

Statistik och dataanalys I

F11 & F12: Osäkerhet och Sannolikhet

Valentin Zulj

Statistiska institutionen



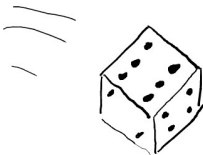
Stockholm
University

- Vi har introducerat en rad olika statistiska begrepp och metoder, t.ex.
 - Centralmått och spridning
 - Korrelationer, linjär regression, och prediktioner
 - Populationer, stickprov, parametrar, och bias
 - Experiment, observationsstudier, och kausalitet
- Vi har diskuterat dessa relativt övergripande, mest skarapat på ytan
- **Men:** för att bli självgående och kreativa statistiker behöver vi mer djuplodad förståelse för statistiska metoder
- Vi ska nu börja arbeta mot precis detta mål!

- Från tidigare: statistik handlar om att dra slutsatser om en population, utifrån information som finns i ett stickprov
- Stickprov dras ofta med hjälp av någon sorts **slumpmekanism**, i någon bemärkelse går vägen från population till stickprov via slumpen
- Vi vill gå åt andra hållet, från stickprovet till populationen, och måste därför ha förståelse för slumpmekanismen som binder samman de två ändpunkterna
- Det visar sig att vi kan formalisera slumpen med hjälp av matematik, och vi kommer nu börja arbeta med att skapa förståelse för denna formalisering
- Vi förknippar ofta slumpen med frågor om **händelser** och **sannolikheter**, t.ex.
 - Sannolikheten att slå krona när vi singlar slant
 - Sannolikheten att Tottenham åker ur PL denna säsong (förhoppningsvis 100%)
- Därför ska vi nu börja ställa upp ett matematiskt ramverk där vi kan definiera just händelser och sannolikheter

Slumpmässighet, slumphändelser och sannolikheter

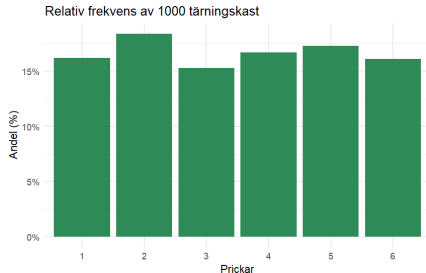
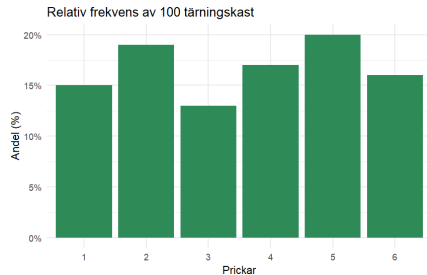
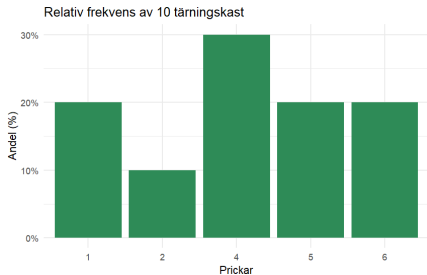
- Går det att förutsäga utfallet av ett tärningskast?



- Går det att förutsäga utfallet av en slantsingling?

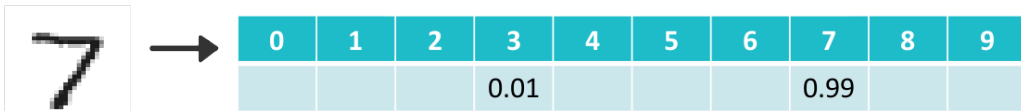


- **Svårt**, eftersom båda är slumpmässiga händelser!
- **Men:** i långa loppet kan vi förutsäga antalet kast som landar på varje utfall!



- Sannolikheter kvantifierar vår osäkerhet kring slumpmässiga fenomen
- På förhand **kan vi inte** säga exakt hur ett tärningskast kommer falla ut
- Men: **vi kan** räkna upp alla möjliga utfall, och beskriva hur troliga de är
- Studiet av sannolikheter kallas **sannolikhetslära**, och kan t.ex. handla om
 - Sannolikheten för en kärnkraftsolycka eller stor översvämning
 - Sannolikheten för att två personer har en matchande DNA-profil
- **Exempel på relevant fråga:**
 - Vi vet att en specifik typ av sjukdom finns hos 0.1% av befolkningen
 - Det finns ett test för sjukdomen i fråga, som har 99% pricksäkerhet
 - Om testet är positivt, vad är sannolikheten att personen faktiskt är sjuk?
 - Vi är osäkra då vi inte vet (i) om personen är sjuk, och (ii) om testet är korrekt

- Vi ska som sagt börja intressera oss för slumpmässiga fenomen och vår osäkerhet kring dem
- Det är relevant för oss då
 - I princip all statistisk metod grundar sig i någon typ av sannolikhetsmodell
 - Det hjälper oss att kvantifiera osäkerheten i en prediktion eller skattning – t.ex. felmarginaler i opinionsundersökningar, osäkerhetsintervall för ränteprognoser, etc
 - Det kan hjälpa oss att fatta (optimala) beslut under osäkerhet

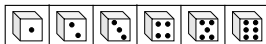


0	1	2	3	4	5	6	7	8	9
			0.01				0.99		

- Vi har använt regressionslinjen, $\hat{y} = b_0 + b_1x$, för att göra prediktioner, $\hat{y}_i = b_0 + b_1x_i$
- Om x_i är medeltemperaturen en given dag, kan vi då vara säkra på $\hat{y}_i$ (skattat antal försålda glassar) kommer vara helt korrekt? Troligtvis inte!
 - **Men:** Med hjälp av sannolikhetslära kan vi ta fram osäkerhetsintervall för $\hat{y}_i$
- Om $b_1 \neq 0$, betyder det då nödvändigtvis att det finns ett samband mellan x och y i populationen?
 - Med sannolikhetslära kan vi få en mer formell behandling av signifikans, och tänka mer på hur vi generaliserar från stickprov till population
 - Vi kan också ta fram osäkerhetsintervall för populationsvarianten av b_1 , som troligtvis täcker det sanna värdet



- När vi t.ex. singlar slant eller kastar tärning så utför vi ett **försök** (*eng: trial*)
- Vi observerar då ett **utfall** (*eng: outcome*), t.ex. krona i slantsingling
- Vi samlar **alla möjliga** utfall i det som kallas för **utfallsrum**, och ofta betecknas med S
- För en slantsingling har vi t.ex. $S = \{\text{Krona, Klave}\}$
- Om vi kastar en tärning kan vi illustrera utfallsrummet som nedan



- Om vi kastar två tärningar, och räknar totalt antal prickar, får vi på motsvarande sätt

						
	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

- En **händelse** består av en mängd av utfall – vi kan t.ex. betrakta händelsen

$$A = \{\text{summan av prickar bli 7 när vi kastar två tärningar}\}$$

- Vilka kombinationer av utfall gör att händelsen A inträffar?

						
	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

- Vi skriver observationspar som (**prickar på tärning 1**, **prickar tärning på 2**), och ser då att händelsen A definieras som

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

- En annan händelse: B = samma antal prickar på två slagna tärningar
- Vilka kombinationer av utfall leder till B ?

						
	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

- Vi har

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

- Ni kan på egen hand testa att reda ut vilka observationspar som hör till händelserna nedan, alla när vi kastar två tärningar
 - $C = \{\text{summan av prickar blir max 4}\}$
 - $D = \{\text{summan av prickar blir minst 10}\}$
 - $G = \{\text{summan av prickar blir mindre än 4 eller större än 8}\}$
 - $F = \{\text{summan av brickar ligger mellan 2 och 7 och är ett jämnt tal}\}$
- Använd bilden på slide 11 om ni behöver hjälp att orientera er bland utfall!

- Vi kan inte förutsäga utfallet av en enskild slump händelse, t.ex. ett tärningskast, på grund av just slumpmässigheten
- **Men:** om vi kan kvantifiera hur troliga de olika utfallen är, så får vi åtminstone en bild av vad vi rimligtvis kan förvänta oss
- Det är av denna anledning vi introducerar **sannolikheter**
- Vi betraktar ett enkelt tärningskast – vad är sannolikheten att slå en sexa?
- Vi definierar utfallsrum och anger händelsen vi är intresserade av – precis som tidigare
 - $S = \{1, 2, 3, 4, 5, 6\}$
 - $A = \{6\}$
- Vi är då intresserade av **sannolikheten för händelsen A** , vilken betecknas med $P(A)$
- Vi har att $0 \leq P(A) \leq 1$ – vad betyder $P(A) = 0$ respektive $P(A) = 1$?

- Vi kan tolka $P(A)$ som en **logisk sannolikhet**, dvs

$$P(A) = \frac{\text{antal utfall i } A}{\text{totalt antal möjliga utfall}}$$

- Det finns sex möjliga utfall vid ett tärningskast, och händelsen $A = \{6\}$ innehåller ett av dem – enligt definitionen ovan ger detta $P(A) = 1/6 \approx 0.1667$
- Vi kan även tolka $P(A)$ som en **empirisk sannolikhet**, som bygger på att vi kan kasta tärningen ett “oändligt” antal gånger

$$P(A) = \frac{\text{antal gånger som } A \text{ inträffar}}{\text{totalt antal försök}} \quad (\text{när antalet försök} \rightarrow \infty)$$

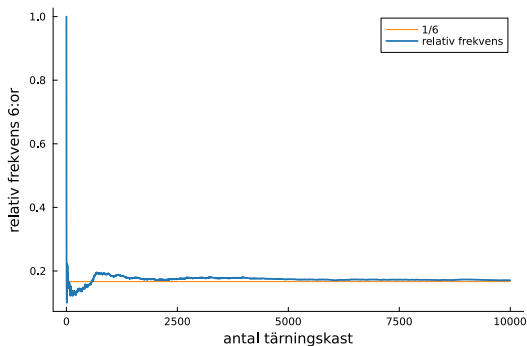
- Till slut kan vi också kalla på vår egna erfarenhet och intuition, och därifrån tolka $P(A)$ som en **subjektiv** sannolikhet
- **Exempel:** Min erfarenhet av och min uppfattning om en tärnings symmetri säger mig att min sannolikhet att slå en sexa är $1/6 \approx 0.1667$

- Vi kan titta lite närmare på empiriska sannolikheter, och använda R för att simulera ett stort antal tärningskast
- Vi kan sedan se hur kvoten nedan utvecklas

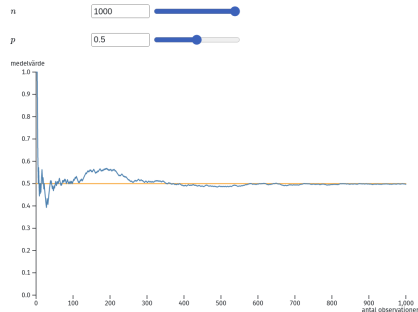
$$P(A) = \frac{\text{antal gånger som } A \text{ inträffar}}{\text{totalt antal försök}} \quad (\text{när antalet försök} \rightarrow \infty)$$

- Vi antar att
 - $P(A)$ är **samma** för varje tärningskast (dvs varje försök)
 - Alla tärningskast är **oberoende** av varandra – detta betyder t.ex. att utfallet i försök 1 inte påverkar sannolikheterna för olika utfall i försök 2, osv
- Då kan vi visa (ingår ej här) att andelen sexor kommer stabiliseras i takt med att vi upprepar försöket ett stort antal gånger
- Kvoten kommer alltså gå mot ett specifikt värde, som vi definierar som den **empiriska sannolikheten** $P(A)$
- Detta är ett exempel på ett resultat som kallas **stora talens lag**

Stora talens lag - slantsingling



(a) Tärningskast



Mattias Villani Stora talens lag för slantsingling

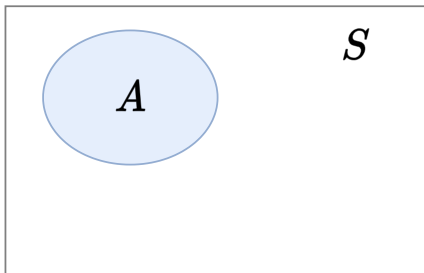
Observable

(a) Slantsingling

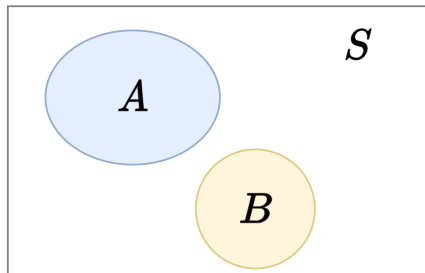
- Bilden till höger är tagen från från [det interaktiva verktyget](#) som finns på [kurshemsidan](#)

Venndiagram

- Det kan vara fördelaktigt att ha ett ramverk för att visualisera olika händelser – dels för inläarning, dels för att underlätta beräkning av mer komplicerade sannolikheter
- Vanligtvis gör vi detta i **Venndiagram**, där
 - **Utfallsrummet** (allt som kan inträffa) visas som en **rektangel**
 - **Händelser** ritas som **cirklar, ellipser eller rektanglar**.



(a) Diagram med händelsen A



(a) Diagram med händelserna A och B

- Kan vi se diagrammen nedan som Venndiagram?



	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

- **Ja!** Hela utfallsrummet representeras av en rektangel (kvadrat), och händelserna ritas in med hjälp av (färgkodade) rektanglar
- Vi kan här notera att det inte finns något överlapp mellan händelserna

- Vi byter nu ut händelsen “summera till sju” mot “summera till tio” och får



1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12



1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12



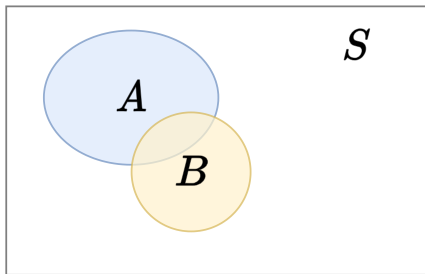
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

- Vi ser att det finns överlapp mellan händelserna – vilket för oss in på en ny distinktion

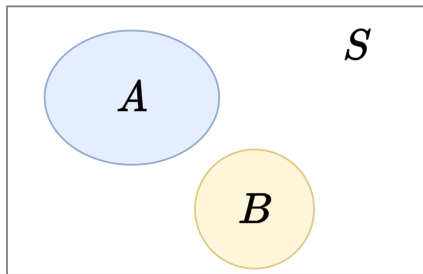


1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

- Vi säger att händelser är
 - **överlappande** om de innehåller gemensamma utfall (t.ex. A = klara tentan och B = klara inluppen)
 - **disjunkta** om de **inte** innehåller gemensamma utfall (t.ex. A = klara tentan och B = kugga tentan)
- Vi har även sett exempel på båda dessa typer ovan!

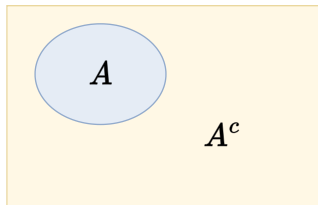


Överlappande händelser



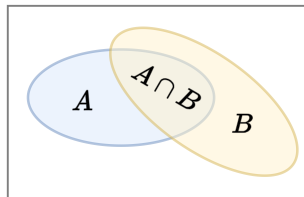
Disjunkta händelser

- Vi utgår återigen ifrån händelsen A , men intresserar oss nu även för händelsen “inte A ”
- Vi kallar “inte A ” för **komplement** till A , och betecknar det med A^c (c för *complement*)



- **Exempel:** Tärningskast
 - $A = \{\text{udda antal prickar}\} = \{1, 3, 5\}$
 - $A^c = \{\text{jämnt antal prickar}\} = \{2, 4, 6\}$
- **Exempel:** Inflation
 - $A = \{\text{inflationen nästa månad} \leq 2\}$
 - $A^c = \{\text{inflationen nästa månad} > 2\}$
- **Fråga:** Är A och A^c överlappande eller disjunkta händelser?

- Snitthändelsen inträffar då **både** A **och** B inträffar – skrivs som $A \cap B$ (alt. A och B)



- **Exempel:** Två tärningskast

- $A = \{\text{samma antal prickar på båda tärningar}\}$
- $B = \{\text{summan av prickarna är tio}\}$

$$A \cap B = \{5:a \text{ på båda tärningarna}\}$$

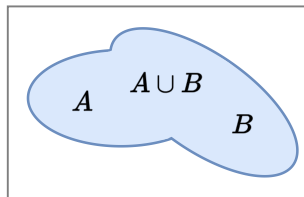
- **Exempel:** Lågkonjunktur

- $A = \{\text{Tillväxt kvartal } 1 < 0\}$
- $B = \{\text{Tillväxt kvartal } 2 < 0\}$

$$A \cap B = \{\text{Negativ tillväxt två kvartal i rad}\}$$

- Om A och B är disjunkta får vi $A \cap B = \emptyset$ (tomma mängden)

- Unionhändelsen inträffar då A **och/eller** B inträffar, och betecknas $A \cup B$
- I bilden nedan ser vi att $A \cup B$ innebär att **minst en av A och B inträffar**



- **Exempel** Universitetsstudier
 - $A = \{\text{Kommer in på en kurs med betyg}\}$
 - $B = \{\text{Kommer in på m.h.a. högscoleprov}\}$

$$A \cup B = \{\text{Kommer in på kursen}\}$$

Venndiagram och sannolikheter

- Vi ska strax föra mer matematiska resonemang kring sannolikheter
- Vi utgår ifrån en given händelse, och uttrycker först några resultat i ord för jämförelse
- Sannolikheten att händelsen inträffar ligger **alltid** mellan 0 och 1 (inklusive 0 och 1)
- Om sannolikheten är 1 är det helt säkert att händelsen **inträffar**, och om sannolikheten är 0 är det helt säkert att händelsen **inte inträffar**
- Sannolikheten att händelsen **inte** inträffar är 1 minus sannolikheten för att den inträffar
- Om två händelser **inte kan inträffa** samtidigt, är sannolikheten att **åtminstone en** inträffar summan av båda händelsernas sannolikheter
- Sannolikheten att två oberoende händelser **båda** inträffar är lika med produkten av händelsernas sannolikheter
- Meningarna blir långa, svårtolkade, och något krystade i ord
- **Men:** vi får mer övergripliga uttryck med hjälp av matematikens språk

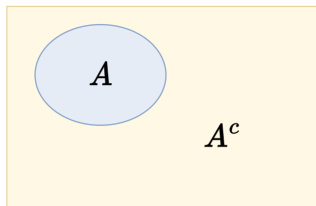
- Vi tar nu motsvarande uttryck/resultat och uttrycker dem i matematik

- För händelserna A och B , i utfallsrummet S , gäller att

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

- Efter lite övning brukar det vara lättare att uttrycka oss på detta sätt – det är dessutom mer exakt och lämnar mindre utrymme för missförstånd
- Vi är nu redo att arbeta fram dessa resultat med hjälp av Venndiagram

- Vi vet att A och A^c är disjunkta, och kan representeras som nedan



- A och A^c täcker tillsammans hela utfallsrummet, vilket ger $P(A) + P(A^c) = P(S) = 1$

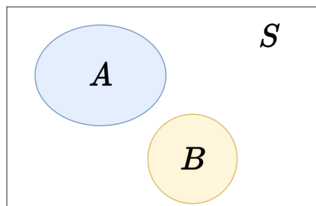
Komplementregeln:

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

Exempel:

$$P(\text{Duplantis slår } \mathbf{inte} \text{ världsrekord}) = 1 - P(\text{Duplantis slår världsrekord})$$

- Vi kan illustrera två **disjunkta** händelser som i diagrammet nedan

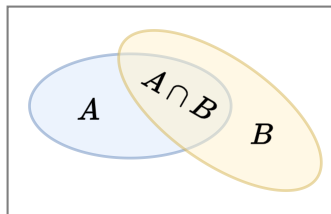
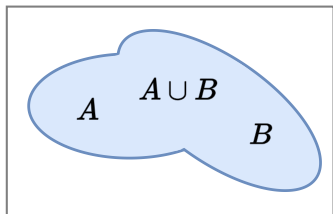


- Vi söker nu sannolikheten att **minst** en av A och B inträffar

Additionsregeln: Om A och B är disjunkta har vi att

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

- Samma utgångspunkt som föregående slide, men nu är A och B **inte** disjunkta



Allmänna additionsregeln: Om A och B är två händelser har vi att

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

Vi måste subtrahera $P(A \cap B)$ för att inte räkna den två gånger!

- **Fråga:** Vad händer om A och B är disjunkta?

- Vi antar att
 - $R = \{\text{En person röstar på S i riksdagsvalet}\}$, med sannolikheten $P(R) = 0.2$.
 - $K = \{\text{En person röstar på S i kommunalvalet}\}$, med sannolikheten $P(K) = 0.3$
 - Sannolikheten att personen röstar på S i båda valen är $P(R \cap K) = 0.1$

Vad är sannolikheten att personen röstar på S i **åtminstone** ett av valen?

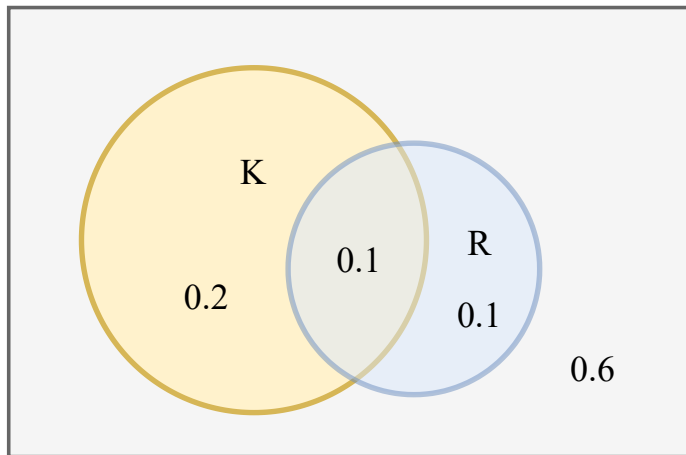
*Vi använder **allmänna additionssatsen** och får*

$$P(R \cup K) = P(R) + P(K) - P(R \cap K) = 0.2 + 0.3 - 0.1 = 0.4$$

Vad är sannolikheten att personen röstar **inte** röstar på S i **något** val?

Detta exempel är något svårare, men försök koppla det till bilden på nästa slide

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



- Händelserna A och B är **oberoende** om vetskapen att B har inträffat **inte påverkar** sannolikheten för A , och vice versa
- Oberoende betyder alltså att händelserna inte påverkar varandra i sannolikhetsväg
- Om A och B är oberoende – vad är då sannolikheten att både A och B inträffar?

Multiplikationsregel: Om A och B är oberoende händelser så har vi att

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

- I F12 kommer vi beräkna motsvarande sannolikhet för beroende händelser A och B

- Vi singlar (samma) slant ett antal gånger i rad, och säger att A = krona första gången, B = krona andra gången, C = krona tredje gången, och så vidare
- Är det rimligt att anta att händelserna är oberoende?

Vad är sannolikheten att få krona två gånger i rad?

A och B är oberoende, och multiplikationsregeln ger

$$P(A \cap B) = P(A) \cdot P(B) = 0.5 \cdot 0.5 = 0.25$$

Vad är sannolikheten att få krona fem gånger i rad?

Slantsinglingarna är oberoende, och multiplikationsregeln ger

$$P(A \cap B \cap \dots \cap E) = P(A) \cdot P(B) \cdot \dots \cdot P(E) = 0.5 \cdot 0.5 \cdot 0.5 \cdot 0.5 \cdot 0.5 = 0.5^5 = 0.03125$$

- Vi drar slumpmässigt två kort ur en vanlig kortlek med 52 kort, och definierar händelserna $A = \{\text{Första kortet är klöver}\}$ och $B = \{\text{Andra kortet är klöver}\}$
- Vi vet att $P(A) = 13/52$, då det finns 13 klöverkort från början
- Om vi drog klöver i första gången har vi att 12 av 51 kvarvarande kort är klöver, alltså:

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

- Om vi **inte** drog klöver första gången har vi att 13 av 51 kvarvarande kort är klöver, dvs:

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

- Vi ser att A **påverkar sannolikheten** för B , vilket betyder att händelserna **inte** är oberoende – vi kan **inte** använda multiplikationsregeln!
- **Fråga:** Vad händer om vi lägger tillbaka det första kortet innan andra dragningen?

- Vi återgår till exemplet där
 - $R = \{\text{En person röstar på S i riksdagsvalet}\}$, med sannolikheten $P(R) = 0.2$.
 - $K = \{\text{En person röstar på S i kommunalvalet}\}$, med sannolikheten $P(K) = 0.3$
 - Sannolikheten att personen röstar på S i båda valen är $P(R \cap K) = 0.1$
- Är R och K oberoende?
- Multiplikationsregeln säger att $P(R \cap K) = P(R) \cdot P(K)$
- Vi har $P(R) \cdot P(K) = 0.2 \cdot 0.3 = 0.06$, och ser så att

$$P(R) \cdot P(K) = 0.2 \cdot 0.3 = 0.06 \neq 0.1 = P(R \cap K)$$

- Vad betyder detta?
- Om händelserna är oberoende **måste** $P(R \cap K) = P(R) \cdot P(K)$ vara sant
- Eftersom likheten inte håller betyder det att händelserna **inte** är oberoende

Kombinatorik

- Vi har pratat om **logisk sannolikhet**, dvs

$$P(A) = \frac{\text{antal utfall i } A}{\text{totalt antal möjliga utfall}}$$

- Denna tolkning av sannolikhet är vanlig i praktiken, och om vi ska räkna med den krävs att vi kan ta fram totalt antal möjliga utfall
- Kombinatorik hjälper oss att göra just detta!
- **Exempel på möjliga frågeställningar**
 - Vi väljer helt slumpmässigt fem kort från en kortlek med 52 kort (4 färger med 13 valörer), vad är sannolikheten att alla har samma färg? **Vi behöver antalet sätt att välja fem kort!**
 - Antag att vi ska välja 3 personer här inne som automatiskt får 10 bonuspoäng till tentan – hur många olika trios kan vi välja? **Vi behöver antalet sätt att välja tre personer!**

- Antag att n personer ska lottas in i en ny bostadskö – hur många olika köer kan bildas?
 - Förstaplatsen kan delas ut på n olika sätt (alla n personer kan få den)
 - Andraplatsen kan delas ut på $n - 1$ olika sätt (alla utom personen på första plats kan få)
 - Tredjeplatsen kan delas ut på $n - 2$ sätt, och så vidare

- **Multiplikationsprincipen** säger att antalet möjliga köer är

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

- $n!$ uttalas *n-fakultet*, och sägs ange antalet möjliga **permutationer** av n objekt
- Antalet permutationer beskriver antalet sätt på vilka vi kan ordna ett antal objekt

Nio löpare springer ett lopp – hur många olika prispallar kan loppet resultera i?

Multiplikationsprincipen säger att pallplatserna kan fördelas på $9 \cdot 8 \cdot 7 = 504$ olika sätt.

- Här stannade vi efter 7, och fortsatte inte multiplicera hela vägen ner till 1 som tidigare
- Anledningen är att vi här räknar antalet **permutationer** av 3 objekt valda bland 9 möjliga – vi väljer alltså inte alla 9, utan stannar när vi valt 3

Antalet **permutationer** av k objekt valda bland n möjliga (k måste vara $\leq n$) är

$${}_nP_k = n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot (n-k+1)$$

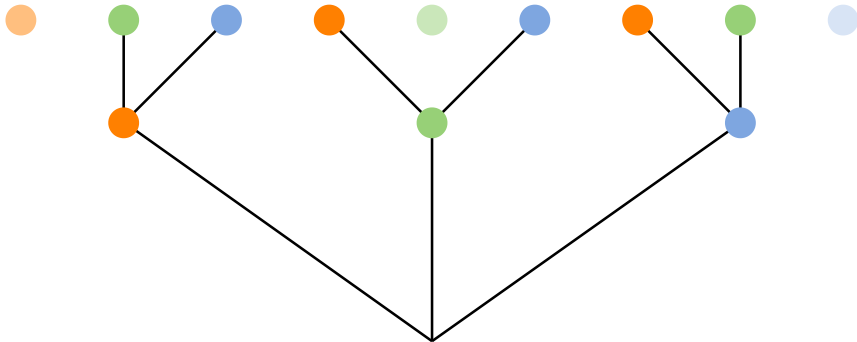
Med lite matematik får vi att

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

välj $k = 2$ bland $n = 3$

utan återläggning - med hänsyn till ordning

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



1:a 2:a



- Hittills har vi dragit personer **utan återläggning** (vi kan ej ha samma etta och trea i kön)
- **Men:** ordningen har alltid spelat roll – det är **inte** samma sak att vara etta och 25:a i kön
- När vi drar tre personer som ska få bonuspoäng så **spelar ordningen ingen roll** – så länge du får 10 extrapoäng spelar det inte så stor om du blev dragen först eller sist
- Vi drar då tre personer **utan återläggning**, men nu **utan hänsyn till ordningen**

- En förälder skickar in sin femåring i en glassaffär, och ber barnet välja tre (blåbär, lakrits, jordgubbe, vanilj, choklad är möjliga)
- Hur många sätt kan barnet välja på?

Vi söker antalet permutationer av tre smaker när de väljs bland 5 möjliga, och får så ${}_5P_3 = 5 \cdot 4 \cdot 3 = 60$ olika val för barnet.

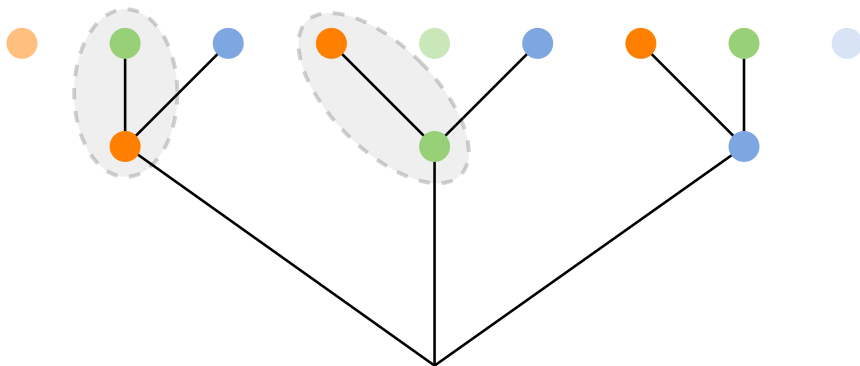
Det finns alltså 60 olika sätt för barnet att **beställa** glassen, men för föräldern som ska äta glassen är “blåbär, jordgubbe, vanilj” likvärdigt med “vanilj, blåbär, jordgubbe”.

Vi är här intresserade av **kombinationer** snarare än permutationer, då det inte spelar någon roll vilken **ordning** smakerna beställs i.

välj $k = 2$ bland $n = 3$

utan återläggning - utan hänsyn till ordning

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



1:a 2:a



- Vi vet att ${}_nP_k$ ger antalet permutationer av k objekt valda bland n möjliga
- Vi vet också att varje mängd av k objekt kan ordnas på $k!$ olika sätt
- Om vi inte bryr oss om ordningen blir antalet **kombinationer** ${}_nP_k/k!$

Antalet **kombinationer** (utan hänsyn till ordning) av k objekt valda ur n utan återläggning ges av

$${}_nC_k = \frac{{}_nP_k}{k!} = \frac{n!}{k!(n-k)!} = \binom{n}{k}$$

Barnet kan **beställa** glass på ${}_5P_3 = 60$ sätt, alltså att det finns 60 *permutationer*.

Men: antalet “unika glassar” ges av antalet *kombinationer*, alltså

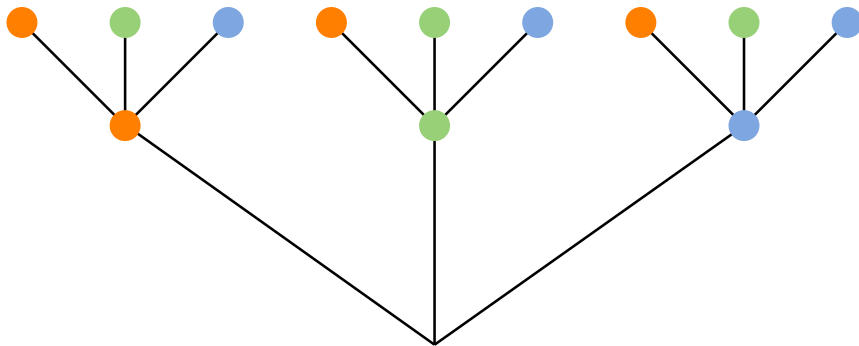
$${}_5C_3 = \frac{5!}{3! \cdot (5-3)!} = \frac{5 \cdot 4 \cdot \mathbf{3!}}{\mathbf{3!} \cdot 2 \cdot 1} = \frac{5 \cdot 4}{2} = 10.$$

- Du har glömt din firsiffriga portkod, och tänker att du ska chansa tills du slår rätt – hur många möjliga portkoder finns att slå?
- Här har vi dragning
 - Med hänsyn till ordning: koden 1995 är **inte** samma som koden 5991
 - Med återläggning: vi kan välja bland **alla** siffror på **varje** position
- Första siffran väljs bland $\{0, 1, 2, \dots, 9\}$, och det samma gäller för övriga siffror
- Antalet möjliga portkoder blir så

$$10 \cdot 10 \cdot 10 \cdot 10 = 10^4 = 10000$$

Vid dragning av k objekt från n möjliga, med återläggning och med hänsyn till ordning, finns n^k möjliga permutationer.

välj $k = 2$ bland $n = 3$
med återläggning - med hänsyn till ordning
 $n^k = 3^2 = 9$



1:a 2:a



Sammanfattning: Att räkna på permutationer och kombinationer

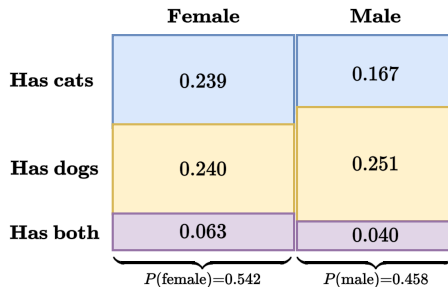
	Med återläggning	Utan återläggning
Med ordning	n^k	${}_nP_k = \frac{n!}{(n-k)!}$
Utan ordning	ej på kurs	${}_nC_k = \binom{n}{k} = \frac{n!}{k!(n-k)!}$

Betingade sannolikheter

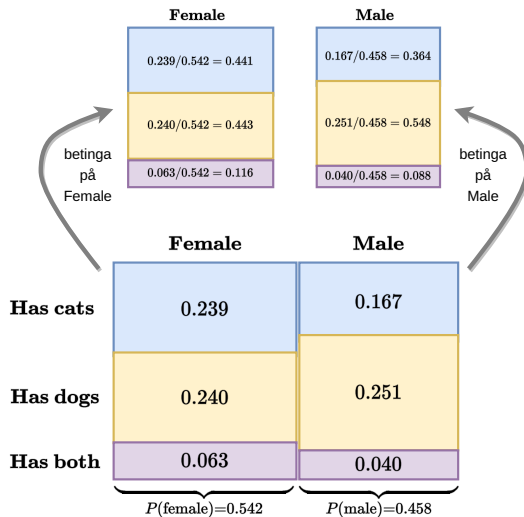
- Vi antar återigen att vi har två olika händelser, A och B , i utfallsrummet S
- Betingade sannolikheter används för att svara på frågor av typen “hur sannolikt är det att händelse A inträffar, givet att händelse B redan har inträffat”
- Vi skriver $P(A|B)$, där $A|B$ läses som A givet B eller A betingat på B
- Händelse B behöver inte ha inträffat när vi räknar på sannolikheten, utan vi kan anta att B inträffar för att se hur det påverkar sannolikheten för A
- **Exempel:**
 - Vad är sannolikheten att fotbollslaget A vinner matchen givet att lag B gör första målet?
 - Vad är sannolikheten att lunchrestaurangen serverar ärtsoppa givet att det är torsdag?
 - Vad är sannolikheten för att lunchrestaurangen serverar ärtsoppa givet att det är tisdag?
- Vi kan ankyta till resonemang i F3 för att reda ut hur vi bör tänka

- Vi kan beskriva fördelningen av variablerna **Pets** och **Gender** med korstabell/mosaic plot

		Gender		
		Female	Male	Total
Pets	Has cats	23.9%	16.7%	40.6%
	Has dogs	24.0%	25.1%	49.1%
	Has both	6.3%	4.0%	10.3%
	Total	54.2%	45.8%	100%

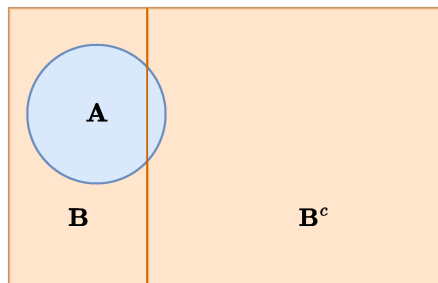


- Om vi vill betinga på **Gender** går vi igenom varje kolumn separat, och beräkna kolumnvisa andelar
- Givet att **Gender** = **Female**, hur stor andel av individerna har katt/hund/båda?

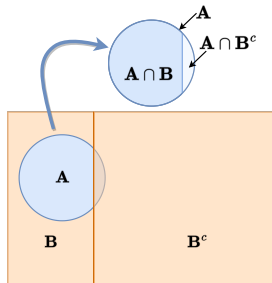


- Vi använder samma procedur/tankesätt när vi vill beräkna $P(A|B)$!

- Under pandemin kunde vi testa oss själva för covid, med viss felrisk
- Vi definierar händelserna $A = \{\text{Positivt hemtest}\}$ och $B = \{\text{Har covid på riktigt}\}$
- Pga risk för felaktiga tester söker vi $P(\text{Har covid på riktigt} | \text{Positivt hemtest}) = P(B|A)$
- Det finns många olika utfall, covidsjuka kan t.ex. både få positivt och negativt tester
- I allmänhet ges utfallsrummet av $S = \{\{A, B\}, \{A^c, B\}, \{A, B^c\}, \{A^c, B^c\}\}$
- Vilken area i diagrammet motsvarar den sannolikhet vi är ute efter?



- När vi betingar så “zoomar vi in” på ett nytt utfallsrum, precis som när vi fokuserade på kolumnen **Female** – vi betingar på A , så cirkeln A blir vårt nya utfallsrum



Betingad sannolikhet: Sannolikheten för B , betingat på A , ges av

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

$P(B|A)$ är alltså andelen av arean A som faller inom arean B .

Allmänna multiplikationsregeln: Vi kan kasta om i formeln för $P(B|A)$ och få

$$\underbrace{P(A \cap B)}_{\text{snittsannolikhet}} = \underbrace{P(A)}_{\text{marginell sannolikhet}} \cdot \underbrace{P(B|A)}_{\text{betingad sannolikhet}}$$

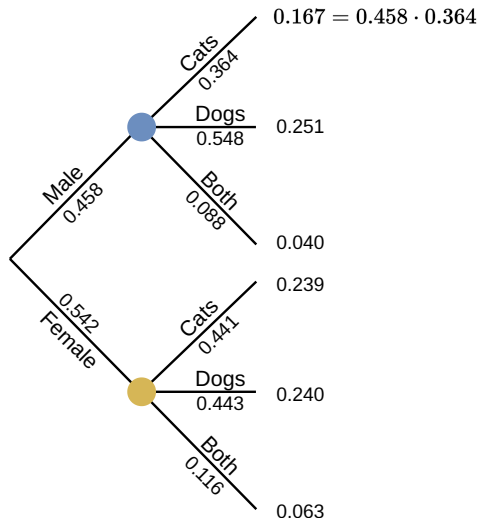
- **Från F11:** A och B är oberoende om (och endast om) $P(A \cap B) = P(A) \cdot P(B)$
- Utifrån regeln ovan, ser vi att detta villkor håller om $P(B|A) = P(B)$

Oberoende händelser (variant): A och B är oberoende om (och endast om)

$$P(B|A) = P(B)$$

- **Rimligt!** Oberoende betyder att A och B inte påverkar varandras sannolikheter
- **Notera:** Oberoende är **inte** liktydigt med disjunkt – oberoende händelser kan inträffa samtidigt, men det kan **inte** disjunkta!

- Beslutsträd är ett bra sätt att illustrera denna typ av information
- Marginella sannolikheter, t.ex. $P(\text{Male})$, finns i första grenen
- Betingade sannolikheter, t.ex. $P(\text{Cats}|\text{Male})$, finns i andra grenen
- Snittsannolikheter, t.ex. $P(\text{Cats} \cap \text{Male})$, finns på slutet



- Om vi **bara** vill förmedla snittsannolikheter och marginella sannolikheter är en korstabell fortfarande det bästa alternativet
- Då finns ingen anledning att lägga till den extra abstraktionsnivån i ett beslutsträd

simultansannolikheter (joint) kön och husdjur

		Gender		Total
		Female	Male	
Pets	Has cats	23.9%	16.7%	40.6%
	Has dogs	24.0%	25.1%	49.1%
	Has both	6.3%	4.0%	10.3%
	Total	54.2%	45.8%	100%

marginalsannolikheter
husdjur

marginalsannolikheter kön

marginalsannolikheter
husdjur

Bayes sats och satsen om total sannolikhet

- Sedan tidigare har vi den **allmänna multiplikationsregeln**

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

- Vi har även formeln för en **betingad sannolikhet**

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

- Vi kan nu kombinera dess, och få **Bayes sats**

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

- Bayes sats låter oss **vända betingningar**: från $P(B|A)$ får vi $P(A|B)$ (och vice versa)
- Detta kan vara värdefullt om en av dem är lättare att beräkna än den andra

- På en flygplats har vi $A = \{\text{metalldetektorn piper}\}$ och $B = \{\text{förbjudet föremål}\}$
- När de ser över sina säkerhetsrutiner vill personalen veta

$$P(B|A) = P(\text{förbjudet föremål} \mid \text{pip}) = \frac{P(\text{förbjudet föremål} \cap \text{pip})}{P(\text{pip})}$$

- Rakt av är denna sannolikhets svår att beräkna – personalen måste vänta på ett stort antal pip, och sedan avgöra om varje pip indikerar förbjudet föremål
- Förbjudna föremål är också relativt ovanliga, så **många** kontroller skulle behövas
- **Men:** Bayes sats låter oss beräkna sannolikheten som

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

- I detta fall är $P(A|B)$ lättare att beräkna – personalen kan själva testa detektorn med olika förbjudna föremål och se hur ofta den piper
- Huvudsaken är att vi **vänder betingningen** till vår fördel!

Vi har A och B som i föregående slide, och har nu fått informationen att

$$P(A) = 0.01, P(B) = 0.001, \text{ och } P(A|B) = 0.99.$$

Hur kan vi beräkna $P(B|A)$?

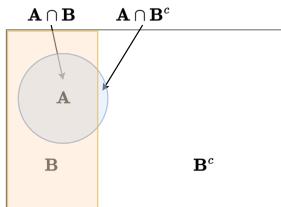
*Vi använder **Bayes sats!** Vi får så att*

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{0.99 \cdot 0.001}{0.01} = 0.099$$

Om detektorn piper är sannolikheten att personen bär ett förbjudet föremål cirka 9.9%.

- Sannolikheten för en händelse A kan delas upp som

$$P(A) = P(A \cap B) + P(A \cap B^c)$$



- De enskilda termerna i uppdelningen kan skrivas om m.h.a. allmänna multiplikationssatsen

$$P(A \cap B) = P(A|B)P(B) \text{ och } P(A \cap B^c) = P(A|B^c)P(B^c)$$

Satsen om total sannolikhet (för två händelser):

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$

- Bayes sats säger att

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

- Lagen om total sannolikhet säger att

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$

Bayes sats (alternativ version):

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

- Denna version av Bayes sats är lämplig om vi inte har $P(A)$ rakt av, utan bara uppdelad i termer av någon annan händelse – detta är vanligare än man tror!

- Vi har händelserna $A = \{\text{Positivt test}\}$ och $B = \{\text{Covid}\}$, och söker $P(B|A)$
- I bruksanvisningen till ett test kan vi läsa

antigentest för detektering av Covid-19 är ett CE-certifierat test för självprovtagning som kan indikera pågående infektion av coronavirus. Snabbtest som utförs genom nästopsning i främre näsan. Med hög **specificitet på 99,20 %** samt hög **sensitivitet på 96,77 %**. Passar för screening av symptomfria individer exempelvis på arbetsplatser.

- **Översättning:**
 - **Prevalens:** $P(\text{Covid})$ - andel med covid i populationen (rörligt över tid)
 - **Sensitivitet:** $P(\text{Pos}|\text{Covid})$ - hur känsligt är testet för att upptäcka covid?
 - **Specificitet:** $P(\text{Neg}|\text{Ej covid})$ - är testet specifikt för covid, eller reagerar det på annat?
- Notera att $P(\text{Neg}|\text{Ej covid}) = 1 - P(\text{Pos}|\text{Ej covid})$ – **varför gäller detta?**

- Bayes sats säger att

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

- Översatt till vårt exempel blir det

$$P(\text{Covid}|\text{Pos}) = \frac{P(\text{Pos}|\text{Covid})P(\text{Covid})}{P(\text{Pos}|\text{Covid})P(\text{Covid}) + P(\text{Pos}|\text{Ej covid})P(\text{Ej covid})}$$

- Vi stoppar in siffrorna från föregående slide

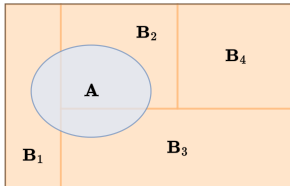
$$P(\text{Covid}|\text{Pos}) = \frac{0.9677 \cdot P(\text{Covid})}{0.9677 \cdot P(\text{Covid}) + (1 - 0.9920) \cdot P(\text{Ej covid})}$$

- Om 5% i populationen har Covid får vi

$$P(\text{Covid}|\text{Pos}) = \frac{0.9677 \cdot 0.05}{0.9677 \cdot 0.05 + (1 - 0.9920) \cdot 0.95} \approx 0.8642$$

- Ni kan titta närmare på detta exempel i vår [widget](#)!

- Ibland kan utfallsrummet vara uppdelat i flera olika händelser B_1, B_2, B_3 , osv

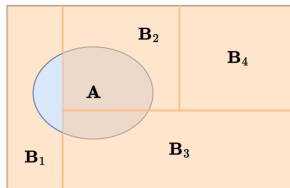


- Data om metalldetektortjutfar kan t.ex. uppdelade enligt $B_1 = \{\text{Farligt föremål}\}$, $B_2 = \{\text{Glömt nycklarna i fickan}\}$, $B_3 = \{\text{Virknål i backfickan}\}$, etc

Satsen om total sannolikhet uppdelad på K händelser $B_1, \dots, B_K$:

$$P(A) = \sum_{j=1}^K P(A|B_j)P(B_j) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \dots + P(A|B_K)P(B_K)$$

- Den allmänna lagen om total sannolikhet ger en mer allmän version av Bayes sats



Bayes sats för utfallsrum uppdelat på K händelser $B_1, \dots, B_K$:

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_{j=1}^K P(A|B_j)P(B_j)}$$

- I vissa situationer är data/information uppdelad i den här typen av kategorier, och då vet vi att vi kan använda de mer generella resultaten

Tack för idag!!