Föreläsning 1: Introduktion

Matematisk statistik

David Bolin
Chalmers University of Technology
August 28, 2017

Lärare

David Bolin:	Kursansvarig och föreläsare
Rum:	H3028
E-mail:	david.bolin@chalmers.se
Anders Hildeman:	Övningsledare
Rum:	L2121
E-mail:	hildeman@chalmers.se
Helga Kristín Ólafsdóttir :	Övningsledare
Rum:	L3070
E-mail:	khelga@chalmers.se

Tider och litteratur

	Dag	Tid	Plats	Notera
Föreläsning	Måndag	13-15	FB	
Övning	Tisdag	10-12	FL71, FL72	Ibland andra salar
Föreläsning	Onsdag	10-12	KB	
Övning	Fredag	8-10	KS11, KS32	v6 $\rightarrow$ Mån v7
Datorövning	Fredag	13-15	HB105, HB110	Endast v3-6
			Vasa A & B	v3
Återlämning	Torsdag	13-15	KD1	v8

Kurslitteratur:
Milton & Arnold: Introduction to probability and statistics (4 ed)

Hemsida:
http://www.math.chalmers.se/Stat/Grundutb/CTH/tma074/1718/

Examination

Examinationen har två moment

- Skriftlig tentamen i slutet av kursen.
- En projektuppgift.

Projektuppgiften:

- Genomförs i grupper om två eller tre studenter.
- Redovisas med en skriftlig rapport.
- Är obligatorisk men påverkar inte graderingen 3,4 eller 5.
- Kommer att jobbas på under datorövningarna.
- Mer information kommer i samband med att projektet delas ut, innan den första datorövningen.

Kursinnehåll

Kursens syfte:
Kursen avser att ge grunderna av sannolikhetsläran och statistiken med speciellt beaktande av sådana moment som är av betydelse för kemister.

Lärandemål:
Efter fullgjord kurs ska studenten behärska grundläggande begrepp inom sannolikhetsteori och statistik. Kursen ska även ge de studerande förmågan att planera statistiska undersökningar och utföra enkla statistiska analyser.

Upplägg:

- Vecka 1-3: Sannolikhetsteori
- Vecka 4-7: Inferens, regressionsanalys och försöksplanering

Statistikens uppgift

Statistikens uppgift är att bidra med metoder för att analysera och dra slutsatser kring data som innehåller olika typer av osäkerhet och variation. Några viktiga områden är

- **Beskrivande statistik:** används för att summera och strukturera data.
- **Utforskande statistik:** används för att få idéer om samband mellan variabler och för att få utökad förståelse hos studerade processer.
- **Inferens:** används för att skatta parametrar i statistiska modeller eller för att dra formella slutsatser genom att motbevisa hypoteser kring data.
- **Försöksplanering:** används för att planera och genomföra statistiska undersökningar.

Typiska problemställningar

Det som typiskt karakteriserar ett statistiskt problem är att vi har en stor grupp (population) som vi vill analysera. Vi kan inte studera populationen i sin helhet och måste därför uttala oss om dess egenskaper baserat på ett urval (sampel).

Några vanliga statistiska problemställningar är att

- studera en parameter i en statistik modell, som till exempel andelen defekta enheter i en produktion.
- jämföra olika grupper, som till exempel att jämföra effektiviteten hos olika produktionsanläggningar.
- analysera samband mellan olika variabler, som till exempel hur höjd över havet påverkar lufttemperaturen.
- optimera en process, för att till exempel maximera utbytet eller minimera utsläppen hos en produktion.

Sannolikhetsteori och dess förhållande till statistiken

I sannolikhetsteori konstruerar och analyserar man matematiska modeller som kan användas för att beskriva fenomen som uppvisar variation.

Inom statistiken vill man baserat på data uttala sig egenskaper hos en tänkt modell för fenomenet.

Antag till exempel att vi har en perfekt tärning, då är sannolikheten att få en femma $1/6$. Med hjälp av detta kan vi uttala oss om sannolikheten för andra händelser, som att få två femmor i rad.

Inom statistikteorin är frågan snarare att vi vill testa om en given tärning är symmetrisk, dvs att sannolikheten att få t.ex. en femma är $1/6$ och inte något annat. Vi kan göra detta genom att kasta tärningen många gånger och räkna andelen femmor vi får.

Beskrivande statistik

När man ska analysera en datamängd är det en bra idé att först illustrera den grafiskt. Detta kan ge idéer om hypoteser att testa eller om samband mellan variabler.

Frekvenstabell:
För diskret data, dvs data som bara kan anta ett ändligt antal värden, kan man sammanfatta datan i en frekvenstabell som visar hur många observationer vi har för varje möjligt utfall.

Stolpdiagram:
Med hjälp av en frekvenstabell kan vi sedan rita ett stolpdiagram (även kallat stapeldiagram) där man till varje värde ritar en stapel vars höjd är proportionell mot antalet observationer för detta värde.

Sekvensdiagram:
Det kan också vara intressant att visa data med hjälp av ett sekvensdiagram för att studera trender i värden. Vi ritar då upp värdena i en följd.

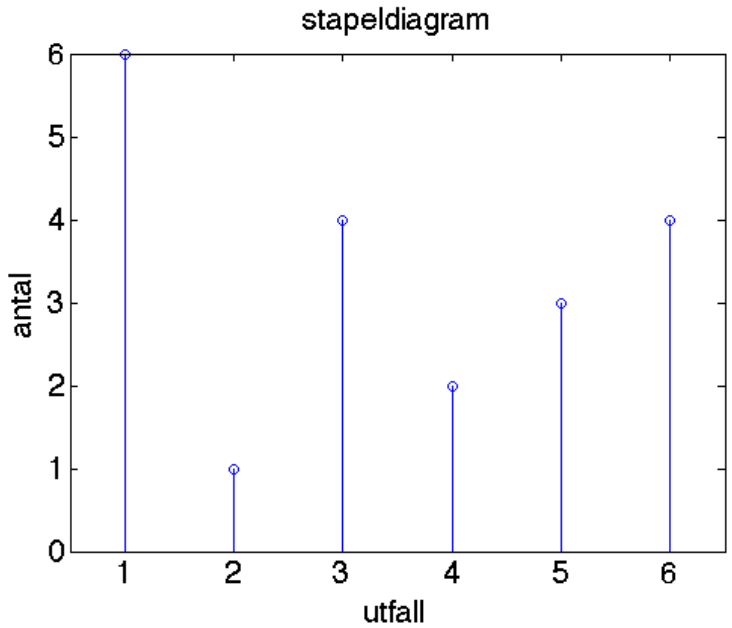
Exempel 1

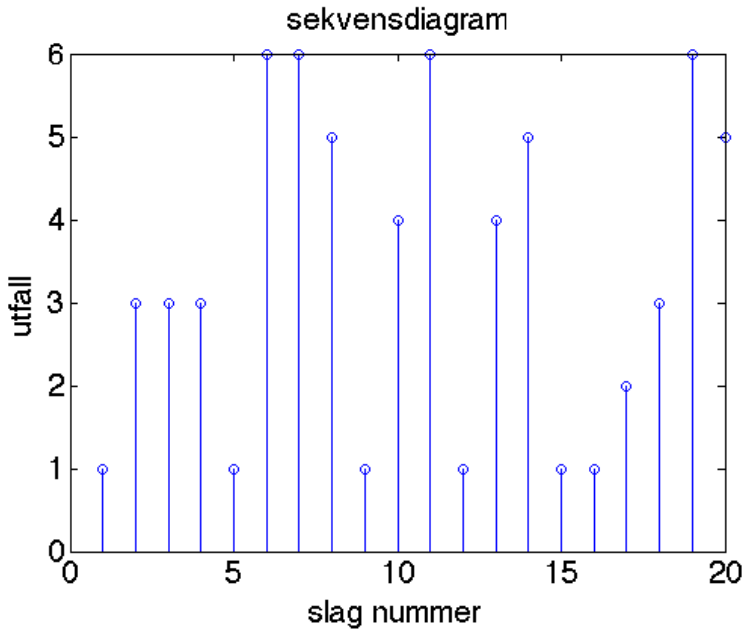
Antag att vi slår en tärning 20 gånger med följande resultat:

$$1, 3, 3, 3, 1, 6, 6, 5, 1, 4, 6, 1, 4, 5, 1, 1, 2, 3, 6, 5$$

Frekvenstabell:

Utfall	1	2	3	4	5	6
Antal gånger	6	1	4	2	3	4
Andel gånger	0.30	0.05	0.20	0.10	0.15	0.20





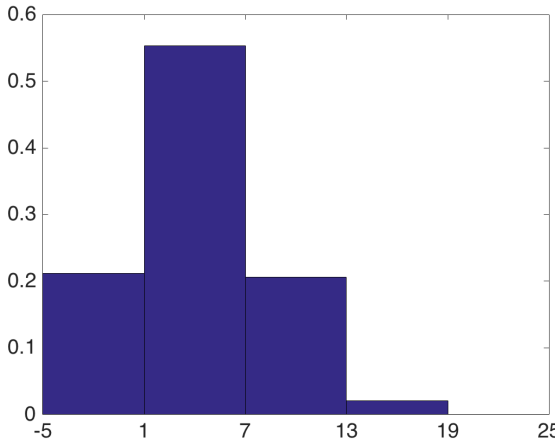
Histogram

För kontinuerlig data, det vill säga data som kan anta ett oändligt (överuppräknerligt) anatal värden, används ofta histogram:

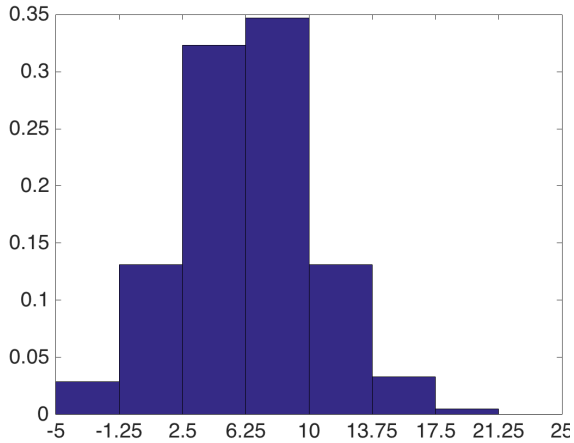
- Dela in datan i ett antal klasser (intervall) och räkna sedan antalet observationer i varje klass.
- Rita ett stolpdiagram där höjden är proportionell mot antalet i klassen och bredden är den samma som intervallbredden.
- Exempel: Sammanfatta följande 1000 reella tal:

$$12.15, 17.33, 0.96, 13.44, 11.27, 4.76, 8.26, 11.37, 24.31, 21.07, \ldots$$

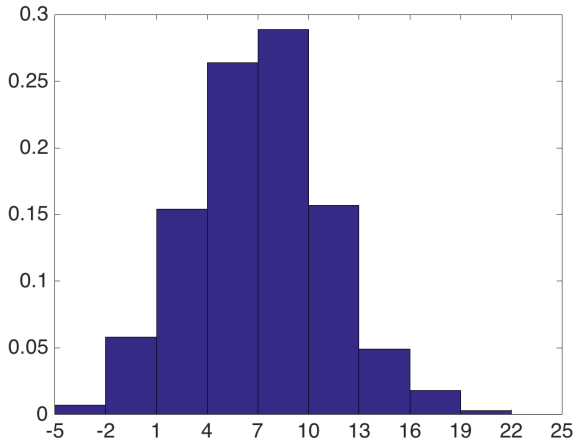
4 klasser



7 klasser



9 klasser



Numerisk beskrivning av data

Förutom att visa data grafiskt brukar man också beräkna vissa numeriska tal som beskriver datamängden. Man brukar beräkna:

- Lägesmått som beskriver vara datan är centrerad.
- Spridningsmått som beskriver variationen kring lägesmåttet.

Vanliga lägesmått är:

Medelvärde:

Medelvärdet (stickprovsmedelvärdet) av en datamängd $x_1, \ldots, x_n$ ges av

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \ldots + x_n)$$

Median:

Medianen av en datamängd $x_1, \ldots, x_n$ ges av det mittersta värdet efter att man har sorterat värdena. Om antalet, n, är jämnt brukar man ta medelvärdet av de två mittvärdena.

Numerisk beskrivning av data

- Medelvärdet är också tyngdpunkten i ett stolpdiagram.
- Medelvärdet påverkas av värdet på alla observationer, och framförallt om man har någon enstaka observation som skiljer sig kraftigt mot de andra (en sk outlier) påverkar den medelvärdet men inte medianen.

Vanliga spridningsmått är

Stickprovsvarians: Variansen av en datamängd $x_1, \ldots, x_n$ ges av

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}((x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2)$$

Variansen kan också beräknas som

$$s^2 = \frac{1}{n-1}(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2) = \frac{1}{n-1}(x_1^2 + \ldots + x_n^2 - n\bar{x}^2)$$

Standardavvikelse: ges av $\sqrt{s^2}$, dvs kvadratroten av variansen.

Exempel 1 (fortsatt)

Antag att vi slår en tärning 20 gånger med följande resultat:

$$1, 3, 3, 3, 1, 6, 6, 5, 1, 4, 6, 1, 4, 5, 1, 1, 2, 3, 6, 5$$

Vi beräknas medelvärde:

$$\bar{x} = (1 + 3 + 3 + \ldots + 3 + 6 + 5)/20 = 67/20 = 3.35$$

Vi beräknar medianen genom att sortera värdena och välja det tioende värdet, vilket är 3.

Variansen fås som

$$s^2 = ((1 - 3.35)^2 + (3 - 3.35)^2 + \ldots + (5 - 3.35)^2)/19 = 3.8184$$

och standardavvikelsen är $s = 1.9541$.

Sannolikhetsteori: Utfall och utfallsrum

Man brukar säga att ett slumpmässigt försök är ett försök som kan upprepas under väsentligen identiska förhållanden där utfallet av försöket inte exakt kan förutsägas i det enskilda fallet. Till exempel

1. kasta en tärning och räkna antalet prickar.
2. singla ett mynt två gånger och registrera resultatet i varje kast.
3. Undersök en enhet från en tillverkningsprocess och se om den är hel.

Resultatet av ett försök brukar kallas för ett *utfall* ω, och mängden av alla möjliga utfall kallas för *utfallsrummet* Ω. I exemplen ovan är utfallsrummet

1. $\Omega = \{1,2,3,4,5,6\}$.
2. $\Omega = \{(\text{krona},\text{krona}),(\text{krona},\text{klave}),(\text{klave},\text{krona}),(\text{klave},\text{klave})\}$.
3. $\Omega = \{\text{defekt},\text{hel}\}$.

Händelser

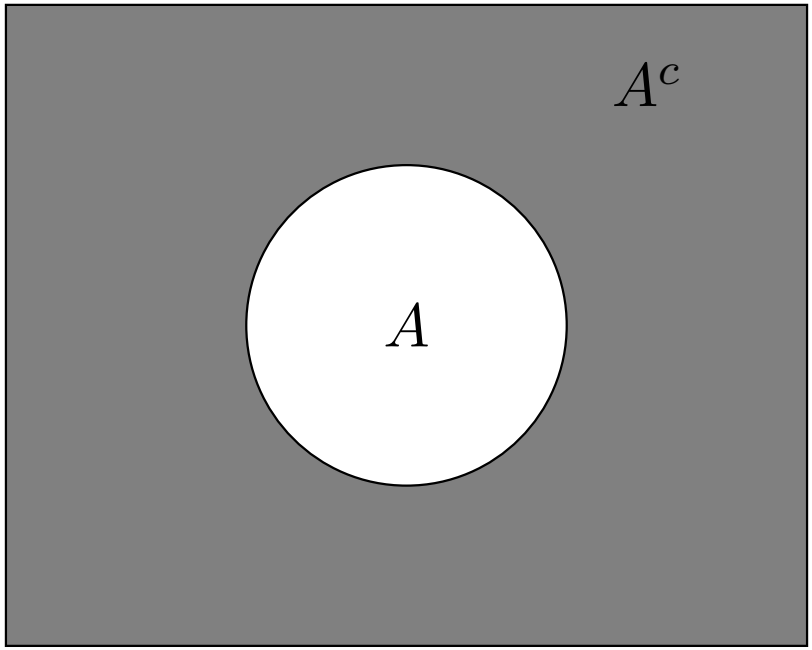
Man kan också sätta samman olika utfall till *händelser*.

En händelse A är en mängd av utfall, det vill säga en delmängd av utfallsrummet Ω.

Från exemplen ovan har vi till exempel händelser

1. $A = \{1,3,5\}$, dvs vi får ett udda tal.
2. $A = \{(\text{krona},\text{krona}),(\text{klave},\text{krona})\}$, dvs andra kastet ger krona.
3. $A = \{\text{hel}\}$, dvs enheten är hel.

Komplement

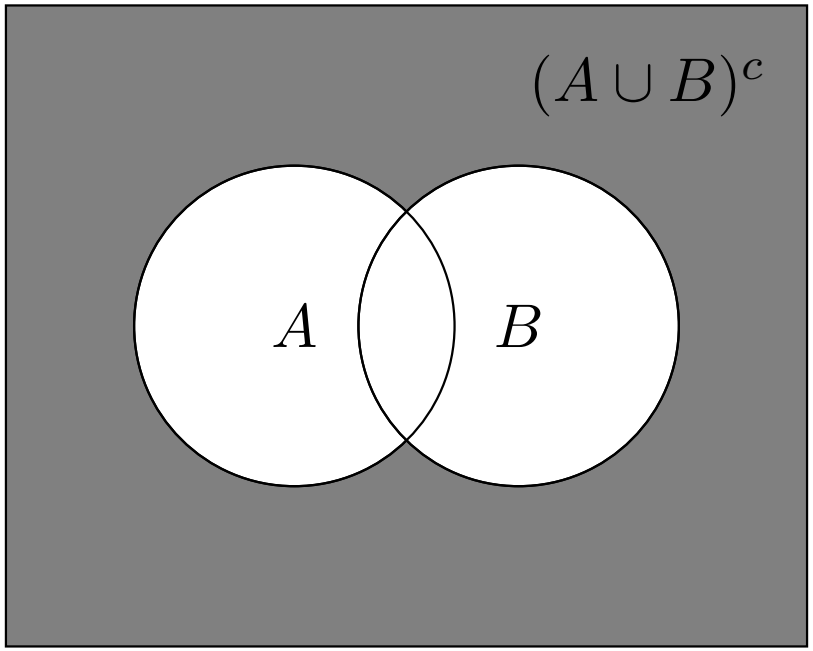


Att händelsen A inträffar betyder att något av utfallen i A inträffar. Om något av utfallen i A inte inträffar säger vi att komplementet till A inträffar

$$A^c = \Omega \setminus A.$$

I exemplet med tärningen inträffar t.ex. A om vi får en etta, om vi däremot får en tvåa säger vi att A^c inträffat.

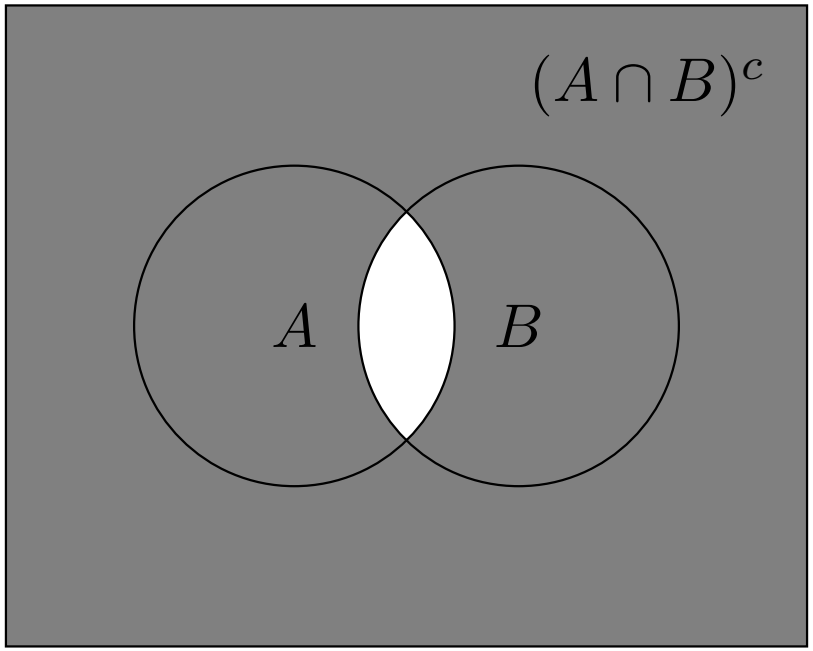
Union



Om vi har två händelser A och B så kan vi även definiera $A \cup B$ som är unionen av A och B.

- Om ett utfall i $A \cup B$ inträffar så säger vi att A eller B inträffar.

Snitt



Vi kan även definiera snittet $A \cap B$ som är mängden av element som finns i A och B.

- Om ett utfall i $A \cap B$ inträffar så säger vi att A och B inträffar.

Den tomma mängden $\emptyset$ brukar kallas för en omöjlig händelse. Om $A \cap B = \emptyset$ så säger vi att A och B är oförenliga, eller utesluter varandra.

Tolkning av sannolikhetsbegreppet

Om vi har en händelse A så menar vi lite löst att sannolikheten för A, $\mathrm{P}(A)$, är troligheten för att A ska inträffa.

- Sannolikheter är tal mellan 0 och 1.
- Om sannolikheten för A är nära 1 betyder det att det är troligt att A inträffar.
- Om sannolikheten för A är nära 0 betyder det att det inte är troligt att A inträffar.
- Om sannolikheten för A är 0.5 så är det lika troligt att A inträffar som att A inte inträffar.
- Ibland beskriver man sannolikheter i procent, dvs en sannolikhet på 0.1 kan skrivas som 10%.

Den klassiska tolkningen av sannolikhet

- Hur man ska relatera sannolikhetsbegreppet till verkligheten har varit en stor filosofisk tvistefråga, och tolkningen görs påverkar hur vi utformar våra statistiska metoder senare.
- Den klassiska tolkningen av sannolikhet kommer från spelteori och sannolikheten för en händelse A definieras i detta fall som

$$\mathsf{P}(A) = \frac{\text{antal sätt } A \text{ kan inträffa på}}{\text{Totalt antal möjliga utfall}}$$

- I exemplet när vi singlar ett mynt två gånger hade vi totalt fyra möjliga utfall i Ω. Om vi har händelsen A att vi får krona på andra kastet kan detta inträffa på två sätt och vi får därför

$$\mathsf{P}(A) = \frac{2}{4} = \frac{1}{2}$$

Permutationer och kombinationer

Användbara begrepp för att beräkna antalet möjliga utfall:

Permutation

En specifik ordning av ett antal objekt

Kombination

Ett urval av ett antal objekt utan hänsyn till ordning

Multiplikationsprincipen

Om det finns a sätt att utföra ett val och b sätt att utföra ett annat val så finns det ab sätt att utföra det kombinerade valet.

Fakultet

För $n \in \mathbb{N}$ definierar vi $n! = n \cdot (n-1) \cdot (n-2) \cdot \cdots \cdot 2 \cdot 1$ där $0! = 1$. $n!$ utläses som "n-fakultet".

Att räkna antalet kombinationer

Sats

Antalet sätt vi kan plocka ut r av totalt n objekt, med hänsyn tagen till ordningen ges av

$$ {}_nP_r = \frac{n!}{(n-r)!} $$

Sats

Antalet sätt vi kan plocka ut r av totalt n objekt, utan hänsyn tagen till ordningen ges av

$$ {}_nC_r = \binom{n}{r} = \frac{n!}{r!(n-r)!} $$

- ${}_nC_r$ brukar kallas för en binomialkoefficient. Vi kommer återkomma till detta senare.

Tolkning av sannolikhetsbegreppet

- Den klassiska tolkningen av sannolikhet fungerar endast om det är rimligt att anta att alla möjliga utfall är lika troliga.
- Vi kommer i regel använda oss av den så kallade *frekventistiska* tolkningen.
- Antag att vi har ett slumpmässigt försök som vi kan upprepa under identiska förhållanden. När antalet försök, n, växer så konvergerar erfarenhetsmässigt den relativa frekvensen, n_A/n för en händelse A mot ett tal mellan noll och ett. Detta tal definierar vi som sannolikheten för A. Det vill säga

$$\frac{n_A}{n} \to \mathsf{P}(A), \text{när } n \to \infty$$