Inferential Statistics
Semester 4
2023/2024

Name:..Group:..................................

Final exam - Duration: 2h30

For each test, the statistical model must be specified and the hypotheses clearly stated. All tests will be performed at the $\alpha = 0.05$ level, and results given to an accuracy of 10^{-3}. The four parts of the following problem can be treated independently. Extracts from statistical tables are given at the end of the problem.

Problem 1 *The first-serve speeds of Novak Djokovic, Rafael Nadal, Stefanos Tsitsipas and Alexander Zverev were recorded during the 2009 Roland Garros semi-finals. The aim is to determine whether they are statistically significantly comparable.*

Part A.

The speeds of the first 10 serves of the four players are recorded. The following results are obtained (in km/h).

For N. Djokovic:	209	204	219	221	189	183	206	188	209	178
For R. Nadal:	193	181	195	205	213	199	218	172	188	175
For S. Tsitsipas:	167	181	185	194	196	199	203	207	212	217
For A. Zverev:	165	178	181	185	188	190	196	199	205	219

1. *According to this data, are the average first serve speeds of the four players comparable?*

We want to test $H_0: m_1 = m_2 = m_3 = m_4$ vs $H_1: \exists i,j = 1..4$ st $m_i \neq m_j$
the distribution of the population is unknown and the size of the samples is less than 30, then we can't use the ANOVA, so we use the non-parametric test of Kruskal and Wallis. For that we order all the values:

0,25

165 - 167 - 172 - 175 - 178 - 178 - 181 - 181 - 181 - 183 - 185 - 185
1 2 3 4 5,5 5,5 8 8 8 10 11,5 11,5
188 - 188 - 188 - 189 - 190 - 193 - 194 - 195 - 196 - 196 - 199 - 199 - 199
14 14 14 16 17 18 19 20 21,5 21,5 24 24 24
203 - 204 - 205 - 205 - 206 - 207 - 209 - 209 - 212 - 213 - 217 - 218
26 27 28,5 28,5 30 31 32,5 32,5 34 35 36 37
219 - 219 - 221
38,5 38,5 40

0,5

Then we calculate the sum of the ranks in each sample.

$r_1 = 5.5 + 10 + 14 + 16 + 27 + 30 + 32.5 + 32.5 + 38.5 + 40 = 246$

$r_2 = 3 + 4 + 8 + 14 + 18 + 20 + 24 + 28.5 + 35 + 37 = 191.5$

$r_3 = 2 + 8 + 11.5 + 19 + 21.5 + 24 + 26 + 31 + 34 + 36 = 213$

$r_4 = 1 + 5.5 + 8 + 11.5 + 14 + 17 + 21.5 + 24 + 28.5 + 38.5 = 169.5$

1

0,25

We have $h = \frac{12}{40(41)} \left[\frac{1}{10} (246^2 + 191.5^2 + 213^2 + 169.5^2) \right] - 3.41 = 2{,}33$

All the n_i are greater than 5 so the r.v. $H \sim \chi^2(k-1=3)$

$P(H \geq \chi^2_{0.05}(3)) = 0.5 \Rightarrow \chi^2_{0.05}(3) = 7.81$

0,5

Since $h < \chi^2_{0.05}(3)$ the we accept H_0 and then the means are significatively equal.

So the first serves of the four tennismen are approximabilly the same.

2. *Using a statistical test, investigate whether Djokovic's average first serve speed is higher than Nadal's.*

We want to test here $H_0: m_1 \leq m_2$ v.s $H_1: m_1 > m_2$.
In this case also we use the non parametric test of Wilcoxon since we have paired sample,

0,25

We the calculate the difference d between each paired values, and then we order their absolute values.

0,5

D.	209	204	219	221	189	183	206	188	209	178
N.	193	181	195	205	213	199	218	172	188	175
d	16	23	24	16	-24	-16	-12	16	21	3

the ordred difference is:

0,25

3	-12	16	16	16	-16	21	23	24	-24
1	2	4,5	4,5	4,5	4,5	7	8	9,5	9,5

and we calculate

$W_+ = 1 + 4,5 + 4,5 + 4,5 + 7 + 8 + 9,5 = 39$

$W_- = 2 + 4,5 + 9,5 = 16$ (0,5)

The relation $W_+ + W_- = 55 = \frac{10(11)}{2} = \frac{N(N+1)}{2}$ is verified

We ~~take~~ $W_c = \min(W_+, W_-) = 16$

And from the Wilcoxon table we have

$W_\alpha = W_{0,05} = 8$ so $W_c > W_{0,05}$ (0,25)

Then we ~~reject~~ H_0 and ~~m_1~~ is [illegible] greater than m_2

The ~~speed~~ mean speed of the first service of Djokovic is [illegible] greater than Nadal's one. (0,25)

3. *What test could be performed if we now wanted to be sure of the independence between the first serve speeds of the two players (Djokovic and Nadal)? Interpret the result obtained.*

In this situation the non-parametric test of independance of Spearman for the hypothesis.

(0,25)

H_0: "the first serve speeds of the two players are independant"

H_1: "the first serve speeds of the two players are not independant"

(0,5)

D	209	204	219	221	189	183	206	188	209	178
x'	7,5	5	9	10	4	2	6	3	7,5	1
N	193	181	195	205	213	199	218	172	188	175
y'	5	3	6	8	9	7	10	1	4	2
$d_i = \lvert x' - y' \rvert$	2,5	2	3	2	5	5	4	2	3,5	1

0,5

We have $r_S = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6.106,5}{10.99} = 0,355$

here $n = 10 < 13$ then r_α correspond to $P(|R_S| \geq r_\alpha) = \alpha$, for $\alpha = 0,05$ we have $r_{0,05} = 0,64$ 0,25

$|r_S| < r_{0,05}$ then we accept H_0 et so the fast serve speeds of the two players are independant 0,25

Part B.

4. *Using a Chi-squared test on the classes*]170, 180],]180, 190],]190, 200],]200, 210],]210, 220],]220, 230], *say whether N. Djokovic can indeed be considered as the realization of a Gaussian distribution sample* $N(m_1, \sigma_1^2)$ *of size* 10 *(with the parameters* m_1 *and* σ_1^2 *unknown, for which care will be taken to show that they can be estimated by* $\widehat{m_1} = 200.6$ *and* $\widehat{\sigma_1^2} = 205.04$*).*

If $X \leadsto N(m_1, \sigma_1^2)$ then $Z = \frac{X - \hat{m}_1}{\sigma_1} \leadsto N(0,1)$,

To do the Khi2 test we calculate $C_i = P(z_1 \leq Z \leq z_2)$ and draw the following table

Speed	z_1	z_2	$P_i = P(z_1 \leq Z \leq z_2)$	$C_i = n \cdot P_i$
]170; 180]	-2,14	-1,44	0,059	0,59
]180; 190]	-1,44	-0,74	0,154	1,54
]190; 200]	-0,74	-0,04	0,254	2,54
]200, 210]	-0,04	0,66	0,261	2,61
]210, 220]	0,66	1,35	0,168	1,68
]220, 230]	1,35	2,05	0,068	0,68
Total				

0,5

All the $C_i < 5$

Then we can't use this test it is prefered to use the Kolmogorov Test

We have then in this case

~~Speed~~	n_i	$P_i = F_i^{th}$	F_i^{e}	Difference
]170, 180]	1	0,059	0,1	0,041
]180; 190]	3	0,213	0,4	0,187
]190, 200]	0	0,467	0,4	0,067
]200, 210]	4	0,728	0,8	0,072
]210, 220]	1	0,896	0,9	0,004
]220, 230]	1	0,964	1	0,036

0,5

$\Rightarrow D_n = 0,187$

Name:..**Group:**..............................

We have for $n=10$ and $\alpha=0{,}05$ $c=1{,}294$

Then $\frac{c}{\sqrt{n}}=0{,}409$ so $D_n < \frac{c}{\sqrt{n}}$ 0,5

$\Rightarrow$ we accept the normality of the population.

5. *Is the estimator $\widehat{\sigma_1^2}$, of the variance, considered in the previous question biased or unbiased? Justify your answer?*

the variance $\hat{\sigma}_1^2 = 205{,}04$ is calculated with the formula $\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{m}_1)^2$ 0,25 then is the unbiased estimator.

Since we have $\mathbb{E}[\hat{\sigma}_1^2] = \mathbb{E}\left[\frac{1}{n}\sum(X_i-\hat{m}_1)^2\right] = \frac{n-1}{n}\sigma^2$

0,25

Part C.

We now assume that the first serve speeds of the four players follow Gaussian distributions $\mathbf{N}(m_i, \sigma_i), i = 1, \cdots, 4$.

6. *Show that the speeds of the first serves of the four players can be assumed to have the same dispersion.*

0,25 We have to test $H_0: \sigma_1^2=\sigma_2^2=\sigma_3^2=\sigma_4^2$ v.s $H_1: \exists i,j=1..4: \sigma_i^2 \neq \sigma_j^2$ using the Bartlett test.

Here we have: $n_1=10$ $m_1=200{,}6$ $S_1^2=205{,}04$

$n_2=10$ $m_2=193{,}80$ $S_2^2=213{,}49$

0,5 $n_3=10$ $m_3=196{,}10$ $S_3^2=206{,}69$

$n_4=10$ $m_4=190{,}60$ $S_4^2=205{,}84$.

the residual variance is $S_R^2=\frac{\sum S_i^2}{4}=207{,}76$ 0,25

We have $\Delta = 1+\frac{1}{3(k-1)}\left[\left(\sum_{i=1}^{4}\frac{1}{n_i-1}\right)-\frac{1}{n-k}\right]=1+\frac{1}{3.3}\left[\frac{4}{9}-\frac{1}{36}\right]=\frac{113}{108}$

0,25

and $b=\frac{108}{113}\left[36\ln(207{,}76)-9\left[\ln 205{,}04+\ln 213{,}49+\ln 206{,}69+\ln 205{,}84\right]\right]$

0,25

$\Rightarrow b = 0{,}056$

we have $\chi^2_{0,05}(k-1=3) = 7{,}81$

0,25

$b < \chi^2_{0,05}(3) \Rightarrow$ we accept H_0 and the the variance are equal.

7. *Show that the average first serve speeds of the four players are comparable.*

0,25 To test $H_0: m_1 = m_2 = m_3 = m_4$ v.s $H_1: \exists i,j : m_i \neq m_j$ we use the ANOVA.

we determine first the total variance

$n = 40$ $m = 195{,}3$ and $S^2 = 220{,}96$. 0,25

Then $S_F^2 = S^2 - S_R^2 = 220{,}96 - 207{,}76 = 13{,}2$. 0,25

so the table of ANOVA is

Source of variation	Sum of squares	ddf	mean of squares	F
Factorial	$nS_F^2 = 528$	$k-1=3$	176	$f = \dfrac{\frac{nS_F^2}{k-1}}{\frac{n \cdot S_R^2}{n-k}} = 0{,}762$ 0,25
Residual	$nS_R^2 = 8310{,}4$	$n-k=36$	230,844	
Total	$nS^2 = 8838{,}4$	$n-1=39$		

1

0,25 We have $f_{0,05}(3,36) = 2{,}866 \Rightarrow f < f_{0,05}(3,36)$.

0,25 the we accept H_0 and the mean of first serve speeds are equal.

8. *Find out whether Tsitsipas' first serve speed is greater than Zverev's. Remember to check that the two variances are equal before you do this.*

0,25 Here we want to test $H_0: m_3 \leq m_4$ v.s $H_1: m_3 > m_4$.

the variances must be equal then we test first

0,25 $H_0: \sigma_3^2 = \sigma_4^2$ v.s $H_1: \sigma_3^2 \neq \sigma_4^2$. using the

0,25 statistics $F = \dfrac{n_1 S_3^2}{n_1 - 1} \dfrac{n_2 - 1}{n_2 S_4^2} \rightsquigarrow \mathcal{F}(n_1-1, n_2-1) = \mathcal{F}(9,9)$.

$f_e = 1{,}024$ and $I_\alpha = [f_{\frac{\alpha}{2}} ; f_{1-\frac{\alpha}{2}}] = [f_{0,025} ; f_{0,975}]$

0,25

$f_{0,975}(9,9) = 4,026 \Rightarrow f_{0,025}(9,9) = \frac{1}{f_{0,975}(9,9)} = 0,248$ 0,5

$\Rightarrow I_{0,05} = [0,248 ; 4,026]$; $f_c \in I_{0,05}$

$\Rightarrow$ We accept H_0 and the variance are equal. 0,25

* Now we test $H_0: m_3 \leq m_4$ v.s $H_1: m_3 > m_4$
0,25 with unknown variances, so we use the statistics

$$T = \frac{\overline{X_3} - \overline{X_4}}{\sqrt{\left(\frac{1}{n_3} + \frac{1}{n_4}\right)\left(\frac{n_3 S_3^2 + n_4 S_4^2}{n_3 + n_4 - 2}\right)}} \rightsquigarrow \mathcal{T}(n_3 + n_4 - 2 = 18),$$

0,5

$$t_c = \frac{196,10 - 199,60}{\sqrt{\left(\frac{1}{5}\right) \frac{10}{18}(206,69 + 205,84)}} \simeq 0,812$$

0,25

$I_{0,05} =]-\infty, t_{1-\alpha}(18)] =]-\infty, 1,734] \Rightarrow t_c \in I_{0,05}$

0,25

then we accept H_0, thus the mean speeds of Tsitsipas is not > than Zverev's one.

9. *Propose a test to ensure independence between the first serve speeds of the two players (Tsitsipas and Zverev)? Interpret the result obtained.*

To ensure de indepence we use here the correlation test so to test H_0: "The speeds are independent"
v.s H_1: "The speeds are not independant" 0,25
we use the statistics

$$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \rightsquigarrow \mathcal{T}(n-2 = 8)$$

0,5

we calculate $r = \frac{Cov(T,Z)}{\sigma_T \cdot \sigma_Z} = \frac{\frac{375770}{10} - 196,1 \times 190,6}{\sqrt{206,69 \cdot 205,84}} \simeq 0,971$

then $t = \frac{0,971\sqrt{8}}{\sqrt{1-0,971^2}} \simeq 11,487$

0,25

~~For~~ $\alpha = 0{,}05$ we have $t_{0{,}05}(8) = 2{,}306$ (0,25)

$\Rightarrow I_{0{,}05} =]-2{,}306\,;\,2{,}306[$ $\qquad t_c \notin I_{0{,}05}$

We reject H_0 and we conclude that the speeds are dependent. (0,25)

Part D.

In this section, we'd like to find out whether or not first serve speed is related to tennis player height. To do this, we record the first serve speeds and heights of 100 tennis players chosen at random from the top 200 of the ATP rankings. The results are as follows:

Speed / Size	*Low*	*Moderate*	*Fast*
Small	8	10	7
Medium	7	22	12
Large	5	12	17

10. *Using a test of independence, to be precisely described and justified, determine whether or not the speed of the first serve is related to the tennis player's height.*

we have qualitative characters then to test the independence we use the Khi2 test. (0,25)

so we calculate $C_{ij} = \frac{n_{i\cdot}\, n_{\cdot j}}{n}$ then we will have the following table

Speed / Size	Low	Moderate	Fast	$n_{i\cdot}$
Small	8 5	10 11	7 9	25
Medium	7 8,2	22 18,04	12 14,76	41
Large	5 6,8	12 14,96	17 12,24	34
$n_{\cdot j}$	20	44	36	100

n_{ij}

C_{ij}

(1)

All the $C_{ij} \geq 5$.

we calculate $K = \sum_{ij} \frac{(n_{ij} - C_{ij})^2}{C_{ij}} \rightsquigarrow \chi^2((k-1)(\ell-1)) = \chi^2(4)$

(0,5)

$\chi_c^2 = 6{,}810$ and $\chi^2_{0{,}05}(4) = 9{,}488$. (0,25)

Since $\chi_c^2 < \chi^2_{0{,}05}(4)$ then we accept H_0 and the characters are independent. (0,5)