UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK1000 — Innføring i anvendt statistikk
Eksamensdag:	Onsdag 25. november 2020
Tid for eksamen:	15:00 – 19:00
Oppgavesettet er på 5 sider.	
Vedlegg:	Ingen
Tillatte hjelpemidler:	Alle hjelpemidler er tillatt, men det er ikke tillatt å kommunisere eller samarbeide med andre.

Kontroller at oppgavesettet er komplett før
du begynner å besvare spørsmålene.

Oppgavesettet har fire oppgaver som til sammen består av elleve deloppgaver. Hver deloppgave teller likt.

Oppgave 1

La høyden til nevøen være $X \sim N(\mu = 124.5, \sigma = 4.8)$.

1a

Beregner sannsynlighet for X vha standard normalfordelt $Z \sim N(0, 1)$.

$$P(X > 128) = P\left(Z > \frac{128 - 124.5}{4.8}\right) = P(Z > 0.73) = 1 - 0.7673 = 0.2327$$

1b

Finner verdi x_{90} slik at $P(X > x_{90}) = 10\% = 0.10$, som er det samme som at $P(X < x_{90}) = 0.90$. Fra standard normal-beregning har vi

$$P(X < x_{90}) = P(Z < \frac{x_{90} - 124.5}{4.8}).$$

Fra Tabell A leser vi at $0.9 = P(Z < 1.28)$. Det følger at $1.28 = \frac{x_{90} - 124.5}{4.8}$, og dermed

$$x_{90} = 124.5 + 1.28 \cdot 4.8 = 130.64.$$

Genseren finnes i klesstørrelsene 116, 122, 128, 134, 140, 146 og 152, og vi velger den minste av disse som er større enn (eller lik) 130.64: Vi velger derfor str 134.

(Fortsettes på side 2.)

Oppgave 2

Påstand: Prøven har konsentrasjon $\mu = 0.60$.

2a

Null-hypotese og alternativ hypotese til grunn for hypotesetesten vi ønsker å gjennomføre (tosidig test):

$$H_0: \ \mu = 0.60, \qquad H_a: \ \mu \neq 0.60.$$

2b

Vi ser på de to måleresultatene gitt i oppgaveteksten som et utvalg av en hypotetisk uendelig sekvens (eller sekvens av betydelig lengde) av repeterte konsentrasjons-målinger gjort av prøven, hver med måleresultat normalfordelt $\mathrm{N}(\mu, \sigma = 0.022)$.

Utvalgsgjennomsnittet er $\bar{x} = \frac{0.58+0.56}{2} = 0.57$.

Velger å gjennomføre statistisk hypotesetest på signifikansnivå $\alpha = 0.05$.

Standardisert test-observator er $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{2}}$.

Under H_0 er $\mu = 0.60$, som gir testobservatoren verdi

$$Z = \frac{0.57 - 0.60}{0.022/\sqrt{2}} = -1.93.$$

Under H_0 er testobservatoren standard normalfordelt, og vi beregner P-verdi

$$\text{p-val} = 2 \cdot P(Z < -1.93) = 2 \cdot 0.0268 = 0.0536.$$

Vi observerer P-verdi større enn statistisk signifikansnivå α, og forkaster ikke H_0. Dataene gir ikke tilstrekkelige bevis mot nullhypotesen til å forkaste den, og vi konkluderer at konsentrasjonen i prøven ikke er statistisk signifikant forskjellig fra 0.60.

2c

Punktestimat for μ er $\bar{x} = 0.57$. Med 95% konfidensnivå får vi feilmargin $m = z_{97.5} \cdot \sigma/\sqrt{2} = 0.03005$. Det gir 95% konfidensintervall for μ:

$$(0.5395, \ 0.6005).$$

Konfidensintervallet for μ er et intervall av mulige verdier for konsentrasjonen i prøven som er i samsvar med dataene (målingene). Videre er 95% konfidensintervallet konstruert med en metode som er slik at i 95% av tilfellene metoden blir brukt, vil det resulterende konfidensintervallet inneholde den sanne verdien av μ (her: konsentrasjonen i prøven).

(Fortsettes på side 3.)

Oppgave 3

Responsvariabel tannlengde (y_i), forklaringsvariabel daglig dose C-vitamin (x_i). Enkel lineær regresjonsmodell:

$$y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i, \quad i = 1, 2, \cdots, 60, \qquad \text{der } \epsilon_i \sim N(0, \sigma), \text{ uavhengige}$$

3a

En oversikt over de angitte populasjonsparameterne, med punkt-estimater og forklaringer finnes i Tabell 1. Leser også av $R^2 = 0.6443 = 64.43\%$, dvs at

Parameter	Estimat	Forklaring
Konstantledd β_0	$b_0 = 7.4225$	Forventa tannlengde ved 0mg C-vitamin daglig
Stigningstall β_1	$b_1 = 9.7636$	Økning i forventa tannlengde per mg C-vitamin daglig
Standardavvik σ	$s = 4.601$	Spredningsmål for feilledd ϵ_i (individuell variasjon)

Tabell 1: Oppgave 3a: populasjonsparametere, estimater og forklaringer

64.43% av variasjonen i tannlengde (y_i) forklares av daglig dose C-vitamin (x_i) i den enkle lineærmodellen.

3b

Hypotesene som tilhører t-testen som er gjort for stigningstallet β_1 i R-utskriften:

$$H_0 : \ \beta_1 = 0, \qquad H_a : \ \beta_1 \neq 0.$$

R utfører tosidig statistisk hypotesetest som standard.

Velger å gjennomføre statistisk hypotesetest på signifikansnivå $\alpha = 0.05$.

Standardisert test-observator er $T = \frac{b_1 - \beta_1}{\sigma_{b_1}}$.

Under H_0 er $\beta_1 = 0$, som gir testobservatoren verdi

$$T = \frac{9.7636 - 0}{0.9525} = 10.25.$$

Tilhørende p-verdi (lest av fra R) er $1.23 \cdot 10^{-14}$, som er mindre enn valgt statistisk signifikansnivå α (også for $\alpha = 0.01$, og ethvert rimelig forhåndsbestemt signifikansnivå). Vi observerer P-verdi$< \alpha$, og forkaster H_0 til fordel for H_a. Vi konkluderer at vi har observert en statistisk signifikant lineær effekt av daglig dose C-vitamin på tannlengden til marsvin.

3c

Forventet tannlengde når marsvinet har fått $x_i = 2$mg (dose) C-vitamin daglig er

$$\mu_i = \beta_0 + \beta_1 \cdot x_i = 7.4225 + 9.7636 \cdot 2 = 26.9497$$

(Fortsettes på side 4.)

3d

95% konfidensintervall for forventet tannlengde for et marsvin som har fått en daglig dose av 2mg C-vitamin: (24.96508, 28.9342).

Et 95% konfidensintervall for forventet tannlengde ved en angitt verdi av forklaringsvariabelen x er konstruert med en metode som i 95% av tilfellene metoden blir brukt vil inneholde den sanne forventningsverdien μ i underpopulasjonen der forklaringsvariabelen har verdi x.

95% prediksjonsintervall for tannlengde for et marsvin som har fått daglig dose av 2mg C-vitamin: (17.52801, 36.37127).

Et 95% prediksjonsintervall er konstruert med en metode som vil i 95% av tilfellene metoden blir brukt inneholde verdien y for en tilleggsobservasjon med kjent verdi av forklaringsvariabelen x.

Begge intervallene er sentrert i forventet tannlengde når marsvinet har fått $x_i = 2$mg (dose) C-vitamin daglig (beregnet i oppgave 3c), men har ulik bredde (prediksjonsintervallet har større feilmargin). Intervallene ender opp med ulik bredde fordi prediksjonsintervallet også inkluderer variabiliteten i en fremtidig observasjon om forventningsverdien til underpopulasjonen der $x_i = 2.0$, og prediksjonsintervallet for tannlengde blir derfor bredere enn konfidensintervallet for forventet tannlengde.

Oppgave 4

4a

Logistisk regresjonsmodell for sammenhengen mellom lommelykt til jul (y_i) og forklaringsvariablene alder (x_{i1}), forberedelsestid (x_{i2}) og ønskeliste (x_{i3}):

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \beta_3 \cdot x_{i3},$$

der p_i er sannsynligheten $P(y_i = 1)$ for lommelykt til jul for individ i med kjent verdi x_1, x_2, x_3 for forklaringsvariablene. Videre er

- β_1 den naturlige logaritmen av odds-ratioen for lommelykt til jul for en økning i alder på ett år
- β_2 er den naturlige logaritmen av odds-ratioen for lommelykt til jul for en økning i forberedelsestid på en dag, og
- β_3 er den naturlige logaritmen av odds-ratioen for lommelykt til jul for å ha sendt ønskeliste til nissen sammenlignet med å ikke ha gjort det.

4b

Tar utgangspunkt i (lest fra R-utskriften) at estimatet for β_1 er $b_1 = 0.24758$ med standardfeil $s_{b_1} = 0.09426$. Et 95% konfidentsintervall for β_1 er gitt ved $b_1 \pm z_{0.975} \cdot s_{b_1} = 0.24758 \pm 1.96 \cdot 0.09426 = (0.0628,\ 0.4323)$.

(Fortsettes på side 5.)

Med utgangspunkt i 95% konfidensintervallet for β_1, har vi at et 95% konfidensintervall for odds ratio for lommelykt til jul for en økning i alder på ett år, når alle andre forklaringsvariabler holdes fast:

$$(e^{0.0628},\ e^{0.4323}) = (1.064814,\ 1.540797).$$