Bayesiaanlik statistika

Sissejuhatus

See õpik on kirjutatud inimestele, kes kasutavad, mitte ei uuri, statistikat. Õpiku kasutaja peaks olema võimeline töötama R keskkonnas (vt meie eestikeelset R-i konspekti). Meie lähenemised statistika õpetamisele on arvutuslikud, mis tähendab, et me eelistame meetodi matemaatilise aluse asemel õpetada selle kasutamist ja tulemuste tõlgendamist. See õpik on bayesiaanlik ja ei õpeta sageduslikku statistikat. Me usume, et nii on lihtsam ja tulusam statistikat õppida ja et Bayesi statistikat kasutades saab rahuldada 99% teie tegelikest statistilistest vajadustest paremini, kui see on võimalik klassikaliste sageduslike meetoditega. Me usume ka, et kuigi praegused kiired arengud bayesi statistikas on tänaseks juba viinud selle suurel määral tavakasutajale kättesaadavasse vormi, toovad lähiaastad selles vallas veel suuri muutusi. Nende muutustega koos peab arenema ka bayesi õpetamine.

Me kasutame järgmisi R-i pakette, mis on kõik loodud bayesi mudelite rakendamise lihtsustamiseks: rethinking, brms, rstanarm, BayesianFirstAid ja bayesplot. Lisaks veel bayesboot bootstrapimiseks. Bayesi arvutusteks kasutavad need paketid Stan ja JAGS mcmc sämplereid (viimast küll ainult BayesianFirstAid paket). Selle õpiku valmimisel on kasutatud McElreathi, Kruschke ja Gelmani õpikuid (VIITED).

Kuna osadele lugejatele on ilmselt õpetatud sageduslikku statistikat, võrdleme järgnevalt lühidalt sageduslikku ja bayesi statistilisi paradigmasid, et oleks paremini arusaadav, millest selle õpiku lugeja ilma jääb. Inimesed, kes ei ole õppinud sageduslikku statistikat võivad rahumeeli selle osa vahele jätta ning otsemaid siirduda osa 1. juurde.

Ajaloost ja tõenäosusest

Bayesiaanlik ja sageduslik statistika sünnitati üksteise järel sama ema poolt. Tema nimi on Laplace ja ta arendas välja kõigepealt bayesi statistika alused ning seejärel sagedusliku statistika omad. Sagedusliku statistika tekkimise ja hilisema õitsengu peamine põhjus oli arvutuslik lihtsus. Lihtsalt bayesi meetoditega ei olnud arvutustehnilistel põhjustel võimalik korralikult teadust teha enne 1990-ndaid aastaid, mil personaalarvutite kiire levik algatas suure buumi nende meetodite arendamises. Praegu on maailmas bayesi ja sageduslikku statistikat umbes pooleks (vähemalt uute meetodite arendustöö poole pealt), Eestis aga bayesi statistika 2017 aasta seisuga peaaegu, et puudub.

Statistika on ainulaadne teadus selle poolest, et see on alates oma lätetest 18. sajandil lahknenud kahes suunas, mis lahendavad samu teaduslikke küsimusi aga sellest hoolimata suures osas üksteisega ei haaku. Need kaks suunda on Bayesi statistika ja sageduslik statistika. Meie oleme bayesiaanid selles mõttes, et võimaluse korral eelistame me oma töös bayesi statistikat, ja ainult siis, kui seda on liiga keeruline rakendada (või kui meile ei meeldi inimene, kellele me statistikat teeme), kasutame sageduslikke meetodeid. Õnneks tuleb seda üha harvemini ette.

```
Statistikast leiab mõned meetodite klassid, mida võib edukalt vaadata nii
sagedusliku kui bayesiaanliku mätta otsast: eriti bootstrappimine ja osad
False Discovery Rate-i arvutamise meetodid.
Aga need on pigem erandid, mis kinnitavad reeglit.
```

Kahe statistika põhiline erinevus ei tule mitte matemaatikast — mõlemad harud lähtuvad samadest tõenäosusteooria aksioomidest ja nende vahel puuduvad matemaatilised lahkarvamused — vaid tõenäosuse tõlgendusest.

Bayesi tõlgenduses on tõenäosus on kas üksikteadlase või (vähemate arvates) teadusüldsuse usu määr mingi hüpoteesi kehtimisse. Hüpotees võib näiteks olla, et järgmise juulikuu sademete hulk Vilsandil jääb vahemikku 22 kuni 34 mm. Kui Bayesi arvutus annab selle hüpoteesi tõenäosuseks 0.57, siis oleme me selle teadmise najal nõus maksma mitte rohkem kui 57 senti kihlveo eest, mille alusel makstakse juhul, kui see hüpotees tõeseks osutub, välja 1 EUR (ja me saame vähemalt 43 senti kasumit).

Sageduslikud statistikud usuvad, et selline tõenäosuse tõlgendus on ebateaduslik, kuna see on "subjektiivne". On võimalik, et n teadlast arvutavad korrektselt samade andmete põhjal n erinevat tõenäosust ja usuvad seega samade tõendite põhjal erinevaid asju. See võib juhtuda siis, kui nad alguses lähtuvad väga erinevatest taustauskumustest oma hüpoteeside kehtimise kohta. Kui te usute, et teie taustateadmised ei tohi mingil juhul mõjutada järeldusi, mis te oma andmete põhjal teete, siis te ei ole bayesiaan. Sel juhul pakub alternatiivi sageduslik tõenäosuse tõlgendus. Sageduslik tõenäosus on defineeritud kui teatud tüüpi andmete esinemise pikaajaline suhteline sagedus. Näiteks, kui me viskame münti palju kordi, siis peaks kullide (või kirjade) suhteline sagedus meile andma selle mündi tõenäosuse langeda kiri üleval. Selline tõenäosus on omistatav ainult sündmustele, mille esinemisel on sagedus. Teaduslik teooria ei ole selline sündmus. Seega ei ole sageduslikus statistikas võimalik rääkida ka hüpoteesi kehtimise tõenäosusest. Sageduslik lahendus on, selle asemel, et rääkida meie hüpoteesi tõenäosusest meie andmete korral, rääkida andmete, mis sarnanevad meie andmetega, esinemise tõenäosusest teatud hüpoteesi (mis aga ei ole meie hüpotees) kehtimise korral. Seega omistatakse sagedus (= tõenäosus) andmetele, mitte hüpoteesidele. Järgnevalt toome näite, kuidas bayesiaan ja sageduslik statistik lahendavad sama ülesande.

Näide: kahe grupi võrdlus

meil on 2 gruppi, katse ja kontroll, millest kummagis n mõõtmist ja me soovime teada, kas katsetingimus mõjutab mõõtmistulemust piisavalt, et olla teaduslikult huvitav. Me eeldame, et andmed on normaaljaotusega ja et n ei ole väga väike.

Bayesiaan

Statistiline küsimus on Bayesiaanil ja sageduslikul statistikul sama: kas ja kui palju erinevad kahe grupi keskväärtused? Bayesiaan alustab sellest, et ehitab kaks mudelit: andmete mudel ja taustateadmiste mudel. Kui andmed on normaaljaotusega siis on ka andmemudel normaaljaotus. Täpsemalt, andmemudel modeleerib, mida meie andmete põhjal saab öelda selle populatsiooni, kust andmed korjati, keskväärtuse kohta. See mudel on normaaljaotus, mille keskväärtus võrdub valimiandmete keskväärtusega ja standardhälve võrdub valimi standardhälvega, mis on jagatud ruutjuurega n-st (SEM = sd(valim)/sqrt(n)).

Taustateadmiste mudel on sageli samuti normaaljaotus. Kui meil on palju taustateadmisi, siis on see jaotus kõrge ja kitsas, kui meil on vähe taustateadmisi, siis on see madal ja lai. Igal juhul järgmise sammuna korrutab bayesiaan need kaks jaotust saades tulemuseks kolmanda normaaljaotuse, mille ta seejärel normaliseerib nii, et jaotuse alune pindala = 1.

See kolmas jaotus on posterioorne tõenäosus, mis sisaldab kogu infot, millest arvutada kõige tõenäolisem katseefekti suurus koos ebakindluse määraga selle ümber (mida rohkem andmeid, seda väiksem ebakindlus) ja tõenäosused, et tegelik katseefekt jääb ükskõik milllisesse meid huvitavasse vahemikku.

Sageduslik statistik

Sageduslik lähenemine seevastu sisaldab ainult ühte mudelit, mida võrreldakse valimi andmetega. Sageduslik statistik alustab täpselt sama moodi nagu bayesiaan, tekitades eelmisega identse anmdemudeli, mis on keskendatud valimi keskväärtusele ja omab sama laiust (SEM). Seejärel nihutab ta oma andmemudelit niipalju, et normaaljaotuse tipp ei ole enam valimi keskväärtuse kohal vaid hoopis 0-väärtuse kohal (näiteks 0-effekti kohal). Samas, jaotuse laius ei muutu. Seda nullile tsentreeritud mudelit kutsutakse null-hüpoteesiks (H0). Järgmine samm erineb samuti bayesiaani omast. Nüüd võrreldakse oma valimi keskväärtust H0 jaotusega. Kui valimi keskväärtuse kohal on H0 jaotus kõrge, siis on andmete tõenäosus H0 kehtimise korral suur. Ja kui valimi keskväärtuse kohal on H0 normaaljaotus madal, siis on andmete esinemise tõenäosus H0 all madal. Seda tõenäosust kutsutakse p väärtuseks. Mida väiksem on p väärtus, seda vähem tõenäolised on teie andmed juhul, kui H0 on tõene ja katseefekt võrdub nulliga. Tehniliselt on p defineeritud kui "teie andmete või 0-st veel kaugemal asuvate andmete esinemise pikaajaline suhteline sagedus tingimusel, et H0 kehtib".

tulemuste tõlgendamine

Kui sageduslik statistik kirjutab oma artiklites, et tema "efekti suurus on statistiliselt oluline 0.05 olulisusnivool" siis ta ütleb sellega, et tema poolt arvutatud $p < 0.05$. Selle korrektne tõlgendus on, et juhul kui statistik pika aja jooksul võtab omaks "statistiliselt olulistena" kõik tulemused, millega kaasnev $p < 0.05$ ja lükkab tagasi kõik tulemused, mille $p > 0.05$, siis sooritab ta 5% sagedusega tüüp 1 vigu. See tähendab, et igast sajast tõesest H0-st, mida ta testib, võtab ta keskeltläbi 5 vastu kui statistiliselt olulised. Sellisel kujul töötab sageduslik statistika väga hästi — see on parim viis tüüp 1 vigade sageduse pikaajaliseks fikseerimiseks. Aga kuna me ei tea ühegi üksiku testi kohta ette, kas see testib kehtivat või mittekehtivat H0-i, siis on selle kasutegur katseseeriate ühekaupa tõlgendamisel vaieldav. Tuletame meelde, et sageduslikus statistikas ei saa rääkida H0 kehtimise tõenäosusest vaid peab rääkima andmete tõenäosusest (= andmete esinemise sagedusest) tingimusel, et H0 kehtib.

Kas p väärtusi saab tõlgendada ühekaupa kui hinnangut tõendusmaterjali hulgale, mida teie valim pakub H0 vastu? Selle üle on vaieldud üle 80 aasta, kuid tundub, et ainus viis seda kas või umbkaudu teha on bayesiaanlik. Kui te mõtlete p väärtuse definitsioonile, siis peaksite mõistma, miks p väärtust, mis on defineeritud pikaajalise sagedusena, on raske rakendada üksiksündmustele. Bayesiaanliku p väärtuste tõlgendamiskalkulaatori leiate aadressilt ...

tüüpiline tulemuse kirjeldus artiklis:

1. sageduslik: the effect size is q ($p < 0.01$).
2. bayesiaanlik: the most likely effect size is q1 (90% CI= q2, q3) and the probability that the true effect is less than zero is q4 percent. [90% CI - credible interval; tähendab, et me oleme 90% kindlad, et tegelik efekti suurus asub vahemikus q2 ... q3]

kahe paradigma erinevused

1. sageduslikus statistikas võrdub parim hinnang tegelikule efekti suurusele valimi keskmise efekti suurusega. Bayesi statistikas see sageli nii ei ole sest taustateadmiste mudel mõjutab seda hinnangut. Paljud mudelid püüavad ekstreemseid valimeid taustateadmiste abil mõistlikus suunas veidi nihutada, niiviisi vähendades ülepaisutatud efektide avaldamise ohtu.
2. sageduslik statistika töötab tänu sellele, et uurija võtab vastu pluss-miinus otsuseid: iga H0 kas lükatakse ümber või jäetakse kehtima. Seevastu bayesiaan mõtleb halli varjundites: sissetulevad andmed kas suurendavad või vähendavad hüpoteeside tõenäosusi (mis jäävad aga alati > 0 ja < 1).
3. p väärtused kontrollivad tüüp 1 vigade sagedust ainult siis, kui katse disaini ja hilisema tulemuste analüüsi detailid on enne katse sooritamist jäigalt fikseeritud (või eelnevalt on täpselt paika pandud lubatud variatsioonid katse- ja analüüsi protokollis). Eelkõige tähendab see, et valimi suurus ja kasutatav(ad) statistiline(sed) test(id) peavad olema eelnevalt fikseeritud. Me saame p väärtuse arvutada vaid üks kord ja kui $p = 0.051$, siis oleme sunnitud H0 paika jätma ning efekti deklareerimisest loobuma. Me ei saa lihtsalt katset juurde teha, et vaadata, mis juhtub. Bayesiaan seevastu võib oma posterioorse tõenäosuse arvutada kasvõi pärast iga katsepunkti kogumist ning katse peatada kohe (või alles siis), kui ta leiab, et tema posterioorne jaotus on piisavalt kitsas, et teaduslikku huvi pakkuda.
4. sagedusliku statistika pluss-miinus iseloom tingib selle, et kui tegelik efekti suurus on küll teaduslikult huvitav, aga siiski liiga väike, et teie katsesisese varieeruvuse ja valimi suuruse juures anda $p < 0.05$, siis annavad statistiliselt olulisi tulemusi ainult populatsiooniga võrreldes ülepaisutatud efekti suurusega ja alandatud varieeruvusega valimid. (Selliseid valimeid tekib tänu juhuslikele valimiefektidele.) Nii saab süstemaatiliselt kallutatud teaduse, mis hindab kordades üle oma efektide suurusi. Bayesi statistikas seda probleemi ei esine, kuna otsused ei ole pluss-miinus tüüpi.
5. bayesi statistika ei hoia tüüp 1 vigade sagedust kontrolli all. See-eest võtleb see nn valehäirete vastu, milleks kaasajal kasutatakse enim hierarhilisi shrinkage mudeleid. Neid käsitleme oma õpikus päris lõpus. See on ka bayesi vaste sageduslikus statistikas kasutatavatele multiple testingu suhtes adjusteeritud p väärtustele.

See on kõik, mida me sagedusliku statistika kohta ütleme. Mitte miski, mis järgneb, ei eelda sagedusliku paradigma tundmist ega valdamist.

1. osa: Mudel ja maailm

Andmeanalüüs ja statistika (siin sünonüümid) on lahutamatu osa igast loodusteadusest. Järgnevalt seletan, miks.

Suur ja väike maailm

Kuna maailm on liiga suur ja keeruline, et seda otse uurida, lõikavad teadlased selle väiksemateks tükkideks, kasutades tordilabidana teaduslike hüpoteese. Tüüpiline hüpotees pakub välja mittematemaatilise seletuse mõnele kitsalt piiritletud loodusnähtusele. Näiteks darvinistlik evolutsiooniteooria püüab seletada evolutsiooni toimemehhanisme. Seda teooriat võib võrrelda empiiriliste andmetega.

Mis juhtub, kui teie lemmikhüpotees on andmetega kooskõlas? Kas see tähendab, et see hüpotees on tõene? Või, et see on tõenäoliselt tõene? Kahjuks on vastus mõlemale küsimusele eitav. Põhjuseks on asjaolu, et enamasti leiab iga nähtuse seletamiseks rohkem kui ühe alternatiivse teadusliku hüpoteesi (näit. lamarksistlik evolutsiooniteooria) ning rohkem kui üks üksteist välistav hüpotees võib olla olemasolevate andmetega võrdses kooskõlas. Asja teeb veel hullemaks, et teoreetiliselt on võimalik sõnastada lõpmatult palju erinevaid teooriaid, mis kõik pakuvad alternatiivseid ja üksteist välistavaid seletusi samale nähtusele.

Olgu peale, kui me vaatame maailma kõiketeadja jumala perspektiivist, siis tema võib vaadelda kõikehõlmava tõendusmaterjali sobivust kõigi võimalike teooriatega ning valida välja selle ainsa teooria, mis kõige paremini tõendusmaterjaliga sobib. Kuigi, see eeldaks, et jumalal on lõpmata palju andmeid, sest muidu ei oleks tal loogiliselt võimalik lõpmata paljude teooriate vahel valida - aga jumala jaoks on kõik võimalik. Igal juhul meie, surelike, jaoks tähendab see, et teaduslikus "faktis" saab alati kahelda sest kunagi ei või kindel olla, et parimad teooriad lõpmata suurest teooriapilvest ei ole meil täiesti tähelepanuta jäänud ning, et meie jaoks eksisteerivad andmed kajastaksid hästi kõiki võimalikke andmeid! On selge nagu seebivesi, et mida vähem aega me kulutame teoorialoomeks ja andmete kogumiseks, seda vähem usutavad on ka meie teaduslikud järeldused. Enamasti on nii, et mida kehvem on olukord andmerindel, seda rohkem vajame statistikat. Kui meil õnnestuks oma andmetest ilma statistikata saia teha, ei kõhkleks me hetkegi! Eriti, kuna statistikaga käivad käsikäes statistilised mudelid.

Mudeli väike maailm

Ülalmainitud teadusliku meetodi puudused tingivad, et meie huvides on oma teaduslikke probleeme veel ühe taseme võrra lihtsustada, taandades need statistilisteks probleemideks. Selleks tuletame me tavakeelsest ja laiahaardelisest teaduslikust teooriast täpselt formuleeritud matemaatilise mudeli ning seejärel asume uurima oma mudelit.

Mudeli maailm erineb päris maailmast selle poolest, et mudeli maailmas on kõikvõimalikud sündmused, mis põhimõtteliselt võivad juhtuda, juba ette teada ja üles loetud (seda sündmuste kogu kutsutakse parameetriruumiks). Seega, tehniliselt on mudeli maailmas üllatused võimatud.

Mudeli eeliseks teooria ees on, et hästi konstrueeritud mudel on lihtsamini mõistetav — erinevalt vähegi keerulisemast teaduslikust hüpoteesist on mudeli eeldused ja ennustused läbinähtavad ja täpselt formuleeritavad. Mudeli puuduseks on aga, et erinevalt teooriast ei ole mingit võimalust, et mudel vastaks tegelikkusele ehk oleks tõene. Seda sellepärast, et mudel on taotluslikult lihtsustav (erandiks on puhtalt ennustuslikud mudelid, mis on aga enamasti läbinähtamatu struktuuriga). Mudel on kas kasulik või kasutu; teooria on kas tõene või väär. Mudeli ja maailma vahel võib olla kaudne "peegeldus", aga mitte kunagi otsene side. Seega, ükski number, mis arvutatakse mudeli raames, ei kandu sama numbrina üle teaduslikku ega päris maailma. Ja kogu statistika (ka mitteparameetriline) toimub mudeli väikses maailmas. Arvud, mida statistika teile

pakub, elavad mudeli maailmas; samas kui teie teaduslik huvi on suunatud päris maailmale. Näiteks 95% usaldusintervall ei tähenda, et te peaksite olema 95% kindel, et tõde asub selles intervallis – sageli ei tohiks te seda nii julgelt tõlgendada isegi kitsas mudeli maailmas.

Näide: Aristoteles, Ptolemaios ja Kopernikus

Aristoteles lõi teooria maailma toimimise kohta, mis domineeris haritud Eurooplase maailmapilti enam kui 1200 aasta vältel. Selle kohaselt asub universumi keskpunktis maakera ning kõik, mida siin leida võib, on tehtud neljast elemendist: maa, vesi, õhk ja tuli. Samas, kogu maailmaruum alates kuu sfäärist on tehtud viiendast elemendist (eeter), mida aga ei leidu maal (nagu nelja elementi ei leidu kuu peal ja sealt edasi). Taevakehad (kuu, päike, planeedid ja kinnistähed) tiirlevad ümber maa kontentrilistes sfäärides, mis on omavahel seotud (mille vahel pole vaba ruumi). Seega on kogu liikumine eetri sfäärides ühtlane ja ringikujuline ja see liikumine põhjustab pika põhjus-tagajärg ahela kaudu kõiki liikumisi, mida maapeal kohtame. Kaasa arvatud meie sündimine, elukäik ja surm (mis on kõik liikumised). Kõik, mis maapeal huvitavat, ehk kogu liikumine, on algselt põhjustatud esimese liikumise poolt, mille käivitab kõige välimises sfääris paiknev meie jaoks mõistetamatu intellektiga "olend".

Aristotelese suur teooria ühendab kogu maailmapildi alates kaasaegses mõistes keemiast ja kosmoloogiast kuni bioloogia, maateaduse ja isegi geograafiani. Samas ühte selle olulist puudust nähti kohe. Nimelt ei suuda Aristoteles seletada, miks osad planeedid teavavõlvil vahest suunda muudavad ja mõnda aega lausa vastupidises suunas liiguvad (retrogression). Kuna astronoomia põhiline kasutusala oli astroloogia, siis põõrati planeetide liikumisele suurt tähelepanu. Lahenduseks ei olnud mitte suure teooria ümbertegemine või ümber lükkamine, vaid nõudlus uue teaduse järele, mis "päästaks fenomenid". Siin tuli appi Ptolemaios, kes lõi matemaatilise mudeli, kus planeedid mitte lihtsalt ei liigu ringtrajektoori mõõda vaid samal ajal teevad ka väiksemaid ringe ümber esimese suure ringi joone. Neid väiksemaid ringe kutsutakse epitsükliteks. See mudel suutis planeetide liikumist taevavõlvil piisavalt hästi ennustada, et astroloogide nõudlik seltskond sellega rahule jäi.

Ptolemaiosel ja tema järgijatel oli tegelikult mitu erinevat mudelit. Osad neist ei sialdanud epitsükleid ja maakera ei asunud tema mudelites universumi keskel, vaid oli sellest punktist eemale nihutatud — nii et päike ei teinud ringe ümber maakera vaid ümber tühja punkti. Oluline oli, et leidus epitsüklitega mudel ja ilma epitsükliteta mudel, mis olid matemaatiliselt ekvivalentsed ja andsid seega võrdseid ennustusi. Oli selge, et Aristotelese teooria ja fenomenide päästmise mudelid olid fundamentaalselt erinevad asjad. Samal ajal kui Aritoteles **seletas** maailma põhiolemust põhjuslike seoste jadana (mitte matemaatiliselt), **kirjeldas/ennustas** Ptolemaios sellesama maailma käitumist matemaatiliste (mitte põhjuslike) struktuuride abil.

Nii tekkis olukord, kus maailma mõistmiseks kasutati 1000 aasta vältel Aristotelese ühendteooriat aga selle kirjeldamiseks ja tuleviku ennustamiseks hoopis Ptolemailisi mudeleid, mida keegi "päriselt" tõeks ei pidanud ja mida hinnati selle järgi, kui hästi need "päästsid fenomene".

See toob meid Koperniku juurde, kes teadusajaloolaste arvates vallandas 17. sajandi teadusliku revolutsiooni avaldades raamatu, kus ta asetab päikse universumi keskele ja paneb maa selle ümber ringtrajektooril tiirlema. Kas Kopernikus tõrjus sellega kõrvale Aristotelese, Ptolemaiose või mõlemad? Kaasaegne seisukoht on, et kuigi Kopernikus soovis teha kolmandat, arvasid tema rängalt matemaatilise teose avaldamisele järgnenud 40 aasta vältel pea kõik asjatundlikud astronoomid, et ta soovis välja pakkuda vaid lihtsama alternatiivi epitsüklitega mudelile, mis selleks ajaks oli muutunud väga keerukaks (aga ka samavõrra ennustustäpseks). Kuna Kopernikuse raamat läks trükki ajal, mil selle autor oli juba surivoodil, kirjutas sellele eessõna üks tema vaimulikust sõber, kes püüdis oodatavat kiriklikku pahameeletormi leevendada väitega, et päikese keskele viimine ei ole muud kui mudeldamise trikk, millest ei tasu järeldada, et maakera ka tegelikult ümber päikese tiirleb (piibel nimelt räägib sellest, kuidas jumal peatas mõneks ajaks päikese taevavõlvil, mitte maa). Ja kuna eessõna oli anonüümne, eeldasid lugejad muidugi, et selle kirjutas autor. Lisaks, kuigi Kopernikus tõstis päikese keskele, jäi ta ringikujuliste trajektooride juurde, mis tähendas, et selleks, et tema mudel fenomenide päästmisel hätta ei jääks ja astroloogidele kasutu ei oleks, oli ta sunnitud maad ja planeete liigutana ümber päikese mõõda epitsükleid. Kokkuvõttes oli Koperniku mudel sama keeruline kui Ptolemaiose standardmudel

(neis oli võrdne arv epitsükleid) ja selle abil tehtud ennustused planeetide liikumise kohta tulid väiksema täpsusega.

> Mudelina seisnes Koperniku eelis selles, et tema mudel suutis ennustada mõningaid nähtusi (planeetide näiva heleduse muutumine, mis jõuab maksimumi nende lähimas asukohas maale), mida Ptolemaiose mudel ei ennustanud. See ei tähenda, et need fenomenid oleksid olnud vastuolus Ptolemaiose mudeliga. Lihtsalt, nende Ptolemaiose mudelisse sobitamiseks oli vaja osad mudeli parameetrid fikseerida nii-öelda suvalistele väärtustele. Seega Koperniku mudel töötas nii, nagu see oli, samas kui Ptolemaiose mudel vajas ad hoc tuunimimst.

Samas, kui vaadata Koperniku produkti teooriana, mitte mudelina, siis oli sel selgeid eeliseid Aristotelese ees. Oli nähtud komeete üle taevavõlvi lendamas (mis Aristotelese järgi asusid kinnistähtede muutumatus sfääris) ja Galileo joonistas oma teleskoobist kraatreid kuu pinnal, mis näitas, et kuu ei saanud koosneda täiuslikust viiendast elemendist ja sellel toimusid ilmselt sarnased füüsikalised protsessid kui maal. On usutav, et kui Kopernikus oleks oma raamatule jõudnud ise essõna kirjutada oleks tema teooria vastuvõtt olnud palju kiirem (ja valulisem). Seega, teooria ja mudeli eristus on tähtis!

> Koperniku teooriast tuleneb loogilise paratamatusena, et tähtedel esineb maapealt vaadates parallaks. See tähendab, et kui maakera koos astronoomiga teeb poolringi ümber päikese, siis kinnistähe näiv asukoht taevavõlvil muutub sest astronoom vaatleb teda teise nurga alt. Pange oma nimetissõrm näost u 10 cm kaugusele, sulgege parem silm, seejärel avage see ning sulgege vasak silm ja te näete oma sõrme parallaksi selle näiva asukoha muutusena. Tähtede parallaksi püüti mõõta juba Aleksandrias 1000 aastat enne Kopernikust, et leida kinnitust teooriale, mille kohaselt maakera tiirleb ümber päikese. Mõõtmised ei näidanud aga parallaksi olemasolu (sest maa trajektoori diameeter on palju lühem kui maa kaugus tähtedest). Parallaksi olemasolu sai kinnitust alles 19. sajandi teisel poolel, siis kui juba ammu iga koolijüts uskus, et maakera tiirleb ümber päikese!

Millest koosneb mudel?

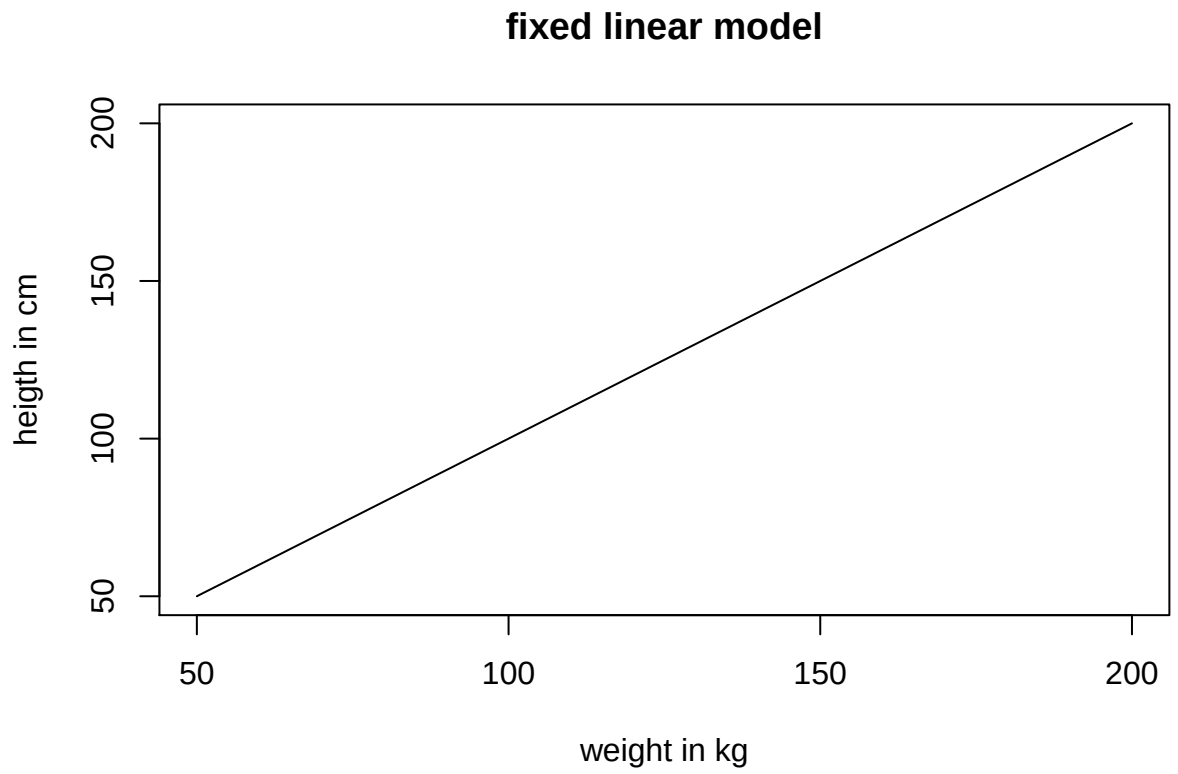
Mudel on matemaatilise formalism, mis püüab kirjeldada füüsikalist protsessi.

Näiteks, sageli kirjeldame me produkti kuhjumist ensüümreaktsioonis eksponentsiaalse funktsioni (mudeli) abil. Kui meie andmed seda tüüpi funktiooniga sobivad, ütleb see meile midagi konkreetse ensüümi töömehhanismi kohta. Teisest küljest, need mudelid, mis on "generatiivsed", suudavad lisaks simuleerida ka uusi andmeid. Sealhulgas ka selliseid, mida päris maailmas ei saa kunagi esineda sest seal puuduvad vastavad tingimused. Mudelisse saab aga sisse kirjutada igasuguseid tingimusi ehk parameetri väärtusi (näit substraadi konsentratsioone, mida me ei suuda "päriselt" saavutada).

Oletame, et me mõõtsime N inimese pikkuse cm-s ja kaalu kg-s ning meid huvitab, kuidas inimeste pikkus sõltub nende kaalust.

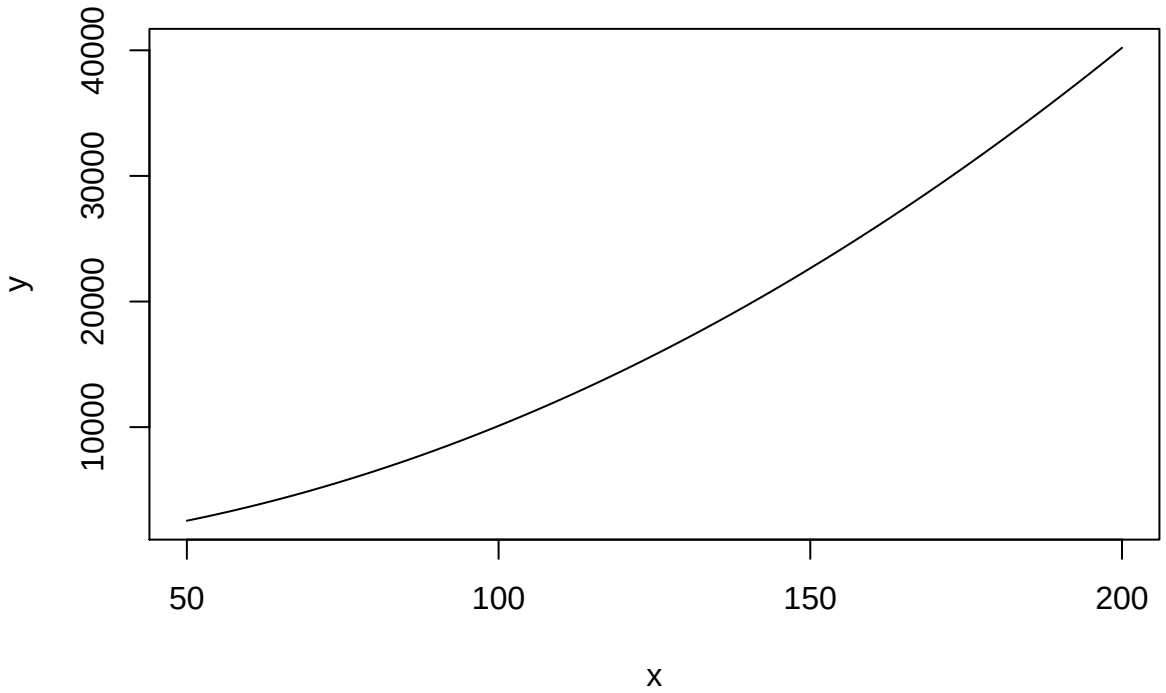
Lihtsaim mudel pikkuse sõltuvusest kaalust on pikkus = kaal (formaliseeritult: y = x) ja see mudel ennustab, et kui Johni kaal = 80 kg, siis John on 80 cm pikkune. Selle mudeli saame graafiliselt kujutada nii

```
x <- 50:200 #y = kaal
y <- x # x = pikkus
plot(y~x, type="l", xlab="weight in kg", ylab="heigth in cm", main="fixed linear model")
```



kõigepealt painutame sirget. See joon on ikka veel täielikult fikseeritud, aga ta pole enam sirge (ehki tehniliselt on meil ikka lineaarne seos x ja y vahel)

```
x <- 50:200
y <- x + x**2
plot(x, y, type="l")
```



Mudeli keeles tähistame me seda, mida me ennustame (antud juhul pikkus) Y-ga ja seda, mille väärtuse

põhjal me ennustame (antud juhul kaal) X-ga. Seega sirge mudeli matemaatiline formalism on Y = X. See on äärmiselt jäik mudel: ta on sirge kujuline ja selle sirge asukoht parameetriruumis on rangelt fikseeritud. Sirge lõikab y telge alati 0-s (ehk mudeli keeles: selle sirge intercept ehk lõikepunkt Y teljel = 0) ja tema tõusunurk saab olla ainult 45 kraadi (mudeli keeles: mudeli slope ehk tõus = 1). Selle mudeli jäikus tuleneb sellest, et selles mudelis ei ole parameetreid, mida me saaksime vabalt muuta ehk tuunida.

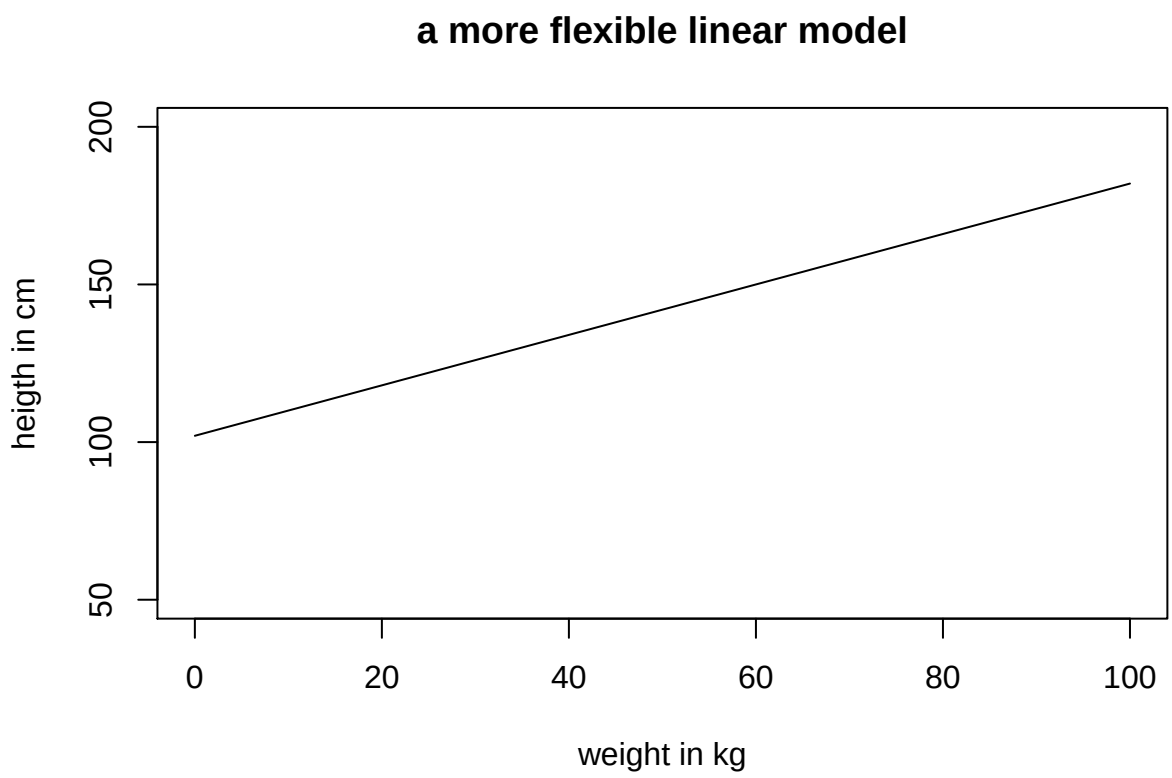
Kuidas aga kirjeldada sirget, mis võib paikneda 2-mõõtmelises ruumis ükskõik millises asendis? Selleks lisame mudelisse kaks parameetrit, intercept (a) ja tõus (b). Kui a=0 ja b=0, saame me eelpool kirjeldatud mudeli y = x. Kui a = 102, siis sirge lõikab y telge väärtusel 102. Kui b = 0.8, siis x-i tõustes 1 ühiku võrra tõuseb y-i väärtus 0.8 ühiku võrra. Kui a = 100 ja b = 0, siis saame sirge, mis on paraleelne x-teljega ja lõikab y telge väärtusel 100 (mis juhtub, kui a = Inf?). Seega, Teades a ja b väärtusi ning omistades x-le suvalise meid huvitava väärtuse, saab ennustada y-i keskmist väärtust. Näiteks, olgu andmete vastu fititud mudel:

pikkus(cm) = 102 + 0.8 * kaal(kg) ehk

y = 102 + 0.8x.

Omistades nüüd kaalule väärtuse 80 kg, tuleb mudeli poolt ennustatud keskmine pikkus 102 + 0.8 * 80 = 166 cm. Iga kg lisakaalu ennustab mudeli kohaselt 0.8 cm võrra suuremat pikkust.

```
a <- 102
b <- 0.8
x <-  0:100
y <-  a + b * x
plot(y~x, type="l", xlab="weight in kg", ylab="heigth in cm", main="a more flexible linear model", ylim=
```



See mudel ennustab, et 0 kaalu juures on pikku 102 cm, mis on rumal, aga mudelite puhul tavaline, olukord. Me tuunime mudelit andmete peal, mis ei sisalda 0-kaalu (sest 0-kaaluga inimesi pole olemas). Meie valimiandmed ei peegelda täpselt inimpopulatsiooni. Sirge mudel ei peegelda täpselt pikkuse-kaalu suhteid vahemikus, kus meil on reaalseid kaaluandmeid; ja ta teeb seda veelgi vähem seal, kus meil mõõdetud kaalusid ei ole. Seega pole mõtet imestada, miks mudeli intercept meie üle irvitab.

4 mõistet

X ja Y on muutujad, a ja b on parameetrid. Muutujate väärtused fikseeritakse andmete poolt, parameetrid fititakse muutujate väärtuste põhjal. Fititud mudel ennustab igale X-i väärtusele vastava kõige tõenäolisema Y väärtuse (Y keskväärtuse sellel X-i väärtusel).

Y - mida me ennustame (dependent variable, predicted varable)

X - mille põhjal me ennustame (independent variable, predictor)

muutuja (variable) - iga asi, mida me valimis mõõdame (X ja Y on kaks muutujat). Muutuja väärtused on fikseeritud andmete poolt. Muutujal on sama palju fikseeritud väärtusi kui meil on selle muutuja kohta mõõtmisandmeid.

parameeter (parameter) - mudeli koefitsient, millele võib omistada suvalisi väärtusi. Parameetreid tuunides fitime me mudeli võimalikult hästi sobituma andmetega.

Mudeli fittimine

Mudelid sisaldavad (1) matemaatilisi struktuure, mis määravad mudeli tüübi ning (2) parameetreid, mida saab andmete põhjal tuunida, niiviisi täpsustades mudeli kuju.

Seda tuunimist nimetatakse mudeli fittimiseks. Mudelit fittides on eesmärk saavutada antud tüüpi mudeli maksimaalne sobivus andmetega. Näiteks võrrand y = a + bx määrab mudeli, kus y = x on on see struktuur, mis tagab, et mudeli tüüp on sirge, ning a ja b on parameetrid, mis määravad sirge asendi. Seevastu struktuur y = x + xˆ2 tagab, et mudeli y = a + b1x + b2xˆ2 tüüp on parabool, ning parameetrite a, b1 ja b2 väärtused määravad selle parabooli täpse kuju. Ja nii edasi

Hea mudel on

(1) võimalikult lihtsa struktuuriga, mille põhjal on veel võimalik teha järeldusi protsessi kohta, mis genereeris mudeli valmistamiseks kasutatud andmeid;

(2) sobitub piisavalt hästi andmetega (eriti uute andmetega, mida ei kasutatud selle mudeli fittimiseks), et olla relevantne andmeid genereeriva protsessi kirjeldus;

(3) genereerib usutavaid simuleeritud andmeid (see näitab mudeli kvaliteeti).

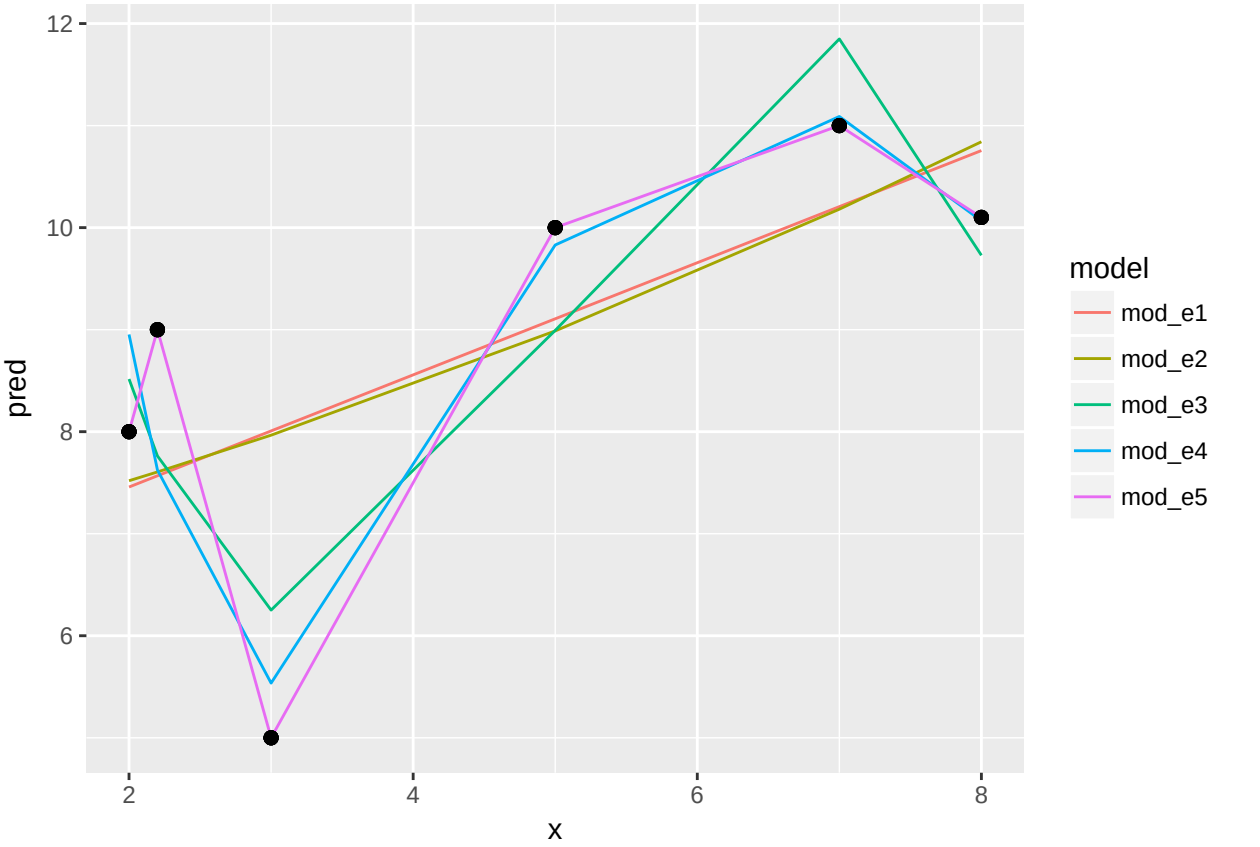
Sageli fititkse samade andmetega mitu erinevat tüüpi mudelit ja püütakse otsustada, milline neist vastab kõige paremini eeltoodud tingimustele. Näiteks, kui sirge suudab kaalu järgi pikkust ennustada paremini kui parabool, siis on sirge mudel kooskõlas teadusliku hüpoteesiga, mis annaks mehhanismi protsessile, mille käigus kilode lisandumine viiks laias kaaluvahemikus inimeste pikkuse kasvule ilma, et pikkuse kasvu tempo langeks.

Üle- ja alafittimine

Osad mudelite tüübid on vähem paindlikud kui teised (parameetreid tuunides on neil vähem liikumisruumi). Kuigi sellised mudelid sobituvad halvemini andmetega, võivad need ikkagi paremini kui mõni paindlikum mudel välja tuua andmete peidetud olemuse. Mudeldamine eeldab, et me usume, et meie andmetes leidub nii müra (mida mudel võiks ignoreerida), kui signaal (mida mudel püüab tabada). Kuna mudeli jaoks näeb müra samamoodi välja kui signaal, on iga mudel kompromiss üle- ja alafittimise vahel. Me lihtsalt loodame, et meie mudel on piisavalt jäik, et mitte liiga palju müra modelleerida ja samas piisavalt paindlik, et piisaval määral signaali tabada.

Üks kõige jäigemaid mudeleid on sirge, mis tähendab, et sirge mudel on suure tõenäosusega alafittitud. Keera sirget kuipalju tahad, ikka ei sobitu ta enamiku andmekogudega. Ja need vähesed andmekogud, mis sirge mudeliga sobivad, on genereeritud teatud tüüpi lineaarsete protsesside poolt. Sirge on seega üks kõige paremini tõlgendatavaid mudeleid. Teises äärmuses on polünoomsed mudelid, mis on väga paindlikud, mida on väga raske tõlgendada ja mille puhul on suur mudeli ülefittimise oht. Ülefititud mudel järgib nii

täpselt valimiandmeid, et sobitub hästi valimis leiduva juhusliku müraga ning seetõttu sobitub halvasti järgmise valimiga samast populatsioonist (sest igal valimil on oma juhuslik müra). Üldiselt, mida rohkem on mudelis tuunitavaid parameetreid, seda paindlikum mudel, seda kergem on seda valimiandmetega sobitada ja seda raskem on seda mudelit tõlgendada. Veelgi enam, alati on võimalik konstrueerida mudel, mis sobitub täiuslikult lõpliku arvu andmepunktidega (selle mudeli parameetrite arv = N). Selline mudel on täpselt sama informatiivne kui andmed, mille põhjal see fititi — ja täiesti kasutu.



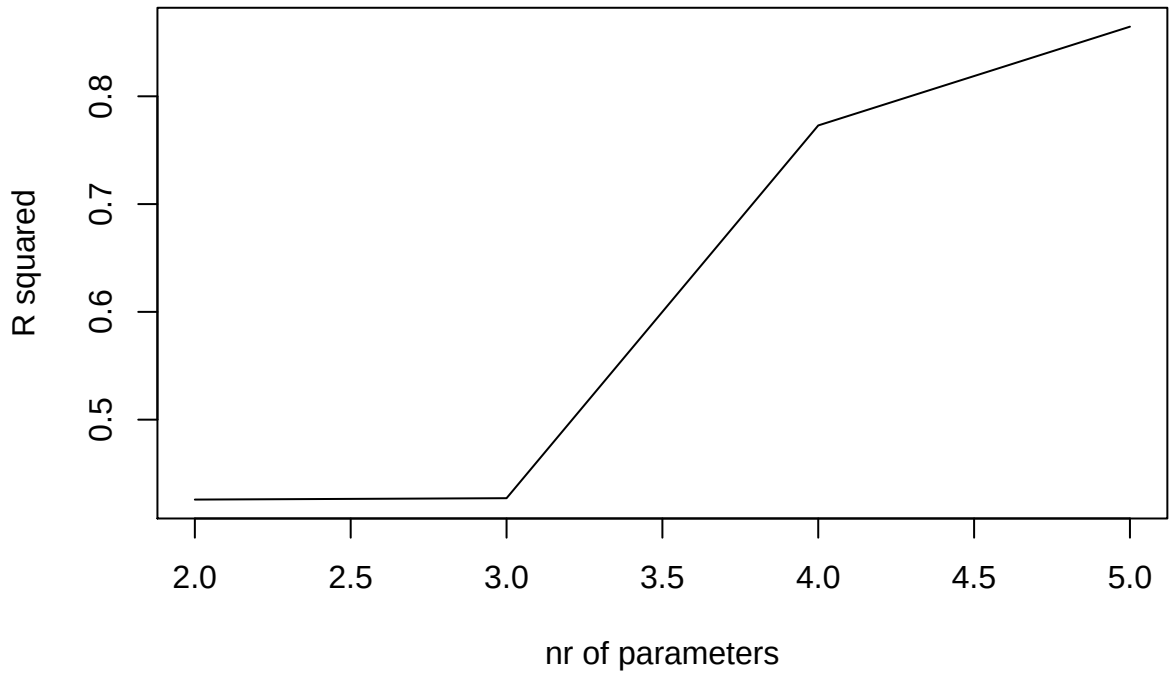
Joonis: Kasvava paindlikusega polünoomsed mudelid. mod_e1 on sirge võrrand y = a + b1x (2 parameetrit: a ja b1), mod_e2 on lihtsaim võimalik polünoom: y= a + b1x + b2x^2 (3 parameetrit), . . . , mod_e5: y= a + b1x + b2x^2 + b3x^3 + b4x^4 + b5x^5 (6 parameetrit). mod_e5 vastab täpselt andmepunktidele (N = 6).

```
AIC(mod_e1, mod_e2, mod_e3, mod_e4, mod_e5)
```

```
##        df      AIC
## mod_e1  3 27.77993
## mod_e2  4 29.76669
## mod_e3  5 26.21330
## mod_e4  6 25.11245
## mod_e5  7     -Inf
```

AIC näitab, et parim mudel on mod_e4. Aga kas see on ka kõige kasulikum mudel? Mis siis, kui 3-s andmepunkt on andmesisestaja näpuviga?

```
AIC - Akaike Informatsiooni Kriteerium - vaatab mudeli sobivust andmetega
ja mudeli parameetrite arvu.
Väikseim AIC tähitab parimat fitti väikseima parameetrite arvu juures
(kompromissi) ja väikseima AIC-ga mudel on eelistatuim mudel. Aga seda ainult
võrreldud mudelite hulgas. AIC-i absoluutväärtus ei loe - see on suhteline näitaja.
```



Joonis: sedamööda kuidas parameetrite arv mudelis kasvab, kasvab ka R ruut. R ruut 0.8 tähendab, et x-i varieeruvus suudab seletada kuni 80% y-i varieeruvusest. Lisaparameetri lisamine ei saa põhimõtteliselt R ruutu vähendada. Aga selle kasvu kiirus on aeglustuv. Ühel hetkel ei õigusta mudeli fiti paranemine enam mudeli paindlikuse kasvu (mis mõlemad saavutatakse parameetreid lisades).

```
Ülefittimise vältimiseks kasutavad Bayesi mudelid informatiivseid prioreid,
mis välistavad ekstreemsed parameetriväärtused.
Vt http://elevanth.org/blog/2017/08/22/there-is-always-prior-information/
```

Veamudel

Eelpool kirjeldatud mudelid on deterministlikud — nad ei sisalda hinnangut andmete varieeruvusele ennustuse ümber. Neid kutsutakse ka **protsessi mudeliteks** sest nad modelleerivad protsessi täpselt. Ehk kui mudel ennustab, et 80 kg inimene on 166 cm pikkune, siis protsessi mudel ei ütle, kui suurt kaalust sõltumatut pikkuste varieeruvust võime oodata 80 kg-ste inimeste hulgas? Selle hinnangu andmiseks tuleb mudelile lisada veel üks komponent, **veamudel** ehk veakomponent, mis sageli tuuakse sisse normaaljaotuse kujul. Veakomponent modelleerib üksikute inimeste pikkuste varieeruvust (mitte keskmise pikkuse varieeruvust) igal mõeldaval ja mittemõeldaval kaalul. Tänu sellele ei ole mudeli ennustused enam deterministlikud, vaid tõenäosuslikud.

Kuidas veakomponent lineaarsesse mudelisse sisse tuua?

ilma veakomponendita mudel: $y = a + bx$

Veakomponent tähendab, et y-i väärtus varieerub ümber mudeli poolt ennustatud keskväärtuse ja seda varieeruvust normaaljaotusega modelleerides saame

$y \sim dnorm(mu, sigma)$

kus mu on mudeli poolt ennustatud keskväärtus ja $sigma$ on mudeli poolt ennustatud standardhälve ehk varieeruvus andmepunktide tasemel. Tilde ~ tähistab seose tõenäosuslikkust.

Sirge mudelisse varieeruvuse sisse toomiseks defineerime mu ümber nõnda:

$mu = a + bx$, mis tähendab, et

$y \sim dnorm(a + bx, sigma)$

See ongi sirge mudel koos veakomponendiga. Peatükis 3 õpime me selliste mudelitega töötama.

Kõik statistilised mudelid on tõenäosusmudelid ning sisaldavad veakomponenti.

Statistiline mudel koosneb 3 komponendist:

```
> (1) matemaatiline struktuur, mis sisaldab muutujaid ja annab mudeli tüübi,

> (2) tuunitavad parameetrid ja

> (3) veamudel.
```

Muide, kõik veamudelid, millega me edaspidi töötame, modelleerivad igale x-i väärtusele (kaalule) sama suure y-i suunalise varieeruvuse (pikkuste sd). Suurem osa statistikast kasutab eeldusi, mida keegi päriselt tõe pähe ei võta, aga millega on arvutuslikus mõttes lihtsam elada.

Enimkasutatud veamudel on normaaljaotus.

Oletame, et meil on kolm andmepunkti ning me usume, et need andmed on juhuslikult tõmmatud normaaljaotusest või sellele lähedasest jaotusest. Normaaljaotuse mudelit kasutades me sisuliselt deklareerime, et me usume, et kui me oleksime olnud vähem laisad ja 3 mõõtmise asemel sooritanuks 3000, siis need mõõtmised sobituksid piisavalt hästi meie 3 väärtuse peal fititud normaaljaotusega. Seega, me usume, et omades 3 andmepunkti me teame juba umbkaudu, millised tulemused me oleksime saanud korjates näiteks 3 miljonit andmepunkti. Oma mudelist võime simuleerida ükskõik kui palju andmepunkte.

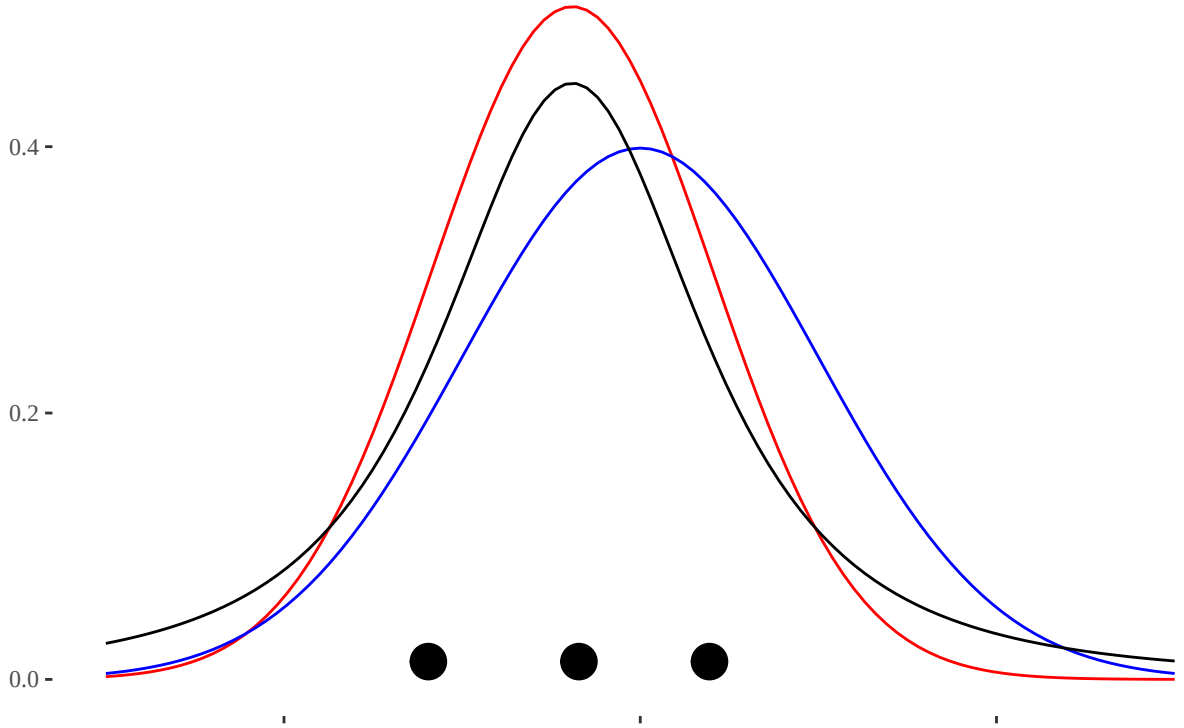
Aga pidage meeles, et selle mudeli fittimiseks kasutame me ainult neid andmeid, mis meil päriselt on — ja kui meil on ainult 3 andmepunkti, on tõenäoline, et fititud mudel ei kajasta hästi tegelikkust.

Halvad andmed ei anna kunagi head tulemust.

Eelnev ei kehti Bayesi mudelite kohta, mis toovad priorite kaudu sisse lisainfot, mis ei kajastu valimiandmetes ja võib analüüsi päästa.

Kuidas panna skeptik uskuma, et statistilised meetodid töötavad halvasti väikestel valimitel. Siin aitab simulatsioon, kus me tõmbame 3-se valimi etteantud populatsioonist ning üritame selle valimi põhjal ennustada selleasama populatsiooni struktuuri. Kuna tegemist on simulatsiooniga, teame täpselt, et populatsioon, kust me tõmbame oma kolmese valimi, on normaaljaotusega, et tema keskväärtus = 0 ja et tema sd = 1. Me fitime oma valimi andmetega 2 erinevat mudelit: normaaljaotuse ja Studenti t jaotuse.

```
## Loading required package: Rcpp

## Loading 'brms' package (version 1.8.0). Useful instructions
## can be found by typing help('brms'). A more detailed introduction
## to the package is available through vignette('brms_overview').

## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```



Joonis: juhuvalim normaaljaotusest, mille keskmine=0 ja sd=1 (n=3; andmepunktid on näidatud mustade munadena). Sinine joon - popualtsioon, millest tõmmati valim; punane joon - normaaljaotuse mudel, mis on fititud valimi andmetel; must joon - Studenti t jaotuse mudel, mis on fititud samade andmetega.

Mõlemad mudelid on süstemaatiliselt nihutatud väiksemate väärtuste poole ja alahindavad varieeruvust. t jaotuse mudel on oodatult paksemate sabadega ja ennustab 0-st kaugele palju rohkem väärtusi kui normaaljaotuse mudel. Kuna me teame, et populatsioon on normaaljaotusega, pole väga üllatav, et t jaotus modeleerib seda halvemini kui normaaljaotus.

Igal juhul, mõni teine juhuvalim annaks meile hoopis teistsugused mudelid, mis rohkem või vähem erinevad algsest populatsioonist.

Mis juhtub kui me kasutame oma normaaljaotuse mudelit uute andmete simuleerimiseks? Kui lähedased on need simuleeritud andmed populatsiooni andmetega ja kui lähedased valimi andmetega, millega me normaaljaotuse mudeli fittisime?

```
set.seed(19) #muudab simulatsiooni korratavaks
#tõmbame 3 juhuslikku arvu normaalhaotusest, mille keskväärtus = 0 ja sd = 1.
df <- tibble(sample_data=rnorm(3))
#fitime normaaljaotuse mudeli valimi keskmise ja sd-ga
mean(df$sample_data); sd(df$sample_data)
```

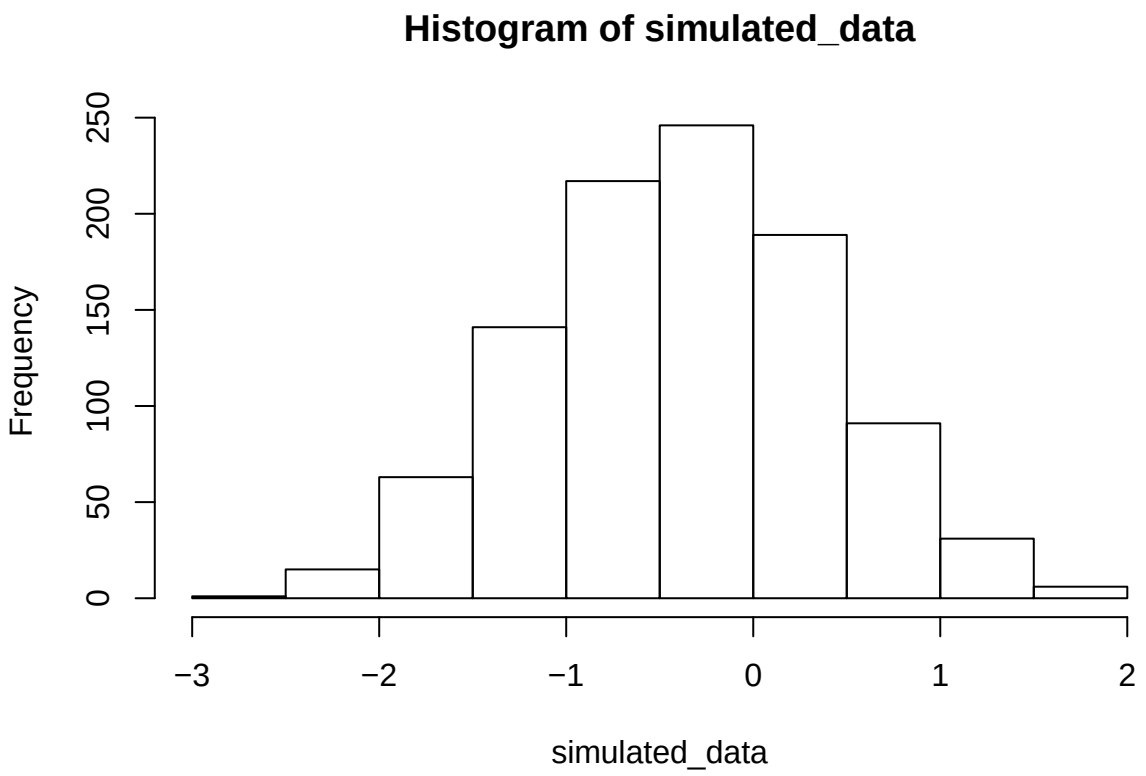
```
## [1] -0.3817353
```

```
## [1] 0.7896821
```

```
#simuleerime 1000 uut andmepunkti fititud mudelist
simulated_data <- rnorm(1000, mean(df$sample_data), sd(df$sample_data))
#arvutame simuleeritud andmete keskmise ja sd ning joonistame neist histogrammi
mean(simulated_data); sd(simulated_data); hist(simulated_data)
```

```
## [1] -0.3848133
```

```
## [1] 0.7749198
```



Nagu näha, on uute (simuleeritud) andmete keskväärtus ja SD väga sarnased algsete andmete omale, mida kasutasime mudeli fittimisel. Kahjuks ei ole need aga kaugeltki nii sarnased algsele jaotusele, mille kuju me püüame oma andmete ja mudeli pealt ennustada. Seega on meie mudel üle-fittitud, mis tähendab, et ta kajastab liigselt neid valimi aspekte, mis ei peegelda algse populatsiooni omadusi. Loomulikult ei vasta ükski mudel päriselt tegelikkusele. Küsimus on pigem selles, kas mõni meie mudelitest on piisavalt hea, et olla kasulik. Vastus sellele sõltub, milleks plaanime oma mudelit kasutada.

```
mean(simulated_data>0); mean(simulated_data>1)
```

```
## [1] 0.317
```

```
## [1] 0.037
```

Kui populatsioonivääртustest on 50% suuremad kui 0, siis mudeli järgi vaevalt 32%. Kui populatsioonivääртustest on 16% suuremad kui 1, siis mudeli järgi vaevalt 4%. See illustreerib hästi mudeli kvaliteeti.

```
library(brms)
sim_t <- rstudent_t(1000, 2, mean(df$sample_data), sd(df$sample_data))
mean(sim_t>0); mean(sim_t>1)
```

```
## [1] 0.338
```

```
## [1] 0.11
```

Samad ennustused t jaotusest on isegi paremad! Aga kumb on ikkagi parem mudel populatsioonile?

normaaljaotuse ja lognormaaljaotuse erilisus

Normaaljaotus ja lognormaaljaotus on erilised sest

(1) keskne piirteoreem ütleb, et olgu teie valim ükskõik millise jaotusega, paljudest valimitest arvutatud **aritmeetilised keskmised** on alati enam-vähem normaaljaotusega (kui n>30). Selle matemaatilise formalismi tuletus füüsikalisse maailma on nn “elementaarsete vigade hüpotees”, mille kohaselt paljude

väikeste üksteisest sõltumatute juhuslike efektide (vigade) summa annab tulemuseks normaaljaotuse. Paraku annavad enamus bioloogilisi mõõtmisi eranditult mitte-negatiivseid tulemusi. Sageli on selliste mõõtmiste tulemuste jaotused ebasümmeetrilised (v.a. siis, kui cv = sd/mean on väike) ja siis on meil sageli tegu lognormaaljaotusega, mis tekkib log-normaalsete muutujate korrutamisest (mitte liitmisest, nagu normaaljaotuse puhul). Keskne piirteoreem 2: suvalise jaotusega muutujate **geomeetrilised keskmised** on lognormaaljaotusega. Elementaarsete vigade hüpotees 2: Kui juhuslik varieeruvus tekib paljude juhuslike efektide korrutamisel, on tulemuseks lognormaaljaotus. Lognormaaljaotuse elementide (arvude) logaritmimisel saame normaaljaotuse.

(2) Mõlemad jaotused (normaal ja lognormaal) on maksimaalse entroopiaga jaotused. Entroopiat vaadeldakse siin informatsiooni/müra kaudu — maksimaalse entroopiaga süsteem sisaldab maksimaalselt müra ja minimaalselt informatsiooni (Shannoni informatsiooniteooria). See tähendab, et väljaspool oma parameetrite tuunitud väärtusi on need normaal- ja lognormaaljaotused minimaalselt informatiivsed. Näiteks normaaljaotusel on kaks parameetrit, mu ja sigma (ehk keskmine ja standardhälve). Seega, andes normaaljaotusele ette keskväärtuse ja standardhälbe fikseerime üheselt jaotuse ehk mudeli kuju ja samas lisame sinna minimaalselt muud (sooviamtut) informatsiooni. Teised maksimaalse entroopiaga jaotused on eksponentsiaalne jaotus, binoomjaotus ja poissoni jaotus. Maksimaalse entroopiaga jaotused sobivad hästi Bayesi prioriteks sest me suudame paremini kontrollida, millist informatsiooni me neisse surume.

Küsimused, mida statistika küsib

Statistika abil saab vastuseid järgmisetele küsimustele:

1) kuidas näevad välja teie andmed ehk milline on just teie andmete jaotus, keskväärtus, varieeruvus ja koos-varieeruvus? Näiteks, mõõdetud pikkuste ja kaalude koos-varieeruvust saab mõõta korrelatsioonikordaja abil.

2) mida me peaksime teie valimi andmete põhjal uskuma populatsiooni parameetri tegeliku väärtuse kohta? Näiteks, kui meie andmete põhjal arvutatud keskmine pikkus on 178 cm, siis kui palju on meil põhjust arvata, et tegelik populatsiooni keskmine pikkus > 185 cm?

3) mida ütleb statistilise mudeli struktuur teadusliku hüpoteesi kohta? Näiteks, kui meie poolt mõõdetud pikkuste ja kaalude koos-varieeruvust saab hästi kirjeldada kindlat tüüpi lineaarse regressioonimudeliga, siis on meil ehk tõendusmaterjali, et pikkus ja kaal on omavahel sellisel viisil seotud ja eelistatud peaks olema teaduslik teooria, mis just sellise seose tekkimisele bioloogilise mehhanismi annab.

4) mida ennustab mudel tuleviku kohta? Näiteks, meie lineaarne pikkuse-kaalu mudel suudab ennustada tulevikus kogutavaid pikkuse andmeid. Aga kui hästi?

   statistika ülesanne on lähtuvalt piiratud hulgast andmetest ja mudelitest kvantifitseerida parimal võimalikul viisil kõhedust, mida peaksime tundma vastates eeltoodud küsimustele.

Statistika ei vasta otse teaduslikele küsimustele ega küsimustele päris maailma kohta. Statistilised vastused jäävad alati kasutatud andmete ja mudelite piiridesse. Sellega seoses peaksime eelistama hästi kogutud rikkalikke andmeid ja paindlikke mudeleid. Siis on lootust, et hüpe mudeli koefitsientidest päris maailma kirjeldamisse tuleb üle kitsama kuristiku. Bayesil on siin eelis, sest osav statistik suudab koostöös teadlastega priori mudelisse küllalt palju kasulikku infot koguda. Samas, amatöör suudab bayesi abil samavõrra käkki keerata. Mida paindlikum on meetod, seda vähem automaatne on selle mõistlik kasutamine.

2 osa. Kuidas näevad välja teie andmed?

summaarsed statistikud

Summaarne statistik = üks number.
Milliseid statistikuid arvutada ja milliseid vältida, sõltub statistilisest mudelist

> summaarse statistika abil iseloomustame a) tüüpilist valimi liiget (keskmist), b) muutuja sisest varieeruvust, c) erinevate muutujate (pikkus, kaal vms) koos-varieeruvust

keskväärtused

Keskväärtust saab mõõta paaril tosinal erineval viisil, millest järgnevalt kasutame kolme või nelja. Enne kui te arvutama kukute, mõelge järele, miks te soovite keskväärtust teada. Kas teid huvitab valimi tüüpiline liige? Kuidas te sooviksite seda tüüpilisust defineerida? Kas valimi keskmise liikmena või valimi kõige arvukama liikmena? või veel kuidagi? See, millist keskväärtust kasutada sõltub sageli andmejaotuse kujust. Sümmeetrilisi jaotusi on lihtsam iseloomustada ja mitmetipulised jaotused on selles osas kõige kehvemad.

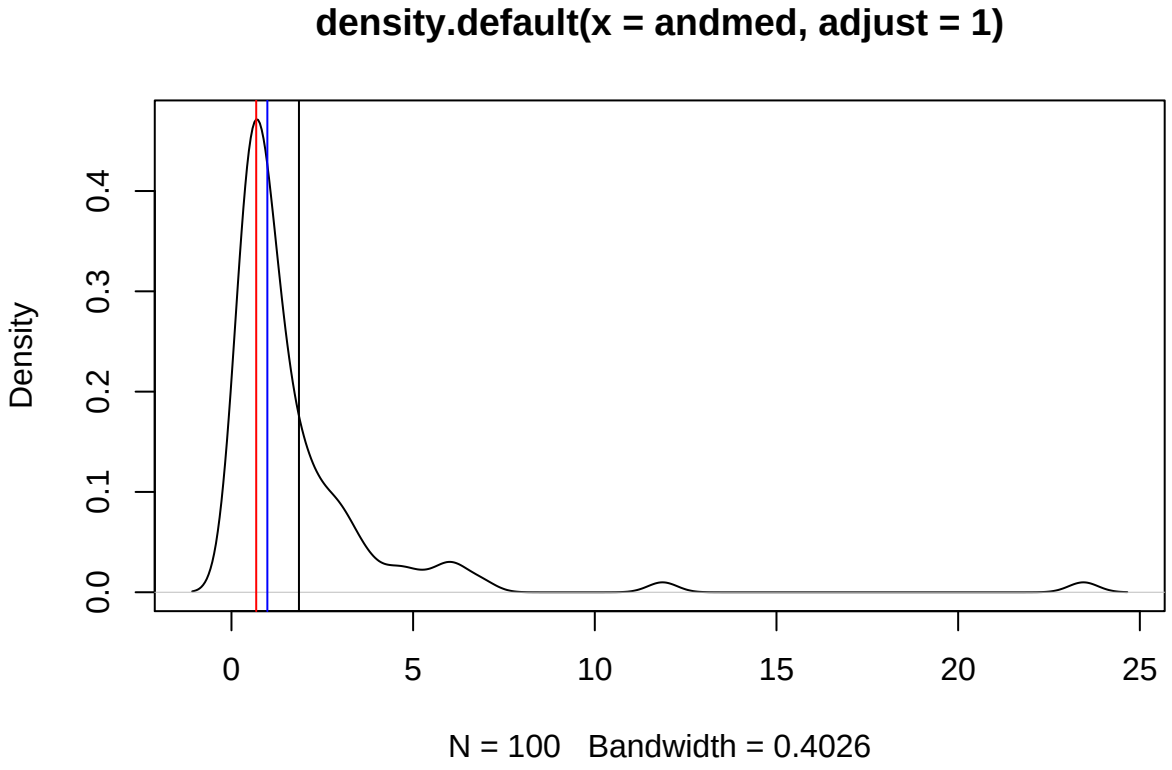
Mina eelistan selliseid nõuandeid (mis on rangelt soovituslikud):

(1) Kui valim on normaaljaotusega (histogramm on sümmeetriline), hinda tüüpilist liiget läbi aritmeetilise keskmise (mean).

(2) Muidu kasuta mediaani (median). Kui valim on liiga väike, et jaotust hinnata (aga > 4), eelista mediaani. Mediaani saamiseks järjestatakse mõõdetud väärtused suuruse järgi ja võetakse selle rea keskmine liige. Mediaan on vähem tundlik ekstreemsete väärtuste (outlierite) suhtes kui mean.

(3) Valimi kõige levinumat esindajat iseloomustab mood ehk jaotuse tipp. Seda on aga raskem täpselt määrata ja mitmetipulisel jaotusel on mitu moodi. Töötamisel posterioorsete jaotustega on mood sageli parim lahendus.

```
## [1] 0.6817168
```



Joonis: Simuleeritud lognormaaljaotusega andmed. Punane joon - mood; sinine joon - mediaan; must joon - aritmeetiline keskmine (mean). Milline neist vastab parimini teie intuitsiooniga nende andmete "keskväärtusest"? Miks?

muutuja sisene varieeruvus

Mean-iga käib kokku standardhälve (SD).

SD on sama ühikuga, mis andmed (ja andmete keskväärtus). Statistikute hulgas eelistatud formaat on mean (SD), mitte mean (+/- SD). 1 SD katab 68% normaaljaotusest, 2 SD – 96% ja 3 SD – 99%. Normaaljaotus langeb servades kiiresti, mis tähendab, et tal on peenikesed sabad ja näiteks 5 SD kaugusel keskmisest paikneb vaid üks punkt miljonist.

Näiteks: inimeste IQ on normaaljaotusega, mean=100, sd=15. See tähendab, et kui sinu IQ=115 (ülikooli astujate keskmine IQ), siis on tõenäosus, et juhuslikult kohatud inimene on sinust nutikam, 18% ((100% - 68%)/2 = 18%).

Kui aga "tegelikul" andmejaotusel on "paks saba" või esinevad outlierid, siis normaaljaotust eeldav mudel tagab ülehinnatud SD ja seega ülehinnatud varieeruvuse. Kui andmed saavad olla ainult positiivsed, siis SD > mean/2 viitab, et andmed ei sobi normaaljaotuse mudeliga (sest mudel ennustab negatiivsete andmete esinemist küllalt suure sagedusega).

```
Normaaljaotus on defineeritud ka mõnede teiste jaotuste jaoks peale normaaljaotuse
(Poissioni jaotus, binoomjaotus).
Funktsioon sd() ja selle taga olev valem, mis on loodud normaaljaotuse tarbeks,
ja neid alternatiivseid standardhälbeid ei arvuta.
Seega tasub meeles pidada, et tavapärane sd kehtib normaaljaotuse mudeli piirides ja ei kusagil mujal!
```

Kui andmed ei sobi normaaljaotusesse siis võib pakkuda kahte alternatiivset lahendust:

(1) logaritmi andmed.

Kui logaritmimine muudab andmed normaalseks, siis saab logaritmitud andmetest arvutada mean-i ja SD ja seejärel mõlemad anti-logaritmida. Sellisel juhul avaldad sa lõpuks geomeetrilise keskmise ja multiplikatiivse SD (multiplikatiivne SD = geom mean x SD; geom mean/SD). Geomeetriline keskmine on alati väiksem kui aritmeetiline keskmine. Lisaks on SD interval nüüd asümmeetriline ja SD on alati > 0. Nagu ennegi, 68% lognormaalsetest andmetest jääb 1SD vahemikku ning 95.5% andmetest jääb 2SD vahemikku.

lognormaalsete andmete tavapärane iseloomustus keskmise ja standardhälbega: mean(sd) on 1.8(1.9). See sd interval on sümmeetriline, ehkki andmete jaotus on vägagi ebasümmeetriline. Lisaks ennustab standardhälve, mis on suurem kui keskväärtus, suure sagedusega negatiivseid väärtusi. Sageli on aga negatiivsed muutuja väärtused võimatud (näiteks nädalas suitsetatud sigarettide arv). See on näide halvast mudelist!

Juhul kui tegu lognormaalsete andmetega on meil võimalus kasutada palju paremat mudelit varieeruvusele - multiplikatiivset standardhälvet:

```
## # A tibble: 4 x 4
##                   SD     MEAN      lower    upper
##                <chr>    <dbl>      <dbl>    <dbl>
## 1  multiplicative_SD 1.084891  0.4010893 2.934481
## 2 multiplicative_2SD 1.084891  0.1482845 7.937367
## 3        additive_SD 1.857924 -0.9636351 4.679482
## 4       additive_2SD 1.857924 -3.7851938 7.501041
```

Tavalise aritmeetitilise keskmise asemel on meil nüüd geomeetriline keskmine. Võrdluseks on antud ka tavaline (aritmeetiline) keskmine ja (aditiivne) SD. Additiivne SD on selle jaotuse kirjeldamiseks selgelt ebaadekvaatne (vt jaotuse pilti ülalpool ja võrdle mulitplikatiivse SD-ga).

Kuidas aga töötab multiplikatiivne standardhälve normaaljaotusest pärit andmetega? Kui normaalsete andmete peal multiplikatiivse sd rakendamine viib katastroofini, siis pole sel statistikul suurt praktilist kasutusruumi.

```
set.seed(5363)
norm_andmed <- rnorm(3, 100, 20)
multiplicative_sd(norm_andmed)
```

```
## # A tibble: 4 x 4
##                   SD     MEAN    lower    upper
##                <chr>    <dbl>    <dbl>    <dbl>
## 1  multiplicative_SD 108.1088 92.80205 125.9403
## 2 multiplicative_2SD 108.1088 79.66252 146.7128
## 3        additive_SD 108.9603 92.08395 125.8367
## 4       additive_2SD 108.9603 75.20756 142.7131
```

Nagu näha, on multiplikatiivse sd kasutamine normaalsete andmetega üsna ohutu (kuigi, ainult niikaua, kuni meil puuduvad negatiivsed andmed). Seega, kui sa ei ole kindel, kas tegu on normaaljaotusega või lognormaaljaotusega, arvesta, et lognormaaljaotus on bioloogias üsna tavaline (eriti ensüümreaktsioonide ja kasvuprotsesside juures). Seega on mõistlik alati kasutada multiplicative_sd() funktsiooni ja kui mõlema SD väärtused on sarnased, siis võib loota, et andmed on normaalsed ning saab avaldada tavapärase additiivse SD refereede rõõmuks.

> kui n < 10, siis mõlemad SD-d alahindavad tehnilistel põhjustel tegelikku sd-d. Ettevaatust väikeste valimitega!

(2) iseloomusta andmeid algses skaalas: median (MAD).

MAD —- median absolute deviation — on vähem tundlik outlierite suhtes ja ei eelda normaaljaotust. Puuduseks on, et MAD ei oma tõlgendust, mille kohaselt ta hõlmaks kindlat protsenti populatsiooni või valimi andmejaotusest. Seevastu sd puhul võime olla kindlad, et isegi kõige hullema jaotuse korral jäävad vähemalt 75% andmetest 2 SD piiridesse.

```
mad(andmed, constant = 1)
```

```
## [1] 0.5950562
```

```
 Ära kunagi avalda andmeid vormis: mean (MAD) või median (SD).
 Korrektne vorm on mean(SD) või median(MAD).
```

muutujate koos-varieeruvus

Andmete koos-varieeruvust mõõdetakse korrelatsiooni abil. Tulemuseks on üks number - korrelatsioonikordaja r, mis varieerub -1 ja 1 vahel.

```
r = 0 - kahte tüüpi mõõtmised (x=pikkus, y=kaal) samadest mõõteobjektidest varieeruvad üksteisest sõltu
r = 1: kui ühe muutuja väärtus kasvab, kasvab ka teise muutuja väärtus alati täpselt samas proportsioon
r = -1: kui ühe muutuja väärtus kasvab, kahaneb teise muutuja väärtus alati täpselt samas proportsiooni
```

Kui r on -1 või 1, saame me x väärtust teades täpselt ennustada y väärtuse (ja vastupidi, teades y väärrust saame täpselt ennustada x väärtuse).
Kuidas tõlgendame aga tulemust r = 0.9? Mitte kuidagi. Selle asemel tõlgendame r2 = 0.9**2 = 0.81 – mis tähendab, et x-i varieeruvus suudab seletada 81% y varieeruvusest ja vastupidi, et Y-i varieeruvus suudab seletada 81% X-i varieeruvusest.

```
#correlation<-cor.test(iris$Sepal.Length, iris$Sepal.Width, na.rm=T, method = "pearson") # a list of 9

#names(correlation)
#str(correlation)
#correlation$conf.int
cor(iris$Sepal.Length, iris$Sepal.Width, use="complete.obs")
```

```
## [1] -0.1175698
```

Korrelatsioonikordaja väärtus sõltub mitte ainult andmete koos-varieeruvusest vaid ka andmete ulatusest. Suurema ulatusega andmed X ja/või Y teljel annavad keskeltläbi 0-st kaugemal oleva korrelatsioonikordaja. Selle pärast sobib korrelatsioon halvasti näiteks korduskatsete kooskõla mõõtmiseks.

Lisaks, korrelatsioonikordaja mõõdab vaid andmete *lineaarset* koos-varieeruvust: kui andmed koos-varieeruvad mitte-lineaarselt, siis võivad ka väga tugevad koos-varieeruvused jääda märkamatuks.

> Kõik summaarsed statistikud kaotavad enamuse teie andmetes leiduvast infost – see kaotus on õigustatud ainult siis, kui teie poolt valitud statistik iseloomustab hästi andmete sügavamat olemust (näiteks tüüpilist mõõtmistulemust või andmete varieeruvust).

```
#Kuidas arvutada correlatsioonimaatriksit koos adjusteeritud p väärtustega
#numeric columns only!
print(psych::corr.test(iris[-5], use="complete"), short = FALSE)
```

2.2 EDA — eksploratoorne andmeanalüüs

Kui ühenumbriline andmete summeerimine täidab eelkõige kokkuvõtliku kommunikatsiooni eesmärki, siis EDA on suunatud teadlasele endale. EDA eesmärk on andmeid eelkõige graafiliselt vaadata, et saada aimu 1) andmete kvaliteedist ja 2) lasta andmetel "sellisena nagu nad on" kõneleda ja sugereerida uudseid teaduslikke hüpoteese. Neid hüpoteese peaks siis testima formaalse statistilise analüüsi abil.

> EDA: mida rohkem graafikuid, seda rohkem võimalusi uute mõtete tekkeks!

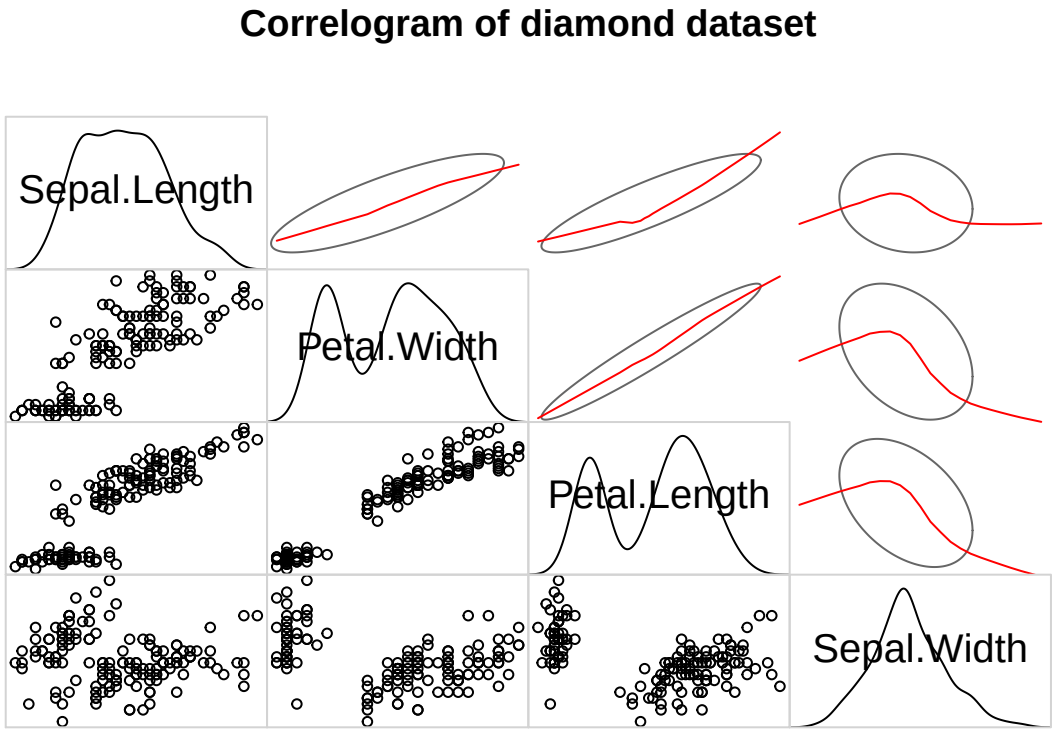
Kõigepealt vaatame andmeid numbrilise kokkuvõttena:

```
psych::describe(iris)
```

```
##              vars   n mean   sd median trimmed  mad min max range  skew
## Sepal.Length    1 150 5.84 0.83   5.80    5.81 1.04 4.3 7.9   3.6  0.31
## Sepal.Width     2 150 3.06 0.44   3.00    3.04 0.44 2.0 4.4   2.4  0.31
## Petal.Length    3 150 3.76 1.77   4.35    3.76 1.85 1.0 6.9   5.9 -0.27
## Petal.Width     4 150 1.20 0.76   1.30    1.18 1.04 0.1 2.5   2.4 -0.10
## Species*        5 150 2.00 0.82   2.00    2.00 1.48 1.0 3.0   2.0  0.00
##              kurtosis   se
## Sepal.Length    -0.61 0.07
## Sepal.Width      0.14 0.04
## Petal.Length    -1.42 0.14
## Petal.Width     -1.36 0.06
## Species*        -1.52 0.07
```

Millised korrelatsioonid võiksid andmetes esineda?

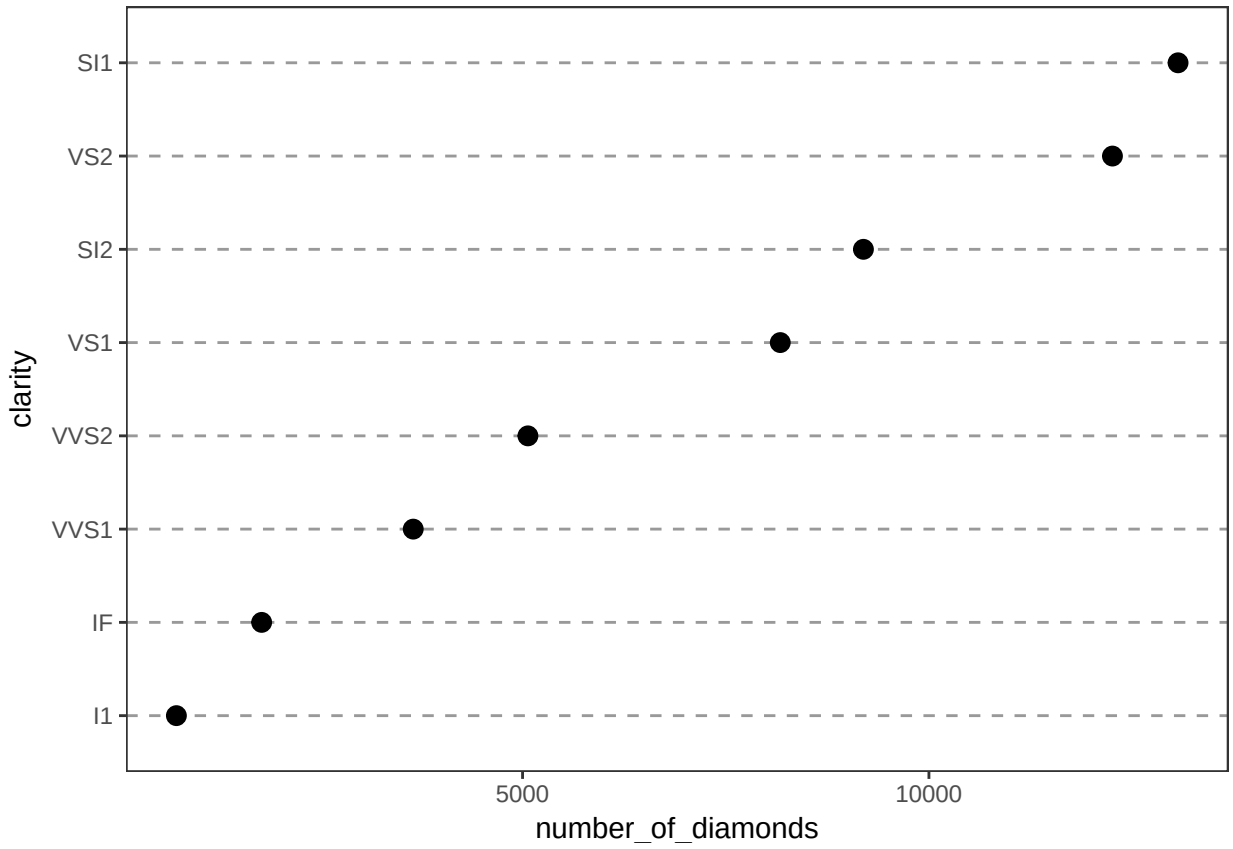
```
library(corrgram) #PCA for ordering

corrgram(iris, order=TRUE,
         lower.panel = panel.pts,
         upper.panel = panel.ellipse,
         diag.panel = panel.density,
         main="Correlogram of diamond dataset")
```



Kuidas uurida muutuja sisest varieeruvust

Muutuja - midagi, mida mõõdeti (näiteks mõõteobjektide kaal). Kui iga muutuja kohta on vaid üks number, mida plottida, kasuta Cleveland plotti:

```
dd <- diamonds %>% group_by(clarity) %>% summarise(number_of_diamonds=n())
dd %>% ggplot(aes(x=number_of_diamonds,
                  y=reorder(clarity, number_of_diamonds))) +
  geom_point(size=3) +
  theme_bw() +
  theme(panel.grid.major.x = element_blank(),
        panel.grid.minor.x = element_blank(),
        panel.grid.major.y = element_line(colour="grey60", linetype="dashed")) +
  labs(y="clarity")
```

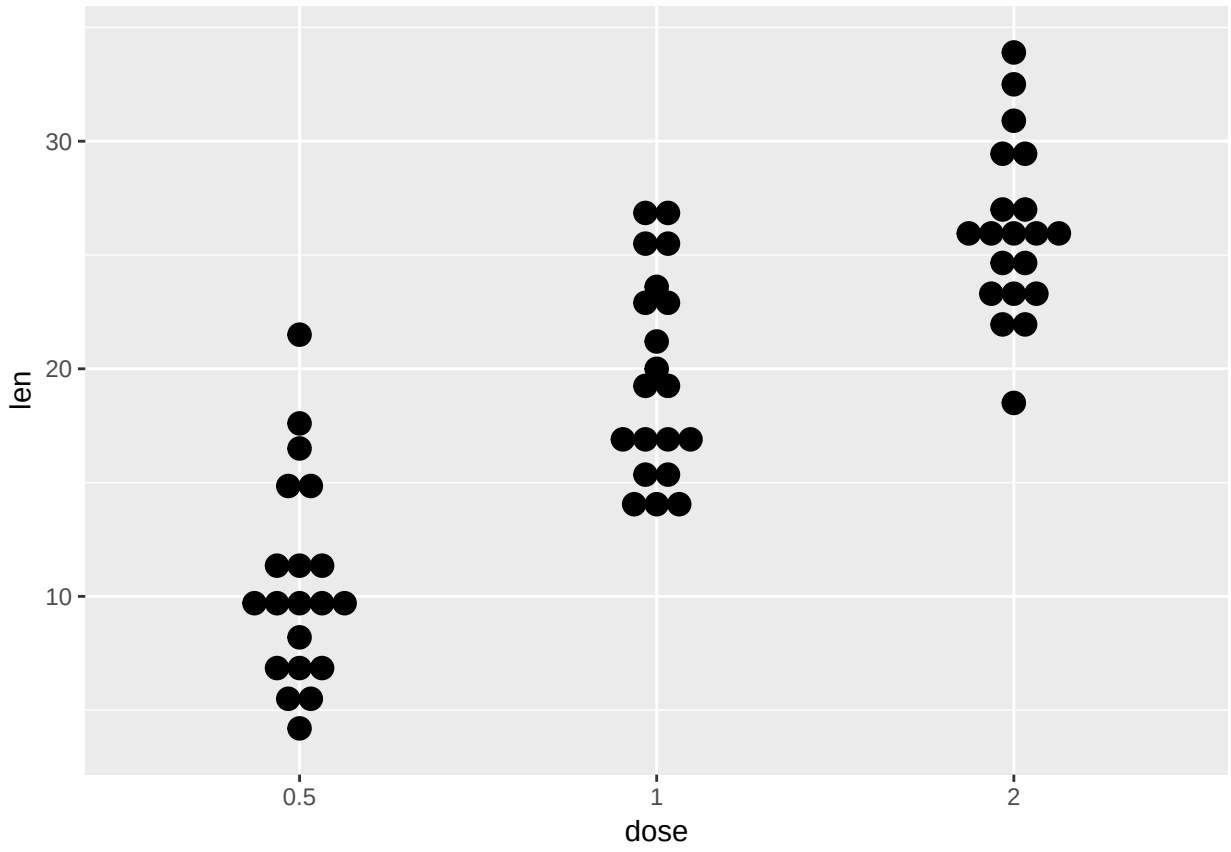


Ploti valik sõltub valimi suurusest.

1) N < 20 - ploti iga andmepunkt eraldi (stripchart(), plot()) ja keskmine või mediaan.
2) 20 > N > 100: geom_dotplot() histogrammi vaates
3) N > 100: geom_histogram(), geom_density() — nende abil saab ka 2 kuni 6 jaotust võrrelda
4) Mitme jaotuse kõrvuti vaatamiseks kui N > 15: geom_boxplot() or, when N > 50, geom_violin(), geom_joy()

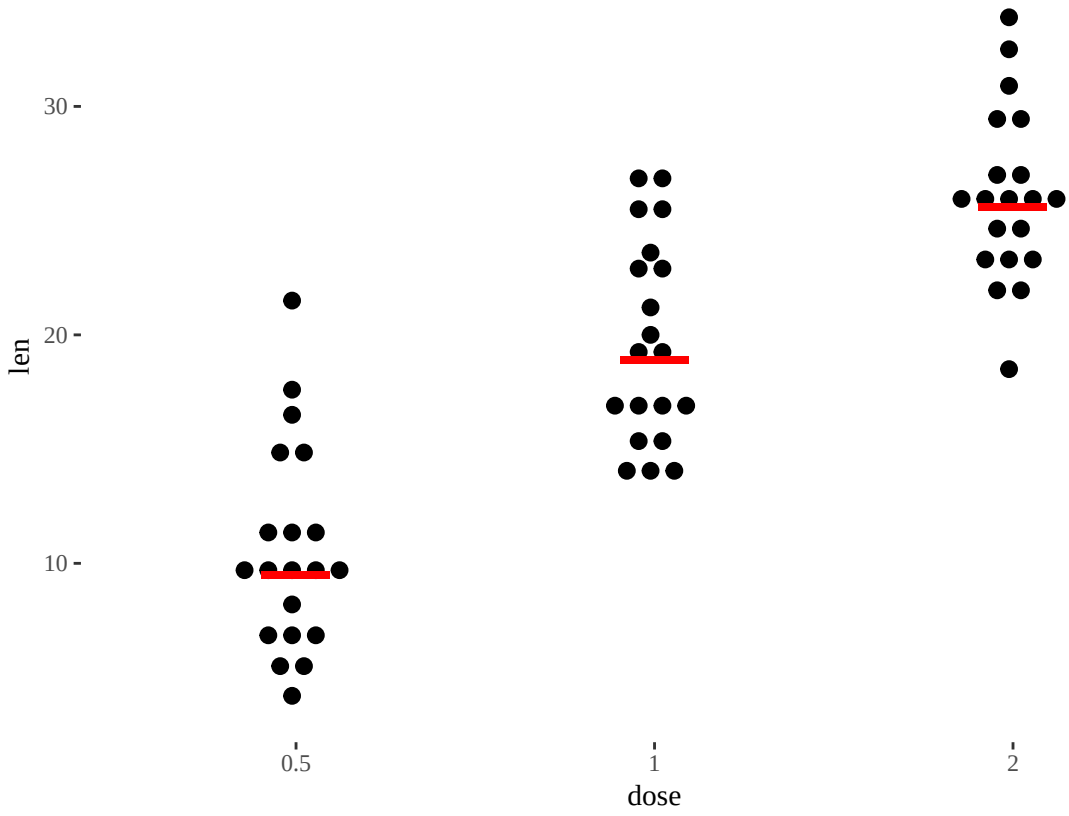
```
ToothGrowth <- ToothGrowth
ToothGrowth$dose <- as.factor(ToothGrowth$dose)
p<-ggplot(ToothGrowth, aes(x=dose, y=len)) +
  geom_dotplot(binaxis='y', stackdir='center')
p
```

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```



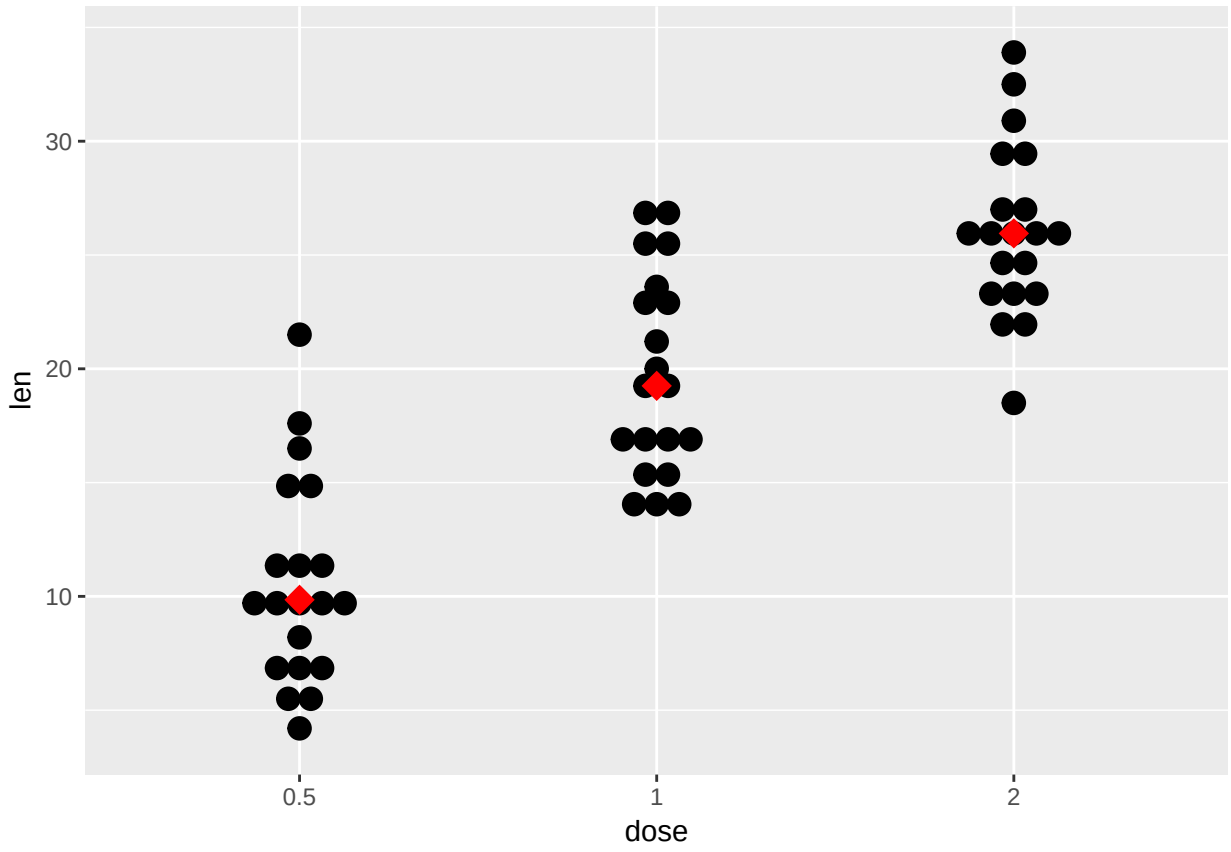

```
# Change dotsize and stack ratio, add line or dot for median
ggplot(ToothGrowth, aes(x=dose, y=len)) +
  geom_dotplot(binaxis='y', stackdir='center',
               stackratio=1.5, dotsize=0.7)+
  stat_summary(fun.y = median, geom = "point", shape = 95,
               color = "red", size = 15) +
  theme_tufte()
```

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```



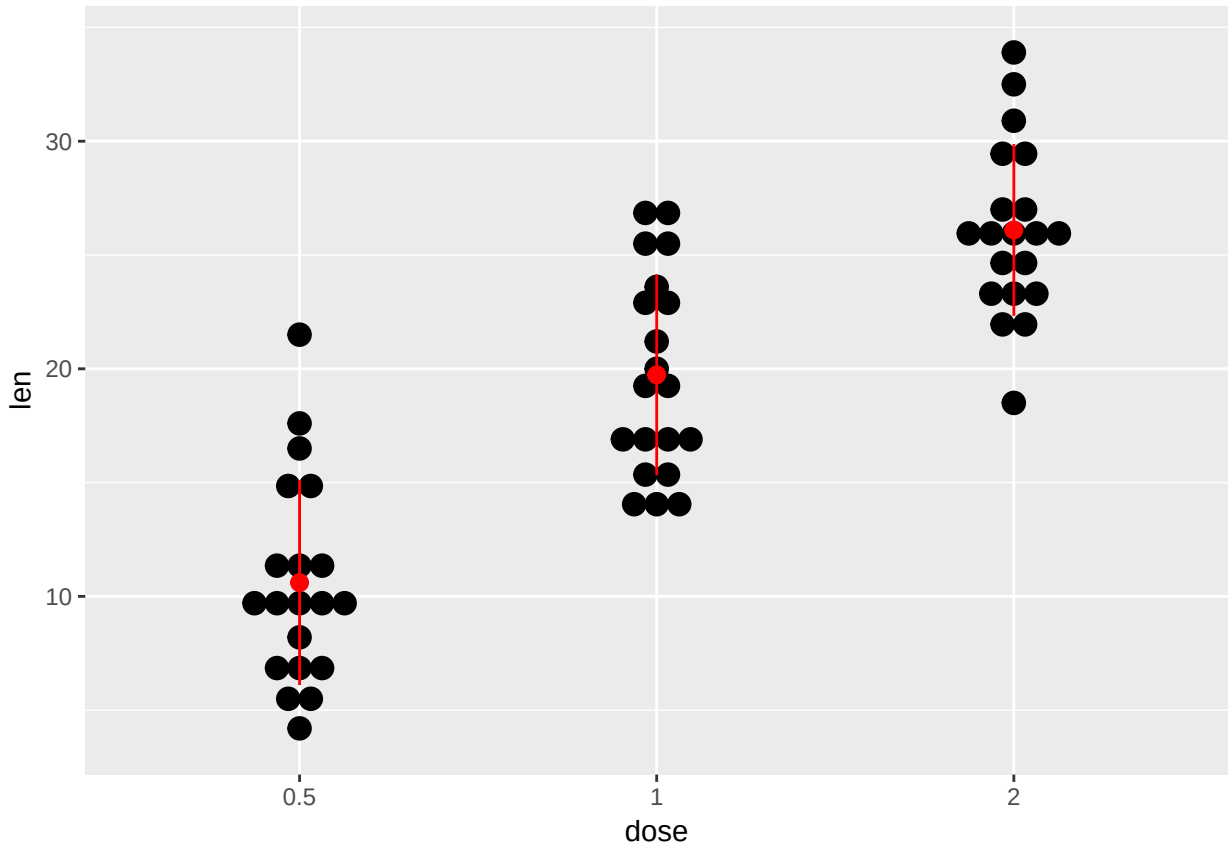

```
p + stat_summary(fun.y=median, geom="point", shape=18,
                 size=5, color="red")
```

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```



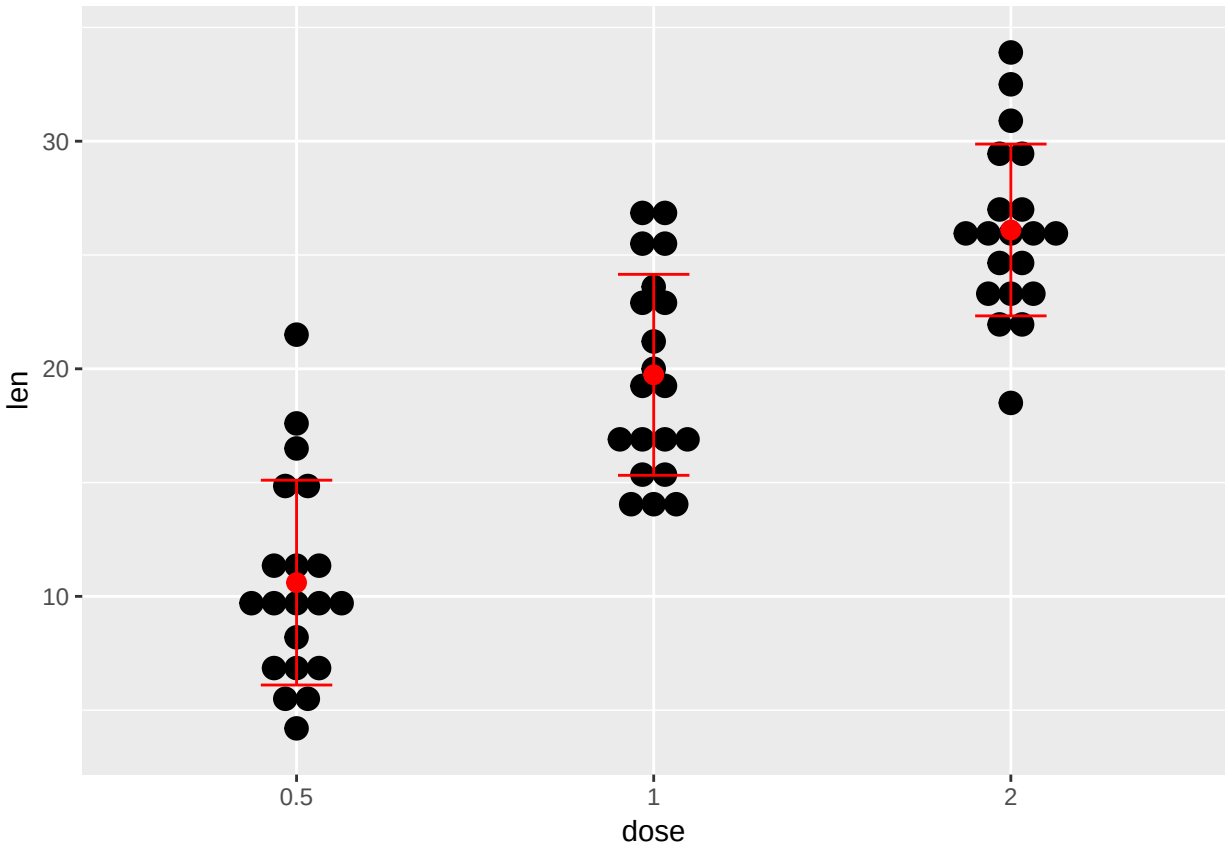

```
#add mean and SD, use pointrange
p + stat_summary(fun.data=mean_sdl, fun.args = list(mult=1),
                 geom="pointrange", color="red")
```

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#use errorbars
p + stat_summary(fun.data=mean_sdl, fun.args = list(mult=1),
        geom="errorbar", color="red", width=0.2) +
  stat_summary(fun.y=mean, geom="point", size=3, color="red")
```

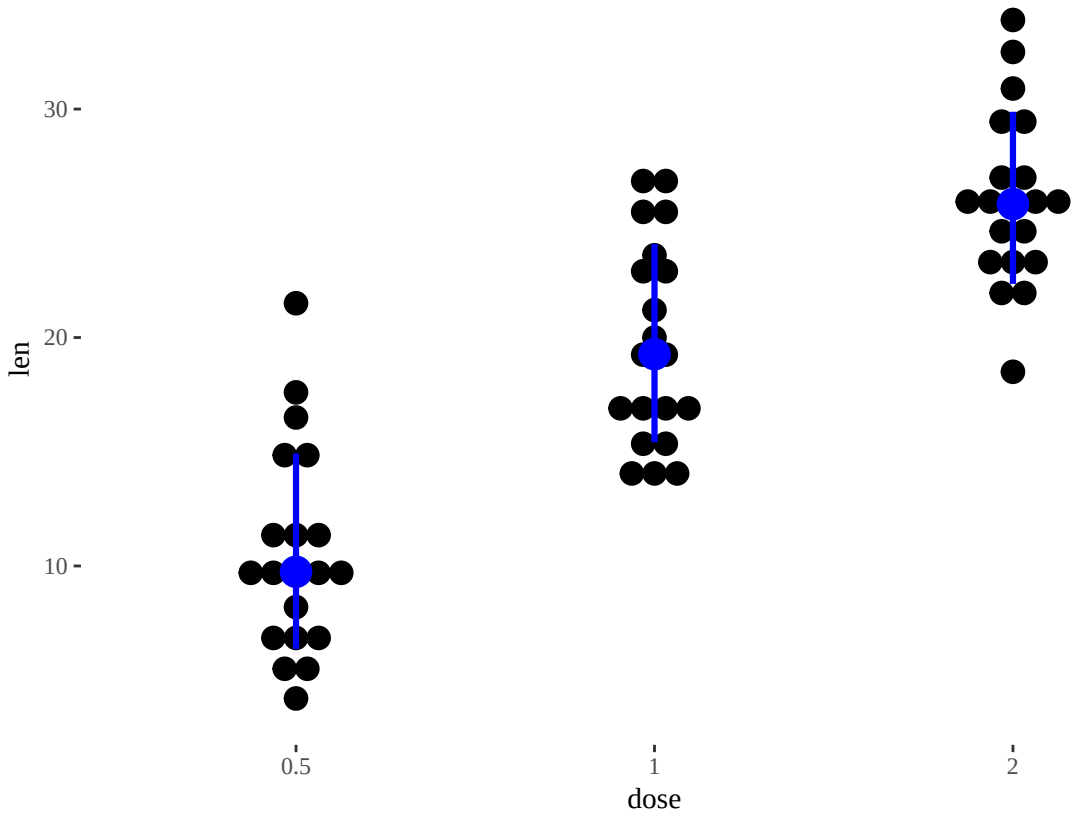
```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```



Use a custom summary function :

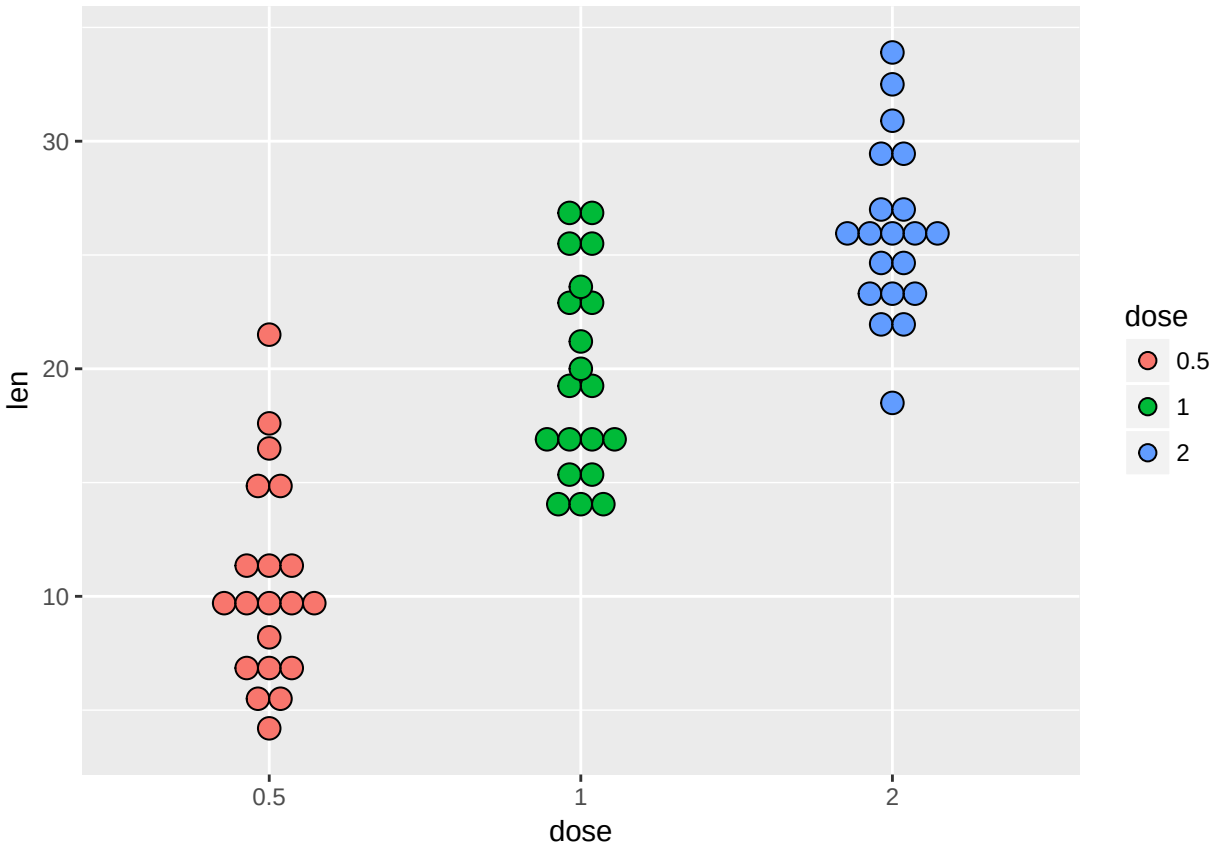
```
# Function to produce summary statistics (geometric mean and multipülicative sd)
multi_sd <- function(x) {
  x <- na.omit(x)
  a <- log10(x)
  b <- mean(a)
  c <- sd(a)
  g_mean <- 10**b
  msd <- 10**c
  ymin <- g_mean/msd
  ymax <- g_mean * msd
 return(c(y = g_mean, ymin = ymin, ymax = ymax))
}
p + stat_summary(fun.data=multi_sd, color="blue", size=1.1) + theme_tufte()
```

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# Change dot plot colors by groups
p<-ggplot(ToothGrowth, aes(x=dose, y=len, fill=dose)) +
  geom_dotplot(binaxis='y', stackdir='center')
p
```

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```



It is also possible to change manually dot plot colors using the functions :

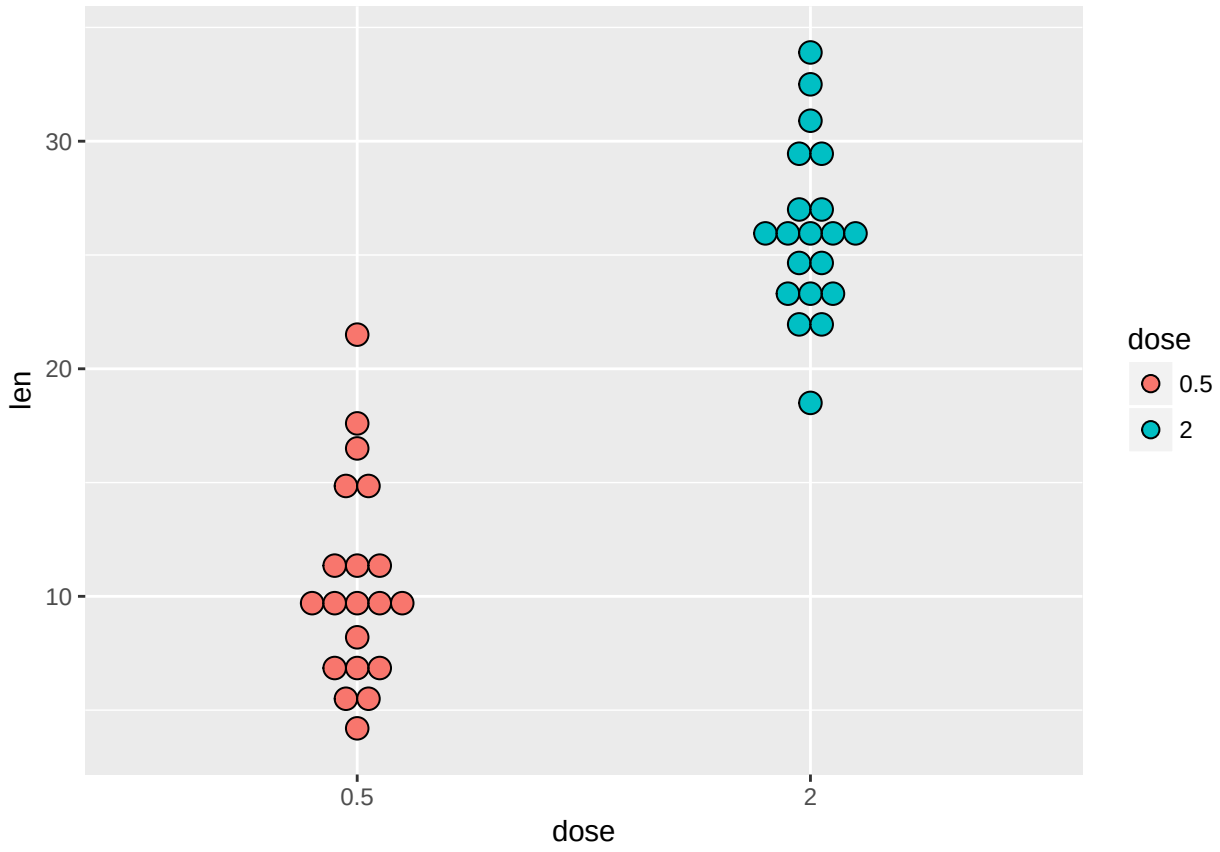
scale_fill_manual() : to use custom colors

scale_fill_brewer() : to use color palettes from RColorBrewer package

scale_fill_grey() : to use grey color palettes

```
#Choose which items to display :
p + scale_x_discrete(limits=c("0.5", "2"))
```

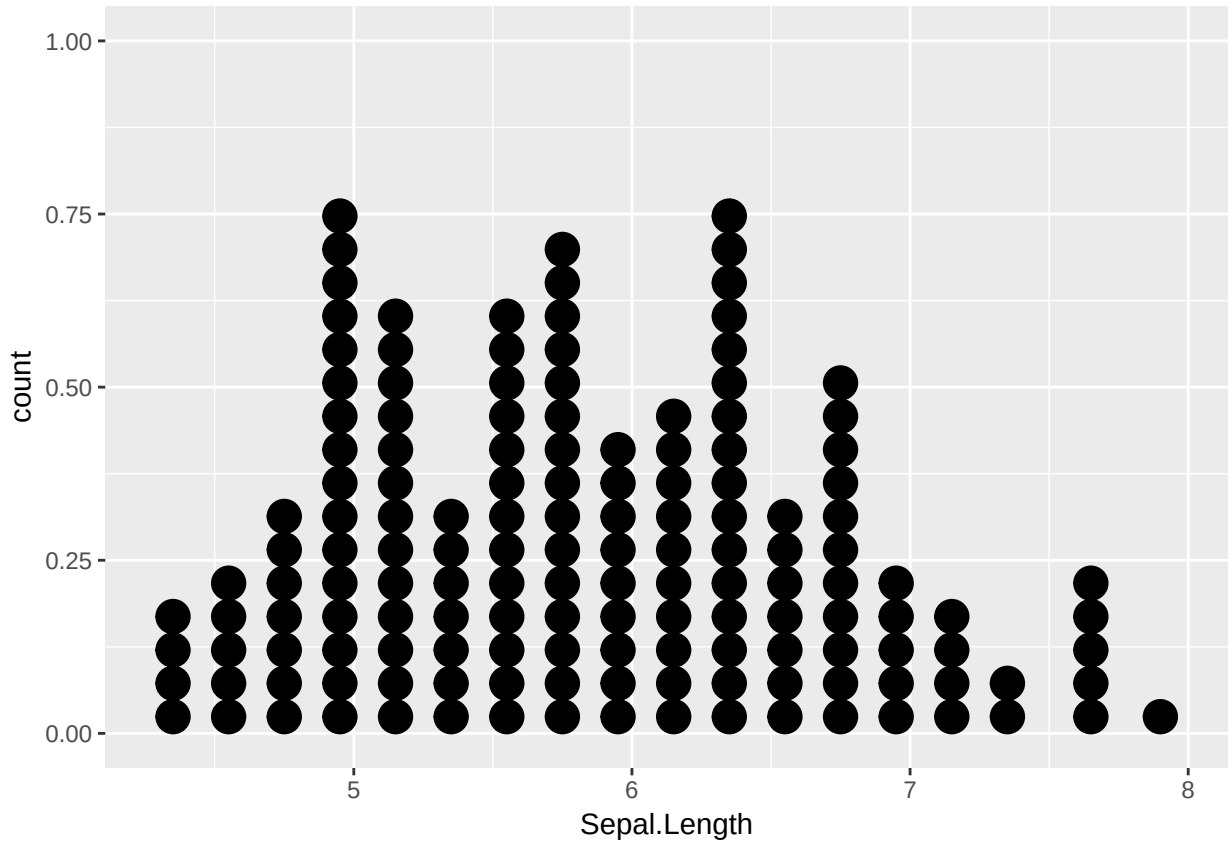
```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
## Warning: Removed 20 rows containing non-finite values (stat_bindot).
```



Dotplot kui histogram:

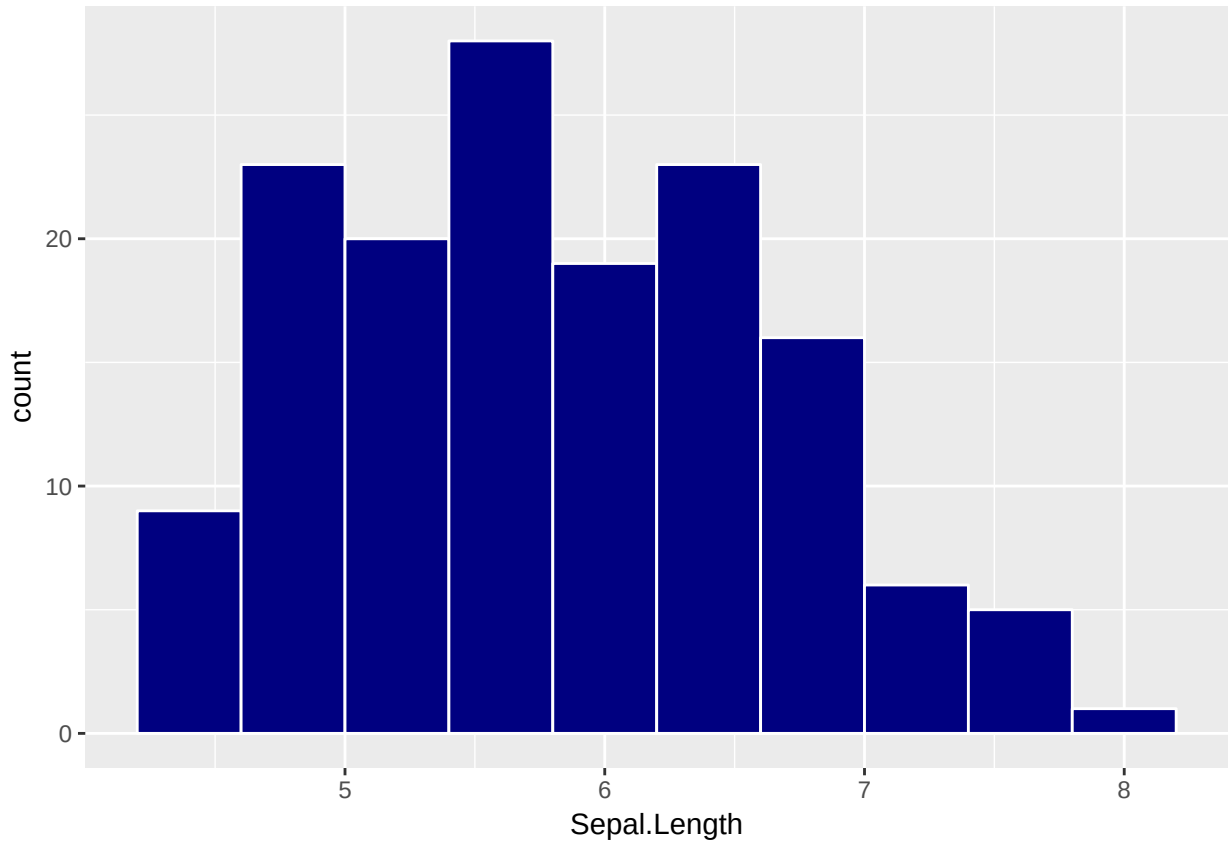
```
ggplot(iris, aes(Sepal.Length)) + geom_dotplot()
```

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```

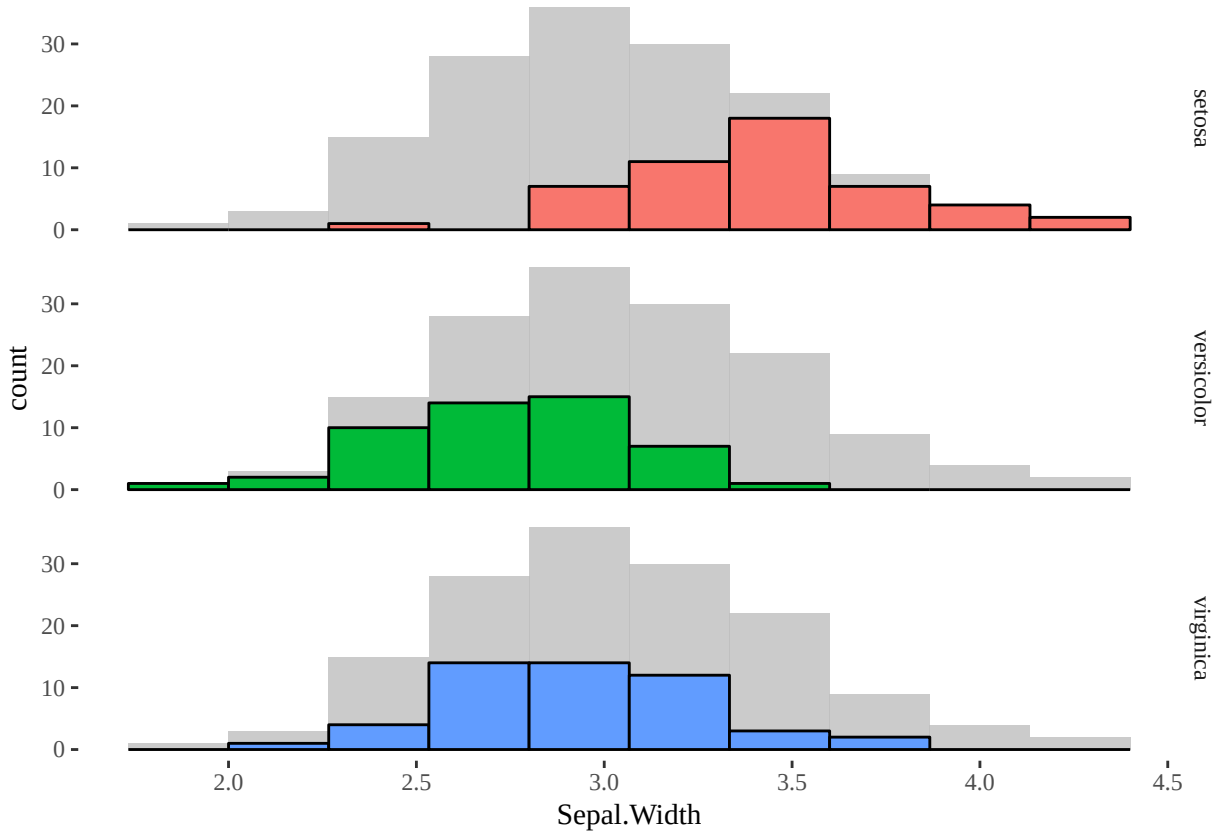


Histogram:

```
ggplot(iris, aes(Sepal.Length)) +
  geom_histogram(bins = 10, color="white", fill = "navyblue")
```

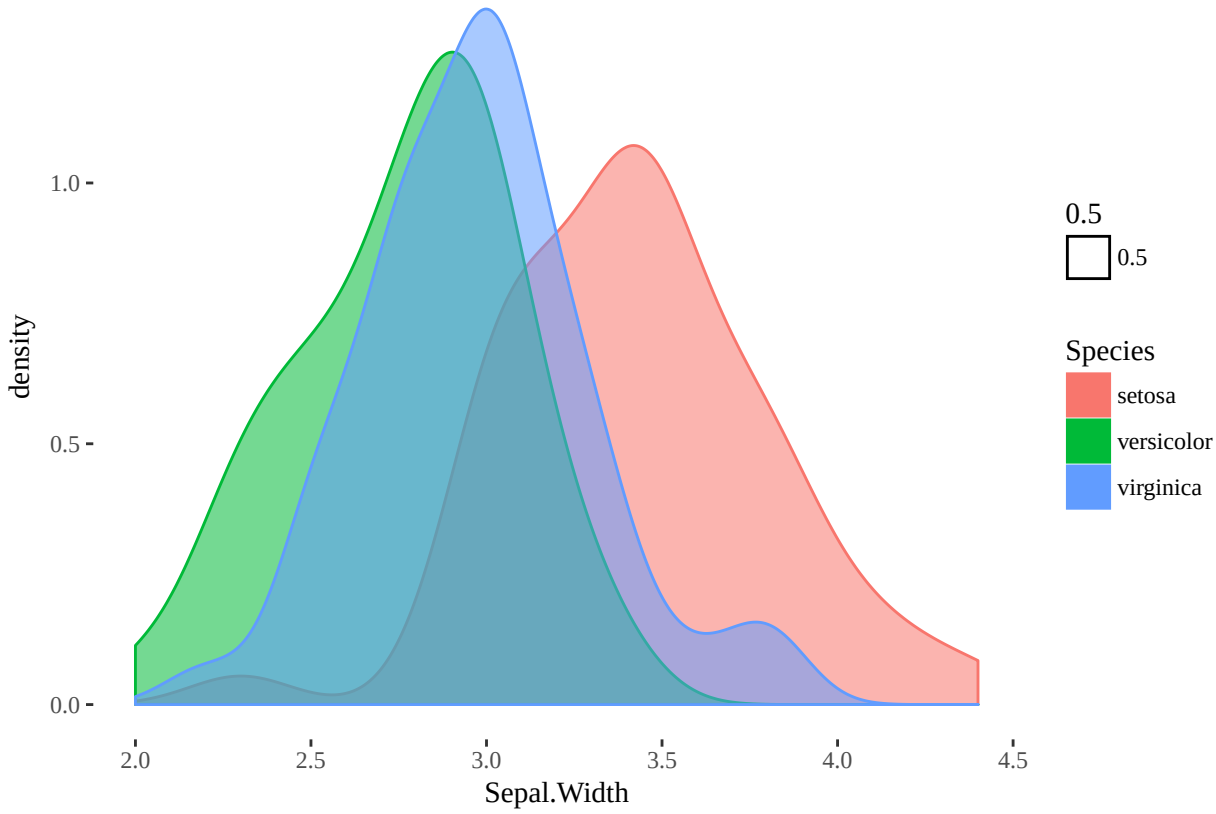



```
library(ggthemes)
d <- iris        # Full data set
d_bg <- d[, -5]  # Background Data - full without the 5th column (Species)

ggplot(data = d, aes(x = Sepal.Width, fill = Species)) +
  geom_histogram(data = d_bg, fill = "grey", alpha=0.8, bins=10) +
  geom_histogram(colour = "black", bins=10) +
  facet_grid(Species~.) +
  guides(fill = FALSE) +  # to remove the legend
  theme_tufte()          # for clean look overall
```



density plot:

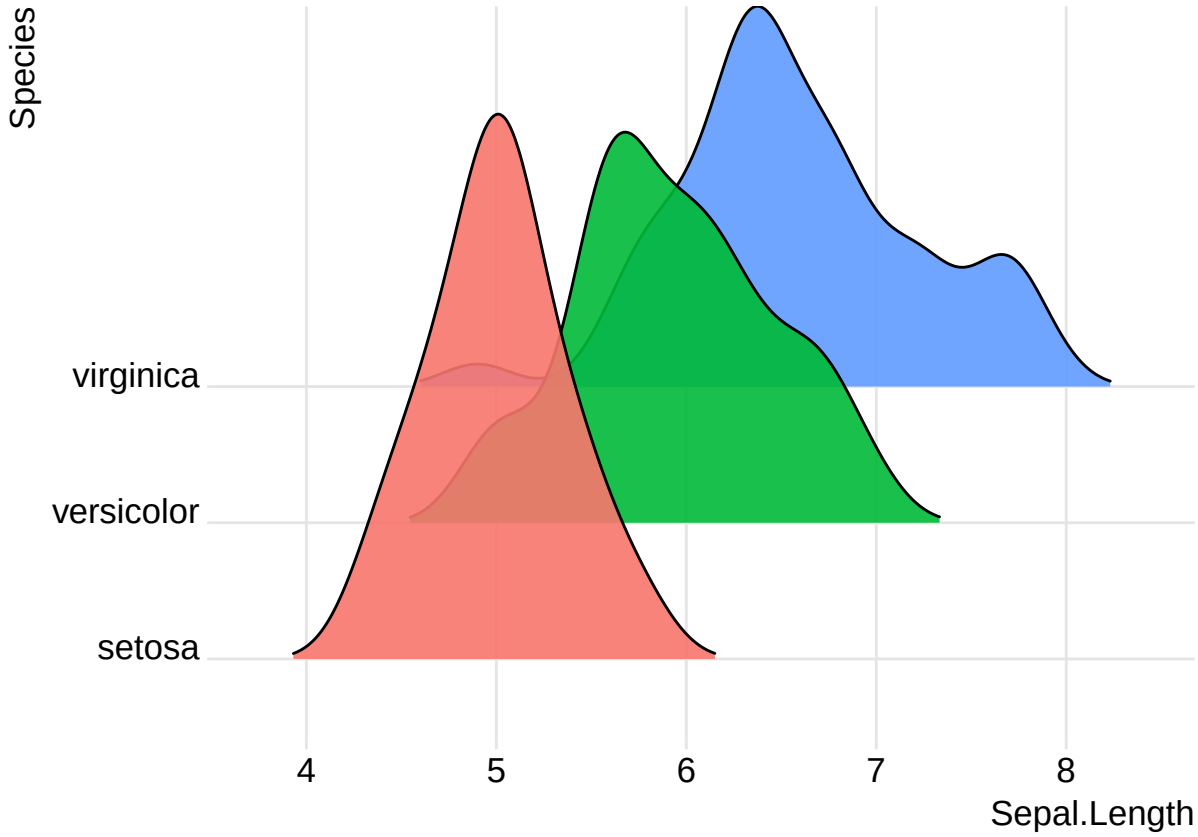
```
iris%>%ggplot()+
  geom_density(aes(Sepal.Width, fill=Species, color=Species, alpha=0.5))+
  theme_tufte()
```



joyplot võimaldab kõrvuti panna isegi sadu density plotte

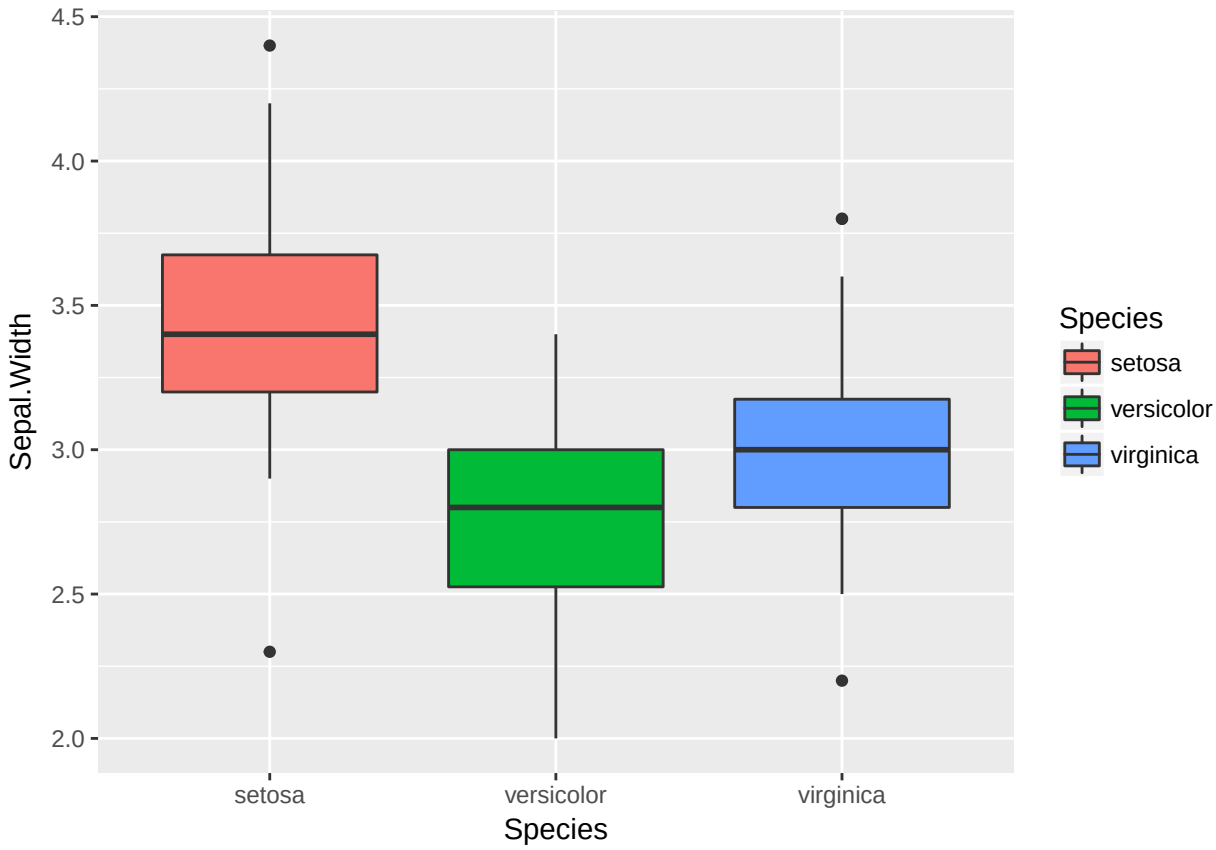
```
library(ggjoy)
ggplot(iris, aes(x=Sepal.Length, y=Species, fill=Species)) +
  geom_joy(scale=4, rel_min_height=0.01, alpha=0.9) +
  theme_joy(font_size = 13, grid=TRUE) +
  theme(legend.position = "none")
```

```
## Picking joint bandwidth of 0.181
```



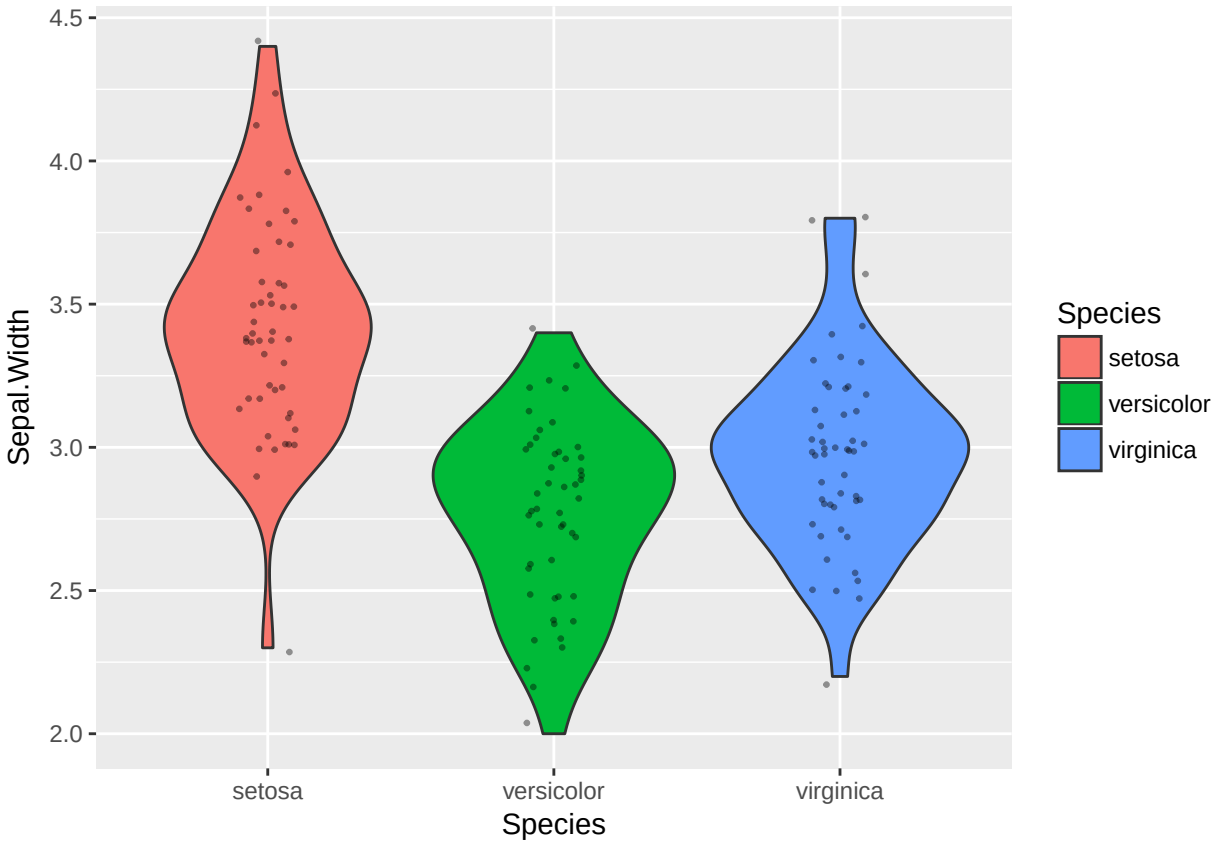
Boxplot:

```
ggplot(iris, aes(Species, Sepal.Width, fill=Species)) + geom_boxplot()
```



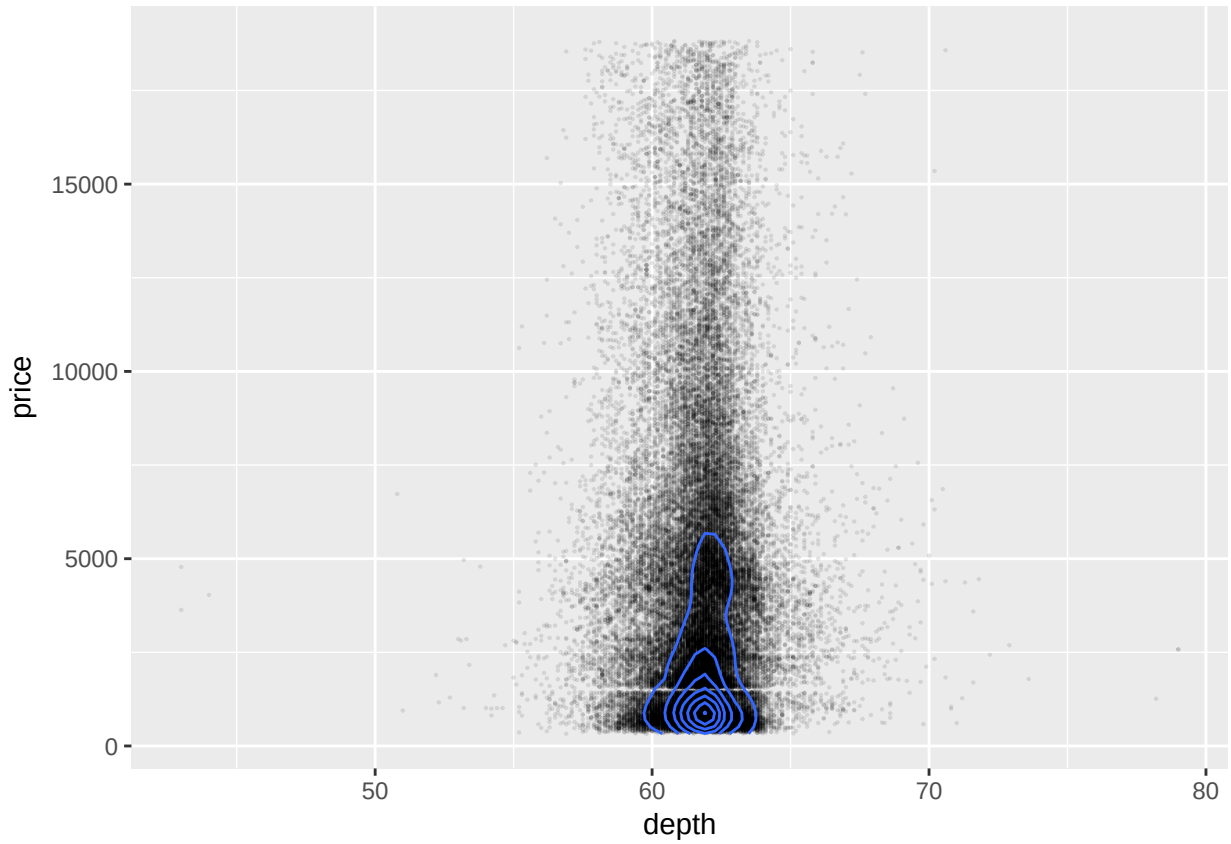
violin plot plus jitterplot:

```
ggplot(iris, aes(Species, Sepal.Width)) +
  geom_violin(aes(fill=Species)) +
  geom_jitter(width = 0.1, alpha=0.4, size=0.5)
```



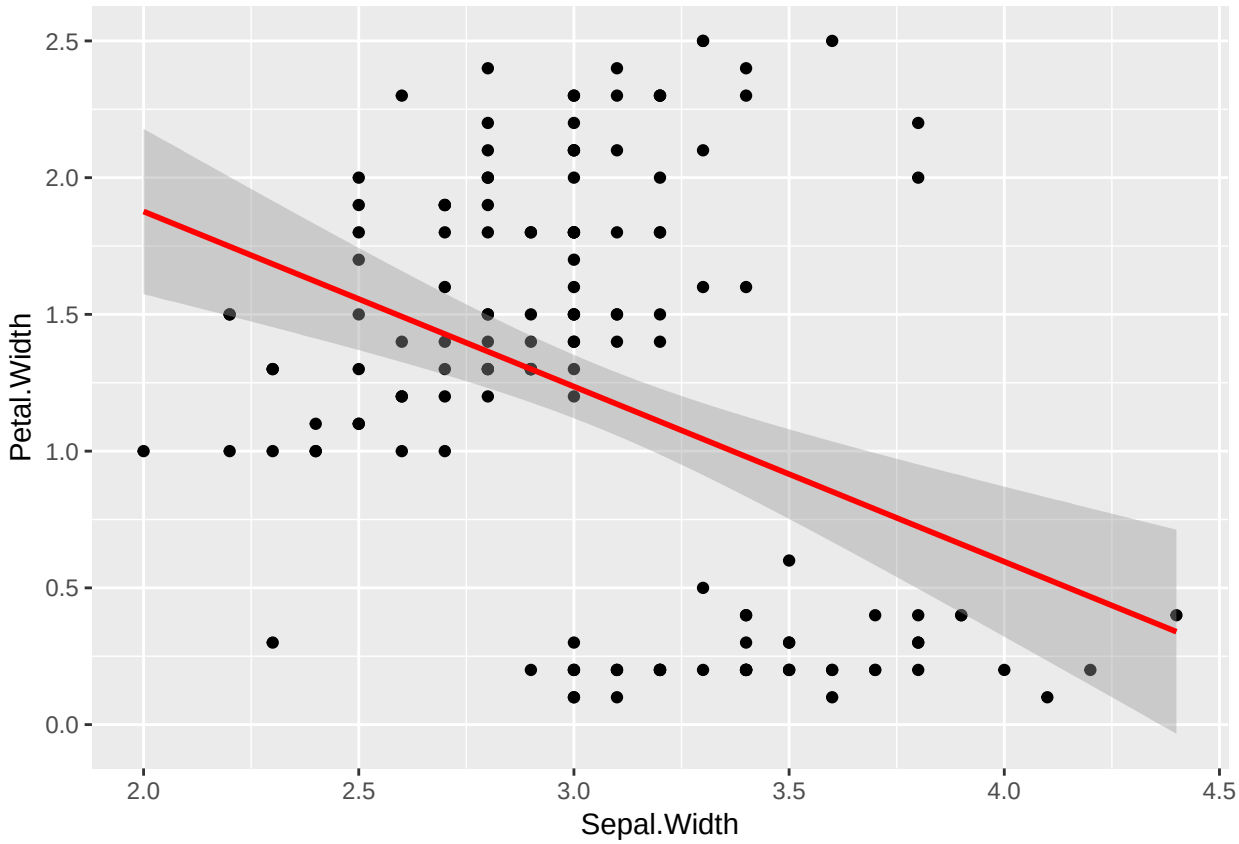
Kahe muutuja koos-varieerumine

```
ggplot(data = diamonds, aes(x = depth, y = price)) +
  geom_point(size=0.1, alpha=0.1) +
  geom_density2d()
```



Fit a linear model and plot the dots and model:

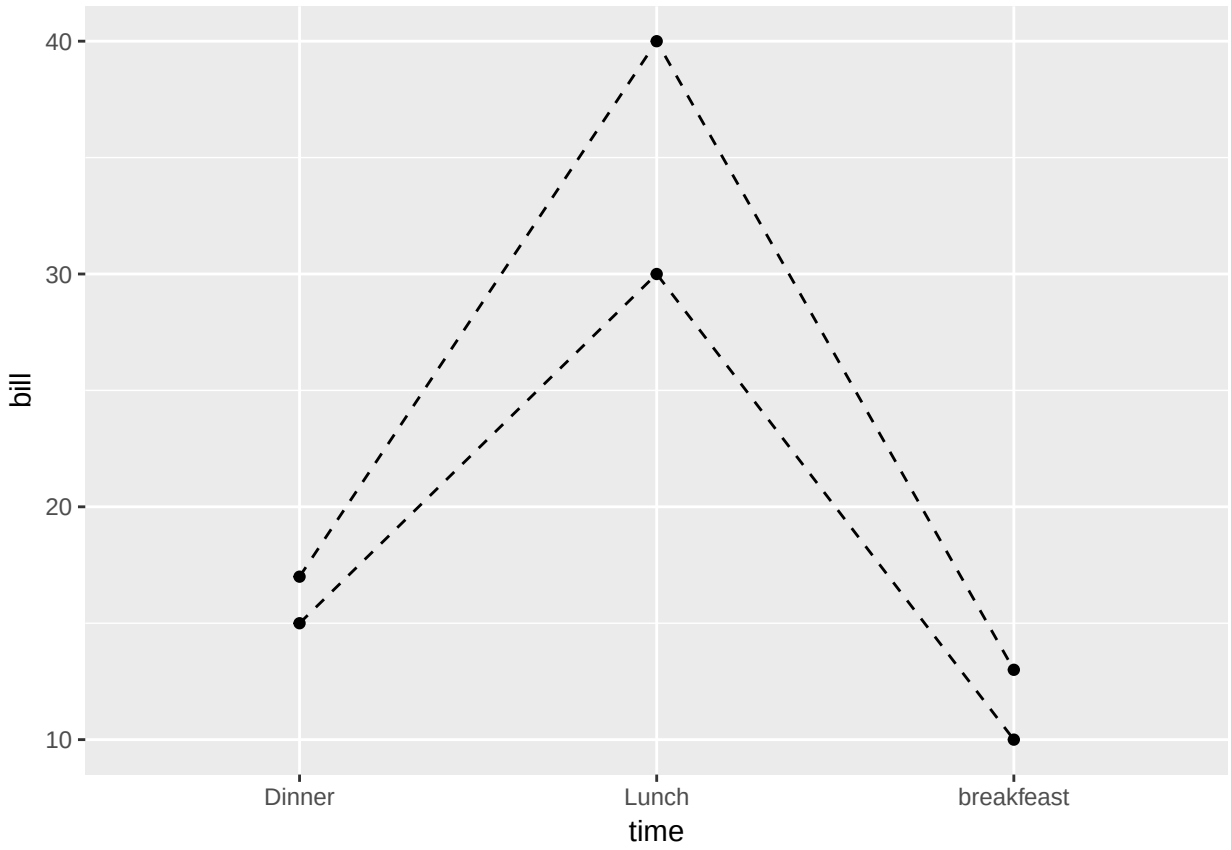
```
ggplot(data=iris, aes(Sepal.Width, Petal.Width))+
  geom_point()+
  geom_smooth(method="lm", color="red")
```



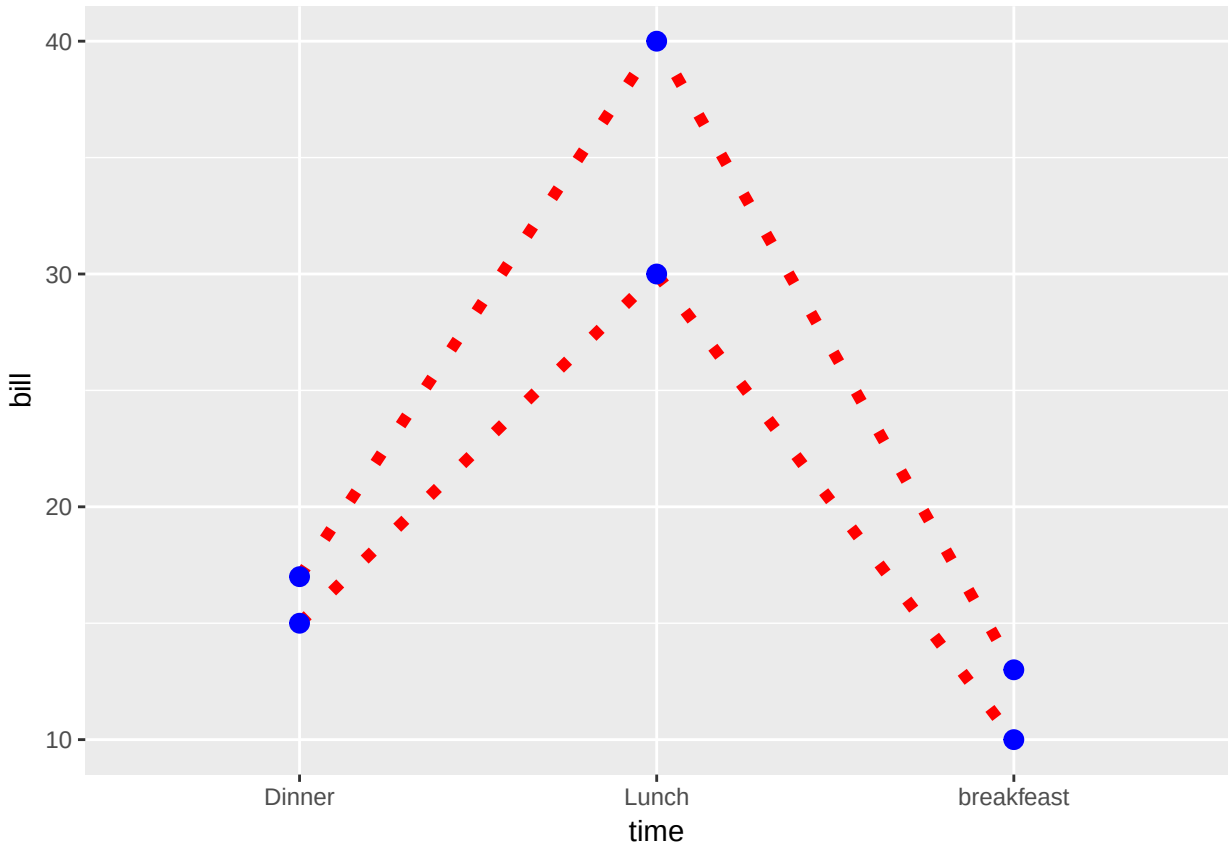
Joongraafikud

Joonetüübid : “blank”, “solid”, “dashed”, “dotted”, “dotdash”, “longdash”, “twodash”.

```
df2 <- data.frame(sex = rep(c("Female", "Male"), each=3),
                  time=c("breakfeast", "Lunch", "Dinner"),
                  bill=c(10, 30, 15, 13, 40, 17) )
# Change line types
ggplot(data=df2, aes(x=time, y=bill, group=sex)) +
  geom_line(linetype="dashed")+
  geom_point()
```

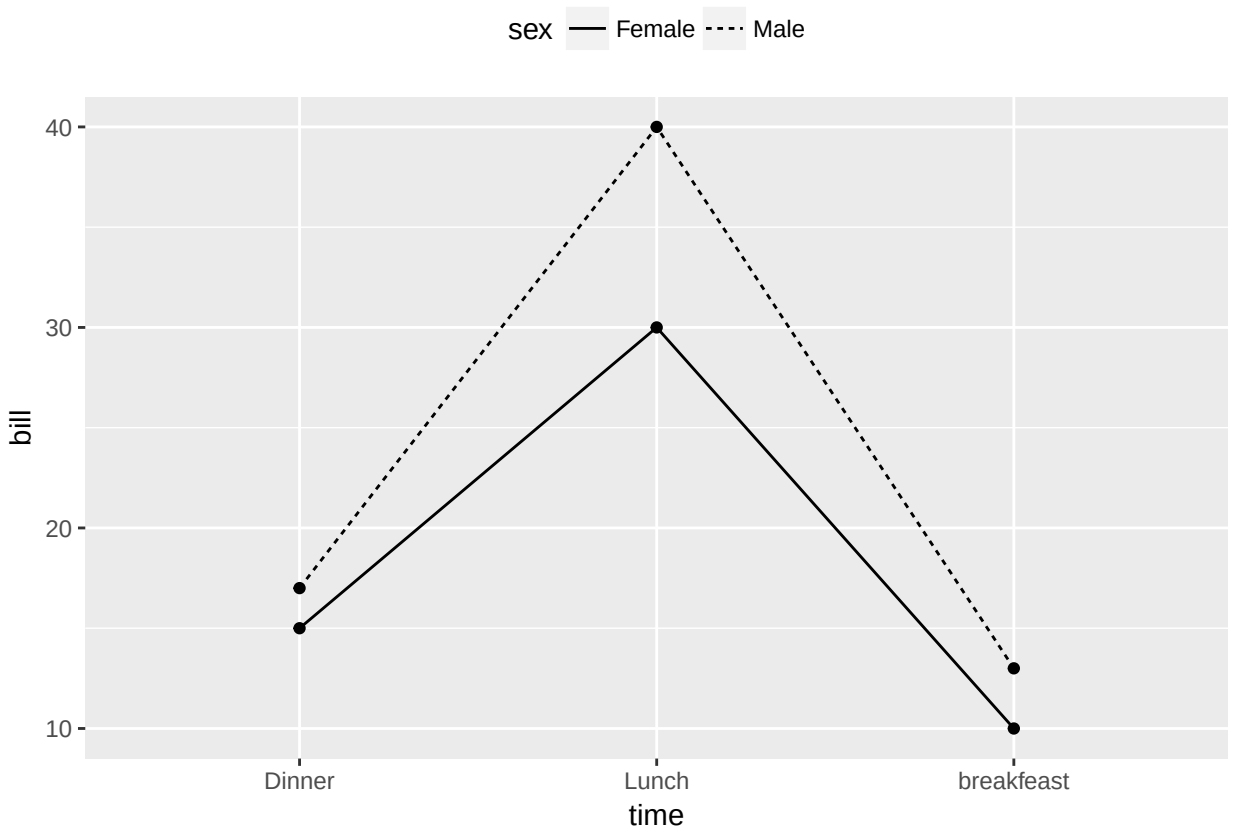



```
# Change line colors and sizes
ggplot(data=df2, aes(x=time, y=bill, group=sex)) +
  geom_line(linetype="dotted", color="red", size=2)+
  geom_point(color="blue", size=3)
```

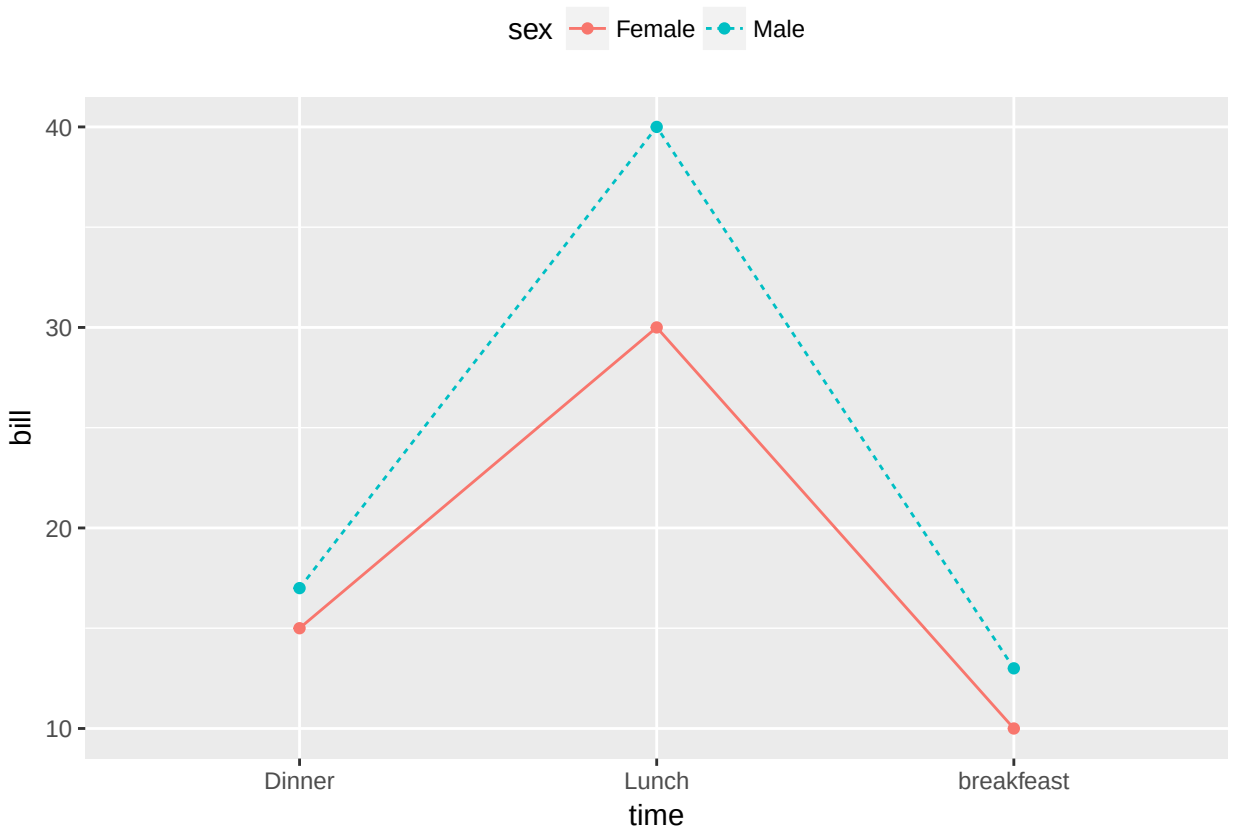


Muudab tüüpi automaatselt muutuja sex taseme järgi

```
# Change line types by groups (sex)
ggplot(df2, aes(x=time, y=bill, group=sex)) +
  geom_line(aes(linetype=sex))+
  geom_point()+
  theme(legend.position="top")
```




```
# Change line types + colors
ggplot(df2, aes(x=time, y=bill, group=sex)) +
  geom_line(aes(linetype=sex, color=sex))+
  geom_point(aes(color=sex))+
  theme(legend.position="top")
```



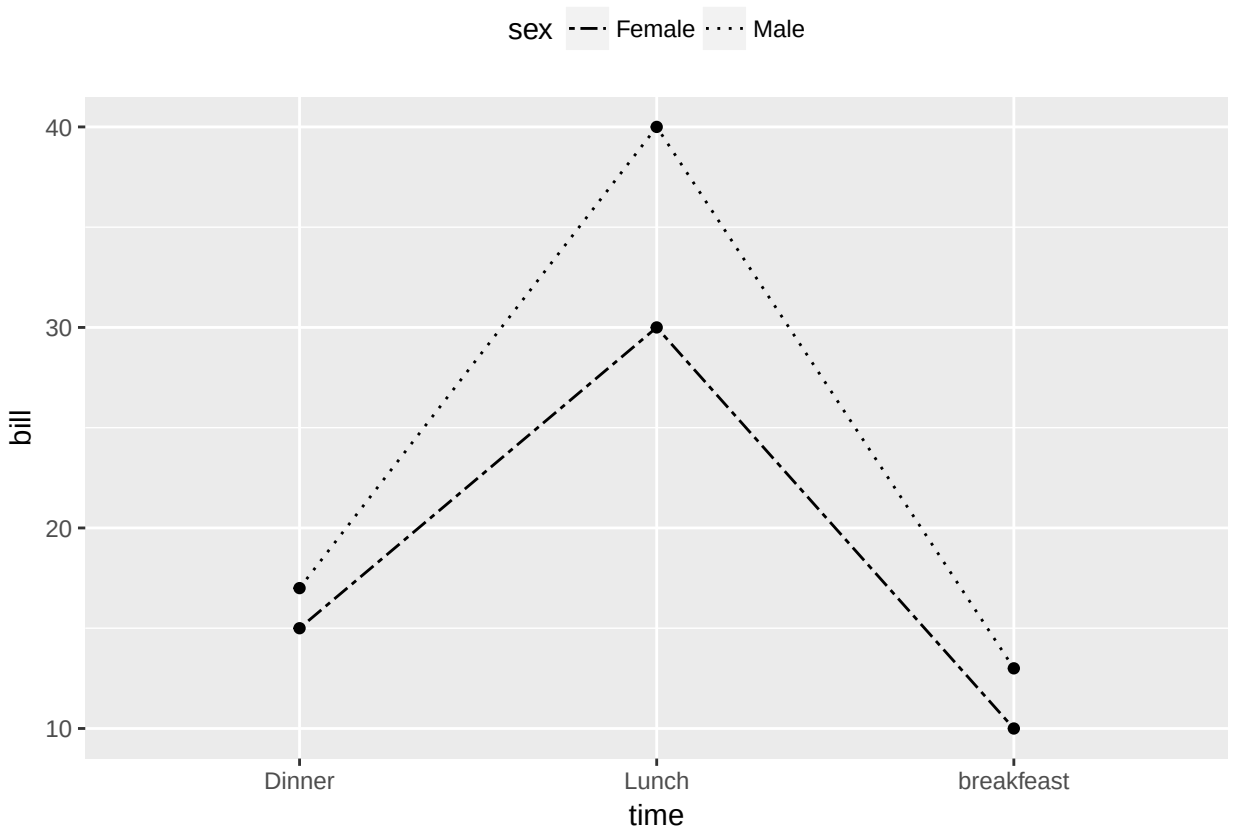
Muuda jooni käsitsi:

scale_linetype_manual() : joone tüüp

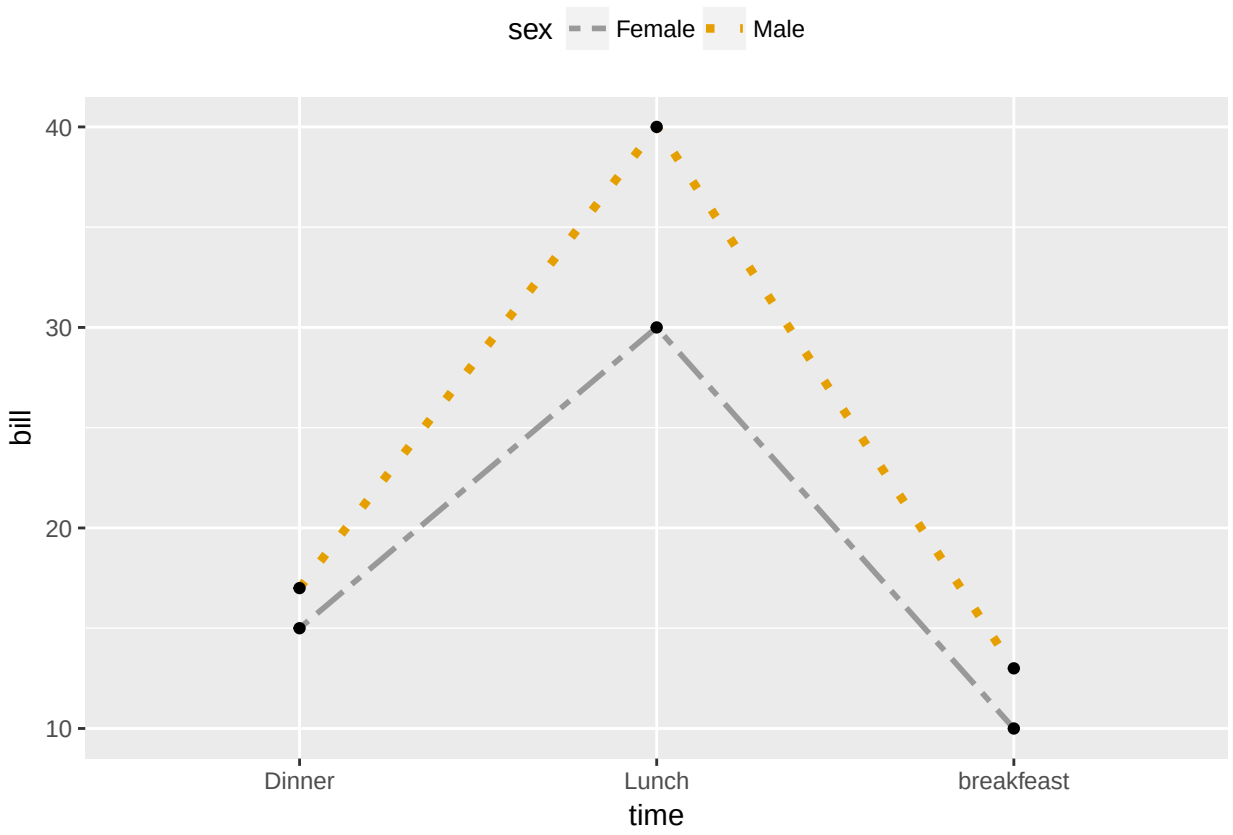
scale_color_manual() : joone värv

scale_size_manual() : joone laius

```
# Set line types manually
ggplot(df2, aes(x=time, y=bill, group=sex)) +
  geom_line(aes(linetype=sex))+
  geom_point()+
  scale_linetype_manual(values=c("twodash", "dotted"))+
  theme(legend.position="top")
```




```
# Change line colors and sizes
ggplot(df2, aes(x=time, y=bill, group=sex)) +
  geom_line(aes(linetype=sex, color=sex, size=sex))+
  geom_point()+
  scale_linetype_manual(values=c("twodash", "dotted"))+
  scale_color_manual(values=c('#999999','#E69F00'))+
  scale_size_manual(values=c(1, 1.5))+
  theme(legend.position="top")
```



Kokkuvõte:

a. Andmepunktide plottimine säilitab maksimaalselt andmetes olevat infot (nii kasulikku infot kui müra). Aitab leida outliereid (valesti sisestatud andmeid, valesti mõõdetud proove jms). Kui valim on väiksem kui 20, piisab täiesti üksikute andmepunktide plotist koos mediaaniga. Dot-plot ruulib.

b. Histogramm – kõigepealt mõõtskaala ja seejärel andmed jagatakse võrdse laiusega binnidesse ja plotitakse binnide kõrgused. Bin, kuhu läks 20 andmepunkti on 2X kõrgem kui bin, kuhu läks 10 andmepunkti. Samas, bini laius/ulatus mõõteskaalal pole teile ette antud – ja sellest võib sõltuda histogrammi kuju. Seega on soovitav proovida erinevaid bini laiusi ja võrrelda saadud histogramme. Histogramm sisaldab vähem infot kui dot plot, aga võimaldab paremini tabada seaduspärasid & andmejaotust & outliereid suurte andmekoguste korral.

c. Density plot. Silutud versioon histogrammist, mis kaotab infot aga toob vahest välja signaali müra arvel. Density plotte on hea kõrvuti vaadelda joy ploti abil.

d. Box-plot – sisaldab vähem infot kui histogramm, kuid neid on lihtsam kõrvuti võrrelda. Levinuim variant (kuid kahjuks mitte ainus) on Tukey box-plot – mediaan (joon), 50% IQR (box) ja 1,5x IQR (vuntsid), pluss outlierid eraldi punktidena.

e. Violin plot – informatiivsuselt box-ploti ja histogrammi vahepeal – sobib paljude jaotuste kõrvuti võrdlemiseks

f. Line plot – kasuta ainult siis kui nii X kui Y teljele on kantud pidev väärtus (pikkus, kaal, kontsentratsioon, aeg jms). Ära kasuta, kui teljele kantud punktide vahel ei ole looduses mõtet omavaid pidevaid väärtusi (näiteks X teljel on katse ja kontroll või erinevad valgumutatsioonid, mille aktiivsust on mõõdetud)

g. Suhete võrdlemine (pie vs bar)

h. Cleveland plot countide jaoks. Kasuta Barplotti ainult siis, kui Cleveland plot vm plot mingil põhjusel ei sobi. Barplot võiks olla viimane valik.

   Informatsiooni hulk kahanevalt: iga andmepunkt plotitud (dot plot) -> histogram -> density plot/violin plot -> box plot -> bar plot standardhälvetega -> Cleveland plot (barplot ilma veapiirideta)

Jäta meelde:

1. statistika uurib formaalseid mudeleid, mitte teooriaid ega päris maailma.
2. Statistika jagatakse kahte ossa: kirjeldav ja järeldav (inferential).
3. Kirjeldav statistika kirjeldab teie andmeid summaarsete statistikute ning graafiliste meetodite abil.
4. Järeldav statistika püüab teie andmete põhjal teha järeldusi statistilise populatsiooni kohta, millest need andmed pärinevad
5. Statistika põhiline ülesanne on kvantifitseerida ebakindlust, mis ümbritseb neid järeldusi.

Sõnastik

- Statistiline populatsioon – objektide kogum, millele soovime teha statistilist üldistust. Näiteks hinnata keskmist ravimi mõju patsiendipopulatsioonis. Või Escherichia coli ensüümi X keskmist Kcat-i.
- Valim – need objektid (patsiendid, ensüümiprepid), mida me reaalselt mõõdame.
- Juhuvalim – valim, mille liikmed on populatsioonist valitud juhuslikult ja iseseisvalt. See tähendab, et kõigil populatsiooni liikmetel (kõikidel patsientidel või kõikidel võimalikel ensüümipreparatsioonidel) on võrdne võimalus sattuda valimisse JA, et valimisse juba sattunud liikme(te) põhjal ei ole võimalik ennustada järgmisena valimisse sattuvat liiget.
- Esinduslik valim – Valim on esinduslik, kui ta peegeldab hästi statistilist populatsiooni. Ka juhuvalim ei pruugi olla esinduslik (juhuslikult).
- Statistik – midagi, mis on täpselt arvutatud valimi põhjal (näiteks pikkuste keskmine)
- Parameetri väärtus – teadmata suurus, mille täpset väärtust me saame ainult umbkaudu ennustada aga mitte kunagi täpselt teada. (näiteks mudeli intercept, populatsiooni keskmine pikkus; efekti suurus = katsegrupi keskmine – kontrollgrupi keskmine)
- Statistiline mudel – matemaatiline formaliseering, mis sageli koosneb 2st osast: determinismlik protsessi-mudel pluss juhuslik vea/varieeruvuse-mudel. Protsessi-mudeli näiteks kujutle, et mõõdad mitme inimese pikkust (X muutuja) ja kaalu (Y muutuja). Sirge võrrandiga Y = a + b * X (kaal = a + b * pikkus) saab anda determinismliku lineaarse ennustuse kaalu kohta: kui X (pikkus) muutub ühe ühiku (cm) võrra, siis muutub Y (kaal) väärtus keskmiselt b ühiku (kg) võrra. Seevastu varieeruvuse-mudel on tõenäosusjaotus (näit normaaljaotus). Selle abil modelleeritakse Y-suunalist andmete varieeruvust igal X väärtusel (näiteks, milline on 182 cm pikkuste inimeste oodatav kaalujaotus). Mudel on seega tõenäosuslik: me saame näiteks küsida: millise tõenäosusega kaalub 182 cm pikkune inimene üle 100 kilo. Mida laiem on varieeruvuse mudeli Y-i suunaline jaotus igal X-i väärtusel, seda kehvemini ennustab mudel, millist Y väärtust võime konkreetselt oodata mingi X-i väärtuse korral. Lineaarsete mudelite eesmärk ei ole siiski mitte niivõrd uute andmete ennustamine (seda teevad paremini keerulised mudelid), vaid mudeli struktuurist lähtuvalt põhjuslike hüpoteeside püstitamine/kontrollimine (kas inimese pikkus võiks otseselt reguleerida/kontrollida tema kaalu?). Kuna selline viis teadust teha töötab üksnes lihtsate mudelite korral, on enamkasutatud statistilised mudelid taotluslikult lihtsustavad ja ei pretendeeri tõelähedusele.

- Tehniline replikatsioon – sama proovi (patsienti, ensüümipreparatsiooni, hiire pesakonna liiget) mõõdetakse mitu korda. Mõõdab tehnilist varieeruvust ehk mõõtmisviga. Seda püüame kontrollida parandades mõõtmisaparatuuri/protokolle või siis juba andmete tasemel, statistilise analüüsiga. Näiteks saame andmeid agregeerida ja arvutada keskväärtuse. Kui andmepunkte on piisavalt ja varieeruvus on sümmeetriline ümber tõelise populatsiooniväärtuse, siis annab keksväärtus meile hea hinnangu parameetri tõelisele väärtusele.
- Bioloogiline replikatsioon – erinevaid patsiente, ensüümipreppe, erinevate hiirepesakondade liikmeid mõõdetakse, igaüht üks kord. Eesmärk on mõõta Bioloogilist varieeruvust, mis tuleneb mõõteobjektide reaalsetest erinevustest: iga patsient ja iga ensüümimolekul on erinev kõigist teistest omasugustest. Bioloogiline varieeruvus on teaduslikult huvitav ja seda saab visualiseerida algandmete tasemel (mitte keskväärtuse tasemel) näiteks histogrammina. Teaduslikke järeldusi tehakse bioloogiliste replikaatide põhjal. Tehnilised replikaadid seevastu kalibreerivad mõõtesüsteemi täpsust. Kui te uurite soolekepikest E. coli, ei saa te teha formaalset järeldust kõigi bakterite kohta. Samamoodi, kui te uurite vaid ühe hiirepesakonna/puuri liikmeid, ei saa te teha järeldusi kõikide hiirte kohta. Kui teie katseskeem sisaldab nii tehnilisi kui bioloogilisi replikaate on lihtsaim viis neid andmeid analüüsida kõigepealt keskmistada üle tehniliste replikaatide ning seejärel kasutada saadud keskmisi edasistes arvutustes üle bioloogiliste replikaatide (näiteks arvutada nende pealt uue keskmise, standardhälve ja/või usaldusintervalli). Selline kahe-etapiline arvutuskäik ei ole siiski optimaalne. Optimaalne, kuid keerukam, on panna mõlemat tüüpi andmed ühte hierarhilisse mudelisse.

Tõenäosuse (P) reeglid on ühised kogu statistikale:

- P jääb 0 ja 1 vahele; P(A) = 1 tähendab, et sündmus A toimub kindlasti.
- kui sündmused A ja B on üksteist välistavad, siis tõenäosus, et toimub sündmus A või sündmus B on nende kahe sündmuse tõenäosuste summa — P(A v B) = P(A) + P(B).
- Kui A ja B ei ole üksteist välistavad, siis P(A v B) = P(A) + P(B) – P(A & B).
- kui A ja B on üksteisest sõltumatud (A toimumise järgi ei saa ennustada B toimumist ja vastupidi) siis tõenäosus, et toimuvad mõlemad sündmused on nende sündmuste tõenäosuste korrutis —- P(A & B) = P(A) x P(B).
- Kui B on loogiliselt A alamosa, siis P(B) < P(A)
- P(A | B) —- tinglik tõenäosus. Sündmuse A tõenäosus, juhul kui peaks toimuma sündmus B. P(vihm | pilves ilm) ei ole sama, mis P(pilves ilm | vihm).
- Juhul kui P(B)>0, siis P(A | B) = P(A & B)/P(B) ehk
- P(A | B) = P(A) x P(B | A)/P(B) —- Bayesi teoreem.

Kuigi kõik statistikud lähtuvad tõenäosustega töötamisel täpselt samadest matemaatilistest reeglitest, tõlgendavad erinevad koolkonnad saadud numbreid erinevalt. Kaks põhilist koolkonda on sageduslikud statistikud ja Bayesiaanid.

- Tõenäosus, sageduslik tõlgendus – pikaajaline sündmuste suhteline sagedus. Näiteks 6-te sagedus paljudel täringuvisetel. Sageduslik tõenäosus on teatud tüüpi andmete sagedus, tingimusel et nullhüpotees (H0) kehtib; ehk P(andmed | H0). Nullhüpotees ütleb enamasti, et uuritava parameetri (näiteks ravimiefekti suurus) väärtus on null. Seega, kui P on väike, ei ole seda tüüpi andmed kooskõlas arvamusega, et parameetri väärtus on null (mis aga ei tähenda automaatselt, et sa peaksid uskuma, et parameetri väärtus ei ole null).
- Tõenäosus, Bayesi tõlgendus – usu määr mingisse hüpoteesi. Näiteks 62% tõenäosus (et populatsiooni keskmine pikkus < 180 cm) tähendab, et sa oled ratsionaalse olendina nõus kulutama mitte rohkem kui 62 senti kihlveo peale, mis võidu korral toob sulle sisse 1 EUR (ja 38 senti kasumit). Bayesi tõenäosus omistatakse statistilisele hüpoteesile (näiteks, et ravimiefekti suurus jääb vahemikku a kuni b), tingimusel, et sul on täpselt need andmed, mis sul on; ehk P(hüpotees | andmed).