EKSTRAOPPGAVE I STK1000 TIL KAPITTEL 3

1. Minitaboppgave til Kapittel 3

Vi skal bare gjennom en eneste Minitaboppgave for kapittel 3, men den er til gjengjeld litt omfattende. Siden det er mange spørsmål og ting å undersøke minner jeg om at Minitaboppgavene skal besvares ved hjelp av en tekstbehandler som Word, og at det ikke er nok å bare gå igjennom skrittene for seg selv og tenke litt på hvert av spørsmålene. Skriv derfor en ryddig rapport over resultatene du kommer frem til.

Oppgave 1.1. La oss si vi er interessert i medianinntekten til familier i byen datasettet "families.mtw" kommer fra, som vi også jobbet med i kursets første oppgavehefte. I denne oppgaven skal vi tenke oss at dette datasettet faktisk er en oversikt over *alle 43886 familier i byen* (en såkalt *census* over familiene i byen), og at datasettet er helt riktig. Du er ansatt statistiker i et underbudsjettert statlig organ, og har fått i oppdrag å undersøke hva en typisk inntekt er, og hvor stor forskjellen på fattig og rik er i byen din. Du har desverre svært begrensede midler til rådighet, og har maksimalt råd til å gjøre en undersøkelse med 200 personer.

i. Men heldigvis *har* du til rådighet en komplett liste over alle familier i byen. Forklar hvorfor en SRS er fornuftig angrepsmåte for problemstillingen din, og hvorfor man burde ta med så mange personer man har ressurser til.

ii. I løpet av en uke har du fullført en SRS med $n = 200$ – gratulerer! Du beregner så medianen (og får 33398) og IQR (og får 32950) til utvalget. Dette er altså din beregnede *statistikk*, som prøver å nå *parameteren* du er ute etter – nemlig medianen og IQR til hele befolkningen. Man sier dessuten ofte at en statistikk *estimerer* en parameter. Altså er utvalgsmedianen et estimat for befolkningsmedianen, og utvalgs-IQR estimerer befolknings-IQR.

   For å skille disse to størrelsene klart og tydelig kaller vi medianen du beregnet for utvalget for *utvalgsmedian*, mens medianen til hele befolkningen kaller vi *befolkningsmedian* og tilsvarende skal vi snakke om utvalgs-IQR og befolknings-IQR.

   Det kan hende du synes at talloppsummeringer for utvalget ditt er interessant i seg selv, men det er egentlig ikke dem du er ute etter. Planen for undersøkelsen var å finne en tilnærming til *befolkningsmedianen* og *befolknings-IQR*, og riktignok er det fint og flott at statistisk teori garanterer at estimatene dine er et sted i nærheten av utvalgsmedianen og utvalgs-IQR, men siden du har tenkt til å bruke disse tilnærmingene som grunnlag for videre avgjørelser er du også veldig interessert i *et feilanslag*!

   Med mer statistisk teori kunne man gitt et usikkerhetsanslag basert på sannsynlighetsteori. Akkurat median og IQR viser seg å ha en komplisert sannsynlighetsstruktur, og vi skal ikke lære hvordan de oppfører seg i STK1000. Men vi skal lære tilsvarende teknikker for andre talloppsummeringer. Spesifikt skal

Date: 10. september 2009.

vi løse dette problemet for gjennomsnittet senere i kurset. Men for vårt befolkningsdatasett er vi ikke interessert i gjennomsnittsinntekt (siden inntektsfordelingen er ganske høyreskjev), og vil utelukkende jobbe med medianinntekt.

En opplagt måte å se hvor godt utvalgsmedianen og utvalgs-IQR tilnærmer befolkningsmedianen og befolknings-IQR er selvsagt å sammenlikne dem med fasiten. I vår tiltenkte fantasiverden, hvor man ikke har denne kunnskapen, er dette selvsagt en umulighet, men for oss som har tilgang til hele "families.mtw"-datasettet kan vi gjøre nettopp dette.

Last derfor inn hele datasettet i Minitab og beregn median og IQR for hele befolkningen. Kommentér hvor forskjellige statistikkene er i forhold til parameterene. Er de i samme størrelsesorden? Og ville du forventet at resultatet var så bra/så dårlig med bare 200 i utvalgsstørrelse?

iii. Vi har ikke bare fasitsvaret på hva befolkningsmedianen og befolknings-IQR er – *vi har hele datasettet!* Dette lar oss *simulere* hele prosessen involvert i en SRS. Hvis vi gjør dette mange ganger får vi inntrykk av hvordan resultatet av en SRS pleier å være.

Planen er å be Minitab om å trekke 200 tilfeldig valgte personer, og lagre resultatet i en ny regnearkskolonne. Hvis vi beregner median og IQR for denne nye kolonnen (som vi tenker på som utvalget vårt) kan vi se hvor gode tilnærminger man typisk får til befolkningsmedian og IQR.

Trekk derfor et tilfeldig utvalg med 200 individer, og beregn dets median og IQR. Minitab trekker individene via

Calc $\mapsto$ **Random Data** $\mapsto$ **Sample from columns** .

Skriv (igjen) inn 200 i "number of rows to sample", velg "income" i "from columns" og C7 (en tom kolonne) i "store samples in". Gi dessuten C7 et fornuftig navn, som "Sample 1". Beregn deretter median og IQR. Sammenlikne disse med befolkningsmedianen og befolknings-IQR.

iv. Median og IQR er bare to aspekter ved et datasett. De aller fleste aspektene man er interessert i kan sees ut i fra histogrammet til datasettet. Man er derfor interessert i hvordan utvalgets histogram ser ut i forhold til befolkningens histogram. La Minitab tegne opp begge, og noter deres likheter og forskjeller. Virker det som at alle aspekter av datasettet er like godt tilnærmet fra utvalget? Kommentér spesifikt skalaen til de to histogrammene, og knytt dette opp i mot hvordan den største inntekten i befolkningen tilnærmes ved den største inntekten i utvalget.

v. Siden SRS tar et *tilfeldig* utvalg av befolkningen vil konklusjonene på de to siste spørsmålene være litt forskjellig fra gang til gang. Hvor mye varierer utvalget? For å undersøke dette, simuler tre nye SRS-er og lagre resultatet i nye kolonner du kaller "Sample 2", "Sample 3" og "Sample 4". Det vil si, gå på

Calc $\mapsto$ **Random Data** $\mapsto$ **Sample from columns** .

Skriv inn 200 i "number of rows to sample", velg "income" i "from columns" og hver gang fyll inn C8, C9 og så C10 i "store samples in". Ser de to siste punktenes konklusjoner ut til å holde for alle utvalgene? Og hvor mye varierer medianen fra utvalg til utvalg? Varierer alle de vanlige talloppsummeringene like mye? F.eks, hvordan oppfører maksimum seg?

vi. Vi skal her gå enda videre, og bruke simulering til å se på *utvalgsfordelingen* til medianen og IQR for dette datasettet. I denne sammenheng er det to punkter som er viktige å tenke over.

(I) Vi simulerer fra *dette* datasettet, og uten å kunne sannsynlighetsteori kan vi ikke være sikre på at eventuelle konklusjoner vi kommer til også gjelder for andre datasett.

(II) Vi skal finne tilnærminger til utvalgsfordelingen til medianen og IQR, noe som vi ikke ville klart med den teorien vi skal igjennom i STK1000. Vi skal lære tilnærminger til noen vanlig brukte statistikkers utvalgsfordeling, men ikke mange. Men simuleringsprosedyren vår blir ikke vanskeligere eller lettere i forhold til om sannsynlighetsstrukturen til medianen er mer komplisert enn gjennomsnittet. Minitab bryr seg ikke om dette, og beregner medianen til utvalgene vi simulerer like lett som i andre kontekster. Simulering gir ofte svar på problemer man ikke klarer med å løse via teori. Men her kommer også forrige punkt inn: Uten teori kan vi ikke generalisere konklusjonene vi får fra simuleringene til andre tilfeller.

La oss repetere hva utvalgsfordelingen til en statistikk funnet ved hjelp av en SRS er. La oss si at man finner mange, mange SRS-er og noterer seg resultatet av en talloppsummering som medianen hver gang. Man lager så et histogram av alle de resulterende medianene, og dette er utvalgsfordelingen. Takket være Minitab skal vi nå følge beskrivelsen til hva utvalgsfordelingen er for å se på hvert skritt av gangen, og til slutt ende opp med en tilnærming av utvalgsfordelingen til medianen og IQR for dette datasettet.

Vi har altså lyst til å lage mange kolonner med simuleringer. Jeg regner med at dere ikke har lyst til å gjenta "sample for columns"-skrittene mange ganger til, så derfor har jeg lagt ut datasettet "sampled_families.mtw" på hjemmesiden, hvor jeg har gjort dette skrittet 30 ganger. Ikke lukk Minitabvinduet du har jobbet med, men åpne opp denne filen. Copy-Paste kolonnene du har simulert inn etter alle kolonnene med simuleringer i "sampled_families.mtw", så du i alt har 34 simuleringer totalt.

Vi ønsker så å finne medianene til alle utvalgene. Man kan selvsagt gjøre dette manuelt, men Minitab kan heldigvis gjøre dette automatisk.

$$\textbf{Stat} \mapsto \textbf{Basic Statistics} \mapsto \textbf{Store basic statistics},$$

i motsetning til den vanlig brukte "Display basic statistics". Sett inn alle kolonnene med simuleringer i "variables", og gå inn på "Statistics". Velg vekk "mean" og "N nonmissing" (som vi for øyeblikket ikke er interessert i), og huk på "Median" og "Interquartile range". Trykk OK.

Nå har Minitab lagt medianene og IQRene til hver av kolonnene inn i regnearket. Dessverre har den lagd en kolonne for hver beregnet median og IQR, mens histogramfunksjonen bare kan ta inn data i en kolonne. Vi må derfor gjøre disse verdiene om til to kolonner, noe vi kan gjøre manuelt men også via Minitab. For å bruke Minitab, gå på

$$\textbf{Data} \mapsto \textbf{Transpose Columns},$$

velg alle kolonnene med medianene i "transpose the following columns", og marker "after last column in use" i "store transpose". Gjør det samme med IQR-kolonnene. Lag endelig nå histogrammer av alle de 34 medianene og IQRene.

vii. Vi har nå en tilnærming av utvalgsfordelingen til median og IQR beregnet via SRS i befolkningsdatasettet. Ved hjelp av de to histogrammene fra forrige punkt, er disse statistikkene forventningsrett, og hvor stor variabilitet ser de

ut til å ha? Beregn talloppsummeringer for senter og spredning for å tallfeste dette mer presist.