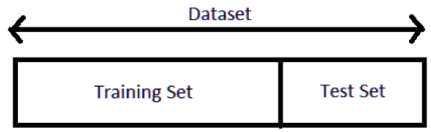
Cross Validation and Bootstrapping

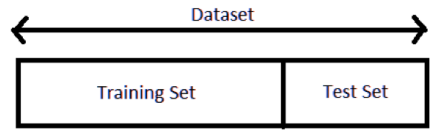
Roberta De Vito

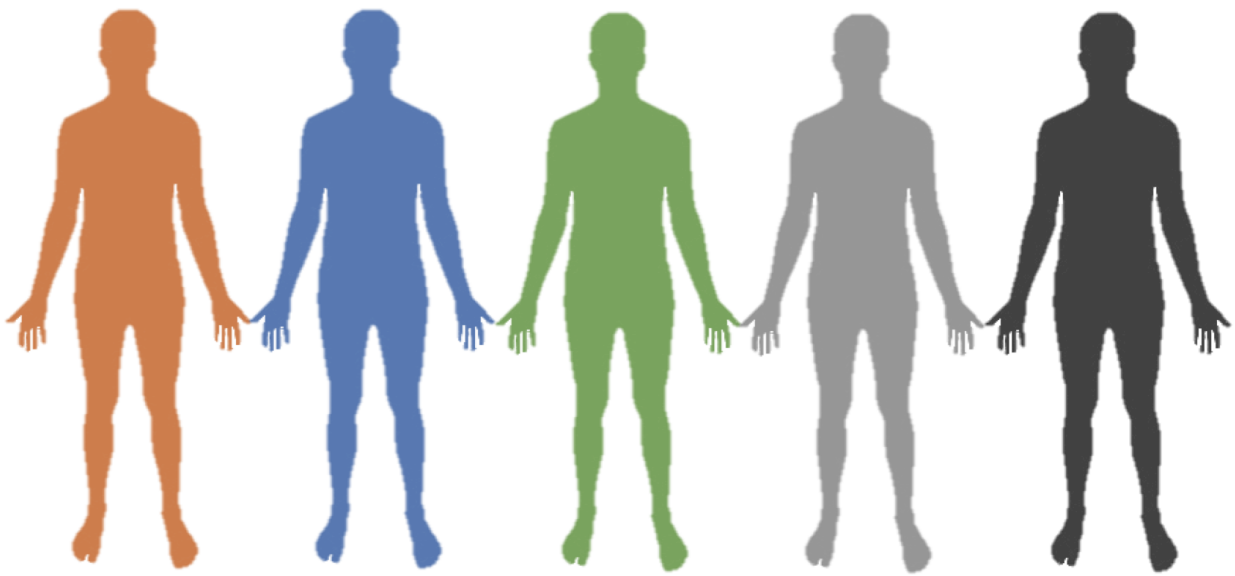
BROWN
Public Health

Division of the data set

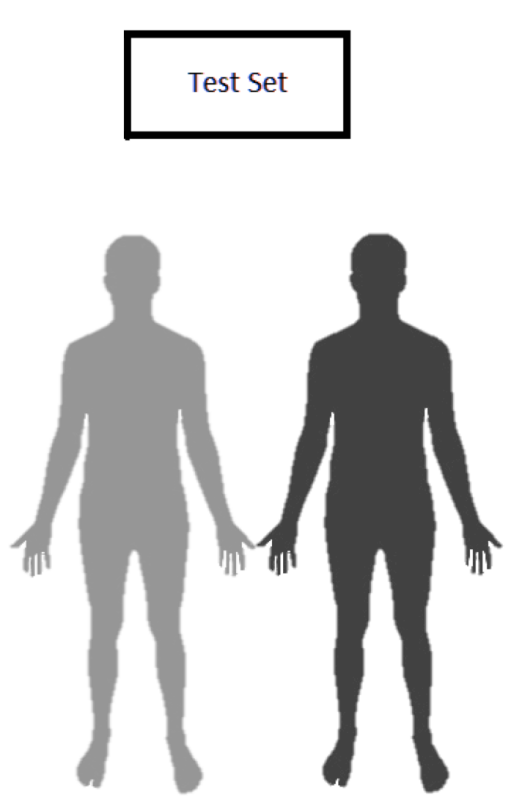


Division of the data set





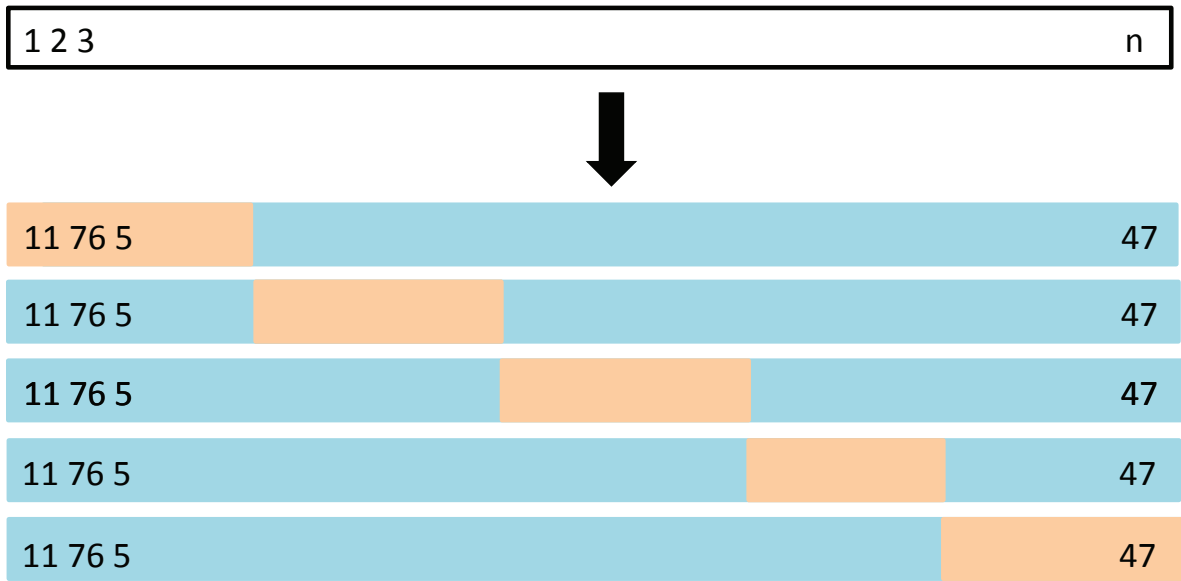
Division of the data set



K-fold Cross-Validation for MSE

- Widely used approach for estimating test error
- Estimates can be used to select best model, and to give an idea of test error of final chosen model
- Algorithm
- Need to repeat entire model development process each time

K-fold Cross-Validation



Leave-One-Out CV: Q1 prismia

Differences between the K-fold, LOO CV, and Normal V

- Q2 Bias: would you prefer LOO CV or Normal Validation?

Differences between the K-fold, LOO CV, and Normal V

- Q2 Bias: would you prefer LOO CV or Normal Validation?
- Does the CV tend to overestimate the test error rate?

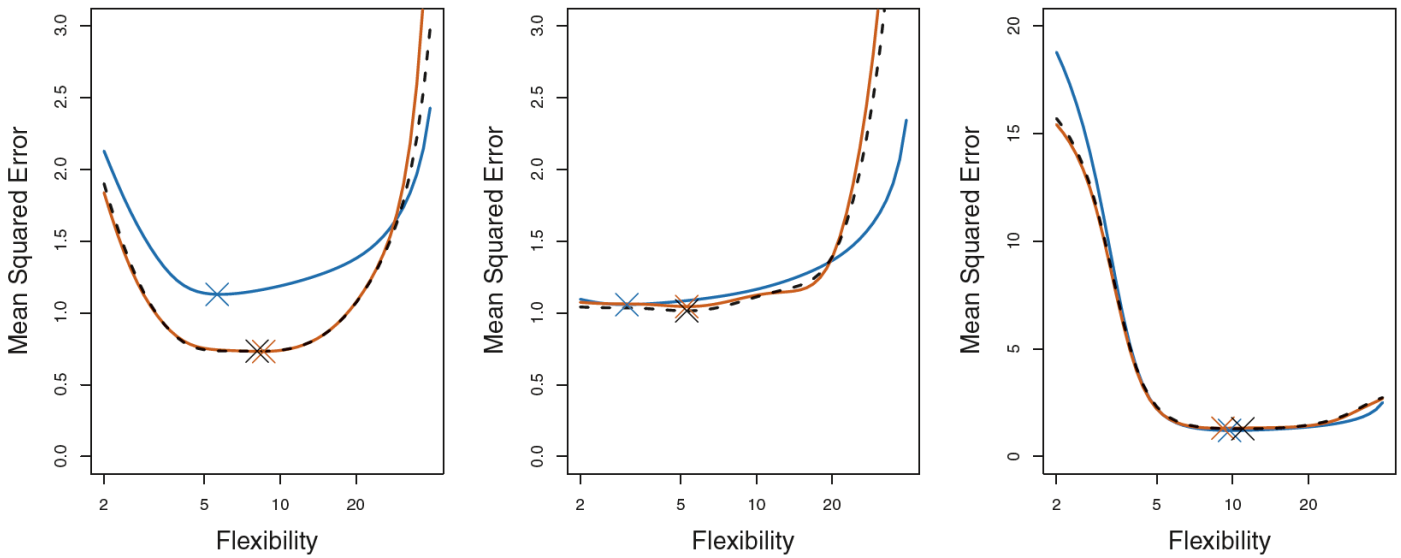
Differences between the K-fold, LOO CV, and Normal V

- Q2 Bias: would you prefer LOO CV or Normal Validation?
- Does the CV tend to overestimate the test error rate?
- Will the normal validation yeald to different Results?

Differences between the K-fold, LOO CV, and Normal V

- Q2 Bias: would you prefer LOO CV or Normal Validation?
- Does the CV tend to overestimate the test error rate?
- Will the normal validation yeald to different Results?
- for LOO with the high leverage: amount that an observation influences its own fit
- What is the advantage of using $k = 5$ or $k = 10$ rather than $k = n$?

Differences between the K-fold, LOO CV, and Normal V



Differences between the K-fold and LOO CV

- Q3 Bias Reduction: would you prefer K-fold or LOO CV?

Differences between the K-fold and LOO CV

- Q3 Bias Reduction: would you prefer K-fold or LOO CV?
- Variance: Does the LOO CV tends to have higher variance than the K-fold?

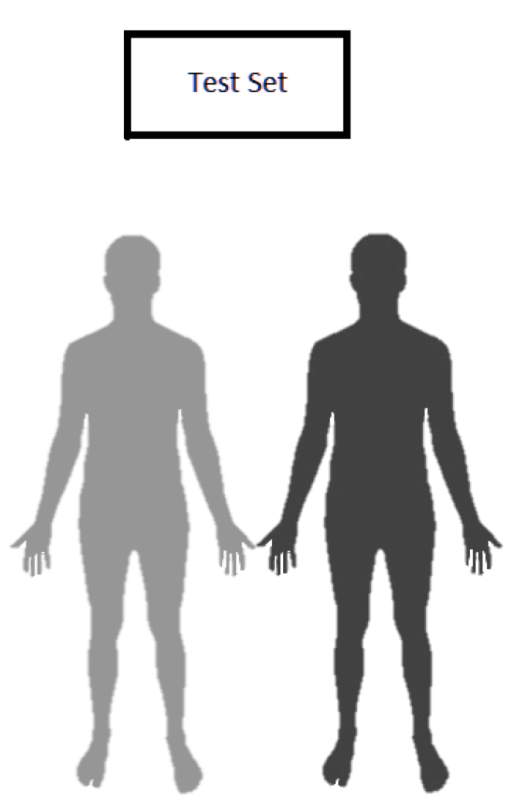
Differences between the K-fold and LOO CV

- Q3 Bias Reduction: would you prefer K-fold or LOO CV?
- Variance: Does the LOO CV tends to have higher variance than the K-fold?
- bias-variance trade-off

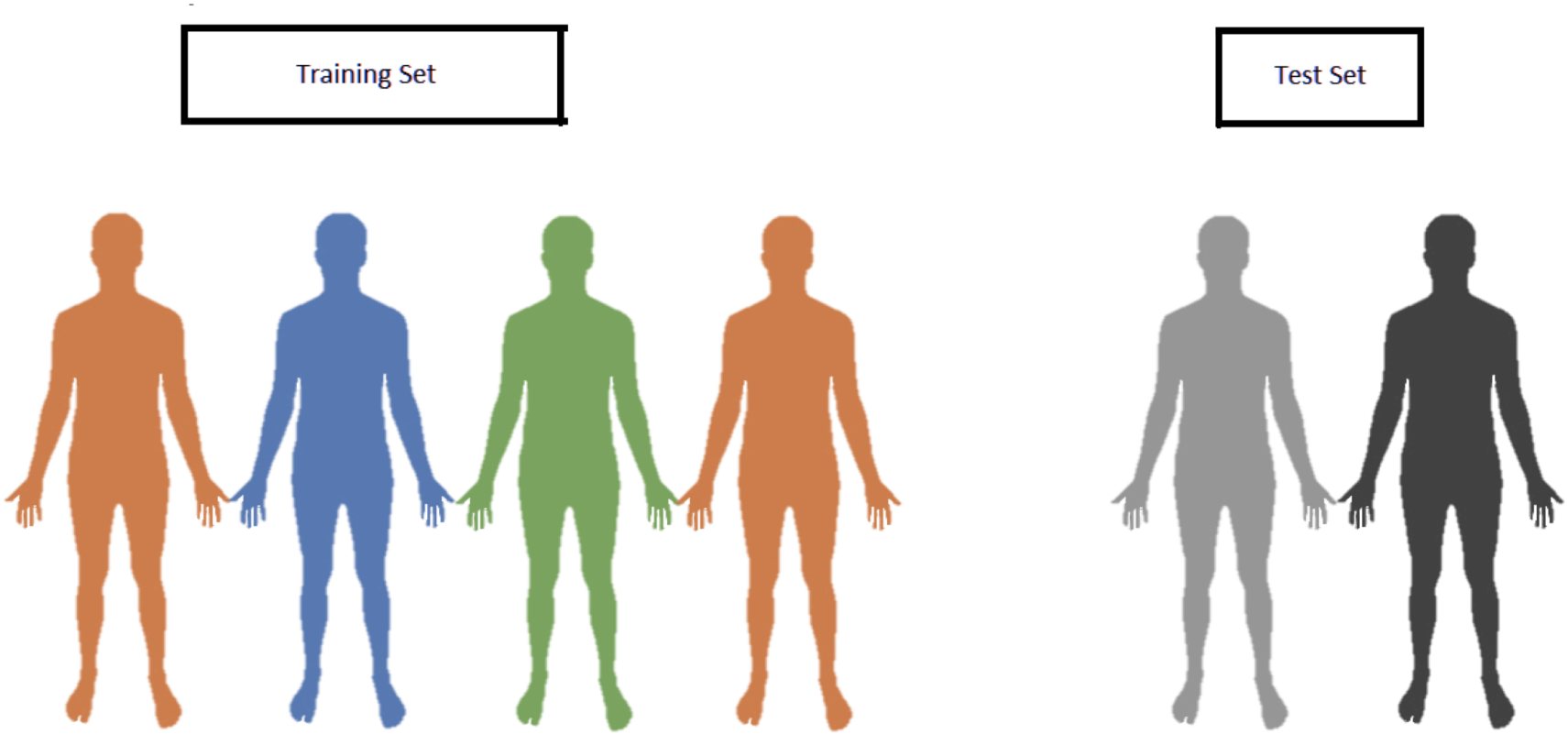
Differences between the K-fold and LOO CV

- Q3 Bias Reduction: would you prefer K-fold or LOO CV?
- Variance: Does the LOO CV tends to have higher variance than the K-fold?
- bias-variance trade-off
- classification problem

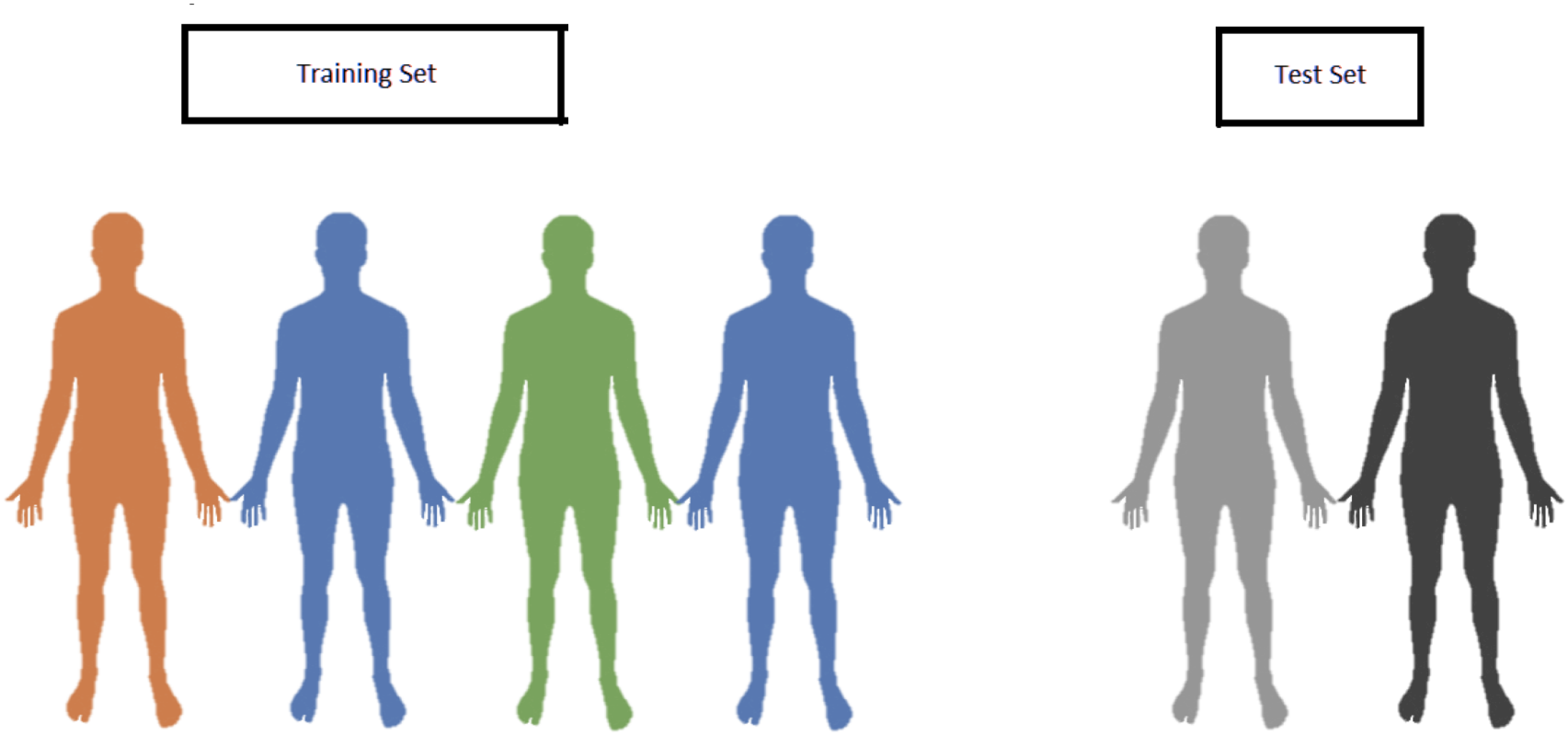
Division of the data set



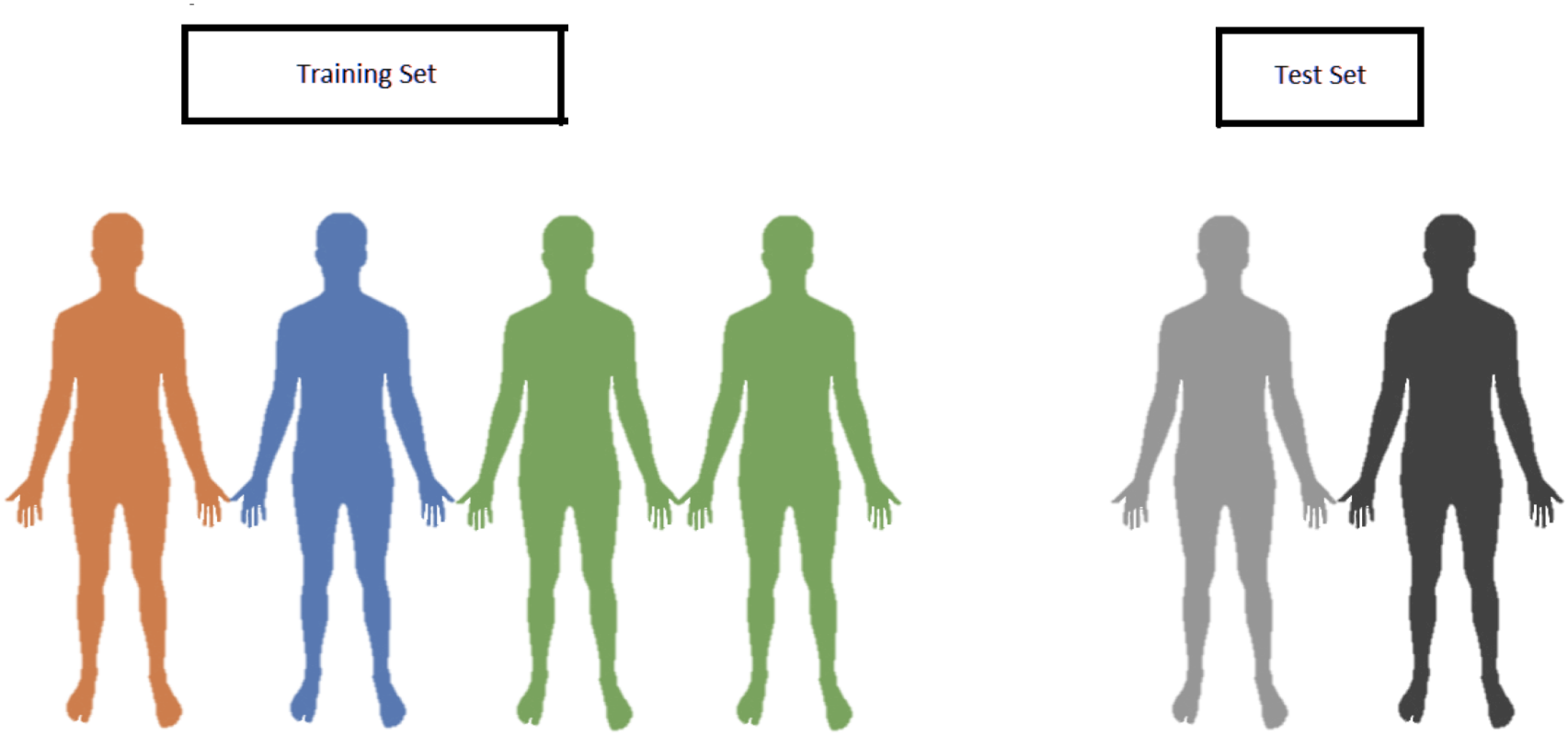
Division of the data set: replacement



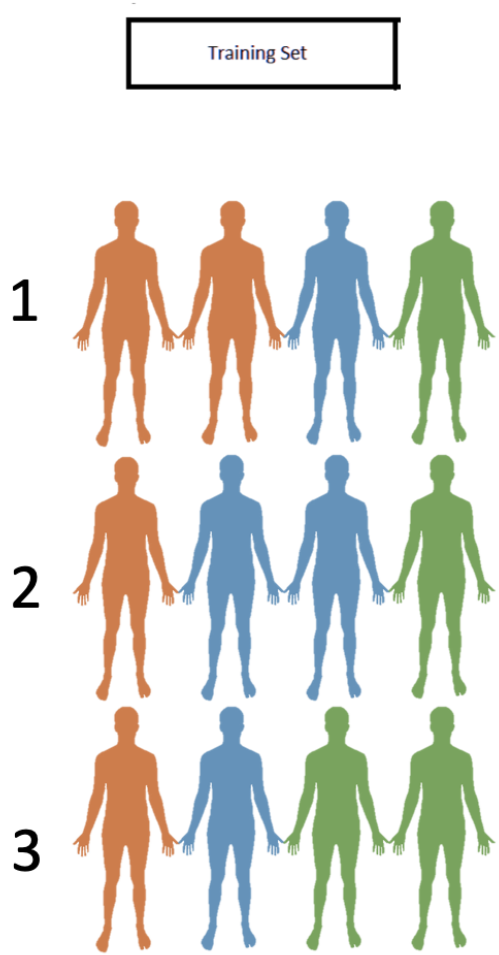
Division of the data set: replacement



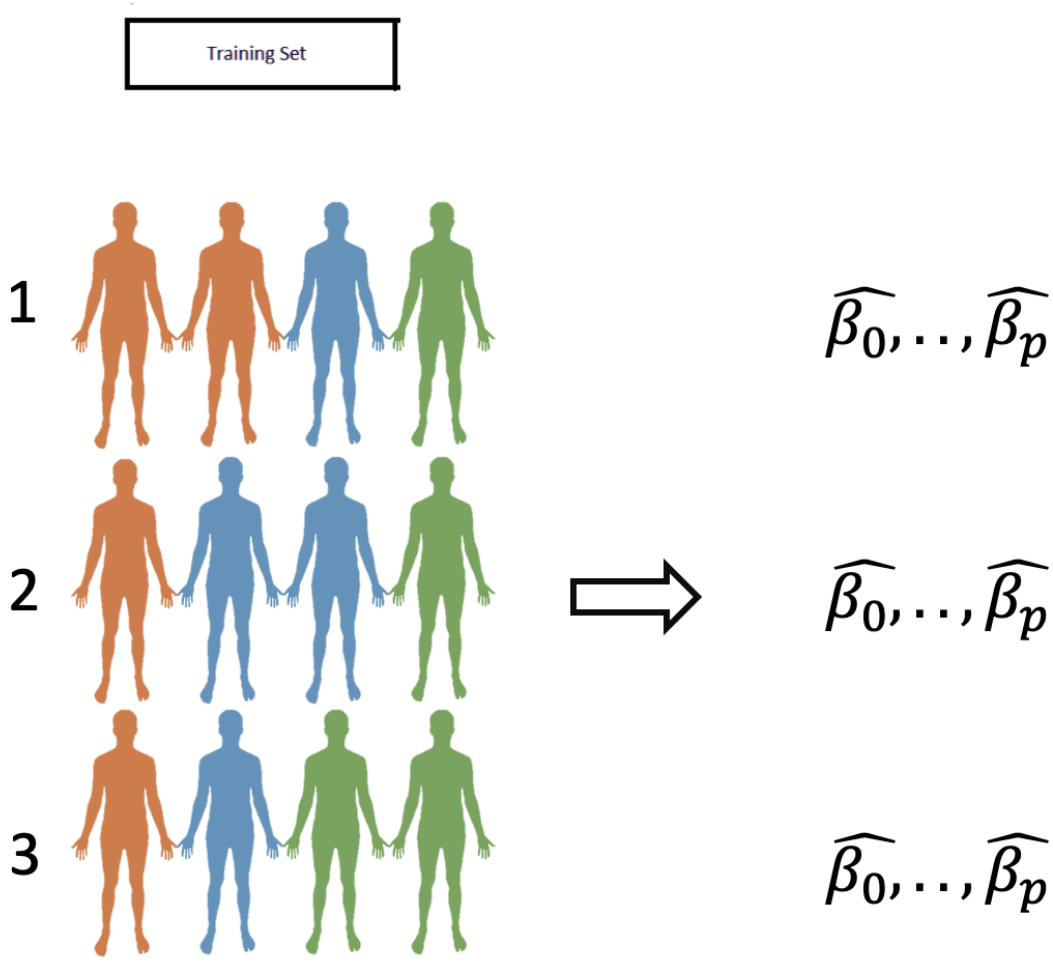
Division of the data set: replacement



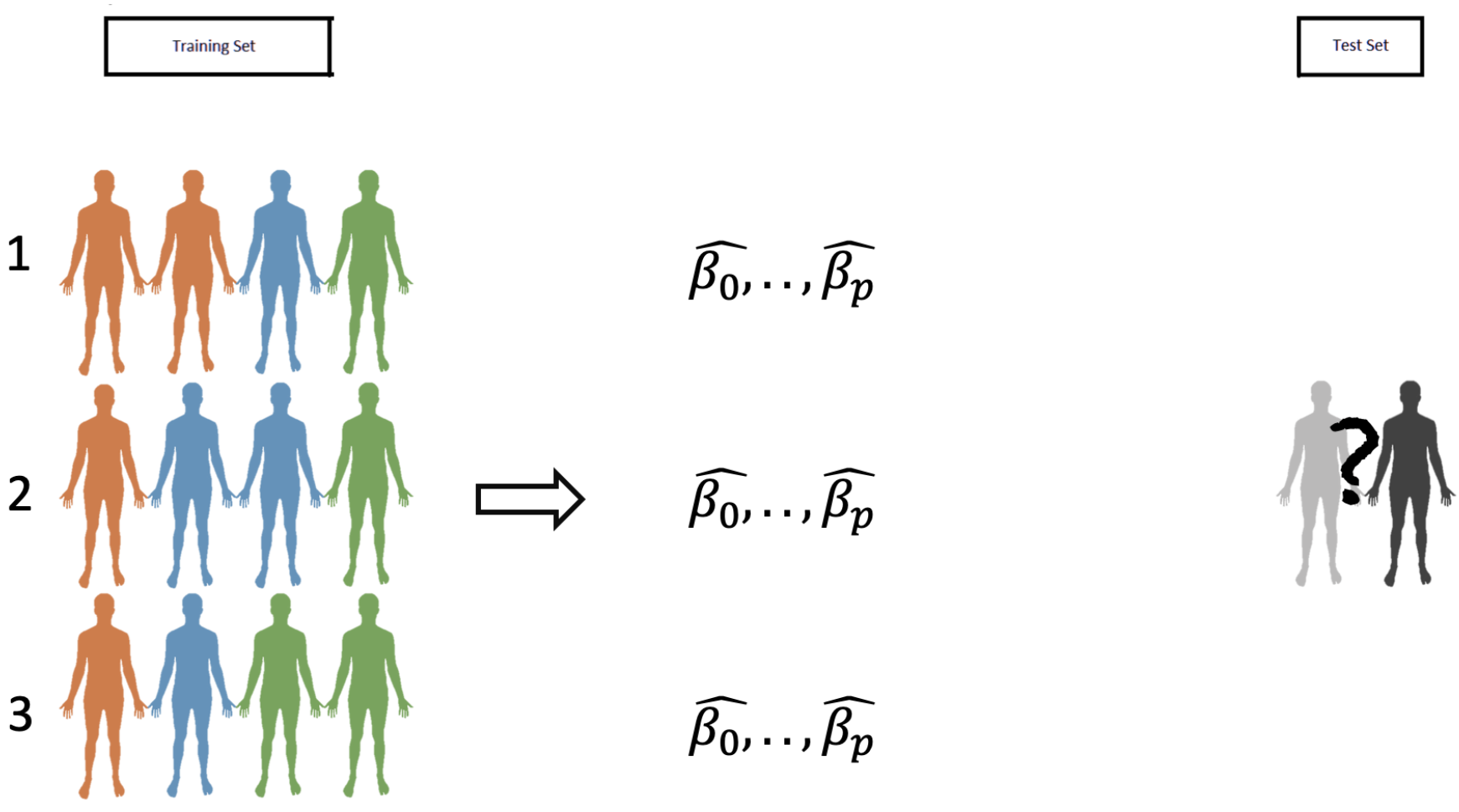
Division of the data set



Division of the data set



Division of the data set



Bootstrap procedure

Bootstrap procedure

```
> sample_class <- seq(1,5, by=1)
> sample_class
[1] 1 2 3 4 5
```

Bootstrap procedure

```
> sample_class <- seq(1,5, by=1)
> sample_class
[1] 1 2 3 4 5
```

```
> train1 <- sample(c(1,2,3),5, replace=T)
> train1
[1] 3 2 1 1 3
```

Bootstrap procedure

```
> sample_class <- seq(1,5, by=1)
> sample_class
[1] 1 2 3 4 5
```

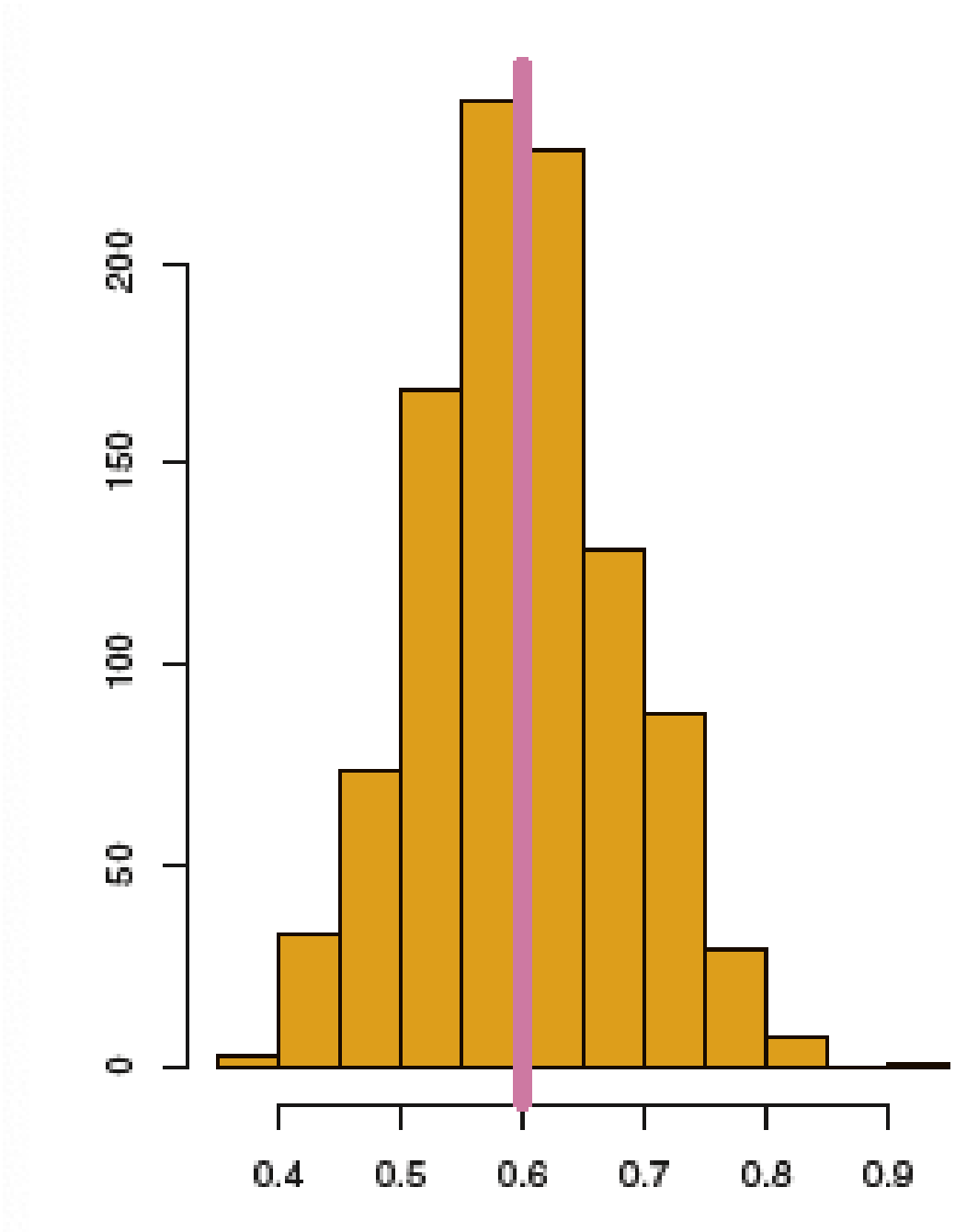
```
> train2 <- sample(c(2,3,4),5, replace=T)
> train2
[1] 2 2 3 3 4
```

Bootstrap procedure

```
> sample_class <- seq(1,5, by=1)
> sample_class
[1] 1 2 3 4 5
```

```
> train3 <- sample(c(3,4,5),5, replace=T)
> train3
[1] 3 4 4 4 5
```

Bootstrap procedure



Bootstrap procedure