Regneøvelser STK1000 Høsten 2017

Dette dokumentet viser hvilke øvelsesoppgavene som passer til stoffet som gjennomgås i neste uke. Tidligere ukesoppgaverligger lenger ned. Dokumentet oppdateres ukentlig.

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:

Spørsmål om installasjon av R og Rstudio 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.

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 åpne dem fra RStudio. Dette blir gjennomgått i forelesningene og plenumsregningene de to første ukene i kurset.

Deretter kan du gjøre analyser, enten ved å skrive kommandoer rett inn i Rstudio, eller ved å åpne et R-script (som er en fil med kommandoer) og kjøre kommandoene derfra. Dette blir også gjennomgått i forelesningene og plenumsregningene de to første ukene i kurset.

Det kan være 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 kløne 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!

Appelsin-scriptet som du finner på nettsiden fungerer som et nybegynnerscript og vil ta deg gjennom mange av de grunnleggende analysene. 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.

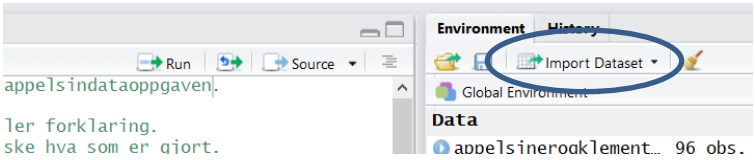
Uke 34-35(36):

Tema: (Innsamling av data,) deskriptiv (beskrivende) statistikk av én og én variabel. Boka, Ch 1.1-1.3

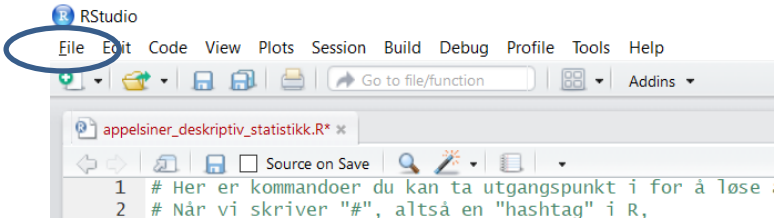
Forelesningene i disse ukene vil bli delt opp med en time forelesning og en time plenumsregning. Det er dessuten åpne grupper i STK1000 som dere kan bruke til å spørre om det dere lurer på.

- Fra boka: Oppgave 1.5
- KFF-oppgave 1:
  I Appelsindata2017-fila har du 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). 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 ned appelsindatafila fra kurssiden og lagre den på din egen datamaskin. Last ned og lagre det tilhørende R-scriptet.
  Åpne Rstudio, og åpne datafila ved å bruke Import Dataset:



  Åpne R-scriptet fra File – Open file – …



  Jobb deg gjennom scriptet og lag oppsummeringer av alle variablene, slik det er forklart. Denne oppgaven vil bli gjennomgått grundig i plenum på fredag 25/8.
- Fra boka: Oppgave 1.73 (R), 1.74 (R), 1.77, 1.90
  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).
- 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?

Tema: Deskriptiv (beskrivende) og eksplorativ (utforskende) statistikk for sammenhengen mellom to variabler.
Boka, Ch 2.1-2.3 + 2.6

- 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.

Tema: Innsamling av data, og design av forskningsprosjekter. Forskningsprosessen.
Boka, Ch 3.1-3.3 + 3.5

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.

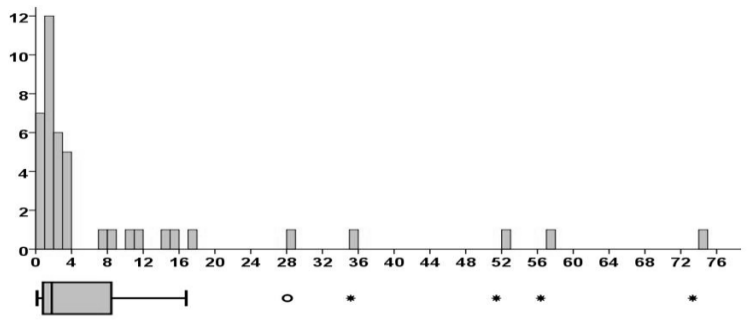
DESIGN AND METHODS: This was an 11-month, randomized double-blind, placebo-controlled study of BoNT-A for the treatment of patients aged 18 to 65 years old with 16 or more headache days per 30 days conducted at 13 North American study centers. Following a 30-day screening period and a 30-day, single-blind, 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.

Relevante oppgaver fra fjorårets midtveiseksamen:

Oppgave 1



Her ser du et histogram og et tilhørende boksplott for hvor mange uker det tar å strikke en genser. Hvilken påstand er feil:

- **A** Fordelingen er skjev.
- **B** Fordelingen har flere outliere eller ekstremverdier.
- **C** Gjennomsnittet og standardavviket er gode oppsummeringstall for denne fordelingen.
- **D** Fordelingen er entoppet eller unimodal.

Oppgave 2

Her ser du deskriptiv statistikk for de samme dataene, men i dager i stedet for uker. Hvilken påstand er riktig:

```
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max. sd
 2.5000   7.0000  14.0000  63.6100  58.0000 520.0000 116.4596
```

- **A** Strikkerne i undersøkelsen har strikket 95% av genserne i løpet av 7-58 dager.
- **B** Strikkerne i undersøkelsen har strikket 50% av genserne i løpet av 7-58 dager.
- **C** Strikkerne i undersøkelsen har brukt mer enn 63 dager på 50% av genserne.
- **D** Ingen av genserne er blitt ferdige på under en uke.

Oppgave 3

Hva er variansen for observasjonene (antall dager det tar å strikke en genser) i oppgave 2?

A 13563 **B** 10.8 **C** 51 **D** 517.5

Oppgave 4

Hvis du får oppgitt at varigheten på et sykehusopphold har et gjennomsnitt på 3 dager og et standardavvik på 4 dager, hvordan ser du for deg at fordelingen er da:

- **A** Den er symmetrisk om gjennomsnittet

B Det er umulig å vite
C Den er skjev med tung høyrehale
D Den er skjev med tung venstrehale

Oppgave 12

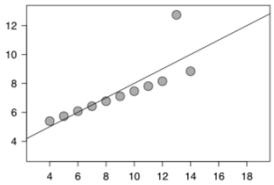
En universitetslærer ønsker å studere effekten av ulike læringsstrategier på ulike personlighetstyper. Han kartlegger studentene og deler dem inn i to grupper ved å trekke lodd om hvem som skal i hvilken gruppe. Han tilbyr konvensjonell undervisning til den ene halvparten av studentene, og en ny undervisningsstrategi til den andre. Dette er

A Et observasjonelt design
B Et dobbelt-blindt forsøk
C Et randomisert forsøk med kontrollgruppe
D Et matchet design

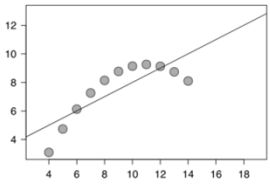
Oppgave 13

I hvilken av disse situasjonene er det fornuftig å bruke korrelasjonskoeffisienten som et oppsummeringstall for sammenhengen mellom de to variablene?

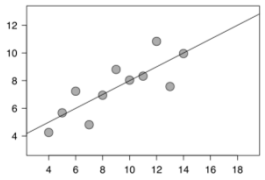
A



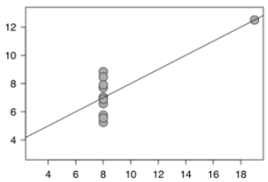
B



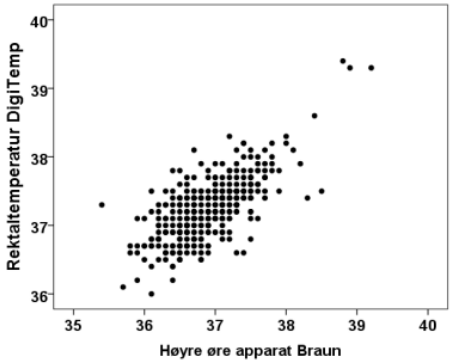
C



D



Oppgave 14



Den empiriske korrelasjonen mellom kroppstemperatur målt rektalt og kroppstemperatur målt i øret (vist i figuren over) er ett tallene oppgitt under. Angi hvilket.

A 0.98 **B** 0.2 **C** -0.65 **D** 0.7

Uke 37

Tema: Normalfordelingen.
Boka, Ch 1.4

Husk å bruke gruppene når det er noe dere står fast på. 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?
- Fra boka: Oppgave 1.133 og 1.152
- Fra boka: 1.122, 1.124, 1.125
- 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.

Melding fra foreleser

Kjære alle!

Nå har vi holdt på i noen uker, og dere er en herlig gjeng å forelese for! Tusen takk til alle som smiler hei, alle som stiller spørsmål, og alle som er behjelpelig når det spørres om ting på facebook-gruppa. Ut i fra de mange gode kommentarene og spørsmålene jeg har fått underveis, har jeg inntrykk av at mange av dere er på god vei inn i statistikkverdenen, og at hvis dere fortsetter på denne måten, kommer det til å gå riktig bra.

Jeg ble derfor ganske skremt på fredag da bare to av dere bekreftet at dere hadde prøvd/gjort ukesoppgavene, for ukesoppgavene er også pensum i kurset. Men det viser seg at vi har misforstått hverandre! Jeg er sjeleglad for at vi oppdaget det nå.

Saken er: Fordi vi har forelesning tidlig mandag og sent fredag, samt plenumsregning helt til slutt på fredag, velger jeg å gi ukesoppgaver til samme uke som det stoffet vi gjennomgår. Dere er derimot vant til at oppgavene gis til og gjennomgås uka etter at temaet har vært presentert på forelesning, og det var derfor dere ikke hadde gjort oppgavene. Tusen takk til deg som torde å si fra om dette! (Nå skjønner jeg hvorfor dere har sett så usikre ut når dere har spurt om jeg mener at oppgavene er til uke 37 når det står «Uke 37». Og dere skjønner sikkert min reaksjon når jeg har svart at «ja, når jeg skriver 37, så mener jeg 37».)

Fra nå av gjør vi det på min måte.

For eksempel: I uke 37 var det normalfordelingen som var hovedtema på mandag. Da ga jeg et sett med oppgaver om normalfordelingen til uke 37, slik at dere både kunne lese og jobbe med dette temaet denne uka, før jeg gjennomgikk oppgavene fredag 15/9, på slutten av uke 37.

I uke 38 er det oblig-innlevering, og dere må komme på riktig kjøl med oppgavene. Jeg gir derfor litt færre ukesoppgaver så dere skal rekke å ta igjen etterslepet fra uke 37.

Fra og med uke 39 gis det normal mengde oppgaver.

.

Uke 38

Tema: En liten introduksjon til inferens. Begrepene parameter og observator, bias og variabilitet. Variasjon i verdiene i utvalget vs variasjon i verdiene til observatoren.

Boka, Ch 3.4. OBS: Dette er en forsmak på bokas kapittel 5 og kapittel 6.1. Ta en titt i dette kapittelet også, og noter overskriftene. På fredag starter vi med sannsynlighetstemaet, Ch 4.

Innlevering av Oblig 1 torsdag 21/9

- 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

Relevante oppgaver fra fjorårets midtveiseksamen (gjennomgås fredag):

Oppgave 5

Hvilken påstand er feil:

- **A** I en normalfordeling er 50% av verdiene mellom $\mu - 1 \cdot \sigma$ og $\mu + 1 \cdot \sigma$
- **B** I en normalfordeling er 68% av verdiene mellom $\mu - 1 \cdot \sigma$ og $\mu + 1 \cdot \sigma$
- **C** I en normalfordeling er 90% av verdiene mellom $\mu - 1.645 \cdot \sigma$ og $\mu + 1.645 \cdot \sigma$
- **D** I en normalfordeling er 95% av verdiene mellom $\mu - 1.96 \cdot \sigma$ og $\mu + 1.96 \cdot \sigma$

Oppgave 6

En IQ-måling kommer fra en N(100,15)-fordeling, en normalfordeling med forventing 100 og standardavvik 15. Hvis du scorer 110, er du

- **A** Smartere enn 75% av befolkningen.
- **B** Smartere enn 68% av befolkningen.
- **C** Smartere enn 50% av befolkningen.
- **D** Smartere enn 90% av befolkningen.

Oppgave 7

Her ser du vekstkurver (som viser sammenhengen mellom vekt og alder), basert på Vekststudien i Bergen (BGS), SYSBARN-undersøkelsen, og WHOs internasjonale vekstkurver for gutter og jenter i alderen 0 - 5 år.

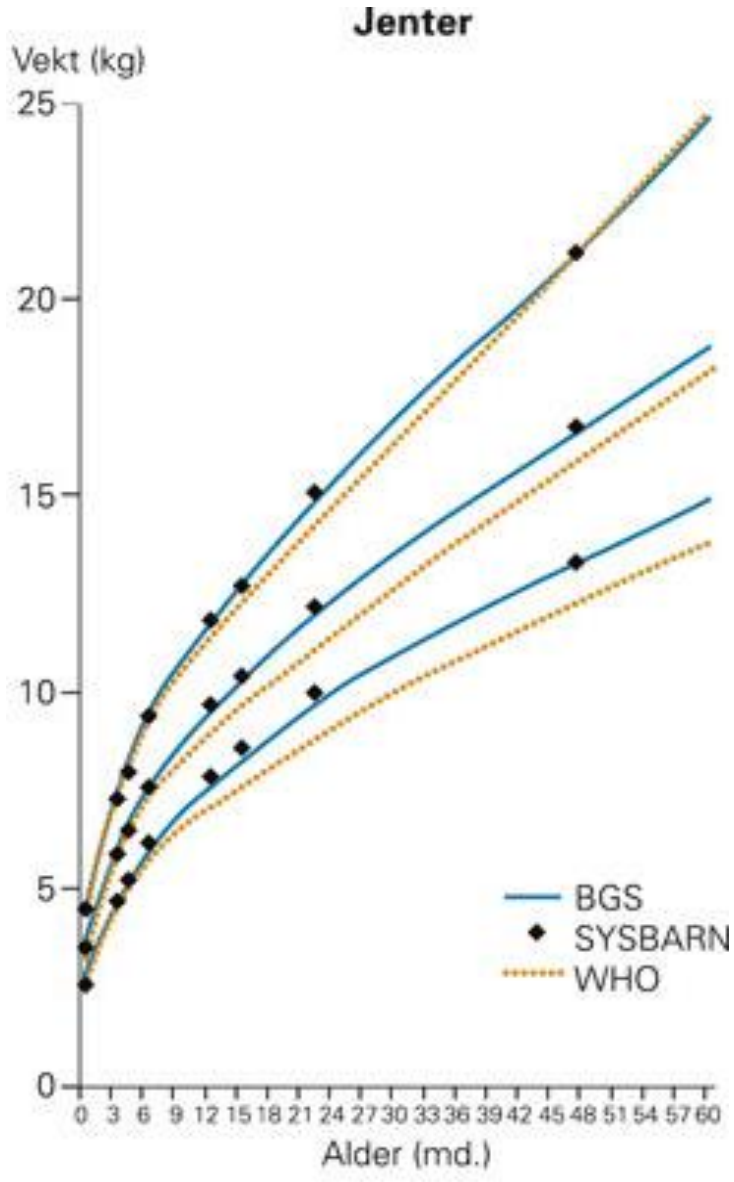
Linjene viser 2.5-, 50- og 97.5-prosentilene (percentilene).

Figuren er hentet fra Tidsskriftet for Den norske legeforening 2009;129: 281-6

Ta utgangspunkt i de nye, norske vekstkurvene fra BGS, altså de heltrukne linjene.

Hvilken påstand stemmer ikke:

- **A** 95% av 2-årige jenter veier mellom 10 og 15.3 kg
- **B** Over 50% av 2-årige jenter veier mer enn 12 kg



C Gjennomsnittsvekten for 2-åringer er ganske lik medianvekten for 2-åringer
D 95% konfidensintervall for forventet vekt for en 2-årig jente er [10,15.3]

Oppgave 8

I en studie av barns utviklingsnivå laget amerikanske forskere et spørreskjema som ga en totalscore. Denne scoren ble undersøkt i et utvalg av amerikanske barn, og beregnet slik at den fulgte en N(50,10)-fordeling. Dersom en score mindre enn 30 blir ansett som utviklingsavvik, hva er den tilhørende standardiserte scoren (z-scoren)?

A 2 **B** 20 **C** -2 **D** -20

Oppgave 11

Bias, eller på norsk, skjevhet, oppstår når studiedesignet gir en systematisk feil i resultatene. Hvilken påstand er feil:

Skjevhet kan skyldes

A At en eksperimentell studie gjøres uten en kontrollgruppe
B At en analyse i en observasjonell studie ikke har tatt hensyn til konfunderende (i boka: Lurking) variabler
C At man ikke har randomisert når man har delt deltakerne i et eksperiment inn i grupper
D At utvalget ikke er stort nok

Uke 39 + 40

Tema: Sannsynlighet, stokastiske variabler, sannsynlighetsfordelinger, forventning og varians. Regneregler for sannsynlighet, forventning og varians. (Sannsynlighetstemaet ble påbegynt på forelesning i uke 38)

Boka, Ch 4

Fredag 29/9 (ikke 6/10 som tidligere annonsert)
Den beste forberedelsen til midtveiseksamen: Midtveisquiz i RF-kjelleren etter forelesning! Matematisk institutt spanderer mat! Møt opp og test hva du kan. Premie til beste lag.

- Fra boka: 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.
- 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

Tema: Fordelingen til et gjennomsnitt, og fordelingen til en andel, motivert av hhv sentralgrenseteoremet og normalfordelingsapproksimasjonen til binomisk fordeling

Boka, Ch 5

- Fra boka:Oppgave 5.14, 5.20, 5.21, 5.23.
- Midtveiseksamen H2015: Hele unntatt oppgave 13 og 14. Dette oppgavesettet ligger under ukesoppgavene for dette semesteret.
- Fra boka: Oppgave 5.14, 5.20, 5.21, 5.23, 5.25, 5.28, 5.50, 5.51, 5.53, 5.60, 5.62.

Relevante oppgaver fra fjorårets midtveiseksamen:

Oppgave 9

Hvis vi samler nok data, altså at utvalgsstørrelsen n blir stor nok, så vil tre av disse påstandene være sanne. Hvilken av dem skjer ikke?

A Fordelingen til gjennomsnittet blir normalfordelt.
B Histogrammet over dataene vil bli normalfordelt.
C Gjennomsnittet vil ligne mer og mer på forventningsverdien i populasjonen.
D Histogrammet over dataene vil ligne på populasjonsfordelingen.

Oppgave 10

Hvis vi har to tilfeldige utvalg, ett på n=25, som gir et gjennomsnitt $\bar{x}_{25}$ og et standardavvik sd_{25}, og ett på n=100, som gir et gjennomsnitt $\bar{x}_{100}$ og et standardavvik sd_{100}, og histogrammene for begge utvalgene ser rimelig symmetriske ut, så vil tre av påstandene være riktige. Hvilken påstand er feil?

A Gjennomsnittene i de to utvalgene er ganske like, altså i samme størrelsesorden.
B Standardavvikene i de to utvalgene er ganske like, altså i samme størrelsesorden.
C Percentilene i de to utvalgene er like, altså i samme størrelsesorden.
D Estimeringsusikkerheten (standardavvikene) til de to gjennomsnittene er ganske like, altså i samme størrelsesorden.

Oppgave 15

En diskret tilfeldig variabel X antar verdiene 4 til 10, med sannsynligheter gitt i følgende tabell

x	4	5	6	7	8	9	10	Sum
P(X=x)	0.02	0.09	0.24	0.31	0.26	0.06	0.02	1

Da er forventningen til X, μ_X, lik

A 7.0 **B** 6.5 **C** 7.5 **D** 6.8

Oppgave 16

For X gitt i forrige oppgave blir standardavviket lik

A 1.5 **B** 2.0 **C** 1.0 **D** 1.2

Oppgave 17

Variabelen X i de to forrige oppgavene viser fordelingen av skostørrelse for kvinner, målt i amerikanske skonummer. La variabelen Y være skostørrelse for kvinner, målt i europeiske skonummer. Da er Y = X + 30.5. Hvilken påstand er feil:

- **A** E(Y) = μ_Y = 37.5
- **B** X og Y er korrelerte.
- **C** Standardavviket til Y er lik standardavviket til X
- **D** Standardavviket til Y er større enn standardavviket til X

Oppgave 18

I en sannsynlighetsmodell der S er utfallsrommet, A og B er disjunkte begivenheter, og P(A) > 0 og P(B) > 0, er en av påstandene feil:

- **A** P(S) = 1
- **B** $P(A^C) = 1 - P(A)$
- **C** P(A eller B) = P(A) + P(B)
- **D** P(A og B) = P(A)·P(B)

Oppgave 19

I to blindtester av sjokolade og cola, definerer vi begivenhetene A og B som
A: Student gjetter riktig sjokolade, og
B: Student gjetter riktig cola.
P(A) = 0.7, P(B) = 0.6, og P(A og B) = 0.4 Hvilket utsagn er feil:

- **A** $P(A^C) = 0.3$
- **B** $P(A^C \text{ og } B^C) = P(A^C) \cdot P(B^C)$
- **C** $P(A^C \text{ og } B^C) = 1 - P(A \text{ eller } B)$
- **D** $P(B^C) = 0.4$

Oppgave 20

Type 2-diabetes er et økende problem på verdensbasis, og det diskuteres hvilke tester som bør brukes for å avdekke sykdomstilstanden. En blodprøve der man måler det såkalte HbA1c, og gir positivt svar på testen dersom HbA1c $\geq$ 7%, ble ikke anbefalt som diagnosekriterium av WHO i 2006. I denne oppgaven skal du regne på sannsynligheter knyttet til dette. Du kan anta at 8.5% av jordas befolkning har type 2-diabetes (tall fra 2014), at andelen syke som får positiv test er 0.78, og andelen friske som får positiv test, er 0.15. Hva er sannsynligheten for at du er syk gitt at du har fått en positiv HbA1c-test?

A 0.33 **B** 0.22 **C** 0.07 **D** 0.85

Uke 41

Ingen undervisning eller oppgaver, men midtveiseksamen onsdag 11/9

Uke 42-43

Tema: Konfidensintervall, CLT og fordelingen til $\bar{x}$
(Oppgaver gitt uka før midtveiseksamen) Fra boka: 5.14, 5.20, 5.21, 5.23, 5.25, 5.28

Tema: Inferens, Konfidensintervall (KI) for μ
Fra boka: Oppgave 6.19, 6.20, 6.27, 6.28, 6.30, 6.33

Tema: Binomisk fordeling og fordelingen til $\hat{p}$
(Oppgaver gitt uka før midtveiseksamen) Fra boka: 5.50, 5.51, 5.53, 5.60, 5.62

Tema: Inferens, Konfidensintervall (KI) for p
Fra boka: 8.13 og 8.16

Uke 43-44

Tema: Hypotesetesting. Vi fokuserer 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.

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).

Ett-utvalgs-tester:

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

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)

Uke 45-47

Tema 1: Analyse av sammenheng mellom to variabler (bivariate analyser), der korrelasjonsanalyse (kontinuerlig variabel mot kontinuerlig variabel, Ch 2) og to-utvalgs t-test (kontinuerlig variabel mot kategorisk variabel med to kategorier, Ch 7) er to av de alternative situasjonene vi ofte kommer borti. En oversikt over bivariate analyser blir gitt på forelesningen på mandag 6/11.

Tema 2: Regresjonsanalyse. Både korrelsajonssitasjonen og toutvalgs-t-testsituasjonen kan også analyseres med regresjonsanalyse, Ch 10. Men regresjonsanalyse er et så stort og generelt rammeverk at det også åppner for å inkludere flere variabler i analysen: Ch 11.

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 tidligere i kurset), 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.

Vi skal også gå grundig gjennom eksamensoppgavene fra i fjor, både ordinær eksamen og konteeksamen. Begge ligger som pdf-er under ukesoppgavene.

Andre aktuelle eksamensoppgaver: Oppgave 3 STK1000 V05, Oppgave 3 STK1000 H06, Oppgave 2 STK1000 H14, Oppgave 1,2,3 desember 2008, Oppgave 3 desember 2011.