3. Datorlaboration 3 handlade om hypotestest och konfidensintervall. För denna del förväntas ni läsa och följa instruktionerna i datorlaborationen.

3.1.1 (fråga 1.1 i lab 3) Vad är stickprovsandelen p_hat?

3.1.2 (fråga 1.2 i lab 3) Spara och infoga ert histogram. Vad har vi visat med denna simulering? Förklara!

3.1.3 (fråga 1.3 i lab 3) Använd resultatet från simuleringen (MI_proportions) för att skatta p-värdet. Jämför p-värdet med vår gräns på 1%. Dra en slutsats.

OBS! För fråga 3.1.3 ska ni kort beskriva hur ni skattade p-värdet.

3.1.4 (fråga 1.4 i lab 3) Upprepa hypotestestet från uppgift 1.3 med prop.test(). Läs om fuktionens argument genom att skriva ?prop.test i Console. Du kommer att behöva använda argumenten x, n, p och alternative. Jämför p-värdet med värdet från uppgift 1.3. Fick du ungefär samma?

Tips: x ska vara antalet giltiga underskrifter, n ska vara stickprovsstorleken, p ska vara andelen som vi antar vara sann i nollhypotesen, alternative ska vara ”greater”, eftersom vi testar större än i den alternativa hypotesen

3.2.1 (fråga 2.1 i lab 3) Här räcker det att skapa ett histogram över medelbetygen. Infoga i er rapport.

3.2.2 (fråga 2.2 i lab 3) Titta på bootstrapkonfidensintervallet. Tolka resultatet med ord. Varför har vi valt percentilerna 2.5% och 97.5%?

Presentera ert konfidensintervall med siffror också.

3.2.3 (fråga 2.3 i lab 3) Beräkna ett konfidensintervall för filmbetygen med formeln ovan. Kan vi anta normalfördelning? Jämför med bootstrapintervallet.

Ni kan beräkna stickprovsstandardavvikelsen s i R med sd() och $t\star$ med qt(.975, df = 99), där 99 är stickprovets storlek minus ett.

3.2.4 (fråga 2.4 i lab 3) Beräkna ett konfidensintervall för filmbetygen med t.test(). Kan vi anta normalfördelning? Jämför med bootstrapintervallet.

Ta screen shot på ert output från t.test() och ha med i rapporten.

3.2.5 (fråga 2.5 i lab 3) Tolka output från testet. Vad kan vi säga om skillnaden? Är det statistiskt säkerställt att engelskspråkiga filmer får högre betyg i genomsnitt?

Ta screen shot på ert output från t.test() och ha med i rapporten.

4. Spridningsdiagram, korrelation och regression (bygger vidare på labb 4).

Vi kommer att använda datasetet **gapm** med länder och sex variabler från Gapminder[1], som ni också träffar på i labb 4. Vi har dessutom lagt till vår variabel ”landlocked” från tidigare i kursen. Du kan ladda ner data här (högerklicka), eller från Githubsidan (labb 4) eller från Datafiler i Athena. Data är från 2022. Variablerna som finns i datasetet är[2]:

country – de länder som finns i Gapminderdata och för vilka det finns kompletta data
child_mort – antal barn som dör före fem års ålder, per 1000 barn födda
fertility – förväntat antal barn per kvinna
co2_cap – antal ton koldioxid som varje individ ”konsumerar”
gdp_cap – BNP per capita i dollar (köpkraftsjusterat)
life_exp – förväntad medellivslängd
landlocked – indikator för om ett land har kust eller inte (1=har inte kust)

Börja med detta:

Sätt arbetskatalogen och ladda mosaicpaketet. Ladda ner data till arbetskatalogen. Läs in data till R, från arbetskatalogen, med read.csv-kommandot och skapa en data frame med era inlästa data, kalla den exv. **gapm**.

Bekanta er med hur data ser ut genom kommandona head(gapm) – titta på de första sex raderna, str(gapm) – vilka variabeltyper vi har, class(gapm) – vilken typ av dataobjekt vi har, summary(gapm) – sammanfattande mått för de olika variablerna. Gör också gärna exv. histogram över de enskilda variablerna för att se hur data är fördelade, exempelvis medellivslängd och koldioxidutsläpp i olika länder (detta behöver inte tas med i redovisningen).

4.1 Ta fram korrelationskoefficienten mellan barnadödlighet och övriga variabler (förutom landlocked)

Med vilken annan variabel är korrelationen högst?

4.2 Gör ett spridningsdiagram för sambandet mellan barnadödlighet och bnp per capita

Beskriv hur sambandet ser ut. Är sambandet linjärt? Beskriv skillnaden mellan denna graf och den graf baserad på liknande data som vi sett på föreläsningarna.

4.3 Gör ett spridningsdiagram för sambandet mellan förväntat antal barn per kvinna och barnadödlighet

Beskriv hur sambandet ser ut. Är sambandet linjärt?

4.4 Gör en regression med förväntat antal barn per kvinna som responsvariabel och barnadödlighet som förklaringsvariabel. Plotta regressionslinjen i det spridningsdiagram ni gjorde i 4.3.

Hur starkt är sambandet mellan de två variablerna (förklaringsgraden R2)? Är sambandet signifikant på 95%-nivån? Tolka lutningskoefficienten. Ta fram ett 95%-igt konfidensintervall för lutningskoefficienten. Kan vi säga något om kausalitet?

4.5 Till regressionen i 4.4, lägg till variabeln landlocked som en andra förklaringsvariabel.

Förändras R2 och lutningskoefficienten från 4.4 nämnvärt? Tolka lutningskoefficienten för variabeln barnadödlighet (obs: multipel regression). Är variabeln landlocked en signifikant förklaringsvariabel?

---

[1] Based on free material from GAPMINDER.ORG, CC-BY LICENSE.
[2] Mer exakta definitioner av vissa av variablerna finns på Gapminders hemsida men är inte viktiga för uppgiften.