

# Eksamensløsninger - HSTAT1101 - november 2024

## Oppgave 1

Oppgaven spør etter standardaviket i en populasjonen, vi bruker derfor følgende formel for å regne ut standardaviket  $\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$ .

$$\mu = 5 \text{ (gjennomsnitt)}$$

$$\sigma^2 = 6.666 \text{ (varians)}$$

$$\sigma = 2.582 \text{ (standardavvik)}$$

## Oppgave 2

a)

- Kontinuerlige variabler er numeriske (tall og kan telles på), og kan ta alle verider i et intervall.
- Kategoriske er oppdelt i distinkte kategorier eller grupper.
- Hvorvidt en variabel er kategorisk eller kontinuerlig har også noe å si for hvilke statistiske metoder du skal bruke (for eksempel for fordelinger)

b)

- Konfidensintervallet sier noe om usikkerheten rundt estimatet og hvorvidt vi tror det sanne populasjonsgjennomsnittet ligger i dette intervallet.
- Du behøver standardavviket, antallet og et signifikansnivå for å regne ut konfidensintervallet. I tillegg må du anta en fordeling.

c)

- $\alpha = 5\%$
- $H_0$ : Kostnadene er lik
- $H_1$ : Kostnadene er ikke lik
- Antar normalfordeling (kritiske verdier fra z-fordelingen)

d)

- z-fordelingen er en såkalt standardnormalfordeling som har et gjennomsnitt på 0 og et standardavvik på 1.
- Alle normalfordelinger, kan per definisjon, blir omgjort til en z-fordeling gjennom konvertering av X til Z. Det betyr at vi kan regne ut sansynnligheter ved bruk av z-fordelingen, som er enklere ved hjelp av tabellen i boken, for alle mulige normalfordelinger.
- z-fordelingen har alle egenskapene til en normalfordeling (symmetrisk, median/snitt/typetall er likt), karakterisert ved gjennomsnitt og standardavvik.

e)

- En fordeling som baserer seg på diskrete variabler (tellevariabler).
- Fordelingen styres av  $\lambda$  som er forventningsverdien/gjennomsnittet

f)

- Hvert Bernoulli-forsøk har to utfall (suksess og ikke-suksess). De er altså gjensidig utelukkende og fulstendig dekkende.
- $P(\text{suksess})$  er den samme i alle Bernoulli-forsøkene
- De enkelte Bernoulli-forsøkene er (stokastisk) uavhengige

### Oppgave 3

g)

$$P(\text{tallkort}) = \frac{36}{52}$$

h)

$$P(\text{tallkort} \mid \text{rødt kort}) = \frac{P(\text{tallkort} \cup \text{rødt kort})}{P(\text{rødt kort})} = \frac{18}{26} = \frac{9}{13}$$

i)

$$P(\text{tre like}) = \frac{52}{52} \cdot \frac{3}{51} \cdot \frac{2}{50} = \frac{312}{132600} = 0.00235 = 0.235\%$$

Spørsmålet var sannsynligheten for å trekke tre like, ikke å trekke tre dronninger. Dronningeksempelet i oppgaven var bare et eksempel.

j)

Sannsynligheten for å trekke en dronning blant 52 kort:  $\frac{4}{52}$   
Sannsynligheten for å trekke et rødt kort blant 52 kort:  $\frac{26}{52}$  Men vi har uten tilbakelegging, som betyr at du kan trekke en rød dronning ved første trekk, som betyr at sannsynligheten for å trekke et rødt kort på trekk nr. 2 er:  $\frac{25}{51}$   
Disse to trekkene (da vi har uten tilbakelegging) er avhengige av hverandre. Det betyr at  $P(A|B) = P(A)$  ikke stemmer i dette tilfellet.

### Oppgave 4

$$P(Z \leq -1.25) = P(Z \geq 1.25) = 1 - P(Z \leq 1.25) = 1 - 0.8944 = 0.1056$$

Noter at dette er en ensidig test.

### Oppgave 5

Ved bruk av den kumulative binomiske fordelingen

$$P(X \leq 4) = 0.583$$

$$P(X \leq 2) = 0.151$$

$$0.583 - 0.151 = 0.432$$

Ved bruk av individuelle binomisk fordelingen

$$P(X = 4) = 0.2367$$

$$P(X = 3) = 0.1954$$

$$0.2367 + 0.1954 = 0.4321$$

## Oppgave 6

a)

Ved bruk av den kumulative Poisson-fordelingen

$$\begin{aligned}P(X \leq 10) &= 0.7060 \\P(X \leq 7) &= 0.3229 \\0.7060 - 0.3229 &= 0.3831\end{aligned}$$

b)

Ved bruk av den kumulative Poisson-fordelingen

$$P(X \leq 14) = 0.5176$$

## Oppgave 7

Formelen for konfidensintervallet for et gjennomsnitt er når vi må bruke t-fordelingen (pga vi bare vet standardavviket i utvalget).

$$\begin{aligned}X \pm t_{n-1, \alpha/2} \cdot \sqrt{\frac{s^2}{n}} \\4.5 \pm 2.660 \cdot \frac{12}{\sqrt{61}} \\(4.09, 4.91)\end{aligned}$$

Hadde man brukt z-fordelingen istedenfor hadde man brukt en kritisk faktor på 2.756 og fått (4.10, 4.90) som svar, men dette er altså feil da vi måtte bruke t-fordelingen pga vi bare visste standardavviket i utvalget.

## Oppgave 8

a)

Spørsmålet var uklart på eksamen. Man ble derfor gitt poeng også om man regnet ut  $0.7 \cdot 0.45$  for P(frukt og boller). Gjorde man det ville ikke sansynlighetene passe. Basert på informasjonen i oppgaven opp følgende tabell:

Table 1:

|                | P(frukt)                  | P(ikke frukt) |      |
|----------------|---------------------------|---------------|------|
| P(boller)      | $0.70 \cdot 0.25 = 0.175$ |               | 0.45 |
| P(ikke boller) |                           |               |      |
|                | 0.25                      |               | 1    |

Ved bruk av sansynlighetsreglene (komplementregelen) kan vi fylle ut resten av tabellen og dermed få ut de marginale og simultane sansynlighetene.

Table 2:

|               | P(frukt)                  | P(ikke frukt)          |                   |
|---------------|---------------------------|------------------------|-------------------|
| P(boller)     | $0.70 \cdot 0.25 = 0.175$ | $0.45 - 0.175 = 0.275$ | 0.45              |
| P(ikke lunsj) | $0.25 - 0.175 = 0.075$    | $0.75 - 0.275 = 0.475$ | $1 - 0.45 = 0.55$ |
|               | 0.25                      | $1 - 0.25 = 0.75$      | 1                 |

b)

Variansen er  $\sigma^2$  Ettersom vi vet at  $\sigma = 140$ , så blir variansen  $\sigma^2 = 140^2 = 19,600$

c)

Vi har fått oppgitt at undersøkelsen ble gjort blant et utvalg studenter. Informasjonen vi har fått utgitt er derfor for utvalget, ikke populasjonen. Men utvalget er relativt stort (250), så vi kan bruke z-fordelingen som en tilnærming til t-fordelingen selvom vi ikke vet populasjonsvariansene.

$$\begin{aligned}(X - Y) \pm z_{\alpha/2} \cdot \sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}} \\(155 - 85) \pm 1.645 \cdot \sqrt{\frac{30^2}{250} + \frac{25^2}{250}} \\70 \pm 1.645 \cdot \sqrt{\frac{900}{250} + \frac{625}{250}} \\70 \pm 1.645 \cdot \sqrt{6.1} \\70 \pm 1.645 \cdot 2.57 \\70 \pm 4.07(65.93, 74.07)\end{aligned}$$

Noter, bruk av t-fordelingen og antagelser på uavhengige og avhengige utvalg gav også poeng og vurdert likeverdig som å bruke z-fordelingen som en tilnærming. Så lenge begrunnelsen var rimelig, ble det gitt poeng.

d)

Vanskelig å få tak i hele populasjonen av praktiske (kostnader, logistikk etc) og etiske årsaker (kan f.eks ikke ta kontakt med hele populasjonen). Vi kan derfor bruke utvalgsdata og statistiske metoder for å kunne si noe om populasjonen.

e)

I gjentatte utvalg vil fordelingen til utvalgsgjennomsnittet gå mot en normalfordeling når antallet gjentatte utvalg blir stort nok.

f)

- $Y = \text{UiO}$
- $X = \text{UiB}$
- $H_0 : \mu_x = \mu_y$  vs.  $H_1 : \mu_x \neq \mu_y$
- $250 = N_x = N_y$
- $\mu_Y = 640$
- $\mu_X = 785$
- $140 = \sigma_y$
- $185 = \sigma_x$
- $\alpha = 0.01$
- $Z_{\text{kritisk}, \alpha/2} = 2.576$

$$Z_{\text{observert}} = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} = \frac{(785 - 640)}{\sqrt{\frac{185^2}{250} + \frac{140^2}{250}}} = 9.88$$

Ettersom den observerte Z-verdien er større enn den kritiske verdien forkaster vi nullhypotesen.

g)

- Begge bruker gjennomsnitt, standardavvik og størrelsen på utvalget/populasjonen
- I begge utregningene må du sette et signifikansnivå ( $\alpha$ )
- Konfidensintervallet sier noe om usikkerheten rundt utvalgsestimatet og hvorvidt vi tror det sanne populasjonsgjennomsnittet ligger i dette intervallet.
- I en hypotesetest setter man to hypoteser opp imot hverandre og regner på sannsynligheten for at en av dem er sann gitt visse forutsetninger. Basert på utfallet av hypotesetesten, avgjør man om man forkaster eller beholder en av hypotesene som sann.