Föreläsning 1: Introduktion

Matematisk statistik

David Bolin
Chalmers University of Technology
March 22, 2014

Lärare och kurslitteratur

David Bolin:	Kursansvarig, föreläsare, övningsledare
Rum:	H3018
E-mail:	david.bolin@chalmers.se
Fredrik Boulund:	Övningsledare
Rum:	L3102
E-mail:	fredrik.boulund@chalmers.se

Kurslitteratur:
Introduction to probability and statistics (4 ed) av J.S. Milton och J.C. Arnold

Hemsida:
http://www.math.chalmers.se/Stat/Grundutb/CTH/tma073/1314/

Föreläsningar och övningar

	Dag	Tid	Plats
Föreläsning	Måndag	13-15	EC
Övning	Tisdag	8-10	KS11 och KS32
Föreläsning	Torsdag	10-12	KB
Övning	Fredag	10-12	KS11 och KS32

Från läsvecka fyra och framåt har vi även datorövningar:

	Dag	Tid	Plats
Datorövning	Måndag	10-12	KD1 (v4) och KD2 (v5-7)

Läsvecka 2 har vi en liten ändring i schemat:

	Dag	Tid	Plats
Föreläsning	Torsdag	8-10	KB
Övning	Torsdag	10-12	KS11 och KS32

Examination

Examinationen har två moment

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

Projektuppgiften:

- Genomförs i grupper av 3-5 studenter.
- Kommer gå ut på att samla in och analysera egen data.
- Redovisas med en skriftlig rapport.
- Projektet är obligatoriskt med påverkar inte graderingen 3,4 eller 5.
- Mer information kommer i samband med att projektet delas ut.

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 sannolikhetslära och statistisk inferens och kunna utföra enkla statistiska analyser.

Upplägg:

- Vecka 1-3: Sannolikhetsteori
- Vecka 4-7: Statistik och regressionsanalys.

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.
- att jämföra olika grupper, som till exempel att jämföra effektiviteten hos olika produktionsanläggningar eller luftförorening i olika städer.
- att analysera samband mellan olika variabler, som till exempel hur höjd över havet påverkar lufttemperaturen.
- att optimera en process, för att till exempel maximera utbytet eller minimera utsläppen hos en produktion.

Exempel 1

Ett mejeri kräver att andelen osterial förpackningar för mjölk i det långa loppet är högst en promille. Vid en undersökning mäter man 2000 förpackningar och finner 1 osteril.

- Kan de med någon säkerhet påstå att kravet är mött?
- Räcker det att undersöka 2000 förpackningar?

Exempel 2

Man vill testa hur en nyanställd handskas med ett laboratoriums utrustning och ber henne därför mäta klorhalten i ett vattenprov fem gånger med följande resultat

$$56.7, 61.3, 58.5, 62.1, 59.5.$$

Man vet att klorhalten är 60 mg/liter. Vi ser att den anställdes mätningar har en viss spridning samt att medelvärdet är

$$(56.7 + 61.3 + 58.5 + 62.1 + 59.5)/5 = 59.62.$$

Medelvärdet är mindre än 60, men räcker detta för att säga att den anställde mäter halten med ett systematiskt fel? Dvs kommer den anställde i det långa loppet mäta för låga halter, eller beror resultatet på slumpmässigheten i värdena?

Exempel 3

Antag att vi mäter halten av PM_{10} luftförorening i två städer A och B vid 5 tillfällen. Medelvärdet av mätningarna vid stad A blir $15.4\mu g/m^3$ och medelvärdet för stat B blir $14.7g/m^3$.

- Kan vi med någon säkerhet säga att stad A har lägre halt av förorening?
- Hur många mätningar skulle vi behöva göra för att kunna uttala oss om detta?

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.

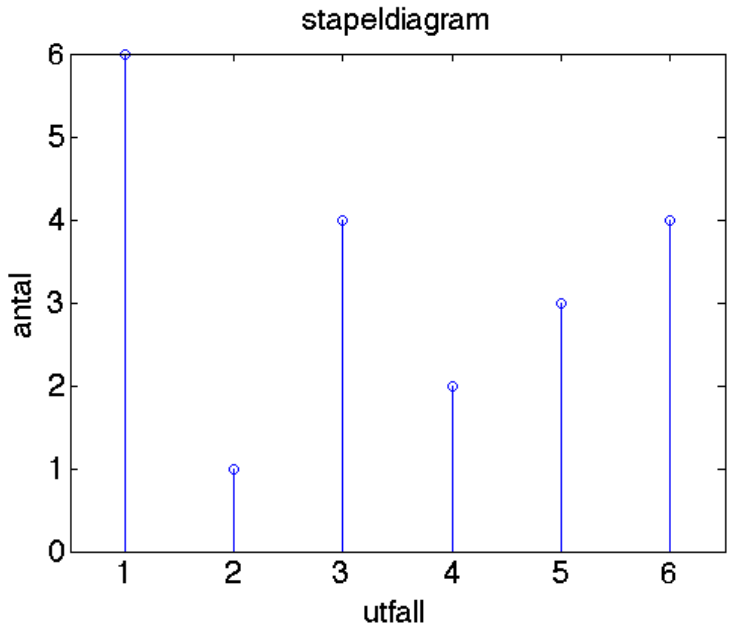
Exempel 4

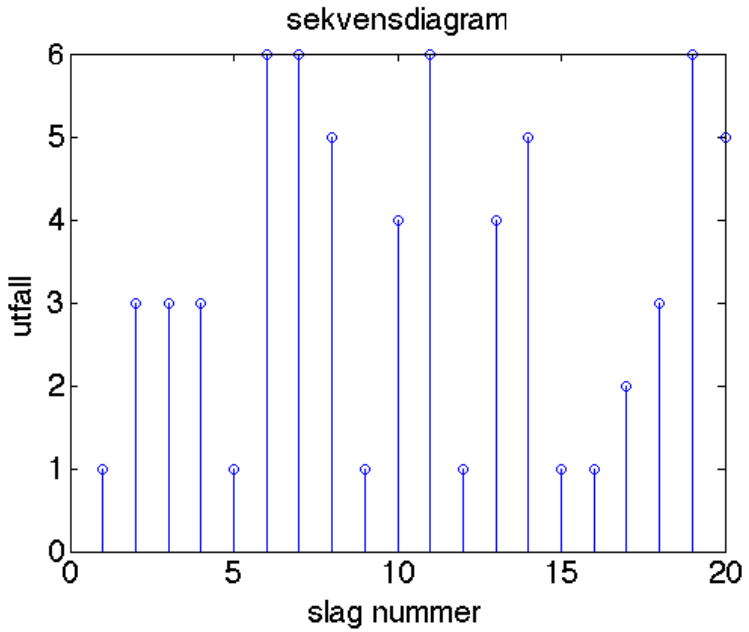
Antag att vi slår en tärning 20 gånger med följande resultat:

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

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



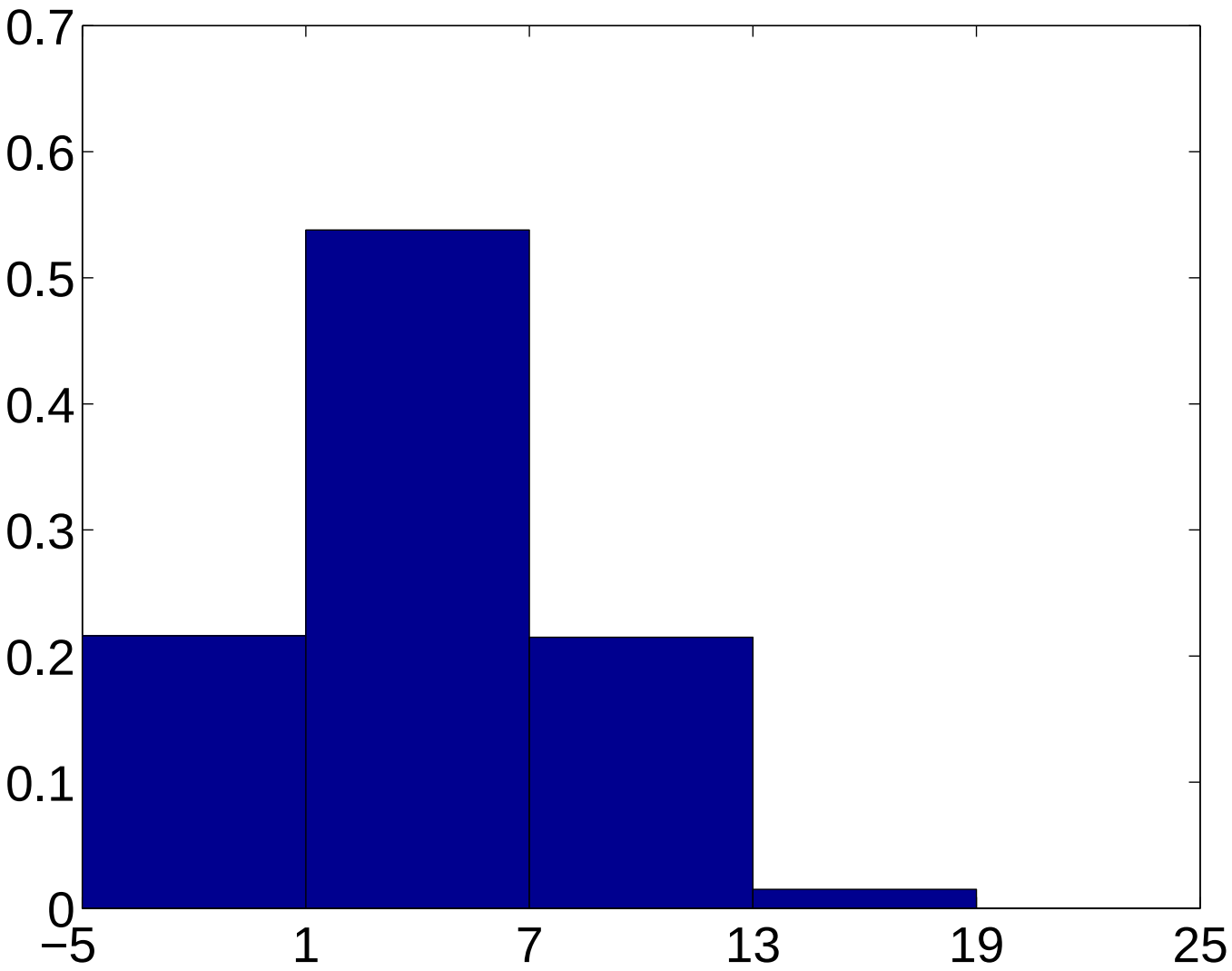


Histogram

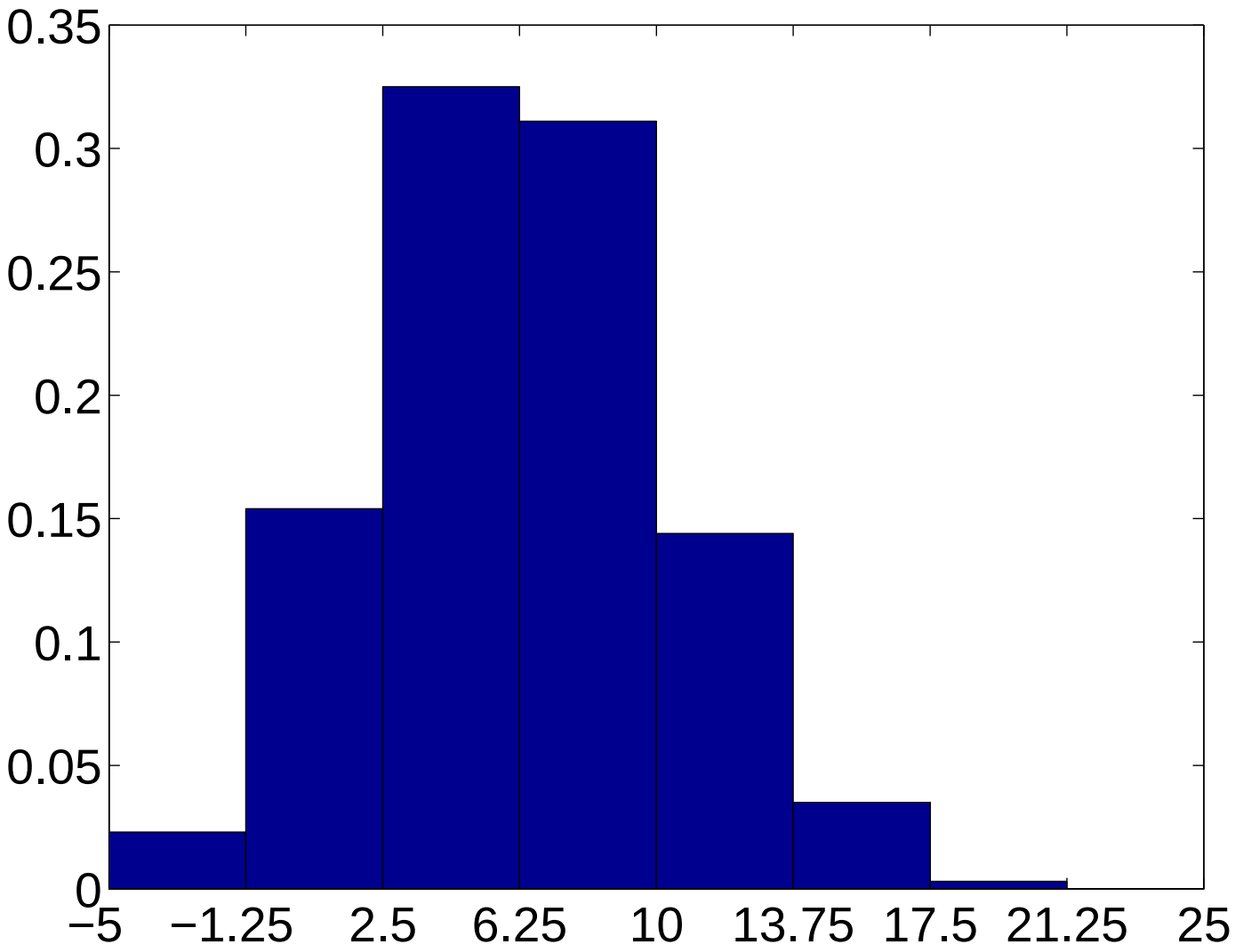
För kontinuerlig data, det vill säga data som kan anta ett oändligt (överuppräknerligt) anatal värden, används ofta histogram:

- Dela in datan i ett antal klasser (intervall) och räknar sedan antalet observationer i varje klass.
- Rita ett stolpdiagram där höjden är proportionell mot antalet i klassen och bredden är den samma som intervallbredden för den klassen.
- I regel väljer man alla klassbredder lika, och ju mer data man har desto smalare kan man göra intervallen.

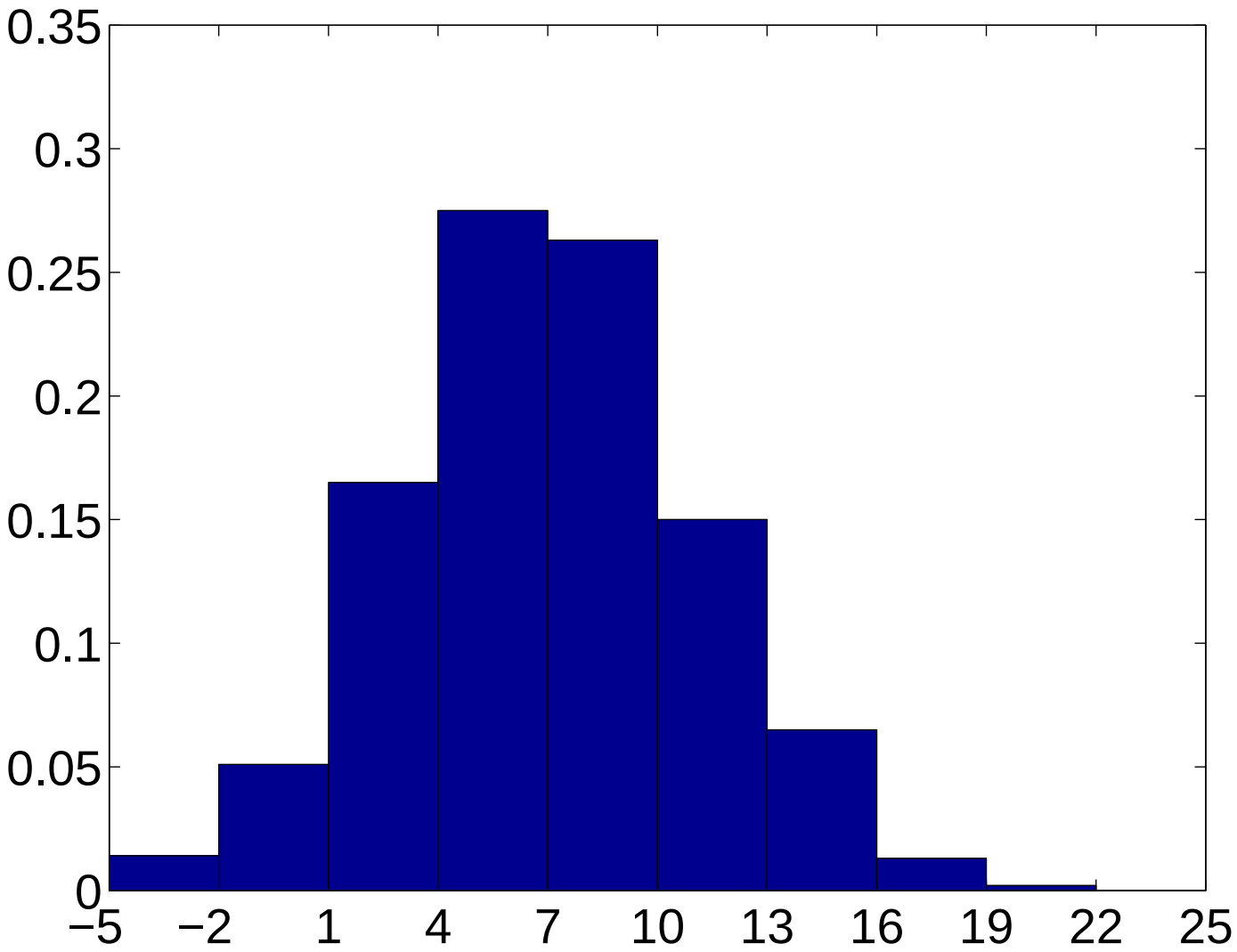
Histogram med 4 klasser



Histogram med 7 klasser



Histogram med 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 4 (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 $s = 1.9541$.

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

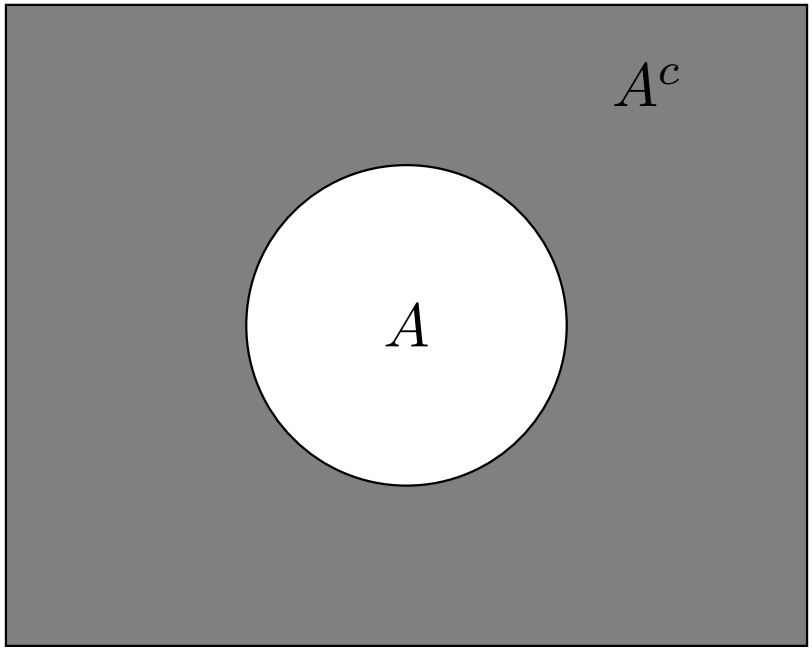
1. $\Omega = \{1,2,3,4,5,6\}$.
2. $\Omega = \{(\text{krona},\text{krona}),(\text{krona},\text{klave}),(\text{klave},\text{krona}),(\text{klave},\text{klave})\}$.
3. $\Omega = \{\text{defekt},\text{hel}\}$.

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

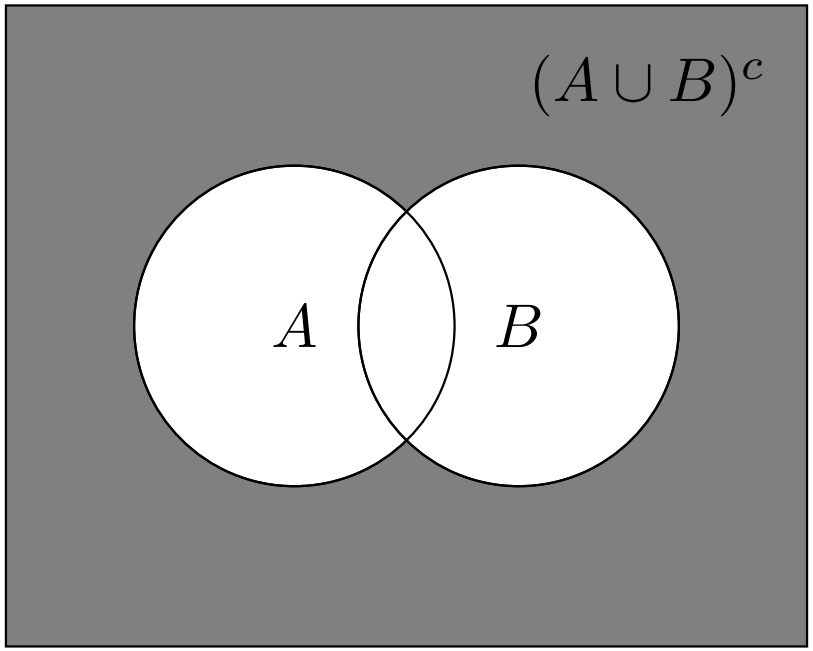


Att händelsen A inträffar betyder att något av utfallen i A inträffar. Om något av utfallen i A inte inträffar säger vi att komplementet till A inträffar

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

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.

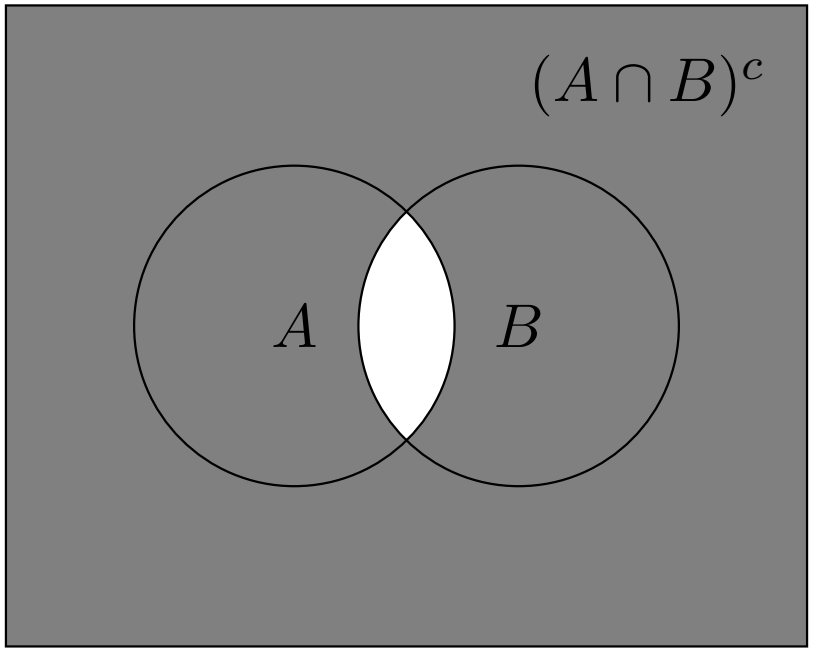
Union



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%, det är ovanligt inom akademin, men desto vanligare i media.

Tolkning av sannolikhetsbegreppet

Exakt hur man ska relatera sannolikhetsbegreppet till verkligheten har varit en stor filosofisk tvistefråga, och hur denna tolkning görs påverkar hur vi utformar våra statistiska metoder senare.

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

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

Den klassiska tolkning av sannolikhet

Den klassiska tolkningen av sannolikhet kommer från spelteori och fungerar endast om det är rimligt att anta att alla möjliga utfall är lika troliga. 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

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