Labb 2 - SF1918

Selma Hagelin
Sakarias Åman Rosengren

December 2023

Innehåll

1 Problem 1 - Simulering av konfidensintervall

Ett konfidensintervall med konfidensgraden $1 - \alpha$ är ett statistiskt verktyg som används för att uppskatta intervallet inom vilket det sanna värdet av en okänd parameter, i detta fall μ (väntevärdet), förväntas ligga med en viss sannolikhet $1 - \alpha$. För att fördjupa förståelsen av detta begrepp och dess tillämpning används simuleringar.

I den aktuella situationen har vi valt en normalfördelning med ett medelvärde $\mu = 2$ och en standardavvikelse $\sigma = 1$. Simuleringarna utförs 100 gånger, där varje simulering innebär att vi tar 25 oberoende observationer från denna fördelning. Syftet är att skatta ett konfidensintervall för väntevärdet med en konfidensgrad på 95%.

Resultaten av simuleringarna presenteras grafiskt genom att plotta 100 konfidensintervall. De horisontella strecken representerar dessa intervall, och det vertikala strecket markerar det sanna värdet av μ (i detta fall 2). De röda strecken indikerar de intervall där det sanna värdet inte är fångat, medan de blåa strecken representerar de intervall där det sanna värdet förväntas finnas inom intervallet.

För att fördjupa insikterna kan man variera parametrarna, såsom att ändra μ, σ, n och α, en i taget, och observera hur dessa förändringar påverkar resultaten. Genom att utföra simuleringarna upprepade gånger kan man också uppskatta variationen i resultaten på grund av slumpmässighet i simuleringarna. Detta ger en praktisk förståelse för hur olika faktorer påverkar konfidensintervallen och därigenom stödjer tolkningen av resultaten. Detta är de samband som observerats:

- Öka σ gör intervallen bredare, och vice versa.
- Öka n gör intervallen smalare, och vice versa.
- Minska α gör intervallen bredare (ger högre konfidens och smalare konfidensintervall), och vice versa.
- Förändring av μ kommer påverka var konfidensintervallen centreras då de utgår ifrån μ, och även hur det beräknade medelvärdet beter sig, vilket i sin tur påverkar bredden på konfidensintervallet.

2 Problem 2: Maximum Likelihood- och Minsta Kvadrat-Skattningar

I denna uppgift står vi inför utmaningen att undersöka två olika metoder för att skatta en parameter i en Rayleighfördelning. För att genomföra denna analys används en kodsnutt som genererar en samling stokastiska variabler från en Rayleighfördelning med en specifik parametervärde på 4. Efter att dessa variabler har genererats, utförs två olika skattningar, benämnda `est_ml` och `est_mk`, vilka kommer från förberedelseuppgift 1.

I själva problemet används en storlek på 10 000 observationer ($M = 10000$) och en given parameterinställning på 4 ($b = 4$). Genom att simulera dessa utfall erhålls en serie Rayleighfördelade stokastiska variabler. Ett histogram skapas för att ge en visuell representation av fördelningen av dessa observationer. Skattningarna `est_ml` och `est_mk`, som för tillfället inte är definierade och markerade som kommentarer, presenteras grafiskt som röda och gröna stjärnor, medan det sanna parametervärdet visas som en blå cirkel.

För att bedöma kvaliteten på dessa skattningar uppmanas vi att granska täthetsfunktionen genom att plotta den för `est_ml`. En ny figur skapas där histogrammet visas igen, och dessutom inkluderas en röd kurva som representerar täthetsfunktionen för den skattade parametern (`est_ml`).

Tolkningen som görs är att ML-skattningen ligger väldigt nära den blåa cirkeln som representerar det sanna värdet, och täthetsfunktionen passar väl överens med histogrammet, vilket indikerar att ML-skattningen är en bra uppskattning vid Rayleigh-fördelning. MK-skattningen ligger inte lika nära det sanna värdet, men relativt. Det visar att även denna enklare metod ger en rimlig, men långt ifrån perfekt, uppskattning. Skillnaden mellan dessa två skattningar och deras närhet till det sanna värdet ger en uppfattning om hur robusta dessa metoder är för att skatta parametern i Rayleighfördelningen, och vi kan dra slutsatsen att ML-skattning ger en bättre bild av det sanna värdet än MK-skattningen, även om den är mer komplicerad.

Denna uppgift ger därmed möjlighet att inte bara jämföra två olika skattningsmetoder utan även att utvärdera hur väl dessa skattningar återspeglar den faktiska fördelningen. Genom att visuellt inspektera histogrammet och täthetsfunktionen får vi en djupare förståelse för hur väl skattningarna presterar och om de ger meningsfulla resultat.

3 Konfidensintervall för Rayleighfördelning

Uppgiften är att analysera vågdata som följer en Rayleigh-fördelning. Först plottar vi de första hundra datapunkterna av vågdatan vilket visar amplituden vid varje datapunkt. Vi plottar ett histogram som visar sannolikhetsdensitetsfunktionen, PDF. Datan normaliseras.

Inledningsvis laddas datan från en .dat-fil och en del av denna data visualiseras som en tidsdomän-signal för att ge insikt i dess karaktär. Ett histogram över hela datamängden plottas sedan för att ge en visuell representation av sannolikhetsdensitetsfunktionen (pdf). Histogrammet används för att bedöma datans fördelning och normaliseras så att det totala området under histogrammet blir 1, vilket representerar en sannolikhet.

Vidare görs en punktskattning av skalfaktorn för Rayleigh-fördelningen genom Maximum Likelihood Estimation (MLE). Skattningen beräknas med hjälp av de observerade värdena, och dess noggrannhet bedöms genom beräkning av ett konfidensintervall. Konfidensintervallet, beräknat med en konfidensnivå på 95%, ger en uppskattning av osäkerheten kring den skattade parametern.

Slutligen visualiseras konfidensintervallet och MLE-skattningen tillsammans med datans histogram. Dessutom plottas den teoretiska täthetsfunktionen för Rayleigh-fördelningen med hjälp av den skattade parametern för att bedöma hur väl skattningen överensstämmer med den observerade datan. Om täthetsfunktionen passar väl över histogrammet indikerar det att den valda fördelningsmodellen är lämplig för datamängden och att den skattade parametern är en god uppskattning.

Denna analys är kritisk för att förstå signalens egenskaper och är avgörande för att tillämpa statistiska modeller på verkliga data.

4 Problem 4: Jämförelse av Fördelningar hos Olika Populationer

I denna omfattande analys av datamängden *birth.dat* syftar vår metodik till att belysa potentiella samband och riskfaktorer för låg födelsevikt bland nyfödda. Genom att använda en kombination av histogram, boxplot och kärnestimatorer ges en djupgående inblick i datan och dess möjliga samband.

Histogrammen för födelsevikt, moderns ålder, längd och vikt ger en översikt över fördelningarna. Dessa visuella representationer gör det möjligt att identifiera centrala tendenser, spridningar och eventuella utstickande värden, vilket är avgörande för att förstå datan.

Fokus riktas sedan mot rökning som en potentiell riskfaktor för låg födelsevikt. Genom att differentiera födelsevikter mellan rökande och icke-rökande mödrar får vi en första indikation på eventuella samband. Boxplot-analysen stärker vår förståelse genom att visa fördelningen och outliers för båda grupperna, vilket kan vara värdefullt för att identifiera potentiella risker.

Vidare integrerar vi kärnestimatorer för att erhålla en smidigare och mer kontinuerlig representation av fördelningarna. Denna teknik ger oss möjlighet att visualisera sannolikhetsdensitetfunktioner och jämföra formen av dessa funktioner för rökande och icke-rökande mödrar.

Slutligen, när vi undersöker andra kategoriska variabler, är det viktigt att vara medveten om NaN-värden och hantera dem på ett lämpligt sätt för att inte snedvrida resultaten. Genom att använda `np.isnan` skapar vi en renad dataset som kan användas för en mer tillförlitlig och nyanserad analys.

Denna noggranna metodik möjliggör inte bara identifiering av samband utan också en fördjupad förståelse för hur olika faktorer kan påverka födelsevikten. Resultaten kan bidra till medicinska insikter och informera om preventiva åtgärder för att minska risken för låg födelsevikt hos nyfödda. Skillnaderna verkar inte vara tillräckligt extrema för att dra definitiva slutsatser, men vi kan åtminstone dra slutsatsen att plottarna utifrån den givna datan verkar indikera att födelsevikten verkar vara något lägre för rökande mödrar än för icke-rökande mödrar.

5 Uppgift 5: Test av normalitet

Uppgiften här är att avgöra om fyra olika variabler - barnets födelsevikt, moderns ålder, moderns längd och moderns vikt - är normalfördelade med en signifikansnivå på 5 procent. För att göra detta så utför vi Jarque-Bera testet på alla variabler. Testresultatet inkluderar en teststatistik och ett p-värde. Ett lågt p-värde (vanligtvis under signifikansnivån 0,05) tyder på att datamängden inte är normalfördelad. Vi ser utifrån datan att moderns längd ser ut att vara normalfördelad men vi kan inte vara säkra medans moderns vikt eller ålder samt barnets vikt inte är normalfördelad.

För att undersöka om en datamängd är normalfördelad används `stats.probplot`-funktionen i Python. Denna metod ger en visuell representation genom att jämföra kvantiler av den empiriska datamängden med kvantilerna för en normalfördelning. Funktionen kräver att returvärdet tilldelas variabeln `_` för att undvika onödig utskrift i notebook-miljön. Dessutom måste `plt`-modulen inkluderas för att specificera platsen för plotten. Det är viktigt att notera att `stats.probplot`-funktionen har en bugg som påverkar visningen av den röda linjen om datamängden innehåller NaN-värden. Det rekommenderas därför att filtrera bort dessa värden med `np.isnan` innan plottningen.

För att undersöka normalfördelningen för variablerna för barnets födelsevikt, moderns ålder, moderns längd och moderns vikt kan man använda `stats.probplot`. Genom att analysera formen på dessa s.k. sannolikhetsplotter kan man få en känsla för hur väl datat överensstämmer med en normalfördelning. Om punkterna på plotterna ligger nära den röda linjen, indikerar det att datat är närmare en normalfördelning.

För en mer formell statistisk bedömning av normalitet används Jarque-Beras test. Denna metod jämför skevhet och kurtosis för den empiriska datan med förväntade värden för en normalfördelning. Skevhet (γ) och kurtosis (κ) definieras för en slumpvariabel X med väntevärde μ och standardavvikelse σ. Jarque-Beras testvariabel (JB) används för att testa nollhypotesen att datat är normalfördelat och förväntas vara approximativt χ^2-fördelad med två frihetsgrader.

För att genomföra Jarque-Beras test i Python används `stats.jarque_bera`-funktionen. Detta test ger ett p-värde, och om p-värdet är mindre än den valda signifikansnivån (oftast 0,05) förkastas nollhypotesen, vilket tyder på att datat inte är normalfördelat. Om p-värdet är större än signifikansnivån har man inte tillräckligt med bevis för att förkasta nollhypotesen och kan därför anta att datat är normalfördelat. Detta ger en kvantitativ indikation på normalitet jämfört med den visuella inspektionen från `stats.probplot`.

6 Uppgift 6: Enkel linjär regression

Denna rapport beskriver användningen av linjär regression för att modellera fenomenet Moores lag, vilket är observationen att antalet transistorer per ytenhet på integrerade kretsar fördubblas ungefär varannat år. Detta är hur vi gått tillväga: En datafil som innehåller år och motsvarande antal transistorer per ytenhet laddas in. Eftersom Moores lag antyder en exponentiell tillväxt, transformeras antalet transistorer med logaritmen för att linearisera relationen, vilket gör det möjligt att använda linjär regression. En designmatris X skapas som består av en kolumn med ettor (för interceptet β_0) och en kolumn med årstal x, vilket möjliggör en linjär modell med formen $w_i = \log(y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i$. Skattningarna av parametrarna β beräknas med hjälp av minsta kvadratmetoden. De förutspådda logaritmerade värdena av transistorerna beräknas med den skattade modellen. Den ursprungliga datan och den anpassade modellen plottas för att visuellt utvärdera modellens passform. Residualerna beräknas och visualiseras med hjälp av en kvantil-kvantil-plot (Q-Q plot) och ett histogram för att bedöma deras fördelning.

För att göra en prediktion för år 2025 med modellen, används följande formel:

$$\log(\hat{y}_{2025}) = \hat{\beta}_0 + \hat{\beta}_1 \times 2025$$

Därefter omvandlas den logaritmerade skattningen tillbaka till den ursprungliga skalan med exponentialfunktionen för att få den förväntade mängden transistorer per ytenhet. Uppskattningen för år 2025 landar på ungefär 135 987 000 stycken.

7 Uppgift 7: Multipel linjär regression

I denna studie utförs både enkel och multipel linjär regression för att undersöka sambandet mellan barnets födelsevikt och olika förklaringsvariabler, inklusive moderns längd och rökvanor. Data laddas från filen `birth.dat`, och relevant information om moderns längd och barnets födelsevikt extraheras.

Enkel Linjär Regression Initialt tillämpas enkel linjär regression för att modellera relationen mellan barnets födelsevikt och moderns längd. En designmatris X konstrueras, bestående av en kolumn med ettor (för interceptet) och en kolumn med moderns längd. Funktionen `tools.regress` används för regression, vilket ger skattningar för parametrarna β_0 och β_1. Modellen visualiseras genom att plotta barnets födelsevikt mot moderns längd, med den linjära modellens förutsägelse överlagrad.

Multipel Linjär Regression Analysen utvidgas sedan till en multipel linjär regressionsmodell. Ytterligare förklaringsvariabler inkluderas: moderns rökvanor, representerade av två binära variabler för icke-rökare och rökare. En ny designmatris X skapas för att inkludera dessa variabler. Regressionen utförs återigen med `tools.regress`, och skattningar för parametrarna $\beta_0, \beta_1, \beta_2, \beta_3$ erhålls. Dessa resultat ger insikt i varje förklaringsvariabels inverkan på barnets födelsevikt.

Residualanalys För att utvärdera modellens anpassning och antaganden genomförs en residualanalys. Residualerna beräknas och analyseras med två metoder: en kvantil-kvantil-plot (Q-Q plot) genereras med `stats.probplot` för att visuellt bedöma om residualerna är normalfördelade, och ett histogram av residualerna plottas. En Q-Q plot som visar att residualerna ligger längs en rät linje indikerar att de är normalfördelade, vilket är en viktig förutsättning för vissa statistiska tester och slutsatser i linjär regression.

Sammanfattning Denna analys ger viktiga insikter i faktorer som kan påverka barnets födelsevikt och demonstrerar effektiviteten av både enkel och multipel linjär regression inom forskningsområdet.