Статистика за аутоматску анализу података

др Марко Обрадовић

Основне информације

Предиспитне и испитне обавезе

- ПИ: Четири теста која вреде по 20 поена, рачунају се три најбоље урађена
- И: Усмени испит, вреди 40 поена
- Бонус поени: Групни семинарски рад, вреди 10 поена (није обавезан, додаје се на резултат усменог)

Литература

- J.S. Milton, J.J. Corbet and P.M. McTeer. Introduction to Statistics, DC Heath & Company, 1986.
- Б. Милошевић. *Основи статистике*, Математички факултет, Београд, 2021.

Увод

Дефиниција

Статистика *је наука о подацима, тј. о њиховом прикупљању, приказивању, анализирању и извођењу закључака на основу њих.*

Статистичке методе деле се на

- дескриптивне (описне)
- методе статистичког закључивања

Увод

Дефиниција

Популација *у статистичком смислу је група објеката о којима треба донети некакав закључак.*
Узорак *је део (или подскуп) објеката извучен из популације.*

Популације – пример

- Конзумирање алкохола међу тинејџерима: Колики је проценат њих који конзумира редовно и с колико година су почели да конзумирају?
- Производња у аутомобилској индустрији: колико хауба може машина у просеку да офарба пре првог сервиса?
- Испитивање јавног мњења: треба ли уводити нове аутобуске линије?

Увод

Дефиниција

Случајна променљива *је променљива чије се вредности одређују исходом случајног експеримента.*
Случајна променљива дефинисана на објектима популације назива се и **обележјем** *те популације.*

Дефиниција

Непрекидна случајна променљива *је случајна променљива, која, пре изведеног експеримента, може узети било коју вредност из неког интервала реалних бројева.*
Дискретна случајна променљива *је случајна променљива, која може узети највише коначно или пребројиво бесконачно много различитих вредности.*

Увод

Случајне променљиве – пример

- Редовно конзумирање алкохола – дискретна променљива с две вредности “да ” и “не”; Старост почетка конзумирања – непрекидна променљива
- Број офарбаних хауба – дискретна променљива с пребројиво бесконачно вредности 0,1,2,...
- Заинтересованост за увођењем аутобуске линије – дискретна променљива – “да” и “не“

Дискретне променљиве могу се сврстати у две групе: **категоричке (фактори)** и **нумеричке**. Категоричке променљиве даље се деле на **номиналне** и **ординалне**. Непрекидне променљиве све припадају групи нумеричких.

- номиналне: ”да“–”не“, пол, боја косе, место рођења...
- ординалне: стручна спрема, одговори на анкетама...

Увод

Дефиниција

Параметар популације *је нека описна мера случајне променљиве (обележја) посматране на целој популацији.*
Статистика *је описна мера случајне променљиве (обележја) посматране само на узорку.*

Параметри популације – пример

- p - удео (проценат) редовних конзумената алкохола; μ - просечна старост почетка конзумирања
- m - просечан број офарбаних хауба до првог сервиса

Кораци у статистичкој анализи

- Одредити популацију која се проучава
- Поставити питања у вези популације на која желимо одговор
- Одредити случајне променљиве (обележја) чије ће проучавање помоћи да дођемо до одговора
- Одредити параметре популације који су од важности
- Извући узорак из популације
- Одредити статистике којима ће се проценити вредности непознатих параметара
- Применити технике статистичког закључивања и одговорити на постављена питања

Прелиминарна анализа података

- Циљ прелиминарне анализе је приказати податке на што разумљивији начин. Укључује графички приказ и рачунање неких важних статистика који могу дати више информација о променљивим од интереса.
- Она нам помаже у конструкцији **статистичког модела** који обухвата све наше претпоставке о променљивим и служи као основа за све методе статистичког закључивања.

Анализа нумеричких података

Старост деце кад је примећен први знак аутизма – пример целе популације, није узорак!

1 6 8 3 2 3 14 24 7 4

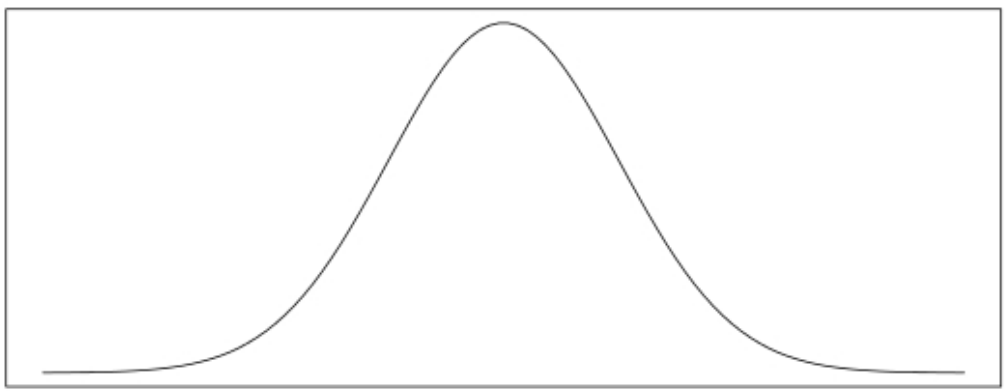
Снага земљотреса у Калифорнији по Рихтеровој скали – пример узорка

1.0	8.3	3.1	1.1	5.1
1.2	1.0	4.1	1.1	4.0
2.0	1.9	6.3	1.4	1.3
3.3	2.2	2.3	2.1	2.1
1.4	2.7	2.4	3.0	4.1
5.0	2.2	1.2	7.7	1.5

Занима нас:

- Какав је облик расподеле? Да ли вредности случајне променљиве чине неку препознатљиву структуру?
- Који је положај података, тј. око које централне вредности су они распоређени?
- Колико има одступања међу подацима? Да ли су они прилично расејани или згуснути око централне вредности?

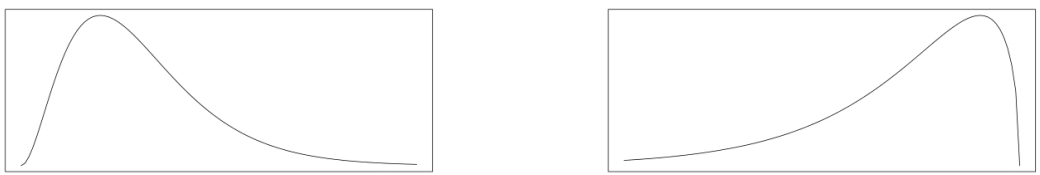
Облици расподела



Слика: симетрична расподела

Дефиниција

За расподелу се каже да је **померена удесно** *уколико има дугачак реп на десној страни. Уколико је тај реп на левој страни, каже се да је* **померена улево**.



Слика: расподеле померене удесно и улево

Дефиниција

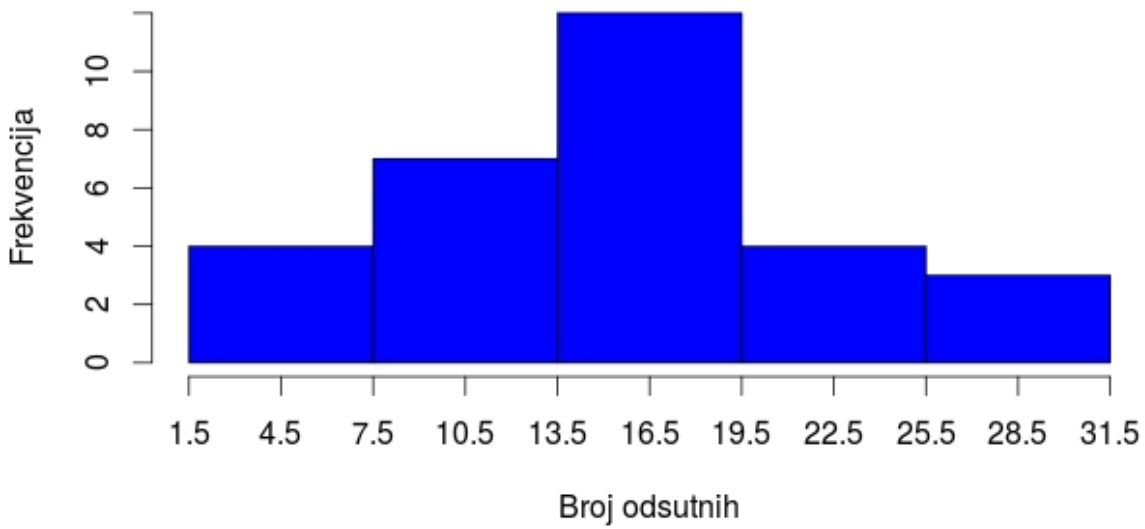
Хистограм фреквенција (учесталости) је график такав да је висина сваког стуба једнака броју елемената из узрока у категорији коју представља.

Конструкција хистограма

- Одредити број класа – препорука је $1 + \log_2 n$, где је n укупан број података, заокружено нагоре на цео број.
- Одредити најмањи и највећи елемент у узорку; Наћи узорачки распон је једнак њиховој разлици
- Наћи минималну ширину стуба дељењем распона с бројем стубова
- Наћи стварну ширину стуба заокруживањем минималне ширине на горе, на онолики број децимала колики имају и подаци
- Одредити леву границу првог стуба, која је мало (за пола јединице) мања од најмањег елемента узорка
- Одредити остале границе и нацртати стубове

Хистограм

Број одсутних радника с посла	15	9	15	5	16	16
	30	7	12	9	23	15
	21	16	17	13	20	18
	2	31	11	12	27	22
	15	14	10	6	19	14



Слика: Хистограм броја одсустава

Хистограм

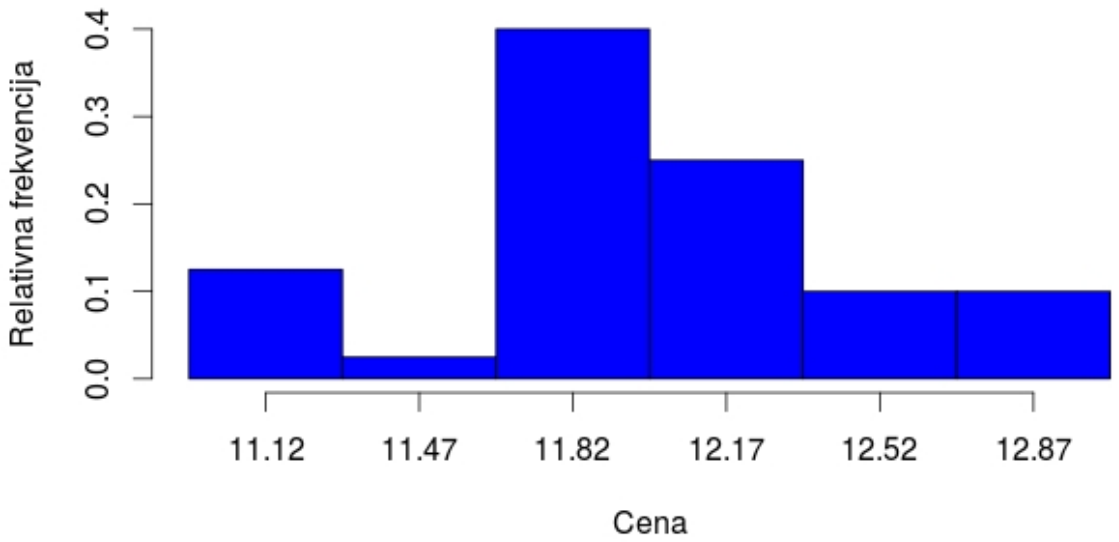
Дефиниција

Хистограм релативних фреквенција *је график такав да је висина сваког стуба једнака уделу (проценту) елемената из категорије коју представља у целом узорку.*

Хистограм

Цене лека у апотекама

12.00	11.98	11.48	12.99	11.20	12.06	11.98	11.20
12.50	13.02	11.75	12.05	11.71	11.10	11.82	11.80
11.75	11.17	12.25	11.90	12.03	11.89	12.15	11.96
11.87	10.95	12.20	11.85	11.70	11.92	13.00	12.40
12.03	12.75	12.69	12.03	11.90	11.72	12.60	12.11



Табеларни приказ

```
> baza$radni.st<-factor(baza$radni.st)
> baza$bracno.st<-factor(baza$bracno.st)
> baza$br.dece<-as.numeric(baza$br.dece)
> baza$godine<-as.numeric(baza$godine)
> baza$obrazov<-factor(baza$obrazov, ordered = TRUE)
> baza$otac.obr<-factor(baza$otac.obr, ordered = TRUE)
> baza$majka.obr<-factor(baza$majka.obr, ordered = TRUE)
> baza$pol<-factor(baza$pol)
> summary(baza)
```

```
   radni.st   bracno.st     br.dece          godine      obrazov otac.obr   majka.obr   pol
1      :237   1:232     Min.   : 1.000   Min.   : 1.00   0: 52   0 :147   1      :179   1:185
2      : 58   2: 31     1st Qu.: 1.000   1st Qu.:16.00   1:208   1 :135   0      :152   2:215
5      : 40   3: 52     Median : 3.000   Median :25.00   2: 29   2 :  5   3      : 24
7      : 29   4:  6     Mean   : 2.868   Mean   :27.14   3: 74   3 : 36   8      : 16
4      : 17   5: 79     3rd Qu.: 4.000   3rd Qu.:36.00   4: 37   4 : 16   2      : 10
3      :  9             Max.   :10.000   Max.   :65.00           8 : 13   4      :  9
(Other): 10                                                      NA: 48   (Other): 10
```

Пре даље анализе требало би избацити недостајуће податке

```
> baza1<-na.omit(baza)
> summary(baza1)
```

За табелирање једне променљиве користи се

```
> levels(baza1$radni.st)=c("puno radno vreme", "skraceno radno vreme", "trenutno ne radi", "nezaposlen, otpusten", "u penziji", "ucenik/ca", "domacin/ca", "drugo")
```

```
> table(baza1$radni.st)
```

puno radno vreme	skraceno radno vreme	trenutno ne radi	nezaposlen, otpusten	u penziji
205	56	7	15	31
ucenik/ca	domacin/ca	drugo		
6	23	3		

или једне променљиве у односу на другу

```
> levels(baza1$pol)=c("M","Z")
```

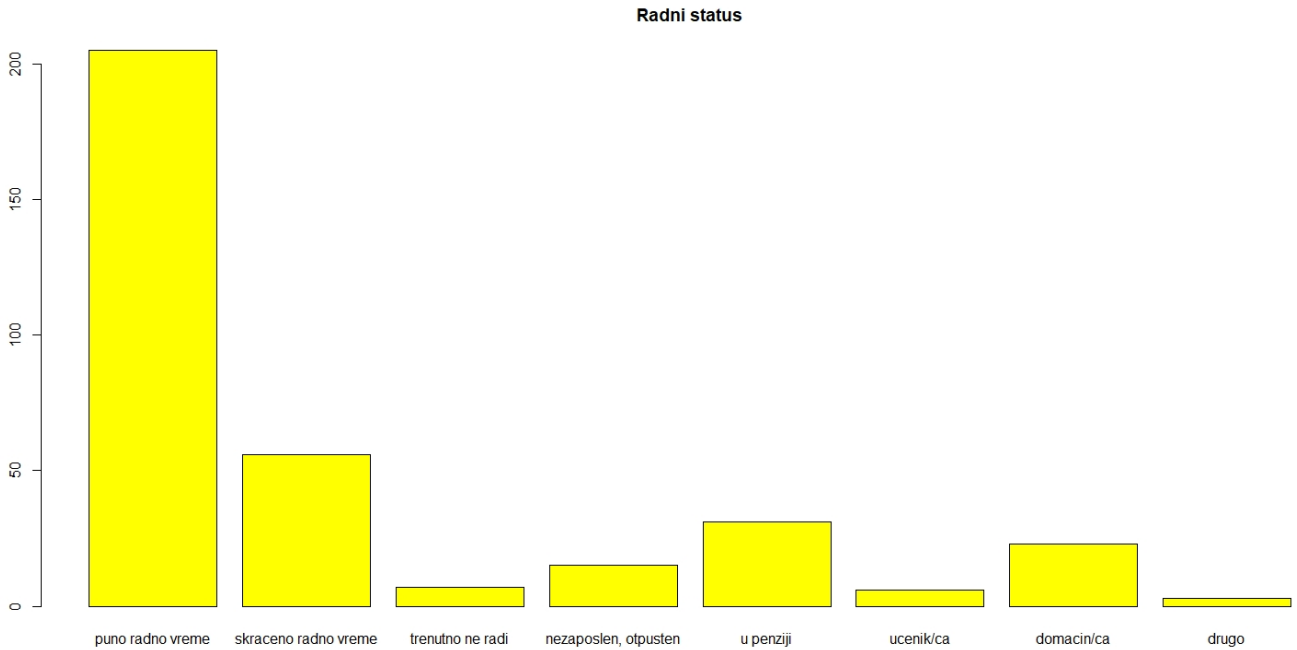
```
> table(baza1$radni.st, baza$pol)
```

	M	Z
puno radno vreme	118	87
skraceno radno vreme	14	42
trenutno ne radi	1	6
nezaposlen, otpusten	9	6
u penziji	15	16
ucenik/ca	1	5
domacin/ca	0	23
drugo	3	0

Тракасти дијаграм

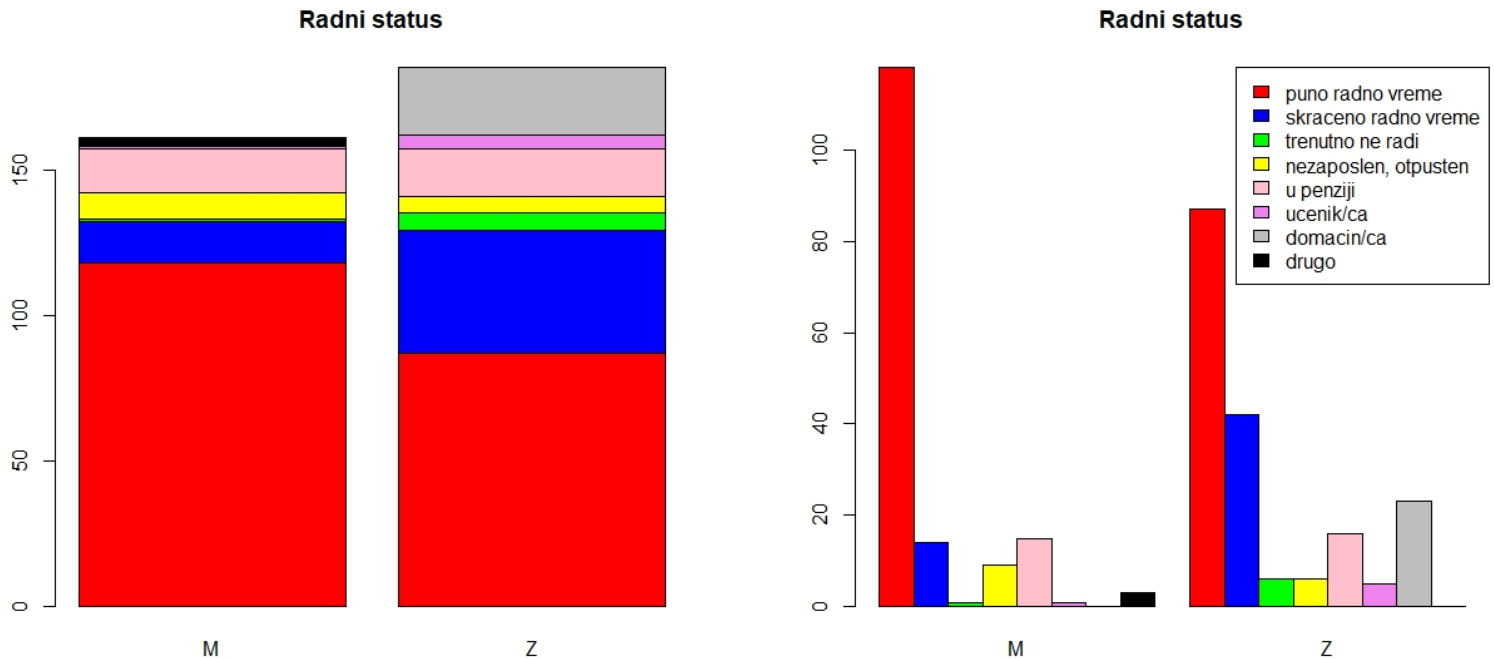
Тракасти дијаграм на x—оси приказује категорије, а на y—оси фраквенције појављивања елемената из узорка у свакој од категорија

```
> barplot(table(baza1$radni.st), col="yellow", main="Radni status")
```



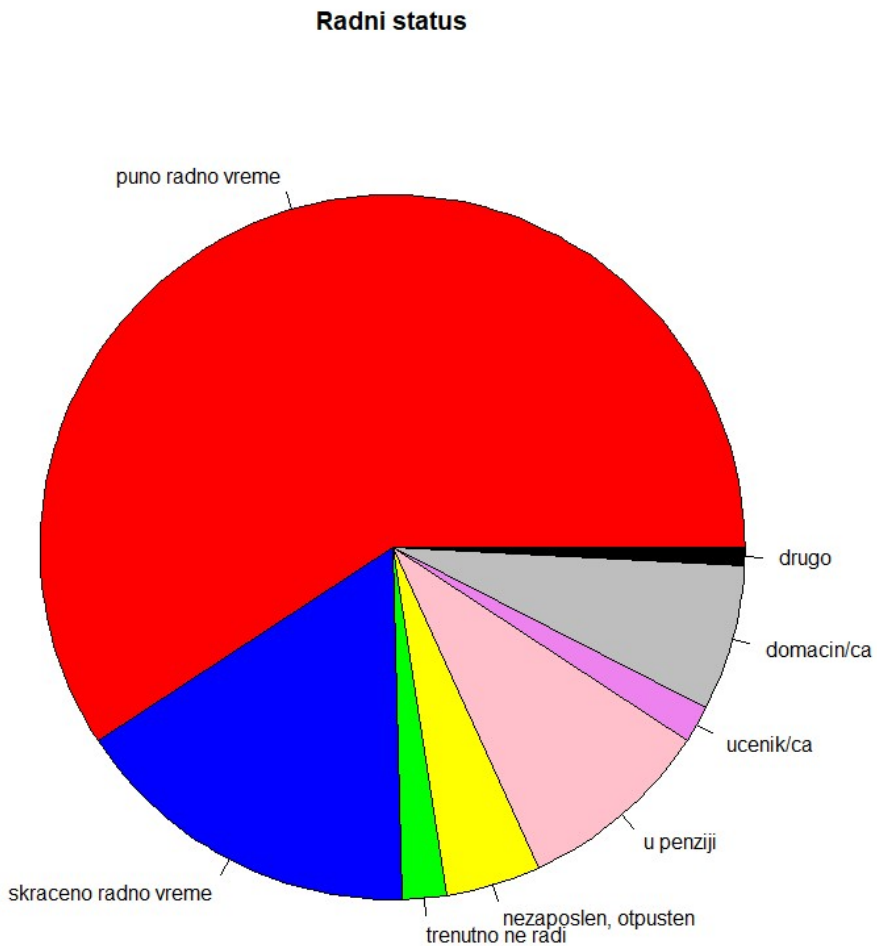
Може се цртати једна променљива у односу на категорије друге променљиве.

```
> boje<-c("red","blue","green","yellow","pink","violet","grey","black")
> barplot(table(baza1$radni.st, baza1$pol), main="Radni status",
col=boje)
> barplot(table(baza1$radni.st, baza1$pol), main="Radni status",
col=boje, beside=TRUE)
> legend("topright", legend=levels(baza1$radni.st),
fill=boje,cex=0.8)
```

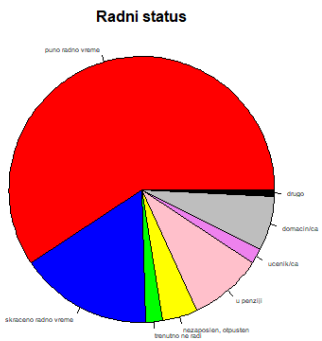


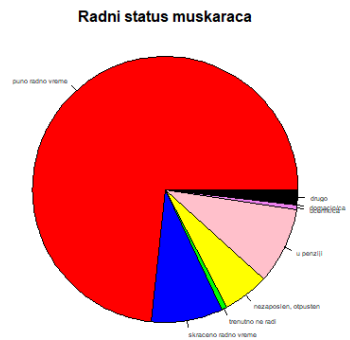
Кружни дијаграм

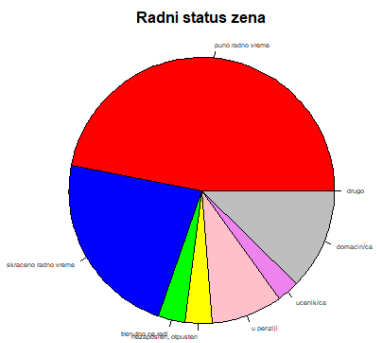
```
> pie(table(baza1$radni.st), main="Radni status",
col=boje)
```

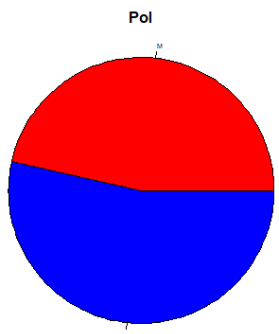


```
> par(mfrow=c(2,2))
> pie(table(baza1$radni.st), main="Radni status", col=boje,
radius=1, cex=0.5)
> pie(table(baza1$radni.st[baza1$pol=="M"]), main="Radni status
muskaraca", col=boje, radius=1, cex=0.5)
> pie(table(baza1$radni.st[baza1$pol=="F"]), main="Radni status
zena", col=boje, radius=1, cex=0.5)
> pie(table(baza1$pol), main="Pol", col=boje, radius=1, cex=0.5)
```



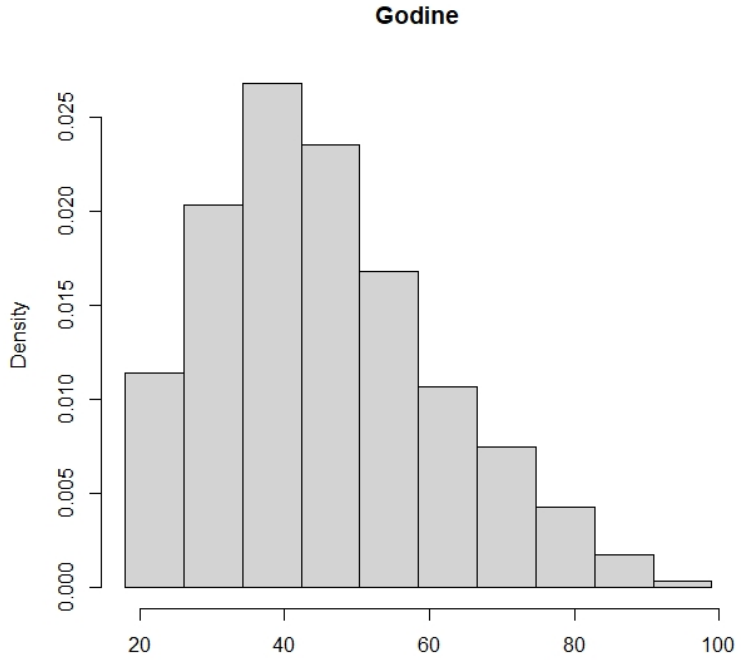






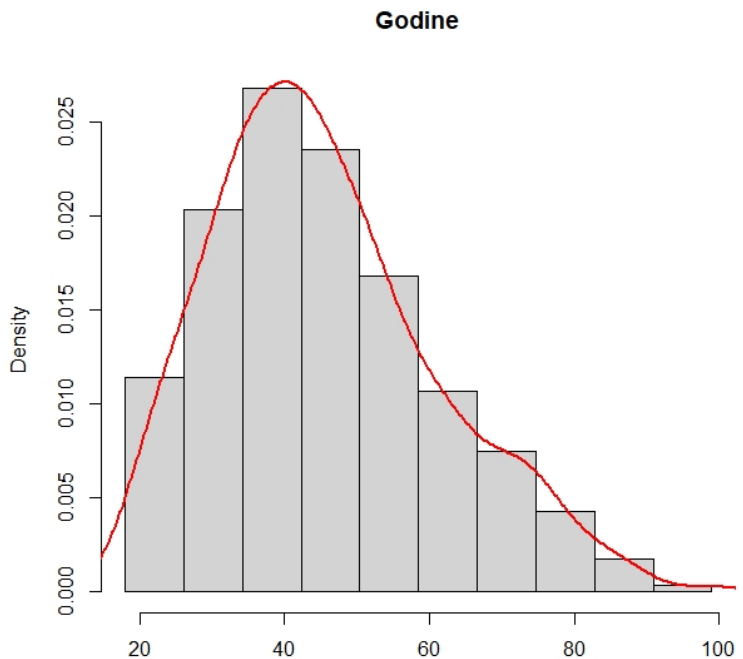
Одређивање граница хистограма по формули

```
> broj.kategorija <-ceiling(log(length(baza1$godine),2))+1
> d<-(max(baza1$godine)-min(baza1$godine))/broj.kategorija
> granice<-min(baza1$godine)-1+(0:broj.kategorija)*(d+0.1)
> hist(baza1$godine, breaks=granice, prob=TRUE)
```



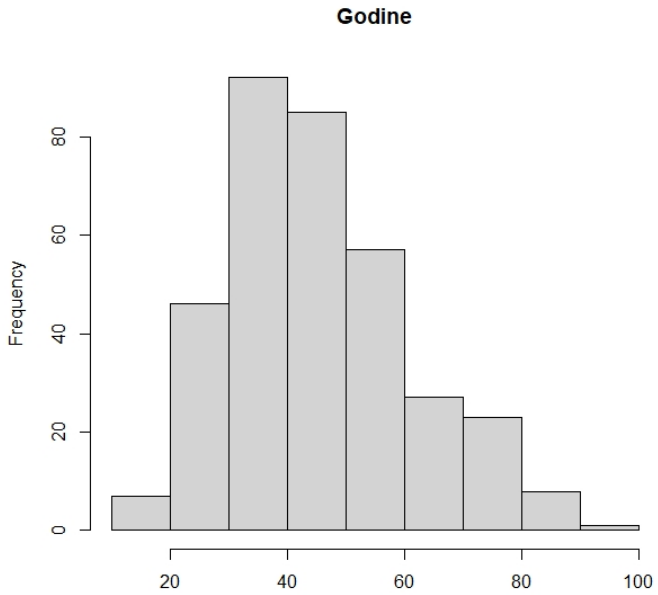
Додавање оцене густине на хистограм

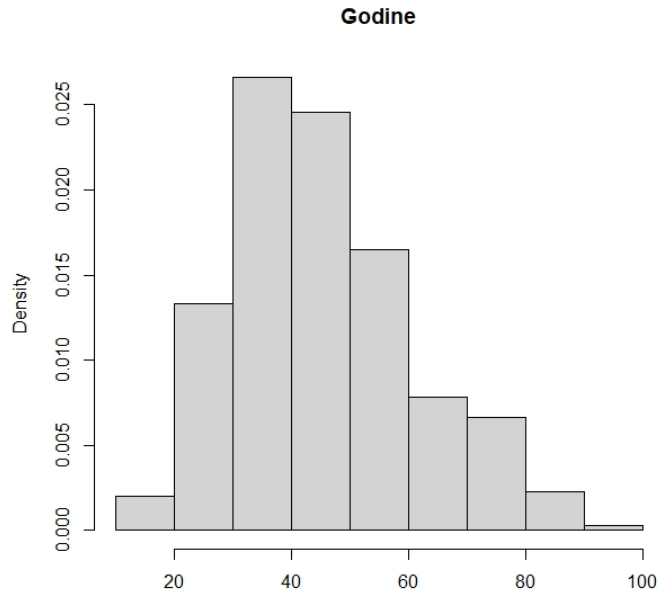
```
> broj.kategorija <-ceiling(log(length(baza1$godine),2))+1
> d<-(max(baza1$godine)-min(baza1$godine))/broj.kategorija
> granice<-min(baza1$godine)-1+(0:broj.kategorija)*(d+0.1)
> hist(baza1$godine, breaks=granice, prob=TRUE)
> lines(density(baza1$godine), col="red", lwd=2)
```



Коришћење подразумеваних граница функције `hist`

```
> hist(baza1$godine, main="Godine", xlab="")
> hist(baza1$godine, main="Godine", prob=TRUE, xlab="")
```





Процена удела популације чија је старост мања од 40 година

```
> H<-hist(baza1$godine, prob=TRUE, plot=FALSE)
> cumsum(H$density*diff(H$breaks))
[1] 0.02023121 0.15317919 0.41907514 0.66473988 0.82947977 0.90751445
0.97398844 0.99710983 1.00000000
```

Тражена процена је 0.42.

Мере положаја

Три важна параметра популације који одређују положај расподеле су:

- средња вредност популације
- медијана популације
- мода популације

Они се називају и **параметри положаја** или **мере централне тенденције**.

Средња вредност

Средња вредност популације μ – непознати параметар
Процењујемо га (приближно) статистиком коју називамо узорачком средњом вредношћу или, краће, узорачком средином.

Дефиниција

Нека су $x_1, x_2, \ldots, x_n$, n вредности случајне величине X добијене у узорку. **Узорачком средином** *називамо $\bar{x}$, аритметичку средину тих вредности, тј.*

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

```
> mean(x)
```

Узорачка средина – примери

Број упамћених речи за два минута

8 2 4 9 7 2 12 5 5 7

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{61}{10} = 6.1.$$

Комбиновање више узорачких средина

Број хитних случајева у једној болници је $\bar{x}_1 = 3$ за $n_1 = 5$, а у другој болници $\bar{x}_2 = 15$ за $n_2 = 100$.

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{5 \cdot 3 + 100 \cdot 15}{5 + 100} = \frac{1515}{105} = 14.4.$$

```
> sredine<-c(3,15)
> tezine<-c(5,100)/105
> weighted.mean(sredine,tezine)
```

Медијана

Медијана популације - непозната вредност од које је пола популације веће, а пола мање
Процењујемо је (приближно) статистиком коју називамо узорачком медијаном.

Дефиниција

Нека је $x_1, x_2, \ldots, x_n$ узорак поређан по величини од најмање до највеће вредности. Уколико је n непаран број, **узорачка медијана** *је број тачно на средини низа. Уколико је n паран број, узорачка медијана је аритметичка средина два броја на средини низа.*

Медијана - примери

Године старости купаца у једној продавници гардеробе

жене		мушкарци
12	27	17
15	30	29
17	35	37
20	42	40
24	60	72

Медијана старости жена је $(24 + 27)/2 = 25.5$, а мушкараца је 37 година.

На већим узорцима рачунамо преко положаја медијане $(n + 1)/2$.

```
> median(x)
```

Средња вредност и медијана

Узорак тржишне вредности (у хиљадама доларима) десет кућа у једном насељу

82 91 78.5 86 80.5 85 82.5 80 77 850

Какав је ово крај?

Средња вредност је 159.25, а медијана је 82.25.

Из вредности $\bar{x}$ извлачимо погрешан закључак о вредности кућа у крају, медијана нам даје много бољу информацију. То је због утицаја неуобичајене вредности 850 коју називамо **аутлајером** (енгл. outlier - онај који се ту налази али не припада).

Мода

Мода популације – непозната вредност која је најчешћа у популацији
Процењујемо је (приближно) статистиком коју називамо узорачком модом, вредношћу која се највише пута појављује у узорку.

Уколико је расподела симетрична, тада се средња вредност, медијана и мода популације поклапају. Одговарајуће статистике, наравно неће се поклапати, али ће имати блиске вредности.

Мере расејања

Важни параметри популације који описују расејање расподеле

- распон популације
- дисперзија (варијанса) популације σ^2
- стандардно одступање (девијација) популације σ
- међуквартилно растојање

Распон

Распон је разлика највећег и најмањег елемента популације
Процењујемо га (приближно) узорачким распоном.

Дефиниција

Узорачки распон *је разлика између највећег и најмањег елемента узорка.*

- није посебно добар као мера расејања

```
> diff(range((x))
```

Пример: резултати студената на испиту у два семестра

	први семестар	други семестар
обим узорка	23	26
средњи број поена $\bar{x}$	75	75
медијана број поена	75	75
распон	50 (од 50 до 100)	50 (од 50 до 100)

Стварна расподела поена

први семестар	други семестар
50 50 50 50 50 50	50
60 60	65 65
70 70	70 70 70
75	74 74 74 74
80 80	75 75 75 75 75 75
85 85 85	76 76 76 76
100 100 100 100 100 100	80 80 80
	85 85
	100

Дисперзија

Параметар популације – средње квадратно одступање случајне величине X од своје средње вредности μ
Приближно је процењујемо узорачком дисперзијом

Дефиниција

Нека је $x_1, \ldots, x_n$ узорак од n елемената. **Узорачка дисперзија** *дефинише се као*

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}.$$

```
> var(x)
```

Рачунање дисперзије

Формула за рачунање узорачке дисперзије

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} x_i^2 - \frac{n}{n-1}\bar{x}^2.$$

Подаци о дужини трајања телефонских разговора

10 20 6 12 15 8 4 9 3 12

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{9} = \frac{(10-10)^2 + \cdots + (13-10)^2}{9} = \frac{244}{9} = 27.11$$

$\bar{x} = 10$ минута; $\sum_{i=1}^{n} x_i^2 = 1244$

$$s^2 = \frac{1}{9}\cdot 1244 - \frac{10}{9}\cdot 10^2 = 27.11.$$

Стандардно одступање

Параметар популације – квадратни корен из дисперзије
Процењујемо га (приближно) узорачким стандардним одступањем

Дефиниција

Узорачко стандардно одступање *једнако је квадратном корену из узорачке дисперзије, тј.* $s = \sqrt{s^2}$.

```
> sd(x)
```

Стандардно одступање – пример

Подаци о дневној температури

$$2°\mathrm{C} \quad 5°\mathrm{C} \quad 8°\mathrm{C} \quad 0°\mathrm{C} \quad 10°\mathrm{C} \quad 20°\mathrm{C} \quad -10°\mathrm{C}$$

$$s^2 = 86.33, \text{ а } s = \sqrt{86.33} = 9.3°\mathrm{C}.$$

Међуквартилно растојање

Међуквартилно растојање IQR – мера расејања неосетљива на аутлајере за разлику од распона и дисперзије Узорачки квартили:

- Први квартил Q_1 је вредност од које је $1/4$ узорка мање, а $3/4$ веће.
- Други квартил Q_2 (медијана) је вредност од које је $2/4$ узорка мање, а $2/4$ веће.
- Трећи квартил Q_3 је вредност од које је $1/4$ узорка мање, а $3/4$ веће.

```
> quantile(x)
```

Међуквартилно растојање $\mathrm{IQR} = Q_3 - Q_1$ је распон у ком се налази средњих 50% узорка.

Међуквартилно растојање

- Одредити положај узорачке медијане, $(n+1)/2$, где је n обим узорка.
- Одредити l, највећи природан број који није већи од $(n+1)/2$ (може бити једнак).
- Наћи положај квартила као $q = (l+1)/2$.
- Одредити q_1, број у узорку који је q-ти по величини почевши од најмањег. Ако q није природан број, тада је q_1 аритметичка средина бројева који су $q - 1/2$ и $q + 1/2$ по реду. Приближно 25% (четвртина) узорка ће бити мање од q_1, па се он назива први квартил узорка.
- Одредити q_3, број у узорку који је q-ти по величини почевши од највећег. Ако q није природан број, тада је q_1 аритметичка средина бројева који су $q - 1/2$ и $q + 1/2$ по реду. Приближно 75% (три четвртине) узорка ће бити мање од q_3, па се он назива трећи квартил узорка.
- Израчунати $\mathrm{IQR} = q_3 - q_1$.

```
> IQR(x)
```

Боксплот

Боксплот (енгл. box - кутија) је дијаграм који нам визуелно обједињује мере положаја, расејања и степен померености расподеле и омогућава нам откривање аутлајера.

Аутлајер је податак који се не уклапа у нас модел, тј. оступа од правила уочених за остатак узорка. Понекад су аутлајер последица грешке, и у том случају их треба уклонити, а у другим случајевима треба извршити прилагођавање модела.

Цртање боксплот дијаграма

- Одредити узорачку медијану, узорачке квартиле q_1 и q_3, и међуквартилно растојање IQR
- Одредити тачке f_1 и f_3, унутрашње границе, као

$$f_1 = q_1 - 1.5 \cdot \mathrm{IQR} \text{ и } f_3 = q_3 + 1.5 \cdot \mathrm{IQR}.$$

- Одредити ивичне вредности a_1 и a_3 тако да је a_1 најближа вредност из узорка до f_1 која није мања од f_1, а a_3 најближа вредност из узорка до f_3 која није већа од f_3.
- Одредити тачке F_1 и F_3, спољашње границе, као

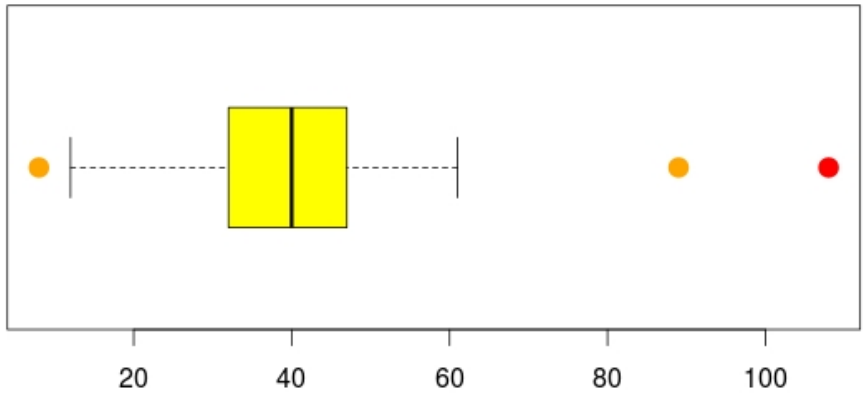
$$F_1 = q_1 - 3 \cdot \mathrm{IQR} \text{ и } F_3 = q_3 + 3 \cdot \mathrm{IQR}.$$

- Нацртати правоугаоник с крајевима у q_1 и q_3, и унутрашњом линијом на медијани
- Повезати ивичне вредности с правоугаоником. Обележити благе аутлајере, тј. све тачке између унутрашњим и спољашјих граница, као и екстремне аутлајере, тј. све тачке изван спољашњих граница.

Боксплот – пример

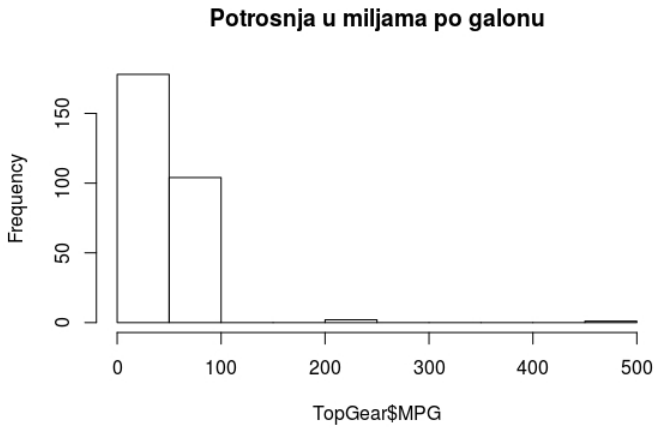
Дужина (у данима) болничког лечења пацијената с амнезијом

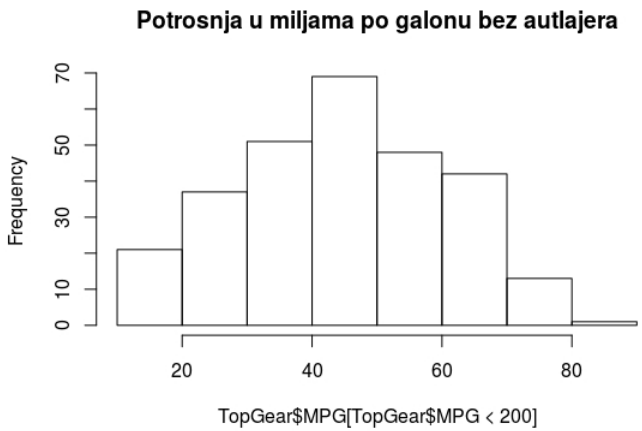
8
12
20 27
30 32 35 36
40 40 40 40 41 42 45 47
50 52
61
89
108

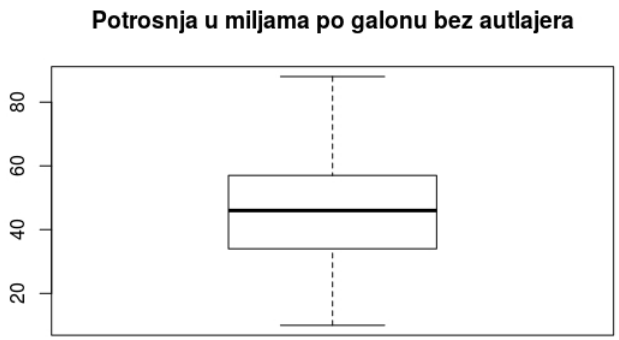



```
> amnezija<-c(8,12,20,27,30,32,35,36,40,40,40,40,41,42,45,47,50,52,61,
89,108)
> boxplot(amnezija,horizontal=T)
> points(amnezija[1],order(amnezija[1]),pch=19,col="orange",lwd=2)
> points(amnezija[20],order(amnezija[20]),pch=19,col="orange",lwd=2)
> points(amnezija[21],order(amnezija[21]),pch=19,col="red",lwd=2)
```

Утицај аутлајера на графички приказ



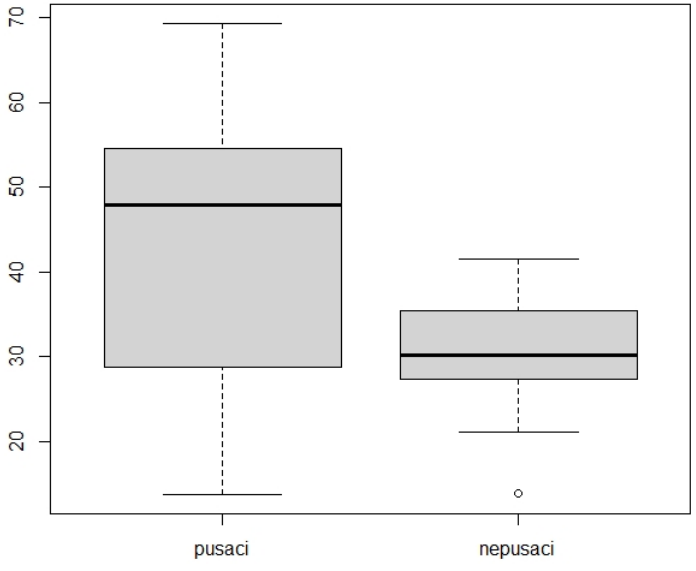




Коришћење боксплота за упоређивање

Проучаван је утицај пушења на спавање. Подаци представљају време у минутима које је било потребно да испитаници заспу.

```
x<-c(69.3,56.0,22.1,47.6,53.2,48.1,52.7,34.4,60.2,43.8,23.2,13.8)
y<-c(28.6,25.1,26.4,34.9,29.8,28.4,38.5,30.2,30.6,31.8,41.6,21.1,36.0,
37.9,13.9)
b1<-boxplot(x,y, names=c("pusaci", "nepusaci"))
b1$stats
```




```
      [,1]  [,2]
[1,] 13.80 21.10
[2,] 28.80 27.40
[3,] 47.85 30.20
[4,] 54.60 35.45
[5,] 69.30 41.60
```

$\leftarrow a_1$
$\leftarrow q_1$
$\leftarrow q_2$ (медијана)
$\leftarrow q_3$
$\leftarrow a_3$

Шта је вероватноћа?

- Вероватноће су бројеви који се налазе између 0 и 1 укључујући и њих. Често се изражавају и у процентима.
- Вероватноће близу нуле указују на то да су мале шансе да се тај догађај догоди. То не значи да се он неће догодити, већ смо да се сматра ретким.
- Вероватноће близу јединице указују на то да су велике шансе да се тај догађај догоди. То не значи да ће се он догодити, већ смо да се сматра уобичајеним.
- Вероватноће близу $1/2$ указују на то да догађај има приближни исту шансу да се догоди и да се не догоди.

Како доделити вероватноће?

1. Субјективно
2. Класично (математички)
3. Статистички

Субјективна вероватноћа: На вечерашњој утакмици вероватноћа да наши победе је 75%.

Класична дефиниција вероватноће

Дефиниција

Нека се изводи експеримент у коме је сваки од његових исхода једнако вероватан. Нека је $n(A)$ број начина на које се може догодити догађај A, а n укупан број исхода експеримента. Тада је

$$P(A) = \frac{n(A)}{n}.$$

У фиоци имамо 25 идентичних батерија од којих су 4 истрошене. На случајан начин узимамо једну батерију. Колика је вероватноћа да је исправна?
$n = 25$, $n(A) = 21$, $P(A) = \frac{21}{25}$.

Статистичка дефиниција вероватноће

Дефиниција

$$P(A) = \frac{\textit{број експеримената у којима се догађај } A \textit{ догодио}}{\textit{укупни број изведених експеримената}}$$

Вероватноћа да је трудноћа близаначка је $\frac{1}{96}$.

Извођење експеримента – бацање једне коцкице.

```
> eksperiment<-sample(c(1,2,3,4,5,6),1000,replace=TRUE)
> table(eksperiment)
   eksperiment
   1    2    3    4    5    6
 151  173  172  160  177  167
```

Статистичка вероватноћа добијања јединице је $\frac{151}{10000} = 0.151$.

Класична вероватноћа истог догађаја је $\frac{1}{6} \approx 0.167$.

Класична вероватноћа – неједнако вероватни исходи

Дефиниција

Нека се изводи експеримент чији могући исходи имају вероватноће редом $p_1, \ldots, p_n$. Вероватноћа догађаја A једнака је збиру вероватноћа исхода који реализују догађај A.

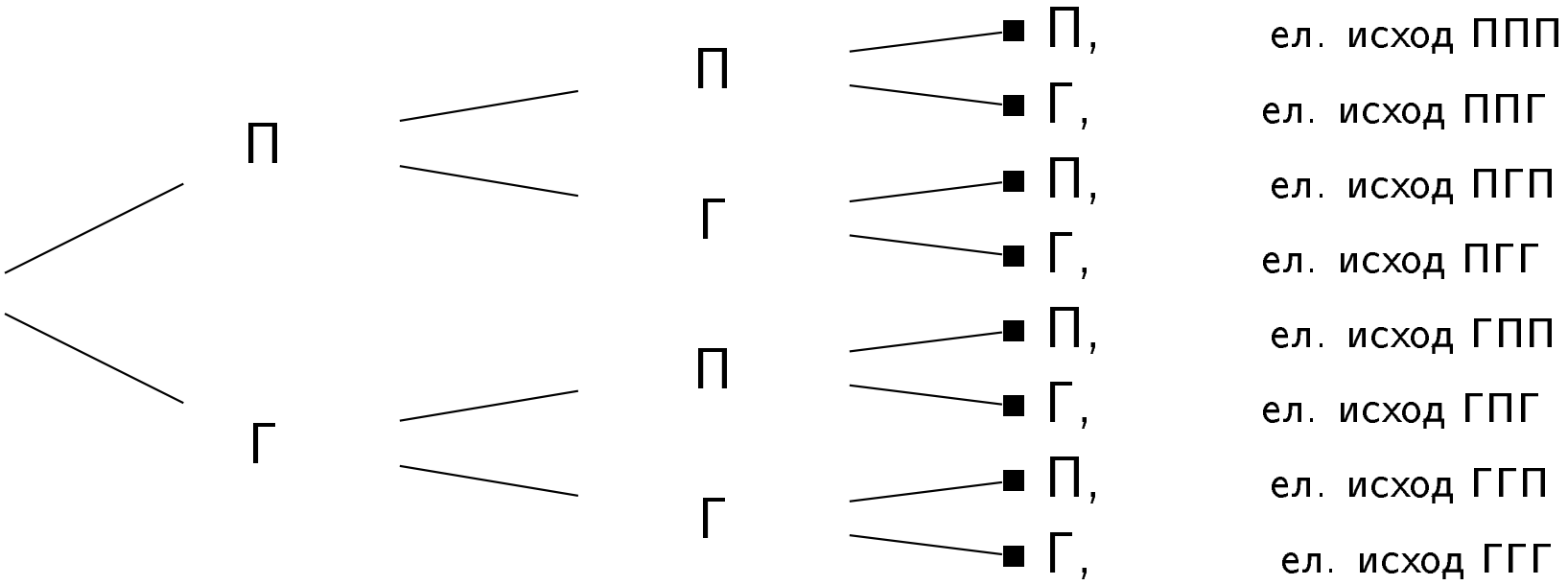
Баца се кутија шибица. Вероватноће добијања њених шест страна су: по $\frac{1}{20}$ за две странице најмање површине, по $\frac{1}{10}$ за две странице "средње" површине, и по $\frac{7}{20}$ за две странице највеће површине. Вероватноћа да шибица не падне на страницу највеће површине је $P(A) = 2 \cdot \frac{1}{20} + 2 \cdot \frac{1}{10} = \frac{3}{10}$.

```
> sibice<-sample(c("a1","a2","b1","b2","c1","c2"),10000,replace=TRUE,
prob=c(0.05,0.05,0.1,0.1,0.35,0.35))
> table(sibice)
   sibice
  a1   a2   b1   b2   c1   c2
 526  507  998 1029 3444 3496
```

$$P(A) \approx \frac{526+507+998+1029}{10000} = 0.306$$

Класична вероватноћа – дијаграми гранања

Сложеније експерименте можемо посматрати у етапама и приказати их на дијаграму гранања.



Слика: Бацање три новчића

Исходи и догађаји

- **Случајни експеримент** је било која појава или процес чији исход не можемо предвидети са сигурношћу.
- **Скуп елементарних исхода** Ω је скуп могућих исхода случајног експеримента. Сваки његов члан назива се **елементарни исход**.
- Сваки подскуп скупа елементарних исхода назива се **догађај**.
- Сам скуп Ω назива се **сигуран догађај**. Празан скуп назива се **немогућ догађај**.

Исходи и догађаји

Скуп елементарних исхода Ω приликом бацања две коцке

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

- A – збир је 7; $P(A) = \frac{6}{36}$
- B – збир је 12; $P(B) = \frac{1}{36}$
- C – збир је 13; $P(C) = 0$
- D – оба броја су мања од 7; $P(D) = P(\Omega) = 1$

Примери скупова исхода

Извлачи се једна карта из стандардог шпила од 52 карте (без џокера). Потенцијални скупови ел. исхода:

- $\Omega_1 = \{\text{црвена, црна}\}$
- $\Omega_2 = \{\clubsuit, \diamondsuit, \heartsuit, \spadesuit\}$
- $\Omega_3 = \{A\clubsuit, A\diamondsuit, A\heartsuit, A\spadesuit, \ldots, K\clubsuit, K\diamondsuit, K\heartsuit, K\spadesuit\}$ (свака карта понаособ)
- $\Omega_4 = \{\text{слика (краљ, дама, жандар), није слика}\}$
- $\Omega_5 = \{\text{слика, карта с бројем}\}$
- $\Omega_6 = \{\text{слика, ас, није слика}\}$

$\Omega_1, \Omega_2, \Omega_3, \Omega_4$ јесу скупови исхода; Ω_5 није – нема исхода који одговара асу; Ω_6 није – асу одговара више од једног исхода.

Операције над догађајима

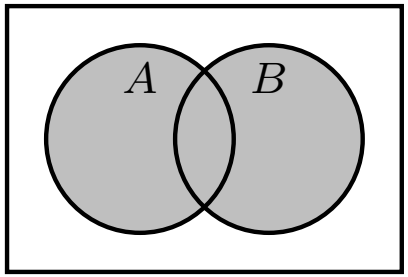
- Унија два догађаја $A \cup B$ садржи све елементарне исходе који се налазе у бар једном од догађаја A или B, тј. у A, у B, или у оба.
- Пресек два догађаја $A \cap B$, или краће AB садржи све елементарне исходе који се налазе и у A и у B.
- Комплемент $\bar{A}$ догађаја A садржи све елементарне исходе који се не налазе у A.

Дефиниција

За два догађаја, A и B, кажемо да су међусобно искључива уколико се не могу истовремено догодити, тј. ако им је пресек немогућ догађај $AB = \emptyset$.

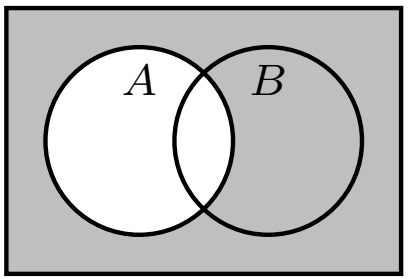
Операције над догађајима

Унија $A \cup B$



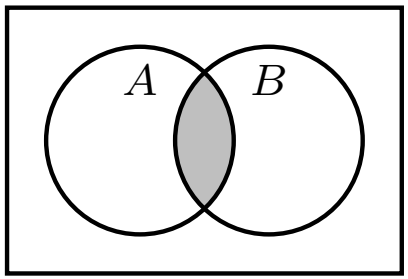
“у A или у B”

Комплемент $\bar{A}$

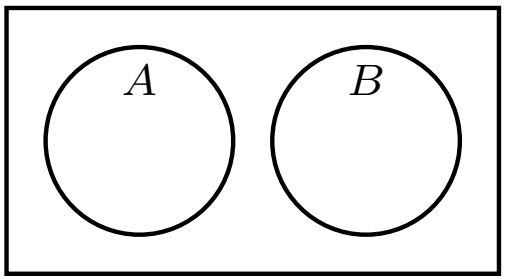


“не у A”

Пресек $A \cap B$ (AB)



“у A и у B”



Међусобно искључиви догађаји

Неке особине вероватноће

Основна својства вероватноће (аксиоме)

- $P(\Omega) = 1$
- $P(A) \geq 0$ за сваки догађај A.
- Ако су догађаји A_1, A_2, A_3, ... међусобно искључиви, онда је

$$P(A_1 \cup A_2 \cup A_3 \cdots) = P(A_1) + P(A_2) + P(A_3) + \cdots$$

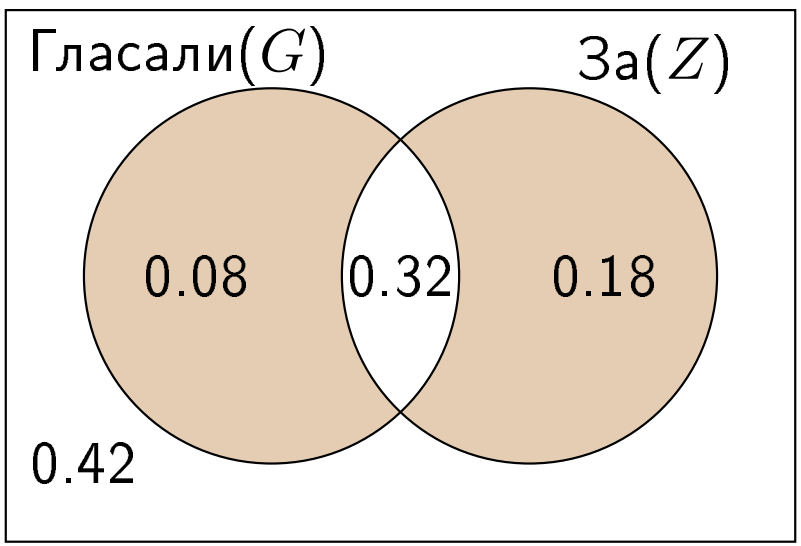
Још својстава вероватноће

- $P(\emptyset) = 0$
- $P(\bar{A}) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(AB)$

Пример

Организује се студентски референдум о изградњи новог терена. Пре гласања, 50% су за (Z) ту изградњу. На гласање (G) је изашло само 40% студената. Укупно је 32% студената гласало "за" (GZ).

Вероватноћа да је случајно изабрани студент гласао или био за је $P(G \cup Z) = P(G) + P(Z) - P(GZ) = 0.4 + 0.5 - 0.32 = 0.58$



Условна вероватноћа

- Колика је вероватноћа да је број добијен на коцкици мањи од 4?
- Колика је вероватноћа да је број добијен на коцкици мањи од 4 ако се зна да је непаран?

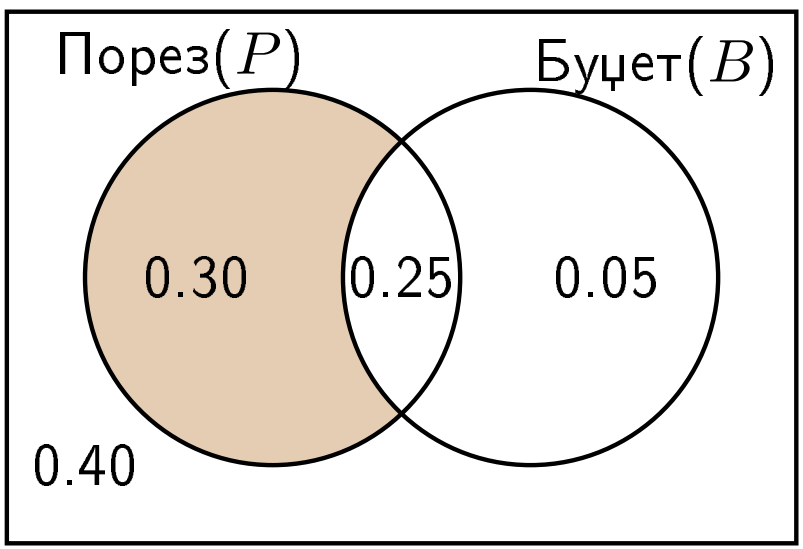
Дефиниција

Нека су A и B догађаји такви да је $P(B) > 0$. **Условна вероватноћа** *догађаја A, под условом оствареног догађаја B је количинику вероватноће да се оба догађаја остваре и вероватноће да се оствари услов B:*

$$P(A|B) = \frac{P(AB)}{P(B)}.$$

Условна вероватноћа

- Колика је вероватноћа да је број добијен на коцкици мањи од 4 ако се зна да је непаран?
  $B = \{1, 3, 5\}$, $AB = \{1, 3\}$, $P(A|B) = \frac{P(AB)}{P(B)} = \frac{n(AB)}{n(B)} = \frac{2}{3}$.
- У парламенту, у циљу сузбијања инфлације, 55% посланика је за смањење одређених пореза, 30% за смањење буџета, а 25% за обе мере. Колика је вероватноћа да је случајно изабрани посланик за смањење буџета, ако знамо да је он за смањење пореза? А колика да је за смањење пореза ако знамо да је против смањења буџета?



$$P(B|P) = \frac{P(BP)}{P(P)} = \frac{0.25}{0.55} = \frac{5}{11}$$

$$P(P|\bar{B}) = \frac{P(P\bar{B})}{P(\bar{B})} = \frac{0.30}{0.70} = \frac{3}{7}$$

Независност догађаја

Два догађаја сматрамо независним уколико остварење једног од њих нема никакав утицај на вероватноћу другог догађаја.

Дефиниција

Нека су A и B догађаји такви да је $P(B) > 0$. За догађаје A и B кажемо да су **независни** *уколико за њих важи да је*

$$P(A|B) = P(A).$$

- Бацају се плава и црвена коцкица. Дати су догађаји: A – добијени су исти бројеви; B – на црвеној је двојка или тројка.
  $P(A) = \frac{6}{36}$, $P(B) = \frac{12}{36}$, $P(AB) = \frac{2}{36}$, $P(A|B) = \frac{2/36}{12/36} = \frac{2}{12}$.
  $P(A|B) = P(A)$, па су догађаји A и B независни.

Независност догађаја

Теорема

Ако су A и B независни, тада је

$$P(AB) = P(A)P(B).$$

- У Америци око 46% људи има крвну групу O, а око 39% негативан Rh-фактор. Ова два обележја сматрају се независним. Колика је вероватноћа да случајно изабрани Американац има крвну групу O^-?

  N – догађај да он има негативан Rh-фактор

  $$P(O^-) = P(O \cap N) = P(O) \cdot P(N) = 0.46 \cdot 0.39 = 0.179 \approx 18\%.$$

Независност догађаја и поузданост система

Размотримо систем од k компоненти у серији. Претпоставимо да су ове компоненте независне у смислу да поузданост једне не утиче на отказивање или успех било које друге. Поузданост целог система је вероватноћа да систем ради када треба. Систем ће бити исправан само ако све компоненте раде добро.

$$R = P\{\text{све компоненте раде исправно}\} = R_1 \cdot R_2 \cdots R_k.$$

Нека имамо систем од 5 компоненти и нека свака компонента има поузданост 0.95 у одређеном тренутку:

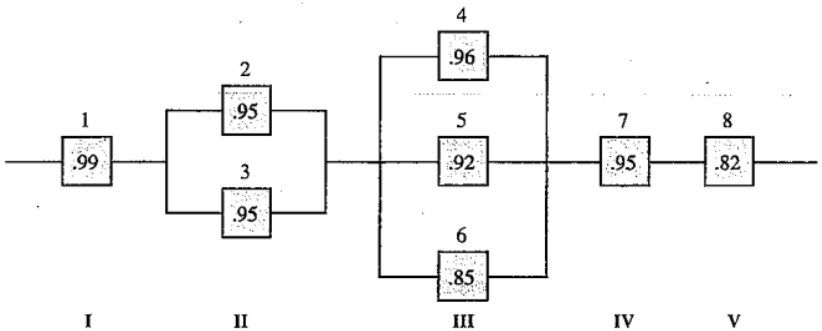
$$R = 0.95^5 = 0.774.$$

Размотримо сада k независних компоненти повезаних у паралелан систем. Поузаност система у

$$\begin{aligned} R &= 1 - P\{\text{све компоненте су отказале}\} \\ &= 1 - (1 - R_1) \cdot (1 - R_2) \cdots (1 - R_k). \end{aligned}$$

Независност догађаја и поузданост система

Размотримо сада систем од 8 независних компоненти повезаних као на слици.



Систем се састоји од 5 група у серији. Да би се одредила поузданост система, прво одређујемо поузданост паралелних група:

$$R_2 = 1 - (1 - 0.95) \cdot (1 - 0.95) = 1 - 0.05 \cdot 0.05 = 0.9975,$$

$$R_3 = 1 - (1 - 0.96) \cdot (1 - 0.92)(1 - 0.85) = 1 - 0.04 \cdot 0.08 \cdot 0.15 = 0.99952.$$

Поузданост целог система је

$$R = R_1 \cdot R_2 \cdot R_3 \cdot R_4 \cdot R_5 = 0.99 \cdot 0.9975 \cdot 0.99952 \cdot 0.95 \cdot 0.82 = 0.7689.$$

Вероватноћа пресека зависних догађаја

Теорема

Нека су A и B догађаји такви да је $P(B) > 0$. Тада важи

$$P(AB) = P(A|B)P(B).$$

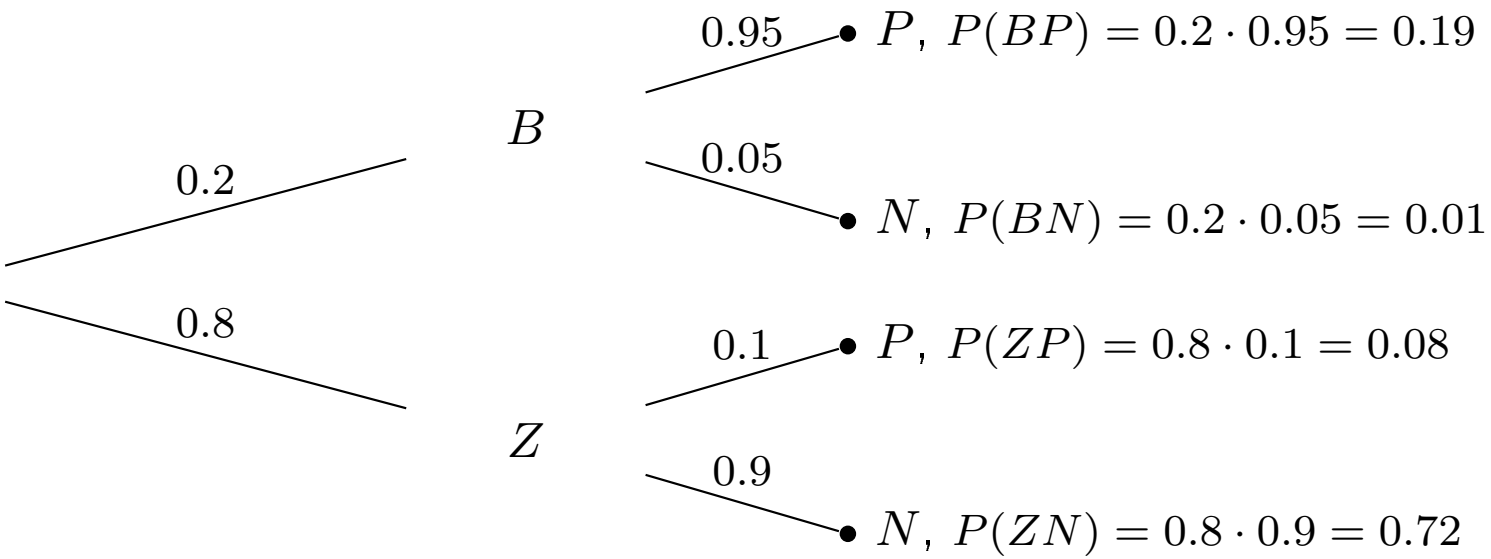
- У Америци око 46% људи има крвну групу O, а у регистрима је 4% оних који имају O грешком забележено као A. Колика је вероватноћа да случајно изабрани Американац стварно има O, али су му забележили A?

  O – има O крвну групу; A – забележено му је A. Дато нам је $P(O) = 0.46$ и $P(A|O) = 0.04$.

  $$P(O \cap A) = P(O) \cdot P(A|O) = 0.46 \cdot 0.04 = 0.018 \approx 2\%.$$

Формула потпуне вероватноће

- Тест на једну болест је такав да 95% болесних има позитиван резултат, а 90% здравих има негативан резултат. Ако 20% пацијената има ту болест, колика је вероватноћа да ће случајно изабраном пацијенту тест бити позитиван?



$P(P) = 0.19 + 0.08 = 0.27$ што је добијено као

$P(P) = P(B)P(P|B) + P(Z)P(P|Z)$

Формула потпуне вероватноће

Теорема (Формула потпуне вероватноће)

Нека су $A_1, \ldots, A_n$ међусобно искључиви догађаји чија је унија скуп Ω и нека је B било који догађај. Тада је

$$P(B) = P(A_1) \cdot P(B|A_1) + \cdots + P(A_n) \cdot P(B|A_n).$$

- Испитаник баца новчић и ако падне писмо, одговара на питање А) "Да ли сте рођени парне године?", а ако падне глава, одговара на питање Б) "Да ли сте пробали дрогу?" Од 500 испитаника 350 је одговорило да. Проценити проценат оних који су пробали дрогу.

  Знамо да је $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{2}$, $P(D|A) = \frac{1}{2}$ и $P(D) \approx \frac{350}{500} = \frac{7}{10}$.

$$P(D) = P(A)P(D|A) + P(B)P(D|B)$$

$$\frac{7}{10} = \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot P(D|B) \quad \Rightarrow \quad P(D|B) = \frac{9}{10} = 90\%.$$

Бајесова формула

- Тест на ретку болест коју има 0.1% популације је такав да 99% болесних има позитиван резултат, а 95% здравих има негативан резултат. Ако је неко позитиван на тесту, колика је вероватноћа да је болестан?

$$\begin{aligned} P(B|P) &= \frac{P(BP)}{P(P)} = \frac{P(B)P(P|B)}{P(B)P(P|B) + P(Z)P(P|Z)} \\ &= \frac{0.001 \cdot 0.99}{0.001 \cdot 0.99 + 0.999 \cdot 0.05} = \frac{0.00099}{0.05094} \\ &= 0.01943 \approx 2\% \end{aligned}$$

Теорема (Бајесова формула)

Нека су $A_1, \ldots, A_n$ међусобно искључиви догађаји чија је унија скуп Ω и нека је B било који догађај. Тада је за сваки A_i

$$P(A_i|B) = \frac{P(A_i) \cdot P(B|A_i)}{P(A_1) \cdot P(B|A_1) + \cdots + P(A_n) \cdot P(B|A_n)}.$$

Пребројавање

- За рачунање класичне вероватноће треба знати укупан број исхода и број начина реализације догађаја
- За експерименте с великим бројем исхода постоје методи за пребројавање исхода тражених догађаја
- Ако се експеримент може поделити у етапе, онда је број исхода једнак производу броја исхода у свакој етапи
- Студент треба да изабере три изборна предмета. Први бира од три понуђене природне науке, други од четири друштвене науке, а трећи од пет спортова. На колико начина он то може да уради?

  $3 \cdot 4 \cdot 5 = 60.$
- Приликом бацања пет коцкица на колико начина се може добити исход с најмање два различита броја?

  $6 \cdot 6 \cdot 6 \cdot 6 \cdot 6 - 6 = 7770.$

Пермутације

Дефиниција

Пермутације су низови објеката у одређеном редоследу.

- На колико начина се 8 спринтера може поставити на стартну линију?
  То је број пермутација од 8 елемената. Први има 8 места, други преосталих 7, итд.

$$8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 8! = 40320$$

- $n! = n(n-1) \cdots 2 \cdot 1; \quad 0! = 1.$

Пермутације

- Колико има пермутација речи БАБА?
  ААББ, АБАБ, АББА, БААБ, БАБА, ББАА
  Кад би слова била различита било би $4!$. Пошто имамо две групе с по два иста слова делимо с $2! \cdot 2!$.

$$\frac{4!}{2!2!} = 6.$$

Теорема

Имамо n објеката у k група, а унутар сваке групе објекти су идентични. Нека је n_j број објеката у j-тој групи, где је $j = 1, 2, \ldots, k$ и $n_1 + n_2 + \cdots + n_k = n$. Број пермутација таквих n објеката је

$$\frac{n!}{n_1! n_2! \cdots n_k!}.$$

Комбинације

Дефиниција

Комбинације су скупови објеката без одређеног редоследа.

- На колико начина можемо изабрати 3 волонтера од 5 пријављених?
  Обележимо их бројевима од 1 до 5. Могуће комбинације су:
  1,2,3 1,2,4 1,2,5 1,3,4 1,3,5
  1,4,5 2,3,4 2,3,5 2,4,5 3,4,5
  Има их $\frac{5\cdot 4\cdot 3}{3!} = \frac{60}{6} = 10.$

Теорема

Број комбинација r објеката изабраних од n различитих објеката $\binom{n}{r}$ је

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}.$$

Комбинације

- Колика је вероватноћа да случајно подељених 5 карата садрже тачно два аса?

  A – 5 подељених карата садрже тачно два аса

  Треба пребројати укупан број комбинација од 5 карата, као и број комбинација које садрже два аса.

  Укупан број комбинација:

$$n = \binom{52}{5} = \frac{52!}{5!47!} = \frac{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48 \cdot 47!}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 \cdot 47!} = 2598960.$$

  Два аса (од 4 могућа) можемо добити на $\binom{4}{2}$ начина. Преостале три карте нису асови и можемо их добити на $\binom{48}{3}$ начина.

$$n(A) = \binom{4}{2}\binom{48}{3} = 6 \cdot 17296 = 103776,$$

$$P(A) = \frac{n(A)}{n} = \frac{103776}{2598960}.$$

Случајне променљиве

Случајна променљива је променљива чије се вредности одређују исходом случајног експеримента. Обележавамо их словима $X, Y, Z, \ldots$

- Бацање две коцкице — X - збир добијених бројева
- Рулет (38 поља, од тога 18 црвених, 18 црвених и 2 зелена) - играч игра сваки пут на зелено — Y - број игара до добитка
- Полицијска станица — Z - време првог позива између 7:30 и 8:00 ујутру
- W - дужина извршавања одређеног рачунарског програма

Дискретне и непрекидне случајне променљиве

Дискретне случајне променљиве су случајне променљиве које могу узети коначно или пребројиво бесконачно много могућих вредности.

- Збир бројева на коцкицама X може бити $2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12$ — коначно много вредности
- Број игара до добитка на рулету Y може бити $1, 2, 3, 4, \ldots$ (није ограничено) — пребројиво бесконачно много вредности

Непрекидне случајне променљиве су случајне променљиве које могу узети вредности с неког интервала реалних бројева, а вероватноћа да узму конкретну вредност је нула.

- Време првог позива у полицији Z може узети било коју вредност из интервала (7:30, 8:00)
- Дужина извршавања рачунарског програма W може узети било коју вредност из интервала $(0, t)$, где је t време за које се програм сигурно извршава

Дискретне случајне променљиве

Дефиниција

Нека је X дискретна случајна променљива. Њена **расподела** *вероватноће је*

$$f(x) = P\{X = x\} \text{ за сваку вредност } x.$$

Теорема (Својства расподеле)

Свака дискретна расподела мора да задовољава

1) $f(x) \geq 0$ *за сваки реалан број* x
2) $\sum f(x) = 1.$

Дискретне случајне променљиве

- Трговац на берзи посматра одређених 5 деоница. Нека је X број деоница којима ће сутра порасти цена. Расподела за X је

x	0	1	2	3	4	5
$P\{X = x\} = f(x)$	?	0.30	0.20	0.10	0.05	0.01

Колика је вероватноћа да ће већини деоница сутра порасти цена?

Да би укупан збир вероватноћа био 1, мора бити $P\{X = 0\} = 0.34$.
Већина деоница значи 3, 4 или 5 деоница.
$P\{X \geq 3\} = P\{X = 3\} + P\{X = 4\} + P\{X = 5\} =$
$0.10 + 0.05 + 0.01 = 0.16.$

Приметимо да је
$P\{X > 3\} = P\{X = 4\} + P\{X = 5\} = 0.06 \neq P\{X \geq 3\}$

Код дискретних расподела мора се пазити да ли је граница укључена или не ($>$ није исто што и $\geq$)!

Дискретне случајне променљиве

- У игри “крепс” бацају се коцкице и играч побеђује у првом бацању уколико добије збир 7 или 11. Колика је вероватноћа да он победи у првом бацању?
  Расподела за X, збир добијених бројева је

$$X : \begin{pmatrix} 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\ \frac{1}{36} & \frac{2}{36} & \frac{3}{36} & \frac{4}{36} & \frac{5}{36} & \frac{6}{36} & \frac{5}{36} & \frac{4}{36} & \frac{3}{36} & \frac{2}{36} & \frac{1}{36} \end{pmatrix}$$

Краће се може записати као

$$f(x) = \begin{cases} \frac{x-1}{36}, & \text{ако је } x = 2, 3, 4, 5, 6, 7 \\ \frac{13-x}{36}, & \text{ако је } x = 8, 9, 10, 11, 12. \end{cases}$$

Из расподеле имамо да је

$$P(\text{победа у првом бацању}) = f(7) + f(11) = \frac{6}{36} + \frac{2}{36} = \frac{8}{36} = \frac{2}{9}.$$

Независност случајних променљивих

Дефиниција

За случајне променљиве X и Y кажемо да су **независне** *уколико је сваки догађај везан за X независан од сваког догађаја везаног за Y, односно ако важи*

$$P\{X = x | Y = y\} = P\{X = x\} \text{ за свако } x \text{ и свако } y.$$

Мере положаја и расејања

Параметри популације

- средња вредност популације μ
- дисперзија популације σ^2
- стандардно одступање популације σ

Како их повезујемо са случајном променљивом?

- $\mu = EX$ математичко очекивање случајне променљиве X
- $\sigma^2 = DX$ дисперзија случајне променљиве X
- $\sigma = \sqrt{DX}$ стандардно одступање случајне променљиве X

Математичко очекивање

Математичко очекивање или очекивана вредност EX, случајне променљиве X представља дугорочну теоретску просечну вредност за X.

- Баца се једна коцкица и X је број добијен на њој. Рецимо да смо понављали експеримент n пута и добили нпр. следеће вредности:
  $1, 3, 2, 5, 2, 1, 1, 6, 5, 4, 2, 3, 6, 4...$

  Ако после сваког бацања рачунамо дотадашњи просек добијамо низ просека
  $1, 2, 2, 2.75, 2.6, 2.33, 2.14, 2.63, 2.89, 3.0, 2.91, 2.92, 3.15, 3.21...$

  Ако наставимо вредности ће бити све приближније једнаке EX.

Математичко очекивање

- Ако бацамо коцкицу велики број пута n, приближно у једној шестини од n бацања добићемо 1, исто важи и за остале бројеве. Тако да ће просек бити приближно једнак

$$\frac{\frac{n}{6}\cdot 1+\frac{n}{6}\cdot 2+\frac{n}{6}\cdot 3+\frac{n}{6}\cdot 4+\frac{n}{6}\cdot 5+\frac{n}{6}\cdot 6}{n}$$
$$=\frac{1}{6}\cdot 1+\frac{1}{6}\cdot 2+\frac{1}{6}\cdot 3+\frac{1}{6}\cdot 4+\frac{1}{6}\cdot 5+\frac{1}{6}\cdot 6$$
$$=3.5$$

Дефиниција

Нека је X дискретна случајна променљива. Тада је

$$EX=\sum xf(x).$$

Математичко очекивање

Рачунање математичког очекивања случајних променљивих $g(X)$ које су функције од X (нпр. X^2, $X+1$, $(3X-2)^2$, итд.)

$$Eg(X) = \sum g(x)f(x).$$

- Рачунамо математичко очекивање квадрата броја добијеног на коцкици

$$\begin{aligned} EX^2 &= \sum x^2 f(x) \\ &= 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + 3^2 \cdot \frac{1}{6} + 4^2 \cdot \frac{1}{6} + 5^2 \cdot \frac{1}{6} + 6^2 \cdot \frac{1}{6} \\ &= \frac{91}{6} \end{aligned}$$

Дисперзија

Дефиниција

Нека је X дискретна случајна променљива. Њена **дисперзија** *DX је*

$$DX = E(X - EX)^2.$$

Теорема (Формула за рачунање дисперзије)

$$DX = EX^2 - (EX)^2.$$

- Дате су случајне променљиве

$$X : \begin{pmatrix} 15 & 45 & 75 \\ 0.4 & 0.20 & 0.40 \end{pmatrix} \text{ и } Y : \begin{pmatrix} 43 & 44 & 45 & 46 & 47 \\ 0.025 & 0.05 & 0.85 & 0.05 & 0.025 \end{pmatrix}.$$

- Можемо израчунати $EX = 45$, а такође и $EY = 45$. Иако су очекивања иста, расподеле се драстично разликују!
- Рачунамо дисперзије

$$\begin{aligned} DX &= E(X - EX)^2 = E(X - 45)^2 \\ &= (15 - 45)^2 \cdot 0.40 + (45 - 45)^2 \cdot 0.20 + (75 - 45)^2 \cdot 0.40 \\ &= 360 + 0 + 360 = 720. \\ DY &= E(Y - EY)^2 = E(Y - 45)^2 \\ &= (43 - 45)^2 \cdot 0.025 + (44 - 45)^2 \cdot 0.05 + (45 - 45)^2 \cdot 0.85 \\ &+ (46 - 45)^2 \cdot 0.05 + (47 - 45)^2 \cdot 0.025 \\ &= 0.1 + 0.05 + 0 + 0.05 + 0.1 = 0.3. \end{aligned}$$

- Дисперзије нам указују на суштинску разлику у расподелама

Дисперзија

- Други начин:

$$
\begin{aligned}
EX^2 &= \sum x^2 f(x) \\
&= 15^2 \cdot 0.40 + 45^2 \cdot 0.20 + 75^2 \cdot 0.40 = 2745 \\
EY^2 &= \sum y^2 f(y) \\
&= 43^2 \cdot 0.025 + 44^2 \cdot 0.05 + 45^2 \cdot 0.85 + 46^2 \cdot 0.05 \\
&+ 47^2 \cdot 0.025 = 2025.3 \\
DX &= EX^2 - (EX)^2 = 2745 - 45^2 = 2745 - 2025 = 720 \\
DY &= EY^2 - (EY)^2 = 2025.3 - 45^2 = 2025.3 - 2025 = 0.3.
\end{aligned}
$$

Особине математичког очекивања и дисперзије

Теорема (Особине математичког очекивања)

Нека су X и Y случајне променљиве и нека је c било који реалан број. Тада важи:

- $Ec = c$;
- $E(cX) = cEX$;
- $E(X + Y) = EX + EY$.

Теорема (Особине дисперзије)

Нека су X и Y случајне променљиве и нека је c било који реалан број. Тада важи:

- $Dc = 0$;
- $D(cX) = c^2 DX$.

Ако су X и Y независне случајне променљиве, онда важи:

- $D(X + Y) = DX + DY$.

Биномна расподела

- Тест с 5 питања и по 4 понуђена одговора — Студент случајно бира одговор — X – број тачних одговора
- Вероватноћа да здраво дете добије заушке у контакту с оболелим дететом је 10% — 15 здраве деце дошло је у контакт с оболелим — Y – број деце која су се разболела
- 20 људи анкетирано је у вези предлога владе — у целој популацији 70% подржава овај предлог — Z – број анкетираних који подржавају предлог
- Експеримент се састоји из фиксног и познатог број етапа n
- У свакој етапи имамо два исхода: “успех“ и “неуспех“
- Исход у једној етапи не утиче на исход у другој, ондосно етапе су независне и вероватноће успеха су исте у свакој етапи
- Случајна променљива од интереса је број ”успеха” у n етапа

Биномна расподела

- Имамо n експеримената и у сваком посматрамо да ли се догодио одређени догађај који називамо успехом. Експерименти су међусобно независни и вероватноћа успеха у сваком од њих је p. За случајну променљиву која представља број успеха у n оваквих експеримената кажемо да има **биномну расподелу** с параметрима n и p.

Теорема

Нека случајна променљива X има биномну расподелу с параметрима n и p. Тада је њена расподела

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \text{ за } x = 0, 1, 2, \ldots, n.$$

- Студент одговара случајно да један од четири понуђена одговора. На тесту има пет питања. Колика је вероватноћа да ће имати тачно три тачна одговора? Колика је да ће имати највише три тачна одговора? А колика да ће имати бар четири тачна одговора?

Нека је X број тачних одговора. Расподела за X је

$$f(x) = \binom{5}{x}\left(\frac{1}{4}\right)^x\left(\frac{3}{4}\right)^{5-x}, \quad x = 0, 1, 2, 3, 4, 5.$$

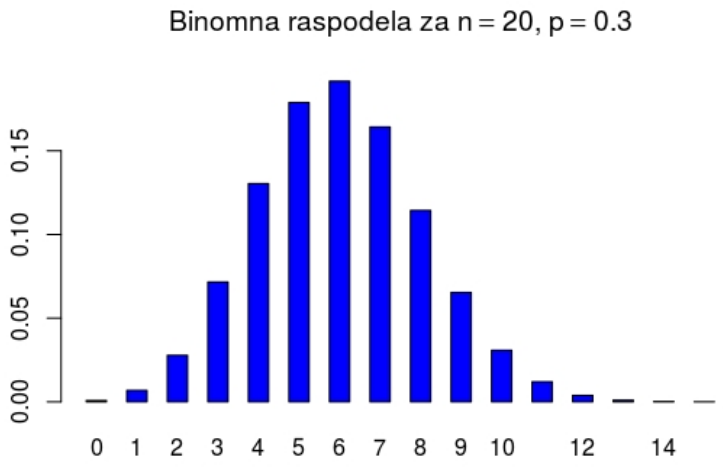
$$P\{X = 3\} = f(3) = \binom{5}{3}\left(\frac{1}{4}\right)^3\left(\frac{3}{4}\right)^2 = 10 \cdot \frac{1}{64}\frac{9}{16} = \frac{90}{1024} \approx 9\%$$

$$P\{X \leq 3\} = f(0) + f(1) + f(2) + f(3) = \frac{1008}{1024} \approx 98.4\%$$

$$P\{X \geq 4\} = 1 - P\{X < 4\} = 1 - P\{X \leq 3\} = \frac{16}{1024} \approx 1.6\%.$$

```
> dbinom(3,size=5,p=1/4)
0.08789063
> pbinom(3,size=5,p=1/4)
0.984375
> rbinom(10,size=20,p=0.3)
6 5 7 4 6 6 3 4 5 5
```

Биномна расподела




```
> x<-dbinom(0:15,size=20,p=0.3)
> barplot(x,names.arg=0:15, space=1,col="blue",main=expression(paste(
"Binomna raspodela za ",n==20,", ",p==0.3)))
```

Биномна расподела

Теорема

Нека случајна променљива X има биномну расподелу. Тада важи

$$EX = np, \quad DX = np(1-p).$$

- Анкетирано је 20 људи у вези с предлогом владе. За сваког од њих нам је 70% шансе да је "за".

  Математичко очекивање броја анкетираних који су "за" је $\mu = np = 20 \cdot 0.7 = 14$.

  Дисперзија је $\sigma^2 = np(1-p) = 20 \cdot 0.7 \cdot 0.3 = 4.2$.

  Стандардно одступање је $\sigma = \sqrt{4.2} = 2.049$.

Геометријска расподела

- Играч рулета сваки пут улаже на црвено – број изгубљених партија пре првог добитка
- Игра “Не љути се човече” – број неуспешних бацања пре појаве шестице
- Гађање у мету – број покушаја пре првог поготка у центар
- У свакој етапи имамо два исхода: “успех“ и “неуспех“
- Експеримент се одвија све до појаве првог ”успеха“
- Исход у једној етапи не утиче на исход у другој, ондосно етапе су независне и вероватноће успеха су исте у свакој етапи
- Случајна променљива од интереса је број остварених ”неуспеха“

Геометријска расподела

Теорема

Нека случајна променљива X има геометријску расподелу с параметром p. Тада је њена расподела

$$f(x) = p(1-p)^x, \text{ за } x = 0, 1, 2, \ldots.$$

- Математичко очекивање и дисперзија случајне променљиве X која има геометријску расподелу с параметром p су

$$EX = \frac{1-p}{p} \text{ и } DX = \frac{1-p}{p^2}$$

- Геометријска расподела може се дефинисати и као расподела броја изведених експеримената, што је број остварених неуспеха плус један успех. Уколико Y има овакву геометријску расподелу, важи $Y = X + 1$, где X има горенаведену геометријску расподелу. Расподела за Y је

$$f(y) = p(1-p)^{y-1}, \text{ за } y = 1, 2, \ldots.$$

Геометријска расподела

- Коцкица се баца до појаве шестице. Израчунати вероватноћу да је пре појаве шестице било четири несупешна покушаја. Колика је вероватноћа да је шестица добијена пре четвртог бацања (тј. да је пре шестице било највише два неуспешна покушаја)?

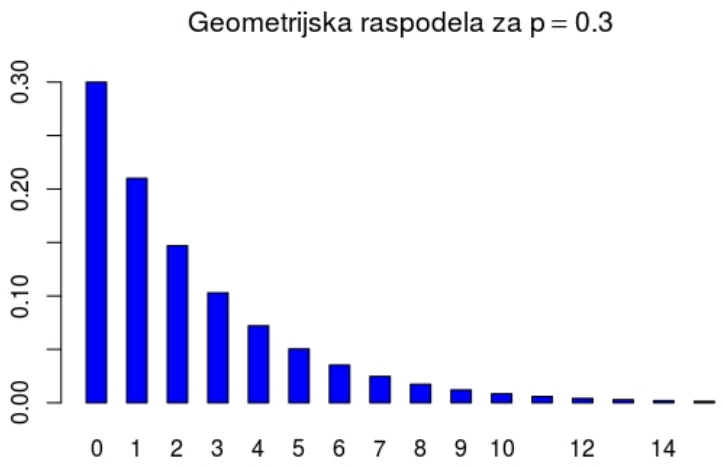
Број бацања пре добијања шестице X има геометријску расподелу с вероватноћом успеха $\frac{1}{6}$, па је

$$P\{X = 4\} = f(4) = \frac{1}{6} \cdot \left(\frac{5}{6}\right)^4 = \frac{625}{7776} = 0.0804.$$

$$P\{X < 3\} = f(0) + f(1) + f(2) = \frac{1}{6} + \frac{5}{36} + \frac{25}{216} = 0.4213.$$

```
> dgeom(4,probability=1/6)
0.08037551
> pgeom(2,probability=1/6)
0.4212963
> rgeom(10,probability=1/6)
3 1 6 0 5 5 1 3 15 19
```

Геометријска расподела




```
> x<-dgeom(0:15,prob=0.3)
> barplot(x,names.arg=0:15, space=1,col="blue",main=expression(paste(
"Geometrijska raspodela za ",p==0.3)))
```

Пуасонова расподела

- Број догађаја који се догоде за неко одређено време често представљамо Пуасоновом расподелом
- Примери: број аутомобила који прођу кроз наплатну рампу за сат времена, број људи који уђу у продавницу у току једног дана, број телефонских позива у полицијској станици у току од два сата итд.
- Пуасонова расподела има параметар λ који представља средњи (очекивани) број таквих догађаја за то време.

Дефиниција

Пуасонова расподела Случајна променљива X има Пуасонову расподелу ако је

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, \ldots,$$

где је $e \approx 2.72$.

- У полицијску станицу стиже у просеку 11 позива на сат. Колика је вероватноћа да у периоду од 7 до 7:15 ујутру неће бити позива?

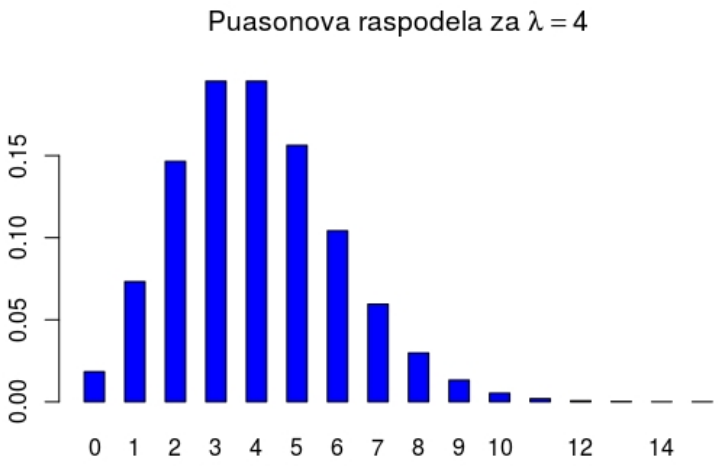
$\lambda = 11 \cdot \frac{1}{4} = 2.75.$

$$P\{X = 0\} = \frac{e^{-2.75}2.75^0}{0!} \approx 2.72^{-2.75} = 0.064.$$

```
> dpois(0,lambda=2.75)
0.06392786
> rpois(10,lambda=3)
1 4 5 2 1 6 3 5 6 3
```

- Ако X има Пуасонову расподелу с параметром λ, тада је $EX = \lambda$, а такође и $DX = \lambda$.

Пуасонова расподела




```
> x<-dpois(0:15,lambda=4)
> barplot(x,names.arg=0:15, space=1,col="blue",main=expression(paste(
"Puasonova raspodela za ",lambda==4)))
```

Рачунање биномних вероватноћа преко Пуасонових

- Уколико је n велико, а p такво да је $np \leq 10$, биномне вероватноће могу приближно да се израчунају коришћењем Пуасонових

$$\binom{n}{x} p^x (1-p)^x \approx \frac{e^{-np}(np)^x}{x!}.$$

- Контигент од 2000 флаша се превози, а за сваку флашу вероватноћа да се разбије је 0.003. Колика је вероватноћа да се разбију две флаше? А бар две флаше?

  X – број разбијених флаша; $n = 2000$ – велико; $np = 2000 \cdot 0.003 = 6 < 10$.

$$P\{X = 2\} = \binom{2000}{2} 0.003^2 (0.997)^{1997} \approx \frac{2.72^{-6} 6^2}{2!} = 0.044$$

$$P\{X \geq 2\} = 1 - P\{X = 0\} - P\{X = 1\} \approx 1 - \frac{2.72^{-6} 6^0}{0!} - \frac{2.72^{-6} 6^1}{1!} = 0.98.$$

Квалитет апроксимације

```
> dbinom(2,2000,0.003)
0.04446124
> dpois(2,6)
0.04461754
```

Непрекидне случајне променљиве

Нека је X непрекидна случајна променљива. Њена густина расподеле $f(x)$ мора да задовољава

- $f(x) \geq 0$ за свако x
- Укупна површина испод графика функције f једнака је 1.
- Вероватноћа да X узме вредност између било које две вредности a и b, $P\{a < X < b\}$ једнака је површини испод графика функције f од a до b.
- Није битно да ли су крајње тачке укључене, вероватноћа је увек иста, тј.

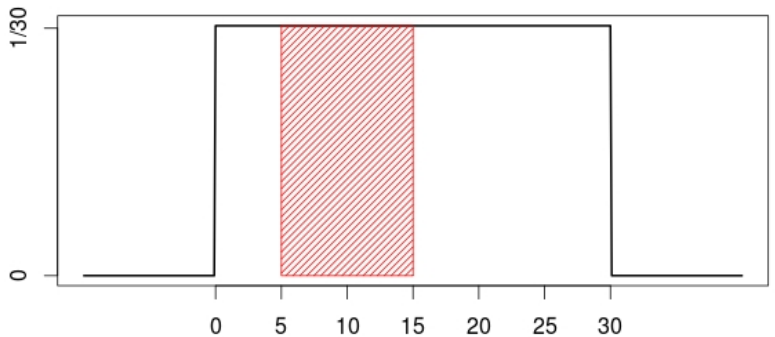
$$\begin{aligned} P\{a < X < b\} &= P\{a \leq X < b\} \\ &= P\{a < X \leq b\} \\ &= P\{a \leq X \leq b\} \end{aligned}$$

- Вероватноће код већине расподела рачунају се из таблица (или коришћењем рачунара)

Непрекидне вероватноће

- X – време првог позива у полицијској станици у првих пола сата радног времена (7:30–8:00). Ниједан период унутар ових пола сата није вероватнији од других. Колика је вероватноћа да први позив буде између 7:35 и 7:45?

Интервал када је позив могућ дуг је 30 минута – сваки део овог интервала је једнако вероватан — $f(x) = \frac{1}{30}$. Оваква расподела назива се **равномерном**.

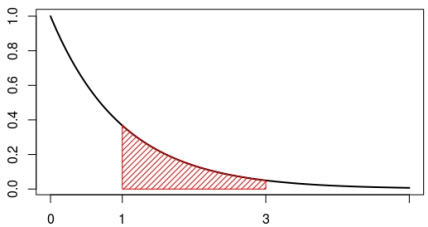


Слика: $P\{5 < X < 15\}$

$$P\{5 < X < 15\} = 10 \cdot \frac{1}{30} = \frac{1}{3}.$$

Непрекидне вероватноће

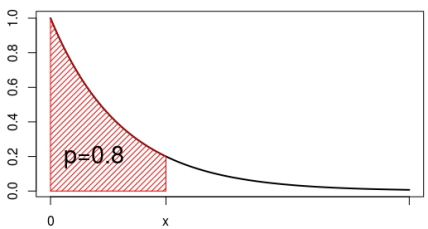
- X – животни век компоненте (у годинама) – густина расподеле $f(x) = e^{-x}$, $x > 0$. Колика је вероватноћа да се конкретна компонента поквари између 1 и 3 године коришћења? После колико времена је вероватноћа да је компонента отказала 80%?
- Ова расподела припада групи **експоненцијалних** расподела чије су густине $f(x) = \lambda e^{-\lambda x}$, $x > 0$, где је $\lambda > 0$ параметар стопе отказа (смртности).



Слика: $P\{1 < X < 3\}$

$$\int_1^3 e^{-x}dx = e^{-1} - e^{-3} = 0.318.$$

```
> pexp(3,rate=1)-pexp(1,rate=1)
0.3180924
```



Слика: $P\{0 < X < x\}$

$$\int_0^x e^{-x}dx = 1 - e^{-x} = 0.8, \quad x = -\ln 0.2 = 1.61$$

```
> qexp(0.8,rate=1)
1.609438
```

Математичко очекивање и дисперзија

- Математичко очекивање и дисперзија непрекидних променљивих дефинишу се као

$$EX = \int_{-\infty}^{\infty} x f(x) dx \quad \text{и} \quad DX = \int_{-\infty}^{\infty} (x - EX)^2 dx$$

- Математичко очекивање или средња вредност представља тежиште расподеле
- Код симетричних расподела математичко очекивање је на средини и једнако је такође и медијани (а често и моди) расподеле
- Дисперзија одређује облик расподеле, што је већа график је "плоснатији", а што је мања график је "суженији" око средње вредности

Нормална расподела

- Откривена у 18. веку као расподела грешке астрономских осматрања
- Једна од најзначајнијих расподела у анализи података, нарочито у природним наукама, медицини и инжењерству
- Већина статистичких метода праве се за податке управо из нормалне расподеле

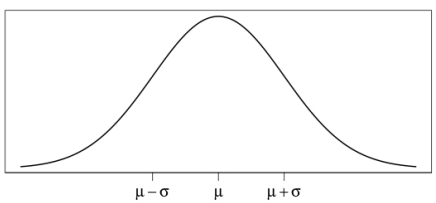
Дефиниција

Случајна променљива има нормалну расподелу $\mathcal{N}(\mu, \sigma^2)$, с математичким очекивањем μ и дисперзијом σ^2, уколико је њена густина расподеле облика

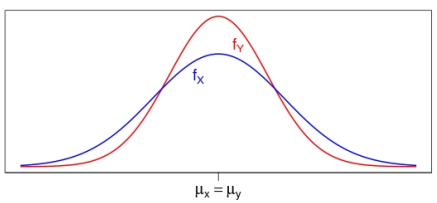
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, \quad \text{за свако реално } x.$$

Особине нормалне расподеле

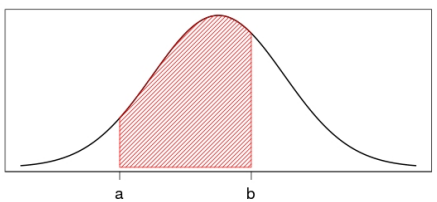
- График сваке нормалне расподеле је симетрична, звоната крива чија је средина једнака μ
- Превоји криве су у тачкама $\mu - \sigma$ и $\mu + \sigma$
- Дисперзија σ^2 одређује облик криве
- Површина испод целе криве једнака је 1
- Вероватноћа да је нормална случајна променљива једнака неком броју је 0, а вероватноће да узме вредност из неког интервала (a, b) је површина испод графика између a и b



Слика: Нормална расподела



Слика: Различите дисперзије



Слика: $P\{a < Z < b\}$

$$X : \mathcal{N}(\mu_X, \sigma_X^2),\ Y : \mathcal{N}(\mu_Y, \sigma_Y^2),\ \mu_X = \mu_Y,\ \sigma_X^2 > \sigma_Y^2$$

Нормалне вероватноће

- Како рачунамо нормалне вероватноће?
- Површина испод графика једнака је интегралу

$$P\{a < Z < b\} = \int_a^b f(x)dx = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx$$

- Овај интеграл не може се одредити, већ се само за конкретне a и b може приближно израчунати
- Израчунате вредности су традиционално табелиране у статистичким таблицама, али данас се углавном одређују помоћу рачунара

Стандардна нормална расподела

Дефиниција

Случајна променљива која има нормалну расподелу са средњом вредношћу $\mu = 0$ и дисперзијом $\sigma^2 = 1$ назива се **стандардном нормалном расподелом**.

- Стандардна нормална расподела је значајна јер све све друге линеарном трансформацијом могу свести на њу
- Ако Z има нормалну расподелу (не обавезно стандардну), на основу стандардне нормалне расподеле обично решавамо следеће две врсте проблема:
  - За дато x рачунамо вероватноће облика $P\{Z < x\}$, $P\{Z > x\}$ и сл.
  - За дату вероватноћу p рачунамо вредности x тако да је $P\{Z < x\} = p$, $P\{Z > x\} = p$ и сл.

Читање таблица стандардне нормалне расподеле

Функција стандардне нормалне расподеле $P\{Z < x\}$

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
...	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
...	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Ако је $-4 < x < 4$, онда се $P\{Z < x\}$ чита из таблице

$$P\{Z < 1.75\} = P\{Z < 1.7 + 0.05\} = 0.9599$$

$$P\{Z < 0.39\} = P\{Z < 0.3 + 0.09\} = 0.6517$$

$P\{Z < x\} = 0$ ако је $x \leq -4$; $P\{Z < x\} = 1$ ако је $x \geq 4$.

Стандардна нормална расподела – R функције

Случајна променљива Z има стандардну нормалну расподелу

- Вероватноћу догађаја $P\{Z < x\}$ рачунамо функцијом pnorm(x).

```
> pnorm(1.25)
0.8943502
> pnorm(-0.73)
0.2326951
```

- Вредност x за коју је $P\{Z < x\} = p$ рачунамо функцијом qnorm(p).

```
> qnorm(0.95)
1.644854
> qnorm(0.50)
0
```

- Вредност густине расподеле у тачки x, $f(x)$, рачунамо функцијом dnorm(x)

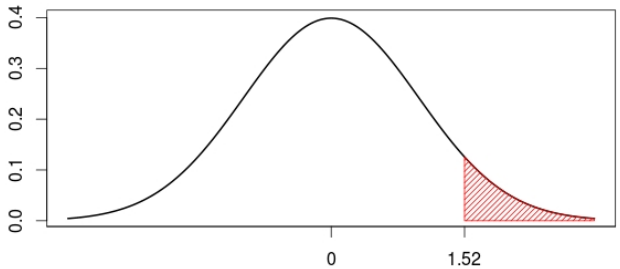
```
> dnorm(0.20)
0.3910427
```

- Случајни узорак обима n (тј. n реализација експеримента случајне променљиве Z) добијамо функцијом rnorm(n)

```
> rnorm(5)
0.8797865 2.4125970 0.9575995 0.6599810 -1.1878357
```

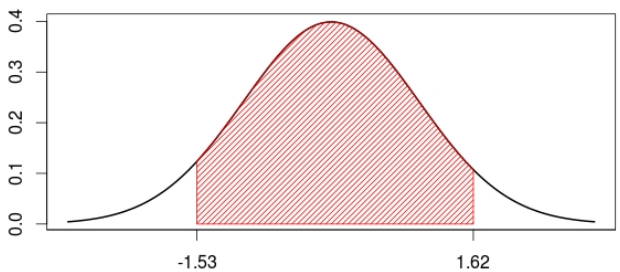
Стандардна нормална расподела

Z – стандардна нормална — желимо да израчунамо следеће вероватноће:
$P\{Z \geq 1.52\}$, $P\{-1.53 < Z < 1.62\}$



Слика: $P\{Z > 1.52\}$

$$\begin{aligned} P\{Z \geq 1.52\} &= 1 - P\{Z < 1.52\} \\ &= 1 - 0.9357 \\ &= 0.0643 \end{aligned}$$

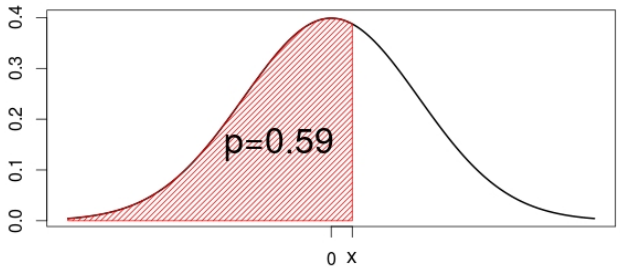


Слика: $P\{-1.53 < Z < 1.62\}$

$$\begin{aligned} & P\{-1.53 < Z < 1.62\} \\ =& P\{Z < 1.62\} - P\{Z < -1.53\} \\ =& 0.9474 - 0.0639 \\ =& 0.8844. \end{aligned}$$

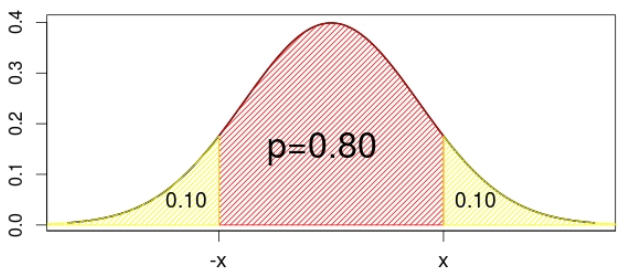
Стандардна нормална расподела

Z – стандардна нормална — желимо да израчунамо следеће вредности x: $P\{Z \leq x\} = 0.59$, $P\{-x < Z < x\} = 0.80$



Слика: $P\{Z < x\} = 0.59$

Тражимо x за које важи да је $P\{Z \leq x\} = 0.59$. Функцијом qnorm(0.59) добијамо 0.23, па је $x = 0.23$.



Слика: $P\{-x < Z < x\} = 0.80$

Тражимо x за које важи да је $P\{-x < Z < x\} = 0.80$. Видимо с графика да је онда $P\{Z < x\} = 0.90$. Функцијом qnorm(0.90) добијамо 1.28, па је $x = 1.28$.

Нормална расподела

- Како рачунамо вероватноће из нормалне расподеле која није стандардна?

Теорема (Теорема стандардизације)

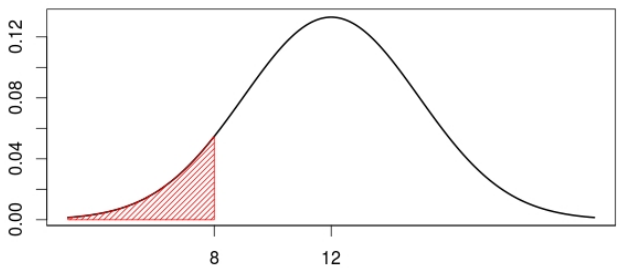
Нека случајна променљива Z има нормалну расподелу са средњом вредношћу μ и дисперзијом σ^2. Тада случајна променљива

$$Z^* = \frac{Z - \mu}{\sigma}$$

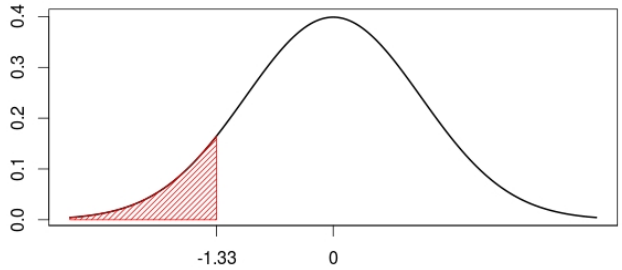
има стандардну нормалну расподелу. Z^* се назива стандардизацијом случајне променљиве Z.

Нормална расподела

- Маса Z (у кг.) изгубљена после једнонедељне дијете има нормалну расподелу са $\mu = 12$ и $\sigma^2 = 9$. Колика је вероватноћа да неко изгуби мање од 8 килограма?



Слика: $P\{Z < 8\}$



Слика: $P\{Z^* < -1.333\}$

$$P\{Z < 8\} = P\left\{Z^* < \frac{8-12}{9}\right\} = P\{Z^* < -1.333\} = 0.0912.$$

У функцију pnorm могу се убацити и параметри нормалне расподеле, програм ће у том случају сам извршити стандардизацију.

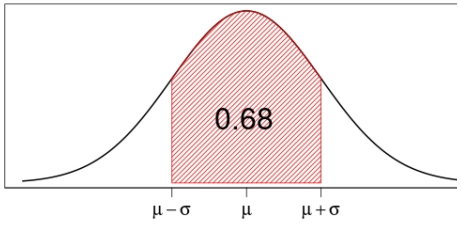
```
> pnorm(8,mean=12,sd=3)
0.09121122
```

Правила 1σ, 2σ и 3σ

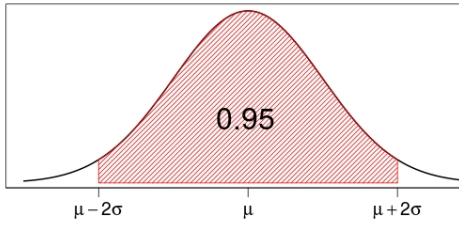
Теорема

Нека случајна променљива Z има нормалну расподелу $\mathcal{N}(\mu, \sigma^2)$. Тада важи:

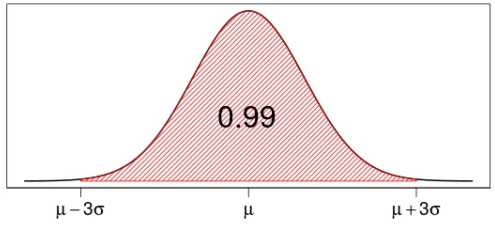
1) *Вероватноћа да Z одступи од свог математичког очекивања највише за једно стандардно одступање је приближно 0.68*
$$P\{\mu - \sigma < Z < \mu + \sigma\} \approx 0.68$$
2) *Вероватноћа да Z одступи од свог математичког очекивања највише за два стандардна одступања је приближно 0.95*
$$P\{\mu - 2\sigma < Z < \mu + 2\sigma\} \approx 0.95$$
3) *Вероватноћа да Z одступи од свог математичког очекивања највише за три стандардна одступања је приближно 0.99*
$$P\{\mu - 3\sigma < Z < \mu + 3\sigma\} \approx 0.99$$



Слика: Правило 1σ



Слика: Правило 2σ



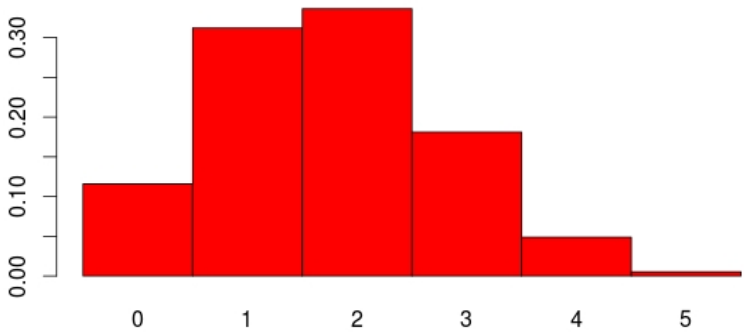
Слика: Правило 3σ

Апроксимација биномне расподеле нормалном

X има биномну расподелу где је $n = 5$ и $p = 0.35$

$$f(x) = \binom{5}{x} 0.35^x 0.65^{5-x}$$

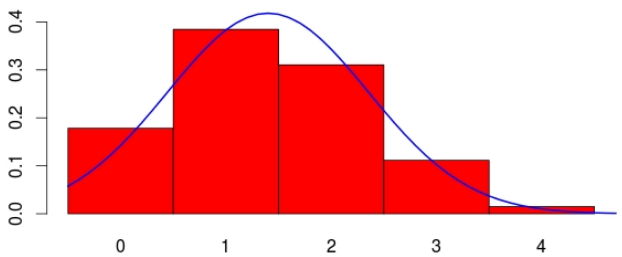
$$f(0) = 0.1160$$
$$f(1) = 0.3124$$
$$f(2) = 0.3364$$
$$f(3) = 0.1811$$
$$f(4) = 0.0488$$
$$f(5) = 0.0052$$



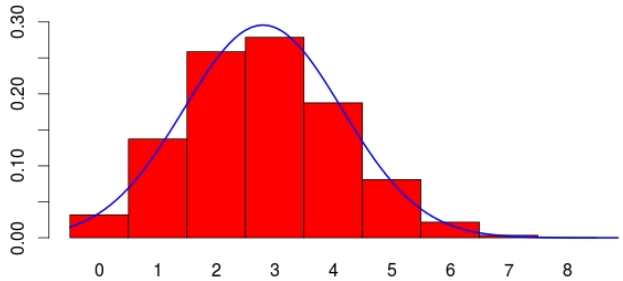
Слика: Графички приказ $f(x)$

Апроксимација биномне расподеле нормалном

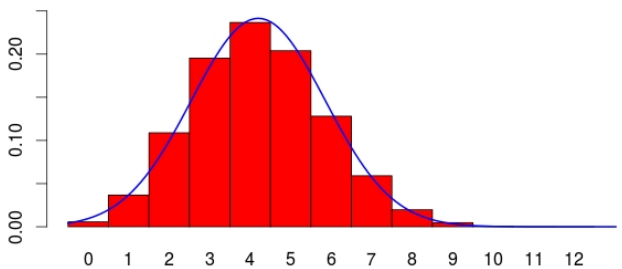
Графички приказ биномних вероватноћа за $p = 0.35$



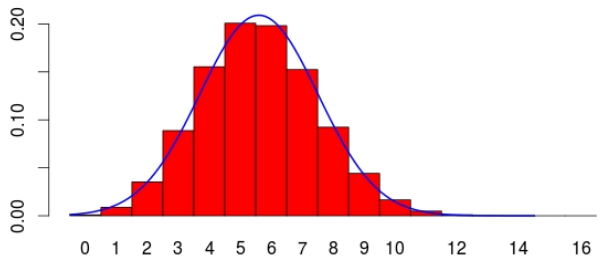
Слика: $n = 4$, $np = 1.4$



Слика: $n = 8$, $np = 2.8$



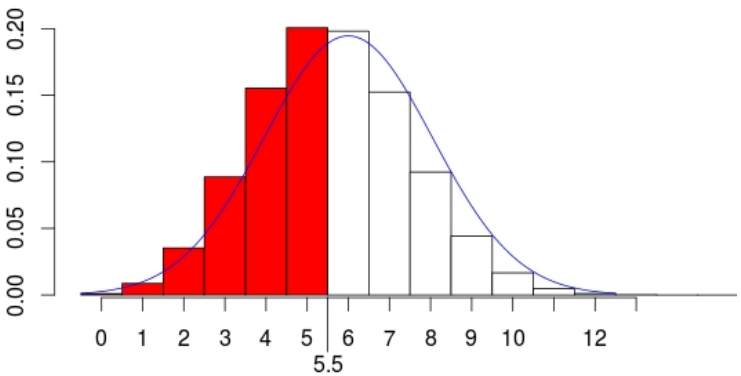
Слика: $n = 12$, $np = 4.2$



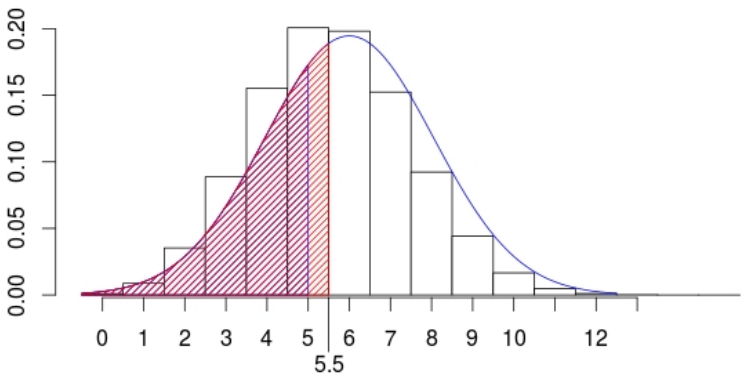
Слика: $n = 16$, $np = 5.6$

Апроксимација биномне расподеле нормалном

X има биномну расподелу где је $n = 20$ и $p = 0.3$. Рачунамо $P\{X \leq 5\}$.



Слика: Биномна вероватноћа



Слика: Нормална вероватноћа

$n = 20,\ p = 0.3$

$$\mu = np = 6, \quad \sigma = \sqrt{np(1-p)} = 2.05$$

$$\begin{aligned} P\{X \leq 5\} &= 0.0008 + 0.0068 + 0.0279 \\ &\quad + 0.716 + 0.1304 + 0.1789 \\ &= 0.4164 \end{aligned}$$

$$\begin{aligned} P\{Z \leq 5.5\} &= P\left\{\frac{Z-\mu}{\sigma} \leq \frac{5.5-6}{2.05}\right\} \\ &= P\{Z^* \leq -0.24\} = 0.4052 \end{aligned}$$

Апроксимација биномне расподеле нормалном

Теорема

Нека X има биномну расподелу с параметрима n и p. Уколико је $p \leq 0.5$ и $np > 5$ или $p \geq 0.5$ и $n(1-p) > 5$, тада, за природне бројеве a и b важи

$$P\{a \leq X \leq b\} \approx P\left\{\frac{a - 0.5 - np}{\sqrt{np(1-p)}} \leq Z^* \leq \frac{b + 0.5 - np}{\sqrt{np(1-p)}}\right\},$$

где Z^* има стандардну нормалну расподелу.

Апроксимација биномне расподеле нормалном

Колика је вероватноћа да је међу 49 ученика њих 7 рођено у недељу? А колика да је таквих ученика више од 10?

X - број ученика рођених у недељу има биномну расподелу где је $n = 49$ и $p = 1/7$.
$p < 0.5$, $np = 7 > 5$ - користимо нормалну апроксимацију

$$P\{X = 7\} = P\{7 \leq X \leq 7\} \approx P\Big\{\frac{6.5 - 7}{\sqrt{6}} \leq Z^* \leq \frac{7.5 - 7}{\sqrt{6}}\Big\}$$
$$= P\{-0.204 \leq Z^* \leq 0.204\} = 0.5808 - 0.4192 = 0.1616.$$

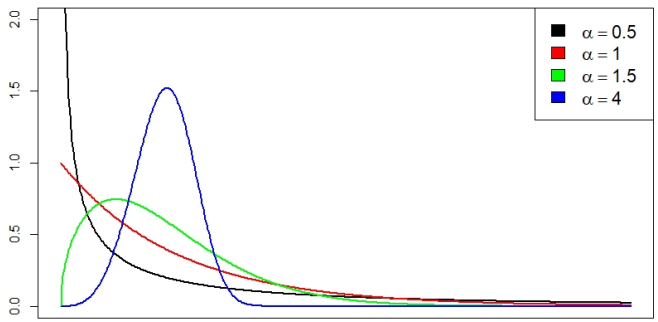
$$P\{X > 10\} = P\{11 \leq X\} \approx P\Big\{\frac{10.5 - 7}{\sqrt{6}} \leq Z^*\Big\}$$
$$= P\{Z^* \geq 1.43\} = 0.0764.$$

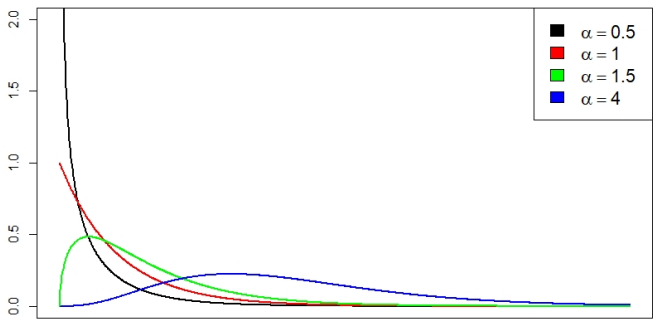
Квалитет апроксимације

```
> dbinom(7,size=49,p=1/7)
0.1608958
> 1-pbinom(10,size=49,p=1/7)
0.0820548
```

Неке расподеле животног века

- Вејбулова расподела $f(x) = \alpha\lambda^{\alpha}x^{\alpha-1}e^{-\lambda x^{\alpha}}, x > 0$
- Гама расподела $f(x) = \frac{\lambda^{\alpha}x^{\alpha-1}}{\Gamma(\alpha)}e^{-\lambda x}, x > 0$





- Обе расподеле имају параметре облика $\alpha > 0$ и размере $\lambda > 0$
- За $\alpha = 1$ обе расподеле своде се на експоненцијалну с параметром стопе отказа λ
- Обе расподеле су померене удесно, што је израженије за мање вредности α
- Код обе расподеле ако је $\alpha > 1$ стопа отказа расте током времена, тј. половна компонента има већу стопу отказа од нове, ако је $\alpha < 1$, онда је обрнуто

```
> pweibull(1,shape=2,scale=1)
0.6321206
> qgamma(0.5,shape=2,rate=1)
1.678347
```

scale $= 1/\lambda$

Врсте статистичког закључивања

- Оцењивање непознатих параметара
- Тестирање статистичких хипотеза

Приликом проучавања криминала популацију чине све особе старије од 16 година који су осуђени због неког кривичног дела. Занима нас:

1) Колики је средњи број година образовања у тој популацији?
2) Да ли је већина чланова популације ухапшена бар једном пре него што је први пут осуђена?

1)- оцењивање параметра - примењује се кад немамо претходна знања о параметру

2) тестирање хипотезе - примењује се када имамо претпоставку од правој вредности непознатог параметра – у примеру да је проценат претходно ухапшених већи од 50%

Заједничко за оба приступа је

- Одређивање популације
- Одређивање случајне променљиве коју проучавамо
- Одређивање параметара од важности
- Извлачење узорка из популације

Узорак

- Пре извођења статистичког закључка треба најпре извући случајни узорак
- Одредимо обим узорка
- Елементе популације на којима меримо вредност случајне променљиве бирамо случајно преко таблице случајних бројева или коришћењем рачунара
- Пре избора елемената популације елементи узорка $X_1, \ldots, X_n$ су случајне променљиве, а кад измеримо вредности добијамо њихове реализације

Дефиниција

Случајни узорак из расподеле за X чине случајне променљиве $X_1, \ldots, X_n$, које су међусобно независне и имају исту расподелу као X.

Тачкасте оцене

- Тачкаста оцена непознатог параметра θ је нека статистика T чије вредности дају добру процену о вредности тог параметра
- Оцена T је случајна променљива јер за различите узорке узима различите вредности, она никад неће бити баш једнака θ, али се надамо да даје добру процену
- Квалитетне тачкасте оцене пожељно је да испуњавају неке услове:
  1) да буду непристрасне, тј. да је математичко очекивање оцене једнако параметру
  2) да им је дисперзија мала кад је n велико (тј. тежи нули кад $n \to \infty$)
- Тачкасте оцене најчешће налазимо принципом замене или методом максималне веродостојности
  - Принцип замене тражи шта представља параметар у популацији и оцењује га одговарајућом статистиком из узорка
  - Метод максималне веродостојности рачуна вероватноћу добијања конкретног узорка као функцију непознатог параметра и оцењује параметар вредношћу за коју функција достиже максимум

Тачкаста оцена параметра μ

- Претпоставимо да имамо популацију и на њој дефинисану случајну променљиву X чији су средња вредност μ и дисперзија σ^2 непознати
- Тачкаста оцена параметра μ је узорачка средина $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$
- Оцена $\bar{X}$ има обе пожељне особине, $EX = \mu$ и $DX = \frac{\sigma^2}{n}$, што је мало када је n велико.

Тражимо оцену посечног броја продатих сендвича у току једне недеље. На узорку обима 16 добили смо следеће вредности

905	975	783	900
1000	950	1003	789
800	600	850	913
795	925	875	810

На основу овог узорка обима 16 добија се $\bar{x} = 867.1$ што је тачкаста оцена параметра μ.

Интервали поверења

Дефиниција

За интервал (G_1, G_2) кажемо да је $100(1-\alpha)\%$ интервал поверења за параметар θ уколико су G_1 и G_2 статистике такве да важи

$$P\{G_1 \leq \theta \leq G_2\} = 1 - \alpha,$$

без обзира на праву вредност параметра θ.

- За одређивање граница интервала (из одговарајућих вероватноћа) треба нам расподела неке случајне променљиве

Интервал поверења за μ

Теорема

Нека је $X_1, \ldots, X_n$ узорак обима n из нормалне расподеле с параметрима μ и σ. Тада $\bar{X}$ има нормалну расподелу чија је средњу вредост μ и дисперзија σ^2/n.

- На основу стандардизације $Z = \frac{\bar{X}-\mu}{\sigma}\sqrt{n}$ има стандардну нормалну расподелу
- Уколико узорак није из нормалне, него из неке друге расподеле чија је средња вредност μ, а дисперзија σ^2, онда Z добијено горњом формулом нема нормалну расподелу, али за велико n $(n > 25)$ има приближно нормалну расподелу

Интервал поверења за μ кад је σ^2 познато

Тражимо 90% интервал поверења за средњи број сендвича продатих у току недеље у једном фаст фуд ресторану. Претпоставимо да је дисперзија, на основу неких старих истраживања једнака 100. У узорку обима 16 израчунали смо $\bar{x} = 867.1$.
Пошто $Z = \frac{\bar{X}-\mu}{\sigma}\sqrt{n}$ има нормалну расподелу, онда важи да је

$$P\{-1.645 < Z < 1.645\} = 0.90 \qquad (\text{qnorm}(0.95))$$

$$P\Big\{-1.645 < \frac{\bar{X}-\mu}{\sigma}\sqrt{n} < 1.645\Big\} = 0.90$$

$$P\Big\{\bar{X} - 1.645\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.645\frac{\sigma}{\sqrt{n}}\Big\} = 0.90,$$

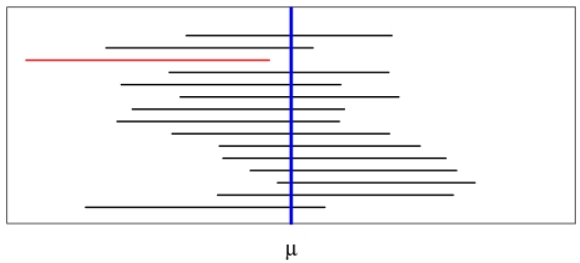
па је тражени интервал поверења

$$\Big(\bar{X} - 1.645\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.645\frac{\sigma}{\sqrt{n}}\Big)$$

За наш узорак добијамо интервал (826.0,908.2). За друге узорке добили бисмо другачије интервале.

Интервал поверења за μ кад је σ^2 познато

Шта значи да имамо поверење од 90%? То значи да ће 90% узорака "ухватити" вредност μ, а 10% ће га промашити. Ми "верујемо" да је наш узорак онај који "хвата" праву вредност непознатог параметра.



Слика: Интервали поверења за μ

Теорема

Нека је $X_1, \ldots, X_n$ случајни узорак обима n из нормалне расподеле с параметрима μ и познатом вредношћу σ^2. $100(1-\alpha)\%$ интервал поверења за μ је

$$\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right),$$

где је $z_{\alpha/2}$ такво да је $P\{Z > z_{\alpha/2}\} = \frac{\alpha}{2}$ (површина **десно** *од $z_{\alpha/2}$ је $\frac{\alpha}{2}$).*

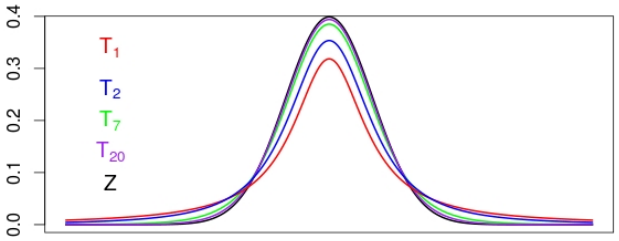
Студентова T расподела

Шта да радимо ако σ није познато? Оцењујемо га узорачким стандардним одступањем S и онда се расподела мења.

Кад оценимо параметар σ статистиком S, тада

$$\frac{\bar{X} - \mu}{S}\sqrt{n}$$

има Студентову T расподелу.



Слика: Студентове расподеле

Особине Студентових расподела

- Свака Студентова расподела има један параметар ν, број степени слободе
- Студентова расподела је непрекидна
- График је симетричан око нуле, средња вредност је нула
- Параметар ν утиче на дисперзију, што је он већи, дисперзија је мања
- Када је ν велико, Студентова расподела је приближна стандардној нормалној

Таблица Студентових расподела

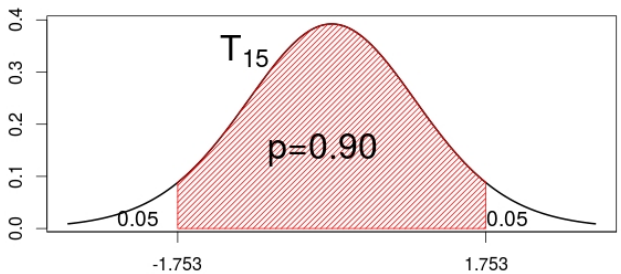
Студентова расподела - вредности x такве да је $P\{T_\nu < x\} = p$

	p									
ν	0.600	0.667	0.750	0.800	0.875	0.900	0.950	0.975	0.990	0.995	0.999
1	0.325	0.577	1.000	1.376	2.414	3.078	6.314	12.706	31.821	63.657	318.31
2	0.289	0.500	0.816	1.061	1.604	1.886	2.920	4.303	6.965	9.925	22.327
3	0.277	0.476	0.765	0.978	1.423	1.638	2.353	3.182	4.541	5.841	10.215
4	0.271	0.464	0.741	0.941	1.344	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.457	0.727	0.920	1.301	1.476	2.015	2.571	3.365	4.032	5.893
6	0.265	0.453	0.718	0.906	1.273	1.440	1.943	2.447	3.143	3.707	5.208
...	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	
28	0.256	0.435	0.683	0.855	1.175	1.313	1.701	2.048	2.467	2.763	3.408
29	0.256	0.435	0.683	0.854	1.174	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.435	0.683	0.854	1.173	1.310	1.697	2.042	2.457	2.750	3.385
...	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	
55	0.255	0.433	0.679	0.848	1.163	1.297	1.673	2.004	2.396	2.668	3.245
60	0.254	0.433	0.679	0.848	1.162	1.296	1.671	2.000	2.390	2.660	3.232
∞	0.253	0.431	0.674	0.842	1.150	1.282	1.645	1.960	2.326	2.576	3.090

- Тражимо x такво да је $P\{T_5 < x\} = 0.95$ — из таблице видимо да је $x = 2.015$.
- Тражимо x такво да је $P\{T_5 < x\} = 0.05$ — површина је мања од $1/2$, па је x негативно $P\{T_5 < -x\} = 0.95$, па је $x = -2.015$.
- Тражимо x такво да је $P\{T_2 > x\} = 0.025$ — онда је $P\{T_2 < x\} = 0.975$, па је $x = 4.303$

Студентова расподела - R функције

- Вероватноћу догађаја $P\{T_\nu < x\}$ рачунамо функцијом $\text{pt}(x, \text{df} = \nu)$.
  ```
  > pt(1.25,df=10)
  0.8801197
  ```
- Вредност x за коју је $P\{T_\nu < x\} = p$ рачунамо функцијом $\text{qt}(p, \text{df} = \nu)$.
  ```
  > qt(0.95,df=5)
  2.015048
  ```
- Тражимо x такво да је $P\{-x < T_{15} < x\} = 0.90$



Слика: $P\{-x < T_{15} < x\} = 0.90$

Видимо да је површина графика лево од x једнака 0.95 па x налазимо као $\text{qt}(0.95, \text{df} = 15)$, а то је 1.753.

Интервал поверења за μ када се σ^2 оцењује

Теорема

Нека је $X_1, \ldots, X_n$ случајни узорак обима n из нормалне расподеле с параметрима μ и σ^2. Тада

$$\frac{\bar{X} - \mu}{S}\sqrt{n}$$

има Студентову T расподелу с $n-1$ степеном слободе.

Теорема

Нека је $X_1, \ldots, X_n$ случајни узорак обима n из нормалне расподеле с параметрима μ и σ^2. $100(1-\alpha)\%$ интервал поверења за μ је

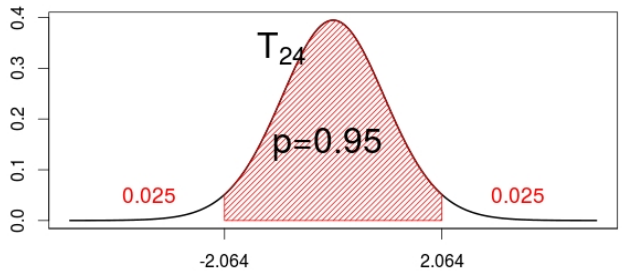
$$\left(\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right),$$

где је $t_{\alpha/2}$ такво да је $P\{T_{n-1} > t_{\alpha/2}\} = \frac{\alpha}{2}$ (површина **десно** *од $t_{\alpha/2}$ је $\frac{\alpha}{2}$).*

Интервал поверења за μ када се σ^2 оцењује

Посматра се процентуална промена у броју студената уписаних на државне универзитете. Може ли се, на основу доњег узорка, тврдити да је, у просеку, дошло до повећања броја студената?

5%	35%	-8%	0.3%	5%
-1%	-30%	12%	0%	3%
-10%	16%	-5%	7%	7%
25%	-15%	2%	-17%	8%
0%	6%	9%	7%	3%



Слика:

$$P\{-2.064 < T_{24} < 2.064\} = 0.95$$

Имамо да је $\bar{x} = 2.6$, $s^2 = 170.36$, $s = 13.1\%$. Вредност $t_{\alpha/2}$ налазимо тако што је површина десно једнака 0.025, односно важи $P\{T_{24} < t\} = 0.975$, те имамо $\text{qt}(0.975, \text{df} = 24) = 2.064$. Интервал поверења је $\left(\bar{X} - 2.064 \cdot \frac{S}{5}, \bar{X} + 2.064 \cdot \frac{S}{5}\right)$.

За наш узорак добија се (-2.8,8.0). Закључак је да верујемо, с поверењем од 95% да је процентуално повећање броја уписаних студената између -2.8 и 8.0%. Пошто је 0 унутар интервала, а имамо и негативне вредности, не можемо тврдити да се број уписаних повећава.

Тестирање статистичких хипотеза

- Имамо две хипотезе: нулту и алтернативну
- Алтернативна (или истраживачка) хипотеза је оно што тврдимо и желимо да статистички проверимо (обично садржи речи као веће, мање, зависи...)
- Нулта хипотеза је супротна алтернативној (обично садржи речи једнако, мање или једнако, не зависи...)
- Тестирање се врши у циљу одбацивања нулте хипотезе, тј. прихватања суштинске алтернативне хипотезе

Проблем тестирања

	стварно стање оптуженог	
одлука пороте	није крив	крив је
крив	грешка прве врсте	исправна одлука
није крив	исправна одлука	грешка друге врсте

	стварно стање ствари	
закључак тестирања	H_0 је тачна	H_0 је нетачна
одбацујемо H_0	грешка прве врсте	исправна одлука
не одбацујемо H_0	исправна одлука	грешка друге врсте

Проблем тестирања

- Закључак доносимо на основу вредности неке статистике, коју називамо **тест статистиком**. Ако она у нашем узорку узме вредност која је неуобичајена ако важи H_0, онда одбацујемо H_0.
- Одбацивање H_0 је статистички значајна одлука, значи да смо скупили довољно доказа у прилог нашој алтернатвиној хипотези
- Неодбацивање H_0 није статистички значајан резултат. То може да значи да стварно важи H_0, или да важи H_1 али да немамо довољно доказа њој у прилог.
- Како донети одлуку? Једна могућност је задати унапред вредност α (најчешће 0.05), који називамо **мером** или **прагом значајности** теста, који ће нам фиксирати вероватноћу грешке прве врсте. Уколико нам тест статистика узме вредност која под H_0 има вероватноћу мању од α, одбацујемо H_0.
- p-вредност теста је вероватноћа да извучемо неки узорак који је бољи доказ у корист наше алтернативне хипотезе од оног који смо већ извукли Уколико је та вероватноћа мала, то значи да су наши докази одлични па одбацујемо H_0. Граница је поново обично на 5%.

Тестирање хипотеза о средњој вредности

Врсте нултих и алтернативних хипотеза

- Машина за постављање чуњева у куглању треба да има просечно време постављања од 4 секунде. Ако је дуже, ствара се нервоз код такмичара, а ако је краће, чуњеви се обарају. Тестирамо машину да ли ради како треба

$$H_1 : \mu \neq 4, \quad H_0 : \mu = 4$$

- Имамо рачунар на коме је за наш програм потребно 45 секунди да се изврши. Приликом куповине новог рачунара желимо да будемо сигурни да је он бољи. Тестирамо

$$H_1 : \mu < 45; \quad H_0 : \mu \geq 45$$

- Разматра се отварање нове продавнице и сматра се да је треба отворити уколико приходи буду већи од 2$ по муштерији. Тестира се

$$H_1 : \mu > 2; \quad H_0 : \mu \leq 2$$

Тестирање хипотеза о средњој вредности

Дефиниција

Постоје три теста о средњој вредности

- $H_0 : \mu = \mu_0$ *против* $H_1 : \mu \neq \mu_0$ *(двострани)*
- $H_0 : \mu \geq \mu_0$ *против* $H_1 : \mu < \mu_0$ *(једностани леви)*
- $H_0 : \mu \leq \mu_0$ *против* $H_1 : \mu > \mu_0$ *(једнострани десни)*

Тест статистика у сва три случаја је

$$T_0 = \frac{\bar{X} - \mu_0}{S}\sqrt{n},$$

која, ако је H_0 *тачно, има Студентову расподелу с* $n - 1$ *степеном слободе.*

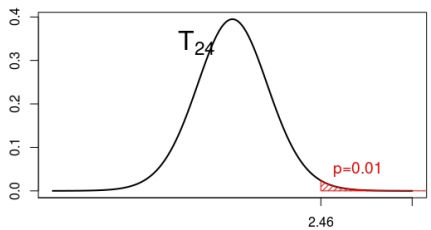
- p*-вредност једностраног левог теста је површина лево од вредности* t_0 *коју статистика* T_0 *узме у узорку.*
- p*-вредност једностраног десног теста је површина десно од вредности* t_0
- p*-вредност двостраног теста је двострука површина лево од вредности* t_0*, ако је* $t_0 < 0$ *или двострука површина десно од вредности* t_0*, ако је* $t_0 > 0$.

Тестирање хипотеза о средњој вредности

Тестирамо $H_0 : \mu \leq 2\$$ (продавница није профитабилна) против $H_1 : \mu > 2\$$ (продавница је профитабилна)

Узорак:

2.75	6.25	3.50	3.01	5.10
5.06	4.50	4.17	2.57	3.15
3.98	2.37	2.03	1.02	5.28
1.57	1.00	1.16	1.07	3.12
0.75	0.10	0.25	3.09	4.10



Слика: $P\{T_{24} > 2.46\} = 0.01$

$n = 25$, $T_0 = \frac{\bar{X}-2}{S}\sqrt{25}$ има Студентову T_{24} расподелу. Из узорка рачунамо $\bar{x} = 2.842$, $s^2 = 1.708$, $s = 2.918$. Вредност тест статистике из узорка је $t_0 = \frac{2.842-2}{2.918}\sqrt{25} = 2.46$.

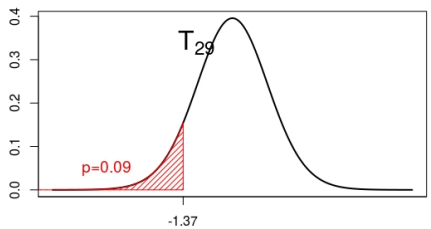
p-вредност теста добијамо као $1-\text{pt}(2.46, \text{df} = 24) = 0.01$, па закључујемо да треба одбацити H_0 и отворити продавницу.

Тестирање хипотеза о средњој вредности

Тестирамо $H_0 : \mu \geq 45$ (нови рачунар није бољи) против $H_1 : \mu < 45$ (нови рачунар је бољи)

У узорку обима 30 добијено је

$\bar{x} = 44.5,\ s = 2$



Слика: $P\{T_{29} < -1.37\} = 0.09$

$n = 30$, $T_0 = \frac{\bar{X}-45}{S}\sqrt{29}$ има Студентову T_{29} расподелу. Вредност тест статистике из узорка је $t_0 = \frac{44.5-45}{2}\sqrt{29} = -1.37$.

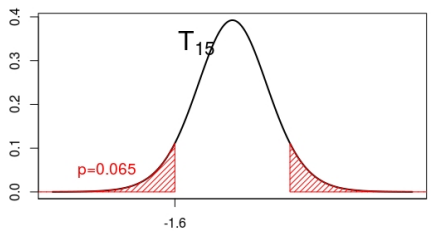
p-вредност теста добијамо као $\text{pt}(-1.37, \text{df} = 29) = 0.09$. Одлука је на нама, ако сматрамо да је грешка од 9% велика, закључићемо да не треба одбацити H_0 и не треба купити нови рачунар, а ако мислимо да је мала, онда ћемо купити нови рачунар.

Тестирање хипотеза о средњој вредности

Тестирамо $H_0 : \mu = 4$ (машина за чуњеве ради како треба) против $H_1 : \mu \neq 4$ (треба јој сервис)

Узорак:

4.1	3.5	3.2	4.1
3.5	4.3	4.0	4.5
2.5	3.8	4.6	3.0
4.1	3.6	3.7	3.9



Слика: $P\{T_{15} < -1.60\} = 0.065$

$n = 16$, $T_0 = \frac{\bar{X}-4}{S}\sqrt{16}$ има Студентову T_{15} расподелу. Из узорка рачунамо $\bar{x} = 3.78$, $s = 0.55$. Вредност тест статистике из узорка је $t_0 = \frac{3.78-4}{0.55}\sqrt{16} = -1.60$.

Како је $\mathrm{pt}(-1.60, \mathrm{df} = 15) = 0.065$, а пошто је тест двострани, p-вредност теста је 0.13. Таква грешка је превелика, па не одбацујемо H_0 и не сервисирамо машину.

Коришћење уграђене R функције

```
> kuglanje<-c(4.1,3.5,2.5,4.1,3.5,4.3,3.8,3.6,3.2,4.0,4.6,3.7,4.1,4.5,
3.0,3.9)
> t.test(kuglanje,mu=4,alternative="two.sided",conf.level=0.95)

 One Sample t-test

data:  kuglanje
t = -1.6234, df = 15, p-value = 0.1253
alternative hypothesis: true mean is not equal to 4
95 percent confidence interval:
 3.479594 4.070406
 sample estimates:
mean of x
    3.775
```

Праг значајности теста

- Праг значајности теста је максимална грешка коју толеришемо при одбацивању нулте хипотезе
- Ако је дат праг α, онда ако је p-вредност теста мања од α одбацујемо нулту хипотезу, а ако је p-вредност теста већа од α, немамо довољно доказа да одбацимо нулту хипотезу
- Кажемо да смо одбацили (или не можемо да одбацимо) H_0 при прагу α

Испитује се дужина паузе између два узастопна светла код једне врсте свитаца. Желимо да потврдимо нашу претпоставку да је средња дужина паузе краћа од 4 секунде, па је $H_0 : \mu \geq 4$, а $H_1 : \mu < 4$. Последице грешке нису катастрофалне па дозвољавамо грешку од $\alpha = 10\%$. Имамо узорак обима 16 у коме је $\bar{x} = 3.77$ и $s = 0.30$.
$t_0 = \frac{\bar{x}-4}{s}\sqrt{16} = -3.06$. Расподела је T_{15}.

Први начин: p-вредност теста је 0.004. Пошто је p-вредност мања од $\alpha = 0.1$, одбацујемо H_0 и закључујемо да смо у праву кад тврдимо да је средња пауза краћа од 4 секунде, при прагу од 10%.

Други начин: Вредност t тако да је $P\{T_{15} < t\} = 0.1$ је -1.34 $(\text{qt}(0.1, \text{df} = 15))$. Како је $t_0 < t$, то значи да је $P\{T_{15} < t_0\} < P\{T_{15} < t\}$ па одбацујемо H_0.

Хи квадрат расподела

Слика: Хи квадрат расподела

Особине хи квадрат расподела

- Свака хи квадрат расподела има један параметар ν, број степени слободе
- Хи квадрат расподела је непрекидна
- Вредности хи квадрат расподеле увек су позитивне
- Хи квадрат расподела је несиметрична
- Математичко очекивање хи квадрат расподеле χ^2_ν је ν, а дисперзија 2ν

Хи квадрат расподела

χ^2 расподела - вредности x такве да је $P\{X^2_\nu < x\} = p$

	p										
ν	0.001	0.005	0.010	0.025	0.050	0.100	0.900	0.950	0.975	0.990	0.995	0.999
1	0.000	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879	10.828
2	0.002	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597	13.816
3	0.024	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838	16.266
4	0.091	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860	18.467
5	0.210	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750	20.515
6	0.381	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548	22.458
7	0.598	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278	24.322
...	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	
20	5.921	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997	45.315
21	6.447	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401	46.797
22	6.983	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796	48.268
...	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	

- Тражимо x такво да је $P\{X^2_5 < x\} = 0.05$ — из таблице видимо да је $x = 1.145$.
- Тражимо x такво да је $P\{X^2_{20} > x\} = 0.025$ — онда је $P\{X^2_{20} < x\} = 0.975$, па је $x = 34.170$

Хи квадрат расподела - R функције

- Вероватноћу догађаја $P\{X_\nu^2 < x\}$ рачунамо функцијом pchisq$(x, \mathrm{df} = \nu)$.

```
> pchisq(4.25,df=3)
0.7642966
```

- Вредност x за коју је $P\{X_\nu^2 < x\} = p$ рачунамо функцијом qchisq$(p, \mathrm{df} = \nu)$.

```
> qchisq(0.90,df=7)
12.01704
```

Интервал поверења за σ^2 код нормалне расподеле

- Тачкаста оцена параметра σ^2 је узорачка дисперзија $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
- Она је непристрасна оцена чија дисперзија тежи нули кад $n \to \infty$

Теорема

Нека је $X_1, \ldots, X_n$ случајни узорак обима n из нормалне расподеле с параметрима μ и σ^2. Тада

$$\frac{(n-1)S^2}{\sigma^2}$$

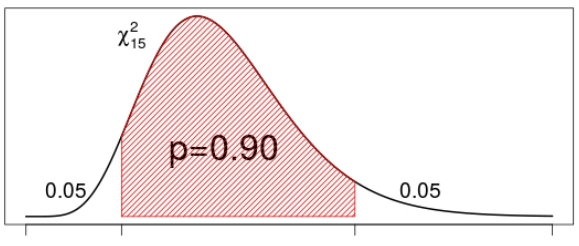
има χ^2 расподелу с $n-1$ степеном слободе.

Интервал поверења за σ^2 код нормалне расподеле

Тражимо 90% интервал поверења за дисперзију паковања једне врсте чипса.

Узорак			
17.86	17.42	15.91	14.19
14.52	17.11	18.11	19.25
15.82	13.27	13.71	15.80
14.85	17.38	14.28	16.85

$s^2 = 3.125$



Слика: $P\{7.26 < X^2_{15} < 25.0\} = 0.9$

Пошто $\frac{(n-1)S^2}{\sigma^2}$ има χ^2_{n-1} расподелу, онда важи да је

$$P\left\{7.26 < \frac{15S^2}{\sigma^2} < 25.0\right\} = 0.90$$

$$P\left\{\frac{15S^2}{25.0} < \sigma^2 < \frac{15S^2}{7.26}\right\} = 0.90,$$

па је тражени интервал поверења

$$\left(\frac{15S^2}{25.0} < \sigma^2 < \frac{15S^2}{7.26}\right)$$

За наш узорак добијамо (1.875,6.456). За друге узорке су другачији интервали.

Интервал поверења за σ^2 код нормалне расподеле

Теорема

Нека је $X_1, \ldots, X_n$ *случајни узорак обима* n *из нормалне расподеле с параметрима* μ *и* σ^2. $100(1-\alpha)\%$ *интервал поверења за* σ^2 *је*

$$\left(\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}}, \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}}\right),$$

где је $\chi^2_{1-\alpha/2}$ *такво да је* $P\{X^2_{n-1} > \chi^2_{1-\alpha/2}\} = \frac{\alpha}{2}$ *(површина* **десно** *од* $\chi^2_{1-\alpha/2}$ *је* $\frac{\alpha}{2}$*), а* $\chi^2_{\alpha/2}$ *такво да је* $P\{X^2_{n-1} < \chi^2_{\alpha/2}\} = \frac{\alpha}{2}$ *(површина* **лево** *од* $\chi^2_{\alpha/2}$ *је* $\frac{\alpha}{2}$*)*

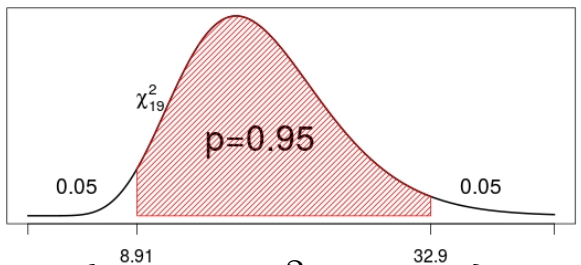
Интервал поверења за стандардно одступање σ *је*

$$\left(\sqrt{\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}}}, \sqrt{\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}}}\right).$$

Интервал поверења за σ^2 код нормалне расподеле

За један психолошки експеримент потребно је да чланови популације која се проучава имају разноврсне године старости, а жељено стандардно одступање је 5 година. Тражимо 90% интервал поверења за дисперзију година старости популације коју испитујемо.

Узорак			
31	26	40	37
35	36	39	37
34	37	38	35
26	41	40	41
35	30	42	36



Слика: $P\{8.91 < X^2_{19} < 32.9\} = 0.9$

Имамо да је $n = 20$, $s^2 = 21.12$. Одговарајући квантили су $\chi^2_{0.95} = 32.9$ и $\chi^2_{0.05} = 8.91$. Интервал поверења за σ^2 је

$$\left(\frac{19S^2}{32.9}, \frac{19S^2}{8.91}\right).$$

За наш узорак добијамо (12.20,45.04). Интервал поверења за стандардно одступање σ је (3.5,6.7). Пошто он обухвата жељену вредност, можемо сматрати да нам популација има задовољавајућу дисперзију.

Тестирање хипотезе о σ^2 код нормалне расподеле

Дефиниција

Постоје три теста о дисперзији σ^2

- $H_0 : \sigma^2 = \sigma_0^2$ *против* $H_1 : \sigma^2 \neq \sigma_0^2$ *(двострани)*
- $H_0 : \sigma^2 = \sigma_0^2$ *против* $H_1 : \sigma^2 < \sigma_0^2$ *(једностани леви)*
- $H_0 : \sigma^2 = \sigma_0^2$ *против* $H_1 : \sigma^2 > \sigma_0^2$ *(једнострани десни)*

Тест статистика у сва три случаја је

$$X_0^2 = \frac{(n-1)S^2}{\sigma_0^2},$$

која, ако је H_0 *тачно, има хи квадрат расподелу с* $n-1$ *степеном слободе.*

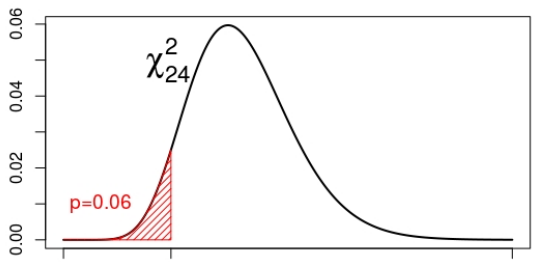
- p*-вредност једностраног левог теста је површина лево од вредности* χ_0^2 *коју статистика* χ_0^2 *узме у узорку.*
- p*-вредност једностраног десног теста је површина десно од вредности* χ_0^2
- p*-вредност двостраног теста је* **приближно** *двострука површина лево или десно од вредности* χ_0^2*, у зависности од тога која је од тих површина мања*

Тестирање хипотезе о σ^2 код нормалне расподеле

Унутрашњи притисак стандардних тениских лоптица има нормалну расподелу са средњом вредношћу 28 и дисперзијом 0.25. Тестирамо да ли лоптице добијене новом техником производње имају мању дисперзију притиска с прагом значајности $\alpha = 0.05$. Тестирамо, дакле,

$$H_0 : \sigma^2 = 0.25 \text{ против } H_1 : \sigma^2 < 0.25$$

		Узорак		
28.20	27.31	28.68	27.98	27.99
28.04	27.47	28.57	28.12	28.75
28.36	27.96	28.30	28.29	28.40
27.46	27.99	27.94	27.76	27.91
27.59	27.71	28.60	27.91	27.82



Слика: $P\{0 < X^2_{24} < 14.37\} = 0.06$

Имамо да је $n = 25$, $s^2 = 0.1497$, $\chi^2_0 = \frac{24 \cdot 0.1497}{0.25} = 14.37$.

p-вредност теста добијамо као $\text{pchisq}(14.37, \text{df} = 24) = 0.062$. Пошто је p-вредност већа од α, немамо доказа да одбацимо H_0, па сматрамо да нове лопте немају мањи притисак од стандардних.

Коришћење уграђене R функције

```
> pritisak.loptica<-c(28.20,27.31,28.68,27.98,27.99,28.04,27.47,
28.57,28.12,28.75,28.36,27.96,28.30,28.29,28.40,27.46,27.99,27.94,
27.76,27.91,27.59,27.71,28.60,27.91,27.82)
> EnvStats::varTest(pritisak.loptica,sigma.squared=0.25,
alternative="less")

Results of Hypothesis Test
--------------------------

Null Hypothesis:  variance = 0.25
Alternative Hypothesis:  True variance is less than 0.25
Test Name:  Chi-Squared Test on Variance
Estimated Parameter(s):  variance = 0.149709
Data:  pritisak.loptica
 Test Statistic:  Chi-Squared = 14.37206
Test Statistic Parameter:  df = 24
P-value:  0.0622042
95% Confidence Interval:  LCL = 0.000000
                          UCL = 0.259453
```

Оцењивање непознатог процента p

- Имамо популацију, и сваки члан класификујемо у односу на то да ли има одређено својство или га нема
- Параметар p представља проценат (тј. удео) популације који има то својство
- Случајна променљива X узима вредност 1 на елементима популације који имају то својство, а 0 на оним који га немају
- Тачкаста оцена параметра p је $\hat{p} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$, где је $\sum_{i=1}^{n} X_i$, у ствари, број елемената узорка који имају испитивано својство

Анкетирано је 500 особа телефоном и 285 њих је против предложених порезних реформи. Ако је p проценат популације који је против реформи, оцена тог процента је

$$\bar{x} = \frac{\sum_{i=1}^{500} x_i}{n} = \frac{285}{500} = 0.57.$$

Закључујемо да је 57% популације против реформи. Када би популација била од милион становника, процена је да је 570000 против.

Оцењивање непознатог процента p

- Случајна променљива $Y = \sum_{i=1}^{n} X_i$, број "успеха" у узорку обима n, има биномну расподелу с параметрима n и p

Теорема

Узорачка средина

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{\textit{број елемената у узорку који имају одређено својство}}{\textit{обим узорка}}$$

непристрасна је оцена непознатог процента p елемената популације који имају то својство. Поред тога важи

$$D\bar{X} = \frac{p(1-p)}{n}.$$

Интервал поверења за p

- С обзиром да Y има биномну расподелу, ако је n велико,

$$\frac{Y - np}{\sqrt{np(1-p)}} = \frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

  има приближно стандардну нормалну расподелу
- Примењујући поступак прављења интервала поверења добили бисмо, за ниво поверења $1 - \alpha$

$$\left(\bar{X} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}, \bar{X} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right)$$

- Ово није добар интервал поверења јер зависи од непознатог параметра p!
- Зато уместо p стављамо његову оцену $\hat{p} = \bar{X}$.

Интервал поверења за p

Теорема

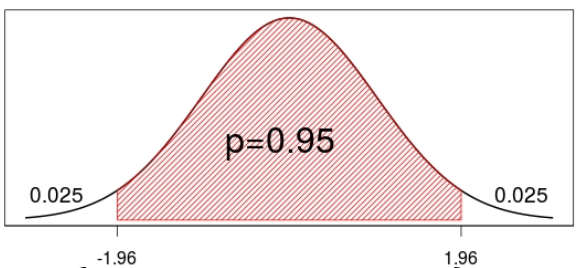
Интервал поверења за непознати проценат p је

$$\left(\bar{X} - z_{\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}, \bar{X} + z_{\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}\right),$$

где је $z_{\alpha/2}$ такво да је $P\{Z > z_{\alpha/2}\} = \frac{\alpha}{2}$.

- Анализирамо популацију гојазних младића (18-24 године). У узорку од 25 њих 20 има висок притисак. Желимо 95% интервал поверења за проценат гојазних младића којимимају висок крвни притисак.

Имамо да је $n = 25$, $\sum x = 20$, $\bar{x} = \frac{20}{25} = 0.80$.
Одговарајући квантил стандардне нормалне расподеле је $z_{0.025} = 1.96$, па је интервал поверења за наш узорак (64.3%,95.7%). Примећујемо да је интервал веома широк!



Слика: $P\{-1.96 < Z < 1.96\} < 0.95$

Обим узорка за оцењивање p

- Како "скратити" интервал да би био смислен?
- Једна могућност је смањити ниво поверења, али онда губимо на његовој поузданости
- Друга могућност је повећати обим узорка, али узорци су скупи па треба да израчунамо колики је најмањи узорак који нам треба

Дужина интервала поверења је $2z_{\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}$. Вредност $\bar{X}(1-\bar{X})$ увек је мања или једнака $1/4$.
Уколико желимо да интервал буде не дужи од $2d$, тада мора да важи

$$2z_{\alpha/2}\frac{1}{\sqrt{4n}} \leq 2d,$$

односно

$$n \geq \frac{z^2_{\alpha/2}}{4d^2}.$$

Обим узорка за оцењивање p

Теорема

Потребан обим узорка за оцењивање p интервалом унапред задате дужине $2d$ је

$$n = \frac{z^2_{\alpha/2}}{4d^2}$$

У примеру о крвном притиску, да бисмо имали интервал дужине највише 0.02 (2 процента), треба да имамо

$$n = \frac{1.96^2}{4 \cdot 0.01^2} = 9604.$$

Значи, треба да испитамо 9604 особе да бисмо с поверењем од 95% проценили проценат гојазних са жељеном прецизношћу.

Тестирање хипотезе о p

Дефиниција

Постоје три теста о непознатом проценту p

- $H_0 : p = p_0$ *против* $H_1 : p \neq p_0$ *(двострани)*
- $H_0 : p = p_0$ *против* $H_1 : p < p_0$ *(једностани леви)*
- $H_0 : p = p_0$ *против* $H_1 : p > p_0$ *(једнострани десни)*

Тест статистика у сва три случаја је

$$Z_0 = \frac{\bar{X} - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n},$$

која, ако је H_0 *тачно, има приближно стандардну нормалну расподелу.*

- p*-вредност једностраног левог теста је површина лево од вредности* z_0 *коју статистика* z_0 *узме у узорку.*
- p*-вредност једностраног десног теста је површина десно од вредности* z_0
- p*-вредност двостраног теста је двострука површина лево или десно од вредности* z_0*, у зависности од тога да ли је* z_0 *негативно или позитивно*

Тестирање хипотезе о p

Проценат мањинског становништва у неком граду је 20%. Желимо да испитамо да код радника у тешкој индустрији који су припадници мањина постоји дискриминација приликом запошљавања (било позитивна или негативна).

Тестирамо $H_0 : p = 0.2$ против $H_0 : p \neq 0.2$.

У узорку од 100 радника било је 17 припадника мањина. Вредност тест статистике је

$$z_0 = \frac{0.17 - 0.2}{\sqrt{0.2 \cdot 0.8}} \sqrt{100} = -0.75.$$

Пошто је $z_0 < 0$, гледамо површину лево од z_0. Она је једнака $P\{Z < -0.75\} = 0.2266$. Пошто је тест двострани, имамо да је p-вредност $2 \cdot 0.2266 = 0.45$.

Закључак је да немамо доказа о дискриминацији код запошљавања радника.

Коришћење уграђене R функције

```
> prop.test(x=17,n=100,p=0.20,alternative="two.sided",conf.level=0.95,
correct=FALSE)

     1-sample proportions test without continuity correction

data:  17 out of 100, null probability 0.2
X-squared = 0.5625, df = 1, p-value = 0.4533
alternative hypothesis: true p is not equal to 0.2
95 percent confidence interval:
 0.1089357 0.2554800
 sample estimates:
   p
0.17
```

Оцењивање разлике процената

- Често треба да упоредимо непознате проценте p_1 и p_2 у две различите популације
- Испитујемо да ли је проценат чланова који имају одређено својство већи у некој од популација и за колико је већи
- Непознати параметар од важности је $p_1 - p_2$

Тачкаста оцена је

$$\widehat{p_1 - p_2} = \hat{p_1} - \hat{p_2} = \frac{\sum X_1}{n_1} - \frac{\sum X_2}{n_2} = \bar{X}_1 - \bar{X}_2,$$

разлика узорачких средина одговарајућих узорака.

Интервал поверења за p_1-p_2

Теорема

Нека су $\bar{X}_1$ *и* $\bar{X}_2$ *оцене процената* p_1 *и* p_2 *засноване на независним узорцима обима* n_1 *и* n_2. *Оцена* $\bar{X}_1 - \bar{X}_2$ *је непристрасна, а њена дисперзија је*

$$D(\bar{X}_1 - \bar{X}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

Теорема

$100(1-\alpha)\%$ *интервал поверења за разлику процената* $p_1 - p_2$ *је*

$$(\bar{X}_1 - \bar{X}_2 \pm z_{\alpha/2}\sqrt{\frac{\bar{X}_1(1-\bar{X}_1)}{n_1} + \frac{\bar{X}_2(1-\bar{X}_2)}{n_2}}),$$

Интервал поверења за p_1-p_2

Од 50 људи 40 је променило мишљење о уметничкој слици на основу критике Пабла Пикаса, а 30 од 60 на основу критике студента ликовне уметности. Тражимо 95% интервал поверења за разлику процената оних на чије мишљење утиче јачи, односно слабији, ауторитет.

Тачкаста оцена је $\widehat{p_1 - p_2} = \frac{40}{50} - \frac{30}{60} = 0.3$. Одговарајући квантил је $z_{0.025} = 1.96$, а интервал поверења на основу нашег узорка је (0.132,0.468).

Закључак је, с обзиром да је интервал позитиван, да је проценат оних на који утиче јачи ауторитет већи, па ауторитет има утицаја на формирање мишљења.

Тестирање хипотеза о разлици процената

Дефиниција

Постоје три теста о разлици процената $p_1 - p_2$

- $H_0 : (p_1 - p_2) = (p_1 - p_2)_0$ *против* $H_1 : (p_1 - p_2) \neq (p_1 - p_2)_0$
- $H_0 : (p_1 - p_2) = (p_1 - p_2)_0$ *против* $H_1 : (p_1 - p_2) < (p_1 - p_2)_0$
- $H_0 : (p_1 - p_2) = (p_1 - p_2)_0$ *против* $H_1 : (p_1 - p_2) > (p_1 - p_2)_0$

Тест статистика у сва три случаја је

$$Z_0 = \frac{(\bar{X}_1 - \bar{X}_2) - (p_1 - p_2)_0}{\sqrt{\bar{X}_1(1 - \bar{X}_1)/n_1 + \bar{X}_2(1 - \bar{X}_2)/n_2}},$$

која, ако је H_0 *тачно, има приближно стандардну нормалну расподелу.*

- p*-вредност једностраног левог теста је површина лево од вредности* z_0 *коју статистика* z_0 *узме у узорку.*
- p*-вредност једностраног десног теста је површина десно од вредности* z_0
- p*-вредност двостраног теста је двострука површина лево или десно од вредности* z_0*, у зависности од тога да ли је* z_0 *негативно или позитивно*

Тестирање хипотеза о разлици процената

Продавци фотокопир апарата тврде да њихова машина прави за 10% више квалитетних копија него конкурентска. Нека је p_1 проценат квалитетних копија њихове машине, а p_2 конкурентске.

Тестирамо $H_0 : p_1 - p_2 = 0.10$ против $H_1 : p_1 - p_2 > 0.10$.

Рекламирана машина је од 1000 направила 900 квалитетних, а конкурентска 711 од 900. Тачкасте оцене су $\hat{p_1} = \bar{x}_1 = 900/1000 = 0.90$, $\hat{p_2} = \bar{x}_2 = 711/900 = 0.79$, $\widehat{p_1 - p_2} = 0.11$.

Вредност тест статистике је

$$z_0 = \frac{0.90 - 0.79 - 0.10}{\sqrt{0.90 \cdot 0.10/1000 + 0.79 \cdot 0.21/900}} = 0.604.$$

Површина десно од z_0 је $P\{Z > z_0\} = 0.2743$.

Пошто је p-вредност велика, закључак је да нема доказа да је проценат квалитетних копија рекламиране машине већи.

Тестирање хипотеза о разлици процената

- Најчешћа примена овог тестирања је када је претпостављена разлика $(p_1 - p_2)_0$ једнака нули, тј. када је $H_0 : p_1 = p_2$, а H_1 може да буде $p_1 \neq p_2$, $p_1 < p_2$ или $p_1 > p_2$

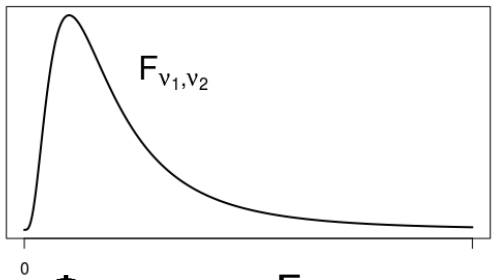
Директор спортког сектора једног универзитета жели да добије статистичку подршку својој тврдњи да студенти спортисти у мањем проценту падају године него остали студенти. У добијеним узорцима 30 студената спортиста од њих 150 је пало годину, док се то догодило и у случају 43 студента од 200 чланова остатка популације.

Тестирамо $H_0 : p_1 = p_2$ против $H_1 : p_1 < p_2$. Тачкаста оцена разлике је $\bar{x_1} - \bar{x_2} = -0.015$. Вредност статистике z_0 је -0.34. Пошто је p-вредност теста $P\{Z < -0.015\} = 0.37$, немамо довољно доказа да је спортисти ређе падају године од осталих студената.

Коришћење уграђене R функције

```
> prop.test(x=c(30,43),n=c(150,200),alternative="less",conf.level=0.95,
correct=FALSE)

     2-sample test for equality of proportions without continuity
correction

data:   c(30, 43) out of c(150, 200)
X-squared = 0.11683, df = 1, p-value = 0.3662
alternative hypothesis:  less
95 percent confidence interval:
 -1.00000000  0.05689613
 sample estimates:
prop 1 prop 2
0.200  0.215
```

Фишерова F расподела



Слика: Фишерова F расподела

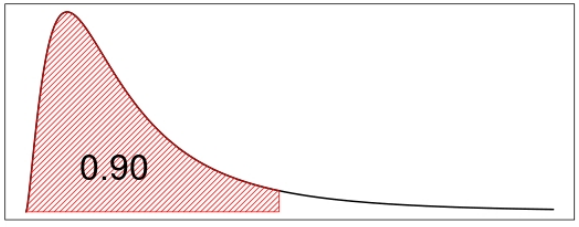
- Свака Фишерова расподела има два параметра, ν_1 и ν_2
- Фишерова расподела је непрекидна
- Вредности Фишерове расподеле увек су позитивне
- Фишерова расподела је несиметрична
- Вероватноћу догађаја $P\{F_{\nu_1,\nu_2} < x\}$ рачунамо функцијом $\mathrm{pf}(x, \mathrm{df1} = \nu_1, \mathrm{df2} = \nu_2)$.

```
> pf(3.5,df1=10,df2=6)
0.9306753
```

- Вредност x за коју је $P\{F_{\nu_1,\nu_2} < x\} = p$ рачунамо функцијом $\mathrm{qf}(p, \mathrm{df1} = \nu_1, \mathrm{df2} = \nu_2)$.

```
> qf(0.99,df1=4,df2=11)
5.6683
```

Фишерова F расподела



Слика: $P\{F_{\nu_1,\nu_2} < x\} = 0.90$

Таблица Фишерове расподеле, $p = 0.90$

$\nu_2 \backslash \nu_1$	2	3	4	5	6	7	8	10	12	15
1	49.5	53.6	55.8	57.2	58.2	59.1	59.7	60.5	61.0	61.5
2	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.39	9.41	9.43
3	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.23	5.22	5.20
4	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.92	3.90	3.87
5	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.30	3.27	3.24
6	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.94	2.90	2.87
7	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.70	2.67	2.63
8	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.54	2.50	2.46
...	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....

- Тражимо x такво да је $F_{3,5} = 0.9$. Из таблице добијамо $x = 3.62$.
- Таблице се праве за сваку вероватноћу посебно - овде је за $p = 0.90$

Упоређивање дисперзија две нормалне популације

Теорема

Нека су $X_1, \ldots, X_{n_1}$ *и* $Y_1, \ldots, Y_{n_2}$ *независни узорци из нормалних расподела* $\mathcal{N}(\mu_1, \sigma_1^2)$ *и* $\mathcal{N}(\mu_2, \sigma_2^2)$. *У случају да важи* $\sigma_1^2 = \sigma_2^2$, *случајна променљива*

$$\frac{S_1^2}{S_2^2}$$

има Фишерову F_{n_1-1, n_2-1} *расподелу.*

Упоређивање дисперзија две нормалне популације

Дефиниција

Приликом тестирања хипотеза $H_0 : \sigma_1^2 = \sigma_2^2$ *против неке од стандардних алтернатива користи се статистика*

$$F_0 = \frac{S_1^2}{S_2^2}.$$

- p*-вредност једностраног левог теста је површина лево од вредности* f_0 *коју статистика* F_0 *узме у узорку.*
- p*-вредност једностраног десног теста је површина десно од вредности* f_0
- p*-вредност двостраног теста је* **приближно** *двострука површина лево или десно од вредности* f_0*, у зависности од тога која је од тих површина мања.*

Упоређивање дисперзија две нормалне популације

Инжењер хортикултуре је направио експеримент с две нове хибридне сорте зимзеленог жбуња и важно му је да дисперзија буде што мања. На основу посматрања има индиција да сорта A има мању дисперзију. На основу узорка од 12 биљака сорте A и 10 биљака сорте B добијено је $s_A^2 = 0.0955$ и $s_B^2 = 0.1831$.

Тестираћемо једнакост дисперзија против алтернативе $\sigma_A^2 < \sigma_B^2$. Вредност тест статистике је $f_0 = \frac{s_B^2}{s_A^2} = 0.521$.

Како је расподела тест статистике под нултом хипотезом Фишерова $F_{11,9}$, одговарајућа p-вредност је $\text{pf}(0.521, \text{df1} = 11, \text{df2} = 9) = 0.15$, па немамо доказа за тврдњу да је дисперзија висине биљака сорте A већа.

Упоређивање средњих вредности две нормалне популације

- Имамо две популације чија облежја имају нормалне расподеле. Желимо да оценимо или тестирамо разлику њихових средњих вредности
- Тачкаста оцена је $\widehat{\mu_1 - \mu_2} = \bar{X}_1 - \bar{X}_2$, разлика узорачких средњих вредности
- За одређивање интервала поверења и тестирање хипотеза разликујемо два основна случаја
  - Случај независних узорака
  - Случај спарених узорака

Случај независних узорака

- Извлачимо два узорка обима n_1 из нормалне расподеле $\mathcal{N}(\mu_1, \sigma_1^2)$ случајне промењљиве X, и обима n_2 из нормалне расподеле $\mathcal{N}(\mu_2, \sigma_2^2)$ случајне промењљиве Y. Желимо да интервално оценимо или тестирамо параметар $\mu_1 - \mu_2$.
- Расподела одговарајућих статистика зависи од тога да ли претпостављамо да су дисперзије σ_1^2 и σ_2^2, иако непознате, једнаке или различите. Зато је важно најпре испитати једнакост дисперзија и примењујемо тест о једнакости дисперзија.
- Овде имамо нестандардни случај тестирања хипотезе о једнакости дисперзија јер је последица већа ако одлучимо да не одбацимо H_0. Стога тестирамо с неуобичајено великим прагом значајности $\alpha = 0.2$, тако да нам треба p-вредност теста од бар 20% да бисмо закључили да су дисперзије једнаке.
- Уколико закључимо да су дисперзије једнаке, примењујемо процедуру с оцењивањем заједничке дисперзије, а у супротном Сатервајтову процедуру.

Случај једнаких дисперзија

Теорема

Нека су $X_1, \ldots, X_{n_1}$ *и* $Y_1, \ldots, Y_{n_2}$ *независни узорци из нормалних расподела* $\mathcal{N}(\mu_1, \sigma_1^2)$ *и* $\mathcal{N}(\mu_2, \sigma_2^2)$. *Уколико је* $\sigma_1^2 = \sigma_2^2 = \sigma^2$, *случајна променљива*

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{1/n_1 + 1/n_2}}$$

има стандардну нормалну расподелу.

- Међутим, σ је непознато па га оцењујемо из узорка. Пошто је оно једнако у оба узорка оцењујемо да **заједничком узорачком дисперзијом**.

Дефиниција

Нека су S_1^2 *и* S_2^2 *узорачке дисперзије узорака обима* n_1 *и* n_2 *из популација с једнаком дисперзијом* σ^2. *Заједничка узорачка дисперзија је тада*

$$S_z^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

Случај једнаких дисперзија - интервал поверења

Теорема

Нека су $X_1, \ldots, X_{n_1}$ *и* $Y_1, \ldots, Y_{n_2}$ *независни узорци из нормалних расподела* $\mathcal{N}(\mu_1, \sigma^2)$ *и* $\mathcal{N}(\mu_2, \sigma^2)$ *и нека је* S_z *заједничка узорачка дисперзија. Тада случајна променљива*

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_z\sqrt{1/n_1 + 1/n_2}}$$

има Студентову расподелу с $n_1 + n_2 - 2$ *степени слободе.*

$100(1-\alpha)\%$ *интервал поверења за разлику средњих вредности* $\mu_1 - \mu_2$ *је*

$$\left(\bar{X}_1 - \bar{X}_2 - t_{\alpha/2}S_z\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \bar{X}_1 - \bar{X}_2 + t_{\alpha/2}S_z\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right),$$

где је $t_{\alpha/2}$ *вредност из* $T_{n_1+n_2-2}$ *расподеле такво да је површина десно од њега једнака* $\alpha/2$.

Случај једнаких дисперзија - интервал поверења

Да би се испитао утицај аспирина на сузбијање главобоље, 22 пацијента случајно је подељено у две групе. Првој групи дат је аспирин, а другој други лек. Затим је мерено време у минутима до престанка главобоље. Добијени су следећи резултати

Аспирин				Други лек			
9.9	8.0	9.5	5.9	8.2	17.3	10.1	10.2
12.2	13.5	9.6	11.5	9.1	10.5	9.7	
12.5	9.5	10.3	11.9	9.0	15.2	11.6	

Из узорка добијамо да је $\bar{x}_1 = 10.36$, и $\bar{x}_2 = 11.09$. Тачкаста оцена разлике је $\widehat{\mu_1 - \mu_2} = 10.36 - 11.09 = -0.73$.

Тестирамо прво једнакост дисперзија, тј. $H_0 : \sigma_1^2 = \sigma_2^2$ против $H_1 : \sigma_1^2 \neq \sigma_2^2$ с прагом значајности од 20%. Имамо да је $s_1^2 = 4.475$ и $s_2^2 = 8.494$, а статистика је $s_2^2/s_1^2 = 1.898$. На основу Фишерове $F_{9,11}$ расподеле тест статистике, површина десно од 1.898 је 15.7%, па је p-вредност двостраног теста већа од 20%, те закључујемо да можемо сматрати да су дисперзије једнаке.

Тражимо сада 90% интервал поверења за $\mu_1 - \mu_2$. Заједничка дисперзија је $s_z^2 = \frac{11 \cdot 4.475 + 9 \cdot 8.494}{12+10-2} = 6.284$, а $s_z = \sqrt{s_z^2} = 2.51$. Одговарајући квантил Студентове T_{20} расподеле је $t_{0.05} = 1.725$. Интервал поверења је (-2.58,1.12). С обзиром да је нула унутар овог интервала, немамо доказа да је аспирин бољи од другог лека за третман главобоље.

Случај једнаких дисперзија - T-тест

Студентов или T-тест — један од најчешће коришћених у примењеној статистици

Дефиниција

Постоје три хипотезе о разлици средњих вредности $\mu_1 - \mu_2$

- $H_0 : \mu_1 = \mu_2$ $(\mu_1 - \mu_2 = 0)$ *против* $H_1 : \mu_1 \neq \mu_2$
- $H_0 : \mu_1 = \mu_2$ *против* $H_1 : \mu_1 < \mu_2$
- $H_0 : \mu_1 = \mu_2$ *против* $H_1 : \mu_1 > \mu_2$

Тест статистика у сва три случаја је

$$T_0 = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{S_z \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

која, ако је H_0 *тачно, има приближно Студентову* $T_{n_1+n_2-2}$ *расподелу.*

- p*-вредност једностраног левог теста је површина лево од вредности* t_0 *коју статистика* T_0 *узме у узорку.*
- p*-вредност једностраног десног теста је површина десно од вредности* t_0
- p*-вредност двостраног теста је двострука површина лево или десно од вредности* t_0*, у зависности од тога да ли је* t_0 *негативно или позитивно*

Случај једнаких дисперзија

Произведене су две нове супстанце за заштиту каросерије од рђе. Случајна променљива која испитује њихов квалитет је број месеци после употребе пре него што се појави рђа. С обзиром да су супстанце нове и још нетестиране немамо никаких предзнања. Желимо да испитамо да ли су у просеку једнако квалитетне.

Тестирамо $H_0 : \mu_1 = \mu_2$ против $\mu_1 \neq \mu_2$. С обе супстанце премазано је по $n_1 = n_2 = 9$ аутомобила и добијено је $\bar{x}_1 = 16$, $s_1 = 10.1$, $\bar{x}_2 = 15$, $s_2 = 10$.

Пре него што тестирамо нашу хипотезу, проверавамо једнакост дисперзија. Имамо да је $s_1^2/s_2^2 = 1.02$, па је p-вредност теста много већа од 20%, те закључујемо да су дисперзије једнаке и рачунамо заједничку дисперзију. Добијамо да је $s_z^2 = 101.005$ и $s_z = 10.05$.

Вредност тест статистике $t_0 = \frac{\bar{x}_1 - \bar{x}_2}{s_z\sqrt{1/n_2 + 1/n_2}} = 0.199$. Како је $P\{T_{16} > 0.258\} = 0.42$, онда је p-вредност двостраног теста 0.84. Како је ова вредност велика, закључујемо да немамо доказе да постоји разлика у квалитету тих супстанци.

Коришћење уграђене R функције

```
> aspirin<-c(9.9,12.2,12.5,8.0,13.5,9.5,9.5,9.6,10.3,5.9,11.5,11.9)
> drugi.lek<-c(8.2,9.1,9.0,17.3,10.5,15.2,10.1,9.7,11.6,10.2)
> var.test(aspirin,drugi.lek,alternative="two.sided")

 F test to compare two variances

data:  aspirin and drugi.lek
F = 0.52687, num df = 11, denom df = 9, p-value = 0.314
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.134677 1.890343
 sample estimates:
ratio of variances
    0.5268664

> t.test(aspirin,drugi.lek,var.equal=TRUE,conf.level=0.90)

 Two Sample t-test

data:  aspirin and drugi.lek
t = -0.68168, df = 20, p-value = 0.5033
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence interval:
 -2.582869 1.119535
 sample estimates:
mean of x mean of y
    10.35833 11.09000
```

Случај неједнаких дисперзија

- У случају неједнаких дисперзија нема смисла да рачунамо заједничку дисперзију па користимо променљиву

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}.$$

Она има **приближно** Студентову расподелу где је број степени слободе

$$\nu = \frac{(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2})^2}{(\frac{S_1^2}{n_1})^2/(n_1 - 1) + (\frac{S_2^2}{n_2})^2/(n_2 - 1)}.$$

У пракси ν неће бити цео број па вредност заокружујемо на најближи цео број.

- Наведени поступак назива се Сатервајтовом процедуром.

Случај неједнаких дисперзија

- Интервал поверења за $\mu_1 - \mu_2$ у случају неједнаких дисперзија је

$$\left(\bar{X}_1 - \bar{X}_2 - t_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, \bar{X}_1 - \bar{X}_2 + t_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right)$$

  где је $t_{\alpha/2}$ вредност из T_ν расподеле таква да је површина десно од ње једнака α_2.

- Тест статистика за тестирање $H_0 : \mu_1 = \mu_2$ у случају неједнаких дисперзија је

$$T_0 = \frac{\bar{X}_1 - \bar{X}_2 - 0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}},$$

  и има приближно Студентову T_ν расподелу.

Случај неједнаких дисперзија

Социолози испитују разлике у генерацијама у једно од обележја од интереса је старост приликом куповине првог аутомобила.

Случајно су изабране две групе по 25 особа. У првој групи, где су особе старости преко 30 година, добијено је да је просечна старост била $\bar{x}_1 = 22.3$ и $s_1^2 = 4.52$. У другој групи, где су особде старости до 30 година добијено је $\bar{x}_2 = 18.7$ и $s_2^2 = 2.00$.

Желимо да тестирамо $H_0 : \mu_1 = \mu_2$ против $H_1 : \mu_1 > \mu_2$, тј. да су у старијој генерацији касније куповали аутомобил.

Тестирамо најпре једнакост дисперзија. Пошто је $s_1^2/s_2^2 = 2.26$ добијамо p-вредност теста од 0.03, па сходно претходном закључујемо да нису једнаке.

Вредност тест статистике $t_0 = 7.05$, а број степени слободе $\nu = 42.5 \approx 42$.

p-вредност теста практично је једнака нули и закључујемо да млађа генерација значајно раније купује свој први аутомобил.

Коришћење уграђене R функције

```
> stariji<-c(22.9,24.0,24.3,23.2,24.7,20.8,19.9,20.8,20.1,21.4,24.7,
26.2,21.1,24.7,26.8,22.4,19.4,22.6,19.3,22.2,21.6,22.9,20.5,21.4,19.6)
> mladji<-c(17.4,17.5,19.1,17.5,19.8,18.6,17.3,17.2,19.5,18.8,16.5,
19.5,18.8,21.5,18.0,19.7,18.6,20.0,19.5,18.3,18.9,16.8,17.3,18.8,22.5)
> t.test(stariji,mladji,var.equal=FALSE,alternative="greater")

 Welch Two Sample t-test

data:   stariji and mladji
t = 7.0669, df = 41.877, p-value = 5.923e-09
alternative hypothesis:  true difference in means is greater than 0
95 percent confidence interval:
 2.746176 Inf
 sample estimates:
mean of x mean of y
    22.300 18.696
```

Случај спарених узорака

- Некада је природно да сваки елемент једног узорка има свој пар у другом узорку
- Спаривање умањује утицај неке спољне променљиве која нам може сметати да откријемо стварну разлику у средњим вредностима
- Испитујемо ефиканост нове креме за сунчање следећим експериментом. Сваком појединцу намажемо једну руку и једну ногу нашом кремом, а другу руку и другу ногу конкурентском. Након три сата излагања јаком сунцу, меримо ниво изгорелости (који зависи од температуре и боје). Овакав експеримент се прави да би се неутралисао утицај различитих типова коже.

Случај спарених узорака

- Имамо два узорка где сваки елемент X_i има свој пар у узорку Y_i. Дефинишемо нову случајну променљиву $D = X - Y$ која представља разлику променљивих које испитујемо.
- Средња вредност променљиве D је $\mu_D = \mu_1 - \mu_2$ па се интервали поверења и тестирање хипотеза у вези параметра $\mu_1 - \mu_2$ своде на интервале поверења и тестирање хипотеза у вези μ_D.
- Случајна променљива

$$\frac{\bar{D} - (\mu_1 - \mu_2)}{S_D}\sqrt{n}$$

има Студентову расподелу с $n - 1$ степеном слободе.

Случај спарених узорака

Теорема

Нека су $X_1, \ldots, X_n$ и $Y_1, \ldots, Y_n$ спарени узорци из две популације чије случајне променљиве имају нормалну расподелу. $100(1-\alpha)\%$ интервал поверења за разлику $\mu_1 - \mu_2$ је

$$\left(\bar{D} - t_{\alpha/2}\frac{S_D}{\sqrt{n}}, \bar{D} - t_{\alpha/2}\frac{S_D}{\sqrt{n}}\right),$$

где је $t_{\alpha/2}$ вредност из T_{n-1} расподеле таква да је површина десно од ње једнака α_2.

- За тестирање хипотеза $H_0 : \mu_1 = \mu_2$ против $H_1 : \mu_1 \neq \mu_2$, $H_1 : \mu_1 < \mu_2$ или $H_1 : \mu_1 > \mu_2$ користи се тест статистика

$$T_0 = \frac{\bar{D}}{S_D}\sqrt{n}$$

која има Студентову расподелу с $n-1$ степеном слободе. Овај тест познат је под именом **спарени T тест**.

Случај спарених узорака

x	y	$d = x - y$
1.3	7.1	-5.8
6.0	7.5	-1.5
4.3	2.0	2.3
19.1	19.3	-0.2
7.5	4.3	3.2
2.0	7.5	-5.5
5.0	6.0	-1.0
7.9	8.3	-0.4
8.9	8.7	0.2
9.2	11.3	-2.1
6.2	7.5	-1.3
3.0	2.5	0.5
6.9	7.1	-0.2
7.6	8.3	-0.7
8.2	6.9	1.3
15.3	15.7	-0.4
14.9	13.8	1.1
6.1	7.3	-1.2
7.9	8.3	-0.4
17.5	17.9	-0.4
6.1	7.3	-1.2
5.1	4.9	0.2
13.7	13.5	0.2
14.2	17.1	-2.9
18.1	19.2	-1.1

Мери се степен изгорелости приликом коришћења средства X и Y. Желимо да покажемо да је X бољи па стога тестирамо

$H_0 : \mu_X = \mu_Y$ против $\mu_X < \mu_Y$.

Из узорка добијамо $\bar{d} = -0.69$, $s_d = 1.98$, па је вредност тест статистике $t_0 = \frac{\bar{d}}{s_d}\sqrt{n} = -1.74$.

На основу расподеле тест статистике, T_{24}, добијамо да је p-вредност теста 0.047, па ако нам је праг 0.05, закључујемо да наша крема, у просеку, ефикасније штити кожу од конкурентске.

Коришћење уграђене R функције

```
> nasa<-c(1.3,6.0,4.3,19.1,7.5,2.0,5.0,7.9,8.9,9.2,6.2,3.0,6.9,7.6,8.2,
15.3,14.9,6.1,7.9,17.5,6.1,5.1,13.7,14.2,18.1)
> konkurent<-c(7.1,7.5,2.0,19.3,4.3,7.5,6.0,8.3,8.7,11.3,7.5,2.5,7.1,
8.3,6.9,15.7,13.8,7.3,8.3,17.9,7.3,4.9,13.5,17.1,19.2)
> t.test(nasa,konkurent,paired=TRUE,alternative="less")

 Paired t-test

data:   nasa and konkurent
t = -1.7504, df = 24, p-value = 0.04641
alternative hypothesis:  true difference in means is less than 0
95 percent confidence interval:
 -Inf -0.01561205
 sample estimates:
mean of the differences
    -0.692
```

Линеарна регресија

- Укључује две променљиве, зависну (Y), и независну (x)
- Зависна променљива (Y) је она коју желимо да испитамо, а њена средња вредност и расподела зависе од друге променљиве x
- Циљ нам је да добијемо линеарну једначину која нам добро описује ту зависности
- Вредности независне променљиве x (тачке) често можемо сами да бирамо и онда узимамо узорак за Y у тим тачкама
- Испитујемо концентрацију извесног лека (Y) у зависности од времена протеклог од узимања лека (x)
- Испитујемо губитак телесне тежине (Y) у зависности од број часова аеробика недељно (x)
- Испитујемо цену пшенице (Y) у зависности од количине падавина за време сезоне (x)
- Уколико сами бирамо x, то је планирани експеримент, а ако не, онда имамо посматрани експеримент

Линеарна регресија

- Желимо да за конкретну вредност x предвидимо Y. На пример, колика је концентрација лека после 5 минута?

Тражимо $Y|x = 5$. С обзиром да људи различито реагују на лек, $Y|x = 5$ је случајна променљива. Њена (теоретска) средња вредност је $\mu_{Y|x=5}$.

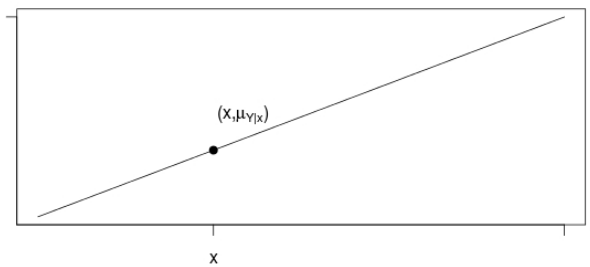
Дефиниција

Нека је x нека променљива и нека је Y случајна променљива. Регресиона крива Y на x је график функције средње вредности Y за различите вредности x, тј график функције $\mu_{Y|x}$.
За регресиону криву Y на x каже се да је линеарна ако је

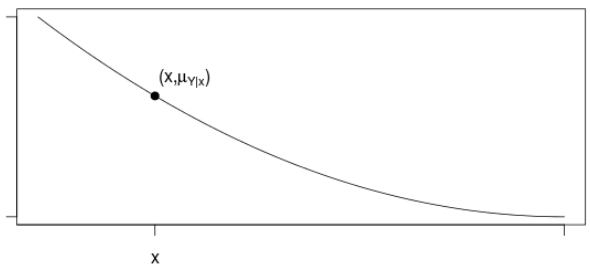
$$\mu_{Y|x} = \alpha + \beta x$$

за неке реалне бројеве α и β. Број β назива се нагибом линеарне регресије.

Линеарна регресија



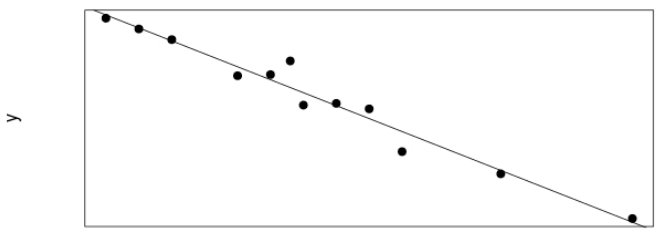
Слика: линерна регресиона крива



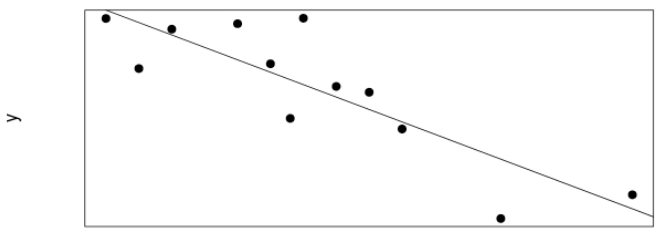
Слика: нелинеарна регресиона крива

Линеарна регресија

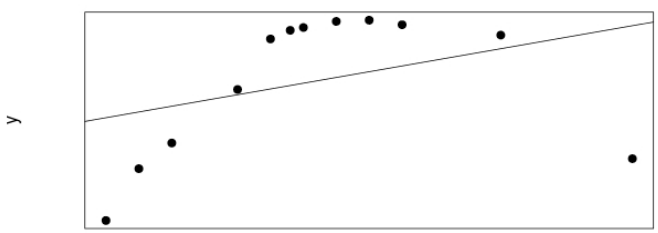
- Пре одређивања праве треба се уверити да се веза Y и x може представити линеарном функцијом
- У том циљу црта се дијаграм тачака (x_i, y_i)



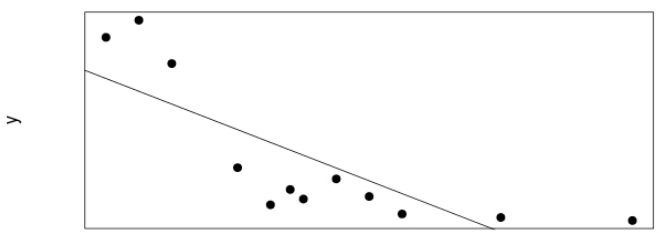
Слика: Веза је линеарна



Слика: Веза је линеарна



Слика: Веза није линеарна



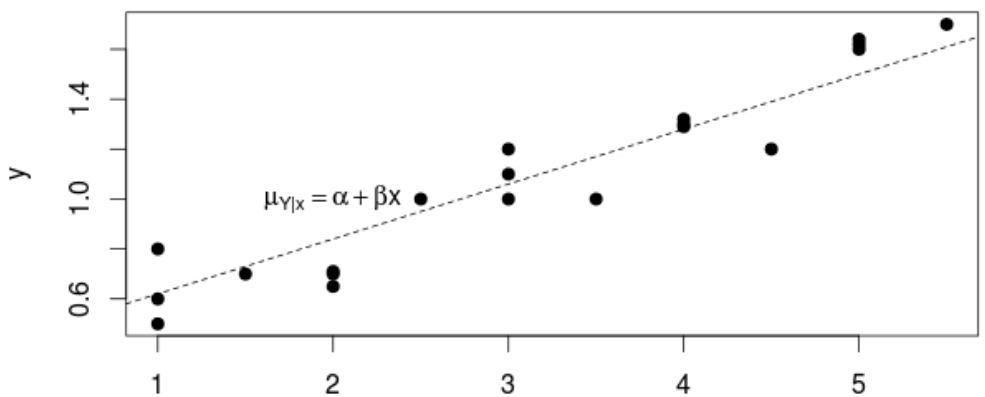
Слика: Веза није линеарна

Линеарна регресија

- Затим треба одредити једначину праве линеарне регресије на основу случајног узорка $(x_1, Y_1), \ldots, (x_n, Y_n)$. Реализација овог узорка је $(x_1, y_1), \ldots, (x_n, y_n)$.

Нека је x број сати вежбања аеробика недељно, а Y број изгубљених килограма за време фитнес програма. Добијени су следећи подаци

(1,0.5)	(2,0.7)	(3,1.1)	(4,1.3)	(5,1.6)
(1,0.8)	(2,0.65)	(3,1.2)	(4,1.29)	(5,1.62)
(1,0.6)	(2,0.71)	(3,1.0)	(4,1.32)	(5,1.64)
(1.5,0.7)	(2.5,1.0)	(3.5,1.0)	(4.5,1.2)	(5.5,1.7)



Слика: број изгубљних килограма у зависности од броја часова аеробика

Метод најмањих квадрата

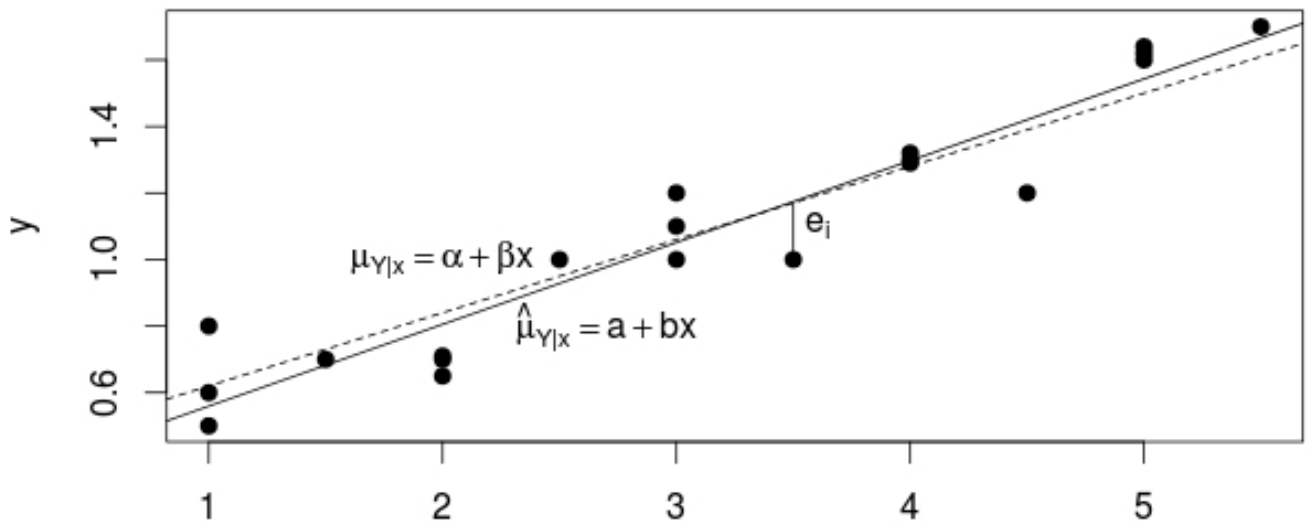
- Како добити оцену праве? Узимамо ону којој су тачке графика најближе. Меримо вертикална растојања тачака од графика, такозване **резидуале**

$$e_i = y_i - (a + bx_i),$$

а оцена праве биће за оно a и b за које је збир квадрата резидуала најмањи.

Збир квадрата обележавамо са

$$\mathrm{SSE} = \sum e^2 = \sum (y - (a + bx))^2$$



Метод најмањих квадрата

Вредности a и b за које је збир квадрата резидуала SSE најмањи су

$$b = \frac{\frac{1}{n}\sum xy - \bar{x}\bar{y}}{\frac{1}{n}\sum x^2 - \bar{x}^2}$$
$$a = \bar{y} - b\bar{x}$$

У нашем случају је $\bar{x} = 3.125$, $\bar{y} = 1.0825$, $\sum xy = 77.66$, $\sum x^2 = 236.25$, па добијамо

$$b = 0.25, a = 0.30,$$

тј.

$$\hat{\mu}_{Y|x} = 0.30 + 0.25x$$

Ако желимо да предвидимо колико се у просеку килограма изгуби ако се вежба 2.1 час недељно, добијамо $\hat{\mu}_{Y|2.1} = 0.30 + 0.25 \cdot 2.1 = 0.83$. Толика би била најбоља процена губитка телесне тежине и за конкретну особу.

Метод најмањих квадрата

Теорема

Оцене непознатих параметара линеарне регресије α и β методом најмањих квадрата су

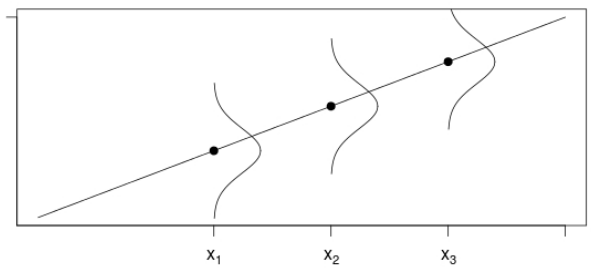
$$\hat{\beta} = B = \frac{n\sum xY - \sum x \sum Y}{n\sum x^2 - (\sum x)^2} = \frac{\frac{1}{n}\sum xY - \bar{x}\bar{Y}}{\frac{1}{n}\sum x^2 - \bar{x}^2}$$

$$\hat{\alpha} = A = \bar{Y} - B\bar{x}$$

- Предвиђање помоћу регресионе линије важи само тамо где су подаци, у нашем примеру за x између 1 и 5.5. Изван овог опсега, немамо евиденцију да је веза и даље линеарна па се не сме користити, а ако бисмо је користили често бисмо добили бесмислене или чак немогуће вредности.
- Методом најмањих квадрата у ствари тачкасто оцењујемо μ_Y (или Y) за сваку вредност x_0. Али за интервалне оцене и тестирање треба нам бољи модел који поред средње вредности описује и одступања од ње.

Прост линеарни регресиони модел

- Претпостављамо да је средња вредност $\mu_{Y|x_i} = \alpha + \beta x_i$ за свако i
- Одступање вредности Y_i од $\alpha + \beta x_i$ називамо грешком регресије и обележавамо E_i
- Претпостављамо да свако E_i има нормалну расподелу са средњом вредношћу 0 и неком дисперзијом σ^2 и да су међусобно независни



Слика: Проста линеарна регресија

Дефиниција

Прост линеарни регресиони модел је

$$Y|x_i = (\alpha + \beta x_i) + E_i,$$

где су E_i независне случајне променљиве с нормалном $\mathcal{N}(0, \sigma^2)$ расподелом.

Прост линеарни регресиони модел

- Из модела Y_i зависи од две ствари, $\alpha + \beta X$, средње вредности, и E_i, необјашњене грешке
- Модел има три непозната параметра α, β и σ^2
- α и β оцењујемо методом најмањих квадрата A и B
- Оцена грешке модела E_i у тачки i је резидуал e_i
- Оцена за σ^2 је

$$\hat{\sigma^2} = \frac{\sum (Y - (A + Bx))^2}{n - 2} = \frac{\mathrm{SSE}}{n - 2}.$$

Рачунске формуле

Дефинишимо

$$S_{yy} = \sum (y - \bar{y})^2 = \frac{n \sum y^2 - (\sum y)^2}{n} = \sum y^2 - n\bar{y}^2$$
$$S_{xx} = \sum (x - \bar{x})^2 = \frac{n \sum x^2 - (\sum x)^2}{n} = \sum x^2 - n\bar{x}^2$$
$$S_{xy} = \sum (x - \bar{x})(y - \bar{y}) = \frac{n \sum xy - (\sum x)(\sum y)}{n} = \sum xy - n\bar{x}\bar{y}$$

На основу овога добијамо

$$B = \frac{S_{xy}}{S_{xx}}$$
$$SSE = S_{yy} - BS_{xy},$$

па можемо лакше израчунати оцене параметара.

Коришћење уграђене R функције

```
> x.aerobik<-c(1,1,1,1.5,2,2,2,2.5,3,3,3,3.5,4,4,4,4.5,5,5,5,5.5)
> y.aerobik<-c(.5,.8,.6,.7,.7,.65,.71,1,1.1,1.2,1,1,1.3,1.29,1.32,
1.2,1.6,1.62,1.64,1.7)
> mod<-lm(y.aerobik~x.aerobik)
> summary(mod)

Call:
lm(formula = y.aerobik ~x.aerobik)

Residuals:
Min 1Q Median 3Q Max
-0.21960 -0.06795 0.02071 0.06113 0.24102

Coefficients:
Estimate Std.  Error t value Pr(>|t|)
(Intercept) 0.31308 0.06190 5.058 8.19e-05 ***
x.aerobik 0.24589 0.01801 13.653 6.16e-11 ***
---
Signif.  codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’  0.1 ‘ ’ 1

Residual standard error:  0.1152 on 18 degrees of freedom
Multiple R-squared:  0.9119, Adjusted R-squared:  0.907
F-statistic:  186.4 on 1 and 18 DF, p-value:  6.159e-11
```

Интервал поверења за $\mu_{Y|x_0}$

Желимо да за конкретно x_0 нађемо интервал поверења за средњу вредност Y.

Теорема

Случајна променљива

$$\frac{\hat{\mu}_{Y|x_0} - \mu_{Y|x_0}}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}}$$

где је $\hat{\sigma} = \sqrt{\frac{\mathrm{SSE}}{n-2}}$, *има Студентову расподелу с* $n-2$ *степена слободе.*

$100(1-\alpha)\%$ *интервал поверења за* $\mu_{Y|x_0}$ *је тада*

$$\left(\hat{\mu}_{Y|x_0} - t_{\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}, \hat{\mu}_{Y|x_0} + t_{\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}}\right),$$

где је $t_{\alpha/2}$ *је вредност Студентове расподеле с* $n-2$ *степена слободе тако да је површина десно од ње* $\alpha/2$.

Интервал поверења за $\mu_{Y|x_0}$

Желимо да интервално оценимо средњи број изгубљених килограма приликом 2.1 часова вежбања недељно.
Из података добијамо $\sum x = 62.50$, $\sum y = 21.63$, $\sum xy = 77.66$, $\bar{x} = 3.125$, $\sum x^2 = 236.25$, $\sum y^2 = 26.11$, $\bar{y} = 1.0825$, $n = 20$.

Рачунамо $S_{xx} = 40.94$, $S_{xy} = 10.07$, $S_{yy} = 2.72$, $b = \frac{S_{xy}}{S_{xx}} = 0.25$, $a = 0.30$, $\mu_{Y|2.1} = 0.83$.

$\mathrm{SSE} = S_{yy} - bS_{xy} = 0.20$, $\hat{\sigma}^2 = \frac{\mathrm{SSE}}{n-2} = 0.01$

За 95% интервал поверења за квантил расподеле T_{18} добијамо $t_{0.05} = 2.101$, па је интервал поверења

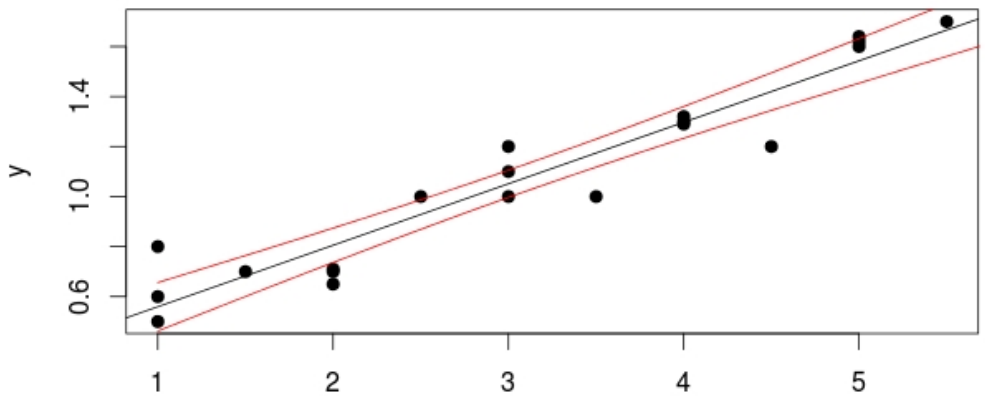
$$\left(0.83 - 2.101 \cdot 0.1\sqrt{\frac{1}{20} + \frac{(2.1-3.125)^2}{40.94}}, 0.83 - 2.101 \cdot 0.1\sqrt{\frac{1}{20} + \frac{(2.1-3.125)^2}{40.94}}\right) = (0.773, 0.887),$$

па верујемо с поверењем од 95% да је просечан број изгубљених килограма оних који вежбају 2.1 сат између 0.773 и 0.887.

```
> predict(mod,data.frame(x=2.1),interval="confidence")
fit lwr upr

1 0.8294595 0.7628636 0.8960555
```

Интервал поверења за $\mu_{Y|x_0}$

Ако направимо интервале поверења за сваку вредност x добијамо тзв. траку поверења.



Слика: Трака поверења за $\mu_{Y|x}$

Видимо да је трака најужа за $x = \bar{x}$, па нам је тада процена најпрецизнија.

```
> yy<-predict(mod,interval="confidence",
data.frame(x.aerobik=seq(1,6,length=51)))
> plot(x.aerobik,y.aerobik,pch=19,xlab="x", ylab="y")
> abline(mod)
> lines(seq(1,6,length=51),yy[,3],col="red")
> lines(seq(1,6,length=51),yy[,2],col="red")
```

Интервал предвиђања за $Y|x_0$

Желимо да за конкретно x_0 нађемо интервал предвиђања за вредност случајне променљиве Y у тој тачки.

Теорема

Случајна променљива

$$\frac{\hat{Y}|x_0 - Y|x_0}{\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}}$$

где је $\hat{\sigma} = \sqrt{\frac{\mathrm{SSE}}{n-2}}$, *има Студентову расподелу с* $n-2$ *степена слободе.*

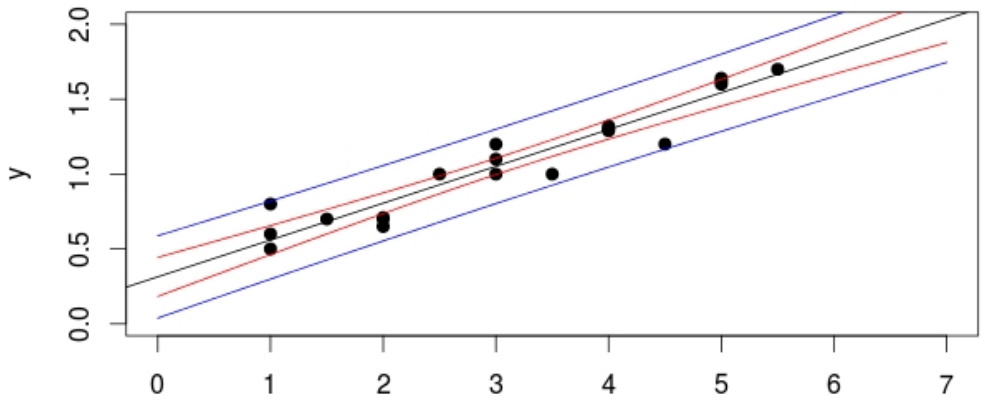
$100(1-\alpha)\%$ *интервал поверења за* $\mu_{Y|x_0}$ *је тада*

$$\left(\hat{Y}|x_0 - t_{\alpha/2}\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}, \hat{Y}|x_0 + t_{\alpha/2}\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}\right),$$

где је $t_{\alpha/2}$ *је вредност Студентове расподеле с* $n-2$ *степена слободе тако да је површина десно од ње* $\alpha/2$.

Интервал предвиђања

Тражимо 95% интервал предвиђања за број изгубљених килограма особе која намерава вежбати 2.1 час недељно. Добијамо интервал (0.612,1.048), па верујемо с поверењем од 95% да ће особа изгубити између 0.612 и 1.048 килограма.



Слика: Трака поверења за $Y|x$

Видимо да је интервал предвиђања у свакој тачки шири него интервал поверења за средњу вредност.

```
> yy2<-predict(mod,interval="prediction",
data.frame(x.aerobik=seq(1,6,length=51)))
> lines(seq(1,6,length=51),yy2[,3],col="red")
> lines(seq(1,6,length=51),yy2[,2],col="red")
```

Тестирање хипотезе о параметрима регресије

- Уколико нисмо сигурни да ли је линеарни модел применљив можемо то да тестирамо
- Тестира се нулта хипотеза
  $H_0 : \beta = 0$, тј. Y је исто за свако x и регресиони модел је непотребан, против
  $H_1 : \beta \neq 0$, тј. линеарни регресиони модел нам је користан за предвиђање Y на основу x
- Тест статистика је

$$T_0 = \frac{B}{\hat{\sigma}} \sqrt{S_{xx}},$$

  која има Студентову расподелу с $n - 2$ степена слободе
- p-вредност теста рачуна се на исти начин као код сваког двостраног теста

Тестирање хипотезе о параметрима регресије

Тестирамо да ли нам је регресиони модел изгубљене тежине у односу на број сати вежбања довољно добар да можемо да на основу њега предвиђамо број изгубљених килограма.

Имамо да је $S_{xx} = 40.94$, $b = 0.25$, $\hat{\sigma} = 0.1$ и $n = 20$, па је $t_0 = \frac{0.25}{0.1}\sqrt{40.94} = 16$. p-вредност теста је онда једнака нули, што значи да одбацујемо H_0, па нам је линеарни модел користан за предвиђање.

```
> summary(mod)

..... .....
Coefficients:
            Estimate Std.Error t value Pr(>|t|)
(Intercept) 0.31308 0.06190 5.058 8.19e-05 ***
x.aerobik 0.24589 0.01801 13.653 6.16e-11 ***
---
Signif.  codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’  0.1 ‘ ’ 1

Residual standard error:  0.1152 on 18 degrees of freedom
Multiple R-squared:  0.9119, Adjusted R-squared:  0.907
F-statistic:  186.4 on 1 and 18 DF, p-value:  6.159e-11
```

Вишеструка линеарна регресија

- Прост линеарни модел заснива се на претпоставци да вредност Y зависи од једне променљиве x
- Вишеструки линеарни модел, који претпоставља да Y зависи од више променљивих $x_1, x_2, \ldots, x_p$, је

$$Y|x_i = \beta_0 + \beta_1 x_1 + \cdots \beta_p x_p + E_i,$$

  где се, за грешке E_i, као и раније, претпоставља да су независне и да имају нормалне расподеле са средњом вредношћу 0 и истом дисперзијом σ^2.
- На основу узорка $(x_{11}, \ldots, x_{p1}, Y_1), \ldots, (x_{1n}, \ldots, x_{pn}, Y_n)$ оцењујемо параметре $\beta_0, \ldots, \beta_p$ методом најмањих квадрата
- Вишеструка линеарна регресија најпопуларнији је метод у статистичкој анализи
- Поред стандардних тачкастих и интервалних оцена, једна од најважнијих ствари у вишеструкој регресији је **избор модела**, тј. одредити које од x_j треба да постоје у формули регресије
- Пошто није могуће да цртамо вишедимензионе променљиве, морамо да добијемо одговор тестирањем; изаберемо неколико x-ева и тестирамо нулту хипотезу да су њихови коефицијенти β једнаки нули, и ако одбацимо ову хипотезу, не треба све те променљиве избацити из модела

Вишеструка линеарна регресија – пример

- База података о биљним врстама на архипелагу Галапагос
- База се састоји из 30 редова који сваки представља податке о једном острву
- Променљиве су:
  - Species која представља број различитих биљних врста на острву
  - Endemics – број ендемских врста на острву
  - Area – површина острва
  - Elevation – надморска висина највише тачке острва
  - Nearest – растојање од најближег суседног острва
  - Scruz, растојање од острва Санта Круз
  - Adjacent – површина суседног острва
- Променљиву Species узећемо за зависну, а последњих 5 као потенцијалне предикторе (независне променљиве)

```
> library(faraway)
> data(gala)
```

Вишеструка линеарна регресија – пример

- Модел са свим променљивим

```
> model <- lm(Species ~ Area + Elevation + Nearest + Scruz + Adjacent, data=gala)
> summary(model)

Call:
lm(formula = Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
    data = gala)

Residuals:
     Min       1Q   Median       3Q      Max
-111.679  -34.898   -7.862   33.460  182.584

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.068221  19.154198   0.369 0.715351
Area        -0.023938   0.022422  -1.068 0.296318
Elevation    0.319465   0.053663   5.953 3.82e-06 ***
Nearest      0.009144   1.054136   0.009 0.993151
Scruz       -0.240524   0.215402  -1.117 0.275208
Adjacent    -0.074805   0.017700  -4.226 0.000297 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60.98 on 24 degrees of freedom
Multiple R-squared:  0.7658, Adjusted R-squared:  0.7171
F-statistic:  15.7 on 5 and 24 DF,  p-value: 6.838e-07
```

Тумачење коефицијената

- У простом линеарном моделу коефицијент уз x представљао је нагиб праве
- У вишеструком линеарном моделу појединачни коефицијенти тумаче се на следећи начин: коефицијент β_j уз x_j представља за колико се повећава μ_Y када се x_j повећа за 1, а остали x-еви остану непромењени
- Проблем је што се често мењањем једног x_j у пракси мењају и други, тј. предиктори нису у општем случају независни

Дијагностика модела

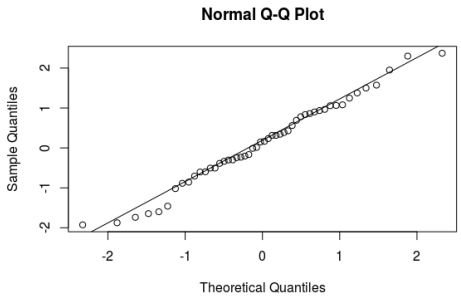
Проверавамо да ли су претпоставке модела испуњене. Имамо три претпоставке које се односе на грешке модела:

- нормална расподела грешака E_i – проверавамо QQ–дијаграмом резидуала
- једнака дисперзија за свако E_i – проверавамо графиком тачкастих оцена $\hat{y}_i$ и резидуала e_i
- независност грешака E_i – проверава се зависност суседних резидуала (ово има смисла ако су Y_i дати неким редоследом, нпр. временски следе један другог)

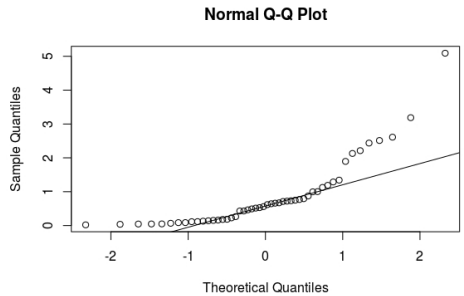
QQ–дијаграм

- QQ–дијаграм служи да се испита да ли се за неки узорак може сматрати да је из нормалне расподеле
- На овом дијаграму на x-оси налазе се теоријски квантили стандардне нормалне расподеле, а на y-оси узорачки квантили
- Користи се чињеница да је свака нормална случајна променљива линеарна функција стандардне нормалне, па и квантили случајног узорка се претпоставља да неће много одступати од праве
- Уколико тачке дијаграма не одступају много од праве, можемо сматрати да је расподела нормална, у супротном не можемо

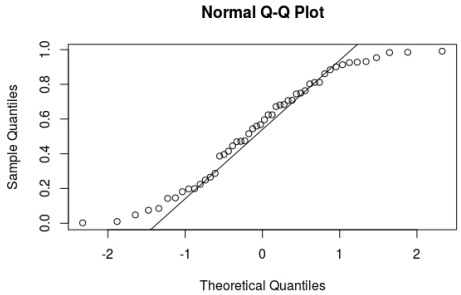
QQ–дијаграм



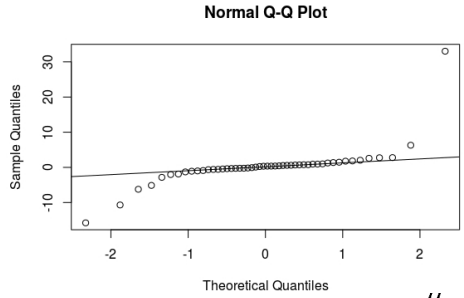
Слика: нормална расподела грешака



Слика: несиметрична расподела



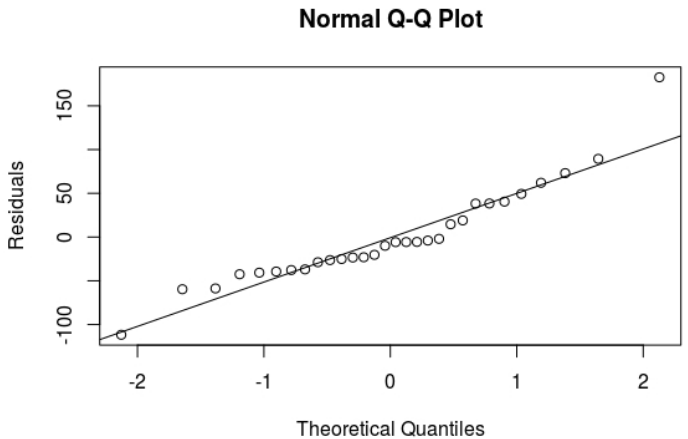
Слика: ограничена симетрична расподела



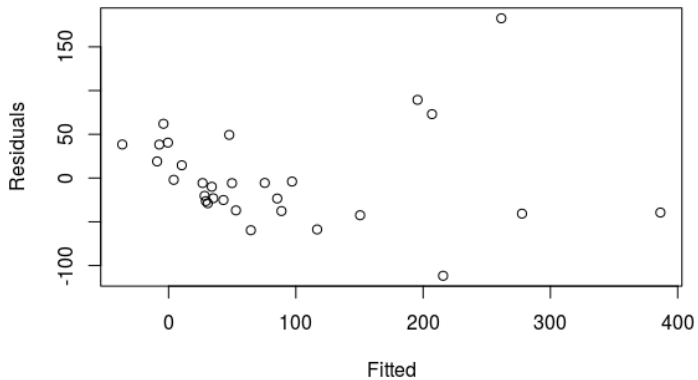
Слика: симетрична расподела “дебелих репова”

- ако је дијаграм као на графицима лево, можемо користити линеарни модел
- ако је дијаграм као горе десно, препоручује се трансформација квадратним кореном или логаритмом
- ако је дијаграм као доле десно, не можемо користити линеарни модел

Дијагностика модела



Слика: QQ-дијаграм резидуала



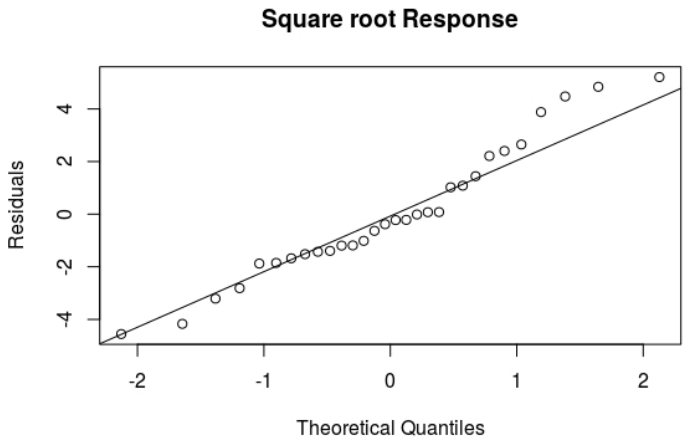
Слика: график тачастих оцена и резидуала

```
> qqnorm(residuals(model),ylab="Residuals")
> qqline(residuals(model))
> plot(fitted(model), residuals(model),xlab="Fitted",ylab="Residuals")
```

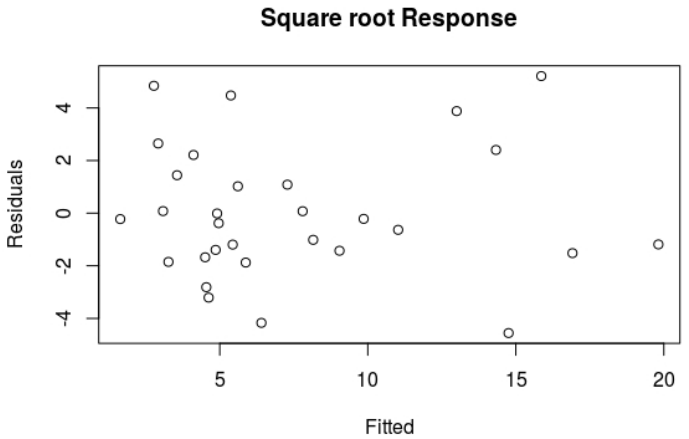
- Оба дијаграма нам говоре да наше претпоставке нису лоше, али можемо пробати да побољшамо модел трансформацијом

Дијагностика модела

Модел после трансформације квадратним кореном



Слика: QQ-дијаграм резидуала



Слика: график тачкастих оцена и резидуала

```
model.koren <- lm(sqrt(Species) ~Area + Elevation + Scruz + Nearest +
Adjacent, gala)
```

- Можемо видети да су сада претпоставке боље задовољене, међутим нови модел је обично тежи за интерпретацију

Избор модела

- Циљ нам је доћи до што једноставнијег модела (тј. са што мање променљивих) који је једнако добар као и модел са свим предикторима
- Најчешће кренемо од модела са свим потенцијалним предикторима па избацујемо један по један
- Избацивање једног предиктора мења модел (вредности коефицијената и њихову значајност), тако да се може догодити да предиктор који није био значајан у ширем моделу сада то буде, и обрнуто
- За упоређивање два модела користимо Фишеров F-тест

Упоређивање модела

- Модел с предикторима Elevation и Adjacent

```
model2 <- lm(Species ~ Elevation +  Adjacent, data=gala)

> summary(model2)

Call:
lm(formula = Species ~ Elevation + Adjacent, data = gala)

Residuals:
    Min      1Q  Median      3Q     Max
-103.41  -34.33  -11.43   22.57  203.65

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.43287   15.02469   0.095 0.924727
Elevation    0.27657    0.03176   8.707 2.53e-09 ***
Adjacent    -0.06889    0.01549  -4.447 0.000134 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60.86 on 27 degrees of freedom
Multiple R-squared:  0.7376, Adjusted R-squared:  0.7181
F-statistic: 37.94 on 2 and 27 DF,  p-value: 1.434e-08
```

Фишеров F-тест за упоређивање модела

Имамо два модела: модел A који укључује p независних променљивих, и модел B који укључује q од ових p, где је $q < p$. Тестирамо нулту хипотезу да су оба модела једнако добра, тј. $H_0 : \beta_{q+1} = 0, \ldots, \beta_p = 0$ против алтернативе да је модела A бољи, тј. да је макар један коефицијент различит од нуле.

Нека су SSE_A и SSE_B редом збирови квадрата грешака оба модела. Тада тест статистика

$$F_0 = \frac{\mathrm{SSE}_B - \mathrm{SSE}_A}{\mathrm{SSE}_A} \cdot \frac{p-q}{n-p}$$

под нултом хипотезом има Фишерову расподелу с параметрима $\nu_1 = p - q$ и $\nu_2 = n - p$.

p-вредност теста је површина **десно** од вредности f_0 коју F_0 узме у узорку.

Напомена: Ако је $q = 0$, онда је $\mathrm{SSE}_B = S_{yy}$ и хипотеза се своди на то да су сви коефицијенти β_j једнаки нули, односно да је било какав линеарни модел потребан.

Фишеров F-тест за упоређивање модела

- Упоређујемо модел са свим предикторима и модел с предикторима Elevation и Adjacent

```
> anova(model,model2)
Analysis of Variance Table

Model 1: Species ~ Area + Elevation + Nearest + Scruz + Adjacent
Model 2: Species ~ Elevation + Adjacent
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     24  89231
2     27 100003 -3    -10772 0.9657  0.425
```

- Видимо да је p-вредност велика па сматрамо да су модели једнако добри, а пошто је други модел једноставнији определићемо се за њега
- Други начин упоређивања је коефицијент детерминације R^2: видимо да се код ових модела незнатно разликује
- У неким случајевима инсистирамо да неки предиктори остану у моделу (због лакше интерпретације, односно знања из конкретне струке)

- Прости модели – Adjacent и Elevation

```
Call:
lm(formula = Species ~ Adjacent, data = gala)
---------------------------------------------
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 84.32743   22.27411   3.786 0.000744 ***
Adjacent     0.00347    0.02505   0.139 0.890831
---
Residual standard error: 116.6 on 28 degrees of freedom
Multiple R-squared:  0.0006847, Adjusted R-squared:  -0.03501
F-statistic: 0.01918 on 1 and 28 DF,  p-value: 0.8908
```

```
Call:
lm(formula = Species ~ Elevation, data = gala)
----------------------------------------------
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.33511   19.20529   0.590     0.56
Elevation    0.20079    0.03465   5.795 3.18e-06 ***
---
Residual standard error: 78.66 on 28 degrees of freedom
Multiple R-squared:  0.5454, Adjusted R-squared:  0.5291
F-statistic: 33.59 on 1 and 28 DF,  p-value: 3.177e-06
```

На основу коефицијента детерминације видимо да је први модел потпуно бесмислен, а други доста лошији него када имамо оба предиктора. Упоређивање F-тестом се своди на тестирање да је један коефицијент једнак нули.

- Прост модел – Area

```
Call:
lm(formula = Species ~ Area, data = gala)
---------------------------------------------
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 63.78286   17.52442   3.640 0.001094 **
Area         0.08196    0.01971   4.158 0.000275 ***
---
Residual standard error: 91.73 on 28 degrees of freedom
Multiple R-squared:  0.3817, Adjusted R-squared:  0.3596
F-statistic: 17.29 on 1 and 28 DF,  p-value: 0.0002748
```

Видимо да је предиктор Area у простом моделу значајан. Међутим, ако га додамо на наш модел с предикторима Elevation и Adjacent, не добијамо значајно бољи модел.

```
> model3 <- lm(Species ~ Area + Elevation + Adjacent, data=gala)
> anova(model3,model2)
Analysis of Variance Table

Model 1: Species ~ Area + Elevation + Adjacent
Model 2: Species ~ Elevation + Adjacent
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     26  96776
2     27 100003 -1   -3227.3 0.8671 0.3603
```

Корелација

- У регресионој анализи посматрали смо везу променљиве x и средње вредности $\mu_{Y|x}$ случајне променљиве Y
- У корелационој анализи, и X и Y су случајне променљиве
- Испитујемо постоји ли линеарна веза међу њима, тј. да ли важи

$$Y = \alpha + \beta X$$

Дефиниција

Нека су X и Y случајне променљиве са средњим вредностима μ_X и μ_Y.
Коваријација *између X и Y је*

$$\mathrm{Cov}(X, Y) = E(X - \mu_X)(Y - \mu_Y).$$

- Коваријација описује на који начин се X и Y истовремено одступају од својих средњих вредности
- Ако су велике вредности X кад су велике вредности Y, $\mathrm{Cov}(X, Y) > 0$
- Ако су велике вредности X кад су мале вредности Y, $\mathrm{Cov}(X, Y) < 0$
- Ако су велике вредности X подједнако повезане и с великим и с малим вредностима Y, $\mathrm{Cov}(X, Y) = 0$

Корелација

Коваријацију оцењујемо узорачком коваријацијом

$$\widehat{\mathrm{Cov}(X,Y)} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{n-1} = \frac{S_{xy}}{n-1}.$$

Испитује се веза између гојазности и крвног притиска код средовечних мушкараца. Добијени су подаци (X,Y) где је X вишак килограма, а Y горњи крвни притисак

(5,115) (20,128) (15,120) (10,118) (25,130) (28,135)

Имамо да је $\sum x = 103$, $\sum y = 746$, $\sum xy = 13145$, па је $S_{xy} = 338.67$, а $\widehat{\mathrm{Cov}(X,Y)} = \frac{S_{xy}}{n-1} = 67.734$.
Пошто је коваријација позитивна, закључујемо да је већи степен гојазности у вези с вишим крвним притиском.

```
> visak<-c(5,20,15,10,25,28)
> pritisak<-c(115,128,120,118,130,135)
> cov(visak,pritisak)
67.73333
```

- Међутим, коваријација нам не мери јачину те везе и њену вредност не можемо лако тумачити

Корелација

Дефиниција

Нека су X и Y случајне променљиве са средњим вредностима μ_X и μ_Y, и дисперзијама σ_X^2 и σ_Y^2. Пирсонов **коефицијент корелације** *између њих дефинишемо као*

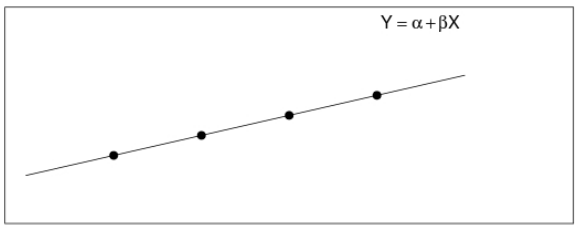
$$\rho = \frac{\operatorname{Cov}(X, Y)}{\sqrt{\sigma_X^2 \sigma_Y^2}}.$$

Теорема

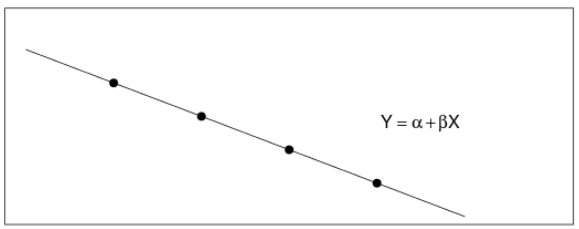
Нека су α и $\beta \neq 0$ реални бројеви. Линеарна веза X и Y постоји, тј. $Y = \alpha + \beta X$ ако и само ако је $\rho = 1$ или $\rho = -1$.

- $\rho = 1$ — постоји савршена линеарна веза с позитивном корелацијом
- $\rho = -1$ — постоји савршена линеарна веза с негативном корелацијом
- $\rho = 0$ — случајне променљиве су некорелисане, па ако постоји веза међу њима, она никако није линеарна

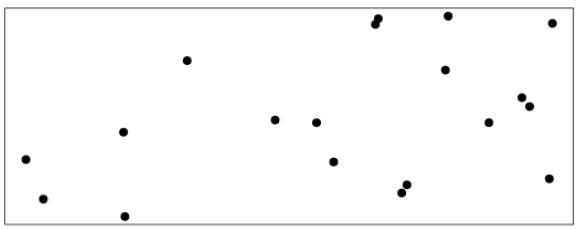
Корелација



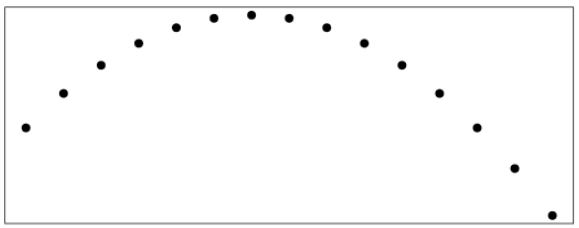
Слика: $\rho = 1$



Слика: $\rho = -1$



Слика: $\rho = 0$



Слика: $\rho = 0$

Корелација

Пирсонов коефицијент оцењује се узорачким коефицијентом корелације.

Дефиниција

Узорачки коефицијент корелације R дефинише се као

$$R = \frac{S_{xx}}{\sqrt{S_{xx}S_{yy}}} = \frac{n\sum XY - \sum x \sum y}{\sqrt{(n\sum X^2 - (\sum x)^2)(n\sum Y^2 - (\sum y)^2)}}$$

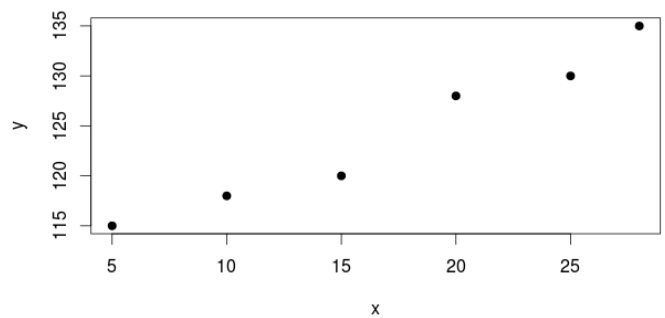
- Вредности R блиске 1 (веће од 0.75) или -1 (мање од -0.75) сматрамо добром линеарном везом
- Вредности R између 0.5 и 0.75, односно, -0.75 и -0.5, сматрамо осредњом линеарном везом
- Остале вредности R сматрамо слабом линеарном везом

Корелација

Код испитивања везе гојазности и крвног притиска имамо $\sum x = 103$, $\sum y = 746$, $\sum xy = 13145$, $\sum x^2 = 2159$, $\sum y^2 = 93058$, па је

$$r = \frac{6 \cdot 13145 - 103 \cdot 746}{\sqrt{(6 \cdot 2159 - 103^2)(6 \cdot 93058 - 746^2)}} = 0.98.$$

Вредност r је близу 1 па постоји снажна позитивна линеарна веза X и Y, што видимо и на графику.
Снажна линеарна веза не значи да гојазност узрокује висок крвни притисак, већ је могуће да постоји трећи заједнички узрочник.



Слика: крвни притисак (Y) у односу на вишак килограма (X)

```
> cor(visak,pritisak)
0.9803689
```

Веза регресије и корелације

- Узорачки коефицијент корелације R у тесној је вези с нагибом регресионе праве B

$$B = \frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}} R$$

- Знак коефицијента корелације одређује нам и знак нагиба регресије: када је $R > 0$, регресиона права расте како расте x, а када је $R < 0$ опада; када је $R = 0$, нагиб је такође једнак нули па регресиони модел није применљив
- Такође, R је у тесној вези и са збиром квадрата грешака SSE

$$R^2 = \frac{S_{yy} - \mathrm{SSE}}{S_{yy}}.$$

- Како је S_{yy} укупно одступање Y, а SSE одступање настало услед необјашене грешке, R^2 нам је проценат објашњеног одступања регресионом линијом.
- R^2 називамо **коефицијентом детерминације**.

У нашем примеру о аеробику $r = 0.95$, па је $r^2 = 0.90$. Значи да је 90% одступања који људи имају у губитку килограма приликом фитнес програма објашњено бројем часова аеробика, што је одличо. Осталих 10% не знамо да објаснимо, а наш модел их сматра случајном грешком.

Веза регресије и корелације

```
> cor(gala)
```

	Species	Endemics	Area	Elevation	Nearest	Scruz	Adjacent
Species	1.00000000	0.970876516	0.6178431	0.73848666	-0.014094067	-0.17114244	0.02616635
Endemics	0.97087652	1.000000000	0.6169791	0.79290437	0.005994286	-0.15426432	0.08265803
Area	0.61784307	0.616979087	1.0000000	0.75373492	-0.111103196	-0.10078493	0.18003759
Elevation	0.73848666	0.792904369	0.7537349	1.00000000	-0.011076984	-0.01543829	0.53645782
Nearest	-0.01409407	0.005994286	-0.1111032	-0.01107698	1.000000000	0.61541036	-0.11624788
Scruz	-0.17114244	-0.154264319	-0.1007849	-0.01543829	0.615410357	1.00000000	0.05166066
Adjacent	0.02616635	0.082658026	0.1800376	0.53645782	-0.116247885	0.05166066	1.00000000

Категорички подаци

- Понекад проучавамо случајне променљиве које не узимају вредности које се природно изражавају бројем (или узимају мали број различитих вредности). У таквим случајевима не можемо испитивати њихову повезаност коефицијентом корелације.
- Такве променљиве називамо категоричким променљивим (или факторима) а њихове вредности су категорије
- Примери су пол (две категорије: мушки и женски), да ли је особа пушач (две категорије: да или не), годишње доба (четири категорије), итд.
- Ако проучавамо повезаност две категоричке променљиве X и Y које имају r и k категорија, онда цео узорак можемо поделити у $r \cdot k$ категорија и направити **табелу контигенције**. Најпре ћемо проучити табеле контигенције 2×2.

	X	
Y	категорија x_1	категорија x_2
категорија y_1	x_1 и y_1	x_2 и y_1
категорија y_2	x_1 и y_2	x_2 и y_2

Табеле контигенције 2×2

Испитује се веза рака плућа и изложености азбесту. Случајне промењљиве да неко има рак плућа и да је изложен азбесту, имају по две категорије: ДА и НЕ. Елементе узорка класификујемо у четири категорије (ДА,ДА), (ДА,НЕ), (НЕ,ДА) и (НЕ,НЕ) и у табелу уписујемо колико је елемената узорка у одговарајућој категорији.

	изложен азбесту		
има рак плућа	да	не	
да	n_{11}	n_{12}	$n_{1\bullet} = n_{11} + n_{12}$
не	n_{21}	n_{22}	$n_{2\bullet} = n_{21} + n_{22}$
	$n_{\bullet 1} = n_{11} + n_{21}$	$n_{\bullet 2} = n_{12} + n_{22}$	n

Нека је испитано 5000 особа од којих 50 има рак плућа. Од њих је 10 било изложено азбесту. Укупно је, од 5000 особа, 500 било изложено азбесту. Добијамо табелу:

	изложен азбесту		
има рак плућа	да	не	
да	$n_{11} = 10$	$n_{12} = 40$	$n_{1\bullet} = 50$
не	$n_{21} = 490$	$n_{22} = 4460$	$n_{2\bullet} = 4950$
	$n_{\bullet 1} = 500$	$n_{\bullet 2} = 4500$	$n = 5000$

Тестирање везе између две категоричке променљиве

Разликујемо два случаја

- Тест независности: испитујемо да ли су неке две случајне променљиве независне — извлачимо узорак обима n и сваки елемент сврставамо у одговарајуће категорије, без претходног знања колико ће их у којој категорији бити.
- Тест хомогености: испитујемо да ли је код обе категорије случајне променљиве X подједнако заступаљена свака од категорија случајне променљиве Y — извлачимо узорак обима $n_1\bullet$ и $n_2\bullet$ из сваке категорије за X (укупно n), а затим их сврставамо у категорије за Y.

Тест независности

- Нулта и алтернативна хипотеза су

  H_0: X и Y су независне; H_1: X и Y нису независне
- Идеја теста је да упореди стварни број елемената узорка у свакој категорији с очекиваним бројем елемената када би X и Y биле независне.
- Оцењени број елемената у i-том реду и j-тој колони је

$$\hat{E}_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n};$$

  другим речима, он је једнак производу збира вредности i-те врсте и збира вредности j-те колоне подељен с укупним збиром.
- Тест статистика је

$$X_0^2 = \sum_{\text{по свим пољима}} \frac{(\hat{E}_{ij} - n_{ij})^2}{\hat{E}_{ij}}.$$

- Статистика X_0^2 има χ^2 расподелу с једним степеном слободе.
- p-вредност теста рачунамо као површину десно од χ_0^2, вредности коју је статистика X_0^2 узела у узорку.

Тест независности

$$\hat{E}_{11} = \frac{50 \cdot 500}{5000} = 5; \quad \hat{E}_{12} = \frac{50 \cdot 4500}{5000} = 45$$
$$\hat{E}_{21} = \frac{4950 \cdot 500}{5000} = 495; \quad \hat{E}_{22} = \frac{4950 \cdot 4500}{5000} = 4455$$

	изложен азбесту		
има рак плућа	да	не	
да	10 (5)	40 (45)	50
не	490 (495)	4460 (4455)	4950
	500	4500	5000

$$\chi_0^2 = \sum_{\text{по свим пољима}} \frac{(\hat{E}_{ij} - n_{ij})^2}{\hat{E}_{ij}} = \frac{(5-10)^2}{5} + \frac{(45-40)^2}{45} + \frac{(495-490)^2}{495} + \frac{(4455-4460)^2}{4455} = 5.61.$$

На основу χ_1^2 расподеле видимо добијамо да је p-вредност теста 0.02. Пошто је ова p-вредност мала, закључујемо да постоји веза између изложености азбесту и рака плућа.

	изложен азбесту		
има рак плућа	да	не	
да	10 (5)	40 (45)	50
не	490 (495)	4460 (4455)	4950
	500	4500	5000

Каква повезаност је у питању? Од 500 људи који су изложено азбесту, очекивали смо да 5 има рак плућа. У узорку смо добили да их је 10, тј. дупло више, па је рак плућа чешћи код оних који су изложени азбесту.

- Довољно је било израчунати само једну од очекиваних вредности, нпр $\hat{E}_{11}$. Остале се могу добити из услова да збирови по редовима и колонама морају бити једнаки вредностима на маргинама. То је повезано с тим што χ^2 расподела има 1 степен слободе.
- Ове тестове можемо примењивати када је n велико. Обично се узима да је n довољно велико ако да је свако $\hat{E}_{ij} > 5$. У супротном су p-вредности непрецизне.

Тест хомогености

- Разлика је та што се вредности на једној маргини фиксирају, тј. узорак делимо у две групе (према једној категоричкој променљивој) унапред одређене величине
- Нулта хипотеза је да исти проценат елемената има одређено својство (друга категоричка променљива) у обе групе, тј. H_0: $p_{11} = p_{21}$, а алтернативна је да је тај проценат различит H_1: $p_{11} \neq p_{21}$
- Тест статистика и рачунање p-вредности је идентично као код теста независности

Тест хомогености

Испитује се да ли је проценат оних који нису преживели операцију исти у две врсте болница: оним при истраживачким институтима и стандарднима. Узет је узорак од 139 пацијената из истраживачких и 528 из стандардних (укупно 667).

	преживели операцију		
врста болнице	не	да	
истраживачка	32	107	139 (унапред одређено)
стандардна	62	466	528 (унапред одређено)
	94	573	667

Рачунањем $\hat{E}_{ij}$ добијамо

	преживели операцију		
врста болнице	не	да	
истраживачка	32 (19.6)	107 (119.4)	139
стандардна	62 (74.4)	466 (453.6)	528
	94	573	667

Вредност тест статистике је $\chi_0^2 = 11.54$, па је p-вредност теста 0.0006 и закључујемо да проценат оних који нису преживели није исти. Из табеле видимо да је тај проценат већи у истраживачким болницама.

Табеле контигенције $r \times k$

- Претпоставимо да X има r категорија, а Y има k категорија. Тада је табела контигенције

	Y				
X	1	2	$\cdots$	k	
1	n_{11}	n_{12}	$\cdots$	n_{1k}	$n_{1\bullet}$
2	n_{21}	n_{22}	$\cdots$	n_{2k}	$n_{2\bullet}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
r	n_{r1}	n_{r2}	$\cdots$	n_{rk}	$n_{r\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet k}$	n

- Нулте и алтернативне хипотезе тестова независности и хомогености остају исте
- Тест статистика је поново

$$X_0^2 = \sum_{\text{по свим пољима}} \frac{(\hat{E}_{ij} - n_{ij})^2}{\hat{E}_{ij}},$$

а сада има χ^2 расподелу с ν степени слободе где је $\nu = (r-1)(k-1)$.

Тест хомогености у случају $r \times k$

Испитује се повезаност чира на дванаестопалачном цреву и крвне групе пацијента. Ранија истраживања указују на то да постоји веза између крвне групе O и појаве ове врсте чира. Узет је узорак од 1301 пацијента који имају чир и 6313 контролне особе и одређена им је крвна група.

Рачунамо очекиване вредности у пољима, нпр.
$\hat{E}_{11} = \frac{n_{1\bullet} \cdot n_{\bullet 1}}{n} = \frac{1301 \cdot 3590}{7614} = 613.42.$

	крвна група				
	O	A	B	AB	
пацијент	698 (613.42)	472 (529.18)	102 (114.82)	29 (43.57)	1301 (фиксно)
контролна	2892 (2976.58)	2625 (2567.82)	570 (557.18)	226 (211.43)	6313 (фиксно)
	3590	3097	672	255	7614

Вредност тест статистике је

$$\chi_0^2 = \frac{(613.42 - 698)^2}{613.42} + \cdots + \frac{(211.43 - 226)^2}{211.43} = 29.12,$$

па на основу χ^2 расподеле с $(r-1)(k-1) = 3$ степена слободе видимо да је p-вредност теста мања једнака нули, те закључујемо да постоји повезаност.

Тест хомогености у случају $r \times k$

	крвна група				
	O	A	B	AB	
пацијент	698 (613.42)	472 (529.18)	102 (114.82)	29 (43.57)	1301 (фиксно)
контролна	2892 (2976.58)	2625 (2567.82)	570 (557.18)	226 (211.43)	6313 (фиксно)
	3590	3097	672	255	7614

- Каква је повезаност у питању? Из табеле видимо да је за O крвну групу стварни број пацијената већи од очекиваног, а за остале мањи. Како је тест указао да повезаност постоји, онда је она у складу с претходним истраживањима, да је ова врста чира чешћа код људи с O крвном групом.
- Било је довољно наћи очекиване вредности у 3 поља (нпр. $\hat{E}_{11}, \hat{E}_{12}$ и $\hat{E}_{13}$, остале се израчунавају на основу збирова. То је у складу с 3 степена слободе χ^2 расподеле тест статистике.
- И овде је тест прецизан само за велике узорке, а n сматрамо довољно великим, ако ни у једном пољу очекивани број није мањи од 1 и у барем 80% поља није мањи од 5.

Коришћење уграђене R функције

```
> tabela <- matrix(c(698,472,102,29,2892,2625,570,226), nrow=2, byrow=TRUE)
> colnames(tabela) <- c("0","A","B","AB")
> rownames(tabela) <- c("Pacijenti","Kontrolna grupa")

> tabela
                   0    A   B  AB
Pacijenti        698  472 102  29
Kontrolna grupa 2892 2625 570 226


> chisq.test(tabela)

        Pearson's Chi-squared test

data:  tabela
X-squared = 29.122, df = 3, p-value = 2.111e-06
```

Једнофакторска дисперзиона анализа

- Уопштење T-теста за упоређивање средње вредности две популације
- Имамо три или више група (популација) или делимо популацију на три или више група
- Тестирамо да ли постоји разлика међу средњим вредностима група
- Уопштење независног T-теста — **једнофакторска дисперзиона анализа**
- Уопштење спареног T-теста — **рандомизирани комплетни блок дизајн**

Једнофакторска дисперзиона анализа

- Имамо k популација на којима проучавамо исто обележје. Извлаче се узорци обима n_1, n_2, ..., n_k. Свакој групи даје се исти третман. Тестирамо нулту хипотезу да су ефекти третмана исти у свим популацијама, док је алтернативна хипотеза да постоји бар нека разлика.
- Имамо једну популацију на којој желио да испитамо ефекте различитих третмана. Случајни узорак обима n делимо на k подузорака обима n_1, n_2, ..., n_k. Свака група добија различит третман. Тестирамо нулту хипотезу да су ефекти свих третмана једнаки, док је алтернативна да постоји бар нека разлика.
- У оба случаја је нулта хипотеза

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k,$$

док је алтернативна

$$H_1 : \mu_i \neq \mu_j \text{ за бар неко } i \text{ и } j.$$

Једнофакторска дисперзиона анализа

С обзиром да су збирови и средње вредности по групама (и по свим групама) важни за даљу анализу, дајемо овде њихове ознаке

- X_{ij} – j-ти елемент у i-тој групи
- n_i – број елемената у i-тој групи
- $T_{i\bullet} = \sum_{j=1}^{n_i} X_{ij}$ – збир елемената у i-тој групи
- $\bar{X}_{i\bullet} = \frac{T_{i\bullet}}{n_i}$ – средња вредност елемената у i-тој групи
- $T_{\bullet\bullet} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij} = \sum_{i=1}^{k} T_{i\bullet}$ – збир свих елемената узорка
- $\bar{X}_{\bullet\bullet} = \frac{T_{\bullet\bullet}}{n}$ – средња вредност свих елемената узорка

Једнофакторска дисперзиона анализа

Дефиниција

Модел је

$$X_{ij} = \mu + (\mu_i - \mu) + (X_{ij} - \mu_i),$$

где је

- μ *средња вредност свих популација (целе популације)*
- $\mu_i - \mu$ *ефекат* i*-те групе (*i*-тог третмана)*
- $X_{ij} - \mu_i$ *случајно (индивидуално) одступање у оквиру* i*-те групе (третмана)*

Претпоставке модела

- k узорака из k група међусобно су независни
- унутар сваке групе случајна променљива која се проучава има нормалну расподелу са средњом вредношћу μ_i и **истом** дисперзијом σ^2.

Једнофакторска дисперзиона анализа

Тачастим оцењивањем непознатих параметара модела добија се

$$X_{ij} - \bar{X}_{\bullet\bullet} = (\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet}) + (X_{ij} - \bar{X}_{i\bullet}),$$

где је $\bar{X}_{i\bullet}$ средња вредност узорка из i-те групе, док је $\bar{X}_{\bullet\bullet}$ средња вредност свих елемената свих узорака.

Сабирањем квадрата ових једначина за свако X_{ij} добија се

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{\bullet\bullet})^2 = \sum_{i=1}^{k}(\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\bullet})^2,$$

где је

- $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{\bullet\bullet})^2 = \mathrm{SST}$, укупно одступање свих елемената узорка од заједничке средње вредности (укупна варијабилност целог узорка)
- $\sum_{i=1}^{k}(\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet})^2 = \mathrm{SSG}$, одступање средњих вредности група од заједничке средње вредности (укупно одступање међу групама) — то је оно што испитујемо да ли постоји
- $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\bullet})^2 = \mathrm{SSE}$, одступање елемената узорка од средње вредности своје групе (индивидуално одступање унутар група) — случајна грешка

Једнофакторска дисперзиона анализа

$$\mathrm{SST} = \mathrm{SSG} + \mathrm{SSE}$$

Уколико је утицај SSG значајнији од SSE одбацићемо нулту хипотезу

Дефиниција

Средње одступање по групама MSG *и средње одступање унутар група* MSE *рачунају се као*

$$\mathrm{MSG} = \frac{\mathrm{SSG}}{k-1}, \quad \mathrm{MSE} = \frac{\mathrm{SSE}}{n-k}.$$

Тест статистика је

$$F_0 = \frac{\mathrm{MSG}}{\mathrm{MSE}},$$

која има Фишерову расподелу с параметрима $\nu_1 = k-1$ *и* $\nu_2 = n-k$.

p*- вредност теста је површина Фишерове* $F_{k-1,n-k}$ *расподеле* **десно** *од вредности* f_0 *коју је тест статистика узела у узорку.*

Рачунске формуле

$$\mathrm{SST} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}^2 - \frac{T_{\bullet\bullet}^2}{n},$$

$$\mathrm{SSG} = \sum_{i=1}^{k} \frac{T_{i\bullet}^2}{n_i} - \frac{T_{\bullet\bullet}^2}{n},$$

$$\mathrm{SSE} = \mathrm{SST} - \mathrm{SSG}.$$

Једнофакторска дисперзиона анализа

Социолог испитује утицај броја деце у породици на самосталност особе, на популацији бруцоша једног универзитета. Популација је подељена на четири групе, породице с једним, двоје, троје или више од троје деце. Узети су узорци обима 15,15,14 и 13 и сваком је дато да попуни анкетни лист на основу чијег је резултата процењена самосталност особе. Тестирамо нулту хипотезу $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$. Добијени подаци су:

Број деце у породици			
једно	двоје	троје	више од 3
59.1	61.2	73.4	73.1
84.4	71.0	69.3	95.7
76.0	46.6	64.9	91.1
59.5	54.0	48.7	49.7
60.1	66.6	67.7	94.9
73.4	56.6	72.5	65.8
64.1	70.5	68.8	75.8
69.4	72.8	79.9	77.2
56.4	58.5	77.7	86.2
67.1	48.7	79.2	61.1
97.6	63.3	56.7	83.1
58.5	74.8	60.1	95.6
70.7	53.1	69.8	83.8
51.8	69.9	58.2	
53.2	65.5		

Из података добијамо најпре

$$T_{1\bullet} = 1001.3$$
$$T_{2\bullet} = 933.1$$
$$T_{3\bullet} = 946.9$$
$$T_{4\bullet} = 1033.1$$
$$T_{\bullet\bullet} = 3914.4.$$

Једнофакторска дисперзиона анализа

Користећи рачунске формуле добијамо

$$\sum_{i=1}^{4}\sum_{j=1}^{n_i} x_{ij}^2 = 59.1^2 + 84.4^2 + \cdots + 83.8^2 = 277845.9$$

$$\frac{T_{\bullet\bullet}^2}{n} = \frac{3914.4^2}{57} = 268816.27$$

$$\mathrm{SST} = 277845.9 - 268816.27 = 9029.63$$

$$\sum_{i=1}^{4}\frac{T_{i\bullet}^2}{n_i} = \frac{1001.3^2}{15} + \frac{933.1^2}{15} + \frac{946.9^2}{14} + \frac{1033.1^2}{13} = 271029.07$$

$$\mathrm{SSG} = 271029.07 - 268816.27 = 2212.80$$

$$\mathrm{SSE} = \mathrm{SST} - \mathrm{SSG} = 9029.63 - 2212.80 = 6816.83$$

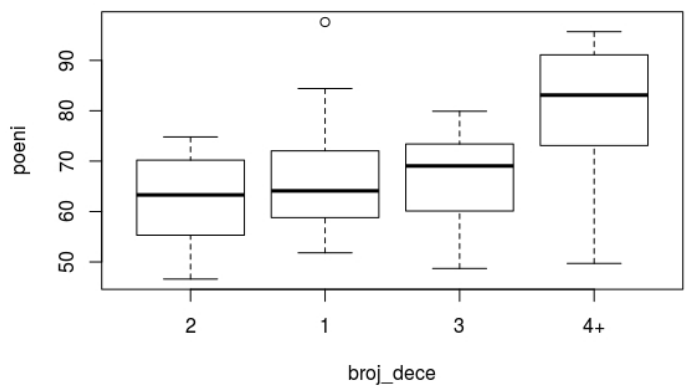
Једнофакторска дисперзиона анализа

$$\mathrm{MSG} = \frac{\mathrm{SSG}}{k-1} = \frac{2212.80}{3} = 737.60$$
$$\mathrm{MSE} = \frac{\mathrm{SSE}}{n-k} = \frac{6816.83}{53} = 128.62$$
$$f_0 = \frac{\mathrm{MSG}}{\mathrm{MSE}} = 5.73$$

p-вредност теста је од 0.002, па закључујемо да постоји утицај броја деце у породици на самосталност особе.

Коришћење уграђене R функције

```
> samostalnost <- read.csv("samostalnost.txt", sep="")
> levels(samostalnost$broj_dece)<-c("2","1","3","4+")
> plot(poeni~broj_dece,samostalnost)

> jda.samostalnost<-aov(poeni~broj_dece,samostalnost)
> summary(jda.samostalnost)
             Df Sum Sq Mean Sq F value  Pr(>F)
broj_dece     3   2213   737.6   5.735 0.00179 **
Residuals    53   6817   128.6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



Слика: Boksplot dijagrami po faktorima

Накнадно упоређивање

- Уколико немамо довољно доказа да одбацимо нулту хипотезу о једнакости средњих вредности, анализа се ту завршава. Нисмо успели да пронађемо разлику међу популацијама.
- Уколико смо одбацили нулту хипотезу, то значи да постоји бар нека разлика међу средњим вредностима, али још увек немамо одговор међу којим групама постоји разлика. Анализа није завршена, потребно је накнадно упоређивање
- Постоји више метода накнадног упоређивања: овде ћемо обрадити Бонферонијеве T-тестове и Тјукијев тест поштених значајних разлика

Бонферонијеви T-тестови

- Ако упоређујемо k група, имамо $\binom{k}{2} = k(k-1)/2$ могућих парова средњих вредности који се разликују.
- Бонферонијев метод упоређује групу сваку са сваком, тј. обавља $k(k-1)/2$ стандардних T-тестова
- Тест статистика за упоређивање μ_i и μ_j је

$$T_0 = \frac{|\bar{X}_{i\bullet} - \bar{X}_{i\bullet}|}{\sqrt{\mathrm{MSE}(\frac{1}{n_i} + \frac{1}{n_j})}},$$

  која има Студентову расподелу с параметром $n - k$.
- Пошто истовремено радимо $k(k-1)/2$ тестова, p-вредност сваког од њих мора бити мања од $\frac{2\alpha}{k(k-1)}$.
- Није увек обавезно обавити свих $k(k-1)/2$ тестова, већ истраживач може изабрати оне где сматра да ће открити разлику.

Бонферонијеви T-тестови

Испитује се утицај различите температуре на избацивање токсичних материја. Добијени су подаци:

Температура				
I	II	III	IV	V
40	36	49	47	55
45	42	51	49	60
42	38	53	51	62
48	39	53	52	63
50	37	52	50	59
51	40	50	51	61

Најпре радимо једнофакторску дисперзиону анализу. Рачунамо $\mathrm{SST} = 1648.80$, $\mathrm{SSG} = 1458.13$, $\mathrm{SSE} = 190.67$, $\mathrm{MSG} = 364.53$, $\mathrm{MSG} = 7.63$, $f_0 = 47.78$, па како је p-вредност мала, одбацујемо нулту хипотезу о непостојању утицаја различитих температура.

Желимо сада, на нивоу $\alpha = 0.1$ да откријемо за које температуре је значајно различита количина избачених материја. Пошто имамо $5 \cdot 4/2 = 10$ комбинација, треба да обавимо 10 тестова на нивоу 0.01. На пример, ако тестирамо $\mu_1 = \mu_2$ против $\mu_1 \neq \mu_2$ имамо да је

$$t_0 = \frac{|46.0 - 38.7|}{\sqrt{7.63(\frac{1}{6} + \frac{1}{6})}} = 4.58.$$

Како је, за T_{25} расподелу, p-вредност теста мања од 0.01, закључујемо да постоји разлика у количини одбачених материја при температурама I и II.

Слично се тестови обављају и у осталим случајевима.

Тјукијев тест поштених значајних разлика

Овај тест је својеврсно побољшање Бонферонијевих тестова. Тјукијева критика састоји се у томе да накнадна упоређивања нису међусобно независна и да се дисперзија оцењује на основу свих група, што мења расподелу тест статистике. Тест статистика за упоређивање μ_i и μ_j је иста као код Бонферонијевог теста,

$$Q_0 = \frac{|\bar{X}_{i\bullet} - \bar{X}_{i\bullet}|}{\sqrt{\mathrm{MSE}(\frac{1}{n_i} + \frac{1}{n_j})}},$$

а разлика је у томе што је сада њена расподела под нултом хипотезом такозвана Тјукијева расподела.

Често се резултат теста приказује преко интервала поверења

$$\bar{X}_{i\bullet} - \bar{X}_{i\bullet} \pm q_{k,N-k,1-\alpha}\sqrt{\mathrm{MSE}(\frac{1}{n_i} + \frac{1}{n_j})}$$

и интервали који не садрже нулу представљају значајне разлике између одговарајућих група.

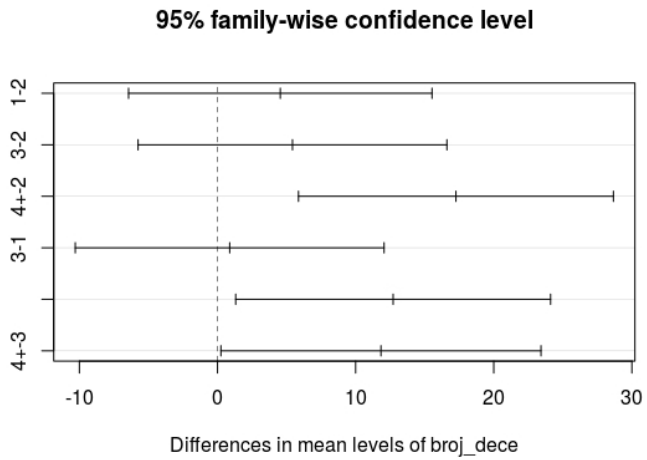
Коришћење уграђене R функције

```
> pairwise.t.test(samostalnost$poeni, samostalnost$broj_dece)

Pairwise comparisons using t tests with pooled SD

data:  samostalnost$poeni and samostalnost$broj_dece

   2      1      3
1  0.6098 -      -
3  0.6098 0.8350 -
4+ 0.0011 0.0230 0.0363

P value adjustment method: holm
```

Закључак је да постоји значајна разлика између породица с двоје и породица с више од четворо деце, док се за остала упоређивања не може рећи да постоје значајне разлике.

Тјукијев тест поштених значајних разлика

```
> TukeyHSD(jda.samostalnost)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = poeni ~ broj_dece, data = samostalnost)

$broj_dece
          diff        lwr      upr     p adj
1-2   4.546667  -6.437592 15.53093 0.6923447
3-2   5.429048  -5.749638 16.60773 0.5744046
4+-2 17.262564   5.863660 28.66147 0.0010458
3-1   0.882381 -10.296304 12.06107 0.9967146
4+-1 12.715897   1.316994 24.11480 0.0231420
4+-3 11.833516   0.247142 23.41989 0.0435942
```



Слика: Тјукијеви интервали

Рандомизирани комплетни блок дизајн

- Уопштење спареног T-теста
- Користи се када је поред утицаја на основу којег делимо у групе постоји и други утицај који желимо да контролишемо
- На основу тог утицаја делимо узорак на блокове тако да у оквиру сваке групе по један елемент узорка припада сваком од блокова

Желимо да испитамо постоји ли разлика у четири врсте асфалта за асфалтирање аутопута. Променљива коју меримо је степен истрошености након годину дана. Међутим, поред квалитета асфалта на истрошеност утичу и други фактори, као што су фреквентност саобраћаја и временске прилике. Зато одређујемо три различитих места (блокове) на којима постављамо четири различите врста асфалта (групе).

Рандомизирани комплетни блок дизајн

У општем случају имамо овакву табелу

	група					
блок	1	2	3	$\cdots$	k	
1	X_{11}	X_{21}	X_{31}	$\cdots$	X_{k1}	$T_{\bullet 1}$
2	X_{12}	X_{22}	X_{32}	$\cdots$	X_{k2}	$T_{\bullet 2}$
3	X_{13}	X_{23}	X_{33}	$\cdots$	X_{k3}	$T_{\bullet 3}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
b	X_{1b}	X_{2b}	X_{3b}	$\cdots$	X_{kb}	$T_{\bullet b}$
	$T_{1\bullet}$	$T_{2\bullet}$	$T_{3\bullet}$	$\cdots$	$T_{k\bullet}$	$T_{\bullet\bullet}$

- X_{ij} – елемент у i-тој групи и j-том блоку
- k — број група (и број елемената у оквиру једног блока)
- b — број блокова (и број елемената у оквиру једне групе)
- kb – укупан број елемената
- $T_{i\bullet} = \sum_{j=1}^{b} X_{ij}$ – збир елемената у i-тој групи
- $T_{\bullet j} = \sum_{i=1}^{k} X_{ij}$ – збир елемената у j-том блоку
- $T_{\bullet\bullet} = \sum_{i=1}^{k} \sum_{j=1}^{b} X_{ij}$ — збир свих елемената узорка

Рандомизирани комплетни блок дизајн

Дефиниција

Модел је

$$X_{ij} = \mu + (\mu_i - \mu) + (\mu_{\bullet j} - \mu) + (X_{ij} - \mu_i - \mu_{\bullet j} + \mu),$$

где је

- μ *средња вредност свих популација (целе популације)*
- $\mu_i - \mu$ *ефекат* i*-те групе (*i*-тог третмана)*
- $\mu_{\bullet j}$ *ефекат* j*-тог блока*
- $X_{ij} - \mu_i$ *случајно (индивидуално) одступање у оквиру* i*-те групе (третмана)*

Нулта и алтернативна хипотеза су исте $H_0 : \mu_1 = \cdots = \mu_k$, H_1 : постоји бар једна разлика

Рандомизирани комплетни блок дизајн

Коришћењем тачкастих оцена параметара модела добија се

$$\mathrm{SST} = \mathrm{SSG} + \mathrm{SSB} + \mathrm{SSE},$$

где је

- SST укупно одступање свих елемената узорка од заједничке средње вредности (укупна варијабилност целог узорка)
- SSG одступање средњих вредности група од заједничке средње вредности (укупно одступање међу групама)
- SSB одступање под утицајем различитих блокова
- SSE случајно индивидуално одступање

Рандомизирани комплетни блок дизајн

Дефиниција

Средње одступање по групама MSG, *средња одступања због блокова* MSB, *и средње одступање унутар група* MSE *рачунају се као*

$$\mathrm{MSG} = \frac{\mathrm{SSG}}{k-1}, \quad \mathrm{MSB} = \frac{\mathrm{SSB}}{b-1}, \quad \mathrm{MSE} = \frac{\mathrm{SSE}}{(k-1)(b-1)}.$$

Тест статистика је

$$F_0 = \frac{\mathrm{MSG}}{\mathrm{MSE}},$$

која има Фишерову расподелу с параметрима $\nu_1 = k-1$ *и* $\nu_2 = (k-1)(b-1)$.

p- *вредност теста је површина Фишерове* $F_{k-1,(k-1)(b-1)}$ *расподеле десно од вредности* f_0 *коју је тест статистика узела у узорку.*

Напомена

Блок дизајн треба користити ако је утицај блокова већи од случајног утицаја, тј. $\mathrm{MSB} > \mathrm{MSE}$, *иначе је боље користити једнофакторску дисперзиону анализу.*

Рачунске формуле

$$\mathrm{SST} = \sum_{i=1}^{k}\sum_{j=1}^{b} X_{ij}^2 - \frac{T_{\bullet\bullet}^2}{kb},$$

$$\mathrm{SSG} = \sum_{i=1}^{k} \frac{T_{i\bullet}^2}{b} - \frac{T_{\bullet\bullet}^2}{kb},$$

$$\mathrm{SSB} = \sum_{j=1}^{b} \frac{T_{\bullet j}^2}{k} - \frac{T_{\bullet\bullet}^2}{kb},$$

$$\mathrm{SSE} = \mathrm{SST} - \mathrm{SSG} - \mathrm{SSB}.$$

Рандомизирани комплетни блок дизајн

	врста асфалта				
блок	1	2	3	4	
1	42.7	39.3	48.5	32.8	$T_{\bullet 1}$ =163.3
2	50.0	38.0	49.7	40.2	$T_{\bullet 2}$ =177.9
3	51.9	46.3	53.5	51.1	$T_{\bullet 3}$ =202.8
	$T_{1\bullet}$ =144.6	$T_{2\bullet}$ =123.6	$T_{3\bullet}$ =151.7	$T_{4\bullet}$ =124.1	$T_{\bullet\bullet}$ =544.0

Најпре израчунамо $\sum_{i=1}^{k}\sum_{j=1}^{b} x_{ij}^2 = 25136.76$, $\frac{T_{\bullet\bullet}^2}{kb} = 24661.33$, $\sum_{i=1}^{k} \frac{T_{i\bullet}^2}{b} = 24.866.61$ и $\sum_{j=1}^{b} \frac{T_{\bullet j}^2}{k} = 24.860.79$.

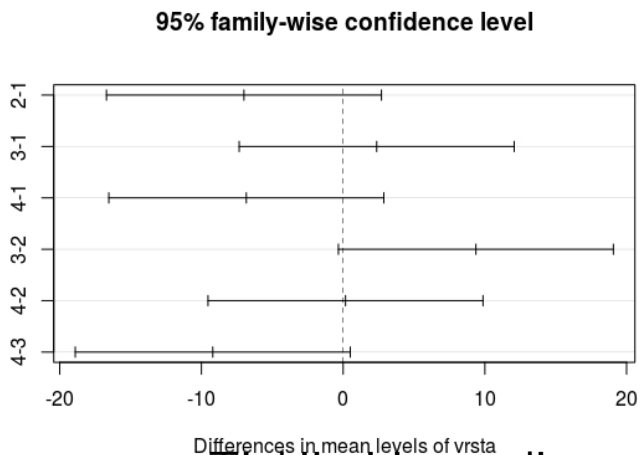
Добијамо да је $\mathrm{SST} = 205.28$, $\mathrm{SSG} = 199.46$, $\mathrm{SSB} = 475.43$, $\mathrm{SSE} = 70.69$

Следи да је $\mathrm{MSG} = 68.43$, $\mathrm{MSB} = 99.73$, $\mathrm{MSE} = 11.78$. Како је $\mathrm{MSB} > \mathrm{MSE}$, блок дизајн је добар избор модела.

Тест статистика узима вредност $f_0 = 5.81$. Добијамо да је површина десно од f_0 једнака 0.03. Значи да је наша p-вредност теста мања од 0.05, па закључујемо да постоји разлика међу различитим врстама асфалта.

Коришћење уграђене R функције

```
> asfalt<-data.frame(ostecenost=c(42.7,39.3,48.5,32.8,50.0,38.0,49.7,40.2,51.9,46.3,53.5,51.1),
blok=c("1","1","1","1","2","2","2","2","3","3","3","3"),
vrsta=c("1","2","3","4","1","2","3","4","1","2","3","4"))

> bd.asfalt <- aov(ostecenost ~ vrsta + blok, asfalt)
> summary(bd.asfalt)
            Df Sum Sq Mean Sq F value Pr(>F)
vrsta        3  205.3   68.42   5.807 0.0330 *
blok         2  199.4   99.73   8.463 0.0179 *
Residuals    6   70.7   11.78
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> plot(TukeyHSD(bd.asfalt,"vrsta"))
```



Слика: Тјукијеви интервали

Дисперзиона анализа преко линеарне регресије

- Модел једнофакторске дисперзионе анализе и линеарне регресије математички су еквивалентни
- Формирамо модел вишеструке линеарне регресије где су предиктори индикатори припадности свим категоријама фактора, осим једној (обично референтној, контролној групи).

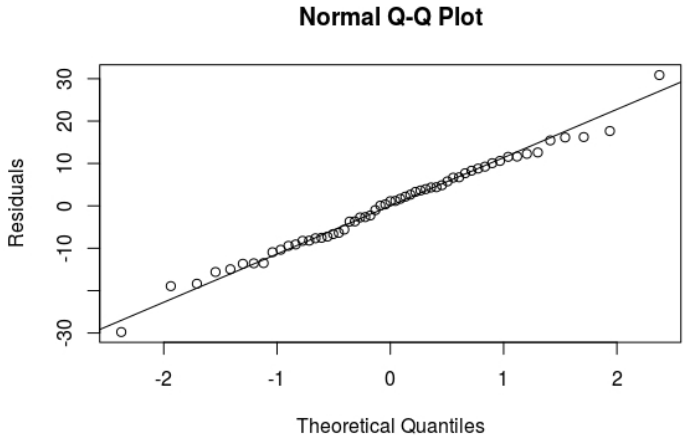
$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{k-1} x_{i(k-1)} + E_i,$$

  где су променљиве x_{ij} једнаке 1 ако елемент узорка припада j-тој групи, а 0 ако не припада, а E_i грешка која има нормалну расподелу с средњом вредношћу 0 и дисперзијом σ^2
- Тумачење је следеће: Коефицијент β_0 (Intercept) представља средњу вредност посматране (зависне) променљиве у подразумеваној (недостајућој) категорији. Коефицијенти β_j говоре за колико се мења средња вредност ако том категоријом заменимо подразумевану
- Уколико је коефицијент значајан, значи да се средња вредност значајно разликује у тој и подразумеваној категорији
- Тест статистика дисперзионе анализе поклапа се с тест статистиком теста да су сви коефицијенти једнаки нули

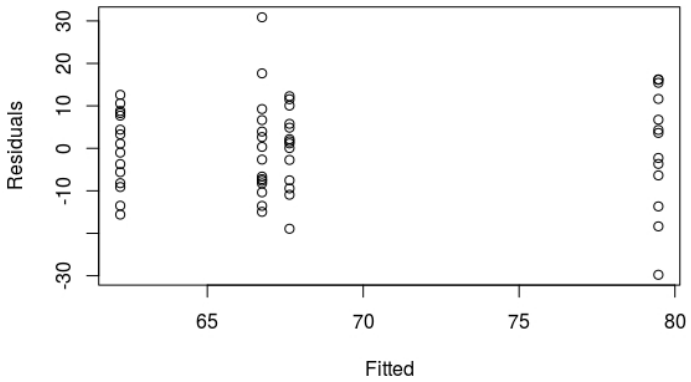
Дисперзиона анализа преко линеарне регресије

```
> lm.samostalnost<-lm(poeni~broj_dece,samostalnost)
> summary(lm.samostalnost)

Call:
lm(formula = poeni ~ broj_dece, data = samostalnost)

Residuals:
    Min      1Q  Median      3Q     Max
-29.769  -7.653   1.093   7.693  30.847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   62.207      2.928  21.244  < 2e-16 ***
broj_dece1     4.547      4.141   1.098 0.277203
broj_dece3     5.429      4.214   1.288 0.203274
broj_dece4+   17.263      4.297   4.017 0.000187 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.34 on 53 degrees of freedom
Multiple R-squared:  0.2451, Adjusted R-squared:  0.2023
F-statistic: 5.735 on 3 and 53 DF,  p-value: 0.001787
```

Дијагностика



Слика: QQ-дијаграм резидуала



Слика: график средњих вредности по групама и резидуала

```
> qqnorm(residuals(jda.samostalnost),ylab="Residuals")
> qqline(residuals(jda.samostalnost))
> plot(fitted(jda.samostalnost),
residuals(jda.samostalnost),xlab="Fitted",ylab="Residuals")
```

- Оба дијаграма нам говоре да су наше претпоставке коректне.

Други линеарни модели

- **Вишефакторска дисперзиона анализа**, или **факторски дизајн** укључује више од једног фактора који имају две или више категорија. Приликом прављења дизајна треба обратити пажњу на комбинације категорија, такозвану интеракцију међу факторима.
- **Коваријациона анализа** је модел у коме имамо комбинацију нумеричких предиктора и фактора. У овом моделу суштински правимо регресионе праве (равни или вишедимензиона уопштења) за сваку категорију понаособ.
- **Уравнотежени некомплетни блок дизајн** се користи када нисмо у могућности да у сваком блоку имамо сваку групу. Тада формирамо дизајн пазећи да у сваком блоку имамо исти број група, као и да су групе равномерно заступљене.
- **Уопштени линеарни регресиони модели** уместо средње вредности као линеарну функцију предиктора изражавају неку њену функцију (тзв. функцију везе g)

$$g(\mu_Y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p,$$

Уопштене линеарне регресије се користе када нам је зависна променљива фактор (логистичка, мултиномна регресија) и ту се користи логит функција $g(x) = \ln(\frac{x}{1-x})$ или дискретна нумеричка променљива која се не може апроксимирати нормалном (нпр. Пуасонова регресија где се користи $g(x) = \ln x$.

Тестови слободни од расподеле

Многе статистичке процедуре као што су

- T-тестови за један и два узорка
- Линеарна регресија и корелација
- Дисперзиона анализа

засновани су на претпоставци да нека случајна променљива, тј. обележје популације има **нормалну расподелу**.

- Ако расподела није нормална, онда користимо тестове **слободне од расподеле**.

Тестови на једном узорку

- Код нормалне расподеле тестирали смо хипотезу о параметру средње вредности μ
- Пошто је нормална расподела симетрична, μ је такође и њена **медијана**
- Тестови слободни од расподеле тестирају хипотезе о непознатој медијани
- Медијана расподеле је број M такав да важи

$$P\{X \leq M\} = P\{X \geq M\} = \frac{1}{2}$$

Тест знакова

- Имамо узорак $X_1, \ldots, X_n$ из расподеле која има непознату медијану M и желимо да тестирамо у вези с њеном вредношћу.
- Нулта хипотеза је $H_0 : M = M_0$, а алтернатива, у зависности шта желимо да испитамо, $M < M_0$, $M > M_0$ или $M \neq M_0$.
- Од сваког елемента узорка одузмемо M_0 и пребројимо колико има позитивних (q^+), а колико негативних разлика (q^-).
- Уколико имамо неку нулу, сматрамо је позитивном ако је $H_1 : M < M_0$, а негативном ако је $H_1 : M > M_0$
- Случајна променљива Q^+ је број позитивних разлика (у општем узорку), а Q^- број негативних разлика.
- Обе статистике Q^+ и Q^- имају биномну расподелу с параметрима n и $\frac{1}{2}$.

Тест знакова

Ако је алтернатива

- $M < M_0$, тада је p-вредност теста вероватноћа да је Q^+ мање од q^+, колико их има у нашем узорку, тј.

$$p = P\{Q^+ \leq q^+\},$$

- $M > M_0$, тада је p-вредност теста вероватноћа да је Q^- мање од q^-, колико их има у нашем узорку, тј.

$$p = P\{Q^- \leq q^-\},$$

- $M \neq M_0$, тада је p-вредност теста двострука вредност мање од ове две вероватноће.

Тест знакова

Ранија социолошка истраживања показала су да је медијална старост девојчица на првом састанку била 14 година. Сматра се да данас почињу млађе да излазе. Испитано је 15 случајно изабраних девојчица и добијен је узорак

13.0	12.5	13.5	14.2	11.5
12.5	15.0	15.5	13.5	13.0
16.0	15.5	13.7	12.0	14.5

Тестирамо $H_0 : M = 14$ против $H_1 : M < 14$. Пошто је 6 бројева из узорка већих од 14, добијамо да је $q^+ = 6$.

p-вредност теста је, на основу биномне расподеле статистике Q^+, $P\{Q^+ \leq 6\} = 0.3016$.

Пошто је p-вредност велика закључак је да немамо довољно доказа да су девојчице на првом састанку у просеку млађе од 14 година.

Коришћење уграђене R функције

```
> prvi.sastanak<-c(13.0,12.5,16.0,12.5,15.0,15.5,13.5,15.5,13.7,
14.2,13.5,12.0,11.5,13.0,14.5)
> BSDA::SIGN.test(prvi.sastanak,md=14,alternative="l")

One-sample Sign-Test

data:  prvi.sastanak
s = 6, p-value = 0.3036
alternative hypothesis: true median is less than 14
95 percent confidence interval:
  -Inf 14.61084
sample estimates:
  median of x
13.5

Achieved and Interpolated Confidence Intervals:

  Conf.Level L.E.pt  U.E.pt
Lower Achieved CI     0.9408   -Inf 14.5000
Interpolated CI       0.9500   -Inf 14.6108
Upper Achieved CI     0.9824   -Inf 15.0000
```

Вилкоксонов тест означених рангова

- Уколико имамо индиција да је расподела симетрична (а није нормална) можемо користити Вилкоксонов тест означених рангова
- Тестира се нулта хипотеза $H_0 : M = M_0$, где је M_0 претпостављена вредност медијане расподеле
- Формирају се разлике $D_1 = X_1 - M_0$, $D_2 = X_2 - M_0$, ... , $D_n = X_n - M_0$
- Апсолутне вредности $|D_i|$ поређају се по величини од најмање до највеће и свакој се додели ранг од 1 до n.
- Ако има једнаких елемената међу $|D_i|$, онда им се додељује средња вредност њихових рангова (нпр. ако су први и други једнаки, онда добијају ранг по 1.5)
- Тест статистике су

$$W_- = \sum_{\text{по негативним } D_i} R_i \quad \text{или } W_+ = \sum_{\text{по позитивним } D_i} R_i$$

- p-вредност се рачуна коришћењем Вилкоксонове расподеле означених рангова

Вилкоксонов тест означених рангова

Ако је алтернатива

- $M > M_0$, онда посматрамо статистику W_-
- $M < M_0$, онда посматрамо статистику W_+
- $M \neq M_0$, онда посматрамо **мању вредност** од W_- и W_+

p-вредност теста за вредност тест статистике w одређујемо функцијом $\mathrm{psignrank(w, n)}$

- Ако је $n = 18$, а тестирамо против алтернативе $M > M_0$ и добијемо $W_- = 35$, тада је p-вредност једнака $\mathrm{psignrank}(35, 18) = 0.013$
- Ако је $n = 21$, а тестирамо против алтернативе $M \neq M_0$ и добијемо $W_+ = 85$ и $W_- = 146$, тада је p-вредност, узимајући у обзир мању статистику W_+, једнака $2 \cdot \mathrm{psignrank}(85, 21) = 0.30$.

Вилкоксонов тест означених рангова

Године 1969. међу белцима у САД проценат неписмених био је 0.7%. Сумња се да је у већим градовима тај проценат већи. Добијен је узорак процената неписмених у 20 великих градова

0.6	0.5	0.62	1.7	0.75
1.0	0.69	0.8	0.8	0.57
0.9	1.5	0.95	0.53	1.1
1.2	2.0	0.65	0.79	0.61

Уз претпоставку о симетрији расподеле, тестирамо $H_0 : M = 0.7$ против $H_1 : M > 0.7$ тестом означених рангова.

Најпре одузмемо 0.7 од сваке вредности у табели и добијемо

-0.1	-0.2	-0.08	1.0	0.05
0.3	-0.01	0.1	0.1	-0.13
0.2	0.8	0.25	-0.17	0.4
0.5	1.3	-0.05	0.09	-0.09

Вилкоксонов тест означених рангова

Затим формирамо табелу

$\lvert D_i \rvert$	0.01	0.05	0.05	0.08	0.09	0.09	0.1	0.1	0.1	0.13
знак	$-$	$+$	$-$	$-$	$+$	$-$	$+$	$+$	$-$	$-$
ранг R_i	1	2.5	2.5	4	5.5	5.5	8	8	8	10

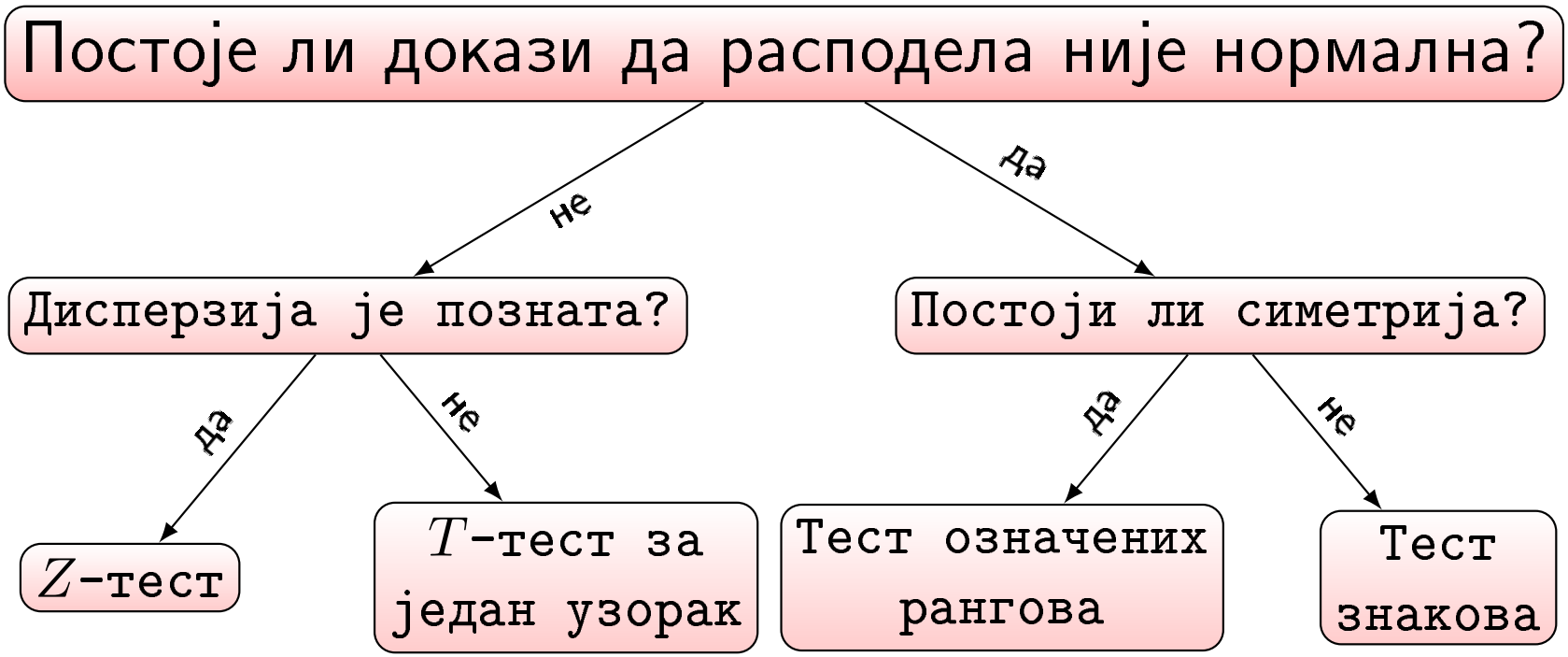
$\lvert D_i \rvert$	0.17	0.2	0.2	0.25	0.3	0.4	0.5	0.8	1.0	1.3
знак	$-$	$+$	$-$	$+$	$+$	$+$	$+$	$+$	$+$	$+$
ранг R_i	11	12.5	12.5	14	15	16	17	18	19	20

Збир рангова с негативним знаком је $W_- = 54.5$. p-вредност теста је 0.03, па на нивоу значајности 0.05 закључујемо да је медијални проценат неписмених у великим градовим већи него на националном нивоу.

Коришћење уграђене R функције

```
> procenat.pismenih<-c(0.6,1.0,0.9,1.2,0.5,0.69,1.5,2.0,0.62,0.8,0.95,
0.65,1.7,0.8,0.53,0.79,0.75,0.57,1.1,0.61)
> wilcox.test(procenat.pismenih,mu=0.7,alternative="g")

Wilcoxon signed rank test with continuity correction

data:  procenat.pismenih
V = 158, p-value = 0.02499
alternative hypothesis: true location is greater than 0.7

Warning message:
  In wilcox.test.default(procenat.pismenih, mu = 0.7, alternative = "g") :
  cannot compute exact p-value with ties
```

Схема бирања одговарајућег теста за хипотезу о параметру положаја



Случај спарених узорака

- Тест знакова и тест означених рангова можемо применити и у случају два спарена узорка $X_1, \ldots, X_n$ и $Y_1, \ldots, Y_n$
- Желимо да тестирамо да су просечне (медијалне) вредности једнаке за ове две променљиве ($H_0 : M_X = M_Y$), а не претпостављамо нормалну расподелу
- Формирамо разлике $D_1 = X_1 - Y_1$, ... , $D_n = X_n - Y_n$ и добијемо нови узорак $D_1, \ldots, D_n$
- Тестирамо хипотезу да је $M_D = 0$ против неке од алтернатива
- У случају да не претпостављамо симетрију користимо тест знакова
- У случају када претпостављамо симетрију користимо тест означених рангова

Тест знакова за спарене узорке

Произвођач хидратантних сапуна жели доказ да је његов сапун бољи од конкурентског. Узет је узорак од 10 жена које су две недеље прале једу половину лица једним, а другу другим сапуном. Затим им је измерен степен влажости коже. Добијени су резултати

сапун произвођача	5.0	4.3	7.3	2.1	9.8	6.9	10.0	1.5	8.2	7.3
конкурентски сапун	6.1	4.5	6.0	2.0	7.5	8.0	9.2	1.0	8.0	6.9
знак разлике	$-$	$-$	$+$	$+$	$+$	$-$	$+$	$+$	$+$	$+$

Тестирамо хипотезу $H_0 : M_D = M_{X-Y} = 0$ против $H_1 : M_D = M_{X-Y} > 0$. Тест статистика Q_- има биномну расподелу с параметрима 10 и $\frac{1}{2}$, а у овом узорку узела је вредност 3. p-вредност теста је

$$P\{Q^- \leq 3\} = 0.1719,$$

одакле закључујемо да немамо доказа да је сапун тог произвођача бољи од конкутентског. Лажна реклама могла имати озбиљне последице, те је не би требало правити.

Коришћење уграђене R функције

```
> sapun.nas<-c(5.0,4.3,7.3,2.1,9.8,6.9,10.0,1.5,8.2,7.3)
> sapun.njihov<-c(6.1,4.5,6.0,2.0,7.5,8.0,9.2,1.0,8.0,6.9)
> BSDA::SIGN.test(sapun.nas,sapun.njihov,alternative="g")

Dependent-samples Sign-Test

data:  sapun.nas and sapun.njihov
S = 7, p-value = 0.1719
alternative hypothesis: true median difference is greater than 0
95 percent confidence interval:
  -0.296    Inf
sample estimates:
  median of x-y
0.3

Achieved and Interpolated Confidence Intervals:

  Conf.Level L.E.pt U.E.pt
Lower Achieved CI     0.9453 -0.200    Inf
Interpolated CI       0.9500 -0.296    Inf
Upper Achieved CI     0.9893 -1.100    Inf
```

Тест означених рангова за спарене узорке

Фармацеутска компанија има два метода за испитивање квалитета лека против пчелињег убода. Сумња се да је метода A "строжија", тј. да се том методом добијају стално ниже мере квалитета лека. Добијени су подаци

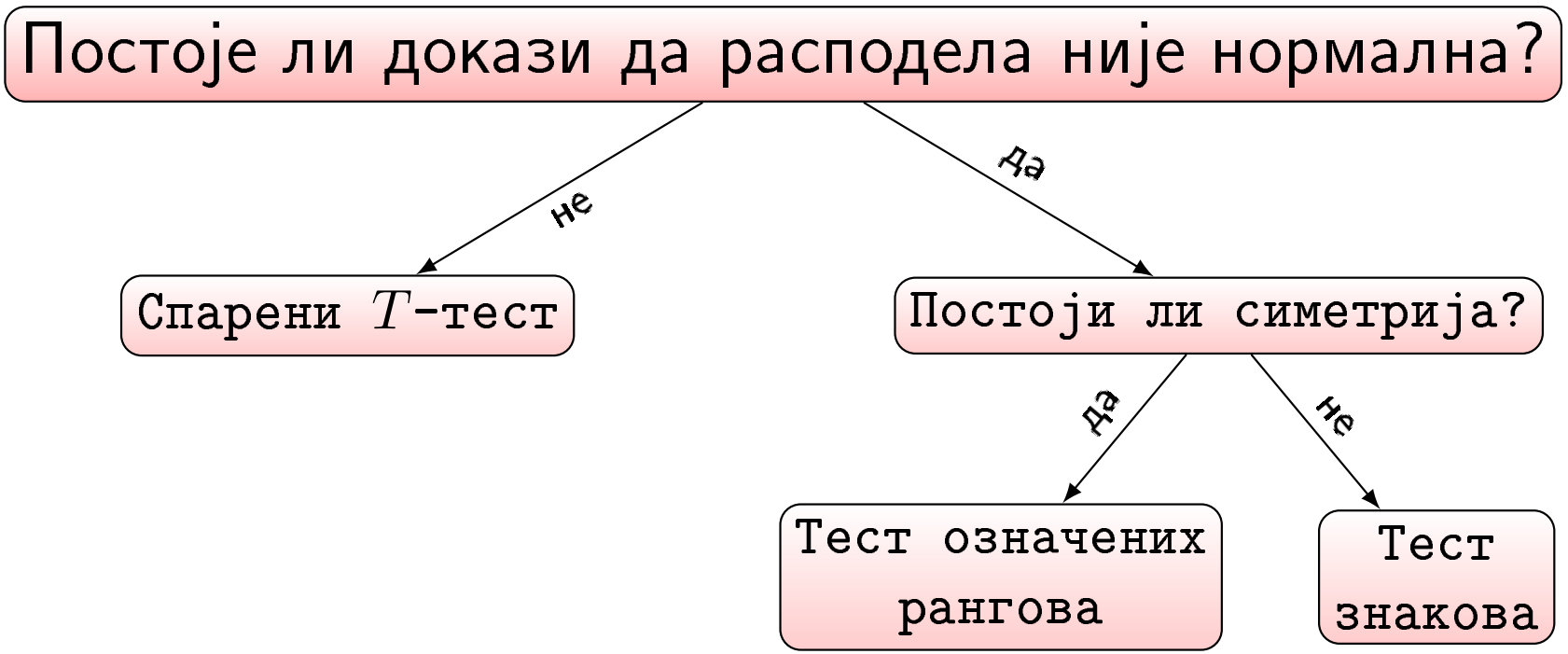
Метод A (X_i)	1.5	1.4	1.4	1.0	1.1	0.9	1.3	1.2	1.1	0.9	0.7	1.8
Метод B (Y_i)	2.0	1.8	0.7	1.3	1.2	1.5	1.1	0.9	1.5	1.7	0.9	0.9
$D_i = X_i - Y_i$	-0.5	-0.4	0.7	-0.3	-0.1	-0.6	0.2	0.3	-0.4	-0.8	-0.2	0.9
$\lvert D_i \rvert$	0.1	0.2	0.2	0.3	0.3	0.4	0.4	0.5	0.6	0.7	0.8	0.9
знак	$-$	$-$	$+$	$-$	$+$	$-$	$-$	$-$	$-$	$+$	$-$	$+$
ранг R_i	1	2.5	2.5	4.5	4.5	6.5	6.5	8	9	10	11	12

Уз претпоставку симетрије, тестирамо $H_0 : M_{A-B} = 0$ против $H_1 : M_{A-B} < 0$. Збир позитивних рангова је $W_+ = 29$. p-вредност је 0.23, те закључујемо да нема довољно доказа да метода A даје ниже мере квалитета лека.

Коришћење уграђене R функције

```
> metod.A<-c(1.5,1.4,1.4,1.0,1.1,0.9,1.3,1.2,1.1,0.9,0.7,1.8)
> metod.B<-c(2.0,1.8,0.7,1.3,1.2,1.5,1.1,0.9,1.5,1.7,0.9,0.9)
> wilcox.test(metod.A,metod.B,paired = TRUE,alternative="l")

Wilcoxon signed rank exact test

data:  metod.A and metod.B
V = 28, p-value = 0.2119
alternative hypothesis: true location shift is less than 0
```

Схема бирања теста за случај спарених узорака



Вилкоксонов тест збира рангова

- У случају да желимо да упоредимо медијане код два независна узорка користимо Вилкоксонов тест збира рангова
- Нека је један узорак $X_1, \ldots, X_m$, а други $Y_1, \ldots, Y_n$ и нека је $m \leq n$
- Тестирамо нулту хипотезу $H_0 : M_X = M_Y$
- Сваком елементу доделимо ранг који би имао у обједињеном узорку
- Статистика W_m је збир рангова елемената из мањег узорка (обима m) умањена за $m(m+1)/2$
- p-вредност се рачуна коришћењем Вилкоксонове расподеле збира рангова

Вилкоксонов тест збира рангова

Ако је алтернатива

- $M_X > M_Y$, онда p-вредност рачунамо као вероватноћу $P\{W_m > w_m\}$, где је w_m вредност из узорка
- $M_X < M_Y$, онда p-вредност рачунамо као вероватноћу $P\{W_m < w_m\}$
- $M_X \neq M_Y$, онда p-вредност рачунамо као двоструку мању вероватноћу од $P\{W_m > w_m\}$ и $P\{W_m < w_m\}$

Одговарајуће вероватноће одређујемо функцијом $\mathrm{pwilcox(w, m, n)}$

- Ако је $m = 18$, а $n = 21$, а алтернатива нам је $M_X > M_Y$, и добијемо у узорку $W_m = 254$, тада је p-вредност једнака $1 - \mathrm{pwilcox}(254, 18, 21) = 0.032$.
- Ако је $m = 21$, а $n = 21$, а алтернатива нам је $M_X \neq M_Y$, и добијемо у узорку $W_m = 174$, тада је p-вредност једнака $2 \cdot \mathrm{pwilcox}(174, 21, 21) = 0.25$.

Вилкоксонов тест збира рангова

Испитује се ефекат обуке на успешност агената осигурања. Узорак од 22 агента приправника подељен је случајно на две групе, X, који зу затим обучавани, и Y, који нису добили додатни тренинг. На крају је свако од њих тестиран у раду с клијентима и добио оцену од 0 до 10. Подаци су дати у табели.

X	8.1	7.9	9.0	4.3	7.0	9.1	7.2	8.0	9.0	3.1	
Y	9.1	6.3	2.5	6.0	0.0	2.0	7.0	5.5	1.0	9.0	9.7	5.1

Тестирамо хипотезу да је $M_X = M_Y$ против алтернативе да је $M_X > M_Y$. Формирамо табелу рангова.

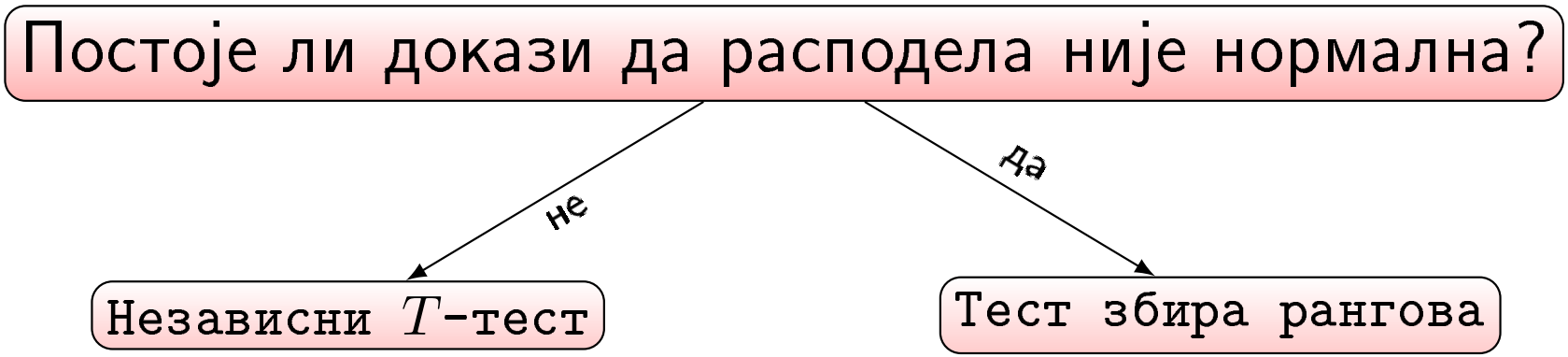
вредност	0.0	1.0	2.0	2.5	3.1	4.3	5.1	5.5	6.0	6.3	7.0
група	Y	Y	Y	Y	X	X	Y	Y	Y	Y	Y
ранг	1	2	3	4	5	6	7	8	9	10	11.5
вредност	7.0	7.2	7.9	8.0	8.1	9.0	9.0	9.0	9.1	9.1	9.7
група	X	X	X	X	X	X	X	Y	X	Y	Y
ранг	11.5	13	14	15	16	18	18	18	20.5	20.5	22

Збир рангова за X је 137, те је $w_m = 82$. p-вредност теста је 0.07, те на нивоу $\alpha = 0.05$ немамо доказа да се обуком постижу бољи резултати.

Коришћење уграђене R функције

```
> trenirani<-c(8.1,7.9,9.0,4.3,7.0,9.1,7.2,8.0,9.0,3.1)
> kontrolni<-c(9.1,6.3,2.5,6.0,0.0,2.0,7.0,5.5,1.0,9.0,9.7,5.1)
> wilcox.test(trenirani,kontrolni,alternative="g")

Wilcoxon rank sum test with continuity correction

data:  trenirani and kontrolni
W = 82, p-value = 0.07779
alternative hypothesis: true location shift is greater than 0

Warning message:
  In wilcox.test.default(trenirani, kontrolni, alternative = "g") :
  cannot compute exact p-value with ties
```

Схема бирања теста за случај независних узорака



Случај више узорака

- Имамо k група и желимо да тестирамо да су просечне вредности неког обележја у свим групама једнаке
- Уколико претпостављамо нормалну расподелу обележја, користимо једнофакторску дисперзиону анализу (ако су узорци независни) или блок дизајн (ако су зависни)
- Ако имамо доказе да расподела обележја није нормална, онда користимо
  - Крускал-Валисов тест ако су узорци независни
  - Фридманов тест ако су узорци зависни

Крускал-Валисов тест

- Имамо k група и независне узорке у свакој од њих обима n_i (укупно n)
- Тестирамо хипотезу да су им медијане једнаке $H_0 : M_1 = \cdots = M_k$
- Одредимо рангове свих елемената узорка и нека је R_i збир рангова у i-тој групи
- Тест статистика је

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1).$$

- H има χ^2 расподелу с параметром $k-1$
- p-вредност теста је површина χ^2_{k-1} расподеле **десно** од h_0, вредности коју је H узело у узорку.

Крускал-Валисов тест

Упоређује се ниво загађености три велике реке. Количина загађења измерена је на по пет места у свакој реци. Добијени су подаци (у заградама је ранг сваког елемента узорка)

прва река	друга река	трећа река
2.7 (13)	2.9 (14)	0.6 (1)
1.4 (4)	2.4 (11.5)	1.2 (2.5)
2.0 (8)	3.7 (15)	1.5 (5)
1.2 (2.5)	1.6 (6)	1.7 (7)
2.1 (9.5)	2.4 (11.5)	2.1 (9.5)

Из табеле добијамо да су збирови рангова $R_1 = 37$, $R_2 = 58$, $R_3 = 25$. Вредност тест статистике у овом узорку је

$$h_0 = \frac{12}{15 \cdot 16}\left(\frac{37^2}{5} + \frac{58^2}{5} + \frac{25^2}{5}\right) = 5.58.$$

- На основу χ^2_2 расподеле добијамо да је p-вредност теста 0.06. Закључак доносимо у зависности од нивоа значајности α. Ако је $\alpha = 0.05$, онда нема довољно доказа о различитом нивоу загађености река.

Коришћење уграђене R функције

```
> zagadjenost<-data.frame(zagadjenje=c(2.7,2.9,0.6,1.4,2.4,1.2,2.0,3.7,1.5,
1.2,1.6,1.7,2.1,2.4,2.1),reka=c(1,2,3,1,2,3,1,2,3,1,2,3,1,2,3))
> kruskal.test(zagadjenje~reka,zagadjenost)

Kruskal-Wallis rank sum test

data:  zagadjenje by reka
Kruskal-Wallis chi-squared = 5.6101, df = 2, p-value = 0.06051
```

Фридманов тест

- Имамо k група и b блокова, укупно kb елемената (исто као код блок дизајна)
- Тестирамо хипотезу да су им медијане једнаке $H_0 : M_1 = \cdots = M_k$
- Одредимо рангове елемената **у оквиру својих блокова** и нека је R_i збир рангова у i-тој групи
- Тест статистика је

$$S = \frac{12}{bk(k+1)} \sum_{i=1}^{k} \left(R_i - \frac{b(k+1)}{2} \right)^2.$$

- S има χ^2 расподелу с параметром $k-1$
- p-вредност теста је површина χ^2_{k-1} расподеле **десно** од s_0, вредности коју је S узело у узорку.

Фридманов тест

Упоређује се квалитет три врсте кочнице за бицикле. Сматра се да и марка бицикле има утицаја на перформансу кочница, па је узето шест познатијих марки и формирано 6 блокова. Свака кочница тестирана је на сваком бициклу и мерен је број недеља коришћења пре првог сервиса. Добијени су подаци (у заградама је ранг по блоку)

	врста кочнице		
марка бицикла	A	B	C
S	5.2 (2)	7.3 (3)	3.0 (1)
V	6.8 (1)	8.9 (3)	7.5 (2)
JH	6.3 (2.5)	6.3 (2.5)	6.0 (1)
R	13.0 (1.5)	14.8 (3)	13.0 (1.5)
C	12.8 (2.5)	12.8 (2.5)	11.0 (1)
Ra	15.0 (2)	15.2 (3)	14.5 (1)

Из табеле добијамо да су збирови рангова $R_1 = 11.5$, $R_2 = 17$, $R_3 = 7.5$. Вредност тест статиситке у овом узорку је

$$s_0 = \frac{12}{6 \cdot 3 \cdot 4}\left((11.5 - 12)^2 + (17 - 12)^2 + (7.5 - 12)^2)\right) = 7.58.$$

- На основу χ^2_2 расподеле добијамо је p-вредност теста 0.02, што значи да закључујемо да постоји разлика између ових врста кочница. Судећи по томе што кочница B има увек највиши ранг, препоручује се њена употреба.

Коришћење уграђене R функције

```
> bicikle<-data.frame(rok.trajanja=c(5.2,7.3,3.0,6.8,8.9,7.5,6.3,6.3,6.0,
13.0,14.8,13.0,12.8,12.8,11.0,15.0,15.2,14.5),blok=c(1,1,1,2,2,2,3,3,3,
4,4,4,5,5,5,6,6,6),vrsta=c(1,2,3,1,2,3,1,2,3,1,2,3,1,2,3,1,2,3))
> friedman.test(rok.trajanja~vrsta|blok,bicikle)

Friedman rank sum test

data:  rok.trajanja and vrsta and blok
Friedman chi-squared = 8.6667, df = 2, p-value = 0.01312
```

Схема бирања теста за случај више узорака