

Industriell matematik och statistik, LMA136 2013/14

29 Januari 2014

Disposition

Repetition

Beskrivande
statistik

Data
Statistik

Repetition

Beskrivande statistik

Data

Statistik

Kontinuerlig stokastisk variabel

Repetition

Beskrivande
statistik

Data
Statistik

- ▶ Fördelningsfunktion $P(X \leq x) = \int_{-\infty}^x f(\xi) d\xi$
- ▶ Frekvensfunktion $f(x)$
 - 1) $f(x) \geq 0$ för alla x
 - 2) $\int_{-\infty}^{\infty} f(x) dx = 1$
- ▶ Väntevärde, $\mathbb{E}[X] = \int_{-\infty}^{\infty} x f(x) dx$
- ▶ Varians, $\text{Var}(X) = \mathbb{E}[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$
- ▶ Kvantil, $P(X \leq \alpha_q) = q$
- ▶ Rektangelfördelning, Exponentialfördelning, Normalfördelningen
- ▶ X är $N(\mu, \sigma^2)$ -fördelad om

$$f(x) = \frac{e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sqrt{2\pi}\sigma}.$$

Repetition

Beskrivande statistik

Data Statistik

Data är en samling fakta ur vilken man kan erhålla information.

Statistik är vetenskapen om att:

- ▶ samla in data,
- ▶ analysera data,
- ▶ tolka data.

- ▶ Kvalitativ data; data som kan sorteras i kategorier men har inte en väldefinierad mätskala
t.ex. ögonfärg, utbildningsnivå, bra/dåligt, av/på.
- ▶ Kvantitativ data; data som har en väldefinierad mätskala
t.ex. längd, ålder, antal, pris

En punkt i en datamängd kallas för en observation.

Data {22, 21, 20, 23, ...}

Ålder	Antal	Relativ frekvens
19	4	$4/28 \approx 0.143$
20	5	$5/28 \approx 0.179$
$\vdots$		
Summa:	28	1

Datat åskådliggörs i form av diagram

- ▶ stapeldiagram
- ▶ cirkeldiagram
- ▶ histogram

- ▶ läge; var ligger datapunkterna huvudsakligen?
- ▶ spridning; hur stor variation finns i datat?
- ▶ avvikare; finns det något eller några värden som avviker betydligt från majoriteten?
- ▶ form; liknar datan någon formen för någon av de parametriska fördelningarna?

Data $\{x_1, x_2, \dots, x_n\}$

- ▶ Typvärde; det värde som förekommer flest gånger,
- ▶ Median; det mittersta värdet, det värde som delar datat i två lika stora delar, om man har ett jämnt antal datavärden brukar man välja medianen som medelvärde av de två mittersta värdena
- ▶ Aritmetiskt medelvärde; summan av alla värden dividerat med antalet värden, betecknas $\bar{x}$

Data $\{x_1, x_2, \dots, x_n\}$

- ▶ variationsbredd, $R = \max(x_i) - \min_i(x_i)$
- ▶ Stickprovsvarians, $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$
- ▶ Kvartilavståndet, $Q = q_{75\%} - q_{25\%}$

Empirisk fördelningsfunktion

Data $\{x_1, x_2, \dots, x_n\}$

$$F_{emp}(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{x_i \leq x}$$

Om Y är en diskret stokastisk variabel likformigt fördelad på $\{x_1, x_2, \dots, x_n\}$ så gäller

$$P(Y \leq x) = F_{emp}(x)$$

Sorterad data $\{y_1, y_2, \dots, y_n\}$

$$P(Y \leq y_1) = 1/n$$

$$P(Y \leq y_2) = 2/n$$

$$\vdots$$

$$P(Y \leq y_n) = 1$$

Repetition

Beskrivande
statistikData
Statistik

Hur väl passar en modell data?

Data $\{x_1, x_2, \dots, x_n\}$ (sorterad data $\{y_1, y_2, \dots, y_n\}$)

- ▶ Modellen definerar en fördelningsfunktion $P(X \leq x)$

t.ex. om vi skattat medel samt standardavvikelse och vill testa om datan är

normalfördelad då gäller att $P(X \leq x) = \int_{-\infty}^x \frac{e^{-\frac{1}{2}(\frac{x-\bar{x}}{s})^2}}{\sqrt{2\pi}s} dx$

- ▶ Kvantil, $q = P(X \leq \alpha_q)$

QQ-plot $(\alpha_{i/n}, y_i)$

PP-plot $(i/n, P(X \leq y_i))$