Kapittel 4.3:
Tilfeldige/stokastiske variable

Litt repetisjon: Sannsynlighetsteori

Stokastisk forsøk og sannsynlighet

- Tilfeldig fenomen
  - Individuelle utfall er usikre, men likevel et regulært mønster for et stort antall repetisjoner
- Sannsynlighet for et utfall av et tilfeldig fenomen
  - Andel ganger et utfall skjer i en veldig lang serie av repetisjoner

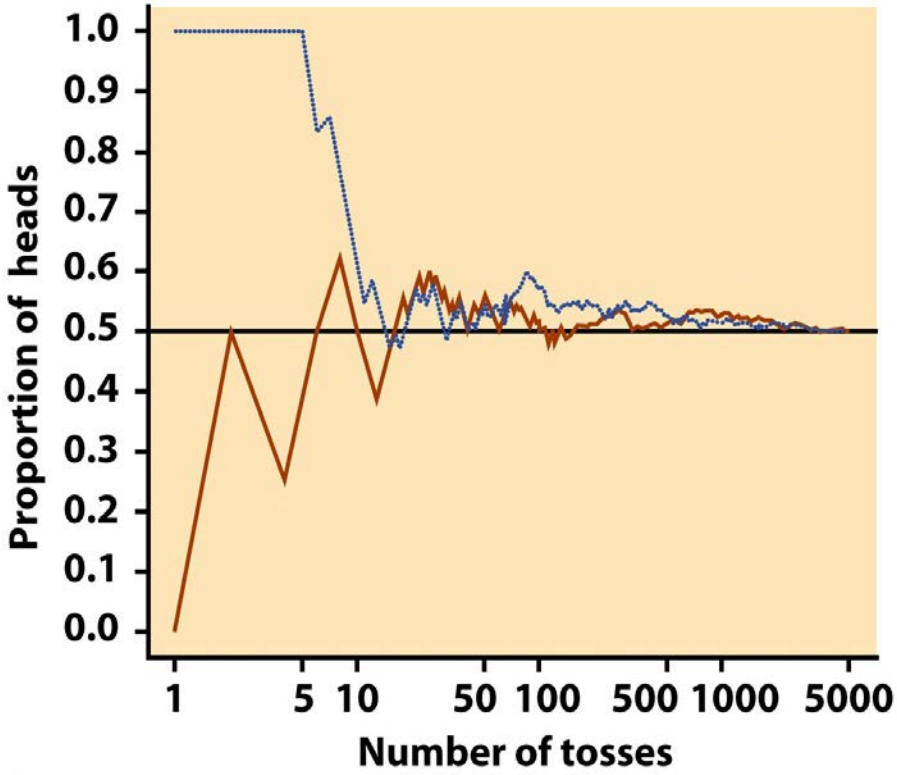


Figure 4-1
Introduction to the Practice of Statistics, Fifth Edition

Utfallsrom

- Utfallsrommet til et tilfeldig fenomen er mengden av alle mulige utfall
- En hendelse er et utfall eller et sett av utfall av et tilfeldig fenomen- det er altså en delmengde (undergruppe) av utfallsrommet
- Eksempel: Ett myntkast
  - Utfallsrom: S={Mynt (M), Kron (K)}
  - Eksempel hendelse: Ett myntkast gir mynt ={M}
- Eksempel: To myntkast
  - Utfallsrom: S={MM,MK,KM,KK}
  - Eksempel hendelse: To myntkast gir minst en mynt ={MM,MK,KM}

Egenskaper sannsynligheter

1:

2:

3:

4:

Disjunkte (ikke-overlappende) og ikke-diskjunkte (overlappende) hendelser

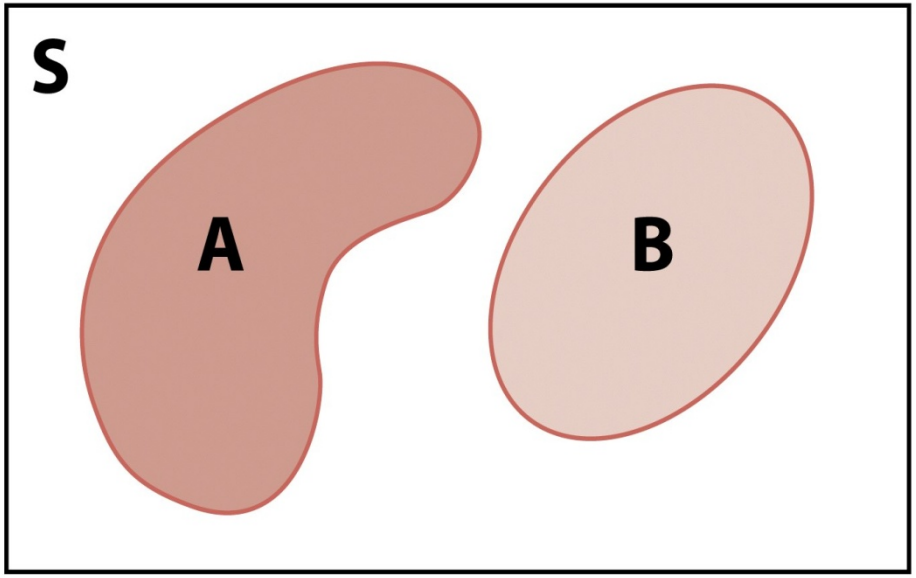


Figure 4-2
Introduction to the Practice of Statistics, Fifth Edition

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

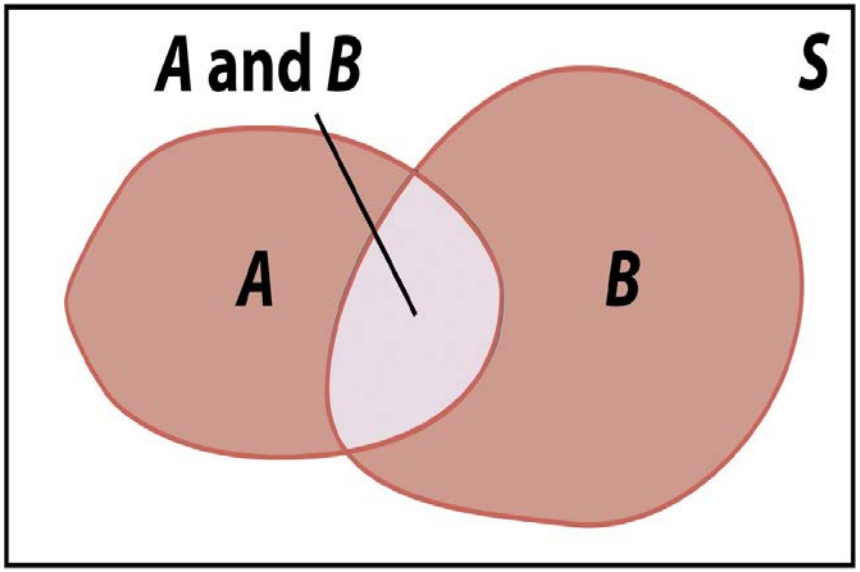


Figure 4-4
Introduction to the Practice of Statistics, Fifth Edition

Her:
P(A eller B) =P(A)+P(B)-P(A og B)

Endelige antall utfall

- De individuelle utfallende til et tilfeldig fenomen er alltid disjunkte
- Addisjonsregelen gir dermed en metode for å gi sannsynligheter til begivenheter med mer enn ett utfall
  - Gi en sannsynlighet til hvert individuelle utfall (tall mellom 0 og 1 som tilsammen summerer seg til 1)
  - Sannsynlighet for hendelse fås ved å summere sannsynlighetene for alle individuelle utfall involvert

Like sannsynlige utfall

Et stokastisk forsøk har *N* utfall
Det er de *mulige utfallene* for forsøket

Vi antar at de *N* utfallene er *like sannsynlige*

Da her hvert utfall sannsynlighet 1/*N*

En begivenhet *A* består av *n* utfall
Det er de *gunstige utfallene* for begivenheten *A*

Sannsynligheten for begivenheten *A* er

$$P(A) = \frac{n}{N} = \frac{\text{antall gunstige utfall}}{\text{antall mulige utfall}}$$

Uavhengighet og multiplikasjonsregelen

5. Multiplikasjonsregel $P(A \text{ og } B)=P(B|A) * P(A)$

Spesialtilfelle når A og B er uavhengige: $P(A \text{ og } B)=P(A) * P(B)$

- Eksempel: Myntkast
  - A = {Første kast er kron}
  - B = {Andre kast er kron}

  - Rimelig å anta at A og B er uavhengige
    - Kunnskap om A endrer ikke sannsynligheten for B

$P(A \text{ og } B) = P(A) * P(B) = 0.5*0.5=0.25$

P(A og B) = P(A) * P(B): Kun ved uavhengighet!

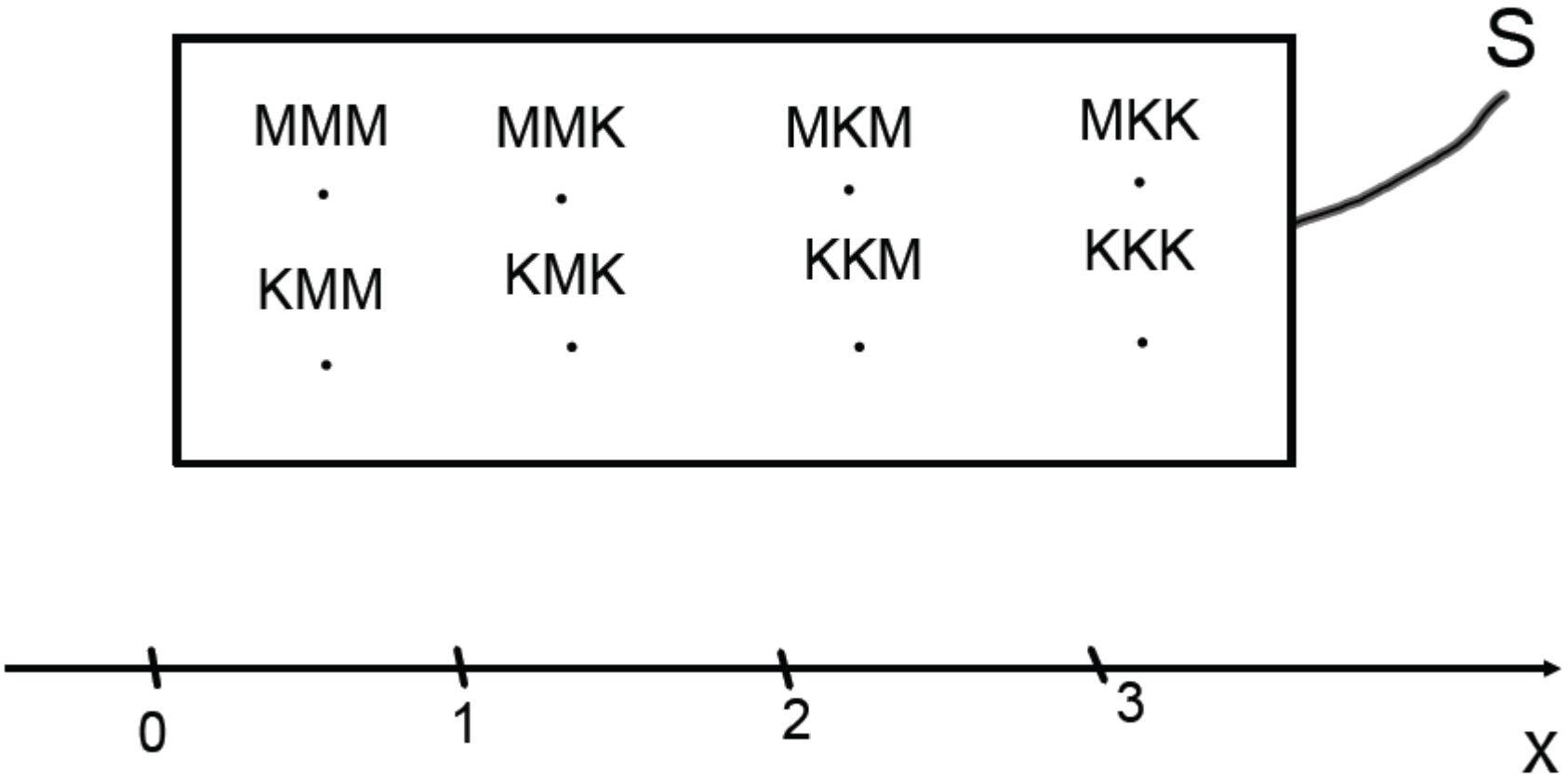
- Krybbedød: 1 av 8500 dør uforklarlig, sanns. 0.000118
- To barn døde i samme familie, A=”1. barn dør”, B=”2. barn dør”
  - Man antok uavhengighet: P(B|A)=P(B) = 0.000118
    P(A)= P(A)* P(B)=0.000118*0.000118=1/72 250 000
  - Flere foreldre siktet for drap i England
  - Royal Statistical Society: Det er ikke rimelig å anta uavhengighet, det kan være en genetisk faktor, som gjør at P(B |A) >> P(B), slik at

    P(A og B)= P(A)* P(B| A) >> 1/72 250 000
  - Det førte til at den britisk regjeringen tok opp 258 saker på nytt

Kap 4.3: Stokastiske variable

- Utfallsrom kan ta ulike former, ikke nødvendigvis tall. Eksempel myntkast:
  - 3 myntkast, utfallsrom: S={MMM,MMK,MKM, MKK,KMM,KMK,...}
  - Ønskelig å uttrykke utfall ved tall
  - Dersom X = antall kron i 3 myntkast, er utfallsrommet for X: S={0,1,2,3}
  - X er en *stokastisk (eller tilfeldig) variabel* som kan ta verdiene 0,1, 2 eller 3

Kast en mynt tre ganger. Da har vi 8 mulige enkeltutfall.



La X være antall kron i de tre kastene. X kalles en tilfeldig variabel

Stokastiske variable

- Stokastisk variabel: Variabel som tar verdier som er det numeriske utfallet av et stokastisk fenomen
- Betegnes ofte med store bokstaver nær slutten av alfabetet, f.eks. X og Y

- Hendelser kan formuleres v.h.a. X, for eksempel

$$P(\text{'minst en kron'}) = P(X \geq 1)$$

$$P(\text{'ingen kron'}) = P(X = 0)$$

Stokastiske variable og sansynlighetsmodeller

- Utfallsrommet til en stokastisk variabel X: Alle mulige verdier som X kan ta
- Vi skiller mellom diskrete og kontinuerlige stokastiske variable, etter om utfallsrommet for variabelen er diskret eller kontinuerlig
- Hendelser kan formuleres v.h.a. X, for eksempel

$$P(\text{'minst en kron'}) = P(X \geq 1)$$

$$P(\text{'ingen kron'}) = P(X = 0)$$

Diskrete stokastiske variable

- En diskret stokastisk variabel X har et endelig antall mulige verdier
- **Sannsynlighetsfordelingen** for X viser de mulige verdiene av X med tilhørende sannsynligheter:

- Sannsynlighetene p_i må tilfredsstille

  1.

  2.

- Sannsynligheten for en hendelse finnes ved å legge sammen sannsynlighetene p_i for de verdiene x_i hendelsen består av

Diskret stokastisk variabel: Eksempel

- I er statistikkurs er karakterfordelingen som følger:

  1% F-er, 5% D-er, 30% C-er, 43% B-er og 21% A-er

- En students karakter X på en heltallsskala fra 0 til 4 (der 0 tilsvarer F, 1 tilsvarer D, ..., og 4 tilsvarer A) er en diskret stokastisk variabel. Sannsynlighetsfordelingen til X er da

- Hendelsen at en student får B eller bedre er det samme som at X=3 eller X=4. Sannsynligheten for at en student får B eller bedre er da

Sannsynlighets-histogrammer

- *Høyden på søylene i histogrammet viser sannsynligheten av utfallet vist på den horisontale aksen. Summen av søylene summerer seg til 1*
- (a) viser et sannsynlighetshistogram for kast med en 9-sidet terning
- (b) viser et sannsynlighetshistogram for Benford's lov

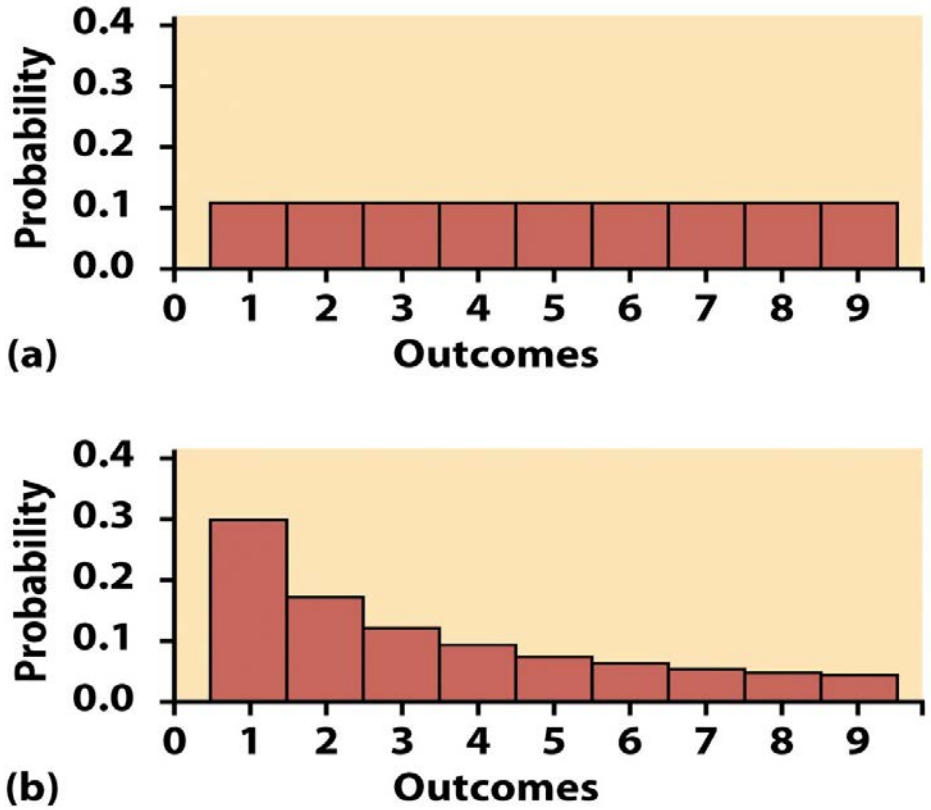


Figure 4-5
Introduction to the Practice of Statistics, Fifth Edition

Første tall	1	2	3	4	5	6	7	8	9
Sannsynlighet	0.301	0.176	0.125	0.097	0.079	0.067	0.058	0.051	0.046

Myntkast

- Hva er sannsynlighetsfordelingen til den diskrete stokastiske variabelen som teller antall kron i fire myntkast?
- X=antall kron (H) i 4 myntkast
- Gjør to rimelige antagelser
  - P(H)=P(T)=1/2, dvs. balansert mynt
  - Uavhengige kast
- Totalt 16 mulige utfall av de fire kastene, og 4 mulige verdier av X

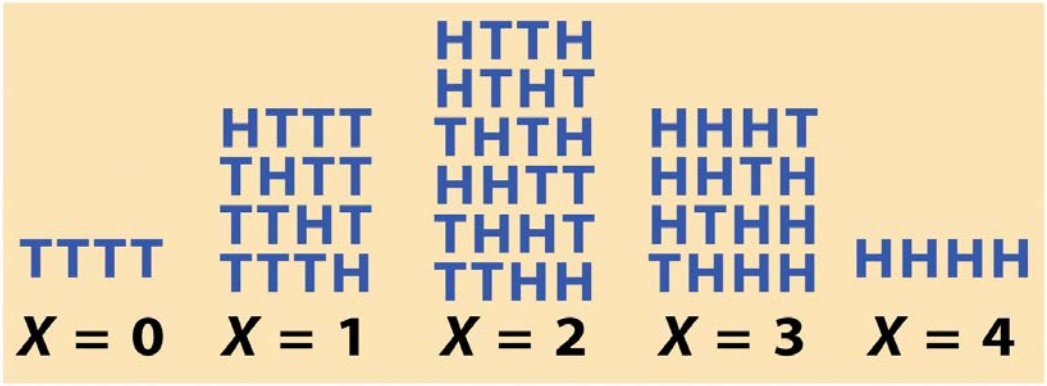


Figure 4-6
Introduction to the Practice of Statistics, Fifth Edition

Myntkast

- Multiplikasjonsregelen for uavhengige hendelser gir oss at f.eks. hendelsen HTTH har sannsynlighet P(HTTH)=1/2*1/2*1/2*1/2=1/16 - alle de seksten mulige utfallene av myntkastene har sannsynlighet 1/16
- Den stokastiske variabelen X (som teller antall kron (H) i de fire myntkastene) har fire mulige verdier (0,1,2,3,4). Disse verdiene er *ikke* like sannsynlige:

P(X=0)=P(TTTT)=

(Antall «gunstige» kombinasjoner)/(Antall mulige kombinasjoner)

=

P(X=2)=

(Antall «gunstige» kombinasjoner)/(Antall mulige kombinasjoner)

=

- Tilsvarende finnes sannsynlighetene for de andre mulige verdiene av X

X	0	1	2	3	4
P(X)	0.0625	0.250	0.375	0.250	0.0625

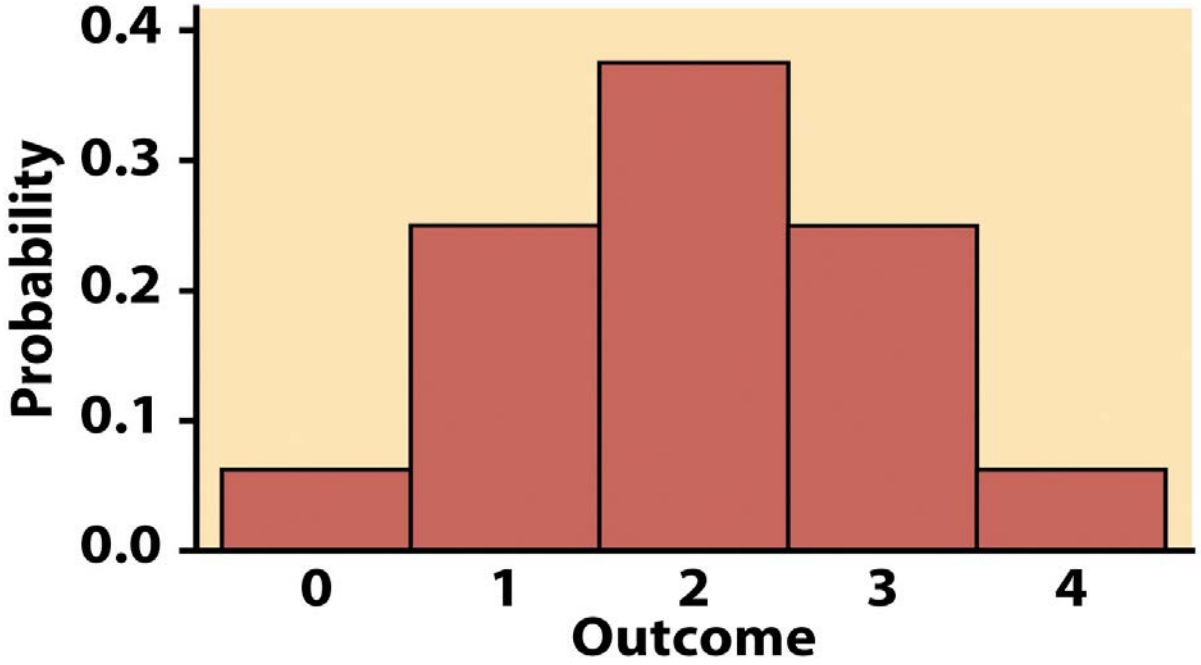


Figure 4-7
Introduction to the Practice of Statistics, Fifth Edition

Symmetrisk!

Idealisering av hva som ville skjedd etter veldig mange repetisjoner

Myntkast: Sannsynligheter for sammensatte hendelser

- $P(X \geq 2) = P(X=2) + P(X=3) + P(X=4)$

  $= 0.375 + 0.25 + 0.0625 = 0.6875$

- $P(X \geq 1) = 1 - P(X=0)$ (komplementregelen)

  $= 1 - 0.0625 = 0.9375$

(Kunne også funnet denne som

$P(X \geq 1) = P(X=1) + P(X=2) + P(X=3) + P(X=4)$,

men enklere/mindre regning å bruke komplementregelen her)

Kontinuerlige stokastiske variable

- Datamaskiner genererer tilfeldige tall mellom 0 og 1
- S={alle tall x slik at $0 \leq x \leq 1$}
- Sprer data uniformt over S
- Hvordan kan vi f.eks. finne en sannsynlighet for hendelsen $0.3 \leq x \leq 0.7$?
- Kan ikke lengre allokere sannsynligheter til hver mulig verdi av x og summere, for det er uendelig mange mulige verdier!
- Bruker da tetthetskurver og areal

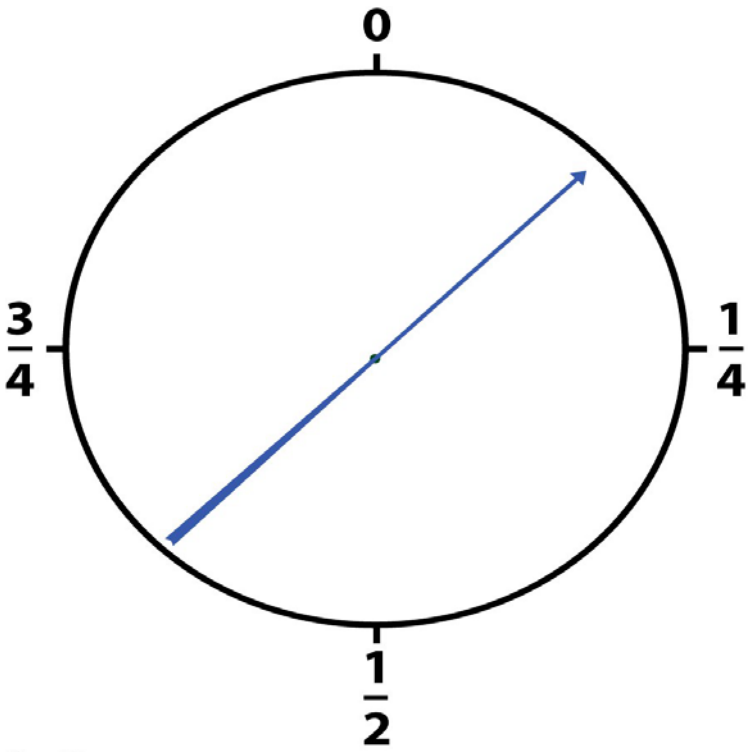


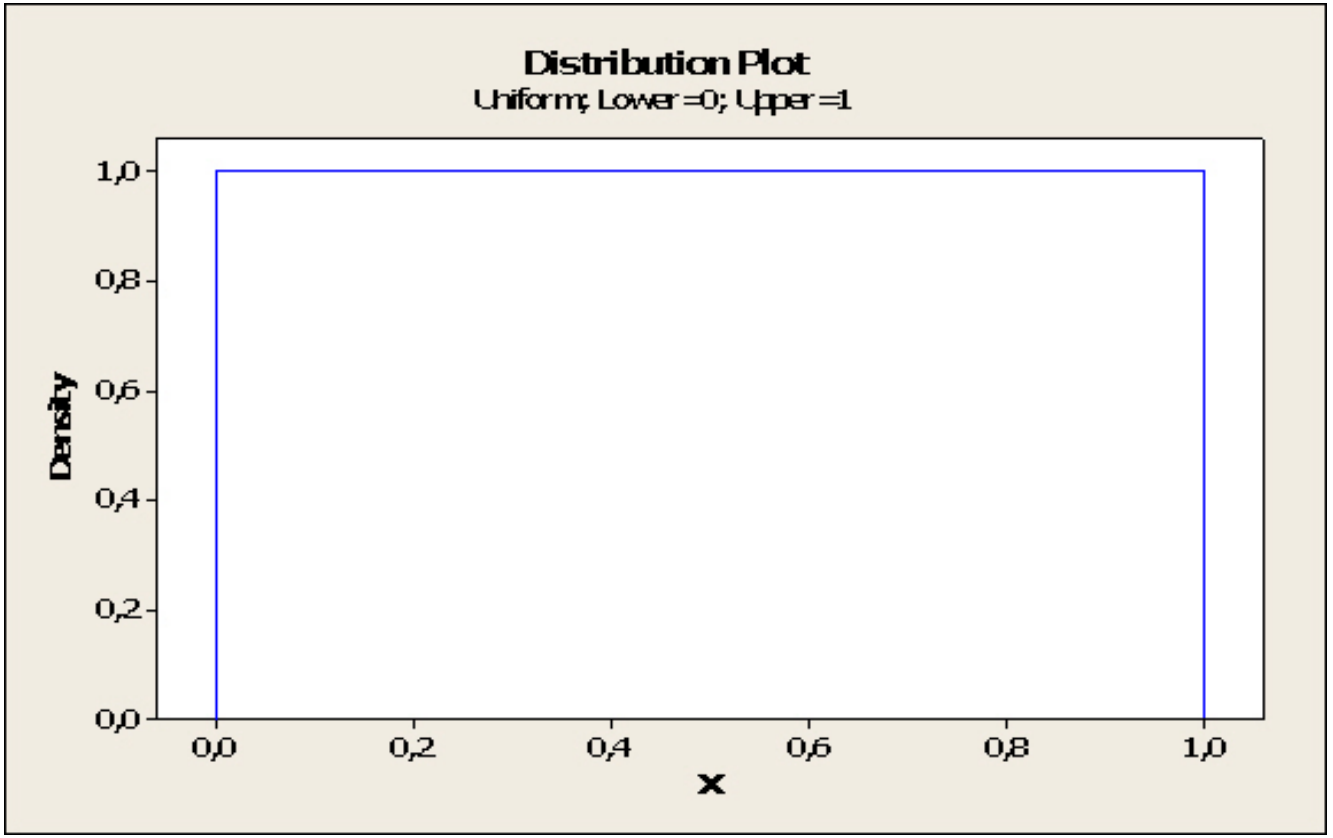
Figure 4-8
Introduction to the Practice of Statistics, Fifth Edition

Kontinuerlige tilfeldige variable

- En variabel X som tar verdier i et intervall av tall (ikke et eksakt, enkelt tall)
- Sannsynlighetsfordelingen til X beskrives av en tetthetskurve ( )
  - Kun ikkenegative verdier ( )
  - Totalt areal under fordelingen lik 1 ( )
- Sannsynligheten for en hendelse er arealet under tetthetskurven og over de verdier av X som beskriver hendelsen

Uniform (lik) tetthetskurve mellom 0 og 1

Fordi en tetthetskurve har areal=1, og denne tetthetskurven er for X mellom 0 og 1, er høyden=1 (areal=1*1=1)



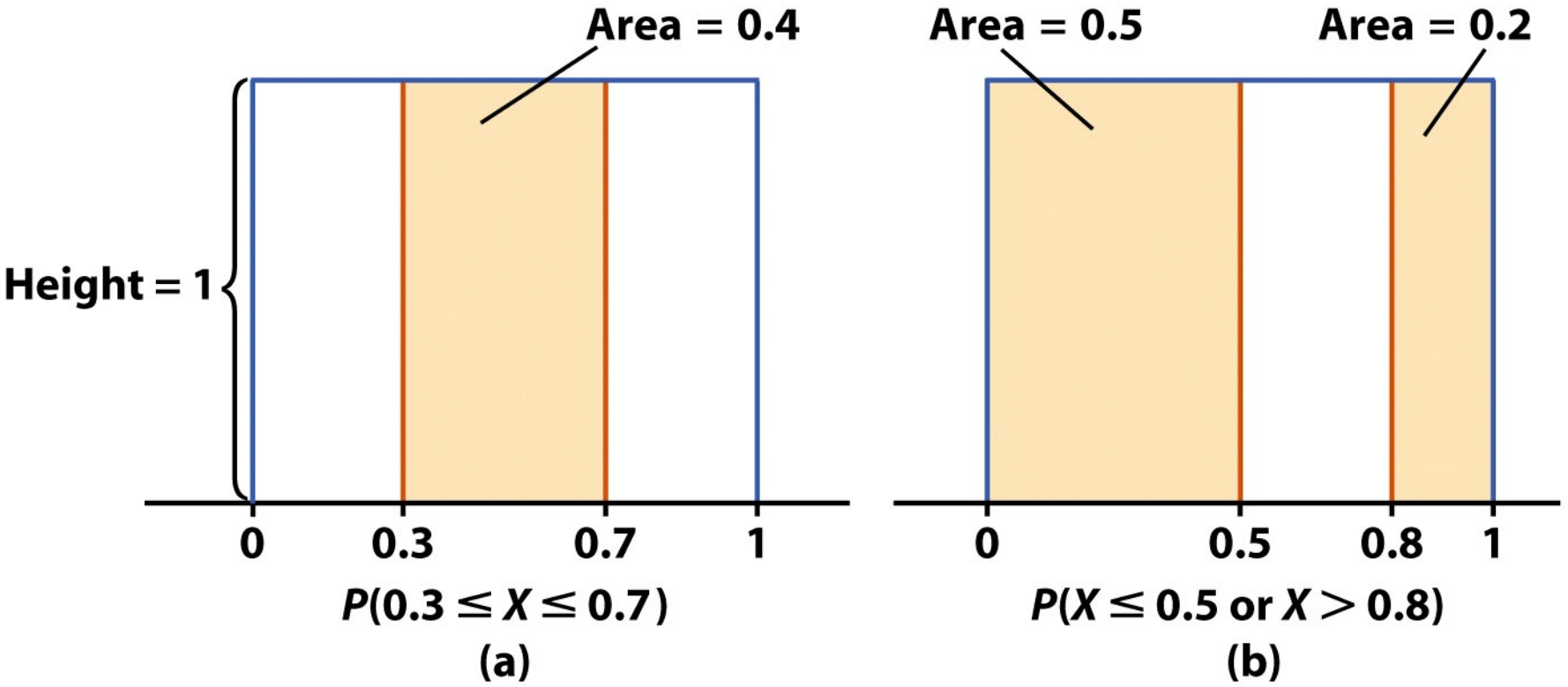


Figure 4-9
Introduction to the Practice of Statistics, Fifth Edition

Generell tetthetskurve: Sannsynligheten for en hendelse

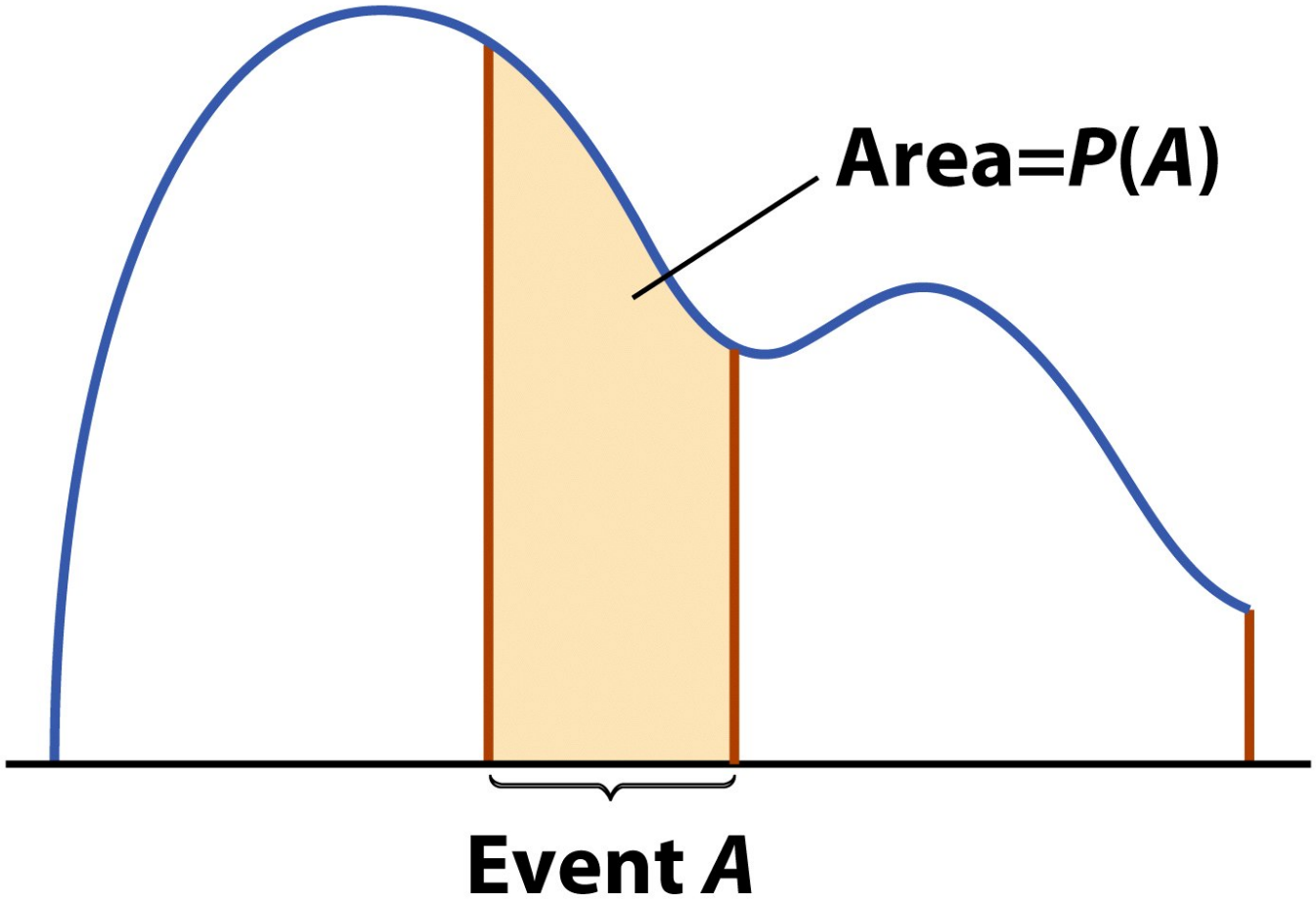


Figure 4-10
Introduction to the Practice of Statistics, Fifth Edition

Sannsynlighet 0 for enkeltverdier

- Angir sannsynligheter for kontinuerlige variable ved intervaller istedet for individuelle utfall
- Et utfall vil i praksis aldri være helt lik en bestemt verdi
- Alle sannsynlighetsfordelinger for kontinuerlige variable gir sannsynlighet 0 til alle individuelle utfall ($P(X=x)=0$)

- Eksempel: Et utfall av en kontinuerlig variabel med tetthetskurve mellom 0 og 1 vil i praksis aldri være helt lik 0.8
  - X=0.8 er en mengde av lengde 0
  - Altså er arealet under kurven=0, og P(X=0.8)=0
  - Intuitiv forståelse: Se på sannsynligheten for intervaller som krymper:
    - Tre desimaler: At X ligger mellom 0.799 og 0.801 har sannsynlighet 0.002
    - Seks desimaler: At X ligger mellom 0.799999 og 0.800001 har sannsynlighet 0.000002
    - Jo trangere vi gjør intervallet, jo nærmere 0.8 vi kommer, jo mer nærmer sannsynligheten seg 0

Kontinuerlige variable: Ingen forskjell på $<$ og $\leq$

- Fordi det ikke er noen sannsynlighet for eksakt X=0.8 når X er en kontinuerlig stokastisk variabel, har hendelsene $X< 0.8$ og $X \leq 0.8$ samme sannsynlighet
- Kan ignorere forskjellen mellom $<$ og $\leq$ for kontinuerlige stokastiske variable (men NB: det gjelder ikke for diskrete stokastiske variable!)
- Tilsvarende for $>$ og $\geq$
- Dvs $P(X< 0.8)= P(X \leq 0.8)$ og $P(X> 0.8)= P(X \geq 0.8)$ når X er en kontinuerlig stokastisk variabel

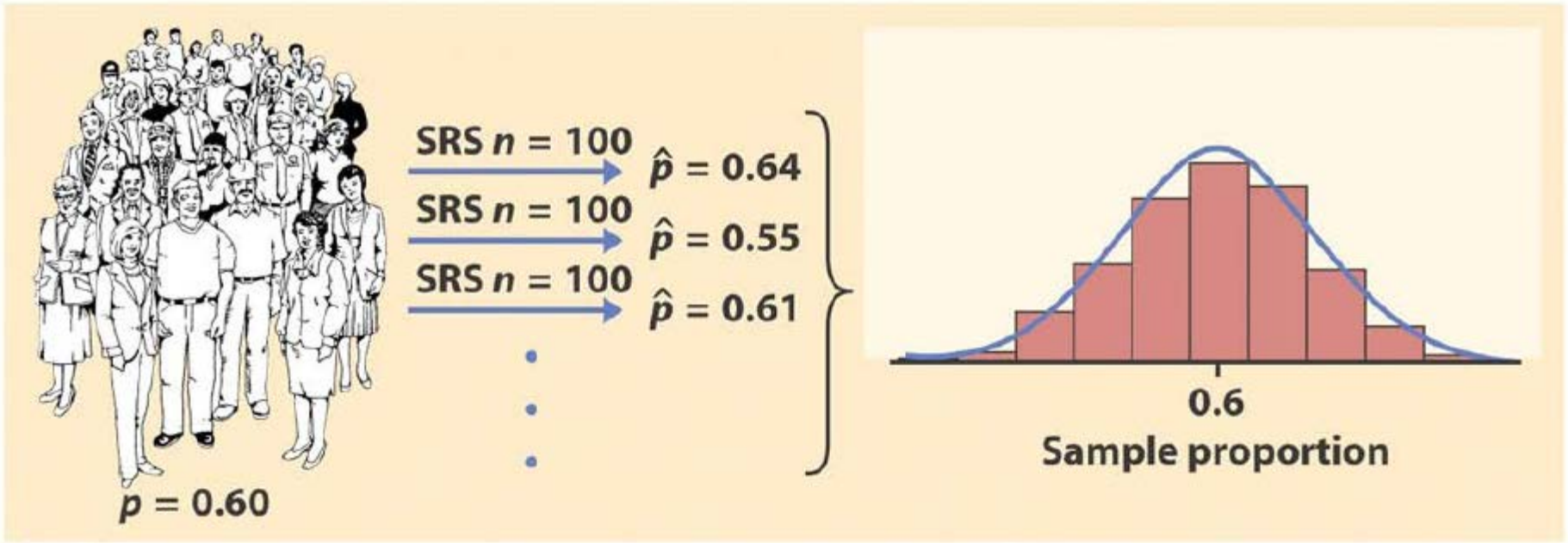
Normalfordeling

- Kjenner fra før tetthetskurven for normalfordelingen
- Normalfordelingen er en sannsynlighetsfordeling
- Hvis X er normalfordelt med forventning μ og standardavvik σ, sier vi at X er $N(\mu,\sigma)$-fordelt
- $Z=(X-\mu)/\sigma$ er da N(0,1)-fordelt
- Skriver

Normalfordeling

- Eksempel:
  - p=andel studenter som jukser på eksamen, parameter i populasjonen
  - $\hat{p}$=andel observert jukset i et tilfeldig utvalg på 400 studenter, observator vi kan bruke til å estimere (anslå) p
  - $\hat{p}$ er en stokastisk variabel, repeterte utvalg vil gi ulike verdier av $\hat{p}$
  - Kan vise (kap. 5): $\hat{p}$ er tilnærmet N(0.12,0.016)-fordelt (hvis p=0.12). Har også sett dette i kap. 3 *
  - Hva er sannsynligheten for at observatoren fra utvalget ikke er mer enn 0.02 forskjellig fra den sanne populasjonsverdien av p? Dvs. hva er sannsynligheten for at $0.10 \leq \hat{p} \leq 0.14$?

* Husk figur fra kap. 3:



Nå:

$P(0.10 \leq \hat{p} \leq 0.14) = P(0.10 < \hat{p} < 0.14) =$

$P((0.10 - 0.12)/0.016 < (\hat{p} - 0.12)/0.016 < (0.14 - 0.12)/0.016) =$

$P(-1.25 < Z < 1.25) = P(Z < 1.25) - P(Z - < 1.25) = 0.8944 - 0.1056 = 0.7888$

- Dette har vi lært å regne på før, men nå bruker vi sannsynlighetsspråket (i kap 1.3 snakket vi om andeler)