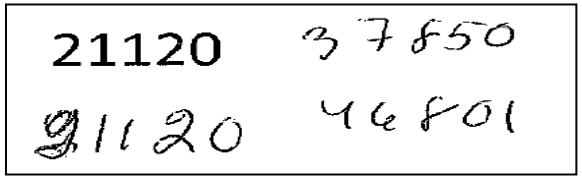
Matemaatikot ja tilastotieteilijät

Matematiikka/tilastotiede ammattina

- Tilastotiede on matematiikan osa-alue, lähinnä todennäköisyyslaskentaa, mutta se on myös itsenäinen tieteenala.
- Tilastotieteen tutkijat kehittävät uusia teorioita ja kaavoja
- Molempia opiskellaan yliopistolla

Matematiikka/tilastotiede ammattina

- Jos opiskelee matematiikkaa tai tilastotiedettä, voi ammatikseen
  - Opettaa
  - Työskennellä esimerkiksi
    - eri tutkimusaloilla, kuten lääketeollisuus, kemian teollisuus
    - vakuutusalalla,
    - kaupan alalla,
    - ohjelmoijana, tietojärjestelmien parissa,
      - Koneoppiminen



    - bioinformatiikka (geeni ym valtavat datamassat)
  - Hyvä työllisyys ja vaihteleva työ

Mitä on tilastollinen päättely?

Esimerkkejä

- Lääketehdas haluaa todistaa, että heidän astmalääkkeensä A on tehokkaampi kuin lääke B
- Tutkitaan voidaanko todistaa, että yskänlääke vähentää kurkkukivun oireita
- Imeytyykö uusi lääke nopeammin kuin vanha lääke?
- Halutaan tutkia, kumpi leikkaustapa A vai B on parempi potilaan toipumisen kannalta
- Voiko liikunnan lisääminen estää lonkkamurtumia?
- Löytääkö uusi allergiatesti luotettavammin allergiat kuin entinen?

Mikä on yhteistä tutkimuksille?

- **Halutaan todistaa** jotakin
  - Tämä tehdään **matemaattisella mallilla = tilastollinen analyysi**

- Täytyy tietää, millä paremmuutta/paranemista ym voidaan **mitata**
  - esimerkiksi lonkkamurtuma, kurkkukipu, pitoisuus veressä (=muuttuja)

- Täytyy arvella, mikä muu voi vaikuttaa asiaan
  - esimerkiksi lonkkamurtumissa ravinto, geeniperimä (=vaikuttavat taustatekijät)

Tutkimuksista yleensä

- **Suunnitellaan** mitä tehdään
  - mikä on tavoite, mitä mitataan, millä mitataan, milloin mitataan => kirjoitetaan tästä suunnitelma
- Tutkimukselle anotaan lupa (mikäli tutkimus liittyy ihmisten tai eläinten terveyteen)
- Suoritetaan tutkimus, **tehdään mittaukset** ja seurannat
- Kerätään ja tallennetaan data
- Tehdään **matemaattiset mallit = tilastolliset analyysit**
- Tehdään **päätelmät** tuloksista
- Kirjoitetaan raportti tai julkaisu

Otos -> päätelmät

- Tutkimuksissa käytännössä aina on **otos**,
  - mittauksia tehdään vähintään muutamasta ihmisestä tai korkeintaan muutamasta tuhannesta
- Kuitenkin halutaan todistaa, että päätelmät voidaan yleistää kaikille eli koko **populaatioon**
  - esimerkiksi kaikkiin suomalaisiin, kaikkiin maailman astmapotilaisiinn

Mitä tunnusluvut ja tilastotiede ylipäätänsä on?

Oikea arvo ja Otos

- Totuus on se Oikea arvo, jota ei ikinä tiedetä
  - Esim suomalaisten miesten keskipituus, pituuden keskihajonta
- Otoksella pyritään saamaan tietoa Oikeasta arvosta
- Mitä Parempi Otos, sen parempi käsitys Oikeasta arvosta tulee

Hyvä Otos

- Onko Otos Hyvä vai onko se Huono?
  - **Otoksen koko** (ja sen suhde hajontaan)
    - **mitä suurempi Otos, sitä parempi käsitys Oikeasta arvosta = pienempi epävarmuus**
    - Esimerkiksi saadaanko arvio, että

      Oikea arvo on välillä 10-1000 vai 510-520
  - **Otoksen edustavuus**
    - Esimerkiksi jos halutaan arvio suomalaisten keskipituudesta, ei voida mitata pelkästään miehiä

Esimerkki

- Ollaan kiinnostuneita *veren hemoglobiinista*
- Hemoglobiini on numeerinen muuttuja

- Kun halutaan kuvailla dataa, lasketaan tunnuslukuja

- Numeerisesta datasta usein lasketaan **keskiarvo ja keskihajonta** (oletetaan normaalijakauma datalle)
- Ne kertovat missä datan arvot ovat keskimäärin ja kuinka paljon ne keskimäärin vaihtelevat

Aritmeettinen keskiarvo

- Jos on data, jossa luvut 121,125,133,134,145,146
- Aritmeettinen **keskiarvo** kertoo mitä suuruusluokkaa yleensä ovat, havaintojen keskimääräisestä sijainnista

$$keskiarvo = \frac{121+125+133+134+145+146}{6} = 134$$

Keskihajonta

- **Keskihajonta** kertoo kuinka paljon keskimäärin jokainen luku poikkeaa keskiarvosta
- Keskihajonta kuvaa **vaihtelun suuruutta**

$$keskihajonta = \sqrt{\frac{(121-134)^2 + (125-134)^2 + \cdots + (146-134)^2}{6-1}} = 10.2$$

Luottamusväli

- Jos halutaan päätellä enemmän hemoglobiinin keskiarvosta populaatiossa datan perusteella, lasketaan **luottamusväli**

- Siihen laskuun vaikuttavat **otoskoko, keskiarvo ja keskihajonta**
  - Jos saadaan keskiarvon 95% luottamusväliksi 115 -145.
  - Voidaan sanoa, että havaitun **datan perusteella, oikea keskiarvo on 95% todennäköisyydellä 115 ja 145 välillä**

- Tämä on jo **tilastollista päättelyä**!

Data analyysi esimerkki – miten todistaminen tehdään?

- Olkoon meillä tutkimus, jossa lumelääke (=Placebo) ja testattava lääke (=Treatment)

- **Halutaan todistaa, että testattavalla lääkkeellä saadaan korkeampi hemoglobiinin keskiarvo**

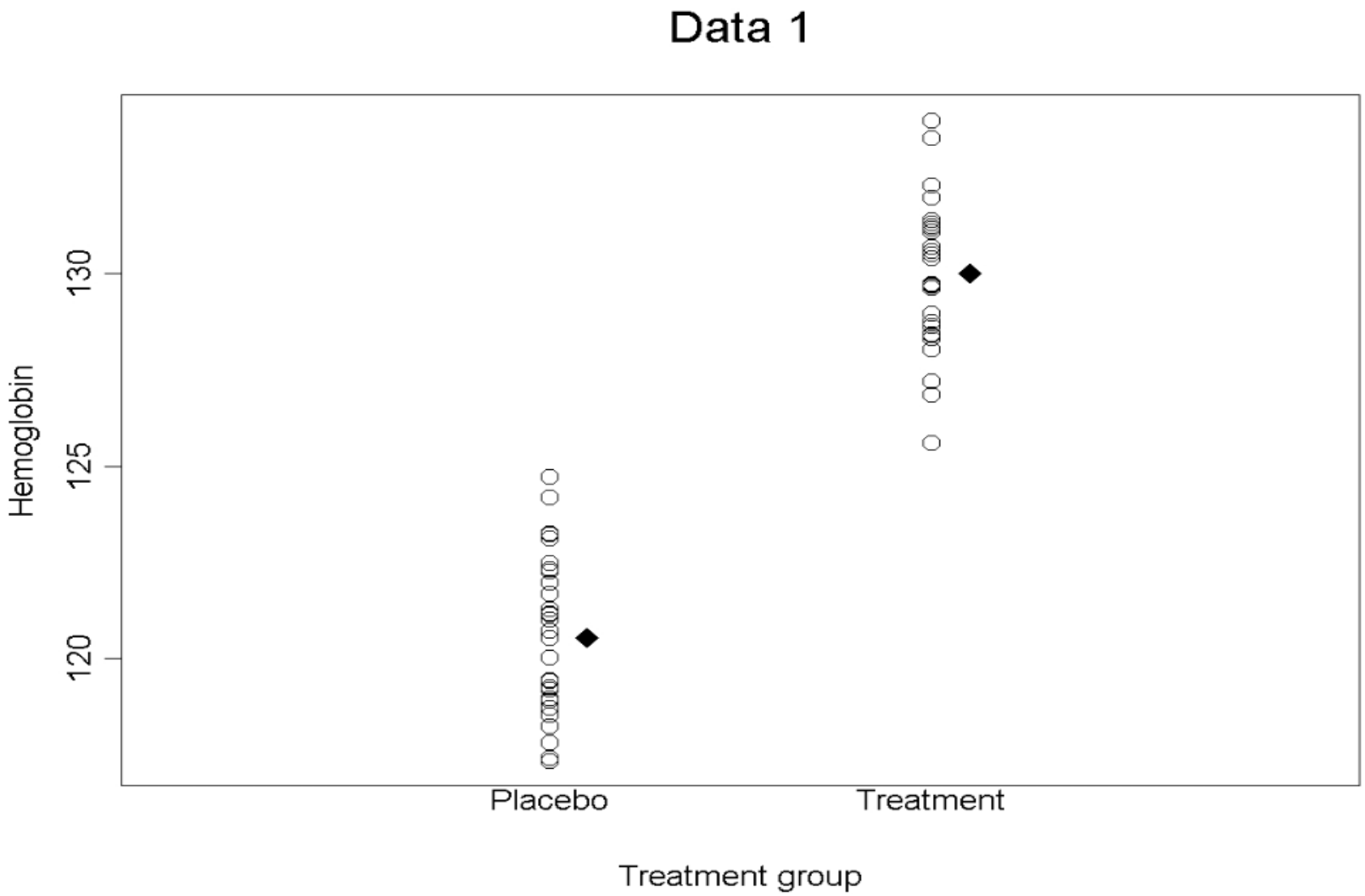
Mitä tehdään?

- Arvotaan puolet henkilöistä lumelääkkeelle ja puolet testattavalle lääkkeelle
  - (=randomisointi, satunnaistaminen)
- Tehdään kaikille samannäköiset pillerit (=sokkoutus)

- Henkilöt syövät pillereitä viikon ajan
- Mitataan hemoglobiini
  - ennen tutkimusta mitataan myös, mutta sitä ei näytetä nyt tässä

- Tutkitaan kolmen erilaisen datan kautta miten päättelyä voi tehdä

Koko data kuvassa: DATA 1



Alustavaa tutkailua

- Kuvasta päättelemme, että lääke vaikuttaa
- Lasketaan tunnuslukuja
  - havaintojen määrä, keskiarvo, mediaani, keskihajonta, luottamusväli, minimi ja maksimi

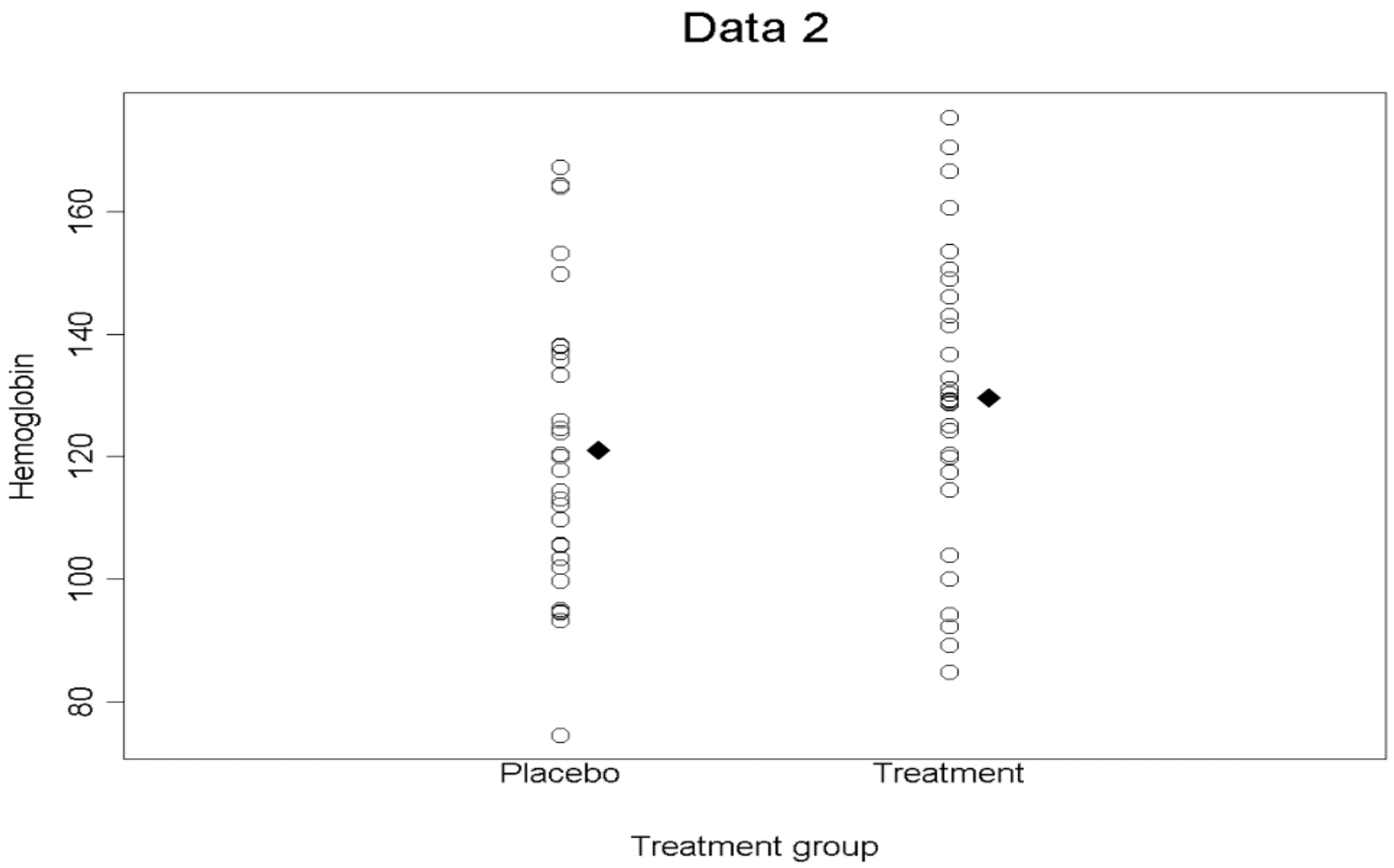
RYHMÄ	N	Keskiarvo	Keski-hajonta	Median	Min	Max	Luottamusvälin alaraja	Luottamusvälin yläraja
Placebo	30	120.5	2.1	120.6	117.3	124.7	119.8	121.3
Treatment	30	130.0	1.9	130.1	125.6	134.0	129.3	130.7

- Myös luottamusvälien perusteella näyttää siltä, että keskiarvot eroavat

Lopullinen testi

- Lopulta voimme tehdä **tilastollisen testin**
  - Otetaan huomioon vähintään ryhmien keskiarvot, ryhmien keskihajonnat ja ryhmien koot
  - Vaikeammissa tapauksissa tehdään paljon vaikeampaa mallinnusta
- **Tutkitaan, kuinka todennäköistä on, että keskiarvot eroavat näin paljon tässä otoksessa vain sattumalta, jos keskiarvot olisivat oikeasti yhtä suuret**
- Jos saamme todennäköisyydeksi 0.0001, uskotko silti että keskiarvot olisivat 'oikeasti' yhtäsuuret?
- Et, eikä, usko sitä enää muutkaan ja näin olet todistanut eron!

Koko data kuvassa: Data 2



Alustavaa tarkastelua

- Kuvan perusteella eroa ei tunnu olevan
- Tunnusluvut

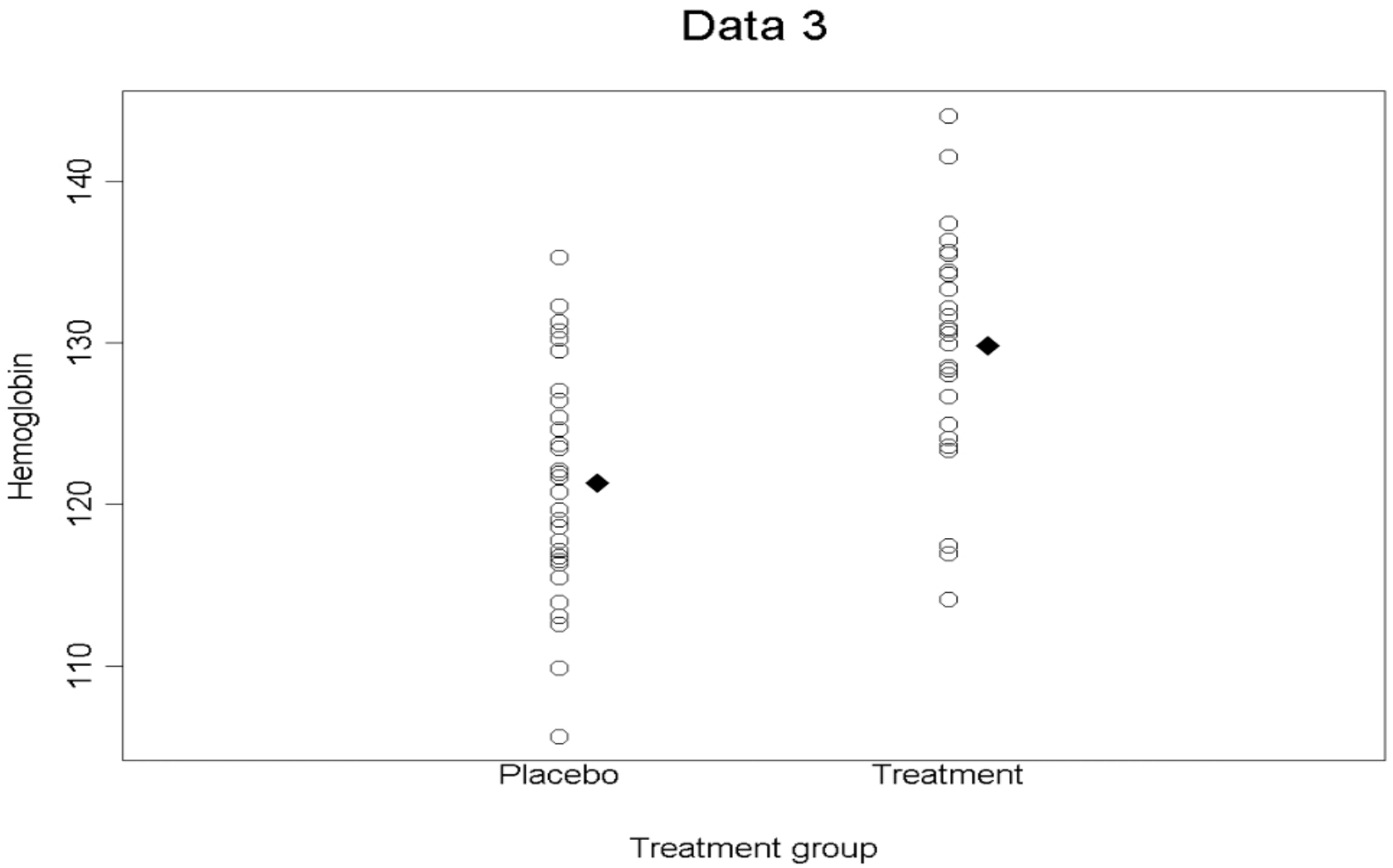
RYHMÄ	N	Keskiarvo	Keski-hajonta	Median	Min	Max	Luottamusvälin alaraja	Luottamusvälin yläraja
Placebo	30	121.0	23.4	118.9	74.5	167.4	112.3	129.8
Treatment	30	129.7	23.9	129.2	84.9	175.4	120.7	138.6

- Keskiarvot ovat melkein samat kuin viimeksi, mutta hajonta on paljon suurempaa kuin aiemmassa datassa!
- Luottamusvälitkin menevät päällekkäin. Pahalta näyttää.

Lopullinen testi

- Lasketaan todennäköisyys sille, että keskiarvot eroavat tämän verran vain sattumalta, jos ne ei oikeasti eroa
- Todennäköisyydeksi saadaan 0.16
- Se on kohtuullisen suuri eli **ei voida päätellä, että lääkkeellä olisi vaikutusta hemoglobinin keskimääräiseen tasoon**
- Tämä päättelyiden ero johtui siis siitä, että hajonta oli nyt niin suurta!
- *Jos hajonta on suuri, tarvitaan isompi otos osoittamaan lääkkeen teho*

Koko data kuvassa: Data 3



- Kuvan perusteella jotain vaikutusta on, mutta riittääkö se todistamaan lääkkeen tehon?
- Tunnusluvut

RYHMÄ	N	Keskiarvo	Keski-hajonta	Median	Min	Max	Luottamusvälin alaraja	Luottamusvälin yläraja
Placebo	30	121.3	7.1	121.2	105.6	135.3	118.6	124.0
Treatment	30	129.8	6.7	130.6	114.1	144.1	127.3	132.3

  - Taas samat keskiarvot, mutta hillitympi hajonta

- Tehdään testi ja todennäköisyys, että keskiarvot olisivat samat, on 0.0001
- Voidaan päätellä, että **lääke tehoaa <u>keskimäärin</u>**

Lääke tehoaa keskimäärin

- Mitä se tarkoittaa?
- Koska kuvissa havainnot menivät päällekkäin?
- Lääke siis nostaa keskiarvoja katsottaessa (=keskimäärin) hemoglobiinitasoa, mutta ei se tarkoita välttämättä, että kaikilla on korkeampi hemoglobiini. Populaatiotasolla keskiarvo on korkeampi!

- **Kuvan perusteella ei voi tehdä päätelmiä vaan tarvitaan matemaattista mallinnusta!**

Lopuksi…

- Tässä oli esimerkkejä hyvin yksinkertaisesta mallinnuksesta, mutta silti nähtiin, että pitää ottaa huomioon monta asiaa yhtä aikaa (**datan määrä, keskiarvot, hajonta, tutkimuksen väite..**datan jakauma, riippuvuus), jotta voi tehdä kunnollisia päätelmiä kerätystä datasta.

- Päätelmien perusteella **voimme muuttaa käsityksiä tutkimusaiheesta**
  - esimerkiksi parempia syöpähoitoja, uusia lääkkeitä