Statistik och Dataanalys I

Föreläsning 11 - Osäkerhet och Sannolikhet

Oskar Gustafsson

Statistiska institutionen
Stockholms universitet

Översikt

- Motivation
- Försök, Utfall och Händelser
- Sannolikheter
- Sannolikhetsberäkningar
- Kombinatorik

Sannolikheter för dataanalys

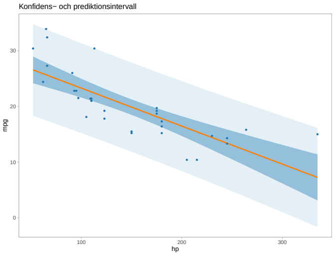
- **Sannolikhetslära** är intressant i sig:
  - Sannolikheten för en kärnkraftsolycka
  - Sannolikheten att två personer har identiska DNA.
  - Sannolikheten att träffa den rätta på dejtingapp.
- Sannolikhetslära viktigt för **dataanalys**:
  - **Statistiska modeller är sannolikhetsmodeller**.
    Bra modell av verkligheten: data sannolika enligt modellen.
  - Kan **kvantifiera osäkerheten** i en **prediktion**.
  - Kan **fatta optimala beslut i en osäker värld**.

 $\longrightarrow$

0	1	2	3	4	5	6	7	8	9
			0.01				0.99		

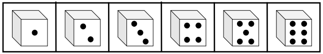
Sannolikheter för regression

- Hittills på kursen:
  - skatta **regressionslinjen**: $\hat{y} = b_0 + b_1 x$
  - **prediktion** för ny observation: $\hat{y}_i = b_0 + b_1 x_i$
- Med sannolikhetslära kan vi göra mycket mer:
  - om $b_1 \neq 0$, finns det **verkligen korrelation** mellan x och y? **Stickprov** vs **Population**.
  - **osäkerhetsintervall för b_1** som troligen täcker sanna värdet.
  - **osäkerhetsintervall för prediktionen** $\hat{y}_i$.



Försök, utfall och utfallsrum

- Vi utför ett **försök** (eng. trial): singlar ett mynt.
- Observerar ett **utfall** (eng. outcome): Krona.
- **Utfallsrummet** är **alla möjliga utfall** som kan inträffa.
- Singla slant $S = \{\text{Krona}, \text{Klave}\}$.
- Kasta en tärning:



- Kasta två tärningar, summera antal prickar.

	⚀	⚁	⚂	⚃	⚄	⚅
⚀	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

Händelse - exakt sju prickar med två tärningar

- En **händelse** är en **mängd av utfall**.
- Händelsen A = få exakt 7 prickar med två tärningar.

$$A = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$$

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

Händelse - samma antal prickar på båda tärningarna

- Händelsen A = {få samma antal prickar på båda tärningarna}

$$A = \{(1,1),(2,2),(3,3),(4,4),(5,5),(6,6)\}$$

	⚀	⚁	⚂	⚃	⚄	⚅
⚀	**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**

Tre sannolikhetsbegrepp

- Vad är **sannolikheten** att få en 6:a med en tärning?
  - Utfallsrum: $S = \{1, 2, 3, 4, 5, 6\}$.
  - Händelse: $A = \{6\}$.
  - Sannolikhet: $P(A)$. Måste uppfylla: $0 \leq P(A) \leq 1$.

1. **Lika sannolika utfall** (logisk sannolikhet).
   En tärnings fysiska egenskaper $\rightarrow$ alla sidor är lika sannolika.

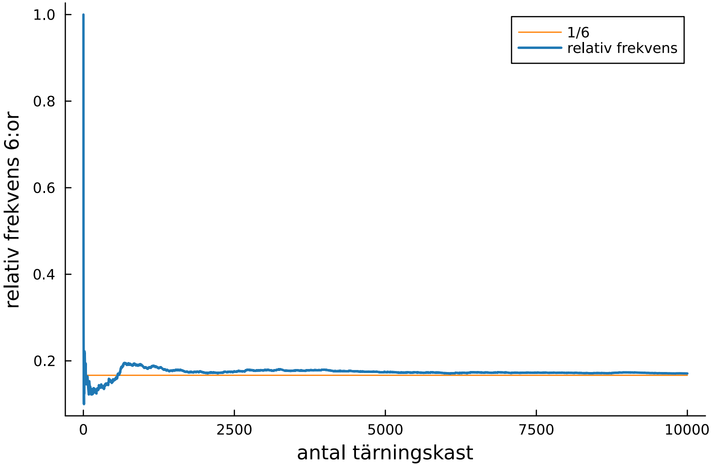
$$P(A) = \frac{\text{antal utfall i } A}{\text{totalt antal möjliga utfall}} = 1/6 \approx 0.1667$$

2. **Empirisk sannolikhet**: andelen 6:or om jag kastar tärningen ett "oändligt" antal gånger.

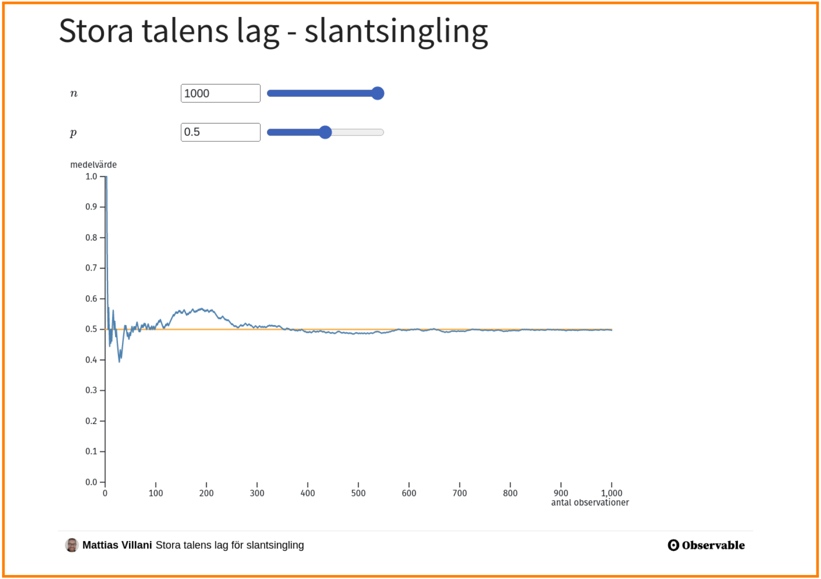
$$P(A) = \frac{\text{antal gånger som A inträffar}}{\text{totalt antal försök}}$$

3. **Subjektiva sannolikheter**. **Min** tidigare erfarenhet av tärningskast och **min** uppfattning om en tärnings symmetri säger mig att **min** sannolikhet att få en 6:a är $1/6 \approx 0.1667$.

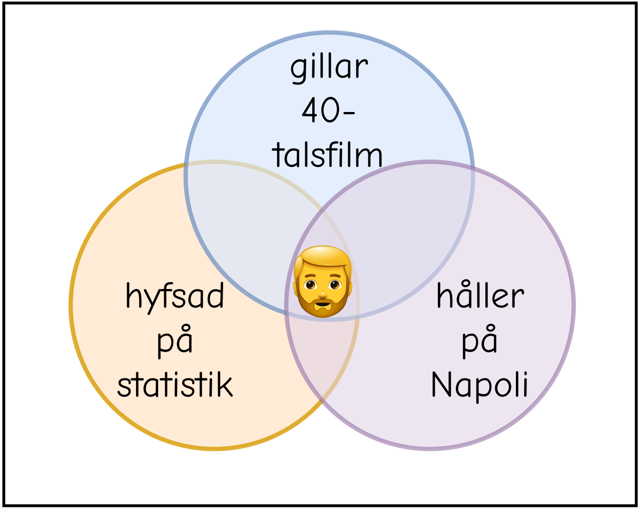
Stora talens lag - få 6:a med tärning



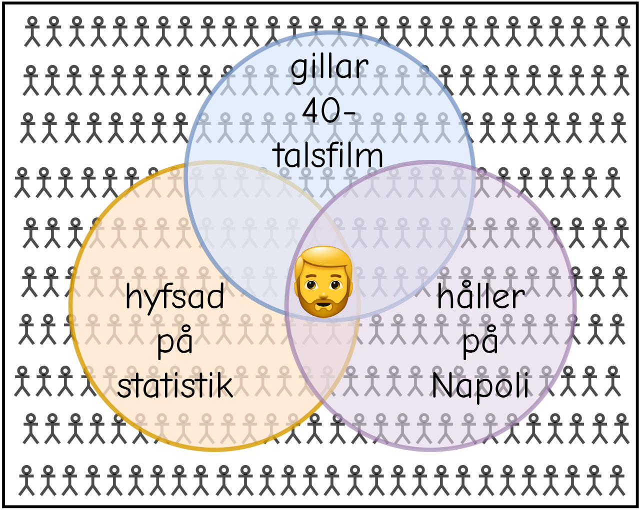
Stora talens lag - slantsingling



Venndiagram

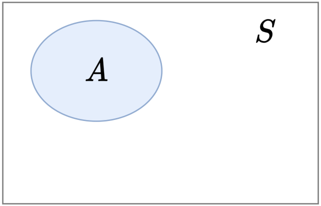


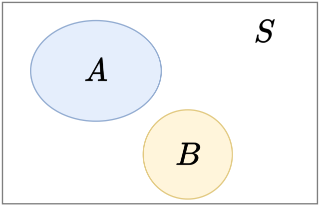
Venndiagram



Händelse - Venndiagram

- Praktiskt att visualisera händelser i ett **Venndiagram**.
- **Utfallsrummet** (allt som kan inträffa) visas med **rektangel**.
- **Händelser** ritas som **cirklar**, **ellipser** eller **rektanglar**.





Venndiagram - summa sju prickar och samma

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

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

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

Venndiagram - summa tio prickar och samma

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

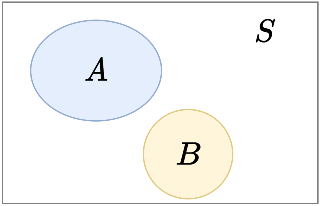
2	3	4	5	6	7
3	4	5	6	7	8
4	5	6	7	8	9
5	6	7	8	9	10
6	7	8	9	10	11
7	8	9	10	11	12

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

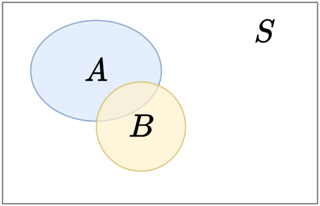
2	3	4	5	6	7
3	4	5	6	7	8
4	5	6	7	8	9
5	6	7	8	9	10
6	7	8	9	10	11
7	8	9	10	11	12

Disjunkta händelser

Disjunkta händelser
inga gemensamma element



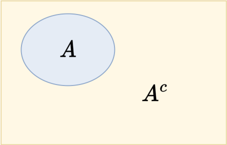
Överlappande händelser
med gemensamma element



US Disjunkt = **Disjoint**

Komplementshändelsen

- **Komplementet** till A inträffar när **A inte inträffar**.
- Vi skriver $\mathbf{A}^c$ där c står för engelskans **C**omplement.

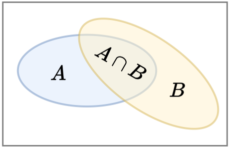


- Tärningar
  - $\mathbf{A} = \{$udda antal prickar på tärning$\} = \{1,3,5\}$.
  - $\mathbf{A}^c = \{$jämnt antal prickar på tärning$\} = \{2,4,6\}$.

- Inflation
  - $\mathbf{A} = \{$inflationen nästa månad $\leq 2\}$.
  - $\mathbf{A}^c = \{$inflationen nästa månad $> 2\}$.

Snitthändelsen

- **Snitthändelsen** är händelsen där **både A och B** inträffar.
- Vi skriver A och B eller $\mathbf{A} \cap \mathbf{B}$. US Snitt = **Intersection**

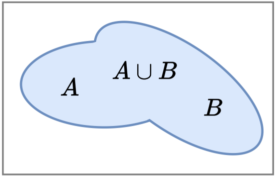


- Två tärningar:
  - $\mathbf{A} = \{$samma prickar på båda tärningar$\}$
  - $\mathbf{B} = \{$totalt 10 prickar$\}$
  - $\mathbf{A} \cap \mathbf{B} = \{$5:a på båda tärningar$\}$
- Lågkonjunktur 😰
  - $\mathbf{A} = \{$BNP-tillväxt kvartal 1 $< 0\}$
  - $\mathbf{B} = \{$BNP-tillväxt kvartal 2 $< 0\}$
  - $\mathbf{A} \cap \mathbf{B} = \{$Negativ BNP-tillväxt två kvartal i rad$\}$
- Disjunkta händelsers snitt är den **tomma mängden** $\emptyset$

$$\mathbf{A} \text{ och } \mathbf{B} \text{ disjunkta} \quad \Longleftrightarrow \quad \mathbf{A} \cap \mathbf{B} = \emptyset$$

Unionhändelsen

- **Unionhändelsen** är händelsen där **A och/eller B** inträffar.
- **Minst en** av händelserna inträffar.



- Universitetstudier
  - $\mathbf{A}$ = {Kommer in på kurs på betyg}
  - $\mathbf{B}$ = {Kommer in på kurs på högskoleprov}
  - $\mathbf{A} \cup \mathbf{B}$ = {Kommer in på kurs}

Formell sannolikhet

Sannolikheten $P(A)$ för händelse A på utfallsrummet S

1. $0 \leq P(A) \leq 1$
2. $P(S) = 1$
3. $P(A^c) = 1 - P(A)$
4. $P(A \cup B) = P(A) + P(B)$ om A och B är **disjunkta**
5. $P(A \cap B) = P(A) \cdot P(B)$ om A och B är **oberoende**

1. En sannolikhet är ett tal mellan 0 och 1.
2. Sannolikheten för en **säker händelse** är 1.
3. Sannolikheten att en händelse **inte** inträffar är 1 minus sannolikheten för händelsen.
4. Sannolikheten att **åtminstone en** av två händelser som **inte kan inträffa samtidigt** är summan av händelsernas sannolikheter.
5. Sannolikheten att två oberoende händelser **båda** inträffar är produkten av händelsernas sannolikheter.

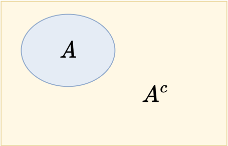
Komplementsregeln

- A och A^c är **disjunkta**. Kan inte inträffa samtidigt.
- Någon av A eller A^c *måste* inträffa.

$$P(A) + P(A^c) = 1$$

Komplementsregeln

$$P(A^c) = 1 - P(A)$$



- $\mathbf{A}$ = {ingen bugg i koden}.
- $\mathbf{A}^c$ = {**åtminstone en** bugg i koden} = {1 bugg, 2 buggar, ...}
- P({minst en bugg i koden}) $= 1 - P$({ingen bugg i koden}).

Den allmänna additionsregeln

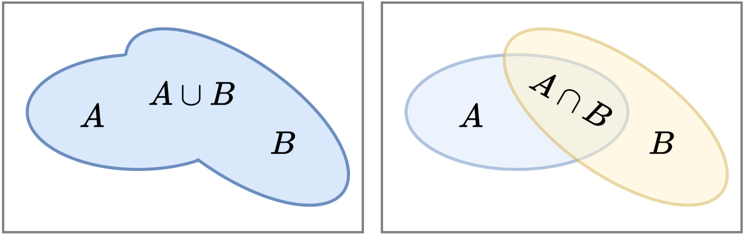
- **Additionsregeln**: Om A och B är **disjunkta**:

$$P(A \cup B) = P(A) + P(B)$$

Allmänna additionsregeln (även överlappande händelser)

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- Måste dra bort snittet $A \cap B$ för det räknas två ggr.



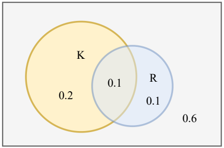
Röst på socialdemokraterna 🌹 additionsregeln

- $\mathbf{R}$ = person röstar 🌹 i riksdagsvalet. $P(\mathbf{R}) = 0.2$
- $\mathbf{K}$ = person röstar 🌹 i kommunalvalet. $P(\mathbf{K}) = 0.3$
- Personen röstar på 🌹 in båda valen: $P(\mathbf{R} \cap \mathbf{K}) = 0.1$
- Röstar 🌹 i åtminstone ett av valen? Additionsregeln:

$$P(\mathbf{R} \cup \mathbf{K}) = 0.2 + 0.3 - 0.1 = 0.4$$

- Röstar inte på 🌹 i något av valen?

$$P(\mathbf{R}^c \cap \mathbf{K}^c) = P((\mathbf{R} \cup \mathbf{K})^c) = 1 - P(\mathbf{R} \cup \mathbf{K}) = 1 - 0.4 = 0.6$$



Multiplikationssregeln för oberoende händelser

- Händelserna A och B är **oberoende** om vetskapen att B har inträffat **inte påverkar** sannolikheten för A. Och vice versa.
- Test: kommer sannolikheten för A förändras om man får veta att B har inträffat? Om inte, så är A och B oberoende.

Multiplikationsregeln. För **oberoende** händelser A och B

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

- Hur beräknar man sannolikheten för snittet $A \cap B$ för händelser som **inte** är oberoende? Stay tuned, kommer i F12. 🫣

Multiplikationsregeln för oberoende händelser

- Vad är sannolikheten att få 2 st krona i rad vid slantsingling?

$$0.5 \cdot 0.5 = 0.5^2 = 0.25$$

- Vad är sannolikheten att få 5 st krona i rad vid slantsingling?

$$0.5 \cdot 0.5 \cdot 0.5 \cdot 0.5 \cdot 0.5 = 0.5^5 = 0.03125$$

- 1% risk att streaming laggar under en kväll. Oberoende kvällar.

$$P(\text{ingen lagg hela veckan}) = (1 - 0.01)^7 = 0.99^7 \approx 0.932.$$

- Sannolikheten att dra två klöver ♣ ur en blandad kortlek?

$$P(\text{1:a kortet klöver}) = \frac{13}{52} = \frac{1}{4}$$

$$P(\text{2:a kortet klöver } \textbf{givet} \text{ 1:a kortet klöver}) = \frac{12}{51}$$

$$P(\text{2:a kortet klöver } \textbf{givet} \text{ 1:a kortet } \textbf{inte} \text{ klöver}) = \frac{13}{51}$$

- $A = \{\clubsuit \text{ på 1:a}\}$ och $B = \{\clubsuit \text{ på 2:a}\}$ är **inte** oberoende.

Röst på socialdemokraterna 🌹 multiplikationsregeln

- $\mathbf{R}$ = person röstar 🌹 i riksdagsvalet. $P(\mathbf{R}) = 0.2$
- $\mathbf{K}$ = person röstar 🌹 i kommunalvalet. $P(\mathbf{K}) = 0.3$
- Personen röstar på 🌹 in båda valen: $P(\mathbf{R} \cap \mathbf{K}) = 0.1$
- Är händelserna $\mathbf{R}$ och $\mathbf{K}$ **oberoende**? Vi måste undersöka om

$$P(\mathbf{R} \cap \mathbf{K}) \stackrel{?}{=} P(\mathbf{R}) \cdot P(\mathbf{K})$$

- Händelserna är **inte** oberoende:

$$P(\mathbf{R}) \cdot P(\mathbf{K}) = 0.2 \cdot 0.3 \neq 0.1 = P(\mathbf{R} \cap \mathbf{K})$$

- Att hen röstat på 🌹 i kommunalvalet ger information om vad hen röstat på i riksdagsvalet.
- **Betingad sannolikhet** för $\mathbf{R}$ **givet** $\mathbf{K}$ är sann: 0.333 (se F13).
- $P(\mathbf{R})$ ökar från 0.2 till 0.333 när vi vet att $\mathbf{K}$ är sann.

Kombinatorik

- Dataset 1: **med ordning**: Krona, Klave, Klave, Krona, Krona.
- Dataset 2: **utan ordning**: 3 st Krona och 2 st Klave.
- Är det lika sannolikt att observera Dataset 1 som Dataset 2?

- **Kombinatorik**: räknar antal sätt/kombinationer.

- **Fakultetet** (eng. factorial). Utläses som n-fakultet.

$$n! = n(n-1)(n-2)\cdots 2\cdot 1$$

Hur många sätt att välja k element bland n element?

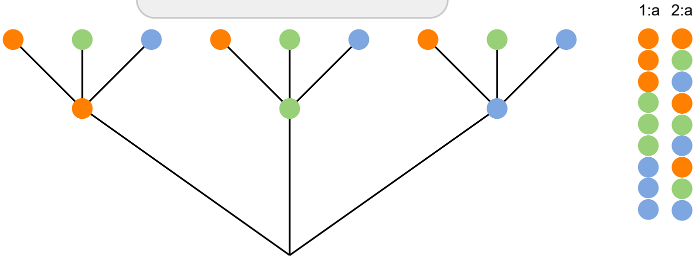
	med återläggning	utan återläggning
med ordning	n^k	${}_nP_k = \frac{n!}{(n-k)!}$
utan ordning	ej på kurs	${}_nC_k = \frac{n!}{(n-k)!k!}$

Med återläggning, med hänsyn till ordning

välj $k = 2$ bland $n = 3$

med återläggning - med hänsyn till ordning

$$n^k = 3^2 = 9$$

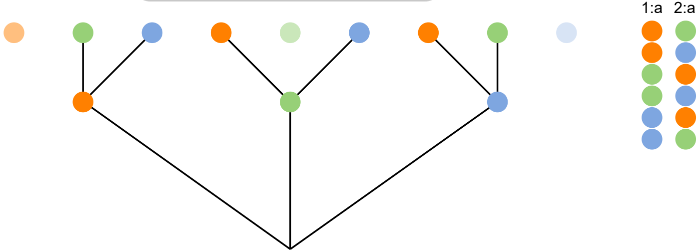


Utan återläggning, med hänsyn till ordning

välj $k = 2$ bland $n = 3$

utan återläggning - med hänsyn till ordning

$${}_nP_k = \frac{n!}{(n-k)!} = \frac{3!}{1!} = \frac{3 \cdot 2 \cdot 1}{1} = 6$$

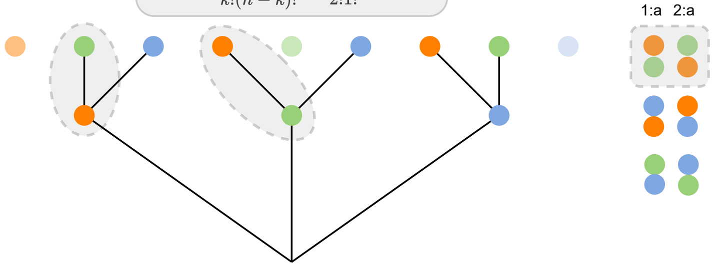


Utan återläggning, utan hänsyn till ordning

välj $k = 2$ bland $n = 3$

utan återläggning - utan hänsyn till ordning

$$_nC_k = \frac{n!}{k!(n-k)!} = \frac{3!}{2!1!} = 3 \text{ sätt}$$



Credits

Dessa slides skapades för kursen statistik och dataanalys 1 av Mattias Villani HT 2023, och har modifierats av Oscar Oelrich VT 2024, och Oskar Gustafsson för VT 2025.