CSC 439/539
Statistical Natural Language Processing
Statistical Significance: Bootstrap Resampling

Mihai Surdeanu

Fall 2017

Example: Is a New Battery Really Better?

- A company has developed a new battery for use in electric cars. They believe that the new battery will allow cars to drive farther before needing to be recharged.
- The distribution of battery life in miles-to-recharge is positively skewed.
- Since the distribution is positively skewed, it might make sense to do a test of the *median*, rather than the mean.
- The median battery life for the current state-of-the-art version is 300 miles.
- What would our null and alternative hypotheses be?

Example: Is a New Battery Really Better?

- We are interested in testing

$$H_0 : \theta \leq 300 \qquad\qquad H_1 : \theta > 300$$

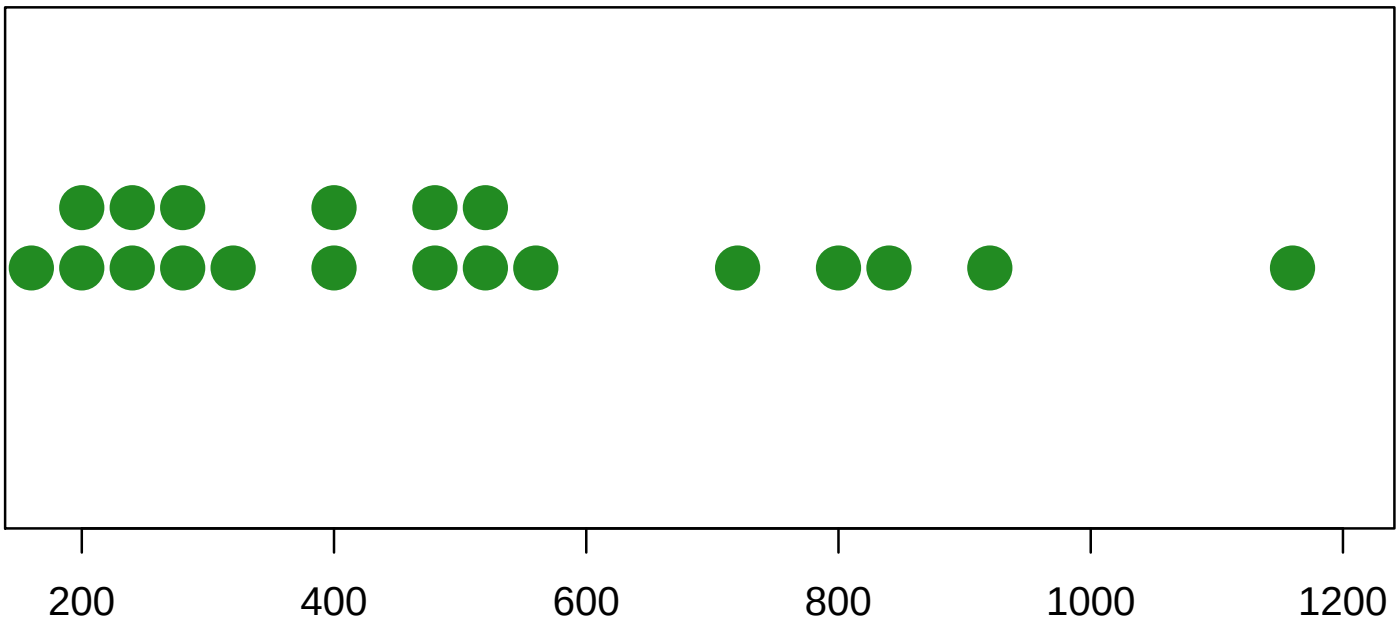
  where θ is the *population* median for the battery life distribution for the new model.
- To test this, we need a random sample of batteries from the population. We can then use the sample median, Q_2, to estimate the population median:
  - $Q_2 = \hat{\theta} \approx \theta$ (the "hat" stands for estimate)
- But because **we only have a sample**, we will be off to some extent. How can we find out the distribution of Q_2?

It Can Be Tricky

- We have no clue what the population distribution is (**same in NLP**). We know it's right skewed, so it's not normal.
- Our samples are small (evaluating batteries is expensive). – **same in NLP**

Example: Is a New Battery Really Better?

All we have is a sample of 20 batteries:



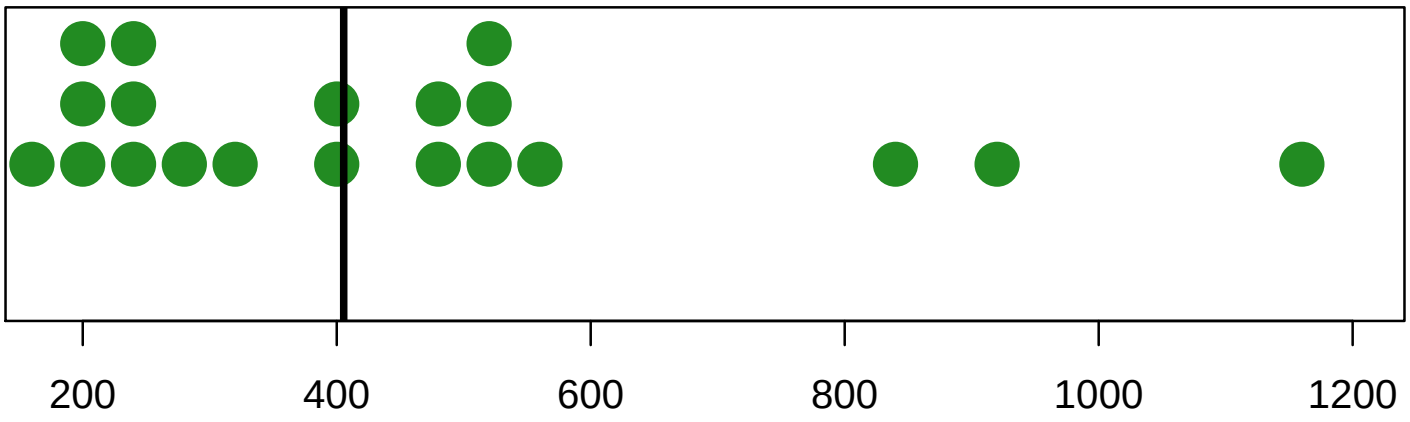
Bootstrap Resampling Intuition

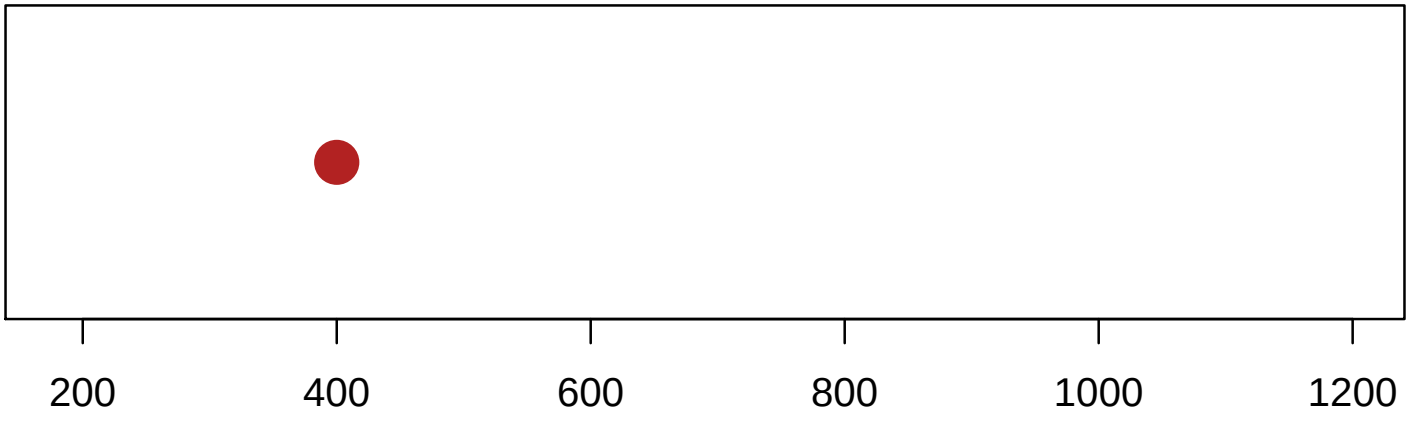
- Ideally, we would like to draw many samples from the population (of all batteries in the world) to construct the distribution of our statistic.
- But we do not have access to the entire population.
- Next best thing: **treat the sample we have as the population**. That is, draw samples from the sample.

Bootstrap Resampling Intuition

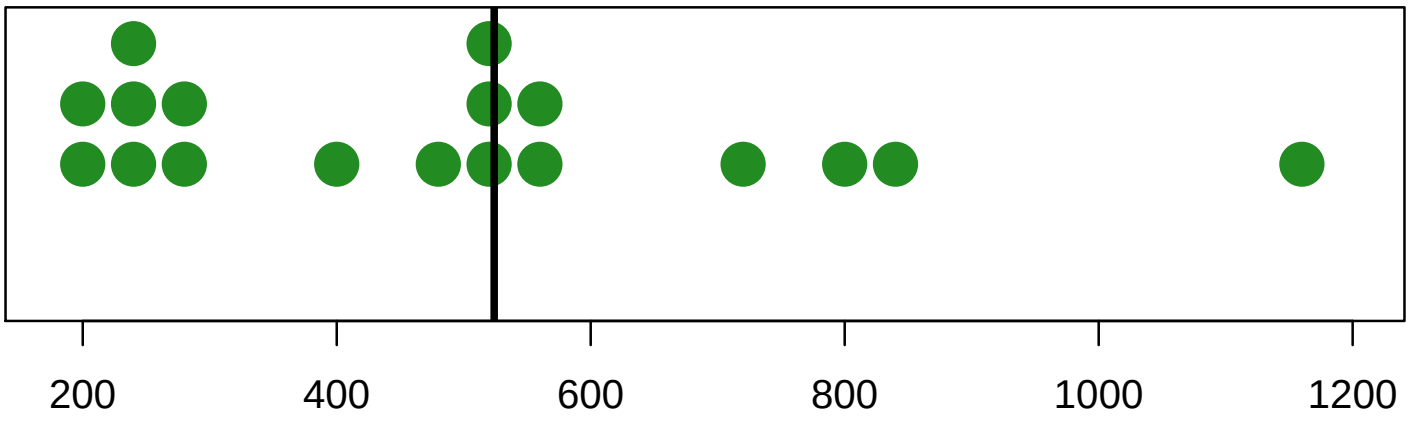
- We want our draws to be *independent* (to minimize dependencies to our sample), though, so what should we do?
- We need to sample *with replacement*, treating the sample as the population.
- If we do this repeatedly, computing our sample statistic each time, we can construct an estimated sampling distribution for that statistic.
- This procedure is known as (non-parametric) **bootstrap resampling** (or just "bootstrapping" for short), because we are using the data to "pull ourselves up by our bootstraps".

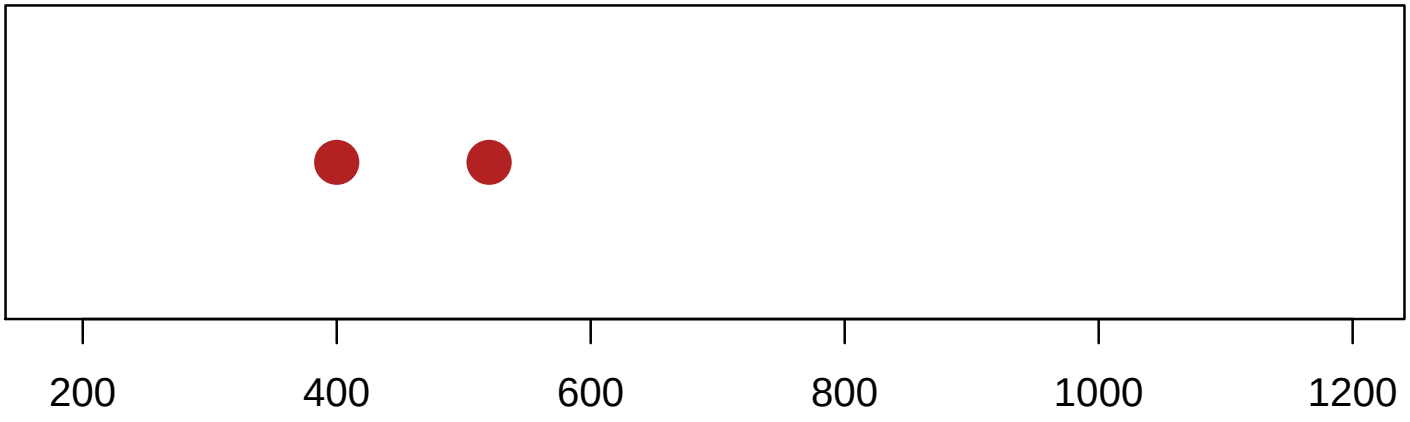
Bootstrap Resampling



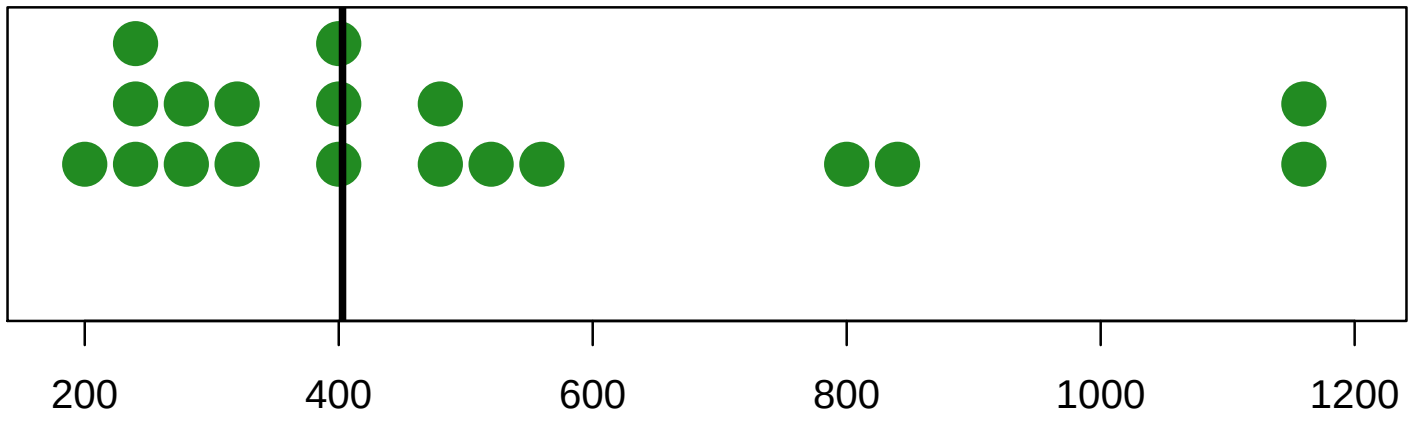


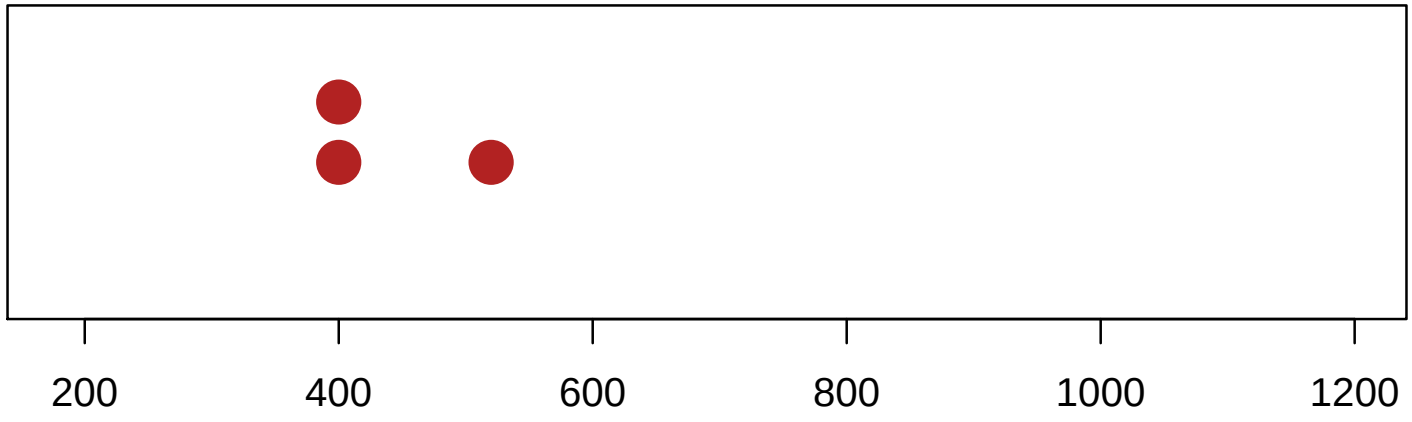
Bootstrap Resampling



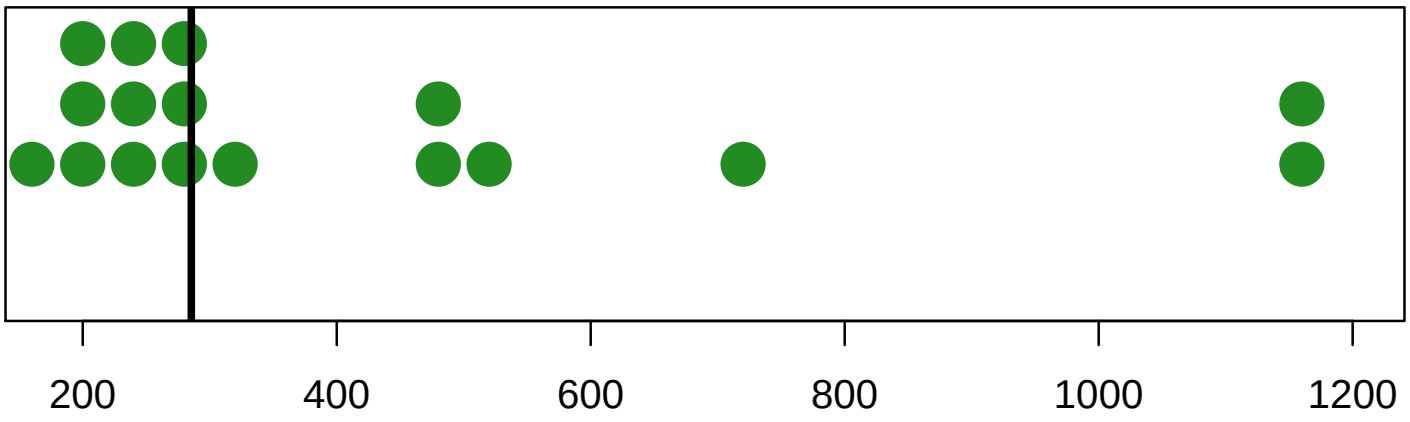


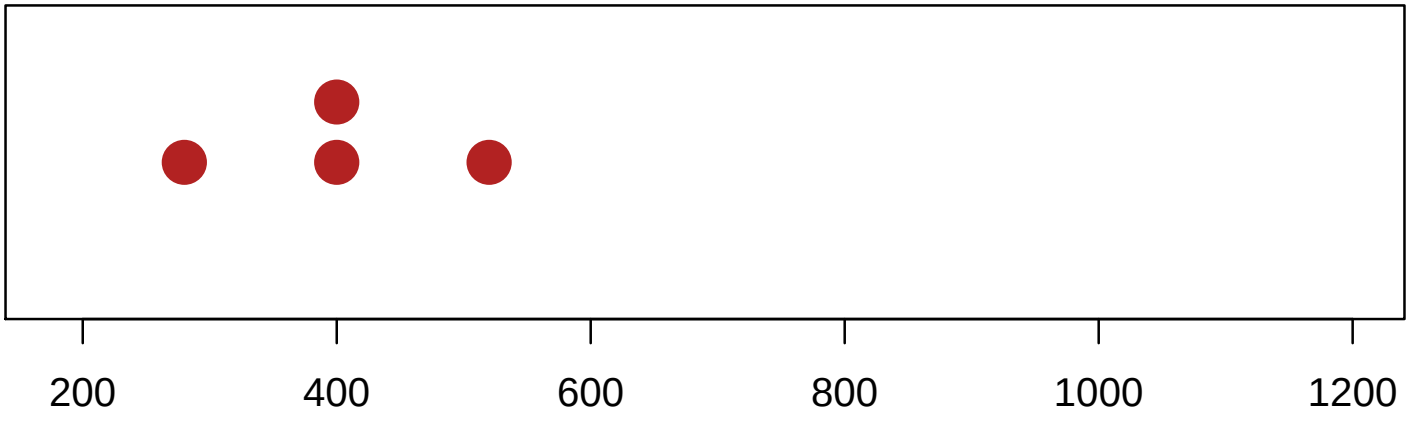
Bootstrap Resampling



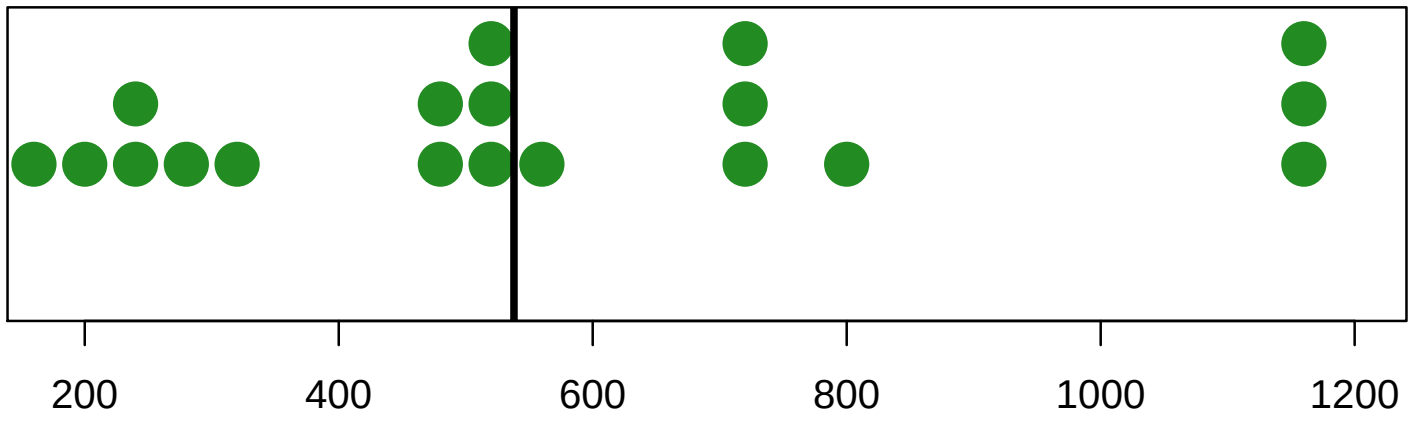


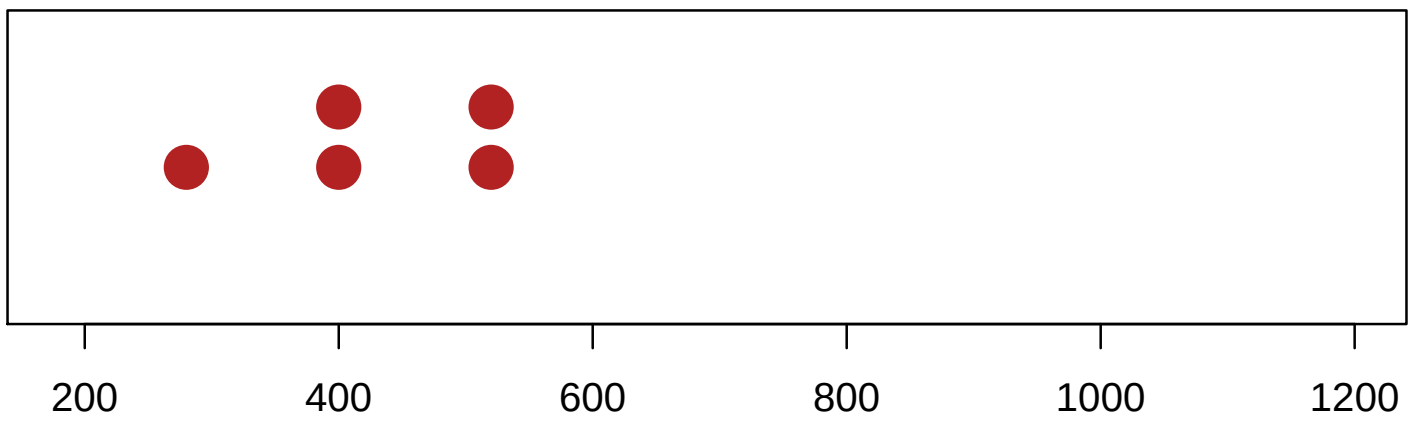
Bootstrap Resampling



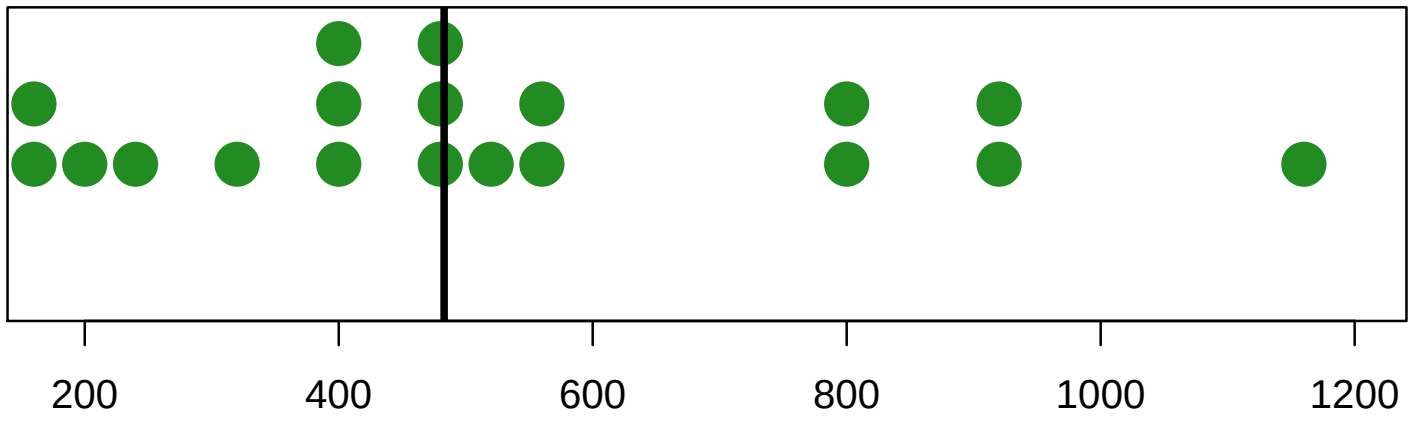


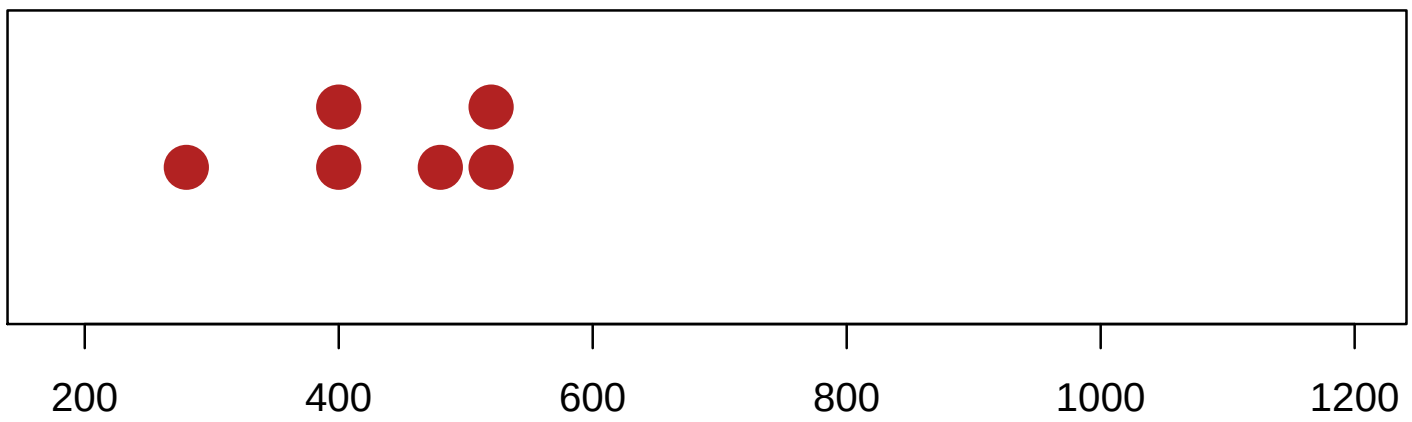
Bootstrap Resampling



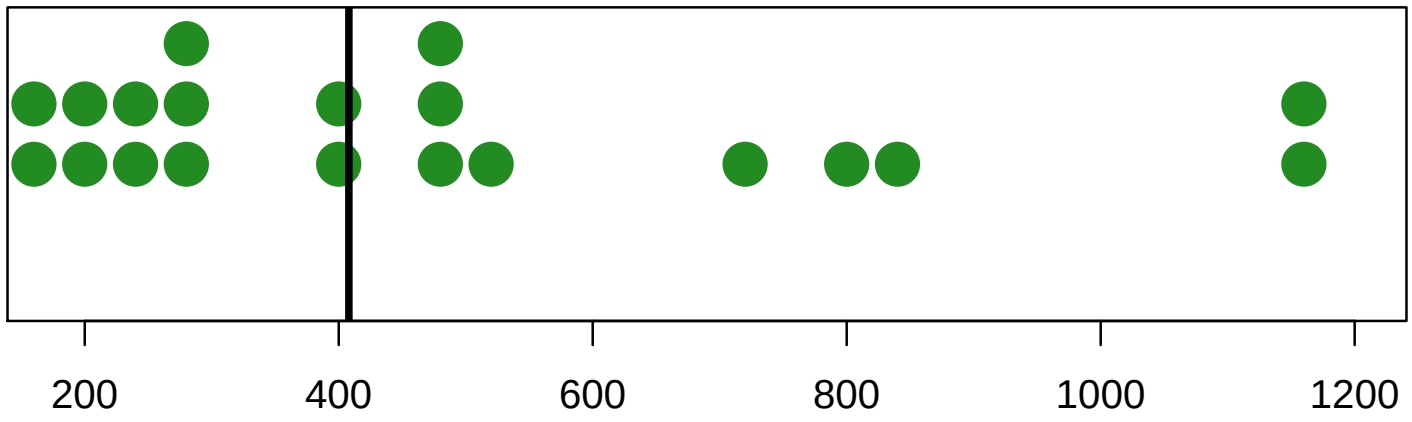


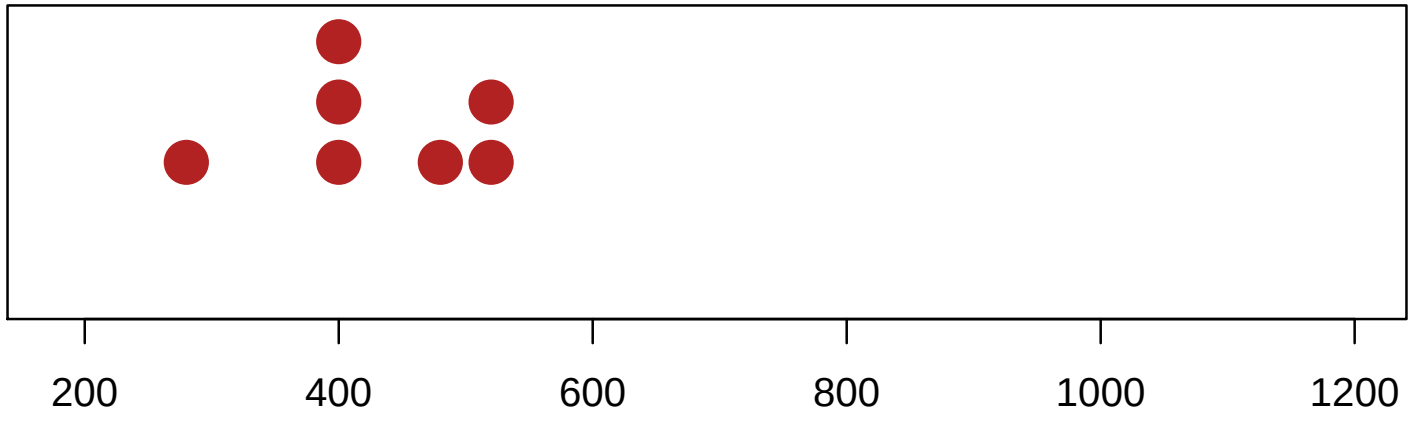
Bootstrap Resampling



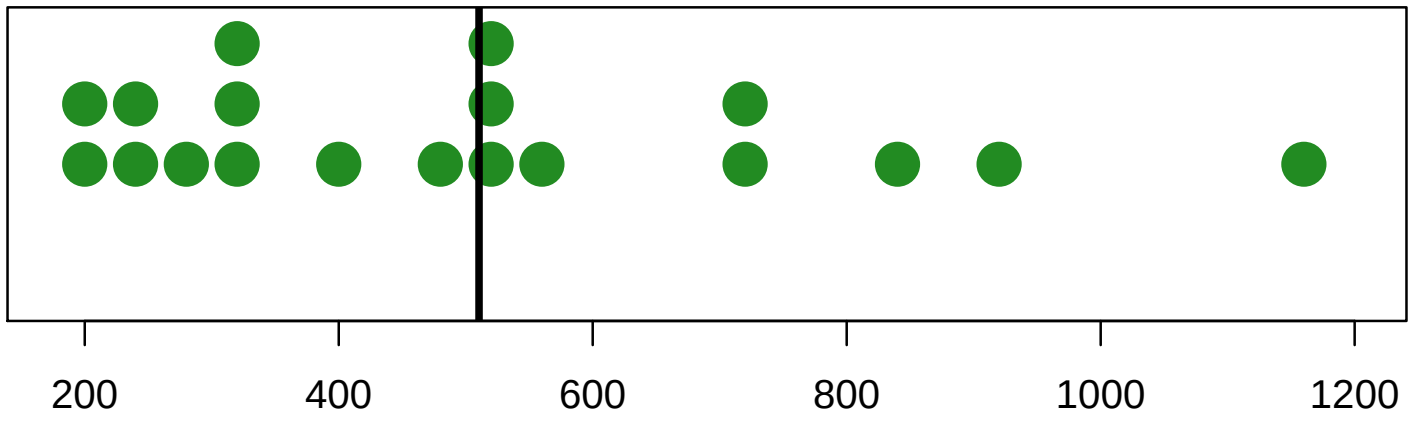


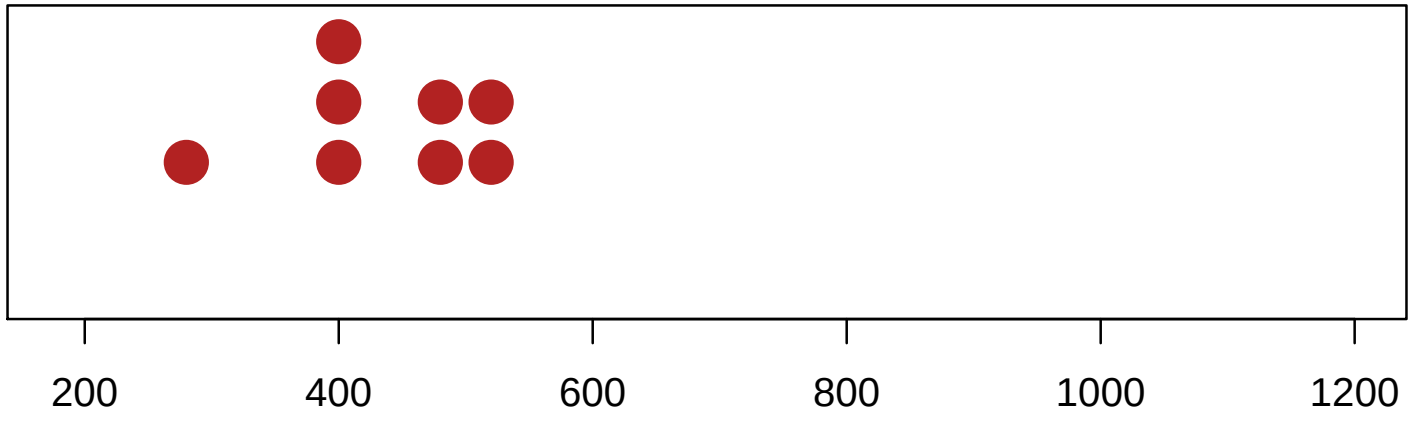
Bootstrap Resampling



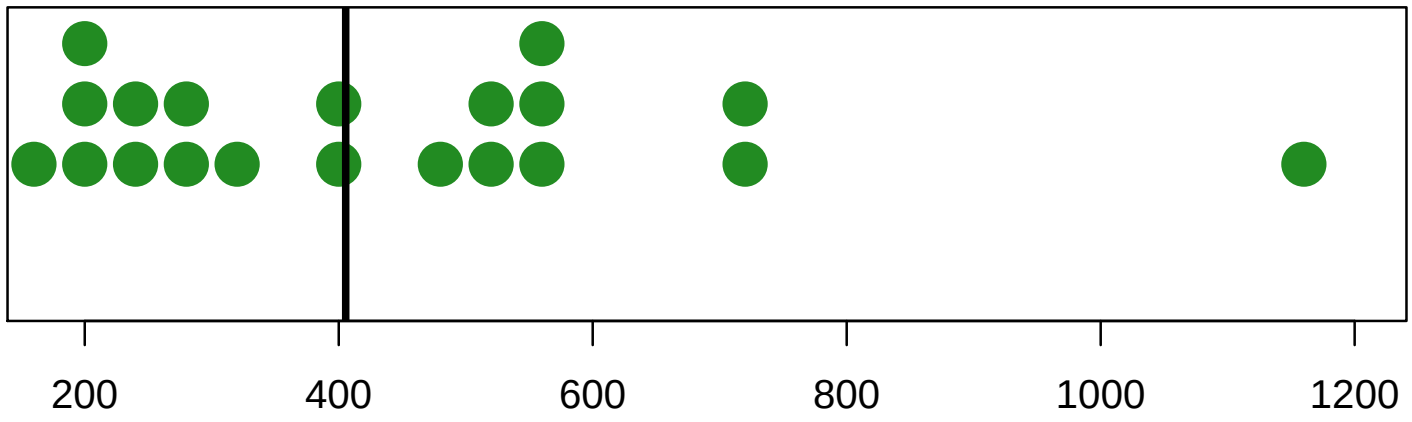


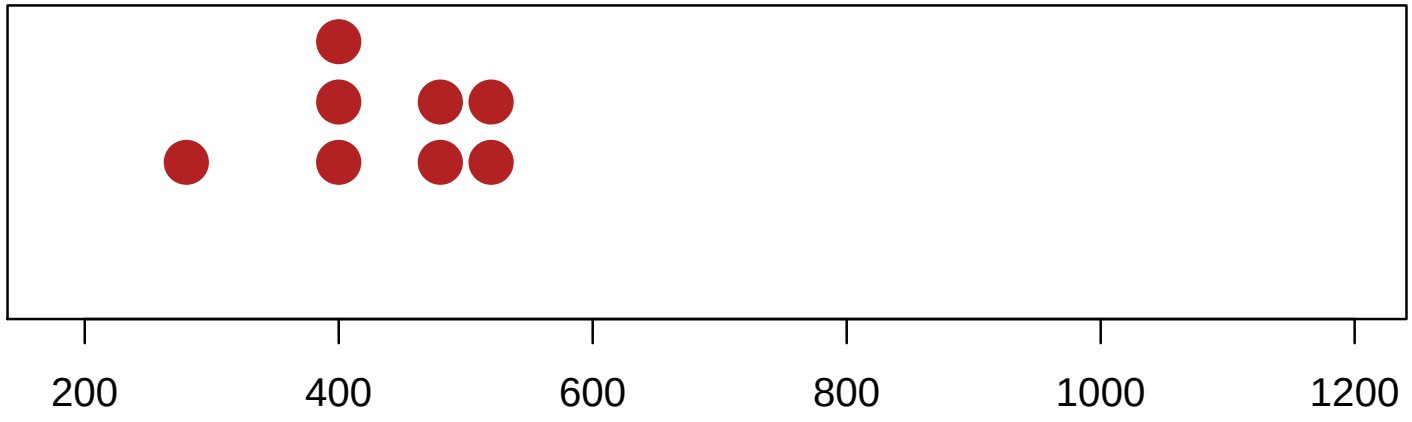
Bootstrap Resampling



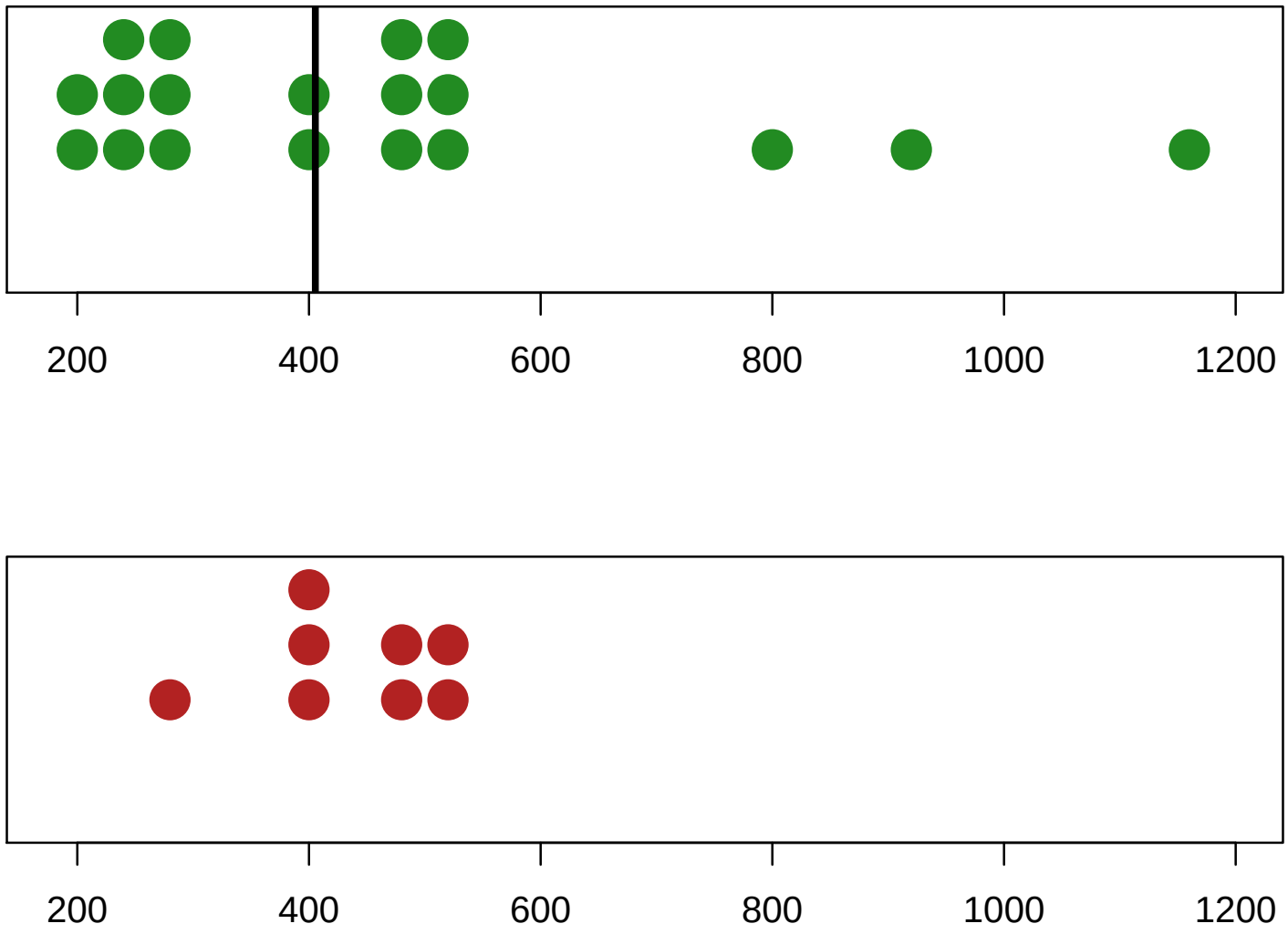


Bootstrap Resampling

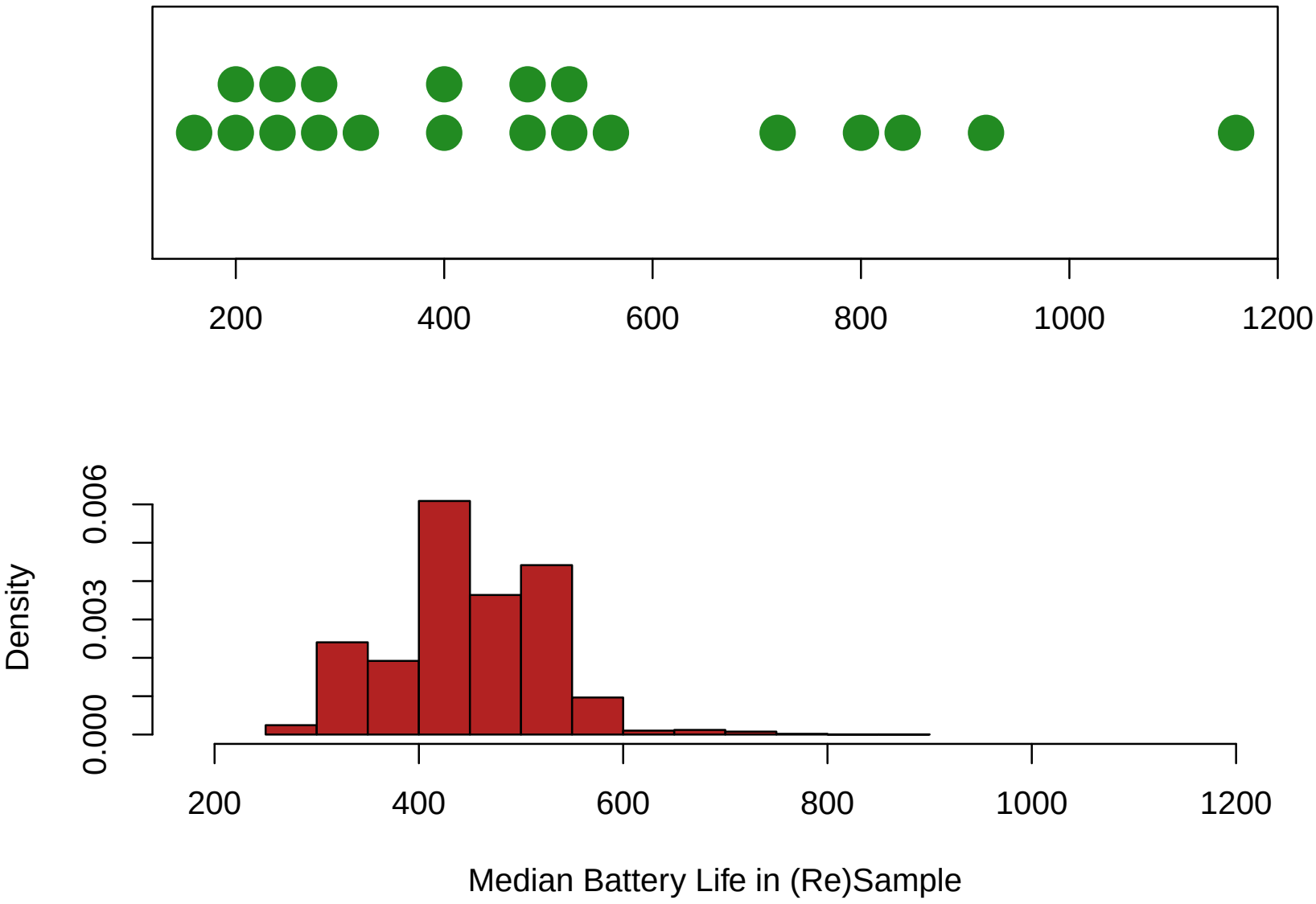




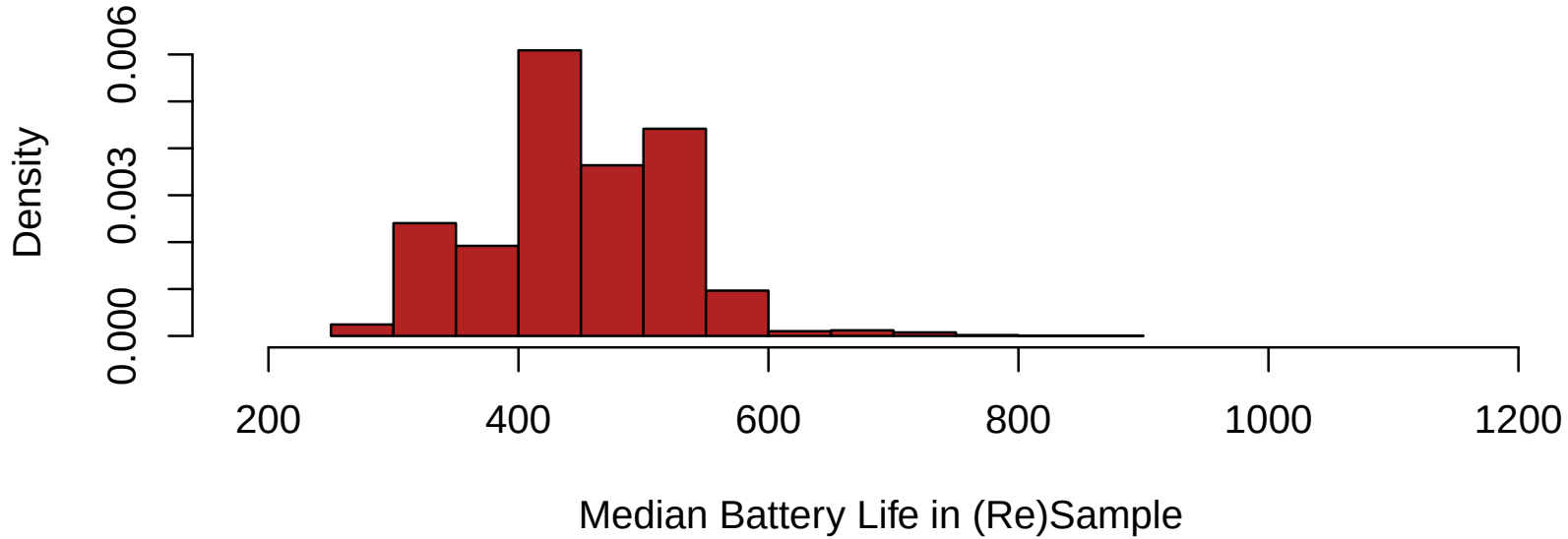
Bootstrap Resampling



Bootstrap Resampling

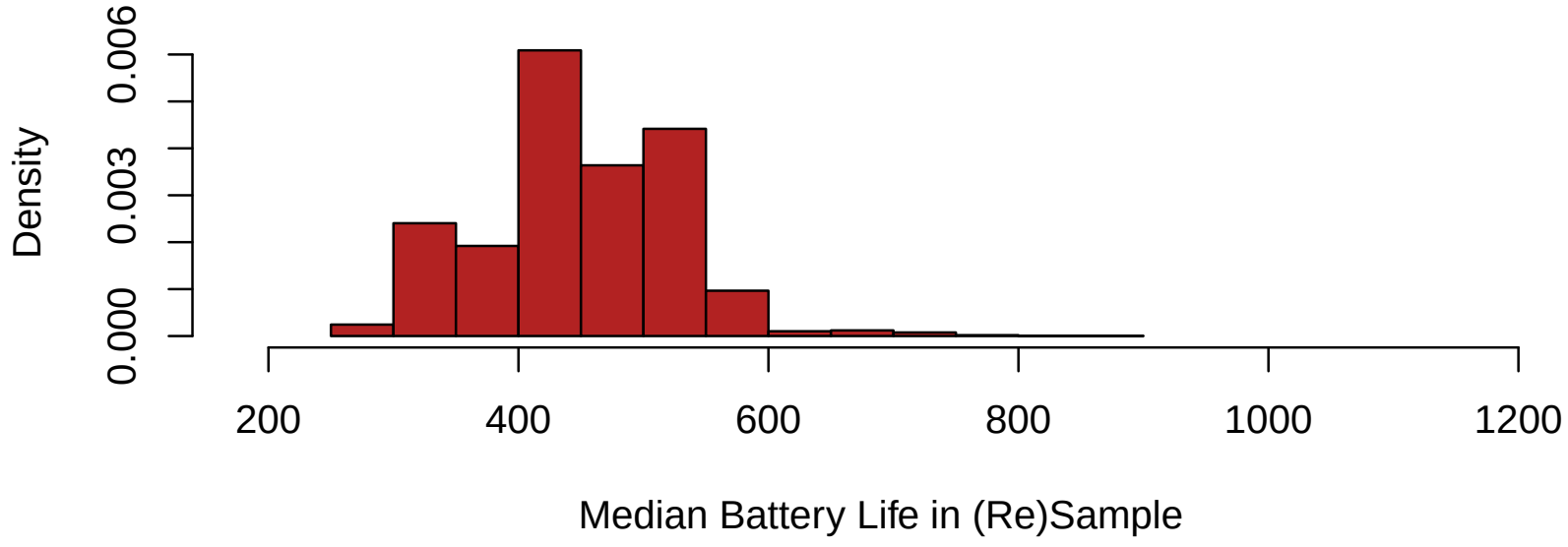


Bootstrap Resampling



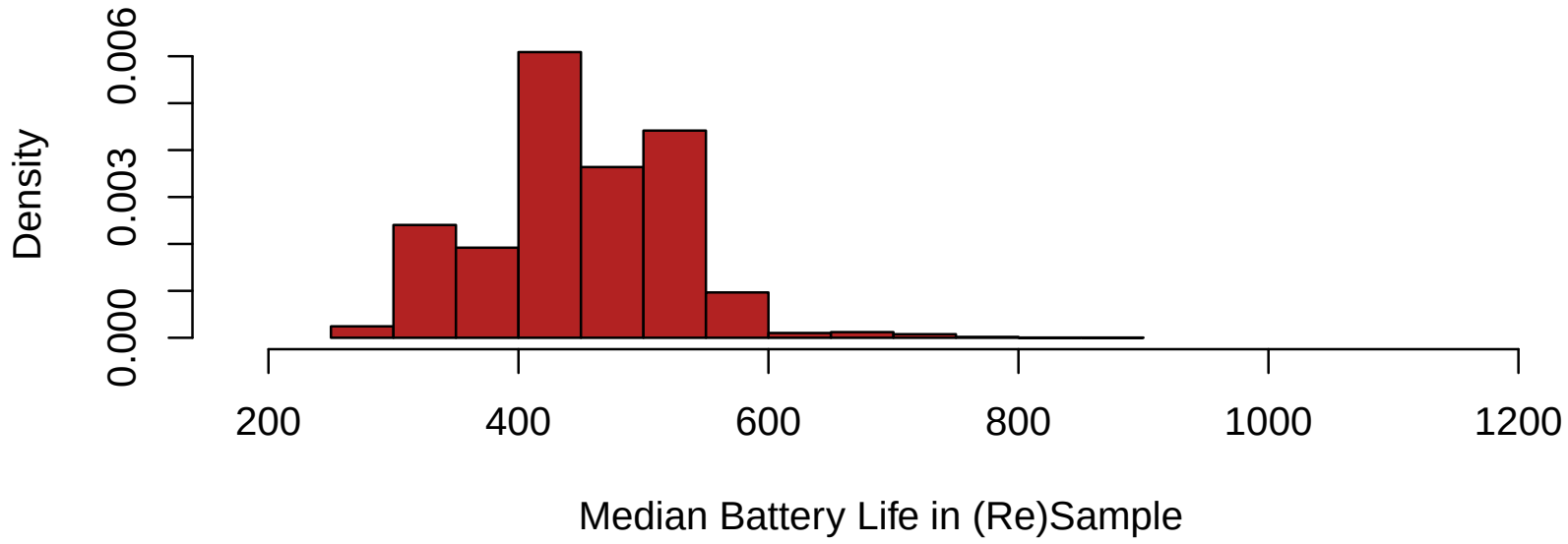
- What can we do with this distribution?
- Where did we make use of our hypotheses?

Bootstrap Resampling



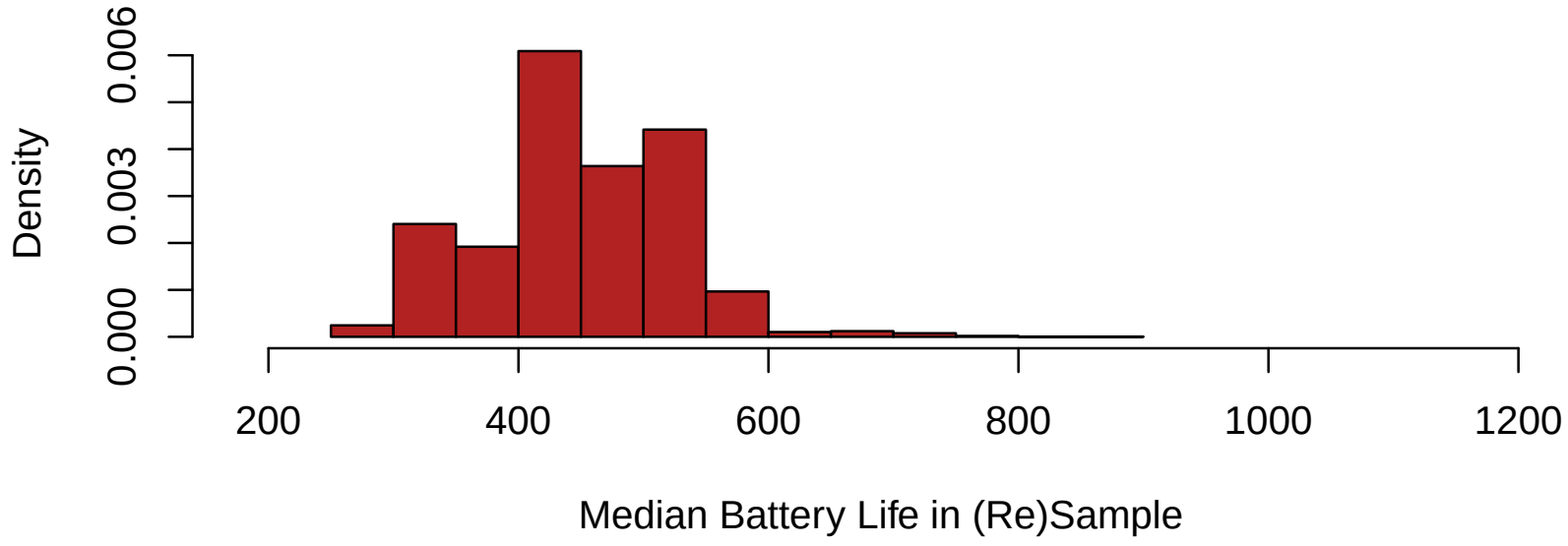
- What can we do with this distribution?
- Where did we make use of our hypotheses?
- We haven't yet! Our reference point is 300 miles. What should the (qualitative) relationship be in order to reject H_0?

Bootstrap Resampling



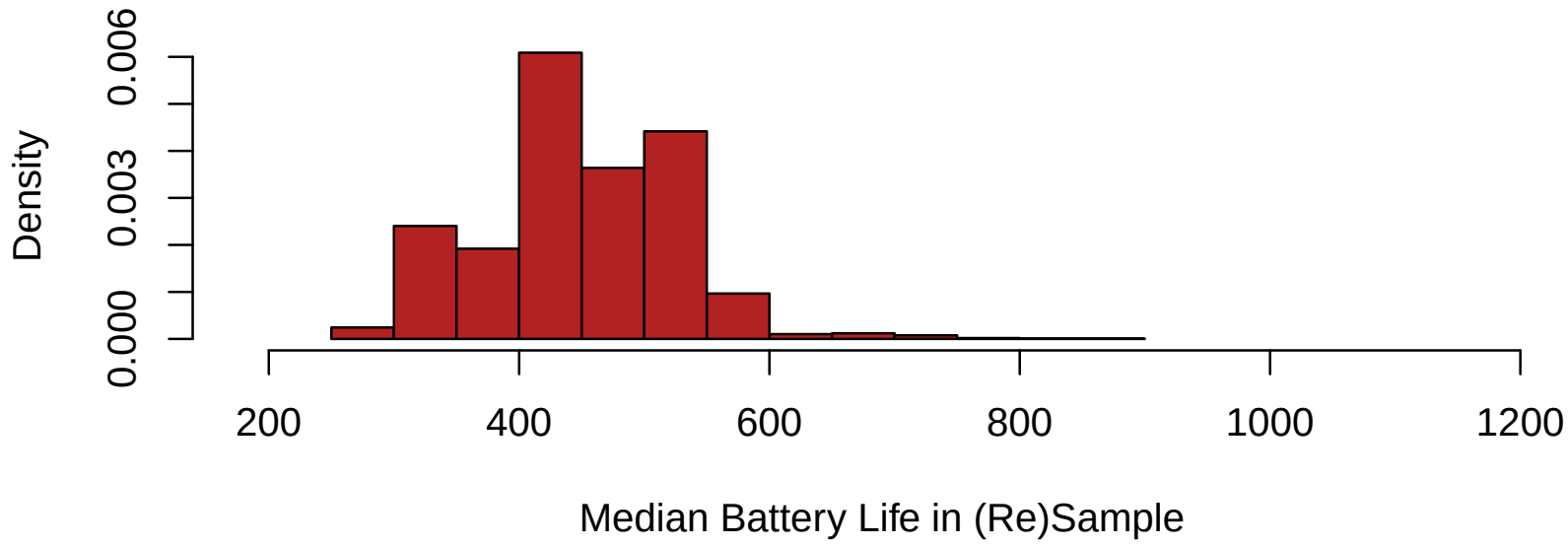
- Intuition: the farther away the H_0 value is from the mass of the bulk of the sampling distribution, the stronger the evidence that it's wrong.

Bootstrap Resampling



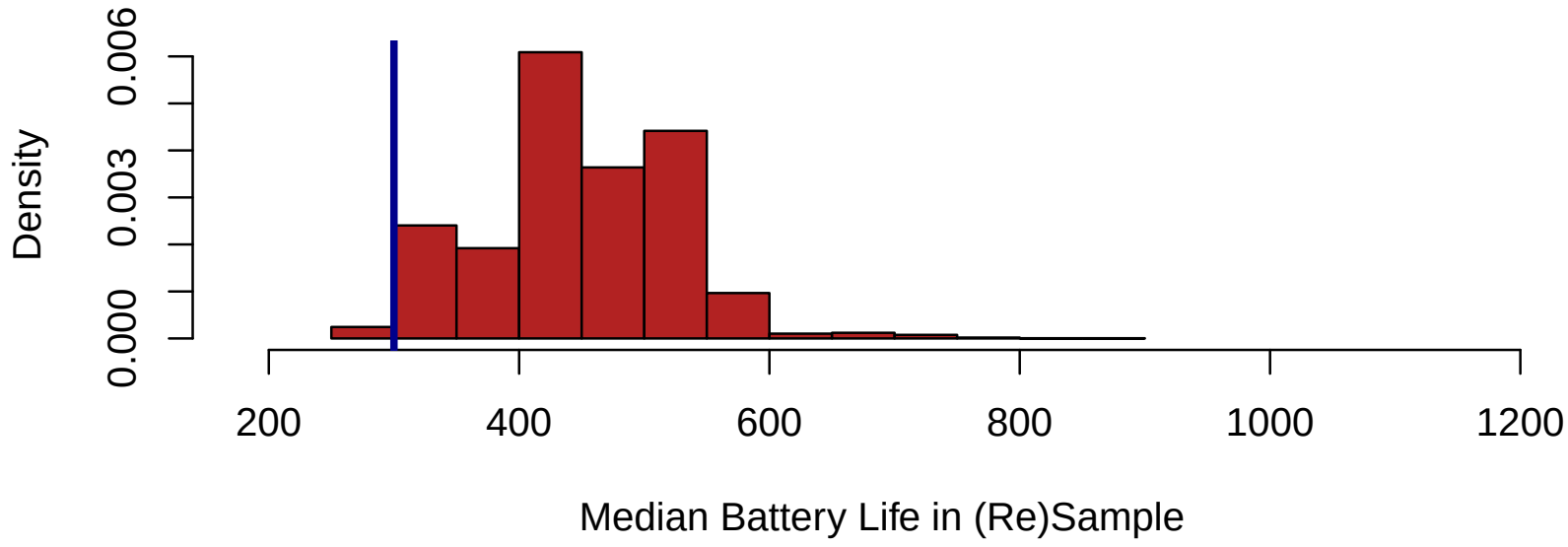
- Intuition: the farther away the H_0 value is from the mass of the bulk of the sampling distribution, the stronger the evidence that it's wrong.
- In this case, since H_1 says the median is *greater* than 300, we are looking for 300 to fall toward the *low* end of the sampling distribution.

Bootstrap Resampling



- If the H_0 value is in the extreme α (typically 0.05) of the distribution, reject H_0.

Bootstrap Resampling



- Here, 0.0122% of our simulated sample medians are below 300.
- Since this is less than α, we reject H_0.

Bootstrap Resampling

Bootstrap Resampling Procedure

1. Identify a **population value**, θ, of interest (e.g., median, 0.75 quantile, ratio of means, ...), and the boundary value, θ_0, between H_0 and H_1.
2. Identify a sample **statistic**, $\hat{\theta}$ that's a good estimator of θ (here it was the sample median, Q_2).
3. Collect a **simple random sample** of size n from the population.
4. Simulate repeated sampling from the population by repeatedly sampling n values **with replacement** from the sample. This gives an estimate of the sampling distribution of $\hat{\theta}$.
5. If θ_0 lies in the most **extreme** α of the simulated distribution (where this is defined based on directionality of H_1), reject H_0.

Bootstrap Resampling for Differences

- In NLP, we typically do bootstrap resampling to compare two groups: a baseline with my fancy new algorithm.
- In this case, the statistic of interest is the difference in means: mean(after) - mean(before).
- Assuming that your H1 says that "after" is better, then H1 is: difference > 0 and H0 is: difference ≤ 0.
- Resample using the standard algorithm and see what percentage of values ≤ 0 exist in the constructed distribution.
- If this percentage is lower than $\alpha = 0.05$ you can claim your algorithm yields statistically significant improvements.

Exercise

- You implemented a "fancy new system"^{TM} for binary classification, and want to compare it against a baseline system on a dataset of 10 points. The labels produced on this dataset are:

	1	2	3	4	5	6	7	8	9	10
Gold	A	B	A	B	A	B	A	B	A	B
Fancy new system	A	B	A	B	A	B	A	A	A	B
Baseline	A	B	A	B	B	B	A	A	A	B

- What are your H0 and H1 hypotheses?
- Can you reject H0 with $\alpha = 0.20$ using bootstrap resampling using 10 samples?