An Introduction to Computational Statistics

Workshop by Brian Spiering, Professor of CS/DS at University of San Francisco

(March 10, 2019)

ENTERPRISE ANALYTICS & DATA

Two Perspectives: Analytical and Computational

Analytical Methods	Computational Methods
Based on Mathematics (Statistics, Probability)	Based on Computer science (software engineering)
Use domain and theoretical knowledge	Use fundamental computing (i.e. looping and random number generator)
Made sense when data collection and computing power were expensive	Make sense because data collection and computing is cheaper
	Using computers we can also solve novel problems
Main tool: probability, statistical tests	Main tool: simulation

Computational Statistics by Example

An Introduction to Computational Statistics

ENTERPRISE
ANALYTICS & DATA

- **Example 1:** Simulating outcomes
  - Efficient simulation of probabilities of common events and statistics
- **Example 2:** Permutation testing
  - Hypothesis (A/B) testing
- More in the workshop Jupyter notebooks
  - Computational approach to linear regression
  - Permutation tests for median differences
  - Inspection paradox
  - Additional resources

Simulating Outcomes

An Introduction to Computational Statistics

ENTERPRISE
ANALYTICS & DATA

Question: If you roll two 6-sided dice,

how likely is it that the sum is greater than 7?

Analytical Method

ENTERPRISE ANALYTICS & DATA

If you roll two 6-sided dice, how likely is it that the sum is greater than 7?

- We have 6^2 potential outcomes (variations with repetition)
- In total 36 pairs of (result of first roll, result of second roll)
- (1, 1), (1, 2), (1, 3), ..., (6, 6) -> how many with the sum greater then 7?

- (2, 6) -> 1
- (3, 5), (3, 6) -> 2
- (4, 4), (4, 5), (4, 6) -> 3
- (5, 3), ..., (5, 6) -> 4
- (6, 2), ..., (6, 6) -> 5

} 15 pairs with sum > 7, so the probability is 15/36 = 0.4166(6)

- What about: If you roll seven 6-sided dice, how likely is it that the sum is greater than 35?
- 6^7 = 279936 (variations)

Computational Method

If you roll two 6-sided dice, how likely is it that the sum is greater than 7?

ENTERPRISE
ANALYTICS & DATA

[1]:
```
from random import choices
from collections import Counter

# 6-sided die
faces = list(range(1, 7))

# A function for rolling a pair of dice
roll_2_dice = (lambda: choices(population=faces, weights=None, k=2))

# Simulate rolling a pair many times and tracking the outcomes
rolls = [sum(roll_2_dice()) for _ in range(50_000)]

# Count those outcomes
rolls_counts = Counter(rolls)

# If you roll two dice, how likely is it that your sum is greater than 7
sum(v for k, v in rolls_counts.items() if k > 7) / sum(rolls_counts.values())
```

[1]:
```
0.41468
```

Analytical vs. Computational

Simulating outcomes approach comparison

ENTERPRISE
ANALYTICS & DATA

- Analytical approach
  - Think about all possible outcomes
  - Enumerate outcomes which meet the condition (>7)
  - Count the outcomes which meet the condition and calculate the likelihood
  - Have problem with the same approach for more complex scenario

- Computational approach
  - Define 6-sided die
  - Simulate rolling a pair of dice (many times) -> how many times should suffice?
  - Count and calculate the likelihood
  - Bonus: I can count, sort, visualize easily
  - No problem with more complex scenario

Permutation Testing

An Introduction to Computational Statistics

ENTERPRISE
ANALYTICS & DATA

Question: Is a drug effective or not?

```
drug = [54, 73, 53, 70, 73, 68, 52, 65, 65]
placebo = [54, 51, 58, 44, 55, 52, 42, 47, 58, 46]
```

Analytical Method

Is a drug effective or not?

ENTERPRISE ANALYTICS & DATA

Is a drug effective or not?

[1]:
```
drug = [54, 73, 53, 70, 73, 68, 52, 65, 65] # treatment group
placebo = [54, 51, 58, 44, 55, 52, 42, 47, 58, 46] # control group
```

$$H_0 : \mu_{drug} - \mu_{placebo} = 0$$
$$H_1 : \mu_{drug} - \mu_{placebo} <> 0$$

Where:

H_0 -> **The null hypothesis:** there is no difference in the drug and placebo groups on average (the drug does not have any effect).

H_1 -> **The alternative hypothesis:** there is a significant difference in the drug and placebo groups on average (the drug works).

I will assume **type I error (alpha)** level at 1%.

[2]:
```
alpha = 0.01
```

Analytical Method

https://github.com/ksatola/Computational-Statistics/blob/master/Example2.html

Computational Method

Is a drug effective or not?

ENTERPRISE
ANALYTICS & DATA

[26]:
```
# Simulate it a bunch of times
n = 10_000
count = 0
simulated_means = []

for _ in range(n):
    shuffle(combined)
    shuffled_diff = mean(combined[:len(drug)]) - mean(combined[len(drug):])
    simulated_means.append(shuffled_diff)
    count += (shuffled_diff >= observed_diff)
```

[27]:
```
print(f"""{n:,} label reshufflings produced only {count} instances
with a difference at least as extreme as the observed difference of {observed_diff:.2f}.""")
```

```
10,000 label reshufflings produced only 9 instances
with a difference at least as extreme as the observed difference of 12.97.
```

[28]:
```
# The p-value is a chance of observing the current difference when there is truly no difference
p = count / n
p
```

[28]:
```
0.0009
```

https://github.com/ksatola/Computational-Statistics/blob/master/Example2.html

Analytical vs. Computational

Hypothesis Testing approach comparison

ENTERPRISE
ANALYTICS & DATA

- Analytical approach
  - Setup testing framework
    - Define null and alternative hypothesis
    - Set alpha (the threshold for rejecting differences)
  - Collect data
  - Pick a sampling distribution, calculate a test statistic, remember about assumptions
  - Calculate p-value
  - Draw conclusion

- Computational approach
  - Setup testing framework
    - Define null and alternative hypothesis
    - Set alpha (the threshold for rejecting differences)
  - Collect data
  - Iterate while shuffling data to simulate null effect and create a sampling distribution
  - Calculate p-value
  - Draw conclusion

Advantages of Computational Approach

Hypothesis Testing approach comparison

ENTERPRISE ANALYTICS & DATA

Analysis	Simulation
Often dictates the test statistic, and tests have assumptions If you have issues like: • censored data, • non-independence, • and long-tailed distributions, you won't find an off-the-shelf test Unless you are a mathematical statistician, you won't be able to make it by yourself	There are no test assumptions to satisfy These are not issues when simulating