Harjoitus 11: Tilastollinen testaus (Excel + R)

MS-C2107 Sovelletun matematiikan tietokonetyöt 2021

Harjoituksen aiheita

- Tilastollinen testaus (t-testi ja varianssianalyysi)
  - Tilastolliset hypoteesit
  - Testisuure ja sen tulkitseminen p-arvon avulla
- Satunnaismuuttujan jakauman tutkiminen
  - Onko tutkittava aineisto peräisin normaalijakaumasta?

Oppimistavoitteet

- Osaat tilastollisen testauksen peruskäsitteet

Tilastollinen testaus

- Tilastollisessa testauksessa tutkitaan satunnaisilmiötä tai perusjoukkoa koskevia väitteitä
  - Satunnaisilmiö: prosessi jonka lopputulos vaihtelee satunnaisesti (esim. nopanheitto)
  - Perusjoukko: tutkimuksen kohteena oleva ”suuri” joukko alkioita (esim. *kaikkien* suomalaisten verenpaineet)
- Testausasetelmaan kuuluu yleinen hypoteesi, joka sisältää taustaoletukset tutkittavasta ilmiöstä sekä nollahypoteesi ja vaihtoehtoinen hypoteesi

Yleinen hypoteesi H

- Yleinen hypoteesi H sisältää oletukset
  - perusjoukosta tai satunnaisilmiöstä
    esim. *perusjoukkoon kuuluu kaikki suomalaiset*
  - käytetystä otantamenetelmästä
    esim. *tutkittava otos on satunnaisotos perusjoukosta*
  - perusjoukon tai satunnaisilmiön jakaumasta
    esim. *suomalaisten verenpaineet ovat normaalijakautuneita*
- Yleisen hypoteesin oletuksista pidetään kiinni koko testauksen ajan
- Yleisen hypoteesin sisältämiä jakaumaoletuksia voidaan ja on yleensä syytä testata erikseen

Nollahypoteesi H_0

- Nollahypoteesi on tutkittavaa perusjoukkoa tai ilmiötä koskeva väite, joka merkitään H_0
  - Nollahypoteesi H_0 voidaan osoittaa vääräksi näyttämällä, että vaihtoehtoinen hypoteesi H_1 on ”selvästi” sitä todennäköisempi
  - Usein sanotaan, että nollahypoteesi ”jätetään voimaan”, jos näin ei käy. Huom! Vaikka nollahypoteesi jäisi voimaan, niin se ei todista nollahypoteesia oikeaksi
- Yksinkertaisissa testausasetelmissa nollahypoteesi on muotoa

$$H_0 : \theta = \theta_0,$$

jossa θ on jokin jakauman parametri esim. suomalaisen keskimäärinen verenpaine ja θ_0 mikä tahansa vakio

Vaihtoehtoinen hypoteesi H_1

- Vaihtoehtoinen hypoteesi H_1 on väite johon nollahypoteesia H_0 verrataan
- Jos vaihtoehtoinen hypoteesi on muotoa

$$H_1 : \theta > \theta_0 \qquad \text{tai} \qquad H_1 : \theta < \theta_0,$$

  vaihtoehtoista hypoteesia kutsutaan yksisuuntaiseksi.
- Jos vaihtoehtoinen hypoteesi on muotoa

$$H_1 : \theta \neq \theta_0$$

  vaihtoehtoista hypoteesia kutsutaan kaksisuuntaiseksi.
- Usein sanotaan, että jos nollahypoteesi hylätään, niin vaihtoehtoinen hypoteesi ”astuu voimaan”

Esimerkki hypoteesien valinnasta

- Yleinen luulo on, että suomalaisten keskimääräinen verenpaine on 105, minkä lähistöllä se on historiallisesti ollut.

  Lääkärit epäilevät, että viime vuosina se on kohonnut ja haluavat osoittaa tämän pätevän, mutta eivät halua mitata jokaisen suomalaisen verenpainetta

- Lääkärit muodostavat testausasetelman, jossa he keräävät n. kpl satunnaiselta suomalaiselta verenpaineet:

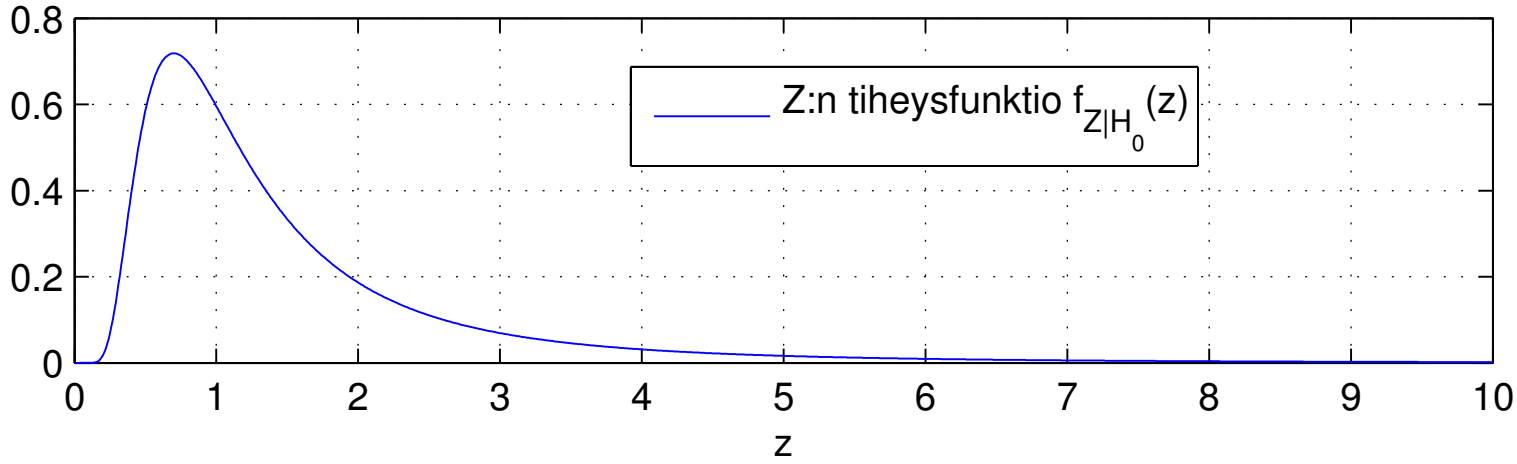
  H: Verenpaineet noudattavat normaalijakaumaa, otos on satunnainen kaikista suomalaisista

  H_0: Keskimääräinen verenpaine $\theta = 105$

  H_1: Keskimääräinen verenpaine $\theta > 105$.

Testisuure Z

- Tilastollinen testi perustuu testisuureeseen Z, joka mittaa havaintojen ja nollahypoteesin H_0 yhteensopivuutta.
- Testisuure on satunnaismuuttuja, jonka arvo riippuu havainnoista ja nollahypoteesista H_0. Niinpä testisuureen tulkitsemiseksi täytyy tuntea testisuureen jakauma sillä ehdolla, että H_0 pätee.

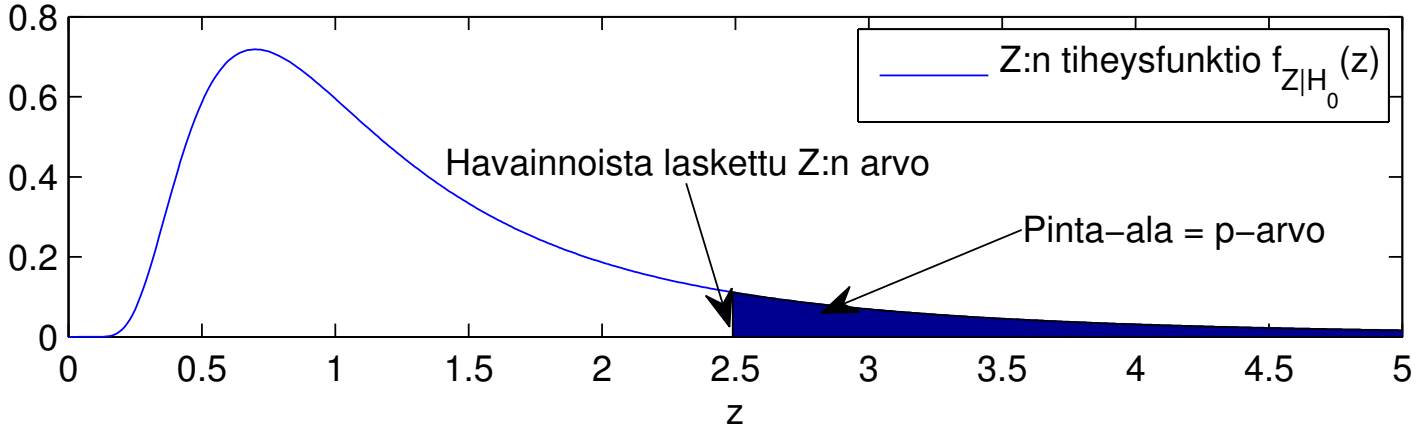


Kuva 1: Huom! Tämä on esimerkki erään testisuureen tiheysfunktiosta. Tämä voisi olla aivan erilainenkin.

p-arvo

- Testisuureen suuruus tulkitaan p-arvon avulla
  - Suuri p-arvo $\rightarrow$ testisuureen arvo Z on todennäköinen H_0:n pätiessä $\rightarrow$ aineisto ei sisällä vahvaa evidenssiä H_0 vastaan
  - Pieni p-arvo $\rightarrow$ testisuureen arvo Z on epätodennäköinen H_0:n pätiessä $\rightarrow$ aineisto sisältää vahvaa evidenssiä H_0 vastaan ja H_1 puolesta
- p-arvo lasketaan eri tavalla riippuen vaihtoehtoisesta hypoteesista:
  - $H_1 : \mu > \mu_0 \rightarrow$ käytä oikeanpuoleista testiä p-arvon laskemiseen
  - $H_1 : \mu < \mu_0 \rightarrow$ käytä vasemmanpuoleista testiä p-arvon laskemiseen
  - $H_1 : \mu \neq \mu_0 \rightarrow$ käytä kaksisuuntaista testiä p-arvon laskemiseen

p-arvo



- Kuvassa käytetty oikeanpuoleista testiä p-arvon laskemiseen.

Tilastollinen päättely p-arvon avulla

- Nollahypoteesi H_0 hylätään, jos testin p-arvo on pieni. Yleisesti joko

  (i) Valitaan etukäteen merkitsevyystaso α, esim 0.05 ja hylätään H_0, jos $p < \alpha$

  (ii) Tehdään johtopäätös vasta, kun p-arvo on laskettu

- Sopivin tapa tehdä tilastollista päättelyä riippuu tulosten käyttötarkoituksesta

Tilastollisen testin suorittamisen vaiheet

- Tilastollisen testin suorittaminen sisältää seuraavat vaiheet:

(1) Asetetaan testin hypoteesit H, H_0, H_1.

(2) Valitaan testisuure Z

(3) Valitaan merkitsevyystaso α

(4) Poimitaan otos niin, että yleisen hypoteesin oletukset pitävät.

(5) Lasketaan havainnoista testisuureen Z arvo ja vastaava p-arvo

(6) Tehdään johtopäätös siitä hylätäänkö nollahypoteesi (p-arvo $< \alpha$) vai jätetäänkö se voimaan (p-arvo $\geq \alpha$)

Testi perusjoukon odotusarvolle, kun otos on normaalijakaumasta (nk. t-testi)

- **Yleinen hypoteesi** H :

  (1) $X_i \sim N(\mu, \sigma^2), \quad i = 1, \ldots, n$

  (2) Satunnaismuuttujat $X_1, \ldots, X_n$ ovat riippumattomia

- **Nollahypoteesi** $H_0 : \mu = \mu_0$
- **Vaihtoehtoiset hypoteesit**

  $H_1 : \mu > \mu_0$, $H_1 : \mu < \mu_0$, $H_1 : \mu \neq \mu_0$
- **Testisuure**

$$Z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}},$$

- Jossa $\bar{X}$ on otoksen keskiarvo ja s otoksen keskihajonta.
- **Testisuureen jakauma** ehdolla että nollahypoteesi pätee on t-jakauma parametrilla $n - 1$, eli $Z \sim t(n - 1)$.

Varianssianalyysi

- Varianssianalyysi voidaan ymmärtää kahden riippumattoman otoksen t-testin yleistykseksi tilanteisiin, joissa

  (i) Perusjoukko koostuu kahdesta tai useammasta ryhmästä

  (ii) Testataan ryhmäkohtaisten odotusarvojen yhtäsuuruutta

  (iii) Havainnot ryhmissä noudattavat normaalijakaumaa

  (iv) Jokaisesta ryhmästä poimitaan toisistaan riippumattomat yksinkertaiset satunnaisotokset

- Nollahypoteesi: ryhmien odotusarvoissa ei ole eroja,

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_i = \mu$$

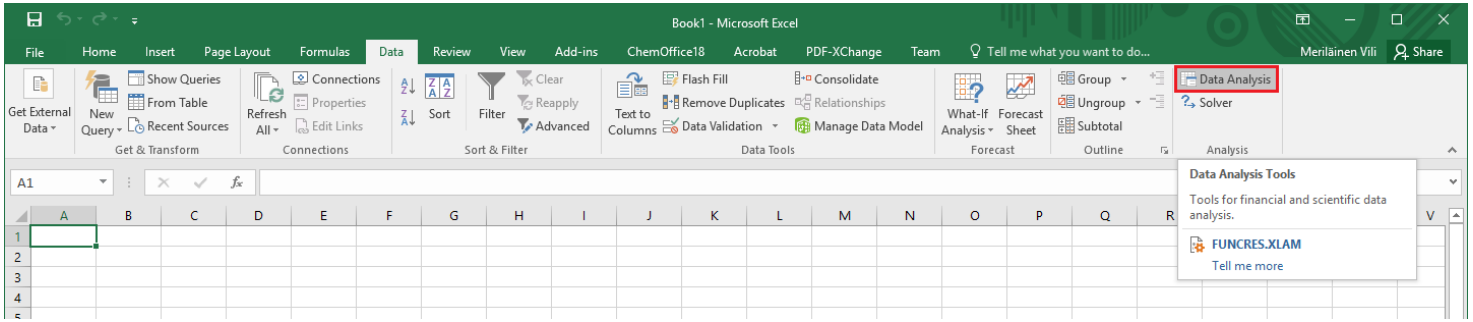
- Vaihtoehtoinen hypoteesi: ainakin kahden ryhmän odotusarvot poikkeavat toisistaan ($H_1 : \mu_i \neq \mu_j,$ joillain i, j)

Ovatko havainnot normaalijakaumasta?

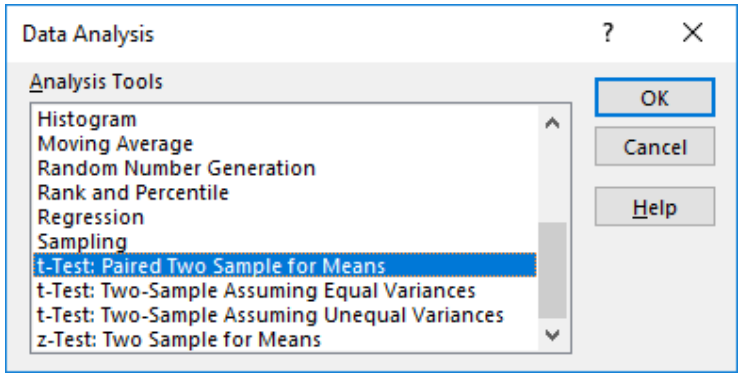
- Yksinkertaisin tapa arvioida aineiston jakautumista on piirtää siitä histogrammi. Mitä suurempi aineisto on, sitä tarkemmin histogrammin tulisi muodoltaan muistuttaa normaalijakauman tiheysfunktiota.
- Huipukkuus (engl. Kurtosis) kuvaa jakauman tai aineiston ”piikikkyyttä”. Mitä suurempi arvo, sitä selvempi piikki aineistossa on. Mitä pienempi arvo, sitä tasaisempi jakauma on. Excelissä normaalijakauma saa kurtosis-tunnusluvun 0.
- Vinous (engl. Skewness) kuvaa jakauman tai aineiston epäsymmetrisyyttä. Negatiiviset arvot kertovat, että aineisto on painottunut oikealle ja häntä vasemmalla puolella on pitempi. Positiiviset arvot vastaavasti kertovat, että aineiston häntä on oikealla pidempi. Normaalijakauman vinous on 0.

Tilastollinen analyysi Excelillä

- Excel sisältää valmiiksi yleisimmät tilastolliset analyysityökalut Analysis ToolPak-lisäosassa
- Analysis ToolPak:n käyttöönotto (Office 2016)
  - File-välilehti → Options → Add-Ins
  - Valitse Manage-tippuvalikosta Excel Add-Ins ja klikkaa Go...
  - Ruksi Analysis ToolPak ja klikkaa OK
- Käyttöönoton jälkeen Analysis ToolPak (Data Analysis) löytyy Data-välilehdeltä



Excelin Analysis ToolPak (Data Analysis)-työkalu



- Klikkaa Data Analysis -valikko auki ja valitse haluamasi tilastollinen testi tai muu työkalu
- Valitse solut, joissa analysoitava data on ja anna tarvittavat parametrit (esim. merkitsevyystaso)
- Valitse mihin haluat tulokset: tiettyyn kohtaan työkirjaa, uudelle välilehdelle tai kokonaan uuteen työkirjaan
- Tulkitse tulokset

R-demo: T-testi parivertailulle

- Tutkitaan kahden eri unilääkkeen vaikutusta koehenkilöiden unenlaatuun. Käytössä on mittauksia koehenkilöiden unen pidentymisestä kummankin lääkkeen vaikutuksen alaisena.
- Käytetään t-testiä parivertailulle, sillä kokeet on suoritettu samoille koehenkilöille. Valitaan merkitsevyystasoksi α=0.05. Nollahypoteesi on H_0: $\mu_1 - \mu_2 = 0$ ja vaihtoehtoinen hypoteesi H_1: $\mu_1 - \mu_2 \neq 0$.

```
> #Ladataan data
> library(datasets)
> data(sleep)
> View(sleep)
> #Tutkitaan ryhmiä
> drug1=sleep$extra[sleep$group == 1]
> drug2=sleep$extra[sleep$group == 2]
> mean(drug1)
[1] 0.75
> mean(drug2)
```

```
[1] 2.33
> #Tehdään parittainen t-testi
> t.test(drug1,drug2, alternative="two.sided", paired = TRUE)

Paired t-test

data:  drug1 and drug2
t = -4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.4598858 -0.7001142
sample estimates:
mean of the differences
                  -1.58
```

$0.002833 < 0.05$ ($p < \alpha$), joten unilääkkeiden vaikutukset unen kestoon poikkeavat toisistaan tilastollisesti merkitsevästi.

Tehtävä A: Ilman kadmiumpitoisuus

- Tutkittaessa ilman kadmiumpitoisuutta tehtiin 35 havaintoa.

Kadmiumpitoisuus (mg/m^3)				
0.064	0.050	0.072	0.064	0.066
0.040	0.086	0.072	0.069	0.050
0.060	0.065	0.059	0.059	0.059
0.077	0.070	0.076	0.081	0.062
0.075	0.057	0.082	0.082	0.090
0.081	0.081	0.078	0.073	0.080
0.067	0.071	0.074	0.062	0.071

- Tehtävänäsi on kuvailla aineistoa ja tehdä siihen liittyvä tilastollinen testaus.

Tehtävä A: Datan kuvailu

1. Vie data Exceliin kopioimalla yllä oleva taulukko pdf-tiedostosta Exceliin (Copy-Paste) ja käytä Excelin Data-välilehdeltä löytyvää Text To Columns -työkalua
2. Laske havaintoaineistolle tunnuslukuja käyttäen Analysis Toolpakin Descriptive Statistics toimintoa

✍ Raportoi kadmiumpitoisuuden keskiarvo, keskihajonta, keskiarvon keskivirhe, vinous ja huipukkuus

Tehtävä A: Datan kuvailu ja normaalisuuden tutkiminen

1. Piirrä histogrammi havainnoista (katso Analysis Toolpak). Piirtämistä varten tarvitsee luoda Exceliin sarake numeroarvoluokista (bins)
   - Luo exceltiedostoon sarake, jossa on tasavälein 11 lukua aineiston minimistä aineiston maksimiin

 Liitä histogrammi vastausdokumenttiin.

2. Ennen tilastollista testaamista olisi hyvä tutkia päteekö yleisen hypoteesin oletukset riittävän hyvin.

 Mitkä asiat puoltavat, että aineisto on normaalijakautunut? Mitkä asiat puhuvat sitä vastaan?

Tehtävä A: Tilastollinen testaus

- Terveyssäädökset vaativat, että kadmiumpitoisuus on keskimäärin alle 0.07. Tehtävänäsi on tutkia tätä tilastollisesti.
- Käytä t-testiä, jossa nollahypoteesinä on $H_0 : \theta = 0.07$.

✍ Mitä vaihtoehtoista hypoteesia H_1 käytät? (Vinkki: katso mallia kalvolta 7. Siellä testataan onko keskiarvo noussut, tässä testataan onko keskiarvo yli rajan.)

1. Laske testisuure kaavalla $T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$, jossa $\bar{X}$ on aineiston keskiarvo, μ_0 on nollahypoteesin mukainen keskimääräinen kadmiumpitoisuus, s aineiston keskihajonta ja n otoskoko.
2. Laske p-arvo Excelin funktiolla T.DIST.RT$(T; n-1)$ (tässä komennossa RT on lyhenne sanoista right tail)

✍ Minkä p-arvon sait testillesi? Hylkäätkö vai hyväksytkö nollahypoteesin, kun käytät merkitsevyystasona $\alpha = 0.05$?

Tehtävä B: Iirislajikkeet (ANOVA) (R)

Tehtävänä on tutkia iirislajikkeiden ominaisuuksien eroja varianssianalyysilla (ANOVA) sekä t-testeillä.

Lataa R:n sisäänrakennettu Iris-datasetti ottamalla käyttöön `datasets`-paketti ja lataamalla data komennolla `data(iris)`. Tulosta yhteenveto iirisaineistosta ja tutustu sen avulla aineiston sisältämiin muuttujiin.

Haluat tutkia, onko eri iirislajien terälehtien pituuksien välillä tilastollisesti merkittävää eroa.

Palauta mieleesi varianssianalyysin (ANOVA) oletukset eli yleinen hypoteesi. Tutki ryhmien jakaumia piirtämällä kaikkien kolmen iirislajin `Petal.Length`-muuttujan histogrammit samaan kuvaikkunaan ja liitä kuva vastaukseesi. Laske myös ryhmien kurtosis- ja skewness-tunnusluvut (tarvitset jonkin tilastopaketin). Miten hyvin ANOVAn yleisen hypoteesin jakaumaoletukset toteutuvat?

- ✍ Mitä nollahypoteesia H_0 käytät? Entä mikä on vaihtoehtoinen hypoteesi H_1?
- ✍ Olkoon käytetty merkitsevyystaso α=0.05. Suorita varianssianalyysi `aov`-komennolla. Mikä on p-arvo ja hylätäänkö nollahypoteesi?

Jos hylkäsit nollahypoteesin, minkä ryhmien väliltä löydät tilastollisesti merkitsevää eroa? Tutki eroja t-testillä (kaksisuuntainen t-testi olettaen erisuuret ryhmävarianssit). Käytetään merkitsevyystasona α=0.01.

- ✍ Löydätkö ryhmien välillä tilastollisesti merkitsevää eroa? Minkä ryhmien?
- ✍ Miksi kahden otoksen vertailussa käytetään pienempää merkitsevyystasoa kuin varianssianalyysissä? (Vinkki: Bonferronin korjaus).

Tehtävä C: Pelinkehitys

- Olet data-analytiikko suomalaisessa pelifirmassa MegaCube. Pelinne 'Dragons and champions' on ensimmäisessä testausvaiheessa.
- Olette kehittäneet pelille kaksi vaihtoehtoista designia. Toinen on nimeltään 'Sininen' ja toinen 'Punainen'.
- Käytössäsi on aineistoa viikon pelitestauksesta ja tehtävänäsi on selvittää designin vaikutus keskimääräiseen peliaikaan ja pelikertojen lukumäärään.
- Käytä kaksisuuntaista T-testiä olettaen erisuuret ryhmävarianssit, kun vastaat kysymyksiin 'onko ero tilastollisesti merkitsevä'. Excelissä T.TEST().
- Löydät datan MyCourses-sivuilta.

Tehtävä C: Pelinkehitys

- ✍ Mikä on ryhmien välinen ero: Keskimääräisessä peliajassa per pelikerta?
- ✍ Mikä on ryhmien välinen ero: Pelikertojen lukumäärässä?
- ✍ Ovatko erot tilastollisesti merkitseviä? Perustele p-arvon avulla.
- ✍ Arvioi, kuinka suuri on designing merkitys kokonaispeliaikaan. Anna vastaus %-lukuna. (Esim. Design A:lla pelataan 5% enemmän kuin B:llä). Perustele.
- ✍ Anna sanallinen yhteenveto tuloksista ja johtopäätöksistäsi.

Kotitehtävä: Datan normaalisuuden tutkiminen

- Lataa kurssin MyCourses-sivuilta (kohdasta Harjoitus 11, Lisämateriaali kotitehtävään) Excel-tiedosto Kotitehtävä_data.xlsx
- Tiedostosta löytyy viisi aineistoa. Kukin aineisto on satunnaisotos jostain perusjoukosta.

1. Laske kaikille aineistoille tilastolliset tunnusluvut (Descriptive Statistics)
2. Piirrä kaikille aineistoille histogrammit. Käytä tässä kaikille aineistoille paitsi vuosituloille numeroarvoluokkina (bins) kymmenentä tasavälein jakautunutta lukua aineiston minimin ja maksimin välillä. Vuosituloille käytä numeroarvoluokkina 20 tasavälein jakautunutta lukua.

Liitä kuva kaikkien aineistojen histogrammeista. Excel-tiedostossa on valmiina esimerkki siitä, miltä histogrammien kuvien pitää näyttää.

Kotitehtävä: Datan normaalisuuden tutkiminen

3. Tehtävänäsi on laittaa aineistot järjestykseen sen perusteella, miten paljon uskot niiden perusjoukkojen jakaumien muistuttavan normaalijakaumaa. Sinun tulee lisäksi perustella valintasi. (Vinkki: googlettamalla löydät paljon tietoa aiheesta.)

- Ota huomioon ainakin seuraavat tiedot: histogrammin ulkonäkö, vinous, huipukkuus ja otoskoko.

✍ Miten otoskoko vaikuttaa päätelmiisi?

✍ Esitä perusteltu näkemyksesi aineistojen järjestyksestä.

✍ Kommentoi tuloksia lisäksi kaupunkilaisjärjelläsi. Mitkä suureista voisivat olla suurinpiirtein normaalisia ja mitkä eivät ainakaan voisi olla normaalisia perustuen yleistietoosi.