

# TDAB01 SANNOLIKHETSLÄRA OCH STATISTIK

Jose M. Peña  
IDA, Linköping University, Sweden

Föreläsning 1

# STATISTIK? ÄR INTE DET TABELLER, TYP?

- Statistik - läran om **osäkerhet**. Kommer det regna imorgon?
- **Sannolikheter** - osäkerhetens språk.  $P(\text{regn imorgon}) = 20\%$
- **Statistisk inferens** - att lära sig om osäkra händelser utifrån **data**.  $\Pr(\text{regn imorgon} \mid \text{regn idag, lågtryck idag}) = 45\%$
- Optimala **beslut** i en osäker värld. Ska jag boka en charterresa?

# SANNOLIKHETSLÄRA, STATISTIK OCH BESLUT

- **Sannolikhetslära:** Ett inkommande email är spam med sannolikheten 5%. Efter din semester har du 78 olästa mejl. Vad är sannolikheten att inget email är spam?
- **Statistik:** Efter din semester har du 78 olästa mejl. Två email visar sig vara spam. Vad är sannolikheten att ett godtyckligt email är spam?
- **Beslut under osäkerhet:** Ska nästa inkommande email skickas till spamkorgen?

# BYGG DITT EGET SPAMFILTER



- **Samla in träningsdata:** Läs in texten från dina inkomna mejl och beräkna intressanta kvantiteter (features) från varje mejl. T ex hur många \$-tecken? hur många 'viagra'? etc.
- **Formulera en sannolikhetsmodell** baserat på features:

$$\text{Pr(email is spam)} = \exp(\alpha + \beta \cdot n\text{Dollar} + \gamma \cdot n\text{Viagra})$$

- **Estimera modellen** på träningsdata. Hur beror spam-sannolikheten på antalet \$-tecken?
- **Prediktion.** Beräkna spam-sannolikheten för ett nytt mejl. [mejl -> features -> Pr(spam)]
- **Beslut.** Hur stor måste spam-sannolikheten vara för att mejlet ska skickas till spamkorgen?
- Se **DigitClassification.R**.



# EXEMPEL - ROBOTIK

- **Lokalisering.** Robot: 'Var är jag?'

Sannolikhetsfördelning över positioner som uppdateras vartefter med brusiga sensordata.



- **Räddningsrobot.**

Ta sig till olycksplats med objekt i vägen, t ex väggar, människor. K olika vägar, alla med olika förväntade färdtider. Osäkerhet. Beslut.



# ROBOTIK RIMMAR MED STATISTIK

*"As robotics is now moving into the open world, the issue of **uncertainty** has become a major stumbling block for the design of capable robot systems. **Managing uncertainty is possibly the most important step towards robust real-world robot systems.**"* från boken *"Probabilistic Robotics"* av Thrun et al.

*"**Statistics** provides the mathematical glue to integrate models and sensor measurements."* från boken *"Probabilistic Robotics"* av Thrun et al.

*"To date, **probabilistic robotics** is one of the most rapidly growing subfields of robotics. Probabilistic techniques have proven their value in practice. They are at the core of dozens of successful robotic systems to date"* från artikeln *"Is Robotics Going Statistics? The Field of Probabilistic Robotics"* av Thrun.

# EXEMPEL - MJUKVARUUTVECKLING

- **Hitta programmeringsbuggar.**

Beslut: Allokera bugg till rätt team.

Data: Text i buggrapport.

Osäkerhet: Sannolikhetsfördelning över kodblock.

$\Pr(\text{bugg i block 1}) = 10\%$ ,  $\Pr(\text{bugg i block 2}) = 30\%$  osv



- **Release-schema för mjukvara.**

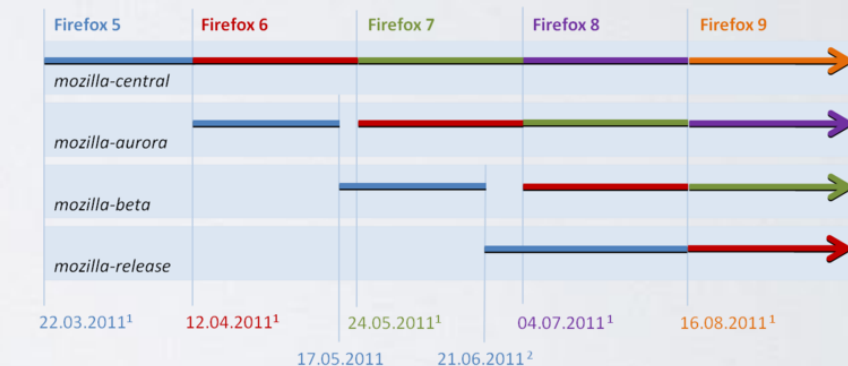
Beslut: Release om x antal månader.

Data: Antalet buggar i tidigare releaser, releasedatum, försäljningsdata vid tidigare releaser etc.

Osäkerhet: Sannolikhetsfördelning över antalet buggar vid olika releasedatum.

$\Pr(\text{minst en allvarlig bugg vid release om } x \text{ månader}) = 1/(1 + x^2)$

Sannolikhetsfördelning över antalet sålda licenser vid olika releasedatum.



¹ Entwicklungsstart der jeweiligen Vorabversion

² Veröffentlichung von Firefox 5



# DATAVETARE MÅSTE FÖRSTÅ STATISTIK



- “I keep saying the **sexy job** in the next ten years will be statisticians.”  
Hal Varian, Chief Economist, Google.
- “But the challenges for **massive data** go beyond the storage, indexing, and querying ... and, instead, hinge on the ambitious goal of inference. Inference is the problem of turning data into knowledge ... **Statistical rigor is necessary to justify the inferential leap from data to knowledge ...**”  
från rapporten *"Frontiers in Massive Data Analysis"*, US National Research Council.
- “**Computer scientists** involved in building big-data systems **must develop a deeper awareness of inferential issues**, while statisticians must concern themselves with scalability, algorithmic issues, and real-time decision-making.”  
från rapporten *"Frontiers in Massive Data Analysis"*, US National Research Council.



# SANNOLIKHETER

- Intuitivt: En **sannolikhet** beskriver chansen/risken att en händelse inträffar.

$\text{Pr}(\text{Norrköping vinner allsvenskan})=0.3$

$\text{Pr}(\text{inflationen överstiger 2\% om ett halvår}) = 0.05$

$\text{Pr}(\text{objektet vid position (x,y) är en bomb}) = 0.001$

$\text{Pr}(\text{person x har cancer}) = 0.01$

- Några saker att reda ut:
  - Vad är en **händelse** ?
  - Vilka **matematiska egenskaper** måste en sannolikhet ha för att de ska vara användbara (inte leda till paradoxer) ?
  - Hur ska en sannolikhet **tolkas** ?

# GRUNDLÄGGANDE TERMINOLOGI FÖR SANNOLIKHETER

- **Experiment.** Kasta två tärningar.



- **Utfall.**



- **Utfallsrum.** Mängden av alla utfall.  $S$  eller  $\Omega$

$$S = \{\text{Hammarby vinner, Halmstad vinner, Oavgjort}\}$$

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

- **Händelse.** En mängd av utfall. En delmängd av utfallsrummet.  $A \subset S$

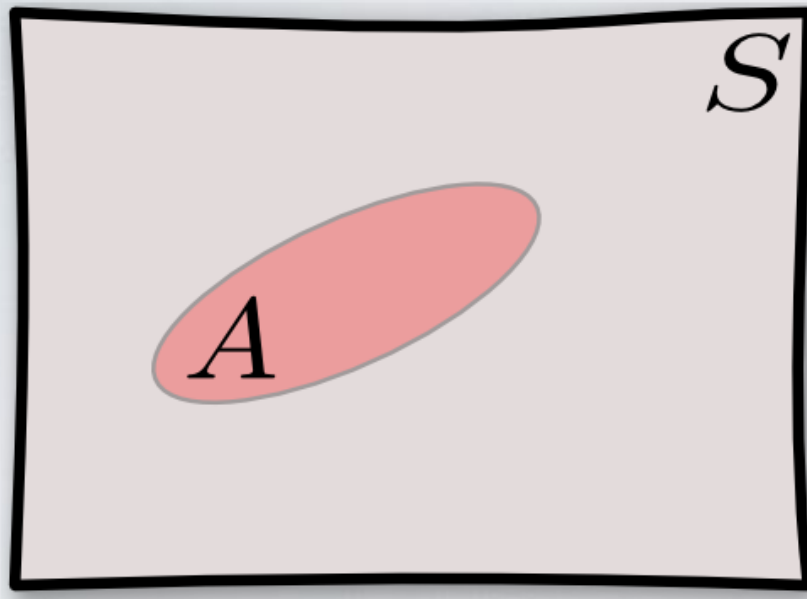
$A$  = Hammarby tar minst en poäng

$A$  = "Att få summan 7".

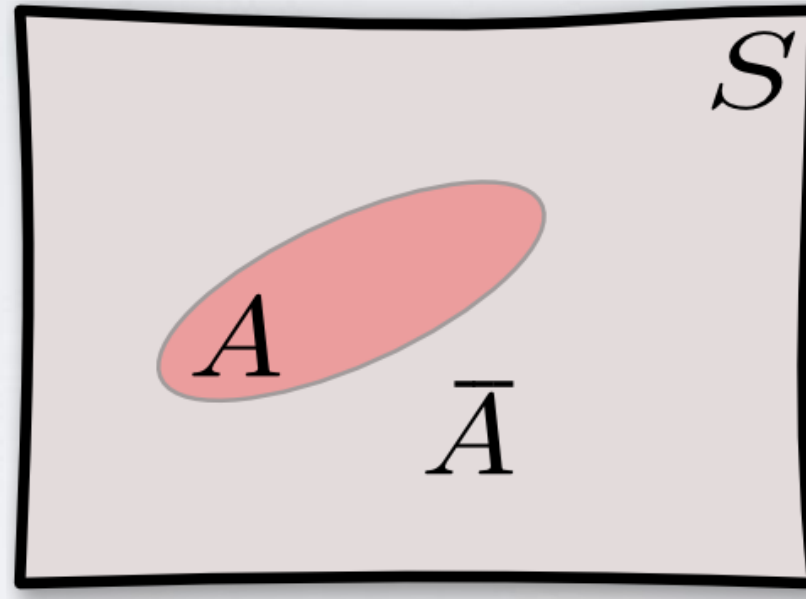
2	3	4	5	6	7
3	4	5	6	7	8
4	5	6	7	8	9
5	6	7	8	9	10
6	7	8	9	10	11
7	8	9	10	11	12

# MÄNGDLÄRA

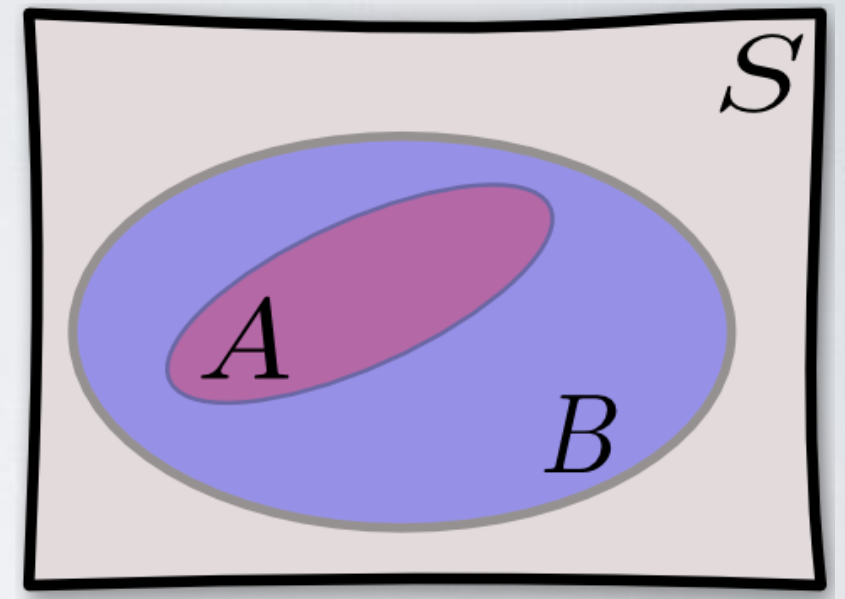
Händelsen  $A$   
ingår i **universalhändelsen**  $S$



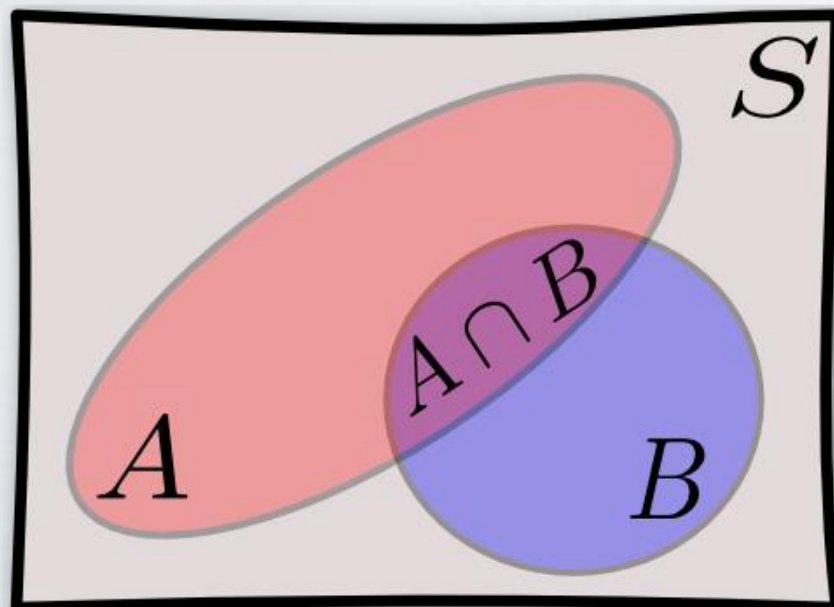
**Komplement** till  $A$   
De element som *inte* ingår i  $A$



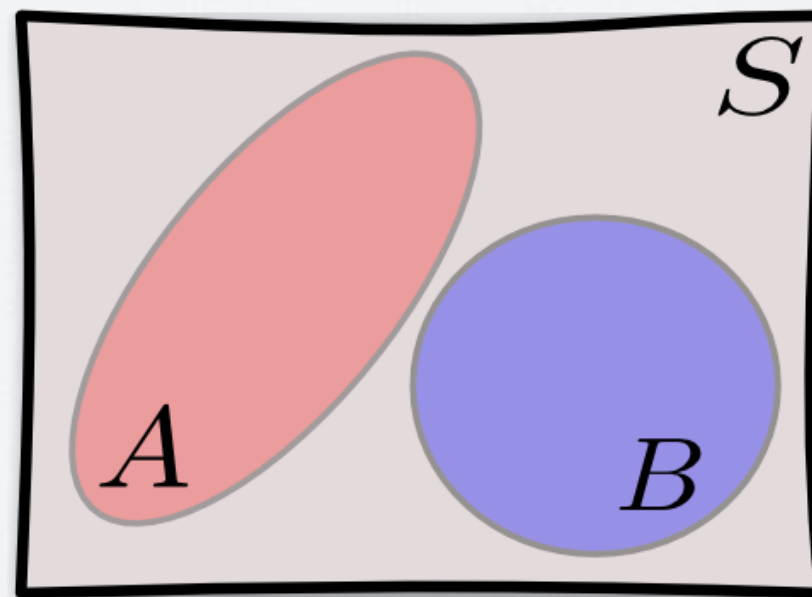
$A$  är en **delmängd** av  $B$   
Alla element i  $A$  är också i  $B$



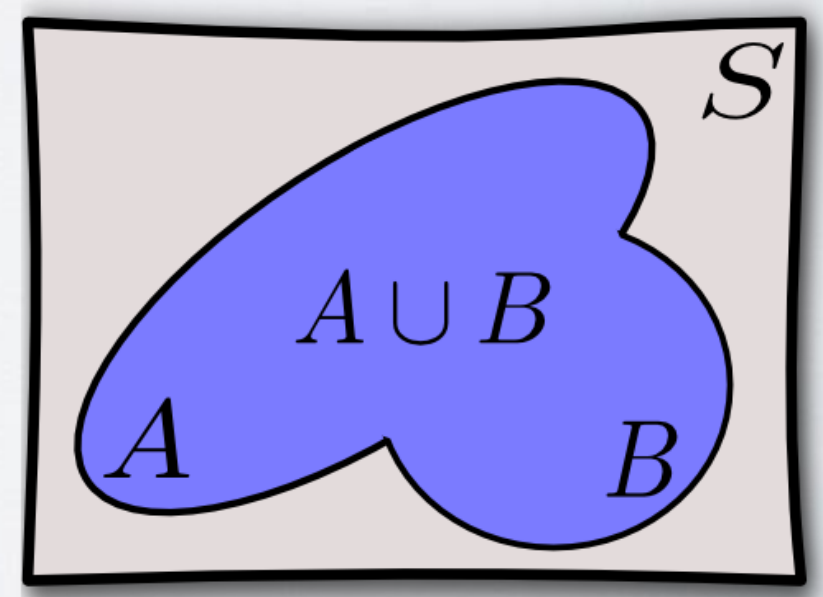
**Snittet** av  $A$  och  $B$   
Alla element som är i *både*  $A$  och  $B$



$A$  och  $B$  är **disjunkta** händelser  
 $A$  och  $B$  har inga gemensamma element



**Unionen** av  $A$  och  $B$   
Alla element som är i  $A$  och/eller  $B$



# RÄKNA MED SANNOLIKHETER

- **Universalhändelsen**, utfallsrummet:  $P(S) = 1$
- **Komplement**:  $P(\bar{A}) = 1 - P(A)$
- **Union**:  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- **Disjunkta händelser**:  $P(A \cup B) = P(A) + P(B)$
- **Oberoende händelser**: om  $A$  har inträffat eller ej påverkar inte  $P(B)$

$$P(A \cap B) = P(A) \cdot P(B)$$

- Från union till snitt. Från snitt till union.

$$\overline{A \cup B} = \bar{A} \cap \bar{B}$$

$$\overline{A \cap B} = \bar{A} \cup \bar{B}$$



# EXEMPEL 1

- $H = \{\text{hårddisk crashar}\}$ ,  $A = \{\text{backup A crashar}\}$  och  $B = \{\text{backup B crashar}\}$ .
- $P(H)=0.01$ ,  $P(A)=0.02$  och  $P(B)=0.02$ .
- Om  $H$ ,  $A$  och  $B$  är oberoende:

$$\begin{aligned} P(\text{fil sparad}) &= 1 - P(\text{fil borta}) = 1 - P(H \cap A \cap B) \\ &= 1 - P(H) \cdot P(A) \cdot P(B) \\ &= 1 - 0.01 \cdot 0.02 \cdot 0.02 = 0.999996 \end{aligned}$$

## EXEMPEL 2

- $A_1 = \{\text{bugg i kodblock 1}\}$ ,  $A_2 = \{\text{bugg i kodblock 2}\}$  och  $A_3 = \{\text{bugg i kodblock 3}\}$ .
- $P(A_1)=0.01$ ,  $P(A_2)=0.05$  och  $P(A_3)=0.01$ . Oberoende händelser.
- Vad är sannolikheten att programmet är fritt från buggar ?

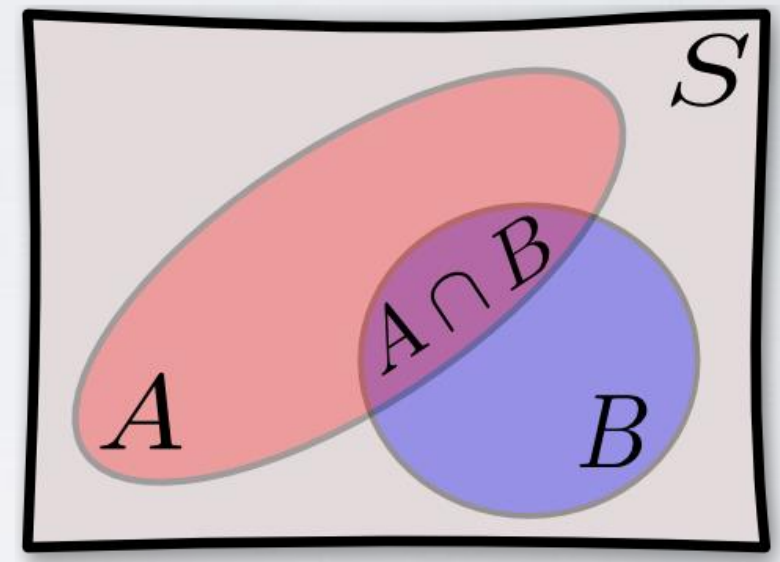
$$\begin{aligned} P(\text{buggfritt program}) &= P(\bar{A}_1 \cap \bar{A}_2 \cap \bar{A}_3) \\ &= P(\bar{A}_1) \cdot P(\bar{A}_2) \cdot P(\bar{A}_3) \\ &= (1 - 0.01) \cdot (1 - 0.05) \cdot (1 - 0.01) \\ &= 0.931 \end{aligned}$$

# BETINGADE SANNOLIKHETER

- Sannolikheten att händelse  $A$  inträffar givet att händelse  $B$  har inträffat.

- Notation:  $P(A|B)$

- Definition:  $P(A|B) = \frac{P(A \cap B)}{P(B)}$



- Händelse  $B$  har inträffat. **Universalmängden krymper** från  $S$  till  $B$ .
- Notera:  $P(A \cap B) = P(B) \cdot P(A|B)$
- **Oberoende** om  $B$  ger ingen information om  $A$ :  $P(A|B) = P(A)$

# BAYES SATS

- Ibland vet vi  $P(B|A)$ , men är intresserade av  $P(A|B)$ .

- Bayes sats: 
$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

- Exempel:  $A = \{\text{har sjukdom}\}$ ,  $B = \{\text{test positivt}\}$ .

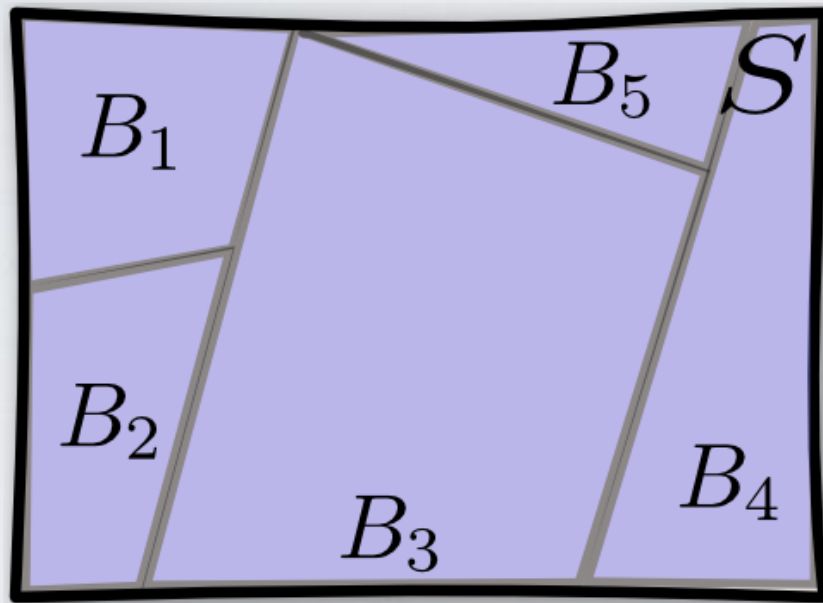
- $P(A|B) = P(\text{har sjukdom} \mid \text{test positivt})$

- $P(B|A) = P(\text{test positivt} \mid \text{har sjukdom})$

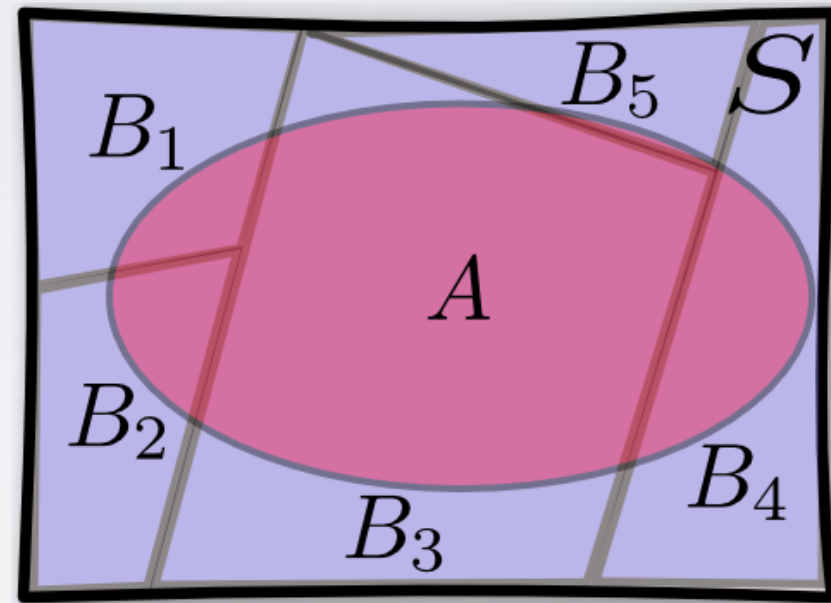


# LAGEN OM TOTAL SANNOLIKHET

$B_1, \dots, B_5$  är en **partitionering** av  $S$



Lagen om total sannolikhet



Partitionering = disjunkta och uttömmande delmängder

$$A = (A \cap B_1) \cup \dots \cup (A \cap B_5)$$

$$P(A) = P(A \cap B_1) + \dots + P(A \cap B_5)$$

$$P(A) = P(A|B_1) \cdot P(B_1) + \dots + P(A|B_5) \cdot P(B_5)$$

## BAYES SATS - ALTERNATIV FORM

- Bayes sats: 
$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

- Lagen om total sannolikhet ger

$$P(B) = P(B|A) \cdot P(A) + P(B|\bar{A}) \cdot P(\bar{A})$$

- Alternativ form av Bayes sats:

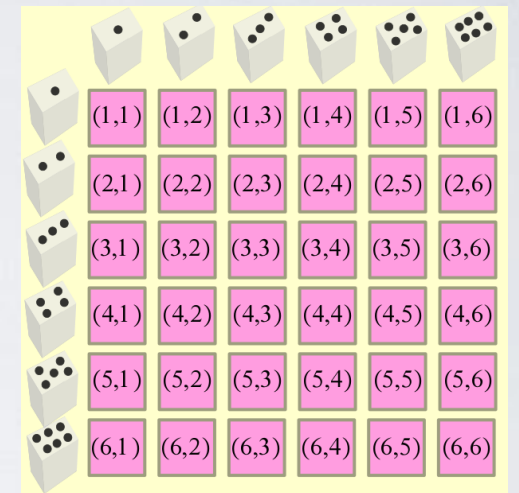
$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|\bar{A}) \cdot P(\bar{A})}$$

# LIKA SANNOLIKA UTFALL - KOMBINATORIK

- Låt  $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$  vara utfallsrummet för  $n$  lika sannolika utfall, t.ex. hasardspel (gambling), kortspel, tärningar, undersökningar.
- Vid lika sannolika utfall:

$$P(E) = \frac{\text{antalet utfall i } E}{\text{antalet utfall i } \Omega}$$

$$P(E) = \frac{\text{antalet lyckade utfall}}{\text{totala antalet utfall}} = \frac{\mathcal{N}_F}{\mathcal{N}_T}$$



(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

- Att räkna antalet möjligheter/utfall blir viktigt. T ex sannolikheten att jag gissar ditt lösenord ? Sannolikheten att 5 av 50 kvinnor är med i en undersökning av 10 av 100 människor ? **Kombinatorik**.

# PERMUTATIONER OCH KOMBINATIONER

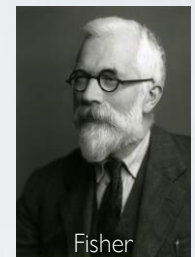
Välj ut  $k$  element från en mängd med  $n$  element

	Med återläggning	Utan återläggning
Med ordning	$n^k$	$\frac{n!}{(n-k)!}$
Utan ordning	$\frac{(k+n-1)!}{k!(n-1)!}$	$\frac{n!}{(n-k)!k!}$



# OLIKA TOLKNINGAR AV EN SANNOLIKHET

- **Lika sannolika händelser.** Kombinatorik.
- **Relativa frekvenser:**  $P(A)=0.25$  betyder att händelsen  $A$  kommer att inträffa 25% av antalet försök i genomsnitt.  
**Frekventistisk statistik.**

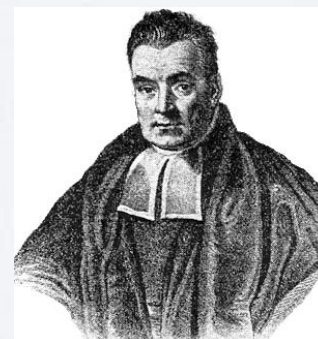


Fisher



Neyman

- **Subjektiv grad av tilltro.**  $P(A)=0.4$  betyder att du skulle acceptera vadet 'vinn 10 kr om  $A$  inträffar' om vadet kostade 4 kr eller mindre.  
**Bayesiansk statistik.**



Bayes



Turing