UiT Norges arktiske universitet

Kort hjemmeeksamen i:	STA-0001, Brukerkurs i statistikk 1
Dato:	1.juni 2021
Tidspunkt:	09:00-13:00 + 30 minutter til innlevering i WISEflow
Kursansvarlig:	Elinor Ytterstad
Antall sider:	5 sider inklusive forsiden.
Support:	Du kan ringe 776 20 880 for support på eksamensdagen.
Vekting av spørsmål, eller annen informasjon:	Alle 10 delspørsmål (a),b), osv.) vektes likt ved bedømming.
Viktig informasjon om sitering og plagiering:	1. Dette er en individuell eksamen som skal besvares uten samarbeid med andre. 2. Alle hjelpemidler er tillatt (egne notater, pdfer fra forelesningene, lærebok, internett etc). 3. Alle eksamener som leveres i WISEflow blir automatisk sjekket for plagiat. Det er ikke tillatt å kopiere medstudenter, nettressurser, kilder, eller litteratur uten referanser.

I oppgaver der det eksplisitt står at en skal vise mellomregninger, vil korrekt svar ikke bli godkjent dersom mellomregninger mangler.

Det brukes desimalpunktum i dette oppgavesettet.
Oppgavesettet består av 10 delpunkter som alle teller likt ved bedømming.

Oppgave 1

Covid-19 og hurtigtester (antigen-test).

La hendelsen C betegne at en person er smittet av Covid-19.
La hendelsen T betegne positivt resultat av en hurtigtest.

Det opplyses at hurtigtestens evne til å oppdage smitte er 0.80, og dens evne til å avsløre om en person er frisk er 0.95.

Dette kan formuleres slik:
$P(T \mid C) = 0.80$ (sensitivitet)
$P(\overline{T} \mid \overline{C}) = 0.95$ (spesifisitet).

Vi skal i denne oppgaven bruke at $P(C) = 0.01$.

a)
- Forklar med ord hva $P(C) = 0.01$ betyr i denne oppgaven.
- Vis at $P(\overline{T} \cap C) = 0.002$. Vis mellomregninger.

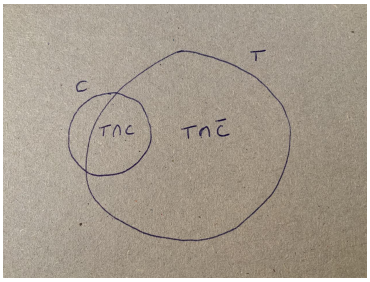
Løsningsforslag:

- $P(C) = 0.01$ betyr at 1 % av befolkningen er Covid-smittet.
- $P(\overline{T} \cap C) = P(\overline{T}|C) \cdot P(C) = (1 - P(T|C)) \cdot P(C) = (1 - 0.80) \cdot 0.01 = 0.002$

b)
- Lag et venndiagram og markér hendelsene $T \cap C$ og $T \cap \overline{C}$.
- Vis at $P(T) = 0.0575$. Her må du ta med mellomregninger.

Løsningsforslag:

- Venndiagram:



- $P(T) = P(T \cap C) + P(T \cap \overline{C}) =$
  $(P(C) - P(\overline{T} \cap C)) + P(T|\overline{C}) \cdot P(\overline{C}) =$
  $(0.01 - 0.002) + (1 - 0.95) \cdot (1 - 0.01) = 0.008 + 0.0495 = 0.0575.$

c)
- Finn sannsynligheten for at en Covid-smittet person skal teste negativt.
- Vis at $P(C \mid \overline{T}) = 0.0021$ (med fire desimalers nøyaktighet).
  Vis mellomregninger.
- Forklar med ord hva dette er sannsynligheten for.

Løsningsforslag:

- $P(\overline{T} \mid C) = 1 - P(T \mid C) = 1 - 0.80 = 0.2$
  Eller $P(\overline{T} \mid C) = \frac{P(\overline{T} \cap C)}{P(C)} = \frac{0.002}{0.01} = 0.2$
- $P(C \mid \overline{T}) = \frac{P(C \cap \overline{T})}{P(\overline{T})} = \frac{P(\overline{T}|C)P(C)}{1-P(T)} = \frac{(1-0.80)\cdot 0.01}{0.9425} = .0021,$
- som er sannsynligheten for at en person som har testet negativt, likevel er smittet.

Hver uke kjører 12000 vogntog over norskegrensen, dvs. 12000 utenlanske yrkessjåfører krysser grensen til Norge hver uke.

Antar alle 12000 yrkessjåfører hurtigtestes på grensen, og kun de med negativt test-resultat kan kjøre inn i landet.
Vi antar videre at 1% av yrkessjåførene som kommer til grensen, er Covid-smittet.

La Y være antall Covid-smittede med negativ hurtigtest ($\overline{T} \cap C$) av $n = 12000$ yrkessjåfører. Bruk fra tidligere at $P(\overline{T} \cap C) = 0.002$

d)
- Hvilken fordeling har Y? Oppgi navnet på fordelingen.
- Finn $E(Y)$.
- Bruk en tilnærmingsformel og finn $P(Y > 24)$.
  Forklar valg av fordeling i tilnærmingen.

Løsningsforslag:

- Y er binomisk fordelt med $n = 12000$ og $p = 0.002$.
- Binomisk forventning: $E(Y) = n \cdot p = 12000 \cdot 0.002 = 24$
- Binomiske sannsynligheter kan enten tilnærmes med poisson (fordi n er stor og p er liten), eller med normalfordeling (fordi varians $np(1-p) = 23.95 \geq 5$).
  Her er det beregningsmessig enklest å bruke normaltilnærming.
  Med heltallskorreksjon:
  $P(Y > 24) = 1 - P(Y \leq 24) = 1 - G(\frac{24+.5-24}{\sqrt{23.95}}) = 1 - G(0.10) = 1 - 0.5398 = 0.4602$.
  Uten heltallskorreksjon;
  $P(Y > 24) = 1 - P(Y \leq 24) = 1 - G(\frac{24-24}{\sqrt{23.95}}) = 1 - G(0) = 1 - 0.5 = 0.5$.

Oppgave 2

Anta fødselsvekt (X) til (levendefødte) barn i Norge er normalfordelt med forventningsverdi 3500 gram og standardavvik 500 gram.

a)
- Finn $P(X < 2500)$, og vis mellomregninger.
- Finn $P(X \leq 2499)$, og vis mellomregninger.
- Finn sannsynlighet for at et tilfeldig valgt barn hadde en fødselsvekt på mer enn 4 kg. Vis mellomregninger.

Løsningsforslag:

- $P(X < 2500) = G(\frac{2500-3500}{500}) = G(-2) = 0.0228$
- $P(X \leq 2499) = G(\frac{2499-3500}{500}) = G(-2.002) \approx G(-2.00) = 0.0228$
- $P(X > 4000) = 1 - G(\frac{4000-3500}{500}) = 1 - G(1) = 1 - 0.8413 = 0.1587$

b) Det har vært fire fødsler på ei fødestue en dag.

- Finn sannsynligheten for at gjennomsnittsvekten av disse $n = 4$ barna er mindre enn 3000 gram. Vis mellomregninger.

Løsningsforslag:

- $P(\bar{X} < 3000) = G(\frac{3000-3500}{\frac{500}{\sqrt{4}}}) = G(-2) = 0.0228$

c) Det er misstanke om at fødselsvekt i et bestemt geografisk område er lavere enn det normale for Norge ($\mu_0 = 3500$), noe vi skal undersøke med hypotesetest.
Bruk kjent $\sigma = 500$ og 1 % signifikansnivå.

Datamaterialet er fra en liten kommune i det aktuelle geografiske området, der det i fjor ble født $n = 25$ barn som i gjennomsnitt veide 3256 gram.

- Formulér hypoteser.
- Regn ut testobservatoren. Vis mellomregninger.
- Konkludér i hypotesetesten og formuler med ord hva konklusjonen sier om fødselsvekt.
- Hvor stort må utvalget være dersom en med 90% sannsynlighet skal kunne forkaste H_0 om forventet fødselsvekt er 100 gram lavere enn i landet forøvrig (altså $\mu_1 = 3400$)?

Løsningsforslag:

- $H_0 : \mu \geq 3500$ mot $H_1 : \mu < 3500$
- Bruker Z-testen siden vi har kjent σ.
  $Z = \frac{3256-3500}{\frac{500}{\sqrt{25}}} = \frac{-244}{100} = -2.44$
- Forkastingsområdet for 1 % signifikansnivå er: $(\infty, \quad -2.326]$.
  Verdien av testobservatoren er innenfor dette forkastingsområdet, og dermed forkastes nullhypotesen.
  Det er grunnlag i data for å hevde at fødselsvekten (i gjennomsnitt) er lavere i det aktuelle geografiske området sammenlignet med landet forøvrig.
- $n = (z_{0.01} + z_{0.10})^2(\frac{500}{3500-3400})^2 = (2.326 + 1.282)^2(5)^2 = 325.4$
  Dvs utvalget må være av størrelse 326 som et minimum.

d) I dette delspørsmålet skal vi bruke datamaterialet fra c), men ikke anta kjent standardavvik.
Vi har følgende: $n = 25, \quad \bar{x} = 3256, \quad \sum_{i=1}^{25}(x_i - \bar{x})^2 = 5412802$

- Vis at $s = 474.9$. Her må du inkludere mellomregninger.
- Finn et 99 % konfidensintervall for forventet fødselsvekt.
- Forklar med ord hva konfidensintervallet sier oss om forventet fødselsvekt. Bruk maksimalt tre setninger.

Løsningsforslag:

- $s^2 = \frac{\sum(x_i-\bar{x})^2}{25-1} = \frac{5412802}{24} = 225533.4$
  $s = \sqrt{225533.4} = 474.9$
- Bruker t-fordeling med 24 frihetsgrader der $t_{0.005} = 2.797$.
  Nedre grense: $NG = 3256 - 2.797 \cdot \frac{474.9}{\sqrt{25}} = 2990.3$
  Øvre grense: $ØG = 3256 + 2.797 \cdot \frac{474.9}{\sqrt{25}} = 3521.6$
  Et 99 % konfidensintervall for μ er da: $[2990.3, \quad 3521.6]$
- Konfidensintervallet $[2990.3, \quad 3521.6]$ er et intervall-estimat for forventet fødselsvekt μ i det aktuelle geografiske området.
  Det er 99 % sjanse for at intervallet $[2990.3, \quad 3521.6]$ inneholder sann (men ukjent) verdi av μ.

Oppgave 3

I denne oppgave skal vi undersøke med hypotesetest om sjansen for tidligfødsel (før uke 37) er avhengig av om det er kvinnens første fødsel eller om hun har født tidligere.

Datamaterialet består at 2708 levendefødsler ved to fødeklinikker i Oslo i 1964.

Svangerskapslengde	Førstegangsfødende	Flergangsfødende	Totalt
Før uke 37	100	72	172
37 uker eller senere	1188	1348	2536
Totalt	1288	1420	2708

a)
- Hva heter testen som må benyttes her?
- Vis med mellomregninger at forventet antall førstegangsfødende før uke 37 er 81.8.
- Bruke dette svaret (81.8) til å finne alle de tre andre forventede verdiene i tabellen.

Løsningsforslag:

- Kjikvadrattest
- $E_{11} = R_1 \frac{K_1}{2708} = 172 \frac{1288}{2708} = 81.808$
- Fordi rad- og kolonnesummer av forventede verdier (E_{ij}) skal være lik rad- og kolonnesummer av observerte verdier (X_{ij}) så finner vi at:
  $E_{12} = 172 - 81.8 = 90.2$
  $E_{21} = 1288 - 81.8 = 1206.2$
  $E_{22} = 2536 - 1206.2 = 1329.8$

Svangerskapslengde	Førstegangsfødende	Flergangsfødende	Totalt
Før uke 37	100 (81.8)	72 (90.2)	172
37 uker eller senere	1188 (1206.2)	1348 (1329.8)	2536
Totalt	1288	1420	2708

b)
- Regn ut testobservatoren og vis med mellomregninger at den er lik 8.24.
- Hva heter fordelingen til testobservatoren, og hvor mange frihetsgrader har den?
- Finn forkastingsområdet til testen og konkludér på 5 % signifikansnivå.
- Hva sier testens konklusjon om tidligfødsler og førstegangs-/flergangsfødende?

- Dersom du får vite at p-verdien er 0.004, forklar hvordan det kan benyttes til å trekke konklusjon i hypotesetesten.

Løsningsforslag:

- $Q = \sum_{\text{alle } ij} \frac{(X_{ij}-E_{ij})^2}{E_{ij}} = \frac{(100-81.8)^2}{81.8} + \frac{(72-90.2)^2}{90.2} + \frac{(1188-1206.2)^2}{1206.2} + \frac{(1348-1329.8)^2}{1329.8} = 8.24$
- Kjikvadratfordeling med 1 frihetsgrad. Frihetsgrader beregnes slik: (antall rader-1)(antall kolonner -1)=(2-1)(2-1)=1
- Kjikvardatfordeling med 1 frihetsgrader (tabell): Forkast nullhypotesen dersom $Q \geq 3.84$
- Nullhypotesen forkastes på 5 % nivå. Datamaterialet indikerer at andel tidligfødsler er ulik hos førstegangsfødende og flergangsfødende.
- Dersom p-verdien er mindre enn signifikansnivået (her 0.05), forkastes nullhypotesen. p-verdi $= 0.004 < 0.05$ og nullhypotesen forkastes.