Turun yliopisto
University of Turku

PTKMY2

KUVAILEVA TILASTOTIEDE 3 OP

Jani Tolonen

Turun Kauppakorkeakoulu, Porin yksikkö

Turun yliopisto
University of Turku

YHTEYSTIEDOT

Jani Tolonen

jttolo@utu.fi

Kurssimateriaali Moodlessa,

Kurssi: PTKMY2 (Tolonen, 2015)

Salasana: PTKMY2_2015

Ottakaa salasana talteen!

Turun yliopisto
University of Turku

KURSSIKUVAUS

Osaamistavoitteet:

Tavoitteena on, että opiskelija oppii ymmärtämään havaintoaineistoa kuvailevien analyysimenetelmien perusteet ja pystyy soveltamaan niitä käytännössä.

Sisältö:

Opintojaksolla käsitellään mm. seuraavia asioita: havaintoaineiston graafiset ja taulukkomuotoiset esitystavat, empiirisen jakauman tunnusluvut sekä regressioanalyysin perusteet. Lisäksi käsitellään mittaamiseen ja otantaan liittyviä peruskysymyksiä.

Turun yliopisto
University of Turku

KURSSIAIKATAULU

Luennot tiistaisin klo 12:15 - 13:45 ja keskiviikkoisin klo 10:15 - 12:45. Ensimmäinen luento on Ti 01.09.2015 ja viimeinen luento on Ke 7.10.2015.

Lisäksi opiskelija osallistuu yhteen harjoitusryhmään:

Ryhmä 1: Tiistaisin klo 14:15 - 15:45 (LK240A)

Ryhmä 2: Tiistaisin klo 16.15: - 17:45 (LK240A)

Ryhmä 3: Keskiviikkoisin klo 12:15 - 13:45 (LK240B)

Ryhmä 4: Keskiviikkoisin klo 14.15: - 15:45 (LK240B)

Harjoitukset alkavat jo ensimmäisellä viikolla! Harjoituksia on myös viikko luentojen jälkeen.

Turun yliopisto
University of Turku

HARJOITUKSET

Tenttioikeuden saa kun kerännyt riittävän määrän harjoituspisteitä (= 25 harjoituspistettä).

Osallistuminen 1. viikon harjoituksiin antaa 2 harjoituspistettä, muuten 1 kotitehtävä=1 harjoituspiste. Ylimenevästä 25 ylittävästä pistemäärästä saa maksimissaan 4 lisäpistettä tentissä.

Harjoitukset voi palauttaa Moodleen. Etäpalautuksen deadline on pitävä.

Turun yliopisto
University of Turku

HARJOITUKSET

- **Etäpalautuksissa pisteitä vain oikein tehdyistä tehtävistä. (Toki inhimillisyys on läsnä).**
- **Harjoituksiin osallistuvilta myös riittävän hyvästä yrityksestä saa pisteet.**
- **Harjoituksissa esitetään oikeat vastaukset.**
- **Mallivastaukset tulevat nettiin myöhemmin.**

Turun yliopisto
University of Turku

MATERIAALIT

- **Luentojen materiaali on tulostettavissa edellisenä päivänä.**
- **Harjoitustehtävät jaetaan edellisellä viikolla sekä ovat ladattavissa moodle alueelta.**
- **Tenttialue on luentokalvot, luentojen esimerkit sekä harjoitustehtävät.**
- **Tentit: 23.10., 4.12., 15.1.**
- **Kurssin opettajana ja demonstraattorina toimii Jani Tolonen.**
- **Vastaanotto luentojen ja demojen yhteydessä sekä sähköpostitse jttolo@utu.fi**

Turun yliopisto
University of Turku

KURSSIN TUEKSI

Netistä löytyy paljon hyvää materiaalia, kuitenkin eri maissa ja tieteissä jotkin merkintätavat ja käytänteet voivat erota toisistaan.

Kirjallisuudesta esimerkiksi:

- **Holopainen, *Tilastolliset menetelmät***
- **Grönruus, *Johdatus tilastotieteeseen***

Turun yliopisto
University of Turku

LASKIMISTA

- **Tentissä omien laskimien käyttö on kielletty.**
- **Demojen ratkaisuun voi hyödyntää Exceliä yms., mutta tentissä vaaditaan kykyä ratkaista tehtävät laskinta hyödyntämällä.**
- **Harjoitustehtävien palautuksissa täytyy nähdä miten laskut on laskettu!**

Turun yliopisto
University of Turku

KURSSISISÄLTÖ

Tilastollisen tutkimuksen rakenne, Mittaaminen ja mitta-asteikot, Otantamenetelmät. Yksiulotteisen jakauman kuvailu: graafiset esitystavat, taulukot, kuvailevat tunnusluvut.

Kaksiulotteisen jakauman kuvailu: graafiset esitystavat, ristiintaulukointi, tilastollinen riippuvuus ja sen tunnusluvut, regressioanalyysi.

Turun yliopisto
University of Turku

MOTIVAATIO?

- **Pohja TKMY3 kurssin asioille.**
- **Tilastotieteen merkitys kasvaa, koska dataa on yhä enemmän.**
- **Menetelmät yhä laajemmin käytössä.**
- **Osaamista vaaditaan, osaajille tarvetta työelämässä.**
- **Markkinointi (asiakkaiden segmentointi), rahoitus (aikasarjat), tuotanto (laadunvalvonta), lääketeollisuus, vakuutusala, ministeriöt…**

Turun yliopisto
University of Turku

MITÄ ON TILASTOTIEDE?

Tilastotieteeseen kuuluvat reaalimaailman ilmiöitä koskevan, havaintoihin perustuvan numeerisen tiedon kerääminen ja tietojen hankinnan suunnittelu.

Tilastotiedettä ovat myös näiden tietojen käsittely, analysointi, esittäminen ja tulkinta. Tilastotieteen keskeisin tavoite kuitenkin on sellaisten käsitteiden ja menetelmien kehittäminen, joita voidaan käyttää edellä mainituissa tehtävissä.

Tilastotiede on luonteeltaan menetelmätiede, mikä tarkoittaa, että se tuottaa välineistöä soveltavien tieteiden tarpeisiin.

Turun yliopisto
University of Turku

MITÄ ON TILASTOTIEDE?

- **Tilastotieteellisiä menetelmiä kutsutaan eri nimillä riippuen tieteenalasta, myös eri traditioilla on omia termejä ja merkintätapoja samoille asioille.**
- **Esim: ekonometria (taloustiede), psykometria (Psykiatria ja psykologia), biostatistiikka (Bio- ja lääketieteet) sosiometria (sosiologia), demometria (väestötiede), epidemologia jne…**

Turun yliopisto
University of Turku

TILASTOTIETEEN REUNA-ALUEITA

- **Finanssimatematiikka**
- **Hahmontunnistus**
- **Hermoverkot**
- **Kaaosteoria**
- **Katastrofiteoria**
- **Kuvankäsittely**
- **Kybernetiikka**
- **Operaatioanalyysi**
- **Peliteoria**
- **Päätösteoria**
- **Riskiteoria**
- **Signaalinkäsittely**
- **Stokastiset prosessit**
- **Todennäköisyyslaskenta**
- **Tulevaisuudentutkimus**
- **Vakuutusmatematiikka**

Turun yliopisto
University of Turku

TILASTOLLISEN TUTKIMUKSEN RAKENNE

1. Tutkimuksen suunnittelu

Valitaan tutkimuskohde ja rajataan tutkimusongelma. Asetetaan tutkimuksen tavoitteet, eli selvitetään, mihin kysymyksiin tutkimuksen toivotaan antavan vastauksen.

2. Havaintoaineiston eli datan hankinta

Määritetään se kohdepopulaatio, josta tiedot kerätään, ja suoritetaan tietojen kerääminen.

3. Aineiston esittäminen ja tiivistäminen

Esitetään saatu aineisto taulukkoina tai graafisesti (kuvioina).

4. Tilastollisen mallin soveltaminen ja hypoteesien testaus. Sovelletaan aineistoon jotakin tai joitakin analysointimenetelmiä.

5. Johtopäätösten tekeminen

Tehdään analyysin tuloksista tiivistetty sanallinen tulkinta, joka vastaa tutkimuksen alussa tehtyihin kysymyksiin. Arvioidaan tulosten luotettavuutta.

Turun yliopisto
University of Turku

YMMÄRRYS

Tiedon saaminen vaatii dataa. Data, joka on hyvin jäsennelty ja ymmärretty kutsutaan informaatioksi. Kaikki data ei silti ole tietoa. Jos datassa on tietoa, niin datan lisääntyminen lisää ymmärrystä aiheesta.

Voidaan ajatella, että polku jonkin ilmiön ymmärrykseen etenee seuraavasti:

Data -> informaatio -> tieto -> ymmärrys

Turun yliopisto
University of Turku

DESKRIPTIIVINEN TILASTOTIEDE

Kuvaileva tilastotiede tarkoittaa tilastollisten aineistojen yleispiirteiden kuvaamista ilman pyrkimystä yleistysten tekemiseen.

Toinen tilastotieteen osa edustaa ns. tilastollista päättelyä, jossa pienempien aineistojen, otosten, avulla tehdään yleistyksiä suurempiin joukkoihin, perusjoukkoihin. (TKMY3)

Turun yliopisto
University of Turku

MITTAAMINEN

Voidaan ymmärtää käsitteenä laajasti. Esimerkkejä mitattavista ominaisuuksista:

- **Pituus ja paino**
- **Yrityksen liikevaihto**
- **Mielipide**
- **Liiketoiminnan riskit**

Kaikkea ei voi mitata suoraan yhdellä mittarilla. Tällaiset ominaisuudet ovat ns. latentteja ominaisuuksia. Näitä mitataan ominaisuuden indikaattorien avulla. Esimerkkinä latentista ominaisuudesta voidaan mainita muodikkuus.

-> Indikaattori-mittarit

Turun yliopisto
University of Turku

TILASTOLLISET MUUTTUJAT

Muuttujien jaotteluja:

- **Jatkuvat ja epäjatkuvat (diskreetit) muuttujat**
- **Jos muuttuja voi periaatteessa saada annetulta lukuväliltä minkä arvon hyvänsä, se on jatkuva.**
- **Numeeriset ja kategoriset muuttujat**
- **Jos yksiköiden väliset erot mitattavan ominaisuuden suhteen ovat määrällisiä eli kvantitatiivisia, puhutaan numeerisesta muuttujasta.**
- **Jos erot taas ovat laadullisia eli kvalitatiivisia, muuttuja on kategorinen. (Paidan väri, kansallisuus, nimi)**
- **Kategoristen muuttujien arvoilla ei voi suorittaa laskutoimituksia. (Esim. mikä on luokassa olevien paitojen värien keskiarvo?)**

Turun yliopisto
University of Turku

MITTA-ASTEIKOT: TILASTOLLISET MUUTTUJAT

- **Luokittelu- eli nominaaliasteikko** (Väri)
- **Järjestys- eli ordinaaliasteikko (**Sija kilpailussa**)**
- **Välimatka-asteikko (**lämpötila celsius-asteikolla**)**
- **Suhdeasteikko (esim. paino)**
- **Kaksi ensimmäistä ovat ns. kategorisia mitta-asteikkoja ja jälkimmäiset numeerisia.**
- **Muuttujan mitta-asteikko vaikuttaa käytettävissä olevien tilastollisten menetelmien valikoimaan! Ero on merkittävin kategoristen ja numeeristen mitta-asteikkojen välillä.**

Turun yliopisto
University of Turku

LUOKITTELUASTEIKKO

Luokitteluasteikollisen muuttujan arvot sisältävät tiedon ainoastaan siitä, mihin luokkaan havainto kuuluu.

- **Esimerkkejä: Sukupuoli, toimialat.**
- **Vaikka itse muuttuja on kategorinen, talletetaan niihin liittyvät arvot tilasto-ohjelmissa numeerisina koodilukuina.**
- **Sallitut muunnokset koodiluvuille: bijektio.**

Tämä tarkoittaa käytännössä sitä, että muuttujan erilaisten arvojen koodilukujen on oltava erilaisia.

Turun yliopisto
University of Turku

JÄRJESTYSASTEIKKO

Järjestysasteikollisen muuttujan arvot sisältävätl uokittelun lisäksi myös tiedon arvojen järjestyksestä.

- **Esimerkkejä: Ns. asenneasteikot, koulutustasot**
- **Vaikka itse muuttuja on kategorinen, talletetaan niihin liittyvät arvot tilasto-ohjelmissa numeerisina koodilukuina.**
- **Sallitut muunnokset: aidosti kasvava. Käytännössä tämä tarkoittaa sitä, että mitattavan ominaisuuden kasvaessa on koodiluvun oltava suurempi.**

Turun yliopisto
University of Turku

VÄLIMATKA-ASTEIKKO

- **Välimatka-asteikko on numeeristen muuttujien mitta-asteikko. Muuttujan arvojen nollakohta on sopimuksenvarainen.**
- **Esimerkkejä: Lämpötila (°C, °F), vuosiluku, .**
- **Sallitut muunnokset: lineaarinen (*Y=aX+b, a>0*).**

Käytännössä tämä tarkoittaa sitä, että muuttujan arvojen välimatkojen suhteet säilyvät. (esim. °C -> °F)

Turun yliopisto
University of Turku

SUHDEASTEIKKO

- **Välimatka-asteikko on numeeristen muuttujien mitta-asteikko.**

Muuttujan arvojen nollakohta on absoluuttinen.

- **Esimerkkejä: rahamäärä**
- **Sallitut muunnokset: lineaarinen, jossa nollakohta säilyy (*Y=aX, a>0*). Käytännössä tämä tarkoittaa sitä, että muuttujien arvojen suhteet säilyvät.**
- **Muunnos, joka on sallittu mitta-asteikolle, on sallittu myös sitä alemmille mitta-asteikoille.**

Turun yliopisto
University of Turku

Luokittelu-asteikko

Ei voida laittaa järjestykseen

Värit

Perusmaut

Järjestys-asteikko

Ei voida sanoa, että kuinka suuri etäisyys on eri arvojen välillä.

Esimerkiksi kulta, hopea ja pronssi sijat

Välimatka-asteikko

Voidaan kertoa, että kuinka monta yksikkö suurempi kuin toinen, mutta ei absoluuttista nollapistettä

Esimerkiksi lämpötila celsiusasteissa

Suhde-asteikko

Absoluuttinen nollapiste, voi kasvaa

Esimerkiksi yksilön pankkitilillä oleva rahamäärä

Turun yliopisto
University of Turku

MITTAUSTEN VIRHELÄHTEET

- **Käytännössä mittaamisen yhteydessä syntyy aina mittausvirhettä mittausvälineiden rajallisen tarkkuuden ja mittausprosessiin vaikuttavien häiriötekijöiden vuoksi.**
- **Jos tietyllä menetelmällä toisistaan riippumattomat samalle tilastoyksikölle tehdyt mittaukset antavat huomattavasti vaihtelevia tuloksia, sanotaan, että mittarin reliabiliteetti on alhainen.**
- **On myös varmistuttava siitä, että mittari mittaa todella sitä ominaisuutta, jota sen on tarkoitus mitata, eli siitä, että mittari on validi. Epävalidi mittari antaa systemaattisesti virheellisiä eli harhaisia tuloksia.**

Turun yliopisto
University of Turku

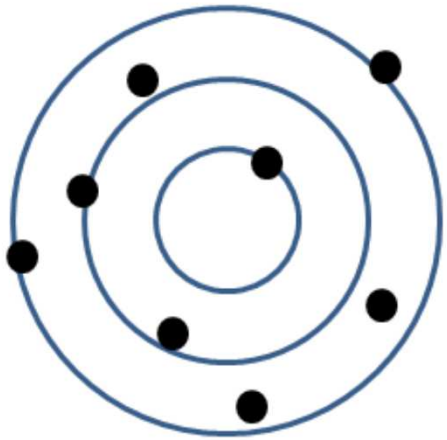
RELIABILITEETTI / VALIDITEETTI

Mittarin **validiteetti** eli mitataanko mitä on tarkoitus mitata. Esim ÄO-testit, koneoppiminen (auto/tankki), sosiaalinen tasa-arvo, masennustesti.

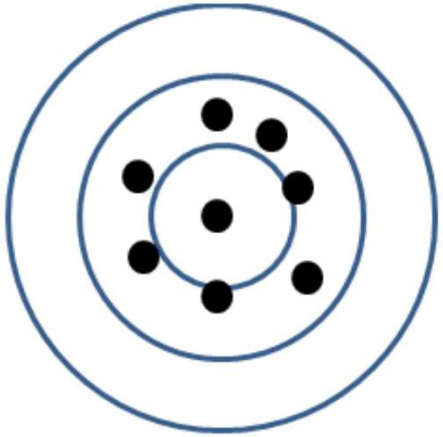
Mittarin **reliabiliteetti** eli mittaako aina samalla tavalla. Kotilämpömittari (uuni, takapiha, Alaska), valintakokeen haastattelijan pisteytys, HIV-testi.

Turun yliopisto
University of Turku

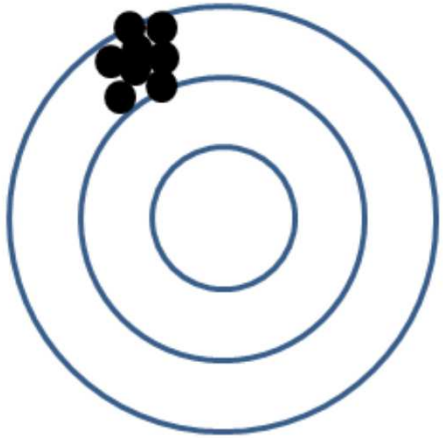
Ei validi, ei reliaabeli



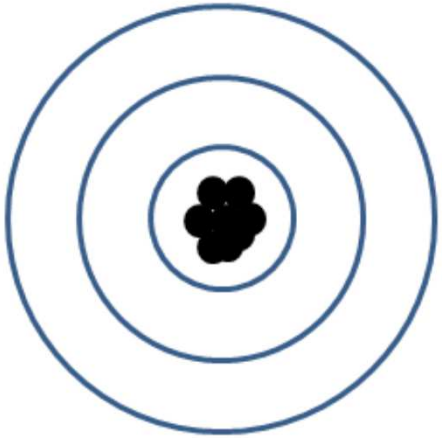
Validi, ei reliaabelia



Ei validi, mutta reliaabeli



Validi ja reliaabeli



Turun yliopisto
University of Turku

VALE, EMÄVALE, TILASTO?

- **Kuvilla ja tilastoilla voidaan johtaa harhaan.**
- **Usein tämä on tarkoituksenmukaista esimerkiksi mainontatarkoitukset tai propaganda.**
- **Vaatii lukijalta ymmärrystä tilastotieteestä.**
- **Tilastotiede pyrkii luomaan yhteisiä pelisääntöjä ja tekee suosituksia (mm. tilastokeskus)**

Turun yliopisto
University of Turku

OTANTA

- **Kokonaistutkimus vs. otantatutkimus**
- **Kokonaistutkimus on luotettavin tapa tutkia asiaa.**
    - Usein kallista
    - Joskus mahdotonta (esim. lääketiede)
    - Useimmiten liian järeä työkalu
    - Ratkaisuna on otoksen tekeminen

Turun yliopisto
University of Turku

PERUSJOUKKO (POPULATION)

- **Joukko, josta päätelmiä tehdään.**
  - Koostuu alkioista
  - Esimerkiksi Suomalaiset, miehet tai Kauppakorkean Porin yksikön 27-vuotiaat naisopiskelijat, joilla on lapsia.
- **Täytyy määritellä ennen otoksen tekemistä. Otokseen valitaan osa tästä perusjoukosta, mutta ei ketään sen ulkopuolelta.**

Turun yliopisto
University of Turku

OTOS

Satunnaiset vs. ei-satunnaiset otantamenetelmät

- **Satunnaisotos = satunnaisella otantamenetelmällä hankittu otos**
  - Eli kaikilla yksiköillä on yhtä suuri todennäköisyys tulla valituksi otokseen.
- **Näyte = ei-satunnaisella otantamenetelmällä hankittu otos.**
  - Tietoisesti tietyn ominaisuuden omaavat yksiköt / yksilöt saavat suuremman todennäköisyyden tulla valituksi.
  - Esim. tyytymättömät asiakkaat.

Turun yliopisto
University of Turku

OTOS

- **Otanta sisältää aina virhettä: Satunnainen vs. systemaattinen harha.**
- **Satunnaisharha: esim. virheellisesti otokseen valittu yksilö**
- **Systemaattinen harha: osaa perusjoukosta ei oteta huomioon otosta laatiessa**
- **Lisäksi on olemassa nk. Karkea virhe**
  - Esimerkiksi viallinen mittauslaite tai väärin suoritettu mittaus.
- **Satunnaisia otantamenetelmiä: Yksinkertainen satunnaisotanta, systemaattinen otanta, ositettu otanta ja ryväsotanta.**

Turun yliopisto
University of Turku

KATO

- **Ongelmana sekä kokonais- että otantatutkimuksessa.**
- **Kaikki joiden pitäisi vastata kyselyyn eivät vastaa tai osa tutkimuskohteista jää tutkimatta.**
- **Ratkaisuna on esimerkiksi täydentää otosta, uudelleen tavoittelemalla kadon kohteita.**
    - Aina ei mahdollista, esim. muutto tai kuolema. Usein myös liian kallista tai muuten haitallista.

Turun yliopisto
University of Turku

YKSINKERTAINEN SATUNNAISOTANTA

- **Perusjoukon kaikilla alkioilla (esim. yksilöillä) on yhtä suuri todennäköisyys tulla valituksi.**
    - Esimerkiksi lotto.
- **Vastaa käytännössä arvontaa esimerkiksi satunnaislukujen avulla.**

Turun yliopisto
University of Turku

SYSTEMAATTINEN OTANTA (ESIM. PISA)

- **Asetetaan alkiot järjestykseen joko sattumanvaraisesti tai jonkin järjestyksen mukaan.**
- **Määritellään otoksen koon perusteella poimintaväli**
- k = N/n
- K (poimintavälin pituus), N (perusjoukon koko) ja n (otoksen koko)
- Ensimmäinen valinta pitää arpoa 1- k väliltä.
- Harha eli systemaattinen virhe? Syklisyys voi olla ongelma, mutta ei välttämättä.

Turun yliopisto
University of Turku

OSITETTU OTANTA (HOMOGEENISET OSITTEET)

- **Tasainen kiintiöinti, esim. Sukupuolet, kansallisuudet, puolueiden edustajat.**
- **Suhteellinen kiintiöinti, esim. asiakassegmenttien perusteella tai kunnan koon perusteella.**
- **Sama %-osuus joka ryhmästä (ositteesta).**
- **Paras kiintiöinti. Usein aina tästä piirteitä. Painotus suurkaupunkeihin, suuriin kouluihin yms. käytännön syistä.**

Turun yliopisto
University of Turku

MIKSI OSITTAA?

- **Voidaan varmistaa, että eri kansallisuudet, ikä-ryhmät tai muut ominaisuudet ovat sopivasti edustettuina otoksessa.**
- **Näin voidaan korjata perusjoukossa olevia ”vääristymiä”. Esimerkiksi voidaan valita 50% joilla on tietty sairaus ja 50% joilla ei ole. Vaikka sairautta esiintyy vain 1% perusjoukosta.**

Turun yliopisto
University of Turku

RYVÄSOTANTA (HETEROGEENISET RYPPÄÄT)

- Ensin toisensa poissulkevat ryppäät.
- Valitaan satunnaisotannalla vain tietyt ryppäät.
- Näihin ryppäisiin kohdistuu tutkimustoimenpiteet.
- Ryväs voisi olla esimerkiksi koulu.

Turun yliopisto
University of Turku

RYVÄSTÄMISEN HYÖDYT

- **Jos Pisa-testaus tehdään yksinkertaisen satunnaisotannan perusteella, niin tiimin pitäisi kulkea ympäri koulua ja järjestää testejä yksittäisille opiskelijoille.**
- **Fiksumpaa tutkia kokonaisia kouluja ja luokkia.**
- **Näin säästetään kustannuksia.**

Turun yliopisto
University of Turku

OTANTA ASETELMIA VOIDAAN MYÖS YHDISTELLÄ

Esimerkiksi Pisa-tutkimus.

Ensin valitaan systemaattisella otannalla koulut (ryppäät), jotka osallistuvat tutkimukseen. Tämän jälkeen koulun luokista valitaan ositetulla otannalla oikeat määrät eri ikäisiä henkilöitä.

Turun yliopisto
University of Turku

HARKINNAN VARAINEN OTANTA (NÄYTE)

- Turun yliopisto testaa opiskelijoita.
- Jokaisella alkiolla ei ole yhtä suuri todennäköisyys tulla valituksi.
- Onko käytännön haittaa (esim. havaintotutkimus / muistitutkimus)
- Helppo, halpa.

Turun yliopisto
University of Turku

OTOSKOKO

- Kun otos pienenee, niin sen informatiivisuus laskee.
- Kun otos kasvaa, niin sen informatiivisuus nousee tiettyyn pisteeseen asti.
- Raha (ei rahaa tutkita kaikkia), etiikka (kaikkia ei ole oikein tutkita), pragmatiikka (syntymä – kuolema)
- Päätelmissä virhemarginaali (voidaan laskea).

Turun yliopisto
University of Turku

KÄYTÄNNÖN TIEDONHANKINTAMENETELMIÄ

- **Postikyselyt**
- **Internet-kyselyt**
- **Puhelinhaastattelut**
- **Henkilökohtaiset haastattelut**
- **Valmiit tietokannat.**

AINEISTON ESITYS: HAVAINTOMATRIISI

Jos tutkittavia tilastoyksiköitä on n kpl ja mitattavia ominaisuuksia eli muuttujia p kpl, esitetään havaintomatriisi A muodossa

$$A = \begin{array}{c} \\ a_1 \\ .. \\ a_n \end{array} \begin{array}{c} \begin{array}{ccc} X_1 & .. & X_p \end{array} \\ \begin{pmatrix} x_{11} & .. & x_{1p} \\ .. & .. & .. \\ x_{n1} & .. & x_{np} \end{pmatrix} \end{array}$$

Tilastoyksiköltä a_i mitatut muuttujien $X_1, \ldots, X_p$ arvot, eli havaintoarvot $x_{i1}, \ldots, x_{ip}$ (matriisin i:s rivi) muodostavat tilastoyksikön a_i havaintovektorin kyseisten muuttujien suhteen. Matriisin A j:s pystyrivi eli sarake sisältää kaikilta n:ltä tilastoyksiköiltä mitatut muuttujan X_j arvot $x_{1j}, \ldots, x_{nj}$ ja sitä sanotaan muuttujan X_j jakaumavektoriksi. Käytännössä käytetään käsitteitä havainto ja jakauma.

KUVAILEVA TILASTOTIEDE: YKSIULOTTEINEN JAKAUMA.

- Yksiulotteisen jakauman tarkastelu merkitsee yksittäisen muuttujat tarkastelua eli havaintomatriisin sarakkeen tarkastelua.
- Taulukot: ns. frekvenssitaulukko.
- Graafiset esitystavat: Pylväs-, piirakka-, sironta- ja laatikkojanakuvioita.
- Kuvailevat tunnusluvut: keskiluvut, hajontaluvut, muut tunnusluvut.

FREKVENSSITAULUKOT

- Yleensä kategoristen muuttujien esitystapa.
- Koostuu absoluuttisista (f_i) ja/tai suhteellisista ($p_i = f_i/n$) frekvensseistä.
- Taulukko sisältää jokaiseen muuttujan arvoon liittyvät frekvenssit.
- Järjestysasteikollisen muuttujan arvot kirjoitetaan yleensä suuruusjärjestykseen.
- Suhteelliset frekvenssit voidaan esittää myös prosenttisina ($p_i\%$) .
- Absoluuttinen, suhteelliset ja prosenttiset frekvenssit voidaan esittää myös kumulatiivisina eli summafrekvensseinä.
- Muuttujan on tällöin oltava vähintään järjestysasteikollinen.

FREKVENSSITAULUKOT

Arvo L_i	*Frekvenssi* f_i	*Suhteellinen frekvenssi* $p_i = f_i/n$	*Prosenttinen frekvenssi* $p_i\,\% = (f_i/n)\ 100\ \%$
L_1	f_1	p_1	$100\ p_1$
L_2	f_2	p_2	$100\ p_2$
. . .	. . .	. . .	. . .
L_r	f_r	p_r	$100\ p_r$
yht.	n	*1*	*100*

FREKVENSSITAULUKOT

Arvo Li	*Frekvenssi* f_i	*Summa-Frekvenssi* $F_i = \sum_{j=1}^{i} f_j$	*Suhteellinen summafrekvenssi* $\frac{F_i}{n}$
L_1	f_1	$F_1 = f_1$	$\frac{F_1}{n}$
L_2	f_2	$F_2 = f_1 + f_2$	$\frac{F_2}{n}$
. . .	. . .	. . .	. . .
L_r	f_r	$F_r = f_1 + f_2 + ... + f_r = n$	$\frac{F_r}{n} = 1$
yht.	n		

FREKVENSSITAULUKOT: ESIMERKKEJÄ

Aineisto 1 (tavaratalo): Alla olevaan frekvenssitaulukkoon on kirjattu erään tavaratalon asiakkaiden lukumäärät (frekvenssit) ja suhteelliset (%) frekvenssit osastoittain eräänä päivänä.

		f_i	f_i%
Osasto	Urheilu	320	16,0
	Elintarvike	680	34,0
	Vaatetus	450	22,5
	Työkalut	200	10,0
	Kahvila	350	17,5
	Yhteensä	**2000**	**100,0**

Aineisto 2 (lapsiluku): Erääseen tutkimukseen poimittiin 50 perhettä. Seuraavassa taulukossa esitetään perheiden lapsilukujen frekvenssit, suhteelliset frekvenssit (%) ja suhteelliset summafrekvenssit (%).

Lasten lukumäärä	f_i	f_i%	F_i%
0	6	12,0	12,0
1	11	22,0	34,0
2	18	36,0	70,0
3	9	18,0	88,0
4	5	10,0	98,0
5	1	2,0	100,0
Yht.	**50**	**100,0**	

AINEISTON LUOKITTELU

- **Numeeristen, etenkin jatkuvien, muuttujien mittauksessa saadaan tyypillisesti hyvin useita erilaisia muuttujan arvoja. Jos frekvenssijakaumaa laadittaessa jokaiselle muuttujan eri arvolle varataan oma luokka, ei jakauma täytä tehtäväänsä aineiston tiivistäjänä. On siis välttämätöntä jollakin tavalla *luokitella* muuttujan arvot.**
- **Kategoristen muuttujien tapauksessa edelliseen on harvemmin tarvetta.**

AINEISTON LUOKITTELU: ESIMERKKI

Aineisto 3 (liikennemelu):
Katujen risteyksessä suoritettiin liikenteen aiheuttaman melun mittauksia desibeleinä yhden desimaalin tarkkuudella. Taulukossa on esitetty erään luokituksen mukaiset frekvenssijakaumat.

pyöristetyt luokkarajat	**Todelliset luokkarajat**	f_i	f_i%	F_i	F_i%
50.8 - 55.2	50.75-55.25	3	6	3	6
55.3 - 59.7	55.25-59.75	15	30	18	36
59.8 - 64.2	59.75-64.25	19	38	37	74
64.3 - 68.7	64.25-68.75	10	20	47	94
68.8 - 73.2	68.75-73.25	2	4	49	98
73.3 - 77.7	73.25-77.75	1	2	50	100
Yht.		**50**	**100**		

GRAAFINEN HAVAINNOLLISTUS

- **Muuttujien tyyppi vaikuttaa kuviotyypin valintaan.**
- **Kuvion tarkoituksena on tiivistää aineiston sisältämää informaatiota.**
- **Kuvio ei saa vääristää aineiston sisältämää informaatiota.**
- **Kuvioiden määrää raportissa kannattaa aina miettiä.**

GRAAFINEN HAVAINNOLLISTUS: KUVIOTYYPPEJÄ

Pylväskuvio

- **Kategorisen tai numeerisen muuttujan absoluuttista tai suhteellista jakaumaa kuvaava kuvio.**
- **Pylväät piirretään yleensä erikseen. Tällä korostetaan muuttujan diskreettiä luonnetta.**

Piirakkakuvio

- **Kategorisen tai numeerisen muuttujan suhteellista jakaumaa kuvaava kuvio.**

Sirontakuvio

- **Kahden numeerisen muuttujat graafinen havainnollistus (vrt. X-Y-koordinaatisto).**

GRAAFINEN HAVAINNOLLISTUS: KUVIOTYYPPEJÄ

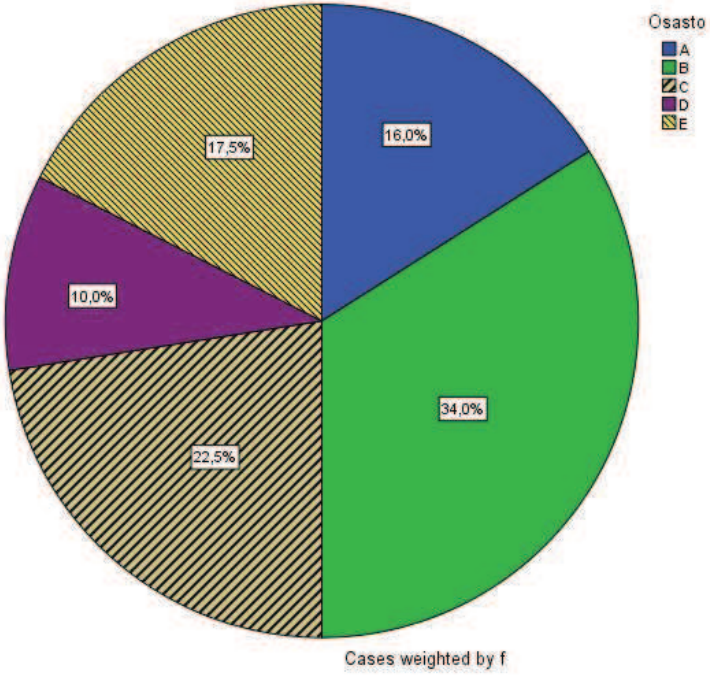
Histogrammi

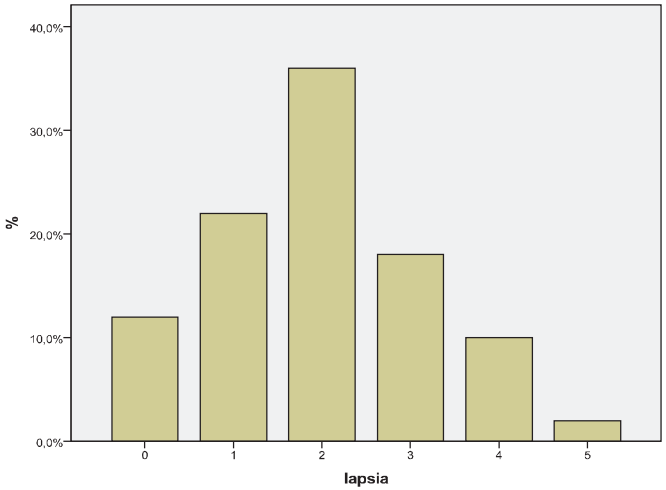
- **Luokitellun usein jatkuvan numeerisen muuttujan absoluuttista tai suhteellista jakaumaa kuvaava pylväskuvio.**
- **Erotuksena tavalliseen pylväskuvioon pylväät piirretään yleensä yhteen. Tällä korostetaan muuttujan jatkuvaa luonnetta.**

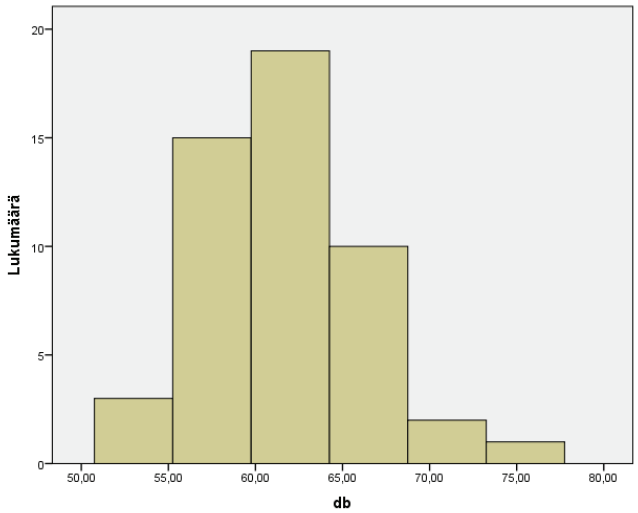
Laatikkojanakuvio

- **Erilaisten kuvailevien tunnuslukujen graafinen esitys.**

KUVIOTYYPPEJÄ: ESIMERKKEJÄ







TILASTOLLISET OHJELMISTOT

- Käytetään tilastollisten aineistojen hallintaan ja analysointiin.
- ”Perus”analyysit: Excel.
- Koululla lisenssi myös SPSS ja SAS-ohjelmistoihin.
- Ilmainen tilasto-ohjelmisto: R.

KUVAILEVAT TUNNUSLUVUT

- **Keskiluvut: Keskiarvo, moodi, mediaani, ala- ja yläkvartiili, fraktiilit.**
- **Hajontaluvut: vaihteluväli- ja sen pituus, kvartiiliväli ja sen pituus, keskihajonta, varianssi.**
- **Muut tunnusluvut: Vinous- ja huipukkuuskertoimet.**

KESKILUVUT: ARITMEETTINEN KESKIARVO (ARITHMETIC MEAN)

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

missä x_i = havainto i.

Laskenta frekvenssejä hyväksi käyttäen:

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + \ldots + f_r x_r}{f_1 + f_2 + \ldots + f_r} = \frac{\sum_{i=1}^{r} f_i x_i}{n}$$

KESKILUVUT: ARITMEETTINEN KESKIARVO: OMINAISUUKSIA

- Vähintään välimatka-asteikollisen muuttujan keskiluku
- Tulkinta: keskimääräinen havainto eli ”painopiste”
- Soveltuu parhaiten yksihuippuisille ja symmetrisille jakaumille.
- Herkkä poikkeaville havainnoille
- Mahdollinen ratkaisu: ns. leikattu keskiarvo
- Muita keskiarvoja: Geometrinen ja harmoninen keskiarvo.

KESKILUVUT: ARITMEETTINEN KESKIARVO: OMINAISUUKSIA

- **Muuttujan havaintoarvojen keskiarvosta laskettujen poikkeamien summa on nolla, eli**

$$\sum_{i=1}^{n}(X_i - \bar{X}) = 0$$

- **Jos muuttuja Y on saatu muuttujasta X lineaarisella muunnoksella $Y = aX + b$ ja muuttujan X keskiarvo on $\bar{X}$, on muuttujan Y keskiarvo**

$$\bar{Y} = a\bar{X} + b$$

KESKILUVUT: MOODI (MODE)

- **Kaikille mitta-asteikoille soveltuva keskiluku.**
- **Tulkinta: Yleisin havainto**
- **Ei välttämättä yksikäsitteinen**
- **Luettavissa frekvenssitaulukosta sen muuttujan arvon kohdalta, jonka frekvenssi on suurin.**

KESKILUVUT: MEDIAANI (MEDIAN)

- Vähintään järjestysasteikolliselle muuttujalle soveltuva keskiluku.
- Mediaani on havaintoarvo, joka jakaa aineiston kahteen osaan, siten että aineiston arvoista korkeintaan puolet on pienempiä ja korkeintaan puolet suurempia kuin mediaani.
- Tulkinta: keskimmäinen havainto.
- Ei välttämättä yksikäsitteinen.
- Jos muuttuja numeerinen ja mediaani ei yksikäsitteinen, voidaan käyttää mediaanina kahden muuttujan arvon keskiarvoa.

KESKILUVUT: FRAKTIILIT (QUANTILES)

Fraktiilit ovat vähintään järjestysasteikollisille muuttujille soveltuvia keskilukuja, jotka rajaavat tietyn osan suuruusjärjestykseen asetetuista muuttujan arvoista. Mediaani on eräs fraktiileista, 50 %:n fraktiili Q_2. Muita kvartiileja ovat alakvartiili eli 25 %:n fraktiili Q_1 ja yläkvartiili eli 75 %:n fraktiili Q_3. Kvintiilit ovat 20 %, 40 %, 60 % ja 80 %:n ja desiilit 10 %, . . ., 90 %:n fraktiilit.

HAJONTALUVUT

Hajontalukujen tarkoitus on kuvata havaintoarvojen vaihtelua toisiinsa nähden. Hajontaluvun voidaan ajatella ilmaisevan, kuinka tiheästi tai harvasti havaintoarvot ovat jakautuneet.

HAJONTALUVUT: VAIHTELUVÄLI JA SEN PITUUS (RANGE)

- **Vaihteluväli on vähintään järjestysasteikolliselle muuttujalle soveltuva hajontaluku.**
- $W = (x_{min}, x_{max})$
- **Vaihteluvälin pituus on vähintään välimatka-asteikolliselle muuttujalle soveltuva hajontaluku.**
- $R = x_{max} - x_{min}$

HAJONTALUVUT: KVARTIILILIVÄLI JA SEN PITUUS (INTERQUARTILE RANGE)

- **Kvartiiliväli Q on vähintään järjestysasteikolliselle muuttujalle soveltuva hajontaluku.**

  $Q = (Q_1, Q_3)$

- **Kvartiilivälin pituus QR on vähintään välimatka-asteikolliselle muuttujalle soveltuva hajontaluku.**

  $QR = Q_3 - Q_1$

HAJONTALUVUT: KESKIHAJONTA JA VARIANSSI (STANDARD DEVIATION, VARIANCE)

- Keskihajonta ja varianssi ovat vähintään välimatka-asteikollisen muuttujan hajontalukuja.
- Perusjoukon varianssi: $\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}$
- Perusjoukon keskihajonta: $\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}}$
- Otoksen varianssi: $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
- Otoksen keskihajonta: $s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$
- Yllä olevat kaavat voidaan sieventää helpommin laskettavaan muotoon, esim. otosvarianssi voidaan laskea kaavalla $s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}$

HAJONTALUVUT: KESKIHAJONTA JA VARIANSSI

- **Keskihajonnan ja varianssin ominaisuuksia: Jos muuttuja Y on saatu muuttujasta X lineaarisella muunnoksella $Y = aX + b$, voidaan muuttujan Y keskihajonta ja varianssi laskea muuttujan X keskihajonnan ja varianssin avulla seuraavasti:**
- **Y:n keskihajonta** $s_Y = |a| s_X$
- **Y:n varianssi** $S_Y^2 = a^2 S_X^2$ **.**

VARIAATIOKERROIN (VARIATION COEFFICIENT)

Variaatiokerroin voidaan määritellä keskiarvon ja keskihajonnan avulla:

$$V = \frac{s}{\bar{x}}$$

joka on pelkästään suhdeasteikolle soveltuva hajontaluku. Variaatiokerroin ei riipu käytetystä mittayksiköstä, joten sen avulla voidaan vertailla keskenään muuttujia, joiden mittayksiköt tai suuruusluokka eroavat toisistaan. Usein variaatiokerroin ilmaistaan prosenttilukuna, ts. muodossa

$$V\% = 100 \cdot \frac{s}{\bar{x}}\%$$

MUUTTUJAN ARVOJEN STANDARDOINTI

- Standardoinnin avulla voidaan verrata toisiinsa muun muassa yksittäisiä havaintoja. Käytännössä standardoitujen arvojen avulla voidaan mm. etsiä poikkeavia havaintoja.
- Edellisessä yhteydessä se lasketaan havainnolle i muodossa:

$$z_i = \frac{x_i - \bar{x}}{s}$$

- Jossa siis standardoitu havainto saadaan vähentämällä alkuperäisestä havainnosta muuttujan keskiarvo ja jakamalla tulos muuttujan keskihajonnalla.
- Standardoitu arvo z_i ilmaisee havaintoarvon x_i ja aineiston keskiarvon välisen etäisyyden keskihajontaa yksikkönä käyttäen.
- Standardointi voidaan suorittaa myös muulla tavoin kuin keskiarvoon ja keskihajontaan perustuen, esimerkiksi kvartiileihin ja kvartiilivälin perustuen.

MUUT TUNNUSLUVUT: VINOUSKERTOIMET (SKEWNESS)

- Jakauman vinous kuvaa jakauman muodon poikkeamista symmetrisestä.
- Positiivisesti vino jakauma: Muuttujan jakauma on ”venynyt” suuriin arvoihin päin.
- Kun jakauman muoto on edellisen peilikuva, on kyseessä negatiivisesti vino jakauma.

MUUT TUNNUSLUVUT: VINOUSKERTOIMET

- Synonyymeinä edellisten kanssa käytetään myös vasemmalle tai oikealle vinon jakauman käsitteitä.
- Kertoimia on useita erilaisia, mutta ne tulkitaan samoin. Ne ovat kaikki skaalattu siten, että negatiiviset arvot liittyvät negatiiviseen vinouteen ja positiiviset positiiviseen. Jos jakauma on symmetrinen, saavat kertoimet arvoja lähellä nollaa.

MUUT TUNNUSLUVUT: VINOUSKERTOIMET

- **Moodiin perustuva kerroin:**

$$\frac{\bar{x} - Mo}{s}$$

- **Mediaaniin perustuva kerroin:**

$$\frac{3(\bar{x} - Md)}{s}$$

- **Vinouskerroin** g_1**:**

$$g_1 = \frac{m_3}{s^3}$$

jossa

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

- **Edellisissä kertoimissa tarvitaan siis välituloksina tapauksesta riippuen moodia, keskiarvoa, mediaania ja keskihajontaa.**

MUUT TUNNUSLUVUT: HUIPUKKUUSKERTOIMET (KURTOSIS)

- Jakauman huipukkuus kuvaa jakauman muodon poikkeamista ns. normaalijakaumasta.
- Kun jakauma on muodoltaan terävähuippuinen ns. normaalijakaumaan verrattuna ts. havainnot ovat keskittyneet pienelle alueelle, puhutaan positiivisesta huipukkuudesta.
- Kun jakauma on muodoltaan litteä ns. normaalijakaumaan verrattuna ts. havainnot ovat keskittyneet suurelle alueelle, puhutaan negatiivisesta huipukkuudesta.

MUUT TUNNUSLUVUT: HUIPUKKUUSKERTOIMET

- **Huipukkuuskertoimet ovat kaikki skaalattu siten, että negatiiviset arvot liittyvät negatiiviseen huipukkuuteen ja positiiviset positiiviseen. Jos jakauma on huipukkuudeltaan lähellä normaalijakaumaa, saavat kertoimet arvoja lähellä nollaa.**
- **Huipukkuuskerroin** g_2**:** $g_2 = \frac{m_4}{s^4} - 3$

jossa $m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$

YHTEENVETO YKSIULOTTEIDEN JAKAUMAN TUNNUSLUVUISTA

Mitta-asteikko	Keskiluvut	Hajontaluvut	Muut tunnusluvut
Luokittelu-asteikko	Moodi		
Järjestys-asteikko	Moodi Fraktiilit	Vaihteluväli Kvartiiliväli	
Välimatka-asteikko	Moodi Aritmeettinen keskiarvo Fraktiilit	Vaihteluväli Vaihteluvälin pituus Kvartiiliväli Kvartiilivälin pituus Keskihajonta Varianssi	Vinousmitat Huipukkuuskerroin
Suhdeasteikko	Moodi Keskiarvot (aritmeetinen, geometrinen, harmoninen) Fraktiilit	Vaihteluväli Vaihteluvälin pituus Kvartiiliväli Kvartiilivälin pituus Keskihajonta Varianssi Variaatiokerroin	Vinousmitat Huipukkuuskerroin

KAKSIULOTTEINEN EMPIIRINEN JAKAUMA

- Kaksiulotteinen jakauma on keino tarkastella kahta muuttujaa samanaikaisesti.
- Kaksiulotteisessa jakaumassa aineisto muodostuu havaintopareista, jotka ilmoittavat tutkimuksen kohteena olevien ominaisuuksien mittaustulokset kullakin tilastoyksiköllä.
- Kuten yksiulotteisessa jakaumassa mittausten tuottama informaatio voidaan tiivistää ja havainnollistaa taulukoiden, graafisten esitysten ja tunnuslukujen avulla.

KAKSIULOTTEISEN JAKAUMAN ESITTÄMINEN: FREKVENSSITAULUKOT (FREQUENCY TABLES)

Merkitään tarkasteltavia muuttujia X:llä ja Y:llä, muuttujan X arvoja $L_1, L_2, \ldots, L_r$ ja muuttujan Y arvoja $E_1, E_2, \ldots, E_s$. Silloin muuttujien X ja Y kaksiulotteinen frekvenssijakauma on taulukko

	Muuttuja Y				
Muuttuja X	E_1	E_2	..	E_s	*Rivisumma*
L_1	f_{11}	f_{12}	..	f_{1s}	$f_{1.}$
L_2	f_{21}	f_{22}	..	f_{2s}	$f_{2.}$
..	..	..	..	..	..
L_r	f_{r1}	f_{r2}	..	f_{rs}	$f_{r.}$
Sarakesumma	$f_{.1}$	$f_{.2}$	..	$f_{.s}$	n

KAKSIULOTTEISEN JAKAUMAN ESITTÄMINEN: FREKVENSSITAULUKOT

- Arvojen L_i ja E_j muodostama paria (L_i, E_j) sanotaan soluksi.
- Solun (L_i, E_j) solufrekvenssi f_{ij} on niiden havaintoparien lukumäärä, joissa muuttujan X arvo on L_i ja muuttujan Y arvo E_j.
- Taulukkoa voidaan sanoa myös kontingenssitauluksi tai ristiintaulukoksi ja taulukon muodostamista sanotaan ristiintaulukoinniksi. Taulukot voidaan laskea myös suhteellisina tai prosenttisina.

RISTIINTAULUKOINNIT: ESIMERKKEJÄ

Y=Pääaineeni tarjoaa hyvän työllistymismahdollisuuden heti valmistumisen jälkeen						
X=Pääaine	*Täysin eri mieltä*	*Jonkin verran eri mieltä*	*Ei eri eikä samaa mieltä*	*Jonkin verran samaa mieltä*	*Täysin samaa mieltä*	*Rivisumma*
MA	*12*	*13*	*33*	*42*	*55*	*155*
JO	*9*	*33*	*25*	*35*	*23*	*125*
LT	*22*	*23*	*24*	*19*	*17*	*105*
TJT	*14*	*9*	*5*	*9*	*9*	*46*
Sarakesumma	*57*	*78*	*87*	*105*	*104*	*n=431*

EHDOLLINEN FREKVENSSIJAKAUMA (CONDITIONAL FREQUENCY TABLE)

- **Muuttujan X ehdollinen frekvenssijakauma muuttujan Y suhteen saadaan, kun aineistosta otetaan tarkasteltavaksi vain ne havaintoparit, joissa muuttujan Y arvo on tietty.**
- **Muuttujan X ehdolliset jakaumat voidaan lukea frekvenssitaulun kultakin sarakkeelta: esimerkiksi luvut $f_{11}, f_{21}, \ldots, f_{r1}$ muodostavat muuttujan X ehdollisen jakauman muuttujan Y arvolla E_1.**
- **Vastaavasti saadaan muuttujan Y ehdolliset jakaumat muuttujan X suhteen ja ne voidaan lukea frekvenssitaulun riveiltä. Esimerkiksi luvut $f_{21}, f_{22}, \ldots, f_{2s}$ muodostavat muuttujan Y ehdollisen jakauman muuttujan X arvolla L_2.**

EHDOLLINEN FREKVENSSIJAKAUMA

- Ehdolliset jakaumat voidaan laskea myös suhteellisina tai prosenttisina.
- Tällöin rivin tai sarakkeen summa on 100%. Edellinen riippuu siitä, lasketaanko muuttujan X vai Y ehdolliset jakaumat.
- Nämä jakaumat ovat tärkeitä muuttujien välistä tilastollista riippuvuutta arvioitaessa.
- Edelleen myös ehdollisia jakaumia voidaan havainnollistaa graafisesti tai tunnuslukujen avulla.

TILASTOLLINEN RIIPPUVUUS (STATISTICAL DEPENDENCY)

- **Muuttujat ovat toisistaan täydellisesti riippuvia, jos tunnettaessa toisen muuttujan arvo voidaan varmuudella sanoa, mikä toisenkin arvo on.**
- **Muuttujat ovat toisistaan tilastollisesti riippumattomia, jos toisen muuttujan arvon tunteminen ei anna mitään informaatiota toisen muuttujan arvosta.**
- **Ehdollisia frekvenssijakaumia tarvitaan, kun tutkitaan muuttujien välisiä riippuvuussuhteita. Käytännössä tällöin verrataan ehdollisten jakaumien suhteellisia (prosenttisia) frekvenssejä.**
- **Frekvenssitaulusta riippumattomuus ilmenee siten, että toisen muuttujan suhteelliset (prosenttiset) ehdolliset frekvenssijakaumat toisen muuttujan kaikilla arvoilla ovat samat.**

TILASTOLLINEN RIIPPUVUUS

- **Riippuvuuden tarkastelu voidaan suorittaa joko rivi- tai sarakeprosenttien avulla riippuen siitä kumpien tulkinta on helpompaa.**
- **Käytännössä muuttujien välinen riippuvuus tai riippumattomuus on harvoin täydellistä.**

TILASTOLLINEN RIIPPUVUUS

Pääaineeni tarjoaa hyvän työllistymismahdollisuuden heti valmistumisen jälkeen						
Pääaine	*Täysin eri mieltä*	*Jonkin verran eri mieltä*	*Ei eri eikä samaa mieltä*	*Jonkin verran samaa mieltä*	*Täysin samaa mieltä*	*Rivisumma*
MA	*7,7 %*	*8,4 %*	*21,3 %*	*27,1 %*	*35,5 %*	*100 %*
JO	*7,2 %*	*26,4 %*	*20,0 %*	*28,0 %*	*18,4 %*	*100 %*
LT	*21,0 %*	*21,9 %*	*22,9 %*	*18,1 %*	*16,1 %*	*100 %*
TJT	*30,4 %*	*19,6 %*	*10,9 %*	*19,5 %*	*19,6 %*	*100 %*

TILASTOLLINEN RIIPPUVUUS

- **Yleensä ollaankin kiinnostuneita siitä, miten voimakasta riippuvuus on, ja jos muuttujat ovat vähintään järjestysasteikollisia, mikä on riippuvuuden suunta.**
- **Jos tilastollisesti riippuvilla muuttujilla X ja Y pieniin X:n arvoihin liittyy pieniä Y:n arvoja, puhutaan positiivisesta riippuvuudesta.**
- **Jos taas pieniin X:n arvoihin on taipumus liittyä suuria Y:n arvoja, on riippuvuus negatiivista.**
- **Huomattavaa on, että tilastollinen riippuvuus ei kerro välttämättä mitään muuttujien välisestä syy-seuraus –suhteesta.**

RIIPPUVUUSTUNNUSLUVUT

- Riippuvuustunnusluvut kuvaavat muuttujien välisen riippuvuuden voimakkuutta ja usein myös sen suuntaa.
- Johtuen tunnuslukujen erilaisista luonteista tulkitaan ne myös eri tavoin.
- Kategoristen muuttujien riippuvuus, ei suuntaa: Kontingenssikerroin.
- Järjestysten riippuvuus ja suunta: Spearmanin järjestyskorrelaatiokerroin.
- Lineaarinen riippuvuus ja suunta: Pearsonin korrelaatiokerroin.

RIIPPUVUUSTUNNUSLUVUT: KONTINGENSSIKERROIN (CONTINGENCY COEFFICIENT)

- **Kontingenssikerroin on kaikille mitta-asteikoille soveltuva riippuvuustunnusluku, joka perustuu havaintoaineiston frekvenssien ja teoreettisten, täydellistä riippumattomuutta edustavien frekvenssien vertailuun.**
- **Riippumattomuustilannetta vastaavien frekvenssien laskemisessa lähtökohdaksi otetaan havainnoista laaditun frekvenssitaulun reunajakaumat.**
- **Mitä enemmän nämä havaitut ja odotetut frekvenssit poikkeavat toisistaan, sitä voimakkaampi on muuttujien välinen riippuvuus.**

RIIPPUVUUSTUNNUSLUVUT: KONTINGENSSIKERROIN

- **Teoreettisessa riippumattomuustilanteessa toisen muuttujan suhteellisten ehdollisten jakaumien tulee olla samat kuin vastaava suhteellinen reunajakauma.**
- **Täten solun ij odotettu frekvenssi on luku, joka sarakkeen j sarakesummalla (tai rivin i rivisummalla) jaettuna on sama kuin rivin i suhteellinen rivisumma (tai sarakkeen j suhteellinen sarakesumma), riippuen siitä kumman muuttujan ehdollisia jakaumia tarkastellaan.**

RIIPPUVUUSTUNNUSLUVUT: KONTINGENSSIKERROIN

- Edellisestä seuraa siis $\frac{e_{ij}}{f_{.j}} = \frac{f_{i.}}{n}$.
- Solun ij odotettu frekvenssi e_{ij} voidaan siis laskea kaavalla

$$e_{ij} = \frac{f_{i.} f_{.j}}{n}$$

- Odotettujen frekvenssien e_{ij} ja havaittujen frekvenssien f_{ij} erotuksia $f_{ij} - e_{ij}$ sanotaan jäännöksiksi.
- Eri solujen jäännökset saadaan paremmin keskenään vertailukelpoisiksi käyttämällä ns. standardoituja jäännöksiä $\frac{f_{ij} - e_{ij}}{\sqrt{e_{ij}}}$.

RIIPPUVUUSTUNNUSLUVUT: KONTINGENSSIKERROIN

- Jäännösten avulla voidaan tutkia, miten havaitut frekvenssit poikkeavat täydellisen riippumattomuuden ääritapauksesta. Eräs standardoitujen jäännösten tarkasteluun perustuva suure on χ^2, joka määritellään kaavalla

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

- Jos muuttujat ovat täysin riippumattomia, jäännökset ovat nollia ja χ^2 saa arvon nolla.

RIIPPUVUUSTUNNUSLUVUT: KONTINGENSSIKERROIN

- **Luku χ^2 on sitä suurempi, mitä enemmän havaitut frekvenssit poikkeavat odotetuista, täydellistä riippumattomuustilannetta vastaavista frekvensseistä.**
- **χ^2 -suureen käytössä on hankaluutena, että sen suuruus riippuu havaintojen määrästä ja taulukon koosta: mitä enemmän havaintoja ja/tai taulukossa rivejä ja sarakkeita, sitä suurempi on χ^2 -arvo. Näin eri aineistoja ei voi verrata toisiinsa χ^2 -arvon perusteella.**

RIIPPUVUUSTUNNUSLUVUT: KONTINGENSSIKERROIN

- Siksi suuretta ei tavallisesti käytetäkään sellaisenaan, vaan se normeerataan ensin jollakin tavalla. Eräs tapa normeerata χ^2 on seuraava:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

- Lukua C sanotaan kontingessikertoimeksi.
- Kontingenssikerroin siis normeeraa χ^2 -arvot nollan ja ykkösen välille, ts. on voimassa

$$0 \leq C < 1$$

- Kertoimen C suuruus kuvaa edelleen vain riippuvuuden voimakkuutta: mitä voimakkaampi riippuvuus muuttujien välillä on, sitä suurempi on C:n arvo.

RIIPPUVUUSTUNNUSLUVUT: KONTINGENSSIKERROIN

- **Normeeraus, jolla kerroin C saadaan, ei kuitenkaan täysin poista tunnusluvun sidonnaisuutta aineistoon.**
- **Voidaan osoittaa, että kontingenssikertoimen laskennallinen maksimiarvo C_{max} riippuu tarkasteltavan frekvenssitaulun koosta, eli tutkittavan aineiston muuttujien luokkien lukumäärästä, siten että**

$$C_{\max} = \sqrt{\frac{q-1}{q}}$$

missä q on taulukon sarakkeiden ja rivien lukumääristä pienempi, ts. $q = min(r,s)$, missä r = taulukon rivien lukumäärä ja s = taulukon sarakkeiden lukumäärä.

RIIPPUVUUSTUNNUSLUVUT: KONTINGENSSIKERROIN

- **Jos aineistosta laskettu kontingenssikerroin C suhteutetaan aineistokohtaiseen maksimiarvoon C_{max} , saadaan luku C/C_{max} , jonka suhteen eri aineistot ovat vertailukelpoisia. Luvun C/C_{max} arvo on aina nollan ja ykkösen välillä, ts.**

$$0 \leq \frac{C}{C_{\max}} \leq 1$$

- **Jos $C/C_{max} = 0$, on $C = 0$, eli muuttujat ovat keskenään täydellisesti riippumattomat. Jos taas $C/C_{max} = 1$, on $C = C_{max}$, eli riippuvuus muuttujien välillä on suurin mahdollinen.**

RIIPPUVUUSTUNNUSLUVUT: KONTINGENSSIKERROIN: TULKINTA

$\frac{C}{C_{\max}} \leq 0.2$ **Ei merkittävää riippuvuutta muuttujien välillä**

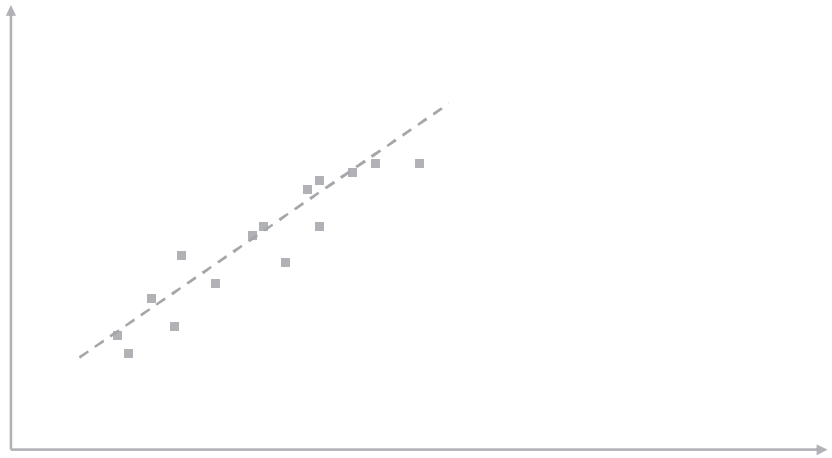
$0.2 \leq \frac{C}{C_{\max}} < 0.3$ **Rajatapaus**

$0.3 \leq \frac{C}{C_{\max}}$ **Riippuvuutta muuttujien välillä, tarkempi tulkinta esim. riviprosenttien avulla.**

Huom! Rajat ohjeellisia.

RIIPPUVUUSTUNNUSLUVUT: PEARSONIN KORRELAATIOKERROIN (CORRELATION COEFFICIENT)

- Kahden numeerisen muuttujan lineaarista riippuvuutta kuvaava tunnusluku.
- Lineaarinen riippuvuus: Eräs kahden numeerisen muuttujan riippuvuuden laji (ks. kuvio).
- Kerroin kuvaa sekä riippuvuuden voimakkuutta että suuntaa.



RIIPPUVUUSTUNNUSLUVUT: PEARSONIN KORRELAATIOKERROIN: KOVARIANSSI (COVARIANCE)

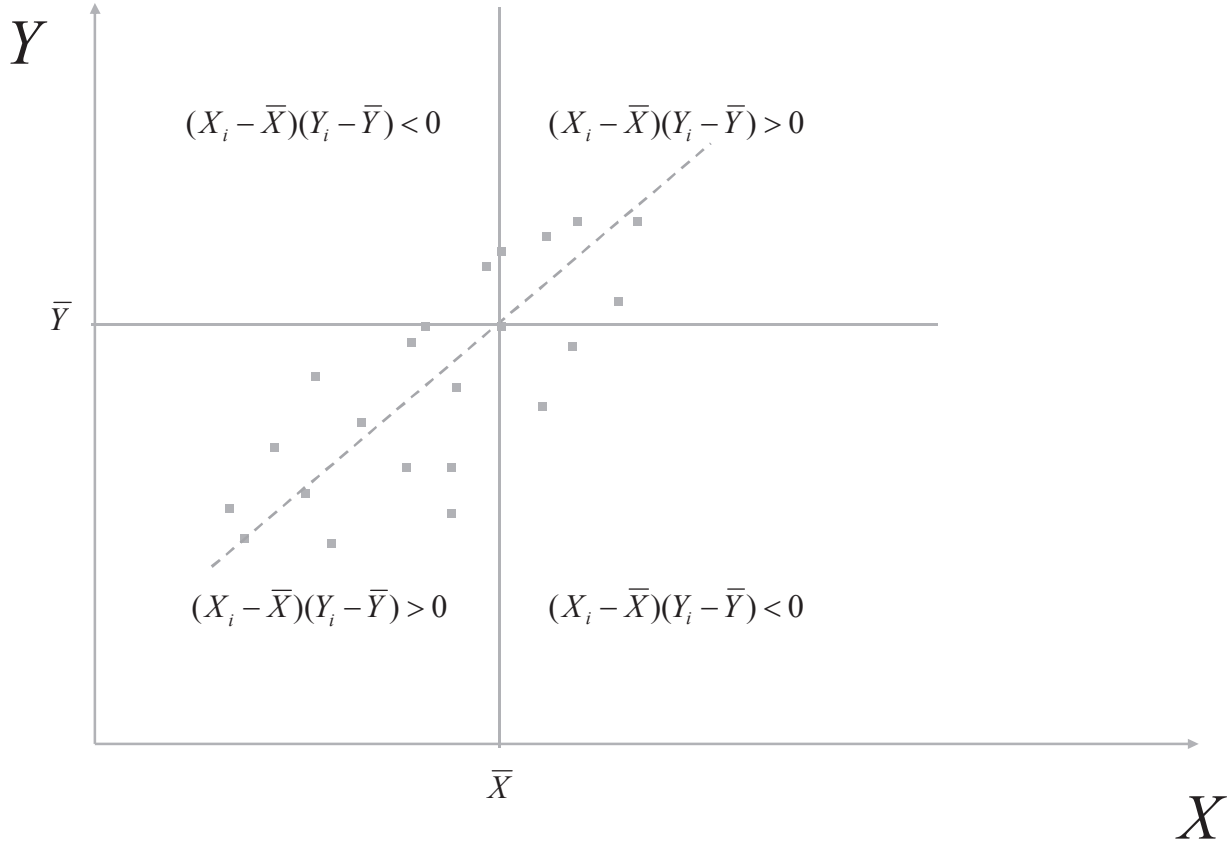
- **Korrelaatiokerroin perustuu muuttujien X ja Y arvojen yhteistä vaihtelua kuvaavaan kovarianssiin, joka puolestaan on yksiulotteisten jakaumien varianssia vastaava suure.**
- **Kun muuttujien X ja Y arvoja merkitään x_i ja y_i, muodostuu tarkasteltava n:n havainnon havaintoaineisto lukupareista (x_i, y_i), jossa indeksi i saa arvot 1:stä n:ään. Muuttujien X ja Y välinen otoskovarianssi määritellään**

$$s_{XY} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{n-1}$$

RIIPPUVUUSTUNNUSLUVUT: PEARSONIN KORRELAATIOKERROIN: KOVARIANSSI

- Kovarianssi kuvaa havaintoparien hajontaa eli havaintoparien poikkeamista kummankin muuttujan keskiarvoa edustavasta pisteestä $(\bar{x}, \bar{y})$. Kovarianssi voi saada myös negatiivisen arvon, jos muuttujien arvot vaihtelevat enimmäkseen erisuuntaisesti: pientä X:n arvoa vastaa suuri Y:n arvo, eli poikkeaman $(x_i - \bar{x})$ ollessa negatiivinen $(y_i - \bar{y})$ on positiivinen, tai päinvastoin.
- Kovarianssin ongelma tunnuslukuna se, että sen suuruus riippuu muuttujien arvojen suuruudesta. Näin muuttujaparien vertailu on vaikeaa.
- Kovarianssi itsensä kanssa = varianssi.

RIIPPUVUUSTUNNUSLUVUT: PEARSONIN KORRELAATIOKERROIN: KOVARIANSSI



RIIPPUVUUSTUNNUSLUVUT: PEARSONIN KORRELAATIOKERROIN

Pearsonin tulomomenttikorrelaatiokerroin, eli lyhyemmin korrelaatiokerroin r_{XY}, jossa kovarianssi jaetaan maksimiarvollaan:

$$r_{XY} = \frac{s_{XY}}{s_X \cdot s_Y} = \frac{\frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{n-1}}{\sqrt{\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}\right) \cdot \left(\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}\right)}}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

RIIPPUVUUSTUNNUSLUVUT: PEARSONIN KORRELAATIOKERROIN

- Käytännön laskuissa on useimmiten helpointa käyttää kaavaa muodossa

$$r_{XY} = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}$$

- Pearsonin korrelaatiokerroin on aina itseisarvoltaan korkeintaan yksi, eli

$$-1 \le r_{XY} \le 1$$

- On herkkä poikkeaville havainnoille pienissä aineistoissa.
- Pearsonin korrelaatiokerroin on symmetrinen tunnusluku.

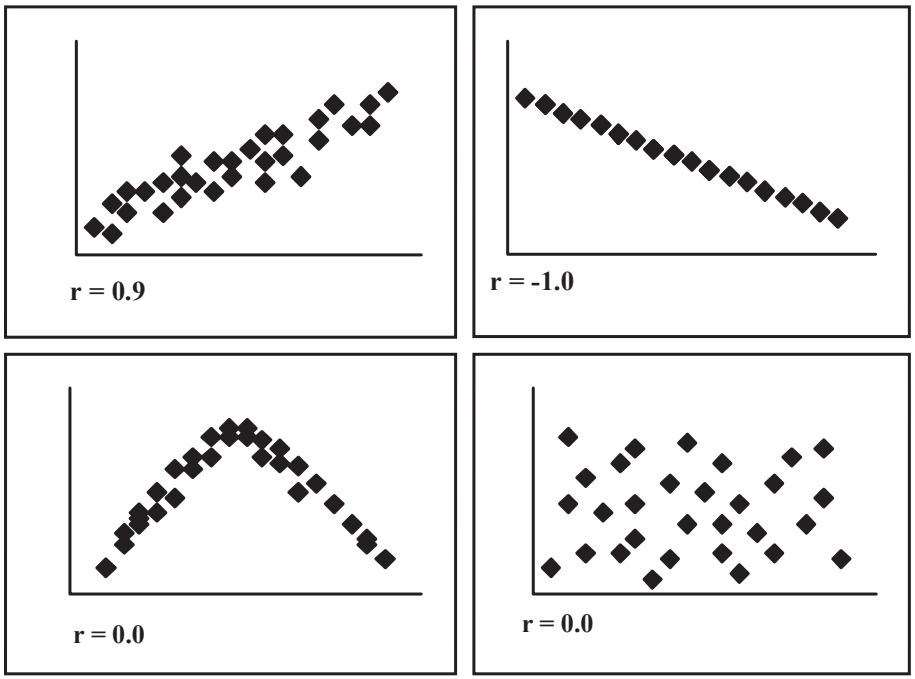
RIIPPUVUUSTUNNUSLUVUT: PEARSONIN KORRELAATIOKERROIN: TULKINTA

$\lvert r_{XY} \rvert \leq 0.3$	**Heikkoa positiivista/negatiivista lineaarista riippuvuutta muuttujien X ja Y arvojen välillä.**
$0.3 < \lvert r_{XY} \rvert < 0.7$	**Kohtalaista positiivista/negatiivista lineaarista riippuvuutta muuttujien X ja Y arvojen välillä.**
$\lvert r_{XY} \rvert \geq 0.7$	**Voimakasta positiivista/negatiivista lineaarista riippuvuutta muuttujien X ja Y arvojen välillä.**

Huom! Rajat ohjeellisia.

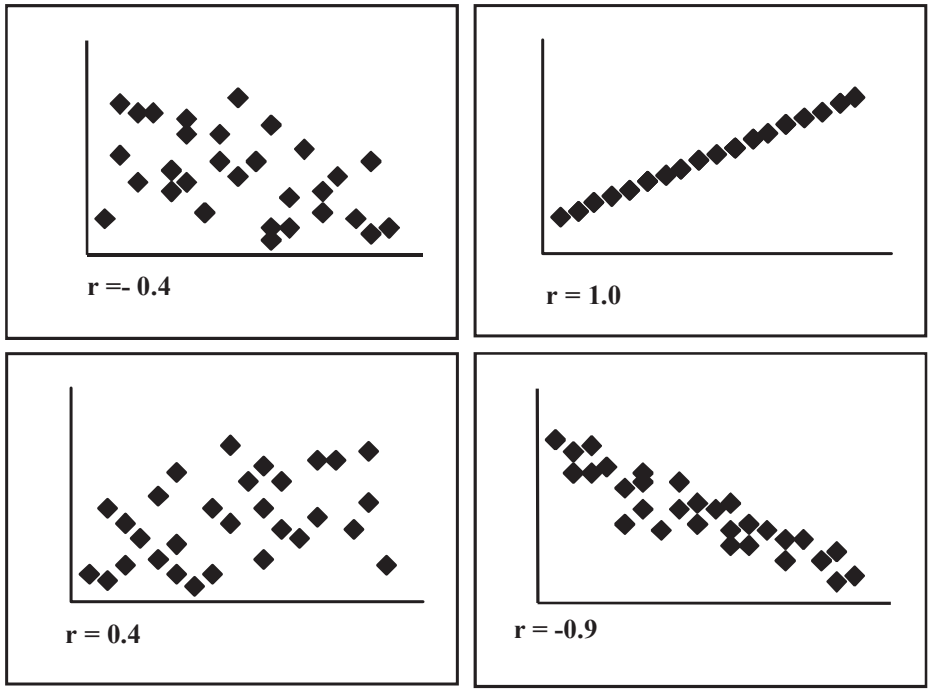
RIIPPUVUUSTUNNUSLUVUT: PEARSONIN KORRELAATIOKERROIN

Kahden muuttujan havaintoaineistoa voi tarkastella graafisesti sirontakuvion avulla:



RIIPPUVUUSTUNNUSLUVUT: PEARSONIN KORRELAATIOKERROIN

Lisää esimerkkejä:



RIIPPUVUUSTUNNUSLUVUT: SPEARMANIN JÄRJESTYSKORRELAATIOKERROIN

- Spearmanin järjestyskorrelaatiokerroin lasketaan kaavalla

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

jossa d on havaintoihin liittyvä järjestyslukujen erotus ja n havaintojen lukumäärä.

- Kerroin kuvaa kahden muuttujan järjestysten riippuvuutta.
- Tunnusluku soveltuu vähintään järjestysasteikollisille muuttujille ja on symmetrinen tunnusluku.
- Ei niin herkkä poikkeaville havainnoille kuin Pearsonin kerroin: Kerroin ”unohtaa” muuttujien arvojen väliset etäisyydet.

RIIPPUVUUSTUNNUSLUVUT: SPEARMANIN JÄRJESTYSKORRELAATIOKERROIN

- Laskettaessa kerrointa kummankin muuttujan arvolle annetaan järjestysluku, jotka saavat arvoja $1, \ldots, n$. Korrelaatiokertoimessa oleva summa $\sum_{i=1}^{n} d_i^2$ on näiden järjestyslukujen erotuksien summa.
- Sillä, annetaanko suurimmalle vai pienimmälle muuttujan arvolle järjestysluku 1, ei ole merkitystä. Olennaista on, että numerointi on tehty samalla periaatteella kummallekin muuttujalle.
- Yhtä suuret muuttujan arvot saavat järjestysluvuikseen kahden niihin liittyvän järjestysluvun keskiarvon.
- Spearmanin korrelaatiokerroin on aina itseisarvoltaan korkeintaan yksi, eli $-1 \leq r_S \leq 1$.

RIIPPUVUUSTUNNUSLUVUT: SPEARMANIN KORRELAATIOKERROIN: TULKINTA

$|r_s| \leq 0.3$ **Heikkoa positiivista/negatiivista arvojen järjestysten riippuvuutta muuttujien X ja Y välillä.**

$0.3 < |r_s| < 0.7$ **Kohtalaista positiivista/negatiivista arvojen järjestysten riippuvuutta muuttujien X ja Y välillä.**

$|r_s| \geq 0.7$ **Voimakasta positiivista/negatiivista arvojen järjestysten riippuvuutta muuttujien X ja Y välillä.**

Huom! Rajat ohjeellisia.

KORRELAATIOMATRIISI

Jos tilastoyksiköistä on mitattu esimerkiksi kolmea eri ominaisuutta, muodostuu havaintoaineisto näitä ominaisuuksia edustavien muuttujien X, Y ja Z arvoista eli kolmikoista (x_i, y_i, z_i). Korrelaatiokerroin voidaan laskea jokaisen muuttujaparin välille erikseen, jolloin saadaan korrelaatiomatriisi

$$\begin{array}{c} \\ X \\ Y \\ Z \end{array} \begin{array}{c} \begin{array}{ccc} X & Y & Z \end{array} \\ \begin{pmatrix} r_{XX} & r_{XY} & r_{XZ} \\ r_{YX} & r_{YY} & r_{YZ} \\ r_{ZX} & r_{ZY} & r_{ZZ} \end{pmatrix} \end{array} \quad \textit{tai} \quad \begin{array}{c} \\ X \\ Y \\ Z \end{array} \begin{array}{c} \begin{array}{ccc} X & Y & Z \end{array} \\ \begin{pmatrix} 1 & r_{XY} & r_{XZ} \\ & 1 & r_{YZ} \\ & & 1 \end{pmatrix} \end{array}$$

OSITTAISKORRELAATIOKERROIN

- **Kahden muuttujan (X ja Y) lineaarinen tai järjestysten riippuvuus, kun kolmannen muuttujan (Z) vaikutus on otettu huomioon.**

$$r_{XY.Z} = \frac{r_{XY} - r_{XZ} r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}}$$

REGRESSIOANALYYSI: PERUSPERIAATE (REGRESSION ANALYSIS)

- Regressio esittää riippuvuuden epäsymmetrisestä näkökulmasta. Ei puhuta keskenään riippuvista muuttujista, vaan katsotaan, että toinen muuttuja riippuu toisesta.
- Regressiossa muuttujat jaetaankin selittäviin ja selitettäviin muuttujiin.
- Selittäviä muuttujia sanotaan myös riippumattomiksi ja selitettäviä riippuviksi muuttujiksi.
- Se että muuttujan Y katsotaan riippuvan muuttujasta X, ei merkitse kannan ottamista muuttujien X ja Y välisiin mahdollisiin syy-seuraussuhteisiin eli kausaalisuhteisiin.

REGRESSIOANALYYSI: PERUSPERIAATE

- Regressioyhtälö kuvaa muuttujien välisen riippuvuuden funktionaalisena yhteytenä, ts. se esittää riippuvan muuttujan riippumattomien muuttujien funktiona.
- Aineiston perusteella arvioidaan funktion käyttökelpoisuutta ja sopivuutta yhteyden kuvaamiseen, ja jos on tarpeen, valitaan sopivampi funktio. Tätä sanotaan funktion sovittamiseksi aineistoon.
- Menettelyn päämääränä ei ole täsmälleen oikean funktionaalisen yhteyden määrittäminen, vaan sen sijaan on tarkoitus löytää sellainen funktio, jonka avulla voidaan luonnehtia yhteyden tärkeimpiä ja oleellisimpia piirteitä.

REGRESSIOANALYYSI: PERUSPERIAATE

- Sovitettavaa funktiota sanotaan regressiofunktioksi, tai sen graafiseen havainnollistukseen viitaten, regressiokäyräksi.
- Funktiotyyppejä ovat esimerkiksi lineaarinen, kvadraattinen (toisen asteen polynomi) tai logaritminen.
- Regressioanalyysissä yleisimmin käytetty funktio on lineaarinen. Tätä sanotaan myös regressiosuoraksi.

REGRESSIOANALYYSI: LINEAARINEN REGRESSIO (LINEAR REGRESSION)

- Tarkastellaan ensin yhden selittävän muuttujan lineaarista regressiota. Merkitään selittävää muuttujaa X:llä ja selitettävää Y:llä. Tarkoituksena on löytää sellainen lineaarinen funktio g, joka parhaiten kuvaisi aineistoa (x_i, y_i), $i=1,\ldots,n$.
- Sanotaan, että lineaarisen funktion g arvo tietyllä muuttujan X arvolla x on sovite, ja merkitään sitä $\hat{y}$:llä, ts.

$$\hat{y} = g(x)$$

- Tilastoyksiköllä i muuttujan X arvo on x_i, joten vastaava sovite on $\hat{y}_i = g(x_i)$ .Kyseisellä tilastoyksiköllä havaitun muuttujan Y arvon y_i ja sitä vastaavan sovitteen erotusta sanotaan jäännökseksi tai residuaaliksi ja merkitään e_i:llä, ts.

$$e_i = y_i - \hat{y}_i$$

- Lineaarinen funktio kirjoitetaan yleisesti muodossa $g(x) = b_0 + b_1 x$, missä b_0 ja b_1 ovat vakioita. Nyt muuttujan X arvoon x liittyvä sovite määräytyy regressiosuoran yhtälöstä $\hat{y} = b_0 + b_1 x$.

REGRESSIOANALYYSI: LINEAARINEN REGRESSIO

Vakiot eli parametrit b_0 ja b_1 määräävät sen, millainen suora on kysymyksessä. Kulmakerroin b_1 ilmoittaa suoran kaltevuuden ja noususuunnan. Mitä suurempi b_1 on itseisarvoltaan, sitä jyrkempi suora. Jos kulmakerroin on positiivinen, suora on nouseva, ja jos kulmakerroin on negatiivinen, suora on laskeva. Parametri b_0 määrää suoran sijainnin koordinaatistossa, eli kertoo, millä etäisyydellä *x*-akselista suora leikkaa *y*-akselin.

REGRESSIOANALYYSI: PIENIMMÄN NELIÖSUMMAN MENETELMÄ (LEAST SQUARES)

- **PNS -menetelmässä määritetään parametrien b_0 ja b_1 arvot siten, että jäännöksien neliöiden summa, ns. jäännösneliösumma**

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

on mahdollisimman pieni. Tehtävänä on siis minimoida funktio

$$w(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

- **Käytännössä tehtävä ratkaistaan siten, että määritetään funktion w osittaisderivaatat kummankin tuntemattoman b_0 ja b_1 suhteen ja merkitään ne nolliksi.**

REGRESSIOANALYYSI: PIENIMMÄN NELIÖSUMMAN MENETELMÄ

Parametrien b_0 ja b_1 arvot ratkaistaan edellisestä kahdesta yhtälöstä, jolloin saadaan

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$b_0 = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i\right) = \bar{y} - b_1 \bar{x}$$

REGRESSIOANALYYSI: PARAMETRIT JA NIIDEN TULKINTA

- b_0:n tulkinta: Selitettävän muuttujan arvo, kun selittäjä saa arvon 0. Tämä parametri ei olennainen.
- b_1:*n* tulkinta: Selitettävän muuttujan keskimääräinen muutos, kun selittäjä kasvaa yhden yksikön.

REGRESSIOANALYYSI: SELITYSASTE

- **Jos regressioyhtälön avulla halutaan laatia ennusteita, on oltava jotakin tietoa siitä, kuinka luotettavia ennusteet ovat, eli kuinka hyvin regressiosuora kuvaa muuttujien välistä yhteyttä. Regressiosuoran "hyvyyttä" voidaan arvioida mm. *selitysasteen* (merkitään R^2) avulla. Yhden selittäjän regressioyhtälölle selitysaste on sama kuin korrelaatiokertoimen neliö.**

$$R^2 = r_{XY}^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{SST - SSE}{SST} = \left(\frac{s_X}{s_Y} b_1\right)^2$$

- **Tulkinta: Selitysaste kuvaa, kuinka suuren osan selitettävän muuttujan vaihtelusta regressiosuoralla voidaan selittää.**

REGRESSIOANALYYSI: ENNUSTEET

- **Kulmakertoimen avulla tapahtuva ennustaminen:**

  $$\Delta y = b_1 \Delta x$$

  eli kulmakertoimen avulla voidaan laskea tiettyä x:n muutosta vastaava keskimääräinen y:n muutos.
- **Regressioyhtälön avulla voidaan laskea tiettyä x:n arvoa vastaava keskimääräinen y:n arvo, eli**

  $$\hat{y} = b_0 + b_1 x$$
- **Huomattavaa on, että ennusteita y:n arvolle kannattaa laskea vain sellaisten x:n arvojen avulla, jotka kuuluvat x:n havaittuun vaihteluväliin.**

REGRESSIOANALYYSI: MONTA SELITTÄJÄÄ

- **K kpl selittäjiä::**

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

- **Nyt kunkin kertoimen laskemisessa on otettu huomioon myös muiden selittäjien ja selitettävän yhteys.**
- **Selittäjä voi olla myös kategorinen kaksiluokkainen muuttuja. Silloin vastaava kerroin estimoi vakiotermien eron eli regressiosuorien etäisyyden kahden eri ryhmän välillä.**
- **Jos kategorisella selittäjällä on k eri arvoa, tarvitaan k-1 kaksiluokkaista muuttujaa sen esittämiseksi.**
- **Kaksiluokkaisia muuttujia kutsutaan myös dummy-muuttujiksi.**

MONILUOKKAINEN SELITTÄVÄ MUUTTUJA (K LUOKKAA)

- **Dummy-muuttujat esitetään seuraavasti:**

	X_1	X_2	X_3	..	X_{k-1}
$X_1=1$, jos $X=1$, muuten $X_1=0$	1	0	0	0	0
$X_2=1$, jos $X=2$, muuten $X_2=0$	0	1	0	0	0
..	..	..	..	..	..
$X_{k-1}=1$, jos $X=k-1$, muuten $X_{k-1}=0$	..	..	..	..	1
$X= k$	0	0	0	0	0

- **Rivistä, jossa on pelkkiä nollia, käytetään nimitystä referenssiluokka. Dummy-muuttujien kertoimet kertovat regressioyhtälöiden vakiotermien erosta referenssiluokkaan verrattuna.**

LOGISTINEN REGRESSIOANALYYSI

- Jos kahden luokittelevan muuttujan yhteys kiinnostaa, voidaan asiaa lähestyä ristiintaulukoinnin avulla. Riippuvuutta voidaan kuvailla esim. kontingenssikertoimen avulla.
- Muuttujan ollessa numeerinen, siis vähintään välimatka-asteikollinen, ei ristiintaulukointi ole järkevää.
- Myös useamman luokittelevan muuttujan välisen yhteyden tarkastelu ristiintaulukoinnilla on vaikeaa.
- Tällaisia tilanteita varten on kehitetty omia menetelmiään.

LOGISTINEN REGRESSIO

Jos 2-luokkaista muuttujaa selitetään muilla muuttujilla, on siirrytty käyttämään *logistisen regressioanalyysin* nimellä tunnettuja menetelmiä.

- **Mallin kaava (yksi selittäjä):**

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x$$

, jossa $\frac{p}{1-p}$ **viittaa ns. vedonlyöntisuhteeseen.**

RISTITULOSUHDE

- **Ristitulosuhde (eli suhteellinen vedonlyöntisuhde) on logistisen regression kannalta tärkeä käsite.**
- **Odds Ratio (OR)**
- **Ristitulo voidaan laskea vain nelikentässä.**
- **Havaintoaineisto esitetään ristiintaulukoinnin avulla seuraavasti:**

Y/X	A	B
1=Kyllä	f_{11}	f_{12}
0=Ei	f_{21}	f_{22}

RISTITULOSUHDE

- **Tapahtuman $Y=1$ vedonlyöntisuhde (odds) O, kun toisen muuttujan arvo on A, on f_{11}/f_{21} ja kun toisen muuttujan arvo on B f_{12}/f_{22}. Arvot O kuvaavat tapahtuman riskiä toisen muuttujan arvoilla verrattuna.**
- **Oddseista voidaan laskea ristitulosuhde odds ratio (OR)**

$$\frac{f_{11}/f_{21}}{f_{12}/f_{22}}$$

- **Se kuvaa siis tapahtuman riskin suhteellista eroa toisen muuttujan eri arvoilla. Ns. referenssiluokkana toimii toisen muuttujan se luokka, johon suureen nimittäjä viittaa. (Esim. edellisessä luokka B, johon A:ta verrataan).**
- **Jos ristitulosuhde saa lähellä ykköstä olevia arvoa, tukee se muuttujien riippumattomuusoletusta.**

REGRESSIOYHTÄLÖN KERTOIMET

- **Yksinkertainen logistinen malli, kun ns. refererenssiluokkana on selittäjän arvo B:**

$$b_0 = \ln(f_{12} / f_{22}) \qquad b_1 = \ln\left(\frac{f_{11}/f_{21}}{f_{12}/f_{22}}\right) = \ln(OR)$$

joten $OR = e^{b_1}$**.**

- **Tämän avulla saadaan siis selittäjän kahden ryhmän eroa kuvaava ristitulosuhde. Tässä referenssiluokkana on siis X:n luokka B. Jos referenssiluokkana olisi ensimmäinen luokka, olisi OR edellisen käänteisluku.**

KAKSILUOKKAINEN SELITETTÄVÄ JA NUMEERINEN SELITTÄVÄ MUUTTUJA

- **Nyt mielenkiinnon kohteena olevalle parametrille voidaan antaa seuraava tulkinta:**

$$b_1 = \ln(\textit{Yhtä } X\textit{:n yksikköä vastaava } OR)$$

- **Tällöin kuvaa sitä, mikä on mallin trendin mukaan ykkösen suuruista kasvua X:n arvoissa vastaava Y:n jakaumaeroa kuvaava OR.**

KAKSILUOKKAINEN SELITETTÄVÄ JA MONTA SELITTÄVÄÄ MUUTTUJAA

- Edellä esitetyt tapaukset voidaan yleistää tapauksiksi, jossa selitettävää muuttujaa selittää useampi kategorinen tai numeerinen muuttuja. Mallin kertoimet on laskettu vastaavasti, kuten “tavallisessa” regressiossa, ottaen huomioon myös muiden selittäjien ja selittävän väliset riippuvuudet.
- Kategorisen selittäjän ollessa moniluokkainen, käytetään dummy-muuttujia vastaavasti kuin ”tavallisessa” regressiossa. Mallin kertoimesta saadaan ristitulosuhteet eri luokkien ja referenssiluokan välille.
- Mallin parametreille voidaan antaa vastaava tulkinta kuin yksinkertaisemmissa tapauksissa.

TILASTOLLINEN PÄÄTTELY

- Todennäköisyyslaskennan sovellus
- Estimointi: Piste ja väliestimointi
- Tilastolliset testit
- Tavoitteena otoksen perusteella tehtävät yleistykset perusjoukkoon.
- Näihin keskitytään kurssilla TKMY3

Turun yliopisto
University of Turku

KIRJALLISUUTTA

Anderson, D.R., Sweeney D.J. , Williams T.A.: Statistics for Business and Economics, 5th ed. tai uudempi, West Publishing Company.

Nummenmaa, T., Konttinen, R., Kuusinen, J., Leskinen, E.: Tutkimusaineiston analyysi, 1997, WSOY.

TKMY2 Kuvaileva tilastotiede

Kaavakokoelma

(1) $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$

(2) $\bar{x} = \frac{f_1 x_1 + f_2 x_2 + \ldots + f_r x_r}{f_1 + f_2 + \ldots + f_r} = \frac{\sum_{i=1}^{r} f_i x_i}{n}$

(3) $W = (x_{\min}, x_{\max})$

(4) $R = x_{\max} - x_{\min}$

(5) $Q = (Q_1, Q_3)$

(6) $QR = Q_3 - Q_1$

(7) $\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{N}$

(8) $\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{N}}$

(9) $s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$

(10) $s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$

(11) $s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}$

(12) $\sigma_Y^2 = a^2 \sigma_{X1}^2 + (1-a)^2 \sigma_{X2}^2 + 2a(1-a)\sigma_{X1}\sigma_{X2}\rho_{X1X2}$

(13) $a^* = \frac{\sigma_{X2}^2 - \sigma_{X1}\sigma_{X2}\rho_{X1X2}}{\sigma_{X1}^2 + \sigma_{X2}^2 - 2\sigma_{X1}\sigma_{X2}\rho_{X1X2}}$

(14) $z_i = \frac{x_i - \bar{x}}{s}$

(15) $z_i = \frac{x_i - Q_3}{QR}$

(16) $z_i = \frac{x_i - Q_1}{QR}$

(17) $V\% = 100 \cdot \frac{s}{\bar{x}}\%$

(18) $\frac{\bar{x} - Mo}{s}$

(19) $\frac{3(\bar{x} - Md)}{s}$

(20) $g_1 = \frac{m_3}{s^3}$

(21) $m_3 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^3$

(22) $g_2 = \frac{m_4}{s^4} - 3$

(23) $$m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$$

(24) $$e_{ij} = \frac{f_{i.} f_{.j}}{n}$$

(25) $$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

(26) $$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

(27) $$C_{\max} = \sqrt{\frac{q-1}{q}}$$

(28) $$q = \min(r, s)$$

(29) $$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\cdot(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

(30) $$r_{XY} = \frac{n\sum_{i=1}^{n}x_i y_i - \left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)}{\sqrt{\left[n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2\right]\left[n\sum_{i=1}^{n}y_i^2 - \left(\sum_{i=1}^{n}y_i\right)^2\right]}}$$

(31) $$r_s = 1 - \frac{6\cdot\sum_{i=1}^{n}d_i^2}{n(n^2-1)},$$

(32) $$r_{XY.Z} = \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}}$$

(33) $$b_1 = \frac{n\sum_{i=1}^{n}x_i y_i - \left(\sum_{i=1}^{n}y_i\right)\left(\sum_{i=1}^{n}x_i\right)}{n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2}$$

(34) $$b_0 = \frac{1}{n}\left(\sum_{i=1}^{n}y_i - b_1\sum_{i=1}^{n}x_i\right) = \bar{y} - b_1\bar{x}$$

(35) $$R^2 = r_{XY}^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$
$$= \frac{SST - SSE}{SST} = \left(\frac{s_X}{s_Y}b_1\right)^2$$

(36) $$OR = \frac{f_{11}/f_{21}}{f_{12}/f_{22}}$$

(37) $$b_0 = \ln(f_{12}/f_{22})$$

(38) $$b_1 = \ln\left(\frac{f_{11}/f_{21}}{f_{12}/f_{22}}\right) = \ln(OR)$$

(39) $$OR = e^{b_1}$$

(40) $$p = \frac{1}{1 + e^{-(b_0 + b_1 x)}}$$