STT 3850 Midterm Study Guide

Andrew Thorp

October 9, 2017

Statistics

Characterizing a set of data (3 S's)

- Shape: how the data is distributed
- Low outliers make a dataset skewed to the Right
- High outliers make a dataset skewed to the Left
- Normal distributions have fairly even outliers on either side

Center: Where the data is centered around

- Normal: If the dataset has a normal distribution (shape) this can be calculated using the `mean($data)` function.
- Skewed (left or right): The mean will be misrepresent the center. Calculate a skewed center using `median($data)`.

Spread: How far the data differs from the center

- Normal: If the dataset has a normal distribution, then the standard deviation applies to both sides of the data and so it represents the spread.
- Skewed (left or right): If the dataset is skewed on either side, the deviation above and below the center will not be the same, so you must calculate it using `IQR($data)` for the interquartile range.

Hypothesis testing (5 step procedure)

- Z-Score: The number of Standard deviations an element is form the mean.
- P-Score:
- $\bar{X} = mean$

1. Specify the Null and ALternative hypothesis
   - Null hypothesis notated as $H_0 : M = value$ or $\bar{M_1} - \bar{M_2} = 0$
   - Alternative hypothesis notated as $H_A : M \neq 0$
2. Test your staitstic using the Z-test or T-test
   - `t.test(variable~catagoricalVariable, data=DF)` will perform the t-test on a set of data. If the data is not tidy your might want to use `dplyr` to tidy it up first.

   Example:

```
DF <- ChickWeight %>%                          #imports data frame
  filter(Diet %in% c(3:4))                     #then removes all collums except for 3 and 4
                                               #weight~Diet means the weight value grouped by the Diet
t.test(weight~Diet, data=DF)                   # Quickly gives us what we need to know
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  weight by Diet
## t = 0.75908, df = 226.16, p-value = 0.4486
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -12.26840  27.64298
## sample estimates:
## mean in group 3 mean in group 4
##        142.9500        135.2627
```

Markdown Dplyr Ggplot2