IMM, 2002-09-14
Poul Thyregod

4 Oversigt over kapitel 4

Introduktion

Hidtil har vi beskæftiget os med *data*. Når data repræsenterer gentagne observationer (i bred forstand) af et fænomen, kan det være bekvemt at beskrive *fordelingen* af observationer ved en idealiseret model.

I dette kapitel gives den formelle baggrund for de statistiske metoder, der behandles i de følgende kapitler. Således introduceres begrebet *stokastisk variabel* (random variable) og *fordelingsmodel* (distibution model) til modellering af en variabel og af fordelingen af observationer af den variable. Begreberne *middelværdi* og *varians* for en stokastisk variabel indføres i relation til en sådan fordelingsmodel, og der angives nogle regneregler for middelværdier og varianser ved lineære transformationer og ved kombinationer af stokastiske variable. Endelig beskrives nogle hyppigt anvendte fordelingsmodeller, binomial- og Poissonfordelingen for diskrete variable, og *normalfordelingen* for kontinuerte variable.

I de følgende kapitler vil vi specielt interessere os for normalfordelingsmodeller (afsnit 4.5 - 4.7), og til slut i kurset vil vi møde binomialfordelingen igen.

Fordelingsmodellerne beskriver *sandsynlighedsfordelinger*, og derfor indledes kapitlet med en kort introduktion til sandsynlighedsregning.

I kapitlet springer forfatterne lidt frem og tilbage mellem modeller og vurdering af overensstemmelsen mellem data og model. Ved læsningen bør man gøre sig klart, hvornår teksten vedrører modeller, og hvornår den foretager et sidespring og beskriver vurdering af data i relation til en model.

4.1 Density Histogram

Discussion questions .

På dansk bruges bare betegnelsen *histogram* for et sådant "density histogram".

Dette indledende afsnit vedrører *data*, dvs observationer (vi tager tilløb til at formulere en *model*).

4.2 Sandsynlighedsbegrebet

Ordliste:

- **trial**: forsøg
- **Bernoulli trial**: Bernoulli forsøg
- **probability** : sandsynlighed
- **random** : tilfældig
- **random variable**: stokastisk variabel
- **event**: hændelse
- **outcome**: udfald

Den matematisk orienterede læser vil bemærke forbindelsen til mængdelæren. En hændelse er en delmængde i mængden af samtlige mulige udfald, udfaldsrummet Ω.

Definitionerne side 139 er blot en repetition af regneregler fra mængdelæren.

- **intersection**: fællesmængde
- **union**: foreningsmængde
- **disjoint** : disjunkte
- **null event** : nulhændelse, den tomme hændelse

Bemærk at additionsreglen for sandsynligheder side 140 kun mugælder for disjunkte hændelser, og at *uafhængighed* mellem hændelser (side 143) er defineret ved at produktreglen gælder for de tilsvarende sandsynligheder. (Betragt fx kast med to terninger og lad den ene hændelse være, at første terning viser to øjne, og den anden hændelse, at summen af øjnene på de to terninger er syv. Er de to hændelser uafhængige ?).

4.3 Stokastiske variable og deres fordelingsmodeller

Her introduceres begrebet *stokastisk variabel* (side 145), som er den grundlæggende byggesten for statistiske modeller.

Afsnit 4.3 omhandler *diskrete stokastiske variable* og afsnit 4.4 omhandler *kontinuerte stokastiske variable.*

Formålet med afsnit 4.3 er dels at introducere de begreber og resultater, der bruges i forbindelse med fordelingsmodeller, og dels at introducere til nogle grundlæggende diskrete fordelingsmodeller.

I afsnit 4.4 resumeres disse begreber og resultater for kontinuerte fordelingsmodeller og normalfordelingsmodellen introduceres.

Som nævnt vil vi i kursets første del fortrinsvis beskæftige os med normalfordelingsmodeller, men det er alligevel nødvendigt at læse afsnit 4.3 for at blive introduceret til begreberne.

Den formelle definition af en *stokastisk variabel* (random variable) er anført på side 145. Den praktiske betydning er, at den stokastiske variabel angiver "det, vi måler", fx. hjernevægten for et tilfældigt udtrukket pattedyr.
Bemærk specielt, at en stokastisk variabel er en *funktion.* Funktionsværdien svarende til en aktuel observation af den stokastiske variable er det tal, som er resultatet af den aktuelle måling, eksempelvis hjernevægten af den mus, som vi fik fat i, da vi tog et tilfældigt pattedyr.

Populært udtrykt kan man opfatte en stokastisk variabel som et overbegreb for det, vi måler (før vi har foretaget målingen), mens en *observeret værdi* er det tal, der er resultatet af målingen. Det væsentlige er, at en

stokastisk variabel har tilknyttet en *sandsynlighedsfordeling*, mens en observeret værdi bare er et tal.

I bogen symboliseres stokastiske variable ved STORE BOGSTAVER, $(X, Y, Z, \ldots, S^2)$, mens observerede værdier symboliseres ved de tilsvarende små bogstaver, $(x, y, z, \ldots, s^2)$

Egenskaber ved stokastiske variable udtrykkes ved egenskaber for den *sandsynlighedsfordeling*, som er knyttet til den pågældende stokastiske variable, fx *middelværdi* (forventningsværdi), *varians*, *standardafvigelse* (spredning).

En sandsynlighedsfordeling er karakteriseret ved sin tæthedsfunktion, eller ved sin fordelingsfunktion (cumulative distribution function). En tæthedsfunktion kan opfattes som beskrivelse af, hvorledes den totale sandsynlighedsmasse er fordelt på punkterne på den reelle akse.
Fortsætter man i billedet med massefordeling, kan middelværdi netop fortolkes som *tyngdepunkt*, og varians som *inertimoment*.

Definitionerne for middelværdi og varians for diskrete (side 156) og for kontinuerte (side 150) stokastiske variable adskiller sig bare ved at for diskrete variable summerer man hen over alle de mulige værdier, men for kontinuerte variable udgør de mulige værdier et kontinuum, så derfor er man nødt til at integrere.

Notation

Notationen kan godt være lidt forvirrende:
Dels optræder symbolerne μ og σ^2 som *parametre* (fx i normalfordelingen), og dels optræder de som "operatorer" , der virker på fordelingen af en stokastisk variabel. Når de optræder som operatorer, er den stokastiske variable angivet som fodtegn. Således betyder μ_Y og σ_Y^2 henholdsvis middelværdien og variansen af den stokastiske variabel Y, ligesom $p_Y(\cdot)$ betyder tæthedsfunktionen for den stokastiske variable Y.

Bogen bruger undertiden også notationen $E(Y)$ for at angive middelværdien, μ_Y, af den stokastiske variable. Tilsvarende bruger man i den statistiske litteratur ofte notationen $V(Y)$ for at angive variansen, σ_Y^2 for den stokastiske

variable.

Regneregler for middelværdier og varianser

Med denne notation kan man direkte udtrykke de vigtige regneregler (4.9) og (4.10) som

$$E(aY+b) = aE(Y)+b \quad (9)$$
$$V(aY+b) = a^2V(Y) \quad (10)$$

og tilsvarende udtrykkes (4.12) som

$$E(Y_1+Y_2+\cdots+Y_n) = E(Y_1)+E(Y_2)+\cdots+E(Y_n) \quad (12)$$

og regnereglerne for varianser af kombinationer af uafhængige stokatiske variable (rammen side 162) udtrykkes som

$$V(Y_1 \pm Y_2) = V(Y_1)+V(Y_2)$$

og

$$V(Y_1+Y_2+\cdots+Y_n) = V(Y_1)+V(Y_2)+\cdots+V(Y_n) \quad (14)$$

Og endelig kan resultatet i rammen side 163 udtrykkes som:
Lad $Y_1, Y_2, \ldots, Y_n$ være uafhængige stokastiske variable, der følger den samme fordeling med $E(Y_i) = \mu$ og $V(Y_i) = \sigma^2$. Da gælder for den stokastiske variabel $\overline{Y}_n = \sum_i Y_i/n$, at

$$E(\overline{Y}_n) = \mu \quad (16)$$
$$V(\overline{Y}_n) = \frac{\sigma^2}{n} \quad (17)$$
$$\sqrt{V(\overline{Y}_n)} = \frac{\sigma}{\sqrt{n}} \quad (18)$$

4.6 Den centrale grænseværdisætning

Denne matematiske sætning danner en formel begrundelse for den udbredte anvendelse af normalfordelingen til beskrivelse af fordelingen af data.

I Lab. 4.3 får vi netop set en konsekvens af sætningen.

Man kan også klikke ind på

`http://www.users.on.net/zhcchz/java/quincunx/central.html`
og fortsætte igennem serien af web-sider til sidste billede, (quincunx8.html)

4.8 Transformationer af variable for at opnå en normalfordeling

Vi vil senere i kurset vise tilbage til forslagene i de fire pinde på side 208.