Kapitel 6: "Descriptive Statistics"

(Detta kapitel handlar mest om grafiskt åskådliggöra ett datamaterial.)

(Kapitlet är klart enklast i boken.)

Det finns fyra olika grafiska metoder lära:

1) "Steam-and-leaf diagram" = stam-och-blad plot
2) "Ogive/Commulative distribution plot" = fördelnings plot
3) "Boxplot" = lådagram
4) "Histogram"

Dessa kan ni läsa om i boken själva.

Det står även om tre saker till i kapitel 6:

*) "Sample mean" = stickprovs medelvärde

*) "Sample variance" = stickprovs varians

(som vi pratar om i kapitel 7 - vänta tills dess, och)

*) <u>median</u>:

$x_1, \dots, x_n$ är ett datamaterial i (vanliga) siffror

$x_{(1)}, \dots, x_{(n)}$ är samma data ordnade i växande ordning

$$\text{medianen} = \tilde{x} = \begin{cases} x_{(k)} & \text{om } n = 2k-1 \text{ är udda} \\ (x_{(k)} + x_{(k+1)})/2 & \text{om } n = 2k \text{ är jämn} \end{cases}$$

är det mittersta värdet bland data

Kapitel 7: "Estimation" = skattning

$X_1, \ldots, X_n$ oberoende (stokastiska variabler) med samma frekvensfunktion $f(x) = f_X(x)$ som en (stokastisk variabel) X

$X_1, \ldots, X_n$ (kallas) stickprov på "sample" observationer av X

(frekvensfunktionen) f beror på en parameter θ (som kan vara en- eller flerdimensionell) vars värde vi ej vet

vi önskar skatta θ's värde mha $X_1, \ldots, X_n$

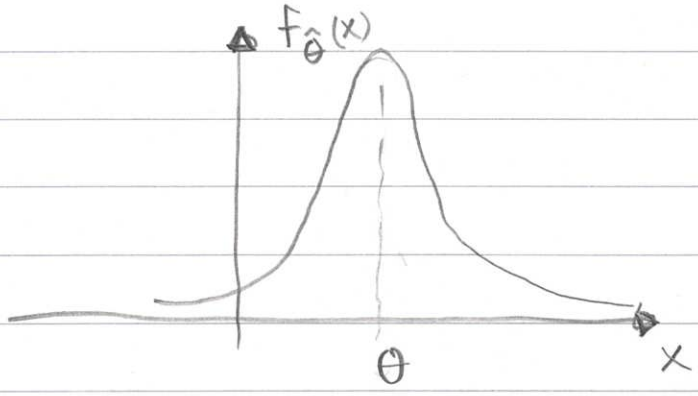
vår skattning $\hat{\theta}$ av θ är en funktion $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ av stickprovet (vår tillgängliga information)

eftersom $X_1, \ldots, X_n$ är stokastiska är $\hat{\theta}(X_1, \ldots, X_n)$ stokastisk

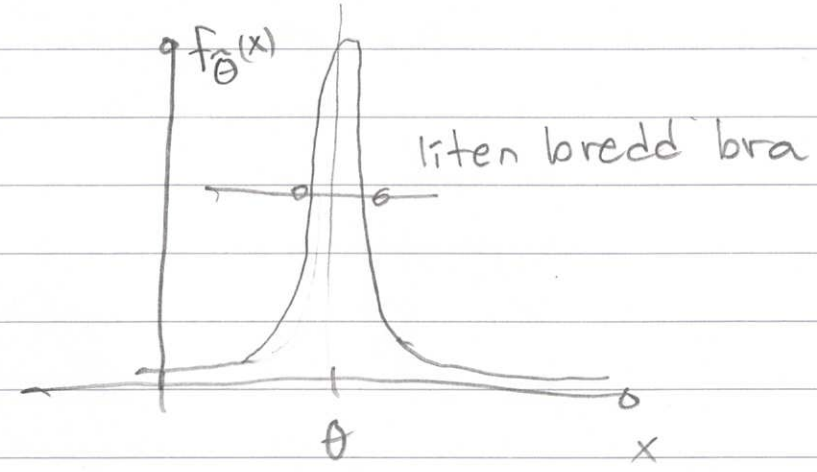
Definition skattningen $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ av parametern θ (med okänt värde) är väntevärdesriktig (vvr) om

$$E(\hat{\theta}) = E(\hat{\theta}(X_1, \ldots, X_n)) = \theta$$

Eftersom $\hat{\theta}$ är en stokastisk variabel har den en frekvensfunktion



Om $\hat{\theta}$ är vvr är en liten varians $\mathrm{Var}(\hat{\theta})$ för $\hat{\theta}$ önskvärd ty det betyder $\hat{\theta} \approx \theta$ med stor sannolikhet



Så har man två olika vvr skattningar av $\hat{\theta}$ är den med minst varians bäst.

Definition stickprovs medelvärdet är

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Sats $\bar{X}$ är en vvr skattning av $\mu = E(X_1) = \ldots = E(X_n)$.

Bevis: $E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$. □

(sats 5.9)

Sats $\mathrm{Var}(\bar{X}) = \sigma^2/n$ då $\sigma^2 = \mathrm{Var}(X_1) = \ldots = \mathrm{Var}(X_n)$

Bevis: $\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$. □

(3 i sats 5.13)

Definition stickprovsvariansen är

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})^2$$

stickprovsstandardavvikelsen är $S = \sqrt{S^2}$

Sats S^2 är en vvr skattning av σ^2

(Beviset är småsvår överkurs.)

(Det finns två allmänt användbara metoder för att hitta skattningar av parametrar:)

1) Momentmetoden

De (nu endimensionella) parametrarna $\theta_1, \dots, \theta_\ell$ önskas skattade.

Frekvensfunktionen $f_X(x)$ för X (och $X_1, \dots, X_n$) är känd förutom den beror på okända $\theta_1, \dots, \theta_\ell$

a) använd $f_X(x)$ till räkna ut momenten

$$E(X^k) = m_k = m_k(\theta_1, \dots, \theta_\ell) = \int_{-\infty}^{+\infty} x^k f_X(x)\, dx$$

för $k = 1, \dots, \ell$

b) räkna ut motsvarande stickprovsmoment

$$M_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k \quad \text{för } k = 1, \dots, \ell$$

c) finn lösningen $(\theta_1, \dots, \theta_\ell) = (\hat{\theta}_1, \dots, \hat{\theta}_\ell)$ till ekvationssystemet

$$\begin{cases} M_1 = m_1(\theta_1, \dots, \theta_\ell) \\ \vdots \\ M_\ell = m_\ell(\theta_1, \dots, \theta_\ell) \end{cases}$$

(Lösningen till ekvationssystemet är sökta skattningen.)

Exempel (Normalfördelning)

$X_1, \ldots, X_n$ är observationer av X som är $N(\mu, \sigma^2)$

$$\begin{cases} m_1(\mu, \sigma^2) = E(X) = \mu \\ m_2(\mu, \sigma^2) = E(X^2) = \mathrm{Var}(X) + (E(X))^2 = \sigma^2 + \mu^2 \end{cases}$$

Lös

$$\begin{cases} \mu = m_1(\mu, \sigma^2) = \bar{X} \\ \sigma^2 + \mu^2 = m_2(\mu, \sigma^2) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \end{cases}$$

Lösningen $(\mu, \sigma^2) = (\hat{\mu}, \hat{\sigma}^2)$ till detta är skattningen

$$\hat{\mu} = \bar{X}, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2$$

2) Maximum likelihood (ML) metoden

(Det är tryckfel mitt på sidan 232 i boken - där skall det stå ML metoden istf momentmetoden fetstilt.)

$f_X(x) = f_X(x; \theta_1, \ldots, \theta_e)$ beror av parametrarna (med okänt värde som önskas skattade)

a) bilda likelihoodfunktionen

$$L(X_1, \ldots, X_n; \theta_1, \ldots, \theta_e) = \prod_{i=1}^{n} f_X(X_i; \theta_1, \ldots, \theta_e)$$

(produkttecken - som summa men gånger istället)

b) finn de värde $(\theta_1, \ldots, \theta_e) = (\hat{\theta}_1, \ldots, \hat{\theta}_e)$ som maximerar $L(X_1, \ldots, X_n; \theta_1, \ldots, \theta_e)$

(Lösningen till detta är skattningen. Därav namnet maximum likelihood metoden.)

(Momentmetoden och ML-metoden för finna skattningar av parametrar behöver ej ge exakt samma skattningar men för det mesta är de åtminstone nästa lika.)

Man väljer ofta maximera
<u>loglikelihooden</u> $\ln(L(X_1,\dots,X_n;\theta_1,\dots,\theta_\ell))$
istf likelihooden $L(X_1,\dots,X_n;\theta_1,\dots,\theta_\ell)$
själv map $\theta_1,\dots,\theta_\ell$.

(Kom ihåg att det är inte maximala
värdet av L eller $\ln L$ vi söker
utan de parametrar $(\theta_1,\dots,\theta_\ell)=(\hat\theta_1,\dots,\hat\theta_\ell)$
för vilket maximala värdet antages
och de är samma för båda.)

Orsaken till maximera $\ln L$ istf L är att

$$\ln(L)=\ln\left(\prod_{i=1}^{n} f_X(X_i;\theta_1,\dots,\theta_\ell)\right)=\sum_{i=1}^{n}\ln\left(f_X(X_i;\theta_1,\dots,\theta_\ell)\right)$$

omvandlar produkten till ofta enklare summa

Exempel (Normalfördelning fortsättning)

$X_1, \dots, X_n$ är observationer av X som är $N(\mu, \sigma^2)$

$$L = L(X_1, \dots, X_n; \mu, \sigma^2) = \prod_{i=1}^{n} f_X(X_i; \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}}$$

$$\ln(L) = -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \sum_{i=1}^{n} \frac{(X_i-\mu)^2}{2\sigma^2}$$

$$\begin{cases} \frac{\partial \ln(L)}{\partial \mu} = \sum_{i=1}^{n} \frac{(X_i-\mu)}{\sigma^2} = 0 \\ \frac{\partial \ln L}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(X_i-\mu)^2}{\sigma^3} = 0 \end{cases} \quad \text{har lösning}$$

$$(\mu, \sigma^2) = (\hat{\mu}, \hat{\sigma}^2) = \left(\frac{1}{n}\sum_{i=1}^{n} X_i, \frac{1}{n}\sum_{i=1}^{n} (X_i-\mu)^2 \right) = \left(\bar{X}, \frac{1}{n}\sum_{i=1}^{n} (X_i-\bar{X})^2 \right)$$

Konfidensintervall

(Bara) en skattning $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ av en (okänd) parameter θ kan vara otillräcklig (då man ej vet något om skattningens precishet).

Definition $[L_1(X_1, \ldots, X_n), L_2(X_1, \ldots, X_n)]$ kallas ett konfidensintervall för θ med konfidensgrad $1-\alpha \in (0,1)$ om

$$P(L_1(X_1, \ldots, X_n) \leq \theta \leq L_2(X_1, \ldots, X_n)) = 1-\alpha$$

Talet α väljes ofta som $\alpha = 0.05, 0.01$ eller 0.001.

För att lära er göra konfidensintervall för normalfördelade data - exemplet vi studerat två gånger - gör jag nu några förberedelser:

Sats Om $X_1, \ldots, X_n$ är oberoende $N(\mu, \sigma^2)$-fördelade är

$$\sum_{i=1}^{n} X_i \quad N(n\mu, n\sigma^2)\text{-fördelad.}$$

Bevis Jag använder MGF (momentgenererande funktion): kom ihåg att enligt exempel 4.3

$$m_{X_i}(t) = E(e^{tX_i}) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$$

så att eftersom $e^{tX_1}, \ldots, e^{tX_n}$ (också) är oberoende (och därför okorrelerade)

$$m_{\sum_{i=1}^{n} X_i}(t) = E(e^{t(X_1 + \ldots + X_n)}) = E(e^{tX_1} \cdot \ldots \cdot e^{tX_n})$$

$$= E(e^{tX_1}) \cdot \ldots \cdot E(e^{tX_n}) = \left(e^{\mu t + \frac{1}{2}\sigma^2 t^2}\right)^n = e^{n\mu t + \frac{1}{2}\sigma^2 n t^2}$$

som är MGF för $N(n\mu, n\sigma^2)$. □

Följdsats Om $X_1, \ldots, X_n$ är oberoende $N(\mu, \sigma^2)$-fördelade är

$$\bar{X} \quad N(\mu, \sigma^2/n)\text{-fördelad.}$$

Bevis $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = (X_1/n) + \ldots + (X_n/n)$

där

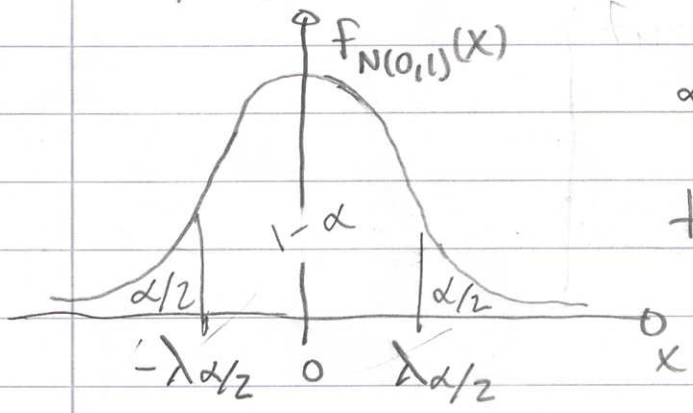
$$m_{(X_i/n)}(t) = E(e^{t(X_i/n)}) = E(e^{(t/n)X_i}) = e^{\mu t/n + \frac{1}{2}\sigma^2 t^2/n^2}$$

är MGF för en $N(\mu/n, \sigma^2/n^2)$-fördelning så att genom byta ut μ och σ^2 mot μ/n och σ^2/n^2 i föregående sats

$\bar{X} = (X_1/n) + \ldots + (X_n/n)$ är $N(\mu, \sigma^2/n)$-fördelad. □

Exempel (Normalfördelning fortsättning)

$X_1, \ldots, X_n$ är observationer av X som är $N(\mu, \sigma^2)$



$\alpha/2, 1-\alpha, \alpha/2$ är areor

talet $\lambda_{\alpha/2}$ fines mha tabell I

Om $\alpha = 0.01$ är $P(N(0,1) \leq \lambda_{\alpha/2}) = 1 - \alpha/2 = 0.995$
vilket enligt tabell I ger $\lambda_{\alpha/2} = 2.575$

$$P\left(\bar{X} - \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

är konfidensintervall för μ med konfidensgrad $1-\alpha$ eftersom händelsen (vi räknar P för) ovan är samma som

$$\left\{-\lambda_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq \lambda_{\alpha/2}\right\}$$

där $(\bar{X}-\mu)/(\sigma/\sqrt{n})$ är $N(0,1)$ enligt föregående sats och sats 4.12.

Konfidensintervallet för μ ovan skrivs ofta kortare

$\mu \in \overline{X} \pm \lambda_{\alpha/2}\, \sigma/\sqrt{n}$ med sannolikhet $1-\alpha$

och fungerar då σ är känt.

Fråga: Vad gör man då σ okänt (= realistiskt)?

Svar: Ersätt σ med S (dvs roten av stickprovsvariansen)!

(Det är en allmänn princip i statistik att om man behöver ett parametervärde som man ej vet så ersätter man det med ett skattat parametervärde.)

Då σ ersättes med S får konfidens-intervallet approximativ konfidensgrad $1-\alpha$ istf exakt (eftersom vi ändrat från vår exakta beräkning).

Kom ihåg sats 4.13 enligt vilken

$$\sum_{i=1}^{n}(X_i-\mu)/\sqrt{n\sigma^2} = (\bar{X}-\mu)/\sqrt{\sigma^2/n} \approx N(0,1)$$

för vilka som helst oberoende likafördelade $X_1,\dots,X_n$ med väntevärde μ och varians σ^2.

Eftersom detta var precis den information vi utnyttjade i exemplet ovan men med (approximativ likhet) $\approx$ utbytt mot (likhet) = följer att

$$\mu \in \bar{X} \pm \lambda_{\alpha/2}\sigma/\sqrt{n} \approx \bar{X} \pm \lambda_{\alpha/2}S/\sqrt{n}$$

är konfidensintervall för μ med approximativ konfidensgrad $1-\alpha$ för alla $X_1,\dots,X_n$ som ovan!

(Detta är ett av de mest grundläggande och mest använda faktumen i statistik.)

Exempel (Binomialapproximation)

Om $X_1, \ldots, X_n$ är oberoende likafördelade Binomial $(1,p)$ (dvs var och en av dem är 1 eller 0 med sannolikhet p respektive $1-p$) är

$$Y = X_1 + \ldots + X_n \text{ Binomial}(n,p)\text{-fördelad}$$

(se text efter Definition 3.14 – ty Y är antalet lyckade försök av n utförda försök då varje försök lyckas med sannolikhet p).

Eftersom $E(X_i) = p$ och $\text{Var}(X_i) = p(1-p)$ (se exempel 3.2) så att enligt föregående teori är

$$(Y - np)/\sqrt{np(1-p)} = \sum_{i=1}^{n}(X_i - p)/\sqrt{np(1-p)} \approx N(0,1)$$

$$p \in \bar{X} \pm \lambda_{\alpha/2}\,\sigma/\sqrt{n} = \bar{X} \pm \lambda_{\alpha/2}\sqrt{p(1-p)/n}$$

med konfidensgrad $\approx 1-\alpha$

där $\sigma = \sqrt{p(1-p)}$

där okända p på högersidan kan ersättas av skattningen $\hat{p} = \hat{\mu} = \bar{X}$.