Matematisk Statistik

SF1910 Tillämpad statistik, HT 2016
Laboration 2 för CSAMHS, CLGYM-TEMI

Introduktion

Detta är handledningen till Laboration 2, ta med en utskriven kopia av den till laborationen. Läs handledningen två gånger. Försäkra dig om att du förstår hur de MATLAB-kommandon som finns i den bifogade koden fungerar. Laborationen bedöms som godkänd eller ej godkänd. För att få deltaga i laborationen skall svar på förberedelseuppgifter kunna redovisas **individuellt**. Arbete i grupp är tillåtet (och uppmuntras) med **högst två** personer per grupp. Godkänd laboration ger 3 poäng till ordinarie tentamenstillfälle.

1 Förberedelseuppgifter

1. Rita täthetsfunktionerna för följande fördelningar
   (a) $N(0,1), N(-1,10), N(100,0.01)$.
   (b) $Exp(1), Exp(2), Exp(10)$.
   (c) $\Gamma(1,2), \Gamma(5,1)$.
2. Reflektera över vad som är karakteristiskt för normalfördelade data.

   Kommentar: .....
   .....
   .....
3. Specificera ett $1-\alpha$ konfidensintervall för μ om data består av n oberoende $N(\mu,\sigma)$-fördelade stokastiska variabler där σ är känd/okänd?

   Kommentar: .....
   .....
   .....
4. Definiera likelihood och log-likelihood samt förklara sambandet mellan dessa begrepp. Beskriv idén bakom Minsta-kvadrat (MK) respektive Maximum-likelihood (ML) metoden.

Kommentar: .....
. . . . . .
. . . . . .
. . . . . .

5. När X har täthetsfunktionen

$$f_X(x) = \frac{x}{b^2} e^{-\frac{x^2}{2b^2}}$$

så säger vi att den är Rayleighfördelad. Antag nu att du har n stycken Rayleighfördelade variabler.

- Bestäm ML-skattningen av b.

  Svar: .....

- Bestäm MK-skattningen av b.

  Svar: .....

6. Beskriv hur du kan ta fram ett approximativt konfidensintervall för parametern b. Motivera varför det är rimligt att göra den approximation som du har gjort. Ledning: Använd MK-skattningen.

   Kommentar: .....
   . . . . . .
   . . . . . .
   . . . . . .

7. Beskriv idén bakom linjär regression. Förklara vad polynom-regression är.

   Kommentar: .....
   . . . . . .
   . . . . . .
   . . . . . .

8. Beskriv hur man i MATLAB mha kommandot `regress` kan skatta parametrarna i modellen

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

9. Förklara idén bakom bootstrap. Läs sid. 272-273 om bootstrap i läroboken om nödvändigt.

Vidare introduktion

Börja laborationen med att ladda ner följande filer frå n kurshemsidan.

- `wave_data.mat`
- `resistorer.mat`
- `moore.mat`
- `poly.mat`
- `birth.dat`
- `birth.txt` - beskrivning av datat `birth.dat`

Se till att de ligger i den mapp du kommer att arbeta i. För att kontrollera att du har lagt filerna rätt, skriv `ls *.*at` och se om filerna ovan listas.
Du kan skriva dina kommandon direkt i MATLAB-prompten men det är absolut att föredra att arbeta i editorn. Om den inte är öppen så kan du öppna den och skapa ett nytt dokument genom att skriva `edit lab2.m`. Koden som ges nedan är skriven i celler. En ny cell påbörjas genom att skriva två procenttecken. `Ctrl+Enter` exkeverar innehållet i en cell.

Problem 1 - Maximum likelihood/Minsta kvadrat

Scriptet nedan generar en samling Rayleigh-fördelade stokastiska variabler och plottar sedan skattningen `my_est` Använd dina två skattningar från förberdelseuppgift 5.

```
%% Problem 1: Maximum likelihood/Minsta kvadrat
    M = 1e4;
    b = 4;
    x = raylrnd(b, M, 1);
    hist_density(x, 40)
    hold on
    my_est_ml = % Skriv in din ML-skattning har
    % my_est_mk =
    plot(my_est, 0, 'r*')
    plot(b, 0, 'ro')
    hold off
```

Ser din skattning bra ut?

Kommentar: .....
.....

Kontrollera hur täthetsfunktionen ser ut genom att plotta den med din skattning:

```
%% Problem 1: Maximum likelihood/Minsta kvadrat (forts.)
    plot(0:0.1:6, raylpdf(0:0.1:6, my_est), 'r')
    hold off
```

Problem 2 - Konfidensintervall

I detta avsnitt kommer en Rayleigh-fördelad signal att undersökas; parameter och konfidensintervall för denna skall skattas. Ladda in data genom att skriva `load wave_data.mat`. Filen innehåller en signal som du kan plotta genom att skriva följande

```
%% Problem 2: Konfidensintervall
    load wave_data.mat
    subplot(211), plot(y(1:100))
    subplot(212), hist_density(y)
```

Om du ändrar `y(1:100)` till `y(1:end)` så kan du se hela signalen.
Skatta parametern på datat på samma sätt som i föregående uppgift. Spara din skattning som `my_est`. Ta fram ett konfidensintervall för skattningen och spara övre respektive undre värdet som `upper_bound` respektive `lower_bound`. Skriv ner dina resulat:

Svar: .....
.....
.....
Plota nu intervallet för din skattning av parametern

```
%% Problem 2: Konfidensintervall (forts.)
    hold on     % Gor sa att ploten halls kvar
    plot(lower_bound, 0, 'g*')
    plot(upper_bound, 0, 'g*')
```

Kontrollera hur täthetsfunktionen ser ut genom att plotta den med din skattning på samma vis som i föregående avsnitt:

```
%% Problem 2: Konfidensintervall (forts.)
    plot(0:0.1:6, raylpdf(0:0.1:6, my_est), 'r')
    hold off
```

Ser fördelningen ut att passa bra?

Svar: .....

Rayleighfördelningen kan t ex användas för att beskriva hur en radiosignal

avtar. Experimentella mätningar på Manhattan har visat att Rayleighfördelningen beskriver radiosignalers fädning (engeleska: fading) på ett bra sätt i den sortens stadsmiljö [1].

Problem 3 - Simulering av konfidensintervall

Ett $1-\alpha$ konfidensintervall för parametern μ täcker den sanna (okända) μ med sannolikhet $1-\alpha$. Syftet med uppgiften är att förstå innebörden av detta begrepp med hjälp av simuleringar. Följande kod använder $n = 25$ oberoende observationer från $N(2,1)$ fördelningen för att skatta ett 95% konfidensintervall för väntevärdet (vi glömmer bort att vi vet vad det sanna värdet är). Detta upprepas 100 gånger så vi har 100 konfidensintervall, hur många av dessa förväntar vi oss ska täcka den sanna parametern?

```
%% Problem 3: Simulering av konfidensintervall
    % Parametrar:
    n = 25; %Antal matningar
    mu = 2; %Vantevardet
    sigma = 1; %Standardavvikelsen
    alpha = 0.05;

    %Simulerar n * 100 observationer. (n observationer for ...
        varje intervall och 100 intervall)
    x = normrnd(mu, sigma,n,100); %n x 100 matris med varden

    %Skattar mu med medelvardet
    xbar = mean(x); %vektor med 100 medelvarden.

    %Beraknar de undre och ovre granserna
    undre = xbar - norminv(1-alpha/2)*sigma/sqrt(n);
    ovre = xbar + norminv(1-alpha/2)*sigma/sqrt(n);

    %Ritar upp alla intervall och markerar de som inte ...
        tacker det sanna vardet roda
    figure(1)
    hold on
    for k=1:100
        if ovre(k) < mu
            plot([undre(k) ovre(k)],[k k],'r')
            elseif undre(k) > mu
            plot([undre(k) ovre(k)],[k k],'r')
        else
            plot([undre(k) ovre(k)],[k k],'b')
        end
    end
    %b1 och b2 ar bara till for att figuren ska se snygg ut.
    b1 = min(xbar - norminv(1 - alpha/2)*sigma/sqrt(n));
    b2 = max(xbar + norminv(1 - alpha/2)*sigma/sqrt(n));
    axis([b1 b2 0 101]) %Minimerar mangden outnyttjat ...
        utrymme i figuren
```

```
34      %Ritar ut det sanna vardet
35      plot([mu mu],[0 101],'g')
36      hold off
```

Vad visar de horisontella strecken och det vertikala strecket? Hur många intervall täcker det sanna värdet av μ? Stämmer resultatet med dina förväntningar? Kör simuleringarna flera gånger och tolka resultatet.

Svar: .....
.....
.....

Ändra nu på μ, σ, n och α (en i taget) och se hur de olika parametrarna på verkar resultatet?

Svar: .....
.....
.....

Problem 4 - Passning av fördelning

Ladda in `resistorer.mat` och studera datat (som beskriver en uppmätt egenskap hos ett antal resistorer) med hjälp av ett histogram. Undersök också hur det ser ut med kommandot `normplot`. Vilken fördelning tror du att resistorernas motstånd har? Är det någon fördelning du kan utesluta? Varför kan man vara intresserad av fördelningen för någon specifik egenskap hos resistorer?

Kommentar: .....
.....

Problem 5a - Linjär regression

Vi kommer att titta på fenomenet som kallas Moores lag. Ladda in datat `moore.mat` på samma sätt som tidigare. I datat så är `y` antal transistorer/yta medan `x` representerar årtal. Det betyder att om vi plottar dem mot varandra så ser vi en plot av utvecklingen över tid av antalet transistorer per yta. Inför modellen

$$w_i = \log(y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i. \tag{2}$$

Skatta β_0 och β_1 med hjälp av MATLABs funktion `regress`.
Om du skattar parametrar mha data från 1971 till 2011, vad är då din prediktion för antalet transistorer år 2020?

Svar: .....

Problem 6b - Polynomregression

Börja med att ladda filen `poly.mat`. Plotta `y1, y2, y3`, var för sig mot `x1, x2`, respektive `x3`. Ser de ut att kunna beskrivas av polynom?

Kommentar: .....
.....

Inför modellen

$$y_k = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_n x^n. \tag{3}$$

Bilda nu, för vart och ett av de tre data-mängderna, en X-matris på ett lämpligt vis. Alltså, studera plottarna och designa sedan ett X sådant att det kan representera ett polynom av den grad du tror passar. I fallet för modellen (3) ovan så ser X ut så här:

$$X = \begin{bmatrix} 1 & x & x^2 & \dots & x^n \\ 1 & x & x^2 & \dots & x^n \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & x & x^2 & \dots & x^n \end{bmatrix}. \tag{4}$$

Alltså bilda din X-matris, ta fram din skattning av $\hat{\beta}$ med hjälp av `regress` och plotta sedan din skattade modell,

$$\hat{y} = X\hat{\beta}, \tag{5}$$

I fallet `y1`, får vi:

```
%% Problem 4: Regression
    y_hat = X*beta_hat;
    plot(y1, '.')
    hold on
    plot(y_hat, 'r.')
    hold off
```

Plotta residualerna

```
%% Problem 4: Regression (forts.)
    res = y_hat - y1;
    subplot(211), normplot(res)
    subplot(212), hist(res)
```

Vilken fördelning ser de ut att komma från?

Svar: .....

Vad kan du dra för slutsatser om modellen?

Svar: .....

Linjär regression utvecklades under sent 1700-tal av en ung Gauss. Metoden fick ett genomslag när den förutspådde banan för den genom tiderna först upptäckta asteroiden, Ceres. Linjär regression används än flitigare idag med tillämpningar inom i stort sett all vetenskap som behandlar data. Fördjupning i ämnet ges i kursen "Regressionsanalys".

Problem 7 - Bootstrap av skattning av skillnad mellan väntevärden för födelsevikter

Vi skall nu gå vidare till att studera skillnaden mellan väntevärden i två populationer, t ex skillnaden i födelsevikt för barn vars mammor röker respektive inte röker under graviditeten. (Om ni vill kan ni ta två andra populationer, och/eller andra variabler att studera!).
I filen `birth.txt` ser man att kolonn 20 i `birth.txt` innehåller rökvanor, och att värdena 1 och 2 betyder att mamman inte röker under graviditeten, medan värdet 3 betyder att hon gör det. Ni kan skapa två variabler `x` och `y` för födelsevikter hörande till icke-rökande respektive rökande mammor enligt

```
>> x = birth(birth(:, 20) < 3, 3);
>> y = birth(birth(:, 20) == 3, 3);
```

Vad som händer här är att `birth(:, 20) < 3` returnerar en vektor av "sant" och "falskt", och att bara de rader av kolonn 3 (födelsevikterna i `birth` för vilka jämförelsen är sann, väljs ut. Använd funktionen `length` eller kommandot `whos` för att se storleken på vektorerna `x` och `y`.
För att skatta skillnaden mellan populationernas väntevärden, använder vi som vanligt skillnaden mellan stickprovsmedelvärdena,

```
mean(x) - mean(y).
```

För att undersöka osäkerheten i denna skattningen ska vi använda bootstrap och simulera M stycken bootstrapreplikat enligt

```
>> thetaboot = bootstrp(M, @mean, x) - bootstrp(M, @mean, y);
```

Ser bootstrapreplikaten ut att komma från en normalfördelning? Vad får du för konfidensintervall för skillnaden θ mellan väntevärdena? Använd

```
>> quantile(thetaboot, [0.025, 0.975])
```

Vad får du med hjälp av metoden i boken, dvs konfidensintervall för skillnad mellan väntevärden?

Svar: .....

Kommentar: .....
.....

Referenser

[1] Dmitry Chizhik, Jonathan Ling, Peter W. Wolniansky, Reinaldo A. Valenzuela, Nelson Costa, and Kris Huber (2003). Multiple-input-multiple-output measurements and modeling in Manhattan *IEEE Journal on Selected Areas in Communications*, Vol **21**, p. 321-331.