

UNIVERSITETET I OSLO

Matematisk Institutt

EKSAMEN I: **STK 1100 – Sannsynlighetsregning
og statistisk modellering**
VED: **Nils Lid Hjort og Emil Stoltenberg**
TIDSPUNKT: **Mandag 1. juni 2026, 15:00 – 19:00**

Løsning & randbemerkninger

Oppgave 1: en fordeling på enhetsintervallet

- (a) Funksjonen $f(x, \theta) = 1 + \theta(x - \frac{1}{2})$ integrerer seg pent til 1 over $[0, 1]$, fordi $\int_0^1 (x - \frac{1}{2}) dx = 0$. For at den skal holde seg ikke-negativ må $|\theta(x - \frac{1}{2})| \leq 1$ for disse x , og da er det ytterpunktene som gjelder: altså $|\frac{1}{2}\theta| \leq 1$, som betyr $\theta \in [-2, 2]$. Grafene er pene og lineære, og for $\theta = 0$ har vi vår gode venn den uniforme.
- (b) Det er like greit å finne den kumulative,

$$F(x) = \int_0^x f(u, \theta) du = x + \theta \int_0^x (u - \frac{1}{2}) du = x - \frac{1}{2}\theta(x - x^2);$$

dennes deriverte er ganske riktig $f(x, \theta)$. Så de to sannsynligheter blir

$$\begin{aligned} F(0.1) &= 0.1 - \frac{1}{2}\theta(0.1 - 0.1^2) = 0.1 - \frac{1}{2}\theta 0.09 = 0.1 - 0.045\theta, \\ 1 - F(0.9) &= 1 - (0.9 - \frac{1}{2}\theta(0.9 - 0.9^2)) = 0.1 + \frac{1}{2}\theta 0.09 = 0.1 + 0.045\theta. \end{aligned}$$

- (c) Forventningen blir riktig pen:

$$E X = \int x f(x, \theta) dx = \frac{1}{2} + \theta \int_0^1 x(x - \frac{1}{2}) dx = \frac{1}{2} + \theta/12.$$

- (d) Momentestimatoren setter det empiriske $\bar{X}_n$ lik forventningen, og løser for θ . Altså $\bar{X}_n = \frac{1}{2} + \theta/12$, og $\hat{\theta}_n = 12(\bar{X}_n - \frac{1}{2})$.
- (e) Vi vet temmelig godt hva som skjer med $\bar{X}_n$ for voksende n . (i) Ved LLN vil $\bar{X}_n \rightarrow_{\text{pr}} \frac{1}{2} + \theta/12$, som per enkel lineær transformasjon viser at $12\bar{X}_n - 6 \rightarrow_{\text{pr}} \theta$; estimatoren konvergerer i sannsynlighet mot den sanne verdi. (ii) Vi vet jo enda mer, per CLT, så

$$Z_n = \frac{\bar{X}_n - \frac{1}{2} - \theta/12}{\sigma/\sqrt{n}} \rightarrow_d N(0, 1),$$

der σ^2 er variansen til X_i . Gange-Rolv ganger med tolv og finner

$$\frac{12\bar{X}_n - 6 - \theta}{\sigma/\sqrt{n}} = \frac{\hat{\theta}_n - \theta}{\sigma/\sqrt{n}} \rightarrow_d 12 N(0, 1) \sim N(0, 12^2),$$

altså standardavvik 12 for denne grensen. Sensorene instrueres til å være tilfredse med essensen av (i), men flinkisene kunne jo kjøre også (ii), og forklare at dette leder til konfidens, av typen

$$\hat{\theta} \pm 1.96 \cdot 12\hat{\sigma}/\sqrt{n},$$

som krever at man også finner variansen over; den blir $\sigma^2 = (12 - \theta^2)/144$.

Oppgave 2: hvor ofte bruker du «z»?

- (a) Mgf-en til X blir

$$M(t) = E \exp(tX) = \sum_{x=0}^{\infty} e^{-\lambda} \frac{(\lambda e^t)^x}{x!} = \exp(-\lambda) \exp(\lambda e^t) = \exp(-\lambda + \lambda e^t).$$

- (b) Vi deriverer, finner $M'(t) = M(t)\lambda e^t$, og dermed $E X = M'(0) = \lambda$. Tilsvarende, via den annenderiverte $M''(t) = M'(t)\lambda e^t + M(t)\lambda e^t$, får vi

$$E X^2 = M''(0) = \lambda^2 + \lambda, \quad \text{og dermed } \text{Var } X = (\lambda^2 + \lambda) - \lambda^2 = \lambda.$$

- (c) La X, Y, Z være uavhengige Poisson-fordelte variable med parametere a, b, c . Da er mgf-en for summen lik

$$\begin{aligned} M_{X+Y+Z}(t) &= E \exp(t(X + Y + Z)) = M_X(t)M_Y(t)M_Z(t) \\ &= \exp(a(e^t - 1)) \exp(b(e^t - 1)) \exp(c(e^t - 1)) \\ &= \exp((a + b + c)(e^t - 1)). \end{aligned}$$

Men dette er sannelig mgf-en til $\text{Pois}(a + b + c)$, og mgf-er karakteriserer fordelinger entydig.

- (d) Modellen sier altså at $Z(n) \sim \text{Pois}(n\lambda)$, med $\lambda = 0.07/100$ (hvilket er et godt estimat, fra millioner av lange dokumenter i engelsk). For ti tusen ord har vi $Z(n) \sim \text{Pois}(7)$, så vi forventer omtrent 7, med et visst Poisson-slingringsmonn (standardavviket er $\sqrt{7} = 2.646$). Sum-egenskapen til Poisson, etablert over, viser at

$$\text{Pois}(n_1\lambda) + \text{Pois}(n_2\lambda) + \text{Pois}(n_3\lambda) \sim \text{Pois}((n_1 + n_2 + n_3)\lambda),$$

som jo stemmer prikk med modellens antagelser. Hadde ikke dette holdt, ville modellen vært dårlig. Mer generelt, hvis $A_1, \dots, A_k$ er passende adskilte dokumenter eller tekster, og vi teller $Z_1, \dots, Z_k$ z-er i disse, stemmer det presist som det skal, med

$$Z(A_1) + \dots + Z(A_k) \text{ samme fordeling som } Z(A_1 \cup \dots \cup A_k),$$

ser vi, der $Z(A) \sim \text{Pois}(\lambda|A|)$, der $|A|$ er antall bokstaver i dokumentet A .

- (e) La Z være antallet «z» i et herlig poem med $n = 4868$ bokstaver. Modellen sier at da bør $Z \sim \text{Pois}(3.408)$. Sjansen for at svaret er null er $p_0 = \exp(-3.408) = 0.033$. Dette er altså sjelden vare, men neppe over grensen til det ekstremt rare. Poe bruker «azure» og «lazily» i et annet kjent dikt, forresten. Nils har f.ø. vist at bokstavfordelingen hos Poe avviker signifikant fra «det normale». *Nevermore!*

Oppgave 3: estimering av en forventning

- (a) Joda: når estimatoren γ^* estimerer parameter γ , har vi som følger. La oss skrive

$$E_{\gamma} \gamma^* = \xi_{\gamma} = \gamma + b_{\gamma},$$

altså med biasen (eller skjevheten) $b_{\gamma} = \xi_{\gamma} - \gamma$:

$$\begin{aligned} \text{mse}(\gamma^*, \gamma) &= E_{\gamma} \{(\gamma^* - \xi_{\gamma}) - (\gamma - \xi_{\gamma})\}^2 \\ &= E_{\gamma} [(\gamma^* - \xi_{\gamma})^2 + (\gamma - \xi_{\gamma})^2] = \text{Var } \gamma^* + b_{\gamma}^2, \end{aligned}$$

idet produktleddet $(\gamma^* - \gamma)(\xi_{\gamma} - \gamma)$ jo har forventning null. Hei på deg, Pythagoras: $\text{mse}^{1/2}$ er hypotenusen i den trekant som har absolutt bias og standardavvik som de to kateter, og statistikere ønsker seg korte hypotenusen (ofte kalt root-mse, $\text{rmse} = \text{mse}^{1/2}$).

- (b) Tilbake til $X_1, \dots, X_n$, uavhengige fra $N(\mu, 1)$: gjennomsnittet $\bar{X}_n$ har forventning μ og varians $(1/n^2) \sum_{i=1}^n 1 = 1/n$ (og har dessuten selv en normalfordeling, men det trenger vi inntil videre ikke).
- (c) Vi bruker det generelle mse-resultatet over, og får

$$\text{mse}(\bar{X}_n, \mu) = \text{Var } \bar{X}_n + (E \bar{X}_n - \mu)^2 = 1/n,$$

siden den er forventningsrett. Så mse-funksjonen er konstant, enten μ er fjern eller nær eller 1066.

- (d) Den justerte estimator $\lambda \bar{X}_n$ får mse lik

$$\text{mse}(\lambda \bar{X}_n, \mu) = \text{Var}(\lambda X_n) + (E \lambda X_n - \mu)^2 = \lambda^2/n + (1 - \lambda)^2 \mu^2.$$

Denne er god og lav for $|\mu|$ ikke langt fra null, men eksploderer jo når $|\mu|$ vokser, og blir langt høyere enn $1/n$.

- (e) Så når er $\lambda^2/n + (1 - \lambda)^2 \mu^2 \leq 1/n$, der altså $\lambda \bar{X}_n$ har lavere risiko enn $\bar{X}_n$? Dette er det samme som

$$\mu^2 \leq \frac{1}{n} \frac{1 - \lambda^2}{(1 - \lambda)^2}, \quad \text{eller } |\mu| \leq \frac{1}{\sqrt{n}} \sqrt{\frac{1 + \lambda}{1 - \lambda}}.$$

Vinnerintervallet krymper mot null når n vokser. Det er stort, som er godt for $\lambda \bar{X}_n$, dersom nevneren i rot-uttrykket er lite, som betyr λ opp mot 1. Man kan f.eks. sjekke mse-ene for $\bar{X}_n$ og $0.99 \bar{X}_n$, for en aktuell n ; den andre vinner, i et kanskje stort område, men er farlig om μ er langt nok unna null.

- (f) Vi ser at $\text{mse}(\lambda \bar{X}_n, \mu)$ er en enkel parabel i λ for fast μ , og den er minst når den deriverte er null, altså $2\lambda/n - 2(1 - \lambda)\mu^2 = 0$, som betyr

$$\lambda_{\text{opt}} = \frac{\mu^2}{\mu^2 + 1/n} = \frac{n\mu^2}{n\mu^2 + 1}.$$

Hvis μ er liten skal det krympes mot null, og motsatt. Vekten går pent mot 1 når n vokser.

- (g) Ved å estimere λ_{opt} , ved å plugge inn $\bar{X}_n$ for μ , får vi en muligens smart ny estimator,

$$\mu^* = \hat{\lambda}_{\text{opt}} \bar{X}_n = \frac{n\bar{X}_n^2}{n\bar{X}_n^2 + 1} \bar{X}_n.$$

Den krymper altså standardestimatorens mot null. Hvor smart er den? Jo, vi kan beregne mse-en (poeng for bare å ha pekt på det, selv om vi ikke rekker å beregne den). – Det følgende går utenfor forventet besvarelse, men de ivrige kan jo kjøre på. I punktene over har vi egentlig bare brukt at X_i -ene er uavhengige med forventning μ og varians 1, men nå, for fin-beregning av mse og annet, kommer normalfordelingen inn. Vi har $Z_n = \sqrt{n}\bar{X}_n = \sqrt{n}\mu + N$, der N er standard normal, og vi har, om ikke annet, at

$$\begin{aligned} \text{mse}(\mu^*, \mu) &= E(\mu^* - \mu)^2 = E\left(\frac{Z_n^2}{Z_n^2 + 1} \frac{Z_n}{\sqrt{n}} - \mu\right)^2 \\ &= E\left\{\frac{(\sqrt{n}\mu + N)^2}{(\sqrt{n}\mu + N)^2 + 1} (\mu + N/\sqrt{n}) - \mu\right\}^2 \\ &= \int_{-\infty}^{\infty} \left\{\frac{(\sqrt{n}\mu + x)^2}{(\sqrt{n}\mu + x)^2 + 1} (\mu + x/\sqrt{n}) - \mu\right\}^2 \phi(x) dx, \end{aligned}$$

der $\phi(x)$ er standard-normal-tettheten. Dette gir ikke en lukket formel, men er lett nok å integrere numerisk. Dermed får man plottet mse-en, for aktuelle n , som funksjon av μ , og kan sammeligne med standard (med mse lik $1/n$) og med f.eks. $0.99 \bar{X}_n$, og atter andre konkurrenter. Denne nye estimatoren gjør det riktig bra: den er bedre enn $\bar{X}_n$ i et område rundt null, har risiko litt men ikke meget over $1/n$ i et par bumper vekk fra null, og for $|\mu|$ stor konvergerer den pent mot $1/n$.

Oppgave 4: gutter og jenter

- (a) Hvis mødrene i denne store populasjon er samplet representativt, har vi N uavhengige fødsler, hver med samme sannsynlighet p for gutt, antatt konstant. Det betyr at $X \sim \text{binom}(N, p)$, altså med $N = nm = 2000 \times 5 = 10000$. Estimatoren $\hat{p} = X/N$ kjennet vi godt; fra forventning og varians for den binomiske modellen (tja, det står kanskje i Formelsamlingen, hvem vet) får vi

$$E\hat{p} = p \quad \text{og} \quad \text{Var}\hat{p} = \frac{p(1-p)}{N}.$$

- (b) Vi vet at den binomiske X er tilnærmet normalfordelt, med meget god presisjon når N er så stor som dette. Så de to stokastiske brøker

$$Z = \frac{\hat{p} - p}{\sqrt{p(1-p)/N}} \quad \text{og} \quad Z' = \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/N}}$$

har begge fordelinger meget lik $N(0, 1)$. Der vet vi, som kjent, at $\Pr(-1.96 \leq N(0, 1) \leq 1.96) = 0.95$. Enkel symbolflytting viser at

$$1.96 \leq Z' \leq 1.96 \text{ er ekvivalent med } \hat{p} - 1.96 R \leq p \leq \hat{p} + 1.96 R,$$

der $R = \{\hat{p}(1 - \hat{p})/N\}$. Altså er det oppgitte $[L, U]$ et 95 prosenters konfidensintervall (med god presisjon, siden N er så stor, og siden p ikke ligger ute på kantene). For det observerte $x = 5150$, for de $N = 10000$ barn, finner vi $\hat{p} = 0.515$, og intervallet $[0.505, 0.525]$. Det bemerkelsesverdige er at intervallet ligger til høyre for 0.50; det er statistiske tegn på at barn født på vår jord har $p = \Pr(\text{gutt})$ litt større en 0.50.

- (c) Under de binomiske antagelser, dermed også innenfor hver familie, med konstant guttesannsynlighet p , er sannsynligheten for å få fem gutter lik p^5 . Estimatet for np^5 er $n\hat{p}^5 = 72.45$, og ved å transformere $[L, U]$ til $[nL^5, nU^5]$, som må ha samme konfidens 95 prosent, finner man $[65.82, 79.61]$.
- (d) Men oi!, virkeligheten er litt annerledes, for det er jo 85 forekomster av fem-gutter av de $n = 2000$ familier (spør ti meter Hjort). Tallet 85 er høyere enn det beregningene over tilsier. Hva kan dette skyldes? Matematikken er riktig, men something is rotten in the state of Binomialia: fødsler på vår jord signe dem er *ikke* uavhengige med prikk lik p for alle fruktsommelige. Den beste alternative forklaring er at p ikke er konstant i populasjonen; se *Overdispersed and Markovian Children* (Nils Lid Hjort, arXiv 2026, og på FocuStat-sidene), der det vises at p omtrent følger en Beta-fordeling, med senter i 0.515, men med en viss spredning rundt denne. Spredningen er ikke stor, med standardavvik omtrent 0.05, men den er altså merkbar når man ser på fødselsregistre med ti tusen eller enda flere barn. Det er ganske riktig også flere fem-jenter blant fembarnsfamilier enn det binomiske forventede $n(1 - p)^5 = 2000 \cdot 0.485^5$.

**Velkommen til flere kurs i statistikk & sannsynlighetsregning
& maskinlæring & data science & forsikringsmatematikk**