Effective Access to Higher Decathlon Scores Based on Statistical Analysis

Name: Minjie Xu

Number: c1898688

Date: 2020.2.13

Abstract

The basic decathlon data set records heights, times, distances and scores for decathletes from 1986 to 2006 and then subjected to statistical analysis. Firstly, the analysis focused on descriptive statistics using line charts, sample means as well as standard deviations. The second part of the analysis concentrated on the influence of scoring rules on scores and training points of focus using normality test, correlation analysis, regression analysis and analysis of variance. The conclusion is that high scores are related to an effective training plan focused on one function and game mentality.

Content

1.Introduction

Decathlon is the event in athletics with high requests of human functions, which combines techniques, physical ability, intellectuals into one. The winner of the competition should get the most points after all ten events. The paper is to find ways to improve the total point of an athlete even to win the game using statistical methods.
The analysis consists of two parts. Firstly, the descriptive statistics method is used to find features and identify research directions. Secondly, the influence of scoring rules on scores and training points of focus are discussed using methods of inferential statistics.

2.background

A Decathlon is a combined event in athletics where an athlete's performance in ten track-field events over two days: 400 m race, long jump, high jump, 100 m race, shot-put for day 1; 1500m race, 110 m hurdles, pole vault, javelin, discus for day 2. The winner of the competition should get the highest score. The dataset of "Decathlon" is the performance of elite decathletes over the period from 1986 to 2006 with 7986 observations on 24 variables.

Statistics methods:
1) "Descriptive Statistics" is concerned mainly with collecting, summarizing and interpreting data.
2) "Inferential Statistics" is concerned with methods for obtaining and analyzing data to make inferences applicable in a broader context. It is also concerned with the precision and reliability of such inferences in so far as this involves probabilistic considerations. These methods, including the normality test, correlation analysis, regression analysis, the analysis of variance are in this paper.

Software methods:
1) Python
Stats models is a library for statistical and econometric analysis in Python. Currently, generalized least squares, weighted least squares, and ordinary least squares of regression are included in the main codebase as one of the statistical models (Seabold S, Perktold J,2010).

2) SPSS

3.Descriptive analysis

In this part, the description of the data set of Decathlon will be given.
Table 1 shows the top ten countries with the highest average points. It can be seen that athletes of American countries and European countries get better results than other countries. Furthermore, table 2 shows that the former is better at races of speed than the latter.

Table 1. Top ten countries with average score for each event

Nationality	Totalpoints	P100m	Plj	Psp	Phj	P400m	P110h	Pdt	Ppv	Pjt	P1500	Country
JAM	7707.06	890.65	826.00	765.65	810.82	846.12	898.06	777.94	588.71	652.41	650.82	AM
BAR	7678.00	931.00	938.00	619.00	874.67	867.33	937.33	539.67	499.67	733.33	738.00	AM
UZB	7668.21	810.54	839.14	704.68	806.14	801.29	831.79	694.61	817.79	685.14	677.21	AS
ICE	7661.88	846.62	859.38	751.62	777.00	814.88	850.75	723.25	763.25	655.00	620.12	EU
DDR	7657.58	810.65	847.60	726.50	768.86	812.36	825.12	715.96	764.76	689.56	696.25	EU
CZE	7617.77	811.35	849.60	702.40	779.87	803.00	849.43	665.15	787.31	675.17	694.57	EU
TUR	7589.71	850.00	879.14	706.86	757.00	800.57	874.86	678.29	793.57	577.57	671.86	EU
UKR	7587.04	804.41	836.47	725.22	776.59	786.00	842.15	710.67	811.90	640.36	653.31	EU
PRI	7577.67	882.33	828.67	745.00	708.33	853.33	888.67	679.67	674.00	781.33	536.33	AM
MDA	7562.00	795.00	900.00	700.00	731.00	756.00	759.00	733.00	702.00	704.00	782.00	EU

- "AM" in the "Country" column means American countries
- "EU" in the "Country" column represents European countries
- "AS" in the "Country" column means Asian Countries
- Red frame represents speed races

Table 2. Average score of speed races of American countries and European countries

Country	P100m	P110h	P400m
AM	901.33	908.02	855.59
AS	810.54	831.79	801.29

Table 3 shows some statistics of the main variables in the data set.
The standard deviation reflects the degree of dispersion. If for instance, the discus has a more significant standard deviation than putting the shot, then the athlete who excels at the former would gain a better total score advantage over the other competitors who excel at the latter. Table 3 shows that the top five events with the most significant standard deviation are the pole vault>1500m>javelin>discus>high jump, indicating that these events can widen the score gap. Fortunately, these events are played on the

last day so that everyone should be at a similar level after the first day's game. Get ready for the next day's game if it does not work well on the first day. A right mentality is also an essential key to getting high scores.

Table 3. Descriptive statistics of points

	Totalpoints	P100m	Plj	Psp	Phj	P400m	P110h	Pdt	Ppv	Pjt	P1500
count	7968.000000	7968.000000	7968.000000	7968.000000	7968.000000	7968.000000	7968.000000	7968.000000	7968.000000	7968.000000	7968.000000
mean	7349.110316	806.584463	807.418675	674.645708	748.954568	788.034890	821.596762	657.763931	734.356175	648.131652	661.678087
std	413.479141	61.131021	75.325331	75.659246	76.463698	63.809945	66.553049	86.039938	106.434221	89.103168	89.923320
min	6800.000000	597.000000	569.000000	418.000000	396.000000	532.000000	526.000000	231.000000	321.000000	330.000000	121.000000
25%	7020.000000	765.000000	757.000000	622.000000	696.000000	746.000000	777.000000	600.000000	673.000000	588.000000	607.000000
50%	7255.000000	806.000000	804.000000	672.000000	749.000000	789.000000	824.000000	655.000000	731.000000	644.000000	669.000000
75%	7608.000000	847.000000	857.000000	726.000000	803.000000	832.000000	868.000000	713.000000	804.000000	707.000000	725.000000
max	9026.000000	1042.000000	1089.000000	941.000000	1061.000000	998.000000	1044.000000	970.000000	1152.000000	1040.000000	946.000000

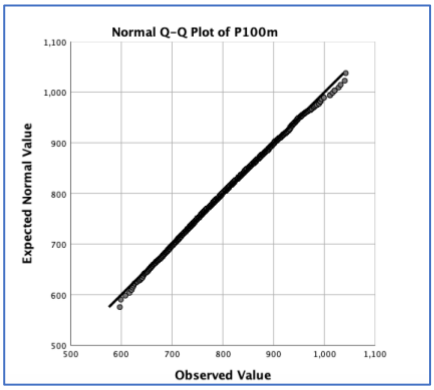
4. The distribution of results

The distribution of the results in the data set should be analyzed first.

It is assumed that the results of the competition in the data set should follow the normal distribution. The situation for competitors to complete decathlon competition is complicated. Many reasons affect the result of the competition. For example, weather and other environmental factors are difficult to be analyzed on how to affect the achievements of athletes. In addition, mental or physical status may affect the results. According to the Central Limit Theory, despite these varieties of effect factors mentioned above which are independent, especially when the sample size is quite more significant, we still can use the normal distribution model theoretically.

Q-Q plots are used to test whether the results of events have approximately the same distribution as a normal distribution. From SPSS, distributions of scores of ten events follow the normal distribution according to Q-Q plots. Fig.1 shows the Q-Q plot of points of 100m, and more details are shown in Appendices[a].

Fig.1 Q-Q plot of points of 100m



5.Regression analysis of effort and scores

The fairness of Decathlon is essential, which means it has the formula to convert times, distances and heights into points reasonably. Regression analysis is used to quantify the influence of times or distance or heights on points of each event and fit a formula.
The sample correlation coefficient quantifies the linear dependence between X and Y. Then we can determine that a simple linear regression equation can be established:

$$\boxed{Y=\alpha+\beta X+\varnothing \quad \text{with} \quad \varnothing \sim N(0,\sigma^2) \quad \text{where} \quad \alpha,\beta \in R \quad \text{and} \quad \sigma^2 \in R}$$

For example, it can also be seen from Fig.2 that 'Discus'(Distance throwing the discus) and 'Pdt'(Points for performance in discus) are linearly related.

Fig.2 Regression analysis for 'Discus' and 'Pdt'

```
for i in range(10):
    X=data[[action[i]]]
    y=data[[action[i+10]]]
    linreg = LinearRegression()
    model=linreg.fit(X, y)
    print(model.score(X,y))
```

```
0.9993924618892405
0.9996363865677682
0.9999717193353348
0.9994962215960801
0.9992780924014488
0.9991289132685813
0.9999121416601924
0.9991772100545665
0.9999386053926203
0.9963812466164692
```

1500m and P1500

```
plt.scatter(data[['Discus']],data[['Pdt']])
```

```
<matplotlib.collections.PathCollection at 0x1a189ae2e8>
```

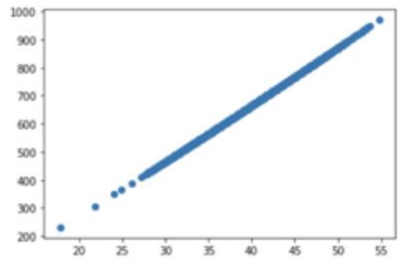


Table 4 shows the results of the ordinary least squares (OLS) regression for these two variables. As we can see, t values of parameters α and β are 2110.637 and 9521.592, respectively. The null hypothesis that the parameter is zero is rejected. For the simple linear model $Y=\alpha+\beta X$, OLS estimators of the model parameters α and β are 7.2427 and 0.0493, respectively. F value=9.066e+07(p-value<0.05) and R-squared=1 indicate that $Y=7.2427+0.0493X$ fits the data perfectly.

Table 4. OLS Regression Results for 'Discus' and 'Pdt'

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Discus   R-squared:                       1.000
Model:                            OLS   Adj. R-squared:                  1.000
Method:                 Least Squares   F-statistic:                 9.066e+07
Date:                Sat, 04 Jan 2020   Prob (F-statistic):               0.00
Time:                        17:30:45   Log-Likelihood:                 14398.
No. Observations:                7968   AIC:                        -2.879e+04
Df Residuals:                    7966   BIC:                        -2.878e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      7.2427      0.003   2110.637      0.000       7.236       7.249
Pdt            0.0493   5.17e-06   9521.592      0.000       0.049       0.049
==============================================================================
Omnibus:                     6218.891   Durbin-Watson:                   1.853
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           382724.955
Skew:                          -3.237   Prob(JB):                         0.00
Kurtosis:                      36.330   Cond. No.                     5.11e+03
==============================================================================
```

Other events also get the same conclusion except 1500m. As we can see, t values of parameters α and β are 5266.108 and -1480.994, respectively. The null hypothesis that the parameter is zero is rejected. For the simple linear model Y=α+βX, OLS estimators of the model parameters α and β are 7.2427 and -0.1655, respectively. F values =2.193e+06 and the p-value is 0.0. indicates that Y=392.9-0.1655X fits the data. However, R-squared=0.996 shows that the model fits worse than other nine events.

Table 5. OLS Regression Results for '1500m' and 'P1500'

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  m1500   R-squared:                       0.996
Model:                            OLS   Adj. R-squared:                  0.996
Method:                 Least Squares   F-statistic:                 2.193e+06
Date:                Sat, 22 Feb 2020   Prob (F-statistic):               0.00
Time:                        18:27:06   Log-Likelihood:                -10438.
No. Observations:                7968   AIC:                         2.088e+04
Df Residuals:                    7966   BIC:                         2.089e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    392.9659      0.075   5266.108      0.000     392.820     393.112
P1500         -0.1655      0.000  -1480.994      0.000      -0.166      -0.165
==============================================================================
Omnibus:                    11708.026   Durbin-Watson:                   2.001
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          10724341.790
Skew:                           8.543   Prob(JB):                         0.00
Kurtosis:                     181.914   Cond. No.                     4.96e+03
==============================================================================
```

It is easy to see that the linear formula is used to convert times, distances and heights into points simply. However, performance is not linear. That is to say, the extra effort needed in throwing the discus to 60m compared with 50m will be higher than that needed in improving the distance from 40m to 50m, the former of which is more unworthy (Cox Dunn,2002). Thus, Athletes should set goals taking the reward of effort into account.

To determine a proper scope of goal for athletes should consider the quartiles of each event. Table 6 describes data of 10 events as reference. For example, if the athlete would like to get a high score of 100m, the lower quartile(11.06s) should be the goal.

Table 6. Descriptive statistics of heights, times, distances

	m100	Longjump	Shotput	Highjump	m400	m110hurdles	Discus	Polevault	Javelin	m1500
count	7968.000000	7968.000000	7968.00000	7968.000000	7968.000000	7968.000000	7968.000000	7968.000000	7968.000000	7968.000000
mean	11.251177	6.969378	13.11821	1.938468	50.606242	15.243671	39.640157	4.405147	53.978697	283.458074
std	0.282724	0.317433	1.24014	0.085385	1.411713	0.562964	4.237995	0.367232	5.978417	14.909321
min	10.220000	5.920000	8.87000	1.510000	46.210000	13.470000	17.780000	2.850000	32.200000	241.000000
25%	11.060000	6.760000	12.25000	1.880000	49.630000	14.850000	36.800000	4.200000	49.960000	273.020000
50%	11.250000	6.960000	13.08000	1.940000	50.550000	15.210000	39.540000	4.400000	53.705000	281.825000
75%	11.440000	7.180000	13.96000	2.000000	51.510000	15.610000	42.380000	4.650000	57.932500	291.902500
max	12.280000	8.110000	17.45000	2.270000	56.700000	17.980000	54.780000	5.760000	79.800000	401.180000

6.The relationship of average total points and event scores

In this part, it will be discussed whether different events make the same contribution to total points. To compare the differences in event scores in the data set, this part uses t-distribution to calculate the confidence intervals for means.

It has already been discussed why the results of events follow the normal distributions. Therefore, all the discussion in this referential part of the analysis is based on the normal distribution. Variables and distribution could be shown as the following:

H0: the score of this event is equal the average of total points

H1: the score of this event is not equal to the average of total points

Firstly, after handling the data set with python, fig.3 shows that the average of total points is calculated to 734.91.

Fig.3 The average of total points for ten events

```
In [43]:  1  pd.DataFrame.mean(data["Totalpoints"])/10
Out[43]: 734.9110316265061
```

Then, the hypothesis test results have been shown in table 7. The table shows that the means of events all have a statistically significant difference from the average of total points except "Ppv”. We can conclude that all events make different contributions to total points.

Table 7. T-test result for events and total points

One-Sample Test

	Test Value = 734.91					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
P100m	104.659	7967	.000	71.674	70.33	73.02
Plj	85.926	7967	.000	72.509	70.85	74.16
Psp	-71.101	7967	.000	-60.264	-61.93	-58.60
Phj	16.396	7967	.000	14.045	12.37	15.72
P400m	74.316	7967	.000	53.125	51.72	54.53
P110h	116.268	7967	.000	86.687	85.23	88.15
Ppv	-.464	7967	.642	-.554	-2.89	1.78
Pdt	-80.037	7967	.000	-77.146	-79.04	-75.26
Pjt	-86.935	7967	.000	-86.778	-88.74	-84.82
P1500	-72.695	7967	.000	-73.232	-75.21	-71.26

7.The focus of the training plan

The training plan for athletes should focus on improvement to one event or function making most contributions. To find which function should be focused on to be an improvement, the correlation coefficient is computed to reflect the influences of 10 events on each other and total points are essential to be considered.
As we conclude, all events are based on normal distributions and fit the simple linear model. In Table 8, the null hypothesis that ten events are not correlated with total points is rejected. More details on the correlations of ten events are shown in Appendices[b]. We can conclude that the Pearson correlation coefficient can be referred to directly.

Table 8. The correlation coefficient of event scores and total points

Correlations

		Totalpoints	P100m	Plj	Psp	Phj	P400m	P110h	Ppv	Pdt	Pjt	P1500
Totalpoints	Pearson Correlation	1	.517**	.624**	.615**	.454**	.515**	.608**	.584**	.609**	.509**	.227**
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	Sum of Squares and Cross-products	1362078158	104152737.3	154916392.0	153202115.4	114365574.9	108238643.3	133261445.4	204808468.9	172558816.5	149478915.3	67112451.96
	Covariance	170965.000	13073.018	19444.759	19229.586	14354.911	13585.872	16726.678	25707.100	21659.196	18762.259	8423.805
	N	7968	7968	7968	7968	7968	7968	7968	7968	7968	7968	7968

Table 9 is the Pearson correlation coefficient of all points. It gives the conclusion that points of 400 meters and 100 meters have a strong correlation. This is likely because all four items have high-speed requirements. They each depend on the high anaerobic function and are based on absolute speed. Discus and putting the shot can also result in the same conclusion, and they are likely to be improved together to some extent. Moreover, points of 100 meters, 400 meters and discus make the most contribution to total points. These events can be treated as the main directions of the training plan.

Table 9. The correlation coefficient of all points

	Totalpoints	P100m	Plj	Psp	Phj	P400m	P110h	Pdt	Ppv	Pjt	P1500
Totalpoints	1.000000	0.765593	0.234636	0.439342	0.087296	0.665981	0.383139	0.584647	0.180260	0.438428	-0.317341
P100m	0.765593	1.000000	0.320082	0.339730	-0.304733	0.697053	0.559973	0.299977	0.140507	0.129084	-0.210875
Plj	0.234636	0.320082	1.000000	-0.001599	-0.341163	-0.042266	0.474526	0.043180	0.310487	-0.155586	-0.349326
Psp	0.439342	0.339730	-0.001599	1.000000	0.059114	0.315513	0.302262	0.948366	-0.302262	-0.135481	-0.632625
Phj	0.087296	-0.304733	-0.341163	0.059114	1.000000	-0.105167	0.125101	0.239204	-0.497195	-0.015580	-0.311377
P400m	0.665981	0.697053	-0.042266	0.315513	-0.105167	1.000000	0.172493	0.344300	-0.246973	0.591958	0.094357
P110h	0.383139	0.559973	0.474526	0.302262	0.125101	0.172493	1.000000	0.296779	-0.269820	-0.486863	-0.644048
Pdt	0.584647	0.299977	0.043180	0.948366	0.239204	0.344300	0.296779	1.000000	-0.278044	0.016678	-0.659127
Ppv	0.180260	0.140507	0.310487	-0.302262	-0.497195	-0.246973	-0.269820	-0.278044	1.000000	0.174320	0.147818
Pjt	0.438428	0.129084	-0.155586	-0.135481	-0.015580	0.591958	-0.486863	0.016678	0.174320	1.000000	0.543523
P1500	-0.317341	-0.210875	-0.349326	-0.632625	-0.311377	0.094357	-0.644048	-0.659127	0.147818	0.543523	1.000000

8.Conclusion

Getting high scores depends on an effective training plan and game mentality.
Athletes' physical fitness is different from each other, which means the training plan must be kept with one's aptitude and talents. Besides, unfairness caused by simple linear scoring rules should be considered when setting goals. Extra efforts to raise scores slowly is unnecessary. In order to improve score more quickly, training should focus on events that can affect the total score significantly, such as 100 meters race and 400 meters race.
For athletes, mentality in games is also the key to win the game. Luckily, the events that can open the gap between the total scores are on the last day of the game. It is better to adjust the mentality to deal with the game on the last day of events on the first day did not go well.

9.Reference

Cox Dunn, T. (2002). An analysis of decathlon data. Journal of the Royal Statistical Society: Series D (The Statistician),21(2), pp. 179-187.

D'Agostino, R., Belanger, A. and D'Agostino, R. (1990). A Suggestion for Using Powerful and Informative Tests of Normality. The American Statistician, 44(4), p.316.
Seabold S, Perktold J: Statsmodels. (2010). Econometric and statistical modelling with python. Proceedings of the 9th Python in Science Conference.

10.Appendices

a. The Q-Q plots of all events:

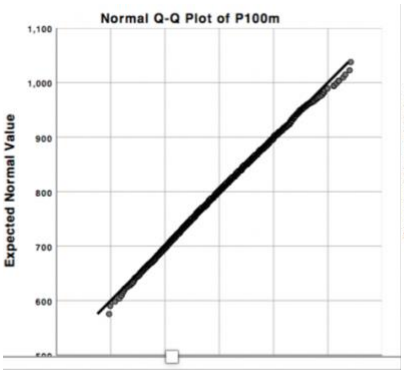
Estimated Distribution Parameters

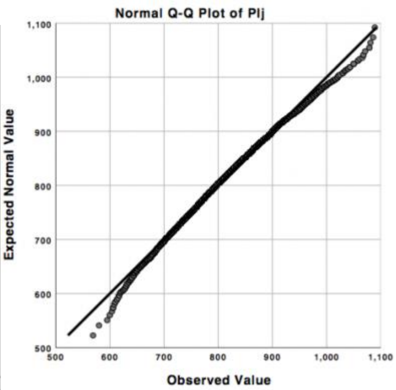
		P100m	Plj	Psp	Phj	P400m	P110h
Normal Distribution	Location	806.58	807.42	674.65	748.95	788.03	821.60
	Scale	61.131	75.325	75.659	76.464	63.810	66.553

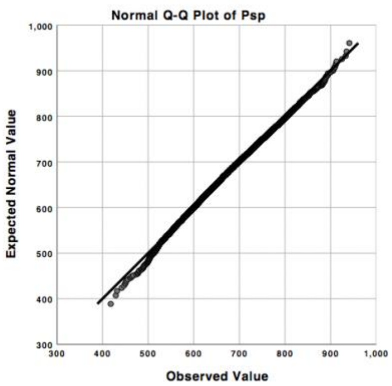
Estimated Distribution Parameters

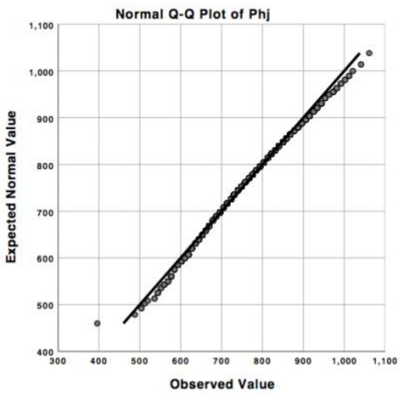
		Ppv	Pdt	Pjt	P1500
Normal Distribution	Location	734.36	657.76	648.13	661.68
	Scale	106.434	86.040	89.103	89.923

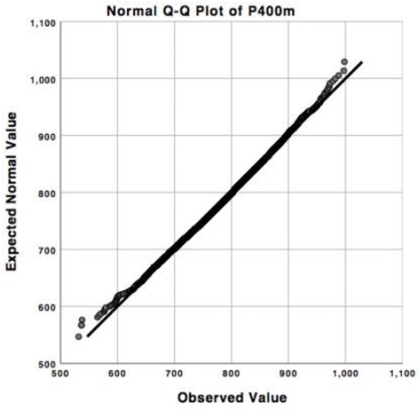
The cases are unweighted.

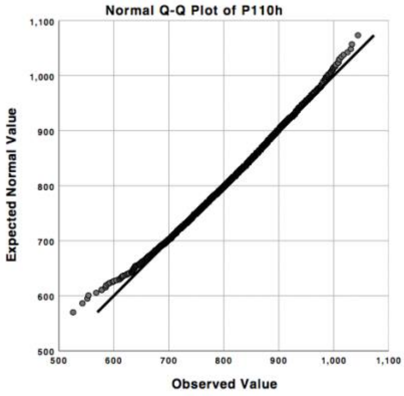


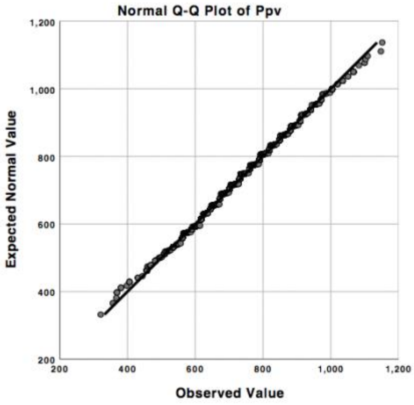


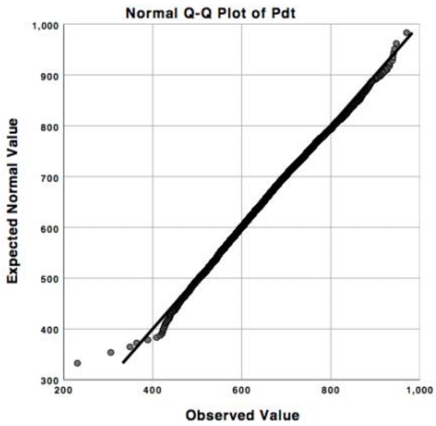


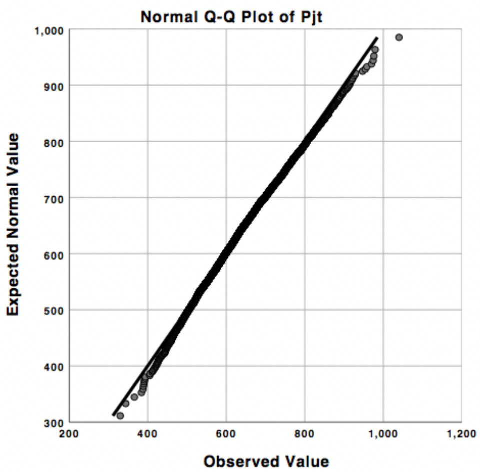


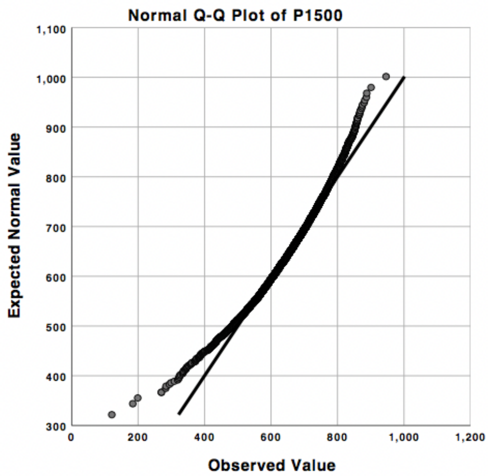












b. The correlation table of all events

Correlations

		Totalpoints	P100m	Plj	Psp	Phj
Totalpoints	Pearson Correlation	1	.517**	.624**	.615**	.454**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	7968	7968	7968	7968	7968
P100m	Pearson Correlation	.517**	1	.487**	.158**	.126**
	Sig. (2-tailed)	.000		.000	.000	.000
	N	7968	7968	7968	7968	7968
Plj	Pearson Correlation	.624**	.487**	1	.253**	.362**
	Sig. (2-tailed)	.000	.000		.000	.000
	N	7968	7968	7968	7968	7968
Psp	Pearson Correlation	.615**	.158**	.253**	1	.158**
	Sig. (2-tailed)	.000	.000	.000		.000
	N	7968	7968	7968	7968	7968
Phj	Pearson Correlation	.454**	.126**	.362**	.158**	1
	Sig. (2-tailed)	.000	.000	.000	.000	
	N	7968	7968	7968	7968	7968
P400m	Pearson Correlation	.515**	.575**	.312**	.036**	.109**
	Sig. (2-tailed)	.000	.000	.000	.001	.000
	N	7968	7968	7968	7968	7968
P110h	Pearson Correlation	.608**	.456**	.388**	.259**	.258**
	Sig. (2-tailed)	.000	.000	.000	.000	.000
	N	7968	7968	7968	7968	7968
Ppv	Pearson Correlation	.584**	.174**	.273**	.254**	.192**
	Sig. (2-tailed)	.000	.000	.000	.000	.000
	N	7968	7968	7968	7968	7968

Correlations

		P400m	P110h	Ppv	Pdt	Pjt
Totalpoints	Pearson Correlation	.515**	.608**	.584**	.609**	.509**
	Sig. (2-tailed)	.000	.000	.000	.000	.000
	N	7968	7968	7968	7968	7968
P100m	Pearson Correlation	.575**	.456**	.174**	.125**	.065**
	Sig. (2-tailed)	.000	.000	.000	.000	.000
	N	7968	7968	7968	7968	7968
Plj	Pearson Correlation	.312**	.388**	.273**	.200**	.177**
	Sig. (2-tailed)	.000	.000	.000	.000	.000
	N	7968	7968	7968	7968	7968
Psp	Pearson Correlation	.036**	.259**	.254**	.719**	.438**
	Sig. (2-tailed)	.001	.000	.000	.000	.000
	N	7968	7968	7968	7968	7968
Phj	Pearson Correlation	.109**	.258**	.192**	.146**	.070**
	Sig. (2-tailed)	.000	.000	.000	.000	.000
	N	7968	7968	7968	7968	7968
P400m	Pearson Correlation	1	.384**	.133**	.039**	.025*
	Sig. (2-tailed)		.000	.000	.000	.026
	N	7968	7968	7968	7968	7968
P110h	Pearson Correlation	.384**	1	.292**	.231**	.137**
	Sig. (2-tailed)	.000		.000	.000	.000
	N	7968	7968	7968	7968	7968
Ppv	Pearson Correlation	.133**	.292**	1	.273**	.196**
	Sig. (2-tailed)	.000	.000		.000	.000
	N	7968	7968	7968	7968	7968