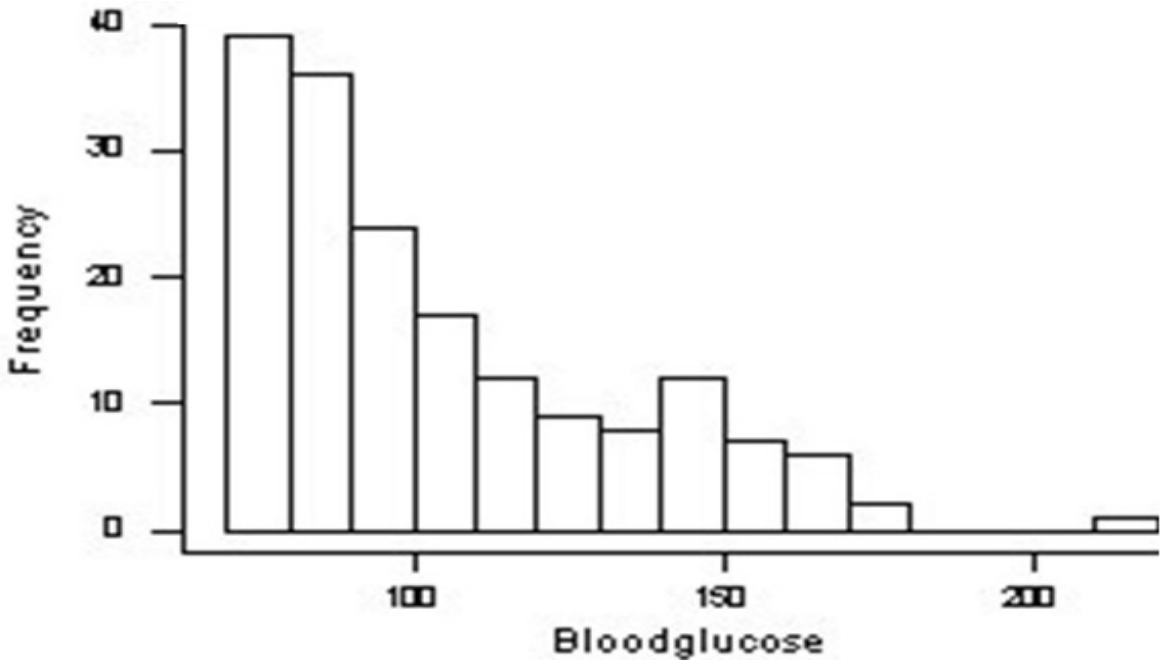
Descriptive Statistics & Graphs

Lecture 13

02/11/2013

Turn on your clickers!



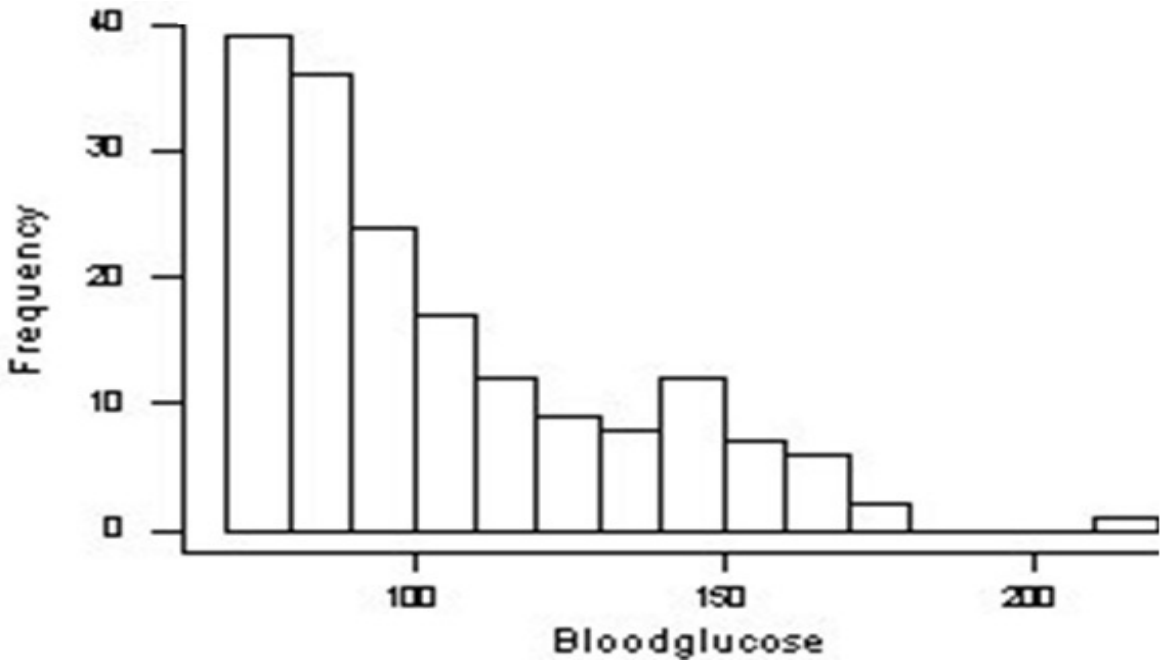
The shape of this distribution is:

(a) unimodal, left skewed.

(b) bimodal.

(c) unimodal, right skewed.

Turn on your clickers!



The shape of this distribution is:

(a) unimodal, left skewed.

(b) bimodal.

(c) unimodal, right skewed.

Turn on your clickers!

A distribution has a mean of 100 and a median of 120. The shape of this distribution is most likely:

(a) skewed left

(b) skewed right

(c) symmetric

Turn on your clickers!

A distribution has a mean of 100 and a median of 120. The shape of this distribution is most likely:

(a) skewed left

(b) skewed right

(c) symmetric

Turn on your clickers!

Which of the following measures is least affected by *outliers*?

(a) mean

(b) standard deviation

(c) IQR (InterQuartile Range)

Turn on your clickers!

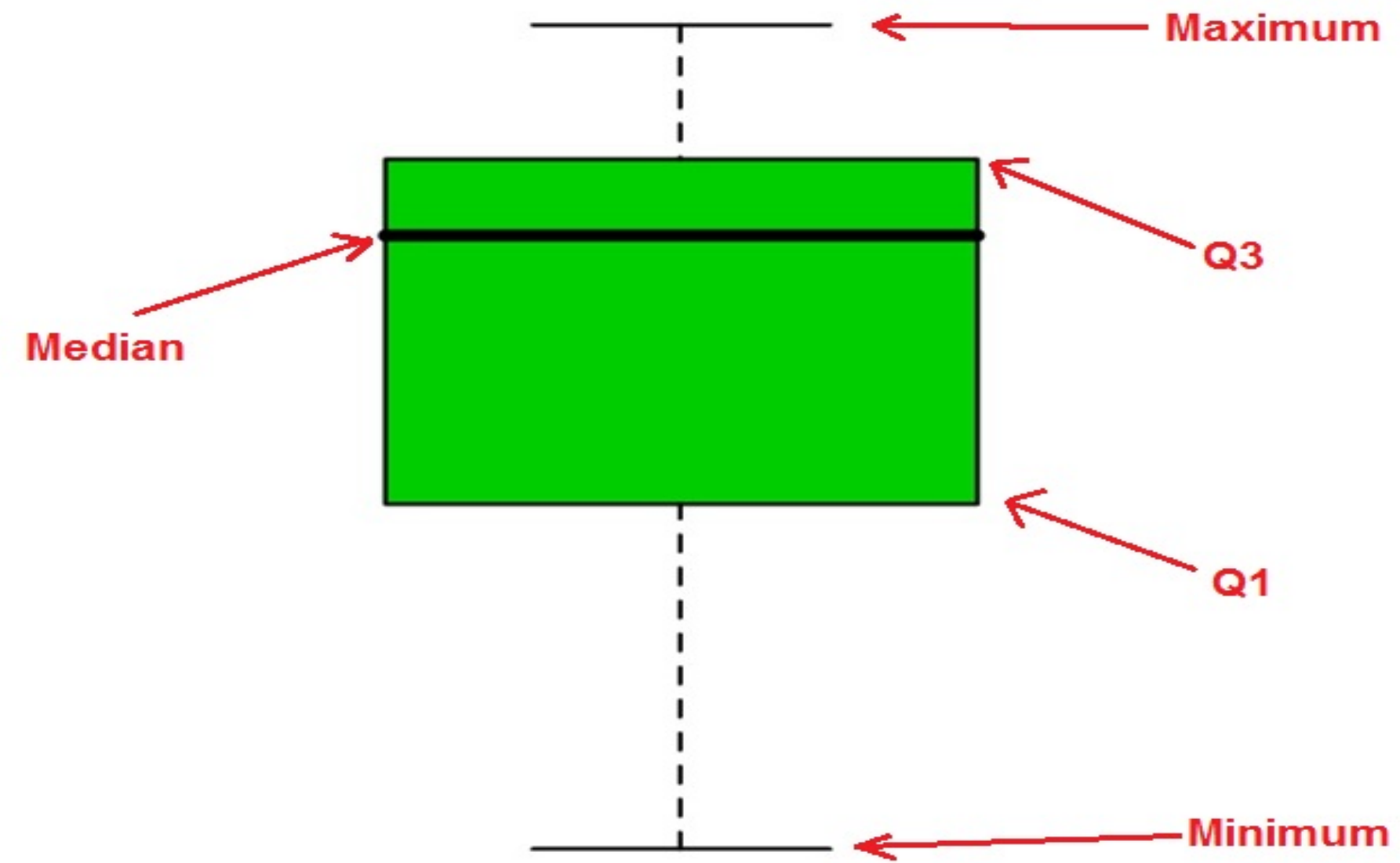
Which of the following measures is least affected by *outliers*?

(a) mean

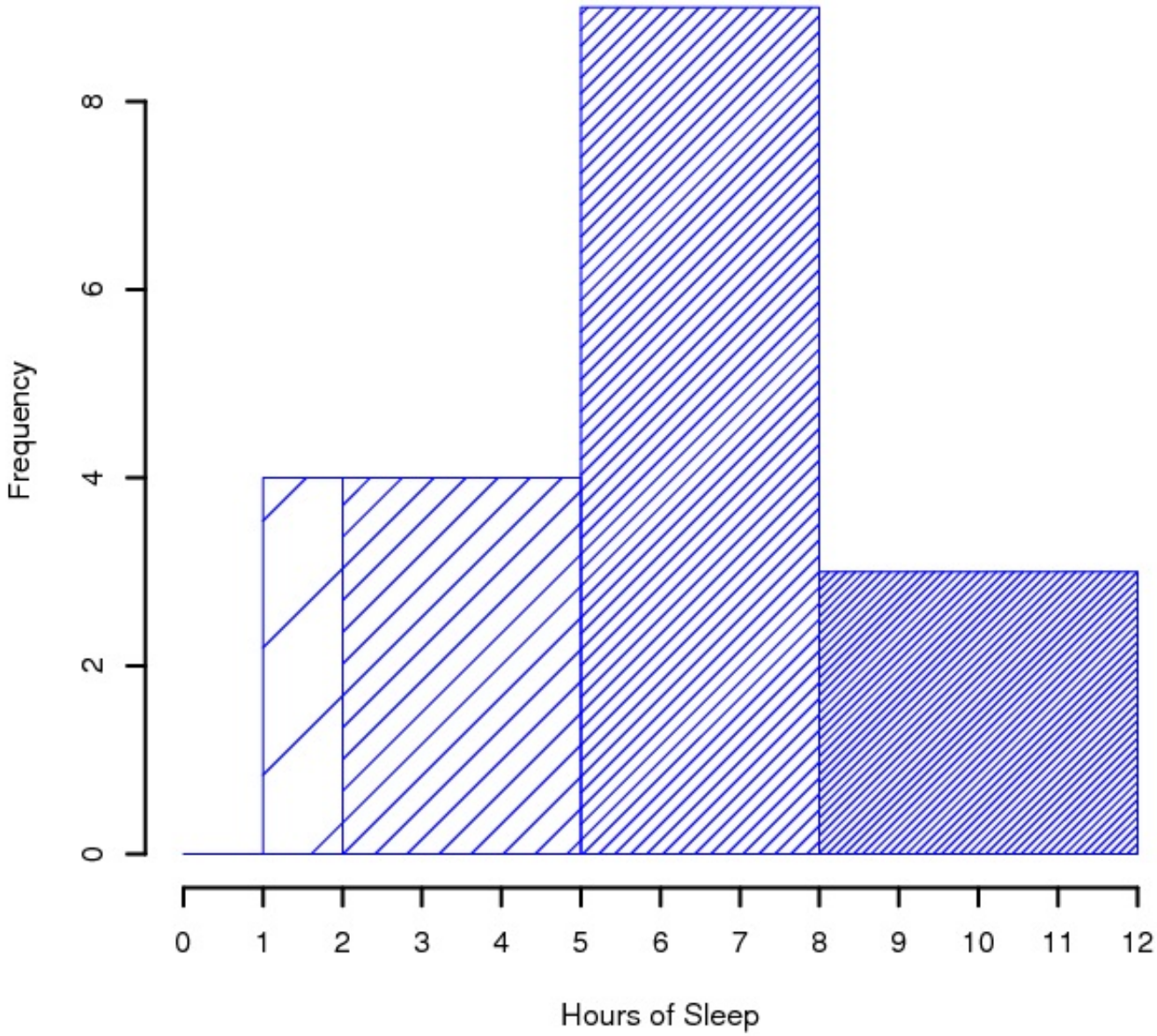
(b) standard deviation

(c) IQR (InterQuartile Range)

Box Pot



The Histogram



The Histogram

Example: *Hours of sleep*

$$\{12, 8.5, 7.2, 7.3, 7.7, 6, 6.5, 4.5, 3, 1.2,$$
$$1.3, 2, 2, 3.8, 6.6, 8.5, 5.9, 4.6, 5.6, 6.7\}$$

Variable: `hours of sleep`,

Values = $[0, 24]$.

1. How many blocks are we going to have?
2. How are we going to determine the length of each block?

Example: *Hours of sleep*

1. Sort the data:

$$\{1.2, 1.3, 2, 2, 3, 3.8, 4.5, 4.6, 5.6, 5.9,$$
$$6, 6.5, 6.6, 6.7, 7.2, 7.3, 7.7, 8.5, 8.5, 12\}$$

2. Choose the desired *class intervals*:

   1-2 hours, 2-5 hours, 5-8 hours, 8-12 hours

   - 4 class intervals
   - 4 unevenly spaced blocks

Example: *Hours of sleep*

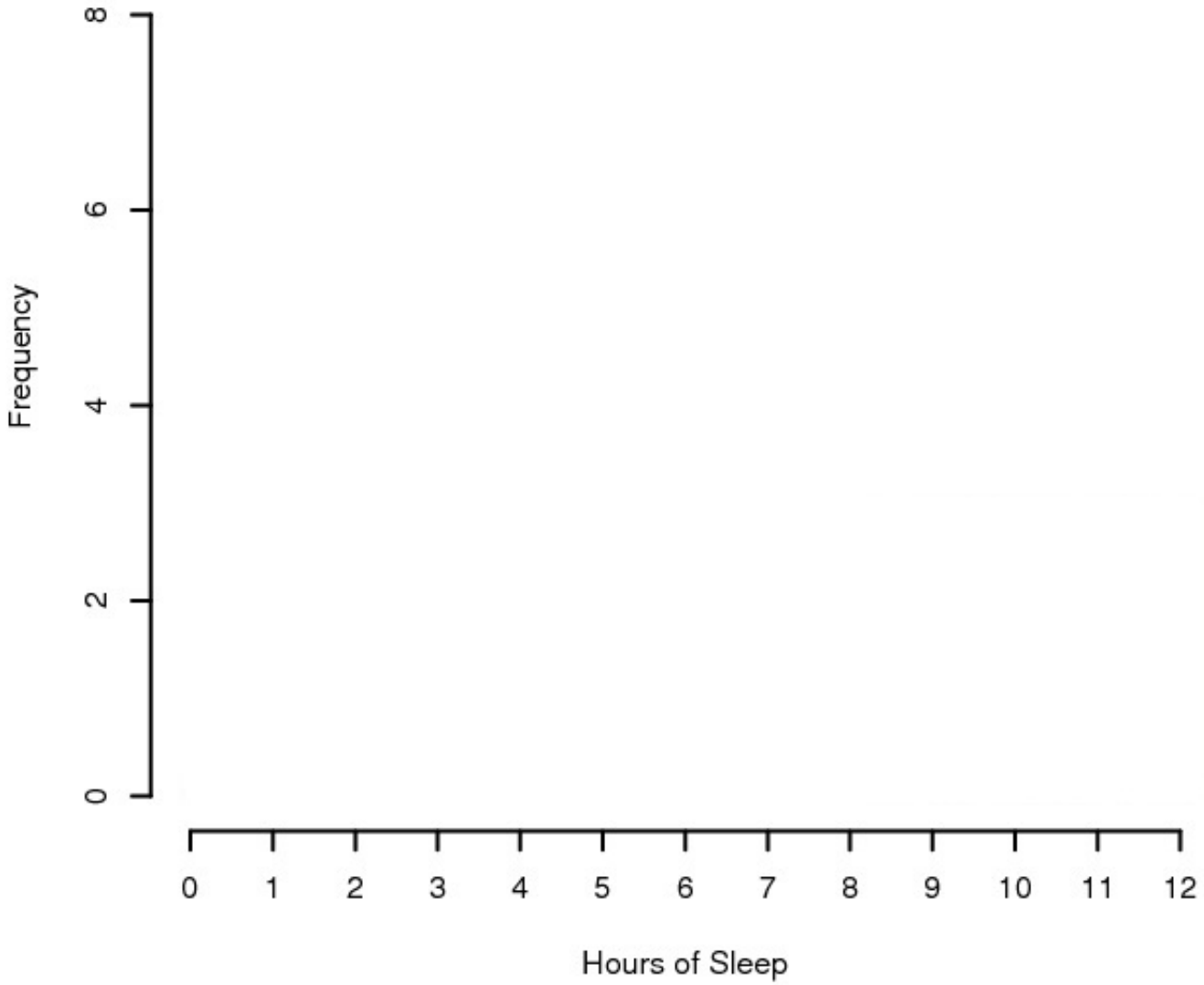
How to draw the block?

- Count the number of datapoints that falls into each class:

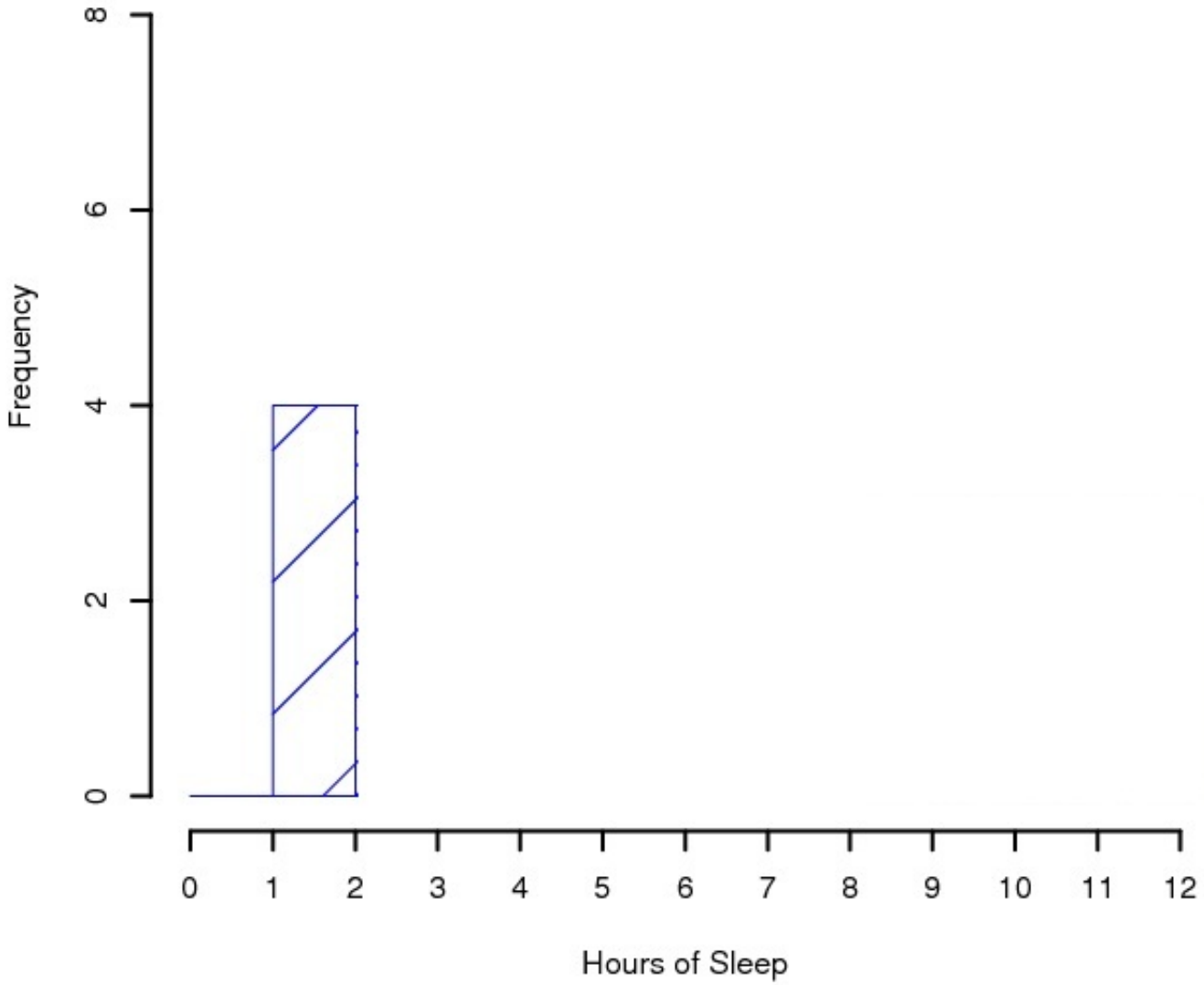
Hours of Sleep (X)	*Counts*	*Proportions*
$1 < X \leq 2$	4	4/20=0.2
$2 < X \leq 5$	4	4/20=0.2
$5 < X \leq 8$	9	9/20=0.45
$8 < X \leq 12$	3	3/20=0.15

The intervals are not necessary to have the same length.

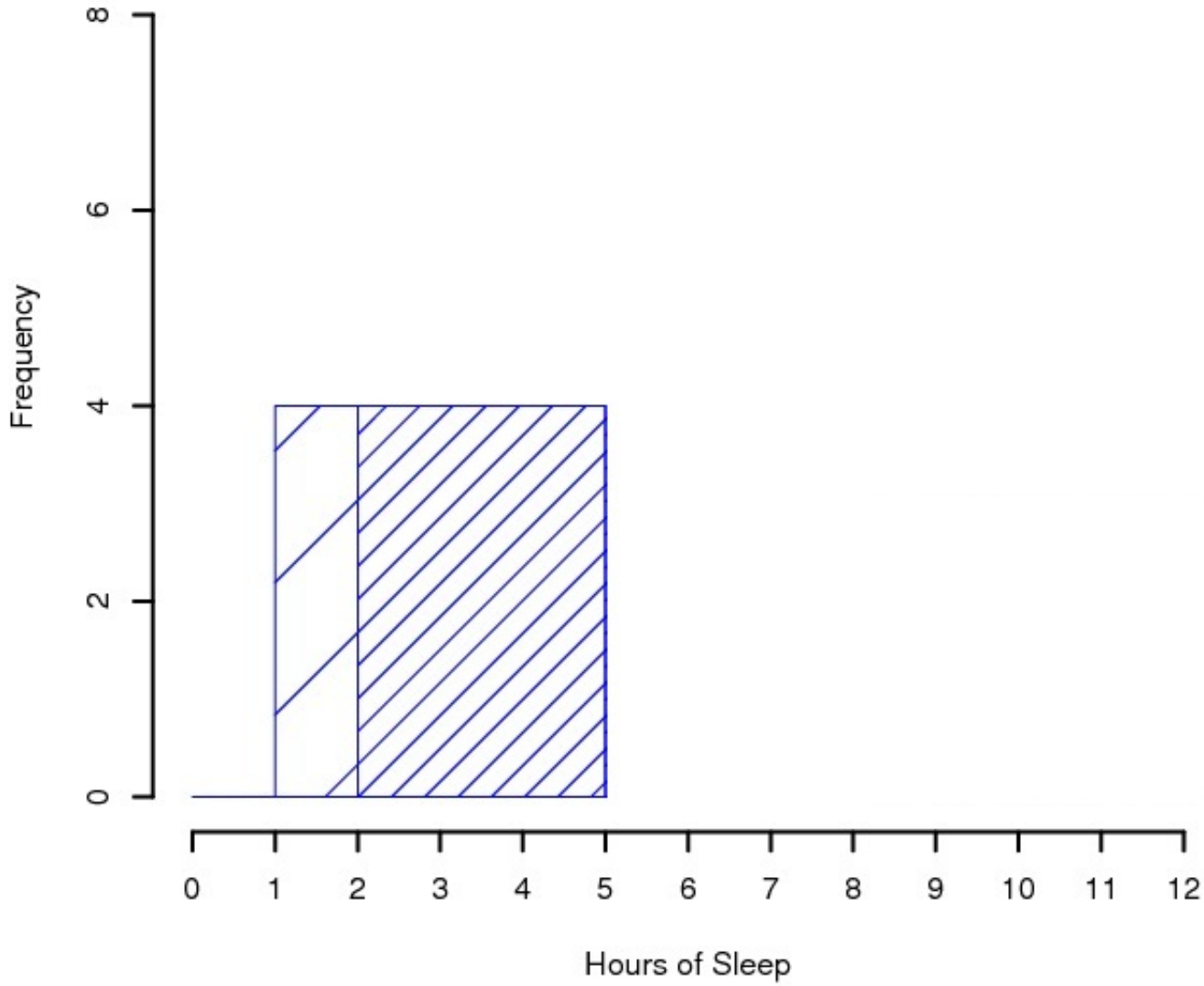
Example: *Hours of sleep*



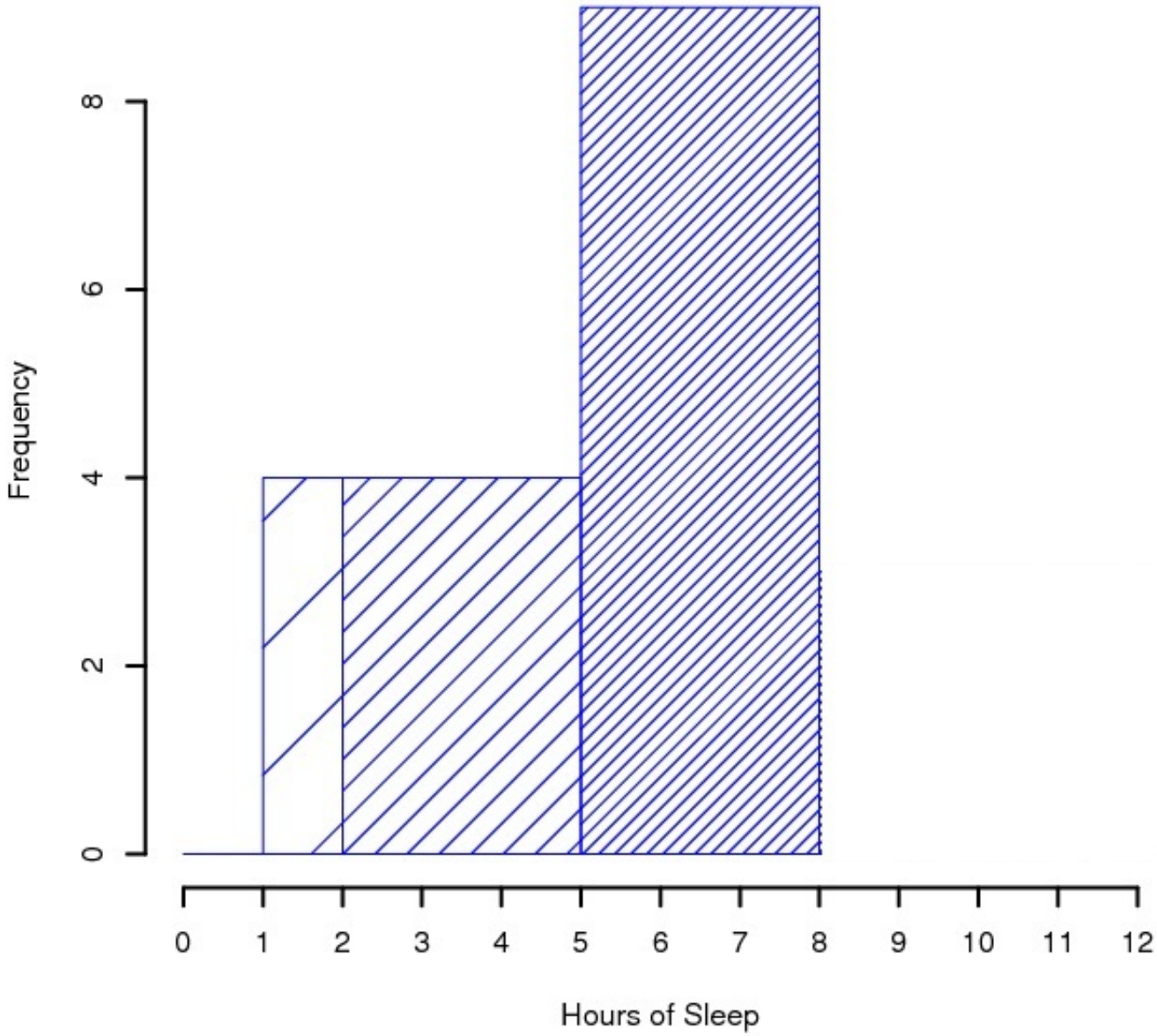
Example: *Hours of sleep*



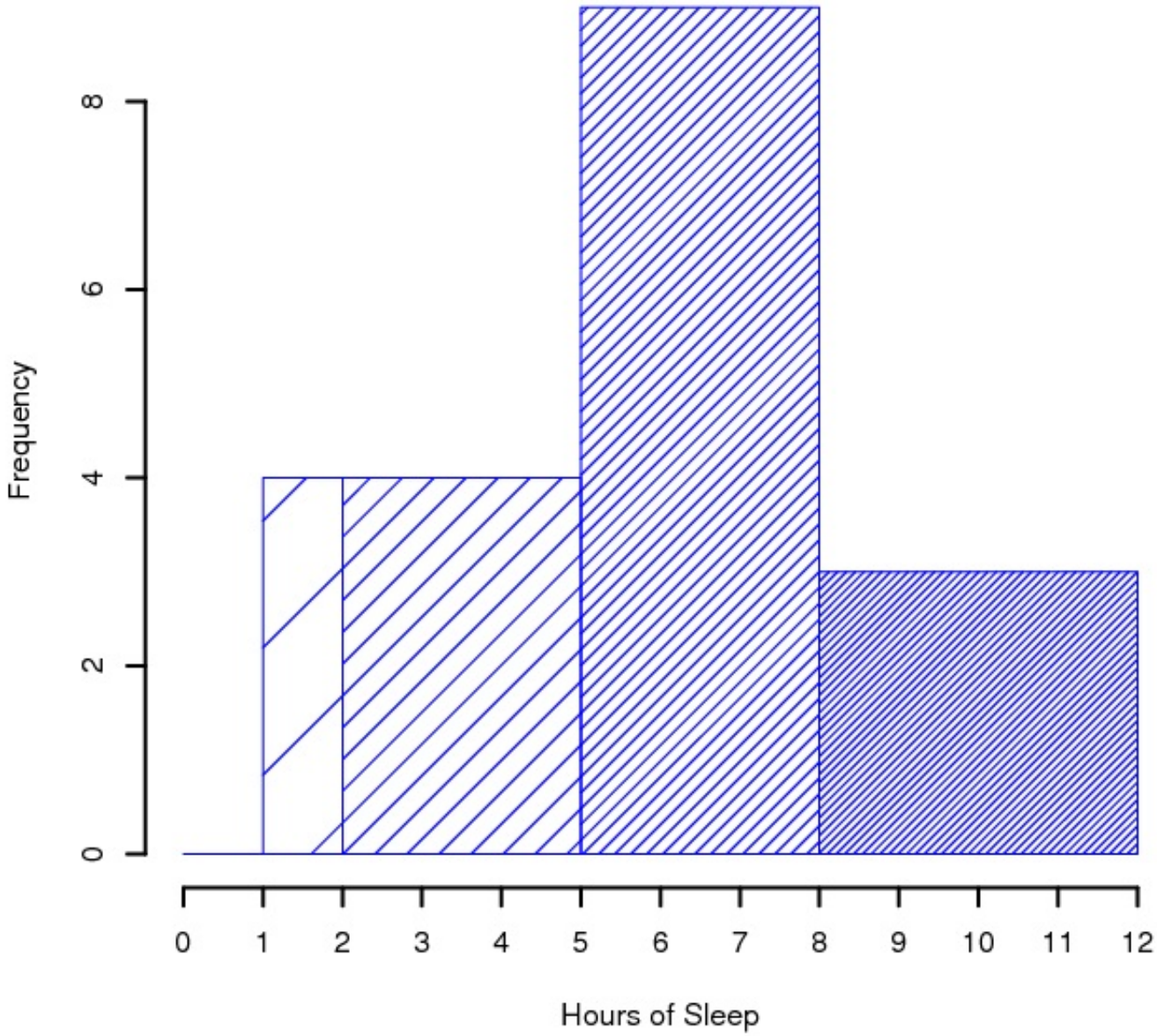
Example: *Hours of sleep*



Example: *Hours of sleep*



Example: *Hours of sleep*



The 13th Biannual Youth Survey on Politics and Public Service
Field Dates: October 28th – November 9th, 2007
Master Questionnaire; N=2,526 18-24 Year Olds

When it comes to most political issues, do you think of yourself as a liberal, moderate or conservative? (If moderate ask: as a moderate, which way do you lean?)

	Total	College	Non-College
Liberal	32%	34%	31%
Moderate leaning liberal	14%	18%	13%
Moderate	21%	17%	23%
Moderate leaning conservative	12%	12%	12%
Conservative	21%	19%	22%

Variables

- Variable: the aspect that differs from subject to subject, individual to individual.
  - Political learning, Age, Sex, Income,....
- Data: the value of the variables
  - Conservative, 19, Male, $15,000, ...

Two types of variables

- Quantitative or numerical variables
  - Numbers, measurements
  - Age, height, miles traveled, hours slept, income
- Categorical variables
  - Classify each observation
  - Political Affiliation, sex, race

Turn on your Clickers!

A survey of college students collected information on several variables: Distance from home, Age, Major, Gender and Class.

The variable *Major* is:

(a) Quantitative

(b) Categorical

(c) Neither categorical nor numeric

Turn on your Clickers!

A survey of college students collected information on several variables: Distance from home, Age, Major, Gender and Class.

The variable *Distance from home* is:

(a) Quantitative

(b) Categorical

(c) Neither categorical nor numeric

Categorical Data

- How do we describe it?
- How do we graph it?

Summary of Categories

Sample Proportion

- Counts (Each category has a number of occurrences.)
- Proportions/Percentages

	Total
Liberal	32%
Moderate leaning liberal	14%
Moderate	21%
Moderate leaning conservative	12%
Conservative	21%

Sample Proportion

For example...

$$\hat{p}_{\text{Liberal}} = \frac{\text{\# of people being Liberals}}{\text{total \# of respondents}} = 0.32$$

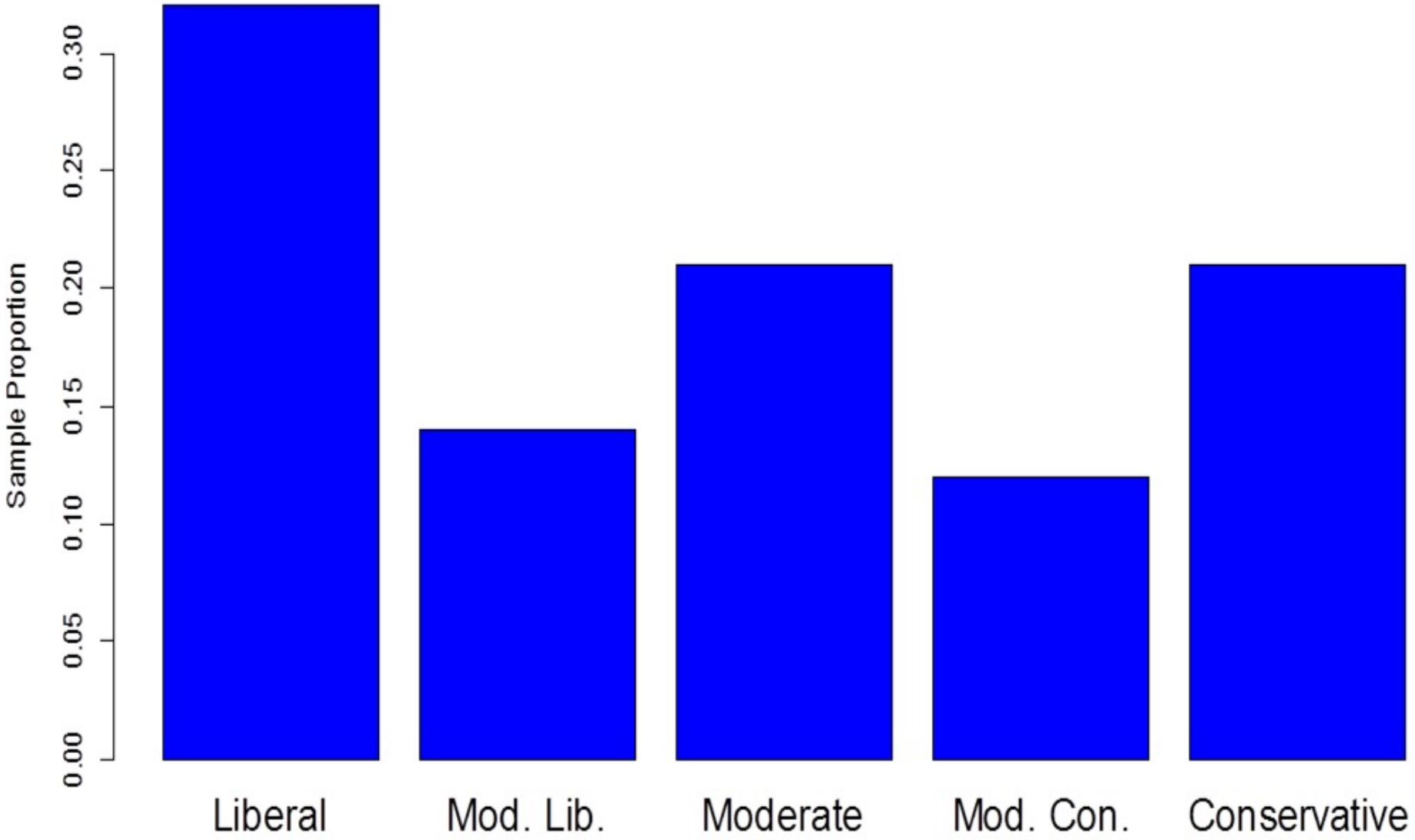
$$\hat{p}_{\text{Moderate}} = \frac{\text{\# of people being Moderate}}{\text{total \# of respondents}} = 0.21$$

$$\hat{p}_{\text{Conservative}} = \frac{\text{\# of people being Conservative}}{\text{total \# of respondents}} = 0.21$$

Visualizing Categorical Data

- Give a clear picture of what the data contains
- Emphasize differences/similarities
- Bar graphs are usually the best

Bar Graph


Sample Proportion
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Liberal
Mod. Lib.
Moderate
Mod. Con.
Conservative

Summary Statistics

Type of Variable	***Statistics***	***Graphs***
Categorical	Proportions	Bar Graph

Summary Statistics

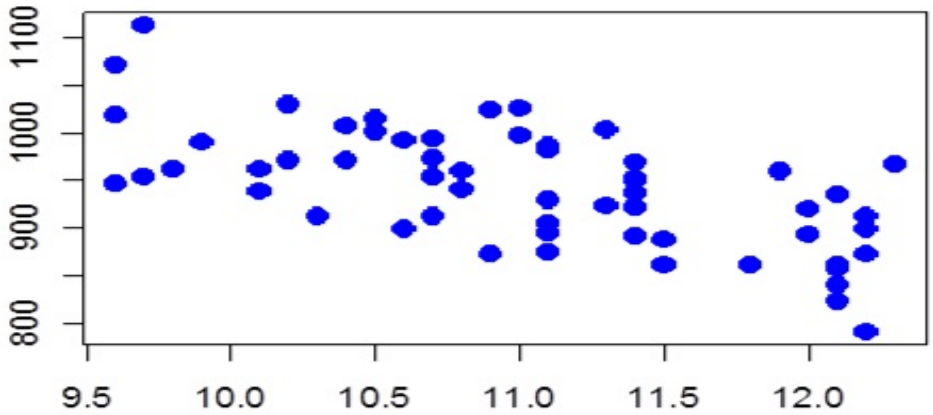
Type of Variable	***Statistics***	***Graphs***
Quantitative Continuous	Mean, St. Deviation Median, IQR 5-Number Summary	Histogram Boxplot

Pair of Measurements

- Two quantitative measurements
- What is their relationship?
- Can we predict one value from the other?

Correlation

- Response and Explanatory variables
- Scatter plot



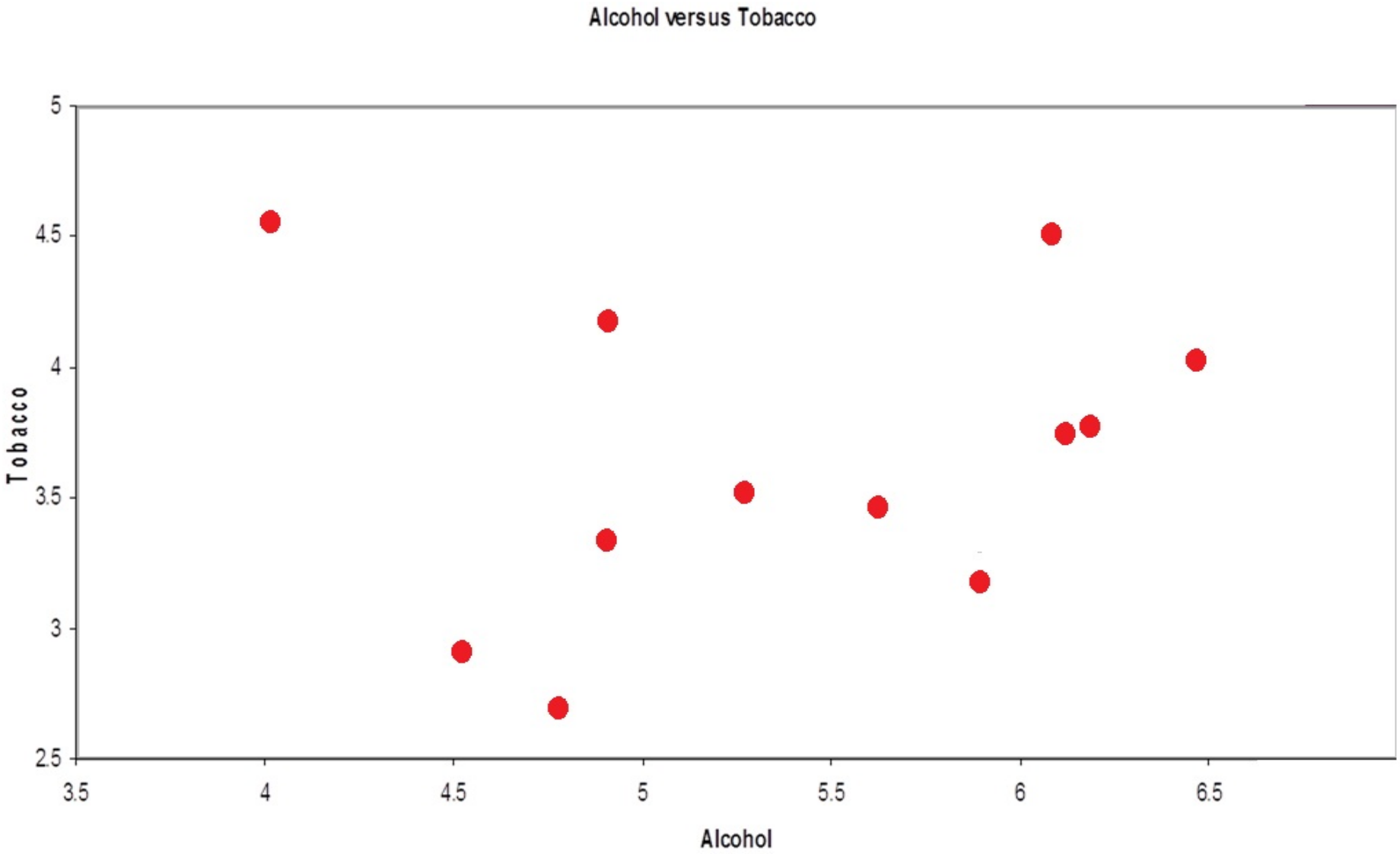
- Positive and Negative Association
- Sample Correlation

$$r = \frac{1}{(n-1)\, s_x\, s_y} \left[\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \right]$$

Example: Average Consumption per Capita by Region in UK

Region	Alcohol	Tobacco
North	6.47	4.03
Yorkshire	6.13	3.76
Northeast	6.19	3.77
East Midlands	4.89	3.34
West Midlands	5.63	3.47
East Anglia	4.52	2.92
Southeast	5.89	3.2
Southwest	4.79	2.71
Wales	5.27	3.53
Scotland	6.08	4.51
Northern Ireland	4.02	4.56

Scatter Plot



Two variables

- Explanatory Variable
  - Input into the system
  - Explains or predicts the other
- Response Variable
  - Output of the system
  - What we want to predict

Turn on your clickers!

A researcher would like to know if mother's height can explain how tall her child will be. Which is the response variable?

(a) Child's height

(b) Mother's height

(c) Father's height

Turn on your clickers!

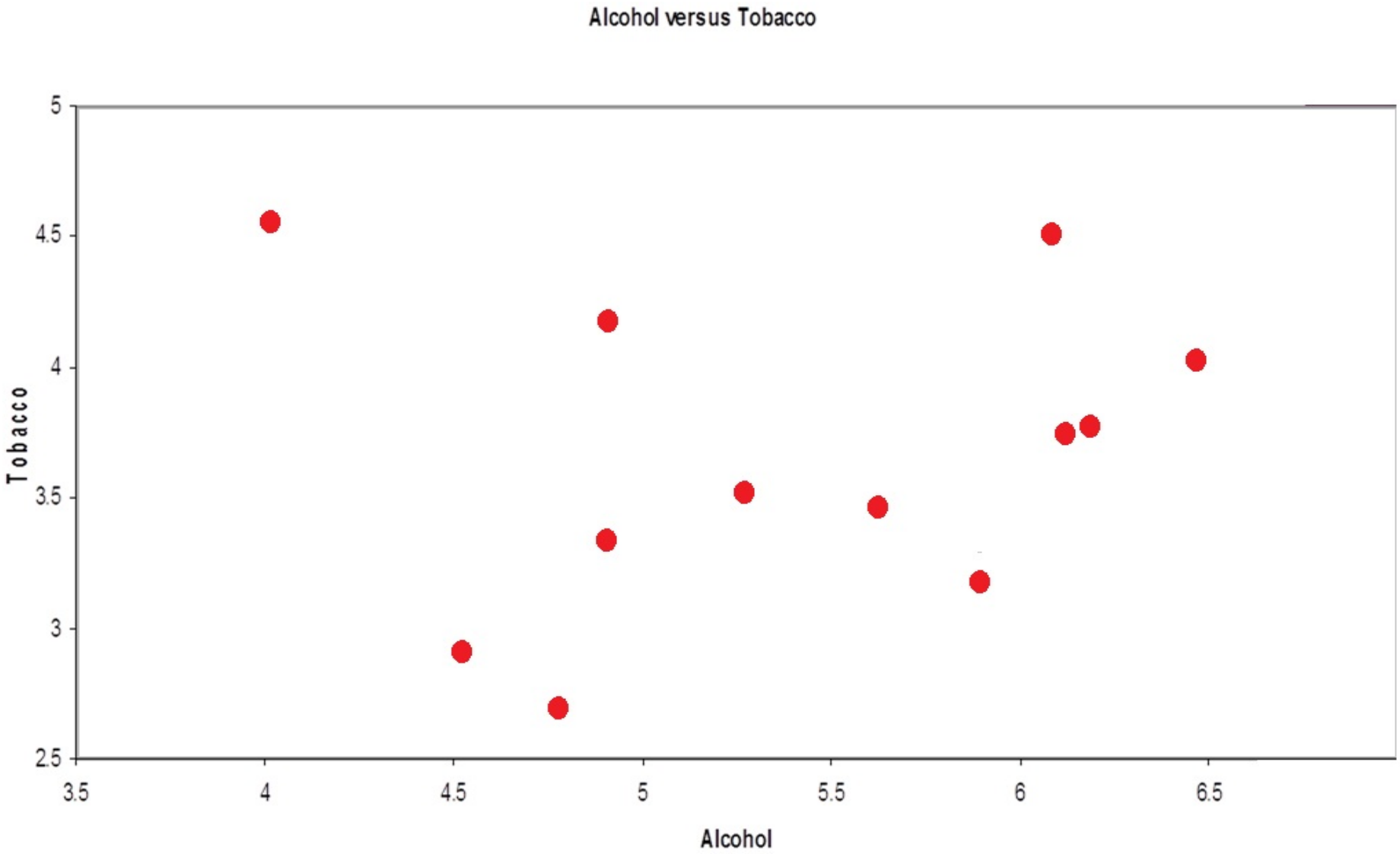
A researcher would like to know if mother's height can explain how tall her child will be. Which is the response variable?

(a) Child's height

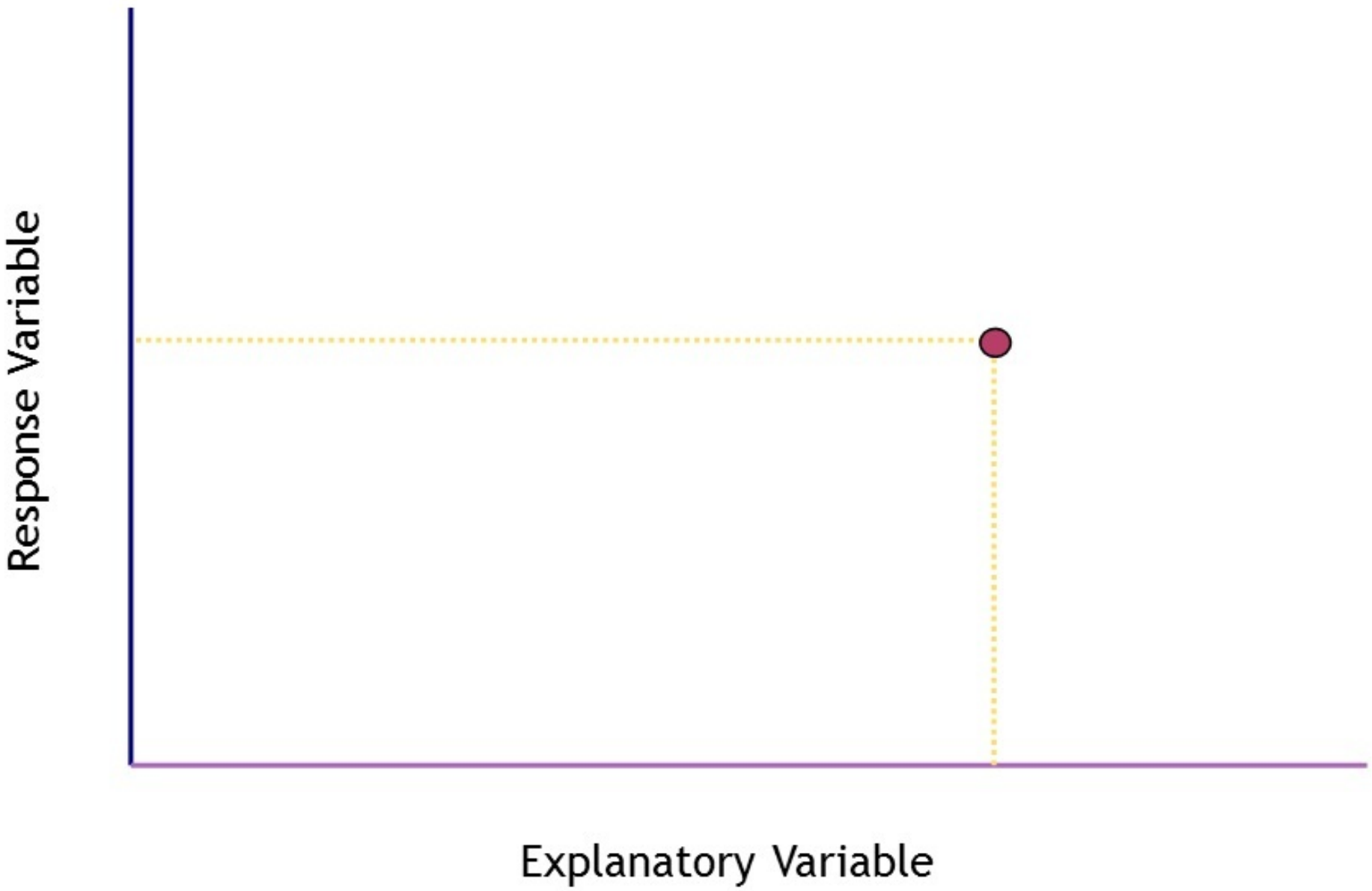
(b) Mother's height

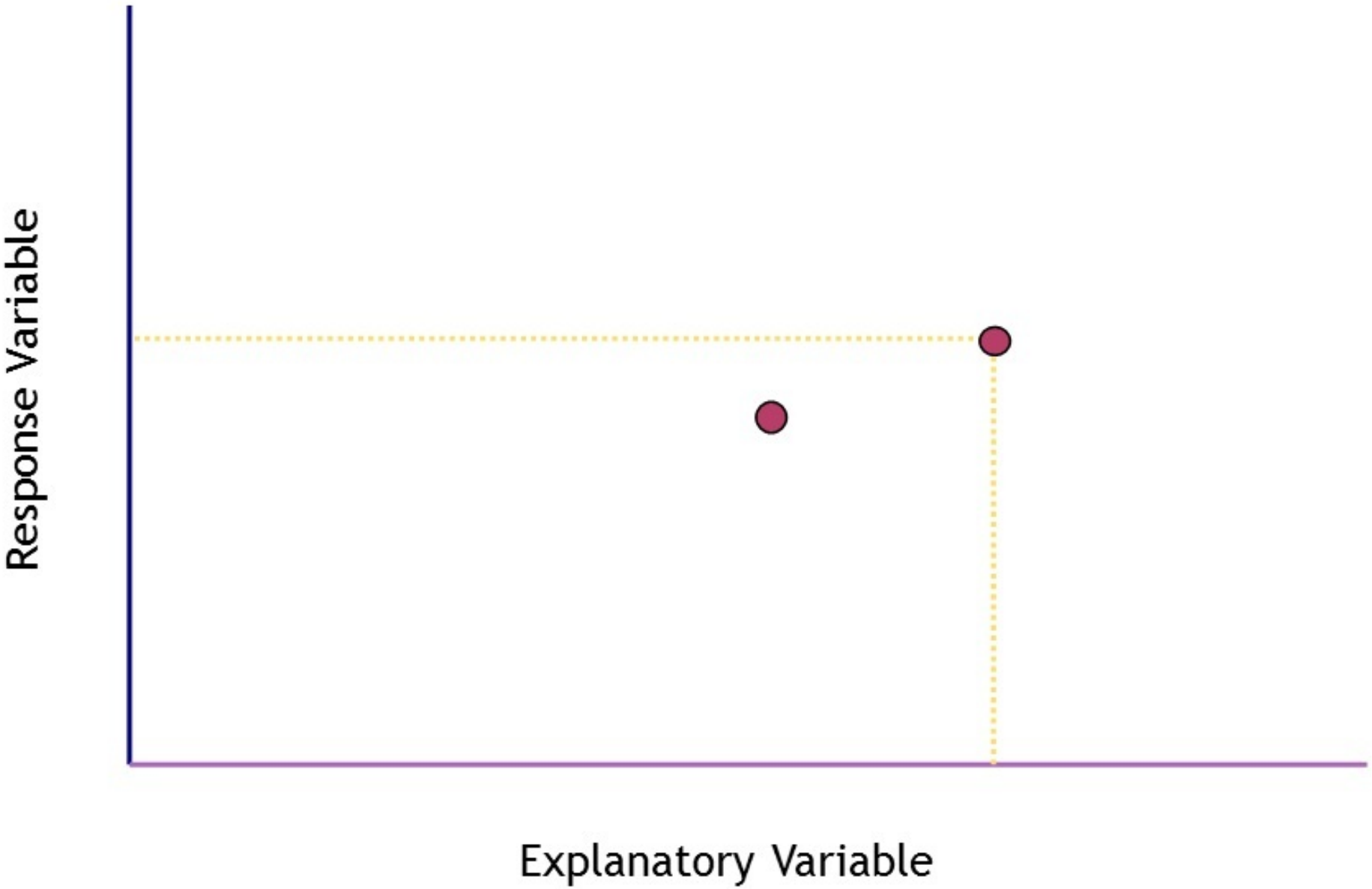
(c) Father's height

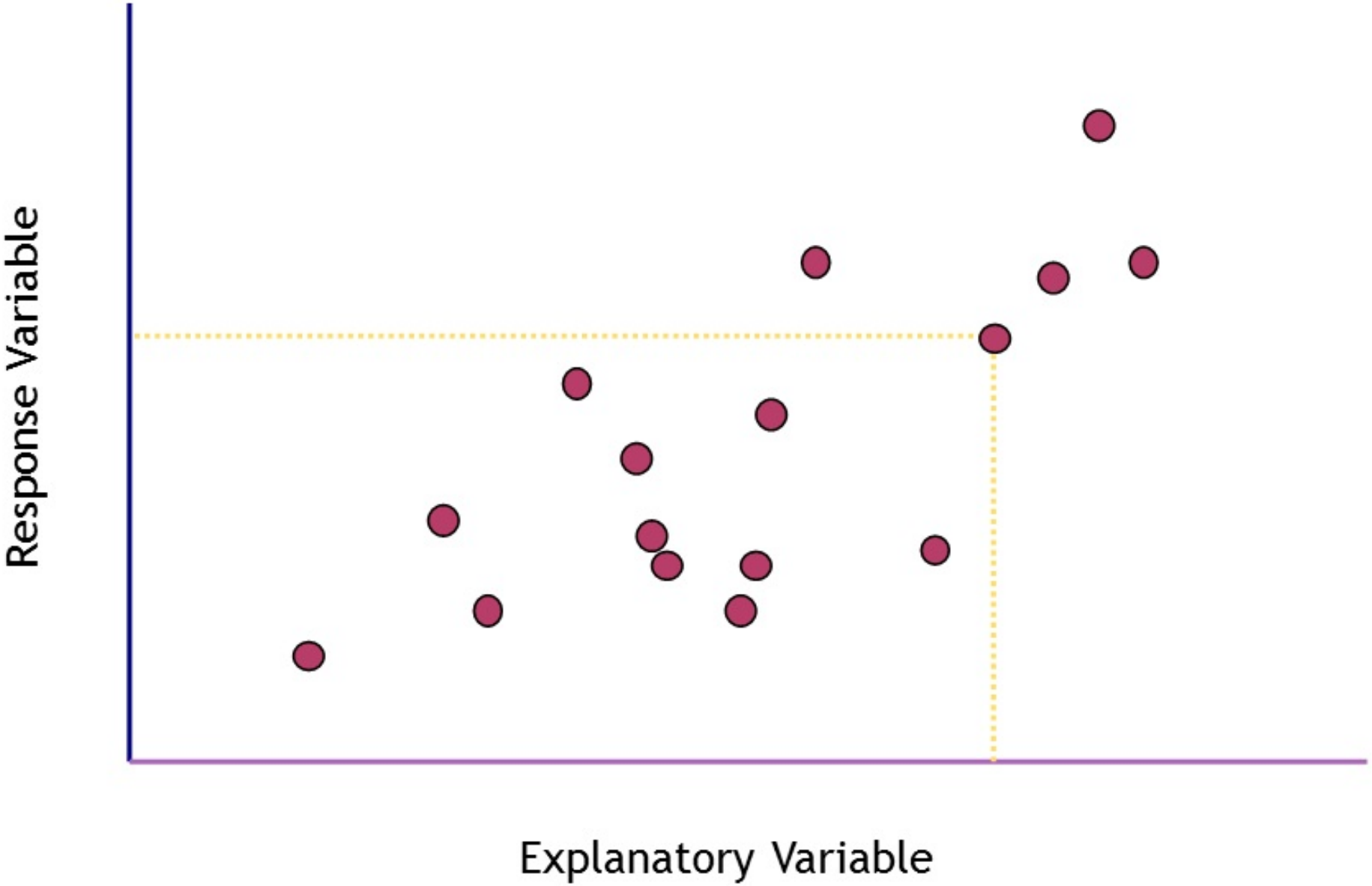
Scatter Plot



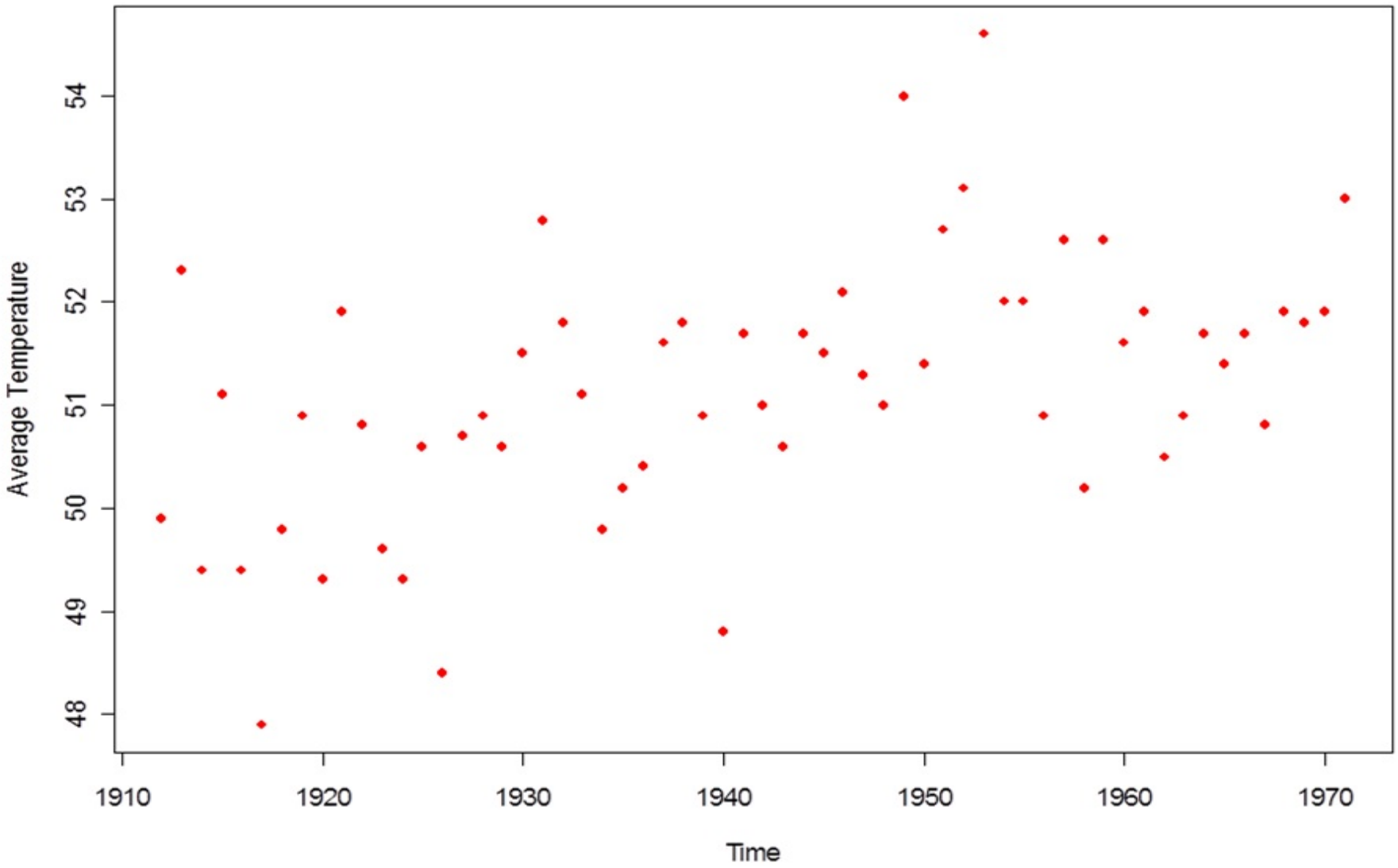
Scatter Plot




Response Variable
Explanatory Variable


Response Variable
Explanatory Variable

Average Temperature through the years



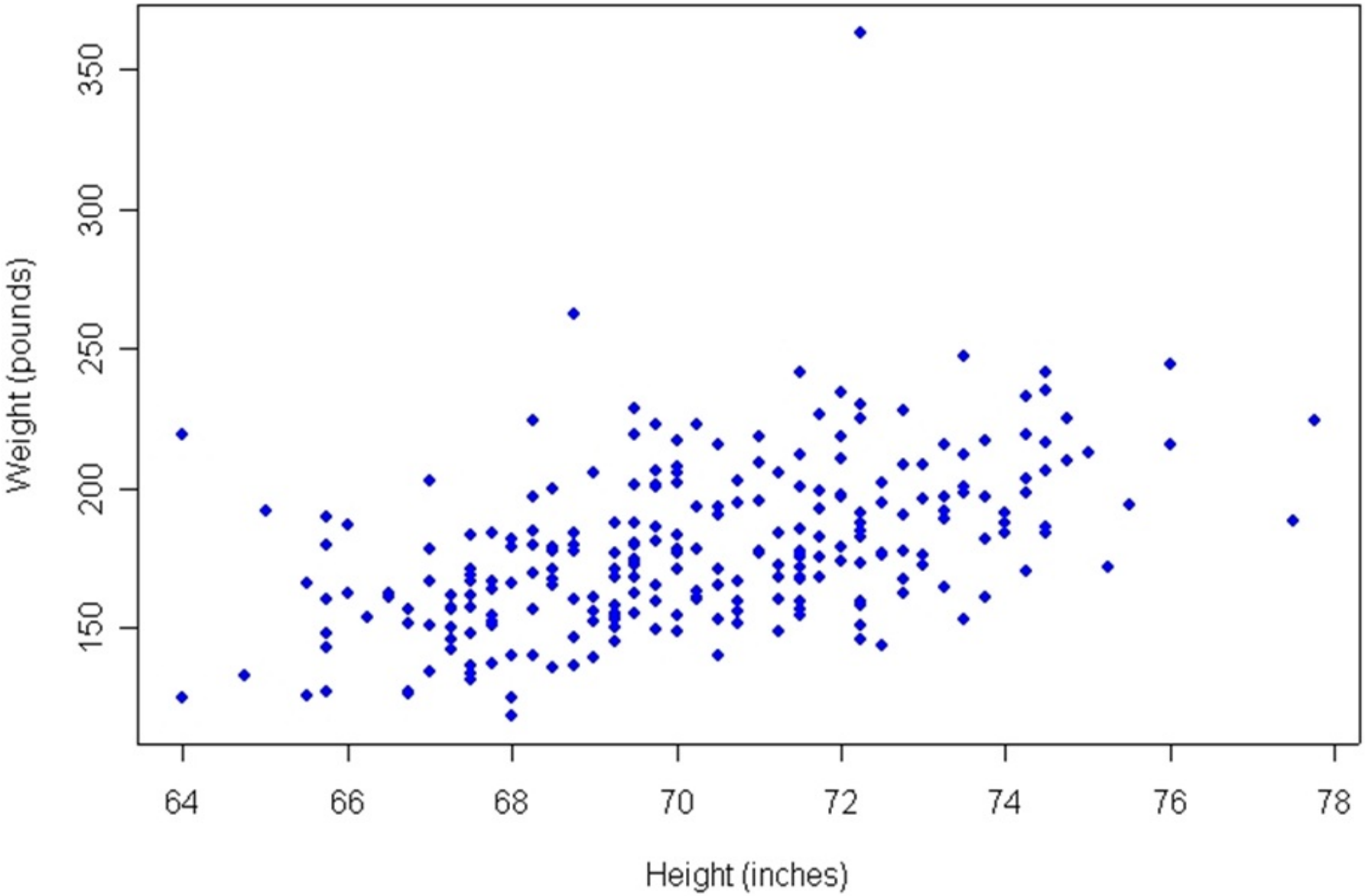
What to look for

- Linear relationship
  - Positive Association
  - Negative Association
- Non-linear relationship
- Outliers

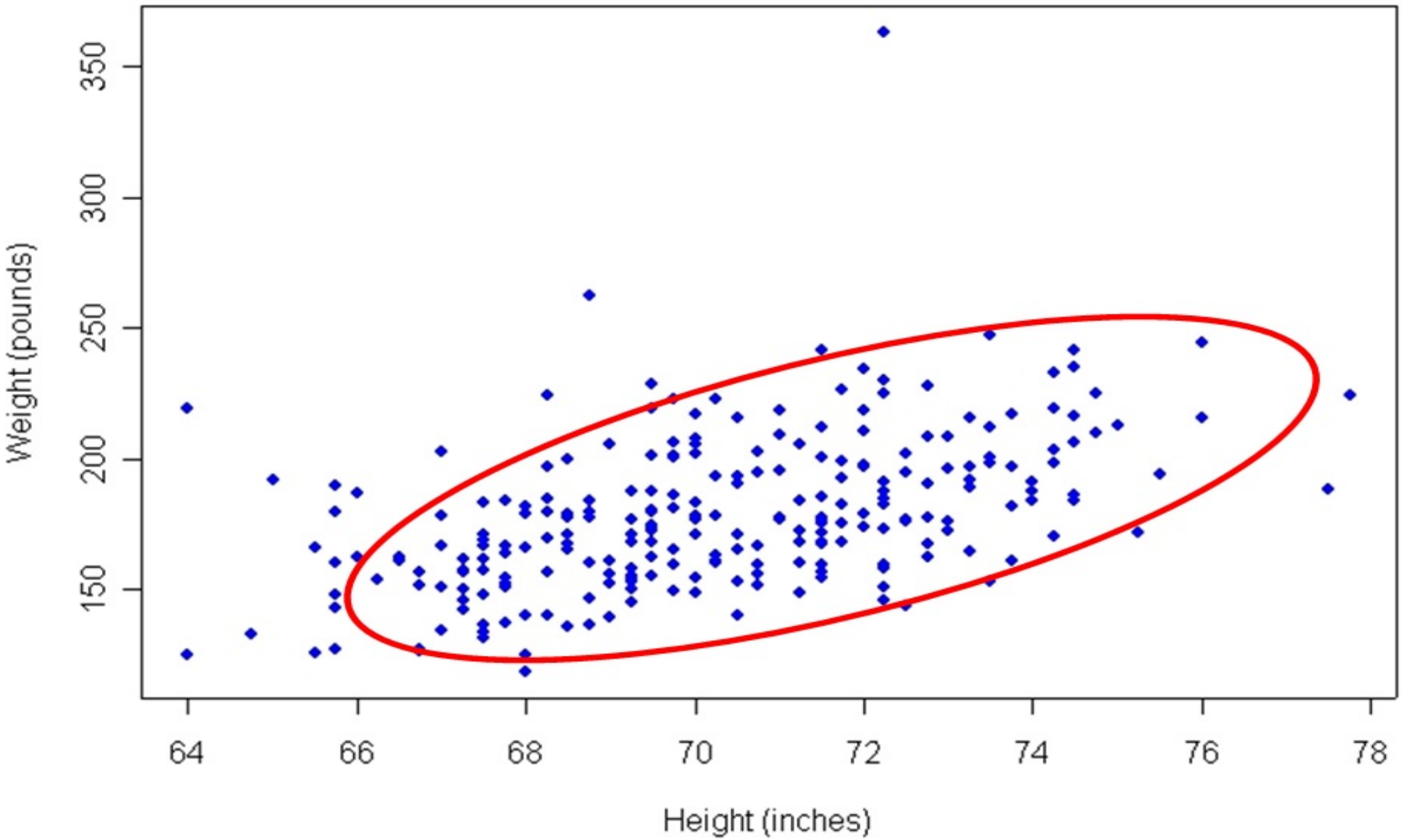
Linear Relationship

- Explanatory variable increases or decreases
- Response increases or decreases proportionally

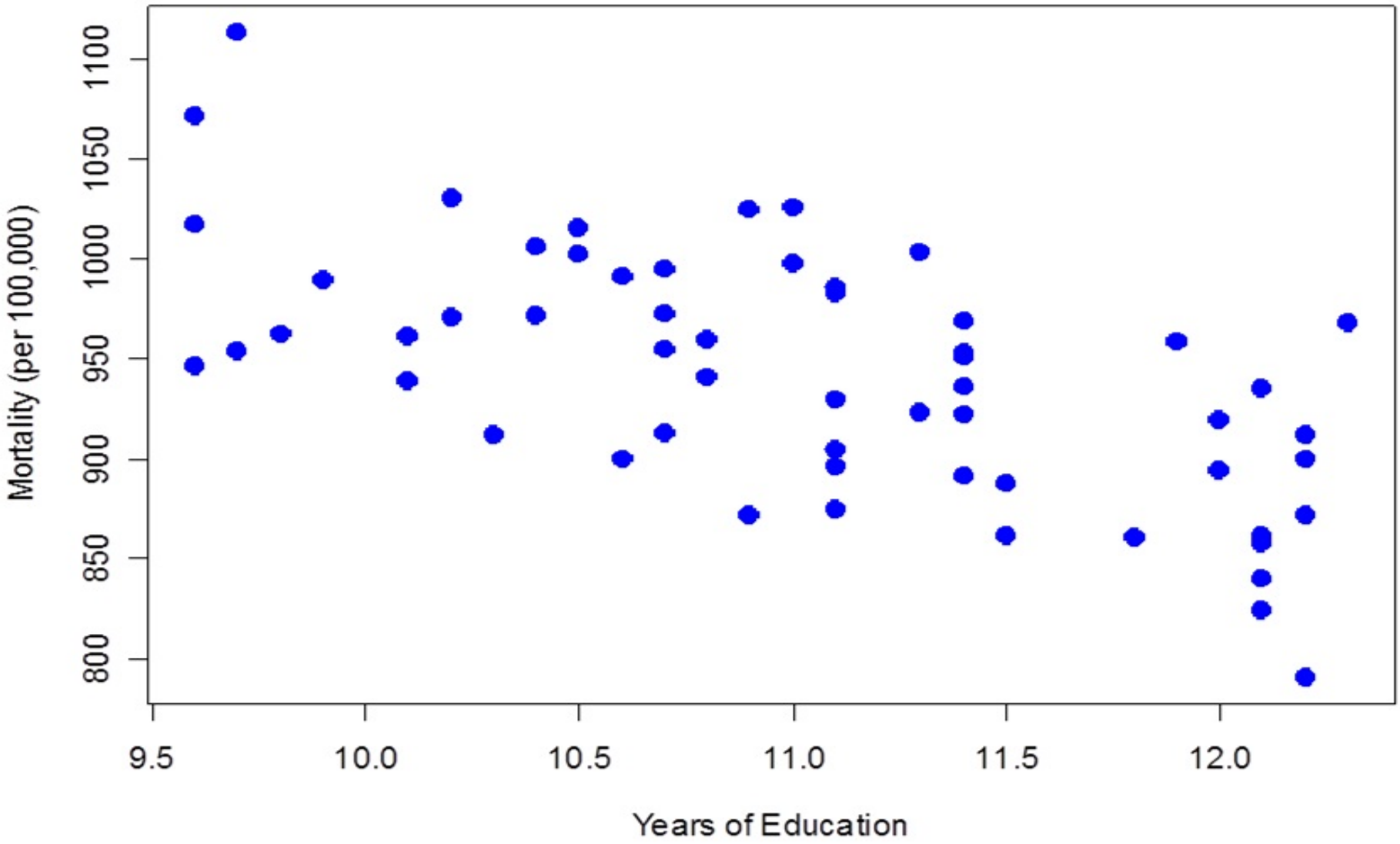
Positive Association



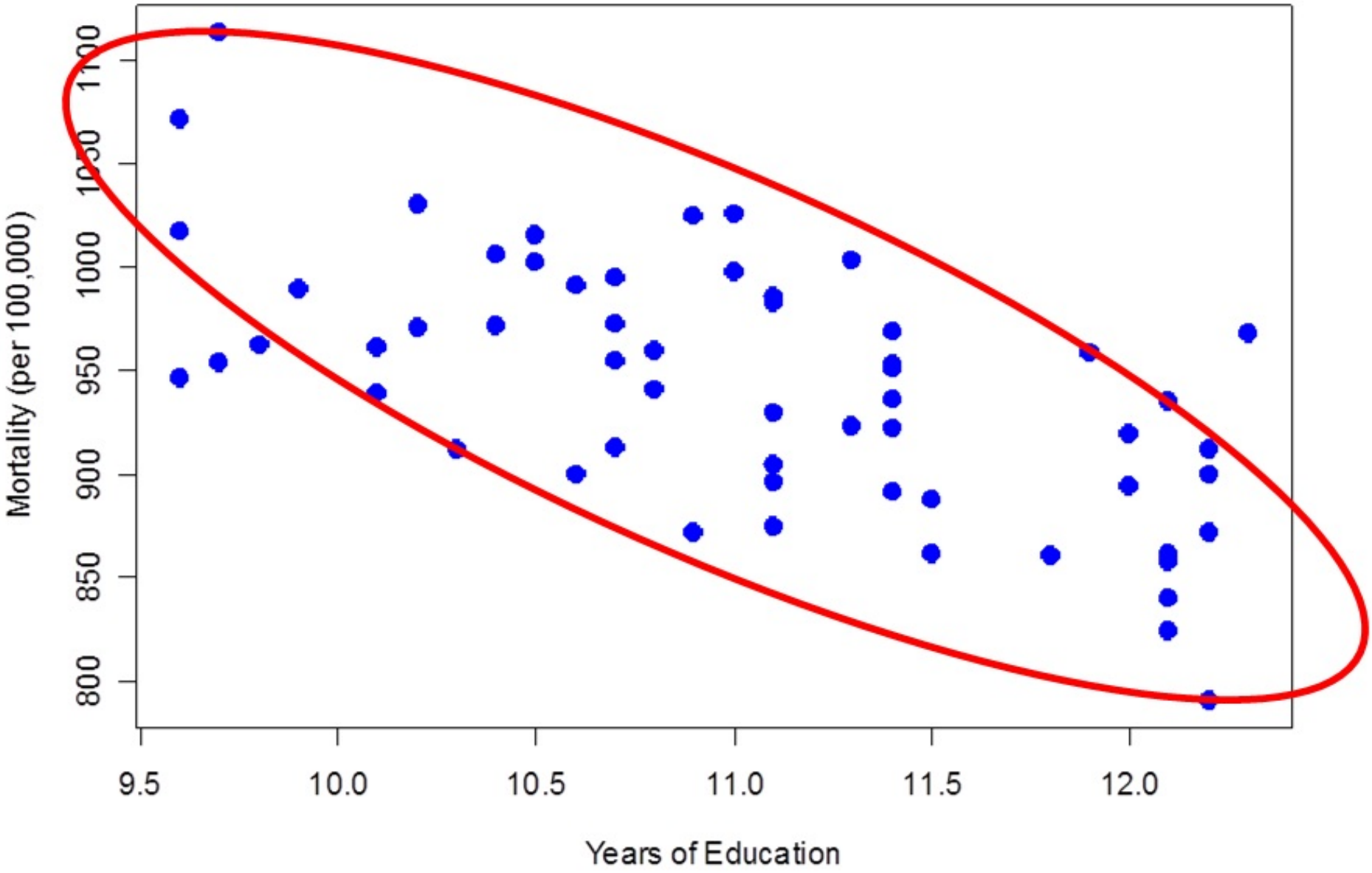
Positive Association



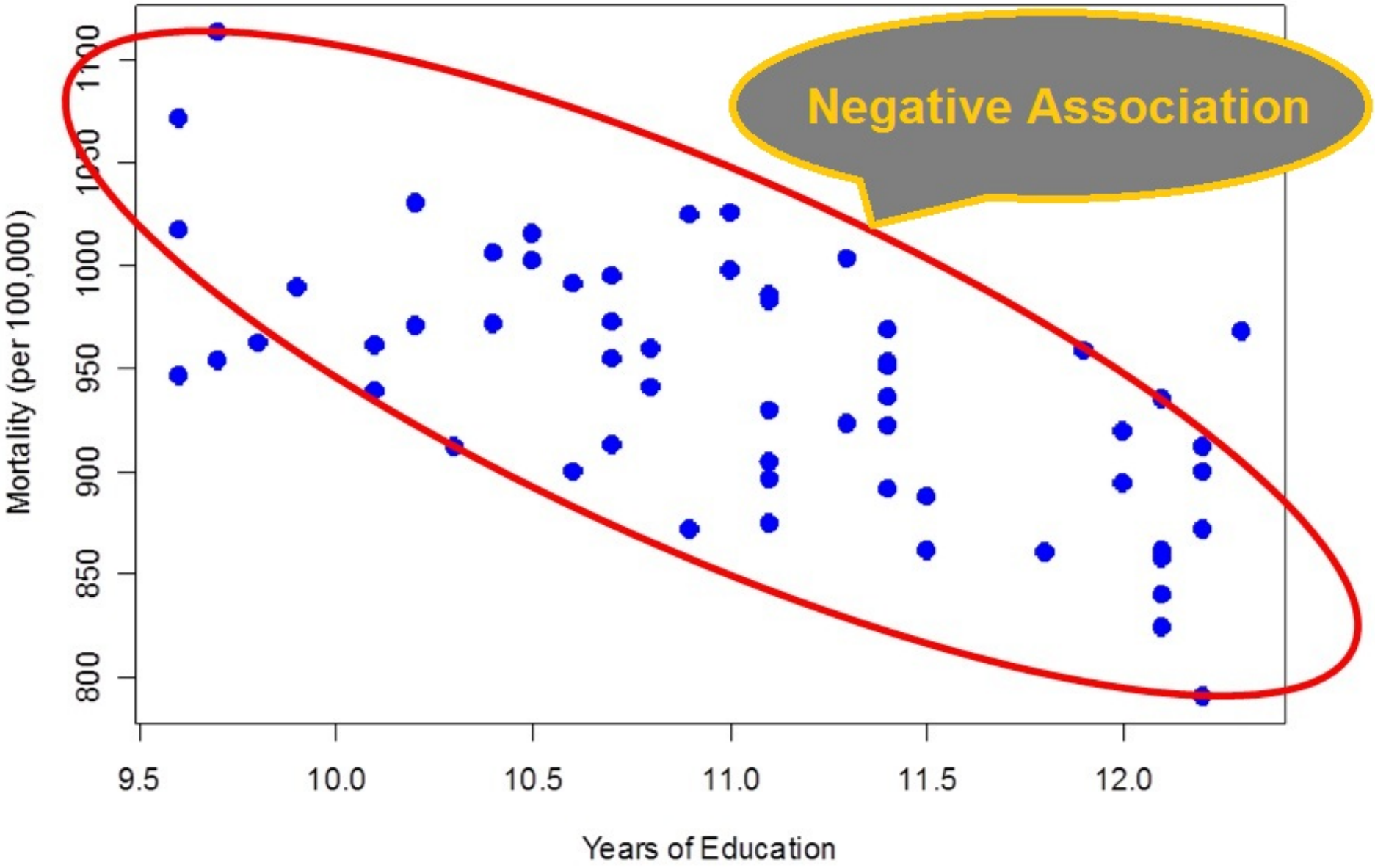
How does education affect health?



How does education affect health?



How does education affect health?



Linear Association

	Explanatory Variable	Response Variable
Positive Association	Increases	Increases
	Decreases	Decreases
Negative Association	Increases	Decreases
	Decreases	Increases