LØSNING: Eksamen 28. mai 2015

"*MAT110 Statistikk 1*", vår 2015

Oppgave 1: (revisjon)

a) Situasjonen som beskrives i oppgaven kan modelleres med en urne.
I denne urnen er fordelingen kjent, M antall bilag med feil og $N = 100\,000$ bilag totalt.
Den stokastiske variabelen X har derfor en hypergeometrisk fordeling.

b) Siden X er hypergeometrisk fordelt så er forventingen:

$$E[X] = n\frac{M}{N} \tag{1}$$

Løser med hensynpå M alane og får:

$$\underline{\underline{M}} = \frac{E[X]N}{n} = \frac{10 \cdot 100\,000}{1000} = \underline{\underline{1000}} \quad , \quad \text{q.e.d.} \tag{2}$$

c) Siden X er hypergeometrisk fordelt så er variansen:

$$\underline{Var[X]} = \frac{N-n}{N-1} n \frac{M}{N}\left(1 - \frac{M}{N}\right) \tag{3}$$

$$= \frac{100\,000 - 1000}{100\,000 - 1} 1000 \frac{1000}{100\,000}\left(1 - \frac{1000}{100\,000}\right) = \underline{9.8} \tag{4}$$

og tilhørende standardavvik:

$$\underline{\underline{\sigma[X]}} \;=\; \sqrt{Var[X]} \;=\; \sqrt{9.8} \;=\; \underline{\underline{3.13}} \quad , \quad \text{q.e.d.} \tag{5}$$

d) i) En hypergeometrisk fordeling kan med god tilnæremelse beskrives av en normalfordeling dersom:

$$N \;\gtrsim\; 20 \cdot n \tag{6}$$

$$n\,\frac{M}{N}\left(1 - \frac{M}{N}\right) \;\gtrsim\; 5 \tag{7}$$

ii) For vårt tilfelle er:

$$100\,000 \;\gtrsim\; 20 \cdot 1000 \tag{8}$$

$$1000\,\frac{1000}{100\,000}\left(1 - \frac{1000}{100\,000}\right) \;\gtrsim\; 5 \tag{9}$$

som gir

$$100\,000 \;\gtrsim\; 20\,000 \tag{10}$$

$$9.9 \;\gtrsim\; 5 \tag{11}$$

dvs. kritereiene er oppfylte i vårt tilfelle.

e) og **g**) Se vedlegg.

f) I oppgaven er det oppgitt at den stokastiske variabelen X er normalfordelt med $X \sim \mathrm{N}\big[\, E[X] = 10, \sigma[X] \,\big]$. Dermed løses denne oppgaven med (omvendt) tabelloppslag [1]. Standardiserer:

$$P(X \geq g) \quad = \quad 0.10 \tag{12}$$

$$P\Bigg(\underbrace{\frac{X - E[X]}{\sigma[x]}}_{\equiv Z} \geq \underbrace{\frac{g - E[X]}{\sigma[X]}}_{\equiv Z_0} \Bigg) \quad \overset{\text{standardiser}}{=} \quad 0.10 \tag{13}$$

$$P(Z \geq Z_0) \quad = \quad 0.10 \tag{14}$$

Ulikheten skal være "rett vei", altså mindre enn eller lik:

$$P(Z \geq Z_0) \;=\; 0.10 \tag{15}$$

$$1 - P(Z < Z_0) \;=\; 0.10 \tag{16}$$

$$P(Z < Z_0) \;=\; 0.90 \tag{17}$$

Ved "omvendt tabelloppslag" ser vi at Z_0 må være 1.28: [2]

$$Z_0 \;=\; 1.28 \tag{19}$$

Vi løser

$$Z_0 \;=\; \frac{g - E[X]}{\sigma[X]} \tag{20}$$

med hensyn på g: ($E[X] = 10$ og $\sigma[X] = 3.13$ fra oppgave **c**)

$$\underline{g} \;=\; E[X] \;+\; Z_0 \cdot \sigma[X] \tag{21}$$

$$=\; 10 \;+\; 1.28 \cdot 3.13 \;=\; \underline{14} \tag{22}$$

---

[1]Husk at tabellen på side 65 i 2015-formelsamlingen dreier seg KUN om normalfordeling.

[2]For en kontinuerlig sannsynlighetsfordeling er sannsynligheten i et punkt lik null. Dermed kan vi skrive:

$$P(Z \leq Z_0) = P(Z < Z_0) + \underbrace{P(Z = Z_0)}_{=0} = P(Z < Z_0) \tag{18}$$

Dermed kan vi finne fortsatt bruke tabellen i formelsamlingen selv om tabellen sier noe om $G(Z_0) \equiv P(Z \leq Z_0)$ mens vi skal finne $P(Z < Z_0)$.

Dette betyr at revisoren underkjenner regnskapet dersom han finner 14 feil eller mer blant de $n = 1000$ bilagene han trekker.

$\blacksquare$

Oppgave 2: (økonomi / logistikk)

a) Siden den stokastiske variabelen D_1 er oppgitt i oppgaven til å være normalfordelt så finner vi sannsynligheten $P(D_1 > 21\,000)$ ved standardisering og tilhørende tabelloppslag:

$$\underline{P(D_1 > 21\,000)} \overset{\text{standardiser}}{=} P\left(\underbrace{\frac{D_1 - \mu_1}{\sigma_1}}_{=\,Z} > \frac{21\,000 - \mu_1}{\sigma_1}\right) \tag{23}$$

$$= P\left(Z > \frac{21\,000 - 15\,000}{3000}\right) \tag{24}$$

$$= P(Z > 2) \tag{25}$$

$$= 1 - P(Z \leq 2) \tag{26}$$

$$= 1 - \underbrace{G(2)}_{=\,0.9772} \tag{27}$$

$$\overset{\text{tabell}}{=} 1 - 0.9772 \tag{28}$$

$$= \underline{0.0228} \tag{29}$$

Sannsynligheten for at mer enn 21 000 trofaste aviskjøpere kjøper Dagbladet en gitt dag er $\underline{\underline{2.28\ \%}}$.

b) Det er oppgitt i oppgaven at den totale etterspørselen D_{tot} av aviser dersom det har skjedd en spesiell nyhet er:

$$D_{tot} = D_1 + D_2 \tag{30}$$

Ved å bruke formel (5.13) i formelsamlingen fra 2015 finner vi da:

$$\underline{\underline{E[D_{tot}]}} = E[D_1 + D_2] \tag{31}$$

$$\overset{\text{alltid}}{=} \underbrace{E[D_1]}_{=\mu_1} + \underbrace{E[D_2]}_{=\mu_2} = \mu_1 + \mu_2 = 15\,000 + 18\,000 = \underline{\underline{33\,000}} \tag{32}$$

NB: Overgangen i lign.(31) til (32) gjelder alltid. Uansett om de stokastiske variablene X_i er uavhengige eller ikke.

c) Variansen til den totale etterspørselen D_{tot} av aviser dersom det har skjedd en spesiell nyhet er:

$$\underline{\underline{Var[\,D_{tot}\,]}} = Var\,[\,D_1 + D_2\,] \tag{33}$$

$$\overset{\text{uavh.}}{=} Var[D_1] + Var[D_2] \tag{34}$$

$$= \sigma_1^2 + \sigma_2^2 = 3000^2 + 4000^2 = \underline{\underline{25\,000\,000}} \tag{35}$$

NB: Overgangen fra lign.(33) til lign.(34) gjelder *kun* dersom de stokastiske variablene D_1 og D_2 er uavhengige.

d) Forventede fortjenesten til Dagbladet $E[F]$ per dag dersom en spesiell nyhet har skjedd:

$$\underline{\underline{E[F]}} \quad = \quad E\Big[\,(p-k)D_{tot} \;-\; c\,\Big] \tag{36}$$

$$= \quad (p-k)E[D_{tot}] \;-\; c \;=\; \Big(\,(25-12)\cdot 33\,000 - 400\,000\,\Big)\ \text{NOK} \tag{37}$$

$$= \quad \underline{\underline{29\,000\ \text{NOK}}} \tag{38}$$

Her har vi brukt at $E[D_{tot}] = 33\,000$ fra oppgave **2b**, se lign.(32).

e) Siden D_{tot} er oppgitt i oppgaven til å være normalfordelt så finner vi sannsynligheten $P(D_{tot} > 21\,000)$ ved standardisering og tilhørende tabelloppslag: (bruker at $\sigma[D_{tot}] = \sqrt{Var[D_{tot}]} = \sqrt{25\,000\,000} = 5000$ fra oppgave **2c**)

$$\underline{P(D_{tot} > 21\,000|S)} \overset{\text{standardiser}}{=} P\left(\underbrace{\frac{D_{tot} - E[D_{tot}]}{\sigma[D_{tot}]}}_{=\,Z_{tot}}\Bigg| S > \frac{21\,000 - 33\,000}{5000}\right) \tag{39}$$

$$= \quad P\left(Z_{tot} \;>\; \frac{21\,000 - 33\,000}{5000}\right) \tag{40}$$

$$= \quad P(\,Z_{tot} > -\,2.4\,) \tag{41}$$

$$= \quad 1 - P(\,Z_{tot} \leq -\,2.4\,) \tag{42}$$

$$= \quad \cancel{1} \;-\; \Big(\cancel{1} - P(\,Z_{tot} \leq 2.4\,)\Big) \tag{43}$$

$$= \quad P(\,Z_{tot} \;\leq\; 2.4\,) \tag{44}$$

$$\overset{\text{tabell}}{=} \quad G(2.4) \tag{45}$$

$$= \quad \underline{\;0.9918\;} \tag{46}$$

Sannsynligheten for at mer enn 21 000 personer kjøper Dagbladet en gitt dag, dersom vi vet at det har skjedd en spesielle nyhet, er $\underline{\underline{99.18\ \%}}$.

f) Oppgaven spør etter sannsynligheten $P(D_{tot} > 21\,000)$.
Oppsplitting av utfallsrom Ω: [3]

$$\underline{P(D_{tot} > 21\,000)} \overset{\text{oppspl.}}{=} \underbrace{P(D_{tot} > 21\,000|S)}_{=0.9918} \cdot \overbrace{P(S)}^{=0.20} + \underbrace{P(D_{tot} > 21\,000|\overline{S})}_{=P(D_1>21\,000)=0.028} \cdot \overbrace{P(\overline{S})}^{=1-0.20} \tag{47}$$

$$= 0.9918 \cdot 0.20 + 0.0228 \cdot (1 - 0.20) = \underline{0.2166} \tag{48}$$

Sannsynligheten for at det en gitt dag etterspørres mer enn 21 000 aviser dersom vi ikke vet om det skjer en spesiell nyhet eller ikke den aktuelle dagen, er $\underline{\underline{21.66\,\%}}$.

■

[3]Se side 32 i formelsamlingen 2015.

Oppgave 3: (petroleumslogistikk)

a) Siden

$$\underline{\underline{\sum_{i=0}^{n} P(X = x_i)}} = P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4) \quad (49)$$

$$= 0.09 + 0.28 + 0.41 + 0.17 + 0.05 = \underline{\underline{1}} \quad (50)$$

b) Sannsynligheten for at det leveres inn 3 rapporter eller mer per dag:

$$\underline{\underline{P(X \geq 3)}} = P(X=3) + P(X=4) = 0.17 + 0.05 = \underline{\underline{0.22}} \quad (51)$$

c) i) Forventet antall innleverte rapporter per dag:

$$\underline{\underline{E[X]}} \stackrel{\text{def.}}{=} \sum_{i=0}^{4} x_i \cdot P(X = x_i) \quad (52)$$

$$= 0 \cdot \underbrace{P(X=0)}_{=0.09} + 1 \cdot \underbrace{P(X=1)}_{=0.28} + 2 \cdot \underbrace{P(X=2)}_{=0.41} + 3 \cdot \underbrace{P(X=3)}_{=0.17} + 4 \cdot \underbrace{P(X=4)}_{=0.05}$$

$$= 0 \cdot 0.09 + 1 \cdot 0.28 + 2 \cdot 0.41 + 3 \cdot 0.17 + 4 \cdot 0.05 = \underline{\underline{1.81}} \quad (53)$$

ii) For å finne variansen $Var[X]$ så regner vi først ut $E[X^2]$:

$$\underline{E[X^2]} \stackrel{\text{def.}}{=} \sum_{i=0}^{3} x_i^2 \cdot P(X = x_i) \quad (54)$$

$$= 0^2 \cdot \underbrace{P(X=0)}_{=0.09} + 1^2 \cdot \underbrace{P(X=1)}_{=0.28} + 2^2 \cdot \underbrace{P(X=2)}_{=0.41} + 3^2 \cdot \underbrace{P(X=3)}_{=0.17} + 4^2 \cdot \underbrace{P(X=4)}_{=0.05}$$

$$= 0^2 \cdot 0.09 + 1^2 \cdot 0.28 + 2^2 \cdot 0.41 + 3^2 \cdot 0.17 + 4^2 \cdot 0.05 = \underline{4.25} \quad (55)$$

Dette innsatt i setningen for '*varianssetningen*': [4]

$$\underline{\underline{Var[X]}} \quad = \quad E[X^2] - E[X]^2 \quad = 4.25 - 1.81^2 \quad = \underline{\underline{0.9739}} \tag{56}$$

d) i) Tolkning:

$$\underline{\underline{E[\,\overline{X}\,] \;=\; \text{forventet antall innleverte rapporter per dag i gjennomsnitt over et helt år}}}$$

ii) Forventet antall innleverte rapporter i gjennomsnitt per dag: ($n = 365$)

$$\underline{\underline{E[\,\overline{X}\,]}} \quad = \quad E\left[\,\frac{1}{n}\left(X_1 + X_2 + \;...\; + X_n\right)\,\right] \tag{57}$$

$$\overset{\text{alltid}}{=} \quad \frac{1}{n}\left(E[X_1] + E[X_2] + \;...\; + \;E[X_n]\right) \tag{58}$$

$$= \quad \frac{1}{n}\left(\underbrace{E[X] + E[X] + \;...\; + E[X]}_{=\; n\cdot E[X]}\right) \quad = \quad \frac{1}{\cancel{n}}\cancel{n}\cdot \underbrace{E[X]}_{=1.81} \quad = \quad \underline{\underline{1.81}} \tag{59}$$

NB: Overgangen i lign.(57) til (58) gjelder alltid. Uansett om de stokastiske variablene X_i er uavhengige eller ikke.

[4] Se side 39 i formelsamlingen fra 2015.

e) i) Tolkning:

$$\underline{\underline{Var[\overline{X}\,]}} \quad = \quad \underline{\underline{\text{forventet variasjon/spredning i antall innleverte rapporter per dag i gjennomsnitt over et helt år}}} \tag{60}$$

ii) Variansen til antall innleverte rapporter per dag i gjennomsnitt over et år: $\quad (\; n = 365 \;)$

$$\underline{\underline{Var[\overline{X}\,]}} \quad = \quad Var\left[\, \frac{1}{n}\Big(X_1 + X_2 + \; ... \; + X_n \Big) \right] \tag{61}$$

$$\overset{\text{uavh.}}{=} \quad \frac{1}{n^2}\Big(\underbrace{Var[X_1] + Var[X_2] + \; ... \; + Var[X_n]}_{n \cdot Var[X]} \Big) \tag{62}$$

$$= \quad \frac{1}{n^{\cancel{2}}} \, \cancel{n} \cdot \underbrace{Var[X]}_{=0.9739} \quad = \quad \frac{0.9739}{365} \quad = \quad \underline{\underline{0.002668}} \tag{63}$$

NB: Overgangen i lign.(61) til (62) gjelder fordi de stokastiske variablene X_i er uavhengige.

f) Siden

1. antall innleverte rapporter per dag er uavhengige: (oppgitt i oppgaveteksten)
   $X_i \;\sim\;$ er uavhengige for alle $i = 1, 2, 3, ..., n$
2. alle X_i har samme sannsynlighetsfordeling: (oppgitt i oppgaveteksten)
   $X_i \;\sim\;$ samme sannsynlighetsfordeling for alle $i = 1, 2, 3, ..., n$
3. antall "forsøk", dvs. antall dager, $n = 365$ er tilstrekkelig stort [5]

så gjelder **sentralgrensesetningen**. Dermed er $\overline{X}$ normalfordelt.

[5] Husk: Antall forsøk n for at sentralgrensesetningen skal gjelde er avhengig av situasjonen. Men en **tommelfingerregel** er at vi bør ha $n \gtrsim 30$.

g) Fra oppgavene **2c**, **2d** og **2e** foran ser vi at:

$$\underbrace{E[\,\overline{X}\,]}_{=\ 1.81} \quad = \quad \underbrace{E[X]}_{=\ 1.81} \tag{64}$$

og at

$$\underbrace{Var[\,\overline{X}\,]}_{=\ 0.002668} \quad \ll \quad \underbrace{Var[X]}_{=\ 0.9739} \tag{65}$$

Det betyr at sannsynlighetfordelingen til X, dvs. $P(X = x)$ gitt ved tabell i oppgaveteksten, og sannsynlighetfordelingen til $\overline{X}$, dvs. $\overline{X} \sim N[\,E[X], \sqrt{\frac{VarX}{n}}\,]$ har samme tyngdepunkt, men mye mindre varians / usikkerhet.

h) Sannsynligheten for at antall innleverte rapporter i løpet av et år overstiger 700: ($n = 365$)

$$\underline{\underline{P(X_1 + X_2 + \ \dots\ + X_n > 700)}} = P\left(\frac{X_1 + X_2 + \ \dots\ + X_n}{n} > \frac{700}{n}\right) \tag{66}$$

$$= P\left(\overline{X} > \frac{700}{n}\right) \tag{67}$$

$$= 1 - P\left(\overline{X} \leq \frac{700}{n}\right) \tag{68}$$

$$\overset{\text{standardiser}}{=} 1 - P\Bigg(\underbrace{\frac{\overline{X} - E[\overline{X}]}{\sigma[\overline{X}]}}_{=\overline{Z}} \leq \frac{\frac{700}{n} - E[\overline{X}]}{\sigma[\overline{X}]}\Bigg) \tag{69}$$

$$= 1 - P\left(\overline{Z} \leq \frac{\frac{700}{365} - 1.81}{\sqrt{0.002668}}\right) \tag{70}$$

$$= 1 - P(\overline{Z} \leq 2.09) \tag{71}$$

$$= 1 - G(2.09) \tag{72}$$

$$= 1 - 0.9817 \tag{73}$$

$$= \underline{\underline{0.0183}} \tag{74}$$

Kommentar:

Legg merke til at det er $\overline{X}$ som skal standardiseres. Ikke X. Det betyr at vi må bruke $E[\overline{X}] = 1.81$ og $\sigma[\overline{X}] = \sqrt{0.002668}$ (ikke $E[X] = 1.81$ og $\sigma[X] = \sqrt{0.9739}$).

■

Oppgave 4: (økonomi)

a) **i**) R_{xy} er normalisert og ligger mellom -1 og 1, dvs.: $\underline{\underline{-1 \leq R_{xy} \leq 1}}$.

ii) R_{xy} er et mål på <u>lineær korrelasjon</u> mellom obervasjonene x og y.
Det er et mål på i hvor stor grad det er en lineær sammmenheng mellom obervasjonene x og y.

iii) R_{xy} er <u>enhetsuavhengig</u>, dvs. ingen enhet. [6]

b) Uttrykket for korrelasjonskoeffisienten finner vi i formelsamlingen. Det gir:

$$\underline{\underline{R_{xy}}} = \frac{S_{xy}}{S_x \cdot S_y} \tag{75}$$

$$= \frac{-110.83}{\sqrt{116.67} \cdot \sqrt{109.54}} = \underline{\underline{-0.98}} \tag{76}$$

c) At $R_{xy} = -0.98$ betyr at det er en <u>sterk negativ korrelasjon</u> mellom x og y, dvs.
store x hører sammen med små y og omvendt.
Det er altså en <u>sterk lineær sammenheng</u> mellom x og y med negativt stigningstall.

[6] Dette betyr at R_{xy} har samme verdi uansett hva slags enhet man bruker for å regne ut $R_{xy} = S_{xy}/(S_x \cdot S_y)$.

d) Vi skal finne regresjonslinjen. Formelen for minste kvadraters lineære regresjonslinje står i formelsamlingen:

$$\hat{y} \;=\; \hat{\alpha} \;+\; \hat{\beta}x \tag{77}$$

hvor parametrene $\hat{\beta}$ og $\hat{\alpha}$ er: (dropper benevningen)

$$\underline{\hat{\beta}} \;=\; \frac{S_{xy}}{S_x^2} \;=\; \frac{-110.83}{116.67} \;=\; \underline{-0.95} \tag{78a}$$

$$\underline{\hat{\alpha}} \;=\; \overline{y} \;-\; \beta\overline{x} \;=\; 40.43 - (-0.95)\cdot 15 \;=\; \underline{54.68} \tag{78b}$$

Minste kvadraters lineære regresjonslinje $\hat{y} \;=\; \hat{\alpha} \;+\; \hat{\beta}\,x$ blir dermed:

$$\underline{\underline{\hat{y} \;=\; 54.68 \;-\; 0.95\cdot x}} \tag{79}$$

e) Bruker regresjonslinjen $\hat{y} = \hat{y}(x)$ fra oppgave **4d**, dvs. lign.(79):

$$\underline{\hat{y}(17)} \;=\; 54.68 \;-\; 0.95\cdot 17 \;=\; \underline{38.53} \tag{80}$$

Regresjonslinjen predikerer at kvadratmeterprisen for en leilighet som er 17 km utenfor sentrum er $\underline{\underline{38\,530 \text{ NOK}/m^2}}$.

f) Gjennomsnittlig kvadratmeterpris i Norge er 35 000 NOK/m^2.
Ifølge regresjonslinjen, når $\hat{y}(x) = 35$, er da:

$$35 \quad = \quad 54.68 \; - \; 0.95 \cdot x \tag{81}$$

Løser med hensyn på x:

$$\underline{x} \quad = \quad \frac{54.68 \; - \; 35}{0.95} \quad = \quad \underline{20.72} \tag{82}$$

Ifølge regresjonslinjen oppnår man landsgjennomsnittet i kvadratmeterpris
20.72 km utenfor sentrum.

g) Forklaringskraften R^2 kan leses direkte fra Excel-utskriften:
(Se cellen som heter "*R Square*" i Excel-utskriften): [7]

$$\underline{\underline{R^2 \; = \; 0.9613}} \tag{83}$$

h) Kommentar til svaret i oppgave **4g**:

At $R^2 = 0.9613$ betyr at regresjonslinjen vil i "i stor grad" forutsi kvadrameterprisen på en leilighet for en gitt avstand fra sentrum.
Modellen har stor forklaringskraft.

■

[7] Man kan også regne ut forklaringskraften R^2 "for hånd" via definisjonen $R^2 = 1 - SSE/SST$. Men det er mye mer arbeidskrevende. Fint at dataprogrammer (som f.eks. Excel) kan hjelpe oss med slikt.

Vedlegg

Studentnummer: ____________________