UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	ST 102 — Videregående kurs i statistikk.
Eksamensdag:	Lørdag 1. juni 1996.
Tid for eksamen:	09.00 – 15.00.
Oppgavesettet er på 5 sider.	
Vedlegg:	Ingen.
Tillatte hjelpemidler:	Formelsamling for ST 101, ST 102, lommekalkulator.

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

Oppgave 1.

a) La F være en kumulativ sannsynlighetsfordeling. Fordelingens p-fraktil, t_p, defineres ved $p = F(t_p)$ eller $t_p = F^{-1}(p)$. Finn p-fraktilen til den kumulative sannsynlighetsfordelingen, G, gitt ved

$$G(t) = \begin{cases} 0 & \text{hvis } t < 0 \\ t^n & \text{hvis } 0 \leq t \leq 1 \\ 1 & \text{hvis } t > 1 \end{cases}$$

der n er et helt, positivt tall.

La i det følgende $X_1, X_2, \ldots, X_n$ være uavhengige og identisk uniformt fordelte over $(0, \theta)$.

b) La $Y = \max\limits_{1 \leq i \leq n}(X_i) = X_{(n)}$. Vis at Y/θ har kumulativ sannsynlighetsfordeling G.

c) Utled et $(1-\alpha)$-konfidensintervall for θ basert på Y.

(Fortsettes side 2.)

d) Hvis $n = 2m + 1$, har den empiriske median, $V = \underset{1 \le i \le n}{\text{med}}(X_i) = X_{(m+1)}$, sannsynlighetstettheten

$$h(v) = \begin{cases} \frac{\Gamma(2m+2)}{[\Gamma(m+1)]^2}\left(\frac{v}{\theta}\right)^m\left(1-\frac{v}{\theta}\right)^m \cdot \frac{1}{\theta} & \text{hvis } 0 \le v \le \theta \\ 0 & \text{ellers} \end{cases}$$

Vis at $W = \frac{V}{\theta - V}$ er F-fordelt med $2m + 2$ og $2m + 2$ frihetsgrader.

Vink: Finn først et uttrykk for den kumulative sannsynlighetsfordelingen til W.

e) Utled et $(1 - \alpha)$-konfidensintervall for θ basert på V.

f) Anta $n = 5$ $(m = 2)$ og at resultatet av de 5 målingene er:

0.03 0.62 0.88 0.09 0.91

Beregn de to konfidensintervallene utledet i punktene c) og e) når $\alpha = 0.05$. (2.5% og 97.5% fraktilene i F-fordelingen med 6 og 6 frihetsgrader er henholdsvis 0.1718 og 5.8198.)

Oppgave 2.

For å undersøke om det er en sammenheng mellom sportslig aktivitet og prestasjoner til eksamen, ble et utvalg på i alt 852 hovedfagstudenter i realfag intervjuet. Hver student måtte svare på om han/hun drev aktiv konkurranse-idrett, trimmet regelmessig eller ingen av delene. Samtidig ble det notert om gjennomsnittskarakter til cand.mag.graden var dårligere enn 2.5 eller ikke. Tabellen nedenfor angir hvor mange studenter som falt i hver gruppe (dataene er konstruert):

	aktiv idrett	trimmer regelmessig	ingen/lite mosjon
2.5 eller bedre	63	125	238
dårligere enn 2.5	42	94	290

a) Angi og begrunn en modell for observasjonene. Still opp en nullhypotese og alternativ hypotese.

(Fortsettes side 3.)

b) Gir resultatet av denne undersøkelsen grunn til å påstå at det er noen sammenheng mellom sportslig aktivitet og eksamensresultat? Bruk 5% signifikansnivå. (95% fraktilen i $\mathcal{X}^2$-fordelingen med 2 frihetsgrader er 5.99.)

c) I dette punktet er vi bare interessert i eksamensresultatene til studenter som driver aktiv idrett. Gir resultatet av denne undersøkelsen grunn til å påstå at sannsynligheten for å få 2.5 eller bedre i cand.mag. gjennomsnitt er større enn 1/2? (95% fraktilen i $N(0,1)$-fordelingen er 1.645.)

Oppgave 3.

For en del år siden ble det på Kypros gjort flere funn av gamle sølvmynter, som viste seg å være preget i 4 forskjellige tidsperioder. Prosent sølvinnhold (% Ag) ble målt i henholdsvis $n_1 = 9$, $n_2 = 7$, $n_3 = 4$ og $n_4 = 7$ av myntene fra hver av de 4 periodene.

La Y_{ij} være % Ag i den i-te målte mynten fra j-te periode. Anta modellen:

$y_{ij} = \mu_j + \varepsilon_j$, $i = 1, \ldots, n_j$, $j = 1, \ldots, 4$ der alle ε_{ij}-ene er uavhengige og identisk $N(0, \sigma^2)$-fordelte, $\mu_1, \ldots, \mu_4$ og σ^2 er ukjente parametre.

For å undersøke om forventet % Ag i mynter fra de 4 tidsperiodene er forskjellig, skal hypotesen H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ testes mot alternativet H_1: ikke alle μ_j er like. Her er forutsetningene for å bruke en F-test oppfylt.

a) Fyll ut nedenstående ufullstendige variansanalysetabell (Anova Table)

Source	*df*	SS	MS	F
Treatment		37.784		
Residual				
Total		48.763		

og gjennomfør testing av H_0 mot H_1. Bruk 5% signifikansnivå. (95%-fraktilen i F-fordelingen med 3 og 23 frihetsgrader er 3.028.)

b) hvis H_0 forkastes, er en interessert i å konkludere med en eller flere av de 12 alternative hypotesene H_{ij}: $\mu_i > \mu_j$, $i \neq j$.

(Fortsettes side 4.)

(i) Angi en test med signifikansnivå α for H_0 mot H_{12}-alternativet. Bruk tilsvarende tester for å teste mot de andre alternative hypotesene.

(ii) En ønsker at sannsynligheten for å påstå minst en av de alternative hypotesene når H_0 er riktig, skal være høyst 5%. Vis at dette kravet tilfredsstilles dersom en bruker signifikansnivå $100 \cdot \frac{0.05}{12}\%$ for hver av testene.

(iii) Hvilke av de alternative hypotesene aksepteres når gjennomsnittlig % Ag fra de 4 periodene er henholdsvis $\overline{y}_{\cdot 1} = 6.744$, $\overline{y}_{\cdot 2} = 8.243$, $\overline{y}_{\cdot 3} = 4.875$, $\overline{y}_{\cdot 4} = 5.614$? ($100 \cdot (1 - \frac{0.05}{12})$%-fraktilen i t-fordelingen med 23 frihetsgrader er 2.886).

Oppgave 4.

En bedrift vil undersøke om det er forskjell i effektiviteten av to produksjonsmetoder, A og B. Effektiviteten måles ved hvor lang tid det tar å produsere en enhet. Bedriften velger å gjennomføre et forsøk med parvise sammenligninger, slik at hver av n arbeidere skal produsere en enhet både ved metode A og ved metode B – i tilfeldig rekkefølge.

a) Gi en kort begrunnelse for hvorfor det er valgt en forsøksplan med parvise sammenligninger.

b) La $X_i(Y_i)$ være tiden i-te arbeider bruker for å produsere en enhet ved metode $A(B)$, $i = 1, 2, \ldots, n$. Gi en begrunnelse for modellen:

$$M\colon \begin{cases} \text{alle observasjoner er uavhengige} \\ X_i \sim N(\mu_i, \sigma^2),\ \ Y_i \sim N(\mu_i + \delta, \sigma^2),\ \ i = 1, 2, \ldots, n \\ \text{der } \mu_1, \ldots, \mu_n,\ \delta \text{ og } \sigma^2 \text{ er ukjente parametre} \end{cases}$$

I modellen M er μ_i forventet tid som i-te arbeider bruker for å produsere en enhet ved metoden A. Noen hevder at det er produksjonsmetodene som avgjør forventet tid det tar å produsere en enhet og at arbeiderne har liten innflytelse. Da en ved senere og lignende undersøkelser ønsker færrest mulig ukjente parametre i modellen, vil en bruke det innsamlede observasjonsmaterialet til å teste

$$H_0\colon \mu_1 = \mu_2 = \cdots = \mu_n$$

mot H_1: ikke alle μ_i-ene er like.

(Fortsettes side 5.)

Sett $Q_1 = \sum_{i=1}^{n}[(X_i + Y_i) - (\overline{X} + \overline{Y})]^2$ og $Q_2 = \sum_{i=1}^{n}[(X_i - Y_i) - (\overline{X} - \overline{Y})]^2$.

c) Forklar hvorfor $Q_2/2\sigma^2$ er $\mathcal{X}^2$-fordelt med $n-1$ frihetsgrader under a priori betingelser, dvs. modellen M. Benytt dette til å finne $E(Q_2)$.

d) Beregn $E(Q_1)$ og vis at $E(Q_1) \geq E(Q_2)$ under M og $E(Q_1) = E(Q_2)$ under H_0.

e) Det kan vises at Q_1 og Q_2 er stokastisk uavhengige (dette skal du ikke vise), men hvorfor er $Q_1/2\sigma^2$ $\mathcal{X}^2$-fordelt med $(n-1)$ frihetsgrader under H_0?

f) Benytt resultatene fra c)–e) til å foreslå en rimelig test for H_0 mot H_1 som har signifikansnivå α.

SLUTT