Confidence Intervals

It's What's in Between

James Hunter

7 November 2019

jameshunterbr@gmail.com

Introduction

First, a Step Backward

Theory of Confidence Intervals

An Example Using an Invented Data Set

Constructing Confidence Intervals

Conclusion

How to Deal with the CHIKV/DENV Problem

Bioinformatics at EPM/UNIFESP

Introduction

- Last year: **p-values**
    - Use and abuse
    - Is $\alpha = 0.05$ a reasonable standard to use?
- This year: **confidence intervals**
    - Define them
    - Show how to calculate a simple CI
    - Evaluate their use and abuse

First, a Step Backward

Description vs. Inference

- We can do 2 things with data
    - *Describe* it
    - Draw *inferences* about populations based on it
- The data we normally collect represents a **sample** of a population
    - a subgroup
    - We can describe this
- To draw inferences, we need to know what the population is we want to study
- A distinction with a very great difference
    - That most people forget or never learned

Example coming from the ever popular “Reviewer #2”

- A study of co-infection of Chikungunya and Dengue in Tocantins
    - Textual description of how many patients were in each of the 4 groups being studied
    - CHIK mono infected
    - DENV-1 mono infected
    - DENV-2 mono infected
    - Coinfected

Data Involved

bank	n	percent
CHIKV	47	46.08%
DENV-1	28	27.45%
DENV-2	22	21.57%
Coinfect	5	4.90%
TOTAL	102	-

> *Prevalence rates of Chikungunya virus infection and Coinfection with Dengue virus should be estimate* [sic] *using CI 95%.*

- Prevalence rates??
    - What's the base for the prevalence?
- CI assumes we are making an inference about some population – *which population*?
- The numbers as a description without CI's or p-values were just fine.
- Problem extends to most of the Table 1's we see in papers
    - Simple descriptions of the sample are pushed by editors and reviewers into meaningless p-values and confidence intervals

I will show a workaround to satisfy Reviewer 2 even when we don’t know why we’re doing a confidence interval

Theory of Confidence Intervals

Some Basic Definitions

- *Population Parameter*:
    - A summary measure representing the true population value for the measure
    - Example: population mean (μ), a measure of the central tendency of a continuous variable
- *Sample Statistic*:
    - A point estimate of the population parameter based on the values of a sample from the population
    - Example: sample mean ($\overline{x}$), a number calculated from the values of the sample under study
- *Standard Error*:
    - The standard deviation of the sampling distribution
    - Measure of uncertainty associated with point estimate
    - i.e., the standard deviation of all the possible means of the distribution of values in our sample

Objective of Using These Quantities

- Have our point estimate (sample statistic) exactly match the population parameter
- Not going to happen
    - Except in extremely rare random (i.e., lucky) cases

What is a Confidence Interval?

- “A plausible range of values for the population parameter”
    - Diez, Barr, Cetinkaya-Rundel, OpenIntro Statistics (and others who have copied from their open source work)

Fishing Metaphor (credit to Diez, Barr, Cetinkaya-Rundel)

- Trying to hit the population parameter with a point estimate (sample statistic) is like fishing with a spear
    - Very unlikely to hit the target
- Using a confidence interval is like fishing with a net
    - Much more likely to capture the fish (population parameter value) in the range that the interval covers

Better than a p-value?

- Expressed in the same units as the statistic and the parameter
    - Easier to interpret than a p-value
- Represents a range of values in which the parameter could fall
    - Shows a bit of humility about your powers of inference

An Example Using an Invented Data Set

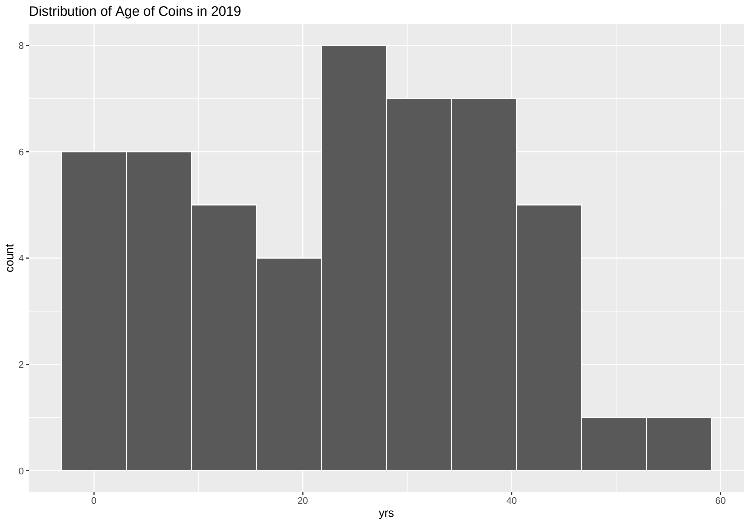
- Objective: replicate process of looking at all potential samples of a population
- We have 50 coins and we know the year they were minted
- Add a variable (`yrs`) that is number of years since minting of the coin

```
## # A tibble: 6 x 3
##        ID  year   yrs
##     <int> <dbl> <dbl>
## 1       1  2002    17
## 2       2  1986    33
## 3       3  2017     2
## 4       4  1988    31
## 5       5  2008    11
## 6       6  1983    36
```

Show Statistics for This Sample

```
## Descriptive Statistics
## ps$yrs
## N: 50
##
##                      yrs
## ------------- -------
##          Mean   23.56
##       Std.Dev   15.18
##           Min    1.00
##            Q1   11.00
##        Median   23.50
##            Q3   36.00
##           Max   57.00
```

Histogram of Data



Bootstrap Resampling

- We can't go out and get the mint year of all the coins in the U.S.
- However, we can take many, **many** samples of the coins we do know about
- **Bootstrapping**
- Technique invented at Stanford in 1980's
- Proofs that effectively imitates drawing samples of unkown coins
- We are going to make 1,000 resamples of our set of coins
- Sampling will be done with replacement
    - This means that when we draw a coin, we put it back so it can be drawn again
    - Coins can repeat within a given sample

One Resample

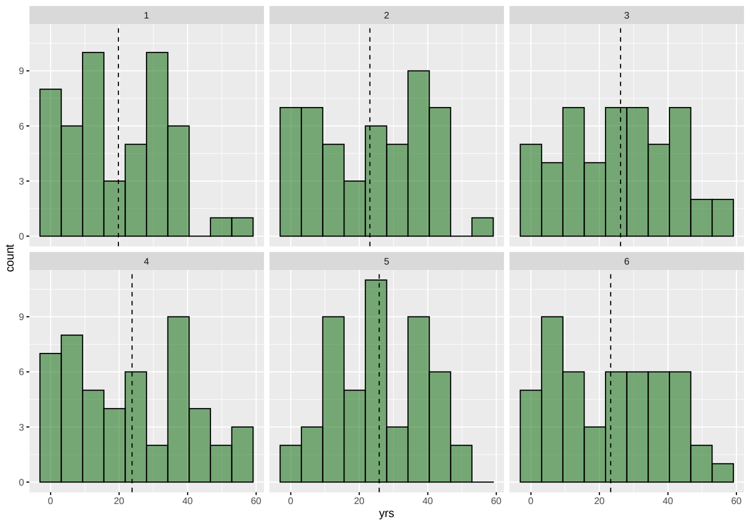
```
## # A tibble: 15 x 4
## # Groups:   replicate [1]
##    replicate    ID  year   yrs
##        <int> <int> <dbl> <dbl>
##  1         1    49  2006    13
##  2         1    37  1962    57
##  3         1     1  2002    17
##  4         1    25  1979    40
##  5         1    10  2000    19
##  6         1    36  2015     4
##  7         1    18  1996    23
##  8         1    49  2006    13
##  9         1    47  1982    37
## 10         1    24  2017     2
## 11         1     7  2008    11
## 12         1    36  2015     4
## 13         1    25  1979    40
## 14         1    37  1962    57
## 15         1    46  2017     2
```

Statistics for Our Resample Compared to Statistics for Original Sample

```
## Descriptive Statistics
## ps$yrs
## N: 50
##
##                    yrs
## ------------- -------
##          Mean   23.56
##       Std.Dev   15.18
```

```
## Descriptive Statistics
## resamp1$yrs
## N: 50
##
##                    yrs
## ------------- -------
##          Mean   24.32
##       Std.Dev   17.06
```

Differences among 6 Samples



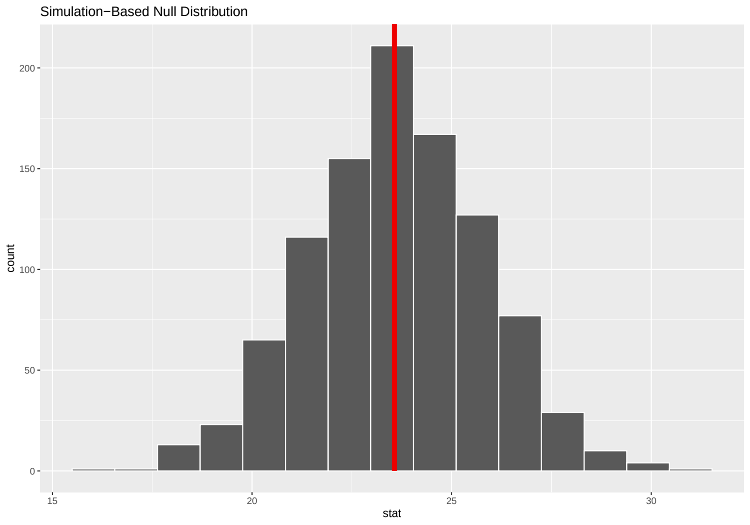
Do the full 1,000 Resamples

```
## # A tibble: 1,000 x 2
##    replicate  stat
##        <int> <dbl>
##  1         1  24.2
##  2         2  22.6
##  3         3  21.0
##  4         4  24.0
##  5         5  22.9
##  6         6  26.6
##  7         7  21.5
##  8         8  26.3
##  9         9  22.1
## 10        10  22.1
## # ... with 990 more rows
```

Statistics for 1,000 Resamples

```
## Descriptive Statistics
## resamples$stat
## N: 1000
##
##                    stat
## ------------- ---------
##          Mean     23.61
##       Std.Dev      2.18
##           Min     16.02
##            Q1     22.18
##        Median     23.59
##            Q3     25.10
##           Max     30.98
##       N.Valid   1000.00
```

Histogram of Means of Resamples



Comparison of Means

- Mean of Sample of 50 Coins = 23.56
- Means of 1,000 Means of Samples = 23.61

Constructing Confidence Intervals

We say we want a 95% confidence interval - which means ??

95% of all the confidence intervals we can create will have the true population mean between the interval's lower and upper limits

- How to determine the upper and lower limits that will enable this

Standard Error Method of Computing Confidence Interval

$$\bar{x} \pm SE * multiplier$$

- Information we need
- $\bar{x}$: mean from the *original* sample
- SE: standard deviation of mean of means of *bootstrap samples*
- *multiplier*: appropriate percentiles of standard normal distribution to cover 95% of the resamples

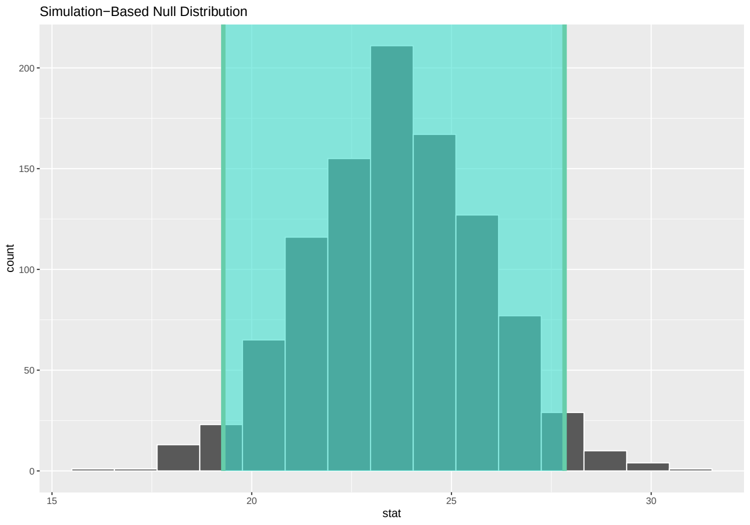
Calculation of Confidence Interval

- $\overline{x} = 23.56$
- $SE = 2.179$
- *multiplier* = approximately 2 (1.96 exactly)

```
lower_ci <- x_bar - (SE * 1.96)
upper_ci <- x_bar + (SE * 1.96)
```

- Confidence Interval is 19.29 to 27.83

Visualize Confidence Interval



Conclusion

- More flexible, interpretable tool to report inferences about population parameters
- Needs to be applied in situations where inference is being undertaken rather than simple description
    - Inference implies we are concerned about the nature of the distribution of values and where our sample data sit in an overall distribution
    - Description is simply describing what you measured

How to Deal with the CHIKV/DENV Problem

bank	n	percent
CHIKV	47	46.08%
DENV-1	28	27.45%
DENV-2	22	21.57%
Coinfect	5	4.90%
TOTAL	102	-

- To satisfy Reviewer 2, we can treat each of the categories as a proportion of the number of cases, which is similar to binomial (Yes/No, True/False, Heads/Tails) problems
- Use the Binomial distribution, which calculates proportions and translate the results back to numbers

The Final Table

```
##     bank cases totalcases proportion    low_num    hi_num
## 1   Chik    47        102 0.46078431 37.1331556 56.866844
## 2  coinf     5        102 0.04901961  0.7261539  9.273846
## 3 DENV-1    28        102 0.27450980 19.1662952 36.833705
## 4 DENV-2    22        102 0.21568627 13.8585017 30.141498
```

Bioinformatics at EPM/UNIFESP

- Center of Bioinformatics is in process of being established
- Desire to meet with as many laboratories as possible to find out
    - What kinds of techniques you are using
    - What you would like to be doing with data
    - What your frustrations with computation and biostatistics are
- My post-doc is focused this year on assisting in getting the Center up and running