Ályktanir um hlutföll

Fyrirlestur í Tölfræði II (SÁL203G)

2009-02-23

Hlutföll

Hlutföll samsvara meðaltölum þegar unnið er með tvíkostabreytur. Það eru tveir möguleikar og hlutfallið segir hversu oft annar þeirra verður sem hlutfall af heildarfjöldanum.

Hlutfallið í þýði er táknað með *p*. Oftast þekkjum við það ekki en verðum að meta það í úrtaki.

Ef úrtakið er stórt og hlutfallið ekki of nálægt 0,0 eða 1,0, má gera ráð fyrir því að úrtakadreifing hlutfalls sé normallaga.

$$\hat{p} = \frac{X}{n}$$

X : Fjöldi annars kostsins

n : Heildarfjöldi í úrtaki

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Öryggisbil í stórum úrtökum

Formúlurnar hér til hliðar miðast við normallaga úrtakadreifingu. Oft er miðað við að formúlurnar gildi ef úrtakið er stórt.

Algengt er að miða við að fjöldinn þurfi að vera 10 í minni hópnum. Í 5. útgáfu bókarinnar miða Moore & McCabe við að 15 séu í minni hópnum.

Engin þessara viðmiða eru traust og nálgunin batnar ekki alltaf við stærra úrtak.

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\hat{p} - z^* \cdot SE_{\hat{p}} < p < \hat{p} + z^* \cdot SE_{\hat{p}}$$

eða

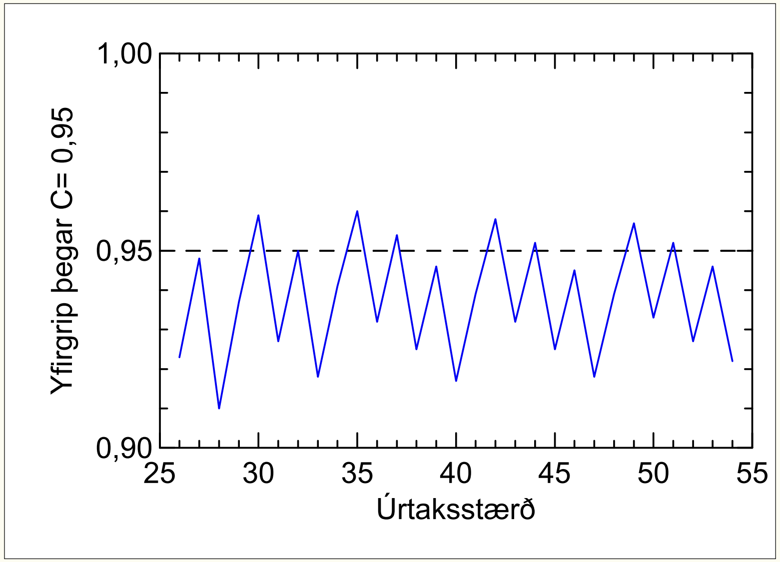
$$\hat{p} \pm z^* \cdot SE_{\hat{p}}$$

Vandi normalnálgunar

Þegar við reiknum 95% öryggisbil viljum við að aðferðin skili okkur bilum sem innihalda þýðihlutfallið í 95% tilvika.

Myndin sýnir hvernig hefðbundin öryggisbil virka í litlum úrtökum þegar p= 0,5. Það eru miklar sveiflur í yfirgripi (*coverage*) öryggisbilanna auk þess sem yfirgripið er of lítið.

Athugið að myndin miðast við ströng skilyrði, þ.e. fjöldinn í fámennari hópnum er að jafnaði meiri en 15.



Leiðrétt öryggisbil

Það er auðvelt að fá betra öryggisbil. Það nægir að bæta fjórum mæligildum við gagnasafnið. Tvö þeirra eru sett í hvorn hóp fyrir sig. Við reiknum nýtt mat á hlutfallinu í þýði sem við setjum inn í formúlurnar.

Við getum annað hvort (a) reiknað þetta í höndunum, (b) bætt fjórum mælingum tímabundið í gagnasafnið eða (c) notað forritlinga á netinu. SPSS býður ekki upp á þessa nálgun.

$$\tilde{p} = \frac{X+2}{n+4}$$

$$SE_{\tilde{p}} = \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$$

$$\tilde{p} - z^* \cdot SE_{\tilde{p}} < p < \tilde{p} + z^* \cdot SE_{\tilde{p}}$$

eða

$$\tilde{p} \pm z^* \cdot SE_{\tilde{p}}$$

Moore og McCabe miða við að þessi nálgun henti þegar $n \geq 10$, þ.e. í öllum nema allra minnstu úrtökunum.

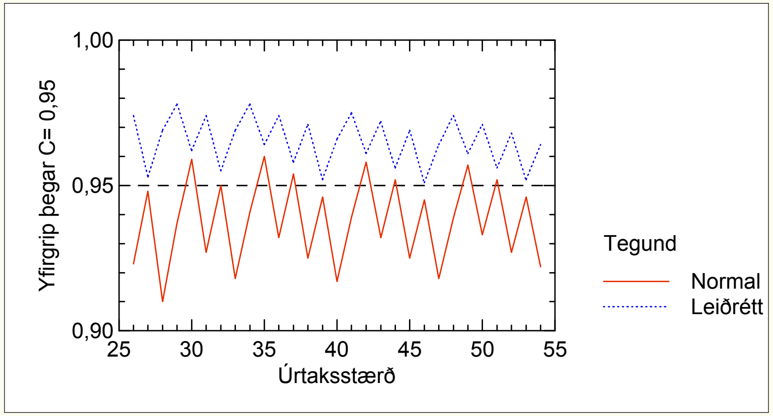
Yfirgrip leiðréttra öryggisbila

Yfirgrip leiðréttu öryggisbilanna eru nær 95% heldur en við normalnálgun. Yfirgrip sveiflast einnig mun minna eftir úrtaksstærð.

Yfirgripið er alltaf meira en 95% á myndinni og að jafnaði nálægt 96%. Mest verður það næstum 98%.

Önnur hlutföll og úrtaksstærðir gefa meiri mun, t.d. fer óleiðrétt bil niður í 0,85 ef úrtaksstærðin er 12.

Leiðréttu öryggisbilin eru ekki fullkomin en mun betri en óleiðrétt.

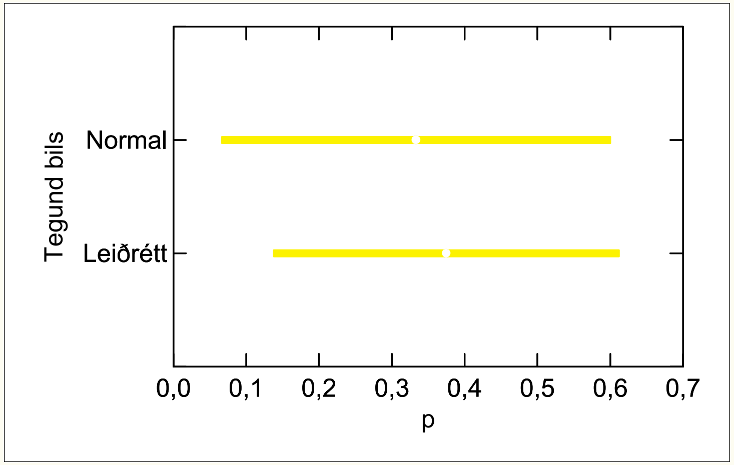


Brown, L.D., Cai, T.T. & DasGupta, A. (2001). Interval estimation for a binominal proportion. *Statistical Science, 16*(2), 101–133.

Tíðni equol hjá konum

Í rannsókn á áhrifum soja kom í ljós að 4 af 12 konum í úrtaki framleiddu efnið equol í líkama sínum við neyslu soja.

Rannsakandinn vildi rannsaka nánar heilsusamleg áhrif soja. Hann vildi því gæta þess að í rannsókninni væru bæði konur sem framleiddu equol og þær sem gerðu það ekki. Fyrsta skrefið í hönnun rannsóknarinnar var að meta hlutfall equol „framleiðenda“ á grundvelli upplýsinganna í þýði.



Við notum forritlinga á netinu til að reikna öryggisbilin.

Normalnálgun gefur öryggisbilið 6,4–59,6% miðað við 95% öryggi. Til samanburðar er leiðrétt öryggisbil 13,7–61,5.

Marktektarpróf á einu hlutfalli

Moore & McCabe mæla með normalnálgun við marktektarprófun eins hlutfalls ef fjöldinn í minni hópnum er meiri en 10.

Marktektarprófið notar viðmiðsgildið í núlltilgátunni til að reikna út staðalvilluna *ekki* úrtakshlutfallið.

Að öðru leyti er þetta eins og venjulegt *z*-próf. Það er einnig hægt að prófa ýmist einhliða eða tvíhliða tilgátur.

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 \cdot (1 - p_0)}{n}}}$$

$$H_0 : p = p_0$$

p_0 er viðmiðsgildið sem núlltilgátan gefur upp. Þetta er mikilvægur munur frá því sem er gert við venjuleg z og t-próf fyrir meðaltöl. Hér er *þýðistalan* samkvæmt núlltilgátunni notuð til að reikna staðalvilluna en ekki úrtakstalan.

Er streitan sambærileg við aðra?

Í veitingahúsakeðju kom í ljós að 68% starfsmanna svöruðu því játandi að streita hefði áhrif á einkalíf þeirra. Til samanburðar svara 75% þessu játandi þegar litið er til allrar (bandarísku) þjóðarinnar.

Fyrirtækið vill vita hvort streitan sé önnur hjá þeirra starfsmönnum en almennt hjá fyrirtækjum. Við reiknum það með *z*-prófi fyrir eitt hlutfall.

Niðurstaðan er ómarktæk. Öryggisbil gefur þó til kynna að streita gæti verið umtalsvert minni en almennt gerist.

$$\hat{p} = 0{,}68 \qquad H_a : p \neq 0{,}75$$

$$n = 100 \qquad H_0 : p = 0{,}75$$

$$z = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0 \cdot (1 - p_0)}{n}}} = \frac{0{,}68 - 0{,}75}{\sqrt{\dfrac{0{,}75 \cdot (1 - 0{,}75)}{100}}}$$

$$= \frac{-0{,}07}{\sqrt{\dfrac{0{,}75 \cdot 0{,}25}{100}}} = \frac{-0{,}07}{\sqrt{0{,}001875}} = \frac{-0{,}07}{0{,}0433}$$

$$= 1{,}62$$

$$p = 0{,}1052 \approx 0{,}11$$

95% leiðrétt öryggisbil: 0,583<p< 0,764

Öryggisbil tveggja hlutfalla

Formúlurnar gefa normalnálgun að öryggisbili fyrir mismun tveggja óháðra hlutfalla.

Óháð hlutföll koma t.d. úr tveimur aðskildum hópum. Hlutföll eru *háð* ef t.d. sömu einstaklingar eru metnir á tveim ólíkum tímum. Hlutfall þeirra sem fylgja ríkisstjórn er augljóslega háð hlutfalli þeirra sem eru á móti.

Ef minnsti undirhópurinn er nógu stór, gildir normalnálgun. Varast ber normalnálgun ef fjöldi er lítill eða eitthvert hlutfall nálægt 0,0 eða 1,0.

$$D = \hat{p}_1 - \hat{p}_2$$

$$SE_D = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$D - z^* SE_D < \mu_D < D + z^* SE_D$$

eða

$$D \pm z^* SE_D$$

Moore & McCabe mæla með þessari aðferð ef a.m.k. 10 stök eru í *minnsta* undirhópnum. Slík ráð hafa þó reynst illa og best að vera mun strangari.

Dottið í það eftir kyni

Rannsókn gefur til kynna að karlar detti fremur í það heldur en konur meðal háskólanema. Til að meta stærð kynjamunarins viljum við reikna öryggisbil. Um 23% af 7.180 karlnemum detta í það en aðeins 17% kvenna.

Bæði fjöldinn og úrtakshlutföllin réttlæta normalnálgun. Niðurstaðan gefur til kynna að kynjamunurinn sé 4,5 til 6,9 prósentustig miðað við 95% öryggi. Ef tekin væru fjöldamörg úrtök, myndu 95% slíkra öryggisbila innihalda þýðismuninn.

$$D = \hat{p}_1 - \hat{p}_2 = 0{,}227 - 0{,}170 = 0{,}057$$

$$SE_D = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$= \sqrt{\frac{0{,}227(1-0{,}227)}{7.180} + \frac{0{,}17(1-0{,}17)}{9.916}}$$

$$= \sqrt{\frac{0{,}175}{7.180} + \frac{0{,}141}{9.916}} = \sqrt{2{,}44 \cdot 10^{-5} + 1{,}42 \cdot 10^{-5}}$$

$$= \sqrt{3{,}86 \cdot 10^{-5}} = 0{,}00622$$

$$D - z^* SE_D < \mu_D < D + z^* SE_D$$

$$0{,}057 - 1{,}96 \cdot 0{,}0062 < \mu_D < 0{,}057 + 1{,}96 \cdot 0{,}0062$$

$$0{,}045 < \mu_D < 0{,}069$$

Leiðrétt öryggisbil tveggja hlutfalla

Normalnálgun verður mjög ónákvæm ef fjöldinn er lítill eða hlutföllin nálægt 0,0 eða 1,0. Almennt ætti ekki að treysta henni nema forsendur séu augljóslega til staðar.

Mun nákvæmara öryggisbil fæst ef einu viðbótarstaki er bætt í hvern undirhóp—fjögur mæligildi í allt eitt af hvorri tegund í hvorn hóp.

Ef normalnálgun á við, fæst sama niðurstaða með leiðréttu öryggisbili og því má í raun nota það *alltaf*.

$$\tilde{D} = \tilde{p}_1 - \tilde{p}_2$$

$$\tilde{p}_1 = \frac{X_1 + 1}{n_1 + 2} \quad \tilde{p}_2 = \frac{X_2 + 1}{n_2 + 2}$$

$$SE_{\tilde{D}} = \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1 + 2} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2 + 2}}$$

$$\tilde{D} - z^* SE_{\tilde{D}} < \mu_{\tilde{D}} < \tilde{D} + z^* SE_{\tilde{D}}$$

eða

$$\tilde{D} \pm z^* SE_{\tilde{D}}$$

Moore & McCabe mæla með þessari aðferð ef úrtaksstærð beggja hópa er a.m.k. 5.

Kynþroski unglinga

Í athugun á kynþroska 12 pilta og 12 stúlkna kom í ljós að 4 piltar og 3 stúlkur náðu Tanner-gildi sem var 4 eða hærra. Hvað segir þetta okkur um mun á kynþroska unglinga?

Hér er augljóslega ekki hægt að nota normalnálgun svo við reiknum leiðrétt öryggisbil.

Greinilega er ekki hægt að fullyrða að kynþroski sé ólíkur eftir kyni því munurinn er á bilinu 27 prósentustig stúlkum í „hag“ til 42 prósentustiga drengjum í hag miðað við 95% öryggi.

$$\tilde{p}_{kk} = \frac{4+1}{12+2} = \frac{5}{14} = 0{,}357$$

$$\tilde{p}_{kvk} = \frac{3+1}{12+2} = \frac{4}{14} = 0{,}286$$

$$\tilde{D} = \tilde{p}_{kk} - \tilde{p}_{kvk} = 0{,}357 - 0{,}286 = 0{,}071$$

$$SE_{\tilde{D}} = \sqrt{\frac{\tilde{p}_{kk}(1-\tilde{p}_{kk})}{n_{kk}+2} + \frac{\tilde{p}_{kvk}(1-\tilde{p}_{kvk})}{n_{kvk}+2}}$$

$$= \sqrt{\frac{0{,}357(0{,}643)}{14} + \frac{0{,}286(0{,}714)}{14}}$$

$$= \sqrt{0{,}0164 + 0{,}0146} = \sqrt{0{,}0301} = 0{,}176$$

$$\tilde{D} - z^* SE_{\tilde{D}} < \mu_D < \tilde{D} + z^* SE_{\tilde{D}}$$

$$0{,}071 - 1{,}96 \cdot 0{,}176 < \mu_D < 0{,}071 + 1{,}96 \cdot 0{,}176$$

$$0{,}071 - 0{,}345 < \mu_D < 0{,}071 + 0{,}345$$

$$-0{,}27 < \mu_D < 0{,}42$$

Marktektarpróf á mismun hlutfalla

Marktektarpróf á mismun hlutfalla byggist á því að að meta hlutfallið sem núlltilgátan gefur sér—við köllum það hér p_0.

Við byrjum á að meta þetta hlutfall á grunni beggja hópanna. Síðan notum við það til að reikna út staðalvilluna.

Mismunur hlutfallanna og staðalvillan fer inn í z-prófið. Niðurstaða þess er borin saman við normaltöflu í venjulegan hátt.

$$\mathrm{H}_0 : p_1 = p_2$$

$\hat{p}_0$: Mat okkar á p undir H_0

$$\hat{p}_0 = \frac{X_1 + X_2}{n_1 + n_2}$$

$$SE_{D_p} = \sqrt{\hat{p}_0 \left(1 - \hat{p}_0\right) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{SE_{D_p}}$$

Þetta er ónákvæm aðferð sem aðeins ætti að nota ef stærð *minni* hópsins er a.m.k. 5 fyrir *bæði* hlutföll.

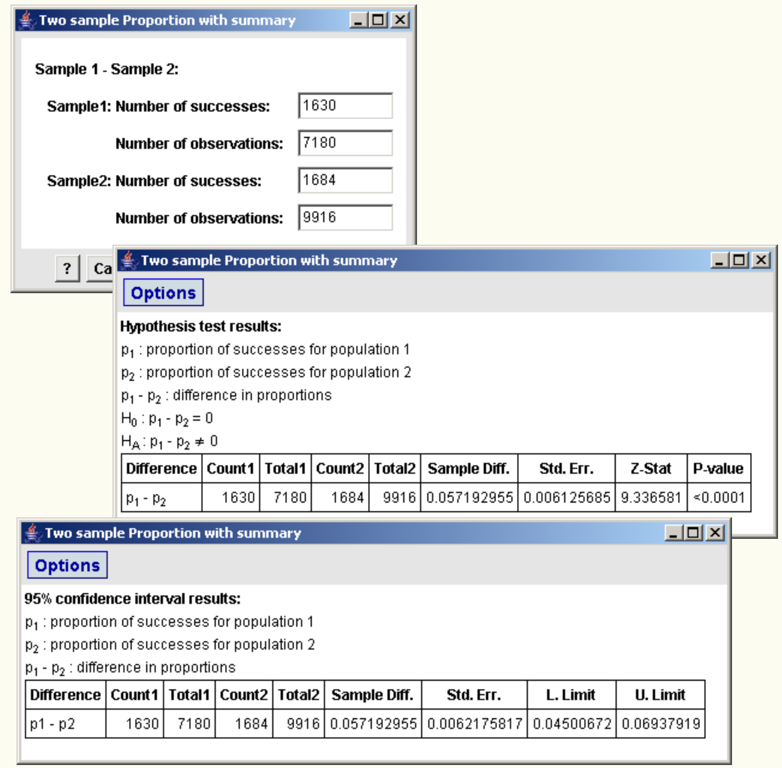
Mismunur hlutfalla í CrunchIt

Í stað þess að reikna þetta í höndum, kýs ég að setja þetta inn í CrunchIt.

Ég slæ ekki inn nein gögn heldur fer á `Stat/ Proportions/ Two sample/ with summary`.

Þar set ég inn bæði fjöldann sem er talinn og heildarfjöldann fyrir hvorn hóp fyrir sig. Smelli svo á `Next`, gæti þess að allt sé rétt skilgreint og smelli svo á `Calculate`.

Niðurstaðan gefur til kynna að konur „detti síður í það“ en karlar, eða a.m.k. viðurkenni það ekki!



Difference	Count1	Total1	Count2	Total2	Sample Diff.	Std. Err.	Z-Stat	P-value
$p_1 - p_2$	1630	7180	1684	9916	0.057192955	0.006125685	9.336581	<0.0001

Difference	Count1	Total1	Count2	Total2	Sample Diff.	Std. Err.	L. Limit	U. Limit
p1 - p2	1630	7180	1684	9916	0.057192955	0.0062175817	0.04500672	0.06937919

Mismunur hlutfalla í VasserStat

Ég get einnig notað vefreikni frá VasserStat.

Hér set ég inn sömu upplýsingar og í CrunchIt og fæ sömu niðurstöður.

Athugaðu að SPSS reiknar ekki hlutföll. Það er hægt að reikna venjulegt *t*-próf í staðinn. Niðurstöður geta orðið lítillega aðrar ef hóparnir eru ekki mjög stórir.