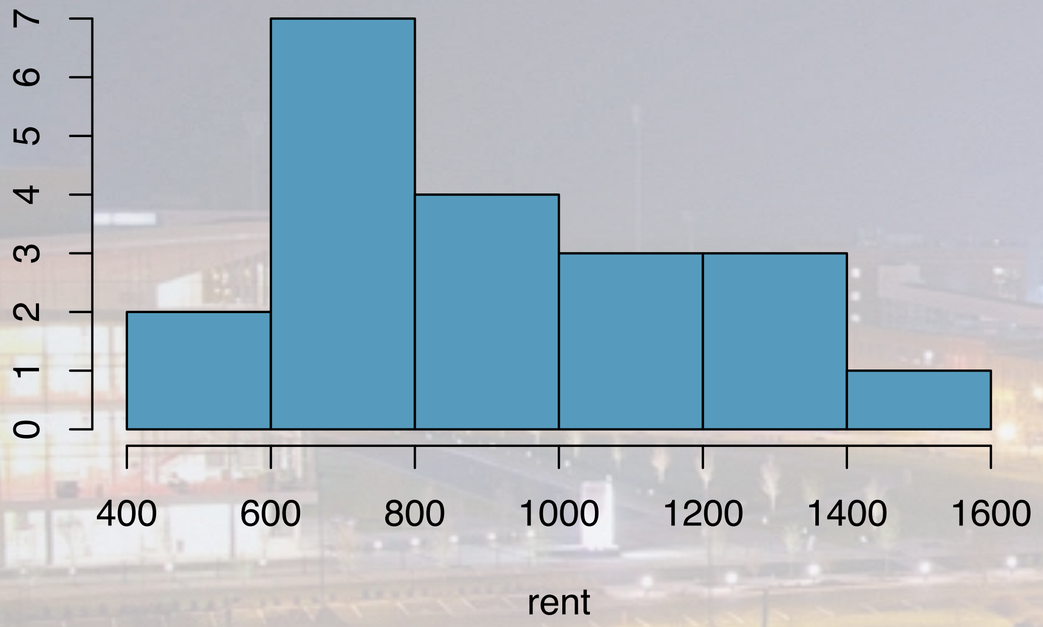
bootstrapping

- what is bootstrapping?
- estimation beyond means
- limitations

Duke
UNIVERSITY

Dr. Mine Çetinkaya-Rundel
Duke University

rent in durham, nc



Twenty 1+ bedroom apartments were randomly selected on raleigh.craigslist.org. (keyword: **`Durham`**). Is the mean or the median a better measure of typical rent in Durham?

Can we apply CLT based methods we have learned so far to construct confidence intervals for both?

Photo by Kiril Kolev http://www.flickr.com/photos/kiril106/3110838732 (CC BY 2.0)

bootstrapping

- *"Pulling oneself up by one's bootstraps":* accomplishing an impossible task without any outside help
- In this case, the ~~im~~possible task is estimating a population parameter, using data from only the given sample.

Boots: http://openclipart.org/detail/26401/-by--26401

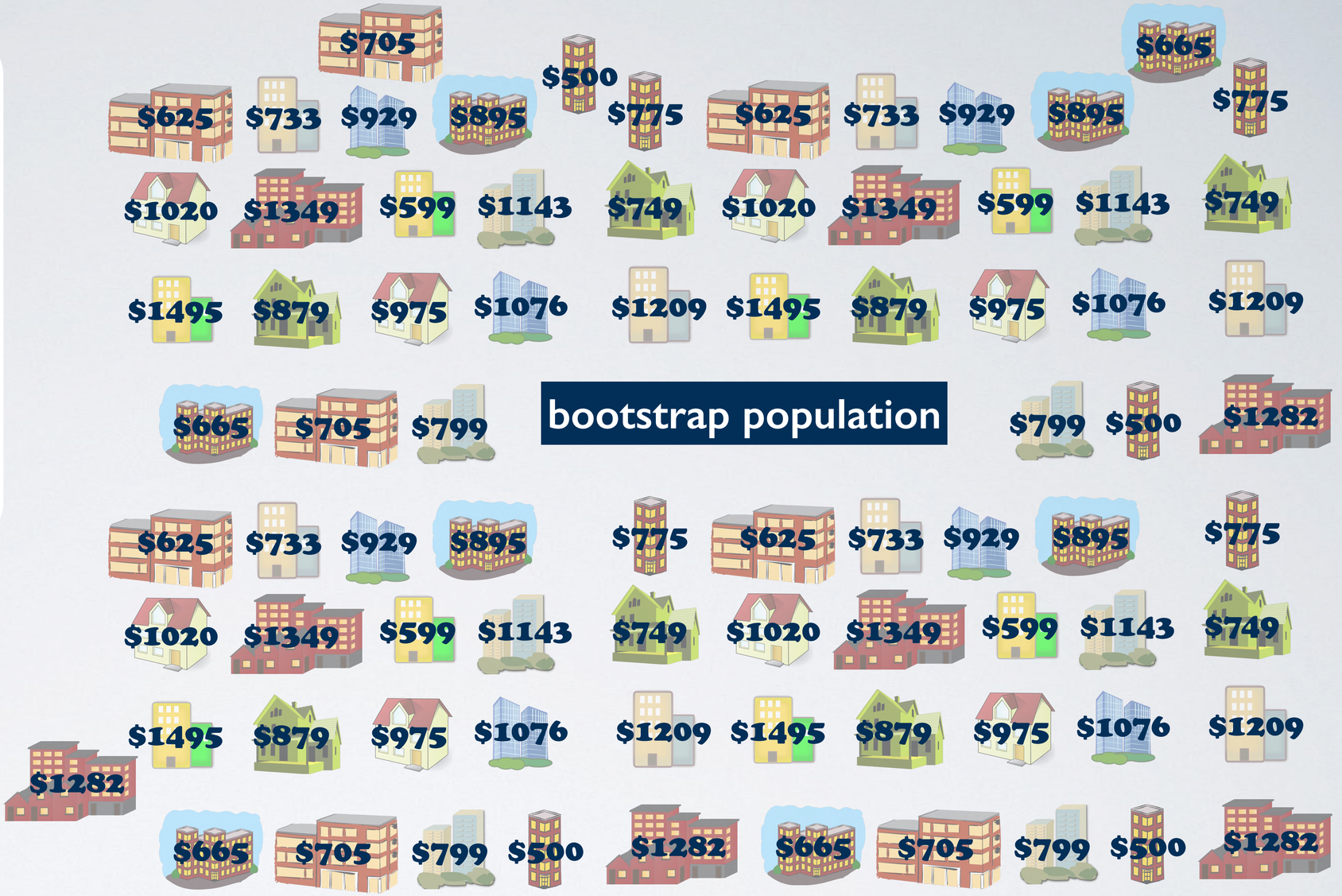
original sample

median = $887

All images from OpenClipArt.org

original sample

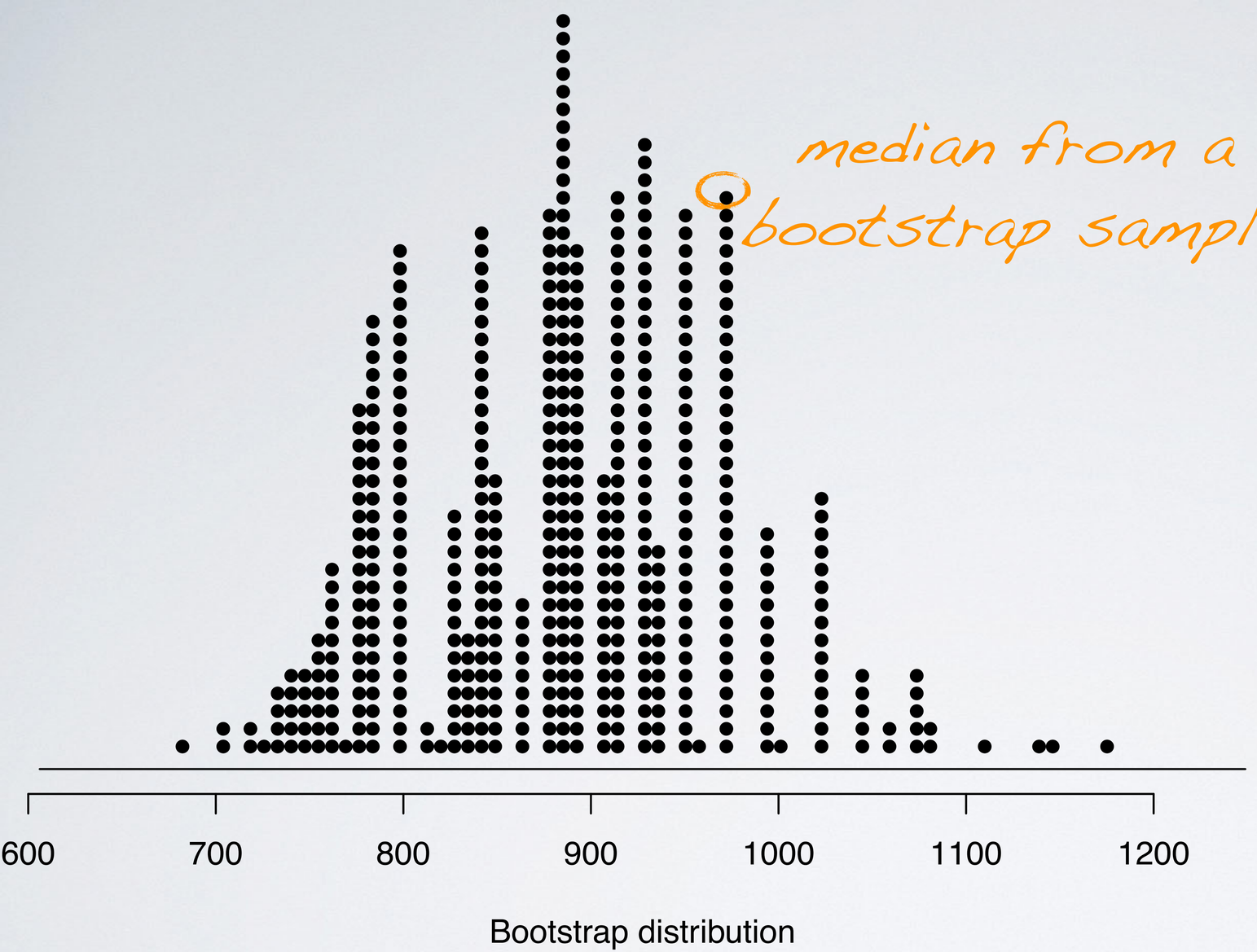
bootstrap population



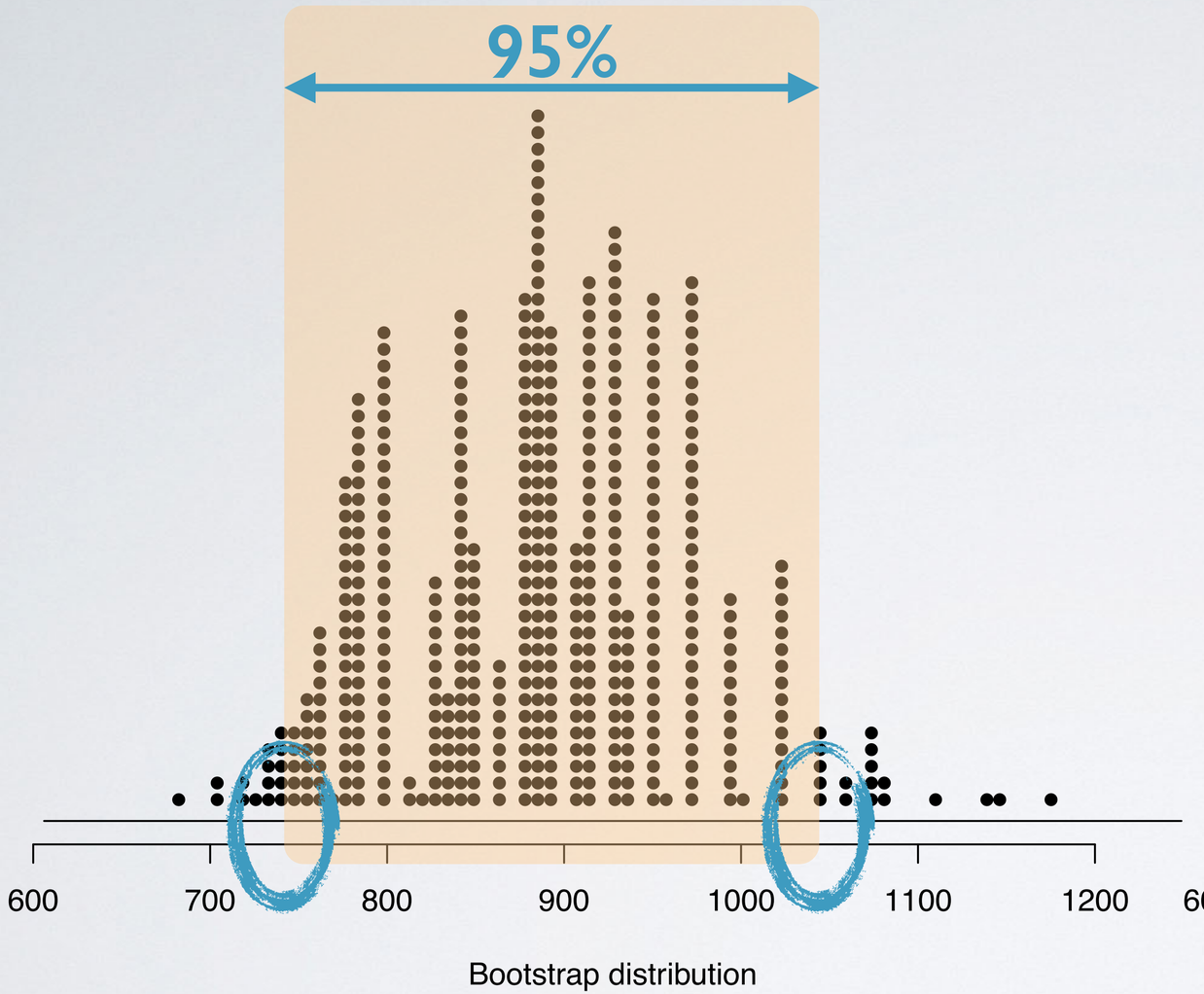
All images from OpenClipArt.org

bootstrapping scheme

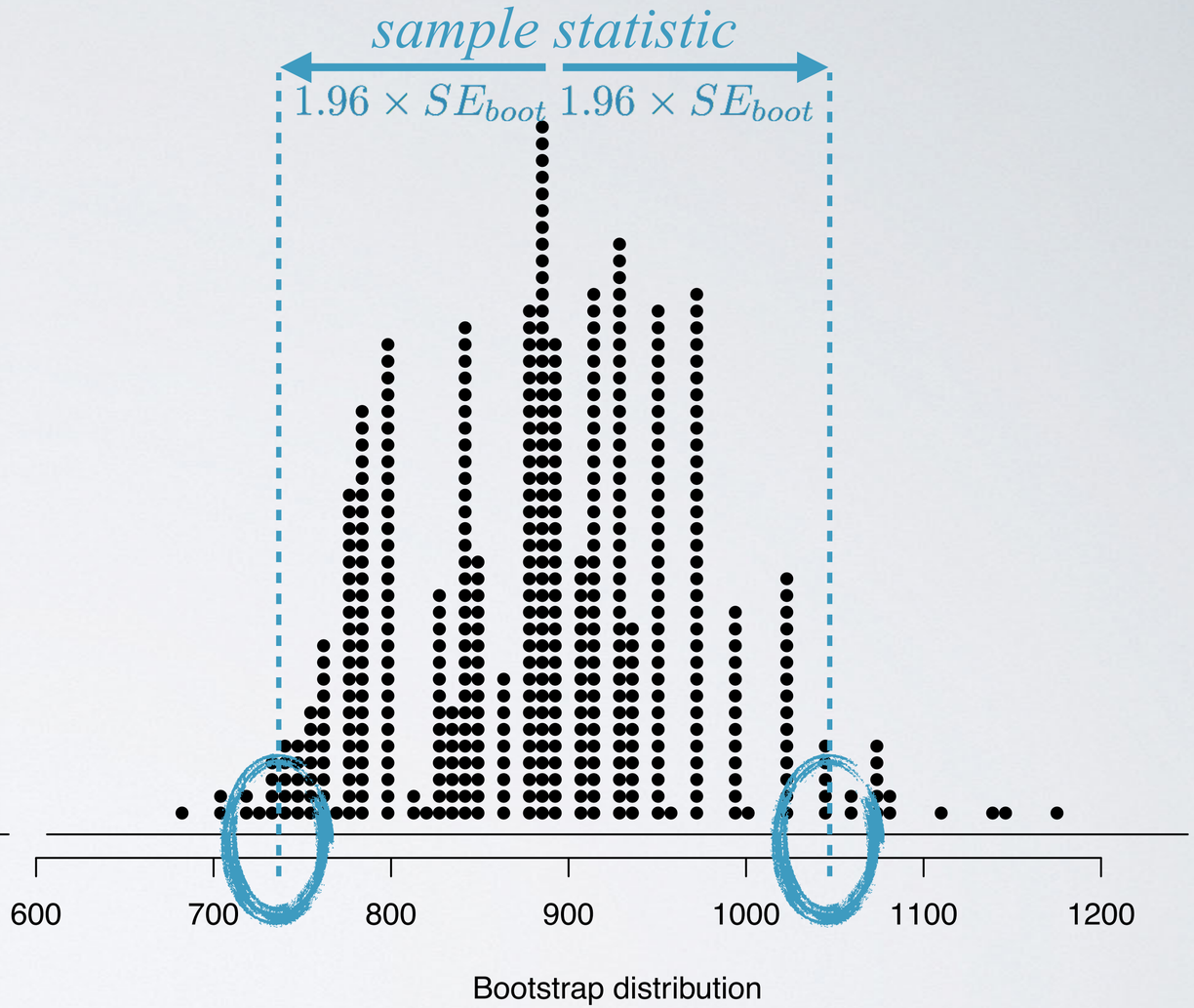
(1) take a bootstrap sample - a random sample taken **with replacement** from the original sample, of the same size as the original sample
(2) calculate the bootstrap statistic - a statistic such as mean, median, proportion, etc. computed on the bootstrap samples
(3) repeat steps (1) and (2) many times to create a bootstrap distribution - a distribution of bootstrap statistics


median from a
bootstrap sample
600
700
800
900
1000
1100
1200
Bootstrap distribution

(1) percentile method



(2) standard error method



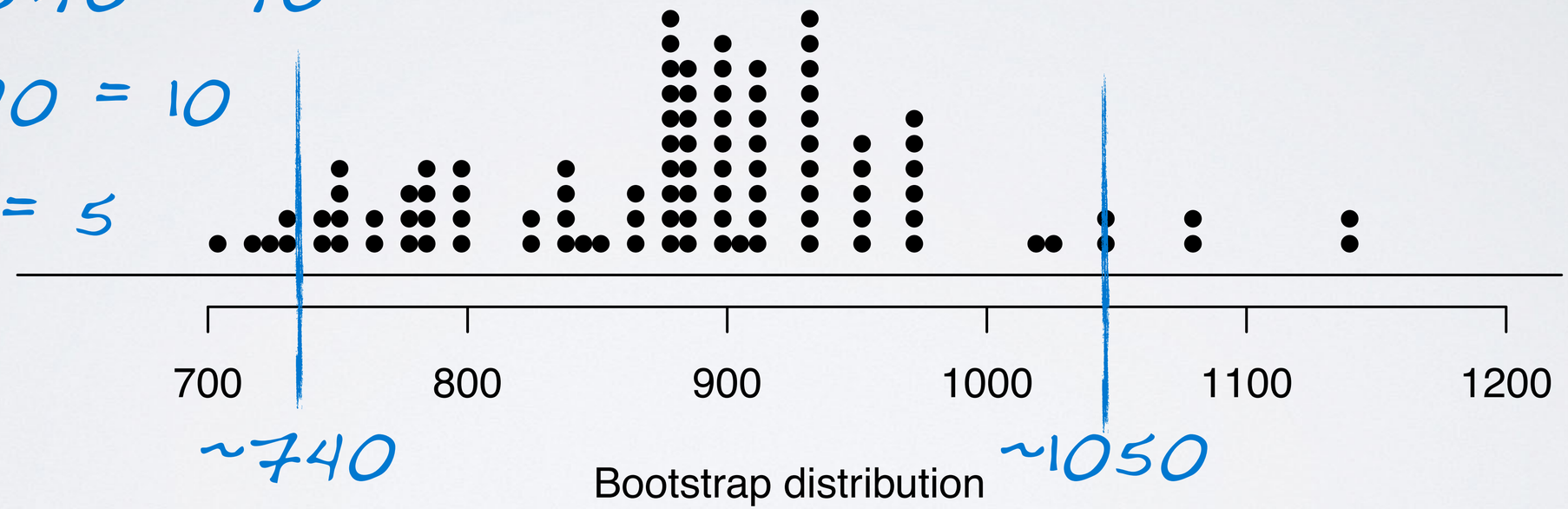
practice

The dot plot below shows the distribution of medians of 100 bootstrap samples from the original sample. Estimate the 90% bootstrap confidence interval for the median rent based on this bootstrap distribution using the percentile method.

$100 \times 0.90 = 90$

$100 - 90 = 10$

$10 / 2 = 5$



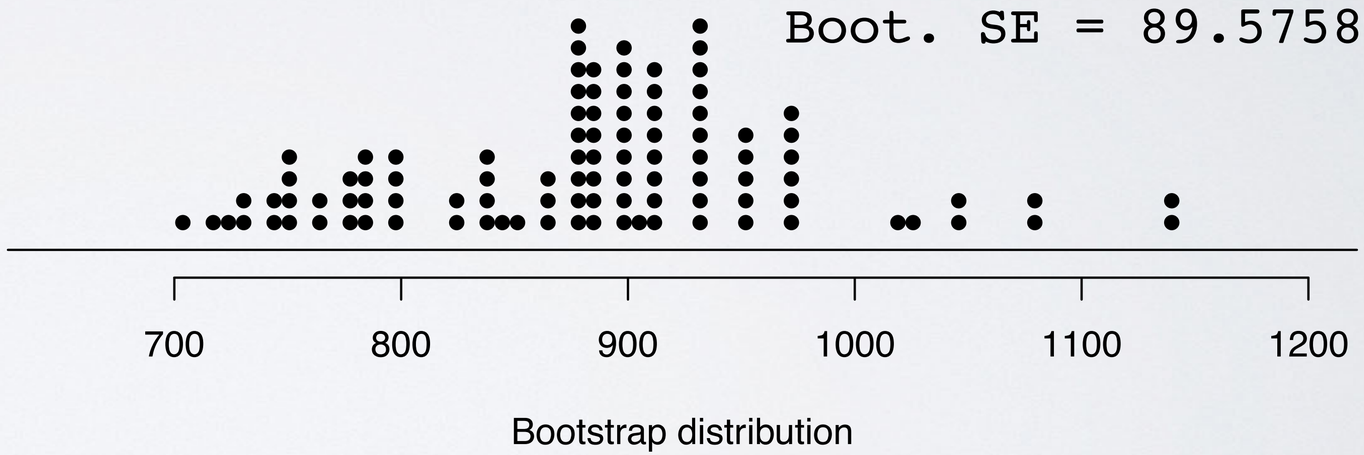
practice

The dot plot below shows the distribution of medians of 100 bootstrap samples from the original sample. Estimate the 90% bootstrap confidence interval for the median rent based on this bootstrap distribution using the standard error method.

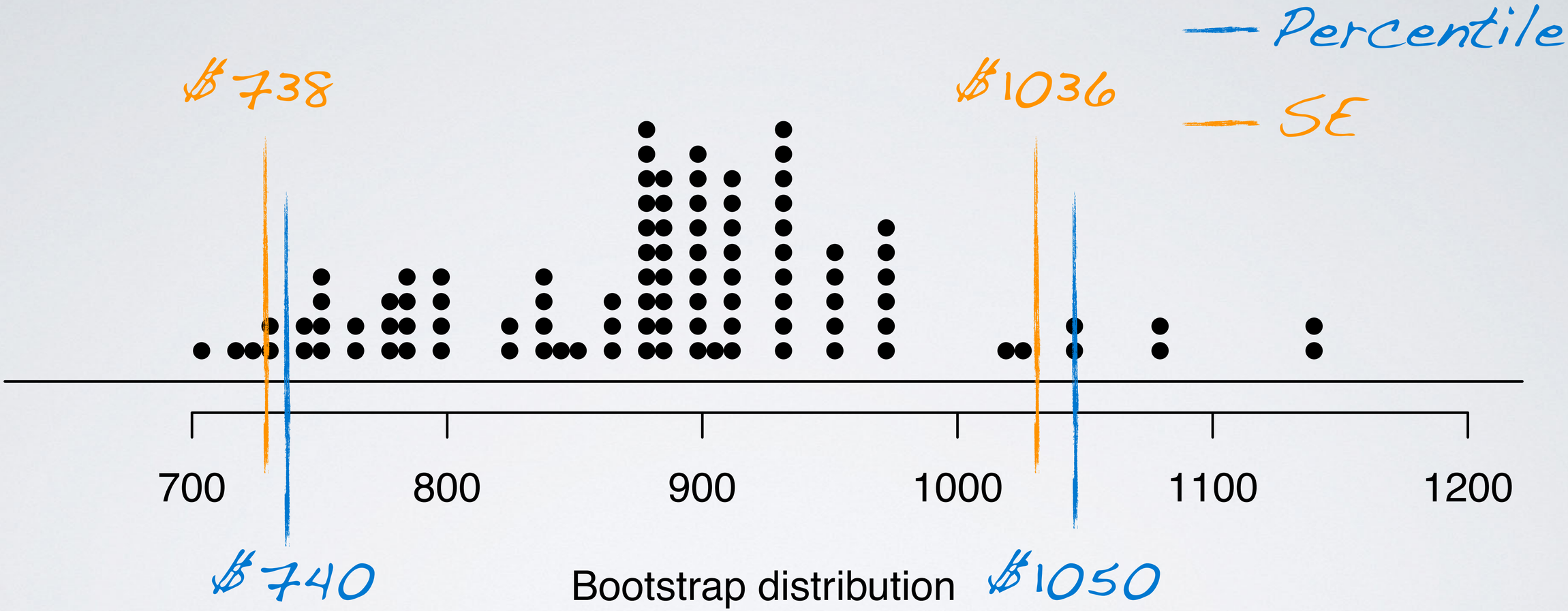
sample median $\pm t^* SE_{boot}$ =

$= 887 \pm 1.66 \times 89.5758$

$\approx (738, 1036)$



comparison: percentile vs. SE methods



bootstrapping limitations

- Not as rigid conditions as CLT based methods
- If the bootstrap distribution is extremely skewed or sparse, the bootstrap interval might be unreliable
- A representative sample is still required — if the sample is biased, the estimates resulting from this sample will also be biased.

bootstrap vs. sampling distribution

- Sampling distribution created using sampling (with replacement) from the population
- Bootstrap distribution created using sampling (with replacement) from the sample
- Both are distributions of sample statistics