TMA4240 Statistikk H2016 [20]

Parutvalg [9.9]
Ett utvalg: estimere en andel [9.10]
To utvalg: estimere differansen mellom to andeler [9.11]
Konfidensintervall for varians [9.12]

Mette Langaas

Institutt for matematiske fag, NTNU

wiki.math.ntnu.no/emner/tma4240/2016h/start/

Estimering

- Mål: gjøre inferens om ukjent parameter, θ, i en valgt fordeling.
- Vi trekker et tilfeldig utvalg fra populasjonen; $X_1, X_2, ..., X_n$ (u.i.f.).
- En estimator gir et anslag for den ukjente parameteren og er en funksjon av stokastiske variabler, $\hat{\theta} = \hat{\theta}(X_1, X_2, ..., X_n)$.
- Hvilke egenskaper bør en god estimator ha?
  - Estimatoren bør være forventningsrett, dvs. $\mathrm{E}(\hat{\theta}) = \theta$.
  - Estimatoren bør ha minst mulig varians, $\mathrm{Var}(\hat{\theta})$, og variansen bør avta når antall observasjoner, n, øker.
- Hvordan kan vi finne estimatorer?
  - ved intuisjon,
  - ved matematisk metode: sannsynlighetsmaksimeringsestimatoren (SME) finner det anslaget som gjør at de observasjonene vi har gjort (utvalget) har maksimal rimelighet!
- Vi har 95% tillit til at sann parameterverdi ligger i et 95% konfidensintervall (KI).
- En ny observasjon har 95% sannsynlighet for å ligge i et 95% prediksjonsintervall.

Hva gjenstår innen estimering?

- Lag et konfidensintervall for differansen i populasjonsgjennomsnittet i kroppsfettprosent etter pedikyr og før pedikyr.
  - Parutvalg (to avhengige utvalg).
- Lag et konfidensintervall for andelen studenter som synes de er flinkere å kjøre bil enn landsgjennomsnittet.
  - Ett utvalg (binomisk): andel
- Synes flere unge menn enn kvinner at de er flinkere enn landsgjennomsnittet til å kjøre bil?
  - To utvalg (binomisk): andeler
- Kraften som trengs til å åpne en vinkork
  - Ett utvalg: undersøke varians.

Effekt av pedikyr på måling av kroppsfettprosent
Eksamen ST0202, desember 2013

Kroppsfettprosenten til en person kan måles ved å sende et svakt elektrisk signal gjennom kroppen for å måle kroppens impedans. Signalet ledes gjennom vannet i kroppen. Muskelmasse inneholder mer vann enn fettvev, og dette kan brukes til å beregne et mål på kroppsfettprosent.

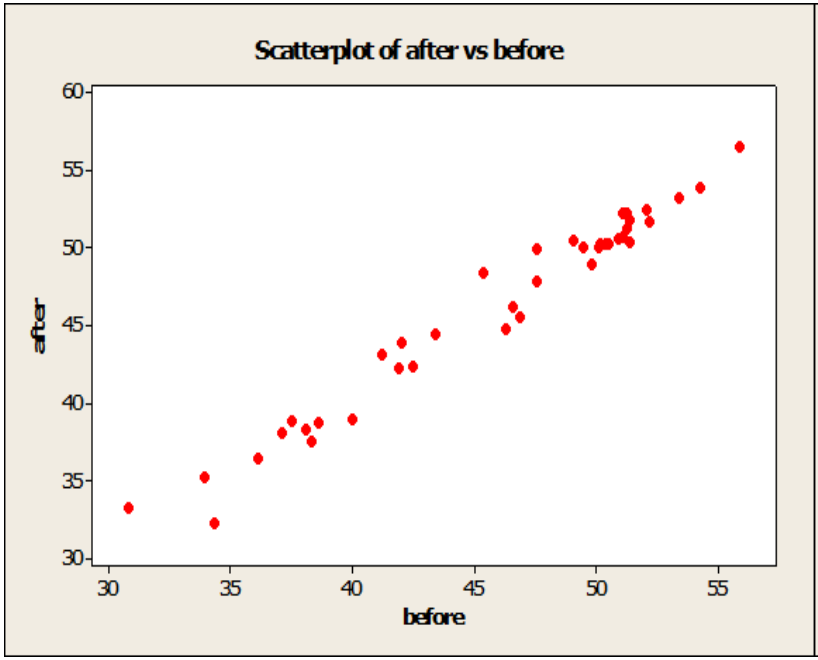
En masteroppgave ved Obesitasklinikken ved St. Olavs Hospital hadde som mål å undersøke om pedikyrbehandling ville påvirke målingen av kroppsfettprosent (som beskrevet over).

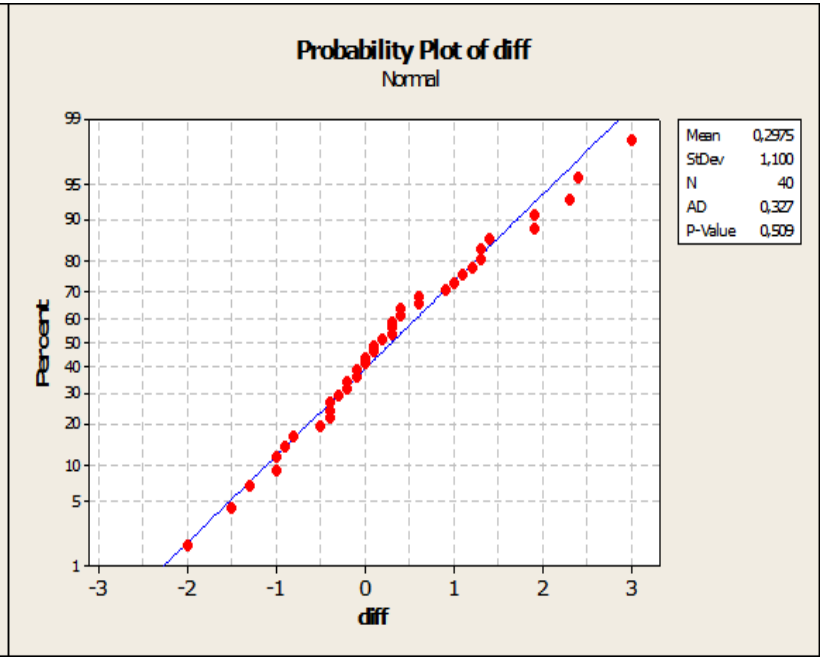
Totalt deltok 40 personer i studien, alle med kroppsmasseindeks (vekt delt på kvadratet av høyde) over 30 kg/m^2. Alle deltakerne ble målt med impedansteknologien to ganger, før og etter en pedikyrbehandling. For hver av de 40 personene, la x_1 angi kroppsfettprosenten målt før pedikyrbehandlingen og x_2 angi kroppsfettprosenten målt etter pedikyrbehandlingen.

Kroppsfettprosent og pedikyr, forts.

Det oppgis følgende. Utvalgsgjennomsnittet før behandling er $\bar{x_1} = 45.80$ og utvalgsstandardavviket før behandling er $s_1 = 6.51$. Utvalgsgjennomsnittet etter behandling er $\bar{x_2} = 46.10$ og utvalgstandardavviket etter behandling er $s_2 = 6.40$. Videre er $s_d = 1.10$ utvalgsstandardavviket til differansen $d = x_2 - x_1$.

a) Er de to utvalgene (før og etter pedikyr) uavhengige? Begrunn svaret.
b) Konstruer et 95% konfidensintervall for differansen mellom populasjonsgjennomsnittet i kroppsfettprosent etter pedikyr og før pedikyr. Forklar hvilke antagelser du gjør for å lage konfidensintervallet.


Scatterplot of after vs before
after
before
60
55
50
45
40
35
30
30
35
40
45
50
55


Probability Plot of diff
Normal
Percent
diff
99
95
90
80
70
60
50
40
30
20
10
5
1
-3
-2
-1
0
1
2
3
Mean 0,2975
StDev 1,100
N 40
AD 0,327
P-Value 0,509

To utvalg: uavhengige eller avhengige?

- Betong: to ulike oppskrifter, A og B, skal sammenlignes. Hvor stor forskjell er det i styrken ("crushing strength") for betong fra oppskrift A og fra oppskrift B?
- Algoritmer: to ulike sorteringsalgoritmer skal sammenlignes. Hvilken algoritme har kortest kjøretid for en spesiell problemstilling?
- Sykdom: tester ut ny blodtrykksmedisin. Hvor mye bedre er den enn nåværende markedsledende blodtrykksmedisin?
- Kosthold: hvor stor vektreduksjon vil man oppleve ved å følge diett A i et halvt år?
- Bildekk: to typer dekk, A og B, skal sammenlignes mhp slitasje. Kan enten sette både A og B-dekk på hver bil eller noen biler med A og noen biler med B.

Studenter og bilkjøring

- Følgende tabell er tatt fra spørreundersøkelse blant studentene i TMA4240 i 2016.
- Her angir n antall studenter i utvalget som hadde sertifikat, og x antall studenter som svarte at de er "bedre enn gjennomsnittet av bilførere i Norge" til å kjøre bil.

	n	x	$\frac{x}{n}$
Menn	200	121	0.61
Kvinner	131	49	0.37
Alle	331	170	0.51

- a) Finn punktestimat og 99% konfidensintervall for andelen av studenter som synes sine kjøreegenskaper er "bedre enn gjennomsnittet".
- b) Finn punktestimat og 99% konfidensintervall for differensen mellom andelen av mannlige studenter og kvinnlige studenter som synes sine kjøreegenskaper er "bedre enn gjennomsnittet".

Konfidensintervall

Anta $X_1, X_2, ..., X_n$ er et tilfeldig utvalg fra en populasjon med fordeling $f(x;\theta)$, med observerte verdier $x_1, x_2, ..., x_n$. Vi vil finne et $(1-\alpha)\cdot 100\%$ konfidensintervall for θ.

1. Finn estimator $\hat{\theta}$ for θ.
2. Bestem $W = h(\theta, \hat{\theta})$, der funksjonen h er slik at W har kjent fordeling (normal, kjikvadrat, Student-t) og ikke er avhengig av andre ukjente parametere enn θ.
3. Vi vet da at $P(w_{1-\frac{\alpha}{2}} < W < w_{\frac{\alpha}{2}}) = 1-\alpha$, der $w_{1-\frac{\alpha}{2}}$ er verdien i fordelingen til W som har areal $\frac{\alpha}{2}$ til venstre, og $w_{\frac{\alpha}{2}}$ har areal $\frac{\alpha}{2}$ til høyre.
4. Løs ulikhetene mhp θ og sett dem sammen igjen slik at vi får

$$P(\hat{\theta}_L < \theta < \hat{\theta}_U) = 1-\alpha.$$

5. Resultat: $(1-\alpha)\cdot 100\%$ konfidensintervall for θ er $[\hat{\theta}_L, \hat{\theta}_U]$ innsatt verdiene $x_1, x_2, ..., x_n$.

Estimering av andel: ett utvalg

- X er antall suksesser i et binomisk forsøk med parametere antallet n og andelen p. Vi vil estimere p. (n er kjent.)
- Estimator $\hat{p} = \frac{X}{n}$ (intuitiv og SME).
- $\mathrm{E}(\hat{p}) = p$ og $\mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n}$.
- Tilnærmet $(1-\alpha)100\%$ konfidensintervall for p (normaltilnærming):

$$[\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$$

Læreboka [9.10]: Dette resultatet er kalt "metode 1". Metode 2 er å løse 2grads ligning for p.

Estimering av andel: to utvalg

- X_1 er antall suksesser i et binomisk forsøk med parametere antallet n_1 og andelen p_1.
- X_2 er antall suksesser i et binomisk forsøk med parametere antallet n_2 og andelen p_2.
- Vi vil estimere $p_1 - p_2$.
- Estimator $\hat{p}_1 - \hat{p}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$.
- $\mathrm{E}(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$ og
- $\mathrm{Var}(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$.
- Tilnærmet $(1-\alpha)100\%$ konfidensintervall for $p_1 - p_2$ (normaltilnærming):

$$[(\hat{p}_1 - \hat{p}_2) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}]$$

- I aldersgruppa under 25 år er faktisk kvinnene bedre til å kjøre bil enn menn, ifølge statistikken, sier Bjørnskau til Dagbladet.no.

Kvinner blir ikke dårligere sjåfører etter fylte 25, men menn blir bedre. Trening betyr nemlig svært mye for kjøreferdighetene.

- Allerede etter ni måneders bilkjøring er skaderisikoen halvert for begge kjønn. Det skjer en voldsom læring den første tida, påpeker Bjørnskau.

LEVER FARLIG: Syklistene kommer dårligst ut av skadestatistikken. Og egentlig er syklistrisikoen enda høyere enn statistikken viser.
Foto SCANPIX

Forsikringsselskapene sluttet i 2006 med å føre detaljert statistikk over kvinner og menns skaderisiko i bil. Da ble det forbudt å gi kvinner og menn forskjellig priser på sin bilforsikring.

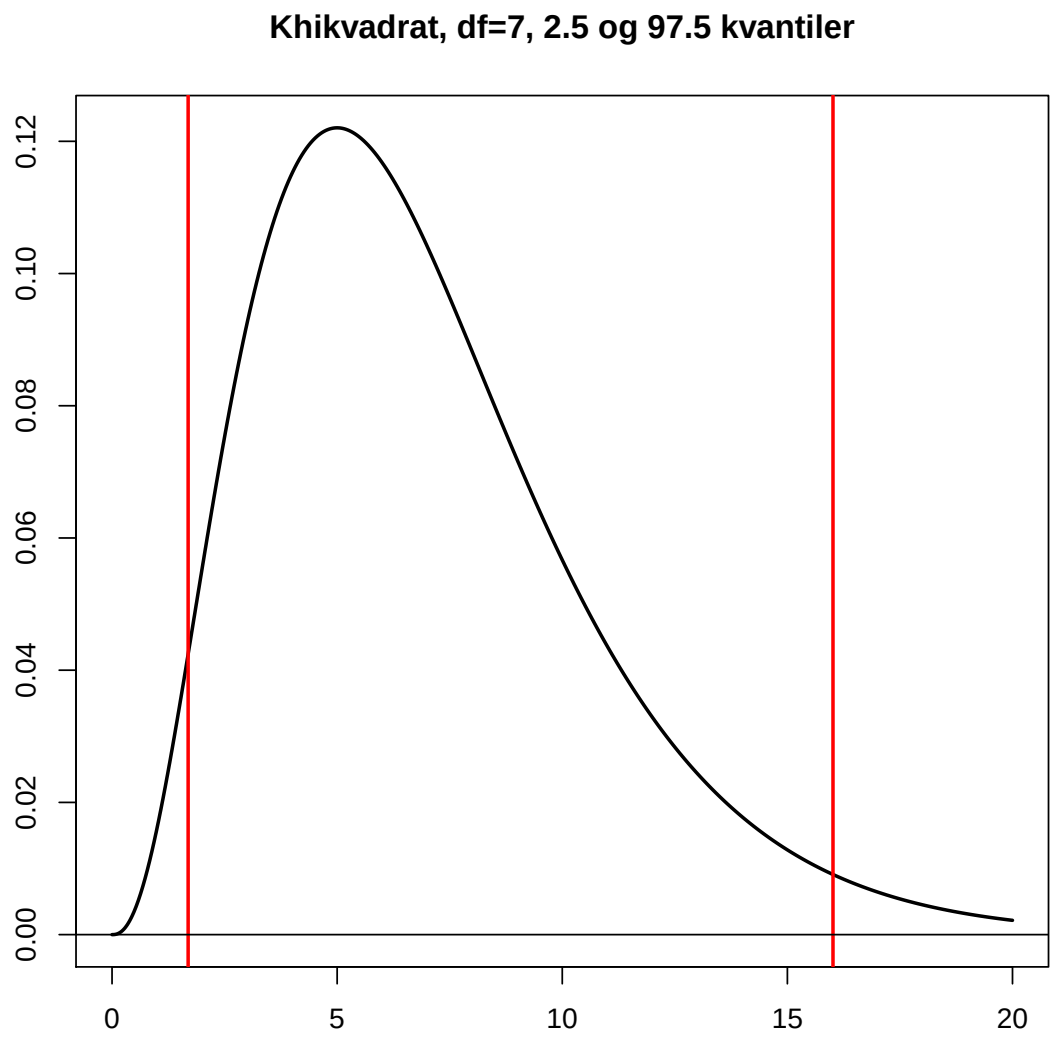
Kraften for å åpne en vinkork
(basert på eksamen desember 2007, oppgave 3)

Kraften som er nødvendig for å trekke ut en kork fra en vinflaske er en viktig egenskap ved korken. Dersom kraften er for liten gir ikke korken nok beskyttelse mot innsig av luft for vinen inni flasken. Dersom kraften er for stor, vil korken være vanskelig å fjerne. Kraften (Newton) for en bestemt korktype kan antas normalfordelt med forventning μ og standardavvik σ.

Et utvalg av 8 tilfeldig valgte flasker med samme korktype ble plukket ut og kreftene som var nødvendig for å fjerne korkene var:
305.98 205.48 322.97 198.58 191.76 288.50 341.18 222.62

Generelt ønsker en lavest mulig standardavvik for korkene, og det er et krav at standardavviket for kraften ikke er større enn 36.0. Produsenten av korkene hevder at dette kravet er oppfylt.

a) Lag et 95% konfidensintervall for standardavviket basert på det tilfeldige utvalget.
b) Er 36 et tall vi tiltro til som standardavvik for kraften til å fjerne en kork av denne korktypen?


Khikvadrat, df=7, 2.5 og 97.5 kvantiler
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0
5
10
15
20

Kritiske verdier i χ^2-fordelingen

$$P(\mathcal{X}^2 > \chi^2_{\alpha,\nu}) = \alpha$$

$\nu \backslash \alpha$	.995	.990	.975	.950	.050	.025	.010	.005
1	.000	.000	.001	.004	3.841	5.024	6.635	7.879
2	.010	.020	.051	.103	5.991	7.378	9.210	10.597
3	.072	.115	.216	.352	7.815	9.348	11.345	12.838
4	.207	.297	.484	.711	9.488	11.143	13.277	14.860
5	.412	.554	.831	1.145	11.070	12.833	15.086	16.750
6	.676	.872	1.237	1.635	12.592	14.449	16.812	18.548
7	.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300

Konfidensintervall for varians [9.12]

- $X_1, X_2, \ldots, X_n$ u.i.f normal, $\mathrm{E}(X_i) = \mu$ og $\mathrm{Var}(X_i) = \sigma^2$.
- Et $(1-\alpha)100\%$ konfidensintervall for σ^2 er

$$\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},(n-1)}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},(n-1)}}$$

hvor $\chi^2_{\frac{\alpha}{2},(n-1)}$ er verdien i kjikvadrat-fordelingen med $n-1$ frihetsgrader som har areal $\frac{\alpha}{2}$ til høyre, dvs. $P(V > \chi^2_{\frac{\alpha}{2},(n-1)}) = \frac{\alpha}{2}$, og $\chi^2_{1-\frac{\alpha}{2},(n-1)}$ er verdien i kjikvadrat-fordelingen med $n-1$ frihetsgrader som har areal $\frac{\alpha}{2}$ til venstre, dvs. $P(V < \chi^2_{1-\frac{\alpha}{2},(n-1)}) = \frac{\alpha}{2}$.

Situasjon	Fokus	Har	Observator	Fordeling
Ett utvalg N	μ	kjent σ^2	$Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$	normal
Ett utvalg N	μ	ukjent σ^2	$T = \frac{\overline{X}-\mu}{S/\sqrt{n}}$	t med $n-1$ frihetsgr.
To uavhengige utvalg N	$\mu_1 - \mu_2$	kjente σ_1^2, σ_2^2	$Z = \frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}$	normal
To uavhengige utvalg N	$\mu_1 - \mu_2$	ukjente σ_1^2, σ_2^2	$T = \frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{S_1^2/n_1+S_2^2/n_2}}$	tilnærmet t-fordelt med ν^* frihetsgr.
To par-utvalg N	$\mu_d = \mu_1 - \mu_2$	ukjent σ_d	$T = \frac{\overline{D}-\mu_d}{S_d/\sqrt{n}}$	t med $n-1$ frihetsgr.
Ett utvalg binomisk	p	-	$Z = \frac{\widehat{P}-p}{\sqrt{p(1-p)/n}}$	tilnærmet normal
To utvalg binomisk	$p_1 - p_2$	-	$Z = \frac{(\widehat{P}_1-\widehat{P}_2)-(p_1-p_2)}{\sqrt{p_1(1-p_1)/n_1+p_2(1-p_2)/n_2}}$	tilnærmet normal
Ett utvalg	σ^2	ukjent μ	$V = \frac{(n-1)S^2}{\sigma^2}$	Kji-kvadrat med $n-1$ frihetsgr.

Læringsmål

- Være med de ulike stegene i utledningen av et konfidensintervall - i en GENERELL situasjon.
- For de situasjonene vi har sett på (tabell forrige slide), kjenne til resultat for estimator og tilhørende fordeling.
- Kunne bruke de to punktene over til å lage konfidensintervall for alle situasjoner - og nye situasjoner.
- Tips: Skriftlig innlevering 3: her er 2c og 3c som gir deg trening i en ny situasjon du ikke har sett før.