MVE055/MSG810 2017 Föreläsning 7

Petter Mostad

Chalmers

September 18, 2017

Överblick

- Deskriptiv statistik
  - Grafiska sammanfattningar.
  - Numeriska sammanfattningar.
- Estimering (skattning)
  - Teori
  - Några exempel

Deskriptiv statistik

- Att sammanställa och illustrera data kallas *deskriptiv statistik*.
- Syftet med att samla i hop data är generellt att hitta någo sorts "underliggande modell" för data: Vi skall snart precisera detta.
- *Vad* man samlar in och val av *sammanfattning och illustration* av data kan aldrig göras helt objektivt. Processen reflekterar ideer man redan har om vad som är möjliga "underliggande modeller". Det bästa man kan göra är att öppet dokumentera sina val.
- När man samlat i hop data bör man alltid, som första uppgift, göra
  - grafiska plots av data
  - numeriska sammanfattningar av data

  Syftet är att försöka leta efter och jämföra möjliga "underliggande modeller".

Numeriska sammanfattningar av data

- Medelvärde $\overline{x}$.
- Median: Om $x_1, x_2, \ldots, x_n$ är data *sorterat i storleksordning* så är medianen $x_{(n+1)/2}$ om n är udda och medelvärdet av $x_{n/2}$ och $x_{(n+1)/2}$ om n är jämn.
- Stickprovs-variansen:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

  Stickprovs-standardavvikelsen är kvadratroten av detta.
- Min och max.

Grafiska sammanfattningar av data

- I läroboken beskrivs ”stem and leaf” och ”ogives”. Värd att läsa genom, men lite gammaldags i dag.
- Använd de plotningsfunktioner som finns tillgänglig i det dataprogram du använder. Till exempel spridningsdiagram (scatterplots).
- Ett möjligt beräkningsverktyg: R (www.r-project.org).
- Exempel: Histogrammer. Märk att vissa val tas (av dig eller av programmet).
- Kom i håg: Syftet är alltid att hitta möjliga underliggande modeller, eller jämföra olika möjliga modeller.

Boxplots

- Exempel på bra sätt att visualisera en-variata data.
- Visualiserer medianen, och *kvartilerna* (värdena så att 25% av data är mindre än första kvartil och 75% av data är mindre än 3dje kvartil).
- Visualiserar även *outliers*: (avvikande värden).
- Detaljerna för hur Boxplots konstrueras varierar mellan olika program; ni behöver inte kunna detaljerna som finns i Milton.

Stokastiska variabler som modeller för data

Vi kan säga att vi använder en stokastisk variabel X som en modell för en del av verkligheten om

- Vi beskriver en *population* av objekt, eller eventuellt möjliga observationer, där våra data representerar en delmängd av populationen.
- Variabeln X representerar värdena i populationen.
- De observerade data $x_1, \ldots, x_n$ är ett *stickprov* (random sample) från populationen, med andra ord:
  - Varje enskild observation x_i modelleras med variabeln X_i.
  - Alla X_i är *oberoende* och *har samma fördelning som* X.
- MÄRK: Att det är rimligt att anta att data är ett stickprov behöver man alltid kolla:
  - Är varje observation slumpmässigt vald med uniform sannolikhet från populationen?
  - Är varje observation slumpmässigt vald oberoende av andra observationer?
- MÄRK: Syftet med modelleringen (och med all vetenskap) är egentligen att kunna göra *prediktioner* av nya observationer.

Att hitta stokastisk modell för dina data

- Steg ett: Att hitta typ av modell. I vårt fall, att välja mellan normalfördelad, Poissonfördelad, Binomialfördelad, Gammafördelad, etc.:
  - Använd kunskap om processen som genererat data. Är data kontinuerlig eller diskret? En sekvens av "försök"? Etc.
  - Man kan också använda observerade data för att välja mellan liknande modeller: Detta ligger utanför vår kurs.
- Steg två: Hitta parametrarna i modellen. (E.g., μ och σ^2 i en normalfördeling, λ i en Poissonfördelning, etc.):
  - Man beräknar ett *estimat* (en skattning) för varje parameter genom att använda data i någon formel tagit fram för denna parametern i denna modellen.
  - Man kan i stället använda sannolikhetsteori för att beskriva hur man lär sig om parametrarna från data. Detta ligger dock utanför vår kurs.

Estimater och estimatorer

Vi antar nu vi har en modell med en okänd parameter θ, och data $x_1, \ldots, x_n$ som är ett stickprov från modellen.

- Vi skiljer mellan
  - En *estimator*: En funktion, eller formel, som från ett stickprov $X_1, \ldots, X_n$ kan bereäkna ett tal som är tänkt att kunna fungera som värde på parametern.
  - Ett *estimat*: Det värdet man får när man använder estimatorn på våra data $x_1, \ldots, x_n$.
- Vi skriver ofta $\hat{\theta}$ för estimatet för parametern θ, men förvirrande nog så används $\hat{\theta}$ ofta också för att beteckna estimatorn som används för att beräkna estimatet. (Om parametern kallas μ så skriver vi $\hat{\mu}$ etc.)
- En funktion av ett stickprov kallas en *statistika*. En estimator är alltså en statistika ("a statistic")
- En statistika är en stokastisk variabel, så vi kan prata om dens fördelning, väntevärde, varians, etc.

Att konstruera en estimator

- Att använda estimatorer är ett ad-hoc tillvägagångssätt; det finns inget matematiskt entydigt svar på hur de skall konstrueras.
- Dock tycker många att en bra estimator bör ha följande egenskaper:
  - Väntevärdesrätt: Om vi ser estimatorn som en statistika och beräknar dens väntevärde, så skall detta vara lika med parametern man estimerar.
  - Minimum varians: Om vi ser estimatorn som en statistika och beräknar dens varians, så skall denna vara så liten som möjligt.
- (Utanför kursen) Vanligaste sättet att konstruera en estimator: Maximum Likelihood (ML).
  - Skriv sannolikheten av alla observerade data som en funktion av alla parametrar. Detta är *likelihood funktionen* (trolighetsfunktionen).
  - Hitta de parametrar som maximerar likelihood funktionen.
  - Formlerna för parametrarna som maximerar likelihood, beräknat från observerade data, används sen som estimatorer för parametrarna.

Exempel: Estimator för väntevärdet

- Antag att data är ett stickprov från fördelningen Normal(μ, σ), alltså att de representeras med oberoende stokastiska variablar

$$X_1, X_2, \ldots, X_n \sim \text{Normal}(\mu, \sigma)$$

  Vi vill hitta en estimator för μ.
- En naturlig estimator är $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$. (Det är en ML-estimator).
- Estimatorn är väntevärdesrätt, i.e., $\mathbb{E}\left[\overline{X}\right] = \mu$. (Enkelt bevis).
- Beviset fungerar lika bra oberoende av fördelningen till X_i. Så estimatorn $\overline{X}$ är alltid en väntevärdesrätt estimator för väntevärdet i en fördelning.
- Exempel: Antag observationerna $x_1, x_2, \ldots, x_n$ är ett stickprov från en Poisson(λ) fördelning, där ju väntevärdet är lika med parametern λ. Då är $\overline{X}$ en väntevärdesrätt estimator för λ.

Varians av en estimator

- Estimatorn $\overline{X}$ har varians σ^2/n, där σ^2 är variansen till fördelningen för $X_1, \ldots, X_n$.
- Beviset är bra att ha koll på. (MÄRK: Det använder att om X och Y är obereoende stokastiska variabler, så har vi $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$).
- Exempel: Bernoulli-fördelningen (Binomialfördelningen med bara ett försök):
  - Fördelningen har en parameter p och X har möjliga värden 0 och 1.
  - Väntevärdet i fördelningen är p och variansen är $p(1-p)$.
  - $\hat{p} = \overline{X}$ är en väntevärdesrätt estimator för p. Estimatorn har varians $p(1-p)/n$.

Estimator för varians

- Om $X_1, X_2, \ldots, X_n$ är ett stickprov från en fördelning med väntevärde μ och varians σ^2, så är

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

  en väntevärdesrätt estimator för σ^2.

- Beviset är bra att ha koll på:

$$\begin{aligned}\mathbb{E}\left[\hat{\sigma}^2\right] &= \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right] \\ &= \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}((X_i - \mu) - (\overline{X} - \mu))^2\right] = \cdots = \sigma^2\end{aligned}$$

- Detta är orsaken till att vi delar på $n-1$ i stället för n när vi beräknar stickprovsvariansen.

En estimators fördelning

- För att ytterligare lära oss om hur "bra" en estimator är, så kan vi studera dens fördelning, inte bara dens väntevärde och varians.
- Exempel: Om $X_1, \ldots, X_n \sim \text{Normal}(\mu, \sigma)$, så vet vi att estimatorn $\overline{X}$ har väntevärde μ och varians σ^2/n: Också:
  - Man kan visa: Om $X_1, \ldots, X_n$ är oberoende normalfördelade, så är även $X_1 + X_2 + \cdots + X_n$ normalfördelad.
  - Vi vet att om Y är normalfördelad så är även Y/n normalfördelad.

  Från detta får vi att $\overline{X} \sim \text{Normal}(\mu, \sigma/\sqrt{n})$.
- Om $X_1, \ldots, X_n$ har en Bernoulli-fördelning med parameter p, så får vi direkt från definitionerna att $X_1 + X_2 + \cdots + X_n$ har en Binomialfördelning med Parametrar n och p. Därmed kan vi också få en explicit beskrivning av fördelningen till $\overline{X} = (X_1 + X_2 + \cdots + X_n)/n$.
- Man kan visa att om $X_1, \ldots, X_n \sim \text{Normal}(\mu, \sigma)$ så har vi för variansestimatorn $\hat{\sigma}^2$

$$(n-1)\hat{\sigma}^2/\sigma^2 \sim \chi^2(n-1)$$

  Alltså: $\hat{\sigma}^2$ har en fördelning som motsvarar σ^2 multiplicerad med en chi-kvadrat fördelning med $n-1$ frihetsgrader.

Vi studerar estimatorer, inte estimat

- Antag vi undersöker en typ försök som varje gång resulterar i success (1) eller misslyckande (0), och sannolikheten för success är en okänd parameter p. Antag vi gör ett antal försök, och får resultaten

$$0, 1, 0, 0, 1, 0, 0, 1$$

  Vi gör sen estimatet $3/8 = 0.375$ för p.
- Hur bra är estimatet? *Det kan vi inte säga något om med metoderna vi använder nu: Vi studerar estimatorer, inte estimat*.
- Hur bra är estimatorn? För att svara måste vi först specificera exakt vilken estimator vi använt:
- ALTERNATIV 1: Estimatorn består i att göra 8 försök, låta x vara antalet successer, och beräkna $\hat{p} = x/8$.
- ALTERNATIV 2: Estimatorn består i att göra försök tills man fått fram 3 lyckade försök, och låta x vara antalet försök man behövde göra. Sen beräknar man $\hat{p} = 3/x$.
- De två estimatorerna har olika egenskaper! Det är kanske inte naturligt att använda några av dessa för att anslå hur nära 0.375 är p givet våra data.

Exempel, fortsättning

- Låt oss till exempel anta att det verkliga värdet för p är 0.6. Vi kan då studera vilka fördelningar våra två alternativa estimatorer har.
- ALTERNATIV 1: Vi har $X \sim \text{Binomial}(8, 0.6)$. De möjliga värdena för $\hat{p} = X/8$ och deras sannolikheter finns i tabellen under:

0/8	1/8	2/8	3/8	4/8	5/8	6/8	7/8	8/8
0.001	0.008	0.041	0.124	0.232	0.279	0.209	0.090	0.017

- Estimatorn har väntevärde 0.6; den är väntevärdesrätt.
- ALTERNATIV 2: Vi får $X \sim \text{Neg-Binomial}(3, 0.6)$. De möjliga värdena för $\hat{p} = 3/X$ och deras sannolikheter finns i tabellen under:

3/3	3/4	3/5	3/6	3/7	3/8	3/9	3/10	3/11
0.216	0.259	0.207	0.138	0.083	0.046	0.025	0.013	0.006

3/12	3/13	3/14	3/15	3/16,3/17,...
0.003	0.001	0.001	0.000	totalt 0.000

- Estimatorn har väntevärde 0.672. Den är INTE väntevärdesrätt.

Centralgränsteoremet (Central Limit Theorem, CLT)

- Antag $X_1, \ldots, X_n$ är ett stickprov från en fördelning med väntevärde μ och varians σ^2. Då gäller som gräns när $n \to \infty$, att

$$\overline{X} \sim \text{Normal}(\mu, \sigma/\sqrt{n})$$

- För ändliga n kan detta användas som en approximation. Hur stor n behöver vara för att approximationen skall vara OK beror på vilken noggrannhet man behöver, och egenskaperna för fördelningen X kommer ifrån.
- Vissa fördelningar har ingen varians; då gäller heller inte CLT!