Viewership Project

Miles Lunn

May 2017

0.1 Introduction

In this project, I aim to test some of the points made in Thorin's Esports Salon video. For context, I am mathematics student and have just completed my first year at University. I have completed two modules based on Statistics, thus I aim to apply them to Esports in this project.

Chapter 1

Nationality & Stream Viewership

1.1 North America

In the video, Noah claimed that NA LCS imports average less viewers than North American players. A statistical regression line test can be formed to test whether this claim is true. Below summarizes a table of data for North American Players in the NA LCS:[1]

Player	Hours Streamed/Month	Avg. viewers	Coefficient	AVC
Contractz	38	765	1.667	1,275
Sneaky	132	13,423	1.842	24,725
Stixxay	15	401	1	401
Akaadian	31	267	1.339	357
Moon	32.5	123	1	123
Hai	21.5	1,005	1.704	1712
Dardoch	28	582	1.339	779
Pobelter	61	2,702	1	2,702
Meteos	48	4,518	1.745	7,885
Stunt	43	66	1.704	112
Lourio	112	640	1.745	1,117
Doublelift	72	12,607	1.250	15,758
WildTurtle	16	2,175	1	2,175

The following players either have streamed League of Legends for less than 10 hours within 30 days. These players will not be used for the test.

- Smoothie
- Balls
- Altec
- LemonNation
- zig
- LOD
- Keith
- Aphromoo

- Darshan
- Xpecial
- Apollo
- Hakuho
- Matt
- Hauntzer

One factor which will effect player viewer count is the time of day at which they stream. Twitch usually peaks around 9:30pm, usually holding around $1,000,000$ viewers [2]. The minimum viewer count is around $500,000$, which occurs around 10:00am. In consideration of this, I will multiply the viewer count by a coefficient for each streamer, depending on typical time of day at which they stream. For example, if somebody typically streams around 10am, there are 50% of the maximum viewer pool on Twitch at that time, thus I will multiply their viewer count by 2, to match somebody who streams during peak hours. AVC stands for adjusted viewer count, which is calculated by multiplying the original viewer count by the coefficient.

Pearson's Product Moment Correlation Coefficient can be used to tell us if there is a relation between monthly hours spent streaming, and viewer count for NA players. This is given by:

$$r = \frac{\sum xy - n\overline{x}\overline{y}}{\sqrt{[\sum x^2 - n\overline{x}^2][\sum y^2 - n\overline{y}^2]}}$$

Here are some values which are needed to solve the equation for r:

$$\sum xy = 5.2344 \times 10^6$$

$$\sum x^2 = 48214$$

$$\sum y^2 = 9.4057^8$$

$$\overline{x} = 50$$

$$\overline{y} = 4548$$

$$n = 13$$

Where $\overline{x}$ and $\overline{y}$ are the mean values of x and y respectively.

$$\Rightarrow r = 0.701$$

Since $r = 0.701$, it can be inferred that there is a strong, positive correlation between how often a player streams on Twitch, and their average viewer count. This means it would be reasonable to produce a linear regression model: an equation which will tell us how many viewers a North American player should get, when we are given how often they stream. This equation will be in the form $y = a + bx$, where a and b are constants, which are calculated using the formulae:

$$b = \frac{\sum xy - n\overline{xy}}{\sum x^2 - n\overline{x}^2}$$

$$a = \overline{y} - b\overline{x}$$

$$\Rightarrow b = 145$$

$$\Rightarrow a = -2701$$

Now we have an equation which tells us roughly how many viewers a North American streamer has, y , when given the hours they stream per month, x. This equation is: $y = 145x - 2701$. This is far from a perfect model, and it only fits specific criteria. For example, if a player streams for less than 18 hours a month, they are given a negative number of viewers. The sample size, n, was very small, and so extreme results have a large impact on the line equation which has been derived. However, this problem cannot easily be resolved due to the nature of the NA LCS, which only holds 50 players, many of which are imports.

1.2 Imports

A second equation must be formed, which can then be compared with the "North American Player" equation above. A similar table below has been made for NA LCS imports.

Player	Hours Streamed/Month	Avg. viewers	Coefficient	AVC
Ray	31	1,045	1.745	1,824
Xmithie	30	91	1.590	145
HuHi	10	178	1.368	244
Froggen	28	2,466	1	2,466
Flame	28	48	1.368	66
Cody Sun	29	50	1.675	84
Olleh	27.5	193	1.842	355
Arrow	136	208	1.745	363
Ssumday	42.5	181	1.704	308
Chaser	14	36	1.704	61
Keane	48	161	1.704	274
lira	11	142	1.745	248
Piglet	19	383	1.704	652
Svenskeren	20	2,456	1.675	4114
Bjergsen	13	21,525	1.842	39641
Biofrost	17.5	944	1	944

The following players either have streamed League of Legends for less than 10 hours within 30 days. These players will not be used for the test.

- Impact
- Jensen
- Looper
- Gate
- Ryu
- Inori
- Seraph
- Ninja
- Reignover

The 3 imports with the highest viewership are all Danish (Froggen, Svenskeren, Bjergsen). These players have around 10× as many viewers as non-danish players. I will not include them in my formula as they will have a very significant impact on the equation.

The Product Moment Correlation Coefficient for the second set of data tells us that $r = -0.023$. This implies that there is no correlation between how frequent an imported player streams, and their mean viewer count. Thus it would be unreasonable to find an equation using the data we have. Because of this, we cannot compare regression models for North American and Imported players, instead, we will compare the means of each set of data via hypothesis testing.

1.3 Hypothesis Testing

Two hypotheses will be tested, which are as follows:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Where μ_1 is the mean viewers for North American players, for the entire population, and μ_2 is the mean viewers for imported players, for the entire population. The results will therefore account for all future NA LCS players, as well as present players. H_0 is the null hypothesis, claiming that North American players average the same viewers as imported players. In contrast, H_1 is the alternative hypothesis, claiming the average viewer count for North American and Imported players is not equal. To test this claim, t_0 must be calculated.

$$t_0 = \frac{\overline{x_1} - \overline{x_2}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

$$s^2 = \frac{\sum(x-\overline{x})^2}{n-1}$$

Where s_1^2 is the sample variance for North American Players, s_2^2 is the sample variance for Imported players, and s_p^2 is the pooled sample variance. N.B: $\overline{x}$ denotes the mean adjusted viewer count for this hypothesis test. It previously denoted monthly hours spent streaming.

$$\overline{x_1} = 4547.77$$

$$\overline{x_2} = 428.31$$

$$s_1 = 7188.15$$

$$s_2 = 467.67$$

$$n_1 = n_2 = 13$$

$$\Rightarrow s_p = 5093.54$$

$$\Rightarrow t_0 = 2.06195$$

For this test, a significance level of 0.01 will be used. This equates to a 99% confidence level. This hypotheses test is considered "Two-tailed". This is because the alternative hypothesis states "Not equal". The test would be a "One tailed test" if the alternative hypothesis stated "Less than" or "Greater than". Because we are using a two-tailed test, to find the critical t values, the significance level must be halved to 0.005. This column corresponding to this value will give our critical t value on the "Student's t Distribution" table.[3]

1.4 Results

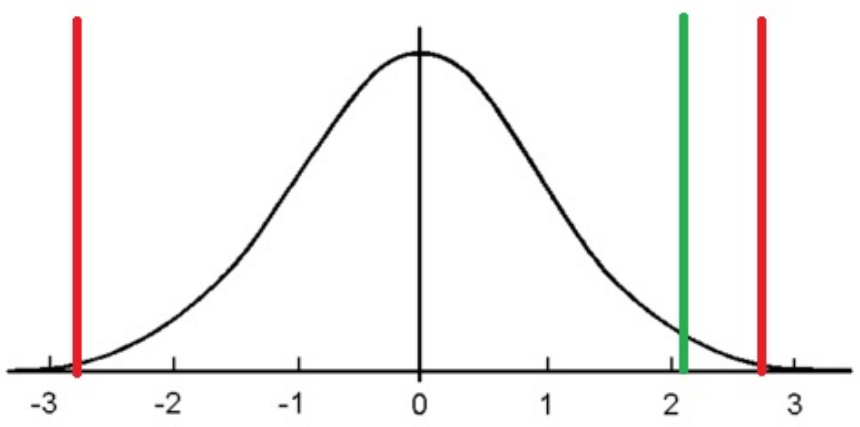

Figure 1.1: Student's t Distribution Graph - $t_{24,0.005}$ is represented by the red lines, t_0 is represented by the green line.

$$t_{24,0.005} = \pm 2.797$$

$$t_0 = 2.06195$$

As shown above, t_0 lies within the two $t_{24,0.005}$ values. This is known as the acceptance region. Because of this, we must accept the null hypothesis, thus statistically, $\mu_1 = \mu_2$. North American players and imported players have the same average viewers on their live streams. This may be caused by the small sample size - A larger sample size would lead to smaller critical t values. A larger sample size would also give more accurate values for $\overline{x_1}$ and $\overline{x_2}$, which would change the observed t value. The large sample variances from both sets of data also lead to uncertainty, causing the null hypothesis to be accepted.

Chapter 2

Spring & Summer Split Correlation

Spring and summer split correlation was not mentioned in the podcast, however I decided to test it myself out of personal interest. Since the summer split has not taken place currently in 2017, I will use data from the splits of 2015. The reason behind not using data from 2016 is that many organizations bought spots for the summer split, making it difficult to merge teams.

N.B: Winterfox were relegated in the spring promotion, and were replaced by Team Dragon Knights. For this table, I will consider them the same organization. Team Coast were replaced by EnemyGG heading into the summer split, so I will also consider those two organizations the same.

Team	Spring Split Rank	Summer Split Rank	$Delta^2$
Team SoloMid	1	2	1
Cloud 9	2	8.5	42.25
Team Liquid	3	3	0
Team Impulse	4	4	0
CLG	5.5	1	20.25
Gravity Gaming	5.5	5.5	0
Team 8	8.5	8.5	0
Winterfox/Team DK	8.5	8.5	0
Team Dignitas	8.5	5.5	9
Team Coast/EnemyGG	8.5	8.5	0

$$\sum d^2 = 63.5$$

"Spearman's Rank Correlation Coefficient" can be used to test if there is a relation between team ranking during the spring & summer split. Since the $5^{th} - 6^{th}$ ranking is shared, the rank was split to 5.5. The same rule applies for ranks $7 - 10$. The coefficient is calculated by:

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

$$\Rightarrow r_s = 0.615$$

Another hypothesis test can be used to indicate a possible relation.

H_0 : Spring and summer split ranks are uncorrelated.
H_1 : There is some correlation between spring and summer split ranks.

The calculated value, r_s can be compared to a critical value[4] to test the claims. This is a two-tailed test, and a significance level of 0.01 will be used. We also have 10 pairs of data, thus the critical values are $r_{0.005,10} = \pm 0.7818$. Since the observed value, r_s lies within the acceptance region, we will accept H_0, and reject H_1, concluding that there is no correlation between spring and summer split rankings.

Bibliography

[1] Twinge: Twitch User Data
http://twinge.tv/

[2] TwitchTools: Total Twitch viewership
https://www.twitchtools.com/stats?type=twitch&method=week

[3] Student's t Distribution Table data
http://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf

[4] Spearman's Rank Correlation Coefficient Table
http://www.bws.wilts.sch.uk/Curriculum/pdfs/geog_SpearmanExplained.pdf