Analyzing Numerical Data with Outliers/Skew

Ryan Miller

Introduction

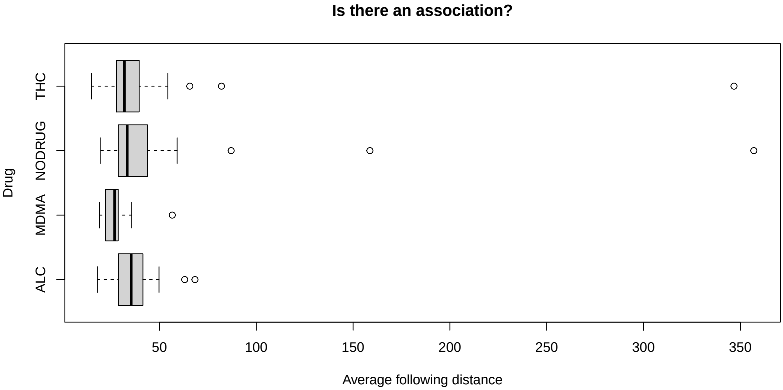
- We've previously learned about the *t*-test as a method for statistical inference on numerical data
  - The *two-sample t-test* compares the difference in means of two separate populations
  - The *one-sample t-test* compares the mean of one population to a null value (often the mean of paired differences)

Introduction

- We've previously learned about the *t*-test as a method for statistical inference on numerical data
    - The *two-sample t-test* compares the difference in means of two separate populations
    - The *one-sample t-test* compares the mean of one population to a null value (often the mean of paired differences)
- There are a few challenges that tend to arise in the analysis of numerical data (that do not occur with categorical data)
    - *Outliers*, or unusual observations with a large impact on the sample/population mean
    - *Skew*, or data whose distribution doesn't meet the assumptions made by the *t*-test

Drug Use and Tailgating

- Do regular users of certain drugs tend to engage in other risky behaviors?
- Participants drove behind a lead vehicle that behaved erratically
  - More cautious drivers should respond by increasing their following distance (“D” in the dataset)



Statistical Comparisons

Recall the *t*-distribution is appropriate in two situations:

1) Small samples from a Normally distributed population
2) Large samples from any population (even highly-skewed or unusual distributions)

Considering the sample sizes shown below, and boxplots shown on the previous slide, you comfortable using *t*-tests to compare these groups?

```
tail <- read.csv("https://remiller1450.github.io/data/Tailgating.csv")
table(tail$Drug)
```

```
##
##    ALC   MDMA NODRUG    THC
##     24     16     40     39
```

The Impact of Outliers

To illustrate the importance of these assumptions, conduct the following tests:

1) Use `t.test()` to compare the MDMA and THC groups
2) Remove the extreme outlier in the THC group (Case #88 with a following distance of 346.72 ft) and repeat this test

You can use the following code to prepare these data:

```r
tail <- read.csv("https://remiller1450.github.io/data/Tailgating.csv")
D_MDMA <- tail$D[tail$Drug == "MDMA"]
D_THC <- tail$D[tail$Drug == "THC"]
D_THC2 <- D_THC[-which(D_THC > 300)]
```

The Impact of Outliers

Below are the *p*-values resulting from the hypothesis tests described on the previous slide:

```r
## Outlier included
t.test(x = D_MDMA, y = D_THC)$p.value
```

```
## [1] 0.08664708
```

```r
## Outlier removed
t.test(x = D_MDMA, y = D_THC2)$p.value
```

```
## [1] 0.02744264
```

Clearly this outlier has a substantial impact, but *should* we remove it?

A Cautionary Tale

- In the 1970s, NASA began monitoring the Earth's atmosphere using satellites
- In 1985, British scientists discovered a large ozone hole above Antarctica
    - Had NASA's monitoring system failed?

A Cautionary Tale

- In the 1970s, NASA began monitoring the Earth's atmosphere using satellites
- In 1985, British scientists discovered a large ozone hole above Antarctica
  - Had NASA's monitoring system failed?
- 1970s technology was prone to measurement irregularities
  - NASA used software that was programmed to flag and set aside data that deviated greatly from the expected measurements
  - Data supporting the existence of the ozone hole existed nearly a decade prior to its discovery!

Outlier Mitigation

Selectively discarding real data is ethically questionable, but there are a few valid reasons to remove data-points:

1) Measurement or recording errors (a pulse of 0, or an age of 166 in study on high students)
2) Not belonging to the target population (a subject lies to get into a study, or a subject doesn't take the study protocol seriously)

Outlier Mitigation

Selectively discarding real data is ethically questionable, but there are a few valid reasons to remove data-points:

1) Measurement or recording errors (a pulse of 0, or an age of 166 in study on high students)
2) Not belonging to the target population (a subject lies to get into a study, or a subject doesn't take the study protocol seriously)

Aside from these examples, a better approach is *mitigation*, or choosing methods that are more robust to outliers:

1) Data transformations
2) Non-parametric tests

Logarithms

- Statisticians use the term "log" to refer to what most call the *natural logarithm*:

$$log(X) = T \leftrightarrow X = e^{T}$$

Logarithms

- Statisticians use the term “log” to refer to what most call the *natural logarithm*:

$$log(X) = T \leftrightarrow X = e^{T}$$

- An important mathematical property of logarithms is that *differences* on the log-scale correspond to *ratios* on the original scale *after exponentiation*:

$$log(X) - log(Y) = log(X/Y)$$

$$e^{log(X/Y)} = X/Y$$

Logarithms - Example

```
LD_THC <- log(D_THC)
LD_MDMA <- log(D_MDMA)
mean(LD_THC) - mean(LD_MDMA)
```

```
## [1] 0.266054
```

- As shown above, the difference in means on the log-scale is 0.26
  - Undoing the log-transformation: $exp(0.26) = 1.30$
  - Thus, average following distances in the THC group are *30% higher* than in the MDMA group

Logarithms - Example

```r
LD_THC <- log(D_THC)
LD_MDMA <- log(D_MDMA)
mean(LD_THC) - mean(LD_MDMA)
```

```
## [1] 0.266054
```

- As shown above, the difference in means on the log-scale is 0.26
  - Undoing the log-transformation: $exp(0.26) = 1.30$
  - Thus, average following distances in the THC group are *30% higher* than in the MDMA group
- This concept applies to confidence intervals too:
  - The 95% CI on the log-scale is (0.05, 0.47), which we can exponentiate to (1.05, 1.60)
  - So we can be 95% confident the mean following distance is somewhere between 5% and 60% higher for THC users in the population these data represent

Logarithms - Practice

- Use the `log()` function in `R` to log-transform the variable `D`, then perform a two-sample *t*-test using the log-transformed data
- How does this compare to our previous results?
  - Recall the two-sample *t*-test *p*-value was 0.027 without the THC outlier and 0.087 with the THC outlier)

Logarithms - Practice (solution)

```r
LD_THC <- log(tail$D[tail$Drug == "THC"])
LD_MDMA <- log(tail$D[tail$Drug == "MDMA"])

t.test(x = LD_THC, y = LD_MDMA)
```

```
##
##  Welch Two Sample t-test
##
## data:  LD_THC and LD_MDMA
## t = 2.5526, df = 49.483, p-value = 0.01383
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.05665175 0.47545634
## sample estimates:
## mean of x mean of y
##  3.547082  3.281028
```

After applying the log-transformation, the *p*-value is actually smaller than either of the previous tests. And we didn't need to selectively discard any data!

Logarithms - Technical Details

- On a technical note, $\sum log(x_i)/n \neq log(\sum x_i/n)$; so the exponentiated mean of the log-transformed data is actually the *geometric mean*
  - In the THC vs. MDMA comparison, 1.30 was actually the ratio of geometric means, not the ratio of arithmetic means
  - I include this only for completeness, it is not an important distinction in a practical sense
- The main take-away is that analyzing the log-transformed data allows us to measure *relative changes* across groups (after the transformation is undone via exponentiation)

Comments

- There are many transformations that statisticians sometimes apply to non-normally distributed data
  - The log-transformation is popular because it retains interpretability (we can use exponentiation make relative comparisons)

Comments

- There are many transformations that statisticians sometimes apply to non-normally distributed data
  - The log-transformation is popular because it retains interpretability (we can use exponentiation make relative comparisons)
- **Non-parametric** tests are a completely different alternative to transforming the data
  - In the slides that follow I will *briefly* introduce a couple of non-parametric analogs to the one-sample and two-sample *t*-tests
  - You *will not be responsible* for knowing the procedural details of these tests, but you should be aware of when they might be used (and you should consider them for your project depending upon the nature of your data)

Wilcoxon Signed-Rank test (one-sample test)

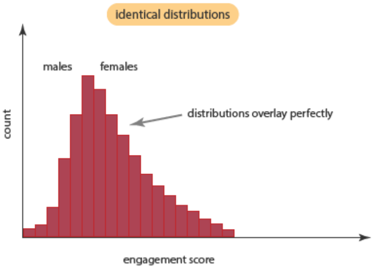
- The **Wilcoxon Signed-Rank test** is a non-parametric analog to the one-sample t-test (single mean)
  - It is most often used to test whether the *median difference* in a paired design is zero

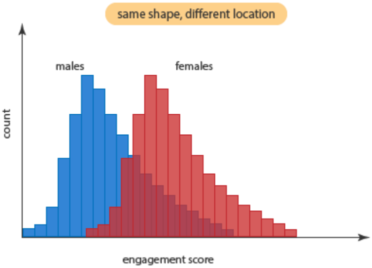
Wilcoxon Signed-Rank test (one-sample test)

- The **Wilcoxon Signed-Rank test** is a non-parametric analog to the one-sample *t*-test (single mean)
  - It is most often used to test whether the *median difference* in a paired design is zero
- Formally, the test specifies $H_0 : m = m_0$, or the median is some theoretical median
  - Next, data-points are ranked (1:N) based upon how far they are from m_0
  - Then, signs (+ or -) are given to these ranks based upon whether the data-point was above m_0 (+) or below m_0 (-)
  - Under the null hypothesis, the sum of the signed-ranks is expected to be zero, which can be used to derive a null distribution and a *p*-value (we won't cover the details

Wilcoxon Rank-Sum (two-sample test)

- The **Wilcoxon Rank-Sum** (synonymous with the Mann-Whitney U-test) is a non-parametric analog to the two-sample *t*-test (difference in means)
  - It tests whether the location of one distribution is *shifted* relative to another
  - In doing so, it makes no assumptions about the shape of the distributions (they could both be skewed, have outliers, etc.)





IMG source: https://statistics.laerd.com/spss-tutorials/mann-whitney-u-test-using-spss-statistics.php

Wilcoxon Rank-Sum (two-sample test)

- Formally, the Wilcoxon Rank-Sum specifies $H_0 : \text{dist}(X_1) = \text{dist}(X_2)$ and $H_A : \text{dist}(X) \neq \text{dist}(Y)$
  - Next, data-points are ranked, regardless of group, from smallest to largest (1:N)
  - Then, these ranks are summed within each group, yielding the quantities R_1 and R_2
  - R_1 and R_2, along with n_1 and n_2 are used to construct a standardized value called the U-statistic
- An exact test or a Z-test can be performed using U (something we won't cover)

Non-parametric Tests in R

In R, the `wilcox.test()` function is used to perform both the Wilcoxon Signed-Rank test and the Wilcoxon Rank-Sum:

```
## Wetsuit Example
wet <- read.csv("https://remiller1450.github.io/data/Wetsuits.csv")
diff <-  wet$Wetsuit -  wet$NoWetsuit

## Two-sample tests
wilcox.test(x = wet$Wetsuit, y = wet$NoWetsuit)$p.value
```

```
## [1] 0.1838276
```

```
t.test(x = wet$Wetsuit, y = wet$NoWetsuit)$p.value
```

```
## [1] 0.1848961
```

```
## One-sample (paired) tests
wilcox.test(x = diff)$p.value
```

```
## [1] 0.00246845
```

```
t.test(x = diff)$p.value
```

```
## [1] 8.885414e-08
```

Practice

- Use the `wilcoxon.test()` function to compare the average following distances in the MDMA and THC groups in the tailgating study.
- How do the results of this test compare to the two-sample t-test on the log-transformed data? (recall this test had a p-value of 0.013)

Practice

- Use the `wilcoxon.test()` function to compare the average following distances in the MDMA and THC groups in the tailgating study.
- How do the results of this test compare to the two-sample *t*-test on the log-transformed data? (recall this test had a *p*-value of 0.013)

```
wilcox.test(x = tail$D[tail$Drug == "THC"], y = tail$D[tail$Drug == "MDMA"])$p.value
```

```
## [1] 0.007328961
```

Conclusion

- This lecture presented strategies you can use to mitigate the influence of outliers in your analysis of numerical data
  - Transformations, such as the log-transformation, make the data better aligned with the assumptions of traditional methods (ie: the *t*-test requiring a Normally distributed population)
  - Non-parametric tests, take a different approach by directly comparing robust measures (ie: medians)
- Overall, you should know when to consider using these methods, and you should have a conceptual understanding their results (ie: a log-transformed difference in means can be exponentiated to provide a ratio of means)