Johdatus tilastolliseen päättelyyn, kevät 2015
Harjoitus 5 (21.–24. 4.)

1. Luennolla ja monisteessa (erityisesti jaksoissa 6.4 ja 6.9) on selostettu p-arvon ja siihen liittyvien johtopäätösten tekoa sekä oikeaa tulkintaa.

a) Eräs suomalainen oppikirja[1] kuvailee asiaa näin: *Hypoteesi ei koskaan jää voimaan absoluuttisesti, vaan se jää voimaan jollain todennäköisyydellä. P-arvot ovat yksinkertaisia todennäköisyyslukuja, jotka vaihtelevat välillä* $[0, 1]$ *kuten klassinen todennäköisyyskin. P-arvot ilmoittavat, kuinka suurella todennäköisyydellä vaihtoehtoinen hypoteesi on väärä. Mitä lähempänä p-arvo on ykköstä, sitä suuremmalla todennäköisyydellä nollahypoteesi on asetettu oikein. Jos p-arvo on taas lähellä nollaa, vaihtoehtoinen hypoteesi on erittäin todennäköisesti oikea. Koska p-arvot ovat todennäköisyyksiä, niitä voidaan ajatella myös prosentteina. Toisin sanoen* $p = .5$ *tarkoittaa* 50 %*:n todennäköisyyttä.*

Millaisia virheitä ja/tai puutteita tässä kuvauksessa on?

b) Eräs englanninkielinen oppikirja[2] selostaa näin: *Most studies require very small p-values, such as* $p \leq 0.05$*, in order to reject* H_0*. In such cases, the results are said to be significant at the* 0.05 *level. –– Making a decision by rejecting or not rejecting a null hypothesis is an optional part of the significance test. –– the study should interpret the p-value in context. The smaller p is, the stronger the evidence against* H_0 *and in favor of* H_1*. –– Why do smaller p-values indicate stronger evidence against* H_0*? Because the data would then be more unusual if* H_0 *were true. –– In practise, it is sometimes necessary to decide whether the evidence against* H_0 *is strong enough to reject it. The decision is based on whether the p-value falls below a prespecified cutoff point. –– The* α*-level is a number such that we reject* H_0 *if the p-value is less than or equal to it. The* α*-level is called the significance level.*

Onko tämä kuvaus mielestäsi oikein ja yhtäpitävä kurssilla oppimasi kanssa?

Seuraavat kaksi tehtävää liittyvät lineaariseen regressiomalliin (ks. monisteen jaksot 9.1–9.3).

2. a) Näytä, että otoksista $\mathbf{x} = (x_1, \ldots, x_n)$ ja $\mathbf{y} = (y_1, \ldots, y_n)$ laskettu otoskovarianssi

$$s(\mathbf{x}, \mathbf{y}) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

voidaan lausua myös seuraavissa muodoissa:

$$s(\mathbf{x}, \mathbf{y}) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}) y_i = \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \right). \qquad (*)$$

b) Näytä, että pistepareihin (x_i, y_i) sovitetun PNS-suoran yhtälö $y = a + bx$ (ks. monisteen sivu 107) voidaan esittää myös muodossa

$$\frac{y - \bar{y}}{s_{\mathbf{y}}} = r(\mathbf{x}, \mathbf{y}) \frac{x - \bar{x}}{s_{\mathbf{x}}},$$

jossa $s_{\mathbf{x}} = \sqrt{s(\mathbf{x}, \mathbf{x})}$ ja $s_{\mathbf{y}} = \sqrt{s(\mathbf{y}, \mathbf{y})}$ ovat otoskeskihajonnat ja $r(\mathbf{x}, \mathbf{y})$ on eräs luvuista x_i ja y_i laskettu kerroin. Mikä on $r(\mathbf{x}, \mathbf{y})$:n lauseke? Entä mitä nimitystä siitä käytetään? (Jos et ole sitä aikaisemmin nähnyt, selvitä asia kirjallisuudesta!)

---

[1] L. Nummenmaa: *Käyttäytymistieteiden tilastolliset menetelmät*, Tammi, 2009. Sivut 148–149.

[2] A. Agresti ja B. Finlay: *Statistical Methods for the Social Sciences*, 4. laitos, Pearson, Lontoo, 2008.

Huom. Kohdan (b) esityksessä on suoritettu eräänlainen standardointi sekä x- että y-havaintoihin: keskiarvon vähentäminen merkitsee keskistämistä ja keskihajonnalla jakaminen skaalausta siten, että muunnettujen havaintojen keskihajonnaksi tulee 1.

3. JTP-kurssilla 2012 tehdyssä kyselytutkimuksessa (otoskoko $n = 37$) saatiin miesopiskelijoiden pituudelle x (cm) ja painolle y (kg) seuraavat yhteenvedot:

$$\bar{x} = 180.1, \qquad \bar{y} = 77.07$$
$$s(\mathbf{x}, \mathbf{x}) = 56.52, \qquad s(\mathbf{x}, \mathbf{y}) = 46.37, \qquad s(\mathbf{y}, \mathbf{y}) = 184.56.$$

Mallinnetaan aineisto lineaarisen regressiomallin avulla. Laske mallin kertomien SU-estimaatit (eli PNS-estimaatit) ja ilmoita PNS-suoran yhtälö.

4. Kertaustehtävä suurimman uskottavuuden menetelmästä. Kallen mökkimatkaan kuuluu 100 km:n soratieosuus, jonka aikana hänen autonsa tuulilasiin usein tulee kiveniskemiä. Neljällä eri ajokerralla Kalle havaitsee kiveniskemien lukumääriksi

$$y_1 = 1,\ y_2 = 3,\ y_3 = 0,\ y_4 = 2,$$

JTN-kurssilla (harjoituksen 6 tehtävä 1) oppimansa perusteella Kalle mallintaa näitä lukumääriä riippumattomina havaintoina Poissonin jakaumasta Poisson(μ), jonka pistetodennäköisyysfunktio on $g(x; \mu) = e^{-\mu}\mu^x/x!$, kun $x = 0, 1, 2, \ldots$. Muodosta Kallen havaintoja vastaava uskottavuusfunktio ja johda sen logaritmia tutkimalla suurimman uskottavuuden estimaatti $\hat{\mu}$. Mikä on parametrin μ tulkinta?

5. Bayesin kaavan mieliinpalautus viimeistä viikkoa varten. Vuonna 1975 uutisoitiin tutkimuksesta, jonka mukaan 50 % kanadalaisista miehistä käytti värillisiä (= muita kuin valkoisia) alushousuja kun taas amerikkalaisista miehistä sellaisia käytti vain 20 %. Bermudalaisen hotellin asiakaskunta koostui yksinomaan amerikkalaisista ja kanadalaisista siten, että miesasiakkaista 80 % oli amerikkalaisia ja 20 % kanadalaisia. Todennäköisyyslaskentaa opiskellut siivooja huomasi miesasiakkaan huoneessa punaiset alushousut. Millä todennäköisyydellä hän päätteli asiakkaan olevan kanadalaisen? Muotoile sopivat tapahtumat ja sovella niihin Bayesin kaavaa.