Regneøvelser STK1000 Høsten 2016

Her kommer øvelsesoppgavene til neste uke, og en oversikt over tidligere oppgaver. Alle oppgaver refererer til 8. utgave av læreboka hvis ikke annet er nevnt. KFF-oppgave er oppgaver Kathrine Frey Frøslie har laget eller tilpasset.

På kursnettsiden vil det regelmessig legges ut pdf-filer av det som ble skriblet på smart-tavlen i løpet av oppgavegjennomgangen på fredag. Men disse notatene er ikke renskrevet, og må tas slik de er. Husk at alle oppgaver som gis i kurset også er på pensumlista, og at å gjøre oppgaver er den beste måten dere kan komme dere gjennom pensum på.

Litt om R og Rstudio:

Vi bruker programvaren R og Rstudio i STK1000. Alle må sørge for å komme i gang med det.

1. Du må laste ned og installere R først.Last ned R herfra: https://www.r-project.org/ ved å trykke på **download R** i første avsnitt, og følge instruksjonene.
2. Last så ned og installer Rstudio herfra: https://www.rstudio.com/ Rstudio gjør det litt enklere å bruke R.
3. Når du skal gjøre en analyse i Rstudio, må du starte Rstudio, og ikke R. Logoen/ikonet for Rstudio ser slik ut:

Før du kan bruke Rstudio (eller R) til noe annet enn en litt diger og upraktisk kalkulator, må du klare å laste ned datafiler fra kursets nettside, lagre dem på din egen maskin, og lese dem inn i R. Da må du vite hvor filene ligger, og hvordan du får tak i filbanen/adressen til filene du har lastet ned.

Spørsmål om dette kan du rette til universitetets IT-hjelpetjeneste, «Houston»: http://www.uio.no/tjenester/it/kontakt/houston/ Send mail, ring, eller møt opp der, hvis det blir for travelt på gruppetimene til slike spørsmål.

Det kan være svært nyttig å gå gjennom det interaktive R-kræsjkurset i 8 steg som du finner her: http://tryr.codeschool.com/ Jeg har gått gjennom hele, både for å sjekke innholdet og tidsbruken. Med unntak av Ch4, som jeg synes var litt klønete satt opp, var innholdet utmerket. Jeg antar at du vil bruke mellom 5 og 20 minutter på hvert steg, avhengig av hvor fort du skriver/leser, og hvor mye statistikk/programmering du kan fra før. Det er vel anvendt tid!

Jeg legger også ut et nybegynnerscript på nettsiden. Jobber du deg gjennom det, får du mye av den samme oppstarthjelpen som i det interaktive kurset på tryr.codeschool.com.

Nettsiden "Getting started with R" gir videre en komprimert liste med tips og nyttige nettsteder for den som ikke kjenner R fra før: https://support.rstudio.com/hc/en-us/articles/201141096-Getting-Started-with-R

OBS: I mange oppgaver bes dere å oppsummere data, eller å analysere sammenhenger i data.

Noen lurer kanskje på helt spesifikt hvilke figurer og tall den som har laget oppgaven forventer å få som svar på slike spørsmål. Noen lurer av og til på om selve datafila skal legges ved. Svaret er at disse vurderingene er en del av det å gjøre oppgaven. I statistikk er det sjelden ett enkelt tallsvar vi er ute etter. Derimot er det nesten alltid en stor og viktig del av analysen å velge hva man synes er den beste måten å oppsummere og analysere et datasett på. Forståelsen for faget viser man altså først når man selv klarer å velge ut det som er essensielt for dataene. Det er altså opp til dere å velge den måten dere synes det er best å beskrive data på.

Uke 45 og 46

I disse to ukene er temaet regresjonsanalyse. Vi går først tilbake til avsnitt 2.4 og 2.5 for definisjoner og tekniske tips, før vi fortsetter med Ch 10 og Ch 11. Oppgavene er som følger:

- Fra boka: 2.79, 2.80 (disse er basert på samme datasett som oppgave 2.26, 2.51, 2.52 som var ukesoppgaver i uke 38), og 2.110.
- Fra boka: 10.32, 10.33, 10.34, 10.35, 10.3 og 10.37 (alle er basert på samme datasett).
- Fra boka: 11.1, 11.2, 11.23, 11.24, 11.25, 11.26 a,b.
- Eksamensoppgaver: Oppgave 3 STK1000 V05, Oppgave 3 STK1000 H06, Oppgave 2 STK1000 H14, Oppgave 1,2,3 desember 2008, Oppgave 3 desember 2011.

Uke 43 og 44

Disse to ukene handler om hypotesetesting. Jeg har valgt å fokusere på det generelle med hypotesetesting først.

I følgende oppgaver er det svært lite regning, og disse kan være en god hjelp for å kontrollere om dere har forstått stoffet. Mange vil dere få svar på på tirsdag i uke 43.

- Fra boka: Oppgave 6.53, 6.54, 6.55, 6.58, 6.73, 6.83, 6.84, 6.87, 6.86, 6.94, 6.95, 6.106, 7.64, 7.65 a).

Så kommer hypotesetester som sammenligner to grupper:

- Fra boka: Oppgave 7.68, 7.70, 7.75, 7.77, 7.85, 7.89, 15.14, 15.15, 15.16, 15.17. (OBS: Kapittel 15 må lastes ned og skrives ut fra hjemmesiden til læreboka: http://www.macmillanlearning.com/Catalog/studentresources/ips8e#t_922171)

Deretter kommer ett-utvalgs-tester:

- Fra boka: Oppgave 7.22, 7.23, 7.38, 7.43.

Uke 42

- Fra boka: Oppgave 5.14, 5.20, 5.21, 5.23, hvis dere ikke gjorde dem i forrige uke. Løsningsforslag ligger ute på kurssiden.

- Fra boka: Oppgave 5.25, 5.28, 5.50, 5.51, 5.53, 5.60, 5.62.
- Fra boka: Oppgave 6.19, 6.20, 6.27, 6.28, 6.30, 6.33.

.

Uke 40

- Fra boka: Oppgave 4.79, 4.80, 4.82, 4.81, 4.88, 4.90, 4.111, 4.112, 4.115, 4.116, 4.131, 5.14, 5.20, 5.21, 5.23.
- Midtveiseksamen H2015: Hele unntatt oppgave 13 og 14. Dette oppgavesettet ligger under ukesoppgavene for dette semesteret.

Uke 39

- KFF-oppgave 12: Gå gjennom R-scriptet Sampling_distributions_Rstudio. Det er basert på simuleringer, så du trenger ikke å laste ned noe datasett. Kjør gjennom scriptet i Rstudio, og bruk i tillegg hjelpefunksjonene i R (for eksempel `?rnorm` og `?abline`) eller Google til å finne ut hva det er som foregår i dette scriptet.
  Hvordan endrer spredningen i histogrammet seg når utvalgsstørrelsen øker?
  Hvordan endrer variabiliteten i gjennomsnittet seg når utvalgsstørrelsen øker?
- KFF-oppgave 13: La oss si at vi har spurt 30 mennesker om de stemmer Høyre, og 9 svarer «Ja». La «Nei» kodes med (oversettes til) 0 og «Ja» kodes med (oversettes til) 1. Vis at en andel er det samme som et gjennomsnitt.
- Fra boka: Oppgave 3.90, 3.92, 4.4, 4.10, 4.11, 4.27, 4.28, 4.29, 4.42, 4.45, 4.52, 4.54, 4.55, 4.57, 4.62, 4.65
- Midtveiseksamen H2014: Oppgave 9, 10, 19. Disse oppgavene finner du i venstremargen på hjemmesiden til kurset, under Oppgaver – Deleksamen.
- Midtveiseksamen H2013: Oppgave 3, 4, 7 og 8.

Uke 38

- Midtveiseksamen H2014: Oppgave 3 og 18. Disse oppgavene finner du i venstremargen på hjemmesiden til kurset, under Oppgaver – Deleksamen.
- Midtveiseksamen H2013: Oppgave 1, 2, 5, 6, 9 og 18.
- KFF-oppgave 8: Last inn datafila iqdata i Rstudio, og gå gjennom R-scriptet Hjelpescript_til_IQdata.R. Begge filene ligger på hjemmesiden til kurset.
- Fra boka: Oppgave 2.26, 2.51, 2.52 (Basert på samme datasett, gjøres i Rstudio)
  - Oppgave 2.28, 2.31, 2.50, 2.53
- KFF-oppgave 9: Hva er Simpson's paradoks?
- KFF-oppgave 10: Søk opp «Anscombe's quartet», og formulér en twittermelding (max 140 tegn) om hva som er essensen i disse dataene.
- KFF-oppgave 11: Gå hjem og tell klær og sko i klesskapet ditt. Undertøy, badetøy, sokker, stillongs og strømpebukser telles ikke. Du skal telle antall plagg eller sko-par i hver kategori;
  kategoriene for klær er

- ✓ Hverdagstøy (f.eks. t-skjorter, singleter, skjorter, gensere, bukser, skjørt, kjoler,… (Skjerf er egen kategori))
- ✓ Festklær
- ✓ Treningstøy (klær du ikke bruker, med mindre du skal på trening. Du må selv avgjøre om et plagg er i kategorien treningstøy eller hverdagstøy)
- ✓ Ytterklær (luer, votter etc telles ikke)
- ✓ Skjerf/sjal

Kategoriene for sko er

- ✓ Hverdagssko
- ✓ Pensko
- ✓ Vintersko
- ✓ Sportssko

Du må selv definere hvilken kategori dine egne klær og sko tilhører. Det viktigste er at du kommer frem til et tall i hver kategori. Fyll så ut nettskjemaet som du finner her: https://nettskjema.uio.no/answer/75702.html

OBS: Denne delen av oppgaven må gjøres senest tirsdag ettermiddag, så datafila kan legges ut for analyse.

Skriv opp hvilke bivariate sammenhenger du kunne tenke deg å undersøke, og sjekk det når fila blir lagt ut.

Uke 37 (12/9-16/9):

Fra boka: Oppgave 3.38, 3.43, 3.52, 3.56, 3.63. Obs: Gjør oppgave 3.63 ved hjelp av R i stedet for å bruke Table B:

```
leilighetsnr    <- 1:33   # Nummererer de 33 leilighetene i lista
tilfeldige_tall <- sample(leilighetsnr,5,replace=FALSE) # Trekker 5 tilfeldige tall
tilfeldige_tall <- sort(tilfeldige_tall)
```

KFF-oppgave 7: Botox mot migrene. Basert på en sann historie.

a) En middelaldrende kvinne på Manhattan går til plastikk-kirurgen sin for å glatte ut «sinnarynken». Hun opplever å bli kvitt migrene. Hun forteller det til en venninne som har samme erfaring, og når hun senere intervjues i et kvinneblad om sine erfaringer med plastisk kirurgi, får dette stor plass. Hvorfor kalles dette anekdotiske data? Hvor nyttig er slike data? Kom med andre eksempler på anekdotiske data.
b) En spesifikk Manhattan-lege har diverse data tilgjengelig i sine pasientregistre. Ikke alle data i registeret er like komplette, for det er ikke alle som har svart på alle spørsmål, og noen ganger har legen hatt dårlig tid når ting har blitt registrert i systemet. I Norge har vi databaser med helseopplysninger fra generelle helseundersøkelser, vi har nasjonale registre som Medisinsk fødselsregister, og databanken til Statistisk sentralbyrå. Hva kan man kalle slike data? Hva er styrken og svakheten med slike data?
c) Man ønsker å gjøre en observasjonell studie for å undersøke sammenhengen mellom botox og migrene, og et spørreskjema om «korrigerende kirurgi» og overgangsplager/migrene sendes ut til et utvalg. Hva kjennetegner en observasjonell studie? Hva er populasjonen i dette tilfellet? Hva mener vi med et utvalg? Hvordan kan man velge hvem som bør inviteres til å delta i studien?
d) Diskutér styrker og svakheter med dette studiedesignet.
e) Anta nå at man finner en viss sammenheng mellom botox-bruk og migrene, i bivariate analyser basert på den observasjonelle studien. Hva er en bivariat analyse? Hva er svakheten med resultater fra slike analyser?
f) Noen forskere ønsker så å sette i gang en RCT-studie av temaet. Hva er en RCT-studie?
g) Hvordan kan personer rekrutteres til å delta i denne studien?
h) Hvordan bør randomiseringen gjøres?
i) Kan en slik studie gjøres blindet? Dobbel-blindet?
j) Diskutér styrker og svakheter med dette studiedesignet.
k) Hvilke etiske vurderinger bør og må gjøres i slike studier?

På Pubmed.com finner man et sammendrag («Abstract») av en vitenskapelig artikkel om denne problemstillingen, skrevet av forskerne Mathew NT, Frishberg BM, Gawel M, Dimitrova R, Gibson J, Turkel C, BOTOX CDH Study Group. Artikkelen har tittel "Botulinum toxin type A (BOTOX) for the prophylactic treatment of chronic daily headache: a randomized, double-blind, placebo-controlled trial", og er publisert i *Headache*, 2005;45:293-307. Les abstractet på neste side, og se hvor mye du forstår. Mer spesifikt:

a) Hva er problemstillingen? Hvilket design er valgt?
b) Hva er forklaringsvariabel, og hva er responsvariabel?
c) Virker botox?

Abstract

OBJECTIVE: The objective of this study was to evaluate the safety and efficacy of botulinum toxin type A (BoNT-A; BOTOX, Allergan, Inc.) for the prophylactic treatment of chronic daily headache (CDH).

BACKGROUND: Several open-label and small controlled trials suggest that BoNT-A may be effective in the prophylactic treatment of headache.

DESIGN AND METHODS: This was an 11-month, randomized double-blind, placebo-controlled study of BoNT-A for the treatment of patients aged 18 to 65 years old with 16 or more headache days per 30 days conducted at 13 North American study centers. Following a 30-day screening period and a 30-day, single-blind, placebo-response period to identify placebo responders, eligible patients from both the placebo responder and placebo nonresponder groups were injected with BoNT-A or placebo every 90 days and assessed every 30 days for 9 months, a period encompassing three treatment cycles. The primary efficacy measure was the change from baseline in the frequency of headache-free days in a 30-day period for the placebo nonresponder group at day 180, the chosen efficacy time point. The secondary efficacy measure was the proportion of patients with a decrease from baseline of 50% or more in the frequency of headache days per 30-day period for the placebo nonresponder group at day 180. The change from baseline in the frequency of headaches (per 30-day period), the proportion of patients with a decrease from baseline of 50% or greater in the frequency of headaches per 30-day period, acute medication use, and adverse events were also assessed.

RESULTS: Of 571 patients assessed over the baseline period, 355 (mean age, 43.5 years; 300/355 [84.5%] female) were enrolled and randomized. At the end of the placebo run-in period, 279 patients (79%) were classified as placebo nonresponders and 76 patients (21%) as placebo responders. Subsequently, patients were randomized within each group to receive either BoNT-A or placebo. In the placebo nonresponder stratum, the mean number of headache-free days at baseline was 5.8 (+/-4.7) for BoNT-A- versus 5.5 (+/-4.7) for placebo-treated patients. At day 180, placebo nonresponders treated with BoNT-A had an improved mean change from baseline of 6.7 headache-free days per 30-day period compared to a mean change from baseline of 5.2 headache-free days for placebo-treated patients. The between-group difference of 1.5 headache-free days favored BoNT-A treatment, although the difference between the groups was not statistically significant. However, a statistically significant difference was observed at day 180 endpoint for the secondary efficacy measure. A significantly higher percentage of BoNT-A patients had a decrease from baseline of 50% or greater in the frequency of headache days per 30-day period at day 180 (32.7% vs. 15.0%, P=.027). Also, the mean change from baseline in the frequency of headaches per 30-day period at day 180 was -6.1 for BoNT-A patients vs. -3.1 for the placebo patients (P=.013). Only 4 of 173 BoNT-A patients (2.3%) discontinued the study due to adverse events. The majority of treatment-related adverse events were transient and mild to moderate in severity.

CONCLUSIONS: BoNT-A treatment resulted in patients having, on average, approximately seven more (1 week) headache-free days compared to baseline. Although at the primary time point (day 180) the BoNT-A treatment resulted in a 1.5 between-group difference compared to placebo, this difference was not statistically significant. The treatment met secondary efficacy outcome measures, including the percentage of patients experiencing a 50% or more decrease in the frequency of headache days, in addition to statistically significant reductions in headache frequency. BoNT-A was also well tolerated in patients with CDH.

Uke 36 (5/9-9/9):

Oppgavene er for øvelsene i uke 36 og blir gjennomgått i plenum 9. september. Det er åpne grupper i STK1000 så dere kan velge den (eller de) gruppa/gruppene som passer best.

- Fra boka: Oppgave 1.109, 1.110, 1.111, 1.118, 1.120, 1.128 og 1.129
- KFF-oppgave 6: Ta følgende IQ-test på nettet:
  http://www.funeducation.com → Free IQ test, og notér resultatet.
  Ta utgangspunkt i at IQ-scorer er konstruert for å være normalfordelt N(100,15).
  a) a) Hva er forventningsverdien i denne fordelingen? Og standardavviket?
  b) Tegn en N(100,15)-fordeling, og markér din egen IQ på figuren.
  c) Beregn din egen z-score, altså din standardiserte IQ-score.
  d) Tegn en N(0,1)-fordeling, og markér din egen z-score på figuren.
  e) Hva er forskjellen på disse to fordelingene/tegneingene?
  f) Hvor stor andel av befolkningen er dummere enn deg?
  g) Fyll ut nettskjemaet https://nettskjema.uio.no/answer/75299.html
- Fra boka: Oppgave 1.133 og 1.152
- Fra boka: 1.122, 1.124, 1.125

Uke 35 (29/8-2/9):

Oppgavene er for øvelsene i uke 35 og blir gjennomgått i plenum 2. september. Det er åpne grupper i STK1000 så dere kan velge den (eller de) gruppa/gruppene som passer best.

- Fra boka: Oppgave 1.5
- KFF-oppgave 1:
  Ta utgangspunkt i appelsindatafila, som har opplysninger om kjønn (M/K), sitrusfrukt-type (Appelsin/Klementin), appelsinvekt (i gram), skrelletid (i sekunder), antall båter, lange negler (ja/nei), og preferanser (Liker/Liker sånn passe/Liker ikke/Vet ikke), og repeter spørsmålene fra oppgave 1.5 for dette datasettet.
- KFF- oppgave 2: Hvordan kan man best tallfeste hvor god en forelesning er? En fotballkamp? En konsert? Er det i det hele tatt mulig? Og hvorfor er i så fall ditt forslag en god måte å gjøre det på?
- Fra boka: Oppgave 1.25 og 1.26
- KFF-oppgave 3: Lag (for hånd) et histogram for dataene i tabell 1.1, men la kategoriene være [70,80), [80,90), [90,100) etc, og ta utgangspunkt i dette når du svarer på spørsmålene i oppgave 1.29.
- Fra boka: Oppgave 1.54 og 1.55
- KFF-oppgave 4: Last inn appelsindatafila i R, og lag oppsummeringer av alle variablene.
- Fra boka: Oppgave 1.73 (R), 1.74 (R), 1.77, 1.90
- KFF-oppgave 5: Hvis du leser en tabell som oppsummerer data om ungdommer som ikke driver med regelmessig fysisk aktivitet, og ser at det står

a) at blodsukkerverdiene har et gjennomsnitt på 5.1, og et standardavvik på 1.0, hvordan ser du for deg at fordelingen til blodsukkerverdiene er da?
b) at insulinverdiene har en median på 30, og kvartiler på 23 og 47, hvordan ser du for deg at fordelingen til insulinverdiene er da?

Dataene som trengs til oppgavene kan lastes ned fra hjemmesiden til læreboka: http://www.macmillanlearning.com/Catalog/studentresources/ips8e#t_922171 (se under Data sets).