Kapittel 4: Sannsynlighet - Studiet av tilfeldighet

Vi så i forrige kapittel at utvalgsfordeling til en statistikk (observator) er fordelingen av verdiene til statistikken over alle utvalg av samme størrelse fra populasjonen.

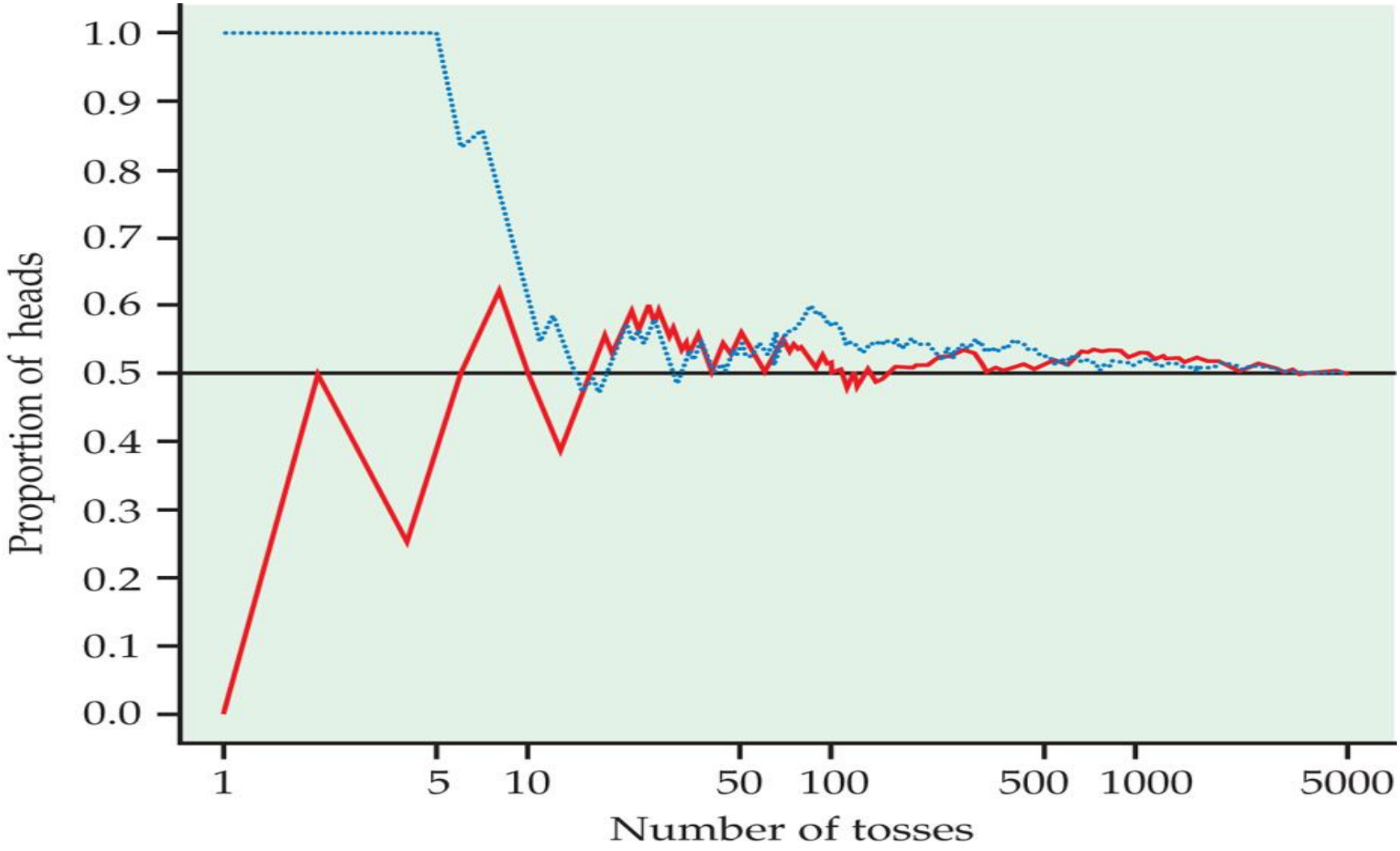
Spesielt la vi merke til at det var trekkingen av ulike utvalg som definerte utvalgsfordelingen til en statistikk.

Men « å trekke utvalg» og «å kaste terninger» er fenomener som kan beskrives ved sannsynlighetsregning.

I begge tilfellene er det slik at et enkelt utfall ikke kan forutsis, men etter mange gjentagelser opptrer det likevel en regelmessighet.

Denne regelmessigheten har vi sett eksempler på at man kan studere ved hjelp av simuleringer. Men det finnes også en matematisk teori for å beskrive regelmessighetene . Vi skal se på noen av viktigste elementene.

Eksempel 5000 myntkast. Utvikling av relativ andel kron.



Tilfeldig fenomen: Hvert enkelt utfall er usikkert, men fordelingen av utfallene er regulær etter mange gjentak.

Sannsynligheten til et utfall er andelen ganger utfallet vil forekomme i mange gjentak.

Sannsynlighetsmodeller

Eksempel: Kasting av mynt
Kan ikke si noe om spesifikt utfall
Kan si noe om mulige utfall (kron/mynt):
Balansert mynt, rimlig å tro hvert utfall har sannsynlighet 0.5

Dette er en eksempel **sannsynlighetsmodell**
Generelt defineres den ved

- En liste av mulige utfall
- Sannsynlighet for hvert mulige utfall

Utfallsrommet S av et tilfeldig fenomen er **mengden** av alle mulige utfall.

Eksempel

a) Kasting av mynt: S= { mynt , kron }.

b) Tilfeldig tall i tabell B: S={ 0 , 1 , 2 , 3 , 4 , 5 , 6 , 7 , 8 , 9 }.

Valg av S:

Har ofte litt valgmuligheter mhp spesifisering av utfallsrommet S, avhengig av hva som er av interesse.

Eksempel (i) Kast av mynt 3 ganger, S kan være

$$S = \{KKK, KKM, KMK, MKK, KMM, MKM, MMK, MMM\}$$

Antall elementer i S er 8

Alternativ: Interessert i antall kron.

S = {0,1,2,3}, antall elementer i S er 4

(ii) Meningsmåling: Tilfeldig utvalg på 3.
Hvert individ svarer Ja eller Nei på spørsmål.

Mulige utfallsrom tilsvarende som kast av 4 mynter.

Sålangt: Intuitiv sannsynlighet ved frekvensfortolkning

Sannsynlighet introdusert gjennom frekvens ved **mange** repetisjoner.

Hvordan beskrive sannsynligheter matematisk?

Vil starte med å sette opp regler frekvenser må følge.

Begivenhet: Et utfall eller et sett av utfall av et tilfeldig fenomen, dvs. en begivenhet er en delmengde av utfallsrommet S.

Eksempel Kasting av 3 mynter: 8 mulige utfall.

La A = 2 kron blant de 3 kastene, dvs.

$$A = \{KKM, KMK, MKK\}$$

så alternativt er A begivenheten to kron og en mynt.

Begivenheter kan knyttes til frekvenser,
som vi tolker som sannsynligheter. Da gjelder:

Grunnleggende egenskaper ved sannsynlighet/frekvenser:

1. Enhver sannsynlighet er et tall mellom 0 og 1, siden en hver andel er et tall mellom 0 og 1.

2. Alle mulige utfall tilsammen må ha sannsynlighet 1.

3. Hvis to begivenheter ikke har noen felles utfall, så er sannsynligheten for at den ene eller den andre inntreffer lik summen av sannsynlighetene for hver av begivenhetene.

4. Sannsynligheten for at en begivenhet ikke inntreffer er 1 minus sannsynligheten for at den inntreffer.
   (Hvis en begivenhet inntreffer i 70% vil den ikke inntreffe i 30% av tilfellene.)

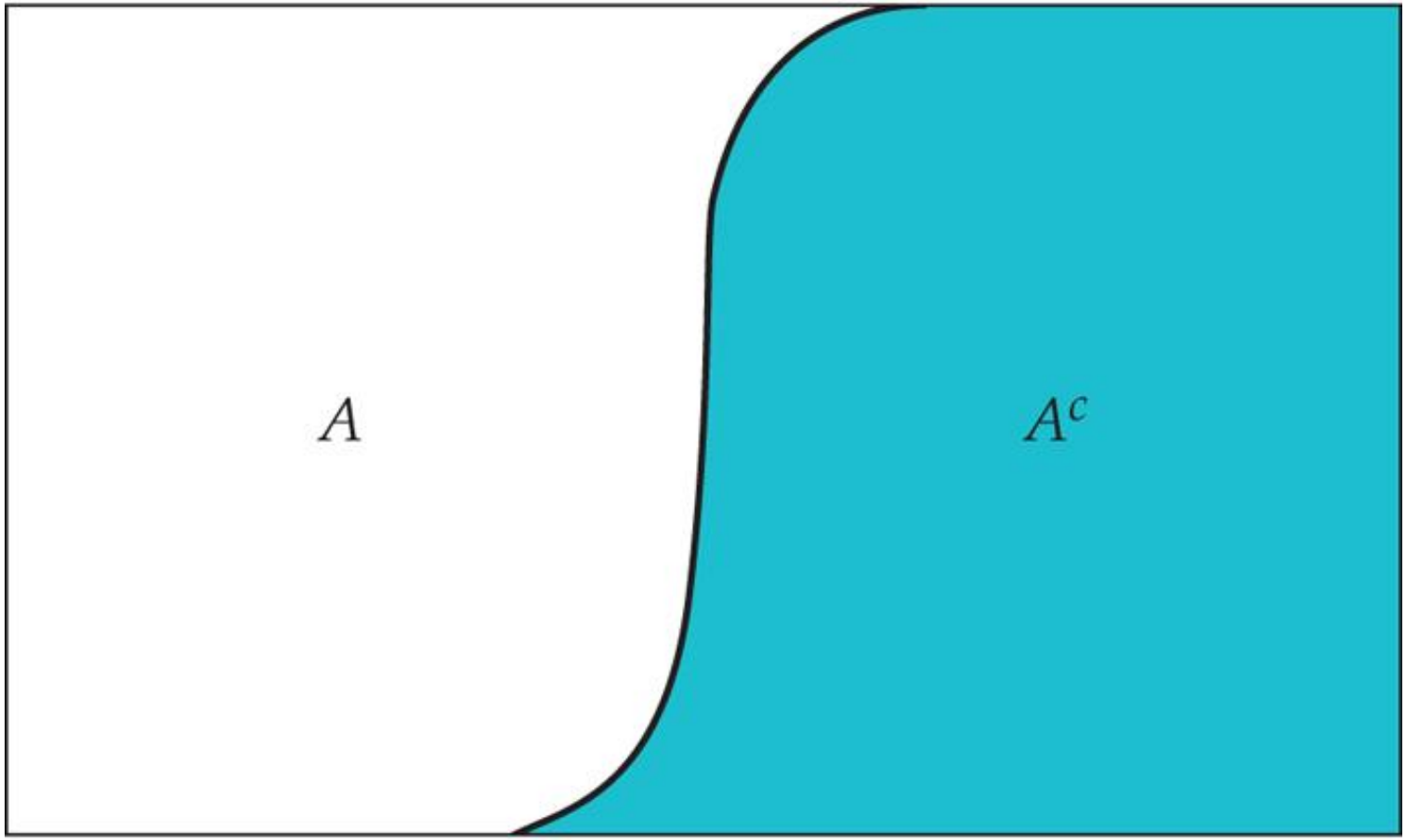
Litt mengdelære:

La A være en begivenhet i utfallsrommet S

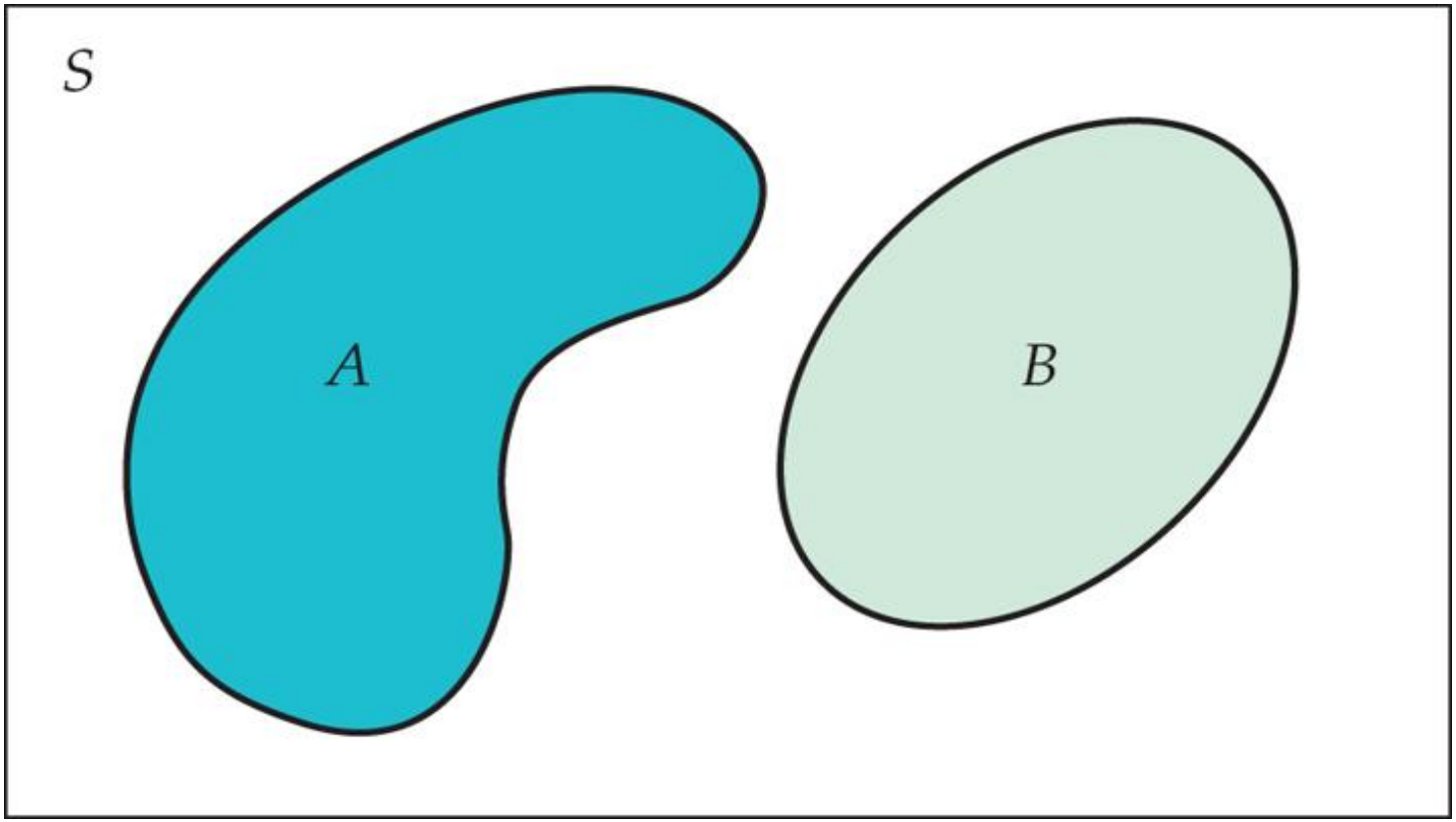
Komplementet til A, beskrevet som A^c er den begivenhet som ikke inntreffer hvis A intreffer.

Alternativt er A^c den mengden av elementer i S som ikke er med i A.

Venndiagrammer er nyttige for å beskrive begivenheter grafisk.


A
A^c

To mengder A og B er **disjunkte** hvis de ikke har noen felles utfall, dvs de inntreffer aldri samtidig.



Regler for sannsynligheter:

La P(A) være sannsynligheten for begivenheten A.

Regel 1 $0 \leq P(A) \leq 1$ for alle A

Regel 2 P(S) = 1

Regel 3 $P(A^C) = 1 - P(A)$ (Komplementær-regelen)

Regel 4 P(A eller B) = P(A) + P(B) når A og B er disjunkte
(Addisjonsregelen for disjunkte begivenheter)

Legg merke til hvordan disse reglene er motivert ut fra hva vi Observerte om frekvenser.

Eksempel

Trekk en tilfeldig dame mellom 25 og 34 år, hva er hennes sivile status?

Sannsynlighetsmodell:

Sivil status :	Aldri gift	Gift	Enke	Skilt
Sannsynlighet	0.298	0.622	0.005	0.075

P(ikke gift) = 1-P(gift)
= 1 - 0.622 = 0.378

P(skilt eller aldri gift) =P(aldri gift) + P(skilt)
= 0.298 + 0.075 = 0.373

Tilordning av sannsynlighet når antall utfall endelig, (altså når antall elementer i utfallsrommet S endelig).

For at reglene for sannsynlighet skal holde må

1. Tilordne en sannsynlighet til hvert individuelt utfall.

2. Disse sannsynlighetene må summere seg opp til 1

3. Sannsynligheten for enhver begivenhet er summen av sannsynligheter av enkeltutfallene i begivenheten.

Eksempel: Innrapportering av gale tall

Første desimal i legale innrapporteringer følger ofte
Benford's lov:

Første desimal :	1	2	3	4	5
Sannsynlighet :	0.301	0.176	0.125	0.097	0.079

Første desimal :	6	7	8	9
Sannsynlighet:	0.067	0.058	0.051	0.046

Kan oppdage svindel ved å se på frekvens av første desimal.

Begivenhet A = første desimal er lik 1
Begivenhet B = første desimal er 6 eller høyere

Fra tabellen over Bensford's lov blir

$$P(A) = P(\{1\}) = 0.301$$

$$P(A^c) = 1 - P(A) = 1 - 0.301 = 0.699$$

$$P(B) = P(\{6\}) + P(\{7\}) + P(\{8\}) + P(\{9\})$$

$$= 0.067 + 0.058 + 0.0501 + 0.046 = 0.222$$

$$P(\text{A eller B}) = P(A) + P(B) = 0.301 + 0.222 = 0.523$$

Tilordning av sannsynlighet når alle utfall like sannsynlige.

Vi så at i en følge av tilfeldige tall har alle sifre like stor sjanse for å forekomme på en plass. Hvis vi begrenser oss til tallene 1,2,...,9, blir S={1,2,3,4,5,6,7,8,9}

$$P(\{1\})=....P(\{9\})=1/9.$$

En falskner som bruker en tabell tilfeldige tall for rapportering vil derfor ha sannsynlighet $3/9= 1/9+1/9+1/9 = 0.333$ for å bruke et av tallene 1,2,3 som første siffer.

I følge Benfords lov er

$$P(1 \text{ eller } 2 \text{ eller } 3) = 0.301+ 0176+0.125= 0.602$$

for legale innrapporteringer.

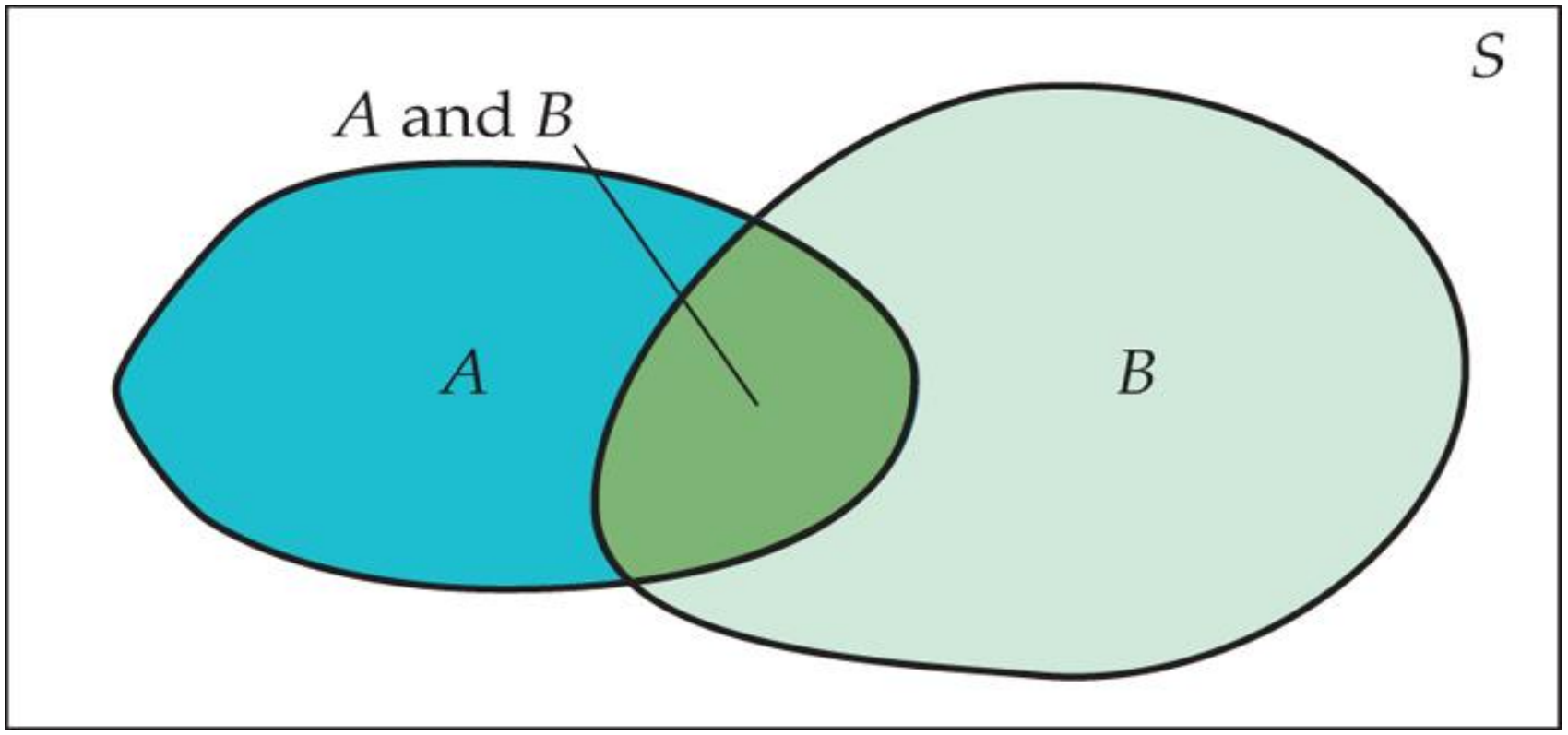
I sannsynlighetsrom der alle utfall har samme sannsynlighet gjelder at sannsynligheten for en begivenhet er

«gunstige»/«mulige».

Her er «gunstige» antall utfall som svarer til at begivenheten inntreffer, og «mulige» alle mulige utfall, dvs. antall elementer i utfallsrommet.

Uavhengighet og multiplikasjonsregel

Skal betrakte situasjonen der de to begivenhetene A og B **ikke** lenger er disjunkte.



Eksempel: Anta myntkast uavhengige, rettferdig mynt

A=første myntkast er kron
B=andre myntkast er kron

P(A)=P(B)=1/2
A og B er ikke disjunkte

Mulige utfall for to kast: KK,KM, MK,MM

Alle utfall like sannsynlige:

P(A og B) = P(KK) = 1/4= (1/2)×(1/2)= P(A)×P(B).

Dette er et eksempel på at at A og B er uavhengige

To begivenheter A og B er **uavhengige**
hvis kunnskap om at den ene inntreffer ikke endrer på sannsynligheten for at den andre inntreffer.

Hvis A og B er uavhengige, så er

$P(A \text{ og } B)=P(A)\,P(B)$

Dette er **produktregelen for uavhengige begivenheter**.

Eksempel: Myntkast er som vist uavhengige.

Eksempel: Trekking av kort – uten tilbakelegging.
Her er begivenhetene ikke uavhengige.

P(Første kort rødt)= 26/52

P(Andre kort rød hvis første kort rødt)=25/51

P(Andre kort rød hvis første kort sort)=26/51

Å vite resultat av første trekking endrer sannsynligheten for resultat av andre trekking.

Eksempel: Mendels lov

To farger på erter: (G=green, Y=yellow).

To planter krysset for å produsere nye. Et gen fra hver.

Plantene har to gener hver som bestemmer farge (GG, GY, YY)
Y er dominant, dvs. ett Y-gen gir gul farve:
GG →G, GY → Y, YY → Y.

Mendels lov: Arver gener uavhengig fra de to foreldre og hvert gen er like sannsynlig.

Anta begge foreldre-plantene har GY.

P(ny plante G) = P(G fra første og G fra andre)
= P(G fra første) P(G fra andre) =(1/2)(1/2)= 1/4.

Eksempel Krybbedød.

I familer der ingen røyker er raten for krybbedød 1 av 8500. Hadde krybbedød vært uavhengige Begivenheter, skulle sannsynligheten for å miste to barn i krybbedød være:

$$(1/8500)(1/8500) = 1/72{,}250{,}000$$

Men dette tallet er alt for lavt. Da ser man bort fra at miljø- og genetiske faktorer kan ha betydning, slik at begivenhetene er avhengige.

A og B disjunkte:

P(A eller B) = P(A) + P(B)

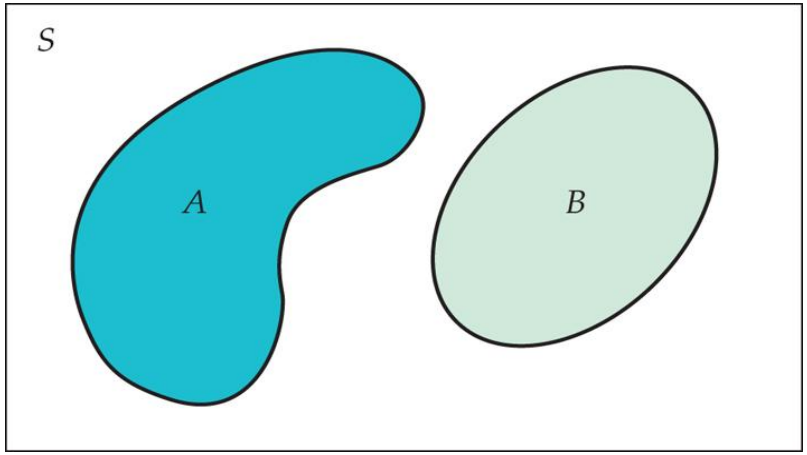
A og B uavhengige:

P(A og B) = P(A) P(B)

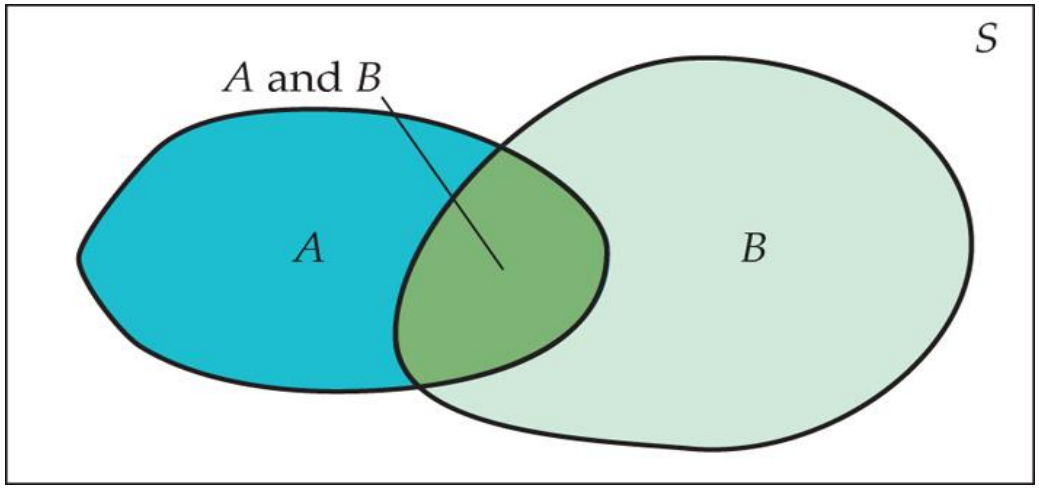
Merk: Hvis A og B er disjunkte kan de ikke være uavhengige.

Dette fordi da må P(A og B) = 0.
Men P(A) P(B) > 0 når både P(A) > 0 og P(B) > 0 .

Disjunkte begivenheter A og B, så P(A og B) = 0



Begivenhetene A og B **ikke** er disjunkte - og rimeligvis P(A og B) > 0



Eksempel : Falske positive

Anta P(falsk positiv i en prøve) =0.02
samt at 50 prøver av samme negative person er uavhengige.

La A =”En eller flere positive blant de 50 prøvene”

Da blir

$$P(A) = 1 - P(A^c) = 1 - P(\text{"Ingen falske positive i 50 prøver"})$$

$$= 1 - P(\text{"50 negative prøver"})$$

$$= 1 - (1-0.02)^{50}$$

$$= 1 - 0.3642 = 0.6358$$

Tilfeldige variable

Som vi har sett behøver ikke utfallsrom bestå av tall.
Men ofte er det numeriske størrelser som er av interesse, f. eks antall kron i fire myntkast.

En **tilfeldig variabel** er en variabel hvis verdi er den numeriske verdien (resultatet) av en tilfeldig prosess.

Omvendt er utfallsrommet S alle mulige verdier den tilfeldige variabelen kan ta.

Notasjon: Tilfeldige variable betegnes vanligvis med store bokstaver som X og Y.

De tilsvarende små, x og y, betegner de faktiske utfallene.

Unntak: $\bar{x}$ for gjennomsnitt
s for standardavvik
$\hat{p}$ for andel

For disse unntakene brukes de to betydningene om hverandre.

Fordelingen til en **diskret variabel** X beskrives ved en liste over de mulige verdiene X kan anta samt en liste over sannsynlighetene for at X antar disse verdiene

Verdier for X: $x_1,\ x_2,\ x_3,\ .....$

Sannsynligheter: $p_1,\ p_2,\ p_3,\ .....$

Krav

1: $0 \leq p_i \leq 1$

2: $p_1 + p_2 + p_3 + ... = 1.$

Eksempel: Karakterfordeling

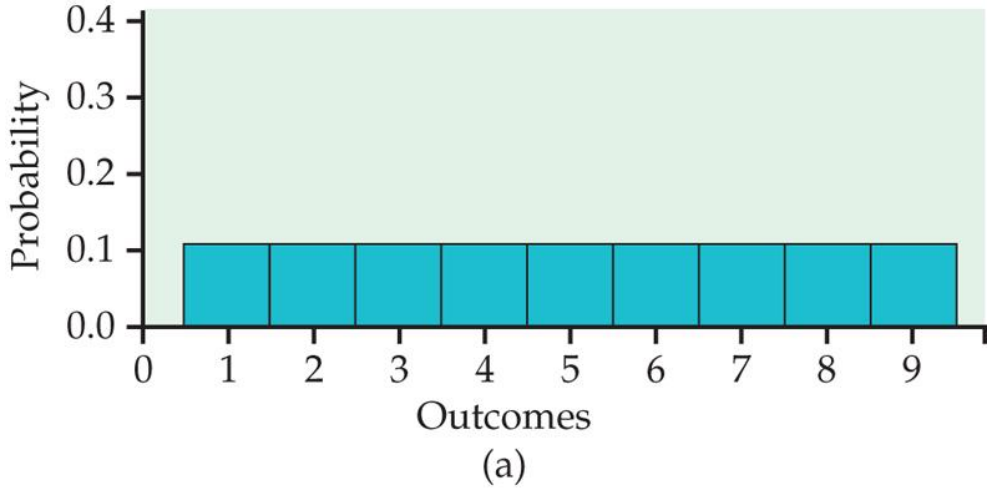
0: F
1: D
2: C
3: B
4: A

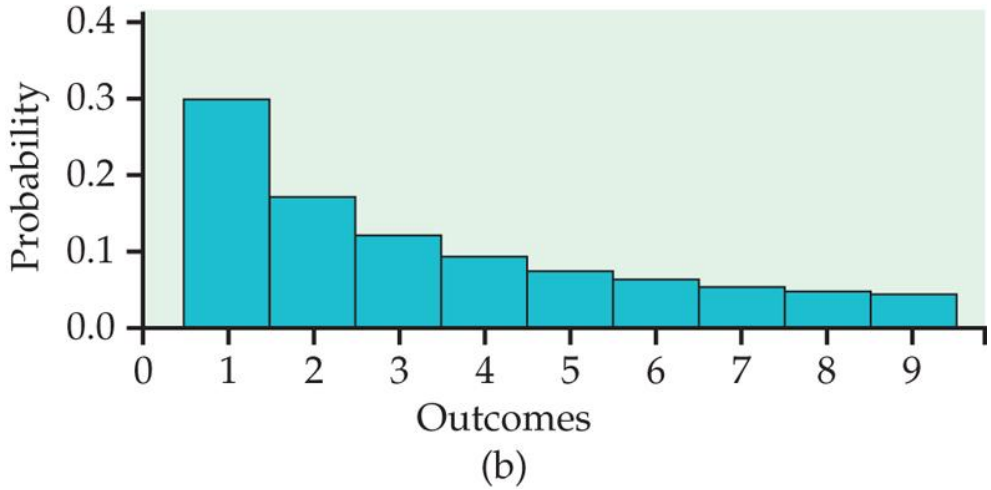
Verdi for X:	0	1	2	3	4
Sannsynlighet :	0.05	0.04	0.20	0.40	0.31

Derfor: $P(\text{A eller B}) = P(X \geq 3) = P(X=3) + P(X=4)$

$= 0.40 + 0.31 = 0.71$

Det kan være nyttig å fremstille fordelingene grafisk med histogrammer, her alle 1,2,...,9 like sannsynlige og Benfords lov





Benford's lov:

Første desimal :	1	2	3	4	5
Sannsynlighet :	0.301	0.176	0.125	0.097	0.079

Første desimal :	6	7	8	9
Sannsynlighet:	0.067	0.058	0.051	0.046

Eksempel: Antall kron i fire myntkast.

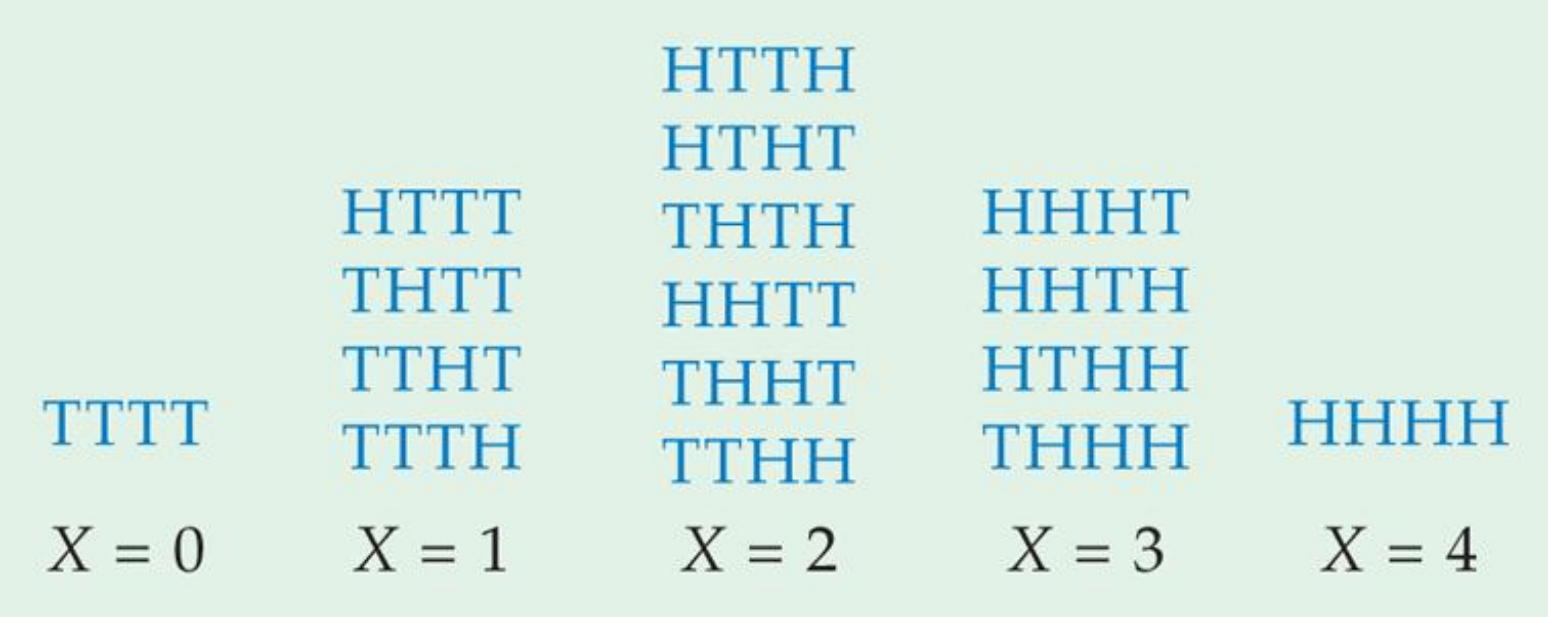
Antar 1: I et kast like sannsynlig med kron som mynt
2: Kastene uavhengige

La X være antall kron i fire kast.

Resultatene av fire kast er en følge av fire K eller M som i KKMK. Det er i alt 16 slike følger og alle er like sannsynlige. Multiplikasjonsregelen gir for en følge $(1/2)\times(1/2)\times(1/2)\times(1/2) = 1/16 = 0.0625$.

Figuren nedenfor angir hvilke av de 16 utfallene som svarer til de mulige verdiene av X.

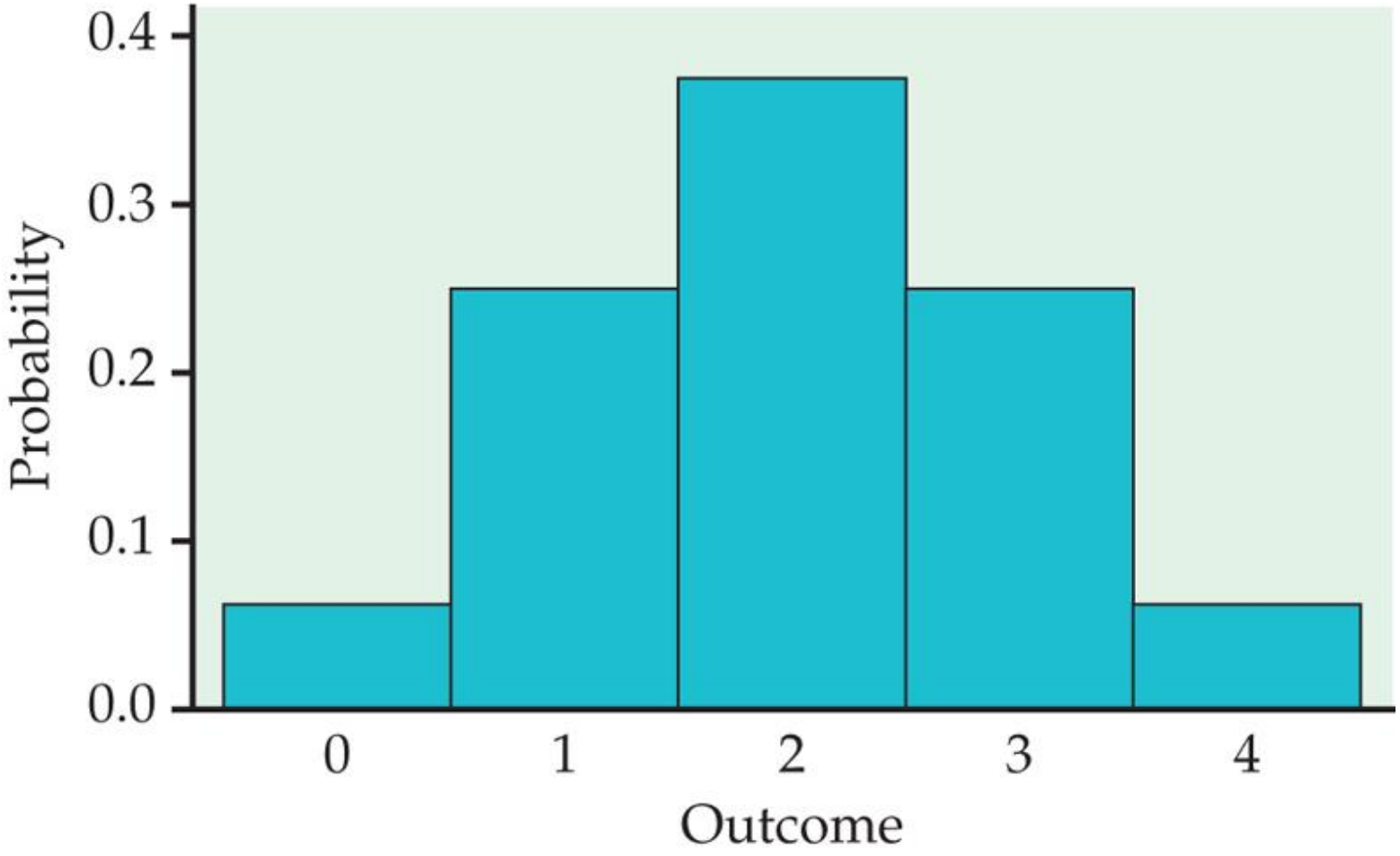
Figuren nedenfor angir hvilke av de 16 utfallene som svarer til de mulige verdiene av X.



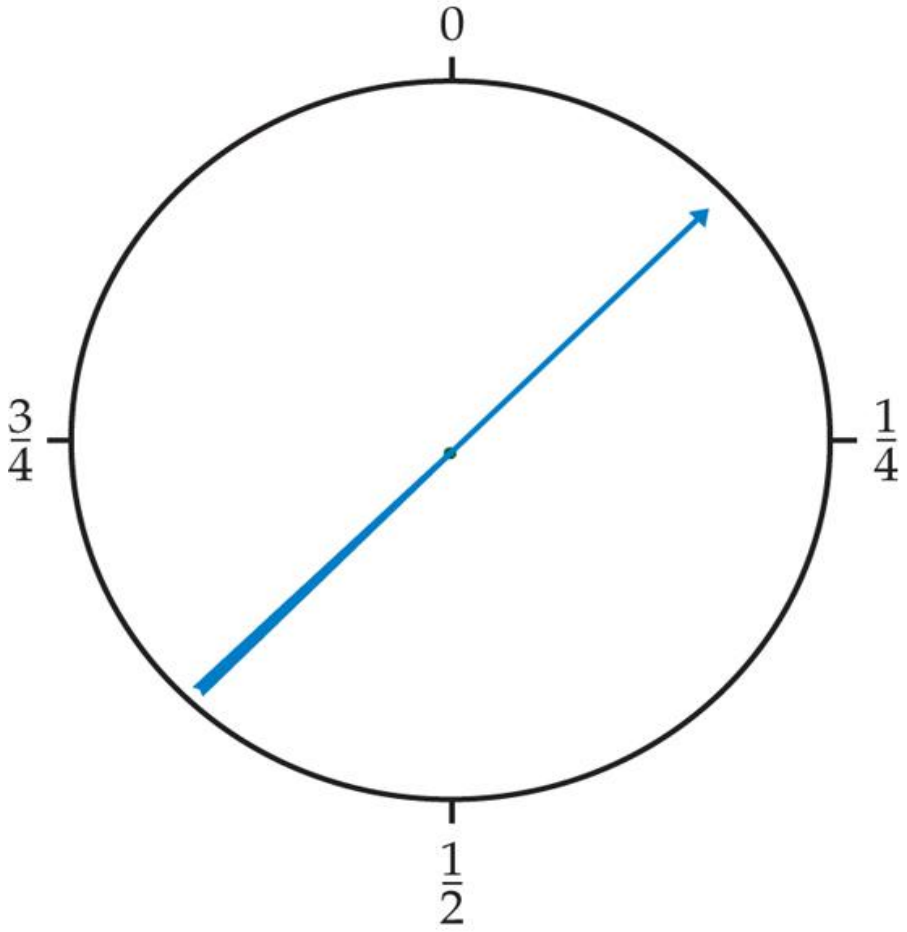
Dette gir fordelingen:

Verdi for X:	0	1	2	3	4
Sannsynlighet :	0.0625	0.25	0.375	0.25	0.0625

og tilhørende histogram



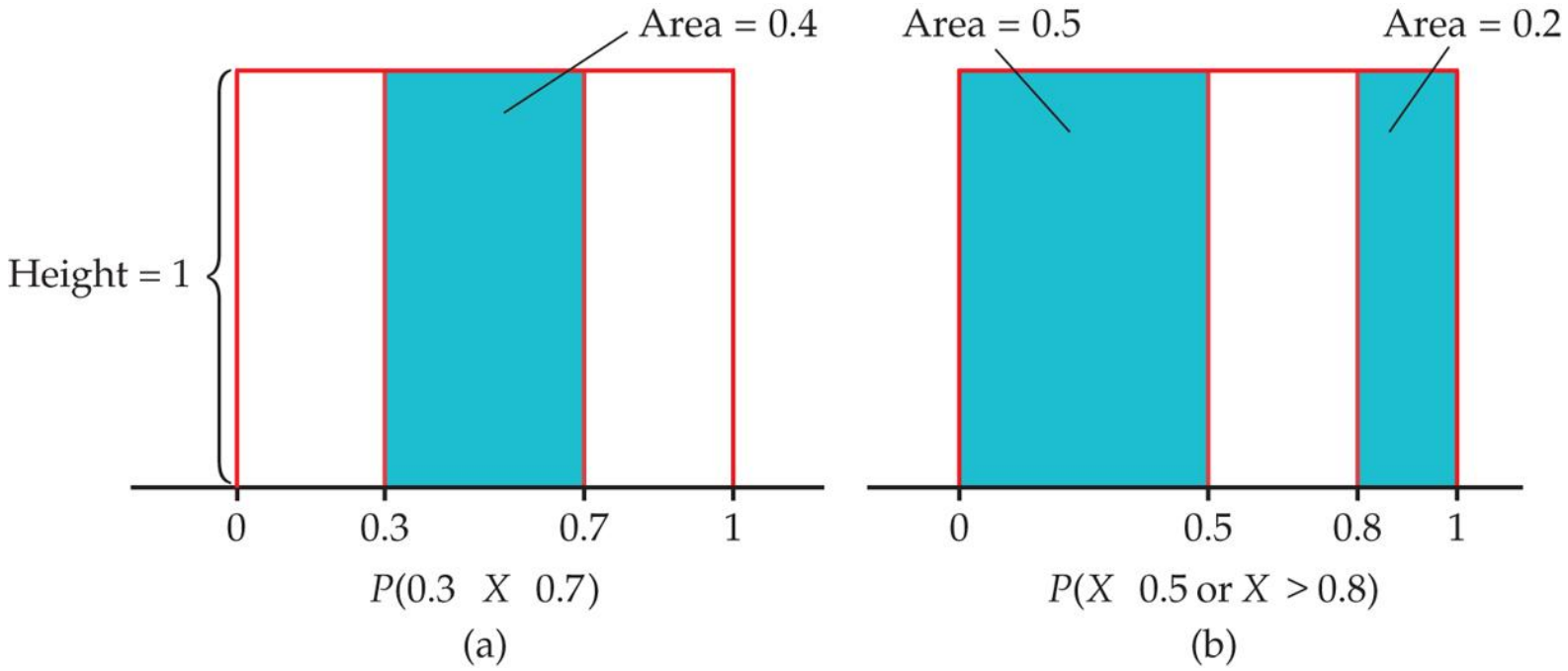
Den andre hovedtypen tilfeldige variabler er **kontinuerlige tilfeldige** variable. Et eksempel vil være å velge et tilfeldig tall mellom 0 og 1. Det kan tenkes som punktet nålen peker på.



Nålen kan stoppe hvor som helst mellom 0 og 1. Det er uendelig mange verdier mellom 0 og 1, så strategien med å lage en liste over de tilhørende sannsynlighetene er ikke mulig. I stedet ser en på begivenheter av typen

$$A = \{ 0.3 \leq x \leq 0.7 \},$$

og lar begivenhetene svare til områder under en tetthetskurve.



Her ser man at

P(0.3 ≤ X ≤ 0.7) = 0.4

P(X ≤ 0.5) = 0.5

P(0.8 > X) = 0.2

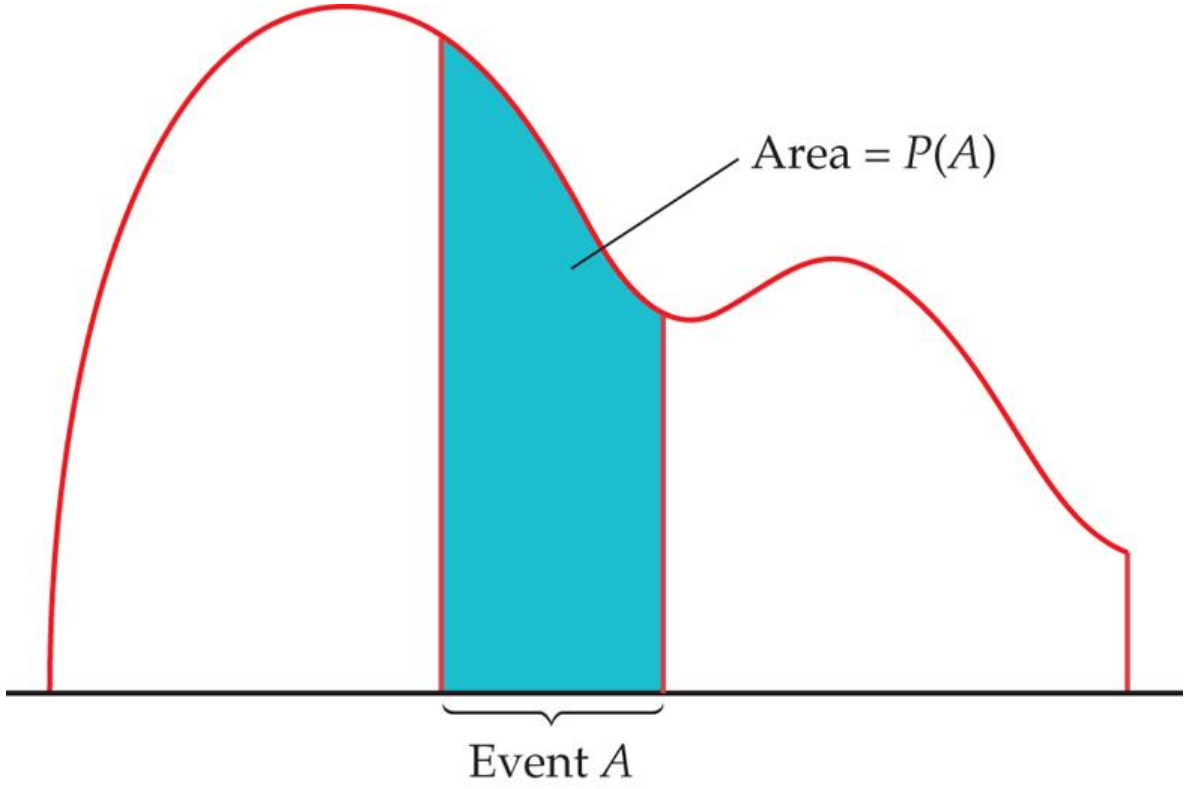
P(X ≤ 0.5 eller X>0.8) = 0.7

Husk at tetthetskurver var definert som kurver som alltid et ikke-negative og som har areal under kurven som er lik 1, side 54 i læreboka.

Dette gir stor frihet i å tilordne sannsynligheter som areal under en tettkhetskurve.

En **kontinuerlig tilfeldig variabel** X kan anta alle verdier i et intervall. **Sannsynlighetsfordelingen** til X beskrives ved hjelp av en tetthetskurve. Sannsynligheten til en begivenhet er arealet under kurven i det intervallet som definerer begivenheten.

Merk Sannsynligheten for at en kontinuerlig tilfeldig variabel skal anta en bestemt verdi er 0. Det må være slik siden et vilkårlig lite intervall som inneholder verdien har villkårlig liten sannsynlighet. Dermed blir f.eks. $P(X \leq 0.8) = P(X < 0.8)$.


Area = $P(A)$
Event A

Vi har sett at hvordan man får regnet ut arealer under en normalfordelingskurve, N(μ,σ).

Nå vil vi oppfatte disse arealene som sannsynligheter for at en normalfordelt tilfeldig variabel X ligger innenfor det spesifiserte området.

Spesielt gjelder at hvis X er N(μ,σ) så er blir den standardiserte variablen

$$Z = \frac{X-\mu}{\sigma}$$

standardnormalfordelt, Z er N(0,1).

Man kan vise at hvis andelen i en populasjon med et kjennetegn er p=0.26 og vi trekker et utvalg på n=500 så vil den observerte andelen i utvalget, $\hat{p}$, ha en tilnærmet normalfordeling angitt som N(0.26 , 0.0196).

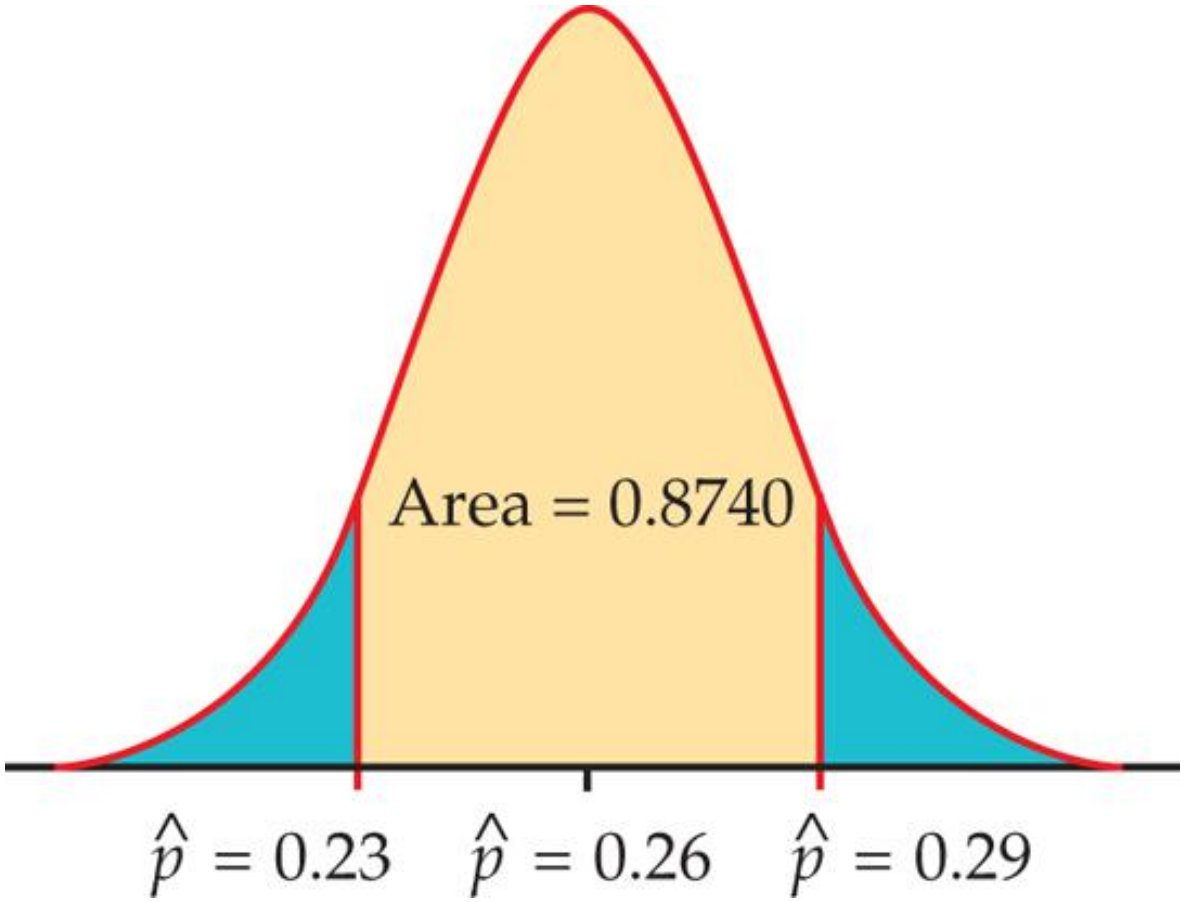
Da kan vi beregne sannsynligheten for at forskjellen mellom parameteren p og andelen i utvalget er mindre enn 3%.

$$\mathrm{P}(|\hat{p} - p| \leq 0.03) = \mathrm{P}(|\hat{p} - 0.26| \leq 0.03) = \mathrm{P}(0.23 \leq \hat{p} \leq 0.29)$$

$$= \mathrm{P}\left(\frac{0.23-0.26}{0.0196} \leq \frac{\hat{p}-0.26}{0.0196} \leq \frac{0.29-0.26}{0.0196}\right)$$

$$= \mathrm{P}(-1.53 \leq Z \leq 1.53) = 0.9379 - 0.0630 = 0.8740$$

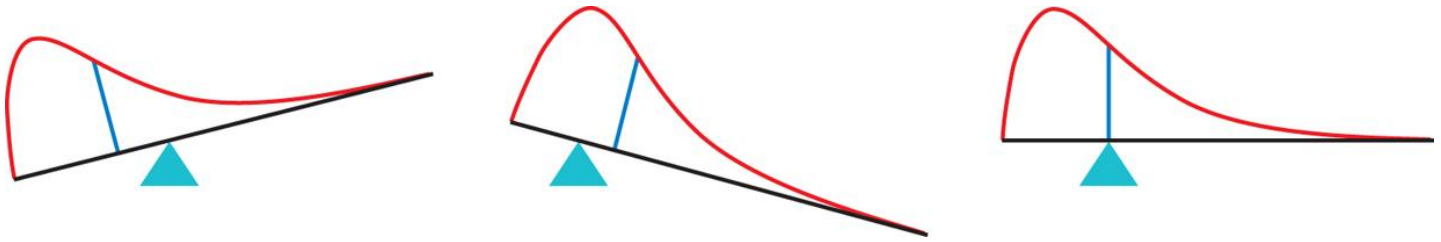
hvilket betyr at hvis vi hadde gjenntatt trekningen mange ganger ville vi havnet i en avstand fra p mindre enn 3% i så mye som 87% av gangene.


Area = 0.8740
$\hat{p} = 0.23$
$\hat{p} = 0.26$
$\hat{p} = 0.29$

Forventning og varians til tilfeldige variable

Vi har tidligere sett på mål for spredning i empiriske fordelinger, dvs. for gitte datasett. Nå skal vi gjøre det samme for fordelingene definert ved tilfeldige variable. Vi kommer til å konsentrere oss om de størrelsene som svarer til (empirisk) gjennomsnitt og standardavvik.

Vi har også definert forventninger som ”tyngdepunktet” i tetthetskurver. For diskrete tilfeldige variable skal vi gi en formell definisjon.



Eksempel: Lotteri, tall mellom 0 og 999 gitt.

Tipp tall mellom 000 og 999.

Gevinst: 5000 kr hvis man gjetter på det gitte tallet.
Sannsynlighet for å vinne: g/m=1/1000

Vinner 0 kr. 999 av 1000 ganger
Vinner 5000 kr. 1 av 1000 ganger

Veid sammen: (0)× (999/1000)+(5000)×(1/1000)= 5kr.
Dette er et eksempel på forventning.

Her kan vi oppfatte den som gevinsten i det lange løp,
dvs hva man vinner i gjennomsnitt hvis man spiller mange ganger.

Notasjon:

Tilfedig variabel X

Forventning: μ eller μ_x

Standardavvik: σ eller σ_x

Forventning for en diskret variabel **X**

Verdier for X: $x_1,\ x_2,\ x_3,\ \ldots\ldots,\ x_k$

Sannsynligheter: $p_1,\ p_2,\ p_3,\ \ldots\ldots,\ p_k$

Da er forventningen for **X**

$$\mu = \mu_x = x_1\, p_1 + x_2\, p_2 + x_3\, p_3 + \ldots. + x_k\, p_k$$

$$= \sum x_i\, p_i$$

Eksempel : Lotteri

Verdi av X:	0	5000
Sannsynlighet:	0.999	0.001

μ_X = (0)×(0.999) + (5000)×(0.001)= 5

Eksempel: like sannsynlige siffre

Første siffer:	1	2	3	4	5	6	7	8	9
Sannsynlighet:	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9

Forventning:

μ_x= 1×(1/9) + 2×(1/9)+ 3×(1/9)+,,,,+ 9×(1/9) = 45×(1/9)=5

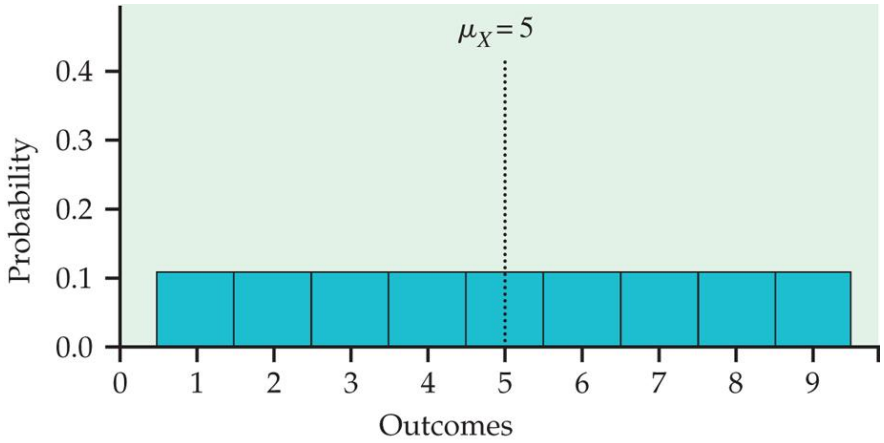
Eksempel : Benfords lov

Første desimal :	1	2	3	4	5
Sannsynlighet :	0.301	0.176	0.125	0.097	0.079

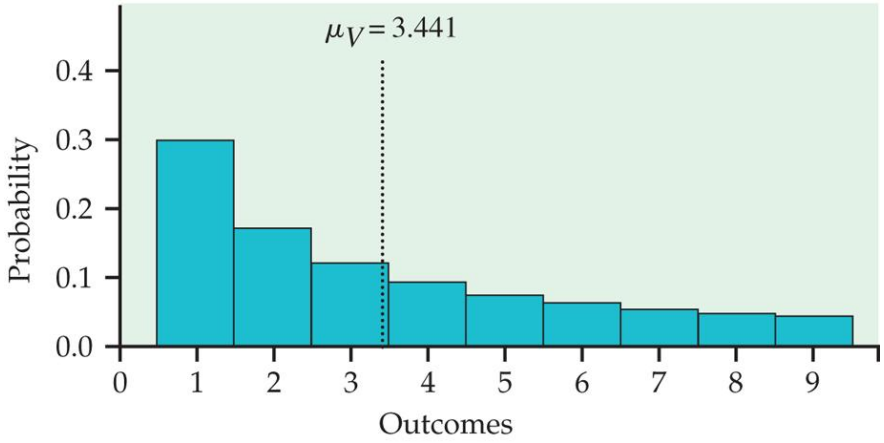
Første desimal :	6	7	8	9
Sannsynlighet:	0.067	0.058	0.051	0.046

Forventning:

μ_x= (1)×(0.301) + (2)×(0.176)+ 3×(0.125)+,,,,+ (9)×(0.046) = 3.441


$\mu_X = 5$
Probability
0.4
0.3
0.2
0.1
0.0
0 1 2 3 4 5 6 7 8 9
Outcomes

(a)


$\mu_V = 3.441$
Probability
0.4
0.3
0.2
0.1
0.0
0 1 2 3 4 5 6 7 8 9
Outcomes

(b)

For kontinuerlige tilfeldige variable defineres forventningen ved hjelp av tetthetskurven. Den kan beskrives som et balansepunkt eller tyngdepunkt.

Mer matematisk kan den beskrives ved et integral: Hvis X har tettheten *f(x)* blir forventningen

$$\mu = \mu_x = \int x\, f(x)\, dx$$

Merk: Forventning og varians brukes til å beskrive utvalgsfordelingen. Den er definert ved populasjonen. Da er forventningen og variansen også det, dvs at både forventning og varians er parametre. Herav de greske bokstavene.

Store talls lov

Fordeling: ``Idealisert relative frekvenser''

Forventing: ``Idealisert gjennomsnitt''

``Idealisert'' består i at resultatet forstås
som basert på et stort antall uavhengige forsøk.

Store talls lov ufomelt:
Trekk uavhengige observasjoner fra en populasjon med forventning μ_X .
Da vil gjennomsnittet: $\bar{x}$= (X_1 + X_2 + …. X_n)/n
nærme seg μ_X når antallet gjentak , n , øker.

Store talls lov formelt:

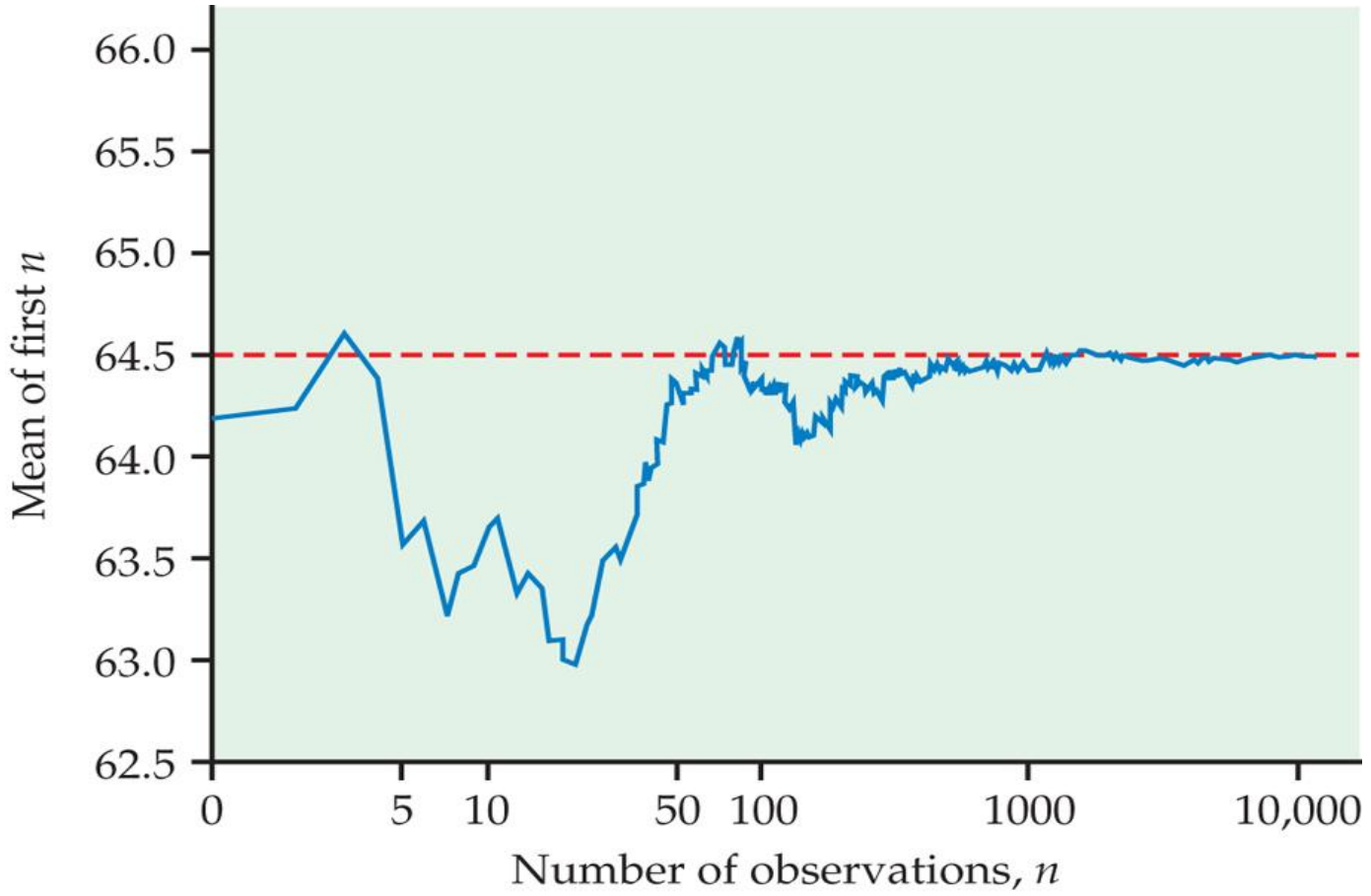
For alle tall $\varepsilon > 0$ vil

$P(| (X_1 + X_2 + \ldots . X_n)/n - \mu_X | > \varepsilon) \rightarrow 0$ når n vokser

Dette betyr at uansett hvor lite et intervall rundt μ_X er, vil sannsynligheten for at gjennomsnittet ligger utenfor intervallet nærme seg null når n vokser.

Store talls lov svarer til måten vi introduserte sannsynlighet. Ved mange gjentak vil andelen utfall som svarer til en verdi nærme seg sannsynligheten til den verdien. På samme måte vil gjennomsnittet av utfallene nærme seg forventningen i fordelingen.

Vi har sett at fordelingen as høyden til unge kvinner er tilnærmet normal N(64.5, 2.5).



Hvordan forstå store talls lov?

Store talls lov sier noe om resultatet ``i det lange løp''.

Vanskelig å skille tilfeldig og systematisk innflytelse. Trenger mer systematisk tilnærming.

Gambling bygger på at oppfatning av usikkerhet ulik hos spiller og casino. Casinoer utnytter store talls lov, det samme gjør forsikringsselskaper.

Regler for forventning

Eksempel

(i) Forventet antall småbulker i kjøleskap: 0.7
Forventet antall blemmer i malingen: 1.4
Forventet antall uregelmessigheter: 2.1

(ii) Forventet lengde av gresshoppe: 1.2 tomme
En tomme = 2.54 cm.
Forventet lengde av gresshoppe: 1.2 x 2.54 = 3.05 cm.

Regel 1: Hvis X er en tilfeldig variabel og a og b er tall er

$$\mu_{a+bX} = a + b\,\mu_X$$

Regel 2: Hvis X og Y er tilfeldige variable, er

$$\mu_{X+Y} = \mu_X + \mu_Y$$

Regel 3: Hvis X og Y er tilfeldige variable, er

$$\mu_{X-Y} = \mu_X - \mu_Y$$

Eksempel: Kurskreditt

Fordeling for antall kurs, X høst, Y vår, Z=X+Y hele året

Antall kurs om høsten:	1	2	3	4	5	6
Sannsynlighet:	0.05	0.05	0.13	0.26	0.36	0.15

Antall kurs om våren:	1	2	3	4	5	6
Sannsynlighet:	0.06	0.08	0.15	0.25	0.34	0.12

μ_X = (1)(0.05)+ (2)(0.05)+(3)(0.13)+ (4)(0.26)+(5)(0.36)+(6)(0.15)=4.28

μ_Y = (1)(0.06)+ (2)(0.08)+(3)(0.15)+ (4)(0.25)+(5)(0.34)+(6)(0.12)=4.09

$\mu_Z = \mu_X + \mu_Y = 4.28 + 4.09 = 8.37$

Eksempel: Kreditt timer

Anta hvert kurs gir 3 time kreditt. La T være antall timer.

Da er T=3 x Z= 3 x (X+Y) og forventningen

$\mu_T = \mu_{3Z} = 3\ \mu_Z = (3)(8.37) = 25.11$

Variansen til en diskret tilfeldig variabel

Forventningen angir senteret i fordelingen til en tilfeldig variabel. Den andre viktige størrelsen vi trenger et mål for er spredningen. For en tilfeldig variabel X er variansen en veiet gjennomsnittsverdi av kvadratet av aviket av X fra μ_X, dvs $(X - \mu_X)^2$.

Variansen betegnes med $\sigma_X{}^2$. Standardavviket er $\sigma_X = \sqrt{\sigma_X{}^2}$.

Variansen til en diskret tilfeldig variabel

Anta at fordelingen til den tilfeldige variabelen X er gitt ved

Verdi av X :	$x_1 , x_2 , \ldots . \quad x_k$
Sannsynlighet	$p_1 , p_2 , \ldots . \quad p_k$

Med forventning μ_x.

Da er variansen …

$$\sigma_X{}^2 = (x_1 - \mu_x)^2 p_1 + (x_2 - \mu_x)^2 p_2 + \ldots + (x_k - \mu_x)^2 p_k$$

$$= \sum (x_i - \mu_X)^2 p_i$$

Eksempel: Kurskreditt

Fordeling for antall kurs, X høst,

Antall kurs om høsten:	1	2	3	4	5	6
Sannsynlighet:	0.05	0.05	0.13	0.26	0.36	0.15

x_i	p_i	$x_i\, p_i$	$(x_i - \mu_x)^2\, p_i$
1	0.05	0.05	$(1-4.28)^2\,(0.05)=$ 0.53297
2	0.05	0.10	$(2-4.28)^2\,(0.05)=$ 0.25992
3	0.13	0.39	$(3-4.28)^2\,(0.13)=$ 0.21299
4	0.26	1.04	$(4-4.28)^2\,(0.26)=$ 0.02038
5	0.36	1.80	$(5-4.28)^2\,(0.36)=$ 0.18662
6	0.15	0.90	$(6-4.28)^2\,(0.15)=$ 0.44376
		μ_X =4.28	$\sigma_X{}^2$= 1.662

og standardavvik $\sigma_X = \sqrt{1.662}$ = 1.289.

Regler for varians

Andre regler for varians/standardavvik enn for forventing:

Eksempel:
X: Andel av studielån til mat og sted å bo
Y: Andel av studielån til andre ting
X+Y =100

Konstant har varians/standardavvik lik 0.

X og Y kan variere

Nødvendig å ta hensyn til **korrelasjonen**
mellom X og Y.

Empirisk korrelasjon: r

Korrelasjon mellom to tilfeldige variable: ρ

Begge er mål for samvariasjon og har samme egenskaper:

1. $-1 \leq \rho \leq 1$

2. Måler styrke av lineær sammenheng

3. Hvis variablene er uavhengige, er $\rho=0$.

Regel 1: Hvis **X** er en tilfeldig variabel og Z = a + b X
for gitte verdier a og b så er variansen til Z

$$\sigma_Z^2 = b^2 \, \sigma_X^2$$

Merk at varians/standardavvik ikke endres hvis det legges til en konstant a.
Men når X multipliseres med en konstant b så vil variansen endres med b^2, mens standardavviket endres med absoluttverdien |b|

Regel 2a: Hvis X og Y er **uavhengige** tilfeldige variable
og Z = X + Y så er variansen til Z

$$\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2$$

Regel 2b: Hvis **X** og Y er **uavhengige** tilfeldige variable
og Z = X - Y så er variansen til Z (fortsatt)

$$\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2$$

Regel 3a: Hvis X og Y er **avhengige** tilfeldige variable med korrelasjon ρ og Z = X + Y så er variansen til Z

$$\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2 + 2\,\rho\,\sigma_X\,\sigma_Y$$

Regel 3b: Hvis X og Y er **avhengige** tilfeldige variable med korrelasjon ρ og Z = X - Y så blir variansen til Z

$$\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2 - 2\,\rho\,\sigma_X\,\sigma_Y$$

Eksempel : Lotteri

Verdi av X:	0	5000
Sannsynlighet:	0.999	0.001

μ_X = (0)(0.999) + (5000)(0.001)= 5

$\sigma_X{}^2$= $(0-5)^2$ 0.999 + $(5000-5)^2$ 0.001= 24975.0

σ_X = 157.96

Hvis loddet koster 10 kr, er forvente utlegg W=X-10
μ_W = μ_X - 10 = -5, men $\sigma_X = \sigma_W$ ut fra den første regelen .

For to lodd er forventet gevinst μ_{X+Y} = μ_X + μ_Y = 10,
mens $\sigma_{X+Y}{}^2$ =$\sigma_X{}^2$ + $\sigma_Y{}^2$ = 24975.0 +24975.0= 49950.0
og σ_{X+Y} = 223.50.

Eksemplene på forrige side var basert på at X og Y var uavhengige. Ofte er ikke dette tilfelle. Da må vi ta hensyn til korrelasjonen ρ.

Eksempel: SAT-scores
SAT matematikk score, $\mu_X = 625$, $\sigma_X = 90$
SAT verbal score, $\mu_Y = 590$, $\sigma_Y = 100$

Da blir $\mu_{X+Y} = \mu_x + \mu_Y = 625 + 590 = 1215$

Men $\sigma^2_{X+Y} \neq \sigma^2_X + \sigma^2_Y$ når X og Y er avhengige.

Anta at korrelasjonen $\rho = 0.70$. Da blir

$$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y + 2\,\rho\,\sigma_X\,\sigma_Y = 90^2 + 100^2 + 2 \cdot 0.70 \cdot 90 \cdot 100 = 30700$$

mens $\sigma_{X+Y} = \sqrt{30700} = 175.2$

Generelle regler for sannsynlighet

Vi tar nå et skritt tilbake til sannsynlighetsregningen og setter denne i en litt mer generell ramme.

Spesiel antar vi som i starten at utfallene ikke nødvendigvis svarer til at tilfeldige varaibler antar spesielle numeriske verdier.

Som tidligere lar vi S betegne utfallsrommet og
A og B begivenheter,

La P(A) og P(B) være sannsynlighetene for begivenhetene A og B

Regel 1: $0 \leq P(A) \leq 1$ for alle A, tillatte verdier
Regel 2: P(S)=1, total sannsynlighet
Regel 3: $P(A^c)$=1-P(A) (Komplementær-regel)
Regel 4: P(A eller B)=P(A)+P(B) hvis A og B er disjunkte begivenheter, addisjons regel
Regel 5: Hvis A og B er uavhengige, er P(A og B)=P(A) ×P(B), multiplikasjonsregel.

Som tidligere lar vi S betegne utfallsrommet og
A og B begivenheter,

La P(A) og P(B) være sannsynlighetene for begivenhetene A
og B

Spørsmål:

Hva er P(A eller B) hvis A og B **ikke** er **disjunkte** begivenheter?

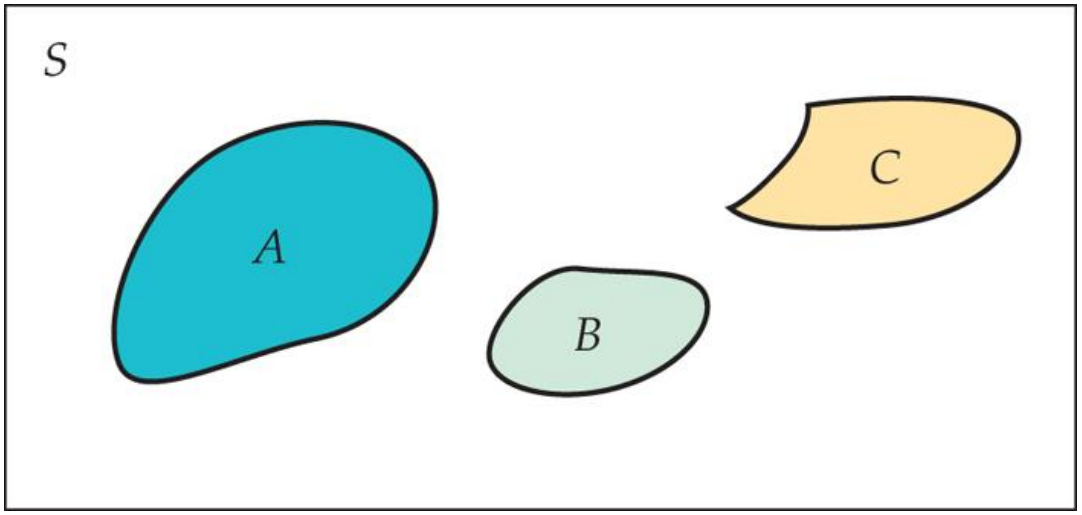
Hva er P(A eller B eller C) hvis A, B og C er disjunkte begivenheter?

Unionen til en samling eller familie begivenheter
er begivenheten at minst en av begivenhetene inntreffer.

Addisjonsregel for disjunkte begivenheter: Hvis A, B og C er disjunkte begivenheter er

$$P(\text{en eller flere av } A,B,C) = P(A)+P(B)+P(C).$$

Dette kan generaliseres til flere begivenheter

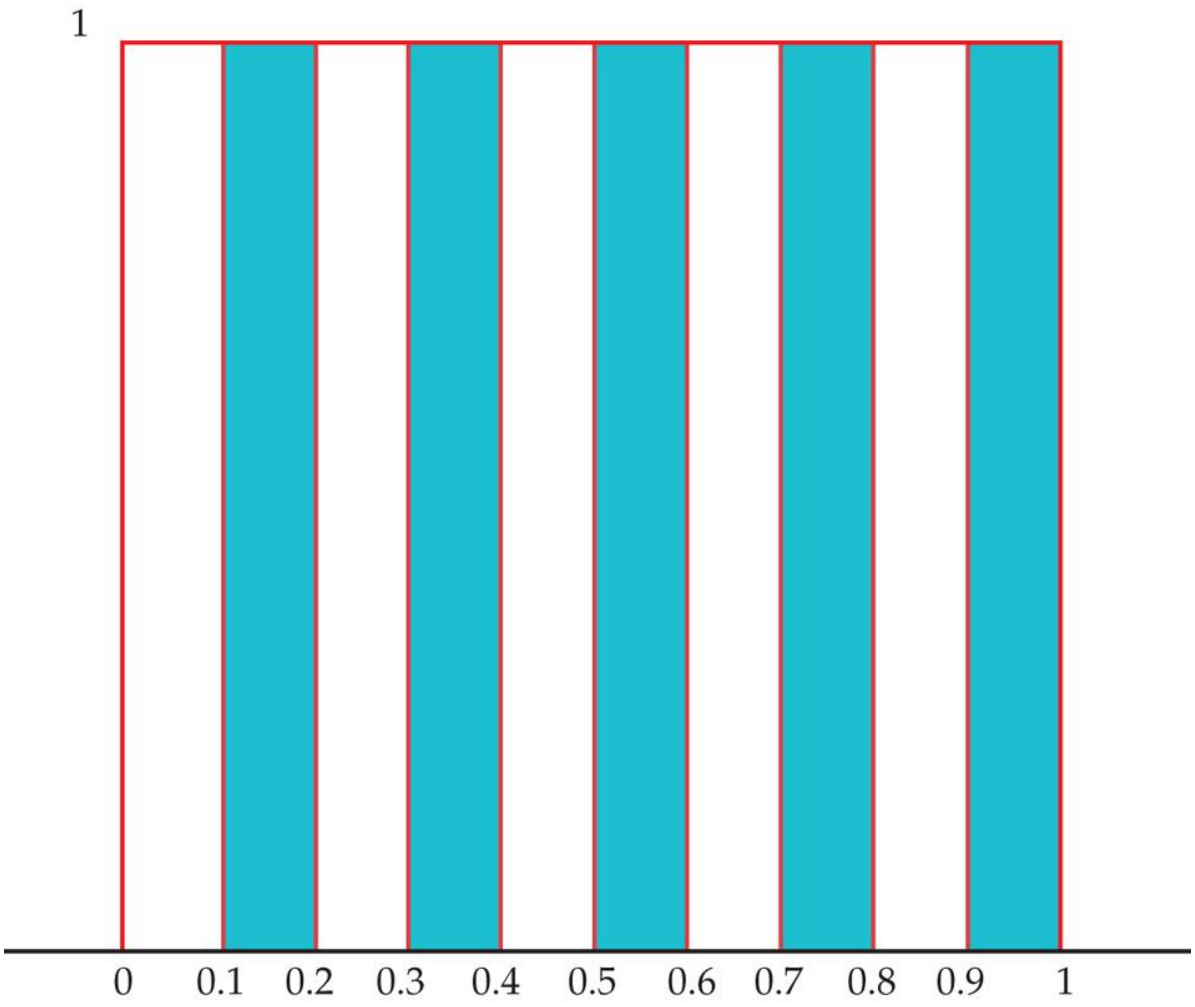


Eksempel: Hva er sannsynligheten for at første siffer i et tilfeldig tall mellom 0 og 1 er odde?

La X være et tilfeldig tall mellom 0 og 1. Da svare «første siffer i X odde» til unionen av de disjunkte begivenhetene

$$0.1 \leq X < 0.2$$
$$0.3 \leq X < 0.4$$
$$0.5 \leq X < 0.6$$
$$0.7 \leq X < 0.8$$
$$0.9 \leq X < 1.0.$$

Hver av disse har sannsynlighet 0.1, slik at ikke overraskende er sannsynligheten for at første siffer er odde 0.5.

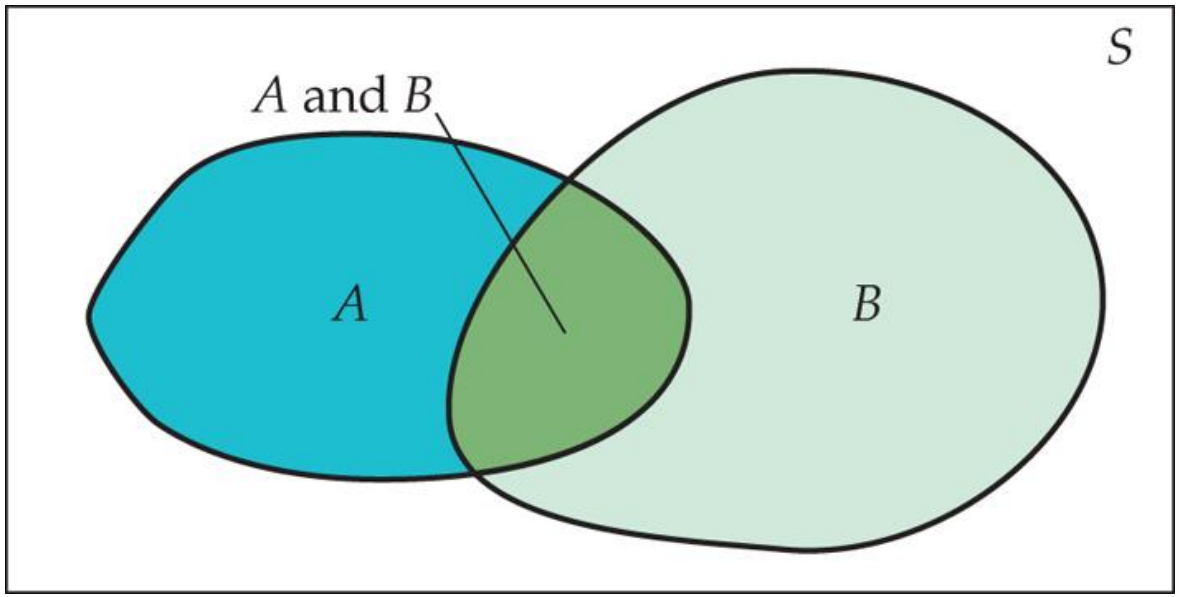

1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Hvis begigivenhetene A og B **ikke** er disjunkte er det utfall som forekommer i begge. Legger man sammen sannsynlighetene vil disse utfallene telle dobbelt.

Det må korrigeres for og det gjøres i

Addisjonsregel for union:

P(A eller B) = P(A)+P(B)-P(A og B).



Merk Hvis A og B er disjunkte er siste ledd i summen lik null, så da er vi tilbake til situasjonen vi allerede har sett.

Eksempel: Anta 40% voksne får nok søvn, 46% trener jevnlig og at 24% gjør begge deler. Da er

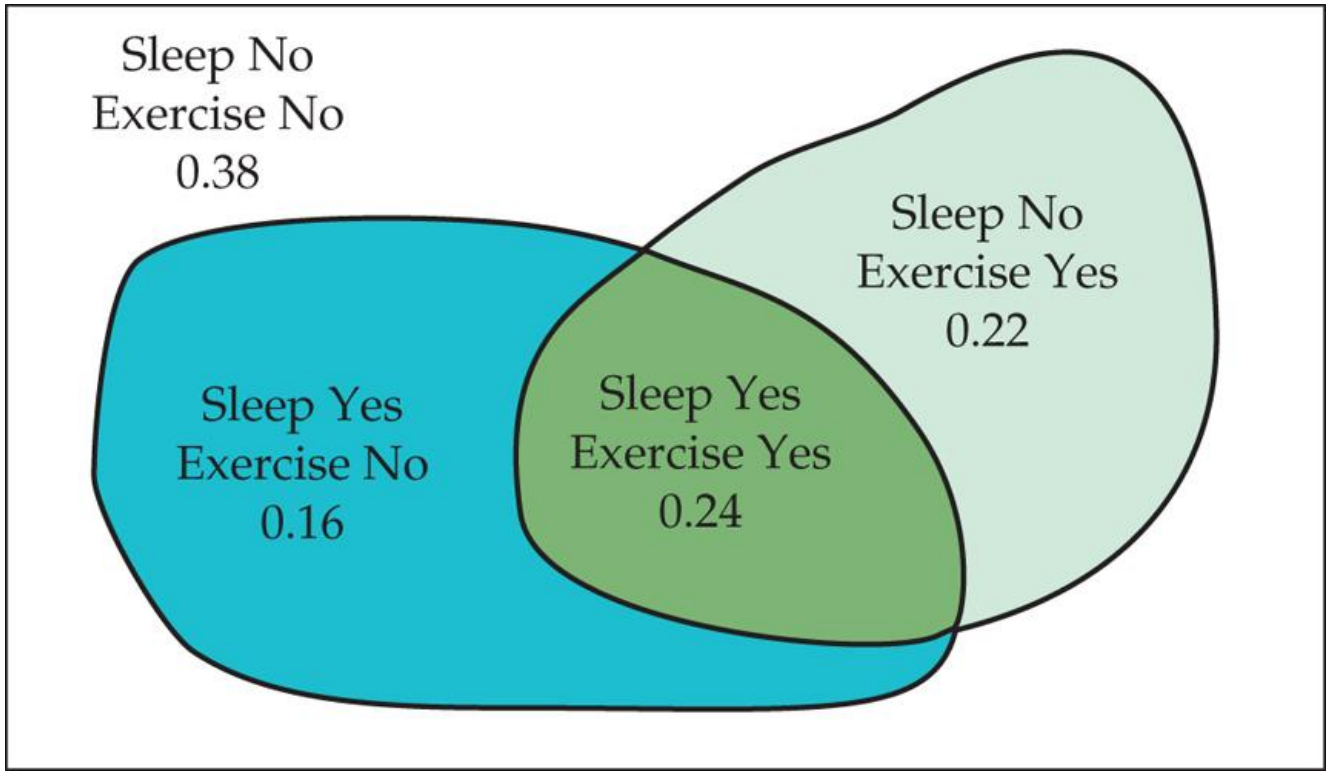
P(«nok søvn» eller «trener regelnmessig»)

= P(«nok søvn»)+ P(«trener regelnmessig»)

-P(«nok søvn» og «trener regelnmessig»)

= 0.40 + 0.46 – 0-24 = 0.62.

Venn diagrammer er nyttige :


Sleep No
Exercise No
0.38
Sleep No
Exercise Yes
0.22
Sleep Yes
Exercise No
0.16
Sleep Yes
Exercise Yes
0.24

Eksempel: Kortspill

P(ess)=g/m= 4/52 =1/13

Men hvis man allerede vet at det er eksakt **ett** ess blant
fire kjente og allerede utdelte kort.

Da endres sannsynlighetene til

m=48, siden det allerede er delt ut 4
g=3, siden det er tre ess igjen i stokken.

Dette er et eksempel på **en betinget sannsynlighet**, sannsynligheten for en begivenhet gitt at vi vet at en annen har inntruffet,

P(ess i nytt kort | 1 ess av fire allerede synlige kort})=3/48=1/16

Dette skrives $P(B \mid A)$.

Generelt gjelder **multiplikasjonsregelen**

$P(A \text{ og } B) = P(A)\, P(B \mid A)$

eller

$P(B \mid A) = P(A \text{ og } B) / P(A)$ hvis $P(A) > 0$.

$P(B \mid A)$ er den betingede sannsynligheten for at
B inntreffer hvis A inntreffer.

Eksempel: Kortspill,

11 kort på bordet herav 4 ruter.

Trekke to kort til
A: Første ruter
B: Andre ruter

$P(A) = g/m = (13-4)/(52-11) = 9/41$
$P(B|A) = (13-5)/(52-12) = 8/40$

$P(A \text{ og } B) = (9/41)\times(8/40) = 0.044$

Ofte er betingede sannsynligheter enn del av informasjonen vi har og må brukes for å finne P(A og B) eller
P(A og B og C og...)

Snittet (intersection) til en samling eller familie begivenheter er begivenheten at alle begivenhetene inntreffer.

For tre begivenheter blir multiplikasjonsregelen

$$P(A \text{ og } B \text{ og } C) = P(A \text{ og } B)\, P(C \mid A \text{ og } B)$$
$$= P(A)\, P(B \mid A)\, P(C \mid A \text{ og } B)$$

For å holde orden på de betingede sannsynlighetene kan **trediagrammer** være nyttige.

Eksempel:

C: Medlem av «chat room»

A_1 : Aldersgruppe 18-29 og internettbruker

A_2 : Aldersgruppe 30-49 og internettbruker

A_3 : Aldersgruppe 50+ og internettbruker

P(A_1)= 0.29, P(A_2)= 0.47, P(A_3)= 0.24

P(C|A_1)= 0.47, P(C|A_2)= 0.21, P(C|A_3)= 0.07

P(C)= P(C og A_1) + P(C og A_2) + P(C og A_3)

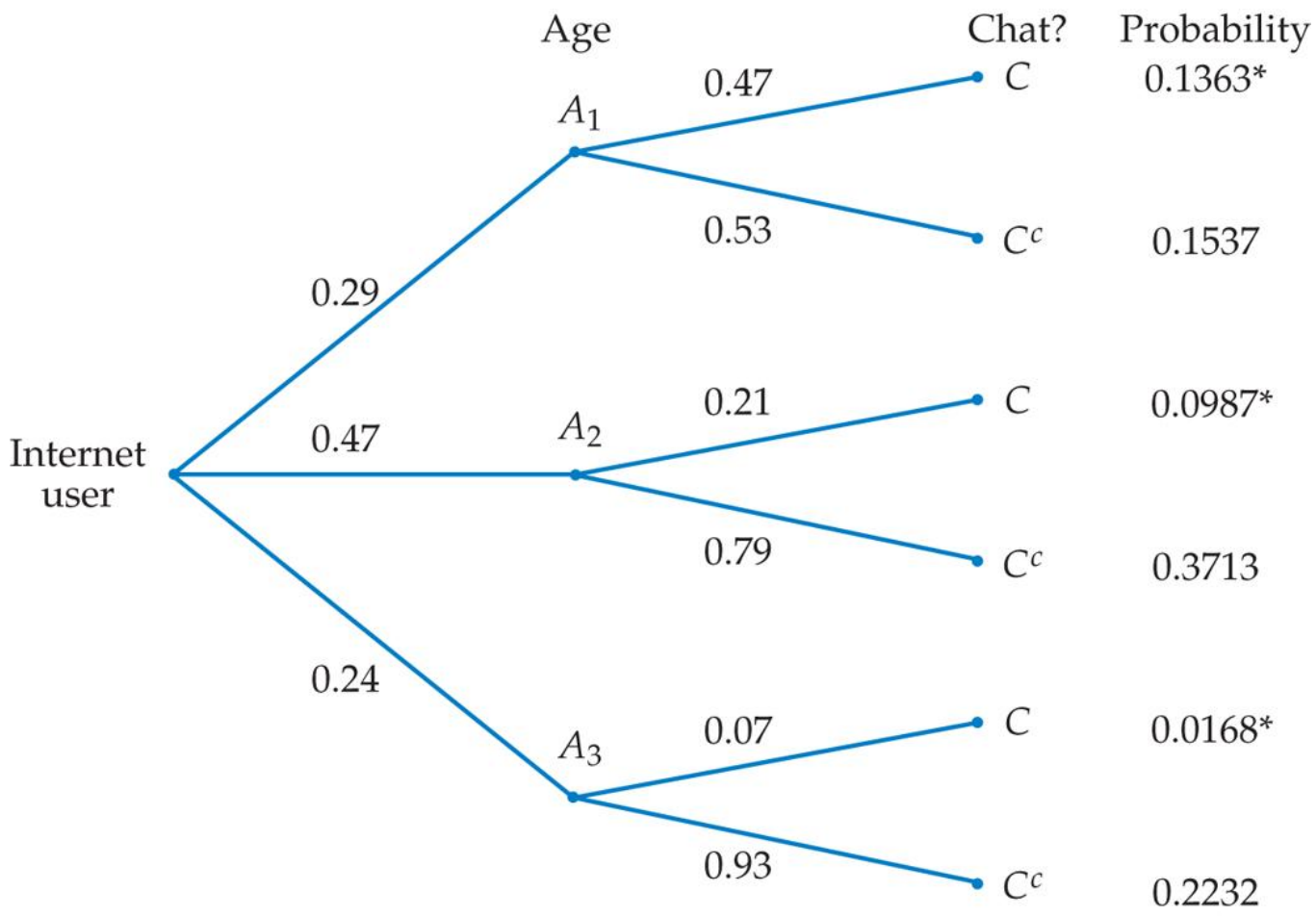
= P(A_1) P(A_1|C)+ P(A_2) P(A_2|C)+ P(A_3) P(A_3|C)

= 0.29×0.47+ 0.47×0.21+ 0.24×0.07

= 0.1363+0.0987+0.0168

= 0.2518

Tilhørende trediagram



Bayes regel (i litt enklere form enn på side 293)

Anta at $0 < P(A) < 1$ og $0 < P(B) < 1$. Da gjelder

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

Utledning: Pr definisjon er $P(A|B) = P(A \text{ og } B) / P(B)$

Men vi kan skrive telleren $P(A \text{ og } B) = P(B \mid A)\,P(A)$

og dessuten neveren $P(B) = P(A \text{ og } B) + P(A^c \text{ og } B)$.

Men dette siste uttrykket kan omformes til

$$P(B) = P(B \mid A)\,P(A) + P(B \mid A^c)\,P(A^c)$$

Bayes regel (generell form, side 293)

Anta at $0 < P(A_i) < 1$, i=1,…,k er disjunkte begivenheter slik at $P(A_1 \text{ eller } A_2 \text{ eller } \ldots \text{ eller } A_k) = 1$ samt at $0 < P(B) < 1$. Da gjelder

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B|A_1)\,P(A_1) + P(B|A_2)\,P(A_2) + \cdots + P(B|A_k)\,P(A_k)}$$

Eksempel: Falske positive

Diagnose:
B: Person faktisk syk
A: Diagnose syk

P(A|B): Diagnosens evne til å avsløre syke
P(A|B^c): Sannsynlighet for feilklassifikasjon

P(B^c|A): Falsk positiv

Fra Bayes regel:

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

P(B) utbredelsen av sykdommen.

Dette kan anvendes på f.eks. mammografiscreening:
La B = positiv mammografiprøve, dvs. prøven viser tegn på kreft
og A = kvinnen har faktisk brystkreft og anta at

sannsynligheten for B hvis brystkreft = $P(B \mid A) = 0.95$
sannsynligheten for B hvis ikke brystkreft = $P(B \mid A^c) = 0.035$
samt at sannsynligheten for brystkreft er lik $P(A) = 0.007$.

Ved Bayes formel finner vi at sannsynligheten for brystkreft gitt positiv mammografiprøve blir

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)\,P(A) + P(B|A^C)\,P(A^C)}$$

$$= \frac{0.95 \cdot 0.007}{0.95 \cdot 0.007 + 0.035 \cdot (1\text{-}0.007)} = 0.16$$

Så selv om mammografitesten er ganske god (dvs. små sannsynligheter for feilklassifisering) så er det likevel så liten sannsynlighet som 16% for å ha brystkreft etter en positiv test.

Dette har samsvar med at risken for brystkreft ved en prøve er nokså liten.

Man regne ut slike sannsynligheter uten formelt å bruke Bayes regel, men ved å tegne opp trediagrammer for problemet.

I slike diagrammer beregenes

$P(B \text{ og } A) = P(B \mid A) P(A)$, $P(B \text{ og } A^c) = P(B \mid A^c) P(A^c)$

$P(B^c \text{ og } A) = P(B^c \mid A) P(A)$ og $P(B^c \text{ og } A^c) = P(B^c \mid A^c) P(A^c)$

Og finner dermed

$P(B \mid A) = P(A \text{ og } B) / P(B) = P(A \text{ og } B) /(P(B \text{ og } A) + P(B \text{ og } A^c))$

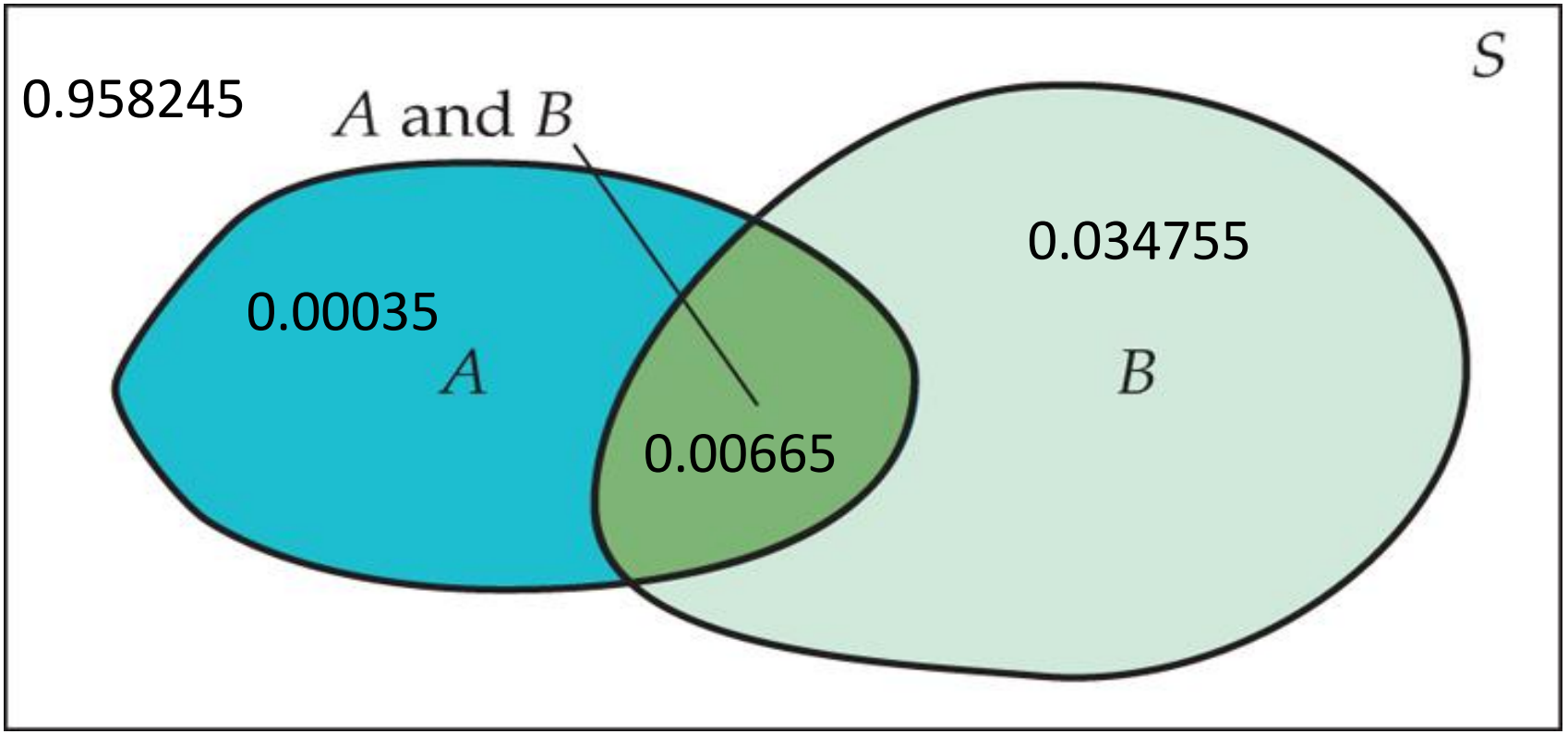
Eks. mammografiscreening:

$P(A \text{ og } B) = P(A)\,P(B \mid A) = 0.007 \times 0.95 = 0.00665$

$P(A \text{ og } B^c) = P(A)\,P(B^c \mid A) = 0.007 \times (1-0.95) = 0.00035$

$P(A^c \text{ og } B) = P(A^c)\,P(B \mid A^c) = 0.993 \times 0.035 = 0.034755$

$P(A^c \text{ og } B^c) = P(A^c)\,P(B^c \mid A^c) = 0.993 \times 0.965 = 0.958245$



Så $P(B) = 0.03475+0.00665 = 0.0414$ og

$P(B \mid A) = 0.0665 / 0.414 = 16\%$

To begivenheter A og B der $P(A) > 0$ og $P(B)>0$ er **uavhengige** hvis

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

Merk: Multiplikasjonsregelen for uavhengige begivenheter er et spesialtilfelle av den generelle multiplikasjonsregelen, $P(A \text{ og } B)= P(B|A)P(A)$.

Eksempel: Tvillingers kjønn

P(«enegget») = 1/3, antagelse

Addisjonsregel:

P(«samme kjønn»)= P(«enegget» og «samme kjønn»)
+ P(«toegget» og «samme kjønn»)
=P(«samme kjønn»| «enegget») P(«enegget»)
+ P(«samme kjønn»| («toegget»)P («toegget»)
= 1×(1/3) + (1/2)x(2/3)= 2/3

Bayes regel:

P(«enegget»|«samme kjønn»)=
P(«samme kjønn| «enegget»)P («enegget»)/
(P(«samme kjønn| «enegget»)P («enegget»)+
P(«samme kjønn| «toegget»)P («toegget»))
= 1×(1/3) / (1×(1/3 + (1/2)×(2/3)) = ½

Alternativt, multiplikasjonsregelen:

P(«enegget»|«samme kjønn») =
P(«enegget» og«samme kjønn»)/P(«samme kjønn»)=
P(«enegget» og«samme kjønn»)/P(«samme kjønn»=
P(«samme kjønn»| «enegget»)P(«enegget»)/P(«samme kjønn)=
1×(1/3)/(2/3)= ½.