

# SANNOLIKHETSLÄRA OCH STATISTIK

## FÖRELÄSNING 2

Mattias Villani

**Avdelningen för Statistik och Maskininlärning  
Institutionen för datavetenskap  
Linköpings universitet**

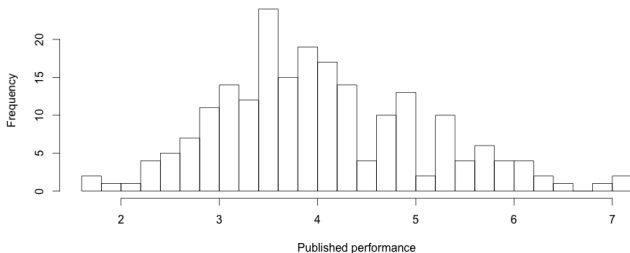


# ÖVERSIKT

- ▶ Deskriptiv statistik
- ▶ Slumpvariabler
- ▶ Sannolikhetsfördelning
- ▶ Väntevärde och varians
- ▶ Kovarians och korrelation
- ▶ Chebyshevs olikhet

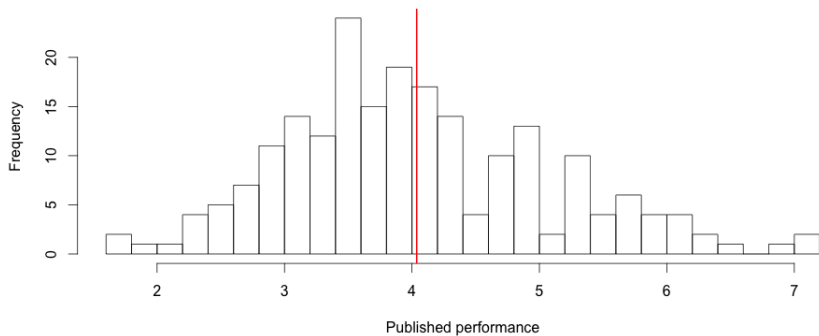
# DESKRIPTIV STATISTIK

- ▶ **Mätningar:**  $x_1, x_2, \dots, x_n$ .
- ▶ Exempel: Prestanda för  $n = 209$  datorer.
- ▶ **Medelvärde:**  $\bar{x} = 4.037$ .
- ▶ **Histogram**

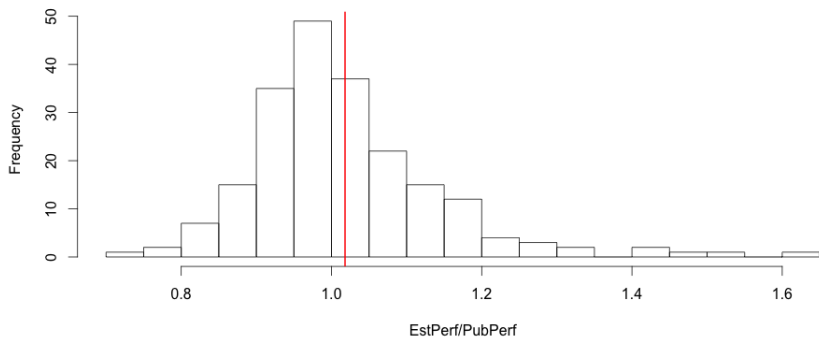


- ▶ 10 mätningar  $> 6$  , dvs ca 2.8% (10/209) av mätningarna hade hög prestanda ( $>6$ ).

# DESKRIPTIV STATISTIK, FORTS



# DESKRIPTIV STATISTIK, FORTS



# SLUMPVARIABLER

**Definition.** En **slumpvariabel**  $X$  är en funktion från utfallsrummet  $\Omega$  till  $\mathbb{R}$

$$X = f(\omega)$$

där  $\omega \in \Omega$  är ett utfall.

- ▶ Slumpvariabler är **praktiska**: vi bryr oss ofta bara om enklare variabler ( $X$ ) vars utfall är en funktion av den underliggande slumpen  $\omega$ .
- ▶ Två typer av slumpvariabler:
  - ▶ **Kontinuerlig**:  $X$  antar värden i  $\mathbb{R}$  (eller  $(0, 1)$ ). Längdhopp.
  - ▶ **Diskret**:  $X$  antar ett ändligt (t ex  $\{0, 1, 2, \dots, n\}$ ) eller uppräknligt ( $\{0, 1, 2, \dots\}$ ) antal värden. Höjdhopp.
- ▶ Ett annat ord för slumpvariabel (eng. random variable) är **stokastisk variabel** (eng. stochastic variable).
- ▶ Funktionen  $f()$  måste vara **mätbar**. Teknikalitet. Hänger ihop med **sigma-algebra** (se sid 14-15 i Baron). **Måtteori**.

# SLUMPVARIABLER - NÅGRA EXEMPEL

**Ex** Kasta två tärningar.

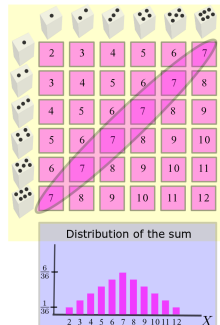
- ▶  $\Omega = \{(1, 1), (1, 2), \dots, (6, 5), (6, 6)\}.$
- ▶  $X =$  antalet prickar på två kast.

**Ex** Singla två mynt.

- ▶  $\Omega = \{(H, H), (H, T), (T, H), (T, T)\}.$
- ▶  $X =$  antalet  $H$  (krona).  $X$  kan anta värdena 0, 1, 2.
  - ▶  $P(X = 0) = 1/4$
  - ▶  $P(X = 1) = 1/2$
  - ▶  $P(X = 2) = 1/4.$

**Ex** Flyga quadcopter.

- ▶  $\Omega =$  Abstrakt utfallsrum med alla möjliga utfall på faktorer som bestämmer quadcopterns resväg.
- ▶  $X =$  Tre-dimensionella koordinater  $(x, y, z)$  över quadcopterns position vid tidpunkt  $t$ .



# SANNOLIKHETSFÖRDELNING

**Definition.** (sannolikhets)fördelningen för en slumpvariabel  $X$  är sannolikheterna för alla dess utfall, dvs

$$P(x) = \mathbf{P}\{X = x\}$$

för alla möjliga utfall  $x$ .

- ▶ Stora och små bokstäver spelar roll:
  - ▶  $X$  är **slumpvariabeln**. Ex. summan av två tärningarna
  - ▶  $x$  är ett **givet utfall**. Ex. 7 prickar.
- ▶ Fet stil eller ej spelar roll:
  - ▶  $\mathbf{P}$  är sannolikheten för ett givet utfall.  $\mathbf{P}\{X = x\}$  betyder egentligen 'Sannolikheten för alla de utfall  $[(1, 6), (2, 5)]$  etc] som ger summan 7'.
  - ▶  $P(x)$  är en enkel reellvärd funktion, precis som i vanlig analys.
- ▶ För diskreta slumpvariabler kallas  $P(x)$  ofta för **pmf**:en (probability mass function).
- ▶ Slumpvariabelns **support**:  $\{x : P(x) > 0\}$ .



# FÖRDELNINGSFUNKTION

**Definition.** **Fördelningsfunktionen** för en slumpvariabel  $X$  definieras som

$$F(x) = \mathbf{P}\{X \leq x\} = \sum_{y \leq x} P(y).$$

- En sannolikhetsfördelning summerar till 1:

$$\sum_{\text{alla } x} P(x) = \sum_{\text{alla } x} \mathbf{P}\{X = x\} = 1.$$

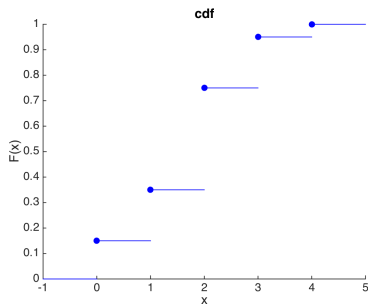
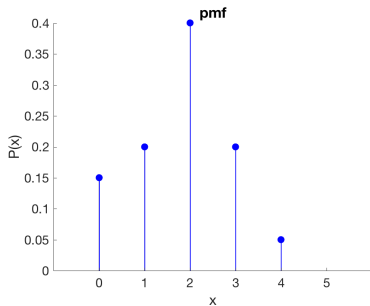
- Fördelningsfunktionen är icke-avtagande mellan 0 och 1:

$$\lim_{x \downarrow -\infty} F(x) = 0 \qquad \lim_{x \uparrow +\infty} F(x) = 1.$$

- Fördelningsfunktionen kallas också för den **kumulativa täthetsfunktionen** (cumulative density function), eller **cdf**.

# SANNOLIKHETS- OCH FÖRDELNINGSFUNKTION

$x$	0	1	2	3	4
$P(x)$	0.15	0.20	0.40	0.20	0.05
$F(x)$	0.15	0.35	0.75	0.95	1.00



# SIMULTANFÖRDELNING

- ▶ Låt  $X$  och  $Y$  vara slumpvariabler.
- ▶  $(X, Y)$  är en **slumpvektor** med typiskt utfall  $(x, y)$ .
- ▶ Fördelningen för  $(X, Y)$  kallas **simultanfördelning**.

$$P(x, y) = \mathbf{P}\{(X, Y) = (x, y)\} = \mathbf{P}\{X = x \cap Y = y\}.$$

- ▶ Simultanfördelningen är en sannolikhetsfördelning:

$$\sum_x \sum_y P(x, y) = 1.$$

**Ex**  $X = \text{Spam/Ham}$  och  $Y = \text{Inbox/Spambox}$ .

	Spam	Ham
Inbox	0.02	0.88
Spambox	0.09	0.01

- ▶ Simultanfördelningen: 'Vad är sannolikheten att få ett ham-mejl och att det hamnar i spamboxen?'

# SIMULTANFÖRDELNING

**Ex**  $X$  =avkastning aktie X och  $Y$  =avkastning aktie Y.

		Aktie Y		
		Låg	Medel	Hög
Aktie X	Låg	0.05	0.05	0.15
	Medel	0.10	0.30	0.20
	Hög	0.05	0.05	0.05

- ▶ Aktieportfölj: 50% i aktie X och 50% i aktie Y.
- ▶ Simultanfördelningen: 'Vad är sannolikheten att min aktieportfölj får medelavkastning?'

# MARGINALFÖRDELNING

- ▶ Fördelningen för bara  $X$  kallas **marginalfördelningen** (för  $X$ ).
- ▶ Fördelningen för bara  $Y$  kallas marginalfördelningen (för  $Y$ ).
- ▶ Marginalfördelningen: 'Vad är sannolikheten att få ett spam-mejl (oavsett var det hamnar)?'
- ▶ Marginalfördelningen fås genom att summera ut den andra variabeln:

$$P_X(x) = \sum_y P(x, y)$$

$$P_Y(y) = \sum_x P(x, y)$$

- ▶ Jämför med Lagen om total sannolikhet (F1).

**Ex**  $X = \text{Spam/Ham}$  och  $Y = \text{Inbox/Spambox}$ .

	Spam	Ham	
Inbox	0.02	0.88	0.9
Spambox	0.09	0.01	0.1
	0.11	0.89	

# MARGINALFÖRDELNING

Ex  $X$  = avkastning aktie X och  $Y$  = avkastning aktie Y.

		Aktie Y			
		Låg	Medel	Hög	
Aktie X	Låg	0.05	0.05	0.15	0.25
	Medel	0.10	0.30	0.20	0.6
	Hög	0.05	0.05	0.05	0.15
		0.20	0.40	0.40	

- Vilka portföljandelar är optimala? Beslut under osäkerhet.

## OBEROENDE

**Definition.** Slumpvariablerna  $X$  och  $Y$  är **oberoende** om

$$P(x, y) = P_X(x) \cdot P_Y(y)$$

för **alla** värden på  $x$  och  $y$ .

**Ex**  $X = \text{Spam/Ham}$  och  $Y = \text{Inbox/Spambox}$ .

	Spam	Ham	
Inbox	0.02	0.88	0.9
Spambox	0.09	0.01	0.1
	0.11	0.89	

- ▶ Valet av box är inte oberoende av om mejlet är ham eller spam:

$$P(\text{inbox}) \cdot P(\text{ham}) = 0.9 \cdot 0.89 = 0.801 \neq 0.88 = P(\text{inbox, ham})$$

- ▶  $P(\text{inbox}|\text{ham}) = \frac{P(\text{inbox, ham})}{P(\text{ham})} = \frac{0.88}{0.89} = 0.988 > 0.9 = P(\text{inbox})$ .
- ▶ Lättare att gissa box om man vet att mejlet är ham.

# LÄGESMÅTT

- ▶ En sannolikhetsfördelning  $P(x)$  beskriver **all** osäkerhet om  $X$ .
- ▶ Kan vara komplicerat att förmedla hela  $P(x)$ , speciellt om  $X$  är en fler-dimensionell slumpvektor.
- ▶ Naturliga **lägesmått**:
  - ▶ **Median**,  $m$ .  $P(X \leq m) = 0.5$ . Hälften av sannolikhetsmassan ligger till vänster om  $m$ .
  - ▶ **Väntevärdet** (eng. expected value),  $\mu$  eller  $\mathbb{E}(X)$ , är det genomsnittliga värdet för  $X$ :

$$\mu = \mathbb{E}(X) = \sum_x x \cdot P(x).$$

- ▶ **Typvärdet** (eng. mode) är det mest sannolika värdet, dvs  $\operatorname{argmax}_x P(x)$ .

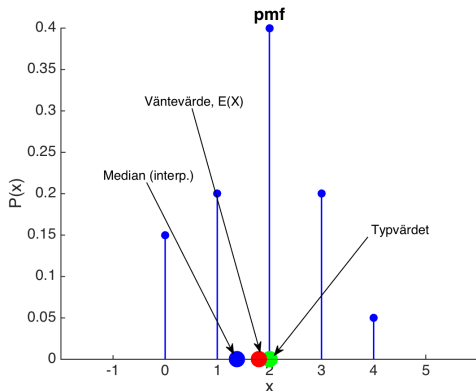


# LÄGESMÅTT - EXEMPEL

$x$	0	1	2	3	4
$P(x)$	0.15	0.20	0.40	0.20	0.05

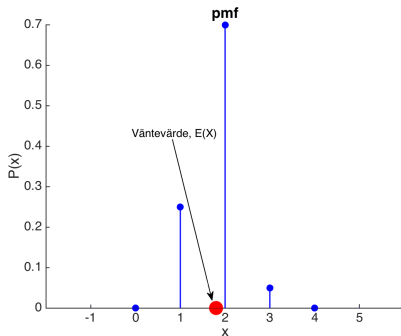
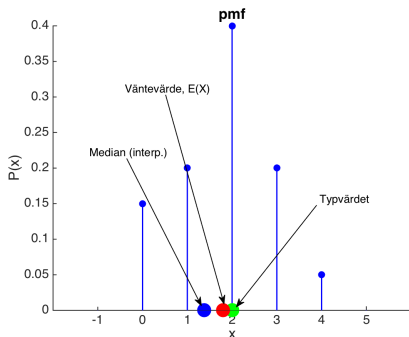
## ► Väntevärdet

$$\mathbb{E}(X) = 0 \cdot 0.15 + 1 \cdot 0.20 + 2 \cdot 0.40 + 3 \cdot 0.20 + 4 \cdot 0.05 = 1.8$$



# LÄGESMÅTT SÄGER INGET OM SPRIDNINGEN

- Väntevärdet är ett lägesmått. Ingen info om fördelningens spridning.



# VARIANS

- ▶ Storleken på avvikelserna  $x - \mathbb{E}(X)$  säger något om spridningen.
- ▶ Idé till spridningsmått: den förväntade avvikelserna:

$$\mathbb{E}(X - \mu) = \sum_x P(x) \cdot (X - \mu)$$

- ▶ Problem:  $\mathbb{E}(X - \mu)$  är alltid exakt noll ... Positiva och negativa avvikelser tar ut varandra.
- ▶ **Varians**: förväntade kvadrerade avvikelserna:

$$\sigma^2 = \text{Var}(X) = \mathbb{E}(X - \mu)^2 = \sum_x (x - \mu)^2 \cdot P(x).$$

- ▶ Alternativ formel

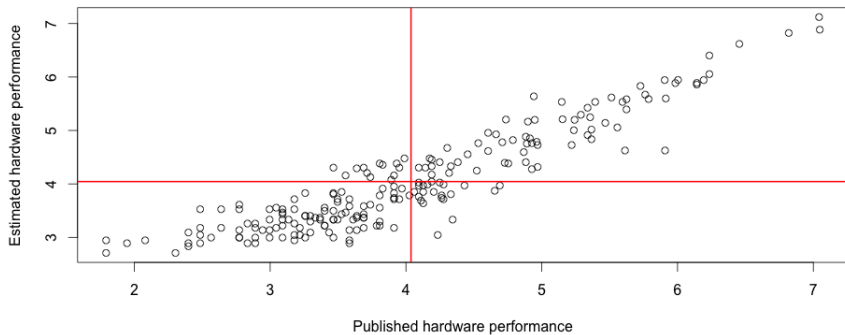
$$\text{Var}(X) = \mathbb{E}X^2 - \mu^2.$$

- ▶ **Standardavvikelse**:  $\sigma = \text{Std}(X) = \sqrt{\text{Var}(X)}$ . Samma skala som  $X$ .

# EGENSKAPER HOS VÄNTEVÄRDE OCH VARIANS

- ▶  $\mathbb{E}(c) = c$ , där  $c$  är en konstant.
- ▶  $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$ ,  $a, b$  konstanter.
- ▶  $\mathbb{E}(X + Y) = \mathbb{E}X + \mathbb{E}Y$
- ▶  $\mathbb{E}(aX + bY + c) = a\mathbb{E}(X) + b\mathbb{E}(Y) + c$ ,  $a, b, c$  konstanter.
- ▶  $\text{Var}(aX + b) = a^2 \cdot \text{Var}(X)$
- ▶ Om  $X$  och  $Y$  oberoende:  $\mathbb{E}(X \cdot Y) = \mathbb{E}(X) \cdot \mathbb{E}(Y)$
- ▶ Om  $X$  och  $Y$  oberoende:  $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$ .

# DESKRIPTIV STATISTIK - BEROENDE



# KOVIARIANS OCH KORRELATION

- ▶ Mått på **samvariation**. Sammanfattning av simultanfördelning.
- ▶ **Kovarians** mellan  $X$  och  $Y$ :

$$\sigma_{XY} = \text{Cov}(X, Y) = \mathbb{E} \{ (X - \mathbb{E}X) (Y - \mathbb{E}Y) \}$$

- ▶ Positiv kovarians:
  - ▶  $X$  tenderar att vara större än  $\mathbb{E}X$  samtidigt som  $Y$  tenderar att vara större än  $\mathbb{E}Y$ .
  - ▶  $X$  tenderar att vara mindre än  $\mathbb{E}X$  samtidigt som  $Y$  tenderar att vara mindre än  $\mathbb{E}Y$ .
- ▶ **Korrelationskoefficienten** mellan  $X$  och  $Y$

$$\rho = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\text{Std}(X) \cdot \text{Std}(Y)}.$$

- ▶  $-1 \leq \rho \leq 1$ .

# EGENSKAPER HOS KOVARIANS

- ▶  $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
- ▶  $\text{Var}(aX + bY + c) = a^2 \cdot \text{Var}(X) + b^2 \cdot \text{Var}(Y) + 2a \cdot b \cdot \text{Cov}(X, Y)$ ,  
 $a, b, c$  konstanter.
- ▶  $\text{Cov}(a \cdot X + b, c \cdot Y + d) = a \cdot c \cdot \text{Cov}(X, Y)$
- ▶ Om  $X$  och  $Y$  oberoende:  $\text{Cov}(X, Y) = 0$ .
- ▶ Om  $X$  och  $Y$  oberoende:  $\rho(X, Y) = 0$ .

# Chebyshevs Olikhet

- ▶ Väntevärdet  $\mu$  och Variansen  $\sigma^2$  innehåller information om sannolikhetsfördelningen.
- ▶ Chebyshevs olikhet: givet  $\mu$  och  $\sigma^2$  så kommer  $X$  ligga i intervallet  $[\mu - \varepsilon, \mu + \varepsilon]$  med en sannolikhet som är åtminstone  $1 - (\sigma/\varepsilon)^2$ .
- ▶ **Chebyshevs olikhet**

$$P\{|X - \mu| > \varepsilon\} \leq \left(\frac{\sigma}{\varepsilon}\right)^2.$$

- ▶ Notera att Chebyshevs olikhet endast kräver vetskap om  $\mu$  och  $\sigma^2$ . Inget andra egenskaper behövs (symmetri, skevhet).
- ▶ Men den lilla information har sitt pris:  $\left(\frac{\sigma}{\varepsilon}\right)^2$  är ofta bra mycket större än den sanna sannolikheten  $P\{|X - \mu| > \varepsilon\}$ .
- ▶ Chebyshevs olikhet är ofta nyttig i teoretiska sammanhang.