

---

**Ilkka Mellin**

**Tilastolliset menetelmät**

**Osa 3: Tilastolliset testit**

**Yhteensopivuuden, homogeenisuuden ja riippumattomuuden testaaminen**

# Yhteensopivuuden, homogeenisuuden ja riippumattomuuden testaaminen

---

- >> Jakaumaoletuksien testaaminen
  - Yhteensopivuuden testaaminen
  - Homogeenisuuden testaaminen
  - Riippumattomuuden testaaminen
  - Normaalisuuden testaaminen

## $\chi^2$ -yhteensopivuustesti 1/2

---

- Tavanomaisen *t*-testin yleisessä hypoteesissa oletetaan, että havainnot muodostavat *satunnaisotoksen normaali-jakaumasta*.
- Siten *t*-testin yleiseen hypoteesiin sisältyy havaintojen *normaalisuutta koskeva jakaumaoletus*.
- *Normaalisuusoletuksesta* pidetään kiinni *t*-testiä tehtäessä, mutta *oletusta havaintojen normaalisuudesta voidaan – ja on tavallisesti myös syytä – testata erikseen*.
- Huomautus:  
Jos havainnot eivät muodosta yksinkertaista satunnaisotosta normaalijakaumasta, tavanomainen *t*-testisuure ei ole jakautunut Studentin *t*-jakauman mukaan.

## $\chi^2$ -yhteensopivuustesti 2/2

---

- Jakaumaoletuksia koskevia tilastollisia testejä kutsutaan usein **yhteensopivuustesteiksi**.
- Nimitys johtuu siitä, että yhteensopivuustesteissä tutkitaan *sopivatko havainnot ja tehty jakaumaoletus toisiinsa eli ovatko havainnot sopuinnussa tehdyn jakaumaoletuksen kanssa*.
- *Yleisenä yhteensopivuustestinä* tilastotieteessä käytetään  **$\chi^2$ -testiä**.
- Koska normaalijakaumalla on niin keskeinen asema tilastotieteessä, *normaalisuuden testaamiseen* on kehitetty nimenomaan tähän erikoistarkoitukseen tarkoitettuja testejä; ks. kappaletta **Normaalisuuden testaaminen**.

## $\chi^2$ -homogeenisuustesti

---

- Oletetaan, että tilastollisen tutkimuksen kohteena oleva perusjoukko voidaan jakaa *kahteen* tai *useampaan ryhmään*.
- Tehtävänä on selvittää noudattaako tutkimuksen kohteena olevaa perusjoukon alkioden ominaisuutta kuvaava muuttuja kaikissa ryhmissä *samaa jakaumaa*.
- Jos muuttuja noudattaa kaikissa ryhmissä *samaa jakaumaa*, havaintoaineisto on tutkimuksen kohteena olevan perusjoukon ominaisuuden suhteen *homogeeninen*.
- *Yleisenä homogeenisuustestinä* tilastotieteessä käytetään  **$\chi^2$ -testiä**.

## $\chi^2$ -riippumattomuustesti

---

- Oletetaan, että tilastollisen tutkimuksen kohteena olevan perusjoukon alkiot voidaan luokitella ristiin kahden *faktorin* eli *tekijän*  $A$  ja  $B$  suhteen.
- Tehtävänä on selvittää ovatko tekijät  $A$  ja  $B$  *riippumattomia*.
- Jos tekijät  $A$  ja  $B$  ovat riippumattomia, tekijöitä  $A$  ja  $B$  voidaan tarkastella *erillisinä*.
- *Yleisenä riippumattomuustestinä* tilastotieteessä käytetään  **$\chi^2$ -testiä**.
- Huomautus:

Testi voidaan yleistää koskemaan *useamman kuin kahden tekijän* riippumattomuutta.

## $\chi^2$ -testien jakaumista riippumattomuus ja ei-parametrisuus

---

- $\chi^2$ -testit yhteensopivuudelle, homogeenisuudelle ja riippumattomuudelle ovat **jakaumista riippumattomia, ei-parametrisia** testejä:
  - (i) Testien yleiset hypoteesit *eivät kiinnitä havaintojen jakaumaa.*
  - (ii) Testeissä *ei testata todennäköisyysjakauman parametreja koskevia hypoteeseja.*

## Normaalisuuden testaaminen

---

- Koska normaalijakaumalla on niin keskeinen asema tilastotieteessä, *havaintojen normaalisuuden* tutkimista varten on kehitetty useita erilaisia menetelmiä.
- Havaintojen normaalisuutta voidaan *testata* yleisellä  $\chi^2$ -*yhteensopivuustestillä* tai erityisesti *normaalisuus-oletuksen testaamista varten konstruoiduilla testeillä*.
- **Bowmanin ja Shentonin testi** normaalisuudelle perustuu havaintojen *vinouden ja huipukkuuden* mittoihin.
- **Wilkin ja Shapiroin testi** normaalisuudelle perustuu ns. *rankit plot -kuviioon*, jota voidaan käyttää havaintojen normaalisuuden *graafiseen* tutkimiseen.



## Testit

---

- Tässä esityksessä tarkastellaan seuraavia sekä *diskreeteille* että *jatkuville* muuttujille tarkoitettuja testejä:
  - $\chi^2$ -yhteensopivuustesti
  - $\chi^2$ -homogeenisuustesti
  - $\chi^2$ -riippumattomuustesti
  - **Bowmanin ja Shentonin testi normalisuudelle**
  - **Wilkin ja Shapiron testi normalisuudelle**

# Yhteensopivuuden, homogeenisuuden ja riippumattomuuden testaaminen

---

Jakaumaoletuksien testaaminen

>> Yhteensopivuuden testaaminen

Homogeenisuuden testaaminen

Riippumattomuuden testaaminen

Normaalisuuden testaaminen

Yhteensopivuuden testaaminen

## $\chi^2$ -yhteensopivuustesti: Testausasetelma 1/2

---

- Tarkastellaan tutkimustilannetta, jossa perusjoukon  $S$  alkioita kuvataan *faktorilla* eli *tekijällä*  $A$ .
- Tekijä  $A$  saa olla *laatuero-*, *järjestys-*, *välimatka-* tai *suhdeasteikollinen muuttuja*.
- Tehdään oletus, että tekijä  $A$  *noudattaa perusjoukossa*  $S$  *jotakin määrättyä todennäköisyysjakaumaa*.

## $\chi^2$ -yhteensopivuustesti: Testausasetelma 2/2

---

- Poimitaan perusjoukosta  $S$  yksinkertainen satunnaisotos.
- Haluamme *testata* otostietojen perusteella tehtyä *jakaumaoletusta*, joka voidaan muotoilla seuraavilla tavoilla:
  - (i) Voidaanko havaintojen jakaumaa *kuvata* oletuksen määrittelemällä todennäköisyysjakaumalla?
  - (ii) Voiko otos olla oletuksen määrittelemän todennäköisyysjakauman *generoima* eli *tuottama*?
- **Yhteensopivuustestissä** tutkitaan ovatko otos ja tehty jakaumaoletus *yhteensopivia*.

## $\chi^2$ -yhteensopivuustesti: Testin suoritus

---

- $\chi^2$ -yhteensopivuustestissä havaintojen ja havaintojen jakaumasta tehdyn oletuksen *yhteensopivuutta mitataan seuraavalla tavalla:*
  - (1) Valitaan havainnoille sopiva *luokitus*.
  - (2) Määrätään *havaintojen luokkafrekvenssit*.
  - (3) Määrätään tehdyn jakaumaoletuksen mukaiset *odotetut luokkafrekvenssit*.
  - (4) *Verrataan* havaittuja ja odotettuja luokkafrekvenssejä toisiinsa  $\chi^2$ -testisuureella.

## $\chi^2$ -yhteensopivuustesti: Hypoteesit

---

- *Yleinen hypoteesi*  $H$  :

Havainnot  $X_1, X_2, \dots, X_n$  on saatu poimimalla yksinkertainen satunnaisotos perusjoukosta  $S$ .

- *Nollahypoteesi*  $H_0$  :

Havainnot  $X_1, X_2, \dots, X_n$  noudattavat *todennäköisyysjakaumaa*  $f(x; \theta)$ , jonka parametrit eivät kuitenkaan välttämättä ole tunnettuja.

- *Vaihtoehtoinen hypoteesi*  $H_1$  :

Havainnot  $X_1, X_2, \dots, X_n$  *eivät noudata* nollahypoteesin  $H_0$  määrittelemää todennäköisyysjakaumaa.

Yhteensopivuuden testaaminen

## $\chi^2$ -yhteensopivuustesti: Havaitut luokkafrekvenssit

---

- *Luokitellaan* havainnot  $X_1, X_2, \dots, X_n$  *toisensa pois-sulkeviin luokkiin*, joiden lukumäärä olkoon  $m$ .

- Olkoon

$$O_k, k = 1, 2, \dots, m$$

niiden havaintojen *frekvenssi* eli *lukumäärä* jotka kuuluvat luokkaan  $k$ .

- Frekvenssi  $O_k$  on luokkaan  $k$  kuuluvien havaintojen **havaittu frekvenssi**.

## $\chi^2$ -yhteensopivuustesti:

### Havaittujen luokkafrekvenssien määrittäminen 1/2

---

- Oletetaan, että havainnot  $X_1, X_2, \dots, X_n$  ovat *diskreetin* satunnaismuuttujan  $X$  havaittuja arvoja ja, että satunnaismuuttujan  $X$  *mahdolliset arvot* ovat

$$y_1, y_2, \dots, y_m$$

- *Luokitellaan* havainto  $X_i$  luokkaan  $k$ , jos

$$X_i = y_k, i = 1, 2, \dots, n, k = 1, 2, \dots, m$$

- Luokkaan  $k$  kuuluvien havaintojen  $X_i$  *havaittu frekvenssi*  $O_k$  on niiden havaintojen *lukumäärä*, jotka saavat arvon  $y_k$ .



## $\chi^2$ -yhteensopivuustesti:

### Havaittujen luokkafrekvenssien määrittäminen 2/2

---

- Oletetaan, että havainnot  $X_1, X_2, \dots, X_n$  ovat *jatkuvan* satunnaismuuttujan  $X$  havaittuja arvoja ja, että

$$X \in (a, b)$$

- Jaetaan väli  $(a, b)$  pisteillä

$$a = a_0 < a_1 < a_2 < \dots < a_{m-1} < a_m = b$$

pistevieraisiin *osaväleihin*  $(a_{k-1}, a_k]$ ,  $k = 1, 2, \dots, m$

- *Luokitellaan* havainto  $X_i$  luokkaan  $k$ , jos

$$X_i \in (a_{k-1}, a_k], i = 1, 2, \dots, n, k = 1, 2, \dots, m$$

- Luokkaan  $k$  kuuluvien havaintojen  $X_i$  *havaittu frekvenssi*  $O_k$  on niiden havaintojen *lukumäärä*, jotka kuuluvat väliin  $k$ .

## $\chi^2$ -yhteensopivuustesti:

### Havaittujen luokkafrekvenssien taulukko

---

- *Havaitut luokkafrekvenssit*  $O_k$  voidaan esittää **frekvenssi-taulukkona** seuraavassa muodossa:

Luokka	1	2	...	$m$	Summa
Havaittu frekvenssi	$O_1$	$O_2$	...	$O_m$	$n$

- Frekvenssiä  $O_k$  kutsutaan tavallisesti *havaituksi solufrekvenssiksi frekvenssitaulukon solussa k*.
- Havaitut solufrekvenssit  $O_k$  toteuttavat yhtälön

$$\sum_{k=1}^m O_k = n$$

## $\chi^2$ -yhteensopivuustesti:

### Odotetut luokkafrekvenssit 1/2

---

- Oletetaan, että havainnot  $X_1, X_2, \dots, X_n$  ovat satunnaismuuttujan  $X$  havaittuja arvoja ja, että havainnot on *luokiteltu* toisensa poissulkeviin luokkiin, joiden lukumäärä on  $m$ .
- Oletetaan, että *nollahypoteesi*  $H_0$  *täysin määrää satunnaismuuttujan  $X$  jakauman*.
- Olkoon  $P_k$  tällöin *todennäköisyys* sille, että satunnaismuuttuja  $X$  saa arvon luokasta  $k$ , kun *nollahypoteesi*  $H_0$  *pätee*.
- Tällöin luokkaan  $k$  kuuluvien havaintojen **odotettu frekvenssi**  $E_k$  on

$$E_k = nP_k, k = 1, 2, \dots, m$$

## $\chi^2$ -yhteensopivuustesti:

### Odotetut luokkafrekvenssit 2/2

---

- Oletetaan, että *nollahypoteesi*  $H_0$  määrää *satunnaisuuttujan*  $X$  jakauman tyypin, mutta jakauman parametrit ovat *tuntemattomia*.
- Koska jakauman parametreja ei tunneta, jakaumasta ei voida määrätä todennäköisyyksiä, ellei jakauman parametreja ensin *estimoida* havainnoista.
- Olkoon  $P_k$  tällöin *estimoitu todennäköisyys* sille, että satunnaisuuttuja  $X$  saa arvon luokasta  $k$ , kun *nollahypoteesi*  $H_0$  pätee.
- Tällöin luokkaan  $k$  kuuluvien havaintojen **odotettu frekvenssi**  $E_k$  on

$$E_k = nP_k, k = 1, 2, \dots, m$$

## $\chi^2$ -yhteensopivuustesti: Luokkatodennäköisyydet 1/2

---

- Oletetaan, että  $X$  on *diskreetti* satunnaismuuttuja, jonka mahdolliset arvot ovat  $y_1, y_2, \dots, y_m$ .

- Tällöin

$$P_k = \Pr(X = y_k), \quad k = 1, 2, \dots, m$$

jossa todennäköisyys  $\Pr(X = y_k)$  määrätään olettaen, että *nollahypoteesi*  $H_0$  pätee.

- Todennäköisyydet  $\Pr(X = y_k)$  voidaan määrätä satunnaismuuttujan  $X$  *kertymäfunktion* tai *pistetodennäköisyysfunktion* avulla.

## $\chi^2$ -yhteensopivuustesti: Luokkatodennäköisyydet 2/2

---

- Oletetaan, että  $X$  on *jatkuva* satunnaismuuttuja, joka saa arvoja väliltä  $(a, b)$  ja, että väli  $(a, b)$  on jaettu pisteillä

$$a = a_0 < a_1 < a_2 < \dots < a_{m-1} < a_m = b$$

pistevieraisiin osaväleihin.

- Tällöin

$$P_k = \Pr(a_{k-1} < X \leq a_k), k = 1, 2, \dots, m$$

jossa todennäköisyys  $\Pr(a_{k-1} < X \leq a_k)$  määrätään olettaen, että *nollahypoteesi*  $H_0$  pätee.

- Todennäköisyydet  $\Pr(a_{k-1} < X \leq a_k)$  voidaan määrätä satunnaismuuttujan  $X$  *kertymäfunktion* tai *tiheysfunktion* avulla.

## $\chi^2$ -yhteensopivuustesti:

### Odottettujen luokkafrekvenssien taulukko

---

- *Odotetut frekvenssit*  $E_k$  voidaan esittää **frekvenssi-taulukkona** seuraavassa muodossa:

<b>Luokka</b>	<b>1</b>	<b>2</b>	<b>...</b>	<b><math>m</math></b>	<b>Summa</b>
<b>Odotettu frekvenssi</b>	$E_1$	$E_2$	<b>...</b>	$E_m$	$n$

- Frekvenssiä  $E_k$  kutsutaan tavallisesti *odotetuksi solufrekvenssiksi frekvenssitaulukon solussa  $k$* .
- Odotetut solufrekvenssit  $E_k$  toteuttavat yhtälön

$$\sum_{k=1}^m E_k = n$$

## $\chi^2$ -yhteensopivuustesti:

### Testin idea

---

- *Testi nollahypoteesille  $H_0$  perustuu havaittujen frekvenssien  $O_k$  ja odotettujen frekvenssien  $E_k$  vertailuun.*
- *Jos havaittujen frekvenssien  $O_k$  ja odotettujen frekvenssien  $E_k$  jakaumat muistuttavat toisiaan, havainnot ovat sopu-soinnussa nollahypoteesin  $H_0$  kanssa.*



## Yhteensopivuuden testaaminen

# $\chi^2$ -yhteensopivuustesti:

### Testisuure – muoto 1

---

- Määritellään  $\chi^2$ -testisuure

$$\chi^2 = \sum_{k=1}^m \frac{(O_k - E_k)^2}{E_k}$$

jossa

$O_k =$  havaittu frekvenssi luokassa  $k$

$E_k =$  odotettu frekvenssi luokassa  $k$

$m =$  luokkien lukumäärä

- Testisuure  $\chi^2$  mittaa havaittujen ja odotettujen frekvenssien jakaumien yhteensopivuutta tai etäisyyttä ja siksi sitä kutsutaan usein  $\chi^2$ -etäisyydeksi.

Yhteensopivuuden testaaminen

## $\chi^2$ -yhteensopivuustesti: Testisuure – muoto 2

---

- $\chi^2$ -testisuure voidaan kirjoittaa myös muotoon

$$\chi^2 = n \sum_{k=1}^m \frac{(\hat{p}_k - P_k)^2}{P_k}$$

jossa

$\hat{p}_k = O_k/n =$  *havaittu suhteellinen frekvenssi*  
luokassa  $k$

$P_k =$  *todennäköisyys, että havainto kuuluu luokkaan  $k$ ,*  
*kun nollahypoteesi  $H_0$  pätee*

$m =$  luokkien lukumäärä

Yhteensopivuuden testaaminen

## $\chi^2$ -yhteensopivuustesti:

### Testisuureen asymptoottinen jakauma

---

- *Jos nollahypoteesi  $H_0$  pätee, testisuure  $\chi^2$  noudattaa suurissa otoksissa approksimatiivisesti  $\chi^2$ -jakaumaa vapausastein  $f = m - 1 - p$ :*

$$\chi^2 \sim_a \chi^2(f)$$

jossa

$m$  = luokkien lukumäärä

$p$  = odotettujen frekvenssien  $E_k$  määrittämiseksi  
estimoitujen parametrien lukumäärä

## $\chi^2$ -yhteensopivuustesti: Jakauma-approksimaation hyvyys

---

- Testisuure  $\chi^2$  noudattaa *suurissa otoksissa approksimatiivisesti*  $\chi^2$ -jakaumaa, jos nollahypoteesi  $H_0$  pätee:

$$\chi^2 \sim_a \chi^2(f), f = m - 1 - p$$

- Approksimaatio on tavallisesti *riittävän hyvä*, jos *odotetut frekvenssit*  $E_k$  toteuttavat ehdot

$$E_k > 5, k = 1, 2, \dots, m$$

- Ehdot saadaan toteutumaan valitsemalla havainnoille *sopiva luokitus*.

## $\chi^2$ -yhteensopivuustesti:

### Testisuureen normaaliarvo ja testi

---

- Testisuureen  $\chi^2$  normaaliarvo eli odotusarvo nollahypoteesin  $H_0$  pätiessä on

$$E(\chi^2) = f$$

jossa

$$f = m - 1 - p$$

- Normaaliarvoaan *merkitsevästi suuremmat*  $\chi^2$ -testisuureen arvot viittaavat siihen, että nollahypoteesi  $H_0$  *ei päde*.
- Normaaliarvoaan *merkitsevästi pienemmät*  $\chi^2$ -testisuureen arvot viittaavat siihen, että nollahypoteesi  $H_0$  *pätee liian hyvin*:

Havainnot saattavat olla *väärennettyjä*.

## $\chi^2$ -yhteensopivuustesti: Kommentteja

---

- $\chi^2$ -yhteensopivuustesti on **jakaumista riippumaton, ei-parametrinen** testi:
  - (i) Testin yleinen hypoteesi *ei kiinnitä havaintojen jakaumaa*, joten se soveltuu kaikille todennäköisyysjakaumille.
  - (ii) Testissä *ei testata todennäköisyysjakauman parametreja koskevaa hypoteesia*, vaan oletusta havaintojen jakaumasta.

Yhteensopivuuden testaaminen

## $\chi^2$ -yhteensopivuustesti: Esimerkki 1/13

---

- Luvuissa

**Tilastollisten aineistojen kuvaaminen**

**Väliestimointi**

**Tilastollinen testaus**

on käsitelty seuraavaa esimerkkiä :

- Kone tekee *ruuveja*, joiden *tavoitepituus* on 10 cm.
- *Ruuvien pituus saa vaihdella satunnaisesti* jonkin verran, *kunhan valmistettujen ruuvien keskimääräinen pituus on mahdollisimman lähellä tavoitearvoaan.*

## Yhteensopivuuden testaaminen

# $\chi^2$ -yhteensopivuustesti: Esimerkki 2/13

---

- Ruuvien laadunvalvonnassa niiden *keskimääräistä pituutta* tutkitaan siten, että jokaisesta ruuvierästä poimitaan *yksinkertainen satunnaisotos* ja otokseen poimittujen ruuvien pituudet mitataan.
- Olemme aikaisemmin soveltaneet näin kerättyyn esimerkkiaineistoon seuraavia tilastollisia menetelmiä:
  - (i) Luvussa **Tilastollisten aineistojen kuvaaminen** näytettiin, miten ruuvien *pituuden jakaumaa otoksessa voidaan kuvata luokitellulla frekvenssijakaumalla* ja sitä vastaavalla *histogrammilla*.
  - (ii) Luvussa **Väliestimointi** näytettiin, miten koneen tekemien ruuvien *keskipituudelle voidaan konstruoida otostietojen perusteella luottamusväli*.
  - (iii) Luvussa **Tilastollinen testaus** näytettiin, miten voidaan *testata ovatko otoksessa saadut tiedot ruuvien keskipituudesta sopusoinnussa ruuvien tavoitepituuden kanssa*.



## Yhteensopivuuden testaaminen

# $\chi^2$ -yhteensopivuustesti: Esimerkki 3/13

---

- Sekä ruuvien keskipituuden *luottamusväli* (ks. lukua **Väliestimointi**) että *testi* ruuvien keskipituuden tavoitearvolle (ks. lukua **Tilastollinen testaus**) perustuivat oletukseen, jonka mukaan *ruuvien pituus vaihtelee normaalijakauman mukaan*.
- Tätä **jakaumaoletusta** voidaan testata  $\chi^2$ -yhteensopivuustestillä seuraavassa esitettävällä tavalla.
- **Yleinen hypoteesi  $H$**  :  
Ruuvit on poimittu *yksinkertaisella satunnaisotannalla* koneen tekemien ruuvien joukosta.
- **Nollahypoteesi  $H_0$**  :  
Koneen tekemien ruuvien pituudet *noudattavat normaalijakaumaa*.
- **Vaihtoehtoinen hypoteesi  $H_1$**  :  
Koneen tekemien ruuvien pituudet *eivät noudata normaalijakaumaa*.

## Yhteensopivuuden testaaminen

# $\chi^2$ -yhteensopivuustesti:

### Esimerkki 4/13

---

- Koneen valmistamien ruuvien joukosta poimittiin siis *yksinkertainen satunnaisotos*, jonka *koko*  $n = 30$  ja otokseen poimittujen ruuvien pituudet mitattiin.
- Ruuvien pituuksien *aritmeettinen keskiarvo* otoksessa oli
$$\bar{X} = 10.09 \text{ cm}$$
ja *otoskeskihajonta* oli
$$s = 0.1038 \text{ cm}$$
- Taulukko oikealla esittää pituuksien *luokiteltua frekvenssi-jakaumaa*.

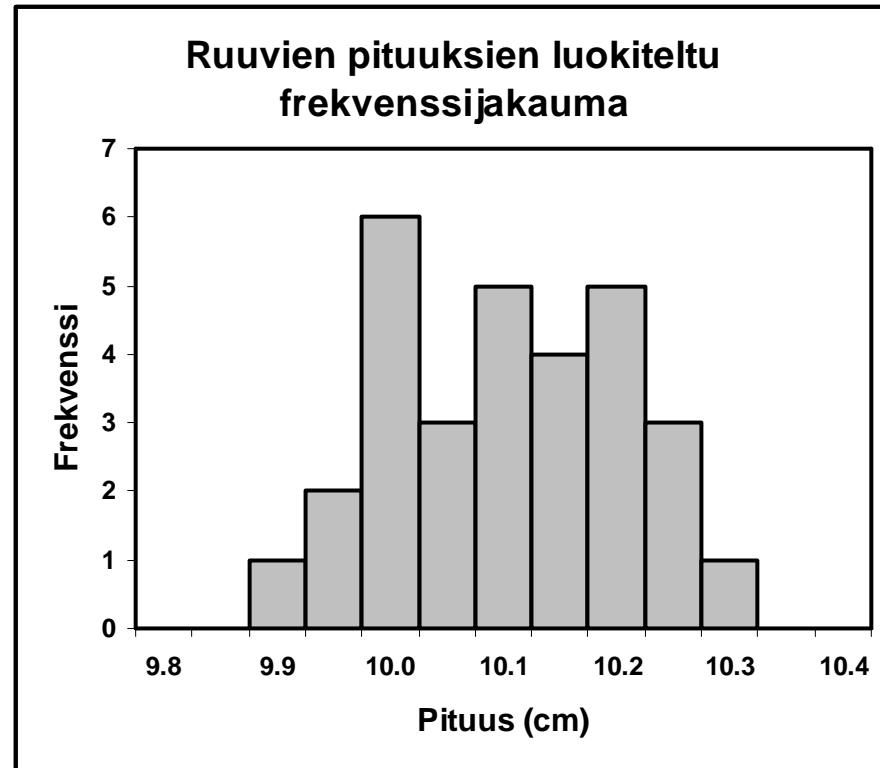
Luokkavälit	Luokkafrekvenssit
(9.85,9.90]	1
(9.90,9.95]	2
(9.95,10.00]	6
(10.00,10.05]	3
(10.05,10.10]	5
(10.10,10.15]	4
(10.15,10.20]	5
(10.20,10.25]	3
(10.25,10.30]	1

## Yhteensopivuuden testaaminen

# $\chi^2$ -yhteensopivuustesti:

### Esimerkki 5/13

- Kuva oikealla esittää otokseen poimitujen ruuvien pituuksien luokiteltua frekvenssijakaumaa vastaavaa *histogrammia*.
- Voisiko tällainen pituuksien jakauma syntyä *normaalijakautuneesta perusjoukosta* poimitusta yksinkertaisesta satunnaisotoksesta?
- Tähän kysymykseen antaa vastauksen  **$\chi^2$ -yhteensopivuustesti**.



## Yhteensopivuuden testaaminen

# $\chi^2$ -yhteensopivuustesti: Esimerkki 6/13

- $\chi^2$ -yhteensopivuustestin vaatimat laskutoimitukset voidaan järjestää seuraavan taulukon muotoon (taulukko on luotu *Microsoft Excel* -ohjelmalla):

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Luokka	Luokan yläraja	Havaittu luokka-frekvenssi	Standardoitu luokan yläraja	Kertymäfunktion arvo	Kertymäfunktion arvojen erotus	Odotettu luokka-frekvenssi	Khi <sup>2</sup> -arvo
$k$	$a_k$	$O_k$	$z_k$	$F(z_k)$	$F(z_k) - F(z_{k-1})$	$E_k$	$\chi_k^2$
1	9.90	1	-1.79444	0.03637	0.03637	1.09115	0.00762
2	9.95	2	-1.31292	0.09460	0.05823	1.74698	0.03665
3	10.00	6	-0.83141	0.20287	0.10827	3.24799	2.33177
4	10.05	3	-0.34990	0.36321	0.16034	4.81010	0.68116
5	10.10	5	0.13161	0.55235	0.18915	5.67443	0.08016
6	10.15	4	0.61313	0.73010	0.17775	5.33245	0.33295
7	10.20	5	1.09464	0.86316	0.13306	3.99177	0.25466
8	10.25	3	1.57615	0.94250	0.07934	2.38026	0.16136
9	10.30	1	2.05766	0.98019	0.05750	1.72487	0.30462
<b>Summa</b>	*	30	*	*	1	30	4.190937

## Yhteensopivuuden testaaminen

# $\chi^2$ -yhteensopivuustesti: Esimerkki 7/13

---

- Kalvon 6/13 taulukon *sarakkeet*:

(1) Luokka:  $k = 1, 2, \dots, m = 9$

(2) Luokan  $k$  yläraja:  $a_k$

(3) *Havaittu luokkafrekvenssi* luokassa  $k$ :  $O_k$

$$\sum_{k=1}^m O_k = n = 30$$

(4) Luokan  $k$  yläraja  $a_k$  standardoituna:

$$z_k = \frac{a_k - \bar{X}}{s} = \frac{a_k - 10.09}{0.1038}$$

jossa  $\bar{X} = 10.09$  cm on ruuvien pituuksien aritmeettinen keskiarvo ja  $s = 0.1038$  cm on pituuksien keskihajonta otoksessa.

## $\chi^2$ -yhteensopivuustesti: Esimerkki 8/13

---

- Kalvon 6/13 taulukon *sarakkeet* (jatkuu):

(5) Standardoidun normaalijakauman  $N(0,1)$  kertymäfunktion  $F(\cdot)$  arvo pisteessä  $z_k$  :  $F(z_k)$

(6) Kertymäfunktion arvojen erotus

$$F(z_k) - F(z_{k-1}) = \Pr(z_{k-1} < z \leq z_k) = P_k$$

on todennäköisyys, että havainto kuuluu luokkaan  $k$ , jos nollahypoteesi  $H_0$  ruuvien pituuden normaalijakautuneisuudesta pätee.

$$\sum_{k=1}^m P_k = 1$$

koska valitun luokituksen ulkopuolelle jääneet normaalijakauman häntäalueiden todennäköisyysmassat on yhdistetty reunaluokkiin.

(7) *Odotettu luokkafrekvenssi* luokassa  $k$ :  $E_k = nP_k$

$$\sum_{k=1}^m E_k = n = 30$$

## Yhteensopivuuden testaaminen

# $\chi^2$ -yhteensopivuustesti: Esimerkki 9/13

---

- Kalvon 6/13 taulukon *sarakkeet* (jatkuu):

(8) Luokan  $k$   $\chi^2$ -arvo:

$$\chi_k^2 = \frac{(O_k - E_k)^2}{E_k}$$

jossa

$O_k$  = havaittu frekvenssi luokassa  $k$

$E_k$  = odotettu frekvenssi luokassa  $k$

- **$\chi^2$ -yhteensopivuustestin** testisuureen arvo saadaan sarakkeen (8) lukujen sarakesummana:

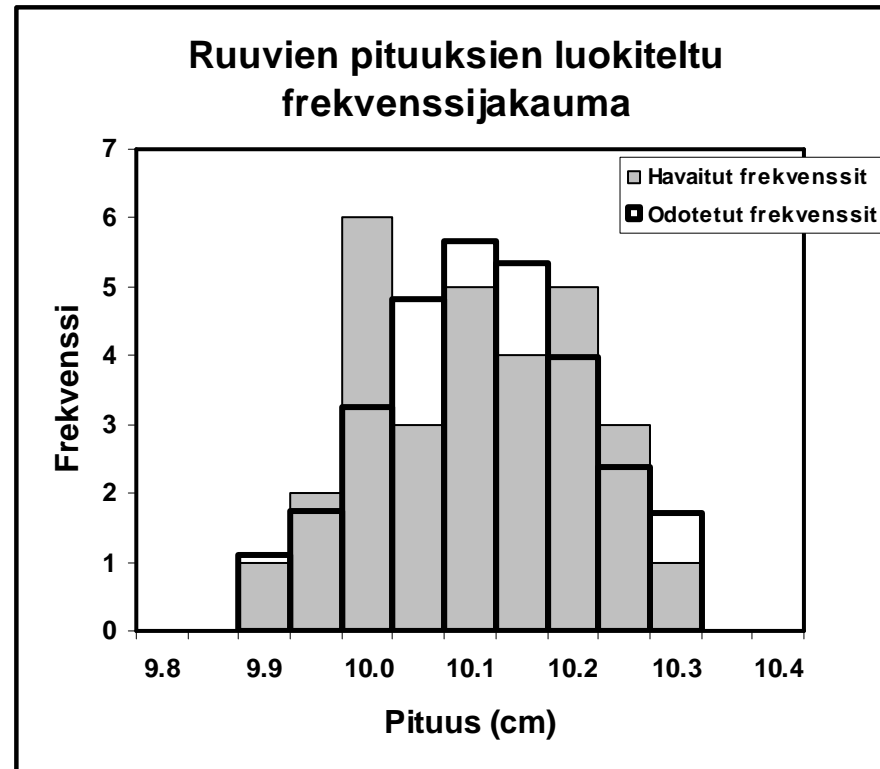
$$\chi^2 = \sum_{k=1}^m \chi_k^2 = \sum_{k=1}^m \frac{(O_k - E_k)^2}{E_k} = 4.20$$

## Yhteensopivuuden testaaminen

# $\chi^2$ -yhteensopivuustesti:

### Esimerkki 10/13

- $\chi^2$ -yhteensopivuustesti vertaa havaittuja frekvenssejä  $O_k$  ja nollahypoteesin mukaan odotettuja frekvenssejä  $E_k$  toisiinsa.
- Geometrisesti vertailu merkitsee havaittuja luokkafrekvenssejä vastaavien suorakaiteiden pinta-alojen  $O_k$  vertaamista odotettuja frekvenssejä  $E_k$  vastaavien suorakaiteiden pinta-aloihin; ks. kuvaa oikealla.





## Yhteensopivuuden testaaminen

# $\chi^2$ -yhteensopivuustesti: Esimerkki 11/13

---

- *Nollahypoteesin*

$H_0$  : Ruuvien pituudet noudattavat *normaalijakaumaa*

*pätiessä testisuure  $\chi^2$  noudattaa  $\chi^2$ -jakaumaa* vapausastein  $m - 1 - p$ :

$$\chi^2 \sim \chi^2(m - 1 - p)$$

jossa

$m$  = luokkien lukumäärä

$p$  = odotettujen frekvenssien  $E_k$  määrittämiseksi estimoitujen parametrien lukumäärä

- Esimerkissä

$$m - 1 - p = 9 - 1 - 2 = 6$$

## Yhteensopivuuden testaaminen

# $\chi^2$ -yhteensopivuustesti: Esimerkki 12/13

---

- Valitaan **merkitsevyystasoksi**

$$\alpha = 0.05$$

- Merkitsevyystasoa  $\alpha = 0.05$  vastaava **kriittinen arvo** on

$$\chi_{0.05}^2 = 12.592$$

koska  $\chi^2$ -jakauman taulukoiden mukaan

$$\Pr(\chi^2 \geq 12.592) = 0.05$$

jossa

$$\chi^2 \sim \chi^2(6)$$

- Merkitsevyystasoa  $\alpha = 0.05$  vastaava **hylkäysalue** on siten muotoa  
(12.592,  $+\infty$ )

## Yhteensopivuuden testaaminen

# $\chi^2$ -yhteensopivuustesti: Esimerkki 13/13

---

- Koska  $\chi^2$ -yhteensopivuustestin testisuureen arvo

$$\chi^2 = 4.20 < 12.592$$

niin *nollahypoteesi jää voimaan* merkitsevyystasolla  $\alpha = 0.05$ :

**Havainnot ovat sopusoinnussa normaalisuusoletuksen kanssa.**

- Huomautuksia:

- Microsoft Excel -ohjelman mukaan  $\chi^2$ -yhteensopivuustestin testisuureen arvoa 4.20 vastaava ***p*-arvo** on 0.65.

Siten normaalisuusoletuksen hylkäämiseen ei ole myöskään testin *p*-arvon mukaan mitään perusteita.

- Tarkkaan ottaen *luokkia olisi pitänyt yhdistää* niin, että ehdot

$$E_k > 5, k = 1, 2, \dots, m$$

olisivat toteutuneet. Tällä ei kuitenkaan pitäisi *tässä* olla vaikutusta testin tulokseen.

# Yhteensopivuuden, homogeenisuuden ja riippumattomuuden testaaminen

---

Jakaumaoletuksien testaaminen

Yhteensopivuuden testaaminen

>> Homogeenisuuden testaaminen

Riippumattomuuden testaaminen

Normaalisuuden testaaminen

## Homogeenisuuden testaaminen

# $\chi^2$ -homogeenisuustesti: Testausasetelma 1/3

---

- Tarkastellaan tutkimusasetelmaa, jossa perusjoukon  $S$  alkioita kuvataan  *yhdellä faktorilla eli tekijällä  $A$ .*
  - Tekijä  $A$  saa olla *laatuero-, järjestys-, välimatka- tai suhdeasteikollinen muuttuja.*
  - Oletetaan, että perusjoukko  $S$  voidaan jakaa *kahteen tai useampaan ryhmään.*
  - Poimitaan ryhmistä *toisistaan riippumattomat yksinkertaiset satunnaisotokset.*
  - Tarkastellaan tekijän  $A$  *vaihtelua* otoksissa.
  - Tehdään oletus, että *tekijä  $A$  noudattaa kaikissa ryhmissä samaa, tarkemmin määrittelemätöntä todennäköisyysjakaumaa.*
-

## $\chi^2$ -homogeenisuustesti: Testausasetelma 2/3

---

- Haluamme *testata* tehtyä *jakaumaoletusta*:
  - (i) Voidaanko eri otoksista (ryhmistä) saatujen havaintoarvojen jakaumia *kuvata samalla todennäköisyysjakaumalla*?
  - (ii) Voivatko otokset olla *saman todennäköisyysjakauman generoimia eli tuottamia*?

## $\chi^2$ -homogeenisuustesti: Testausasetelma 3/3

---

- Jos tehty jakaumaoletus *pätee* ja tekijä *A* noudattaa kaikissa ryhmissä samaa jakaumaa, perusjoukko on **homogeeninen** ja perusjoukkoa ei tarvitse jakaa tekijää *A* koskevissa tarkasteluissa erillisiksi ryhmiksi.
- Jos tehty jakaumaoletus *ei päde* ja tekijä *A* noudattaa eri ryhmissä eri jakaumia, perusjoukko on **heterogeeninen** ja ryhmiä on syytä tarkastella erillisinä.
- Tällaisten *jakaumaoletuksien* testaamiseen tarkoitettuja testejä kutsutaan **homogeenisuustesteiksi**.
- Huomautus:  
Monissa tutkimusasetelmissa toivotaan, että homogeenisuusoletus *tulee* testissä *hylätyksi*.

## $\chi^2$ -homogeenisuustesti: Testin suoritus

---

- $\chi^2$ -homogeenisuustestissä havaintojen ja niiden jakaumasta eri ryhmissä tehdyn homogeenisuusoletuksen *yhteensopivuutta mitataan* seuraavalla tavalla:
  - (1) Valitaan havainnoille *yhteinen luokitus*, jota siis käytetään jokaisessa otoksessa.
  - (2) Määrätään *havaintojen luokkafrekvenssit* jokaisesta otoksesta.
  - (3) Määrätään tehdyn homogeenisuusoletuksen mukaiset *odotetut luokkafrekvenssit*.
  - (4) *Verrataan* havaittuja ja odotettuja frekvenssejä toisiinsa  $\chi^2$ -testisuureella.



## $\chi^2$ -homogeenisuustesti: Hypoteesit

---

- *Yleinen hypoteesi  $H$  :*  
Perusjoukko on jaettu  $r$  ryhmään, joista on poimittu *toisistaan riippumattomat yksinkertaiset satunnaisotokset.*
- *Nollahypoteesi  $H_0$  :*  
Otokset  $i = 1, 2, \dots, r$  on poimittu *samasta todennäköisyysjakaumasta.*
- *Vaihtoehtoinen hypoteesi  $H_1$  :*  
Otokset  $i = 1, 2, \dots, r$  on poimittu *eri todennäköisyysjakaumista.*

## $\chi^2$ -homogeenisuustesti: Havaitut frekvenssit

---

- Oletetaan, että tutkimuksen kohteena oleva perusjoukko  $S$  on jaettu  $r$  ryhmään.
- Poimitaan ryhmistä *toisistaan riippumattomat yksinkertaiset satunnaisotokset*, joiden koot ovat  $n_i$ ,  $i = 1, 2, \dots, r$ .
- *Luokitellaan* havainnot jokaisessa otoksessa *samaa luokitusta käyttäen* toisensa poissulkeviin luokkiin, joiden lukumäärä olkoon  $c$ .
- Määrätään ryhmässä  $i$  luokkaan  $j$  kuuluvien havaintojen **havaittu frekvenssi** eli *lukumäärä*  $O_{ij}$ , kun  $i = 1, 2, \dots, r$  ja  $j = 1, 2, \dots, c$ .

## Homogeenisuuden testaaminen

### $\chi^2$ -homogeenisuustesti:

### Havaittujen frekvenssien taulukko 1/3

---

- Muodostetaan *havaituista frekvensseistä*  $O_{ij}$   $(r \times c)$ -frekvenssitaulukko  $[O_{ij}]$  :

		Luokat				
		1	2	...	$c$	Summa
Ryhmät	1	$O_{11}$	$O_{12}$	...	$O_{1c}$	$n_1$
	2	$O_{21}$	$O_{22}$	...	$O_{2c}$	$n_2$
	...	...	...	...	...	...
	$r$	$O_{r1}$	$O_{r2}$	...	$O_{rc}$	$n_r$
	Summa	$C_1$	$C_2$	...	$C_c$	$n$

## $\chi^2$ -homogeenisuustesti:

### Havaittujen frekvenssien taulukko 2/3

---

- Olkoon  $[O_{ij}]$  havaittujen frekvenssien  $O_{ij}$  muodostama  $(r \times c)$ -frekvenssitaulukko.

$r$  = ryhmien lukumäärä

$c$  = luokkien lukumäärä

$O_{ij}$  = havaittu frekvenssi ryhmän  $i$  luokassa  $j$ ,  
 $i = 1, 2, \dots, r, j = 1, 2, \dots, c$

$n_i$  = otoskoko ryhmässä  $i$

$C_j$  = havaittu frekvenssi yhdistetyn havainto-  
aineiston luokassa  $j$

$n$  = havaintojen kokonaislukumäärä

- Frekvenssiä  $O_{ij}$  kutsutaan tavallisesti *havaituksi solu-*  
*frekvenssiksi frekvenssitaulukon solussa  $(i, j)$ .*

## $\chi^2$ -homogeenisuustesti:

### Havaittujen frekvenssien taulukko 3/3

---

- *Havaittujen frekvenssien*  $O_{ij}$  frekvenssitaulukossa pätee:

- (i) *Rivisummat* yhtyvät ryhmäkohtaisiin otoskokoihin:

$$\sum_{j=1}^c O_{ij} = n_i, i = 1, 2, \dots, r$$

- (ii) *Sarakesummat* yhtyvät yhdistetyn havaintoaineiston luokkafrekvensseihin:

$$\sum_{i=1}^r O_{ij} = C_j, j = 1, 2, \dots, c$$

- (iii) *Havaintojen kokonaislukumäärä*:

$$\sum_{i=1}^r \sum_{j=1}^c O_{ij} = \sum_{i=1}^r n_i = \sum_{j=1}^c C_j = n$$

## $\chi^2$ -homogeenisuustesti: Nollahypoteesin tulkinta

---

- *Jos nollahypoteesi  $H_0$  pätee, havaintojen pitäisi jakautua (satunnaisvaihtelua lukuun ottamatta) jokaisessa ryhmässä  $i = 1, 2, \dots, r$  samalla tavalla luokkiin  $j = 1, 2, \dots, c$ .*
- *Jos nollahypoteesi  $H_0$  pätee, havaintojen jakautuminen luokkiin  $j = 1, 2, \dots, c$  ei saa riippua siitä, mihin ryhmään  $i = 1, 2, \dots, r$  ne kuuluvat.*
- *Jos nollahypoteesi  $H_0$  pätee, todennäköisyys, että havainto kuuluu luokkaan  $j = 1, 2, \dots, c$  ei saa riippua siitä, mihin ryhmään  $i = 1, 2, \dots, r$  se kuuluu.*

## $\chi^2$ -homogeenisuustesti: Odotetut frekvenssit 1/4

---

- Olkoon  $x$  on tarkastelun kohteena olevan perusjoukon  $S$  alkio.
- Määritellään seuraavat todennäköisyydet:

$$p_{ij} = \Pr(x \text{ kuuluu ryhmään } i \text{ ja luokkaan } j)$$

$$p_{j|i} = \Pr(x \text{ kuuluu luokkaan } j | x \text{ kuuluu ryhmään } i)$$

$$p_i = \Pr(x \text{ kuuluu ryhmään } i)$$

$$p_j = \Pr(x \text{ kuuluu luokkaan } j)$$

- Todennäköisyyslaskennan *yleisen tulosäännön* mukaan

$$p_{ij} = p_{j|i} p_i$$

## Homogeenisuuden testaaminen

# $\chi^2$ -homogeenisuustesti: Odotetut frekvenssit 2/4

---

- *Jos nollahypoteesi  $H_0$  pätee, todennäköisyys, että perusjoukon  $S$  alkio  $x$  kuuluu luokkaan  $j$  ei saa riippua siitä, mihin ryhmään  $i$  alkio  $x$  kuuluu.*
- *Siten nollahypoteesi  $H_0$  perusjoukon homogeenisuudesta voidaan ilmaista muodossa*

$$H_0 : p_{j|i} = p_j, i = 1, 2, \dots, r, j = 1, 2, \dots, c$$

*tai muodossa*

$$H_0 : p_{ij} = p_i p_j, i = 1, 2, \dots, r, j = 1, 2, \dots, c$$



## Homogeenisuuden testaaminen

# $\chi^2$ -homogeenisuustesti: Odotetut frekvenssit 3/4

---

- Todennäköisyydet  $p_{ij}$ ,  $p_i$ ,  $p_j$  voidaan estimoida havaituista frekvensseistä  $O_{ij}$ ,  $n_i$ ,  $C_j$  kaavoilla

$$\hat{p}_{ij} = \frac{O_{ij}}{n} \quad \hat{p}_i = \frac{n_i}{n} \quad \hat{p}_j = \frac{C_j}{n}$$

- *Jos nollahypoteesi  $H_0$  pätee, solutodennäköisyydet  $p_{ij}$  voidaan estimoida kaavoilla*

$$P_{ij} = \frac{n_i}{n} \times \frac{C_j}{n} = \hat{p}_i \hat{p}_j, \quad i = 1, 2, \dots, r, \quad j = 1, 2, \dots, c$$

## $\chi^2$ -homogeenisuustesti: Odotetut frekvenssit 4/4

---

- Määrätään *nollahypoteesin*  $H_0$  *pätiessä odotetut solu-*  
**frekvenssit**  $E_{ij}$  yhtälöillä

$$E_{ij} = nP_{ij} = \frac{C_j}{n} \times n_i = \frac{n_i C_j}{n}, i = 1, 2, \dots, r, j = 1, 2, \dots, c$$

- Tällöin *odotetut suhteelliset frekvenssit*  $E_{ij}/n_i$  *jakautuvat*  
*jokaisessa ryhmässä*  $i$  *samalla tavalla luokkiin*  
 $j = 1, 2, \dots, c$ :

$$\frac{E_{ij}}{n_i} = \frac{C_j}{n}, j = 1, 2, \dots, c, i = 1, 2, \dots, r$$

## Homogeenisuuden testaaminen

### $\chi^2$ -homogeenisuustesti:

### Odottujen frekvenssien taulukko 1/3

---

- Muodostetaan *odotetuista frekvensseistä*  $E_{ij}$   
( $r \times c$ )-frekvenssitaulukko  $[E_{ij}]$  :

**Luokat**

	1	2	...	$c$	Summa	
Ryhmät	1	$E_{11}$	$E_{12}$	...	$E_{1c}$	$n_1$
	2	$E_{21}$	$E_{22}$	...	$E_{2c}$	$n_2$
	...	...	...	...	...	...
	$r$	$E_{r1}$	$E_{r2}$	...	$E_{rc}$	$n_r$
	Summa	$C_1$	$C_2$	...	$C_c$	$n$

## $\chi^2$ -homogeenisuustesti:

### Odottettujen frekvenssien taulukko 2/3

---

- Olkoon  $[E_{ij}]$  odottettujen frekvenssien  $E_{ij}$  muodostama  $(r \times c)$ -frekvenssitaulukko.

$r$  = ryhmien lukumäärä

$c$  = luokkien lukumäärä

$E_{ij}$  = odotettu frekvenssi ryhmän  $i$  luokassa  $j$ ,  
 $i = 1, 2, \dots, r, j = 1, 2, \dots, c$

$n_i$  = otoskoko ryhmässä  $i$

$C_j$  = havaittu frekvenssi yhdistetyn havainto-  
aineiston luokassa  $j$

$n$  = havaintojen kokonaislukumäärä

- Frekvenssiä  $E_{ij}$  kutsutaan tavallisesti *odotetuksi solu-*  
*frekvenssiksi frekvenssitaulukon solussa*  $(i, j)$ .

## $\chi^2$ -homogeenisuustesti:

### Odottujen frekvenssien taulukko 3/3

---

- *Odottujen frekvenssien*  $E_{ij}$  frekvenssitaulukossa pätee:

- (i) *Rivisummat* yhtyvät ryhmäkohtaisiin otoskokoihin:

$$\sum_{j=1}^c E_{ij} = n_i, \quad i = 1, 2, \dots, r$$

- (ii) *Sarakesummat* yhtyvät yhdistetyn havaintoaineiston luokkafrekvensseihin:

$$\sum_{i=1}^r E_{ij} = C_j, \quad j = 1, 2, \dots, c$$

- (iii) *Havaintojen kokonaislukumäärä*:

$$\sum_{i=1}^r \sum_{j=1}^c E_{ij} = \sum_{i=1}^r n_i = \sum_{j=1}^c C_j = n$$

## $\chi^2$ -homogeenisuustesti: Testisuure – muoto 1

---

- Määritellään  $\chi^2$ -testisuure

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

jossa

$O_{ij}$  = havaittu frekvenssi solussa  $(i, j)$

$E_{ij}$  = odotettu frekvenssi solussa  $(i, j)$

$r$  = ryhmien lukumäärä

$c$  = luokkien lukumäärä

- Testisuure  $\chi^2$  mittaa havaittujen ja odotettujen frekvenssien jakaumien yhteensopivuutta tai etäisyyttä ja siksi sitä kutsutaan usein  $\chi^2$ -etäisyydeksi.

## Homogeenisuuden testaaminen

# $\chi^2$ -homogeenisuustesti:

### Testisuure – muoto 2

---

- $\chi^2$ -testisuure voidaan kirjoittaa myös muotoon

$$\chi^2 = n \sum_{i=1}^r \sum_{j=1}^c \frac{(\hat{p}_{ij} - P_{ij})^2}{P_{ij}}$$

jossa

$$\hat{p}_{ij} = \frac{O_{ij}}{n}$$

$$P_{ij} = \frac{E_{ij}}{n} = \frac{n_i}{n} \times \frac{C_j}{n} = \hat{p}_i \hat{p}_j$$

$$\hat{p}_i = \frac{n_i}{n} \quad \hat{p}_j = \frac{C_j}{n}$$

$$i = 1, 2, \dots, r, \quad j = 1, 2, \dots, c$$

## Homogeenisuuden testaaminen

### $\chi^2$ -homogeenisuustesti:

### Testisuureen asymptoottinen jakauma

---

- *Jos nollahypoteesi  $H_0$  pätee, testisuure  $\chi^2$  noudattaa suurissa otoksissa approksimatiivisesti  $\chi^2$ -jakaumaa vapausastein  $f = (r - 1)(c - 1)$ :*

$$\chi^2 \sim_a \chi^2(f)$$

jossa

$r =$  ryhmien lukumäärä

$c =$  luokkien lukumäärä



## $\chi^2$ -homogeenisuustesti: Jakauma-approksimaation hyvyys

---

- Testisuure  $\chi^2$  noudattaa *suurissa otoksissa approksimatiivisesti*  $\chi^2$  -jakaumaa, jos nollahypoteesi  $H_0$  pätee:

$$\chi^2 \sim_a \chi^2(f), f = (r-1)(c-1)$$

- Approksimaatio on tavallisesti *riittävän hyvä*, jos *odotetut frekvenssit*  $E_{ij}$  ja *keskimääräiset odotetut frekvenssit*  $C_j/r$  toteuttavat ehdot

$$E_{ij} > 1, i = 1, 2, \dots, r, j = 1, 2, \dots, c$$

$$C_j / r > 5, j = 1, 2, \dots, c$$

- Ehdot saadaan toteutumaan valitsemalla havainnoille *sopiva luokitus*.

## $\chi^2$ -homogeenisuustesti:

### Testisuureen normaaliarvo ja testi

---

- Testisuureen  $\chi^2$  normaaliarvo eli odotusarvo nollahypoteesin  $H_0$  pätiessä, on

$$E(\chi^2) = f$$

jossa

$$f = (r - 1)(c - 1)$$

- Normaaliarvoaan *merkitsevästi suuremmat*  $\chi^2$ -testisuureen arvot viittaavat siihen, että nollahypoteesi  $H_0$  *ei päde*.
- Normaaliarvoaan *merkitsevästi pienemmät*  $\chi^2$ -testisuureen arvot viittaavat siihen, että nollahypoteesi  $H_0$  *pätee liian hyvin*:

Havainnot saattavat olla *väärennettyjä*.

## $\chi^2$ -homogeenisuustesti: Kommentteja

---

- $\chi^2$ -homogeenisuustesti on **jakaumista riippumaton, ei-parametrinen** testi:
  - (i) Testin yleinen hypoteesi *ei kiinnitä havaintojen jakaumaa*.
  - (ii) Testissä *ei testata todennäköisyysjakauman parametreja koskevaa hypoteesia*, vaan oletusta havaintojen jakaumasta.

## $\chi^2$ -homogeenisuustesti ja $\chi^2$ -riippumattomuustesti

---

- $\chi^2$ -homogeenisuustesti ja seuraavassa kappaleessa esitettävä  $\chi^2$ -riippumattomuustesti muistuttavat toisiaan.
- Frekvenssitaulukosta *ei voi sellaisenaan nähdä* kummasta testausasetelmasta on kyse.
- $\chi^2$ -homogeenisuustesti ja  $\chi^2$ -riippumattomuustesti tehdään teknisesti täsmälleen *samalla tavalla*:
  - *Odotetut frekvenssit määrätään samalla kaavalla.*
  - *Testisuureet lasketaan samalla kaavalla.*
  - *Testisuureet noudattavat nollahypoteesin pätiessä approksimatiivisesti samaa jakaumaa.*
- Testien testausasetelmat ovat kuitenkin *täysin erilaiset*.

## Homogeenisuustesti vs riippumattomuustesti 1/2

---

- *Homogeenisuustestin* testausasetelma:
  - (i) Perusjoukko koostuu  $r$  ryhmästä ja testissä tarkastellaan perusjoukon alkioiden *jakautumista luokkiin eri ryhmissä*, kun *luokittelu on tehty kaikissa ryhmissä samaa luokitusta käyttäen*.
  - (ii) Havaintoaineisto muodostuu *toisistaan riippumattomista ryhmäkohtaisista satunnais-otoksista*.
  - (iii) Sekä ryhmäkohtaiset otoskoot  $n_i$  että havaintojen kokonaislukumäärä  $n$  ovat *kiinteitä eli ei-satunnaisia* (valittuja) lukuja, kun taas *sattuma määrää miten havainnot jakautuvat luokkiin ryhmien sisällä*.

## Homogeenisuustesti vs riippumattomuustesti 2/2

---

- *Riippumattomuustestin* testausasetelma:
  - (i) Testissä tarkastellaan *kahden tekijän A ja B assosiaatiota eli riippuvuutta*, kun havainnot on luokiteltu tekijöiden A ja B suhteen *ristiin*.
  - (ii) Havaintoaineisto muodostuu *yhdestä satunnaisotoksesta*.
  - (iii) Vain havaintojen kokonaislukumäärä *n* on *kiinteä* eli *ei-satunnainen* (valittu) luku, kun taas *sattumamäärää miten havainnot jakautuvat luokkiin tekijöiden A ja B ristiluokituksen suhteen*.

# Yhteensopivuuden, homogeenisuuden ja riippumattomuuden testaaminen

---

Jakaumaoletuksien testaaminen

Yhteensopivuuden testaaminen

Homogeenisuuden testaaminen

>> Riippumattomuuden testaaminen

Normaalisuuden testaaminen

## Riippumattomuuden testaaminen laatueroasteikollisilla muuttujilla

---

- Tarkastellaan kahden *laatueroasteikollisen muuttujan riippumattomuuden testaamista*.
- Esitettävä testi on  **$\chi^2$ -testi riippumattomuudelle**.
- $\chi^2$ -testi riippumattomuudelle sopii myös *järjestys-, välimatka- tai suhdeasteikollisille muuttujille*.
- $\chi^2$ -testi riippumattomuudelle on *jakaumista riippumaton, ei-parametrinen testi*.



## $\chi^2$ -riippumattomuustesti: Testausasetelma 1/2

---

- Tarkastellaan tutkimusasetelmaa, jossa perusjoukon  $S$  alkioita kuvataan *kahdella faktorilla* eli tekijällä  $A$  ja  $B$ .
- Tekijät  $A$  ja  $B$  saavat olla *laatuero-, järjestys-, välimatka- tai suhdeasteikollisia muuttujia*.
- Poimitaan perusjoukosta  $S$  *yksinkertainen satunnaisotos*.
- Tarkastellaan tekijöiden  $A$  ja  $B$  *vaihtelua* otoksessa.
- Tehdään oletus, että *tekijät  $A$  ja  $B$  ovat riippumattomia*.
- Haluamme *testata* tehtyä *riippumattomuusoletusta*:  
Ovatko havainnot *sopusoinnussa* tehdyn riippumattomuus-oletuksen kanssa?

## $\chi^2$ -riippumattomuustesti: Testausasetelma 2/2

---

- Jos tehty oletus *pätee* ja tekijät *A* ja *B* ovat riippumattomia, tekijöitä *A* ja *B* voidaan tarkastella erillisinä.
- Jos tehty oletus *ei päde* ja tekijät *eivät ole riippumattomia*, tekijät *A* ja *B* ovat **assosioituneita**.
- Riippumattomuusoletuksen testaamiseen tarkoitettuja testejä kutsutaan **riippumattomuustesteiksi**.
- Huomautus:  
Monissa tutkimusasetelmissa toivotaan, että riippumattomuus-oletus *tulee testissä hylätyksi*.

## $\chi^2$ -riippumattomuustesti: Testin suoritus

---

- $\chi^2$ -riippumattomuustestissä havaintojen ja tehdyn riippumattomuusoletuksen yhteensopivuutta mitataan seuraavalla tavalla:
  - (1) Valitaan havainnoille sopivat *luokitukset* tekijöiden *A* ja *B* suhteen.
  - (2) Luokitellaan havainnot tekijöiden *A* ja *B* suhteen *ristiin* ja määrätään *havaitut luokkafrekvenssit*.
  - (3) Määrätään tehdyn riippumattomuusoletuksen mukaiset *odotetut luokkafrekvenssit*.
  - (4) *Verrataan* havaittuja ja odotettuja frekvenssejä toisiinsa  $\chi^2$ -testisuureella.

## $\chi^2$ -riippumattomuustesti: Hypoteesit

---

- *Yleinen hypoteesi  $H$  :*  
Perusjoukosta on poimittu *yksinkertainen satunnaisotos* ja havaintoyksiköt voidaan *luokitella ristiin kahden tekijän  $A$  ja  $B$  suhteen.*
- *Nollahypoteesi  $H_0$  :*  
Tekijät  $A$  ja  $B$  ovat *riippumattomia.*
- *Vaihtoehtoinen hypoteesi  $H_1$  :*  
Tekijät  $A$  ja  $B$  eivät ole *riippumattomia.*

## $\chi^2$ -riippumattomuustesti: Havaitut frekvenssit 1/2

---

- Poimitaan tutkimuksen kohteena olevasta perusjoukosta  $S$  yksinkertainen satunnaisotos, jonka koko on  $n$ .
- *Luokitellaan* havaintoyksiköt tekijän  $A$  suhteen toisensa poissulkeviin luokkiin, joiden lukumäärä on  $r$ .
- *Luokitellaan* havaintoyksiköt tekijän  $B$  suhteen toisensa poissulkeviin luokkiin, joiden lukumäärä on  $c$ .
- *Luokitellaan* havaintoyksiköt tekijöiden  $A$  ja  $B$  suhteen ristiin toisensa poissulkeviin luokkiin, joiden lukumäärä on  $r \times c$ .

## $\chi^2$ -riippumattomuustesti: Havaitut frekvenssit 2/2

---

- Määrätään tekijän  $A$  luokkaan  $i$  kuuluvien havaintojen **havaittu frekvenssi** eli lukumäärä  $R_i$ , kun  $i = 1, 2, \dots, r$ .
- Määrätään tekijän  $B$  luokkaan  $j$  kuuluvien havaintojen **havaittu frekvenssi** eli lukumäärä  $C_j$ , kun  $j = 1, 2, \dots, c$ .
- Määrätään tekijän  $A$  luokkaan  $i$  ja tekijän  $B$  luokkaan  $j$  kuuluvien havaintojen **havaittu frekvenssi** eli *lukumäärä*  $O_{ij}$ , kun  $i = 1, 2, \dots, r$  ja  $j = 1, 2, \dots, c$ .

## $\chi^2$ -riippumattomuustesti:

### Havaittujen frekvenssien taulukko 1/3

- Muodostetaan *havaituista frekvensseistä*  $O_{ij}$   $(r \times c)$ -frekvenssitaulukko  $[O_{ij}]$  :

		<b>B-luokat</b>				
		<b>1</b>	<b>2</b>	<b>...</b>	<b>c</b>	<b>Summa</b>
<b>A-luokat</b>	<b>1</b>	$O_{11}$	$O_{12}$	<b>...</b>	$O_{1c}$	$R_1$
	<b>2</b>	$O_{21}$	$O_{22}$	<b>...</b>	$O_{2c}$	$R_2$
	<b>...</b>	<b>...</b>	<b>...</b>	<b>...</b>	<b>...</b>	<b>...</b>
	<b>r</b>	$O_{r1}$	$O_{r2}$	<b>...</b>	$O_{rc}$	$R_r$
	<b>Summa</b>	$C_1$	$C_2$	<b>...</b>	$C_c$	$n$

## $\chi^2$ -riippumattomuustesti:

### Havaittujen frekvenssien taulukko 2/3

---

- Olkoon  $[O_{ij}]$  havaittujen frekvenssien  $O_{ij}$  muodostama  $(r \times c)$ -frekvenssitaulukko.

$r$  = A-luokkien lukumäärä

$c$  = B-luokkien lukumäärä

$O_{ij}$  = havaittu frekvenssi luokassa, jonka määrää  
A-luokka  $i$  ja B-luokka  $j$ ,

$i = 1, 2, \dots, r, j = 1, 2, \dots, c$

$R_i$  = havaittu frekvenssi A-luokassa  $i$

$C_j$  = havaittu frekvenssi B-luokassa  $j$

$n$  = havaintojen kokonaislukumäärä

- Frekvenssiä  $O_{ij}$  kutsutaan tavallisesti *havaituksi solu-*  
*frekvenssiksi frekvenssitaulukon solussa*  $(i, j)$ .



## $\chi^2$ -riippumattomuustesti:

### Havaittujen frekvenssien taulukko 3/3

---

- *Havaittujen frekvenssien*  $O_{ij}$  frekvenssitaulukossa pätee:

- (i) *Rivisummat* yhtyvät havaittuihin frekvensseihin *A*-luokituksessa:

$$\sum_{j=1}^c O_{ij} = R_i, \quad i = 1, 2, \dots, r$$

- (ii) *Sarakesummat* yhtyvät havaittuihin frekvensseihin *B*-luokituksessa:

$$\sum_{i=1}^r O_{ij} = C_j, \quad j = 1, 2, \dots, c$$

- (iii) *Havaintojen kokonaismäärä*:

$$\sum_{i=1}^r \sum_{j=1}^c O_{ij} = \sum_{i=1}^r R_i = \sum_{j=1}^c C_j = n$$

## $\chi^2$ -riippumattomuustesti: Nollahypoteesin tulkinta

---

- *Jos nollahypoteesi  $H_0$  pätee, havaintojen jakautuminen A-luokkiin ei saa riippua siitä, mihin B-luokkaan havainnot kuuluvat.*

*Jos nollahypoteesi  $H_0$  pätee, todennäköisyys, että havainto kuuluu A-luokkaan  $i = 1, 2, \dots, r$  ei saa riippua siitä, mihin B-luokkaan  $j = 1, 2, \dots, c$  se kuuluu.*

- *Jos nollahypoteesi  $H_0$  pätee, havaintojen jakautuminen B-luokkiin ei saa riippua siitä, mihin A-luokkaan havainnot kuuluvat.*

*Jos nollahypoteesi  $H_0$  pätee, todennäköisyys, että havainto kuuluu B-luokkaan  $j = 1, 2, \dots, c$  ei saa riippua siitä, mihin A-luokkaan  $i = 1, 2, \dots, r$  se kuuluu.*

## $\chi^2$ -riippumattomuustesti: Odotetut frekvenssit 1/5

---

- Olkoon  $x$  on tarkastelun kohteena olevan perusjoukon  $S$  alkio.
- Määritellään seuraavat todennäköisyydet:

$$p_{ij} = \Pr(x \text{ kuuluu } A\text{-luokkaan } i \text{ ja } B\text{-luokkaan } j)$$

$$p_i = \Pr(x \text{ kuuluu } A\text{-luokkaan } i)$$

$$p_j = \Pr(x \text{ kuuluu } B\text{-luokkaan } j)$$

## $\chi^2$ -riippumattomuustesti: Odotetut frekvenssit 2/5

---

- Jos nollahypoteesi  $H_0$  pätee, tapahtumat

$$\{x \in S \mid x \text{ kuuluu } A\text{-luokkaa } i\}$$

$$\{x \in S \mid x \text{ kuuluu } B\text{-luokkaa } j\}$$

ovat riippumattomia kaikille  $i = 1, 2, \dots, r, j = 1, 2, \dots, c$ .

- Siten nollahypoteesi  $H_0$  tekijöiden  $A$  ja  $B$  riippumattomuudesta voidaan ilmaista muodossa

$$H_0 : p_{ij} = p_i p_j, i = 1, 2, \dots, r, j = 1, 2, \dots, c$$

## $\chi^2$ -riippumattomuustesti: Odotetut frekvenssit 3/5

---

- Todennäköisyydet  $p_{ij}$ ,  $p_i$ ,  $p_j$  voidaan estimoida havaituista frekvensseistä  $O_{ij}$  kaavoilla

$$\hat{p}_{ij} = \frac{O_{ij}}{n} \quad \hat{p}_i = \frac{R_i}{n} \quad \hat{p}_j = \frac{C_j}{n}$$

- *Jos nollahypoteesi  $H_0$  pätee, solutodennäköisyydet  $p_{ij}$  voidaan estimoida kaavoilla*

$$P_{ij} = \frac{R_i}{n} \times \frac{C_j}{n} = \hat{p}_i \hat{p}_j, \quad i = 1, 2, \dots, r, \quad j = 1, 2, \dots, c$$

Riippumattomuuden testaaminen

## $\chi^2$ -riippumattomuustesti: Odotetut frekvenssit 4/5

---

- Määrätään *nollahypoteesin*  $H_0$  *pätiessä* **odotetut solu-**  
**frekvenssit**  $E_{ij}$  yhtälöillä

$$E_{ij} = nP_{ij} = \frac{R_i C_j}{n}, i = 1, 2, \dots, r, j = 1, 2, \dots, c$$

## $\chi^2$ -riippumattomuustesti: Odotetut frekvenssit 5/5

---

- Odotettujen solufrekvenssien määritelmästä seuraa, että *odotetut suhteelliset frekvenssit*  $E_{ij}/R_i$  jakautuvat jokaisessa *A-luokassa*  $i$  samalla tavalla *B-luokkiin*  $j = 1, 2, \dots, c$ :

$$\frac{E_{ij}}{R_i} = \frac{C_j}{n}, \quad j = 1, 2, \dots, c, \quad i = 1, 2, \dots, r$$

- Odotettujen solufrekvenssien määritelmästä seuraa, että *odotetut suhteelliset frekvenssit*  $E_{ij}/C_j$  jakautuvat jokaisessa *B-luokassa*  $j$  samalla tavalla *A-luokkiin*  $i = 1, 2, \dots, r$ :

$$\frac{E_{ij}}{C_j} = \frac{R_i}{n}, \quad i = 1, 2, \dots, r, \quad j = 1, 2, \dots, c$$

## $\chi^2$ -riippumattomuustesti:

### Odottettujen frekvenssien taulukko 1/3

- Muodostetaan *odotetuista frekvensseistä*  $E_{ij}$   
 $(r \times c)$ -frekvenssitaulukko  $[E_{ij}]$  :

**B-luokat**

		1	2	...	c	Summa
A-luokat	1	$E_{11}$	$E_{12}$	...	$E_{1c}$	$R_1$
	2	$E_{21}$	$E_{22}$	...	$E_{2c}$	$R_2$
	...	...	...	...	...	...
	r	$E_{r1}$	$E_{r2}$	...	$E_{rc}$	$R_r$
	Summa	$C_1$	$C_2$	...	$C_c$	$n$



## $\chi^2$ -riippumattomuustesti:

### Odotettujen frekvenssien taulukko 2/3

---

- Olkoon  $[E_{ij}]$  odotettujen frekvenssien  $E_{ij}$  muodostama  $(r \times c)$ -frekvenssitaulukko.

$r$  =  $A$ -luokkien lukumäärä

$c$  =  $B$ -luokkien lukumäärä

$E_{ij}$  = odotettu frekvenssi luokassa, jonka määrää  
 $A$ -luokka  $i$  ja  $B$ -luokka  $j$ ,

$i = 1, 2, \dots, r, j = 1, 2, \dots, c$

$R_i$  = havaittu frekvenssi  $A$ -luokassa  $i$

$C_j$  = havaittu frekvenssi  $B$ -luokassa  $j$

$n$  = havaintojen kokonaislukumäärä

- Frekvenssiä  $E_{ij}$  kutsutaan tavallisesti *odotetuksi solu-*  
*frekvenssiksi frekvenssitaulukon solussa  $(i, j)$ .*

## $\chi^2$ -riippumattomuustesti:

### Odottettujen frekvenssien taulukko 3/3

---

- *Odottettujen frekvenssien*  $E_{ij}$  frekvenssitaulukossa pätee:

- (i) *Rivisummat* yhtyvät havaittuihin frekvensseihin *A*-luokituksessa:

$$\sum_{j=1}^c E_{ij} = R_i, \quad i = 1, 2, \dots, r$$

- (ii) *Sarakesummat* yhtyvät havaittuihin frekvensseihin *B*-luokituksessa:

$$\sum_{i=1}^r E_{ij} = C_j, \quad j = 1, 2, \dots, c$$

- (iii) *Havaintojen kokonaismäärä*:

$$\sum_{i=1}^r \sum_{j=1}^c E_{ij} = \sum_{i=1}^r R_i = \sum_{j=1}^c C_j = n$$

## $\chi^2$ -riippumattomuustesti: Testisuure – muoto 1

---

- Määritellään  $\chi^2$ -testisuure

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

jossa

$O_{ij}$  = havaittu frekvenssi solussa  $(i, j)$

$E_{ij}$  = odotettu frekvenssi solussa  $(i, j)$

$r$  = A-luokkien lukumäärä

$c$  = B-luokkien lukumäärä

- Testisuure  $\chi^2$  mittaa havaittujen ja odotettujen frekvenssien jakaumien yhteensopivuutta tai etäisyyttä ja siksi sitä kutsutaan usein  $\chi^2$ -etäisyydeksi.

## Riippumattomuuden testaaminen

# $\chi^2$ -riippumattomuustesti:

### Testisuure – muoto 2

---

- $\chi^2$ -testisuure voidaan kirjoittaa myös muotoon

$$\chi^2 = n \sum_{i=1}^r \sum_{j=1}^c \frac{(\hat{p}_{ij} - P_{ij})^2}{P_{ij}}$$

jossa

$$\hat{p}_{ij} = \frac{O_{ij}}{n}$$

$$P_{ij} = \frac{E_{ij}}{n} = \frac{R_i}{n} \times \frac{C_j}{n} = \hat{p}_i \hat{p}_j$$

$$\hat{p}_i = \frac{R_i}{n} \quad \hat{p}_j = \frac{C_j}{n}$$

$$i = 1, 2, \dots, r, \quad j = 1, 2, \dots, c$$

Riippumattomuuden testaaminen

## $\chi^2$ -riippumattomuustesti:

### Testisuureen asymptoottinen jakauma

---

- *Jos nollahypoteesi  $H_0$  pätee, testisuure  $\chi^2$  noudattaa suurissa otoksissa approksimatiivisesti  $\chi^2$ -jakaumaa vapausastein  $f = (r - 1)(c - 1)$ :*

$$\chi^2 \sim_a \chi^2(f)$$

jossa

$r$  = A-luokkien lukumäärä

$c$  = B-luokkien lukumäärä

## $\chi^2$ -riippumattomuustesti: Jakauma-approksimaation hyvyys

---

- Testisuure  $\chi^2$  noudattaa *suurissa otoksissa approksimatiivisesti*  $\chi^2$ -jakaumaa, jos nollahypoteesi  $H_0$  pätee:

$$\chi^2 \sim_a \chi^2(f), f = (r-1)(c-1)$$

- Approksimaatio on tavallisesti *riittävän hyvä*, jos *odotetut frekvenssit*  $E_{ij}$  ja *keskimääräiset odotetut frekvenssit*  $R_j/c$  ja  $C_j/r$  toteuttavat ehdot

$$E_{ij} > 1, i = 1, 2, \dots, r, j = 1, 2, \dots, c$$

$$R_i/c > 5, i = 1, 2, \dots, r \quad C_j/r > 5, j = 1, 2, \dots, c$$

- Ehdot saadaan toteutumaan valitsemalla havainnoille *sopiva luokitus*.

## $\chi^2$ -riippumattomuustesti:

### Testisuureen normaaliarvo ja testi

---

- Testisuureen  $\chi^2$  normaaliarvo eli odotusarvo nollahypoteesin  $H_0$  pätiessä on

$$E(\chi^2) = f$$

jossa

$$f = (r - 1)(c - 1)$$

- Normaaliarvoaan *merkitsevästi suuremmat*  $\chi^2$ -testisuureen arvot viittaavat siihen, että nollahypoteesi  $H_0$  *ei päde*.
- Normaaliarvoaan *merkitsevästi pienemmät*  $\chi^2$ -testisuureen arvot viittaavat siihen, että nollahypoteesi  $H_0$  *pätee liian hyvin*:

Havainnot saattavat olla *väärennettyjä*.

## $\chi^2$ -riippumattomuustesti: Kommentteja

---

- $\chi^2$ -riippumattomuustesti on **jakaumista riippumaton, ei-parametrinen** testi:
  - (i) Testin yleinen hypoteesi *ei kiinnitä havaintojen jakaumaa*.
  - (ii) Testissä *ei testata todennäköisyysjakauman parametreja koskevaa hypoteesia*, vaan oletusta havaintojen jakaumasta.



## $\chi^2$ -riippumattomuustesti ja $\chi^2$ -homogeenisuustesti

---

- $\chi^2$ -riippumattomuustesti ja edellisessä kappaleessa esitetty  $\chi^2$ -homogeenisuustesti muistuttavat toisiaan.
- Frekvenssitaulukosta *ei voi sellaisenaan nähdä* kummasta testausasetelmasta on kyse.
- $\chi^2$ -riippumattomuustesti ja  $\chi^2$ -homogeenisuustesti tehdään teknisesti täsmälleen *samalla tavalla*:
  - *Odotetut frekvenssit määrätään samalla kaavalla.*
  - *Testisuureet lasketaan samalla kaavalla.*
  - *Testisuureet noudattavat nollahypoteesin pätiessä approksimatiivisesti samaa jakaumaa.*
- Testien testausasetelmat ovat kuitenkin *täysin erilaiset*.

## Riippumattomuustesti vs homogeenisuustesti 1/2

---

- *Riippumattomuustestin* testausasetelma:
  - (i) Testissä tarkastellaan *kahden tekijän A ja B assosiaatiota eli riippuvuutta*, kun havainnot on luokiteltu tekijöiden A ja B suhteen *ristiin*.
  - (ii) Havaintoaineisto muodostuu *yhdestä satunnaisotoksesta*.
  - (iii) Vain havaintojen kokonaislukumäärä  $n$  on *kiinteä eli ei-satunnainen* (valittu) luku, kun taas *sattumamäärää miten havainnot jakautuvat luokkiin tekijöiden A ja B ristiluokituksen suhteen*.

## Riippumattomuustesti vs homogeenisuustesti 2/2

---

- *Homogeenisuustestin* testausasetelma:
  - (i) Perusjoukko koostuu  $r$  ryhmästä ja testissä tarkastellaan perusjoukon alkioiden *jakautumista luokkiin eri ryhmissä*, kun *luokittelu on tehty kaikissa ryhmissä samaa luokitusta käyttäen*.
  - (ii) Havaintoaineisto muodostuu *toisistaan riippumattomista ryhmäkohtaisista satunnais-otoksista*.
  - (iii) Sekä ryhmäkohtaiset otoskoot  $n_i$  että havaintojen kokonaislukumäärä  $n$  ovat *kiinteitä eli ei-satunnaisia* (valittuja) lukuja, kun taas *sattuma määrää miten havainnot jakautuvat luokkiin ryhmien sisällä*.

# Yhteensopivuuden, homogeenisuuden ja riippumattomuuden testaaminen

---

Jakaumaoletuksien testaaminen

Yhteensopivuuden testaaminen

Homogeenisuuden testaaminen

Riippumattomuuden testaaminen

>> Normaalisuuden testaaminen

## Normaalisuusoletuksen tutkiminen

---

- *Normaalijakaumalla on keskeinen asema tilastotieteessä.*
- Esimerkiksi tavanomaisen *t-testin yleisessä hypoteesissa oletetaan, että havainnot noudattavat normaalijakaumaa.*
- Siksi tilastotieteessä on kehitetty useita erilaisia menetelmiä *havaintojen normaalisuuden tutkimiseen.*
- Normaalisuutta voidaan testata edellä esitetyllä  **$\chi^2$ -testillä**, joka on *yleinen yhteensopivuustesti.*
- Seuraavassa tarkastellaan kahta erityisesti normaalisuuden tutkimiseen tarkoitettua menetelmää:
  - **Bowmanin ja Shentonin testi**
  - **Rankit Plot -kuvi**o sekä **Wilkin ja Shapiron testi**

## Bowmanin ja Shentonin testi: Testausasetelma

---

- Olkoot  $\gamma_1$  ja  $\gamma_2$  tavanomaiset keskusmomentteihin perustuvat tunnusluvut todennäköisyysjakaumien *vinoudelle ja huipukkuudelle*.

- *Normaalijakaumalle*

$$\gamma_1 = \gamma_2 = 0$$

- *Bowmanin ja Shentonin testissä* havaintojen normaalisuuden testaaminen perustuu testisuureeseen, joka on *vastaavien otossuureiden funktio*.
- Testisuure saa *suuria arvoja*, jos havaintojen vinous ja/tai huipukkuus *poikkeavat* paljon normaalijakautuneen satunnaismuuttujan vinoudesta ja/tai huipukkuudesta.

## Bowmanin ja Shentonin testi: Satunnaismuuttujan momentit

---

- Satunnaismuuttujan  $X$   **$k$ . origomomentti** on

$$\alpha_k = E(X^k), k = 1, 2, 3, \dots, K$$

- Satunnaismuuttujan  $X$   **$k$ . keskusmomentti** on

$$\mu_k = E[(X - \alpha_1)^k], k = 1, 2, 3, \dots, K$$

- Erityisesti

$$\alpha_1 = E(X) = \mu_X$$

on satunnaismuuttujan  $X$  *odotusarvo* ja

$$\mu_2 = E[(X - \mu_X)^2] = \sigma_X^2$$

on satunnaismuuttujan  $X$  *variانسsi*.

## Bowmanin ja Shentonin testi: Satunnaismuuttujan vinous ja huipukkuus

---

- Tunnuslukua

$$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}}$$

käytetään *todennäköisyysjakauman vinouden mittana*.

- Tunnuslukua

$$\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3$$

käytetään *todennäköisyysjakauman huipukkuuden mittana*.

- *Normaalijakaumalle*

$$\gamma_1 = \gamma_2 = 0$$



## Bowmanin ja Shentonin testi: Havaintojen momentit 1/2

---

- Olkoot  $X_1, X_2, \dots, X_n$  välimatka- tai suhdeasteikollisen satunnaismuuttujan  $X$  havaittuja arvoja.
- Havaintojen  $X_1, X_2, \dots, X_n$   **$k$ . origomomentti** on

$$a_k = \frac{1}{n} \sum_{i=1}^n X_i^k, \quad k = 1, 2, \dots, K$$

- Havaintojen  $X_1, X_2, \dots, X_n$   **$k$ . keskusmomentti** on

$$m_k = \frac{1}{n} \sum_{i=1}^n (X_i - a_1)^k, \quad k = 1, 2, \dots, K$$

## Bowmanin ja Shentonin testi: Havaintojen momentit 2/2

---

- Erityisesti

$$a_1 = \bar{X}$$

on havaintojen  $X_1, X_2, \dots, X_n$  *aritmeettinen keskiarvo* ja

$$m_2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \hat{\sigma}_X^2$$

on havaintojen  $X_1, X_2, \dots, X_n$  *otosvarianssi*.

## Bowmanin ja Shentonin testi: Havaintojen vinous ja huipukkuus

---

- Tunnuslukua

$$c_1 = \frac{m_3}{m_2^{3/2}}$$

käytetään *havaintoarvojen jakauman* **vinouden** *mittana*.

- Tunnusluku

$$c_2 = \frac{m_4}{m_2^2} - 3$$

käytetään *havaintoarvojen jakauman* **huipukkuuden** *mittana*.

## Bowmanin ja Shentonin testi: Hypoteesit

---

- *Yleinen hypoteesi*  $H$  :  
Havainnot  $X_1, X_2, \dots, X_n$  on poimittu yksinkertaisella satunnaisotannalla perusjoukosta  $S$ .
- *Nollahypoteesi*  $H_0$  :  
Havainnot  $X_1, X_2, \dots, X_n$  noudattavat normaalijakaumaa.
- *Vaihtoehtoinen hypoteesi*  $H_1$  :  
Havainnot  $X_1, X_2, \dots, X_n$  eivät noudata normaalijakaumaa.

## Bowmanin ja Shentonin testi: Testisuure ja sen jakauma

---

- Määritellään  $\chi^2$ -testisuure

$$\chi^2 = \frac{n}{6}c_1^2 + \frac{n}{24}c_2^2$$

- *Jos nollahypoteesi  $H_0$  pätee, testisuure  $\chi^2$  noudattaa suurissa otoksissa approksimatiivisesti  $\chi^2$ -jakaumaa 2:lla vapausasteella:*

$$\chi^2 \sim \chi^2(2)$$

## Bowmanin ja Shentonin testi: Testisuure ja sen jakauma

---

- Testisuureen  $\chi^2$  normaaliarvo eli odotusarvo nollahypoteesin  $H_0$  pätiessä on
$$E(\chi^2) = 2$$
- Normaaliarvoaan *merkitsevästi suuremmat*  $\chi^2$ -testisuureen arvot viittaavat siihen, että nollahypoteesi  $H_0$  *ei päde*.
- Normaaliarvoaan *merkitsevästi pienemmät*  $\chi^2$ -testisuureen arvot viittaavat siihen, että nollahypoteesi  $H_0$  *pätee liian hyvin*:  
Havainnot saattavat olla *väärennettyjä*.

## Normaalisuusoletuksen tutkiminen graafisesti

---

- Tietokonegrafiikka mahdollistaa havaintoaineiston normaalisuuden tutkimisen *graafisin keinoin*.
- Normaalisuuden tutkimiseen tarkoitettut graafiset menetelmät perustuvat kuvioihin, joiden ideana on se, että *normaalijakautuneet havainnot asettuvat (satunnaisvaihtelua lukuun ottamatta) suoralle viivalle ja havaintojen epänormaalisuus tulee kuviossa esille poikkeamina tästä suorasta*.
- Tällainen kuvio voidaan muodostaa usealla eri periaatteella; tässä tarkastellaan ns. *Rankit Plot* -kuviota.

## Rankit Plot -kuvio:

### Kuvion idea 1/2

---

- Olkoot

$$Z_1, Z_2, \dots, Z_n$$

havainnot  $X_1, X_2, \dots, X_n$  *suuruusjärjestyksessä* pienimmästä suurimpaan.

- Olkoon  $E(Y_i)$   $i$ . havainnon  $Y_i$  *odotusarvo*, jossa  $Y_i$  on suuruusjärjestyksessä  $i$ . havainto standardoidusta normaalijakaumasta  $N(0,1)$  poimitusta satunnaisotoksesta,  $i = 1, 2, \dots, n$ .

- Piirretään *pistediagrammi*

$$(E(Y_i), Z_i), i = 1, 2, \dots, n$$



## Rankit Plot -kuvio:

### Kuvion idea 2/2

---

- Jos havainnot  $X_1, X_2, \dots, X_n$  ovat *normaalijakautuneen satunnaismuuttujan  $X$  havaittuja arvoja, pisteet*

$$(E(Y_i), Z_i), i = 1, 2, \dots, n$$

*asettuvat (satunnaisvaihtelua lukuun ottamatta) suoralle viivalle.*

- *Poikkeamat* suorasta viittaavat epänormaalisuuteen.
- Kuvioista voidaan tunnistaa:
  - Havaintoarvojen jakauman *vinous*
  - Havaintoarvojen jakauman *huipukkuus*
  - *Poikkeavat havainnot (engl. outliers)*

## Wilkin ja Shapiroon testi: Testin idea

---

- **Wilkin ja Shapiroon testisuure** on *Rankit Plot -kuvion pisteistä*

$$(E(Y_i), Z_i), i = 1, 2, \dots, n$$

lasketun *otoskorrelaatiokertoimen neliö*.

- *Pienet* testisuureen arvot viittaavat siihen, että *normaalisuusoletus ei päde*.
- *Suuret* testisuureen arvot ovat sopusoinnussa *normaalisuusoletuksen kanssa*.
- Testisuureen jakauma on *epästandardi*, mutta monet tilasto-ohjelmistot osaavat laskea Wilkin ja Shapiroon testisuureen arvoa vastaavia *p-arvoja*.

## Normaalisuuden testaaminen

# Wilkin ja Shapiron testi: Esimerkki 1/5

---

- Kappaleessa

### Yhteensopivuuden testaaminen

tarkasteltiin seuraavaa esimerkkiä:

- (i) Kone tekee *ruuveja*, joiden *tavoitepituus* on 10 cm.
- (ii) Ruuvien *pituus vaihtelee satunnaisesti* jonkin verran.
- (iii) Ruuvien pituuksien oletetaan kuitenkin noudattavan *normaalijakaumaa*.

## Normaalisuuden testaaminen

# Wilkin ja Shapiron testi:

### Esimerkki 2/5

---

- Koneen valmistamien ruuvien joukosta poimittiin *yksinkertainen satunnaisotos*, jonka *koko*  $n = 30$  ja otokseen poimittujen ruuvien pituudet mitattiin.
- Ruuvien pituuksien *aritmeettinen keskiarvo* otoksessa oli
$$\bar{X} = 10.09 \text{ cm}$$
ja *otoskeskihajonta* oli
$$s = 0.1038 \text{ cm}$$
- Taulukko oikealla esittää pituuksien *luokiteltua frekvenssi-jakaumaa*.

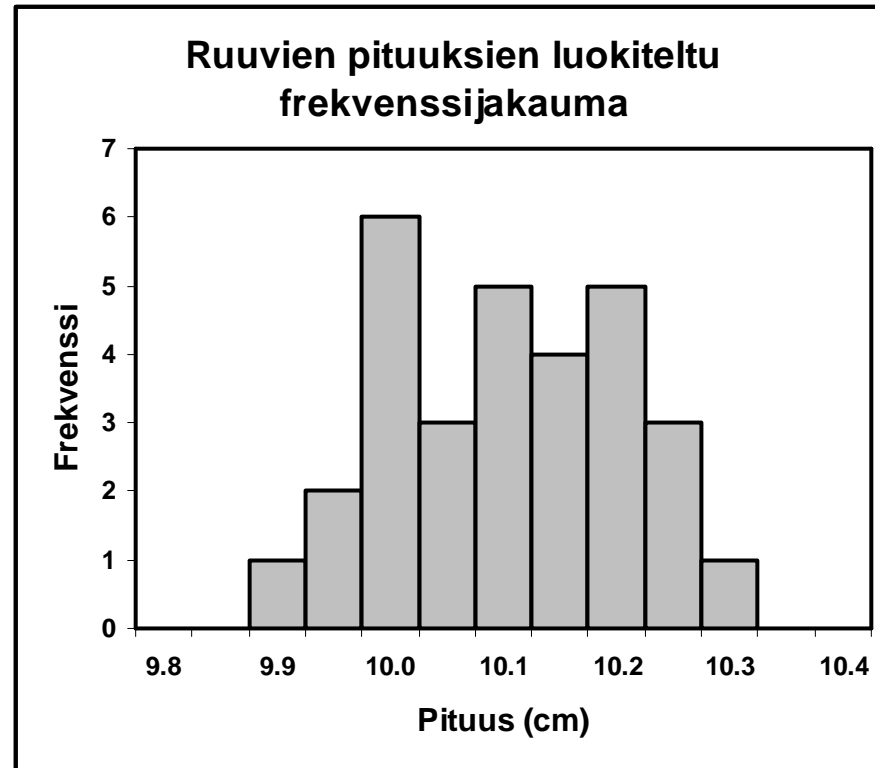
Luokkavälit	Luokkafrekvenssit
(9.85,9.90]	1
(9.90,9.95]	2
(9.95,10.00]	6
(10.00,10.05]	3
(10.05,10.10]	5
(10.10,10.15]	4
(10.15,10.20]	5
(10.20,10.25]	3
(10.25,10.30]	1

## Normaalisuuden testaaminen

# Wilkin ja Shapiron testi:

### Esimerkki 3/5

- Kuva oikealla esittää otokseen poimittujen ruuvien pituuksien luokiteltua frekvenssijakaumaa vastaavaa *histogrammia*.
- Kappaleessa **Yhteensopivuuden testaaminen** todettiin, että *havainnot ovat sopusoinnussa normaalisuusoletuksen kanssa yleisen  $\chi^2$ -yhteensopivuustestin perusteella*.
- Tarkastellaan normaalisuusoletusta vielä *Shapiron ja Wilkin testin* valossa.



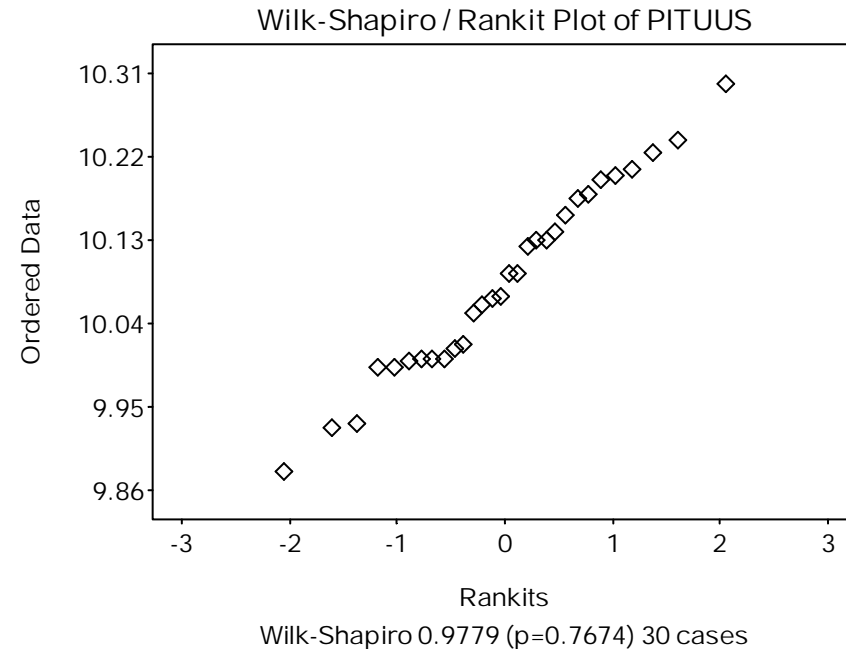
## Normaalisuuden testaaminen

# Wilkin ja Shapiroon testi:

### Esimerkki 4/5

---

- Oikealla on otokseen poimittujen ruuvien pituuksista piirretty *Rankit Plot* -kuvio.
- Havaintoja vastaavien pisteiden poikkeamat suorasta viivasta ovat niin vähäisiä, että *normaalisuusoletusta ei ole mitään syytä asettaa kyseenalaiseksi Rankit Plot -kuvion perusteella.*



## Normaalisuuden testaaminen

# Wilkin ja Shapiroon testi:

### Esimerkki 5/5

- Rankit Plot -kuvioon liittyvä *Wilkin ja Shapiroon testisuureen arvo* on esimerkkiaineiston tapauksessa

0.9779

ja testisuureen arvoa vastaava *p-arvo* on

0.7674

- Siten *normaalisuusoletusta ei ole syytä asettaa kyseenalaiseksi myöskään Wilkin ja Shapiroon testin perusteella.*

