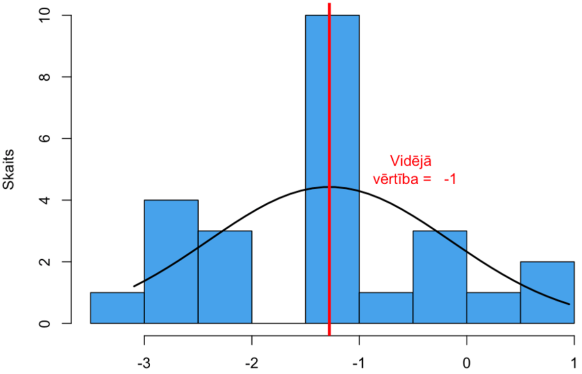
# Kā neapmaldīties datu jūrā?

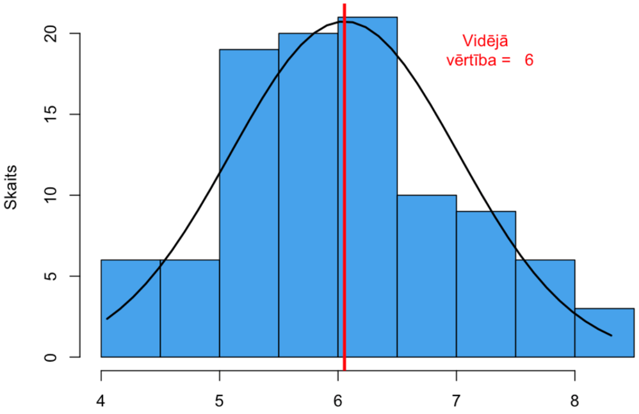
**Ievads statistiskajos testos**
**Mazā Matemātikas universitāte**

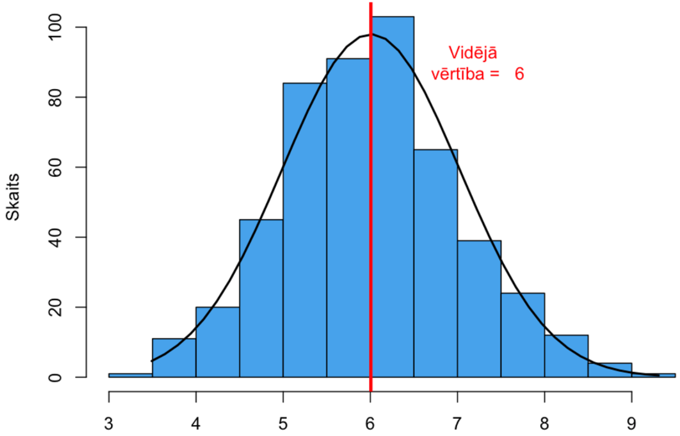
Rolands Aleksandrs Rudenko
4. kursa matemātiķis-statistiķis

# Normālais sadalījums

- Definīcija: Datus, kas histogrammā sakārtojas simetriski ap kopas vidējo vērtību, sauc par **normāli sadalītiem** datiem.

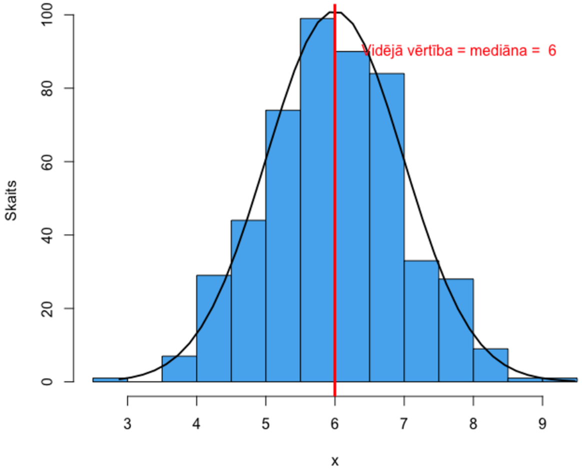


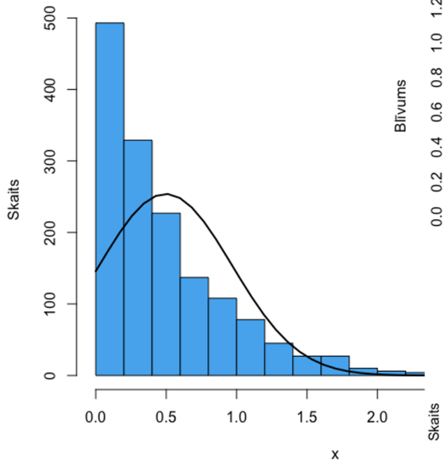



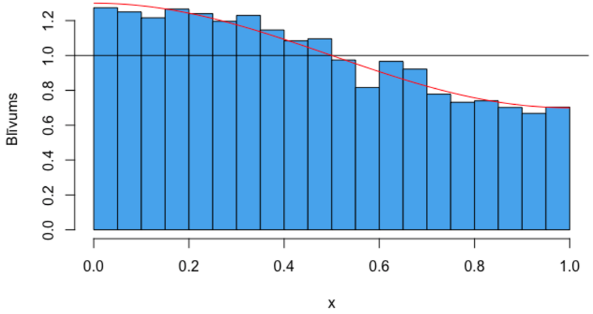
# Normālā sadalījuma noteikšanas metodes: Histogramma

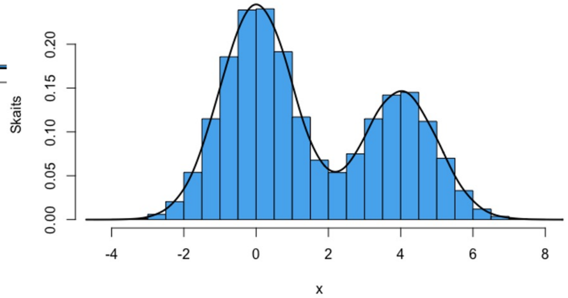
## Ir normālais sadalījums




## Nav normālais sadalījums



# Hipotēžu testi I

- *Definīcija*: **Statistiskā hipotēze** ir kāds apgalvojums par populācijas parametru.

- *Definīcija*: Hipotēžu pārbaudē piedalās divas komplimentāras (papildinošas) hipotēzes: nulles hipotēze $H_0$ un alternatīvā hipotēze $H_1$

- *Piemērs*: $\theta$ ir populācijas parametrs. *
  - divpusējā hipotēze: $H_0 : \theta = 0, H_1 : \theta \neq 0;$
  - vienpusējā hipotēze: $H_0 : \theta = 0, H_1 : \theta > 0 \text{ vai } H_0 : \theta = 0,\ H_1 : \theta < 0.$

- *Definīcija*: **Testa statistika** (kritērijs) ir tāda statistika, kas tiek atbilstoši izvēlēta tā, lai pārbaudītu konkrēto hipotēzi.

- *Definīcija*: **Kritiskais apgabals** - tās statistiskās vērtības pie kurām **noraida** nulles hipotēzi.

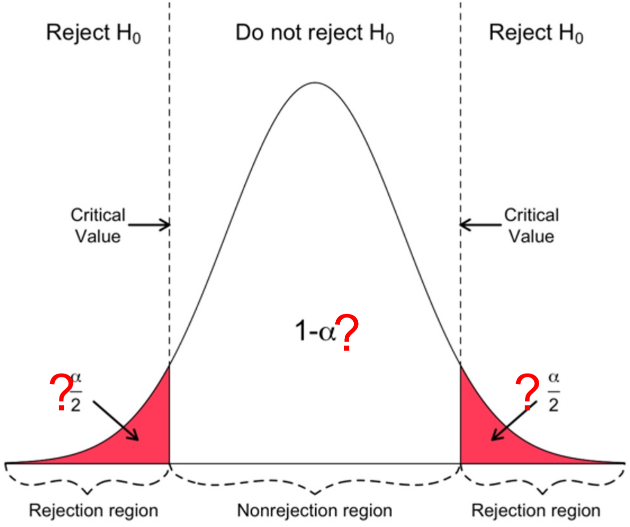
- *Definīcija*: **Pieņemšanas apgabals** - tās statistiskās vērtības pie kurām **pieņem** nulles hipotēzi.

* Vienkāršības pēc, šodien rēķināsim statistikas tikai **divpusējām hipotēzēm**

# Hipotēžu testu vizuālā reprezentācija vienai izlasei

Divpusējā hipotēze

$$H_0 : \theta = 0, H_1 : \theta \neq 0;$$

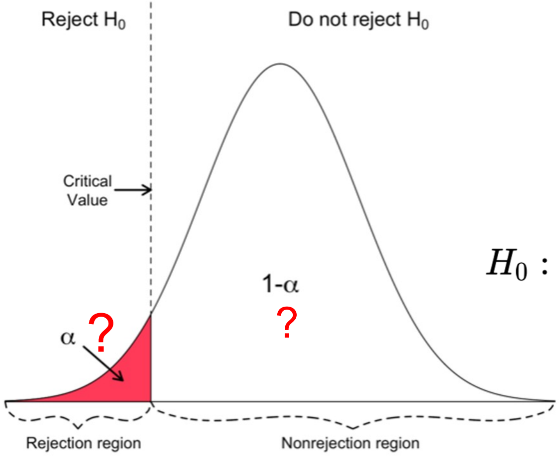



Bet kas ir $\alpha$?

Sauc arī par nozīmības līmeni un tā ir **varbūtība**, ka tiks pieļauta kļūda.

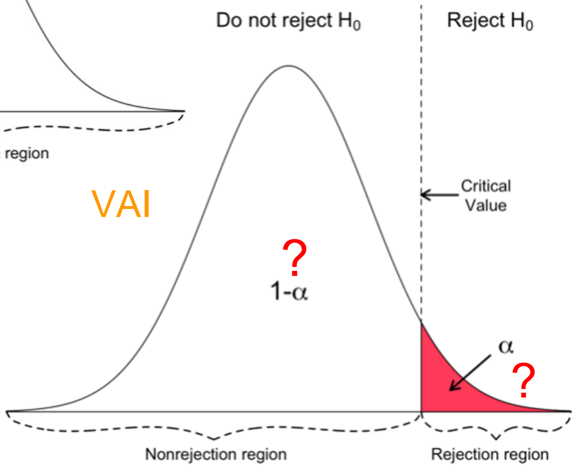
Vienpusējā hipotēze

$$H_0 : \theta = 0, H_1 : \theta > 0$$

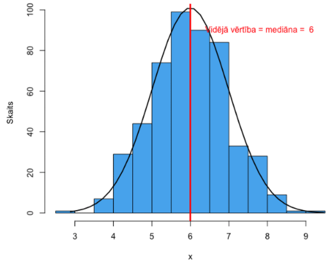


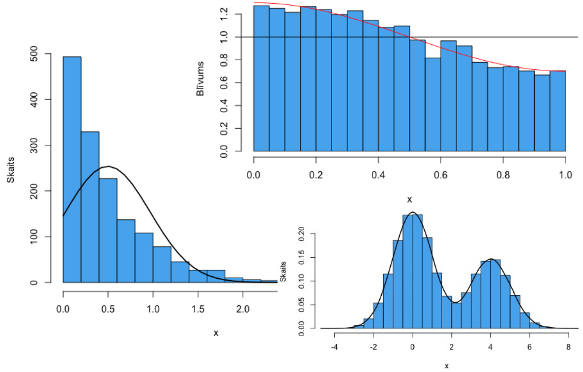

VAI

$$H_0 : \theta = 0,\ H_1 : \theta < 0.$$



# Hipotēžu testi II




## Ir divu veidu testi

### Parametriskie testi

- Izmanto, kad dati **IR** normāli sadalīti
- Uzrāda **labākus** rezultātus (biežāk uzrāda pareizus rezultātus)
- Parametrisko testu piemēri:
  - **T-tests**
  - One Way Anova
  - Z-tests
  - F-tests

### Neparametriskie testi

- Izmanto, kad dati **NAV** normāli sadalīti
- Uzrāda **sliktākus** rezultātus (biežāk pieļauj kļūdas)
- Neparametrisko testu piemēri:
  - **Vilkoksonu rangu tests**
  - Kolmogorova-Smirnova tests
  - Hī-kvadrāta tests
  - Kruskal-Wallis H tests

# Hipotēžu testu pielietojuma piemēri


Hipotēžu testus pielieto sekojošajās nozarēs:

- Medicīnā
- Farmācijā
- Bioloģijā
- Fizikā
- Optometrijā
- Socioloģijā/socioloģiskajos pētījumos
- Psiholoģijā

Hipotēžu testus izmanto, lai atbildētu uz sekojošiem jautājumiem:

- Kuras zāles ir efektīvākas, lai samazinātu asinsspiedienu? Zāles A vai zāles B?
- Vai tikai tāpēc, ka cilvēks ir republikānis, no tā var spriest, ka viņš balsos PRET abortiem? (Hī-kvadrāta tests)
- Vai šķidruma ieliešanas iekārta nav sabojājusies?
- Novērtētu covid19 vakcīnas efektivitāti

# T-tests

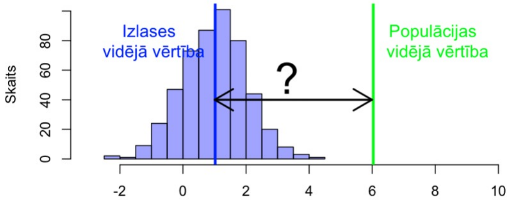
- Populācijas vērtība, kas tiek novērtēta: **vidējā vērtība**
- Pieņem, ka dati **IR** normāli sadalīti
- T-testa statistikas vērtība: **T**

!!! Mēs apskatīsim situācijas, kad populācijas vērtības mums **netiek** dotas, līdz ar to viss būs jānovērtē ar izlases datiem !!!

## Hipotēzes

Vienas izlases vidējās vērtības pārbaude

$$H_0 : \mu = \mu_1 \text{ pret } H_1 : \mu \neq \mu_1$$

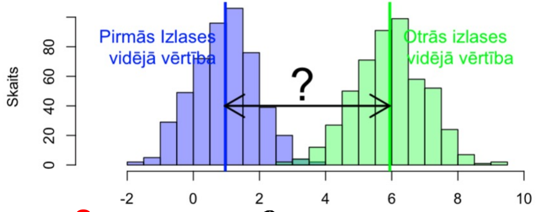



$$T = \sqrt{n}\frac{\bar{X} - \mu_1}{s}$$

$$df = n - 1$$

Divu izlašu vidējo vērtību vienādību pārbaude

$$H_0 : \mu_1 = \mu_2 \text{ pret } H_1 : \mu_1 \neq \mu_2$$




? Dati ir atkarīgi ?

$$T = \sqrt{n}\frac{\bar{D} - 0}{s_D}$$

$$df = \frac{n}{2} - 1$$

- !! Dati normalitāte tiek pārbaudīta uz datu pāru starpības

? Dati ir neatkarīgi* ?

$$T = \frac{\bar{X} - \bar{Y}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$df = n_1 + n_2 - 2$$

- !! Datu normalitāte tiek pārbaudīta katrai grupai

## Atkarīgi dati

Atšķirība tiek meklēta subjekta iekšienē


## Neatkarīgi dati

Atšķirība starp diviem dažādiem subjektiem


* T-testam ir vēl viens apakšgadījums, kad populācija dispersijas abām izlasēm nav vienādas, bet vienkāršības pēc šodien mēs apskatīsim situāciju, kad dispersijas ir vienādas

# Vilkoksona ranga tests

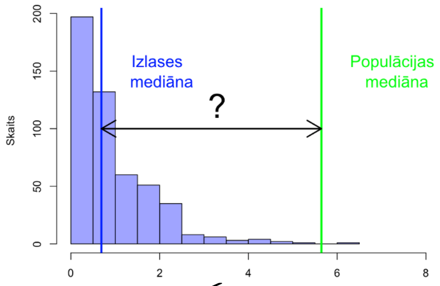
- Populācijas vērtība, kas tiek novērtēta: **mediāna**
- Pieņem, ka dati **NAV** normāli sadalīti
- Vilkosona ranga testa statistika: **W**

Bet kas ir rangs?

## Hipotēzes

Vienas izlases mediānas pārbaude

$$H_0 : \nu = \nu_1 \text{ pret } H_1 : \nu \neq \nu_1$$



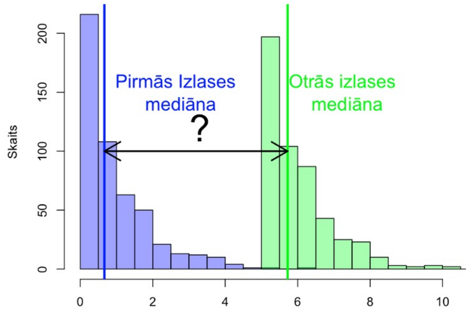

Izmanto Vilkoksona **zīmju** ranga testu

Statistikas aprēķināšanas algoritms:

1. Izpilda darbību $X_i - \nu_1$
2. 1. soļa darbībai piešķir rangu $R_i$
3. Aprēķina $R^+ \text{ un } R^-$
4. Nosaka statistikas vērtību $W = \min(R^+; R^-)$

Divu izlašu mediānu vienādību pārbaude

$$H_0 : \nu_1 = \nu_2 \text{ pret } H_1 : \nu_1 \neq \nu_2$$




Izlašu dati ir atkarīgi

Izmanto Vilkoksona **zīmju** ranga testu

Statistikas aprēķināšanas algoritms:

1. Izpilda darbību $X_i - Y_i$
2. 1. soļa *absolūtajai vērtībai* piešķir rangu $R_i$
3. Aprēķina $R^+ \text{ un } R^-$
4. Nosaka statistikas vērtību $W = \min(R^+; R^-)$

Izlašu dati ir neatkarīgi

Izmanto Manna Vitnija U testu (Vilkoksona **summas** ranga tests)

Statistikas aprēķināšanas algoritms:

1. Pirmo un otro izlasi apvieno vienā kolonnā
2. 1. soļa darbībai piešķir rangu $R_i$
3. Aprēķina $R_1, R_2, U_1 \text{un}\, U_2$
4. Nosaka statistikas vērtību $U = \min(U_1; U_2)$

Vienas izlases hipotēžu pārbaude

1) Ir aizdomas, ka iekārta nelej pieteikami daudz piena iepakojumā

5) Vai eksakto zinātņu studentiem IQ ir lielāks par populācijas IQ, kas ir 100

4) Vai cilvēki ir laimīgāki no rītiem vai vakarā?

1) Ir aizdomas, ka iekārta nelej pieteikami daudz piena

Divu izlašu hipotēžu pārbaude un dati ir atkarīgi

2) Vai zāles A palīdz nomest svaru?

) Vai eksakto zinātņu
tudentiem IQ ir
elāks par populācijas
Q, kas ir 100

Divu izlašu hipotēžu pārbaude un dati ir neatkarīgi

Minerālvielas Alpha vai Beta?

rālvielas
palīdz izaudzēt auglīgākus burkānus? Minerālvielas Alpha vai Beta?

# Kritiskā vērtība un tās noteikšana

- T-tests:
  - Ja $|T| > T_{krit}$, tad nulles hipotēze tiek noraidīta jeb populācijas vidējās vērtības ir atšķirīgas
- Vilkoksona **zīmju** ranga tests:
  - Ja $W < W_{krit}$, tad nulles hipotēze tiek noraidīta jeb populācijas mediānas ir atšķirīgas
- Manna Vitnija U tests (Vilkoksona **summas** ranga tests):
  - Ja $U \leq U_{krit}$, tad nulles hipotēze tiek noraidīta, jeb populācijas mediānas ir atšķirīgas

Veids kā noteikt kritisko vērtību:
Noz. līmenis = 0.025 un df = 7

| df \ Noz. līm. | 0.01 | 0.025 | 0.05 |
|---|---|---|---|
| 1 | 63.657 | 25.452 | 12.706 |
| 2 | 9.925 | 6.205 | 4.303 |
| 3 | 5.841 | 4.177 | 3.182 |
| 4 | 4.604 | 3.495 | 2.776 |
| 5 | 4.032 | 3.163 | 2.571 |
| 6 | 3.707 | 2.969 | 2.447 |
| 7 | 3.499 | 2.841 | 2.365 |
| 8 | 3.355 | 2.752 | 2.306 |
| 9 | 3.250 | 2.685 | 2.262 |
| … | | | |

# Izmantoto lielumu skaidrojums un formulas

$n -$ Izlases apjoms (datu apjoms izlasē)

$X_i -$ Pirmās izlases/grupas i-tais novērojums

$Y_i -$ Otrās izlases/grupas i-tais novērojums

$\bar{X} = \frac{1}{n}\sum_{n=1}^{i} X_i -$ Izlases vidējā vērtība

$D_i = X_i - Y_i$

$\bar{D} = \frac{1}{n}\sum_{i=1}^{n} D_i -$ Starpības vidējā vērtība

$\mu -$ Populācijas vidējā vērtība

$s^2 = \frac{1}{n}\sum_{n=1}^{i}(X_i - \bar{X})^2 -$ Izlases dispersija

$\sigma^2 -$ Populācijas dispersija

$s = \sqrt{s^2} -$ Izlases standartnovirze

$s_D = \sqrt{\frac{1}{n}\sum_{n=1}^{i}(D_i - \bar{D})^2}$ Starpības standartnovirze

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

$df -$ brīvības pakape, ko izmanto kritisko vērtību noteikšanai

$\nu -$ Populācijas mediāna

$m -$ Izlases mediāna

$R_i -$ Rangs i-tajam novērojumam

$R^+ = \sum_{i=1}^{n} R_i -$ Visu pozitīvo starpību ranku summa

$R^- = \sum_{i=1}^{n} R_i -$ Visu negatīvo starpību ranku summa

$R_1 = \sum_{i=1}^{n} R_i -$ Pirmās izlases rangu summa

$R_2 = \sum_{i=1}^{n} R_i -$ Otrās izlases rangu summa

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

# Praktiskā daļa

1) Sadalāmies grupā pa **2 vai 3**

## Pirmā daļa

2) Noteikt datu normalitāti un izsecināt, kuru testu izmantot: t-testu vai vilkoksona ranga testu

3) Veikt spriedumu par datu atkarību

## Otrā daļa

4) Aprēķināt statistikas vērtību

5) Noteikt kritisko vērtību

6) Sniegt secinājumu par populācijas vērtībām

# Materiālus, prezentāciju, formulas atradīsiet šeti

https://ej.uz/MMU_stat

VAI