Биостатистика и анализа 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

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

Дефиниција

Непрекидна случајна промењива *је случајна промењива, која, пре изведеног експеримента, може узети било коју вредност из неког интервала реалних бројева.*
Дискретна случајна промењива *је случајна промењива, 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

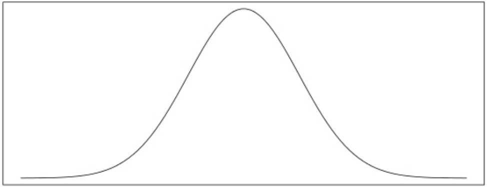
Снага земљотреса у Калифорнији по Рихтеровој скали – пример узорка

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 промењиве 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.

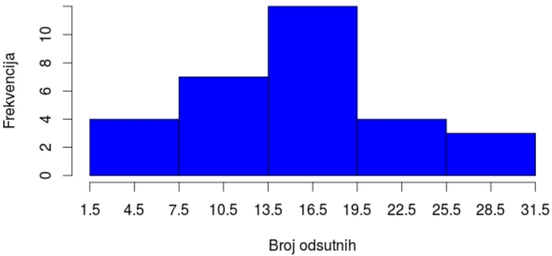
Конструкција хистограма

- 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

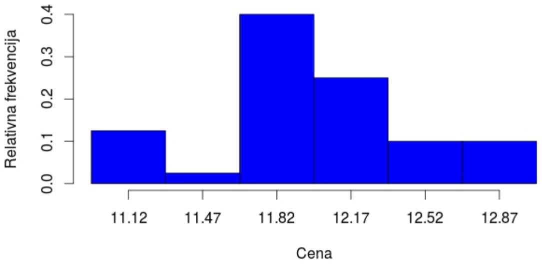
Дефиниција

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

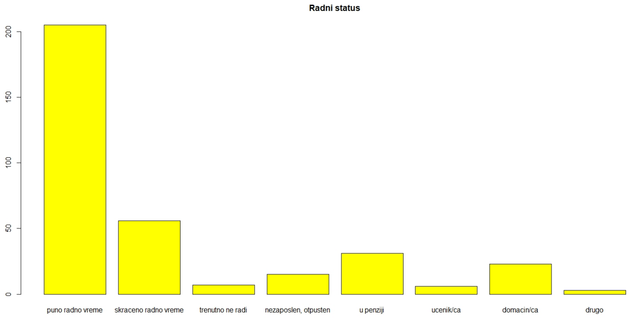
```
> levels(baza1$pol)=c("M","Z")

> 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

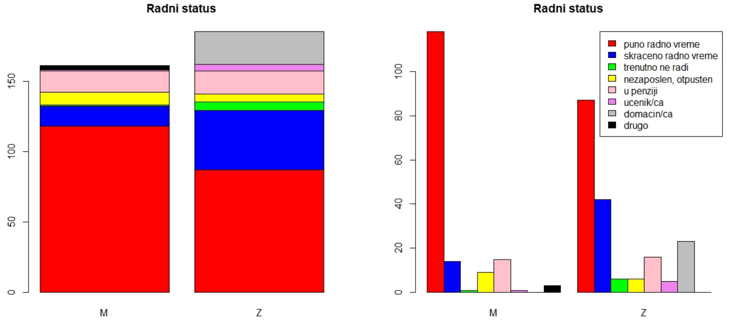
Тракасти дијаграм на $x-$оси приказује категорије, а на $y-$оси фраквенције појављивања елемената из узорка у свакој од категорија

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



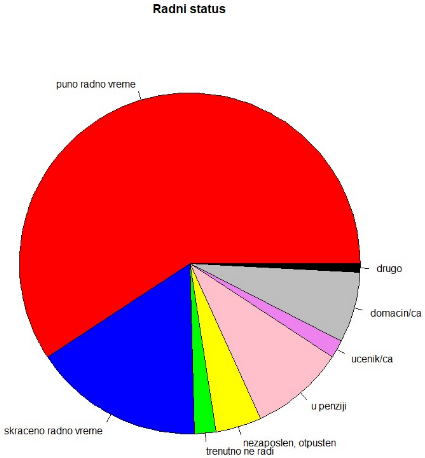
Може се цртати једна променљива у односу на категорије друге променљиве.

```
> 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)
```

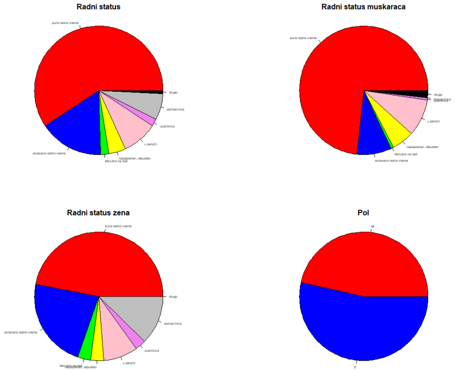


Кружни дијаграм

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

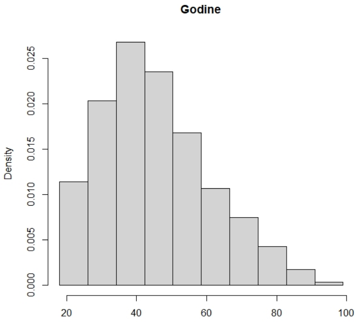


```
> par(mfrow=c(2,2))
> pie(table(baza1$radni.st), main="Radni status", col=boje,
radius=1, cex=0.5)
> pie(table(baza1$radni.st[baza1$pol=="M"]), main="Radni status
muskaraca", col=boje, radius=1, cex=0.5)
> pie(table(baza1$radni.st[baza1$pol=="F"]), main="Radni status
zena", col=boje, radius=1, cex=0.5)

> pie(table(baza1$pol), main="Pol", col=boje, radius=1, cex=0.5)
```



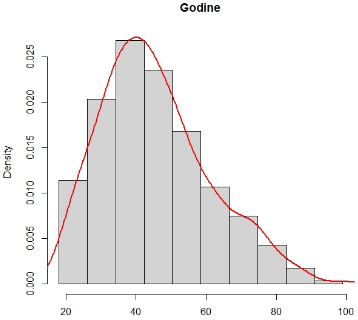
Одређивање граница хистограма по формули

```
> 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)
```



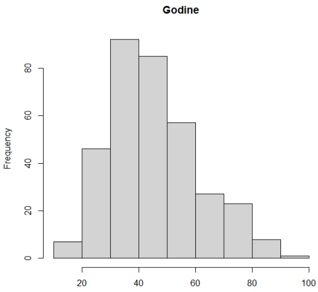
Додавање оцене густине на хистограм

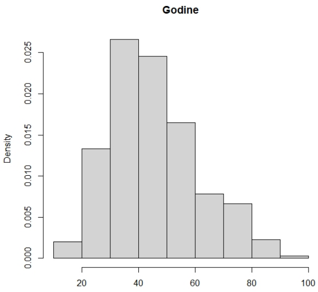
```
> broj.kategorija <-ceiling(log(length(baza1$godine),2))+1
> d<-(max(baza1$godine)-min(baza1$godine))/broj.kategorija
> granice<-min(baza1$godine)-1+(0:broj.kategorija)*(d+0.1)
> hist(baza1$godine, breaks=granice, prob=TRUE)
> lines(density(baza1$godine), col="red", lwd=2)
```



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)

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 $3/4$ узорка мање, а $1/4$ 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 наш модел, тј. одступа од правила уочених за остатак узорка. Понекад су аутлајери 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

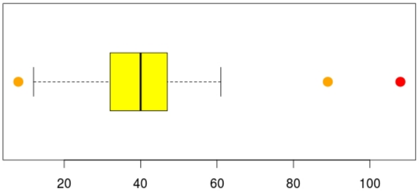
$$F_1 = q_1 - 3 \cdot \mathrm{IQR} \text{ и } F_3 = q_3 + 3 \cdot \mathrm{IQR}.$$

- 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

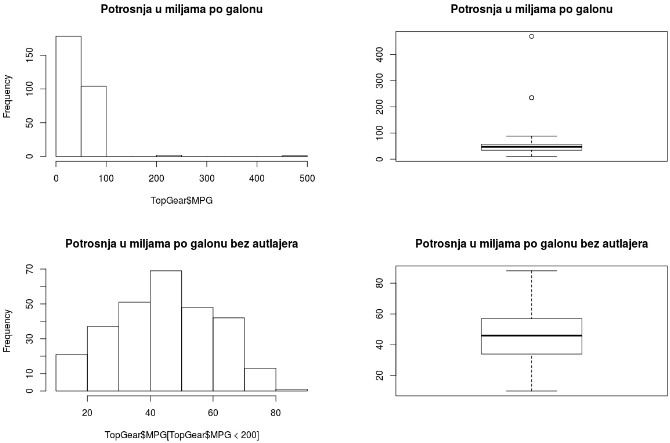
Дужина (у данима) болничког лечења пацијената с амнезијом

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)
```

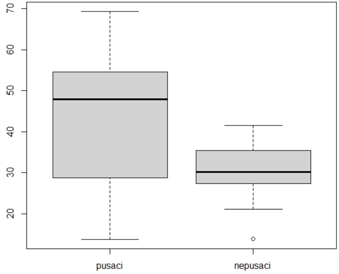
Утицај аутлајера на графички приказ



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	← a_1
[2,]	28.80	27.40	← q_1
[3,]	47.85	30.20	← q_2 (медијана)
[4,]	54.60	35.45	← q_3
[5,]	69.30	41.60	← 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.

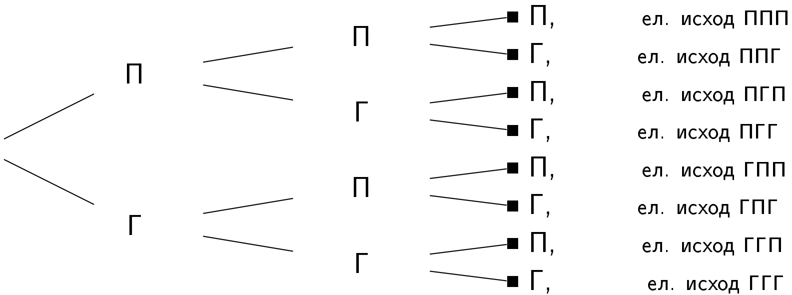
Баца се кутија шибица. Вероватноће добијања њених шест страна су: по $\frac{1}{20}$ за две странице најмање површине, по $\frac{1}{10}$ за две странице "средње" површине, и по $\frac{7}{20}$ за две странице највеће површине. Вероватноћа да шибица не падне на страницу највеће површине је $P(A) = 2 \cdot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

```
> 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

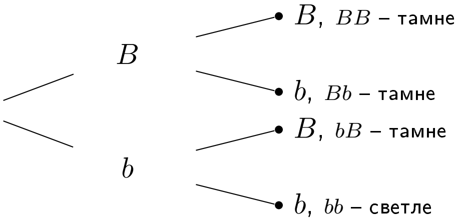
Сложеније експерименте можемо посматрати у етапама и приказати их на дијаграму гранања.



Слика: Бацање три новчића

Елементарна генетика – примена класичне веровантоће

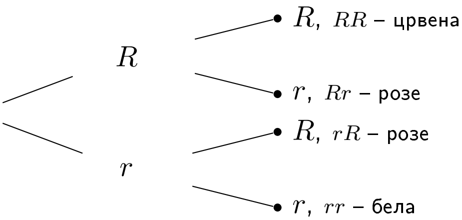
Обоје родитеља имају алеле и за тамне и светле очи, тј. хетерозиготни су према боји очију. Алел за тамне очи B је доминантан у односу на алел за светле очи b.
$P($дете има тамне очи $= \frac{3}{4})$.



Слика: Исходи наслеђивања боје очију код детета хетерозиготних родитеља

Елементарна генетика – примена класичне верoвантоће

Једна биљка има црвене, розе или беле цветове. Алели за црвену боју су R, а за белу r. Црвени цвет има RR, бели rr, а хетерозиготни су розе. Вероватноћа белог цвета након укрштања два хетерозиготна је $\frac{1}{4}$.



Слика: Исходи укрштања два хетерозиготна цвета

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:

- $\Omega_1 = \{$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)
- $\Omega_4 = \{$слика (краљ, дама, жандар), није слика$\}$
- $\Omega_5 = \{$слика, карта с бројем$\}$
- $\Omega_6 = \{$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

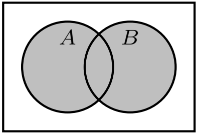
- Унија два догађаја $A \cup B$ садржи све елементарне исходе који се налазе у бар једном од догађаја A или B, тј. у A, у B, или у оба.
- Пресек два догађаја $A \cap B$, или краће AB садржи све елементарне исходе који се налазе и у A и у B.
- Комплемент $\bar{A}$ догађаја A садржи све елементарне исходе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

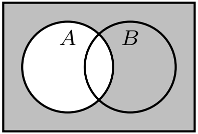
Операције над догађајима

Унија $A \cup B$



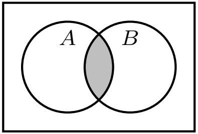
"у A или у B"

Комплемент $\bar{A}$

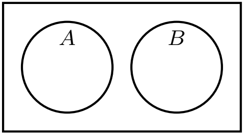


"не у A"

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

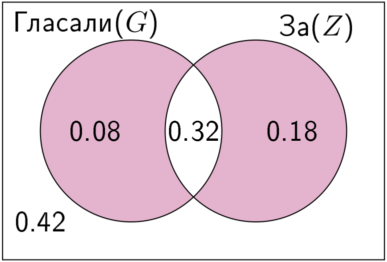
$$P(A_1 \cup A_2 \cup A_3 \cdots) = P(A_1) + P(A_2) + P(A_3) + \cdots$$

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?

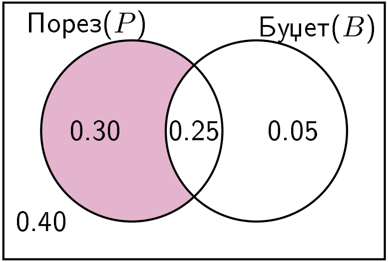
Дефиниција

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

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

Вероватноћа пресека зависних догађаја

Теорема

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

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

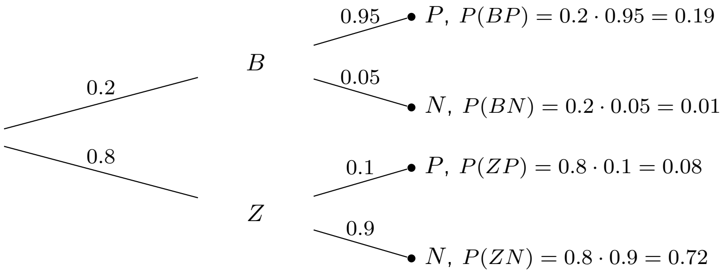
- У Америци око 46% људи има крвну групу O, а у регистрима је 4% оних који имају O грешком забележено као A. Колика је вероватноћа да случајно изабрани Американац стварно има O, али су му забележили A?

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, ... ,A_n međусобно 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, ... ,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 $\frac{5\cdot4\cdot3}{3!}=\frac{60}{6}=10.$

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 промењљивих $g(X)$ 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

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

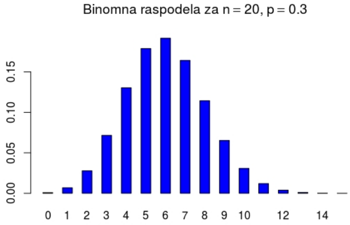
$$P\{X = 3\} = f(3) = \binom{5}{3}\left(\frac{1}{4}\right)^3\left(\frac{3}{4}\right)^2 = 10 \cdot \frac{1}{64}\frac{9}{16} = \frac{90}{1024} \approx 9\%$$

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

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

```
> 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)?

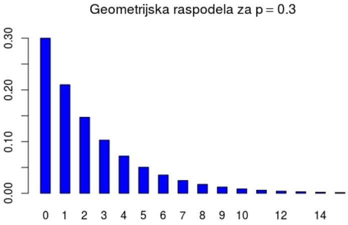
Број бацања пре добијања шестице X има геометријску расподелу с вероватноћом успеха $\frac{1}{6}$, па је

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

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

```
> 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?

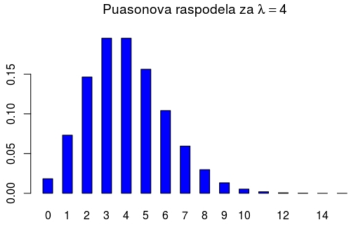
  $\lambda = 11 \cdot \frac{1}{4} = 2.75.$

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

```
> 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.

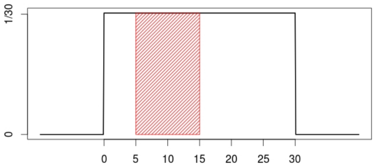
$$\begin{aligned} P\{a < X < b\} &= P\{a \leq X < b\} \\ &= P\{a < X \leq b\} \\ &= P\{a \leq X \leq b\} \end{aligned}$$

- 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**.

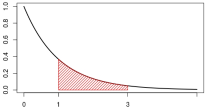


Слика: $P\{5 < X < 15\}$

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

Непрекидне вероватноће

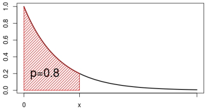
- X – животни век јединке (нпр. у годинама) – густина расподеле $f(x) = e^{-x}$, $x > 0$. Колика је вероватноћа да конкретна јединка умре између прве и треће године живота? После колико времена је вероватноћа да јединка није жива 80%?
- Ова расподела припада групи **експоненцијалних** расподела чије су густине $f(x) = \lambda e^{-\lambda x}$, $x > 0$, где је $\lambda > 0$ параметар стопе 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 промењљивих 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 "пљоснатији", а што је мања график је "сужениjи" 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

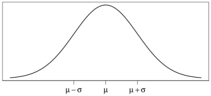
Дефиниција

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

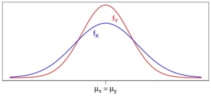
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad \text{за свако реално } x.$$

Особине нормалне расподеле

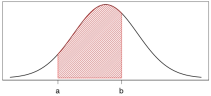
- График сваке нормалне расподеле је симетрична, звонаста крива чија је средина једнака μ
- Превоји криве су у тачкама $\mu - \sigma$ и $\mu + \sigma$
- Дисперзија σ^2 одређује облик криве
- Површина испод целе криве једнака је 1
- Вероватноћа да је нормална случајна променљива једнака неком броју је 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)

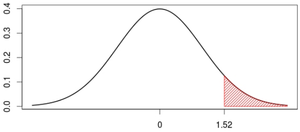
```
> dnorm(0.20)
0.3910427
```

- 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
```

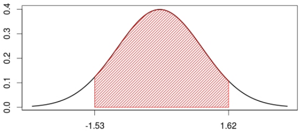
Стандардна нормална расподела

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



Слика: $P\{Z \geq 1.52\}$

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

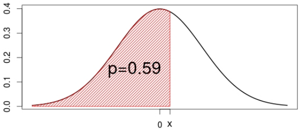


Слика: $P\{-1.53 < Z < 1.62\}$

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

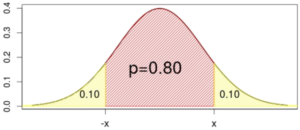
Стандардна нормална расподела

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



Слика: $P\{Z \leq x\} = 0.59$

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)

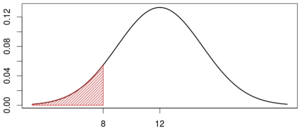
Нека случајна променљива Z *има нормалну расподелу са средњом вредношћу* μ *и дисперзијом* σ^2. *Тада случајна променљива*

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

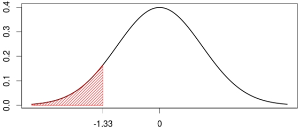
има стандардну нормалну расподелу. Z^* *се назива стандардизацијом случајне променљиве* Z.

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.

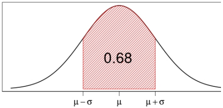
```
> pnorm(8,mean=12,sd=3)
0.09121122
```

Правила 1σ, 2σ и 3σ

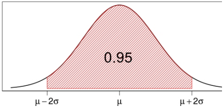
Теорема

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

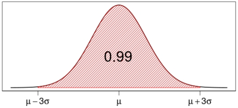
1) *Вероватноћа да Z одступи од свог математичког очекивања највише за једно стандардно одступање је приближно 0.68*
$$P\{\mu - \sigma < Z < \mu + \sigma\} \approx 0.68$$
2) *Вероватноћа да Z одступи од свог математичког очекивања највише за два стандардна одступања је приближно 0.95*
$$P\{\mu - 2\sigma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



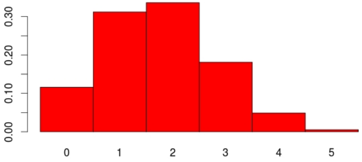
Слика: Правило 3σ

Апроксимација биномне расподеле нормалном

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

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

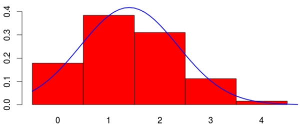
$$f(0) = 0.1160$$
$$f(1) = 0.3124$$
$$f(2) = 0.3364$$
$$f(3) = 0.1811$$
$$f(4) = 0.0488$$
$$f(5) = 0.0052$$



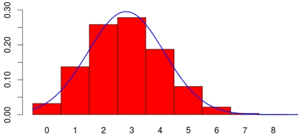
Слика: Графички приказ $f(x)$

Апроксимација биномне расподеле нормалном

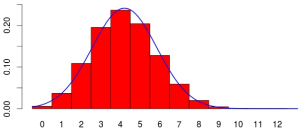
Графички приказ биномних вероватноћа за $p = 0.35$



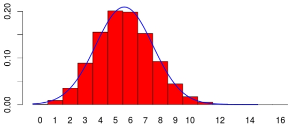
Слика: $n = 4,\ np = 1.4$



Слика: $n = 8,\ np = 2.8$



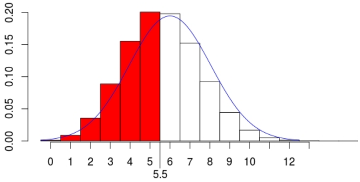
Слика: $n = 12,\ np = 4.2$



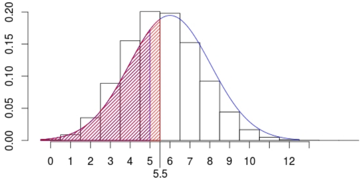
Слика: $n = 16,\ np = 5.6$

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

$n = 20,\ p = 0.3$

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

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

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

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

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

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
```

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 $X_1, \dots, X_n$ су случајне променљиве, а кад измеримо вредности добијамо њихове реализације

Дефиниција

Случајни узорак из расподеле за X чине случајне променљиве $X_1, \dots, X_n$, 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 промењиву 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

$$
\begin{aligned}
&P\{-1.645 < Z < 1.645\} = 0.90 && (\mathrm{qnorm}(0.95))\\
&P\Big\{-1.645 < \frac{\bar{X}-\mu}{\sigma}\sqrt{n} < 1.645\Big\} = 0.90\\
&P\Big\{\bar{X} - 1.645\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.645\frac{\sigma}{\sqrt{n}}\Big\} = 0.90,
\end{aligned}
$$

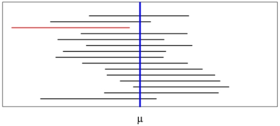
па је тражени интервал поверења

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

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 $\frac{\alpha}{2}$).*

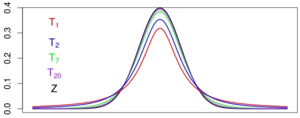
Студентова T расподела

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

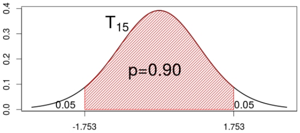
Студентова расподела - вредности x такве да је $P\{T_\nu < x\} = p$

ν \ p	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 pt$(x, \mathrm{df} = \nu)$.
```
> pt(1.25,df=10)
0.8801197
```
- Вредност x за коју је $P\{T_\nu < x\} = p$ рачунамо функцијом qt$(p, \mathrm{df} = \nu)$.
```
> 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 qt$(0.95, \mathrm{df} = 15)$, 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

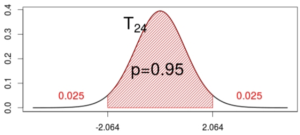
$$\left(\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right),$$

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 qt$(0.975, \mathrm{df} = 24) = 2.064$. 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.*

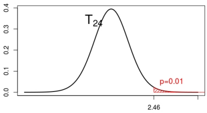
- p*-вредност једностраног левог теста је површина лево од вредности* t_0 *коју статистика* T_0 *узме у узорку.*
- p*-вредност једностраног десног теста је површина десно од вредности* t_0
- p*-вредност двостраног теста је двострука површина лево од вредности* t_0*, ако је* $t_0 < 0$ *или двострука површина 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: $P\{T_{24} > 2.46\} = 0.01$

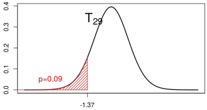
$n = 25$, $T_0 = \frac{\bar{X}-2}{S}\sqrt{25}$ има Студентову T_{24} расподелу. Из узорка рачунамо $\bar{x} = 2.842$, $s^2 = 1.708$, $s = 2.918$. Вредност тест статистике из узорка је $t_0 = \frac{2.842-2}{2.918}\sqrt{25} = 2.46$.

p-вредност теста добијамо као $1-\mathrm{pt}(2.46, \mathrm{df} = 24) = 0.01$, 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 $t_0 = \frac{44.5-45}{2}\sqrt{29} = -1.37$.

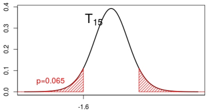
p-вредност теста добијамо као $\mathrm{pt}(-1.37, \mathrm{df} = 29) = 0.09$. Одлука је на нама, ако сматрамо да је грешка од 9% велика, закључићемо да не треба одбацити H_0 и не треба купити нови рачунар, а ако мислимо да је мала, онда ћемо купити нови рачунар.

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 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 (qt$(0.1, \mathrm{df} = 15)$). 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 $P\{X_\nu^2 < x\} = p$

ν \ p	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 $P\{X_5^2 < x\} = 0.05$ — из таблице видимо да је $x = 1.145$.
- Тражимо x такво да је $P\{X_{20}^2 > x\} = 0.025$ — онда је $P\{X_{20}^2 < x\} = 0.975$, 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.

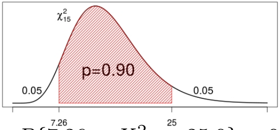
Интервал поверења за σ^2 код нормалне расподеле

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

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

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

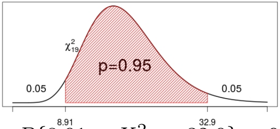
$$\Big(\sqrt{\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}}}, \sqrt{\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}}}\Big).$$

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*

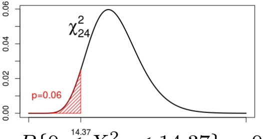
Тестирање хипотезе о σ^2 код нормалне расподеле

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 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

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

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

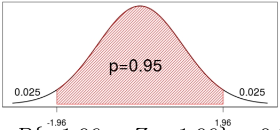
Интервал поверења за непознати проценат p је

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

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

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

Обим узорка за оцењивање p

Теорема

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

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

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 $p_1 > p_2$

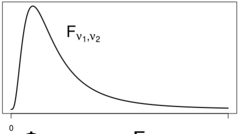
Директор спорткого сектора једног универзитета жели да добије статистичку подршку својој тврдњи да студенти спортисти у мањем проценту падају године него остали студенти. У добијеним узорцима 30 студената спортиста од њих 150 је пало годину, док се то догодило и у случају 43 студента од 200 чланова остатка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

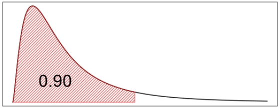
Фишерова F расподела



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 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 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 $X_1, \dots, X_{n_1}$ *и* $Y_1, \dots, Y_{n_2}$ *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 $\mathrm{pf}(0.521, \mathrm{df1} = 11, \mathrm{df2} = 9) = 0.15$, па немамо доказа за тврдњу да је дисперзија висине биљака сорте A већа.

Упоређивање средњих вредности две нормалне популације

- Имамо две популације чија обележја 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 промењиве X, и обима n_2 из нормалне расподеле $\mathcal{N}(\mu_2, \sigma_2^2)$ случајне промењиве 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 $X_1, \dots, X_{n_1}$ и $Y_1, \dots, Y_{n_2}$ 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 промењљива

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

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 промењљива 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 суособде 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
```