Høgskolen i Molde

Testeksamen i

MAT110 Statistikk 1

Versjon 03

Eksamensdag	:	**Mai 2019**
Tid	:	**09:00 – 13:00 (4 timer)**
Faglærer/telefonnummer	:	**Molde + Kristiansund:** **Per Kristian Rekdal / 924 97 051**
Hjelpemidler	:	**KD + formelsamling** **+ generell ordbok** **(morsmål/norsk/engelsk i papirformat)**
Antall sider inkl. forsiden	:	**xx**
Målform	:	**Norsk (bokmål)**

Noen generelle råd:

- **Skriv rett inn. Ikke bruk så mye tid på kladding.**
- **Det er totalt 4 oppgaver.**

Hver uke er alle oppdrettsanlegg i Norge pålagt å rapportere til forvaltningsmyndighetene om hvor mange voksne hunnlus som i gjennomsnitt er på alle fiskene i anlegget. Hensikten med tellingen er tofoldig:

1. Ha en oversikt over spredningen av voksne hunnlus i norske oppdrettsanlegg.
2. Kunne regulere oppdrettsanlegg som har for høye lusetall. [1]

Forvaltningsmyndighetene opererer derfor med en lusegrense, a, som det sanne lusetallet må ligge under:

$$\boxed{\begin{aligned} a &= \text{lusegrense} \\ &= 0.2 \text{ voksne hunnlus i gjennomsnitt per fisk per anlegg} \end{aligned}} \quad (1)$$

Veterinærinstituttet har skrevet en rapport med tittelen:

"*Telling av lakselus - Hvordan forstå og håndtere usikkerheten i telleresultatene*"

som handler om hvordan den ukentlige tellingen av lakselus kan brukes til å forvalte lusegrensa. [2]

Figur 1: Oppdrettsanlegg. Lakselus. Rapport.

[1] Lusetall = forventet antall voksne hunnlus per fisk per oppdrettsanlegg.

[2] Denne rapporten finnes her. Den er også lastet opp på Canvas under "materiale for emnet". Vi anbefaler at man tar en titt på denne rapporten. Mye av eksamen er basert på denne rapporten med hensikt å aktualisere eksamen.

Oppgave 1: (sannsynlighetsteori - telling av lakselus)

Anta nå at vi ser på et oppdrettsanlegg med kun 3 merder.
En merd er en innhengning for oppdrettsfisk, se figur 2.
Vi gjør følgende *eksperiment*:

1. Nummerer de 3 merdene fra 1 til 3.
2. Én fisk velges tilfeldig fra hver av de 3 merdene.
   La i representere den valgte fisken fra merd nr. i, for $i = 1, 2, 3$.
3. For hver valgte fisk registreres det om den har lakselus på seg eller ikke.
   Notasjon:

$$l_i = \text{en tilfeldig valgt fisk fra merd nr. } i \text{ har lakselus} \tag{2}$$

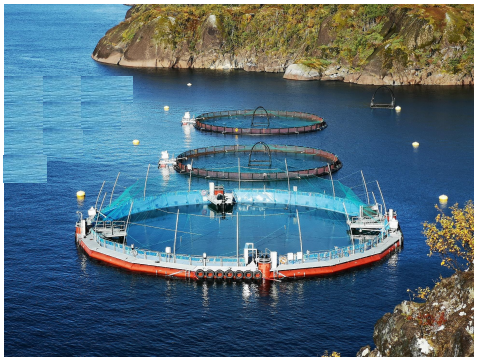
$$\overline{l_i} = \text{en tilfeldig valgt fisk fra merd nr. } i \text{ har } \underline{\text{ikke}} \text{ lakselus} \tag{3}$$

I vårt tilfelle er altså:

$$\text{eksperiment} = \text{sekvens med 3 forsøk} \tag{4}$$

Anta at vi kjenner:

$$p = \underline{\text{sannsynligheten}} \text{ for at en vilkårlig valgt fisk har } \underline{\text{lakselus}} \tag{5}$$



Figur 2: 3 merder.

Definer et utfall u av eksperimentet som:

$$u \;=\; (\; a_1, a_2, a_3 \;) \tag{6}$$

hvor $a_i \in \{l_i, \overline{l_i}\}$ for $i = 1, 2, 3$.

a) Definer utfallsrommet Ω_3 for eksperimentet. [3]
Bruk notasjonen som i lign.(6).

b) Er utfallsrommet Ω_3 diskret eller kontinuerlig?
Begrunn svaret kort.

Definer begivenheten:

$$S \;=\; \text{minst to av de trekte fiskene har lakselus} \tag{7}$$

c) Beskriv begivenheten S som en delmengde av utfallsrommet Ω_3.

d) Finn et uttrykk for sannsynligheten for begivenheten S beskrevet i lign.(7).
Anta $p = 0.5$. Regn ut sannsynligheten for begivenheten.

[3] Skriv gjerne opp utfallene eksplisitt.

For de resterende oppgavene ser vi på et annet oppdrettsanlegg som har 10 merder.
Anta de gjør det samme eksperimentet men at de trekker 10 fisker istedet for 3, en fra hver merd, og registrerer om fiskene har lakselus eller ikke. Definer et utfall u av eksperimentet som:

$$u \;=\; (\; a_1, a_2, \ldots, a_{10} \;) \tag{8}$$

hvor $a_i \in \{l_i, \overline{l_i}\}$ for $i = 1, 2, \ldots 10$.

Anta $p = 0.5$ for resten av oppgavene.

e) Definer utfallsrommet for Ω_{10} for dette eksperimentet. [4]

f) Er sannsynlighetsmodellen uniform? Begrunn svaret.

Hadde sannsynlighetsmodellen vært uniform dersom vi hadde hatt en annen verdi enn $p = 0.5$? Begrunn svaret.

g) Regn ut sannsynligheten for at nøyaktig 7 av de 10 trekte fiskene har lakselus. [5]

h) Anta vi får vite at fisk nr. 1 og fisk nr. 10 har lakselus.

Bruk betinget sannsynlighet for å regne ut sannsynligheten for at minst 8 av de 10 trekte fiskene har lakselus.

■

[4] Hint: *Ikke* skriv opp utfallene eksplisitt - bruk en formel for å definere utfallsrommet, dvs. $\Omega_{10} = \{$ formel $\}$.
[5] Hint: Bruk urnemodellen.

Oppgave 2: (stokastiske variabler og forventningverdi - telling av lakselus)

Avsnitt 4.1 i rapporten til Veterinærinstituttet:

"*For å beskrive fordelingen av lus mellom fiskene brukte vi et datasett fra 20 oppdrettsanlegg i Rogaland, med lusetall på enkeltfisknivå fra lusetellinger utført som ledd i daglig drift. Tellingene var utført fra og med uke 1 i 2012 til og med uke 37 i 2017. Til sammen hadde vi lusetellinger fra 352 228 enkeltfisker.*"

Fra dette datasettet ble det trekt ut et tilfeldig utvalg av fisker som kom fra merder med gjennomsnittlig antall hunnlus per fisk mellom 0.1 og 0.3. Det tilfeldige utvalget utgjorde 44 001 fisker. Gjennomsnittlig antall hunnluser per fisk $\overline{x}$ for dette utvalget samt tilhørende empirisk varians s_x^2 er vist i tabell 1. [6]

Antall fisker N	$\overline{\mathbf{x}}$	$\mathbf{s_x^2}$
44 001	0.1970	0.2970

Tabell 1: Nøkkeltall for utvalgte merder.

Figur 3: Rapport.

[6] Disse nøkkeltallene kommer fra rapporten og derfor faktiske nøkkeltall.

Populasjonsvariabel:

$$\boxed{X \quad = \quad \text{antall hunnlus på en } \underline{\text{tilfeldig valgt}} \text{ fisk fra et vilkårlig oppdrettsanlegg med lusetall mellom 0.1 og 0.3} \qquad (9)}$$

Forsøksvariabeler:

$$\boxed{X_i \quad = \quad \text{antall hunnlus på fisk nr. } i \text{ fra utvalget} \qquad (10)}$$

hvor

$$i \;=\; 1, 2, \ldots, N \tag{11}$$

med $N = 44\,001$ antall fisker.

I denne oppgaven antar vi en statistisk modell for forsøksvariablene basert på den *empiriske* fordelingen, dvs. vi antar:

$$\boxed{X_1, X_2, \ldots, X_N \overset{\text{i.i.d.}}{\sim} X \;\sim\; P(X = j) \qquad (12)}$$

hvor

$$j \;=\; 0, 1, 2, 3, 4, 5 \quad \text{(antall lakseluser)} \tag{13}$$

med fordelingen: [7]

	$j = 0$	$j = 1$	$j = 2$	$j = 3$	$j = 4$	$j = 5$
$P(X = j)$	0.85	0.12	0.0185	0.0070	0.0035	0.0010

Tabell 2: Sannsynlighetsfordeling $P(X = j)$.

[7] Vi antar her for enkelthetsskyld at det ikke ble registrert fisker med flere enn 5 lakselus. Tallene i tabell 2 er fiktive da selve rådataene ikke er tilgjengelige.

a) Forklar kort med ord hva sannsynligheten $P(X \leq j)$ betyr. [8]

b) Vis at $P(X = j)$ er en gyldig sannsynlighetsfordeling.

c) Regn ut forventningsverdien $E[X]$.

d) Regn ut variansen $Var[X]$ og standardavviket $\sigma[X]$.

[8] En setning er nok.

Et av de vanligste virkemidlene for å redusere lusetallet på er å gi fiskene medisin i fôret. [9] Anta at vi har funnet følgende sammenheng mellom mengden M av legemiddelet, som må blandes i fôret, og antall hunnlus per fisk:

$$\boxed{M = aX^2 + bX + c \qquad (14)}$$

hvor M er antall gram medisin per 100 gram fôr, og

$$a = 1.2 \tag{15}$$
$$b = 2.3 \tag{16}$$
$$c = 1.4 \tag{17}$$

e) Regn ut E[M].
Forklar med ord hva E[M] betyr.

Figur 4: Fôr.

[9] Fôr = mat for fisken.

Hver uke må hvert oppdrettsanlegg telle antall voksne hunnlus på 50 tilfeldig valgte fisker og rapportere lusetallet til Mattilsynet. Basert på gjennomsnittet av antall hunnlus per fisk for disse 50 fiskene må Mattilsynet vurdere om hvorvidt lusegrensa er overskredet eller ikke.

Anta vi ser på ett av disse oppdrettsanleggene for en vilkårlig uke t.

Forsøksvariablene er

$$\boxed{X_{kt} \;=\; \text{antall hunnlus på valgt fisk nr. } k \text{ i uke } t \qquad (18)}$$

hvor

$$k \;=\; 1, 2, \ldots, n \qquad \text{(antall fisk)} \tag{19}$$

med $n = 50$ antall fisk.

Anta videre at:

$$\boxed{X_{1t}, X_{2t}, \ldots, X_{50t} \overset{\text{i.i.d.}}{\sim} X \;\sim\; P(X = j) \qquad (20)}$$

hvor *i.i.d.* betyr uavhengige identiske fordelte variabler og hvor X er definert ved lign.(9) med fordeling som vist i tabell 2.

f) Regn ut sannsynligheten for at det målte lusetallet $\overline{X}_t$ til oppdrettsanlegget i uke t er lik eller ligger under lusegrensa a, dvs. regn ut:

$$\boxed{P(\ \overline{X}_t \leq a\ ) \qquad (21)}$$

hvor $a = 0.2$.

■

Oppgave 3: (statistisk modell og estimatorer)

I rapporten formulerte forfatterne to kandidater som mulig statistisk modell for forsøksvariablene X_i fra lign.(10), den ene var den velkjente Poisson-familien mens den andre var en annen diskret fordeling kalt *Negativ Binomial*:

Alternativ 1: Poisson (X og X_i er definert i lign.(9) og lign.(10))

$$\boxed{X_1, X_2, \ldots, X_N \overset{\text{i.i.d.}}{\sim} X \sim \text{Poi}[\lambda] \qquad (22)}$$

hvor

$$\lambda = \text{forventningsverdien} \tag{23}$$

Dersom $X \sim \text{Poi}[\lambda]$ så gjelder:

$$\begin{aligned} E[X] &= \lambda \\ Var[X] &= \lambda \end{aligned} \tag{24}$$

hvor $\lambda > 0$.

Alternativ 1: Negativ binomial

$$\boxed{X_1, X_2, \ldots, X_N \overset{\text{i.i.d.}}{\sim} X \sim \text{NBin}[\lambda, \theta] \qquad (25)}$$

hvor

$$\lambda = \text{forventningsverdien} \tag{26}$$

$$\theta = \text{spredningsparameteren} \tag{27}$$

Dersom $X \sim \text{NBin}[\lambda, \theta]$ så gjelder:

$$E[X] = \lambda \tag{28}$$

$$Var[X] = \lambda + \frac{\lambda^2}{\theta} \tag{29}$$

hvor $\lambda > 0$ og $\theta > 0$.

Poissonfordelingen, også kalt "*loven om sjeldne begivenheter*", er en statistisk fordeling som kan *utledes* dersom en del *forutsetninger* er oppfylt.

En viktig forutsetning som bør være oppfylt i den sammenheng er det som kalles *homogenitet* eller *ekvispredning*, dvs. at forventningsverdien og variansen er lik. Med $X \sim \text{Poi}[\lambda]$ har vi da:

$$\boxed{E[X] \;=\; Var[X] \;=\; \lambda \qquad (30)}$$

Dersom:

$$E[X] \quad < \quad Var[X]: \qquad \text{overspredning} \tag{31}$$
$$E[X] \quad > \quad Var[X]: \qquad \text{underspredning} \tag{32}$$

Det kan være flere grunner til overspredning i et datasett, men den mest vanlige er at det er *heterogenitet* i dataene. Heterogenitet betyr i prinsippet at dataene <u>ikke</u> er fremskaffet under like betingelser/omgivelser.

Dersom dataene viser tegn til overspredning, kan Negativ Binomial modellen være en passende modell å bruke siden den har en ekstra parameter i *spredningsparameteren* θ definert i lign.(27).

Fra lign.(28) og (29) ser vi at θ gjør at Negativ Binomial fordelingen har rom for overspredning.

Siden datasettet i rapporten faktisk viste tegn til overspredning, valgte forfatterne å bruke Negativ Binomial modellen.

Vi skal videre i denne oppgaven finne estimatorer for parameterne λ og θ og et asymptotisk $(1-\alpha)\,100\,\%$ konfidensintervall for λ for Negativ Binomial modellen definert i lign.(25).

Anta vi definerer følgende estimator $\hat{\lambda}$ for parameteren λ:

$$\boxed{\hat{\lambda} = \overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \qquad (33)}$$

a) Vis at $\hat{\lambda}$ i lign.(33) er forventningsrett.
Hva er realiseringen av $\hat{\lambda}$?

Vi skal nå konstruere en estimator $\hat{\theta}$ for θ.
Først definerer vi parameteren:

$$\sigma^2 = \lambda + \frac{\lambda^2}{\theta} \qquad (34)$$

Siden $Var[X] = \sigma^2$, kan vi bruke S_X^2 som estimator $\hat{\sigma}^2$ for σ^2.

b) Finn et uttrykk for $\hat{\theta}$ basert på lign.(34).
Regn ut realiseringen av $\hat{\theta}$?

c) Vis at [10]

$$\boxed{\left[\, LB_N^{\lambda} \,,\, UB_N^{\lambda} \,\right] \;=\; \left[\, \overline{X} \;-\; z_{1-\alpha/2}\sqrt{\frac{\overline{X} + \frac{\overline{X}^2}{\hat{\theta}}}{N}} \;,\; \overline{X} \;-\; z_{\alpha/2}\sqrt{\frac{\overline{X} + \frac{\overline{X}^2}{\hat{\theta}}}{N}} \,\right] \qquad (35)}$$

er et asymptotisk $(1-\alpha)\,100\,\%$ konfidensintervall for λ,
hvor $z_{\alpha/2}$ og $z_{1-\alpha/2}$ er de tilhørende kvantilene til normalfordelingen.

d) Vis at konfidensintervallet i lign.(35) kan skrives om på formen: [11]

$$\boxed{\left[\, LB_N^{\lambda} \,,\, UB_N^{\lambda} \,\right] \;=\; \left[\, \overline{X} \;-\; z_{1-\alpha/2}\,\frac{S_X}{\sqrt{N}} \;,\; \overline{X} \;-\; z_{\alpha/2}\,\frac{S_X}{\sqrt{N}} \,\right] \qquad (36)}$$

■

Figur 5: Lakselus.

[10] Se definisjon av asymptotiske $(1-\alpha)\,100\,\%$ konfidensintervall i formelsamlingen.
[11] Bruke uttrykket for $\hat{\theta}$ fra oppgave **2b**.

Oppgave 4: (krav som indikerer overskridelse av lusegrensa)

Vi skal i denne oppgaven se på kravene foreslått i rapporten som indikerer overskridelse av lusegrensa $a = 0.2$.

I oppgave **2** definerte vi forsøksvariablene fra en telling av 50 fisker på et vilkårlig oppdrettsanlegg en vilkårlig uke t:

$$\boxed{X_{kt} \;=\; \text{antall hunnlus på valgt fisk nr. } k \text{ i uke } t \qquad (37)}$$

hvor

$$k \;=\; 1, 2, \ldots, n \quad \text{(antall fisk)} \tag{38}$$

med $n = 50$ fisker.

Anta nå at vi bruker Negativ Binomial-modellen i lign.(25) fra oppgave **3** som statistisk modell for forsøksvariablene i lign.(37), dvs. anta:

$$\boxed{X_{1t}, X_{2t}, \ldots, X_{nt} \overset{\text{i.i.d.}}{\sim} X \;\sim\; \text{NBin}[\lambda, \theta] \qquad (39)}$$

Oppdrettsanlegget regner ut så gjennomsnittet $\overline{X}_t$ for uke t og sender resultatet til Mattilsynet. Mattilsynet vurderer deretter, på bakgrunn av den målte verdien for $\overline{X}_t$, om lusegrensa er overskredet eller ikke.

Veterinærinstituttet har foreslått *to* krav på $\overline{X}_t$ som Mattilsynet kan bruke når de skal vurdere om lusegrensa er overskredet eller ikke. Hovedmålet med disse kravene er å unngå feile konklusjoner av typen "*falske positiver*", dvs.

konkludere med at lusetallet er over lusegrensa, gitt at det sanne lusetallet ligger under

siden det å overskride lusegrensa har store øknomiske konsekvenser for oppdretterne.

De to kravene som Veterinærinstituttet foreslår er:

1. For å unngå *falske positiviteter* så ønsker man å teste $\overline{X}_t$ mot et kriterium på formen

$$\overline{X}_t \;>\; b \tag{40}$$

som er slik at det kun er $(1 - 0.96)\,100\,\% = 4\,\%$ sannsynlighet for at $\overline{X}_t$ er ligger over lusegrensen $a = 0.2$ gitt at det sanne lusetallet er under grensen, dvs. gitt at $E[X] < 0.2$.

2. Lusegrensa $a = 0.2$ anses som overskredet dersom lusetellingen ligger over lusegrensa 4 uker på rad, gitt at det sanne lusetallet ligger under, dvs. gitt at $E[X] < 0.2$.

a) Formuler krav nr. 1 matematisk. [12]

b) Formuler et matematisk uttrykk for *sannsynligheten* for at krav 2 inntreffer.

■

[12] Dette er en A-oppgave.