Matematisk statistik V
Föreläsningsanteckningar
Tommy Norberg
16 mars 2005

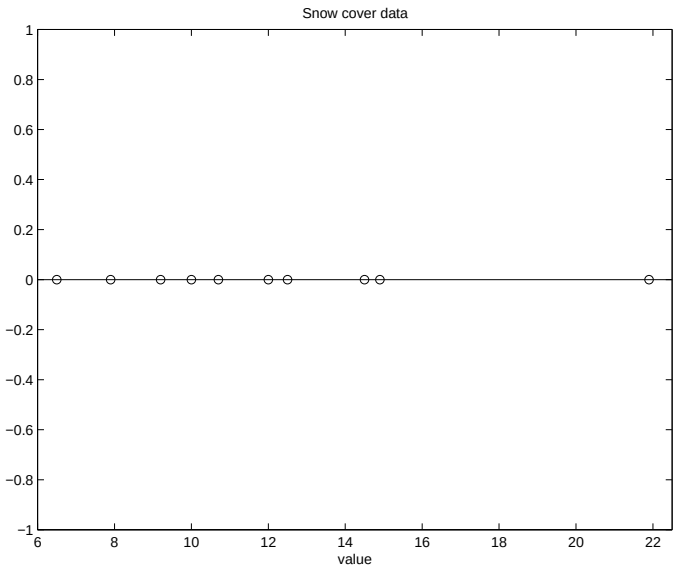
F3: Ch 2 Numerical Summary Measures

Nyckelord: Lägesmått för data och för fördelningar medelvärde, median, väntevärde, varians, standardavvikelse, kvartil, IQR, kvantil-plot.

Övning 5 (s 68) Yearly October snow cover for Eurasia (in 10^6 km^2) during the period 1970-1979:

6.5 12.0 14.9 10.0 10.7 7.9 21.9 12.5 14.5 9.2

What would you report as a representive, or typical, value of October snow cover from this period, and what prompted your choice?



Figur 1: En illustration av snödatamängden i övning 5.

2.1 Lägesmått för data (s 59-62) och fördelningar (s 62-67)

Antag att vi har n observationer eller mätningar

$$x_1, \ldots, x_n$$

av en variabel x.

Definition 1 (s 59) Observationernas *medelvärde* är

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Ordna data i storleksordning:

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

Definition 2 (s 61) Observationernas *median* är

$$\tilde{x} = \begin{cases} x_{(k+1)} & \text{om} \quad n = 2k+1 \\ \dfrac{x_{(k)} + x_{(k+1)}}{2} & \text{om} \quad n = 2k \end{cases}$$

Övning 5 (s 68) (forts) Medelvärdet av snödata är

$$\begin{aligned} \bar{x} &= \tfrac{6.5+12.0+14.9+10.0+10.7+7.9+21.9+12.5+14.5+9.2}{10} \\ &= 12.01 \end{aligned}$$

Vi ordnar snö-data i storleksordning:

$$6.5 \quad 7.9 \quad 9.2 \quad 10.0 \quad 10.7 \quad 12.0 \quad 12.5 \quad 14.5 \quad 14.9 \quad 21.9$$

och ser direkt att median är

$$\tilde{x} = \frac{10.7 + 12.0}{2} = 11.35$$

Observationernas median är den experimentella (eller *empiriska*) motsvarigheten till den teoretiska medianen. Också det observerade medelvärdet har en teoretisk motsvarighet. Antag att vi har en variabel x med antingen diskret fördelning med massfunktion $p(x)$ eller kontinuerlig fördelning med täthet $f(x)$.

Definition 3 (s 63, 65) Väntevärdet av x är

$$\mu = \sum_x xp(x)$$

om x är diskret, respektive

$$\mu = \int_R xf(x)\,dx$$

om x är kontinuerlig. Om det är nödvändigt, indicerar vi med x för att göra tydligt vilken variabel μ är medelvärdet av.

Om man har infört en stokastisk variabel X, så skriver man ofta $E[X]$ istället för μ eller μ_x.

Det är säkrast att tillägga att det finns fördelningar som inte har något väntevärde! Man måste kräva i ovanstående definition att summan respektive integralen är absolutkonvergent.

Att en variabel x är likformigt fördelad på (a, b) innebär att den är kontinuerlig och har täthetsfunktionen

$$f(x) = \frac{1}{b-a}, \quad a < x < b$$

Denna fördelning ska vi beteckna med $\mathrm{U}(a, b)$.

Övning 10 (s 68) Låt $X \sim \mathrm{U}(a, b)$. Bestäm X:s väntevärde och median.

Lösning: Väntevärdet är

$$E[X] = \int_a^b x\,\frac{1}{b-a}\,dx = \frac{a+b}{2}$$

Medianen m fås ur villkoret

$$\int_a^m \frac{1}{b-a}\,dx = \frac{1}{2}$$

Härur fås att

$$m = \frac{a+b}{2}$$

2.2 Mätning av variabilitet och motsvarande teoretiska storheter (s 69-75)

Definition 4 (s 70) Observationernas *varians* är

$$s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$$

Deras *standardavvikelse* är $s = \sqrt{s^2}$.

När man beräknar s^2, använd att

$$(n-1)s^2 = \sum_i x_i^2 - \frac{1}{n}\left(\sum_i x_i\right)^2 = \sum_i x_i^2 - n\bar{x}^2$$

Definition 5 (s 73-74) Variabeln x:s (teoretiska) *varians* är

$$\sigma^2 = \int_R (x-\mu)^2\, f(x)\,dx$$

i det kontinuerliga fallet, och

$$\sigma^2 = \sum_x (x-\mu)^2\, p(x)$$

i det diskreta fallet. *Standardavvikelsen* σ är den positiva kvadratroten ur variansen.

Om man har infört motsvarande stokastiska variabel X, så kan man skriva $\mathrm{Var}[X]$ istället för σ^2 eller σ_x^2.

Egentligen är observationernas varians medelvärdet av de kvadratiska avvikelserna från medelvärdet, alltså lika med $\sum_i (x_i - \bar{x})^2/n$. Det som boken

och många andra (lite slarvigt) kallar för observationernas varians, borde egentligen kallas skattad varians eftersom man använder s^2 till att skatta den (teoretiska) variansen σ^2.

Notera att

$$\sigma^2 = \sum_x (x-\mu)^2\, p(x) = \sum_x \left(x^2 - 2x\mu + \mu^2\right) p(x)$$

$$= \sum_x x^2 p(x) - 2\mu \sum_x x p(x) + \mu^2 \sum_x p(x)$$

$$= \sum_x x^2 p(x) - 2\mu^2 + \mu^2 = \sum_x x^2 p(x) - \mu^2$$

Om x är en variabel, så är naturligtvis även x^2 en variabel. Man kan visa att den första summan i det sista ledet ovan är lika med väntevärdet av x^2. (Detta är kanske inte så överaskande, men observera att ett bevis krävs och att det är icke-trivialt.) Således gäller

$$\sigma_x^2 = \mu_{x^2} - \mu_x^2$$

Formeln blir enklare om den uttrycks i de stokastiska variablerna X och X^2:

$$\mathrm{Var}[X] = E[X^2] - E[X]^2$$

Analoga beräkningar (man byter bara ut summationen ovan mot integration) visar att detta resultat även gäller för kontinuerliga variabler.

Övning 26 (s 68) Likformig fördelning på $[a, b]$. Bestäm variansen σ^2 och standardavvikelsen σ.

Lösning: Vi har redan sett att

$$E[X] = \frac{a+b}{2}$$

Enligt ovan behöver vi också

$$E[X^2] = \int_a^b x^2 \, \frac{1}{b-a}\, dx = \frac{1}{b-a}\left(\frac{b^3}{3} - \frac{a^3}{3}\right) = \frac{b^2+ab+a^2}{3}$$

Vi ser nu att

$$\mathrm{Var}[X] = \frac{b^2 + ab + a^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}$$

Normalfördelningens täthet är

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\, e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty < x < \infty$$

där $-\infty < \mu < \infty$ och $\sigma > 0$. Man kan visa att om $X \sim \mathrm{N}(\mu, \sigma)$, så gäller att

$$\mu = E[X]$$

och

$$\sigma^2 = \mathrm{Var}[X]$$

Om $Z \sim \mathrm{N}(0,1)$, så gäller alltså att $E[Z] = 0$ och $\mathrm{Var}[Z] = 1$.

2.3 Andra numeriska storheter av intresse (s 78-84)

I kap 1 diskuterades teoretiska percentiler. Det finns naturliga experimentella (eller empiriska) storheter. Läs själva om detta i avs 2.3 i Devore & Farnum. Här tar vi bara upp

Den 25:e percentilen kallas ofta för den *undre kvartilen*. Analogt kallas den 75:e percentilen för den *övre kvartilen*.

Kvartilerna och medianen delar in utfallsrummet för variabeln ifråga i fyra delar som alla har sannolikheten 1/4. De definieras i det kontinuerliga fallet utav att

$$\frac{1}{4} = \int_{-\infty}^{c_l} f(x)\, dx$$

$$\frac{1}{2} = \int_{-\infty}^{m} f(x)\, dx$$

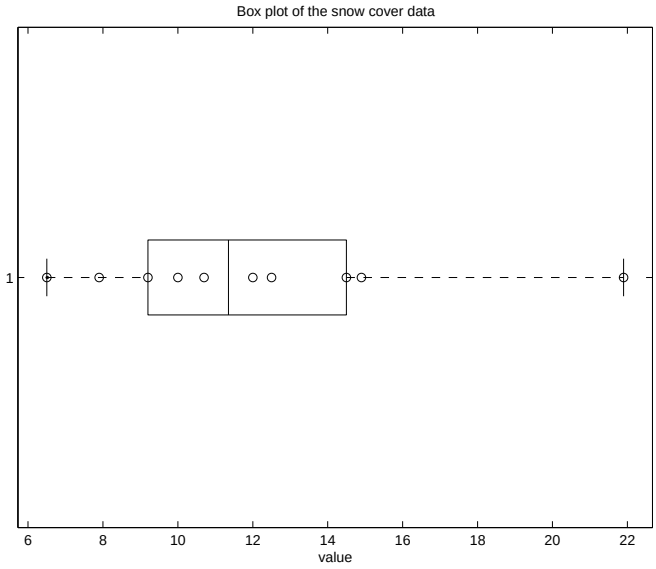
$$\frac{3}{4} = \int_{-\infty}^{c_u} f(x)\, dx$$

och det gäller att

$$P(X \leq c_l) = P(c_l < X \leq m)$$

$$= P(m < X \leq c_u) = P(c_u < X) \quad = \tfrac{1}{4}$$

Definition 6 (s 78) Dela in den ordnade datamängden i en undre och övre halva. Om antalet observationer är udda, inkludera medianen $\tilde{x}$ i båda halvorna. Den *undre kvartilen* är lika med medianen i den undre datahalvan. Den *övre kvartilen* är lika med medianen i den övre datahalvan. IQR ("Inter Quartile Range") är avståndet mellan dessa två kvartiler.



Figur 2: En s k box-plot av snödatamängden i övning 5. Obs att normalt är inte datapunkterna markerade.

2.4 Kvantilplottar (s 87-91)

Definition 7 (s 88) Betrakta den ordnade datamängden

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

Den i:te observationen i storleksordning, $x_{(i)}$, ska vi nu kalla för den empiriska (observerade, uppmätta) $(i - 0.5)/n$*-kvantilen*.

Tanken med definitionen är att vi tänker oss att mellan varje par av på varandra följande observationer lägger vi sannolikhetsmassan $1/n$. Det finns $n-1$ sådana par. Av den återstående sannolikhetsmassan lägger vi hälften $(= 1/2n)$ till vänster om den minsta och lika mycket till höger om den största.

I snödata-plotten på OH 1 kan vi alltså tänka oss att vi har sannolikhetsmassan $1/20 = 0.05$ till vänster om den minsta datapunkten $x_{(1)} = 6.5$, sannolikhetsmassan $1/10 = 0.1$ mellan $x_{(1)}$ och $x_{(2)}) = 7.9$, mellan $x_{(2)}$ och $x_{(3)} = 9.2$, etc, och slutligen $1/20 = 0.05$ till höger om den största datapunkten, som är $x_{(10)} = 21.9$.

Vi gör detta i syfte att jämföra datamängden med en teoretisk modell. Antag att man tror att tätheten $f(x)$, som vi antar är kontinuerlig, är en bra teoretisk modell för data. Låt $y_{(i)}$ vara den teoretiska $(i-0.5)/n$-kvantilen. Då gäller

$$\frac{i-0.5}{n} = \int_{-\infty}^{y_{(i)}} f(x)\,dx = F(y_{(i)})$$

Definition 8 I en *kvantilplot* jämföres den observerade kvantilen $x_{(i)}$ med motsvarande teoretiska storhet $y_{(i)}$. Man plottar alltså paren $(y_{(i)}, x_{(i)})$, för $i = 1, \ldots, n$.

Övning 5 (s 68) (forts) Snödata ordnade i storleksordning:

$$6.5 \quad 7.9 \quad 9.2 \quad 10.0 \quad 10.7 \quad 12.0 \quad 12.5 \quad 14.5 \quad 14.9 \quad 21.9$$

Låt oss kolla om det verkar rimligt att den teoretiska modellen är $\mathrm{LN}(\mu, \sigma)$, för något parameterpar (μ, σ).

Vi logaritmerar data, d.v.s. låter $x_{(1)}, \ldots, x_{(10)}$ vara lika med $\ln(6.5), \ldots, \ln(21.9)$. Vi får då följande loggade data:

$$\begin{array}{rrrrr} 1.872 & 2.067 & 2.219 & 2.303 & 2.370 \\ 2.485 & 2.526 & 2.674 & 2.701 & 3.086 \end{array}$$

Detta är de experimentella $0.05, 0.15, \ldots, 0.95$-kvantilerna.

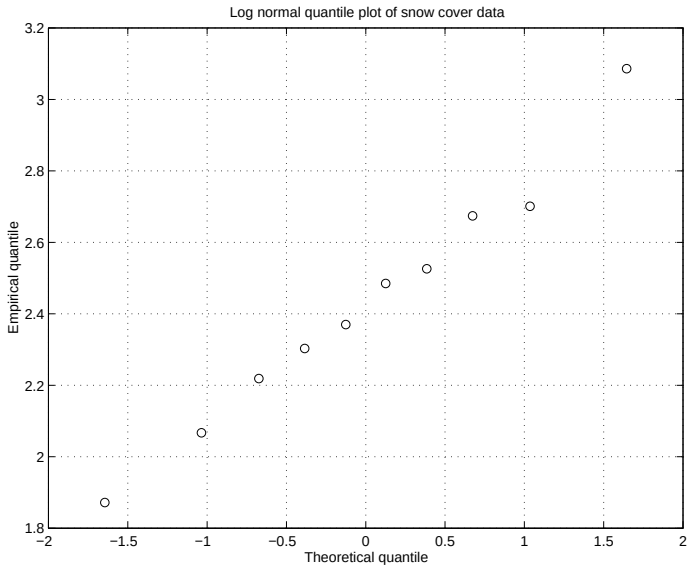
Motsvarande kvantiler i den standardiserade normalfördelningen är

$$\begin{array}{rrrrr} -1.645 & -1.036 & -0.674 & -0.385 & -0.126 \\ 0.126 & 0.385 & 0.674 & 1.036 & 1.645 \end{array}$$

Obs att

$$\begin{aligned} -1.645 &= \Phi(0.05) \\ -1.036 &= \Phi(0.15) \\ &\vdots \\ 1.645 &= \Phi(0.95) \end{aligned}$$

Vi plottar nu uppmätta kvantiler mot standardiserade, och får



Figur 3: En kvantilplot av snödatamängden i övning 5. Teoretisk modell är den lognormala.

Obs att om modellen är ok, så ska punkterna ungefär ligga utefter en rät linje.

Hade vi trott att den teoretiska modellen var $\mathrm{N}(\mu, \sigma)$ för något par μ, σ, så hade vi gjort precis som ovan fast med ursprungsdata istället för de logaritmerade.

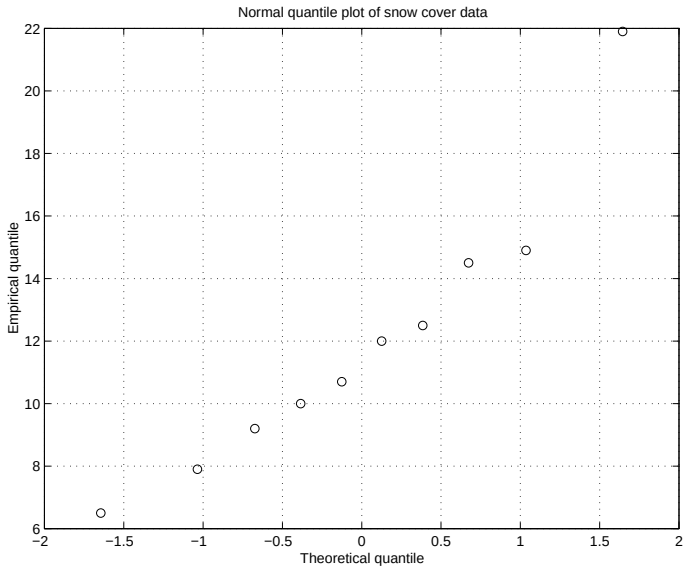
Då hade vi jämfört de empiriska $0.05, 0.15, \ldots, 0.95$-kvantilerna

$$6.5 \quad 7.9 \quad 9.2 \quad 10.0 \quad 10.7 \quad 12.0 \quad 12.5 \quad 14.5 \quad 14.9 \quad 21.9$$

med de teoretiska

$$\begin{array}{rrrrr} -1.645 & -1.036 & -0.674 & -0.385 & -0.126 \\ 0.126 & 0.385 & 0.674 & 1.036 & 1.645 \end{array}$$

och fått plotten



Figur 4: En kvantilplot av snödatamängden i övning 5. Teoretisk modell är normalfördelningen.

Helst vill man att punkterna i en kvantilplott ska ligga ungefär utefter en rät linje. För snödatamängden ser kvantilplotten relativt lognormalfördelningen bra ut. Observera att när vi plottar relativt normalfördelningen så hamnar den största observationen långt ifrån den räta linje som man kan ana sig till att de övriga ligger utefter. Detta är en inte alltför ovanlig situation och slutsatsen är att data förmodligen ej är normalfördelade, men att de kan vara lognormalfördelade. Men observera att vi endast har 10 observationer. För att göra tillförlitliga uttalanden om vilka fördelningar som passar resp. inte passar ihop med en viss datamängd ska man nog helst ha fler observationer än 10.

Värt att nämna är att man kan också jämföra data och teori i en s.k *probability-plot*. I den plottas de teoretiska sannolikheterna

$$\frac{i-0.5}{n} = F(y_{(i)})$$

mot de empiriska

$$\hat{p}_i = F(x_{(i)})$$

Även punkterna i denna plot skall ligga ungefär utefter en rät linje om modellen ska kunna anses vara ok.

Obs att man brukar reservera beteckningen z_p för $(1-p)$-kvantilen m a p den standardiserade normalfördelningen. Således definieras t.ex. $z_{0.05}$ utav att

$$\int_{-\infty}^{z_{0.05}} \varphi(z)\, dz = 0.95$$

Vi har redan sett att $z_{0.05} = 1.645$.

En annan bra kvantil att komma ihåg är $z_{0.025} = 1.96$.

Man kallar (tyvärr) ofta $z_{0.05}$, resp $z_{0.025}$ för 5%, resp 2.5%-kvantilen.

Observera inkonsekvensen. Vi har ju lärt oss tidigare att p-kvantilen definieras av

$$\int_{-\infty}^{z_p} \varphi(z)\, dz = p$$

Men i statistiska sammanhang är det brukligt att istället använda definitionen

$$\int_{z_p}^{\infty} \varphi(z)\, dz = p$$

Om tveksamhet råder, var noga med att påpeka vilken definition du avser.

Titta själva på **exempel 2.19** där den teoretiska modellen är Weibullfördelningen.