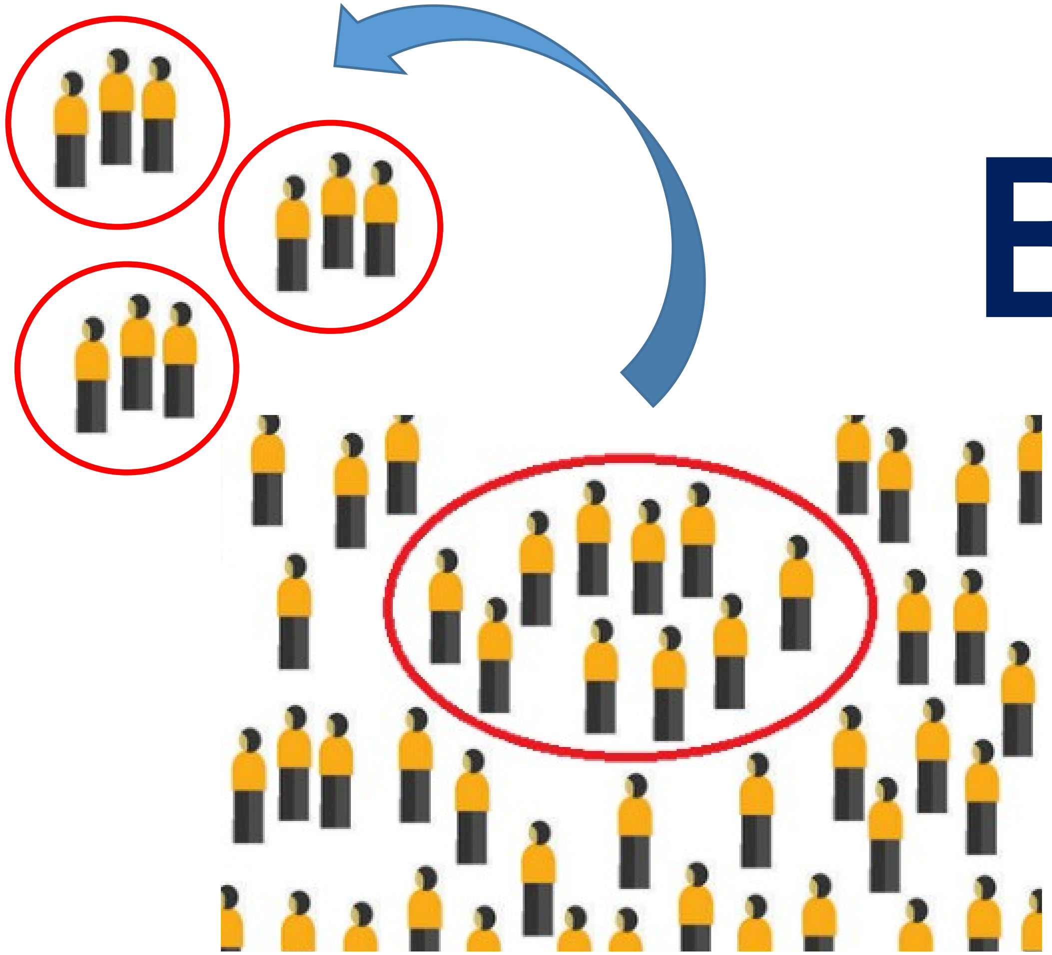


Bootstrapping
Method

Elaborado por Guilherme Souza Rodrigues
Docente Donald Matthew Pianto

Why use the bootstrap?

- ✓ Good question.

- ✓ Small sample size.
- ✓ Non-normal distribution of the sample.
- ✓ A test of means for two samples.
- ✓ Not as sensitive to N.

Why use the bootstrap?

✓Small sample size.

✓Non-normal distribution of the sample.

✓A test of means for two samples.

✓Not as sensitive to N.

What’s it all about?

- ✓Actuaries compute points estimates of statistics all the time.
  - ✓Loss ratio/claim frequency for a population
  - ✓Outstanding Losses
  - ✓Correlation between variables
  - ✓GLM parameter estimates …

- ✓A point estimate tells us what the data indicates.
- ✓But how can we measure our *confidence* in this indication?

What's it all about?

- ✓Actuaries compute points estimates of statistics all the time.
  - ✓Loss ratio/claim frequency for a population
  - ✓Outstanding Losses
  - ✓Correlation between variables
  - ✓GLM parameter estimates …

- ✓A point estimate tells us what the data indicates.
- ✓But how can we measure our *confidence* in this indication?

More Concisely…

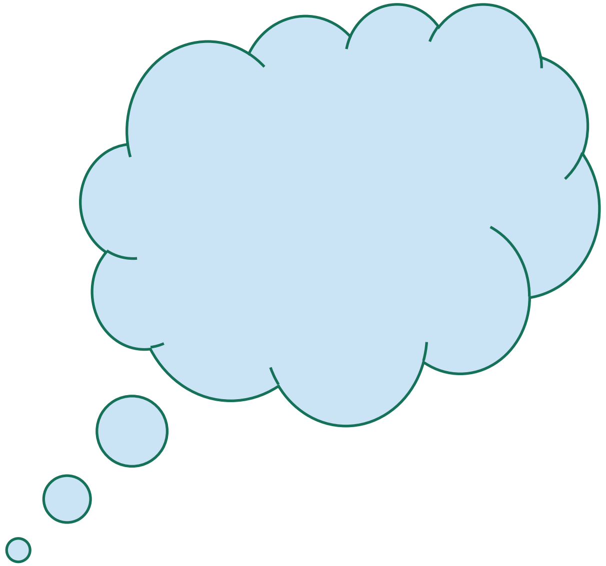
- ✓Point estimate says: “what do you think?”
- ✓Variability of the point estimate says: “how sure are you?”
- ✓Traditional approaches
  - ✓Credibility theory
  - ✓Use distributional assumptions to construct confidence intervals
- ✓Is there an easier – and more flexible – way?

Enter the Bootstrap

- ✓In the late 70's the statistician Brad Efron made an ingenious suggestion.
- ✓Most (sometimes all) of what we know about the "true" probability distribution comes from the data.
- ✓So let's treat the data as a *proxy* for the true distribution.
- ✓We draw multiple samples from this proxy…
  - ✓This is called "resampling".
- ✓And compute the statistic of interest on each of the resulting pseudo-datasets.

Philosophy

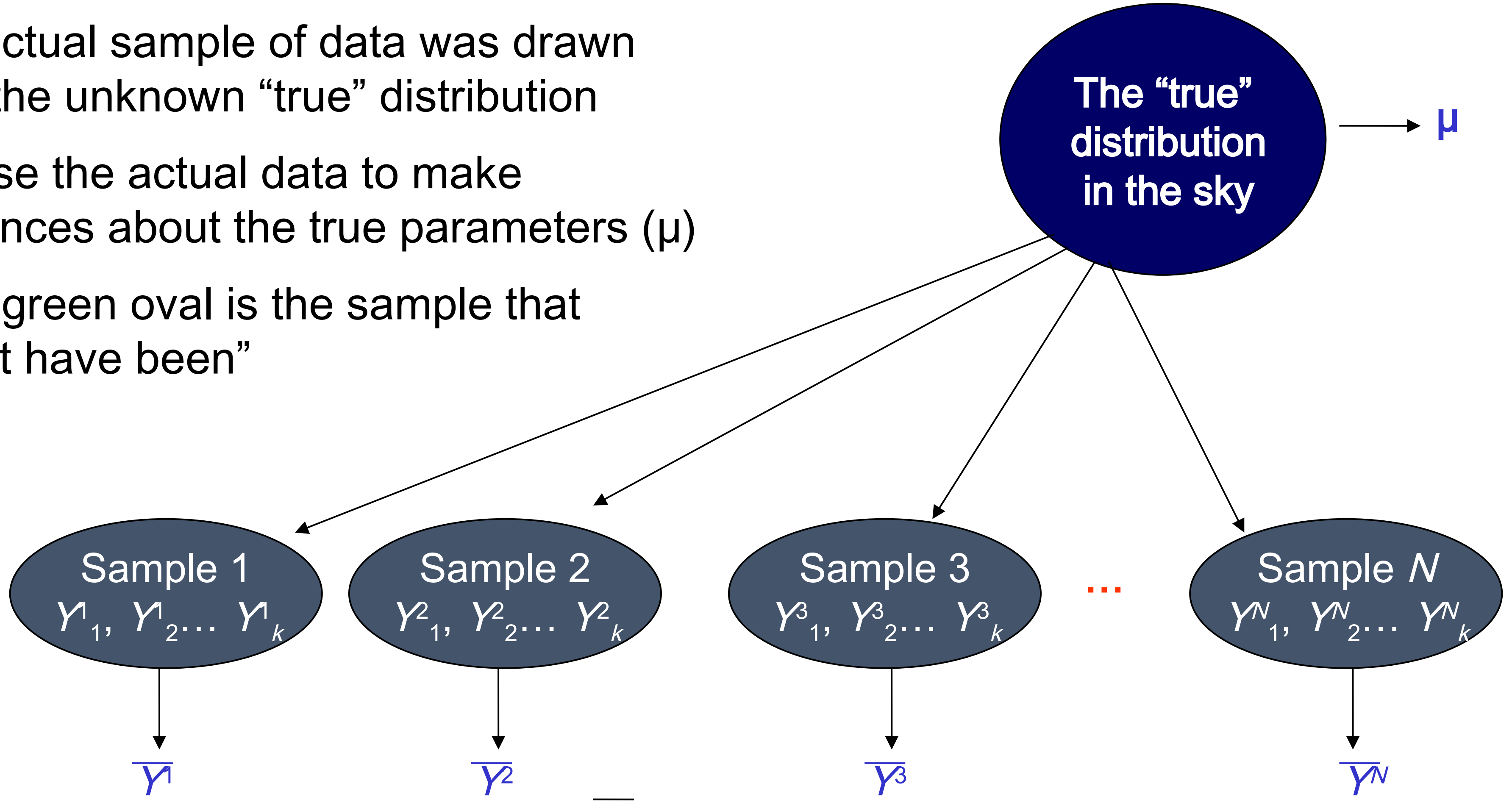
- ✓"[Bootstrapping has] requires very little in the way of modeling, assumptions, or analysis, and can be applied in an automatic way to any situation, no matter how complicated".
- ✓"An important theme is the substitution of raw computing power for theoretical analysis"
  - ✓--Efron and Gong 1983
- ✓Bootstrapping fits very nicely into the "data mining" paradigm.



The Basic Idea

- ✓ Any actual sample of data was drawn from the unknown “true” distribution
- ✓ We use the actual data to make inferences about the true parameters (μ)
- ✓ Each green oval is the sample that “might have been”

Theoretical Picture

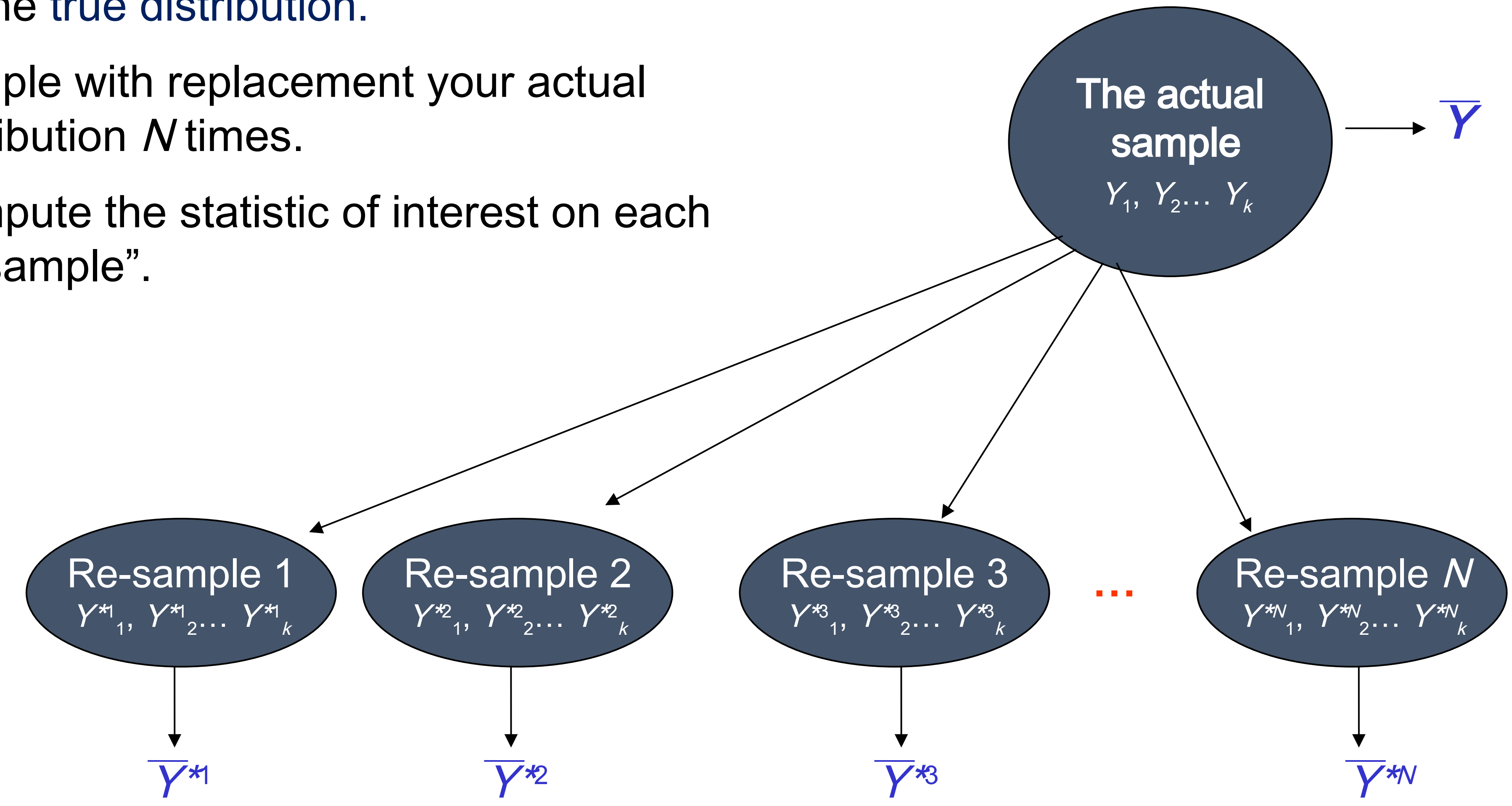


- The distribution of our estimator ($\overline{Y}$) depends on both the true distribution *and* the size (k) of our sample

The Basic Idea

- ✓ Treat the actual distribution as a proxy for the true distribution.
- ✓ Sample with replacement your actual distribution *N* times.
- ✓ Compute the statistic of interest on each "re-sample".

The Bootstrapping Process



$\{\overline{Y}^*\}$ constitutes an estimate of the *distribution* of $\overline{Y}$.

Sampling With Replacement

✓In fact, there is a chance of

$$(1-1/500)^{500} \approx {}^{1}/_{e} \approx .368$$

that any one of the original data points won't appear at all if we sample with replacement 500 times.

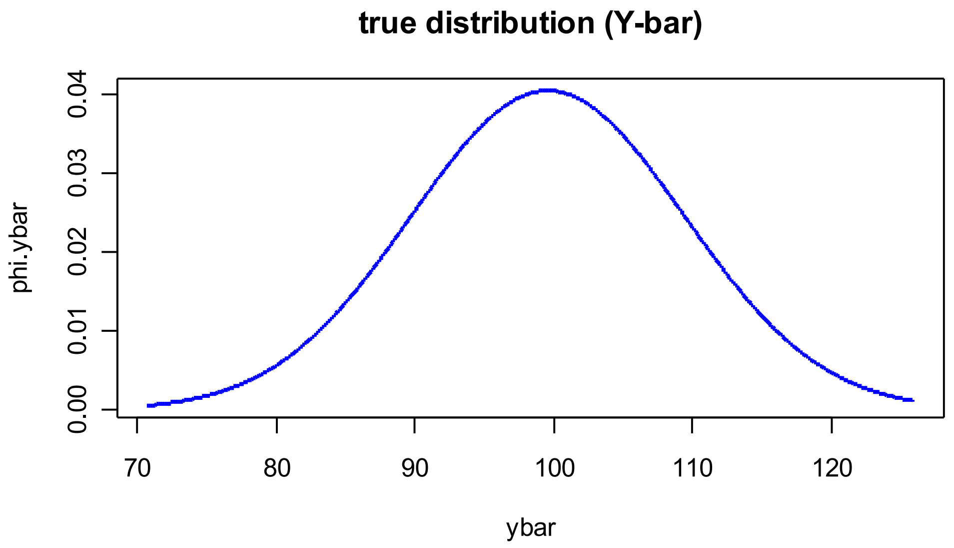
➔any data point is included with Prob ≈ .632

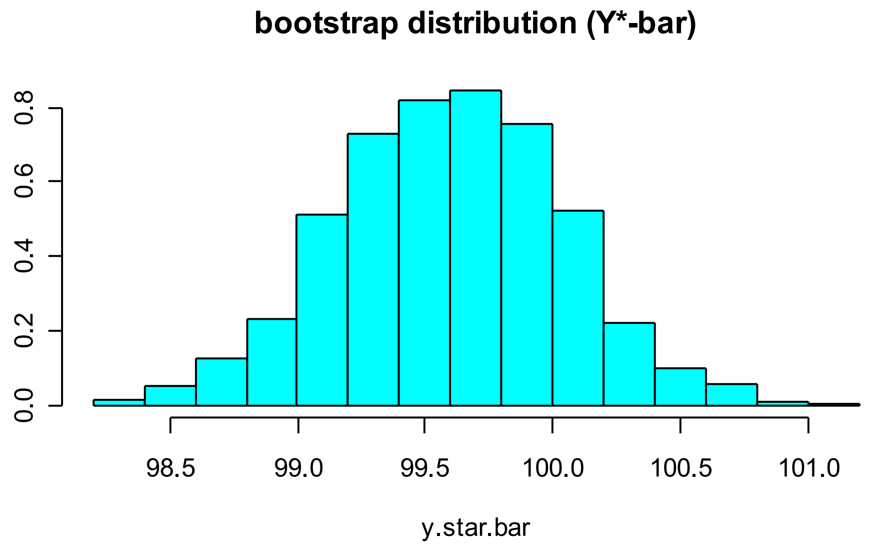
✓Intuitively, we treat the original sample as the "true population in the sky".

✓Each *resample* simulates the process of taking a sample from the "true" distribution.

Theoretical vs. Empirical

- Graph on left: Y-bar calculated from an ∞ number of samples from the "true distribution".
- Graph on right: {Y*-bar} calculated in each of 1000 re-samples from the *empirical* distribution.
- Analogy: $\mu : \overline{Y} \; :: \; \overline{Y} : \overline{Y}^*$





Summary

- ✓ The empirical distribution – your data – serves as a proxy to the “true” distribution.

- ✓ “Resampling” means (repeatedly) sampling with replacement.

- ✓ Resampling the data is analogous to the process of drawing the data from the “true distribution”.

- ✓ We can resample multiple times
  - Compute the statistic of interest T on each re-sample
  - We get an estimate of the distribution of T.

Motivating Example

- ✓Let's look at a simple case where we all know the answer in advance.

- ✓Pull 500 draws from the *n*(5000,100) dist.

- ✓The sample mean ≈ 5000
  - ✓Is a point estimate of the "true" mean μ.
  - ✓But how sure are we of this estimate?

- ✓From theory, we know that:

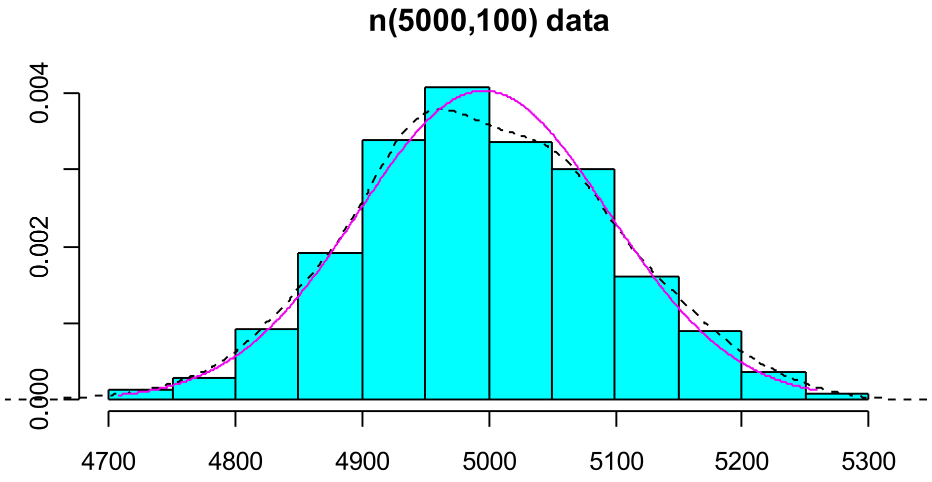
$$s.d.(\overline{X}) = \sigma / \sqrt{N} \approx 100 / \sqrt{500} \approx 4.47$$

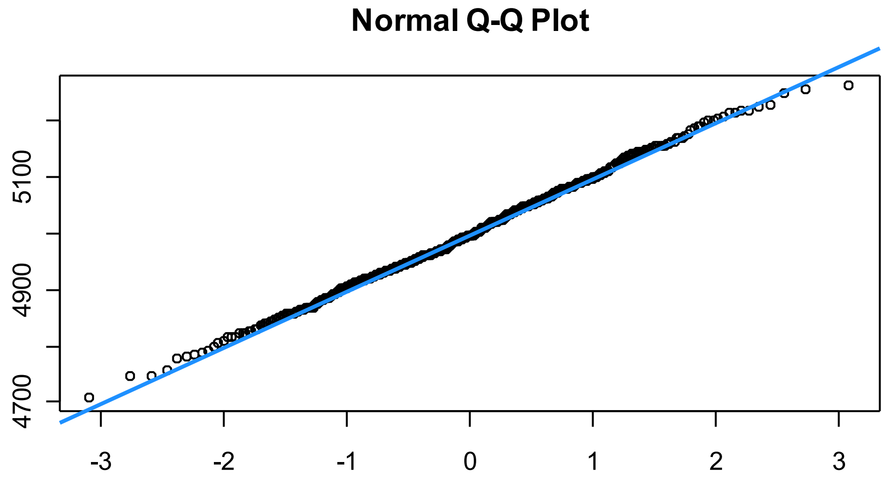
raw data	
statistic	**value**
#obs	500
mean	**4995.79**
sd	**98.78**
2.5%*ile*	4812.30
97.5%*ile*	5195.58

Visualizing the Raw Data

- ✓500 draws from *n*(5000,100)
- ✓Look at summary statistics, histogram, probability density estimate, QQ-plot.
- ✓… looks pretty normal

raw data	
statistic	**value**
#obs	500
mean	**4995.79**
sd	**98.78**
2.5%*ile*	4812.30
97.5%*ile*	5195.58





Sampling With Replacement

Now let's use resampling to estimate the s.d. of the sample mean (≈4.47)

- ✓ Draw a data point at random from the data set.
  Then throw it back in
- ✓ Draw a second data point.
  Then throw *it* back in…
- ✓ Keep going until we've got 500 data points.
  You might call this a "pseudo" data set.
- ✓ This is not merely re-sorting the data.
  Some of the original data points will appear more than once; others won't appear at all.

Sampling With Replacement

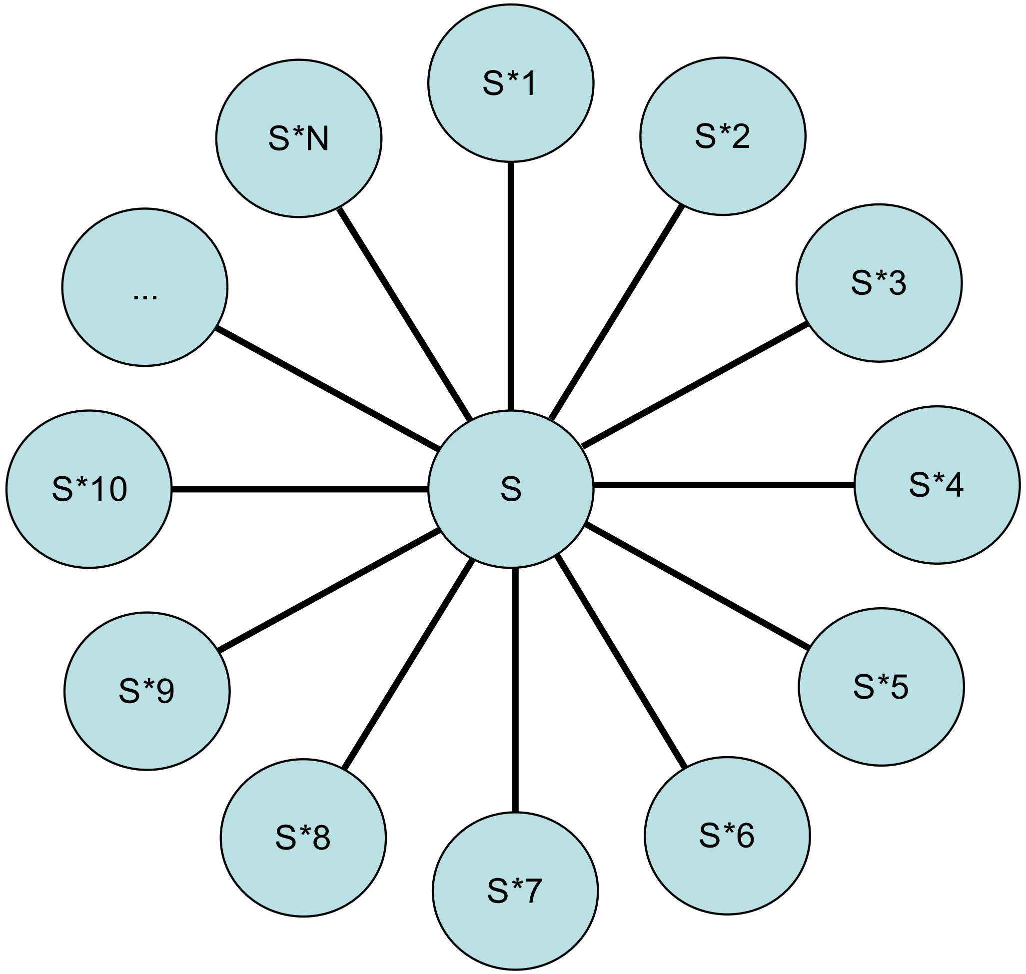
✓Sample with replacement 500 data points from the original dataset S

Call this S^*_1

✓Now do this 999 more times!

$S^*_1, S^*_2, \ldots, S^*_{1000}$

✓Compute X-bar on each of these 1000 samples.



Exemplo no R

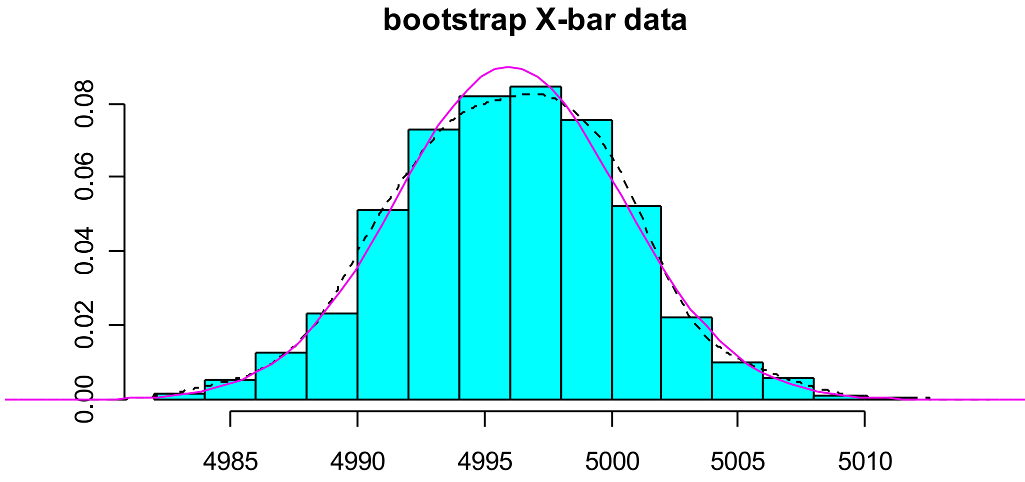

```
norm.data <- rnorm(500, mean=5000, sd=100)
boots <- function(data, R){
b.avg <<- c(); b.sd <<- c()
for(b in 1:R) {
        ystar <- sample(data,length(data),replace=T)
        b.avg <<- c(b.avg,mean(ystar))
        b.sd  <<- c(b.sd,sd(ystar))}
}
boots(norm.data, 1000)
```

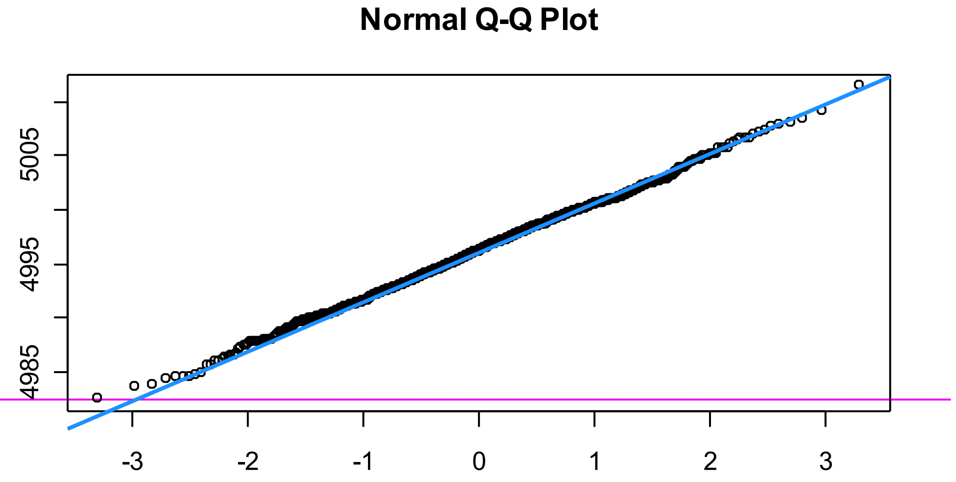
Results

- ✓From theory we know that

  X-bar ~ *n*(5000, 4.47)
- ✓Bootstrapping estimates this pretty well!
- ✓And we get an estimate of the *whole distribution*, not just a confidence interval.

raw data		X-bar	
statistic	**value**	**theory**	**bootstrap**
#obs	500	1,000	1,000
mean	**4995.79**	**5000.00**	**4995.98**
sd	**98.78**	**4.47**	**4.43**
2.5%*ile*	4705.08	4991.23	4987.60
97.5%*ile*	5259.27	5008.77	5004.82





Two Ways of Looking at a Confidence

Approximate normality assumption

- X-bar ±2*(bootstrap dist s.d.)

Percentile method

- Just take the desired percentiles of the bootstrap histogram.
- More reliable in cases of asymmetric bootstrap histograms.

```
 mean(norm.data) - 2 * sd(b.avg)
[1] 4986.926
 mean(norm.data) + 2 * sd(b.avg)
[1] 5004.661
```

raw data		X-bar	
statistic	**value**	**theory**	**bootstrap**
#obs	500	1,000	1,000
mean	**4995.79**	**5000.00**	**4995.98**
sd	**98.78**	**4.47**	**4.43**
2.5%*ile*	4705.08	4991.23	4987.60
97.5%*ile*	5259.27	5008.77	5004.82

OBRIGADO!