UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	ST101 — Innføring i statistikk og sannsynlighetsregning.
Eksamensdag:	Mandag 28. november 1994.
Tid for eksamen:	09.00 – 15.00.
Oppgavesettet er på 5 sider.	
Vedlegg:	Opptrykk av tabell over kumulative normalfordelinger.
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.

Ved uttesting av legemidler ("kliniske forsøk") går man mange ganger fram på omtrent følgende måte. Anta at $n+m$ pasienter står til disposisjon og at virkningen av to legemidler A og B skal sammenlignes. Vi anvender A på n (tilfeldig valgte) personer og måler virkningen. B prøves ut på de m øvrige. Dette gir observasjonsmaterialet

$X_1, \ldots, X_n$ (for medisin A)
$Y_1, \ldots, Y_m$ (for medisin B).

Som modell antas at $X, \ldots, X_n, Y, \ldots, Y_m$ er uavhengige. Alle $X_1, \ldots, X_n$ har samme fordeling. Det har også $Y_1, \ldots, Y_m$. Dessuten er

$$E(X_j) = \mu_A\,, \qquad E(Y_k) = \mu_B$$
$$\mathrm{var}(X_j) = \sigma^2\,, \qquad \mathrm{var}(Y_k) = \sigma^2\,.$$

μ_A og μ_B kan da fortolkes som gjennomsnittsvirkningene av de to medisinene.

La

$$\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j\,, \qquad \bar{Y} = \frac{1}{m}\sum_{k=1}^{m} Y_k$$

(Fortsettes side 2.)

være estimatorer for μ_A og μ_B henholdsvis.

a) Gjør rede for at de to estimatorene er forventningsrette. Hva blir deres varians?

b) Under hvilken forutsetning vil $\bar{X}$ og $\bar{Y}$ være tilnærmet normalfordelte? Angi parametrene i normalfordelingen.

Ofte vil interessen rette seg mot *forskjellen* i de to medisinenes effekt. La

$$D = \bar{Y} - \bar{X}$$

c) Vis at $E(D) = \mu_B - \mu_A$ og at $\text{var}(D) = \sigma^2(1/n + 1/m)$.

Legg til grunn at $\bar{X}$ og $\bar{Y}$ begge er tilnærmet normalfordelte.

d) Hvorfor blir da D også tilnærmet normalfordelt? Kort begrunnelse.

Anta at $\mu_B = \mu_A$, $\sigma = 1$ og at $n = m = 18$.

e) Beregn sannsynligheten (tilnærmet) for at $D \geq 2/3$.

f) Sett at vi har observert at $\bar{X} = 1/3$ og at $\bar{Y} = 1$. Ville du da tro at de to medisinene har samme gjennomsnittsvirkning? Legg til grunn at $\sigma = 1$ og at $n = m = 18$, og bruk beregningen i e) som støtte for din vurdering.

Oppgave 2.

Gitt en populasjon med N individer hvorav r har et kjennetegn A og $N - r$ ikke har det. Anta at vi trekker tilfeldig n individer *uten tilbakelegging*. For j-te trekning la

$$\begin{aligned} I_j &= 1, \quad && \text{dersom } j\text{-te uttrukne individ er av type A} \\ &= 0, && \text{ellers.} \end{aligned}$$

Da er $X = I_1 + \cdots + I_n$ antall individer i utvalget med kjennetegn A.

a) Finn punktsannsynligheten til I_j.

b) Vis at $E(I_j) = r/N$ og at $\text{var}(I_j) = \frac{r}{N}\left(1 - \frac{r}{N}\right)$.

c) For $j \neq k$, vis at den simultane punktsannsynligheten for (I_j, I_k) blir

$$\begin{aligned} P(I_j = 0, I_k = 0) &= \frac{N-r}{N} \cdot \frac{N-r-1}{N-1} \\ P(I_j = 1, I_k = 1) &= \frac{r}{N} \cdot \frac{r-1}{N-1} \\ P(I_j = 1, I_k = 0) &= P(I_j = 0, I_k = 1) = \frac{r}{N} \cdot \frac{N-r}{N-1} \end{aligned}$$

(Fortsettes side 3.)

d) Hvorfor blir

$$E(I_j I_k) = P(I_j = 1, I_k = 1) = \frac{r}{N} \cdot \frac{r-1}{N-1}\,?$$

Bruk dette til å vise at

$$\operatorname{cov}(I_j, I_k) = -\frac{r}{N}\Big(1 - \frac{r}{N}\Big)\frac{1}{N-1}\,.$$

e) Bevis at

$$\begin{aligned} E(X) &= n\frac{r}{N} \\ \operatorname{var}(X) &= n\frac{r}{N}\Big(1 - \frac{r}{N}\Big)\frac{N-n}{N-1} \end{aligned}$$

[Hint: Anvend (uten bevis) at $\operatorname{var}(X) = \sum_{j=1}^{n} \operatorname{var}(I_j) + \sum_{j \neq k} \operatorname{cov}(I_j, I_k)$.]

La $r = pN$. *Hold p (og n) fast* og la $N \to \infty$.

f) Hvilke uttrykk vil $E(X)$ og $\operatorname{var}(X)$ da nærme seg mot? Kjenner du igjen disse uttrykkene? Forklar intuitivt hvorfor disse fremkommer.

Oppgave 3.

I ST101-læreboka diskuteres et eksempel (Collins-saken) der det var kjent at den skyldige S i en forbrytelse var bærer av et bestemt kjennetegn E. Spørsmålet var hvilken beviskraft som lå i at en anklaget A også hadde dette kjennetegnet. I denne oppgaven skal vi analysere problemstillingen fra en annen synsvinkel.

Legg til grunn at S og A begge er med i en populasjon med N individer. Det er ikke kjent hvor mange individer som har kjennetegnet E. La X være dette ukjente antallet. Forutsett at $\theta = P(E)$ for hvert individ samt at populasjonens størrelse N er kjent.

a) Forklar hvorfor X kan være binomisk fordelt. Hvilken tilleggsforutsetning i forhold til teksten må da innføres?

Ofte vil E i en slik situasjon være sjeldent forekommende slik at θ er liten.

b) Gjør rede for at X under denne forutsetning kan ansees å være poissonfordelt med parameter $\lambda = N\theta$.

(Fortsettes side 4.)

Vi legger poisson-modellen til grunn i det etterfølgende. λ ansees kjent. Dersom X hadde vært kjent, hadde det vært naturlig å si at

$$P(A \text{ er } S \mid X = x) = \frac{1}{x} \tag{1}$$

I praksis er X ukjent. Det vi vet er at $X \geq 1$ (siden den skyldige har E).

c) Vis at

$$P(A \text{ er } S \mid X \geq 1) = \sum_{x=1}^{\infty} \frac{1}{x} P(X = x \mid X \geq 1)$$

d) Gjør rede for at

$$P(X = x \mid X \geq 1) = \frac{\lambda^x}{x!}(e^{\lambda} - 1)^{-1}$$

under poisson-modellen, og herved også at

$$P(A \text{ er } S | X \geq 1) = (e^{\lambda} - 1)^{-1} \sum_{x=1}^{\infty} \frac{1}{x} \frac{\lambda^x}{x!}\,.$$

Summen kan ikke bestemmes i endelig form, men en god tilnærming for små λ er

$$P(A \text{ er } S \mid X \geq 1) \approx 1 - \frac{\lambda}{4} + \frac{\lambda^2}{72} \tag{$*$}$$

Dette resultat, som vil bli brukt mot slutten av oppgaven, skal ikke bevises.

Resonnementet ovenfor er grunnleggende det samme som i ST101-boka, selv om det avviker i detalj. En alternativ fremgangsmåte er følgende. Det vi vet er at den skyldige S har E. Antallet X med E i populasjonen er derfor

$$X = 1 + Y \tag{2}$$

der Y er antallet med E blant de $N - 1$ personene som er uskyldige. Y kan da ansees å være poissonfordelt med parameter $(N - 1)\theta \approx N\theta = \lambda$ om N er stor. Dette antas i fortsettelsen.

e) Finn punktsannsynligheten til X under antakelse (2).

f) Vis at under den nåværende betraktningsmåte blir

$$\begin{aligned} P(A \text{ er } S) &= \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{x!} e^{-\lambda} \\ &= \lambda^{-1}(1 - e^{-\lambda}) \end{aligned} \tag{$**$}$$

[Hint: Utnytt $P(A \text{ er } S) = \sum_x P(A \text{ er } S \mid X = x) P(X = x)$].

(Fortsettes side 5.)

La q^* være sannsynligheten for at A ikke er S under beregningen $(*)$ og q^{**} tilsvarende under $(**)$. Vi har da at

$$q^* \approx \lambda/4 - \lambda^2/72, \qquad q^{**} = 1 - \lambda^{-1}(1 - e^{-\lambda})$$

h) Beregn q^* og q^{**} når $\lambda = 0.01, 0.1$ og 0.2. Sammenlign tallene og kommenter. Kan det ha praktisk betydning hvilken av de to metodene som følges? Hvilken av dem finner du mest opplysende?

SLUTT