R: Оценка на доверителни интервали (2020-11-30)

```
library(MASS)
attach(survey)
summary(Height)

samp1=sample(Height, size=70, replace=T)
summary(samp1)
samp2=sample(Height, size=70, replace=T)
samp3=sample(Height, size=70, replace=T)
samp4=sample(Height, size=70, replace=T)

hist(samp4, breaks=10)
abline(v=mean(Height, na.rm=T), lwd=2, col="red2")
abline(v=mean(samp1, na.rm=T), lwd=2, col="green1")
abline(v=mean(samp2, na.rm=T), lwd=2, col="green2")
abline(v=mean(samp3, na.rm=T), lwd=2, col="green3")
abline(v=mean(samp4, na.rm=T), lwd=2, col="green4")
```

Как да оценим интервал, по такъв начин, по който да заявим 95% сигрност, че популационното средно ще попадне в него..

Доверителни интервали за средното

z-test: Когато знаем какво е средното отклонение.

Ако $X \in \mathcal{N}(\mu, \sigma^2)$ и **знаем** σ, тогава $\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \in \mathcal{N}(0,1)$.

$$\mathbb{P}\left(z_{\frac{\alpha}{2}} < \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < z_{1-\frac{\alpha}{2}} \right) = 1-\alpha; \quad \mathbb{P}\left(-z_{-\frac{\alpha}{2}} < \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < z_{1-\frac{\alpha}{2}} \right) = 1-\alpha;$$

$$\mathbb{P}\left(-z^* < \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < z^* \right) = 1-\alpha; \quad \mathbb{P}\left(-z^* \frac{\sigma}{\sqrt{n}} < \overline{X}-\mu < z^* \frac{\sigma}{\sqrt{n}} \right) = 1-\alpha;$$

$$\mathbb{P}\left(\overline{X} - z^* \underbrace{\frac{\sigma}{\sqrt{n}}}_{\text{стандартна грешка}} < \mu < \overline{X} + \underbrace{z^* \frac{\sigma}{\sqrt{n}}}_{\text{максимална сохастична грешка}} \right) = \underbrace{1-\alpha}_{\text{доверитрелен интервал за } \mu};$$

```r
weight=c(75, 76, 73, 75, 74, 73, 73, 76, 73, 79)
sigma=1.5

function(weight, sigma, conf.level=0.95){
    n=length(weight)
    xbar=mean(weight)
    alpha=1-conf.level
    zstar=qnorm(1-alpha/2)
    SE=sigma/sqrt(n)
    xbar+c(-zstar*SE, zstar*SE)
}

z.test(weight, sigma=1.5)
# [1] 73.77031 75.62969

library(UsingR)
simple.z.test(weight, sigma=1.5, conf.level=0.95)
# [1] 73.77031 75.62969

h=survey$Height[!is.na(survey$Height)]
simple.z.test(h, sigma=10, conf.level=0.95)
# [1] 171.0251 173.7366
```

t-test: Когато не знаем какво е стандартното отклонение.

Обикновено не знаем какво е стандартното отклонение на данните от популацията, която изследваме и искаме да направим доверителен интервал за средното. Извадъчното стандартно отклонение бележим със s и го използваме него.

Ако $X \in \mathcal{N}(\mu, \sigma^2)$ и **НЕ знаем** σ, тогава $\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \in t(n-1)$.

Аналогично на предходните сметки, $1-\alpha$ доверителен интервал за μ е

$$\left(\overline{X} - t* \frac{s}{\sqrt{n}}, \overline{X} + t* \frac{s}{\sqrt{n}}\right),$$

където $t*$ е $1 - \frac{\alpha}{2}$ квантил на $t(n-1)$ разпределението. В този случай стандартната грешка е $SE := \frac{s}{\sqrt{n}}$.

$$(\overline{X} - t* SE, \overline{X} + t* SE)$$

Ако n е достатъчно голямо и отклонението на наблюдаваната случайна величина е добре дефинирано (крайно), то по ЦГТ може да използваме същия доверителен интервал.

Сега,

```
qqnorm(weight)
qqline(weight)

sh=shapiro.test(weight)
if(sh$p.value>0.05) print("normal") else print("not normal")

# може да използваме и qqplot.das за да прверим дали нашите данни се вписват в
# доверителните интервали на нормалното разпределение

library(StatDA)
qqplot.das(weight, "norm")

# сега вече може да използваме t-test, за да определим доверителен интервал за
# средното на изследваната от нас популация

t.test(weight)

# оценката с t-test-а ще даде по-широк доверителен интервал за средното на
популацията от оценката с z-test, тъй като тя прави оценка на стандартното
отклонение, а z-test-а го ползва като даденост (т.е. го взима точнчо).

x.norm=rnorm(1000)
x.t=rt(1000,9)
boxplot(x.norm, x.t)

qqplot(x.norm)
qqline(x.norm)

qqplot(x.t)
qqline(x.t)

xvals=seq(-4, 4, 0.01)
plot(xvals, dnorm(xvals), type="l", lwd=2)
for(i in c(2,5,10,20,50))
        points(xvals, dt(xvals, df=i), type="l, lty=i, col="darkblue")
```

С нарастването на степените на свобода на t-разпределението - все повече се приближава до нормалното разпределение. Като първоначално при по ниски степени на свобода - опашките на t-разпределението са по-тежки от тези на нормалното разпределение.

Доверителни интервали за популационната пропорция

$$p = \frac{\text{\#of successes in the POPULATION}}{\text{size of the POPULATION}}; \qquad \hat{p} = \frac{\text{\#of successes in the SAMPLE}}{\text{size of the SAMPLE}}.$$

За произволна извадка знаем, че $n\hat{p} \in Bin(n, p)$

```
n=100; p=0.42
prop.test(x=n*p, n=n, conf.level=0.99)
```

Доверителни интервали за медианата

Wilcox test:

```
x=c(110, 12, 2.5, 98, 1017, 540, 54, 4.3, 150, 432)
range(x)
# [1] 2.5 1017.0

wilcox.test(x, conf.int=T)
# 95 percent confidence interval:
#         33.0 514.5
```

Доверителния интервал е доста широк, тъй като броя на извадката е малък, а рейнджа е голям. **Не може да използваме t-test, тъй като данните не са нормално разпределени.**

Доверителни интервали за стандартното отклонение

Chi-square test:

Ние вече знаем, че ако наблюдаваната случайна велизина е $X \in \mathcal{N}(\mu, \sigma^2)$, то извадките от нея ще са $X_i \in \mathcal{N}(\mu, \sigma^2)$ и $\frac{X_i - \mu}{\sigma} \in \mathcal{N}(0,1),\ i = 1, 2, \ldots, n.$

Има два случая:

- Ако **знаем** μ, тогава $\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \in \mathcal{X}^2(n)$ или $\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \in \mathcal{X}^2(n)$

  Тук може да построим двустранен или едностранен доверителен интервал;

  - Едностранен или двустранен дов. инт.

    двустранен:

    ```
    height=survey$Height[is.na(survey$Height)]
    n=length(height)
    s=sqrt(sum(height-172)^2)
    alpha=0.05
    xstar1=chisq(1-alpha/2,n)
    xstar2=chisq(alpha/2,n)
    leftend=sqrt(s^2/xstar1)
    rightend=sqrt(s^2/xstar2)
    c(leftend, rightend)
    # [1] 8.972384 10.873545
    ```

едностранен:

```
n=length(height)
s=sqrt(sum((height-172)^2))
alpha=0.05
xstar=qchisq(1-alpha, n)
leftend=sqrt(s^2/xstar)
leftend
# [1] 9.104016
```

- Ако **НЕ знаем** μ, тогава може да използваме средното на извадката и да поставим степените на свобода на ХИ-квадрат разпределението да са $n-1$

$$\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{\sigma}\right)^2 \in \mathcal{X}^2(n-1) \text{ или } \sum_{i=1}^{n}\frac{(X_i-\overline{X})^2}{\sigma^2} \in \mathcal{X}^2(n-1)$$

  - Едностранен или двустранен дов. инт.

двустранен:

```
n=length(survey$Height)
s=sd(survey$Height, na.rm=T)
alpha=0.05
xstar1=chisq(1-alpha/2, n-1)
xstar2=chisq(alpha/2, n-1)
leftend=sqrt(s^2*(n-1)/xstar1)
rightend=sqrt(s^2*(n-1)/xstar2)
c(leftend, rightend)
# [1] 9.033603 10.823901
```

едностранен:

```
n=length(survey$Height)
s=sd(survey$Height, na.rm=T)
alpha=0.05
xstar=chisq(1-alpha, n-1)
leftend=sqrt(s^2*(n-1)/xstar)
leftend
# [1] 9.158673
```

Важно: доверителните интервали не винаги са верни.

Зад. 1 Острата левкимия е една от най-смъртоносните форми на рак. Предишни изследвания показват, че времето на преживяване след първоначалното откриване на левкимия е нормално разпределена сл. вел. с математическо очакване 13 месеца и стандартно отклонение 3 месеца. Въвежда се ново лечение, като се очаква то да удължи средното време на живот без да полияе на дисперсията. Наблюдавани са 16 пациента:

10.0 13.6 13.2 11.6 12.5 14.2 14.9 14.5 13.4 8.6 11.5 16.0 14.2 16.8 17.9 17.0

Да се намери оценка за очакването. Да се построи 95% доверителен интервал за средното време на живот на болните.

Реш:

I н/н:

```
X=scan()
# copy data with spaces
10.0 13.6 13.2 11.6 12.5 14.2 14.9 14.5 13.4 8.6 11.5 16.0 14.2 16.8 17.9 17.0

mu=13; s=3
simple.z.test(X, sigma=s, conf.level=0.95)
# [1] 12.27378 15.21372
```

II н/н:

```
X=c(10.0, 13.6, 13.2, 11.6, 12.5, 14.2, 14.9,14.5, 13.4, 8.6, 11.5, 16.0, 14.2, 16.8, 17.9,
17.0)

z.test=function(X, sigma, conf.level=0.95){
    n=length(X)
    xbar=mean(X)
    alpha=1-conf.level
    zstar=qnorm(1-alpha/2)
    SE=sigma/sqrt(n)
    xbar+c(-zstar*SE,zstar*SE)
}

z.test(X, sigma=3)
# [1] 12.27378 15.21372
```

Зад. 2 Генерирайте 20 наблюдения над случайна величина, която е нормално разпределена с очакване 5 и дисперсия 4. Постройте 90%-тов доверителен интервал за математическото очакване. Повторете опита 100 пъти. Проверете, в колко от случайте математическото очакване принадлежи на доверителния интервал.

Реш:

I н/н:

```
sol1=function(N=100, n=20, mu=5, s=4){
    cnt=0
    for(i in 1:N){
        X=rnorm(n=n, mean=mu, sd=s)
        z=simple.z.test(X, sigma=s, conf.level=0.90)
```

```
        left=z[1]; right=z[2]
        M=mean(X)
        if(left<mu && mu<right)
            cnt=cnt+1
    }
    cnt/N
}
sol()

II н/н:
sol2=function(N=100, n=20, mu=5, s=2){
    CI=function(X, alpha=0.05){
        n=length(X)
        mean(X)+c(-qnorm(1-alpha/2)*sd(X)/sqrt(n), qnorm(1-alpha/2)*sd(X)/sqrt(n))
    }
    cnt=0
    for(i in 1:N){
        X=rnorm(n=n, mean=mu, sd=s)
        ci=CI(X, 0.10)
        left=ci[1]; right=ci[2]
        if(left<mu && mu<right[])
            cnt=cnt+1
    }
    cnt/N
}
```

Зад. 3 Постройте 95%-тов доверителен интервал за средното време на живот на болните, ако и дисперсията **НЕ** е известна.

Реш:

Доверителния интервал за средното на **НОРМАЛНО РАЗПРЕДЕЛЕНА** (иначе трябва да проверяваме дали е нормална или дали са много наблюденията и от ЦГТ да приближим с нормално) случайна величина с неизвестна дисперсия е:

```
X=c(10.0, 13.6, 13.2, 11.6, 12.5, 14.2, 14.9, 14.5, 13.4, 8.6, 11.5, 16.0, 14.2, 16.8, 17.9,
17.0)
I н/н:
tTest=function(X, conf.level=0.95){
    n=length(X)
    xbar=mean(X)
    sigma=sd(X)    alpha=1-conf.level

    zstar=qt(1-alpha/2, n-1)
    SE=sigma/sqrt(n)
    xbar+c(-zstar*SE, zstar*SE)
}
# [1] 12.39066 15.09684
II н/н:
t.test(X, conf.level=0.95)
# [1] 12.39066 15.09684
```

Зад. 4 Направете графика с плътността на стандартно нормално разпределение и разпределение на "Student's t" с 5, 10, 30, 100 степени на свобода.

Реш:

Колкото повече степпените на свобода на "Student's t" разпределената сл. вел. растат, толкова по-близо ще става до стандартното нормално разпределение: t(n)->N(0,1)

```
sol=function(){
   x=seq(from=-3, to=3, by=0.01)

   l=c(5,10,30,100)
   c=c("red1", "red2", "red3", "red4")
   i=1
   plot(x, dnorm(x), type="l")
   for(d in l){
      lines(x, dt(x, df=d), lty=i+1, col=c[i])
      i=i+1
   }
}
sol()
```

```
n <- c(5, 10, 30, 100)
col <- c("#00F5FF", "#4F94CD", "#836FFF", "#473C8B")
plot(x, y1, type = "l")
for (i in 1:4){
        y <- dt(x, df = n[i])
        lines(x, y, lty = i+1, col = col[i])
}
temp <- legend("topright",
        legend = c(" ", " ", " ", " ", " "),
        text.width = 1,
        lty = 1:5,
        xjust = -1,
        yjust = 2,
        col = c("black", col))
text(temp$rect$left + temp$rect$w,
        temp$text$y,
        c("N(0,1)", "t(5)", "t(10)", "t(30)", "t(100)"),
        pos = 2,
        col = c("black", col))
```

Зад. 5 За данните rat от пакета UsingR postrojte 96% доверителен интервал за очакването.

Реш:

Първо трябва да проверим дали данните са нормално разпределени.

```
qqnorm(rat)
qqline(rat)
```

```
sh=shapiro.test(rat)
if(sh$p.value>0.05) print("normal") else print("not normal")
# [1] "normal"
```

Можеше да се обедим в това, че данните са нормално разпределени и като използваме qqplot.das(rat, "norm") и проверим дали опашките на нашите данни попадат в доверителните интервали на нормалното разпределение.

```
length(rat)
# [1] 20

mean(rat)
# [1] 113.45

t.test(rat, conf.level=0.96)
```

Зад. 6 При провеждане на анкета 87 от 150 анкетирани са отговорили, че са използвали даден продукт. Постройте 92% доверителен интервал за броя на хората използвали продукта.

Реш:
I н/н:

```
sol=function(n=150, k=87, per=92){
    alpha=1-per/100
    phat <- k/n
    SE <- sqrt((phat * (1 - phat)) / n)
    MaxE <- qnorm(1 - alpha/2) * SE
    (phat + c(-MaxE, MaxE))*n
}
```

II н/н:

```
N=150; succ=87
ci=prop.test(x=succ, n=N, conf.level=0.92)
left=ci$conf.int[1]*N; right=ci$conf.int[2]*N
left; right
# [1] 75.77986
# [1] 97.7171
```