Grundläggande matematisk statistik

Punktskattning

Uwe Menzel, 2018

uwe.menzel@matstat.org

www.matstat.org

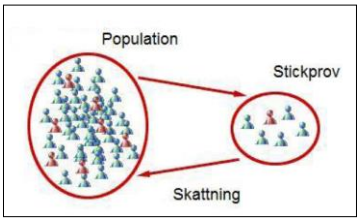
Sannolikhetsteori och statistikteori

Sannolikhetsteori:

- vad vi har gjort t.o.m. nu
- vi hade en given fördelning, t.ex. $X \sim N(\mu, \sigma)$, där μ och σ var kända.
- utifrån detta beräknades bl.a. väntevärde, varians och slh.:er, t.ex. $P(a < X \leq b)$

Statistikteori:

- ”det verkliga livet”
- vi har inte μ och σ (men möjligtvis en uppfattning över den föreliggande typen av fördelning, alltså t. ex. $X \sim N$).
- vi måste skatta fördelningsparametrarna μ och σ pga. av ett stickprov → punktskattning, intervallskattning (inferens)



I ett slumpmässigt stickprov måste alla mätvärden vara **oberoende**.

Intuitiva skattningar

Normalfördelning:

Skattning av μ och σ hos en normalfördelning (vi måste alltså veta/visa att normalfördelning föreligger).

$X \sim N(\mu, \sigma)$ men μ eller/och σ okända

Stickprov: $x = (x_1, x_2, \ldots x_n)$

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

skattning för väntevärdet $\mu = E(X)$ hos en normalfördelad slumpvariabel $X \sim N(\mu, \sigma)$

OBS: en skattning betecknas ofta med en stjärna, t. ex. μ^*

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

skattning för variansen $\sigma^2 = V(X)$ hos en normalfördelad slumpvariabel $X \sim N(\mu, \sigma)$

Intuitiva skattningar

Binomialfördelning:

Skattning av parametrern p för binomialfördelningen. Vi måste alltså veta att en binomialfördelning föreligger, t.ex. som hos myntkast).

$X \sim Bin(n, p)$ … men p okänd

Stickprov: x = antalet ”lyckade”; n = antalet (Bernoulli-) försök

$$p^* = \frac{x}{n}$$

skattning för slh. att ”lyckas”, p, för Bin(n, p)

Exempel 1: 10 myntkast ($n = 10$), därav 4 ”lyckade” ($x = 4$)

$p^* = \frac{x}{n} = \frac{4}{10} = 0.4$ **OBS**!: kanske var n för liten?

Exempel 2: Opinionsundersökning: 1000 personer frågas om de skulle välja partiet ”De bästa”. 3011 svarar med ”Ja”. Vi kan skatta andelen ”De bästa”-väljare i hela populationen med ungefär 30%.

Skattningen som slumpvariabel

stickprov 1: $x^1 = (x_1^1, x_2^1, \ldots x_n^1)$

stickprov 2: $x^2 = (x_1^2, x_2^2, \ldots x_n^2)$

stickprov 3: $x^3 = (x_1^3, x_2^3, \ldots x_n^3)$

alla stickprov ger olika $\bar{X}$ och S^2 → skattningar måste själva betraktas som slumpvariabler

- olika numeriska värden för olika stickprov → skattningen beror av slumpen
- observationerna $x_1, x_2, \ldots, x_3$ ses som utfall av oberoende slumpvariabler $X_1, X_2, \ldots, X_3$ som antas ha samma fördelning.
- likaså anses det konkreta värdet $\bar{x}$ för en enskild skattning som en realisation av en slumpvariabel $\bar{X}$

$$\underbrace{\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i}_{\text{stickprovsvärde}} \Longrightarrow \underbrace{\bar{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i}_{\text{slumpvariabel}}$$

Skattningen som slumpvariabel

stickprov 1: $x^1 = (x_1^1, x_2^1, \ldots x_n^1)$

stickprov 2: $x^2 = (x_1^2, x_2^2, \ldots x_n^2)$

stickprov 3: $x^3 = (x_1^3, x_2^3, \ldots x_n^3)$

alla stickprov ger olika $\bar{X}$ och S^2 → skattningar måste själva betraktas som slumpvariabler

$$\underbrace{\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i}_{\text{stickprovsvärde}} \Longrightarrow \underbrace{\bar{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i}_{\text{slumpvariabel}}$$

Exempel *Bin*: $p^* = \frac{x}{n} \Longrightarrow p^* = \frac{X}{n}$ $\quad X \sim Bin(n, p)$

$E(p^*)$, $V(p^*)$, … kan beräknas

Skattningar

observationer $x_1, x_2, \ldots, x_3$ → slumpvariabler $X_1, X_2, \ldots, X_3$

stickprovsvärde	slumpvariabel	fördelning
$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	$X_i \sim N(\mu, \sigma)$
$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$	$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})^2$	$X_i \sim N(\mu, \sigma)$
$p^* = \frac{x}{n}$	$p^* = \frac{X}{n}$	$X \sim Bin(n, p)$

Egenskaper hos skattningar

En skattning borde uppfylla flera **krav:**

- väntevärdesriktigthet
- konsistens
- effektivitet

Med hjälp av dessa egenskaper kan man bedöma hur lämplig en skattning är.

Väntevärdesriktighet

θ^* = skattning för θ θ står för en parameter i en fördelning ($\mu, \sigma, p, \ldots$)

När väntvärdesriktighet föreligger, så måste gälla $E(\theta^*) = \theta$

dvs. väntevärdet för skattningen måste stämma överens med parametern (i den förknippade slumpvariabeln) som ska skattas.

Ex1: skattning av väntevärde och varians för en normalfördelning:

$$E(\mu^*) = E(\bar{X}) = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n \cdot \mu = \mu$$ ✓

Även om vi inte känner det sanna värdet på μ, så vet vi dock att skattningens väntevärde stämmer överens med μ !

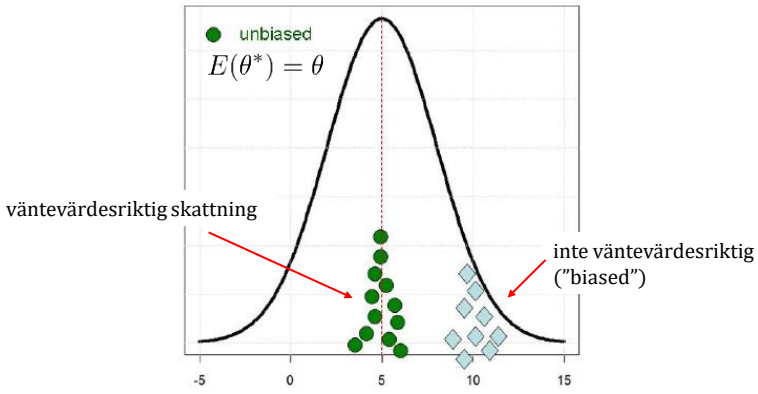
$E(S^2) = \cdots = \sigma^2$ ✓ (det är därför vi måste ha $n-1$ in nämnaren på S^2!)

Ex2: skattning för sannolikheten ”att lyckas” i en binomialfördelning:

$$E(p^*) = E\left(\frac{X}{n}\right) = \frac{1}{n} \cdot E(X) = \frac{1}{n} \cdot n \cdot p = p$$ ✓

$E(X)$ för $Bin(n, p)$

Väntevärdesriktighet



- En väntevärdesriktig skattning θ^*är väl koncentrerad kring det sanna värdet på parametern θ.
- En väntevärdesriktig skattning ger rätt parametervärde i genomsnittet om man gör ett stort antal försök.

Konsistens

θ_n^* = skattning baserad på n observationer (mätvärden)

$$\lim_{n\to\infty} V(\theta_n^*) = 0$$

Ju större n, desto mindre ska variansen hos skattningen vara. Om n går mot ∞ ska variansen gå mot noll.

Om en skattning är konsistent lönar det sig att samla fler observationer – vi får ju mindre varians hos skattningen, dvs. mindre osäkerhet.

Ex 1: skattning för väntevärdet och variansen för en normalfördelning:

$$V(\mu^*) = V(\bar{X}) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n^2}\cdot n\cdot\sigma^2 = \frac{\sigma^2}{n}$$

Låter det konstigt? μ^* är en slumpvariabel, har alltså en varians! $V(\mu^*)$ är variansen för skattningen av medelvärdet μ^*.

$$\lim_{n\to\infty} V(\mu^*) = \lim_{n\to\infty} \frac{\sigma^2}{n} = 0 \quad ✓ \qquad \lim_{n\to\infty} V(S^2) = \cdots = 0 \quad ✓$$

Konsistens

Ex 2: skattning för sannolikheten ”att lyckas” i en binomialfördelning:

$$p^* = \frac{X}{n}$$ skattning för p i Bin(n,p) $\quad X \sim Bin(n,p)$

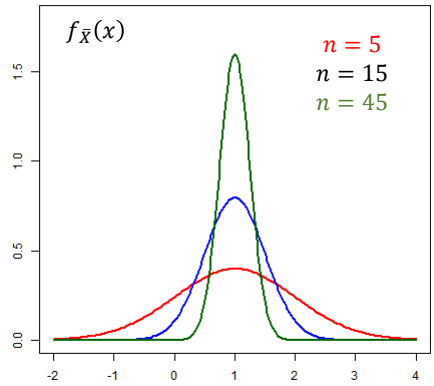
$$V(p^*) = V\left(\frac{X}{n}\right) = \frac{1}{n^2}V(X) = \frac{1}{n^2}\cdot n\cdot p\cdot(1-p) = \frac{p\cdot(1-p)}{n}$$

$$\lim_{n\to\infty} V(p^*) = \lim_{n\to\infty} \frac{p\cdot(1-p)}{n} = 0 \quad ✓$$

Sammanfattning: De intuitiva skattningarna för μ, σ i $N(\mu,\sigma)$ respektive p i $Bin(n,p)$ är både väntevärdesriktiga och konsistenta.

Konsistens

Täthetsfunktion för skattning $\bar{X}$



Skattningen koncentreras mer och mer kring det sanna värdet när stickprovet blir större och större, dvs. när n växer.

www.matstat.org

Effektivitet

En skattning borde ha en så liten varians som möjligt. Stor varians betyder ju stor osäkerhet. Kan man välja mellan olika skattningar borde man använda *den* skattning som har minst varians, alltså den mest effektiva.

Exempel: stickprov $x = (x_1, x_2, \ldots x_{10})$ där $X \sim N(\mu, \sigma)$

$\mu_1^* = \bar{X}_{10}$ **Skattning nr. 1**: det aritmetiska medelvärdet av alla 10 observationer

$$V(\mu_1^*) = V\left(\bar{X}_{10}\right) = \frac{\sigma^2}{10}$$

$\mu_2^* = \dfrac{X_1 + X_{10}}{2}$ **Skattning nr. 2:** medelvärdet av det största och det minsta värdet

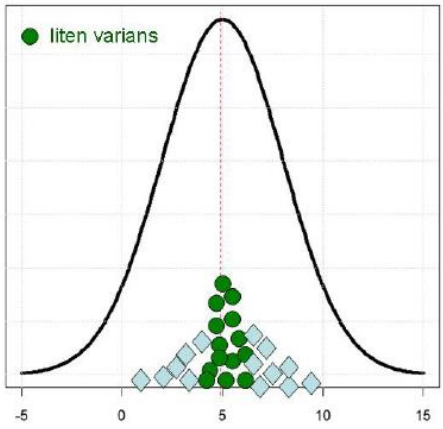
$$V\left(\mu_2^*\right) = V\left(\frac{X_1 + X_{10}}{2}\right) = \frac{1}{4}\left[V(X_1) + V(X_{10})\right] = \frac{1}{4} \cdot 2\sigma^2 = \frac{\sigma^2}{2}$$

$V(\mu_1^*) < V(\mu_2^*)$ $\Longrightarrow$ μ_1^* är mest effektiv

Bäst: **MVUE** = Minimum Variance Unbiased Estimator

Effektivitet

θ^*är effektivare än $\hat{\theta}$ om $V(\theta^*) < V(\hat{\theta})$



mest effektiv

mindre effektiv

www.matstat.org

Metoder för att hitta skattningar

(… för att hitta en formel som används på mätvärdena)

a) Momentmetoden
b) Minsta-kvadrat-metoden
c) Maximum-Likelihood-metoden
d) …..

www.matstat.org

Momentmetoden

Det k:e momentet av en slumpvariabel X är enligt definition väntevärdet för den k:e potensen av denna slumpvariabel:

$$m_k = E(X^k)$$

För normalfördelningen gäller t.ex.: → föreläsning F3: $V(X) = E(X^2) - E(X)^2$

$$m_1 = E(X) = \mu \qquad m_2 = E(X^2) = V(X) + E(X)^2 = \sigma^2 + \mu^2$$

Momenten är alltså parametrar som beskriver fördelningen; har man alla moment, så har man en entydig beskrivning av fördelningen. Analogt kan ett stickprov beskrivas genom empiriska moment:

$$\hat{m}_k = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^k$$

Idéen är nu att skatta de teoretiska momenten för en fördelning med hjälp av stickprovets empiriska moment → momentmetoden.

Momentmetoden

fördelningens 1:a moment = stickprovets 1:a moment

$$E(X;\ \theta) = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

fördelningens 1:a moment
(som beror på parametern som ska skattas)

stickprovets 1:a moment

Exempel: $Bin(n, p)$ $\qquad n \cdot p = x_1$

p är parametern som ska skattas (här: $\theta = p$)

bara ett mätvärde
(antalet ”lyckade”)

$$\Rightarrow \quad p^* = \frac{X_1}{n}$$ skattning för p i $Bin(n, p)$

Minsta-kvadrat-metoden

Man bestämmer fördelningssparametern θ så att summan av alla mätvärdens kvadratiska avstånd till väntevärdet $Q(\theta)$ minimeras (jämför föreläsning linjär regression, **F10**)

$$Q(\theta) = \sum_{i=1}^{N} \left(x_i - E(X;\ \theta)\right)^2 \ \rightarrow \ Min$$

x_i: stickprov

$E(X)$ beror på fördelnings-parametern θ

Exempel: $Bin(n, p)$

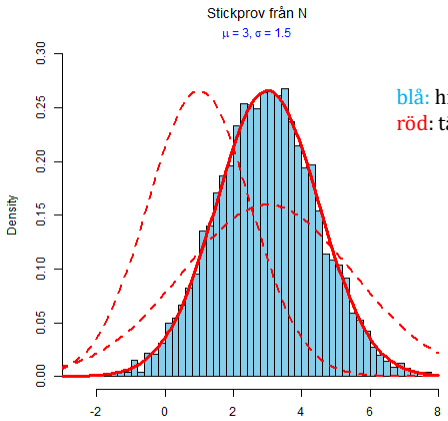
$X \sim Bin(n, p)$ där p ska skattas $\quad E(X) = n \cdot p$

bara ett värde i stickprovet: x_1

$Q(p) = (x_1 - n \cdot p\ )^2 \ \rightarrow \ Min$

$$\frac{dQ}{dp} = -2 \cdot (x_1 - n \cdot p) \cdot n = 0 \ \rightarrow \ p = \frac{x_1}{n}$$

Maximum-Likelihood-Metoden



blå: histogram över mätvärdena
röd: täthetsfunktion för olika μ, σ

ML_plot.R

Idé: välj parametrarna μ, σ så att mätvärdena förklaras bäst → välj de parametrarna som ger den tjocka röda kurvan.

Maximum-Likelihood-Metoden

Likelihood: funktion av en fördelningsparameter

stickprov: $x = (x_1, x_2, \ldots, x_n)$

$$L(\theta) = \begin{cases} p_X(x_1) \cdot p_X(x_2) \cdot \ldots \cdot p_X(x_n) & \text{diskret, } p_x \text{ känd sannolikhetsfunktion} \\ f_X(x_1) \cdot f_X(x_2) \cdot \ldots \cdot f_X(x_n) & \text{kontinuerlig, } f_x \text{ känd täthetsfunktion} \end{cases}$$

Likelihoodfunktionen kan skrivas som produkt eftersom mätvärdena är oberoende!

Exempel: normalfördelning x_i kända mätvärden

$$L(\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i-\mu)^2/2\sigma^2} = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum(x_i-\mu)^2/2\sigma^2}$$

Bestäm μ och σ så att $L(\mu, \sigma)$ blir maximalt :

$$\frac{\partial L}{\partial \mu} = 0 \qquad \frac{\partial L}{\partial \sigma} = 0$$

ger μ och σ som funktion av mätvärdena x_i, alltså en formel för skattningen

www.matstat.org

Parametriska punktskattningar

Fördelning	Parameter	Skattning för parametern	Metod	Blom
$X \sim N(\mu, \sigma)$	μ	$\mu^* = \bar{X}$	ML, MK, MM	289
$X \sim N(\mu, \sigma)$	σ^2	$S^2 = \frac{1}{n-1} S_{xx}$	MM, ML	277, 290
$X \sim N(\mu, \sigma)$	σ	$S = \sqrt{S^2}$	ML	291 Anm. 7.14
$X \sim Po(\mu)$	μ	$\mu^* = \bar{X}$	MM, MK, ML	295
$X \sim Bin(n, p)$	p	$p^* = \frac{X}{n}$	ML	293
$X \sim Hyp(N, n, m)$	p	$p^* = \frac{X}{n}$	MM, MK, ML	294
$X \sim Exp(\lambda)$	λ	$\lambda^* = \frac{1}{\bar{X}}$	ML	282
$X \sim U(0, \theta)$	θ	$\theta^* = \frac{n+1}{n} \cdot max(X_i)$	ML	284

$\bar{X} = \sum_{i=1}^{n} X_i$ $\quad S_{xx} = \sum_{i=1}^{n} (X_i - \bar{X})^2$ $\quad$ www.matstat.org

Parametriska punktskattningar, forts.

Fördelning	Parameter	Skattning för parametern	Metod	Blom
$X \sim N(\mu_x, \sigma)$ $Y \sim N(\mu_y, \sigma)$	μ_x	$\mu_x^* = \bar{X}$	ML	289
	μ_y	$\mu_y^* = \bar{Y}$	ML	289
	σ^2	$S^2 = \frac{S_{xx} + S_{yy}}{(n_x - 1) + (n_y - 1)}$	ML	292
$X_1 \sim Bin(n_1, p)$ $X_2 \sim Bin(n_2, p)$	p	$p^* = \frac{X_1 + X_2}{n_1 + n_2}$	ML	293
$X_1 \sim Po(\lambda \cdot t_1)$ $X_2 \sim Po(\lambda \cdot t_2)$	λ	$\lambda^* = \frac{X_1 + X_2}{t_1 + t_2}$	ML	Problem 7.2.19

$\bar{X} = \sum_{i=1}^{n} X_i$ $\quad S_{xx} = \sum_{i=1}^{n} (X_i - \bar{X})^2$

Medelfelet

Medelfelet är en approximation för standardavvikelsen av en skattning (större medelfel → större osäkerhet).

1. Medelfelet för skattning av μ in $N(\mu, \sigma)$:

$\mu^* = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{X}$ skattning för μ i $N(\mu, \sigma)$

$V(\bar{X}) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n^2} \cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}$ skattningens varians

$D(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ standardavvikelse för skattningen av μ, men σ okänd!

⇩ $\sigma \rightarrow s$ (σ ersätts med sin skattning s) $\quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$

$d(\bar{X}) = \frac{s}{\sqrt{n}}$ medelfelet, numeriskt värde kan beräknas, approximering för standardavvikelsen (för skattning av μ i $N(\mu, \sigma)$ med $\bar{X}$)

Medelfelet

2. Medelfelet för skattning av p in Bin(n, p):

$p^* = \frac{X}{n}$ skattar p i $Bin(n, p)$

$$V(p^*) = V\left(\frac{X}{n}\right) = \frac{1}{n^2}V(X) = \frac{1}{n^2} \cdot n \cdot p \cdot (1-p) = \frac{p \cdot (1-p)}{n}$$

$D(p^*) = \sqrt{\frac{p \cdot (1-p)}{n}}$ men p okänd (det är ju parametern vi vill skatta)

$p \implies p^* = \frac{x}{n}$ ersätter p med skattning för p

$d(p^*) = \sqrt{\frac{p^* \cdot (1-p^*)}{n}} = \sqrt{\frac{\frac{x}{n} \cdot (1-\frac{x}{n})}{n}}$ medelfelet, numeriskt värde kan beräknas

$d(p^*)$ = medelfelet för skattning av p i $Bin(n, p)$ med $p^* = {}^{x}/_{n}$

Medelfelet för skattning av en proportion

Stickprov: $n = 1000$ personer, $x = 350$ svarade ”Ja” (observation).
Sökes: skattning för andel ”Ja”-svarare i hela populationen och motsvarande medelfel.

Observation: $p^*_{obs} = x/n = 35\%$

Slumpvariabel $p^* = X/n$ där $X \sim Bin(n, p)$

Medelfelet:

$$d(\hat{p}) = \sqrt{\frac{p_{obs} \cdot (1-p_{obs})}{n}} = \sqrt{\frac{0.35 \cdot (1-0.35)}{1000}} = 1.5\%$$

Resultat: andelen ”Ja”-svarare: $35 \pm 1.5\%$

Medelfelet

Medelfelet fås genom att ersätta okända parametrar i formeln för standardavvikelsen av en skattning med skattningar av dessa parametrar:

Förd.	att skatta	estimator	Standardavv. estimatorn	ersätta	Medelfelet hos estimatorn
$N(\mu, \sigma)$	μ	$\bar{X}$	$\frac{\sigma}{\sqrt{n}}$	$\sigma \to s$	$d = \frac{s}{\sqrt{n}}$
$Bin(n, p)$	p	$\frac{x}{n}$	$\sqrt{\frac{p \cdot (1-p)}{n}}$	$p \to p^*$	$\sqrt{\frac{p^* \cdot (1-p^*)}{n}}$

$p^* = \frac{x}{n}$

www.matstat.org