

Statistical/Hypothetical Question

My statistical question is to find if Fire Type Pokemon have higher stats compared to other Pokemon types. With this information, I will be able to conclude as to why Fire Type Pokemon are so well-received by the Pokemon fandom.

To test my question I will utilize the following variables: Type1, Total, HP, Attack, Defense, Speed, and Legendary. I will perform statistical testing to include the usage of histograms, line charts, and scatter plots. To understand why I have selected these variables I need to explain the other variables within the dataset. The data set utilized contains the following variables:

1. #
2. Name
3. Type 1
4. Type 2
5. Total
6. HP
7. Attack
8. Defense
9. Sp. Atk
10. Sp. Def
11. Speed
12. Generation
13. Legendary

is the Pokemon's assigned number in the Pokedex, something not relevant to my statistical question but important to mention because Pokemon are numbered 1 to 801 in the Pokedex (Note this dataset is incomplete as there are now 1,025 Pokemon). Name is the name of the Pokemon. Type 1 is the Pokemon's main type, Type 1 defines a Pokemon's main characteristics. Type 2 is its subtype, defining its

minor characteristics. I will omit Type 2 from the data set because there are many empty cells, not every Pokemon has a Type 2 and every Pokemon has a Type 1.

Next is Total, one of the variables I will be using. Total represents the attribute total of a particular Pokemon. For example, Bulbasaur, Pokemon # 1, has a Total of 318. Breaking down Bulbasaur's stats we have 45 Health Points (HP), 49 Attack, 49 Defense, 65 Special Attack (Sp. Atk.), 65 Special Defense (Sp. Def), and 45 Speed, all added together equal the 318 Total or attribute total.

HP is how much damage a particular Pokemon can take before fainting and becoming unable to continue the battle. Attack is the power of a Pokemon's attack and Defense is a Pokemon's resistance to an attack. The resultant attack damage is the difference between Attack and Defense. Sp. Atk. and Sp. Def. works the same as Attack and Defense but is for attacks that are type-specific, recall Type 1 and Type 2. Bulbasaur, for example, is a Grass Type with a Type 2 of Poison, and can use Grass Type moves that carry the Sp. Atk. statistic, when used, their effectiveness is reduced by Sp. Def.

Speed is the order in which Pokemon will attack. The higher the speed the greater chances for the Pokemon to attack first. However, there are many Pokemon moves (think general Dungeons and Dragons abilities like Fireball) that carry attributes that may boost the chances of a Pokemon with lower speed to attack first.

Lastly, we have Generation and Legendary. Generation is a grouping a Pokemon is tied to, there 6 Generations at this point each containing its own unique Pokemon created specifically for that Generation. Last is Legendary, a one-of-a-kind Pokemon. I chose Legendary as my last variable because it will lend an ad hoc nature to my statistical question.

The outcome of my EDA

The regression analysis gives me an $r\text{-squared} = 0.003$ or about .3%. This tells me that .3% of the variance in Total Stat is explained by whether a Pokemon is a Fire Type or not. This is a very low percentage suggesting Fire Type vs. Other Type is not a strong predictor of Total.

The hypothesis test tells me that Fire Type Pokemon have a slightly higher Total on average seen in my T-statistic of 1.429, but this does not suggest evidence against my null hypothesis. Since my $p = 0.153$, I failed to reject my null hypothesis. This means I do not have enough evidence to support that Fire Type Pokemon have significantly higher stats compared to all other Pokemon types.

My Thoughts

As far as things missed, the breadth of my analysis could be deeper. I could have sought a more complete dataset that includes all 1,025 Pokemon. The regression model only considers Type 1 as the predictor, adding more predictors may have led to an increased R^2 . Many other variables could also be influencing Total as well such as Generation or Type 2.

The inclusion of Type 2 as a variable might have had an impact but would have resulted in removing many Pokemon. Additionally, Type 2 could also be considered an outlier. A Pokemon with Type 2 gains influence on their stats, as seen by Fire Type Pokemon tending to have a higher Speed and Attack than Other Types. This could be alleviated by treating Type 2 as a stand-alone category, but wouldn't assist in answering my statistical question. If I decided to treat Type 2 as a stand-alone category then I could conduct t-tests to compare them to Pokemon without a Type 2 and see if the differences in Total are significant.

I came into this research project with assumptions, and some are present in the models. The regression assumes a linear relationship between Fire Type and Total. If the relationship is nonlinear other models might be better. I assumed there was favoritism towards Fire Type Pokemon. I started to see this after the creation of the "Comparison of Fire Type Legendary Pokemon to Other Legendary Types" bar plot.

While bar plots are not incredibly statistical, they do break down the data in a simple manner. I assumed there would be more Fire Type Legendary Pokemon to skew the data in favor of the standard Fire Type Pokemon, but as it turns out Psychic and Dragon Type 1 might be worth looking into. This brings me to my greatest challenge, the utilization of Type 1, Type 2, or both. Even with 12 weeks to work on this project, there can always be more EDA. But with more time I could have avoided concluding,

generalizing the data, and performed more tests on Type 2 to potentially find statistical significance to whether Fire Type Pokemon have higher stats compared to other Pokemon types.