Stat 306: Finding Relationships in Data.

Lecture 1

Introduction to Course

Stat 306: Finding Relationships in Data.

The main topic of this course is **regression**, which means fitting prediction equations.

Regression is a common statistical method in scientific research.

Statistics – Recap: the two sample t-test

Age vs. Money

Age vs. Money

Age vs. Money

Dependent variable

Independent variable

Age vs. Money

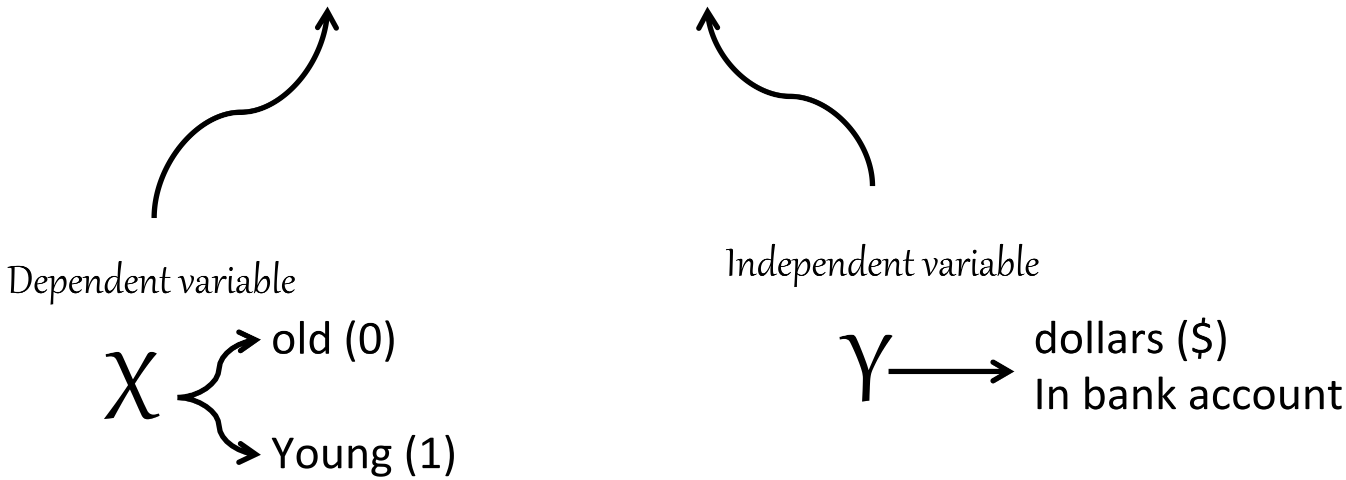
Dependent variable

X

Independent variable

Y

Age vs. Money


Dependent variable
X
old (0)
Young (1)
Independent variable
Y
dollars ($)
In bank account

Age vs. Money

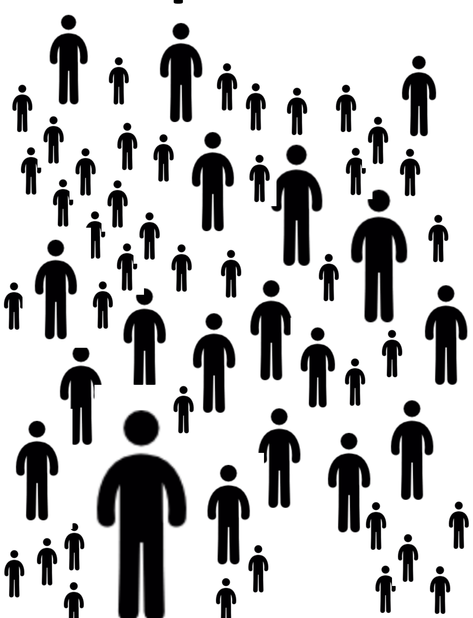
Dependent variable

X → old (0)
X → young (1)

Independent variable

Y → dollars ($)
In bank account

Population



Age vs. Money

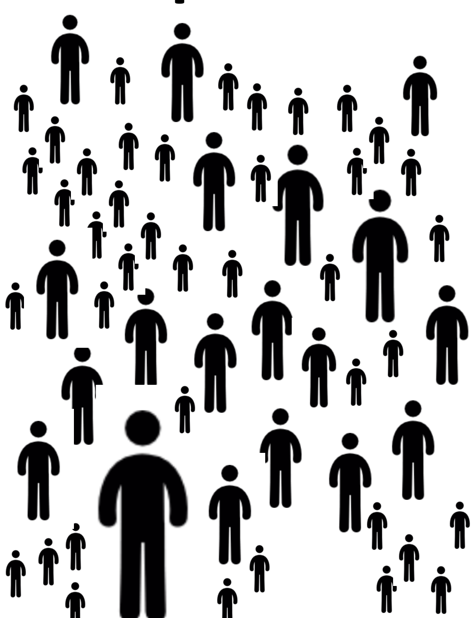
Dependent variable

X → old (0), young (1)

Independent variable

Y → dollars ($) In bank account

Population



Population parameters

μ_0

μ_1

σ^2

Age vs. Money

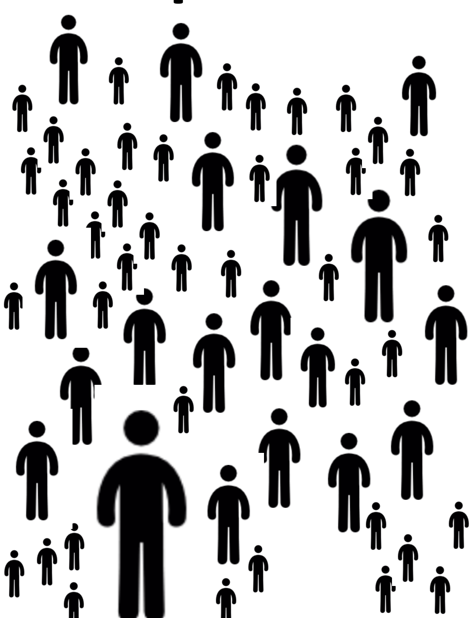
Dependent variable

X → old (0), young (1)

Independent variable

Y → dollars ($)
In bank account

Population



Population parameters

μ_0 ← *Mean money ($) for old people*

μ_1 ← *Mean money ($) for young people*

σ^2 ← *Variance ($) for everyone*

Age vs. Money

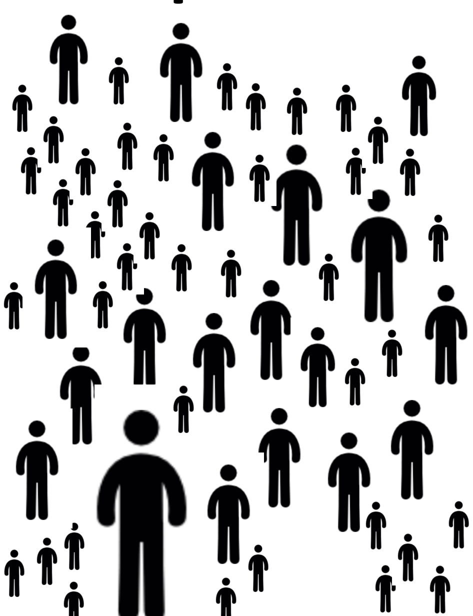
Dependent variable

X → old (0), young (1)

Independent variable

Y → dollars ($)
In bank account

Population



Population parameters

$$\mu_0 , \mu_1 , \sigma^2$$

Hypothesis Test

$$H_0 : \mu_0 = \mu_1$$
$$H_1 : \mu_0 \neq \mu_1$$

Age vs. Money

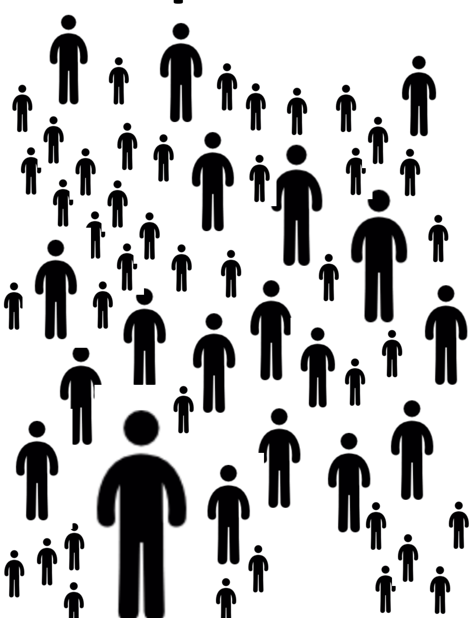
Dependent variable

X → old (0)
X → young (1)

Independent variable

Y → dollars ($) In bank account

Population



Population parameters

$$\mu_0 , \mu_1 , \sigma^2$$

Hypothesis Test

$$H_0 : \mu_0 = \mu_1$$

"Null" hypothesis

$$H_1 : \mu_0 \neq \mu_1$$

"Alternative" hypothesis

Age vs. Money

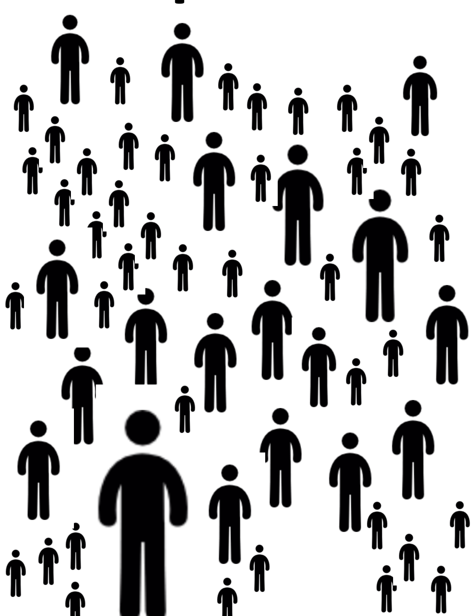
Dependent variable

X → old (0), young (1)

Independent variable

Y → dollars (\$) In bank account

Population



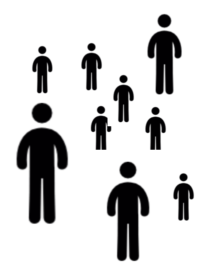
Population parameters

$$\mu_0 , \mu_1 , \sigma^2$$

Hypothesis Test

$$H_0 : \mu_0 = \mu_1$$
$$H_1 : \mu_0 \neq \mu_1$$

Sample



Age vs. Money

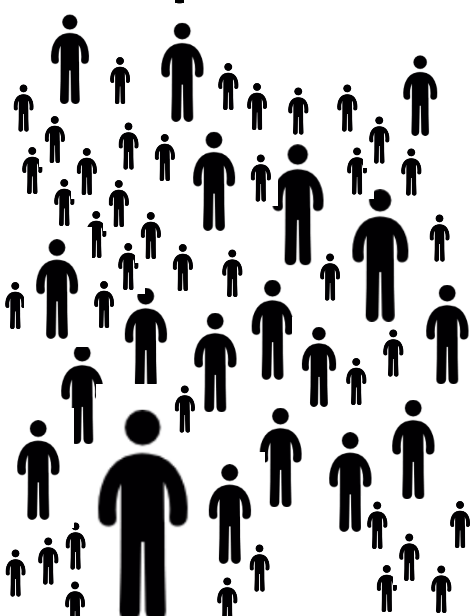
Dependent variable

X → old (0)
X → young (1)

Independent variable

Y → dollars ($) In bank account

Population



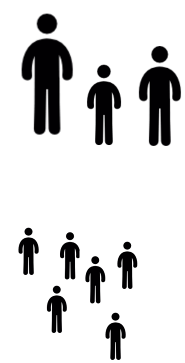
Population parameters

$$\mu_0 , \mu_1 , \sigma^2$$

Hypothesis Test

$$H_0 : \mu_0 = \mu_1$$
$$H_1 : \mu_0 \neq \mu_1$$

Sample



Age vs. Money

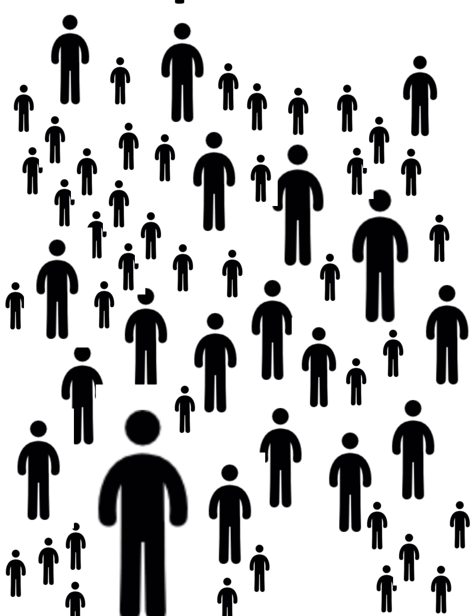
Dependent variable

X → old (0), young (1)

Independent variable

Y → dollars ($) In bank account

Population



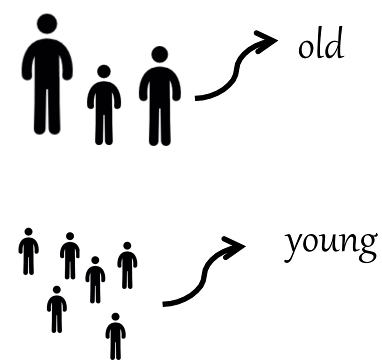
Population parameters

$$\mu_0 , \mu_1 , \sigma^2$$

Hypothesis Test

$$H_0 : \mu_0 = \mu_1$$
$$H_1 : \mu_0 \neq \mu_1$$

Sample



Age vs. Money

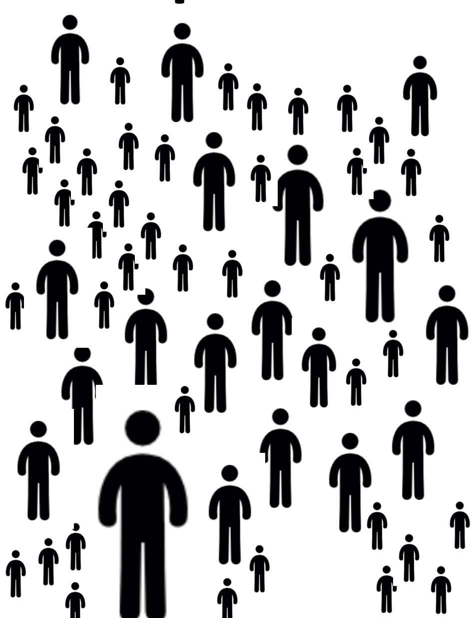
Dependent variable

X → old (0)
X → young (1)

Independent variable

Y → dollars ($) In bank account

Population



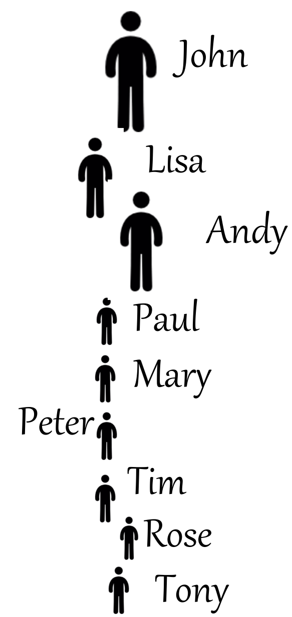
Population parameters

$$\mu_0 , \mu_1 , \sigma^2$$

Hypothesis Test

$$H_0 : \mu_0 = \mu_1$$
$$H_1 : \mu_0 \neq \mu_1$$

Sample, n=9



Age vs. Money

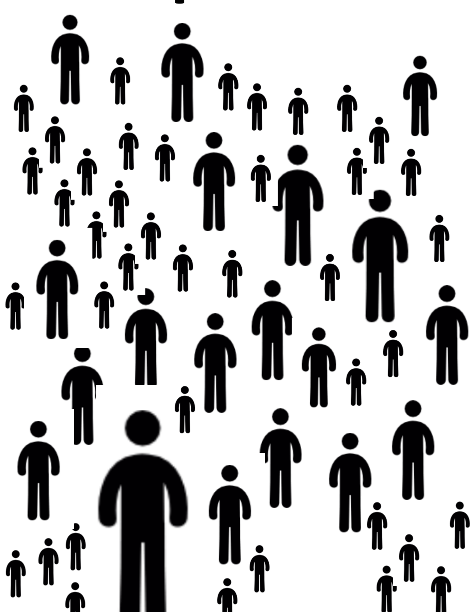
Dependent variable

χ → old (0) / young (1)

Independent variable

γ → dollars ($) In bank account

Population



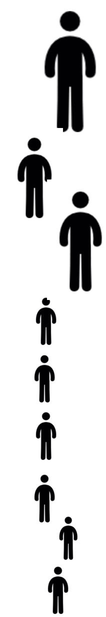
Population parameters

$$\mu_0 , \mu_1 , \sigma^2$$

Hypothesis Test

$$H_0 : \mu_0 = \mu_1$$
$$H_1 : \mu_0 \neq \mu_1$$

Sample, n=9



χ	y
old	71
	54
old	43
old	
young	45
young	21
young	11
young	30
young	45
young	10

Age vs. Money

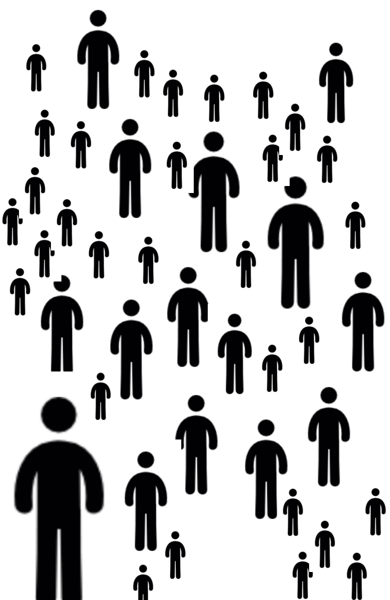
Dependent variable

X → old (0)
X → young (1)

Independent variable

Y → dollars ($) In bank account

Population



Population parameters

μ_0 , μ_1 , σ^2

Hypothesis Test

$H_0 : \mu_0 = \mu_1$

$H_1 : \mu_0 \neq \mu_1$

Sample statistics

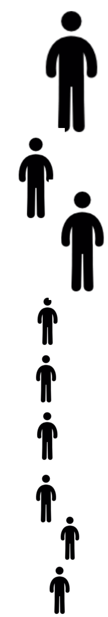
$\bar{y}_0 = 56$

$\bar{y}_1 = 27$

$\bar{y}_0 - \bar{y}_1 = 29$

$s_p = 10.81$

Sample, n=9



x	y
old	71
old	54
old	43
young	45
young	21
young	11
young	30
young	45
young	10

Age vs. Money

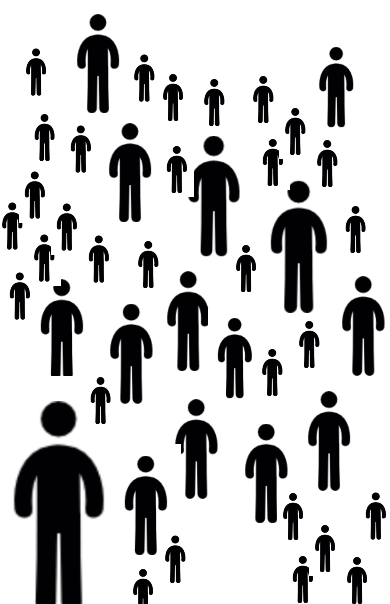
Dependent variable

X → old (0), young (1)

Independent variable

Y → dollars ($) In bank account

Population



Population parameters

μ_0, μ_1, σ^2

Hypothesis Test

$H_0 : \mu_0 = \mu_1$

$H_1 : \mu_0 \neq \mu_1$

Sample statistics

$\bar{y}_0 = 56$

$\bar{y}_1 = 27$

$\bar{y}_0 - \bar{y}_1 = 29$

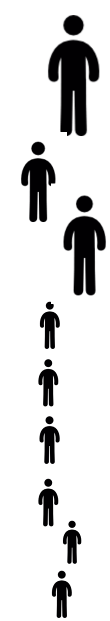
$s_p = 10.81$

$t = 2.68$, $df = 7$

p-value = 0.03

95% C.I. = [3.4, 54.6]

Sample, n=9



X	y
old	71
old	54
old	43
young	45
young	21
young	11
young	30
young	45
young	10

Age vs. Money

$\bar{y}_0 = 56$
$\bar{y}_1 = 27$
$\bar{y}_0 - \bar{y}_1 = 29$
$s_p = 10.81$

Objective: The purpose of this observational study was to demonstrate if, and to what extent, age is associated with money.

Design and Methods: We surveyed a number individuals and for each determined approximate age (recorded as “old” or “young”) and the amount of money (in dollars) in their bank accounts. Comparison of the two groups was done using a Student two sample *t*-test.

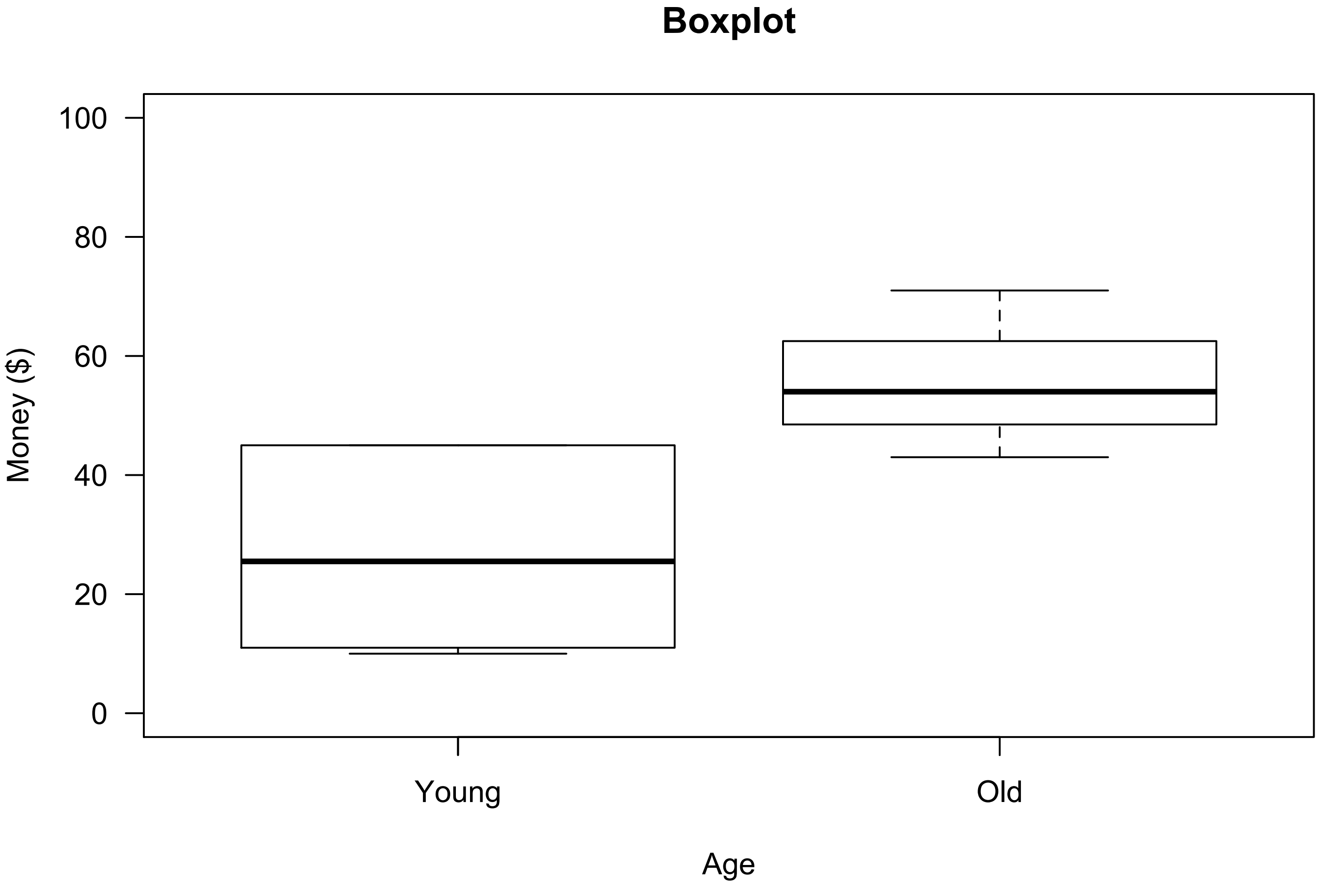
t = 2.68, *df* = 7
p-value = 0.03
95% C.I. = [3.4, 54.6]

Results: We obtained a random sample of *n* = 9 subjects. The “young” group had an average of $27, while the “old” group had an average of $56. This estimated difference of $29 (95% C.I. = [$3.4, $54.6]) is statistically significant, *t* = 2.68, *df* = 7; *p*-value = 0.03.

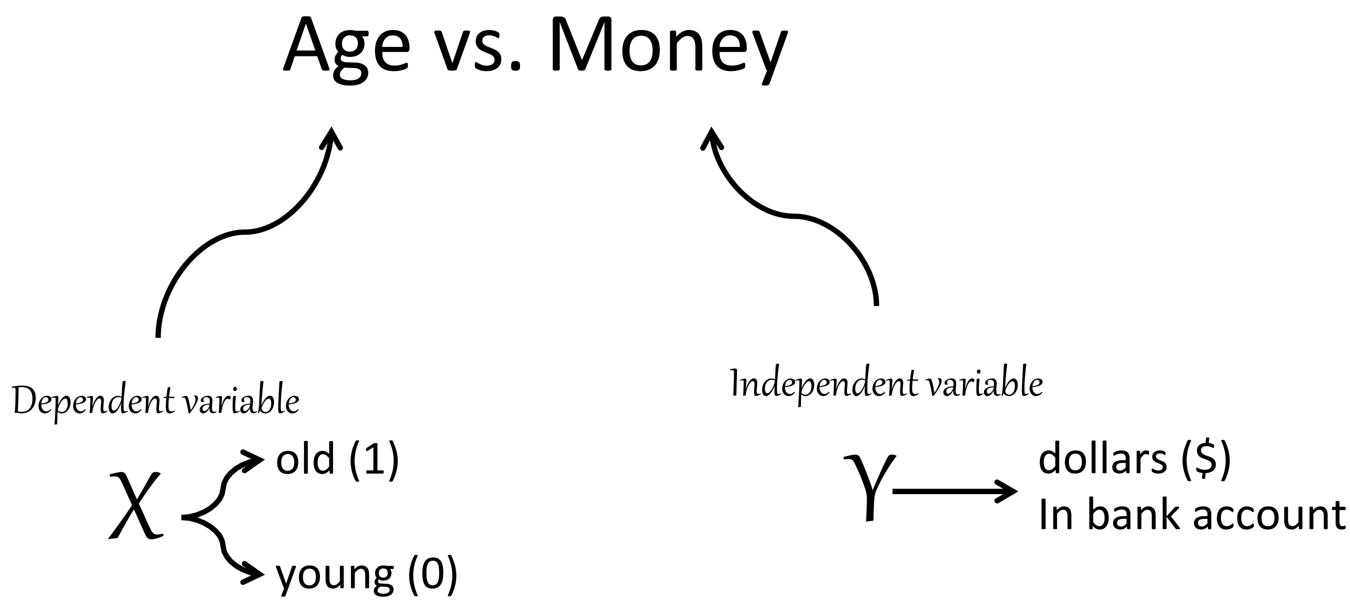
Conclusions: We found that, as hypothesized, age is associated with money. On average, younger people have less in their accounts than older people.

Small Print: The analysis rests on the following assumptions:

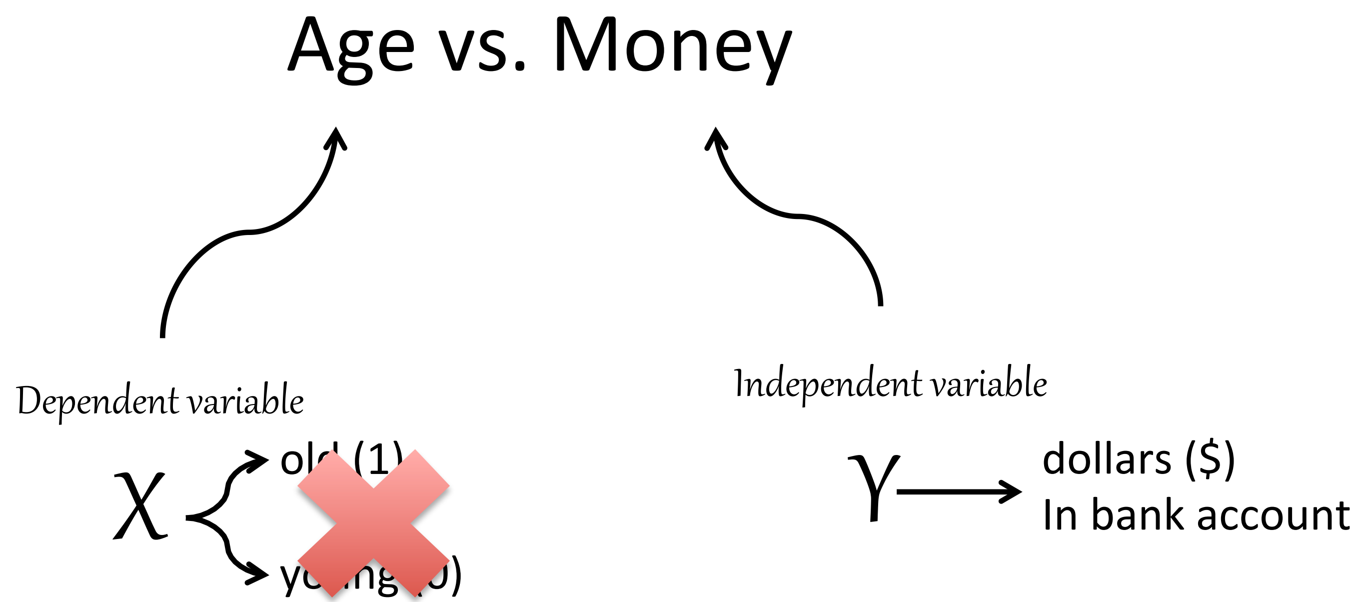
- the observations are independently and identically distributed.
- the independent variable, money, is normally distributed.
- the two populations being compared have the same variance.


Boxplot
100
80
60
40
20
0
Money ($)
Young
Old
Age

Linear Regression



Linear Regression



Linear Regression

Age vs. Money

Dependent variable

X → Age in Years

Independent variable

Y → dollars ($) In bank account

Linear Regression

Age vs. Money

PREDICTOR *variable*

X → Age in Years

RESPONSE *variable*

Y → dollars ($) In bank account

Age vs. Money

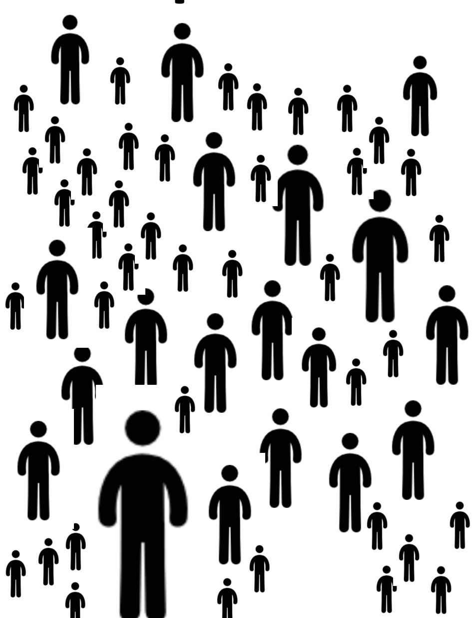
PREDICTOR *variable*

$X \longrightarrow$ Age in Years

RESPONSE *variable*

$Y \longrightarrow$ dollars ($) In bank account

Population



Population parameters

β_0

β_1

σ^2

Age vs. Money

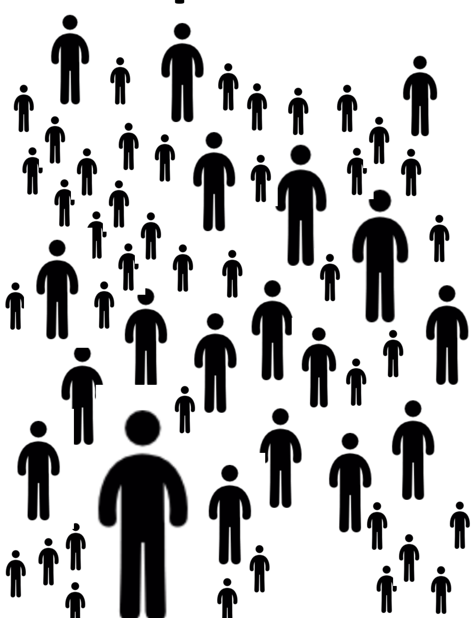
PREDICTOR *variable*

$X \longrightarrow$ Age in Years

RESPONSE *variable*

$Y \longrightarrow$ dollars ($) In bank account

Population



Population parameters

$$\beta_0 , \beta_1 , \sigma^2$$

Hypothesis Test

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Age vs. Money

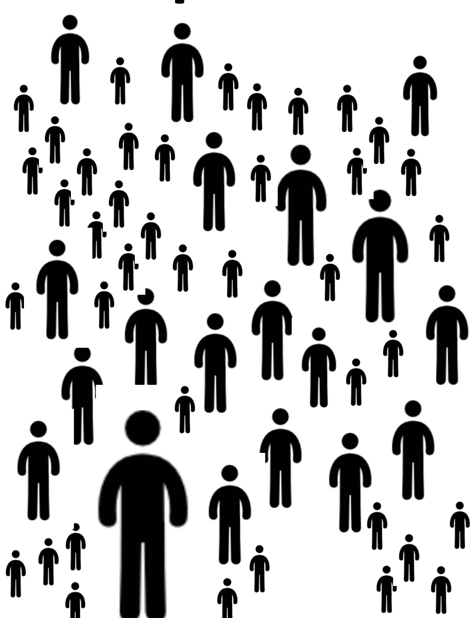
PREDICTOR variable

$X \longrightarrow$ Age in Years

RESPONSE variable

$Y \longrightarrow$ dollars ($) In bank account

Population



Population parameters

$$\beta_0 , \beta_1 , \sigma^2$$

Hypothesis Test

$$H_0 : \beta_1 = 0$$

“Null” hypothesis

$$H_1 : \beta_1 \neq 0$$

“Alternative” hypothesis

Age vs. Money

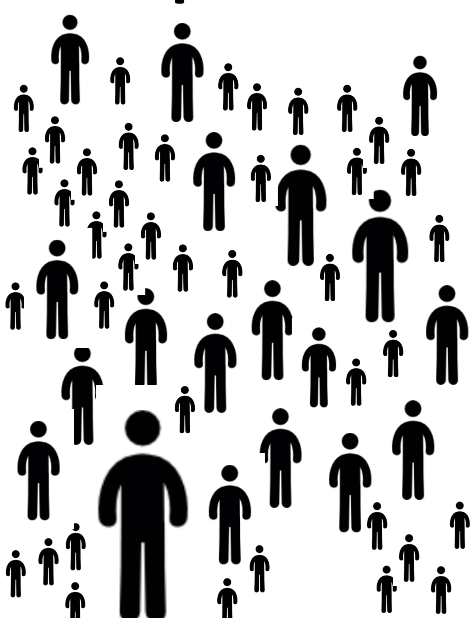
PREDICTOR variable

$X \longrightarrow$ Age in Years

RESPONSE variable

$Y \longrightarrow$ dollars ($) In bank account

Population



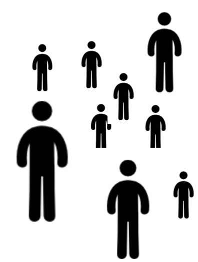
Population parameters

$$\beta_0 , \beta_1 , \sigma^2$$

Hypothesis Test

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Sample



Age vs. Money

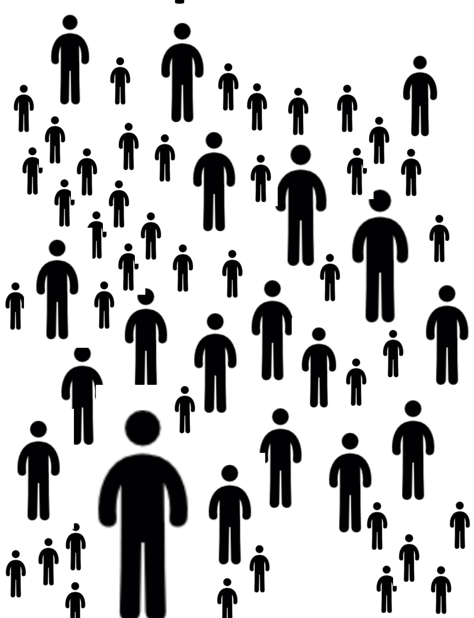
PREDICTOR *variable*

$X \longrightarrow$ Age in Years

RESPONSE *variable*

$Y \longrightarrow$ dollars ($) In bank account

Population



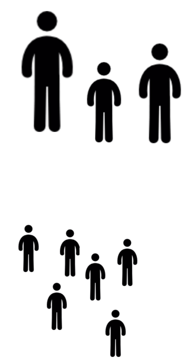
Population parameters

$$\beta_0 , \beta_1 , \sigma^2$$

Hypothesis Test

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Sample



Age vs. Money

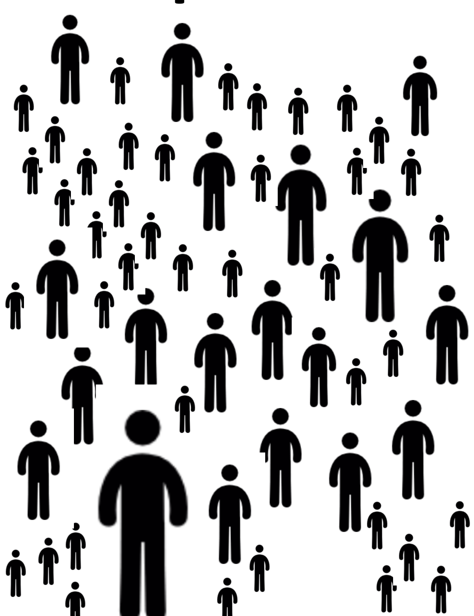
PREDICTOR *variable*

$X \longrightarrow$ Age in Years

RESPONSE *variable*

$Y \longrightarrow$ dollars ($) In bank account

Population



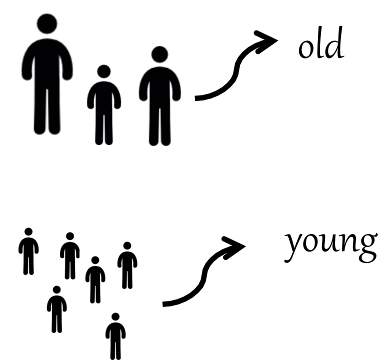
Population parameters

$$\beta_0 , \beta_1 , \sigma^2$$

Hypothesis Test

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Sample



Age vs. Money

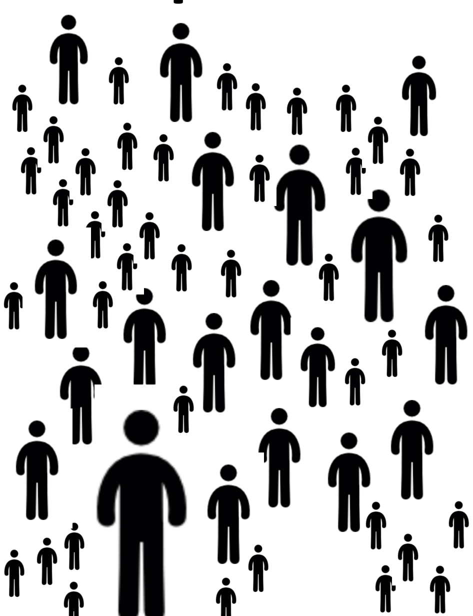
PREDICTOR *variable*

$X \longrightarrow$ Age in Years

RESPONSE *variable*

$Y \longrightarrow$ dollars ($) In bank account

Population



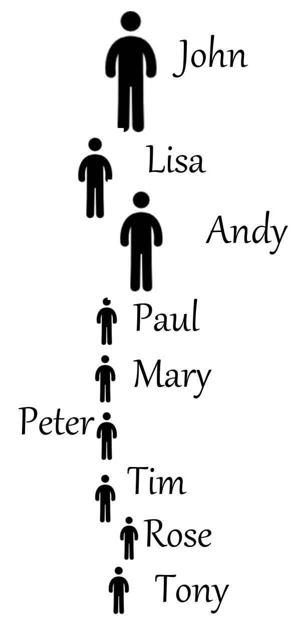
Population parameters

$$\beta_0 , \beta_1 , \sigma^2$$

Hypothesis Test

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Sample, n=9



Age vs. Money

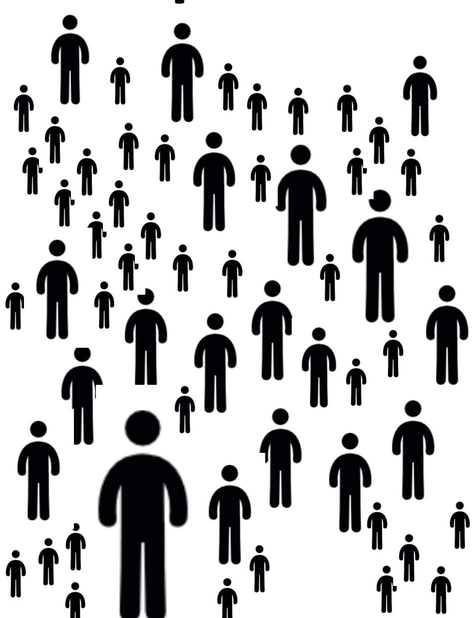
PREDICTOR *variable*

$\chi \longrightarrow$ Age in Years

RESPONSE *variable*

$\gamma \longrightarrow$ dollars ($) In bank account

Population



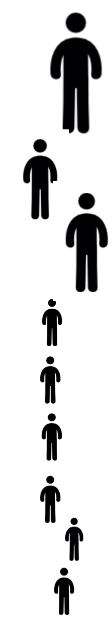
Population parameters

$$\beta_0 , \beta_1 , \sigma^2$$

Hypothesis Test

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Sample, n=9



χ	y
82	71
45	54
71	43
22	45
29	21
9	11
12	30
18	45
24	10

Age vs. Money

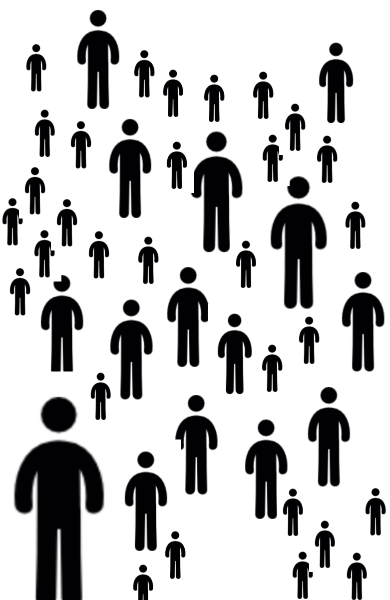
PREDICTOR variable

$X \longrightarrow$ Age in Years

RESPONSE variable

$Y \longrightarrow$ dollars ($) In bank account

Population



Population parameters

$$\beta_0 , \beta_1 , \sigma^2$$

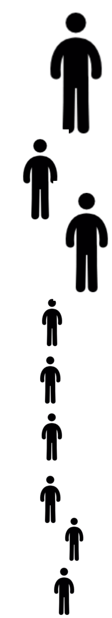
Hypothesis Test

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Sample statistics

$b_0 = 17.7$
$b_1 = 0.55$
$s = 15.5$
$R^2 = 0.49$

Sample, n=9



χ	y
82	71
45	54
71	43
22	45
29	21
9	11
12	30
18	45
24	10

Age vs. Money

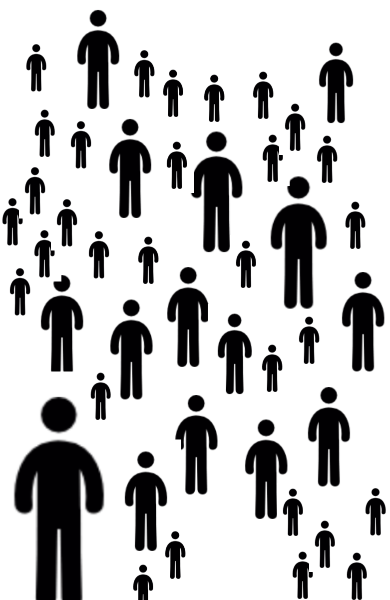
PREDICTOR variable

$X \longrightarrow$ Age in Years

RESPONSE variable

$Y \longrightarrow$ dollars ($) In bank account

Population



Population parameters

$$\beta_0, \beta_1, \sigma^2$$

Hypothesis Test

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

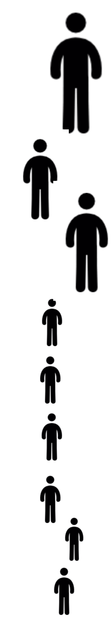
Sample statistics

$b_0 = 17.7$
$b_1 = 0.55$
$s = 15.5$
$R^2 = 0.49$

For parameter β_1 :

95% C.I. $= [0.05, 1.05]$
p-value $= 0.036$

Sample, n=9



x	y
82	71
45	54
71	43
22	45
29	21
9	11
12	30
18	45
24	10

Age vs. Money

Sample statistics

b_0 = 17.7
b_1 = 0.55
s = 15.5
R^2 = 0.49

For parameter β_1 :

95% C.I. $= [0.05, 1.05]$
p-value $= 0.036$

Objective: The purpose of this observational study was to demonstrate if, and to what extent, age is associated with money.

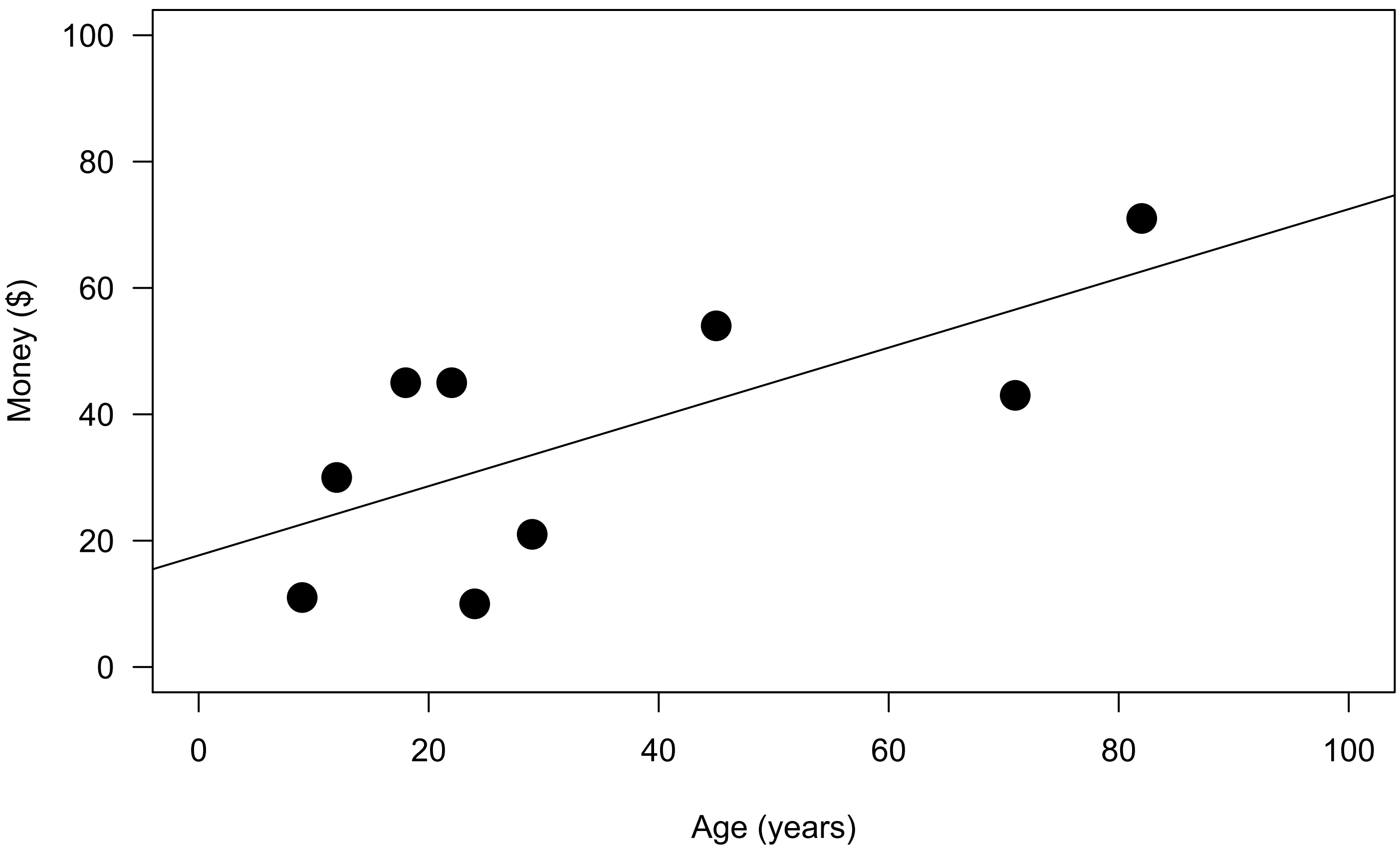
Design and Methods: We collected a random sample of individuals and for each determined their age **(recorded in years)** and the amount of money (in dollars) in their accounts. Analysis of the data was done using **linear regression**.

Results: We obtained a random sample of n = 9 subjects. There is a statistically significant association between age and money (p-value =0.036). For every additional year in age, an individual's amount of money increases on average by an estimated of $0.55 (95% C.I. = [$0.05, $1.05]).

Conclusions: We found that, as hypothesized, age is associated with money. In our sample age accounted for about half of the variability observed in money (R^2=0.49). We **predict** that a 50 year old will have $45.1 (95% P.I. = [$5.6, $84.5]), whereas a 40 year old will have $39.6 (95% P.I. = [$0.8, $78.4]).

Small Print: The analysis rests on the following assumptions:

- the observations are independently and identically distributed.
- the **response** variable, money, is normally distributed.
- Homoscedasticity of residuals or equal variance.
- the relationship between **response** and **predictor** variables is linear.



100
80
60
40
20
0
Money ($)
0
20
40
60
80
100
Age (years)