Biomeetria

Populatsioon, valim, punkt- ja vahemikhinnangud.
Tõenäosus. Teoreetilised jaotused

Populatsioon *versus* valim

Üldkogum (**populatsioon**) on reaalselt olemasolev või ka abstraheeritud objektihulk, mille/kelle kohta soovitakse uurimistöö tulemusena sisulisi järeldusi teha (<u>peab olema enne uuringut fikseeritud nii ajas kui ka ruumis</u>).

Valim on teatava eeskirja järgi moodustatud hulk üldkogumisse kuuluvatest objektidest e mõõtmiseks valitud üldkogumi osa;
kui valim on piisavalt esinduslik, siis võime eeldada seal esinevate seaduspärade kehtimist enam-vähem sarnasel kujul ka üldkogumis.
<u>Lihtsaim reegel</u>: igal üldkogumi objektil on võrdne tõenäosus valimisse sattuda.

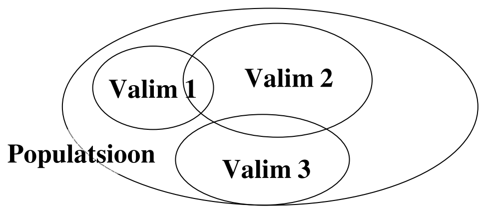
Näiteid:

	<u>Populatsioon</u>	<u>Valim</u>
1.	Eesti vetes kudevad lõhed	100 Eesti vetest kinni püütud ja ära mõõdetud lõhet
2.	Eesti maatõugu veised	Eesti maatõugu veised
3.	Eesti sigade pekipaksuse muutus ajavahemikul 1995-2005	34521 ajavahemikus 1995-2005 sooritatud sigade pekipaksuse mõõtmist

Populatsioon *versus* valim

Kõikne uuring: mõõdetakse kõiki üldkogumi objekte.
Valikuuring: mõõdetakse teatud väikest osa üldkogumist (valimit).

Statistika olemus: võtame teatud reeglite järgi osa üldkogumist (valimi), analüüsime seda ja teeme järeldusi kogu üldkogumi kohta!



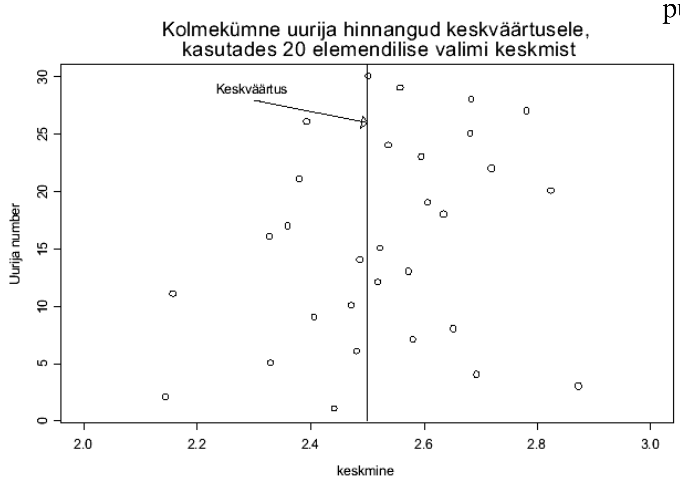
Populatsioon *versus* valim, punkthinnangud

Valimi alusel leitud arvkarakteristikud on vastavate populatsiooni parameetrite punkthinnangud.

Näiteks:

$\overline{x} = \hat{\mu}$, kus $\mu = \mathrm{E}(X)$;

$s = \hat{\sigma}$, kus $\sigma = \sqrt{\mathrm{D}(X)}$.



Tõenäosus

Statistika valetab alati!
Küsimus on üksnes selles, kui palju või vähe.
Viimase leidmiseks on vaja tõenäosuse mõistet ning eeskirju selle leidmiseks.

Tõenäosus [*probability*] on sündmuse toimumise mõõt skaalal 0-st 1-ni.
Võimatu sündmuse tõenäosus on 0 ja kindla sündmuse tõenäosus on 1.

Tõlgendused:

tõenäosus – osakaal e protsent
(kui teatud haigus esineb 5%-l uuritavast populatsioonist, siis tõenäosus, et selle populatsiooni juhuslikult valitud esindaja on haige, on 0,05 ehk 5%);

tõenäosus – usk sündmuse või nähtuse võimalikkusesse.

Tõenäosus

Kuidas leida praktikas sündmuse toimumise tõenäosust?

• Teha väga palju (teoreetiliselt lõpmatu arv) katseid/vaatlusi – nn statistiline tõenäosus – sündmuse toimumise suhteline sagedus annab tulemuseks ligikaudse tõenäosuse.

• Tundes uuritavat objekti hästi või mõistes täielikult, mis katse käigus toimub, on võimalik tõenäosust leida mõttetöö tulemusena (st, et teatavate sündmuste toimumise tõenäosuse kohta saab teha äärmiselt usutavaid eeldusi) – nn klassikaline tõenäosus (see, mida koolis õpitakse).

• Vahel on võimalik sündmuse tõenäosust arvutada, kui teame mõne teise sündmuse toimumise tõenäosust.

Statistiline tõenäosus

Suurte arvude seadus: katseseeria lõpmatul pikenemisel läheneb sündmuse suhteline sagedus tema tõenäosusele.

Suhtelise sageduse kaudu leitud nn statistiline tõenäosus on klassikalise tõenäosuse hinnanguks (st, et ei ole konstant – muutub katseseeria pikenedes).

Näiteks veeretate 75 korda täringut ja saate 52 korda 6 silma. Antud täringuga 6 silma saamise tõenäosus on siis hinnatav suhtest

$$\hat{P}(6\ \text{silma}) = 52/75 \approx 0{,}693.$$

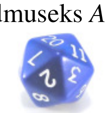
Klassikaline tõenäosus

Juhuslik katse – katse, mille tulemus pole ette teada.
Juhuslik sündmus – juhusliku katse tulemus.

$$\text{Tõenäosus} = \frac{\text{Sündmuse jaoks soodsate katsetulemuste arv}}{\text{Kõigi katsetulemuste arv}}$$

Näide. Katseks on 20-tahulise täringu veeretamine, sündmuseks A on 10-ga jaguva silmade arvu saamine.

$$P(A) = 2 / 20 = 0{,}1 .$$



Šansid

Sündmuse toimumise šansid [*odds*] näitavad, mitmel juhul sündmus toimub võrreldes sellega, mitmel juhul ta ei toimu.

Näide. Kui sündmus toimub tõenäosusega 0,2 (20%) e ühel juhul viiest, siis selle sündmuse toimumise šansid on üks nelja vastu e 1:4.

Tehted tõenäosustega

1. $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$
2. Tinglik tõenäosus $\mathrm{P}(A \mid B) = \mathrm{P}(A \cap B) / \mathrm{P}(B)$,
   millest $\mathrm{P}(A \cap B) = \mathrm{P}(B) \times \mathrm{P}(A \mid B) = \mathrm{P}(A) \times \mathrm{P}(B \mid A)$
3. Kui A ja B on sõltumatud, siis
   $\mathrm{P}(A \cap B) = 0$,
   $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$,
   $\mathrm{P}(A \mid B) = \mathrm{P}(A)$ ja $\mathrm{P}(B \mid A) = \mathrm{P}(B)$
4. Täistõenäosuse valem: $\mathrm{P}(A) = \sum_{i=1}^{k} \mathrm{P}(H_i) \times \mathrm{P}(A \mid H_i)$
5. Bayesi valem: $\mathrm{P}(H_i \mid A) = \dfrac{\mathrm{P}(A \mid H_i) \times \mathrm{P}(H_i)}{\sum_{j=1}^{k} \mathrm{P}(A \mid H_j) \times \mathrm{P}(H_j)}$

Teoreetilised jaotused

• Teoreetilised jaotused leitakse analüütiliselt või arvutisimulatsioonide abil, **baseeruvana uuritava tunnuse** (=juhusliku suuruse) või selle väärtuste funktsiooni (e statistiku) tekkemehhanismil e **olemusel**.

Näiteks on matemaatika mõistes oma olemuselt sarnased tunnused 'bakterite arv 1 ml piimas', 'emise pesakonna suurus', 'edukalt talvitunud mesitarude arv mesilas' jne, või siis 'lõhe kasvukiirus', 'õhu liikumise kiirus laudas', 'mulla happesus' jne.

• **Teoreetilised jaotused kirjeldatakse parameetritest sõltuvate eeskirjadega**, mille abil on võimalik leida vastava jaotusega tunnuste (statistikute) väärtuste esinemise tõenäosused.

• **Teoreetilised jaotused on aluseks teaduslike järelduste tegemisel** (statistiliste hüpoteeside kontrollimisel, sageli ka parameetrite väärtuste hindamisel ja nende hinnangute usaldusväärsuse leidmisel).

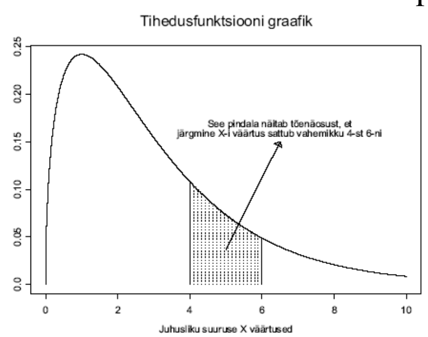
Seejuures on järeldused õiged üksnes siis, kui nad on tehtud andmetega sobivatele teoreetilistele jaotustele tuginedes (seda eriti väikeste, $n < 100$, valimite puhul)!

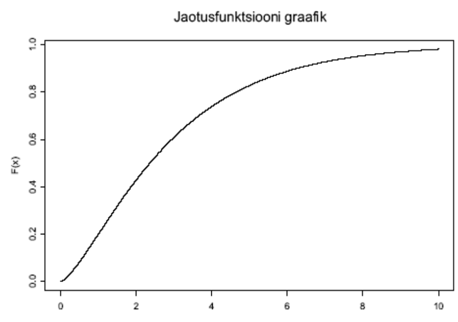
Teoreetilised jaotused

Diskreetne jaotus esitatakse **tõenäosusfunktsiooniga** $\boldsymbol{p(k) = \mathrm{P}(X{=}k)}$, kus k on jaotuse võimalik väärtus.

Pidev jaotus esitatakse **tihedusfunktsiooniga** $\boldsymbol{f(x) = \mathrm{d}F(x)/\mathrm{d}x}$, kus $\boldsymbol{F(x) = \mathrm{P}(X \le x)}$ on **jaotusfunktsiooni** väärtus kohal x,

$$\mathrm{P}(a < X < b) = \int_a^b f(x)dx = F(b) - F(a).$$





Diskreetsed jaotused

Bernoulli jaotusega on kõik binaarsed tunnused, $X \sim Be(p)$, kus p on Bernoulli jaotuse parameeter (tõenäosus, et uuritav suurus omandab väärtuse 1).

Seejuures $\mathrm{E}(X) = p$ ja $\mathrm{D}(X) = p \times (1 - p)$.

Binoomjaotus

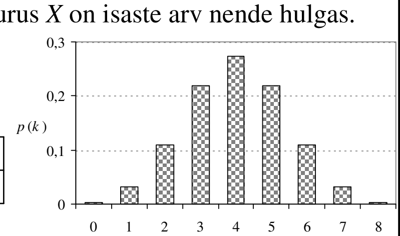
Sündmuse toimumiste arv n-katselises katseseerias, kus igal üksikul katsel on sündmuse toimumise tõenäosus p: $X \sim B(n;p)$.

Tõenäosusfunktsioon: $p(k) = C_n^k p^k (1-p)^{n-k}, \quad C_n^k = \dfrac{n!}{k!(n-k)!}$

$\mathrm{E}(X) = np$ ja $\mathrm{D}(X) = np(1 - p)$

Näited. Koeral sündis 8 kutsikat. Huvi pakkuv suurus X on isaste arv nende hulgas. Lihtsuse mõttes on eeldatud, et isase ja emase kutsika sündimise tõenäosus on võrdne ($p = 0{,}5$).

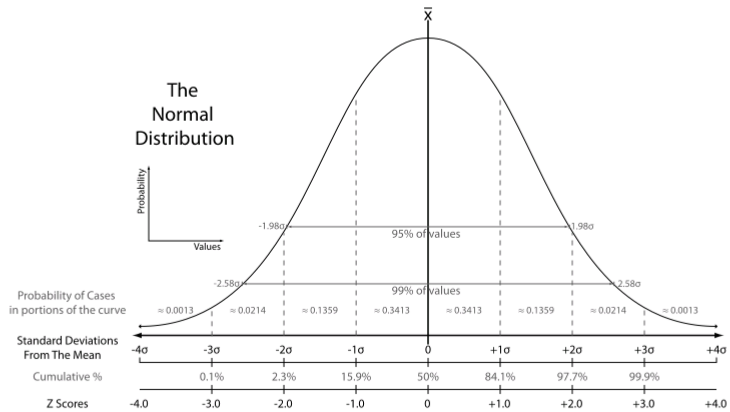
k	0	1	2	3	4	5	6	7	8
$p(k)$	0,004	0,031	0,109	0,219	0,273	0,219	0,109	0,031	0,004



Normaaljaotus

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

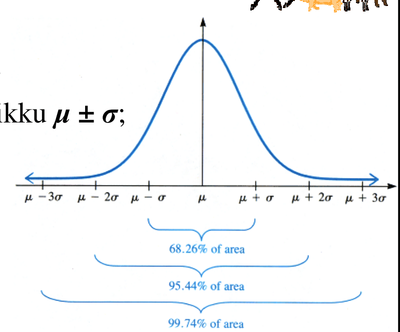
kus μ on uuritava tunnuse keskväärtus ja σ^2 dispersioon, $X \sim N(\mu,\sigma^2)$.



Normaaljaotus

Kui uuritav tunnus on normaaljaotusega, siis

- ligikaudu **68,3%** väärtustest jäävad vahemikku $\boldsymbol{\mu \pm \sigma}$;
- **95,4%** vahemikku $\boldsymbol{\mu \pm 2\sigma}$ ja
- **99,7%** vahemikku $\boldsymbol{\mu \pm 3\sigma}$.



Normaaljaotuse tähtsus statistikas:

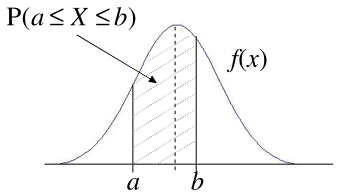
- Paljud mõõdetud tunnused on ligikaudu normaaljaotusega (näiteks juhul, kui uuritavat tunnust mõjutab hulk üksikult võttes väikese mõjuga tegureid).
- Paljud matemaatilise statistika meetodid eeldavad tunnuse jaotumist vastavalt normaaljaotuse seaduspäradele.
- Suurte valimite korral on paljud normaaljaotusega tunnuste tarvis loodud meetoditest rakendatavad sõltumata jaotusest – näiteks läheneb mistahes jaotusega tunnuse keskmise (ja ka summa) jaotus valimi mahu kasvades normaaljaotusele.

Normaaljaotus

Normaaljaotusega juhuslike suuruste lineaarkombinatsioon on samuti normaaljaotusega (muutuvad vaid parameetrite väärtused).
Sagedasemaks lineaarteisenduseks on standardiseerimine

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

kus $N(0,1)$ on standardne normaaljaotus, mille jaotusfunktsiooni $\Phi(z) = \mathrm{P}(Z \le z)$ väärtused on tabuleeritud. Seejuures kehtivad seosed



$$F(x) = \Phi(\tfrac{x-\mu}{\sigma}) \text{ ja } \Phi(-z) = 1-\Phi(z)$$

$$\mathrm{P}(a \le X \le b) = \mathrm{P}\left(\frac{a-\mu}{\sigma} \le \frac{X-\mu}{\sigma} \le \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Tabel. Standardse normaaljaotuse enamkasutatavad jaotusfunktsiooni väärtused

$\Phi(z) = \alpha$	**0,005**	**0,025**	**0,05**	**0,5**	**0,95**	**0,975**	**0,995**
z_α (q_α)	**-2,58**	**-1,96**	**-1,64**	**0**	**1,64**	**1,96**	**2,58**

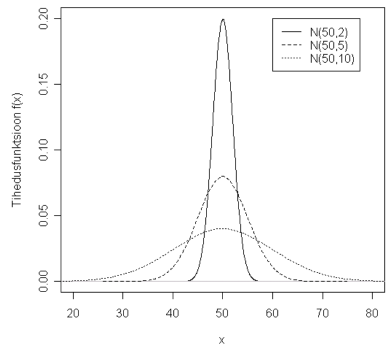
Normaaljaotus

Normaaljaotuse jaotusfunktsioon $\Phi(x)$

x	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990

Normaaljaotus

Näide. Vere kogus indiviidi 50 ml vereproovis $X \sim N(50; \sigma^2)$, kus σ iseloomustab proovivõtmise täpsust.



Kui suur on tõenäosus, et proovi maht erineb 50 ml-st enam kui 5 ml võrra?

$$\mathrm{P}(X<45 \cup X>55)=1-\mathrm{P}(45 \le X \le 55)=?$$

$\sigma=1$:

$$\mathrm{P}(45 \le X \le 55)=\Phi\left(\frac{55-50}{1}\right)-\Phi\left(\frac{45-50}{1}\right)$$

$$=\Phi(5)-\Phi(-5)=\Phi(5)-[1-\Phi(5)]=2\Phi(5)-1=2*0,9999997-1=0,9999994$$

$$\mathrm{P}(X<45 \cup X>55)=1-0,9999994=0,00000057=5,76\mathrm{E}-07$$

$\sigma=5$:

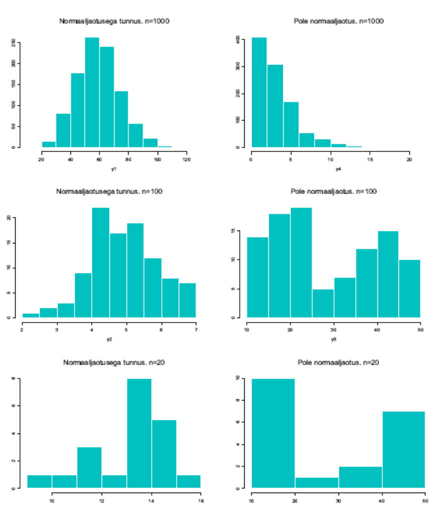
$$\mathrm{P}(45 \le X \le 55)=\Phi\left(\frac{55-50}{5}\right)-\Phi\left(\frac{45-50}{5}\right)=2\Phi(1)-1=2*0,8413-1=0,6827$$

$$\mathrm{P}(X<45 \cup X>55)=1-0,6827=0,3173$$

Normaaljaotus

Normaaljaotuse eelduse kontrollimine graafiliselt

Histogramm



Tõenäosuspaber [*Normal prabability plot*]