Kurskod: TAMS24 (Statistisk teori)/Provkod: TEN1 2017-10-17 08:00-12:00

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 (4 points)

The surface roughness was determined for four different materials used for encapsulation. Results are:

Material	surface roughness						$\bar{x}_i$	s_i
Type 1	0.50	0.55	0.55	0.36			0.4900	0.0898
Type 2	0.31	0.07	0.25	0.18	0.56	0.20	0.2617	0.1665
Type 3	0.20	0.28	0.12				0.2000	0.0800
Type 4	0.10	0.16					0.1300	0.0424

Assume that these four samples are from independent normal distributions $N(\mu_i, \sigma), i = 1, 2, 3, 4$.
(1.1). (1p) Construct a confidence interval for σ^2 with confidence coefficient 0.95.
(1.2). (1p) Construct a confidence interval for μ_1 with confidence coefficient 0.95.
(1.3). (2p) Is it reasonable to conclude that $\mu_1 > 1.2\mu_2$? Answer the question by constructing an appropriate confidence interval with confidence coefficient 0.95.

Solution. (1.1).

$$s^2 = \frac{(n_1-1)s_1^2 + \ldots + (n_4-1)s_4^2}{n_1 + \ldots + n_4 - 4} = 0.016.$$

$$I_{\sigma^2} = \left(\frac{(n_1+\ldots+n_4-4)s^2}{\chi^2_{\alpha/2}(n_1+\ldots+n_4-4)}, \quad \frac{(n_1+\ldots+n_4-4)s^2}{\chi^2_{1-\alpha/2}(n_1+\ldots+n_4-4)} \right) = \left(\frac{11\cdot 0.016}{21.93}, \quad \frac{11\cdot 0.016}{3.81} \right) = (0.008, 0.046).$$

(1.2). Now we have

$$\frac{\bar{X} - \mu_1}{S/\sqrt{n_1}} \sim t(n_1 + \ldots + n_4 - 4)$$

Then we have CI for $\mu_1 : I_{\mu_1} = \bar{x} \mp t_{\alpha/2}(n_1+\ldots+n_4-4) \cdot \frac{s}{\sqrt{n_1}} = 0.4900 \mp t_{0.025}(11) \cdot \frac{\sqrt{0.016}}{\sqrt{4}} = (0.35, 0.63)$, where $t_{0.025}(11) = 2.20$.
(1.3). Note that $\bar{X} \sim N(\mu_1, \sigma/\sqrt{n_1})$ and $\bar{Y} \sim N(\mu_2, \sigma/\sqrt{n_2})$, then $\bar{X} - 1.2\bar{Y} \sim N(\mu_1 - 1.2\mu_2, \sigma\sqrt{1/n_1 + 1.2^2/n_2}$. Then we get

$$\frac{(\bar{X} - 1.2\bar{Y}) - (\mu_1 - 1.2\mu_2)}{S\sqrt{\frac{1}{n_1} + \frac{1.2^2}{n_2}}} \sim t(n_1 + \ldots + n_4 - 4)$$

$$I_{\mu_1 - 1.2\mu_2} = (a, \infty) = ((\bar{x} - 1.2\bar{y}) - t_\alpha(n_1+\ldots+n_4-4) \cdot s \cdot \sqrt{\frac{1}{n_1} + \frac{1.2^2}{n_2}}, \quad \infty) \approx (0.02, \infty)$$

where $t_{0.05}(11) = 1.8$. So $\mu_1 - 1.2\mu_2 > 0$, that is, we can conclude that $\mu_1 > 1.2\mu_2$ with 95% confidence. □

2 (2.5 points)

Ten observations $\{4.1, 4.4, 3.9, 3.8, 4.0, 4.3, 4.0, 3.9, 4.4, 4.2\}$ are from $N(\mu, 0.2)$.
(2.1) (1p) Test the hypothesis $H_0 : \mu = 4$ versus $H_1 : \mu > 4$ with a significance level 5%.
(2.2) (1.5p) Calculate the power for $\mu = 3.9$.

Solution. (2.1) Since σ is known, we have $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$,
then the test statistic is

$$TS = \frac{\bar{x} - 4}{\sigma/\sqrt{n}} = \frac{4.1 - 4}{0.2/\sqrt{10}} \approx 1.58.$$

The rejection region is

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

Because $TS \notin C$, we do NOT reject H_0.
(2.2)

$$\begin{aligned} h(3.9) &= P(\text{reject } H_0 \text{ when } H_0 \text{ is false if } \mu = 3.9) \\ &= P(\frac{\bar{X} - 4}{\sigma/\sqrt{n}} > 1.645 \text{ if } \mu = 3.9) \\ &\text{(need to change } \frac{\bar{X} - 4}{\sigma/\sqrt{n}} \text{ to } \frac{\bar{X} - 3.9}{\sigma/\sqrt{n}} \text{ since } \frac{\bar{X} - 3.9}{\sigma/\sqrt{n}} \sim N(0,1)) \\ &= P(\bar{X} > 4 + 1.645\frac{\sigma}{\sqrt{n}}, \text{ if } \mu = 3.9) \\ &= P(\frac{\bar{X} - 3.9}{\sigma/\sqrt{n}} > 1.645 + \frac{4 - 3.9}{\sigma/\sqrt{n}} \text{ if } \mu = 3.9) \\ &= P(N(0,1) > 3.23) = 1 - 0.9994 = 0.0006. \end{aligned}$$

□

3 (2.5 points)

The normal random vector $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$ has a mean vector and a covariance matrix

$$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \text{ respective } \begin{pmatrix} 7 & 1 & -2 \\ 1 & 1 & 0 \\ -2 & 0 & 1 \end{pmatrix}.$$

(3.1). (0.5p) Are X_1 and X_2 independent? Why?
(3.2). (0.5p) Are X_3 and X_2 independent? Why?
(3.3). (1.5p) Determine the mean vector and the covariance matrix for $\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix}$, where

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

Solution. (3.1). X_1 and X_2 are Not independent, since $Cov(X_1, X_2) = 1 \neq 0$.
(3.2). X_3 and X_2 are independent since $Cov(X_3, X_2) = 0$.
(3.3). We can write $\mathbf{Y}$ as

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = A \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}, \quad \text{where the matrix } A = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$

Thus the mean vector for $\mathbf{Y}$ is

$$\mu_{\mathbf{Y}} = A\mu_{\mathbf{X}} = A \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

The covariance matrix for $\mathbf{Y}$ is

$$C_{\mathbf{Y}} = AC_{\mathbf{X}}A^T = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 4 & 6 \\ 0 & 6 & 10 \end{pmatrix}.$$

□

4 (3 points)

A type of measuring error has a uniform distribution on $[0, \theta]$, the probability density function is $f_X(x) = \frac{1}{\theta}$, where $0 \leq x \leq \theta$ and $\theta > 0$. We have measuring errors $\{x_1, x_2, \ldots, x_n.\}$
(4.1). (0.5p) Find a point estimate $\hat{\theta}_{MM}$ of θ using method of moments.
(4.2). (0.5p) Is the point estimate $\hat{\theta}_{MM}$ unbiased? Why?
(4.3). (2p) Find a point estimate $\hat{\theta}_{ML}$ of θ using Maximum-Likelihood method.

Solution. (4.1) The first equation from the Method of Moments is $E(X) = \bar{x}$. Since $E(X) = \int_0^\theta \frac{x}{\theta} dx = \frac{\theta}{2}$, we have $\frac{\theta}{2} = \bar{x}$, thus $\hat{\theta}_{MM} = 2\bar{x}$.
(4.2) $E(\hat{\Theta}_{MM}) = E(2\bar{X}) = 2E(X) = 2\frac{\theta}{2} = \theta$, so $\hat{\theta}_{MM}$ is unbiased.
(4.3) The likelihood function is

$$L(\theta) = f(x_1)f(x_2)\ldots f(x_n) = \frac{1}{\theta^n}, \text{ where } 0 \leq x_i \leq \theta, \quad i = 1, 2, ...n$$

Then we can get $\ln L(\theta) = -n\ln(\theta)$, and note that $\frac{d \ln L(\theta)}{d\theta} = -\frac{n}{\theta} < 0$, so $\ln L(\theta)$ is decreasing as θ is increasing. Thus $L(\theta)$ is decreasing as θ is increasing. To maximiaze the $L(\theta)$, we only need to find minimum value of θ. We know that $0 \leq x_i \leq \theta$ for $i = 1, 2, ...n$, so we get $\theta_{min} = max\{x_1, x_2, ..., x_n\}$. Therefore $\hat{\theta}_{ML} = \theta_{min} = max\{x_1, x_2, ..., x_n\}$. □

5 (4 points)

A company has a warehouse where goods are transported by trucks. At 500 different randomly chosen intervals of length one hour the number of trucks arriving was observed. Result:

Number of forklifts	0	1	2	3	4	5	6	7	8
Frequency	52	151	130	102	45	12	5	1	2

Use a χ^2-test with a significance level 5% to determine if the number of trucks arriving during one hour is $Po(\mu)$.

Solution.

$$H_0 : X \sim Po(\mu)$$
$$H_1 : X \nsim Po(\mu)$$

There is an unknown parameter μ in the $Po(\mu)$, and we can estimate as following:

$$\hat{\mu} = \bar{x} = \frac{0 \times 52 + 1 \times 151 + ... + 8 \times 2}{500} = 2.02 \approx 2.0$$

Then we get the following information, note that $n = 500$.

Number of trucks	0	1	2	3	4	5	6	7	8
N_i	52	151	130	102	45	12	5	1	2
p_i	0.1353	0.2707	0.2707	0.1804	0.0902	0.0361	0.0120	0.0034	0.0009
np_i	67.65	135.35	135.35	90.2	45.1	18.05	6	1.7	0.45

We find theoretical probabilities by calculating, for example, $p_1 = P(Po(2.0) = 0) = 0.1353$ which you can calculate either by applying the probability mass function or by checking the table for Poisson distribution.
We can see that $\sum_{i=1}^{9} p_i < 1$, then we add one group and get the following:

Number of trucks	0	1	2	3	4	5	6	7	8	≥ 9
N_i	52	151	130	102	45	12	5	1	2	0
p_i	0.1353	0.2707	0.2707	0.1804	0.0902	0.0361	0.0120	0.0034	0.0009	0.0003
np_i	67.65	135.35	135.35	90.2	45.1	18.05	6	1.7	0.45	0.15

Since $np_8 < 5$, $np_9 < 5$ and $np_{10} < 5$, we need to combine these three groups to group 7. Thus we get new groups

Number of trucks	0	1	2	3	4	5	≥ 6
N_i	52	151	130	102	45	12	8
p_i	0.1353	0.2707	0.2707	0.1804	0.0902	0.0361	0.0166
np_i	67.65	135.35	135.35	90.2	45.1	18.05	8.3

Therefore the test statistic is

$$TS = \sum_{i=1}^{7} \frac{(N_i - np_i)^2}{np_i} \approx 9.22,$$

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 can assume it is $Po(\mu)$. □

6 (2 points)

In a study of the survival time for patients with prostate cancer, for each patient the treatment type (x_1), age in years (x_2) at the time of the start of the treatment, the halt (x_3) of a certain characteristic substance, AP, in the blood, the occurrence of of skeletal metastasis (x_4) and the survival time (y) from the start of the treatment, was observed. We have

$$x_1 = \begin{cases} 0 & \text{at placebo treatment (=no treatment)} \\ 1 & \text{at estrogen treatment.} \end{cases}$$

$$x_4 = \begin{cases} 0 & \text{for no skeletal metastasis} \\ 1 & \text{presence of skeletal metastasis.} \end{cases}$$

An analysis of variance for the data for the models

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon, \qquad \varepsilon \sim N(0, \sigma)$$

and

$$Y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \varepsilon, \qquad \varepsilon \sim N(0, \sigma)$$

Analysis of variance
Estimated regression line: $y = 104.1 + 10.0x_1 - 0.96x_2 - 0.03x_3 - 4.32x_4$

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	104.076	37.7264
1	10.0219	6.69623
2	-0.956706	0.506024
3	-0.0262767	0.00907681
4	-4.31777	7.18044

	Degrees of freedom	Sum of squares
REGR	4	6665.0
RES	45	24082.1
TOT	49	30747.1

Analysis of variance
Estimated regression line: $y = 32.9 + 9.17x_1 - 0.022x_3$

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	32.8860	4.97944
1	9.17159	6.65964
2	-0.0220795	0.00858002

	Degrees of freedom	Sum of squares
REGR	2	4692.56
RES	47	26054.6
TOT	49	30747.1

(6.1). (1p) Choose appropriate model. Does the skeletal metastasis seem to affect the survival time? Justify your answer by constructing a suitable confidence interval or test with confidence level 90%.
(6.2). (1p) Is it useful to have x_2 or x_4? Perform a suitable test on 5% significance level.

Solution. (6.1). We have $\frac{\hat{B}_4-\beta_4}{d(\hat{\beta}_4)} \sim t(n-k-1)$, thus we have
90% Confidence interval of β_4 is:
$I_{\beta_4} = \hat{\beta}_4 \mp t_{\alpha/2}(n-k-1)\cdot d(\hat{\beta}_4) = -4.31777 \mp t_{0.05}(45)(7.18044) = -4.31777 \mp (1.68)(7.18044) \approx (-16.38, 7.75)$. Since $0 \in I_{\beta_4}$, we believe that β_4 might be 0. So the variable x_4 (the skeletal metastasis) doesn't affect the survival time.
Or we test $H_0 : \beta_4 = 0$ vs $H_1 : \beta_4 \neq 0$ has

$$\begin{cases} \text{TS} = \frac{\hat{\beta}_4}{d(\hat{\beta}_4)} \approx -0.6 \\ C = (-\infty, -t_{\alpha/2}(n-k-1)) \ \cup \ (t_{\alpha/2}(n-k-1), +\infty) = (-\infty, -1.68) \cup (1.68, \infty). \end{cases}$$

Since $TS \notin C$, we don't reject H_0, namely, So the variable x_4 (the skeletal metastasis) doesn't affect the survival time.
(6.2). We want to test $H_0 : \beta_2 = \beta_4 = 0$ vs $H_1 :$ at least one $\beta_2, \beta_4 \neq 0$, then we have

$$\frac{(SS_E^{(1)} - SS_E^{(2)})/p}{SS_E^{(2)}/(n-k-p-1)} \sim F(p, n-k-p-1)$$

$$\begin{cases} \text{TS} = \frac{(SS_E^{(1)} - SS_E^{(2)})/p}{SS_E^{(2)}/(n-k-p-1)} = \frac{(26054.6-24082.1)/2}{24082.1/45} \approx 1.84 \\ C = (F_\alpha(p, n-k-p-1), +\infty) = (F_{0.05}(2,45), \infty) \approx (3.2, \infty). \end{cases}$$

Since $TS \notin C$, we don't reject H_0, namely, x_2 and x_4 might not be useful. So the model with two variables x_1 and x_3 is sufficient. □

Svensk Version

1 (4 poäng)

Ytojämnheten har bestämts fär fyra olika material som används för inkapsling . Resultat är:

Material	Ytojämnhet						$\bar{x}_i$	s_i
Typ 1	0.50	0.55	0.55	0.36			0.4900	0.0898
Typ 2	0.31	0.07	0.25	0.18	0.56	0.20	0.2617	0.1665
Typ 3	0.20	0.28	0.12				0.2000	0.0800
Typ 4	0.10	0.16					0.1300	0.0424

Anta att datamaterialet härrör från oberoende normalfördelning $N(\mu_i, \sigma), i = 1, 2, 3, 4$.
(1.1). (1p) konstruera ett konfidensintervall för σ^2 och med konfidensgraden 0.95.
(1.2). (1p) konstruera ett konfidensintervall för μ_1 och med konfidensgraden 0.95.
(1.3). (2p) Förefaller det troligt att $\mu_1 > 1.2\mu_2$? Besvara frågan genom att konstruera ett lämpligt konfidensintervall med konfidensgraden 0.95.

2 (2.5 poäng)

Tio observationer {4.1, 4.4, 3.9, 3.8, 4.0, 4.3, 4.0, 3.9, 4.4, 4.2 } är från $N(\mu, 0.2)$.
(2.1) (1p) Pröva hypotesen $H_0 : \mu = 4$ mot $H_1 : \mu > 4$ på nivån 5%.
(2.2) (1.5p) Beräkna styrkan för $\mu = 3.9$.

3 (2.5 poäng)

Den stokastiska variabeln $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$ har en tredimensionell normalfördelning med väntevärdesmatris och kovariansmatris

$$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \text{ respektive } \begin{pmatrix} 7 & 1 & -2 \\ 1 & 1 & 0 \\ -2 & 0 & 1 \end{pmatrix}.$$

(3.1). (0.5p) Är X_1 och X_2 oberoende? Varför?
(3.2). (0.5p) Är X_3 och X_2 oberoende? Varför?
(3.3). (1.5p) Bestäm väntevärdesmatris och kovariansmatris för $\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix}$, där

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

4 (3 poäng)

Ett visst mätfel har likformig fördelning på $[0, \theta]$, täthesfunktionen $f_X(x) = \frac{1}{\theta}$, där $0 \leq x \leq \theta$ och $\theta > 0$. Vi har mätfelen $\{x_1, x_2, \ldots, x_n.\}$
(4.1). (0.5p) Hitta en punktskattning $\hat{\theta}_{MM}$ av θ genom att använda momentmetoden.
(4.2). (0.5p) Är den punktskattningen $\hat{\theta}_{MM}$ väntevärdesriktig? Varför?
(4.3). (2p) Hitta en punktskattning $\hat{\theta}_{ML}$ av θ genom att använda Maximum Likelihood metoden.

5 (4 poäng)

Ett företag har ett lager, där varor hämtas med truckar. Under 500 skilda, slumpmässigt valda tidsintervall av längden en timme noterades hur många truckar som anlände. Resultat:

Antal truckar	0	1	2	3	4	5	6	7	8
Frekvens	52	151	130	102	45	12	5	1	2

Undersök med att ett χ^2-test på nivån 5% om antalet truckar som anländer under en timme är $Po(\mu)$.

6 (2 poäng)

I en studie av överlevnadstiden för patienter med prostatacancer har man för varje patient noterat behandlingstypen (x_1), åldern i år (x_2) vid behandlingens början, halten (x_3) av ett visst karakteristiskt ämne, AP, i blodet, förekomsten av skelettmetastaser (x_4) samt överlevnadstiden (y) i månader räknat från behandlingens början. Vi har

$$x_1 = \begin{cases} 0 & \text{vid placebobehandling (= ingen behandling)} \\ 1 & \text{vid östrogenbehandling .} \end{cases}$$

$$x_4 = \begin{cases} 0 & \text{om skelettmetastaser ej förekommer} \\ 1 & \text{om skelettmetastaser förekommer.} \end{cases}$$

En variansanalys av datamaterialet för modellerna

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon, \qquad \varepsilon \sim N(0, \sigma)$$

and

$$Y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \varepsilon, \qquad \varepsilon \sim N(0, \sigma)$$

Variansanalys
Skattad regressionslinje: $y = 104.1 + 10.0x_1 - 0.96x_2 - 0.03x_3 - 4.32x_4$

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	104.076	37.7264
1	10.0219	6.69623
2	-0.956706	0.506024
3	-0.0262767	0.00907681
4	-4.31777	7.18044

	Frihetsgrader	Kvadratsumma
REGR	4	6665.0
RES	45	24082.1
TOT	49	30747.1

Variansanalys
Skattad regressionslinje: $y = 32.9 + 9.17x_1 - 0.022x_3$

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	32.8860	4.97944
1	9.17159	6.65964
2	-0.0220795	0.00858002

	Frihetsgrader	Kvadratsumma
REGR	2	4692.56
RES	47	26054.6
TOT	49	30747.1

(6.1). (1p) Välj lämplig modell. Påverkar skeletmetastaser överlevnadstiden? Motivera ditt svar genom att konstruera ett lämpligt konfidensintervall eller test med konfidensgraden 90%.
(6.2). (1p) Är det meningsfullt att ta x_2 eller x_4? Genomför ett lämpligt test på nivån 5%.