Ανάλυση Μιας Μεταβλητής (έλεγχοι)

με την R

Κλίμακες μέτρησης τυχαίων μεταβλητών

Τυχαία Μεταβλητή X (random variable): *Το χαρακτηριστικό που θέλουμε να μελετήσουμε.*

Ποσοτικές (quantitative)

Διαστήματος (interval) : θερμοκρασία, ηλικία *(δεν λέμε ότι μία σαραντάρα είναι ίση με δυο εικοσάρες).*

Αναλογίας (ratio) :ύψος, βάρος *(μπορούμε να πούμε ότι έχει διπλάσιο βάρος).*

Διακριτές – Συνεχείς (discrete) – (continuous)

Πλήθος ασθενών - βάρος

Ποιοτικές (qualitative)

Ονομαστικές (Nominal) : φύλο, ομάδα αίματος *(καμία μορφή διάταξης).*

Διάταξης (ordinal) : σειρά προτίμησης, αξιολόγηση της κατάστασης των ασθενών *(κακή, μέτρια, καλή, πολύ καλή). Διάταξη με ασαφείς αποστάσεις*

Τύποι υπολογισμού του n ή του h

Τύπος Sturge	$n = 1 + \log_2 N$	N πλήθος δεδομ.
Τύπος Doane	$n = 1 + \log_2 N + \log_2(1 + \hat{\gamma}\sqrt{N/6})$	$\hat{\gamma}$ εκτίμηση της κύρτωσης (χωρίς -3)
Τύπος Scott	$h = 3.5\hat{\sigma} N^{-1/3}$	$\hat{\sigma}$ εκτίμηση της τυπ. απόκλ.
Τύπος Freedman & Diaconis	$h = 2RN^{-1/3}$	R ενδοτεταρτη-μοριακό πλάτος

Περιγραφικά Στατιστικά Μέτρα

```
statistics <- function(x) {
  x1 <- na.omit(x)
  n <- length(x1)
  m <- mean(x1)
  s <- sqrt(var(x1))
  sk <- (sqrt(n/(n-1)^3)*sum((x1-m)^3))/s^3
  kr <- ((n/(n-1)^2)*sum((x1-m)^4))/s^4
  y <- c(n,m,s,sk,kr)
  names(y) <- c("N","mean","std", "skewness",
        "kurtosis") ; y }
```

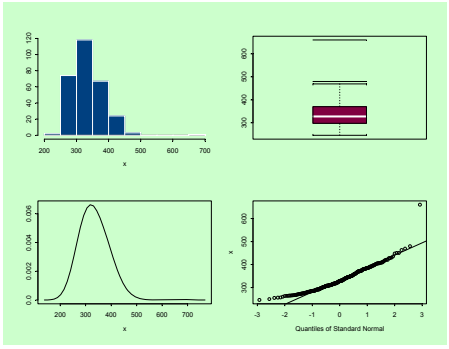
$$\lambda = \frac{\sqrt{n}\sum (x_i - \bar{x})^3}{\sqrt{\left[\sum (x_i - \bar{x})^2\right]^3}} \qquad \kappa = \frac{\nu \sum (x_i - \bar{x})^4}{\left[\sum (x_i - \bar{x})^2\right]^2}$$

Από το συντελεστή κύρτωσης κ συνήθως αφαιρείται το 3 και συγκρίνουμε με το 0 (η συνάρτηση kurtosis του Splus αφαιρεί το 3).

Άμεση συνοπτική περιγραφή μιας ποσοτικής μεταβλητής

Ορίζουμε μία συνάρτηση

```
eda.shape <- function(x) {
   par(mfrow=c(2,2))
   hist(x) ; boxplot(x)
   iqd <- summary(x)[5]-summary(x)[2]
   plot(density(x,width=2*iqd), xlab="x", ylab="", type="l")
   qqnorm(x)
   qqline(x)
   par(mfrow=c(1,1))
   invisible( )
   }
```



```
eda.shape(na.omit(time1000))
```

επειδή υπάρχουν ελλείπουσες τιμές

```
time1000[time1000 == "NA"]
```


Έλεγχος για τη μέση τιμή κανονικού πληθυσμού

t-test (σύνταξη)

```
t.test(x, y = NULL, alternative = c("two.sided",
   "less", "greater"), mu = 0, paired = FALSE,
   var.equal = FALSE, conf.level = 0.95, ...)
```

Παράδειγμα. Είναι μέση τιμή της time1000 ίση με τη διάμεσό της;

```
med<-median(time1000,na.rm=T)
t.test(time1000,mu=med)
```

($n \geq 30$)

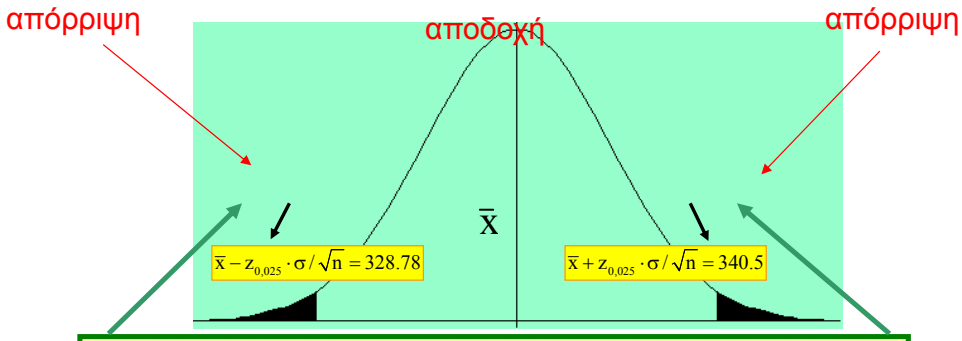
```
t = 2.2322, df = 288,
            p-value = 0.0264
alternative hypothesis: true
mean is not equal to 328
95 percent confid interval:
 328.7857 340.5015
sample estimates:
 mean of x
  334.6436
```

Απαραίτητη προϋπόθεση για να εφαρμόσουμε το t-test είναι να ακολουθεί η μεταβλητή κανονική κατανομή (έλεγχοι) ή να έχουμε μεγάλο δείγμα.

Η πιθανότητα εσφαλμένης απόρριψης της μηδενικής υπόθεσης.

(Αν p-value<a, όπου a συνήθως 0.05, απορρίπτουμε τη μηδενική υπόθεση σε στάθμη σημαντικότητας a. Η σ.σ. προϋπάρχει του ελέγχου.)

Μια γραφική ερμηνεία του ελέγχου



Η p-τιμή του ελέγχου είναι **0.0264** άρα είμαστε στην **απορριπτική περιοχή του τεστ** (τιμή ελέγχου 328, μεγάλο μέγεθος δείγματος ανιχνεύει μικρές διαφορές).

Έλεγχος Αναλογίας σε ένα δείγμα (binom.test)

Σύνταξη εντολής για έλεγχο διωνυμικής κατανομής

```
binom.test(x, n, p=0.5, alternative="two.sided")
```

Βρίσκει αν x επιτυχίες σε n δοκιμές προσαρμόζονται στην B(n,p)

Το αποτέλεσμα είναι λίστα με συνιστώσες τον αριθμό επιτυχιών , το πλήθος δοκιμών, το p.value κ.ά.

Παράδειγμα 2.36.

Σε 800 ρίψεις ενός νομίσματος εμφανίστηκαν 380 φορές «γράμματα». Είναι το νόμισμα κανονικό; Αν ήταν 360 ;

```
binom.test(380, 800, p=.5)$p.value
[1] 0.1678986
```

Δεν ΑΠΟΡΡΙΠΤΕΤΑΙ ότι είναι κανονικό

```
binom.test(360, 800)$p.value
[1] 0.005189051
```

ΑΠΟΡΡΙΠΤΕΤΑΙ ότι είναι κανονικό

Έλεγχος Αναλογιών (prop.test)

Σύνταξη εντολής για έλεγχο αναλογιών

```
prop.test(x, n, p, alternative="two.sided",conf.level=.95,
   correct=T)
```

x είναι το διάνυσμα επιτυχιών, n το διάνυσμα των δοκιμών, p το διάνυσμα των αντίστοιχων πιθανοτήτων

Αν δε δοθεί p, ελέγχεται η υπόθεση ότι όλες οι αναλογίες είναι ίσες.
Αν τα x και n είναι βαθμωτά(μονοδιάστατα) το αποτέλεσμα είναι ίδιο με αυτό της binom.test.

`prop.test(380,800,conf.level=.99)` **(για το παράδειγμα 2.36)**

```
1-sample proportions test with continuity correction
data:  380 out of 800, null probability 0.5
X-square = 1.9013, df = 1, p-value = 0.1679
alternative hypothesis: true P(success) in Group 1 is not
equal to 0.5
99 percent confidence interval:
 0.4292979 0.5211190
sample estimates:
 prop'n in Group 1
             0.475
```


Παραδείγματα

Είναι οι τιμές της μεταβλητής ypsos ισοκατανεμημένες γύρω από τη μέση τιμή της;

```
rypsos <- na.omit(ypsos)
fr <- sum(rypsos <=
   mean(rypsos))
prop.test(fr,length(rypsos))
```

```
.............................
X-square = 0.4933, df = 1,
          p-value = 0.4825

95 percent confidence interval:
 0.4724631 0.5600462
sample estimates:
 prop'n in Group 1
         0.5163776
```

Είναι το πλήθος παιδιών από κάθε νομό στο talenta.df ανάλογο του συνολικού πληθυσμού;

```
pop<-c(36797,96554,139934,
   946864,135937,52685,
   81710,150386,138741,
   116763,192828,53147,92117)

fr<-as.numeric(table(nomos))
prop.test(fr,pop)
```

```
.............................
X-square = 35.0618, df = 12,
           p-value = 0.0005

sample estimates:
0.0001902329      0.0002899932
.....................
```

Έλεγχος για τη διασπορά του πληθυσμού

Ο έλεγχος στηρίζεται στο ότι αν ένα δείγμα προέρχεται από κανονικό πληθυσμό, τότε η μηδενική υπόθεση H_0: $\sigma^2=\sigma_0^2$ με εναλλακτική H_1: $\sigma^2>\sigma_0^2$ απορρίπτεται, όταν ισχύει η συνθήκη

$$\hat{\chi}^2 = \frac{\sum(x_i-\bar{x})^2}{\sigma_0^2} = \frac{(n-1)s^2}{\sigma_0^2} > \hat{\chi}^2_{n-1;\alpha}$$

όταν μ άγνωστη

$$\hat{\chi}^2 = \frac{\sum(x_i-\mu)^2}{\sigma_0^2} = \frac{n\,s^2}{\sigma_0^2} > \hat{\chi}^2_{n;\alpha}$$

όταν μ γνωστή

Στην περίπτωση που η μ είναι <u>άγνωστη</u>, το 100(1-α)% διάστημα εμπιστοσύνης για τη διασπορά είναι:

$$\frac{(n-1)s^2}{\hat{\chi}^2_{n-1;\alpha/2}} \le \sigma^2 \le \frac{(n-1)s^2}{\hat{\chi}^2_{n-1;1-\alpha/2}}$$

Η συνάρτηση var1.test

```
var1.test <- function(x, sigma, mu = NULL, a = 0.05)  {
        nax <- na.omit(x)  ;    n <- length(nax)
        if(missing(mu)) { mu <- mean(nax)  ;      n <-  n - 1     }
        csq <- (num <- sum((nax - mu)^2))/sigma^2
        cr.chi <- qchisq(1 - a, n)
        ret.val <- list(Chi.square = csq, Significance = a,   "Critical
                        value" =   cr.chi,    Inference = "Reject the null
                        hypothesis if Chi.square  is GREATER than
                        the Critical value")
        if(missing(mu)) { lb <- num/qchisq(1 - a/2, n)
                        ub <- num/qchisq(a/2, n)
                        ret.val <- c(ret.val, list("Confidence
                        Interval" = c(lb, ub)))  }
        return(ret.val)
}
```

```
x<- c(40, 60, 60, 70, 50, 40, 50, 30 )
var1.test(x,60)
```

```
var1.test(ypsos,35)
```

Κριτήριο χ^2 προσαρμογής για διακριτά δεδομένα.

Κατεβάζουμε το πακέτο vcd που έχει τη συνάρτηση godfit()

goodfit(x, type = c("poisson", "binomial", "nbinomial"), method = c("ML", "MinChisq"), par = NULL)

```
dummy <- rnbinom(200, size = 1.5, prob = 0.8)
gf <- goodfit(dummy, type = "nbinomial", method =
  "MinChisq")
summary(gf)
plot(gf)
```

Αν γνωρίζουμε παραμέτρους χρησιμοποιούμε τη συνάρτηση chisq.test()

chisq.test(x, y = NULL, correct = TRUE, p = rep(1/length(x), length(x)), rescale.p = FALSE, simulate.p.value = FALSE, B = 2000)

π.χ p=(0.11639097, 0.34943563, 0.29057144, 0.15393984, 0.08966212)

x=c(26, 76, 56, 25, 15)

```
chisq.test(x,p) ## chi-square test
```


Κριτήριο Kolmogorov-Smirnov 1/2

Σύνταξη εντολής ks.test που προσαρμόζει ένα δείγμα x σε γνωστή κατανομή, ή συγκρίνει δύο δείγματα

```
ks.test(x, y, ...,
   alternative = c("two.sided","less","greater"),
        exact = NULL)
```

x: a numeric vector of data values.

y: either a numeric vector of data values, or a character string naming a cumulative distribution function or an actual cumulative distribution function such as pnorm.
Only continuous CDFs are valid.

Κριτήριο Kolmogorov-Smirnov 2/2

Βρίσκει την εμπειρική συνάρτηση κατανομής $F_E(x)$ που προκύπτει από το δείγμα και ελέγχει πόσο «απέχει» από τη θεωρητικής σ.κ. $F_0(x)$

δίπλευρο τεστ	$ks = \sup_x \left\{ \left\| F_0(x) - F_E(x) \right\| \right\}$	$F_E(x_{(k)}) = \frac{k}{n}, k = 1, \ldots, n$

Μονό-πλευρα	$ks^+ = \sup_x \left\{ F_E(x) - F_0(x) \right\}$	$ks^- = \sup_x \left\{ F_0(x) - F_E(x) \right\}$

Αν δεν δώσουμε y πρέπει να δώσουμε την κατανομή (αν είναι διαφορετική από την κανονική) και τις ειδικές παραμέτρους της, για προσαρμογή ενός δείγματος στην κατανομή αυτή.

Αν δώσουμε y τότε θα γίνει σύγκριση δύο δειγμάτων.

Έλεγχοι τυχαιότητας

Μέση τετραγωνική διαδοχική διαφορά

Neymann 1941

$$\Delta^2 = \frac{1}{n-1}\left((x_1 - x_2)^2 + (x_2 - x_3)^2 + \ldots + (x_{n-1} - x_n)^2 \right), \quad \frac{\Delta^2}{s^2} \approx 2$$

```
mssd <- function(x) {
 y <- diff(x) ;   n <- length(x)
  mssd <- sum(y^2)/(n - 1)
  z <- c(3.09, 2.326, 1.645)
  bound <- 2 - 2*z*sqrt((n-2)/(n^2-1))
  names(bound) <- c("0.1%", "1%", "5%")
  param <- c(mssd, var(x), mssd/var(x))
  names(param) <- c("Delta squared",
        "Variance", "Delta/Var")
  par<-list(parameters = param,
       "Upper bound for large n" = bound)
  par1<-list(parameters = param)
  ifelse(n>22,return(par),return(par1))
}
```

```
> v1<-c(3,4,5,10,13)
> mssd(v1)
$parameters:
 Delta squared Variance Delta/Var
             9    18.5    0.4864865
```

```
> v2 <- c(3,4,13,10,5)
> mssd(v2)
$parameters:
 Delta squared Variance Delta/Var
            29    18.5   1.567568
```

```
> mssd(na.omit(baros))
$parameters:                                   $"Upper bound for large n":
 Delta squared Variance Delta/Var        0.1%        1%        5%
      125.8801   62.0033  2.030216    1.729251 1.796193 1.855863
```

Κάτω φράγματα του λόγου Δ^2/s^2

n	0.10%	1%	5%		n	0.10%	1%	5%
4	0.5898	0.6256	0.7805		14	0.6223	0.8931	1.1816
5	0.4161	0.5379	0.8204		15	0.6532	0.9221	1.2053
6	0.3634	0.5615	0.8902		16	0.6826	0.9491	1.2272
7	0.3695	0.6140	0.9359		17	0.7104	0.9743	1.2473
8	0.4036	0.6628	0.9825		18	0.7368	0.9979	1.2660
9	0.4420	0.7088	1.0244		19	0.7617	1.0199	1.2834
10	0.4816	0.7518	1.0623		20	0.7852	1.0406	1.2996
11	0.5197	0.7915	1.0965		21	0.8073	1.0601	1.3148
12	0.5557	0.8280	1.1276		22	0.8283	1.0785	1.3290
13	0.5898	0.8618	1.1558		...	...	...	...

Αν ο παρατηρούμενος λόγος είναι μικρότερος από την τιμή του πίνακα τότε η υπόθεση της τυχαιότητας απορρίπτεται στην αντίστοιχη σ.σ.

Για μεγαλύτερα n χρησιμοποιείται ως κάτω φράγμα η τιμή $2-2\cdot z\cdot\sqrt{\frac{n-2}{n^2-1}}$

Κριτήριο ροών (runs test) ή των Wald-Wolfowitz

Εφαρμόζεται σε διχοτομικά δεδομένα π.χ. 0, 1

r οι ροές (διαδοχές ίδιων συμβόλων)

n_1 τα 0 και n_2 τα 1

```
runs.test <- function(x, m = NULL) {
  if(missing(m))  m <- median(x)
  n <- length(x)
  rbar <- 1 + sum(abs(diff(x >= m)))
  s <- sum(x >= m) ; f <- n - s
  denom <- sqrt((2*f*s*(2*f*s-n))/(n-1))
  zbar <- abs(n * (rbar - 1) –
              2 * s * f)/denom
  p.value.2sided <- 2 * (1-pnorm(zbar))
  counts <- c(f, s,rbar)
  names(counts) <- c("fails",
   "successes", "runs")
  stats <- c(zbar, p.value.2sided)
  names(stats) <- c("zbar",
        "p.value (2-sided)")
  list(counts = counts, stats = stats)
}
```

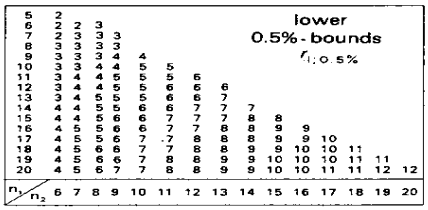
Για πλήθη $n_1, n_2 \le 10$ ο πίνακας IX σελίδα 478 του βιβλίου Στατιστική των Μπόρα & Κολυβά, δίνει τις πιθανότητες $P(U\le k_i)$. Η υπόθεση απορρίπτεται αν $r\le k_1$ ή $r\ge k_2$, όπου $P(U\le k_1)= P(U \ge k_2)=a/2$

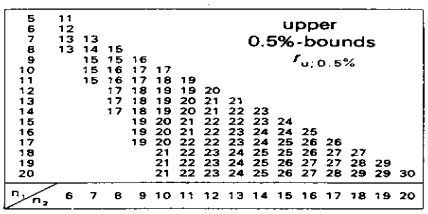
Για μεγαλύτερα n χρησιμοποιούμε το στατιστικό

$$z=\frac{\left|n(r-1)-2n_1n_2\right|}{\sqrt{\frac{2n_1n_2(2n_1n_2-n)}{n-1}}},\quad n=n_1+n_2$$

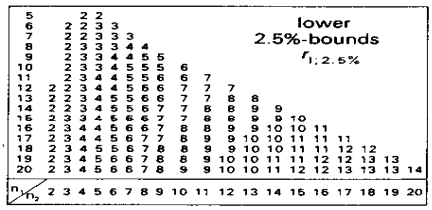
και η μηδενική υπόθεση απορρίπτεται όταν $|z|>z_{a/2}$

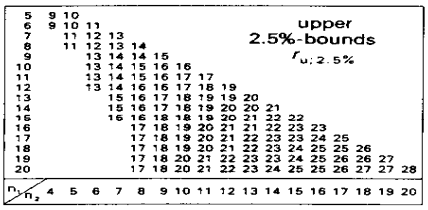
P = 0.01



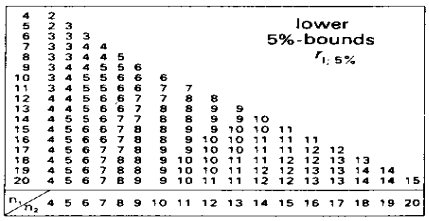


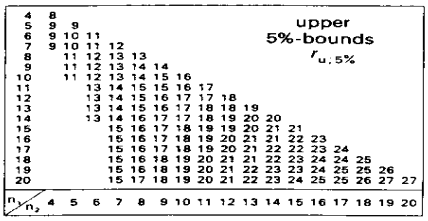
P = 0.05





P = 0.10





$n_1 = n_2 = n$

n	P = 0,10	P = 0,05	P = 0,02	P = 0,01	n	P = 0,10	P = 0,05	P = 0,02	P = 0,01
20	15–27	14–28	13–29	12–30	60	51– 71	49– 73	47– 75	46– 76
21	16–28	15–29	14–30	13–31	61	52– 72	50– 74	48– 76	47– 77
22	17–29	16–30	14–32	14–32	62	53– 73	51– 75	49– 77	48– 78
23	17–31	16–32	15–33	14–34	63	54– 74	52– 76	50– 78	49– 79
24	18–32	17–33	16–34	15–35	64	55– 75	53– 77	51– 79	49– 81
25	19–33	18–34	17–35	16–36	65	56– 76	54– 78	52– 80	50– 82
26	20–34	19–35	18–36	17–37	66	57– 77	55– 79	53– 81	51– 83
27	21–35	20–36	19–37	18–38	67	58– 78	56– 80	54– 82	52– 84
28	22–36	21–37	19–39	18–40	68	58– 80	57– 81	54– 84	53– 85
29	23–37	22–38	20–40	19–41	69	59– 81	58– 82	55– 85	54– 86
30	24–38	22–40	21–41	20–42	70	60– 82	58– 84	56– 86	55– 87
31	25–39	23–41	22–42	21–43	71	61– 83	59– 85	57– 87	56– 88
32	25–41	24–42	23–43	22–44	72	62– 84	60– 86	58– 88	57– 89
33	26–42	25–43	24–44	23–45	73	63– 85	61– 87	59– 89	57– 91
34	27–43	26–44	24–46	23–47	74	64– 86	62– 88	60– 90	58– 92
35	28–44	27–45	25–47	24–48	75	65– 87	63– 89	61– 91	59– 93
36	29–45	28–46	26–48	25–49	76	66– 88	64– 90	62– 92	60– 94
37	30–46	29–47	27–49	26–50	77	67– 89	65– 91	63– 93	61– 95
38	31–47	30–48	28–50	27–51	78	68– 90	66– 92	64– 94	62– 96
39	32–48	30–50	29–51	28–52	79	69– 91	67– 93	64– 96	63– 97
40	33–49	31–51	30–52	29–53	80	70– 92	68– 94	65– 97	64– 98
41	34–50	32–52	31–53	29–55	81	71– 93	69– 95	66– 98	65– 99
42	35–51	33–53	31–54	30–56	82	71– 95	69– 97	67– 99	66–100
43	35–53	34–54	32–56	31–57	83	72– 96	70– 98	68–100	66–102
44	36–54	35–55	33–57	32–58	84	73– 97	71– 99	69–101	67–103
45	37–55	36–56	34–58	33–59	85	74– 98	72–100	70–102	68–104
46	38–56	37–57	35–59	34–60	86	75– 99	73–101	71–103	69–105
47	39–57	38–58	36–60	35–61	87	76–100	74–102	72–104	70–106
48	40–58	38–60	37–61	36–63	88	77–101	75–103	73–105	71–107
49	41–59	39–61	38–62	36–64	89	78–102	76–104	74–106	72–108
50	42–60	40–62	38–64	37–65	90	79–103	77–105	74–108	73–109
51	43–61	41–63	39–65	38–66	91	80–104	78–106	75–109	74–110
52	44–62	42–64	40–66	39–67	92	81–105	79–107	76–110	75–111
53	45–63	43–65	41–67	40–68	93	82–106	80–108	77–111	75–113
54	45–65	44–66	42–68	41–69	94	83–107	81–109	78–112	76–114
55	46–66	45–67	43–69	42–70	95	84–108	82–110	79–113	77–115
56	47–67	46–68	44–70	42–72	96	85–109	82–112	80–114	78–116
57	48–68	47–69	45–71	43–73	97	86–110	83–113	81–115	79–117
58	49–69	47–71	46–72	44–74	98	87–111	84–114	82–116	80–118
59	50–70	48–72	46–74	45–75	99	87–113	85–115	83–117	81–119
60	51–71	49–73	47–75	46–76	100	88–114	86–116	84–118	82–120

```
> a1
 [1] 1 0 0 1 0 0 1 0 1 1 1 0 0 1 0 0 1 0 1 0 0 1 0 0 0
> runs.test(a1,.5)
$counts:                              $stats:
 fails successes runs                 zbar      p.value (2-sided)
    15        10   16                  1.279204          0.2008251
```

```
> a2<- c("H","H","H","T","H","H","T","T")
> runs.test(as.numeric(factor(a2)),1.5)
$counts:
 fails successes runs
     5         3    4
```

```
> runs.test(na.omit(time1000))
$counts:
 fails successes runs
   144       145  127

$stats:
     zbar p.value (2-sided)
 2.180081         0.02925143
```

```
> runs.test(y<-na.omit(time1000),
    mean(y))
$counts:
 fails successes runs
   160       129  137

$stats:
      zbar p.value (2-sided)
 0.8152092          0.4149526
```


Κριτήριο προσήμου διαφορών ή των Wallis-Moore

Εφαρμόζεται σε διαδοχικές (στο χρόνο ή στο χώρο) μετρήσεις.

Αν οι μετρήσεις είναι ανεξάρτητες (από το χρόνο ή την απόσταση) τότε θα κυμαίνονται γύρω από μια μεσαία τιμή. Έτσι οι διαδοχικές διαφορές θα εναλλάσσουν πρόσημο τυχαία. Αν h (phase) είναι το πλήθος ροών των προσήμων χωρίς την πρώτη και την τελευταία, τότε το στατιστικό z παρακάτω, ακολουθεί την κανονική κατανομή.

$$\hat{z} = \frac{\left| h - \frac{2n-7}{3} \right| - 0.5}{\sqrt{\frac{16n-29}{90}}}, \quad \gamma\iota\alpha \;\; n > 10 \qquad \acute{\eta} \qquad \hat{z} = \frac{\left| h - \frac{2n-7}{3} \right|}{\sqrt{\frac{16n-29}{90}}}, \quad \gamma\iota\alpha \;\; n > 30$$

Δεδομένα		5 6 2 3 5 6 4 3 7 8 9 7 5 3 4 7 3 5 6 7 8 9
Πρόσημα		+ - + + + - - + + + - - - + + - + + + + +
Φάσεις		1 2 3 4 5 6 7
Δηλαδή	**h=7**	οπό $\hat{z}$ = 2.56 άρα η ανεξαρτησία ΑΠΟΡΡΙΠΤΕΤΑΙ σε α=0.05

Ορίσατε συνάρτηση στην R που να υπολογίζει το $\hat{z}$

Shapiro-Wilk τεστ κανονικότητας

Το τεστ Shapiro - Wilk (W-statistic) ελέγχει την προσαρμογή τυχαίου δείγματος στην **κανονική κατανομή (είναι ισχυρότερο από το Kolmogorov-Smirnov όταν N<50)**. Έχει μια απλή γραφική επεξήγηση: μπορεί να θεωρηθεί ως προσεγγιστικό μέτρο της συσχέτισης των ποσοστιαίων σημείων από την κανονική κατανομή με τα ποσοστιαία σημεία από τα δεδομένα.

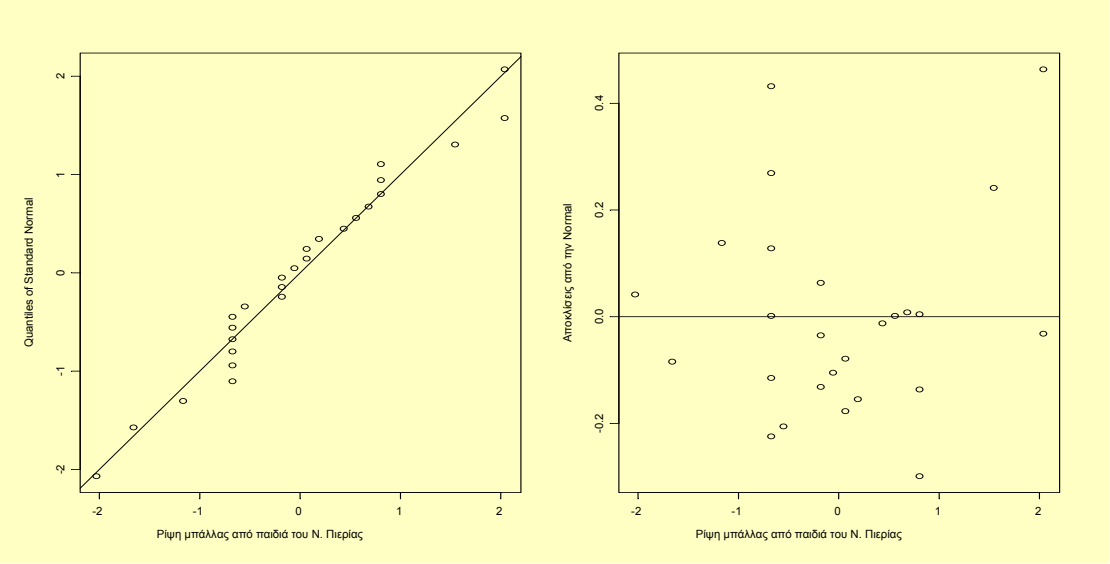
Ορίζεται $$W = \left(\sum_{i=1}^{n} a_i x_{(i)} \right)^2 \Big/ \sum_{i=1}^{n} (x_{(i)} - \bar{x})^2$$

όπου $$\mathbf{a}' = \frac{\mathbf{m}' V^{-1}}{\sqrt{\mathbf{m}' V^{-1} V^{-1} \mathbf{m}}}$$ και $\mathbf{m}=Ey_{(i)}$, y_i τυχ. καν. τιμές

Σύνταξη εντολής του S-Plus `shapiro.test(x)`

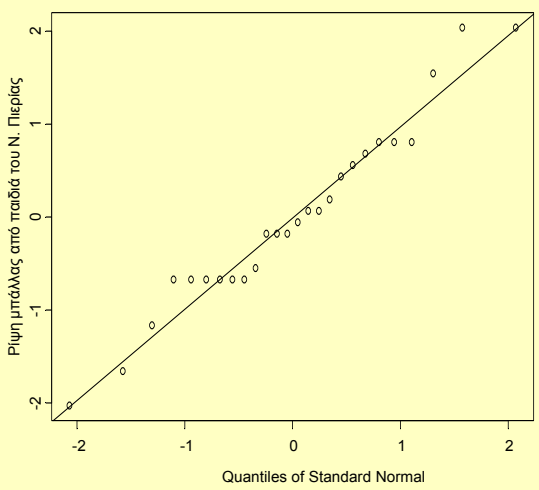
Γραφικοί έλεγχοι προσαρμογής

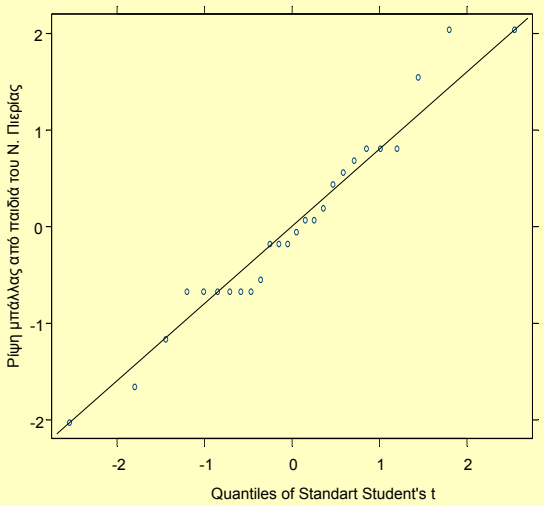
QQ-Plot και αποκλίσεις από την κανονική κατανομή για τη μεταβλητή ball στο νομό Πιερίας.



Γραφικοί έλεγχοι προσαρμογής

QQ-Plots με την qqnorm και την qqmath για τη μεταβλητή ball στο νομό Πιερίας.





Ανάλυση Δύο Μεταβλητών

με την R

Μωυσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

Περιγραφή δύο ποιοτικών μεταβλητών

Πίνακες Συνάφειας

Οι πίνακες συνάφειας σχηματίζονται όταν οι μεταβλητές μας εκφράζουν **ποιοτικά** χαρακτηριστικά ή όταν έχουν **ομαδοποιηθεί**. Στο S-Plus συνίσταται να τις ορίσουμε ως παράγοντες (factors).

```
Pet <- factor(c("Cat","Dog","Cat","Dog","Cat","?"),
exclude="?")
Food <- factor(c("Dry","Dry","Dry","Wet","Wet","Wet"))
table(Pet) ; table(Food) ;  PetFood<-table(Pet, Food)
```

Δεν συμπεριλαμβάνει την κατηγορία "?" στη δημιουργία factor

```
Cat Dog
  3   2
```

```
Dry Wet
  3   3
```

```
    Dry Wet
Cat   2   1
Dog   1   1
```

Εκτελέστε τις εντολές χωρίς να τις κάνετε factor.

Συσσωρευμένο Ραβδόγραμμα

Πίνακας συνάφειας

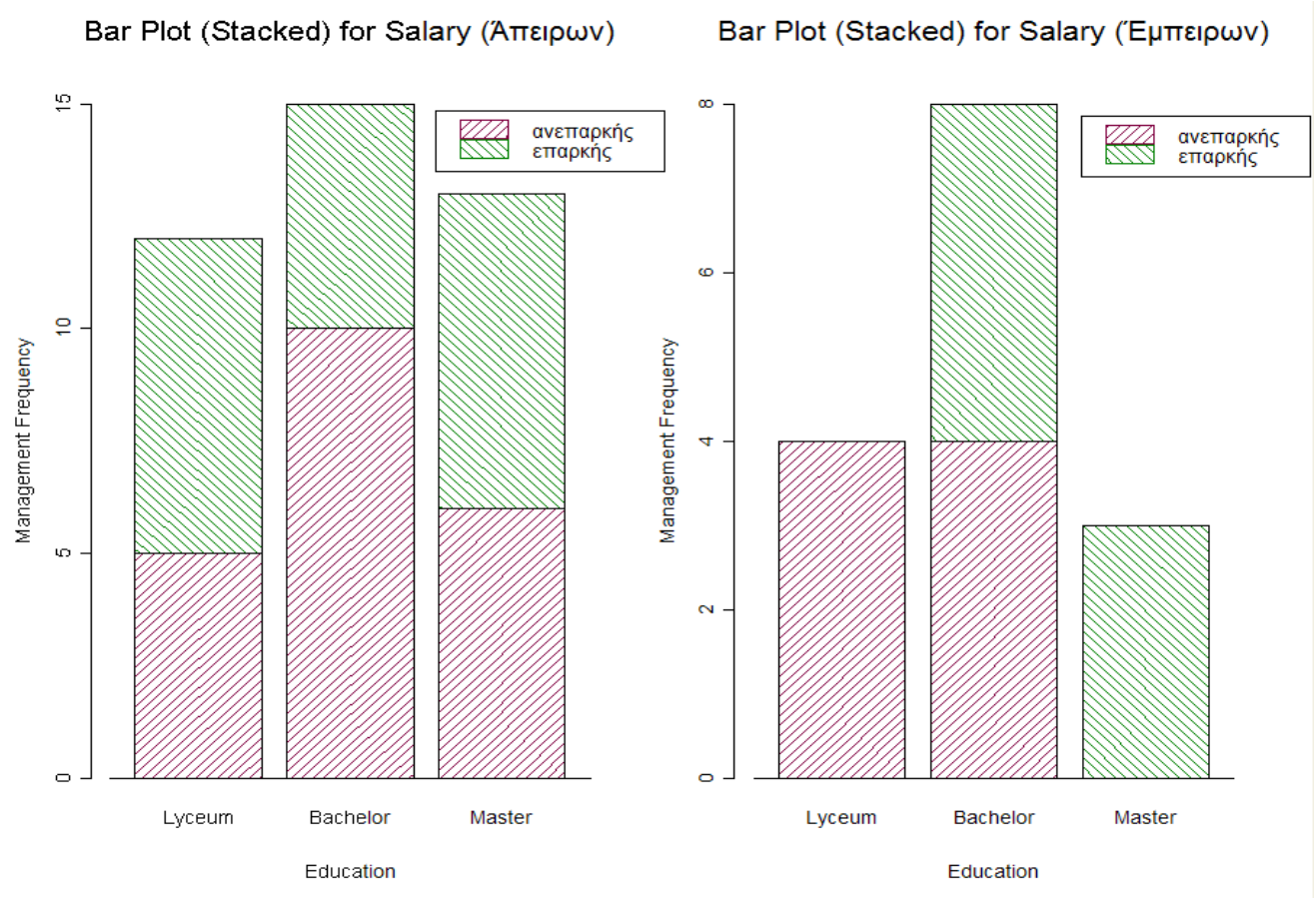
όσες οι γραμμές του πίνακα συνάφειας

```
par(mfrow=c(1,2))
barplot(tab[,,1], angle=seq(45,135,len=2),
density=16,col=3:4,
    names=c("Lyceum","Bachelor","Master"),
xlab="Education",
    ylab="Management Frequency")
legend(locator(1), legend = c("ανεπαρκής","επαρκής"),
    col=3:4, angle=seq(45,135,len=2), density=16)
  title("Bar Plot (Stacked) for Salary (Άπειρων)")
```

```
barplot(tab[,,2], angle=seq(45,135,len=2),
density=16,col=3:4,
    names=c("Lyceum","Bachelor","Master"),
xlab="Education",
    ylab="Management Frequency")
legend(locator(1), legend = c("ανεπαρκής","επαρκής"),
    col=3:4, angle=seq(45,135,len=2), density=16)
  title("Bar Plot (Stacked) for Salary (Έμπειρων)")
par(mfrow=c(1,1))
```


Συσσωρευμένο Ραβδόγραμμα

Μωυσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ



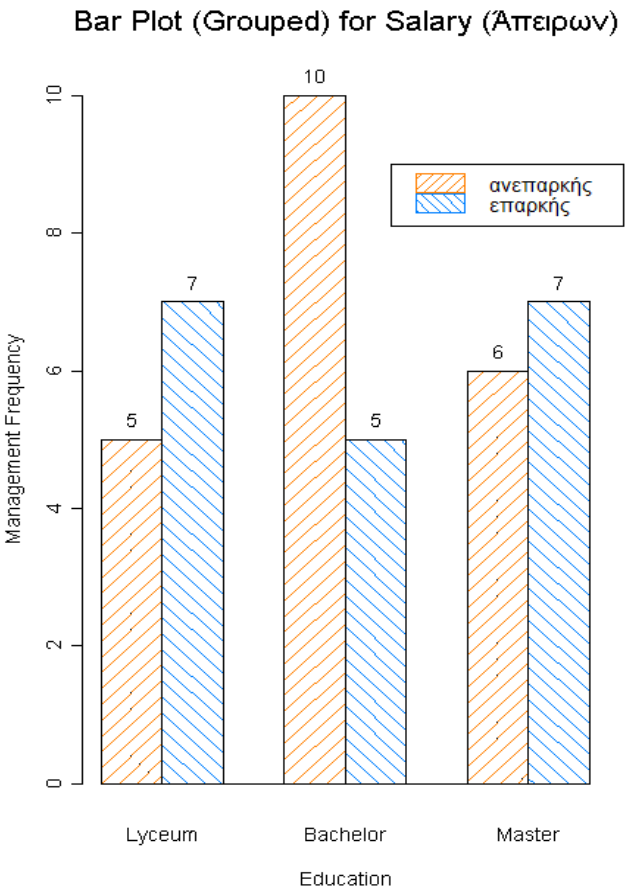
Ομαδοποιημένο Ραβδόγραμμα

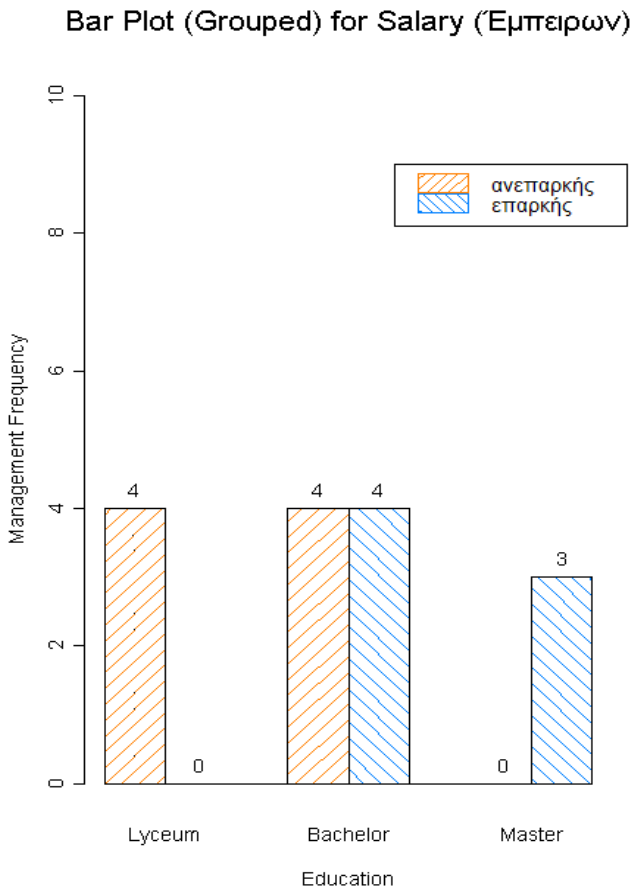

```
par(mfrow=c(1,2))
  x<-barplot(tab[,,1], beside=T, width=2, space=sp,
     angle=rep(c(45,135),3), col=rep(5:6,3),
     density=10, xlab="Education", ylim=c(0,10),
     ylab="Management Frequency")
  x1=c(x[1]+x[2],x[3]+x[4],x[5]+x[6])/2
  text(x1,-0.7,c("Lyceum","Bachelor","Master"))
legend(9.5,9,c("ανεπαρκής","επαρκής"),angle=c(45,135),
     density=16,col=5:6)
  ht<-as.numeric(tab[,,1])
  text(x,ht+.3,ht)
  title("Bar Plot (Grouped) for Salary (Άπειρων)")
```

```
  z<-barplot(tab[,,2], beside=T, width=2, space=sp,
     angle=rep(c(45,135),3), col=rep(5:6,3),
     density=10, xlab="Education",ylim=c(0,10),
     ylab="Management Frequency")
  ……..
  title("Bar Plot (Grouped) for Salary (Έμπειρων)")
par(mfrow=c(1,1))
```


Ομαδοποιημένο Ραβδόγραμμα

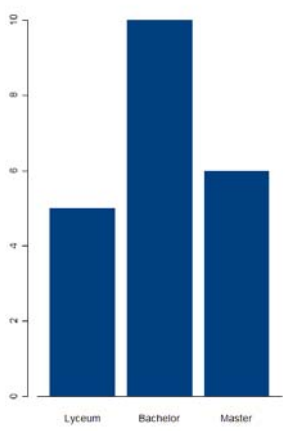
Μωυσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

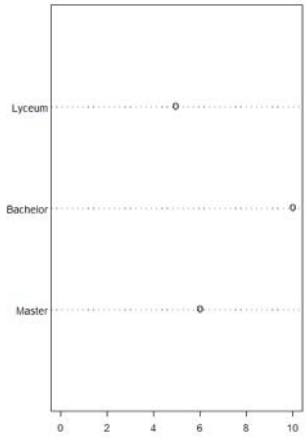




Σημειογράφημα

Μωυσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ






```
par(mfrow=c(1,2))
barplot(tab[1,,1],names=c("L
yceum","Bachelor","Master"))
dotchart(tab[1,,1],names=c("
Lyceum","Bachelor","Master")
    , xlim=c(0,10))
par(mfrow=c(1,1))
dotchart(tab[1,,1],names=c("
Lyceum","Bachelor","Master")
 ,ylim=c(0,10),horiz=F)
```

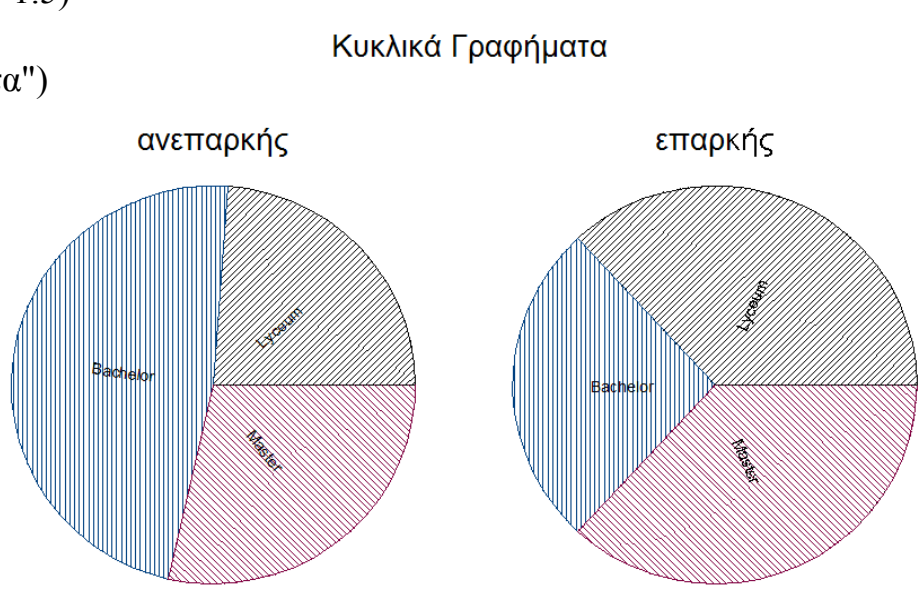
Ομαδοποιημένο σημειογράφημα

```
educ.fac<- factor(col(tab[,,1]),
label=c("Lyceum","Bachelor",
"Master"))
dotchart(tab[,,1],  labels=
c("ανεπαρκής","επαρκής"),
  group=educ.fac)
title("Σημειογράφημα για
  Salary (Άπειρων)")
```

Κυκλικά γραφήματα

Μωυσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

```
par(mfrow=c(1,2))
educn=c("Lyceum","Bachelor","Master")
pie(tab[1,,1], names=educn, angle= seq(45,135, len=3), size=1, density=20)
text(0,2.7,"ανεπαρκής",cex=1.5)
pie(tab[2,,1], names=educn, angle= seq(45,135, len=3), size=1, density=20)
text(0,2.7,"επαρκής",cex=1.5)
par(mfrow=c(1,1))
title("Κυκλικά Γραφήματα")
```



Έλεγχος ομοιογένειας σε πίνακα συνάφειας

```
chisq.test(x, y=NULL, correct=T)
```

Ελέγχουμε αν οι άνδρες και οι γυναίκες καπνίζουν με την ίδια αναλογία τους τρεις τύπους τσιγάρων.

Φύλο	Τσιγ 1	Τσιγ 2	Τσιγ 3
Άνδρες	36	10	14
Γυναίκες	12	9	19

```
prefer.mat <- rbind(c(36,10,14),c(12,9,19))
```

```
chisq.test(prefer.mat)
```

```
X-square = 9.1773,
df = 2,
p-value = 0.0102
```

p.value<0.05 άρα απορρίπτουμε τη μηδενική υπόθεση των ίδιων αναλογιών ως προς τα δύο φύλα

Για να είναι αξιόπιστο το συμπέρασμα του ελέγχου, θα πρέπει το ποσοστό των κελιών με αναμενόμενες συχνότητες <5 να μην ξεπερνά το 20%

```
fisher.test(prefer.mat)
```

```
p-value = 0.0098
```


Η εντολή xtabs

Η σύνταξη της εντολής είναι:

```
xtabs(formula = ~., data = parent.frame(), subset, sparse
= FALSE, na.action, exclude = c(NA, NaN),
drop.unused.levels = FALSE)
```

π.χ.

```
B=xtabs(~manag+educ+cexper);B
```

```
B=xtabs(~manag+educ+cexper);B
, , cexper = 0

      educ
manag  1  2  3
    0  5 10  6
    1  7  5  7

, , cexper = 1

      educ
manag  1  2  3
    0  4  4  0
    1  0  4  3
```

```
> apply(B,c(1,2),sum)
      educ
manag 1  2  3
    0 9 14  6
    1 7  9 10
> apply(B,c(1,2),mean)
      educ
manag   1   2 3
    0 4.5 7.0 3
    1 3.5 4.5 5
```

Μωυσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

Επί μέρους αθροίσματα - Περιθώριες Συχνότητες

apply(tab,1,sum) →

apply(tab,2,sum) →

apply(tab,3,sum) →

ανεπαρκής επαρκής
29 26
Lyceum Bachelor Master
16 23 16
άπειρος έμπειρος
40 15

$\sum_{j,k} a_{ijk}$

Αν θεωρήσουμε ως α_{ijk} τα στοιχεία του πίνακα tab

i= *1(=ανεπαρκής), 2(=επαρκής)*

j= 1(=Lyceum),2(= Bachelor),3(= Master)

k= *1(=άπειρος), 2(=έμπειρος)*

η `apply(tab,1,sum)` *παίρνει για i σταθερό το*

Συσχέτιση μεταξύ ποιοτικών μεταβλητών

1. Συντελεστής Συνάφειας του Pearson

Ορίζεται: $$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

με τιμές από 0 (ανεξάρτητες μεταβλητές)
μέχρι μία μέγιστη τιμή (<1) (εξαρτημένες μεταβλητές)
με σημαντικότητα p-value αυτήν του χ^2.

πχ. για n=4, και α>=0.01
είναι C <0.87

$$C = \sqrt{\frac{\chi^2_{n;a}}{\chi^2_{n;a} + n}}$$

```
sqrt(qchisq(1-a,4)/(qchisq(1-a,4)+4))
```

π.χ. στο salary για τους άπειρους η ΕΠΑΡΚΕΙΑ και ΕΚΠΑΙΔΕΥΣΗ

```
h<-chisq.test(tab[,,1])
csq=as.numeric(h$s)# το s είναι το στατιστικό X2
n=sum(tab[,,1])
cc.pearson=sqrt(csq/(csq+n));cc.pearson;h$p.value
```

CC του Pearson 0.2172739 με σημαντικότητα 0.37

2. Συντελεστές phi (ϕ) και V του Cramer

Ο συντελεστής *phi (φ)* $\varphi = \sqrt{\chi^2 / n}$

ορίζεται μόνο σε 2x2 πίνακες συνάφειας με τιμές από 0 (ανεξάρτητες μεταβλητές) μέχρι 1 (απόλυτη συνάφεια).

phi=sqrt(csq/n);phi

φ= 0.2225914 με σημαντικότητα 0.37

Ο συντελεστής ***V του Cramer*** $V = \sqrt{\frac{\chi^2}{n \cdot \min\{r-1, c-1\}}}$

επεκτείνει τον συντελεστή phi (φ) σε μεγαλύτερους πίνακες συνάφειας με τιμές από 0 (ανεξ. Μεταβλ.) μέχρι 1 (απόλυτη συνάφεια).

v.cramer=sqrt(csq/(n*min(d[1],d[2])));v.cramer

V του Cramer 0.1573959 με σημαντικότητα 0.37

3. Συντελεστής phi (ϕ) για 2×2 πίνακες και τ_b του Kendall

Ο συντελεστής *phi (φ)* σε 2×2 πίνακα συνάφειας (δίπλα) ταυτίζεται με την απόλυτη τιμή του τ_b του Kendall, και δίνεται ισοδύναμα από τη σχέση:

$$\varphi = \frac{x_{11}x_{22} - x_{12}x_{21}}{\sqrt{n_1 n_2 m_1 m_2}}$$

x_{11}	x_{12}	m_1
x_{21}	x_{22}	m_2
n_1	n_2	n

```
t2=tab[,2,]
n=sum(t2)
h<-chisq.test(t2,correct=F)
csq=as.numeric(h$s)
phi=sqrt(csq/n);phi
ms=apply(t2,1,sum);ms
ns=apply(t2,2,sum);ns
frac=(t2[1,1]*t2[2,2]-t2[1,2]*t2[2,1])/sqrt(
     prod(ms)*prod(ns));frac
```

Πρέπει να μη γίνει η διόρθωση συνέχειας του Yates για να είναι ίδιο το αποτέλεσμα

```
> phi
[1] 0.16265
> frac
[1] 0.16265
```

Περιγραφή δύο ποσοτικών μεταβλητών Ανεξάρτητα δείγματα

```
attach(salary.eu)
feduc=factor(educ, labels=c("Lyceum","Bachelor","Master"))
boxplot(salary~feduc)
title(main="Boxplot της salary για τα διαφορετικά επίπεδα μόρφωσης")
#ή διαφορετικά
boxplot(split(salary,feduc))
```

Τα ανεξάρτητα δείγματα σε ένα φύλλο δεδομένων δίνονται σε ένα διάνυσμα (αφορούν το ίδιο χαρακτηριστικό) και διαχωρίζονται από τις στάθμες κάποιου παράγοντα.

Αριθμητικά μέτρα για δύο ή περισσότερα ανεξάρτητα δείγματα

```
fmanag_as.factor(manag)
levels(fmanag)<-c("ανεπαρκής«,"επαρκής")
tapply(salary,fmanag,mean)

feduc_as.factor(educ)
levels(feduc)<-
c("Lyceum","Bachelor","Master")
tapply(salary,feduc,mean)

fcexper_as.factor(cexper)
levels(fcexper)<-c("άπειρος","έμπειρος")
tapply(salary,fcexper,mean)
```

```
ανεπαρκής επαρκής
  569.1348  829.9035

Lyceum Bachelor   Master
590.197  734.246  734.474

άπειρος έμπειρος
633.6547 849.0807
```

```
> tapply(salary, fmanag, statistics)
$insufficient:
  N      mean       std  skewness kurtosis
 29 569.1348 100.9995 0.6648002 2.386397

$sufficient:
  N      mean       std   skewness kurtosis
 26 829.9035 167.7446 -0.2714809 2.137568
```

Συγκεντρωτικά μέτρα για την περιγραφή δύο ανεξάρτητων δειγμάτων

Μέσες τιμές για συνδυασμούς δύο ή περισσοτέρων χαρακτηριστικών

tapply(salary,list(fmanag,feduc),mean)

```
              Lyceum Bachelor  Master
ανεπαρκής 577.7022 587.0657 514.445
  επαρκής 606.2629 963.1933 866.491
```

tapply(salary,list(fmanag,feduc,fcexper),stdev)

```
, , άπειρος
              Lyceum Bachelor   Master
ανεπαρκής 63.31079 56.10010 48.51005
  επαρκής 67.38650 46.95309 65.07193

, , έμπειρος
              Lyceum Bachelor   Master
ανεπαρκής 77.65505 41.42206       NA
  επαρκής       NA 40.15421 72.90357
```


Έλεγχος μέσων τιμών δύο δειγμάτων από κανονικούς πληθυσμούς

Ανεξάρτητα Δείγματα

Ο έλεγχος της ισότητας ή μη των μέσων τιμών δύο ανεξάρτητων δειγμάτων, που προέρχονται από κανονικούς πληθυσμούς με ίσες (αλλά άγνωστες συνήθως) διασπορές γίνεται με την εντολή

```
> t.test(x, y=NULL, alternative="two.sided",mu=0,
     paired=F, var.equal=T, conf.level=.95)
```

με άνισες (αλλά άγνωστες συνήθως) διασπορές γίνεται με την εντολή

```
> t.test(x, y=NULL, alternative="two.sided",mu=0,
     paired=F, var.equal=F, conf.level=.95)
```

Έλεγχος ισότητας διασπορών δύο ανεξάρτητων δειγμάτων από κανονικούς πληθυσμούς

Για τον έλεγχο της ισότητας διασπορών χρησιμοποιούμε την επόμενη εντολή, η οποία επιπλέον καθορίζει την επιλογή ενός t.test που πρόκειται να εφαρμόσουμε

```
> var.test(x, y, alternative="two.sided",
        conf.level=.95)
```


Παράδειγμα

Είναι το μέσο ύψος των κοριτσιων ίσο με το μέσο ύψος των αγοριών για τα δεδομένα του αρχείου talenta;

```
# Προϋποθέσεις
ypsF<-na.omit(ypsos[sex=="F"])
ypsM<-na.omit(ypsos[sex=="M"])
length(ypsF);length(ypsM)
eda.shape(ypsF);eda.shape(ypsM)
ks.gof(ypsF)
ks.gof(ypsM)
var.test(ypsF,ypsM)

# Έλεγχος
t.test(ypsF, ypsM,
alternative="two.sided", mu=0,
paired=F,var.equal=T,
conf.level=.95)
```

```
p-value = 0.5
p-value = 0.0652
p-value = 0.4369
```

Άρα δέχομαι ότι οι 2 κατηγορίες ακολουθούν κανονική κατανομή (που είναι προϋπόθεση του ελέγχου)

Δέχομαι ότι οι διασπορές είναι ίσες

```
Standard Two-Sample t-Test

data:  ypsF and ypsM
 = 1.2621, df = 517, p-value = 0.2075
alternative hypothesis: true difference in
means is not equal to 0
95 percent confidence interval:
 -0.4220525  1.9386300
sample estimates:
 mean of x mean of y
  141.5765  140.8182
```

Περιγραφή δυο ποσοτικών μεταβλητών

Ζευγαρωτές Παρατηρήσεις - Εξαρτημένα Δείγματα

Οι βαθμολογίες 10 αθλητών από τους Κριτές Α και Β παρουσιάζονται στον παρακάτω πίνακα

Κριτής Α	28.7	26.2	24.8	25.3	25.1	23.9	26.1	25.8	30.3	31.4
Κριτής Β	25.4	25.8	24.9	25.0	23.9	23.3	26.6	24.8	28.8	30.3

Μωυσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

```
boxplot(list(judge.A,judge.B),names=c("Judge A","Judge B"))
title(main="Κατανομή των βαθμολογιών",cex=1.3)
```

Με τις εντολές:

summary,

statistics

και

full.summary

παίρνουμε τα αντίστοιχα αριθμητικά μέτρα.

Έλεγχος μέσων τιμών δύο δειγμάτων από κανονικούς πληθυσμούς

Ζευγαρωτές Παρατηρήσεις - Εξαρτημένα Δείγματα

Κριτής Α	28.7	26.2	24.8	25.3	25.1	23.9	26.1	25.8	30.3	31.4
Κριτής Β	25.4	25.8	24.9	25.0	23.9	23.3	26.6	24.8	28.8	30.3

Διαφέρει σημαντικά η βαθμολογία που οι δύο κριτές δίνουν στους δέκα αθλητές;

Μωυσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

Μικρό πλήθος παρατηρήσεων,
το test έδειξε κανονικότητα

```
shapiro.test(judge.A)
shapiro.test(judge.B)
```

```
t.test(judge.A,
judge.B, paired = T)
```

```
        Paired t-Test

data:  judge.A and judge.B
t = 2.6512, df = 9, p-value = 0.0264
alternative hypothesis: true mean of
differences is not equal to 0
95 percent confidence interval:
 0.1291213 1.6308787
sample estimates:
 mean of x - y
          0.88
```

Μη παραμετρικός Έλεγχος για δείγματα που δεν προέρχονται από κανονική κατανομή

Ο παρακάτω έλεγχος χρησιμοποιείται και για μεταβλητές διάταξης (**ordinal**)

```
wilcox.test(x, y, alternative="two.sided",
 mu=0, paired=F, exact=T, correct=T)
```

Με `paired=F` το χρησιμοποιούμε για ανεξάρτητα δείγματα και είναι ισοδύναμο με το test των Mann-Whitney

Με `paired=T` το χρησιμοποιούμε για ζευγαρωτές παρατηρήσεις και είναι το test του Wilcoxon

Χρησιμοποιούμε την επιλογή `exact=T` για μικρά δείγματα

Παράδειγμα(1)

Ανεξάρτητα δείγματα

Σε δείγμα 7 ποντικών εφαρμόσαμε τη δίαιτα Α και μετα από ένα μήνα είχαν τα παρακάτω βάρη

```
diet.A<-c(76.1,  72.4,  76.2,  74.3,  77.4,  78.4,  76.0)
```

ενώ σε ένα άλλο δείγμα 9 ποντικών εφαρμόσαμε στο ιδιο διάστημα τη δίαιτα Β και πήραμε τα παρακάτω βάρη

```
diet.B<-c(79.1,  73.0,  77.3,  79.1,  80.0,  79.1,  79.1,  77.3,  80.2)
```

Υπάρχει διαφορά μεταξύ των δύο τύπων δίαιτας;

```
shapiro.test(diet.A)
shapiro.test(diet.B)

wilcox.test(diet.A,
   diet.B, paired = F)
```

```
        Wilcoxon rank-sum test

data:  diet.A and diet.B
rank-sum normal statistic with
correction Z = -2.2411, p-value = 0.025
alternative hypothesis: true mu is not
equal to 0
```

Παράδειγμα (2)

Ζευγαρωτές παρατηρήσεις

Μετρήσαμε το χρόνο που χρειάζεται ένας άνθρωπος τις 6 εργάσιμες ημέρες μιας εβδομάδας για να πάει από το σπίτι του στη δουλειά του επιλέγοντας δύο διαφορετικές διαδρομές

```
route.A<-c(76.1,  72.4,  76.2,  74.3,  77.4,  78.4)
route.B<-c(79.1,  73.0,  77.3,  79.1,  80.0,  79.1)
```

Υπάρχει διαφορά ως προς το χρόνο ανάμεσα στις δύο διαδρομές που επέλεξε;

```
shapiro.test(route.A)

shapiro.test(route.B)


wilcox.test(route.A,route.
B,paired=T)
```

```
        Exact Wilcoxon signed-rank test

data:  route.A and route.B
signed-rank statistic V = 0, n = 6,
 p-value = 0.0312
alternative hypothesis: true mu is
not equal to 0
```


Μωυσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

Γραμμική Παλινδρόμηση

με την R

Μωυσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

Δεδομένα των Sams και Shadman (1986)

Για τη μελέτη απόδοσης σε φυσικό αέριο κοιτασμάτων άνθρακα έγινε ένα πείραμα στο οποίο μετρήθηκε η απόδοση (y) σε σχέση με την περιεκτικότητα σε άνθρακα (x) 22 δειγμάτων.

Μωυσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

```
KperC <- c(0.05, 0.05, 0.25, 0.25, 0.50, 0.50, 0.50,
     1.25, 1.25, 1.25, 1.25, 1.25, 2.10, 2.10,
     2.10, 2.10, 2.10, 2.10, 2.50, 2.50, 2.50, 2.50)
CO.des <- c(0.05, 0.10, 0.25, 0.35, 0.75, 0.85, 0.95,
     1.42, 1.75, 1.82, 1.95, 2.45, 3.05, 3.19,
     3.25, 3.43, 3.50, 3.93, 3.75, 3.93, 3.99, 4.07)
samsad <- data.frame(CO.des,KperC)
```

Συσχέτιση δύο μεταβλητών

Στις διάφορες έρευνες, μας ενδιαφέρει συχνά η ένταση της σχέσης μεταξύ ποσοτικών μεταβλητών (π.χ. μας ενδιαφέρει όταν αυξάνεται η μία, αν αυξάνεται ή ελαττώνεται η άλλη και με ποιο τρόπο) .

Η πιο απλή μορφή σχέσης στα μαθηματικά είναι η γραμμική σχέση (y=a+bx).
Ο συντελεστής συσχέτισης r του Pearson μετρά τη γραμμική συσχέτιση δύο ποσοτικών μεταβλητών. Παίρνει τιμές στο [-1,1].

- **Τιμές κοντά στο 1 δηλώνουν ισχυρή θετική γραμμική συσχέτιση.**
- **Τιμές κοντά στο -1 δηλώνουν ισχυρή αρνητική γραμμική συσχέτιση.**
- **Τιμές κοντά στο 0 δηλώνουν ανυπαρξία γραμμικής συσχέτισης.**

```
cor(KperC,CO.des)
```

Η συνάρτηση cor.test κάνει έλεγχο σημαντικότητας της συσχέτισης. Για να είναι αξιόπιστα τα αποτελέσματα του ελέγχου αυτού, θα πρέπει οι μεταβλητές να ακολουθούν κανονική κατανομή.

H_0: Ο συντ. συσχέτισης R=0
H_1: Ο συντ. συσχέτισης R≠0

```
cor.test(KperC,CO.des)
```

Αν δεν ακολουθούν κανονική κατανομή, καταλληλότερος είναι ο συντελεστής συσχέτισης του Spearman.

```
cor.test(KperC,CO.des,method="spearman")
```


Απλή Γραμμική Παλινδρόμηση
Αντικείμενα lm

```
lm(formula,
   data="πλαίσιο δεδομένων",
   weights="διάνυσμα",
   subset="έκφραση",
   na.action=na.fail,
   method="qr", model=T,
   x=F, y=F, qr=F,
   contrasts=NULL, ...)
```

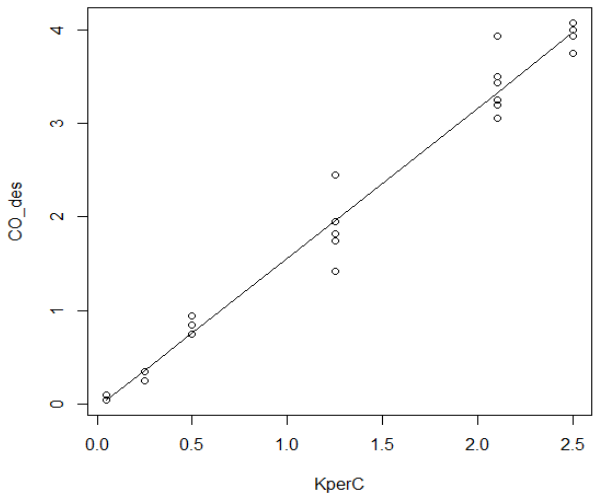
y~x1+x2+…
(Πρώτα η Υ!!!)

για σταθμισμένη παλινδρόμηση

για περαιτέρω μελέτη

Παλινδρόμηση με τα δεδομένα samsad

```
plot(KperC,CO.des, xlab="KperC",ylab="CO_des")
samsad.lm <- lm(CO.des ~ KperC)
opar <- par(mfrow = c(2,2), oma = c(0, 0, 1.1, 0))
plot(samsad.lm, las = 1)
par(opar)
anova(samsad.lm)
summary(samsad.lm)
```



πίνακας ANOVA

συντελεστές παλινδρόμησης, προσδιορισμού, τυπικά σφάλματα κλπ.

Τα βήματα της γραμμικής παλινδρόμησης αναλυτικά (1/2)

```
x <- KperC
y <- CO.des
X <- cbind(1,x)
XtX <- t(X) %*% X
XtXin <- solve(XtX)
XtXin
XtY <- t(X) %*% y
bhat <- XtXin %*% XtY
n <- length(x)
YtY <- sum(y^2)
sY1 <- (sum(y))^2/n
ssr <- t(bhat) %*% XtY - sY1
```

Σχηματισμός του πίνακα σχεδιασμού X

Υπολογισμός του πίνακα $X'X$

Υπολογισμός του πίνακα $(X'X)^{-1}$

Υπολογισμός του πίνακα $X'Y$

Εκτίμηση συντελεστών παλινδρόμησης

Το πλήθος παρατηρήσεων

Το άθροισμα τετραγώνων $Y'Y$

Το άθροισμα τετραγώνων $(Y'1)^2/n$

Το άθροισμα τετραγώνων SSR

$$SSR = \hat{\beta}' X' Y - \frac{1}{n}\left(Y' 1\right)^2$$

$$\hat{\beta} = \left(X' X\right)^{-1} X' Y$$

Τα βήματα της γραμμικής παλινδρόμησης αναλυτικά (2/2)

```
sst <- YtY-(sum(y))^2/n          Το άθροισμα τετραγώνων SST
c(ssr,sst,sst-ssr)               Το άθροισμα τετραγώνων SSE
sse <- YtY-t(bhat) %*% XtY       Το άθροισμα τετραγώνων SSE απευθείας

mse <- sse/(n-2)                 Το άθροισμα τετραγώνων MSE (k=1=>n-k-1=n-2)
ste <- sqrt(mse); c(mse, ste)    Τα MSE και s
V.mat <- mse[1,1]*XtXin          Πίνακας διασπορών – συνδιασπορών

x0 <- c(1, 2.05)                 Διάνυσμα για πρόβλεψη
y0hat <- t(x0)%*% bhat           Πρόβλεψη
vary0 <- mse * t(x0) %*% XtXin%*% x0
vary0[1,1]                       Var(Y0)
F.ratio <- ssr/mse;F.ratio[1,1]  Παρατηρούμενο F
qf(1-0.01,1,20)                  Κρίσιμο F για α=0.99
```

$$SST = \underset{\sim}{Y}'\underset{\sim}{Y} - \frac{1}{n}\left(\underset{\sim}{Y}'\underset{\sim}{1}\right)^2$$

$$SSE = \underset{\sim}{Y}'\underset{\sim}{Y} - \hat{\underset{\sim}{\beta}}'X'\underset{\sim}{Y}$$

$$Var(\hat{\underset{\sim}{\beta}}) = \sigma^2 X'X^{-1}$$

$$Var(\hat{y}_0) = \sigma^2 \underset{\sim}{x}_0'(X'X)^{-1}\underset{\sim}{x}_0$$

Επαναλαμβανόμενες μετρήσεις Καθαρά σφάλματα

Καθαρά σφάλματα

α/α	x	y	$\bar{y}$	SSE_i	β.ε
1 2	0.05	0.05 0.10	0.075	0.0012	1
3 4	0.25	0.25 0.35	0.300	0.0050	1
5 6 7	0.50	0.75 0.85 0.95	0.850	0.0200	2
8 9 10 11 12	1.25	1.42 1.75 1.82 1.95 2.45	1.878	0.5619	4
13 14 15 16 17 18	2.10	3.05 3.19 3.25 3.43 3.50 3.93	3.392	0.4805	5
19 20 21 22	2.50	2.75 3.93 3.99 4.07	3.935	0.0555	3
				1.1241	16

$$SSE_i = \sum_{j=1}^{n_j}\left(y_{ij} - \bar{y}_i\right)^2 = = (n_j - 1)\cdot s_i^2$$

$$\beta.\varepsilon. = n_j - 1$$

$$SS_e = \sum_{i=1}^{r} SSE_i = = 1.1241$$

$$n_e = \sum_{i=1}^{r}(n_j - 1) = = n - r = = 22 - 6 = = 16$$

Επαναλαμβανόμενες μετρήσεις
Καθαρά σφάλματα

```
ord <- order(KperC)
x <- KperC[ord]
y <- CO.des[ord]                    Διατάσσονται ως προς αύξοντα x
lev <- levels(factor(x))            Εύρεση διάφορων σταθμών του x
n <- length(lev)                    Πλήθος διάφορων σταθμών του x
lsd <- NULL                         Προετοιμασία διανύσματος
for(i in 1:n) lsd[i] <-
var(y[x==lev[i]])*(length(x[x==lev[i]])-1)
```

$$SSE_i = (n_j - 1)s_i^2$$

Για κάθε διαφορετική τιμή της x το lsd περιέχει τα αθροίσματα τετραγώνων των αποκλίσεων από την αντίστοιχη μέση τιμή

```
ssp <- sum(lsd, na.rm=T)            Αθροίζουμε για να βρούμε τα καθαρά σφάλματα
ssf <- sse-ssp                      Τα σφάλματα προσαρμογής
Msf <- ssf/(dff<-dfe-dfp); fr.ratio<-round(msf/msp,3)
ss<-round(c(ssr,sse,ssf,ssp,sst),5)
ms<-round(c(ssr/1,mse,msf,msp),5)
anov<-c(ss,1,dfe,dff,dfp,dfe+1,ms,"",F.ratio,"",Fr.ratio,"","")
dim(anov)<-c(5,4);
dimnames(anov)<-list(c("Παλινδρόμηση","Υπόλοιπα","Σφ.Προσαρμογής",
   "Σφ. Καθαρά","Σύνολο"),c("Αθροίσματα","β.ε.",
   "Μέσα Τετράγωνα","Λόγοι F"))
```


Ο πίνακας ANOVA για επαναλαμβανόμενες μετρήσεις

```
print.default(anov,quote=F)
```

	Αθροίσματα	β.ε.	Μέσα Τετράγωνα	Λόγοι F
Παλινδρόμηση	42.69209	1	42.69209	697.614
Υπόλοιπα	1.22395	20	0.0612	
Σφ.Προσαρμογής	0.09983	4	0.02496	0.355
Σφ. Καθαρά	1.12411	16	0.07026	
Σύνολο	43.91604	21		

Συγκρίνετε με αυτό που βρήκαμε στο μάθημα

Πηγή	Αθρ. Τετραγώνων	β.ε.	Μέσα τετράγ.	F
Παλινδρόμηση	42.69209	1	42.69209	697.6
Υπόλοιπα	1.22395	20	0.06120	
Σφάλμ. Προσαρμ.	0.099837	4	0.02496	0.355
Καθαρά Σφάλματα	1.124113	16	0.07026	
Σύνολο	43.91604			

Συναρτήσεις που εφαρμόζονται σε ένα lm αντικείμενο

Οι συναρτήσεις που δέχονται ως όρισμα ένα lm αντικείμενο πχ
`dat.lm <- lm(y~x1+x2+…, …)` είναι:

`print(dat.lm):`
τυπώνεται μία απλή εκτύπωση

`summary(dat.lm):`
τυπώνεται εκτενέστερη κατάσταση των παραμέτρων της παλινδρόμησης.

`coef(dat.lm):`
τυπώνονται οι εκτιμώμενοι συντελεστές παλινδρόμησης.

`resid(dat.lm):`
τυπώνονται τα υπόλοιπα.

`fitted(dat.lm):`
υπολογίζονται οι εκτιμήσεις (προβλέψεις) των αρχικών δεδομένων.

`deviance(dat.lm):`
υπολογίζεται το άθροισμα τετραγώνων των σφαλμάτων

`anova(dat.lm):`
δίνεται ο πίνακας ανάλυσης της διασποράς.

`predict(dat.lm, newdata):`
υπολογίζονται προβλέψεις για νέα δεδομένα

`plot(dat.lm):`
τυπώνονται διαγνωστικά γραφήματα

Παράδειγμα με δεδομένα της βιβλιοθήκης MASS Σύγκριση δυο ευθειών παλινδρόμησης

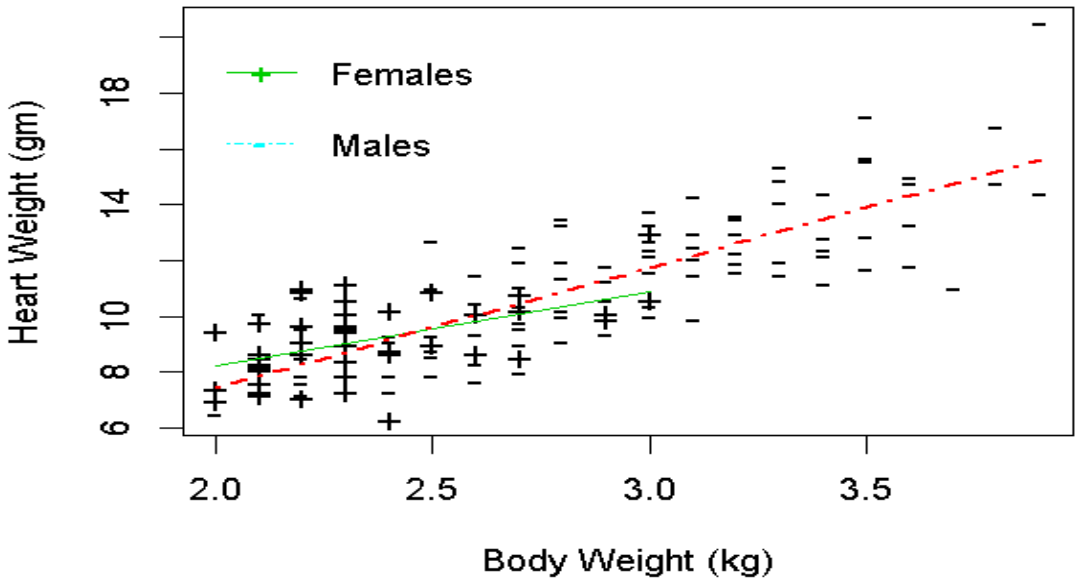
```
library(mass)
attach(cats)
catsF <- lm(Hwt~Bwt,data=cats,
            subset=Sex=="F")
catsM <- update(catsF, subset = Sex == "M")

nF <- catsF$df.resid; nM <- catsM$df.resid

vF <- deviance(catsF)/nF

vM <- deviance(catsM)/nM

c(Male=vM,Female=vF, F=(f<-vM/vF), crit=qf(1-
  0.05,nM,nF),sig= 1-pf(f,nM,nF))
```

```
   Male    Female         F      crit       sig
2.423778 1.350839 1.794276 1.55767 0.01552139
```

Πολλαπλές γραμμές παλινδρόμησης

```
plot(Bwt, Hwt, xlab="Body Weight (kg)",
     ylab="Heart Weight (gm)",type="n")
text(Bwt, Hwt, c("+","-")[Sex])
legend(2.0,20,c("Females"," ","Males"),
      pch="+ -", lty = c(1,-1,4),
      col=c(3,-1,5),bty="n")
lines(Bwt[Sex == "F"], fitted(catsF),
          lty=1,col=3)
lines(Bwt[Sex == "M"], fitted(catsM),
          lty=4,col=5,lwd=2)
```


Συγκριτικό γράφημα των δύο μοντέλων



Μωυσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

Πρόβλεψη με την predict

```
yhat <- fitted(samsad.lm)

eres <- CO.des-yhat ; sum(eres^2)
```

1.223946

```
samsad.newx <- data.frame(KperC=c(1,1.5,2,3))
samsad.newx
predict(samsad.lm, samsad.newx)
predict(samsad.lm, samsad.newx,se.fit=T)
```

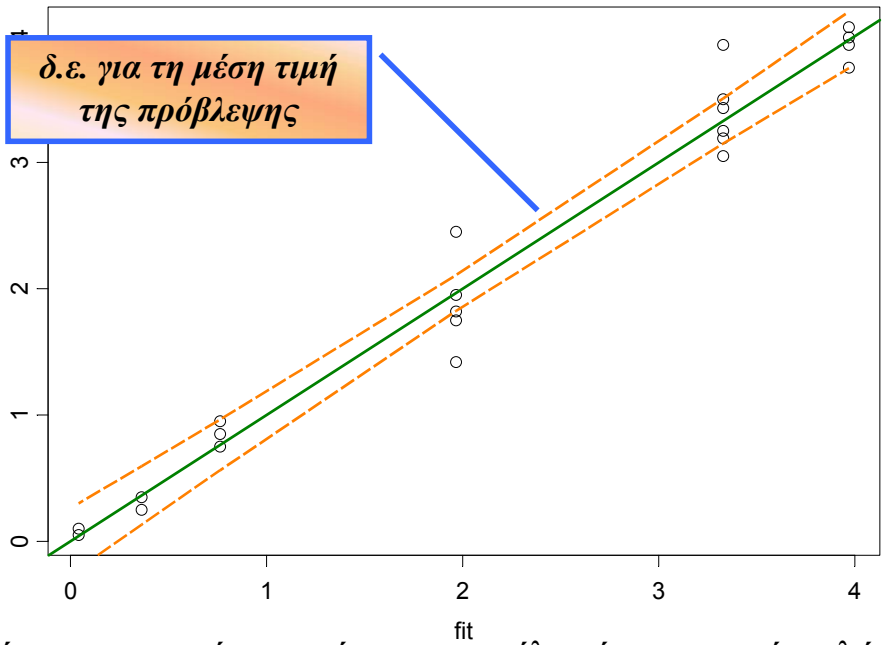
δ.ε. για τη μέση τιμή της πρόβλεψης

```
$fit:
        1        2        3        4
 1.565089 2.366656 3.168222 4.771355
$se.fit:
        1        2        3        4
0.058236 0.053044 0.063860 0.110148
$residual.scale:  [1] 0.2473809
$df:   [1] 20
```

```
$upper:
        1        2        3        4
 1.73079 2.517584 3.349923 5.084764
$fit:
        1        2        3        4
 1.565089 2.366656 3.168222 4.771355
$lower:
        1        2        3        4
 1.399388 2.215727 2.986521 4.457946
```


Η συνάρτηση confint.lm

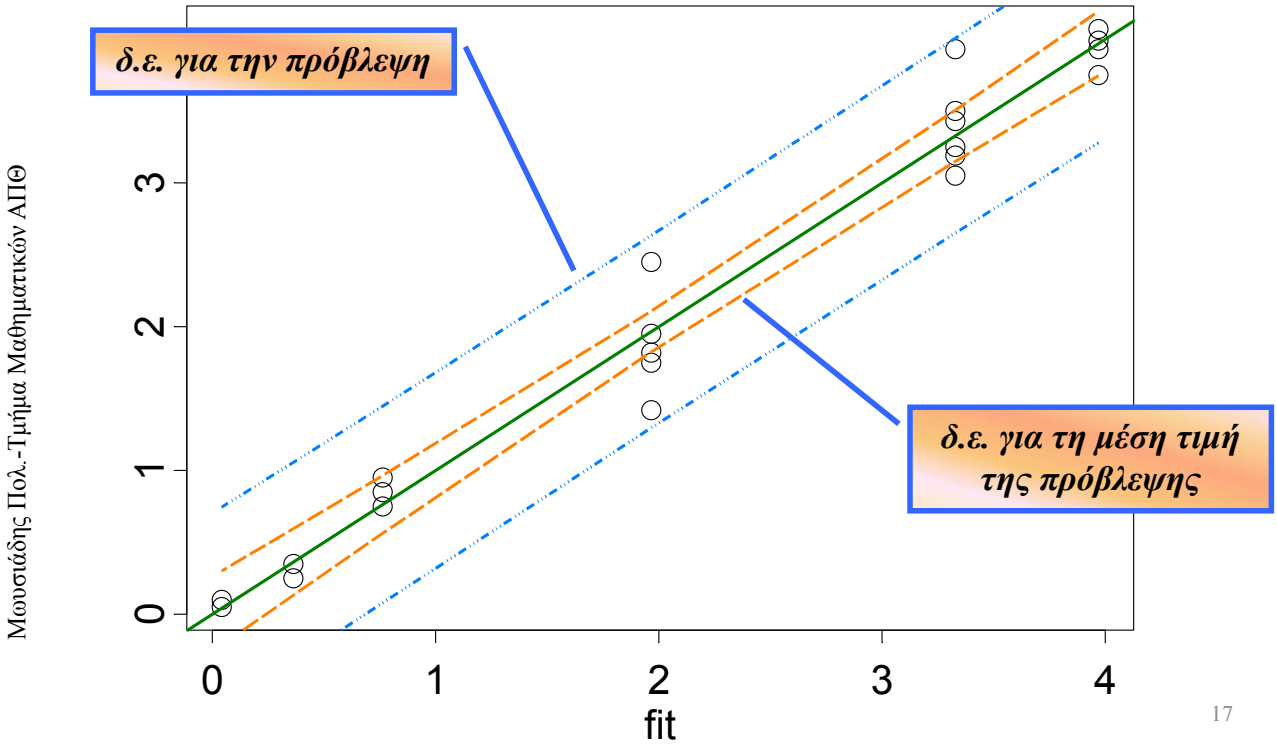
Διάγραμμα διασποράς με ζώνη εμπιστοσύνης



Από το σχήμα συμπεραίνουμε ότι το μοντέλο είναι αρκετά καλό, αφού έξω από τη ζώνη εμπιστοσύνης για τη μέση πρόβλεψη υπάρχουν λίγα σημεία.

Η συνάρτηση confint.adj.lm

Διάγραμμα διασποράς με ζώνες εμπιστοσύνης - πρόβλεψης



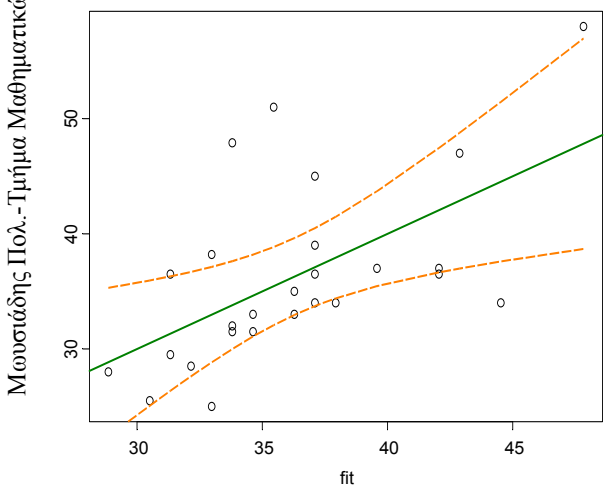
Άλλο παράδειγμα

```
#library(statlab)
#attach(talenta)
nm <- nomos==40
talenta.lm <- lm(
   baros[nm]~ ypsos[nm])
```

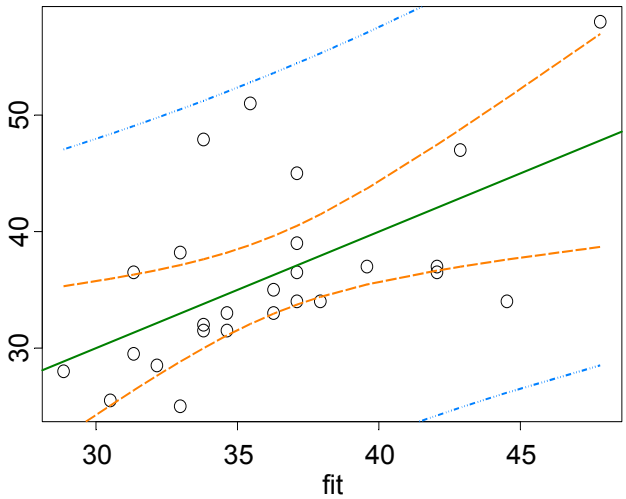
confint.lm(talenta.lm)

confint.adj.lm(talenta.lm)

Διάγραμμα διασποράς με ζώνη εμπιστοσύνης



Διάγραμμα διασποράς με ζώνες εμπιστοσύνης - πρόβλεψης



Πολλαπλή Γραμμική Παλινδρόμηση

Δεδομένα του Hald

```
x1<-c(7,1,11,11,7,11,3,1,2,21,1,11,10)
x2<-c(26,29,56,31,52,55,71,31,54,47,40,66,68)
x3<-c(6,15,8,8,6,9,17,22,18,4,23,9,8)
x4<-c(60,52,20,47,33,22,6,44,22,26,34,12,12)
y<-c(78.5,74.3,104.3,87.6,95.9,109.2,102.7,72.5,
  93.1, 115.9,83.8,113.3,109.4)
hald.y<-y
hald.x<-cbind(x1,x2,x3,x4)
hald.df <- data.frame(hald.y,hald.x)
```

```
hald.lm <- lm(hald.df)   ή
hald.lm <- lm(hald.y~., data=hald.df)  ή
hald.lm <- lm(y ~ x1+x2+x3+x4)
```


Μοντέλα στο S-Plus

`y ~ x1+x2+x3` ή `y ~ 1+x1+x2+x3`

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \varepsilon$$

`y ~ -1+x1+x2+x3` χωρίς σταθερά

`y ~ .` Όλες οι μεταβλητές του data frame (προσοχή στην εξαρτημένη)

`y ~ I(x1+x3)+x2`

$$y = \beta_0 + \beta_1 (x_1 + x_3) + \beta_2 x_2 + \varepsilon$$

`y ~ I(x/z)+x+x*w`

$$y = \beta_0 + \beta_1 (x / z) + \beta_2 x + \beta_3 xw + \varepsilon$$

`y ~ x+z+x*z`

z = 0 ή 1
βωβή μεταβλητή

$$y = \begin{cases} \beta_{10} + \beta_{11} x + \varepsilon \\ \beta_{20} + \beta_{21} x + \varepsilon \end{cases}$$

`y ~ x+x2+(1+x+x2)*z, x2=x^2`

anova και summary

```
hald.lm <- lm(y ~ x1+x2+x3+x4)
anova(hald.lm)
```

Analysis of Variance Table

Response: hald.y Terms added sequentially (first to last)

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
x1	1	1450.076	1450.076	242.3679	0.0000003
x2	1	1207.782	1207.782	201.8705	0.0000006
x3	1	9.794	9.794	1.6370	0.2366003
x4	1	0.247	0.247	0.0413	0.8440715
Residuals	8	47.864	5.983		

```
summary(hald.lm)
```

Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	62.4054	70.0710	0.8906	0.3991
x1	1.5511	0.7448	2.0827	0.0708
x2	0.5102	0.7238	0.7049	0.5009
x3	0.1019	0.7547	0.1350	0.8959
x4	-0.1441	0.7091	-0.2032	0.8441

Residual standard error: 2.446 on 8 degrees of freedom
Multiple R-Squared: 0.9824
F-statistic: 111.5 on 4 and 8 degrees of freedom, the p-value is 4.756e-007
Correlation of Coefficients:

Η συνάρτηση r.squared

Από τον πίνακα ANOVA που δίνει το S-Plus έχουμε το SSR κάθε μεταβλητής αλλά όχι το γνωστό συνολικό SSR.

```
r.squared <- function(x){
      # x is lm type object (x<-lm(y~x1+...))
      y <- anova(x)[, 2]
      z<-anova(x)[,1]
      sse <- y[length(y)]
      sst <- sum(y)
      rsq <- (sst - sse)/sst
      frat<-((sst-sse)/sum(z[-length(y)]))/
                                  (sse/z[length(z)])
      output <- c(sst - sse, sse, sst, rsq, frat)
      names(output) <- c("SSR ", "  SSE ", "  SST ",
                  "Coefficient R^2", "F Ratio")
      output
}
```

```
r.squared(hald.lm)
```

SSR	SSE	SST	Coefficient R^2	F Ratio
2667.899	47.86364	2715.763	0.9823756	111.479

Σύγκριση πλήρους - περιορισμένου μοντέλου

$$F = \frac{(SSR - SSR_{\Pi\varepsilon})/k_1}{SSE/(n-k-1)} = \frac{(SSE_{\Pi\varepsilon} - SSE)/k_1}{SSE/(n-k-1)}$$

$$F \sim F_{k_1, n-k-1}$$

Συγκρίνουμε το πλήρες μοντέλο με το περιορισμένο μοντέλο που περιέχει σαν ανεξάρτητες μεταβλητές μόνο την x3 και την x4.

k_1: πλήθος μεταβλητών που φεύγουν

μικρότερο του 0.05 άρα απορρίπτεται η υπόθεση της ισοδυναμίας των μοντέλων

```
hald.lm<-lm(hald.df)
hald.restricted<-lm(y~x3+x4)
n<-length(y)
k<-4
k1<-2  ##πλήθος μεταβλητών που φεύγουν
sse<-anova(hald.lm)[k+1,2]
sse.restricted<-anova(hald.restricted)[k-k1+1,2]
f.ratio<-(sse.restricted-sse)/k1/(sse/(n-k-1)); f.ratio
p.value<-1-pf(f.ratio,k1,n-k-1);p.value
#P(F>f.ratio)=1-P(F<=f.ratio)
```


Μέθοδοι forward και backward

```
drop1(hald.lm)
```
← Επιλογή μεταβλητής για διαγραφή

```
hald0.lm <- lm(y~1)
```
← Μοντέλο θέσης

```
add1(hald.lm0, ~. +x1+x2+x3+x4)
```
← Επιλογή μεταβλητής για προσθήκη

```
hald.lm2 <- lm(y~x1+x3)
update(hald.lm2,~.+x2)
update(hald.lm2,~.-x3)
```
← προσθήκη ή διαγραφή μεταβλητής

Επιλογή του καλύτερου μοντέλου

```
hald.lm0 <- lm(y~1)
step(hald.lm0, ~x1+x2+x3+x4, trace=T)
```

τυπώνει τα ενδιάμεσα

```
direction="both"
ή "back"  ή "for"
```

η επιλογή στηρίζεται στο κριτήριο AIC (βασίζεται στο Cp), που όσο μικρότερο τόσο καλύτερη προσαρμογή

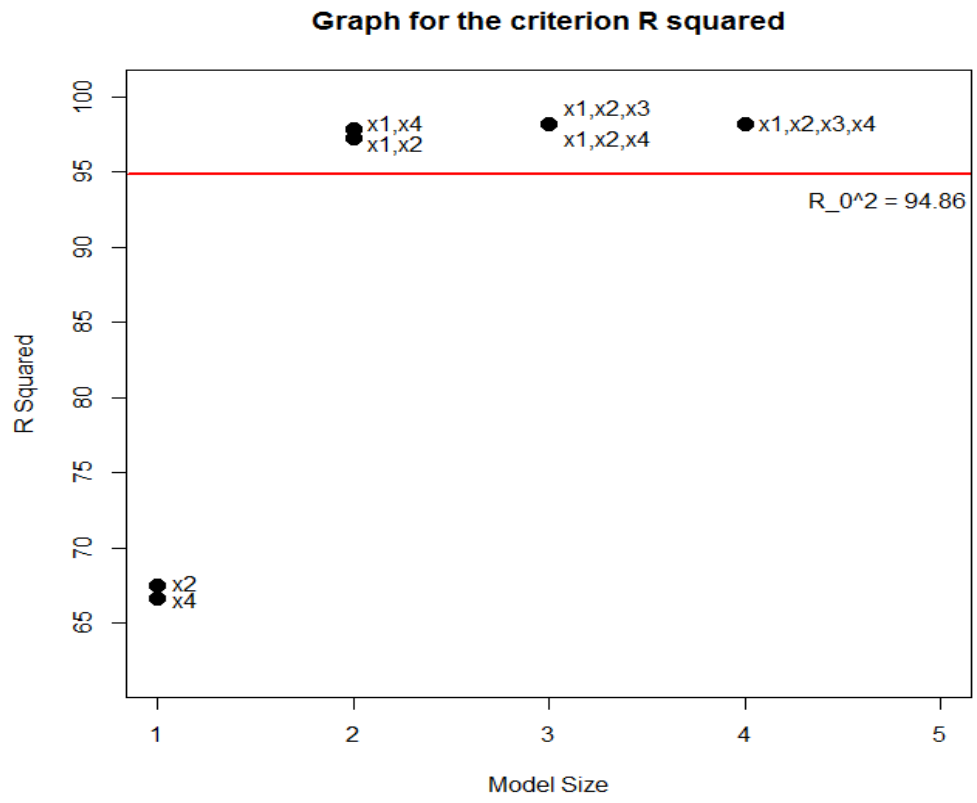
Επιλογή του καλύτερου μοντέλου

```
hald.lm0 <- lm(y~1)
step(hald.lm0, ~x1+x2+x3+x4,trace=T, direction = "forward")

step(hald.lm0, ~x1+x2+x3+x4,trace=T,direction = "both")


hald.lm <-lm(y~x1+x2+x3+x4)
step(hald.lm, ~x1+x2+x3+x4,trace=T,direction = "backward")

step(hald.lm, ~x1+x2+x3+x4,trace=T,direction = "both")
```


Επιλογή με το κριτήριο R^2

```
library(leaps)
lp<-leaps(hald.x,hald.y,nbest=2,method="r2")
        #για κάθε κλάση βρίσκει τα καλύτερα nbest μοντέλα
        # και για κάθε μοντέλο βρίσκει το R^2 (ή το cp, κλπ)
k<-dim(hald.x)[2];a<-0.05
r0<-(1-(1-lp$r2[length(lp$r2)])*(1+k * qf(1-a,k,n-k-1)/(n-k-1)))*100

plot(lp$size-1,lp$r2*100,xlab="Model Size",ylab="R Squared",axes=F,
  xlim=c(1,k+1),ylim=c(min(lp$r2*100,r0)-5,max(lp$r2*100)+2),pch=16,cex=1.5)
m=lp$which
lab=NULL
for( i in 1:nrow(m)){
   mod<-paste("x",colnames(m)[m[i,]],sep="")
   lab[i]<-paste(as.factor(mod),sep=",",collapse =",")
   }
text(lp$size[c(1,3,5)]-1,100*lp$r2[c(1,3,5)]-1,lab[c(1,3,5)],pos=4)
text(lp$size[c(2,4,6)]-1,100*lp$r2[c(2,4,6)]+1,lab[c(2,4,6)],pos=4)
text(lp$size[7]-1,100*lp$r2[7],lab[7],pos=4)
axis(1,at=1:(k+1),labels=0:k+1)
axis(2,at=pretty(c(lp$r2*100,r0)))
title(main="Graph for the criterion R squared")
box()
lines(c(0.85,k+1.15),c(r0,r0),col=2,lwd=2)
legend(k+.1,r0,paste("R_0^2 =",as.character(round(r0,2))),bty="n")
```

Γράφημα με το κριτήριο R^2



Η συνάρτηση leaps και το C_p

```
> leaps(hald.x,hald.y,nbest=3)

$label
[1] "(Intercept)" "1"   "2"   "3"   "4"

$size
 [1] 2 2 2 3 3 3 4 4 4 5

$Cp
 [1] 138.730833 142.486407 202.548769
 [4]   2.678242   5.495851  22.373112
 [7]   3.018233   3.041280   3.496824
[10]   5.000000

$which
      1     2     3     4
1 FALSE FALSE FALSE  TRUE
1 FALSE  TRUE FALSE FALSE
1  TRUE FALSE FALSE FALSE
2  TRUE  TRUE FALSE FALSE
2  TRUE FALSE FALSE  TRUE
2 FALSE FALSE  TRUE  TRUE
3  TRUE  TRUE FALSE  TRUE
3  TRUE  TRUE  TRUE FALSE
3  TRUE FALSE  TRUE  TRUE
4  TRUE  TRUE  TRUE  TRUE
```

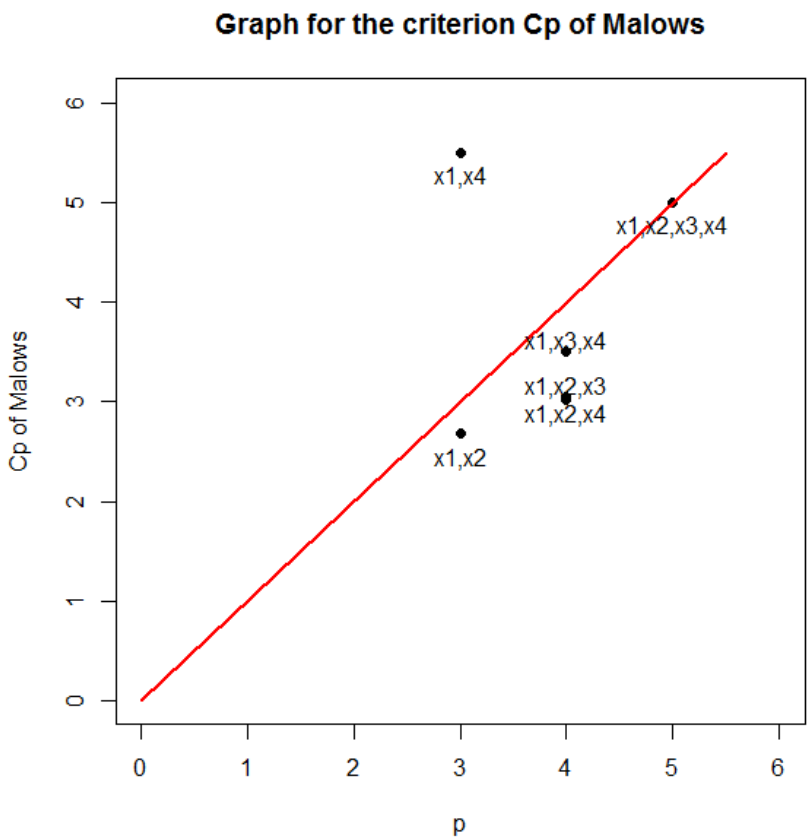

Μπορούμε να εργαστούμε με τη συνάρτηση αυτή για την επιλογή του καλύτερου μοντέλου με το κριτήριο C_p του Mallows

Επιλογή με το κριτήριο C_p του Malows

```
lp3=leaps(hald.x,hald.y,nbest=3)
# με το nbest=3 θα συμπεριλάβει τα 3 καλύτερα μοντέλα από κάθε
# περίπτωση, δηλ 3 καλύτερα με 1 μεταβλητή (με βάση το κριτήριο Cp),
# 3 καλύτερα με 2 μεταβλητές κ.τ.λ.
# στο size μετρά και τη σταθερά κ όχι μόνο το πλήθος των μεταβλητών
size <- lp3$size ; cp <- lp3$Cp
m=lp3$which
lab=NULL
 for( i in 1:nrow(m)){
  mod<-paste("x",colnames(m)[m[i,]],sep="")
  lab[i]<-paste(as.factor(mod),sep=",",collapse =",")
}
lab
# [1] "x4"          "x2"          "x1"          "x1,x2"       "x1,x4"
# [6] "x3,x4"       "x1,x2,x4"    "x1,x2,x3"    "x1,x3,x4"    "x1,x2,x3,x4"

par(mai=c(1.2,1.2,.8,.5))#η mai=c(κάτω,αριστερά,πάνω,δεξιά) δίνει τα
περιθώρια στο γράφημα
plot(size, cp ,xlab="p", ylab="Cp of Malows",
xlim=c(0,6),ylim=c(0,6),pch=16,cex=1)
lines(c(0,5.5),c(0,5.5),lwd=2,type="l",col=2)
text(size[-c(7,8,9)],cp[-c(7,8,9)], lab[-c(7,8,9)],pos=1)
text(size[7],cp[7]+0.1, lab[7],pos=1)
text(size[8],cp[8]-0.1, lab[8],pos=3)
text(size[9],cp[9]-0.1, lab[9],pos=3)
title(main="Graph for the criterion Cp of Malows")
```


Μωυσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

Γράφημα με το κριτήριο C_p του Malows



Μωυσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

Παλινδρόμηση με βωβές μεταβλητές

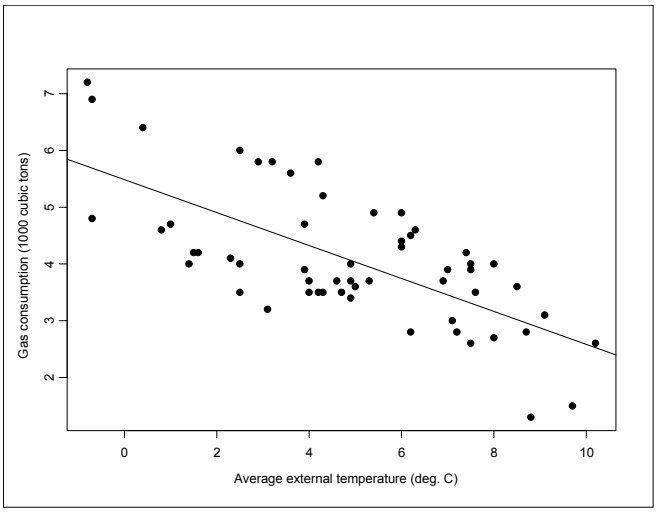
Mr Derek Whiteside of the UK Building Research Station recorded the weekly gas consumption and average external temperature at his own house in south-east England for two heating seasons, one of 26 weeks before, and one of 30 weeks after cavity-wall insulation was installed. The object of the exercise was to assess the effect of the insulation on gas consumption.

Insul
A factor, before or after insulation.

Temp
The weekly average outside temperature in degrees Celsius.

Gas
The weekly gas consumption in 1000s of cubic feet.

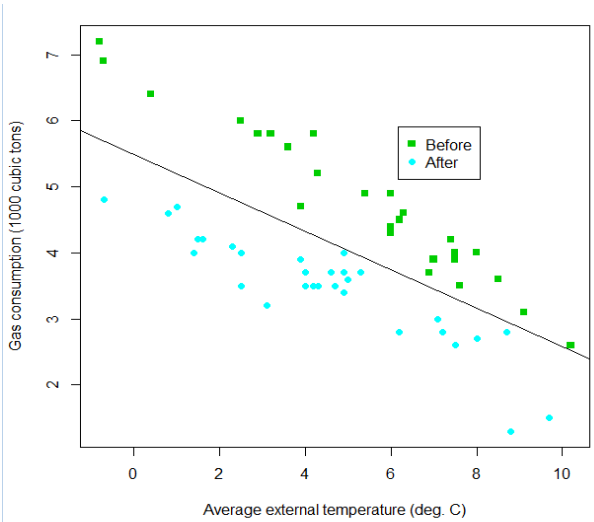
```
plot(whiteside$Temp,whiteside$Gas,xlab="Average
external temperature (deg. C)",    ylab="Gas
consumption (1000 cubic tons)")
abline(lm(Gas ~ Temp, data = whiteside))
```



Το γράφημα, λαμβάνοντας υπόψη τον παράγοντα

```
attach(whiteside)
plot(Temp,Gas,xlab="Average external temperature (deg. C)", ylab="Gas
consumption (1000 cubic tons)",type="n")
points(Temp[Insul=="Before"],Gas[Insul=="Before"],pch=15,col=3)
points(Temp[Insul=="After"],Gas[Insul=="After"],pch=16,col=5)
abline(lm(Gas ~ Temp, data = whiteside))
legend(locator(1),legend=c("Before", "After"),
  col=c(3,5),pch=15:16)
```

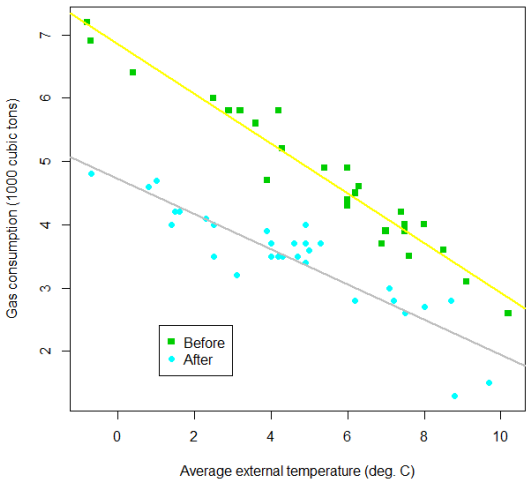
Μωυσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ



Υπάρχει εμφανής διαφορά μεταξύ Before και After.

Το γράφημα, με ξεχωριστές ευθείες παλινδρόμησης

```
plot(Temp,Gas,xlab="Average external temperature (deg. C)",
    ylab="Gas consumption (1000 cubic tons)",type="n")
points(Temp[Insul=="Before"],Gas[Insul=="Before"],pch=15,col=3)
abline(lm(Gas ~ Temp, data = whiteside,
subset=Insul=="Before"),lwd=2,col=7)
points(Temp[Insul=="After"],Gas[Insul=="After"],pch=16,col=5)
abline(lm(Gas ~ Temp, data = whiteside,
subset=Insul=="After"),lwd=2,col=8)
legend(locator(1),legend=c("Before", "After"),
col=c(3,5),pch=15:16)
```



Θα εξετάσουμε αν οι δύο αυτές ευθείες ταυτίζονται, δηλαδή αν ο παράγοντας Insul επηρεάζει την σχέση των Temp και Gas.

Χρήση βωβών μεταβλητών

Μετασχηματίζουμε την Insul σε βωβή μεταβλητή (με τιμές 0 και 1).

```
Insul<-as.numeric(factor(Insul))-1
```

Θεωρούμε το πλήρες μοντέλο

$$\mathbf{Gas = b_0 + b_1 Temp + b_2\ Insul + b_3\ Temp*Insul}$$

Για Insul = 0 (Before) παίρνουμε το μοντέλο

$$\mathbf{Gas = b_0 + b_1 Temp}$$

Για Insul = 1 (After) παίρνουμε το μοντέλο

$$\mathbf{Gas = (b_0+b_2) + (b_1+b_3)Temp}$$

Για να ελέγξουμε αν τα δύο αυτά μοντέλα είναι ισοδύναμα, θα ελέγξουμε τη μηδενική υπόθεση $\mathbf{H_0: b_0=b_0+b_2}$ **και** $\mathbf{b_1=b_1+b_3}$, δηλαδή την υπόθεση $\mathbf{H_0: b_2 = b_3 = 0}$

Θα ελέγξουμε δηλαδή την ισοδυναμία του πλήρους μοντέλου

$$\mathbf{Gas = b_0 + b_1 Temp + b_2\ Insul + b_3\ Temp*Insul}$$

και του περιορισμένου

$$\mathbf{Gas = b_0 + b_1 Temp}$$

Οι εντολές για τη σύγκριση των μοντέλων στην R

```
> lm.ful<-lm(Gas ~ Temp+Insul+InsTemp) # Το πλήρες μοντέλο
> bt <- coef(lm.ful);bt
 (Intercept)        Temp      Insul    InsTemp
    6.853828 -0.3932388 -2.129978 0.1153039
> c(bt[1] + bt[3], bt[2] + bt[4])
 (Intercept)       Temp
     4.72385 -0.277935
> lm.restr<-lm(Gas ~ Temp)     # Το περιορισμένο μοντέλο


 d1_deviance(lm.restr)  # το SSE του περιορισμένου
 # anova(lm.restr)[2,2]
 d2_deviance(lm.ful);d2 # το SSE του πλήρους

 f.ratio_((d1-d2)/2)/(d2/lm.ful$df.resid);f.ratio

 qf(1-0.05,2,52)  # αν f.ratio μεγαλύτερο από αυτήν την τιμή
 απορρίπτουμε
                  # την υπόθεση ισοδυναμίας πλήρους-περιορισμένου

 1-pf(f.ratio,2,52 )# το αντίστοιχο p-value
```

Το p-value είναι πολύ μικρό, επομένως απορρίπτεται η υπόθεση ισοδυναμίας των δύο μοντέλων, δηλαδή η επίδραση του παράγοντα Insul είναι σημαντική.

Έλεγχος παραλληλίας των δύο ευθειών παλινδρόμησης

Θεωρούμε το πλήρες μοντέλο

$$Gas = b_0 + b_1 Temp + b_2\ Insul + b_3\ Temp*Insul$$

Για Insul = 0 $Gas = b_0 + b_1 Temp$

Για Insul = 1 $Gas = (b_0+b_2) + (b_1+b_3) Temp$

Για να ελέγξουμε αν οι δύο ευθείες παλινδρόμησης είναι παράλληλες, ελέγχουμε τη μηδενική υπόθεση $H_0: b_3 = 0$

Αυτός ο έλεγχος μπορεί να γίνει όπως και προηγουμένως, συγκρίνοντας το πλήρες μοντέλο με το περιορισμένο

$$Gas = b_0 + b_1 Temp + b_2\ Insul$$

Εναλλακτικά `summary(lm.ful)`

```
Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)   6.8538   0.1360    50.4091   0.0000
       Temp  -0.3932   0.0225   -17.4874   0.0000
      Insul  -2.1300   0.1801   -11.8272   0.0000
    InsTemp   0.1153   0.0321     3.5907   0.0007
```

Απορρίπτεται η υπόθεση $\beta_3=0$

ANOVA

με την R

Ανάλυση Διασποράς - Εισαγωγή

Η Ανάλυση Διασποράς (Analysis of Variance – ANOVA) εξετάζει αν η μέση τιμή μιας ποσοτικής μεταβλητής είναι ίδια για τις διάφορες ομάδες (περισσότερες από 2) στις οποίες χωρίζεται ο πληθυσμός με βάση μία ή περισσότερες ποιοτικές μεταβλητές (παράγοντες).

Η ANOVA αποτελεί γενίκευση του t-test. Χρησιμοποιείται γιατί η πολλαπλή χρήση του t-test οδηγεί σε αύξηση του σφάλματος τύπου II. Επίσης, τα ίδια αποτελέσματα που παίρνουμε με την ANOVA μπορούμε να τα πάρουμε και εκτελώντας παλινδρόμηση με ποιοτικές (βωβές) μεταβλητές.

Παράδειγμα: *Έχουμε ένα δείγμα 50 ανθρώπων. Τους χωρίζουμε σε τρεις ομάδες (όχι κατ' ανάγκη ισοπληθείς) και σε κάθε ομάδα δίνουμε διαφορετική δίαιτα* ***(ποιοτική μεταβλητή)****. Ενδιαφερόμαστε για το βάρος* ***(ποσοτική μεταβλητή)*** *που χάνουν σε 2 εβδομάδες ακολουθώντας τη δίαιτα που τους δόθηκε.*

Μωυσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

ANOVA στην R

Τρεις ομάδες των 8 ατόμων διδάχτηκαν την ίδια ύλη με 3 διαφορετικές μεθόδους. Στη συνέχεια υποβλήθηκαν σε κοινή εξέταση και μετρήσαμε τις επιδόσεις τους.

method1	method2	method3
3	4	6
5	4	7
2	3	8
4	8	6
8	7	7
4	4	9
3	2	10
9	5	9

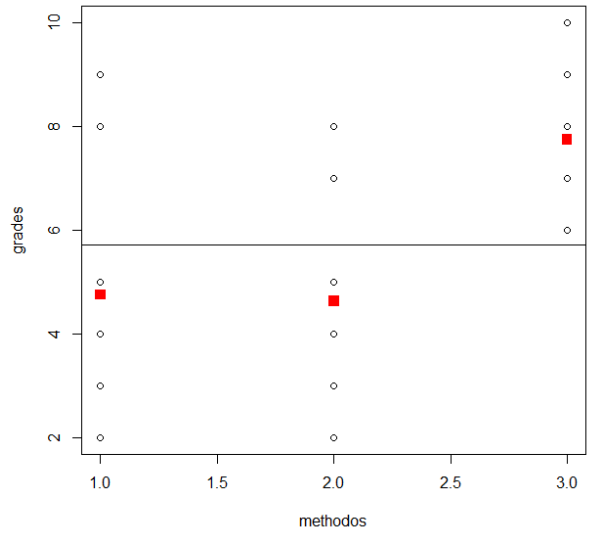
```
m1<-c(3,5,2,4,8,4,3,9)
m2<-c(4,4,3,8,7,4,2,5)
m3<-c(6,7,8,6,7,9,10,9)
grades<-c(m1,m2,m3)
```

Ορίζουμε έναν **παράγοντα** (με στάθμες 1, 2, 3) για την μέθοδο που χρησιμοποιήθηκε

```
methodos<-rep(1:3,each=8)
methodos<-factor(methodos,labels=paste("method",1:3),
                 levels=1:3)
```


Γράφημα

```
plot(methodos,grades,xlim=c(1,3),ylim=range(grades))
abline(mean(grades),0)
points(1:3,
c(mean(m1),mean(m2),mean(m3)),pch=15,col=2,cex=1.5)
```



Μωυσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

Συμβολισμοί

Κάθε παρατήρηση συμβολίζεται με Y_{ij}, όπου ο πρώτος δείκτης αναφέρεται στην ομάδα (στάθμη του παράγοντα) και ο δεύτερος στον αριθμό του ατόμου μέσα στην ομάδα.

πχ η Y_{23} παριστάνει την τιμή της (ποσοτικής) μεταβλητής για το τρίτο άτομο της δεύτερης ομάδας.

Με $\overline{Y}_{i\cdot}$ συμβολίζουμε τη μέση τιμή της μεταβλητής στην i ομάδα, ενώ με $\overline{Y}_{\cdot\cdot}$ τη μέση τιμή για το σύνολο των παρατηρήσεων.

Υπολογίζουμε το γενικό μέσο όρο ($\overline{Y}_{\cdot\cdot}$) της μεταβλητής grades και τους μέσους όρους για κάθε στάθμη του παράγοντα methodos $\left(\overline{Y}_{1\cdot}, \overline{Y}_{2\cdot}, \overline{Y}_{3\cdot}\right)$

```
y..m  <- mean(grades);y..m
means <- tapply(grades,methodos,mean);means
```

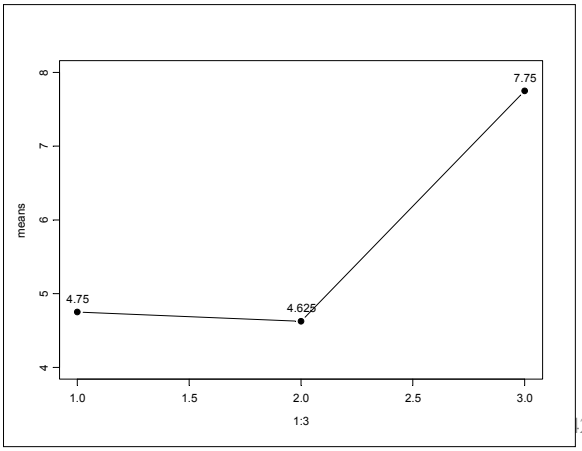

Διαγνωστικά γραφήματα

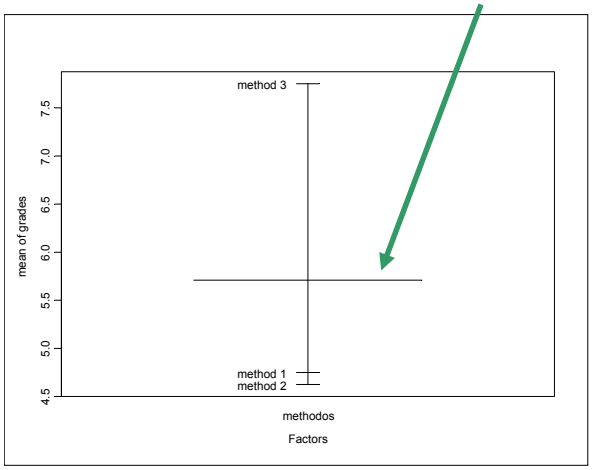
Κατασκευάζουμε ένα απλό γράφημα των επιμέρους μέσων όρων.

```
plot(1:3,means,type='b',
     ylim=c(4,8))
text(1:3,means+.2,means)
```

Εναλλακτικά, μπορούμε να χρησιμοποιήσουμε την εντολή **`plot.design(grad.df)`**.

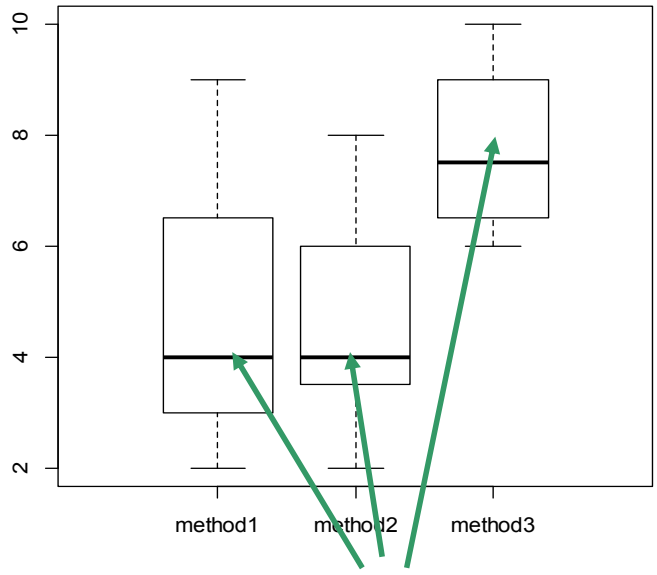
Στο γράφημα τώρα φαίνεται και ο **γενικός μέσος όρος**.





Διαγνωστικά γραφήματα (2)

```
boxplot(split(grades,methodos))    ;h
plot(methodos,grades)
```



διάμεσοι – όχι μέσες τιμές

Ανάλυση διασποράς αναλυτικά

$$SSA = \sum_i r_i \left(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\right)^2$$ ← **μεταξύ των ομάδων**

$$SSE = \sum_i \sum_j \left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2$$ ← **ανάμεσα στις ομάδες**

$$SST = \sum_i \sum_j \left(Y_{ij} - \bar{Y}_{\cdot\cdot}\right)^2$$

«ανάμεσα» = «μέσα»

```
means<-tapply(grades,methodos,mean)
y1.m<-means[1];y1.m
y2.m<-means[2];y2.m
y3.m<-means[3];y3.m
y..m<-mean(grades);y..m

Sizes<-table(methodos);sizes
```

```
sst<-sum((grades-y..m)^2);sst
ssa<-sum(sizes*(means-y..m)^2);ssa
vars<-tapply(grades,methodos,var);

sse<-sum( vars*(sizes-1));sse

# ή απλά
sse<-sst-ssa;sse
```

Κατασκευή του πίνακα ANOVA

```
ss<-round(c(ssa,sse,sst),4)
df<-c(2,length(grades)-3,length(grades)-1)
ms<-round(c(ss[1]/df[1],ss[2]/df[2]),4)
f<-round(ms[1]/ms[2],4)
p.value<-round(1-pf(f,df[1],df[2]),4)
grades.anova<-c(ss,df,ms,"",f,"","",p.value,"","")
dim(grades.anova)<-c(3,5)
dimnames(grades.anova)<-
list(c("Παράγοντας","Σφάλματα","Σύνολο"),c("Άθροισμα
Τετραγώνων","Βαθμοί ελευθερίας","Μέσα τετράγωνα","F","p.value"))
print.default(grades.anova,quote=F)
```

$$Y_{ij} = \mu + a_i + \varepsilon_{ij}$$

	Άθροισμα Τετραγώνων	Βαθμοί ελευθερίας	Μέσα τετράγωνα	F	p.value
Παράγοντας	50.0833	2	25.0416	6.0532	0.0084
Σφάλματα	86.875	21	4.1369		
Σύνολο	136.9583	23			

$H_0: \alpha_1=\alpha_2=\alpha_3$

$H_1: \alpha_1\neq\alpha_2\neq\alpha_3$

Απορρίπτεται η μηδενική υπόθεση ότι η μέση τιμή της grades για τις τρεις ομάδες είναι ίδια (δηλαδή οι κύριες επιδράσεις είναι διαφορετικές.

ANOVA με εντολές της R

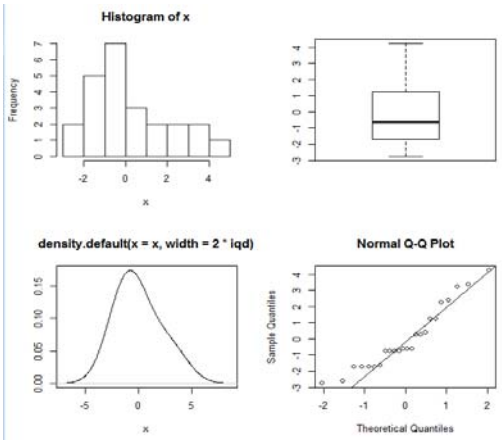
Η εντολή της R για τη διεξαγωγή ανάλυσης διασποράς είναι η aov. Συντάσσεται με τρόπο εντελώς ανάλογο της lm.

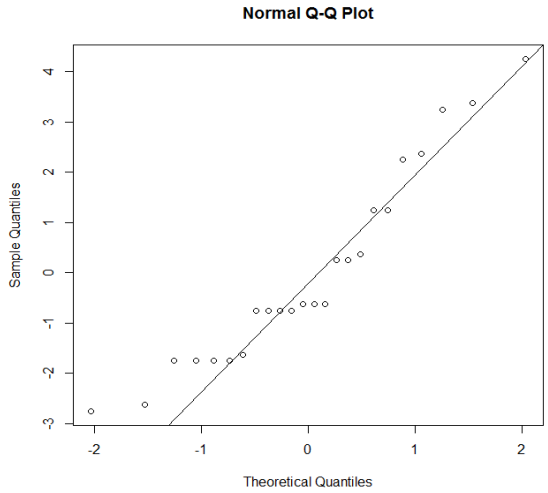
```
grad.aov <- aov(grades ~ methodos); grad.aov
Call:    aov(formula = grades ~ methodos)
Terms:          methodos Residuals
 Sum of Squares 50.08333  86.87500
Deg. of Freedom        2        21
Residual standard error: 2.033938
Estimated effects are balanced

summary(grad.aov)
            Df Sum of Sq  Mean Sq  F Value      Pr(F)
 methodos    2  50.08333 25.04167 6.053237 0.00839879
Residuals 21  86.87500  4.13690
model.tables(grad.aov)      ##κύριες επιδράσεις
    method 1 method 2 method 3
     -0.9583  -1.0833   2.0417
model.tables(grad.aov, type="means") ##μέσοι όροι των ομάδων
    method 1 method 2 method 3
       4.750    4.625    7.750
```

Έλεγχος του μοντέλου - Γραφήματα

```
res<-grad.aov$res
qqnorm(res)
qqline(res)
```

```
> shapiro.test(res)

Shapiro-Wilk  Normality Test

data:  res
W = 0.9221, p-value = 0.0648
```






```
eda.shape(res)
```

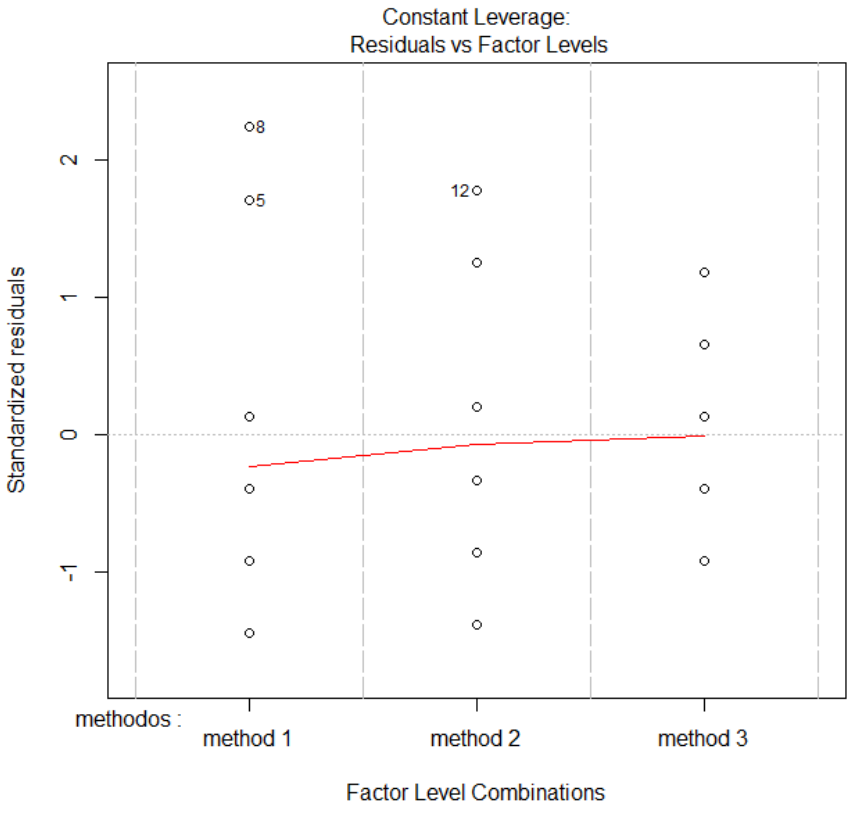
Επίσης, όπως και στην παλινδρόμηση μπορούμε να χρησιμοποιήσουμε την **`plot(grad.aov)`**

Μωυσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

Οπτικός έλεγχος για ετεροσκεδαστικότητα

```
x1<-rep(1:3,each=8)
res1<-res/stdev(res)
plot(x1,res1)
```

Παρατηρώ ότι δεν έχω ετεροσκε-δαστικότητα



Μωυσιάδης Πολ.-Τμήμα Μαθηματικών ΑΠΘ

Σχέση ανάλυσης διασποράς -παλινδρόμησης

Η ανάλυση διασποράς με έναν παράγοντα που έχει k στάθμες ισοδυναμεί με το μοντέλο παλινδρόμησης που έχει σαν προβλέπουσες k βωβές μεταβλητές X_i με $X_i = 1$ μόνο αν η παρατήρηση προέρχεται από την i στάθμη του παράγοντα

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Κατά την εφαρμογή του παραπάνω μοντέλου, ο πίνακας X΄X είναι μη αντιστρέψιμος και χρειάζεται η επιπλέον υπόθεση

$$r_1\beta_1 + r_2\beta_2 + \ldots + r_k\beta_k = 0$$

Επομένως δεν μπορούμε να χρησιμοποιήσουμε στην R τη σύνταξη `lm(Y~X1+X2+…+Xk)`

Χρησιμοποιούμε εναλλακτικά τη σύνταξη `lm(Y~factor)` όπου factor ένας παράγοντας που δηλώνει από ποια στάθμη προέρχεται κάθε παρατήρηση

```
> anova(lm(grades ~ methodos))
Analysis of Variance Table
Response: grades
          Df Sum Sq Mean Sq F value   Pr(>F)
methodos   2 50.083 25.0417  6.0532 0.008399 **
Residuals 21 86.875  4.1369
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```


Παράδειγμα 4.1 – Αντοχή υλικών

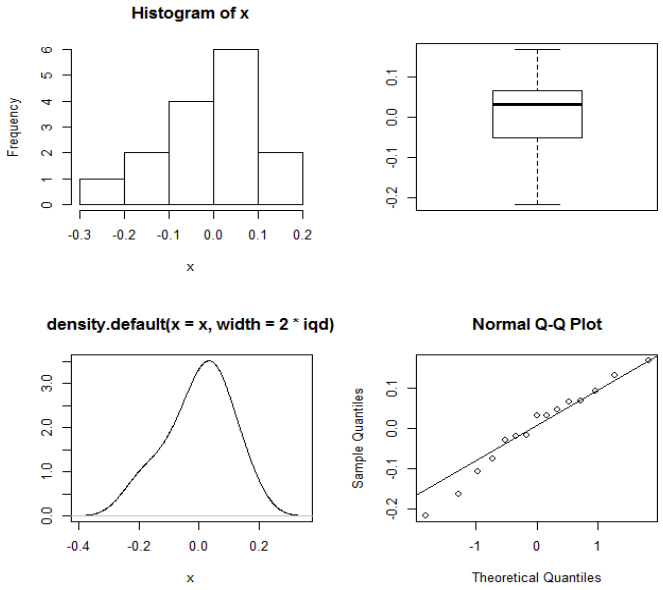
```
antoxi_c(1.700,1.505,1.800,1.725,1.825,1.440,1.640,
   1.735,1.935,1.815,1.890,2.025,2.075,2.110,1.970)
material_rep(paste("mat",LETTERS[1:4],sep=""),c(3,4,4,4))
antoxi.df_data.frame(material,antoxi);antoxi.df
```

```
> antoxi.df
   material antoxi
 1     matA  1.700
 2     matA  1.505
 3     matA  1.800
 4     matB  1.725
 5     matB  1.825
 6     matB  1.440
 7     matB  1.640
 8     matC  1.735
 9     matC  1.935
10     matC  1.815
11     matC  1.890
12     matD  2.025
13     matD  2.075
14     matD  2.110
15     matD  1.970
```

```
> antox.aov <- aov(antoxi ~ material)
> summary(antox.aov)
          Df Sum of Sq   Mean Sq  F Value      Pr(F)
 material  3 0.3785829 0.1261943 8.702487 0.00304325
Residuals 11 0.1595104 0.0145009

> antox.aov$coef
 (Intercept)    material1  material2  material3
    1.803646 -0.005416667 0.06027778 0.08045139

> antox.aov$fitted.values
       1        2        3      4      5      6
1.668333 1.668333 1.668333 1.6575 1.6575 1.6575
       7        8        9     10     11     12
  1.6575  1.84375 1.84375 1.84375 1.84375  2.045
      13       14       15
   2.045    2.045    2.045
```

Παράδειγμα 4.1 – Αντοχή υλικών (2)

```
res<-antox.aov$res
```




```
> shapiro.test(res)

        Shapiro-Wilk Normality Test

data:  res
W = 0.9689, p-value = 0.8415
```


Ισοδυναμία Anova με παλινδρόμηση

4 είδη υλικών => 3 βωβές μεταβλητές z1, z2, z3 όπου:

$$z1[i]=\begin{cases}1 & \alpha\nu \quad \text{είναι το υλικό Α}\\ 0 & \text{αλλού}\end{cases} \qquad z2[i]=\begin{cases}1 & \alpha\nu \quad \text{είναι το υλικό Β}\\ 0 & \text{αλλού}\end{cases}$$

$$z3[i]=\begin{cases}1 & \alpha\nu \quad \text{είναι το υλικό Γ}\\ 0 & \text{αλλού}\end{cases}$$

Το μοντέλο $Y=\beta_0+\beta_1Z_1+\beta_2Z_2+\beta_3Z_3$

Στο υλικό Α: $Y=\beta_0+\beta_1Z_1$ $(Y=\mu+\alpha_1)$

Στο υλικό Β: $Y=\beta_0+\beta_2Z_2$ $(Y=\mu+\alpha_2)$

Στο υλικό Γ: $Y=\beta_0+\beta_3Z_3$ $(Y=\mu+\alpha_3)$

Στο υλικό Δ: $Y=\beta_0$ $(Y=\mu+\alpha_4)$

Το αντίστοιχο μη-παραμετρικό τεστ

Το αντίστοιχο μη-παραμετρικό τεστ της ανάλυσης διασποράς είναι το Kruskal-Wallis test. Το χρησιμοποιούμε όταν τα δεδομένα ανήκουν στην κλίμακα διάταξης ή αν τα σφάλματα δεν ακολουθούν κανονική κατανομή


```
>kruskal.test(grades, methodos)

        Kruskal-Wallis rank sum test

data:  grades and methodos
Kruskal-Wallis chi-square = 8.077, df = 2, p-value = 0.0176
alternative hypothesis: two.sided
```

Και πάλι απορρίπτεται η μηδενική υπόθεση της ισότητας των μέσων τιμών

Ανάλυση Διασποράς με 2 παράγοντες

ΠΑΡΑΔΕΙΓΜΑ

Δύο κατηγορίες ασθενών εξετάσθηκαν ως προς την αντίδρασή τους σε τρία φάρμακα. Για κάθε φάρμακο και κάθε κατηγορία επελέγησαν τυχαία τρεις ασθενείς. Πριν και μετά τη χορήγηση μετρήθηκε με κάποιο κριτήριο η αντίδραση των ασθενών και η διαφορά δίνεται παρακάτω.

Y_{ijk}	Φάρμακο 1 B_1	Φάρμακο 2 B_2	Φάρμακο 3 B_3
Κατηγ. 1 A_1	8, 4, 0	10, 8, 6	8, 6, 4
Κατηγ. 2 A_2	14, 10, 6	4, 2, 0	15, 12, 9

Τα δεδομένα κατά γραμμές

```
apotel<-c(8,4,0,10,8,6,8,6,4,14,
          10,6,4,2,0,15,12,9)

A<-rep(1:2,each=9)
A<-factor(A,labels=c("A1","A2"))
B<-rep(rep(1:3,each=3),2)
B<-factor(B,labels=c("B1","B2","B3"))

apoteld<-data.frame(A,B,apotel)
```

```
> apoteld
    A  B apotel      A  B apotel
1  A1 B1      8  10 A2 B1      14
2  A1 B1      4  11 A2 B1      10
3  A1 B1      0  12 A2 B1       6
4  A1 B2     10  13 A2 B2       4
5  A1 B2      8  14 A2 B2       2
6  A1 B2      6  15 A2 B2       0
7  A1 B3      8  16 A2 B3      15
8  A1 B3      6  17 A2 B3      12
9  A1 B3      4  18 A2 B3       9
```

Ανάλυση Διασποράς με 2 παράγοντες

Ανάλυση διασποράς χωρίς να λάβουμε υπόψη τις αλληλεπιδράσεις

```
> apot1 <- aov(apotel ~ A + B, data = apoteld)

> summary(apot1)
            Df Sum of Sq  Mean Sq F Value     Pr(F)
          A  1        18 18.00000   1.008 0.3324207
          B  2        48 24.00000   1.344 0.2924598
Residuals 14       250 17.85714
```

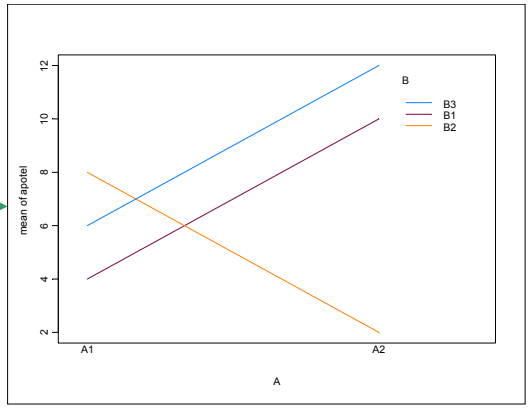
Ανάλυση διασποράς λαμβάνοντας υπόψη τις αλληλεπιδράσεις

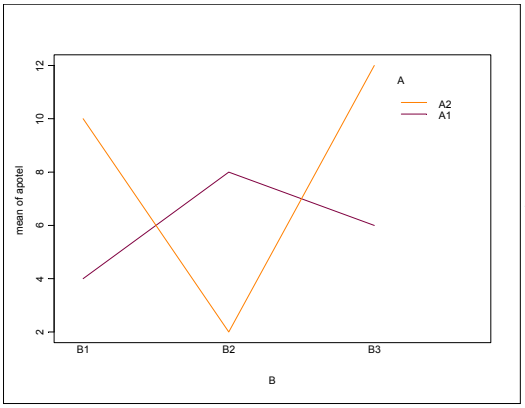
```
> apot2 <- aov(apotel ~ A * B, data = apoteld)
> summary(apot2)
            Df Sum of Sq  Mean Sq  F Value     Pr(F)
          A  1        18 18.00000 2.037736 0.1789399
          B  2        48 24.00000 2.716981 0.1063435
        A:B  2       144 72.00000 8.150943 0.0058103

Residuals 12       106  8.83333
```

Οι κύριες επιδράσεις των δύο παραγόντων δεν είναι σημαντικές, αλλά η αλληλεπίδρασή τους είναι σημαντική

Γραφήματα αλληλεπιδράσεων

```
attach(apoteld)
interaction.plot(A,B,apotel,lwd=2,lty=1,col=c(3,5,6))
interaction.plot(B,A,apotel,lwd=2,lty=1,col=c(3,5))
```





Αν δεν υπάρχουν αλληλεπιδράσεις, περιμένουμε οι γραμμές να είναι περίπου παράλληλες.