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 (2.5 points)

Julia's doctors suspect that she was suffering from hypokalaemia, i.e., low levels of potassium in the blood. Repeated measurements of the potassium value of a person gives different results, partly because of individual variations from day to day, partly due to measurement error. It has been found that it is reasonable to assume that a measured potassium value of a person is normally distributed with parameters μ and σ, where μ is the characteristic potassium value of the person and $\sigma = 0.2$. A person classed as potassium hypokalaemic if the value is below 3.5. Assume that Julia has $\mu = 3.7$.

(1.1). (1p) What is the probability that Julia is classified as hypokalaemic if you make a single potassium measurement?
(1.2). (1.5p) What is the probability that Julia is classified as hypokalaemic if one makes four independent measurements at appropriate time intervals and the mean of these measurements are compared with 3.5?

Solution. (1.1). Let X be the measured potassium value of Julia. Then $X \sim N(3.7, 0.2)$, Then $P(X < 3.5) \approx 0.16$.
(1.2). Let $\overline{X}$ be the mean of four measured potassium values of Julia. Then $\overline{X} \sim N(3.7, \frac{0.2}{\sqrt{4}})$, Then $P(\overline{X} < 3.5) \approx 0.02$. $\square$

2 (3 points)

Suppose that the distribution of a population X has the probability mass function as follows

X	0	1
$p(x)$	$1-p$	p

where p is unknown. We have a sample $\{x_1, \ldots, x_n\}$ from this distribution.
(2.1). (1.5p) Find a point estimate $\hat{p}_{MM}$ of p using Method of Moments. Is it unbiased?
(2.2). (1.5p) Find a point estimate $\hat{p}_{ML}$ of p using Maximum-Likelihood method (Hint: $P(X = x) = p^x(1-p)^{1-x}$). Is it unbiased?

Solution. (2.1). For Method of Moments, the first equation is $E(X) = \bar{x}$. The mean $E(X)$ can be calculated as

$$E(X) = 0 \times (1-p) + 1 \times p = p.$$

By solving $E(X) = \bar{x}$, we have $p = \bar{x}$ which yields $\hat{p}_{MM} = \bar{x}$. The corresponding point estimator is $\hat{P}_{MM} = \overline{X}$. This is unbiased since it can be easily checked that $E(\hat{P}_{MM}) = E(\overline{X}) = E(X) = p$.
(2.2). For the Maximum-Likelihood method, we write the likelihood function as

$$L(p) = f(x_1) \cdot f(x_2) \ldots f(x_n) = p^{\sum x_i}(1-p)^{\sum(1-x_i)}.$$

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

$$\ln L(p) = \sum x_i \cdot \ln p + \sum (1-x_i) \cdot \ln(1-p).$$

By $\frac{d \ln L(p)}{dp} = 0$, we have $\frac{\sum x_i}{p} - \frac{\sum(1-x_i)}{1-p} = 0$, therefore $\hat{p}_{ML} = \frac{\sum x_i}{n} = \bar{x}$. (The second derivative $\frac{d^2 \ln L(p)}{dp^2} < 0$ which yields that $\hat{p}_{ML}$ is indeed a maximal point). The same reasoning as in (1.1) implies that it is unbiased. $\square$

3 (2.5 points)

The transmission of a digital image with a certain system takes an average of 3.45 seconds. By compressing the data (which need not lead to a worse picture of the recipient) one can cut down transmission time. A new algorithm that compresses the information, gives transit times that are $N(\mu, \sigma)$, where $\sigma = 0.32$ seconds. Fifteen independent image transfers gave the average transfer time $\bar{x} = 2.42$ seconds.
(3.1). (1p) Test at level 0.05 $H_0 : \mu = 2.5$ against $H_1 : \mu < 2.5$.
(3.2). (1.5p) Calculate the power of the test in (3.1) if $\mu = 2.40$.

Solution. (3.1) Since $\sigma = 0.32$, we have $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0, 1)$ then the test statistic is $TS = \frac{\bar{x}-2.5}{0.32/\sqrt{15}} = -0.97$. The rejection region is

$$C = (-\infty,\ -\lambda_{0.05}) = (-\infty,\ -1.645).$$

Because $TS \notin C$, we do NOT reject H_0.
(3.2)

$$\begin{aligned} h(2.4) &= P(\text{reject } H_0 \text{ when } H_0 \text{ is false if } \mu = 2.40) \\ &= P(\frac{\bar{X} - 2.5}{\sigma/\sqrt{n}} < -1.645 \text{ if } \mu = 2.40) \\ &= P(N(0, 1) < -0.43) = 0.3336. \end{aligned}$$

□

4 (3.5 points)

Assume that X_1 and X_2 are independent, where $X_1 \sim N(0, 1)$ and $X_2 \sim N(0, 2)$. Let $\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$, where

$$Y_1 = X_1 + X_2, \qquad Y_2 = X_1 - X_2.$$

(4.1). (1p) Show that $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ is a normal vector.
(4.2). (1.5p) Determine the mean vector and the covariance matrix for $\mathbf{Y}$.
(4.3). (1p) Find $P(Y_1 > Y_2 + 2)$.

Solution. (4.1). Let $X_2' = 1/2X_2$, then $X_2' \sim N(0, 1)$, so $\begin{pmatrix} X_1 \\ X_2' \end{pmatrix}$ is a standard normal vector. Since $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2' \end{pmatrix}$, so $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ is a normal vector.
(4.2). For $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$, we have $\mu_{\mathbf{X}} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and $C_{\mathbf{X}} = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}$. So,

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix},$$

Thus we get the mean and covariance matrix for $\mathbf{Y}$

$$\mu_{\mathbf{Y}} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad C_{\mathbf{Y}} = AC_{\mathbf{X}}A^T = \begin{pmatrix} 5 & -3 \\ -3 & 5 \end{pmatrix}.$$

(4.3). $P(Y_1 > Y_2 + 2) = P(X_2 > 1) = P(N(0, 1) > 1/2) = 0.3085$.

□

Indicator	Mean	Sample standard deviation	Number of measurements
Methyl red	$\bar{x} = 0.08686$	$s_x = 0.00098$	16
Bromocresol green	$\bar{y} = 0.08641$	$s_y = 0.00113$	26

5 (2 points)

One has made repeated measurements of the concentration of HCl in the solution by titration. Two different color indicators have been utilized to find the end point of the titration. Results:
Model: We have two independent random samples from $N(\mu_1, \sigma)$ and $N(\mu_2, \sigma)$, respectively.
Does it seem that both indicators give equivalent results? Justify your answer using a suitable confidence interval or test with $\alpha = 0.05$.

Solution. $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$ Note that

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

$$TS = \frac{(\bar{x} - \bar{y}) - (0)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = 1.90, \qquad C = (-\infty, -t_{0.025}(40)) \cup (t_{0.025}(40), \infty) = (-\infty, -2.02) \cup (2.02, \infty)$$

where

$$s^2 = \frac{(16-1)s_1^2 + (26-1)s_2^2}{16+26-2}.$$

$TS \notin C$, don't reject H_0, that is, it seems that both indicators could give equivalent results.

□

6 (4.5 points)

At a wastewater treatment plant in the laboratory conducted a series of experiments to determine the phosphate reduction is y in percent because of the waste water pH-value x. Results:

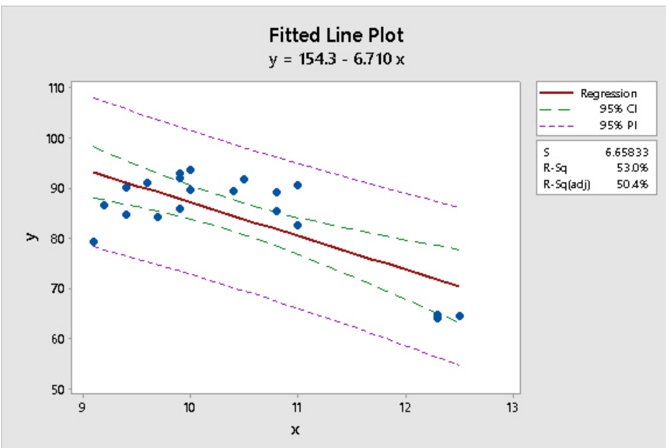
Row	x	y
1	9.2	86.5
2	9.9	93.0
3	11.0	90.5
4	10.4	89.5
5	10.8	89.2
6	12.5	64.5
7	12.3	64.0
8	12.3	64.6
9	10.5	91.7
10	9.4	90.2
11	9.6	91.0
12	9.4	84.6
13	10.0	89.7
14	10.8	85.3
15	11.0	82.6
16	9.9	85.8
17	9.1	79.4
18	9.7	84.2
19	9.9	91.9
20	10.0	93.6

The data were analyzed according to two different models

$$\begin{aligned} \text{Model 1}: \quad Y &= \tilde{\beta}_0 + \tilde{\beta}_1 x + \tilde{\epsilon}, \\ \text{Model 2}: \quad Y &= \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon, \end{aligned}$$

where ϵ-variables are assumed to be independent and $N(0, \sigma)$ distributed.
Outputs and plots are given below.
(6.1). (0.5p) Explain briefly on the basis of plots why model 2 describes the data better than the model 1.
(6.2). (1.5p) How does it appear from the analysis that term x^2 is essential to the model 2. Motivate your answer with help of appropriate 95% confidence interval.
(6.3). (1p) Which value of pH is optimal according to the model 2. Motivate your answer using the appropriate calculations.
(6.4). (1.5p) Find a 99% prediction interval for pH value $x = 11.1$ in model 2.

```
MODEL 1...............................
```

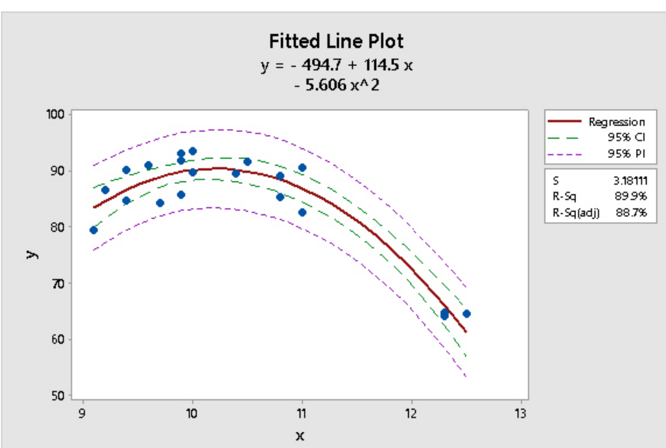



```
Analysis of variance

Estimated regression equation: y = 154.3 - 6.710 x
```

	Degrees of freedom	Sum of squares
REGR	1	900.68
RES	18	798.00
TOT	19	1698.68

```
MODEL 2...............................
```



```
Analysis of Variance

Estimated regression equation: y = - 494.7 + 114.5 x - 5.606 x^2

           Degrees of freedom        Sum of squares
REGR               2                     1526.65
RES                17                    172.03
TOT                19                    1698.68
```

$$(X^T X)^{-1} = \begin{pmatrix} 678.164 & -126.211 & 5.8116 \\ -126.211 & 23.5343 & -1.08585 \\ 5.8116 & -1.08585 & 0.050207 \end{pmatrix}$$

Solution. (6.1). Model 2 fits xy-plot better.
(6.2). $I_{\beta_2} = \hat{\beta}_2 \mp t_{0.025}(17) \cdot \sqrt{\frac{172.03}{17}} \cdot \sqrt{0.050207} \approx (-7.11, -4.10)$, where $t_{0.025}(17) = 2.11$. Since $0 \notin I_{\beta_2}$, we believe that β_2 is not 0. So the variable x^2 is essential to the model 2.
(6.3). $y = -494.7 + 114.5x - 5.606x^2$, then $\frac{dy}{dx} = 114.5 - 2 \cdot 5.606x = 0$, so $x = 10.21$. That is when $pH = 10.21$, we get optimal according to Model 2.
(6.4). P.I.

$$I_Y = \hat{\mu} \mp t_{0.005}(17) \cdot s \cdot \sqrt{1 + \mathbf{x}^T \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{x}} = (75.73, 95.34).$$

Where $\hat{\mu} = -494.7 + 114.5(11.1) - 5.606(11.1)^2$, $t_{0.005}(17) = 2.90$, $s = \sqrt{\frac{172.03}{17}}$, $\mathbf{x} = (1 \quad 11.1 \quad 11.1^2)$ and $\sqrt{1 + \mathbf{x}^T \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{x}} = \sqrt{1.1287430487}$.

□

1 (2.5 poäng)

Julias läkare misstänker att hon lider av hypokaliemi, d.v.s. låg kaliumhalt i blodet. Upprepade mätningar av kaliumvärdet för en person ger olika resultat, dels beroende på individuella variationer från dag till dag, dels beroende på mätfel- Man har funnit att det är rimligt att anta att ett uppmätt kaliumvärde för en person är normalfördelat med parametrar μ och σ, där μ är det karakteriska kaliumvärdet för personen och $\sigma = 0.2$. En person klassas som hypokaliemisk om kaliumvärdet är under 3.5. Anta att Julia har $\mu = 3.7$.

(1.1). (1p) Vad är sannolikheten at Anna klassas som hypokaliemisk om man gör en enda kaliummätning?
(1.2). (1.5p) Vad är sannolikheten att Julia klassas som hypokaliemisk om man gör fyra oberoende mätningar med lämpliga tidsavstånd och medelvärdet av dessa mätningar jämförs med 3.5?

2 (3 poäng)

Antag att fördelningen för en population X har sannolikhetsfunktionen enligt följande

X	0	1
$p(x)$	$1-p$	p

där p är okänd. Vi har ett stickprov $\{x_1, \ldots, x_n\}$ från denna fördelning.
(2.1). (1.5p) Hitta en punktskattning $\hat{p}_{MM}$ av p genom att använda momentmetoden. Är det väntevärdesriktigt?
(2.2). (1.5p) Hitta en punktskattning $\hat{p}_{ML}$ av p genom att använda Maximum Likelihood-metoden (Ledning: $P(X = x) = p^x(1-p)^{1-x}$). Är det väntevärdesriktigt?

3 (2.5 poäng)

Överföring av en digital bild med ett visst system tar i genomsnitt 3.45 sekunder. Genom att komprimera informationen (vilket inte behöver leda till en sämre bild hos mottagaren) kan man skära ned överföringstiden. En ny algoritm, som komprimerar informationen, ger överföringstider som är $N(\mu, \sigma)$, där $\sigma = 0.32$ sekunder. Femton oberoende bildöverföringar gav den genomsnittliga överföringstiden $\bar{x} = 2.42$ sekunder.
(3.1) (1p) Pröva på nivå 0.05 $H_0 : \mu = 2.5$ mot $H_1 : \mu < 2.5$.
(3.2) (1.5p) Beräkna styrkan för testet i (3.1), då $\mu = 2.40$.

4 (3.5 poäng)

Antag att X_1 och X_2 är oberoende, där $X_1 \sim N(0, 1)$ och $X_2 \sim N(0, 2)$. Den stokastiska variabeln $\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$, där

$$Y_1 = X_1 + X_2, \qquad Y_2 = X_1 - X_2.$$

(4.1). (1p) Visa att $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ är en normalvektor.
(4.2). (1.5p) Bestäm väntevärdesmatrix och kovariansmatrix för $\mathbf{Y}$.
(4.3). (1p) Beräkna $P(Y_1 > Y_2 + 2)$.

5 (2 poäng)

Man har gjort upprepade mätningar av koncentrationen av HCl i en lösning med hjälp av titrering. Två olika färgindikatorer har utnyttjats för att hitta slutpunkten i titreringen. Resultat:

Indikator	Medelvärde	Stickprovsstand.avv.	Antal mätningar
Metylrött	$\bar{x} = 0.08686$	$s_x = 0.00098$	16
Bromkresol grönt	$\bar{y} = 0.08641$	$s_y = 0.00113$	26

Modell: Vi har två oberoende slumpmässiga stickprov från $N(\mu_1, \sigma)$ respektive $N(\mu_2, \sigma)$.
Verkar de båda indikatorerna ge likvärda resultat? Motivera ditt svar med hjälp av ett lämpligt konfidensintervall eller test på nivå $\alpha = 0.05$.

6 (4.5 poäng)

Vid ett avloppsreningsverk i laboratorieskala genomfördes en serie experiment för att fastställa hur fosfatreduktionen y mätt i procent beror av avloppsvattnets pH-värde x. Resultat:

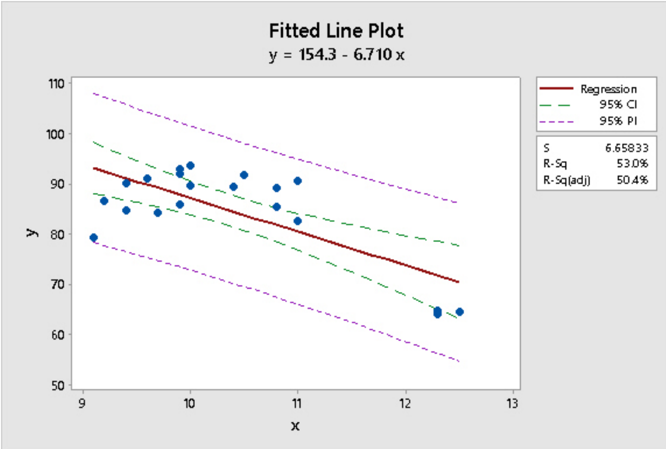
Row	x	y
1	9.2	86.5
2	9.9	93.0
3	11.0	90.5
4	10.4	89.5
5	10.8	89.2
6	12.5	64.5
7	12.3	64.0
8	12.3	64.6
9	10.5	91.7
10	9.4	90.2
11	9.6	91.0
12	9.4	84.6
13	10.0	89.7
14	10.8	85.3
15	11.0	82.6
16	9.9	85.8
17	9.1	79.4
18	9.7	84.2
19	9.9	91.9
20	10.0	93.6

Datamaterialet har analyserats enligt två olika modeller

$$\begin{aligned} \text{Model 1}: \quad Y = & \quad \tilde{\beta}_0 + \tilde{\beta}_1 x + \tilde{\epsilon}, \\ \text{Model 2}: \quad Y = & \quad \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon, \end{aligned}$$

där ϵ-variablerna antas vara oberoende $N(0, \sigma)$.
Utskrift och plottar finns på nästa sida.
(6.1). (0.5p) Förklara kortfattat utifrån plotterna varför modell 2 beskriver datamaterialet bättre än modell 1.
(6.2). (1.5p) Hur framgår det av analysen att x^2-termen är väsentlig i modell 2. Motivera ditt svar med hjälp av ett lämpligt 95% konfidensintervall.
(6.3). (1p) Vilket pH-värde är optimalt enligt analysen för modell 2. Motivera ditt svar med hjälp av lämpliga beräkningar.
(6.4). (1.5p) Konstruera ett 99% prediktionsintervall för pH-värde $x = 11.1$ för modell 2.

MODEL 1..............................

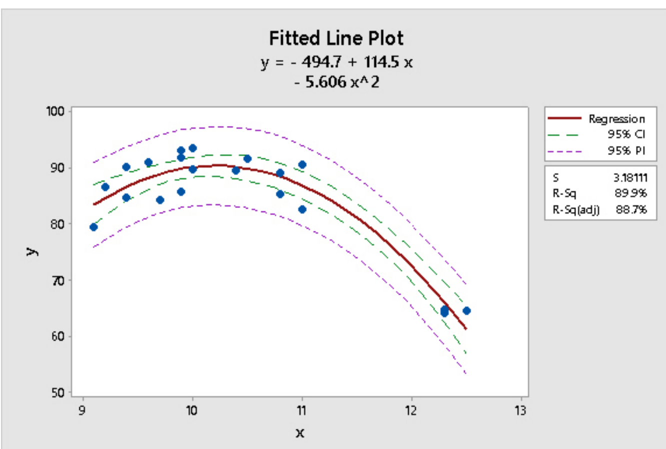


Analysis of variance

Estimated regression equation: y = 154.3 - 6.710 x

	Degrees of freedom	Sum of squares
REGR	1	900.68
RES	18	798.00
TOT	19	1698.68

MODEL 2..............................



Analysis of Variance

Estimated regression equation: y = - 494.7 + 114.5 x - 5.606 x^2

	Degrees of freedom	Sum of squares
REGR	2	1526.65
RES	17	172.03
TOT	19	1698.68

$$(X^T X)^{-1} = \begin{pmatrix} 678.164 & -126.211 & 5.8116 \\ -126.211 & 23.5343 & -1.08585 \\ 5.8116 & -1.08585 & 0.050207 \end{pmatrix}$$