UNIVERSITETET I OSLO

Matematisk Institutt

Eksamen i:	**STK 1110 – Statistiske metoder og dataanalyse 1**
Tid for eksamen:	**Mandag 28. november 2016, kl. 14:30–18:30**
Hjelpemidler:	**Formelsamling til STK 1100 og STK 1110, godkjent kalkulator**

Dette eksamenssettet inneholder fire oppgaver og er på seks sider (inkludert et kort appendiks på siste side, som kan være til hjelp under løsningen av visse punkter).

Oppgave 1

I denne oppgaven skal vi se på problemstillinger knyttet til den eksponentielle fordelingen, med først ett og deretter to utvalg.

(a) Anta at observasjoner $X_1, \ldots, X_n$ er uavhengige og eksponentialfordelte med rate θ, altså at tettheten for en enkeltobservasjon er på formen

$$f(x, \theta) = \frac{1}{\theta} \exp\Big(-\frac{x}{\theta}\Big) \quad \text{for } x > 0.$$

Vis at variablene

$$V_i = \frac{2X_i}{\theta}$$

har χ^2-fordelingen med 2 frihetsgrader. – Her minner jeg om at χ^2-fordelingen med ν frihetsgrader har tettheten

$$h(z) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} z^{\nu/2-1} \exp(-\tfrac{1}{2}z) \quad \text{for } z > 0,$$

og at dennes forventning og varians er ν og 2ν.

(b) Vis at X_i har forventning θ og varians θ^2.

(c) Vis at $\widehat{\theta} = \bar{X}$, gjennomsnittet $(1/n)\sum_{i=1}^n X_i$ av de n observasjonene, er en forventningsrett estimator for θ. Finn dens varians.

(d) Forklar hva sentralgrenseteoremet (the central limit theorem) medfører for fordelingen for $\widehat{\theta}$, og bruk dette til å sette opp et konfidensintervall for θ, med dekningsgrad tilnærmet lik 95%.

(e) Anta vi observerer $X_1, \ldots, X_{20}$ fra laboratorium A, antatt uavhengige og eksponentielle med rate θ_1, samt $Y_1, \ldots, Y_{20}$ fra laboratorium B, uavhengige og eksponentielle med rate θ_2. Vi vil teste hypotesen H_0 at $\theta_1 = \theta_2$. Her har X_i-ene gjennomsnitt $\bar{X} = 2.222$ og empirisk standardavvik 2.468, mens Y_i-ene har gjennomsnitt $\bar{Y} = 4.444$ og empirisk standardavvik 4.987. Lag en testobservator for H_0, av typen

$$t = \frac{\bar{Y} - \bar{X}}{W},$$

der du skal konstruere W slik at t har en tilnærmet $\mathrm{N}(0,1)$-fordeling dersom H_0 er korrekt. Beregn t for denne situasjonen, og kommenter det du finner ut.

(f) Finn også et 95% konfidensintervall, eksakt eller tilnærmet, for brøkparameteren $\rho = \theta_2/\theta_1$, basert på disse dataverdiene. Du kan her få bruk for verdier gitt i tabellen bakerst i oppgavesettet.

Oppgave 2

En bestemt type operasjon blir jevnlig utført på barn under ett år med en viss type dramatisk hjertelidelse (der barna vil dø om de ikke blir operert). Operasjonen ender av og til med at barnet ikke kan reddes. Tabellen under viser dødsratene, i prosent, for elleve britiske sykehus, over en periode på noen få år. I tillegg til å forstå risikoen for død, samt hvilken type omstendigheter som øker denne risikoen, er man interessert i å finne ut om det er systematiske forskjeller mellom de elleve sykehus, eller om dødsrisikoen for små barn med denne ekstreme hjertelidelsen essensielt er den samme ved hvert av dem.

1	Leicester	13.37
2	Leeds	7.43
3	Oxford	18.85
4	Guys	15.24
5	Liverpool	10.37
6	Southampton	10.04
7	Great Ormond St	11.00
8	Newcastle	13.33
9	Harefield	14.12
10	Birmingham	9.98
11	Brompton	10.30

En mer detaljert analyse er mulig, men i denne oppgaven skal vi for enkelhets skyld se på disse estimatene som resultatet av uavhengige stokastiske variable

$$\widehat{\theta}_i \sim \mathrm{N}(\theta_i, \sigma^2) \quad \text{for } i = 1, \ldots, 11,$$

der standardavviket $\sigma = 2.112$ er såpass godt estimert at den kan anses som en kjent verdi. Parameterverdiene $\theta_1, \ldots, \theta_{11}$ ses på som de reelle, underliggende rater, som vi altså ikke kan observere eksakt, kun estimere, med de data vi har (som er det vi har gjort over).

(a) Finn to 95% konfidensintervall, ett for parameteren θ_1 og ett for θ_2, dødsfrekvensen for slike operasjoner utført ved sykehusene i henholdsvis Leicester og Leeds.

(b) For å analysere i hvilken grad θ_i-ene varierer fra sykehus til sykehus, eller om de kanskje er essensielt like, tenker vi oss at $\theta_1, \ldots, \theta_{11}$ selv er uavhengige stokastiske variable, fra fordelingen

$$\theta_i \sim \mathrm{N}(\theta_0, \tau^2).$$

Her er θ_0 og τ faste, men ukjente, parametre. Vi kan skrive

$$\theta_i = \theta_0 + \delta_i \quad \text{og} \quad \widehat{\theta}_i = \theta_i + \varepsilon_i \quad \text{for } i = 1, \ldots, 11,$$

der $\delta_i \sim \mathrm{N}(0, \tau^2)$ og $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ er uavhengige. Forklar hvorfor dette leder til

$$\widehat{\theta}_i \sim \mathrm{N}(\theta_0, \sigma^2 + \tau^2) \quad \text{for } i = 1, \ldots, 11.$$

(c) For estimatene over kan vi beregne

$$S^2 = \frac{1}{10}\sum_{i=1}^{11}(\widehat{\theta}_i - \bar{\theta})^2 = 10.0607 = 3.1719^2,$$

der $\bar{\theta}$ er gjennomsnittet $(1/11)\sum_{i=1}^{11}\widehat{\theta}_i = 12.1857$. Bruk dette til å estimere τ.

– For punktene (d) og (e) under kan du anvende at kvantilene 0.025, 0.050, 0.500, 0.950, 0.975 for χ^2_{10}, altså χ^2-fordelingen med 10 frihetsgrader, er
`3.2470 3.9403 9.3418 18.3070 20.4832`

(d) Lag en test med nivå 0.05 for hypotesen om at θ_i-ene er helt like, mot alternativet at det altså er forskjeller. Utfør testen.

(e) Lag til sist et 95% konfidensintervall for τ.

Oppgave 3

Anta at X er binomisk fordelt (n, p), altså med de klassiske punktsannsynligheter

$$f(x,p) = \binom{n}{x}p^x(1-p)^{n-x} \quad \text{for } x = 0, 1, \ldots, n.$$

(a) For et gitt utfall x, sett opp log-likelihood-funksjonen $\ell(p)$, og vis at denne maksimeres for $\widehat{p} = x/n$.

(b) Anta at vi har tre uavhengige binomiske eksperimenter,

$$X \sim \text{bin}(n,p), \quad Y \sim \text{bin}(n,q), \quad Z \sim \text{bin}(n,r).$$

Sett opp et uttrykk for log-likelihood-funksjonen $\ell(p,q,r)$. Dersom hypotesen H_0 at de tre sannsynlighetene er like, altså $p = q = r$, holder, hva er da maximum-likelihood-estimatet for den felles verdi av denne sannsynligheten?

(c) Anta at $n = 50$ og at man i de tre eksperimenter observerer $X = 17$, $Y = 22$, $Z = 14$. Finn verdier for

$$\ell_{\max}(\text{big}) \quad \text{og} \quad \ell_{\max}(H_0),$$

der $\ell_{\max}(\text{big})$ er maksimum av log-likelihood-funksjonen under modellen der p, q, r er frie parametre, og $\ell_{\max}(H_0)$ tilsvarende maksimum av samme funksjon under H_0. Utfør en test for H_0 og kommenter det du finner.

Oppgave 4

Tabellen under viser et helsepolitisk og historisk viktig datasett, fra 1963, idet det ble benyttet for å arbeide frem og etter noen år gjennomføre vedtak, i flere land, om at sigarettpakker skulle utstyres med en advarsel. For hvert av de angitte land vises

x = gjennomsnittlig antall sigaretter pr. person pr. år,

y = antall døde, av hjerte-og-karsykdommer, pr. hundre tusen innbyggere.

Gjennomsnittstallet x omfatter altså både røykere og ikkerøykere, av alle over 18 år. Populasjonene det ble rapportert om for y, i disse undersøkelsene, gjelder dem fra 35 til 64 år.

	x	y	
1962	3350	211.6	Canada
1962	3220	238.1	Australia
1962	3220	211.8	New Zealand
1962	2790	194.1	United Kingdom
1962	2780	124.5	Switzerland
1962	2770	187.3	Ireland
1962	2290	110.5	Iceland
1962	2160	233.1	Finland
1963	1890	150.3	West Germany
1962	1810	124.7	Netherlands
1962	1800	41.2	Greece
1962	1770	182.1	Austria
1962	1700	118.1	Belgium
1962	1680	31.9	Mexico
1963	1510	114.3	Italy
1961	1500	144.9	Denmark
1962	1410	59.7	France
1962	1270	126.9	Sweden
1961	1200	43.9	Spain
1962	1090	136.3	Norway
1962	3900		USA

En enkel lineær regresjonsmodell setter

$$y_i = a + bx_i + \varepsilon_i \quad \text{for } i = 1, \ldots, n,$$

der ε_i-ene tenkes uavhengige og $\mathrm{N}(0, \sigma^2)$. Her tar vi med de $n = 20$ land i tabellen over, der vi har både x og y; USA er altså ikke med i denne delen av analysen. Den vanlige kommandoen `lm(y ~ x)` i programpakken **R** gir bl.a. følgende:

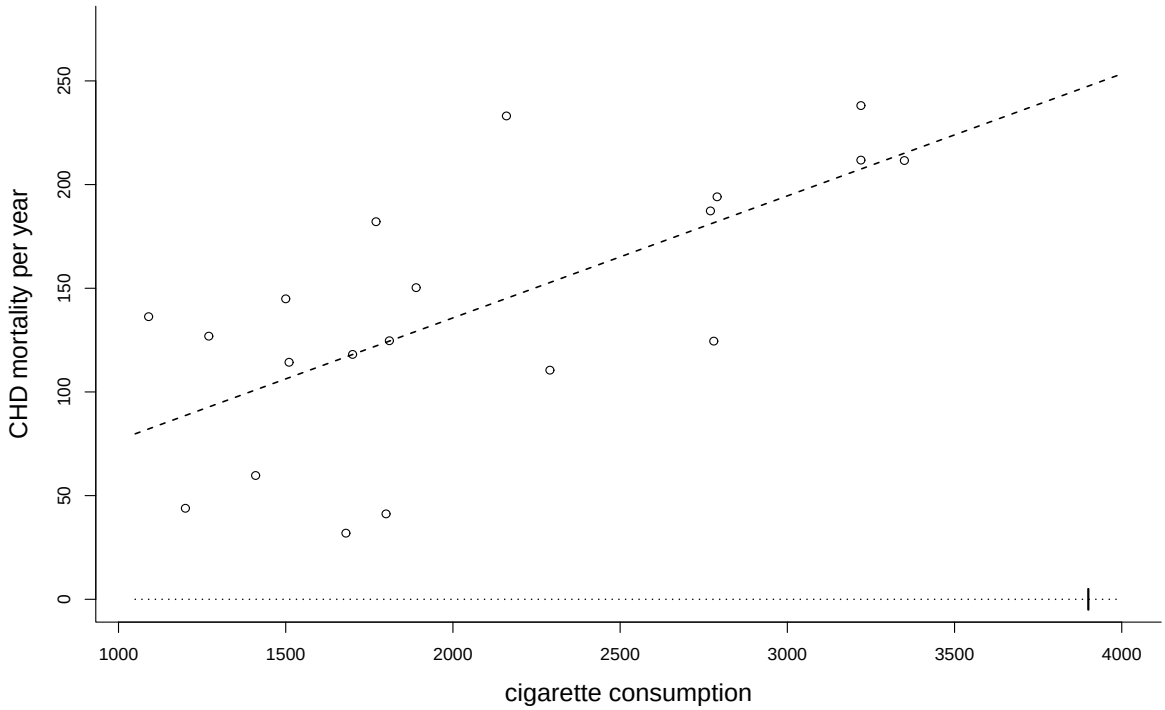
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.03462   33.28179   0.542  0.59455
x            0.05884    0.01529   3.848  0.00118 **
```

Jeg opplyser også om at $\sum_{i=1}^n (y_i - \widehat{a} - \widehat{b}x_i)^2 = 41388.09$, der $\widehat{a}$ og $\widehat{b}$ er minste-kvadratsums-estimatene. Dataene, med regresjonslinjen, er også vist frem i figuren neste side.

(a) Gi verdiene for $\widehat{a}$ og $\widehat{b}$, og vis at det vanlige estimatet for σ, det som tar utgangspunkt i forventningsretthet for $\widehat{\sigma}^2$, blir $\widehat{\sigma} = 47.951$.

(b) Dersom sigarettkonsumet i et land går ned, så meget at gjennomsnittstallet for antall sigaretter går ned med 100 (pr. person pr. år), hvor mange færre døde vil da dette landet senere kunne forvente, av hjerte-og-karsykdommer, i dette alderssegmentet fra 35 til 64 år?

(c) Sambandsstatene (USA) hadde altså så høyt sigarettkonsum i 1962 som 3900 pr. voksenperson pr. år. La Y_0 være antallet døde av hjerte-og-karsykdommer i 1962, i USA, per hundre tusen, i alderen 35 til 64 år. Gi et estimat for Y_0.



(d) I tillegg til estimatet $\widehat{y}_0$ for $Y_0 = a + bx_0 + \varepsilon_0$ ønsker vi oss et fullt prediksjonsintervall. Vi innfører $Z_0 = Y_0 - \widehat{a} - \widehat{b}x_0$, der $x_0 = 3900$. Finn forventning og varians til Z_0. Her kan du bruke følgende, fra pensum:

$$\widehat{a} + \widehat{b}x_0 = \bar{Y} + \widehat{b}(x_0 - \bar{x}),$$

der $\bar{x} = (1/n)\sum_{i=1}^n x_i = 2060.500$ og $\bar{Y} = (1/n)\sum_{i=1}^n Y_i = 139.265$ er gjennomsnittene, og $\widehat{b} = \sum_{i=1}^n (x_i - \bar{x})Y_i/M$, med $M = \sum_{i=1}^n (x_i - \bar{x})^2 = 9833895$.

(e) Beregn et intervall som med sannsynlighet 90% (eksakt eller tilnærmet) inneholder Y_0. Kommenter kort forutsetninger som ligger til grunn. Du kan få bruk for at 0.95-kvantilen for t-fordelingen med 18 frihetsgrader er `qt(0.95,18)=1.734`.

(f) Ifølge offisielt tilgjengelige kilder har Ungarn i 2014 en dødsrate på 172.6 pr. hundre tusen, relatert til hjerte-og-karsykdommer. Hva tror du sigarettkonsumet er i Ungarn?

Appendiks: en liten F-tabell

Her er en tabell over (low, up), 0.025-kvantilen og 0.975-kvantilen i F-fordelingen (Fisher-fordelingen) med frihetsgrader (m_1, m_2), der vi i denne situasjonen kun ser på $m_1 = m_2 = m$, og for m-verdier fra 15 til 45.

m1	m2	low	up
15	15	0.349	2.862
16	16	0.362	2.761
17	17	0.374	2.673
18	18	0.385	2.596
19	19	0.396	2.526
20	20	0.406	2.464
21	21	0.415	2.409
22	22	0.424	2.358
23	23	0.433	2.312
24	24	0.441	2.269
25	25	0.448	2.230
26	26	0.456	2.194
27	27	0.463	2.161
28	28	0.470	2.130
29	29	0.476	2.101
30	30	0.482	2.074
31	31	0.488	2.049
32	32	0.494	2.025
33	33	0.499	2.002
34	34	0.505	1.981
35	35	0.510	1.961
36	36	0.515	1.942
37	37	0.520	1.924
38	38	0.524	1.907
39	39	0.529	1.891
40	40	0.533	1.875
41	41	0.538	1.860
42	42	0.542	1.846
43	43	0.546	1.833
44	44	0.550	1.820
45	45	0.553	1.807