Examinator/Examiner: Zhenxia Liu (Tel: 070 0895208). Please answer in ENGLISH if you can.

a. You are permitted to bring:
   - a calculator;
   - formel -och tabellsamling i matematisk statistik (from MAI);
   - TAMS24: Notations and Formulas (by Xiangfeng Yang)

b. Scores rating: 8-11 points giving rate 3; 11.5-14.5 points giving rate 4; 15-18 points giving rate 5.

English Version

1 (3 points)

To measure the state of health of a lake, sometimes one uses the number of cell clumps in a unit volume. From 100 different water samples which are taken randomly from a lake, one has counted the number of cell clumps $x_1, \ldots, x_{100}$. Results:

Number of clumps i	0	1	2	3	4	5	6	7	≥ 8
Frequencies N_i	7	15	18	26	20	6	5	3	0

Use χ^2-test to test the following hypothesis with a significance level $\alpha = 0.05$

$$H_0 : X \sim Po(\mu).$$

Solution. There is an unknown parameter μ in the $Po(\mu)$, and we can estimate (from method of moments or maximum-likelihood method) as

$$\hat{\mu} = \bar{x} = \frac{0 \times 7 + 1 \times 15 + 2 \times 18 + 3 \times 26 + 4 \times 20 + 5 \times 6 + 6 \times 5 + 7 \times 3}{100} = 2.9$$

We can find theoretical probabilities as follows

$$\begin{aligned}
p_1 &= P(Po(2.9) = 0) = 0.0550, & np_1 &= 5.50;\\
p_2 &= P(Po(2.9) = 1) = 0.1596, & np_2 &= 15.96;\\
p_3 &= P(Po(2.9) = 2) = 0.2314, & np_3 &= 23.14;\\
p_4 &= P(Po(2.9) = 3) = 0.2237, & np_4 &= 22.37;\\
p_5 &= P(Po(2.9) = 4) = 0.1622, & np_5 &= 16.22;\\
p_6 &= P(Po(2.9) = 5) = 0.0940, & np_6 &= 9.40;\\
p_7 &= P(Po(2.9) = 6) = 0.0455, & np_7 &= 4.55;\\
p_8 &= P(Po(2.9) = 7) = 0.0188, & np_8 &= 1.88;\\
p_9 &= P(Po(2.9) \geq 8) = 0.0098, & np_9 &= 0.98;
\end{aligned}$$

We can see that $\sum_{i=1}^{9} p_i = 1$. Since $np_7 < 5$, $np_8 < 5$ and $np_9 < 5$, we need to combine these three groups. Thus the new groups are

$$\begin{aligned}
p_1 &= P(Po(2.9) = 0) = 0.0550, & np_1 &= 5.50;\\
p_2 &= P(Po(2.9) = 1) = 0.1596, & np_2 &= 15.96;\\
p_3 &= P(Po(2.9) = 2) = 0.2314, & np_3 &= 23.14;\\
p_4 &= P(Po(2.9) = 3) = 0.2237, & np_4 &= 22.37;\\
p_5 &= P(Po(2.9) = 4) = 0.1622, & np_5 &= 16.22;\\
p_6 &= P(Po(2.9) = 5) = 0.0940, & np_6 &= 9.40;\\
p_7 &= P(Po(2.9) \geq 6) = 0.0741, & np_7 &= 7.41.
\end{aligned}$$

Therefore the test statistic is

$$TS = \sum_{i=1}^{7} \frac{(N_i - np_i)^2}{np_i} = 4.36,$$

and the rejection region is

$$C = (\chi^2_\alpha(k - 1 - \#\text{of unknown parameters}), \quad \infty) = (\chi^2_{0.05}(7 - 1 - 1), \quad \infty) = (11.07, \quad \infty).$$

Since $TS \notin C$, we don't reject H_0, namely, we do not know if it is a Poisson distribution $Po(\mu)$. □

2 (3 points)

Suppose that the distribution of a population has the probability density function

$$f(x) = \begin{cases} 3\theta x^{3\theta-1} & \text{if } 0 \leq x \leq 1; \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is an unknown parameter. A sample $\{x_1, x_2, \ldots, x_n\}$ from this population is now given.
(2.1). (1.5p) Find a point estimate $\hat{\theta}_{MM}$ of θ using Method of Moments.
(2.2). (1.5p) Find a point estimate $\hat{\theta}_{ML}$ of θ using Maximum-Likelihood method.

Solution. (2.1) The first equation from the Method of Moments is $E(X) = \bar{x}$. Since $E(X) = \int_0^1 x f(x)dx = \frac{3\theta}{3\theta+1}$, we have

$$\frac{3\theta}{3\theta + 1} = \bar{x}, \text{ thus } \hat{\theta}_{MM} = \frac{\bar{x}}{3 - 3\bar{x}}.$$

(2.2) The likelihood function is

$$L(\theta) = f(x_1)f(x_2)\ldots f(x_n) = 3\theta x_1^{3\theta-1} \cdot 3\theta x_2^{3\theta-1} \ldots 3\theta x_n^{3\theta-1} = (3\theta)^n \cdot (x_1 x_2 \ldots x_n)^{3\theta-1}.$$

Maximizing $L(\theta)$ is equivalent to maximize $\ln L(\theta)$ where

$$\ln L(\theta) = n \ln(3\theta) + (3\theta - 1) \ln(x_1 x_2 \ldots x_n).$$

By taking $\frac{d \ln L(\theta)}{d\theta} = 0$, we get $\frac{3n}{3\theta} + 3 \ln(x_1 x_2 \ldots x_n) = 0$. Therefore

$$\hat{\theta}_{ML} = -\frac{n}{3 \ln(x_1 x_2 \ldots x_n)}.$$

□

3 (3 points)

Suppose that the distribution of heights of all male college students in Sweden is a normal distribution $N(\mu_{\text{male}}, \sigma_{\text{male}})$, and the distribution of heights of all female college students in Sweden is also a normal distribution $N(\mu_{\text{female}}, \sigma_{\text{female}})$. Assume that $N(\mu_{\text{male}}, \sigma_{\text{male}})$ and $N(\mu_{\text{female}}, \sigma_{\text{female}})$ are independent, and the variance $\sigma_{\text{male}} = \sigma_{\text{female}} = \sigma$ which is unknown. Now we choose two independent random samples from $N(\mu_{\text{male}}, \sigma_{\text{male}})$ and $N(\mu_{\text{female}}, \sigma_{\text{female}})$ respectively, which yield the following data (in cm).

Male:	179	186	182	178	185;
Female:	181	178	182	179.	

(3.1). (1p) Find a two-sided 95% confidence interval of μ_{male}.
(3.2). (2p) Is it reasonable to conclude that $\mu_{\text{male}} > 1.05\mu_{\text{female}}$? Answer the question by constructing a two-sided 95% confidence interval of the difference $\mu_{\text{male}} - 1.05\mu_{\text{female}}$.

Solution. (3.1). A two-sided 95% confidence interval of μ_{male} is $\bar{x}_{\text{male}} \pm t_{0,025}(7)\frac{s}{\sqrt{5}} = 182 \pm 2.36\frac{2.93}{\sqrt{5}} = (178.9, 185.1)$.
(3.2). We have $\bar{X} - 1.05\bar{Y} \sim N(\mu_{\text{male}} - 1.05\mu_{\text{female}}, \sigma\sqrt{\frac{1}{n_{\text{male}}} + \frac{1.05^2}{n_{\text{female}}}})$, which gives

$$\frac{(\bar{X} - 1.05\bar{Y}) - (\mu_{\text{male}} - 1.05\mu_{\text{female}})}{\sigma\sqrt{\frac{1}{n_{\text{male}}} + \frac{1.05^2}{n_{\text{female}}}}} \sim N(0,1).$$

But σ is unknown, so we need to replace σ by its point estimator S, thus

$$\frac{(\bar{X} - 1.05\bar{Y}) - (\mu_{\text{male}} - 1.05\mu_{\text{female}})}{S\sqrt{\frac{1}{n_{\text{male}}} + \frac{1.05^2}{n_{\text{female}}}}} \sim t(n_{\text{male}} + n_{\text{female}} - 2).$$

We construct a two-sided confidence interval for $\mu_{\text{male}} - 1.05\mu_{\text{female}}$.

$$\begin{aligned} I_{\mu_{\text{male}} - 1.05\mu_{\text{female}}} &= (\bar{x} - 1.05\bar{y}) \mp t_{\alpha/2}(n_{\text{male}} + n_{\text{female}} - 2) \cdot s \cdot \sqrt{\frac{1}{n_{\text{male}}} + \frac{1.05^2}{n_{\text{female}}}} \\ &= (182 - 1.05 \times 180) \mp t_{.025}(7) \cdot \sqrt{60/7} \cdot \sqrt{\frac{1}{5} + \frac{1.05^2}{4}} = (-11.77, -2.23) \end{aligned}$$

Since the confidence interval is negative, we say that it is not reasonable to conclude that $\mu_{\text{male}} > 1.05\mu_{\text{female}}$. But it is reasonable to say $\mu_{\text{male}} < 1.05\mu_{\text{female}}$.
Here $s^2 = \frac{4s^2_{\text{male}} + 3s^2_{\text{female}}}{7}$. □

4 (3 points)

Assume that X_1, X_2 and X_3 are independent standard normal random variables $N(0,1)$. The random vector $\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix}$ is defined as

$$Y_1 = X_1 + X_2, \qquad Y_2 = X_1 + X_3, \qquad Y_3 = X_2 + X_3.$$

(4.1). (2p) Determine the mean vector and the covariance matrix for $\mathbf{Y}$
(4.2). (1p) Find $P(2Y_1 > Y_2 + 2)$.

Solution. (4.1). It is known that

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}, \quad \mu_{\mathbf{X}} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad C_{\mathbf{X}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Thus for $\mathbf{Y} = A\mathbf{X}$ with

$$A = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix},$$

we have

$$\mu_{\mathbf{Y}} = A\mu_{\mathbf{Y}} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad C_{\mathbf{Y}} = AC_{\mathbf{Y}}A' = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

(4.2).

$$\begin{aligned} P(2Y_1 > Y_2 + 2) &= P(2X_1 + 2X_2 > X_1 + X_3 + 2) = P(X_1 + 2X_2 - X_3 > 2) \\ &= P(N(0, \sqrt{6}) > 2) = P(N(0,1) > 2/\sqrt{6}) = 1 - 0.7939 = 0.2061. \end{aligned}$$

□

5 (3 points)

The minimal daily demand on zinc of a male person over 30 years of age is 15 mg. A scientist conjectures that the expected value is lower and wants to conduct a study in order to show that. Assume that the scientist measures the zinc intake of 25 randomly selected male person over 30 years of age and uses these data in order to test the hypotheses

$$H_0 : \mu = 15 \qquad H_a : \mu < 15.$$

Assume that the observations are independent and from a population $N(\mu, \sigma)$. The sample mean is $\bar{x} = 13$ and the sample standard deviation is $s = 6$.
(5.1). (1p) If σ is unknown, do you reject H_0 given a significance level $\alpha = 0.01$? and why ?
(5.2). (1p) If σ is known $\sigma = 4$, do you reject H_0 given a significance level $\alpha = 0.01$? and why ?
(5.3). (1p) If σ is known $\sigma = 4$, based on (5.2), what is the probability of not concluding that $\mu < 15$ when the actual $\mu = 12$?

Solution. (5.1) Since σ is unknown, the test statistic is $TS = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{13-15}{6/\sqrt{25}} = -1.67$. The rejection region is

$$C = (-\infty,\ -t_\alpha(n-1)) = (-\infty,\ -t_{0.01}(25-1)) = (-\infty,\ -2.49).$$

Because $TS \notin C$, we do NOT reject H_0.
(5.2) Since σ is known $\sigma = 4$, the test statistic is $TS = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} = \frac{13-15}{4/\sqrt{25}} = -2.5$. The rejection region is

$$C = (-\infty,\ -\lambda_\alpha) = (-\infty,\ -\lambda_{0.01}) = (-\infty,\ -2.33).$$

Because Because $TS \in C$, we reject H_0.
(5.3) This is a Type II error, namely

$$\begin{aligned}
\beta(12) &= P(\text{don't reject } H_0 \text{ when } H_0 \text{ is false if } \mu = 12) \\
&= P(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > -2.33 \text{ if } \mu = 12) \\
&(\text{need to change } \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \text{ to } \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \text{ since } \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)) \\
&= P(\bar{X} > -2.33\frac{\sigma}{\sqrt{n}} + \mu_0 \text{ if } \mu = 12) \\
&= P(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} > \frac{-2.33\frac{\sigma}{\sqrt{n}} + \mu_0 - \mu}{\sigma/\sqrt{n}} \text{ if } \mu = 12) \\
&= P(N(0,1) > -2.33 + \frac{15-12}{4/\sqrt{25}}) \\
&= P(N(0,1) > 1.42) = 1 - 0.9222 = 0.0778.
\end{aligned}$$

□

6 (3 points)

The following table shows the expenses for private comsumption (y) and the disposable income (x_1) both expressed in billions of USD. The variable x_2 denotes the war state

$$x_2 = \begin{cases} 1, & \text{when the country is at war} \\ 0, & \text{otherswise} \end{cases}$$

This data is for USA during the years 1935 - 1949.

x_1	58.5	66.3	71.2	65.5	70.3	75.7	92.7	99.6	116.9	133.5	146.3	150.2	160.0	169.8	189.2
x_2	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0
y	56	62	67	64	67	71	81	89	94	99	108	120	144	162	175

An analysis of the data according to the model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, where $\varepsilon \sim N(0, \sigma)$.

Estimated regression line: $y = 1.00 + 0.92x_1 - 23.34x_2$.

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	1.0016	2.3855
1	0.9241	0.0196
2	-23.3432	2.0608

	Degrees of freedom	Sum of squares
REGR	2	25868.1
RES	12	139.7
TOT	14	26007.8

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0.48891 & -0.00363 & 0.00673 \\ -0.00363 & 0.00003 & -0.00089 \\ 0.00673 & -0.00089 & 0.36486 \end{pmatrix}.$$

(6.1). (1p) Is the private consumption affected by the war state according to this analysis? Explain your answer. Use significance level 1%.
(6.2). (2p) Construct a 99% prediction interval for the private consumption a year when $x_1 = 100$ and the country is not at war.

Solution. (6.1). Method 1: By Hypothesis testing:

$$H_0 : \beta_2 = 0, \qquad H_1 : \beta_2 \neq 0.$$

Then we get

$$TS = \frac{\hat{\beta}_2 - 0}{d(\hat{\beta}_2)} = \frac{-23.3432}{2.0608} = -11.29,$$

$C = (-\infty, -t_{\alpha/2}(n-k-1)) \cup (t_{\alpha/2}(n-k-1), +\infty) = (-\infty, -t_{0.005}(12)) \cup (t_{0.005}(12), +\infty) = (-\infty, -3.05) \cup (3.05, +\infty)$

Since $TS \in C$, reject H_0. So we think $\beta_2 \neq 0$, namely, the private consumption is affected by the war state x_2.
Method 2: We can construct a confidence interval for β_2 as follows

$$\begin{aligned} I_{\beta_2} &= \hat{\beta}_2 \mp t_{\alpha/2}(n-k-1) \cdot s \cdot \sqrt{h_{22}} = \hat{\beta}_2 \mp t_{\alpha/2}(n-k-1) \cdot d(\hat{\beta}_2) \\ &= -23.3432 \mp 3.05 \cdot 2.0608 = (-29.63, -17.06). \end{aligned}$$

Since $0 \notin I_{\beta_2}$, we think $\beta_2 \neq 0$. Namely, tthe private consumption is affected by the war state x_2.
(6.2). A prediction interval for Y as follows

$$I_Y = \hat{\mu} \mp t_{\alpha/2}(n-k-1) \cdot s \cdot \sqrt{1 + \mathbf{x}' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}},$$

where

$$\begin{aligned} &\hat{\mu} = 1.00 + 0.92 \times 100 - 23.34 \times 0; \\ &t_{\alpha/2}(n-k-1) = t_{0.005}(12) = 3.05; \\ &s^2 = \frac{SS_E}{n-k-1} = \frac{139.7}{15-2-1}; \\ &\mathbf{x}' = (1, \quad 100, \quad 0). \end{aligned}$$

Thus a 99% prediction interval for the private consumption $I_Y = (82.27, \quad 103.73)$. □

Here if you use $\hat{\mu} = 1.0016 + 0.9241 \times 100 - 23.3432 \times 0$, then a 99% prediction interval for the private consumption $I_Y = (82.68, \quad 104.14)$.

Svensk Version

1 (3 poäng)

Som ett mått på en sjös hälsotillstånd använder man ibland antalet cellklumparper volymsenhet. För hundra olika vattenprover tagna på slumpmässiga platser i en sjö har man räknat antalet cellklumpar $x_1, \ldots, x_{100}$. Resultat:

Antal klumpar i	0	1	2	3	4	5	6	7	≥ 8
Frekvens N_i	7	15	18	26	20	6	5	3	0

Pröva med ett χ^2-test på nivån $\alpha = 0.05$ hypotesen

$$H_0 : X \sim Po(\mu).$$

2 (3 poäng)

Antag att fördelningen för en population har täthetsfunktionen

$$f(x) = \begin{cases} 3\theta x^{3\theta-1} & \text{om } 0 \leq x \leq 1; \\ 0 & \text{annars,} \end{cases}$$

där $\theta > 0$ är en okänd parameter. $\{x_1, x_2, \ldots, x_n\}$ är ett stickprov från populationen.
(2.1). (1.5p) Hitta en punktskattning $\hat{\theta}_{MM}$ av θ genom att använda momentmetoden.
(2.2). (1.5p) Hitta en punktskattning $\hat{\theta}_{ML}$ av θ genom att använda Maximum-Likelihood-metoden.

3 (3 poäng)

Antag att längderna på manliga studenter i Sverige är normalfördelade $N(\mu_{\text{male}}, \sigma_{\text{male}})$, och att längden på kvinnliga studenter i Svergie följer en normalfördelning $N(\mu_{\text{female}}, \sigma_{\text{female}})$. Antag att $N(\mu_{\text{male}}, \sigma_{\text{male}})$ och $N(\mu_{\text{female}}, \sigma_{\text{female}})$ är oberoende, och variansen $\sigma_{\text{male}} = \sigma_{\text{female}} = \sigma$ är okänd. Vi tar nu två oberoende stickprov från $N(\mu_{\text{male}}, \sigma_{\text{male}})$, respektive $N(\mu_{\text{female}}, \sigma_{\text{female}})$, och får följdande data. (i cm).

Män:	179	186	182	178	185;
Kvinnor:	181	178	182	179.	

(3.1). (1p) Bilda ett tvåsidigt 95% konfidensintervall för μ_{male}.
(3.2). (2p) Förefaller det troligt att $\mu_{\text{male}} > 1.05\mu_{\text{female}}$? Besvara frågan genom att konstruera ett tvåsidigt 95% konfidensintervall för differensen $\mu_{\text{male}} - 1.05\mu_{\text{female}}$.

4 (3 poäng)

Antag att X_1, X_2 och X_3 är oberoende standard normalfördelad $N(0,1)$. Den stokastiska variabeln $\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix}$ är

$$Y_1 = X_1 + X_2, \qquad Y_2 = X_1 + X_3, \qquad Y_3 = X_2 + X_3.$$

(4.1). (2p) Bestäm väntevärdesmatris och kovariansmatris för $\mathbf{Y}$
(4.2). (1p) Beräkna $P(2Y_1 > Y_2 + 2)$.

5 (3 poäng)

Minsta dagliga behov av zink är 15 mg för män över 30 år. I själva verket misstänker man att det förväntade värdet är lägre och man will genomföra en studie för att påvisa detta. Antag att man mäter zinkintaget för 25 slumpmässigt utvalda män över 30 år och använder data för att testa hypoteserna

$$H_0 : \mu = 15 \qquad H_a : \mu < 15.$$

Antag att observationerna är oberoende och från en population $N(\mu, \sigma)$. Stickprovsmedelvärdet är $\bar{x} = 13$ och stickprovsstandardavvikelsen är $s = 6$.
(5.1). (1p) Om σ är okänd, förkastar du H_0 givet en signifikansnivån $\alpha = 0.01$? Varför ?
(5.2). (1p) Om σ är känd $\sigma = 4$, förkastar du H_0 givet en signifikansnivån $\alpha = 0.01$? Varför ?
(5.3). (1p) Om σ är känd $\sigma = 4$, baserat på (5.2), vad är sannolikheten att inte dra slutsatsen att $\mu < 15$ men $\mu = 12$?

6 (3 poäng)

Följande tabell visar utgifterna för privat konsumtion (y) samt den disponibla inkomsten (x_1) båda uttryckta i miljarder dollar. Variabeln x_2 anger krigstillståndet

$$x_2 = \begin{cases} 1, & \text{då landet är i krig} \\ 0, & \text{annars} \end{cases}$$

Uppgifter gäller USA under åren 1935 - 1949.

x_1	58.5	66.3	71.2	65.5	70.3	75.7	92.7	99.6	116.9	133.5	146.3	150.2	160.0	169.8	189.2
x_2	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0
y	56	62	67	64	67	71	81	89	94	99	108	120	144	162	175

Analys av datamaterialet enligt modellen $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$, där $\varepsilon \sim N(0, \sigma)$.

Estimated regression line: $y = 1.00 + 0.92x_1 - 23.34x_2$.

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	1.0016	2.3855
1	0.9241	0.0196
2	-23.3432	2.0608

	Degrees of freedom	Sum of squares
REGR	2	25868.1
RES	12	139.7
TOT	14	26007.8

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0.48891 & -0.00363 & 0.00673 \\ -0.00363 & 0.00003 & -0.00089 \\ 0.00673 & -0.00089 & 0.36486 \end{pmatrix}.$$

(6.1). (1p) Är den privata konsumtionen påverkas av landets krigstillstånd enlight den här analysen ? Förklara ditt svar. Använd signifikansnivå 1%.
(6.2). (2p) Konstruera ett 99% prediktionsintervall för den privata konsumtionen ett år, då $x_1 = 100$ och då landet inte är i krig.