Oulun yliopiston matemaattisten tieteiden laitos/tilastotiede

806113P TILASTOTIETEEN PERUSTEET, kl 2011 (Esa Läärä)

M-harjoitus 1, viikot 3-4 (21.1.-27.1): Johdatus R-kieleen.

Harjoituksessa tutustutaan R-kielen interaktiiviseen käyttöön ja perehdytään mm. reaalilukujen ja vektorien tavanomaisiin laskutoimituksiin, funktioiden käsittelyyn ja niiden kuvaajien piirtämiseen.

1. Mikä R on?

R on ohjelmointikieli ja -ympäristö tilastollista laskentaa ja grafiikkaa varten. Ohjelmointikielenä se on luonteeltaan sekä funktionaalinen että olio- eli objektiorientoitunut. Tämä tarkoittaa, että R:n komennot koostuvat erilaisten funktioiden kutsuista, joiden syötteinä ja tuloksina on erityyppisiä datarakenteita eli olioita.

R on vapaan lähdekoodin ohjelmisto, joka on lisenssimaksutta imuroitavissa R-projektin kotisivulta `www.r-project.org/`.

Suomenkielisiä R-oppaita ovat mm. Jari Oksasen (`http://cc.oulu.fi/~jarioksa/opetus`), sekä Marja-Leena Hannilan ja Vesa Kiviniemen (`http://www.uef.fi/tike/ohjelmistot`, linkki `'R-opas.pdf'`) kirjoittamat.

R käynnistetään klikkaamalla työpöydän ao. ikonia, joka aukaisee ikkunan `RGui` ja sen sisällä ns. konsoli-ikkunan `R Console`.

Interaktiivisessa käytössä operoidaan ensisijaisesti konsoli-ikkunassa. Käyttäjä antaa omat komentonsa kehotemerkillä `'>'` alkavalle riville tässä ikkunassa. Ohjelma antaa vastauksen komentoon tätä seuraavalla rivillä. Alkupuolen yksinkertaisissa esimerkeissä vastausrivi alkaa tavallisesti merkinnällä `'[1]'` (joka viittaa tulostettavan vektorin ensimmäiseen koordinaattiin). Grafiikkakomentojen tulokset näkyvät erillisessä grafiikkaikkunassa.

Mikäli kehoteriville annettava komento on sijoituskomento eikä sisällä tulostuskäskyä, ohjelma ei kirjoita enterin painamisen jälkeen mitään vaan tallettaa halutun sijoituksen työmuistiin.

R-istunnon aikaista komentohistoriaa (eli aiemmin annettuja R-komentoja) voi selata nuolinäppäimiä $\uparrow$ ja $\downarrow$ käyttäen. Niiden avulla yksittäisen komennon mahdolliset syntaksivirheet on nopea korjata tai muuten ajaa komento uudestaan esim. uusilla argumenttien ja parametrien arvoilla.

`RGui`-ikkunan vasemmalla ylälaidalla olevan valikkorivin `help`-valikosta löytyy mm. html-pohjainen help-järjestelmä sekä mahdollisuus käyttää erilaisia hakutoimintoja. Nopea tapa löytää opastusta jostakin funktiosta on kirjoittaa komentoriville kysymysmerkin jälkeen funktion nimi, esim. piirrosfunktio `?plot`. Tämä aukaisee selaimen ja kyseisen funktion help-sivun.

2. R laskimena

R:ää voi käyttää tavanomaisen laskimen tapaan. Esim. kirjoittamalla komentoriville $2+3$ saadaan enterin painamisen jälkeen summa seuraavalle riville:

```
> 2+3
[1] 5
```

R käyttää peruslaskutoimituksissa tavanomaisia operaattoreita ja noudattaa laskujärjestyksen ja sulkujen käytön tuttuja sääntöjä; esim. lausekkeen $2^3 \times (5-2)/6 + 1$ arvo saadaan

```
> 2^3*(5-2)/6+1
[1] 5
```

R sisältää myös kaikki tärkeät perusfunktiot, kuten neliöjuuri-, eksponentti-, Napierin logaritmi- (ln; R:ssä `log`), Briggsin logaritmi- (lg; R:ssä `log10`), trigonometriset ym. funktiot sekä vakioita kuten π.

```
> sqrt(2)
[1] 1.414214
> exp(1)
[1] 2.718282
```

Yhdelle riville voidaan kirjoittaa puolipisteellä toisistaan erotettuna useampi laskutoimitus tai muu funktiokutsu, joiden tulokset ilmestyvät alekkaisille riveille.

```
> log(2) ; log10(2) ; pi ; sin(pi/4)
[1] 0.6931472
[1] 0.30103
[1] 3.141593
[1] 0.7071068
```

3. Sijoitusoperaattori

Yksittäisen luvun tai laskutoimituksen tulos voidaan myöhempää käyttöä varten tallettaa muistiin sijoittamalla se omaan muuttujaansa sijoitusoperaattoria käyttäen. Muuttujan tunnukseksi kelpaa isolla tai pienellä kirjaimella alkava yhtenäinen merkkijono, jossa alkukirjaimen jälkeen voi esiintyä muita kirjaimia, numeroita sekä piste '.' ja alaviiva '`-`'. R erottaa isot ja pienet kirjaimet toisistaan, joten esim. '`a`' ja '`A`' eivät tarkoita samaa.

Ensisijainen sijoitusoperaattori '`<-`' koostuu kahdesta merkistä: '`<`' ja '`-`', jotka kirjoitetaan yhteen. (Sijoitusoperaattoriksi kelpaa myös '`=`'-merkki.) Kirjoittamalla komentoriville `a <- 2` sijoitetaan numeerisen (skalaari)muuttujan `a` arvoksi luku 2. Haluttaessa muuttujan sisältö tai eri muuttujia sisältävän laskutoimituksen lopputulos näkyviin kirjoitetaan yksinkertaisesti ao. muuttujan nimi tai vastaava lauseke.

```
> a <- 2
> b <- 3
> a
[1] 2
> a+b
[1] 5
> a^b
[1] 8
> n <- 10+2
> n
[1] 12
> n/a
[1] 6
```

4. Vektorit ja laskutoimitukset niillä

R:n perustietorakenne on **vektori** (`vector`), joka on matematiikasta tuttuun tapaan samantyyppisten alkioiden (esim. reaalilukujen) järjestetty joukko. (Edellisten esimerkkien muuttujat sisälsivät kukin skalaarin eli vain yhden alkion sisältävä vektorin.) Perusfunktio vektorien muodostamiseksi on funktio `c()`, joka yhdistää (= "*concatenate*") yksittäiset alkiot pilkulla toisistaan erotettuina samaan vektoriin.

Talletetaan kuuden opiskelijan painot (kg) ja pituudet (cm) vektoreihin `paino` ja `pituus`.

```
> paino <- c(60,72,57,90,95,72)
> paino
[1] 60 72 57 90 95 72
> pituus <- c(175,180,165,190,174,191)
```

Tiettyjä yksinkertaisia sääntöjä noudattavia vektoreita voidaan muodostaa myös mm. komentojen `seq()` ja `rep()` avulla. Kutsumalla `seq(1,10)` saadaan muodostetuksi tasavälinen jono 1:stä 10:een 1:n välein. Sama saadaan aikaan kirjoittamalla `1:10`. Jos kuitenkin peräkkäisten lukujen väli on jotain muuta kuin 1, pitää tämä ilmoittaa parametrin `by` avulla

```
> seq(1,10)
 [1]  1  2  3  4  5  6  7  8  9 10
> ykskym <- 1:10
> ykskym
 [1]  1  2  3  4  5  6  7  8  9 10
> x2 <- seq(1,5,by=0.5)
> x2
[1] 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0
```

Huomaa, että R:ssä desimaalierottimena on piste eikä pilkku. Funktion `rep()` kutsulla `rep(1, times=10)` toistetaan puolestaan lukua 1 kymmenen kertaa.

```
> rep(1,times=10)
 [1] 1 1 1 1 1 1 1 1 1 1
```

Mikäli vektorit ovat samanmittaiset, kaikki matemaattiset operaatiot tehdään alkioittain. Muunnetaan esimerkiksi senttimetreinä mitattu pituus metreiksi:

```
> pituus.m <- pituus/100
> pituus.m
[1] 1.75 1.80 1.65 1.90 1.74 1.91
```

Nyt voidaan laskea kunkin opiskelijan *body mass index* eli BMI = paino (kg)/pituus^2 (m^2); myös pyöristää tulostus kahteen desimaaliin.

```
> bmi <- paino/pituus.m^2
> bmi
[1] 19.59184 22.22222 20.93664 24.93075 31.37799 19.73630
> round(bmi,digits=2)
[1] 19.59 22.22 20.94 24.93 31.38 19.74
```

Vektorin `paino` sisältämien arvojen summa saadaan laskettua funktiolla `sum()`

```
> sum(paino)
[1] 446
```

ja aritmeettinen keskiarvo ($= \sum x_i/n$) kirjoittamalla

```
> sum(paino)/length(paino)
[1] 74.33333
```

jossa funktio `length()` laskee argumenttivektorin alkioiden lukumäärän. Keskiarvo voidaan laskea myös suoraan funktion `mean()` avulla:

```
> mean(paino)
[1] 74.33333
```

Edellä määriteltyihin vektoreihin talletettujen data-alkioiden tyyppi on kaikissa ollut **numeerinen** (`numeric`). Alkioiden tyyppi voi olla myös **merkkimuotoinen** (`character`), kuten seuraavassa

```
> sukupuoli <- c("nainen","mies","nainen","mies","mies","mies")
> sukupuoli
[1] "nainen" "mies"   "nainen" "mies"   "mies"   "mies"
```

Kolmas tärkeä vektorin tyyppi on **looginen** (`logical`). Tällaisella vektorilla kuvataan mm. vektorin alkioiden arvoja koskevien yhtäsuuruus- ja epäyhtälötyyppisten väitteiden paikkansapitävyyttä, jolloin mahdollisia arvoja ovat `TRUE` ja `FALSE`.

Jos esimerkiksi halutaan poimia erikseen tutkittavaksi ne yksilöt, joille pätee epäyhtälö BMI < 18 kg/m^2 tai jotka toteuttavat ehdon $20 \leq$ BMI < 25, voidaan muodostaa näitä vastaavat loogiset muuttujat `hoikka` ja `normpaino`

```
> hoikka <- bmi < 20 ; hoikka
[1]  TRUE FALSE FALSE FALSE FALSE  TRUE
> normpaino <- (20 <= bmi) & (bmi < 25)
> normpaino
[1] FALSE  TRUE  TRUE  TRUE FALSE FALSE
```

Lukujen suuruusjärjestyksen vertailuja merkitään `<=, <, >=, >, <>, ==`. Huomaa myös loogisen konjunktion "ja" merkki '`&`'. Loogista disjunktiota "tai" merkitään '`|`' ja negaatiota "ei" huutomerkillä '`!`'.

5. Uudet funktiot ja R graafisena laskimena

Valmiiden funktioiden lisäksi käyttäjä voi määritellä niiden avulla uusia funktiota omiin tarpeisiinsa. Otetaan esimerkiksi L-harjoituksessa 1 käsitelty $N(0,1)$-jakauman tiheysfunktio $\varphi(z) = (2\pi)^{-1/2}\exp(-z^2/2)$, jota vastaavalle R:n funktiolle annamme nimen `fii` ja määrittelemme sen seuraavasti

```
> fii <- function(z) (2*pi)^(-1/2)*exp(-z^2/2)
```

Tämän jälkeen voimme laskea $\varphi(z)$:n arvon haluamillamme z:n arvoilla kuten pisteessä 0 kutsumalla funktiotamme ao. argumentilla.

```
> fii(0)
[1] 0.3989423
```

Jos haluamme tietää $\varphi(z)$:n suuruuden useilla z:n arvoilla samalla kertaa, kuten esim. L-harjoituksen kotitehtävässä 2.(a), voimme koota kaikki arvot yhteen vektoriin ja kutsua funktiota käyttäen tätä vektoria argumenttina.

```
> zz <- c(-3, -1, 0, 0.5, 2)
> round(fii(zz), 4)
[1] 0.0044 0.2420 0.3989 0.3521 0.0540
```

Vertaa tulostetun rivin lukuja em. kotitehtävän vastaukseen.

Funktion kuvaaja halutulla välillä kuten $[-3.5, 3.5]$ voidaan piirtää komennolla `curve()`. Se aukaisee grafiikkaikkunan, johon kuvaaja ilmestyy.

```
curve(fii, from = -3.5, to = 3.5)
```

Numeerinen integrointi on myös mahdollista R:ssä, kunhan integroitava funktio on riittävän siisti. Jos esim. halutaan laskea $N(0,1)$-jakauman kertymäfunktion $\Phi(z)$ arvo pisteessä $z = 2$ eli tiheysfunktion määrätty integraali $\Phi(2) = \int_{-\infty}^{2} \varphi(z)dz$, niin tämä onnistuu kirjoittamalla

```
> integrate(fii, lower = -Inf, upper=2)
0.9772499 with absolute error < 1.6e-06
```

R osaa myös derivoida riittävän siistejä funktiota. Nyt täytyy kuitenkin kirjoittaa derivoitavan funktion lauseke kokonaan auki. Derivoidaan malliksi $N(0,1)$-tiheysfunktion ydin eli funktio $g(z) = e^{-z^2/2}$, jonka derivaatta on $g'(z) = -ze^{-z^2/2}$:

```
D(expression(exp(-z^2/2)), "z")
-(exp(-z^2/2) * (2 * z/2))
```

6. Todennäköisyysjakaumat

Oman funktion määrittely normaalijakauman tiheydelle ja sen integrointi edellä esitettyyn tapaan ei kuitenkaan ole tarpeen, koska R:ssä on valmiiksi ohjelmoituina tärkeimpien todennäköisyysjakaumien avainfunktiot

R:n sisältämät jakaumafunktiot ovat neljää eri tyyppiä:
- `d`*jak* ; pistetodennäköisyys- tai tiheysfunktio,
- `p`*jak* ; kertymäfunktio,
- `q`*jak* ; kvantiili- eli fraktiilifunktio – kertymäfunktion käänteisfunktio,
- `r`*jak* ; jakaumasta satunnaislukuja generoiva funktio,

jossa "*jak*" viittaa jakauman R-nimeen. Esimerkiksi normaalijakauman R-nimi on `norm`, jolloin sen tiheys- ym. funktiot ovat oikealta nimeltään `dnorm, pnorm, qnorm` ja `rnorm`. Binomijakauman R-nimi on puolestaan `binom`, jolloin sen pistetodennäköisyysfunktio on `dbinom` ja muut jakaumafunktiot `pbinom, qbinom` ja `rbinom`. Kullakin näistä on omanlaisensa parametrointi, joka on syytä selvittää ao. jakauman `help`-sivulta (kysymysmerkki eteen, esim. `?rnorm`). Simuloinnissa tarvitaan erityisesti `r`*jak*-tyyppisiä funktioita.

Suoritetaanpa nyt R:n työkaluilla samat laskelmat kuin tehtiin edellä funktion `fii` avulla. Argumentiksi voidaan antaa useamman z:n arvon sisältävä vektori, jolloin myös vastaukseksi saadaan vastaava vektori kutsutun funktion arvoista.

```
> dnorm(0)
[1] 0.3989423
> round(dnorm(zz), 4)
[1] 0.0044 0.2420 0.3989 0.3521 0.0540
> pnorm(2)
[1] 0.9772499
> round(pnorm(zz), 4)
[1] 0.0013 0.1587 0.5000 0.6915 0.9772
```

Kertymäfunktion $\Phi(z)$ kuvaajan piirto onnistuu tuttuun tapaan.

```
> curve(pnorm, -4, 4)
```

$N(0,1)$-jakauman kertymäfunktion käänteisfunktio Φ^{-1} on jakauman kvantiili- eli fraktiilifunktio, josta saadaan tämän jakauman kvantiileja z_p halutuissa pisteissä; esim. kun $p \in \{0.025, 0.1, 0.67, 0.95\}$ kuten L-harjoituksen tehtävässä 2.(b). Kvantiilifunktion kuvaaja voidaan myös piirtää ja tarkastella sen kulkua grafiikkaikkunassa

```
> qnorm( c(0.025, 0.1, 0.67, 0.95) )
[1] -1.9599640 -1.2815516  0.4399132  1.6448536
> curve(qnorm, 0.001, 0.999)
```

Yleisen normaalijakauman $N(\mu, \sigma^2)$ käsittely onnistuu samoilla funktioilla kuin edellä, kunhan antaa funktion kutsussa lisäargumentit `mean` ja `sd` vastaten odotusarvoa μ ja keskihajontaa σ (Huom: σ eikä σ^2!). Esimerkiksi L-harjoituksen 1 tehtävässä 3. on $\mu = 166$ cm ja $\sigma^2 = 5^2$ cm^2, jolloin $\sigma = 5$ cm. Tehtävän kohdan (a) todennäköisyydet lasketaan seuraavilla riveillä

```
> pnorm(150, mean=166, sd=5)
[1] 0.000687138
> pnorm(180, 166, 5) - pnorm(150, 166, 5)
[1] 0.9967577
> 1 - pnorm(180, 166, 5)
[1] 0.002555130
```

95% viitevälin rajat saamme kvantiilifunktiosta

```
> qnorm( c(0.025, 0.975), 166, 5)
[1] 156.2002 175.7998
```

Lisäksi voimme mm. katsoa jakauman tiheysfunktion kuvaajaa välillä [140 cm, 190 cm]

```
curve( dnorm(x, 166, 5), 140, 190)
```

L-harjoituksen 2 tehtävässä 4.(a) kohteena oli binomijakauma $\mathrm{Bin}(6, 1/3)$. Lasketaan tästä jakaumasta pistetodennäköisyydet $\mathbb{P}(X = 0)$ ja $\mathbb{P}(X = 1)$ kuin myös kertymäfunktion arvo pisteessä 1 eli $P(X \leq 1)$:

```
> dbinom(0, size=6, prob=2/3)
[1] 0.001371742
> dbinom(1, 6, 2/3)
[1] 0.01646091
> pbinom(1, 6, 2/3)
[1] 0.01783265
```

Tulostetaan seuraavaksi kaikki pistetodennäköisyydet $p_k = \mathbb{P}(X = k),\ k = 0, 1, \ldots, 6$ ja piirretään ptnf:n kuvaaja:

```
> k <- 0:6
> pk <- dbinom(k, 6, 2/3)
> round(rbind(k, pk), 4)
     [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]
k  0.0000 1.0000 2.0000 3.0000 4.0000 5.0000 6.0000
pk 0.0014 0.0165 0.0823 0.2195 0.3292 0.2634 0.0878
> plot( k, pk, type="h")
```

7. Tutustuminen perusfunktioihin

Sijoita vektoreihin `a`, `b` ja `c` kuhunkin viisi lukua seuraavasti

```
> a <- c(2, 4, 3, 1, 5)
> b <- c(2.2, 4.5, 3.1, 1.8, 5.3)
> c <- c(1, 1, 0, 0, 1)
```

Kokeile, mitä seuraavat funktiokutsut saavat aikaan:

```
b[1]        # (yksittäisen alkion valinta indeksin mukaan)
b[2:4]      # (perättäisten alkioiden valinta)
b[c(1,3,5)] # (valinta indeksien osajoukon mukaan)
b[-1]       # (yksittäisen alkion poisjättäminen)
b[-c(2,4)]  # (alkioiden osajoukon poisjättäminen)
b[b>4]      # (alkioiden valinta niitä koskevan kriteerin mukaan)
a[c==1]     # (valinta toisen vektorin määräämän kriteerin mukaan)

length(b)   # (vektorin alkioiden lukumäärä)
max(a)      # (maksimi vektorin luvuista)
min(b)      # (minimi vektorin luvuista)
range(b)    # (vektorin arvojen vaihteluväli)
sum(a)      # (vektorin alkioiden summa)
prod(a)     # (vektorin alkioiden tulo)
mean(a)     # (vektorin alkioiden aritmeettinen keskiarvo)

sort(b)     # (järjestää vektorin alkiot suuruusjärjestykseen)
rev(b)      # (kääntää vektorin alkiot päinvastaiseen järjestykseen)
floor(b)    # (pyöristää vektorin alkiot lähimpään kokonaislukuun alaspäin)
ceiling(b)  # (pyöristää vektorin alkiot lähimpään kokonaislukuun ylöspäin)
round(b,0)  # (pyöristää lähimpään kokonaislukuun)
7%/%2       # (jakolaskun kokonaisosa)
7%%2        # (jakolaskun jakojäännös)
```

8. HARJOITUSTEHTÄVIÄ

Tee R-ohjelmalla seuraavat tehtävät. Käytä ratkaisuissa apunasi edellä esitettyjä R:n ominaisuuksia.

1. Yksinkertaisia vektoreita, niiden summia ja tuloja

$$\text{(a)} \sum_{i=1}^{10} i \qquad \text{(b)} \sum_{j=1}^{7} 2^j \qquad \text{(c)} \prod_{i=1}^{10} i \qquad \text{(d)} \prod_{j=1}^{7} 2^j \qquad \text{(e)} \prod_{k=1}^{5} 3.$$

2. Alkioiden valinta kokonaislukujen jonosta.

(a) Talleta `luvut1`-nimiseen vektoriin kaikki parittomat kokonaisluvut väliltä 1–1000.

(b) Talleta `luvut2`-nimiseen vektoriin kaikki parilliset kokonaisluvut väliltä 1–1000.

(c) Laske kohtien (a) ja (b) vektoreiden lukujen summat.

3. Tallenna alla olevilla sijoituskomennoilla kymmenen opiskelijan sukupuoli- ja painotiedot vektoreihin `sukupuoli` (koodeina 1 = nainen ja 2 = mies) ja `paino`:

```
sukupuoli <- c(1,1,2,1,2,2,2,1,2,2)
paino <- c(60,69,75,55,90,68,80,61,74,70)
```

(a) Tulosta yli 70 kiloa painavien opiskelijoiden painot.

(b) Tulosta kaikkien miesten painot.

(c) Laske miesten painojen summa.

4. Piirrä seuraavien funktioiden kuvaajat välillä [-4,4]:

$$(a)\ f(x) = x^3 - 3x, \quad (b)\ f(x) = e^x, \quad (c)\ f(x) = \sin(x), \quad (d)\ f(x) = \cos^2(3x)$$

5. Olkoon $X \sim N(180, 6^2)$ mallina pituuden (cm) jakaumalle miesopiskelijoiden keskuudessa.

(a) Piirrä X:n tiheysfunktio välillä [150 cm, 210 cm]

(b) Laske todennäköisyydet

$$(i)\ \mathbb{P}(X \leq 170), \quad (ii)\ \mathbb{P}(170 < X \leq 185), \quad (iii)\ \mathbb{P}(X > 185).$$

(c) Laske 5% ja 95% fraktiilit, jotka rajaavat 90% viitevälin.

Oulun yliopiston matemaattisten tieteiden laitos/tilastotiede

806113P TILASTOTIETEEN PERUSTEET, kl 2011 (Esa Läärä)

M-harjoitus 2, viikot 5-6 (4.-9.2.): mikroluokkatehtävät

Kun R:llä alkaa tehdä vähänkään vaativampia analyysejä kuin tämän kurssin ensimmäisessä M-harjoituksessa, niin on hyvä perehtyä seuraaviin ohjelmateknisiin seikkoihin: (A) työhakemisto, (B) R-skripti, (B) ulkoisen datatiedoston lukeminen.

(A) Työhakemisto. Kussakin erillisessä tilastollisessa analyysiprojektissa tarvittavat skriptit ja datatiedostot kannattaa sijoittaa tietokoneessa omaan hakemistoonsa. Tälle kurssille voit luoda `paju`-koneen Samba-hakemistoosi oman alihakemiston esim. nimeltä `TPR`.

Kun sitten käynnistät R:n niin ensimmäisenä tehtävänäsi on asettaa ao. projektille nimetty hakemisto istuntosi työhakemistoksi (*working directory*) seuraamalla `RGui`-ikkunan vasemmasta ylänurkasta lähtien valikkopolkua `File - Change dir ...` Näin menetellen voit lukea istunnon aikana tarvitsemasi ulkoiset tiedostot ohjelman käytettäväksi kirjoittamalla lukukomennon pääargumentiksi pelkästään ao. tiedoston nimen. Jos tiedosto sijaitsee jossain muussa kansiossa, niin koko hakemistopolku pitää antaa lukukomennossa.

(B) R-skripti (*R script*) on ohjelmatiedosto, johon talletetaan ne R-komennot, jotka tarvitaan haluttujen analyysitehtävien suorittamiseksi. Skripti voidaan kirjoittaa esimerkiksi R:n omalla skriptieditorilla, jonka ikkuna aukeaa kun `RGui`-ikkunan vasemmasta ylänurkasta lähtien seurataan valikkopolkua `File - New script ...` Avoimeen ikkunaan voi alkaa kirjoittaa R-komentoja rivi kerrallaan. Rivien korjaus, muokkaaminen ja skriptin tallentaminen onnistuu `RGui`-ikkunan ylärivin valikkojen `File` ja `Edit` sisältämien tavallisten työkalujen avulla.

Yhden tai useammankin komentorivin aktivoiminen toteutetaan siten, että ensin nämä rivit maalataan, sitten kursori viedään `RGui`-ikkunan ylänurkassa olevan valikkorivin alapuolella olevan ikonirivin keskimmäisen ikonin ('*Run line or selection*') päälle ja lopuksi klikataan. Maalatut komentorivit siirtyvät konsoli-ikkunaan, ja ohjelma alkaa suorittaa niitä.

(C) Datakehikko. R-istunnossa tarvittavat ulkoiset datatiedostot luetaan sisään R:n datakehikoksi (*data frame*) joko komennolla `read.table()` tai muilla lukukomennoilla riippuen luettavan tiedoston formaatista. Oletusarvona on tavanomainen vapaan formaatin ascii-tiedosto, joka on järjestetty havaintomatriisin muotoon siten, että sen rivit liittyvät eri havaintoyksiköihin, sarakkeet eri muuttujiin, ja sarakkeiden erottimena on välilyönti.

1. Työhakemiston käyttöönotto, skriptin kirjoittaminen ja ulkoisen datatiedoston lukeminen.

(a) Luo alihakemisto `TPR` oman Samba-hakemistosi sisälle, käynnistä R ja vaihda em. alihakemisto R-istuntosi työhakemistoksi.

(b) Aukaise R:n skriptieditori-ikkuna ja ala kirjoittaa siihen tämän istunnon R-komentoja. Voit kirjoittaa kommentteja merkin '`#`' jälkeen niin komentorivien väliin kuin myös rivien loppuun. Talleta skripti esim. nimellä `tp-mh2.R` työhakemistoosi.

(c) Kopioi nyt omaan työhakemistoosi mikroluokan P-levyn hakemistosta `P:\TP2011\` siellä oleva ascii-muotoinen datatiedosto nimeltä `tpdata.txt`. Avaa tiedosto esim. *Notepad*-ohjelmalla tai vastaavalla ja tarkastele sen sisältöä. Kirjoita skriptiisi ja aja seuraava komento, joilla em. datatiedosto luetaan R:n datakehikoksi `tp`:

```
tp <- read.table('tpdata.txt', header = T)
```

Oulun yliopiston matemaattisten tieteiden laitos/tilastotiede
TILASTOTIETEEN PERUSTEET, kl 2011

L-harjoitus 1, viikko 3 (to 20.1.): **Mittaus- ja datankeruu**

Harjoituksessa kerätään kultakin osallistujalta arvot seuraavista 19 muuttujasta, joiden tunnukset on kirjoitettu isoilla kirjaimilla. Täytä lomake, tee mittaukset ja tallenna omat arvosi tiedostoon harjoituksen vetäjän antamien ohjeiden mukaan.

1. SUKUP: Sukupuoli (rengasta) 1 mies 2 nainen

2. SYNTV: syntymävuosi 19_____ **3.** IKA: Ikä _____ vuotta

4. VERIRYHM: Veriryhmä (rengasta) A B AB O X (en tiedä)

5. KOTITAL: Kuinka monta henkeä kotitalouteesi kaikkiaan kuuluu? _____ henkeä

6. ISANPIT: Isäsi pituus _____cm **7.** AIDINPIT: Äitisi pituus_____cm

8. VELJET: Veljien lukumäärä_____ **9**. SISKOT: Siskojen lukumäärä_____

10. LIIKUNTA: Kuinka usein harrastat liikuntaa vähintään ½ tuntia kerrallaan niin, että ainakin lievästi hengästyt ja hikoilet? (rengasta)

0	en lainkaan tai hyvin harvoin	3	2-3 kertaa viikossa
1	1-3 kertaa kuukaudessa	4	4-6 kertaa viikossa
2	noin kerran viikossa	5	päivittäin

11. PAINEM1Y: Verenpaine, 1. mittaus, systolinen eli yläpaine _____ /mmHg

12. PAINEM1A: Verenpaine, 1. mittaus, diastolinen eli alapaine _____ /mmHg

13. SYKEM1: Leposyke, 1. mittaus _____ /min

14. PAINEM2Y: Verenpaine, 2. mittaus, systolinen eli yläpaine _____ /mmHg

15. PAINEM2A: Verenpaine, 2. mittaus, diastolinen eli alapaine _____ /mmHg

16. SYKEM2: Leposyke, 2. mittaus _____ /min

17. PITUUSA: Oma arvio pituudestasi ennen mittausta _____ cm

18. PITUUSM: Pituus, mittauksen tulos _____ cm

19. TUPAKKA: Tupakoitko nykyisin?

0 en lainkaan
1 kyllä, harvemmin kuin kerran viikossa
2 kyllä, joka viikko, mutta en päivittäin
3 kyllä, päivittäin

2. Tällä kurssilla toteutetussa datankeruu- ja mittausharjoituksessa 31 osallistujaa antoi itsestään tiedot 20 muuttujasta. Vuonna 2010 samat tiedot saatiin vastaavan kurssin 49 osallistujalta, ja havaintoaineistomme sisältää molempien vuosien havainnot. Muistin virkistämiseksi datankeruussa käytetty lomake on viereisellä sivulla. Havainnot on siis tallennettu tiedostoon `tpdata.txt`, joka edellä luettiin R-datakehikoksi `tp`.

Kirjoita skriptiisi ja aja seuraavat komennot, joilla datakehikon `tp` rakennetta kuvataan funktiolla `str()`, listataan sen ensimmäiset 5 riviä, lasketaan funktiolla `summary()` itse kunkin muuttujan suoran jakauman eräitä tunnuslukuja. Lopuksi funktiolla `attach()` kiinnitä datakehikko R:n sisäiseen hakupolkuun, jotta sen sisäisiin muuttujanimiin voidaan viitata suoraan myöhemmissä komennoissa. Skripti kannattaa tallettaa uudelleen aina muutaman uuden rivin kirjoittamisen jälkeen. – Pysähdy hetkeksi katsomaan komentojen tuloksia.

```
str(tp)     # rivien ja sarakkeiden määrät; muuttujat ja niiden tyypit
tp[1:5, ]   # ensimmäiset 5 riviä, kaikki sarakkeet
summary(tp)
attach(tp)
```

3. Tarkastellaan leposykkeen arvojen jakaumaa 1. mittauskerralta (muuttuja `SYKEM1`).

(a) Piirrä sykearvojen runko-lehtikuvio

```
stem(SYKEM1)
```

(b) Laske leposykkeen minimi, maksimi, mediaani, kvartiilit, keskiarvo sekä keskihajonta. Huomaa, että nämä laskevien funktioiden kutsussa pitää antaa lisäargumentiksi `na.rm=TRUE`, joka poistaa joiltakin osallistujilta puuttuvat sykearvot laskentaa häiritsemästä.

```
min(SYKEM1, na.rm=T) ; max(SYKEM1, na.rm=T)
median(SYKEM1, na.rm=T)
quantile( SYKEM1, probs = c(0.25, 0.75), na.rm=T)
mean(SYKEM1, na.rm=T) ; sd(SYKEM1, na.rm=T)
```

(c) Totea, että keskihajontaa lukuunottamatta em. tunnusluvut saadaan samanaikaisesti funktion `summary()` kutsulla, jonka yksi tulos on puuttuvien havaintojen lukumäärä.

```
summary(SYKEM1)
```

(d) Verrataan sykkeen jakauman tunnuslukuja miesten ja naisten välillä.

```
tapply(SYKEM1, SUKUP, summary)
```

Sukupuoli on koodattu numeroin 1 = 'mies' ja 2 = 'nainen'. Muodostetaan sukupuolesta uusi luokkamuuttuja (`factor`) nimeltä `sukup`, jossa arvoina käytetään luokkien selväkielisiä nimiä, ja lisätään se datakehikkoon `tp`. Tulostetaan sen jälkeen sukupuolittaiset tunnusluvut toistamiseen. – Mitä yhtäläisyyksiä ja mitä eroja havaitset sykkeen jakaumassa miesten ja naisten välillä?

```
tp$sukup <- factor(tp$SUKUP, labels = c("mies", "nainen") )
attach(tp)
tapply(SYKEM1, sukup, summary)
```

HUOM. Luotaessa uusia muuttujia datakehikon alkuperäisistä muuttujista, niin ne kannattaa heti sitoa ao. datakehikkoon kuten yllä eikä jättää irrallisiksi muuttujiksi. Niinpä uuden muuttujan `sukup` nimi ("etunimi") em. sijoituksessa kirjoitettiin datakehikon nimen (muuttujan "sukunimi") ja dollarimerkin perään. Sen jälkeen täydennetty datakehikko voidaan jälleen kiinnittää hakupolkuun funktiolla `attach()`.

4. Piirretään leposykkeen pistekuvio, laatikko-janakuvio ja otoskertymäfunktion kuvaaja.

(a) Kutsu funktiota `stripchart()` sen oletusarvoilla

```
stripchart(SYKEM1)
```

Päällekkäiset havaintopisteet eivät erotu, joten on parempi panna niitä pinoon eli käyttää optiota '`stack`'. Vaihdetaan myös havaintopisteen symboli umpipalloksi (`pch=16`).

```
stripchart(SYKEM1, method='stack', pch=16, xlab='Leposyke (/min)' )
```

(b) Piirrä miesten ja naisten sykearvot alekkain samaan kuvioon

```
stripchart(SYKEM1 ~ sukup, method='stack', pch=16, xlab='Leposyke (/min)' )
```

Metodin '`stack`' asemesta kokeile metodia '`jitter`' ja vertaa vaikutelmaa:

```
stripchart(SYKEM1 ~ sukup, method='jitter', pch=16, xlab='Leposyke (/min)' )
```

(c) Piirrä sykkeen jakauman laatikko-janakuvio vaakasuoraan.

```
boxplot(SYKEM1, horizontal=T)
```

ja sen jälkeen miehille ja naisille alekkain samaan kuvaan.

```
boxplot(SYKEM1 ~ sukup, horizontal=T)
```

(d) Piirrä otoskertymäfunktion porraskuvio, jossa oikealta jatkuvuus kussakin hypyssä osoitetaan oletusarvosta puoleen pienennetyllä (`cex=0.5`) umpipallolla

```
plot(ecdf(SYKEM1), cex=0.5, xlab= 'Leposyke (/min)',
    main = 'Otoskertymäfunktio' )
abline( h = 0.25*(1:3), col= 'gray', lty=2 ) # hilaviivat 25% välein
```

5. Jatkuvan muuttujan luokittelu ja taulukointi.

(a) Funktiolla `cut()` voidaan jatkuva muuttuja luokitella haluttuihin luokkiin, joiden todelliset luokkarajat annetaan vektorina argumentin `br` arvoksi. Jos leposykkeen pyöristetyiksi luokkarajoiksi valitaan vaikkapa 45, 60, ..., 90, 100 ja 120, niin todelliset luokkarajat `todraj` saadaan määrätyksi esim. seuraavasti

```
todraj <- c(44.5, 59.5+10*(0:4), 119.5)
todraj    # tarkistetaan tulos
```

Seuraavaksi toteutetaan muuttujan `SYKEM1` luokittelu näillä todellisilla rajoilla antaen eri luokille kuitenkin pyöristettyjen rajojen mukaiset nimet (`labels`).

```
tp$sykeluok <- cut(tp$SYKEM1, br = todraj,
   labels = c('45-59', '60-69', '70-79', '80-89', '90-99', '100-119') )
attach(tp)
```

(b) Laske kunkin luokan absoluuttiset ja suhteelliset frekvenssit kuin myös kumulatiiviset absoluuttiset ja suhteelliset frekvenssit vektoreihin `m, pr, M` ja `Pr`. Tulosta nämä samaan taulukkoon vektoreita sarakkeittain yhteen liittävän funktion `cbind()` avulla.

```
m <- table(sykeluok) ; m
n <- sum(m) ; n
pr <- 100*m/n ; pr
M <- cumsum(m) ; M
Pr <- 100*M/n ; Pr
cbind( m, "p(%)" = round(pr,1), M, "P(%)"=round(Pr,1) )
```

(c) Absoluuttiset ja suhteelliset frekvenssit saadaan siististi taulukoiduksi myös R-paketista `Epi` löytyvää funktiota `stat.table()` hyödyntäen:

```
library(Epi)
stat.table(sykeluok,
      contents = list( "Lkm" = count(), "%" = percent(sykeluok) ),
       margins = T )
```

(d) Piirrä kumulatiivisten frekvenssien pohjalta luokitellun sykeaineiston summakäyrä.

```
rajat <- c(34.5, todraj, 129.5) # x-koordinaatit
Pros <- c(0, 0, Pr, 100)        # y-koordinaatit
plot( Pros ~ rajat, type = 'l', yaxs='i', # y-akseli alkaa 0:sta
      xlab = 'Leposyke (/min)', ylab = 'Kertymäosuus (%)' )
```

Täydennetään kuviota y-akselin neljänneksiin jakavilla hilaviivoilla.

```
abline( h= 25*(1:3), lty=2, col="gray" ) # hilaviivat
```

6. Leposykearvojen histogrammi

(a) Piirrä histogrammi funktiolla `hist()` nojautuen sen oletusarvoihin

```
hist(SYKEM1)
```

Oletusarvoinen luokittelu on siis tasavälinen, ja y-akselilla pylvään korkeudet kuvaavat luokkien absoluuttisia frekvenssejä, mutta luokittelu ei noudata edellisen tehtävän luokkarajoja.

(b) Piirrä uusi versio, jossa luokat ja pylväiden x-koordinaatit määrätään edellisen tehtävän todellisten luokkarajojen mukaisesti. Vertaa edelliseen; mikä on nyt y-akselin asteikko?

```
hist(SYKEM1, br = todraj,
     xlim= c(35, 125), xlab="Leposyke (per min)")
```

(c) Piirrä nyt histogrammi funktiolla `truehist()`, joka löytyy paketista `MASS`, nojautuen sen oletusarvoihin ja vertaa edellisiin.

```
library(MASS)
truehist(SYKEM1 )
```

(d) Piirrä myös tiheysestimaatti eli silotettu histogrammi, jonka taustalle harmain katkoviivoin tulee kohdassa (b) piirretty histogrammi.

```
plot(density(SYKEM1, na.rm=T))
hist(SYKEM1, br = todraj,  lty= 3, border="gray", add=T )
```

(e) Piirrä vielä kaikki edellisten kohtien (a)–(d) histogrammit samaan grafiikkaikkunaan, 2 alekkain ja 2 rinnakkain kirjoittamalla ensin

```
par(mfrow=c(2,2))
```

jonka jälkeen maalaa ja aja kaikkien neljän histogrammin piirroskomennot samanaikaisesti. Vertaa lopputuloksia. Lopuksi palauta grafiikkaikkuna perustilaan:

```
par(mfrow=c(1,1))
```

7. Datatiedostossa numeerisesti koodatun muuttujan määrittely luokkamuuttujaksi onnistuu käyttäen funktiota `factor()`, jonka argumentilla `labels` voi antaa myös luokille selväkieliset nimilaput (vrt. tehtävä 5.(a) edellä).

(a) Määritellään nyt muuttuja `LIIKUNTA` luokkamuuttujaksi eli tyypin `factor` muuttujaksi. Katsotaan sen jälkeen sen yksinkertaista frekvenssijakaumaa.

```
tp$liikunta <- factor(LIIKUNTA,
         labels = c('ei/harv', '1-3/kk', '1/vko',
                    '2-3/vko', '4-6/vko', 'päiv') )
table(tp$liikunta)
```

(b) Kaikkein alimmassa luokassa on vain yksi havainto. Luokitusta voisi ehkä tiivistää liittämällä kaksi alinta luokkaa yhteen. Tämä onnistuu esim. paketin `Epi` funktiolla `Relevel()`, minkä jälkeen voidaan laatia frekvenssitaulukko.

```
tp$liik5 <- Relevel( tp$liikunta, list( 1:2, 3,4,5,6 ) )
attach(tp)
stat.table( liik5,
        contents = list( 'Lkm' = count(), '%' = percent(liik5) ),
         margins = T )
```

(c) Verrataan liikuntaharrastuksen jakaumia miesten ja naisten välillä. Mitä eroja havaitset?

```
stat.table( index = list(liik5, sukup),
        contents = list( 'Lkm' = count(), '%' = percent(liik5) ),
         margins = T )
```

8. Jatkoa edelliseen (jos aikaa on vielä jäljellä, tai sitten omatoimisesti harjoituksen jälkeen tehtäväksi). Verrataan liikuntaharrastuksen intensiivisyyden jakaumia miesten ja naisten välillä graafisesti pylväs- ja palkkikuvioiden avulla käyttäen funktiota `barplot()`.

(a) Ensin kannattaa viimeksi tulostettu taulukko tallettaa omaan olioonsa

```
tab <- stat.table( index = list(liik5, sukup),
                contents = percent(liik5) )
```

Tarkastellaan taulukko-olion rakennetta

```
str(tab)
```

Se on rakenteeltaan periaatteessa 3-ulotteinen taulukko, jossa vain 2. ja 3. dimensio ovat tässä tapauksessa olennaiset. Muunnetaan taulukko siis kaksiulotteiseksi ottamalla vain jälkimmäiset dimensiot mukaan, minkä jälkeen tarkistetaan lopputulos.

```
tab <- tab[1 , , ] ; str(tab) ; tab
```

(b) Piirretään aluksi ositetut pylväskuviot. Ensiksi sellainen, jossa miesten kaikki pylväät ovat vierekkäin ja erillään naisten pylväiden rypäästä.

```
barplot(tab, beside=T, ylim=c(0,60) ) ; box()
```

(c) Sitten vaihtoehtoinen järjestys, jossa kunkin liikuntaluokan kohdalla on vierekkäin miesten ja naisten osuudet kummankin sukupuolen omasta jakaumasta. Tässä käytetään taulukon transponoivaa funktiota `t()`, joka vaihtaa rivit ja sarakkeet keskenään.

```
barplot(t(tab), beside=T, ylim=c(0,60) ) ; box()
```

(d) Kolmanneksi tarkastelemme miesten ja naisten osiin jaettuja pylväskuvioita rinnakkain

```
barplot(tab); box()
```

(e) Pystyssä olevien rinnakkaisten pylväiden asemesta alekkain asetetut palkit voivat olla havainnollisempia

```
barplot(tab, horiz=T) ; box()
```

(f) Tästä kehittyneempi versio, jossa palkkeja kavennetaan ja niiden väliin jätetään enemmmän tilaa (`space`) ja y-akselin tekstit kirjoitetaan horisontaalisesti (`las=1`)

```
barplot(tab, horiz=T, space=0.8, las=1, xlab="Osuus (%)") ; box()
```

(g) Kehitetään lopuksi kuvaa vielä niin, että väljennetään kuvion reuna-alueita y-akselin suunnassa ja määritellään palkkien osille toivotut harmauden asteet itse skaalalla 0:sta (musta) 100:aan (valkoinen)

```
barplot(tab, horiz=T, space=0.8, las=1, ylim=c(0, 4.5), yaxs = "i",
     col=c("gray55", "gray65", "gray75", "gray82", "gray90") )
rect(0,0,100,4.5)
```

Kuvaa voi tarvittaessa edelleen rikastaa monin tavoin, mm. lisäämällä sen sisään tekstinpätkiä, lukuja ym. merkkijonoja haluttuihin koordinaattipisteisiin.

Oulun yliopiston matemaattisten tieteiden laitos/tilastotiede

806113P TILASTOTIETEEN PERUSTEET, kl 2011 (Esa Läärä)

M-harjoitus 3, viikot 7-8 (18.-23.2.): mikroluokkatehtävät

Kuten edellisissäkin M-harjoituksissa R:n käynnistämisen jälkeen vaihda työhakemistoksi oman Samba-hakemistosi se alihakemisto, jonka olet luonut tätä kurssia varten. Avaa sen jälkeen `script editor` -ikkuna, johon kirjoitat istunnon aikana tarvittavat komennot, ja tallenna nämä tiedostoon esim. nimellä `tp-mh3.R`.

1. Jatkoa L-harjoituksen 5 tehtävään 1. Hannu Hanhi heitti omaa noppaansa $n = 30$ kertaa, ja näistä $m = 2$ kertaa silmäluvuksi tuli yksi. Kohdeparametri on

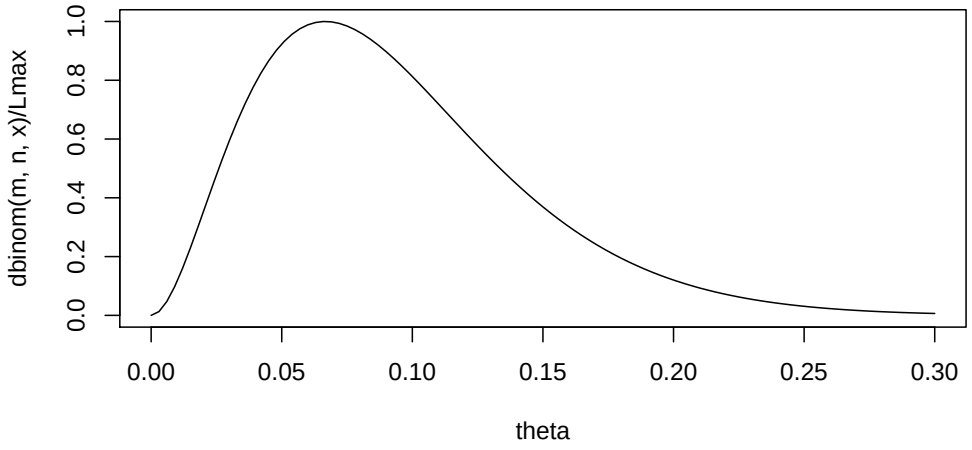
$$\theta = \mathbb{P}(\text{“kerran noppaa heitettäessä silmäluvuksi tulee yksi”}).$$

(a) Laske parametrin θ suurimman uskottavuuden estimaatti $\widehat{\theta} = m/n$ ja sijoita se muuttujaan `that`. Laske θ:n uskottavuusfunktion $L(\theta)$ arvo tässä pisteessä

$$L_{\max} = L(\widehat{\theta}) = \mathbb{P}(M = m; \widehat{\theta}) = \binom{n}{m} \widehat{\theta}^m (1 - \widehat{\theta})^{n-m}$$

käyttäen R-funktiota `dbinom()` ja sijoita se muuttujaan `Lmax`. Piirrä tämän jälkeen jälkeen θ:n suhteellisen uskottavuusfunktion kuvaaja, joka siis näyttää kuinka osamäärä $L(\theta)/L(\widehat{\theta})$ riippuu θ:sta. Pysähdy tarkastelemaan uskottavuusfunktion kuvaajaa. Mitä informaatiota se antaa?

```
> n <- 30
> m <- 2
> that <- m/n
> Lmax <- dbinom(m, n, that)
> round(c(that, Lmax), 4)
> curve(dbinom(m, n, x)/Lmax, from = 0, to = 0.3, xlab = "theta")
```



(b) Jos noppa on reilu, niin voidaan odottaa, että $\theta = 1/6 \approx 0.167$. Pidetään tätä nollahypoteesina. Laske, kuinka suuri on tämän arvon suhteellinen uskottavuus $\text{LR} = L(1/6)/L(\widehat{\theta})$

Hannun heittotulosten nojalla. Laske ja tulosta myös tämän nollahypoteesin testaamisessa käytettävän testisuureen Z arvo samoin kuin sen neliö Z^2 ja arvioi normaalijakauman kertymäfunktiota hyväksi käyttäen vastaava likimääräinen P-arvo. (Funktio `abs()` laskee argumenttinsa itseisarvon.)

```
> theta0 <- 1/6
> LR <- dbinom(m, n, theta0)/dbinom(m, n, that)
> Zhav <- (that - theta0)/sqrt(theta0 * (1 - theta0)/n)
> P <- 2 * (1 - pnorm(abs(Zhav)))
> round(c(LR, Zhav, P, Z2 = Zhav^2), 4)
```

(c) Laske seuraavaksi 95% likimääräinen luottamusväli parametrille θ käyttäen AC-menetelmää.

```
> that.ac <- (m + 2)/(n + 4)
> SE.ac <- sqrt(that.ac * (1 - that.ac)/(n + 4))
> CI.ac <- that.ac + c(-1.96, 1.96) * SE.ac
> round(CI.ac, 4)
```

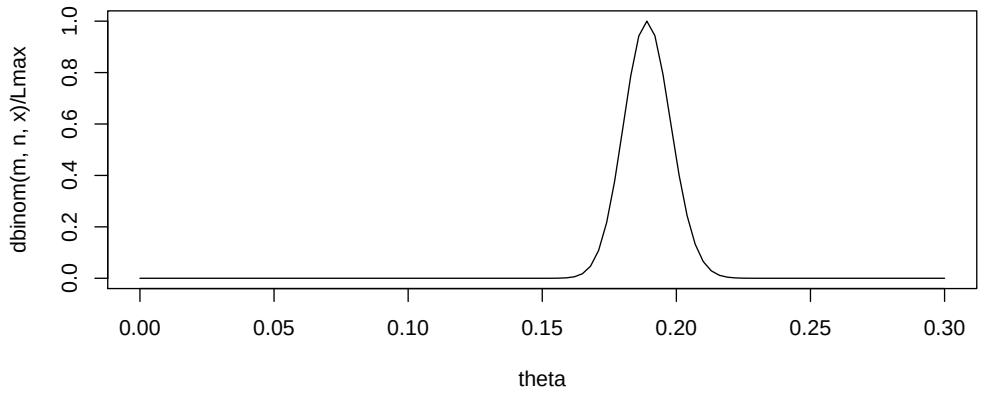
(d) Toteuta lopuksi testaus ja luottamusvälin laskenta R-funktiolla `prop.test()` ja tutki saamaasi tulostusta

```
> prop.test(m, n, 1/6, correct = F)

        1-sample proportions test without continuity correction

data:  m out of n, null probability 1/6
X-squared = 2.16, df = 1, p-value = 0.1416
alternative hypothesis: true p is not equal to 0.1666667
95 percent confidence interval:
 0.01847702 0.21323458
sample estimates:
         p
0.06666667
```

Tämän funktion raportoiman khiin neliö -suureen arvo 2.16 on siis täsmälleen sama kuin edellä laskemamme Z-suureen -1.47 neliö, joten myös $P = 0.1416$ on sama. Sen sijaan luottamusvälin rajat poikkeavat hieman edellä lasketuista, koska `prop.test()` soveltaa AC-menetelmän asemesta Wilsonin menetelmää (ks. L-harjoitus 5 tehtävä 4).

2. Jatkoa L-harjoituksen 5 tehtävään 2. Taloustutkimuksen puoluekannatusmittauksessa 1/2011 raportoitiin SDP:n kannatusosuudeksi 18.9% niiden 2000 henkilön joukossa, jotka ilmoittivat kannattavansa jotakin puoluetta. Vuoden 2007 eduskuntavaaleissa SDP sai kaikista äänistä 21.4%. Ota mallia edellisen tehtävän R-komennoista ja sovella niitä seuraavien laskelmien tekemiseen – muistaen kuitenkin, että SDP:tä äänestävien lukumäärä m pitää ensin laskea otoskoosta n ja SDP:n kannatusosuudesta $\widehat{\theta}$.

(a) Piirrä ensin tuntemattoman kannatusosuuden suhteellisen uskottavuusfunktion kuvaaja samalla välillä $[0, 0.3]$. Vertaa tätä kuvaajaa tehtävän 1 vastaavaan. Mistä arvelet pääasiassa johtuvan sen, että tämä kuvaaja on edellistä huomattavasti kapeampi?



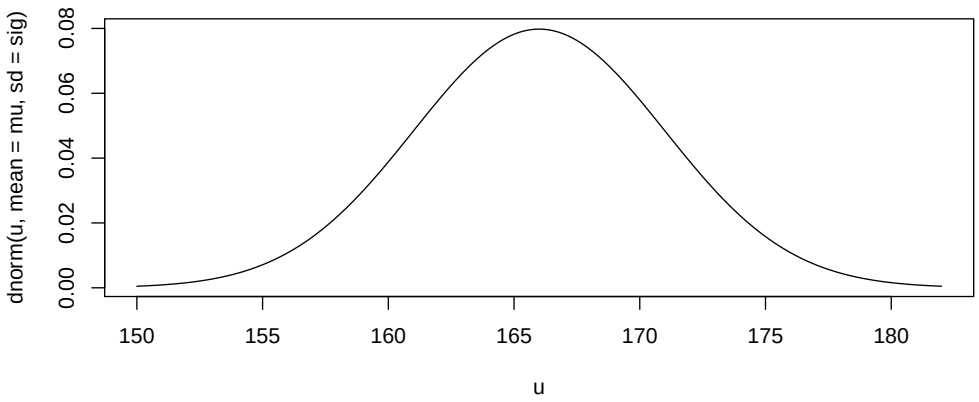
(b) Käyttäen funktiota `prop.test()` (a) laske SDP:n kannatusosuuden 95% luottamusväli tuoreimman kannatusmittauksen tulosten pohjalta ja (b) testaa nollahypoteesia, jonka mukaan SDP:n kannatusosuus olisi muuttunut siitä mikä se oli vuoden 2007 eduskuntavaaleissa, ja arvioi vastaava P-arvo. Vertaa tuloksia L-harjoituksessa saatuihin.

Seuraavissa tehtävissä tutkimme simuloiden eli Monte Carlo -menetelmällä sitä, kuinka normaalijakaumasta poimittavan otoksen tunnuslukujen otantajakaumat käyttäytyvät. Havainnollistamme, kuinka otoskeskiarvo $\bar{Y}$, otoskeskihajonta S, testisuure T, vastaava P-arvo sekä 90% luottamusvälin ala- ja ylärajat vaihtelisivat mahdollisesta otoksesta toiseen tilanteessa, jossa tunnemme kohdemuuttujan Y jakauman, sen odotusarvon μ ja varianssin σ^2 siinä populaatiossa, josta otoksia poimittaisiin.

3. Oletetaan, että suomalaisten naispuolisten korkeakouluopiskelijoiden populaatiossa pituus Y noudattaa normaalijakaumaa odotusarvolla $\mu = 166$ cm. Lisäksi oletamme Y:n jakauman varianssin olevan tunnettu: $\sigma^2 = 5^2$ cm^2, eli keskihajonta on $\sigma = 5$ cm.

Piirrä jakauman $N(166, 5^2)$ tiheysfunktion kuvaaja vaihteluvälille [150, 182] cm:

```
> u <- seq(150, 182, by = 0.1)
> mu = 166
> sig = 5
> plot(u, dnorm(u, mean = mu, sd = sig), type = "l")
```



Vektori `u` sisältää hilan mahdollisia pituusarvoja 0.1 cm välein: $u_1 = 150, u_2 = 150.1, \ldots, u_{321} = 182$. Komento `plot()` piirtää murtoviivan (`type = 'l'` joka on "äll" eli "*line*" eikä "yksi") pisteiden $(u_i,\ \frac{1}{5}\varphi[(u_i - 167)/5])$ kautta, jossa $\varphi(z)$ on $N(0,1)$-jakauman tiheysfunktio.

4. Jatkoa edelliseen tehtävään. Simuloimme nyt satunnaisotantaa naisopiskelijoiden pituuden oletetusta populaatiojakaumasta.

(a) Poimi $n = 4$ havainnon satunnaisotos $y_1, \ldots, y_4$ jakaumasta $N(166, 5^2)$ funktiolla `rnorm()` ja sijoita vektoriin `otos`. Piirrä otosarvojen pistekuvio sekä laske ja tulosta tästä otoksesta keskiarvo $\bar{y}$, keskihajonta s_y ja keskiarvon keskivirhe $\mathrm{SE}(\bar{y}) = s_y/\sqrt{n}$:

```
> n <- 4
> otos <- rnorm(n, mu, sig)
> stripchart(otos, xlim = c(150, 182))
> round(c(mean(otos), sd(otos), sd(otos)/sqrt(n)), 2)
```

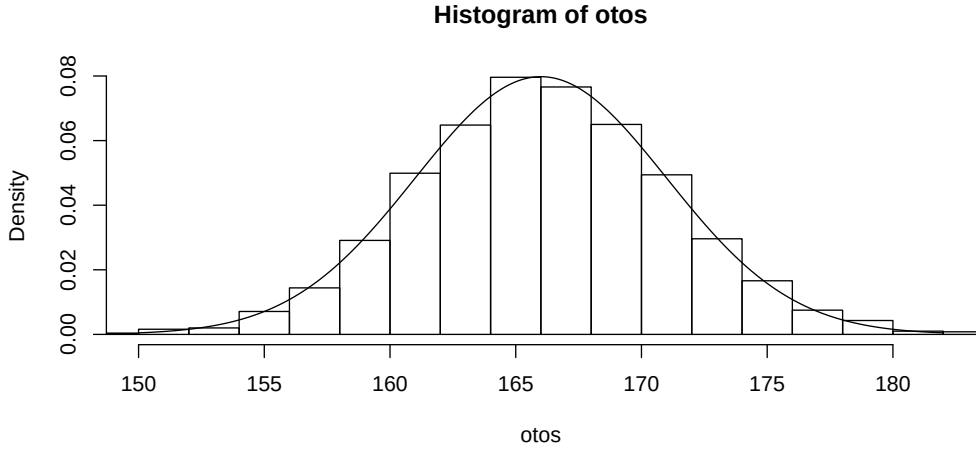
Mitä havaintoja teet otoskeskiarvon ja -hajonnan arvojen poikkeamista teoreettisiin arvoihin $\mu = 166$ cm, $\sigma = 5$ cm ja $\sigma/\sqrt{4} = 2.5$ cm verrattuna?

(b) Poimi uusi otos kooltaan $n = 4$, piirrä pistekuvio ja laske tunnusluvut kuten edellä sekä vertaile tuloksia näiden kahden otoksen välillä.

(c) Tee sama uudelleen kaksi kertaa mutta suuremmalla otoskoolla eli $n = 25$. Mitä nyt havaitset? Miten esim. otosarvojen vaihteluväli muuttuu pieniin otoksiin verrattuna?

(d) Poimi nyt suuri otos, jossa $n = 5000$, ja laske tunnusluvut kuten edellä. Piirrä tämän otoksen arvoista histogrammi välille [150, 182] 2 cm luokkaleveyksin. Piirrä histogrammin päälle oletetun populaatiojakauman eli $N(166, 5^2)$-jakauman tiheysfunktion kuvaaja kuten tehtävässä 3. mutta nyt komennolla `lines()`, jolloin ei tarvita argumenttia `type`. Mitä havaitset?

```
> n <- 5000
> otos <- rnorm(n, mu, sig)
> hist(otos, freq = F, br = seq(130, 210, by = 2), xlim = c(150,
+     182))
> lines(u, dnorm(u, mu, sig))
```



(Argumentilla `br` määrätään pylväiden leveydet ja varaudutaan laajaankin vaihteluväliin otosarvoissa, mutta itse kuvioon säädetään kapeampi väli argumentilla `xlim`.)

5. Jatkamme satunnaisotannan simulaatiotutkimuksia ja tarkastelemme nyt, miten keskeiset otostunnusluvut käyttäytyvät toistettaessa otantaa monta kertaa. Käytämme funktiota `normotos.sim()`, joka on FM Timo Knürrin alun perin laatima. Tällä funktiolla on seuraavat argumentit

- `n` = otoskoko n yksittäisessä otoksessa,
- `mu` = populaatiojakauman odotusarvo μ, ja `sig` = jakauman keskihajonta σ,
- `mu0` = nollahypoteesin $H_0 : \mu = \mu_0$ olettama odotusarvo,
- `level` = luottamustaso, oletusarvona `0.9` eli 90%,
- `nsim` = simuloitavien otosten lukumäärä, oletusarvona `nsim=20`,
- `kuva` = looginen muuttuja oletusarvona `TRUE`, jolloin otoksista piirretään graafinen esitys,
- `loc` = oletusarvona `TRUE`, jolloin käytetään graafisessa esityksessä interaktiivista `locator()`-funktiota, joka mahdollistaa kuvaan tulevien alkioiden piirtämisen otos kerrallaan.

Funktio tuottaa tuloksenaan datakehikon, joka sisältää kustakin otoksesta lasketut tunnuslukujen arvot. Kun `kuva = T`, se myös piirtää samaan kuvaan kaikkien otosten havainnot ja niistä lasketut luottamusvälit odotusarvolle μ.

(a) Lataa luennoitsijan laatimien funktioiden kokoelma R-istuntoosi ja säädä tulostuksen desimaalitarkkuus:

```
> source("P:/TP2011/Esanfunktiot.R")
> options(digits=4)
```

(b) Poimi 20 kpl kooltaan $n = 4$ suuruisia otoksia naisten pituusjakaumasta $N(166, 5^2)$ ja sijoita tulokset datakehikkoon `otos4`:

```
> otos4 <- normotos.sim(4, mu, sig, level = 0.95)
```

Siirry grafiikkaikkunaan ja klikkaa hiiren vasemmanpuoleista näppäintä, jolloin 1. otoksen havaintojen pistekuvio ilmestyy koordinaatistoon. Klikkaa toisen kerran, jolloin vasemmalle reunalle tulostuvat otoskeskiarvo ja -hajonta, ja lisäksi kuvioon ilmestyy μ:n luottamusväli.

Jatka klikkaamista rauhalliseen tahtiin ja seuraa, kuinka otosarvot, tunnusluvut ja luottamusväli vaihtelevat otoksesta toiseen, kunnes kaikkien 20 otoksen tulokset ovat näkyvillä. Mitä havaintoja teet? Kuinka moni luottamusväli ei peittänyt μ:tä?

(c) Listaa datakehikon `otos4` sisältö ja tulosta sen muuttujien suorat jakaumat. Mitä havaintoja teet eri otostunnuslukujen vaihteluvälien suuruuksista?

```
> otos4
> summary(otos4)
```

(d) Toista kohdat(b)–(c) mutta nyt käyttäen otoskokoa $n = 25$

```
> otos25 <- normotos.sim(25, mu, sig, level = 0.95)
> summary(otos25)
```

Vertaile luottamusvälien ja muiden otostunnuslukujen vaihtelua kohtien (b)-(c) tuloksiin. Mitkä ovat keskeiset havaintosi tästä vertailusta?

6. Simuloidaan nyt peräti 10000 samankokoista otosta naisopiskelijoiden pituuden mallista $N(166, 5^2)$ ja tarkastellaan otostunnuslukujen jakautumista.

(a) Toteuta simulaatio otoskoolla $n = 4$, talleta datakehikkoon, kiinnitä ja tulosta tunnuslukujen suorien jakaumien tunnusluvut:

```
> otos4.10k <- normotos.sim(4, mu, sig, nsim = 10000, level = 0.95,
+     kuva = F, loc = F)
> attach(otos4.10k)
> summary(otos4.10k)
```

Mitä huomioita teet? Tutki erityisesti, kuinka lähellä otosvarianssien ja otoshajontojen keskiarvot ovat teoreettisia arvoja $\sigma^2 = 25$ ja $\sigma = 5$. Mitä havaitset?

(b) Piirrä simuloiduista otoksista laskettujen *otoskeskiarvojen* $\bar{y}$ histogrammi 1 cm luokkavälein. Piirrä samaan kuvioon otoskeskiarvon $\bar{Y}$ teoreettisen otantajakauman $N(\mu, \sigma^2/n)$ kuin myös alkuperäisen muuttujan Y jakauman $N(\mu, \sigma^2)$ tiheysfunktioiden kuvaajat

```
> hist(keskiarvo, freq = F, br = 150:180)
> lines(u, dnorm(u, mu, sig/sqrt(4)))
> lines(u, dnorm(u, mu, sig), lty = 3)
```

Mitä havaintoja teet simuloitujen otoskeskiarvojen jakautumisesta suhteessa teoreettiseen otantajakaumaansa?

(c) Piirrä myös *otoskeskihajontojen* s histogrammi. Onko sen otantajakauma symmetrinen vai vino?

```
> hist(hajonta, freq = F)
> detach(otos4.10k)
```

(d) Toteuta kohdat (a)-(c) uudelleen mutta olkoon nyt yksittäisen otoksen koko $n = 25$. Millainen on keskiarvojen otantajakauma tällä otoskoolla verrattuna tilanteeseen, jossa $n = 4$? Entä keskihajontojen otantajakauma; onko symmetrisempi?

7. Jatkoa edelliseen tehtävään. Tarkastelemme seuraavaksi luottamusvälien ja testitunnuslukujen käyttäytymistä simuloidusta otoksesta toiseen, kun nollahypoteesina oletetaan $H_0 : \mu = 166$ ja otoskoko on $n = 25$.

(a) Simuloiduista otoksista laskettujen 95% luottamusvälien $\bar{Y} \pm t_{0.975}(24) \times \mathrm{SE}(\bar{Y})$ ala- ja ylärajat on talletettu datakehikon muuttujiin `mu.alar` ja `mu.ylar`. Laske, kuinka moni alempi luottamusraja on suurempi kuin odotusarvo μ. Laske vastaavasti, kuinka moni yläraja on pienempi kuin μ. Mikä on siten μ:n peittävien luottamusvälien osuus kaikista simuloiduista otoksista?

```
> alaryli <- length(mu.alar[mu.alar > mu])
> ylarali <- length(mu.ylar[mu.ylar < mu])
> peitto <- 1 - (alaryli + ylarali)/10000
> c(alaryli, ylarali, peitto)
```

Peitto-osuuden pitäisi olla lähellä arvoa 0.95.

(b) Datakehikon sarake `T.suure` sisältää otoksista lasketut arvot testisuureelle $T = (\bar{Y} - 166)/\mathrm{SE}(\bar{Y})$, ja sarake `P.arvo` vastaavat 2-tahoiset P-arvot. Piirrä histogrammi simuloitujen otosten T-arvojen jakautumisesta välille [-4, 4] luokkavälein 0.2

```
> hist(T.suure, freq = F, br = seq(-50, 50, by = 0.2), xlim = c(-4,
+     4))
```

Piirrä samaan kuvioon $N(0, 1)$-jakauman tiheysfunktion kuvaaja punaisella värillä

```
> tval <- seq(-4, 4, by = 0.1)
> lines(tval, dnorm(tval), col = "red")
```

Kuinka hyvin standardinormaalijakauma kuvaa T:n otantajakaumaa tällä vapausasteluvulla? Edelleen piirrä vapausastein $n-1$ Studentin jakauman tiheysfunktion kuvaaja sinisellä

```
> lines(tval, dt(tval, df = 25 - 1), col = "blue")
```

Kuinka hyvin tämä otantajakauma kuvaa simuloitujen T-arvojen jakaumaa? Kumpi jakaumista, normaali- vai Student, on paremmin yhteensopiva simuloitujen T-arvojen jakauman kanssa?

(c) Piirrä simuloitujen P-arvojen histogrammi. Laske, kuinka suuri osa niistä on pienempiä kuin 0.05. Piirrä myös P-arvojen otoskertymäfunktio

```
> hist(P.arvo, freq = F)
> sum(P.arvo < 0.05)/10000
> plot(ecdf(P.arvo))
```

Mitä päättelet suureen P otantajakaumasta H_0:n vallitessa?

8. Tarkastellaan lopuksi tilannetta, jossa pituuden todellinen odotusarvo kohdepopulaatiossa onkin $\mu = 167$ cm, mutta nollahypoteesina oletetaan edelleen $H_0 : \mu = \mu_0 = 166$ cm. Millaiset ovat T-testisuureen ja P-arvon otantajakaumat nyt?

```
> detach(otos25.10k)
> otos25.10k <- normotos.sim(25, 167, sig, mu0 = 166, nsim = 10000,
+     level = 0.95, kuva = F, loc = F)
> attach(otos25.10k)
> summary(otos25.10k)
> hist(T.suure, freq = F, br = seq(-50, 50, by = 0.2), xlim = c(-4,
+     4))
> lines(tval, dt(tval, df = 25 - 1), col = "blue")
> hist(P.arvo, freq = F)
> sum(P.arvo < 0.05)/10000
```

T-suureen otantajakauma ei enää olekaan keskittynyt 0:n ympärille vaan on siirtynyt siihen suuntaan, mikä on μ:n oikea arvo verrattuna oletettuun μ_0:n arvoon. Myöskään P-arvojen jakauma ei ole enää tasainen, vaan on todennäköisempää saada pieniä P-arvoja kuin suuria. Kuitenkaan todennäköisyys saada $P < 0.05$ ei sentään ole 95% vaan paljon alhaisempi.

Jos aikaa on, voit vielä simuloida saman määrän T- ja P-arvoja asetelmassa, jossa oikea odotusarvo on $\mu = 168$ mutta nollahypoteesi on edelleen $H_0 : \mu = 166$, todetaksesi että testisuureen jakauma on siirtynyt vielä enemmän oikealle 0:sta, ja että nyt on entistä todennäköisempää saada pieniä P-arvoja. Se, millaisia tuloksia näiden tunnuslukujen otantajakaumista on siis odotettavissa, riippuu siis vahvasti siitä, mikä on parametrin μ arvo siinä tilanteessa kun se ei ole H_0:n mukainen.

Oulun yliopiston matemaattisten tieteiden laitos/tilastotiede

806113P TILASTOTIETEEN PERUSTEET, kl 2011 (Esa Läärä)

M-harjoitus 4, viikko 9-10 (4.-9.3.): tehtävät

1. Jatkoa L-harjoituksen 6 tehtävään 1. Ohessa vielä nastanheitossa saadut yhdistetyt tulokset eri harjoitusryhmissä:

Ryhmä	Heittäjiä	Heittoja	Jäi selälleen
1 (Hanna)	7	175	84
2 (Hanna)	9	225	131
3 (Päivi)	15	375	219
Yhteensä	31	775	434

Merkitään θ_k = "todennäköisyys, että nasta jää selälleen yksittäisessä heitossa harjoitusryhmässä k", jossa $k = 1, 2, 3$.

Käyttäen R-paketin `Epi` sisältämää funktiota `twoby2()` laske seuraavat tunnusluvut:

(a) Todennäköisyyksien θ_k:n piste-estimaatit ja likimääräiset 95% luottamusvälit erikseen kahdessa ensimmäisessä ryhmässä; ts. $k = 1, 2$

(b) Vertailuparametrin $\delta = \theta_1 - \theta_2$ piste-estimaatti ja likimääräinen 95% luottamusväli.

(c) Nollahypoteesia $H_0 : \delta = 0$ koskeva likimääräinen P-arvo.

Funktiolle `twoby2()` annetaan syötteenä 2×2-taulukko tai matriisi, joka sisältää riveittäin kummassakin ryhmässä k ($k = 1, 2$) "onnistumisten" lukumäärät m_k ja "epäonnistumisten" lukumäärät $n_k - m_k$.

```
> library(Epi)
> m1 <- 84
> n1 <- 175
> m2 <- 131
> n2 <- 225
> taulu <- matrix(c(m1, n1 - m1, m2, n2 - m2), nrow = 2, byrow = T)
> taulu
> twoby2(taulu)
```

Pysähdy tarkastelemaan tuloksia ja vertaa niitä L-harjoituksissa saatuihin. Pienet erot johtuvat siitä, että `twoby2()` käyttää yksinkertaista menetelmää tarkempia approksimaatiokaavoja.

2. Jatkoa L-harjoituksen 6 tehtävään 4. Keski-ikäisten miesten ryhmälle ($n = 16$) tehtiin rasitustesti. Miesten verenpaineet mitattiin sekä ennen rasitusta että rasituksen jälkeen. Systolisen verenpaineen (mmHg) mittaustulokset olivat henkilöittäin seuraavat:

Henkilö	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Ennen	148	142	136	134	138	140	132	144	128	170	162	150	138	154	126	116
Jälkeen	152	152	134	148	144	136	144	150	146	174	162	162	146	156	132	126
Erotus	+4	+10	−2	+14	+6	−4	+12	+6	+18	+4	0	+12	+8	+2	+6	+10

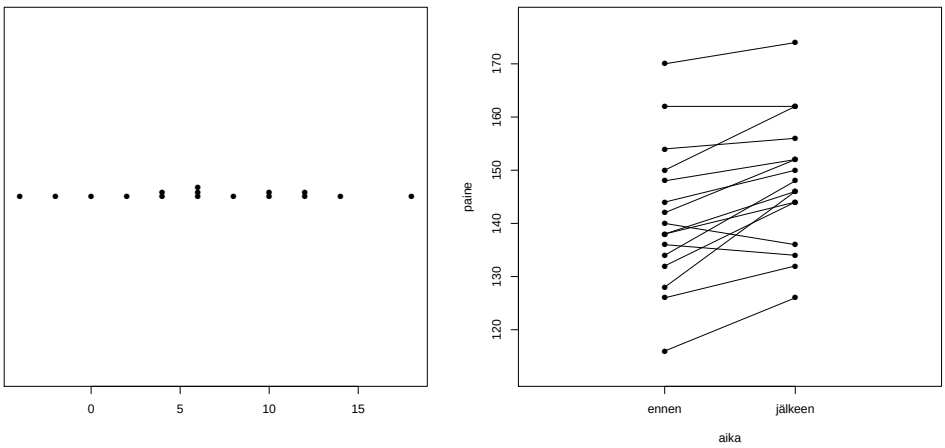
Nämä luvut löytyvät henkilöittäin tiedostosta `P:\TP2011\paineet.txt`, josta ne on suoraan luettavissa R:n datakehikoksi.

(a) Lue tiedosto datakehikoksi, listaa, laske muuttujien keskiarvot ja keskihajonnat sekä kiinnitä kehikko.

```
> rasi <- read.table('P:/TP2011/paineet.txt', header=T)
> rasi
> mean(rasi) ; sd(rasi)
> attach(rasi)
```

(b) Piirrä rinnakkain kaksi pistekuviota: (1) yksinkertainen horisontaalinen pistekuvio havaituista erotuksista ja (2) vertikaalinen pistekuvio, jossa vasemmalla puolella ovat `ennen`-havainnot ja oikealla puolella `jalkeen`-havainnot siten, että saman henkilön havaintopisteiden välille on piirretty yhdysjana. Jälkimmäistä kuvaa varten kaikki mittaustulokset on hyvä sijoittaa samaan vektoriin `paine`, joista 16 ensimmäistä muodostaa `ennen`-mittausten lohkon ja 16 viimeistä `jälkeen`-lohkon siten, että kummassakin lohkossa henkilöiden järjestys on sama, ja lohkoille annetaan koodit 1 ja 2.

```
> par(mfrow = c(1, 2))
> stripchart(erotus, method = "stack", pch = 16)
> paine <- c(ennen, jalkeen)
> aika <- c(rep(1, 16), rep(2, 16))
> plot(paine ~ aika, xlim = c(0, 3), pch = 16, axes = F, ylim = c(112,
+     178))
> axis(1, at = c(1, 2), labels = c("ennen", "jälkeen"))
> axis(2, at = 10 * (12:17))
> box()
> segments(aika[1:16], paine[1:16], aika[17:32], paine[17:32])
```



Vertaile näitä kuvioesityksiä. Kumpi on mielestäsi informatiivisempi?

(c) Oletetaan, että tällaisessa asetelmassa rasituksen aiheuttamat systolisen verenpaineen muutokset noudattavat ao. kohdepopulaatiossa jakaumaa, jonka odotusarvo on Δ ja varianssi τ^2 ovat tuntemattomat. Käyttäen funktiota `t.test()` laske 95 % luottamusväli parametrille Δ sekä nollahypoteesia $H_0 : \Delta = 0$. koskeva testisuureen arvo ja P-arvo.

```
> t.test(erotus)
```

3. Jatkoa edelliseen tehtävään. Tarkastellaan `ennen`- ja `jalkeen`-mittausten välistä lineaarista riippuvuutta.

(a) Piirrä sirontakuvio, jossa x-koordinaattina on `ennen` ja y-koordinaattina `jalkeen`

```
> par(mfrow = c(1, 1))
> plot(jalkeen ~ ennen, xlim = c(110, 180), ylim = c(110, 180),
+     pch = 16)
```

(b) Laske korrelaatiokerroin muuttujien `ennen` ja `jalkeen` välillä. Laske myös funktiolla `lm()` (*"linear models"*) regressiosuoran kulmakerroin ja vakiokerroin sekä näiden 95% luottamusvälit mallille, jossa Y-muuttujana on `jalkeen` ja X-muuttujana `ennen`. Piirrä myös uudelleen sirontakuvio, joka lisäksi sisältää sovitetun regressiosuoran.

```
> cor(ennen, jalkeen)
> malli <- lm(jalkeen ~ ennen)
> cbind(coef(malli), confint(malli))
> plot(jalkeen ~ ennen, xlim = c(110, 180), ylim = c(110, 180),
+     pch = 16)
> abline(malli)
```

Mitä havaintoja teet ennen rasitusta ja rasituksen jälkeen mitattujen verenpaineiden keskinäisestä riippuvuudesta?

4. Jatkoa L-harjoituksen 7 tehtävään 1. Analysoidaan sykekokeen tuloksia eli vertaillaan loppusykkeen jakaumia koe- ja vertailuryhmän välillä.

(a) Lue tiedosto `P:\TP2011\sykkeet.txt` datakehikoksi `syk`, listaa ja kiinnitä:

```
> syk <- read.table('P:/TP2011/sykkeet.txt', header=T)
> syk
> attach(syk)
```

(b) Piirrä havainnot ryhmittäin alekkaisiin pistekuvioihin.

```
> stripchart(loppusyke ~ ryhma, method = "stack", pch = 16, ylim = c(0,
+     3))
```

(c) Laske loppusykkeen keskiarvot ryhmittäin sekä piste-estimaatti loppusykkeen odotusarvojen erotukselle käsittelyjen välillä. Laske myös keskihajonnat kummassakin ryhmässä

```
> means <- tapply(loppusyke, ryhma, mean)
> round(means, 1)
> round(means[1] - means[2], 1)
> round(tapply(loppusyke, ryhma, sd), 1)
```

(d) Laske nollahypoteesia "ei eroa odotusarvoissa" koskeva testisuureen arvo ja sitä vastaava 2-tahoinen P-arvo sekä 95% luottamusväli loppusykkeiden odotusarvojen erotukselle käsittelyjen välillä.

```
> t.test(loppusyke ~ ryhma, var.equal = T)
```

Vertaa saamiasi tuloksia L-harjoituksissa saatuihin.

5. Jatkoa L-harjoituksen 7 tehtävään 3. Analysoidaan älykkyysosamäärän IQ riippuvuutta aivojen kokoa edustavasta muuttujasta MRI. Aineisto on tiedostossa `P:\TP2011\aivot.txt`.

(a) Lue aineisto datakehikkoon `iq` ja listaa. Luo uusi datakehikko `miq`, johon valitaan koko aineistosta vain miesten (`suku=0`) havainnot, ja samalla jätetään sukupuolta kuvaava muuttuja pois.

```
> iq <- read.table("P:/TP2011/aivot.txt", header=T)
> iq
> miq <- subset(iq, suku == 0 )[, -2]
> attach(miq)
```

(b) Piirrä vierekkäin sirontakuviot, joissa x-koordinaattina ovat vuorollaan `MRI` ja `pituus` ja y-koordinaattina `IQ`.

```
> par(mfrow = c(1, 2))
> plot(IQ ~ MRI, pch = 16)
> plot(IQ ~ pituus, pch = 16)
```

(c) Laske datakehikon sisältämien muuttujien `IQ`, `MRI` ja `pituus` väliset parittaiset korrelaatiokertoimet samanaikaisesi. Mitä havaintoja teet?

```
> round(cor(miq), 3)
```

(d) Sovita regressiomalli, jossa X-muuttujana on `pituus` ja Y-muuttujana `IQ`. Tulosta regressiokertoimet, niiden keskivirheet ja regressiokertoimien 95% luottamusvälit.

```
> iqpit <- lm(IQ ~ pituus)
> round(cbind(summary(iqpit)$coef, confint(iqpit)), 2)
> summary(iqpit)$sigma
```

Mitä päättelet tuloksista. Onko älykkyysosamäärä kääntäen verrannollinen pituuteen?