STK1110: Statistiske metoder og dataanalyse

Innledning til STK1110 Høsten 2026

Forelesninger ved Sven Ove Samuelsen
Matematisk institutt
Universitetet i Oslo

Litt om STK1110

- STK1100 omhandlet først og fremst **sannsynlighetsregning**.
- STK1110 tar for seg en rekke viktige statistiske modeller og metoder.
- Innføring av begreper og metoder for enkle modeller
  - men i de fleste tilfeller overførbart til mer komplekse settinger
- STK1110 fortsetter der STK1100 slapp.
- Alt stoff fra STK1100 forutsettes kjent i STK1110 (men vi vil selvfølgelig repetere litt når det er behov for det).

Litt om statistikk

- Statistikk handler om å tallfeste usikkerhet:
  - Hva er sannsynligheten for regn i morgen?
  - Hvor stor sum bør forsikringsselskapet sette til side for at det med 99.95% sannsynlighet vil kunne dekke alle sine forpliktelser?
  - Virker denne nye medisinen bedre enn den gamle?
  - Gjør salting av veiene luftkvaliteten bedre om vinteren?
- Statistikk blir brukt mye i andre fagdisipliner
  - men er også en egen fagdisiplin!
- Statistikerens jobb er å
  - lære nok om anvendelsen til å forstå problemet
  - oversette problemet til et statistisk problem
  - løse det statistiske problemet
  - oversette resultatene til noe oppdragsgiver forstår.

Hvorfor trenger vi statistikere?

- Statistiske metoder brukes ofte av ikke-statistikere.
- Statistikere lærer ikke bare hvordan en bruker statistiske metoder, men også hvordan de fungerer og hvilke antakelser de forutsetter.
- Statistikere er spesielt nyttige i mer kompliserte situasjoner, der standardantakelsene ikke gjelder og standardmetoder derfor ikke kan brukes.
- Som statistiker skal du være i stand til å endre en metode slik at den også kan brukes i en mer komplisert situasjon, eller til og med finne på en ny metode.

Hva slags type jobb kan du få som statistiker?

- Etter som det stadig samles inn større mengder data, øker også behovet for folk som kan analysere dem.
  - Maskinlæring, Kunstig intelligens
- Statistikere er veldig nyttige og ettertraktet (over 300 utlyste jobber på Finn, bare i Oslo)!
- Tidligere studenter jobber f.eks. i SSB, Folkehelseinstituttet, Skatteetaten, Gjensidige, DnB, Norsk Regnesentral, NRK, Ruter,...



25-åringen jobber i landets best betalte yrke. Her er hele listen

(Fra VG,16. februar 2022)

Foreløpig pensum

- Lærebok:
  From *Jay L. Devore, Kenneth N. Berk & Matthew A. Carlton: Modern Mathematical Statistics with Applications, third edition, 2021. Springer. ISBN: 978-3-030-55155-1.*
  Boka er tilgjengelig som e-bok hvis du arbeider på en datamaskin på universitetet.
  Alt stoff fra STK1100 regnes som kjent.
  1. Kapittel 7: Avsnitt 7.4 (unntatt beviset på sidene 441-442).
  2. Kapittel 6: Avsnitt 6.3 og 6.4.
  3. Kapittel 8: Avsnitt 8.2-8.5.
  4. Kapittel 9: Avsnitt 9.1-9.4, 9.5 (bare sidene 547-551) og 9.6 (bare sidene 553-555).
  5. Kapittel 10: Avsnitt 10.1-10.5.
  6. Kapittel 11: Avsnitt 11.1 og 11.3 (bare sidene 662-663).
  7. Kapittel 12: Avsnitt 12.3-12.10.
  8. Kapittel 14: Avsnitt 14.1-14.2.
  9. Kapittel 15: Avsnitt 15.1-15.2.

Kapittel 7: Punktestimering

- Metoder for punktestimering
- Informasjon og effisiens.

Kapittel 6: Fordelinger knyttet til normalfordelte utvalg

- Støtte til kapittel 8, 9, 10, 11 og 12.
- χ^2-, t- og F-fordelingen.

Kapittel 8: Konfidensintervall

- Konfidensintervall for forventningsverdi, populasjonsandel, standardavvik og varians.
- Metoder for store og små utvalg.
- Konfidensintervall basert på bootstrapping.

Kapittel 9: Hypotesetesting

- Tester for forventningsverdi og populasjonandel
- P-verdier
- Valg av testprosedyre.

Kapittel 10: Statistiske metoder for to utvalg

- Statistisk inferens (konfidensintervall og testing) om to forventningsverdier, populasjonsandeler og to varianser
- Metoder for store og små utvalg
- Parforsøk.

Kapittel 11: Variansanalyse

- Bare 11.1 og 11.3: Enveis variansanalyse.
- Sammenligning av 3 og flere utvalg

Kapittel 12: Lineær regresjon

- Mer om enkel lineær regresjon
- Multippel lineær regresjon
- Korrelasjon
- Estimering og inferens
- Logistisk regresjon for binære responser

Kapittel 14: Ikke-parametriske metoder

- Ikke-parametriske hypotesetester
- for sammenligning av to behandlinger

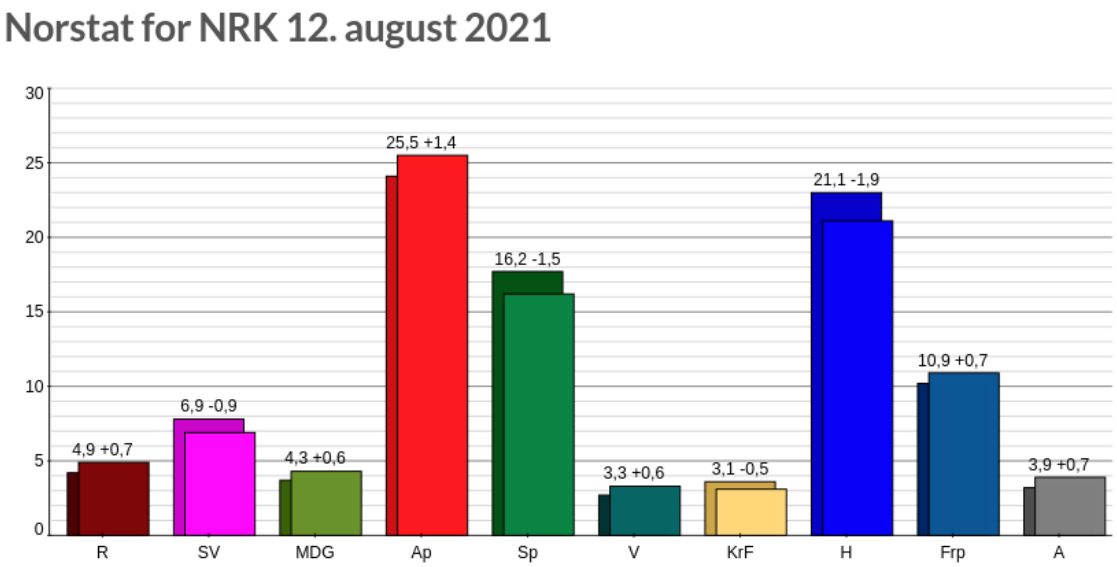
Kapittel 15: Introduksjon til bayesiansk estimering

- Apriori- og aposteriorifordelinger
- Bayesiansk estimering

Sannsynlighetsmodeller og statistiske metoder

- Vi ønsker å få kunnskap om en populasjon basert på observerte data fra denne populasjonen, kalt utvalg.
- Vi må da ha en modell som angir hvordan dataene vi har observert framkommer fra populasjonen.
- Vi antar at dataene våre er observerte verdier av stokastiske variabler, og at vi kjenner fordelingen til de stokastiske variablene (med unntak av en eller flere parametere).
- Videre antar vi at problemstillingene vi er interessert i "oversettes" til utsagn om parameterne i modellen.

Eksempel 1



- Av de 691 som ville ha stemt om det var stortingsvalg i morgen, ville 41 ha stemt på Sp.
- Sps oppslutning er da $41/691 \approx 0.059$.
- Hvor sikkert er dette estimatet?

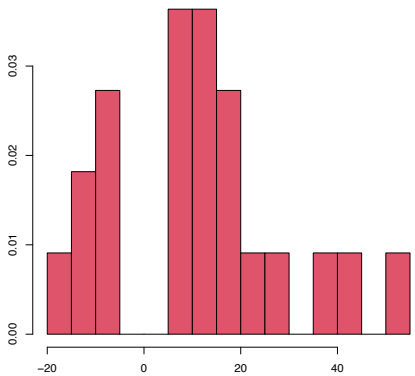
Eksempel 1

- Vi har data $x_1, \ldots, x_n$, der $x_i = 1$ hvis person nummer i ville ha stemt Sp og $x_i = 0$ ellers.
- Vi vil anta at $x_1, \ldots, x_n$ er observerte verdier av $X_1, \ldots, X_n \overset{uif}{\sim} Bernoulli(p)$, dvs. $P(X_i = 1) = p$ og $P(X_i = 0) = 1 - p$.
- Her er p den faktiske andelen av populasjonen som ville ha stemt på Sp.
- Vi har at $Y = \sum_{i=1}^{n} X_i \sim Binomisk(n, p)$.
- En naturlig estimator for p er da $\hat{p} = Y/n$ (Kap. 7).
- Det gir $\hat{p} = \frac{41}{691} \approx 0.059$.
- Basert på antakelsene over kan vi lage et konfidensintervall for p, som angir hvor sikkert estimatet er (Kap. 8).

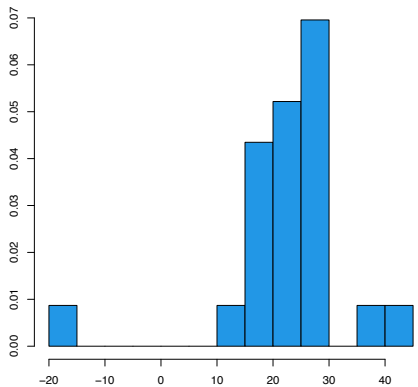
Eksempel 2

I et forsøk ble 22 rotter utsatt for ozon (behandlede), mens 23 andre ikke ble det (kontroll). Så registrerte en rottenes vektøkning i gram i løpet av en uke.

Behandlede rotter



Kontrollgruppe



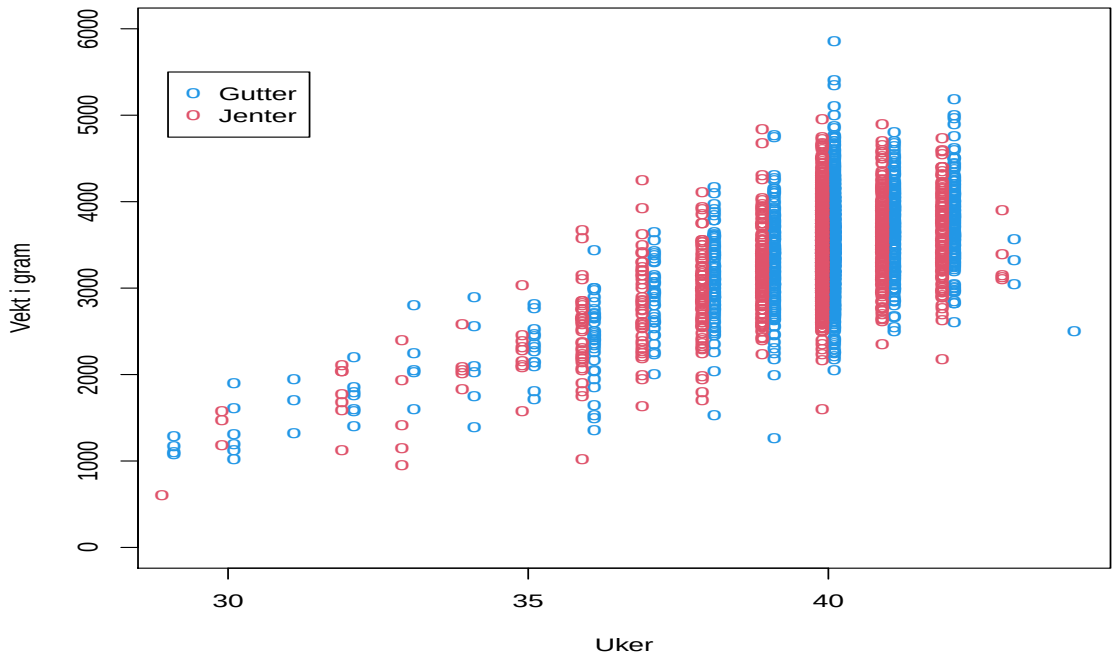
- Kan vi med rimelig grad av sikkerhet si at det er en forskjell i vektøkning mellom behandlede og ubehandlede rotter?
- Kan vi gi et anslag på forskjellen i vektøkning?
- Hvor sikkert er dette anslaget?

Eksempel 2, forts.

- Vi har data $x_1, \ldots, x_n$ og $y_1, \ldots, y_n$, der x_i er vektøkningen til den i-te rotta i behandlingsgruppa og y_j er vektøkningen til den j-te rotta i kontrollgruppa.
- Vi antar at $x_1, \ldots, x_n$ og $y_1, \ldots, y_n$ er observerte verdier av henholdsvis $X_1, \ldots, X_n \overset{uif}{\sim} N(\mu_1, \sigma_1)$ og $Y_1, \ldots, Y_n \overset{uif}{\sim} N(\mu_2, \sigma_2)$.
- Hovedformålet med forsøket er å avgjøre om μ_1 og μ_2 er forskjellige. Det kan vi gjøre ved hjelp av hypotesetesting (Kap 9 og 10).
- Vi vil også være interessert i å estimere differansen $\mu_1 - \mu_2$ og å si noe om hvor sikkert estimatet for differansen er (Kap 10).
- Da trenger vi å estimere variansene σ_1^2 og σ_2^2 (Kap 7 og 10).

Eksempel 3

Figuren viser fødelsvekten til et utvalg av 4066 barn, delt inn etter kjønn, mot svangerskapsvarighet i uker:



- Kan vi anslå forventet fødelsvekt for ei jente som blir født i uke 38 av svangerskapet?
- Hvor sikkert er anslaget?

Eksempel 3, forts.

- Vi har data $y_1, \ldots, y_n$, $u_1, \ldots, u_n$ og $s_1, \ldots, s_n$, der y_i, u_i og s_i er henholdvis fødselsvekten, svangerskapslengden og kjønnet til den i-te barnet ($s_i = 0$ hvis det er en gutt og $s_i = 1$ hvis det er ei jente).
- Vi antar at $y_1, \ldots, y_n$ er observerte verdier av $Y_1, \ldots, Y_n$, med

$$Y_i = \beta_0 + \beta_1 u_i + \beta_2 s_i + \epsilon_i, \ \epsilon_i \overset{uif}{\sim} N(0, \sigma^2).$$

- Modellen over er en lineær regresjonsmodell (Kap 12).
- Forventet fødelsvekt for ei jente som blir født i uke 38 av svangerskapet er da

$$E(Y|u = 38, s = 1) = \beta_0 + \beta_1 \cdot 38 + \beta_2.$$

Eksempel 3, forts.

- Et anslag av denne forventningen er $\hat{\beta}_0 + \hat{\beta}_1 \cdot 38 + \hat{\beta}_2$, der $\hat{\beta}_0$, $\hat{\beta}_1$ og $\hat{\beta}_2$ er estimater av β_0, β_1 og β_2, som vi må finne (Kap. 12).
- Usikkerheten i den anslåtte forventningen kan beskrives ved hjelp av et konfidensintervall (Kap. 12).

Eksempel 4

Vi har en gruppe pasienter med en spesifikk sykdom har og en kontroll gruppe uten sykdommen. For begge grupper kjenner vi diverse bakgrunnsvariable som kjønn, alder, utdannelse, røykevaner, fysisk aktivitet.

- Hvordan kan vi finne ut hvilke variable som samsvarer med sykdom?

Her er responsvariablen om det er sykdom eller ikke, altså binær. Slike data kan analyseres ved logistisk regresjon som introduseres i Kapittel 12.

Regresjon for binære responer samt for responser fra andre typer fordelinger som Poisson-fordelinger, Negativt binomisk fordelinger (begge telledata) samt gammafordelte variable behandles i STK3100 - Innføring i generaliserte lineære modeller

Eksempel 5

En gruppe pasienter med en spesifikk sykdom har blitt gentypet. Tilsvarende er gjort for en kontroll gruppe uten sykdommen. Antall individer er igjen i størrelsesorden ≈ 100. Antall gener er ≈ 21000.

- Er det mulig å si noe om hvilke gener som trigger en sykdom?

Merk at her vil vi ha flere forklaringsvariable (alle gener, 21000) enn pasienter (ca. 100).

Både statistisk og numerisk utfordrende. Tema for STK2100 - Maskinlæring og statistiske metoder for prediksjon og klassifikasjon

Eksempel 6

En ønsker å si noe om levetiden etter en ny type kreftbehandling.
Noen pasienter er allerede døde, men mange lever fremdeles

- Hvordan si noe om levetiden når data er ufullstendige?

Tema for STK4080 - Forløpsanalyse

Eksempel 7

En ønsker å studere hvordan aksjekurser varierer over tid og hvordan de kan tenkes å utvikle seg over tid.

- Hvordan studere slike data med avhengighet mellom påfølgende dager?

Tema for STK4060 - Tidsrekkeanalyse