Izlases, populācijas un centrālā robežteorēma

Māra Vēliņa, pētniece, LU FMF

Mazā Matemātikas Universitāte, 03.02.2018

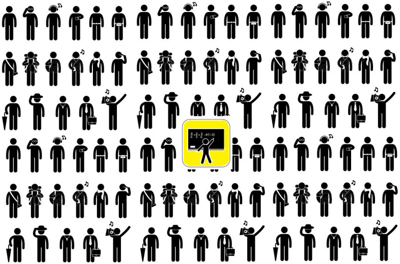
LU FMF Statistikas un datu analīzes laboratorija

Laboratorijas mērķi

- **Īstenot koppētījumus** dzīvības zinātņu, ekonomikas, socioloģijas u.c. jomās LU,
- **Konsultēt LU pētniekus un studentus**, kā arī **privātus uzņēmumus** ar statistisko analīzi saistītos jautājumos,
- **Iesaistīt LU FMF studentus** pētniecības darbā.
- **Līdzšinējā sadarbība:** LU Datorikas un medicīnas fakultātes, RTU, Latvijas Vides, ģeoloģijas un meteoroloģijas aģentūra, Eiropas Hitu radio, u.c.

Ievada piemērs: Aptauja par matemātiku

- Vēlamies uzzināt:
  Cik lielā mērā Latvijas vidusskolēniem patīk matemātika?



- Latvijā ir **36820 vidusskolēni**[1] -
  - Gandrīz neiespējami aptaujāt visus skolēnus!
- Nolemjam aptaujāt **200 skolēnus** -
  - Vai tam ir jēga?..

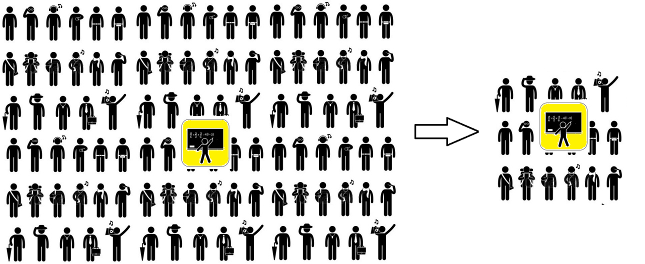
[1] 2017. gada 1. septembrī

Lekcijas saturs

- Izlase un populācija,
- Aprakstošā un secinošā statistika,
- Izlases veidošanas principi,
- Ilases sadalījumi,
- Normālais sadalījums,
- Centrālā robežteorēma un tās pielietojums.

Izlase un populācija

Izlase un populācija



Visi 36820 skolēni - POPULĀCIJA   200 skolēni - IZLASE

- **Populācija** (jeb *ģenerālkopa*) ir visu pētāmo elementu kopums (visi vidusskolēni),
- **Izlase** ir populācijas apakškopa, kas atlasīta praktiskai novērošanai (200 aptaujātie vidusskolēni).

Izlase

200 skolēni

- Var pētīt dažādas **pazīmes**, izdarot izlases elementu mērījumus:
  - **kvantitatīvas pazīmes**: skolēnu augums, vecums, atzīme matematikā,
  - **kvalitatīvas pazīmes**: dzimums, apmeklētā skola, utt.
- Izlases pazīmēm var aprēķināt dažādus skaitliskus lielumus:
  - *centrālās tendences* mēri: (aritmētiskais) vidējais, moda, mediāna,
  - *izkliedes* mēri: amplitūda, standartnovirze,
  - *pazīmju saistības analīze*: korelācijas koeficients.
- Šos lielumus sauc par **aprakstošās statistikas mēriem** jeb vienkārši **statistikām**.

Centrālā tendence: izlases (aritmētiskais) vidējais

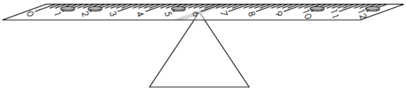
- **Izlases (aritmētiskais) vidējais** $\bar{x}$ ir novērojumu summa, dalīta ar to novērojumu skaitu izlasē, n:

$$\bar{x} = \frac{\sum x_i}{n}$$

- Piemērs. Pieci cilvēki atrisināja mīklu 5, 2, 12, 1 un 10 sekundēs. Tad, *izlases vidējais* risināšanas laiks ir

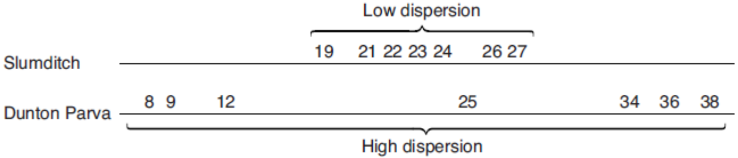
$$\bar{x} = \frac{5 + 2 + 12 + 1 + 10}{5} = 6 \text{ sekundes.}$$

- Izlases vidējais ir *tipisks novērojums* jeb *balansa punkts* starp izlases novērojumiem!



Dispersijas mēri: Izlases standartnovirze

- Izlases **dispersijas mēri** raksturo, cik tuvu novērojumu vērtības izkliedētas ap izlases centru.
- **Piemērs.** Viedokļa aptaujas rezultāti par lapsu medībām divos Anglijas ciematos:



  - Slumditch ciematā viedokļi ir ar *zemu dispersiju (izkliedi)*, taču Dunton Parva ciematā - ar *augstu dispersiju*!

- Bieži izmantots izkliedes mērs ir **izlases standartnovirze**:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

Aprakstošā un secinošā statistika

- **Aprakstošā statistika** sniedz izlases elementu vienkāršus raksturlielumus (skaitliskus vai vizuālus), pētot izlasi pēc vienas vai vairākām pazīmēm.
  - aprakstošā statistika raksturo izlasi!
- **Secinošā statistika** palīdz izdarīt secinājumus par visu pētāmo elementu kopumu jeb populāciju.
  - veic prognozes, veido dažādus modeļus!
  - galvenais pētnieku mērķis!
- Secinošās statistikas metodes ir **novērtēšana** un **hipotēžu pārbaude**.
  - Šodien aplūkosim **novērtēšanu**!

No izlases uz populācju

Izlase	$\rightarrow$	**Populācija**
Statistikas	SECINOŠĀ	Parametri
latīņu burti	STATISTIKA $\rightarrow$	grieķu burti
$\bar{x}$=izlases vidējais		μ=populācijas vidējais
s= izlases standartnovirze		σ =populācijas standartnovirze

Piemērs: Aptauja par matemātiku (turpinājums)

Pieņemsim, ka tika sastādīta *reprezentatīva izlase* no 200 vidusskolēniem, un tika uzdots jautājums:

- **Kā Tev patīk matemātikas priekšmets? Atbildi skalā no 0 līdz 10, kur**
  - **0** nozīmē *nemaz nepatīk*,
  - **10** nozīmē *mīļākais priekšmets*.

0 1 2 3 4 5 6 7 8 9 10

- Tā kā šis ir **kvantitatīvs** mainīgais,
  - aprēķinām *izlases vidējo*, pieņemism, ka iegūstam punktu skaitu
  - $\bar{x} = 6,12$
- BET: kāds ir *visu vidusskolēnu (populācijas)* videjais punktu skaits?
  - $\mu = ?$
  - Atbildi var sniegt *secinošā statistika*.

Atlases principi

Gadījumizlase

Izlase ir *populācijas apakškopa*, bet svarīgi, lai tā *labi reprezentētu populāciju*.

- **Vienkārša gadījumizlase** ir izlase, kur ikvienam populācijas elementam pastāv vienāda iespēja tikt iekļautam izlasē.
  - sastādīt visu 36820 Latvijas vidusskolēnu sarakstu (**atlases ietvars**),
  - likt datoram **nejauši izlozēt** 200 skaitļus no 36820,
  - izvēlēties **aptaujas metodi** - klātienes vai telefona intervija, aptauja internetā utml.,
  - aptaujāt izlozētos vidusskolēnus un **aprēķināt izlases statistikas**.

Sistemātiskās kļūdas

- Pārklajuma kļūda
  - sastādītais vidusskolēnu saraksts nav pilnīgs,
- Atlases kļūda
  - visiem elementiem nav vienāda varbūtība iekļūt izlasē,
  - piemēram, t.s. ērtības izlase - "nejauša" aptauja uz ielas,
- Nerespondences (neatbildētības) kļūda
  - atteikšanās atbildēt uz jautājumu
- Respondences kļūda
  - sociāli vēlamu atbilžu sniegšana

Mērķis: Sastādīt vienkāršu gadījumizlasi un pēc iespējas samazināt sistemātisko kļūdu iespējamību!

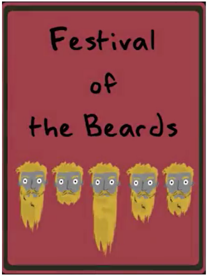
- Izveidot nejaušu gadījumizlasi ir teju neiespējami...

Nejaušas gadījumizlases alternatīvas

1. Ligzdveida izlase
   - Sadala populāciju *līdzīgu elementu apakškopās* (**ligzdas**), piemēram skolas, nami, preču kastes,
   - Nejauši atlasa noteiktu ligzdu skaitu un apseko visus tās elementus.
2. Stratificētā izlase
   - Sadala populāciju *atšķirīgu elementu apakškopās* (**stratas**), piemēram, vidusskolas klase, preču partija,
   - nejauši atlasa noteiktu elementu skaitu no katras stratas, kas ir proporcionāls stratas lielumam populācijā.

Izlases sadalījumi

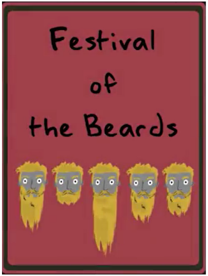
Piemērs: bārdaiņu festivāls[2]

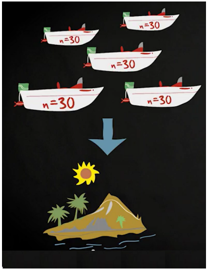


- Pieņemsim, ka Oslo pilsētā uz kādas salas tiek rīkots bārdaiņu festivāls

[2]W.W. Norton. Naked Statistics: Stripping the Dread from the Data. 2013

Piemērs: bārdaiņu festivāls





- Tiek pārdotas $\sim$ 5000 biļetes
- Sadala dalībniekus laivās pa 30 dalībniekiem
- Pieņemsim, ka mēs *zinām*, ka vidējais festivāla dalībnieku bārdas garums $\mu = 10.3$mm.

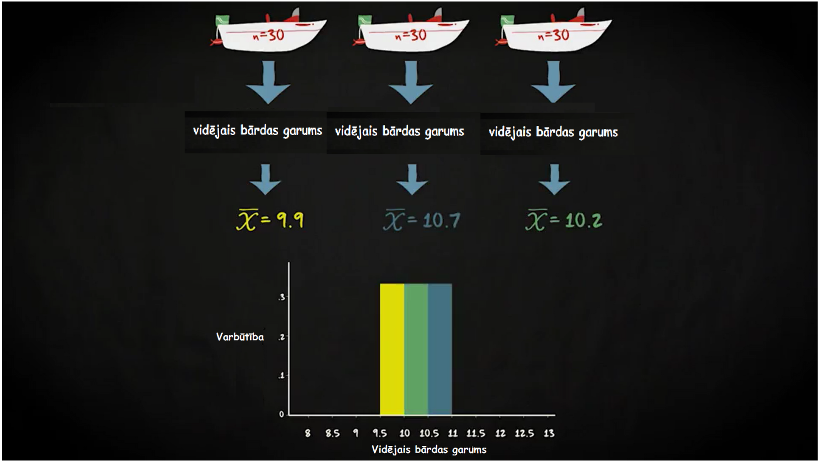
Piemērs: bārdaiņu festivāls

Piemērs: Vidējais bārdas garums laivā

Piemērs: Vidējais bārdas garums laivā

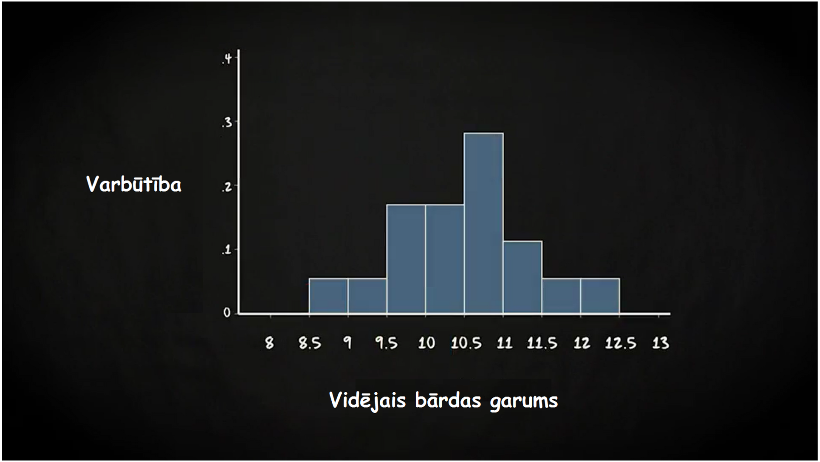
Piemērs: Vidējais bārdas garums laivā

- Aplūkosim vidējo vērtību sadalījumu **3 laivām**:



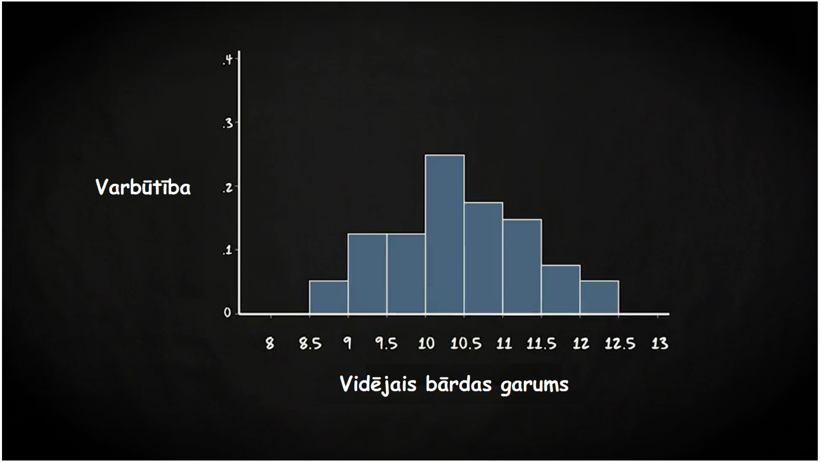
Piemērs: Vidējais bārdas garums laivā

- Aplūkosim vidējo vērtību sadalījumu **15 laivām**:



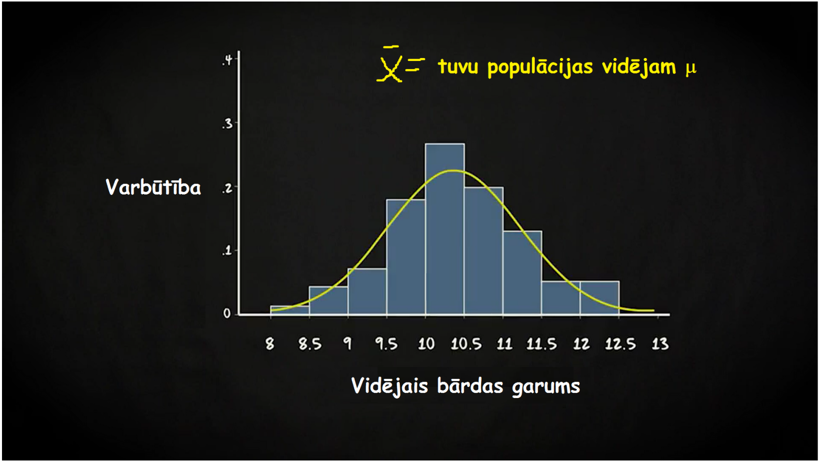
Piemērs: Vidējais bārdas garums laivā

- Aplūkosim vidējo vērtību sadalījumu **40 laivām**:



Piemērs: Vidējais bārdas garums laivā

- Aplūkosim vidējo vērtību sadalījumu **100 laivām**:



- Vidējo vērtību **sadalījuma forma** tiecas uz īpašu **zvanveida** jeb **Gausa funkciju**!

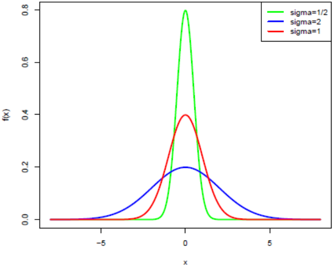
Normālais jeb Gausa sadalījums

Normālais sadalījums

Daudzi procesi dabā pakļaujas normālajam (**Gausa**) sadalījumam. Normālā sadalījuma *blīvuma funkcija*:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

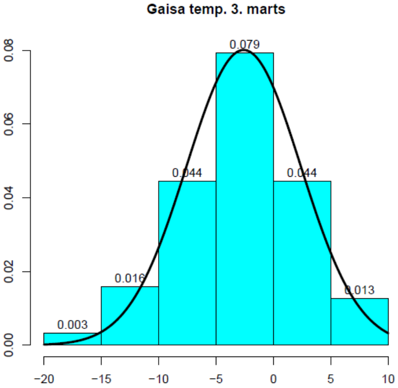
kur μ – vidējā vērtība, σ^2 – dispersija (σ – standartnovirze).



Gausa līknes laukums ir 1!!!

Piemēram: μ – vidējā temperatūra, σ^2 – izkliede ap vidējo temperatūru.

Normālais sadalījums

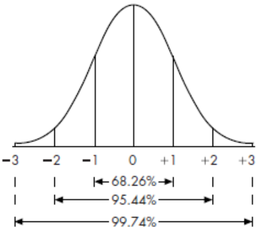


Vidējā temperatūra: $\boxed{\mu = -2,6°\text{C}}$

Standartnovirze: $\boxed{\sigma \approx 4,98°\text{C}}$

Ja gadījuma lielums X ir normāli sadalīts, to pieraksta $\boxed{X \sim N(\mu, \sigma^2)}$

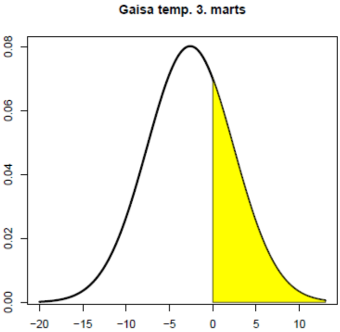
Trīs sigma likums



- Normālajam sadalījumam ir spēkā:
  - **68.2%** no novērojumiem atrodas intervālā +/- **vienu standartnovirzi** no vidējā,
  - **95.4%** no novērojumiem atrodas intervālā +/- **divas standartnovirzes** no vidējā,
  - **99.7%** no novērojumiem atrodas intervālā +/- **trīs standartnovirzes** no vidējā.
- **Sekas**. Ja dati ir *aptuveni normāli sadalīti*, gandrīz visas datu vērtības atrodas ± 3 standartnoviržu attālumā no vidējā!
- **Piezīme**. Šie procenti raksturo *laukumu zem Gausa līknes* jeb *varbūtību* novērojumam piederēt konkrētajam intervālam.

Laukums zem līknes: Piemērs

- Izmantojot blīvuma funkciju, laukumu zem līknes jeb varbūtību iespējams aprēķināt *jebkuram* brīvi izvēlētam intervālam!
- Kāda ir varbūtība, ka gaisa temperatūra 3. martā pārsniedz 0'C?
  - Ja $\mu = -2.6$ un $\sigma = 4.9$, tad
  - $P(X > 0) = 0.3$.

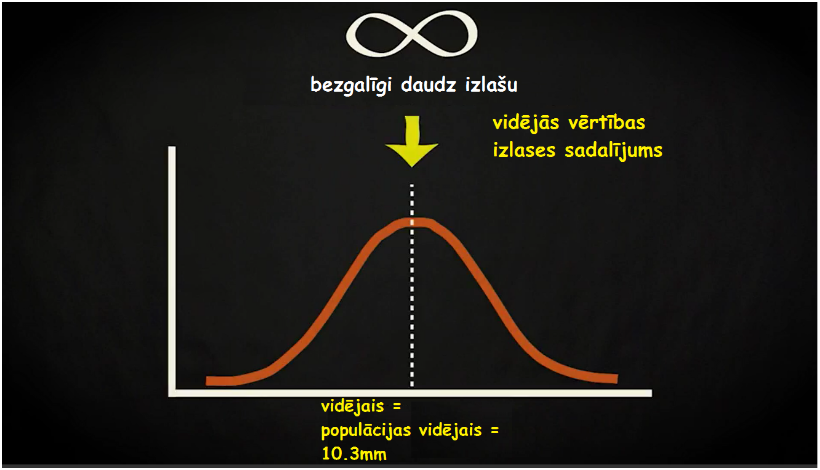


$\boxed{P(X > 0) = 0.3}$

Saliekot visu kopā... Centrālā robežteorēma

Vidējās vērtības izlases sadalījums

- Sadalījumu, kas raksturo visas vērtības, kādas var pieņemt izlases vidējais, sauc par **vidējās vērtības izlases sadalījumu** izlasēm ar apjomu n.



Centrālā robežteorēma

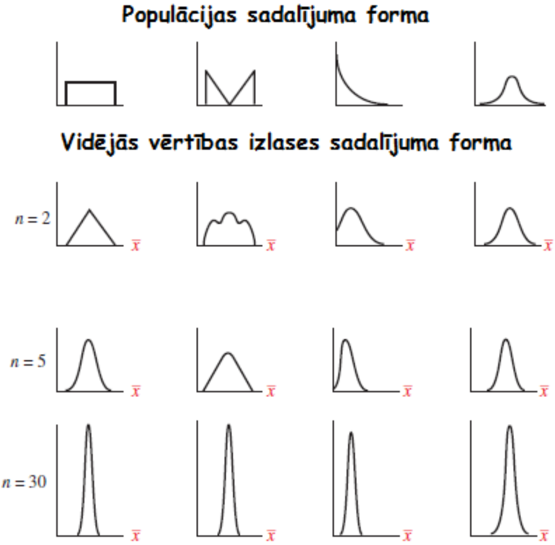
Centrālā robežteorēma. Pie nosacījuma, ka n ir pietiekoši liels, vidējās vērtības izlases sadalījums ir normālais, t.i.,

$$\bar{X} \sim N(\mu, \sigma^2/n),$$

kur μ-populācijas vidējā vērtība, σ - populācijas standartnovirze.

- 1. Piezīme. Izlases sadalījuma standartnovirze ir vienāda ar $\sigma/\sqrt{n}$. Šo lielumu sauc par **vidējās vērtības standartkļūdu** (se).
- 2. Piezīme. CRT ir ļoti spēcīgs rezultāts, jo tā ir **spēkā jebkurai populācijai** ar **galīgu standartnovirzi** ($\sigma < \infty$), neatkarīgi no populācijas sadalījuma veida!

CRT ir spēkā neatkarīgi no populācijas sadalījuma veida!



Centrālās robežteorēmas pielietojums: Ticamības intervāli

Vidējās vērtības variācija

- Centrālā robežteorēma ļauj kvantificēt *vidējās vērtības variāciju*, ko izsaka ar *standartkļūdu*:

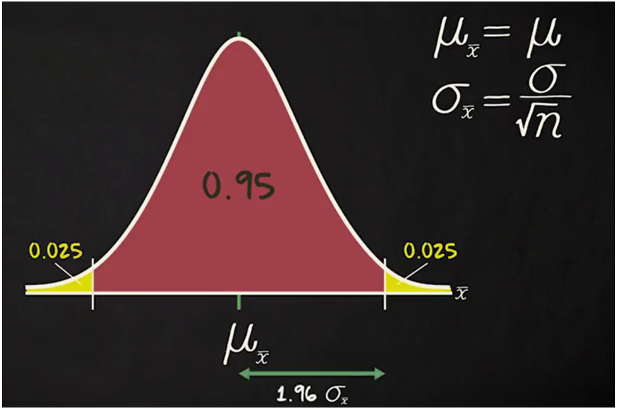
$$se = \frac{\sigma}{\sqrt{n}}$$

- *Vidējās vērtības standartkļūda* ir atkarīga no:
  - Populācijas standartnovirzes: $\sigma \uparrow se \uparrow$,
  - Izlases apjoma: $n \uparrow se \downarrow$.

Ticamības intervāli

- Ideja: uzdot *vērtību intervālu*, kurā ar kādu noteiktu *ticamibas līmeni* atrodas populācijas vidējais
- Vēl viens veids kā novērtēt populācijas parametru μ!
- Izvēlas ticamības līmeni tuvu 1, visbiežāk, 0.95.
- Izmantosim **centrālo robežteorēmu**!

Ticamības intervāli



- Saskaņā ar **normālā sadalījuma īpašībām**, 95.4% novērojumu atrodas $\pm$ **2 standartnovirzes** attēlumā no vidējā.
- Saskaņā ar **CRT** un normālā sadalījuma īpašībām*, **tieši 95% vērtību** atrodas ± 1.96 standartnoviržu $\sigma/\sqrt{n}$ attālumā no μ.

Ticamības intervāla konstruēšana, kad σ zināms

- Pieņemsim, ka populācijas standartnovirze σ **ir zināma**.
- Tad, **95% ticamības intervāls populācijas vidējai vērtībai** ir formā

$$\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}; \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right).$$

- Iespējams konstruēt arī α% **ticamības intervālus**:

$$\left(\bar{x} - z_\alpha\frac{\sigma}{\sqrt{n}}; \bar{x} + z_\alpha\frac{\sigma}{\sqrt{n}}\right),$$

$+z_\alpha$ ir **normālā sadalījuma kvantile**, piem.:

Tic. līmenis α	z_α
0,9	1,65
0,95	1,96
0,99	2,58

Piemērs. Aptauja par matemātiku

- Aptaujājot 200 skolēnu *reprezentatīvu izlasi*, tika iegūts vidējais punktu skaits $\bar{x} = 6,12$.
- Pieņemsim, ka pirms gada *jau bija veikts plašs pētījums* par skolēnu attieksmi pret matemātiku, kur tika noskaidrots, ka
  - $\sigma = 1,12$
- Aprēķināt **95% ticamības intervālu** populācijas vidējai vērtībai μ!
  - $n = 200$
  - $\sigma/\sqrt{n} = 1,12/\sqrt{200} = 0,079$
  - $z_{0.95} = 1,96$
  - Ticamības intervāls ir

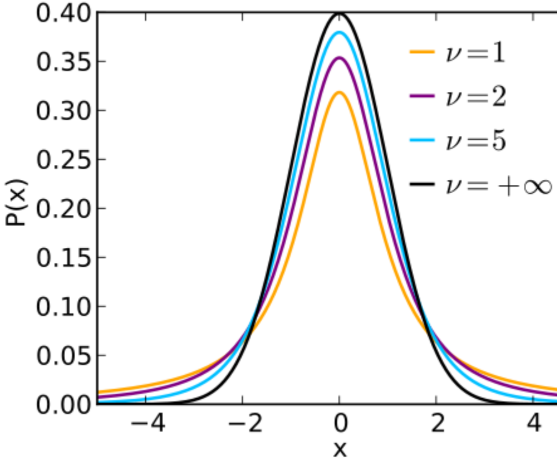
$$(6,12 - 1,96 \cdot 0,079;\ 6,12 + 1,96 \cdot 0,079) = (5,97;\ 6,27).$$

- **Interpretācija.** Ar 95% pārliecību varam teikt, ka visu Latvijas vidusskolēnu vidējā attieksme pret matemātiku 10 punktu skalā pieder intervālam $(5,97;\ 6,27)$.

Ticamības intervāla konstruēšana, kad σ nav zināma

- Visbiežāk, populācijas standartnovirze σ **nav zināma**!
- *Novērtē* σ ar **izlases standartnovirzi** s.
  - Rodas *papildus nenoteiktība*!
- Tādēļ, $\bar{x}$ sadalījumu raksturo nevis normālais sadalījums, bet **Stjūdenta t-sadalījums**!

Stjūdenta t-sadalījums



- Parametrs ν raksturo **brīvības pakāpes**, $\nu = n - 1$.
- Tātad, katram izlases apjomam ir *cits Stjūdenta t-sadalījums*!
- *Kas bija Stjūdents?
    - Viljams Gosets, kas strādāja Ginesa alusdarītavā, tāpēc

Ticamības intervāls balstīts uz Stjūdenta sadalījumu

- Ticamības intervāls ir formā:

$$\left(\bar{x} - t_{\alpha,n-1}\frac{s}{\sqrt{n}}; \bar{x} + t_{\alpha,n-1}\frac{s}{\sqrt{n}}\right),$$

kur $t_{\alpha,n-1}$ ir Stjūdenta sadalījuma α kvantile!

Piemērs. Holesterīna līmenis asinīs

- Pētnieki vēlējās noteikt vidējo holesterīna līmeni smēķējošiem vīriešiem ar paaugstinātu asinsspiedienu.
- Tika izveidota 12 vīriešu izlase un iegūts
  - izlases vidējais $\bar{x} = 217$mg/100ml,
  - izlases standartnovirze $s = 46$mg/100ml.
- Aprēķināt 95% ticamības intervālu **populācijas vidējam holesterīna līmenim** smēķējošiem vīriešiem ar paaugstinātu asinsspiedienu!
- Tabulā atrod $t_{0,95,11} = 2,20$,
- Ticamības intervāls:

$$(217 - 2,2 \cdot 46/\sqrt{12};\ 217 + 2,2 \cdot 46/\sqrt{12}) = (187,8;\ 246,2).$$