Innledning til STK1110
Statistiske metoder og dataanalyse 1

høsten 2012

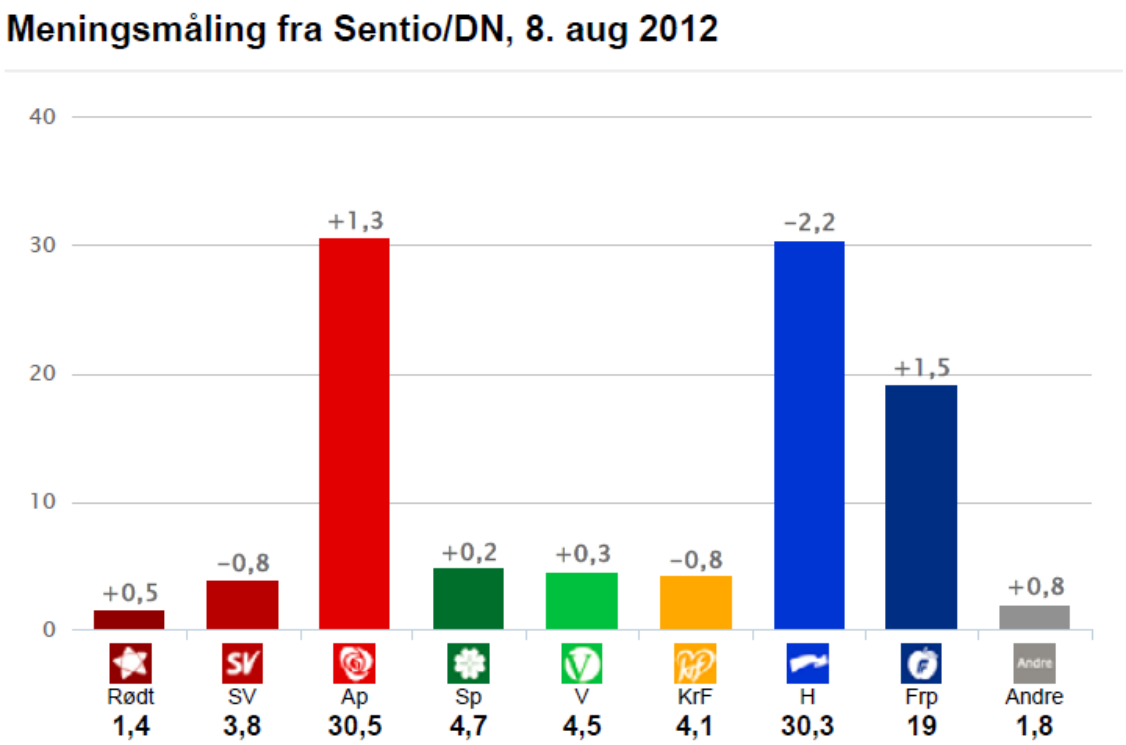
I denne innledningen vil vi først vise fire eksempler på noen av problemstillingene vi skal se på i STK1110.

Felles for eksemplene er at vi har samlet inn data (tall) som skal hjelpe oss til å svare på problemene våre.

Videre skal vi antyde hvordan sannsynlighetsmodeller fra STK1100 kan brukes til å beskrive usikkerheten i dataene våre, og hvordan problemstillingene vi er interessert i kan "oversettes" til utsagn om modellparameterne.

Det motiverer at vi i STK1110 vil studere statistiske metoder der en på grunnlag av observerte data kan trekke konklusjoner om verdiene av modellparametrene.

Eksempel 1: Partibarometer august 2012



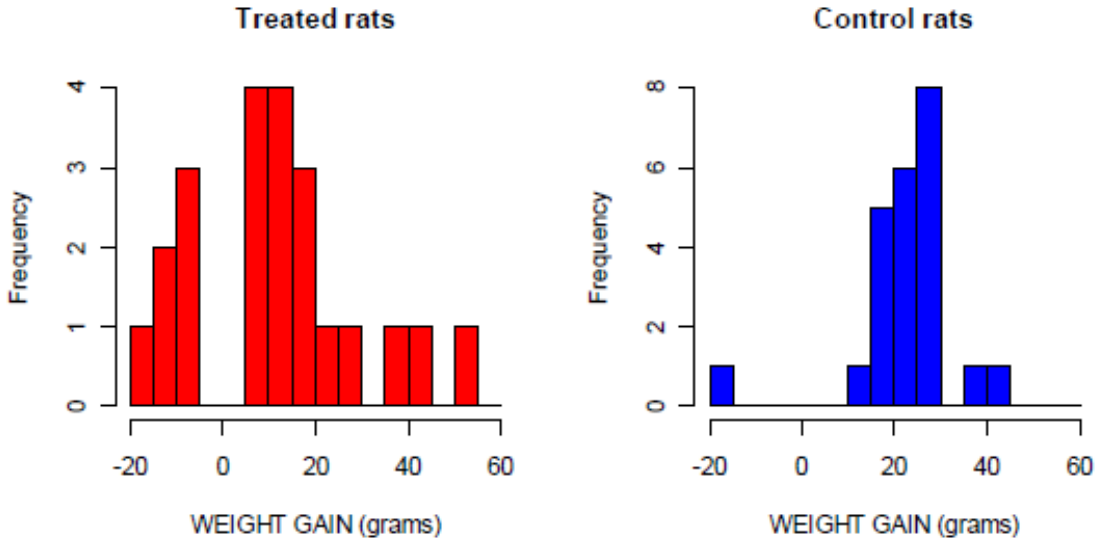
Av 914 spurte, som hadde stemt hvis det var valg dagen etter, var det 279 som ville ha stemt Ap.

I barometeret får Ap dermed en oppslutning på 279/914 = 30.5 %

Hvor sikkert er dette anslaget?

Eksempel 2: Vektøkning og ozon

I et forsøk lot en 22 rotter være i et miljø med ozon (behandlede rotter) og 23 rotter være i et ozonfritt miljø (kontroll rotter).
En registrerte så vektøkningen for rottene 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?
Og hvor sikkert er dette anslaget?

Eksempel 3: Trafikkulykker

Det har vært en nedgang i antall alvorlige trafikkulykker de senere årene, jf. www.ssb.no/vtuaar

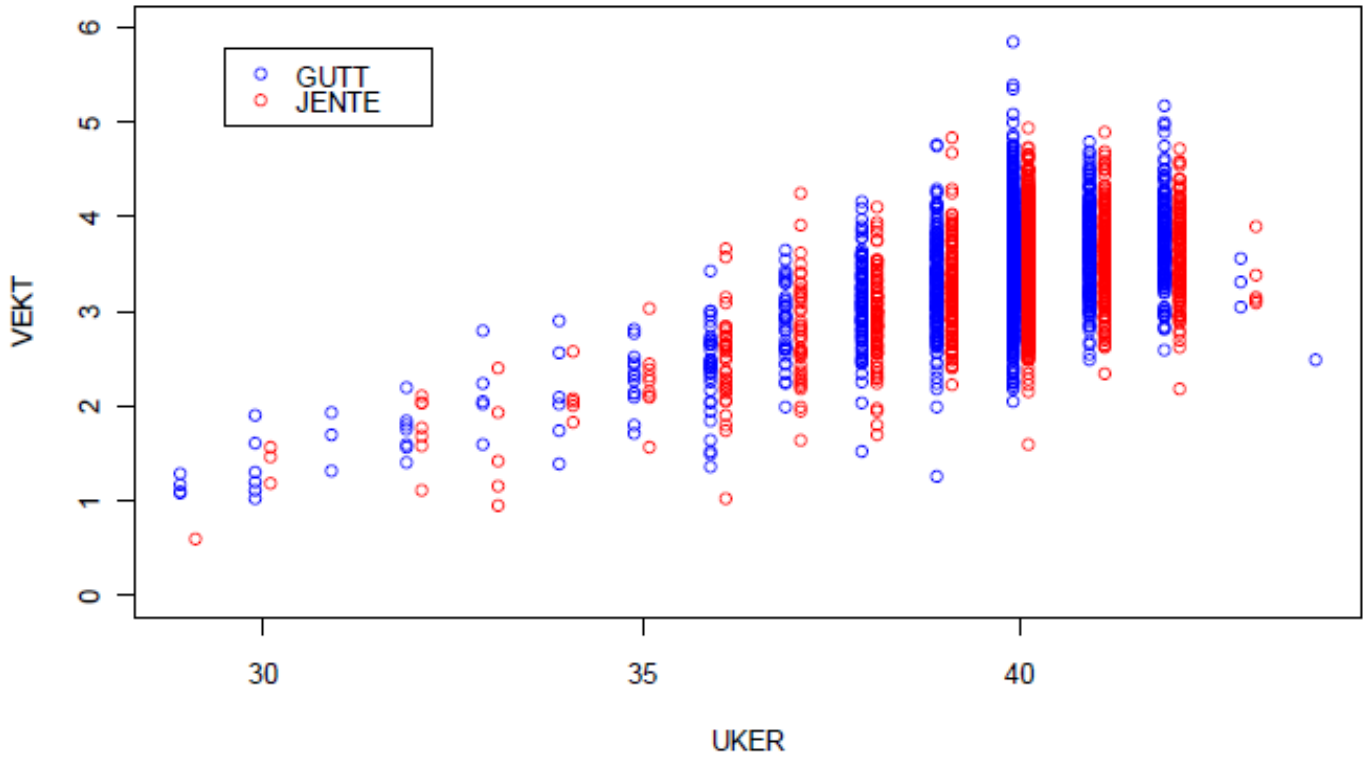
Antall dødsulykker de fire siste årene var

- 2008: 237
- 2009: 186
- 2010: 190
- 2011: 158

Kan vi med rimelig sikkerhet si at det har vært en nedgang i risikoen for dødsulykker fra 2010 til 2011? Har det vært en nedgang i risikoen fra 2008 til 2011?

Eksempel 4: Fødselsvekt og lengden av svangerskapet

Figuren viser vekten for et utvalg av omtrent 4000 nyfødte barn:



Kan vi finne en sammenheng som beskriver hvordan fødselsvekten henger sammen med varigheten av svangerskapet og barets kjønn?

Sannsynlighetsmodeller og statistiske metoder

De dataene vi har sett på i eksemplene er usikre (i den forstand at en ny studie ikke vil gi akkurat de samme resultatene selv om den faktiske situasjonen er den samme).

For å ta hensyn til denne usikkerheten, trenger vi en matematisk modell som beskriver data beheftet med usikkerhet, og det er nettopp det en sannsynlighetsmodell gjør.

Den grunnleggende ideen i statistikk er at vi tenker oss at dataene våre er generert fra en sannsynlighetsmodell. Da kan vi vurdere usikkerheten i de virkelige dataene i lys av den variasjonen en vil ha i data som er generert fra sannsynlighetsmodellen.

Videre kan de problemstillingene vi er interessert i "oversettes" til utsagn om parametrene i modellen.

Eksempel 1: Partibarometer for august 2012

For meningsmålingen kan vi bruke følgende sannsynlighetsmodell:

Vi antar at Ap på det aktuelle tidspunktet har oppslutning fra 100 p % av dem som ville ha stemt hvis det hadde vært stortingsvalg.

Vi antar videre at de n = 914 som ville ha stemt, er et tilfeldig utvalg av alle som ville ha stemt hvis det hadde vært valg.

La X være antall som vil stemme Ap ved en slik meningsmåling. Da vil X være binomisk fordelt:

$$X \sim \text{bin}(n,p).$$

Det gir en beskrivelse av den variasjonen en vil ha fra en meningsmåling til en annen (hvis styrkeforholdene mellom partiene er uendret).

Formålet med meningsmålingen er å anslå (eller estimere) verdien til parameteren p . Vi ønsker også å si noe om hvor sikkert anslaget (eller estimatet) er.

Eksempel 2: Vektøkning og ozon

For rotteforsøket kan vi bruke følgende sannsynlighetsmodell:

Vi antar at vektøkningene for de m = 22 behandlete rottene er observerte verdier av stokastiske variable $X_1, \ldots, X_m$ som er uavhengige og $N(\mu_1, \sigma_1^2)$ - fordelte

Tilsvarende antar vi at vektøkningene for de n = 23 kontroll rottene er observerte verdier av stokastiske variable $Y_1, \ldots, Y_n$ som er uavhengige og $N(\mu_2, \sigma_2^2)$ - fordelte

Formålet med forsøket er å avgjøre om μ_1 og μ_2 er forskjellige, og også å estimere differansen $\mu_1 - \mu_2$

Vi er også interessert i å si noe om hvor sikkert estimatet for differansen er

For å kunne gjøre dette må vi også estimere variansene σ_1^2 og σ_2^2

Eksempel 3: Trafikkulykker

For trafikkulykkene kan vi bruke følgende sannsynlighetsmodell:

Vi antar at antall dødsulykker i 2008, 2009, 2010 og 2011 er observerte verdier av uavhengige og Poisson-fordelte stokastiske variable

$$X_{2008}, X_{2009}, X_{2010} \text{ og } X_{2011}$$

med forventningsverdier λ_{2008}, λ_{2009} og λ_{2010} og λ_{2011} .

Da angir λ_{2008}, λ_{2009} og λ_{2010} og λ_{2011} den "underliggende risikoen" for dødsulykker i hver av de fire årene, og avvik fra disse skyldes tilfeldige variasjoner.

Vi er interessert i å avgjøre om λ_{2011} er mindre enn λ_{2010}.
Vi er også interessert i å avgjøre om λ_{2011} er mindre enn λ_{2008} .

Eksempel 4: Fødselsvekter

For fødselsvektene kan vi bruke følgende sannsynlighetsmodell:

La Y være fødselsvekten til et barn av kjønn s ($s = j, g$) der svangerskapet har vart i u uker. Vi vil anta at Y er normalfordelt med forventningsverdi $\mu = \alpha_s + \beta_s(u - 40)$ og varians σ^2

Her er α_s forventet fødselsvekt for et barn av kjønn s ved fullgått svangerskap (40 uker), mens β_s er forventet vektøkning per uke

Vi er interessert i å estimere parametrene $\alpha_j, \alpha_g, \beta_j$ og β_g

Vi kan også være interessert i å avgjøre om α_j og α_g er like og om β_j og β_g er like

Oversikt over STK1110

- Kap 7: Punktestimering for ett utvalg (eks 1)
- Kap 8: Konfidensintervall for ett utvalg (eks 1)
- Kap 9: Hypotesetesting for ett utvalg
- Kap 10: Statistiske metoder for to utvalg (eks 2 og 3)
- Kap 12: Lineær regresjon (eks 4)

I tillegg kommer avsnitt 6.4 om fordelinger som er viktige for statistiske metoder for normalfordelte utvalg