ECON2130: EKSAMEN 2016v

SENSORVEILEDNING.
Det anbefales at de 9 deloppgavene merket med A, B,… teller likt – uansett variasjon i vanskelighetsgrad. Svarene er gitt i << … >> Grensen til bestått bør ligge på ca 30-33%.

Oppgave 1

I en urne er det 10 kuler hvorav 5 er røde, 4 er blå og 1 er hvit. Kulene er ellers like store og tunge. Det trekkes ut et rent tilfeldig utvalg på 2 kuler fra urnen som betyr at alle mulige (ikke-ordnete) utvalg på 2 kuler er like sannsynlige.

Det kan vises (som du ikke trenger å gjøre) at utvalget blir rent tilfeldig hvis kulene trekkes en og en slik at alle de 10 kulene i urnen har samme sjanse for å bli trukket i første trekning, mens alle de 9 gjenværende kulene har samme sjanse for å bli trukket i annen trekning.

A. La, R_1, B_1, H_1 betegne begivenhetene at den første kulen som blir trukket er henholdsvis rød, blå eller hvit. La R_2, B_2, H_2 betegne de tilsvarende begivenhetene for den andre kulen som blir trukket ut. For eksempel, $R_1 \cap H_2$ betyr begivenheten at det blir trukket en rød kule i første trekning og en hvit i andre, mens $B_1 \cap B_2$ betyr at både første og andre kule som blir trukket ut er blå. $\bar{B}_1$ betyr at den første kulen som blir trukket ut ikke er blå.

Finn de 5 sannsynlighetene, $P(B_1)$, $P(\bar{B}_1)$, $P(B_2 | B_1)$, $P(B_2 | \bar{B}_1)$, $P(B_1 \cap B_2)$

<< Svar:

$$P(B_1) = \frac{4}{10} = 0.4, \qquad P(\bar{B}_1) = 1 - P(B_1) = 0.6, \qquad P(B_2 | B_1) = \frac{3}{9} = 0.33..., \qquad P(B_2 | \bar{B}_1) = \frac{4}{9} = 0.44...$$

$$P(B_1 \cap B_2) = P(B_1)P(B_2 | B_1) = \frac{4}{10} \cdot \frac{3}{9} = 0.133....$$

>>

B. La Y være antall blå kuler i et rent tilfeldig utvalg på 2 kuler.

i. Forklar hvorfor Y er hypergeometrisk fordelt og sett opp et uttrykk for sannsynlighetsfordelingen.

ii. Beregn sannsynlighetene $P(Y = y)$, $y = 0,1,2$. [**Hint.** Vis først at $\binom{10}{2} = 45$.]

iii. Beregn $E(Y)$, $\operatorname{var}(Y)$ ved, for eksempel, å bruke formlene for forventning og varians i en hypergeometrisk fordeling.

<< **Svar:** **i:** Populasjonen består av $N = 10$ kuler hvorav $M = 4$ er blå. Det trekkes et rent tilfeldig utvalg på $n = 2$ kuler hvorav Y er blå. Siden utvalget er rent tilfeldig, blir Y hypergeometrisk fordelt med punktsannsynligheter

$$P(Y = y) = \frac{\binom{4}{y}\binom{6}{2-y}}{\binom{10}{2}}, \quad y = 0,1,2$$

ii. Vi har $\binom{10}{2} = \frac{10 \cdot 9}{2} = 45$ og sannsynligheter gitt i tabellen

y	$\binom{4}{y}$	$\binom{6}{2-y}$	$P(Y = y)$
0	1	15	$\frac{15}{45} = 0.33...$
1	4	6	$\frac{24}{45} = 0.533...$
2	6	1	$\frac{6}{45} = 0.133...$

iii:

$$E(Y) = n\frac{M}{N} = 2\frac{4}{10} = 0.8, \qquad \text{var}(Y) = n\frac{M}{N}\left(1 - \frac{M}{N}\right)\frac{N-n}{N-1} = 2(0.4)(0.6)\frac{10-2}{9} = 0.4267$$

>>

C. La X være antall røde kuler i utvalget. Det kan vises (som du ikke trenger å gjøre) at den simultane fordelingen for X og Y er gitt i tabell 1.

Tabell 1 Den simultane fordelingen for *X* og *Y* gitt ved $f(x, y) = P(X = x \cap Y = y)$

		y			
		0	1	2	Sum
x	0	0	4/45	6/45	10/45
	1	5/45	20/45	0	25/45
	2	10/45	0	0	10/45
	Sum	15/45	24/45	6/45	1

i. Er X og Y stokastisk uavhengige? Begrunn svaret ditt.

ii. Vis at kovariansen mellom X og Y er $\text{cov}(X,Y) = -0.3556$.

iii. Finn sannsynlighetene $P(Y=1|X=0)$ og $P(Y=1|X=1)$.

<< **Svar:** **i:** De marginale fordelingene finnes i margen i tabellen. *X* og *Y* er ikke uavhengige hvis det fins en kombinasjon (x,y) slik at $P(X=x\cap Y=y)\neq P(X=x)P(Y=y)$. Dette gjelder f.eks. for kombinasjonen $(0,0)$ siden $P(X=0\cap Y=0)=0$, mens $P(X=0)P(Y=0)\neq 0$ i følge tabellen. Noen vil vise til at *X* og *Y* må være avhengige siden kovariansen er $\neq 0$, som bør godtas.

ii: Vi har $\text{cov}(X,Y)=E\left(XY\right)-E(X)\cdot E(Y)=E\left(XY\right)-1\cdot 0.8$, og tabellen gir

$E\left(XY\right)=\sum_x\sum_y xyf(x,y)=1\cdot 1\cdot\frac{20}{45}=0.4444...$, som gir $\text{cov}(X,Y)=0.4444-0.8=-0.3556$.

iii: $P(Y=1|X=0)=\dfrac{f(0,1)}{P(X=0)}=\dfrac{4/45}{10/45}=\dfrac{4}{10}=0.4$

$P(Y=1|X=1)=\dfrac{f(1,1)}{P(X=1)}=\dfrac{20/45}{25/45}=\dfrac{20}{25}=0.8$

>>

D. Mot å betale en spilleavgift på 10 kr. blir Jens tilbudt et spill som består i å trekke 2 kuler fra urnen som beskrevet i innledningen. For hver blå kule i utvalget får Jens 40 kr, men må betale 25 kr. for hver rød kule i utvalget. Hvis det dukker opp en hvit kule i utvalget, får ikke Jens noe for denne, men slipper også å betale noe. Fortjenesten til Jens, *V*, for et spill kan dermed uttrykkes som $V=40Y-25X-10$. De mulige fortjenestene som kan inntreffe i et enkelt spill er angitt i tabell 2.

Tabell 2 **Mulige fortjenester for et spill med utfall** (x,y)

		y		
		0	1	2
x	0	---	30	70
	1	–35	5	---
	2	–60	---	---

i. Finn sannsynligheten for at Jens får positiv fortjeneste i et spill (dvs. finn $P(V>0)$).

ii. Finn forventet fortjeneste i et spill, $E(V)$. Forklar kort hvordan du ville tolke dette resultatet.

iii. Vis at variansen til fortjenesten i et spill er, $\text{var}(V)=1671.556$.

[**Hint.** Variansen til *X*, som du ikke trenger å beregne her, oppgis til $4/9$]

<< **Svar:** **i:** Positiv fortjeneste skjer ved utfallene, $(0,1),(0,2)$ og $(1,1)$, som har sannsynlighet $P(V>0)=\frac{4+6+20}{45}=\frac{30}{45}=\frac{2}{3}$.

ii: Regel 4.12 i boka (utg.3) gir
$E(V)=40E(Y)-25E(X)-10=40(0.8)-25\cdot 1-10=32-25-10=-3$.
Siden forventet fortjeneste er et uttrykk for gjennomsnittlig fortjeneste i det lange løp når Jens spiller mange ganger, viser resultatet at Jens vil tape på spillet i lengden.

iii: Regel 4.15 i boka (utg.3) gir

$$\mathrm{var}(V)=40^2\,\mathrm{var}(Y)+(-25)^2\,\mathrm{var}(X)-2\cdot 40\cdot 25\cdot \mathrm{cov}(X,Y)=1600(0.4267)+625(4/9)-2000(-0.3556)$$

$$=1671.556$$

\>>

E. Jens spiller spillet i punkt **D** 30 ganger. La V_i betegne fortjenesten i spill nr. i, $i=1,2,\ldots,30$. Den totale fortjenesten er $T=V_1+V_2+\cdots+V_{30}$, der $V_1,V_2,\ldots,V_{30}$ antas å være uavhengige og identisk fordelte (*uid*) med samme fordeling som *V*.

i. Finn sannsynligheten tilnærmet for at *T* blir positiv (dvs. finn $P(T>0)$ tilnærmet). [**Hint.** Du trenger bl.a. $E(V)$. Hvis du ikke fant denne i punkt **D**, gjett på en verdi og bruk denne verdien i beregningen.]

ii. Anta at sannsynligheten for at et enkelt spill gir positiv fortjeneste er $2/3$. La *U* være antall ganger i løpet 30 spill V_i blir positiv ($V_i>0$). Finn sannsynligheten tilnærmet for at U blir minst 20 (dvs. finn $P(U\ge 20)$ tilnærmet).

<< **Svar:** **i.** I følge sentralgrenseteoremet er *T* tilnærmet normalfordelt, $N(30E(V),\ \sqrt{30}\ SD(V))=N\left(-90,\ \sqrt{30\cdot 1671.556}\right)=N(-90,\ 223.9345)$, hvorav (der $G(z)=P(Z\le z)$ for $Z\sim N(0,1)$):

$$P(T>0)=1-P\left(\frac{T+90}{223.9345}\le\frac{90}{223.9345}\right)\approx 1-G\left(\frac{90}{223.9345}\right)=1-G(0.40)=1-0.6554=0.3446$$

ii: Enkeltspillene kan ses på som uavhengige binomiske forsøk med suksessbegivenhet $S=(V_i>0)$ og konstant sannsynlighet $p=2/3$. *U* blir dermed binomisk fordelt, $U\sim\mathrm{bin}(30,\ 2/3)$, som har en varians $\mathrm{var}(U)=30\cdot\frac{2}{3}\cdot\frac{1}{3}=\frac{20}{3}>5$. Dette, ifølge regel 5.20 i boka, garanterer akseptabel tilnærmelse til normalfordelingen,

$U \overset{\text{tilnærmet}}{\sim} N\left(np,\ \sqrt{np(1-p)}\right)=N(20,\ 2.5820)$. Dermed (uten heltallskorreksjon):

$$P(U\ge 20)=1-P(U\le 19)\approx 1-P\left(Z\le\frac{19-20}{2.582}\right)=1-P\left(Z\le -0.39\right)=1-0.3483=0.6517\ .$$

Noen vil foretrekke heltallskorreksjon siden variansen ligger nær tilnærmelses-kriteriet, som vel bør berømmes. Dette gir

$$P(U \geq 20) = 1 - P(U \leq 19) = 1 - P(U \leq 19.5) \approx 1 - P\left(Z \leq \frac{19.5 - 20}{2.582}\right) =$$
$$= 1 - P\left(Z \leq -0.19\right) = 1 - 0.4247 = 0.5753$$

>>

Oppgave 2

Tabell 3 viser antall trafikkulykker med personskade i Norge for to perioder, mai 2015 (periode 1) og juni-juli 2015 (periode 2)[1].

Tabell 3

Periode	Mai 2015	Juni –juli 2015	Sum
Antall trafikkulykker med personskade	577	1274	1851

La X betegne antall trafikkulykker i mai og Y summen av antall ulykker i juni og juli der X og Y er sett på som stokastiske variable. Anta at X og Y er uavhengige og poissonfordelte med forventninger, λ og 2λ henholdsvis (kort skrevet: $X \sim Pois(\lambda),\ Y \sim Pois(2\lambda)$), der λ tolkes som forventet ulykkesrate pr. måned som antas å være den samme i de tre månedene i perioden mai – juli.

For enkelthets skyld, bortsett fra i punkt **D** nedenfor, antar vi også at de tre månedene er like lange, for eksempel 30 dager hver.

X og Y gir to uavhengige og forventningsrette estimatorer for λ, nemlig $\hat{\lambda}_1 = X$ og $\hat{\lambda}_2 = Y/2$.

A. **i.** Hvis $V \sim Pois(t\lambda)$ for en periode på t måneder, forklar kort hvorfor $\lambda^* = V/t$ er en forventningsrett estimator for λ med varians, $\mathrm{var}(\lambda^*) = \lambda/t$.

ii. Det foreslås to estimatorer for λ basert på $\hat{\lambda}_1$ og $\hat{\lambda}_2$:

$$\hat{\lambda} = \frac{\hat{\lambda}_1 + 2\hat{\lambda}_2}{3} \quad \text{og} \quad \tilde{\lambda} = \frac{\hat{\lambda}_1 + \hat{\lambda}_2}{2}$$

Vis at både $\hat{\lambda}$ og $\tilde{\lambda}$ er forventningsrette og at $\hat{\lambda}$ har minst varians lik $\lambda/3$.

iii. Beregn de to estimatene for λ basert på $\hat{\lambda}$ og $\tilde{\lambda}$ og velg det estimatet du mener er best. Angi hvorfor du mener valget ditt er best.

<< **Svar:** **i:** Siden $E(V) = \mathrm{var}(V) = t\lambda$, får vi $E(\lambda^*) = E\left(V/t\right) = E(V)/t = (t\lambda)/t = \lambda$

[1] Kilde SSB.

og $\text{var}(\lambda^*) = \text{var}(V/t) = \frac{\text{var}(V)}{t^2} = \frac{t\lambda}{t^2} = \frac{\lambda}{t}$

ii: Siden både $\hat{\lambda}_1$ og $\hat{\lambda}_2$ er forventningsrette med varianser henholdsvis, λ og $\lambda/2$, får vi

$$E(\hat{\lambda}) = E\left(\frac{\hat{\lambda}_1 + 2\hat{\lambda}_2}{3}\right) = \frac{E(\hat{\lambda}_1) + 2E(\hat{\lambda}_2)}{3} = \frac{3\lambda}{3} = \lambda$$

$$E(\tilde{\lambda}) = E\left(\frac{\hat{\lambda}_1 + \hat{\lambda}_2}{2}\right) = \frac{E(\hat{\lambda}_1) + E(\hat{\lambda}_2)}{2} = \frac{2\lambda}{2} = \lambda$$

$$\text{var}(\hat{\lambda}) = \text{var}\left(\frac{\hat{\lambda}_1 + 2\hat{\lambda}_2}{3}\right) = \frac{1}{9}\text{var}(\hat{\lambda}_1) + \frac{4}{9}\text{var}(\hat{\lambda}_2) = \frac{\lambda}{9} + \frac{4}{9}\frac{\lambda}{2} = \frac{3}{9}\lambda = \frac{\lambda}{3}$$

$$\text{var}(\tilde{\lambda}) = \text{var}\left(\frac{\hat{\lambda}_1 + \hat{\lambda}_2}{2}\right) = \frac{\text{var}(\hat{\lambda}_1) + \text{var}(\hat{\lambda}_2)}{4} = \frac{\lambda}{4} + \frac{\lambda}{8} = \frac{3\lambda}{8} > \frac{\lambda}{3} = \text{var}(\hat{\lambda})$$

iii: Estimater:

$$\hat{\lambda}_{obs} = \left[\frac{X + 2Y/2}{3}\right]_{obs} = \frac{577 + 1274}{3} = 617, \qquad \tilde{\lambda}_{obs} = \left[\frac{X + Y/2}{2}\right]_{obs} = \frac{577 + 637}{2} = 607$$

For to forventningsrette estimatorer foretrekkes den med minst varians. Dermed blir det mest troverdige estimatet det første, 617.
>>

B. Det kan vises (som du ikke trenger å gjøre) at $\frac{\hat{\lambda} - \lambda}{\sqrt{\hat{\lambda}/3}}$ er tilnærmet standard normalfordelt (N(0, 1)) uansett λ.

i. Bruk dette til å utlede et konfidensintervall for λ med konfidensgrad tilnærmet 0.90, basert på $\hat{\lambda}$.

ii. Beregn det observerte konfidensintervallet basert på data i tabell 3.

<< **Svar:** i: Hvis $U \sim N(0,1)$, er $P(-1.645 \le U \le 1.645) = 0.90$, der $1.645 = z_{0.05}$ er 5% kvantilen i N(0, 1). Dermed

$$0.90 \approx P\left(-1.645 \le \frac{\hat{\lambda} - \lambda}{\sqrt{\hat{\lambda}/3}} \le 1.645\right) = P\left(\hat{\lambda} - 1.645\sqrt{\frac{\hat{\lambda}}{3}} \le \lambda \le \hat{\lambda} + 1.645\sqrt{\frac{\hat{\lambda}}{3}}\right)$$, som gir et

tilnærmet 90% konfidensintervall: $\hat{\lambda} \pm 1.645\sqrt{\frac{\hat{\lambda}}{3}}$.

ii: Observert: $\left[\hat{\lambda} \pm 1.645\sqrt{\frac{\hat{\lambda}}{3}}\right]_{obs} = 617 \pm 1.645\sqrt{\frac{617}{3}} = 617 \pm 23.59 = [593,\ 641]$

>>

C. Vi ønsker å sjekke antakelsen om konstant ulykkesrate i hele tre-månedsperioden mai – juli med en statistisk test. For å få til dette utvider vi modellen litt ved å anta at forventet ulykkesrate for mai er λ_1 og for juni-juli λ_2, der λ_1 og λ_2 ikke behøver å være like. Vi ønsker å teste nullhypotesen at λ_1 og λ_2 er like, dvs.
$H_0 : \lambda_2 - \lambda_1 = 0$ mot $H_1 : \lambda_2 - \lambda_1 \neq 0$. Innfører vi parameteren $\delta = \lambda_2 - \lambda_1$, kan vi skrive hypotesen vi ønsker å teste som $H_0 : \delta = 0$ mot $H_1 : \delta \neq 0$

Forventningsrette estimatorer, $\hat{\lambda}_1$ og $\hat{\lambda}_2$, er gitt i innledningen. La σ^2 betegne variansen til $\hat{\delta} = \hat{\lambda}_2 - \hat{\lambda}_1$ som avhenger av de ukjente λ_1 og λ_2. Estimatoren $\hat{\sigma}^2$ er dannet ved å erstatte de ukjente λ_1 og λ_2 med sine respektive estimatorer.

i. Forklar kort hvorfor $\hat{\delta} = \hat{\lambda}_2 - \hat{\lambda}_1$ er tilnærmet normalfordelt.

ii. Det kan vises (som du ikke trenger å gjøre) at $Z = \dfrac{\hat{\delta}}{\hat{\sigma}}$ er tilnærmet standard normalfordelt, N(0, 1), dersom $\lambda_1 = \lambda_2$ (dvs. $\delta = 0$). Bruk dette til å sette opp et forkastningskriterium for en test for H_0 med signifikansnivå tilnærmet 0.05.

iii. Gjennomfør testen i **ii.** og formuler en konklusjon.

<< **Svar:** **i:** I følge regel 5.20 i boka er både *X* og *Y* tilnærmet normalfordelte med varianser godt over 5 som er tommelfingerkriteriet for normaltilnærmelsen. Siden en lineærkombinasjon av uavhengige normalfordelte variable selv er normalfordelt, må $\hat{\delta} = \hat{\lambda}_2 - \hat{\lambda}_1 = Y/2 - X$ være tilnærmet normalfordelt.

ii: Forkast H_0 hvis $Z \leq -1.96$ eller $Z \geq 1.96$, der $1.96 = z_{0.025}$ er 2.5%-kvantilen i N(0, 1).

iii: $\hat{\delta}_{obs} = \left[\hat{\lambda}_2 - \hat{\lambda}_1\right]_{obs} = \left[Y/2 - X\right]_{obs} = 637 - 577 = 60$

$\sigma^2 = \mathrm{var}(\hat{\delta}) = \mathrm{var}\left(\hat{\lambda}_2 - \hat{\lambda}_1\right) = \lambda_2/2 + \lambda_1$ og $\hat{\sigma}^2 = \hat{\lambda}_2/2 + \hat{\lambda}_1 = Y/4 + X$

Observert: $\hat{\sigma}_{obs} = \sqrt{Y_{obs}/4 + X_{obs}} = \sqrt{1274/4 + 577} = 29.925$

$Z_{obs} = \dfrac{60}{29.925} = 2.01$ → Konklusjon: Forkast H_0. Dvs. på 5% (tilnærmet) nivå er det evidens i data for at antakelsen om konstant ulykkesrate er feil.
>>

D. Til tross for eventuell forkastning av H_0 i punkt **C.**, går vi nå tilbake til den opprinnelige modellen med konstant forventet ulykkesrate pr. tidsenhet i hele tre-månedsperioden, men ønsker å justere for at månedslengden varierer i perioden. I alt omfatter perioden 92 dager siden mai og juli har 31 dager hver. Istedenfor λ vil vi nå estimere θ definert som forventet ulykkesrate pr. 30 dager uansett måned.

En måte å gjøre dette på er ved å forutsette at $T = X + Y$ er poisson-fordelt med forventning 92μ, der $\mu = \theta/30$ er forventet ulykkesrate pr. dag som antas å være konstant i perioden.

i. Sett opp en forventningsrett estimator, $\hat{\theta}$, for θ og finn et uttrykk for variansen til denne.

ii. Beregn estimatet basert på $\hat{\theta}$ og standardfeilen, $SE(\hat{\theta})$ (som betyr estimert standardavvik for $\hat{\theta}$).

iii. Beregn også et tilnærmet 95% konfidensintervall basert på at $\frac{\hat{\theta}-\theta}{SE(\hat{\theta})}$ er tilnærmet standard normalfordelt (som du ikke trenger å begrunne her).

<< **Svar:** **i:** En forventningsrett estimator for μ er $\hat{\mu}=T/92$ med varians $\mu/92$. Siden $\theta=30\mu$, blir $\hat{\theta}=30\hat{\mu}=30T/92$ forventningsrett med varians

$\mathrm{var}(\hat{\theta})=900\mathrm{var}(\hat{\mu})=900\mu/92=30^2\,\theta/(30\cdot 92)=\frac{30}{92}\theta$ (eller $\frac{900}{92}\mu$ som også bør godtas).

ii: Estimat: $\hat{\theta}_{obs}=\left[\frac{30}{92}T\right]_{obs}=\frac{30\cdot 1851}{92}=603.6$

Standardfeil = $\sqrt{\left[\frac{30}{92}\hat{\theta}\right]_{obs}}=\sqrt{\left[\frac{30}{92}603.6\right]}=14.029$

iii: Konfidensintervallet blir $\hat{\theta}\pm 1.96\cdot SE(\hat{\theta})$.

Observert: $603.6\pm(1.96)(14.029)=[576,\ 631]$

>>
