

## F9 STATISTIK, PUNKTSKATTNING

Kap.  
10.1-10.4, 11.1-11.9 | Linnéa Gustafsson  
linneag2@kth.se

- Beskrivande statistik
- Lägesmått, spridning och korrelation
- Definition av punktskattning
- Konsistens
- Maximum likelihood-metoden för punktskattning

## Kap. 10 STATISTIK

SANNOLIKHETSLÄRA	STATISTIK
Väntevärde $\mu$	Medelvärdet $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
Varians $\sigma^2$	Stickprovsvariancen $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$
<i>Hur stor betydelse standardavvikelsen har i relation till väntevärdet</i>	Populationsvariancen $= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$
Variationskoefficienten $R(X) = \frac{D(X)}{E(X)}$	<b>OBS!</b> Blanda ej ihop på räknaren. (Använder nästan bara stickprovsvariancen i denna kurs.) + finns andra ev. problem.
Medianen / mittvärdelet $F_X(\tilde{x}) = \frac{1}{2}$	Variationskoefficienten $= \frac{s}{\bar{x}} \cdot 100$ Uttrycks i procent
Kovariansen $C[X, Y] = E[(X - \mu_X)(Y - \mu_Y)]$	Medianen / mittvärdelet $\tilde{x}$
Korrelationskoefficienten $r[X, Y] = \frac{C[X, Y]}{D(X)D(Y)}$	Kovariansen $c_{x,y} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$
	Korrelationskoefficienten $r_{x,y} = \frac{c_{x,y}}{s_x s_y}$

### PRESERATION AV DATA

Anta att vi har följande data som är ogrupperade:

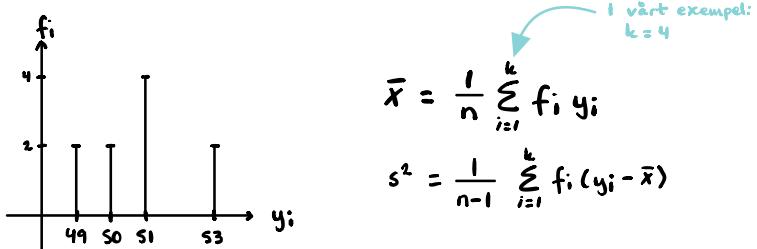
51	49	51	50	49	53	50	53	51	51
$x_1$	$x_2$	$x_3$		...			$x_9$	$x_{10}$	

### Grupperade data

absoluta frekvensen	relativa frekvensen
$y_1 = 49$	$f_1 = 2$
$y_2 = 50$	$f_2 = 2$
$y_3 = 51$	$f_3 = 4$
$y_4 = 53$	$f_4 = 2$
	$p_1 = \frac{2}{10} = 0.2$
	$p_2 = 0.2$
	$p_3 = 0.4$
	$p_4 = 0.2$

Som sannolikhetsfunktion fast inte teoretiska värdena utan vad vi faktiskt fick

Vi presenterar data m.h.a. stolpdiagram:



### Klassindelade data

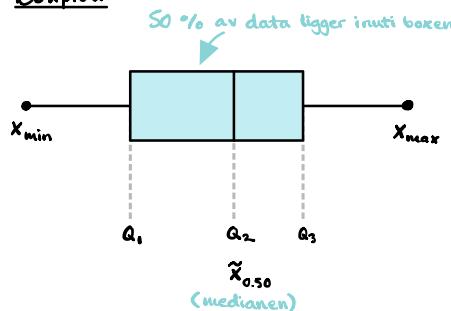
Anta att vi har många data. Då delar vi in dem i klasser, där varje klass innehåller de data som ligger inom ett visst interval.

- 1) Vi räknar som om alla data har klassmittens värde.
- 2) Inga öppna klasser      En klass kan t.ex. inte vara 50-oo
- 3) Klassbredden bör vara konstant
- 4) Vi ritar histogram där antalet data är proportionellt mot arean av varje rektangel.  
Om 3) uppfylls: proportionellt mot höjden

### Histogram, ex:



### Boxplot



$Q_1$  är första kvartilen; 25 % av data ligger till vänster  
 $Q_2 = \bar{x}$  är andra kvartilen; 50 % ————— || —————  
 $Q_3$  är tredje kvartilen; 75 % ————— || —————

$(Q_1, Q_3)$  är kvartilintervallet

$Q_3 - Q_1$  är kvartilavståndet

$(x_{\min}, x_{\max})$  är variationsintervallet

$x_{\max} - x_{\min}$  är variationsbredden

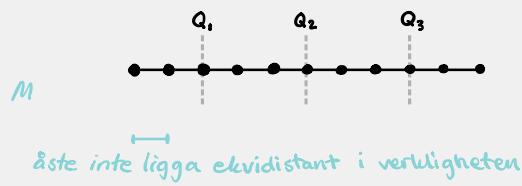
$Q_1$  kallas också 25 %-percentilen  
 $Q_2$  ————— || ————— 50 % ————— || —————  
 $Q_3$  ————— || ————— 75 % ————— || —————

Hur tar man fram  $Q_1$ ?

Om vi har  $n$  st mätdata borde det vara mätdata  $x_k$  där  $\frac{k}{n} = 0.25$

Vi väljer  $k$  till det heltal som uppfyller  $0.25n \leq k \leq 0.25n + 1$  Eftersom  $k$  inte alltid är delbart med  $n$

Ex | Anta  $n=11$

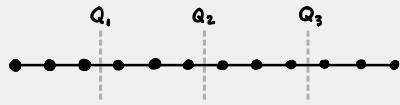


$$\begin{array}{l} Q_1 \\ Q_2 \\ Q_3 \end{array} \quad \left. \begin{array}{l} 0.25 \cdot 11 = \frac{11}{4} = 2.75 \\ 0.25 \cdot 11 + 1 = 3.75 \end{array} \right\} \Rightarrow k=3 \Rightarrow Q_1 = x_3$$

$$\begin{array}{l} Q_1 \\ Q_2 \\ Q_3 \end{array} \quad \left. \begin{array}{l} 0.75 \cdot 11 = \frac{33}{4} = 8.25 \\ 0.75 \cdot 11 + 1 = 9.25 \end{array} \right\} \Rightarrow k=9 \Rightarrow Q_3 = x_9$$

$$Q_2 \quad \dots \quad \Rightarrow Q_2 = x_6$$

Ex | Anta  $n=12$



$$\begin{array}{l} Q_1 \\ Q_2 \\ Q_3 \end{array} \quad \left. \begin{array}{l} 0.25 \cdot 12 = 3 \\ 0.25 \cdot 12 + 1 = 4 \end{array} \right.$$

Nu har vi två heltal som uppfyller  
 $0.25n \leq k \leq 0.25n + 1$

Då väljer vi  $Q_1 = \frac{x_3 + x_4}{2}$

P.S.S. för  $Q_2, Q_3$

Fungerar på samma sätt för alla percentiler (inte bara kvartiler).

5%-percentilen: välj det  $x_k$  där  $k$  uppfyller  $0.05n \leq k \leq 0.05n + 1$

## Kap. 11 PUNKTSKATTNING

En skattning av  $\theta$  kallas  $\theta_{\text{obs}}^*$  och är ett utfall av den stokastiska variabeln  $\theta^*$ .

$\theta^*$  kallas även stickprovsvariabeln;  $\theta^* = \theta^*(X_1, X_2, \dots, X_n)$

$\theta_{\text{obs}}^*$  kallas även punktskattningen av  $\theta$ ;  $\theta_{\text{obs}}^* = \theta_{\text{obs}}^*(x_1, x_2, \dots, x_n)$   
alla mätdata

$$\text{Ex 1} \quad E(X_i) = \mu$$

$$\mu_{\text{obs}}^* = \bar{x}$$

$$\mu^* = \bar{X}$$

$$\text{Ex 2} \quad E(X_i) = \mu$$

$$\mu_{\text{obs}}^* = \frac{2x_1 + 9x_2}{11}$$

$$\mu^* = \frac{2\bar{X}_1 + 9\bar{X}_2}{11}$$

$$\text{Ex 3} \quad D(X_i) = \sigma^2$$

$$\sigma_{\text{obs}}^* = s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

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

### SKATTNINGAR: VANLIGA FÖRDELNINGAR

#### Diskreta

👑 Binomialfördelningen  $X \in \text{Bin}(n, p)$

$$P_{\text{obs}}^* = \frac{x}{n}$$

👑 Hypergeometriska fördelningen  $X \in \text{Hyp}(N, n, p)$

$$P_{\text{obs}}^* = \frac{x}{n}$$

👑 Poissonfördelningen  $X \in \text{Po}(\mu)$

$$\text{Enl. F.S. § 3: } \mu = E(X_i)$$

$$\mu_{\text{obs}}^* = \bar{x}$$

👑 "För-första-gången"- fördelningen  $X \in \text{ffg}(p)$

$$E(X_i) = \frac{1}{p}$$

$$\text{Enl. F.S. § 3: } E(X) = \frac{1}{p}$$

$$P_{\text{obs}}^* = \frac{1}{\bar{x}}$$

Kontinuerliga  Likformiga fördelningar: finns inget enkelt sätt att skatta

 Exponentialfördelningen  $X \in \text{Exp}(\lambda)$

$$\text{Enl. F.S. § 4: } E(X_i) = \frac{1}{\lambda} \quad \text{d.v.s.} \quad \lambda = \frac{1}{E(X_i)}$$

$$\lambda_{\text{obs}}^* = \frac{1}{\bar{x}}$$

 Normalfördelningen  $X \in N(\mu, \sigma)$

$$\mu_{\text{obs}}^* = \bar{x}$$

$$\sigma_{\text{obs}}^* = s$$

### KONSISTENS

Def. En skattning är konsistent  $\Leftrightarrow \lim_{n \rightarrow \infty} P(|\theta_n^* - \theta| > \varepsilon) \rightarrow 0, \theta_n^* = \theta^*(X_1, \dots, X_n)$

### MAXIMUM-LIKELIHOODMETODEN

"Största trolighets-metoden"

Idén är att eftersom man fått de mätdata man fått så borde sannolikheten vara stor att få just dessa mätdata. Då maximeras vi denna sannolikhet w.a.p. den parameter vi vill skatta. Det värde på parameter som maximeras denna sannolikhet sätts till ML-skattningen.

Ex 11.10 Anta att vi vet att  $X_i$ :na är obero. och  $\in Po(\mu)$ .

Vi får 5 mätdata 10, 12, 7, 10, 4.

Vi vill ta fram ML-skattningen av  $\mu$ .

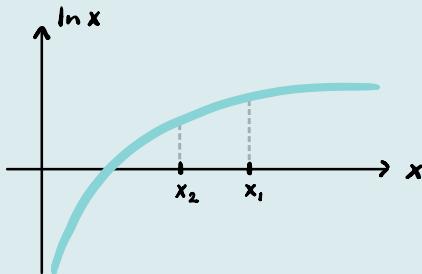
$$P(X_1=10 \cap X_2=12 \cap \dots \cap X_5=4) = [\text{enl. F.S. § 9.1}] = L(\mu) = [\text{ober.}] =$$

$$= P_{X_1}(10) \cdot P_{X_2}(12) \cdot \dots \cdot P_{X_5}(4) = [\text{se. F.S. § 3}] = \frac{\mu^{10}}{10!} e^{-\mu} \cdot \frac{\mu^{12}}{12!} e^{-\mu} \cdot \dots \cdot \frac{\mu^4}{4!} e^{-\mu}$$

Det  $\mu$  som maximeras  $L(\mu)$  väljs till ML-skattning av  $\mu$ .

FORTSÄTTNING ↓

### Metod att ta fram $\mu_{\text{obsML}}^*$



$$x_1 > x_2 \Leftrightarrow \ln x_1 > \ln x_2$$

$$f(x_1) > f(x_2) \Leftrightarrow \ln f(x_1) > \ln f(x_2)$$

d.v.s. där  $\ln f(x)$  har max har även  $f(x)$  sitt max, ty  $\ln$ -funktionen är kontinuerlig och strängt växande.

$$L(\mu) = \frac{\mu^{10+12+\dots+4}}{10! 12! \dots 4!} e^{-5\mu} = \frac{\mu^{43}}{10! 12! \dots 4!} e^{-5\mu}$$

$$\ln(L(\mu)) = \ln \mu^{43} + \ln e^{-5\mu} - \ln(10! 12! \dots 4!) = 43 \ln \mu - 5\mu - \ln(10! 12! \dots 4!)$$

$$\frac{d}{d\mu} \ln(L(\mu)) = \frac{43}{\mu} - 5 = 0 \quad \text{Ska eg. även kolla 2:a derivata så det verkligen är max}$$

$$\Rightarrow \mu = \frac{43}{5} \text{ maximiserar } L(\mu)$$

$$\Rightarrow \underline{\underline{\mu_{\text{obsML}}^* = \frac{43}{5}}} \quad \text{Ej "räta" } \mu \text{ som är } \frac{43}{5} \text{ utan vår ML-skattning } \underline{\underline{\mu_{\text{obsML}}^*}}$$

$$\mu_{\text{obsML}}^* = \sum \frac{x_i}{n}$$

### $L(\theta)$

Vid diskret fördelning

$$L(\theta) = p_{x_1, x_2, \dots, x_n}^{(x_1, x_2, \dots, x_n; \theta)}$$

Vid kontinuerlig fördelning

$$L(\theta) = f_{x_1, x_2, \dots, x_n}^{(x_1, x_2, \dots, x_n; \theta)}$$

Kärnan med ML-skattning:

Hitta det  $\theta$  som maximiserar likelihoodfunktionen  $L(\theta)$ .