ITHS
Statistiska metoder 21-22
Mattias Sundén

Inlämningsuppgift

Arbetet kan genomföras i grupper om två personer (men det är ok att arbeta ensam) och skall redovisas i en skriftlig rapport. Rapporten skall mejlas i pdf-format till mattias.sunden@iths.se senast kl.23.59, 25/1. Det går inte att lämna in efter den tiden. Alla frågor ska besvaras. Inga långa rapporter ska skrivas utan endast svar, i vissa fall motiverade med utskrifter eller grafer från Jupyter notebooks/Python, krävs.

Tabeller och/eller plottar som ligger till grund för era svar ska kopieras in i dokumentet för de frågor som är markerade med en asterisk (*). Man skall kunna förstå vad diagram eller tabeller beskriver genom att det i stället för koder står klartext. Diagram skall ha tydliga axelbeteckningar (blanda t.ex. inte svenska och engelska). ***Ekvationseditorn i Word skall användas för alla formler.***

OBS! Glöm inte ange gruppmedlemmarnas namn, personnummer och e-postadress.

Avsikten med inlämningsuppgiften är att belysa centrala gränsvärdessatsen, hypotestest och konfidensintervall.

Uppgift 1

Här är tanken att du lära dig förstå hur Centrala gränsvärdessatsen fungerar genom att se hur ett histogram av observationer av stickprovsmedelvärden påverkas av hur många observationer som ingår i varje medelvärde. Dessutom behandlas konfidensintervall och styrka hos ett hypotestest.

I Python kan man, t ex med numpy.random.random_integers, generera observationer av en likformigt fördelad diskret slumpvariabel som tar värden 4,5,6,7. För en sådan slumpvariabel, kalla den X, gäller alltså att $P(X = 4) = P(X = 5) = P(X = 6) = P(X = 7) = 1/4$. I deluppgifterna 1-7 menas slumpvariabel/slumpvariabler definierade enligt ovan även om det bara står slumpvariabel/slumpvariabler.

1) Vi ska alltså skapa observationer av stickprovsmedelvärden $\bar{X}$. Hur ska du göra/skriva i Python för att få observationer av stickprovsmedelvärden baserade på fem observationer av slumpvariabler? Svara gärna med hjälp av en skärmdump och förklarande text.

2) Skapa 1000 standardiserade (dvs subtrahera $\mu_{\bar{X}}$ och dividera med $\sigma_{\bar{X}}$) stickprovsmedelvärden, vart och ett baserat på 2 observationer av slumpvariabler, och gör histogram för de 1000 standardiserade stickprovsmedelvärdena. Upprepa proceduren för stickprovsstorlekarna 10, 20, 30 och 50. (Tips: Gör en array/dataframe för varje stickprovsstorlek, alltså en med 1000 stickprovsmedelvärden baserade på stickprov av storlek 2, en med 1000 stickprovsmedelvärden baserade på stickprov av storlek 10 osv…) Histogrammen ska vara inkluderade i rapporten och det ska tydligt framgå vilka stickprovsstorlekar som använts för stickprovsmedelvärdena som histogrammen baseras på.*

3) Hur beräknar du $\mu_{\bar{X}}$ i 2)? (Svara med en formel, uträkning baserad på formeln och ett numeriskt svar)

4) Hur beräknar du $\sigma_{\bar{X}}$ i 2)? (Svara med en formel, uträkning baserad på formeln och ett numeriskt svar)

5) Vad händer med fördelningen för de standardiserade stickprovsmedelvärdena då antalet observationer som stickprovsmedelvärdena baseras på ökar? Finns det något teoretiskt stöd för detta och i så fall vilket?

6) Gör 1000 95%-konfidensintervall för populationsmedelvärdet μ, vart och ett baserat på 50 observationer av slumpvariabler. Hur många av dessa täcker populationsmedelvärdet? Är detta vad du förväntade dig? Motivera?

7) Antag att vi vill göra hypotestest för populationsmedelvärdet med

$$H_0: \mu \leq 5.1$$
$$H_A: \mu > 5.1$$

Testet ska utföras på signifikansnivån 0.05 och med stickprovsstorleken 50. Med hjälp av de 1000 raderna och 50 kolonnerna med observerade slumpvariabler vill vi

undersöka testets styrka genom att utföra testet 1000 gånger. Vad blir styrkan, dvs hur stor andel av testerna förkastar den falska nollhypotesen?

Uppgift 2

Syftet med den här uppgiften är att belysa det vanliga misstaget att tro att bara för att man får stora stickprov så blir data normalfördelade. Vad CGS säger är ju att om stickproven blir stora så blir stickprovsmedelvärdena/stickprovsproportionerna normalfördelade! I den här uppgiften kommer du att skapa stickprovsdatadata av olika storlekar. Du ska använda slumptal som är **kontinuerligt** likformigt fördelade på intervallet $[7,11]$. Sådana slumptal kan genereras med numpy.random.uniform. Nedan kallar vi dessa bara slumptal.

8) Generera 200 stickprov av storlek fem och gör ett histogram av alla 1000 observationerna. Tyder histogrammet på att data är normalfördelade?*

9) Skapa 200 standardiserade medelvärden vart och ett baserat på stickprov av storlek fem genererade i uppgift 8 och skapa ett histogram av dessa medelvärden. Tyder histogrammet på att medelvärdena är normalfördelade?*

10) Hur beräknar du väntevärdet för stickprovsmedelvärdena, $\mu_{\bar{X}}$, i 9)? (Svara med en formel, uträkning baserad på formeln och ett numeriskt svar)

11) Hur beräknar du populationsstandardavvikelsen för stickprovsmedelvärdena, $\sigma_{\bar{X}}$, i 9)? (Svara med en formel, uträkning baserad på formeln och ett numeriskt svar)

12) Generera 200 stickprov av storlek 20 och gör ett histogram av alla 4000 observationerna. Tyder histogrammet på att data är normalfördelade?*

13) Skapa 200 standardiserade medelvärden vart och ett baserat på stickprov av storlek 20 genererade i uppgift 12 och skapa ett histogram av dessa medelvärden. Tyder histogrammet på att medelvärdena är normalfördelade?*

14) Generera 200 stickprov av storlek 50 och gör ett histogram av alla 10000 observationerna. Tyder histogrammet på att data är normalfördelade?*

15) Skapa 200 standardiserade medelvärden vart och ett baserat på stickprov av storlek 50 genererade i uppgift 14 och skapa ett histogram av dessa medelvärden. Tyder histogrammet på att medelvärdena är normalfördelade?*

16) Slutsatsen blir att

    a) Det spelar ingen roll hur många slumptal jag genererar så blir varken slumptalen eller stickprovsmedelvärdena normalfördelade.
    b) Om jag bara genererar tillräckligt många slumptal så blir slumptalen normalfördelade.
    c) Om jag skapar stickprovsmedelvärden av tillräckligt många slumptal så blir stickprovsmedelvärdena likformigt fördelade.
    d) Om jag skapar stickprovsmedelvärden av tillräckligt många slumptal så blir stickprovsmedelvärdena normalfördelade.