Andmetüübid I

Empiirilised andmed — kogemusest (vaatlusest, katsest) pärinevad andmed.

Teoreetilised andmed — teatud teoreetilistest printsiipidest tuletatud või mudelist arvutatud andmed, mis kehtivad enamasti vaid teatud eelduste korral (juhul kui ...).

Andmed koosnevad reeglina
- **vaatluste** (objektide) kirjeldustest **tunnuste** abil ja
- **metaandmetest** (tunnuste kirjeldus, vaatluste päritolu).

	Tunnused			
Vaatlused	1	2	3	4
1	v11	v12	v13	v14
2	v21	v22	v23	v24
3	v31	v32	v33	v34
4	v41	v42	v43	v44

Tunnuste väärtused

Jaotused

Andmetüübid II

Tunnuste andmemudelile vastavad tüübid

Arvuline ehk **kvantitatiivne** tunnus
- pidev
- diskreetne

↕ Klassifitseerimine ja kodeerimine

Mittearvuline ehk **kvalitatiivne**
- järjestatav
- **nominaalne** ehk nimeline
  - binominaalne
  - multinominaalne

Formaalselt järjestada saab ikka!

Jaotused

Andmetüübid III

Tunnuste funktsionaalsed tüübid

Millises rollis tunnus esineb?

Seletav tunnus ehk **kirjeldav tunnus** ehk **argumenttunnus** ehk **sõltumatu tunnus**

Prognoositav tunnus ehk **funktsioontunnus** ehk **sõltuv tunnus**

Sõltuv tunnus sõltub sõltumatust. Kas seletavad tunnused on omavahel sõltumatud või mitte, on teine teema.

Jaotused

Jaotused

Tunnuste väärtused on vaatlustel erinevad.
Väärtused jaotuvad väärtusvahemikesse või väärtusklassidesse.
Vaatlused jaotuvad ka (geograafilises) **ruumis** (ühed on siin ja teised seal).

⇒ **jaotus ehk jaotumus**

Kui väärtused või asukohad ei ole üheselt määratud (neid mõjutavad mitteteadaolevad tegurid ehk/või juhus), siis on **juhuslik jaotus.**

Juhuslik jaotus võib olla osaliselt mittejuhuslik. Juhuslikkus tekib paljude väikese mõjuga faktorite koostoimel.

Jaotused

Juhuslikkus I

Täringuviske tulemus
Libedalt teelt väljasõit
Metsas hulkuva põdra koordinaadid mingil ajahetkel
Kas ka märkilaskmise ja märkiviskamise tulemus?
Neid sündmusi saab käsitleda juhuslikena

Suurte arvude seadus:
Juhuslike sündmuste suure arvu korral ilmnevad seaduspärasused
ehk:
suure hulga juhuslike sündmuste koosmõju ei sõltu enam vaid juhusest

Suurte arvude seaduse formuleeris Jakob Bernoulli

Jaotused

Juhuslikkus II

Juhuslikud üksikvaatlused üldistuvad seaduspärasusteks (jaotusteks, statistilisteks seosteks).
Andmetöötluse käigus lihtsustatakse ja üldistatakse tegelikkuse lõputut keerukust ja detailsust.

Üldistatakse ja lihtsustatakse ka kartograafias. Kartograaf tegeleb ka andmetöötlusega.

Maastik on lõpmata keeruline, kaart on lihtne ja mõistetav (nende jaoks, kes oskavad kaarti lugeda)

Tegelikkus on keeruline, andmetöötlus ei ole! ☺

Suurte arvude seaduse formuleeris Jakob Bernoulli

Jaotused

Jakob Bernoulli (1655…1705)

Bernoullide suguvõsas oli 11 tuntud matemaatikut !

Suurte arvude seaduse formuleeris Jakob Bernoulli (1655…1705)

Daniel Bernoulli (Jakobi vennapoeg) 1700…1782 oli hüdrodünaamika ja matemaatilise füüsika teooria rajaja.

Mis maa on Helveetsia?

Confœderatio Helvetica asutati 1291

Jaotused

Juhuslikkusega seotud mõisteid I

Juhuslikkus — määramatus ehk mittedetermineeritus.

Juhus — mitteteadaolev põhjus (juhul, kui maailma käsitleda deterministlikult).

Juht(um) — (*case,* kaasus*)*, mitmete tunnustega kirjeldatav **sündmus** (andmetöötluse aspektist).

Juhuslik muutuja — muutuja, mille väärtusi mõjutab juhus.

Juhuslikud arvud — juhusliku muutuja sõltumatud väärtused. Juhuslikud arvud moodustavad juhusliku jaotuse.

Tõenäosus — võimalikkuse mõõt. Mõõdab, mil määral on üks või teine sündmus võimalik.

Jaotused

Juhuslikkusega seotud mõisteid II

Vaatlus — nähtuste jälgimine ja jälgimistulemuste talletamine. Andmetöötluses kasutatakse hästi üldises tähenduses.

Katse — nähtuste esilekutsumine nende vaatlemiseks.

Sündmus — katse tulemus.

Jaotus — seaduspärasus, mille järgi muutuja igale väärtusele vastab selle väärtuse esinemise tõenäosus või esinemissagedus.

Tõenäosusjaotus — tõenäosuse jagunemine sündmuse variantide või väärtuste vahel. Tõenäosus on pidev muutuja vahemikus 0 ... 1.

Sagedusjaotus — sündmuse variantide või väärtuste esinemise sagedus.

Absoluutne sagedus — mitu korda esines?

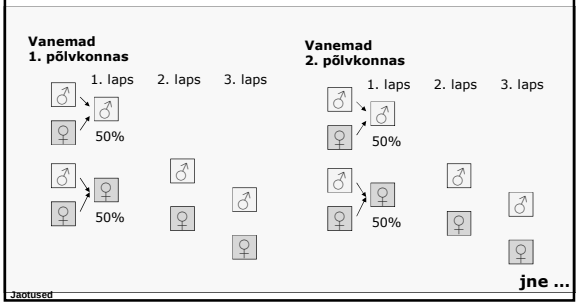
Suhteline sagedus — esinemiskordade osakaal (protsent).

Mille suhtes?

Jaotused

Arvutused tõenäosustega

Oletagem, et poisi või tüdruku sündimise tõenäosus on võrdselt 50%. Oletagem veel, et 100 000 elanikuga maal otsustavad kõik pered saada lapsi seni, kuni sünnib poeg ja **seejärel** nad rohkem lapsi ei saa. Milline on meeste ja naiste vahekord sellel maal 100 põlvkonna järel?



jne …

Jaotused

Arvutused tõenäosustega

Juhuslike ja omavahel sõltumatute sündmuste koos esinemise tõenäosus võrdub nende üksiksündmuste tõenäosuste korrutisega

Üks kantpea oli mures, et lennukis, millega ta reisib, võib olla pomm. Ta uuris välja, et pommi olemasolu tõenäosus lennukis on väike, kuid mitte piisavalt väike tema jaoks. Nüüd võtab ta alati väikese lõhkekeha endaga kaasa, sest kahe pommi samas lennukis oleku tõenäosus oleks ju lõpmata väike. Kas ta käitus **õigesti**? Kui ei, siis kuidas talle selgitada, et ta eksis?

Kahe teineteisest sõltumatu ja juhusliku sündmuse (pommi olemasolu) tõenäosus võrdub üksiksündmuste tõenäosuste korrutisega. Näiteks $0{,}001 \times 0{,}001 = 0{,}00001$.

Kas siin on tegemist sõltumatute ja juhuslike sündmustega?

Jaotused

Arvutused tõenäosustega II

Jääpangal triivivatele polaaruurijatele kavatsetakse visata varustust kahelt erinevalt lennukilt. Tõenäosus ühelt lennukilt visatud varustuse sattumiseks jääpangale on 70% ja teiselt lennukilt 80%.
Milline on tõenäosus, et uurijad saavad varustust?
Milline on tõenäosus, et nad jäävad varustusest ilma?
Milline on tõenäosus, et nad saavad topeltvarustuse?

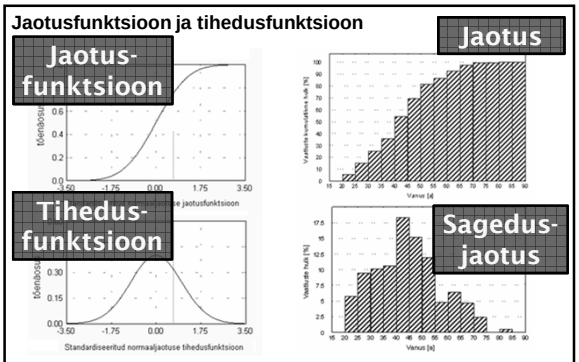
Topeltvarustuse saamise tõenäosus = kahe sõltumatu juhusliku sündmuse tõenäosuse korrutis = 0,7 · 0,8 = 0,56 = 56%

Varustusest ilmajäämise tõenäosus = (1 – 0,7) · (1 – 0,8) = 0,3 · 0,2 = 0,06 = 6%

Varustuse saamise tõenäosus = 1 – varustusest ilmajäämise tõenäosus = 1 – 0,06 = 0,94 = 94%

Jaotused

Jaotusfunktsioon ja tihedusfunktsioon



Pideva tunnuse **jaotusfunktsioon** ja **tihedusfunktsioon**

Diskreetse tunnuse **jaotus** ja **sagedusjaotus**

Mis on kumulatiivse jaotuse eelis?

Jaotused

Jaotuste parameetrid

Jaotusi saab kirjeldada arvuliste tunnuste — **jaotusparameetrite** abil

Põhilised jaotusparameetrid:

- **keskväärtus**,
- dispersioon ehk **keskmine ruuthälve**,
- standardhälve ehk **ruutkeskmine hälve**.

Andmebaase **normaliseeritakse**.

Jaotusi **normeeritakse** (jagatakse standardhälbega või mingi muu parameetriga, mille suhtes parajasti normeeritakse) ja

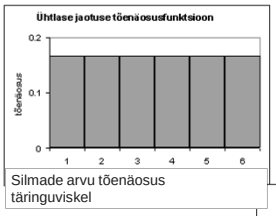
standardiseeritakse (lahutatakse keskväärtus ja jagatakse standardhälbega).

Aga milleks?

Ikka selleks, et oleks neid mugavam võrrelda.

Jaotused

Ühtlane jaotus

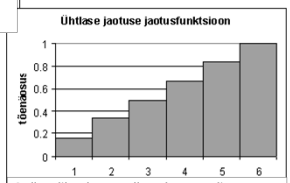


Silmade arvu tõenäosus täringuviskel

Sündmuse võrdvõimalike variantide oodatava sageduse jaotus

Milline peab olema täring, et silmade arvu jaotus oleks ühtlane jaotus?

Täring peab olema võrdkülgne ja ühetaolise massiga. Siis on igale küljele langemine sama tõenäoline.

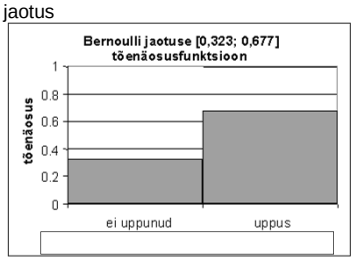


Selle või vähema silmade arvu tõenäosus

Jaotused

Bernoulli jaotus

Kaheväärtuseline jaotus:
sündmuse variantide jaotus ehk sündmuse toimumise ja mittetoimumise jaotus



Titanicu pardal olnud isikute uppunute või mitteuppunute hulka kuulumise tõenäosus.

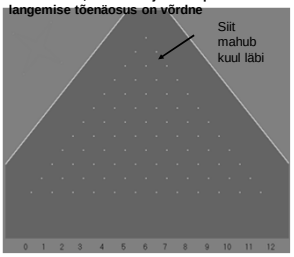
Miks mitte sagedus?

Jaotused

Binoomjaotus I

Näitab tulemuste tõenäosust kaheväärtuseliste sõltumatute katsete **kindla pikkusega** seerias.

Galtoni võre, kus ühele ja teisel poole langemise tõenäosus on võrdne



Parameetrid:

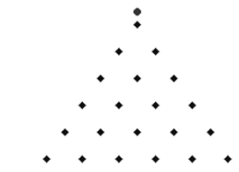
- **katseseeria pikkus ja**
- **ühe katsetulemuse** (eduka katse) **tõenäosus.**

Iga naela peal võib kuulike põrkuda kas vasakule või paremale. Kõik trajektoorid on võrdse tõenäosusega, aga keskmistesse lahtritesse kukuvad kuulid sagedamini, sest

keskmistesse lahtritesse viib rohkem trajektoore.

Jaotused

Iga naela peal võib kuulike põrkuda kas vasakule või paremale. Kõik trajektoorid on võrdse tõenäosusega, aga keskmistesse lahtritesse kukuvad kuulid sagedamini.



Keskmistesse lahtritesse viib rohkem trajektoore.

Kas Galtoni võre modelleerib ka vihmapiiskade teekonda palgivirnas?

Ei, sest vihm ei lange vaid virna harjale.

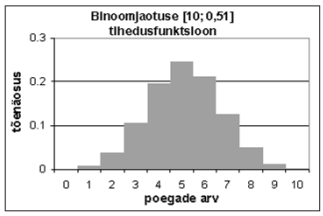
Kas binoomjaotus sobib poegade ja tütarde arvu modelleerimiseks peres?

Sobib, kui pere suurus on ette antud ja kui lepime eeldusega, et kas poja või tütre sünd on juhuslik.

Paul Trow, 2007. http://ptrow.com/articles/Galton_June_07.htm

Binoomjaotus II

NB! Binoomjaotus on sümmeetriline vaid siis, kui variantide tõenäosused on võrdsed.



NB! Poegade arv võib olla ka null.

Poegade arvu tõenäosuse jaotus 10-lapselises peres juhul, kui poja sündimise tõenäosus on 51%. Poegade ja tütarde arv peaks olema tasakaalus.

Mida võib suurte perede kohta järeldada, kui vaatlusandmetes ei vasta poegade ja tütarde sagedus binoomjaotusele?

Teoreetiline jaotus eeldab juhuslikkust. Ju siis ei ole sündmus juhuslik.

Jaotused

sir Francis Galton (1822…1911)

Eugeenika rajaja. Oli veendunud, et intelligentsus on **eelkõige** pärilik.
Kas eliitkoolid ja diferentseeritud suurusega vanemapalk on eugeenika teooria elluviimine?

Tõestas, et paljude omavahel normaaljaotusega paiknevate normaaljaotuste liitmisel saadakse normaaljaotus.

Võttis kasutusele küsitlusmeetodi, korrelatsioon- ja regressioonanalüüsi.

Leiutas sõrmejälgede võtmise.

Tegeles geograafiaga (troopiliste alade uurimise) ja meteoroloogiaga (võttis kasutusele ilmakaardid ja antitsükloni mõiste).

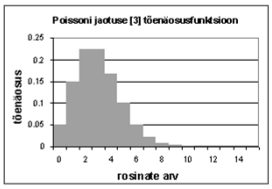
Ch. Darwini sugulane (ühine vanaisa — Erasmus Darwin).

Intelligentsuse määr sõltub mõõtmisviisist. Intelligentsitestid on subjektiivsed ja (lääne)kultuurikeskses.

Loomulik (looduslik) valik on inimeste puhul ehk **õiglasem** otsustaja kui teine inimene.

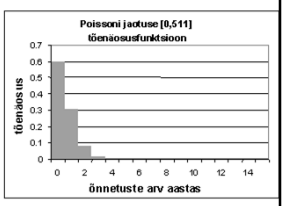
Jaotused

Poissoni jaotus **Vaid üks parameeter (λ) = keskväärtus = dispersioon**



Ühes rosinakuklis esinevate rosinate arvu tõenäosused eeldusel, et **keskmiselt** on ühes kuklis kolm rosinat (λ = 3)

Katseseeria on pikk



Ühe aasta jooksul liiklusõnnetusse sattumise kordade tõenäosus juhul, kui on 60% tõenäosust, et teiega aasta jooksul liiklusõnnetust üldse ei juhtu (λ = 0,511). Viimasel juhul on siiski ~8% tõenäosust, et vaadeldaval aastal satute koguni kahte õnnetusse. Eeldatud on, et õnnetusse sattumine on muutumatu tõenäosusega ja juhuslik.

Poissoni jaotust saab muuhulgas kasutada riskiprognoosides. Eeldades, et sündmused on juhuslikkud ja üksteisest sõltumatud.

Jaotused

Siméon Denis Poisson (1781…1840)

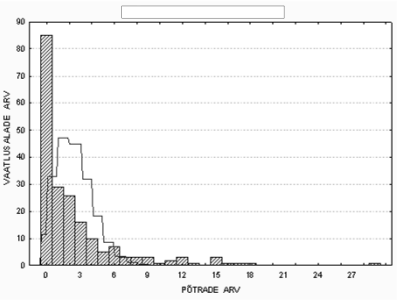
Isa oli sõdur (mitteaadlik). Prantsuse revolutsiooni järel sai maakonna juhiks.
Toetas igati poja hariduse omandamist ja karjääri.
Koordinatsioonihäirete tõttu loobus Siméon tegelemast kujutava geomeetriaga ja eksperimentaalse füüsikaga. Piirdus teoreetilise füüsika, astronoomia ja matemaatikaga.
300…400 publikatsiooni, kuid töötas alati **korraga vaid ühe** probleemiga.

Olevat kolleegile öelnud:
"Elus on vaid kaks head asja, matemaatika uurimine ja selle õpetamine."

Jaotused

Poissoni jaotus ja empiiriline jaotus

Põtrade asustustiheduse hinnang Ida-Viru vaatlusaladel



Objektide juhupaiknemise korral on vaatluste arv Poissoni jaotusega.

Kui empiiriline jaotus erineb Poissoni jaotusest, siis mida see näitab?

Näitab, et teoreetilise jaotuse mõni eeldus (eelkõige juhuslikkus) ei kehti. Näitab, et põtrade paiknemine ei ole juhuslik.

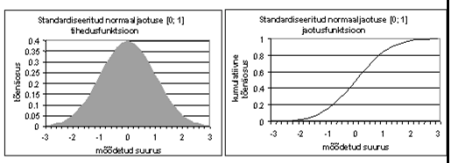
Poissoni jaotus

Normaaljaotus

Oodatavalt normaaljaotusega on juhuslikud suurused, mille väärtus
- sõltub paljudest faktoritest (katseseeria on lõpmata pikk),
- kõik faktorid avaldavad vaid ühetaolist nõrka mõju ja
- kõigi faktorite mõju võib olla nii positiivne kui ka negatiivne ning (positiivse mõju tõenäosus 50%)
- mõjud liituvad.

Näiteks mõõtmisvead:
Kui kahe kindelpunkti vahemaad korduvalt ja lõpmata väikestes mõõtühikutes mõõta, saadakse iga kord erinev tulemus. Mõõtmistulemust mõjutab palju faktoreid. Enamik tulemusi koondub siiski keskväärtuse ümber.

Gaussi kõver — normaaljaotuse tihedusfunktsiooni graafik



Normaaljaotus on teoreetiline mudel, mida tegelikkuses ei esine. Nagu ei ole olemas ei keskmist inimest ega ideaalset inimest.

Jaotused

Carl Friedrich Gauss (1777—1855)

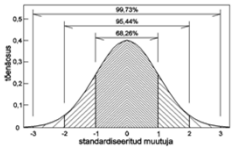
Saksa matemaatik, füüsik ja astronoom. Göttingeni ülikooli professor ja observatooriumi direktor.
Tõestas algebra põhiteoreemi (igal kopleksarvuliste kordajatega algebralisel võrrandil on vähemalt üks kompleksarvuline lahend),
võttis kasutusele vähimruutude meetodi,
lõi kompleksarvude teooria,
lõi võrrandsüsteemide lahendamise meetodi,
uuris **maamagnetismi**,
tegeles planeetide orbiitide arvutamise teooriaga,
tegeles telegraafi loomisega,
jõudis mitte-eukleidilise geomeetriani, aga ei avaldanud sel teemal midagi.
Kartis, et mitteeukleidilise geomeetria võimalikkuse uskumise avaldamine võib kahjustada tema reputatsiooni.

Normaaljaotus

Parameetrid: keskväärtus ja dispersioon (või standardhälve).

Miks on dispersiooni eespool standardhälvet?

Teoreetiliselt võib normaaljaotusega muutuja omada väärtusi vahemikus +∞ … −∞, seejuures:
- keskväärtusest 1 standardhälbe ulatuses **68,**26**%** väärtustest,
- 2 standardhälbe ulatuses **95**,44% väärtustest,
- 3 standardhälbe ulatuses **99**,73% väärtustest.



Ehk:
***) umbes 2 / 3**
***) umbes 95%**
***) peaaegu kõik**

Keskväärtus (matemaatiline ootus) — juhusliku suuruse jaotuse paiknemist iseloomustav arv. Teoreetiline suurus, mida püütakse võimalikult täpselt hinnata

Keskmine — keskväärtuse hinnang valimi järgi. Konkreetsetest arvudest arvutatud arv

Jaotused

Võiks meelde jääda

- Suurte arvude seadus ütleb, et juhuslike sündmuste suure arvu korral ilmnevad seaduspärasused.
- Omavahel sõltumatute sündmuse koosesinemise tõenäosus võrdub üksiksündmuste tõenäosuste korrutisega.
- Paljude ülesannete puhul on lihtsam esmalt arvutada vastandsündmuse tõenäosus.
- Binoomjaotus näitab loendi tõenäosust kindla pikkusega juhusliku tulemusega kaheväärtuselises katseseerias.
- Poissoni jaotus näitab loendi tõenäosust pikas katseseerias, kui loendi keskmine on teada.
- Normaaljaotus on pidevale muutujale mõjuvate paljude väikese juhusliku mõjuga faktorite koosmõju oodatav tulemus.
- Teoreetiliselt võib normaaljaotusega muutuja omada väärtusi vahemikus +∞ … −∞, seejuures on:
  - keskväärtusest ±1 standardhälbe ulatuses umbes 2/3 väärtustest,
  - ±2 standardhälbe ulatuses umbes 95% väärtustest,
  - ±3 standardhälbe ulatuses peaaegu kõik väärtused.
- Kui empiiriline jaotus erineb oluliselt juhuslikkust eeldavast teoreetilisest jaotusest, siis teoreetilise jaotuse eeldused ilmselt nende empiiriliste andmete puhul ei kehti – jaotust loonud protsess ei ole juhuslik.

Jaotused