OPPGAVEHEFTE I STK1000 TIL KAPITTEL 7

1. Regneoppgaver til kapittel 7

Oppgave 1. Anta at man har resultatet av et randomisert forsøk med to grupper, og observerer

$$X_{1,1}, X_{1,2}, \ldots, X_{1,n_1}$$

fra gruppe 1, mens man observerer

$$X_{2,1}, X_{2,2}, \ldots, X_{2,n_2}$$

fra gruppe to. Vi antar at alle observasjonene er uavhengige, og at observasjonene fra gruppe 1 er $N(\mu_1, \sigma_1)$-fordelte, mens observasjonene fra gruppe 2 antar vi er $N(\mu_2, \sigma_2)$-fordelte. Vi er interessert i å estimere μ_1 og μ_2, og se om de er forskjellige. Derfor jobber vi med

$$\bar{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1,i}, \qquad \bar{X}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2,i}$$

og lar

$$D = \bar{X}_1 - \bar{X}_2.$$

På onsdagsforelesningen påstod jeg at da må

$$D \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) \tag{1}$$

i. Hvorfor er $\bar{X}_1$ og $\bar{X}_2$ normalfordelte? Hvorfor må D også være normalfordelt? (Hint: Siden alle observasjonene er uavhengige må også $\bar{X}_1$ og $\bar{X}_2$ være uavhengige!)

ii. Vis at

$$\mu_D = \mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$$

og at

$$\sigma_D^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Konkluder med at likning (1) stemmer.

Oppgave 2. Følgende oppgave er ment å supplementere seksjon 7.1 og 7.1, samt være en forberedelse til de inferensmetodene vi møter i kapittel 10 og 11, som blant annet inkluderer andre typer konfidensintervaller og hypotesetester enn det vi til nå har sett. Poenget med oppgaven blir å vise at de faktisk er likere enn det man kan se ved første øyekast, og er veldig koblet til helt vanlige t-tester og konfidensintervall. Det er mulig at mange syns denne oppgaven ser ganske skummel ut, men det lønner seg å komme igjennom den for å forstå konfidensintervallene vi bruker i kapittel 7 og de vi skal bruke i kapittel 10 og 11 – så det er en vel verdt investering. Et hint er at den egentlig *ikke er så vanskelig som den kanskje ser ut som!* Det er nemlig slik at man egentlig skal regne *nesten samme regnestykke* i deloppgave ii, iii og iv.

Date: 9. november 2009.

i. Vi minner først om standardtilfellet vi har nå jobbet med en stund: Man har observasjoner $X_1, X_2, \ldots, X_n$ som er uavhengige, og som vi for enkelhetsskyld antar er normalfordelte $N(\mu, \sigma)$. Mens kapittel 6 jobber med

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}},$$

som bare kan beregnes hvis man antar σ kjent – jobber kapittel 7 med

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

som alltid kan beregnes fra data.

På forelesningen minnet jeg på at motivasjonen for å jobbe med Z er at den er *standardisert*, det vil si at den er standard normalfordelt. Vis at dette er tilfellet, ved hjelp av Oppgave 2 i oppgaveheftet for kapittel 5 og 6.

ii. Motivasjonen for å jobbe med en standardisert versjon av $\bar{X}$ og ikke $\bar{X}$ direkte, er at man alltid forholder seg til samme skala. Hvis man får en observert z-verdi på 4.5, vet man med en gang at det er veldig høyt, siden en standard normalfordelt variabel stort sett holder seg mellom -3 og 3 fra $68, 95, 99.7$-reglen.

Motivasjonen for å jobbe med $T = (\bar{X} - \mu)/(s/\sqrt{n})$ når vi ikke kjenner σ, er at det er en tilnærmet standardisering. Vi vet at s estimerer σ, og at for stor n er $s \approx \sigma$ takket være store talls lov. Hvis man introduserer

$$SE_{\bar{x}} = s/\sqrt{n}$$

er $SE_{\bar{x}}$ en *estimator for*

$$\sigma_{\bar{X}} = \sigma/\sqrt{n},$$

så akkurat som at man kan skrive

$$z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}},$$

kan man skrive

$$T = \frac{\bar{X} - \mu}{SE_{\bar{x}}}.$$

Denne tilnærmede standardiseringen gjør at man alltid har t på samme skala – uansett hva standardavviket til X er. Men i motsetning til at $Z \sim N(0,1)$, vil fordelingen til T avhenge av n. Vi vet at $t \sim t(n-1)$, det vil si har en t-fordeling med $n-1$-frihetsgrader. La $t^\star_C$ være det tallet slik at

$$P(-t^\star_C \leq T \leq t^\star_C) = C\%.$$

Jeg sa på forelesningen at med helt analoge skritt som da vi fant konfidensintervaller basert på Z, kan vi vise at man må ha

$$C\% = P(-t^\star_C \leq T \leq t^\star_C) = P\left(\bar{X} - t^\star_C SE_{\bar{x}} \leq \mu \leq \bar{X} + t^\star_C SE_{\bar{x}}\right),$$

slik at når man beregner

$$\bar{x} \pm t^\star_C SE_{\bar{x}}$$

fra data, så er det et C% konfidensintervall. Gjør denne utregningen.

iii. I kapittel 10 og 11 skal vi jobbe med konfidensintervaller på formen

$$\text{estimat} \pm t^\star_C SE_{\text{estimat}}$$

hvor estimatene ikke lenger er et enkelt gjennomsnitt som lar oss rettferdiggjøre konfidensintervaller rett fra sentralgrenseteoremet. Men til tross for at estimatene vi skal jobbe med har en litt mer komplisert form, viser det seg at man ofte kan bruke konfidensintervaller på akkurat samme form som før!

La oss så tenke oss at vi har en estimator b for β (den greske bokstaven "beta") og har $\mu_b = \beta$ og standardavvik σ_b som er ukjent. Det vil si at $b \approx \beta$ hvis man beregner b fra data – når antall observasjoner n er stor nok.

Man vet ikke σ_b, men la oss si at man kan estimere den, og la oss kalle dette estimatet SE_b. Anta så at

$$S = \frac{b - \beta}{\mathrm{SE}_b}$$

er tilnærmet $t(n-2)$ fordelt (altså, $n-2$ og ikke $n-1$ som i kapittel 7.1). Dette betyr at man kan rettferdiggjøre skritt av typen

$$P(S \leq x) \approx P(T \leq x) \qquad \text{og} \qquad P(a \leq S \leq b) \approx P(a \leq T \leq b)$$

for en tilfeldig variabel $T \sim t(n-2)$ og hvilke som helst tall a, b, x. Ved å gå igjennom tilsvarende skritt som forrige deloppgave, vis at man har

$$P\left(b - t^\star_C \mathrm{SE}_{\bar{x}} \leq \beta \leq b + t^\star_C \mathrm{SE}_{\bar{x}}\right) \approx C,$$

som vil si at $b \pm t^\star_C \mathrm{SE}_b$ er et tilnærmet C% konfidensintervall for β.

iv. La oss så gå over til seksjon 7.2, hvor man observerer

$$X_{1,1}, X_{1,2}, \ldots, X_{1,n_1}$$

fra gruppe 1, mens man observerer

$$X_{2,1}, X_{2,2}, \ldots, X_{2,n_2}$$

fra gruppe to. Vi antar at alle observasjonene er uavhengige, og at observasjonene fra gruppe 1 er $N(\mu_1, \sigma_1)$-fordelte, mens observasjonene fra gruppe 2 antar vi er $N(\mu_2, \sigma_2)$-fordelte. På onsdagsforelesningen fikk vi oppgitt at for $D = \bar{X}_1 - \bar{X}_2$, er

$$T = \frac{D - \mu_D}{\mathrm{SE}_D} = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

tilnærmet $t(k)$-fordelt for en bestemt k (hvor man enten bruker den enkle reglen at k er den minste av $n_1 - 1$ og $n_2 - 1$, eller lar k tilnærmes via f.eks Minitab på en bedre men mer komplisert måte), og hvor

$$\bar{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1,i}, \qquad \bar{X}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2,i}$$

og standardfeilen, som estimerer (se Oppgave 1!)

$$\sigma_D = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}},$$

er gitt helt analogt med forrige oppgave som

$$\mathrm{SE}_D = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

definert ut i fra

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}\left(X_{1,i} - \bar{X}_1\right)^2, \qquad s_2^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}\left(X_{2,i} - \bar{X}_2\right)^2.$$

Ved å gå igjennom tilsvarende skritt som forrige deloppgaver, vis at man har

$$P\left((\bar{X}_1 - \bar{X}_2) - t^\star_C \mathrm{SE}_D \leq \mu_1 - \mu_2 \leq (\bar{X}_1 - \bar{X}_2) + t^\star_C \mathrm{SE}_D\right) \approx C,$$

som vil si at $(\bar{X}_1 - \bar{X}_2) \pm t^\star_C SE_D$ er et tilnærmet C% konfidensintervall for $\mu_1 - \mu_2$.

Oppgave 3. Da vi lærte om hypotesetester i kapittel 6 fikk vi oppgitt at det å forkaste nullhypotesen om $H_0 : \mu = \mu_0$ mot en tosidig alternativ hypotese $H_A : \mu \neq \mu_0$ på nivå α er det samme som at μ_0 ikke er med i et konfidensintervall for μ med nivå $1 - \alpha$. Vi skal her først regne oss frem til at dette er tilfellet, og så se at det også stemmer for t-tester.

i. Gitt $X_1, X_2, \ldots, X_n$ som er uavhengige normalfordelte variable med forventning μ og (for øyeblikket) kjent standardavvik σ, anta vi er interessert i en tosidig hypotesetest med

$$H_0 : \mu = \mu_0$$

og

$$H_A : \mu \neq \mu_0.$$

Man aksepterer H_0 på nivå α hvis

$$2P\left(Z \geq \left|\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right|\right) \geq \alpha,$$

der $Z \sim N(0,1)$, altså der stor Z er en standard normalfordelt tilfeldig variabel. Dette er det samme som at

$$P\left(Z \geq \left|\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right|\right) \geq \alpha/2.$$

Forklar hvorfor

$$P\left(Z \geq z^\star_{1-\alpha}\right) = \alpha/2,$$

der $z^\star_{1-\alpha}$ er tallet vi også møtte i konstruksjonen av konfidensintervaller – nemlig det tallet som gjør at

$$P(-z^\star_{1-\alpha} \leq Z \leq z^\star_{1-\alpha}) = 1 - \alpha.$$

Konkluder med at å akseptere H_0 på nivå α er det samme som at

$$P\left(Z \geq \left|\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right|\right) \geq P\left(Z \geq z^\star_{1-\alpha}\right).$$

ii. Forklar hvorfor

$$P\left(Z \geq \left|\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right|\right) \geq P\left(Z \geq z^\star_{1-\alpha}\right) \tag{2}$$

er det samme som at

$$\left|\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right| \leq z^\star_{1-\alpha}$$

Hint: Tegn opp tetthetskurven til normalfordelingen, og tegn hva det vil si at likning (2) er sann.

iii. Overbevis deg selv om at for to tall x og z, er

$$|x| \leq z$$

det *samme* som at

$$-z \leq x \leq z.$$

Bruk dette til å konkludere med at

$$\left|\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right| \leq z^\star_{1-\alpha}$$

er det samme som at

$$-z^\star_{1-\alpha} \leq \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \leq z^\star_{1-\alpha}.$$

iv. Bruk konklusjonen til forrige oppgave for å regne deg frem til at å akseptere H_0 på nivå α er *det samme som* at

$$\bar{X} - z^\star_{1-\alpha}\frac{\sigma}{\sqrt{n}} \leq \mu_0 \leq \bar{X} + z^\star_{1-\alpha}\frac{\sigma}{\sqrt{n}},$$

og konkluder med at dette er det samme som at μ_0 er innenfor konfidensintervallet

$$\bar{X} \pm z^\star_{1-\alpha}\frac{\sigma}{\sqrt{n}}$$

med nivå $1-\alpha$.

v. Følgende deloppgave antar at du allerede har gjort Oppgave 1 (ii) over. Anta nå samme oppsett som i deloppgave (i), men *at* σ *er ukjent*. Hvis man nå i stedet jobber med tallet $t^\star_{1-\alpha}$, som er slik at

$$P(-t^\star_{1-\alpha} \leq T \leq t^\star_{1-\alpha}) = 1-\alpha,$$

der $T \sim t(n-1)$. Vis at det å akseptere H_0 på et nivå α er det samme som at μ_0 er innenfor det vanlige t-baserte konfidensintervallet

$$\bar{X} \pm t^\star_{1-\alpha}\frac{s}{\sqrt{n}}$$

på nivå $1-\alpha$ ved å gå igjennom nesten nøyaktig de samme skrittene som de forrige deloppgavene.

Et addendum til oppgaven er at man pleier å si at det vi viste over sier at det er en *dualitet* mellom konfidensintervaller og tosidige hypoteser om μ. Dualiteter viser at to ting med forskjellig *tolkning* (nemlig konfidensintervaller og tosidige hypotesetester), faktisk er like.

2. Minitaboppgaver til kapittel 7

Oppgave 4. En juiceprodusent ønsker å teste sin påstand om deres multijuiceblanding har et C-vitamininnhold på 25 mg per 100 ml juice. Man måler derfor C-vitamininnholdet i 20 appelsinsaftflasker for å se om disse målingene gir oss grunn til å tvile på C-vitamininnholdet.

Resultatene ble følgende:

30	24	19	23	16	18	29	19	31	14
22	16	29	15	26	17	27	14	30	15

(A) Skriv inn dataene i C1 i Minitab. Gi kolonnen navnet Cvit. Finn estimater for forventningen μ og variansen σ^2 for resultatene. Finn også et estimat for variansen til estimatoren for μ.

I Minitab, via rullegardinmenyen:

Stat $\mapsto$ **Basic Statistics** $\mapsto$ **Display basic statistic.**

Skriv deretter Cvit i *variables*. Alternativt kan du skrive describe 'Cvit' i *Session* vinduet.

(B) Angi et 95 % konfidensintervall for μ.

I Minitab:

$$\textbf{Stat} \mapsto \textbf{Basic Statistics} \mapsto \textbf{1-Sample t}$$

Skriv Cvit i *Samples in columns*. Sjekk via *options* om du har riktig konfidenskoeffisient.

(C) Vi skal som sagt teste produsentenes påstand. Formuler nullhypotese og en alternativ hypotese. Utfør testen og formuler en konklusjon. Bruk 5 % signifikansnivå. Ville konklusjonen blitt den samme om vi på forhånd hadde valgt 1 % signifikansnivå?

I Minitab:

$$\textbf{Stat} \mapsto \textbf{Basic Stat} \mapsto \textbf{1 sample t},$$

og tast inn riktige tall i *Test Mean*. Gå i tillegg inn på *options* og velg riktig alternativ (avhenger av hvordan du formulerte din alternative hypotese).

Oppgave 5. Gjør oppgave 7.2 i boken. Utfør t-test- og konfidens-prosedyrene ved hjelp av

$$\textbf{Stat} \mapsto \textbf{Basic Statistics} \mapsto \textbf{1-Sample t}$$

i Minitab.

Oppgave 6. Gjør oppgave 7.3 i boken.

I oppgave 7.3 skal vi se på logaritmen av et datasett, slik vi gjorde i Oppgave 8 i oppgaveheftet til kapittel 2. Men, siden det er observasjoner som formelt sett er 0 (men se Oppgave 1 (A) i oppgaveheftet for kapittel 1!), blir vi bedt om å legge til 1 på alle observasjonene[1]. Dette gjøres lettest ved å gå på

$$\textbf{Calc} \mapsto \textbf{Calculator},$$

så klikk inn "C1 CRP" (altså første kolonne) både til "Store result in variable" og til "Expression". Fyll inn "+1" etter 'CRP' i "Expression"-feltet, og trykk OK. Følg så Minitaboppskriften fra Oppgave 8 i oppgaveheftet til kapittel 2 for å ta logaritmen til observasjonene. Husk å legg dem i en egen kolonne!

Merk at en raskere måte å gjøre det over på ville vært å rett og slett skrive inn

```
log('CRP'+1)
```

i "Expression" når man er i Minitabs kalkulator! Husk å lagre resultatet i en ny kolonne, så datasettet ikke overskrives; gi dessuten denne kolonnen et fornuftig navn.

Oppgave 7. Ved Odense sykehus ble 205 pasienter operert for hudkreft (maligne melanomer) i løpet av en periode på 15 år. Det ble registrert ulike opplysninger om pasientene. Vi skal se på noen av disse, nemlig pasientens kjønn, alder og tumorens

[1] Grunnen til at man ikke vil ta logaritmen til 0 er at den ikke er definert. Man legger til 1, siden $\log 1 = 0$. Så dermed blir den minste observasjonen 0.

tykkelse. Vi ser dessuten ikke på hele datamaterialet, men kun på data for pasienter som er 45 år eller yngre. Dette datamaterialet er hentet fra Andersen, Borgan, Gill og Keiding (1993). Datasettet heter *hudkreft.txt* og finnes linket til på kurshjemmesidens oversikt over ukesoppgaver. 1. kolonne inneholder tumortykkelsen i mm, 2. kolonne inneholder pasientens kjønn (1= kvinne, 2=mann) og 3. kolonne inneholder pasientens alder. Les datasettet inn i Minitab, og sett på fornuftige kolonnenavn.

(A) Beregn gjennomsnittlig tumortykkelse, median og empirisk standardavvik for kvinner og menn. Er det noen forskjeller? Kommenter.

I Minitab, under rullegardinmenyen:

Stat $\mapsto$ **Basic Statistics** $\mapsto$ **Display Descriptive Statistics.**

Velg *Variable* C1 *by variable* C2.

(B) Man lurer på om det kan være en forskjell på tumortykkelsen til kvinner og menn. Sett opp nullhypotese og alternativhypotese. Sett opp to-utvalgs-t-testobservator (se side 529, læreboka). Hva blir antall frihetsgrader?

I Minitab, under rullegardinmenyen:

Stat $\mapsto$ **Basic Statistics** $\mapsto$ **2 sample t.**

Under *Samples in one column* legges C1 bak *Samples* og C2 bak *Subscripts*.

Vil du fra utskriften i Minitab si at det ser ut til å være forskjell mellom kjønnene når det gjelder tumortykkelse?

(C) Vi skal nå undersøke om det er noen forskjell på tumortykkelsen for pasienter som er 30 år eller yngre og de over 30 år. Du må først lage en kolonne C4 med 0 og 1 verdier (for eksempel 1 for alder $\leq$ 30 og 0 for alder $>$ 30). Gjør (B) om igjen, nå med C4 istedenfor C2.
Vil du si at det ser ut til å være forskjell mellom unge($\leq$ 30 år) og eldre ($>$ 30 år) når det gjelder tumortykkelse?

I Minitab, under rullegardinmeny:

Calc $\mapsto$ **Calculator**

I *Store results in variable* skriver du C4, og i *Expression* skriver du

```
1*WHEN(C3 <= 30),
```

Gjør deretter som i (B).