Resampling Techniques and their Application

-Class 4-

Frank Konietschke

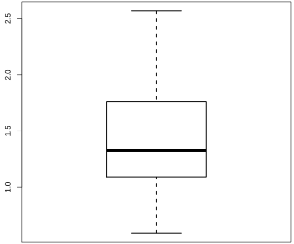
Institut für Biometrie und Klinische Epidemiologie

Charité - Universitätsmedizin Berlin, Berlin

frank.konietschke@charite.de

Motivation and Examples-II

The diameter of cork of a Champagne bottle is supposed to be 1.5 cm. If the cork is either too large or too small, it will not fit in the bottle. The manufacturer measures the diameter in a random sample of **n=36** bottles and obtains:

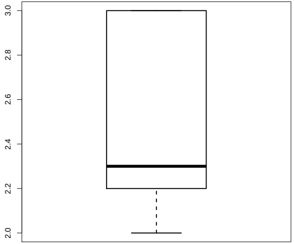



```
x=c(
0.59, 1.23, 1.00, 0.84, 0.88, 1.71,
1.81, 1.84, 2.03, 1.39, 1.30, 1.31,
1.96, 1.33, 2.57, 1.19, 1.01, 2.06,
1.32, 1.55, 1.28, 0.93, 1.63, 1.24,
1.83, 1.81, 0.94, 1.46, 1.25, 1.56,
0.61, 0.83, 1.17, 2.24, 1.68, 1.51)
```

- Data Analysis: Confidence interval and t-test

Motivation and Examples-III

A researcher measures the reaction time of $n = 10$ mice to signal pain when a stitch is applied to their tail.




```
x=c(
2.4, 3.0, 3.0, 2.2, 2.2,
2.2, 2.2, 2.8, 2.0, 3.0)
```

- Data Analysis: Confidence interval and t-test

Example

The diameter of cork of a Champagne bottle is supposed to be 1.5 cm. If the cork is either too large or too small, it will not fit in the bottle. The manufacturer measures the diameter in a random sample of n=36 bottles. Is there evidence at 1% level that the true mean diameter has moved away from the target?

- Hypothesis: $H_0 : \mu = 1.5$ vs. $H_1 : \mu \neq 1.5$

```
mx <- mean(x)
vx <- var(x)
T <- sqrt(36)*(mx-1.5)/sqrt(vx)
p <- 2*min(pt(T,35),1-pt(T,35))
```

Example

The diameter of cork of a Champagne bottle is supposed to be 1.5 cm. If the cork is either too large or too small, it will not fit in the bottle. The manufacturer measures the diameter in a random sample of n=36 bottles. Is there evidence at 1% level that the true mean diameter has moved away from the target?

- Hypothesis: $H_0 : \mu = 1.5$ vs. $H_1 : \mu \neq 1.5$
- Estimator: $\widehat{\mu} = 1.4136$

```
mx <- mean(x)
vx <- var(x)
T <- sqrt(36)*(mx-1.5)/sqrt(vx)
p <- 2*min(pt(T,35),1-pt(T,35))
```

Example

The diameter of cork of a Champagne bottle is supposed to be 1.5 cm. If the cork is either too large or too small, it will not fit in the bottle. The manufacturer measures the diameter in a random sample of n=36 bottles. Is there evidence at 1% level that the true mean diameter has moved away from the target?

- Hypothesis: $H_0 : \mu = 1.5$ vs. $H_1 : \mu \neq 1.5$
- Estimator: $\widehat{\mu} = 1.4136$
- Standard deviation: $\widehat{\sigma} = 0.46$

```
mx <- mean(x)
vx <- var(x)
T <- sqrt(36)*(mx-1.5)/sqrt(vx)
p <- 2*min(pt(T,35),1-pt(T,35))
```

Example

The diameter of cork of a Champagne bottle is supposed to be 1.5 cm. If the cork is either too large or too small, it will not fit in the bottle. The manufacturer measures the diameter in a random sample of n=36 bottles. Is there evidence at 1% level that the true mean diameter has moved away from the target?

```
mx <- mean(x)
vx <- var(x)
T <- sqrt(36)*(mx-1.5)/sqrt(vx)
p <- 2*min(pt(T,35),1-pt(T,35))
```

- Hypothesis: $H_0 : \mu = 1.5$ vs. $H_1 : \mu \neq 1.5$
- Estimator: $\widehat{\mu} = 1.4136$
- Standard deviation: $\widehat{\sigma} = 0.46$
- Test statistic: $T = \sqrt{36} * \frac{1.4136 - 1.5}{0.46} = -1.13$

Example

The diameter of cork of a Champagne bottle is supposed to be 1.5 cm. If the cork is either too large or too small, it will not fit in the bottle. The manufacturer measures the diameter in a random sample of n=36 bottles. Is there evidence at 1% level that the true mean diameter has moved away from the target?

```
mx <- mean(x)
vx <- var(x)
T <- sqrt(36)*(mx-1.5)/sqrt(vx)
p <- 2*min(pt(T,35),1-pt(T,35))
```

- Hypothesis: $H_0 : \mu = 1.5$ vs. $H_1 : \mu \neq 1.5$
- Estimator: $\widehat{\mu} = 1.4136$
- Standard deviation: $\widehat{\sigma} = 0.46$
- Test statistic: $T = \sqrt{36} * \frac{1.4136 - 1.5}{0.46} = -1.13$
- p-value: 2 times the area to the **left** of T under the t curve with 35 df. (Here, p=0.27)

Example

The diameter of cork of a Champagne bottle is supposed to be 1.5 cm. If the cork is either too large or too small, it will not fit in the bottle. The manufacturer measures the diameter in a random sample of n=36 bottles. Is there evidence at 1% level that the true mean diameter has moved away from the target?

```
mx <- mean(x)
vx <- var(x)
T <- sqrt(36)*(mx-1.5)/sqrt(vx)
p <- 2*min(pt(T,35),1-pt(T,35))
```

- Hypothesis: $H_0 : \mu = 1.5$ vs. $H_1 : \mu \neq 1.5$
- Estimator: $\widehat{\mu} = 1.4136$
- Standard deviation: $\widehat{\sigma} = 0.46$
- Test statistic: $T = \sqrt{36} * \frac{1.4136 - 1.5}{0.46} = -1.13$
- p-value: 2 times the area to the **left** of T under the t curve with 35 df. (Here, p=0.27)
- Quality of the estimator? Is the method valid? Is $p = 0.27$ a good estimate?

Example

A researcher measures the reaction time of $n = 10$ mice to signal pain when a stitch is applied to their tail. Compute a 95% confidence interval.

```
x=c(2.4, 3.0, 3.0, 2.2, 2.2,
2.2, 2.2, 2.8, 2.0, 3.0)
mx <- mean(x)
vx <- var(x)
crit <- qt(0.975,9)
lower <- mx-crit/sqrt(10)*sqrt(vx)
upper <- mx+crit/sqrt(10)*sqrt(vx)
```

- Estimator: $\widehat{\mu} = 2.5$

Example

A researcher measures the reaction time of $n = 10$ mice to signal pain when a stitch is applied to their tail. Compute a 95% confidence interval.

```
x=c(2.4, 3.0, 3.0, 2.2, 2.2,
2.2, 2.2, 2.8, 2.0, 3.0)
mx <- mean(x)
vx <- var(x)
crit <- qt(0.975,9)
lower <- mx-crit/sqrt(10)*sqrt(vx)
upper <- mx+crit/sqrt(10)*sqrt(vx)
```

- Estimator: $\widehat{\mu} = 2.5$
- Standard deviation: $\widehat{\sigma} = 0.40$

Example

A researcher measures the reaction time of $n = 10$ mice to signal pain when a stitch is applied to their tail. Compute a 95% confidence interval.

```
x=c(2.4, 3.0, 3.0, 2.2, 2.2,
2.2, 2.2, 2.8, 2.0, 3.0)
mx <- mean(x)
vx <- var(x)
crit <- qt(0.975,9)
lower <- mx-crit/sqrt(10)*sqrt(vx)
upper <- mx+crit/sqrt(10)*sqrt(vx)
```

- Estimator: $\widehat{\mu} = 2.5$
- Standard deviation: $\widehat{\sigma} = 0.40$
- Quantile: $t_9(0.975) = 2.26$

Example

A researcher measures the reaction time of $n = 10$ mice to signal pain when a stitch is applied to their tail. Compute a 95% confidence interval.

```
x=c(2.4, 3.0, 3.0, 2.2, 2.2,
2.2, 2.2, 2.8, 2.0, 3.0)
mx <- mean(x)
vx <- var(x)
crit <- qt(0.975,9)
lower <- mx-crit/sqrt(10)*sqrt(vx)
upper <- mx+crit/sqrt(10)*sqrt(vx)
```

- Estimator: $\widehat{\mu} = 2.5$
- Standard deviation: $\widehat{\sigma} = 0.40$
- Quantile: $t_9(0.975) = 2.26$
- CI: $[2.21; 2.79]$

Example

A researcher measures the reaction time of $n = 10$ mice to signal pain when a stitch is applied to their tail. Compute a 95% confidence interval.

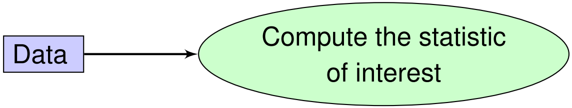
```
x=c(2.4, 3.0, 3.0, 2.2, 2.2,
2.2, 2.2, 2.8, 2.0, 3.0)
mx <- mean(x)
vx <- var(x)
crit <- qt(0.975,9)
lower <- mx-crit/sqrt(10)*sqrt(vx)
upper <- mx+crit/sqrt(10)*sqrt(vx)
```

- Estimator: $\widehat{\mu} = 2.5$
- Standard deviation: $\widehat{\sigma} = 0.40$
- Quantile: $t_9(0.975) = 2.26$
- CI: $[2.21; 2.79]$
- Quality of the confidence interval?

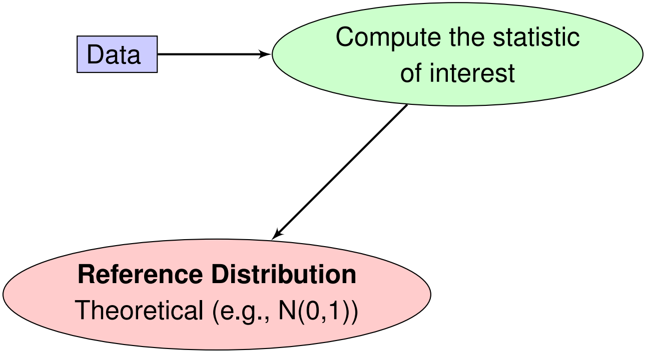
The Resampling Work Flow

Data

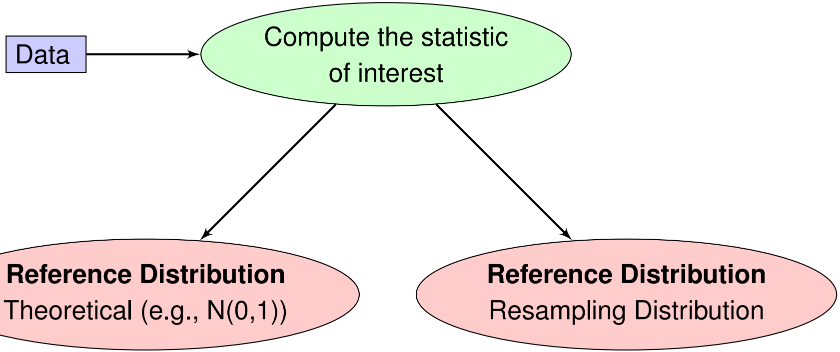
The Resampling Work Flow



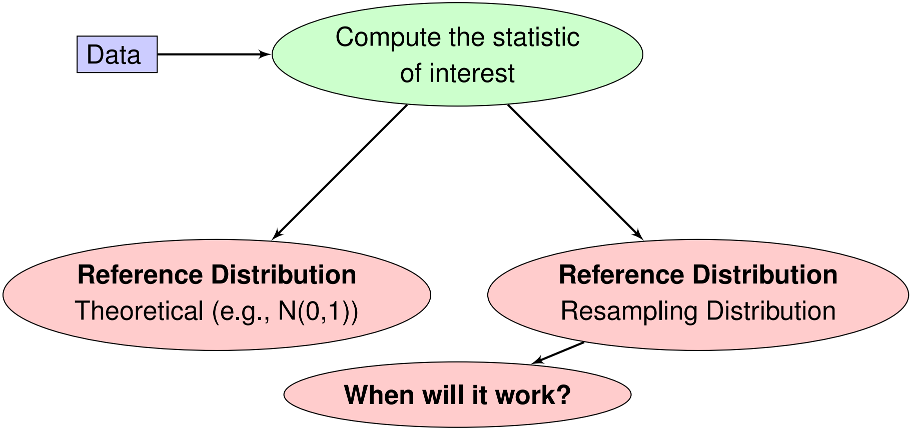
The Resampling Work Flow



The Resampling Work Flow



The Resampling Work Flow



Resampling Distribution

The resampling test will control the type-I error α if and only if the resampling distribution of the statistic mimics the distribution of the test, at least asymptotically.

- Generation of resampling variables

Resampling Distribution

The resampling test will control the type-I error α if and only if the resampling distribution of the statistic mimics the distribution of the test, at least asymptotically.

- Generation of resampling variables
  - We do NOT resample the distribution of the sample

Resampling Distribution

The resampling test will control the type-I error α if and only if the resampling distribution of the statistic mimics the distribution of the test, at least asymptotically.

- Generation of resampling variables
  - We do NOT resample the distribution of the sample
  - We resample the distribution of the statistic

Resampling Distribution

The resampling test will control the type-I error α if and only if the resampling distribution of the statistic mimics the distribution of the test, at least asymptotically.

- Generation of resampling variables
  - We do NOT resample the distribution of the sample
  - We resample the distribution of the statistic
  - Endless different ways to generate resampling variables

Resampling Distribution

The resampling test will control the type-I error α if and only if the resampling distribution of the statistic mimics the distribution of the test, at least asymptotically.

- Generation of resampling variables
  - We do NOT resample the distribution of the sample
  - We resample the distribution of the statistic
  - Endless different ways to generate resampling variables
  - Drawing with replacement is just one of them

Resampling Distribution

The resampling test will control the type-I error α if and only if the resampling distribution of the statistic mimics the distribution of the test, at least asymptotically.

- Generation of resampling variables
  - We do NOT resample the distribution of the sample
  - We resample the distribution of the statistic
  - Endless different ways to generate resampling variables
  - Drawing with replacement is just one of them
  - Today, we will study few more

Resampling Distribution

- Randomly sample/generate from data $\mathbf{X} = (X_1, \ldots, X_n)$

$$X_1^*, \ldots, X_n^*$$

such that

$$T^* = \sqrt{n}\frac{\overline{X}_.^* - E(\overline{X}_.^* | \mathbf{X})}{\widehat{\sigma}^*}$$

has a $N(0, 1)$ distribution

Resampling Distribution

- Randomly sample/generate from data $\mathbf{X} = (X_1, \dots, X_n)$

$$X_1^*, \dots, X_n^*$$

  such that

$$T^* = \sqrt{n}\frac{\overline{X}_{\cdot}^* - E(\overline{X}_{\cdot}^* | \mathbf{X})}{\widehat{\sigma}^*}$$

  has a $N(0, 1)$ distribution
- Compute p-values, critical values etc from the **Resampling Distribution**

Nonparametric Bootstrap

- **Data $\mathbf{X} = (X_1, \ldots, X_n)$ (fixed values)**

Nonparametric Bootstrap

- **Data $\mathbf{X} = (X_1, \ldots, X_n)$ (fixed values)**
- **Drawing with Replacement:** randomly draw n observations X_k^* from **X** with replacement such that

$$P(X_1^* = X_1) = \frac{1}{n}$$

Nonparametric Bootstrap

- **Data $\mathbf{X} = (X_1, \ldots, X_n)$ (fixed values)**
- **Drawing with Replacement:** randomly draw n observations X_k^* from $\mathbf{X}$ with replacement such that

$$P(X_1^* = X_1) = \frac{1}{n}$$

- Example $\mathbf{X} = (1, 2, 3, 4, 5) \Rightarrow$

Nonparametric Bootstrap

- **Data $\mathbf{X} = (X_1, \ldots, X_n)$ (fixed values)**
- **Drawing with Replacement:** randomly draw n observations X_k^* from $\mathbf{X}$ with replacement such that

$$P(X_1^* = X_1) = \frac{1}{n}$$

- Example $\mathbf{X} = (1, 2, 3, 4, 5) \Rightarrow$
  $\mathbf{X}^* = (2, 2, 4, 3, 2)$

Nonparametric Bootstrap

- **Data $\mathbf{X} = (X_1, \dots, X_n)$ (fixed values)**
- **Drawing with Replacement:** randomly draw n observations X_k^* from **X** with replacement such that
$$P(X_1^* = X_1) = \frac{1}{n}$$
- Example $\mathbf{X} = (1, 2, 3, 4, 5) \Rightarrow$
$\mathbf{X}^* = (2, 2, 4, 3, 2)$
$\mathbf{X}^* = (1, 1, 2, 3, 3)$

Nonparametric Bootstrap

- **Data $\mathbf{X} = (X_1, \dots, X_n)$ (fixed values)**
- **Drawing with Replacement:** randomly draw n observations X_k^* from **X** with replacement such that

$$P(X_1^* = X_1) = \frac{1}{n}$$

- Example $\mathbf{X} = (1, 2, 3, 4, 5) \Rightarrow$
  $\mathbf{X}^* = (2, 2, 4, 3, 2)$
  $\mathbf{X}^* = (1, 1, 2, 3, 3)$
  $\mathbf{X}^* = (2, 5, 5, 3, 3)$

Nonparametric Bootstrap

- **Data $\mathbf{X} = (X_1, \ldots, X_n)$ (fixed values)**
- **Drawing with Replacement:** randomly draw n observations X_k^* from **X** with replacement such that

$$P(X_1^* = X_1) = \frac{1}{n}$$

- Example $\mathbf{X} = (1, 2, 3, 4, 5) \Rightarrow$
  $\mathbf{X}^* = (2, 2, 4, 3, 2)$
  $\mathbf{X}^* = (1, 1, 2, 3, 3)$
  $\mathbf{X}^* = (2, 5, 5, 3, 3)$
  ...

Nonparametric Bootstrap

- **Data X** $= (X_1, \ldots, X_n)$ (**fixed values**)
- **Drawing with Replacement:** randomly draw n observations X_k^* from **X** with replacement such that

$$P(X_1^* = X_1) = \frac{1}{n}$$

- Example $\mathbf{X} = (1, 2, 3, 4, 5) \Rightarrow$
  $\mathbf{X}^* = (2, 2, 4, 3, 2)$
  $\mathbf{X}^* = (1, 1, 2, 3, 3)$
  $\mathbf{X}^* = (2, 5, 5, 3, 3)$
  ...
- In R: sample(x,replace=TRUE)

Permutation

- **Data** $\mathbf{X} = (X_1, \ldots, X_n)$ **(fixed values)**

Permutation

- **Data $\mathbf{X} = (X_1, \ldots, X_n)$ (fixed values)**
- **Drawing without Replacement:** randomly draw n observations X_k^* from $\mathbf{X}$ without replacement such that

$$P(X_1^* = X_1) = \frac{1}{n}$$

Permutation

- **Data $\mathbf{X} = (X_1, \ldots, X_n)$ (fixed values)**
- **Drawing without Replacement:** randomly draw n observations X_k^* from $\mathbf{X}$ without replacement such that

$$P(X_1^* = X_1) = \frac{1}{n}$$

- Example $\mathbf{X} = (1, 2, 3, 4, 5) \Rightarrow$

Permutation

- **Data $\mathbf{X} = (X_1, \ldots, X_n)$ (fixed values)**
- **Drawing without Replacement:** randomly draw n observations X_k^* from $\mathbf{X}$ without replacement such that

$$P(X_1^* = X_1) = \frac{1}{n}$$

- Example $\mathbf{X} = (1, 2, 3, 4, 5) \Rightarrow$
  $\mathbf{X}^* = (4, 1, 3, 2, 5)$

Permutation

- **Data $\mathbf{X} = (X_1, \ldots, X_n)$ (fixed values)**
- **Drawing without Replacement:** randomly draw n observations X_k^* from $\mathbf{X}$ without replacement such that

$$P(X_1^* = X_1) = \frac{1}{n}$$

- Example $\mathbf{X} = (1, 2, 3, 4, 5) \Rightarrow$
  $\mathbf{X}^* = (4, 1, 3, 2, 5)$
  $\mathbf{X}^* = (5, 1, 2, 3, 4)$

Permutation

- **Data $\mathbf{X} = (X_1, \ldots, X_n)$ (fixed values)**
- **Drawing without Replacement:** randomly draw n observations X_k^* from $\mathbf{X}$ without replacement such that
$$P(X_1^* = X_1) = \frac{1}{n}$$
- Example $\mathbf{X} = (1, 2, 3, 4, 5) \Rightarrow$
  $\mathbf{X}^* = (4, 1, 3, 2, 5)$
  $\mathbf{X}^* = (5, 1, 2, 3, 4)$
  $\mathbf{X}^* = (3, 1, 2, 5, 4)$

Permutation

- **Data X** $= (X_1, \ldots, X_n)$ **(fixed values)**
- **Drawing without Replacement:** randomly draw n observations X_k^* from **X** without replacement such that
$$P(X_1^* = X_1) = \frac{1}{n}$$
- Example $\mathbf{X} = (1, 2, 3, 4, 5) \Rightarrow$
  $\mathbf{X}^* = (4, 1, 3, 2, 5)$
  $\mathbf{X}^* = (5, 1, 2, 3, 4)$
  $\mathbf{X}^* = (3, 1, 2, 5, 4)$
  ...

Permutation

- **Data $\mathbf{X} = (X_1, \ldots, X_n)$ (fixed values)**
- **Drawing without Replacement:** randomly draw n observations X_k^* from $\mathbf{X}$ without replacement such that
$$P(X_1^* = X_1) = \frac{1}{n}$$
- Example $\mathbf{X} = (1, 2, 3, 4, 5) \Rightarrow$
$\mathbf{X}^* = (4, 1, 3, 2, 5)$
$\mathbf{X}^* = (5, 1, 2, 3, 4)$
$\mathbf{X}^* = (3, 1, 2, 5, 4)$
...
- In R: sample(x)

Permutation

- **Data $\mathbf{X} = (X_1, \ldots, X_n)$ (fixed values)**
- **Drawing without Replacement:** randomly draw n observations X_k^* from $\mathbf{X}$ without replacement such that
$$P(X_1^* = X_1) = \frac{1}{n}$$
- Example $\mathbf{X} = (1, 2, 3, 4, 5) \Rightarrow$
  $\mathbf{X}^* = (4, 1, 3, 2, 5)$
  $\mathbf{X}^* = (5, 1, 2, 3, 4)$
  $\mathbf{X}^* = (3, 1, 2, 5, 4)$
  ...
- In R: sample(x)
- Only for more than one group applicable. Why?

Parametric Bootstrap

- **Data $\mathbf{X} = (X_1, \ldots, X_n)$ (fixed values)**

Parametric Bootstrap

- **Data X** $= (X_1, \ldots, X_n)$ (**fixed values**)
- **Resampling** randomly draw n observations X_k^* from

$$N(0, \widehat{\sigma}^2)$$

Parametric Bootstrap

- **Data** $\mathbf{X} = (X_1, \ldots, X_n)$ (**fixed values**)
- **Resampling** randomly draw n observations X_k^* from

$$N(0, \widehat{\sigma}^2)$$

- In R: $rnorm(n, 0, sd(x))$

Parametric Bootstrap

- **Data $\mathbf{X} = (X_1, \ldots, X_n)$ (fixed values)**
- **Resampling** randomly draw n observations X_k^* from

$$N(0, \widehat{\sigma}^2)$$

- In R: $rnorm(n, 0, sd(x))$
- For more than one group. Why?

Parametric Bootstrap-II

- **Data $\mathbf{X} = (X_1, \ldots, X_n)$ (fixed values)**

Parametric Bootstrap-II

- **Data X** $= (X_1, \ldots, X_n)$ (**fixed values**)
- Estimate the skewness of the sample by

$$\widehat{\mu}_3 = \frac{n}{(n-1)(n-2)} \sum_{k=1}^{n} \left(\frac{X_k - \overline{X}.}{\widehat{\sigma}} \right)^3$$

Parametric Bootstrap-II

- **Data X** $= (X_1, \ldots, X_n)$ (**fixed values**)
- Estimate the skewness of the sample by

$$\widehat{\mu}_3 = \frac{n}{(n-1)(n-2)} \sum_{k=1}^{n} \left(\frac{X_k - \overline{X}_.}{\widehat{\sigma}} \right)^3$$

- **Resampling** randomly draw n observations X_k^* from

$$sign(\widehat{\mu}_3)\widehat{\sigma}\frac{\chi_f^2 - f}{\sqrt{2f}}$$

Parametric Bootstrap-II

- **Data X** $= (X_1, \ldots, X_n)$ **(fixed values)**
- Estimate the skewness of the sample by

$$\widehat{\mu}_3 = \frac{n}{(n-1)(n-2)} \sum_{k=1}^{n} \left(\frac{X_k - \overline{X}.}{\widehat{\sigma}} \right)^3$$

- **Resampling** randomly draw n observations X_k^* from

$$sign(\widehat{\mu}_3)\widehat{\sigma}\frac{\chi_f^2 - f}{\sqrt{2f}}$$

- $f = 8/\widehat{\mu}_3^2$

Parametric Bootstrap-II

- **Data** $\mathbf{X} = (X_1, \ldots, X_n)$ (**fixed values**)
- Estimate the skewness of the sample by

$$\widehat{\mu}_3 = \frac{n}{(n-1)(n-2)} \sum_{k=1}^{n} \left(\frac{X_k - \overline{X}_.}{\widehat{\sigma}} \right)^3$$

- **Resampling** randomly draw n observations X_k^* from

$$sign(\widehat{\mu}_3)\widehat{\sigma}\frac{\chi_f^2 - f}{\sqrt{2f}}$$

- $f = 8/\widehat{\mu}_3^2$
- Note that $E(X_k^*) = 0$, $Var(X_k^*) = \widehat{\sigma}^2$ and $\mu_3(X_k^*) = \widehat{\mu}_3$. (Why?)

Wild Bootstrap

- **Data $\mathbf{X} = (X_1, \ldots, X_n)$ (fixed values)**

Wild Bootstrap

- **Data $\mathbf{X} = (X_1, \ldots, X_n)$ (fixed values)**
- Fix the values $Z_k = X_k - \overline{X}$.

Wild Bootstrap

- **Data X** $= (X_1, \ldots, X_n)$ (**fixed values**)
- Fix the values $Z_k = X_k - \overline{X}$.
- **Resampling** randomly generate iid weights W_k with $E(W_k) = 0$ and $Var(W_k) = 1$ and generate X_k^* by

$$X_k^* = W_k * Z_k$$

Wild Bootstrap

- **Data X** $= (X_1, \dots, X_n)$ (**fixed values**)
- Fix the values $Z_k = X_k - \overline{X}$.
- **Resampling** randomly generate iid weights W_k with $E(W_k) = 0$ and $Var(W_k) = 1$ and generate X_k^* by

$$X_k^* = W_k * Z_k$$

- Examples: $W_k \sim N(0, 1)$

Wild Bootstrap

- **Data X** $= (X_1, \ldots, X_n)$ (**fixed values**)
- Fix the values $Z_k = X_k - \overline{X}$.
- **Resampling** randomly generate iid weights W_k with $E(W_k) = 0$ and $Var(W_k) = 1$ and generate X_k^* by

$$X_k^* = W_k * Z_k$$

- Examples: $W_k \sim N(0, 1)$
- Rademacher: $P(W_k = 1) = P(W_k = -1) = 1/2$

Wild Bootstrap

- **Data X** $= (X_1, \ldots, X_n)$ (**fixed values**)
- Fix the values $Z_k = X_k - \overline{X}$.
- **Resampling** randomly generate iid weights W_k with $E(W_k) = 0$ and $Var(W_k) = 1$ and generate X_k^* by

$$X_k^* = W_k * Z_k$$

- Examples: $W_k \sim N(0, 1)$
- Rademacher: $P(W_k = 1) = P(W_k = -1) = 1/2$

  ...

Generation of the Resampling Distribution

- Using any of the above methods, we do the following:

Generation of the Resampling Distribution

- Using any of the above methods, we do the following:
    1. Generate the resampling variables $X_1^*, \ldots, X_n^*$

Generation of the Resampling Distribution

- Using any of the above methods, we do the following:
    1. Generate the resampling variables $X_1^*, \ldots, X_n^*$
    2. Compute $\overline{X}_{\cdot}^* = \frac{1}{n}\sum_{k=1}^{n} X_k^*$ and $\widehat{\sigma}^{2,*} = \frac{1}{n-1}\sum_{k=1}^{n}(X_k^* - \overline{X}_{\cdot}^*)^2$

Generation of the Resampling Distribution

- Using any of the above methods, we do the following:
  1. Generate the resampling variables $X_1^*, \ldots, X_n^*$
  2. Compute $\overline{X}_\cdot^* = \frac{1}{n}\sum_{k=1}^n X_k^*$ and $\widehat{\sigma}^{2,*} = \frac{1}{n-1}\sum_{k=1}^n (X_k^* - \overline{X}_\cdot^*)^2$
  3. Compute the statistic

$$T^* = \sqrt{n}\frac{\overline{X}_\cdot^* - E(\overline{X}_\cdot^* | \mathbf{X})}{\widehat{\sigma}^*}$$

Generation of the Resampling Distribution

- Using any of the above methods, we do the following:
  1. Generate the resampling variables $X_1^*, \ldots, X_n^*$
  2. Compute $\overline{X}_\cdot^* = \frac{1}{n}\sum_{k=1}^n X_k^*$ and $\widehat{\sigma}^{2,*} = \frac{1}{n-1}\sum_{k=1}^n (X_k^* - \overline{X}_\cdot^*)^2$
  3. Compute the statistic
  $$T^* = \sqrt{n}\frac{\overline{X}_\cdot^* - E(\overline{X}_\cdot^* | \mathbf{X})}{\widehat{\sigma}^*}$$
  4. Safe the value of T^* in T_ℓ

Generation of the Resampling Distribution

- Using any of the above methods, we do the following:
  1. Generate the resampling variables $X_1^*, \dots, X_n^*$
  2. Compute $\overline{X}_\cdot^* = \frac{1}{n}\sum_{k=1}^n X_k^*$ and $\widehat{\sigma}^{2,*} = \frac{1}{n-1}\sum_{k=1}^n (X_k^* - \overline{X}_\cdot^*)^2$
  3. Compute the statistic
  $$T^* = \sqrt{n}\frac{\overline{X}_\cdot^* - E(\overline{X}_\cdot^* | \mathbf{X})}{\widehat{\sigma}^*}$$
  4. Safe the value of T^* in T_ℓ
  5. Repeat the steps above a large number of times, e.g. $n_{boot} = 10,000$

Generation of the Resampling Distribution

- Using any of the above methods, we do the following:
  1. Generate the resampling variables $X_1^*, \ldots, X_n^*$
  2. Compute $\overline{X}_\cdot^* = \frac{1}{n}\sum_{k=1}^n X_k^*$ and $\widehat{\sigma}^{2,*} = \frac{1}{n-1}\sum_{k=1}^n (X_k^* - \overline{X}_\cdot^*)^2$
  3. Compute the statistic

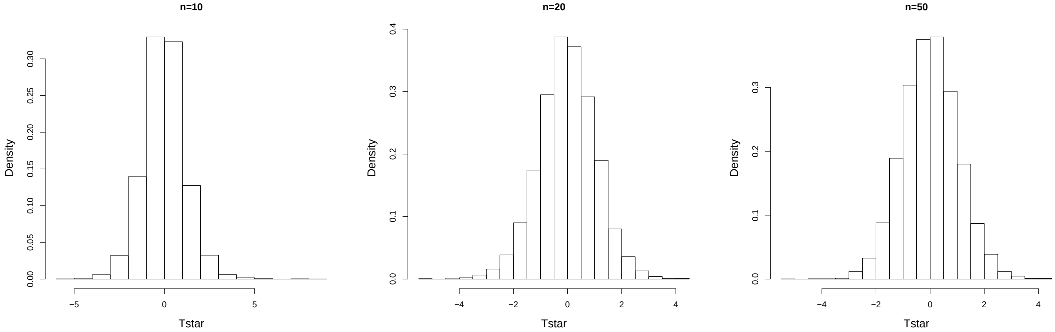
$$T^* = \sqrt{n}\frac{\overline{X}_\cdot^* - E(\overline{X}_\cdot^* | \mathbf{X})}{\widehat{\sigma}^*}$$

  4. Safe the value of T^* in T_ℓ
  5. Repeat the steps above a large number of times, e.g. $n_{boot} = 10,000$
  6. Compute the critical value $c^*(1-\alpha/2)$ or the p-value as

$$p = 2\min\left\{\frac{1}{n_{boot}}\sum_{\ell=1}^{n_{boot}} \mathcal{I}(T_\ell \leq T), \frac{1}{n_{boot}}\sum_{\ell=1}^{n_{boot}} \mathcal{I}(T_\ell \geq T)\right\}$$

$\mathcal{I}(x)$ is an indicator

Illustration Resampling Distribution



Compute the $(1-\alpha)$-quantile

What is $E(\overline{X}^*_{.}|\mathbf{X})$?

- The term $E(\overline{X}^*_{.}|\mathbf{X})$ needed to ensure that $E(\overline{X}^*_{.} - E(\overline{X}^*_{.}|\mathbf{X})) = 0$

What is $E(\overline{X}^*_{.}|\mathbf{X})$?

- The term $E(\overline{X}^*_{.}|\mathbf{X})$ needed to ensure that $E(\overline{X}^*_{.} - E(\overline{X}^*_{.}|\mathbf{X})) = 0$
- Given the data $\mathbf{X}$, you will observe the value $E(\overline{X}^*_{.}|\mathbf{X})$, on average.

What is $E(\overline{X}^*_{\cdot}|\mathbf{X})$?

- The term $E(\overline{X}^*_{\cdot}|\mathbf{X})$ needed to ensure that $E(\overline{X}^*_{\cdot} - E(\overline{X}^*_{\cdot}|\mathbf{X})) = 0$
- Given the data $\mathbf{X}$, you will observe the value $E(\overline{X}^*_{\cdot}|\mathbf{X})$, on average.
- Consider drawing with replacement, then

$$E(X_1^*|\mathbf{X}) = \sum_{k=1}^{n} X_k \cdot P(X_k^* = X_k) = \sum_{k=1}^{n} X_k \cdot \frac{1}{n} = \frac{1}{n}\sum_{k=1}^{n} X_k = \overline{X}.$$

What is $E(\overline{X}^*_\cdot|\mathbf{X})$?

- The term $E(\overline{X}^*_\cdot|\mathbf{X})$ needed to ensure that $E(\overline{X}^*_\cdot - E(\overline{X}^*_\cdot|\mathbf{X})) = 0$
- Given the data $\mathbf{X}$, you will observe the value $E(\overline{X}^*_\cdot|\mathbf{X})$, on average.
- Consider drawing with replacement, then

$$E(X_1^*|\mathbf{X}) = \sum_{k=1}^{n} X_k \cdot P(X_k^* = X_k) = \sum_{k=1}^{n} X_k \cdot \frac{1}{n} = \frac{1}{n}\sum_{k=1}^{n} X_k = \overline{X}_\cdot$$

- On average, X_1^* will be $\overline{X}_\cdot$

What is $E(\overline{X}^*_{\cdot}|\mathbf{X})$?

- The term $E(\overline{X}^*_{\cdot}|\mathbf{X})$ needed to ensure that $E(\overline{X}^*_{\cdot} - E(\overline{X}^*_{\cdot}|\mathbf{X})) = 0$
- Given the data $\mathbf{X}$, you will observe the value $E(\overline{X}^*_{\cdot}|\mathbf{X})$, on average.
- Consider drawing with replacement, then

$$E(X_1^*|\mathbf{X}) = \sum_{k=1}^{n} X_k \cdot P(X_k^* = X_k) = \sum_{k=1}^{n} X_k \cdot \frac{1}{n} = \frac{1}{n}\sum_{k=1}^{n} X_k = \overline{X}.$$

- On average, X_1^* will be $\overline{X}$.
- $E(\overline{X}^*_{\cdot}) =$

What is $E(\overline{X}^*_{.}|\mathbf{X})$?-II

- Compute the following terms

What is $E(\overline{X}^*_{\cdot}|\mathbf{X})$?-II

- Compute the following terms
  - $E(X^*_1|\mathbf{X})$

What is $E(\overline{X}_{.}^{*}|\mathbf{X})$?-II

- Compute the following terms
  - $E(X_1^*|\mathbf{X})$
  - $E(X_1^{2,*}|\mathbf{X}) =$

What is $E(\overline{X}^*_{\cdot}|\mathbf{X})$?-II

- Compute the following terms
  - $E(X_1^*|\mathbf{X})$
  - $E(X_1^{2,*}|\mathbf{X}) =$
  - $E(\overline{X}^*_{\cdot}|\mathbf{X}) =$

What is $E(\overline{X}_{\cdot}^{*}|\mathbf{X})$?-II

- Compute the following terms
  - $E(X_1^*|\mathbf{X})$
  - $E(X_1^{2,*}|\mathbf{X}) =$
  - $E(\overline{X}_{\cdot}^{*}|\mathbf{X}) =$
  - $E(\widehat{\sigma}^{2,*}|\mathbf{X}) =$

What is $E(\overline{X}_{\cdot}^{*}|\mathbf{X})$?-II

- Compute the following terms
  - $E(X_1^{*}|\mathbf{X})$
  - $E(X_1^{2,*}|\mathbf{X}) =$
  - $E(\overline{X}_{\cdot}^{*}|\mathbf{X}) =$
  - $E(\widehat{\sigma}^{2,*}|\mathbf{X}) =$
- For all of the different resampling variables

Which method to use?

- Assess simulation studies w.r.t. the type-1 error rate control

Which method to use?

- Assess simulation studies w.r.t. the type-1 error rate control
- Numerical implementation is important

Which method to use?

- Assess simulation studies w.r.t. the type-1 error rate control
- Numerical implementation is important
- Use matrix techniques to generate the resampling distribution

Which method to use?

- Assess simulation studies w.r.t. the type-1 error rate control
- Numerical implementation is important
- Use matrix techniques to generate the resampling distribution
    - R functions *colMeans(x)*, *colSums(x)*

Which method to use?

- Assess simulation studies w.r.t. the type-1 error rate control
- Numerical implementation is important
- Use matrix techniques to generate the resampling distribution
  - R functions *colMeans(x)*, *colSums(x)*
  - Variance formula $\widehat{\sigma}^2 = (\sum_{k=1}^{n} X_k^2 - n\overline{X}_{\cdot}^2)/(n-1)$

Which method to use?

- Assess simulation studies w.r.t. the type-1 error rate control
- Numerical implementation is important
- Use matrix techniques to generate the resampling distribution
  - R functions *colMeans(x)*, *colSums(x)*
  - Variance formula $\widehat{\sigma}^2 = (\sum_{k=1}^n X_k^2 - n\overline{X}_\cdot^2)/(n-1)$
  - Generate a matrix of resampling variables (example next slide)

```
skew<-function(x){
n<-length(x)
mx<-mean(x)
sdx<-sd(x)
n/((n-1)*(n-2))*sum(((x-mx)/sdx)^3)}
```

```
myboot <- function(n,Distribution,nboot,nsim){
T=Tboot = Tbootsp =c()
#------------Test Statistic of the Sample------------#
if(Distribution == "Normal"){
x <- matrix(rnorm(n*nsim),ncol=nsim)}
if(Distribution == "Exp"){
x <- matrix(rexp(n*nsim)-1,ncol=nsim)}
mx <- colMeans(x)
vx <- (colSums(x^2)-n*mx^2)/(n-1)
T=sqrt(n)*(mx)/sqrt(vx)
```

```
#------------Simulate the Nonparametric Bootstrap-----#
B <- apply(matrix(1:n,ncol=nboot,nrow=n),2,sample,replace=TRUE)
for(i in 1:nsim){
xstar = matrix(x[,i][B],ncol=nboot,nrow=n)
mxstar = colMeans(xstar)
vxb = (colSums(xstar^2)-n*mxstar^2)/(n-1)
Tstar = sqrt(n)*(mxstar - mx[i])/sqrt(vxb)
p1 = mean(Tstar >= T[i])
p2 = mean(Tstar <= T[i])
Tboot[i] = (2*min(p1,p2)<0.05)
```

```
#-------------Simulate the skew parametric Bootstrap---#
mu3=skew(x[,i])
f=8/mu3^2
xstarsp=matrix(sign(mu3)*sqrt(vx)*((rchisq(n*nboot,f)-f)/sqrt(2*f)),ncol=nboot)
mxbootsp=colMeans(xstarsp)
vxbsp=(colSums(xstarsp^2) - n*mxbootsp^2)/(n-1)
Tstarnp = sqrt(n)*mxbootsp/sqrt(vxbsp)
p1sp = mean(Tstarnp >= T[i])
p2sp = mean(Tstarnp <= T[i])
Tbootsp[i] = (2*min(p1sp,p2sp)<0.05)}
```

```
result = data.frame(n=n,nsim=nsim,nboot=nboot,Dist=Distribution,
T=mean(abs(T)>qt(0.975,n-1)),
Tboot = mean(Tboot), Tbootsp = mean(Tbootsp))
result}
myboot(10,"Exp",1000,1000)
```

Parameter Estimation

- Parameter estimation is key feature in statistical sciences

Parameter Estimation

- Parameter estimation is key feature in statistical sciences
- Often, pretty involved

Parameter Estimation

- Parameter estimation is key feature in statistical sciences
- Often, pretty involved
- Think about the following cases

Parameter Estimation

- Parameter estimation is key feature in statistical sciences
- Often, pretty involved
- Think about the following cases
    - Mean

Parameter Estimation

- Parameter estimation is key feature in statistical sciences
- Often, pretty involved
- Think about the following cases
  - Mean
  - Variance

Parameter Estimation

- Parameter estimation is key feature in statistical sciences
- Often, pretty involved
- Think about the following cases
    - Mean
    - Variance
    - Variance of an estimator (e.g. mean)

Parameter Estimation

- Parameter estimation is key feature in statistical sciences
- Often, pretty involved
- Think about the following cases
  - Mean
  - Variance
  - Variance of an estimator (e.g. mean)
  - Variance of correlation coefficients

Parameter Estimation

- Parameter estimation is key feature in statistical sciences
- Often, pretty involved
- Think about the following cases
  - Mean
  - Variance
  - Variance of an estimator (e.g. mean)
  - Variance of correlation coefficients
  - Overdispersion parameters, variance of their estimators,...

Parameter Estimation - II

- Statistical model

$$X_1, \ldots, X_n \sim F(\theta)$$

Parameter Estimation - II

- Statistical model

$$X_1, \ldots, X_n \sim F(\theta)$$

- F is a distribution

Parameter Estimation - II

- Statistical model

$$X_1, \ldots, X_n \sim F(\theta)$$

- F is a distribution
- θ are parameters of this distribution

Parameter Estimation - II

- Statistical model

$$X_1, \ldots, X_n \sim F(\theta)$$

- F is a distribution
- θ are parameters of this distribution
- How to estimate $f(\theta)$?

Parameter Estimation - III

Estimation of θ	Properties
Maximum-Likelihood	F must be known
	Algorithm can be difficult
	Algorithm might not converge
	Large sample for distribution
Moment based	F can be unknown
	Computation usually feasible
	Usually exist (no converging issues)
	Small sample approximations
Resampling Methods	

Parameter Estimation - Resampling. But How?

- Data $\mathbf{X} = (X_1, \ldots, X_n)'$

Parameter Estimation - Resampling. But How?

- Data $\mathbf{X} = (X_1, \ldots, X_n)'$
- **Draw observations with replacement** from $\mathbf{X}$:

$$X_1^*, \ldots, X_n^*$$

Parameter Estimation - Resampling. But How?

- Data $\mathbf{X} = (X_1, \ldots, X_n)'$
- **Draw observations with replacement** from $\mathbf{X}$:

$$X_1^*, \ldots, X_n^*$$

  - Distribution of $X_1^*, \ldots, X_n^*$: $\widehat{F}_n$ (empirical distribution)

Parameter Estimation - Resampling. But How?

- Data $\mathbf{X} = (X_1, \dots, X_n)'$
- **Draw observations with replacement** from $\mathbf{X}$:

$$X_1^*, \dots, X_n^*$$

  - Distribution of $X_1^*, \dots, X_n^*$: $\widehat{F}_n$ (empirical distribution)
  - We draw observations from $\widehat{F}_n$

Parameter Estimation - Resampling. But How?

- Data $\mathbf{X} = (X_1, \ldots, X_n)'$
- **Draw observations with replacement** from $\mathbf{X}$:

$$X_1^*, \ldots, X_n^*$$

  - Distribution of $X_1^*, \ldots, X_n^*$: $\widehat{F}_n$ (empirical distribution)
  - We draw observations from $\widehat{F}_n$
  - We know

$$\widehat{F}_n \to F, n \to \infty$$

Parameter Estimation - Resampling. But How?

- Data $\mathbf{X} = (X_1, \ldots, X_n)'$
- **Draw observations with replacement** from $\mathbf{X}$:

$$X_1^*, \ldots, X_n^*$$

  - Distribution of $X_1^*, \ldots, X_n^*$: $\widehat{F}_n$ (empirical distribution)
  - We draw observations from $\widehat{F}_n$
  - We know

$$\widehat{F}_n \to F, n \to \infty$$

  - Basically, we simulate data from $\widehat{F}_n$

Parameter Estimation - Resampling -II

- Fix the data $\mathbf{X}$

Parameter Estimation - Resampling -II

- Fix the data **X**
  - Generate a bootstrap sample (drawing with replacement from **X**): $X_1^*, \ldots, X_n^*$

Parameter Estimation - Resampling -II

- Fix the data **X**
  - Generate a bootstrap sample (drawing with replacement from **X**): $X_1^*, \ldots, X_n^*$
  - Compute the estimator $\widehat{\theta}^*$ and safe this value

Parameter Estimation - Resampling -II

- Fix the data **X**
  - Generate a bootstrap sample (drawing with replacement from **X**): $X_1^*, \ldots, X_n^*$
  - Compute the estimator $\widehat{\theta}^*$ and safe this value
  - Repeat the previous steps n_{boot} times

Parameter Estimation - Resampling -II

- Fix the data **X**
  - Generate a bootstrap sample (drawing with replacement from **X**): $X_1^*, \ldots, X_n^*$
  - Compute the estimator $\widehat{\theta}^*$ and safe this value
  - Repeat the previous steps n_{boot} times
  - Estimate the parameter of interest using the values of $\widehat{\theta}_1, \ldots, \widehat{\theta}_{n_{boot}}$

Parameter Estimation - Resampling -III

- Model: $X_1, \ldots, X_n \sim F(\theta)$ (iid)

Parameter Estimation - Resampling -III

- Model: $X_1, \ldots, X_n \sim F(\theta)$ (iid)
- Parameters $E(X_k) = \mu$ and $Var(X_k) = \sigma^2$

Parameter Estimation - Resampling -III

- Model: $X_1, \ldots, X_n \sim F(\theta)$ (iid)
- Parameters $E(X_k) = \mu$ and $Var(X_k) = \sigma^2$
- Task: Estimate $f(\theta) = \tau^2 = Var(\overline{X}.)$

Parameter Estimation - Resampling -III

- Model: $X_1, \ldots, X_n \sim F(\theta)$ (iid)
- Parameters $E(X_k) = \mu$ and $Var(X_k) = \sigma^2$
- Task: Estimate $f(\theta) = \tau^2 = Var(\overline{X}.)$

- Data **X**

Parameter Estimation - Resampling -III

- Model: $X_1, \ldots, X_n \sim F(\theta)$ (iid)
- Parameters $E(X_k) = \mu$ and $Var(X_k) = \sigma^2$
- Task: Estimate $f(\theta) = \tau^2 = Var(\overline{X}.)$

- Data $\mathbf{X}$
- Generate $X_1^*, \ldots, X_n^*$

Parameter Estimation - Resampling -III

- Model: $X_1, \ldots, X_n \sim F(\theta)$ (iid)
- Parameters $E(X_k) = \mu$ and $Var(X_k) = \sigma^2$
- Task: Estimate $f(\theta) = \tau^2 = Var(\overline{X}_.)$

- Data $\mathbf{X}$
- Generate $X_1^*, \ldots, X_n^*$
- Compute $\overline{X}_.^*$ (safe in $\widehat{\theta}_\ell$)

Parameter Estimation - Resampling -III

- Model: $X_1, \ldots, X_n \sim F(\theta)$ (iid)
- Parameters $E(X_k) = \mu$ and $Var(X_k) = \sigma^2$
- Task: Estimate $f(\theta) = \tau^2 = Var(\overline{X}_.)$

- Data $\mathbf{X}$
- Generate $X_1^*, \ldots, X_n^*$
- Compute $\overline{X}_.^*$ (safe in $\widehat{\theta}_\ell$)
- Repeated the steps $\ell = 1, \ldots, n_{boot}$ times

Parameter Estimation - Resampling -III

- Model: $X_1, \dots, X_n \sim F(\theta)$ (iid)
- Parameters $E(X_k) = \mu$ and $Var(X_k) = \sigma^2$
- Task: Estimate $f(\theta) = \tau^2 = Var(\overline{X}_.)$

- Data $\mathbf{X}$
- Generate $X_1^*, \dots, X_n^*$
- Compute $\overline{X}_.^*$ (safe in $\widehat{\theta}_\ell$)
- Repeated the steps $\ell = 1, \dots, n_{boot}$ times
- Estimator of $\tau^2 = Var(\overline{X}_.)$ is

$$\widehat{\tau}^2 = \frac{1}{n_{boot} - 1} \sum_{\ell=1}^{n_{boot}} (\widehat{\theta}_\ell - \overline{\widehat{\theta}}_.)^2, \quad \overline{\widehat{\theta}}_. = \frac{1}{n_{boot}} \sum_{\ell=1}^{n_{boot}} \widehat{\theta}_\ell$$

Parameter Estimation - Resampling -III

- Model: $X_1, \dots, X_n \sim F(\theta)$ (iid)
- Parameters $E(X_k) = \mu$ and $Var(X_k) = \sigma^2$
- Task: Estimate $f(\theta) = \tau^2 = Var(\overline{X}_.)$

- Data $\mathbf{X}$
- Generate $X_1^*, \dots, X_n^*$
- Compute $\overline{X}_.^*$ (safe in $\widehat{\theta}_\ell$)
- Repeated the steps $\ell = 1, \dots, n_{boot}$ times
- Estimator of $\tau^2 = Var(\overline{X}_.)$ is

$$\widehat{\tau}^2 = \frac{1}{n_{boot} - 1} \sum_{\ell=1}^{n_{boot}} (\widehat{\theta}_\ell - \overline{\widehat{\theta}}_.)^2, \quad \overline{\widehat{\theta}}_. = \frac{1}{n_{boot}} \sum_{\ell=1}^{n_{boot}} \widehat{\theta}_\ell$$

```
X=c(2.4, 3.0, 3.0, 2.2, 2.2,
2.2, 2.2, 2.8, 2.0, 3.0)
n <- 10
nboot <-10000
B<- apply(matrix(1:n,
ncol=nboot,nrow=n),
2,sample,replace=TRUE)
xstar <- matrix(X[B],
ncol=nboot,nrow=n)
mxstar <- colMeans(xstar)
tauhat2 <- var(mxstar)
```

Project: How Good is the Estimator?

- Use computer simulations to assess the quality of the estimator $\widehat{\tau}^2$
- Compare with $\widehat{\tau}^2_{emp} = \widehat{\sigma}^2 / n$
- Compare the bias and MSE of the estimators
  - Generate n_{sim} random samples from different distributions
  - Compute the estimator upon the sample and safe the value in $\widehat{\theta}_s$
  - Assess the bias and MSE

$$Bias = \frac{1}{n_{sim}} \sum_{s=1}^{n_{sim}} (\widehat{\theta}_s - \theta) \text{ and } MSE = \frac{1}{n_{sim}} \sum_{s=1}^{n_{sim}} (\widehat{\theta}_s - \theta)^2$$

Project

- Task: Estimate the variance of **correlation coefficients** using resampling strategies
- How to get an idea about the true variance?
- Is bootstrap a good way?