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.
   - Alternatively, if we have an expected mean, and our data has the experimental mean, we can call the `t.test` function as such: `t.test(experimental_data, mu = <expected_mean>)` 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 across Diets value.
t.test(weight~Diet, data=DF)           # Quickly gives us what we need to know
```

3. Determine rejection region
   - Rejection region is the percent of data points on either end of a data set.
   - Percentile is represented by alpha (α) (probability of a type 1 error)
   - Unless otherwise specified, assume rejection region to be 5% (0.05).

4. Statistical conclusion.

   You can either

   - Reject the null hypothesis H_0 if $P < \alpha$

   or

   - Fail to reject the null hypothesis H_0 if $P \geq \alpha$

   Based on the output from the previous example pictured below, P = 0.45, and we assume α=0.05. P is far greater than α, so we **fail to reject the null hypothesis** H_0.

```
##
##  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
```

5. English conclusion (or Spanish, if that's your thing)
   - The statistical conclusion deals with the null hypothesis H_0, but the English conclusion only deals with things in terms of the Alternative Hypothesis H_A (written out)

   In the case of our example from above, we failed to reject the null hypothesis, so we **failed to find evidence supporting a weight difference in chicks between Diets 3 and 4**.
   - The English hypothesis must ALWAYS be written this way. You either find evidence to support H_A or you fail to find evidence to support H_A.
   - There is a chance for error:
   - type 1 error: Assuming the Alternative hypothesis when the Null hypothesis is true.
   - type 2 error: Assuming the Null hypothesis when the Alternative hypothesis is true.

Cohens'd (effect size) Not really used in this class

- The difference between a control group's mean, and an experiment group's mean, measured in standard deviations.
- Measured using the following formula: $d = \frac{|M_0 - M_1|}{SD}$
- How large is the effect size?
  - small if |d| < 0.20
  - medium if |d| < 0.5
  - large if |d| > 0.5
- Compute using R with `cohensD(post_data, pre_data, method = "paired")`

```
Diet_3 <- filter(ChickWeight, Diet == 3)
Diet_4 <- filter(ChickWeight, Diet == 4)
#method should normally be "paired", but the sample size was different
cohensD(Diet_4$weight, Diet_3$weight, method = "unequal")
```

Analysing a DataSet

- An analysis consists of
  - t-value

  - p-value
  - effect size (Cohen's D)

```
DF <- ChickWeight %>%
  filter(Diet %in% c(3:4))
t.test(weight ~ Diet, data = DF)

Diet_3 <- filter(DF, Diet == 3)
Diet4 <- filter(DF, Diet == 4)
  cohensD(Diet_3$weight, Diet_4$weight, method = "unequal")
```

```
From this we can gather the effect size is roughly 10% and, with a t-value of 75% and a p-value of 45%.
TODO: ask about reading these statistics.
```

Probability

- Binomial probablility: The probablility of a binary outcome (not always 50%)
  - Written X~Bin(#,)
  - Programmed into R with `sum(dbinom(low#:high#, <number_of_trials>, <probablity>))`

```
#Probablity of making at LEAST 7 shots out of 10, with a 70% success rate (statistically)
prob <- 0.7
shots <- 7
trials <- 10
sum(dbinom(shots:trials, trials, prob))
```

```
## [1] 0.6496107
```

Dplyr

Ggplot2

-