OPPGAVEHEFTE I STK1000 TIL KAPITTEL 5 OG 6

1. Regneoppgaver til kapittel 5 og 6

Oppgave 1. Mange som kommer til STK1000 med dårlige erfaringer fra tidligere mattefag er livredd ulikheter, og selv om man har syns matte har vært greit, er det mange som ikke husker hvordan regning med ulikheter fungerer. Ulikheter er kjempeviktig å ha god peiling på nå som vi skal jobbe med konfidensintervaller, og det er derfor viktig at alle klarer denne oppgaven.

i. Vi skal se hvorfor man *må snu* ulikhetstegnet i ulikheter når man ganger med -1 på begge sider. Altså ønsker vi å vise at

$$a \leq b$$

er det samme som

$$-a \geq -b.$$

Husk at det er lov til å plusse på eller trekke fra samme tall på begge sider av en ulikhet (hvorfor er dette sant?). *Ved å bare bruke dette* vis at $a \leq b$ virkelig er det samme som $-a \geq -b$.

ii. En annen måte å tenke på forrige punkt er at man kan tegne opp et hvert tall på *tallinjen*. Når man ganger et tall med -1, speiler man plasseringen til tallet om null. Tegn først opp tallinjen, marker av et tall, og dens minus og overbevis deg selv om at dette stemmer.

Hvorfor viser også dette at $a \leq b$ er det samme som $-a \geq -b$? (Hint: Tegn opp a, b og $-a$, $-b$!) (Hint: Hvis først hva som skjer hvis a og b begge er positive eller begge er negative, og så ta hensyn til hva som skjer hvis de er på forskjellig side av null på tallinjen.)

iii. De to siste punktene viser at $a \leq b$ er det samme som $-a \geq -b$. Anta vi også vet at et tall c er slik at $b \leq c$ er det samme som $-b \geq -c$. Altså er

$$a \leq b \leq c.$$

Siden vi fra $a \leq b$ vet at $-a \geq -b$, og fra $b \leq c$ vet at $-b \geq -c$ må altså også

$$-a \geq -b \geq -c.$$

En annen måte å skrive dette på er at

$$-c \leq -b \leq -a.$$

På forelesningen gav jeg et 95% usikkerhetsanslag for μ med utgangspunkt i $68, 95, 99.7$-prosentsreglen. Da vi regnet oss frem til dette, hadde vi overgangen

$$P\left(-\bar{X}_n - 2\frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{X}_n + 2\frac{\sigma}{\sqrt{n}}\right) = P\left(\bar{X}_n - 2\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + 2\frac{\sigma}{\sqrt{n}}\right) \quad (1)$$

Gå igjennom denne utregningen en gang til ved hjelp av forklaringen over, og overbevis deg selv om at dette virkelig stemmer ved å bruke så mange skritt at du syns hvert skritt *er enkelt*.

Date: 13. november 2009.

iv. La oss så repetere skrittene i utregningen som ledet til

$$95\% \approx P\left(\bar{X}_n - 2\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + 2\frac{\sigma}{\sqrt{n}}\right). \tag{2}$$

Dette sier altså at det er ca 95% sannsynlig at tallet μ (som er et tall man ikke kjenner, *men som er fastsatt*, f.eks kan det være tallet 1.5) er i intervallet

$$\left[\bar{X}_n - 2\frac{\sigma}{\sqrt{n}}, \bar{X}_n + 2\frac{\sigma}{\sqrt{n}}\right].$$

Her er altså *grensene på intervallet tilfeldig* siden de er beregnet fra observasjonene $X_1, X_2, \ldots, X_n$ som er tilfeldige variable, mens μ er ikke tilfeldig og er bare et tall.

Vi startet med at vi hadde observasjoner $X_1, X_2, \ldots, X_n$ som alle er uavhengige og har samme fordeling. Alle X'ene har *samme* forventning (la oss kalle den μ) og standardavvik (la oss kalle den σ), siden disse er definert ut i fra fordelingen til X'ene.

Vi vet så – enten nøyaktig, hvis observasjonene er normalfordelt eller tilnærmingsvis ellers – at

$$\bar{X}_n \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

Da sier $68, 95, 99.7$-prosentsreglen at det er rundt 95% sannsynlig at $\bar{X}_n$ er mellom to standardavvik (dvs $2\sigma/\sqrt{n}$) av sin forventning (dvs μ). Altså

$$95\% \approx P\left(\mu - 2\frac{\sigma}{\sqrt{n}} \leq \bar{X}_n \leq \mu + 2\frac{\sigma}{\sqrt{n}}\right).$$

Vi skal så regne oss frem til konklusjonen vår, nemlig at

$$P\left(\mu - 2\frac{\sigma}{\sqrt{n}} \leq \bar{X}_n \leq \mu + 2\frac{\sigma}{\sqrt{n}}\right) = P\left(\bar{X}_n - 2\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + 2\frac{\sigma}{\sqrt{n}}\right).$$

Altså har vi fått vekk $\bar{X}_n$ fra midten ("ryddet den bort til siden") og heller fått μ alene i midten. Gå frem via følgende skritt:

a. Rydd vekk $\bar{X}_n$ fra midten ved å trekke fra $\bar{X}_n$ i alle tre ulikheter.
b. Trekk fra μ i alle tre ulikhetene, og end opp med venstresiden av likning (1). (Tilleggsspørsmål: Hvorfor trekker vi fra og ikke plusser?)
c. Bruk deloppgave (iii) til å konkludere med at likheten i likning (1) stemmer.

Gratulerer, du forstår nå alle delene av utregningen som leder til likheten i likning (2)!

v. *OBS: følgende punkt er ikke like viktig at alle får til, men passer inn naturlig i oppgaven.* Generaliser punkt (iii) til at hvis man vet at

$$a_1 \leq a_2 \leq a_3 \leq \cdots \leq a_n,$$

så kan man "gange med -1" overalt, og få at også

$$-a_n \leq -a_{n-1} \leq \cdots \leq -a_2 \leq -a_1.$$

Oppgave 2. På onsdagsforelesningen fikk dere oppgitt at hvis

$$X \sim N(\mu_X, \sigma_X)$$

og

$$Y \sim N(\mu_Y, \sigma_Y)$$

er *uavhengige*, er både

$$Z = aX + bY$$

og

$$W = cX + d$$

normalfordelte hvis ikke både a og b er null og hvis ikke c er null. Vi skal i denne oppgaven se hvilken normalfordeling de har, og de som er litt ekstra tøffe kan også se på noen interessante konsekvenser av dette resultatet i de to siste punktene til oppgaven.

i. Vis at

$$\mu_Z = a\mu_X + b\mu_Y, \qquad \mu_W = c\mu_X + d.$$

og

$$\sigma_Z = \sqrt{a^2\sigma_X^2 + b^2\sigma_Y^2}, \qquad \sigma_W = c\sigma_X.$$

Konkluder med at

$$Z \sim N\left(a\mu_X + b\mu_Y, \sqrt{a^2\sigma_X^2 + b^2\sigma_Y^2}\right)$$

og

$$W \sim N(c\mu_X + d, c\sigma_X).$$

ii. Finn a, b, c og d-verdier som gjør at både Z og W er standard normale *uttrykt ved hjelp av* μ_X, μ_Y *og* σ_X, σ_Y. Obs: det er uendelig mange av dem – det holder å finne ett sett verdier, men klarer du å beskrive alle?

iii. *De to siste punktene av oppgaven er frivillig, og telles ikke som eksamenspensum.* Anta $X_1, X_2, \ldots, X_n$ alle er $N(\mu, \sigma)$-fordelt og er uavhengige. Som vi så i forelesningen, har

$$\bar{X}_n = \frac{1}{n}\sum X_i$$

fordelingen $N(\mu, \sigma/\sqrt{n})$. Anta at vi også har en X_0, som også er $N(\mu, \sigma)$-fordelt, og også uavhengig av de andre X_i'ene. Vis at fordelingen til

$$Q_n = \sum_{i=0}^{n} \left(\frac{1}{2}\right)^i X_i$$

er

$$N\left(2\mu\left(1 - \left(\frac{1}{2}\right)^{n+1}\right), \frac{2}{\sqrt{3}}\sigma\sqrt{1 - \left(\frac{1}{4}\right)^{n+1}}\right)$$

Hint: Husk at $\sum_{i=0}^{n}\left(\frac{1}{2}\right)^i$ er en *geometrisk sum* som det finnes en formel for.

iv. Tenk deg nå at n får lov til å vokse og vokse, og at den går mot uendelig. Hvorfor tyder forrige punkt på at

$$\lim_{n\to\infty} P(Q_n \leq x) = P\left(Z \leq \frac{\sqrt{3}}{2}\frac{x - 2\mu}{\sigma}\right),$$

der $Z \sim N(0, 1)$?

2. Minitaboppgave til Kapittel 5

Oppgave 3. Det finnes folk som tror at alt tilfeldig er normalfordelt, og det er selvfølgelig feil. Men under visse antagelser, som ofte er (tilnærmet) oppfylt i virkeligheten, vil summer av observasjoner som kan sees på som tilfeldige være normalfordelt. Dette er kjent som Sentralgrenseteoremet[1].

[1] Sentralgrenseteoremet er ofte formulert som at gjennomsnittet $\frac{1}{n}\sum X_i$ er tilnærmet normalfordelt. Men det er det samme som at summer av variable er normalfordelt, siden $\frac{1}{n}$ kan ganges inn i summetegnet.

Det er stort sett dette matematiske resultatet vi lener oss på når vi gjør statistiske tester på datasett som faktisk *ikke er normalfordelt*; ofte vil betingelsene til Sentralgrenseteoremet være oppfylt, og man kan så vise at testene er tilnærmet riktige allikevel! Selv om vi altså ikke kjenner fordelingen til datasettet.

Dette er overraskende, og slett ikke opplagt; hvorfor skulle summer oppføre seg så annerledes enn hvert av leddene? Man kan, ved hjelp av en del matematikk, vise dette resultatet. Vi, derimot, skal i denne oppgaven se at dette faktisk fungerer ved hjelp av et *simuleringsforsøk*.

Planen for denne oppgaven er som følger. Gjør først følgende to skritt.

i. Få Minitab til å generere $N = 500$ binomiske variable. Hver binomiske variabel er antall suksesser i et forsøk som repeteres $n = 20$ ganger, og hver gang har $p = 1/2$ i sannsynlighet for å resultere i en "suksess". F.eks kan vi tenke oss 20 myntkast, og at man noterer ned antall kron, f.eks. Altså får vi en tabell med 500 tall mellom 0 og 20.
ii. Ta summen av disse 500 forsøkene, og kall denne summen S_1.

Repeter så disse to skrittene, la oss si $M = 200$ ganger, men hvor summen kalles S_2 når man gjør dette andre gang, S_3 når man gjør det tredje gang, osv opp til S_{200}.

Sentralgrenseteoremet sier at $S_1, S_2, \ldots, S_{200}$ er tilnærmet normalfordelt!

Når vi har lagd disse S'ene, vil vi altså sjekke Sentralgrenseteoremets noe modige påstand. I uke 13 lærte vi om kvantilplott (side 80 i boken), som ser om et datasett avviker fra Normalfordelingen. Derfor setter vi Sentralgrenseteoremet på prøve ved å se på kvantilplottet til $S_1, S_2, \ldots, S_{200}$. Hvis kvantilplottene ser greie ut har vi altså ikke grunn til å tvile på Sentralgrenseteoremets utsagn. Men vi har heller ikke *påvist* at Sentralgrenseteoremet gjelder i dette tilfellet: vi har bare påvist at vi *ikke har motbevist* sentralgrenseteoremet. (Obs: Denne logikken kommer igjen i seksjon 6.2, som vi har om i neste forelesning, i form av *statistiske hypotesetester*, og det kan derfor være greit å skjønne denne logikken først som sist.)

Minitab kan gjøre dette via de følgende skrittene.

i. Gå på

$$\textbf{Calc} \mapsto \textbf{Random Data} \mapsto \textbf{Binomial}\ .$$

Vi har lyst til å lagre hver serie av 500 forsøk kolonnevis, og å lage 200 slike kolonner.

Dermed vil vi ha "Number of rows of data to generate" satt til 500, og at vi skal lagre dette i 200 kolonner. Videre er "Event probability" tilsvarende suksesssannsynlighet p og "Number of trials" er n – antall myntkast i eksemplet over.

Fyll derfor inn 500 for "Number of rows of data to generate", 20 for "Number of trials" og $0,5$ for "Event probability". Skriv videre inn "C1-C200" (altså kolonne 1 til 200) i "Store in columns".

ii. I neste skritt skal vi bruke en kommando som trenger datasettet formatert på en litt annen måte. Dette er litt uheldig, men det ser ut til at de minitabkommandoen vi vil bruke i neste skritt krever at vi må "transponere" datasettet, som vil si å bytte om rad og kolonne-markeringen. Gjør dette via

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

Sett inn "C1-C200" på "Transpose the following columns", og velg "In new worksheet" for "Store transpose".

iii. Under det grafiske brukergrensesnittet, har Minitab også kraftige kommandomuligheter. Vi får nå bruk for å skrive direkte inn en enkel kommando på denne måten. Finn vinduet som heter "Sessions", det vil si der hvor f.eks resultatet for summary statistics kommer opp, og klikk på det. Gå så på

Editor $\mapsto$ Enable Commands

for å skru på disse kommandoverktøyene. Finn så tilbake til "Sessions"-vinduet. Den forrige kommandoen gjør at vi kan skrive inn mer avanserte kommandoer til Minitab. Skriv deretter

```
rsum c2-c501 c502
```

som betyr at Minitab skal ta radsummen ("Row sum" $\mapsto$ "rsum") for hver av kolonnene[2] C2, C3, ..., C501. Den skal så plassere resultatet i C502, den 502'e kolonnen på regnearket. Den 502'e kolonnen er nå en realisering av $S_1, S_2, \ldots, S_{200}$.

Består sentralgrenseteoremet testen? Sjekk dette ved å lage et kvantilplott og histogram av C502.

Lag så et nytt "worksheet" og eksperimenter så med andre verdier av N – antall simuleringer, n – antall forsøk på det binomiske forsøket og p – suksessansynligheten for hvert av de n forsøkene. Sentralgrenseteoremet trenger stor N hvis fordelingen til det man summerer er *ganske skjev*. Når n er liten, og p er for nær null eller en, er nettopp den binomiske fordelingen ganske skjev. Men hvor er smertegrensen til Sentralgrenseteoremet? Gjenta oppsettet over, men sett inn følgende n og p-verdier:

1. Klarer den seg på $n = 10$ og $p = 0.1$?
2. Klarer den seg på $n = 5$ og $p = 0.0001$?

[2] Vi tar ikke med den første kolonnen da Minitab bruker denne til radnummerering.