STA 104 Applied Nonparametric Statistics

Chapter 6: Bootstrap

Xiner Zhou

Department of Statistics, University of California, Davis

Table of contents

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

2. Bootstrap Confidence Intervals

Bootstrap Confidence Intervals

What is a confidence interval?

$$\underbrace{P\big(L(\underline{\text{data}}) < \overset{\text{fixed}}{\theta} < U(\underline{\text{data}})\big) = \underbrace{1-\alpha}_{95\%}}_{\text{defines a } 100(1-\alpha)\%\ \underbrace{\text{C.I.}}\ \text{for } \theta}$$

random

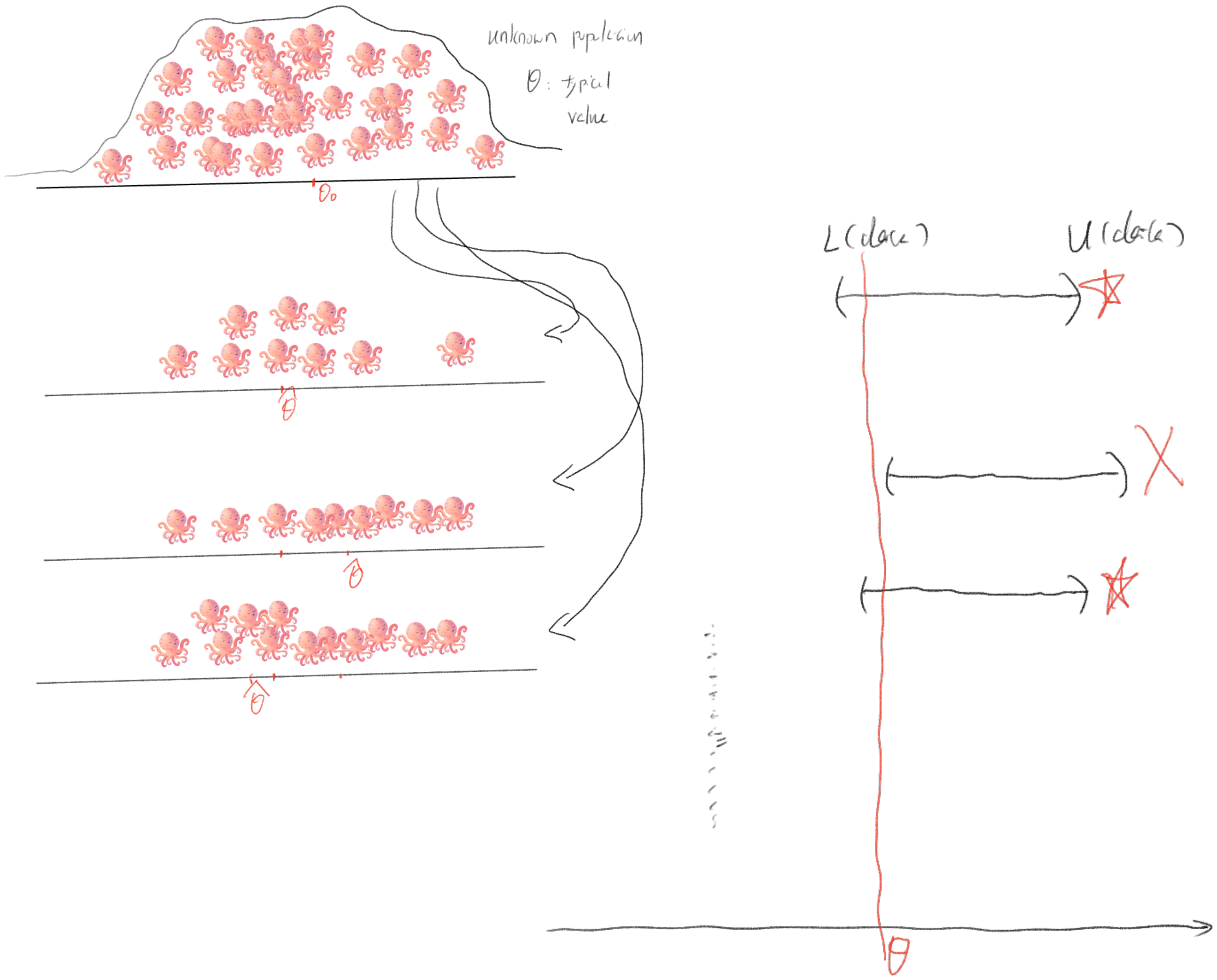
⚠

$. \ \alpha = 0.05$

a range of values, depending on data,

s.t. true θ lies within this interval

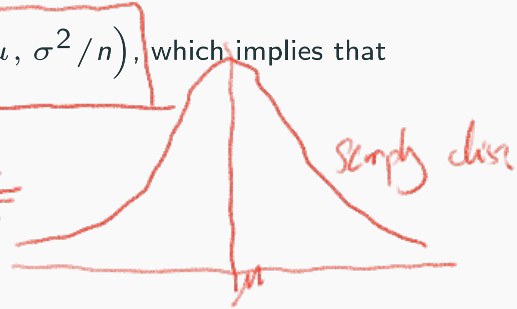
is $100(1-\alpha)\%$


unknown population
θ: typical value
θ₀
θ̂
θ̂
θ̂
L(data)
U(data)
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 $\boxed{\bar{x} \sim N\left(\mu, \sigma^2/n\right)}$, which implies that $\sqrt{n}(\bar{x} - \mu)/\sigma \sim N(0, 1)$. As a result, we have that

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



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

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

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

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

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
```

← check $\theta \in CI$.

Example 2: Confidence Interval for σ^2

Suppose that $x_i \overset{\text{iid}}{\sim} N\left(\mu, \boxed{\sigma^2}\right)$ 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 $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$. Assuming that $\boxed{x_i \overset{\text{iid}}{\sim} N\left(\mu, \sigma^2\right)}$, we know that $\boxed{(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}.}$
As a result, we have that

need to know

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

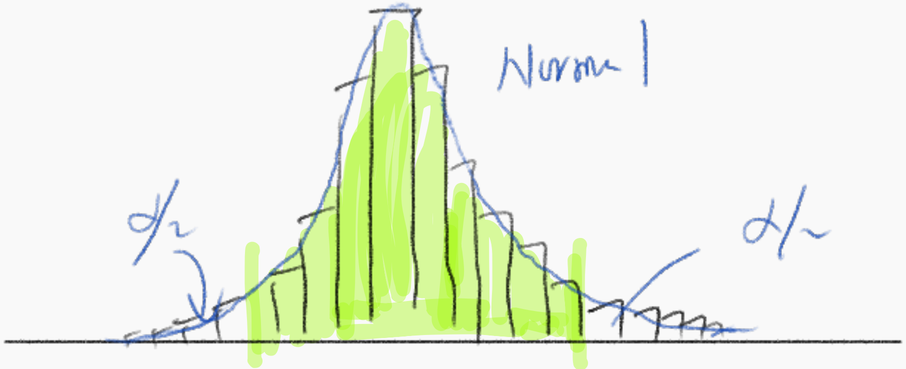
where $q_{n-1;\alpha} = Q_{n-1}(\alpha)$ with $Q_{n-1}(\cdot)$ denoting the quantile function for the χ^2_{n-1} distribution. Rearranging the terms inside the above probability statement gives

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

here

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



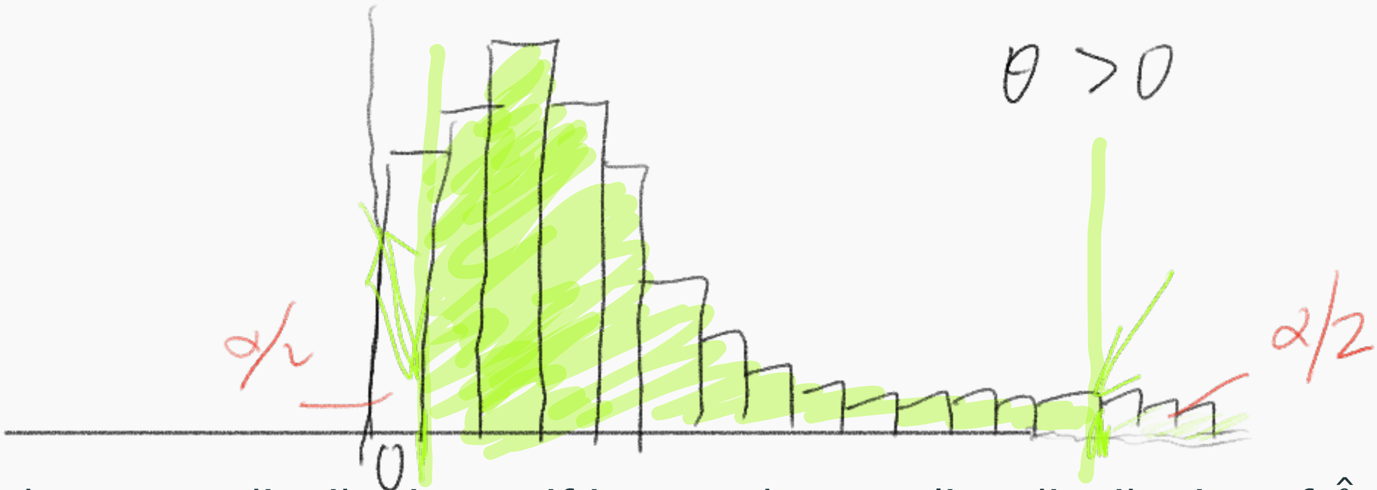
The normal approximation confidence interval uses the classic confidence interval formula (for the mean), but replaces the standard error with the bootstrap estimate of the standard error.

Specifically, the normal approximation interval has the form

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

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

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

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
```

index of re-samples

data.frame

resample

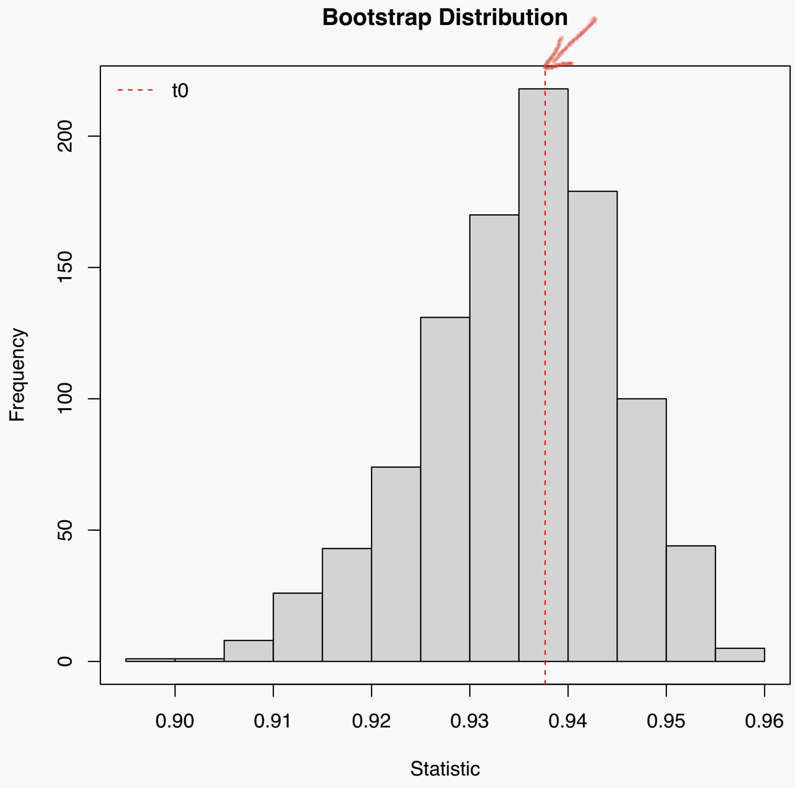
orig'l data

```
##
## 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

```
# 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.

```r
library(boot)
```

← package

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

generate data

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

← median

calculate median $\hat{\theta}$ for each Bootstrap samples.

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

↑ original data ↑ ↑ 5000

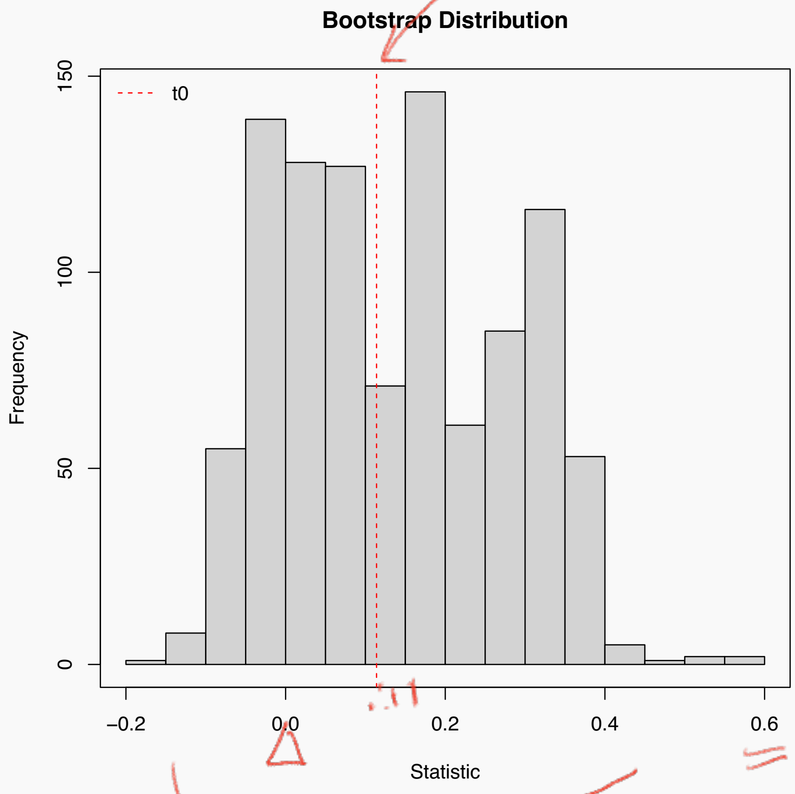
```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = sim.data, statistic = median.fun, R = 1000)
##
##
## Bootstrap Statistics :
##      original      bias    std. error
## t1* 0.1139092 0.02243135   0.1413482
```

$\widehat{SE}_{boot} = \widehat{SE}(\hat{\theta})$

$\theta_0 = 0$. $\hat{\theta} = 0.11$

```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")
```

$\hat{\theta}$



< 0

> 0

output object from boot( )

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

norm < percentile < 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
```