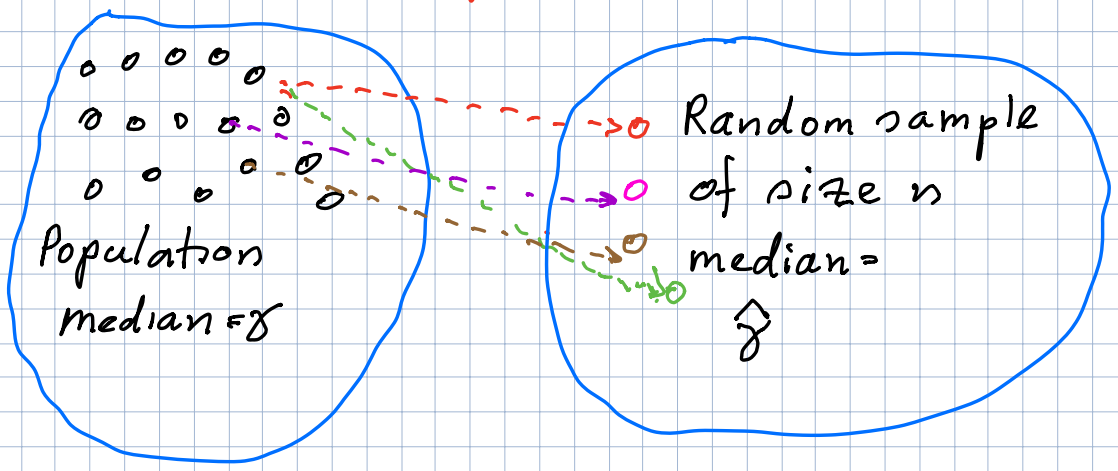
The Bootstrap



Estimate population median γ by sample median $\hat{\gamma}$

If we had picked a different sample we would have obtained a different estimate $\hat{\gamma}$

Want to know the sampling variability of $\hat{\gamma}$

If we could obtain additional samples from the population, we might

- draw a large number m of samples
- compute their medians $\hat{\gamma}_1 \ldots \hat{\gamma}_m$

We could then

- draw a histogram of $\hat{\gamma}_1 \ldots \hat{\gamma}_m$
- measure their variability by their standard deviation $\text{std}(\hat{\gamma}_1 \ldots \hat{\gamma}_m)$

Two methods for getting a 95% confidence interval for the population median γ:

1. "Normal theory" intervals

$$\hat{\gamma} \pm 2 * \text{std}(\hat{\gamma}_1 \ldots \hat{\gamma}_m)$$

2. Percentile intervals

Find interval $[u, v]$ such that 2.5% of sample medians $\hat{\gamma}_1 \ldots \hat{\gamma}_m$ are $\leq u$, and 2.5% are $> v$.

"Normal theory" intervals are based on the assumption that the distribution (histogram) of $\hat{\gamma}_1 \ldots \hat{\gamma}_m$ looks like a Gaussian distribution (bell shaped curve). Will usually be true if the sample size n is large enough.

Percentile intervals seem reasonable but are not so easy to justify.

Note: If we were estimating the population mean μ by by the sample mean $\hat{\mu}$ the situation would be simpler because we know that the std of sample means (the standard error) is $\sigma/\sqrt{n}$, where σ is the population standard deviation.

There is no such simple rule for the median, or the standard deviation, or most other statistics.

The catch!

The idea of using additional samples to assess the variability of $\hat{\gamma}$ is silly.

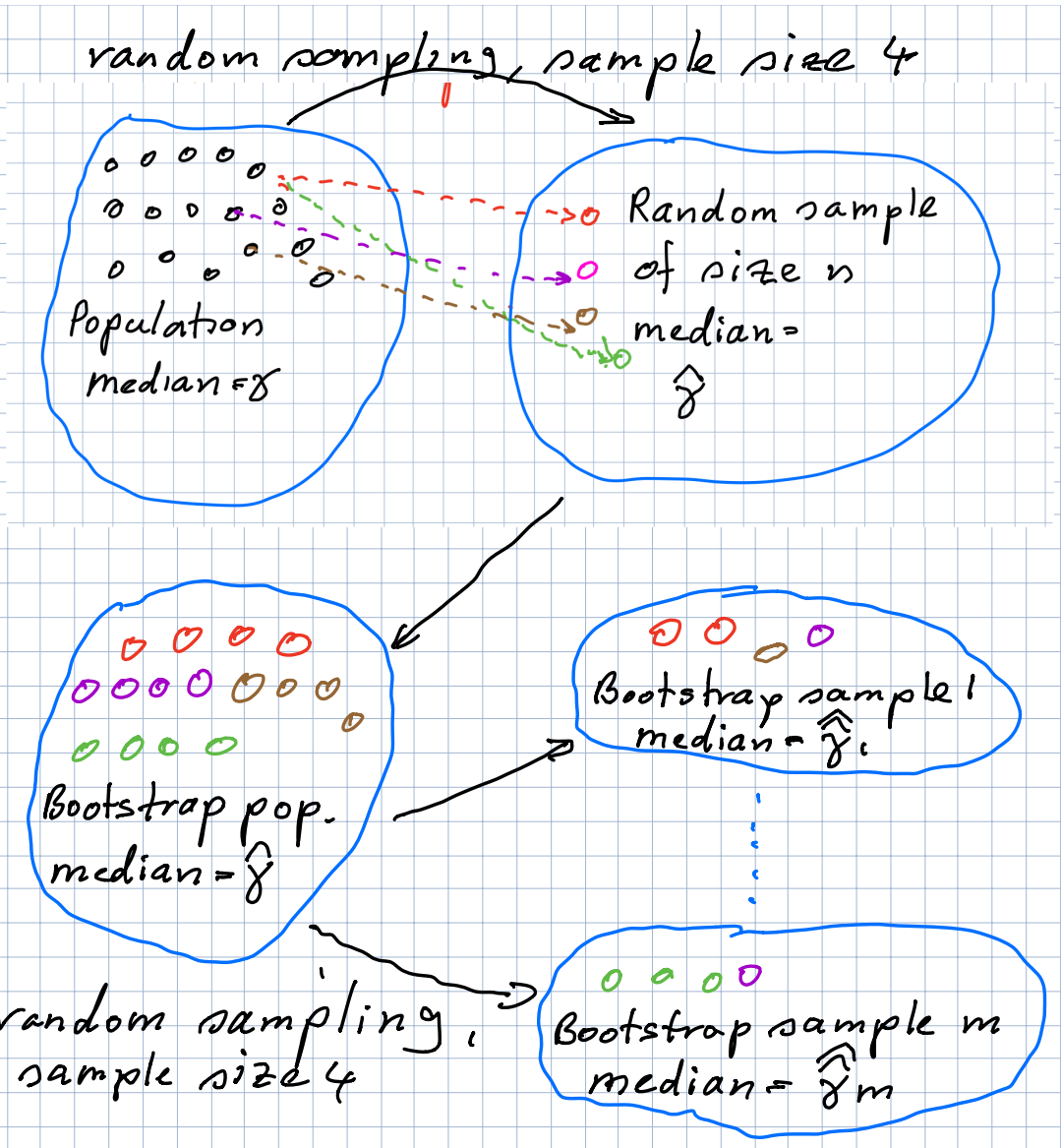
In practice, if we really could get 100 additional samples, we would pool all our data into a single sample of size 100 n and estimate the population median by the median of all 100 n observations.

What to do? Bootstrap!

Instead of drawing more samples from the population, draw "Bootstrap samples" from the original sample

In other words

For the purpose of assessing variabity, substitute the sample for the population



- Draw histogram of $\hat{\hat{\gamma}}_1 \ldots \hat{\hat{\gamma}}_m$
- Estimate the sampling variabity of $\hat{\gamma}$ by std ($\hat{\hat{\gamma}}_1 \ldots \hat{\hat{\gamma}}_m$)

Note: In the figure we replicated the observations in the original sample so that the Bootstrap population had the same size as the original population.

This is not necessary because we are sampling with replacement.

Two methods for getting a 95% confidence interval for the population median γ

1. "Normal theory" intervals

$$\hat{\gamma} \pm 2 \text{ std}(\hat{\gamma}_1^* \ldots \hat{\gamma}_m^*)$$

2. Percentile intervals

Find $[u, v]$ such that 2.5% of the Bootstrap sample medians are $< u$, and 2.5% are $> v$