CompSci 116: Lab 6: Resampling & Bootstrap

Jeff Forbes
March 7, 2019
Duke University

Machine Learning Day

Women in Data Science

Sat March 23, Schiciano Auditorium

Register now at dukeml.org/register

HOSTED BY DUKE UNDERGRADUATE MACHINE LEARNING
The 80th percentile is the value in a set that is at least as large as 80% of the elements in the set.

For $s = [1, 7, 3, 9, 5]$, \( \text{percentile}(80, s) \) is 7.

The 80th percentile is ordered element 4: \( (80/100) * 5 \).

For a percentile that does not exactly correspond to an element, take the next greater element instead.
The percentile Function

- The $p$th percentile is the value in a set that is at least as large as $p\%$ of the elements in the set.
- Function in the `datascience` module:
  ```python
  percentile(p, values)
  ```
- $p$ is between 0 and 100.
- Returns the $p$th percentile of the array.

(Readiness Assurance)
How many enemy planes?
Assumptions

- Planes have serial numbers 1, 2, 3, ..., N.
- We don’t know N.
- We would like to estimate N based on the serial numbers of the planes that we see.

The main assumption

- The serial numbers of the planes that we see are a uniform random sample drawn with replacement from 1, 2, 3, ..., N.
Discussion question

If you saw these serial numbers, what would be your estimate of $N$?

```
170 271 285 290 48
235 24 90 291 19
```

One idea: 291. Just go with the largest one.
The largest number observed

- Is it likely to be close to N?
  - How likely?
  - How close?

**Option 1.** We could try to calculate the probabilities and draw a probability histogram.

**Option 2.** We could simulate and draw an empirical histogram.
Verdict on the estimate

- The largest serial number observed is likely to be close to $N$.
- But it is also likely to underestimate $N$.

Another idea for an estimate:
Average of the serial numbers observed $\sim \frac{N}{2}$

New estimate: 2 times the average

(Lab)
Inference: Estimation

- How big is an unknown parameter (e.g., number of planes)?

- If you have a census (that is, the whole population):
  - Just calculate the parameter and you’re done

- If you don’t have a census:
  - Take a random sample from the population
  - Use a statistic as an estimate of the parameter
Variability of the Estimate

- One sample $\rightarrow$ One estimate
- But the random sample could have come out differently
- And so the estimate could have been different
- Main question:
  - How different could the estimate have been?
- The variability of the estimate tells us something about how accurate the estimate is:
  \[
  \text{estimate} = \text{parameter} + \text{error}
  \]
Where to Get Another Sample?

- One sample ➔ One estimate
- To get many values of the estimate, we needed many random samples
- Can’t go back and sample again from the population:
  - No time, no money
- Stuck?
The Bootstrap

- A technique for simulating repeated random sampling

- All that we have is the original sample
  - ... which is large and random
  - Therefore, it probably resembles the population

- So we sample at random from the original sample!
Why the Bootstrap Works

population

sample

resamples

All of these look pretty similar, most likely.
Key to Resampling

- From the original sample,
  - draw at random
  - with replacement
  - as many values as the original sample contained

- The size of the new sample has to be the same as the original one, so that the two estimates are comparable