UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	ST100 — Sannsynlighetsregning og statistikk.
Eksamensdag:	Mandag 3. desember 2001.
Tid for eksamen:	09.00 – 15.00.
Oppgavesettet er på 5 sider.	
Vedlegg:	Utskrifter fra MINITAB til oppgave 3. Tabell over kumulative t-fordelinger.
Tillatte hjelpemidler:	Formelsamling for ST100, lommeregner, Haugens "Formler og tabeller", Jahren og Knutsens "Formelsamling i matematikk", Rottmanns "Mathematische Formelsammlung."

Kontroller at oppgavesettet er komplett
før du begynner å besvare spørsmålene.

Oppgave 1.

En kan måle lungekapasiteten til en person ved å registrere den kraften personen kan blåse luften ut av lungene med. Et bestemt mål for lungekapasitet har den engelske betegnelsen "peak expiratory flow rate", som ofte forkortes PEF og angis i liter per minutt.

Tabellen nedenfor gir PEF-målinger for 10 menn og 11 kvinner. Vi er interessert i å studere forskjellen i lungekapasitet for menn og kvinner.

Menn	603	628	610	639	579	637	653	612	623	550	
Kvinner	455	456	523	458	460	435	489	399	525	464	450

Vi vil først studere generelt hvordan vi kan sammenligne to grupper, A og B, og i punkt d) vende tilbake til tallene i tabellen.

(Fortsettes side 2.)

La $X_1, X_2, \ldots, X_n$ representere observasjonene fra gruppe A, og $Y_1, Y_2, \ldots, Y_m$ representere observasjonene fra gruppe B. Vi antar at alle disse stokastiske variablene er uavhengige, og at X_i-ene er normalfordelte med forventning μ_X og standardavvik σ, mens Y_j-ene er normalfordelte med forventning μ_Y og standardavvik σ. Vi er interessert i forskjellen $\mu_X - \mu_Y$ i forventningsverdi mellom de to gruppene.

a) Som estimator for $\mu_X - \mu_Y$ bruker vi $\overline{X} - \overline{Y}$. Forklar hvorfor $\overline{X} - \overline{Y}$ er normalfordelt med forventning $\mu_X - \mu_Y$ og varians $\sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)$.

La $S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ og $S_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(Y_i - \overline{Y}\right)^2$ være estimatorer for σ^2 basert på observasjonene fra gruppe A og gruppe B, henholdsvis. Vi estimerer σ^2 ved den vektete estimatoren

$$S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$$

Det er kjent at $(n+m-2)S_p^2/\sigma^2$ er kjikvadrat-fordelt med $n+m-2$ frihetsgrader, og at S_p^2 og $\overline{X} - \overline{Y}$ er uavhengige. (Du skal ikke vise dette.)

b) Forklar hvorfor

$$\frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

er t-fordelt med $n + m - 2$ frihetsgrader.

c) Utled et $100(1-\alpha)\%$ konfidensintervall for $\mu_X - \mu_Y$.

Vi ser så på PEF-målingene for menn og kvinner. Ved hjelp av MINITAB har vi beregnet en del beskrivende statistiske mål for disse dataene:

```
Descriptive Statistics: menn; kvinner

Variable           N      Mean    Median     StDev   SE Mean
menn              10    613,40    617,50     30,57      9,67
kvinner           11     464,9     458,0      36,4      11,0

Variable     Minimum   Maximum        Q1        Q3
menn          550,00    653,00    597,00    637,50
kvinner        399,0     525,0     450,0     489,0
```

d) Bruk MINITAB-resultatene til å finne et estimat for forskjellen i forventet PEF-verdi for menn og kvinner, og til å bestemme et 99% konfidensintervall for denne forskjellen. Kommenter resultatet.

(Fortsettes side 3.)

Oppgave 2.

Det er rimelig å anta at antall fødsler X i løpet av t timer ved en bestemt fødeavdeling er Poisson-fordelt med punktsannsynlighet

$$P(X = x) = \frac{(\lambda t/24)^x}{x!} e^{-\lambda t/24} \qquad x = 0, 1, 2, 3, \ldots$$

I spørsmål a) og c) vil vi anta at $\lambda = 6$.

a) Hva er forventet antall fødsler pr. døgn? Hva er sannsynligheten for at det blir født minst ett barn i løpet av et døgn? Hva er sannsynligheten for at det blir født flere enn ett barn i løpet av en time?

b) La T være tiden i timer fra midnatt (kl. 00.00) til første fødsel skjer. Vis at

$$P(T > t) = e^{-\lambda t/24} \qquad t > 0$$

Finn sannsynlighetstettheten til T, og utled et uttrykk for forventningsverdien til T som funksjon av λ. Hva er forventet tid fram til første fødsel dersom $\lambda = 6$?

c) Finn sannsynligheten for at det ikke blir født noen barn i løpet av de fire første timene etter midnatt. Anta så at det ikke har blitt født noen barn mellom midnatt og klokka 04.00. Finn sannsynligheten for at en da må vente minst fire timer til før det skjer en fødsel.

d) Sannsynligheten for at et barn som fødes er en jente, er 48.6%. La Y være antall jenter som fødes på avdelingen i løpet av et døgn. Gitt at det i løpet av et døgn er født akkurat x barn, er Y binomisk fordelt med parametre $(x, 0.486)$. (Du skal ikke vise dette.)

Vis at (den ubetingede) fordelingen til Y er Poisson med parameter 0.486λ.

Oppgave 3.

I forbindelse med påvisning av sykdommen artritt (leddbetennelse), får pasienter en injeksjon med radioaktivt gull. For å undersøke hvor lenge pasientenes blodserum vil inneholde dette stoffet, har en målt konsentrasjonen av radioaktivt gull i blodserumet (i prosent av initialkonsentrasjonen) på ulike tidspunkt etter injeksjonen (dag 0).

Tall fra ni slike målinger er gitt i tabellen nedenfor:

Dag etter injeksjon	1	1	2	2	3	5	6	6	7
Serumgull i %	94.5	93.7	80.5	81.4	67.4	49.3	46.8	42.3	45.6

(Fortsettes side 4.)

Vi ønsker å tilpasse en lineær regresjonsmodell

$$Y_i = \beta_0 + \beta_1 x_i + e_i, \qquad i = 1, ..., 9 \tag{1}$$

til observasjonene. I vedlegg 1 finner du resultatet av en slik regresjon utført i MINITAB.

a) Hvilke forutsetninger og presiseringer mangler i modellformuleringen (1)? Gjør kort rede for minste kvadraters metode. Finn estimerte verdier $\widehat{\beta}_0$ og $\widehat{\beta}_1$ for β_0 og β_1 fra MINITAB-utskriften.

b) Hva er et residual? Regn ut residualet for observasjonen ved dag 7 for den lineære modellen.

c) Diskuter, i lys av vedlagte residualplott (vedlegg 1), hvor godt modellen (1) passer til observasjonene. Foreslå eventuelt en alternativ modell som du mener er bedre.

Oppgave 4.

Vi minner om at en stokastisk variabel V er kjikvadrat-fordelt med m frihetsgrader hvis den har sannsynlighetstettheten

$$f_V(v) = \begin{cases} \frac{1}{2^{m/2}\Gamma(m/2)} v^{(m/2)-1} e^{-v/2} & \text{hvis} \quad v > 0 \\ 0 & \text{ellers} \end{cases}$$

For en slik kjikvadrat-fordelt variabel gjelder det at

$$E\left(V^r\right) = 2^r \frac{\Gamma\left(\frac{m}{2} + r\right)}{\Gamma(\frac{m}{2})} \qquad \text{hvis} \quad r > -m/2 \tag{2}$$

Du skal selv vise (2) i spørsmål f). Men uansett om du klarer å gjøre det eller ikke, kan du bruke dette resultatet i oppgaven.

La nå X være inntekten til en tilfeldig valgt lønnsmottaker i en bestemt befolkningsgruppe. Det er vanlig å anta at X er Pareto-fordelt, det vil si at X har sannsynlighetstettheten

$$f_X(x) = \begin{cases} \theta k^\theta \left(\frac{1}{x}\right)^{\theta+1} & \text{hvis} \quad x > k \\ 0 & \text{ellers} \end{cases} \tag{3}$$

Her er k minsteinntekten i den aktuelle befolkningsgruppen, mens $\theta > 1$ er en parameter som avhenger av lønnsforskjellene i gruppen. Vi vil i hele oppgaven regne med at minsteinntekten k er kjent.

(Fortsettes side 5.)

a) Vis at den kumulative sannsynlighetsfordelingen til X blir

$$F_X(x) = \begin{cases} 1 - \left(\frac{k}{x}\right)^\theta & \text{hvis} \quad x > k \\ 0 & \text{ellers} \end{cases}$$

b) Vis at $Y = 2\theta(\log X - \log k)$ er kjikvadrat-fordelt med 2 frihetsgrader.

For å bestemme parameteren θ for den aktuelle befolkningsgruppen, gjøres det $n \geq 3$ observasjoner, $X_1, X_2, \ldots, X_n$, av inntektene i denne gruppen. Du kan gå ut fra at $X_1, X_2, \ldots, X_n$ er uavhengige og identisk fordelte stokastiske variable med sannsynlighetstettheten (3).

c) Forklar hvorfor

$$\widehat{\theta} = \frac{n}{\sum_{i=1}^{n} \log X_i - n \log k} \tag{4}$$

er en rimelig estimator for θ.

d) Begrunn at $2n\theta/\widehat{\theta}$ er kjikvadrat-fordelt med $2n$ frihetsgrader, og bruk dette og (2) til å vise at $E(\widehat{\theta}) = [n/(n-1)]\theta$. Finn $\mathrm{Var}(\widehat{\theta})$.

e) Estimatoren (4) er ikke forventningsrett. Foreslå en estimator θ^* som er forventningsrett. Vis at θ^* er konsistent, dvs. at for enhver $\epsilon > 0$ vil $P(|\theta^* - \theta| > \epsilon) \to 0$ når $n \to \infty$.

f) Vis (2).

SLUTT