UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	ST101 — Innføring i statistikk og sannsynlighetsregning.
Eksamensdag:	Mandag 29. november 1993.
Tid for eksamen:	09.00 – 15.00.
Oppgavesettet er på 6 sider.	
Vedlegg:	Ingen.
Tillatte hjelpemidler:	ST101 formelsamling, kalkulator, Rottmanns "Mathematische Formelsamlung", Jahren og Knutsens "Formelsamling i matematikk" (Tapir forlag).

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

Oppgave 1.

Brystkreft er en av de vanligste kreftformer hos kvinner. Ved hjelp av en spesiell form for røntgenundersøkelse, mamografi, kan kreftsvulsten iblant oppdages i et tidligere stadium enn det som ellers ville vært tilfellet. Dermed øker sjansen for helbredelse. Vi vil i denne oppgaven foreta noen beregninger knyttet til slike mamografiundersøkelser.

De følgende (noenlunde realistiske) anslag vil bli benyttet i beregningene: Hvis en kvinne har brystkreft vil sannsynligheten for at dette oppdages ved en mamografiundersøkelse være 0,80. Hvis kvinnen ikke lider av brystkreft, vil det likevel være en sannsynlighet på 0,10 for at undersøkelsen indikerer at kvinnen har brystkreft. Sannsynligheten for at en tilfeldig valgt kvinne over 40 år har brystkreft anslås til 0,005.

a) En kvinne over 40 år møter til mamografiundersøkelse. Hva er sannsynligheten for at undersøkelsen gir et positivt resultat (dvs. indikerer at hun har brystkreft)? Hva er sannsynligheten for at den gir et negativt resultat (dvs. indikerer at hun er frisk)?

(Fortsettes side 2.)

b) Hvis kvinnen får et positivt resultat på mamografiundersøkelsen, hva er da sannsynligheten for at hun virkelig lider av brystkreft? Hvis kvinnen får et negativt resultat, hva er sannsynligheten for at hun virkelig er frisk?

c) En dag møter 25 kvinner over 40 år til mamografiundersøkelse, og for alle gir undersøkelsen et negativt resultat. Hva er sannsynligheten for at minst én av dem likevel har brystkreft?

d) Tenk deg at 50 000 kvinner over 40 år blir innkalt (og møter) til mamografiundersøkelse. Hva er det forventede antallet med brystkreft blant disse kvinnene? Hva er det forventede antallet ekte positive? Hva er det forventede antallet falske positive? (Ekte positiv betyr at undersøkelsen indikerer brystkreft hos en kvinne som faktisk har brystkreft. Falsk positiv betyr at undersøkelsen indikerer brystkreft hos en kvinne som faktisk er frisk.)

e) Det utføres i dag ikke masseundersøkelser med mamografi i Norge hvor alle kvinner over 40 år jevnlig innkalles til undersøkelse. Diskuter hvilken betydning beregningene over har når en skal vurdere om slike masseundersøkelser bør settes i gang.

Oppgave 2.

Anta at X er Laplace-fordelt (eller dobbelteksponensielt fordelt) med parametre μ og σ, dvs. at X har sannsynlighetstettheten

$$f_X(x) = \frac{1}{\sqrt{2}\sigma} e^{-\sqrt{2}\,|x-\mu|\,/\,\sigma} \qquad -\infty < x < \infty, \tag{1}$$

hvor $-\infty < \mu < \infty$ og $\sigma > 0$.

a) Vis at den kumulative fordelingen til X blir

$$F_X(x) = \begin{cases} \frac{1}{2} e^{\sqrt{2}\,(x-\mu)\,/\,\sigma} & \text{for } x < \mu \\ 1 - \frac{1}{2} e^{-\sqrt{2}\,(x-\mu)\,/\,\sigma} & \text{for } x \geq \mu \end{cases}$$

b) Bestem medianen og øvre og nedre kvartil til Laplace-fordelingen (1).

c) Vis at $\mathrm{E}X = \mu$ og $\mathrm{Var}X = \sigma^2$. (Du kan her ta for gitt at $\int_0^\infty x e^{-x} \mathrm{d}x = 1$ og $\int_0^\infty x^2 e^{-x} \mathrm{d}x = 2$; jfr. resultatene om gamma-funksjonen i starten av Oppgave 3.)

(Fortsettes side 3.)

La nå $X_1, X_2, \ldots, X_n$ være uavhengige og identisk fordelte med sannsynlighetstetthet (1). Vi vil først betrakte to estimatorer for μ, nemlig gjennomsnittet $\widehat{\mu} = \overline{X}$ og den empiriske medianen $\widetilde{\mu}$. Det kan vises at

$$\mathrm{E}\widetilde{\mu} \approx \mu \text{ og } \mathrm{Var}\widetilde{\mu} \approx \frac{1}{4nf_X(\mu)^2}$$

når n er tilstrekkelig stor. Du kan ta dette resultatet for gitt nedenfor.

d) Hva blir (tilnærmet) den relative effisiensen av $\widehat{\mu}$ relativt til $\widetilde{\mu}$? Hvilken av de to estimatorene vil du foretrekke?

Vi vil så studere to estimatorer for standardavviket σ, nemlig det empiriske standardavviket

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{2}$$

og den (normerte) empiriske kvartildifferansen

$$\widehat{\sigma} = \frac{Q_3 - Q_1}{\sqrt{2}\log 2}. \tag{3}$$

Her er Q_1 og Q_3 nedre og øvre empiriske kvartil.

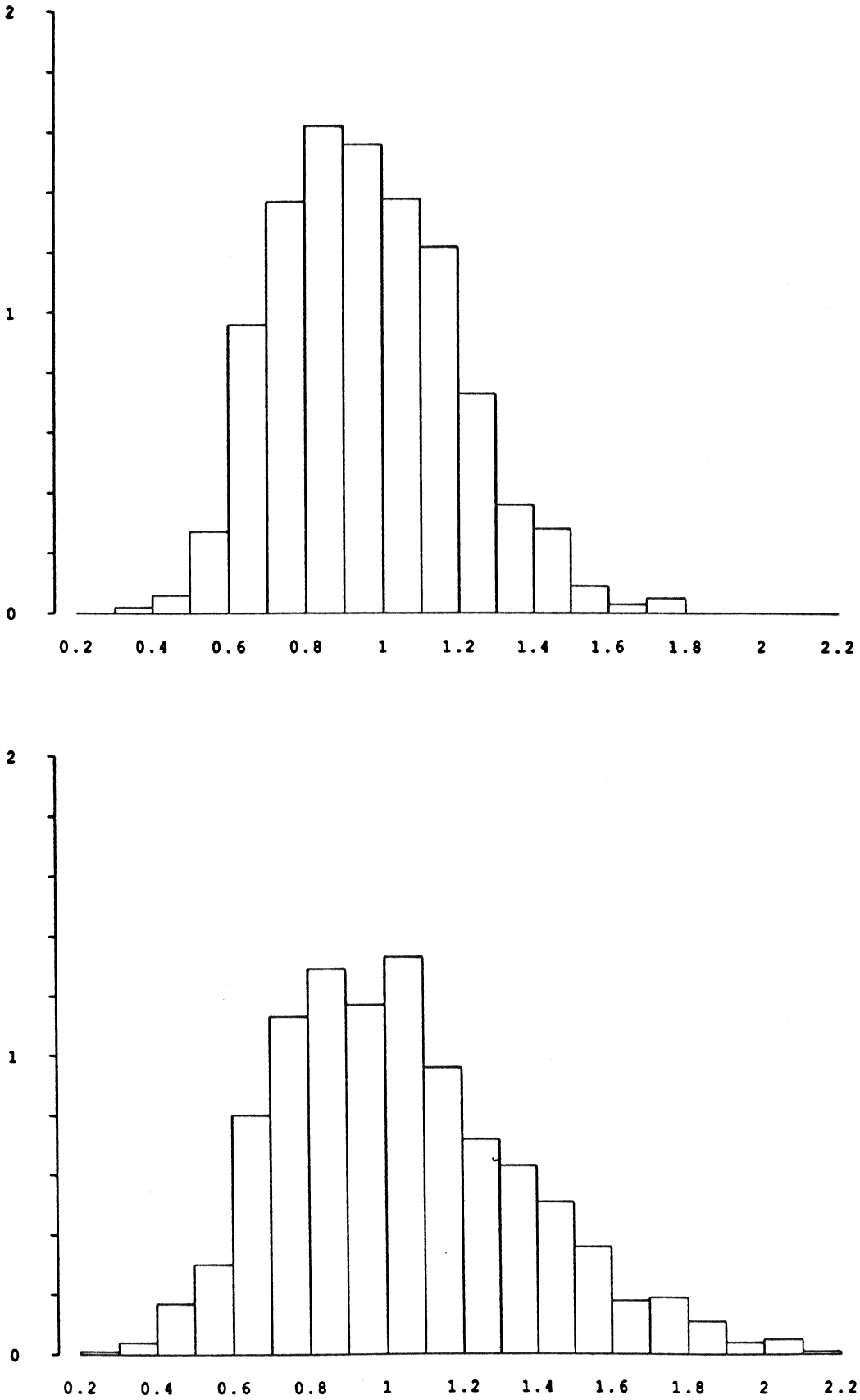
e) Gi en begrunnelse for at $\widehat{\sigma}$ er en rimelig estimator for σ.

Vi vil basere sammenligningen av de to estimatorene på stokastisk simulering, og vil i denne forbindelse nøye oss med å betrakte situasjonen der $n = 20$ og $\mu = \sigma = 1$. (Verdien av μ spiller forøvrig ingen rolle her.) Med datamaskin er det derfor generert $n = 20$ uavhengige (pseudo) stokastiske variable med sannsynlighetstetthet (1) med $\mu = \sigma = 1$. På grunnlag av disse er så verdien av estimatorene (2) og (3) beregnet. Dette er gjentatt 1000 ganger. På figuren på neste side er det gitt histogrammer over disse 1000 estimatene.

f) Forklar hvordan et slikt simuleringsforsøk kan oss informasjon om fordelingene til estimatorene (2) og (3) (når $n = 20$ og $\mu = \sigma = 1$). Bruk histogrammene til å gi et (grovt) anslag på sannsynligheten for at estimatorene (2) og (3) vil avvike høyst 0.2 fra den sanne verdien $\sigma = 1$. Hvilken av de to estimatorene vil du foretrekke?

g) Mange programpakker (blant annet BLSS) har ikke innebygd muligheten for å generere (pseudo) stokastiske variable fra Laplace-fordelingen (1). Hvordan kan vi likevel trekke slike (pseudo) stokastiske variable såsant programpakken kan generere uniformt (0,1)-fordelte (pseudo) stokastiske variable?

(Fortsettes side 4.)



Figur: Histogram over 1000 simulerte verdier av estimatorene S (øverst) og $\widehat{\sigma}$ (nederst).

(Fortsettes side 5.)

Oppgave 3.

Vi minner om at en stokastisk variabel X med sannsynlighetstetthet

$$f_X(x) = \begin{cases} \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x} & \text{for } x > 0 \\ 0 & \text{ellers} \end{cases} \tag{4}$$

sies å være *gammafordelt* med parametre $r > 0$ og $\lambda > 0$. Her er

$$\Gamma(r) = \int_0^\infty u^{r-1} e^{-u} \mathrm{d}u$$

gamma-funksjonen. Vi minner også om at gamma-funksjonen (blant annet) tilfredsstiller $(r, s > 0)$

(i) $\Gamma(1) = 1$

(ii) $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

(iii) $\Gamma(r+1) = r\Gamma(r)$

(iv) $\frac{\Gamma(r)\Gamma(s)}{\Gamma(r+s)} = \int_0^1 u^{r-1}(1-u)^{s-1}\mathrm{d}u$

Hvis vi i (4) spesielt har $r = n/2$ og $\lambda = 1/2$, hvor n er et heltall, sier vi at X er *kjikvadratfordelt* med n frihetsgrader.

Vi vil først i denne oppgaven utlede noen resultater om gammafordelingen.

a) La X være gammafordelt med parametre r og λ. Vis at

$$\mathrm{E}(X^k) = \frac{\Gamma(r+k)}{\lambda^k \Gamma(r)}.$$

For hvilke verdier av k gjelder dette resultatet? Bestem EX og Var$X$.

b) La X_1 og X_2 være uavhengige og gammafordelte stokastiske variable med parametre r_1 og λ for X_1 og r_2 og λ for X_2. Vis at da er $X_1 + X_2$ gammafordelt med parametre $r = r_1 + r_2$ og λ. Generaliser resultatet til n uavhengige gammafordelte variable. (Det kreves ikke et formelt induksjonsbevis her.)

Vi vil så benytte disse resultatene til å studere noen transformasjoner av standard normalfordelte stokastiske variable.

c) La Z være standard normalfordelt. Vis at da er Z^2 kjikvadratfordelt med 1 frihetsgrad (dvs. gammafordelt med parametre $r = 1/2$ og $\lambda = 1/2$).

d) La $Z_1, Z_2, \ldots, Z_n$ være uavhengige og standard normalfordelte. Vis at da er $\sum_{i=1}^n Z_i^2$ kjikvadratfordelt med n frihetsgrader (dvs. gammafordelt med parametre $r = n/2$ og $\lambda = 1/2$). Bruk dette til å vise at $\mathrm{E}(\sum_{i=1}^n Z_i^2) = n$ og $\mathrm{Var}(\sum_{i=1}^n Z_i^2) = 2n$.

(Fortsettes side 6.)

Vi vil til slutt i denne oppgaven betrakte en produsent av vekter som ønsker å fastlegge hvilken presisjon vektene har. For dette formålet tar han et tilfeldig utvalg på n vekter av den løpende produksjonen. Med hver av disse veier han én gang et lodd med *kjent* vekt μ_0 kg. Vi lar Y_i være den registrerte vekten (i kg) med i-te vekt, og antar at $Y_1, Y_2, \ldots, Y_n$ er uavhengige og identisk $N(\mu_0, \sigma^2)$-fordelte. Her er altså μ_0 en kjent størrelse, mens σ er en ukjent parameter som vi ønsker å estimere.

e) Vis at

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_0)^2$$

er en forventningsrett estimator for σ^2.

f) Vis at $n\widehat{\sigma}^2/\sigma^2$ er kjikvadratfordelt med n frihetsgrader og bestem $\mathrm{Var}(\widehat{\sigma}^2)$. Er $\widehat{\sigma}^2$ konsistent?

g) Utled et $100(1-\alpha)\%$ konfidensinterval for σ. Intervallet skal uttrykkes ved hjelp av n, $\widehat{\sigma}$, c_1 og c_2. Her er c_1 og c_2 gitt ved at $\mathrm{P}(\chi^2_n \leq c_1) = \mathrm{P}(\chi^2_n \geq c_2) = \alpha/2$, hvor χ^2_n er kjikvadratfordelt med n frihetsgrader.

SLUTT