιθανότητα λοιπόν να παρατηρήσουμε μια τιμή τόσο "ακραία" ή και ακόμα περισσότερο από το |z|=3.03, δηλαδή η πιθανότητα της περιοχής δεξιά από το 3.03 συν την πιθανότητα της περιοχής αριστερά από το -3.03, είναι περίπου 0.002. Άρα κάτω από την μηδενική υπόθεση η τιμή του δειγματικού μέσου 1700h είναι πάρα πολύ απίθανη και άρα έχουμε πολύ σοβαρές ενδείξεις εναντίον της μηδενικής υπόθεσης, και οπότε την απορρίπτουμε.
- Η παραπάνω πιθανότητα καλείται P-τιμή του ελέγχου.
- Προσοχή! Η P-τιμή του ελέγχου δεν είναι η πιθανότητα η μηδενική υπόθεση να είναι αληθής.
- Η P-τιμή του ελέγχου είναι η πιθανότητα το στατιστικό ελέγχου που χρησιμοποιούμε να πάρει σε κάποιο άλλο δείγμα μία τόσο "ακραία" ή και ακόμα περισσότερο τιμή με αυτή που έχουμε παρατηρήσει, δεχόμενοι την μηδενική υπόθεση.
- Συνήθως όταν η P-τιμή < α απορρίπτουμε την μηδενική υπόθεση.

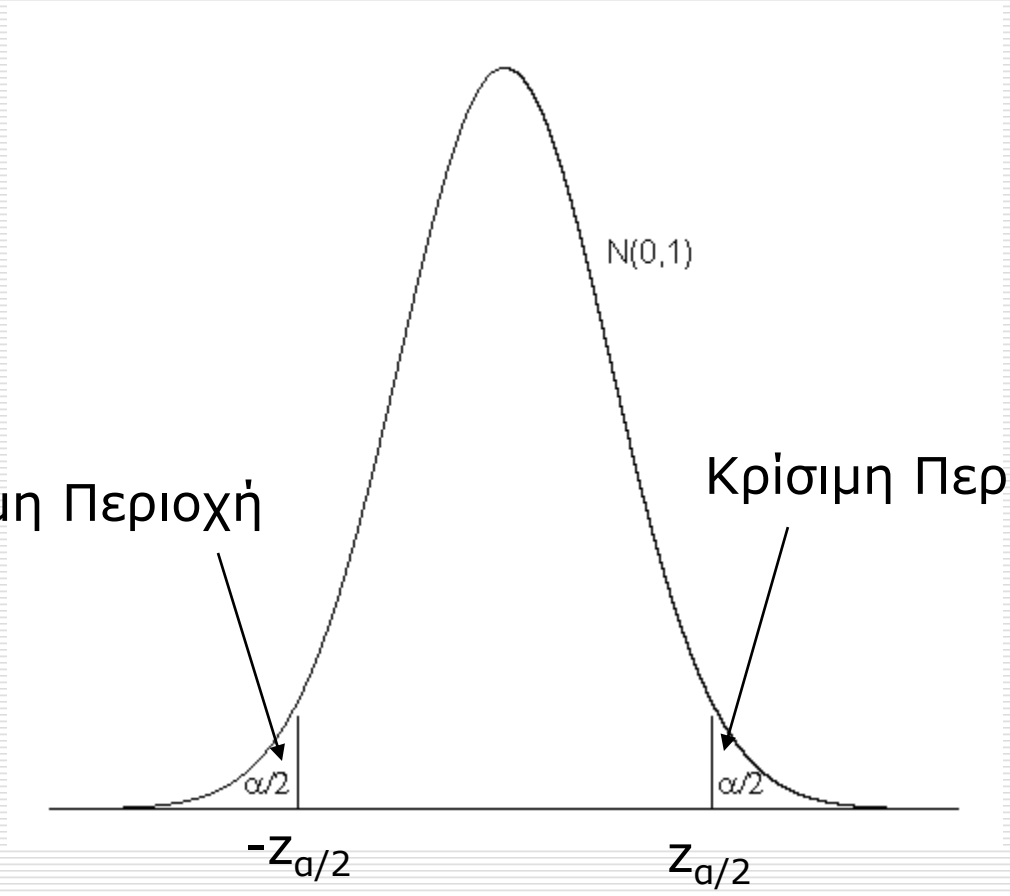
Έλεγχοι Υποθέσεων

- Αν θέλουμε για το εν λόγω παράδειγμα να δημιουργήσουμε την κρίσιμη περιοχή σε ε.σ. έστω α=0.05 θα ήταν λογικό να είχαμε τον εξής κανόνα για το στατιστικό ελέγχου μας:

Απορρίπτω την H_0 αν:

$z<-z_{\alpha/2}$ ή $z > z_{\alpha/2}$

Έλεγχοι Υποθέσεων



Έλεγχοι Υποθέσεων

- Για α=0.05 εύκολα βρίσκουμε από τους πίνακες της τυποποιημένης Κανονικής κατανομής ότι $z_{0.025}=1.96$.
- Άρα $z=-3.03<-z_{0.025}=-1.96$ οπότε απορρίπτουμε την μηδενική υπόθεση.
- Ισοδύναμα θα μπορούσαμε να κατασκευάζαμε ένα συμμετρικό 95% Δ.Ε. για το μ, το οποίο δεν είναι τίποτα άλλο από την περιοχή αποδοχής του αμφίπλευρου ελέγχου:

$$\left(\bar{x}-z_{\alpha/2}\sigma/\sqrt{n},\ \bar{x}+z_{\alpha/2}\sigma/\sqrt{n}\right)=(1700-1.96\cdot 500/\sqrt{50},1700-1.96\cdot 500/\sqrt{50})=$$

$$=(1506,1894)$$

- Παρατηρούμε ότι το παραπάνω ΔΕ δεν περιέχει την υποτιθέμενη τιμή με βάση την H_0 των 2000h και άρα οδηγούμαστε στο συμπέρασμα να απορρίψουμε την H_0.

Ένα δείγμα

- Έλεγχος για την μέση τιμή μιας ποσοτικής μεταβλητής:
  - Ας υποθέσουμε ότι έχουμε μια τυχαία μεταβλητή X από πληθυσμό με μέση τιμή μ και διασπορά σ^2 (μ, σ^2 άγνωστα) και ενδιαφερόμαστε να ελέγξουμε την υπόθεση H_0: $\mu = \mu_0$ έναντι της H_1: $\mu \neq \mu_0$ σε ε.σ. α. Έστω $X_1,\ldots,X_n$ τυχαίο δείγμα. Το στατιστικό ελέγχου μας τότε είναι

$$Z = \frac{\bar{X} - \mu_0}{SE(\bar{X})} = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

το οποίο ακολουθεί την N(0,1). Επειδή το σ είναι άγνωστο το εκτιμούμε από την δειγματική τυπική απόκλιση S και το στατιστικό ελέγχου γίνεται

$$T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}} \sim St(n-1).$$

Ένα δείγμα

- Απαραίτητη προϋπόθεση για τα παραπάνω είναι τα δεδομένα μας να είναι Κανονικά κατανεμημένα ή το μέγεθος του δείγματος είναι μεγάλο (n>50).
- Στην περίπτωση μάλιστα που το μέγεθος του δείγματος είναι μεγάλο η κατανομή του Student προσεγγίζεται από την Κανονική κατανομή, οπότε μπορούμε να θεωρήσουμε ότι το Τ ακολουθεί την N(0,1). Με βάση λοιπόν τους πίνακες της Student κατανομής (ή της N(0,1) για μεγάλο δείγμα) μπορούμε να βρούμε την P-τιμή του παραπάνω ελέγχου. Ο εν λόγω έλεγχος καλείται one sample t-test.
- Αν το μέγεθος του δείγματος δεν είναι μεγάλο και δεν ισχύει η κανονικότητα, τότε ή μετασχηματίζουμε κατάλληλα τα δεδομένα ώστε να επιτευχθεί η κανονικότητα ή χρησιμοποιούμε τον αντίστοιχο μη παραμετρικό έλεγχο όπως θα δούμε παρακάτω.
- Ισοδύναμα με τον παραπάνω αμφίπλευρο έλεγχο θα μπορούσαμε να κατασκευάζαμε ένα συμμετρικό (1-α)% Δ.Ε. για το μ, και να ελέγχαμε αν η υποτιθέμενη τιμή μ_0 ανήκει στο εν λόγω διάστημα. $\left(\bar{x} - t_{n-1,\alpha/2}\, s/\sqrt{n},\ \bar{x} + t_{n-1,\alpha/2}\, s/\sqrt{n}\right)$

Ένα δείγμα

- Η P-τιμή για τον εν λόγω έλεγχο προκύπτει με βάση την εναλλακτική υπόθεση:
  - Αν $H_1 : \mu \neq \mu_0$ τότε η P-τιμή είναι 2 φορές η πιθανότητα δεξιά του |T| (ή ισοδύναμα 2 φορές η πιθανότητα αριστερά του -|T|).
  - Αν $H_1 : \mu > \mu_0$ τότε η P-τιμή είναι πιθανότητα δεξιά του T.
  - Αν $H_1 : \mu < \mu_0$ τότε η P-τιμή είναι πιθανότητα αριστερά του T.

Ένα δείγμα

- Παράδειγμα στην R. Έστω X η διάρκεια ζωής συγκεκριμένων λαμπτήρων φθορίου. Ενδιαφερόμαστε να ελέγξουμε αν η μέση τιμή της τυχαίας μεταβλητής X, έστω μ, είναι 2000h, δηλαδή αν H_0: μ=2000 με εναλλακτική υπόθεση την H_1: μ≠2000 σε ε.σ. 0.05. Έστω πήραμε τις ακόλουθες παρατηρήσεις από τυχαίο δείγμα 20 τέτοιων λαμπτήρων:

  2229.697 2117.296 1907.911 2092.719 1972.783
  2435.007 1866.102 2042.217 2097.514 1733.638
  1885.404 2304.311 2201.424 2035.156 2384.666
  2487.807 1961.931 1711.411 1974.978 1674.833

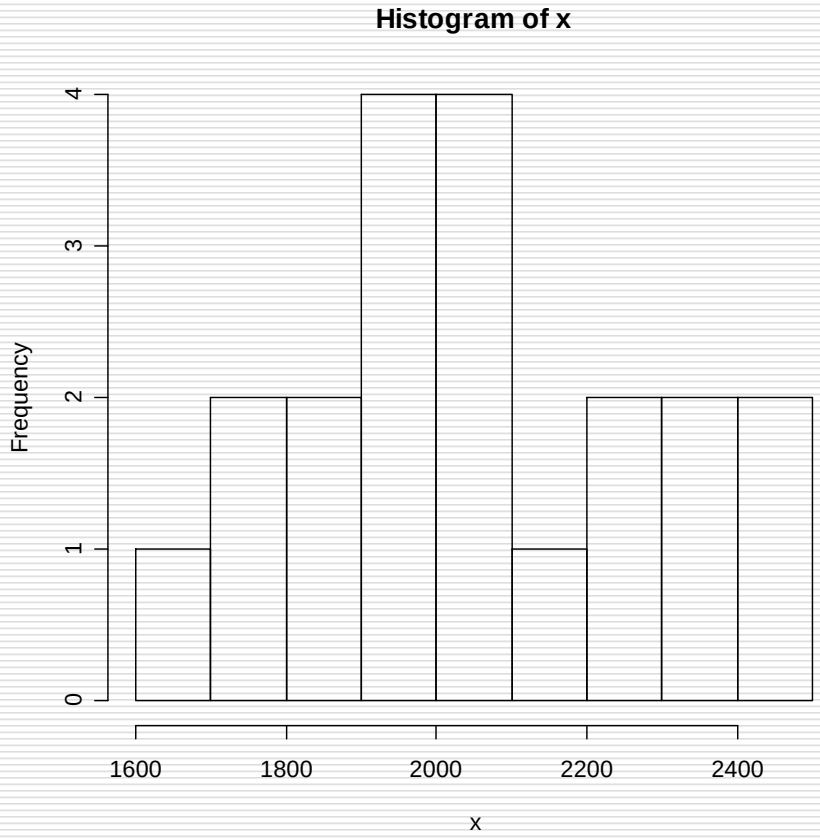
Ένα δείγμα

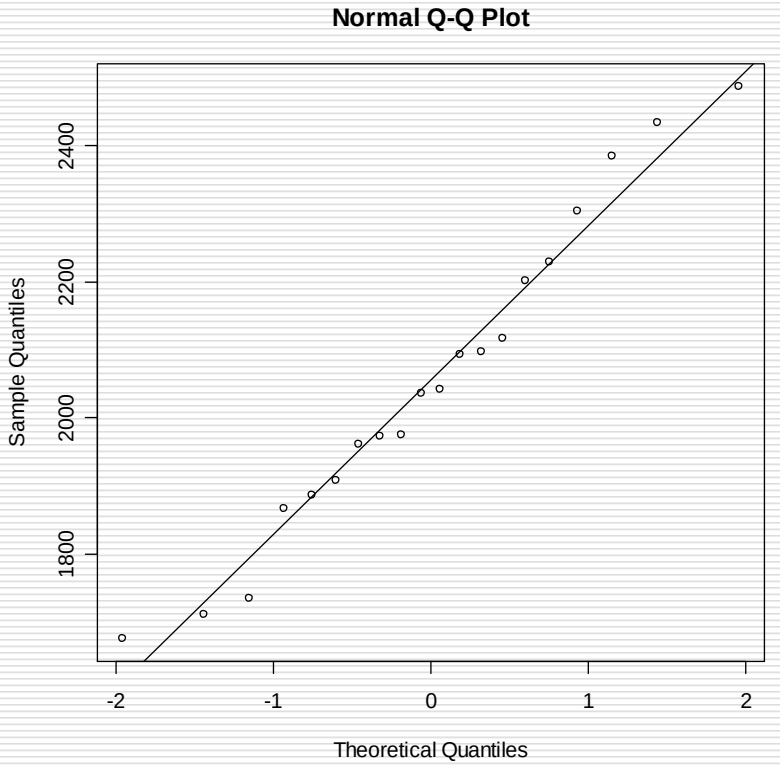
- Μιας και το μέγεθος του δείγματός μας δεν είναι τόσο μεγάλο στην αρχή ελέγχουμε αν η υπόθεση της κανονικότητας είναι λογική.

```
> x<-c(2229.697, 2117.296, 1907.911, 2092.719, 1972.783,
  2435.007, 1866.102, 2042.217,  2097.514, 1733.638,
  1885.404, 2304.311, 2201.424, 2035.156, 2384.666,
  2487.807, 1961.931, 1711.411, 1974.978, 1674.833)
> hist(x)
> qqnorm(x)
> qqline(x)
```

- Από τα επόμενα γραφήματα βλέπουμε ότι η υπόθεση της κανονικότητας δεν είναι παράλογη.

Ένα δείγμα


Histogram of x
Frequency
0
1
2
3
4
1600
1800
2000
2200
2400
x


Normal Q-Q Plot
Sample Quantiles
1800
2000
2200
2400
-2
-1
0
1
2
Theoretical Quantiles

Ένα δείγμα

- Εν συνεχεία χρησιμοποιούμε την έτοιμη συνάρτηση t.test. Ως βασικά ορίσματα δίνουμε στην συνάρτηση τα δεδομένα και την υποτιθέμενη τιμή του μ.

```
> t.test(x, mu=2000)

        One Sample t-test

data:  x
t = 1.0695, df = 19, p-value = 0.2983
alternative hypothesis: true mean is not equal to 2000
95 percent confidence interval:
 1946.558 2165.122
sample estimates:
mean of x
 2055.840
```

Τιμή του στατιστικού ελέγχου Τ

Βαθμοί Ελευθερίας της Student κατανομής

P – τιμή του ελέγχου

Συμμετρικό 95% ΔΕ για το μ

Δειγματικός μέσος

Ένα δείγμα

- Από τα αποτελέσματα του παραπάνω ελέγχου καταλήγουμε ότι δεν έχουμε σοβαρές ενδείξεις εναντίον της μηδενικής υπόθεσης, οπότε δεν την απορρίπτουμε.
- Η παραπάνω συνάρτηση στην R έχει ως προκαθορισμένη τιμή για το α (το ε.σ. του ελέγχου) το 5%. Αν θέλουμε μια άλλη τιμή την περνάμε σαν όρισμα. Αυτό θα έχει ως αποτέλεσμα να αλλάξει το Δ.Ε.

Ένα δείγμα

```
> t.test(x, mu=2000, conf.level=0.99)

    One Sample t-test
data:  x
t = 1.0695, df = 19, p-value = 0.2983
alternative hypothesis: true mean is not equal to 2000
99 percent confidence interval:
1906.464 2205.217
sample estimates:
mean of x
2055.840
```

Ένα δείγμα

- Επίσης θα μπορούσαμε να είχαμε θεωρήσει και κάποιον μονόπλευρο έλεγχο, π.χ. αν H_0: μ=2000 με εναλλακτική υπόθεση την H_1: μ< 2000, με την βοήθεια του ορίσματος alternative. Η προκαθορισμένη του τιμή είναι "two.sided" (≠) ενώ εναλλακτικά μπορούμε να χρησιμοποιήσουμε "greater" (>) or "less" (<). Προσέξτε σε αυτήν την περίπτωση πως αλλάζει η P-τιμή καθώς επίσης και το Δ.Ε.

Ένα δείγμα

```
> t.test(x, mu=2000, alternative="less")

        One Sample t-test

data:  x
t = 1.0695, df = 19, p-value = 0.8509
alternative hypothesis: true mean is less than 2000
95 percent confidence interval:
     -Inf 2146.123
sample estimates:
mean of x
 2055.840
```

$$\left(-\infty,\ \bar{x} + t_{n-1,\alpha}\, s/\sqrt{n}\right)$$

Ένα δείγμα

- Ο αντίστοιχος μη παραμετρικός έλεγχος όταν δεν ισχύουν οι προϋποθέσεις καλείται Wilcoxon test και στην R χρησιμοποιούμε την εντολή wilcox.test.

```
> wilcox.test(x,mu=2000)

        Wilcoxon signed rank test

data:  x
V = 131, p-value = 0.3488
alternative hypothesis: true location is not equal to 2000
```

P – τιμή του ελέγχου

- Με την βοήθεια και πάλι των ορισμάτων conf.level και alternative μπορούμε να αλλάξουμε το α καθώς και την εναλλακτική υπόθεση.

Ένα δείγμα

- Όταν στο δείγμα υπάρχουν ισοπαλίες (παρατηρήσεις με την ίδια τιμή) ή μηδενικά στα μετασχηματισμένα δεδομένα που προκύπτουν αφαιρώντας από τα αρχικά δεδομένα την υποτιθέμενη κάτω από την μηδενική υπόθεση τιμή του μ, η εντολή wilcox.test της R μας δίνει προειδοποιητικό μήνυμα και δεν υπολογίζει την ακριβή P-τιμή του ελέγχου αλλά αυτή που προκύπτει από μια κανονική προσέγγιση.
- Σε περιπτώσεις όπως η παραπάνω μπορούμε λαμβάνοντας υπόψιν ισοπαλίες και μηδενικά να υπολογίσουμε ακριβή P-τιμή, χρησιμοποιώντας την εντολή wilcox.exact που βρίσκεται στο πακέτο exactRankTests.
- Το πακέτο exactRankTests όταν το φορτώσετε στην R, θα σας βγάλει ένα προειδοποιητικό μήνυμα ότι "is not longer under development". Παρόλα αυτά μπορείτε ακόμα να τρέξετε την εντολή wilcox.exact.

Ένα δείγμα

Έλεγχος ποσοστών:

- Ας υποθέσουμε ότι έχουμε μια κατηγορική δίτιμη τυχαία μεταβλητή X με τιμές 0 και 1 και P(X=1)=p (άγνωστο). Προφανώς τότε η X~Bernoulli(p). Ενδιαφερόμαστε να ελέγξουμε την υπόθεση H_0: $p = p_0$ έναντι της H_1: $p \neq p_0$ σε ε.σ. α. Έστω $X_1,\ldots,X_n$ τυχαίο δείγμα. Το στατιστικό ελέγχου μας τότε είναι

$$Z = \frac{\hat{P} - p_0}{SE(\hat{P})} = \frac{\hat{P} - p_0}{\sqrt{p_0(1-p_0)/n}}, \text{ όπου } \hat{P} = \bar{X} \text{ η σχετική}$$

συχνότητα της τιμής 1 στο δείγμα μας.

Ένα δείγμα

- Με βάση το Κ.Ο.Θ., για μεγάλο n, το $Z \dot{\sim} N(0,1)$ κάτω από την μηδενική υπόθεση. Συχνά στην προκειμένη περίπτωση για να είναι πιο ικανοποιητική η προσέγγισή μας προχωράμε σε μια διόρθωση γνωστή ως διόρθωση συνέχειας του Yates (Yates continuity correction). Το στατιστικό ελέγχου γίνεται

$$Z = \frac{\left|\hat{P} - p_0\right| - 1/2n}{\sqrt{p_0(1-p_0)/n}}.$$

- Υπολογίζουμε την τιμή του Z με βάση τις παρατηρήσεις μας και με την βοήθεια του πίνακα της τυποποιημένης Κανονικής κατανομής βρίσκουμε την P-τιμή.
- Αντίστοιχα μπορούμε να υπολογίσουμε το Z^2 το οποίο ακολουθεί την X^2 κατανομή με 1 βαθμό ελευθερίας και να βρούμε την P-τιμή με την βοήθεια του πίνακα της X^2 κατανομής με 1 βαθμό ελευθερίας.

Ένα δείγμα

- Ισοδύναμα με τον παραπάνω αμφίπλευρο έλεγχο θα μπορούσαμε να κατασκευάσουμε ένα συμμετρικό (1-α)% Δ.Ε. για το p, **εκτιμώντας** το τυπικό σφάλμα με την βοήθεια της τιμής $\hat{p}$, και να δούμε αν η υποτιθέμενη τιμή p_0 ανήκει στο εν λόγω διάστημα.

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n},\ \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\right)$$

Ένα δείγμα

- Η P-τιμή για τον εν λόγω έλεγχο προκύπτει και πάλι με βάση την εναλλακτική υπόθεση:
  - Αν $H_1 : p \neq p_0$ τότε η P-τιμή είναι 2 φορές η πιθανότητα δεξιά του |Z| (ή ισοδύναμα 2 φορές η πιθανότητα αριστερά του -|Z|). Ισοδύναμα αν δουλέψουμε με το Z^2 η P-τιμή είναι η πιθανότητα δεξιά του Z^2.
  - Αν $H_1 : p > p_0$ τότε η P-τιμή είναι πιθανότητα δεξιά του Z. Ισοδύναμα αν δουλέψουμε με το Z^2 η P-τιμή είναι το 1/2 της πιθανότητας δεξιά του Z^2.
  - Αν $H_1 : p < p_0$ τότε η P-τιμή είναι πιθανότητα αριστερά του Z. Ισοδύναμα αν δουλέψουμε με το Z^2 η P-τιμή είναι ίση με το συμπλήρωμα του 1/2 της πιθανότητας δεξιά του Z^2.

Ένα δείγμα

- Για να ισχύει το Κ.Ο.Θ. και όλα τα προηγούμενα θα πρέπει το μέγεθος του δείγματος να είναι μεγάλο. Το πόσο μεγάλο πρέπει να είναι το n σχετίζεται με το πόσο συμμετρική είναι η διωνυμική κατανομή που στην πραγματικότητα έχουμε και για αυτό στην πράξη ελέγχουμε αν

$$n \cdot p_0 \geq 5$$

$$\text{ΚΑΙ}$$

$$n \cdot (1 - p_0) \geq 5$$

Ένα δείγμα

- Παράδειγμα στην R. Κατασκευαστής ισχυρίζεται ότι το 2% των προϊόντων του είναι ελαττωματικά. Σε τυχαίο δείγμα n=1000 τέτοιων προϊόντων βρέθηκαν κ=30 ελαττωματικά ($\hat{p} = 30/1000$). Θέλουμε σε ε.σ. 5% να ελέγξουμε τον ισχυρισμό του κατασκευαστή (H_0: p=0.02) με εναλλακτική H_1: p≠0.02.

  Παρατηρούμε ότι np_0= 20 και $n(1-p_0)$= 980

Ένα δείγμα

- Ο έλεγχος στην R μπορεί να γίνει με την βοήθεια της εντολής prop.test.

p_0

Διόρθωση συνέχειας

```
> prop.test(30,1000, p=0.02)

        1-sample proportions test with continuity correction

data:  30 out of 1000, null probability 0.02
X-squared = 4.6046, df = 1, p-value = 0.03189
alternative hypothesis: true p is not equal to 0.02
95 percent confidence interval:
0.02067971 0.04308482
sample estimates:
   p
0.03
```

Μέγεθος δείγματος

Συχνότητα στο δείγμα της τιμής 1

P – τιμή του ελέγχου

z^2

Συμμετρικό 95% ΔΕ για το p

$\hat{p}$

Ένα δείγμα

- Από τα αποτελέσματα του παραπάνω ελέγχου καταλήγουμε ότι, σε ε.σ. 5%, έχουμε σοβαρές ενδείξεις εναντίον της μηδενικής υπόθεσης, οπότε την απορρίπτουμε.
- Μπορούμε να ζητήσουμε στην R να μην γίνει η διόρθωση συνέχειας καθώς επίσης και να αλλάξουμε την προκαθορισμένη τιμή του ε.σ. που και εδώ είναι 5%. Τέλος μπορούμε να ζητήσουμε μονόπλευρο έλεγχο.

Ένα δείγμα

```
> prop.test(30,1000, p=0.02,correct=FALSE)

        1-sample proportions test without continuity
        correction

data:  30 out of 1000, null probability 0.02
X-squared = 5.102, df = 1, p-value = 0.0239
alternative hypothesis: true p is not equal to 0.02
95 percent confidence interval:
 0.02109374 0.04250341
sample estimates:
   p
0.03
```

Ένα δείγμα

```
> prop.test(30,1000, p=0.02,conf.level=0.99)

        1-sample proportions test with continuity correction

data:  30 out of 1000, null probability 0.02
X-squared = 4.6046, df = 1, p-value = 0.03189
alternative hypothesis: true p is not equal to 0.02
99 percent confidence interval:
 0.01851851 0.04789404
sample estimates:
   p
0.03
```

Ένα δείγμα

```
> prop.test(30,1000, p=0.02,alternative="greater")

        1-sample proportions test with continuity correction

data:  30 out of 1000, null probability 0.02
X-squared = 4.6046, df = 1, p-value = 0.01594
alternative hypothesis: true p is greater than 0.02
95 percent confidence interval:
0.02188911 1.00000000
sample estimates
   p
0.03
```

$$\left(\hat{p} - z_{\alpha}\sqrt{\hat{p}(1-\hat{p})/n},\ 1\right)$$

Ένα δείγμα

- Όταν δεν ισχύουν οι προϋποθέσεις του Κ.Ο.Θ. και παρόλα αυτά εμείς εφαρμόσουμε την εντολή prop.test παίρνουμε ένα προειδοποιητικό μήνυμα λάθους από την R.
- Σε αυτές τις περιπτώσεις πρέπει να εφαρμόσουμε το Διωνυμικό κριτήριο (binomial test) με την βοήθεια της εντολής binom.test. Και στην περίπτωση αυτή με ανάλογο τρόπο όπως και πριν μπορούμε να αλλάξουμε την προκαθορισμένη τιμή του ε.σ. του 5%, καθώς και να ζητήσουμε μονόπλευρο έλεγχο.
- Ας υποθέσουμε στο προηγούμενο παράδειγμα ότι $n=100$, $k=10$ και $p_0=0.01$. Τότε $np_0= 1$.

Ένα δείγμα

```
> prop.test(10,100, p=0.01)

        1-sample proportions test with continuity correction

data:  10 out of 100, null probability 0.01
X-squared = 72.9798, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is not equal to 0.01
95 percent confidence interval:
 0.0516301 0.1803577
sample estimates:
  p
0.1

Warning message:
In prop.test(10, 100, p = 0.01) :
 Chi-squared approximation may be incorrect
```

Ένα δείγμα

```
> binom.test(10,100,p=0.01)

        Exact binomial test

data:  10 and 100
number of successes = 10, number of trials = 100, p-value = 7.632e-
    08
alternative hypothesis: true probability of success is not equal to 0.01
95 percent confidence interval:
 0.04900469 0.17622260
sample estimates:
probability of success
                   0.1
```

Δύο ανεξάρτητα δείγματα

- Έλεγχος για την διαφορά των μέσων τιμών δύο ποσοτικών μεταβλητών:
  - Έστω ότι έχουμε μετρήσεις της ίδιας ποσοτικής μεταβλητής σε δύο ομάδες διαφορετικών ατόμων (δύο διαφορετικούς πληθυσμούς). Είναι εύλογη τότε η αναζήτηση πιθανής τους σχέσης. Πιο συγκεκριμένα έστω το χαρακτηριστικό Χ από έναν πληθυσμό με μέση τιμή μ_1 και τυπική απόκλιση σ_1 (μ_1, σ_1 άγνωστα) και έστω Υ το ίδιο χαρακτηριστικό από έναν άλλο πληθυσμό με μέση τιμή μ_2 και τυπική απόκλιση σ_2 (μ_2, σ_2 άγνωστα). Ας θεωρήσουμε ότι Χ, Υ ανεξάρτητες, δηλαδή η τιμή που παίρνει το υπό μελέτη χαρακτηριστικό στον πρώτο πληθυσμό δεν επηρεάζει την τιμή που παίρνει το ίδιο χαρακτηριστικό στο δεύτερο πληθυσμό. Έστω $X_1,\ldots,X_{n1}$ τυχαίο δείγμα από τον πρώτο πληθυσμό και $Y_1,\ldots,Y_{n2}$ τυχαίο δείγμα από τον δεύτερο πληθυσμό. Για να ελέγξουμε πιθανή διαφοροποίηση του υπό μελέτη χαρακτηριστικού στους δύο πληθυσμούς είναι λογικό τότε να ελέγξουμε την υπόθεση H_0: $\mu_1 = \mu_2$ έναντι της H_1: $\mu_1 \neq \mu_2$ σε ε.σ. α, δηλαδή να ελέγξουμε αν το υπό μελέτη χαρακτηριστικό έχει την ίδια μέση τιμή στους δύο πληθυσμούς.

Δύο ανεξάρτητα δείγματα

- Το στατιστικό ελέγχου μας τότε είναι

=0 με βάση την H_0

$$\frac{\left(\bar{X}-\bar{Y}\right)-\left(\mu_1-\mu_2\right)}{SE(\bar{X}-\bar{Y})}=\frac{\left(\bar{X}-\bar{Y}\right)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$$

- Στον παραπάνω στατιστικό έλεγχο μπορούμε να αντικαταστήσουμε τις άγνωστες διασπορές των πληθυσμών στον παρονομαστή με τις αντίστοιχες δειγματικές διασπορές, οπότε το στατιστικό ελέγχου γίνεται

$$Z=\frac{\left(\bar{X}-\bar{Y}\right)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$$

Δύο ανεξάρτητα δείγματα

- Όταν τα μεγέθη των 2 δειγμάτων είναι μεγάλα με βάση το Κ.Ο.Θ. το Z ακολουθεί προσεγγιστικά την N(0,1) και άρα η P-τιμή του ελέγχου είναι 2 φορές η πιθανότητα της περιοχής της N(0,1) δεξιά από το |Z|. Αν ο έλεγχος είναι μονόπλευρος η P-τιμή του ελέγχου είναι η πιθανότητα της περιοχής της N(0,1) δεξιά ή αριστερά από το Z ανάλογα με την εναλλακτική.
- Ισοδύναμα με τον παραπάνω αμφίπλευρο έλεγχο θα μπορούσαμε να είχαμε κατασκευάσει ένα συμμετρικό (1-α)% Δ.Ε. της διαφοράς των μέσων και να ελέγχαμε αν περιέχει το μηδέν.

$$\left[\bar{x}-\bar{y}-z_{\alpha/2}\left\{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right\}^{\frac{1}{2}} < \mu_1-\mu_2 < \bar{x}-\bar{y}+z_{\alpha/2}\left\{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right\}^{\frac{1}{2}}\right].$$

Δύο ανεξάρτητα δείγματα

- Όταν τα δεδομένα προέρχονται από Κανονικούς πληθυσμούς θεωρούμε τις εξής 2 περιπτώσεις:

1. Οι πληθυσμοί έχουν ίσες τυπικές αποκλίσεις, δηλαδή $\sigma_1=\sigma_2=\sigma$ (άγνωστη).

   Στην περίπτωση αυτή υπολογίζουμε την συγχωνευμένη (pooled) δειγματική διασπορά

$$S^2 = \{(n_1-1)S_1^2 + (n_2-1)S_2^2\} / (n_1+n_2-2)$$

όπου S_i^2 οι δύο δειγματικές διασπορές, $i=1,2.$

Δύο ανεξάρτητα δείγματα

και έχουμε το ακόλουθο στατιστικό ελέγχου

$$T = \frac{\left(\bar{X} - \bar{Y}\right)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim St\left(n_1 + n_2 - 2\right).$$

Υπολογίζουμε το Τ και η P-τιμή του ελέγχου είναι 2 φορές η πιθανότητα της περιοχής της $St(n_1+n_2-2)$ δεξιά από το $|T|$. Αν ο έλεγχος είναι μονόπλευρος η P-τιμή του ελέγχου είναι η πιθανότητα της περιοχής της $St(n_1+n_2-2)$ δεξιά ή αριστερά από το Τ ανάλογα με την εναλλακτική. Ο εν λόγω έλεγχος καλείται two sample t-test.

Ισοδύναμα με τον παραπάνω αμφίπλευρο έλεγχο θα μπορούσαμε να είχαμε κατασκευάσει ένα συμμετρικό (1-α)% Δ.Ε. για την διαφορά των μέσων και να ελέγχαμε αν περιέχει το 0.

$$\left(\bar{x} - \bar{y}\right) \pm t_{n_1+n_2-2,\alpha/2} s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Δύο ανεξάρτητα δείγματα

2. Οι πληθυσμοί έχουν άνισες και άγνωστες τυπικές αποκλίσεις.

   Στην περίπτωση αυτή το στατιστικό ελέγχου είναι

$$T = \frac{\left(\bar{X} - \bar{Y}\right)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \dot{\sim} St(\nu),$$

προσεγγιστικά

$$\nu = \frac{\left\{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right\}^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}.$$

Δύο ανεξάρτητα δείγματα

Υπολογίζουμε το Τ και η P-τιμή του ελέγχου είναι 2 φορές η πιθανότητα της περιοχής της St(ν) δεξιά από το |T|. Αν ο έλεγχος είναι μονόπλευρος η P-τιμή του ελέγχου είναι η πιθανότητα της περιοχής της St(ν) δεξιά ή αριστερά από το Τ ανάλογα με την εναλλακτική.

Ο τελευταίος αυτός έλεγχος δίνει προσεγγιστικά αποτελέσματα και έχει την ονομασία Welch Two Sample t-test.

Το συμμετρικό (1-α)% Δ.Ε. στην προκειμένη περίπτωση είναι

$$\left[\bar{x}-\bar{y}-t_{\nu,\alpha/2}\left\{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right\}^{\frac{1}{2}} < \mu_1-\mu_2 < \bar{x}-\bar{y}+t_{\nu,\alpha/2}\left\{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right\}^{\frac{1}{2}}\right].$$

Δύο ανεξάρτητα δείγματα

- Με ποια κριτήρια όμως αποφασίζουμε αν οι διασπορές των πληθυσμών είναι ίσες ή όχι (περίπτωση 1 ή 2); Συχνά η απόφασή μας λαμβάνεται με βάση το αποτέλεσμα του ελέγχου

$$H_0 : \sigma_1^2 = \sigma_2^2 \ \ \mu\varepsilon\ \varepsilon\nu\alpha\lambda\lambda\alpha\kappa\tau\iota\kappa\acute{\eta}\ H_1 : \sigma_1^2 \neq \sigma_2^2,\ \sigma\varepsilon\ \varepsilon.\sigma.\ \alpha.$$

- Αν δεν έχουμε σοβαρές ενδείξεις από τα δεδομένα για να απορρίψουμε την παραπάνω μηδενική υπόθεση, τότε θεωρούμε ισότητα διασπορών και πηγαίνουμε με βάση την περίπτωση 1. Αν έχουμε σοβαρές ενδείξεις εναντίον της H_0 τότε την απορρίπτουμε, θεωρούμε δηλαδή ότι οι διασπορές είναι άνισες και προχωράμε με βάση την περίπτωση 2.

Δύο ανεξάρτητα δείγματα

- Κάτω από την μηδενική υπόθεση αναμένεται ο λόγος

$$F = \frac{S_1^2}{S_2^2} \sim F(n_1 - 1,\ n_2 - 1).$$

Κατανομή του Snedecor

- Υπολογίζουμε λοιπόν το στατιστικό ελέγχου F και η P-τιμή του ελέγχου είναι 2 φορές η πιθανότητα της περιοχής της $F(n_1 -1,n_2-1)$ δεξιά από το F αν $F \geq 1$ ή 2 φορές η πιθανότητα της περιοχής της $F(n_1 -1,n_2-1)$ αριστερά από το F αν $F<1$. Ο εν λόγω έλεγχος καλείται variance test ή F-test.
- Ισοδύναμα με τον παραπάνω αμφίπλευρο έλεγχο θα μπορούσαμε να είχαμε κατασκευάσει ένα συμμετρικό (1-α)% Δ.Ε. του λόγου των διασπορών και να ελέγχαμε αν περιέχει την μονάδα.

$$\frac{1}{F_{n_1-1,n_2-1,\alpha/2}} \cdot \frac{s_1^2}{s_2^2} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1}{F_{n_1-1,n_2-1,1-\alpha/2}} \cdot \frac{s_1^2}{s_2^2}$$

Δύο ανεξάρτητα δείγματα

- Η R για το t-test χρησιμοποιεί τον τύπο με την συγχωνευμένη διασπορά αν δηλωθεί από τον χρήστη ότι οι διασπορές είναι ίσες ή τον τύπο του Welch αν δηλωθεί από τον χρήστη ότι οι διασπορές είναι άνισες. Δεν εφαρμόζει δηλαδή το Κ.Ο.Θ. Για τον λόγο αυτόν πρέπει πάντα να ελέγχεται η Κανονικότητα. Αν τα μεγέθη των δειγμάτων είναι μεγάλα, τότε τυχόν αποκλίσεις από την Κανονικότητα δεν επηρεάζουν τα αποτελέσματα, και το στατιστικό του Welch είναι περίπου ίσο με το αντίστοιχο στατιστικό Z, υπό την προϋπόθεση του Κ.Ο.Θ.

Δύο ανεξάρτητα δείγματα

- Παράδειγμα στην R. Έστω Χ η διάρκεια ζωής λαμπτήρων πυρακτώσεως και Υ η διάρκεια ζωής λαμπτήρων φθορίου. Ενδιαφερόμαστε να ελέγξουμε αν η μέση τιμή της τυχαίας μεταβλητής Χ, έστω μ_1, ισούται με την μέση τιμή της τ.μ Υ, έστω μ_2, δηλαδή αν H_0: $\mu_1=\mu_2$ με εναλλακτική υπόθεση την H_1: $\mu_1\neq\mu_2$ σε ε.σ. 0.05. Έστω τα εξής δεδομένα

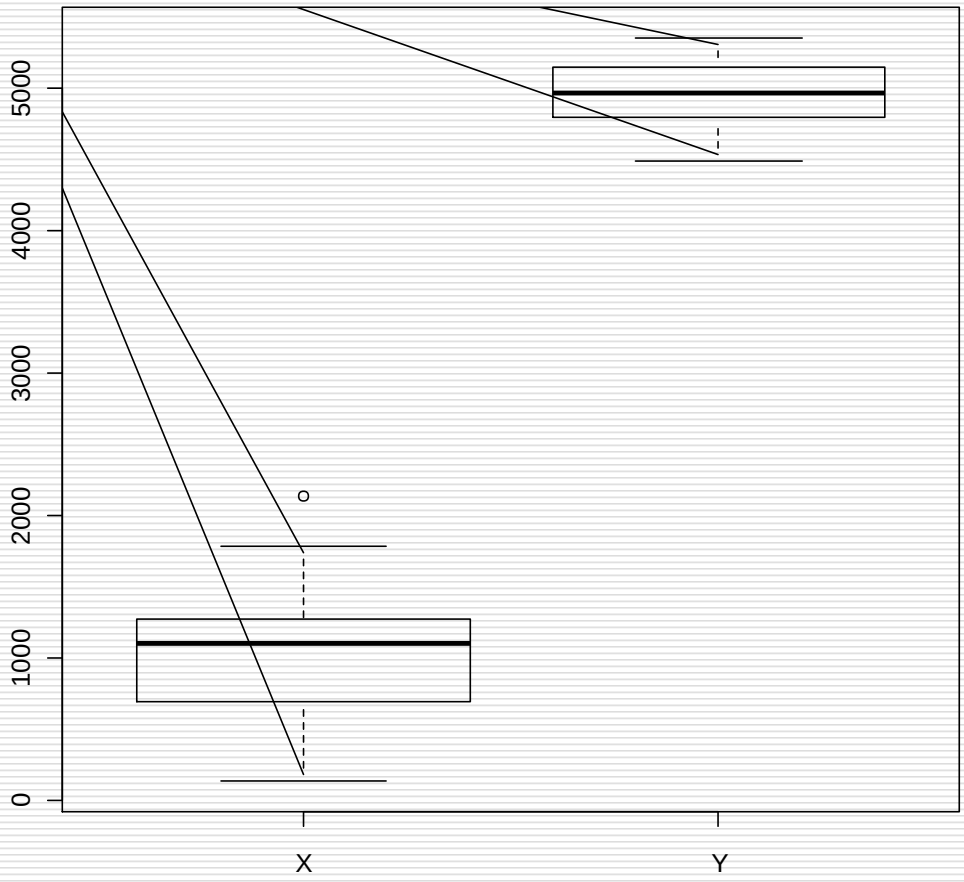
Δύο ανεξάρτητα δείγματα

```
> X
1101.03772 1786.32947 1277.69708 1109.69338 1197.77010 1686.87452
861.60858  133.78242 1261.79457  457.88062  400.99297  796.58186
1101.53622  664.15931 1235.71573 1439.75277 2140.81105  688.02609
700.07342 1138.60536
> Y
5275.439 4894.054 4992.239 4739.001 4946.254 4493.245 4842.237 5256.085
4780.912 4727.235 5357.285 5143.249 5020.929 4790.350 5013.482 5230.848
4929.022 5049.023

> length(X)
[1] 20
> length(Y)
[1] 18
```

```
>  summary(X)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
133.8   697.1  1106.0  1059.0  1266.0  2141.0
> summary(Y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 4493    4803    4969    4971    5120    5357
> boxplot(X,Y,names=c("X","Y"))
```

Δύο ανεξάρτητα δείγματα

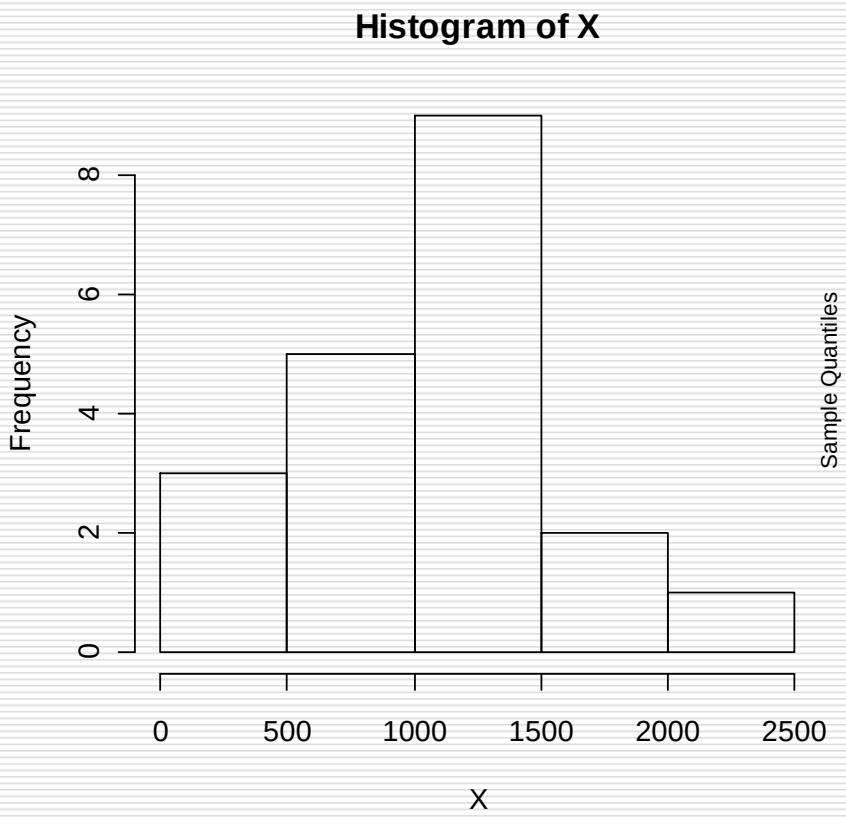


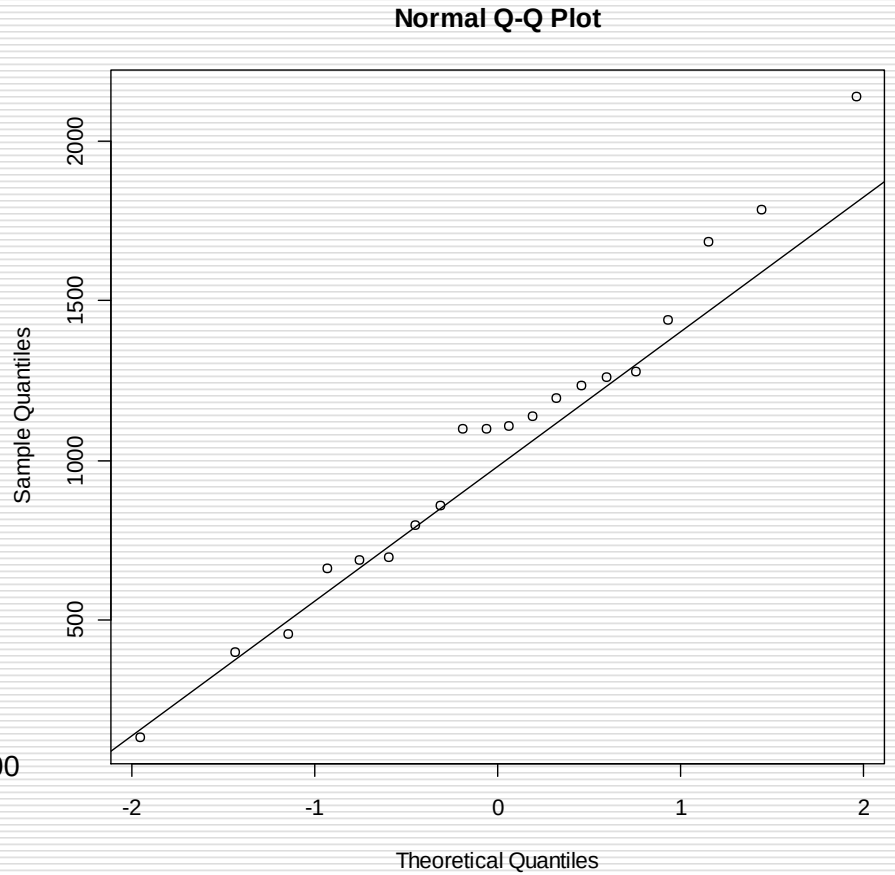
Δύο ανεξάρτητα δείγματα

- Η μέση διάρκεια ζωής στο δείγμα των λαμπτήρων φθορίου είναι 4971h ενώ των αντίστοιχων πυρακτώσεως είναι 1059h. Είναι αυτή η διαφορά που βλέπουμε στο δείγμα στατιστικά σημαντική, μπορούμε δηλαδή να βγάλουμε αντίστοιχα συμπεράσματα και για τον πληθυσμό ή οφείλεται στα συγκεκριμένα δείγματα;
- Θα ελέγξουμε αρχικά αν τα δεδομένα προέρχονται από Κανονικούς πληθυσμούς.

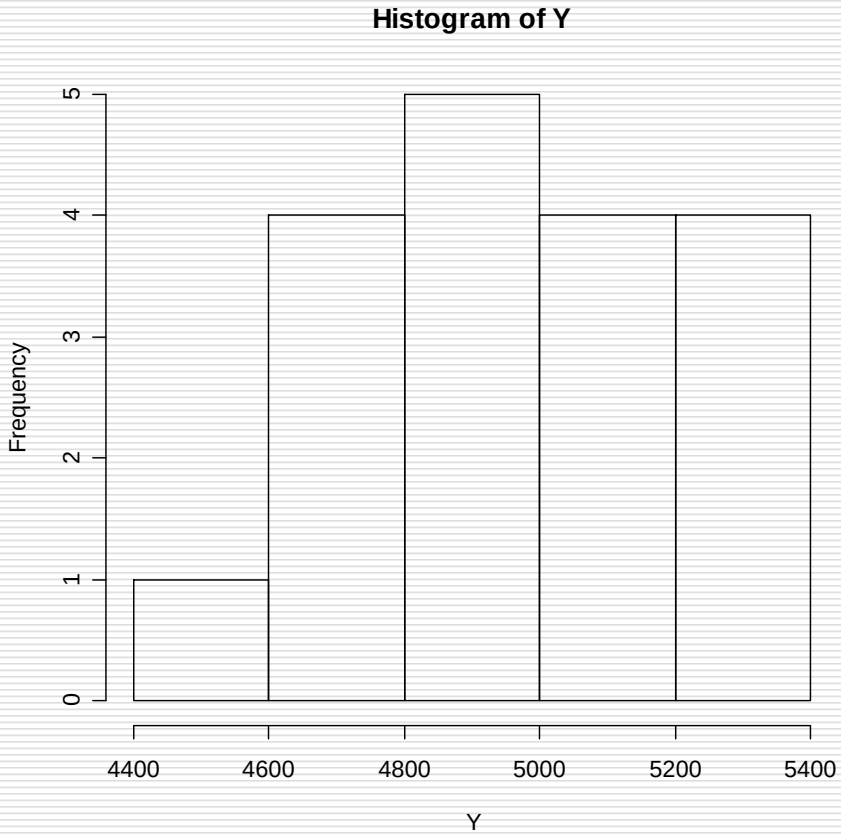
```
> hist(X)
> qqnorm(X)
> qqline(X)
> hist(Y)
> qqnorm(Y)
> qqline(Y)
```

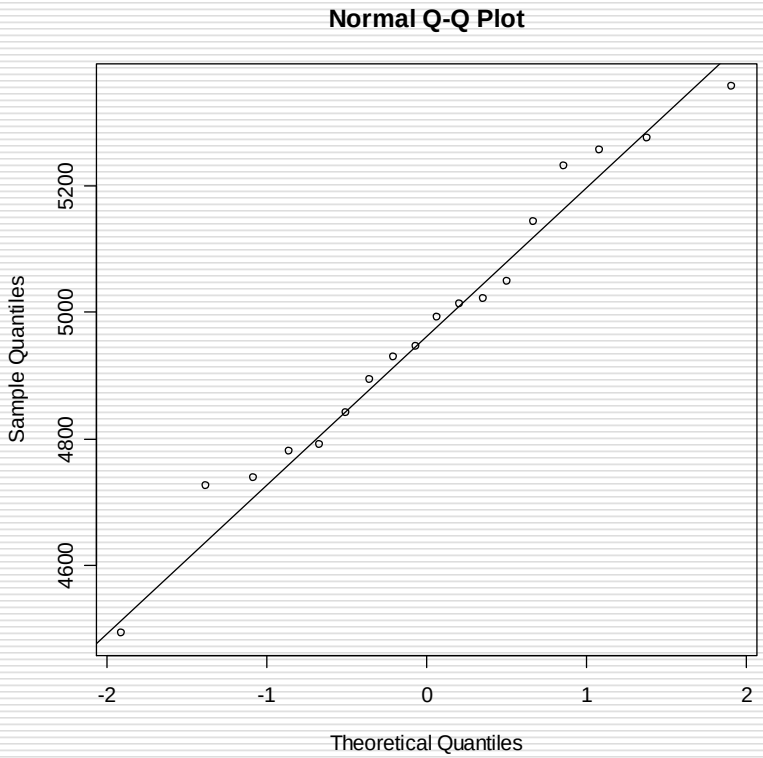
Δύο ανεξάρτητα δείγματα





Δύο ανεξάρτητα δείγματα





Δύο ανεξάρτητα δείγματα

- Παρατηρούμε ότι δεν υπάρχουν μεγάλες αποκλίσεις από την Κανονική κατανομή και για τα δύο μεγέθη.
- Εν συνεχεία ελέγχουμε την ισότητα διασπορών με την βοήθεια της εντολής var.test.

Δύο ανεξάρτητα δείγματα

```
> var.test(X,Y)

        F test to compare two variances

data:  X and Y
F = 4.7208, num df = 19, denom df = 17, p-value = 0.0002266
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  1.792851 12.118288
sample estimates:
ratio of variances
          4.72081
```

F στατιστικό ελέγχου

Βαθμοί ελευθερίας F κατανομής

P-τιμή

Συμμετρικό 95% ΔΕ για τον λόγο των Διασπορών. Παρατηρήστε ότι δεν περιέχει το 1 και μπορούμε να συμπεράνουμε ότι $\sigma_1 > \sigma_2$

$$\frac{s_1^2}{s_2^2}$$

Δύο ανεξάρτητα δείγματα

- Παρατηρούμε ότι η P-τιμή είναι πολύ μικρή οπότε έχουμε σοβαρές ενδείξεις εναντίον της H_0 και άρα απορρίπτουμε την υπόθεση για ισότητα διασπορών. Εν συνεχεία εφαρμόζουμε το Welch Two Sample t-test.

Δύο ανεξάρτητα δείγματα

```
> t.test(X,Y, var.equal=FALSE)

        Welch Two Sample t-test

data:  X and Y
t = -32.0924, df = 27.306, p-value < 2.2e-16
alternative hypothesis: true difference in means is not
     equal to 0
95 percent confidence interval:
 -4162.115 -3662.134
sample estimates:
mean of x   mean of y
 1059.036  4971.161
```

T στατιστικό ελέγχου

Βαθμοί ελευθερίας της Student

P-τιμή

Συμμετρικό 95% ΔΕ για τον διαφορά των μέσων. Παρατηρήστε ότι δεν περιέχει το 0 και μπορούμε να συμπεράνουμε ότι $\mu_1 < \mu_2$

$\bar{y}$

$\bar{x}$

Δύο ανεξάρτητα δείγματα

- Η P-τιμή του ελέγχου είναι πολύ μικρή οπότε απορρίπτουμε την μηδενική υπόθεση. Με την βοήθεια του Δ.Ε. συμπεραίνουμε ότι η διάρκεια ζωής των λαμπτήρων πυρακτώσεως είναι κατά μέσο όρο αρκετά μικρότερη της διάρκειας ζωής των λαμπτήρων φθορίου.
- Αν υποθέσουμε ισότητα διασπορών τότε

Δύο ανεξάρτητα δείγματα

```
> t.test(X,Y, var.equal=TRUE)

        Two Sample t-test

data:  X and Y
t = -30.9836, df = 36, p-value < 2.2e-16
alternative hypothesis: true difference in means is not
     equal to 0
95 percent confidence interval:
 -4168.201 -3656.048
sample estimates:
mean of x mean of y
1059.036  4971.161
```

Δύο ανεξάρτητα δείγματα

- Και πάλι με το όρισμα conf.level μπορούμε να αλλάξουμε το ε.σ. του ελέγχου ενώ με το όρισμα alternative να εφαρμόσουμε μονόπλευρο έλεγχο.
- Όταν δεν ισχύει η υπόθεση της κανονικότητας και τα μεγέθη είναι μικρά χρησιμοποιούμε το αντίστοιχο μη παραμετρικό έλεγχο με την ονομασία Wilcoxon rank sum test.
- Ισχύουν και εδώ οι αντίστοιχες παρατηρήσεις για τις ισοπαλίες και τα μηδενικά που είχαμε στο έλεγχο ενός δείγματος. Σε τέτοιες περιπτώσεις μπορείτε και εδώ να χρησιμοποιήσετε την εντολή wilcox.exact από το πακέτο exactRankTests.
- Σας θυμίζω ότι το πακέτο exactRankTests όταν το φορτώσετε στην R, θα σας βγάλει ένα προειδοποιητικό μήνυμα ότι "is not longer under development". Παρόλα αυτά μπορείτε ακόμα να τρέξετε την εντολή wilcox.exact. Ενναλακτικά εδώ στα δύο ανεξάρτητα δείγματα μπορείτε να χρησιμοποιήσετε την εντολή wilcox_test από το πακέτο coin.

Δύο ανεξάρτητα δείγματα

```
> wilcox.test(X,Y)

        Wilcoxon rank sum test

data:  X and Y
W = 0, p-value = 5.956e-11
alternative hypothesis: true location
    shift is not equal to 0
```

Δύο ανεξάρτητα δείγματα

```
> library(exactRankTests)
 Package ‘exactRankTests’ is no longer under development.
 Please consider using package ‘coin’ instead.

> wilcox.exact(X,Y)

        Exact Wilcoxon rank sum test

data:  X and Y
W = 0, p-value = 5.956e-11
alternative hypothesis: true mu is not equal to 0
```

Δύο ανεξάρτητα δείγματα

```
> library(coin)
> data<-c(X,Y)
> group<-c(rep(1,length(X)), rep(2, length(Y)))
> group<-as.factor(group)
> group
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
Levels: 1 2
> wilcox_test(data~group, distribution = "exact")

        Exact Wilcoxon-Mann-Whitney Test

data:  data by group (1, 2)
Z = -5.2623, p-value = 5.956e-11
alternative hypothesis: true mu is not equal to 0
```

Δύο ανεξάρτητα δείγματα

```
> cbind(data,group)
          data     group
 [1,] 1101.0377     1
 [2,] 1786.3295     1
 [3,] 1277.6971     1
 [4,] 1109.6934     1
 [5,] 1197.7701     1
 [6,] 1686.8745     1
 [7,]  861.6086     1
 [8,]  133.7824     1
 [9,] 1261.7946     1
[10,]  457.8806     1
[11,]  400.9930     1
[12,]  796.5819     1
[13,] 1101.5362     1
[14,]  664.1593     1
[15,] 1235.7157     1
[16,] 1439.7528     1
[17,] 2140.8110     1
[18,]  688.0261     1
[19,]  700.0734     1
[20,] 1138.6054     1
[21,] 5275.4390     2
[22,] 4894.0540     2
[23,] 4992.2390     2
[24,] 4739.0010     2
[25,] 4946.2540     2
[26,] 4493.2450     2
[27,] 4842.2370     2
[28,] 5256.0850     2
[29,] 4780.9120     2
[30,] 4727.2350     2
[31,] 5357.2850     2
[32,] 5143.2490     2
[33,] 5020.9290     2
[34,] 4790.3500     2
[35,] 5013.4820     2
[36,] 5230.8480     2
[37,] 4929.0220     2
[38,] 5049.0230     2
```

Στην τελευταία εντολή τα δεδομένα πρέπει να είναι σε μακρά μορφή. Το group δηλώνει αν οι διάρκειες ζωής (data) αφορούν λαμπτήρες πυρακτώσεως (1) ή φθορίου (2) και πρέπει υποχρεωτικά να είναι factor.

Δύο ανεξάρτητα δείγματα

- Έλεγχος διαφοράς ποσοστών
  - Ας υποθέσουμε ότι έχουμε μια δίτιμη τυχαία μεταβλητή X με τιμές 0 και 1 και $P(X=1)=p_1$ (άγνωστο) και μία άλλη δίτιμη τυχαία μεταβλητή Y, ανεξάρτητη της X, με τιμές 0 και 1 και $P(Y=1)=p_2$ (άγνωστο). Προφανώς τότε η X και η Y ακολουθούν την κατανομή Bernoulli με παράμετρο p_1 και p_2 αντίστοιχα. Ενδιαφερόμαστε να ελέγξουμε την υπόθεση H_0: $p_1 = p_2$ έναντι της H_1: $p_1 \neq p_2$ σε ε.σ. α. Έστω $X_1,...,X_{n1}$ τυχαίο δείγμα από τον πρώτο πληθυσμό και $Y_1,...,Y_{n2}$ τυχαίο δείγμα από τον δεύτερο πληθυσμό.

Δύο ανεξάρτητα δείγματα

- Το στατιστικό ελέγχου μας τότε είναι

=0 με βάση την H_0

$$Z = \frac{\left(\hat{P}_1 - \hat{P}_2\right) - (p_1 - p_2)}{SE\left(\hat{P}_1 - \hat{P}_2\right)},$$

όπου $\hat{P}_1 = \bar{X}$ η σχετική συχνότητα της τιμής 1 στο $1^{ο}$ δείγμα και $\hat{P}_2 = \bar{Y}$ η σχετική συχνότητα της τιμής 1 στο $2^{ο}$ δείγμα.

Δύο ανεξάρτητα δείγματα

Κάτω από την μηδενική υπόθεση $p_1 = p_2 = p$ και τότε

$$Z = \frac{\left(\hat{P}_1 - \hat{P}_2\right)}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Δύο ανεξάρτητα δείγματα

- Με βάση το Κ.Ο.Θ., για μεγάλα μεγέθη δειγμάτων, το $Z \backsim N(0,1)$ κάτω από την μηδενική υπόθεση. Για τον υπολογισμό του τυπικού σφάλματος στον παρονομαστή εκτιμούμε την κοινή αναλογία p συνδυάζοντας τις πληροφορίες των δύο δειγμάτων, υπολογίζοντας δηλαδή την συγχωνευμένη (pooled) σχετική συχνότητα της τιμής 1 από τα δύο δείγματα

$$\tilde{P} = \frac{n_1 \hat{P}_1 + n_2 \hat{P}_2}{n_1 + n_2} = \frac{\sum_{i=1}^{n_1} X_i + \sum_{i=1}^{n_2} Y_i}{n_1 + n_2}.$$

- Συχνά στην προκειμένη περίπτωση για να είναι πιο ικανοποιητική η προσέγγισή μας προχωράμε όπως και στην περίπτωση του ενός δείγματος στην διόρθωση συνέχειας του Yates (Yates continuity correction). Το στατιστικό ελέγχου τότε γίνεται:

$$Z = \frac{\left|\hat{P}_1 - \hat{P}_2\right| - \left(\frac{1}{2n_1} + \frac{1}{2n_2}\right)}{\sqrt{\tilde{P}(1-\tilde{P})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Δύο ανεξάρτητα δείγματα

- Υπολογίζουμε την τιμή του Z με βάση τις παρατηρήσεις μας και με την βοήθεια του πίνακα της τυποποιημένης κανονικής κατανομής βρίσκουμε την P-τιμή. Αντίστοιχα μπορούμε να υπολογίσουμε το Z^2 το οποίο ακολουθεί την X^2 κατανομή με 1 βαθμό ελευθερίας και βρίσκουμε την P-τιμή με την βοήθεια του πίνακα της X^2 κατανομής με 1 βαθμό ελευθερίας.

- Ισοδύναμα με τον παραπάνω αμφίπλευρο έλεγχο θα μπορούσαμε να κατασκευάσουμε ένα συμμετρικό (1-α)% Δ.Ε. για το p_1-p_2 και αν δούμε αν η υποτιθέμενη τιμή της διαφοράς με βάση την μηδενική υπόθεση (το μηδέν δηλαδή) ανήκει στο εν λόγω διάστημα.

$$\left[\hat{p}_1-\hat{p}_2 \pm z_{\alpha/2}\left\{\frac{\hat{p}_1\left(1-\hat{p}_1\right)}{n_1}+\frac{\hat{p}_2\left(1-\hat{p}_2\right)}{n_2}\right\}^{\frac{1}{2}}\right].$$

Δύο ανεξάρτητα δείγματα

- Η P-τιμή για τον εν λόγω έλεγχο προκύπτει και πάλι με βάση την εναλλακτική υπόθεση:
  - Αν $H_1 : p_1 \neq p_2$ τότε η P-τιμή είναι 2 φορές η πιθανότητα δεξιά του $|Z|$ (ή ισοδύναμα 2 φορές η πιθανότητα αριστερά του $-|Z|$). Ισοδύναμα αν δουλέψουμε με το Z^2 η P-τιμή είναι η πιθανότητα δεξιά του Z^2.
  - Αν $H_1 : p_1 > p_2$ τότε η P-τιμή είναι πιθανότητα δεξιά του Z. Ισοδύναμα αν δουλέψουμε με το Z^2 η P-τιμή είναι το 1/2 της πιθανότητας δεξιά του Z^2.
  - Αν $H_1 : p_1 < p_2$ τότε η P-τιμή είναι πιθανότητα αριστερά του Z. Ισοδύναμα αν δουλέψουμε με το Z^2 η P-τιμή είναι ίση με το συμπλήρωμα του 1/2 της πιθανότητας δεξιά του Z^2.

Δύο ανεξάρτητα δείγματα

- Για να ισχύει το ΚΟΘ και όλα τα προηγούμενα θα πρέπει τα μεγέθη των δειγμάτων να είναι μεγάλα. Στην πράξη ελέγχουμε αν

$$n_1 \cdot \tilde{p} \geq 5 \quad \text{και} \quad n_2 \cdot \tilde{p} \geq 5$$

ΚΑΙ

όπου $$\tilde{p} = \frac{\sum_{i=1}^{n_1} x_i + \sum_{i=1}^{n_2} y_i}{n_1 + n_2}$$

$$n_1 \cdot (1 - \tilde{p}) \geq 5 \quad \text{και} \quad n_2 \cdot (1 - \tilde{p}) \geq 5$$

Δύο ανεξάρτητα δείγματα

- Παράδειγμα στην R. Δύο ανεξάρτητες παραγωγικές διαδικασίες έδωσαν 12 (k_1) και 20 (k_2) ελαττωματικά αντικείμενα σε τυχαία δείγματα των 300 (n_1) και 400 (n_2) αντικειμένων αντίστοιχα. Θέλουμε σε ε.σ. 5% να ελέγξουμε αν το ποσοστό των ελαττωματικών προϊόντων από τις 2 παραγωγικές διαδικασίες είναι το ίδιο (H_0: $p_1 = p_2$) έναντι της εναλλακτικής ότι δεν είναι το ίδιο (H_1: $p_1 \neq p_2$).

  Παρατηρούμε ότι

$$\hat{p}_1 = 12/300 = 0.04 \ \text{και} \ \hat{p}_2 = 20/400 = 0.05.$$

  Παρατηρούμε επίσης ότι

$$\tilde{p} = \frac{12+20}{300+400} = 0.046 \ \text{ενώ} \ n_i\tilde{p} \geq 5 \ \text{και} \ n_i(1-\tilde{p}) \geq 5, \ i = 1,2.$$

Δύο ανεξάρτητα δείγματα

- Ο έλεγχος στην R μπορεί να γίνει με την βοήθεια της εντολής prop.test.

```
> x<-c(12,20)
> n<-c(300,400)
> prop.test(x,n)

        2-sample test for equality of proportions with continuity
        correction

data:  x out of n
X-squared = 0.1972, df = 1, p-value = 0.657
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.04370433  0.02370433
sample estimates:
prop 1 prop 2
  0.04   0.05
```

Συχνότητες στα δείγματα της τιμής 1

Μέγεθος δειγμάτων

Διόρθωση συνέχειας

z^2

P – τιμή του ελέγχου

Συμμετρικό 95% ΔΕ για το p_1-p_2

$\hat{p}_1$ $\hat{p}_2$

Δύο ανεξάρτητα δείγματα

- Από τα αποτελέσματα του παραπάνω ελέγχου καταλήγουμε ότι, σε ε.σ. 5%, δεν έχουμε σοβαρές ενδείξεις εναντίον της μηδενικής υπόθεσης, οπότε δεν την απορρίπτουμε.
- Όπως και στην στην περίπτωση με ένα δείγμα μπορούμε να ζητήσουμε στην R να μην γίνει η διόρθωση συνέχειας καθώς επίσης και να αλλάξουμε την προκαθορισμένη τιμή του ε.σ. που και εδώ είναι 5%. Τέλος μπορούμε να ζητήσουμε μονόπλευρο έλεγχο.

Δύο ανεξάρτητα δείγματα

- Τα προηγούμενα δεδομένα μπορούσαμε να τα βλέπαμε υπό μορφή ενός 2×2 πίνακα συχνοτήτων (contingency table).

	# Ελαττωματικών	# Μη ελαττωματικών
Παραγωγική Διαδικασία 1	12	288
Παραγωγική Διαδικασία 2	20	380

Δύο ανεξάρτητα δείγματα

- Ισοδύναμα με τον προηγούμενο αμφίπλευρο έλεγχο θα ήταν να ελέγχαμε αν ο αριθμός των ελαττωματικών προϊόντων είναι ανεξάρτητος της παραγωγικής διαδικασίας (με εναλλακτική ότι δεν είναι). Σε τέτοιες περιπτώσεις υπολογίζουμε το στατιστικό ελέγχου

$$X^2 = \sum \frac{(\text{παρατηρηθείσες συχνότητες - αναμενόμενες συχνότητες})^2}{\text{αναμενόμενες συχνότητες}},$$

$$\text{όπου οι αναμενόμενες συχνότητες} = \frac{(\text{άθροισμα γραμμής}) \times (\text{άθροισμα στήλης})}{\text{μέγεθος δείγματος}}$$

και το άθροισμα είναι ως προς όλα τα κελιά.

Δύο ανεξάρτητα δείγματα

	# Ελαττωματικών	# Μη ελαττωματικών
Παραγωγική Διαδικασία 1	12 (**αναμενόμενη συχνότητα** = (32×300)/700 = 13.7)	288 (**αναμενόμενη συχνότητα** = (668×300)/700 = 286.3)
Παραγωγική Διαδικασία 2	20 (**αναμενόμενη συχνότητα** = (32×400)/700 = 18.3)	380 (**αναμενόμενη συχνότητα** = (668×400)/700 = 381.7)

Δύο ανεξάρτητα δείγματα

- Το στατιστικό ελέγχου X^2, κάτω από την μηδενική υπόθεση της ανεξαρτησίας, ακολουθεί προσεγγιστικά την X^2 κατανομή με 1 βαθμό ελευθερίας. Συχνά στην προκειμένη περίπτωση για να είναι πιο ικανοποιητική η προσέγγισή μας προχωράμε στην διόρθωση συνέχειας του Yates. Το στατιστικό ελέγχου τότε γίνεται:

$$X^2 = \sum \frac{(|\text{παρατηρηθείσες συχνότητες - αναμενόμενες συχνότητες}| - 0.5)^2}{\text{αναμενόμενες συχνότητες}}.$$

- Υπολογίζουμε λοιπόν την τιμή του στατιστικού ελέγχου με βάση τις παρατηρήσεις μας και εν συνεχεία η P-τιμή του ελέγχου είναι η πιθανότητα δεξιά της τιμής αυτής με βάση την X^2 κατανομή με 1 β.ε.

Δύο ανεξάρτητα δείγματα

- Απαραίτητη προϋπόθεση είναι όλες οι αναμενόμενες συχνότητες να είναι ≥ 5.
- Ο εν λόγω έλεγχος καλείται X^2 independence test.
- Μπορούμε να τον εφαρμόσουμε στην R με την βοήθεια της εντολής chisq.test αφού πρώτα γράψουμε τα δεδομένα μας υπό μορφή πίνακα συχνοτήτων.
- Μπορούμε να ζητήσουμε στην R να μην γίνει η διόρθωση του Yates με το όρισμα correct="false".

Δύο ανεξάρτητα δείγματα

```
> mash<-rbind(c(12,288),c(20,380))
> mash
     [,1] [,2]
[1,]   12  288
[2,]   20  380
> chisq.test(mash)

        Pearson's Chi-squared test with Yates' continuity
        correction

data:  mash
X-squared = 0.1972, df = 1, p-value = 0.657
```

Ίδια p-τιμή με πριν

Δύο ανεξάρτητα δείγματα

- Παρατηρούμε λοιπόν ότι σε ε.σ. 5%, δεν έχουμε σοβαρές ενδείξεις εναντίον της μηδενικής υπόθεσης της ανεξαρτησίας, οπότε δεν την απορρίπτουμε.

Δύο ανεξάρτητα δείγματα

- Όταν δεν ισχύουν οι προϋποθέσεις του Κ.Ο.Θ. (ισοδύναμα οι αναμενόμενες συχνότητες δεν είναι όλες μεγαλύτερες ή ίσες του 5) και παρόλα αυτά εμείς εφαρμόσουμε την εντολή prop.test ή την εντολή chisq.test παίρνουμε ένα προειδοποιητικό μήνυμα λάθους από την R.
- Σε αυτές τις περιπτώσεις εφαρμόζουμε το μη παραμετρικό Fisher exact test με την βοήθεια της εντολής fisher.test.
- Ας υποθέσουμε στο προηγούμενο παράδειγμα ότι $n_1 = 10$, $n_2 = 15$, $k_1 = 5$ και $k_2 = 10$. Τότε

$$\tilde{p} = \frac{5+10}{10+15} = 0.6 \text{ και } n_1(1-\tilde{p}) = 4 < 5.$$

Δύο ανεξάρτητα δείγματα

	# Ελαττωματικών	# Μη ελαττωματικών
Παραγωγική Διαδικασία 1	5 (**αναμενόμενη συχνότητα** = (10×15)/25 = 6)	5 (**αναμενόμενη συχνότητα** = (10×10)/25 = 4)
Παραγωγική Διαδικασία 2	10 (**αναμενόμενη συχνότητα** = (15×15)/25 = 9	5 (**αναμενόμενη συχνότητα** = (15×10)/25 = 6

Δύο ανεξάρτητα δείγματα

```
> x<-c(5,10)
> n<-c(10,15)
> prop.test(x,n)
        2-sample test for equality of proportions with continuity correction
data:  x out of n
X-squared = 0.1736, df = 1, p-value = 0.6769
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.6410844  0.3077510
sample estimates:
   prop 1    prop 2
0.5000000 0.6666667
Warning message:
In prop.test(x, n) : Chi-squared approximation may be incorrect
> mash<-rbind(c(5,5),c(10,5))
> chisq.test(mash)
        Pearson's Chi-squared test with Yates' continuity correction
data:  mash
X-squared = 0.1736, df = 1, p-value = 0.6769
Warning message:
In chisq.test(mash) : Chi-squared approximation may be incorrect
```

Δύο ανεξάρτητα δείγματα

```
> fisher.test(mash)

        Fisher's Exact Test for Count Data

data:  mash
p-value = 0.4422
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.07251472 3.43835655
sample estimates:
odds ratio
 0.5144931
```

P-τιμή

95% Δ.Ε. για τον λόγο συμπληρωματικών πιθανοτήτων

$$\frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)}$$ λόγος συμπληρωματικών πιθανοτήτων στο δείγμα

Δυο Εξαρτημένα Δείγματα

- Αρκετές φορές στις στατιστικές μελέτες συναντάμε το φαινόμενο των εξαρτημένων δειγμάτων. Π.χ. ας υποθέσουμε ότι έχουμε μετρήσεις της ίδιας ποσοτικής μεταβλητής για τα ίδια άτομα σε 2 διαφορετικές χρονικές περιόδους. Ας καλέσουμε Χ την μεταβλητή την χρονική τιμή 1 και ας θεωρήσουμε ότι προέρχεται από πληθυσμό με άγνωστη μέση τιμή μ_1 και άγνωστη τυπική απόκλιση σ_1 και Υ την μεταβλητή την χρονική στιγμή 2 και ας θεωρήσουμε ότι προέρχεται από πληθυσμό με άγνωστη μέση τιμή μ_2 και άγνωστη τυπική απόκλιση σ_2. Έστω το τυχαίο δείγμα που αποτελείται από ζεύγη συσχετισμένων τυχαίων μεταβλητών $(X_1,Y_1),...,(X_n,Y_n)$. Ενδιαφερόμαστε να δούμε αν η υπό μελέτη τυχαία μεταβλητή διαφοροποιείται κατά μέσο όρο στις 2 χρονικές περιόδους, δηλαδή να ελέγξουμε την υπόθεση H_0: $\mu_1 = \mu_2$ έναντι της H_1: $\mu_1 \neq \mu_2$ σε ε.σ. α. Ο εν λόγω έλεγχος καλείται *paired t-test*.

Δυο Εξαρτημένα Δείγματα

- Δημιουργούμε τις διαφορές των παραπάνω ζευγών $(Z_1=X_1-Y_1),\ldots,(Z_n= X_n-Y_n)$ και έτσι καταλήγουμε σε ένα τυχαίο δείγμα που προέρχεται από πληθυσμό με άγνωστη μέση τιμή $\mu_1-\mu_2$ και άγνωστη διασπορά $\sigma_1^2+\sigma_2^2-2\mathrm{Cov}(X,Y).$ Μπορούμε λοιπόν με βάση τα όσα είπαμε στην περίπτωση του ελέγχου για την μέση τιμή μιας ποσοτικής μεταβλητής (one sample t-test) να ελέγξουμε τώρα την υπόθεση H_0: $\mu_1 - \mu_2=0$ έναντι της H_1: $\mu_1 - \mu_2 \neq 0$.

Δυο Εξαρτημένα Δείγματα

- Παραδείγματα εξαρτημένων δειγμάτων:

  Μετρήσεις για τα ίδια άτομα

  I. σε 2 παρόμοιες μεταβλητές (π.χ. με ίδιες μονάδες μέτρησης).
  II. της ίδιας μεταβλητής στην ίδια μονάδα μελέτης αλλά σε διαφορετικές χρονικές στιγμές.
  III. της ίδιας μεταβλητής αλλά σε διαφορετικά σημεία της ίδιας μονάδας μελέτης.
  IV. της ίδιας μεταβλητής σε διαφορετικές μονάδες μελέτης που σχετίζονται (δίδυμα, συγγενείς, φίλοι).

Δυο Εξαρτημένα Δείγματα

- Παράδειγμα: Τα παρακάτω δεδομένα εκφράζουν τις εβδομαδιαίες πωλήσεις σε €, 20 προϊόντων πριν και μετά την διαφημιστική καμπάνια. Σε ε.σ. 5% θέλουμε να ελέγξουμε αν η διαφημιστική καμπάνια κατά μέσο όρο δεν επηρεάζει τις εβδομαδιαίες πωλήσεις με εναλλακτική ότι τις βελτιώνει.

Δυο Εξαρτημένα Δείγματα

Διαφημ. Καμπάνια	Εβδ. Πωλ. Πριν	Εβδ. Πωλ. Μετά	Διαφορά
1	220	313	-93
2	300	316	-16
3	390	400	-10
4	270	290	-20
5	510	504	6
6	330	370	-40
7	260	280	-20
8	200	200	0
9	210	230	-20
10	250	300	-50

Διαφημ. Καμπάνια	Εβδ. Πωλ. Πριν	Εβδ. Πωλ. Μετά	Διαφορά
11	300	370	-70
12	280	300	-20
13	330	390	-60
14	310	350	-40
15	400	420	-20
16	140	130	10
17	50	50	0
18	160	210	-50
19	250	300	-50
20	200	230	-30

Δυο Εξαρτημένα Δείγματα

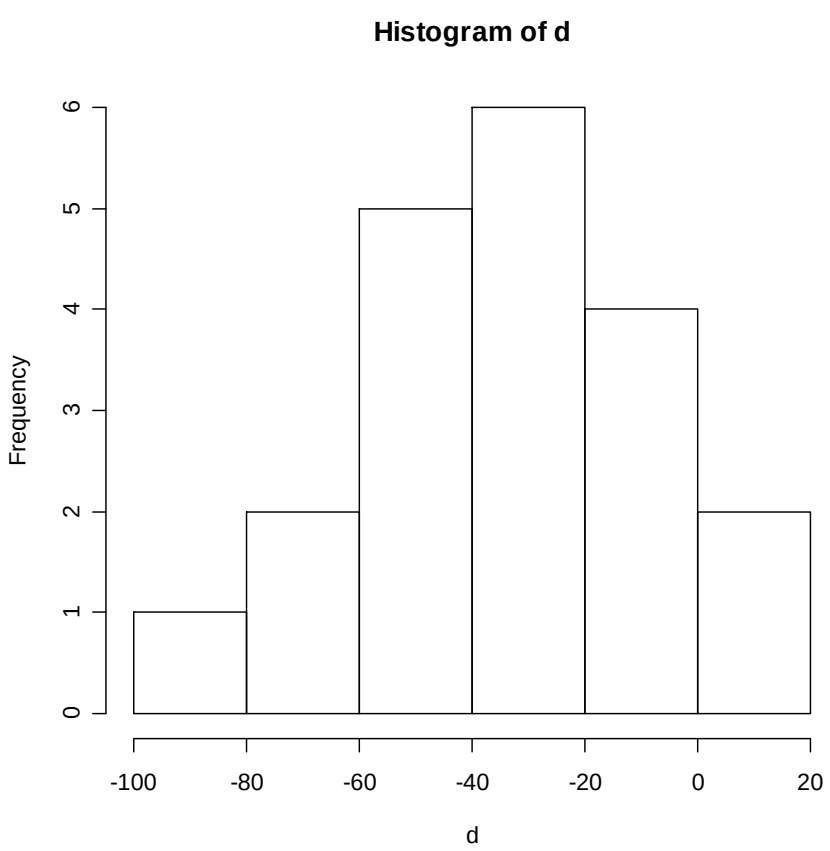
- Αν λοιπόν μ_1 είναι οι μέσες εβδομαδιαίες πωλήσεις πριν και μ_2 οι μέσες εβδομαδιαίες πωλήσεις μετά την διαφημιστική καμπάνια θέλουμε να ελέγξουμε την υπόθεση H_0: $\mu_1 - \mu_2 = 0$ έναντι της H_1: $\mu_1 - \mu_2 < 0$.

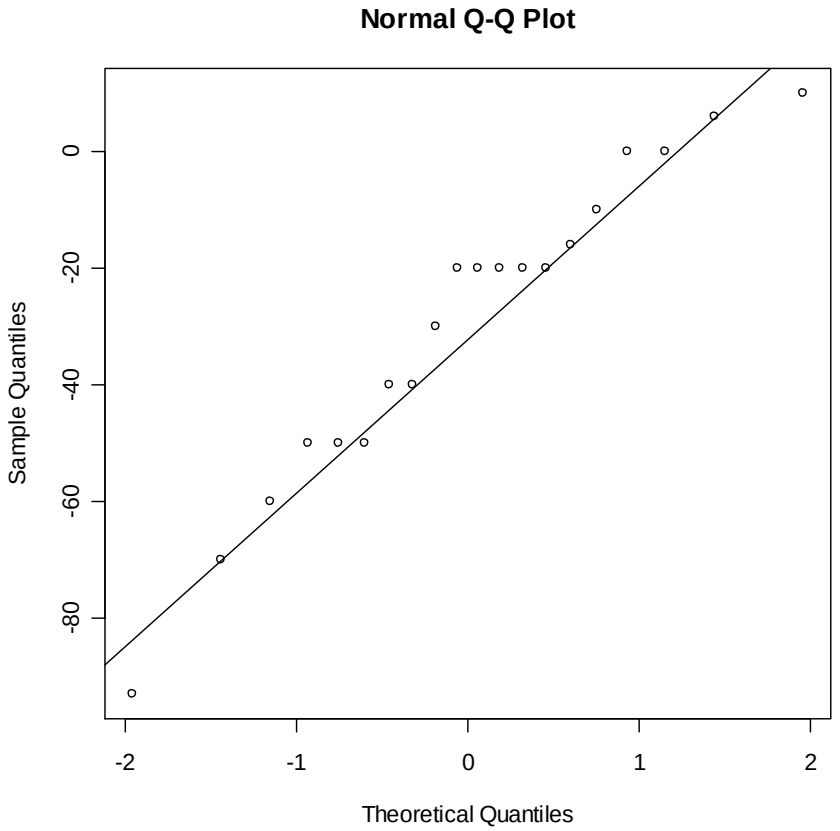
- Μιας και το μέγεθος του δείγματός μας δεν είναι τόσο μεγάλο στην αρχή ελέγχουμε αν η υπόθεση της κανονικότητας για τις διαφορές είναι λογική.

```
>d<-c(-93, -16, -10, -20, 6, -40, -20, 0, -20, -50, -70, -20, -60, -40, -
    20, 10, 0, -50, -50, -30)
>hist(d)
>qqnorm(d)
```

- Από τα επόμενα γραφήματα βλέπουμε ότι η υπόθεση της κανονικότητας δεν είναι παράλογη.

Δυο Εξαρτημένα Δείγματα





Δυο Εξαρτημένα Δείγματα

```
> t.test(d, mu=0, alternative="less")

        One Sample t-test
data:  d
t = -4.9455, df = 19, p-value = 4.490e-05
alternative hypothesis: true mean is less than 0
95 percent confidence interval
      -Inf -19.28334
sample estimates:
mean of x
   -29.65
```

$\left(-\infty,\ \bar{d} + t_{n-1,\alpha} s_d / \sqrt{n}\right)$

Δειγματικός μέσος των διαφορών

Δυο Εξαρτημένα Δείγματα

- Άρα σε ε.σ. 5% απορρίπτουμε την μηδενική υπόθεση ότι η διαφημιστική καμπάνια κατά μέσο όρο δεν επηρεάζει τις εβδομαδιαίες πωλήσεις και δεχόμαστε την εναλλακτική ότι υπήρξε αύξηση των πωλήσεων.

Δυο Εξαρτημένα Δείγματα

- Αν ισοδύναμα θέλαμε να συγκρίνουμε ποσοστά δύο συσχετισμένων δειγμάτων θα εφαρμόζαμε το McNemar's test με την βοήθεια της συνάρτησης mcnemar.test.

Δυο Εξαρτημένα Δείγματα

- Παράδειγμα:

Συμφωνείτε με τους χειρισμούς της κυβέρνησης στα θέματα οικονομικής πολιτικής;			
	Δεύτερη Έρευνα (1 βδομάδα μετά από συγκεκριμένα μέτρα)		
Πρώτη Έρευνα	Ναι	Όχι	Σύνολο
Ναι	10	15	25
Όχι	12	17	29
Σύνολο	22	32	54

Δυο Εξαρτημένα Δείγματα

```
> x<-rbind(c(10,15), c(12,17))
> x
     [,1] [,2]
[1,]   10   15
[2,]   12   17
> mcnemar.test(x)

        McNemar's Chi-squared test with continuity
    correction

data:  x
McNemar's chi-squared = 0.1481, df = 1, p-value =
    0.7003
```

Έλεγχος Καλής Προσαρμογής

- Μέχρι τώρα ο τρόπος που ελέγχαμε την καταλληλότητα του επιλεγμένου μοντέλου (Κανονικού στις περισσότερες περιπτώσεις) ήταν με την βοήθεια γραφικών παραστάσεων, π.χ. ιστογράμματα και QQ-plots.
- Υπάρχει ένας έλεγχος υποθέσεων (Kolmogorov-Smirnov test), κατά τον οποίον ελέγχουμε την μηδενική υπόθεση ότι τα δεδομένα ακολουθούν μια συγκεκριμένη κατανομή, με εναλλακτική ότι δεν την ακολουθούν. Με βάση λοιπόν την P-τιμή του εν λόγω ελέγχου φτάνουμε σε τελικά συμπεράσματα σε σχέση με την καταλληλότητα ή μη του μοντέλου (κατανομή) που έχουμε επιλέξει.
- Για τα δεδομένα της διάρκειας ζωής συγκεκριμένων λαμπτήρων φθορίου που είδαμε στο παράδειγμα με το ένα δείγμα έχουμε:

Έλεγχος Καλής Προσαρμογής

> ks.test(x,"pnorm", 2055, 233)

$\hat{\sigma}$

Δεδομένα x

$\hat{\mu}$

Ελέγχουμε αν προέρχονται από την Κανονική κατανομή

One-sample Kolmogorov-Smirnov test

Δεν απορρίπτουμε τον ισχυρισμό μας για Κανονικότητα

data: x

D = 0.0946, p-value = 0.9842

alternative hypothesis: two-sided

Έλεγχος Καλής Προσαρμογής

- Ένας άλλος έλεγχος καλής προσαρμογής αποκλειστικά για κανονικότητα είναι των Shapiro και Wilk.
- Έχει μεγαλύτερη ισχύ για συγκεκριμένο επίπεδο σημαντικότητας και επομένως προτιμάται.

```
> shapiro.test(x)

        Shapiro-Wilk normality test

data:  x
W = 0.9699, p-value = 0.753
```

Δεν απορρίπτουμε τον ισχυρισμό μας για Κανονικότητα