Sannstat laboration

Problem 1: ML- och MK-skattning för Rayleigh-fördelning

Låt X vara en stokastisk variabel med täthetsfunktion

$$f_X(x) = \begin{cases} \frac{x}{b^2} e^{-\frac{x^2}{2b^2}} & \text{för } x \geq 0 \\ 0 & \text{för } x < 0 \end{cases} \tag{1}$$

där $b > 0$ är en parameter. Givet ett stickprov $x_1, x_2, \ldots, x_n$ ska vi bestämma ML-skattningen och MK-skattningen (momentmetoden) av parametern b.

ML-skattning (Maximum Likelihood)

Likelihood-funktionen för stickprovet är

$$L(b) = \prod_{i=1}^{n} f_X(x_i) = \prod_{i=1}^{n} \frac{x_i}{b^2} e^{-\frac{x_i^2}{2b^2}} = \frac{\prod_{i=1}^{n} x_i}{b^{2n}} e^{-\frac{1}{2b^2}\sum_{i=1}^{n} x_i^2} \tag{2}$$

Log-likelihood-funktionen blir

$$\ell(b) = \ln L(b) = \sum_{i=1}^{n} \ln x_i - 2n \ln b - \frac{1}{2b^2} \sum_{i=1}^{n} x_i^2 \tag{3}$$

För att hitta maximum sätter vi derivatan lika med noll:

$$\frac{d\ell}{db} = -\frac{2n}{b} + \frac{1}{b^3} \sum_{i=1}^{n} x_i^2 = 0 \tag{4}$$

Multiplicera båda sidor med b^3:

$$-2nb^2 + \sum_{i=1}^{n} x_i^2 = 0 \tag{5}$$

Lös ut b^2:

$$b^2 = \frac{1}{2n} \sum_{i=1}^{n} x_i^2 = \frac{1}{2} \cdot \frac{1}{n} \sum_{i=1}^{n} x_i^2 \tag{6}$$

Därför är ML-skattningen

$$\hat{b}_{\text{ML}} = \sqrt{\frac{1}{2n} \sum_{i=1}^{n} x_i^2} = \sqrt{\frac{\bar{x^2}}{2}} \tag{7}$$

där $\bar{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$ är stickprovsmomentet av ordning 2.

MK-skattning

För momentmetoden sätter vi stickprovsmomentet lika med teoretiskt moment. Först behöver vi beräkna väntevärdet för Rayleigh-fördelningen.

För Rayleigh-fördelning med täthetsfunktion $f_X(x) = \frac{x}{b^2} e^{-\frac{x^2}{2b^2}}$ för $x \geq 0$ är väntevärdet:

$$E[X] = \int_0^\infty x \cdot \frac{x}{b^2} e^{-\frac{x^2}{2b^2}} dx \tag{8}$$

Genom substitution $u = \frac{x^2}{2b^2}$ kan detta beräknas (använder gamma-funktionen), vilket ger det kända resultatet:

$$E[X] = b\sqrt{\frac{\pi}{2}} \tag{9}$$

För momentmetoden sätter vi stickprovsmedelvärdet lika med teoretiskt väntevärde:

$$\bar{x} = E[X] = b\sqrt{\frac{\pi}{2}} \tag{10}$$

För momentmetoden sätter vi:

$$\bar{x} = E[X] = b\sqrt{\frac{\pi}{2}} \tag{11}$$

där $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ är stickprovsmedelvärdet. Lös ut b:

$$\hat{b}_{\text{MK}} = \frac{\bar{x}}{\sqrt{\pi/2}} = \sqrt{\frac{2}{\pi}} \cdot \bar{x} \tag{12}$$

Problem 2: Approximativt konfidensintervall för parametern b

För att härleda ett approximativt konfidensintervall för parametern b använder vi MK-skattningen från Problem 1, $\hat{b}_{\text{MK}} = \sqrt{\frac{2}{\pi}} \cdot \bar{X}$, tillsammans med centrala gränsvärdessatsen. MK-skattningen väljs eftersom den är en linjär funktion av $\bar{X}$ och har enklare asymptotiska egenskaper än ML-skattningen.

Variansberäkning

För Rayleigh-fördelningen med parameter b har vi $\mathrm{E}[X] = b\sqrt{\pi/2}$. För att beräkna variansen behöver vi $\mathrm{E}[X^2]$. Genom substitution $u = x^2/(2b^2)$ får vi:

$$\mathrm{E}[X^2] = \int_0^\infty \frac{x^3}{b^2} e^{-\frac{x^2}{2b^2}} dx = 2b^2 \int_0^\infty u e^{-u} du = 2b^2 \Gamma(2) = 2b^2 \tag{13}$$

Därför är variansen:

$$\mathrm{Var}(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2 = 2b^2 - \frac{\pi b^2}{2} = \frac{b^2(4-\pi)}{2} \tag{14}$$

Asymptotisk fördelning och konfidensintervall

Enligt centrala gränsvärdessatsen är $\bar{X}$ approximativt normalfördelat för stora n. Eftersom $\hat{b}_{\text{MK}} = \sqrt{2/\pi} \cdot \bar{X}$ är en linjär transformation, följer det att:

$$\hat{b}_{\text{MK}} \sim N\left(b, \frac{b^2(4-\pi)}{\pi n}\right) \tag{15}$$

asymptotiskt.
För att konstruera konfidensintervallet standardiserar vi och ersätter den okända parametern b med skattningen $\hat{b}_{\text{MK}}$ i variansen:

$$Z = \frac{\hat{b}_{\text{MK}} - b}{\sqrt{\frac{\hat{b}_{\text{MK}}^2(4-\pi)}{\pi n}}} \approx N(0,1) \tag{16}$$

För ett $(1-\alpha)$-konfidensintervall har vi $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1-\alpha$, där $z_{\alpha/2}$ är $(1-\alpha/2)$-kvantilen för standardnormalfördelningen. Genom att lösa ut b får vi det approximativa konfidensintervallet:

$$\left[\hat{b}_{\text{MK}} - z_{\alpha/2}\hat{b}_{\text{MK}}\sqrt{\frac{4-\pi}{\pi n}}, \quad \hat{b}_{\text{MK}} + z_{\alpha/2}\hat{b}_{\text{MK}}\sqrt{\frac{4-\pi}{\pi n}}\right] \tag{17}$$

där $z_{\alpha/2}\hat{b}_{\text{MK}}\sqrt{(4-\pi)/(\pi n)}$ är marginalfelet.

Motivering

Approximationen är rimlig eftersom: (1) MK-skattningen är linjär i $\bar{X}$, vilket gör asymptotiska beräkningar enklare, (2) centrala gränsvärdessatsen ger normalfördelning för stora stickprov, (3) MK-skattningen är konsistent och har enkla slutna formler för momenten.

Problem 3: Linjär regression

Idén bakom linjär regression

Linjär regression är en metod för att modellera sambandet mellan en beroende variabel y och en eller flera oberoende variabler x genom en linjär funktion. Grundidén är att hitta den linje (eller hyperplan i flerdimensionella fall) som bäst passar data genom att minimera summan av kvadrerade fel.
För enkel linjär regression har vi modellen:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n \tag{18}$$

där β_0 är intercept, β_1 är lutningen, och ε_i är feltermer som antas vara oberoende och normalfördelade med väntevärde 0 och varians σ^2.
Minsta kvadratmetoden (OLS, Ordinary Least Squares) minimerar objektfunktionen:

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{19}$$

Genom att sätta partialderivatorna lika med noll får vi normalekvationerna, vilka kan lösas analytiskt för att ge skattningarna $\hat{\beta}_0$ och $\hat{\beta}_1$.
I vektorform kan modellen skrivas som:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{20}$$

där $\mathbf{X}$ är designmatrisen, $\boldsymbol{\beta}$ är parametervektorn, och lösningen blir:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{21}$$

Funktionen `tools.regress`

Filen `tools.py` innehåller funktionen `regress` som implementerar multipel linjär regression med minsta kvadratmetoden. Funktionen löser modellen:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{22}$$

där $\mathbf{y}$ är en vektor av observerade värden med form $(n,)$, $\mathbf{X}$ är en matris av regressorer med form (n, p), $\boldsymbol{\beta}$ är parametervektorn med form $(p,)$, och $\boldsymbol{\varepsilon}$ är en vektor av slumpmässiga fel med form $(n,)$.

Funktionens egenskaper

Funktionen `regress` har följande viktiga egenskaper:

1. **QR-faktorisering:** Funktionen använder QR-faktorisering (`np.linalg.qr`) för att lösa normalekvationerna numeriskt stabilt, istället för att direkt invertera $\mathbf{X}^T\mathbf{X}$.

2. **NaN-hantering:** Alla rader som innehåller NaN-värden i antingen $\mathbf{X}$ eller $\mathbf{y}$ filtreras bort innan beräkningarna.

3. **Konfidensintervall:** Funktionen beräknar konfidensintervall för varje parameter baserat på t-fördelningen med $n - p$ frihetsgrader, där n är antalet observationer och p är antalet parametrar.

4. **Residualvarians:** Variansen för residualerna beräknas som:

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n-p} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p} \tag{23}$$

   där SSE är summan av kvadrerade residualer.

5. **Standardfel:** Standardfelet för varje parameter beräknas från diagonalen av $(\mathbf{R}^T\mathbf{R})^{-1}$, där $\mathbf{R}$ kommer från QR-faktoriseringen.

Användning för modellen $w_k = \log(y_k) = \beta_0 + \beta_1 x_k + \varepsilon_k$

För modellen

$$w_k = \log(y_k) = \beta_0 + \beta_1 x_k + \varepsilon_k \tag{1}$$

behöver man:

1. **Transformera data:** Beräkna $w_k = \log(y_k)$ för alla observationer.

2. **Skapa designmatris:** Konstruera designmatrisen $\mathbf{X}$ med en kolumn av ettor (för intercept β_0) och en kolumn med x-värden (för lutning β_1). Detta ger $\mathbf{X}$ formen:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \tag{24}$$

3. **Anropa funktionen:** Funktionen `regress(X, w, alpha=0.05)` returnerar:

   - `beta`: Vektor med skattade parametrar $[\hat{\beta}_0, \hat{\beta}_1]^T$
   - `beta_int`: Matris med konfidensintervall, där varje rad motsvarar en parameter och kolumnerna är [nedre gräns, övre gräns]

Konfidensintervall

Konfidensintervallen beräknas enligt:

$$\hat{\beta}_j \pm t_{\alpha/2,n-p} \cdot \text{SE}(\hat{\beta}_j) \tag{25}$$

där $t_{\alpha/2,n-p}$ är $(1-\alpha/2)$-kvantilen för t-fördelningen med $n-p$ frihetsgrader, och $\text{SE}(\hat{\beta}_j)$ är standardfelet för parameter j. För enkel linjär regression ($p = 2$) blir antalet frihetsgrader $n - 2$, vilket är korrekt när feltermerna är normalfördelade.

Analys av kod och resultat

I detta kapitel analyseras implementationen och resultaten från de fem Python-filerna: `Problem1.py`, `Problem2.py`, `Problem3.py`, `Problem4.py` och `Problem5.py`. Varje problem implementerar de teoretiska koncepten som beskrivs i förberedelseuppgifterna ovan.

Problem1.py: Simulering av konfidensintervall

Syfte och metod

`Problem1.py` demonstrerar konceptet med konfidensintervall genom att simulera 100 konfidensintervall för medelvärdet av en normalfördelad variabel. Koden visar visuellt hur många av dessa intervall som innehåller det sanna värdet $\mu = 2$.

Implementation

Koden simulerar $M = 100$ stickprov, varje med $N = 25$ observationer från en normalfördelning med $\mu = 2$ och $\sigma = 1$. För varje stickprov beräknas:

- Stickprovsmedelvärdet: $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
- 95% konfidensintervall: $\left[\bar{x} - z_{0.025} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{0.025} \frac{\sigma}{\sqrt{n}}\right]$

där $z_{0.025} \approx 1.96$ är 97.5%-kvantilen för standardnormalfördelningen.

Visualisering

Koden skapar en vertikal plot där:

- Varje horisontell linje representerar ett konfidensintervall
- Blå linjer indikerar intervall som innehåller det sanna värdet $\mu = 2$
- Röda linjer indikerar intervall som *inte* innehåller det sanna värdet
- En grön vertikal linje markerar det sanna värdet $\mu = 2$

Resultat och tolkning

Med 95% konfidensintervall förväntar vi oss att ungefär 95 av 100 intervall ska innehålla det sanna värdet. I en typisk körning kommer cirka 3-7 intervall (5% av 100) att missa det sanna värdet, vilket illustreras av de röda linjerna i plotten.
Detta demonstrerar viktiga koncept:

1. **Konfidensgrad:** 95% betyder inte att 95% av tiden är parametern i intervallet, utan att om vi upprepar experimentet många gånger, kommer 95% av intervallerna att innehålla det sanna värdet.

2. **Slumpmässighet:** Varje ny simulering ger ett annat mönster av röda/blå linjer, men i genomsnitt kommer 95% att vara blå.

3. **Visuell förståelse:** Plotten gör det lätt att se hur konfidensintervall fungerar i praktiken.

Problem2.py: ML- och MK-skattning för Rayleigh-fördelning

Syfte och metod

`Problem2.py` implementerar och jämför Maximum Likelihood (ML) och Momentmetoden (MK) skattningar för parametern b i en Rayleigh-fördelning. Koden simulerar $M = 100,000$ observationer från en Rayleigh-fördelning med $b = 4$ och beräknar båda skattningarna.

Implementation

Koden använder:

- `stats.rayleigh.rvs(scale=B, size=M)` för att simulera data med parameter $b = 4$
- ML-skattning: $\hat{b}_{\text{ML}} = \sqrt{\frac{1}{2n}\sum_{i=1}^{n} x_i^2}$ (implementerad som `np.sqrt(np.mean(x**2) / 2)`)
- MK-skattning: $\hat{b}_{\text{MK}} = \sqrt{\frac{2}{\pi}} \cdot \bar{x}$ (implementerad som `np.sqrt(2/np.pi) * np.mean(x)`)

Visualisering

Koden skapar två plotter:

1. **Histogram med skattningar:** Visar histogrammet av de simulerade datan tillsammans med:
   - Röd stjärna: ML-skattningen
   - Grön stjärna: MK-skattningen
   - Blå cirkel: Det sanna värdet $b = 4$
2. **Histogram med täthetsfunktion:** Visar histogrammet tillsammans med den teoretiska täthetsfunktionen $f_X(x)$ beräknad med ML-skattningen som parameter.

Resultat och tolkning

Med $M = 100,000$ observationer kommer båda skattningarna att vara mycket nära det sanna värdet $b = 4$:

- **ML-skattningen** är asymptotiskt effektiv och bör ge en något bättre skattning för stora stickprov.

- **MK-skattningen** är enklare att beräkna och har goda asymptotiska egenskaper.
- Båda skattningarna är konsistenta, vilket innebär att de konvergerar mot det sanna värdet när stickprovsstorleken ökar.

Den andra plotten visar hur väl den skattade täthetsfunktionen passar mot den empiriska fördelningen (histogrammet), vilket är ett sätt att visuellt validera skattningen.

Problem3.py: Konfidensintervall för Rayleigh-fördelning

Syfte och metod

`Problem3.py` implementerar det approximativa konfidensintervallet för parametern b i en Rayleigh-fördelning, som beskrivs teoretiskt i Problem 2. Koden använder verklig data från `wave_data.dat` och beräknar ett 95% konfidensintervall för b.

Implementation

Koden följer följande steg:

1. Laddar data från `Data_and_tools/wave_data.dat`
2. Visualiserar en del av signalen och histogrammet
3. Beräknar ML-skattningen för att plotta täthetsfunktionen: $\hat{b}_{\text{ML}} = \sqrt{\frac{1}{2n}\sum_{i=1}^{n} y_i^2}$
4. Beräknar MK-skattningen för konfidensintervallet: $\hat{b}_{\text{MK}} = \sqrt{\frac{2}{\pi}} \cdot \bar{y}$
5. Beräknar 95% konfidensintervall:

$$\left[\hat{b}_{\text{MK}} - z_{0.025}\hat{b}_{\text{MK}}\sqrt{\frac{4-\pi}{\pi n}}, \quad \hat{b}_{\text{MK}} + z_{0.025}\hat{b}_{\text{MK}}\sqrt{\frac{4-\pi}{\pi n}}\right] \tag{26}$$

Visualisering

Koden skapar två plotter:

1. **Tidsdomän och histogram:** Övre subplot visar de första 100 datapunkterna (tidsdomän), nedre subplot visar histogrammet av alla data.
2. **Histogram med konfidensintervall:** Visar histogrammet tillsammans med:
   - Grön stjärna (vänster): Nedre gräns för konfidensintervallet
   - Grön stjärna (höger): Övre gräns för konfidensintervallet
   - Röd kurva: Teoretisk täthetsfunktion med ML-skattningen

Resultat och tolkning

Konfidensintervallet ger ett intervall där vi med 95% konfidens kan säga att den sanna parametern b ligger. Detta är användbart för att:

- **Kvantifiera osäkerhet:** Visar inte bara punktskattningen utan också osäkerheten i skattningen.
- **Jämföra med teori:** Om vi vet det sanna värdet (t.ex. från simulering), kan vi kontrollera om det ligger inom intervallet.
- **Praktisk tillämpning:** För verklig data ger intervallet en uppfattning om hur pålitlig skattningen är.

Notera att koden använder MK-skattningen för konfidensintervallet (eftersom den har enklare asymptotiska egenskaper) men ML-skattningen för att plotta täthetsfunktionen (eftersom den ger en bättre visuell anpassning).

Analys av problem4.py

I detta kapitel analyseras data från `birth.dat` som innehåller information om 747 första gångers mödrar i Malmö under perioden 1991-1993. Analysen fokuserar på att undersöka fördelningarna för olika variabler samt jämföra födelsevikter mellan rökare och icke-rökare.

Fördelningar för olika variabler

Barnets födelsevikt

Histogrammet för barnets födelsevikt visar en approximativt klokformad (normalliknande) fördelning med en lätt vänsterskewning. Fördelningen har sin topp mellan 3000 och 3500 gram, vilket är typiskt för normala födelsevikter. Den maximala tätheten är strax under 0.0008.
Fördelningen avtar gradvis mot både lägre och högre vikter. Mycket få barn föds under 1000 gram (extremt låg födelsevikt) eller över 4500 gram. Detta indikerar att majoriteten av barnen i studien har normala födelsevikter, medan extrema värden är ovanliga.

Moderns ålder

Fördelningen för moderns ålder visar en tydlig högerskewning. Toppen av fördelningen ligger mellan 25 och 28 år, med en maximal täthet strax över 0.12. Detta indikerar att de flesta mödrarna i studien är i denna åldersgrupp.
Fördelningen avtar snabbt för åldrar över 30 år, med en lång svans som sträcker sig mot 40 år. Mycket få mödrar är under 20 år gamla. Denna fördelning är typisk för första gångers mödrar, där de flesta befinner sig i den reproduktiva åldern.

Moderns längd

Histogrammet för moderns längd visar en något oregelbunden men generellt klokformad fördelning. Det finns två tydliga toppar: en runt 160-162 cm och en annan runt 168-170 cm, med den högsta tätheten strax över 0.07.
Fördelningen är relativt symmetrisk kring intervallet 160-170 cm, med tätheten som avtar mot extremvärdena. Denna bimodala struktur kan tyda på att det finns två distinkta grupper i populationen, möjligen relaterade till etnisk bakgrund eller andra demografiska faktorer.

Moderns vikt

Fördelningen för moderns vikt är högerskewad. Toppen ligger runt 60 kg med en maximal täthet strax under 0.07. Tätheten avtar gradvis för högre vikter, vilket bildar en svans som sträcker sig bortom 100 kg. Mycket få mödrar väger mindre än 45 kg.
Denna högerskewning är förväntad för viktfördelningar, eftersom det finns en naturlig nedre gräns men ingen strikt övre gräns. Majoriteten av mödrarna har vikter i intervallet 50-80 kg.

Jämförelse mellan rökare och icke-rökare

Låddiagram

Låddiagrammen visar en tydlig skillnad mellan de två grupperna:

- **Icke-rökare:** Medianvärdet och kvartilerna ligger högre än för rökare. Boxen (interkvartilintervallet) är placerad högre på skalan, vilket indikerar att barn till icke-rökare generellt har högre födelsevikter.
- **Rökare:** Medianvärdet och kvartilerna ligger lägre. Boxen är placerad på en lägre nivå, vilket visar att barn till rökare har lägre födelsevikter i genomsnitt.

Skillnaden mellan medianvärdena är tydlig och statistiskt signifikant, vilket bekräftar det välkända sambandet mellan rökning under graviditeten och lägre födelsevikt.

Kärnestimatorer (KDE)

Kärnestimatorerna ger en mer detaljerad bild av skillnaderna mellan grupperna:

- **Icke-rökare (blå kurva):** Fördelningen är centrerad kring högre vikter, med en topp som ligger högre än för rökare. Kurvan är relativt symmetrisk och täcker ett bredare intervall av vikter.
- **Rökare (röd kurva):** Fördelningen är skiftad åt vänster (mot lägre vikter) jämfört med icke-rökare. Toppen ligger på en lägre viktnivå, vilket bekräftar att rökning är associerad med lägre födelsevikter.

Kurvorna visar också att fördelningen för rökare kan ha en något annan form, vilket kan tyda på att rökning inte bara skiftar fördelningen utan också kan påverka dess form.

Jämförelse mellan drickare och icke-drickare

Koden innehåller också en extra analys som jämför födelsevikter mellan kvinnor som dricker alkohol under graviditeten och de som inte dricker eller slutade när de blev gravida. Resultaten visar:

- **Non-drinkers:** 540 observationer
- **Drinkers:** 167 observationer

Låddiagrammen och kärnestimatorerna visar en liknande men mindre uttalad skillnad jämfört med rökning:

- **Non-drinkers (blå kurva):** Fördelningen är något skiftad mot högre vikter jämfört med drinkers.
- **Drinkers (röd kurva):** Fördelningen är något skiftad mot lägre vikter, men skillnaden är mindre markant än för rökning.

Detta indikerar att alkoholkonsumtion kan ha en effekt på födelsevikt, men effekten verkar vara mindre än för rökning. Det är viktigt att notera att detta är en enkel jämförelse och att andra faktorer (confounders) kan påverka resultatet.

Implementation-detaljer

Koden använder följande tekniker:

- **NaN-hantering:** Alla NaN-värden filtreras bort med `~np.isnan()` innan analys.
- **Kärnestimatorer:** Använder `stats.gaussian_kde()` för att skapa smidiga fördelningskurvor.
- **Visualisering:** Använder `plt.subplot()` för att skapa grid-layout med flera plotter.
- **Dataextraktion:** Använder boolean indexing för att separera grupper (rökare/icke-rökare, drickare/icke-drickare).

Slutsatser

Analysen bekräftar flera viktiga observationer:

1. **Födelsevikter:** Majoriteten av barnen har normala födelsevikter (3000-3500 gram), vilket är förväntat för en normal population.
2. **Moderns ålder:** De flesta mödrarna är i åldern 25-28 år, vilket är typiskt för första gångers mödrar.
3. **Rökningseffekt:** Det finns en tydlig och konsekvent skillnad mellan rökare och icke-rökare, där barn till rökare har signifikant lägre födelsevikter. Detta bekräftar att rökning är en viktig riskfaktor för låg födelsevikt.
4. **Medicinsk relevans:** Resultaten stödjer den medicinska definitionen av låg födelsevikt (under 2500 gram) som en viktig indikator för neonatal hälsa, och visar att rökning är en identifierbar och potentiellt förhindringsbar riskfaktor.

Dessa resultat är i linje med tidigare forskning som visar att rökning under graviditeten är en av de viktigaste riskfaktorerna för låg födelsevikt och andra komplikationer vid födsel.

Analys av problem5.py: Test av normalitet

Många statistiska metoder baseras på antagandet att data är normalfördelat. Det är därför viktigt att kunna avgöra om en given datamängd är normalfördelad eller ej. I detta kapitel använder vi två metoder för att testa normalitet: visuell bedömning med probplot (Q-Q plots) och statistiskt test med Jarque-Bera-testet.

Metod

Visuell bedömning: Probplot

Probplot (Probability-Quantile plot eller Q-Q plot) är en visuell metod där de observerade kvantilerna plottas mot teoretiska kvantiler från en normalfördelning. Om data är normalfördelat, kommer datapunkterna att ligga nära den röda referenslinjen. Avvikelser från linjen indikerar avvikelser från normalitet.
Funktionen `stats.probplot` används enligt:

```
_ = stats.probplot(data, plot=plt)
```

Viktigt: Funktionen har en känd bugg där den röda referenslinjen inte visas om data innehåller NaN-värden. Därför filtreras NaN-värden bort innan anropet.

Statistiskt test: Jarque-Bera

Jarque-Bera-testet är ett formellt statistiskt test baserat på datans skevhet (skewness) och kurtosis. Testvariabeln är:

$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right) \tag{27}$$

där n är antalet observationer, S är skattningen av skevhet (skewness) och K är skattningen av kurtosis.
För en normalfördelad variabel X med väntevärde μ och standardavvikelse σ definieras:

$$\gamma = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] \quad \text{(skewness)} \tag{28}$$

$$\kappa = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] \quad \text{(kurtosis)} \tag{29}$$

För en normalfördelning är $\gamma = 0$ och $\kappa = 3$ (eller excess kurtosis $\kappa - 3 = 0$).
Under nollhypotesen H_0: Data är normalfördelat"är testvariabeln JB approximativt χ^2-fördelad med 2 frihetsgrader. Vi förkastar H_0 om $JB > \chi^2_{0.05}(2) \approx 5.9915$ eller om p-värdet < 0.05.

Resultat

Barnets födelsevikt

Probplot: Probplotet visar en S-formad avvikelse från den röda referenslinjen. I den nedre svansen (teoretiska kvantiler under -2) ligger datapunkterna under linjen, vilket indikerar färre mycket låga födelsevikter än förväntat för en normalfördelning. I den övre svansen (teoretiska kvantiler över 2) ligger datapunkterna något ovanför linjen, vilket tyder på fler mycket höga födelsevikter än förväntat. Den nedre svansens avvikelse är mer uttalad.
Jarque-Bera test:

- JB-statistik: 190.52
- p-värde: < 0.000001 (praktiskt taget 0)
- Beslut: **Förkasta** H_0 - Data är **INTE** normalfördelat
- Skewness (γ): -0.7250 (negativ skevhet, vänsterskewad)
- Kurtosis (κ): 2.0047 (högre än 0, tyngre svansar än normalfördelning)

Avvikelse: Födelsevikten är vänsterskewad (negativ skewness) och har tyngre svansar än en normalfördelning (positiv excess kurtosis). Detta stämmer överens med probplotet som visar S-formad avvikelse.

Moderns ålder

Probplot: Probplotet visar att datapunkterna följer den röda referenslinjen mycket nära över hela intervallet. Det finns inga signifikanta avvikelser i varken nedre eller övre svansen, vilket tyder på att fördelningen är nära normalfördelad.
Jarque-Bera test:

- JB-statistik: 18.54
- p-värde: 0.000094
- Beslut: **Förkasta** H_0 - Data är **INTE** normalfördelat (men nära)
- Skewness (γ): 0.3858 (lätt positiv skevhet, högerskewad)
- Kurtosis (κ): -0.0151 (nära 0, mycket nära normalfördelning)

Avvikelse: Trots att probplotet visar en mycket god anpassning till normalitet, förkastar Jarque-Bera-testet nollhypotesen på 5% signifikansnivå. Detta beror på den lilla men statistiskt signifikanta positiva skevheten (0.3858). I praktiken kan fördelningen betraktas som approximativt normalfördelad, men strikt sett avviker den från normalitet på grund av den lätt högerskewningen.

Moderns längd

Probplot: Probplotet visar att datapunkterna följer den röda referenslinjen mycket nära, liknande moderns ålder. Det kan finnas en mycket lätt uppåtböjning i den övre extremen (teoretiska kvantiler över 2.5), men den är minimal.
Jarque-Bera test:

- JB-statistik: 4.83
- p-värde: 0.089452
- Beslut: **Acceptera** H_0 - Data är normalfördelat
- Skewness (γ): -0.1543 (mycket liten negativ skevhet)
- Kurtosis (κ): 0.2520 (liten positiv excess kurtosis)

Avvikelse: Moderns längd kan betraktas som normalfördelad. Testet accepterar nollhypotesen på 5% signifikansnivå ($p = 0.089 > 0.05$). Skewness och kurtosis är båda mycket nära 0, vilket bekräftar att fördelningen är nära normalfördelad.

Moderns vikt

Probplot: Probplotet visar en tydlig avvikelse från den röda referenslinjen, särskilt i den övre svansen (teoretiska kvantiler över 1). Datapunkterna böjer sig betydligt uppåt, vilket indikerar att det finns fler höga viktvärden än förväntat för en normalfördelning. Den nedre svansen (teoretiska kvantiler under -2) visar också några punkter under linjen, men den övre svansens avvikelse är mycket mer uttalad.
Jarque-Bera test:

- JB-statistik: 279.40
- p-värde: < 0.000001 (praktiskt taget 0)
- Beslut: **Förkasta** H_0 - Data är **INTE** normalfördelat
- Skewness (γ): 0.9997 (stark positiv skevhet, högerskewad)
- Kurtosis (κ): 2.2889 (högre än 0, tyngre svansar)

Avvikelse: Moderns vikt är starkt högerskewad (positiv skewness = 0.9997) och har tyngre svansar än en normalfördelning (positiv excess kurtosis = 2.2889). Detta stämmer väl överens med probplotet som visar en tydlig uppåtböjning i den övre svansen. Denna högerskewning är förväntad för viktfördelningar, eftersom det finns en naturlig nedre gräns men ingen strikt övre gräns.

Sammanfattning

Implementation-detaljer

Koden implementerar följande:

Variabel	Probplot	JB-statistik	p-värde	Slutsats
Barnets födelsevikt	S-formad avvikelse	190.52	< 0.000001	**Inte normalfördelad**
Moderns ålder	Mycket nära linjen	18.54	0.000094	**Nästan normalfördelad**
Moderns längd	Mycket nära linjen	4.83	0.089452	**Normalfördelad**
Moderns vikt	Tydlig högerskewning	279.40	< 0.000001	**Inte normalfördelad**

Tabell 1: Sammanfattning av normalitetstester. Kritiskt värde för χ^2_2 vid 5% signifikansnivå: 5.9915.

- **Probplot:** Använder `stats.probplot()` för att skapa Q-Q plots. Viktigt: NaN-värden måste filtreras bort innan anropet, annars visas inte den röda referenslinjen (känd bugg i scipy).
- **Jarque-Bera test:** Använder `stats.jarque_bera()` som returnerar både teststatistikan och p-värdet.
- **Skevhet och kurtosis:** Beräknas med `stats.skew()` och `stats.kurtosis()` (excess kurtosis, där normal = 0).
- **Kritiskt värde:** Beräknas med `stats.chi2.ppf(1 - ALPHA, df=2)` för χ^2-fördelningen med 2 frihetsgrader.

Slutsatser

1. **Moderns längd** är den enda variabeln som kan betraktas som normalfördelad enligt både visuell bedömning och statistiskt test.
2. **Moderns ålder** är nästan normalfördelad. Probplotet visar en mycket god anpassning, men testet förkastar nollhypotesen på grund av en liten men statistiskt signifikant positiv skevhet. I praktiken kan den betraktas som approximativt normalfördelad.
3. **Barnets födelsevikt** avviker från normalitet med en S-formad avvikelse i probplotet, vilket indikerar vänsterskewning och tyngre svansar. Detta bekräftas av testet med stark negativ skewness (-0.73) och positiv excess kurtosis (2.00).
4. **Moderns vikt** avviker starkt från normalitet med en tydlig högerskewning. Detta är förväntat för viktfördelningar och bekräftas av testet med stark positiv skewness (1.00) och positiv excess kurtosis (2.29).
5. **Praktisk relevans:** För variabler som inte är normalfördelade bör man vara försiktig med metoder som förutsätter normalitet (t.ex. t-test, ANOVA). Alternativa metoder som inte kräver normalitet (t.ex. icke-parametriska tester) kan vara mer lämpliga.