Obligatorisk øving for STK1110, Høsten 2021

Øving 2 av 2

Innleveringsfrist

Torsdag 21. oktober 2021, klokken 14:30 i Canvas (`canvas.uio.no`).

Instruksjoner

Du velger selv om du skriver besvarelsen for hånd og scanner besvarelsen eller om du skriver løsningen direkte inn på datamaskin (for eksempel ved bruk av Latex). Besvarelsen skal leveres som **én PDF-fil**. Scannede ark må være godt lesbare. Besvarelsen skal inneholde navn, emne og oblignummer.

Det forventes at man har en klar og ryddig besvarelse med tydelige begrunnelser. Husk å inkludere alle relevante plott og figurer. Studenter som ikke får sin opprinnelige besvarelse godkjent, men som har gjort et reelt forsøk på å løse oppgavene, vil få én mulighet til å levere en revidert besvarelse. Samarbeid og alle slags hjelpemidler er tillatt, men den innleverte besvarelsen skal være skrevet av deg og reflektere din forståelse av stoffet. Er vi i tvil om du virkelig har forstått det du har levert inn, kan vi be deg om en muntlig redegjørelse. I oppgaver der du blir bedt om å programmere må du legge ved programkoden og levere den sammen med resten av besvarelsen.

Søknad om utsettelse av innleveringsfrist

Hvis du blir syk eller av andre grunner trenger å søke om utsettelse av innleveringsfristen, må du ta kontakt med studieadministrasjonen ved Matematisk institutt (e-post: studieinfo@math.uio.no) i god tid før innleveringsfristen. For å få adgang til avsluttende eksamen i dette emnet, må man bestå alle obligatoriske oppgaver i ett og samme semester.

Spesielt om dette oppgavesettet

Du skal bruke programpakken R til å gjøre beregninger i oppgavene, og du må angi hvilke kommandoer du har brukt for å komme fram til svarene dine. For å få godkjent besvarelsen, må du ha minst 65% riktig på hver av de tre oppgavene.

For fullstendige retningslinjer for innlevering av obligatoriske oppgaver, se her:

`www.uio.no/studier/admin/obligatoriske-aktiviteter/mn-math-oblig.html`

LYKKE TIL!

Oppgave 1

En av de viktigste ingrediensene i såkalt 'supermat' er blåbær. At blåbærene regnes som supermat, skyldes at fargestoffet antocyan, som gjør bærene blå, er en kraftig antioksidant.

I forbindelse med en studie av antioksidanter og antocyaner, ble innholdet av antocyan i 15 beger med blåbær målt. De målte verdiene var (i mg per 100 gram bær):

525 587 547 558 591 531 571 551 566 622 561 502 556 565 562

Vi antar at målingene kan betraktes som realisasjoner av uavhengige normalfordelte variabler med forventning μ og varians σ^2.

a) Lag et 95% konfidensintervall for forventet antocyaninnhold μ basert på målingene over.

b) På Wikipedia kan vi lese at forventet antocyaninnhold i blåbær er 558 mg/100g. Nå skal du bruke simuleringer til å late som om du måler antocyan i 15 prøver med blåbær veldig mange ganger. Generér 10000 datasett, hvert av størrelse $n = 15$, bestående av observasjoner er av de stokastiske variablene $X_1, \ldots, X_{15} \overset{uif}{\sim} N(558, 30^2)$. Du kan bruke `rnorm()`-funksjonen i R til dette. Selv om du har simulert fra en fordeling med kjent forventning og varians, skal du late som om begge disse er ukjent i det følgende. Lag et 95% konfidensintervall for μ som i punkt a), basert på hvert av de simulerte datasettene, slik at du får 10000 intervaller. Tell opp andelen av disse intervallene som inneholder verdien 558. Kommentér og forklar.

c) Bruk nå i stedet det tilnærmede intervallet for store utvalg, altså

$$\left(\bar{X} - 1.96\frac{S}{\sqrt{15}}, \bar{X} + 1.96\frac{S}{\sqrt{15}}\right)$$

med

$$\bar{X} = \frac{1}{15}\sum_{i=1}^{15} X_i \text{ og } S^2 = \frac{1}{15-1}\sum_{i=1}^{15}\left(X_i - \bar{X}\right)^2,$$

og beregn dette for hvert av 10000 datasett, generert som i b). Hvor stor andel av intervallene inneholder $\mu = 558$? Kommentér og forklar resultatet.

d) Trekk 10000 datasett som i b) og lag et 95% konfidensintervall for σ for hvert av dem. Hvor stor andel av intervallene inneholder $\sigma = 30$?

e) Under antakelsen om normalfordeling er $Z_i = \frac{X_i - \mu}{\sigma} \sim N(0,1)$, $i = 1, \ldots, n$, med $\mu = 558$ og $\sigma = 30$. Anta nå at $Z_1, \ldots, Z_{15}$ i virkeligheten er t-fordelt med 7 frihetsgrader, altså $Z_1, \ldots, Z_{15} \overset{uif}{\sim} t_7$. Trekk nå 10000 datasett fra denne fordelingen ved å

1. trekke $z_1, \ldots, z_{15}$ fra t_7 med R-funksjonen `rt()`

2. la $x_i = \mu + \sigma z_i$, $i = 1, \ldots, n$.

Gjenta deretter oppgave b) med de nye datasettene. Hvor robust er metoden for å lage konfidensintervall for forventningsverdien for antakelsen om normalfordeling?

f) Trekk datasett som i e) og lag deretter 95%konfidenintervall for standardavviket til X_i slik som i d). Merk imidlertid at $\text{Var}(Z_i) = \frac{7}{7-2}$ slik at variansen til X_i nå er $\tilde{\sigma}^2 = \text{Var}(X_i) = \text{Var}(\mu + \sigma Z_i) = \sigma^2 \text{Var}(Z_i) = 1.4\sigma^2$. Det er altså $\tilde{\sigma}$ du skal lage konfidensintervall for, og sjekke andelen intervaller som inneholder $\tilde{\sigma}$. Sammenlign med resultatene fra d) og kommentér.

Oppgave 2

Følgende tabell viser 10 målinger av kroppstemperatur for kvinner, $x_1, \ldots, x_{10}$, og 10 målinger for menn, $y_1, \ldots, y_{10}$. Hensikten med denne oppgaven er å undersøke om det er tilstrekkelig informasjon i tabellen til å kunne konkludere med at kroppstemperaturen er forskjellig for menn og kvinner. Dataene finnes som `"temp.txt"` i mappen `https://www.uio.no/studier/emner/matnat/math/STK1110/data/`.

Kroppstemperatur	
Menn	Kvinner
36.1	36.6
36.3	36.7
36.4	36.8
36.6	36.8
36.6	36.7
36.7	37.0
36.7	37.1
37.0	37.3
36.5	36.9
37.1	37.4

a) Lag et boksplott som viser fordelingen av observasjonene. Kommentér hva du finner.

b) Lag normalfordelingsplott for de to observasjonssettene, altså ett for menn og ett for kvinner. Kommentér hva du ser.

I resten av oppgaven antar vi at observasjonene er realisasjoner av normalfordelte variabler. I c) og d) skal du forklare hvordan tester og konfidensintervaller konstrueres, og sette inn i formlene du utleder. Sjekk deretter svarene du får mot R-prosedyren `t.test()`.

c) Anta at variansen er den samme for de to utvalgene, og test med signifikansnivå 5% om det er noen forskjell i forventet kroppstemperatur. Beregn P-verdien, og lag et 95% konfidensintervall for denne forskjellen.

d) Gjennomfør testen og beregn P-verdien også i det tilfellet der en ikke antar felles varians. Diskutér og forklar resultatene.

e) Utled og gjennomfør en F-test for å sjekke om det er noen grunn til å påstå at variansene er forskjellige. Sjekk mot `var.test()` i R .

f) Se nå på situasjonen der en vurderer å innhente to nye målinger. La X_{11} være verdien for kvinnen og Y_{11} verdien for mannen, slik at forskjellen er $X_{11} - Y_{11}$. Vi antar nå at alle observasjonene er normalfordelte med samme varians. Begrunn at et rimelig anslag for $X_{11} - Y_{11}$ er differansen mellom gjennomsnittet av de 10 eksisterende målingene for kvinner og menn, altså $\bar{X} - \bar{Y}$. Hva er fordelingen til $X_{11} - Y_{11} - (\bar{X} - \bar{Y})$? Bruk dette til å lage et 95% *prediksjonsintervall* for $X_{11} - Y_{11}$, altså et intervall som med sannsynlighet 0.95 inneholder $X_{11} - Y_{11}$. Dette er gjennomgått for ett-utvalgs-situasjonen på forelesning. Forklar hva som er forskjellen mellom et slikt intervall og et konfidensintervall for $\mu_1 - \mu_2$. Hvordan skal et prediksjonsintervall tolkes? [Hint: Siden alle variablene er normalfordelte, er $X_{11} - Y_{11} - (\bar{X} - \bar{Y})$ også det. Det er derfor nok å beregne forventning og varians for å finne fordelingen til denne størrelsen.]

Oppgave 3

En undersøkelse presentert i Aftenposten slo opp på førstesiden at småbarnsfedrene nå opplever tidsklemma (mellom familie og arbeidsliv) sterkere enn småbarnsmødrene. Undersøkelsen bygde på intervjuer med 3000 kvinner og 3000 menn som har barn i rett alder. Det var 16.2% av fedrene (dvs. 486 personer) som ofte opplevde tidsklemmeproblemer, mens 14.7% (dvs. 441 personer) av mødrene opplevde det samme.

a) Er forskjellen mellom mødre og fedre signifikant? Formulér hypoteser, beregn en p-verdi, og konkludér. Kommentér kort.

b) Kontrollér svaret ditt ved å bruke `prop.test()` i R.