Obligatorisk øving for STK1110, Høsten 2023

Øving 2 av 2

Innleveringsfrist

Torsdag 9. november 2023, 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. Det er kun **ett forsøk**, og det er dermed ikke mulighet til å levere en revidert besvarelse dersom en ikke består. 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 en 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 4 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

Et oppdrettsanlegg for kveite har testet to ulike fôrtyper i en sommersesong. Småfisk i en merde ble fordelt tilfeldig på to andre merder, og ble fôret med fôr av ulik type. I september trekkes det ut 14 tilfeldige fisk fra hver merde. Gir dataene grunnlag for å konkludere at fôr B gir fisk med størst vekt?

Dataene[1] finnes som `"kveite_for.txt"` i mappen `https://www.uio.no/studier/emner/matnat/math/STK1110/data/` og kan leses inn med komandoene

```
dir = "https://www.uio.no/studier/emner/matnat/math/STK1110/data/"
d = read.table(paste(dir,"kveite_for.txt",sep=""),header=T)
```

Dataene er også gitt i tabellen nedenfor.,

Fortype	Vekt (kg) av kveite
A	15.16 17.28 12.33 11.86 16.31 12.08 9.96 16.18 15.96 18.75 15.35 19.52 16.88 13.90
B	17.73 16.03 14.85 18.18 17.08 19.06 15.64 18.29 23.41 14.22 17.35 16.89 16.59 18.54

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

(b) Lag normalfordelingsplott for de to observasjonssettene, altså ett for fôrtype A og ett for fôrtype. Kommentér hva du ser.

I resten av oppgaven antar vi at observasjonene er realisasjoner av normalfordelte variabler.

(c) Anta at variansen er den samme for de to utvalgene.

Formuler en null-hypotese H_0 og en alternativ hypotese H_a for å kunne konkludere om fôr B gir fisk med størst vekt.

Skriv ned testobservatoren du vil bruke for å teste hypotesene mot hverandre. Hva slags fordeling har denne testobservatoren?

(d) Bruk testobservatoren fra (c) til å utføre testen med signifikansnivå 5%. Beregn også P-verdien.

Sammenlikn resultatene du får med et direkte kall på R-prosedyren `t.test()`.

Diskutér og forklar resultatene.

(e) Gjennomfør testen og beregn P-verdien også i det tilfellet der en ikke antar felles varians.

Gjennomfør også en test for å sjekke om det er noen grunn til å påstå at variansene er forskjellige.

Diskutér og forklar resultatene.

[1] Hentet fra Gunnar Løvås: Statistikk: for universiteter og høgskoler 1999

(f) Lag nå en vektor $\boldsymbol{y}$ som består av alle de 14+14=28 observasjonene. Definer en vektor $\boldsymbol{x}$ av lengde 28 der $x_i = 0$ hvis observasjon y_i tilhører fôrtype A mens $x_i = 1$ hvis observasjon y_i tilhører fôrtype B. Analyser nå dataene med en lineær regresjonsmodell

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

der vi antar $\varepsilon_i \overset{uif}{\sim} N(0, \sigma^2)$.

Forklar hvorfor P-verdien knyttet til forklaringsvariabelen x blir dobbelt så stor som P-verdien du fikk i (c).

Oppgave 2

Siden eneggede tvillinger har samme genetiske materiale, brukes såkalte tvillingstudier til å kartlegge hvordan miljøet virker inn på ulike egenskaper. I en bok av den amerikanske forskeren Susan Faber finner vi data for $n = 31$ tvillingpar, der den ene tvillingen vokste opp hos biologiske foreldre (Twin A) og den andre vokste opp hos andre familiemedlemmer, foster- eller adoptivforeldre (Twin B). Nedenfor finnes en oppsummering av målt IQ for disse personene. Spørsmålet vi ønsker å belyse er om det er forskjell i IQ hos eneggede tvillinger der den ene tvillingen har vokst opp hos biologiske foreldre, og den andre ikke.

	N	Mean	StDev	SE Mean
Twin A	31	93.32	15.41	2.77
Twin B	31	96.58	13.84	2.49
Difference	31	-3.26	8.81	1.58

(a) Begrunn hvorfor en paret sammenligning er best egnet i denne situasjonen. Beskriv kort hvilke antakelser vi må legge til grunn for videre analyse.

(b) Kall forventet forskjell mellom Twin A og Twin B for μ_D. Sett opp nullhypotese og alternativ hypotese for å besvare spørsmålet om forskjell i IQ. Finn en egnet testobservator, og beregn dennes verdi. Beregn så tilhørende p-verdi. Spesifisér antall frihetsgrader i fordelingen du bruker. Formulér din konklusjon på testen.

(c) Finn et 95% konfidensintervall for μ_D. Hva betyr det at dette intervallet dekker kun negative verdier? Forklar kort om sammenhengen mellom tosidig testing og konfidensintervaller.

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. 16.2% av fedrene (dvs. 486 personer) opplevde ofte 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.

Oppgave 4

Tabellen nedenfor angir 18 målinger av snømengde om vinteren i et fjellområde og vannstanden i en elv i samme område etter snøsmelting om våren. De 18 målingene representerer 18 sesonger spredd over en lengre tidsperiode. Her er vannstanden, som er angitt i tommer, respons- eller avhengig variabel, mens snømengden, målt ved noe som heter vannekvivalens, er forklaringsvariabel. Sammenhengen mellom snømengde og vannstand er viktig for bl.a. prediksjon av vannføring og flomfare. Dataene finnes i fila `snoe_vann.txt` på kursets hjemmeside. De kan leses inn i R med følgende kommandoer:

```
dir = "https://www.uio.no/studier/emner/matnat/math/STK1110/data/"
d = read.table(paste(dir,"snoe_vann.txt",sep=""),header=F)
names(d) = c("Snoinnhold","Vannstand")
```

Snøinnhold	Vannstand
23.1	10.5
32.8	16.7
31.8	18.2
32.0	17.0
30.4	16.3
24.0	10.5
39.5	23.1
24.2	12.4
52.5	24.9
37.9	22.8
30.5	14.1
25.1	12.9
12.4	8.8
35.1	17.4
31.5	14.9
21.1	10.5
27.6	10.5
27.6	16.1

(a) Beskriv en enkel lineær regresjonsmodell for sammenhengen mellom snømengde og vannstand. Tilpass en regresjonslinje til dataene ovenfor ved hjelp av R-funksjonen `lm()`. Plott observasjonene og den tilpassede regresjonslinja. Virker estimatene for koeffisientene rimelige?

(b) Plott residualene mot forklaringsvariabelen. Lag også et normalfordelingsplott av residualene. Hvordan vurderer du modellens egnethet?

(c) Beregn et estimat for variansen til feilleddene. Konstruér et 95% konfidensintervall for stigningstallet β_1.

(d) Utled en test for $H_0 : \beta_0 = 0$ mot $H_1 : \beta_0 \neq 0$ med signifikansnivå 5%. Gjennomfør testen. Hva er p-verdien? Sammenlign med resultatene fra `lm()`.

Hvorfor kan det være rimelig med $\beta_0 = 0$ i dette tilfellet?

(e) Prøv så ut en lineær regresjonsmodell med andregrads og tredjegrads polynomer.

Hvilken modell vil du foretrekke? Begrunn svaret.

SLUTT