UNIVERSITETET I OSLO

Matematisk Institutt

Eksamen i:	**STK 1100 – Sannsynlighetsregning og statistisk modellering**
ved:	**Nils Lid Hjort og Emil Stoltenberg**
Hjelpemidler:	**Godkjent kalkulator (samt vedlegget)**
Vedlegg:	**Formelsamling til STK1100/STK-FYS1110**
Tidspunkt:	**Mandag 1. juni 2026, 15:00 – 19:00**

Dette eksamenssettet inneholder fire oppgaver og er på tre sider. Vedlegget er altså ti siders Formelsamling (og som du med en viss høy sannsynlighet ikke vil trenge). Du kan levere din besvarelse på norsk, dansk, svensk, eller engelsk.

Oppgave 1: en fordeling på enhetsintervallet

Studer funksjonen gitt ved

$$f(x,\theta) = \begin{cases} 1+\theta(x-\frac{1}{2}) & \text{for } 0 \le x \le 1, \\ 0 & \text{ellers,} \end{cases}$$

der θ er en parameter i modellen.

(a) Hva er parameterområdet for denne θ, altså de verdier som gjør at $f(x,\theta)$ er en sannsynlighetstetthet? Skisser grafen til f, for verdiene -0.2, 0.0, 0.2.

(b) Med X som betegnelse på en stokastisk variabel trukket fra denne fordelingen, finn sannsynlighetene for at $X \in [0.0, 0.1]$ og $X \in [0.9, 1.0]$ (disse må få lov til å avhenge av θ).

(c) Finn en formel for forventningen for X.

(d) Vi har uavhengige observasjoner $X_1, \ldots, X_n$ fra fordelingen, med empirisk gjennomsnitt $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$. Lag en estimator for θ basert på dette, si $\widehat{\theta}_n$.

(e) Når samplestørrelsen n vokser, forklar kort hva som skjer med $\bar{X}_n$, og hva dette betyr for estimatoren $\widehat{\theta}_n$.

Oppgave 2: hvor ofte bruker du «z»?

I løpende tekst, på norsk eller andre språk, har de ulike bokstavene naturligvis nokså forskjellige grader av bruk. De mest sjeldent brukte bokstaver i norske tekster, for eksempel, er *q, x, z, w*. Vi skal ikke ut i lange analyser her (det gjør vi en annen dag), men la oss se på $Z(n)$, antallet ganger vi ser «z» i løpet av n tilfeldig samplede bokstaver, i løpende engelsk tekst. Det viser seg at slike variable, altså for de mindre brukte bokstaver, er tilnærmet Poisson-fordelte, når man sampler fra store tekstkorpora.

(a) La $X \sim \text{Pois}(\lambda)$, som betyr at X altså har Poisson-fordelingen, med parameter λ, med punktsannsynligheter

$$\Pr(X = x) = e^{-\lambda}\frac{\lambda^x}{x!} \quad \text{for } x = 0, 1, 2, \ldots.$$

Vis at dens momentgenererende funksjon blir

$$M(t) = \exp(\lambda(e^t - 1)).$$

(b) Bruk dette til å etablere at en slik X har forventning λ, og finn også variansen.

(c) Vis at summen av tre uavhengige Poisson-fordelte variable også blir Poisson-fordelt.

(d) For variabelen $Z(n)$ definert over, anta at denne er godt tilnærmet med en $\text{Pois}(n\lambda)$, der $\lambda = 0.07/100$, altså 0.07 prosent. Omtrent hvor mange «z» forventer du å se, i en lang tekst med 10000 bokstaver? Forklar at modellantagelsen er «logisk konsistent», ved at $Z(n_1) + Z(n_2) + Z(n_3)$, det totale antallet «z» over tre uavhengige utvalgte disjunkte tekster, med henholdsvis n_1, n_2, n_3 bokstaver, har samme fordeling som $Z(n_1 + n_2 + n_3)$. Sett opp et generelt utsagn om denne logiske konsistens.

(e) I Edgar Allan Poes berømte *The Raven* (fra 1845, klassisk skolepensum i siviliserte land), er det $n = 4868$ bokstaver (ja, vi har tellet). Det viser seg at han faktisk ikke brukte «z» i det hele tatt, i dette diktet. Hva er sjansen for det?

Oppgave 3: estimering av en forventning

La $X_1, \ldots, X_n$ være uavhengige og identisk normalfordelte variable med forventning μ og varians 1. Forventningen μ er et reelt tall, for oss ukjent, og som vi ønsker å estimere.

(a) Før vi starter arbeidet med å estimere akkurat denne μ ser vi kort på mse, the mean squared error eller forventet kvadrattap, i en generell situasjon. Hvis det er en parameter γ som skal estimeres, og estimatoren som vurderes er γ^*, vis at

$$\text{mse}(\gamma^*, \gamma) = \text{E}_\gamma\,(\gamma^* - \gamma)^2 = \text{Var}_\gamma\,\gamma^* + b_\gamma^2,$$

der $b_\gamma = \text{E}_\gamma\,\gamma^* - \gamma$ er biasen, skjevheten, og der fotskriften betegner at forventning og varians tas for den aktuelle γ; mse-en for γ^* er altså en funksjon av hvor vi er i parameterrommet.

(b) Vi går nå tilbake til situasjonen over, med de uavhengige $X_i \sim \text{N}(\mu, 1)$. En naturlig estimator her er $\widehat{\mu}_n = \bar{X}_n = (1/n)\sum_{i=1}^n X_i$, det empiriske gjennomsnittet. Finn forventning og varians for denne.

(c) Finn et uttrykk for mse-en for $\bar{X}_n$. Kommenter gjerne hvordan denne mse-en avhenger av den ukjente sannheten.

(d) Se nå på estimatoren $\lambda\bar{X}_n$, der $\lambda \in (0,1)$ er en justeringsparameter (et tall) statistikeren selv velger. Finn et uttrykk for $\text{mse}(\lambda\bar{X}_n, \mu)$, altså mse-en til denne estimatoren.

(e) Finn det intervallet av verdier for den ukjente μ der

$$\text{mse}(\lambda\bar{X}_n, \mu) \le \text{mse}(\bar{X}_n, \mu).$$

Hvordan påvirkes dette intervallet av justeringsparameteren, og hvordan påvirkes det av samplestørrelsen?

(f) Den verdien av λ som minimerer $\text{mse}(\lambda\bar{X}_n, \mu)$ vil avhenge av μ, er derfor ukjent, og følgelig til liten hjelp når det er μ vi skal estimere. Men det er likevel interessant å finne et utrykk for denne verdien, si λ_{opt}. Finn et slikt uttrykk for λ_{opt}, og forklar hvorfor dette uttrykk gir god statistisk mening.

(g) Parameteren λ_{opt} er altså ukjent, men lar seg estimere, ved å plugge inn $\bar{X}_n$ for μ. Skriv ned en formel for den dermed nye og potensielt smarte estimatoren $\mu_n^* = \widehat{\lambda}\bar{X}_n$. Hvordan vil du vurdere hvor smart den er?

Oppgave 4: gutter og jenter

Vi skal bry oss om sannsynligheten p for at et nyfødt barn er en gutt (og dermed også sannsynligheten $q = 1 - p$ for at det er en jente). Konkret skal vi se på et datamateriale, som vi har konstruert representativt via et større publisert datasett, for $n = 2000$ familier, hver med $m = 5$ barn. For variabelen X, antall gutter blant disse $N = nm = 10000$ barn, finner vi etter telling for dette datasettet at $x = 5150$.

(a) Forklar hvilke rimelige antagelser som må settes opp for at denne X skal være binomisk (N, p). Hva er forventning og varians for estimatoren $\widehat{p} = X/N$?

(b) Se nå på intervallet $[L, U]$, der

$$L = \widehat{p} - 1.96\sqrt{\frac{\widehat{p}(1-\widehat{p})}{N}}, \quad U = \widehat{p} + 1.96\sqrt{\frac{\widehat{p}(1-\widehat{p})}{N}}.$$

Gi argumenter som viser, under de binomiske antagelser, at dette blir et konfidensintervall for p med dekningsgrad tilnærmet 95 prosent. Beregn også $[L, U]$ for den observerte x, og kommenter ditt svar. (For å spare deg for et eksamensminutt opplyser vi at størrelsen $\sqrt{\widehat{p}(1-\widehat{p})/N}$ over, for de oppgitte x og N, er lik 0.0050, til fire desimalers presisjon.)

(c) Presiser de antagelser som skal til for at sannsynligheten for å få fem gutter, i en fem-barn-familie, er lik p^5. Av de $n = 2000$ familier, omtrent hvor mange vil ha fem gutter? Lag et 95 prosents konfidensintervall for np^5, og beregn dette, ut fra det som er funnet frem til over, for den observerte x.

(d) Det viser seg, for dette datasettet, at det er 85 slike familier, av de 2000, som har fem gutter. Kommenter dette.

Her slutter oppgavesettet