Homework 1

Steve Bronder
Statistical Inference

October 21, 2014

Exercise 1. Find a dataset and evaluate whether Punxsutawney Phil can accurately predict six more weeks of winter

Answer

To evaluate whether Punxsutawney Phil can accurately predict six more weeks of winter we have to define a criteria for evaluation. Our criteria will be whether or not Phil accurately predicted a greater than average number of days with snowfall for the months February and March of each year. If Phil predicts a long winter for a year we will give him 1 point out of a possible 65 points. Phil is limited to 65 points because our data ranges from 1950 to 2014 ($T = 65$). We gather data from the National Climate Data Center[1] for the county Punxsutawney Phil is located in, Jefferson County, PA. To find whether Phil saw his shadow we gather data from stormfax[2] on whether or not he saw his shadow for our target years.

Our data set is comprised of three variables, The year, the number of days it snowed in the months of February and March, and whether or not Phil saw his shadow for a given year. Lets have a look at our data.

```
ghog.data<-read.csv("./ghog_clean.csv",header=TRUE)

str(ghog.data)

## 'data.frame': 65 obs. of  3 variables:
##  $ Year    : int  1950 1951 1952 1953 1954 1955 1956 1957 1958 1959 ...
##  $ Snowfall: num  15 13 5 13 15 13 8 13 19 10 ...
##  $ ghog    : Factor w/ 2 levels "NOShadow","Shadow": 1 1 2 2 2 2 2 2 2 2 ...

summary(ghog.data[,2:3])
```

[1] http://www.ncdc.noaa.gov/cdo-web/datatools/findstation
[2] http://www.stormfax.com/ghogday.htm

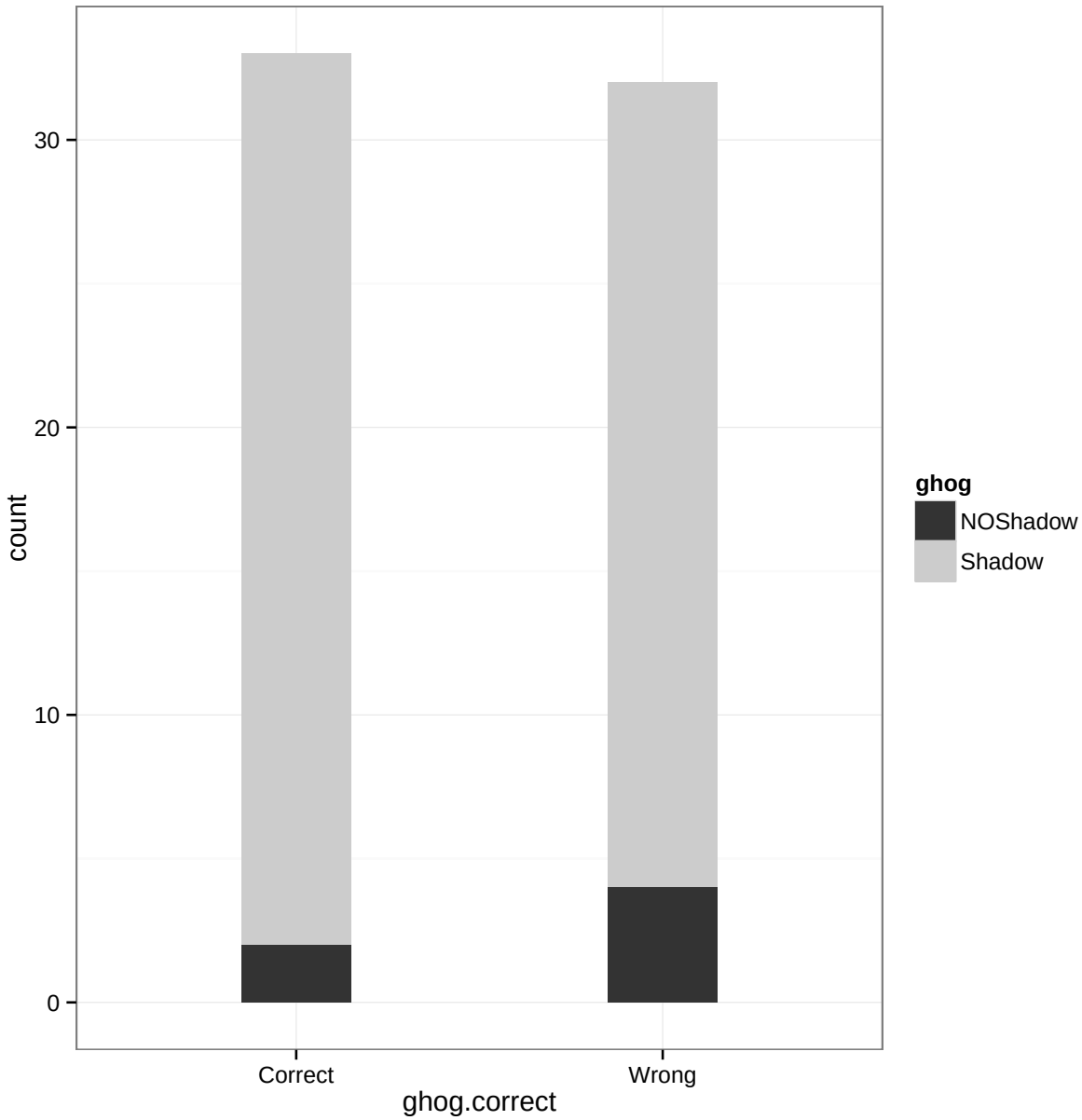
```
##     Snowfall          ghog
##  Min.   : 5.0   NOShadow: 6
##  1st Qu.:15.0   Shadow  :59
##  Median :18.0
##  Mean   :19.1
##  3rd Qu.:25.0
##  Max.   :31.0
```

Notice that we do not have very many years Phil does not see his shadow (No six weeks of winter). Next we will create a variable for whether the amount of days it snowed was greater than the average amount of days for a given year, whether Phil predicted a Long Winter, and whether his prediction was correct

```
 ghog.data$gr.snowfall <- ifelse(ghog.data$Snowfall<mean(ghog.data$Snowfall),
                                 "Long Winter","Short Winter")
 ghog.data$gr.snowfall <- as.factor(ghog.data$gr.snowfall)

ghog.data$gr.ghog.predict <- ifelse(ghog.data$ghog=="Shadow",
                                    "Long Winter","Short Winter")
ghog.data$gr.ghog.predict <- as.factor(ghog.data$gr.ghog.predict)

ghog.data$ghog.correct <- ifelse(ghog.data$gr.ghog.predict==ghog.data$gr.snowfall,
                                 "Correct","Wrong")
ghog.data$ghog.correct <- as.factor(ghog.data$ghog.correct)
```

Now that we have created our variables lets generate a bar graph in ggplot2 that lets us examine the data further

```
 library(ggplot2)

ghog.graph <- ggplot(ghog.data, aes(ghog.correct,fill=ghog)) + geom_bar(width=.3) +
   theme_bw() + scale_fill_grey(end=.8)

ghog.graph
```



From this graph and table we can see he is correct 33 times and incorrect 32 times. Most of his errors are due to lacking short winter predictions. In fact, these prediction values are so close you could infer that Phil is probably as good at predicting the weather as the flip of a coin. This analysis find that Punxsutawney Phil is wrong about fourty nine percent of the time. This is deemed a low score and so we conclude that Phil is a bad predictor of longer winters.

	gr.snowfall	gr.ghog.predict	ghog.correct
1	Long Winter :35	Long Winter :59	Correct:33
2	Short Winter:30	Short Winter: 6	Wrong :32

Table 1: Values for Number of Snowfalls, Groundhog Predictions, and Correct Groundhog Predictions

	Prediction	Accuracy
1	Correct	0.507
2	Incorrect	0.493

Table 2: Percent of Correct and Incorrect Predictions