ECON2130: EKSAMEN 2015v

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

En klasse på 25 elever på videregående består av 15 jenter og 10 gutter. 4 av jentene og 2 av guttene røyker. En elev trekkes ut rent tilfeldig (slik at alle 25 har samme sannsynlighet).

A. **i)** Sett opp en frekvenstabell over de fire mulige kombinasjonene av kjønn og røykestatus. Dvs. fyll inn

	Røyker	Ikke - røyker	Sum
Jente			
Gutt			
Sum			25

ii) Definer begivenhetene G og R ved, G = «den uttrukne er gutt» og R = «den uttrukne røyker». Finn følgende 4 sannsynligheter

$$P(G),\quad P(G\cap R),\quad P(G\cup R)\ \text{ og }\ P(G\,|\,R)$$

<< **Svar:**

i)

	R	$\bar{R}$	Sum
J	4	11	15
G	2	8	10
Sum	6	19	25

ii)

$$P(G) = 10/25 = 0.40 \qquad P(G\cap R) = 2/25 = 0.08$$

$$P(G\cup R) = (2+4+8)/25 = 14/25 = 0.56$$

(Eventuelt $P(G\cup R) = P(G) + P(R) - P(G\cap R) = 10/25 + 6/25 - 2/25 = 14/25$)

$$P(G\,|\,R) = \frac{P(G\cap R)}{P(R)} = \frac{2/25}{6/25} = \frac{2}{6} = 0.333$$

>>>

B. Det trekkes 3 elever fra klassen til en komité rent tilfeldig (dvs. slik at alle ikke-ordnete utvalg på 3 er like sannsynlige). La *X* være antallet i komiteen som røyker.

i) Forklar hvorfor *X* er hypergeometrisk fordelt.

ii) Beregn $P(X = 2)$ (med 3 desimalers nøyaktighet).

[**Hint:** For å lette regningen oppgis $\binom{25}{3} = 2300$.]

iii) Bruk kjente egenskaper ved den hypergeometriske fordelingen til å vise at

$$E(X) = 0.72 \quad \text{og} \quad \text{Var}(X) = 0.5016$$

<< **Svar:** **i)** Populasjonen består av *N* = 25 enheter hvorav *M* = 6 er *R*. Under forutsetning av at utvalget på *n* = 3 er rent tilfeldig, vil *X* = antall *R* i utvalget, være hypergeometrisk med punktsannsynlighet

$$P(X = x) = \frac{\binom{6}{x}\binom{19}{3-x}}{\binom{25}{3}}, \quad x = 0,1,2,3$$

(ii) som gir $P(X = 2) = \frac{\binom{6}{2}\binom{19}{1}}{\binom{25}{3}} = \frac{15 \cdot 19}{2300} = 0.1239....$

(iii) Formler i Løvås gir

$$E(X) = np = 3 \cdot (0.24) = 0.72, \text{ der } p = \frac{M}{N} = \frac{6}{25} = 0.24, \text{ og}$$

$$\text{Var}(X) = np(1-p)\frac{N-n}{N-1} = 3(0.72)(0.28)\frac{22}{24} = 0.5016$$

\>>>

C. I tillegg til *X* innfører vi *Y* = antall jenter i komiteen. Den simultane fordelingen for (X,Y), bestemt ved $f(x,y) = P(X = x \cap Y = y)$, er gitt i tabell 1 (som du ikke trenger å vise):

Tabell 1 **Tabell over** $f(x,y)$

		y				
		0	1	2	3	sum
x	0	0.024	0.134	0.191	0.072	0.421
	1	0.024	0.125	0.201	0.096	0.446
	2	0.003	0.033	0.059	0.029	0.124
	3	0.000	0.002	0.005	0.002	0.009

	sum	0.051	0.294	0.456	0.199	1.000

i) Forklar hvorfor $f(3,0)=0$.

ii) Beregn $P(X=Y)$.

iii) Anta vi vet at det er 2 jenter og 1 gutt i den uttrukne komiteen, men ikke om de røyker eller ikke. Hva er da sannsynligheten for at høyst en av dem røyker?
[**Hint:** Finn $P(X \le 1 | Y=2)$]

<<< **Svar:** **i)** Begivenheten $(X,Y)=(3,0)$ impliserer at det bare er gutter i komiteen og at alle 3 røyker. Dette er umulig siden bare 2 av guttene i klassen røyker, og sannsynligheten for en umulig begivenhet er 0.

ii)

$$P(X=Y)=P\left[(X=0\cap Y=0)\cup(X=1\cap Y=1)\cup\cdots\cup(X=3\cap Y=3)\right]=$$
$$=f(0,0)+f(1,1)+\cdots+f(3,3)=0.024+0.125+0.059+0.002=0.210$$

iii)

$$P(X\le 1|Y=2)=\frac{P(X\le 1\cap Y=2)}{P(Y=2)}=\frac{f(0,2)+f(1,2)}{\sum_{x=0}^{3}f(x,2)}=\frac{0.191+0.201}{0.456}=\frac{0.392}{0.456}=0.8596...$$

>>>

D. Beregn korrelasjonskoeffisienten mellom *X* og *Y*.

[**Hint:** For å lette regningen oppgis
$E(Y)=1.8,\quad \mathrm{Var}(Y)=0.7$ og $\sum_{x=0}^{3}\sum_{y=0}^{3}xyf(x,y)=1.34$]

<<< **Svar:** Korr.koef. mellom X og Y er $\rho=\rho(X,Y)=\dfrac{\mathrm{cov}(X,\mathrm{Y})}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$. Fra hintet og **Biii**, får vi:

$$\mathrm{cov}(X,Y)=E(XY)-E(X)\cdot E(Y)=\sum_{x=0}^{3}\sum_{y=0}^{3}xyf(x,y)-E(X)\cdot E(Y)=1.34-(0.72)(1.8)=0.044$$

hvorav $\rho=\rho(X,Y)=\dfrac{\mathrm{cov}(X,\mathrm{Y})}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}=\dfrac{0.044}{\sqrt{(0.5016)(0.7)}}=0.076..$ >>>

E. Utvalget i punkt **B** blir trukket *uten* tilbakelegging på følgende måte. Alle elevene skriver navnet sitt på en lapp som legges i en kurv (til sammen 25 navnelapper). Utvalget trekkes så ved å trekke en og en navnelapp av gangen uten å legge den uttrukne lappen tilbake i kurven før neste trekning.

Anta vi i stedet trekker 3 elever *med* tilbakelegging (dvs. vi trekker lappene en og en, men legger hver lapp som blir trukket tilbake i kurven før neste trekning). Med denne metoden er det naturligvis en viss risiko for at en og samme elev (dvs. navnelapp) blir trukket ut flere ganger.

Finn sannsynligheten for at de 3 navnelappene (trukket ut med tilbakelegging) er forskjellige.

> [**Hint:** Det er flere måter å finne denne sannsynligheten på. Du kan selve velge den du synes er best. En av dem er å beregne antall gunstige utfall og antall mulige utfall.
> En annen metode er følgende: La *A* være begivenheten at de to første navnelappene som blir trukket er forskjellige, og *B* begivenheten at alle tre er forskjellige (der altså *B* er den begivenheten vi ønsker sannsynligheten for).
> Forklar hvorfor $B = A \cap B$ og $P(A) = 24/25$. Bruk så multiplikasjonssetningen på $P(A \cap B)$.]

<<< **Svar:** Gunstig-på-mulig-metoden: Det er 25^3 mulige utfall som alle er like sannsynlige. Av disse er $25 \cdot 24 \cdot 23$ gunstige muligheter, som gir sannsynlighet for B lik 0.8832..
Metode 2: Vi har $B \Rightarrow A$. I et Venndiagram er dette uttrykt ved at *B* er inneholdt i *A*, som impliserer $B = A \cap B$.
Alternativt: Hvis $A \cap B$ inntreffer, må *B* inntreffe. På den annen side, hvis B inntreffer, må også A inntreffe siden $B \Rightarrow A$. Dermed $B = A \cap B$.
Vi har $P(A) = \dfrac{24}{25}$ og $P(B \mid A) = \dfrac{23}{25}$, hvorav

$$P(B) = P(A \cap B) = P(A)P(B \mid A) = \frac{24 \cdot 23}{25^2} = 0.8832$$ >>>

Oppgave 2

I 2014 ble det i Sverige trukket et representativt utvalg på $n = 1313$ ungdommer i aldersklassen 16-29 år. Av disse var det 107 som røykte jevnlig. I 2012 røykte 10% av ungdommer i denne aldersklassen i Sverige. Vi ønsker å teste om tallene tyder på at røyking blant ungdommer i Sverige har gått ned fra 2012 til 2014.

La *X* være antall som røyker jevnlig i et representativt utvalg på *n* ungdommer. Anta at *X* er binomisk fordelt med parametre *n* og *p* (kort: $X \sim \text{bin}(n, p)$), der *p* er andelen av ungdommer i Sverige som røyker jevnlig i 2014, og som tolkes som sannsynligheten for at en tilfeldig valgt ungdom i Sverige røyker.

A.

i) Sett opp en test med signifikansnivå 5% for $H_0: p \geq 0.1$ mot $H_1: p < 0.1$. Skriv testen på formen, «Forkast H_0 hvis $X \leq k$» og bestem den kritiske verdien *k*.

ii) Beregn p-verdien (tilnærmet) for testen i **i)** basert på tallene i innledningen og formuler en konklusjon. Bruk heltallskorreksjon ved beregning av p-verdien.

<<< **Svar:**

i) Testobservatoren med $p_0 = 0.1$ er i utgangspunktet

$$Z = \frac{\hat{p} - 0.1}{\sqrt{\frac{0.1 * 0.9}{1313}}} = \frac{X - 131.1}{\sqrt{np_0(1-p_0)}} = \frac{X - 131.1}{\sqrt{118.17}} = \frac{X - 131.1}{10.87} \overset{\text{tilnærmet}}{\sim} N(0,1) \qquad \text{hvis } p = 0.1$$

En tilnærmet 5% test er derfor:

Forkast H_0 hvis $\dfrac{X - 131.1}{10.87} \leq -1.645$ (=5%-prosentilen i N(0,1)).

Dette kriteriet er ekvivalent med Forkast H_0 hvis $X \leq 131.1 - (1.645)(10.87)$, dvs.

Forkast H_0 hvis $X \leq 113.2189$

som er ekvivalent med (siden *X* er heltallig):

Forkast H_0 hvis $X \leq 113 \quad (= k)$

ii) Den observerte verdien av X er $x_o = X_{obs} = 107$, som gir forkastning på 5% - dvs. det er sterk evidens i data for at røyking blant ungdommer i Sverige har gått ned.

$$\text{p-verdi} = P_{p=0.1}(X \leq 107) \overset{\text{heltallskorreksjon}}{=} P_{p=0.1}(X \leq 107.5) = P_{p=0.1}\left(\frac{X - 131.1}{10.87} \leq \frac{107.5 - 131.1}{10.87}\right)$$

$$= P_{p=0.1}(Z \leq -2.17) \approx G(-2.17) = 0.015 \qquad \text{(p-verdi} \approx 0.0132 \text{ uten heltallskorreksjon)}$$

der $G(z)$ er den kumulative fordelingsfunksjonen i N(0,1).

\>>>

B. **i)** Utled en tilnærmet formel for styrkefunksjonen for testen din i punkt **Ai,** der du tilnærmer den binomiske fordelingen med en normalfordeling**.**

ii) Forklar kort hva som menes med begrepet «feil av type 2» ved bruk av en test.

iii) Beregn (tilnærmet) sannsynligheten for feil av type 2 for testen din i punkt **Ai** dersom den ukjente *p* er lik 0.095, og dersom $p = 0.105$.

<<< **Svar:**

i) Løvås, regel 5.20, sier at X er tilnærmet normalfordelt $(np,\ \sqrt{np(1-p)})$ dersom $\mathrm{Var}(X) = np(1-p) \geq 5$, noe som ikke er noe problem her siden n er så stor).

Styrkefunksjonen:

$$\gamma(p) = P_p(\text{forkast } H_0) = P_p(X \leq 113) = P_p\left(\frac{X-np}{\sqrt{np(1-p)}} \leq \frac{113-np}{\sqrt{np(1-p)}}\right) \approx G\left(\frac{113-np}{\sqrt{np(1-p)}}\right)$$

(eventuelt $\gamma(p) \approx G\left(\frac{113.5-np}{\sqrt{np(1-p)}}\right)$ med heltallskorreksjon. Begge varianter bør godtas her.)

ii) Feil av type 2 kan kun oppstå dersom H_0 er gal, og man med testen ikke forkaster H_0. Hvis H_0 er sann, er feil av type 2 umulig og har sannsynlighet 0.

iii) Hvis $p = 0.105$, gjelder H_0, og $P(\text{feil type 2}) = 0$.
Hvis $p = 0.095$, er H_0 gal og

$$P(\text{feil type 2}) = P(\text{ikke forkast } H_0) = 1-\gamma(0.095) \approx 1-G\left(\frac{113-np}{\sqrt{np(1-p)}}\right)\Bigg|_{p=0.095} = 1-G(-1.10) = 0.8643$$

(Med heltallskorreksjon blir

$$P(\text{feil type 2}) \approx 1-G\left(\frac{113.5-np}{\sqrt{np(1-p)}}\right)\Bigg|_{p=0.095} = 1-G(-1.06) = 0.8554 \quad)$$

>>>

Oppgave 3

Resultatene fra den svenske undersøkelsen i oppgave 2 er splittet opp på kvinner og menn som vist i tabell 2.

Tabell 2 Antall kvinner og menn, 16-29 år, som røyker - basert på et representativt utvalg fra Sverige 2014.

	Kvinner	Menn	sum
Røyker	67	40	107
Røyker ikke	674	532	1206
sum	741	572	1313

Vi kan se på dette som to utvalg, et for kvinner ($n_K = 741$ personer), og et for menn ($n_M = 572$ personer). La X_K, X_M betegne henholdsvis antall kvinner som røyker og antall menn som røyker i to slike utvalg. De observerte verdiene av X_K, X_M i utvalget er 67 og 40 henholdsvis.

Anta X_K, X_M er uavhengige stokastiske variable som begge er binomisk fordelte, $X_K \sim \text{bin}(n_K, p_K)$ og $X_M \sim \text{bin}(n_M, p_M)$, der p_K, p_M er andelen henholdsvis av kvinner og menn i alderen 16-29 år i Sverige som røyker i 2014. Andelene p_K, p_M betraktes som ukjente, og tallene n_K, n_M som gitte tall (ikke-stokastiske).

Vi ønsker å bruke data til å estimere forskjellen i andel, $\theta = p_K - p_M$, samt beregne et konfidensintervall for θ.

A. **i)** Vis at $\hat{\theta} = \dfrac{X_K}{n_K} - \dfrac{X_M}{n_M}$ er en forventningsrett estimator for θ. Beregn estimatet $\hat{\theta}_{obs}$ (der indeksen *obs* indikerer den observerte verdien).

ii) Vis at standardavviket (*SD*) for $\hat{\theta}$ er gitt ved

$$SD(\hat{\theta}) = \sqrt{\frac{p_K(1-p_K)}{n_K} + \frac{p_M(1-p_M)}{n_M}}$$

iii) Forklar hvorfor $\hat{\theta}$ er tilnærmet normalfordelt.

<<< **Svar:**

i) Av regler for forventning i Løvås følger

$$E(\hat{\theta}) = E\left(\frac{X_K}{n_K} - \frac{X_M}{n_M}\right) = \frac{1}{n_K}E(X_K) - \frac{1}{n_M}E(X_M) = \frac{1}{n_K}n_K p_K - \frac{1}{n_M}n_M p_M = p_K - p_M = \theta$$

Estimat: $\hat{\theta}_{obs} = \dfrac{67}{741} - \dfrac{40}{572} = 0.090 - 0.070 = 0.020$

ii) Siden X_K, X_M er uavhengige

$$\text{Var}(\hat{\theta}) = \text{Var}\left(\frac{X_K}{n_K} - \frac{X_M}{n_M}\right) = \frac{1}{n_K^2}n_K p_K(1-p_K) + \frac{1}{n_M^2}n_M p_M(1-p_M) = \frac{p_K(1-p_K)}{n_K} + \frac{p_M(1-p_M)}{n_M}$$

Dermed $SD(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})}$ som gir uttrykket i oppgaven.

iii) Følger av regler for normalfordeling: Hvis X_1, X_2 er uavhengige og normalfordelte, vil enhver lineærkombinasjon, $a_1X_1 + a_2X_2$ være normalfordelt. $\hat{\theta}$ er en lineærkombinasjon av to uavhengige variable som begge er tilnærmet normalfordelte og må derfor selv være tilnærmet normalfordelt.
>>>

B. **i)** Det kan vises (som du ikke trenger å gjøre) at $W = \dfrac{\hat{\theta}-\theta}{SE(\hat{\theta})}$ er tilnærmet standard normalfordelt (N(0,1)) uansett θ, der standardfeilen, $SE(\hat{\theta})$, er estimert standardavvik som i punkt **Aii,** der de ukjente p_K, p_M er erstattet med estimater. Bruk dette til å utlede en formel for et konfidensintervall for θ med konfidensgrad tilnærmet 0.95.

ii) Beregn det observerte konfidensintervallet utledet i punkt **Bi** ut fra data.

<<< **Svar: i)** Med konfidensdgrad 0.95, brukes N(0,1)-kvantilen $z_{0.025} = 1.96$ og vi har $0.95 \approx P(-1.96 \le \dfrac{\hat{\theta}-\theta}{SE(\hat{\theta})} \le 1.96) = \cdots = P\left(\hat{\theta} - 1.96 \cdot SE(\hat{\theta}) \le \theta \le \hat{\theta} + 1.96 \cdot SE(\hat{\theta})\right)$, som gir konfidensintervallet

$$\hat{\theta} \pm 1.96 \cdot SE(\hat{\theta}) = \hat{\theta} \pm 1.96 \cdot \sqrt{\frac{\hat{p}_K(1-\hat{p}_K)}{n_K} + \frac{\hat{p}_M(1-\hat{p}_M)}{n_M}}$$

der $\hat{p}_K = \dfrac{X_K}{n_K}$ og $\hat{p}_M = \dfrac{X_M}{n_M}$, med estimater $\hat{p}_{K,obs} = 0.090$ og $\hat{p}_{\mathrm{M},obs} = 0.070$, som gir $SE(\hat{\theta}) = 0.0150$

ii) Observert KI

$$\left[\hat{\theta} \pm 1.96 \cdot SE(\hat{\theta})\right]_{obs} = \hat{\theta}_{obs} \pm 1.96 \cdot SE(\hat{\theta}) = 0.020 \pm (1.96)(0.015) = [-0.009,\ 0.049]$$

>>>
