3.1 Havaintoarvot, odotusarvot ja khiin neliön kertymäarvot

Khiin neliön riippumattomuustestillä verrataan toisiinsa kahta muuttujaa tai kahta otosta. Tyypillinen nollahypoteesi tälle testille on, että tutkittavat muuttujat ovat toisistaan riippumattomia (tai että otokset ovat peräisin samasta jakaumasta). Jotta muuttujien jakaumia voitaisiin testata, niiden havaintoaineisto tulee ristiintaulukoida ensin. Katso videosta ”Ristiintaulukointi”, miten ristiintaulukointi tapahtuu EXCEL:issä.

Oletetaan, että meillä on satunnaisotos, jossa on tietoa miesten ja naisten tupakoinnista (keksitty esimerkki). Ristiintaulukoidaan aineisto kuten yllä videossa ja tarkastellaan näin saadun taulukon jakaumia.

	mies	Nainen	summa
ei tupakoi	20	32	52
tupakoi	30	18	48
summa	50	50	100

Tutkimuskysymys voisi olla, riippuuko tupakoiminen sukupuolesta. Nollahypoteesi on, että sukupuoli ja tupakointi ovat toisistaan riippumattomia. Khiin neliö -testillä ei suoraan pysty vastaamaan esimerkiksi kysymykseen, tupakoivatko miehet todennäköisemmin kuin naiset. Sen voimme saada kuitenkin selville arvioimalla asiaa pylväskaavion avulla. Huomaa myös, että tutkimuskysymys ”tupakoivatko miehet useammin kuin naiset” vaatisi toisenlaista aineistoa, esimerkiksi tietoa siitä, miten monta savuketta kukin otokseen osallistuvista henkilöistä polttaa keskimäärin päivässä. Käyttämämme kuvitteellinen aineisto kertoo ainoastaan, tupakoivatko henkilöt ylipäätään vai eivät.

Khiin neliö -testissä lasketaan, millä todennäköisyydellä havaintoarvot pystyttäisiin arvaamaan ristiintaulukoidun taulukon sarakesummien ja rivisummien perusteella eli ns. marginaalien perusteella. Ennen testin tekemistä täytyy siis laskea havaintoarvojen sarakesummat ja rivisummat. Niiden avulla lasketaan odotusarvot jokaiselle ristiintaulukoidun havaintoaineiston solulle. Itse testissä lasketaan, miten paljon havaitut

arvot poikkeavat odotusarvoista. Kullekin ristiintaulukoidun havaintoaineiston solulle lasketaan odotusarvo seuraavalla kaavalla.

$$\frac{\text{sarakesumma}_i * \text{rivisumma}_j}{\text{otoskoko}}$$

Kaavassa alaindeksi i tarkoittaa sarakkeen numeroa ja se saa arvoksi positiivisia kokonaislukuja yhdestä ylöspäin, niin monta saraketta kuin kyseisellä muuttujalla on eri luokkia. Yllä taulukossa on molemmilla muuttujilla kaksi luokkaa (esim. muuttujalla tupakointi on luokat ”tupakoi” ja ”ei tupakoi”) ja siten myös kaksi saraketta, joten i voi saada arvoksi 1 tai 2. Alaindeksi j tarkoittaa rivin numeroa ja se saa arvoksi positiivisia kokonaislukuja yhdestä ylöspäin, niin monta riviä kuin kyseisellä muuttujalla on eri luokkia. Yllä taulukossa on kaksi riviä, joten j voi saada arvoksi 1 tai 2.

Tietylle solulle khiin neliön odotusarvo lasketaan sen sarakkeen sarakesumman ja sen rivin rivisumman perusteella, missä kyseinen solu sijaitsee ristiintaulukoidussa taulukossa. Esimerkiksi odotusarvo sille, että miehet tupakoivat lasketaan kaavalla $50 * 48/100 = 24$ (50 on 1. sarakkeen sarakesumma ja 48 on 2. rivin rivisumma). Eli tämän aineiston perusteella odotusarvo on, että 24 miestä tupakoi. Näin tehdään kaikille soluille ja saadaan seuraavanlaiset odotusarvot neljälle eri solulle.

	mies	nainen	summa
ei tupakoi	26	26	52
tupakoi	24	25	48
summa	50	50	100

Khiin neliö -testissä idea on laskea, miten paljon havaitut arvot (20, 32, 30, 18) poikkeavat näistä odotusarvoista (26, 26, 24, 24). Mitä lähempänä havaitut arvot ovat odotusarvoja, sitä pienempi tämä poikkeama on ja sitä todennäköisemmin nollahypoteesi jää voimaan.

Poikkeama lasketaan laskemalla kullekin solulle khiin neliön (χ^2:n) kertymäarvo. Khiin neliön kertymäarvo lasketaan yksittäiselle solulle havaintoarvojen ja odotusarvojen avulla seuraavasti:

$$\frac{(\text{havaintoarvot-odotusarvot})^2}{\text{odotusarvot}}$$

Esimerkiksi χ^2:n kertymäarvo sille, että miehet tupakoivat saadaan laskemalla $(30-24)^2/24 = 36/24 = 1{,}5$. Sama tehdään kaikille soluille ja näin saadaan seuraavat solukohtaiset kertymäarvot:

	mies	nainen
ei tupakoi	1,385	1,385
tupakoi	1,5	1,5

Koko taulukolle χ^2:n kertymäarvo saadaan laskemalla yhteen kunkin solun kertymäarvo, joka yllä taulukossa olevien lukujen perusteella on pyöristettynä $\chi^2 = 5{,}77$. Koko taulukon kertymäarvo saadaan siis matemaattisella kaavalla:

$$\sum \frac{(havaintoarvo - odotusarvo)^2}{odotusarvo}$$

Katso videosta ”Odotusarvojen laskeminen”, miten odotusarvojen, khiin neliön kertymäarvojen ja merkitsevyyden laskeminen tapahtuu EXCEL:issä.