Datum: 260331

Inlämningsuppgift, Statistisk översiktskurs, VT26

Introduktion, inlämning, format

I denna inlämningsuppgift ska ni självständigt i grupper analysera ett datamaterial i programmeringsspråket R. Till skillnad från datorlaborationerna finns det få kodexempel.

Datorlaborationerna går igenom de flesta momenten som behandlas i inlämningsuppgiften. Inlämningsuppgiften har fyra uppgifter (1, 2, 3, 4), med deluppgifter därunder. De fyra uppgifterna tar upp frågor som liknar det som gås igenom på laborationerna 1-4. När du är klar med laboration 1 ska du kunna göra uppgifterna under 1 i inlämningsuppgiften, osv.

Efter att ni lämnat in uppgiften kan ni antingen få godkänt eller komplettering (i systemet registreras kompletteringsbehov som underkänt). Därefter finns en möjlighet att komplettera, se datum nedan. Efter inlämningsdatum 2 finns ingen ytterligare chans att komplettera, inlämningsuppgiften behöver då göras om vid ett framtida kurstillfälle. Kursbeskrivningen innehåller ytterligare information om bedömningsgrunder, etc., läs igenom detta dokument. All typ av plagiering är otillåten och textmatchningsverktyg används.

1. Inlämning, senast **Tisdag 21 april kl. 16.59. Återlämnas 28 april.**

2. Komplettering, senast **Fredag 8 maj kl. 16.59.**

När ni lämnar in är det viktigt att svara på alla frågor och delfrågor. Om svar inte finns på vissa frågor vid första inlämningen kan ni inte få feedback på den uppgiften. Vi rekommenderar att ni går igenom hela uppgiften och svarar på alla frågor redan vid inlämning 1.

Inlämningen ska innehålla svar i form av siffersvar, grafer, etc., beroende på ställd fråga. Varje uppgift ska också innehålla ett textsvar. Grafer ska ha rubrik, namn på axlarna, etc.

Ni kan skriva i exv. Microsoft Word, eller annat dokument. Koden ni använder ska också finnas med i inlämningen, ni kan antingen klippa in den sist i ert dokument (börja på en ny sida och markera tydligt, i koden, exv. med radbrytningar och förklarande rubriker, var varje deluppgift finns i koden). Om ni tycker det är bättre kan ni istället klistra in koden vid respektive uppgift, i huvuddokumentet, eller lämna in en separat kodfil (som ska gå att följa, uppgift för uppgift).

Skriv fullständiga namn, gruppnummer och inlämningsdatum längst upp i de dokument ni lämnar in. Namnge er fil som Grupp_XX, där XX är ert gruppnummer (använd inga specialtecken i filnamnet såsom t.ex. punkt, kommatecken, parantes, hakparentes, semikolon, kolon, å, ä, ö.) Om ni vill lämna in ett annat filformat än .docx eller .pdf – kolla med lärarna först om det är OK (begränsningar finns i inlämningssystemet).

På alla moment i kursen, och speciellt på jour och datorlabbar, finns möjlighet att ställa frågor och lärarna finns tillgängliga för att hjälpa till. Kontakta lärarna i god tid vid eventuella problem.

Alla gruppmedlemmar ska vara delaktiga och bidra till alla delar av rapporten och arbetet som leder upp till rapporten, dvs skriva kod, analysera data, tolka resultat, dra slutsatser och skriva rapporten.

1. I labb 1 jobbade vi bland annat med data på kapitalavkastning över tid.

1.1 Definiera en vektor med tio värden (0.05, 0.07, 0.054, 0.07, 0.074, 0.087, 0.074, 0.094, 0.091, 0.089), där värdena är andelen av de tillfrågade som säger att de skulle rösta på XYZ-partiet, från opinionsundersökningar i januari för åren 1-10.

1.2 Gör en lämplig graf som visar opinions**andels**utvecklingen för XYZ-partiet, med lämpliga värden etc. på axlarna och en tydlig rubrik.

1.3 Modifiera vektorn ovan på lämpligt sätt och gör en graf som visar opinionsutvecklingen, i **procent**, för XYZ-partiet, med lämpliga värden på axlarna och en tydlig rubrik.

1.4 I uppgift 1.1 ovan var data från åren 2010-2019. Istället för att skapa en vektor, använd följande kommando för att skapa en data frame med två kolumner:

```
data.frame(Year=c(2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019), Vote_intention = c(0.05, 0.07, 0.054, 0.07, 0.074, 0.087, 0.074, 0.094, 0.091, 0.089))
```

För att inte bara skriva ut ovanstående på skärmen, välj ett lämpligt namn på objektet (er data frame) och lagra ovanstående i objektet.

1.5 Skapa en graf som i 1.2 men använd istället kolumnerna i er data frame för att göra grafen.

1.6 Lägg till en tredje kolumn med riksdagsspärren i er data frame (vi antar här att er data frame heter AAA), med följande kommando:

```
AAA$threshold <- rep(0.04, 10)
```

(Kommandot repeterar värdet 0.04 tio gånger och lägger in som en ny variabel i AAA)

Titta på de första raderna i er data frame för att se att den nya kolumnen lagts till

```
head(AAA)
```

1.7 Plotta riksdagsspärren i samma graf som ni gjort i 1.5. Direkt efter ert plottkommando i uppgift 1.5, kör följande kommando (byt till ert namn, istället för AAA):

```
lines(AAA$threshold~AAA$Year)
```

lägg också till någon färg i kommandot så att linjen skiljer sig från opinionsdata

Uppdatera rubrik etc., I grafen, på lämpligt sätt.

1.8 Använd ett lämpligt R-kommando för att ta fram femte värdet i första kolumnen av er data frame.

2. I laboration 2 jobbade vi bland annat med deskriptiv statistik, tabeller och grafer, fördelningar, och central- och spridningsmått.

2.1 Ta fram medelvärde, median, varians och standardavvikelse, med ett eller flera R-kommandon, för den vektor ni skapade i uppgift 1.1 ovan

Läs in datasetet titanic. Svara på följande frågor med hjälp av lämpliga R-kommandon:

2.2 Hur många män respektive kvinnor reste med Titanic?

2.3 Är alla passagerare (inkl. besättning) registrerade som antingen man eller kvinna? (Tips – finns saknade värden?)

2.4 Ta fram en lämplig tabell som visar överlevnadsfrekvensen (andelen överlevande) för män respektive kvinnor.

2.5 Ta fram en lämplig figur som visar överlevnadsfrekvensen (andelen överlevande) för män respektive kvinnor.

2.6 Ta fram en tabell som visar antal överlevande och antal icke-överlevande, uppdelat på kön. Vilken grupp är störst: antalet män som överlevde eller antalet kvinnor som överlevde?

2.7 Gör ett histogram över variabeln ”Paid” och beskriv hur fördelningen ser ut (modalitet, ev. skevhet, jämför median vs. medel). Testa att variera hur många staplar histogrammet har med argumentet breaks (labb2).

3. Punktskattning, konfidensintervall, hypotestest (från labb 3).

Kör först följande två kommandon, där vi skapar variabeln training. Var noggrann med att alltid ha med första raden (set.seed(7)-kommandot), exempelvis om ni skulle definiera er variabel training igen, etc. Första raden kommer göra att ni alltid får samma data trots att det andra kommandot (rnorm) använder en slumpfunktion.

```
set.seed(7)
training <- rnorm(400,300,100)
```

Modifiera sedan variabeln training genom att addera ert gruppnummer (se Excelbladet). Numren är mellan 1 och 13. Om det exempelvis skulle funnits en grupp 20 skulle denna grupp ha använt följande kommando:

```
training <- training + 20
```

Den uppdaterade variabeln är den variabel ni ska jobba med.

Vi tänker att våra data / vår variabel representerar träningsmängder i minuter per månad för ett slumpmässigt och oberoende urval av individer från en viss population vi är intresserade av.

3.1 Illustrera fördelningen av data med ett lämpligt diagram. Beskriv fördelningen kort.

3.2 Baserat på ert urval, ta fram en punktskattning för populationsmedelvärdet för träningsmängd.

3.3 Ta fram ett 90%-igt konfidensintervall för skattningen i 3.2. Tolka konfidensintervallet.

3.4 Ett institut rekommenderar en medelträningstid på 300 min/månad. Formulera ett hypotestest där hypotesen att populationsmedelvärdet skiljer sig från 300 min/månad testas mot en relevant nollhypotes. Använd signifikansnivån 5%. Utför hypotestestet. Tolka resultatet.

3.5 Ett institut rekommenderar en medelträningstid på 300 min/månad. Formulera ett hypotestest där hypotesen att populationsmedelvärdet är större än 300 min/månad testas mot en relevant nollhypotes. Använd signifikansnivån 5%. Utför hypotestestet. Tolka resultatet.

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.