Effect Size Mock Lesson

Andrew Sage

November 30, 2017

Welcome!

- Please fold index card into a "tent" and write your name, as you would like me to address you.
- Access materials for the mock lesson using the following steps
  - Navigate to andrewjsage.github.io
  - Choose *Mock Lesson* from *Teaching* menu
  - Open *Slides*

Learning Outcomes

After participating in this lesson, you will be able to:

- Define effect size
- Explain why it is important to consider effect size when performing statistical inference
- List various ways to measure effect size
- Decide which measure of effect size is most appropriate for a given situation
- Assess the validity of a claim, using effect size to defend your reasoning

Outline

1. Motivating example
2. Discussion of ways to measure effect size
3. Group activity using effect size and other information to make informed decisions
4. Reflection on important points

Flights Out of New York City

The Shiny App available at andrewjsage.shinyapps.io/Flights can be used to compare flight delays for flights leaving LaGuardia Airport (LGA) in New York City, or the nearby Newark Liberty International Airport (EWR) in New Jersey.

How heavily would this information impact your decision of which airport to fly out of if you are:

1. flying into Columbus (CMH) ?
2. flying into Dayton (DAY) ?
3. giving general advice to friends, without knowing their destinaton ?

Choose from: A. Heavily B. Moderately C. A little D. Not at All

Kahoot

1. Go to https://kahoot.it/
2. Enter PIN displayed.
3. Choose nickname (these will be displayed so choose nickname that you are comfortable with and will not be offensive to others)
4. Answer questions as they are asked. You will have 20 seconds for each question.

Tests for Statistical Significance

If we consider these flights to be a random sample of all flights originating at these airports, we might ask whether there is evidence of a difference in mean arrival delays between the airports.

Null Hypothesis: There is no difference in mean arrival times between airports.

Hypothesis Test for Columbus Flights

```
library(tidyverse)
library(nycflights13)
data(flights)
EWR <- flights %>% filter(origin=="EWR" & dest=="CMH")
LGA <- flights %>% filter(origin=="LGA"& dest=="CMH")
t.test(EWR$arr_delay, LGA$arr_delay)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  EWR$arr_delay and LGA$arr_delay
## t = 1.6911, df = 1123.8, p-value = 0.09109
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5022667  6.7719814
## sample estimates:
## mean of x mean of y 
##  9.602041  6.467183
```

Hypothesis Test for Dayton Flights

```
EWR <- flights %>% filter(origin=="EWR" & dest=="DAY")
LGA <- flights %>% filter(origin=="LGA"& dest=="DAY")
t.test(EWR$arr_delay, LGA$arr_delay)
```

```
##
##  Welch Two Sample t-test
##
## data:  EWR$arr_delay and LGA$arr_delay
## t = 8.6037, df = 653.53, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  18.00658 28.65637
## sample estimates:
## mean of x mean of y
## 18.384106 -4.947368
```

Hypothesis Test for All Flights

```
EWR <- flights %>% filter(origin=="EWR")
LGA <- flights %>% filter(origin=="LGA")
t.test(EWR$arr_delay, LGA$arr_delay)
```

```
##
##  Welch Two Sample t-test
##
## data:  EWR$arr_delay and LGA$arr_delay
## t = 17.344, df = 215670, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.947987 3.699146
## sample estimates:
## mean of x mean of y
##  9.107055  5.783488
```

Difference in Test Statistic

Why is the test statistic for all flights so much larger than the test statistic for Columbus-bound flights?

Difference in Test Statistic

Why is the test statistic for all flights so much larger than the test statistic for Columbus-bound flights?

Hint:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Effect Size

Effect size is a measure of the magnitude of the differences between groups or treatments.

While hypothesis testing answers questions of statistical significance, effect size can be used to assess practical significance.

Example: The difference in sample means $\bar{x}_1 - \bar{x}_2$ is a measure of effect size.

Standardized Effect Size

When units are differ between groups, or are difficult to interpret, it can be helpful to use standardized measures of effect size, called **Cohen's d**

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s_p}$$

where $s_p = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$

A value of $d \approx 0.5$ is typically associated with a moderate effect size. A value of $d \approx 0.8$ or higher is typically associated with a large effect size.

Flights Example

The average difference in arrival delay time when flying into Columbus is 9.60-6.47=2.13 minutes ($d = \frac{2.13}{40.13} = 0.053$)

The average difference in arrival delay time when flying into Dayton is 18.38-(-4.95)=23.33 minutes ($d = \frac{23.33}{46.63} = 0.50$)

The average difference in arrival delay time for all airports is 15.11-10.35=4.76 minutes ($d = \frac{4.76}{44.76} = 0.106$)

Effect Size for Proportions

In problems where a proportion is of interest, measures of effect size might include:

- difference in proportions
- relative risk
- odds ratio

Difference in Proportions and Relative Risk for Flights

For flights into Columbus, the probability of arriving at least one hour late is 3 percentage points higher when flying out of Newark than LaGuardia. This difference is 8.6 percentage points when flying into Dayton.

A flight to Columbus is 1.28 times as likely to arrive at least one hour late if it originated in Newark than LaGuardia. A flight to Dayton is more than twice as likely to arrive one hour late if it originated in Newark.

Odds Ratios for Flights

The odds of a flight from Newark to Columbus arriving more than 1 hour late are approximately $91 : 9 \approx 10.1 : 1$.

The odds of a flight from LaGuardia to Columbus arriving more than 1 hour late are approximately $92 : 8 \approx 11.5 : 1$.

Thus the odds ratio is $\frac{\frac{11.5}{1}}{\frac{10.1}{1}} \approx 1.14$.

Odds of arriving at least an hour late are 14% higher when flying out of Newark than LaGuardia.

This compares to odds ratios of 2.51 for flights to Dayton, and 1.27 for all flights.

Examples

For each scenario, the task is to decide whether the data provide convincing evidence to support the claim. Be prepared to defend your answer using an appropriate measure of effect size, and any other information you believe is relevant.

You may choose

A: Agree

B: Need more information

C: Disagree

Group Activity

1. Think individually and decide on your answer. Write this on your white index card.
2. Discuss answers and reasoning with your group. Decide on a group answer and prepare to defend it to the class.
3. Decide on a spokesperson and cardholder for your group.
4. When asked, the cardholder should hold up the color indicating your group's choice.
   A: Agree (Green Card)

   B: Need more information (Yellow Card)

   C: Disagree (Orange Card)
5. I will ask some groups to explain and defend their reasoning.

Fluvoxamine and Childhood Anxiety

A study by The Research Unit on Pediatric Psychopharmacology Anxiety Study Group (RUPP) assessed the impact of fluvoxamine on childhood anxiety. A total of 128 children or adolescents with anxiety disorders were treated with either fluvoxamine or a placebo. Each subject's level of anxiety was measured before and after treatment using the Pediatric Anxiety Rating Scale. Scores range from 0 to 25.

Source: https://www.ncbi.nlm.nih.gov/pubmed/11323729/

Fluvoxamine and Childhood Anxiety Results

	Fluvoxamine Group	Placebo Group
Mean Decrease	9.7	3.1
St. Dev	6.9	4.8
N	63	65

Pooled Standard Deviation Estimate: 5.93

A test of the null hypothesis that there is no difference between treatments results in a p-values less than 0.001.

Claim: Fluvoxamine is an effective treatment for anxiety in children and young adults.

Faculty Teaching Time

A 1992-93 survey by Williams compared the proportion of time that Human Resource Education and Development faculty at U.S. colleges and universities spend teaching to the amount of time they would like to spend teaching. The sample consisted of 155 college faculty members.

Source: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.331.4489&rep=rep1&type=pdf

Faculty Teaching Time

- The study finds that on average faculty spend 52.71% of their time teaching, while they would prefer to spend 49.74% on average.
- A hypothesis test for differences between preferred and actual teaching times results in a test statistic of t = 2.20 and p-value of p=.03.
- A value of 0.09 is reported for Cohen's d.

Claim: Colleges and Universities should hire more Human Resource Education and Development faculty to reduce the time spent teaching by current faculty.

Aspirin and Heart Attacks

The 1987 Physicians' Health Study involved 22,071 male physicians and sought to test whether aspirin, taken regularly in low doses, reduces the risk of mortality from cardiovascular disease. The study tracked the physicians over time and recorded the number who experienced myocardial infaraction (MI).

Source http://www.nejm.org/doi/full/10.1056/NEJM198907203210301

Aspirin and Heart Attacks Result

	Aspirin Group	Placebo Group
N	11,037	11,034
Experienced MI	139	239
Proportion Experiencing MI	0.0126	0.02166

A test of the null hypothesis that there is no difference between treatments results in a p-values less than 0.00001.

Claim: Taking low doses of aspirin regularly reduces the heart attack risk for men.

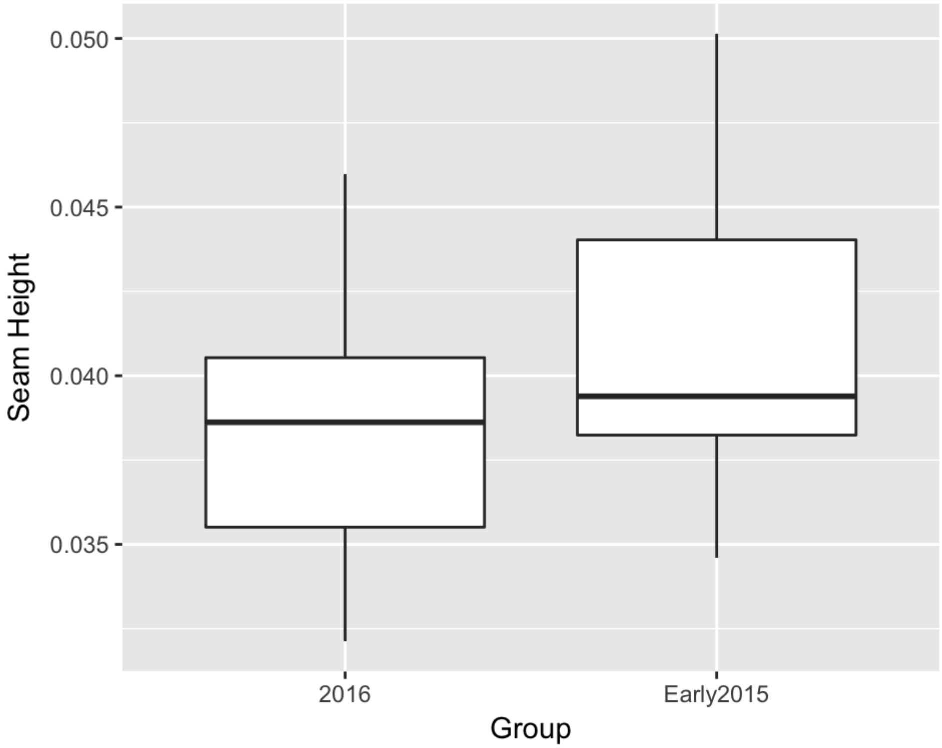
Change in Seam Height in Baseballs?

An article by Ben Lindbergh, titled "The Juiced Ball is Back" explores the surge in homeruns in Major League Baseball, that has occurred since July 2015. The article speculates that changes to the composition of baseballs, including lower seam heights, might be responsible, although MLB denies this claim.

Through EBay, an investigator purchased 17 used gameballs from before July 2015, and 10 from after July 2015. The average seam height was measured on each baseball.

Baseball Example

```
ggplot(data=baseballs, aes(x=Group, y=Seam))+
  geom_boxplot()+xlab("Group")+ylab("Seam Height")
```



Baseball t-test

```
t.test(data=baseballs, Seam~Group)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  Seam by Group
## t = -1.4957, df = 20.981, p-value = 0.1496
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.0058784940  0.0009601411
## sample estimates:
##      mean in group 2016 mean in group Early2015
##              0.03843200              0.04089118
```

Baseball Effect Size: Cohen's d

```
xbar1 <- baseballs %>% filter(Group=="Early2015") %>%
        summarize(x1=mean(Seam))
xbar2 <- baseballs %>% filter(Group=="2016") %>%
        summarize(x2=mean(Seam))
S1 <- baseballs %>% filter(Group=="Early2015") %>%
        summarize(S1=sd(Seam))
S2 <- baseballs %>% filter(Group=="2016") %>%
        summarize(S2=sd(Seam))
Sp=sqrt((16*S1^2+9*S2^2)/25)
d=(xbar1-xbar2)/Sp
d
```

```
##          S1
## 1 0.5765701
```

Claim: Lower seam heights are leading to more homeruns in MLB.

Reflection Questions

1. Explain how a small p-value might be obtained even when effect size is small.
2. If a study shows a large effect size, but also a high p-value what would you recommend and why?