II

MATEMAATILINE STATISTIKA, TEOREETILISED JAOTUSED, PARAMEETRITE HINDAMINE

Antud peatükk püüab anda lühiülevaate matemaatilise statistika olemusest, teoreetilistest jaotustest, jaotuste peamistest parameetritest ja nende hindamismeetoditest.

2.1 JUHUSLIK SUURUS

Statistika jagatakse sageli kaheks – **kirjeldavaks** ehk **lihtsaks statistikaks** ning **matemaatiliseks statistikaks**. Neist esimese all mõistetakse andmete kokkuvõtlikku esitamist olulisemate tendentside tuvastamiseks, teine on aga mõeldud teaduslike järelduste tegemiseks empiiriliste (vaatlustest, katsetest, mõõtmistest pärinevate) andmete (**valimi**) põhjal.

Matemaatilise statistika meetodid baseeruvad tõenäosusteoorial ja on defineeritud juhusliku katse tulemuse, nn **juhusliku suuruse**, kaudu. Et juhuslik katse on iga katse, mille täpset tulemust pole võimalik ette näha, on juhuslikuks katseks ka näiteks kõikvõimalikud mõõtmised ning seega on järgnevast ehk lihtsam aru saada, kui mõista juhusliku suuruse all mingit tunnust. Tunnuste mõõtmistulemusi võib siis tõlgendada juhuslike suuruste realiseerunud väärtustena. Tähistatakse juhuslikke suurusi traditsiooniliselt suurte tähtedega X, Y, Z jmt.

Teoreetiline matemaatiline statistika tegeleb teoreetiliste juhuslike suuruste (üldkogumit iseloomustavate lõpmatul hulgal objektidel mõõdetud tunnuste) ja nende funktsioonide omaduste uurimisega tuginedes juhuslike suuruste realisatsioonidele (valimil mõõdetud tunnuste väärtustele).

2.2 OLULISEMAD JUHUSLIKKE SUURUSI ISELOOMUSTAVAD PARAMEETRID

2.2.1 Keskväärtus

Juhusliku suuruse **keskväärtuseks** (ooteväärtuseks e oodatavaks väärtuseks, inglise keeles *expected value*, *expectation*) nimetatakse tema lõpmatu hulga väärtuste keskmist. Juhusliku suuruse X keskväärtust tähistatakse $\mathrm{E}(X)$, sageli kasutatakse ka tähte μ.

Omadused

1. $\mathrm{E}(k) = k$, k on konstant.
2. $\mathrm{E}(kX) = k\,\mathrm{E}(X)$.
3. $\mathrm{E}(X_1 + X_2) = \mathrm{E}(X_1) + \mathrm{E}(X_2)$.
4. Kui X_1 ja X_2 on sõltumatud, siis $\mathrm{E}(X_1X_2) = \mathrm{E}(X_1)\,\mathrm{E}(X_2)$.
5. Kui $\mathbf{X} = (X_1 \;\; X_2 \;\ldots\; X_n)^{\mathrm{T}}$, siis $\mathrm{E}(\mathbf{X}) = \big(\mathrm{E}(X_1) \;\; \mathrm{E}(X_2) \;\ldots\; \mathrm{E}(X_n)\big)^{\mathrm{T}}$.

2.2.2 Dispersioon

Juhusliku suuruse X **dispersioon** $\mathrm{var}(X)$ (ka $\mathrm{V}(X)$, $\mathrm{D}(X)$, σ_X^2 , inglise keeles *variance*) on defineeritud seosega

$$\mathrm{var}(X) = \mathrm{E}[X - \mathrm{E}(X)]^2 = \mathrm{E}(X^2) - \mathrm{E}(X)\,\mathrm{E}(X) . \qquad (2.1)$$

Omadused

1. $\mathrm{var}(k) = 0$, k on konstant.
2. $\mathrm{var}(kX) = k^2\,\mathrm{var}(X)$.
3. Kui X_1 ja X_2 on sõltumatud, siis $\mathrm{var}(X_1 + X_2) = \mathrm{var}(X_1) + \mathrm{var}(X_2)$.

Ruutjuurt juhusliku suuruse dispersioonist nimetatakse **standardhälbeks** – $\sqrt{\mathrm{var}(X)} = \sigma_X$.

2.2.3 Kovariatsioon

Juhuslike suuruste X_1 ja X_2 vaheline **kovariatsioon** $\mathrm{cov}(X_1, X_2)$ (ka $\sigma_{X_1X_2}$, inglise keeles *covariance*) kirjeldab nende juhuslike suuruste vahelist lineaarset statistilist sõltuvust ja on defineeritud seosega

$$\mathrm{cov}(X_1, X_2) = \mathrm{E}(X_1X_2) - \mathrm{E}(X_1)\,\mathrm{E}(X_2)\,. \tag{2.2}$$

Omadused

1. $\mathrm{cov}(X_1, X_2) = \mathrm{cov}(X_2, X_1)$.
2. $\mathrm{cov}(X, X) = \mathrm{var}(X)$.
3. $\mathrm{cov}(aX_1, bX_2) = ab\,\mathrm{cov}(X_1, X_2)$, a, b on konstandid.
4. $\mathrm{var}(X_1 \pm X_2) = \mathrm{var}(X_1) + \mathrm{var}(X_2) \pm 2\,\mathrm{cov}(X_1, X_2)$.
5. Kui X_1 ja X_2 on sõltumatud, siis $\mathrm{cov}(X_1, X_2) = 0$ (sellest ja eelmisest omadusest järeldub dispersiooni 3. omadus).

Enam kui kahe juhusliku suuruse korral on nende dispersioone ja kovariatsioone mugav esitada maatriksina: juhuslike suuruste vektori $\mathbf{X} = (X_1\ X_2\ \ldots\ X_n)^{\mathrm{T}}$ **dispersioonimaatriksiks** (kovariatsioonimaatriksiks) nimetatakse maatriksit

$$\begin{aligned}\mathrm{var}(\mathbf{X}) &= \mathrm{E}(\mathbf{X} - \mathrm{E}\,\mathbf{X})(\mathbf{X} - \mathrm{E}\,\mathbf{X})^{\mathrm{T}} \\ &= \begin{pmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1, X_2) & \cdots & \mathrm{cov}(X_1, X_n) \\ \mathrm{cov}(X_1, X_2) & \mathrm{var}(X_2) & \ldots & \mathrm{cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(X_1, X_n) & \mathrm{cov}(X_2, X_n) & \ldots & \mathrm{var}(X_n) \end{pmatrix} = \begin{pmatrix} \sigma^2_{X_1} & \sigma_{X_1X_2} & \cdots & \sigma_{X_1X_n} \\ \sigma_{X_1X_2} & \sigma^2_{X_2} & \cdots & \sigma_{X_2X_n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{X_1X_n} & \sigma_{X_2X_n} & \cdots & \sigma^2_{X_n} \end{pmatrix},\end{aligned} \tag{2.3}$$

kus peadiagonaalil paiknevad juhuslike suuruste dispersioonid ja väljaspool peadiagonaali juhuslike suuruste vahelised kovariatsioonid.

Dispersioonimaatriksi omadused on analoogsed kovariatsiooni ja dispersiooni omadustega, väikesed erinevused tulenevad vaid sellest, et maatriksite korrutamisel ei kehti kõik üksikelementide korrutamise reeglid. Näiteks dispersioon konstantse maatriksiga **A** korrutatud juhuslike suuruste vektorist **X** avaldub kujul $\mathrm{var}(\mathbf{AX}) = \mathbf{A}\,\mathrm{var}(\mathbf{X})\mathbf{A}^{\mathrm{T}}$.

2.2.4 Korrelatsioon

Jagades juhuslike suuruste vahelise kovariatsiooni läbi nende standardhälvete korrutisega saame normeeritud kovariatsiooni, mida nimetatakse **korrelatsioonikordajaks**:

$$\mathrm{r}(X_1, X_2) = \frac{\mathrm{cov}(X_1, X_2)}{\sqrt{\mathrm{var}(X_1)\,\mathrm{var}(X_2)}}\,. \tag{2.4}$$

Omadused

1. $\mathrm{r}(X_1, X_2) = \mathrm{r}(X_2, X_1)$.
2. $\mathrm{r}(X, X) = 1$.
3. Kui X_1 ja X_2 on sõltumatud, siis $\mathrm{r}(X_1, X_2) = 0$
4. $\left|\mathrm{r}(X_1, X_2)\right| \le 1$.

Sarnaselt kovariatsioonimaatriksiga defineeritakse juhuslike suuruste vektori $\mathbf{X} = \left(X_1\ X_2\ \ldots\ X_n\right)^{\mathrm{T}}$ **korrelatsioonimaatriks** $\mathrm{r}(\mathbf{X})$ elementidega r_{ij}, $i, j = 1,\ldots,n$, kus

$$r_{ij} = r(X_i, X_j) = \mathrm{cov}(X_i, X_j) \big/ \sqrt{\mathrm{var}(X_i)\,\mathrm{var}(X_j)} \quad \text{ja} \quad r_{ii} = r(X_i, X_i) = 1\,.$$

Nii kovariatsiooni- kui ka korrelatsioonimaatriksid on sümmeetrilised maatriksid.

2.2.5 Lineaarne regressioon

Prognoosimaks juhusliku suuruse Y käitumist tingimusel, et sellega lineaarses sõltuvuses olev teine juhuslik suurus X omandab mingi kindla väärtuse, kasutatakse lineaarset regressiooniseost kujul

$$\mathrm{E}(Y) = a + bX \ .$$

Võrrandi vasakul poolel olev suurus $\mathrm{E}(Y)$ näitab, et tulemuseks on juhusliku suuruse Y oodatav (keskmine) väärtus juhusliku suuruse X mingi väärtuse korral.

Regressioonivõrrandi kordajad a ja b hinnatakse seostest

$$\hat{b} = \frac{\mathrm{cov}(X,Y)}{\mathrm{var}(X)} \quad \text{ja} \quad \hat{a} = \mathrm{E}\,Y - \hat{b}\,\mathrm{E}\,X \ . \tag{2.5}$$

Kasuks tuleb teadmine, et kordaja b väljendab juhusliku suuruse Y muutumise suurust juhusliku suuruse X muutumisel ühe ühiku võrra.

2.3 TEOREETILISED JAOTUSED

Kõikvõimalikke erinevaid juhuslikke suurusi (tunnuseid) on lõpmatu palju ja ei ole mõeldav tuletada iga uue tunnuse puhul uusi eeskirju, iseloomustamaks seda tunnust juhuslike mõõdetud väärtuste (valimi) alusel. Seetõttu jagatakse juhuslikud suurused gruppidesse vastavalt nende tekkemehhanismile e olemusele ning kirjeldatakse igat gruppi parameetritest sõltuva eeskirjaga, mille abil on võimalik leida vastavasse gruppi kuuluvate juhuslike suuruste (tunnuste) väärtuste esinemise tõenäosused. Selliseid eeskirju nimetatakse **teoreetilisteks jaotusteks**. Teoreetilised jaotused on aluseks statistiliste hüpoteeside kontrollimisel, sageli ka parameetrite väärtuste ja nende hinnangute usaldusväärsuse hindamisel.

Jaotusi on kaht tüüpi – diskreetsed ja pidevad jaotused. **Diskreetne jaotus** esitatakse tõenäosusfunktsiooniga

$$p(k) = \mathrm{P}(X=k)$$

või jaotustabeliga $\{p(k), k\}$, mis defineerib täpselt ära iga üksiku väärtuse esinemise tõenäosuse (k on juhusliku suuruse võimalik väärtus).

Pidev jaotus esitatakse tihedusfunktsiooniga

$$f(x) = \mathrm{d}F(x)/\mathrm{d}x,$$

mille abil on defineeritud juhusliku suuruse iga väärtuse mingisse fikseeritud vahemikku (a, b) sattumise tõenäosus:

$$P(a < X < b) = \int_a^b f(x)dx = F(b) - F(a),$$

$F(x) = \mathrm{P}(X \le x)$ on jaotusfunktsiooni väärtus kohal x.

Diskreetsed jaotused, mis tekivad millegi kokkulugemisel, subjektiivsel hindamisel või mingi nähtuse toimumise või mittetoimumise fikseerimisel, vajavad analüüsil siinses kursuses käsitletavate meetodite edasiarendusi – nn üldistatud lineaarseid mudeleid – mistõttu on järgnevalt tutvustatud vaid traditsioonilises üldiste lineaarsete mudelite teoorias ja geneetiliste parameetrite hindamisel kasutust leidvaid pidevaid jaotusi (ausalt öeldes neid ehk selle lühikursuse raames vaja ei lähegi, aga kui tekib soov või vajadus peale aretusväärtuste hindamise tutvuda ka dispersioonikomponentide ja päritavuskoefitsientide hindamisega, on vaja teadmisi normaaljaotusest, ning kui huvi pakub hinnangute täpsus ja hüpoteeside kontrollimine mingi parameetri nullist erinemise kohta, kulub ära ettekujutus χ^2-, F- ja t-jaotusest).

2.3.1 Normaaljaotus

Normaaljaotusega juhuslikku suurust X keskväärtusega $\mathrm{E}(X) = \mu$ ja dispersiooniga $\mathrm{var}(X) = \sigma^2$ tähistatakse $X \sim N(\mu, \sigma^2)$ ja tema tihedusfunktsioon esitub valemiga

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}. \qquad (2.6)$$

Keskväärtuse ja dispersiooni, kui normaaljaotuse ainukeste parameetrite läbi, on tihedusfunktsioon iga juhusliku suuruse X väärtuse tarvis määratud.

Tähtsaks omaduseks on, et normaaljaotusega juhuslike suuruste lineaarkombinatsioon on samuti normaaljaotusega (muutuvad vaid parameetrite väärtused). Sagedasimaks lineaarteisenduseks on standardiseerimine

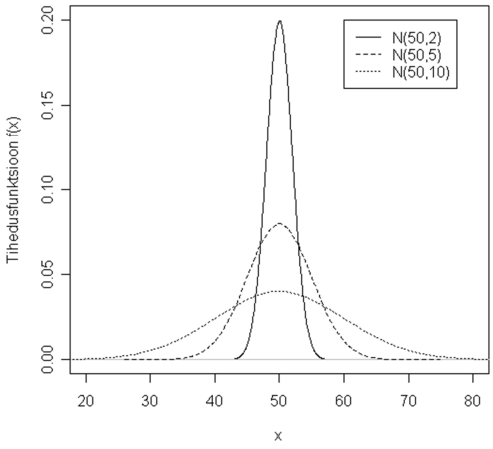
$$Z = \frac{X - \mu}{\sigma} \sim N(0,1),$$

kus $N(0,1)$ on standardne normaaljaotus, mille jaotusfunktsiooni $\Phi(x)$ väärtused on tabuleeritud (vt tabel 2.1). Seejuures kehtivad seosed $F(x) = \Phi(\frac{x-\mu}{\sigma})$ ja $\Phi(-x) = 1 - \Phi(x)$.

Tabel 2.1. Standardse normaaljaotuse enamkasutatavad jaotusfunktsiooni väärtused; $\Phi(x) = P(X \le x)$, kus $X \sim N(0,1)$.

$\Phi(x)$	0,005	0,025	0,05	0,5	0,95	0,975	0,995
x	-2,58	-1,96	-1,64	0	1,64	1,96	2,58

Näide. Normaaljaotusega $N(50,\sigma^2)$ on näiteks vere kogus indiviidi 50 ml vereproovis, kus σ^2 iseloomustab proovivõtmise täpsust.



Joonis 2.1. Näiteid normaaljaotuse $N(50,\sigma)$ tihedusfunktsiooni graafikuist erinevate parameetri σ väärtuste korral.

2.3.2 χ^2-jaotus

Sõltumatute standardse normaaljaotusega juhuslike suuruste $X_1,\ldots,X_n$ ruutude summa on χ^2-jaotusega vabadusastmete arvuga *n*. Sümbolite kaudu defineeritult: kui $X_1,\ldots,X_n$ on sõltumatud juhuslikud suurused, kus $X_i \sim N(0,1)$, $i = 1,...,n$, siis

$$\sum_{i=1}^{n} X_i^2 \sim \chi^2(n) \quad \text{ja} \quad \sum_{i=1}^{n} (X_i - \bar{X})^2 \sim \chi^2(n-1),$$

kus $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Kui $X \sim \chi^2(n)$, siis $\mathrm{E}(X) = n$ ja $\mathrm{var}(X) = 2n$.

2.3.3 *F*-jaotus

Kui juhuslikud suurused $U \sim \chi^2(k_1)$ ja $V \sim \chi^2(k_2)$ ning *U* ja *V* on sõltumatud, siis on juhuslik suurus *Z* *F*-jaotusega:

$$Z = \frac{U/k_1}{V/k_2} \sim F(k_1, k_2).$$

Jaotuse parameetrid k_1 ja k_2 on positiivsed täisarvud, mida nimetatakse *F*-jaotuse vabadusastmeteks (inglise keeles *degrees of freedom*).

2.3.4 *Student*'i *t*-jaotus

Kui juhuslik suurus $X \sim N(0,1)$ ja juhuslik suurus $Y \sim \chi^2(n)$, kusjuures *X* ja *Y* on sõltumatud, siis

$$Z = \frac{X}{\sqrt{Y/n}} \sim t(n),$$

ehk *Z* on *t*-jaotusega vabadusastmete arvuga *n*.

Matemaatilise statistika rakenduste tarvis on oluline tulemus, et kui $X_i \sim N(\mu,\sigma^2)$, $i = 1,...,n$, siis

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t(n-1),$$

kus $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ ja $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})^2}$.

2.4 PARAMEETRITE HINDAMINE

Kui me ka teame, mis teoreetilisele jaotusele meie poolt uuritav suurus oma olemuselt vastab, ei tea me siiski selle jaotuse parameetreid. Viimased tuleb hinnata tuginedes andmetele (valimile). Vastav arvutusvalem on lihtsamal juhul leitav intuitiivselt (nö mõistlikkuse printsiibist lähtudes – näiteks on loomulik võtta keskväärtuse hinnanguks valimi andmete aritmeetiline keskmine). Keerulisemal juhul ei pruugi aga intuitsioonist piisata või kui ka piisab, siis peab seda intuitsiooni kontrollima. Matemaatiline statistika pakub mitmeid võimalusi parameetrite hindamisvalemite teoreetiliseks konstrueerimiseks. Levinuimad meetodid on suurima tõepära meetod ja vähimruutude meetod.

2.4.1 Suurima tõepära meetod (*maximum likelihood method*, ML-meetod)

Suurima tõepära meetodit kasutatakse siis, kui teoreetiline jaotus on teada ja hinnatav parameeter kujutab enesest selle teoreetilise jaotuse tihedus- või tõenäosusfunktsiooni parameetrit (argumenti). Hinnanguks on siis loomulik valida see parameetri väärtus, mis realiseerunud juhul (st uuritavate andmete korral) kõige paremini sobib ehk teisisõnu on antud valimi jaoks tõepäraseim väärtus. Et eelduse kohaselt sõltus hinnatavast parameetrist ka üldkogumi jaotus, siis on ka hinnangule vastav jaotus tõepäraseim antud valimi jaoks.

Teoreetilise definitsioonina sõnastades: parameetri θ suurima tõepära hinnanguks nimetatakse väärtust $\hat{\theta}$, mille korral tõepärafunktsioon $L(\theta)$ saavutab maksimaalse väärtuse oma parameeterruumis,

$$L(\theta) = \begin{cases} f(x_1;\theta)\cdot f(x_2;\theta)\cdot\ldots\cdot f(x_n;\theta), \text{ pideval juhul,} \\ p(x_1;\theta)\cdot p(x_2;\theta)\cdot\ldots\cdot p(x_n;\theta), \text{ diskreetsel juhul.} \end{cases}$$

Tõepärafunktsioon kujutab enesest sama valemit, mis tihedusfunktsioongi. Erinevus seisneb selles, et tõepärafunktsioonis loetakse vastupidiselt teoreetilistele jaotustele fikseerituks andmete osa (meil on ju olemas mingid realiseerunud väärtused) ja juhuslikuks, kirjeldamist vajavaks, parameetrite osa. Juhul, kui tihedus- või tõenäosusfunktsiooni abil defineeritud teoreetiline jaotus vastab tegelikkusele, on parameetrite suurima tõepära hinnangud täpseimad.

Näide. Olgu meil vaatluse all dialleelne lookus alleelide *a* ja *A* sagedustega populatsioonis vastavalt *p* ja 1 – *p*. Olgu eelnevalt teada, et alleeli *a* sagedus on kas ½ või ¼, st $p \in \{½;¼\}$. Olgu meil kaks vaatlust: x_1='*a*' ja x_2='*a*'. Kumb on tõepärasem hinnang *p*-le, kas ½ või ¼?

Tõepärafunktsioon: $L(p) = P(X=x_1)\times P(X=x_2) = p^2$, millest L(½) = ¼, L(¼) = 1⁄16. Kuna L(½) > L(¼), siis $\hat{p} = ½$ on suurima tõepära hinnang *p*-le.

2.4.2 Vähimruutude meetod (*least square method*, LS-meetod)

Vähimruutude meetod ei eelda mingi tihedus- või tõenäosusfunktsiooni kasutamist, mistõttu on selle abil saadavad hinnangud sageli lihtsamal kujul võrreldes teiste hindamismeetoditega. Nagu suurima tõepära meetod, püüab vähimruutude meetodki valida hinnanguks seda parameetri väärtust, mis realiseerunud juhul (uuritavate andmete korral) kõige paremini sobib. Ainult sobivus on defineeritud pisut teisel kujul – parim hinnang on selline, mille korral ruuterinevus realiseerunud väärtuste ja parameetri hinnangule vastavate väärtuste vahel on minimaalne. Vähimruutude meetodil on tuletatud näiteks lineaarse regressioonivõrrandi parameetrite hinnangufunktsioonid (2.5).

2.4.3 Hinnangute omadused

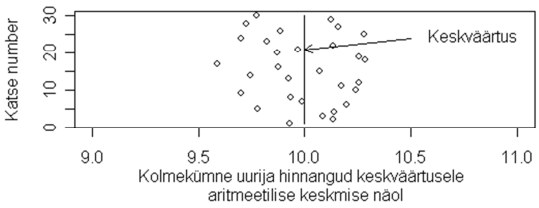
Parameetri θ hinnangut nimetatakse **nihketa hinnanguks** (*unbiased*), kui $E(\hat{\theta}) = \theta$; ehk hinnang on „keskmiselt õige", puudub süstemaatiline viga (vt ka joonis 2.2).

Parameetri θ hinnangut $\hat{\theta}$ nimetatakse **efektiivseks hinnanguks**, kui $\mathrm{var}(\hat{\theta})$ on vähim kõigi parameetri θ nihketa hinnangute dispersioonide hulgas; ehk – efektiivne hinnang on täpseim hinnang.

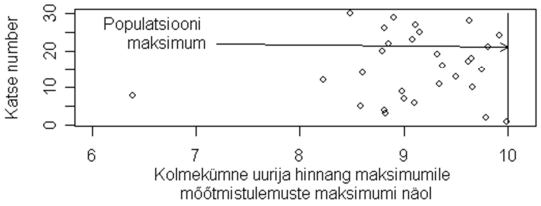
Näide. Valimi (juhusliku suuruse realiseerunud väärtuste) keskmine $\bar{x} = \frac{1}{n}\sum x_i$ on nihketa hinnang populatsiooni (juhusliku suuruse) keskväärtusele E(X):

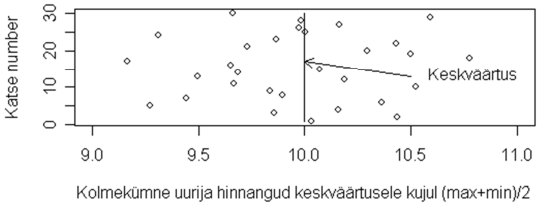
$$\mathrm{E}(\bar{x}) = \mathrm{E}(\tfrac{1}{n}\sum_{i=1}^{n} x_i) = \tfrac{1}{n}\sum_{i=1}^{n} \underbrace{\mathrm{E}(x_i)}_{\mathrm{E}(X)} = \tfrac{1}{n} n\,\mathrm{E}(X) = \mathrm{E}(X)\,.$$

Samuti on $\bar{x}$ populatsiooni keskmise efektiivne hinnang (vt ka joonis 2.2).



Joonis 2.2. Näide nihketa ja nihkega hinnangust.





Joonis 2.3. Näide täpsemast (efektiivsest) ja vähemtäpsemast hindamismeetodist

2.4.4 Hinnangu standardviga

Et andmete alusel leitud parameetri θ hinnang $\hat{\theta}$ on juhuslik suurus, siis eksisteerib tal ka dispersioon $\mathrm{var}(\hat{\theta})$. Viimane on aga jällegi tundmatu üldkogumi parameeter. Seega, et saada tegelikkuses aimu oma andmete alusel leitud parameetri hinnangu täpsusest, tuleb andmetest hinnata ka hinnangu dispersioon, millest reeglina parema mõistetavuse huvides võetakse veel ruutjuur (et saada varieeruvuse hinnangut samal skaalal parameetri endaga).

Hinnangu standardhälbe hinnangut nimetatakse hinnangu **standardveaks**:

$$se(\hat{\theta}) = \sqrt{\hat{\mathrm{var}}(\hat{\theta})}\,.$$

Näide. Keskväärtuse $\mathrm{E}(X) = \mu$ hinnangu $\hat{\mu} = \bar{x}$ dispersiooni hinnang on $\hat{\mathrm{var}}(\hat{\mu}) = s^2/n$ ja standardviga on

$$se(\hat{\mu}) = \frac{s}{\sqrt{n}}\,,$$

kus s^2 on valimi dispersioon.