STK1110 Høsten 2021

Innledning til STK1110

Ingrid Hobæk Haff
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**.
- 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?
- Det er et hjelpefag for andre fagdisipliner.
- 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.

Litt om statistikk

Problemstilling 1:

- En gruppe forskere var interessert i hvordan mennesker oppfatter at en lyd virker menneskelig.
- De gjorde et forsøk der et utvalg personer ble bedt om å lytte til forskjellig syntetiske lyder, og sammenligne dem to og to for å avgjøre hvilken som er mest menneskelig.

Problemstilling 2:

- Når du bruker en strømmetjeneste, slik som Netflix, etterlater du en masse klikkedata.
- Strømmetjenestene ønsker å bruke disse klikkedataene for å komme med best mulig personlig forslag til hva du bør se neste gang.

Hva har disse problemstillingene til felles?

Litt om statistikk

- Svar: de dreier seg begge om å analysere rangerte data, som kan være inkonsistente over tid.
- Anvendelser som tilsynelatende er veldig forskjellige kan være nokså like fra et statistisk ståsted.
- Dermed kan de løses med samme statistiske metoder.
- Statistikere er veldig nyttige!

Foreløpig pensum

- Lærebok:
  From *Jay L. Devore & Kenneth N. Berk: Modern Mathematical Statistics with Applications, second edition, 2012. Springer. ISBN: 978-1-4614-0390-6.*
  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 7.2 og 7.4 (unntatt bevisene på sidene 374-375 og side 376).
  2. Kapittel 6: Avsnitt 6.4.
  3. Kapittel 8: Avsnitt 8.3-8.5.
  4. Kapittel 9: Avsnitt 9.1-9.4 og 9.5 (bare sidene 467-469 og 475-478).
  5. Kapittel 10: Avsnitt 10.1-10.5.
  6. Kapittel 11: Avsnitt 11.1 og 11.3 (bare sidene 572-573).
  7. Kapittel 12: Hele.

Kapittel 7: Punktestimering

- Metoder for punktestimering
- Informasjon og effisiens.

Kapittel 6: Fordelinger knyttet til normalfordelte utvalg

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

Kapittel 8: Konfidensintervaller

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

Kapittel 9: Hypotesetesting

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

Kapittel 10: Statistiske metoder for to utvalg

- Inferens (konfidensintervaller og testing) om to forventningsverdier, populasjonsandeler og to varianser
- Metoder for store og små utvalg
- Parforsøk.

Kapittel 11: Variansanalyse

- Enveis variansanalyse.

Kapittel 12: Lineær regresjon

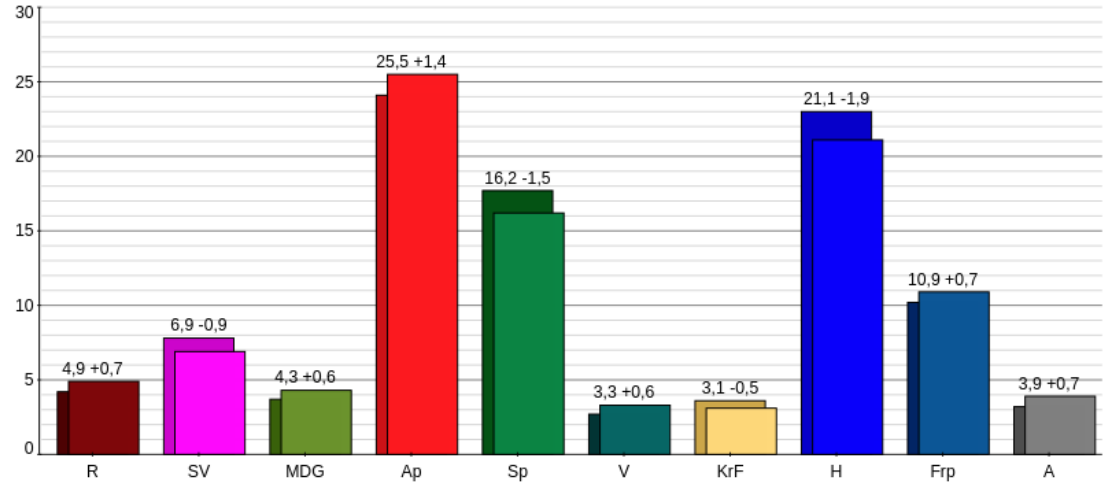
- Enkel lineær regresjon
- Multippel lineær regresjon
- Korrelasjon
- Estimering og inferens.

Sannsynlighetsmodeller og statistiske metoder

- Vi ønsker å få kunnskap om en **populasjon** basert 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

Norstat for NRK 12. august 2021



Fakta om meningsmålingen

Valg	Stortingsvalg
Område	Hele landet
Institutt	Norstat
Antall spurte	11400
Tatt opp	28/7 - 11/8

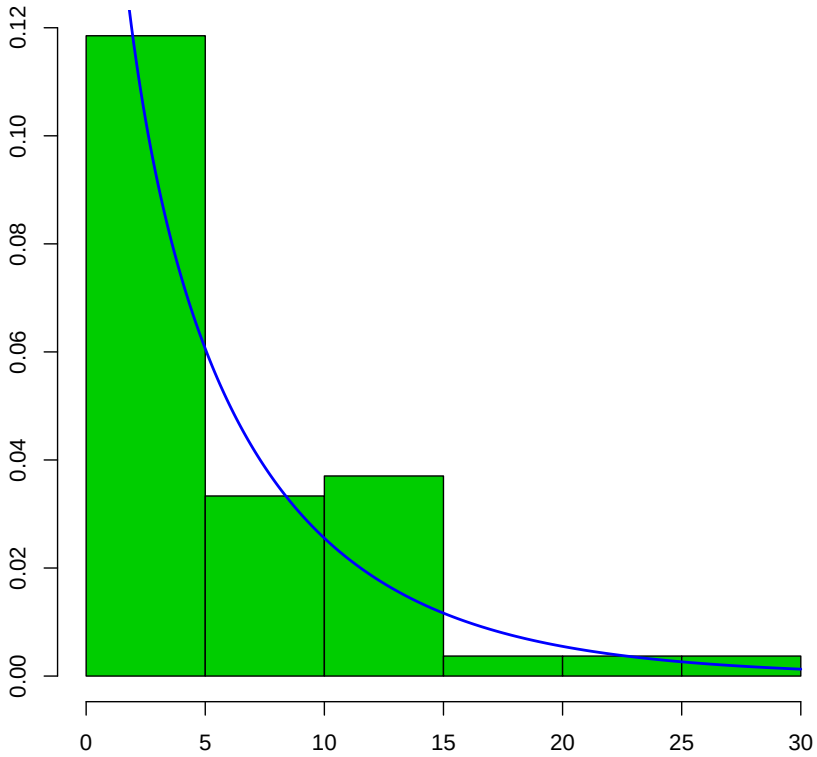
- Av de 8208 som ville ha stemt om det var stortingsvalg i morgen, ville 1330 ha stemt på Sp.
- Sps oppslutning er da $1330/8208 \approx 0.162$.
- Hvor sikkert er dette estimatet?

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 $\hat{p} = \frac{1330}{8208} \approx 0.162$.
- Basert å antakelsene over kan vi lage et **konfidensintervall** for p, som angir hvor sikkert estimatet er (Kap. 8).

Eksempel 2

Histogrammet viser mengden nedbør per døgn fra mai til juli 2016 på Blindern (de 54 dagene det regnet):



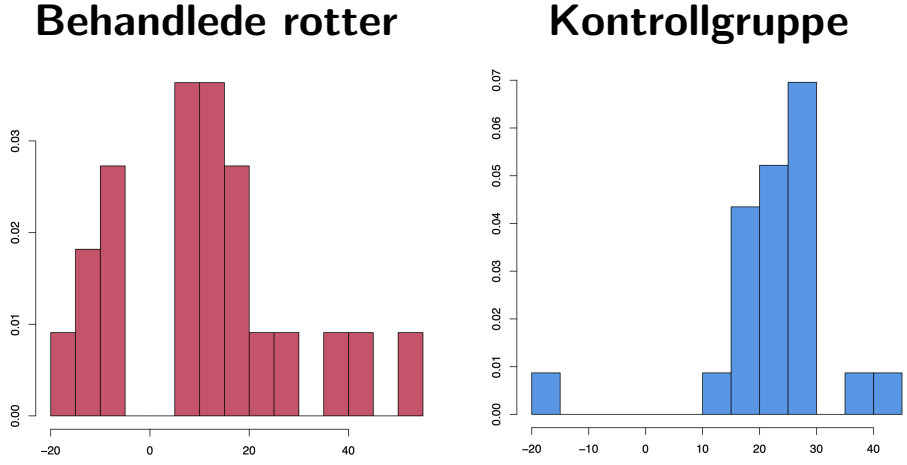
- Gammafordelingen med $\alpha = 0.72$ og $\beta = 7.4$ gir en nokså god beskrivelse av dataene.
- Hvordan finner vi disse estimatene for α og β?

Eksempel 2

- Vi har data $x_1, \ldots, x_n$, der x_i er mengen nedbør den i-te dagen det regner.
- Vi antar at $x_1, \ldots, x_n$ er observerte verdier av $X_1, \ldots, X_n \overset{uif}{\sim} Gamma(\alpha, \beta)$.
- Vi har ikke noen "naturlige" estimatorer for α og β.
- Vi har derfor behov for **metoder** vi kan bruke for å finne estimatorer i situasjoner der vi ikke uten videre har "naturlige" estimatorer (Kap. 7).

Eksempel 3

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.



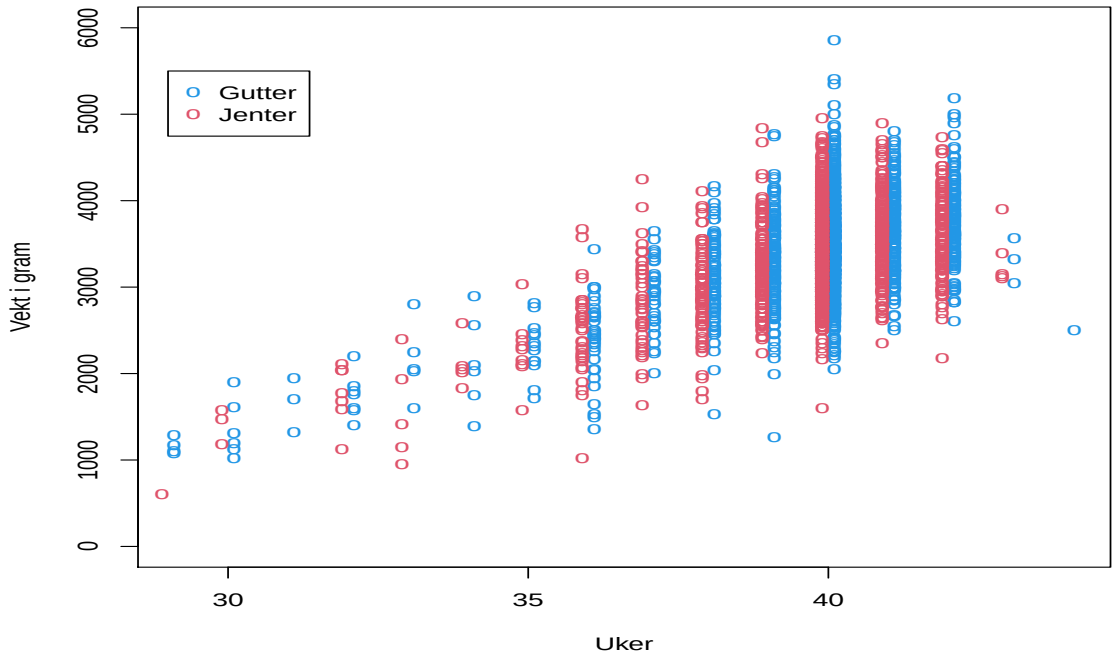
- Kan vi med rimelig grad av sikkerhet si at det er en forskjell i vektøkning mellom behandlede og ubehandlede rotter?
- Kan vi gi et anslag på forskjellen i vektøkning?
- Hvor sikkert er dette anslaget?

Eksempel 3

- 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 $X_1, \ldots, X_n \overset{uif}{\sim} N(\mu_1, \sigma_1^2)$ og $Y_1, \ldots, Y_n \overset{uif}{\sim} N(\mu_2, \sigma_2^2)$.
- 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 i 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 4

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 4

- 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 4

- 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).