Begreper vi har diskutert

- ⋆ Stokastisk forsøk
- ⋆ Utfallsrom
- ⋆ Hendelse
- ⋆ Venndiagram
- ⋆ Sannsynlighet
- ⋆ Betinget sannsynlighet
- ⋆ Uavhengige hendelser
- ⋆ Stokastisk variabel
- ⋆ Punktsannsynlighet
- ⋆ Sannsynlighetstetthet
- ⋆ Kumulativ fordeling
- ⋆ Simultanfordeling
- ⋆ Betinget fordeling
- ⋆ Uavhengige stokastiske variabler
- ⋆ Forventingsverdi
- ⋆ Varians
- ⋆ Standardavvik
- ⋆ Kovarians
- ⋆ Korrelasjon
- ⋆ Bernoulliprosess
- ⋆ Poissonprosess
- ⋆ Transformasjon
- ⋆ Momentgenererende funksjoner
- ⋆ Ekstremvariabler
- ⋆ Ordningsvariabler
- ⋆ Populasjon

Begreper vi har diskutert

- Utvalg
- Tilfeldig utvalg
- Observator
- Estimator
- Forventningsrett estimator
- Forventningsskjev estimator
- Mest effisient estimator
- Rimelighetsfunksjon
- Sannsynlighetsmaksimerings-prinsippet
- Sannsynlighetsmaksimerings-estimator
- Konfidensintervall
- Ett-utvalg
- To-utvalg
- Prediksjonsintervall
- Hypotesetesting
- Nullhypotese
- Alternativ hypotese
- Testobservator
- Signifikansnivå
- Forkastningsområde
- p-verdi
- Ensidig test
- Tosidig test
- Enkel lineær regresjon
- Minste kvadraters metode

Hva har vi gjort med disse begrepene?

- ⋆ Definert (og ofte motivert definisjonen)
- ⋆ Diskutert tolkning
- ⋆ Introdusert regneregler
- ⋆ Regnet på eksempler
- ⋆ Tolket svaret i regneeksempler

Hva har vi gjort med disse begrepene?

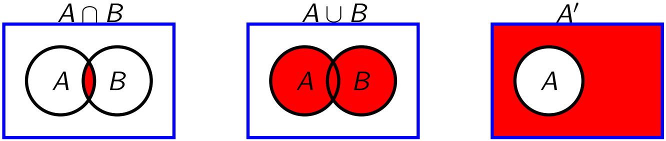
- ⋆ Definert (og ofte motivert definisjonen)
- ⋆ Diskutert tolkning
- ⋆ Introdusert regneregler
- ⋆ Regnet på eksempler
- ⋆ Tolket svaret i regneeksempler

Hva forventer vi av dere?

- ⋆ Skal kunne anvende begrepene til å svare på spørsmål
- ⋆ Skal kunne regne
- ⋆ Skal kunne tolke svarene
- ⋆ Skal kunne forklare begrepene

Stokastisk forsøk, utfallsrom og venndiagram

- Stokastisk forsøk: Et forsøk hvor resultatet er underlagt tilfeldigheter
- Utfallsrom: Mengden av alle mulige resultater, S
- Enkeltutfall: Et mulig utfall, e
- Hendelse: En delmengde av utfallsrommet, $A, B, C, \ldots$
- Venndiagram: Illustrerer hendelser og operasjoner på hendelser



- Operasjoner på hendelser
  - Snitt
  - Union
  - Komplement
- Disjunkte hendelser

Sannsynlighet

- ⋆ Sannsynlighet: Reell funksjon definert på hendelser, P
  1. $0 \leq P(A) \leq 1$
  2. $P(S) = 1$
  3. Hvis $A_1, A_2, \ldots$ parvis disjunkte:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

    – tenk på sannsynlighet som areal i venndiagram

- ⋆ Tolkning: må tenke oss at vi gjentar forsøket uendelig mange ganger

Sannsynlighet

- ⋆ Sannsynlighet: Reell funksjon definert på hendelser, P

  1. $0 \leq P(A) \leq 1$
  2. $P(S) = 1$
  3. Hvis $A_1, A_2, \ldots$ parvis disjunkte:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

  – tenk på sannsynlighet som areal i venndiagram

- ⋆ Tolkning: må tenke oss at vi gjentar forsøket uendelig mange ganger

- ⋆ Uniform sannsynlighetsmodell:
  - $S = \{e_1, e_2, \ldots, e_m\}$
  - $P(\{e_1\}) = P(\{e_2\}) = \ldots = P(\{e_m\})$

$$P(A) = \frac{g}{m} = \frac{\text{antall gunstige}}{\text{antall mulige}}$$

Kombinatorikk/Telleregler

- Multiplikasjonssetningen: $n = n_1 \cdot n_2 \cdot \ldots \cdot n_p$
- Urnemodell: Trekker kuler fra en urne
  - ordnet utvalg eller ikke-ordnet utvalg
  - trekker med tilbakelegging eller trekker uten tilbakelegging
- Inndeling i grupper

Kombinatorikk/Telleregler

- ⋆ Multiplikasjonssetningen: $n = n_1 \cdot n_2 \cdot \ldots \cdot n_p$
- ⋆ Urnemodell: Trekker kuler fra en urne
  - ordnet utvalg eller ikke-ordnet utvalg
  - trekker med tilbakelegging eller trekker uten tilbakelegging
- ⋆ Inndeling i grupper

- ⋆ I formelsamling, nederst på side 32:
  - $n! = 1 \cdot 2 \cdots n$ er antall måter å permutere (ordne) n elementer
  - ${}_nP_r = \frac{n!}{(n-r)!} = n \cdot (n-1) \cdots (n-r+1)$ er antall ordnede utvalg når r elementer velges blant n uten tilbakelegging
  - $\binom{n}{r} = \frac{n!}{r!(n-r)!} = \frac{{}_nP_r}{r!} = {}_nC_r$ er antall ikke-ordnende utvalg når r elementer velges blant n uten tilbakelegging

Regneregler for sannsynlighet

- Additiv regel: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  - A, B disjunkte: $P(A \cup B) = P(A) + P(B)$
- Komplementærsetningen $P(A) = 1 - P(A')$
- Betinget sannsynlighet: $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- Multiplikasjonssetningen: $P(A \cap B) = P(B)P(A|B)$
- Uavhengiget: A, B uavhengige $\Leftrightarrow P(A|B) = P(A)$
  - $P(A \cap B) = P(A) \cdot P(B)$

Regneregler for sannsynlighet

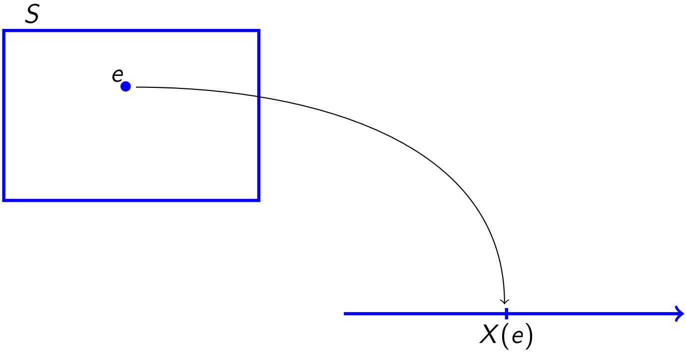
- ⋆ Additiv regel: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  - A, B disjunkte: $P(A \cup B) = P(A) + P(B)$
- ⋆ Komplementærsetningen $P(A) = 1 - P(A')$
- ⋆ Betinget sannsynlighet: $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- ⋆ Multiplikasjonssetningen: $P(A \cap B) = P(B)P(A|B)$
- ⋆ Uavhengiget: A, B uavhengige $\Leftrightarrow P(A|B) = P(A)$
  - $P(A \cap B) = P(A) \cdot P(B)$

- ⋆ Bayes regel: $P(A|B) = \frac{P(A)P(B|A)}{P(B)}$
- ⋆ Setning om total sannsynlighet: $B_1, B_2, \ldots, B_n$ partisjon av S

$$P(A) = \sum_{i=1}^{n} P(B_i \cap A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

Stokastiske variabler og sannsynlighetsfordelinger

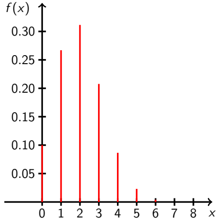
- Stokastisk variabel: Funksjon fra utfallsrom til den reelle tallinja



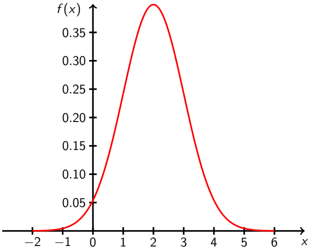
- Diskret og kontinuerlig stokastisk variabel
  - diskret: Endelig eller tellbart uendelig antall mulige verdier
  - kontinuerlig: Ikke-tellbart uendelig mange mulige verdier

$f(x)$: Punktsannsynlighet og sannsynlighetstetthet

- ⋆ Punktsannsynlighet, $f(x) = P(X = x)$

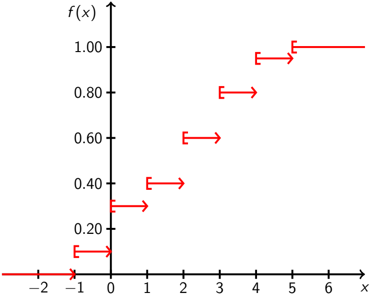


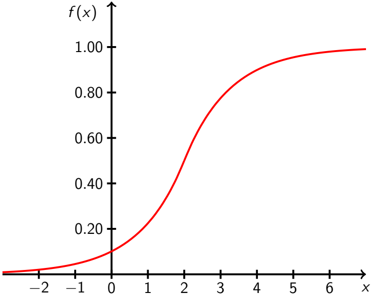
- ⋆ Sannsynlighetstetthet, $f(x)$: $P(a < X \leq b) = \int_a^b f(x)\mathrm{d}x$



$F(x)$: Kumulativ sannsynlighetsfordeling

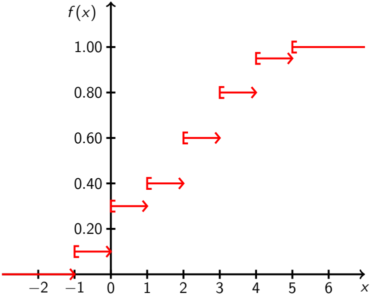
- Kumulativ fordeling, $F(x) = P(X \leq x)$

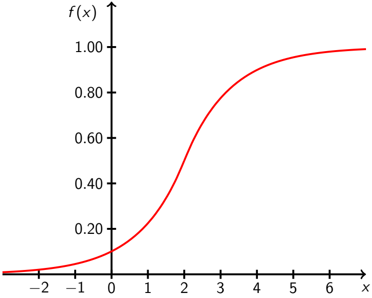




$F(x)$: Kumulativ sannsynlighetsfordeling

- ⋆ Kumulativ fordeling, $F(x) = P(X \leq x)$





- ⋆ Merk: Det er en en-til-en sammenheng mellom $f(x)$ og $F(x)$

Simultanfordeling

- Simultan punktsannsynlighet, $f(x,y)$

$$f(x,y) = P(X = x, Y = y)$$

- Simultan sannsynlighetstetthet, $f(x,y)$

$$P((X,Y) \in A) = \iint_A f(x,y)\mathrm{d}x\mathrm{d}y$$

- Marginalfordeling
  - summerer/integrerer bort variabelen vi ikke er interessert i
- Betinget fordeling

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)}$$

- X og Y er uavhengige stokastiske variabler hvis

$$f_{XY}(x,y) = f_X(x) \cdot f_Y(y)$$

Forventingsverdi og varians

- Forventingsverdi, $\mu = \mathrm{E}[X]$

$$\mathrm{E}[X] = \begin{cases} \sum_x x f(x) & \text{hvis } X \text{ er diskret SV} \\ \int_{-\infty}^{\infty} f(x)\mathrm{d}x & \text{hvis } X \text{ er kontinuerlig SV} \end{cases}$$

  - Tolkning: gjennomsnitt av uendelig mange observasjoner
  - regneregler: $\mathrm{E}[\cdot]$ er en lineær operator

- Varians, $\sigma^2 = \mathrm{Var}[X] = \mathrm{E}\left[(X-\mu)^2\right]$
  - Tolkning: Mål på spredning/variasjon
  - Regneregler:

$$\mathrm{Var}[X] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2$$

$$\mathrm{Var}[aX + b] = a^2 \mathrm{Var}[X]$$

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}[X, Y]$$

- Kovarians og korrelasjon

Viktige diskrete sannsynlighetsfordelinger

- ⋆ Binomisk fordeling
  - n uavhengige forsøk
  - hvert forsøk gir suksess eller fiasko
  - sannsynlighet p for suksess i hvert forsøk
  - X: antall suksess

$$X \sim b(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}; x = 0, 1, \ldots, n$$

  - multinomisk fordeling: k mulige resultat i hvert forsøk

- ⋆ Hypergeometrisk fordeling
  - trekker n kuler fra urne <u>uten</u> tilbakelegging
  - N kuler totalt, k røde, $N-k$ gule
  - X: antall røde kuler trukket

$$X \sim h(x; N, n, k) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}; x = 0, 1, \ldots, \min\{n, x\}$$

Viktige diskrete sannsynlighetsfordelinger

- ⋆ Negativt binomisk fordeling
  - som ved binomisk fordeling, men gjør forsøk inntil k suksesser
  - X: antall trekninger

$$X \sim b^{\star}(x; k, p) = \binom{x-1}{k-1} p^k (1-p)^{x-k}; x = k, k+1, \ldots$$

  - Spesielt: $k = 1$ kalles geometrisk fordeling

$$X \sim g(x; p) = (1-p)^{x-1} p; x = 1, 2, \ldots$$

- ⋆ Poissonfordeling
  - X: antall hendelser i $[0, t]$ i en poissonprosess

$$X \sim p(x; \lambda t) = \frac{e^{-\lambda t}(\lambda t)^x}{x!}; x = 0, 1, \ldots$$

Viktige diskrete sannsynlighetsfordelinger

- Mye finnes i formelsamlinga
  - formel for $f(x)$
  - formel for $\mathrm{E}[X]$
  - formel for $\mathrm{Var}[X]$
  - formel for $M_X(t)$
  - tabell over $F(x) = P(X \leq x)$ (ikke for geometrisk fordeling)

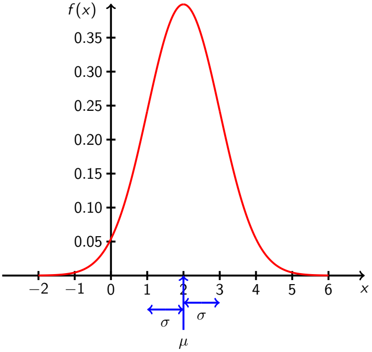
Viktige diskrete sannsynlighetsfordelinger

- ⋆ Mye finnes i formelsamlinga
  - formel for $f(x)$
  - formel for $\mathrm{E}[X]$
  - formel for $\mathrm{Var}[X]$
  - formel for $M_X(t)$
  - tabell over $F(x) = P(X \leq x)$ (ikke for geometrisk fordeling)

- ⋆ Approksimasjoner
  - hypergeometrisk $\approx$ binomisk hvis N er stor i forhold til k
    - intuitivt rimelig, kan vises matematisk
  - binomisk $\approx$ poisson hvis n stor og p liten
    - intuitivt rimelig, kan vises matematisk
  - binomisk $\approx$ normal hvis n stor
    - spesialtilfelle av sentralgranseteoremet

Viktige kontinuerlige sannsynlighetsfordelinger

⋆ Normalfordeling: $X \sim n(x;\mu,\sigma)$



– en lineærkombinasjon av uavhengige normalfordelte variabler er normalfordelt

– standardisering:

$$X \sim n(x;\mu,\sigma) \Leftrightarrow Z = \frac{X-\mu}{\sigma} \sim n(z;0,1)$$

Viktige kontinuerlige sannsynlighetsfordelinger

- Eksponensialfordeling
  - tid til første hendelse i en poissonfordeling/tid mellom etterfølgende hendelser i en poissonfordeling

$$X \sim f(x;\lambda) = \lambda e^{-\lambda x}; x \geq 0$$

- Gammafordeling: $X \sim f(x;\alpha,\beta) = \ldots$
  - $\alpha = 1, \beta = \frac{1}{\lambda}$: eksponensialfordeling
  - $\alpha = \frac{\nu}{2}, \beta = 2$: χ^2_ν-fordeling
  - tid til k-te hendelse i en poissonprosess: $\alpha = k, \beta = \frac{1}{\lambda}$

- χ^2-fordeling

- Student t-fordeling

Viktige kontinuerlige sannsynlighetsfordelinger

- ⋆ Mye finnes i formelsamlinga
  - formel for $f(x)$
  - formel for $\mathrm{E}[X]$
  - formel for $\mathrm{Var}[X]$
  - formel for $M_X(t)$
  - tabell over $F(x) = P(X \leq x)$ (ikke for eksponensialfordeling)

Sentralgrenseteoremet

* Anta $X_1, X_2, \ldots, X_n$ uavhengige og identisk fordelt
* Anta $\mu = \mathrm{E}[X_i]$ og $\sigma^2 = \mathrm{Var}[X]$
* La $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ og

$$Z = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}}$$

* Resultat: Når $n \to \infty$ til fordelingen til Z konvergere mot $n(z; 0, 1)$
* Dette betyr at når n er stor har vi

$$\bar{X} \approx n\left(\bar{x}; \mu, \sqrt{\frac{\sigma^2}{n}}\right)$$

Transformasjoner av stokastiske variabler

- Anta $X \sim f(x)$ og definer Y ved

$$Y = u(X) \Leftrightarrow X = w(X)$$

- Hvis X (og Y) er diskrete stokastiske variabler:

$$g(y) = f(w(y))$$

- Hvis X (og Y) er kontinuerlige stokastiske variabler:

$$g(y) = f(w(y)) \cdot |w'(y)|$$

Transformasjoner av stokastiske variabler

- Anta $X \sim f(x)$ og definer Y ved

$$Y = u(X) \Leftrightarrow X = w(X)$$

- Hvis X (og Y) er diskrete stokastiske variabler:

$$g(y) = f(w(y))$$

- Hvis X (og Y) er kontinuerlige stokastiske variabler:

$$g(y) = f(w(y)) \cdot |w'(y)|$$

- Merk:
  - formelen for kontinuerlig SV står i formelsamlinga
  - alternativt kan man regne ut $G(y) = P(Y \leq y)$ i stedet, og så bruke den til å finne $g(y)$

Ekstremvariabler (og ordningsveriabler)

- Anta $X_1, X_2, \ldots, X_n$ tilfeldig utvalg fra $f(x)$
  - La $V = \max\{X_1, X_2, \ldots, X_n\}$

$$
\begin{aligned}
F_V(v) &= P(V \leq v) = P(\max(\{X_1, X_2, \ldots, X_n\} \leq v) \\
&= P(X_1 \leq v, X_2 \leq v, \ldots, X_n \leq v) \\
&= P(X_1 \leq v) \cdot P(X_2 \leq v) \cdot \ldots \cdot P(X_n \leq v) \\
&= F_X(v) \cdot F_X(v) \cdot \ldots \cdot F_X(v) = (F_X(v))^n
\end{aligned}
$$

Ekstremvariabler (og ordningsveriabler)

- Anta $X_1, X_2, \ldots, X_n$ tilfeldig utvalg fra $f(x)$
  - La $V = \max\{X_1, X_2, \ldots, X_n\}$

$$\begin{aligned} F_V(v) &= P(V \leq v) = P(\max(\{X_1, X_2, \ldots, X_n\} \leq v) \\ &= P(X_1 \leq v, X_2 \leq v, \ldots, X_n \leq v) \\ &= P(X_1 \leq v) \cdot P(X_2 \leq v) \cdot \ldots \cdot P(X_n \leq v) \\ &= F_X(v) \cdot F_X(v) \cdot \ldots \cdot F_X(v) = (F_X(v))^n \end{aligned}$$

  - La $U = \min\{X_1, X_2, \ldots, X_n\}$

$$\begin{aligned} F_U(u) &= P(U \leq u) = P(\min\{X_1, X_2, \ldots, X_n\} \leq u) \\ &= 1 - P(\min\{X_1, X_2, \ldots, X_n\} > u) \\ &= 1 - P(X_1 > u, X_2 > u, \ldots, X_n > u) \\ &= 1 - P(X_1 > u) \cdot P(X_2 > u) \cdot \ldots \cdot P(X_n > u) \\ &= 1 - (1 - F_X(u)) \cdot (1 - F_X(u)) \cdot \ldots \cdot (1 - F_X(u)) \\ &= 1 - (1 - F_X(u))^n \end{aligned}$$

Ekstremvariabler (og ordningsveriabler)

- ⋆ Anta $X_1, X_2, \ldots, X_n$ tilfeldig utvalg fra $f(x)$
  - La $V = \max\{X_1, X_2, \ldots, X_n\}$

$$\begin{aligned}
F_V(v) &= P(V \leq v) = P(\max(\{X_1, X_2, \ldots, X_n\} \leq v) \\
&= P(X_1 \leq v, X_2 \leq v, \ldots, X_n \leq v) \\
&= P(X_1 \leq v) \cdot P(X_2 \leq v) \cdot \ldots \cdot P(X_n \leq v) \\
&= F_X(v) \cdot F_X(v) \cdot \ldots \cdot F_X(v) = (F_X(v))^n
\end{aligned}$$

  - La $U = \min\{X_1, X_2, \ldots, X_n\}$

$$\begin{aligned}
F_U(u) &= P(U \leq u) = P(\min\{X_1, X_2, \ldots, X_n\} \leq u) \\
&= 1 - P(\min\{X_1, X_2, \ldots, X_n\} > u) \\
&= 1 - P(X_1 > u, X_2 > u, \ldots, X_n > u) \\
&= 1 - P(X_1 > u) \cdot P(X_2 > u) \cdot \ldots \cdot P(X_n > u) \\
&= 1 - (1 - F_X(u)) \cdot (1 - F_X(u)) \cdot \ldots \cdot (1 - F_X(u)) \\
&= 1 - (1 - F_X(u))^n
\end{aligned}$$

  - også fordelingen for $X_{(k)}$, den k-te minste av X_i-ene

Momentgenererende funksjoner

- Definisjon: $M_X(t) = \mathrm{E}\left[e^{tX}\right]$
- Regneregler:
  - $M_{X+a}(t) = e^{at} M_X(t)$
  - $M_{aX}(t) = M_X(at)$
  - Hvis $X_1, X_2, \ldots, X_n$ er uavhengige er

$$M_{X_1+X_2+\ldots+X_n}(t) = M_{X_1}(t) \cdot M_{X_2}(t) \cdot \ldots \cdot M_{X_n}(t)$$

- Egenskaper:
  - $M_X^{(r)}(0) = \mathrm{E}[X^r]$
  - X og Y har samme fordeling dersom $M_X(t) \equiv M_Y(t)$

Momentgenererende funksjoner

- Definisjon: $M_X(t) = \mathrm{E}\left[e^{tX}\right]$
- Regneregler:
  - $M_{X+a}(t) = e^{at} M_X(t)$
  - $M_{aX}(t) = M_X(at)$
  - Hvis $X_1, X_2, \ldots, X_n$ er uavhengige er

$$M_{X_1+X_2+\ldots+X_n}(t) = M_{X_1}(t) \cdot M_{X_2}(t) \cdot \ldots \cdot M_{X_n}(t)$$

- Egenskaper:
  - $M_X^{(r)}(0) = \mathrm{E}[X^r]$
  - X og Y har samme fordeling dersom $M_X(t) \equiv M_Y(t)$

- Vi har brukt momentgenererende funksjoner til å vise en del sammenhenger mellom fordelinger
  - lineærkombinasjon av uavhengige normalfordelte variabler er normalfordelt
  - sum av uavhengige poissonfordelte variabler er poissonfordelt

Estimator

- $X_1, X_2, \ldots, X_n$ er et tilfeldig utvalg fra $f(x)$-populasjonen
  - $X_1, X_2, \ldots, X_n$ er uavhengige og $X_i \sim f(x)$ for $i = 1, 2, \ldots, n$

- Observator: En observerbar funksjon av SV er en observator
  - ikke funksjon av ukjente parametre

- Estimator: Observator som brukes til å anslå verdien til ukjent parameter
  - parameter: θ, μ, λ, p
  - estimator: $\widehat{\theta}$, $\widehat{\mu}$, $\widehat{\lambda}$, $\widehat{p}$

Estimator

- ⋆ $X_1, X_2, \ldots, X_n$ er et tilfeldig utvalg fra $f(x)$-populasjonen
  - $X_1, X_2, \ldots, X_n$ er uavhengige og $X_i \sim f(x)$ for $i = 1, 2, \ldots, n$

- ⋆ Observator: En observerbar funksjon av SV er en observator
  - ikke funksjon av ukjente parametre

- ⋆ Estimator: Observator som brukes til å anslå verdien til ukjent parameter
  - parameter: θ, μ, λ, p
  - estimator: $\widehat{\theta}$, $\widehat{\mu}$, $\widehat{\lambda}$, $\widehat{p}$

- ⋆ Merk: En estimator er en stokastisk variabel
  - har sannsynlighetsfordeling
  - har forventningsverdi
  - har varians

Viktige egenskaper for estimatorer

- ⋆ $\widehat{\theta}$ er en forventningsrett estimator for θ dersom

$$\mathsf{E}[\widehat{\theta}] = \theta$$

  - hvis ikke er den forventingsskjev

- ⋆ Foretrekker en estimator som er forventingsrett

  - tenker oss at vi gjentar forsøket uendelig mange ganger

- ⋆ Av flere forv.rette estimatorer foretrekker vi den med minst varians

  - tenker oss at vi gjentar forsøket uendelig mange ganger

Viktige egenskaper for estimatorer

- $\widehat{\theta}$ er en forventningsrett estimator for θ dersom

$$\mathsf{E}[\widehat{\theta}] = \theta$$

  - hvis ikke er den forventingsskjev

- Foretrekker en estimator som er forventingsrett
  - tenker oss at vi gjentar forsøket uendelig mange ganger

- Av flere forv.rette estimatorer foretrekker vi den med minst varians
  - tenker oss at vi gjentar forsøket uendelig mange ganger

- Hvordan bestemme om $\mathsf{Var}[\widehat{\theta}]$ eller $\mathsf{Var}[\widetilde{\theta}]$ er størst?
  - se på $\mathsf{Var}[\widehat{\theta}] - \mathsf{Var}[\widetilde{\theta}]$
  - se på $\mathsf{Var}[\widehat{\theta}]/\mathsf{Var}[\widetilde{\theta}]$
  - merk: kan avhenge av verdien til θ

Sannsynlighetsmaksimeringestimator (SME)

- Idé: Maksimer sannsynligheten for å observere det vi har observert
- Rimelighetsfunksjon: Sannsynligheten for å observere det vi har observert, som funksjon av ukjent parameter θ
  - typisk: $$L(\theta) = f(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$
- For å finne SME må vi maksimere $L(\theta)$ m.h.p. θ
  - typisk: $$l(\theta) = \ln L(\theta) = \sum_{i=1}^{n} \ln f(x_i; \theta)$$
    $$l'(\theta) = 0 \Rightarrow \theta = u(x_1, x_2, \ldots, x_n)$$
    $$\widehat{\theta} = u(X_1, X_2, \ldots, X_n)$$

Sannsynlighetsmaksimeringestimator (SME)

- Idé: Maksimer sannsynligheten for å observere det vi har observert
- Rimelighetsfunksjon: Sannsynligheten for å observere det vi har observert, som funksjon av ukjent parameter θ
  - typisk: $$L(\theta) = f(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$
- For å finne SME må vi maksimere $L(\theta)$ m.h.p. θ
  - typisk: $$l(\theta) = \ln L(\theta) = \sum_{i=1}^{n} \ln f(x_i; \theta)$$
    $$l'(\theta) = 0 \Rightarrow \theta = u(x_1, x_2, \ldots, x_n)$$
    $$\widehat{\theta} = u(X_1, X_2, \ldots, X_n)$$
- Husk: Det "typiske" er ikke alltid korrekt, så man må også tenke!

Konfidensintervall

- Anta $X_1, X_2, \ldots, X_n$ tilfeldig utvalg fra $f(x;\theta)$-populasjonen
- Intervallet $[\widehat{\theta}_L(x_1, x_2, \ldots, x_n), \widehat{\theta}_U(x_1, x_2, \ldots, x_n)]$ er da et $(1-\alpha)\cdot 100\%$ konfidensintervall for θ dersom

$$P(\widehat{\theta}_L(X_1, X_2, \ldots, X_n) \leq \theta \leq \widehat{\theta}_U(X_1, X_2, \ldots, X_n)) = 1 - \alpha$$

Konfidensintervall

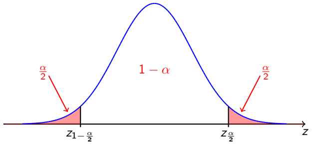
- Anta $X_1, X_2, \ldots, X_n$ tilfeldig utvalg fra $f(x;\theta)$-populasjonen
- Intervallet $[\widehat{\theta}_L(x_1, x_2, \ldots, x_n), \widehat{\theta}_U(x_1, x_2, \ldots, x_n)]$ er da et $(1-\alpha)\cdot 100\%$ konfidensintervall for θ dersom

$$P(\widehat{\theta}_L(X_1, X_2, \ldots, X_n) \leq \theta \leq \widehat{\theta}_U(X_1, X_2, \ldots, X_n)) = 1-\alpha$$

- Hva betyr dette?
  - husk tolkningen av sannsynlighet
  - det er $X_1, X_2, \ldots, X_n$ som er stokastiske, θ er (ukjent) tall
  - må tenke oss at vi gjentar forsøket uendelig mange ganger

Hvordan utlede et konfidensintervall

- ⋆ La $X_1, X_2, \ldots, X_n$ være tilfeldig utvalg fra $f(x;\theta)$-populasjonen
- ⋆ Estimator for θ: $\widehat{\theta}$
- ⋆ Finn en $Z = g(\widehat{\theta}, \theta)$ slik at har en kjent fordeling
- ⋆ Bestemmer $\frac{\alpha}{2}$ og $1 - \frac{\alpha}{2}$-kvantilene



$$P(z_{1-\frac{\alpha}{2}} \leq g(\widehat{\theta}, \theta) \leq z_{\frac{\alpha}{2}}) = 1 - \alpha$$

- ⋆ Løser hver ulikhet m.h.p. θ, setter sammen igjen med θ i midten

$$P(\widehat{\theta}_L(X_1, X_2, \ldots, X_n) \leq \theta \leq \widehat{\theta}_U(X_1, X_2, \ldots, X_n)) = 1 - \alpha$$

- ⋆ Konkluder: $(1 - \alpha) \cdot 100\%$ konfidensintervall for θ er

$$[\widehat{\theta}_L(x_1, x_2, \ldots, x_n), \widehat{\theta}_U(x_1, x_2, \ldots, x_n)]$$

Hvilke konfidensintervall har vi diskutert

- Ett-utvalg normalfordeling
  - Konf.int. for μ, σ^2 kjent: $Z = \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim n(z; 0, 1)$
  - Konf.int. for μ, σ^2 ukjent: $T = \frac{\bar{X}-\mu}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1}$
  - Konf.int. for σ^2, μ ukjent: $V = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$
- To-utvalg normalfordeling
  - Konf.int. for $\mu_1 - \mu_2$, σ_1^2 og σ_2^2 kjente
  - Konf.int. for $\mu_1 - \mu_2$, σ_1^2 og σ_2^2 ukjente
- Ett-utvalg, binomisk fordeling
  - Konf.int. for p
- To-utvalg, binomisk fordeling
  - Konf.int. for $p_1 - p_2$
- Regresjonsmodell: α, β

Prediksjonsintervall

- ⋆ Anta $X_1, X_2, \ldots, X_n$ tilfeldig utvalg fra $f(x;\theta)$-populasjonen. $X^\star$ er en ny observasjon fra samme fordeling uavhengig av de gamle
- ⋆ Intervallet $[X_L^\star(x_1, x_2, \ldots, x_n), X_U^\star(x_1, x_2, \ldots, X_n)]$ er da et $(1-\alpha)\cdot 100\%$ prediksjonsintervall for $X^\star$ dersom

$$P(X_L^\star(X_1, X_2, \ldots, X_n) \le X^\star \le X_U^\star(X_1, X_2, \ldots, X_n)) = 1-\alpha$$

Prediksjonsintervall

- ⋆ Anta $X_1, X_2, \ldots, X_n$ tilfeldig utvalg fra $f(x;\theta)$-populasjonen. $X^\star$ er en ny observasjon fra samme fordeling uavhengig av de gamle
- ⋆ Intervallet $[X_L^\star(x_1, x_2, \ldots, x_n), X_U^\star(x_1, x_2, \ldots, X_n)]$ er da et $(1-\alpha)\cdot 100\%$ prediksjonsintervall for $X^\star$ dersom

$$P(X_L^\star(X_1, X_2, \ldots, X_n) \leq X^\star \leq X_U^\star(X_1, X_2, \ldots, X_n)) = 1-\alpha$$

- ⋆ Hva betyr dette?
  - husk tolkningen av sannsynlighet
  - både $X_1, X_2, \ldots, X_n$ og $X^\star$ er stokastiske
  - må tenke oss at vi gjentar forsøket uendelig mange ganger

Prediksjonsintervall

- Anta $X_1, X_2, \ldots, X_n$ tilfeldig utvalg fra $f(x;\theta)$-populasjonen. $X^\star$ er en ny observasjon fra samme fordeling uavhengig av de gamle
- Intervallet $[X_L^\star(x_1, x_2, \ldots, x_n), X_U^\star(x_1, x_2, \ldots, X_n)]$ er da et $(1-\alpha)\cdot 100\%$ prediksjonsintervall for $X^\star$ dersom

$$P(X_L^\star(X_1, X_2, \ldots, X_n) \leq X^\star \leq X_U^\star(X_1, X_2, \ldots, X_n)) = 1-\alpha$$

- Hva betyr dette?
  - husk tolkningen av sannsynlighet
  - både $X_1, X_2, \ldots, X_n$ og $X^\star$ er stokastiske
  - må tenke oss at vi gjentar forsøket uendelig mange ganger
- Hvordan utlede et prediksjonsintervall
  - Starte med $X^\star - \bar{X}$ (eller noe lignende) og regne tilsvarende som for konfidensintervall

Hypotesetesting

- To hypoteser
  - nullhypotese, H_0: Hypotesen vi skal undesøke om det er grunnlag for å forkaste
  - alternativ hypotese, H_1: Det vi ønsker å undersøke om det er grunnlag for å påstå
- To typer feil

	H_0 er sann	H_1 er sann
Forkast H_0	Type I-feil	Korrekt
Ikke forkast H_0	Korrekt	Type II-feil

- Testobservator, Z: En observator som kan brukes til å formulere en beslutningsregel, med en kjent fordeling når H_0 er riktig
- Beslutningsregel: Forkaster H_0 dersom $Z > k$ (for eksempel)
- Bestemmer kritisk verdi k fra kravet

$$P(\text{Forkast } H_0 | H_0 \text{ er riktig}) = \alpha$$

Hypotesetesting

- To hypoteser
  - nullhypotese, H_0: Hypotesen vi skal undesøke om det er grunnlag for å forkaste
  - alternativ hypotese, H_1: Det vi ønsker å undersøke om det er grunnlag for å påstå
- To typer feil

	H_0 er sann	H_1 er sann
Forkast H_0	Type I-feil	Korrekt
Ikke forkast H_0	Korrekt	Type II-feil

- Testobservator, Z: En observator som kan brukes til å formulere en beslutningsregel, med en kjent fordeling når H_0 er riktig
- Beslutningsregel: Forkaster H_0 dersom $Z > k$ (for eksempel)
- Bestemmer kritisk verdi k fra kravet

$$P(\text{Forkast } H_0 | H_0 \text{ er riktig}) = \alpha$$

- p-verdi: Sannsynligheten for å observere det vi har observert eller noe mer ekstremt, når H_0 er sann

Hypotesetesting ved forkastningsområdemetoden

- Anta $X_1, X_2, \ldots, X_n$ tilfeldig utvalg fra $n(x; \mu, \sigma)$-populasjonen
- Anta $n = 30$ og at observerte verdier gir $\bar{x} = 121.6$ og $s^2 = 8.83^2$
- Ønsker å teste

$$H_0 : \mu = 120 \qquad \text{mot} \qquad H_1 : \mu > 120$$

- Estimator for μ: $\widehat{\mu} = \bar{X}$
- Vi vet at

$$\frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1} \qquad \text{(både under } H_0 \text{ og } H_1\text{)}$$

- Testobservator:

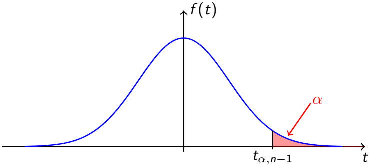
$$T = \frac{\bar{X} - 120}{\sqrt{\frac{S^2}{n}}}$$

- Når H_0 er sann er $T \sim t_{n-1}$

Hypotesetesting ved forkastningsområdemetoden (forts.)

- Testobservator: T
- Vet at $T \sim t_{n-1}$ når H_0 er sann
- Forkastningsregel: Forkaster H_0 dersom $T > k$, der k bestemmes fra kravet

$$P(\text{Forkast } H_0 | H_0 \text{ er sann}) = P(T > k | H_0 \text{ er sann}) = \alpha$$



- Fra figuren ser vi at $k = t_{\alpha,n-1}$
- Forkastningsregelen blir dermed: Forkast H_0 dersom $T > t_{\alpha,n-1}$
- Setter inn tall, og konkluderer

Hypotesetesting ved å regne ut p-verdi

- Anta $X_1, X_2, \ldots, X_n$ tilfeldig utvalg fra $n(x; \mu, \sigma)$-populasjonen
- Anta $n = 30$ og at observerte verdier gir $\bar{x} = 121.6$ og $s^2 = 8.83^2$
- Ønsker å teste

$$H_0 : \mu = 120 \qquad \text{mot} \qquad H_1 : \mu > 120$$

- Estimator for μ: $\widehat{\mu} = \bar{X}$
- Vi vet at

$$\frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1} \qquad \text{(både under } H_0 \text{ og } H_1\text{)}$$

- Testobservator:

$$T = \frac{\bar{X} - 120}{\sqrt{\frac{S^2}{n}}}$$

- Når H_0 er sann er $T \sim t_{n-1}$

Hypotesetesting ved å regne ut p-verdi (forts.)

- Testobservator: T
- Vet at $T \sim t_{n-1}$ når H_0 er sann
- Observert verdi for testobservatoren:

$$t = \frac{\bar{x} - 120}{\sqrt{\frac{s^2}{n}}} = \frac{121.6 - 120}{\sqrt{\frac{8.83^2}{30}}} = 0.992$$

- p-verdien er da: $p = P(T > 0.992 | \text{når } H_0 \text{ er sann}) = 0.1647$

Hypotesetesting — teststyrke

- Anta $X_1, X_2, \ldots, X_n$ tilfeldig utvalg fra $n(x; \mu, \sigma_0)$-populasjonen
- Ønsker å teste: $H_0 : \mu = 3.5$ mot $H_1 : \mu < 3.5$
- Estimator for μ: $\widehat{\mu} = \bar{X}$
- Vi vet at

$$\frac{\bar{X} - \mu}{\sqrt{\frac{\sigma_0^2}{n}}} \sim n(z; 0, 1) \quad \text{(både under } H_0 \text{ og } H_1\text{)}$$

- Testobservator:

$$Z = \frac{\bar{X} - 3.5}{\sqrt{\frac{\sigma_0^2}{n}}}$$

- Når H_0 er sann er $Z \sim n(z; 0, 1)$
- Forkastningskriterium: Forkaster H_0 når $Z < -z_\alpha$
- Spørsmål:
  - Hva blir $P(\text{Forkast } H_0 | \mu = 3.3)$?

Hypotesetesting — teststyrke

- Anta $X_1, X_2, \ldots, X_n$ tilfeldig utvalg fra $n(x; \mu, \sigma_0)$-populasjonen
- Ønsker å teste: $H_0 : \mu = 3.5$ mot $H_1 : \mu < 3.5$
- Estimator for μ: $\widehat{\mu} = \bar{X}$
- Vi vet at

$$\frac{\bar{X} - \mu}{\sqrt{\frac{\sigma_0^2}{n}}} \sim n(z; 0, 1) \quad \text{(både under } H_0 \text{ og } H_1\text{)}$$

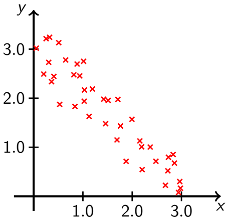
- Testobservator:

$$Z = \frac{\bar{X} - 3.5}{\sqrt{\frac{\sigma_0^2}{n}}}$$

- Når H_0 er sann er $Z \sim n(z; 0, 1)$
- Forkastningskriterium: Forkaster H_0 når $Z < -z_\alpha$
- Spørsmål:
  - Hva blir $P(\text{Forkast } H_0 | \mu = 3.3)$?
  - Hvor stor må n være for at $P(\text{Forkast } H_0 | \mu = 3.3) \geq 0.2$?

Enkel lineær regresjon

- ⋆ Har observasjonspar $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

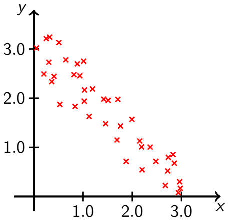


- ⋆ Modell: $Y_1, Y_2, \ldots, Y_n$ er uavhengige og

$$Y_i \sim n(y_i; \alpha + \beta x_i, \sigma)$$

Enkel lineær regresjon

- ⋆ Har observasjonspar $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$



- ⋆ Modell: $Y_1, Y_2, \ldots, Y_n$ er uavhengige og

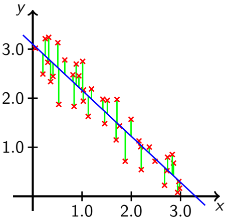
$$Y_i \sim n(y_i; \alpha + \beta x_i, \sigma)$$

- ⋆ Minste kvadraters metode: Finner estimatorer for α og β ved å minimere

$$SSE = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \widehat{\alpha} - \widehat{\beta} x_i \right)^2$$

Enkel lineær regresjon

- ⋆ Har observasjonspar $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$



- ⋆ Modell: $Y_1, Y_2, \ldots, Y_n$ er uavhengige og

$$Y_i \sim n(y_i; \alpha + \beta x_i, \sigma)$$

- ⋆ Minste kvadraters metode: Finner estimatorer for α og β ved å minimere

$$SSE = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \widehat{\alpha} - \widehat{\beta} x_i \right)^2$$

Hva mangler i denne repetisjonen

- Mange detaljer — utledninger
- Regning på konkrete situasjoner
  - de fleste eksamensoppgaver har fokus her

Noen råd før og på eksamen

Noen råd før eksamen

- ⋆ Få oversikt over emnet
  - – mer enn en lang rekke av spesialtilfeller
- ⋆ Regn tidligere eksamensoppgaver
  - – bruk tillatte hjelpemidler
  - – bruk løsningsforslagene fornuftig!
- ⋆ Spør hverandre, kom på spørretimer (gjerne i grupper)
  - – ikke send faglige spørsmål på e-post!
- ⋆ Installer SEB på din maskin og test at det fungerer
  - – hvis du har problemer med SEB, ikke kontakt faglærer
  - – om nødvendig lån en maskin

Noen råd på eksamensdagen

⋆ Husk å ta med tillatte hjelpemidler

⋆ Vær ute i (veldig) god tid

⋆ Alle oppgaver med multiple-choice teller like mye, til sammen 20%

⋆ Alle deloppgaver (1a,1b,1c,2a,2b) på tradisjonelle oppgaver teller 10% hver
  - alle er ikke like vanskelige
  - alle innebærer ikke like mye arbeid

⋆ Faglærere går rundt:
  - det er lov å spørre om hva som helst
  - men dere får ikke svar på hva som helst

⋆ Les oppgaveteksten nøye: Det er forskjell på “skriv opp” og “utled”

⋆ **Skriv mellomregning** slik at det er mulig å forstå hva du har gjort/hvordan du har tenkt — før besvarelsen fornuftig

⋆ Skriv slik at det er mulig å lese håndskriften din!

⋆ Pass på at du besvarer (alle) spørsmålene

Tillatte hjelpemidler på eksamen

- Tabeller og formler i statistikk
- Bestemt, enkel kalkulator
- Et stemplet gult A5-ark med egne håndskrevne formler og notater

Statistikk er mer enn TMA4240

Hva har vi gjort i TMA4240?

- Sannsynlighetsregning:
  - definert sannsynlighet og stokastiske variabler
  - innført regneregler for sannsynligheter og sannsynlighetsfordelinger
  - definert forventning og varians, og sett på regneregler for disse
- Statistisk inferens:
  - har stort sett antatt $X_1, \ldots, X_n$ tilfeldig utvalg (uavhengige og identisk fordelt)
  - punktestimator, konfidensintervall, hypotesetesting
- Hva mer finnes av statistikk?

Hva har vi gjort i TMA4240?

- ⋆ Sannsynlighetsregning:
  - definert sannsynlighet og stokastiske variabler
  - innført regneregler for sannsynligheter og sannsynlighetsfordelinger
  - definert forventning og varians, og sett på regneregler for disse
- ⋆ Statistisk inferens:
  - har stort sett antatt $X_1, \ldots, X_n$ tilfeldig utvalg (uavhengige og identisk fordelt)
  - punktestimator, konfidensintervall, hypotesetesting
- ⋆ Hva mer finnes av statistikk?
  - andre forsøksopplegg, ikke identisk fordelt og/eller avhengige

Hva har vi gjort i TMA4240?

* Sannsynlighetsregning:
  - definert sannsynlighet og stokastiske variabler
  - innført regneregler for sannsynligheter og sannsynlighetsfordelinger
  - definert forventning og varians, og sett på regneregler for disse
* Statistisk inferens:
  - har stort sett antatt $X_1, \ldots, X_n$ tilfeldig utvalg (uavhengige og identisk fordelt)
  - punktestimator, konfidensintervall, hypotesetesting
* Hva mer finnes av statistikk?
  - andre forsøksopplegg, ikke identisk fordelt og/eller avhengige
  - ikke-parametrisk statistikk

Hva har vi gjort i TMA4240?

- Sannsynlighetsregning:
  - definert sannsynlighet og stokastiske variabler
  - innført regneregler for sannsynligheter og sannsynlighetsfordelinger
  - definert forventning og varians, og sett på regneregler for disse
- Statistisk inferens:
  - har stort sett antatt $X_1, \dots, X_n$ tilfeldig utvalg (uavhengige og identisk fordelt)
  - punktestimator, konfidensintervall, hypotesetesting
- Hva mer finnes av statistikk?
  - andre forsøksopplegg, ikke identisk fordelt og/eller avhengige
  - ikke-parametrisk statistikk
  - modeller/problemer hvor man ikke har analytiske svar

Hva har vi gjort i TMA4240?

- Sannsynlighetsregning:
  - definert sannsynlighet og stokastiske variabler
  - innført regneregler for sannsynligheter og sannsynlighetsfordelinger
  - definert forventning og varians, og sett på regneregler for disse
- Statistisk inferens:
  - har stort sett antatt $X_1, \ldots, X_n$ tilfeldig utvalg (uavhengige og identisk fordelt)
  - punktestimator, konfidensintervall, hypotesetesting
- Hva mer finnes av statistikk?
  - andre forsøksopplegg, ikke identisk fordelt og/eller avhengige
  - ikke-parametrisk statistikk
  - modeller/problemer hvor man ikke har analytiske svar
  - bayesiansk statistikk (dere har lært frekventistisk statistikk)

Naturlige neste statistikk-kurs

- TMA4265 Stokastisk modellering
  - sannsynlighetsregning
  - regning på prosesser som utvikler seg i tid, markovegenskap
  - mye kan løses analytisk
- TMA4255 Anvendt statistikk
  - statistisk inferens
  - multippel lineær regresjon, modellvalg, forsøksplanlegging, ikke-parametriske metoder
- TMA4267 Lineære statistiske modeller
  - statistisk inferens
  - mye av det samme som TMA4255, men undervises mer matematisk
- TMA4268 Statistisk læring
  - regresjon og klassifikasjon, modellvalg/regularisering, ikke-linearitet
  - bruker funksjoner i R
  - nytt kurs, undervist første gang vårsemesteret 2018