

Eksamen i

MAT110 Statistikk 1

Eksamensdag	: Tirsdag 21. 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	: 16
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

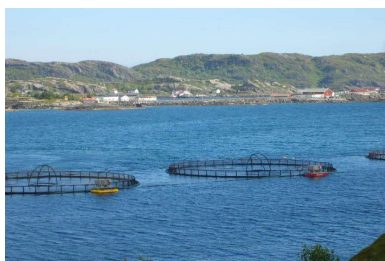
Forvaltningsmyndighetene opererer derfor med ei lusegrense, a , som det sanne lusetallet må ligge under:

$$\begin{aligned} a &= \text{lusegrense} \\ &= 0.5 \text{ voksne hunnlus i gjennomsnitt per fisk per anlegg} \quad (1) \end{aligned}$$

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.



Figur 1: Oppdrettsanlegg. Lakselus. Rapport.

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

Oppgave 1: (sannsynlighetsteori - telling av lakselus)

Anta nå at vi ser på et oppdrettsanlegg med kun 4 merder. En merd er en innhengning for oppdrettsfisk, se figur 2. Vi gjør følgende *eksperiment*:

1. Nummerer de 4 merdene fra 1 til 4.
2. Én fisk velges tilfeldig fra hver av de 4 merdene.
La i representere den valgte fisken fra merd nr. i , for $i = 1, 2, 3, 4$.
3. For hver valgte fisk registreres det om den har lakselus på seg eller ikke.
Notasjon:

$$\textcolor{red}{l}_i = \text{en tilfeldig valgt fisk fra merd nr. } i \text{ har lakselus} \quad (2)$$

$$\overline{\textcolor{green}{l}}_i = \text{en tilfeldig valgt fisk fra merd nr. } i \text{ har } \underline{\text{ikke}} \text{ lakselus} \quad (3)$$

I vårt tilfelle er altså:

$$\text{eksperiment} = \text{sekvens med 4 forsøk} \quad (4)$$

Anta at vi kjenner:

$$\textcolor{red}{p} = \underline{\text{sannsynligheten}} \text{ for at en vilkårlig valgt fisk har } \underline{\text{lakselus}} \quad (5)$$



Figur 2: 4 merder.

Utfallsrommet Ω for eksperimentet er:

$$\begin{aligned} \Omega = \bigg\{ & (\bar{l}_1, \bar{l}_2, \bar{l}_3, \bar{l}_4) , \quad (l_1, l_2, l_3, l_4) , \\ & (l_1, \bar{l}_2, \bar{l}_3, \bar{l}_4) , \quad (\bar{l}_1, l_2, l_3, l_4) , \\ & (\bar{l}_1, l_2, \bar{l}_3, \bar{l}_4) , \quad (l_1, \bar{l}_2, l_3, l_4) , \\ & (\bar{l}_1, \bar{l}_2, l_3, \bar{l}_4) , \quad (l_1, l_2, \bar{l}_3, l_4) , \\ & (\bar{l}_1, \bar{l}_2, \bar{l}_3, l_4) , \quad (l_1, l_2, l_3, \bar{l}_4) , \\ & (\bar{l}_1, \bar{l}_2, l_3, l_4) , \quad (l_1, l_2, \bar{l}_3, \bar{l}_4) , \\ & (\bar{l}_1, l_2, \bar{l}_3, l_4) , \quad (l_1, \bar{l}_2, l_3, \bar{l}_4) , \\ & (\bar{l}_1, l_2, l_3, \bar{l}_4) , \quad (l_1, \bar{l}_2, \bar{l}_3, l_4) \bigg\} \end{aligned} \quad (6)$$

- a) Er utfallsrommet Ω diskret eller kontinuert?
Begrunn svaret kort.
- b) Fra Ω i lign.(6) ser vi at det er 16 utfall i utfallsrommet.
Dersom det hadde vært n merder istedet for 4, hvor mange utfall hadde det vært i utfallsrommet da?
- c) Anta at eksperimentet gir følgende resultat:

$$l_1 , l_2 , l_3 , \bar{l}_4 \quad (7)$$

Skriv kort med ord hva resultatet betyr.

Anta $p = 0.5$.

Regn ut sannsynligheten for utfallet.

Begivenheten:

$$S = \text{minst halvparten av de trekte fiskene har lakselus} \quad (8)$$

er gitt ved

$$S = \left\{ \begin{array}{l} (l_1, l_2, l_3, l_4) \text{ ,} \\ (\bar{l}_1, l_2, l_3, l_4) \text{ ,} \\ (l_1, \bar{l}_2, l_3, l_4) \text{ ,} \\ (l_1, l_2, \bar{l}_3, l_4) \text{ ,} \\ (l_1, l_2, l_3, \bar{l}_4) \text{ ,} \\ (\bar{l}_1, \bar{l}_2, l_3, l_4) \text{ ,} \quad (l_1, l_2, \bar{l}_3, \bar{l}_4) \text{ ,} \\ (\bar{l}_1, l_2, \bar{l}_3, l_4) \text{ ,} \quad (l_1, \bar{l}_2, l_3, \bar{l}_4) \text{ ,} \\ (\bar{l}_1, l_2, l_3, \bar{l}_4) \text{ ,} \quad (l_1, \bar{l}_2, \bar{l}_3, l_4) \end{array} \right\} \quad (9)$$

- d) Finn et uttrykk for sannsynligheten for begivenheten S beskrevet i lign.(9), dvs. finn $P(S)$ uttrykt i ” p ”.
- e) Anta at vi vet at minst halvparten av de trekte fiskene har lakselus. Regn ut sannsynligheten for at minst to av de trekte fiskene *ikke* har lakselus. Bruk at $p = 0.5$.²

²Med begivenheten:

$$A = \text{minst to av de trekte fiskene ikke har lakselus} \quad (10)$$

kan man vise at

$$P(A \cap S) = 6 \cdot 0.5^4 \quad (11)$$

Bruk dette for å løse oppgave 1e.

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 0.4 og 0.6. Det tilfeldige utvalget utgjorde 7971 fisker. Gjennomsnittlig antall hunnluser per fisk $\bar{x}$ for dette utvalget samt tilhørende empirisk varians s_x^2 er vist i tabell 1.³

Antall fisker N	$\bar{x}$	s_x^2
7971	0.499	0.894

Tabell 1: Nøkkeltall for utvalgte merder.



Figur 3: Rapport.

³Disse nøkkeltallene kommer fra rapporten og derfor faktiske nøkkeltall.

Populasjonsvariabel:

$$X = \text{antall hunn lus} \text{ på en tilfeldig valgt fisk fra et vilkårlig oppdrettsanlegg} \quad (12)$$

med lusetall mellom 0.4 og 0.6

Forsøksvariabler:

$$X_i = \text{antall hunn lus} \text{ på fisk nr. } i \text{ fra utvalget} \quad (13)$$

hvor

$$i = 1, 2, \dots, N \quad (14)$$

med $N = 7971$ antall fisker.

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

$$X_1, X_2, \dots, X_N \stackrel{\text{i.i.d.}}{\sim} X \sim P(X = j) \quad (15)$$

hvor

$$j = 0, 1, 2, 3 \quad (\text{antall lakseluser}) \quad (16)$$

med fordelingen: ⁴

	$j = 0$	$j = 1$	$j = 2$	$j = 3$
$P(X = j)$	0.72	0.13	0.08	0.07

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

⁴Vi antar her for enkelthetsskyld at det ikke ble registrert fisker med flere enn 3 lakseluser. Tallene i tabell 2 er fiktive da selve rådataene ikke er tilgjengelige.

- a) Forklar kort med ord hva sannsynligheten $P(X = j)$ betyr. ⁵
- b) Vis at $P(X = j)$ er en gyldig sannsynlighetsfordeling.
- c) Regn ut forventningsverdien $E[X]$.
- d) Regn ut variansen $Var[X]$ og standardavviket $\sigma[X]$.

⁵En setning er nok.

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

$$X_{kt} = \text{antall hunnlus på valgt fisk nr. } k \text{ i uke } t \quad (17)$$

hvor

$$k = 1, 2, \dots, n \quad (\text{antall fisk}) \quad (18)$$

med $n = 50$ antall fisk.

Anta videre at:

$$X_{1t}, X_{2t}, \dots, X_{50t} \stackrel{\text{i.i.d.}}{\sim} X \sim P(X = j) \quad (19)$$

hvor *i.i.d.* betyr uavhengige identiske fordelte variabler og hvor X er definert ved lign.(12) med fordeling som vist i tabell 2.

- e) Bruk sentralgrensesetningen til å vise at sannsynligheten for at det målte lusetallet $\bar{X}_t$ til oppdrettsanlegget i uke t ligger over lusegrensa a , er 50 %, dvs. vis at:

$$P(\bar{X}_t > a) = 0.5 \quad (20)$$

hvor $a = 0.5$.

- f) Bør Mattilsynet konkludere at lusegrensa er overskredet dersom $\bar{X}_t > 0.5$? Begrunn svaret kort.

■

Oppgave 3: (statistisk modell og estimatorer)

I rapporten bestemte forfatterne seg for å formulere en statistisk modell for forsøksvariablene X_i fra lign.(13) basert på Poisson-familien av fordelinger, dvs. de antok at:

$$X_1, X_2, \dots, X_N \stackrel{\text{i.i.d.}}{\sim} X \sim \text{Poi}[\lambda] \quad (21)$$

hvor

$$N = 7971 \text{ antaller fisker} \quad (22)$$

$$\lambda = \text{forventet antall hunn lus per fisk} \quad (23)$$

(ukjent parameter som er positiv, dvs. $\lambda > 0$)

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

$$\hat{\lambda} = \bar{X} = \frac{1}{N} \sum_{i=1}^N X_i \quad (24)$$

a) Vis at $\hat{\lambda}$ i lign.(24) er forventningsrett.



Figur 4: Lakselus.

b) Vis at ⁶ ⁷

$$[LB_N^\lambda, UB_N^\lambda] = \left[\bar{X} - z_{1-\alpha/2} \sqrt{\frac{\bar{X}}{N}}, \bar{X} - z_{\alpha/2} \sqrt{\frac{\bar{X}}{N}} \right] \quad (27)$$

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

⁶Bruk at forventningen og standardavviket til $\hat{\lambda} = \bar{X}$ er, jfr. lign.(21), gitt ved:

$$E[\bar{X}] = \lambda \quad (25)$$

$$\sigma[\bar{X}] = \sqrt{\frac{\lambda}{N}} \quad (26)$$

⁷Se definisjon av asymptotiske $(1 - \alpha) 100\%$ konfidensintervall i formelsamlingen.

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:

$$E[X] = Var[X] = \lambda \quad (28)$$

Dersom:

$$E[X] < Var[X] : \quad \text{overspredning} \quad (29)$$

$$E[X] > Var[X] : \quad \text{underspredning} \quad (30)$$

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 ikke er fremskaffet under like betingelser/omgivelser.

- c) Viser dataene fra rapporten en tendens til *overspredning*?
Begrunn svaret kort. ⁸

⁸Tips: Bruk tabell 1.

For å ta høyde for overspredningen i datasettet foreslo forfatterne en ny statistisk modell basert på en annen diskret fordeling kalt *Negativ Binomial*:

$$X_1, X_2, \dots, X_N \stackrel{\text{i.i.d.}}{\sim} X \sim \text{NBin}[\lambda, \theta] \quad (31)$$

hvor

$$\begin{aligned} \lambda &= \text{forventningsverdien} \\ \theta &= \text{spredningsparameteren} \end{aligned} \quad (32)$$

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

$$\begin{aligned} E[X] &= \lambda \\ \text{Var}[X] &= \lambda + \frac{\lambda^2}{\theta} \end{aligned} \quad (33)$$

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

- d) Synes du negativ binomial fordelingen er en passende modell dersom en har **overspredning** i dataene?

Begrunn svaret kort.

■

Oppgave 4: (krav som indikerer overskridelse av lusegrensa)

Vi skal i denne oppgaven se på kravene foreslått i rapporten som indikerer overskridelse av lusegrensa $a = 0.5$.

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

$$X_{kt} = \text{antall hunnlus på valgt fisk nr. } k \text{ i uke } t \quad (34)$$

hvor

$$k = 1, 2, \dots, n \quad (\text{antall fisk}) \quad (35)$$

med $n = 50$ fisker.

Anta nå at vi bruker Poisson-modellen fra oppgave 3 som statistisk modell for forsøksvariablene i lign.(34), dvs. anta:

$$X_{1t}, X_{2t}, \dots, X_{nt} \stackrel{\text{i.i.d.}}{\sim} X \sim \text{Poi}[\lambda] \quad (36)$$

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

Veterinærinstituttet har foreslått noen krav på $\bar{X}_t$ 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 positive*”, dvs.

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

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

I rapporten foreslås to krav:

1. For å unngå *falske positiviteter* så ønsker man å teste $\bar{X}_t$ mot et kriterium som er slik at det kun er $(1 - 0.96) \cdot 100\% = 4\%$ sannsynlighet for at $\bar{X}_t$ er ligger over lusegrensen $a = 0.5$ gitt at det sanne lusetallet er under grensen, dvs. $E[X] < 0.5$. Man må da finne ei grense, si $b_{0.96}$, som oppfyller kravet:

$$P(\bar{X}_t > b_{0.96}) = 0.04 \quad (37)$$

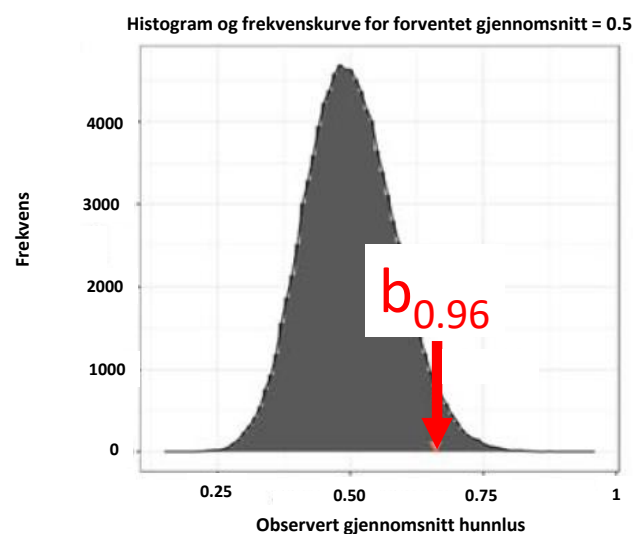
gitt at det sanne lusetallet er $E[X] < 0.5$, hvor

$$b_{0.96} = 96\% \text{-kvantilen til } \bar{X}_t \quad (38)$$

2. Lusegrensa $a = 0.5$ ansees som overskredet dersom lusetellingen ligger over lusegrensa 4 uker på rad, gitt at det sanne lusetallet ligger under, dvs. $E[X] < 0.5$. Begivenheten B som beskriver dette er:

$$B = \left\{ \bar{X} \mid \bar{X}_t > a \text{ og } \bar{X}_{t+1} > a \text{ og } \bar{X}_{t+2} > a \text{ og } \bar{X}_{t+3} > a \right\} \quad (39)$$

hvor $a = 0.5$ er lusegrensa.



Figur 5: Kvantilen $b_{0.96}$ til $\bar{X}_t$.

- a) Anta at ei lusetelling for et vilkårlig oppdrettsanlegg med $E[X] = \lambda = 0.49$ for ei vilkårlig uke t , gav $\bar{x}_t = 0.68$.

Vil Mattilsynet konkludere med at lusegrensa er overskredet dersom de legger krav 1 til grunn, dvs. lign.(37)?

Begrunn svaret ved regning. ⁹

- b) Anta at $E[X] = 0.49$, dvs. anta det sanne lusetallet ligger rett under lusegrensa. Anta også at lusetellingene er uavhengige av hverandre fra uke til uke.

Regn ut sannsynligheten $P(B)$ for at lusetellingen ligger over lusegrensa 4 uker på rad. ¹⁰

■

⁹Dette er en A-oppgave.

¹⁰Dvs. regn ut:

$$P(B) = P(\bar{X}_t > a \text{ og } \bar{X}_{t+1} > a \text{ og } \bar{X}_{t+2} > a \text{ og } \bar{X}_{t+3} > a) \quad (40)$$

Anta at $P(\bar{X}_t > a) = 0.5$.