



Matematisk Statistik

SF1901 Sannolikhhetsteori och statistik: VT 2016

Lab 2 för
CTFYS, CELTE

Introduktion

Detta är handledningen till Laboration 2, ta med en utskriven kopia av den till laborationen. 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 delta i laborationen skall svar på förberedelseuppgifter kunna redovisas **individuellt**. Arbete i grupp är tillåtet (och uppmuntras) om **högst två** personer per grupp. Godkänd laboration ger 3 poäng till ordinarie tentamenstillfälle.

1 Förberedelseuppgifter

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

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

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

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

5. Definiera likelihood, log-likelihood, förklara sambandet mellan dessa. Beskriv idén bakom Minsta-kvadrat (MK) respektive Maximum-likelihood (ML) metoden.

Kommentar:

6. 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:

7. Beskriv hur du kan ta fram ett approximativt konfidensintervall för parametern b . Motivera varför det är rimligt att göra denna approximation. Ledning: Använd MK-skattningen.

Kommentar:

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

Kommentar:

9. Beskriv hur man i MATLAB mha kommandot **regress** kan skatta parametrarna i modellen

$$w = \log(y_k) = \beta_0 + \beta_1 x_k + \varepsilon_k \quad (1)$$

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

Vidare introduktion

Börja med att ladda ner filerna

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

från kurshemsidan. Se till att de ligger i den mapp du kommer att arbeta i. För att kontrollera att du har lagt filerna rätt, skriv `ls *.mat` och försäkra dig om att 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`.

Problem 1 - Maximum likelihood/Minsta kvadrat

Scriptet nedan genererar en samling Rayleigh-fördelade stokastiska variabler och plottar sedan skattningen `my_est`. Använd dina två skattningar från förberedelseuppgift 6.

```
1      %% Problem 1: Maximum likelihood/Minsta kvadrat
2      M = 1e4;
3      b = 4;
4      x = raylrnd(b, M, 1);
5      hist_density(x, 40)
6      hold on
7      my_est_ml = % Skriv in din ML-skattning här
8      % my_est_mk =
9      plot(my_est, 0, 'r*')
10     plot(b, 0, 'ro')
11     hold off
```

Ser din skattning bra ut?

Kommentar:

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

```
1 plot(0:0.1:6, raylpdf(0:0.1:6, my_est), 'r')
2 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

```
1 %% Problem 2: Konfidensintervall
2 load wave_data.mat
3 subplot(211), plot(y(1:100))
4 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 resultat:

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

Plota nu intervallet för din skattning av parametern

```
1 % ...
2 hold on % Gör så att ploten halls kvar
3 plot(lower_bound, 0, 'g*')
4 plot(upper_bound, 0, 'g*')
```

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

```
1 % ...
2 plot(0:0.1:6, raylpdf(0:0.1:6, my_est), 'r')
3 hold off
```

Ser fördelningen ut att passa bra?

Svar:

Rayleighfördelningen kan t ex användas för att beskriva hur en radiosignal

avatar. Experimentella mätningar på Manhattan har visat att Rayleighfördelningen beskriver radiosignalers fädning (engelska: 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?

```
1 %% Simulering av konfidensintervall
2 % Parametrar:
3 n = 25; %Antal matningar
4 mu = 2; %Vantevardet
5 sigma = 1; %Standardavvikelsen
6 alpha = 0.05;
7
8 %Simulerar n * 100 observationer. (n observationer for ...
   varje intervall och 100 intervall)
9 x = normrnd(mu, sigma,n,100); %n x 100 matris med varden
10
11 %Skattar mu med medelvardet
12 xbar = mean(x); %vektor med 100 medelvarden.
13
14 %Beraknar de undre och ovre granserna
15 undre = xbar - norminv(1-alpha/2)*sigma/sqrt(n);
16 ovre = xbar + norminv(1-alpha/2)*sigma/sqrt(n);
17
18 %Ritar upp alla intervall och markerar de som inte tacker ...
   det sanna vardet roda
19 figure(1)
20 hold on
21 for k=1:100
22     if ovre(k) < mu
23         plot([undre(k) ovre(k)], [k k], 'r')
24     elseif undre(k) > mu
25         plot([undre(k) ovre(k)], [k k], 'r')
26     else
27         plot([undre(k) ovre(k)], [k k], 'b')
28     end
29 end
30 %b1 och b2 ar bara till for att figuren ska se snygg ut.
31 b1 = min(xbar - norminv(1 - alpha/2)*sigma/sqrt(n));
32 b2 = max(xbar + norminv(1 - alpha/2)*sigma/sqrt(n));
33 axis([b1 b2 0 101]) %Minimerar mangden outnyttjat utrymme i ...
   figuren
```

```
34
35 %Ritar ut det sanna värdet
36 plot([mu mu],[0 101],'g')
37 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) hur påverkar de olika parametrarna resultatet?

Svar:

Problem 4: Passning av fördelning

Ladda in `resistorer.mat`, 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 följande modell

$$w_i = \log(y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad (2)$$

Skatta β_0 och β_1 med hjälp av MATLABs `regress`.

Om du skattar parametrar mha data från 1971 till 2011, vad är då din prediktion för antalet transistorer år 2020?

Svar:

Problem 5b: 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

$$y_k = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_n x^n. \quad (3)$$

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:

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

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

$$\hat{y} = X\hat{\beta}, \quad (5)$$

i fallet `y1`, får vi:

```
1      %% Problem 4: Regression
2      y_hat = X*beta_hat;
3      plot(y1, 'r.')
4      hold on
5      plot(y_hat, 'b.')
6      hold off
```

Plotta residualerna

```
1      res = y_hat - y1;
2      subplot(211), normplot(res)
3      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 av ämnet ges i kursen "Tillämpad Matematisk Statistik".

Problem 6- 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 bootstrapreplikater 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 den metod i boken, dvs konfidensintervall för skillnad mellan väntevärden, som du skulle använda?

Svar:

Kommentar:
.....

Referenser

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