A Computational Approach to Hypothesis Testing

Embezzlement of Parking Meter Tolls

By Brandon Rozek

Background

- The county hires a company to collect parking meter tolls from a subset of parking meters throughout the county.
- The company reports to us how much money they receive from each toll.

Suspicion

- An accountant looking through the records suspects that a contracting company is embezzling parking meter funds by not reporting the full amount.
- How can we use their reportings and compare it to ours to provide evidence of embezzlement?

We Need Statistics

- We don't know what funds the company would have reported had they collected from our subset.
- If we did, then we could compare it to our own data and see if there's a difference.
- Statistics will help us avoid making bold claims of embezzlement if we have a decent chance of being wrong.

Caveats/Misconceptions

Keeping you from making mistakes

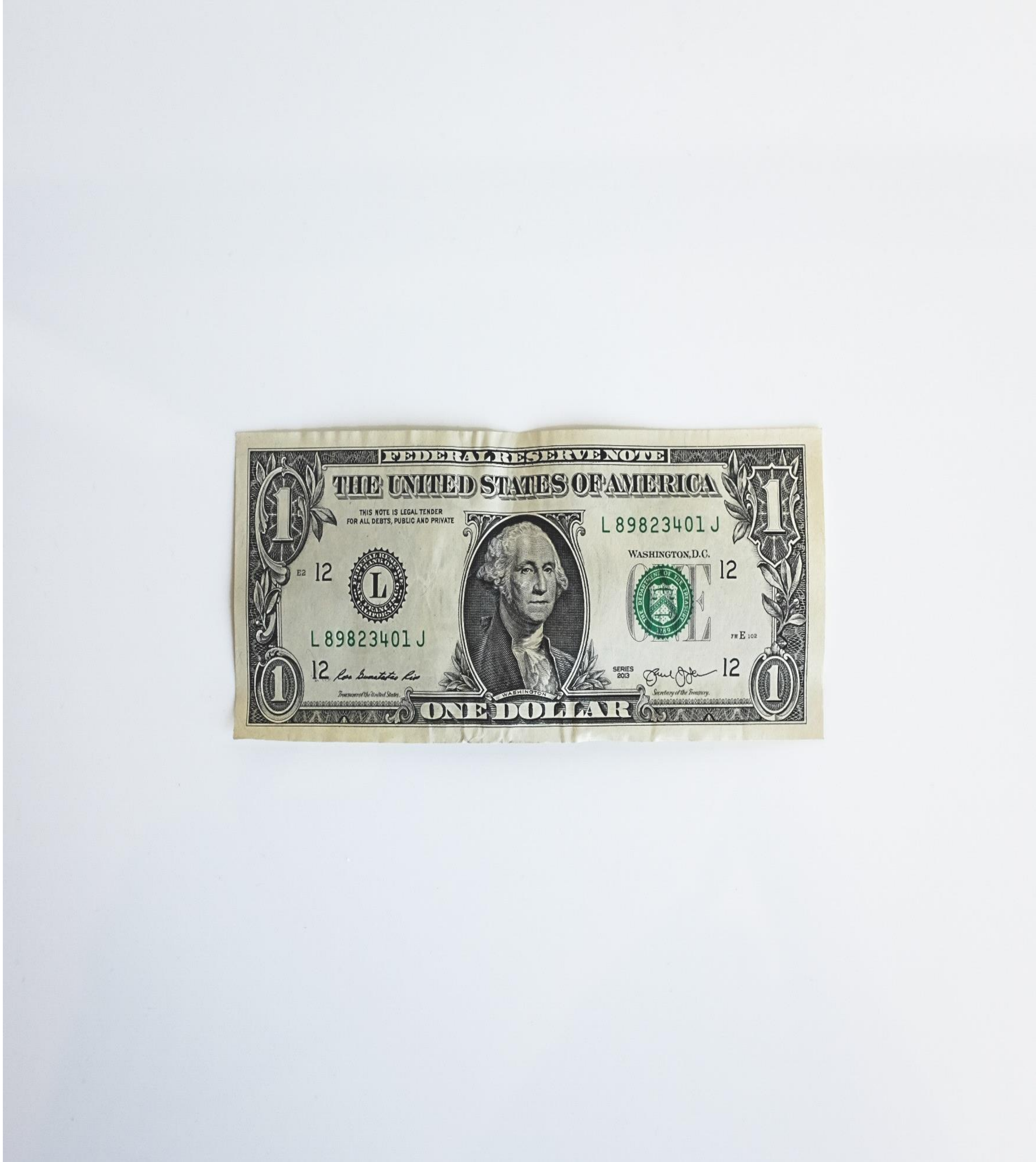
Sample Statistics vs Population Parameters

- If both parties were able to report the collection of funds at each parking meter, then we wouldn't need Statistics.
- We just need to see the difference in the amounts reported.
- But it's because we both collected from different meters that we don't have the full picture.

Statistical Significance
vs
Practical Significance

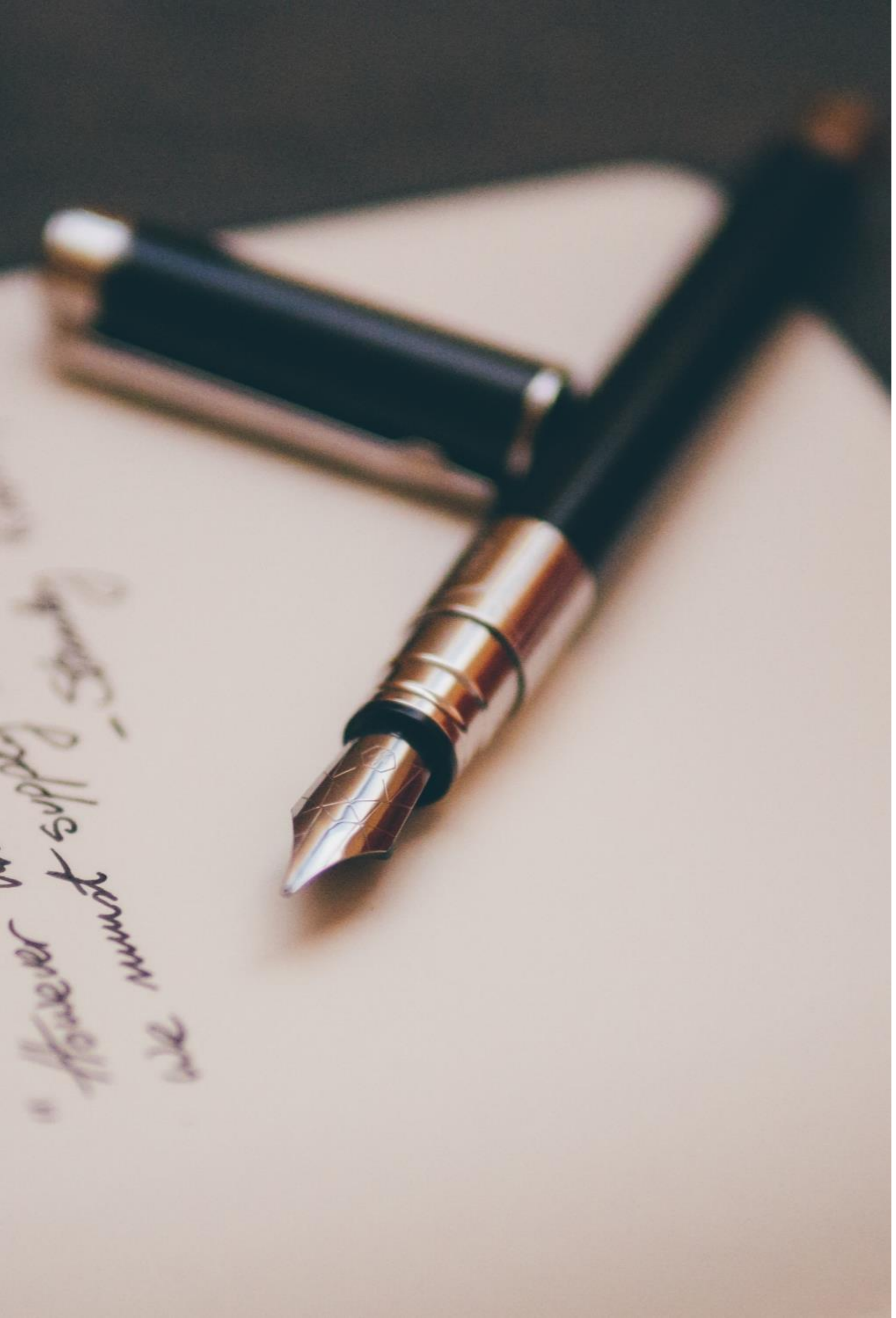
- From a statistical test, we can conclude that the company is reporting less and we see a $100 difference between the county and the company.
- Is it really embezzlement if we see a difference of $100?

It costs more to go to court!



Introduction to Hypothesis Testing

Uneducated guessing is not enough



Defining the case

Before we go straight into testing, we need to establish what we can test.

<u>One Sample Test</u>

You suspect something to be different than what was previously believed.

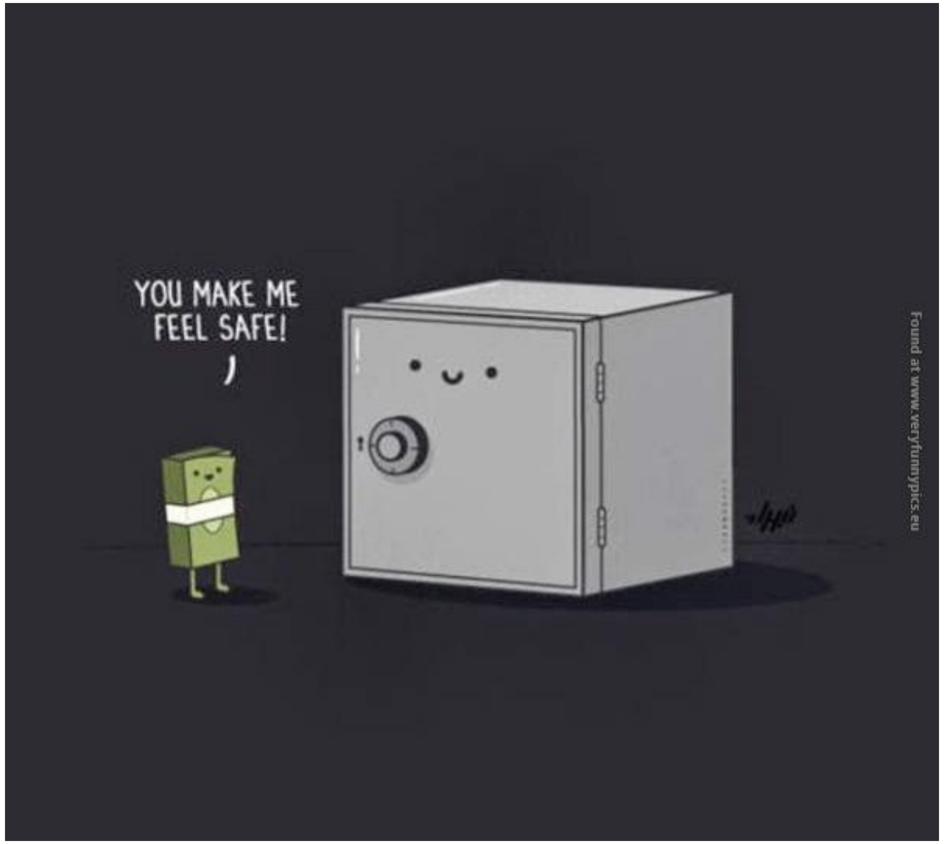
<u>Two Sample Test</u>

You suspect that a parameter, such as the average, is significantly difference between two datasets.

In our case, we suspect that the company collects on average significantly less parking meter funds than the county.

The Safe Default Known as the Null Hypothesis

- In the US Court and in Hypothesis Testing, the company is innocent until proven guilty.
- The null hypothesis is that there is no significant difference between the amounts that the company reported and the amounts that the county reported.



The Suspicion of the Alternative Hypothesis

- This is where the fun begins. If we don't have strong enough evidence for the null hypothesis, then we favor the alternative.
- The alternative hypothesis can come in the following forms:
  - You can say that the average is different. (Two-tailed)
  - The average reported by the company is less than the county's or vice versa. (One-tailed)

- We are trying to prove that the company is embezzling funds. That is that the average amount reported by the company is significantly less than the average amount reported by the county. (One-tailed)

Why this approach?

We're creating a framework of two competing hypotheses.

Due to innocence until proven guilty, we're going to simulate the situation of non-embezzlement with our data.

Then we're going to ask "Is the difference observed likely under the null hypothesis?"

If it's not, then we're going to reject the notion that they're innocent.

If it is, then we have failed to reject the null hypothesis and have to assume they're innocent.

Performing the test

Bootstrapping our results

It all starts off in their differences

- From the data we can compute the differences between the average reported by the company and the county.

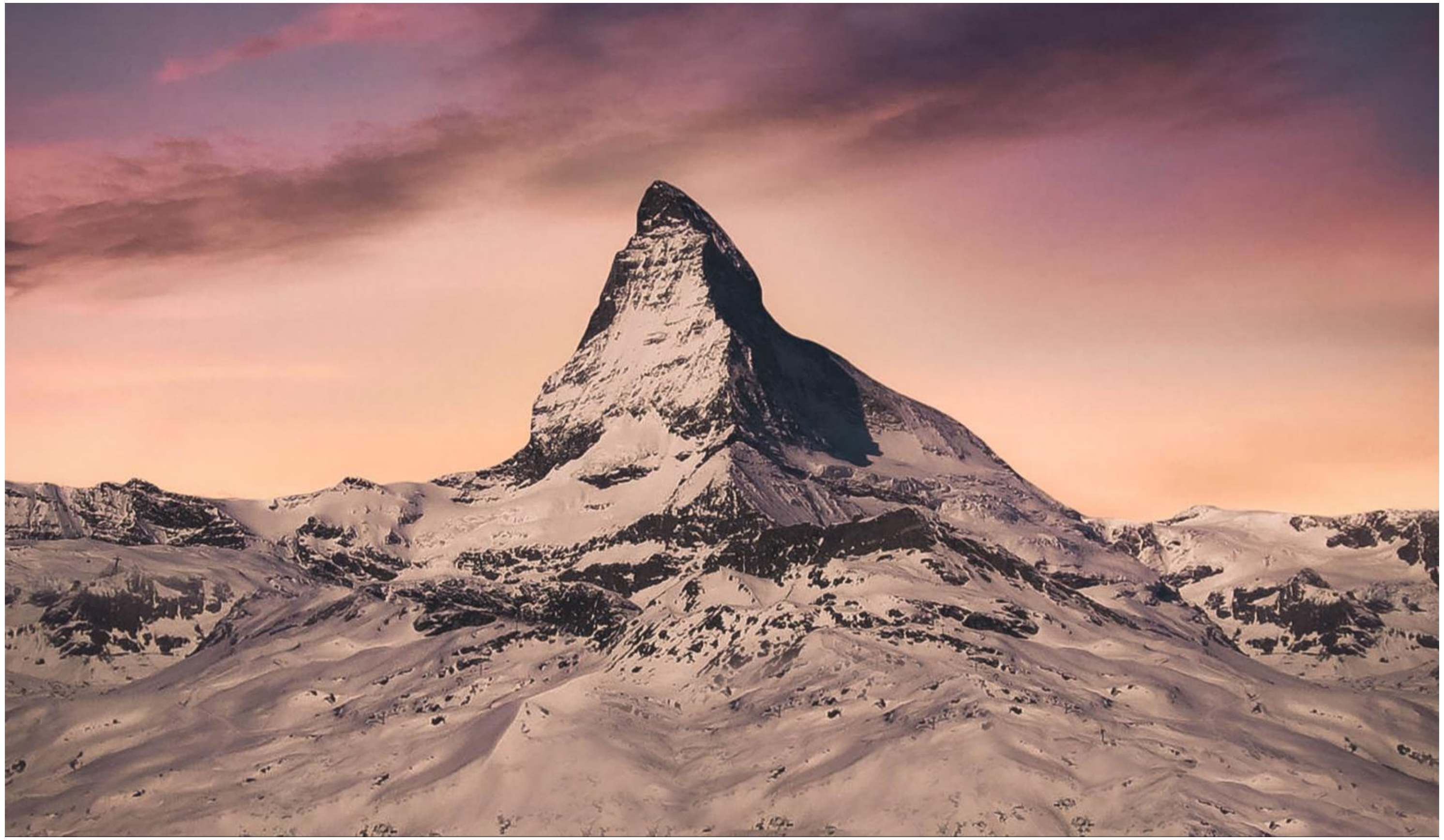
```
observed_difference = mean(company_data) - mean(county_data)
```

- Is the fact that the company reported on average less due to chance or is there actually some truth in the possibility of embezzlement?

Simulating the Null Hypothesis

- We're going to take the company's and county's data and subtract their respective means and add their pooled mean.
- This way we can have both of the datasets shifted to have the same mean which is the null hypothesis.

```
combined_data = c(county_data, company_data)
pooled_mean = mean(combined_data)
county_shifted = county_data - mean(county_data) + pooled_mean
company_shifted = company_data - mean(company_data) + pooled_mean
```



The Bootstrap Distribution

- We can't assume what each group would have collected in the other subset.
- But we just made two datasets that represent what could have been collected had they both had the same average.

Generating the Bootstrap Sample

- Let's sample from these datasets with replacement, which means we can report the same parking meter multiple times.
- Then we recompute the average toll funds collected for both the county and the company and find the difference between those two.

Generating Bootstrap Sample Code

```
CalculateStatistic = function(company_shifted, county_shifted) {
    # Get Number of Observations
    num_county_observations = length(county_shifted)
    num_company_observations = length(company_shifted)
    # Sample with Replacement
    county_bootstrap_sample = sample(county_shifted,
                                     num_county_observations,
                                     replace = TRUE)
    company_bootstrap_sample = sample(company_shifted,
                                      num_company_observations,
                                      replace = TRUE)
    # Find Different Observed
    return (
        mean(company_bootstrap_sample) - mean(county_bootstrap_sample)
    )
}
```

Bootstrap Replicants

- Now it doesn't make sense to only do this once. You can't get any reliable measurements this way.
- Instead, we will do this technique a large number of times. (like 10,000)
- That way, we have 10,000 different differences from the world of no embezzlement to look at.

Generating Bootstrap Replicants Code

```
number_trials = 10000
replicates = numeric(number_trials)
for (i in 1:number_trials) {
    replicates[i] = CalculateStatistic(company_shifted, county_shifted)
}
```

What are we getting out of this simulation?

Recall that our original goal is to state whether or not the data we have is likely under the null hypothesis. We can reject the null hypothesis if...

<u>One Sided Greater/Less Alternative</u>

- The difference in the bootstrap replicants is greater/less than the difference observed no more than 5% of the time.

<u>Two Sided Test</u>

- The absolute difference of the bootstrap replicants are greater than the absolute value of the difference observed no more than 5% of the time.

What is a P-Value?

- The p-value will tell us the probability of obtaining the difference observed or even more embezzlement if the null hypothesis was true.

- For us to go into court, we want to know that there is less than a 5% chance of the difference seen occurring if they were not embezzling.

Code for Calculating Empirical P-Values

```
p.value = sum(replicates <= observed_difference) / number_trials
```

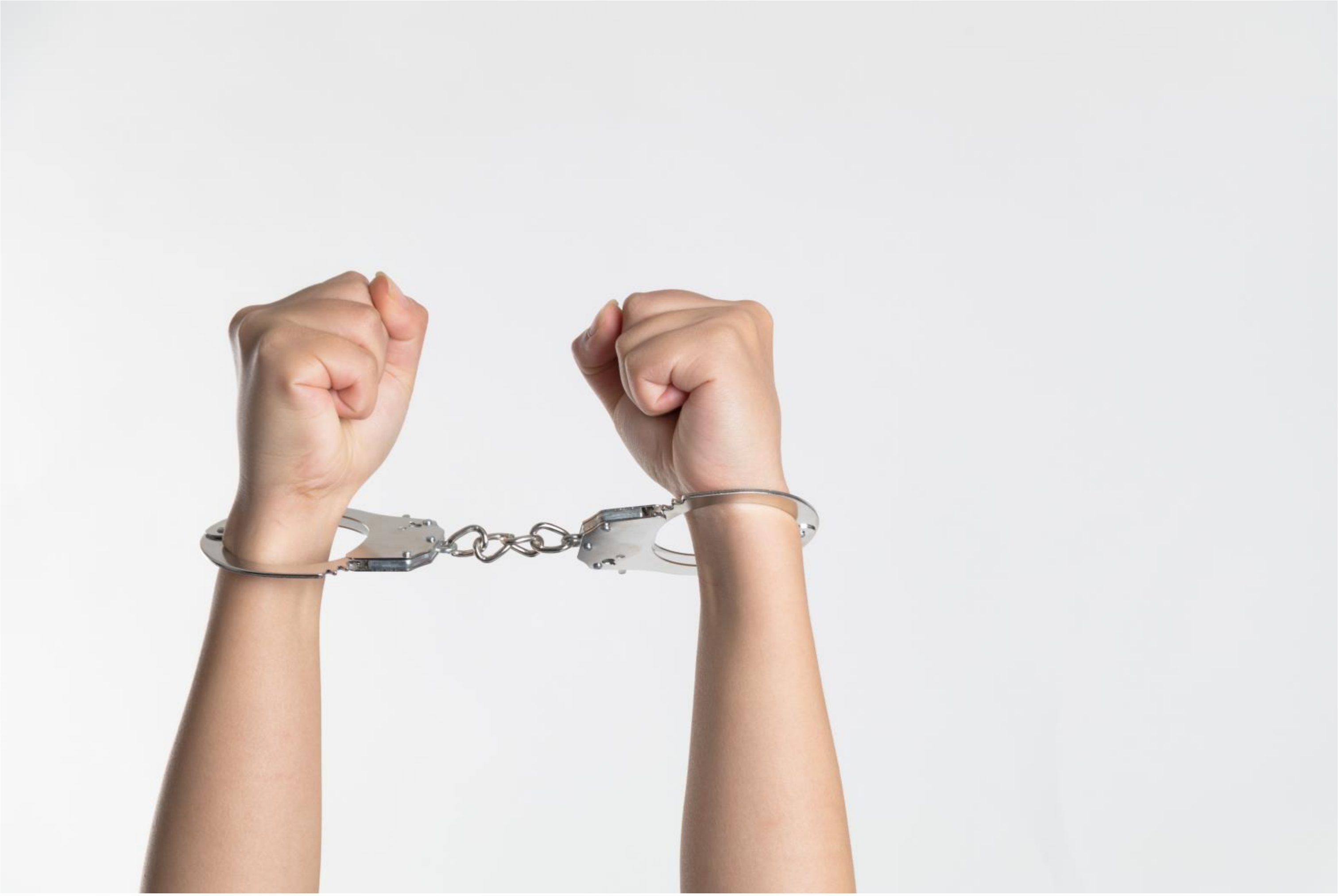
Drawing Conclusions

Getting meaning out of the numbers

Parking Meter Conclusion

- We reject the notion that the average amount reported by the company is equal to the average amount reported by the county.
- It is highly likely that the company is embezzling funds.

Caught



Resources to Learn More

- You can contact me at my email: brozek@mail.umw.edu

- Datacamp Module "Statistical Thinking in Python Part 2"

- You can search the Internet for
  - Resampling
  - Bootstrap Testing
  - Permutation Tests

Unsplashed Images (In Order of Appearance)

- San Francisco 1.4 by Pine Watt

https://unsplash.com/photos/3_Xwxya43hE

- Sign, Coffee, Café, and Coffee Shop by Nathan Dumlao

https://unsplash.com/photos/pULHq4kr7DQ

- Man with Wallet and Laptop by Artem Beliaikin

https://unsplash.com/photos/e--geRd5eCQ

- Statistics on a Laptop by Carlos Muza

https://unsplash.com/photos/hpjSkU2UYSU

- Cash Money by NeONBRAND

https://unsplash.com/photos/8fDhgAN5zG0

More Unsplashed Images

- Going to New York by Steve Richey
https://unsplash.com/photos/wVK0ypTn61Y
- Fountain Pen on Stationary by Alvaro Serrano
https://unsplash.com/photos/hjwKMkehBco
- Mountain, Sky, View, and Nature by Uroš Jovičić
https://unsplash.com/photos/euaQi-5dTj0
- A hand holding baby tomatoes by Chinh Le Duc
https://unsplash.com/photos/EEuC0Q6g7WY
- Row of Shoes by Dương Trần Quốc
https://unsplash.com/photos/ngKbU5tb2m0
- Handcuffed Criminal by Niu Niu
https://unsplash.com/photos/5HzOtV-FSlw