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)

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

$$f(x) = \begin{cases} \frac{1}{\theta} & \text{if } 0 \le x \le \theta; \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$ is an unknown parameter. A sample $\{x_1, x_2, \dots, x_4\}$ from this population is now given.
(1.1). (1p) Find a point estimate $\hat{\theta}_{MM}$ of θ using Method of Moments.
(1.2). (1p) Is $\hat{\Theta}_{MM}$ unbiased?
(1.3). (1p) Which of the estimates $\hat{\Theta}_{MM}$ in (1.1.) and $\hat{\Theta} = X_1 + X_2$ is more effective?

Solution. (1.1) The first equation from the Method of Moments is $E(X) = \bar{x}$. Since $E(X) = \int_0^\theta x f(x)dx = \frac{\theta}{2}$, we have

$$\frac{\theta}{2} = \bar{x}, \text{ thus } \hat{\theta}_{MM} = 2\bar{x}.$$

(1.2) The point estimator $\hat{\Theta}_{MM} = 2\bar{X}$. Then we have

$$E(\hat{\Theta}_{MM}) = E(2\bar{X}) = 2E(X) = 2 \times \frac{\theta}{2} = \theta$$

Therefore, $\hat{\Theta}_{MM}$ is unbiased.
(1.3) Because

$$V(\hat{\Theta}_{MM}) = V(2\bar{X}) = V(2 \times \frac{X_1 + X_2 + X_3 + X_4}{4}) = \frac{1}{2^2} \times 4 \times V(X) = E(X^2) - (E(X))^2 = \int_0^\theta x^2 f(x)dx - (\frac{\theta}{2})^2 = \frac{\theta^2}{12}$$

and

$$V(\hat{\Theta}) = V(X_1 + X_2) = 2 \times V(X) = 2(E(X^2) - (E(X))^2) = 2 \times \frac{\theta^2}{12} = \frac{\theta^2}{6}$$

So $\hat{\Theta}_{MM}$ is more effective than $\hat{\Theta}$ since $V(\hat{\Theta}_{MM}) < V(\Theta)$. □

2 (3 points)

At a comparison of two opinion surveys it was seen that out of 1704 participants 46.5% supported the borgerliga parties in October. In November, 45.6% out of 1689 participants supported the borgerliga parties. Find a confidence interval with approximate confidence level 95% for the change of the support proportion between the two surveys. The number of eligible voters can be thought of as infinite compared to the sample sizes.

Solution. Here we have two Binomial distributions: $X \sim Bin(N_1, p_1)$ and $Y \sim Bin(N_2, p_2)$ with unknown p_1 and p_2. Then we have the following formula for the confidence interval of $p_1 - p_2$ which is

$$Ip_1 - p_2 = (\hat{p_1} - \hat{p_2}) \pm \lambda_{\alpha/2}\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{N_1} + \frac{\hat{p_2}(1-\hat{p_2})}{N_2}} = (-0.025, \quad 0.043).$$

Here $\hat{p_1} = 46.5\%$, $\hat{p_2} = 45.6\%$, $N_1 = 1704$, $N_2 = 1689$ and $\lambda_{\alpha/2} = \lambda_{0.025} = 1.96$.

□

3 (3 points)

Lucas plays on a slot machine that gives a prize with the unknown probability p. The number of games X including the one when the first prize is awarded has the probability function

$$p_X(k) = p(1-p)^{k-1}, \text{ for } k = 1, 2, 3...$$

It is claimed that $p = 0.1$ but Lucas doubts that p is too small and wishes to test the hypothesis $H_0 : p = 0.1$ against $H_1 : p > 0.1$. Is it possible, using the significance level 0.05, to reject H_0 if he loses the first three games and wins on the fourth? Use the P-value method.

Solution. Let X be the distribution with the probability function $p_X(k) = p(1-p)^{k-1}$, for $k = 1, 2, 3...$, then we assume H_0 is ture, then $p = 0.1$, so we have

$$P - \text{value} = P(X \leq 4) = \sum_{k=1}^{4} p(1-p)^{k-1} = 34.39\% > \alpha$$

Here $\alpha = 0.05$, so we don't reject H_0. Namely, it is not possible to reject H_0.

□

4 (3 points)

Sample results of car preferences for male and female users (observed frequencies) are as follows

	Saab	Volvo	Toyota
Male	30	40	20
Female	20	30	30

Do Male and Female show different preferences (with a significance level $\alpha = 5\%$)?

Solution.

$$H_0 : \text{Male and Female have the same preference;}$$
$$H_1 : \text{Male and Female have different preferences.}$$

It is easy to get

$$\text{Male proportion: } p_{male} = \frac{90}{170};$$
$$\text{Female proportion: } p_{female} = \frac{80}{170};$$
$$\text{Saab proportion: } q_{saab} = \frac{50}{170};$$
$$\text{Volvo proportion: } q_{volvo} = \frac{70}{170};$$
$$\text{Toyota proportion: } q_{toyota} = \frac{50}{170}.$$

Thus

$$\begin{aligned}
np_{11} &= 170 \times p_{male} \times q_{saab} = 26.47;\\
np_{12} &= 170 \times p_{male} \times q_{volvo} = 37.06;\\
np_{13} &= 170 \times p_{male} \times q_{toyota} = 26.47;\\
np_{21} &= 170 \times p_{female} \times q_{saab} = 23.53;\\
np_{22} &= 170 \times p_{female} \times q_{volvo} = 32.94;\\
np_{23} &= 170 \times p_{female} \times q_{toyota} = 23.53.
\end{aligned}$$

What is more, $N_{11} = 30, N_{12} = 40, N_{13} = 20, N_{21} = 20, N_{22} = 30$ and $N_{23} = 30$. Therefore,

$$TS = \sum_{j=1}^{3}\sum_{i=1}^{2} \frac{(N_{ij} - np_{ij})^2}{np_{ij}} = 4.86.$$

$$C = (\chi^2_\alpha((2-1)(3-1)), \quad +\infty) = (5.99, \quad +\infty).$$

Since $TS \notin C$, we don't reject H_0. Namely, we don't know if Male and Female have different preferences □

5 (3 points)

At a certain moment in a communications system the received signal Y can be written on the form $Y = X + Z$, where X is the transmitted signal and Z is noise independent of X. Furthermore $X \sim N(5, 2)$ and $Z \sim N(0, 1)$.
(5.1). (1.5p) Find the distribution for the random vector with the components X and Y.
(5.2). (1.5p) We wish to reconstruct X by using a linear function $aY + b$ of the received signal. Find constants a and b such that $E[aY + b] = E[X]$ and such that $V[X - aY - b]$ is minimized.

Solution. (5.1). We let W be the random vector with the components X and Y, then we have

$$\mathbf{W} = \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} X \\ X+Z \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} X \\ Z \end{pmatrix} = A\begin{pmatrix} X \\ Z \end{pmatrix}.$$

We know that X and Z are independent, and we let $\mathbf{U} = \begin{pmatrix} X \\ Z \end{pmatrix}$, it is known that

$$\mu_{\mathbf{U}} = \begin{pmatrix} 5 \\ 0 \end{pmatrix}, \quad C_{\mathbf{U}} = \begin{pmatrix} 2^2 & 0 \\ 0 & 1 \end{pmatrix}.$$

Thus for $\mathbf{W} = A\mathbf{U}$ with

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

Therefore $\mathbf{W} = \begin{pmatrix} X \\ Y \end{pmatrix}$ is a normal vector with

$$\mu_{\mathbf{W}} = A\mu_{\mathbf{U}} = \begin{pmatrix} 5 \\ 5 \end{pmatrix}, \quad C_{\mathbf{W}} = AC_{\mathbf{U}}A' = \begin{pmatrix} 4 & 4 \\ 4 & 5 \end{pmatrix}.$$

(5.2). According to the equation $E[aY + b] = E[X]$, we get $5a + b = 5$.

$$V[X - aY - b] = V[(1-a)X - aZ - b] = (1-a)^2V(X) + (-a)^2V(Z) = 2^2(1-a)^2 + a^2$$

To get the minimal value of $V[X - aY - b]$, we get the critical point $a = 0.8$ by finding the first derivative.
Then we can get $5 \times 0.8 + b = 5$, which produce $b = 1$.

□

6 (3 points)

In a study of the profitability of movie companies, 20 Hollywood films were selected randomly and for each film we observed values on

$$y = \text{gross revenue (unit: million dollar)},$$
$$x_1 = \text{production costs (unit: million dollar)},$$
$$x_2 = \text{marketing costs (unit: million dollar)},$$
$$x_3 = \begin{cases} 1, & \text{for a film based on a book,} \\ 0, & \text{others.} \end{cases}$$

There is a data which has been analyzed according to the model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon,$$

where ε is assumed to be $N(0, \sigma)$. The analysis of variances is as follows:

Estimated regression line: $y = 7.84 + 2.85x_1 + 2.28x_2 + 7.17x_3$

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	7.836	2.333
1	2.8477	0.3923
2	2.2782	0.2534
3	7.166	1.818

	Degrees of freedom	Sum of squares
REGR	3	6325.2
RES	?	?
TOT	19	6542.9

(6.1). (1p) Estimate σ^2.
(6.2). (1p) Test

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \text{ against } H_1 : \text{ at least one of } \beta_1, \beta_2 \text{ and } \beta_3 \neq 0$$

with a significance level $\alpha = 5\%$.
(6.3). (1p) Is the variable x_3 useful/important for the model? Answer the question by an appropriate test with a significance level $\alpha = 1\%$.

Solution. (6.1). $\hat{\sigma}^2 = s^2 = \frac{SS_E}{n-k-1} = \frac{SS_{TOT} - SS_R}{20-3-1} = 13.61.$
(6.2). According the hypothesis testing, we have the following test statistic TS and rejection region (critical region) C:

$$TS = \frac{SS_R/k}{SS_E/(n-k-1)} = \frac{6325.2/3}{13.61} = 154.92 \text{ and } C = (F_\alpha(k, n-k-1), \quad \infty) = (F_{0.05}(3, 16), \infty) = (3.24, \infty)$$

Since $TS \in C$, we reject H_0.
(6.3). We have the hypothesis testing

$$H_0 : \beta_3 = 0 \qquad \textit{against} \qquad H_1 : \beta_3 \neq 0$$

$TS = \frac{\hat{\beta}_3 - 0}{d(\hat{\beta}_3)} = 7.166/1.818 = 3.94$, and $C = (-\infty, -t_{\alpha/2}(n-k-1)) \cup (t_{\alpha/2}(n-k-1), +\infty) = (-\infty, -2.92) \cup (2.92, +\infty)$.
Since $TS \in C$, we reject H_0. So we believe that $\beta_3 \neq 0$, and the variable x_3 is useful/important. □

Svensk Version

1 (3 poäng)

Antag att fördelningen för en population har täthetsfunktionen

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

där $\theta > 0$ är en okänd parameter. $\{x_1, x_2, \ldots, x_4\}$ är ett stickprov från populationen.
(1.1). (1p) Hitta en punktskattning $\hat{\theta}_{MM}$ av θ genom att använda momentmetoden.
(1.2). (1p) Är $\hat{\Theta}_{MM}$ väntevärdesriktig?
(1.3). (1p) Vilken av skattningarna $\hat{\Theta}_{MM}$ i (1.1.) och $\hat{\Theta} = X_1 + X_2$ är effektivare?

2 (3 poäng)

Vid jämförelse av två opinionsundersökningar framgår att av de 1704 intervjuade i oktober sympatiserade 46.5% med det borgerliga blocket och 45.6% av de 1689 i november. Beräkna ett konfidensintervall för förändringen av andelen mellan de två undersökningarna med approximativ konfidensgrad 95%. Väljarkåren kan betraktas som oändligt stor jämfört med urvalens storlek.

3 (3 poäng)

Lucas spelar på en spelautomat som ger vinst med den okända sannolikheten p. Antalet spel X till och med första vinsten har då sannolikhetsfunktionen

$$p_X(k) = p(1-p)^{k-1}, \text{ för } k = 1, 2, 3...$$

Det påstås att $p = 0.1$ men Lucas betvivlar att p är så liten och vill pröva hypotesen $H_0 : p = 0.1$ mot $H_1 : p > 0.1$. Kan han på signifikansnivån 0.05 förkasta H_0 om han finner att han förlorar de tre första spelen och vinner det fjärde? Använd P-värdesmetoden.

4 (3 points)

Resultat från bil preferenser för manliga och kvinnliga användare (observerade frekvenser) finns nedan

	Saab	Volvo	Toyota
Male	30	40	20
Female	20	30	30

Har man och kvinna olika preferenser (med nivån $\alpha = 5\%$)?

5 (3 poäng)

I ett kommunikationssystem kan den i ett visst ögonblick mottagna signalen Y skrivas på formen $Y = X + Z$, där X är den verkliga utsända signalen och Z en störning som är oberoende av X. Vidare gäller att $X \sim N(5, 2)$ och $Z \sim N(0, 1)$.
(5.1). (1.5p) Bestäm fördelningen för den stokastiska vektorn med komponenter X och Y.
(5.2). (1.5p) Man vill rekonstruera X med hjälp av en linjär funktion $aY + b$ av den mottagna signalen. Bestäm konstanterna a och b så att $E[aY + b] = E[X]$ och $V[X - aY - b]$ är minimal.

6 (3 poäng)

I en studie av lönsamheten för filmbolag har man valt ut 20 hollywoodfilmer slumpmässigt och för varje film tagit fram observerade värden på

$$
\begin{aligned}
y &= \text{bruttointäkt (enhet: miljoner dollar)}, \\
x_1 &= \text{produktionskostnad (enhet: miljoner dollar)}, \\
x_2 &= \text{marknadsföringskostnad (enhet: miljoner dollar)}, \\
x_3 &= \begin{cases} 1, & \text{för film baserad på en bok}, \\ 0, & \text{annars}. \end{cases}
\end{aligned}
$$

Det finns en data som har analyserats enligt modellen

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon,$$

där ε antas vara $N(0, \sigma)$. Variansanalyser finns nedan:

Skattad regressionslinje: $y = 7.84 + 2.85x_1 + 2.28x_2 + 7.17x_3$

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	7.836	2.333
1	2.8477	0.3923
2	2.2782	0.2534
3	7.166	1.818

	Frihetsgrader	Kvadratsumma
REGR	3	6325.2
RES	?	?
TOT	19	6542.9

(6.1). (1p) Skatta σ^2.
(6.2). (1p) Pröva på nivån 0.05

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \text{ mot } H_1 : \text{ minst en av } \beta_1, \beta_2 \text{ och } \beta_3 \neq 0.$$

(6.3). (1p) Är den x_3 nyttiga/viktiga för modellen? Motivera svaret med ett lämpligt test på nivån $\alpha = 1\%$.