

Kvalitatiivsete andmete analüüs  
1. praktikum  
Meenutame antud aine jaoks olulisi jaotuseid ja R'i kasutamist

Antud aine praktikumides kasutame R tarkvara. Kellel varasemat kogemust R-iga pole, see võiks esmalt tutvuda järgneva antud tarkvara tutvustava materjaliga:

<http://www-1.ms.ut.ee/mart/SissejuhatusRi.pdf>

Kvalitatiivsete andmete analüüsi käigus mängivad olulist rolli binoomjaotus, Poissoni jaotus ja negatiivne binoomjaotus (lisaks veel mõni jaotus, nagu näiteks multinomiaalne jaotus – temaga tegeleme aga veidi hiljem). Järgnevalt proovime meenutada, millised need jaotused välja näevad ja kuidas käituvad. Joonistame nende tähtsate jaotuste tõenäosusfunktsioonid – praegu nii, et kõigi jaotuste korral juhusliku suuruse keskväärtus oleks 4:

```
x=0:20
p_binom = dbinom(x, 20, 0.2)
p_Pois = dpois(x, 4)
p_nb = dnbinom(x, prob=1-4/7, size=3)

names(p_binom)=names(p_Pois)=names(p_nb)=x

par(mfrow=c(2,2))
barplot(p_binom, main="binoomjaotus B(n=20; p=0.2)")
barplot(p_Pois, main="Poissoni jaotus")
barplot(p_nb, main="Negatiivne binoomjaotus")
```

Meeles tuleb pidada, et eksisteerib erinevaid binoomjaotuseid ja erinevaid negatiivseid binoomjaotuseid mille puhul keskväärtus on võrdne 4-ga. Kuna binoomjaotuse  $X \sim B(n; p)$  keskväärtus on  $E(X) = np$ , siis nii binoomjaotuse  $B(20; 0,2)$  ja  $B(8; 0,5)$  korral on juhusliku suuruse keskväärtuseks 4. Näiteks võime eelmisele joonisele lisada veel ühe binoomjaotuse, mille keskväärtus on ka 4:

```
p_binom2=dbinom(x, 8, 0.5)
names(p_binom2)=x
barplot(p_binom2, main="binoomjaotus B(8;0,5)")
```

Kui soovime oma binoomjaotusega/Poissoni jaotusega või negatiivse binoomjaotusega juhuslike suuruste jaotust iseloomustada, siis võiksime ju joonistada jaotuse iseloomustamiseks mingi graafiku. Mille poolest erinevad alltoodud kolm graafikut teineteisest:

```
# Tekitame „valimi“. Algväärtustame juhuslike suuruste generaatori - siis tekivad kõigil
# samasugused andmed ja joonised.
```

```
set.seed(1)
Y=rpois(35, 4)
```

```
# Joonis 1
hist(Y)

# Joonis 2
barplot(table(Y))

# Joonis 3
barplot(table(factor(Y, levels=0:20)))
```

Kolmekümne vaatluse pealt võib olla raske näha, millise jaotusega on tegemist. Võrdluseks sama joonis 1000 juhusliku suuruse pealt (lisaks võib jaotuse ära tundmisel olla abi keskmise ja dispersiooni võrdlemisest):

```
set.seed(1)
Y=rpois(1000, 4)
barplot(table(factor(Y, levels=0:20)))
mean(Y)
var(Y)
```

### Ülesanne 1.

Vaata järgmiseid juhuslikke suuruseid (Y) ja otsusta, kas tegemist võiks olla binoomjaotusega, Poissoni jaotusega või negatiivse binoomjaotusega (võrdle saadud tulpdiagrammi varem joonistatud teoreetiliste jaotustega).

```
# Variant A
windows()
lambda=rgamma(10000, shape=4)
Y=rpois(10000, lambda=lambda)
barplot(table(factor(Y, levels=0:20)))

# Variant B
windows()
Y1=rpois(10000, lambda=1)
Y2=rpois(10000, lambda=3)
Y=Y1+Y2
barplot(table(factor(Y, levels=0:20)))

# Variant C
Y1=rbinom(10000, 12, 0.2)
Y2=rbinom(10000, 8, 0.2)
Y=Y1+Y2
barplot(table(factor(Y, levels=0:20)))

# Variant D
Y1=rnbinom(100000, prob=1-4/7, size=1)
Y2=rnbinom(100000, prob=1-4/7, size=2)
Y=Y1+Y2
barplot(table(factor(Y, levels=0:20)))
```

```

# Variant E
Y1=rpois(100000, lambda=4)
Y2=rpois(100000, lambda=4)
Y=Y1[Y1+Y2==8]
barplot(table(factor(Y, levels=0:20)))

# Variant F
Y1=rnbinom(1000000, prob=1-4/7, size=1)
Y2=rnbinom(1000000, prob=1-4/7, size=2)
Y=Y1[Y1+Y2==12]
barplot(table(factor(Y, levels=0:20)))

# Variant G
Y=rnbinom(10000, size=1000, prob=250/251)
barplot(table(factor(Y, levels=0:20)))
var(Y)
mean(Y)

```

### **Teoreetilise ja empiirilise jaotuse sarnasus - keskmine ruutviga (MSE)**

Vahel tekib soov mõista, kui hästi uuritava tunnuse valimi põhjal hinnatud jaotus sobida mõne teoreetilise jaotusega. Sellise võrdluse tegemiseks on erinevaid võimalusi, üks võimalus on kasutada keskmist ruutviga – vaadata kui erinevad on osakaalud teoreetilistest tõenäosustest:

```

# Valim
set.seed(1)
Y=rnbinom(100, prob=1-4/7, size=3)
table(factor(Y, levels=0:20))
osakaalud=prop.table(table(factor(Y, levels=0:20)))
osakaalud

# Keskmised ruutvead, MSE
# Võrdlus kindlate jaotustega

# Poissoni jaotus lambda=4,
p_Pois = dpois(x, 4)
mean((p_Pois-osakaalud)**2)
# Negatiivne binoomjaotus, p=3/7, r=3
p_nb = dnbinom(x, prob=1-4/7, size=3)
mean((p_nb -osakaalud)**2)
# binoomjaotus, n=20; p=0,2
p_binom = dbinom(x, 20, 0.2)
mean((p_binom -osakaalud)**2)

```

```

# Keskmised ruutvead, MSE
# Kui jaotuste parameetreid ei tea

# Poissoni jaotus, kui parameetrit lambda ei tea (hindame)
lambda_hinnang=mean(Y)
p_Pois_2=dpois(0:20, lambda=lambda_hinnang)
mean((p_Pois_2-osakaalud)**2)

# NB jaotus
#  $E(X)=r*(1-p)/p$ ;  $DX=r*(1-p)/p^2$ 
# - > avaldame r, p  $EX/DX=p$ ;  $E(X)*p/(1-p)=r$ 
p_hinnang=mean(Y)/var(Y)
r_hinnang=mean(Y)**2/(var(Y)-mean(Y))
p_nb_2=dnbinom(0:20, prob=p_hinnang, size=r_hinnang)
mean((p_nb_2-osakaalud)**2)

# Binoomjaotus, parameetrid MM abil,
#  $E(X)=n*p$ ;  $D(X)=n*p(1-p)$ ;  $D(X)/E(X)=1-p$ ;  $p=1-D(X)/E(X), \dots$ 
p_hinnang=1-var(Y)/mean(Y);
n_hinnang=var(Y)/(p_hinnang*(1-p_hinnang))

p_binom_2=dbinom(0:20, prob=p_hinnang, size=n_hinnang)
mean((p_binom_2-osakaalud)**2)

```

Mis juhtub siis, kui peame jaotusparameetreid hindama?

## Ülesanne 2

Uuritava tunnuse jaotus valimis on järgmine:

|       |      |      |      |      |      |     |      |      |      |       |
|-------|------|------|------|------|------|-----|------|------|------|-------|
| 0     | 1    | 2    | 3    | 4    | 5    | 6   | 7    | 8    | 9    | 10    |
| 0.015 | 0.06 | 0.14 | 0.21 | 0.19 | 0.12 | 0.1 | 0.08 | 0.05 | 0.03 | 0.005 |

Milline Poissoni jaotus (millise parameetriga lambda) sobitub kõige paremini (keskmise ruutvea järgi otsustades) antud jaotusega?

## Sagedustabelid.

Antud kursuse raames tuleb teha palju tööd sagedustabelitega. Sestap meenutame veidi ka sagedustabelitega seonduvat R-is.

Variant 1. Andmed on olemas objekt-tunnus maatriksi kujul:

```
load(url("http://www-1.ms.ut.ee/mart/kval_1.RData"))
head(andmed)
attach(andmed)
tab=table(Suitsetab, Infarkt)
tabjah``
addmargins(tab)
addmargins(tab,1)
plot(tab)
```

## Ülesanne 3

Tee sarnane tabel kasutades tunnuseid tunnus1 (Suitsetab? 1- ei; 2-jah) ja tunnus2 (Infarkt? 0 – jah; 1 - ei). Tee tabel nii, et tabeli veergudel oleksid inimese jaoks arusaadavad päised, mitte mingid kodeeringud).

## Ülesanne 4 – algandmeteks ongi sagedustabel.

Sisesta R-i järgmine andmetabel:

Tabel 10: Tunnuste  $X$ =”aspiriin” ja  $Y$ =”südame infarkt” sagedustabel

| $X \backslash Y$ | 1 (jah) | 2 (ei) | summa  |
|------------------|---------|--------|--------|
| 1 (platseebo)    | 189     | 10 845 | 11 034 |
| 2 (aspiriin)     | 104     | 10 933 | 11 037 |