An Investigation of Punxsutawney Phil: A Statistical Analysis of Groundhog Day Predictions and Temperature Trends

A Computational Methods Final Project

Jordan Patten Hard
Undergraduate Student of Mechanical Engineering
Brigham Young University

June 2025

Abstract

This project evaluates the accuracy of Punxsutawney Phil's Groundhog Day predictions using historical weather data. Specifically, it investigates whether Phil's shadow-based forecasts correlate with actual changes in average temperatures between February and March. Using confidence intervals and hypothesis testing for proportions on a sample of 118 predictions, this analysis tests the claim that Phil predicts the arrival of spring better than chance. The null hypothesis states that Phil is correct less than or equal to 50% of the time, while the alternative hypothesis is that he is correct more than half of the time. I hypothesize that Punxsutawney Phil is correct more than half of the time. Multiple temperature thresholds were evaluated to define the end of winter, and therefore define whether or not Phil's prediction was correct. The tests all resulted in very large p values (i.e. $p = 1.0$, 0.999, and 0.5), and all confidence intervals contained 0.5. Therefore, we fail to reject the null hypothesis. The data shows no significant evidence that Punxsutawney Phil's forecasts are more accurate than random guessing.

Introduction

Every year on February 2nd, thousands gather in Punxsutawney, Pennsylvania, awaiting the spring forecast from the world's most famous groundhog, Punxsutawney Phil. According to tradition, if Phil sees his shadow, six more weeks of winter are expected. However, if he doesn't see his shadow, an early spring is on the way.

While the event is a beloved cultural ritual, many question the accuracy of Phil's predictions. Is there any truth behind the shadow-based forecasting, or is it simply superstition? To investigate, I will analyze historical average temperature data alongside Phil's annual predictions to evaluate whether his forecasts align with actual weather patterns. Leaning toward belief rather than skepticism, I hypothesize that Punxsutawney Phil is correct more than half of the time.

Methods

To test this hypothesis, a one-sample hypothesis test for proportions was used. The null hypothesis was that Phil is correct less than or equal to 50% of the time. The alternative hypothesis was that he is correct more than 50% of the time.

A prediction was considered correct under the following conditions:

- If Phil saw his shadow and March not significantly warmer than February (i.e., more winter)
- If Phil did not see his shadow and March was significantly warmer than February (i.e., early spring)

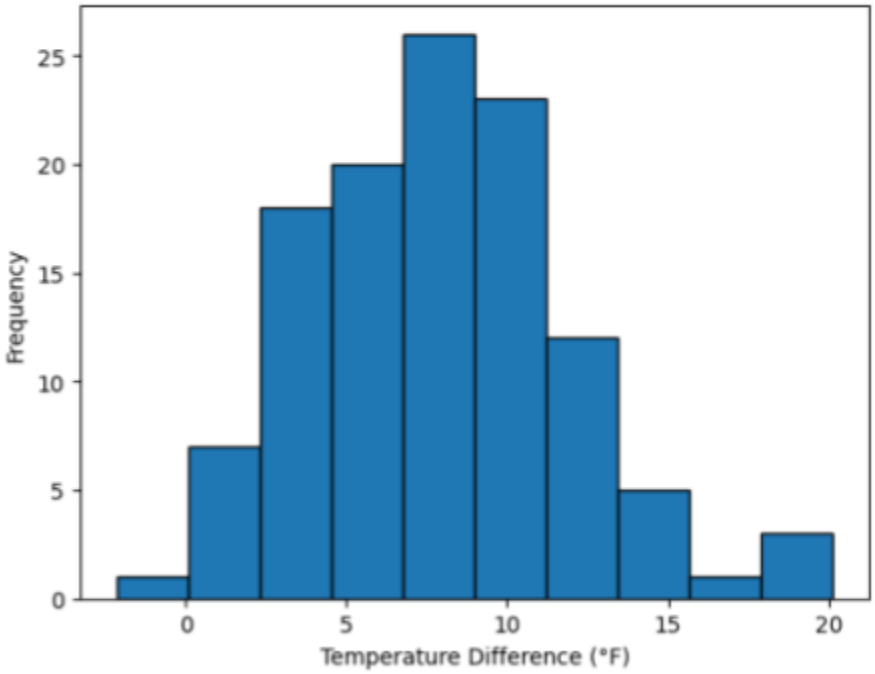

Figure 1. A histogram showing the distribution of average temperature differentials between the months of March and February from the year 1894 to 2016.

Since the definition of "early spring" is subjective, four different thresholds were used to define a significant temperature increase between March and February. The distribution of average temperature differences is shown in Figure 1. The thresholds are as follows:

- A difference of 0°C (no change)
- The mean of the distribution (7.88℉ increase)
- The median (7.96℉ increase)
- The 25th percentile (5.29℉ increase)

Each threshold represents a different interpretation of what constitutes a "warm-up" in March. The dataset originally contained 132 entries. After removing years with missing temperature data

or ambiguous predictions (e.g., “No Record,” “Partial Shadow”), 118 usable data points remained.

Using ***if*** statements in Python, I determined whether each prediction was correct according to the defined threshold, and calculated the proportion of correct predictions for each method. For hypothesis testing and confidence intervals, I used ***norm.cdf*** or ***norm.ppf*** from the ***scipy.stats*** library. A plus-4 adjustment was applied to the confidence intervals for improved accuracy on proportions.

Results

Results for the evaluation of the proportion of correct predictions for each definition of winter are shown in Figure 2. After calculating the proportion of correct predictions, I used a one-sample hypothesis test for proportions, with hypotheses outlined in the methods section. The results are outlined in Table 1, along with the results for a 95% confidence interval for the true population proportion of accurate predictions. None of the p values for the proportion of predictions were statistically significant at a level of $\alpha = 0.05$.

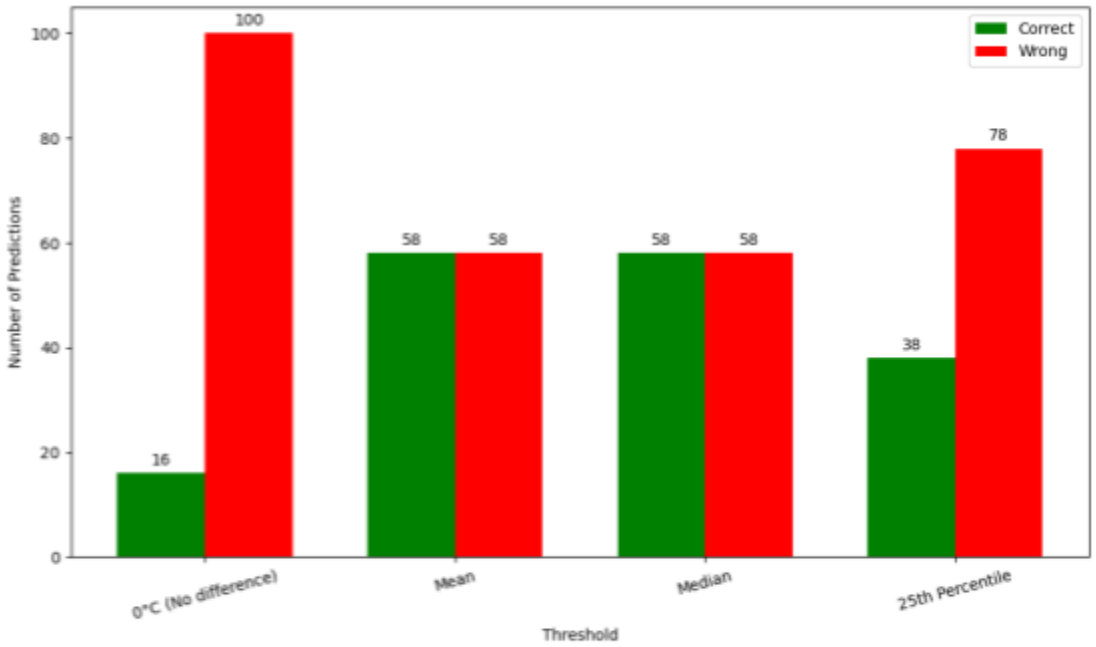

Figure 2. A bar chart illustrating the number of times that Phil the groundhog made a correct prediction regarding winter, based on different definitions of what qualifies as winter.

Table 1. A summary of statistics obtained from a hypothesis test for proportions and a computed confidence interval for a population proportion of correct predictions made by Punxsutawney Phil.

Method	95% Confidence Interval	P Value	Statistically Significant (α = 0.05)
Mean/Median	[0.425, 0.575]	0.5	No
25th Percentile	[0.263, 0.404]	0.999	No
Difference of Zero	[0.096, 0.204]	1.0	No

Discussion

Although this study was rooted in curiosity and cultural folklore, it illustrates the application of statistical methods in evaluating claims. Because none of the p values are large, we have no basis for rejecting our null hypothesis. Therefore, we fail to reject the null hypothesis, and cannot conclude that Phil's predictions are correct more than 50 percent of the time. It appears that Punxsutawney Phil's predictions do not significantly outperform random chance under any of the thresholds tested. While the folklore is entertaining and charming, the data suggests that Phil may not be a reliable source of meteorological insight. Similarly, the calculated confidence intervals provide a region in which we can be 95% confident that the true population proportion of correct predictions lies.

However, this analysis is limited by its simplicity. It does not account for broader climatic trends, geographic variation, or multi-month comparisons. Additionally, the thresholds chosen for defining "spring", or a significant warm-up, could be refined using more objective or meteorologically established and sophisticated metrics.

AI and Academic Honesty

I used AI tools to help debug my Python code and to assist with editing the report. I have verified the aspects of my report that were generated by Artificial Intelligence.

References

Groundhog Club. (2023). *Groundhog Day* [Dataset]. Kaggle. https://www.kaggle.com/datasets/groundhogclub/groundhog-day