Week 11

Foundations

Slides from Guerilla Section F17

Confidence Intervals

Recall:

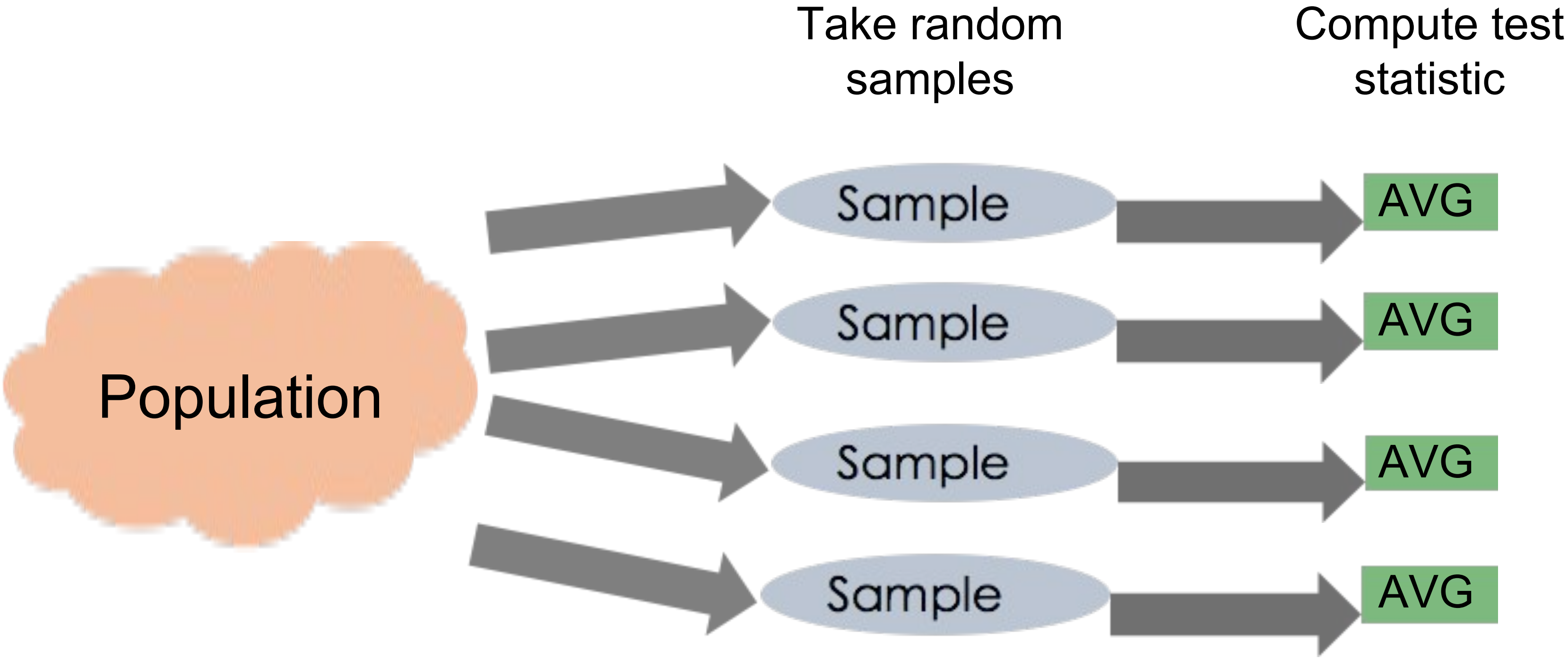
Parameter: a numerical value associated with a **population**

Statistic: any number computed using the data in a **sample**

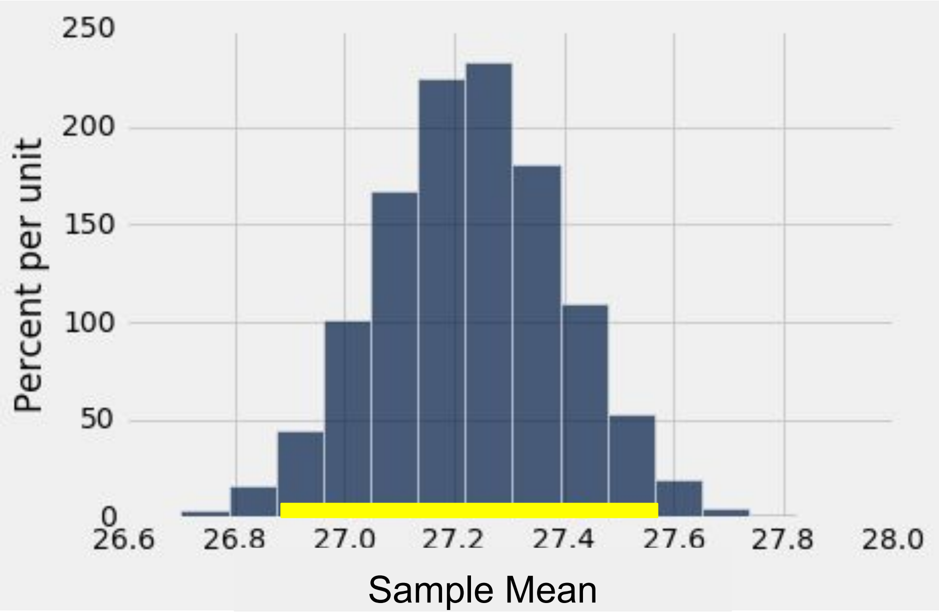
Use confidence intervals to estimate a parameter using a sample(s) from the population

How confident are we that our process generated the right estimate?

Confidence Intervals



Confidence Intervals



Create a histogram of the test statistic

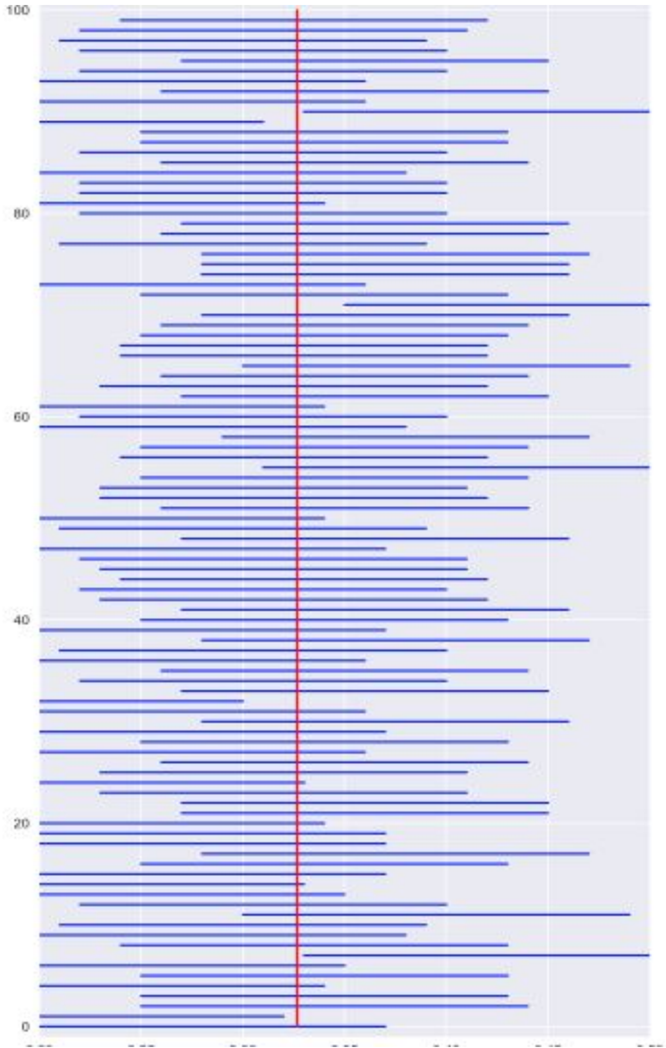
Cut off the outermost 2.5 percentiles on each side to create 95% confidence interval

What does this mean?

Confidence Intervals

How to interpret a 95% confidence interval:

We expect 95% of the confidence intervals we make will contain the true parameter



Confidence Intervals

Duality between Confidence Intervals and Hypothesis Testing

If we conduct a hypothesis test at the x% significance level, then the (1-x)% confidence interval contains all the values that we would fail to reject the null hypothesis for

Ex: We conduct a hypothesis test at the 5% significance level. If our observed statistic is not in the 95% confidence interval, we reject the null hypothesis.

Bootstrapping

- Ideally, we can draw as large of a sample / as many samples as we like, but sometimes that's not possible
  - Most of the time, we don't have access to the population data
  - Logistical problems: too expensive, too difficult, not enough time

- Instead, we use **bootstrapping**, which allows us to simulate the population we are drawing the sample from

Bootstrapping

Steps Conceptually

1) Treat the original sample as if it’s the population
2) Draw from sample, at random **with replacement**
   a) Every sample now can be **different**, but since it’s drawn from the same sample, will also resemble the original sample
3) This way, we create our own population dataset to test a hypothesis on
4) Every sample must be the same size as the original sample
   a) We want to keep variability the same since it depends on sample size

Bootstrapping

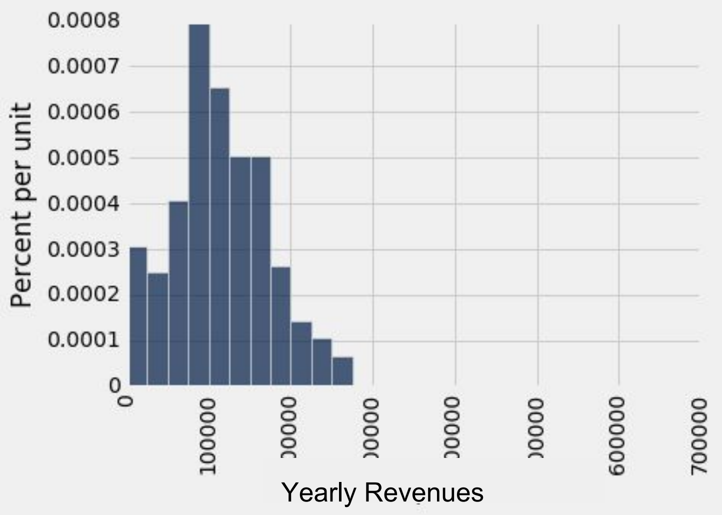
- Works because we assume the sampling method for the original sample is sound and lacks bias
- Law of Averages: By sampling more and more, we approach a similar looking distribution as the population
- Common Problem:
  - Sampling **without replacement**, you'll get the same sample every time, which defeats the purpose of simulating the whole population. It would turn out to be a simulated population distribution with the same ratio of data as the original sample

Bootstrapping Uses

- Create confidence intervals
- Run p-value tests
- You can do anything with a bootstrapped population distribution as you normally do, except you have to create your own distribution.

Bootstrapping Example

Suppose we want to see median yearly revenues for a group of street vendors. We have a sample size of 500 street vendors. Create a 95% confidence interval to estimate the population median.



Step 1 Analyze the Problem

We don’t have a bigger sample, nor do we have the population parameter. So, we use bootstrapping to simulate trials to build a confidence interval.

Step 2 Code it

```
def bootstrap_median(original_sample, label, replications):
    """Returns an array of bootstrapped sample medians:
    original_sample: table containing the original sample
    label: label of column containing the variable
    replications: number                ap samples
    """
    just_one_column = orig            ple.select(lab
    medians = make_array()
    for i in np.arange(replications
        bootstrap_sample = just_o              ple()
        resampled_median = percenti          otstrap_sample.column(0))
        medians = np.append(medians, resampled_median)

    return medians
bstrp_medians = bootstrap_median(street_vendors, 'revenue',
```

Empty array

Sample with_replacement

Find median

Add to medians array

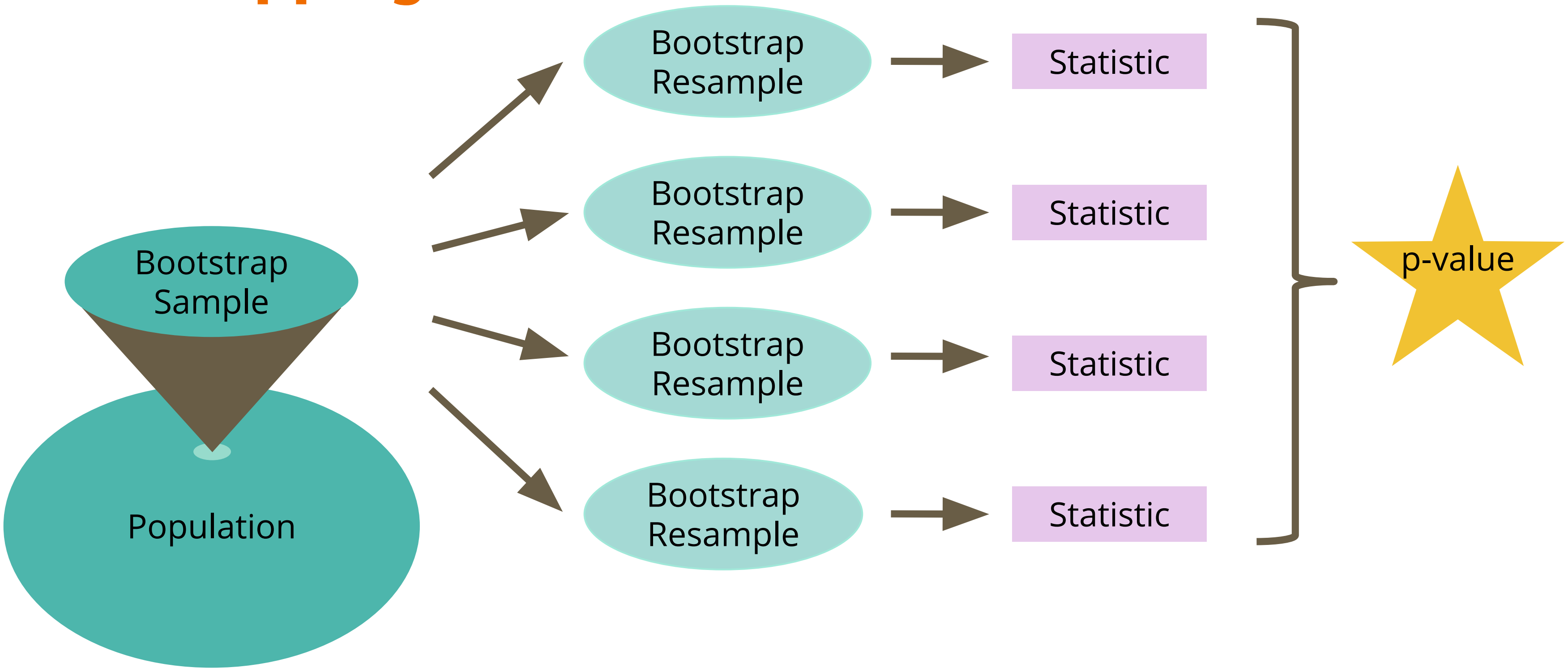
Step 3 Construct the Interval

```
left = percentile(2.5, bstrap_medians)
left
107652.71000000001
```

```
right = percentile(97.5, bstrap_medians)
right
119256.73
```

```
**Percentile function: percentile (percent, array)
```

Bootstrapping



Bootstrap Sample

Population

Bootstrap Resample

Bootstrap Resample

Bootstrap Resample

Bootstrap Resample

Statistic

Statistic

Statistic

Statistic

p-value

Fall 2016 Final Problem

12. In a large random sample of U.S. households, the median annual income is $54,000. This original sample is bootstrapped 5,000 times and the sample median is recorded for each of the bootstrap samples. The middle 95% interval of these values is ($53,000, $55,000).

(a) True or false (explain your answer):

The interval ($53,000, $55,000) is an approximate bootstrap 95% confidence interval for the median income of all the households in the sample.

Fall 2016 Midterm Problem

(b) Pick the option that you think best completes the sentence, and explain your choice.

The percent of all U.S. households with annual incomes in the range ($53,000, $55,000)

(i) is about 95%. (ii) is about 50%. (iii) cannot be approximated based on the information given.

Fall 2016 Midterm Problem

(c) Pick the option that you think best completes the sentence, and explain your choice.

If you calculate the mean of each of the 5,000 bootstrap samples and take the middle 95% interval of the 5,000 means, the center of the new interval will be

(i) less than $54,000. (ii) about $54,000. (iii) more than $54,000.