TILASTOTIETEEN JATKOKURSSI. 5+5 OP. 21.1.–6.5.2016. Luennoi yliopistonlehtori Pekka Pere.

2. välikokeen uusintakuulustelun 20.5.2016 ratkaisuehdotukset

1. Tehtävässä määritetään luottamusvälejä ja testataan hypoteesia suhteelliselle osuudelle. Tehtävän mukaan perusteettomien tukipäätösten osuus kaikissa päätöksissä on $\hat{\pi} = 9/152 = 5.9$ %. Peukalosäännöt luottamusvälin normaalisuusapproksimaation toimivuudelle ($n\pi > 10$ ja $n(1-\pi) > 10$ tai $n\pi(1-\pi) \geq 10$) eivät ole voimassa: $n\hat{\pi} = 152 \times 0.059 \approx 8.97 < 10$ ja $n\hat{\pi}(1-\hat{\pi}) = 152 \times 0.059 \times (1 - 0.059) \approx 8.44 < 10$.[1] Approksimaation toimivuuteen kannattaa suhtautua varauksellisesti.

a) Kaksisuuntainen luottamusväli luottamustasolla $1-\alpha$ on

$$\hat{\pi} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}.$$

Tehtävän arvoilla 95 %:n luottamusväli on

$$\frac{9}{152} \pm 1.96 \times \sqrt{\frac{\frac{9}{152} \times \frac{143}{152}}{152}} \approx 0.059 \pm 0.0375$$

eli noin

$$(2.2\ \%,\ 9.7\ \%)\,.$$

Koska on syytä epäillä normaalisuusapproksimaation toimivuutta tehtävän tilanteessa, lasketaan plus neljä -luottamusväli:

$$\frac{n\hat{\pi}+2}{n+4} - 1.960\sqrt{\frac{\frac{n\hat{\pi}+2}{n+4}\left(1-\frac{n\hat{\pi}+2}{n+4}\right)}{n+4}}, \quad \frac{n\hat{\pi}+2}{n+4} + 1.960\sqrt{\frac{\frac{n\hat{\pi}+2}{n+4}\left(1-\frac{n\hat{\pi}+2}{n+4}\right)}{n+4}}$$

Sijoitetaan lukuarvot:

$$\frac{9+2}{152+4} - 1.960\sqrt{\frac{\frac{9+2}{152+4}\left(1-\frac{9+2}{152+4}\right)}{152+4}} \approx 0.0303$$

ja

$$\frac{9+2}{152+4} - 1.960\sqrt{\frac{\frac{9+2}{152+4}\left(1-\frac{9+2}{152+4}\right)}{152+4}} \approx 0.1107.$$

[1] Toisaalta peukalosäännöt $\hat{\pi} - 2\sqrt{\hat{\pi}(1-\hat{\pi})/n} = 0.059 - 2\sqrt{0.059 \times (1-0.059)/n} \approx 0.021 \in (0,1)$ ja $\hat{\pi} + 2\sqrt{\hat{\pi}(1-\hat{\pi})/n} = 0.059 + 2\sqrt{0.059 \times (1-0.059)/n} \approx 0.097 \in (0,1)$ toteutuvat.

Luottamusväli poikkeaa edellä lasketusta:

$$(3.0\ \%,\ 11.1\ \%)\,.$$

Molemmat luottamusvälit peittävät 4 %:n osuuden.

b) Yksisuuntaisen luottamusvälin luottamustasolla $1-\alpha$ alaraja on

$$\hat{\pi} - z_{1-\alpha}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}.$$

Tehtävän tilanteessa yksisuuntainen 95 %:n luottamusväli on

$$\frac{9}{152} - 1.645 \times \sqrt{\frac{\left(\frac{9}{152}\right)\left(\frac{143}{152}\right)}{152}} \approx 0.059 - 0.031$$

eli noin

$$(2.8\ \%,\ 100\ \%).$$

Luottamusväli peittää 4 %:n osuuden.

c) Testataan 5 %:n riskitasolla nollahypoteesia H_0: $\pi = \pi_0 = 0.04$. Vastahypoteesi on H_1: $\pi \neq \pi_0 = 0.04$. Testisuure

$$\frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

noudattaa approksimatiivisesti standardinormaalijakaumaa, kun nollahypoteesi pätee. Testisuureen kriittiset arvot ovat standardinormaalijakauman 2.5. ja 97.5. persentiilit $\pm$ 1.960.

Testisuure saa arvon

$$\frac{0.059 - 0.040}{\sqrt{\frac{0.04 \times 0.96}{152}}} \approx 1.20.$$

Se on pienempi kuin standardinormaalijakauman 97.5. persentiili 1.960 (`qnorm(0.975)`), joten nollahypoteesi jää voimaan.

d) Testataan 5 %:n riskitasolla nollahypoteesia H_0: $\pi = \pi_0 = 0.04$, kun vastahypoteesi on H_1: $\pi > \pi_0 = 0.04$. Testisuure on sama kuin c)-kohdassa, mutta sen kriittinen arvo on nyt standardinormaalijakauman 95. persentiili 1.645 (`qnorm(0.95)`). Saatu arvo on edelleen pienempi kuin kriittinen arvo ($1.20 < 1.645$), joten nollahypoteesi jää voimaan myös yksisuuntaisessa testauksessa.

e) 95 %:n luottamusvälit ja 5 %:n riskitason testit eivät tarjoa riittäviä todisteita nollahypoteesia vastaan. Luottamusvälit sisältävät nollahypoteesin mukaisen arvon, eivätkä lasketut testisuureet yllä hylkäysalueelle.

Oikeuden argumentti oli, että 95 %:n kaksisuuntaisen luottamusvälin yläraja on 9.7 %. Tämän aineiston perusteella ei siten voida hylätä väittämää, että todellinen virheiden määrä oli ollut sallittu 4 %. Päätös on tilastotieteellisesti perusteltu. Muut laskut edellä tukevat samaa johtopäätöstä. Esimerkiksi nykytietämyksen mukaan luotettavamman plus neljä -luottamusvälin yläraja on 11.1 %. Se tekee oikeuden päätöksestä alkuperäistäkin perustellumman.

2. Testattava nollahypoteesi on H_0: $\mu_1 = \mu_2$. Sijoitetaan arvot $\hat{\mu}_1 = 14.4$, $\hat{\mu}_2 = 16.5$, $s_1 = 2.28$, $s_2 = 2.45$ ja $n_1 = n_2 = 16$ testisuureen kaavaan:

$$\frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = -2.51.$$

Welchin eli Satterthwaiten t-testiä käytettäessä testisuureen vapausasteet lasketaan kaavalla:

$$\begin{aligned}
\nu &= \mathsf{int}\left[\frac{\left(\dfrac{s_1^2}{n_1}+\dfrac{s_2^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{s_1^2}{n_1}\right)^2}{n_1-1}+\dfrac{\left(\dfrac{s_2^2}{n_2}\right)^2}{n_2-1}}\right] \\
&= \mathsf{int}\left[\frac{\left(\dfrac{2.28^2}{16}+\dfrac{2.45^2}{16}\right)^2}{\dfrac{\left(\dfrac{2.28^2}{16}\right)^2}{16-1}+\dfrac{\left(\dfrac{2.45^2}{16}\right)^2}{16-1}}\right] \\
&= \mathsf{int}\,[29.84618] \\
&= 29.
\end{aligned}$$

Kaksisuuntaisessa testauksessa 5 %:n riskitasolla testisuureen kriittiset arvot ovat ± 2.045 (`qt(0.975,29)`). Koska testisuureen arvo $2.51 > 2.045$, nollahypoteesi hylätään 5 %:n riskitasolla kaksisuuntaisessa testauksessa.

3.

a) Yksityisten voittoa tavoittelevien yliopistojen alumneissa näyttää olevan muita vähemmän erityisen hyvin menestyviä. Juuri muita eroja ei vaikuta olevan. Parhaiden yliopistojen alumneissa ei vaikuta olevan mainittavasti enempää erityisen hyvin menestyviä alumneja kuin muiden yliopistojen alumneissa.

b) Ryhmän ”100 parasta” yliopistojen alumnit kuuluvat myös taulukon muihin ryhmiin. He eivät ole muista riippumaton ryhmä. χ^2-yhteensopivuustesti edellyttää, että vertailtavat ryhmät ovat riippumattomia. Oletus ei pidä paikkaansa tehtävässä. Testiä ei voi tehdä.

c) Lasketaan taulukkoon liittyvä χ^2-testisuure. Muodostetaan frekvenssitaulukko:

kokemusten lkm	"erittäin samaa mieltä"	ei "erittäin samaa mieltä"	yhteensä
0	507	6735	7242
1	646	6153	6799
2	680	5232	5912
3	459	3532	3991
4	384	2572	2956
5	276	1350	1626
6	151	736	887
yhteensä	3103	26310	29413

Esimerkiksi 1. rivin frekvenssit on laskettu näin: $0.07000829 \times 7242 = 507$ ja $(1 - 0.07000829) \times 7242 \approx 6735$.

Merkitään n_{i+}:lla i. rivin frekvenssien summaa ja n_{+j}:llä j. sarakkeen frekvenssien summaa ($i = 1 \ldots, 7$ ja $j = 1, 2$). Taulukon solujen odotetut frekvenssit ovat $\mu_{ij} = n_{i+}n_{+j}/n$. Esimerkiksi odotettu frekvenssi solussa (4,2) on $3991 \times 26310/29413 = 3569.959$ (havaittu frekvenssi on 3532). Odotetut frekvenssit ovat:

```
          [,1]      [,2]
[1,] 764.01340 6477.9866
[2,] 717.27797 6081.7220
[3,] 623.70163 5288.2984
[4,] 421.04080 3569.9592
[5,] 311.85081 2644.1492
[6,] 171.53905 1454.4610
[7,]  93.57634  793.4237
```

Ne on laskettu R-ohjelmiston käskyillä

```
count <- matrix(c(507,646,680,459,384,276,151,6735,6153,5232,3532,2572,1350,736),
        nrow=7)
chisq.test(count, correct=F) $expected
```

χ^2-testisuureen arvo on

$$\sum_{i=1}^{4}\sum_{j=1}^{2}\frac{(n_{ij}-\mu_{ij})^2}{\mu_{ij}} = \frac{(507-764.0134)^2}{764.0134} + \cdots + \frac{(736-793.4237)^2}{793.4237} \approx 243.25.$$

Merkitään sarakkeiden ja rivien lukumäärä I:llä ja J:llä. Vapausasteet ovat $(I-1) \times (J-1)$ eli $(7-1) \times (2-1) = 6$. $\chi^2(6)$-jakauman kriittinen arvo 0.1 %:n riskitasolla on 22.458 (`qchisq(0.999,6)`). Koska testisuureen arvo (243.25) on suurempi kuin kriittinen arvo (22.458), nollahypoteesi hylätään 0.1 %:n riskitasolla.

Testisuureen p-arvo on viidentoista desimaalin tarkkuudella 0. P-arvo tulostuu automaattisesti `chisq.test`-käskyllä edellä, jos siitä poistetaan lopusta ohje `$expected`.

Erityisen hyvin kaikilla hyvinvoinnin ulottuvuuksilla menestyvien alumnien osuudet vaihtelevat sen mukaan, kuinka moneen kokemusten lukumäärään (lkm) he vastasivat ”erittäin samaa mieltä”. Mitä useampaan kokemukseen he vastasivat ”erittäin samaa mieltä”, sitä suurempi osuus heistä menestyy erittäin hyvin.

4.

a) Mallin tulkintaa:

- Jos lakosta aiheutuneet taloudelliset menetykset ovat 0 euroa, niin ammattiyhdistyksen sakon odotusarvo on mallin vakio eli 2441.80 euroa.
- Taloudellisen menetyksen kasvaessa yksiköllä (sadallatuhannella eurolla) kasvaa sakon odotusarvo 40.3 eurolla.
- Selitysosuus määritellään kaavalla

$$\mathsf{R}^2 = 1 - \frac{\mathsf{RNS}}{\mathsf{KNS}} = \frac{\mathsf{KNS} - \mathsf{RNS}}{\mathsf{KNS}},$$

  jossa $\mathsf{RNS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ on residuaalineliösumma ($\hat{y}_i$ on sovite i. havainnolle) ja $\mathsf{KNS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ on kokonaisneliösumma ($\bar{y}$ on y_i-havaintojen keskiarvo). Selitysosuus on siten osuus vasteen y vaihtelusta (kokonaisneliösummasta), jonka malli pystyy selittämään. Tehtävän aineistolle selitysosuus on 0.182, eli malli selittää 18.2 prosenttia vasteen (sakko) vaihtelusta.

b) Mallin selittäjien — tässä yhden — tilastollista merkitsevyyttä voidaan testata F-testillä. Tehtävän tilanteessa F-testisuure saa arvon

$$\begin{aligned} F_{p,n-p-1} &= \frac{\mathsf{R}^2/p}{(1-\mathsf{R}^2)/(n-p-1)} = \frac{0.182/1}{(1-0.182)/(258-1-1)} \\ &= \frac{0.182}{(1-0.182)/256} = 56.96. \end{aligned}$$

Tehtävän mukaan jäännökset ε_j ovat normaalijakautuneita ja keskenään korreloimattomia. Testisuure noudattaa tällöin F-jakaumaa vapausasteilla 1 ja 256.

R-ohjelmiston käsky `qf(0.99,1,256)` tuottaa tarkan kriittisen arvon 6.735. F-jakauman taulukoista yleensä ei löydy kriittistä arvoa juuri näille vapausasteille mutta löytyy vapausasteille 1 ja 200. Vapausasteilla 1 ja 200 kriittinen arvo 1 %:n riskitasolla on 6.763 (`qf(0.99,1,200)`). Nämä kriittiset arvot ovat hyvin lähellä toisiaan.

Nollahypoteesi hylätään selvästi 1 %:n riskitasolla kumpaa tahansa kriittistä arvoa käyttäen: $56.96 > 6.763 > 6.735$. Aiheutetun taloudellisen tappion suuruus kasvattaa ammattiyhdistykselle määrättyä sakkoa.

c) Yhden selittäjän tilanteessa pätee $\mathsf{R}^2 = r^2$ eli $r = \pm\sqrt{\mathsf{R}}$, jossa r on otoskorrelaatio. Tehtävässä

$$r = \pm\sqrt{0.182} \approx \pm 0.427.$$

Kuvion perusteella on selvää, että otoskorrelaatio on positiivinen, joten $r \approx 0.427$. Laittomista lakoista aiheutuneiden taloudellisten menetysten ja sakkojen otoskorrelaatio on 0.427.

Ylimääräistä kommentointia. Malli ei ilmeisesti toteuta regressiomallin oletuksia: Sakko ei voi olla negatiivinen eikä siksi mallin jäännöskään negatiivinen "kovin paljoa". Kuvio antaa vaikutelman, että hyvin suuret positiiviset jäännökset ovat mahdollisia mutta itseisarvoltaan suuret negatiiviset eivät ja eritoten niin, että jälkimmäisenlaiset eivät ole mahdollisia, kun selittävä muuttuja on arvoltaan pieni. Jäännöksen varianssi ei ilmeisesti ole vakio (riippumaton selittävän muuttujan arvosta.) Estimoitu regressiomalli voi silti olla "kohtuullinen" kuvaus sakkojen ja taloudellisen vahingoon yhteydestä. Epäselväksi jää, miten oletusten rikkoontuminen vaikuttaa estimaattoreiden ja testisuureiden tilastollisiin ominaisuuksiin (testien kokoon jne.).

5*. Opastuksen mukaan testistä voitaisiin päätellä todennäköisyys erojen olemassaololle perusjoukossa eli vastahypoteesille. Frekventistisessä tilastotieteessä

- ei arvioida hypoteesien todennäköisyyksiä.
- nollahypoteesi oletetaan todeksi, ja sen pätiessä lasketaan todennäköisyyksiä testisuureen arvoille.

Opastus on väärä.