Ei-parametriset ja robustit menetelmät

Jyrki Möttönen
Helsingin yliopisto, Sosiaalitieteiden laitos

9. maaliskuuta 2015

Sisältö

Luku 1

Johdanto

- **Otanta- tai koeasetelma**

  Riippumattomat otokset; kaltaistetut parit tai lohkot; satunnaistaminen.

- **Parametriset menetelmät**

  Koetuloksen jakauma oletetaan tunnetuksi lukuunottamatta äärellistä määrää tuntemattomia parametreja (esim. normaalijakaumaoletus tuntemattomalla odotusarvolla μ ja tuntemattomalla varianssilla σ^2). Päättely koskee tuntemattomia parametreja.

- **Ei-parametriset menetelmät**

  Etsitään estimaatteja ja testejä, jotka ovat valideja ja luotettavia mahdollisimman niukoin oletuksin (parametriton malli). Ääritapaus: hyödynnetään vain koejärjestelyyn liittyvää satunnaistamista.

- **Robustit menetelmät**

  Etsitään estimaatteja ja testejä, jotka ovat tehokkaita parametrisen mallin vallitessa. mutta joihin liittyvät tulokset eivät ole herkkiä (muutamille) poikkeaville havainnoille.

- **Tekniikoita**

  Keskiarvo- t-testi- tyyppiset menetelmät (optimaalinen normaalijakaumavasteen tapauksessa); mediaani- merkkitesti-tyyppiset menetelmät; HL-estimaatti-järjestysluku-tyyppiset menetelmät; M-estimaatti-testimenetelmät

- **Ohjelmisto**

  R-ohjelmisto (`http://cran.r-project.org/`)

Luku 2

Kahden käsittelyn vertailu

2.1 Kaltaistetut parit

2.1.1 Koeasetelma

Taulukko 2.1: Havaintoaineisto kaltaistettujen parien tapauksessa.

Pari	Käsittely		Erotus
	A	B	
1	x_1	y_1	$d_1 = y_1 - x_1$
2	x_2	y_2	$d_2 = y_2 - x_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n	$d_n = y_n - x_n$

- Verrattaessa kahta käsittelyä kullekin koeyksikölle etsitään verrokki, **kaltaistettu pari**
- Kunkin parin kohdalla arvotaan, kumpi saa käsittelyä A ja kumpi käsittelyä B

Kuvatun koeasetelman tapauksessa on usein realistista , että

$$d_1, \ldots, d_n$$

on satunnaisotos jatkuvasta symmetrisestä jakaumasta, jonka symmetriapiste Δ (tuntematon) antaa käsittelyjen vaikutusten keskimääräisen eron.

Esimerkki 2.1.1
Kliinisessä hoitokokeessa tutkitaan, kuinka klofibraattilääkitys [1] alentaa plasman fibrinogeenipitoisuutta [2] keski-ikäisillä miehillä, joiden seerumin kolesterolipitoisuus on suuri. Näin määritellystä perusjoukosta poimittiin 10 hengen otos. Kokeen alussa mitattiin fibrinogeenipitoisuus ensimmäisen kerran ja 8 viikon lääkehoidon jälkeen toisen kerran. Mittauksien tulokset on koottu seuraavaan taulukkoon.

Taulukko 2.2: Fibrinogeenipitoisuudet (mg/100 ml) 10 potilaalla ennen hoitoa ja hoidon jälkeen.

Potilas	Pitoisuus		Erotus
	Ennen	Jälkeen	
1	379	325	-54
2	351	333	-18
3	420	391	-29
4	303	275	-28
5	346	311	-35
6	370	323	-47
7	381	370	-11
8	349	354	5
9	284	249	-35
10	380	315	-65

2.1.2 Normaalijakaumaoletus

$d_1, \ldots, d_n$ on satunnaisotos normaalijakaumasta $N(\Delta, \sigma^2)$ tuntemattomalla odotusarvolla Δ ja tuntemattomalla varianssilla $\sigma^2 > 0$. Käsittelyjen eroa kuvaava parametri on

$$\Delta = \text{'käsittelyjen vaikutusten keskimääräinen ero'}.$$

Silloin

- AINEISTO: $d_1, \ldots, d_n \quad \rightarrow \quad \bar{d}, s_d$
- TESTAUS (t-testi):

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \ \sim \ t(n-1), \text{ kun } H_0 : \Delta = 0 \text{ tosi}$$

[1] Klofibraatti: eräs veren kolesteroli- ja triglyseridipitoisuutta vähentävä ateroskleroosilääke

[2] Fibrinogeeni: veriplasman valkuaisaine, joka aiheuttaa veren hyytymisen pilkkoutumalla fibriiniksi

- PISTE-ESTIMOINTI (suluissa keskivirhe, estimaatin keskihajonta):

$$\Delta : \quad \bar{d} \quad \left(\pm \frac{s_d}{\sqrt{n}} \right)$$

- LUOTTAMUSVÄLI:

$$\left(\bar{d} - t_{\alpha/2} \cdot \frac{s_d}{\sqrt{n}} \ , \ \bar{d} + t_{\alpha/2} \cdot \frac{s_d}{\sqrt{n}} \right)$$

Esimerkki 2.1.2 Normaalijakaumaoletus

```
> esim.dat<-read.table("fibrinogeeni.dat",header=TRUE)
> esim.dat
     x   y
1  379 325
2  351 333
3  420 391
4  303 275
5  346 311
6  370 323
7  381 370
8  349 354
9  284 249
10 380 315
> d<-esim.dat$y-esim.dat$x
> d
 [1] -54 -18 -29 -28 -35 -47 -11   5 -35 -65
> n<-length(d)
> md<-mean(d)
> md
[1] -31.7
> sd<-sqrt(var(d))
> sd
[1] 20.67231
> t<-sqrt(n)*md/sd
> t
[1] -4.849202
> pt(t,n-1)
[1] 0.0004546704
> 2*pt(t,n-1)
[1] 0.0009093409
> dala<-md-qt(0.975,n-1)*sd/sqrt(n)
> dala
```

```
[1] -46.48808
> dyla<-md+qt(0.975,n-1)*sd/sqrt(n)
> dyla
[1] -16.91192
> t.test(d)

        One Sample t-test

data:  d
t = -4.8492, df = 9, p-value = 0.0009093
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -46.48808 -16.91192
sample estimates:
mean of x
    -31.7

>
```

2.1.3 Parametriton malli A

$d_1, \ldots, d_n$ on satunnaisotos (tuntemattomasta) jatkuvasta jakaumasta, jonka tuntematon mediaani on Δ. Käsittelyjen eroa kuvaava parametri on jälleen

$$\Delta = \text{'käsittelyjen vaikutusten keskimääräinen ero'}.$$

Silloin

- AINEISTO: $d_1, \ldots, d_n$
- TESTAUS (merkkitesti):

  $$S = \#\{d_i > 0\} \ \sim \ Bin(n, 1/2), \quad \text{kun } H_0 : \Delta = 0 \text{ tosi.}$$

  S on siis niiden tapausten lukumäärä, joilla $d_i > 0$. Nollahypoteesin vallitessa pätee myös keskeisen raja-arvolauseen nojalla likimain

  $$z = \frac{S - n/2}{\sqrt{n/4}} \ \sim \ N(0,1)$$

- PISTE-ESTIMOINTI :

  $$\Delta : \quad m_d \ = \ Med\{d_1, \ldots, d_n\}$$

- LUOTTAMUSVÄLI: Olkoot $d_{(1)} \leq d_{(2)} \leq \ldots \leq d_{(n)}$ havainnot suuruusjärjestyksessä. Tällöin luottamusvälit ovat muotoa

  $$(d_{(i)}, d_{(n+1-i)})$$

peitetodennäköisyytenä

$$P_i = 1 - 2P_0(S \leq i-1)$$

(todenäköisyydet jakaumasta $Bin(n, 1/2)$).

Esimerkki 2.1.3 Parametriton malli A

```
> esim.dat<-read.table("fibrinogeeni.dat",header=TRUE)
> d<-esim.dat$y-esim.dat$x
> d
 [1] -54 -18 -29 -28 -35 -47 -11   5 -35 -65
> n<-length(d)
> median(d)
[1] -32
> S<-sum(d>0)
> S
[1] 1
> pbinom(1,n,0.5)
[1] 0.01074219
> 2*pbinom(1,n,0.5)
[1] 0.02148438
> do<-sort(d)
> do
 [1] -65 -54 -47 -35 -35 -29 -28 -18 -11   5
> i<-1
> c(do[i],do[n+1-i],1-2*pbinom(i-1,n,0.5))
[1] -65.0000000   5.0000000   0.9980469
> i<-2
> c(do[i],do[n+1-i],1-2*pbinom(i-1,n,0.5))
[1] -54.0000000 -11.0000000   0.9785156
> i<-3
> c(do[i],do[n+1-i],1-2*pbinom(i-1,n,0.5))
[1] -47.000000 -18.000000   0.890625
> i<-4
> c(do[i],do[n+1-i],1-2*pbinom(i-1,n,0.5))
[1] -35.00000 -28.00000   0.65625
```

2.1.4 Parametriton malli B

$d_1, \ldots, d_n$ on satunnaisotos (tuntemattomasta) jatkuvasta **symmetrisestä** jakaumasta, jonka tuntematon mediaani on Δ. Käsittelyjen eroa kuvaava parametri on jälleen

$$\Delta = \text{'käsittelyjen vaikutusten keskimääräinen ero'}$$

Nyt

- AINEISTO: $d_1, \ldots, d_n \ \rightarrow \ w_1, \ldots, w_N$, missä $w_1, \ldots, w_N$, $N = n(n+1)/2$ ovat kaikki mahdolliset pareittaiset keskiarvot (Walshin keskiarvot)
$$\frac{d_i + d_j}{2}, \quad 1 \leq i \leq j \leq n.$$

- TESTAUS (Wilcoxonin pareittaisten otosten testi):
$$W = \#\{w_i > 0\}.$$
Nollahypoteesin $H_0 : \Delta = 0$ vallitessa W:n jakauma ei riipu lainkaan tuntemattoman symmetrisen jakauman muodosta. Nollahypoteesin vallitessa pätee myös likimain
$$z = \frac{W - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \ \sim \ N(0,1).$$

- PISTE-ESTIMOINTI : Niin sanottu Hodgesin-Lehmannin estimaatti (HL-estimaatti) on
$$\Delta : \quad m_w \ = \ Med\{w_1, \ldots, w_N\}$$

- LUOTTAMUSVäLI: Olkoot $w_{(1)} \leq w_{(2)} \leq \ldots \leq w_{(N)}$ Walshin keskiarvot suuruusjärjestyksessä. Tällöin luottamusvälit ovat muotoa
$$(w_{(i)}, w_{(N+1-i)})$$
peitetodennäköisyytenä
$$P_i = 1 - 2P_0(W \leq i-1) \approx 1 - 2\Phi\left(\frac{i - \frac{n(n+1)}{4} - \frac{1}{2}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}\right).$$

Esimerkki 2.1.4 Parametriton malli B

```
> esim.dat<-read.table("fibrinogeeni.dat",header=TRUE)
> d<-esim.dat$y-esim.dat$x
> d
 [1] -54 -18 -29 -28 -35 -47 -11   5 -35 -65
> n<-length(d)
> do<-sort(d)
> apu<-matrix(rep(do,n),n)
> W<-(apu+t(apu))/2
```

```
> W
       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10]
 [1,] -65.0 -59.5 -56.0 -50.0 -50.0 -47.0 -46.5 -41.5 -38.0 -30.0
 [2,] -59.5 -54.0 -50.5 -44.5 -44.5 -41.5 -41.0 -36.0 -32.5 -24.5
 [3,] -56.0 -50.5 -47.0 -41.0 -41.0 -38.0 -37.5 -32.5 -29.0 -21.0
 [4,] -50.0 -44.5 -41.0 -35.0 -35.0 -32.0 -31.5 -26.5 -23.0 -15.0
 [5,] -50.0 -44.5 -41.0 -35.0 -35.0 -32.0 -31.5 -26.5 -23.0 -15.0
 [6,] -47.0 -41.5 -38.0 -32.0 -32.0 -29.0 -28.5 -23.5 -20.0 -12.0
 [7,] -46.5 -41.0 -37.5 -31.5 -31.5 -28.5 -28.0 -23.0 -19.5 -11.5
 [8,] -41.5 -36.0 -32.5 -26.5 -26.5 -23.5 -23.0 -18.0 -14.5  -6.5
 [9,] -38.0 -32.5 -29.0 -23.0 -23.0 -20.0 -19.5 -14.5 -11.0  -3.0
[10,] -30.0 -24.5 -21.0 -15.0 -15.0 -12.0 -11.5  -6.5  -3.0   5.0
> w<-c(W[row(W)<=col(W)])
> w
 [1] -65.0 -59.5 -54.0 -56.0 -50.5 -47.0 -50.0 -44.5 -41.0 -35.0
-50.0 -44.5 -41.0 -35.0
[15] -35.0 -47.0 -41.5 -38.0 -32.0 -32.0 -29.0 -46.5 -41.0 -37.5
-31.5 -31.5 -28.5 -28.0
[29] -41.5 -36.0 -32.5 -26.5 -26.5 -23.5 -23.0 -18.0 -38.0 -32.5
-29.0 -23.0 -23.0 -20.0
[43] -19.5 -14.5 -11.0 -30.0 -24.5 -21.0 -15.0 -15.0 -12.0 -11.5
-6.5  -3.0   5.0
> median(w)
[1] -32
> length(w)
[1] 55
> wtest<-sum(w>0)
> wtest
[1] 1
> z<-(wtest-n*(n+1)/4+1/2)/(sqrt(n*(n+1)*(2*n+1)/24))
> z
[1] -2.752095
> pnorm(z)
[1] 0.002960769
> psignrank(wtest,n)
[1] 0.001953125
> 2*psignrank(wtest,n)
[1] 0.00390625
> i<-3
> 1-2*pnorm((i-n*(n+1)/4-1/2)/(sqrt(n*(n+1)*(2*n+1)/24)))
[1] 0.989173
> 1-2*psignrank(i-1,n)
[1] 0.9941406
> i<-5
```

```
> 1-2*pnorm((i-n*(n+1)/4-1/2)/(sqrt(n*(n+1)*(2*n+1)/24)))
[1] 0.9809411
> 1-2*psignrank(i-1,n)
[1] 0.9863281
> i<-9
> 1-2*pnorm((i-n*(n+1)/4-1/2)/(sqrt(n*(n+1)*(2*n+1)/24)))
[1] 0.947213
> 1-2*psignrank(i-1,n)
[1] 0.9511719
> wo<-sort(w)
> wo
 [1] -65.0 -59.5 -56.0 -54.0 -50.5 -50.0 -50.0 -47.0 -47.0 -46.5
-44.5 -44.5 -41.5 -41.5
[15] -41.0 -41.0 -41.0 -38.0 -38.0 -37.5 -36.0 -35.0 -35.0 -35.0
-32.5 -32.5 -32.0 -32.0
[29] -31.5 -31.5 -30.0 -29.0 -29.0 -28.5 -28.0 -26.5 -26.5 -24.5
-23.5 -23.0 -23.0 -23.0
[43] -21.0 -20.0 -19.5 -18.0 -15.0 -15.0 -14.5 -12.0 -11.5 -11.0
-6.5  -3.0   5.0
> c(wo[i],wo[length(wo)+1-i])
[1] -47 -15
>
```

2.1.5 Estimaattien vertailua

Keskiarvo, mediaani ja Hodgesin-Lehmannin estimaatti minimoivat tappiofunktiot:

$$D_I(\Delta) = \sum_{i=1}^{n} (d_i - \Delta)^2,$$

$$D_{II}(\Delta) = \sum_{i=1}^{n} |d_i - \Delta|$$

ja

$$D_{III}(\Delta) = \sum_{i \le j} |d_i + d_j - 2\Delta|.$$

Todistus. Tappiofunktio D_I: $\frac{d}{d\Delta} D_I(\Delta) = -2 \sum_{i=1}^{n} (d_i - \Delta) \doteq 0 \Rightarrow \hat{\Delta} = \bar{d}$

Tappiofunktio D_{II}: Kun $\Delta \notin \{d_1, \ldots, d_n\}$, niin

$$\begin{aligned}\frac{d}{d\Delta} D_{II}(\Delta) &= -\sum_{i=1}^{n} \text{sign}(d_i - \Delta) \\ &= \#\{(d_i - \Delta) < 0\} - \#\{(d_i - \Delta) > 0\} \\ &= -2\left(\#\{d_i > \Delta\} - \frac{n}{2}\right).\end{aligned}$$

$n = 2k$: $\frac{d}{d\Delta} D_{II}(\Delta) = 0$, kun $d_{(k)} < \Delta < d_{(k+1)}$
$\Rightarrow$ Voidaan valita $\hat{\Delta} = (d_{(k)} + d_{(k+1)})/2$.

$n = 2k - 1$: $\frac{d}{d\Delta} D_{II}(\Delta)$ vaihtaa merkkiä kohdassa $\Delta = d_{(k)}$, joten $\hat{\Delta} = d_{(k)}$.

Tappiofunktion D_{II} minimoi siis otosmediaani $\hat{\Delta} = Med\{d_i \ : \ i = 1, \ldots, n\}$.

Tappiofunktio D_{III}: Kun $\Delta \notin \{(d_i + d_j)/2 \ : \ i \leq j\}$, niin

$$\frac{d}{d\Delta} D_{III}(\Delta) = -2 \sum_{i \leq j} \text{sign}\left(\frac{d_i + d_j}{2} - \Delta\right)$$

Kun verrataan derivaattoja D_{II} ja D_{III}, niin huomataan välittömästi, että $\hat{\Delta} = Med\{(d_i + d_j)/2 \ : \ i \leq j\}$. □

Niinkutsutut **M-estimaatit** (katso kappale 2.1.13) minimoivat funktion

$$D(\Delta) = \sum_{i=1}^{n} \rho(d_i - \Delta),$$

missä $\rho(\cdot)$ on origon suhteen symmetrinen, konveksi funktio (minimi origossa).

TULOS: Oletetaan, että $d_1, \ldots, d_n$ on satunnaisotos Δ:n suhteen symmetrisestä jakaumasta, jonka tiheysfunktio on $f(x)$. Olkoot

$$\hat{\Delta}_I, \quad \hat{\Delta}_{II} \quad \text{ja} \quad \hat{\Delta}_{III}$$

otoksesta lasketut keskiarvo, mediaani ja HL-estimaatti. Silloin

$$\hat{\Delta}_I \ \sim \ AN\left(\Delta, \frac{1}{n}\sigma^2\right),$$

$$\hat{\Delta}_{II} \ \sim \ AN\left(\Delta, \frac{1}{n}\frac{1}{4f^2(\Delta)}\right)$$

ja

$$\hat{\Delta}_{III} \ \sim \ AN\left(\Delta, \frac{1}{n}\frac{1}{12[\int f^2(x)dx]^2}\right),$$

missä σ^2 on jakauman f varianssi.

HUOM. Merkintä

$$\hat{\Delta}_I \ \sim \ AN\left(\Delta, \frac{1}{n}\sigma^2\right)$$

luetaan '$\hat{\Delta}_I$:n jakauma on asymptoottisesti normaalinen (tai likimain normaalinen) odotusarvolla ...' ja täsmällisesti ilmaisten tarkoittaa, että muuttujan

$$\sqrt{n}(\hat{\Delta}_I - \Delta)$$

rajajakauma n:n kasvaessa on $N(0, \sigma^2)$.

MÄÄRITELMÄ. Oletetaan, että $\hat{\Delta}_1$ ja $\hat{\Delta}_2$ ovat kilpailevia $\Delta = \Delta(F)$:n estimaatteja ja että

$$\hat{\Delta}_1 \ \sim \ AN\left(\Delta, \frac{1}{n}\sigma_1^2(F)\right)$$

ja

$$\hat{\Delta}_2 \ \sim \ AN\left(\Delta, \frac{1}{n}\sigma_2^2(F)\right).$$

Silloin rajajakaumien varianssien suhdetta

$$\text{Eff}(\hat{\Delta}_1, \hat{\Delta}_2) = \frac{\sigma_2^2(F)}{\sigma_1^2(F)}$$

kutsutaan $\hat{\Delta}_1$:n asymptoottiseksi suhteelliseksi tehokkuudeksi $\hat{\Delta}_2$:n suhteen (jakauman F tapauksessa).

Tulkinta: Olkoon n "suuri". Jos $\hat{\Delta}_1$:n laskemiseen käytetään n havaintoa, niin $\hat{\Delta}_2$:n laskemiseen tarvitaan $n \cdot \text{Eff}(\hat{\Delta}_1, \hat{\Delta}_2)$ havaintoa, jotta estimaattorit olisivat yhtä tarkkoja.

Esimerkki 2.1.5
$d_1, \ldots, d_n$ satunnaisotos $N(0,1)$-jakaumasta:

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}, \quad -\infty < x < \infty$$

Silloin $\sigma^2 = 1$,

$$f(0) = \frac{1}{\sqrt{2\pi}}$$

ja

$$\int_{-\infty}^{\infty} f^2(x)dx = \frac{1}{\sqrt{4\pi}} \quad \text{(Osoita!)}$$

ja lopulta

$$\text{Eff}(\hat{\Delta}_{II}, \hat{\Delta}_I) = \frac{2}{\pi} \approx 0.637$$

ja

$$\text{Eff}(\hat{\Delta}_{III}, \hat{\Delta}_I) = \frac{3}{\pi} \approx 0.955.$$

2.1.6 Estimaatin tehokkuus pienillä otoskoilla

Oletetaan, että $X = \{x_1, \ldots, x_n\}$ on satunnaisotos jakaumasta F ja että $\hat{\Delta} = \hat{\Delta}(X)$ on parametrin $\Delta = \Delta(F)$ estimaatti. Estimaatin $\hat{\Delta}$ ominaisuuksia otoskoolla n jakauman F tapauksessa voidaan tutkia **simuloimalla**:

- Generoi N otosta otoskoolla n jakaumasta F,

  $$X_1, \ldots, X_N,$$

  ja laske otoksiin liittyvät estimaattien arvot

  $$\hat{\Delta}(X_1), \ldots, \hat{\Delta}(X_N).$$

- Estimaatin odotusarvon estimaatti:

  $$\widehat{E_F(\hat{\Delta})} = \frac{1}{N} \sum_{i=1}^{N} \hat{\Delta}(X_i)$$

- Estimaatin varianssin estimaatti:

  $$\widehat{\text{Var}_F(\hat{\Delta})} = \frac{1}{N} \sum_{i=1}^{N} [\hat{\Delta}(X_i)]^2 - \left[\frac{1}{N} \sum_{i=1}^{N} \hat{\Delta}(X_i)\right]^2 = \frac{1}{N} \sum_{i=1}^{N} (\hat{\Delta}(X_i) - \widehat{E_F(\hat{\Delta})})^2$$

- Harhan $E_F(\hat{\Delta}) - \Delta$ estimaatti

  $$\widehat{B_F(\hat{\Delta})} = \widehat{E_F(\hat{\Delta})} - \Delta(F)$$

- Keskivirheen (Mean Square Error, MSE) $E_F[(\hat{\Delta} - \Delta)^2]$ estimaatti

  $$\widehat{\text{Var}_F(\hat{\Delta})} + [\widehat{B_F(\hat{\Delta})}]^2$$

```
Simulointikokeita keskiarvon, mediaanin ja midrange-estimaattien
vertaamiseksi:
--------------
> luvut<-matrix(c(1,2,3,4,5,6), nrow=2)
> luvut
     [,1] [,2] [,3]
```

```
[1,]    1    3    5
[2,]    2    4    6

> apply(luvut,1,sum)
[1]  9 12
> apply(luvut,2,sum)
[1]  3  7 11
---------------
```

Normaalijakauma:

```
> data<-matrix(rnorm(500*20), nrow=500)
> dim(data)
[1] 500  20
> t1<-apply(data,1,mean)
> mean(t1)
[1] 0.02095357
> var(t1)
[1] 0.05520245
> t2<-apply(data,1,median)
> mean(t2)
[1] 0.02331353
> var(t2)
[1] 0.08238773
> var(t1)/var(t2)
[1] 0.6700324
> midrange<-function(x) {(min(x)+max(x))/2}
> t3<-apply(data,1,midrange)
> mean(t3)
[1] 0.01723710
> var(t3)
[1] 0.1430654
> var(t1)/var(t3)
[1] 0.3858547
-----------------
```

Tasainen jakauma:

```
> data<-matrix(runif(500*20), nrow=500)
> t1<-apply(data,1,mean)
> t2<-apply(data,1,median)
> t3<-apply(data,1,midrange)
> var(t1)/var(t2)
[1] 0.3895521
```

```
> var(t1)/var(t3)
[1] 3.643573
-----------------

Laplace jakauma

rlaplace<-function(n)
{
#
# Generoidaan Laplace-jakaumasta n havaintoa
#
  rexp(n)*(2*rbinom(n,1,0.5)-1)
}

> data<-matrix(rlaplace(500*20), nrow=500)
> t1<-apply(data,1,mean)
> t2<-apply(data,1,median)
> t3<-apply(data,1,midrange)
> var(t1)/var(t2)
[1] 1.478895
> var(t1)/var(t3)
[1] 0.1264577
> data<-matrix(rlaplace(500*20), nrow=500)
> t1<-apply(data,1,mean)
> t2<-apply(data,1,median)
> t3<-apply(data,1,midrange)
> var(t1)/var(t2)
[1] 1.954667
> var(t1)/var(t3)
[1] 0.01313633
```

2.1.7 Estimaatin varianssin BOOTSTRAP-estimointi

Oletetaan, että $X = \{x_1, \ldots, x_n\}$ on satunnaisotos jakaumasta F ja että $\hat{\Delta} = \hat{\Delta}(X)$ on parametrin $\Delta = \Delta(F)$ estimaatti. Miten voidaan estimoida varianssia ja harhaa,

$$B_F(\hat{\Delta}) \text{ ja } \mathrm{Var}_F(\hat{\Delta}),$$

kun F on tuntematon.

Huomaa ensin, että tuntemattoman kertymäfunktion F luonnollinen estimaatti (parametrittomassa mallissa) on otoskertymäfunktio

$$F_n(x) = \frac{1}{n}\#\{x_i \leq x\}.$$

Estimaatin $\hat{\Delta}$ harhaa ja varianssia otoskoolla n voidaan tutkia ns. **BOOTSTRAP-tekniikalla**:

- Generoi M otosta otoskoolla n jakaumasta F_n (M n:n suuruista satunnaisotosta palauttaen joukosta X),

$$X_1^*, \ldots, X_M^*,$$

- Laske otoksiin liittyvät estimaattien arvot

$$\hat{\Delta}(X_1^*), \ldots, \hat{\Delta}(X_M^*).$$

- Estimaatin varianssin estimaatti:

$$\frac{1}{M}\sum_{i=1}^{M}[\hat{\Delta}(X_i^*)]^2 - \left[\frac{1}{M}\sum_{i=1}^{M}\hat{\Delta}(X_i^*)\right]^2 = \frac{1}{M}\sum_{i=1}^{M}\left(\hat{\Delta}(X_i^*) - \frac{1}{M}\sum_{j=1}^{M}\hat{\Delta}(X_j^*)\right)^2$$

- Harhan $E_F(\hat{\Delta}) - \Delta$ estimaatti

$$\frac{1}{M}\sum_{i=1}^{M}\hat{\Delta}(X_i^*) - \hat{\Delta}(X)$$

Funktioita:

```
HL<-function(d)
{
  n<-length(d)
  apu<-matrix(rep(d,n),n)
  W<-(apu+t(apu))/2
  w<-c(W[row(W)<=col(W)])
  median(w)
}
boots1<-function(d,M)
{
  n<-length(d)
  u<-NULL
  for (i in 1:M) u[i]<-mean(sample(d,size=n,replace=T))
  list(EST=mean(d), BIAS=mean(u)-mean(d), VAR=var(u))
}
boots2<-function(d,M)
{
  n<-length(d)
  u<-NULL
```

```
  for (i in 1:M) u[i]<-median(sample(d,size=n,replace=T))
  list(EST=median(d), BIAS=mean(u)-median(d), VAR=var(u))
}
boots3<-function(d,M)
{
  n<-length(d)
  u<-NULL
  for (i in 1:M) u[i]<-HL(sample(d,size=n,replace=T))
  list(EST=HL(d), BIAS=mean(u)-HL(d), VAR=var(u))
}

------
> y<-rnorm(200)
> boots1(y,1000)$VAR
[1] 0.005748212
> boots2(y,1000)$VAR
[1] 0.007527701
> boots3(y,1000)$VAR
[1] 0.005936744
> y<-rlaplace(50)
> boots1(y,5000)$VAR
[1] 0.04830046
> boots2(y,5000)$VAR
[1] 0.02755389
> boots3(y,5000)$VAR
[1] 0.04080261
```

2.1.8 Estimaatin murtumispiste

Olkoot

$$X = \{x_1, \ldots, x_n\}$$

alkuperäiset 'hyvät' havainnot ja

$$\hat{\Delta} = \hat{\Delta}(X)$$

valittu estimaatti. Olkoon

$$X' = \{x'_1, \ldots, x'_m, x_{m+1}, \ldots, x_n\}$$

uusi 'saastunut' havaintoaineisto, missä m ensimmäistä havaintoa on korvattu uusilla 'huonoilla' havainnoilla.
Maksimiharha, joka saadaan aikaan huonoilla havainnoilla on

$$\mathrm{Bias}(X; m) = \sup_{X'} |\hat{\Delta}(X') - \hat{\Delta}(X)|.$$

Silloin estimaatin $\hat{\Delta}$ **murtumispiste** (finite-sample breakdown point) aineiston X tapauksessa on

$$BP(\hat{\Delta}) = \frac{1}{n} \inf_{1 \leq m \leq n} \{m \ : \ \text{Bias}(X; m) = \infty\}.$$

Huomaa, että murtumispiste riippuu otoskoosta n ja aineistosta X. Tavallisesti murtumispiste määritellään raja-arvona $\lim_{n\to\infty} BP(\hat{\Delta})$.

Esimerkki 2.1.6
Keskiarvon $\bar{x}$, mediaanin $Med(X)$ ja HL-estimaatin

$$Med_{i<j}\Big(\frac{x_i + x_j}{2}\Big)$$

murtumispisteet ovat 0, 1/2 ja $1 - 1/\sqrt{2} \approx 0.29$.

2.1.9 Influenssifunktio

Olkoon T funktionaali, $F(y)$ kertymäfunktio ja $\delta_x(y) = 0$ kun $y < x$ ja $\delta_x(y) = 1$ kun $y \geq x$. Funktionaalin T **influenssifunktio** on tällöin

$$\begin{aligned} IF(x; F, T) &= \lim_{t\to 0} \frac{T((1-t)F + t\delta_x) - T(F)}{t} \\ &= \frac{d}{dt} T((1-t)F + t\delta_x)\Big|_{t=0} \end{aligned}$$

Influenssifunktio kuvaa yhden lisähavainnon x vaikutusta funtionaalin T arvoon.

Esimerkki 2.1.7
Jakauman F odotusarvo ja mediaani voidaan kirjoittaa muodossa $T(F) = \int u dF(u)$ ja $T(F) = F^{-1}(\frac{1}{2})$. Olkoon F $N(0,1)$-jakauman kertymäfunktio. Tällöin saadaan odotusarvon influenssifunktioksi $IF(x; F, T) = x$ ja mediaanin influenssifunktioksi $IF(x; F, T) = -\sqrt{\pi/2}\ I(x < 0) + \sqrt{\pi/2}\ I(x > 0)$.

Oletetaan, että estimaattori T_n on funktionaali ($T_n(x_1, \ldots, x_n) = T(F_n)$ kaikilla n, kaikilla otoksilla $(x_1, \ldots, x_n)$ ja vastaavalla otoskertymäfunktiolla F_n). Oletetaan myös, että $T(F_n) \to T(F)$ missä F on x_i:n kertymäfunktio.

Alkuperäisen otoksen $(x_1, \ldots, x_n)$ otoskertymäfunktio on

$$F_n(y) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq y)$$

Otoksen $(x_1, \ldots, x_n, x)$ otoskertymäfunktio on

$$\begin{aligned} F_{n+1}(y) &= \frac{1}{n+1}\left[\sum_{i=1}^{n} I(x_i \leq y) + I(x \leq y)\right] \\ &= \frac{1}{n+1}[nF_n(y) + \delta_x(y)] \\ &= \left(1 - \frac{1}{n+1}\right)F_n(y) + \frac{1}{n+1}\delta_x(y) \end{aligned}$$

Tällöin estimaatin T_n **empiirinen influenssifunktio** on

$$\begin{aligned} IF_n(x; F_n, T_n) &= \frac{T((1-\frac{1}{n+1})F_n + \frac{1}{n+1}\delta_x) - T(F_n)}{1/(n+1)} \\ &= (n+1)\left[T\left(\left(1 - \frac{1}{n+1}\right)F_n + \frac{1}{n+1}\delta_x\right) - T(F_n)\right] \\ &= (n+1)[T_{n+1}(x_1, \ldots, x_n, x) - T_n(x_1, \ldots, x_n)] \end{aligned}$$

Empiirisen influenssifunktion avulla tarkastellaan yhden (poikkeavan) lisähavainnon vaikutusta estimaattiin $\hat{\Delta}(X)$, missä jälleen $X = \{x_1, \ldots, x_n\}$.

Robusteilla estimaateilla influenssifunktio on **rajoitettu** (suoja outlier-havaintoja vastaan) ja **jatkuva** (suoja inlier-havaintoja vastaan).

Esimerkki 2.1.8
Otoskeskiarvo: $T_n(X) = \frac{1}{n}\sum_{i=1}^{n} x_i$ ja $IF_n(x; F_n, T_n) = x - T_n(X)$.

Otosmediaani: $T_n(X) = Med\{x_1, \ldots, x_n\}$.
Kun $n = 2k + 1$, niin

$$\begin{aligned} IF_n(x; F_n, T_n) &= (n+1)\left[\frac{x_{(k)} - x_{(k+1)}}{2} I(x < x_{(k)})\right. \\ &\quad + \frac{x - x_{(k+1)}}{2} I(x_{(k)} \leq x \leq x_{(k+2)}) \\ &\quad \left. + \frac{x_{(k+2)} - x_{(k+1)}}{2} I(x > x_{(k+2)})\right] \end{aligned}$$

Kun $n = 2k$, niin

$$\begin{aligned} IF_n(x; F_n, T_n) &= (n+1)\left[\frac{x_{(k)} - x_{(k+1)}}{2} I(x < x_{(k)})\right. \\ &\quad + \frac{2x - x_{(k)} - x_{(k+1)}}{2} I(x_{(k)} \leq x \leq x_{(k+1)}) \\ &\quad \left. + \frac{x_{(k+1)} - x_{(k)}}{2} I(x > x_{(k+1)})\right] \end{aligned}$$

```
Keskiarvon, mediaanin ja HL-estimaatin influenssifunktiot

>IF1<-function(x,d)
 {n<-length(d)
  (n+1)*(mean(c(d,x))-mean(d))}
>IF2<-function(x,d)
 {n<-length(d)
  (n+1)*(median(c(d,x))-median(d))}
>IF3<-function(x,d)
 {n<-length(d)
  (n+1)*(HL(c(d,x))-HL(d))}

> d<-c(-54,-18,-29,-28,-35,-47,-11,5,-35,-65)
> x<-((1:200)-100)
> y1<-NULL
> y2<-NULL
> y3<-NULL
> for (i in 1:200) y1[i]<-IF1(x[i],d)
> for (i in 1:200) y2[i]<-IF2(x[i],d)
> for (i in 1:200) y3[i]<-IF3(x[i],d)

> plot(x,y1,xlim=c(-100,100),ylim=c(-100,100),type="l")
> lines(x,y2,col="red")
> lines(x,y3,col="blue")

> postscript(file="kuva.ps",width=4,height=4,horizontal=FALSE)
> plot(x,y1,xlim=c(-100,100),ylim=c(-100,100),type="l")
> lines(x,y2,col="red")
> lines(x,y3,col="blue")
> dev.off()

> d<-rnorm(200)
> for (i in 1:200) y1[i]<-IF1(x[i],d)
> for (i in 1:200) y2[i]<-IF2(x[i],d)
> for (i in 1:200) y3[i]<-IF3(x[i],d)
> plot(x,y1,xlim=c(-2,2),ylim=c(-2,2),type="l")
> lines(x,y2,col="red")
> lines(x,y3,col="blue")
```

2.1.10 Havainnon empiirinen standardoitu vaikutus

Yksittäisen havainnon x_i 'empiirinen standardoitu vaikutus' estimaatin arvoon on

$$\hat{\Delta}_i = n\hat{\Delta}(x_1, \ldots, x_n) - (n-1)\hat{\Delta}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$$

(engl. jackknifed pseudovalues). Merkitään edelleen

$$\hat{\Delta}^*(X) = \frac{1}{n}\sum_{i=1}^{n}\hat{\Delta}_i$$

Estimaatit

$$\hat{\Delta}^*(X) \quad \text{ja} \quad \hat{\Delta}(X)$$

ovat tällöin 'asymptoottisesti ekvivalentteja, mutta estimaatilla $\hat{\Delta}^*(X)$ on pienempi harha.
Edelleen

$$\frac{1}{n(n-1)}\sum_{i=1}^{n}(\hat{\Delta}_i - \hat{\Delta}^*(X))^2,$$

on $\hat{\Delta}(X)$:n varianssin **jackknife-estimaatti.**

Δ:n likimääräinen $100(1-\alpha)\%$ luottamusvälin jackknife-estimaatti on

$$\hat{\Delta}^*(X) \pm t_{\alpha/2,n-1}\sqrt{\frac{1}{n(n-1)}\sum_{i=1}^{n}(\hat{\Delta}_i - \hat{\Delta}^*(X))^2}$$

2.1.11 Testien ja luottamusvälien tarkastelua

Testien validisuus ja tehokkuus

Tarkastellaan kaksisuuntaisen testisuureen $|T|$ tuottaman p-arvon jakaumaa. Jos havaittu testisuureen arvo on $t = T(X)$, $X = \{x_1, \ldots, x_n\}$, niin p-arvo on

$$p = p(X) = P_0(|T| \geq |t|)$$

missä siis todennäköisyys lasketaan nollahypoteesin mukaisesta jakaumasta. Jos testisuure on validi, p-arvon (satunnaismuuttuja) todennäköisyysjakauma on $U(0,1)$, tasainen jakauma välillä $(0,1)$. Tätä voi tutkia simuloimalla:

- Generoi N otosta otoskoolla n nollahypoteesijakaumasta F:

$$X_1, \ldots, X_N.$$

- Laske otoksiin liittyvät p-arvot

$$p(X_1), \ldots, p(X_N).$$

- Tutki validisuutta graafisesti p-arvon kertymäfunktio- tai tiheysfunktioestimaatin avulla, p-p-plotilla tai konstruoimalla yhteensopivuustesti.

Mitä voimakkaammasta testistä on kysymys, sitä pienempiä p-arvoja se tuottaa vastahypoteesin vallitessa. Tätä tutkitaan simuloimalla seuraavasti:

- Generoi N otosta otoskoolla n vastahypoteesijakaumasta G:

$$X_1, \ldots, X_N.$$

- Laske otoksiin liittyvät p-arvot

$$p(X_1), \ldots, p(X_N).$$

- Tutki voimakkuutta graafisesti p-arvon kertymäfunktioestimaatin avulla, esim. kertymäfunktion arvo pisteessä 0.10 estimoi tason 0.10 testin voimakkuutta kyseisen vastahypoteesijakauman tapauksessa. Voit myös estimoida p-arvon odotusarvoa ja/tai mediaania.

Luottamusvälien validisuus ja tehokkuus

Jos luottamusväli

$$(\hat{\Delta}_L(X), \hat{\Delta}_U(X))$$

on validi, sen peitetodennäköisyys

$$P(\hat{\Delta}_L < \Delta < \hat{\Delta}_U)$$

on haluttu 0.90 (tai 0.95, tms.) eikä riipu tuntemattomasta parametrin Δ arvosta. Tehokas luottamusväli on puolestaan mahdollisimman lyhyt. Näitä ominaisuuksia voi jälleen tutkia simuloimalla seuraavasti.

- Generoi N otosta otoskoolla n nollahypoteesijakaumasta F:

$$X_1, \ldots, X_N.$$

- Laske otoksiin liittyvät luottamusvälit

$$(\hat{\Delta}_L(X_1), \hat{\Delta}_U(X_1)), \ldots, (\hat{\Delta}_L(X_N), \hat{\Delta}_U(X_N)).$$

- Tutki validisuutta peitetodennäköisyysestimaatin

$$\frac{1}{N}\#\{\hat{\Delta}_L(X_i) < \Delta < \hat{\Delta}_U(X_i)\}$$

  avulla.

- Tutki tehokkuutta estimoimalla keskimääräistä välin pituutta estimaattina

$$\frac{1}{N}\sum_{i=1}^{N}\{\hat{\Delta}_U(X_i) - \hat{\Delta}_L(X_i)\}$$

P-arvojen tarkastelu:

```
p1<-function(d)
{
#2-suuntainen yhden otoksen t-testi
  n<-length(d)
  t<-sqrt(n)*mean(d)/sd(d)
  2*min(pt(t,n-1),1-pt(t,n-1))
}
p2<-function(d)
{
#2-suuntainen yhden otoksen merkkitesti
  n<-length(d)
  S<-sum(d>0)
  2*min(pbinom(S,n,0.5),1-pbinom(S-1,n,0.5),0.5)
}
p3<-function(d)
{
#2-suuntainen Wilcoxonin pareittaisten otosten testi
  n<-length(d)
  apu<-matrix(rep(d,n),n)
  W<-(apu+t(apu))/2
  w<-c(W[row(W)<=col(W)])
  test<-sum(w>0)
  E<-n*(n+1)/4
  SD<- sqrt(n*(n+1)*(2*n+1)/24)
  z<-(test-E)/SD
  2*min(pnorm(z),1-pnorm(z))
}
```

```
> cdf<-function(x,d)
  {
  #otoskertymafunktion arvo pisteessa x
    n<-length(d)
    sum(d<=x)/n
  }
> n<-50
> N<-1000
> p1data<-apply(matrix(rnorm(n*N), nrow=N),1,p1)
> mean(p1data)
[1] 0.4987704
> median(p1data)
[1] 0.5023782
> mean(p1data<0.10)
```

```
[1] 0.087
> mean(p1data<0.20)
[1] 0.197
>x<-(1:100)/101
> y1<-NULL
> for (i in 1:100) y1[i]<-cdf(x[i],p1data)
> plot (x,y1,xlim=c(0,1),ylim=c(0,1),type="l")
```

Luottamusvalien tarkastelu:

```
t.ALA<-function(d)
{
  n<-length(d)
  mean(d)+qt(0.025,n-1)*sd(d)/sqrt(n)
}
t.YLA<-function(d)
{
  n<-length(d)
  mean(d)+qt(0.975,n-1)*sd(d)/sqrt(n)
}
s.ALA <-function(d)
{
  n<-length(d)
  i<-1
  apu<-pbinom(i-1,n,0.5)
  while(apu<0.025)
  {
    i<-i+1
    apu<-pbinom(i-1,n,0.5)
  }
  ii<-i-1
  CL<-1-2*pbinom(ii-1,n,0.5)
  d0<-sort(d)
  d0[ii]
}
s.YLA <-function(d)
{
  n<-length(d)
  i<-1
  apu<-pbinom(i-1,n,0.5)
  while(apu<0.025)
  {
    i<-i+1
    apu<-pbinom(i-1,n,0.5)
```

```
  }
  ii<-i-1
  CL<-1-2*pbinom(ii-1,n,0.5)
  d0<-sort(d)
  d0[n+1-ii]
}


> n<-20
> N<-1000
> data<-matrix(rnorm(n*N), nrow=N)
> mean(apply(data,1,t.ALA))
[1] -0.4613165
> ALA1<-apply(data,1,t.ALA)
> YLA1<-apply(data,1,t.YLA)
> ALA2<-apply(data,1,s.ALA)
> YLA2<-apply(data,1,s.YLA)
> mean(YLA1-ALA1)           #lasketaan luottamusvalin pituus
[1] 0.9274356
> mean(YLA2-ALA2)
[1] 1.185247
> mean((ALA1<0)*(YLA1>0))   #lasketaan peitetodennakoisyys
[1] 0.955
> mean((ALA2<0)*(YLA2>0))
[1] 0.96
> data<-matrix(rlaplace(n*N), nrow=N)
> ALA1<-apply(data,1,t.ALA)
> ALA2<-apply(data,1,s.ALA)
> YLA1<-apply(data,1,t.YLA)
> YLA2<-apply(data,1,s.YLA)
> mean(YLA1-ALA1)
[1] 1.295329
> mean(YLA2-ALA2)
[1] 1.247369
> mean((ALA1<0)*(YLA1>0))
[1] 0.957
> mean((ALA2<0)*(YLA2>0))
[1] 0.963
```

2.1.12 Yhden otoksen järjestyslukutestit

- $X = \{x_1, \ldots, x_n\}$ on satunnaisotos Δ:n suhteen symmetrisestä jakaumasta; kiinnostava nollahypoteesi on $H_0 : \Delta = 0$.

- Järjestyslukufunktio (engl. *Rank function*)

$$R(x, X) = \frac{1}{2} + \sum_{i=1}^{n} \left\{ I(x_i < x) + \frac{1}{2} I(x_i = x) \right\}$$

  Huom! $R(x, X)$ ottaa huomioon myös sidokset.

- Järjestysluvut

$$R_i = R(x_i, X), \quad i = 1, \ldots n$$

- Merkitään

$$x_i^+ = |x_i| \ \text{ ja } \ s_i = \text{sign}(x_i), \quad i = 1, \ldots n.$$

  Siis $x_i = s_i x_i^+$, $i = 1, \ldots, n$.

- Merkitään

$$X^+ = \{x_1^+, \ldots, x_n^+\} \ \text{ ja } \ R_i^+ = R(x_i^+, X^+), \ i = 1, \ldots, n.$$

- Wilcoxonin yhden otoksen testisuure (*signed-rank test*)

$$W_0 = \sum_{i=1}^{n} [s_i R_i^+]$$

- Aikaisemmin

$$W = \sum_{i \le j} I(x_i + x_j > 0) \ \text{ ja } \ S = \sum_{i=1}^{n} I(x_i > 0)$$

- **Tulos.**

$$W - \frac{n(n+1)}{4} + \frac{1}{2} \sum_{i \le j} I(x_i + x_j = 0) = W_0/2.$$

- **Seuraus.** Kun havainnot tulevat jatkuvasta jakaumasta, niin

$$P\left(\sum_{i \le j} I(x_i + x_j = 0) = 0 \right) = 1$$

  ja

$$W - \frac{n(n+1)}{4} \sim W_0/2.$$

  Huom! Vaikka havainnot tulevat jatkuvasta jakaumasta, niin havaintoarvojen mittaukset voidaan suorittaa vain äärellisellä tarkkuudella. Oletetaan, että mittaustarkkuus on ”riittävä”, jotta sidoksia ei syntyisi.

- Valitaan pisteluvut

$$a(1) \leq a(2) \leq \ldots \leq a(n).$$

Usein $a(i) = \psi(i/(n+1))$, missä $\psi(u)$, $u \in (0,1)$, on niin sanottu pistelukufunktio (*score function*)

- Valittuja pistelukuja vastaava järjestyslukutestisuure on

$$T = \sum_{i=1}^{n} [s_i a(R_i^+)].$$

- Nollahypoteesin H_0 vallitessa s_i ja R_i^+ ovat riippumattomia ja T on jakautunut kuten $\sum[s_i a(i)]$. Siis T:n jakauma ei nollahypoteesin vallitessa riipu x_i:n jakaumasta. Erityisesti

$$E_0(T) = 0 \;\text{ ja }\; \mathrm{Var}_0(T) = \sum (a(i))^2.$$

Normaaliaproksimaatiota tai simulointia voi käyttää likimääräisen p-arvon laskemiseen.

- Erikoistapauksina saadaan merkkitesti ($a(i) = 1$, $i = 1, \ldots, n$) ja Wilcoxonin testi ($a(i) = i$, $i = 1, \ldots, n$). Niin sanottu van der Waerdenin testi, pistelukufunktiona $\psi(u) = \Phi^{-1}((u+1)/2)$ on optimaalinen normaalijakauman tapauksessa.

- Pistelukuja $a(1) \leq \ldots \leq a(n)$ vastaava estimaatti ja luottamusväli on mahdollista konstruoida kuten Wilcoxonin testin tapauksessa. (Pisteestimaatti: Arvo, jota (kaksisuuntaisella testillä) testattaessa saadaan suurin mahdollinen p-arvo. 95 %:n luottamusväli: Ne arvot, joita vastaava kaksisuuntaisen testin p-arvo on suurempi kuin 0.05.)

'SIGN CHANGE'-testejä

```
p.sc<-function(d,N)
{
  n<-length(d)
  kerroin<-matrix(2*rbinom(n*N,1,0.5)-1,nrow=N)
  t0<-sum(d)
  t<-kerroin%*%d
  apu<-mean(t<t0)
  2*min(apu,1-apu)
}

p1000<-function(d)
{
```

```
  n<-length(d)
  N<-1000
  kerroin<-matrix(2*rbinom(n*N,1,0.5)-1,nrow=N)
  t0<-sum(d)
  t<-kerroin%*%d
  apu<-mean(t<t0)
  2*min(apu,1-apu)
}

p.waerden<-function(d,N)
{
  n<-length(d)
  s<-(d>0)-(d<0)
  r<-rank(abs(d))
  r<-qnorm(0.5+0.5*r/(n+1))
  kerroin<-matrix(2*rbinom(n*N,1,0.5)-1,nrow=N)
  t0<-sum(s*r)
  t<-kerroin%*%r
  apu<-mean(t<t0)
  2*min(apu,1-apu)
}

p.waerden1000<-function(d)
{
  n<-length(d)
  N<-1000
  s<-(d>0)-(d<0)
  r<-rank(abs(d))
  r<-qnorm(0.5+0.5*r/(n+1))
  kerroin<-matrix(2*rbinom(n*N,1,0.5)-1,nrow=N)
  t0<-sum(s*r)
  t<-kerroin%*%r
  apu<-mean(t<t0)
  2*min(apu,1-apu)
}


> n<-20
> N<-1000

> data<-matrix(rnorm(n*N),nrow=N)
> mean(apply(data,1,p1)<=0.1)
[1] 0.105
> mean(apply(data,1,p1000)<=0.1)
```

```
[1] 0.111
> mean(apply(data,1,p.waerden1000)<=0.1)
[1] 0.112

> data<-matrix(rnorm(n*N)+0.3,nrow=N)
> mean(apply(data,1,p1)<=0.1)
[1] 0.369
> mean(apply(data,1,p1000)<=0.1)
[1] 0.369
> mean(apply(data,1,p.waerden1000)<=0.1)
[1] 0.37

> data<-matrix(rlaplace(n*N),nrow=N)
> mean(apply(data,1,p1)<=0.1)
[1] 0.095
> mean(apply(data,1,p1000)<=0.1)
[1] 0.096
> mean(apply(data,1,p.waerden1000)<=0.1)
[1] 0.115

> data<-matrix(rlaplace(n*N)+0.3,nrow=N)
> mean(apply(data,1,p1)<=0.1)
[1] 0.254
> mean(apply(data,1,p1000)<=0.1)
[1] 0.252
> mean(apply(data,1,p.waerden1000)<=0.1)
[1] 0.276
```

2.1.13 M-estimaatit ja testit

- $X = \{x_1, \ldots, x_n\}$ on satunnaisotos Δ:n suhteen symmetrisestä jakaumasta; kiinnostava nollahypoteesi on $H_0 : \Delta = 0$.

- Valitaan kasvava ja pariton pistelukufunktio $\psi(x)$

- Testisuure testattaessa nollahypoteesia H_0 on

$$T = \sum_{i=1}^{n} \psi(x_i) = \sum_{i=1}^{n} [s_i \psi(x_i^+)]$$

- Ehdollinen testi (*sign change test*): Ehdollistetaan havaituille arvoille $x_1^+, \ldots, x_n^+$. Muuttujan T nollahypoteesin mukainen ehdollinen jakauma voidaan nyt konstruoida kuten järjestyslukutestien tapauksessa (2^n

erilaista yhtä todennäköistä arvoa). Erityisesti

$$E_0(T) = 0 \text{ ja } \mathrm{Var}_0(T) = \sum_{i=1}^{n} [\psi(x_i)]^2.$$

Havaittuun T:n arvoon liittyvä likimääräinen p-arvo voidaan laskea normaaliaproksimaation avulla tai simuloimalla.

- **Määritelmä 1.** M-estimaatti ratkaisee yhtälön

$$\sum_{i=1}^{n} \psi(x_i - \hat{\Delta}) = 0.$$

Ratkaisu löytyy usein iteroimalla. Edellä oleva yhtälö antaa yksikäsitteisen estimaatin, jos $\psi(x)$ on aidosti kasvava ($x > y \Rightarrow \psi(x) > \psi(y)$). Seuraavalla tavalla määritelty M-estimaatti on aina yksikäsitteinen:

- **Määritelmä 2.** M-estimaatti on

$$\hat{\Delta} = \frac{\Delta^* + \Delta^{**}}{2},$$

missä

$$\Delta^* = \sup\left\{\Delta \ : \ \sum_{i=1}^{n} \psi(x_i - \Delta) > 0\right\}$$

ja

$$\Delta^{**} = \inf\left\{\Delta \ : \ \sum_{i=1}^{n} \psi(x_i - \Delta) < 0\right\}.$$

- M-estimaatille $\hat{\Delta}(x_1, \ldots, x_n)$ pätee aina:

  (a) $\hat{\Delta}(-x_1, \ldots, -x_n) = -\hat{\Delta}(x_1, \ldots, x_n)$

  (b) $\hat{\Delta}(x_1 + c, \ldots, x_n + c) = \hat{\Delta}(x_1, \ldots, x_n) + c$

  (c) Jos x_i:n jakauma on symmetrinen Δ:n suhteen, niin myös $\hat{\Delta}(x_1, \ldots, x_n)$:n jakauma on symmetrinen Δ:n suhteen.

Määritelmien 1 ja 2 mukaiset estimaatit eivät ole välttämättä skaalausekvivariantteja, eli ominaisuus $\hat{\Delta}(cx_1, \ldots, cx_n) = c\hat{\Delta}(x_1, \ldots, x_n)$ ei aina pidä paikkaansa. Seuraavalla tavalla määritelty M-estimaatti on myös skaalausekvivariantti:

- **Määritelmä 3.** M-estimaatti on

$$\hat{\Delta} = \frac{\Delta^* + \Delta^{**}}{2},$$

missä

$$\Delta^* = \sup\left\{\Delta \ : \ \sum_{i=1}^{n} \psi\left(\frac{x_i - \Delta}{S}\right) > 0\right\}$$

ja

$$\Delta^{**} = \inf\left\{\Delta \ : \ \sum_{i=1}^{n} \psi\left(\frac{x_i - \Delta}{S}\right) < 0\right\}.$$

ja S on skaalaestimaattori jolle pätee $S(cx_1+d, \ldots, cx_n+d) = cS(x_1, \ldots, x_n)$.

- Optimaalinen pistelukufunktio on

$$L(y) = -\frac{f_0'(y)}{f_0(y)},$$

missä $f_0(y)$ on nollahypoteesin mukainen tiheysfunktio. Pistelukufunktion L antama estimaatti on **suurimman uskottavuuden estimaatti** (*maximum likelihood estimate*). Normaalijakauman ($N(0,1)$) tapauksessa $L(y) = y$ ja Laplacen jakauman ($f(y) = (1/2)\exp(-|y|)$) tapauksessa $L(y) = \text{sign}(y)$.

- Oletetaan, että $S(x_1, \ldots, x_n) \to S_{F_0}$, kun $n \to \infty$. Yleisten oletusten vallitessa valintaa ψ vastaava estimaatti

$$\hat{\Delta} \sim AN\left(\Delta, \frac{1}{n}\frac{b}{a^2}\right)$$

missä

$$a = \text{Cov}_0(\psi(y/S_{F_0}), L(y/S_{F_0})) \quad \text{ja} \quad b = S_{F_0}^2 \text{Var}_0(\psi(y/S_{F_0})).$$

Asymptoottinen suhteellinen tehokkuus suurimman uskottavuuden estimaattiin verrattuna on $a^2/(b \cdot i)$ missä $i = \text{Var}_0(L(y/S_{F_0}))$.

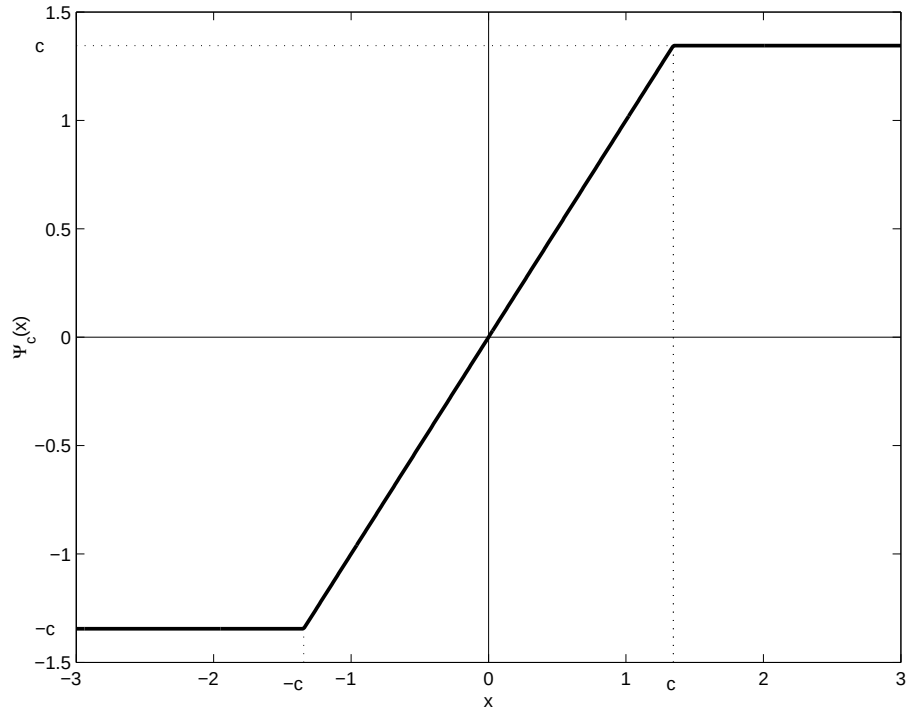
Huberin M-estimaatti

Huberin (1964) ψ-funktio (katso kuva 2.1) määritellään seuraavasti:

$$\psi_c(x) = \begin{cases} x, & |x| \le c \\ \text{sign}(x)c, & |x| > c \end{cases}$$

Käytettäessä M-estimaatin määritelmiä 1 ja 2 ja Huberin ψ-funktiota, niin saatava estimaattori ei ole skaalausekvivariantti. Huberin M-estimaatti lasketaan käyttämällä määritelmää 3.

```
huber.psi<-function(x,c)
{
#Huberin psi-funktio
```



Kuva 2.1: Huberin Ψ-funktio.

```
  -c*(x< -c)+x*(abs(x)<= c)+c*(x> c)
}

dhuber.psi<-function(x,c)
{
#Huberin psi-funktion derivaatta
  1*(abs(x)<= c)
}

huber<-function(d,c)
{
  s<-1.483*median(abs(d-median(d)))   #skaalaestimaattina MAD
  d<-d/s
  theta<-median(d)  #alkuarvona mediaani
  delta<-0.1
  iter<-1
  #lasketaan estimaatti Newtonin menetelmalla
  while(abs(delta) >= 0.0001 & iter<100)
  {
    delta<-mean(huber.psi(d-theta,c))/mean(dhuber.psi(d-theta,c))
```

```
    theta<-theta+delta
    iter<-iter+1
  }
  s*theta
}
```

```
> hub<-function(d) {huber(d,1.345)}  #valinta c=1.345 antaa 95%
> n<-20                              #asympt.suht.tehokkuuden
> N<-1000                            #keskiarvoon verrattuna
> data<-matrix(rnorm(n*N),nrow=N)    #normaalijakaumatapauksessa
> var(apply(data,1,mean))
[1] 0.06005056
> var(apply(data,1,hub))
[1] 0.06209799
> data<-matrix(rlaplace(n*N),nrow=N)
> var(apply(data,1,mean))
[1] 0.09506568
> var(apply(data,1,hub))
[1] 0.06813584
```

2.1.14 Kertausta

- L_2-normi: $||\boldsymbol{d}||_2 = \sqrt{\sum d_i^2}$
- L_1-normi: $||\boldsymbol{d}||_1 = \sum d_i^+$
- Painotettu L_1-normi: $||\boldsymbol{d}||_3 = \sum[R_i^+ d_i^+]$
- Yleistetty painotettu L_1-normi: $||\boldsymbol{d}||_\psi = \sum[\psi(R_i^+/(n+1))d_i^+]$, missä ψ on valittu pistelukufunktio
- Sijaintiestimaatti minimoi normin $||\boldsymbol{d} - \Delta\mathbf{1}_n||_2$, normin $||\boldsymbol{d} - \Delta\mathbf{1}_n||_1$, jne., missä $\mathbf{1}_n$ on n-vektori alkioina 1.
- Testit saadaan 'objektifunktion' derivaatan avulla,

$$T = \sum[\psi(R_i^+)\text{sign}(d_i)]$$

- Luottamusvälit saadaan 'kääntämällä' testi. Ensin etsitään sellainen arvo c, että

$$P_0(-c \leq T \leq c) = 1 - \alpha.$$

  Tason $1-\alpha$ luottamusväli koostuu niistä Δ:n arvoista, joita testattaessa testisuure $T(\Delta) \in [-c, c]$

2.2 Riippumattomat otokset

2.2.1 Koeasetelma

Taulukko 2.3: Havaintoaineisto riippumattomien otosten tapauksessa.

Käsittely	Otos
A	$x_1, \ldots, x_m$
B	$y_1, \ldots, y_n$

- Yksilöt arvottu käsittelyille; kaikki $\binom{m+n}{m}$ 'jakoa' yhtä todennäköisiä.
- Parametrinen malli: Satunnaisotokset riippumattomia ja peräisin parametrisista jakaumista (esim. normaalijakaumista), jotka poikkeavat toisistaan vain sijainnin suhteen.
- Parametriton malli: Satunnaisotokset riippumattomia ja peräisin jakaumista, jotka poikkeavat vain sijainnin suhteen.

Huomaa, että edellä kuvaillussa koeasetelmassa mittaustuloksiin liittyviä satunnaisuuden (vaihtelun) lähteitä ovat (I) yksilöiden välinen vaihtelu populaatiossa, (II) käsittelyjen erosta johtuva vaihtelu ja (III) arvonnasta (satunnaistamisesta) johtuva vaihtelu. On mahdollista konstruoida testejä, niin sanottuja permutaatiotestejä, jotka nojaavat vain viimeksimainittuun vaihteluun.

Merkitään

$$N = m + n \quad \text{ja} \quad \lambda = \frac{m}{N}.$$

Myöhemmin nähdään, että λ:n valinnalla on vaikutusta estimaattien ja testien tehokkuuteen.

Esimerkki 2.2.1
Tutkittaessa pölyaltistuksen vaikutusta keuhkoihin eräässä koe-eläinlaboratoriossa satunnaistettiin 28 rottaa ryhmiin A ja B, kumpaankin 14. Ryhmän B (koeryhmä) rottia pidettiin kolmen kuukauden ajan kopeissa, joiden ilmakehässä ylläpidettiin tietty pölypitoisuus, kun taas ryhmän A (vertailuryhmä) rottia pidettiin pölyttömissä kopeissa samanpituinen aika. Muuten rottien olosuhteet olivat samanlaiset kummassakin ryhmässä. Altistusajan kuluessa kuoli koeryhmästä yksi rotta ja vertailuryhmästä kaksi. Kolmen kuukauden kuluttua jäljellä olevat rotat tapettiin ja niiden keuhkojen painot mitattiin. Tulokset olivat seuraavat (keuhkojen paino grammoina):

Taulukko 2.4: Havaintoaineisto riippumattomien otosten tapauksessa.

Ryhmä	Keuhkojen painot
A	5.1, 3.8, 5.0, 6.4, 5.0, 4.0, 3.2, 4.4, 4.1, 5.6, 4.8, 4.5
B	5.4, 5.6, 5.4, 6.5, 6.8, 6.0, 4.2, 4.4, 4.8, 4.9, 6.9, 4.9, 5.2

2.2.2 Normaalijakaumaoletus

Oletukset:

- $x_1, \ldots, x_m$ satunnaisotos jakaumasta $N(\mu, \sigma^2)$
- $y_1, \ldots, y_n$ satunnaisotos jakaumasta $N(\mu + \Delta, \sigma^2)$
- otokset riippumattomia

Käsittelyjen eroa kuvaava parametri on Δ.
Silloin

- AINEISTO:

$$x_1, \ldots, x_m \quad \rightarrow \quad \bar{x}, s_x^2$$
$$y_1, \ldots, y_n \quad \rightarrow \quad \bar{y}, s_y^2$$

  ja edelleen

$$s_x^2, s_y^2 \quad \rightarrow \quad s^2 = \frac{1}{m+n-2}[(m-1)s_x^2 + (n-1)s_y^2]$$

  ($(\bar{x}, \bar{y}, s^2)$ on tyhjentävä ja täydellinen tunnusluku.)

- TESTAUS (t-testi):

$$t = \sqrt{N\lambda(1-\lambda)} \cdot \frac{\bar{y} - \bar{x}}{s} \ \sim \ t(m+n-2), \text{ kun } H_0 : \Delta = 0 \text{ tosi}$$

- PISTE-ESTIMOINTI (suluissa keskivirhe, estimaatin keskihajonta):

$$\Delta : \quad \bar{y} - \bar{x} \quad \left(\pm \frac{s}{\sqrt{N\lambda(1-\lambda)}} \right)$$

- LUOTTAMUSVÄLI:

$$\left([\bar{y} - \bar{x}] - t_{\alpha/2} \cdot \frac{s}{\sqrt{N\lambda(1-\lambda)}} \ , \ [\bar{y} - \bar{x}] + t_{\alpha/2} \cdot \frac{s}{\sqrt{N\lambda(1-\lambda)}} \right)$$

Esimerkki 2.2.2 Normaalijakaumaoletus

```
> x<-c(5.1,3.8,5.0,6.4,5.0,4.0,3.2,4.4,4.1,5.6,4.8,4.5)
> y<-c(5.4,5.6,5.4,6.5,6.8,6.0,4.2,4.4,4.8,4.9,6.9,4.9,5.2)
> m<-length(x)
> n<-length(y)
> mx<-mean(x)
> my<-mean(y)
> sx<-sd(x)
> sy<-sd(y)
> s<-sqrt(((m-1)*sx*sx+(n-1)*sy*sy)/(m+n-2))
> N<-m+n
> lambda<-m/N
> t<-sqrt(N*lambda*(1-lambda))*(my-mx)/s
> t
[1] 2.321517
> apu<-pt(t,N-2)
> 2*min(apu,1-apu)
[1] 0.02947281
> ala<-(my-mx)-qt(0.975,N-2)*s/sqrt(N*lambda*(1-lambda))
> ala
[1] 0.0874851
> yla<-(my-mx)+qt(0.975,N-2)*s/sqrt(N*lambda*(1-lambda))
> yla
[1] 1.518925
> my-mx
[1] 0.8032051
> t.test(y,x)

 Welch Two Sample t-test

data:  y and x
t = 2.3232, df = 22.899, p-value = 0.02941
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.08782429 1.51858597
sample estimates:
mean of x mean of y
 5.461538  4.658333
```

2.2.3 Parametriton malli A

Oletukset:

- $x_1, \ldots, x_m$ satunnaisotos jakaumasta, jonka kf on $F(x)$

- $y_1, \ldots, y_n$ satunnaisotos jakaumasta, jonka kf on $F(y - \Delta)$
- otokset riippumattomia

Siis otokset ovat peräisin jakaumista, jotka poikkeavat ainoastaan sijainnin suhteen; parametri Δ kuvaa siis käsittelyjen eroa, s.o., jakaumien sijainnin poikkeavuutta.
Nyt

- AINEISTO:

$$x_1, \ldots, x_m \quad \rightarrow \quad m_1 = Med\{x_1, \ldots, x_m\}$$
$$y_1, \ldots, y_n \quad \rightarrow \quad m_2 = Med\{y_1, \ldots, y_n\}$$

  ja edelleen

$$m_0 = Med\{x_1, \ldots, x_m, y_1, \ldots, y_n\}$$

  Olkoon edelleen k niiden havaintojen lukumäärä, jotka ylittävät tämän yhdistetyn aineiston mediaanin. Silloin siis joko $N = 2k$ tai $N = 2k+1$.

- TESTAUS (kahden otoksen merkkitesti; Moodin testi):

$$S = \#\{y_j > m_0\} \ \sim \ HypGeo(k; n, m), \quad \text{kun } H_0 : \Delta = 0 \text{ tosi.}$$

  S on siis niiden toisen otoksen tapausten lukumäärä, jotka ylittävät yhdistetyn aineiston mediaanin. Tällöin nollahypoteesin mukainen jakauma

$$P_0(S = s) = \frac{\binom{n}{s}\binom{m}{k-s}}{\binom{N}{k}}, \quad max(0, k - m) \leq s \leq min(k, n)$$

  riippuu vain otoskoista n ja m. Edelleen

$$E_0(S) = k(1 - \lambda) \ \text{ ja } \ \mathrm{Var}_0(S) = \frac{(N - k)k}{N - 1}\lambda(1 - \lambda) \approx \frac{1}{4}\frac{mn}{N}$$

  ja isoilla otoskoilla voidaan jälleen käyttää normaaliaproksimaatiota.

- PISTE-ESTIMOINTI:

$$\Delta : \ m_2 - m_1$$

- LUOTTAMUSVÄLI: Luottamusväli konstruoidaan jälleen 'kääntämällä' testi, s.o., etsitään ne Δ:n arvot, joita kaksisuuntainen tason α testi ei hylkäisi. Jos

$$\sum_{s=a}^{b} P_0(S = s) = 1 - \alpha$$

  niin satunnaisvälin

$$(y_{(n-b)} - x_{(m-k+b+1)} \ , \ y_{(n-a+1)} - x_{(m-k+a)})$$

  peitetodennäköisyys on $1 - \alpha$.

Tapaus $n = m$
Tarkastellaan huolellisemmin tapausta, jolloin otoskoot ovat samat. Kun $n = m$ ja n on suuri, niin

$$z = \frac{S - n/2}{\sqrt{n/8}} \sim N(0,1)$$

Luottamusvälin konstruoimiseksi riittää tarkastella Δ:n arvoja

$$y_{(1)} - x_{(n)} \leq y_{(2)} - x_{(n-1)} \leq \ldots \leq y_{(n-1)} - x_{(2)} \leq y_{(n)} - x_{(1)}.$$

Luottamusvälit ovat nyt muotoa

$$(y_{(i)} - x_{(n-i+1)}, y_{(n-i+1)} - x_{(i)})$$

peitetodennäköisyydellä

$$1 - 2P_0[S \leq i-1] \approx 1 - 2\Phi\left(\frac{i - n/2 - 1/2}{\sqrt{n/8}}\right).$$

Siis likimääräisen $100(1-\alpha)$ %:n luottamusvälin antaa valinta $i \approx n/2 - z_{\alpha/2}\sqrt{n/8}$.

Esimerkki 2.2.3 Parametriton malli A

```
s2.test<-function(x,y)
{
#
#Kahden otoksen merkkitesti; Moodin testi
#
  n<-length(y)
  m<-length(x)
  N<-m+n
  k<-round(N/2)
  m1<-median(x)
  m2<-median(y)
  m0<-median(c(x,y))
  S<-sum(y>m0)
  p<-2*min(phyper(S,n,m,k),1-phyper(S-1,n,m,k))
  a<-qhyper(0.025,n,m,k)
  b<-qhyper(0.975,n,m,k)
  x0<-sort(x)
  y0<-sort(y)
  estala<-y0[n-b]-x0[m-k+b+1]
```

```
  estyla<-y0[n-a+1]-x0[m-k+a]
  CL<-phyper(b,n,m,k)-phyper(a-1,n,m,k)
  list(est=m2-m1,test=S,p=p,estala=estala,estyla=estyla,CL=CL)
}

> x<-c(5.1,3.8,5.0,6.4,5.0,4.0,3.2,4.4,4.1,5.6,4.8,4.5)
> y<-c(5.4,5.6,5.4,6.5,6.8,6.0,4.2,4.4,4.8,4.9,6.9,4.9,5.2)
> median(x)   [1] 4.65
> median(y)   [1] 5.4
> median(c(x,y))  [1] 5
> s2.test(x,y)
$est
[1] 0.75
$test
[1] 8
$p
[1] 0.3131319
$estala
[1] -0.2
$estyla
[1] 1.9
$CL
[1] 0.9830683
```

2.2.4 Parametriton malli B

Oletukset jälleen:

- $x_1, \ldots, x_m$ satunnaisotos jakaumasta, jonka kf on $F(x)$
- $y_1, \ldots, y_n$ satunnaisotos jakaumasta, jonka kf on $F(y - \Delta)$
- otokset riippumattomia

Nyt

- AINEISTO:

$$x_1, \ldots, x_m, \;\; y_1, \ldots, y_n \quad \rightarrow \quad d_1, \ldots, d_M,$$

missä $d_1, \ldots, d_M$ ovat kaikki mahdolliset pareittaiset erotukset

$$y_j - x_i, \quad i = 1, \ldots, m; \; j = 1, \ldots, n.$$

Pareittaisia erotuksia on siis $M = mn$ kappaletta.

- TESTAUS (kahden otoksen Wilcoxonin testi; Mannin-Whitneyn testi):
$$W = \#\{d_k > 0\}$$
W:n nollahypoteesin mukainen jakauma ei riipu lainkaan jakauman muodosta; $E_0(W) = mn/2$ ja $\mathrm{Var}_0(W) = mn(m+n+1)/12$ ja normaaliaproksimaatio on hyvä jo pienillä otoskoilla.

- PISTE-ESTIMOINTI:
$$\Delta : \ Med\{d_1, \ldots, d_M\}$$

- LUOTTAMUSVÄLI: Luottamusväli konstruoidaan jälleen 'kääntämällä' testi, jolloin välin
$$(d_{(i)} \ , \ d_{(M-i+1)})$$
peitetodennäköisyys on
$$P_i = 1 - 2 \cdot P_0(W \leq i-1) \approx 1 - 2 \cdot \Phi\left(\frac{i - mn/2 - 1/2}{\sqrt{mn(m+n+1)/12}}\right).$$

Esimerkki 2.2.4 Parametriton malli B

```
----
Funktioita:
> i<-(0:6)
> pwilcox(i,2,3)
[1] 0.1 0.2 0.4 0.6 0.8 0.9 1.0
> pwilcox(i,m,n)
> qwilcox(0.025,n,m)
[1] 0
> qwilcox(0.13,n,m)
[1] 1


> n<-5
> m<-6
> i<-(0:30)
> pwilcox(i,m,n)
 [1] 0.002164502 0.004329004 0.008658009 0.015151515 0.025974026
0.041125541
 [7] 0.062770563 0.088744589 0.123376623 0.164502165 0.214285714
0.268398268
[13] 0.331168831 0.396103896 0.465367965 0.534632035 0.603896104
0.668831169
```

```
[19] 0.731601732 0.785714286 0.835497835 0.876623377 0.911255411
0.937229437
[25] 0.958874459 0.974025974 0.984848485 0.991341991 0.995670996
0.997835498
[31] 1.000000000
> qwilcox(pwilcox(i,m,n),m,n)
 [1]  0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19
20 21 22 23 24
[26] 25 26 27 28 29 30
---------------


w2.test<-function(x,y)
{
  n<-length(y)
  m<-length(x)
  M<-m*n
  d<-c(matrix(rep(y,m),n)-t(matrix(rep(x,n),m)))
  W<-sum(d>0)
  p<-2*min(pwilcox(W,n,m),1-pwilcox(W-1,n,m))
  a<-qwilcox(0.025,n,m)
  d0<-sort(d)
  CL<-1-2*pwilcox(a-1,n,m)
  list(est=median(d),test=W,p=p,estala=d0[a],estyla=d0[M+1-a],CL=CL)
}


> x
 [1] 5.1 3.8 5.0 6.4 5.0 4.0 3.2 4.4 4.1 5.6 4.8 4.5
> y
 [1] 5.4 5.6 5.4 6.5 6.8 6.0 4.2 4.4 4.8 4.9 6.9 4.9 5.2
> w2.test(x,y)
$est
[1] 0.8
$test
[1] 115
$p
[1] 0.04571121
$estala
[1] 0.1
$estyla
[1] 1.6
$CL
[1] 0.9542888
```

```
> xx<-rlaplace(100)
> yy<-rlaplace(100)+0.1
> t.test(yy,xx)
 Welch Two Sample t-test
data:  yy and xx
t = 0.0718, df = 197.542, p-value = 0.9428
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4096544  0.4406161
sample estimates:
  mean of x   mean of y
-0.01147489 -0.02695573

> w2.test(xx,yy)
$est           [1] 0.1726839
$test          [1] 5354
$p             [1] 0.3885385
$estala        [1] -0.2260240
$estyla        [1] 0.5487337
$CL            [1] 0.9502466
> s2.test(xx,yy)
$est           [1] 0.3180473
$test          [1] 56
$p             [1] 0.1195807
$estala        [1] -0.04384384
$estyla        [1] 0.6357128
$CL            [1] 0.966364

Raskaudenaikaisen tupakoinnin vaikutus syntymapainoon:
> data<-read.table("lapset.dat")
> paino<-data[,16]
> mean(paino)              [1] 348.6307
> median(paino)            [1] 354
> HL(paino)                [1] 352
> aidintup<-data[,5]
> paino0<-paino[aidintup==0]
> paino1<-paino[aidintup>0]
> length(paino0)           [1] 1593
> length(paino1)           [1] 365
> x<-sample(paino0,100,replace=T)
> y<-sample(paino1,100,replace=T)
> t.test(y,x)              Welch Two Sample t-test
data:  y and x
t = -3.214, df = 187.117, p-value = 0.001542
```

```
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -44.04028 -10.53972
sample estimates:
mean of x mean of y
   333.62    360.91
> s2.test(x,y)
$est           [1] -16
$test          [1] 43
$p             [1] 0.06572644
$estala        [1] -35
$estyla        [1] 3
$CL            [1] 0.966364
> w2.test(x,y)
$est           [1] -23
$test          [1] 3789
$p             [1] 0.002970013
$estala        [1] -38
$estyla        [1] -8
$CL            [1] 0.9502466
```

TULOS:

Oletukset jälleen:

- $x_1, \ldots, x_m$ satunnaisotos jakaumasta, jonka kf on $F(x)$
- $y_1, \ldots, y_n$ satunnaisotos jakaumasta, jonka kf on $F(y - \Delta)$
- otokset riippumattomia

Olkoot

$$\hat{\Delta}_I, \quad \hat{\Delta}_{II} \quad \text{ja} \quad \hat{\Delta}_{III}$$

otoksista lasketut keskiarvojen erotus, mediaanien erotus ja kahden otoksen HL-estimaatti. Silloin (yleisten oletusten vallitessa)

$$\hat{\Delta}_I \sim AN\left(\Delta, \frac{1}{N}\frac{1}{\lambda(1-\lambda)}\sigma^2\right),$$

$$\hat{\Delta}_{II} \sim AN\left(\Delta, \frac{1}{N}\frac{1}{\lambda(1-\lambda)}\frac{1}{4f^2(\mu)}\right)$$

ja

$$\hat{\Delta}_{III} \sim AN\left(\Delta, \frac{1}{N}\frac{1}{\lambda(1-\lambda)}\frac{1}{12[\int f^2(x)dx]^2}\right),$$

missä $N = m+n$, $\lambda = m/N$, μ ja σ^2 ovat jakauman f mediaani ja varianssi.

Pienillä m:n ja n:n arvoilla tehokkuuksia voi jälleen vertailla simuloimalla.
Estimaatin $\hat{\Delta}$ varianssin estimoiminen BOOTSTRAP-tekniikalla

Oletetaan, että $X = \{x_1, \ldots, x_m\}$ ja $Y = \{y_1, \ldots, y_n\}$ ovat riippumattomia satunnaisotoksia jakaumasta $F(x)$ ja $F(y - \Delta)$ ja että $\hat{\Delta} = \hat{\Delta}(X, Y)$ on parametrin Δ estimaatti. Miten voidaan estimoida harhaa (bias) ja varianssia,

$$B_F(\hat{\Delta}) = E_F(\hat{\Delta}) - \Delta \text{ ja } \operatorname{Var}_F(\hat{\Delta}),$$

kun F on tuntematon.

Huomaa ensin, että tuntemattoman kertymäfunktion F luonnollinen estimaatti (parametrittomassa mallissa) on otoskertymäfunktio

$$F_N(z) = \frac{1}{N}\left[\#\{x_i \le z\} + \#\{y_j - \hat{\Delta} \le z\}\right],$$

missä $N = m + n$.

Estimaatin $\hat{\Delta}$ harhaa ja varianssia otoskoilla m ja n voidaan tutkia ns. **BOOTSTRAP-tekniikalla**:

- Generoi M riippumatonta satunnaisotosta otoskoolla N jakaumasta F_N (M N:n suuruista satunnaisotosta palauttaen),

$$Z_i^* = \{z_{i1}^*, \ldots, z_{iN}^*\}, \quad i = 1, \ldots, M.$$

- Merkitse

$$X_i^* = \{z_{i1}^*, \ldots, z_{im}^*\} \text{ ja } Y_i^* = \{\hat{\Delta}+z_{i,m+1}^*, \ldots, \hat{\Delta}+z_{i,m+n}^*\}, \ i = 1, \ldots, M$$

- Laske otoksiin liittyvät estimaattien arvot

$$\hat{\Delta}(X_1^*, Y_1^*), \ \ldots \ , \hat{\Delta}(X_M^*, Y_M^*).$$

- Estimaatin varianssin estimaatti:

$$\frac{1}{M}\sum_{i=1}^{M}[\hat{\Delta}(X_i^*, Y_i^*)]^2 - \left[\frac{1}{M}\sum_{i=1}^{M}\hat{\Delta}(X_i^*, Y_i^*)\right]^2$$

- Harhan estimaatti

$$\frac{1}{M}\sum_{i=1}^{M}\hat{\Delta}(X_i^*, Y_i^*) - \hat{\Delta}(X, Y)$$

2.2.5 Kahden otoksen järjestyslukutestit

- $X = \{x_1, \ldots, x_m\}$ ja $Y = \{y_1, \ldots, y_n\}$ ovat riippumattomia satunnaisotoksia jakaumista $F(x)$ ja $F(y - \Delta)$. Nollahypoteesin $H_0 : \Delta = 0$ vallitessa

$$Z = \{z_1, \ldots, z_N\} = \{x_1, \ldots, x_m, y_1, \ldots, y_n\}$$

  on satunnaisotos jakaumasta $F(z)$.

- Järjestysluvut

$$R_i = R(z_i, Z), \quad i = 1, \ldots N$$

- Wilcoxonin kahden otoksen testisuure

$$W_0 = \sum_{i=m+1}^{N} R_i$$

- Aikaisempi (Mannin-Whitneyn) versio

$$W = \sum_{i=1}^{m} \sum_{j=i}^{n} I(y_j - x_i > 0) = W_0 - \frac{n(n+1)}{2}$$

- Kahden otoksen merkkitesti

$$S = \sum_{j=m+1}^{N} I(y_i > z_{(N/2)}) = \sum_{j=m+1}^{N} I(R_i > N/2)$$

- Nollahypoteesin vallitessa

$$(R_1, \ldots, R_N)$$

  on tasaisesti jakautunut; mahdollisia arvoja ovat kaikki $(1, \ldots, N)$:n permutaatiot todennäköisyyksillä $1/N!$

- Nollahypoteesin vallitessa yksi- ja kaksiulotteiset reunajakaumat ovat

$$P_0(R_1 = i) = \frac{1}{N}, \quad i = 1, \ldots, N$$

  ja

$$P_0(R_1 = i, R_2 = j) = \frac{1}{N(N-1)}, \quad i, j = 1, \ldots, N; \quad i \neq j.$$

  Silloin

$$E_0(R_1) = \frac{N+1}{2}, \qquad \mathrm{Var}_0(R_1) = \frac{N^2 - 1}{12},$$

  ja

$$\mathrm{Cov}_0(R_1, R_2) = -\frac{N+1}{12}.$$

- Lopulta

$$E_0\left[\sum_{i=m+1}^{N} R_i\right] = \frac{n(N+1)}{2} \quad \text{ja} \quad \mathrm{Var}_0\left[\sum_{i=m+1}^{N} R_i\right] = \frac{nm(N+1)}{12}$$

2.2.6 Yleinen järjestyslukutesti

- Valitaan pisteluvut

$$a(1) \leq a(2) \leq \ldots \leq a(N).$$

  Usein $a(i) = \psi(i/(N+1))$, missä $\psi(u)$, $u \in (0,1)$, on niin sanottu pistelukufunktio (*score function*)

- Valittuja pistelukuja vastaava järjestyslukutestisuure on

$$T = \sum_{i=m+1}^{N} a(R_i).$$

- Jälleen T:n jakauma ei nollahypoteesin vallitessa riipu kertymäfunktiosta F. Erityisesti

$$E_0(T) = n\bar{a} \quad \text{ja} \quad \mathrm{Var}_0(T) = \frac{mn}{N} s_a^2.$$

  Normaaliaproksimaatiota tai simulointia voi käyttää likimääräisen p-arvon laskemiseen.

- Erikoistapauksina saadaan kahden otoksen merkkitesti ($a(i) = 1$, kun $i > N/2$ ja 0 muulloin) ja Wilcoxonin-Mannin-Whitneyn testi ($a(i) = i$, $i = 1, \ldots, N$). Niin sanottu van der Waerdenin testi, pistelukufunktiona $\Phi^{-1}(u)$ ($T = \sum_{i=m+1}^{N} \Phi^{-1}(\frac{R_i}{N+1})$) on optimaalinen normaalijakauman tapauksessa.

- Pistelukuja $a(1) \leq \ldots \leq a(N)$ vastaava estimaatti ja luottamusväli on mahdollista konstruoida kuten Wilcoxonin testin tapauksessa.

```
JARJESTYSLUKUTESTEJA:

p.waerden2<-function(x,y)
{
#
# van der Waerdenin kahden otoksen jarjestyslukutesti
#
    m<-length(x)
```

```
    n<-length(y)
    z<-c(x,y)
    N<-m+n
    a<-qnorm(rank(z)/(N+1)) #pisteluvut
    T0<-sum(a[(m+1):N])     #testisuureen arvo

    #lasketaan simuloimalla likim. p-arvo
    T<-NULL
    for (i in 1:1000)
       T[i]<-sum(sample(a,n,replace=F))
    apu<-mean(T<T0)
    2*min(apu,1-apu)
}

p.s2<-function(x,y)
{
#
# Moodin kahden otoksen merkkitesti
#
    m<-length(x)
    n<-length(y)
    z<-c(x,y)
    N<-m+n
    k<-round(N/2)
    a<-(rank(z)>N/2)    #pisteluvut
    S<-sum(a[(m+1):N])  #testisuureen arvo

    p<-2*min(phyper(S,n,m,k),1-phyper(S-1,n,m,k),0.5)
    p
}

p.w2<-function(x,y)
{
#
# Mannin-Whitneyn-Wilcoxonin kahden otoksen jarjestyslukutesti
#
    m<-length(x)
    n<-length(y)
    z<-c(x,y)
    N<-m+n
    a<-rank(z)            #pisteluvut
    W0<-sum(a[(m+1):N])   #Wilcoxonin testisuure
    W<-W0-n*(n+1)/2       #Mann-Whitneyn testisuure
```

```
    p<-2*min(pwilcox(W,n,m),1-pwilcox(W-1,n,m),0.5)
    p
}
```

2.2.7 Kertausta

Kirjoitetaan ylläoleva teoria "formaalimmin". Merkitään yhdistettyä otosta

$$\boldsymbol{y} = (y_1, \ldots, y_m, y_{m+1}, \ldots, y_{m+n})^T$$

ja

$$X = \begin{pmatrix} 1_m & 0_m \\ 1_n & 1_n \end{pmatrix} \text{ ja } \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \Delta \end{pmatrix}$$

Silloin

$$\boldsymbol{e} = \boldsymbol{y} - X\boldsymbol{\beta}$$

on satunnaisotos jakaumasta, jonka kf on F.

- L_2-normi: $||\boldsymbol{e}||_2 = \sqrt{\sum e_i^2}$
- L_1-normi: $||\boldsymbol{e}||_1 = \sum |e_i|$
- Painotettu L_1-normi: $||\boldsymbol{e}||_3 = \sum [R_i e_i]$
- Yleistetty painotettu L_1-normi: $||\boldsymbol{d}||_\psi = \sum [\psi(R_i/(N+1))e_i]$, missä ψ on valittu pistelukufunktio
- Parametrin β estimaatti minimoi normin $||\boldsymbol{y} - X\boldsymbol{\beta}||_2$, normin $||\boldsymbol{y} - X\boldsymbol{\beta}||_1$, jne.
- Testit hypoteesille $H_0 : \ \Delta = 0$ saadaan 'objektifunktion' derivaatan avulla, järjestyslukutestien tapauksessa

$$T = \sum_{i=m+1}^{m+n} \psi(R_i/(N+1))$$

- Luottamusvälit Δ:lle saadaan 'kääntämällä' testi. Ensin etsitään arvo sellainen arvo c, että

$$P_0(-c \leq T \leq c) = 1 - \alpha.$$

  Tason $1 - \alpha$ luottamusväli koostuu niistä Δ:n arvoista, joita testattaessa testisuure $T(\Delta) \in [-c, c]$.

Luku 3

Usean käsittelyn vertailu

3.1 Kaltaistetut otokset

3.1.1 Koeasetelma

Lohkokoe:

- Verrattaessa k käsittelyä, koeyksiköistä muodostetaan n lohkoa siten, että kunkin lohkon jäsenet ovat relevanttien muuttujien osalta mahdollisimman samankaltaisia (joskus sama yksilö).
- Kunkin lohkon sisällä sovelletaan jokaista käsittelyä; käsittelyt arvotaan koeyksilöille.

Kyseessä on siis kaltaistettujen parien suunnitelman yleistys useamman kuin kahden menetelmän samanaikaiseen vertaamiseen. Samoilla havaintomäärillä kaltaistetujen otosten suunnitelma saattaa tuottaa huomattavasti tehokkaamman kokeen kuin riippumattomien otosten suunnitelma.

Taulukko 3.1: Vastemuuttujan arvot.

Lohko	Käsittely			
	1	2	...	k
1	y_{11}	y_{12}	$\cdots$	y_{1k}
2	y_{21}	y_{22}	$\cdots$	y_{2k}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	y_{n1}	y_{n2}	$\cdots$	y_{nk}

Tilastollinen malli:

$$y_{ij} = \mu + \tau_i + \Delta_j + \epsilon_{ij}, \quad i = 1, \ldots, n;\ j = 1, \ldots, k$$

missä $\tau_1 = \Delta_1 = 0$ (esimerkiksi) ja satunnaismuuttujat ϵ_{ij} muodostavat satunnaisotoksen jakaumasta, jonka kertymäfunktio on F.

3.1.2 Normaalijakaumaoletus

$$y_{ij} = \mu + \tau_i + \Delta_j + \epsilon_{ij}, \quad i = 1, \ldots, n;\ j = 1, \ldots, k$$

missä satunnaismuuttujat ϵ_{ij} muodostavat satunnaisotoksen $N(0, \sigma^2)$-jakaumasta.
Kiinnostava nollahypoteesi $H_0 : \ \Delta_2 = \ldots = \Delta_k = 0$ (oletuksen mukaan $\Delta_1 = 0$).

Kaksisuuntainen varianssianalyysi voidaan toteuttaa seuraavasti:

- Keskistä havainnot riveittäin (lohkoittain). Merkitään

$$z_{ij} = y_{ij} - \bar{y}_{i\cdot}, \quad i = 1, \ldots, n;\ j = 1, \ldots k$$

- Laske (nollahypoteesioletuksen mukainen) varianssiestimaatti

$$s^2 = \frac{1}{n(k-1)} \sum_{i=1}^{n} \sum_{j=1}^{k} z_{ij}^2$$

- Laske z-arvojen summat käsittelyittäin

$$z_{\cdot j} = \sum_{i=1}^{n} z_{ij}, \quad j = 1, \ldots k$$

- ANOVA-testisuure on

$$Q = \frac{1}{n} \sum_{j=1}^{k} \frac{z_{\cdot j}^2}{s^2}$$

  on likimain χ^2-jakautunut $k-1$ vapausasteella, kun H_0 tosi.

HUOM: Yleensä testisuureena käytetään asymptootisti ekvivalenttia suuretta (eri varianssiestimaatti)

$$F = \frac{\sum_{i=1}^{n} \sum_{j=1}^{k} (\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2 / (k-1)}{\sum_{i=1}^{n} \sum_{j=1}^{k} (y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot})^2 / [(n-1)(k-1)]}$$

joka noudattaa $F(k-1, (n-1)(k-1))$ jakaumaa, kun H_0 tosi.

3.1.3 Parametriton malli

$$y_{ij} = \mu + \tau_i + \Delta_j + \epsilon_{ij}, \quad i = 1, \ldots, n;\ j = 1, \ldots, k$$

missä satunnaismuuttujat ϵ_{ij} muodostavat satunnaisotoksen tuntemattomasta jakaumasta F.
Kiinnostava nollahypoteesi jälleen $H_0 : \ \Delta_2 = \ldots = \Delta_k = 0$ (oletuksen mukaan $\Delta_1 = 0$).
Konstruoidaan parametriton testisuure jälleen yleisillä pisteluvuilla

$$a(1) \leq \ldots \leq a(k).$$

Oletetaan, että pisteluvut on keskistetty, s.o., $\sum_{i=1}^{k} a(i) = 0$.
Toimitaan seuraavasti:

- Korvataan havainnot ensin **riveittäisillä** järjestysluvuillaan ($y_{ij} \to R_{ij}$) ja sitten järjestysluvut vastaavilla pisteluvuilla ($R_{ij} \to a(R_{ij})$); saadaan taulukko

Taulukko 3.2: Pisteluvut

Lohko	Käsittely			
	1	2	$\cdots$	k
1	a_{11}	a_{12}	$\cdots$	a_{1k}
2	a_{21}	a_{22}	$\cdots$	a_{2k}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	a_{n1}	a_{n2}	$\cdots$	a_{nk}

- Lasketaan (nollahypoteesioletuksen mukainen) pistelukujen varianssi (aina sama riippumatta estimaatista)

$$s^2 = \frac{1}{n(k-1)} \sum_i \sum_j a_{ij}^2 = s_a^2$$

- Laske pistelukujen summat käsittelyittäin

$$a_{\cdot j} = \sum_{i=1}^{n} a_{ij}, \quad j = 1, \ldots k$$

- Jakaumasta riippumaton testisuure on

$$Q = \frac{1}{n} \sum_{j=1}^{k} \frac{a_{\cdot j}^2}{s_a^2}$$

on likimain χ^2-jakautunut $k-1$ vapausasteella, kun H_0 tosi.

HUOM: Kun $a(i) = i - (k+1)/2$, $i = 1, \ldots, k$, saadaan niin sanottu **Friedmanin testi** (1937), testisuureena

$$Q = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_{\cdot j}^2 - 3n(k+1),$$

joka on siis Wilcoxonin testin yleistys kaksisuuntaisen varianssianalyysin tapaukseen. Luonnollisesti myös Moodin testi voidaan yleistää tähän tapaukseen. Friedmanin testin tehokkuus normaalijakauman tapauksessa ei ole yhtä hyvä kuin Wilcoxonin testin (0.64, 0.72, 0.76, 0.80 ja 0.87, kun $k = 2,\ 3,\ 4,\ 5$ ja 10). Muita (tehokkaampia, mutta ei aidosti jakaumasta riippumattomia testejä) saadaan, kun ensin vähennetään jokaisesta havainnosta riveittäinen keskiarvo, sen jälkeen muodostetaan järjestysluvut koko aineistosta, ja edetään lopuksi kuten edellä.

Mahdollista trendiä paljastamaan voidaan konstruoida niin sanottu **Pagen testi** (1963), testisuureena

$$L = \sum_{j=1}^{k} jR_{\cdot j},$$

jonka nollahypoteesin mukaiset odotusarvo ja varianssi ovat

$$E_0(L) = \frac{nk(k+1)^2}{4} \quad \text{ja} \quad \mathrm{Var}_0(L) = \frac{n(k^3-k)^2}{144(k-1)}.$$

Normaaliaproksimaatiota voi jälleen käyttää p-arvon löytämiseen.

```
p.Friedman<-function(x,g,b)
{
  k<-max(g)
  n<-max(b)
  NN<-length(x)
  X<-matrix(rep(0,n*k),nrow=n)
  for (i in 1:NN) X[b[i],g[i]]<-x[i]
  A<-t(apply(X,1,rank))
  ma<-apply(A,1,mean)
  A<-A-matrix(rep(ma,k),nrow=n)
  ss<-sum(A*A)/(n*(k-1))
  T<-apply(A,2,sum)
  Q<-sum(T*T)/(n*ss)
  list(Q=Q,p=1-pchisq(Q,k-1))
}
```

```
p.2anova<-function(x,g,b)
{
  k<-max(g)
  n<-max(b)
  NN<-length(x)
  X<-matrix(rep(0,n*k),nrow=n)
  for (i in 1:NN) X[b[i],g[i]]<-x[i]
  A<-X
  ma<-apply(A,1,mean)
  A<-A-matrix(rep(ma,k),nrow=n)
  ss<-sum(A*A)/(n*(k-1))
  T<-apply(A,2,sum)
  Q<-sum(T*T)/(n*ss)
  list(Q=Q,p=1-pchisq(Q,k-1))
}
```

Kuvailu tutkimuksesta:
'In a study of hypnosis, the emotions of fear, happiness, depression and calmness were requested (in random order) from each of eight subjects during hypnosis. The following measurements are observed values of skin potential (adjusted for initial level) in millivolts (Damaser,Shore and Orne, 1963)'

```
> x<-c( 23.1,57.6,10.5,23.6,11.9,54.6,21.0,20.3,
+ 22.7,53.2,9.7,19.6,13.8,47.1,13.6,23.6,
+ 22.5,53.7,10.8,21.1,13.7,39.2,13.7,16.3,
+ 22.6,53.1,8.3,21.6,13.3,37.0,14.8,14.8)
> length(x)
[1] 32
> subject<-c(rep(1:8,4))
> subject
 [1] 1 2 3 4 5 6 7 8 1 2 3 4 5 6 7 8 1 2 3 4 5 6 7 8 1 2 3 4 5 6 7 8

> g<-c(rep(1,8),rep(2,8),rep(3,8),rep(4,8))
> g
 [1] 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4
> X<-t(matrix(x,nrow=8))
> X
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
[1,] 23.1 57.6 10.5 23.6 11.9 54.6 21.0 20.3
[2,] 22.7 53.2  9.7 19.6 13.8 47.1 13.6 23.6
[3,] 22.5 53.7 10.8 21.1 13.7 39.2 13.7 16.3
[4,] 22.6 53.1  8.3 21.6 13.3 37.0 14.8 14.8
> R<-(apply(X,2,rank))
```

```
> R
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
[1,]    4    4    3    4    1    4    4    3
[2,]    3    2    2    1    4    3    1    4
[3,]    1    3    4    2    3    2    2    2
[4,]    2    1    1    3    2    1    3    1
>
> apply(R,1,sum)
[1] 27 20 19 14
> p.2anova(x,g,subject)
$Q
[1] 7.949385

$p
[1] 0.04706942

> p.Friedman(x,g,subject)
$Q
[1] 6.45

$p
[1] 0.09165537
```

3.2 Riippumattomat otokset

3.2.1 Koeasetelma

- Yksilöt arvottu $k \geq 2$ käsittelylle; käsittely i kohdistuu n_i yksilöön; yksilöiden kokonaismäärä $N = n_1 + \ldots + n_k$; kaikki

$$\frac{N!}{n_1! \cdot \ldots \cdot n_k!}$$

  'jakoa' yhtä todennäköisiä.

- Parametrinen malli: Satunnaisotokset riippumattomia ja peräisin parametrisesta jakaumasta (esim. normaalijakaumasta), jotka poikkeavat toisistaan vain sijainnin suhteen.

- Parametriton malli: Satunnaisotokset riippumattomia ja peräisin jakaumista, jotka poikkeavat vain sijainnin suhteen.

Mittaustuloksiin liittyviä satunnaisuuden (vaihtelun) lähteitä ovat

1. yksilöiden välinen vaihtelu populaatiossa,
2. käsittelyjen erosta johtuva vaihtelu ja
3. arvonnasta (satunnaistamisesta) johtuva vaihtelu.

Permutaatiotestit nojaavat vain viimeksimainittuun vaihteluun.

Merkitään

$$N = n_1 + \ldots + n_k \quad \text{ja} \quad \lambda_i = \frac{n_i}{N}.$$

Osuuksien λ_i valinnoilla on vaikutusta estimaattien ja testien tehokkuuteen. Merkitään edelleen

$$N_i = n_1 + \ldots + n_i, \quad i = 1, \ldots, k.$$

3.2.2 Normaalijakaumaoletus

Oletukset:

- $y_1, \ldots, y_{N_1}$ satunnaisotos jakaumasta $N(\mu, \sigma^2)$
- $y_{N_1+1}, \ldots, y_{N_2}$ satunnaisotos jakaumasta $N(\mu + \Delta_2, \sigma^2)$
- $\cdots$
- $y_{N_{k-1}+1}, \ldots, y_{N_k}$ satunnaisotos jakaumasta $N(\mu + \Delta_k, \sigma^2)$
- otokset riippumattomia

Käsittelyjen eroja kuvaava $(k-1)$-ulotteinen parametri on $\Delta = (\Delta_2, \ldots, \Delta_k)$.

AINEISTO:

$$y_1, \ldots, y_{N_1} \quad \to \quad \bar{y}_1, \ s_1^2$$

$$\vdots$$

$$y_{N_{k-1}+1}, \ldots y_{N_k} \quad \to \quad \bar{y}_k, \ s_k^2$$

ja lopulta

$$s_1^2, \ldots, s_k^2 \quad \to \quad s^2 = \frac{1}{N-k}[(n_1 - 1)s_1^2 + \ldots + (n_k - 1)s_k^2]$$

TESTAUS: Testisuure nollahypoteesille

$$H_0 : \quad \Delta_2 = \ldots = \Delta_k = 0$$

konstruoidaan seuraavasti:

- Muodostetaan kahden otoksen t-testisuure

$$t_i = \sqrt{N\lambda_i(1-\lambda_i)}\frac{\bar{y}_i - \bar{y}_{-i}}{s}$$

ongelmaan "i:s otos vs. muut otokset"($\bar{y}_{-i}$ on keskiarvo niistä $N - n_i$ havainnosta, jotka **eivät** kuulu otokseen i).

- Konstruoidaan yhdistetty testisuure

$$Q = \sum_{i=1}^{k}[(1-\lambda_i)t_i^2]$$

Nollahypoteesin vallitessa $Q/(k-1) \sim F(k-1, N-k)$ (tai $Q \sim \chi^2(k-1)$ likimain)

ESTIMOINTI: Testiä vastaava piste-estimaatit, $\hat{\Delta}_2, \ldots, \hat{\Delta}_k$, ja estimaattien luottamusvälit saadaan kahden otoksen testeistä. Esim. $\hat{\Delta}_2$ on toisen otoksen ja ensimmäisen otoksen keskiarvojen erotus. Likimain pätee:

$$\hat{\Delta}_i \sim AN\left(\Delta_i, \frac{1}{N}\frac{\lambda_1+\lambda_i}{\lambda_1\lambda_i}\sigma^2\right), \quad i = 2, \ldots, k.$$

Samanaikainen luottamusalue: Etsitään ne arvot $\Delta = (\Delta_2, \ldots, \Delta_k)$, joilla

$$F_{1-\alpha/2}(k-1, N-k) \;\leq\; \frac{Q(\Delta)}{k-1} \;\leq\; F_{\alpha/2}(k-1, N-k),$$

missä $Q(\Delta_0)$ testisuure testattaessa nollahypoteesia $H_0 : \Delta = \Delta_0$. Luottamusalue on $(k-1)$-ulotteinen ellipsoidi.

3.2.3 Parametriton malli

Oletukset:

- $y_1, \ldots, y_{N_1}$ satunnaisotos jakaumasta $F(y)$
- $y_{N_1+1}, \ldots, y_{N_2}$ satunnaisotos jakaumasta $F(y - \Delta_2)$
- ...
- $y_{N_{k-1}+1}, \ldots, y_{N_k}$ satunnaisotos jakaumasta $F(y - \Delta_k)$
- otokset riippumattomia

Siis jakaumat poikkeavat ainoastaan sijannin suhteen; käsittelyjen eroja kuvaava $(k-1)$-ulotteinen parametri on $(\Delta_2, \ldots, \Delta_k)$.

Silloin **järjestyslukutestisuure** nollahypoteesille $H_0 : \; \Delta_2 = \ldots = \Delta_k = 0$ yleisillä pisteluvuilla

$$a(1) \leq \ldots \leq a(N)$$

konstruoidaan seuraavasti:

- Muodostetaan kahden otoksen 'standardoitu' järjestyslukutestisuure

$$z_i = \frac{\sum_{j=N_{i-1}+1}^{N_i} a(R_j) - n_i\bar{a}}{\sqrt{N\lambda_i(1-\lambda_i)s_a^2}} = \sqrt{N\lambda_i(1-\lambda_i)}\frac{\bar{a}_i - \bar{a}_{-i}}{s_a}$$

  ongelmaan i:s otos vs. muut otokset.

- Konstruoidaan yhdistetty testisuure

$$Q = \sum_{i=1}^{k}[(1-\lambda_i)z_i^2]$$

  Nollahypoteesin vallitessa $Q \sim \chi^2(k-1)$ likimain.

- Erikoistapauksia: **Kruskalin-Wallisin testi**, Wilcoxonin testin laajennus usean riippumattoman otoksen tapaukseen, saadaan valinnalla $a(i) = i$. Pistelukufunktio $a(i) = 1$, kun $i > N/2$, ja nolla muulloin, antaa usean otoksen **Moodin testin**

Estimointi: Testiä vastaavat piste-estimaatit, $\hat{\Delta}_2, \ldots, \hat{\Delta}_k$, saadaan esimerkiksi kahden otoksen testeistä. Esim. Kruskalin-Wallisin testin tapauksessa $\hat{\Delta}_2$ on toisen otoksen ja ensimmäisen otoksen pareittaisten erotusten mediaani, jne.

Joskus vertailtavat käsittelyt ovat sellaisia, että voidaan odottaa (jos H_0 : $\Delta_2 = \ldots = \Delta_k = 0$ ei ole tosi), että

$$\text{joko} \quad 0 < \Delta_2 < \ldots < \Delta_k \quad \text{tai} \quad \Delta_k < \ldots < \Delta_2 < 0$$

(esimerkiksi kasvava annostus). Tämä voidaan ottaa huomioon testiä konstruoidessa.

Jonckheeren-Terpstran testissä konstruoidaan ensin kaikki pareittaiset kahden riippumattoman otoksen Wilcoxonin testisuureet

$$W_{uv}, \quad 1 \le u < v \le k,$$

ongelmille 'u:s otos vs. v:s otos'. Jakaumasta riippumaton testisuure on

$$J = \sum_{u=1}^{k-1}\sum_{v=u+1}^{k} W_{uv}$$

Sen nollahypoteesin mukainen odotusarvo on

$$E_0(J) = \frac{N^2 - \sum_{i=1}^{k} n_i^2}{4}$$

ja nollahypoteesin mukainen varianssi on

$$\mathrm{Var}_0(J) = \frac{N^2(2N+3) - \sum n_j^2(2n_j+3)}{72}.$$

Standardoitu testisuure on nollahypoteesin vallitessa likimain $N(0,1)$-jakautunut.

Funktioita usean riippumattoman otoksen tapaukseen.

```
p.anova<-function(y,g)
{
  k<-max(g)
  NN<-length(y)
  a<-y
  ma<-mean(a)
  ssa<-var(a)
  T<-NULL
  n<-NULL
  for (i in 1:k)
  {
    n[i]<-sum(g==i)
    T[i]<-(sum(a[g==i])-n[i]*ma)**2*NN/(ssa*n[i]*(NN-n[i]))
  }
  w<-1-n/NN
  Chi<-sum(w*T)
  list(Chi=Chi,p=1-pchisq(Chi,k-1))
}

p.Kruskal<-function(y,g)
{
  k<-max(g)
  NN<-length(y)
  a<-rank(y)
  ma<-mean(a)
  ssa<-var(a)
  T<-NULL
  n<-NULL
  for (i in 1:k)
  {
    n[i]<-sum(g==i)
    T[i]<-(sum(a[g==i])-n[i]*ma)**2*NN/(ssa*n[i]*(NN-n[i]))
  }
  w<-1-n/NN
  KW<-sum(w*T)
  list(KW=KW,p=1-pchisq(KW,k-1))
}

p.Mood<-function(x,g)
{
  k<-max(g)
  NN<-length(x)
```

```
  a<-(rank(x)>NN/2)
  ma<-mean(a)
  ssa<-var(a)
  T<-NULL
  n<-NULL
  for (i in 1:k)
  {
    n[i]<-sum(g==i)
    T[i]<-(sum(a[g==i])-n[i]*ma)**2*NN/(ssa*n[i]*(NN-n[i]))
  }
  w<-1-n/NN
  Mood<-sum(w*T)
  list(Mood=Mood,p=1-pchisq(Mood,k-1))
}
```

Esimerkki 3.2.1
Verrataan neljää ruokavaliota; kokeessa 25 rottaa.

```
> x<- c(257,205,206,164,190,214,228,203,
  + 201,231,197,185,
  + 248,265,187,220,212,215,281,
  + 202,276,207,204,230,227)
> length(x)                    [1] 25
> rank(x)
 [1] 22 10 11  1  4 14 18  8
  +  6 20  5  2
  +  21 23  3 16 13 15 25
  +  7 24 12  9 19 17
> g<-c(rep(1,8),rep(2,4),rep(3,7),rep(4,6))
> p.anova(x,g)
$Chi
[1] 3.947586
$p
[1] 0.2671798
> p.Kruskal(x,g)
$KW
[1] 4.212967
$p
[1] 0.2393668
> kruskal.test(x,g)
 Kruskal-Wallis rank sum test
data:  x and g
Kruskal-Wallis chi-squared = 4.213, df = 3, p-value = 0.2394
```

```
> p.Mood(x,g)
$Mood
[1] 4.837912
$p
[1] 0.1840581
```

3.2.4 Kertausta

Merkitään yhdistettyä otosta

$$\boldsymbol{y} = (y_1, \ldots, y_{N_1}, y_{N_1+1}, \ldots, y_{N_2}, \ldots, y_{N_{k-1}+1}, \ldots, y_{N_K})^T$$

ja

$$X = \begin{pmatrix} 1_{n_1} & 0_{n_1} & 0_{n_1} & \ldots & 0_{n_1} \\ 1_{n_2} & 1_{n_2} & 0_{n_2} & \ldots & 0_{n_2} \\ 1_{n_3} & 0_{n_3} & 1_{n_3} & \ldots & 0_{n_3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1_{n_k} & 0_{n_k} & 0_{n_k} & \ldots & 1_{n_k} \end{pmatrix} \quad \text{ja} \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \Delta_2 \\ \Delta_3 \\ \vdots \\ \Delta_k \end{pmatrix}$$

Silloin

$$\boldsymbol{e} = \boldsymbol{y} - X\boldsymbol{\beta}$$

on satunnaisotos jakaumasta, jonka kf on F.

- L_2-normi: $||\boldsymbol{e}||_2 = \sqrt{\sum e_i^2}$
- L_1-normi: $||\boldsymbol{e}||_1 = \sum |e_i|$
- Painotettu L_1-normi: $||\boldsymbol{e}||_3 = \sum [R_i e_i]$
- Yleistetty painotettu L_1-normi: $||\boldsymbol{d}||_\psi = \sum [\psi(R_i/(N+1))e_i]$, missä ψ on valittu pistelukufunktio
- Parametrin $\boldsymbol{\beta}$ estimaatti minimoi normin $||\boldsymbol{y} - X\boldsymbol{\beta}||_2$, normin $||\boldsymbol{y} - X\boldsymbol{\beta}||_1$, jne.
- Testit hypoteesille $H_0: \ \Delta_i = 0$ saadaan 'objektifunktion' derivaatan avulla,

  $$T = \sum_{j=N_{i-1}+1}^{N_i} \psi(R_j/(N+1))$$

  ja niitä kombinoimalla saadaan 'globaali testi' hypoteesille $H_0: \ \Delta_2 = \ldots = \Delta_k = 0$.

Luku 4

Regressioanalyysi

4.1 Yhden selittäjän tapaus

4.1.1 Koeasetelma

Aineisto:

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad \text{ja} \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Malli:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{e}$$

missä $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ on tuntematon, $\boldsymbol{e} = (e_1, e_2, \ldots, e_n)^T$ on satunnaistos origon suhteen symmetrisestä jakaumasta, jonka kertymäfunktio on F. Selittäjämatriisi X kiinteä ja tunnettu.

Merkitään

$$m_1 = \frac{1}{n}\sum x_i, \quad m_2 = \frac{1}{n}\sum x_i^2, \quad \text{and} \quad s_x^2 = m_2 - m_1^2.$$

Silloin

$$D = \frac{1}{n}X^T X = \begin{pmatrix} 1 & m_1 \\ m_1 & m_2 \end{pmatrix} \quad \text{ja} \quad D^{-1} = \frac{1}{s_x^2}\begin{pmatrix} m_2 & -m_1 \\ -m_1 & 1 \end{pmatrix}.$$

4.1.2 Normaalijakaumaoletus

Oletukset: $\boldsymbol{y} = X\boldsymbol{\beta}+\boldsymbol{e}$, missä $\boldsymbol{e} = (e_1, \ldots, e_n)$ on satunnaisotos jakaumasta $N(0, \sigma^2)$.

Estimointi: Minimoidaan

$$||\boldsymbol{e}||_2^2 \;=\; \boldsymbol{e}^T\boldsymbol{e} \;=\; \sum_{i=1}^{n} e_i^2 \;=\; \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

(pienimmän neliösumman menetelmä). Saadaan ehto $X^T(\boldsymbol{y} - X\boldsymbol{\beta}) = \mathbf{0}$ tai

$$\sum_{i=1}^{n} e_i = 0 \;\text{ ja }\; \sum_{i=1}^{n} x_i e_i = 0$$

ja ratkaisu on

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}$$

Silloin

$$\hat{\boldsymbol{\beta}} \;\sim\; N\left(\boldsymbol{\beta}, \frac{1}{n}\sigma^2 D^{-1}\right).$$

HUOM.

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}$$

Edelleen $\hat{\beta}_1$:n keskivirhe (SE) on

$$SE(\hat{\beta}_1) = \frac{1}{\sqrt{n}}\frac{\sigma}{s_x}$$

ja σ^2:n harhaton estimaatti on

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

Testaus: Kun nollahypoteesi $H_0 : \beta_1 = 0$ on tosi,

$$T = \sqrt{n}\frac{\hat{\beta}_1}{s/s_x} = \sqrt{n}\frac{s_{xy}}{s \cdot s_x} \;\sim t(n-2).$$

Luottamusväli:

$$\hat{\beta}_1 \;\pm\; t_{\alpha/2}\frac{s}{\sqrt{n}s_x}$$

4.1.3 Parametriton malli A

Oletukset: $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{e}$, missä $\boldsymbol{e} = (e_1, \ldots, e_n)$ on satunnaisotos jakaumasta, jonka kf on F.

Estimointi: Minimoidaan

$$||\boldsymbol{e}||_1 \;=\; \sum_{i=1}^{n} |e_i| \;=\; \sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_i|$$

(LAD, *Least Absolute Deviation*). Saadaan ehdot

$$\sum_{i=1}^{n} \text{sign}(e_i) = 0 \;\; \text{ja} \;\; \sum_{i=1}^{n} x_i \, \text{sign}(e_i) = 0.$$

Ratkaisua $\hat{\boldsymbol{\beta}}$ ei voi antaa suljetussa muodossa. Nyt pätee

$$\hat{\boldsymbol{\beta}} \;\sim\; AN\left(\boldsymbol{\beta}, \frac{1}{n}\frac{1}{4f^2(0)} D^{-1}\right).$$

ja $\hat{\beta}_1$:n keskivirhe (SE) on

$$SE(\hat{\beta}_1) = \frac{1}{\sqrt{n}}\frac{1}{2f(0)}\frac{1}{s_x}.$$

Testaus: Testisuure nollahypoteesille $H_0 : \beta_1 = 0$ konstruoidaan seuraavasti. Olkoon

$$S_i = \text{sign}(y_i - Med(y_1, \ldots, y_n)), \quad i = 1, \ldots, n.$$

Nollahypoteesin vallitessa jakaumasta riippumaton testisuure on silloin

$$T = \frac{1}{n}\sum_{i=1}^{n} x_i S_i.$$

Silloin $E_0(T) = 0$ ja $\text{Var}_0(T) \approx s_x^2/n$ ja standardoitu suure

$$\sqrt{n}\frac{T}{s_x} = \sqrt{n}\; \widehat{Corr(x, S)}$$

on likimain $N(0,1)$-jakautunut.

Tarkat jakaumasta riippumattomat luottamusvälit parametrille β_1 saadaan jälleen 'kääntämällä' testi.

4.1.4 Parametriton malli B

Oletukset: $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{e}$, missä $\boldsymbol{e} = (e_1, \ldots, e_n)$ on satunaisotos jakaumasta, jonka kf on F.

Estimointi: Olkoot keskistetyt residuaalien järjestysluvut

$$R_i = R(e_i; E) - \frac{n+1}{2}, \quad i = 1, \ldots, n,$$

missä $E = \{e_1, \ldots, e_n\}$. Minimoidaan

$$||\boldsymbol{e}||_3 \;=\; 2 \cdot \sum_{i=1}^{n} R_i e_i \;=\; \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} |e_j - e_i|$$

$$= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} |(y_j - y_i) - \beta_1 (x_j - x_i)|$$

$$= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} |x_j - x_i| \left| \frac{y_j - y_i}{x_j - x_i} - \beta_1 \right|.$$

Merkitään nyt

$$w_{ij} = |x_j - x_i| \;\text{ ja }\; \hat{\beta}_{ij} = \frac{y_j - y_i}{x_j - x_i}.$$

Siis estimoiva yhtälö on

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_{ij} \,\operatorname{sign}(\hat{\beta}_{ij} - \beta_1) \;= 0.$$

Huomaa, että tappiofunktio ei riipu lainkaan vakiotermistä β_0 eikä sen estimointi näin ollen onnistu ko tekniikalla.
Ratkaisuna saadaan havaintopareista laskettujen estimaattien $\hat{\beta}_{ij}$ **painotettu mediaani** painoina w_{ij}. Nyt pätee

$$\hat{\beta}_1 \;\sim\; AN\left(\beta_1, \frac{1}{n} \frac{1}{12[\int f^2]^2} \frac{1}{s_x}\right).$$

HUOM. Havaintopareihin liittyvien estimaattien $\hat{\beta}_{ij}$ tavallinen (ei painotettu) mediaani tunnetaan nimellä **Theilin estimaatti**. Huomaa myös, että pienimmän neliösumman estimaatti on painotettu keskiarvo

$$\frac{1}{\sum_i \sum_j w_{ij}^2} \sum_i \sum_j w_{ij}^2 \hat{\beta}_{ij}$$

Testaus: Testisuure nollahypoteesille $H_0 : \beta_1 = 0$ konstruoidaan seuraavasti. Olkoot (vastemuuttujaan liittyvät keskistetyt järjestysluvut)

$$R_i = R(y_i; Y) - \frac{n+1}{2}, \quad i = 1, \ldots, n.$$

Nollahypoteesin vallitessa jakaumasta riippumaton testisuure on

$$T = \frac{1}{n} \sum x_i R_i.$$

Silloin standardoitu suure

$$\sqrt{n}\ \widehat{Corr(x, R)}$$

on likimain $N(0, 1)$-jakautunut.

Tarkat jakaumasta riippumattomat luottamusvälit parametrille β_1 saadaan jälleen 'kääntämällä' testi.

HUOM. Myös niin sanottuja **järjestyskorrelaatiokertoimia** käytetään usien riippuvuuden tutkimiseen: Tunnetuimmat ovat Blomqvistin testi, Spearmanin ρ ja Kendallin τ.

```
Regressioanalyysi: Yhden selittajan tapaus.

lse<-function(x,y)
{
cov(x,y)/var(x)
}

Theil<-function(x,y)
{
  n<-length(x)

  xapu<-matrix(rep(x,n),n)
  xero<-(xapu-t(xapu))/2
  xx<-c(xero[row(xero)<col(xero)])

  yapu<-matrix(rep(y,n),n)
  yero<-(yapu-t(yapu))/2
  yy<-c(yero[row(yero)<col(yero)])

  beta<- yy/xx
  median(beta)
}
```

```
rreg<-function(x,y)
{
  n<-length(x)
  xapu<-matrix(rep(x,n),n)
  xero<-(xapu-t(xapu))/2
  xx<-c(xero[row(xero)<col(xero)]) #{(x_i-x_j)/2  : i<j}
  yapu<-matrix(rep(y,n),n)
  yero<-(yapu-t(yapu))/2
  yy<-c(yero[row(yero)<col(yero)]) #{(y_i-y_j)/2  : i<j}
  beta<- yy/xx   #{(y_i-y_j)/(x_i-x_j)  : i<j}
  w<-abs(xx)     #{|(x_i-x_j)/2|  : i<j}
  ind<-order(beta)
  wcum<-cumsum(w[ind])/sum(w)
  beta<-beta[ind]
  i<-1
  while(wcum[i]<0.5)
  {
    i<-i+1
  }
  (beta[i-1]+beta[i])/2  #painotettu mediaani
}

lad<-function (x,y)
{
#
# Lasketaan LAD-estimaatti parametrille beta1
# R:n funktiolla rq (package: quantreg)
#
  rr<-rq(y~x)
  coef(rr)[[2]]
}
```

```
> x<-1:20
> y<-1+x+rnorm(20)
> lse(x,y)
[1] 0.9989793
> Theil(x,y)
[1] 1.005228
> rreg(x,y)
[1] 0.9973954
> library(quantreg)
> lad(x,y)
```

```
[1] 1.012800
Warning message:
Solution may be nonunique in: rq.fit.br(x, y, tau = tau, ...)
> theil1<-function(y) {Theil(x,y)}
> lse1<-function(y) {lse(x,y)}
> rreg1<-function(y) {rreg(x,y)}
> lad1<-function(y) {lad(x,y)}
> N<-1000
> data<-t(matrix(rep(x,N)+rnorm(20*N),nrow=20))
> varlse<-var(apply(data,1,lse1))
> varlse
[1] 0.001453307
> vartheil<-var(apply(data,1,theil1))
> vartheil
[1] 0.001610356
> varrreg<-var(apply(data,1,rreg1))
> varrreg
[1] 0.001595936
> varlad<-var(apply(data,1,lad1))
There were 50 or more warnings (use warnings() to see the first 50)
> varlad
[1] 0.002186614
> varlse/vartheil
[1] 0.9024756
> varlse/varrreg
[1] 0.9106298
> varlse/varlad
[1] 0.6646379
>
> data<-t(matrix(rep(x,N)+rlaplace(20*N),nrow=20))
> varlse<-var(apply(data,1,lse1))
> varlse
[1] 0.002905298
> vartheil<-var(apply(data,1,theil1))
> vartheil
[1] 0.002291751
> varrreg<-var(apply(data,1,rreg1))
> varrreg
[1] 0.002266289
> varlad<-var(apply(data,1,lad1))
There were 50 or more warnings (use warnings() to see the first 50)
> varlad
[1] 0.002459699
> varlse/vartheil
```

```
[1] 1.267720
> varlse/varrreg
[1] 1.281963
> varlse/varlad
[1] 1.181160
>
```

M-estimointi: Regressiokertoimien M-estimaatti saadaan minimoimalla

$$\sum_i \rho(e_i)$$

jollakin konveksilla parillisella ($\rho(-e) = \rho(e)$) funktiolla ρ tai vaihtoehtoisesti ratkaisemalla

$$\sum \psi(e_i) = 0 \quad \text{ja} \quad \sum x_i \psi(e_i) = 0,$$

missä $\psi(e) = \rho'(e)$.

Erikoistapauksena saadaan pienimmän neliösumman estimaatti ($\rho(e) = e^2$) ja LAD-estimaatti ($\rho(e) = |e|$). Ratkaisu yleisessä tapauksessa löytyy esim. Newton-Raphson tekniikalla, askeleena

$$\begin{pmatrix} \sum \psi'(e_i) & \sum x_i \psi'(e_i) \\ \sum x_i \psi'(e_i) & \sum x_i^2 \psi'(e_i) \end{pmatrix}^{-1} \begin{pmatrix} \sum \psi(e_i) \\ \sum x_i \psi(e_i) \end{pmatrix}$$

Estimaatin varianssin bootstrap-estimointi: Estimaatin $\hat{\beta} = \hat{\beta}(X, y)$ varianssin bootstrap-estimointi voidaan toteuttaa seuraavasti.

1. Etsi estimaatin arvo $\hat{\beta} = \hat{\beta}(X, y)$
2. Muodosta residuaalit $\hat{e} = y - X\hat{\beta}$
3. Muodosta M bootstrap-otosta residuaalien joukosta (palauttamatta otoskoolla n): Saadaan residuaalivektorit
$$e_1^*, \ldots, e_M^*$$
4. Muodosta M vastevektoria $y_i^* = X\hat{\beta} + e_i^*$, $i = 1, \ldots, M$
5. Etsi bootstrap-estimaatit $\hat{\beta}_i^* = \hat{\beta}(X, y_i^*)$, $i = 1, \ldots, M$
6. Estimaatin $\hat{\beta}$ varianssin bootstrap-estimaatti on (jos estimaatti harhaton)
$$\frac{1}{N} \sum (\hat{\beta}_i^* - \hat{\beta})^2$$

4.2 Usean selittäjän tapaus

Aineisto:

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \quad \text{ja} \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Malli:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{e}$$

missä $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ on tuntematon, $\boldsymbol{e} = (e_1, e_2, \ldots, e_n)^T$ on satunnaisotos origon suhteen symmetrisestä jakaumasta, jonka kertymäfunktio on F. Selittäjämatriisi X kiinteä ja tunnettu. Merkitään

$$D = \frac{1}{n} X^T X$$

Estimointi: Estimaatit eri tapauksissa saadaan ehdosta

$$X^T \boldsymbol{e} = \boldsymbol{0} \quad \text{tai} \quad X^T \boldsymbol{S} = \boldsymbol{0} \quad \text{tai} \quad X^T \boldsymbol{R} = \boldsymbol{0}$$

missä $S = (\text{sign}(e_1), \ldots, \text{sign}(e_n))$ ja $R = (R_1, \ldots, R_n)$ ovat residuaaleihin liittyvä merkki- ja (keskistetty) järjestyslukuvektori.

Testaus: Testi hypoteesille $\beta_p = 0$ (esimerkiksi) konstruoidaan seuraavasti.

1. Estimoi parametrit $\beta_0, \ldots, \beta_{p-1}$ mallista

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{i,p-1} + e_i, \quad i = 1, \ldots, n.$$

2. Merkitse

$$\hat{e}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_{p-1} x_{i,p-1}, \quad i = 1, \ldots, n.$$

3. Testisuure on joko

$$T = \sum [x_{ip} \hat{e}_i] \quad \text{tai} \quad T = \sum [x_{ip} \text{sign}(\hat{e}_i)] \quad \text{tai} \quad T = \sum [x_{ip} \hat{R}_i],$$

missä $\hat{R}_1, \ldots, \hat{R}_n$ ovar residuaaleihin $\hat{e}_1, \ldots, \hat{e}_n$ liittyvät järjestysluvut.

Kirjallisuutta

[1] Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.

[2] Fraser, D.A.S. (1957). *Nonparametric Methods in Statistics*. New York: Wiley.

[3] Gibbons, J.D. (1976). *Nonparametric Methods for Quantitative Analysis*. New York: Holt.

[4] Hajek, J. and Sidak, Z. (1967). *Theory of Rank Tests*. Prague: Academica.

[5] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel W.A. (1986). *Robust Statistics*. New York: Wiley.

[6] Hettmansperger, T.P. (1984). *Statistical Inference Based on Ranks*. New York: Wiley.

[7] Hettmansperger, T.P. and McKean, J.W. (1998). *Robust Nonparametric Statistical Methods*. London: Arnold.

[8] Huber, P.J. (1981). *Robust Statistics*. New York: Wiley.

[9] Hollander, M. and Wolfe, D.A. (1973). *Nonparametric Statistical Methods*. New York: Wiley.

[10] Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco: Holden-Day.

[11] Maritz, J.S. (1981). *Distribution-Free Statistical Methods*. London: Chapman and Hall.

[12] Puri, M.L. and Sen, P.K. (1981). *Nonparametric Methods in Multivariate Analysis*. New York: Wiley.

[13] Puri, M.L. and Sen, P.K. (1985). *Nonparametric Methods in General Linear Models*. New York: Wiley.

[14] Randles, R.H. and Wolfe, D.A. (1979). *Introduction to the Theory of Nonparametric Statistics*. New York: Wiley.

[15] Rousseeuw, P.J. and Leroy, M. (1987). *Robust Regression and Outlier Detection*. New York: Wiley.

[16] Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

Liite A

Liite

A.1 Kahden otoksen Wilcoxonin testiä (Moodin testiä) vastaavan estimaatin johtaminen

Oletetaan yksinkertaisuuden vuoksi, että havaintoaineistossa ei ole sidoksia. Etsitään sellainen Δ, jolla testisuure

$$S = S(\Delta) = \#\{y_j - \Delta > m_0\}$$

antaa kaksisuuntaiselle testille suurimman P-arvon. Nollahypoteesin vallitessa testisuureen S jakauma on symmetrinen odotusarvon $E_0(S) = k(1-\lambda)$ suhteen, joten suurimman P-arvon antaa sellainen Δ, jolla $S(\Delta)$ on lähimpänä odotusarvoa $k(1-\lambda)$. Oletetaan jatkossa, että $N = 2k$ (tapaus $N = 2k+1$ voidaan käsitellä vastaavalla tavalla). Tällöin nollahypoteesin mukaisen jakauman odotusarvoksi saadaan

$$E_0(S) = k(1-\lambda) = (N/2)(1-m/N) = (N/2)(n/N) = n/2.$$

Käsitellään ensin tapaus $n = 2p$ (eli y-havaintoja on parillinen määrä). Tällöin suurin P-arvo saavutetaan, kun $S(\Delta) = (2p)/2 = p$.

$$\begin{aligned}
\sum_{j=1}^{n} \operatorname{sign}(y_j - \Delta - m_0) &= \#\{y_j - \Delta > m_0\} - \#\{y_j - \Delta < m_0\} \\
&= \#\{y_j - \Delta > m_0\} - [n - \#\{y_j - \Delta > m_0\} - \#\{y_j - \Delta = m_0\}] \\
&= 2 \cdot \#\{y_j - \Delta > m_0\} - n + \#\{y_j - \Delta = m_0\} \\
&= 2S(\Delta) - n + \#\{y_j - \Delta = m_0\}
\end{aligned}$$

Kappaleen 2.1.12 perusteella yhtälön

$$\sum_{i=1}^{n} \operatorname{sign}(d_i - \Delta) \doteq 0 \tag{A.1}$$

ratkaisuna saadaan aineiston $D = \{d_1, \ldots, d_n\}$ otosmediaani $\operatorname{Med}(D)$. Huom! Merkintä "$\doteq 0$" tarkoittaa tässä "$= 0$ tai vaihtaa merkkiä".