1. obligatoriske oppgave i STK1000

Høst 2010

Oppgavesettet består av fire oppgaver. For å løse oppgavene trenger du hjelp av MINITAB eller annen statistisk programvare. I forbindelse med bruk av MINITAB kan du ha nytte av notatet *Starthjelp i MINITAB* (kalt "innføringsheftet" nedenfor). Det er tilgjengelig på hjemmesiden til kurset. Datasett brukt i oppgaven vil være direkte tilgjengelig fra hjemmesiden til kurset (under "Informasjon om obligatoriske oppgaver")

I den skriftlige besvarelsen av oppgavene skal du kort forklare hvordan de enkelte punktene er løst. Oppgaven skal i utgangspunktet skrives med et tekstbehandlingsprogram (oppgave 4 kan godt leveres håndskrevet). Hvis du velger å skrive for hånd bør du begrunne hvorfor. Der du bruker MINITAB, må relevante utskrifter og plott legges ved eller limes inn i besvarelsen. Instruksjoner for utskrift fra MINITAB finner du i avsnitt 10 i innføringsheftet.

Obligen skal leveres med en egen forside som du finner under
`http://www.math.uio.no/academics/obligforsideMI.pdf`.
Det er lov å samarbeide og å bruke hjelpemidler. Den innleverte besvarelsen skal imidlertid skrives av deg og gjenspeile din forståelse av stoffet. Er vi i tvil om at du virkelig har forstått det du har levert inn, kan vi be deg om en muntlig redegjørelse.

Besvarelsen leveres på instituttkontoret ved Matematisk Institutt i 7. etasje, Niels Henrik Abels hus (Matematikkbygningen).

Frist for innlevering er **torsdag 30. september 2010 kl 14.30**.

Oppgave 1

De ulike stater i USA varierer svært mhp værtyper og de konsekvenser ekstremt vær kan ha. Tabell 1.5 i boka viser gjennomsnittlige eiendomsskader forårsaket av tornadoer per år over en periode fra 1950 til 1999 i hver av de 50 statene samt Puerto Rico (konvertert til 1999 dollar). Dataene er gitt på fil i minitab-format:
`http://www.uio.no/studier/emner/matnat/math/STK1000/h10/ta01_005.MTP`
og tekst-format
`http://www.uio.no/studier/emner/matnat/math/STK1000/h10/ta01_005.txt`

(*a*) Identifiser de stater med de 5 høyeste skadeverdier og de 5 laveste (sortering etter skadeverdier kan være nyttig her, **Data -> sort**)

(*b*) Del verdiområdet opp i intervaller [0,10], [10,20] osv og tell opp antall observasjoner i hvert delintervall. Tegn for hånd et histogram for de 51 observasjonene. Kommentér histogrammets form.

(*c*) Lag et histogram over de 51 observasjonene ved hjelp av MINITAB (kommando: **Graph -> Histogram**; jfr. avsnitt 8.3 i innføringsheftet). Sammenlign med histogrammet du lagde i (*b*). Er det noen observasjoner som skiller seg ut?

(*d*) Hvilke oppsummerende mål bør brukes for å beskrive en fordeling som den du ser i (*a*) og (*b*)? Begrunn svaret. Beregn disse ved hjelp av MINITAB (kommando: **Stat -> Basic Statistics -> Display Descriptive Statistics**; jfr. avsnitt 8.1 i innføringsheftet - plukk ut det du trenger fra utskriften).

(*e*) Bruk MINITAB til å lage et boksplot over dataene. Forklar hva plottet viser.

Oppgave 2

I denne oppgaven skal vi se på data som er samlet inn under et innføringskurs i statistikk ved et amerikansk universitet. Studentene i kurset gjennomførte et enkelt forsøk. Hver student noterte sin høyde og vekt og målte pulsen sin (under hvile). Så kastet hver av studentene en mynt. De som fikk krone løp på stedet i ett minutt, mens de som fikk mynt ble sittende stille i ett minutt. Så målte alle pulsen en gang til (for de som satt stille, er dette en måling til av pulsen under hvile). Dataene for de 92 studentene finnes som en Minitab-fil under
`http://www.uio.no/studier/emner/matnat/math/STK1000/h10/oblig1data.MTW`
og som en vanlig tekstfil under
`http://www.math.uio.no/avdc/kurs/STK1000/data/PULSDATA.TXT`,
som du må kopiere inn i et arbeidsark i MINITAB. På datafilen er det en linje for hver av de 92 studentene, der variablene i kolonnene har følgende betydning:

* ***Pulse1:*** Første pulsmåling (antall slag per minutt)

* ***Pulse2:*** Andre pulsmåling (antall slag per minutt)

* ***Ran:*** 1=løp på stedet; 2=satt stille

* ***Sex:*** 1=mann; 2=kvinne

* ***Height:*** høyde i inches (1 inch = 2,54 cm)

* ***Weight:*** vekt i pounds (1 pound =0,454 kg)

Når du har lastet dataene inn i MINITAB, registrerer du dine egne verdier for variablene Sex, Height og Weight i linje 93 i arbeidsarket (husk å regne om til hhv. inches og pounds). Mål pulsen din under hvile og registrer den som *Pulse1*. Kast så et kronestykke. Hvis det viser krone, løper du på stedet i ett minutt. Hvis ikke sitter du stille ett minutt. Så måler du pulsen din igjen og registrer den som *Pulse2* i linje 93 i arbeidsarket. Skriv også inn din verdi for *Ran* (1 hvis du løp, 2 hvis du satt stille).

Med dette modifiserte datasettet:

(*a*) Lag histogram for variabelen *Pulse1* og merk av din egen verdi.

(*b*) Beregn enkle oppsummerende mål for den samme variabelen. Hvordan ligger din egen verdi i dette bildet?

(*c*) Lag et kryssplott med *Pulse1* på x-aksen og *Pulse2* på y-aksen, med forskjellige symboler for de som løp og de som ikke løp. Forklar hva plottet viser. Lag andre grafiske fremstillinger som viser effekten av aktivitet på *Pulse2*.

(*d*) Gjennomfør en regresjonsanalyse med vekt som responsvariabel og høyde som forklaringsvariabel. Forklar hva resultatene av regresjonsanalysen forteller deg. (Kommandoer: **Stat -> Regression -> Regression** og **Stat -> Regression -> Fitted Line Plot**.)

(*e*) Forklar hva **R-Sq** (r-kvadrert) i utskriften betyr.

Oppgave 3

Denne oppgaven er essensielt oppgave 2.106 på side 186 i læreboka. De to datasettene i oppgave 2.106 (ett for kvinner og ett for menn) er fremkommet ved at man har latt et antall eliteløpere løpe på tredemølle ved bestemte hastigheter ('Speed', som her betraktes som forklaringsvariabel), og målt stegfrekvensen (antall steg per sekund) ved hver hastighet. Responsvariabel er gjennomsnittlig stegfrekvens for henholdsvis kvinnelige og mannlige løpere ('Stride rate'). En enkel måte å organisere dataene på i MINITAB er å lage seks kolonner: speed, stride rate kvinner, stride rate menn, en dobbelt så lang kolonne der speed ligger to ganger etter hverandre og en dobbelt så lang kolonne der stride rate for kvinner og stride rate for menn ligger etter hverandre. Til slutt lager du en kolonne med en kategorisk variabel som angir om individene i den sistnevnte kolonnen er kvinner eller menn. De tre første kolonnene må du skrive inn manuelt. De to neste kan du lage manuelt eller ved å bruke **Data -> Stack -> Columns**. Den siste kolonnen må du lage manuelt.

(*a*) Plott først dataene for hastighet og stegfrekvens. Her lager du tre figurer: Først en der dataene plottes i samme figur, men med ulike plottesymboler for menn og kvinner, dernest en der det legges inn en felles regresjonslinje og til slutt en der det legges inn separate regresjonslinjer. Plottene kan du lage med kommandoen **Graph -> Scatterplot**. På bildet som kommer fram kan du spesifisere de tre typene av plott som er beskrevet ovenfor ved **With Groups**, **With Regression** og **With Groups and Regression** henholdsvis. Du trenger den sjette kolonnen, som identifiserer kvinner og menn, for å lage det første og siste av plottene.

(*b*) Anta nå at du fikk dataene uten identifikasjon av kjønn. Beregn koeffisientene for minste kvadraters linje for alle dataene. Bruk de to kolonnene der dataene for kvinner og dataene for menn er slått sammen.

(*c*) Lag plott av residualene fra linjen i punkt (*b*) mot rekkefølgen av observasjonene og mot hastighet. Forklar hvordan det vises i plottet at dataene kommer fra to forskjellige grupper. Kommandoene er her **Stat -> Regression -> Regression**. Klikk på **Graphs** og be om å få plottet (i) residualene mot rekkefølgen av observasjonene og (ii) residualene mot hastighet.

(*d*) Beregn så regresjonslinjer separat for hvert kjønn. Hva blir korrelasjonen mellom hastighet og stegfrekvens i hvert av tilfellene? Hva sier disse korrelasjonene om regresjonslinjene?

(*e*) Lag plott av residualene fra hver av de to linjene fra punkt (*d*) mot hastighet. Kommenter plottene!

Oppgave 4

Summer og gjennomsnitt inngår i statistikken i ulike sammenhenger. I denne oppgaven vil vi se litt nærmere på egenskaper til summer og gjennomsnittet. I det etterfølgende vil vi bruke at summen av x_i kan skrives som $\sum_{i=1}^{n} x_i$ og at gjennomsnittet $\bar{x}$ kan skrives som $\frac{1}{n}\sum_{i=1}^{n} x_i$.

(*a*) Anta at summen av $(x_i - a)$ er 0, dvs $\sum_{i=1}^{n}(x_i - a) = 0$. Hva er a?

(*b*) Vis at summen av $(x_i - \bar{x})$ er 0, dvs $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.

(*c*) Spredningsmålet varians, s^2, er definert ved $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Diskuter hvorfor et spredningsmål gitt ved $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})$ blir meningsløst.

Lykke til!