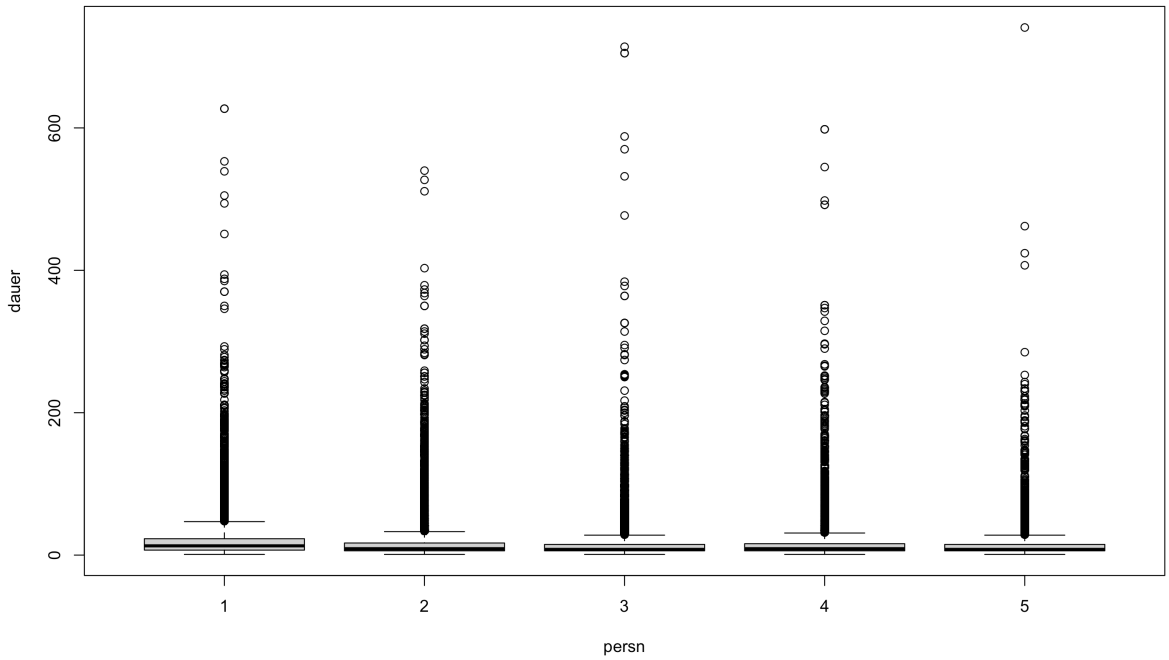
Регресійний аналіз
4 курс, статистика, Шкляр Ірина Володимирівна

Завдання 5, варіант 7

Потрібно перевірити чи є залежність між середнім і/або дисперсією змінної dauer (інтервал між двома черговими покупками кави) та змінною persn (кількість осіб у домогосподарстві покупця).

Розпочнемо з діаграми скриньок з вусами для наших даних:

```
> data<-read.table(file="~/Downloads/regrasympt/kaffee.txt",header=T)
> boxplot(dauer~persn,data=data)
```



З діаграми видно, що значення інтервалів між двома черговими покупками кави – dauer, є приблизно однаковими, незалежно від кількості осіб у домогосподарстві покупця – persn. Використаємо тест Фішера для однофакторного дисперсійного аналізу:

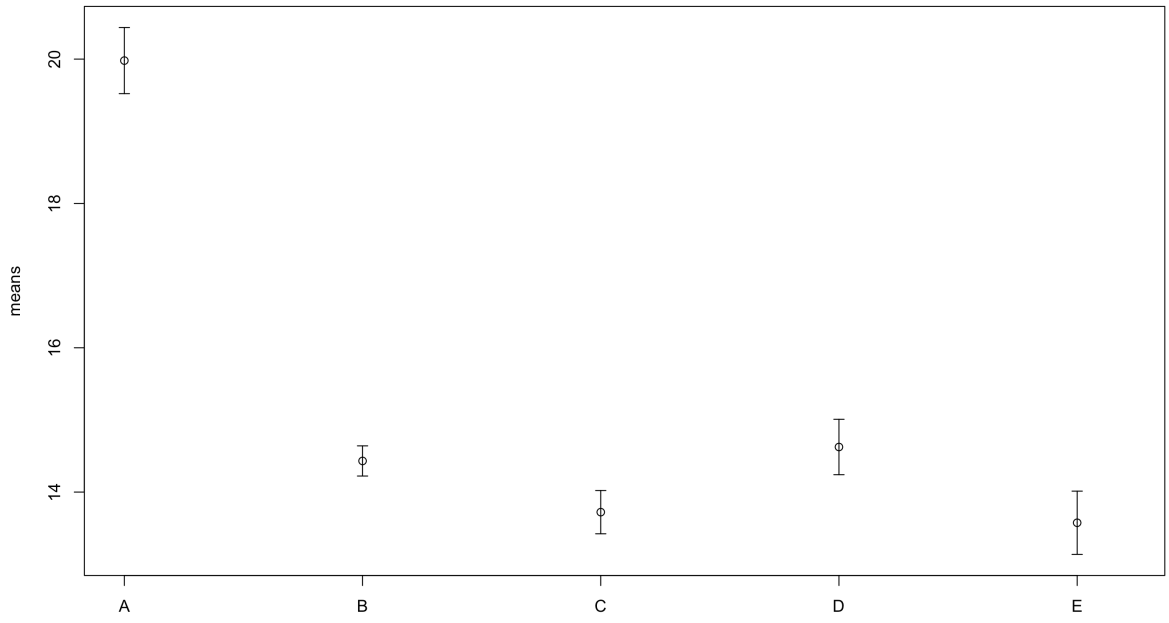
```
> res<-aov(dauer~persn,data=data)
> summary(res)
               Df   Sum Sq Mean Sq F value Pr(>F)
persn           1   265863  265863   630.6 <2e-16 ***
Residuals   130984 55221054     422
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

M-1 = 1, тому M=2.

З таблиці дисперсійного аналізу бачимо, що досягнутий рівень значущості тесту для перевірки гіпотези про рівність математичних сподівань дорівнює < 2e-16, це майже 0. Гіпотезу слід відхилити - математичні сподівання dauer є різними при різній кількості осіб у домогосподарстві покупця.

Тест Фішера не вказує, для якої саме кількості осіб у домогосподарстві покупця є відмінності. Щоб побачити їх, побудуємо довірчі інтервали для математичних сподівань. Скористаємось формулою для довірчих інтервалів для середніх. Будуємо одночасні довірчі інтервали для всіх кількостей осіб у домогосподарстві зі стандартним рівнем значущості 0.05, тому номінальний рівень значущості розраховуємо за формулою:

```
> library(plotrix)
> alpha0<-0.05
> alpha<-1-(1-alpha0)^(1/5)
> l<-tapply(data$dauer,data$persn,length)
> m<-tapply(data$dauer,data$persn,mean)
> s<-tapply(data$dauer,data$persn,sd)
> tf<-qt(1-alpha/2,l-1)
> h<-s*tf/sqrt(l)
> plotCI(1:5,y=m,uiw=h,xlab=" ",ylab="means",xlim=c(1,5.2),xaxt="n")
> axis(1,at=1:5,labels=c("A","B","C","D","E"))
```



Отримані інтервали показують, що математичні сподівання dauer для кількості осіб A і інших кількостей (B, C, D, E) значущо відрізняються.

Перевіримо однорідність дисперсій за допомогою теста Левена:

> #library(car)
> leveneTest(data$dauer,data$persn)

```
Levene's Test for Homogeneity of Variance (center = median)
          Df F value    Pr(>F)
group      4  227.52 < 2.2e-16 ***
      130981
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Досягнутий рівень значущості дорівнює 2.2e-16, отже відхиляємо основну гіпотезу про однорідність дисперсій (є значущі відмінності дисперсій dauer при різній кількості осіб у домогосподарстві покупця.)