# Statistikk og Simulering

## Dataingeniørstudiet i Ålesund våren 2018

Hans Georg Schaathun

3. mai 2019

Velkomen til Statistikk og Simulering i dataingeniørstudiet. Desse vevsidene er *eit kompendium* til emnet, og inneheld alle oppgåvene og mange føredrag. Strukturen i kompendiet fylgjer kalenderen, med avsnitt for kvar time, samt avsnitt for for- og etterarbeid mellom timane. Kompendiet er ikkje ein fullstendig presentasjon av kurset; ein må bruka læreboka ved sidan av. Oppsummeringa av førelesingane inneheld berre høgdepunkt.

Kompendiet finst òg i ein PDF-versjon, som du finn nedst på sida, men PDF-versjonen kan vera rotut fordi det primært er skrive for å lesast i ein vevlesar.

Kompendiet er tilgjengeleg frå to uavhengige tenarar:

1. `https://kerckhoffs.hials.no/StatSim/`
2. `http://www.hg.schaathun.net/StatSim/`

Dersom der skulle oppstå tekniske problem med den eine tenaren, er det greitt å ha notert seg den andre.

Kompendiet er forfatta av Hans Georg Schaathun og Siebe van Albada i samarbeid over fleire år. Herverande utgåve er redigert av Hans Georg Schaathun aleine, men mykje av teksta er stadig fellesarbeid.

# 1. Praktisk info

*Statistikk og Simulering* omfattar på sett og vis to tema, men dei to temaa er avhengige av kvarandre. Me skal bare sjå på simulering av tilfeldige prosessar i dette emnet. Statistikken handlar om teoretiske analysar av tilfeldige (gjerne kalla stokastiske) prosessar, og simulering vil handla om tilfeldige prosessar på datamaskina. Det er viktig å sjå samanhengen mellom teori og praksis, og me vil ofte studera dei same problemstillingane parallelt som statistikk og som simulering. Det er difor viktig å halda seg à jour med båe øvingsopplegga (rekneøving og lab). Dersom ein heng etter i det eine, kan det få konsekvensar for det andre.

Me har tre undervisingsøkter:

**Førelesing** Onsdag 8-10. Hovudsakleg statistikk.

**Rekneøving** Fredag 8-10.

**Labøving** Fredag 10-14. Praktisk på datalab. Hovudsakleg simulering.

Dersom du står fast, spør med ein gong. Læring er ein dialog.

## 1.1. Nota Bene

**Eksamen** Emnet har gått nokre år, men ein skal ikkje leggja for stor vekt på gamle eksamensoppgåver. Eg skal freista å leggja større vekt på simulering på eksamen, og ein må rekna med oppgåver som føreset at dei obligatoriske arbeidskrava er gjort og gjort godt.

Dette skal eg koma tilbake til etter kvart. Spør gjerne, men helst ikkje før i mars, når eg har tenkt meir på saka.

**Emneskildring** `https://www.ntnu.no/studier/emner/IR201812#tab=omEmnet`

**Arbeidsbyrde** Modulen er verd 10 studipoeng (ECTS credits), som svarer til ein normert arbeidsbyrde på 250-300 timar. Avhengig av eksamensdatoen, gjev dette ei arbeidsveke på 16-20 effektive studietimar. Det er like greitt å plotta det i timeplanen med ein gong.

**Læremateriellet** Alt læremateriellet ligg på desse vevsidene. Det er omfatta av åndsverkslova. All eigen, personleg bruk er lov. *All vidareformidling av materialet er forbudt* (utan etter avtale).

**Kalender** Me skal ha fjorten veker organisert undervising totalt. Me tek undervisingsfri i vinterferie på grunnskulen (25. februar–1. mars), samt 29. mars. Desse dagane bør ein setja av til å koma à jour med dei større prosjekta. Påskeveka er ferie, men det seier seg sjølv.

## 1.2. Pensum

Frede Frisvold and Jan Gunnar Moe: *Statistikk for Ingeniører*

Alt innhald frå kompendiet (desse sidene), undervisingstimane og øvingane er pensum. Heile statistikkboka er pensum med nokre unntak.

- Delkapittel 7.8 og 8.8 er kursorisk.
- Kapittel 15 er ikkje pensum.
- Bilag A er ikkje pensum.
- Nokre fordelingar (frå kapittel 7-8) er viktigare enn andre.

- Uniform fordeling, binomialfordelinga, normalfordelinga, $t$-fordelinga og $\chi^2$-fordelinga er dei viktigaste fordelingane å kjenna.
- Dei andre fordelingane er ikkje like sentrale som dei fem som nemnde over, men eg vil vurdera å førelesa éi eller to av dei.
- Fordelingar som ikkje er diskutert korkje i førelesing eller i øvingsmaterialet er kursorisk pensum.

### 1.2.1. Annan anbefalt literatur

**Barnes and Kölling:** *Objects First with Java. A Practical Introduction Using BlueJ*

Although focusing on a more fundamental topic, Barnes and Kölling uses a predator/prey model in one extended example. Thus the book may be useful both to refresh object-oriented modelling and programming skills, and as a source of inspiration when implementing predator/prey models.

(Page references are given wrt the fifth edition.)

**Angela B. Shiflet and George W. Shiflet:** *Introduction to Computational Science: Modeling and Simulation for the Sciences*

We will draw many ideas for simulation projects from this book, which has a much wider scope than the module. It may be useful to have this book to cover theory concerning simulation.

**Gouri K. Bhattacharyya and Richard A. Johnson:** *Statistical Concepts and Methods*

This is a classic, introductory textbook in statistics, more thorough than Frisvold and Moe. It is a useful alternative to the corebook if you either want more depth or prefer English over Norwegian.

# 2. Oblig

Labøvingane omfattar 4–6 prosjekt som skal leverast inn. Me arbeider med prosjekta i labøvingane på fredagar, og kvart prosjekt er delt opp i øvingar som svarer til øktene.

| # | Tema | Veker | Språk (råd) | Innl.-frist | Siste frist |
|---|---|---|---|---|---|
| 1 | Stokastisk simulering (5.5) | 2–4 | Matlab | 29. jan 23:59 | 14. feb 15:00 |
| 2 | *Predator-Prey* og agent-basert modellering (6.7) | 5, 6, 10 | Java | 14. mars 23:59 | 29. mars 15:00 |
| 3 | Estimering av feilsann-syn | 7–8 | Matlab | 5. mars 23:59 | 29. mars 15:00 |
| 4 | Diffusjon | 11–12 | | 2. april 23:59 | 24. april 23:59 |
| 5 | Ymse statistikk | 14,15,17 | Matlab | 29. april 23:59 | 5. mai 23:59 |

Alle innleveringane er individuelle, men på alle prosjekta gjeld at de *bør* eller *skal* arbeida i gruppe. De finn grupper sjølv; det har me tid til å gjera på labben. Sjølv om ein arbeider i gruppe, skal kvar medlem individuelt gjera greia for konklusjonane og kva som er lært.

Dei fleste prosjektoppgåvene inneber modellering eller andre vurderingar som krev skjønn. Der er ingen fasit, og det er viktig å diskutera alternative tolkingar med medstudentane.

## Om fristane

For kvart prosjekt vert der sett ein *innleveringsfrist* og ein *siste frist*. Dei som leverer innan innleveringsfristen skal få reletivt kjapp tilbakemelding, og dersom arbeidet vert underkjend får dei ny sjanse før siste frist. Dersom arbeidet ikkje er levert til godkjend standard innan «siste frist», får ein *ikkje ta eksamen*.

Merk at meininga er at alle skal levera innan innleveringsfristen. Dersom ein hevdar sjukdom eller andre spesielle omstende, må ein *dokumentera* både at ein ikkje kunne klara den opp-rinnelege innleveringsfristen pga. gyldig fråver, og også at utsetjing til «siste frist» ikkje er tilstrekkeleg.

## Prosjektinnleveringa

De skal ikkje levera inn alt de har gjort i prosjektet. I staden har eg vald ut nokre få spørs-mål som de skal svare på. (Sjå eige avsnitt under kvart prosjekt.) På dei spørsmåla vil eg til gjengjeld ha godt gjennomtenkte og grunngjevne svar.

Rapportane skal leverast i BlackBoard.

*Presentasjon.* I tillegg til rapporten skal sjølve implementasjonen og køyringa demonstrerast personleg på labben. Ein kan gjera dette før ein leverer rapporten, for å diskutera dei siste detaljane før ein konkluderer; då treng ein ikkje vera heilt ferdig. Eller ein kan demonstrera på ei av dei to fyrste øvingane etter innlevering.

Ein gong i løpet av semesteret skal ein demonstrera prosjektet sitt for heile klassa, dvs. eitt av prosjekta. Me bruker den siste halvtimen av økta til dette; 5 min. kvar.

Det må gjerne samarbeida om modellering, implementasjon og simulering i gruppene, men kvar student skal stå inne for det arbeidet som vert levert inn og kunna demonstrera simuleringane sjølv. Refleksjonen skal vera ut frå eiga læring og eigen ståstad.

# 3. Statistikk

## 3.1. Veke 2. Stokastiske hendingar og variablar

Målet denne veka er

1. å forstå kvifor me treng studera statistikk og simulering, og kvar dette vert brukt.
2. ei intuitiv forståing av fylgjande omgrep:
    - Tilfeldig
    - Eksperiment (forsøk)
    - Utfall og utfallsrom
    - Stokastisk variabel
    - Uniform fordeling
    - Hending
    - Sannsyn (sannsynlegheit)
    - Fordeling og kummulativ fordeling
3. sjå korleis me kan simulera stokastiske hendingar i Matlab.
4. sjå kva ein kan venta i gjennomsnitt når stokastiske eksperiment vert gjentekne.

### 3.1.1. Lesestoff og heimearbeid

**Les 1 (Repetisjon)** *Frå Frisvold og Moe: Kapittel 1.*

**Les 2 (Introduksjon)** *Frå Frisvold og Moe: Kapittel 2.*

**Les 3 (Sannsynsrekning)** *Frå Frisvold og Moe: §3.1-3.7*

Skaff deg tilgang til MATLAB *før fredag*. Sjå innsida. Du må i alle fall *prøva* å installera før labøvinga. Dersom du har problem med installasjonen, ver snill å gje meg melding **før** fredag (hasc@ntnu.no). Eg kan hjelpa med installasjon på linux; spørsmål om installasjon på Mac OS og Windows må rettast til Orakel.

### 3.1.2. Onsdag (Førelesing)

Me skal bruka quiz i mange av førelesingane.

1. Opna `https://moodle.uials.no`
2. Du må laga deg ein konto og logga inn. Du kan bruka Google, eller laga eigen konto på Moodle.
3. Vel kurset «Statistikk og Simulering».
4. Finn riktig veke og åpna aktiviteten «Tilfeldige og Vilkårlege tal».

#### Tilfeldige og vilkårlege tal

**Quiz** Tilfeldige tal frå 1 til 10.

**Oppgåve 3.1 (Diskusjon)** *Diskuter kva det vil seia at eit tal er* tilfeldig. *Bruk quizzen som døme.*

Terningkastet er eit typisk døme på eit *eksperiment* eller *forsøk* i statistikk. Dette forsøket har ti moglege *utfall*, $1, 2, \ldots, 10$. Mengda av moglege utfall kaller me for *utfallsrommet*

$$\{1, 2, \ldots, 10\}$$

Me kan tenkja på talet som terninga viser som ein variabel. Før me kastar terninga er dette ein *stokastisk variabel*. Me veit ikkje kva verdi variabelen har, men han har ein sannsynsfordeling. Når me har kasta og lest av verdien, er ikkje variabelen lenger stokastisk; i staden har me ein *observasjon* av den stokastiske variabelen.

**Oppgåve 3.2 (Diskusjon)** *Kva meiner me med sannsyn?*

Stokastiske variablar (og forsøk) har ei sannsynsfordeling, som er ein funksjon som tilordnar eit sannsyn (i intervallet $[0, 1]$) til kvart mogleg utfall frå utfallsrommet. For diskrete stokastiske variablar kan me visa sannsynsfordelinga som ein tabell eller eit histogram.

#### Fleire forsøk

**Quiz** Mynt og kron. Eit og to kast.

**Oppgåve 3.3** *Me kastar mynt og kron to gongar, der utfalla er mynt-mynt, mynt-kron, kron-mynt og kron-kron. Kva er sannsynsfordelinga?*

Når me kastar ein rettferdig mynt fleire gongar er kasta *uavhengige*. Uansett kva som skjer i fyrste kast, so er sannsynsfordelinga 50-50 i neste kast.

**Definisjon 1 (Hending)** *Ei hending er eit eller anna vilkår som anten skjer eller ikkje skjer i eit forsøk (eller serie av forsøk). Me kan definera ei hending som ei delmengd av utfallsrommet.*

**Definisjon 2** *Me seier at eit forsøk (eller ein stokastisk variabel) har uniform sannsynsfordeling dersom alle utfalla er like sannsynlege.*

**Oppgåve 3.4** *Me kastar mynt og kron to gongar og observerer kor mange gongar me får kron. Kva er sannsynsfordelinga?*

**Oppgåve 3.5** *Du har ein vanleg stokk med spelkort (52 kort), og deler ei pokerhand (fem kort). Kva er sannsynet for å få royal straight flush (10-Kn-D-K-E i same farge)?*

**Merknad 1** *I tilfellet at alle utfall har lik sannsynlighet, kan sannsynligheten av en hendelse regnes ut som:*

$$\mathsf{P}(\textit{hendelse}) = \frac{\textit{antall gunstige utfall}}{\textit{antall mulige utfall}} \tag{1}$$

*hvor et utfall er "gunstig" om det ligger i hendelsen.*

**Oppgåve 3.6** *Me kastar mynt og kron tre gongar. Kva er sannsynet for å få*

1. *mynt på begge dei to fyrste kasta (Hending $A$)?*
2. *kron på begge dei siste to kasta (Hending $B$)?*
3. *mynt på dei to fyrste og kron på dei to siste (Hending $A \cap B$)?*
4. *mynt på begge dei siste to kasta (Hending $C$)?*
5. *mynt på dei to fyrste og mynt på dei to siste (Hending $A \cap C$)?*

**Definisjon 3** *To hendingar $A$ og $B$ er gjensidig utelukkande (disjunkte) når det er uråd at båe skjer. Mao., dersom $A$ skjer, so kan ikkje $B$ skje og vice versa.*

Om to hendelser er gjensidig utelukkende, er sannsynligheten for at begge inntreffer lik $0$:

$$\mathsf{P}(A \cap B) = 0 \tag{2}$$

Me kan sjå på eit utfall som ei hending, men to moglege utfall er alltid gjensidig utelukkande.

## Uavhengige og avhengige hendingar

**Oppgåve 3.7 (Diskusjon)** *Vi har en hvit og en svart terning (D6) og kaster begge samtidig. Hvordan kan vi regne ut sannsynligheten for at den hvite terningen viser 3 øyne og den svarte et partall?*

**Quiz** To kular

**Oppgåve 3.8 (Diskusjon)** *Vi har en pose med 2 hvite og 2 svarte kuler. Vi drar først en tilfeldig kule, legger den* ikke *tilbake, og velger igjen en tilfeldig kule. Hva er sannsynligheten for at vi velger to hvite på rad?*

### Samansette forsøk og hendingar

**Oppgåve 3.9 (Drøfting)** *Ein analytikar estimerer*

- *30% sannsyn for at rentenivået går opp i 2019.*
- *20% sannsyn for at bustadprisane fell i 2019.*
- *15% sannsyn for at rentenivået går opp* og *bustadprisane fell.*

*Kva sannsyn meiner han at det er for at bustadprisane fell utan at rentenivået går opp?*

### 3.1.3. Fredag (rekneøving)

**Oppgåve 3.10** *Me slår mynt og kron to gongar, og registrerer fylgjande hendingar:*

- *0 gongar mynt*
- *1 gongar mynt*
- *2 gongar mynt*

*Svar på fylgjande:*

1. *Kva er sannsynet for kvar av hendingane?*
2. *Er hendingane gjensidig utelukkande? Kvifor/kvifor ikkje?*
3. *Illustrer sannsynsfordelinga med eit søylediagram.*

**Oppgåve 3.11** *Oppgaver fra Frisvold og Moe: 3.1, 3.4, 3.6, 3.8*

**Oppgåve 3.12 (Ekstra)** *Oppgaver fra Frisvold og Moe: 3.9, 3.10, 3.5.*

**Oppgåve 3.13 (Ekstra)** *Vi kaster tre terninger, én ad gangen. Hva er sannsynligheten at:*

1. *summen av de første to terningene er mindre enn 5 (hendelse $A$)?*
2. *summen av de siste to terningene er større enn 9 (hendelse $B$)?*
3. *summen av de første to terningene er mindre enn 5, mens summen av de siste to terningene er større enn 9 (hendelse $A \cap B$)?*

## 3.2. Veke 3. Betinga sannsyn

Målet denne veka er

1. forstå og kunna rekna med fylgjande omgrep
    - sannsyn

- betinga sannsyn
- sannsynsfordeling og punktsannsyn
- gjennomsnitt

2. forstå skilnaden mellom og samanhengen mellom populasjon og utval
3. vita korleis tilfeldige tal og stokastiske prosessar vert simulert på ei datamasin

### 3.2.1. Lesestoff

**Les 4 (Sannsynsrekning)** *Frå Frisvold og Moe: Kapittel 3 (f.o.m §3.5)*

**Les 5 (Diskrete stokastiske variablar)** *Frå Frisvold og Moe: Kapittel 4 (ikkje §4.2)*

### 3.2.2. Onsdag (førelesing)

#### Betinga sannsyn

**Oppgåve 3.14 (Drøfting)** *Me tek utgangspunkt i siste oppgåve frå forrige førelesing (Oppgåve 3.9). Ein analytikar estimerer*

- *30% sannsyn for at rentenivået går opp i 2019.*
- *20% sannsyn for at bustadprisane fell i 2019.*
- *15% sannsyn for at rentenivået går opp* og *bustadprisane fell.*

*Sett at me veit at rentenivået går opp. Kva kan me då seia om sanssynsfordelinga for om bustadprisane går opp eller ned?*

*Kva om me veit at rentenivået ikkje går opp?*

#### Diskrete Stokastiske Variablar

**Døme 1** *Me kjenner den vanlege terninga, D6, med utfallsrom* $U_1 = \{1, 2, 3, 4, 5, 6\}$.

*Somme spel bruker terningar med andre utfallsrom, som t.d.*

$$U_2 = \{\textit{raud}, \textit{grøn}, \textit{blå}, \textit{gul}, \textit{fiolett}, \textit{oransje}\}$$

*Båe terningane har mange av dei same eigenskapane, og spesielt er dei uniformt fordelte, men* $U_1$ *har mange ekstra eigenskapar fordi* $U_1 \subset \mathbb{R}$.

#### Andre sannsynsfordelingar

**Oppgåve 3.15** *Sjå på eit kast med to vanlege terningar (2D6). Sant/Usant: Talet på augo totalt på terningane er uniformt fordelt.*

**Quiz** Forventingsverdi 2D6.

**Oppgåve 3.16** *Kva regel har de brukt for å koma fram til* gjennomsnittet *for 2D6 (over)?*

### 3.2.3. Fredag (rekneøving)

**Oppgåve 3.17 (Ekstra)** *Frisvold og Moe: Oppgåve 3.11, (3.12),*

**Oppgåve 3.18** *Frisvold og Moe: 3.14, 3.16, 3.18, 3.22*

**Oppgåve 3.19 (Diskusjon)** *Frisvold og Moe: Oppgåve 3.13*

*Det er sterkt tilrådd å løysa denne oppgåva i gruppe. Prøv å visualisera hendingane og moglege kombinasjonar av hendingar, og drøft alternative situasjonar. Ikkje leit etter ein algebraisk mønsterløysing.*

**Oppgåve 3.20** *Oppgåver 4.3 frå Frisvold og Moe*

**Oppgåve 3.21** *Frisvold og Moe: Oppgave 5.1*

**Oppgåve 3.22 (Ekstra)** *Oppgåver 4.1 frå Frisvold og Moe*

## 3.3. Veke 4. Snitt og spredning

Målet denne veka er å forstå spreidningsmåla varians og standardavvik. Me skal kunna rekna varians og standardavvik både for hand og vha. Matlab, og me skal forstå kva standardavviket fortel oss om ei sannsynsfordelinga og kva me me ventar å sjå i gjentakne forsøk.

### 3.3.1. Lesestoff

**Les 6 (Forventingsverdi)** *Frå Frisvold og Moe: Kapittel 5 og 6.1–6.2.*

### 3.3.2. Onsdag (førelesing)

**Innleiiande døme** Me skal sjå på to ulike eksperiment.

1. Du kastar ei terning (D6). Lat resultatet vera den stokastiske variabelen $X$.

2. Du kastar to terningar (2D6) og deler på to. Lat resultatet vera den stokastiske variabelen $Y$.

**Oppgåve 3.23** *Kva er utfallsrommet for $X$?*

**Oppgåve 3.24** *Kva er utfallsrommet for $Y$?*

**Oppgåve 3.25** *Kva er gjennomsnittet for $X$? Dette vert òg kalt forventingsverdien $E(X)$.*

**Oppgåve 3.26** *Kva er gjennomsnittet for $Y$? (Dvs. forventingsverdien $E(Y)$.)*

**Oppgåve 3.27** *Skisser sannsynsfordelingane for $X$ og $Y$ som histogram. Kva skilnader ser du mellom dei to fordelingane? Korleis kan du forklara dei?*

## Populasjonsvariansen og -standardavviket

**Oppgåve 3.28** *Rekn ut variansen for $X$ og $Y$.*

**Definisjon 4 (Populasjonsvarians)** *Populasjonsvariansen for ein variabel $X$ med moglege utfall $x_1, x_2, \ldots, x_n$ er definert som*

$$\sigma^2 = \sum_{i=1}^{n} \mathsf{P}(X = x_i)(x_i - \bar{x})^2. \tag{3}$$

*Standardavviket er $\sigma = \sqrt{\sigma^2}$.*

Legg merke til at $(x_i - \bar{x})$ er avstanden mellom eit utfall og gjennomsnittet. Ved å kvadrera får me alltid eit positivt tal. Stor spreidning i utvalet tyder at utfall med stor kvadratavvvik er (relativt) hyppige, og variansen vert stor.

**Definisjon 5 (Populasjonsstandardavvik)** *Kvadratroten av variansen,*

$$\sigma = \sqrt{\sum_{i=1}^{n} \mathsf{P}(X = x_i)(x_i - \bar{x})^2} \tag{4}$$

*vert kalt for standardavviket.*

**Oppgåve 3.29** *Kva er standardavviket for $X$ og for $Y$?*

## Populasjon og utval (repetisjon)

### Utvalsvariansen

**Definisjon 6 (Utvalsvarians)** *Utvalsvariansen for observasjonane $x_1, x_2, \ldots, x_n$ er definert som*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{5}$$

Legg merke til at $(x_i - \bar{x})$ er avstanden mellom observasjonen og gjennomsnittet. Ved å kvadrera får me alltid eit positivt tal. Stor spreidning i utvalet tyder at mange av desse kvadratavstandane er store. Det er forvirrande at me deler på $n-1$. Dersom hadde delt på $n$, so hadde me sagt at $s^2$ er gjennomsnittet av kvadratavstandane, men det viser seg at $n-1$ gjev eit betre mål.

**Definisjon 7 (Utvalsstandardavvik)** *Kvadratroten av variansen,*

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{6}$$

*vert kalt for standardavviket.*

Kast terningane fem gongar ($n = 5$), slik at du får fem observasjonar av $X$ og fem av $Y$.

**Oppgåve 3.30** *Rekna ut variansen og standardavviket for utvalet $x_1, x_2, \ldots, x_5$.*

**Oppgåve 3.31 (Socrative)** *Kva er variansen for $y_1, y_2, \ldots, y_5$.*

**Oppgåve 3.32 (Socrative)** *Kva er standardavviket for $y_1, y_2, \ldots, y_5$.*

På same måte som me skil mellom utvals- og populasjonsgjennomsnitt, so skil me òg mellom utvals- og populasjonsvarians, og tilsvarande for standardavvik.

**Merknad 2** *Legg merke til at $Y$ er ein lineær kombinasjon av $X$:*

$$Y = \frac{1}{2}X_1 + \frac{1}{2}X_2$$

*der $X_1$ og $X_2$ er to uavhengige variablar (to terningar) med same fordeling som $X$.*

$$\mathsf{var}(Y) = \frac{1}{4}\mathsf{var}(X_1) + \frac{1}{4}\mathsf{var}(X_2) = \frac{1}{2}\mathsf{var}(X)$$

**Merknad 3** *Det fylgjer av merknaden over at*

$$\mathsf{S.Dev.}(Y) = \frac{1}{\sqrt{2}}\mathsf{S.Dev.}(X)$$

**Oppsummering omgrep** Definer omgrepa

1. Punktsannsyn
2. Fordelingsfunksjon (kummulativ sannsynsfordeling)
3. Kvifor bruker me utval?
4. Deskriptiv statistikk og statistisk inferens

### Utvalssannsyn (???)

### 3.3.3. Fredag (rekneøving)

**Oppgåve 3.33** *Frisvold og Moe: Oppgave 5.2, 5.3*

**Oppgåve 3.34** *Frisvold og Moe: Oppgave 5.6, 5.8*

**Oppgåve 3.35** *Frisvold og Moe: Oppgave 5.13*

**Oppgåve 3.36 (Diskusjon)** *Frisvold og Moe: Oppgave 6.1–2*

**Oppgåve 3.37 (Ekstra)** *Frisvold og Moe: Oppgave 5.17*

**Oppgåve 3.38 (Ekstra)** *Oppgåver frå Frisvold og Moe: 5.4, 5.5*

**Oppgåve 3.39 (Ekstra)** *Oppgåver frå Frisvold og Moe: 4.2, 4.4, 4.5*

**Oppgåve 3.40 (Ekstra)** *Frisvold og Moe: Oppgave 6.10 og 6.13*

## 3.4. Veke 5. Feilsannsyn og binomialfordeling

Målet denne veka er å forstå binomialfordelinga, og korleis ho kan brukast til å modellera feilsannsyn i t.d. kommunikasjonssystem.

Merk at me berre går gjennom nokre få fordelingar i denne omgangen. Foruten den uniforme fordelinga ser me på éi diskret fordeling denne veka – binomialfordelinga – og éi kontinuerleg fordeling – normalfordelinga – neste veke. Der er fleire fordelingar i kapittel 7–8, og me vil koma tilbake til nokre av dei seinare.

Ein kan sjå statistikkpensumet som to delar som på mange måtar er uavhengige av kvarandre.

- Sannsynsfordelingar (eller modellar)
- Metodar (estimering, hypotesetest, regresjon)

Metodane er det viktig å læra og forstå. Det krev litt tid og modning å læra å tolka resultata rett. Ein må ha ei fordeling å bruka metoden på, men ein må ofte slå opp nye fordelingar når ein møtar nye problem i praksis. Det er derimot enklare enn å læra metodane. Difor går me i gang med metodane so snart me har nokre få fordelingar å bruka dei på.

### 3.4.1. Lesestoff

**Les 7 (Repetisjon)** *Frå Frisvold og Moe: Kapittel 7.1 (med intro)*

**Les 8 (Binomialfordelinga)** *Frå Frisvold og Moe: Kapittel 7.4–7.5*

### 3.4.2. Onsdag (førelesing)

#### Bernoulli-fordelinga

**Døme 2** *Mobiltelefonen sender éi datapakke til basestasjonen. Der er alltid støy på lina, og det mottekne signalet er aldri heilt identisk med det som vart sendt, men systemet bruker feilrettande kodar slik at basestasjonen* sannsynlegvis *kan rekonstruera pakka.*

*Der er like fullt to moglege utfall: korrekt overføring eller feil. Me kan tala om eit feilsannsyn* $\pi$.

*Dette er eit døme på eit Bernoulli-forsøk.*

Andre døme:

- Mynt og kron
- Du sender éin bit på ein kommunikasjonskanal. Bitten vert overført anten rett eller feil.
- Produksjonsfeil. Kvar eining produsert har eit sannsyn $\pi$ for å vera defekt.

**Oppgåve 3.41** *Lat 0 og 1 vera utfalla i ei Bernoulli-fordeling med punktsannsyn* $\pi = P(X = 1)$. *Finn* $\mu$ *og* $\sigma^2$.

#### Binomialfordelinga

**Døme 3** *I løpet av ei telefonsamtale må mobiltelefonen senda* $N$ *pakker til basestasjonen. Lat oss gå ut frå at pakkene representerer uavhengige Bernoulli-forsøk med sannsyn* $\pi$ *for feil. Lat* $Y$ *vera talet på feil over* $N$ *pakker.*

*Den stokastiske variabelen* $Y$ *er no binomialfordelt med* $N$ *forsøk og punktsannsyn* $\pi$.

Andre døme

- $N$ myntkast
- Produksjonsfeil. Kor mange einingar må kasserast når fabrikken lagar $N$ stk.?
- Kor mange bitfeil får du når du sender eit ord med $N$ bits over ein kanal (t.d. BSC$(p)$)

**Oppgåve 3.42** *Du kastar mynt og kron fire gongar. Kva er sannsynet for å få mynt eksakt ein gong (og kron ein gong)?*

**Oppgåve 3.43** *Du kastar mynt og kron fire gongar. Kva er sannsynet for å få mynt eksakt to gongar (og kron to gongar)?*

**Oppgåve 3.44** *Du kastar mynt og kron $n$ gongar. Kva er sannsynet for å få mynt eksakt $k$ gongar?*

### Lineærkombinasjonar

Lat

$$X = X_1 + X_2 + \ldots + X_n$$

vera summen av $n$ uavhengige stokastiske variablar. Forventingsverdien er då gjeve som

$$\mathsf{E}(X) = \mathsf{E}(X_1) + \mathsf{E}(X_2) + \ldots + \mathsf{E}(X_n).$$

Lat

$$X = X_1 + X_2 + \ldots + X_n$$

vera summen av $n$ uavhengige stokastiske variablar. Variansen er då gjeve som

$$\mathsf{var}(X) = \mathsf{var}(X_1) + \mathsf{var}(X_2) + \ldots + \mathsf{var}(X_n).$$

Lat $X = b \cdot X_1$ vera resultatet av å observera $X_1$ éin gong og ganga med skalaren $b$. Forventingsverdien er då gjeve som

$$\mathsf{E}(X) = b\mathsf{E}(X_1).$$

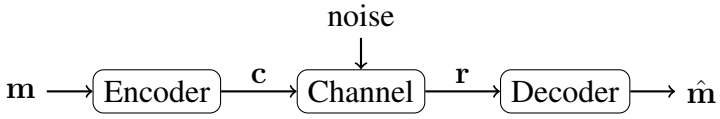
Lat $X = b \cdot X_1$ vera resultatet av å observera $X_1$ éin gong og ganga med skalaren $b$. Variansen er då gjeve som

$$\mathsf{var}(X) = b^2\mathsf{var}(X_1).$$

Når me sender eit $n$-bits ord over BSC$(p)$ er feiltalet $X \sim B(n, p)$ binomialfordelt med $n$ forsøk og suksessannsyn $p$. Merk at $X$ er summen av $n$ uavhengige forsøk,

$$X = X_1 + X_2 + \ldots + X_n,$$

der kvar $X_i$ er fordelt som $Z \sim B(1, p)$ i oppgåva over. Då kan me bruka desse to satsane.




Figur 1: Communication system.

**Oppgåve 3.45** *Tenk deg at du sender eit $n$-bits ord over BSC$(p)$. Lat $X \sim B(n,p)$ vera talet på bitfeil.*

1. *Kva er forventingsverdien (populasjonsgjennomsnittet) $E(X)$?*
2. *Kva er variansen $\mathsf{var}(X)$?*

*Bruk resultatet ditt frå oppgåve 3.56 og dei to satsane over for å koma fram til svaret.*

**Oppgåve 3.46** *Lat $X \sim B(n,p)$ vera fordelt som i oppgåva over. Finn standardavviket $\sigma$ for $X$. Bruk svaret frå oppgåva over.*

Somme tider ser me på $Y' = Y/N \in [0,1]$ heller enn på $Y$.

**PDF og CDF** For store verdiar av $N$ vert utfallsrommet so stort at ein i praksis reknar som om $Y$ er kontinuerleg.

Det vert òg upraktisk å rekna ut punktsannsyn, og me går over til å bruka CDF og tabellar.

### 3.4.3. Utvida døme: Kommunikasjonssystem

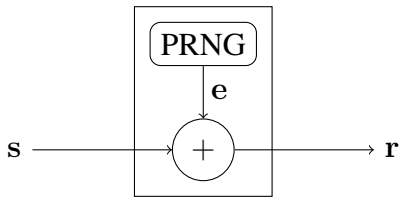
**Ein kommunikasjonsmodell eller to** All digital kommunikasjon er offer for støy. Dvs. at det mottekne signalet ikkje er identisk med det sendte signalet. Kodeteori er løysinga på dette problemet. Ved å koda meldingane er det som regel råd å finna (estimera) den korrekte bodskapen, sjølv om der er feil i overføringa.

Figuren viser dette. Meldinga vert koda som kodeordet $\mathbf{c}$ før han vert sendt på kanalen. Det mottekne signalet $\mathbf{r}$ som kjem ut frå kanalen er sjelden identisk med $\mathbf{c}$, men når me dekodar so får me eit estimat $\hat{\mathbf{m}}$ for den opprinnelege meldinga $\mathbf{m}$. Det store spørsmålet, som me ofte treng statistikk for å svara på, kva er feilsannsynet $\mathsf{P}(\mathbf{m} \neq \hat{\mathbf{m}})$?

*Mange kanalar vert modellert additivt. Dvs. at det mottekne signalet $\mathbf{r}$ er summen av det sendte signale $\mathbf{s}$ og ein feilvektor $\mathbf{e}$.*

*Den binærsymmetriske kanalen (BSC), dvs. at $\mathbf{s}$, $\mathbf{r}$, og $\mathbf{e}$ er binære (vektorar over $\mathsf{GF}(2) = \{0,1\}$), og addisjon er modulo 2. På BSC er feilvektoren $\mathbf{e}$ generert tilfeldig, med* uavhengige *bits, der kvar bit $X$ har sannsyn $P(X=1) = p$ for feil, og $P(X=0) = 1-p$ for korrekt overføring. Me skriv gjerne BSC$(p)$ for ein binærsymmetrisk kanal med feilsannsyn $p$.*

**Definisjon 8 (Binary Symmetric Channel)**




Figur 2: The binary symmetric channel.

*Figuren viser dette skjematisk. PRNG står for «Pseudo-Random Number Generator», dvs. slumptalsgenerator.*

**Døme 4** *Når me sender éin bit over BSC har me to moglege utfall:* feil *eller* ikkje feil. *Alle bits som vert sende er uavhengige av kvarandre.*

*Dette er eit døme på eit* Bernoulli-forsøk.

**Døme 5** *Mynt og kron har mykje til felles med dømet over. Der er to utfall* mynt *eller* kron*, og kasta er uavhengige av kvarandre. Dette er òg eit* Bernoulli-forsøk.

*Sjølv om mynten er skeiv, og eit utfall vert meir sannsynleg enn det andre, so er det eit Bernoulli-forsøk; det er bare sannsynsfordelinga som er endra.*

**Definisjon 9 (Bernoulli-forsøk)** *Eit Bernoulli-forsøk er eit eksperiment som har to moglege utfall, me kaller dei gjerne* suksess *(suksess) og* mislukka *(failure), og der eksperimenta er uavhengige av kvarandre. Me skriv gjerne $p$ for suksess-sannsynet.*

Det er forvirrande, men det som er bitfeilsannsynet $p$ på BSC, vert gjerne kalt suksessannsynet $p$ når me studerer Bernoulli-forsøket. Ein skal hugsa at namna *suksess* og *mislukka* er vilkårlege. Poenget er at der er to utfall. Suksess viser gjerne til det utfallet som me vel å fokusera på, og dermed gjerne til feil i kodeteori.

**Døme 6** *Lat $Z$ vera talet på bitfeil, når me sender $n$ bits over BSC$(p)$. Dvs. at me gjer $n$ Bernoulli-forsøk med suksessannsyn $p$ og tel suksessane. Me seier at $Z$ er binomialfordelt med $n$ forsøk og punktsannsyn $p$, og me kan skriva $Z \sim B(n, p)$.*

**Oppgåve 3.47** *Lat $X$ vera talet på mynt når du kastar mynt og kron $n$ gongar med ein rettferdig mynt. Kva fordeling har $X$?*

**Oppgåve 3.48** *Sett at mynten er bøyd og har 40% sannsyn for å visa mynt. Lat $X$ vera talet på mynt når du kastar mynt og kron $n$ gongar. Kva fordeling har $X$?*

### 3.4.4. Fredag (rekneøving)

**Oppgåve 3.49** *Tenk deg at du sender eit $4$-bits ord på BSC$(p)$ med feilsannsyn $p = 0{,}1$. Me kaller talet for feil for $Z$.*

1. *Kva er sannsynet* $\mathsf{P}(Z = 0)$ *for å få ingen feil?*
2. *Kva er sannsynet* $\mathsf{P}(Z = 4)$ *for å få berre feil?*
3. *Kva er sannsynet* $\mathsf{P}(Z = 1)$ *for å nøyaktig éin feil?*
4. *Kva er sannsynet* $\mathsf{P}(Z = 2)$*?*
5. *Kva er sannsynet* $\mathsf{P}(Z = 3)$*?*

*Set opp sannsynsfordelinga i ein tabell, og bruk gjerne same tabell til dei to neste oppgåvene.*

**Oppgåve 3.50** *Sjå vidare på sannsynsfordeling frå Oppgåve 3.49. Finn forventingsverdien* $\mathsf{E}(Z)$.

**Oppgåve 3.51** *Sjå vidare på sannsynsfordeling frå Oppgåve 3.49. Finn variansen* $\sigma^2 = \mathsf{var}(Z)$ *og standardavviket* $\sigma$.

**Oppgåve 3.52** *Tenk deg at du sender eit* $4$*-bits ord på BSC*$(p)$ *for ein vilkårleg verdi av* $p$*. Me kaller talet for feil for* $Z$.

*Finn sannsynet* $\mathsf{P}(Z = z)$ *for* $z = 0, 1, 2, 3, 4$.

**Oppgåve 3.53** *Sjå vidare på sannsynsfordeling frå Oppgåve 3.52. Finn eit uttrykk for forventingsverdien* $\mathsf{E}(Z)$.

**Oppgåve 3.54** *Sjå vidare på sannsynsfordeling frå Oppgåve 3.52. Finn uttrykk for variansen* $\sigma^2 = \mathsf{var}(Z)$ *og standardavviket* $\sigma$.

**Oppgåve 3.55** *Oppgåve 7.8 og 7.10–7.12 i Frisvold og Moe.*

**Oppgåve 3.56 (Ekstra)** *Tenk deg at du sender éin einskild bit over BSC*$(p)$*. Lat* $Z = 1$ *indikera ein bitfeil, og* $Z = 0$ *null feil. No er* $Z \sim B(1, p)$*, dvs. binomialfordelt med eitt forsøk og suksessannsyn* $p$.

1. *Lag ein tabell som viser (den diskrete) sannsynsfordelinga for* $Z$.
2. *Kva er forventingsverdien (populasjonsgjennomsnittet)* $E(Z)$*?*
3. *Kva er variansen* $\mathsf{var}(Z)$*?*

**Oppgåve 3.57 (Ekstra)** *Tenk deg at du sender eit* $n$*-bits ord på BSC*$(p)$ *for ein vilkårlege verdiar av* $n$ *og* $p$*. Me kaller talet for feil for* $Z$.

*Finn sannsynet* $\mathsf{P}(Z = z)$ *for* $z = 0, 1, \ldots, n$.

## 3.5. Veke 6. Normalfordelinga

### 3.5.1. Lesestoff og heimearbeid

**Les 9 (Forventingsverdi)** *Frå Frisvold og Moe: Kapittel 4.2, 8.2, 8.3, 9.1, 9.2.*

*Framstillinga i læreboka er teoretisk og detaljert. Mange av døma som me har sett og kjem til å sjå illustrerer viktige poeng som er spredd utover boka. De vil truleg måtte lesa avsnitta om igjen fleire gongar for å få meining, men det løner seg med eing gong å skumma gjennom avsnitta over for å få oversikt over kva som står kvar.*

### 3.5.2. Onsdag (førelesing)

**Sentralgrensesetninga** Hittil har me berre arbeidd med diskrete stokastiske variablar. *Kva skil kontinuerlege variablar frå diskrete?*

Korleis kan me definera ei kontinuerleg sannsynsfordeling? Kva vert punktsannsynet i ei kontinuerleg fordeling?

Fordelingsfunksjon versus tettheitsfunksjon. Probability density versus probability distribution (PDF).

**Sentralgrensesetninga** Binomialfordelinga for stor $n$

Sentralgrensesatsen generelt

**Sats 1 (The Central Limit Theorem)** *Let* $X = X_1 + X_2 + \ldots + X_n$ *be a sum of* identically *distributed variables* $X_i$*. regardless of the exact distribution of* $X_i$

*When as* $n \to \infty$*,* $X$ *has always the same distribution, namely the normal distribution.*

Det er sentralgrensesetninga som let oss bruka normalfordelinga som tilnærming til binomialforderlinga for store verdiar av $n$. Ho let oss òg bruka normalfordelinga som tilnærming i mange andre tilfelle med store utval, uansett kva fordeling som ligg i botnen.

#### Lineære kombinasjonar av stokastiske variablar

### 3.5.3. Fredag (rekneøving)

**Oppgåve 3.58** *Oppgåve 8.1 og 8.3 frå Frisvold og Moe.*

**Oppgåve 3.59** *Oppgåve 8.4 og 8.5 frå Frisvold og Moe.*

**Oppgåve 3.60** *Oppgåve 8.7 frå Frisvold og Moe.*

Dersom du har tid til overs, bør du gå tilbake og gjera ekstra-oppgåvene på tidlegare rekne-øvingar.

## 3.6. Veke 7. Estimering av feilsannsyn

**Under utvikling**

### 3.6.1. Lesestoff og heimearbeid

**Les 10 (Forventingsverdi)** *Frå Frisvold og Moe: Kapittel*

### 3.6.2. Onsdag (førelesing)

#### Oppsummering

1. Fordelingar
    a) Normalfordelinga
    b) Binomialfordelinga
2. PDF, CDF
3. Gjennomsnitt $\mu$, standardavvik $\sigma$
4. Estimatorar
    a) Utvalsgjennomsnitt: $\bar{x}$ for $\mu$
    b) Standardavvik: $s$ for $\sigma$
5. Estimator som stokastisk variabel
    a) Forventingsverdi $E(\bar{X})$
    b) Standardavvik $\mathsf{S.Dev.}(\bar{X})$

#### Estimering av gjennomsnitt i normalfordelinga

#### Standardfeilen

### Estimering av punktsannsyn i binomialfordelinga

**Definisjon 10 (Feilsannsyn)** *Feilsannsynet er sannsynet for at ein feil vil oppstå i eit eksperiment som me enno ikkje har observert.*

**Definisjon 11 (Feilraten)** *Feilraten er andelen feil som er observert i ein serie med utførte eksperiment.*

**Merknad 4** *Feilsannsynet gjeld populasjonen — eller framtida.*

*Feilraten gjeld utvalet — eller fortida som me har observert.*

**Oppgåve 3.61 (Drøfting)** *Kva fortel fortida oss om framtida?*

Den observerte parameteren $\hat{p}$ i øving 7.13 er feilraten.

### $t$-fordelinga

### 3.6.3. Fredag (rekneøving)

**Oppgåve 3.62** *Me ynskjer å finna gjennomsnittsvekta på torsk i eit visst havområde. Me reknar med at vekta er normalfordelt, med standardavvik $\sigma = 0{,}5$. Me fangar åtte torsk, og måler vektene til*

$$1{,}2,\ 2{,}3,\ 2{,}4,\ 2{,}9,\ 3{,}1,\ 3{,}5,\ 4{,}4,\ 6{,}0$$

1. *Rekn ut gjennomsnittet $\bar{x}$ av utvalet.*
2. *Korleis vil du estimera gjennomsnittsvekta $\mu$?*
3. *Kva er standardavviket til estimatoren din?*

**Oppgåve 3.63** *Rekn ut utvalsstandardavviket $s$ for torskevektene i forrige oppgåve.*

**Oppgåve 3.64 (Drøfting)** *Ta for deg torskevektene igjen. Sett at me ikkje har peiling på standavviket $\sigma$. Korleis kan me då estimera standardavviket for estimatoren?*

**Oppgåve 3.65** *Me har utvikla ein algoritme for andletsgjenkjenning i bilete. Systemet er ikkje perfekt, og me må rekne med at for kvart bilete er der eit sannsyn $\pi$ for at biletet vert kopla til feil person. Sett at me tester systemet på 1000 bilete og får feil 110 gongar.*

*Korleis vil du estimera feilsannsynet $\pi$?*

**Oppgåve 3.66** *For å ha nytte av estimatoren i forrige oppgåve, må me ha eit idé om standardavviket (standardfeilen).*

*Lat $X$ vera talet på feil, som er binomialfordelt $B(n, \pi)$. Me veit at*

$$\sigma = \mathsf{S.Dev.}(\frac{X}{n}) = \sqrt{\frac{\pi(1-\pi)}{n}}. \tag{7}$$

*Me kan ikkje rekna ut dette exact, sidan $\pi$ er ukjend, men dersom me set inn $\hat{\pi}$ for $\pi$ får me eit høveleg estimat $\hat{\sigma}$ for $\sigma$.*

*Rekn ut dette estimatet for standardfeilen.*

**Oppgåve 3.67** *Oppgåve 8.1 og 8.13 frå Frisvold og Moe*

**Oppgåve 3.68 (Ekstra)** *Oppgåve 5.7 og 5.14 frå Frisvold og Moe*

## 3.7. Veke 8. Intervallestimering

**Under utvikling**

### 3.7.1. Lesestoff og heimearbeid

**Les 11 (Forventingsverdi)** *Frå Frisvold og Moe: Kapittel 9*

### 3.7.2. Onsdag (førelesing)

#### Repetisjon

**Definisjon 12** *Dersom $\hat{x}$ er ein estimator for $x$, so kallar me standardavviket $\sigma$ åt $\hat{x}$ for standardfeilen åt estimatoren, og skriv*

$$\mathsf{S.E.}(\hat{x}) = \sigma.$$

**Sats 2** *Feilraten $\hat{p}$ er ein estimator for feilsannsynet $p$, og standardfeilen er gjeve som*

$$\mathsf{S.E.}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}},$$

*når feilraten er rekna over $n$ forsøk. Ein estimator for standardfeilen er*

$$\widehat{\mathsf{S.E.}}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

**Konfidensintervall** Dersom me tek fylgjande intervall rundt punktestimatoren $\hat{p}$

$$(\hat{p} - \widehat{\mathsf{S.E.}}(\hat{p}), \hat{p} + \widehat{\mathsf{S.E.}}(\hat{p}))$$

går det an å visa at sannsynet for at intervallet omfattar parameteren $p$ er cirka 68%. Meir presist

$$P(p > \hat{p} + \widehat{\mathsf{S.E.}}(\hat{p})) = 0.1587 \tag{8}$$

$$P(\hat{p} - \widehat{\mathsf{S.E.}}(\hat{p}) < p < \hat{p} + \widehat{\mathsf{S.E.}}(\hat{p})) = 0.683 \tag{9}$$

$$P(p < \hat{p} - \widehat{\mathsf{S.E.}}(\hat{p})) = 0.1587 \tag{10}$$

Me kaller intervallet for eit 68.3% *konfidensintervall* for dekodingsfeilssannsynet $p$. Talet 68.3% *konfidensnivået*.

Merk at det er intervallet som er stokastisk, medan parameteren $p$ er konstant (men ukjent). Me kan difor ikkje tala om sannsynet for at $p$ ligg i intervallet.

**Oppgåve 3.69** *Me skal estimera ein feilrate, og testar systemet 1000 gongar, og finn 120 feil. Finn eit 68,3% konfidensintervall for feilsannsynet $\pi$.*

**Oppgåve 3.70** *Me skal finna gjennomsnittsvekta for ein viss dyreart i eit visst område. Me veit at vekta er normalfordelt med standardavvik $\sigma = 4$, men gjennomsnittsvekta varierer frå område til område avhengig av mattilgang m.m.*

*Me måler ni dyr, og finn vektane*

$$3{,}2; 3{,}8; 4{,}2; 4{,}4; 4{,}4; 4{,}5; 4{,}7; 5{,}1; 5{,}2$$

*Finn eit $95{,}4\%$ konfidensintervall for gjennomsnittsvekta.*

**Oppgåve 3.71** *Tilsvarande forrige oppgåve, men denne gongen er $\sigma$ ukjend. Oservasjonene er dei same som over. Finn eit Finn eit $95\%$ konfidensintervall for gjennomsnittsvekta.*

**One pitfall to avoid** Consider the following to statements:

1. When you are going to calculate a 95% confidence interval for $p$, the probability is 95% that you get an interval which encloses $p$.
2. When you have calculated a 95% confidence interval $(l, u)$ for $p$, the probability is 95% that $l \geq p \geq u$.

**Oppgåve 3.72** *Compare the two statements above. Are they equivalent or not? Is the first statement true? Is the second statement true?*

### $t$-fordeling

### 3.7.3. Fredag (rekneøving)

**Oppgåve 3.73** *Me ynskjer å finna gjennomsnittsvekta på torsk i eit visst havområde. Me reknar med at vekta er normalfordelt, med standardavvik $\sigma = 0{,}5$. Me fangar åtte torsk, og måler vektene til*

$$1{,}2,\ 2{,}3,\ 2{,}4,\ 2{,}9,\ 3{,}1,\ 3{,}5,\ 4{,}4,\ 6{,}0$$

1. *Rekn ut eit 95% konfidensintervall for gjennomsnittsvekta $\mu$.*
2. *Rekn ut eit 98% konfidensintervall for gjennomsnittsvekta $\mu$.*

*Du kan bruka resultata frå oppgåva 3.62 som mellomrekning. Datasettet er det same.*

**Oppgåve 3.74** *Me har utvikla ein algoritme for andletsgjenkjenning i bilete. Systemet er ikkje perfekt, og me må rekne med at for kvart bilete er der eit sannsyn $\pi$ for at biletet vert kopla til feil person. Sett at me tester systemet på 1000 bilete og får feil 110 gongar.*

- *Finn eit 95% konfidensinterval for punktsannsynet $\pi$.*
- *Finn eit 99% konfidensinterval for punktsannsynet $\pi$.*

*Du kan bruka resultata frå oppgåva 3.65 og 3.66 som mellomrekning. Datasettet er det same.*

**Oppgåve 3.75** *Ta for deg torskevektene i oppgåve 3.73 igjen. Det viser seg at det oppgjevne standardavviket ikkje er til å lita på. Me må rekna $\sigma$ som ukjend.*

1. *Rekn ut eit 95% konfidensintervall for gjennomsnittsvekta $\mu$.*
2. *Rekn ut eit 98% konfidensintervall for gjennomsnittsvekta $\mu$.*

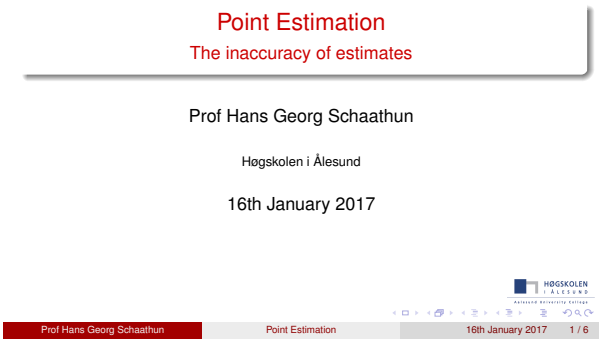
**Oppgåve 3.76** *Frisvold og Moe: oppgåve 9.1-9.2.*

**Oppgåve 3.77** *Frisvold og Moe: oppgåve 8.2 og 8.11*

**Oppgåve 3.78 (Ekstra)** *Frisvold og Moe: oppgåve 9.10-9.11.*

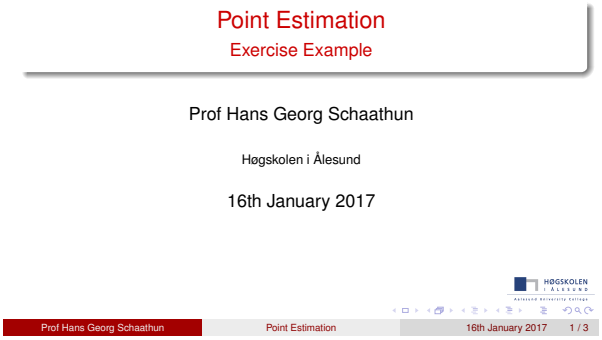
**Oppgåve 3.79 (Ekstra)** *Frisvold og Moe: oppgåve 9.3, 9.12 og 9.16.*

### 3.7.4. Etterarbeid (videføredrag)




**Les:** Frisvold og Moe s. 145-147

**Foilar:** PDF

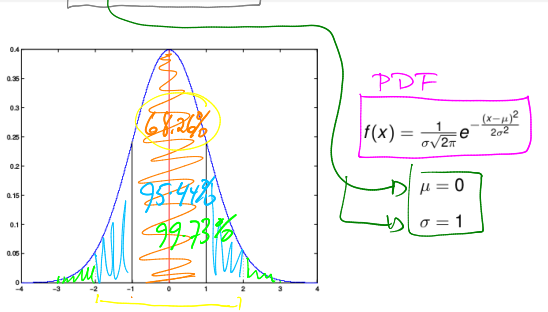



**Problem 3.1** *Suppose you are testing a system with error probability of* $0.01$*. How many trials do you need to make your estimator* $\hat{p}_e$ *fall between* $0.011$ *and* $0.009$ $99.75\%$ *of the time?*

**Foilar:** PDF

## The Gauss Curve

The PDF of the standard normal distribution






**Les:** Frisvold og Moe s. 120ff, 132

**Foilar:** PDF

## Warning! Pitfall

Confidence level versus probability

$$P_D(\hat{\theta}_{\text{low}}(D) \leq \theta \leq \hat{\theta}_{\text{high}}(D)) \geq \beta$$

- The confidence level is *a priori* probability
  - that the confidence interval will enclose the paramter $\theta$
- It is not
  - the probability that $\theta$ is within the interval
  - because $\theta$ is not a stochastic variable

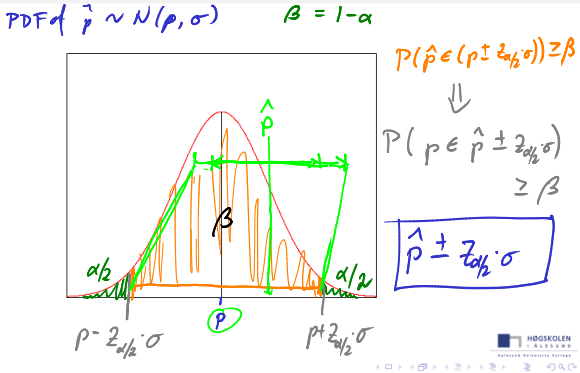
~~$$P_\theta(\hat{\theta}_{\text{low}}(D) \leq \theta \leq \hat{\theta}_{\text{high}}(D)) \geq \beta$$~~



**Les:** Frisvold og Moe s. 147-148, 163-165

**Foilar:** PDF

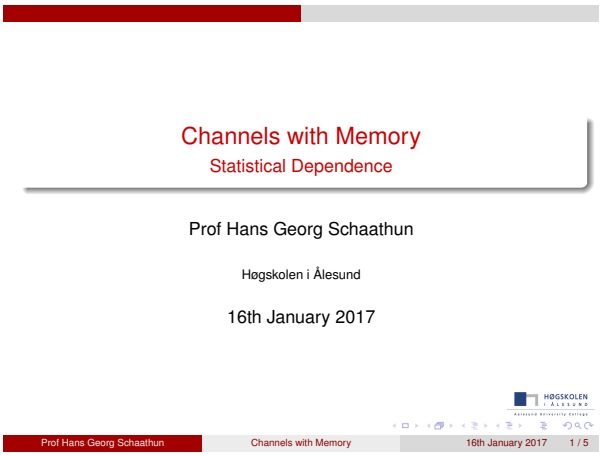
## The Point Estimator as a Start






**Les:** Frisvold og Moe s. 167-169

**Foilar:** PDF




Den binærsymmetriske kanalen har ikkje hugs, dvs. han hugsar ikkje om han har laga feil i tidlegare bits og bitfeilsannsynet er dermed konstant og uavhengig. Lagring på optiske og magnetiske platar vil ofte gje kanalar med hugs, fordi ei av dei største feilkjeldene er riper og hakk som øydelegg fleire bit på rad.

**Les:** Frisvold og Moe s. 36-38(ff)

**Foilar:** PDF

## 3.8. Veke 10. Forsøksplanlegging

**Les 12** *Frå Frisvold og Moe: Kapittel 13–14.*

### 3.8.1. Onsdag (Førelesing)

### Oversyn

1. Statistisk forsøk
    a) Datasanking
    b) Simulering
2. Stokastisk prosess
3. Modell
    a) Domenekunnskap
    b) Modellar og teoriar frå design
    c) Design

4. Ko-evolusjon av problem og design

**Praktisk problemløysing** Statistikk og simulering handlar om komplekse og samansette problem. De har ein del svært ulike problem i labøvingane. Oppgåvene på rekneøvingane er noko forenkla, med svært presise spørsmål, men òg her er der hovudvekt på tekstoppgåver der ein må skilja vesentleg informasjon frå uvesentleg. Desse oppgåvene etterliknar arbeidsoppgåver frå røynda, og det er viktig å ta dei alvorleg og prøva å forstå korleis ein ville tenkje om det faktisk var ein arbeidssituasjon i røynda.

Det som kjenneteiknar røynda er at ho er komplisert. Der er alltid for mange, både kjende og ukjende, faktorar til at eit enkelt fag eller ein einskild teori kan gje fullstendige svar. Kvart problem er individuelt, og ein må vurdera kva ein kan læra med kvar teori og med kvar disiplin.

Fem spørsmål er nyttige å ta med seg når ein ser på slike problem:

1. Kva veit eg?
2. Kva er eg beden om å finna ut?
3. Kva teorigrunnlag kan eg bruka?
4. Korleis kan problemet modellerast for å passa i teorien?
5. Kva opplysingar manglar eg?

Det løner seg å teikna opp problemet og ordna opplysingane som trengst i teikninga. Ein kan teikna perspektivteikningar, flytdiagram (boksdiagram), kurver (t.d. PDF) eller anna. Gjerne fleire teikningar om det hjelper.

Modellar er eit vesentleg hjelpemiddel. Der er mange typar modellar.

1. Sannsynsfordelingar.
2. Rutenettet som modell for landskapet i *predator/prey*
3. Lotka-Voltera-modellen (differentiallikningar) for populasjonsstorleikane i *predator/prey*
4. Den lineære kongruensgeneratoren som modell for ein stokastisk prosess.

Ein (ingeniørfagleg) modell er ein *selektiv, men presis og nøyaktig, representasjon av eit system, laga for støtta ein viss type analyse*. Modellen er aldri fullstendig (det ville ha vore ein kopi og ikkje ein modell). Det som ikkje er relevant for analysen vert abstrahert bort. Modellen er presis og nøyaktig, slik at me kan rekna på han, men det er relativt; han skal vera nøyaktig nok for den planlagde analysen.

Mange modellar definerer samanhengar mellom aktuelle variablar, t.d. populasjonsstorleikar. Samanhengane kan vera probabilistiske eller deterministiske. Når me reknar på ein modell kan me seia at me simulerer systemet som er modellert.

Ein simulator, implementert som eit dataprogram, er eit døme på modell. Dataprogrammet er

ein representasjon av røynda.

Når me bruker ein teori eller matematisk formel, ligg der alltid ein modell til grunn, implisitt eller eksplisitt. Det er ved å samanlikna modellen, og omstendeleg sjekka at modellen faktisk er ein høveleg representasjon av røynda, at me kan vita om formelene er relevant. Difor løner det seg å bruka tid på modelleringa, og forsikra seg om at ein skjøner modellen og systemet som vert modellert, samt samanhengen mellom dei, før ein reknar eller programmerer.

Praktiske problem krev breid kompetanse. Skal me studera eit økosystem, treng me både biologar/økologar som kjenner systemet i røynda, matematikarar/statistikarar som har eit repertoir av gode modelltypar og som veit kva det er råd å rekna på, samt datavitarar som veit kva som kan simulerast maskinelt og korleis.

Som regel veit ein ikkje alt ein burde ha visst. Ein må vera tydleg på føresetnadene som ein gjer, og gjerne simulera fleire alternativ med ulike føresetnader. Det ser ein t.d. når forskarane gjev klimaprediksjonar. Dei reknar gjerne eit pessimistisk og eit optimistisk alternativ, samt eit imellom.

### Døme

1. Prosjekt 2: Predator/Prey
2. Prosjekt 4: Diffusjon Sjå avsnitt 8.1.

## 3.8.2. Fredag (rekneøving)

**Oppgåve 3.80 (Gruppearbeid)** *Ein* random walker *flyttar seg på eit 1D raster. Startposisjonen kallar med $x = 0$. Lat den stokastiske variabelen $X$ vera forflyttinga til partikkelen i løpet av eitt tidssteg. Sannsynsfordelinga for $X$ er:*

$$P(X = -1) = 0{,}2 \tag{11}$$
$$P(X = 0) = 0{,}6 \tag{12}$$
$$P(X = 1) = 0{,}2. \tag{13}$$

*Lat $Y_t$ vera posisjonen etter $i$ tidssteg.*

1. *Rekn ut sannsynsfordelinga for posisjonen etter $t = 2$ tidssteg.*
2. *Rekn ut sannsynsfordelinga for posisjonen etter $t = 3$ tidssteg.*
3. *Bruk resultatet frå forrige oppgåve til å rekna ut standardavviket for posisjonen etter $t = 3$ tidssteg.*

*Legg merke til at posisjonen $Y_t$ etter $t$ steg er summen av $t$ identisk fordelte variablar ($X$), éin for kvart tidssteg.*

4. *Rekn ut populasjonsstandardavviket for posisjonen etter eitt tidssteg.*

5. *Rekn ut populasjonsstandardavviket for posisjonen etter 100 tidssteg.*
6. *Bruk same metode for å finna standardavviket for posisjonen etter $t = 3$ tidssteg og samanlikna med resultatet frå spørsmål 3.*

**Oppgåve 3.81 (Ekstra)** *Frisvold og Moe: oppgåve 9.11 og 9.13.*

**Liknande eksamensoppgåver (Ekstra)** Nokre av oppgåvene under vil vera vanskelege å gjera no, men gje meir meining når de har gjort prosjekt 4.

**Oppgåve 3.82 (Frå eksamen våren 2015)** *Ein partikkel finst i eit firkanta, todimensjonalt raster med uendeleg utstrekking. For kvart tidssteg går partikkelen eitt steg til ein av dei fire naboposisjonane med fylgjande sannsynsfordeling:*

| | | | | |
|---|---|---|---|---|
| | | | | |
| | | *40%* | | |
| | *30%* | start | *30%* | |
| | | *0%* | | |
| | | | | |

1. *Rekn ut sannsynsfordelingen for posisjonen etter to steg.*
2. *Rekn ut sannsynsfordelingen for posisjonen etter tre steg.*

**Oppgåve 3.83 (Frå eksamen våren 2017)** *Ein* random walker *flyttar seg på eit 1D raster. Lat den stokastiske variabelen $X$ være forflytningen til partikkelen i løpet av et tidssteg. Sannsynlighetsfordelingen for X er:*

$$P(X = -2) = P(X = 2) = 0,1 \tag{14}$$

$$P(X = -1) = P(X = 1) = 0,4. \tag{15}$$

1. *Rekn ut populasjonsstandardavviket for posisjonen etter eitt tidssteg.*
2. *Rekn ut populasjonsstandardavviket for posisjonen etter 100 tidssteg.*
3. *Me er interesserte i den eksakte sannsynsfordelinga for posisjonen til partikkelen etter 100 tidssteg. Forklar korleis me kan finna denne sannsynsfordelinga ved hjelp av ei datamaskin.*
4. *Forklar korleis me kan generera eit utval med ti observasjonar av posisjonen etter 100 tidssteg.*

## 3.9. Veke 11. Hypotesetesting med gjennomsnitt

**Related reading:** Denne økta byggjer på Kapittel 11.1 i Frisvold og Moe. Ho føreset og at stoffet frå forrige økt (Kapittel 10 i Frisvold og Moe) er forstått.

**Førelesing onsdag** Repetisjon av fakultet, binomisk fordeling, varians og standardavvik for binomisk fordeling. Tankegang bak hypoteseprøving: Signifikansnivå, p-verdi (=signifikanssannsynlighet), testing ved binomisk fordeling, Underlag for det som kommer onsdag 20.

**Rekneøving onsdag** Regneøving, sannsynlighetsberegninger ved normalfordeling og sentralgrensesetning, normalfordelte variabler, summer og gjennomsnitt av normalfordelte variabler.

**Oppgåve 3.84** *Alle oppgaver fra og med "Oppgave 8.1" side 124 til og med "Oppgave 8.8" side 131.*

**Oppgåve 3.85 (Ekstra)** *Blandede oppgaver fra de siste sidene i kapitel 8, etter eget valg.*

### 3.9.1. Nokre øvingar frå i fjor

#### Einsidig test med kjent $\sigma$

**Oppgåve 3.86** *Ein produsent hevdar at levetida på lysepærene han produserer er minst 1150h i gjennomsnitt. Me veit, generelt, at levetida på slike lyspærer er normalfordelt med standardavvik $\sigma = 62{,}5\,\mathrm{h}$. Tenk deg at me kan testa $n$ pærer til dei går. Korleis kan me gå fram for å testa påstanden frå produsenten? (Kva relevante statistikkar kan me observera? Korleis skal me vurdera observasjonanr?)*

**Oppgåve 3.87** *Tenk deg at me testar $n$ lyspærer og observerer levetida $X$ i timar. Kva fordeling har gjennomsnittet $\bar{X}$?*

**Oppgåve 3.88** *Me testar $n = 20$ pærer som i forrige oppgåve og får fylgjande observasjonar i timar:* $1064{,}464\,008\,3$, $1120{,}006\,031\,44$, $1052{,}047\,925\,79$, $1216{,}546\,943\,16$, $1049{,}867\,459\,64$, $1182{,}072\,895\,52$, $1078{,}326\,096\,66$, $1203{,}812\,842\,65$, $1026{,}011\,540\,81$, $1181{,}024\,915\,34$, $1123{,}430\,758\,23$, $1111{,}794\,327\,89$, $1195{,}845\,007\,38$, $1111{,}744\,430\,65$, $1171{,}080\,278\,17$, $1154{,}858\,451\,04$, $1063{,}504\,736\,86$, $1245{,}843\,696\,43$, $1154{,}683\,771\,04$, $1219{,}354\,634\,99$.

*Det gjev eit gjennomsnitt på $\bar{x} = 1136{,}32\,\mathrm{h}$. Kan me tru på påstanden om ein gjennomsnittleg levetid på minimum 1150h?*

**Oppgåve 3.89** *Me testar $n = 200$ pærer på same måte som i forrige oppgåve, og får eit gjennomsnitt på $\bar{x} = 1136{,}5\,\mathrm{h}$. Kan me tru på påstanden om ein gjennomsnittleg levetid på minimum 1150h?*

#### Tosidig test med kjent $\sigma$

**Oppgåve 3.90** *Det vert påstått at gjennomsnittshøgda på mannlege studentar i Noreg er* $180{,}5\,\mathrm{cm}$. *Me (latar som om me) veit at menneskeleg høgd er normalfordelt med standardavvik* $\sigma = 3{,}2\,\mathrm{cm}$ *Korleis kan me testa påstanden?*

1. *Kva nullhypotese har me?*
2. *Kva er den alternative hypotesa?*
3. *Kva statistikk kan me observera?*
4. *Korleis er statistikken fordelt?*
5. *For kva verdiar av statistikken skal me forkasta nullhypotesen ved* $5\%$ *signifikansnivå?*

**Oppgåve 3.91** *Sett at me målet høgda på* $n = 8$ *studentar, og fann* $186{,}59$, $177{,}39$, $180{,}54$, $178{,}23$, $178{,}55$, $178{,}73$, $180{,}68$, $173{,}84$. *Kva fortel det oss om hypotesen vår?*

**Oppgåve 3.92** *Sett at me aukar talet på observasjonar til* $n = 16$. *Kva er då den kritiske verdien for å forkasta nullhypotesen på* $5\%$ *signifikansnivå.*

*Kva vert dei kritiske verdiane med* $n = 32$*?*

### Einsidig test med ukjent $\sigma$

**Oppgåve 3.93** *Gå tilbake til lyspæredømet i øving 3.86. Sett at me ikkje har nokon informasjon om standardavviket* $\sigma$. *Korleis påverkar det hypotesetesten vår? (Utgangspunktet er same påstand som før.)*

**Definisjon 13** *Statistikken*

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

*er* $t$*-fordelt med* $n - 1$ *fridomsgradar dersom* $X$ *er tilnærma normalfordelt.*

**Oppgåve 3.94** *Sjå på sannsynsfordelinga for* $t$*-fordelinga for ulike tal på fridomsgradar, og samanlikn med normalfordelinga. T.d.*

```
fplot( @(x)pdf('t',x,4), [-5 5] )
hold
fplot( @(x)pdf('t',x,9), [-5 5] )
fplot( @(x)pdf('t',x,12), [-5 5] )
fplot( @(x)pdf('t',x,20), [-5 5] )
fplot( @(x)pdf('Normal',x,0,1), [-5 5] )
```

*Bruk* `help` *eller* `doc` *for å sjå nøyaktig kva funksjonane gjer.*

**Oppgåve 3.95** *Me testar $n = 6$ pærer og får fylgjande observasjonar i timar:* $1064{,}464\,008\,3$*,* $1120{,}006\,031\,44$*,* $1052{,}047\,925\,79$*,* $1216{,}546\,943\,16$*,* $1049{,}867\,459\,64$*,* $1182{,}072\,895\,52$*, Kan me forkasta nullhoptesen med* $5\%$ *signifikansnivå?*

### Eigenøving

**Oppgåve 3.96** *Gjer alle oppgåvene i Kapittel 11.1 i Frisvold og Moe.*

Eg har ein serie med Videoar frå 2015. Økt (session) 4-5 gjeld hypotesetesting. Økt 4 er det stoffet som me har gått gjennom i Økt 27. Økt 5 gjev tre nye fall, gjennomsnitt med ukjent standaravvik, samanlikning av to gjennomsnitt og binomialproporsjonen. Desse tre falla err ikkje veldig vanskelege å forstå dersom de maktar å overføra det som me har diskutert i klassa, til nye problem.

## 3.10. Veke 12. Meir Hypotesetesting

**Førelesing (onsdag)** Kapittel [10.1.3, 11.1.3) og [11.1.4,11.2.3) Parentesene betyr "fra og med" og "til". Delkapittel 11.1.3 blir utsatt.

**Heimearbeid mot fredag** Les teori og løs oppgavene nedenfor etterhvert som du støter på dem. Hver oppgave skal løses mer enn én gang, med mindre du finner den lett og vet du vil huske prosedyren. Gjenta løsning inntil du løser oppgaven "i flytsonen". De fleste lærer raskere ved å gjenta løsning av samme oppgave to eller flere ganger, sammenliknet med å straks gå løs på ny når én oppgave er løst. Gjentatt løsning fester prosedyren i minnet. Når oppgaven er løst: Tenk gjennom hvorfor løsningsmåten virket.

De som "tar faget lett", kan nøye seg med å løse hver oppgave én gang, bør da lete etter utfordringer i oppgavesamling ved kapittel-slutt. Det er for slike studenter en sunn utfordring å selv finne fram til oppgaver som lar seg løse ved allerede lært teori.

Oppgaver plukket fra kapittelslutt vil senere i kurset bli gitt som repetisjonsarbeid.

Oppgaver å løse fram mot og kommende fredag blir

Oppgave 10.1, 10.2 og 10.3, her er det lurt å samarbeide og diskutere. Oppgave 11.2 (Les først eksempel 11.2) Oppgave 11.3 (Les først eksempel 11.4) Oppgave 11.4 (Les først eksempel 11.5)

Teorien skal læres så prinsippene, begreper og prosedyrer huskes! Det er mye teori denne gangen, så dette vil kreve tid. Spesielt må alle nye begreper læres, ellers vil det bli vanskelig å forstå det du leser og hører senere i kurset. Kommende onsdag vil begrepene, forutsettes kjent. (Jeg vil nok likevel gjenta noen definisjoner. Begreper må masseres inn over tid, så vi blir fortrolige med dem.)

Jeg tror dette er mer enn nok. Men hvis tid og lyst: Oppgaver ved kapittelslutt.

**Førelesing (onsdag)** Samarbeid og regn oppgaver gitt onsdag 20. mars.

## 3.11. Veke 13.

Vanleg førelesing onsdag.

Undervisingsfri fredag. Romma er reserverte som vanleg for dei som vil arbeida i lag.

**Førelesing (onsdag)** Gjennomgang av teori og eksempler, læreboka delkapitlene [11.2.2, 11.4). Det blir ganske intensivt, for nå skal du allerede være fortrolig med de underliggende prinsippene for hypoteseprøving. Jeg forutsetter dette og kan dermed øke farta.

**Rekneøving (på eiga hand)** Etter forelesningen, før forelesning onsdag 3. april:

1. Arbeid deg gjennom stoffet. Etter å ha lest teori og eksempel regner du først gjennom eksempel selv, så reflekterer du over hvorfor du gjorde det du gjorde, hvorfor dette er rett og hva som var hensikten. Jeg gjentar med andre ord: Ta noen minutter til å forklare tankegangen for deg selv.

   Deretter regner du påfølgende oppgave i læreboka. Oppgaven er prinsipielt lik forutgående eksempel. Etter å ha regnet oppgaven, sammenlikner du med eksempelet i læreboka, og retter opp om du uttrykte deg dårlig. Husk å avslutte med tekstsvar. Du avslutter du ved å forklare for deg selv tankegangen du brukte da du løste oppgava.

2. Så går du tilbake til teori, eksempler og oppgaver fra forrige uke. Regn oppgavene på nytt. Dette vil bidra til at du fester det du skal lære i hjernen, ikke glemmer. Denne repetisjonen er tidseffektiv: Det er lettere å vedlikeholde enn å ta noe igjen når det er kommet på avstand, er helt glemt. Totalt sett sparer du tid på å repetere allerede nå, etter bare én uke. Husk å forklare for deg selv hver gang du har løst en oppgave. Det tar få minutter sammenliknet med å regne oppgaven, det og gir stor læringseffekt.

3. Regn følgende oppgaver: 11.23, 11.16, 11.17, 11.19, 11.24. Dette er det laveste antall oppgaver du må regne. Det tar tid å få en følelse for hypotseprøving, de fleste må legge ned mye arbeid. Om du ikke rekker å regne alle, regn noen. Noen er bedre enn ingen.

Det vil selvsagt være bra å regne flere oppgaver enn nevnt, men jeg plukker ikke ut flere, vet at enkelte studenter mister motet om de ser for mange anbefalte oppgaver opplistet.

## 3.12. Veke 14.

Delkapittel [11.3, 11.4.2) og [11.5, 2. 12) Teori gjennomgås ved eksemplene i læreboka. Som vanlig skal alle oppgaver flettet inn i teorien løses så snart forutgående eksempel er studert og gjennomtenkt. Regn i tillegg oppgavene 11.27 og 1.28. Hvis du ikke har regnet oppgavene jeg tidligere har anbefalt, prøv å ta igjen de du ikke har fått gjort, så du kommer a jour.

### 3.12.1. Reknøving (fredag)

Rekn oppgåvene som vart gjevne på onsdag.

## 3.13. Veke 15.

Jeg går gjennom kapittel [12, 12.4). Jeg går gjennom prinsippene, men skriver ikke opp alle formlene på tavla, det blir for mange. Jeg går gjennom kriteriene for hvilke formler som skal brukes og jeg slår opp formlene på aktuelle sider i læreboka, studentene som er på forelesning gjør det sammen med meg. Jeg forteller at mange av oppgavene forutsetter at kalkulatoren er tilstrekkelig avansert til å beregne ulike størrelser ved innskriving av tall i tabell, jeg vet ikke om nåværende eksamenskalkulator makter dette.

Jeg vil fortelle at regresjonsanalyse i vår tid utføres ved bruk av datamaskin med hensiktsmessig programvare, at uten slik programvare blir regresjonsanalyse som å bygge flyplass med spade, i 2019. Lærebokas presentasjon bygger på hjelpemidler tillat til eksamen da boka ble skrevet, er dermed ikke "tidsriktig" når det gjelder det beregningstekniske. Prinsippene er imidlertid viktige.

Regneeksemplene i kap 12-12.4 bør studeres, og du bør regne oppgavene som følger eksemplene.

### 3.13.1. Reknøving (fredag)

Rekn oppgåvene som vart gjevne på onsdag.

## 3.14. Veke 17.

Jeg gjennomgår resten av kapittel 12, du bør gå grundig gjennom regneeksemplene.

### 3.14.1. Reknøving (fredag)

Repetisjon, muligens gjennomgang av et par punkter jeg har tatt lett på.

## 3.15. Veke 18.

Inga førelesing pga. fyrste mai.

Repetisjonsøving fredag tredje mai.

# 4. Skisser

## 4.1. Veke 12. Korrelasjon og regresjon

### 4.1.1. Lesestoff og heimearbeid

**Les 13 (Regresjon og Korrelasjon)** *Frå Frisvold og Moe: Kapittel 12.1, 12.2 og 12.5.*

### 4.1.2. Onsdag (førelesing)

Merk at notata nedanfor er meint å gje eit kort overblikk over hovedpunkta som vert gjennomgått. Det er ikkje meininga at du skal læra stoffet utan å vera til stades og lesa læreboka.

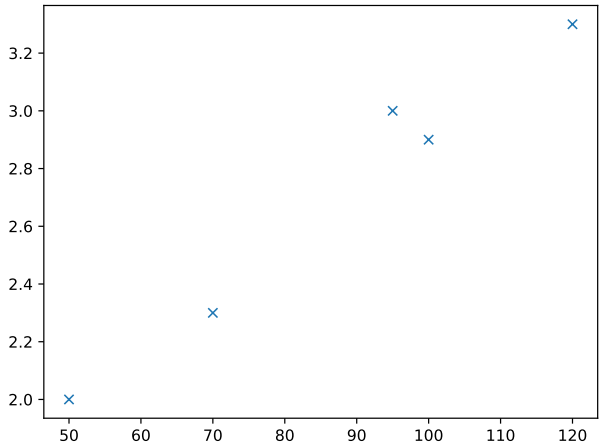
#### Regresjon

**Oppgåve 4.1** *Me ynskjer å forstå samanhengen mellom areal og pris på bustader. Me har observert sal av fem bustadar:*

| *Areal* | *50* | *70* | *95* | *100* | *120* |
|---|---|---|---|---|---|
| *Pris* | *2 mill.* | *2,3 mill* | *3 mill* | *2,9 mill* | *3,3 mill* |

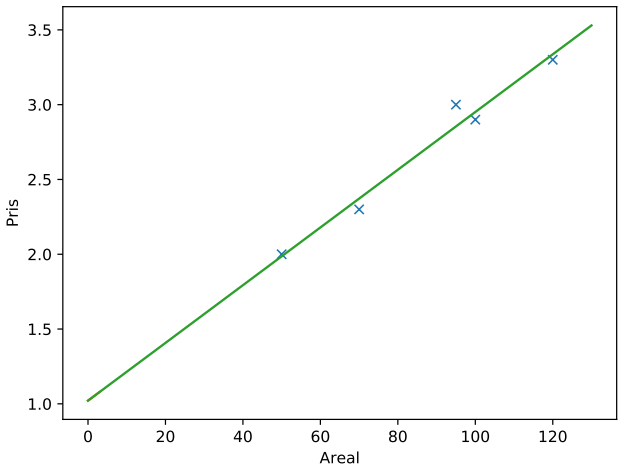
*Kva teknikkar og modellar kan me bruka for å forklara samanhengen?*

**Døme 7** *Me har observert to stokastiske variablar: areal, som me skriv $X$ og pris, som me skriv $Y$. Observasjonane kjem i par $(X, Y)$, der me har observert pris og areal på den same bustaden. Då er det naturleg å plotta dei to variablane saman i $(x, y)$-planet.*




**Oppgåve 4.2 (Drøfting)** *Sjå på plottet over. Går det an å skriva prisen som ein funksjon (omtrentleg eller eksakt) av arealet? Kva slags funksjon vil du føreslå?*

**Døme 8** *Ein lineær funksjon er ein høveleg god tilnærming, som me ser her:*




*Me bruker minste kvadrats metode for å finna den beste lina eksakt (sjå læreboka).*

## Korrelasjon

**Døme 9** *Me kan rekna ut variansen for dei to variablane som fylgjer:*

| $x$ | *50* | *70* | *95* | *100* | *120* | *Sum* |
|---|---|---|---|---|---|---|
| $x - \bar{x}$ | *-37* | *-17* | *8* | *13* | *33* | *0* |
| $(x - \bar{x})^2$ | *1369* | *289* | *64* | *169* | *1089* | *2980* |
| $y$ | *2 mill.* | *2,3 mill* | *3 mill* | *2,9 mill* | *3,3 mill* | |
| $y - \bar{y}$ | *-0,7 mill.* | *-0,4 mill* | *0,3 mill* | *0,2 mill* | *0,6 mill* | *0* |
| $(y - \bar{y})^2$ | *0,49* | *0,16* | *0,09* | *0,04* | *0,36* | *1,14* |

*Mao.* $s_X^2 = 2980$ *og* $s_Y^2 = 1{,}14$.

Dei to variablane er openbert ikkje uavhengige og variasjonen i kvar variabel er langt mindre interessant enn samanhengen mellom dei.

Variansen er

$$\sigma_X^2 = E((X - \mu_X)^2)$$

for $X$ og

$$\sigma_Y^2 = E((Y - \mu_Y)^2)$$

for $Y$. Utfall som er svært forskjellig frå gjennomsnittet trekk forventingsverdien (variansen) mykje opp. Utfall nær gjennomsnittet har liten innverknad.

Me kan òg studera kovariansen

$$\sigma_{XY} = E((X - \mu_X)(Y - \mu_Y)).$$

Her ser me at utfall som er svært forskjellig frå gjennomsnittet for $X$ berre påverkar forventingsverdien når dei opptrer saman med $Y$-verdiar som òg avvik frå gjennomsnittet. Kovariansen kan ha negativt forteikn dersom $X$ plar vera stor når $Y$ er liten og omvendt.

**Døme 10** *Me kan rekna ut utvalskovariansen som fylgjer:*

| $x$ | *50* | *70* | *95* | *100* | *120* | *Sum* |
|---|---|---|---|---|---|---|
| $x - \bar{x}$ | *-37* | *-17* | *8* | *13* | *33* | *0* |
| $y$ | *2 mill.* | *2,3 mill* | *3 mill* | *2,9 mill* | *3,3 mill* | |
| $y - \bar{y}$ | *-0,7 mill.* | *-0,4 mill* | *0,3 mill* | *0,2 mill* | *0,6 mill* | *0* |
| $(x - \bar{x})(y - \bar{y})$ | *25,9* | *6,8* | *2,4* | *2,6* | *19,8* | *57,5* |

*Mao.* $s_{XY} = 57{,}5/4 = 14{,}375$.

Et problem med kovariansen som mål er at høg varians også bidreg til høg kovarians (i absoluttverdi). To variablar med høg kovarians treng difor ikkje vera svært avhengige av kvarandre. For å få eit godt mål for avhenget uavhengig av variansen, kan me normalisera og få den sokalla korrelasjonskoeffisienten:

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

### 4.1.3. Tysdag (rekneøving)

**Oppgåve 4.3** *Frisvold og Moe: oppgåve 12.1*

**Oppgåve 4.4** *Frisvold og Moe: oppgåve 12.5*

**Oppgåve 4.5 (Predicting Mental Ability)** *Er der ein lineær samanheng mellom alderen når eit barn tek til å tala, og mentale evnar seinare?*

*For å svara på dette har me samla data om ti born og registrert alderen i månader då dei fyrst talte, og score på ein evnetest som tenåring.*

| *Alder (i månader)* | *Score* |
|---:|---:|
| *15* | *95* |
| *26* | *71* |
| *10* | *83* |
| *9* | *91* |
| *15* | *102* |
| *20* | *87* |
| *18* | *93* |
| *11* | *100* |
| *8* | *104* |
| *20* | *94* |

*Teikn eit spreidingplott (scatterplot) og avgjer om du synest der ser ut til å vera ein lineær samanheng mellom dei to variablane. Beskriv evt. samanhengen.*

*Rekn ut korrelasjonskoeffisienten ($r = \hat{\rho}$) Kor stor andel av variasjonen i evnenivå (testresultat) kan forklarast med modellen?*

**Oppgåve 4.6** *Sjå på fylgjande datasett:*

| $x$ | *0.00* | *1.00* | *2.00* | *3.00* | *4.00* | *5.00* |
|---|---|---|---|---|---|---|
| $y$ | *0.03* | *0.15* | *0.89* | *2.79* | *6.42* | *12.5* |

*Bruk minste kvadrats metode for å finna ei rett line $y = a + bx$ som tilnærmer datasettet.*

# 5. Prosjekt 1

## 5.1. Heimearbeid. Programvare

Målet med den den fyrste labøvinga er å koma i gang med enkle stokastiske simuleringar på datamaskina. Før labøvinga må de (i alle fall prøva å) installera programvaren de treng.

Øvingane dei fyrste vekene er lagt opp for Matlab, men det er lov å bruka andre språk i staden.

Python er m.a. eit godt val, og eg kan stort sett vegleia ogso i python, men likevel vil de måtta bruka litt tid på søkja fram funksjonar for python

Øvinga er lagt opp som ei serie enkle oppgåver. Utfordringa er å forstå kva kvar komando tyder. Prøv deg fram, reflekter, og drøft tolkingar med andre studentar. Ikkje nøl med å spørja meg om de står fast.

Mot slutten av øvinga vil vi be om presentasjon av resultatane frå utvalde oppgåver. Sjå til at du noterer og lagrer alt du gjer slik at du raskt kan demonstrera kva du har gjort på nytt.

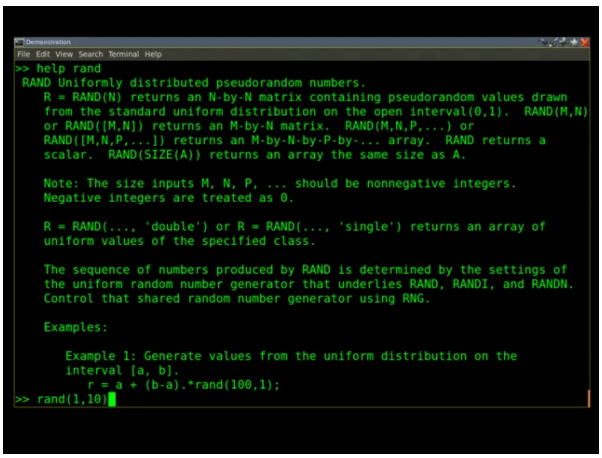
### 5.1.1. Installasjon

Det fyrste du må gjera er å installera Matlab. Sjå innsida for informasjon. Når du vert spurt om kva pakkar (toolboxes) du vil installera, skal du ta med Communications og Statistics som minimum.

Dersom du vil satsa på python, skal du passa på å installera pakkane for vitskapleg rekning. Sjå SciPy.org. Instruksjonane for installasjon ved hjelp av pakkeverkty på linux eller vha. macports er fullstendige og installerer python i tillegg til naudsynte pakkar. Installasjon vha. pip føreset at Python allereie er installert.

Eg kan bistå ved installasjon på Linux. For andre system må de be Orakel om hjelp.

### 5.1.2. Tilfeldige tal




Videoen er laga 2015, og ikkje heilt tilpassa opplegget i år. Han viser éin mogleg måte å bruka Matlab på. Tema er korleis ein genererer tilfeldig tal og tilfeldige bits.

**Foilar:** PDF

**Les:** `help rand` i Matlab

## 5.2. Veke 2. Simulering i Matlab

### 5.2.1. Nokre grunnleggjande komandoar (Matlab)

**Oppgåve 5.1** *Prøv fylgjande kommandoar eit par gongar i Matlab, og forsøk å finna ut kva dei gjer.*

1. `rand`
2. `rand<0.5`
3. `rand(2,3)`
4. `rand(2,3)<0.5`

*Bruk gjerne hjelpefunksjonen for å forstå komandoen. Der er to utgåver:*

- `help rand`
- `doc rand`

**Oppgåve 5.2 (Diskusjon)** *Bruk* `rand`*-komandoen over til å simulera ti myntkast. Korleis kan du gjera det med éin komando?*

**Oppgåve 5.3** *Prøv fylgjande komandoar:*

```
n = 5
x = rand(1,n)<0.5
histogram(x,'BinMethod','integers')
```

*Den siste komandoen lagar eit plot (histogram). Kva informasjon gjev dette histogrammet?*

**Oppgåve 5.4 (Diskusjon)** *Kva skjer når me aukar verdien til variabelen* `n`*? Dvs. korleis ser histogrammet ut for ulike tal på myntkast, t.d.* $n = 5, 10, 25, 100, 500$*?*

**Oppgåve 5.5** *Prøv fylgjande Matlab-kode:*

```
n=2
trials=5
for i=1:trials
  t = rand(1,n)<0.5;
  x(i) = sum(t);
end
x
```

*Dette er òg ein simulering av ein serie myntkast. Forklar kva eksperiment som er simulert, i.e. korleis ville du ha gjort same eksperiment med fysiske myntar? Variabelen* `x` *er ein vektor som*

*oppsummerer resultatet av eksperimentet. Kva er det me har observert i eksperimentet?*

Her er det nyttig å innføra funksjonar. I Matlab kan ein skriva sine eigne funksjonar vha. m-filar, dvs. filar med filnamn som endar på .m.

**Oppgåve 5.6** *Lag ein m-fil,* `cointrial.m` *som gjer simuleringa frå problemet over, dvs. innhaldet kan sjå slik ut:*

```
function [x] = cointrial(n, trials)

for i=1:trials
  t = rand(1,n)<0.5;
  x(i) = sum(t);
end
```

*Test funksjonen på komandolina, som fylgjer:*

```
cointrial(2, 5)
```

*Liknar resultatet på tidlegare test?*

**Oppgåve 5.7** *Me kan teikna eit histogram som før:*

```
  x = cointrial(2, 5)
  histogram(x,'BinMethod','integers')
```

*Test koden eit par gongar med fleire forsøk, og med fleire myntar per forsøk. Kva ser du?*

**Oppgåve 5.8** *Prøv fylgjande kode og samanlikn med forrige oppgåve*

```
  x = cointrial(2, 5)
  histogram(x,'BinMethod','integers','Normalization','probability')
```

*Kva gjer dei to siste argumenta i histogram-lina?*

### 5.2.2. Ein myntsimulering

No tek me for oss eksperimentet der me kastar $n$ myntar, og let den stokastise variabelen $X$ vera talet på myntar som viser kron. Me ynskjer å studera sannsynsfordelinga åt $X$ for ulike verdiar av $n$. For å gjera det, må me gjenta eksperimentet mange gongar.

**Oppgåve 5.9 (Diskusjon)** *Korleis bruker du* `cointrial`*-funksjonen for å gjenta eksperimentet med $n = 3$ myntar 100 gongar?*

**Oppgåve 5.10** *Bruk kommandoane som me har testa over til å laga eit histogram som viser fordelinga for $X$ over 100 forsøk med $n = 3$ myntar.*

**Oppgåve 5.11** *Gjenta forrige oppgåve for $n = 2, 5, 20, 100, 1000$, og samanlikna histogrammet. Korleis påverkar $n$ formen på histogrammet.*

**Merknad 5** *Forrige oppgåve illustrerer eit viktig resultat som er kjend som sentralgrensesatsen. Dén skal me koma tilbake til fleire gongar i løpet av semesteret.*

### 5.2.3. Teoretisk fordeling

Over har me simulert stokastiske forsøk. Me kan òg bruka matlab til å rekna på teoretisk sannsynsfordelign.

**Oppgåve 5.12** *Sett at me kastar mynt og kron $n$ gongar, og let utfallsrommet vera alle ordna kombinasjonar av mynt/kron. Dvs. at me har $2^n$ utfall. Talet på utfall som gjev mynt $k$ gongar kan me finna i Matlab med funksjonen:*

```
    nchoosek(n, k)
```

*Svar på fylgjande:*

1. *Test funksjonen og sjekk hjelpesida. Kva matematisk uttrykk svarer til Matlab-funksjonen* `nchoosek`*?*
2. *Kva uttrykk kan du bruka i Matlab for å rekna sannsynet for å få $k$ gongar mynt på $n$ kast? Lag ein m-fil med ein funksjon* `coinprob(n,k)` *som reknar ut dette sannsynet.*

**Oppgåve 5.13** *Du kan bruka fylgjande kode for å teikna eit histogram over den teoretiske sannsynsfordelinga:*

```
for k = 0:(n)
  prob(k+1) = coinprob(n, k) ;
end
bar((0:n), prob, 'grouped')
```

*Svar på fylgjande:*

1. *Forklar kva løkka gjer. Kva slags objekt er* `prob`*?*
2. *Forklar kva* `bar`*-funksjonen gjer. Bruk gjerne hjelpesidane.*
3. *Bruk koden til å laga eit histogram for sannsynsfordelinga ved 10 myntkast.*

**Oppgåve 5.14** *Bruk* `cointrial`*-funksjonen frå tidlegare til å simulera ti myntkast, og lag histogram over den empiriske fordelinga når du køyrer 10, 100, 1000 og 10 000 forsøk. Samanlikna dei empiriske fordelingane med den teoretiske fordelinga i forrige oppgåve. Kva ser du? Liknar den teoretiske fordelinga på den empiriske?*

## 5.3. Veke 3. Introduksjon til slumptalsgeneratorar (labøving)

Den mest populære metoden for å generera slumptal på ein datamaskin er lineære kongruensgeneratorar. Dei har mange lytar, men dei har vore velkjende over lang tid, dei er enkle å implementera, og dei er effektive. Der er ein del alternativ, og mange er betre sjølv om dei heller ikkje er lytefrie. Me tek for oss den lineære kongruensgeneratoren for å illustrera nokre generelle fenomen og utfordringar.

**Definisjon 14** *Ein lineær rekurrens er gjeve ved formelen*

$$s_i = a \cdot s_{i-1} + c \mod m,$$

*for konstante heiltal* $a$*,* $c$ *og* $m$*. Dersom me startar med eit frø (*seed*)* $s_0$*, kan me bruka rekkurensen til å generera ei fylgje av slumptal* $x_1, x_2, x_3, \ldots$*. Denne slumptalsgeneratoren er kjend som* Lehmers algoritme *og som en* lineær kongruensgenerator.

**Oppgåve 5.15 (Refleksjon)** *Sjå på definisjon over. Hugsar du kva* $\mod$ *-operatoren tyder? Sjå tilbake på pensum frå* Diskret Matematikk *om det trengst.*

### 5.3.1. Fyrste test i Matlab

Me kan implementera kongruensgeneratoren i Matlab som fylgjer:

```
function [x] = rng25(n,x0)

a = int32(6) ;      % Multiplicative constant
c = int32(7) ;      % Additive constant
m = int32(25) ;     % Modulus
persistent s        % State

if nargin > 1,
   % If x0 is given, we use it as the seed
   s = int32(x0) ;
elseif isempty(s),
   % If no state is set, we use a hardcoded one.
   s = int32(11)
end
```

```
15
16 % Start with an empty array
17 x = [] ;
18
19 % Iterate n times to generate an array of random numbers
20 for i=1:n,
21    s = mod(a*s + c, m ) ;
22    x = [ x s ] ;
23 end
```

Legg merke til eit par finessar i Matlab.

1. Kodeordet `persistent` let oss definera lokale variablar som funksjonen hugsar frå kall til kall. Dette svarer til `static` i C.
2. Me reknar med 32-bits heiltal. Sidan Matlab elles bruker flyttal, må alt konverterast med `int32()`-funksjonen.
3. Me kan sjekka om me har fått alle argumenta, ved å sjekka `nargin` som gjev talet på argument.

**Oppgåve 5.16** *Last ned fila rng25.m. Dette er den same funksjonen som er gjengjeve over. Test funksjonen på kommandolina og genererer 25 tilfeldige tal. Vel din eigen verdi for frøet $s$.*

```
1 rng25(25,s)
```

*Drøft med sidemannen: ser tala tilfeldige ut?*

**Oppgåve 5.17** *Prøv funksjonen over igjen, og generer 100 tilfeldige tal. Drøft med sidemannen: ser tala tilfeldige ut?*

**Oppgåve 5.18** *Kopier fila `rng25.m` og lag ein ny funksjon der du endrar den multiplikative parameteren til $a = 9$. Kva skjer? Kva fortel dette oss om desse slumptalsgeneratorane?*

**Oppgåve 5.19** *Lag ein tredje variant, med $a = 10$. Kva skjer denne gongen?*

### 5.3.2. Perioden

**Definisjon 15** *Perioden i ei fylgje $[x_1, x_2, \ldots]$ er det minste talet $m > 0$ slik at $x_i = x_{i+m}$ for alle $i > k$ for ein eller annan $k$.*

Ein fylgje som har periode $m$ repeterer seg sjølv for kvart $m$te element. Merk at der kan vera ein serie med meir enn $m$ distinkte tal i starten av fylgja, men frå eit eller anna punkt $k$ kan me sjå periodiske syklar med fast lengd.

**Oppgåve 5.20** *Sjå på fylgjene som du genererte med dei tre kongruensgeneratorane over (for $a = 6, 9, 10$). Kva periode har fylgjene? Er periode uavhengig av frøet $s$?*

**Oppgåve 5.21** *Kva er den største perioden du kan tenkja deg for ein kongruensgenerator med parameter $a$, $c$ og $m$.*

**Sats 3** *Den lineære kongrunsgeneratoren $s_i = as_{i-1} + c \mod m$ har periode $m$ dersom, og berre dersom,*

1. *$c$ og $m$ er relativt prim (mao. $\text{hcf}(c, m) = 1$)*
2. *$p$ deler $b = a - 1$ for alle primtal $p$ som deler $m$*
3. *$4$ deler $b = a - 1$ dersom $4$ som deler $m$*

### 5.3.3. Uniform fordeling

Ein god slumptalsgenerator skal gje slumptal som er uavhengige og uniformt fordelte. Dersom me ser på ein heil periode frå ein lineære kongruensgenerator, er dette aldri tilfredsstilt. Generatoren kan ikkje gjenta tal innanfor perioden, so når periode går mot slutten veit me at neste tal må vera eit av dei som me ikkje har sett før. Dersom me ser på ein liten del av perioden (og parametra $a$, $c$ og $m$ er velvalde), kan me derimot få ein fordeling som er tilfredsstillande nær uniform og uavhengig.

**Oppgåve 5.22** *Lag to nye matlabfunksjonar `rng1.m` med $a = 219$, $c = 7$ og $m = 32749$ og `rng2.m` med same $a$ og $c$ men $m = 2^{15} = 32768$. Merk at den fyrste $m$-verdien er eit primtal.*

Me skal testa dei to slumptalsgeneratorane, og vurdera om fordelinga er uniform. Dette kan me ikkje gjera når utfallsrommet er større enn utvalet. Difor kan me ikkje sjå eit slumptal $x$ direkte for generatoren, men dersom $x$ er uniformt fordelt, so vil ogso $x \mod q$ vera uniformt fordelt.

**Oppgåve 5.23** *Bruk `rng1.m` og generer ein vektor $x$ med 1000 tilfeldige tal. Lag so ein ny vektor $y$ som er $x$ redusert elementvis modulo 16. Lag eit histogram av $y$. Dvs.*

```
x = rng1(25,s)
y = mod(x,16)
histogram(y, 'BinMethod','integers')
```

*Ser det ut som om tala er uniformt fordelte?*

**Oppgåve 5.24** *Gjenta øvinga over for `rng2.m`. Er dette uniformt fordelt?*

Testane over ser på dei fire minst signifikante bitsa i slumptala. Me kan òg sjå på dei fire mest signifikante sifra.

**Oppgåve 5.25** *Bruk* `rng1.m` *og generer ein vektor* $x$ *med 1000 tilfeldige tal. Lag so ein ny vektor* $y$ *som er* $x$ *med dei fire mest signifikante sifra, ved å dela på* $2^{11}$*. Lag eit histogram av* $y$*. Dvs.*

```
x = rng1(25,s)
y = x/2^11
histogram(y, 'BinMethod','integers')
```

*Ser det ut som om tala er uniformt fordelte?*

*Gjenta øvinga med* `rng2.m`*. Kva ser du?*

### 5.3.4. Frøet i Matlabs slumptalsgenerator

**Oppgåve 5.26** *Finn fram* `cointrial`*-funksjonen din frå forrige veke. Simuler åtte kast med tre myntar per kast. Gjenta denne simuleringa tre gongar. Får du same resultat kvar gong?*

**Oppgåve 5.27** *Gjennta forrige oppgåve, men køyr kommandoen* `rng(243)` *før kvart kall til* `cointrial`*. Kva skjer no? Kvifor?*

I stokastisk simulering er det mogleg å setja frøet manuelt før simuleringa. Ved å bruka same frø, kan ein gjenta eksakt same simulering determistisk. Somme tider er det nyttig, men det legg òg eit stort ansvar på brukaren for å sjå til at frøet er fornuftig (tilfeldig) valgt.

### 5.3.5. Statistiske testar (Ekstra)

(Dette avsnittet er ein forsmak på noko som me kjem tilbake til seinare i semesteret. Dersom du ikkje rekk over det, so er det ikkje noko problem.)

Øving 5.22 tok utgangspunkt i ein hypotese, at slumptala frå `rng1.m` er uniformt fordelte. Dersom denne hypotesen er sann, so er det òg sant at slumptala modulo 16 er sann. Når me ser på datamaterialet, i form av eit histogram, kan me vurdera om denne hypotesen er rimeleg eller ikkje.

Dersom histogrammet ikkje ser ut som ei uniform fordeling, er det rimeleg å tru at det ikkje er generert av ei uniform fordeling, og hypotesen er usann. Me går då ut frå at slumptalsgeneratoren er dårleg.

Dersom histogrammet ser ut som ei uniform fordeling, er det rimeleg å gå ut frå at hypotesen er sann, og me generatoren har i alle fall ikkje denne dårlege eigenskapen.

Slik visuell vurderinga av datamaterialet vert mykje synsing. For å gjera ein objektiv vurdering ynskjer me enkle kvantitative svar.

Lat oss ta utgangspunkt i histogrammet igjen. Lat $[y_1, y_2, \ldots, y_n]$ vera utvalet vårt, dvs. ei fylgje av slumptal modulo 16.

1. Lat $F_y$ vera frekvensen av verdien $y$ i utvalet.

   Dvs. $F_y$, for $0 \leq y \leq 15$ er talet på gongar $y$ førekjem i utvalet. Histogrammet plottar $F_y$ for kvar $y$.

2. Lat $E_y$ vera forventingsverdien til $F_y$, dersom hypotesen vår er sann.

   Hypotesa seier uniform fordelinga, og då er $E_y = n/16$ der $n$ er storleiken på utvalet.

3. Me reknar ut den stokastiske variabelen

$$G = \sum_{y=0}^{15} \frac{(F_y - E_y)^2}{E_y}.$$

Variabelen $G$ er eit standardverkty for å samanlikna ein empirisk fordeing ($F_u$) med ein hypotetisk fordeling (uniform i dette tilfellet). Det er lettare å sjå avvik i ein enkelt skalarvariabel $G$, enn å sjå på heile histogrammet med seksten forskjellige frekvensar.

**Oppgåve 5.28** *Sjå på uttrykket for $G$. Kva verdiar kan $G$ ta? Korleis ser histogrammet ut når $G$ tek minste mogleg verdi?*

**Oppgåve 5.29** *Kva verdiar ventar du at $G$ har når hypotesen om uniform fordeling held? Kva når ho ikkje held?*

**Oppgåve 5.30** *Bruk `rng1.m` som i førre avsnitt og lag eit utval på $n = 1000$ tilfeldige tal modulo 16. Finn frekvensane $F_y$ vha. funksjonen*

```
f = histcounts(y,'BinMethod','integers')
```

*Rekna ut $G$ som forklart over. Kva verdi får du?*

*Gjer det same for `rng2.m`.*

**Oppgåve 5.31** *Variabel $G$ er stokastisk med $\chi^2$-fordeling med 15 fridomsgradar. Plott sannsynsfordelinga*

```
fplot( @(x)chi2pdf(x,15), [0 40] )
```

*Det merkelege uttrykket `@(x)chi2pdf(x,15)` er eit lambdauttrykk og lagar ein ny funksjon med ein parameter $x$ vha. den eksisterande funksjonen som har 2.*

*Samanlikna dine observasjonar av $G$ frå forrige oppgåve med sannsynsfordelinga. Synest du observasjonane dine ser sannsynlege ut dersom slumptala er uniformt fordelte?*

### 5.3.6. Avslutning

I denne økta har me berre skrapt i overflata på eit stort og viktig tema. Dei generatorane som me har sett på er ganske primitive, og det er tydeleg at dei ikkje er perfekte. Det er viktig å hugsa på at dei same problema går igjen i meir sofistikerte slumptalsgeneratorar. Dei er aldri perfekte, og det vesentlege spørsmålet vil alltid vera om dei er gode nok for eit bestemt bruksområde.

Slumptalsgeneratorar er eit kontroversielt tema. Der er delte meiningar i literaturen om kva som er viktig og om kva som er godt nok.

Det er viktig ikkje å vera naiv i valet av slumptalsgeneratorar. Mange standardbibliotek har implementert generatorar med dårlege statistiske eigenskapar, sjølv om det kanskje ikkje er like ille som det var for ein generasjon sidan. På ein del aktuelle område er dei statistiske eigenskapane kritiske. Tenk deg nettpoker der korta ikkje er uniformt fordelte? Eller eit nettbanksystem der banditar veit at nokre nyklar er meir sannsynlege enn andre? Eller ein vitskapleg simulering der sannsynsfordelinga er ei heilt anna enn det som forskaren føreset?

Me avslutta økta med eit døme på statistisk hypotesetest. Det er eit sentralt tema som me går vidare med neste økt. Hypotesetestar vert brukt på mange andre problem enn kvaliteten til slumptalsgeneratorar.

## 5.4. Veke 4. Gjennomsnitt og varians i Matlab

I denne oppgåva skal me arbeida med gjennomsnitt over $m$ terningkast (D6), og samanlikna standardavvika og histogramma når $m$ varierer. Lat $X$ vera den stokastiske variabelen som gjev gjennomsnittet av $m$ terningkast.

### 5.4.1. Terningkast

**Oppgåve 5.32** *Eit vanleg triks for å simulera 1D6 er å ganga opp eit tilfeldig tal i intervallet* $(0, 1)$ *og runda opp, slik:*

```
x = ceil(6*rand())
```

*Lag ein funksjon* `diceavg(m)` *som kastar* $m$ *terningar og returnerer gjennomsnittet.*

**Oppgåve 5.33** *Bruk funksjonen over til å laga eit utval av* $n = 20$ *observasjonar av gjennomsnittet over* $m = 2$ *terningar. Utvalet bør vera ein vektor med lengd 20.*

- *Rekna ut utvalsvariansen vha.* `var`*.*
- *Rekna ut utvalsstandardavviket vha.* `std`*.*

- *Plot histogrammet som viser den empiriske fordelinga.*

**Oppgåve 5.34** *Vurder histogrammet i oppgåva over. Liknar det på den teoretiske fordelinga? Forsøk å auka utvalsstorleiken $n$. Kor stor må $n$ vera før du ikkje ser forskjell på den empiriske og den teoretiske fordelinga? Denne verdien kallar me $n_2$ og me skal bruka den som utvalsstorleik vidare.*

**Oppgåve 5.35** *Gjenta øvinga over med $n_2$ forsøk og ulike verdiar av $m$. Plott histogrammet for $m = 2, 3, 5, 7, 10, 25, 100$. Kva ser du?*

**Oppgåve 5.36** *Rekna ut standardavviket $s$ for kvart utval i forrige oppgåve ($m = 2, 3, 5, 7, 10, 25, 100$). Plott $s$ som ein funksjon av $m$. Dvs. dersom $S$ er ein vektor med standardavvika, kan du skriva*

```
M = [2,3,5,7,10,25,100]
plot(M,S)
```

*Korleis vil du skildra samanhengen mellom $m$ og $s$?*

**Oppgåve 5.37** *Du hugsar kanskje frå førelesinga at $\sigma^2 = 2 + \frac{11}{12}$ for ein D6, og $\sigma^2 = \frac{1}{m}(2 + \frac{11}{12})$ for snittet av $m$ D6. Me kan finna sannsynsfordelinga for normalfordelinga med same standardavvik og forventing vha. fylgjande komando:*

```
m = 100
mu = 3.5
sigma2 = (2+11/12)/m
sigma = sqrt(sigma2)
fplot(@(x)pdf('Normal',x,mu,sigma),[1,6])
```

*Samanlikna dette plottet med histogrammet ($m = 100$) frå tidlegare oppgåver. Kva ser du?*

Det er vanleg å skriva $\mu = E(X)$ for populasjonsgjennomsnittet. Den greske bokstaven $\mu$ vert kalt *mu* på engelsk; difor `mu` i matlabkoden over.

### 5.4.2. Mynt og kron

Me kan gjera eit tilsvarande eksperiment med mynt og kron. Denne gongen får du mindre detaljert hjelp, men du kan fylgja same metode som med terningkastet over.

**Oppgåve 5.38** *Simuler eit kast med $m$ myntar; gjenta forsøket $n$ gongar, og plott eit histogram for fordelinga for $m = 2, 5, 10, 25, 50$. Vel $n$ slik at histogrammet ser ut til å gje eit godt inntrykk av den teoretiske fordelinga. Samanlikn resultat med det du hadde for terningane. Kva likskap og skilnad ser du?*

### 5.4.3. Eksterne datasett

I dei fylgjande oppgåvene skal du gjera deg kjend med nokre av Matlab sine funksjonar for å lasta og arbeida med eksterne datasett.

**Oppgåve 5.39** *Last ned datasettet* `http://jmaurit.github.io/data/oil_fields_cross.csv` *Datasettet inneheld ei liste over oljefelt. Opn fila i ein teksteditor og sjå på søyletitlane. Kva trur du søylene inneheld?*

**Oppgåve 5.40** *Last fila i Matlab:*

```
tbl = readtable('oil_fields_cross.csv')
```

*Kor mange oljefelt har du data om? Du kan bruka* `size`*-funksjonen.*

**Oppgåve 5.41** *Lat oss sjå nærmare på storleiken på felta, definert etter mengda utvinnbar olje:*

```
x = tbl.recoverable_oil
```

*Dette gjev deg ein matrise. Kva type har elementa i matrisa?*

**Oppgåve 5.42** *Konverter dataa til flyttal:*

```
y = str2double(x)
```

*No har du ei numerisk matrise, men nokre felt har verdien* `NaN` *(not a number). Kvifor? Samanlikn med den opprinnelege filen.*

**Oppgåve 5.43** *Prøv funksjonen*

```
ismissing(y)
```

*Kva er resultatet?*

**Oppgåve 5.44** *For å få ein bolsk matrise med 1 for kvar rekkje der* `y` *manglar data, kan du bruka funksjonen:*

```
~ ismissing(y)
```

*Tilda tyder negasjon.*

*Denne bolske matrisa kan brukast som indeks:*

```
z = y(~ ismissing(y))
```

*Samanlikn* `y` *og* `z`*. Kva ser du? Kor mange observasjonar har du i* `z`*? (Bruk* `size`*.)*

**Oppgåve 5.45** *Rekn ut gjennomsnittet og standardavviket for storleiken på dei oljefelta du har data om.*

### 5.4.4. Varians og andre spreidingsmål

**Oppgåve 5.46** *I denne oppgaven studerer skal vi samanlikna variansen med andre spreidingsmål som me kunne ha brukt.*

1. *Lag eit utval med ti terningkast:*

```
n=10
x=ceil(6*rand(1,n))
t=1:n
plot(t,x,'LineStyle','none', 'Marker','diamond')
```

   *No er* `x` *ein vektor med ti observerte terningkast. Plottet visualiserer dei ti kasta.*

2. *Rekn ut utvalsmiddelverdien* `xbar = mean(x)`*.*

3. *Rekn ut avvika* `xdiff = x - xbar`*. Kva vil det seia å ta differansen mellom ein vektor og ein skalar?*

4. *Rekn ut gjennomsnitleg avvik:*

   (16) $$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}).$$

   *Ta utgangspunkt i* `xdiff` *som du rekna ut over, og hugs* `mean`*-funksjonen.*

5. *Rekn ut gjennomsnittet av absoluttverdien på avvika:*

   (17) $$\frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|.$$

   *Du kan bruka* `abs`*-funksjonen i Matlab.*

6. *Variansen er, som me hugsar*

   (18) $$s = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

   *Dette kan reknast ut som*

```
s2 = sum( xdiff .^ 2 )
```

*Punktumet i* `.^` *tyder at operasjonen skal utførast elementvis på matrisa. Kva er variansen på terningkasta dine?*

7. *Rekn ut standardavviket $s$ òg.*

**Merknad 6** *Det er kanskje ikkje openbert kvifor variansen er eit betre mål enn gjennomsnittleg absoluttavvik. Ein grunn er at variansen er kontinuerleg deriverbar, det er ikkje absoluttverdien.*

## 5.5. Innlevering 1

I prosjekt 1 skal du levera inn ein rapport som svarer på fylgjande.

1. Drøft og svar på fylgjande spørsmål:
   a) *Kva er viktig å tenkja på når du skal bruka ein lineær kongruensgenerator for å simulera stokastiske prosessar?*
   b) *Korleis påverkar utvalsstorleiken (t.d. talet på terningar) standardavviket?*
2. Illustrer svar over med døme frå simuleringane dine (hhv. veke 3 og veke 4).
3. Skriv eit refleksjonsnotat (ca. éi side), som oppsummerer det du har lært gjennom dette prosjektet. Svar t.d. på
   - Kva har du lært?
   - Kva er vesentleg for vidare studium og karriere?
   - Kva er vanskeleg?
   - Korleis kan du arbeida for å læra mest mogleg?
   - Korleis fungerer læringsaktivitetane? Kan dei leggjast betre til rette?

# 6. Prosjekt 2

## 6.1. Heimearbeid: agentbasert modellering

Dette prosjektet går over tre labøvingar, og det er tiltenkt grupper på 3–4 personar. De skal gå gjennom eit fullstendig simuleringsscenario, med stega

1. Modellering (labøkt 1. februar 2019)
2. Implementasjon (labøkt 8. februar 2019)

3. Analyse (labøkt 8. mars). Det er kritisk at de har ein simulator som fungerer *før* analyseøkta.

Legg merke til at dette prosjektet startar før, men vert avslutta etter Prosjekt 3. Grunnen til det er at dette prosjektet gjev stor fridom, og mange studentar utnyttar det til å gjera mykje ut av implementasjonen.

Det overordna målet kan oppsummerast slik:

**Oppgåve 6.1** *Implement a predator-prey simulator with visualisation for an eco-system with species of your choice. You should use the* rabbit and fox *example from Barnes and Kolling as a starting point, but also add a few new features of your own, in order to make the model more realistic. Such additions could include food for the prey species (e.g. grass), more detailled movement patterns, or other things. You can take inspiration from Project 4 (wolf/sheep/grass) of Shiflet & Shiflet, page 512 (see attachment).*

*Use the similator to predict (estimate) average life span of each of the species involved. Optionally, you can extract other statistical data for analysis as well. Work in a group of three (or two if numbers don't fit).*

## 6.2. Teori: agentbasert modellering

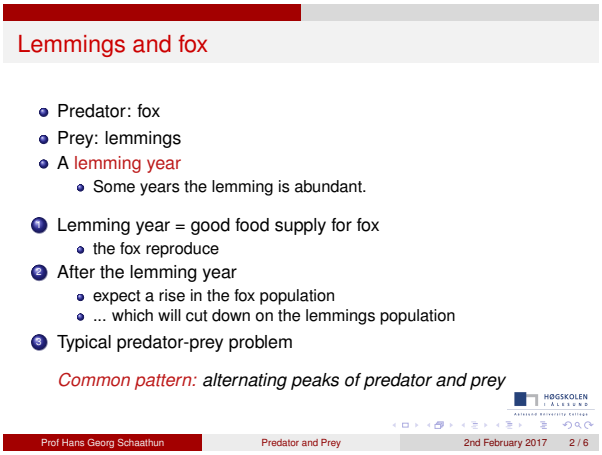
Agentar er eit konsept både innanfor modellering og innanfor programvarearkitektur. Agentbaserte modellar kan implementerast vha. programvareagentar, men òg på andre måtar. Programvareagentar kan likeeins brukast til meir enn berre modellering og simulering.

Teorigrunnlaget som me presiserer her er pensum. Det vil nok gje mest meining om ein ser det i samanheng med labprosjektet, men i år er det forventa at ein kan gjera presist greie for agentbasert modellering og samanlikna med alternative modelleringsformar.

Videoane er gamle. Dei kan vera nyttige, men er ikkje direkte tilpassa framstillinga i år.

### 6.2.1. Ulike tilnærmingar til modellering

Lat oss ta rovdyr/byttedyr-problemet som døme. I eit område lever der rev og kanin. Båe artane formerer seg og reven et kanin. Korleis kan me modellera og simulera utviklinga i populasjonen? Simuleringa kan brukast til å vurdera sannsynet for at ein eller båe artane døyr ut.




**Les:** Shiflet og Shiflet s. 224ff

**Foilar:** PDF

**Alt. 1. Differentiallikningar** Dersom me berre er interessert i populasjonstala, og ikkje geografien der artane lever, kan me laga ein enkel modell med differentiallikningar. Lotka-Volterra-modellen er den mest kjende.

Lat $x$ og $y$ vera talet på hhv. kanin og rev. Observasjonane som ligg til grunn er at

- når der er mykje rev (stor $y$), går kaninpopulasjonen ned (negativ $x'$).
- når der er mykje kanin (stor $x$), er endringa stor (negativ $|x'|$).
  - Mange kaninforeldre gjev mange kaninungar.
  - Reven finn lettare mat når der er mange kaninar og et meir.
- når der er mykje kan (stor $x$), går revepopulasjonen opp (positiv $y'$) pga. mattilgangen.

Dette gjev likningane

$$\frac{dx}{dt} = x(\alpha - \beta y) \tag{19}$$

$$\frac{dy}{dt} = -y(\gamma - \delta x), \tag{20}$$

der $\alpha$, $\beta$, $\gamma$ og $\delta$ er positive konstantar.

Legg merke til at denne modellen er deterministisk. Ingenting er tilfeldig. Han modellerer berre på makronivået, med heile populasjonstal. Individet er uinteressant.

Ein kan simulera Lotka-Voltera-modellen ved iterera over tida $t$. For kvart tidssteg kan ein estimera ut populasjonane for neste steg ved hjelp av forrige verdi av $x$ og $y$ saman med $x'$ og $y'$.

**Alt. 2. Celleautomaton** Ein meir detaljert modell krev at me modellerer landskapet der reven og kaninane bur. Det er vanleg å bruka eit rutenett, slik at dyra alltid flyttar seg i diskrete steg. Det er enklare enn ein kontinuerleg modell. Rutene kallar me gjerne for celler.

Ein automaton er ei tilstandsmaskin. Ein celleautomaton er eit rutenett der kvar celle har eit endeleg tal moglege tilstandar. Til dømes kan tilstandane vera rev, kanin eller tom.

Modellen må ha ein starttilstand, dvs. kvar celle er anten rev, kanin eller tom når modellen startar. Dynamikken i systemet vert avgjort av ein serie med reglar, t.d.

1. Rev med kanin i nabocelle

   $\rightarrow$ reven flyttar til nabocella og kaninen forsvinn

2. Kanin med ein tom nabo

   $\rightarrow$ kaninen flyttar til den tomme cella (gjerne tilfeldig)

3. Kanin når aldersgrensa

   $\rightarrow$ tom celle (kaninen døyr av alderdom)

4. Tom celle med to naboar med kanin.

   $\rightarrow$ kanin (fødsel)

For å køyra modellen itererer ein over tida $t$. For kvart tidsteg vert kvar regel køyrd etter tur, og regelen vert brukt parallelt på alle cellene i rutenettet.

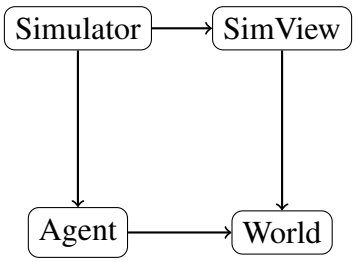
Dette er ein modell som er lett å implementera på ein parallell arkitektur (t.d. GPU) og køyra raskt med relativt store rutenett og mange dyr. Det er derimot ikkje lett å definera eit detaljert regelsett. I dømet over, t.d., kan me sjå at reven sit i ro til der kjem ein kanin forbi og to kaninar kan lett flytta inn i same celle slik at den eine kaninen forsvinn.

**Alt. 3. Agent-basert simulering** Agent er ein som handlar (frå latin *agere*). Kaninar og revar er agentar, dei handlar, dvs. dei et, jaktar, flyktar, osv.

Agentbaserte modellar fokuserer på individet, dvs. kvar einskild agent. I prinsippet kan ein modellera ein unik oppførsel for kvar agent, og det er lett å modellera mykje meir enn to ulike artar i økosystemet.

Dersom me skal modellera rev/kanin-problemet med agentar vil me gjerne bruka eit rutenett tilsvarande celleautomatonen, men i staden for å bruka eit lite regelsett uniformt over heile rutenettet, vil me tenkja på kvar agent som ei autonom skapning som bestemmer kva ho vil gjera. Me skal utdjupa dette i dei neste avsnitta.

Agentbaserte modellar gjer det enkelt å skriva detaljerte modellar for korleis kvart dyr oppfører seg, slik at ein får gode mikromodellar. Simulering kan derimot krevja meir maskinressursar dersom der er mange agentar i modellen, fordi koden for kvar agent i prinsippet er unik og gjerne komplisert.




Figur 3: Agentbasert simulatormodel med visualisering.

### 6.2.2. Programvareagentar

Programvareagentar er ein autonom programvaremodul. Der er ein del sprik i definisjonane, men ein reknar gjerne med at agenten ikkje er bunden til ei spesifikk maskin, men kan flyttast t.d. gjennom lastbalanse. Der er låg kopling mellom agentane, men ein eller annan kommunikasjonsprotokoll. Agenten sjølv bestemmer når og korleis han skal handla.

JADE er eit bibliotek og programmeringsrammeverk for agentar i Java. Me skal ikkje gå inn i detaljane for programvareagentar utanom simuleringssamanheng.

### 6.2.3. Agent-basert simulering

Agent-baserte modellar fokuserer på individetet (agenten), og når me implementerer ein slik modell for simulering, må me halda dette fokuset. Agentane er autonome og handlar ut frå si eiga (lokale) oppfatting av verda. Det tilseier låg kopling mellom agentane.

Det er mogleg å bruka agent-rammeverk som JADE for å implementera agentbaserte modellar, men ein kan òg velja ein enklare programvaremodell der agentane er objekt. Grovarkitekturen er vist i Figur 3, med fire kritiske klasser.

*Agent* er gjerne ei abstrakt klasse og representerer alle moglege agentar i modellen. T.d. vil rev og kanin vera underklasser av agent. Agentklassa treng ein `act()`-metode som simulatorklassa kaller kvar gong agenten får handla, og denne metoden bestemmer oppførselen som gjerne er forskjellig for kvar underklasse.

*World* er modellerer verda der agentane bur. Agentane må kunna spørja verda kva dei ser rundt seg, og det vil seia at verda må halda greie på kvar agentane er, og evt. andre ting som finst i verda.

*Simulator* er selve kjerna i simulatoren. Normalt vil der berre vera eitt simulatorobjekt (gjerne singleton-mynster), og det har ansvar for å instantiera verda og agentane, halda greie på tida, og gje alle agentane høve til å handla (kalla `act()`) for kvart tidssteg.

Desse tre klassene utgjer ein fullstendig simulator, men ingen visualisering og helst ikkje annan I/O heller. Der kan vera andre hjelpeklasser, men dei bør vera knytt til éi av klassene og

ikkje bidra til sterk kopling mellom dei.

*SimViz* er visualiseringa. Dette objektet må for all del ikkje påverka dei andre tre klassene (simulatoren), men tek informasjon frå dei for å visualisera tilstanden i simulatoren.

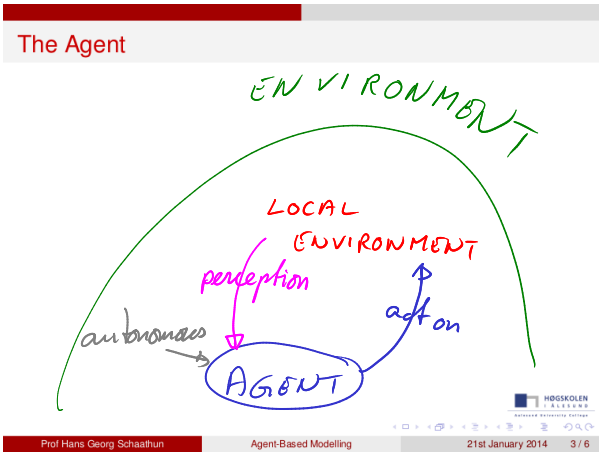
Me kjenner igjen grunnleggjande designprinsipp, med *low coupling* og *high cohesion*. Ei klasse for I/O og berre I/O. Simuleringa (utrekningane) er delt mellom tre klasser for henholdsvis tida, individa og landskapet. Dette er den mest grunnleggjande og grovkorna oppgåvedelinga. Komplekse modellar kan krevja rafinering og meir finkorna inndeling.

### 6.2.4. Vidare lesing og prøving

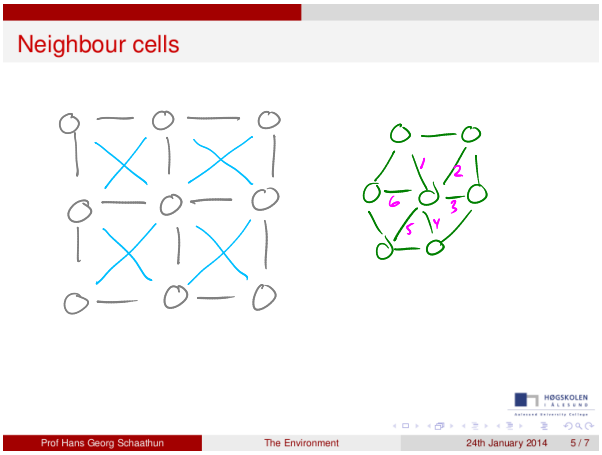
1. Barnes og Kölling har eit døme med simulering av *foxes and rabbits* (sjå neste labøving).
2. Greenfoot er eit pedagogisk verkty for å læra Java. Det har ein agentmodell i botn, og det er lett å arbeida med simuleringar som *foxes and rabbits*.
3. JADE er eit profesjonelt bibliotek og programmeringsrammeverk for agentar i Java.

### 6.2.5. Innføringsvideoar

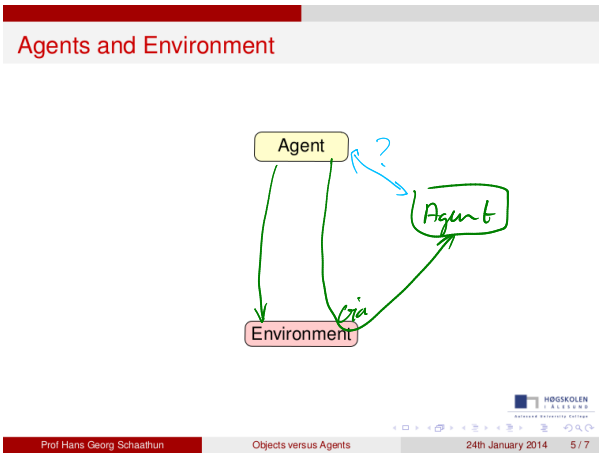
*Videoane er utvikla for kurset i 2015, men er stadig relevante som ein introduksjon.*




**Foilar:** PDF




**Foilar:** PDF




**Les:** Barnes og Kölling s. 326-342

**Foilar:** PDF

## 6.3. Veke 5. Modellering

Målet i labøvinga i dag er å laga ein agent-basert modell for eit økosystem med rov- og bytte-dyr. Neste veke skal de implementera denne modellen som ein simulator. Det er best å arbeida i grupper på tre, kanskje fire personar.

Alle oppgåvene inneber at de må diskutera modellen i gruppa og gjera designval. Det er viktig å dokumentera vala etter kvart, og skrive ned dei vurderingane de har gjort. Helst skal ein skrive ned alternativ som vart forkasta òg, med grunnar både for og imot.

**Oppgåve 6.2 (Plenumsdiskusjon)** *Diskuter kva me meiner med* simulering.

**Oppgåve 6.3 (Plenumsdiskusjon)** *Kva meiner me med agentbasert simulering?*

### 6.3.1. Real world scenario

**Oppgåve 6.4** *Velg artar for rov- og byttedyr som de vil modellera. Slå opp nokre relevante fakta og detaljar om artane, t.d. fruktbarheit, levetid, osv.*

*Målet er ikkje ein fullt ut realistisk simuleringsmodell, det vil kreva fagkompetanse i biologi som me ikkje har. Formålet med denne øvinga er å reflektera på samanhengen mellom røynd og modell og å vurdera kor realistisk modell som er hensiktsmessig.*

**Oppgåve 6.5** *Rov- og byttedyra som de har funne er agentane i modellen. Diskuter kva eigenskapar dei må ha.*

**Oppgåve 6.6** *Diskuter oppførselen åt byttedyragentane. Formuler oppførselsreglar som de trur de kan imlementera. Prioriter desse reglane i fire avsnitt, for* must have, should have, can have, *og* not this project. *(Dette er ein vanleg prioriteringsmetode i prosjektstyring.)*

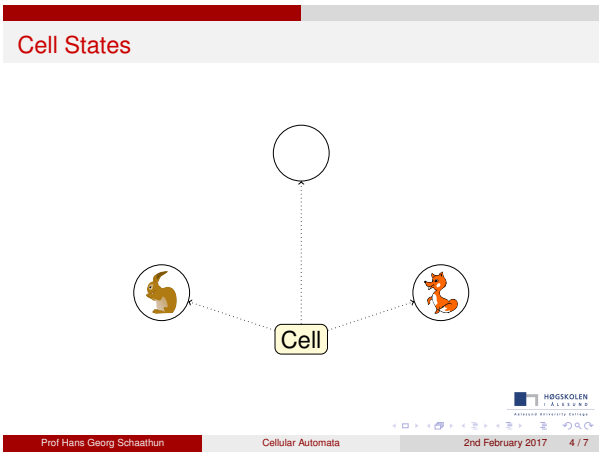
**Oppgåve 6.7** *Diskuter oppførselen åt rovdyragentane. Formuler oppførselsreglar tilsvarande forrige oppgåve.*

**Oppgåve 6.8** *Eit kritisk punkt er modellen for landskapet der dyra beveger seg. Det er vanleg å bruka eit rutenett.*

1. *Kor stort skal rutenettet vera?*
2. *Kor stort areal i røynda svarer til ei rute i modellen? (Mål i meter eller kilometer)*
3. *Kva skjer i ytterkantane av rutenettet?*

*Teori og døme finn du i avsnitt 6.4.*

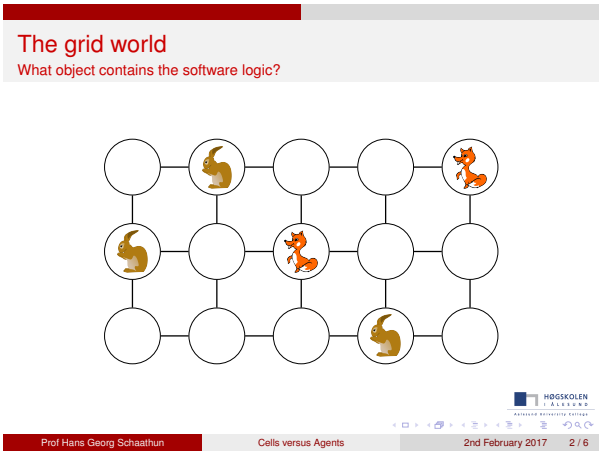
## 6.4. Video: landskap- og cellemodellar




**Les:** Shiflet og Shiflet s. 510ff

**Foilar:** PDF


**Foilar:** PDF




**Foilar:** PDF

## 6.5. Veke 6. Implementasjon av predator/prey

Denne delen av prosjektet er venta å for stor til å verta fullført på éi labøkt. Difor tek me ei pause frå dette prosjektet for å gje tid til å fullføra implementasjonen. I labøkta veke 10 skal de bruka implementasjonen.

### 6.5.1. Getting started

**Oppgåve 6.9** *Download the Barnes and Kölling's «Book projects» from the BlueJ web page, and find the «foxes and rabbits» project.*

**Oppgåve 6.10** *Review the «foxes and rabbits» example from Barnes and Kölling. Check that it compiles and runs. Even if it does not simulate 'your' species, it is still your first running*

*prototype simulator.*

It is suggested that you use Java and build on the example by Barnes and Kölling. You should keep the following requirements in mind throughout the project.

1. The simulator must be able to produce a log file with comma-separated values (CVS) recording statistically data.

   As a minimum, you should record

   - the age and time of death for every agent must be recorded.
   - the population sizes for each timestep.

2. Make some (not all) of the model features configurable, so that you can experiment with variations of the model (e.g. starting population size, grid size, etc.).
3. Visualise the location of individual agents at each time step. (The example from Barnes and Kolling tells you how to do this.)

**Oppgåve 6.11** *Review the OO model from last week's lab session. Discuss what amendments you need to make to the prototype simulator to implement* your *model.*

1. *Write the amendments down in order of priority.*
2. *Allocate one amendment (task) to each person to do first.*
3. *Whenever one amendment is made, check that the prototype still runs.*

*As tasks (amendments) are complete, allocate new ones to keep busy.*

### 6.5.2. Continued development

Since Barnes and Kolling gives you a working prototype, you are well placed for an iterative (agile) development approach. The following exercises are intended as an aid to structure the process.

**Oppgåve 6.12** *Review your runnning prototype and discuss how to implement the features above.*

1. *Divide the work into tasks, in order of priority.*
2. *Allocate one amendment (task) to each person to do first.*
3. *Whenever one amendment is made, check that the prototype still runs.*

*Continue allocating and completing tasks until all the* necessary *features are implemented.*

**Oppgåve 6.13** *Evaluate your simulator. Which features have been implemented to satisfaction? What remains to be desired?*

1. *Write down all features that you would like to add as potential tasks.*
2. *Prioritise the tasks, and order them accordingly.*
3. *Split the list into three sections, of (1) tasks which* must be complete *before the presentation, (2) tasks which you* hope to complete*, and (3) the rest.*
4. *Take one task each and start implementing new features.*
5. *Whenever one amendment is made, check that the prototype still runs and discuss whether it is satisfactory.*

*Continue allocating and completing tasks until you are satisfied.*

## 6.6. Veke 10. Analyse av predator/prey

I denne siste delen av prosjektet, skal me sjå på data frå simulering. Det er difor det er so viktig at simulatoren kan produsera ei loggfil.

**Oppgåve 6.14** *Køyr simulatoren med ulike parametrar. Prøv å skapa desse tre scenarioa:*

1. *Reven døyr ut medan kaninen overlever.*
2. *Båe artane døyr ut.*
3. *Båe artane overlever over lang tid.*

*Ta vare på dei tre loggfilene.*

**Oppgåve 6.15** *Ta dei tre datasetta frå forrige oppgåve og plott populasjonstala for kvar art (eitt plott per scenario). Drøft kva plotta fortel oss.*

**Oppgåve 6.16** *For kvart scenario, finn gjennomsnitt og standardavvik for levealderen åt kvar art.*

**Oppgåve 6.17 (Diskuter)** *Kva er samanhengen mellom populasjonstala for dei to artane?*

**Oppgåve 6.18 (Diskuter)** *Prøv å gjenta simuleringane med dei same parametrane som du brukte i oppgåve 6.14. Gjev same parametrar alltid same scenario (ingen/ein/båe dyra overlever)? Eller er det meir eller mindre tilfeldig kva som skjer? Drøft resultata.*

## 6.7. Innlevering 2

Innleveringa frå prosjekt 2 skal innehalda svar på alle oppgåvene i analysedelen. Legg som vanleg mest vekt på drøftingane. I tillegg skal du (som alltid) ha med ein refleksjon over kva du har lært av prosjektet, dei største utfordringane, samt relevanse for ein karriere som dataingeniør.

# 7. Prosjekt 3

## 7.1. Heime. Domenekunnskap

Temaet for dette prosjektet er estimering av feilsannsyn i kommunikasjonssystem. Me bruker to øvingar

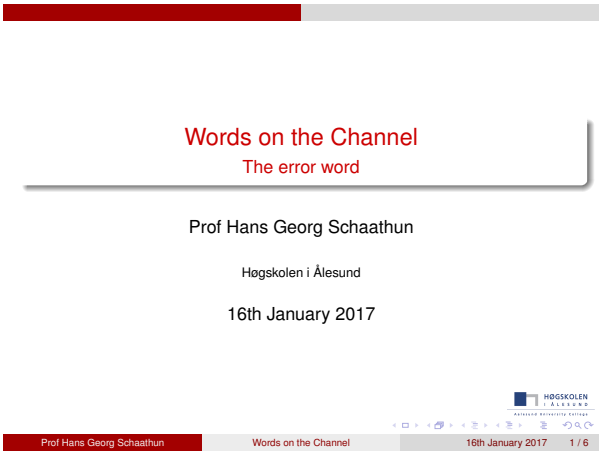
**Veke 6** Implementer ein simulator.

**Veke 7** Estimer feilsannsynet vha. data frå simuleringane.

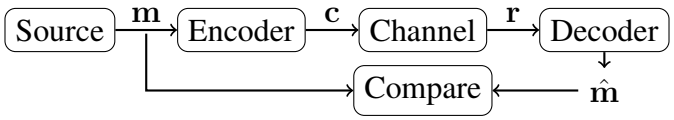
Videoane under gjev ein liten introduksjon til det som er spesielt for domenet (telekommunikasjon og kodeteori).

All telekommunikasjon er utsatt for støy. Det signalet du mottar er *aldri* identisk med det som vart sendt. Vha. feilrettande kodar er det likevel mogleg å koda signalet slik mottakaren kan dekoda det rett med høgt sannsyn.

**Foilar:** PDF




Den binærsymmetriske kanalen (BSC) modellerer sending av éin bit. Stort sett er me interessert i å senda svært mange bits, og det er nyttig å studera kva som skjer når ein sendar eit *ord* på $n$ bits.




Figur 4: Coding system for simulation.

**Foilar:** PDF

## 7.2. Veke 7. Simulering av kommunikasjonssystem

Figur 4 viser kommunikasjonssystemet som me skal simulera.

- $\mathbf{m}$ er meldinga (klartekst) (m for *message*).
- $\mathbf{c}$ er kodeordet (c for *codeword*) som vert sendt på kanalen. Det er lenger enn meldinga. Redundansen vert brukt til å korrigera for feil.
- $\mathbf{r}$ er det mottekne ordet (r for *received*). Dersom kanalen er feilfri får me $\mathbf{r} = \mathbf{c}$.
- $\hat{\mathbf{m}}$ er den dekoda meldinga (hatten $\hat{}$ tyder estimat). Dersom $\hat{\mathbf{m}} \neq \mathbf{m}$ har me ein dekodingsfeil.

Me må modellera og simulera kjelda (source) som genererer tilfeldige meldingar og kanalen (BSC) som tilfører støy mellom sendar og mottakar. Til koding og dekoding kan me bruka implementasjonar av eit ekte kodesystem som me ynskjer å testa. Desse skal me difor ikkje modellera, og me treng ikkje implementera dei sjølve.

«Compare»-boksen er ikkje ein del av kommunikasjonssystemet, men vert brukt av simulatoren for å sjekka resultatet av kvart forsøk og telja feil.

*Viktig.* Me implementerer kvar boks i systemet som ein funksjon (m-fil). Kvar m-fil bør ikkje bruka globale variablar heller skriva utdata på skjermen anna enn ved feilsøking. Kommunikasjon med boksen (funksjonen) skjer ved inn- og utparametrar. Dette gjer at boksane kan brukast igjen i fleire simuleringar.

### 7.2.1. Simulering av kommunikasjonssystemet

**Oppgåve 7.1 (Kjeldesimulator)** *Ein reknar normalt med at meldinga* $\mathbf{m}$ *er uniformt fordelt, over mengda av alle binærvektorar av lengd* $k$*. Dersom meldinga ikkje er uniformt fordelt, løner det seg å komprimera ho fyrst.*

*Skriv ein funksjon (m-fil) som tek ordlengda* $k$ *som argument og returnerer eit tilfeldig meldingsord* $\mathbf{m}$*.*

*Test funksjonen eit par gongar. Er resultata rimelege?*

Når det gjeld kanalmodellen, so skal me simulera ein svært enkel modell, den *binærsymmetriske kanalen* BSC($p$). I røynda er der stor skilnad mellom ulike kanalar. Trådlaust nettverk er forskjellig frå kabla nett, som igjen er ulikt lagringsmedium som optiske og magnetiske plater. Stasjonære antennar er òg svært ulikt mobilt utstyr. Dette kan ein lære meir om i kurs i *telekommunikasjon* og i *kodeteori*. Her klarer me oss med den enkle modellen, og fokuserer på simulering og statistikk.

**Definisjon 16 (Den binærsymmetriske kanalen)** *Den binærsymmetriske kanalen med bitfeilsannsyn $p$ (BSC($p$)) tek ein kodeord $\mathbf{c} = (c_1, c_2, \ldots, c_k)$ som input og returnerer eit motteke ord $\mathbf{r} = (r_1, r_2, \ldots, r_k)$.*

*Kanalen lagar ein tilfeldig* feilvektor $\mathbf{e} = (e_1, e_2, \ldots, e_k)$ *ved å dra kvar bit $e_i$ uavhengig slik at $P(e_i = 1) = p$ og $P(e_i = 0) = 1 - p$.*

*Det mottekne order er*

$$\mathbf{r} = \mathbf{c} + \mathbf{e} \mod 2.$$

**Oppgåve 7.2 (Kanalsimulator)** *Me skal implementera ein funksjon (m-fil) som tek ein meldingsvektor og eit bitfeilsannsyn $p$ som argument og returnerer ein motteken vektor med same lengd, som om han var sendt over BSC($p$). Dette kan gjerast på ulike måtar; det fylgjande er eit forslag:*

1. *Trekk ein tilfeldig feilvektor $\mathbf{e}$. Lag gjerne ein eigen funksjon for det.*
2. *Returner $\mathbf{r} = \mathbf{m} + \mathbf{e} \mod 2$.*

*Test funksjonen/funksjonane eit par gongar. Er resultata rimelege?*

**Simuleringsresultat** Kvar melding som me sender på kanalen er eitt forsøk. Oppgåva for *Compare*-boksen er å rekna ut det resultatet som me ynskjer å observera frå forsøket. Der er to vanlege alternativ:

**Ordfeil** Eitt ord vert sendt i førsøket. Dersom $\mathbf{m} = \hat{\mathbf{m}}$ har me null ordfeil. Dersom $\mathbf{m} \neq \hat{\mathbf{m}}$ har me éin ordfeil.

**Bitfeil** Ordet består av $k$ bits. Kvar bit $\hat{m}_i$ som er ulik den sendte biten $m_i$ gjev éin bitfeil. Talet på bitfeil er ofte eit interessant resultat.

Lat oss testa eit ukoda system før me innfører feilretting. Då har me

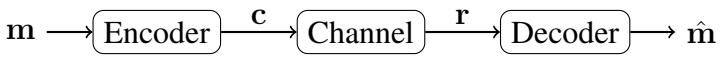
$$n = k, \tag{21}$$

$$\mathbf{c} = \mathbf{m}, \tag{22}$$

$$\hat{\mathbf{m}} = \mathbf{r}. \tag{23}$$

**Definisjon 17 (Hamming-vekta)** *Hamming-vekta $w(\mathbf{x})$ på ein vektor $\mathbf{x}$ er talet på bits som er ulik 0. (Dvs., for ein binær vektor, er Hamming-vekta lik talet på bits som er 1.)*

**Merknad 7** *Talet på bitfeil er hammigvekta $w(\hat{\mathbf{m}} - \mathbf{m})$.*




Figur 5: Channel with coding.

**Oppgåve 7.3** *Skriv ein* Compare*-funksjon som returnerer talet på ordfeil i simuleringa. Input må vera den sendte og den dekoda meldinga.*

**Oppgåve 7.4** *Lag ein funksjon som testar det ukoda systemet $m$ gongar med ordlengd $k$ og returnerer talet på ordfeil. Test funksjonen med nokre ulike ordlengder $k$. Bruk bitfeilsannsyn $p = 0{,}1$. Ser det rimeleg ut?*

*Korleis utviklar ordefeiltalet seg når du aukar $k$?*

**Oppgåve 7.5** *Skriv ein* Compare*-funksjon som returnerer talet på bitfeil i simuleringa. Input må vera den sendte og den dekoda meldinga.*

***Hint** Du kan rekna ut feilordet fyrst, å bruka absoluttverdi og* `sum` *for å løysa problemet.*

**Oppgåve 7.6** *Lag ein funksjon som testar det ukoda systemet $m$ gongar med ordlengd $k$ og registrerer talet på bitfeil.*

*Test funksjonen og plot bitfeiltala som histogram. Prøv nokre ulike ulike ordlengder $k$ og ulike bitfeilsannsyn $p$.*

### 7.2.2. Feilrettande kodar

Feilrettande kodar vert brukt for å hindra kommunikasjonsfeil, som vist i figur 5. Vha. kodeteori er det mogleg å kommunisera påliteleg over svært dårlege kanalar; dvs. sjølv om bitfeilsannsynet $p$ er neste 50%, er det mogleg å få neglisjerbar ordfeilsannsyn. Prisen ein betaler er at svært få meldingsbits, krev mange bits på kanalen.

Her er to kodesystem som me kan testa:

- $[7, 4]$ Hamming-kode
  - Kodar: hammingenc.m
  - Dekodar: hammingdec.m
- $[31, 11]$ BCH-kode
  - Kodar: bchenc.m
  - Dekodar: bchdec.m

Parametrane åt kodane er $[n, k]$, der $n$ er lengda på kodeordet (sendt på kanalen) og $k$ er lengda på meldinga. Hamming-koden over tek altso $4$ bits inn, og lagar eit 7-bits kodeord.

**Oppgåve 7.7** *Last ned kodaren og dekodaren for hammingkoden, og test dei i Matlab.*

1. *Generer ei tilfeldig fire-bits melding* $\mathbf{m}$.
2. *Kod meldinga med* `hammingenc(m)` *og få kodeordet* $\mathbf{c}$. *Korleis ser det ut?*
3. *Dekod* $\mathbf{c}$ *slik at du får* $\hat{\mathbf{m}}$. *Er* $\hat{\mathbf{m}}$ *lik* $\mathbf{m}$ *eller ikkje?*
4. *Lag eit kodeord med éin bitfeil, og prøv å dekoda det:*

```
c1 = mod(c + [0 0 0 1 0 0 0],2)
m1 = hammingdec(c1)
```

*Samanlikn resultatet med den opprinnelege melding* $\mathbf{m}$. *Er det korrekt dekoda?*

Begge testane i øvinga skal gje eit dekoda ord lik det opprinnelege ordet $\mathbf{m}$. I det fyrste tilfellet har du ingen bitfeil på kanalen, og i det andre har du éin. Hammingkoden dekodar alltid korrekt når der er høgst éin bitfeil.

### 7.2.3. Simulator med koding

Ta fram att simulatoren frå forrige veke. I dag skal me utvida han med koding som i figur 5. Dvs. at me må leggja til koding mellom meldingsgeneratoren og kanalen, og dekoding mellom kanalen og samanlikninga. Dette skal me gjera to gongar; både med hammingkoden og BCH-koden.

**Oppgåve 7.8** *Skriv ein funksjon som simulerer* $m$ *forsøk med hammingkoden på BSC, og som tel antall bitfeil* $X$. *Funksjonen må ta* $p$ *(bitfeilsannsynet) som innparameter og returnera ein observasjon av* $X$. *(Hugs at meldinga alltid er* $k = 4$ *bits med hammingkoden.) Simuler* $m = 100$ *sendte meldingar med bitfeilsannsyn* $\pi = 0{,}1$ *og plot eit histogram for* $X$.

**Oppgåve 7.9** *Skriv liknande funksjonar, tilsvarande forrige oppgåve, for BCH-koden og for ordfeil. Totalt skal du ha fire systemsimulatorar:*

1. *Bitfeil i Hamming-koden.*
2. *Bitfeil i BCH-koden.*
3. *Ordfeil i Hamming-koden.*
4. *Ordfeil i BCH-koden.*

*Tenk gjennom API-et slik at du bruker parametrane konsistent i alle fire funksjonane.*

Test gjerne alle fire funksjonane med ulike parametrar dersom du har tid. Det viktigste er derimot å ha dei fire systemsimulatorane klare, slik at me kan bruka dei til statistisk analyse neste veke.

### 7.2.4. Diskusjon

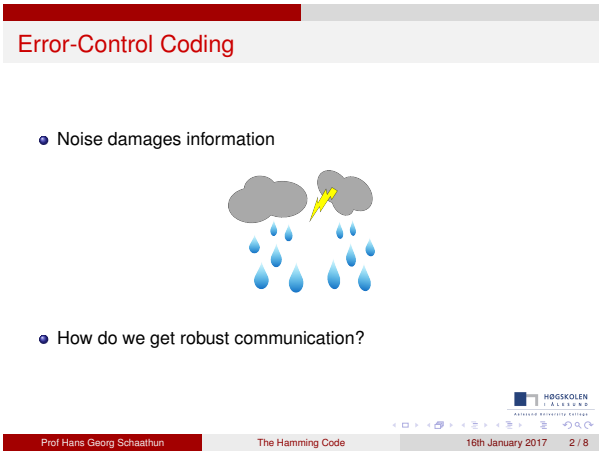
Fylgjande er ein alternativ, og svært vanleg, definisjon på BCH-kanalen.

**Definisjon 18 (Den binærsymmetriske kanalen)** *Den binærsymmetriske kanalen med bit-feilsannsyn $p$ (BSC($p$)) tek ein meldingsord $\mathbf{m} = (m_1, m_2, \ldots . m_n)$ som input og returnerer eit motteke ord $\mathbf{r} = (r_1, r_2, \ldots, r_n)$, der kvar bit $r_i$ er lik $m_i$ med sannsyn $1 - p$ og ulik (feil) med sannsyn $1 - p$. Kvar bit er uavhengig av alle foregåande bits.*

**Oppgåve 7.10 (Ekstra, Diskusjon)** *Er Definisjon 16 og 18 ekvivalent? Korleis kan du vera sikker?*

## 7.3. Video. Kodeteori

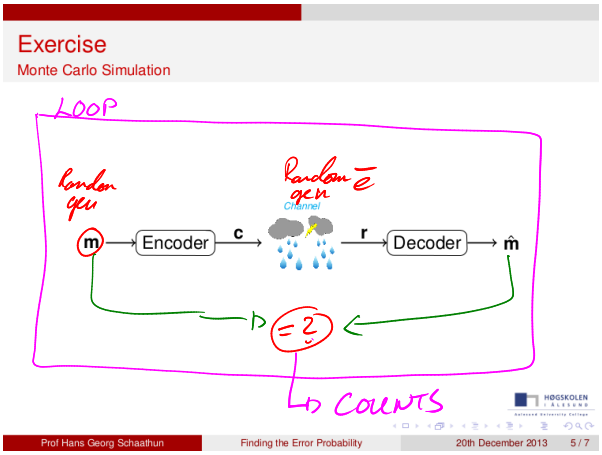
Videoane nedanfor er laga til tidlegare år, og går gjennom stoffet i ein litt annan rekkjefylgje enn i år. Dei er presentert her som repetisjon og mogleg utdjuping.




**Definisjon 19** *Hammingkoden med parametrar $[7, 4, 3]$ er definert ved matrisa*

$$G = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} \tag{24}$$

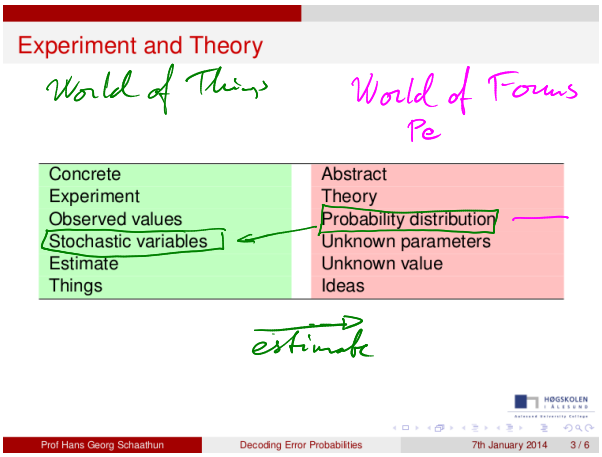
**Foilar:** PDF




**Definisjon 20** *Ein Monte Carlo-simulering er ein stokastisk simulering, dvs. ein simulering av tilfeldige hendingar.*

**Les gjerne** Shiflet og Shiflet ss. 358-360(f).

**Foilar:** PDF




Feilsannsynet er ein av dei viktigaste ytingsparametrane i eit kommunikasjonssystem. Korleis kan me få kunnskap om feilsannsynet?

**Foilar:** PDF

## 7.4. Veke 8. Estimering av feilsannsyn

I denne øvinga skal du bruka simulatoren frå forrige veke.

### 7.4.1. Ordfeil

**Oppgåve 7.11 (Drøfting)** *Tenk deg at du sender $m$ ord og tel talet på ordfeil $Y$. Er $Y$ binomialfordelt? Kvifor (ikkje)?*

**Oppgåve 7.12** *Bruk simulatorane dinne frå førre veka og køyr fylgjande fire simuleringar:*

1. *Hammingkoden (fire bits melding)*
2. *Ukoda, med fire bits melding*
3. *BCH-koden (elleve bits melding)*
4. *Ukoda, med elleve bits melding*

*Køyr $m = 100$ forsøk med bitfeilsannsyn $p_b = 0{,}1$. Registrert talet $Y$ på* ordfeil *i kvar simulering.*

**Oppgåve 7.13** *Ordfeilsannsynet vert som regel estimert som $\hat{P} = Y/m$. Rekn ut estimatet for kvart av dei fire scenarioa i forrige oppgåve.*

**Oppgåve 7.14** *Finn standardavviket for estimatoren $\hat{P} = Y/m$ i kvar simulering.*

**Oppgåve 7.15 (Drøfting)** *Samanlikna scenarioa frå forrige oppgåve. Kva gir mest/minst robust kommunikasjon? Kvifor?*

*Kva fortel standardavviket om samanlikninga.*

**Oppgåve 7.16 (Konfidensinterval)** *Rekn ut eit $95\%$ konfidensinterval for ordfeilsannsynet $p$, for kvar av dei fire eksperimenta.*

**Oppgåve 7.17 (Konfidensinterval)** *Gjenta simuleringane i oppgåve 7.12 med $m = 1000$ sendte ord. Rekn ut eit $95\%$ konfidensinterval for ordfeilsannsynet $p$, for kvar av dei fire eksperimenta.*

*Samanlikn konfidensintervalla med forrige oppgåve. Kva ser du?*

### 7.4.2. Teoretisk fordeling i Matlab

**The probability distribution function (PDF)** I Matlab kan du skriva `pdf('binom',`$x, n, p$`)` for å finna sannsynet $P(Z = x)$ for $Z \sim B(n, p)$. PDF står for *probability distribution function*.

**Oppgåve 7.18** *Gå tilbake til overføringa av fire-bits ord over BSC$(0.1)$. Lat $Z$ vera talet på bitfeil. Bruk Matlab for å finna sannsynet $\mathsf{P}(Z = 0)$, som fylgjer*

```
pdf('binom',0,4,0.1)
```

*Samanlikn svaret med di eiga utrekning tysdag.*

*På same måte, finn* $\mathsf{P}(Z = z)$ *for* $z = 1, 2, 3, 4$*, og samanlikn med di eiga utrekning.*

**Oppgåve 7.19** *Lat* $Z \sim B(4, 0{,}1)$*. Me skal visualisera sannsynsfordelinga for* $Z$ *i Matlab.*

*Fyrst, merk at me kan bruk* `pdf` *på ein vektor eller matrise. Køyr fylgjande i Matlab*

```
zv = 0:4
pv = pdf('binom',zv,4,0.1)
```

*Vektoren* `zv` *er utfallsrommet. Den andre lina reknar ut* $\mathsf{P}(Z = z)$ *for* $z = 0, 1, 2, 3, 4$ *og returnerer ein vektor.*

*No kan me plotta fordelinga*

```
bar(zv,pv)
figure
plot(zv,pv)
```

*I den midterste lina opnar* `figure` *eit nytt figurvindauga slik at du kan sjå begge plotta ved sidan av kvarandre.*

**Oppgåve 7.20** *Bruk teknikken over og plott den teoretiske fordeling for talet på bitfeil når du sender* $n = 10$ *bits over BSC*$(p)$ *for* $p = 0{,}01, 0{,}05, 0{,}1$*. Samanlikn plottet med histogramma du fekk i oppgåve* **??**. *Kva ser du?*

**The cummulative distribution function (CDF)** The `pdf` function (for a discrete distribution) gives you the probability $P(X = x)$. Another important function is `cdf` (*Cummulative Distribution Function*) which gives the probability $P(X \leq x)$.

**Oppgåve 7.21** *Suppose you send a word of 1000 bits over a BSC bit bit error probability* $p = 0.02$*. What is the probability of getting at most ...*

1. $2\%$ *bit errors?*
2. $5\%$ *bit errors?*

*(Use matlab to find the answer.)*

### 7.4.3. Bitfeil

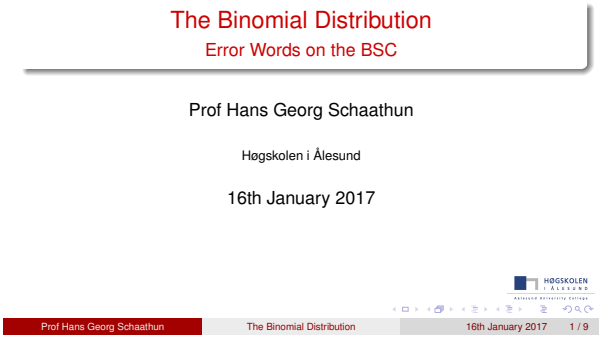
**Oppgåve 7.22** *Bruk simulatorane dinne frå førre veka og køyr fire simuleringar simuleringar, med ukoda og BCH-koden (elleve bits melding) med bitfeilsannsyn* $p_b = 0{,}1$ *og* $p_b = 0{,}2$*.*

*Denne gongen skal du registrera talet på bitfeil $X$ for* kvart *forsøk (sendte ord). Køyr $m = 200$ forsøk og plott eit histogram for kvar simulering (fire plott).*

*Samanlikna dei fire simuleringane. Ser det ut som om bitfeila binomialfordlete? Er der skilnad på fordelinga med og utan koding? Dersom det er vanskeleg å sjå, kan du freista simuleringane på nytt med større verdi for $m$.*

## 7.5. Video. Statistikk

Videoane nedanfor er laga til tidlegare år, og går gjennom stoffet i ein litt annan rekkjefylgje enn i år. Dei er presentert her som repetisjon og mogleg utdjuping.




Talet på bitfeil i eit ord sendt over BSC er eit døme på binomialfordelinga.

**Les:** Frisvold and Moe pp. 100-104.

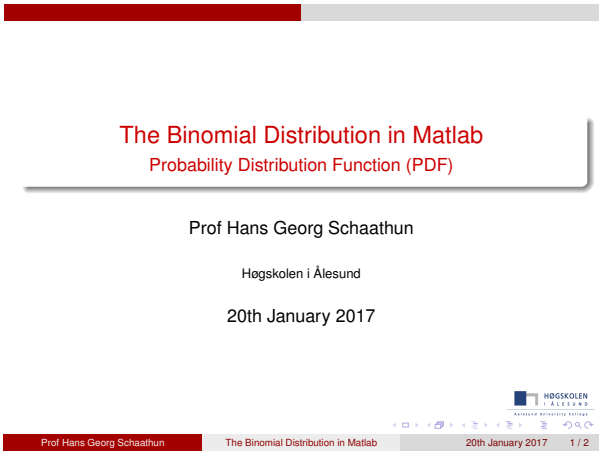
**Foilar:** PDF


Kva er forventingsverdia åt $Z$ når $Z \sim B(n, p)$, dvs. når $Z$ er binomialfordelt med $n$ forsøk og suksessannsyn $p$?

**Foilar:** PDF


Kva er variansen åt $Z$ når $Z \sim B(n,p)$, dvs. når $Z$ er binomialfordelt med $n$ forsøk og suksessannsyn $p$?

**Foilar:** PDF

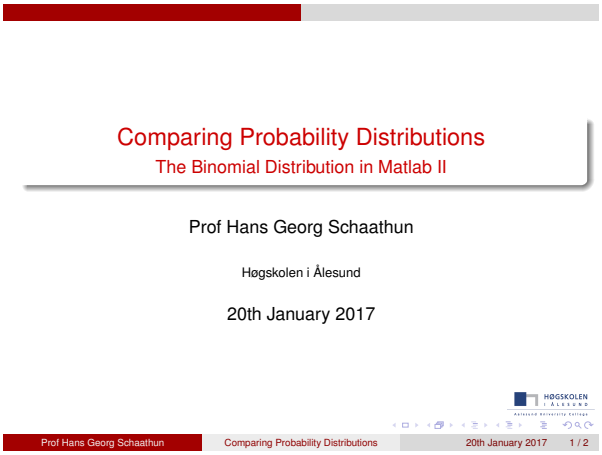



**Døme 11**

*Korleis kan me slå opp $P(Z = z)$ for ein gjeven verdi av $z$ når $Z$ er binomialfordelt, i Matlab?*

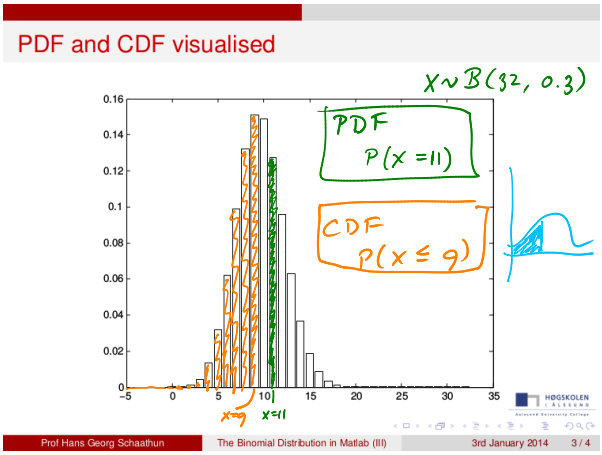
***Les** Matlab* `help`*:* `pdf`*,* `plot`*,* `bar`*,* `figure`*,* `hold`

***Foilar:** PDF*




Korleis kan me samanlikna sannsynsfordelingar i Matlab?

**Foilar:** PDF




**Døme 12**

*Korleis kan me slå opp $P(Z \leq z)$ for ein gjeven verdi av $z$ når $Z$ er binomialfordelt, i Matlab?*

***Les:*** *Frisvold and Moe pp. (55), 56-59, «vanlige forkortelser» p. 61;* `help cdf` *i Matlab*

***Foilar:*** *PDF*

## 7.6. Innlevering 3

Innleveringa frå prosjekt 3 skal innehalda

1. Resultata frå oppgåve 7.16 og 7.17. Legg særleg vekt på å drøfta kva du ser når du samanliknar simuleringane med $m = 100$ og $m = 1000$.
2. Resultata frå oppgåve 7.22. Legg vekt på drøftinga.
3. Ein refleksjon over kva du har lært av prosjektet, dei største utfordringane, samt relevanse for ein karriere som dataingeniør.

# 8. Prosjekt 4

## 8.1. Diffusjonsproblemet

Prosjekt 4 går over to veker. Me skal sjå på eit velkjend fysisk problem som me skal analysera både teoretisk og ved simulering. Der er fleire måtar å løysa oppgåva på, og for å kunna velja er det viktig at de veit kva som skal leverast inn til slutt. Les difor denne introduksjonen nøye.

**Døme 13** *Tenk deg ei glaskrukke med klart vatn. Du slepp ei dråpe med raud konditorfarge ned i krukka. Med ein gong er dråpa ein liten flekk med konsentrert raudfarge, men ganske raskt vil fargen spreia seg utover samstundes som han vert mindre intens. Dette fenomenet vert kalla* diffusjon *i fysikken.*

Diffusjon handlar om eitt stoff som spreier seg i eit anna stoff, slik som fargestoffet i vatnet i dømet over. Stoffet består av partiklar, og diffusjonen er eit resultat av stokastisk rørsle.

Den enklaste måten å skildra diffusjon er å ta utgangspunkt i ein agent-basert modell på ein raster, der éin partikkel er éin agent. Partikkelen flyttar seg tilfeldig. I ein enkel eindimensjonal modell kan t.d. sannsynsfordeling vera 10% sannsyn for eitt steg til høgre, 10% for eit steg til venstre og 80% for at partikkelen står i ro.

Ein kan simulera i ein, to eller tre dimensjonar. Krukka vår er sjølvsagt tredimensjonal. Ein eindimensjonal modell kan representera farge som diffunderer på ein våt bomullstråd, eller bakteriar som spreier seg i eit vassrøyr.

**Oppgåve 8.1 (Innlevering 4)** *De skal studera diffusjon i to dimensjonar (eller tre dimensjonar dersom de klarer å visualisera det). Innleveringa skal omfatta*

1. *Ein sannsynsfordeling for korleis éin partikkel rører seg på eitt tidssteg. De vel kva fordeling de vil bruka.*
2. *Ein visualisering av partikkelkonsentrasjonen (fargeintensiteten i dømet med konditorfarge) over tid basert på simulering.*
3. *Ein sannsynsfordeling for kvar partikkelen finst etter $t = 1, 2, 3, 4, 5$ tidssteg. Denne reknar de ut analytisk basert på sannsynsfordelinga for eitt tidssteg (1).*
4. *Ein refleksjon der du*
    a) *ein forklaring på korleis de har simulert og kvifor. Kor mange partiklar må simuleringa starta med for å få gode resultat?*
    b) *samanliknar dei analytiske resultata (3) med simuleringa (2).*
    c) *vurderer kva du har lært og kva du kan ha nytte av vidare.*

**Randvilkår.** Me hoppa over randvilkår då me introduserte *predator/prey*, men det er eit viktig spørsmål. Kva skjer når ein partikkel (agent) kjem til kanten (randen) av landskapet?

**Absorberande** Forsvinn agenten ut av landsskapet?

**Syklisk** Dukkar agenten opp på motsett side av landskapet, som om landskapet er ei kuleoverflate?

**Eksplisitt** Ser agenten veggen og bestemmer seg for å snu? I kva retning snur ein?

**Reflekterande** Dette kan minna om eksplisitt vegg, men logikken vert lagt i landskapet. Agenten ser ein rute der veggen er, men denne ruten er den same som ein annan rute.

Der kan vera fleire variasjonar over randvilkåra. Det viktigaste her er å gjera eit reflektert val som tek omsyn til problemet ein ynskjer å simulera.

Syklisk er mykje brukt, ikkje berre fordi verda er ein klode, men òg fordi ein kan tenkja seg at ein simulerer eit område i ei større verd der alle områda er einsarta. Då førestiller ein seg at agentane som forsvinn ut på den eine sida tilsvarer (omtrent) agentane som kjem inn frå naboområdet på motsett side.

På ein liknande måte kan absorberande randvilkår vera fornuftig dersom ein simulerer eit lite område i ei stor verd, der naboområda er svært forskjellige frå det området ein simulerer.

Der kan vera både konkrete og abstrakte grunnar for valet ein gjer.

## 8.2. Veke 11. Modellering

**NB.** Les introduksjonen (Diffusjonsproblemet i avsnitt 8.1) før du startar på øvinga.

**Oppgåve 8.2** *Vel ei sannsynsfordeling for korleis éin partikkel rører seg i 2D (eller 3D) på eitt tidssteg. Den same fordelinga gjeld for kvart tidssteg.*

*Det er mogleg å definera rørsle på $x$- og $y$-aksen som uavhengige prosessar, eller de kan definera éin modell direkte i 2D. Det vel de sjølve.*

**Oppgåve 8.3** *Diskuter korleis de vil simulera. Både ein agent-basert modell og ein celleautomaton kan vera velegna. Sjå avsnitt 6.1. Kva er føremonane ved kvart alternativ? Kva alternativ vil de bruka?*

**Oppgåve 8.4** *Uansett om de simulerer agent-basert eller ved celleautomaton, må de ha ein landsskapsmodell. Kva randvilkår vil de bruka?*

**Oppgåve 8.5** *Legg ein implementasjonsplan. Korleis implementerer de ein simulator for modellen som de har laga? Kva språk vil de bruka? Hugs at simulatoren må visualisera partikkelkonsentrasjonane, gjerne som fargeintensitet.*

**Oppgåve 8.6** *Start på implementasjonen.*

## 8.3. Veke 12. Simulering

**NB.** Les introduksjonen (Diffusjonsproblemet) og gjer forrige labøving, før du startar på denne.

**Oppgåve 8.7** *Gjer ferdig implementasjonen frå forrige veke.*

**Oppgåve 8.8** *Køyr simulatoren og visualiser partikkelkonsentrasjonen etter 5, 10, 100 og 1000 steg.*

**Oppgåve 8.9** *Test kor godt simulatoren skalerer. Kor stort rutenett kan de bruka? Kor mange partiklar kan de simulera?*

**Oppgåve 8.10** *Prøv gjerne alternative sannsynsmodellar. Korleis påverkar det diffusjonsprosessen om de aukar sannsynet for at partikkelen står i ro?*

**Oppgåve 8.11** *Rekn teoretisk på sannsynsfordelinga. Visualiser sannsynsfordelinga for kvar ein partikkel er etter 1, 2, 3, 4 og 5 tidssteg. Samanlikna dette med simuleringsresultata.*

**Oppgåve 8.12 (Ekstra)** *Kan du programmera den teoretiske utrekninga i forrige oppgåve slik at du kan finna sannsynsfordelinga etter 100 og etter 1000 steg?*

## 8.4. Teori

Nedanståande oppgåvesett frå i fjor inneheld litt meir omfattande teori for diffusjonsproblemet. Dette er pensum, men det er best forstått på bakgrunn av øvingane som de har gjort.

1. Introduksjon
2. Diffusjonskoeffisienten
3. Analyse

# 9. Prosjekt 5

Det siste prosjektet vil omfatta tre øvingar:

1. statistisk testing (evaluering) av slumptalsgeneratorar
2. analyse av eksterne datasett
3. bootstrap

## 9.1. Veke 14 Bootstrap

**Les 14** *Frå Frisvold og Moe: Kapittel 14.*

Sett at me studerer fordelinga åt vekta til ein bestemt fiskeart. Vekta til ein tilfeldig fisk er ein stokastisk variabel $X$ med ei viss sannsynsfordeling. Sett at me har fiska $n = 20$ fisk, og målt fylgjande vekter:

$$\begin{aligned} &15{,}6; 12{,}6; 13{,}7; 13{,}8; 17{,}0; \\ &12{,}9; 11{,}5; 6{,}9; 7{,}8; 3{,}7; \\ &13{,}0; 14{,}4; 6{,}6; 11{,}7; 11{,}1; \\ &1{,}8; 14{,}9; 16{,}5; 12{,}3; 10{,}5 \end{aligned}$$

Dette er eit *utval*, med $n$ *observasjonar* $x_1, x_2, \ldots, x_n$ av $X$.

Me veit korleis me kan rekna ut utvalsgjennomsnittet $\bar{x} = 11{,}4$ og utvalsstandardavviket $s = 4{,}02$ for dette utvalet.

Utvalsgjennomsnittet $\bar{x}$ vert brukt for å estimera populasjonsgjennomsnittet $\mu$. Sidan $\bar{x}$ er rekna ut frå observasjonane $x_i$, er ogso $\bar{x}$ ein observasjon av ein stokastisk variabel som me noterer $\bar{X}$. Dvs. utvalsgjennomsnittet har ein sannsynsfordeling, og kvar gong me finn gjennomsnittet i eit nytt utval får me ein ny observasjon og som regel eit nytt tal. Me vil (nesten) aldri treffa populasjonsgjennomsnittet $\mu$ akkurat, men som regel vil me treffa nær.

Kor nær me treff avheng av standardavviket åt $\bar{X}$. Standardavviket åt ein estimator vert òg kalt standardfeilen. Me veit at standardfeilen her er gjeve som

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}, \tag{25}$$

der $\sigma$ er standardavviket åt $X$. Me kan estimera standardfeilen som

$$\hat{\sigma}_{\bar{X}} = \frac{\hat{\sigma}}{\sqrt{n}} = \frac{s}{\sqrt{n}}. \tag{26}$$

Dette gjev eit mål for kor presis $\bar{X}$ er som estimator for $\mu$.

Sett no at me ynskjer å studera standardavviket $\sigma$ like djupt som me kan studera $\mu$. Me har utvalsstandardavviket $S$ som estimator for $\sigma$, men korleis kan me estimera standardfeilen $\sigma_S$ åt $S$?

Med mindre me kjenner den underliggjande sannsynsfordelinga åt $X$ finst der inga analytisk løysing på dette. I mange tilfelle må me rett og slett observera $S$ mange gongar, slik at me har eit utval å rekna med. Me kan gjenta forsøket $m$ gongar, og kvar gong observera eit utval på $n = 20$ fisk. For kvart utval kan me rekna ut eit utvalsstandardavvik, slik at me til slutt har $m$ observasjonar $s_1, \ldots, s_m$ av $S$. Då kan me rekna ut gjennomsnittet $\bar{s}$ og utvalsstandardavviket $s_S$.

Problemet med dette er at det er kostbart å samla data. Me treng $m$ gongar so mykje data for å estimera standardfeilen for det opprinnelege forsøket. Bootstrap er ei vanleg løysing som går ut på å simulera $m$ gjentakne utval basert på det eine opprinnelege utvalet.

For å laga eit *bootstrap*-utval, trekk me 20 tilfeldige fisk frå det fyrste utvalet *med tilbakelegging*; dvs. same måling kan verta utvald fleire gongar. Dersom det opprinnelege utvalet er representativt for populasjonen, so har *bootstrap*-utvalet òg ei rimeleg sannsynsfordeling.

$$\begin{aligned} &13{,}0;\ 12{,}6;\ 10{,}5\ 13{,}7;\ 12{,}9;\\ &14{,}4;\ 10{,}5\ 3{,}7;\ 10{,}5\ 12{,}6;\\ &13{,}7;\ 13{,}7;\ 11{,}7;\ 14{,}9;\ 12{,}6;\\ &11{,}7;\ 14{,}4;\ 13{,}0;\ 13{,}7;\ 14{,}9; \end{aligned}$$

I dette *bootstrap*-utvalet finn me $\bar{x} = 12{,}4$ og $s = 2{,}4$. Gjentek med dette eksperimentet $m$ gongar, kan me få eit utval med $m$ observasjonar $s$ av $S$, og rekna ut utvalsstandardavviket $s_s$ for utvalet av observasjonar av $S$, og bruka det som estimat for standardfeilen ved estimering av $\sigma$.

*Bootstrap* er mykje rekning og vert sjelden gjort for hand. Det er ei typisk simuleringsøving, og enkelt å gjera på maskin.

### 9.1.1. Forundersøking

**Oppgåve 9.1** *Last ned bootstrapgen.m, som du skal bruka til å laga syntetiske datasett.*

**Oppgåve 9.2** *Test funksjonen*

```
X = bootstrapgen(200)
```

*Dette dannar eit utval* `X` *med* $n = 200$ *observasjonar.*

**Oppgåve 9.3** *Lag eit histogram over datasettet* `X`*. Bruk minst 20 søyler for å få eit godt inntrykk av fordelinga.*

**Oppgåve 9.4** *Prøv å tippa på gjennomsnittet* $\bar{x}$ *og utvalsstandardavviket* $s$ *på augamål frå histogrammet. Kva verdiar vil du venta å finna når du startar å rekna?*

**Oppgåve 9.5** *Bruk Matlab til å rekna ut gjennomsnittet* $\bar{x}$ *og utvalsstandardavviket* $s$ *for* `X`*.*

### 9.1.2. Bootstrap

No skal me analysera standardavviket $\sigma$ i dømet over vha. *bootstrap*.

**Oppgåve 9.6** *Estimer standardfeilen for gjennomsnittet* $\bar{X}$ *i datasettet* `X` *over.*

**Oppgåve 9.7** *Lag ein matlabfunksjon som tek eit utval* `X` *som argument, og returnerer eit* bootstrap-*utval med same storleik. Test funksjonen på datasettet* `X` *som du har brukt over. Finn* $s$ *og* $\bar{x}$ *for* bootstrap-*utvalet. Ser tala fornuftige ut?*

**Oppgåve 9.8** *Skriv ein funksjon som genererer* $m$ bootstrap-*utval frå det same datasettet* `X` *og reknar ut utvalsstandardavviket* $s$ *kvar gong. Returverdien skal vera ein matrise (vektor) med* $m$ *observasjonar av* $s$*.*

**Oppgåve 9.9** *Test funksjonen frå forrige oppgåve på datasettet* `X`*, og lag eit datasett* `S` *med* $m$ *observasjonar av* $s$*. Vel* $m$ *sjølv. Plott* `S` *i eit histogram.*

**Oppgåve 9.10** *Rekn ut gjennomsnitt og utvalsstandardavviket for datasettet* `S`*.*

### 9.1.3. Kontroll

Får å validera *bootstrap* som metode, skal me no gjenta oppgåvene 9.8–9.10 med ein liten variasjon. I staden for å generera $m$ *bootstrap*-utval skal me generera «ekte» utval ved hjelp av `bootstrapgen.m`.

**Oppgåve 9.11** *Skriv ein funksjon som genererer* $m$ *utval vha.* `bootstrapgen.m`*, kvart med* $n = 200$ *observasjonar. Rekn ut utvalsstandardavviket* $s$ *for kvart utval og returner ein matrise med* $m$ *observasjonar av* $s$*.*

**Oppgåve 9.12** *Test funksjonen frå forrige oppgåve og lag eit datasett* `S2` *med* $m$ *observasjonar av* $s$*. Bruk same* $m$ *som i oppgåve 9.9. Plott* `S2` *i eit histogram.*

**Oppgåve 9.13** *Rekn ut gjennomsnitt og utvalsstandardavviket for datasettet* `S2`*.*

### 9.1.4. Eit døme til (Ekstra)

**Oppgåve 9.14** *Sjå på datasettet som me brukte som døme i starten:*

$$\begin{aligned} &15{,}6;\ 12{,}6;\ 13{,}7;\ 13{,}8;\ 17{,}0;\\ &12{,}9;\ 11{,}5;\ 6{,}9;\ 7{,}8;\ 3{,}7;\\ &13{,}0;\ 14{,}4;\ 6{,}6;\ 11{,}7;\ 11{,}1;\\ &1{,}8;\ 14{,}9;\ 16{,}5;\ 12{,}3;\ 10{,}5 \end{aligned}$$

*Estimer standardavviket* $\sigma$ *og standardfeilen for estimatoren vha. bootstrap.*

### 9.1.5. Rekneøving (Ekstra)

**Oppgåve 9.15** *Ta utgangspunkt i fylgjande datasett:*

$$10, 11, 11, 13, 15.$$

*Svar på fylgjande*

1. *Estimer standardavviket for populasjonen. (Punktestimat er tilrekkeleg.)*
2. *Vis korleis du bruker* bootstrap *for å estimera standardfeilen for estimatoren du brukte over.*

**Oppgåve 9.16** *Eksamen våren 2015, oppgåve 1.*

**Oppgåve 9.17** *Eksamen våren 2015, oppgåve 7 og 9.*

**Oppgåve 9.18** *Eksamen våren 2015, oppgåve 5 og 6.*

## 9.2. Veke 15. Test av slumptalsgeneratorar

I denne øvinga skal me sjå på eit klassisk problem i informatikk: statistisk testing av slumptalsgeneratorar. Me har allereie studert fordelinga frå ulike slumptalsgeneratorar og sett at nokon er svært dårlege og andre mindre dårlege, men det heile har vore skjønn og synsing.

I denne øvinga skal me bruka statistikk til å vurdera kvaliteten på slumptalsgeneratorar. Dette føreset at de har lært hypotesetesting gjennom rekneøvingane. Eg tek gjerne ein prat om korleis me får dei statistiske modellane til å passa saman med det praktiske problemet med slumptalsgeneratorar. Ikkje nøl med å spørja om du sit fast. Eg har ikkje gjeve nokon detaljert oppskrift, simpelthen fordi de treng å vera med på prosessen, men de treng ikkje ta han aleine.

All bruk av slumptalsgeneratorer byggjer på ein hypotese:

$$H_0 : \text{slumptala er uavhengige og uniformt fordelte}$$

Dette er ein hypotese som me kan testa.

### 9.2.1. Rafinering av nullhypotesen (Fyrste omgang)

Me skal bruka $\chi^2$-testen. Før me ser på korleis me utfører han, skal me sjå på kva utval me treng, og finna hypotesar som me kan bruka.

Nullhypotesen over er føreset to ting; at slumptala er (1) uavhengige og (2) uniformt fordelte. I fyrste omgang skal me fokusera berre på fordelinga, og sjå på hypotesen

$$H_0^{(1)} : \text{slumptala er uniformt fordelte}$$

Dersom $H_0$ er sann, so er $H_0^{(1)}$ sann. Dersom me forkastar $H_0^{(1)}$ må me forkasta $H_0$. Me skal koma tilbake til uavhenge seinare.

Føresetnaden for $\chi^2$-testen er at me kan trekkja eit utval som er mange gongar større enn utfallsrommet.

Dvs. at me kan ikkje bruka heile utfall frå slumptalsgeneratoren direkte. Då er utfallsrommet altfor stort. Me kan derimot slå saman utfall for å skapa eit mindre utfallsrom.

**Døme 14** *Lat $X$ vera eit slumptal frå generatoren, og lat $Y = X \mod 16$. Hypotesen vår er*

$$H_0^{(1')} : X \text{ er uniformt fordelt}$$

*Dersom denne hypotesen er sann, er det òg sant at*

$$H_0^{(2)} : Y \text{ er uniformt fordelt}$$

*Utfallsrommet åt $Y$ er $\{0, 1, \ldots, 15\}$, altso 16 element. Det kan me bruka.*

I eksempelet tek me ein nullhypotese som testar dei fire minst signifikante bitsa i slumptalet. Me kan trekkja ut, og testa på, ein kvan bitmaske. Dersom slumptala er uniformt fordelte, vil alle bitmaskar òg vera uniformt fordelte.

(OK. Der er eit aber. Dersom storleiken på utfallsrommet for $X$ ikkje er deleleg med 16, fordeler ikkje tala seg jamnt. Der vil vera eitt tal meir som vert 1 modulo 16 enn dei som vert 0. Dette kan me sjå bort frå dersom utfallsrommet er stort og me maskar ut få bits.)

### 9.2.2. Frå intuisjon til statistisk hypotesetest

Me skal starta med å testa $H_0^{(2)}$ frå døme . Dersom du trekk eit par hundre slumptal $X$ og reknar ut $Y$, kan du plotta observasjonar av $Y$ i eit histogram. Visuelt kan du so vurdera om $H_0^{(2)}$ verkar rimeleg.

Dersom histogrammet ikkje ser ut som ei uniform fordeling, er det rimeleg å tru at det ikkje er generert av ei uniform fordeling, og hypotesen er usann. Me går då ut frå at slumptalsgeneratoren er dårleg.

Dersom histogrammet ser ut som ei uniform fordeling, er det rimeleg å gå ut frå at hypotesen er sann, og me generatoren har i alle fall ikkje denne dårlege eigenskapen.

Den visuelle vurderinga er derimot reint skjønn. Synsing vil nokon seia. For å gjera ein objektiv vurdering ynskjer me enkle kvantitative svar.

### 9.2.3. $\chi^2$-testen

Lat oss ta utgangspunkt i histogrammet igjen. Lat $[y_1, y_2, \ldots, y_n]$ vera utvalet vårt, dvs. ei fylgje av slumptal modulo 16.

1. Lat $F_y$ vera frekvensen av verdien $y$ i utvalet.

   Dvs. $F_y$, for $0 \leq y \leq 15$ er talet på gongar $y$ førekjem i utvalet. Histogrammet plottar $F_y$ for kvar $y$.

2. Lat $E_y$ vera forventingsverdien til $F_y$, dersom hypotesen vår er sann.

   Hypotesa seier uniform fordelinga, og då er $E_y = n/16$ der $n$ er storleiken på utvalet.

3. Me reknar ut den stokastiske variabelen

$$G = \sum_{y=0}^{15} \frac{(F_y - E_y)^2}{E_y}.$$

Variabelen $G$ er eit standardverkty for å samanlikna ein empirisk fordeing ($F_u$) med ein hypotetisk fordeling (uniform i dette tilfellet). Det er lettare å sjå avvik i ein enkelt skalarvariabel $G$, enn å sjå på heile histogrammet med seksten forskjellige frekvensar.

**Oppgåve 9.19** *Sjå på uttrykket for $G$. Kva verdiar kan $G$ ta? Korleis ser histogrammet ut når $G$ tek minste mogleg verdi?*

**Oppgåve 9.20** *Kva verdiar ventar du at $G$ har når hypotesen om uniform fordeling held? Kva når ho ikkje held?*

**Oppgåve 9.21** *Bruk `rng1.m` frå avsnitt 5.3 og lag eit utval på $n = 1000$ tilfeldige tal modulo 16. Finn frekvensane $F_y$ vha. funksjonen*

```
f = histcounts(y,'BinMethod','integers')
```

*Rekna ut $G$ som forklart over. Kva verdi får du?*

*Gjer det same for `rng2.m`.*

**Oppgåve 9.22** *Variabel $G$ er stokastisk med $\chi^2$-fordeling med 15 fridomsgradar. Plott sannsynsfordelinga*

```
fplot( @(x)chi2pdf(x,15), [0 40] )
```

*Det merkelege uttrykket `@(x)chi2pdf(x,15)` er eit lambdauttrykk og lagar ein ny funksjon med ein parameter $x$ vha. den eksisterande funksjonen som har 2.*

*Samanlikna observasjonar av $G$ frå forrige oppgåve med sannsynsfordelinga. Synest du observasjonane dine ser sannsynlege ut dersom slumptala er uniformt fordelte?*

**Oppgåve 9.23** *Rekn ut $p$-verdien frå testane av `rng1.m` og `rng2.m`. Kva kan du seia om kvaliteten på dei to slumptalsgeneratorane? (Der er $\chi^2$-tabell i læreboka.)*

### 9.2.4. Nullhypotesar for uavhengig fordeling

Lat oss gå tilbake til urhypotesen vår

$$H_0 : \text{slumptala er uavhengige og uniformt fordelte}$$

Når me krev uavhengig fordeling, rekk det ikkje å sjå på éin verdi $X$. Me må sjå på ein serie med verdiar $X_1, \ldots, X_n$.

Dersom $H_0$ er sann, so er det òg sant at

$$H_0^{(3)} : X_i \text{ er uavhengig og uniformt fordelt}$$

Dersom $X_1$ og $X_2$ er uavhengig og uniformt fordelt, so vil det seia at parret $(X_1, X_2)$ er uniformt fordelt. Likeeins vil tuplar $(X_1, \ldots, X_n)$ vera uniformt fordelte. Me kan difor danna hypotesen

$$H_0^{(4)} : (X_1, X_2) \text{ er uniformt fordelt}$$

Dersom $H_0^{(3)}$ er sann, so må $H_0^{(4)}$ vera sann.

Utfallsrommet er no sjølvsagt altfor stort, men me kan kombinera med modulustrikset som me brukte i stad. No er $X_1$ og $X_2$ to stokastiske variablar. Me kan definera

$$Y' = 4 \cdot (X_1 \mod 4) + (X_2 \mod 4)$$

Effektivt tek me då to bits frå $X_1$ og to bits frå $X_2$ og set dei saman til eit firebits tal. Utfallsrommet er altso 16 element som før.

Dersom $H_0^{(4)}$ er sann, er det altso sant at

$$H_0^{(5)} : Y' \text{ er uniformt fordelt}$$

Denne hypotesen kan testast på same måte som $H_0^{(2)}$. Det er berre $Y$ som er rekna ut på ein litt annan måte.

### 9.2.5. Testing for uavhengig fordeling

**Oppgåve 9.24** *Gjenta oppgåvene i avsnitt 9.2.3 med utgangspunkt i $Y'$ og $H_0^{(5)}$ i staden for $Y$ og $H_0^{(2)}$.*

### 9.2.6. Meir lesing

Det er verd å lesa Donald Knuths klassikar, *The Art of Computer Programming*. I band 2, *Seminumerical Algorithms*, har han eit kapittel om slumptalsgeneratorar, med ei rekkje døme på statistiske testar.

## 9.3. Veke 17. Dataanalyse

### 9.3.1. Eit eksterne datasett

Machine Learning Repository ved UCI er ei god kjelde for datasett til testing og øving i dataanalyse.

**Oppgåve 9.25** *Last ned datasettet for sykkedeling. Du finn sjølve datasettet som PKzip-fil under «Data Folder».*

Dette datasettet illustrerer samanhengen mellom sykkelutleige og vêrdata. Me skal bruka lineær regresjon for å studera denne samanhengen. Datasettet har tre variablar som me kan prøva å predikera ($y$-verdiar), totalt utleigedal, utleige til medlemmer og utleige til andre (*casual*). Der er òg fleire ulike forklarignsvariablar ($x$-verdiar), som er forklart i datasettbeskrivinga (i Readme-filen eller på vevsida), inklusive temperatur og vind. Der er to datasett; me bruker det for daglege data, fordi det er enklast.

**Oppgåve 9.26** *Opna* `day.csv` *(t.d. i eit rekneark) og les gjennom datasettbeskrivinga. Kva inneheld dei ulike søylene?* Spør dersom du er usikker på noko.

### 9.3.2. Data i Matlab

I denne øvinga skal me alltid sjå på to søyler åt gongen, ein $x$-verdi og ein $y$-verdi. Du kan velja kva for ei vêrdatasøyle og kva for ein utleigesøyle du vil bruka, og det er lurt å prøva med litt ulike kombinasjonar.

**Oppgåve 9.27** *Last fila i Matlab:*

```
tbl = readtable('day.csv')
```

*For å henta ut to søyler kan du bruka t.d.*

```
x = tbl.temp
y = tbl.cnt
```

**Merknad 8** *Der er fleire måtar å lasta datasett på i Matlab. Den som er føreslått over er den same som vart presentert i avsnitt 5.4.3.*

**Oppgåve 9.28** *Det hender at der er rot og manglar i datasett som vert importert. Det bør det ikkje vera her, men det er verd å sjekka:*

```
size(x)
size(y)
```

*Dette gjev storleiken på vektorane. Er dei like store?*

*For å sjekka for manglande verdiar, kan du prøva*

```
any(ismissing(y))
```

*Dersom minst éin verdi manglar i* `y`*, vil* `any()` *returnera sann. Manglar der noko? Test* `x`*-vektoren òg.*

**Oppgåve 9.29** *Plott $x$- og $y$-verdiane mot kvarandre (*`plot`*-funksjonen i Matlab).*

```
plot(x,y,'.')
```

*Kva ser du? Er der samanheng mellom variablane. Dersom du ikkje ser nokon samanheng, prøv ein annan kombinasjon av søyler.*

**Merknad 9** *Dersom du vil sjå både på tilfeldige og registrerte brukarar, kan du bruka to fargar i plottet:*

```
plot(x,tbl.casual,'r.',x,tbl.registered,'k.')
```

### 9.3.3. Regresjon

No har me sett på samanhengen mellom $x$ og $y$ i datasettet visuelt. Matematisk kan me modellera samanhengen som en linær funksjon

$$y = b + ax + \epsilon(x)$$

der $\epsilon$ er eit lite støyledd som me kan sjå bortifrå. Regresjon handlar om å estimera $a$ og $b$ slik at $\epsilon$ vert minst mogleg.

Metoden er forklart i kapittel 12.2 i læreboka.

**Oppgåve 9.30** *Implementer matlab-funksjonar som reknar ut $b$ og $a$ i tråd med likning (12.9) i læreboka.*

**Merknad 10** *Der finst kanskje matlab-funksjonar som gjer heile jobben, men eg fann ikkje dokumentasjon som forklarer nøyaktig kva som skjer med same terminologi som læreboka. Dersom du finn ein funksjon som du trur gjer jobben, må du dermed samanlikna resultata med utrekning etter boka.*

**Oppgåve 9.31** *Plott regresjonslina saman med rådata. Dette kan du t.d. gjera slik:*

```
plot(x,y,'r.')
y2 = a*x + b
plot(x,y2,'k')
```

**Oppgåve 9.32** *Tolk resultata. Kva kan me slutta om korleis potentielle syklistar oppfører seg med varierande vêr.*

## 9.4. Innlevering

Innleveringa skal innehalda

1. For kvar av dei tre øvingane, ei oppsummering av dei statistiske resultata med tolking og konklusjon.

   a) Frå bootstrap. Kva kan me seia om standardavviket i datasettet?

   b) Om slumptalsgeneratorane. I kva grad kan me lita på dei generatorane som er testa?

   c) Frå datanalysen. Kva kan me seia om samanhengen mellom dei to variablane.

   Bruk konkrete data frå køyringane dine for å underbyggja konklusjonen din, men hugs at målet ikkje er å visa at dataprogrammet ditt produserer data. Poenget er å visa at du kan forhalda deg til verkelege data og tolka dei.

2. Eit refleksjonsnotat for kurset som heilskap, der du mellom anna svarer på: *Kva er den mest interessante øvinga i kurset, og kvifor?*

# 10. Siste veke

Siste undervisingsdag er 3. mai. Som vanleg har me klasserommet 8-10 og labben 10-14.

Rekneøvinga bruker me til å løysa eksamensoppgåver frå det siste settet, hausten 2018. Dersom nokon treng repetisjon av spesielle tema, ver snill å senda meg eit ord, om tema og/eller spesifikke oppgåver. Dersom der ikkje kjem inn spesifikke ynskjer, tek me for oss eksamenssettet frå sist haust. Det er nettopp lagt ut.

Me held fram med eksamensoppgåver på labben, men me må der prioritera godkjenning av og spørsmål til laboppgåvene. Det er viktig at dei som manglar godkjenningar får gjort det 3. mai.

# A. Gamle eksamensoppgåver

1. Hausten 2018 (løysingsforslag manglar)
2. Våren 2018 (Bokmål / Løysingar)
3. Våren 2017 (Bokmål / Løysingar)
4. Hausten 2016
5. Våren 2016

6. Hausten 2015
7. Våren 2015 (Bokmål / Løysingar)
8. Våren 2014 (Bokmål / Løysingar)
9. Eksamensdøme nr. 1 (2014). Ufullstendige løysingar
10. Eksamensdøme nr. 2 (2014). Ufullstendige løysingar

# B. Kjelder