Lab 7

Lab Overview

For this lab we will use a candy dataset collected by www.fivethirtyeight.com. Additional details about the dataset are available below (courtesy of Kaggle).

```
candy <- read_csv('http://math.montana.edu/ahoegh/teaching/stat446/candy-data.csv')
candy
```

```
## # A tibble: 85 x 13
##    competitorname chocolate fruity caramel peanutyalmondy nougat
##    <chr>              <dbl>  <dbl>   <dbl>          <dbl>  <dbl>
##  1 100 Grand              1      0       1              0      0
##  2 3 Musketeers           1      0       0              0      1
##  3 One dime               0      0       0              0      0
##  4 One quarter            0      0       0              0      0
##  5 Air Heads              0      1       0              0      0
##  6 Almond Joy             1      0       0              1      0
##  7 Baby Ruth              1      0       1              1      1
##  8 Boston Baked ~         0      0       0              1      0
##  9 Candy Corn             0      0       0              0      0
## 10 Caramel Apple~         0      1       1              0      0
## # ... with 75 more rows, and 7 more variables: crispedricewafer <dbl>,
## #   hard <dbl>, bar <dbl>, pluribus <dbl>, sugarpercent <dbl>,
## #   pricepercent <dbl>, winpercent <dbl>
```

Context

What's the best (or at least the most popular) Halloween candy? That was the question this dataset was collected to answer. Data was collected by creating a website where participants were shown presenting two fun-sized candies and asked to click on the one they would prefer to receive. In total, more than 269 thousand votes were collected from 8,371 different IP addresses.

Content

`candy-data.csv` includes attributes for each candy along with its ranking. For binary variables, 1 means yes, 0 means no. The data contains the following fields:

- chocolate: Does it contain chocolate?
- fruity: Is it fruit flavored?
- caramel: Is there caramel in the candy?
- peanutalmondy: Does it contain peanuts, peanut butter or almonds?
- nougat: Does it contain nougat?
- crispedricewafer: Does it contain crisped rice, wafers, or a cookie component?
- hard: Is it a hard candy?
- bar: Is it a candy bar?
- pluribus: Is it one of many candies in a bag or box?
- sugarpercent: The percentile of sugar it falls under within the data set.
- pricepercent: The unit price percentile compared to the rest of the set.
- winpercent: The overall win percentage according to 269,000 matchups.

Acknowledgements:

This dataset is Copyright (c) 2014 ESPN Internet Ventures and distributed under an MIT license. Check out the analysis and write-up here: The Ultimate Halloween Candy Power Ranking. Thanks to Walt Hickey for making the data available.

Questions

Assume we are interested in understanding the `winpercentage` for four groups of candies:

1. chocolate and pluribus
2. chocolate and not pluribus
3. no chocolate and pluribus
4. no chocolate and not pluribus

1. (5 points)

Compare and contrast stratified sampling with domain estimation. How are they similar and how are they different.

2. (5 points)

A stratified sample with ten samples from each strata has been taken for you. Compute the point estimates for mean `winpercentage` for each strata.

```
stratified_sample <- candy %>% group_by(chocolate, pluribus) %>% sample_n(10) %>% ungroup()
```

3. (5 points)

An SRS sample of size 40 is also taken. Compute the point estimates for mean `winpercentage` within each strata.

```
srs_sample <- candy %>% sample_n(40)
```

4. (5 points)

Compute the variance of the mean `winpercentage` for each domain. You can assume that N and N_d are known.