NTNU

Kunnskap for en bedre verden

Institutt for matematiske fag

# Eksamensoppgave i **ST1201/ST6201 Statistiske metoder**

**Faglig kontakt under eksamen:** Nikolai Ushakov

**Tlf:** 45128897

**Eksamensdato:** 04. desember 2015

**Eksamenstid (fra–til):** 09:00 – 13:00

**Hjelpemiddelkode/Tillatte hjelpemidler:** C:
– Tabeller og formler i statistikk, Tapir forlag,
– K.Rottman. Matematisk formelsamling,
– Ett gult ark (A4 med stempel) med egne håndskrevne formler og notater,
– Kalkulator: HP30S, Citizen SR-270X, Citizen SR-270X College eller Casio fx-82ES PLUS.

**Annen informasjon:**

**Målform/språk:** bokmål

**Antall sider:** 4

**Antall sider vedlegg:** 0

**Kontrollert av:**

Dato Sign

**Oppgave 1**

La $X$ være $\chi^2$-fordelt med $2n$ frihetsgrader ($n > 2$).

**a)** Vis at

$$\mathrm{E}(X^{-1}) = \frac{1}{2(n-1)} \quad \text{og} \quad \mathrm{E}(X^{-2}) = \frac{1}{4(n-1)(n-2)}.$$

**Oppgave 2**

Litteraturen oppgir at gjennomsnittlig lengde i Norge av halen til en pattedyrart er 30 cm. En biolog mener at gjennomsnittet (det vil si forventningsverdien for halelengden til et tilfeldig valgt individ) er større, og hun gjør et forsøk for å undersøke dette. Hun får målt halelengdene $y_i$ til et tilfeldig utvalg på 10 individer, og får disse resultatene (i cm):

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $y_i$ | 32.8 | 36.8 | 30.9 | 34.0 | 38.2 | 33.4 | 21.0 | 33.7 | 34.6 | 26.2 |

Det oppgis at $\bar{y} = 32.16$ og $\sum(y_i - \bar{y})^2 = 233.324$.

**a)** Utfør en test for å undersøke om forventet halelengde er større enn 30 cm. Anta at halelengden er normalfordelt. Bruk signifikansnivå $\alpha = 0.05$.

**b)** Finn et 95%-konfidensintervall for forventet halelengde.

Også breddegraden $x_i$ der dyret oppholdt seg da halen ble målt ble registrert. Tabellen viser breddegrad og halelengde for hvert dyr:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $x_i$ | 63.9 | 61.5 | 64.8 | 65.5 | 59.0 | 58.5 | 68.5 | 66.0 | 66.0 | 66.8 |
| $y_i$ | 32.8 | 36.8 | 30.9 | 34.0 | 38.2 | 33.4 | 21.0 | 33.7 | 34.6 | 26.2 |

Gå ut fra en lineær regresjonsmodell, der breddegrad er forklaringsvariabel og halelengde responsvariabel. Biologen har en mistanke om at halelengden minker med økende breddegrad.

Det oppgis at $\bar{x} = 64.05$, $\sum(x_i - \bar{x})^2 = 100.465$, $\sum(x_i - \bar{x})(y_i - \bar{y}) = -105.88$.

**c)** Estimer regresjonslinja (finn $\hat{\beta}_0$ og $\hat{\beta}_1$). Utfør en test for å undersøke biologens mistanke. Bruk signifikansnivå 0.05. Bruk at $\sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = 121.737$.

**Oppgave 3**

Anta at $X_1, ..., X_n$ er et tilfeldig utvalg fra en normalfordeling med forventning $\mu_x$ og varians $\sigma^2$, og at $Y_1, ..., Y_m$ er et tilfeldig utvalg fra en normalfordeling med forventning $\mu_y$ og varians $k\sigma^2$. Anta dessuten at $X$-ene og $Y$-ene alle er uavhengige av hverandre. Parametrene $\mu_x$ og $\mu_y$ er ukjente og vi skal i denne oppgaven se på hvordan vi kan estimere og lage konfidensintervall for differansen $\mu_x - \mu_y$.

Som estimator for $\mu_x$ og $\mu_y$ benyttes henholdsvis $\bar{X}$ og $\bar{Y}$.

**a)** Begrunn at $\bar{X} - \bar{Y}$ er normalfordelt og vis at

$$\mathrm{E}(\bar{X} - \bar{Y}) = \mu_x - \mu_y, \quad \mathrm{Var}(\bar{\mathrm{X}} - \bar{\mathrm{Y}}) = \sigma^2 \left( \frac{1}{n} + \frac{k}{m} \right).$$

**b)** Anta i dette punktet at begge parametrene $\sigma^2$ og $k$ er kjente. Utled da et $(1 - \alpha)$-konfidensintervall for differansen $\mu_x - \mu_y$.

I resten av oppgaven skal vi anta at parameteren $\sigma^2$ er ukjent, mens parameteren $k$ er kjent.

La $S_x^2$ og $S_y^2$ være empirisk varians for henholdsvis $X$-ene og $Y$-ene, dvs.

$$S_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2, \quad S_y^2 = \frac{1}{m-1} \sum_{i=1}^{m} (Y_i - \bar{Y})^2.$$

Da har vi som kjent at

$$\frac{(n-1)S_x^2}{\sigma^2} \sim \chi^2_{n-1}, \quad \frac{(m-1)S_y^2}{k\sigma^2} \sim \chi^2_{m-1},$$

samt at $S_x^2$ og $\bar{X}$ er uavhengige og $S_y^2$ og $\bar{Y}$ er uavhengige.

**c)** Vis at

$$S_p^2 = \frac{n-1}{n+m-2} S_x^2 + \frac{m-1}{k(n+m-2)} S_y^2$$

er en forventningsrett estimator for $\sigma^2$ og at

$$(n + m - 2) S_p^2 / \sigma^2 \sim \chi^2_{n+m-2}.$$

Vis videre at

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{\sqrt{S_p^2 (\frac{1}{n} + \frac{k}{m})}}$$

er Student $t$-fordelt med $n+m-2$ frihetsgrader. I utledningene i dette punktet, spesifiser spesielt hvilke kjente egenskaper og sammenhenger mellom ulike typer fordelinger du benytter, og forklar hvorfor vilkårene for disse egenskapene er oppfylt i den aktuelle situasjonen.

**d)** Utled et $(1-\alpha)$-konfidensintervall for differansen $\mu_x - \mu_y$.

Anta at du får lov til å være med i planleggingen av forsøket. Totalt antall målinger som kan utføres, $N = n + m$, er gitt ut fra økonomien i prosjektet, men under denne bibetingelsen kan du bestemme antall målinger av hver type, $n$ og $m$. Hvordan vil du velge $n$ og $m$ for at forventet lengde på konfidensintervallet skal bli minst mulig?

## Oppgave 4

Som regel er mars kaldere enn april i Norge. La $X$ være gjennomsnittstemperaturen i mars og $Y$ gjennomsnittstemperaturen i april ved Værnes et tilfeldig valgt år, begge målt i °C. Anta at $X$ er normalfordelt med forventningsverdi $\mu_X$ og varians $\sigma^2$, og at $Y$ er normalfordelt med forventningsverdi $\mu_Y$ og varians $\sigma^2$. Gjennomsnittstemperaturen i °C ved Værnes for årene 2001-2012 var slik:

| | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 |
|---|---|---|---|---|---|---|
| $x_i$ (mars) | -2.5 | 0.5 | 3.3 | 2.6 | -0.7 | -4.6 |
| $y_i$ (april) | 4.1 | 7.2 | 5.0 | 7.9 | 5.8 | 4.9 |

| | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 |
|---|---|---|---|---|---|---|
| $x_i$ (mars) | 3.3 | 0.8 | 1.9 | -0.5 | 1.2 | 3.8 |
| $y_i$ (april) | 5.0 | 5.9 | 6.9 | 4.8 | 6.7 | 3.2 |

Det oppgis at

$$\sum_{i=1}^{12} x_i = 9.10,$$

$$\sum_{i=1}^{12} y_i = 67.40,$$

$$\sum_{i=1}^{12} x_i^2 = 77.07,$$

$$\sum_{i=1}^{12} y_i^2 = 399.30,$$

$$\sum_{i=1}^{12} (x_i - y_i)^2 = 364.53,$$

der $i = 1$ står for år 2001, $i = 2$ for 2002 osv.

**a)** Anta at mars-temperaturene fra år til år er uavhengige, og april-temperaturene fra år til år er uavhengige. Vi ønsker ved hypotesetesting å prøve å påvise at differansen mellom forventet gjennomsnittstemperatur i april og mars er mindre enn 5°C. Vi kan enten bruke en to-utvalgstest eller en partest. Hva vil du gjøre? Argumenter for valget ditt, og utfør testen du velger. Bruk signifikansnivå $\alpha = 0.05$. $\sigma^2$ er ukjent.

## Oppgave 5

Vi har to tilfeldige (stokastiske) variabler $X$ og $Y$. La $X$ ha varians $\mathrm{Var}(X) = 5$, og $Y$ ha varians $\mathrm{Var}(Y) = 9$. Videre er kovariansen mellom $X$ og $Y$ gitt som $\mathrm{Cov}(X, Y) = 1$.

**a)** Regn ut $\mathrm{Cov}(2X + Y, X - Y)$.