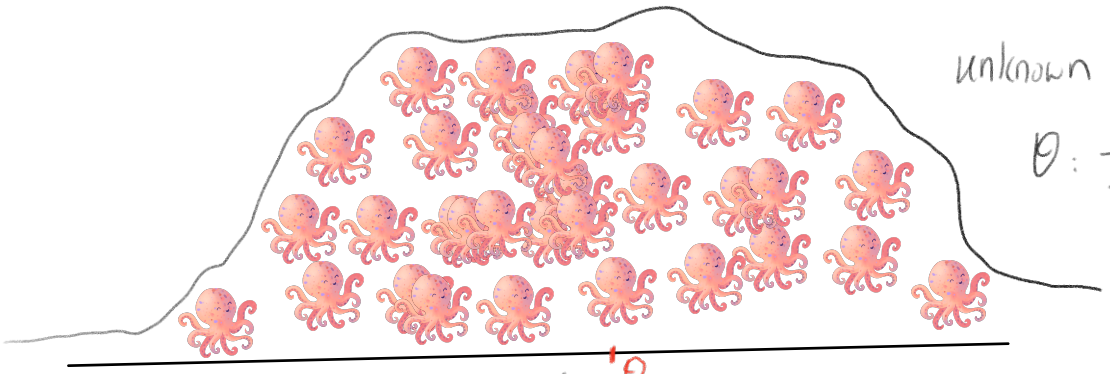
STA 104 Applied Nonparametric Statistics

Chapter 6: Bootstrap

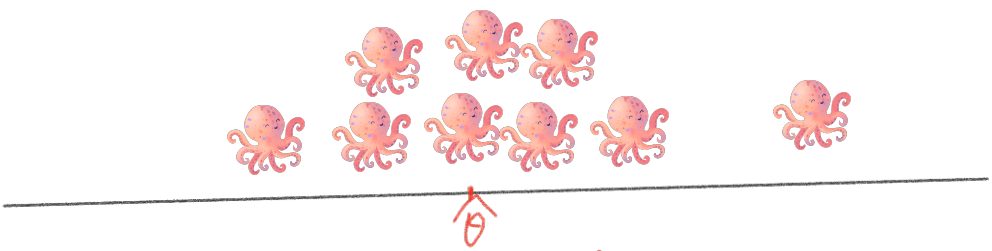
Xiner Zhou

Department of Statistics, University of California, Davis

Table of contents



observe one sample



how good is your estimate $\hat{\theta}$ for the truth θ_0 ?

① typically how far away your estimate $\hat{\theta}$ from θ_0 ?

$SE(\hat{\theta})$

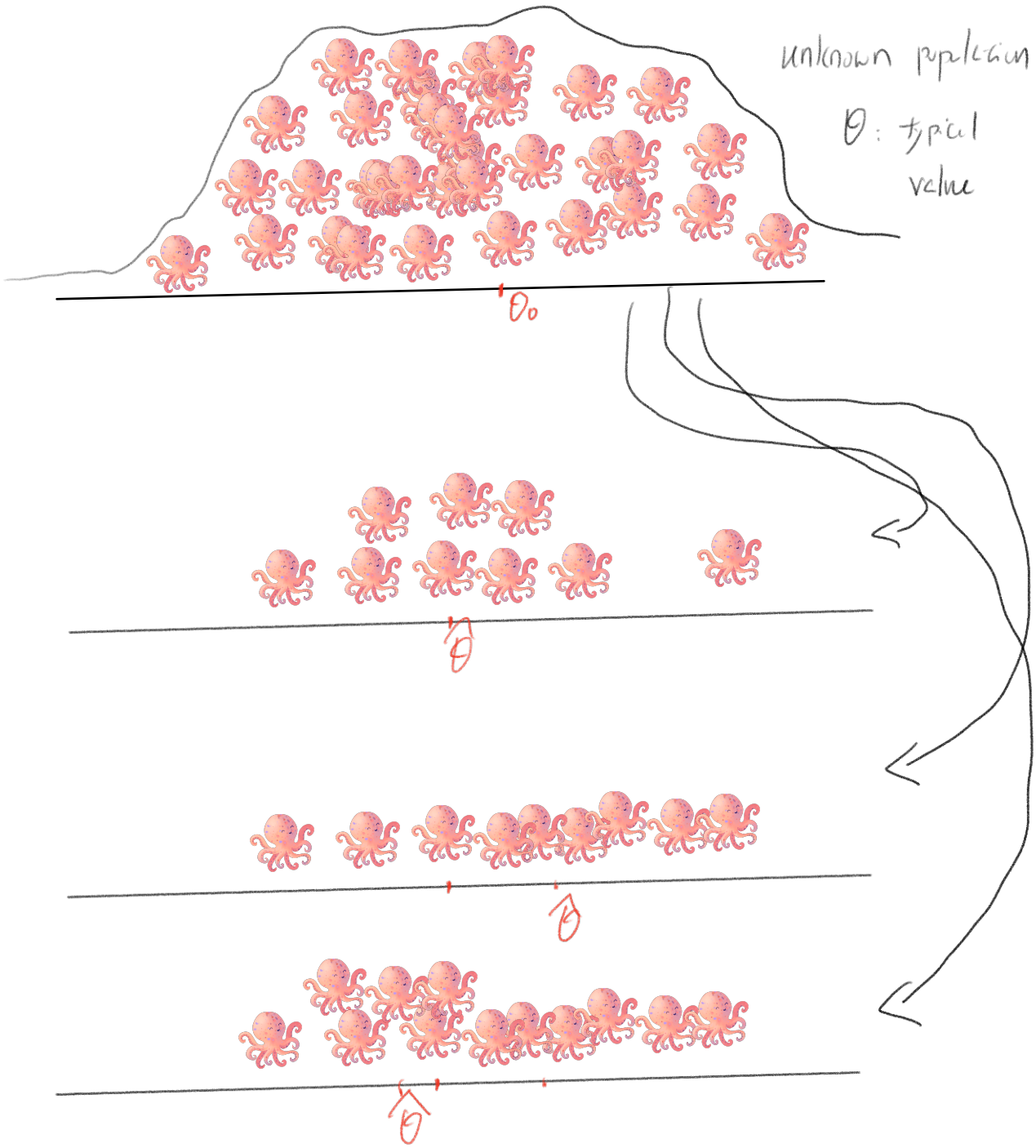
② Confidence interval : give an interval that contains the truth θ_0 almost surely / with very high probability ?

two measures to assess the "quality" of your estimate

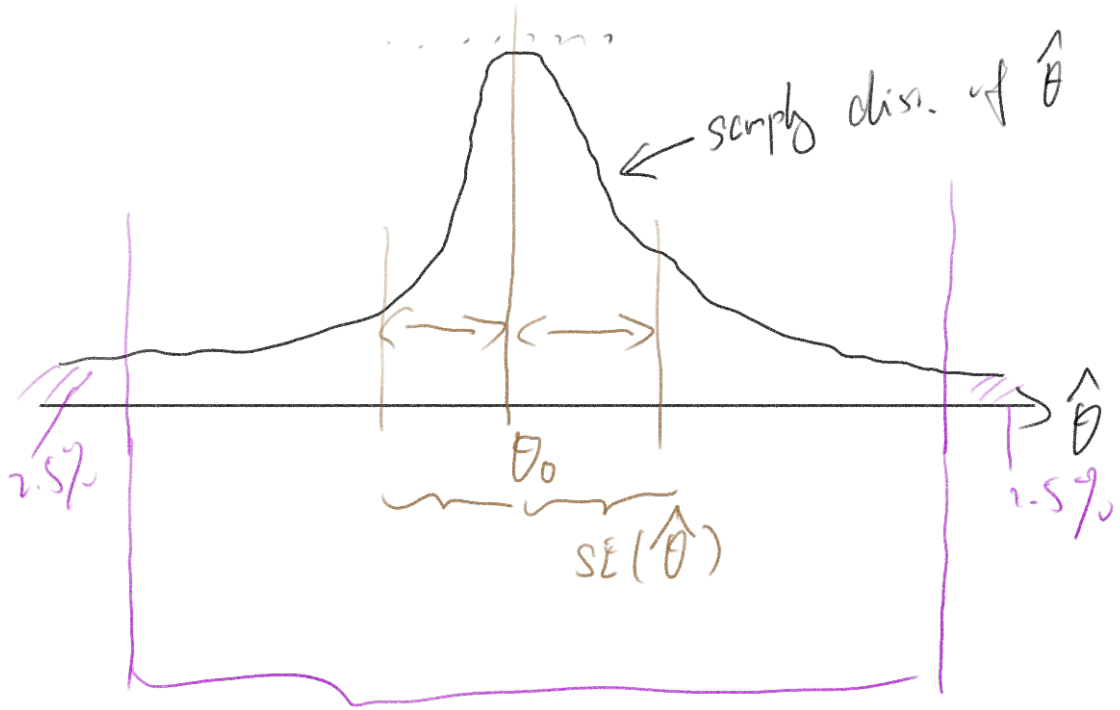
Frequentist's way = Repeated Sampling idea:

if we could repeatedly draw random samples from the underlying population, and estimate $\hat{\theta}$ each time.

⇒ "sampling distribution" of $\hat{\theta}$ describes its behavior!



⇒ repeatedly 1,000,000 times



Problem:

We can't repeated sample !

Example: One-sample mean μ

① $X_1 \cdots X_n \sim N(\mu, \sigma^2=1)$ △

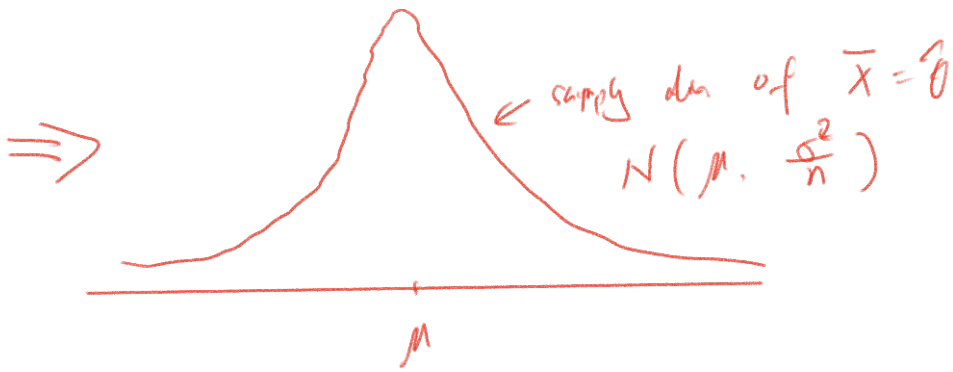
$\hat{\theta} : \bar{X}_n$

$SE(\hat{\theta}) = \sqrt{\frac{\sigma^2}{n}}$

CI for θ: $\bar{X}_n \pm 1.96 \sqrt{\frac{\sigma^2}{n}}$ ← s^2

↑ $t_{1-\alpha}$ if σ^2 unknown

CLT: Central Limit Theorem


⇒
sampling dist of $\bar{X} = \hat{\theta}$
$N(\mu, \frac{\sigma^2}{n})$
μ

A central element of frequentist inference is uncertainty quantification through the standard error or confidence interval.

- no theoretical result
- large sample approximations
  - in many cases (such as with the sample median), requires knowledge about the underlying distribution unknown in real data situations.
- Direct standard error formulas exist for various forms of averaging (sample mean, linear regression), but for hardly anything else.

$\Rightarrow$ modern computer-intensive, nonformulaic, statistical method for estimating quantities like standard error and confidence interval

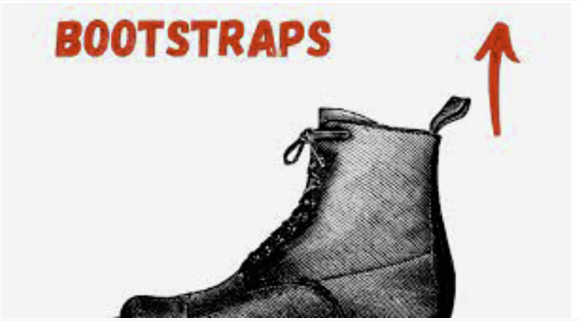


Figure 1: get oneself out of some situation using existing resources

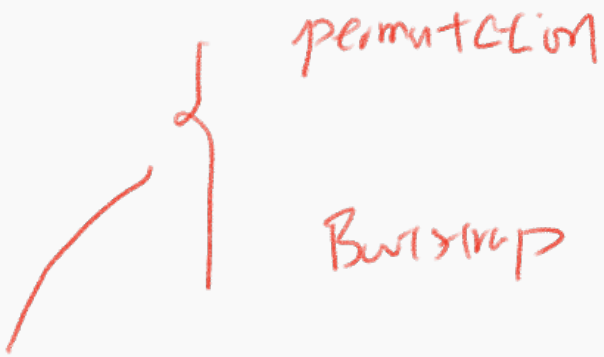
Figure 2: The word "bootstrap" comes from an old story about a hero - Baron Munchausen - who is riding around on his horse in a forest and suddenly gets stuck in a swamp. He screams for help but there is no one around who hears his voice! Luckily our hero does not give up and gets a great idea: "what if I just pull myself out of this swamp?". He grabs the straps of his boots and pulls himself loose. Fantastic - he just invented bootstrapping. Physics-defying stories aside, bootstrapping has become a common term for something seemingly impossible or counterintuitive.

Why Bootstrap?

- modern computer power
- automates a wide variety of inferential calculations, including standard errors, confidence interval.
- sparing statisticians the exhaustion of tedious routine calculations
- opened the door for more complicated estimation algorithms, so that their accuracy would be easily assessed.

Bootstrap for Assessing the Quality of Estimators: Variance and Standard Error

Motivation



Intuition: Resampling from your data to approximate resampling from a population.

- The standard error of an estimate $\hat{\theta} = s(\boldsymbol{x})$ is, ideally, the standard deviation we would observe by repeatedly sampling new versions of $\boldsymbol{x}$ from F.
- This is impossible since F is unknown.
- Instead, the bootstrap substitutes an estimate $\hat{F}$ for F and then estimates by direct simulation, a feasible tactic only since the advent of electronic computation.

Motivation

$\hat{\theta}$ is obtained in two steps: first $\boldsymbol{x}$ is generated by iid sampling from probability distribution F, and then $\hat{\theta}$ is calculated from $\boldsymbol{x}$ according to algorithm $s(\cdot)$,

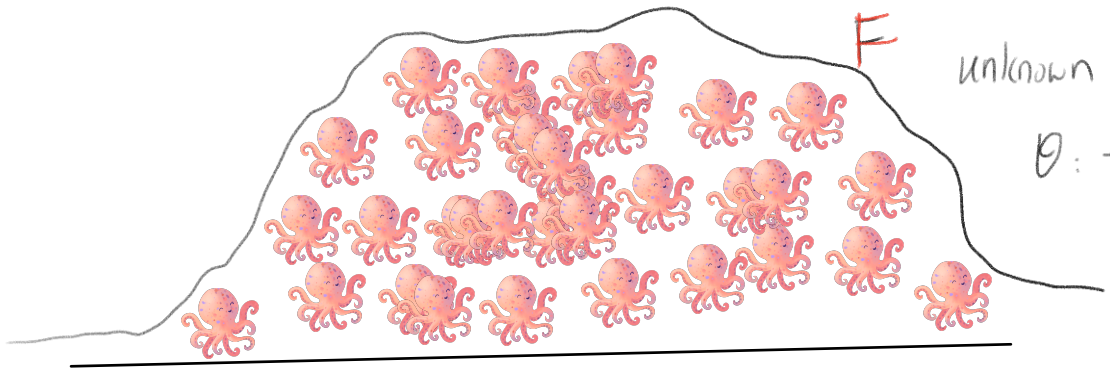
$$F \xrightarrow{\text{iid}} \boldsymbol{x} \xrightarrow{s} \hat{\theta}.$$

We don't know F, but we can estimate it by the empirical probability distribution $\hat{F}$ that puts probability $1/n$ on each point x_i.

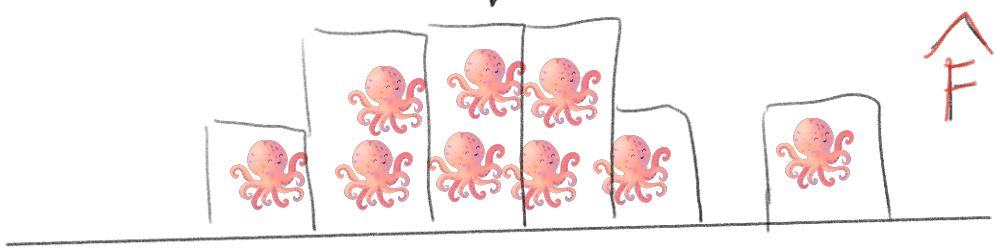
Bootstrap replications $\hat{\theta}^*$ are obtained

$$\hat{F} \xrightarrow{\text{iid}} x^* \xrightarrow{s} \hat{\theta}^*.$$

In the real world we only get to see the single value $\hat{\theta}$, but the bootstrap world is more generous: we can generate as many bootstrap replications $\hat{\theta}^{*b}$ as we want, or have time for, and directly estimate their variability.

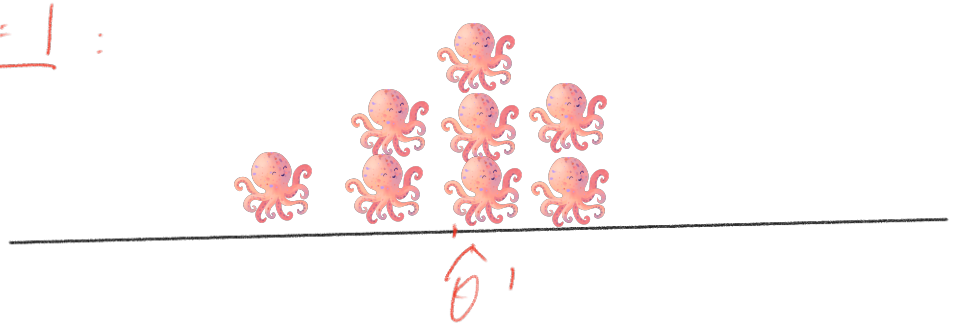


observe one sample

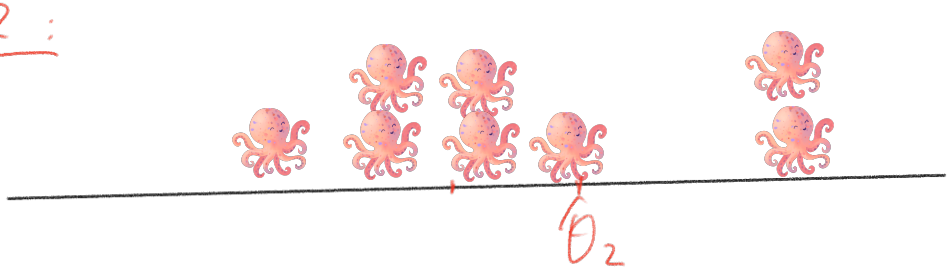


Bootstrap:

b=1:

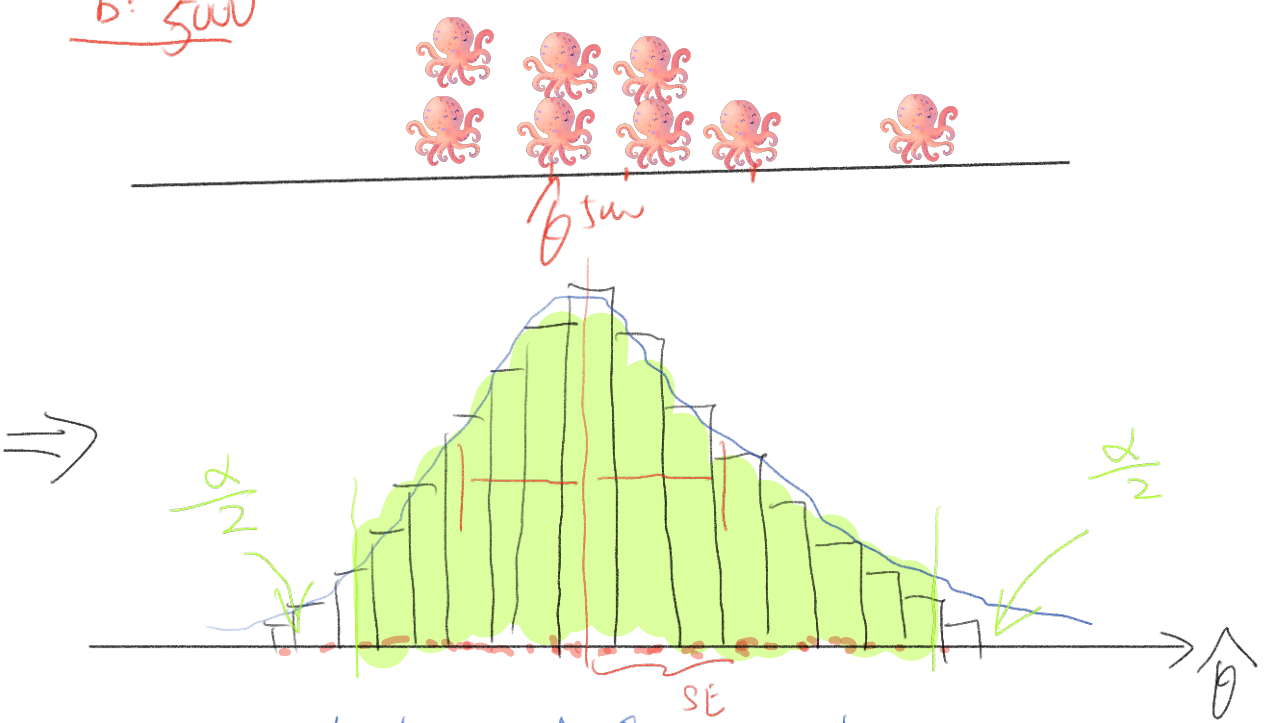


b=2:



repeate for B: 500 times;

b: $\hat{\theta}_{sw}$



distribution of Bootstrap replications

$\approx$ sampling distribution of $\hat{\theta}$

idea of Bootstrap CI:

Motivation

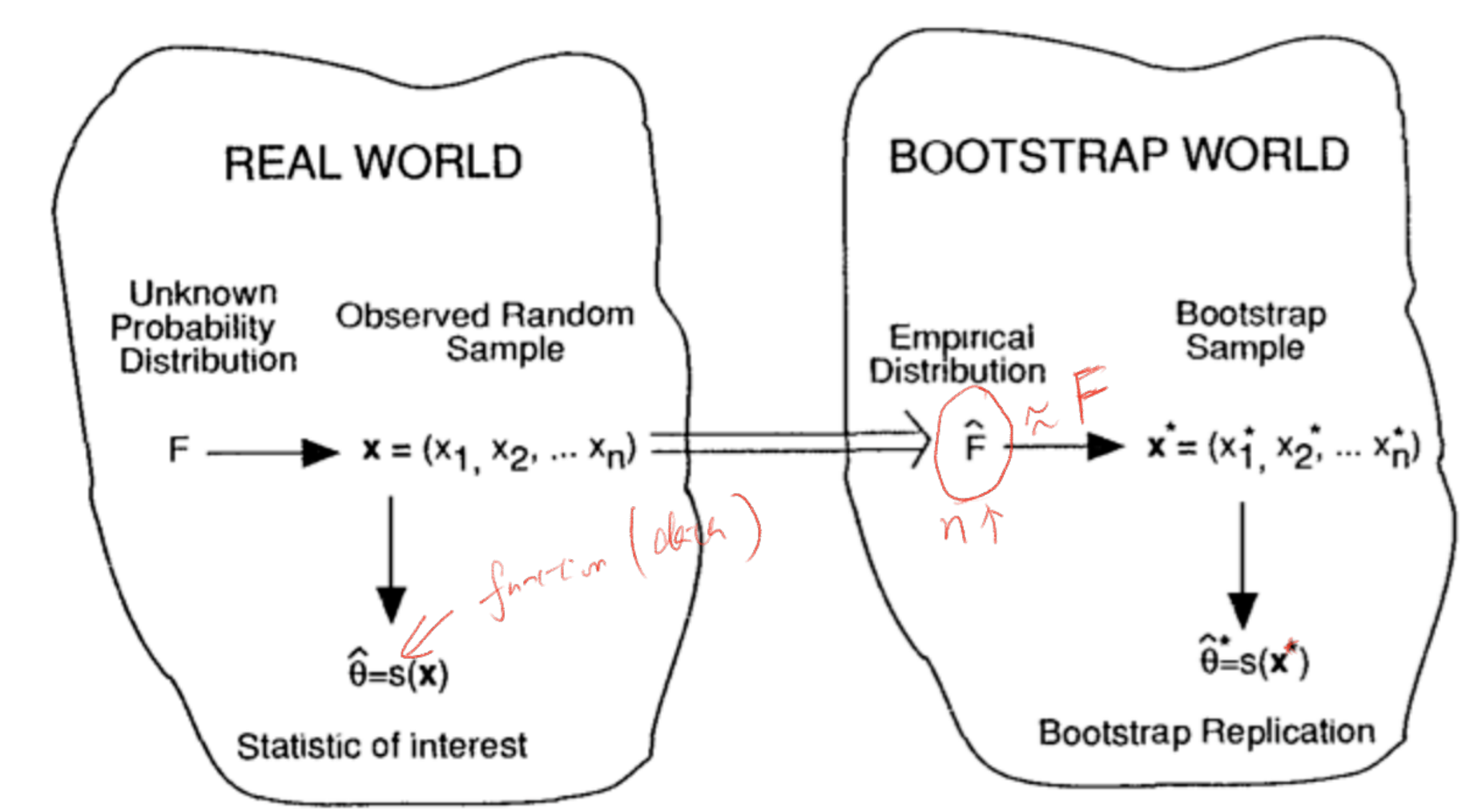


Figure 3: An Introduction to the Bootstrap (Efron Tibshirani, 1993).

Procedure

Bootstrap sample

The bootstrap estimate of standard error for a statistic $\hat{\theta} = s(\boldsymbol{x})$ computed from a data set $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ begins with the notion of a bootstrap sample

$$\boldsymbol{x}^* = (x_1^*, x_2^*, \ldots, x_n^*),$$

where each x_i^* is drawn randomly with equal probability and with replacement from $\{x_1, x_2, \ldots, x_n\}$.

Each bootstrap sample provides a bootstrap replication of the statistic of interest,

$$\hat{\theta}^* = s(\boldsymbol{x}^*).$$

Procedure

Bootstrap for SE

- Some large number B of bootstrap samples are independently drawn (say $B = 1000$). The corresponding bootstrap replications are calculated, say

$$\hat{\theta}^{*b} = s\left(\boldsymbol{x}^{*b}\right) \quad \text{for } b = 1, 2, \ldots, B.$$

- The resulting bootstrap estimate of standard error for $\hat{\theta}$ is the empirical standard deviation of the $\hat{\theta}^{*b}$ values,

$$\widehat{\text{se}}_{\text{boot}} = \left[\sum_{b=1}^{B}\left(\hat{\theta}^{*b} - \hat{\theta}^{*\cdot}\right)^2/(B-1)\right]^{1/2}, \quad \text{with } \hat{\theta}^{*\cdot} = \sum_{b=1}^{B}\hat{\theta}^{*b}/B.$$

$\approx SE(\hat{\theta})$

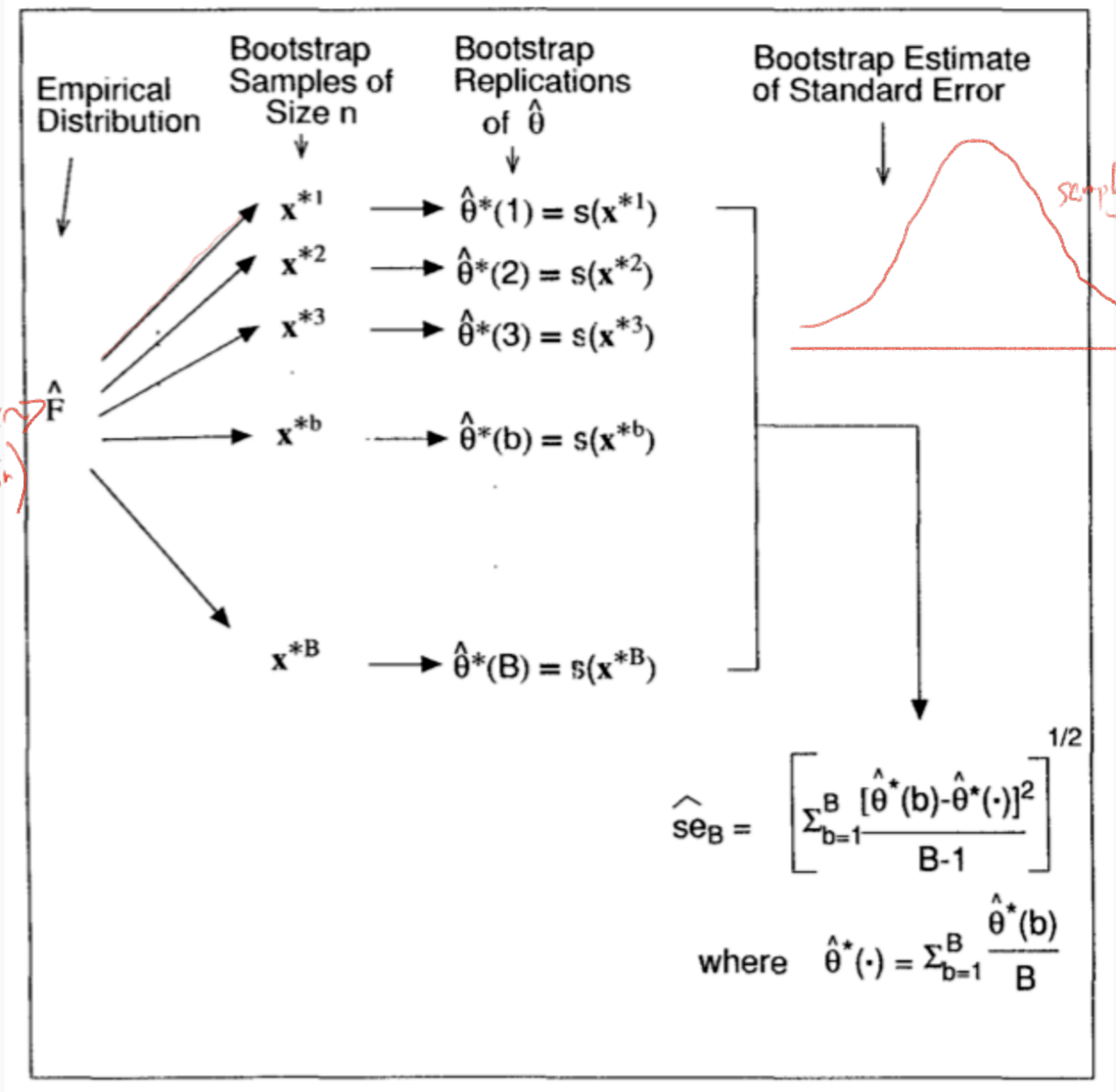
definition of SD

data : $X_1 \cdots X_n$

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

$$\sqrt{\frac{\sum_{b=1}^{B}(\hat{\theta}^{*b} - \hat{\theta}^{*\cdot})^2}{B-1}}$$

Procedure


Empirical Distribution
Bootstrap Samples of Size n
Bootstrap Replications of $\hat{\theta}$
Bootstrap Estimate of Standard Error
$\hat{F}$
x^{*1} → $\hat{\theta}^*(1) = s(x^{*1})$
x^{*2} → $\hat{\theta}^*(2) = s(x^{*2})$
x^{*3} → $\hat{\theta}^*(3) = s(x^{*3})$
x^{*b} → $\hat{\theta}^*(b) = s(x^{*b})$
x^{*B} → $\hat{\theta}^*(B) = s(x^{*B})$
$\widehat{se}_B = \left[\Sigma_{b=1}^{B} \frac{[\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)]^2}{B-1}\right]^{1/2}$
where $\hat{\theta}^*(\cdot) = \Sigma_{b=1}^{B} \frac{\hat{\theta}^*(b)}{B}$

Bootstrap Confidence Intervals

What is a confidence interval?

$P(L(\text{data}) < \theta < U(\text{data})) = \underbrace{1-\alpha}_{95\%}$. $\alpha = 0.05$

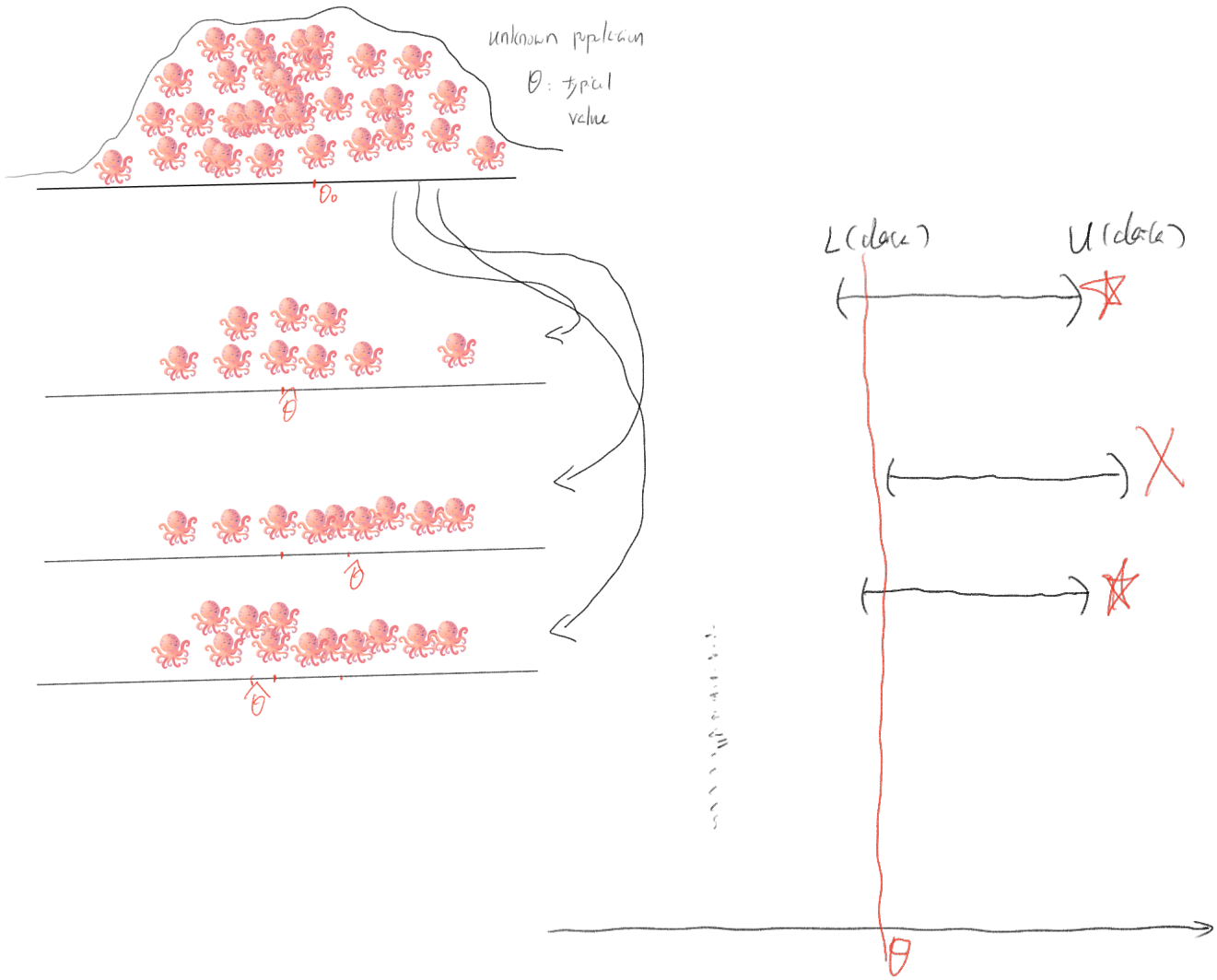
(fixed: θ; random: $L(\text{data})$, $U(\text{data})$)

defines a $100(1-\alpha)\%$ C.I. for θ

a range of values, depending on data,

s.t. true θ lies within this interval

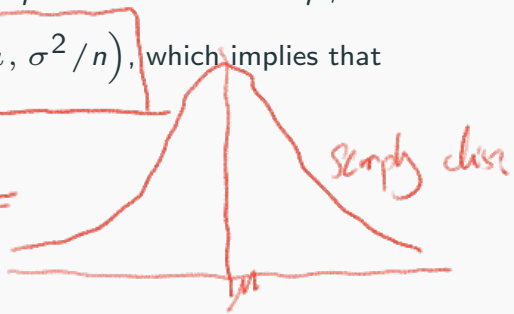
is $100(1-\alpha)\%$


unknown population
θ: typical value
θ₀
θ
θ
θ
L(data)
U(data)
X
θ

Example 1: Confidence Interval for μ

Suppose that $x_i \overset{\text{iid}}{\sim} N\left(\mu, \sigma^2\right)$ for $i = 1, \ldots, n$ and we want to form a confidence interval for μ. As an estimate of μ, we will use the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Assuming that $x_i \overset{\text{iid}}{\sim} N\left(\mu, \sigma^2\right)$, we know that $\bar{x} \sim N\left(\mu, \sigma^2/n\right)$, which implies that $\sqrt{n}(\bar{x} - \mu)/\sigma \sim N(0, 1)$. As a result, we have that

$$P\left(z_{\alpha/2} < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < z_{1-\alpha/2}\right) = 1 - \alpha$$



where $z_\alpha = \Phi^{-1}(\alpha)$ with $\Phi^{-1}(\cdot)$ denoting the quantile function for the standard normal distribution. Rearranging the terms inside the above probability statement gives

$$\begin{aligned}
1 - \alpha &= P\left(z_{\alpha/2}\sigma/\sqrt{n} < \bar{x} - \mu < z_{1-\alpha/2}\sigma/\sqrt{n}\right) \\
&= P\left(z_{\alpha/2}\sigma/\sqrt{n} - \bar{x} < -\mu < z_{1-\alpha/2}\sigma/\sqrt{n} - \bar{x}\right) \\
&= P\left(\bar{x} + z_{\alpha/2}\sigma/\sqrt{n} > \mu > \bar{x} - z_{1-\alpha/2}\sigma/\sqrt{n}\right)
\end{aligned}$$

which implies that a $100(1-\alpha)\%$ confidence interval for μ defines $a(\bar{x}) = \bar{x} - z_{1-\alpha/2}\sigma/\sqrt{n}$ and $b(\bar{x}) = \bar{x} - z_{\alpha/2}\sigma/\sqrt{n}$. Note that since $-z_{\alpha/2} = z_{1-\alpha/2}$ we can write the two endpoints of the confidence interval as

$$\bar{x} \pm z_{1-\alpha/2}\mathrm{SE}(\bar{x})$$

where $\mathrm{SE}(\bar{x}) = \sigma/\sqrt{n}$ is the standard error of the sample mean. In practice, it is typical to form a 90% confidence interval (i.e., $\alpha = 0.1$), which corresponds to $z_{0.95} \approx 1.65$, a 95% confidence interval (i.e., $\alpha = 0.05$), which corresponds to $z_{0.975} \approx 1.96$, or a 99% confidence interval (i.e., $\alpha = 0.01$), which corresponds to $z_{0.995} = 2.58$.

Forming a confidence interval for μ with $x_i \overset{\text{iid}}{\sim} N\left(\mu, \sigma^2\right)$, here is a simple demonstration of forming a 95% confidence interval using $R = 10000$ replications with $n = 25$ observations.

```
 R <- 10000
n <- 25
set.seed(1)
xbar <- replicate(R, mean(rnorm(n)))
ci.lo <- xbar - qnorm(.975) / sqrt(n)      # 95% CI lower bound
ci.up <- xbar - qnorm(.025) / sqrt(n)      # 95% CI upper bound
ci.in <- (ci.lo <= 0) & (0 <= ci.up)
mean(ci.in)

## [1] 0.9501
```

Example 2: Confidence Interval for σ^2

Suppose that $x_i \overset{\text{iid}}{\sim} N\left(\mu, \sigma^2\right)$ for $i = 1, \ldots, n$ and we want to form a confidence interval for σ^2. As an estimate of σ^2, we will use the sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$. Assuming that $x_i \overset{\text{iid}}{\sim} N\left(\mu, \sigma^2\right)$, we know that $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$. As a result, we have that

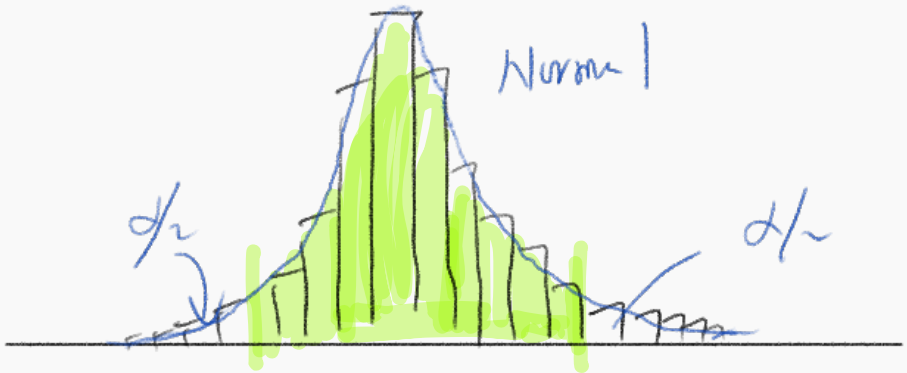
$$P\left(q_{n-1;\alpha/2} < (n-1)\frac{s^2}{\sigma^2} < q_{n-1;1-\alpha/2}\right) = 1 - \alpha$$

where $q_{n-1;\alpha} = Q_{n-1}(\alpha)$ with $Q_{n-1}(\cdot)$ denoting the quantile function for the χ^2_{n-1} distribution. Rearranging the terms inside the above probability statement gives

$$\begin{aligned}
1 - \alpha &= P\left(\frac{q_{n-1;\alpha/2}}{n-1} < \frac{s^2}{\sigma^2} < \frac{q_{n-1;1-\alpha/2}}{n-1}\right) \\
&= P\left(\frac{q_{n-1;\alpha/2}}{s^2(n-1)} < \frac{1}{\sigma^2} < \frac{q_{n-1;1-\alpha/2}}{s^2(n-1)}\right) \\
&= P\left(\frac{s^2(n-1)}{q_{n-1;\alpha/2}} > \sigma^2 > \frac{s^2(n-1)}{q_{n-1;1-\alpha/2}}\right)
\end{aligned}$$

which implies that a $100(1-\alpha)\%$ confidence interval for σ^2 defines $a\left(s^2\right) = (n-1)s^2/q_{n-1;1-\alpha/2}$ and $b\left(s^2\right) = (n-1)s^2/q_{n-1;\alpha/2}$.

Bootstrap CI: Normal Approximation



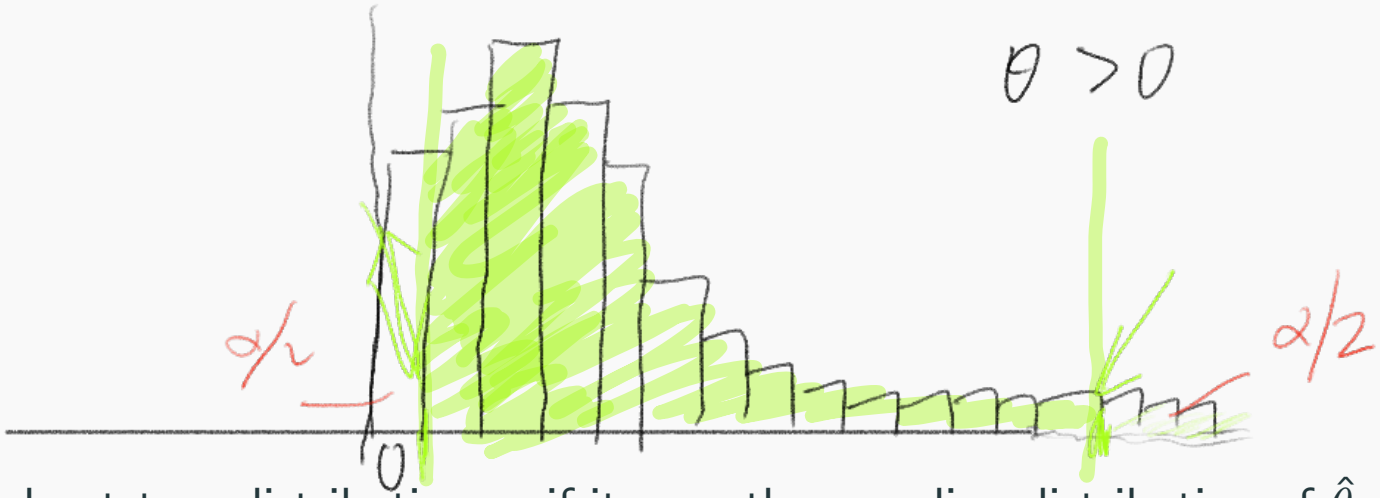
The normal approximation confidence interval uses the classic confidence interval formula (for the mean), but replaces the standard error with the bootstrap estimate of the standard error.

Specifically, the normal approximation interval has the form

$$\hat{\theta} \pm Z_{1-\alpha/2} \widehat{SE}(\hat{\theta})$$

where $Z_{1-\alpha/2}$ is the quantile of the standard normal distribution that cuts-off $\alpha/2$ in the upper tail (e.g., $Z_{1-\alpha/2} = 1.96$ for a 95% interval), and $\widehat{\mathrm{SE}}(\hat{\theta})$ is the bootstrap estimate of the standard error of $\hat{\theta}$.

Bootstrap CI: Percentile Method



Simply uses the bootstrap distribution as if it were the sampling distribution of $\hat{\theta}$.

The percentile method defines the $100(1-\alpha)\%$ confidence interval for θ as

$$\left[Q^*_{\alpha/2}, Q^*_{1-\alpha/2}\right]$$

where $Q^*_{\alpha/2}$ and $Q^*_{1-\alpha/2}$ denote the quantiles of the bootstrap distribution of $\hat{\theta}$.

Example: To estimate the correlation between Petal Length and Petal Width

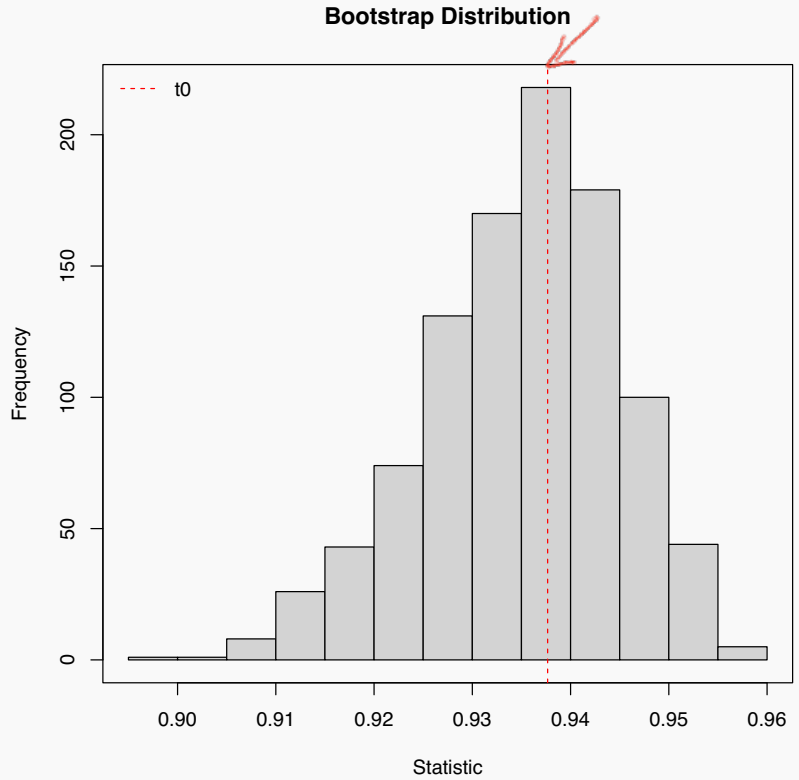
```
library(boot)

# Custom function to find correlation
# between the Petal Length and Width
corr.fun <- function(data, idx)
{
  df <- data[idx, ]

  # Find the spearman correlation between
  # the 3rd and 4th columns of dataset
  c(cor(df[, 3], df[, 4], method = 'spearman'))
}

bootstrap <- boot(iris, corr.fun, R = 1000)
bootstrap
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = iris, statistic = corr.fun, R = 1000)
##
##
## Bootstrap Statistics :
##      original       bias    std. error
## t1* 0.9376668 -0.00274143 0.00980653
```

spearman correlation

```r
# bootstrap distribution
hist(bootstrap$t, xlab = "Statistic", main = "Bootstrap Distribution")
box()
abline(v = bootstrap$t0, lty = 2, col = "red")
legend("topleft", "t0", lty = 2, col = "red", bty = "n")
```

observed



```r
# Function to find the bootstrap Confidence Intervals
boot.ci(boot.out = bootstrap,
        type = c("norm",
                 "perc", "bca"))
```

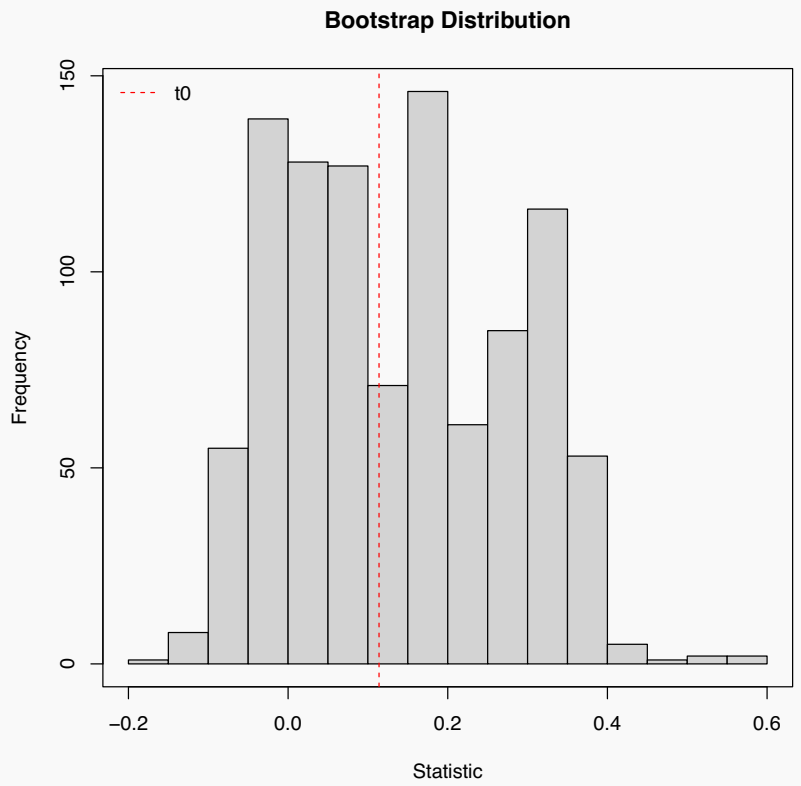
```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = bootstrap, type = c("norm", "perc", "bca"))
##
## Intervals :
## Level      Normal             Percentile            BCa
## 95%   ( 0.9212,  0.9596 )   ( 0.9126,  0.9522 )   ( 0.9173,  0.9539 )
## Calculations and Intervals on Original Scale
```

Example: To estimate Median

We will generate $n = 100$ observations from a standard normal distribution, and use the median as the parameter/statistic of interest.

```
library(boot)

# generate 100 standard normal observations
set.seed(1)
n <- 100
x <- rnorm(n)
sim.data=data.frame(x=x)

median.fun <- function(data, idx)
{
  df <- data[idx, ]
  quantile(df,prob=0.5)
}

bootstrap <- boot(sim.data, median.fun, R = 1000)
bootstrap


##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = sim.data, statistic = median.fun, R = 1000)
##
##
## Bootstrap Statistics :
##      original     bias    std. error
## t1* 0.1139092 0.02243135   0.1413482
```

```r
# bootstrap distribution
hist(bootstrap$t, xlab = "Statistic", main = "Bootstrap Distribution")
box()
abline(v = bootstrap$t0, lty = 2, col = "red")
legend("topleft", "t0", lty = 2, col = "red", bty = "n")
```



```r
# Function to find the bootstrap Confidence Intervals
boot.ci(boot.out = bootstrap,
        type = c("norm",
                 "perc", "bca"))


## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = bootstrap, type = c("norm", "perc", "bca"))
##
## Intervals :
## Level      Normal             Percentile            BCa
## 95%   (-0.1856,  0.3685 )   (-0.0593,  0.3788 )   (-0.0811,  0.3692 )
## Calculations and Intervals on Original Scale
```