Self Marking Exercise

Fill this part before printing

Data File number:	**Dream Team data**
Name:	**Nuo**
Surname:	**Chen**
Student number:	**17011039**

Complete this part at lecture during student discussion

Self marking **Mark yourself and discuss your marking within the group**	Put marks between 0-20
Is the abstract/introduction/summary clearly reporting the essential points?	
Is the methodology explained well?	
Are results and findings clear, correct and significant?	
Have figures and tables comprehensible captions and labels?	
Are conclusions reporting the main findings and highlighting the open questions?	
Total	

Rank your work with respect to the other reports

Top 5%	
Top 10%	
Top 25%	
Top 50%	
Bottom 50%	

Have you learned something new preparing this assignment? ☐ **No** ☐ **Yes**

Comments:

Leave this part unfilled

Final mark by academic examiner: _____

Data Analytics Coursework 1

Name:Nuo Chen
ID:17011039

Section A: Data Description

The Data Dream is an e-sports recruiting platform. It focuses on helping users to quickly find their matching players and teams, increasing their chances to become professional players.

In this project, we have their data captured from 2017-11-08 to 2018-01-05. Their main variables can be divided into three parts, data about players, data about teams and others.

1. Player
   Each player has a profile, recording their general information, such as id, gender, nationality, language, etc. There are different game roles for users to play with. Players' game performance is recorded by maps and by weapons. Their skills level are evaluated against their total number of kills, kill-deaths, kill-headshot, hit, round played, and time played etc. They are the core statistics to analyze a player's skill. Players can follow each other and leave reviews.

2. Team
   Each team also has a profile, containing general information, such as id, country, language, social links, etc. Teams are ranked according to the members' performance. Players can join a team by receiving/requesting invitations. Team vacancies are public with requirements about the players' skill levels and ages. The best map played for each team is calculated.

3. Other
   There is other information, for example, the kind of hardware used (CPU, server, headset, etc), the goal of each player/team, profile ratings.

Section B: Fitting Distribution

In this part, we have analyzed three variables concerning the players' activities. They are total kill-headshot, total time played and the total deaths.

a. Total kill-headshot

1. Data Cleaning
   We regard a piece of data as a cheating record if it matches one of the following conditions:

   (a) Null value
   Null value has provides no information, so remove them.

(b) More than 9 kills per round
CS: GO is a competitive game with two teams. Each team has 5 people in the competitive mode and 10 people in the casual mode. Both modes play in rounds. A player will be kicked out of the game if he keeps shooting his teammates. Thus, the average of kill-headshot per round should range from 0 to 9 (one cannot kill himself).

(c) Kill-headshot is larger than total-kill
By definition, a kill is a shot resulting in a death. Kill-headshot is a kill in the head. So the total-kill-headshot can't be larger than the total-kill.

(d) More than 4,000,000 headshots
4,000,000 headshots is an average of 5 kills in a 2-minute game and play 12 hours every day for six years. We believe data larger than this are considered as cheaters.

After cleaning the data, there are 71967 records left which are around 85.7% of the original data. The conjugate cumulative distribution function of the sample data in log-log scale is in Figure 1. We can see that the irregular tail has been removed, leaving a rather linear-look-tail.

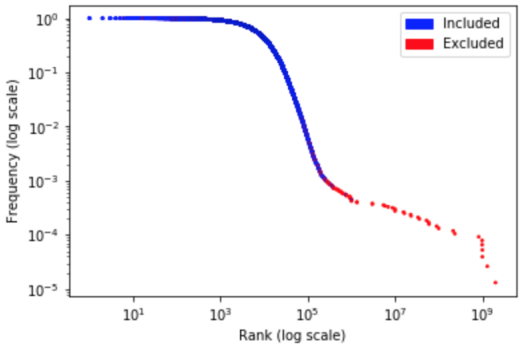

Figure 1: Conjugate cumulative distribution function of total kill-headshot. The blue dots are data we are going to analyze. The red dots are considered as fake data.

2. Explore Data

The first four moments are computed following the equation in the Figure 2. They are 13,230(mean), 280,851,486(variance), 7.8(skew), 255.8(kurtosis). The skew and kurtosis are much larger than the normal distribution. This suggests that the data distribution has a very long right tail. We can confirm this by visualizing its histogram using 100 bins (Figure 3). From the plot, it's clear to see that the data has a wide range, from one

to over one million. There are few observations larger than 100 thousand. The majority of observations are below 50 thousand. This suggests that we should fit the distribution into two parts. As the value of total-kill-headshot has a large span, it's better to visualize data by using its rank/frequency plot as in Figure 4.

$$\gamma_k = \frac{\mu_k}{\sigma^k}$$

Figure 2: Stardardized moments. When the k=3, it is called the skewness. When the k=4, it is called the kurtosis.

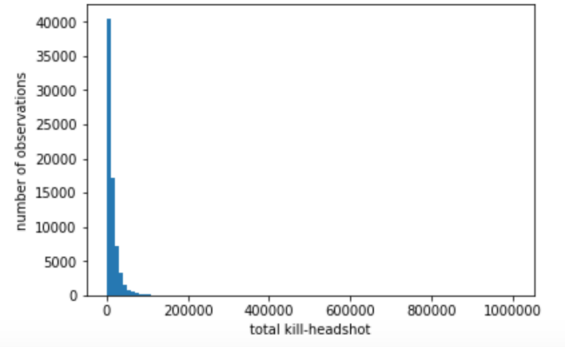

Figure 3: The distribution of total-kill-headshot, without normalization.

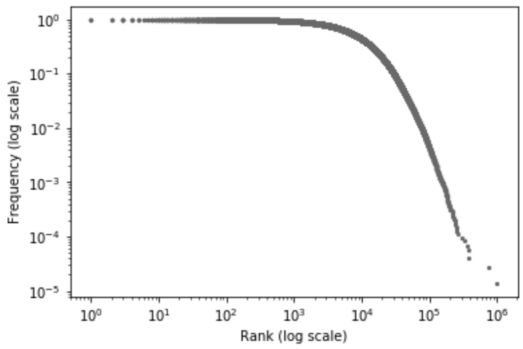

Figure 4: The rank/frequency plot of total-kill-headshot in log-log scale.

3. Fit Distribution

(a) Set clipping point
The clipping point is the point that fits the body as well as fits the tail distribution. Finding the clipping point is tricky as it based on the fitted distribution. To begin with, we select this point by visualizing the Figure 3. The initial guess is around 1500. Thus, we fit below 1500 and over 1500 with two different distributions.

(b) Fit body
We use the 'fit' function in the 'scipy' package. In basic, it tries to find the best distribution parameters that minimize the negative log-likelihood. We have tried to fit against common distributions, such as Beta, Gamma, Weibull, etc. The exponential distribution has the minimal negative log-likelihood. This means this distribution fits the most number of points in the sample. So we decide to fit the body with it.

(c) Fit tail
Similarly, we use the same technique to fit the tail as in fitting the body. The tail is roughly linearly distributed, regardless of few points at the bottom. So we fit it with a pow-law distribution.

(d) Adjust clipping point and re-fit
Once we settle down the two distributions, we try to adjust the position of the clipping point by minimizing the negative log-likelihood of the body distribution. The point at 1495 best fits the body distribution, so we settle the clipping point at this position. We didn't take the tail distribution into account when adjusting the clipping point. First, it's easy to manipulate. Second, the majority of the data are in the body part. So we would like the body be fitted as well as possible. The integrated fitting can be checked in Figure 5.

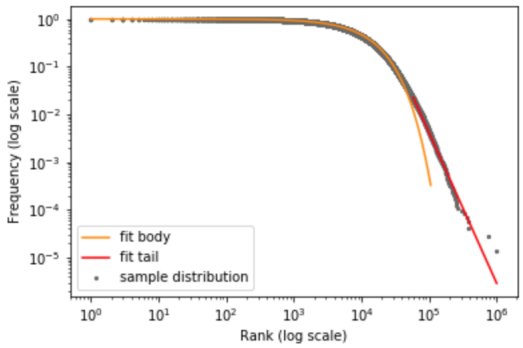

Figure 5: The plot of fitted distribution in rank/frequency plot.

(e) Test and Evaluate

i. P-TEST
By definition, it looks for the probability of observing the extreme value. In our case, the value is extremely small, around 0. This strongly rejects the null hypothesis.

ii. KS-Test
The Kolmogorov–Smirnov test follows equation in Figure 6. The 'D' statistics gives the maximal difference between the fitted distribution and sample distribution. As KS-Test is tolerant, we only use this test in evaluating the fit of the body. In our case, the D is 0.006102, which is small.

$$\hat{F}_N(x) = \frac{1}{N+1}\sum_{i=1}^{N} I_{X_i \le x}$$

$$D_N = \sup_x \left| \hat{F}_N(x) - F(x) \right|$$

Figure 6: KS-Test equation

4. Summary
The total-kill-headshot has been fitted with exponential in the body and power-law in the tail. It suggests that most of the players have kill-headshot less than 60 million, only a few players can have kill-headshot between 60 million and 100 million. The exponential body distribution suggests that the more the kill-headshot, the less number of players. Thus, it is a good way to measure how good a player is. In the tail part, the power-law distribution suggests there are always players who have extremely high kill-headshot.

II. Total time played

As we have discussed in detail how to fit distribution in part a, I would like to only present result in the following sections. This follows the same procedure as discussed in analyzing the 'total-kill-headshot'.

1. Data Cleaning
We regard the data as a fake record if it meets one of the following requirements:

   (a) Null value

   (b) Total-time-played larger than 169,862,400 seconds.
   This is the total seconds that the game has existed to the public. Any player who has played the game longer than this will be treated as cheaters. However, this cutting approach may be too generous. This could be further investigated by combining other data, such as

their number of played rounds and so on. Because of the time limit, we didn't dig this further.

2. Explore Data
   After cleaning the data, we plot the histogram using 100 bins and get Figure 7.
   The first four moments are computed as 203,0211(mean), 7,242,632,689,787 (variance), 21(skew), 903(kurtosis). We can see that the majority of players have played less than 10 million seconds, about 2778 hours. There are long-time players. Each count near to 0. To analyze all the data in one plot, we plot the rank-frequency plot in log-log scale as in Figure 8. The tail looks rather irregular, neither linear-look nor follow the normal distribution. I think this may due to the potential cheaters in our data.

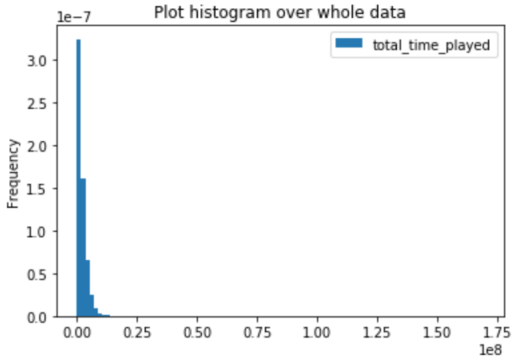

Figure 7: Histogram of total-time-played distribution using 100 bins.

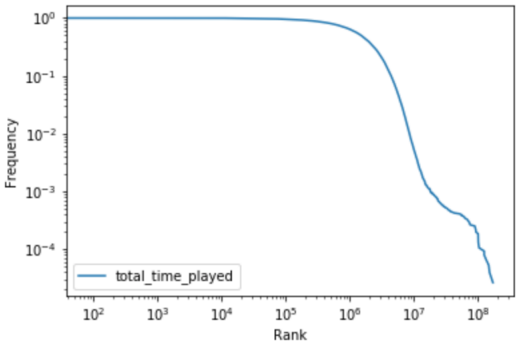

Figure 8: Rank-frequency plot of total-time-played in log-log scale.

3. Fit Distribution

(a) Set clipping point
We set the initial clipping point at 10 million.

(b) Fitting body and tail
Same as in part a, we use the 'fit' function in the 'scipy' package which uses the MLE method to find parameters of a given distribution. Iteratively this method against common distributions, such as beta, gamma, etc. We find the 'exponential' best fits the body with the minimum negative log-likelihood. Similarly, we find 'chi-square' best fits the tail.

(c) Adjust clipping point and re-fit
We compare the negative log-likelihood while moving the clipping point around. Finally, we find the clipping point at 13,380,830 works best. Then we fit again using this clipping point. The integrated fitting can be checked in Figure 9.

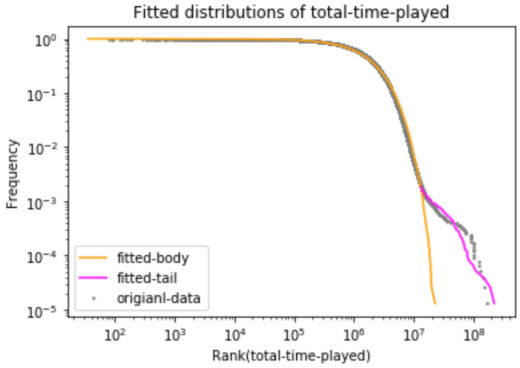

Figure 9: Fitted Distribution

4. Test and Evaluate

(a) P-Test
Our p-value is 0.022933, which is small and can be used to reject the null hypothesis.

(b) KS-Test
We evaluate the body fitting using the KS-test, it gives D value as 0.0229, which is acceptable.

5. Summary
The distribution of total-time-played can be fitted with exponential in the body and chi-2 in the tail. The body has the same distribution as the body of 'total-kill-headshot'. It suggests there might be a correlation between headshot skill and time-played. This makes sense, because the more you practice, the better you kill. Unlike the total-kill-headshot, the

tail of total-time-played is fitted with 'chi-2'. This is weird because the chi-square distribution is mainly used for testing. I think the reason for this is the way we define the cheating records. We only use the information of total-time-played without considering other information, for example, players who have played a considerable amount of time but have suspicious winning records.

III. Total Deaths

As we have discussed in detail how to fit distribution in part a, I would like to only present result in the following sections. Analyzing total-deaths follows the same procedure as the discussed total-kill-headshot.

1. Data Cleaning
   We regard and remove the data as a fake record if it meets one of the following requirments:
   (a) Null values
   (b) Total MVP more than total rounds played
   MVP is a title that could be obtained at most once per round.
   (c) Total-time-played larger than 1963 days
   Same reason as discussed in part a.
   (d) Has average larger than 9 kills per round
   Same reason as discussed in part a.
2. Explore Data
   We plot the histogram of total death in Figure 10. Compared to the previous two distributions, the total-death has a relatively narrow range. This is sensible as people will cheat to win not cheat to die. From the graph, the majority of players have less than 0.23 million deaths. The maximal number of total deaths is 0.6 million. The first four moments are computed as 31583(mean), 33307(standard deviation), 25(kurtosis), 3.4(skewness).

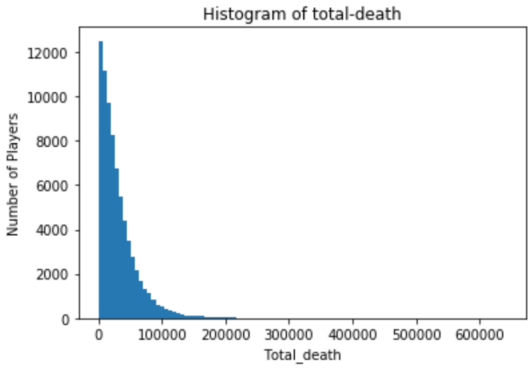

Figure 10: Histogram of total-death

3. Fit Distribution

   (a) Set clipping point
   As its skewness is similar to normal, the tail should not be too long. So we set the initial clipping point at the last 50 point.

   (b) Fit body and tail
   Same as in part a, we find the 'exponential' best fits the body with the minimum negative log-likelihood. Similarly, we find 'chi-square' best fits the tail.

   (c) Adjust clipping point and re-fit Same as in part a, we find the best clipping point and re-fitted the distribution as in Figure 11.

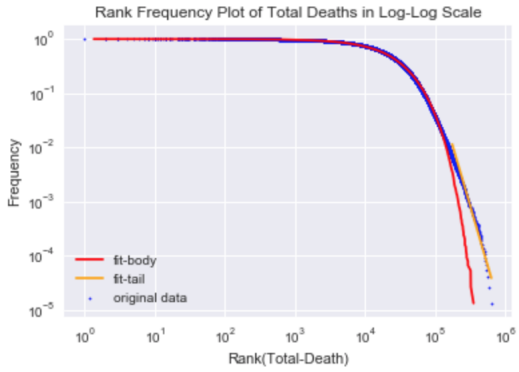

Figure 11: Rank-frequency plot with fitted distributions.

4. Test and Evaluate

(a) P-Test
Our p-value is 3.7e-35, which is very small. So we reject the null hypothesis.

(b) KS-Test
We evaluate the body fitting using the KS-test, it gives D value as 0.0230, which is acceptable.

5. Summary
The total-deaths is fitted with exponential in the body and power-law in the tail. This is the same as total-kill-headshot.They might have correlations which need to be further proved. We believe there's less cheated data in the total-death data, as people will cheat to win rather cheat to lose.