Innledning til STK1110
Statistiske metoder og dataanalyse 1

høsten 2014

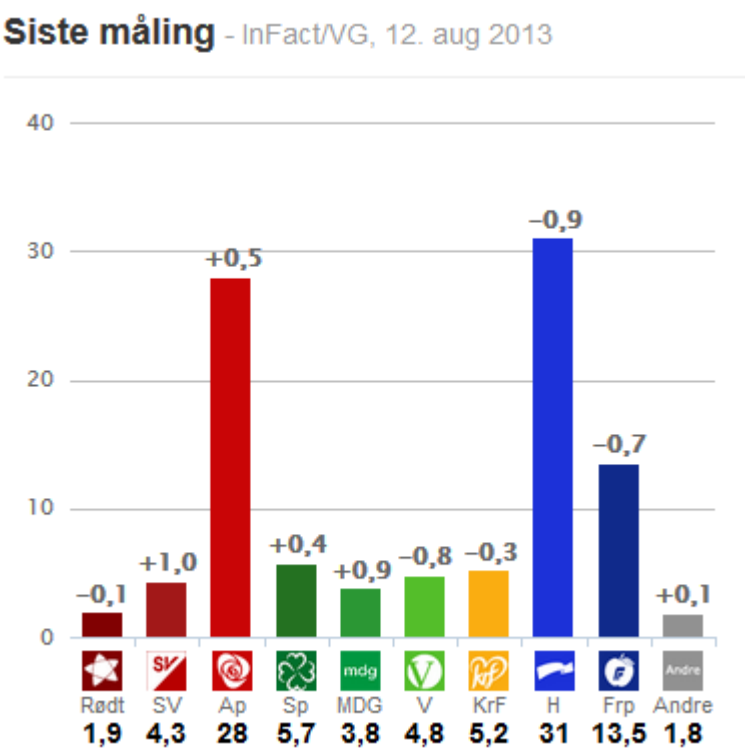
I denne innledningen vil vi først vise fire svært enkle eksempler på noen av problemstillingene vi skal se på i STK1110.

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

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

I STK1110 skal vi studere statistiske metoder der en på grunnlag av observerte data kan trekke konklusjoner om verdiene av modellparametrene, og kvantifisere usikkerheten knyttet til konklusjonene.

Eksempel 1: Partibarometer august 2013 – like før valget!!!



Av 914 spurte, som hadde stemt hvis det var valg dagen etter, var det 256 som ville ha stemt Ap (målt vha. 'robopoll', automatisk talemaskin)

I barometeret får Ap dermed en oppslutning på 256/914 = 28 %

Hvor sikkert er dette anslaget?

Resultatsammendrag [rediger | rediger kilde]

Oversiktene nedafor er basert på endelige valgresultater fra Kommunal- og regionaldepartementet og fra Statistisk sentralbyrå.[5]

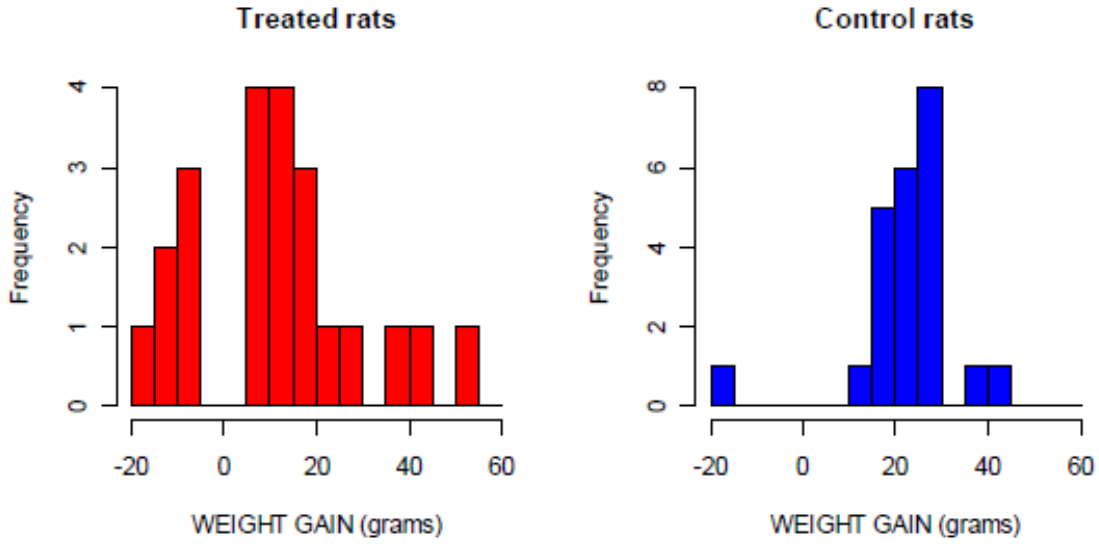
Stemmetall fordelt på partiene (i promille):

Stortingsmandater fordelt på partiene:

Partier[6][7]	Stemmer			Mandater	
	#	%	± pp	#	±
Arbeiderpartiet	874 769	30,8	+4,5 ▼	55	−9 ▼
Høyre	760 232	26,8	+9,6 ▲	48	+18 ▲
Fremskrittspartiet	463 560	16,3	−6,6 ▼	29	−12 ▼
Kristelig Folkeparti	158 475	5,6	0,0 ▬	10	0 ▬
Senterpartiet	155 357	5,5	−0,7 ▼	10	−1 ▼
Venstre	148 275	5,2	+1,4 ▲	9	+7 ▲
Sosialistisk Venstreparti	116 021	4,1	−2,1 ▼	7	−4 ▼
Miljøpartiet De Grønne	79 152	2,8	+2,4 ▲	1	+1 ▲
Rødt	30 751	1,1	−0,3 ▼		
De Kristne	17 731	0,6	+0,6 ▲		
Pensjonistpartiet	11 865	0,4	0,0 ▬		
Piratpartiet	9 869	0,3	+0,3 ▲		
Kystpartiet	3 311	0,1	−0,1 ▼		
Demokratene i Norge	2 214	0,1	0,0[8] ▬		
Kristent Samlingsparti	1 722	0,1	−0,1 ▼		
Det Liberale Folkepartiet	909	0,0	0,0 ▬		
Norges Kommunistiske Parti	611	0,0	0,0 ▬		
Sykehus til Alta	467	0,0	0,0 ▬		

Eksempel 2: Vektøkning og ozon

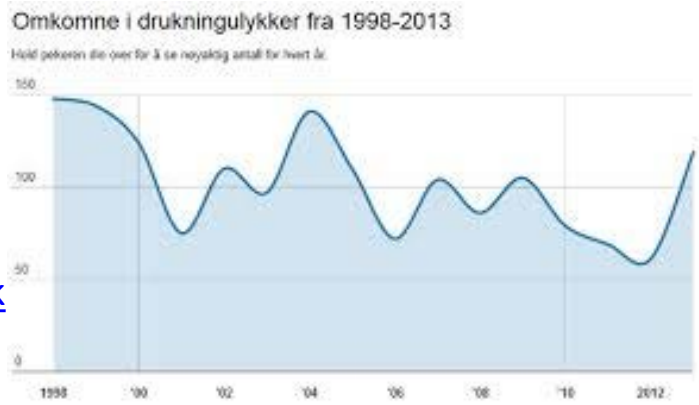
I et forsøk lot en 22 rotter være i et miljø med ozon (behandlede rotter) og 23 rotter være i et ozonfritt miljø (kontroll rotter).
En registrerte så vektøkningen for rottene i løpet av en uke.



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: Drukningsulykker

De siste årene har det vært en nedgang i antall druknede, men antallet steg betraktelig i fjor, se feks.



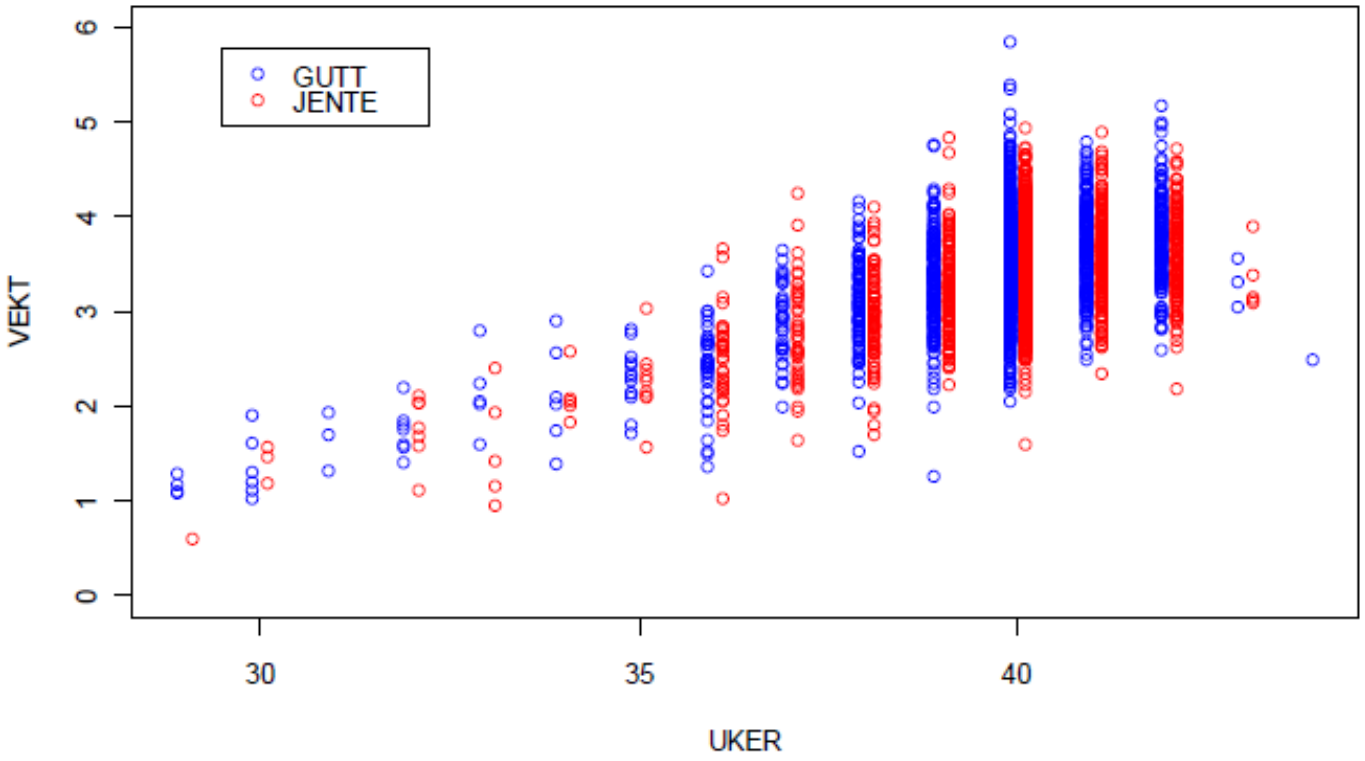
http://www.folkehjelp.no/Presse/Drukningsstatistikk

- 2009: 105
- 2010: 79
- 2011: 69
- 2012: 61
- 2013: 119
- 2014: ? (64 så langt i år)

Kan vi med rimelig sikkerhet si at det har vært en økning i risikoen for drukning fra 2012 til 2013? Har det vært en nedgang i risikoen fra 2009 til 2012?

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.

En av de grunnleggende ideene i statistikk er at vi tenker oss at dataene våre er generert fra en slik 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 ofte "oversettes" til utsagn om parametrene i modellen, som i de enkle eksemplene her.

[I mange reelle situasjoner trenger vi mer komplekse modeller, der vi f.eks. tar høyde for at modellen i seg selv også er usikker!]

Eksempel 1: Partibarometer for august 2013

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.

Mer avansert...:

FOTO: Aas, Erlend

Stortingsvalget

Statistikere: 91 prosent sannsynlighet for blåblått flertall

Inspirert av amerikanske valg-statistikere har tre forskere ved Norsk Regnesentral laget en egen valgmodell.

Lars Molteberg Glomnes
Publisert: 11.jul. 2013 08:40 **Oppdatert:** 11.jul. 2013 08:40

– Det er en statistisk modell, basert på tidligere meningsmålinger fra stortingsvalgene 2005 og 2009, som brukes til å gi en prognose for valget i år, sier seniorforsker Clara-Cecilie Günther til Aftenposten.

Sammen med en gruppe forskere fra Norsk Regnesentral har hun satt sammen en utregning som de mener vil være mer presist enn vanlige meningsmålinger.

Siste saker om politiske partier

Arbeiderpartiet

- Dette må man kunne tillate seg i en valgkamp
- Ønsker lønnstilskudd i Norge
- Jonas Gahr Støre:- Høyre-forslag vil gi mer sykdom
- – Giske eier ikke skamvett

www.nr.no 169.no

Nate Silver

STAY IN THE SHADE

Is Sunscreen A Lifesaver Or A Poison?

By EMILY OSTER

FEATURES

DATALAB

AUG 15
Thousands Of Pictures Are Worth The World

AUG 6
What's The Best Age To Be?

AUG 5
Are Female Scientists Hiding?

AUG 1
Dear Mona, How Many Kids Break A Bone?

JUL 30
We Still Don't Know How Deadly the Ebola Outbreak in West Africa Will Be

JUL 12
More Twins, Fewer Triplets

JUN 26
World Cup Crib Notes: Day 15

RITCHIE KING
Ritchie King is a visual journalist and science reporter for FiveThirtyEight.

EMILY OSTER
Emily Oster is an associate professor of economics at the University of Chicago Booth School of Business.

TOP STORIES

BECHDEL TEST

The Dollar-And-Cents Case Against Hollywood's Exclusion of Women

By WALT HICKEY

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: Drukningsulykker

For drukningsulykkene kan vi bruke følgende sannsynlighetsmodell:

Vi antar at antall drukningsulykker i 2009, 2010, 2011, 2012 og 2013 er observerte verdier av uavhengige og Poisson-fordelte stokastiske variable

$$X_{2009}, X_{2010}, X_{2011}, X_{2012} \text{ og } X_{2013}$$

med forventningsverdier $\lambda_{2009}, \lambda_{2010}, \lambda_{2011}, \lambda_{2012}$ og λ_{2013} .

Da angir $\lambda_{2009}, \lambda_{2010}, \lambda_{2011}, \lambda_{2012}$ og λ_{2013} den "underliggende risikoen" for dødsulykker i hver av de fem årene, og avvik fra disse skyldes tilfeldige variasjoner.

Vi er f.eks. interessert i å avgjøre om λ_{2013} er større enn λ_{2012}.
Vi er også interessert i å avgjøre om λ_{2012} er mindre enn λ_{2009} .

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