SECI 2143 PROBABILITY AND STATISTICAL DATA ANALYSIS
Project II – Inferential Statistics

World Wushu Championships in 2017

ZHANG XIAOMENG
A19EC2024
Section 05
School Of Computing, Faculty of Engineering
Universiti Teknologi Malaysia

Abstract- Martial arts are both art and competition. There was holding competition in World Wushu Champion in 2017. Therefore this study according to this year competition to analysis that athletes' performance wouldn't be affected by time spend and the number of attempted degree of difficulty movements.. during competition. What it would affect them in this study which are gender and regions.

I. INTRODUCTION

Wushu means Chinese Martial Arts was an inherited technique of military warfare in ancient China. Practicing martial arts can strengthen the body and defend against attacks from the enemy. People who practice martial arts are guided by the technique of "stopping the invasion" and lead the practitioners into the traditional education (materialization) method of knowing people and nature, and the objective laws of society. It is the guidance and guarantee of human material civilization, and is the contemporary traditional martial arts, also is performance.

This is data from the 2017 World Wushu Championships. Which is scored by the International Wushu Federation(IWUF) and Justtool. This competition is from Contemporary Wushu whrer athletes perform a routine on a carpet. Their routine is made up of punches, kicks, jumps, tumbling, stances, and balances. The athletes are scored on three components: 1) A score: Quality of Movements. 2) B Score: Evaluation of Overall Performance. 3) C Score: Degree of Difficulty. Together, these three components determine the overall score. And the main purpose to find out if there are some factors effect overall score.

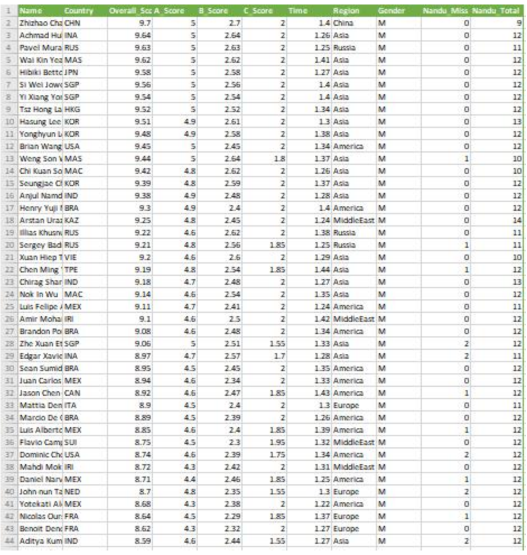
II. METHODOLOGY

This data is secondary data which was collected through an online data science community that called *kaggle.com(https://www.kaggle.com/predact/world-wushu-championships-2017)*. And this data is about 2017 World Wushu Championships. There recorded 201 athletes who are from 177 countries who participated in the competition. Therefore, 201 samples with 17 variables from this data set to analysis some potential factors which affect their performance then lost scores. This data set is analyzed through four types of statistic test -- expected outcome: 1) Hypothesis testing -- 2 sample testing: to analyse the relationship between Overall scores and Gender(Males have higher Overall scores than females in level of significance 0.05). 2) Correlation: to analyse the relationship between Overall score and Time spend(There is strong positive linear correlation between Overall and Time Spend）. 3) Regression: to analyse the relationship between Overall score and Nandu_Tota(Nandu_Total will strongly affect Overall score). 4) ANOVA: to analyse the relationship between Region (Asia, America, Europe) and Overall Score(Region affects scores). And through these statistic tests, there are five variables are going to be used(all ratio):

Overall_Score: Final score of the athletes.
Gender: The athlete’s gender -- Male or Female
Time: The duration of the athlete’s performance
Nandu_Total: The number of attempted degree of difficulty movements.
Region: The region of the athlete’s country.

And meanwhile R studio is also going to be used to help me to carry out graphically statistical analysis with the data set.

III. **RESULT AND ANALYSIS**



1	Name	Country	Overall_Scc	A_Score	B_Score	C_Score	Time	Region	Gender	Nandu_Miss	Nandu_Total
2	Zhizhao Cha	CHN	9.7	5	2.7	2	1.4	China	M	0	9
3	Achmad Hu	INA	9.64	5	2.64	2	1.26	Asia	M	0	12
4	Pavel Mura	RUS	9.63	5	2.63	2	1.25	Russia	M	0	11
5	Wai Kin Yea	MAS	9.62	5	2.62	2	1.41	Asia	M	0	12
6	Hibiki Bettc	JPN	9.58	5	2.58	2	1.27	Asia	M	0	12
7	Si Wei Jowo	SGP	9.56	5	2.56	2	1.4	Asia	M	0	12
8	Yi Xiang Yo	SGP	9.54	5	2.54	2	1.4	Asia	M	0	12
9	Tsz Hong La	HKG	9.52	5	2.52	2	1.34	Asia	M	0	12
10	Hasung Lee	KOR	9.51	4.9	2.61	2	1.3	Asia	M	0	13
11	Yonghyun L	KOR	9.48	4.9	2.58	2	1.38	Asia	M	0	12
12	Brian Wang	USA	9.45	5	2.45	2	1.34	America	M	0	12
13	Weng Son V	MAS	9.44	5	2.64	1.8	1.37	Asia	M	1	10
14	Chi Kuan So	MAC	9.42	4.8	2.62	2	1.26	Asia	M	0	10
15	Seungjae Cl	KOR	9.39	4.8	2.59	2	1.37	Asia	M	0	12
16	Anjul Namd	IND	9.38	4.9	2.48	2	1.28	Asia	M	0	12
17	Henry Yuji I	BRA	9.3	4.9	2.4	2	1.4	America	M	0	12
18	Arstan Uraz	KAZ	9.25	4.8	2.45	2	1.24	MiddleEast	M	0	14
19	Illias Khusn	RUS	9.22	4.6	2.62	2	1.38	Russia	M	0	11
20	Sergey Bad	RUS	9.21	4.8	2.56	1.85	1.25	Russia	M	1	11
21	Xuan Hiep T	VIE	9.2	4.6	2.6	2	1.29	Asia	M	0	10
22	Chen Ming '	TPE	9.19	4.8	2.54	1.85	1.44	Asia	M	1	12
23	Chirag Shar	IND	9.18	4.7	2.48	2	1.27	Asia	M	0	13
24	Nok In Wu	MAC	9.14	4.6	2.54	2	1.35	Asia	M	0	12
25	Luis Felipe /	MEX	9.11	4.7	2.41	2	1.24	America	M	0	11
26	Amir Moha	IRI	9.1	4.6	2.5	2	1.42	MiddleEast	M	0	12
27	Brandon Po	BRA	9.08	4.6	2.48	2	1.34	America	M	0	12
28	Zhe Xuan Et	SGP	9.06	5	2.51	1.55	1.33	Asia	M	2	12
29	Edgar Xavic	INA	8.97	4.7	2.57	1.7	1.28	Asia	M	2	11
30	Sean Sumid	BRA	8.95	4.5	2.45	2	1.35	America	M	0	12
31	Juan Carlos	MEX	8.94	4.6	2.34	2	1.33	America	M	0	12
32	Jason Chen	CAN	8.92	4.6	2.47	1.85	1.43	America	M	1	12
33	Mattia Den	ITA	8.9	4.5	2.4	2	1.3	Europe	M	0	11
34	Marcio De (	BRA	8.89	4.5	2.39	2	1.26	America	M	0	12
35	Luis Albertc	MEX	8.85	4.6	2.4	1.85	1.39	America	M	1	12
36	Flavio Cam	SUI	8.75	4.5	2.3	1.95	1.32	MiddleEast	M	0	12
37	Dominic Chc	USA	8.74	4.6	2.39	1.75	1.34	America	M	2	12
38	Mahdi Mok	IRI	8.72	4.3	2.42	2	1.31	MiddleEast	M	0	12
39	Daniel Nar	MEX	8.71	4.4	2.46	1.85	1.25	America	M	1	12
40	John nun Ta	NED	8.7	4.8	2.35	1.55	1.3	Europe	M	2	12
41	Yotekati Al	MEX	8.68	4.3	2.38	2	1.22	America	M	0	12
42	Nicolas Our	FRA	8.64	4.5	2.29	1.85	1.37	Europe	M	1	12
43	Benoit Denc	FRA	8.62	4.3	2.32	2	1.27	Europe	M	0	12
44	Aditya Kum	IND	8.59	4.6	2.44	1.55	1.27	Asia	M	2	12

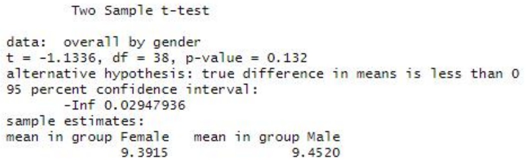
Table 1: Secondary Data of 2017 World Wushu Championships

Hypothesis Testing(2 Samples)

This was a martial arts competition for men and women, so I would like to analysis whether there is any possible the mean of females is less than males from 40 samples by using 5% significance level. And the number of males is 20 while number of females is 20 in 2017 World Wushu Championship. We assume that the variances of the two samples are equal.

Two sample test on 2017 World Wushu Championship with overall score by gender.

Let μ1 = Sample mean of female scored overall score on 2017 World Wushu Championship.
Let μ2 = Sample mean of male scored overall score on 2017 World Wushu Championship.
Hypothesis test
H0 : $\mu 1 - \mu 2 = 0$
H1: $\mu 1 < \mu 2$



```
        Two Sample t-test

data:  overall by gender
t = -1.1336, df = 38, p-value = 0.132
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf 0.02947936
sample estimates:
mean in group Female   mean in group Male
            9.3915               9.4520
```

Figure 1: R calculate for hypothesis test on two sample

This is a left-tailed test, so the rejection area of p-value is the area to the left and right of the test statistic t = -1.1336, and its p-value is 0.132. Since p-value of test statistic (0.132) is greater than p-value of significant level (0.05), it reject to null hypothesis. Thus, it is concluded that the sample mean of mean and female are not the same, other words, there is a evidence to support that mean of female scored overall scores is less than mean of male scored overall scores on 2017 World Wushu Championship.

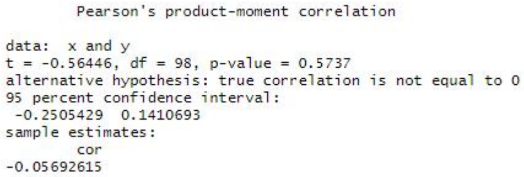
Correlation

This statistic test is to determine whether time spend affects overall score. A random sample of 100 athletes is tested for correlation. And the type of correlation coefficients is used is that PEARSON'S product-moment correlation coefficient under 0.05 significance level.

Hypothesis test

H0 : There is no linear correlation between time spend and overall score.

H1: There is a linear correlation between time spend and overall score.

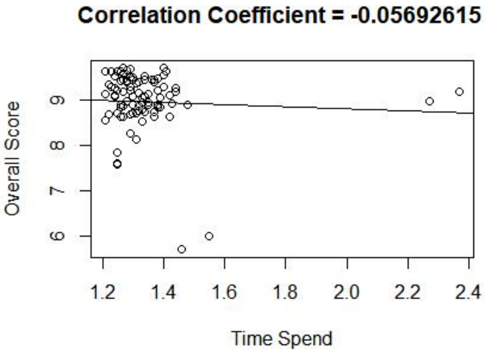


```
        Pearson's product-moment correlation

data:  x and y
t = -0.56446, df = 98, p-value = 0.5737
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2505429  0.1410693
sample estimates:
        cor
-0.05692615
```

Figure 2: R calculate for correlation test

From figure 2, it is clear to tell the significance test of the correlation of x(time spend) and y(overall score). The test statistic t = -0.56446 is greater than t critical region on 0.05 significance level which is -1.990 . Thus it fail to reject null hypothesis.



Graph 1: Scatter plot with correlation of -0.05

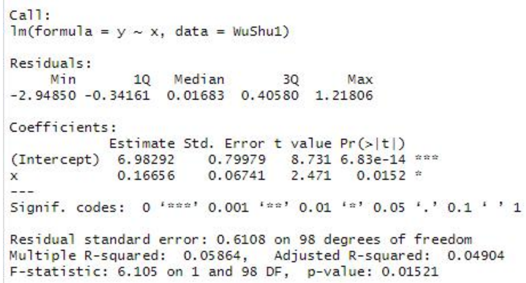
From above, we get correlation coefficient is -0.05 which means is strongly no linear relationship. Graph 1 shows that the line seems slightly going downhill but the those points are somewhat scattered in wider band, thus basically does not show much of anything happening.

According to hypothesis test which is failed to reject and scatter plot by time spend and overall score, we can make a conclusion that there is a strong evidence shows that there is no linear relationship between time spend and overall score. In other words, overall score is not affected by how much time spend.

Regression

This test to see the relationship between overall score and Nandu_Total which means the number of attempted degree of difficulty movements.. Since Nandu_Total straightway impact the C score, logically there is no doubt Nandu_Total will strongly affect overall score, therefore I would to like to determine what I think if it is true on 0.05 significance level. A random sample of 100 is used to analysis. The response variable(y) represents overall score and the predictor variable(x) represents Nandu_Total.

Hypothesis test
H0 : β1 = 0 (no linear relationship)
H1: β1 ≠ 0 (linear relationship does exit)



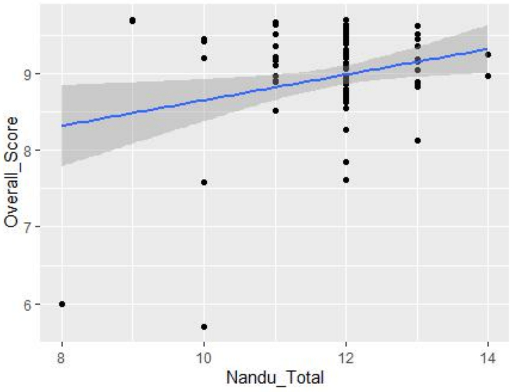
```
Call:
lm(formula = y ~ x, data = WuShu1)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.94850 -0.34161  0.01683  0.40580  1.21806 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.98292    0.79979   8.731 6.83e-14 ***
x            0.16656    0.06741   2.471   0.0152 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6108 on 98 degrees of freedom
Multiple R-squared:  0.05864,	Adjusted R-squared:  0.04904 
F-statistic: 6.105 on 1 and 98 DF,  p-value: 0.01521
```

Figure 3: R calculate for Regression test

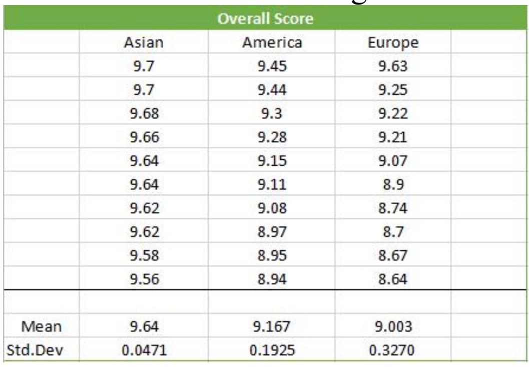


Graph 2: Plot of Nandu_Total with Overall_Score

From figure 3, the regression model shows overall score and Nandu_Total(the number of successful degree of difficulty movements) in a positive linear relationship and its line equation is written as y = 6.9829 + 0.1666*x. The indicates(b0) is 6.9829, it can be interpreted as the predicted overall score for a zero Nandu_Total score. And regression coefficient for the Nandu_Total(b1), as the regression slope is 0.1666 which means it will start to slightly affect/improve overall score. The p-value of this regression test is less than 0.05 significance level, so we reject the null hypothesis. Thus it can be concluded that there is a linear relationship between overall score and Nandu_Total. No matter from regression slope or Graph 2, both mean it does exit relationship,but it just slight relationship, it is not as strong as I predicted Nandu_Total directly impact overall score before.

ANOVA

The birthplace of martial arts is China. so I personally think the Asian region represented by China is ahead of Europe and America. There is a least one mean is different. Therefore, to determine whether the significant differences between the means of overall scores for three main regions. I would like to use ANOVA method. It is with equal sample sizes. Therefore I select top 10 samples for each region . I will use a 0.05 significance level to test the null hypothesis. Table below lists the overall scores for three different region.



	Overall Score			
	Asian	America	Europe	
	9.7	9.45	9.63	
	9.7	9.44	9.25	
	9.68	9.3	9.22	
	9.66	9.28	9.21	
	9.64	9.15	9.07	
	9.64	9.11	8.9	
	9.62	9.08	8.74	
	9.62	8.97	8.7	
	9.58	8.95	8.67	
	9.56	8.94	8.64	
Mean	9.64	9.167	9.003	
Std.Dev	0.0471	0.1925	0.3270	

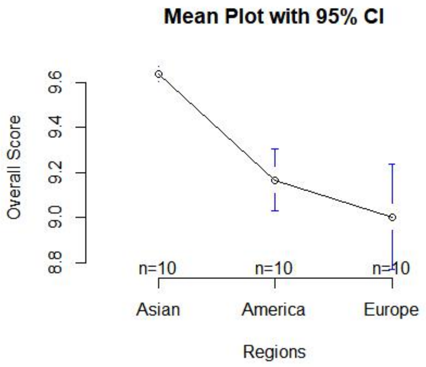
Table 2: Overall Score for 3 different regions

Let µ1 = The mean of overall score for Asian
Let µ2 = The mean of overall score for America
Let µ3 = The mean of overall score for Europe

Hypothesis test

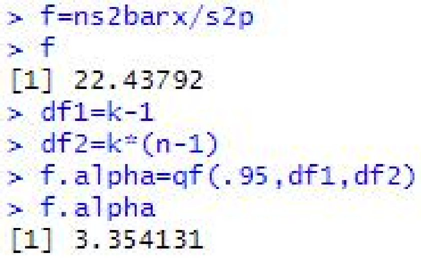
H0: µ1 = µ2 = µ3

H1: At least one of the mean is different



Graph 3: Mean Plot show the mean of overall score for three regions

And the mean plot is a visual representation of what it shows in the compare means outputs. There are three points on the graph are the mean of each region. From this chart, we can see the Asian is with the highest mean value meanwhile has the slowest mean sprint time. But even though Europe is with lowest, still has fastest mean sprint time. From here, we can see there is different for at least one of mean of the regions.



```
> f=ns2barx/s2p
> f
[1] 22.43792
> df1=k-1
> df2=k*(n-1)
> f.alpha=qf(.95,df1,df2)
> f.alpha
[1] 3.354131
```

Figure 3: R calculate for ANOVA test

And since F0 = 22.43792 > F critical value = 3.354131, we reject the null hypothesis. Therefor, I would like to conclude that combine the above mean sprint time and the F distribution there is a significantly evidence that a least one of mean of regions is different on 0.05 significance level. Meanwhile, it might show there is a strong relationship between overall score and regions.

IV. CONCLUSTION

From all the statistical test I have done, there are some short conclusions can be declared. First of all, the overall performance of athletes will be divided into men and women. From the above analysis, male athletes are more dominant in the competition. Second of all, however, the performance time of athletes does not have an impact. It may be possible to make standard actions in a short period of time without mistakes to score higher than trying to perform difficult actions for long time. And The number of attempted degree of difficulty movements. which is one of part of overall score it shows that does not affect overall score as much as everyone thinks. Which means even though some athletes do not get high scores on it,it does not mean they will lose until end, they should be focus on other parts of the whole performance. And the last, it can be seen that there is an irresistible factor -- region, which also affects the performance of athletes.

V. Appendix

1. World Wushu Championships 2017 Chinese martial Arts competition data
https://www.kaggle.com/predact/world-wushu-championships-2017