Contents

1 Sattuma ja satunnaisuus

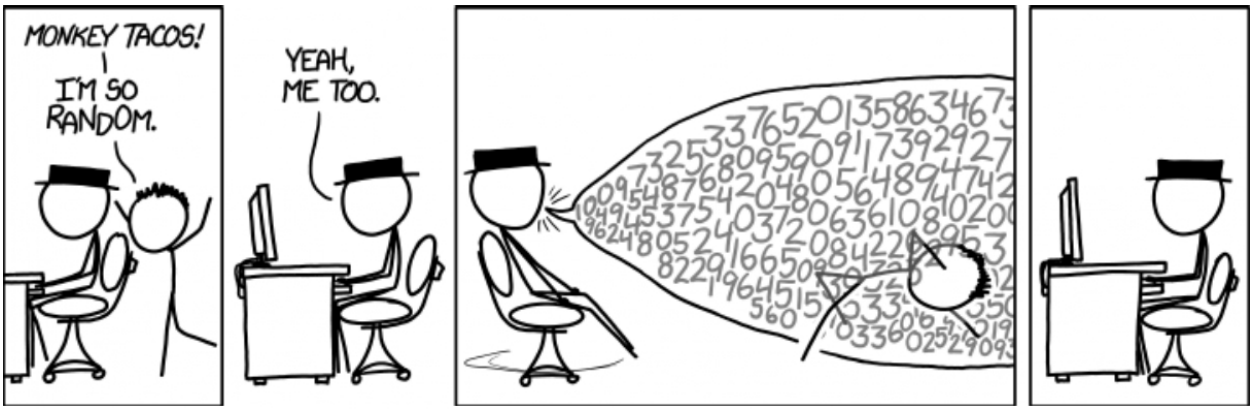

Figure 1: Hauska kuva satunnaisuudesta.

Tässä luvussa pohdimme sattuman ja satunnaisuuden roolia tilastotieteessä ja tieteessä ylipäätään. Satunnaisuudella tarkoitetaan yleensä säännönmukaisuuden puuttumista ja ennustamattomuutta ja kenties juuri siksi sitä voidaan pitää yhtenä maailman vaikuttavammista ilmiöistä. Jokainen haluaisi tietää mitä tuleman pitää ja siksi sattuma on myös filosofisesti mielenkiintoinen: se vaikuttaa ja muokkaa niin meitä itseämme kuin ympäröivää maailmaa mitä merkityksellisimmin tavoin - joskus jopa vasten tahtoamme ja usein vailla täyttä ymmärtystämme!

Ihmisen oma kokemus on kuitenkin altis kaikenlaisille virhepäätelmille, joita kutsutaan myös kognitiivisiksi vinoumiksi. Haluamme löytää systematiikkaa ja tarkoitusta kaaoksesta sekä merkityksiä ja

syy-seuraussuhteita sellaisista tapahtumista, jotka kuuluvat normaalivaihtelun piiriin. Tällaisissa tilanteissa usein tilastollinen tarkastelu paljastaakin ilmiön todellisen, alkuperäisestä kuvitelmasta poikkeavan luonteen. Osatakseen erottaa systemaattisen vaihtelun ja ymmärtääkseen oikeasti merkityksellisiä syy-seuraussuhteita, on välttämöntä ymmärtää satunnaisuutta. Tämä välttämättömyys pätee erityisesti tiedeyhteisön jäseniin, jotka pyrkivät tutkimaan ympäröivän maailman satunnaisia ilmiöitä. Tilastotiede perustuu satunnaisilmiöiden ja satunnaisen aineiston tutkimiseen, joten sen ymmärtäminen on keskeisessä roolissa tieteen ja maailman ymmärtämisessä.

1.1 Satunnaisilmiöt ja satunnaismuuttujat tilastotieteessä

- Edellisestä luvusta muistamme, että tilastotieteellisen tutkimuksen kohteena on aina jokin tilastoyksikköjen tutkimusmuuttujista koostuva havaintoaineisto, jonka pohjalta tehdään päätelmiä perusjoukosta/populaatiosta.
- Nämä tilastolliset muuttujat tulkitaan satunnaisiksi, ja täten tilastollisen tutkimuksen tavoite on tutkia satunnaisilmiötä, joka on generoinut nämä havaitut eli toteutuneet arvot.
  - Yksi tilastotieteen olennainen tehtävä onkin kehittää **tilastollisia malleja**, joiden avulla satunnaisilmiöitä voidaan kuvata, selittää ja ennustaa.
  - Tilastollisen mallin satunnaisten piirteiden kuvaus perustuu johonkin **todennäköisyysmalliin**.

Satunnaisilmiö
Reaalimaailman ilmiö on satunnaisilmiö, jos seuraavat ehdot pätevät:

- Ilmiöllä on useita erilaisia tulosvaihtoehtoja.
- Sattuma määrää mikä tulosvaihtoehdoista toteutuu, eli yksittäistä tulosta ei voida tietää etukäteen.
- Vaikka tulos vaihtelee ilmiön toistuessa satunnaisesti, käyttäytyy tulosvaihtoehtojen suhteellisten osuuksien jakauma tilastollisesti stabiilisti ilmiön toistokertojen lukumäärän kasvaessa.

- **Tilastollisella stabiiliudella** tarkoitetaan sitä, että on mahdollista arvioida kuinka **todennäköisiä** erilaiset tapahtumat, eli satunnaisilmiön tulosvaihtoehdot ovat.
  - Toisin sanoen satunnaisilmiön tulosvaihtoehtoihin on liityttävä säännönmukaisuutta, jonka on tultava esille ilmiön toistuessa.

Esimerkkejä satunnaisilmiöistä uudistettava...

- Kvanttimekaniikan ja hiukkasfysiikan ilmiöt ovat perusluonteeltaan satunnaisia.
- Luonnontieteellisiin mittauksiin liittyvien mittausvirheiden syntymekanismit ovat (ainakin osittain) satunnaisprosesseja.
- Uhkapeleissä kuten arpajaisissa, lotossa, ruletissa, korttipeleissä ja noppapeleissä sattumalla on keskeinen rooli.
- Perinnöllisyys noudattaa sattuman lakeja.
- Eliöiden ominaisuuksien jakautuminen populaatiossa on satunnaista.
- Ihmisten, ihmisryhmien ja ihmisten muodostamien organisaatioiden sosiaalisessa ja taloudellisessa käyttäytymisessä on monia satunnaisia elementtejä.
- Teknisten prosessien tuloksien ominaisuudet jakautuvat satunnaisesti.

Satunnaismuuttujat

- Satunnaisilmiöitä koskevan tutkimuksen kohteena olevat tilastolliset muuttujat tulkitaan **satunnaismuuttujiksi** ja havainnot (havaintoarvot) voidaan näin ollen tulkita näiden satunnaismuuttujien realisoituneiksi arvoiksi. Satunnaismuuttuja siis kuvaa tarkasteltavan mitattavan ominaisuuden (satunnais)vaihtelua tutkimuksen kohteiden, eli tilastoyksiköiden joukossa.
  - Mitattavan ominaisuuden mahdolliset arvot määräävät satunnaismuuttujan luonteen. Yleisesti satunnaismuuttujat jaetaan kahteen luokkaan: **jatkuviin** ja **diskreetteihin**.
  - Satunnaismuuttujan **todennäköisyysjakauma**, määrää erilaisten tulosvaihtoehtojen todennäköisyyden ja mahdollistaa täten tilastollisen analyysin ja päättelyn.
    * Satunnaisuus eroaa mielivaltaisesta prosessista siinä, että satunnaista ilmiötä voidaan kuvata jollakin **tilastollisella lailla** kun taas mielivaltaista prosessia ei.

Satunnaismuuttuja
Satunnaismuuttuja (usein lyhyesti sm., englanniksi random variable, merkitään esim. Y, ja kutsutaan ajoittain myös stokastiseksi muuttujaksi) on todennäköisyyslaskennan peruskäsite, jolla tarkoitetaan satunnaisilmiön määräämää lukua.

- Satunnaismuuttujan Y realisoituvaa arvoa y kutsutaan realisaatioksi tai toteumaksi.
- Tilastollinen aineisto muodostuu useiden satunnaismuuttujien (tilastoyksiköiden tutkimusmuuttujien) realisoituneista arvoista.
- Realisoituneiden arvojen vaihtelua tilastoyksiköiden välillä kutsutaan satunnaisvaihteluksi.

Jatkuvat ja diskreetit satunnaismuuttujat

- Satunnaismuuttuja Y on jatkuva, jos se voi saada ylinumeroituvan määrän arvoja tai ts. minkä tahansa arvon joltain väliltä, kuten tyypillisesti minkä tahansa arvon joltain reaalilukuväliltä.
- Satunnaismuuttuja Y on diskreetti, jos se voi saada vain joitain mahdollisia arvoja (vain yksittäisiä, äärellisen tai numeroituvasti äärettömän määrän, arvoja). Yksinkertaisimmillaan diskreetti satunnaismuuttuja Y on kaksiarvoinen (binäärinen), jolloin sen mahdollisia arvoja tyypillisesti merkitään $y = 0$ tai $y = 1$.

Esimerkki: satunnaismuuttuja
Ihmisen pituutta voidaan pitää (ennen mittaukseen tulemista) satunnaismuuttujana Y ja lopullista pituutta täten pituuden realisaationa y. Pituutta kohdellaan jatkuvana muuttujana senttimetreissä, mutta mikäli määritetään toteumaksi jonkin pituuden raja-arvon, esimerkiksi 170cm, ylittävä pituus, on kyseessä kaksiarvoinen (binäärinen) satunnaismuuttuja (pituus on joko yli tai alle 170 cm).

- Muuttujat voidaan luokitella myös **kvalitatiivisiin** ja **kvantitatiivisiin** muuttujiin.
  - Kvalitatiivisiin muuttujiin liittyy luokittelu- tai järjestysasteikko
  - Kvantitatiivisiin muuttujiin välimatka- ja suhdeasteikko.
- Tilastolliset menetelmät perustuvat todennäköisyyslaskennan[1] tuloksiin ja tarjoavat keinon hallita satunnaisuuden aiheuttamaa epävarmuutta sekä tavan erottaa systemaattinen ja satunnainen vaihtelu, eli signaali ja kohina, toisistaan.

[1] Todennäköisyyslaskentaa käsitellään välillisesti tulevissa luvuissa mutta varsinaisesti tarkemmin 2. periodin kurssilla TILM3553 Todennäköisyyslaskennan peruskurssi ja (erityisesti sivuaineopiskelijoille) TILM3568 Todennäköisyyslaskenta sivuaineopiskelijoille.

- Tilastollisen aineiston **tilastollisella mallilla** tarkoitetaan täten niiden satunnaismuuttujien todennäköisyysjakaumaa, jonka ajatellaan generoineen havainnot.
  - Yksinkertaisimmillaan esimerkiksi yksinkertaiseen satunnaisotantaan takaisinpanolla perustuva satunnaismalli (palaamme tähän otantaa käsittelevässä luvussa 2).
  - Satunnaisuus perustuu siihen, että satunnaismuuttujien toteutuvat arvot (ja niistä lasketut tunnusluvut kuten keskiarvo) vaihtelevat satunnaisesti otoksesta toiseen.
- Todennäköisyyslaskennan tehtävä on tuottaa **matemaattisia ja tilastollisia malleja** satunnaisilmiöissä havaittavalle tilastolliselle stabiliteetille.

1.2 Tilastotieteen suhde satunnaisuuteen ja todennäköisyyksiin

- Tilastotieteessä **tutkimusaineiston keräämistä** voidaan pitää hyvänä esimerkkinä satunnaisilmiöstä.
  - Voimme ajatella, että tilastollisen tutkimuksen kohteet on aina valittu arpomalla.
  - Arvonta on mainio esimerkki satunnaisilmiöstä, sillä siihen liittyy aina ennustamattomuutta: vaikka yksittäisen arvonnan tulosta ei voi tietää etukäteen, noudattaa se kuitenkin todennäköisyyden lakeja.
  - Koska arvonnan tulos vaihtelee satunnaisesti arvontakerrasta toiseen, myös tutkimuksen kohteita kuvaavat tiedot vaihtelevat satunnaisesti arvontakerrasta toiseen.
  - Tutkimuksen kohteita kuvaavien tietojen käyttäytymisessä havaitaan kuitenkin arvontaa toistettaessa juuri sitä säännönmukaisuutta, jota kutsutaan tilastolliseksi stabiliteetiksi. **Tämä säännönmukaisuus on tilastollisen tutkimuksen kohde.**
- Esimerkkejä tilastollisten aineistojen keräämisen menetelmistä, jotka perustuvat arvontaan:
  - **Satunnaistetut kokeet**: Kokeellisessa tutkimuksessa tavoitteena on vertailla erilaisten käsittelyiden vaikutuksia kokeen kohteisiin. Erilaisten virhelähteiden kontrolloimiseksi käsittelyt on syytä arpoa kohteille.
  - **Satunnaisotanta**: Otannalla[2] tarkoitetaan laveasti tutkimusaineistojen keräämisen menetelmiä. Erilaisten virhelähteiden kontrolloimiseksi tutkimuksen kohteet on syytä valita arpomalla. (Ks. Luku 2)
- Kerätyn (tai havaitun) aineiston pohjalta tehdään päätelmiä sen generoineesta satunnaisilmiöstä esimerkiksi testaamalla erilaisia siihen liittyviä hypoteeseja.
  - Tilastotiede voidaan jakaa kahteen suureen paradigmaan sen mukaan, miten tilastolliseen päättelyyn, ml. hypoteeseihin ja niiden testaamiseen, suhtaudutaan. Näitä ovat **klassinen eli frekventistinen tilastotiede** sekä **Bayesilainen tilastotiede**. Tarkastellaan seuraavaksi minkälaisia eroja ja yhtäläisyyksiä näiden koulukuntien välillä on.

Frekventistinen tilastotiede

- Klassisessa eli frekventistisessä tilastotieteessä ajatellaan että hypoteesien testaaminen tulee perustua yksinomaan havaittuun aineistoon ja siihen liitettävään tilastolliseen malliin.
- Nimi "frekventistinen" juontuu siitä, että tämä todennäköisyysjakauma määrittää satunnaismuuttujan mahdollisten arvojen todennäköisyydeksi niiden suhteellisen osuuden äärettömästä määrästä realisaatioita, ts. niiden suhteellisen frekvenssin.
- Klassisessa tilastotieteessä havaittuun aineistoon *sovitetaan* sitä kuvaavaan todennäköisyysjakaumaan perustuva tilastollinen malli, joka vastaa saatua aineistoa parhaiten.

[2] Erityisesti erilaisten otantamenetelmien yhteydessä, joita tarkastellaan tarkemmin luvussa 2.

  - Tämä tilastollinen malli perustuu nk. **uskottavuusfunktioon**, joka on *aineiston* sekä yhden tai useamman *parametrin* funktio ja joka saavuttaa suurimman arvonsa nk. "suurimman uskottavuuden pisteessä".
  - Uskottavuusfunktio kertoo kuinka todennäköisenä havaittua aineistoa voidaan pitää, mikäli sen oletetaan olevan peräisin vastaavasta mallista jollain parametriarvolla. Täten ne parametriarvot, joilla uskottavuusfunktion arvo maksimoituu, kuvaavat aineiston generoimaa prosessia parhaiten, annettuna malli- eli jakaumaoletus.
  - Uskottavuusfunktioista, tilastollisten mallien estimoinnnista ja parametreista lisää seuraavassa alaluvussa sekä luvussa **??**.

- Perusjoukkoa koskevia hypoteeseja testataan tämän tilastollisen mallin avulla: havaittu aineisto määrittää uskottavuusfunktion perusteella sellaiset hypoteesit, jotka jäävät joko voimaan tai tulevat hylätyiksi.

- Klassisessa tilastotieteessä hypoteesien testaus perustuu siis vain aineistoon eli tilastollinen päättely on induktiivista: aineiston avulla otosta koskeva päätelmä voidaan yleistää koskemaan perusjoukkoa.

  - Toki kaikki päättely on alisteista tehdyille oletuksille koskien käytettävää tilastollista mallia.

Bayesilainen tilastotiede

- Bayesilainen tilastotiede on tilastotieteen toinen suuri paradigma ja on saanut nimensä englantilaiselta harrastelijamatemaatikko ja presbyteeripappi Thomas Bayesilta, jota pidetään Bayesilaisen tilastotieteen isänä.
- Bayesilainen tilastotiede ulottaa todennäköisyyskäsityksen, eli tn-jakauman, myös aineistoa koskevien hypoteesien puolelle: kuinka todennäköisenä jotain hypoteesia voidaan pitää jo ennen tutkimusaineiston keräämistä?
  - Myös Bayesilaisessa tilastotieteessä hyödynnetään uskottavuusfunktiota, mutta hypoteesien testaus ei perustu niinkään frekventistiseen ajatukseen todennäköisyyksistä suhteellisina osuuksina äärettömässä sarjassa.
  - Bayesilaiset perustavat sen sijaan hypoteesien testaamisen tutkimuskysymystä koskevien ennakkokäsitysten päivittämiselle sen jälkeen, kun aineiston on havaittu.
  - Nämä ennakkokäsitykset voidaan kuvata todennäköisyysjakaumana, priorijakaumana, jota päivitetään ns. posteriorijakaumaksi kun aineisto havaitaan. Näin päättely perustuu priorijakauman ja aineiston uskottavuusfunktion väliselle kompromissille!
- Ajatusta ennakkokäsityksistä todennäköisyyksinä käytetään niin Bayesilaisen tilastotieteen kritiikkinä kuin puolustuksena.
  - Lopulta olemme kaikki Bayesilaisia: jokaisella on sisäisiä ennakkokäsityksiään, myös tutkijoilla! Nämä ennakkokäsitykset voivat perustua esimerkiksi aiempaan tutkittuun tietoon, mutta myös uskomuksiin.
  - Prioritiedon hyödyntäminen tilastollisessa tutkimuksessa on usein perusteltua.
  - Bayesilaista tilastotiedettä tarkastellaan tarkemmin esimerkiksi kursseilla TILM3577 Bayes-päättely sekä TILM3601 Bayes-laskenta.

1.3 Tilastolliset mallit, jakaumat ja parametrit

- Tilastolliset mallit perustuvat satunnaismuuttujan mahdollisten tulosvaihtoehtojen todennäköisyyksiä kuvaavalle **todennäköisyysjakaumalle**, joka määrää millä todennäköisyydellä satunnaismuuttuja saa erilaisia arvoja.

- Toisaalta ajoittain tietyn suureen/ilmiön mallinnuksessa voidaan perustellusti käyttää molempiin luokkiin kuuluvien satunnaismuuttuja- ja tilastollisen mallityypin vaihtoehtoja.
  - Esimerkki: Esimerkiksi COVID19-tartuntatapausten lukumäärä Suomessa on periaatteessa diskreetti satunnaismuuttuja, joka saa yksittäisen (kokonaisluku)arvon joka kuukausi, mutta käytännössä lukumäärät ovat tässä tapauksessa sen verran suuria, että niitä (saatetaan) mallintaa jatkuva-arvoisena muuttujana.
  - Vastaavasti esimerkiksi potilaan jonotusaika päivystyksessä voi periaatteessa saada minkä tahansa arvon tietyltä reaalilukuväliltä (tällöin käytettäisiin jatkuviin sm:jiin perustuvia tilastollisia menetelmiä).
- Satunnaismuuttujan mahdolliset arvot, ja täten todennäköisyysjakauma, määräävät myös käytettävän tilastollisen mallin.
  - **Diskreetin satunnaismuuttujan** jakauma voidaan usein esittää taulukkomuodossa. Eri arvojen todennäköisyydet muodostavat kyseisen satunnaismuuttujan todennäköisyysjakauman (**pistetodennäköisyysfunktion**), jota voidaan havainnollistaa esimerkiksi pylväsdiagrammilla.
  - Jatkuvan satunnaismuuttujan Y arvot muodostavat jonkin reaaliakselin välin, joka sisältää äärettömän määrän lukuja. Tämän vuoksi jatkuvan satunnaismuuttujan jakauman esittäminen taulukossa ei ole luontevaa, vaan jakauma esitetään yleensä satunnaismuuttujan **tiheysfunktion** avulla.
    - Pistetodennäköisyys- ja tiheysfunktio siis määräävät satunnaismuuttujan mahdollisille arvoille todennäköisyydet väliltä $[0, 1]$ ja näin voidaan arvioida havaitun aineiston uskottavuutta ja testata siihen liitettäviä hypoteeseja suhteessa estimoituun suuriman uskottavuuden estimaattiin.
- Tilastolliset mallit approksimoivat "todellista" aineiston generoinutta ilmiötä. Tilastolliset mallit riippuvat **parametreista** ja keskeinen oletus erityisesti klassisessa tilastotieteessä on, että aineiston generoinutta satunnaisilmiötä kuvaa jokin vakioinen mutta tuntematon parametriarvo (tai niiden joukko).

Parametrien estimointi ja niiden testaus

- Satunnaisilmiötä kuvaava tilastollinen malli perustuu siis johonkin parametriseen todennäköisyysjakaumaan, joka yhdessä havaintojen kanssa määrittää uskottavuusfunktion.
  - Aineistoa kuvaavan tilastollisen mallin uskottavuus pyritään maksimoimaan, mikä tarkoittaa valitun todennäköisyysjakauman sovittamista havaintoaineistoon mahdollisimman hyvin.
  - Tässä nk. "suurimman uskottavuuden estimoinnissa" aineiston generoiman (oletetun) todennäköisyysjakauman parametriarvot **estimoidaan** (eli arvioidaan) käytettävän otoksen/aineiston avulla.
  - Perusjoukkoa parhaiten kuvaavan (eli "aineiston generoineen") parametrin arvo pyritään siis estimoimaan aineiston perusteella.
- Parametrien estimoinnin lisäksi usein **testataan** parametreja koskevia oletuksia (eli hypoteeseja).
- Estimointi ja testaus ovat tilastolliseen tutkimukseen liittyvän **tilastollisen päättelyn** keskeisiä välineitä, joiden avulla tutkittavasta ilmiöstä pyritään tekemään johtopäätöksiä siitä kerätyn havaintoaineiston perusteella.
  - Estimoitujen parametrien testaus voi vastata esimerkiksi seuraavanlaisiin kysymyksiin:
    - Onko suomalaisten miesten keskipituus 180cm?
    - Vaikuttaako yliopistokoulutus tulevaisuuden ansioihin?
    - Auttaako tietty lääkeaine jonkin sairauden hoidossa?
    - Voiko osakemarkkinoiden tuottoja ennustaa?
- Parametrien testaus on osa tilastollista päättelyä, johon palataan tarkemmin luvussa **??**

- Hevosen potkuun kuolleiden Preussin armeijan sotilaiden lukumäärä 20 vuoden aikana
- Guinnes -oluen valmistusprosessin hiivasolujen lukumäärä
- Bakteerien lukumäärä litrassa järvivettä
- Viimeisen 10 vuoden lento-onnettomuuksien lukumäärä

- Kaikille yhteistä: lasketaan ***harvinaisten tapahtumien lukumäärä*** tietyssä ajassa tai tilavuudessa
- Jakaumalla ***parametrit***, joiden arvot vaihtelevat ja jotka halutaan estimoida

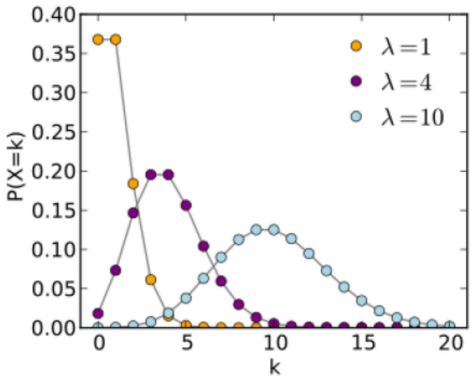

Figure 2: Esimerkki: Poisson-jakauman sovelluskohteita ja sen pistetodennäköisyysfunktio eri parametrin arvoilla. Poisson-jakaumaa esitellään tarkemmin alaluvussa 4.5.

1.4 Odotusarvo ja varianssi

- Satunnaismuuttujan todennäköisyysjakauman tietoa voidaan tiivistää tunnuslukuihin, joista keskeisimpiä ovat **odotusarvo**, **varianssi** ja **keskihajonta**.

Odotusarvo
Satunnaismuuttujan Y odotusarvo $\mathrm{E}(Y)$ kuvaa satunnaismuuttujan odotettavissa olevaa arvoa.

- Muodostamalla satunnaiskokeen tulosten **painotettu keskiarvo**, jossa kunkin tuloksen painona on vastaavan tapauksen todennäköisyys, niin saatua arvoa sanotaan odotusarvoksi $\mathrm{E}(Y)$.
- Odotusarvo kuvaa jakauman painopistettä.
- Merkinnän $\mathrm{E}(Y)$ käyttö juontaa juurensa englannin kielen sanoihin "odotus", expectation, ja 'odotusarvo', expected value.

Esimerkki: Odotusarvo
tähän joku esimerkki tosiaan.

- Odotusarvon lisäksi kiinnostuksen kohteena on usein jakauman keskittyneisyys (hajaantuneisuus). Ts. kun halutaan puolestaan kuvata satunnaismuuttujan arvojen vaihtelua, tutkitaan todennäköisyysjakauman **varianssia** ja **keskihajontaa**.

Varianssi
Satunnaismuuttujan Y hajontaa voidaan mitata varianssilla

$$\mathrm{Var}(Y) = \mathrm{E}\Big[\Big(Y - \mathrm{E}(Y)\Big)^2\Big],$$

tai sen neliöjuuren eli **keskihajonnan** avulla

$$\mathrm{D}(Y) = \sqrt{\mathrm{Var}(Y)}.$$

- Mitä lähempänä nollaa keskihajonta ja varianssi ovat, sitä todennäköisempää on, että satunnaismuuttujan arvo on lähellä odotusarvoa. - Merkintöjen $\mathrm{Var}(Y)$ ja $\mathrm{D}(Y)$ taustalla on englannin kielen sanat variance (varianssi) ja deviation, joka tarkoittaa poikkeamaa, hajontaa.

- Odotusarvon ja varianssin (keskihajonnan) tavanomaiset estimaattorit ovat otoskeskiarvo ja otosvarianssi (otoshajonta), joihin palataan vielä myöhemmin.

1.5 Joitain jakaumia

Tarkastellaan seuraavassa muutamia keskeisiä tilastollisia jakaumia. Esittelemme ensin keskeisintä jatkuvien satunnaismuuttujien jakaumaa, normaalijakaumaa, ennen muutamien diskreettien satunnaismuuttujien jakaumia.

1.5.1 Normaalijakauma

- Jos satunnaismuuttuja Y noudattaa **normaalijakaumaa** odotusarvolla $\mathrm{E}(Y) = \mu$ ja varianssilla $\mathrm{Var}(Y) = \sigma^2$, niin tällöin merkitään $Y \sim \mathrm{N}(\mu, \sigma^2)$.
- Y:n tiheysfunktio on muotoa (ks. kuva alla)

$$f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2},$$

  jossa e viittaa Neperin lukuun $e \approx 2,71828$

  - Ylläoleva tf. määrittelee parven normaalijakaumia kun parametreille (vakioille) μ ja σ^2 annetaan erilaisia arvoja. Nämä kaksi parametria määräävät normaalijakauman tarkemman muodon.

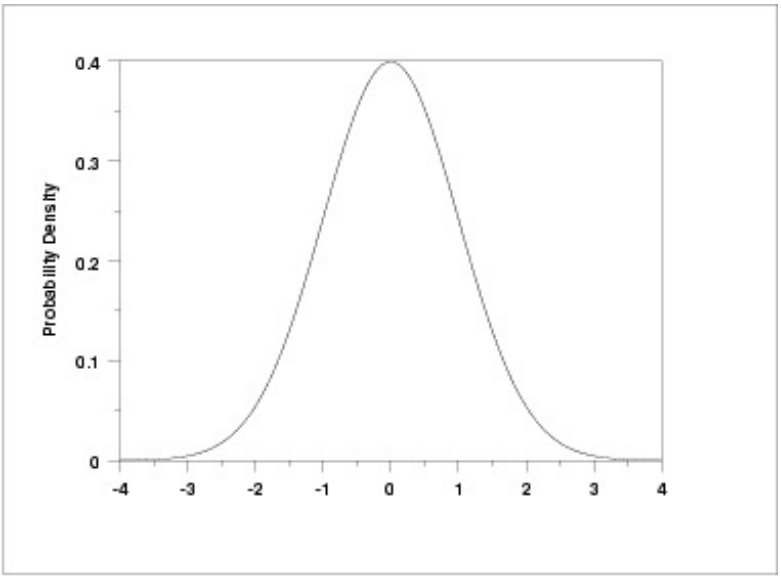

Figure 3: Normaalijakauma

Esimerkki: Miesten pituus

- Tutkitaan miesten pituutta hyvin määritellyssä joukossa, kuten varusmiespalvelusta tiettynä vuonna suorittavien joukossa.
  - Pituus on ominaisuus, jonka voidaan nähdä määräytyvän monista perintö- ja ympäristötekijöistä. Pituutta voidaan siis pitää satunnaismuuttujana.
  - Oletetaan, että pituus noudattaa normaalijakaumaa. Näin ollen sm. Y on valitun miehen pituus ja $Y \sim \mathrm{N}(\mu, \sigma^2)$.
- Tuntemattomien parametrien μ ja σ^2 tulkinta:
  - Odotusarvo $\mu = \mathrm{E}(Y)$ on satunnaisesti valitun miehen pituuden odotettavissa oleva arvo.
  - Varianssi $\sigma^2 = \mathrm{Var}(Y) = \mathrm{E}\left[\left(Y - \mu\right)^2\right]$ kuvaa valitun miehen pituuden odotusarvostaan määrätyn poikkeaman (keskihajonnan) neliön odotettavissa olevaa arvoa (kuvaten ts. pituuksien jakauman keskittyneisyyttä/hajaantuneisuutta pituuksien odotusarvon ympärillä).

1.5.2 Bernoulli-, binomi- ja Poisson-jakauma

- **Bernoulli-jakauma** on todennäköisyysjakauma, jossa satunnaismuuttujalla Y on kaksi mahdollista tulosvaihtoehtoa $Y = 1$ tai $Y = 0$.
  - Yleensä $Y = 0$ tarkoittaa, että jokin tapahtuma ei tapahdu ja $Y = 1$ että tapahtuu.
  - Todennäköisyys tapahtumalle $Y = 1$ on $\mathrm{P}(Y = 1) = p$ ja vastaavasti vastatodennäköisyys $\mathrm{P}(Y = 0) = 1 - p$.
  - Bernoulli-jakaumaa merkitään $Y \sim B(p)$, jossa siis $0 < p < 1$.
  - Bernoulli-jakauman **pistetodennäköisyysfunktio** on muotoa

$$f(y; p) = \mathrm{P}(Y = y) = p^y(1 - p)^{(1-y)},$$

jossa y on sm:n Y realisaatio (havaittu arvo) ja parametri p on tuntematon (voidaan estimoida otoksen avulla).

- Bernoulli-jakauman odotusarvo $\mathrm{E}(Y) = p$ ja varianssi $\mathrm{Var}(Y) = p(1 - p)$.

- **Binomijakauma**
  - Olkoon $Y_1, \ldots, Y_n$ riippumattomia satunnaismuuttujia ja $Y_i \sim B(p)$, $i = 1, \ldots, n$.
  - Jos $X = Y_1 + Y_2 + \ldots + Y_n$, niin $X \sim \mathrm{Bin}(n, p)$. Ts. sm. X noudattaa **binomijakaumaa** parametrein n ja p.
  - Pistetodennäköisyysfunktio:

$$\mathrm{P}(X = k) = \binom{n}{k} p^k (1 - p)^{(n-k)}.$$

- Jakauman odotusarvo $\mathrm{E}(X) = np$ ja varianssi $\mathrm{Var}(X) = np(1 - p)$.
- Binomijakaumalla kyetään vastaamaan mm. kysymykseen millä todennäköisyydellä n:n kokoisessa otoksessa tapahtuu k onnistumista.

Esimerkki: Miesten lukumäärä Saksin osavaltion perheissä 1876–1885[a]
Vuosien 1876–1885 aikana Saksin osavaltiossa rekisteröitiin yli neljä miljoonaa syntynyttä lasta. Tällöin vanhempien tuli ilmoittaa lapsen sukupuoli (mies tai nainen) heidän syntymätodistuksessaan. Myöhemmässä tutkimuksessa tutkittiin tarkemmin 6115 perhettä, joissa asui 12 lasta ja tarkemmin miesten (poikien) lukumäärää näissä perheissä.
Seuraavassa taulukossa taulukoidaan miesten (poikien) lukumäärät näissä 12 lapseen perheissä:
|Miesten lkm. (k)|0|1|2|3|4|5|6|7|8|9|10|11|12|
|Perheiden lkm. (n_k)|3|24|104|286|670|1033|1343|1112|829|478|181|45|7|
Tarkasteltava jakauma esitetään vielä erikseen allaolevassa kuviossa.

Tässä tilantessa mielenkiinnon kohteena saattaisi olla hypoteesi, jonka mukaan pojan (miehen) syntymätodennäköisyys $\mathrm{P}(\text{mies}) = p$ on $p = 0.5$.

[a]Ks. tarkemmin esimerkki 3.2 kirjassa (s. 67-68) Friendly, M., ja D. Meyer (2015). *Discrete Data Analysis with R. Visualization and Modeling Techniques for Categorical and Count Data.* Chapman & Hall/CRC.

Poisson-jakauma

- Jos satunnaismuuttuja Y on Poisson-jakautunut, merkitään $Y \sim P(\lambda)$, jossa parametri $\lambda > 0$ on Poisson-jakauman parametri, jota kutsutaan myös ajoittain intensiteettiparametriksi.
- Poisson-jakaumaa voidaan käyttää tilanteissa, joissa sm. Y on jokin lukumäärä ja sen pistetodennäköisyysfunktio on muotoa
$$\mathrm{P}(Y = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$
- Odotusarvo ja varianssi ovat Poisson-jakauman tapauksessa samat: $\mathrm{E}(Y) = \mathrm{Var}(Y) = \lambda$.

Esimerkki: Poisson-jakauma
Tarkastellaan Englannin Valioliigakauden 1995–1996 otteluissa tehtyjä maalimääriä. Valioliiga (The F.A. Premier League) on korkein Englannin jalkapalloliigan sarjataso, jossa ensi kerran juuri kaudella 1995-1996 20 joukkuetta (aiemmin Valioliigan perustamisen kauden 1992–1993 alussa 22 joukkuetta) pelasivat keskenään kerran toisiaan vastaan koti- ja vieraskentällä. Otteluita oli siis yhteensä 380.
Tämä esimerkki perustuu edellä mainittuun Friendlyn ja Meyerin (2015) kirjan esimerkkiin 3.9 (s. 78-79), joka vastaavasti perustuu Alan J. Leen (1997) artikkeliin[a], jonka esittämään kysymykseen (hypoteesiin) vastaus on tietenkin ilmeinen! Näin ollen seuraavassa tarkastellaankin kotijoukkueiden ja vierasjoukkueiden maalintekointensiteettiä Poisson-jakaumaan perustuen. Seuraavassa emme siis pyri mallintamaan tietyn spesifin ottelun lopputulosta vaan tarkastelemme "keskimääräisen" kotijoukkueen ja vierasjoukkueen "edustavaa" ottelua.
Seuraava taulukko raportoi tehtyjen maalimäärien jakaumat pelatuissa 380 ottelussa. Neljän tai yli neljän maalin tapaukset kirjataan 4+:nä maalina. Ts. esim. kys. kauden lopputulokset *Blackburn Rovers- Nottingham Forest* 7-0 ja *Bolton Wanderers - Manchester United* 0-6 tulevat aineistoon tuloksina 4+ vs. 0 ja 0 vs. 4+.

Kotij. maalien lkm.	Vierasj. maalien lkm.					Yht.
	0	1	2	3	4+	
0	27	29	10	8	2	76
1	59	53	14	12	4	142
2	28	32	14	12	4	90
3	19	14	7	4	1	45
4+	7	8	10	2	0	27
Yht.	140	136	55	38	11	380

Olettamalla, että koti- ja vierasjoukkueen todennäköisyys tehdä maali ottelun aikana on vakio, niin tällöin koti- ja vierasjoukkueen ottelun aikana tekemien maalien lukumäärää (ilman edellä käytettyä maalimäärien "katkaisua" neljään) voidaan melko hyvin approksimoida oletuksella, että nämä lukumäärät ovat Poisson-jakautuneita. Ts. $Y_i^H \sim P(\lambda_H)$ on sm., joka kuvaa i:n ottelun kotijoukkueen tekemien maalien lukumäärää ja intensiteettiparametrin λ_H arvon määrittäminen kuuluu tilastollisen päättelyn ja erityisesti estimointiteorian piiriin. Vastaavasti vierasjoukkueen maalimäärät: $Y_i^A \sim P(\lambda_A)$.

Osoittautuu, että parametreille λ_H ja λ_A saatavat estimaatit ovat $\lambda_H = 1,49$ ja $\lambda_A = 1,06$ ja ne vastaavat tässä yksinkertaistetussa tilanteessa koti- ja vierasjoukkueen keskimääräisiä maalimääriä:

	kotijoukkue (home)	vierasjoukkue (away)	Yht.
keskiarvo	1,486	1,063	2,550
varianssi	1,316	1,172	2,618

Tuloksista voidaan siis päätellä, että kotijoukkueen (odotettavissa oleva) maalimäärä on vierasjoukkuetta korkeampi (osoittaen osaltaan kotiedun merkitystä jalkapallossa). Lisäksi edellä todetun Poisson-jakauman teoreettisten ominaisuuksien mukaisesti keskimääräiset maalimäärät ovat lähellä niiden variansseja, mikä osoittaa osaltaan tarkasteltavassa yksinkertaisessa tilanteessa, että Poisson-jakaumaan perustuva jakaumaoletus on kelvollinen.
On syytä todeta lopuksi, että tämän vahvasti yksinkertaistetun tilanteen sijaan tilastotieteessä on laaja ja kasvava kirjallisuuden haara jalkapalloa ja muuta urheilua koskevien tilastollisen menetelmien saralla. Nämä vaativat kuitenkin syvällisemmän ymmärryksen kannalta jälleen huomattavasti laajempia tilastotieteen (aine- ja syventäviä) opintoja.

[a] Alan J. Lee (1997). Modeling Scores in the Premier League: Is Manchester United Really the Best? *Chance* 10(1), 15-19.

1.6 Sattuman rooli tieteenteossa: Vale-emävale-tilasto?

Erityisesti nykypäivänä ei-tieteellinen tieto ja tarkoituksellinen disinformaatio, joita perustellaan heppoisin havainnoin, leviävät internetissä kulovalkean tavoin. On tiedeyhteisön ja tutkijoiden moraalinen vastuu taistella näitä uskomuksia vastaan popularisoimalla tutkimustietoa, mikä saattaa ajoittain jopa pahentaa ongelmaa, sillä popularisoinnissa päteviltäkin tutkijoilta voi unohtua satunnaisuuden voima.[3]

- Ei-tieteellinen tieto on juurtunut syvälle ja tutkijat pyrkivät taistelemaan tätä vastaan **popularisoimalla tiedettä**.
- Kuten todettua, tilastollisessa tutkimuksessa mielenkiinnon kohteena on satunnaisilmiöiden tutkiminen ja erityisesti systemaattisen ja satunnaisen vaihtelun (signaalin ja kohinan) erottaminen sekä muuttujien välisten riippuvuuksien tutkiminen.
  - Kiinnostuksen kohteena on siis hyvin harvoin vain jokin yksittäinen tunnusluku, kuten keskiarvo, varianssi tai korrelaatio (palaamme näihin myöhemmin luvussa **??**).
  - Tieteen popularisointi on yksi tutkijoiden ja yliopistojen tiedeyhteisön tärkeimmistä yhteiskunnallisista tehtävistä, mutta valitettavan usein se typistyy yksittäisen viimeisimmän tutkimustuloksen esittelyksi.
- Yliopistoyhteisössä kuitenkin luonnollisesti luotamme kumuloituneeseen tutkittuun tietoon ja tiedämme, että **yksittäinen tutkimus on vasta hyvä alku**.
  - Ihmistieteitä, kuten ilmeisesti erityisesti psykologiaa sekä osin myös muiden ohella lääke- ja taloustiedettä, on viimeisen vuosikymmenen ajan puhuttanut paljon niin sanottu **replikaatiokriisi**, sillä useaa arvostettuakaan tutkimusta ei ole saatu **toistettua eli replikoitua**.
  - On ymmärrettävää, että replikaatiokriisi, varsinkin jos se on (alakohtaisesti) laajalle levinnyttä, murentaa kansalaisten luottamusta tieteellisiin tuloksiin.
  - Toistettavuus on yksi tutkimuksen peruskriteereistä, joka erottaa tieteellisen tiedon muista tietolähteistä, joten sen puuttuminen herättää ymmärrettävästi huolta tieteellisen prosessin toimivuudesta.
  - Replikaatiokriisin voi kuitenkin myös tulkita toisin: ilman kriittisyyttä omia (ja muiden) tuloksia kohtaan, ei mitään kriisiä olisikaan, joten silkka sen olemassaolo on osoitus tieteellisen prosessin toimivuudesta.

[3] Tämä jakso perustuu osin psykometriikan yliopisto-opettajan Jari Lipsasen blogiin vuodelta 2021.

- Kun tuntee ja tunnistaa sattuman voiman ja ymmärtää kaikki mahdolliset satunnaisuuden lähteet, jotka altistavat tutkimusprosessin virheille, tulee samalla ymmärtäneeksi että eri tavoin koeteltu, useassa tutkimuksessa kumuloitunut tieto tulisi olla kaiken tieteen popularisoinnin keskiössä yksittäisten, mahdollisesti uusien ja yllättävien tutkimustulosten sijaan.
  - Tähän mennessä olemme jo oppineet, että tälle on myös vahvat tilastolliset perustelut: satunnaisen tiedon maailmassa mikään ei ole täysin varmaa, ei edes kaikkein edistyneimpien tilastomenetelmien avulla!

2 Tilastolliset aineistot, niiden kerääminen ja mittaaminen

Edellisessä luvussa käsiteltiin tilastotieteen suhtautumista satunnaisilmiöihin. Tässä luvussa tarkastelemme lähemmin miten reaalimaailman satunnaisilmiöistä kerätään tietoa ja miten niitä voidaan mitata. Tilastotieteen perusoppimäärä rakentuu ajatukselle ilmiöiden tutkimisesta rajallisen ja epävarman tiedon vallitessa. Käytännössä tämä tarkoittaa sitä, että tutkimuksen kohteena olevat rajalliset aineistot sisältävät niin systemaattista kuin satunnaisuudesta johtuvaa vaihtelua. Tilastollisten menetelmien avulla pyrimme erottamaan systemaattisen vaihtelun satunnaisesta sekä tekemään tilastollista päättelyä aineiston generoimasta mekanismista. Lyhyesti tämä tarkoittaa aineiston systemaattisen vaihtelun tilastollista mallintamista ja sen parametrien estimointia otoksesta, joka kattaa vain (pienen) osajoukon koko populaation (perusjoukon) tilastoyksiköistä.

Voidaksemme tehdä uskottavaa päättelyä "havainnoista parametreihin", tulee otoksen olla riittävän **edustava**. Tämän luvun keskeisin oppi onkin, että miten otanta tulisi suorittaa, jotta havaintoaineisto olisi **edustava otos** populaatiosta, silloin kun aineisto kerätään otannalla. Vaikka aineiston hankinta vaatii yleensä runsaasti käytännön työtä, kannattaa se tehdä huolellisesti, sillä huonosti toteutetun otannan vuoksi tutkimusongelman kannalta keskeisiä johtopäätöksiä ei voida tehdä!

2.1 Kertausta: Data eli aineisto

- **Tilastollinen tutkimus** aloitetaan tutkimusaineiston keruun suunnittelulla.
- Kertauksen vuoksi: tilastollinen tutkimusaineisto (havaintoaineisto) kostuu tilastoyksiköiden populaatiosta havaituista tilastomuuttujien arvoista.
- Havaintoaineisto voidaan koota taulukoksi, johon listataan tilastoyksiköt riveille ja tilastomuuttujat sarakkeisiin. Jos havaintoaineisto koostuu n tilastoyksiköstä, joista jokaisesta on kerätty esim. m tilastomuuttujasta havainnot, niin havainnot voidaan kirjoittaa taulukon muotoon

	tilastomuuttuja 1	tilastomuuttuja 2	…	tilastomuuttuja m
tilastoyksikkö 1	$x_{1,1}$	$x_{1,2}$		$x_{1,m}$
tilastoyksikkö 2	$x_{2,2}$	$x_{2,2}$		$x_{2,m}$
…	…	…	…	
tilastoyksikkö n	$x_{n,1}$	$x_{1,2}$		$x_{n,m}$

Tässä siis rivillä i on i. **tilastoyksikön** havainto ja sarakkeessa j on j. tilastollisesta muuttujasta havaitut arvot $x_{i,j}$. Ts. yhdellä rivillä on yhden tilastoyksikön tiedot kaikista tilastomuuttujista ja yksi sarake on kaikkien tilastoyksiköiden tiedot yhdestä tilastomuuttujasta.

- Usein (varsinkin parhaillaan kiihtyvällä vauhdilla) kerättävät havaintoaineistot ovat niin suuria, ettei edellisenkaltaisesta havaintotaulukosta voida usein suoraan tarkastelemalla nähdä aineiston pääpiirteitä.

- Tällöin voi olla tarpeen luokitella aineistoa taulukon muodostamiseksi.

- Luokittelussa on kysymys aineiston tiivistämisestä kohtuullisen kokoiseksi ja havainnollisempaan muotoon. Luokittelussa tilastomuuttujan arvot sijoitetaan eri luokkiin siten, että yhden tilastomuuttujan arvo voi kuulua vain yhteen luokkaan. Luokka ilmoitetaan yleensä luokkavälinä, kuten reaalilukuvälinä. Esimerkiksi henkilön ikä on tapana luokitella ikäjakauman kuvaamisessa 10-vuotisluokkiin (15-24, 25-34, …), vaikka periaatteessa ikä voitaisiin ilmoittaa minuutinkin tarkkuudella.

- Luokkien lukumäärään vaikuttavat muun muassa tilastomuuttujan arvojen vaihteluväli ja havaintoaineiston laajuus. Luokittelussa pyritään siihen, että luokkien lukumäärä saadaan tarvittaessa luokkia yhdistämällä kohtuulliseksi ja että luokat valitaan tasavälisesti eli siten, että kahden peräkkäisen luokan alarajojen erotus on vakio. Kun aineistoa luokitellaan, aineiston luettavuus paranee mutta toisaalta osa tiedoista menetetään eivätkä yksittäiset havaintoarvot ole enää tiedossa.

- Emme vielä tällä kurssilla etene tämän pidemmälle tilastografiikan esittämisessä ja siihen liittyvissä pohdinnoissa. Muun muassa tilastollisen päättelyn peruskurssi (TILM3555) vastaa näihin kysymyksiin tarkemmin. Graafiset menetelmät ovat joka tapauksessa erittäin tärkeä osa aineiston havainnollistamista. Kuvat helpottavat aineiston tulkitsemista ja toimivat usein perusteltuna lähtökohtana monimutkaisempien tilastollisten mallien (ja algoritmien) sovittamiselle.

- Kvantitatiivisen tutkimuksen aineistoksi kelpaa periaatteessa kaikki havaintoihin perustuva informaatio, joka on **mittauksen** avulla muutettavissa numeeriseen muotoon.
  - Havaintoyksiköiden tilastollisten muuttujien numeerisia arvoja kutsutaan **havaintoarvoiksi** tai **havainnoiksi**.
  - Kaikki havaitut tilastolliset muuttujat eivät ole aina mielenkiintoisia. Tutkimuksen kannalta mielenkiintoisia muuttujia kutsutaan **tutkimusmuuttujiksi**, joiden lisäksi havaintoaineisto pitää mahdollisesti sisällään **taustamuuttujia**.
    - Esimerkiksi, jos tutkimuksella halutaan tietoa suomalaisen aikuisväestön mielipiteistä, havaintoyksikköinä ovat aikuisväestöön kuuluvat henkilöt. Jos halutaan tietoa suomalaisista kunnista, havaintoyksikköinä ovat Suomen kunnat jne.
    - Ensimmäisessä tapauksessa tilastollisina muuttujina on aikuisväestön mielipiteet, joita voidaan selvittää esimerkiksi kyselytutkimuksella. Toisaalta voidaan myös kerätä taustamuuttujiksi haastatelluista muita tietoja, kuten asuinpaikka, ikä ja ammatti.
  - Kaikkia mielenkiintoisia muuttujia ei kuitenkaan välttämättä voida havaita, eli niille ei voida määrittää numeerista arvoa.
  - Tällöin puhutaan nk. **latenteista muuttujista**, eli muuttujista joita ei suoraan havaita mutta joiden oletetaan vaikuttavan havaittavien muuttujien taustalla. Latentteja muuttujia voidaan rakentaa tilastollisten mallien avulla käyttäen hyödyksi niihin liittyviä havaittuja muuttujia.
  - Latentteja muuttujia ovat esimerkiksi elämänlaatu, onnellisuus, konservatiivisuus, yms.

- Tilastollinen tutkimus voi olla joko **kokonaistutkimus** tai **otantatutkimus**.

Kokonaistutkimus
Kokonaistutkimus on tutkimus, jossa tutkitaan kaikki tutkimuksen kohteena olevan perusjoukon alkiot, ts. kaikki ajateltavissa olevat kohteet tutkitaan.

- Kokonaistutkimus on yleinen tutkimustapa silloin, kun kohdeperusjoukko on selvästi määritelty ja sen alkioita koskevat tilastolliset muuttujat ovat helposti mitattavissa.
- Esimerkiksi jos tutkitaan Suomen kuntia, niin kokonaistutkimuksessa tutkitaan kaikki kunnat, joista on helppoa kerätä mielenkiinnon kohteena olevia tilastollisia muuttujia useimmissa tilanteissa.
- Toisaalta jos tutkitaan jonkin lääkeaineen vaikutuksia ihmisiin, niin kokonaistutkimuksessa tutkittaisiin jokainen ihminen erikseen. Selvää on, että tällainen kokonaistutkimus olisi liian vaikeaa toteuttaa.

Otantatutkimus
Otantatutkimuksessa tutkimus kohdistetaan johonkin (populaation/perusjoukon) osajoukkoon, joka pooimitaan sopivaa **otantamenetelmää** käyttäen (ks. alaluku 2.5) ja populaatiota/perusjoukkoa koskevat johtopäätelmät tehdään tähän otokseen perustuen.

- Otantatutkimus on usein luonnollinen valinta, sillä koko populaation tutkiminen ei useinkaan ole mahdollista tai kannattavaa.
  - Esimerkiksi aseiden patruunoita valmistava tehtailija ei voi tutkia toimivatko kaikki ammukset. Myöskään valaisimien valmistaja tuskin tekee kokonaistutkimuksia valmistamiensa tuotteiden kestoajan selvittämiseksi.
- Perusjoukosta otokseen poimittuja alkioita kutsutaan **otosyksiköiksi** ja niiden muodostama osajoukko, eli **otos**, on se osa perusjoukkoa, joka tutkitaan tutkimusaineiston keräämisen jälkeen.
  - Lääketutkimusta tehdäänkin poikkeuksetta otantatutkimuksena (ja kontrolloituina kokeina, ks. alempaa), jolloin lääkettä testataan vain osajoukolla koko ihmispopulaatiosta ja tämän osajoukon alkiot ovat otosyksiköitä.
  - Näin toimimalla, ja riittävän edustavalla otoksella, saadaan kuitenkin tarpeeksi tietoa lääkeaineen vaikutuksista ja tulokset voidaan yleistää populaatiotasolle ja lääke ottaa käyttöön.
- Otantatutkimus on halvempi kuin kokonaistutkimus ja tulokset saadaan nopeammin!

- Otantatutkimuksessa keskitytään siis perusjoukkoa edustavan pienemmän, mieluusti satunnaisesti valitun otoksen tutkimiseen.
  - Otantatutkimuksissa tiedot kerätään useimmiten haastattelemella, kirjallisella/sähköisellä kyselyllä tai suoraan tietorekistereistä. Tiedonkeruun toteuttaminen (eri sovelluksissa) määrää osaltaan käytettävän otantamenetelmän.
  - Teoriassa äärelliseen perusjoukkoon kohdistuvat kokonaistutkimukset voidaan aina tulkita otantatutkimuksiksi (perusjoukko tulkitaan otokseksi hypoteettisesta äärettömästä perusjoukosta)!
    - Esimerkiksi Galilein tekemät painovoiman vaikutusta kappaleiden putoamisaikaan liittyneet mittaukset. Koetuloksia (mittauksia) voidaan pitää otoksena äärettömästä mahdollisten koetulosten joukosta. Tällöin ainoa mahdollisuus ilmiön tutkimiseen on käyttää otantaa.
- Otantatutkimuksen tulokset voivat olla luotettavampia kuin kokonaistutkimuksen.

  - Otantatutkimuksessa voidaan panostaa enemmän huolelliseen ja tarkkaan mittaamiseen sekä valitun otoksen tavoittamiseen.
  - Kokonaistutkimuksessa vastauskato ja tarkasteltavan populaation valintavirhe ovat mahdollisia siinä kuin otantatutkimuksessakin.

- Otantateoria on yksi tilastotieteen keskeisimpiä oppeja ja tarjoaa teoreettisen kehikon empiiristen tutkimusten tulosten yleistämiseen. Tarkastellaan siis tarkemmin otannan ideaa ja toteuttamista seuraavassa alaluvussa.

2.2 Otannan idea

- Otantatutkimuksen (karkeat) suunnittelu- ja työvaiheet ovat seuraavat:

  1. Tavoitteiden asettaminen
  2. Perusjoukon (populaation) asettaminen
  3. Kehikko
  4. Kerättävän informaation sisältö (mitä tietoa todella tarvitaan, mitä voidaan jättää pois, suunnitellaan kysymykset ja mahdollinen kyselylomake)
  5. Otoskoon määrittäminen
  6. Suoritetaan otoksen poiminta, tietojen keräys ja tarkastus
  7. Aineiston taulukointi ja analysointi
  8. Raportin laatiminen

- Otantatutkimuksessa ajatuksena on siis poimia **edustava otos** siitä populaatiosta (perusjoukosta), joka on mielenkiinnon kohteena eli jota halutaan tutkia ja josta halutaan tietoja.
  - **Tavoiteperusjoukko** on joukko, johon otannan myötä saatavat tutkimustulokset halutaan yleistää. Toisin sanoen, se mistä haluamme tietoja määrää populaation.
  - **Kohdeperusjoukko** on joukko, jota koskevia tietoja halutaan kerätä.
    * Esimerkiksi äänestysikäiset Suomen kansalaiset.
    * Usein tavoiteperusjoukko = kohdeperusjoukko.
    * Tavoiteperusjoukko voi joskus olla laajempi (esim. “ihmiset” vs. “suomalaiset”).

- Tutkimuksessa (edustavaan) otokseen poimitut tilastoyksiköt, näiden tilastolliset muuttujat ja niiden arvot muodostavat **otosaineiston** eli siis tutkimus- tai havaintoaineiston (datan).
  - Tutkimuskysymykseen vastatakseen tutkija valitsee sopivan tilastollisen mallin ja estimoi sen parametrit tähän otokseen perustuen.
  - Perusoletuksena on otoksen ja valitun tilastollisten mallin pohjalta suoritettavan tilastollisen päättelyn **yleistettävyys koko populaatioon**.
  - Otos valitaan **otantaa** ja erilaisia **otantamenetelmiä** hyödyntäen pyrkien varmistamaan otoksen **edustavuus** (perusjoukko pienoiskoossa, ks kuva **??**(fig:otanta)).

Edustavuus
Tutkimukseen valitut yksiköt edustavat koko populaatiota, ts. tutkimukseen valittu osajoukko kuvaa perusjoukon ominaisuuksia kattavasti.

- Keskeistä tutkimuksen ja sen edustavuuden kannalta on, että tutkija osaa kerätä sisällöllisesti ja määrällisesti **sopivan kokoisen** aineiston.
- Tietyn otoksen edustavuutta arvioidessa voi käyttää apuna seuraavia kysymyksiä:
  - Miksi päädyttiin tämän kokoiseen otokseen?

    - ∗ **Otoskoko** vaikuttaa siihen miten hyvin otoksesta tehdyt johtopäätökset voidaan yleistää koskemaan koko perusjoukkoa, ts. kuinka luotettavia ne ovat. Tämä johtuu siitä että yksittäisten otosyksiköiden ominaisuudet saattavat vaihdella suuresti ja kasvattamalla otoskokoa perusjoukon systemaattiset piirteet tulevat otoskoon kasvaessa yhä paremmin esille. Kun otoskoko vastaa populaation kokoa, on kyseessä tietenkin kokonaistutkimus, joka kertoo kaiken perusjoukosta. Otoskoon valintaan ja määräämiseen palataan myöhemmin luvussa **??**.
  - Käytettiinkö apuna tilastotieteellisesti vankkaa suunnittelua otoskoon määrittämiseksi ja/tai miten pyrittiin varmistamaan tutkimuksen kannalta tärkeisiin analyysiryhmiin kuuluvien riittävä määrä aineistossa?
  - Harkittiinko muita otantamenetelmiä ja miksi päädyttiin juuri käytössä olleeseen menetelmään?
- Edustavuuteen vaikuttaa keskeisesti se, millä tavoin otanta pystytään suorittamaan, ts. mihin kohdeperusjoukkoon otanta kohdistetaan.
  - **Kehikkoperusjoukko** on rekisterin, luettelon tms. peittämä osa kohdeperusjoukkoa. Kyseessä on siis se osa kohdeperusjoukkoa, josta otanta ylipäänsä pystytään suorittamaan.
  - **Otantakehikon alipeitto** esiintyy, kun otantakehikosta puuttuu osa kohdeperusjoukon alkioista (esim. tutkimus suoritetaan puhelinhaastattelulla, mutta osa aiottuun otokseen kuuluvista haastateltavista ei omista puhelinta).

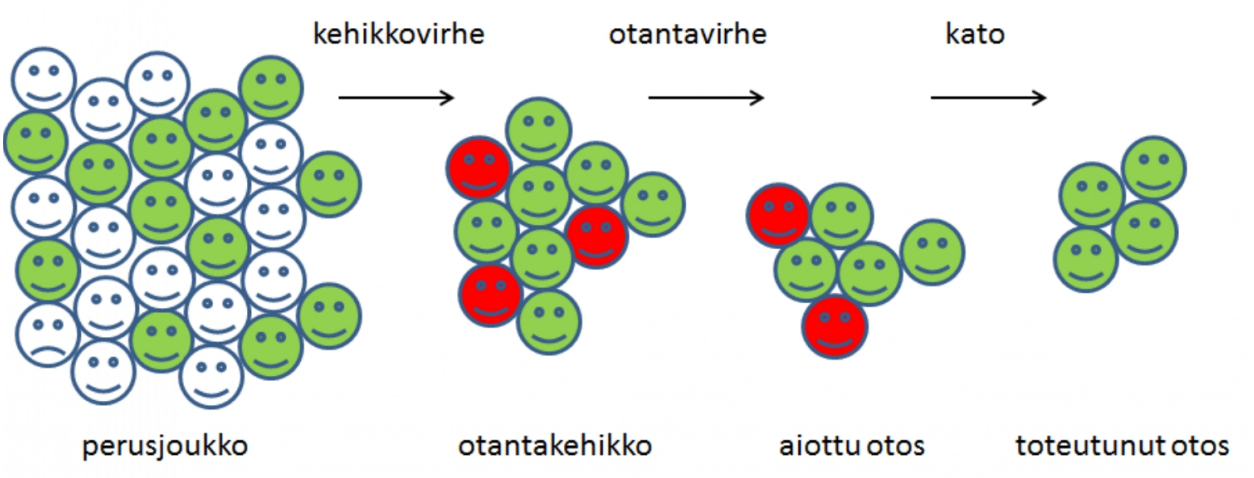

Figure 4: Otannan idea.

- Edustavan otoksen avulla on mahdollista tehdä perusjoukkoa koskevaa tilastollista päättelyä, sillä otos kuvaa perusjoukon ominaisuuksia riittävän hyvin. Tämä on yksi tilastotieteen keskeisimpiä oppeja mutta myös kriittisen tiedelukutaidon ja arkijärjen kannalta tärkeää.

Esimerkki: Kotitalouksien tulot, tuloerot ja pienituloisuusrajan kehitys 1987-2005 (Tilastokeskus)

- Tilastotyksikkö kotitalous, joten kaikkien kotitalouksien tutkiminen (kokonaistutkimus, ks. alla) olisi vaikeaa ja aikaavievää.
- Tutkittavaksi valitaan vain muutama tuhat kotitaloutta (ts. otantatutkimus) ja selvitetään näiden tulot.
- On mahdollista tehdä **kaikkia** suomalaisia kotitalouksia koskevia johtopäätöksiä, jos tutkitut yksiköt olivat **edustava otos** suomalaisista kotitalouksista. Ts. osajoukkoa koskevat päätelmät voidaan yleistää koskemaan perusjoukkoa, mikäli osajoukko on edustava otos perusjoukosta.

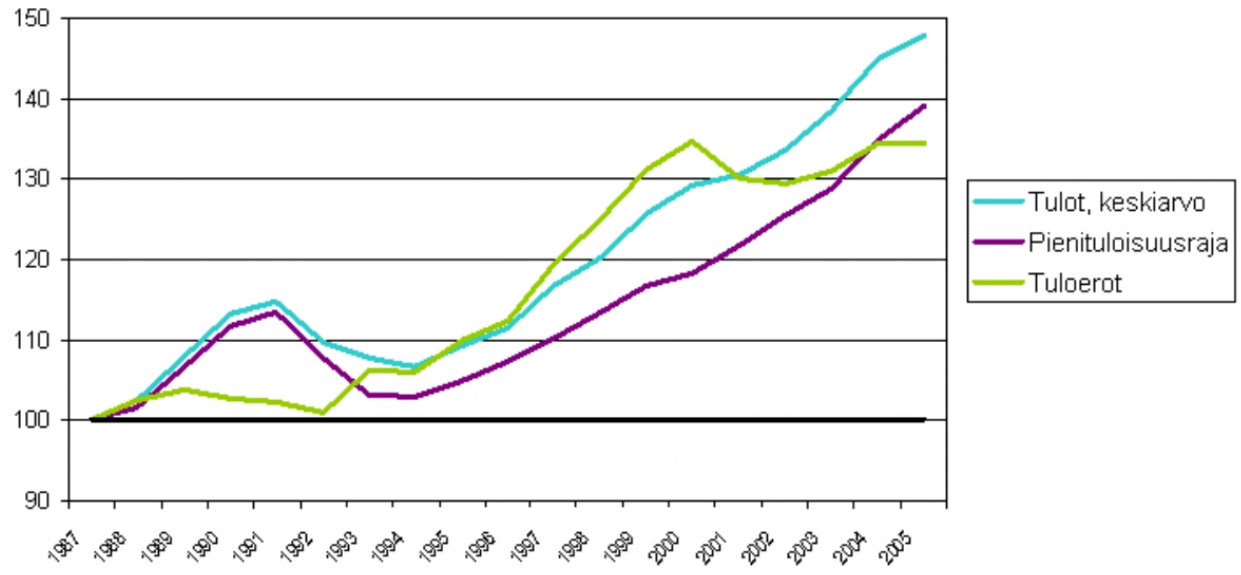

Figure 5: Tuloerot.

2.3 Tilastollisten muuttujien mittaaminen ja mitta-asteikot

Mittaaminen

- Tilastotieteellinen tutkimus perustuu aina mitattaviin satunnaisilmiöihin: tavoitteena on mittaamalla liittää jokin luku ilmiötä kuvaavaan ominaisuuteen, ts. mitata kyseisen satunnaismuuttujan havaittua arvoa.
- Kumpaa tahansa tutkimusotetta (kokonais- tai otantatutkimus) noudatettaessa tietojen keräämisessä on olennaisena osana kohteiden ominaisuuksien **mittaaminen**.
  - Mittaaminen vaatii aina mittauksen kohteen, hyvin määritellyn mitattavan ominaisuuden ja **mittarin**, joka liittää mielekkäät lukuarvot mitattavaan ominaisuuteen.
  - Erilaiset mittarit heijastavat ilmiön ominaisuuksia eri tavoin ja eri tarkkuudella
    - Esimerkiksi jos tutkitaan opiskelijoiden pituuden kehitystä niin mitataan pituutta eri aikoina. Pituudet voidaan mitata senttimetreissä, metreissä, kilometreissä tai vaikkapa tuumissa.
    - Mittari on hyvä jos sen antama mittaus on
      - **(i) validi** eli mittaus esittää oikein mitattavaa ominaisuutta (senttimetri mittaa pituutta, gramma ei) ja
      - **(ii) luotettava** eli mittaus on **harhaton** ja **toistettavissa**.
    - Määritellään nämä termit vielä erikseen, sillä ne ovat keskeisiä tilastotieteessä.

Harhattomuus
Mittari on harhaton, jos se ei systemaattisesti ali- tai yliarvioi mitattavan ominaisuuden määrää.

- Harhaton mittari siis antaa keskimäärin oikeita mittauksia mitattavasta ominaisuudesta.
- Harhattomuutta pidetään myös hyvänä ominaisuutena tilastollisten mallien parametrien estimaattoreille. Tähän palataan myöhemmin luvussa **??**.

Toistettaavuus
Mittari on toistettava, jos se tuottaa keskimäärin samanlaisia mittauksia samanlaisista otoksista eli se on johdonmukainen ja mittausvirheet ovat pieniä.

- Huonosti toistettava mittari antaa tilastoyksiköiden samankaltaisille ominaisuuksille hyvin erilaisia arvoja riippuen otoksesta.
- **Mittausten reliabiliteettia/luotettavuutta** arvioidessa voidaan pohtia esimerkiksi seuraavia kysymyksiä:
  - Kuinka hyvin mittaustulokset ovat toistettavissa, kuinka paljon niissä on ei-sattumanvaraisuutta?
  - Mittausten validiteetti: kuinka hyvin pystyttiin mittaamaan sitä, mitä oli tarkoitus mitata?
- Kun mittaaminen on luotettavaa ja validia, tutkimusaineisto on **sisäisesti luotettavaa**.
- Aineiston **ulkoinen luotettavuus** toteutuu silloin, kun tutkittu otos edustaa perusjoukkoa eli on edustava. Validi mittaaminen ei pelasta epäedustavaa otosta!
- Jokaisen tutkimuksen tulosten luotettavuuden perusteena on käytetty aineisto, kuinka se on hankittu ja mistä lähteestä. Kun käytetään luotettavaksi havaittuja mittareita, voidaan kustakin aineistosta laskea erikseen tunnuslukuja mittauksen luotettavuudelle. Esimerkkinä **luottamusväli**:
  - Luottamusväliä käytetään määrittämään estimaatin luotettavuutta.
  - Väli, joka vaihtelee otoksesta toiseen ja joka usein sisältää mielenkiinnon kohteena olevan parametrin, kun otantakoetta toistetaan!

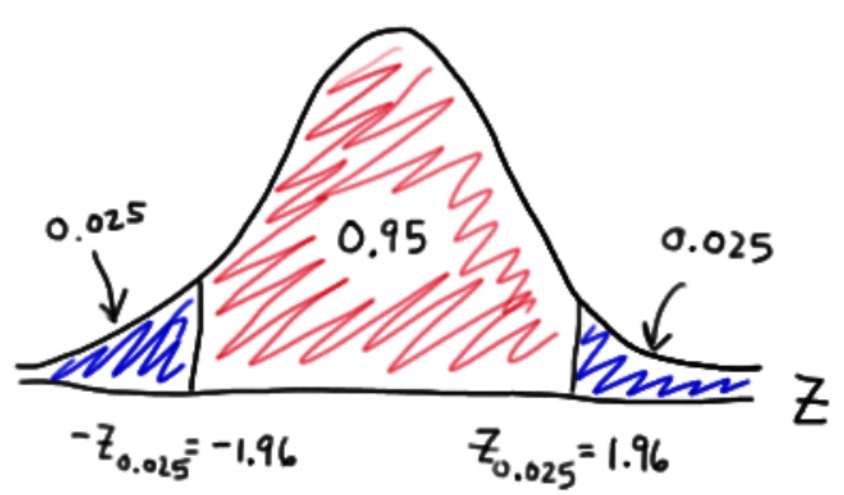

Figure 6: Normaalijakauman luottamusväli. Väliestimointia tarkastellaan tarkemmin seuraavassa luvussa.

- Luotettavuudella voidaan tarkoittaa myös tutkimuksen **objektiivisuutta / puolueettomuutta**
  - **Objektiivinen totuus**, tutkimustulokset ovat samat riippumatta siitä kuka pätevä tutkija tutkimuksen on tehnyt.
  - Tulosten tulisi olla luotettavia, mutta luotettavatkin havainnot voivat olla puolueellisia siinä mielessä, että ne tarkastelevat asiaa vain yhdeltä näkökannalta!
  - Esim. tarkastellaan yrityksen henkilöstökysymyksiä, työn organisointia ja työmoraalia, ongelmien tarkastelu johdon vs. henkilöstön näkökulmasta.

Esimerkki: C-vitamiinin vaikutus syövän hoidossa

- Annettiin C-vitamiinia 100 terminaalivaiheen syöpäpotilaalle ja seurattiin kuolleisuutta (Cameron and Pauling, 1976).
  - Pyrittiin luomaan tärkeiden ominaisuuksien suhteen samanlaisia verrokkiryhmiä ja valittiin kutakin potilasta kohden 10 verrokkia, jotka olivat samanlaisia iän, sukupuolen, primääri-kasvaimen sijaintipaikan ja histologisen kasvaintyypin suhteen.
  - Seuranta-aika: aika hetkestä, jolloin todettiin tavanomaisten hoitojen olevan tehottomia, kuolinhetkeen saakka.
  - Tulos: C-vitamiinia saaneet käsittelyryhmän potilaat elivät 4 kertaa kauemmin ($p < 0.0001$).
- Ristiriitaista evidenssiä saatiin tutkimuksessa, jossa vastaava tutkimusongelma, mutta toteutettu satunnaistettuna kokeena (Moertel et al. 1985).
  - Satunnaistettiin potilaat, joilla pitkälle edennyt paksunsuolen tai peräsuolen syöpä, C-vitamiinia saavien ja lumelääkettä saavien ryhmiin.
  - Tulos: kontrolliryhmän potilaat elivät keskimäärin hieman pidempään, mutta ero ei merkitsevä.
- Mistä kahden tutkimuksen erot johtuivat? Huonolla tuurilla kaltaistetut verrokit erosivat käsittelyryhmän potilaista joillakin merkittävillä tavoilla, joita ei oltu mitattu! Miten kvantifioida "huonoa tuuria"?
- Tilastolliset menetelmät tekevät juuri tämän: "Mikä on todennäköisyys, että havaittu tulos (tai sitä enemmän nollahypoteesista poikkeava tulos) olisi syntynyt vain sattumalta?"
  - Ilman satunnaistamista tuota kenties merkittävää ei-mitattua eroa ei pystytä varmuudella kontrolloimaan.
  - Todellisuudessa ero johtui siitä, että ensin mainitun tutkimuksen kontrollit valittiin jo kuolleista syöpäpotilaista, eikä heihin liittynyt enää mitään satunnaisuutta!

_Mitta-asteikot___

- Kuten satunnaismuuttujia koskeneessa luvussa 1 opittiin, satunnaisilmiöillä on erilaisia tulosvaihtoehtoja jotka kantavat satunnaismuuttujien todennäköisyysjakaumia.
  - On syytä huomauttaa, että vaikka mitattava ilmiö ei olisikaan numeerinen, se voidaan aina "koodata" eli muuntaa numeeriseksi. Esimerkiksi perinteinen kaksiarvoinen mies-nainen -muuttujan tapauksessa voidaan käyttää tunnuksia 0 ja 1.
- Ilmiön luonteesta riippuen voidaan näille tulosvaihtoehdoille käyttää erilaisia **mitta-asteikkoja**.
  - **Laatueroasteikko/luokitteluasteikko** (nominaaliasteikko): Muuttujan mittaustaso on tällöin sellainen, että sen arvot voidaan luokittaa toisistaan eroaviin luokkiin. Ts. mihin luokkaan kohde kuuluu mitattavan ominaisuuden perusteella?
    - Tilastoyksiköt luokitellaan ennaltamääriteltyihin luokkiin. Luokkien järjestyksellä ei ole merkitystä.
    - Kukin tilastoyksikkö kuuluu vain yhteen luokkaan. Tällöin kahdesta tilastoyksiköstä/havainnosta voidaan päätellä vain kuuluvatko ne saamaan luokkaan vai eivät.
    - Emme pysty määrittelemään empiirisesti mielekästä järjestystä havaintoarvojen välillä.

    * Esimerkkejä: Sukupuoli, veriryhmä tai kotikunta.
  - **Järjestysasteikko** (ordinaaliasteikko): Tällöin muuttujan arvot voidaan luokittelun lisäksi asettaa empiirisesti mielekkääseen järjestykseen. Tällöin siis mittauksen kohteella on "enemmän mitattavaa ominaisuutta" kuin jollakin toisella kohteella
    * Tilastoyksiköt luokitellaan ennalta määrättyihin luokkiin, joilla on yksikäsitteinen järjestys.
    * Esimerkkejä: Sotilasarvo, sosiaaliryhmä, kilpailun tulos tai sairauksien tarttuvuus.
  - **Välimatka-asteikko** (intervalliasteikko): Luokittamisen ja järjestyksen asettamisen lisäksi havaintoarvojen välimatkalla on empiirisesti mielekäs tulkinta. Ts. intervalliasteikon tasoisen muuttujan arvoista voidaan sanoa, kuinka paljon toinen arvo on toista suurempi (pienempi).
    * Välimatka-asteikolla pystytään mittaamaan yksittäisten luokkien tai havaintoarvojen ero. Esimerkiksi: Lämpötilan mittaaminen esim. celcius-asteina. Pystymme numeroarvoina ilmoittamaan onko tänään lämpimämpi, yhtä lämmin vai kylmempi sää kuin eilen ja kuinka monta astetta muutos on.
    * Kuinka paljon kahden mittauksen kohteen ominaisuudet eroavat toisistaan.
    * Intervalliasteikon tasoisen muuttujan arvoista voidaan sanoa, kuinka paljon toinen arvo on toista suurempi (pienempi). Mittarin nollapiste on kuitenkin "keinotekoinen" ja siten vapaasti valittavissa. Samoin voidaan valita käytettävä mittayksikkö vapaasti. Oleellista on vain se, että havaintojen välisellä välimatkalla on aina empiirisesti mielekäs tulkinta.
    * Yhteen- ja vähennyslasku ovat sallittuja.
  - **Suhdeasteikko**: Jos intervalliasteikon ominaisuuksien lisäksi on määriteltynä yksikäsitteinen mittalukujen absoluuttinen nollapiste.
    * Esimerkiksi kuuden euron hintainen tuote on kaksi kertaa niin kallis kuin kolmen euron tuote.
    * Kunnan veroäyri tai henkilön pituus: Absoluuttinen nollapiste on 0.
    * Nollapisteen ollessa absoluuttinen, se "pysyy paikallaan" ja mittalukujen suhteet pysyvät samoina.

- Mitta-asteikot voidaan jakaa kahteen luokkaan: **Luokittelu- ja järjestysasteikkoa kutsutaan kvalitatiivisiksi asteikoiksi**. Tällöin muuttujien arvot kuvaavat vain tilastoyksiköiden laadullisia piirteitä.

- Vastavasti **välimatka- ja suhdeasteikkoa kutsutaan kvantitatiivisiksi asteikoiksi**, koska tällöin mittaluvut kuvaavat jonkin ominaisuuden määrää.

- Tilastollisen analyysin kannalta mitta-asteikkojen merkitys on siinä, että tilastollisten (matemaattisten) operaatioiden sallittavuus määräytyy muuttujan mitta-asteikon mukaan. Mitä korkeampi mitta-asteikko, sitä enemmän on käytettävissä olevia analyysimenetelmiä. Esimerkiksi keskiarvon laskeminen on eräs tilastollinen operaatio, ja se ei ole sallittu kvalitatiivislle muuttujille.

- **Aineistotyyppejä**: Käsitellään tarkemmin vielä myöhemmin (Luvussa **??**), joiden yhteydessä mitattavat muuttujat voivat olla kvalitatiivisia tai kvantitatiivisia.

  - Poikkileikkausaineisto: Tietoja yhdeltä ajanhetkeltä tai aikaväliltä
  - Aikasarja-aineisto: Tietoja samasta tutkimuskohteesta eri ajanhetkiltä
  - Paneeliaineisto: Tietoja useilta ajanhetkiltä useista tutkimuskohteista
  - Tapahtumahistoria-aineisto: Tietoja tapahtumahetkiltä

2.4 Kontrolloidut kokeet ja suorat havainnot

- Tilastollinen tutkimusaineisto voidaan kerätä:

  - **Kontrolloiduilla kokeilla**, joissa tutkimuksen kohteet altistetaan suunnitelmallisesti erilaisiin koeolosuhteisiin selvittääkseen miten kohteet reagoivat muutoksiin.
  - **Suoria havaintoja** tehtäessä koeolosuhteita ei pyritä aktiivisesti muuttamaan vaan ainoastaan seurataan miten erilaiset olosuhteet ja niissä tapahtuvat muutokset vaikuttavat kohteisiin.

- Näistä tutkimusasetelmista kontrolloidut kokeet ovat tietenkin ihanteellisempia tutkimuksen tekemiselle, sillä tutkijan on mahdollista tarkastella tutkittavaa asiaa koeolosuhteissa "eristyksissä".
- Kontrolloidut kokeet eivät kuitenkaan ole aina mahdollisia, jolloin on käytettävä suoria havaintoja.
  - Tällöin tutkimuskohdetta ei suunnitelmallisesti altisteta koeolosuhteille ("käsittelyille") vaan muuttuvien olosuhteiden vaikutuksia tilastoyksikköihin seurataan passiivisesti.
  - Toisin sanoen tutkimuksen kohteena olevat tilastoyksiksöt eivät välttämättä edes tiedä osallistuvansa tutkimukseen.
- Lisäksi usein tehdään hoito/käsittelyvastetta koskevia vertailuja erilaisissa olosuhteissa, joka osaltaan vaikuttaa tulosten uskottavuuteen, sillä tutkittaviin tilastoyksiköihin voi vaikuttaa olosuhteiden muutosten lisäksi muut ulkopuoliset tekijät.
  - Näiden **selittävien** ja **sekoittavien tekijöiden** vaikutusten kontrollointi on suoria havaintoja tehtäessä vaativa tehtävä.
  - Mikäli ulkopuolisia tekijöitä ei havaita ja/tai pystytä mittaamaan, tai muuten jostain syystä olla lisätty ja käytetty käytettävässä tilastollisessa mallissa, voi kyseeseen tulla ns. **puuttuvien selittäjien harha**, joka tarkoittaa sitä että havaittuihin tuloksiin vaikuttaa jokin havaitsematon tekijä, mutta jonka vaikutusta ei kyetä kvantifioimaan puutteellisten havaintoarvojen vuoksi.
- Suoria havaintoja tehtäessä ei voida (usein) selvittää vasteen ja olosuhteiden **kausaalista** yhteyttä. Suorilla havainnoilla voidaan lähinnä saada selville onko vasteella ja olosuhteilla jokin yhteys (korrelaatio) (ks. luku **??**).
- Suorien havaintojen keräämiseen liittyy olennaisesti joitain riskejä ja toisaalta rajoituksia. Riskit liittyvät käytännössä otoksen harhaisuuteen (erit. valikoitumisharha)
  - Esimerkiksi jos havaintoja tehtäessä suositaan systemaattisesti joitakin tulosvaihtoehtoja. Tämä suosiminen voi olla tahallista tai tahatonta.
  - Tämä tilastoyksiköiden **valikoituminen** otokseen aiheuttaa harhaa, sillä otokseen valikoituva osajoukko saattaa ylikorostaa perusjoukon jotain ominaisuuksia.

Valikoituminen
Valikoitumista tapahtuu, jos otokseen poiminta ei ole riippumatonta tilastoyksikön ominaisuuksista. Tätä kutsutaan valikoitumisharhaksi.

- Esimerkiksi verrattaessa sydän- ja verisuonitautipotilaiden hoitotoimenpiteitä potilaat eivät mahdollisesti ole valikoituneet yhtä todennäköisesti pallolaajennukseen, ohitusleikkaukseen tai lääkehoitoryhmään, sillä taudin vakavuus saattaa jo määritellä mikä hoitotoimenpide valitaan.
- Valikoituminen on iso ongelma seurantatutkimuksissa, sillä harhaisten havaintotulosten, eli harhaisen otoksen, perusteella ei voida tehdä luotettavia johtopäätöksiä perusjoukosta!

- Harhan syntymistä pyritään välttämään valitsemalla havaintojen kohteet perusjoukosta satunnaisesti (ellei tavoitteena ole tutkia kaikkia perusjoukon alkioita). Tämä merkitsee satunnaisotannan soveltamista havaintojen kohteiden valintaan, eli otokseen poimittavien tilastoyksiköiden valintaan sovelletaan **satunnaistamista**, jolloin sattuma määrää mitkä perusjoukon alkioista tulevat poimituksi otokseen (tutkimuksen kohteiksi)!

Satunnaistaminen
Tilastoyksiköiden poimimista populaatiosta otokseen riippumatta muiden yksiköiden poiminnasta tai

kyseisten (poimittavien) yksiköiden ominaisuuksista.

- Satunnaistaminen takaa sen, että mahdolliset sekoittavat tekijät ovat jakaantuneet tasaisesti tutkittavassa joukossa. Tällöin sekoittavat tekijät eivät aiheuta harhaa otokseen ja tutkimuksen tulokset voidaan yleistää koko populaatioon.
- Satunnaistaminen poistaa otannasta valikoitumisharhan, sillä otokseen poiminta suoritetaan riippumatta tilastoyksiköiden ominaisuuksista. Satunnaistaminen on ainoa puolueeton tapa poimia otos (ei suosi mitään perusjoukon osaa)!

- Satunnaistaminen (osaltaan) mahdollistaa **tilastollisen päättelyn**, jonka avulla otoksesta saatuja tietoja voidaan hyödyntää tehtäessä päätelmiä koko perusjoukosta.
  - Tilastollisen päättelyn avulla voidaan muodostaa esimerkiksi jakaumien ja tilastollisten mallien tuntemattomille parametreille arviot (piste-estimaatit) ja arvioida niiden epävarmuutta (keskivirheet ja luottamusvälit) sekä testata tarkasteltavaan ilmiöön liittyviä hypoteeseja (ks. luku **??**).
- Johtopäätelmien pätevyys riippuu mm. siitä, kuinka hyvin otanta on suoritettu. Tämän vuoksi on tärkeää ymmärtää otannan perusperiaatteet ja erilaisten otantamenetelmien luonne.
- Kontrolloiduissa kokeissa satunnaistaminen jakaa yksilöt **riippumatta yksilön omista ilmiöön vaikuttavista muuttujista** joko **käsittely- tai kontrolliryhmään** (eng. treatment ja control).
  - Se takaa, ettei valikoitumista jonkin käsittelyä edeltävän ominaisuuden mukaan esiinny
  - Tämä tarkoittaa **altisteen** (käsittely / "treatment' ') antamista (täysin) satunnaisesti kokeeseen valituille yksilöille, riippumatta näiden taustamuuttujien arvoista.
  - Nämä yksilöt sinänsä voivat olla satunnaisotos jostain populaatiosta (tai ainakin niiden toivotaan olevan), mutta satunnaistaminen tarkoittaa siis käsittelyn kohdentamista koeyksilöille, ei satunnaisotantaa sinänsä
  - Esimerkiksi tutkittavat voidaan satunnaistaa lääkehoito- ja placeboryhmiin, jotta mahdolliset erot tutkittavien iässä, sukupuolessa ja muissa taustamuuttujissa eivät aiheuta systemaattista harhaa, kun tutkitaan lääkehoidon vaikutusta.

2.5 Otantamenetelmät

- Tässä jaksossa tarkastellaan erilaisia **otantamenetelmiä**. Näiden menetelmien tarkoitus on suorittaa otosaineiston (tutkimusaineiston) kerääminen niin, että se huomioi aiemmin esitellyt hyvän otannan kriteerit, ts. että sen tuottama otos on edustava ja luotettava. Näin ollen otos kuvaa koko perusjoukkoa.
  - Otantamenetelmän, joskus myös **otanta-asetelman**, valinta on tietenkin vahvasti sovellusalakohtainen: käytettävät aineistot ja täten otantamenetelmät määräytyvät pitkälti tehtävän tutkimuksen luonteen perusteella. Ts. käytännön tilanteet poikkeavat toisistaan lopulta varsin paljon ja eri tilanteisiin tarvitaan omat menetelmänsä.
  - Otanta-asetelmalla tarkoitetaan erityisesti otoksen poimintaan käytettyä **satunnaistuksen menetelmää**.
- Otannan tavoitteena on tietenkin edustava otos. Otoksen edustavuuteen vaikuttaa käytännön otannassa se, miten todennäköistä kullakin perusjoukon alkiolla (populaation tilastoyksiköllä) on tulla poimituksi otokseen. Tätä kutsutaan **sisältymistodennäköisyydeksi**.

Sisältymistodennäköisyys
Sisältymistodennäköisyys kuvaa sitä (tunnettua) todennäköisyyttä, jolla perusjoukon alkio tulee poimituksi otokseen.

- Käytännössä otoksen poiminta suoritetaan niin, että n:n alkion otos (n on otoskoko) poimitaan jollakin satunnaisotannan menetelmällä N:n alkion perusjoukosta (N on siis perusjoukon koko).
- Perusjoukon yksittäinen alkio (tilastoyksikkö) k tulee poimituksi n:n alkion otokseen (tutkimusaineistoon) tunnetulla **sisältymistodennäköisyydellä** π_k,

$$0 < \pi_k \leq 1, \quad k = 1, ..., N,$$

  jossa siis N on perusjoukon alkioiden lukumäärä. Toisin sanoen, kaikilla perusjoukon alkioilla on oma nollaa suurempi todennäköisyytensä (voi olla 1), π_k, tulla poimituksi otokseen.
  - Sisältymistodennäköisyys voi olla sama kaikille perusjoukon alkioille tai vaihdella perusjoukon eri osajoukkojen (alkioryhmien) välillä. Tämä tulee huomioida otantamenetelmän valinnassa, jotta saadun otoksen edustavuus ei vaarannu.
  - Sisältymistodennäköisyyttä voidaan käyttää monimutkaisemmassa otantateoriassa **asetelma-** ja **analyysipainojen** muodostamisessa sekä uudelleenpainotuksessa (vastauskadon korjaus).
- Tässä luvussa käsitellään erilaisia perinteisiä otantamenetelmiä sekä sitä, minkälaisten perusjoukkojen tilanteissa mikäkin otantamenetelmä on sopivin.
  - **Yksinkertainen satunnaisotanta** (YSO): perinteisin otantamenetelmä, jossa jokaisella tietyn kokoisella otoksella sama mahdollisuus tulla valituksi.
  - **Systemaattinen otanta** (SYS):
  - **Ositettu otanta**: perusjoukko (populaatio) jaetaan ominaisuuksiltaan yhtenäisiin eli homogeenisiin **ositteisiin**, joista jokaisesta poimitaan erillinen otos.
  - **Ryväsotanta** tai joskus myös **moniasteinen otanta**: Hyödynnetään perusjoukossa esiintyvää kerroksellisuutta, eli hierarkkisuutta otannassa.

2.5.1 Yksinkertainen satunnaisotanta

- **Yksinkertaisessa satunnaisotannassa** (YSO) jokaisella tilastoyksiköllä (perusjoukon alkiolla) on nollasta poikkeava todennäköisyys tulla valituksi otokseen.
  - Otanna satunnaisuus tulee siis siitä, että jokainen tilastoyksikkö poimitaan otokseen *satunnaisesti*! (Ks. luku 1)
  - YSOa pidetään otannan perusmuotona, jossa jokaisella perusjoukon alkiolla on lähtökohtaisesti yhtä suuri todennäköisyys tulla valituksi otokseen.
    - Yksinkertainen satunnaisotanta on periaatteiltaan intuitiivinen ja helppo ymmärtää. Lisäksi se on tietyissä tilanteissa usein helppo toteuttaa.
  - Tällöin on selvää että myös jokaisella perusjoukon samankokoisella osajoukolla on sama todennäköisyys tulla valituksi.
  - Toisin sanoen, todennäköisyys tulla poimituksi ei riipu tilastoyksikön ominaisuuksista tai siitä minkälaisia ominaisuuksia jo poimituilla otosyksiköillä on.
  - Satunnaisotanta siis selvästi korjaa valikoitumisharhaa (viittaus aiempaan lukuun) satunnaistamalla otokseen valikoitumisen täysin! YSO voidaankin aina tulkita arvonnaksi. Käytännön työssä arvonta onkin oiva satunnaistamisen keino.
- **YSO:n toteuttaminen**
  - Käytännössä yksinkertainen satunnaisotanta etenee vaiheittain:
    - Tutkimuksen alussa tutkijalla tulisi olla käytettävänään (ts. tulisi koostaa) lista kaikista perusjoukon havaintoyksiköistä (alkioista). Tämä muodostaa tutkimuksen **otantakehikon**.

    - Tämän jälkeen jokaiseen perusjoukon alkioon voidaan liittää numeeriset tunnukset.
    - Sitten valitaan haluttu otoksen koko. Otoskoon määrittäminen on keskeinen osa koesuunnittelua, ks. luku **??**
    - Otantakehikosta arvotaan perusjoukon alkiot otokseen yksi kerrallaan.
    - Käytännössä arvonta voidaan toteuttaa satunnaislukuja generoimalla (tuottamalla) niin että jokaisen otantakehikon alkion sisältymistodennäköisyys on yhtä suuri.[4]

- YSO:n **poimintastrategiat**: Käytännössä yksinkertainen satunnaisotanta voidaan suorittaa kahdella eri tavalla: **palauttaen** tai **palauttamatta**.
  - Tarkastellaan, aiemman mukaisesti, **äärellistä populaatiota** (perusjoukkoa), jossa on N alkiota ja tarkoituksena on poimia n:n alkion kokoinen otos (huom. $n < N$). Olkoon i yksittäisen alkion indeksiluku (ts. jokainen alkio on numeroitu esimerkiksi tavalla $i = 1, \ldots, N$).

YSO:n poiminta palauttaen

- Kun poiminta suoritetaan **palauttaen**, niin poimittu alkio palautetaan aina ennen uuden alkion arpomista takaisin perusjoukkoon, jolloin alkio voi tulla poimituksi otokseen useita kertoja.
  - Kyseessä on siis otanta **takaisinpanolla** (with replacement).
  - Tällöin alkioiden arvonnat ovat riippumattomia: alkion todennäköisyys tulla poimituksi otokseen ei riipu siitä kuinka monta alkiota otokseen on jo poimittu.
  - Alkion i sisältymistodennäköisyys on tällöin selvästi

$$\pi_i = \frac{1}{N}, \quad \forall\, i$$

- Otantaan palauttaen liittyviä todennäköisyyksiä hallitaan **binomijakauman** avulla (ks. luku 4), joka johtaa yksinkertaiseen **tilastolliseen malliin** YSO:a käytettäessä.
- Poiminta palauttaen, tai otanta takaisinpanolla, on toisaalta varsin epärealistinen otantamenetelmä useassa tutkimuksessa. Esimerkiksi lienee mahdotonta testata samaa lääkettä useaan otteeseen samaan aikaan yhdellä koehenkilöllä.

YSO:n poiminta palauttamatta

- Kun poiminta suoritetaan **palauttamatta**, poimittua alkiota ei palauteta perusjoukkoon poiminnan jälkeen eikä se täten voi tulla poimituksi otokseen kuin kerran.
  - Kyseessä on siis otanta **ilman takaisinpanoa** (without replacement).
  - Tällöin alkioiden arvonnat eivät enää ole riipumattomia: alkion todennäköisyys tulla poimituksi otokseen riippuu siitä kuinka monta alkiota otokseen on jo poimittu.
  - Alkion i sisältymistodennäköisyys on tällöin vastaavasti

$$\pi_i = \frac{1}{N - A_i},$$

- Tässä A_i on jo poimittujen alkioiden lukumäärä ennen kyseistä **otositeraatiota**: ensimmäisen poiminnan kohdalla $A_i = 0$, toisen kohdalla $A_i = 1$ ja niin edespäin.
  - Ilman takaisinpanoa populaatiosta voidaan poimia $\binom{N}{n}$ erilaista otosta.[5]
  - Otantaan palauttamatta liittyviä todennäköisyyksiä hallitaan **hypergeometrisen jakauman** avulla, joka johtaa (melko) yksinkertaiseen **tilastolliseen malliin** YSO:a käytettäessä.

[4] Satunnaislukujen generointia käsitellään ja opetellaan mm. R-kurssilla ja kurssilla TILM3705 Johdatus laskennalliseen tilastotieteeseen.

[5] Kun otosyksiköiden järjestyksellä ei ole merkitystä. $\binom{N}{n}$ on ns. binomikerroin, joka saadaan kaavasta $\binom{N}{n} = \frac{N!}{n!(N-n)!}$, jossa $N! = N \cdot (N-1) \cdot (N-2) \cdots 1$ on N:n kertoma.

Esimerkki: Yksinkertaisen satunnaisotannan poimintastrategiat

- Esimerkki: Poimitaan palloja kulhosta satunnaisesti.
  - Jos yksittäinen pallo (alkio) voi tulla poimituksi useammin kuin kerran, eli pallo palautetaan kulhoon sen poiminnan jälkeen, on kyseessä yksinkertainen satunnaisotanta takaisinpanolla.
  - Vastaavasti jos pallo voi tulla valituksi vain kerran, eli pallo poistetaan kulhosta sen poiminnan jälkeen, on kyseessä otanta ilman takaisinpanoa.

Otoskoon vaikutus YSO:n

- Yksinkertaisen satunnaisotannan erot takaisinpanolla ja ilman takaisinpanoa riippuvat otantakehikon (tai yleisemmin perusjoukon) koosta. Mikäli poimittava otos muodostaa suuren osan perusjoukosta (ts. $\frac{n}{N}$ on "suuri", eli lähellä yhtä) menetelmät poikkeavat olennaisesti.
- Toisaalta, jos perusjoukko on ääretön niin menetelmillä ei ole käytännössä eroa (ts. kun $N \longrightarrow \infty$ niin $\frac{n}{N} \longrightarrow 0$ eli todennäköisyys että sama alkio poimittaisiin otokseen useammin kuin kerran lähestyy nollaa otoskoon lähestyessä ääretöntä).
  - Monesti onkin (teoreettiselta) kannalta järkevää olettaa että otos poimitaan äärettömästä perusjoukosta vaikka perusjoukko tosiasiallisesti olisikin äärellinen (mutta riittävän "iso").
  - Tällöin voidaan olettaa käytettävän otantaa takaisinpanolla, sillä siinä käytettävät tilastolliset mallit ovat yksinkertaisempia kuin otannassa ilman takaisinpanoa ja tämä helpottaa tilastollisessa päättelyssä käytettäviä kaavoja.

_YSO: Potentiaaliset ongelmat___

- Monissa tapauksissa ei kuitenkaan ole helppoa saada listaa kaikista perusjoukon havaintoyksiköistä (jolloin menetelmän käyttö on mahdotonta).
- Kyselytutkimuksissa perusjoukko on usein suuri ja laajalle alueelle hajaantunut. Henkilökohtaisten, kasvotusten toteutettavien, haastattelujen tekeminen vaatisi suuria resursseja (haastattelijat joutuisivat esim. matkustamaan ympäri Suomea satunnaisotokseen valikoituneiden henkilöiden asuinpaikkojen mukaan).
- Tällaisissa tutkimustilanteissa käytetäänkin usein muunlaisia otantamenetelmiä.

2.5.2 Systemaattinen otanta

- Systemaattisessa, eli tasavälisessä, otannassa poimintakehikkoon (perusjoukkoon) kuuluvat alkiot järjestetään jonoon ja siitä poimitaan otokseen joka k. alkio.
  - Esimerkiksi jos oletetaan että perusjoukkoon kuuluu 1000 tilastoyksikköä ja valittu otoskoko on 100, niin otos voidaan poimia perusjoukon alkioiden järjestetystä listasta poimimalla siitä joka kymmenes yksikkö.
  - Systemaattinen otanta ei oikeastaan kuulu satunnaisotannaksi laskettaviin menetelmiin, koska siinä ei sovelleta arvontaa.
  - Yksinkertainen satunnaisotanta voidaan kuitenkin nähdä systemaattisen otannan erikoistapauksena (eli systemaattinen otanta voidaan toteuttaa satunnaisotantana), missä perusjoukon alkiot järjestetään jonoon **satunnaistamalla**. - Ts. jonon järjestys on satunnainen, eli joka k. jonon alkio on "satunnaisotos" otantakehikosta.
  - Systemaattinen otanta tuottaa tällöin samat johtopäätelmät kuin yksinkertainen satunnaisotanta, jos perusjoukon alkioiden järjestys on tutkittavan ilmiön kannalta satunnainen! Toisin sanoen, harhaa ei synny mikäli perusjoukon alkioiden järjestys ei riipu sellaisesta ominaisuudesta, jota tutkitaan.

  - Systemaattisen otannan suhteen potentiaaliseksi ongelmaksi muotoutuu havaintoyksikkölistan mahdollinen säännöllinen jaksollisuus, jota se ei havaitse ja jolloin satunnaisotanta toimisi (kenties) paremmin.
    * Ongelmia syntyy esimerkiksi silloin, jos tiedot perusjoukosta koostuvat pariskunnista ja poimintaintervalli on parillinen luku. Tällöin seurauksena voi olla, että otokseen saattaisi valikoitua ainoastaan joko miehiä tai naisia.

- Myös systemaattisessa otannassa tarvitaan siis lista tai rekisteri kaikista perusjoukon havaintoyksiköistä ja sitä sovelletaankin tavallisesti YSO:n sijasta silloin, kun perusjoukon alkioista on käytettävissä tietorekisteri, luettelo tai havaintoja kerätään ajassa tai tilassa.
  - Esimerkiksi mielipidekyselyn kohteet poimitaan (voitiin poimia) puhelinluettelosta (tai vastaavasta rekisteristä) valitsemalla haastateltavaksi jokaiselta aukeamalta ensimmäisenä esiintyvä henkilö tai jotain tuotetta valmistavan tehtaan laaduvalvonnassa valitsemalla laatuarviointiin joka sadas tuote, joka hihnalta valmistuu. Muita esimerkkejä ovat esim. liikenteen, jäsenrekisteriin tai kassajonossa seisovien otantayksiköiden poiminta otokseen.

2.5.3 Ositettu otanta

- Ositettu otanta on sopiva menetelmä tilanteisiin, joissa perusjoukko koostuu jonkin ominaisuuden suhteen homogeenisista ryhmistä, ts. alkioryhmistä (osista). Ositettu otanta pyrkii varmistamaan, että tutkittava otos on edustava kaikkien (tutkimuksen kannalta) olennaisten ryhmien osalta.
  - Esimerkiksi jos tavoitteena on tutkia jonkin maan erilaisten ja usein hyvin eri kokoisten kieliryhmien taloudellista asemaa. Kaikista ryhmistä tulisi saada edustava otos.
  - Tällöin maan koko populaatioon kohdistettu yksinkertainen satunnaisotanta ei olisi järkevää, sillä otoskoon pitäisi olla (todennäköisesti) hyvin suuri, että jokaisesta kieliryhmästä saataisiin poimittua edustava otos.
  - Ositetun otannan avulla otos voitaisiin kerätä niin, että jokaisesta ryhmästä (ositteesta) poimitaan osaotos yksinkertaisella satunnaisotannalla tai systemaattisella otannalla ja nämä osaotokset yhdistetään yhdeksi otokseksi.
- Ositettu otanta voi (oikein toteutettuna ja sopivassa asetelmassa) tuottaa paljon tarkempaa tietoa kuin yksinkertainen satunnaisotanta samaa otoskokoa käytettäessä! Voidaan esimerkiksi käyttää tietoa siitä, että otosyksiköt ovat joka ositteessa keskenään samankaltaisia.
- Ositetun otannan käyttöön suurissa kyselytutkimuksissa liittyy samoja ongelma kuin yksinkertaiseen ja systemaattiseen satunnaisotantaan.
  - Otokseen valikoituneet vastaajat voivat olla mm. levittäytyneinä suurelle maantieteelliselle alueelle. Näin ollen otannan suorittaminen vaatii suuria kustannuksia.
  - Onko (järkevä) osittaminen ylipäätään mahdollista toteuttaa tarkasteltavassa sovelluskohteessa?

2.5.4 Ryväsotanta

- Ryväsotanta soveltuu tilanteisiin, joissa perusjoukko on “ryvästeistä” eli se voidaan jakaa luonnollisiin ryhmiin eli rypäisiin (eng. *clusters*).

- Rypäkset indikoivat aineiston luontaista hierarkkista, eli monitasoista- tai asteista rakennetta.
  - Esimerkkejä tällaisista ryhmistä ovat erilaiset yritykset tai koululuokat. Esimerkiksi yritykset muodostavat luonnollisesti eri ryppäitä, joiden alkiot ovat työntekijöitä ja koululuokat muodostavat koulun sisällä omia luonnollisia ryppäitään ja opiskelijat ovat alkioita näissä ryppäissä.

- Huomionarvoista onkin, että toisin kuin ositetussa otannassa, ryväsotannassa ryppäiden oletetaan olevan toistensa kanssa riittävän samankaltaisia, että jokaista rypästä ei tarvitse erikseen tutkia.

  - Tämä onkin yksi ryväsotannan tärkeimpiä motivointeja, sillä sitä usein perustellaan kustannustehokkuudella: sen sijaan että poimitaan satunnaisia koululaisia mahdollisesti suuresta määrästä kouluja, voidaan poimia satunnaisia ryppäitä (kouluja), joista tutkimusyksiköt eli koululaiset poimitaan.
  - Lisäksi koulun sisällä koululuokat muodostavat aliryppäitä, joista voidaan edelleen poimia satunnaisotos, jotta päästään tutkimaan perusjoukon alkioita eli koululaisia esim. haastattelututkimuksen muodossa.
  - Tavoitteena on vähentää tietojen keruun aiheuttamia kustannuksia samalla varmistaen, että otos on kuitenkin mahdollisimman edustava!

- Ryväsotannan voi suorittaa **yksi-** tai **kaksivaiheisena** (**yksiasteinen/kaksiasteinen ryväsotanta**).
  - **Kaksivaiheisessa ryväsotannassa**
    * **Ensimmäisessä vaiheessa** poimitaan joukko ryppäitä kaikkien ryppäiden joukosta, eli vain osa ryppäistä on mukana lopullisessa otoksessa.
    * **Toisessa vaiheessa** poimitaan ensimmäisessä vaiheessa poimituista ryppäistä alkiotason otokset.
  - **Yksivaiheisessa ryväsotannassa** toisessa vaiheessa valitaan kaikki ensimmäisen vaiheen otosryppäiden alkiot, jolloin toisen vaiheen otanta typistyy ensimmäisen vaiheen ryppäiden alkioiden kokonaistutkimukseksi.
  - Poiminnan eri vaiheissa voidaan soveltaa yksinkertaista satunnaisotantaa tai systemaattista otantaa.

- Ryväsotantaa käytetään usein suuria haastattelututkimuksia tehtäessä. Erityisesti, ryväsotantaa voidaan hyödyntää myös silloin, kun tutkijalla ei ole käytettävissään kattavaa listaa kaikista havaintoyksiköistä, mutta näiden muodostamat ryppäät on määritettävissä.

- Ryväsotannan heikkoutena pidetään sitä, ettei aina ole helppoa muodostaa rypäitä, jotka ovat toistensa kaltaisia. Tulosten tarkkuus myös riippuu moninpaikoin siitä, kuinka hyvin rypäisiin jako onnistuu.

Esimerkkejä ryväsotannasta

- Esimerkki 1:
  - Poimitaan oppilaitoksen opiskelijoista otos arpomalla ensin otos luokkahuoneista (=rypäistä).
  - Arvotuissa luokkahuoneissa käydään sitten suorittamassa kysely.
    * Esim. Oppilaitoksen opiskelijoista voidaan poimia otos arpomalla ensin otos luokkahuoneista, jolloin luokkahuoneet ovat nk. ryppäitä.
    * Mahdollisia ongelmia? Miten huomoida päivä- ja iltaopiskelijat? Tämän voisi toteuttaa arpomalla otos luokkahuoneista päiväsaikaan ja toinen otos ilta-aikaan. Tässä yhdistetään ryväsotantaan ositettu otanta, jolla taataan päivä- ja iltaopiskelijoiden edustus.
- Esimerkki 2: Tutkittaessa tänä vuonna peruskoulun aloittavia voidaan ensin poimia otos kouluista, jolloin koulut ovat rypäitä. Tämän jälkeen arvotaan kustakin otokseen tulleesta koulusta tietty määrä tutkimuksen kohderyhmään kuuluvia oppilaita.

2.6 Otantaesimerkkejä

Esimerkki: Työllisyys ja työttömyys, Tilastokeskuksen työvoimatutkimus

- Työvoimatutkimus on otostutkimus, jonka avulla tilastoidaan 15–74-vuotiaan väestön työmarkkinoille osallistumista, työllisyyttä, työttömyyttä ja työaikaa (yhden viikon aikana) kuukausittain, neljännesvuosittain ja vuosittain.
  - Työvoimatilastoja käytetään työvoimapoliittisten ennusteiden ja suunnitelmien laadinnassa, toimien seurannassa ja päätöksenteon tukena.
  - Työmarkkina-aseman perusluokittelussa väestö jaetaan työllisiin, työttömiin ja työvoiman ulkopuolisiin.
    - Työlliset ja työttömät muodostavat työvoiman.
  - Työvoimatutkimuksen **perusjoukon** muodostavat Suomessa vakinaisesti asuvat 15–74-vuotiaat henkilöt.
  - Työvoimatutkimuksen otos poimitaan **ositetulla satunnaisotannalla** väestön keskusrekisteriin perustuvasta Tilastokeskuksen väestötietokannasta kahdesti vuodessa.
- Ositetun satunnaisotoksen poiminta:
  - Tutkimus on paneelitutkimus, jossa samaa henkilöä haastatellaan viisi kertaa.
  - Joka kuukauden otokseen kuuluu noin 12 000 henkilöä, keskimäärin noin joka 300. henkilö perusjoukosta.
  - Yhden tutkimuskuukauden otos koostuu viidestä rotaatioryhmästä, jotka ovat tulleet tutkimukseen mukaan eri aikoina. Otos vaihtuu asteittain siten, että kolmena peräkkäisenä kuukautena vastaamisvuorossa ovat eri henkilöt.
- Julkisuudessa seurataan useimmiten kuukausittain työllisyyden ja työttömyyden muutoksia edellisen vuoden vastaavasta kuukaudesta. Vaihtoehtoisesti voidaan käyttää kausitasoitettuja lukuja, jolloin tilannetta voidaan verrata edelliseen kuukauteen.

Esimerkki: Terveys 2000

- Terveys 2000 -tutkimuksen tavoite oli tuottaa ajankohtainen kattava kuva työikäisen ja iäkkään väestön terveydestä ja toimintakyvystä selvittämällä tärkeimpien terveysongelmien yleisyyttä ja syitä sekä niihin liittyvän hoidon, kuntoutuksen ja avun tarvetta.
- Tutkimus koskee (koski) 18 vuotta täyttänyttä Suomen aikuisväestöä (perusjoukko), josta valitaan valtakunnallisesti edustava 10 000 henkilön otos.
- Poimittiin kaksivaiheinen ryväsotos terveyskeskuspiireistä.
  - Ositus perustui yliopistosairaaloiden vastuualuiden väestömäärään suhteutettuun kiintiöintiin.
  - Suurimmat 15 terveyskeskuspiiriä poimittiin otokseen ja lopuista 65:stä piiristä poimittiin loppuotos kussakin ositteessa systemaattisella (PPS) otannalla (sisältymistodennäköisyys suhteessa alkion kokoon).

2.7 Otannan haasteita vielä kootusti

- Poimintaharha: Otos ei edusta populaatiota. Vaarana varsinkin silloin, kun otokseen tulleet populaation alkiot ovat valikoituneet tai ovat itse valinneet itsensä otokseen. Vastaavasti toisinaan otoksen peitto ei ole hyvä eli tällöin otanta ei kata koko perusjoukkoa tai se kattaa perusjoukon ja vähän muutakin.
- Jos poimitaan tutkimukseen ne perusjoukon alkiot, jotka ovat tutkimuksen tekemishetkellä 'saatavilla', niin kyseessä on **näyte**. Näyte ei siis kata ilmiön koko vaihtelua edustavan satunnaisotoksen tapaan.
  - Esimerkiksi perinteiset katukyselyt eivät ole hyvä otantatapa, sillä kadulla liikkujat eivät välttämättä kovin hyvin edusta tutkittavaa perusjoukkoa, ellei perusjoukkona ole kyseisellä kadulla kyseiseen aikaan liikkuvat ihmiset.
  - Jos television ajankohtaisohjelma pyytää katsojia twiittaamaan mielipiteensä ajankohtaisesta asiasta, kyseessä on itse valikoituva näyte (osallistujat valitsevat itse itsensä).
- Vajaapeittävyys: Populaation alkioista ei ole välttämättä täydellistä luetteloa
- Vastauskato: Tutkimuksen kohteita ei tavoiteta tai he kieltäytyvät vastaamatta. Kadon vuoksi lopullinen otoskoko saattaa jopa karsiutua pois tai jokin osajoukko on aliedustettuna.
- Vastausharha: Kysymykset voivat olla huonosti muotoiltuja tai vastaajat voivat antaa vääriä tietoja.