Huomioita tehtävän 2 raporteista

Yleisesti tämän vuoden kurssilaiset ovat suoriutuneet melko hyvin tehtävä 2 kanssa. Muutama seikka kuitenkin on aiheuttanut päänvaivaa.

* On aina hyvä, ainakin itsensä takia, tehdä analysoitavasta datasta kuva. Silloin saattaa huomata analyysin selkeät virheet. Tällä kertaa joku oli harmillisesti tulkinnut t-testin p-arvot täsmälleen väärinpäin, löytäen eroa datasarjojen välillä juuri silloin, kun sitä ei ollut. Jos olisi ollut datasta kuva, tästä olisi voinut välttyä.
* Tosiaan, tässä harjoituksessa, käytetyissä t- ja f-testeissä pieni p-arvo johtaa nollahypoteesin (= ei eroa) hylkäämiseen. Voi tuntua nurinkuriselta tehdä päättely nollahypoteesin kautta, kun oikeasti olemme kiinnostuneita siitä, onko eroa, mutta siihen on selkeä syy. Nollahypoteesi on mahdollisimman yksinkertainen; eroa on mittaustarkkuuden rajoissa nolla. Sitä vastoin, on vaikeampaa muodostaa yhtä selkeä hypoteesi julistamaan datojen välistä eroa; onko ero yhden, viiden vai neljänkymmenenkahden yksikön verran? On olemassa vaihtoehtoinen tilastotieteen koulukunta, niin sanottu Bayesilainen tilastotiede, jonka menetelmien avulla saadaan todennäköisyysjakauma datojen väliselle erolle, mutta siinäkin on omat ongelmansa, se vaatii nimittäin a priori jakauman. Toisin sanoen ennen kun mitään mittausdataa on olemassakaan pitää olla alkuarvaus tulevien datojen väliselle erolle jota sitten modifioidaan mittausten myötä.
* Excelissä on olemassa kolme eri t-testin muunnelmaa: t-testi parivertailulle, testi kun varianssit ovet samat ja testi kun varianssit eroavat. Parillinen testi pitää käyttää ainoastaan silloin kun jokaisella ensimmäisen datan pisteellä on yksiselitteinen parinsa toisessa datassa. Esimerkiksi, jos testataan lääkkeen vaikutusta ja meillä on kymmenen potilasta, joista haluamme selvittää onko potilaiden terveydentila parempi lääkityksen jälkeen, mittaamme vaikkapa jotain veriarvoja sekä ennen lääkitystä että sen jälkeen. Jos vertaamme tuloksia t-testillä on luonnollista yhdistää saman potilaan lääkitystä edeltävät ja sen jälkeiset tulokset käyttämällä parivertailu t-testiä. Jos sen sijaan olisi eri ryhmä potilaita, joista otamme veriarvoja lääkityksen jälkeen ei ole mielekästä käyttää parivertailutestiä. Jos datojen varianssiestimaatit ovat suunnilleen yhtä suuret, voimme antaa Excelin t-testin tyyppiparametrille arvon 2, jos toisen varianssi on selkeästi suurempi, annamme arvon 3. Siinä tapauksessa, että on vaikeaa päättää, onko varianssien ero suuri vai ei, luultavasti eri t-testien tulosten välinen ero on myös pieni.
* Kahden keskiarvon vertailussa kannattaa käyttää t-testiä ja kahden varianssin vertailussa F-testiä. ANOVA (analysis of variance) on myös eräänlainen F-testi, mutta se tutkii ryhmien sisäisen ja välisen varianssien eroa ja näin ollen sitä käytetäänkin useampien keskiarvojen vertailuun. (Voidaan toki myös käyttää kahden keskiarvon vertailuun.) Jotkut opiskelijat olivat oma-aloitteisesti käyttäneet edistyneempiä testejä kuten Levenen testi ja Brown-Forsythen testi tutkimaan useamman datajoukon varianssien eroja.
* Testi voi olla yksisuuntainen tai kaksisuuntainen (one-tailed vs two-tailed). Edellisessä tapauksessa olemme esimerkiksi kiinnostuneita vain siitä, onko ensimmäisen datajoukon keskiarvo (tai varianssi f-testin tapauksessa) suurempi kuin toisen. Meille sen sijaan on yhdentekevää, jos toisen datajoukon keskiarvo on suurempi kuin ensimmäisen. Kaksisuuntaisen testin tapauksessa taas haluamme vain tietää onko datajoukkojen keskiarvojen välillä eroa, riippumatta kumpaan suuntaan. Silloin eron tilastollinen merkitsevyys tarkoittaa sitä, että datojen keskiarvo- ja varianssiestimaateista laskettu t-arvo menee jompaankumpaan t-jakauman hännistä, jonka alle jäävä yhteenlaskettu pinta-ala on sama kuin testin merkitsevyys (esim. 0.05). Jos sitä vastoin käy niin että t-arvo jää häntien ulkopuolelle, voi syntyä houkutus sanoa, että teemmekin yksisuuntaisen testin koska silloin meillä on vain yksi häntä, jonka ala kaksi kertaa isompi kuin kaksisuuntaisen testin toinen häntä. Vaikka näin saattaisimme saada eron näyttämään merkitsevältä, tällainen toiminta olisi kuitenkin epärehellistä, koska olisimme voineet valita myös toisen hännän, jos laskettu t-arvo olisi ollut lähellä sitä ja tällä tavalla tuplata merkitsevyyden todennäköisyys. 



* Vaikka F-testiin liittyvä jakauma ei ole symmetrinen siinäkin tapauksessa voi valita joko yksi- tai kaksisuuntaisen testin. Hieman hämmentävästi Excelissä ftest-funktiolla laskettu testi on kaksisuuntainen, kun taas Data Analysis työkaluista valittu F-testi on yksisuuntainen. Meidän tapauksessamme tulisi valita kaksisuuntainen testi, koska haluamme vain tietää onko varianssien välillä eroa. Tästä ei kuitenkaan ole pisteitä vähennetty.
* Korrelaatio-osiossa piti ottaa kantaa myös siihen, voiko yhdistää datapisteet koivun ja männyn osalta. Yhtäältä, jos molemmissa ryhmissä kahden muuttujan välinen lineaarinen suhde on suunnilleen sama, ryhmien yhdistäminen tuo lisää varmuutta korrelaatiopäätelmiin, toisaalta jos muuttujilla on ryhmissä erilaiset lineaariset suhteet, ryhmien yhdistäminen vähentää korrelaatiota. Korrelaatiolaskijan pitää siis arvioida ovatko ryhmät tarpeeksi samanlaiset, jotta niitä voisi yhdistää.