# Kíkvaðrat ($\chi^2$)

## Fyrirlestur í Aðferðafræði II

# Aldursdreifing í úrtaki

## Er aldursdreifing önnur en í þýði?

Í rannsókn á sjúkdómsótta Íslendinga við sjúkdóma var svarhlutfallið 77,5%. Þar sem sjúkdómar tengjast aldri er mikilvægt að vita hvort brottfallið hafi haft áhrif á aldursdreifinguna.

Taflan sýnir hvernig þátttakendur skiptast á þrjá aldursflokka. Ef allt væri með felldu, ætti skiptingin að vera hlutfallslega sú sama og í þjóðskrá.

Frávik frá aldursskiptingu samkvæmt þjóðskrá getur verið vegna breytileika úrtaka, þ.e. frávik í þessu tiltekna úrtaki en ekki kerfisbundin skekkja í rannsókninni.

| Aldurs flokkur | Fjöldi | Vænti tíðni | Þjóðskrá |
|---|---|---|---|
| 16–29 | 255 | 276 | 35,6% |
| 30–59 | 401 | 383 | 49,4% |
| 60–75 | 119 | 116 | 14,9% |
| Samtals | 775 | 775 | 100,0% |

Fjöldinn í úrtakinu er *rauntíðnin*, þ.e. sá fjöldi sem er í þessu tiltekna úrtaki. Ef úrtakið væri með sömu aldursskiptiningu og þýðið, væri fjöldinn í samræmi við *væntitíðnina*.

Ég vil bera rauntíðni og væntitíðni saman og ákvarða hvort úrtakadreifingin nægi sem skýring á frávikunum.

# Kíkvaðratpróf

## Samræmis rauntíðni væntitíðninni?

Ef úrtakið er nákvæmlega í samræmi við þýðið (þjóðskrá) væri rauntíðnin eins og væntitíðnin. Ég vil vita hvort frávikin séu það mikil að ég verði að gera ráð fyrir að rauntíðnin sé úr öðru þýði.

Ef raun- og væntitíðni eru úr sama þýði, fylgja frávikin kíkvaðratdreifingu. Ég get því prófað frávikin með prófi, kíkvaðratprófi, og reiknað út hversu líkleg þessi frávik væru ef þau eru eingöngu vegna úrtakadreifingar.

Ég þarf að setja fram tilgátur, reikna prófið og meta líkindin á niðurstöðunni.

**Formúla**

$$\chi^2\left(df, N = fjöldi\right) = \sum \frac{\left(f_o - f_e\right)^2}{f_e}$$

**Tilgátur**

$T_1$: Rauntíðnin er úr *öðru* þýði en væntitíðnin

$T_0$: Rauntíðnin er úr *sama* þýði og væntitíðnin

A:8.2, bls. 255–256

# Útreikningur kíkvaðrats

## Samanburður með marktektarprófi

Með því að reikna kíkvaðratprófið er ég að athuga hve mikið rauntíðnin víki frá væntitíðninni. Ef niðurstaðan er há, er rauntíðnin mjög ólík væntitíðninni en ef hún er lág er hún lík væntitíðninni.

Vandinn er sá að ég veit ekki hvað er hátt og hvað er lágt. Ég þarf að vita hve líklegt er að fá þetta háa eða hærri niðurstöðu úr prófinu ef þátttakendur hefðu verið valdir með tilviljunarvali úr þjóðskrá.

Ef niðurstaðan er ólíkleg, dreg ég þá ályktun að rauntíðnin samræmist ekki núlltilgátunni að þátttakendur komi með tilviljunarvali úr þjóðskrá.

Útreikningar

$$\chi^2(2, N = 775) = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$= \frac{(255-276)^2}{276} + \frac{(401-383)^2}{383} + \frac{(119-116)^2}{116}$$

$$= \frac{-21^2}{276} + \frac{18^2}{383} + \frac{3^2}{116} = \frac{441}{276} + \frac{324}{383} + \frac{9}{116}$$

$$= 1{,}598 + 0{,}846 + 0{,}0776 = 2{,}522 \approx 2{,}5$$

Nafn tölfræðiprófsins: $\chi^2$

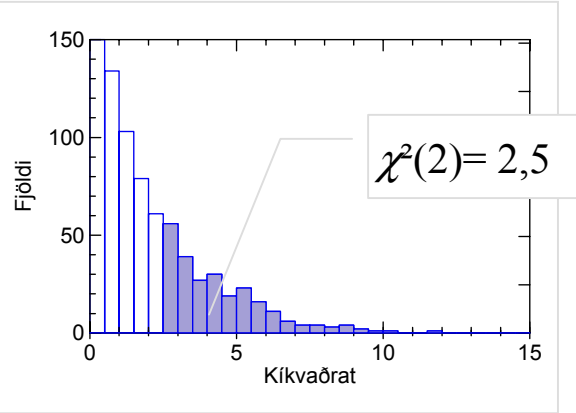
Frígráður: $(2$

Fjöldi þátttakenda: $N = 775)$

Niðurstaða prófsins: $= 2{,}5$

# Monte-Carlo athugun

Monte-Carlo athugun 800 úrtaka

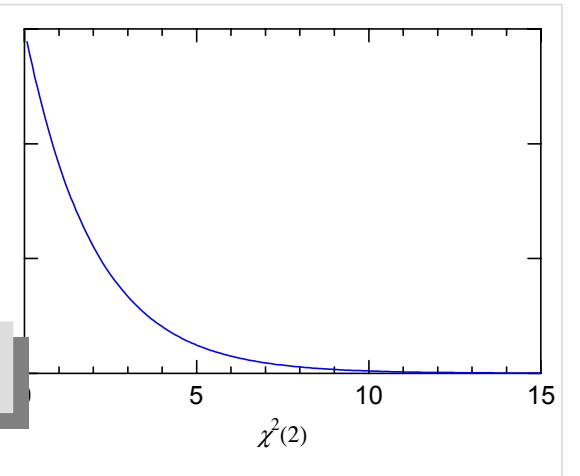



Niðurstaðan er greinilega líkleg undir núlltilgátunni. Ég get því ekki hafnað núlltilgátunni að þrátt fyrir brottfall hópurinn úrtak úr þjóðskrá.

Aldursskiptingin í þessu 775 manna úrtaki samræmist því aldursskiptingu í þjóðskrá.

Þetta þýðir ekki að við *vitum* að úrtakið sé rétt dregið úr þjóðskrá; niðurstaða prófsins samræmist einnig þýði með ýmsum minniháttar frávikum frá aldursskiptingu þjóðskrár.

Fræðidreifing miðað við 2 frígráður



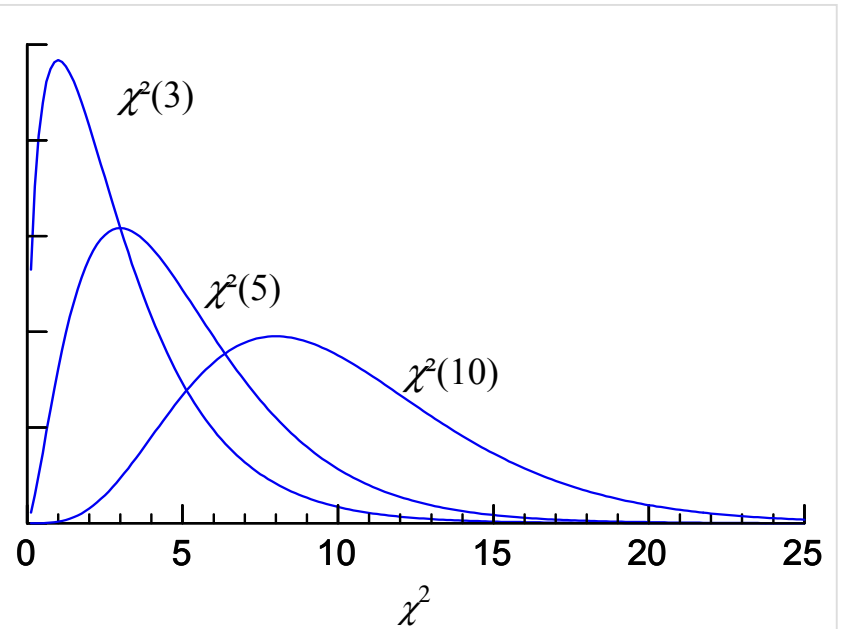

A:8.2, bls. 256–257

# Kíkvaðratdreifing

## Eiginleikar kíkvaðrats

- Á við þegar úrtök eru stór
  - Almennt viðmið að væntitíðni sé aldrei minni en 5
- $\chi^2$ verður aldrei lægra en 0
- Jákvæð skekkja
- Lögun fer eftir frígráðum
  - Samhverfari eftir því sem frígráðum fjölgar
- Meðaltal jafnt frígráðum
  - Vendigildi hækkar með auknum frígráðum

## Ólíkir kíkvaðratferlar




A:8.2, bls. 256–257

# Uppfletting í kíkvaðrattöflu

## Hversu líkleg er niðurstaðan undir núlltilgátunni?

Í okkar tilviki var niðurstaðan mjög líkleg undir núlltilgátunni og því gátum við ekki hafnað henni.

Með því að fletta upp í kíkvaðrattöflu getum við séð líkurnar fyrir niðurstöðu prófsins miðað við gefnar frígráður.

Taflna sýnir að niðurstaða prófsins hefði þurft að vera 5,99 eða hærri til að líkurnar séu 5% eða minni á að frávik raun- og væntitíðni séu jafnmikil eða meiri en þau sem við sjáum í úrtakinu.

| | Líkindi (flatarmál í hægri hala) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Frígráður | 0,25 | 0,2 | 0,15 | 0,1 | 0,05 | 0,025 | 0,02 | 0,01 |
| 1 | 1,32 | 1,64 | 2,07 | 2,71 | 3,84 | 5,02 | 5,41 | 6,63 |
| 2 | 2,77 | 3,22 | 3,79 | 4,61 | 5,99 | 7,38 | 7,82 | 9,21 |
| 3 | 4,11 | 4,64 | 5,32 | 6,25 | 7,81 | 9,35 | 9,84 | 11,34 |
| 4 | 5,39 | 5,99 | 6,74 | 7,78 | 9,49 | 11,14 | 11,67 | 13,28 |
| 5 | 6,63 | 7,29 | 8,12 | 9,24 | 11,07 | 12,83 | 13,39 | 15,09 |
| 6 | 7,84 | 8,56 | 9,45 | 10,64 | 12,59 | 14,45 | 15,03 | 16,81 |
| 7 | 9,04 | 9,80 | 10,75 | 12,02 | 14,07 | 16,01 | 16,62 | 18,48 |
| 8 | 10,22 | 11,03 | 12,03 | 13,36 | 15,51 | 17,53 | 18,17 | 20,09 |
| 9 | 11,39 | 12,24 | 13,29 | 14,68 | 16,92 | 19,02 | 19,68 | 21,67 |
| 10 | 12,55 | 13,44 | 14,53 | 15,99 | 18,31 | 20,48 | 21,16 | 23,21 |
| 11 | 13,70 | 14,63 | 15,77 | 17,28 | 19,68 | 21,92 | 22,62 | 24,73 |
| 12 | 14,85 | 15,81 | 16,99 | 18,55 | 21,03 | 23,34 | 24,05 | 26,22 |
| 13 | 15,98 | 16,98 | 18,20 | 19,81 | 22,36 | 24,74 | 25,47 | 27,69 |
| 14 | 17,12 | 18,15 | 19,41 | 21,06 | 23,68 | 26,12 | 26,87 | 29,14 |
| 15 | 18,25 | 19,31 | 20,60 | 22,31 | 25,00 | 27,49 | 28,26 | 30,58 |
| 16 | 19,37 | 20,47 | 21,79 | 23,54 | 26,30 | 28,85 | 29,63 | 32,00 |
| 17 | 20,49 | 21,61 | 22,98 | 24,77 | 27,59 | 30,19 | 31,00 | 33,41 |
| 18 | 21,60 | 22,76 | 24,16 | 25,99 | 28,87 | 31,53 | 32,35 | 34,81 |
| 19 | 22,72 | 23,90 | 25,33 | 27,20 | 30,14 | 32,85 | 33,69 | 36,19 |
| 20 | 23,83 | 25,04 | 26,50 | 28,41 | 31,41 | 34,17 | 35,02 | 37,57 |
| 21 | 24,93 | 26,17 | 27,66 | 29,62 | 32,67 | 35,48 | 36,34 | 38,93 |
| 22 | 26,04 | 27,30 | 28,82 | 30,81 | 33,92 | 36,78 | 37,66 | 40,29 |
| 23 | 27,14 | 28,43 | 29,98 | 32,01 | 35,17 | 38,08 | 38,97 | 41,64 |
| 24 | 28,24 | 29,55 | 31,13 | 33,20 | 36,42 | 39,36 | 40,27 | 42,98 |
| 25 | 29,34 | 30,68 | 32,28 | 34,38 | 37,65 | 40,65 | 41,57 | 44,31 |
| 26 | 30,43 | 31,79 | 33,43 | 35,56 | 38,89 | 41,92 | 42,86 | 45,64 |
| 27 | 31,53 | 32,91 | 34,57 | 36,74 | 40,11 | 43,19 | 44,14 | 46,96 |
| 28 | 32,62 | 34,03 | 35,71 | 37,92 | 41,34 | 44,46 | 45,42 | 48,28 |
| 29 | 33,71 | 35,14 | 36,85 | 39,09 | 42,56 | 45,72 | 46,69 | 49,59 |
| 30 | 34,80 | 36,25 | 37,99 | 40,26 | 43,77 | 46,98 | 47,96 | 50,89 |
| 31 | 35,89 | 37,36 | 39,12 | 41,42 | 44,99 | 48,23 | 49,23 | 52,19 |
| 32 | 36,97 | 38,47 | 40,26 | 42,58 | 46,19 | 49,48 | 50,49 | 53,49 |
| 33 | 38,06 | 39,57 | 41,39 | 43,75 | 47,40 | 50,73 | 51,74 | 54,78 |
| 34 | 39,14 | 40,68 | 42,51 | 44,90 | 48,60 | 51,97 | 53,00 | 56,06 |
| 35 | 40,22 | 41,78 | 43,64 | 46,06 | 49,80 | 53,20 | 54,24 | 57,34 |

A:4.2, bls. 88–91

# Krosstafla

Taflan sýnir greinileg tengsl kynferðis og námsgreinar; konur eru hlutfallslega algengar í sumum greinum en karlar í öðrum.

Þetta getur verið vegna þess að konur (eða karlar) velji sumar greinar umfram aðrar; þá eru tengsl milli breytanna í þýði.

Þetta getur líka verið *eingöngu* vegna tilviljunarvals kynanna á námsgreinar; þá eru engin tengsl í þýði og breyturnar því *óháðar*.

| | Kyn | | |
|---|---|---|---|
| Námsgrein | Karl | Kona | Samtals |
| Bókasafnsfræði | 4 | 22 | 26 |
| Sálarfræði | 35 | 86 | 121 |
| Uppeldisfræði | 0 | 2 | 2 |
| Námsráðgjöf | 0 | 1 | 1 |
| Félagsfræði | 14 | 40 | 54 |
| Mannfræði | 4 | 8 | 12 |
| Stjórnmálafræði | 27 | 15 | 42 |
| Annað/óskilgreint | 2 | 5 | 7 |
| Samtals | 86 | 179 | 265 |

$T_1$: Breyturnar tengjast í þýði

$T_0$: Breyturnar eru óháðar í þýði

A:8.1, bls. 249–252; A:8.2, bls. 253–254

# Væntitíðni eftir kyni og greinum

Ef núlltilgátan er rétt og breyturnar því óháðar, má nota margföldunarregluna til að reikna þann fjölda sem ætti að vera í hverju hólfi töflunnar fyrir sig.

Taflan sýnir væntitíðni (*expected frequency*), þann fjölda sem má búast við ef breyturnar eru óháðar.

$f_e = \frac{26 \cdot 86}{265} = 8{,}4$

| | Kyn | | |
|---|---|---|---|
| Námsgrein | Karl | Kona | Samtals |
| Bókasafnsfræði | 8,4 | 17,6 | 26 |
| Sálarfræði | 39,3 | 81,7 | 121 |
| Uppeldisfræði | 0,6 | 1,4 | 2 |
| Námsráðgjöf | 0,3 | 0,7 | 1 |
| Félagsfræði | 17,5 | 36,5 | 54 |
| Mannfræði | 3,9 | 8,1 | 12 |
| Stjórnmálafræði | 13,6 | 28,4 | 42 |
| Annað/óskilgreint | 2,3 | 4,7 | 7 |
| Samtals | 86 | 179 | 265 |

$f_e$ $T_r$ $T_c$ $T_t$

Einfaldast er þó að nota formúluna hér til hliðar

**Væntitíðni**

$$f_{e_{rc}} = T_r T_c / T_t$$

A:8.2, bls. 254–255

# Einfölduð tafla

## Fella flokka niður eða saman

Ég kýs að einfalda töfluna þar sem væntitíðnin var undir 5 í mörgum hólfum töflunnar.

Almennt viðmið er að innan við 20% hólfanna sé með væntitíðni undir 5.

Þar sem ég hef mestan áhuga á vinsælustu greinunum felli ég hinar út. Oftast er þó hentugra að fella saman flokka.

Með þessu móti verður væntitíðnin undir 5 í einu hólfi, þ.e. í aðeins 10% hólfanna. Með því móti stenst taflan formkröfur kíkvaðratprófsins.

| | Kyn | | |
|---|---|---|---|
| Námsgrein | Karl | Kona | Samtals |
| Bókasafnsfræði | 4 | 22 | 26 |
| Sálarfræði | 35 | 86 | 121 |
| Félagsfræði | 14 | 40 | 54 |
| Mannfræði | 4 | 8 | 12 |
| Stjórnmálafræði | 27 | 15 | 42 |
| Samtals | 84 | 171 | 255 |

Rauntíðni

| | Kyn | | |
|---|---|---|---|
| Námsgrein | Karl | Kona | Samtals |
| Bókasafnsfræði | 8,6 | 17,4 | 26 |
| Sálarfræði | 39,9 | 81,1 | 121 |
| Félagsfræði | 17,8 | 36,2 | 54 |
| Mannfræði | 4,0 | 8,0 | 12 |
| Stjórnmálafræði | 13,8 | 28,2 | 42 |
| Samtals | 84 | 171 | 255 |

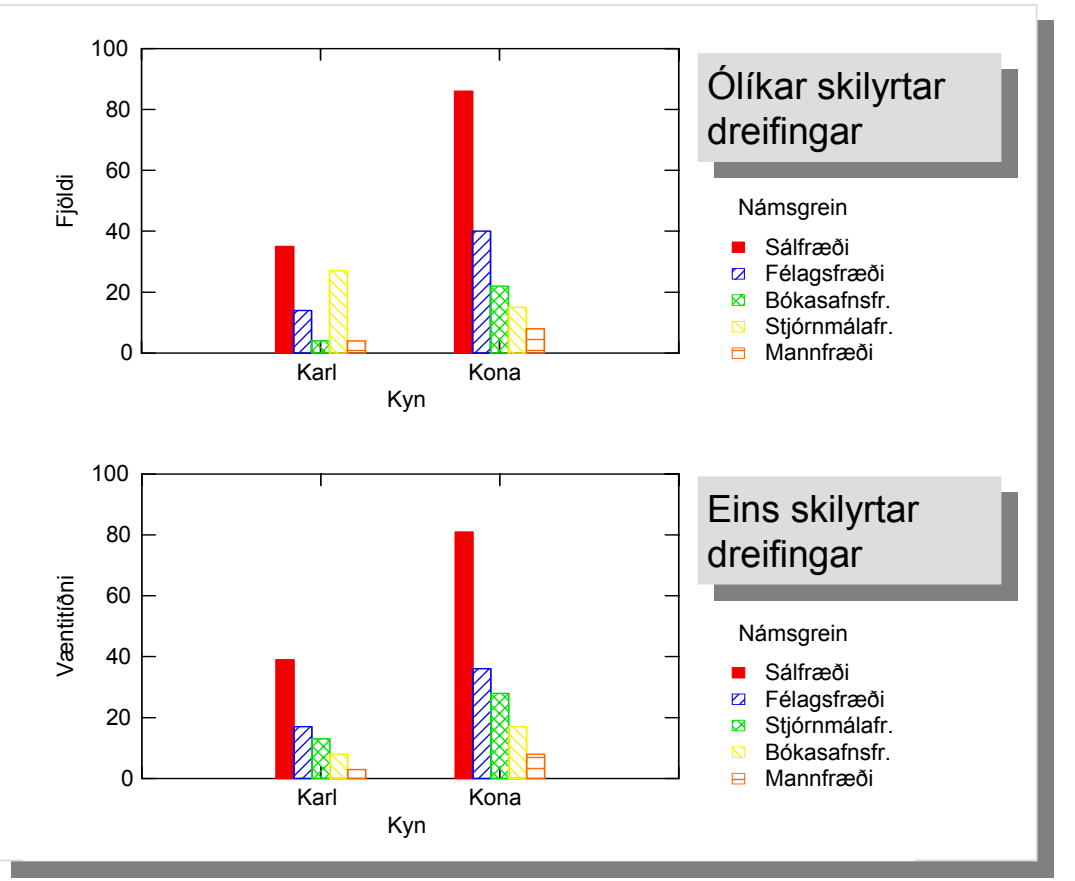
Væntitíðni

# Jaðar- og skilyrt dreifing

Efri myndin sýnir að skilyrtar dreifingar eru ólíkar fyrir karla en konur. T.d. er langflestar konur í sálfræði en hjá körlum eru næstum jafnmargir í stjórnmálafræði.

Neðri myndin sýnir að skilyrtu dreifingarnar eru nákvæmlega eins fyrir væntitíðni; þær eru báðar eins og jaðardreifingin.

Tengsl breyta birtast sem munur skilyrtra dreifinga, það er sem frávik frá væntitíðni.




A:8.1, bls. 250–253; A:8.2, bls. 253–254

# Útreikningur kíkvaðrats

$$df = (c-1)\cdot(r-1)$$

$$\chi^2(4, N=255) = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$= \frac{(4-8{,}56)^2}{8{,}56} + \frac{(22-17{,}44)^2}{17{,}44} + \cdots + \frac{(27-13{,}84)^2}{13{,}84} + \frac{(15-28{,}16)^2}{28{,}16}$$

$$= \frac{20{,}84}{8{,}56} + \frac{20{,}84}{17{,}44} + \cdots + \frac{173{,}31}{13{,}84} + \frac{173{,}31}{28{,}16}$$

$$= 2{,}43 + 1{,}20 + 0{,}59 + 0{,}29 + 0{,}81 + 0{,}4 + 0 + 0 + 12{,}53 + 6{,}15 = 24{,}395$$

Vendigildi

$$\chi^2_{0{,}05}(4) = 9{,}49$$

# Túlkun kíkvaðrats

Niðurstaðan er hærri en vendigildið. Ég hafna því núlltilgátunni miðað við marktektarmörkin $\alpha = 0{,}05$.

Ég dreg því þá ályktun að tengsl séu á milli kynferðis og þeirri námsgrein sem er aðalgrein viðkomandi.

Eðlilegast er að líta svo á að kynferði sé orsakabreytan og því ráði kynferði vali á námsgrein.

Efnisleg túlkun töflunnar fæst með því að

- Skoða töfluna og sjá mynstrið
  - T.d. skoða í hvaða hólfum konur eða karlar eru hlutfallslega flest eða fæst
- Skoða mismun raun- og væntitíðni
  - Athuga hvort hólf þar sem tíðnin er miklu hærri eða lægri en væntitíðnin
- Skoða leiðrétta leif (*adjusted residual*)
  - Túlka má hana á kvarða staðalvillu; *tölugildi* sem er hærra en 2 gefur til kynna litlar (<5%) líkur undir $T_0$

A:8.3, bls. 261–262

# Niðurstöður úr SPSS

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) |
|---|---|---|---|
| Pearson Chi-Square | 24,395[a] | 4 | ,000 |
| Likelihood Ratio | 23,504 | 4 | ,000 |
| Linear-by-Linear Association | 13,805 | 1 | ,000 |
| N of Valid Cases | 255 | | |

a. 1 cells (10,0%) have expected count less than 5. The minimum expected count is 3,95.

**Námsgrein * Kyn Crosstabulation**

| | | | Kyn | | |
|---|---|---|---|---|---|
| | | | Karl | Kona | Total |
| Námsgrein | Bókasafnsfræði | Count | 4 | 22 | 26 |
| | | Expected Count | 8,6 | 17,4 | 26,0 |
| | | Adjusted Residual | -2,0 | 2,0 | |
| | Sálarfræði | Count | 35 | 86 | 121 |
| | | Expected Count | 39,9 | 81,1 | 121,0 |
| | | Adjusted Residual | -1,3 | 1,3 | |
| | Félagsfræði | Count | 14 | 40 | 54 |
| | | Expected Count | 17,8 | 36,2 | 54,0 |
| | | Adjusted Residual | -1,2 | 1,2 | |
| | Mannfræði | Count | 4 | 8 | 12 |
| | | Expected Count | 4,0 | 8,0 | 12,0 |
| | | Adjusted Residual | ,0 | ,0 | |
| | Stjórnmálafræði | Count | 27 | 15 | 42 |
| | | Expected Count | 13,8 | 28,2 | 42,0 |
| | | Adjusted Residual | 4,7 | -4,7 | |
| Total | | Count | 84 | 171 | 255 |
| | | Expected Count | 84,0 | 171,0 | 255,0 |

# Forsendur kíkvaðrats

- Óháðar mælingar
- Mælingar byggjast á tíðni
- Væntitíðni 1,0 eða hærri í öllum hólfum
- Væntitíðni undir 5 í mest 20% hólfanna

A:8.2, bls. 258–260

# Krosstöflur með lága væntitíðni

- Ef taflan er 2x2 má nota Fisher's Exact Test
  - SPSS birtir það ef N<20
- Fella brott eða fella saman flokka
  - Getur verið vandasamt
  - Tengslin geta breyst
  - Myndast geta merkingarlausir flokkar

A:8.3, bls. 264–265; A:7.3, bls. 224–225

# 2x2 töflur í SPSS

Þetta er tafla frá því snemma í námskeiðinu.

Leiðrétt leif (*adjusted residual*) sýnir að óvenjumargir drengir (og óvenjufáar stúlkur) þekkja ekki heimilisfang sitt.

Taflan stenst ekki forsendur kíkvaðrats þar sem tvö hólf af fjórum hafa væntitíðni undir 5.

**KYN * Þekking á heimilisfangi Crosstabulation**

| | | | Þekking á heimilisfangi | | |
|---|---|---|---|---|---|
| | | | Þekkir ekki | Þekkir | Total |
| KYN | Drengir | Count | 5 | 10 | 15 |
| | | Expected Count | 3,1 | 11,9 | 15,0 |
| | | Adjusted Residual | 1,7 | -1,7 | |
| | Stúlkur | Count | 1 | 13 | 14 |
| | | Expected Count | 2,9 | 11,1 | 14,0 |
| | | Adjusted Residual | -1,7 | 1,7 | |
| Total | | Count | 6 | 23 | 29 |
| | | Expected Count | 6,0 | 23,0 | 29,0 |

Tölugildi leiðréttrar leifar er sú sama í öllum hólfum 2x2 töflu

A:8.3, bls. 263–264

# Fisher's Exact Test í SPSS

Niðurstöðu kíkvaðrats er ekki að treysta vegna þess að væntigildi eru undir 5

Þetta er leiðrétting Yates (*Yates's correction*); vinsæl leiðrétting vegna lágra væntigilda en ekki alltaf fyllilega nákvæm.

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
|---|---|---|---|---|---|
| Pearson Chi-Square | 3,027[b] | 1 | ,082 | | |
| Continuity Correction[a] | 1,641 | 1 | ,200 | | |
| Likelihood Ratio | 3,269 | 1 | ,071 | | |
| Fisher's Exact Test | | | | ,169 | ,099 |
| Linear-by-Linear Association | 2,923 | 1 | ,087 | | |
| N of Valid Cases | 29 | | | | |

a. Computed only for a 2x2 table

b. 2 cells (50,0%) have expected count less than 5. The minimum expected count is 2,90.

Leitaðu alltaf að þessari athugasemd þegar þú reiknar kíkvaðrat fyrir krosstöflur.

Líkindi undir núlltilgátunni samkvæmt Fisher's Exact Test. Einfaldast er að nota þetta eingöngu sem tvíhliða próf.

A:8.3, bls. 263–265