# STAT/SOC/CSSS 221
# Statistical Concepts and Methods for the Social Sciences

## Relationships in Data: A first pass

Christopher Adolph

Department of Political Science

*and*

Center for Statistics and the Social Sciences

University of Washington, Seattle

# Aside on mathematical notation

$\bar{x}$ a "bar" indicates this is the mean of a variable

# Aside on mathematical notation

| | |
|---|---|
| $\bar{x}$ | a “bar” indicates this is the mean of a variable |
| $\lvert x \rvert$ | the absolute value of $x$ (drop any minus signs) |

# Aside on mathematical notation

| | |
|---|---|
| $\bar{x}$ | a "bar" indicates this is the mean of a variable |
| $\lvert x \rvert$ | the absolute value of $x$ (drop any minus signs) |
| $x_i^3$ | sometimes, superscripts tell us to raise a variable to a power; this says raise $x_i$ to the third power |

# Aside on mathematical notation

| | |
|---|---|
| $\bar{x}$ | a "bar" indicates this is the mean of a variable |
| $\lvert x \rvert$ | the absolute value of $x$ (drop any minus signs) |
| $x_i^3$ | sometimes, superscripts tell us to raise a variable to a power; this says raise $x_i$ to the third power |
| $x_i^{\text{label}}$ | other times, a superscript is just a label distinguishing this variable from another $x$ (common when there is already an index as a subscript, so we need a different place to put our label) |

# Assessing relationships between variables

Last week, we focused on variation within variables

But most of statistics is concerned with relationships between variables

Most important question: Does variation in $X$ cause variation in $Y$?

Hard question we won't tackle today

Instead, when $X$ varies, do we consistently see similar variation in $Y$?

That is, are $X$ and $Y$ correlated?

# The right tool for the job

This week, we introduce basic tools for understanding correlation

The right tool for our data depends on the order of measurement of the "dependent variable" and the covariate

"Dependent variable", "response variable", & "outcome variable" are synonyms

If outcome is continuous and the covariate is discrete, consider box plots

If both are continuous, consider scatterplots

If both are discrete, consider a contingency table ("cross-tabulation")

# Outline

Comparing two samples with box plots
*Example: GDP and partisan government*

Exploring continuous relationships with scatterplots
*Examples: Height and Weight of 20-year old males; Challenger Launch Decision*

Best fit lines for scattterplots
*Example: Cross-national fertility*

Relationships between ordered variables in tables
*Example: Voting and Education*

# Naïve use of these methods may produce misleading results

Three most important reasons:

**Confounders** If we think $X$ causes $Y$, but we have left out the *real* causal variable $Z$, we could be mislead by this confounding factor.

## Naïve use of these methods may produce misleading results

Three most important reasons:

**Confounders** If we think $X$ causes $Y$, but we have left out the *real* causal variable $Z$, we could be mislead by this confounding factor.

**Sampling Error** Small samples may create a misleading impression of the relation between $X$ and $Y$

## Naïve use of these methods may produce misleading results

Three most important reasons:

**Confounders** If we think $X$ causes $Y$, but we have left out the *real* causal variable $Z$, we could be mislead by this confounding factor.

**Sampling Error** Small samples may create a misleading impression of the relation between $X$ and $Y$

**Correlation does not always imply causation** If $X$ and $Y$ are correlated, either $X$ may cause $Y$, or $Y$ may cause $X$, or both, or *neither*

## Example 1: US Economic growth

Let's investigate an old question in political economy:

Are there partisan cycles, or tendencies, in economic performance?

Does one party tend to produce higher growth on average?

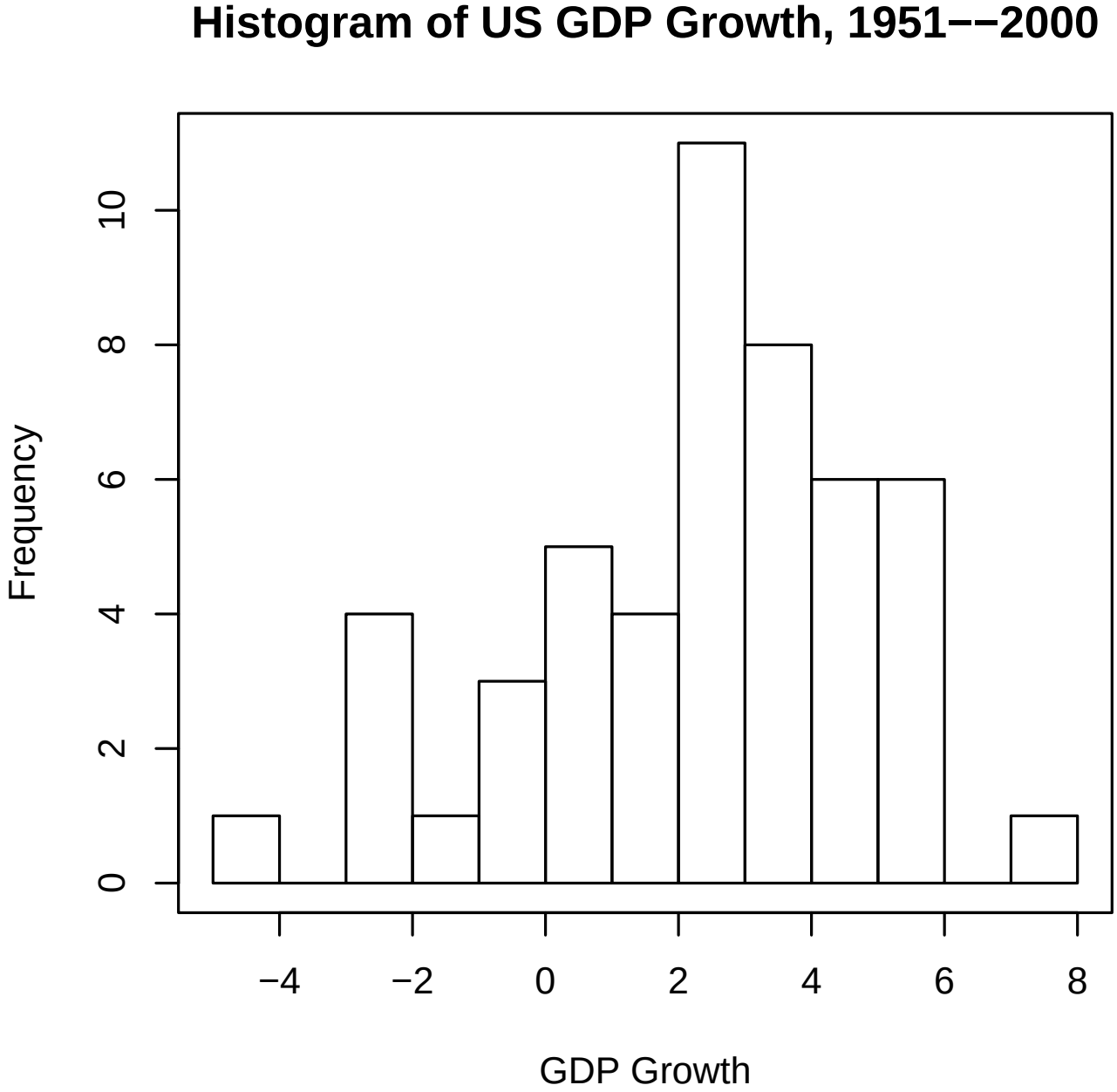
(Theory: Left cares more about growth vis-à-vis inflation than the Right

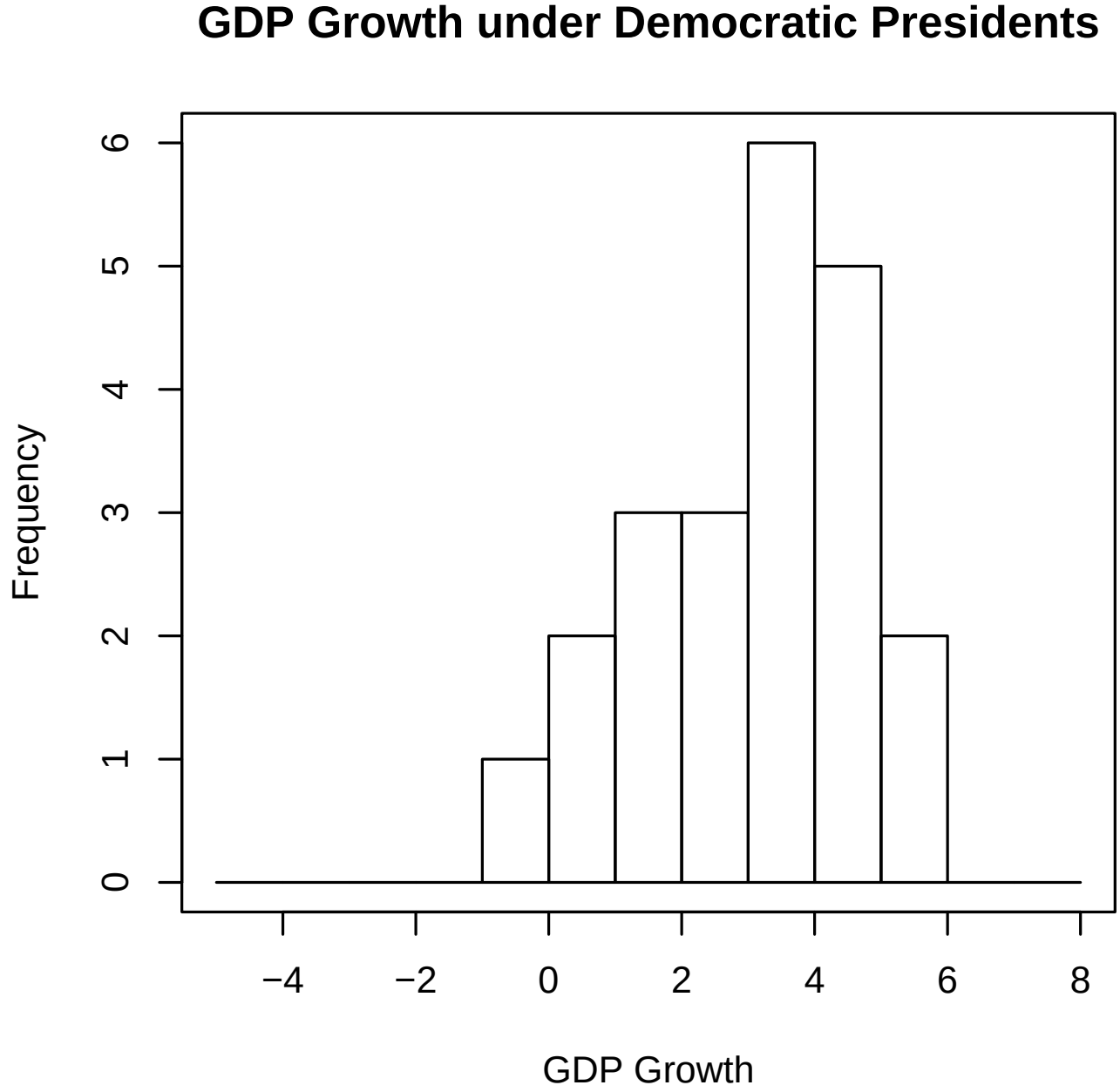
If there is partisan control of the economy,
then Left should have higher growth all else equal)

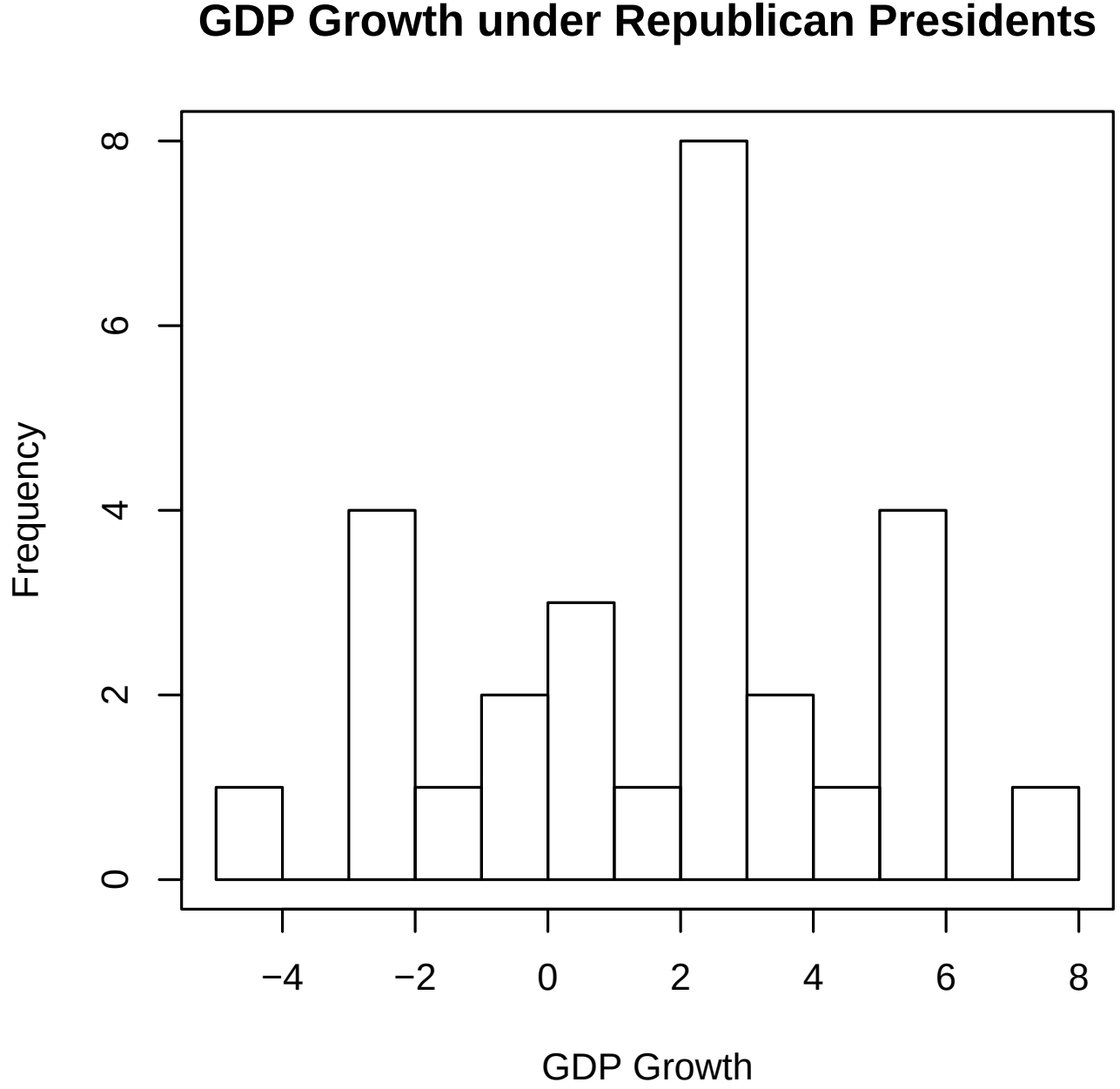
Data from the Penn World Tables (Annual growth rate of GDP in percent)

Two variables:

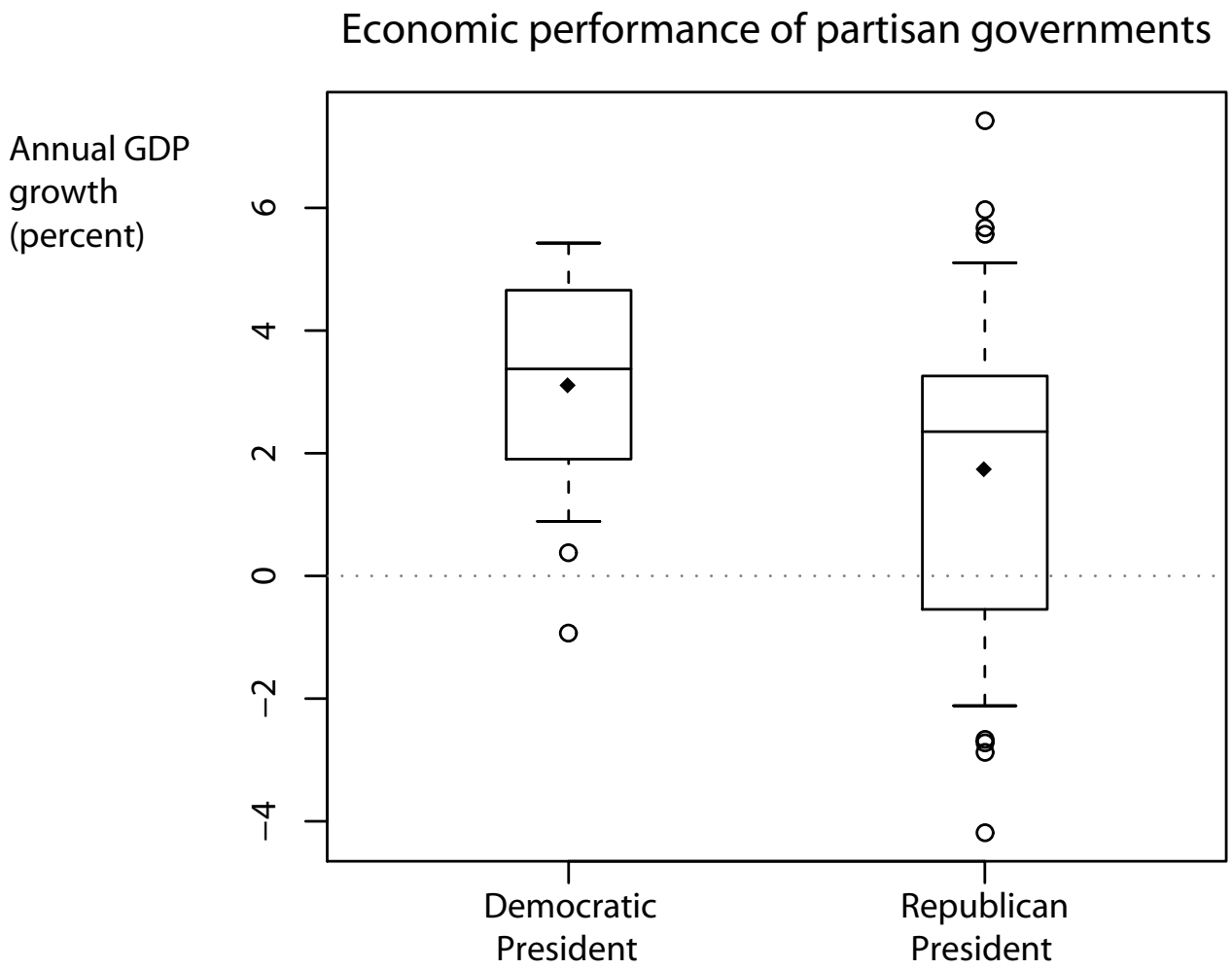
| | |
|---|---|
| GDP Growth | The per capita GDP growth rate |
| Party | The party of the president (Democrat or Republican) |


Histogram of US GDP Growth, 1951−−2000
Frequency
0
2
4
6
8
10
−4
−2
0
2
4
6
8
GDP Growth

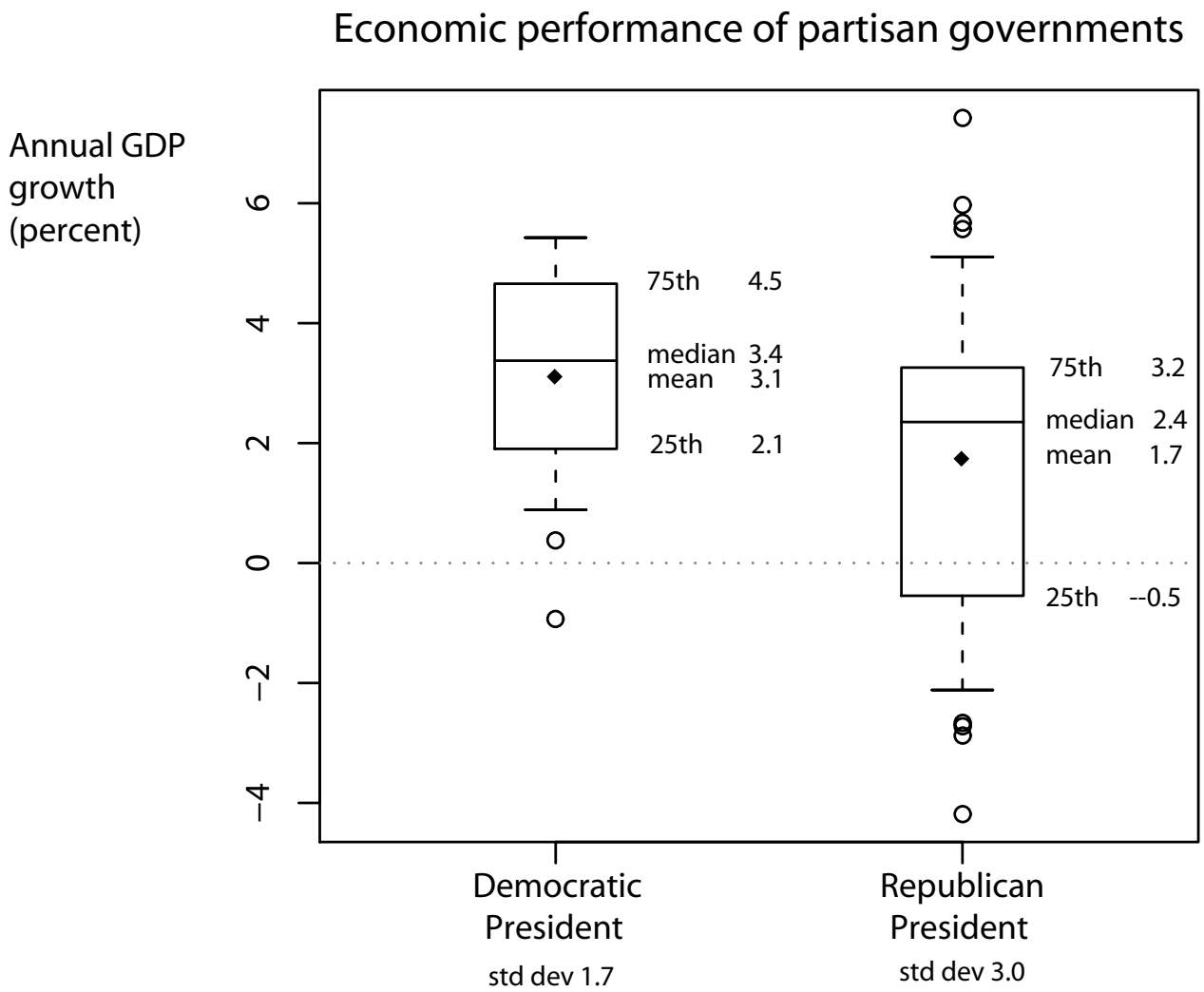

GDP Growth under Democratic Presidents
Frequency
0
1
2
3
4
5
6
−4
−2
0
2
4
6
8
GDP Growth


GDP Growth under Republican Presidents
Frequency
0
2
4
6
8
−4
−2
0
2
4
6
8
GDP Growth

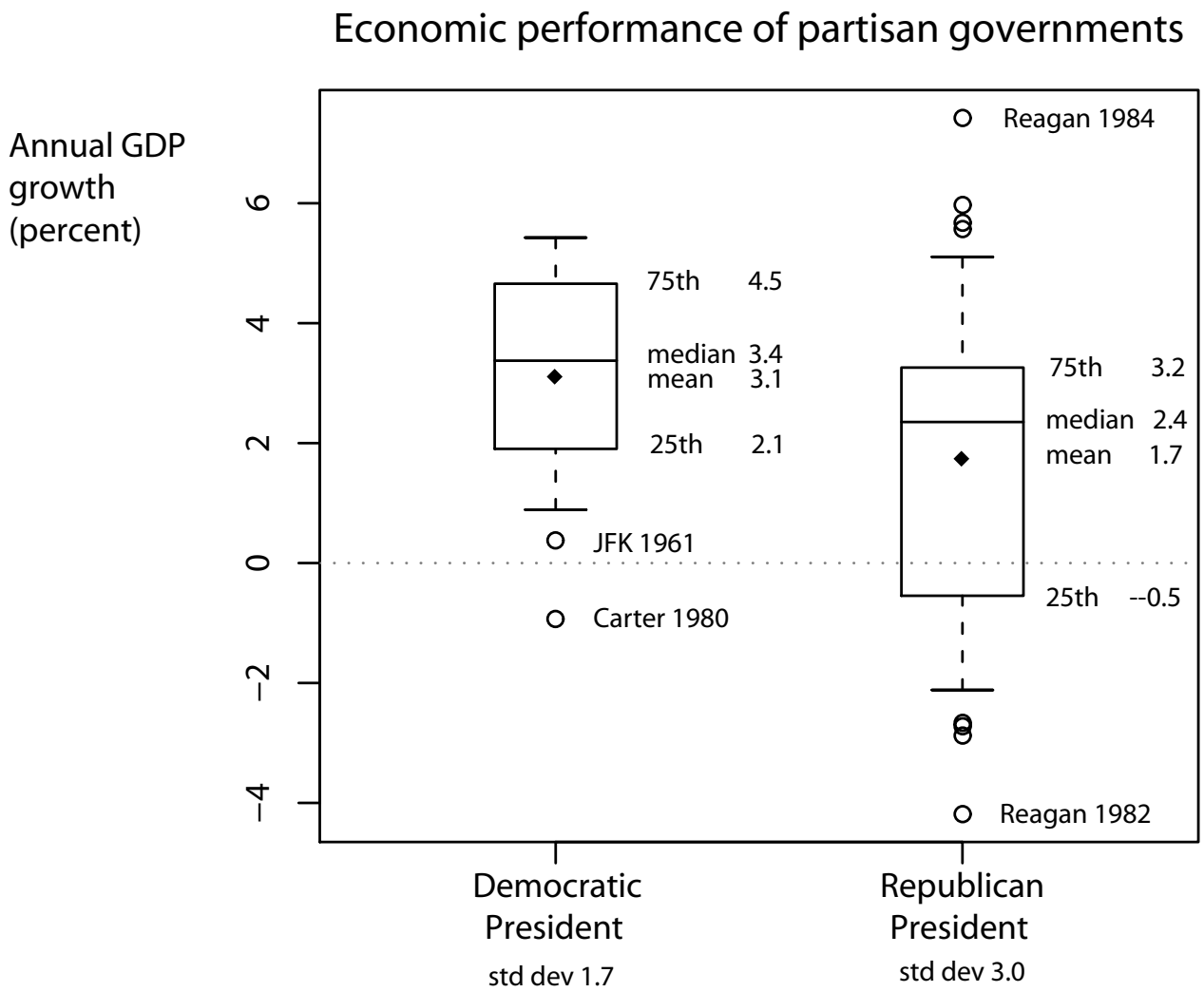
# Box plots: Annual US GDP growth, 1951–2000



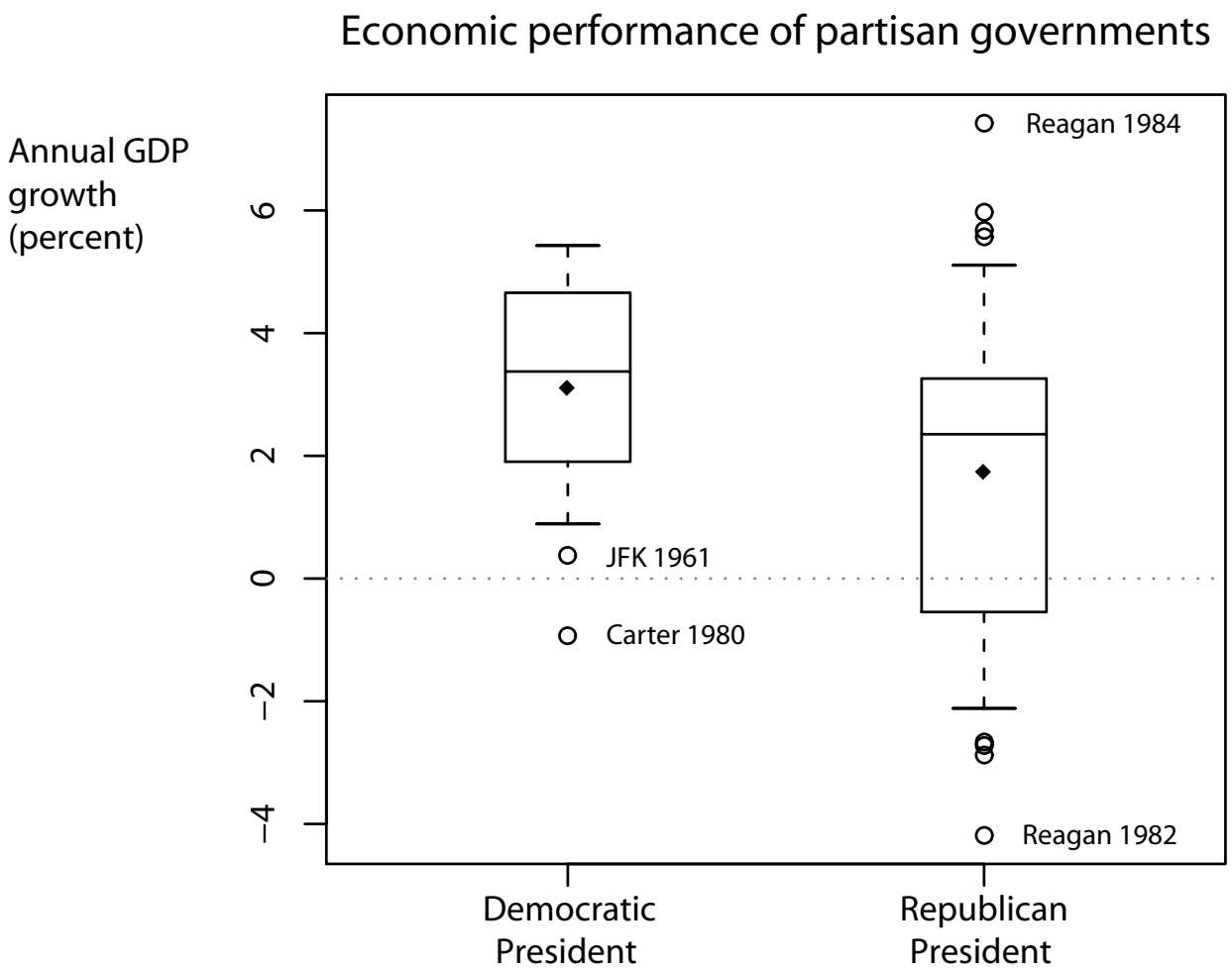
# Box plots: Annual US GDP growth, 1951–2000



# Box plots: Annual US GDP growth, 1951–2000



# Box plots: Annual US GDP growth, 1951–2000



# GDP and Partisan Government

Are you persuaded by this analysis? How might it have gone wrong?

# GDP and Partisan Government

Are you persuaded by this analysis? How might it have gone wrong?

**Confounders** What if other factors, omitted from the analysis, really drive growth? (Partisan control of Congress, or international economic conditions, or the past party in power)

# GDP and Partisan Government

Are you persuaded by this analysis? How might it have gone wrong?

**Confounders** What if other factors, omitted from the analysis, really drive growth? (Partisan control of Congress, or international economic conditions, or the past party in power)

**Sample Error** What if we just don't have enough data to determine the relationship?

# GDP and Partisan Government

Are you persuaded by this analysis? How might it have gone wrong?

**Confounders** What if other factors, omitted from the analysis, really drive growth? (Partisan control of Congress, or international economic conditions, or the past party in power)

**Sample Error** What if we just don't have enough data to determine the relationship?

**Causation** Could we have the direction of the causal arrow wrong? What if voters prefer Democrats when the economy is strong, and Republicans when it is weak?

We haven't introduced the tools to solve these problems yet
we will need to learn some probability first (middle of qtr)

# Stochastic and deterministic relationships

Some relationships are **deterministic**

They always work, without any error, noise, or surprises

1. $2 + 2 = 4$, always. Mathematical laws are deterministic
2. The fruit of a peach tree is always a peach, not an orange.

# Stochastic and deterministic relationships

Some relationships are **deterministic**

They always work, without any error, noise, or surprises

1. $2 + 2 = 4$, always. Mathematical laws are deterministic
2. The fruit of a peach tree is always a peach, not an orange. (But maybe a mutant peach tree could make something new?)
3. Opening the refrigerator turns on a light.

## Stochastic and deterministic relationships

Some relationships are **deterministic**

They always work, without any error, noise, or surprises

1. $2 + 2 = 4$, always. Mathematical laws are deterministic
2. The fruit of a peach tree is always a peach, not an orange.
   (But maybe a mutant peach tree could make something new?)
3. Opening the refrigerator turns on a light.
   (But what if the light burns out?)

A **stochastic** process contains at least *some* natural random error, perhaps in addition to a pattern

Real world relationships are almost always stochastic

# Correlation between two random variables

We often want to summarize the amount of signal vs. noise in a real world relationship.

One way to do that is with a correlation coefficient.

We will work up to correlation coefficients by first exploring:

- Scatterplots
- Standardization

# Height and weight example

## Summary Statistics

| | Height in feet | Weight in pounds |
|---|---|---|
| Mean | 5.89 | 177.3 |
| Standard deviation | 0.25 | 28.5 |

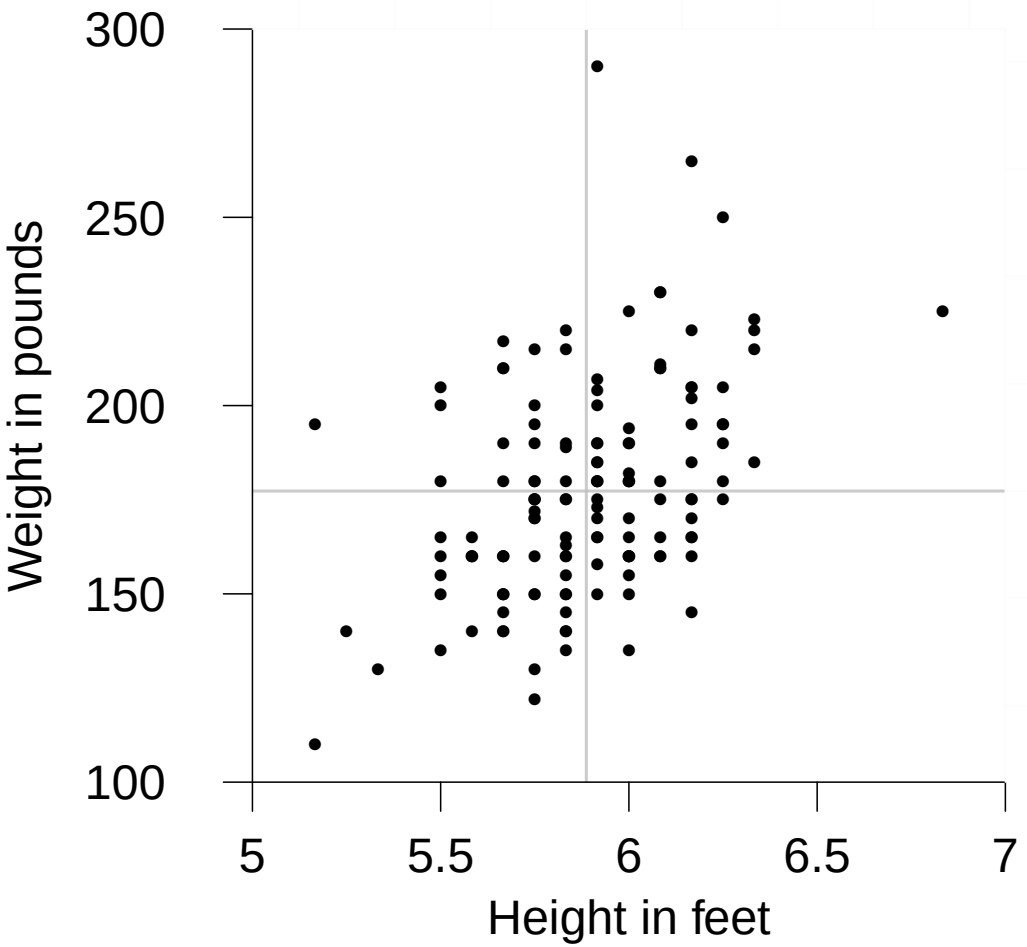
The CDC provides data from 2010 on the height (in feet) and weight (in pounds) for 21-year-old males

We have 137 cases in our sample

Question: to what extend do greater height & weight go together?

Best way to start exploring a relationship is graphically

# Height and Weight: Scatterplot

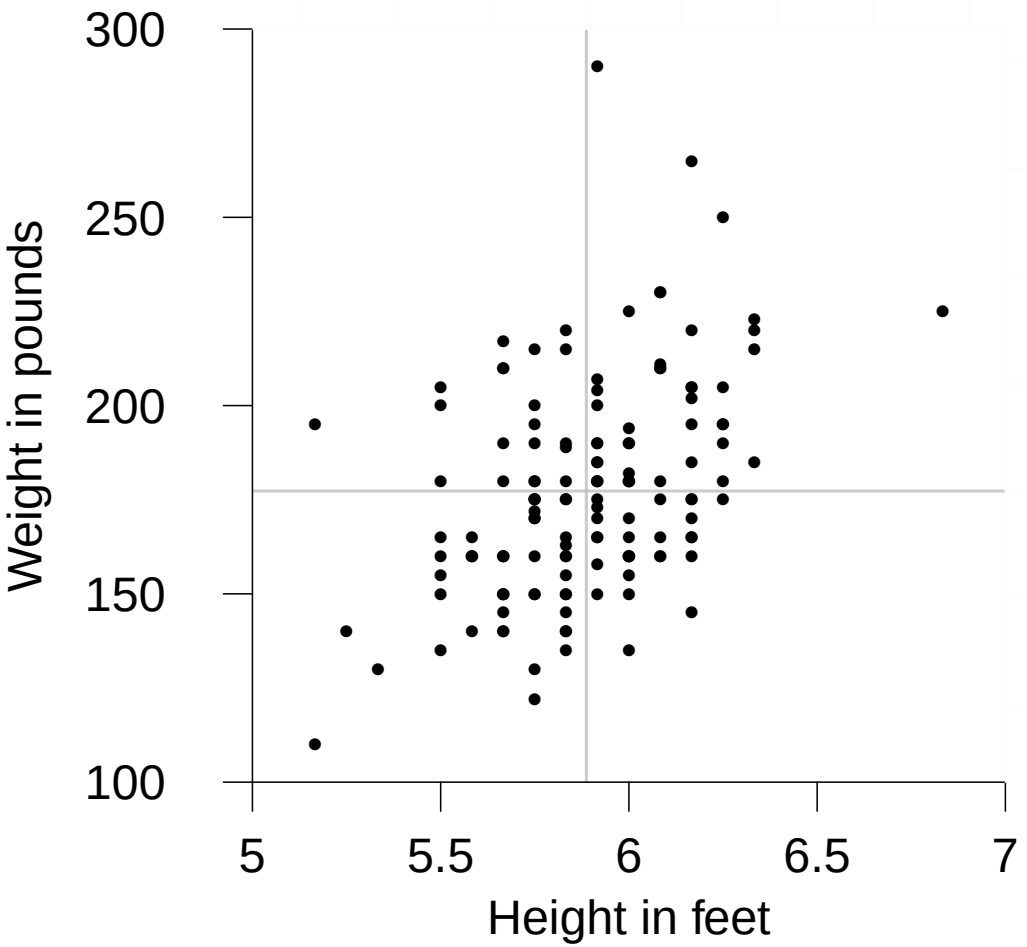



Gray lines mark the means of each variable

Mean height is $\bar{x} = 5.89$ feet

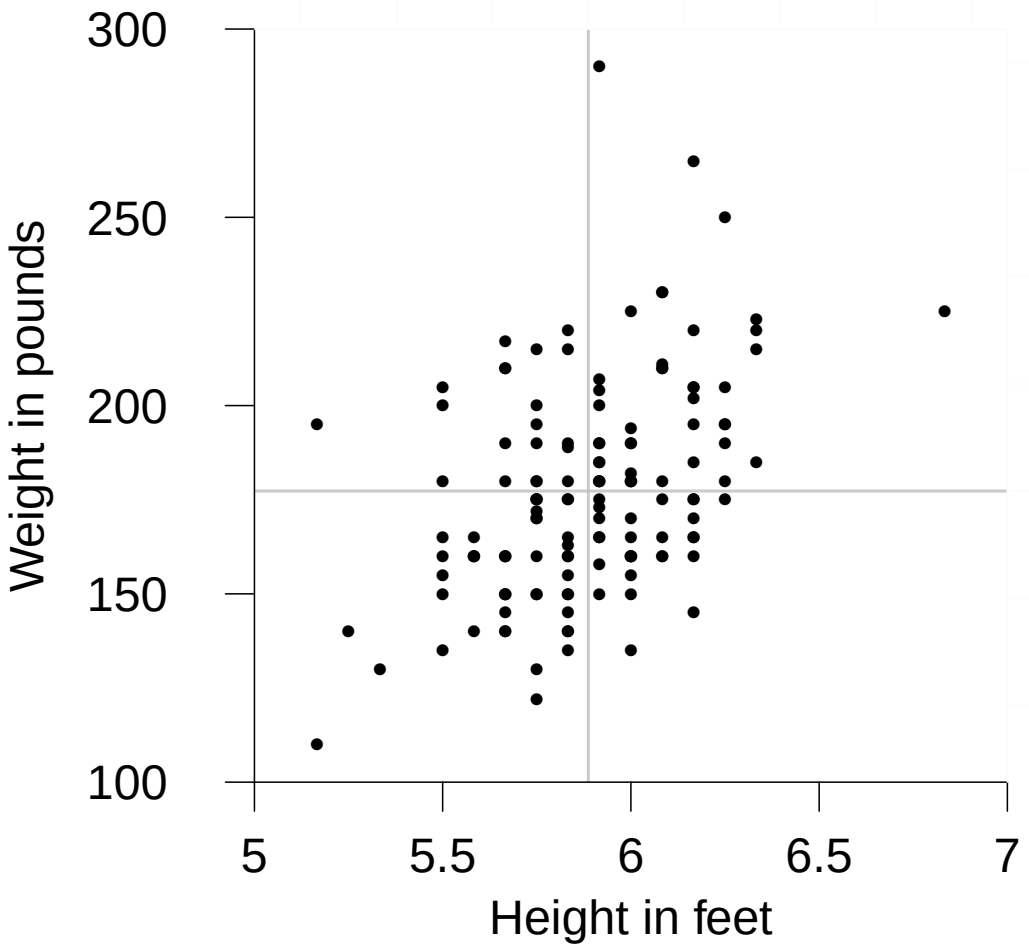
Mean weight is $\bar{y} = 177.3$ pounds

# Height and Weight: Scatterplot




Is there a relationship between height and weight?

# Height and Weight: Scatterplot




Is there a relationship between height and weight?

Height and weight appear moderately *positively* correlated

More of one usually means more of the other (but not always)

# Scatterplots

Most powerful tool for bivariate data analysis

Need additive level variables though!

No matter what advanced methods we learn,
scatterplots will *always* be useful

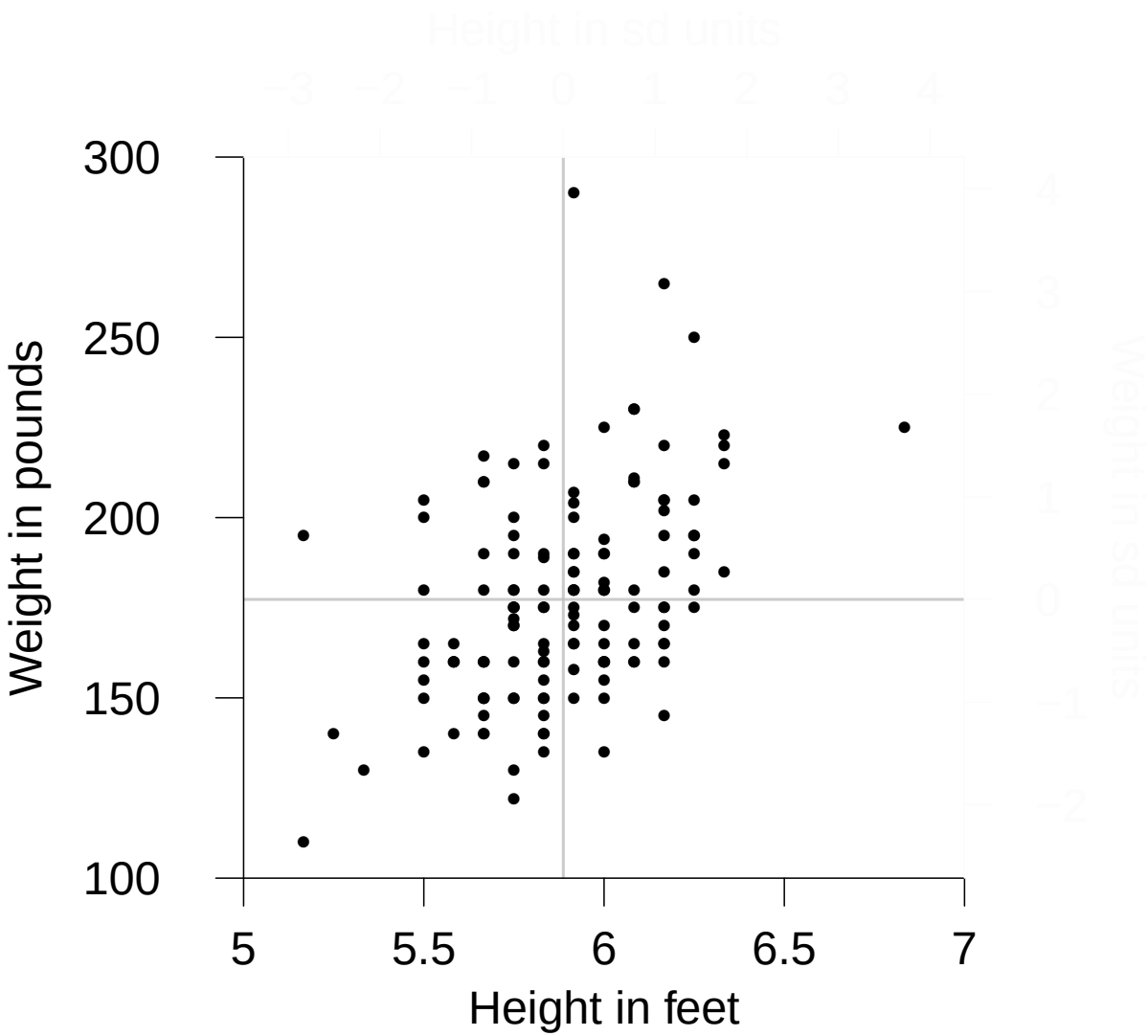
# Standardization

The relation between height and weight would be easier to understand if height and weight were in the same units

But that seems impossible! How do you convert “pounds” and “feet” to a common unit?

Not impossible. The first step is to mean-center the variables

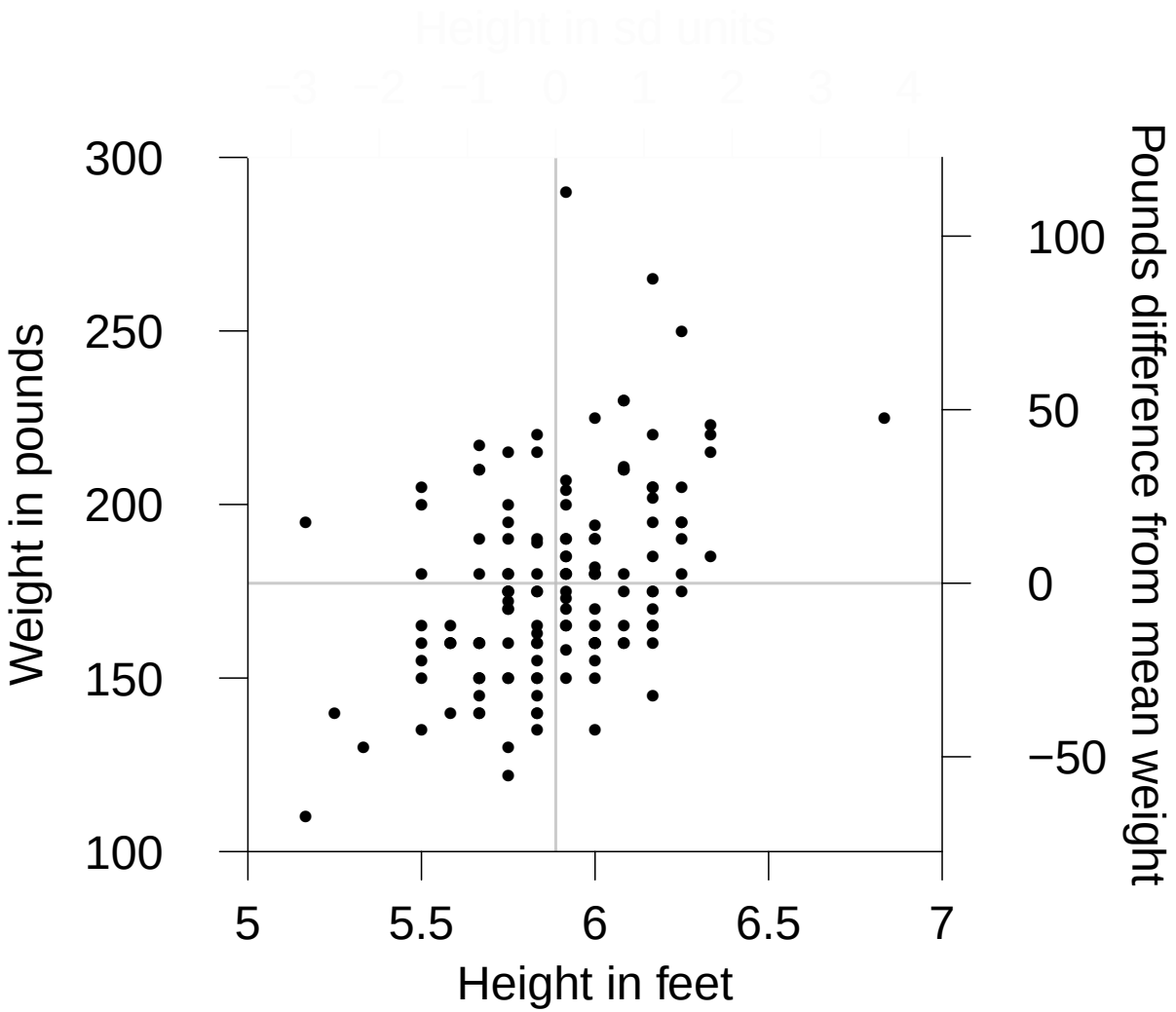
# Height and Weight: Standardization




Let's mean-center weight

This means we need to "remove" the average weight from each observation

# Height and Weight: Standardization




The right axis shows the *deviation* of each individual from the mean weight

$$y_i - \bar{y}$$

This doesn't change the data: we've just translated to a new unit

## Standardization

It would be easier to understand the relationship between height and weight if height and weight were in the same units

How can this be done?

First step: Mean-centering

$$y_i - \bar{y}$$

## Standardization

It would be easier to understand the relationship between height and weight if height and weight were in the same units

How can this be done?

First step: Mean-centering

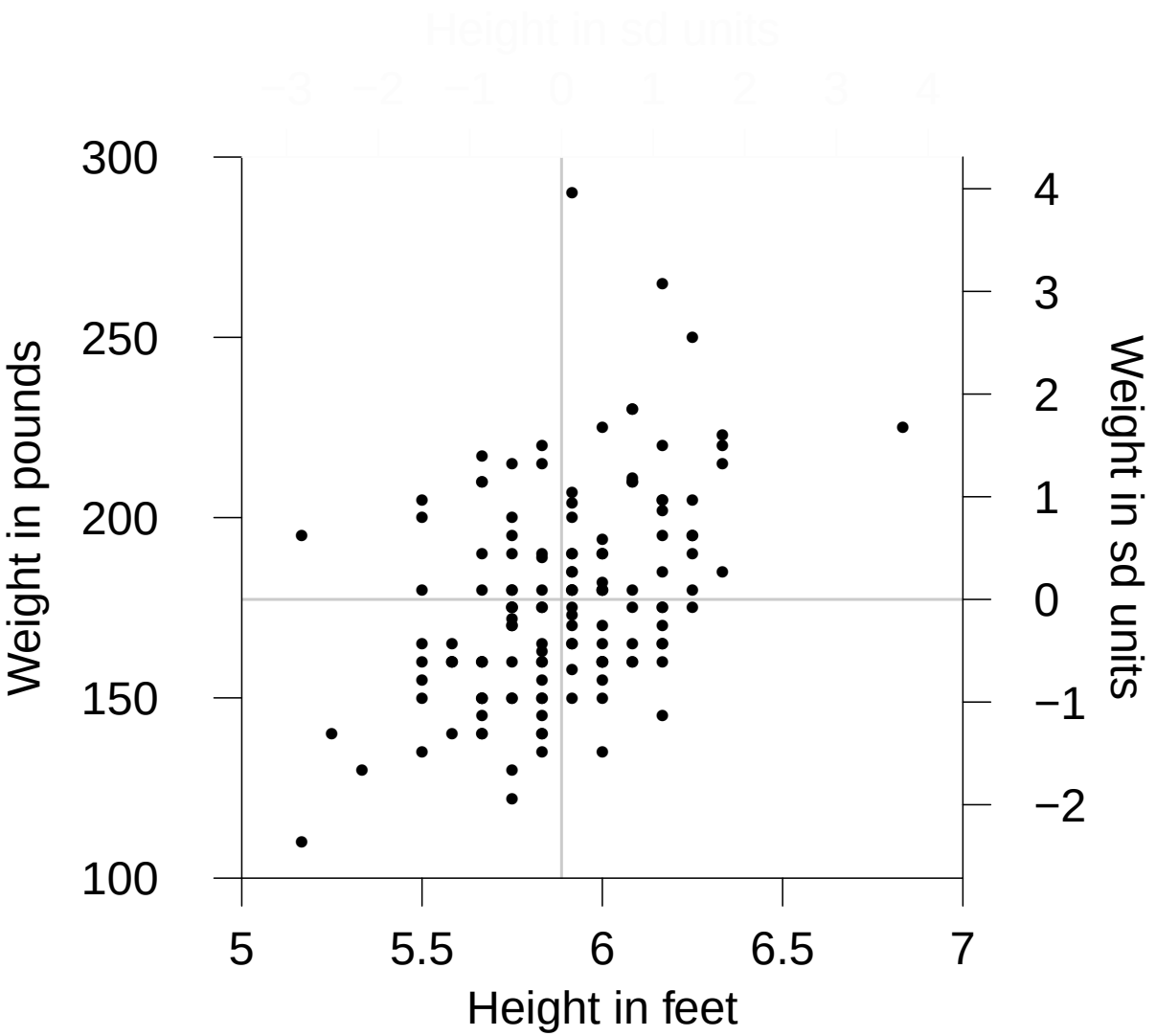
$$y_i - \bar{y}$$

Second step: Convert to standard deviation units

$$\frac{y_i - \bar{y}}{\sigma_y}$$

We can use this procedure to convert any continuous variable to a standardized unit

# Height and Weight: Standardization

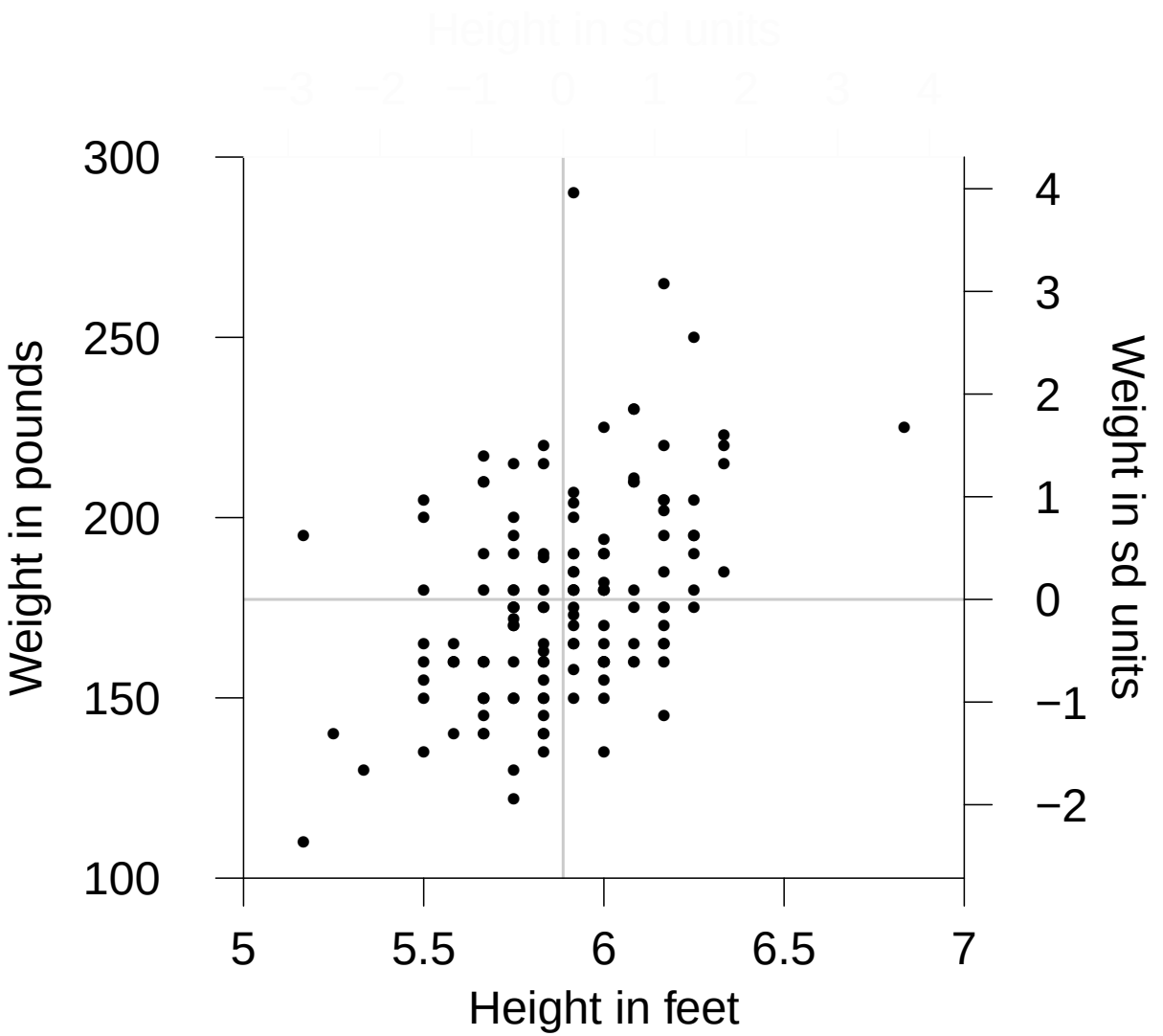



The right axis shows weight in *standard deviation units*

$$\frac{y_i - \bar{y}}{\sigma_y}$$

This still doesn't change the data: again, we've just translated to a new unit

# Height and Weight: Standardization

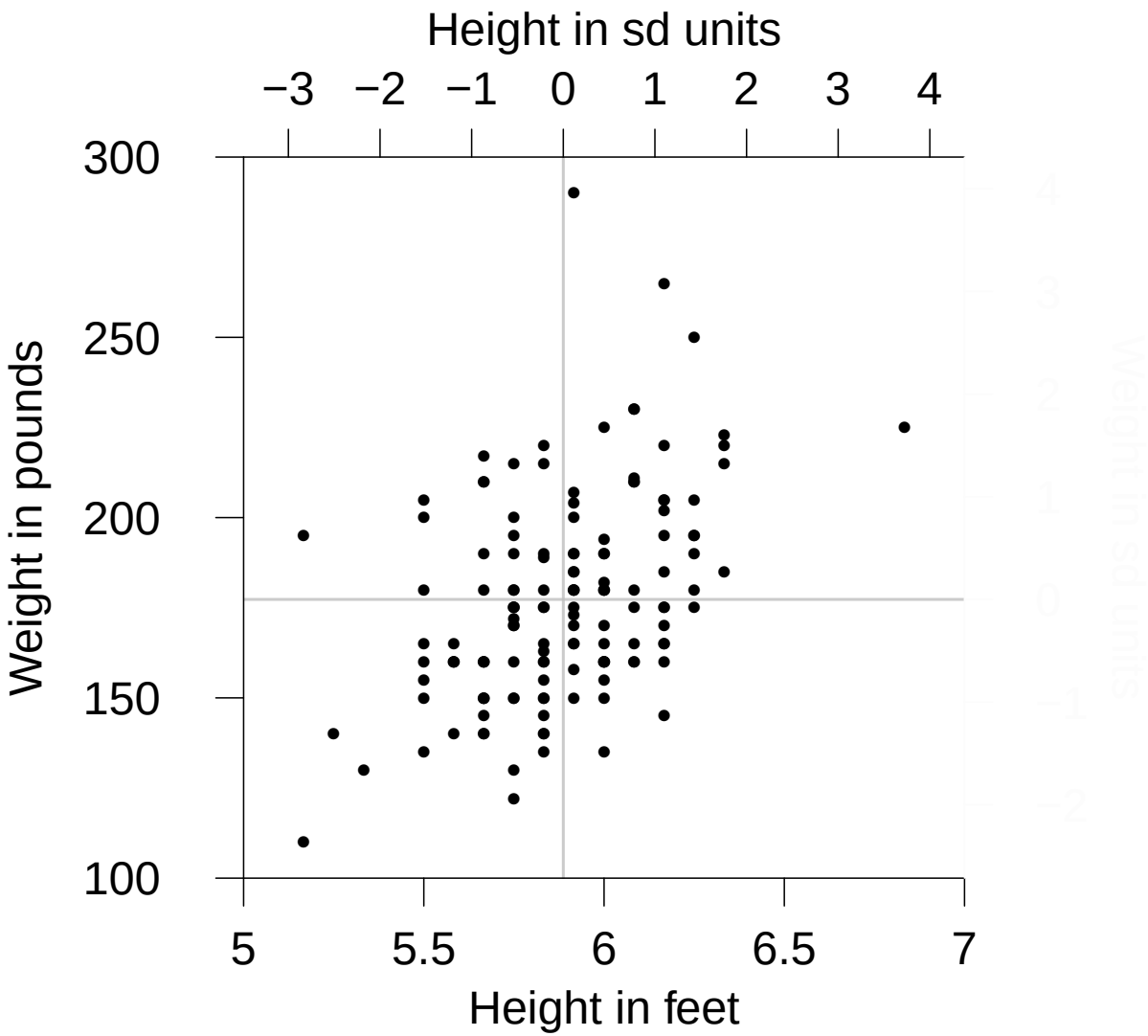



The right axis shows weight in *standard deviation units*

*What does that mean?*

One unit is now the average gap between two randomly drawn individuals

# Height and Weight: Standardization

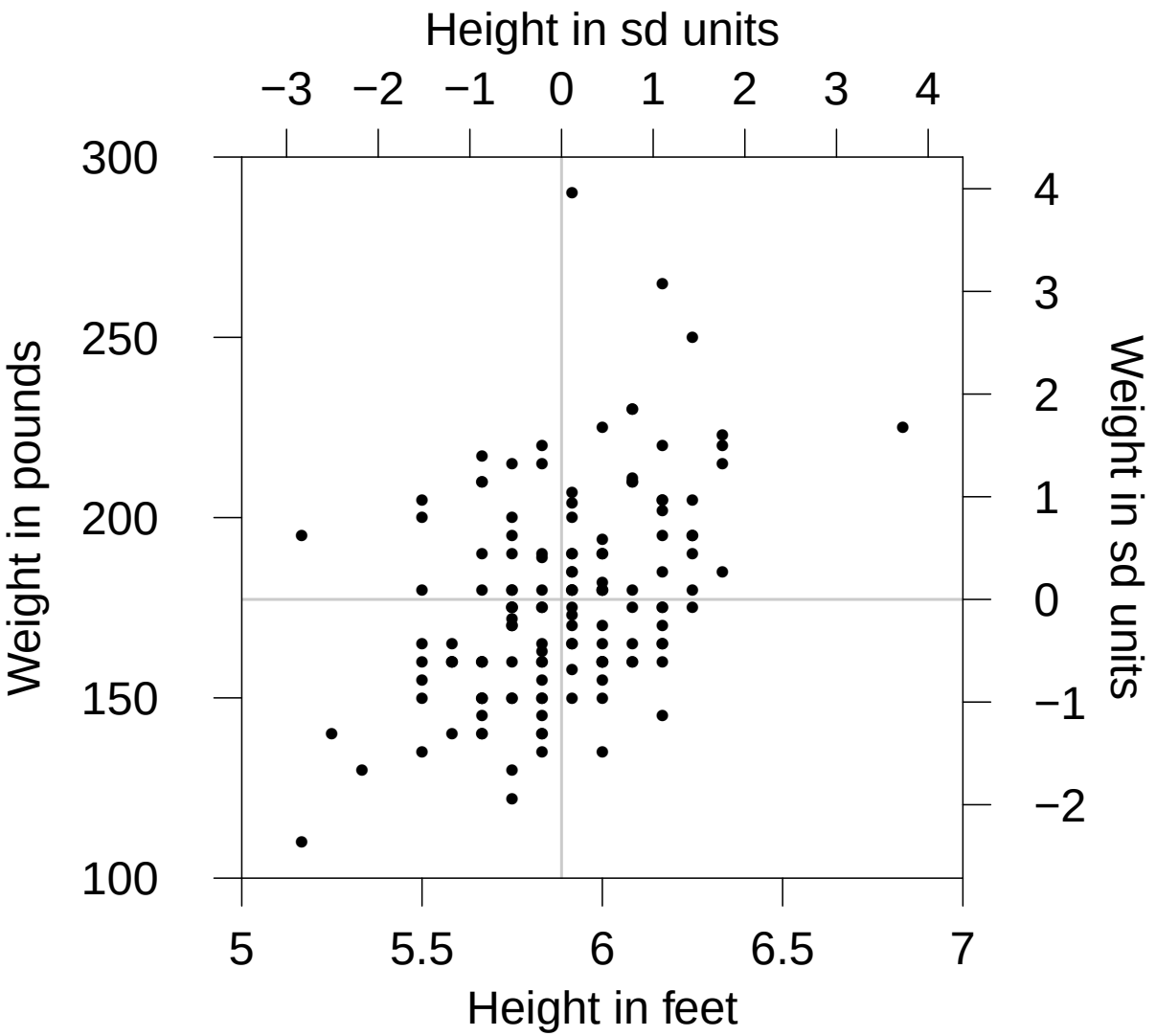



One unit is now the average gap between two randomly drawn individuals

We could convert *anything* to standard deviation units

The top axis shows height in *standard deviation units*

# Height and Weight: Standardization

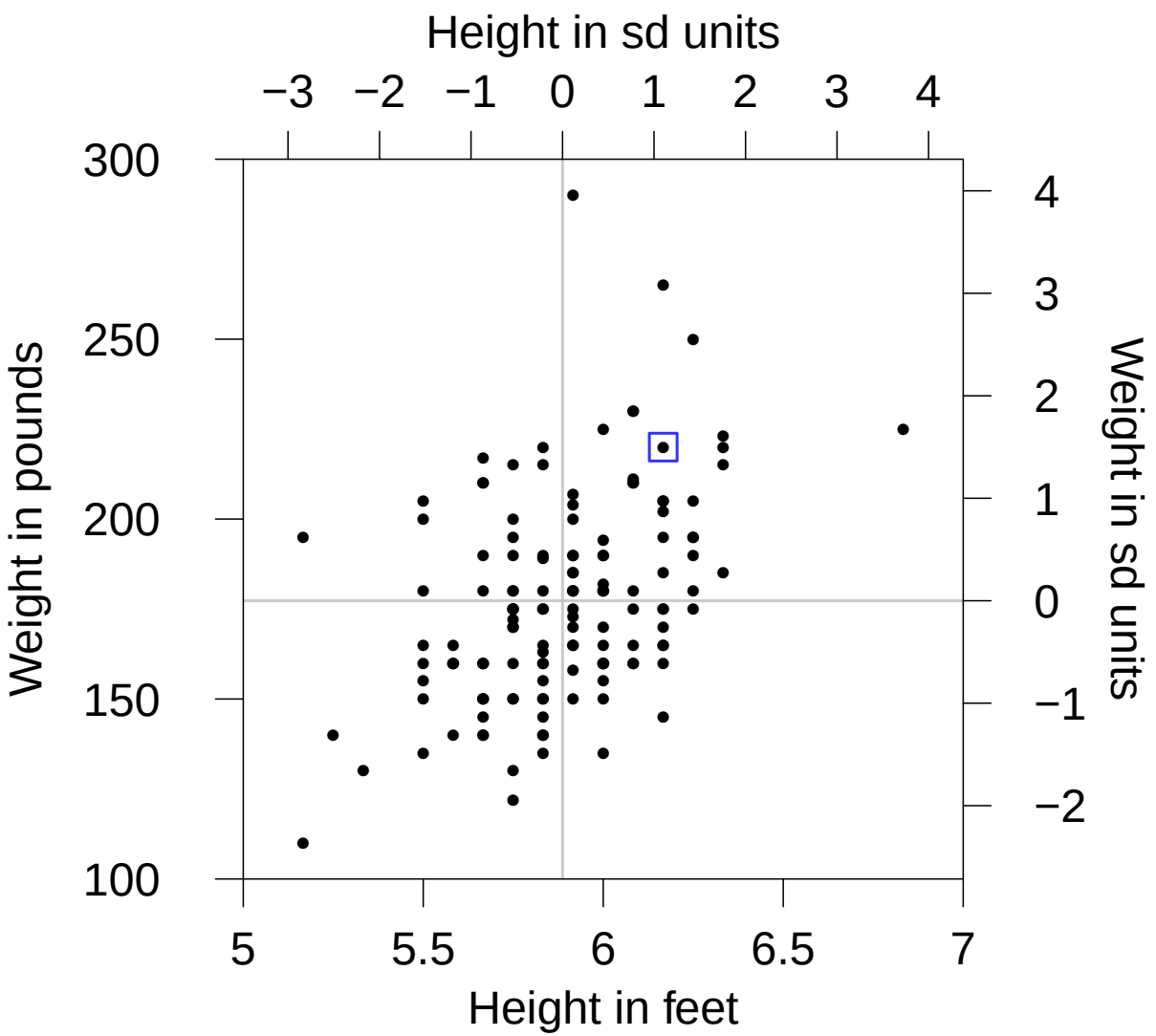



Standardization hasn't changed the pattern in the data at all

We've just relabeled units

Why did we do this? To help calculate the amount of correlation between height and weight

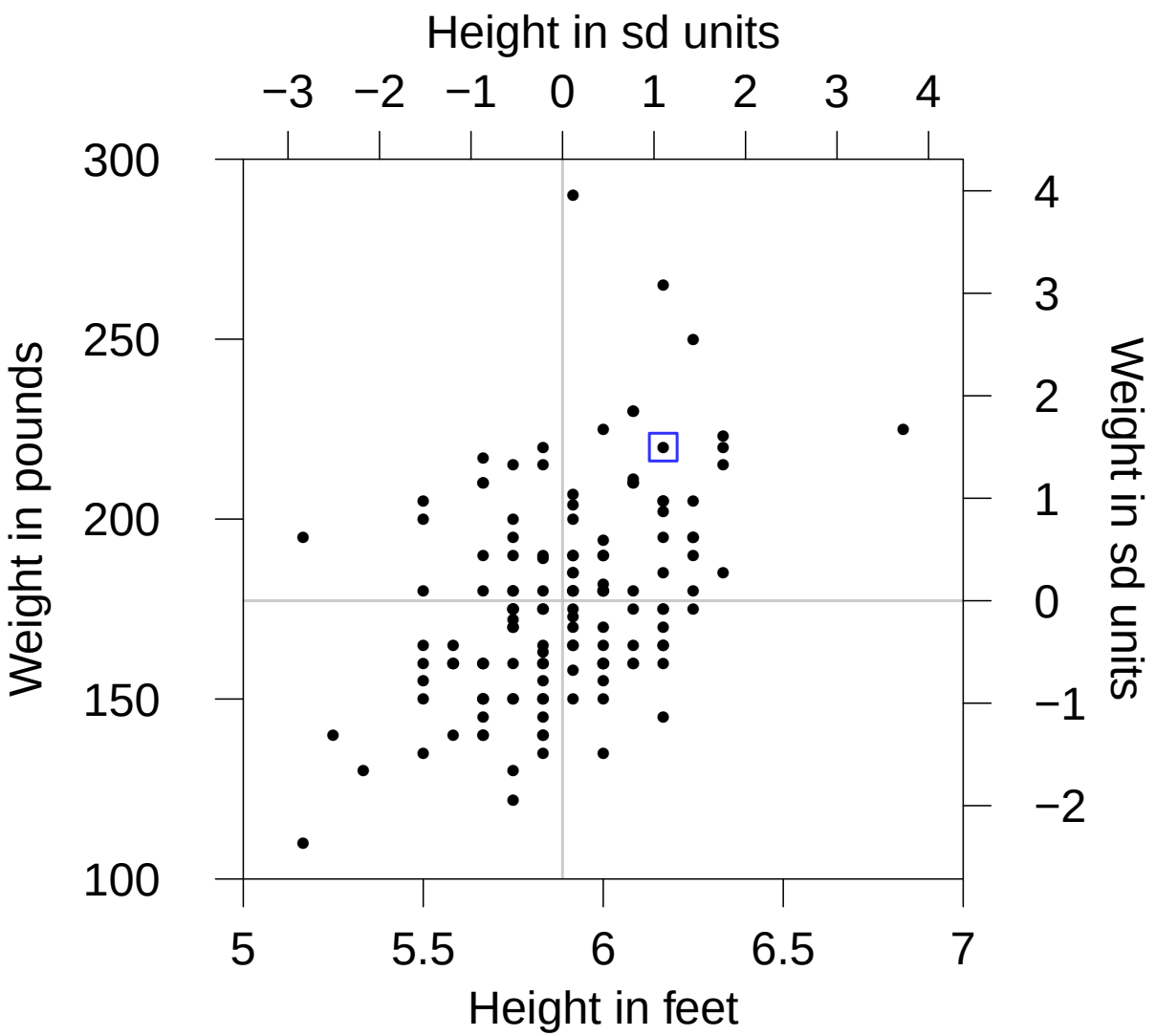
# Height and Weight: Standardization




Standardized weight for blue obs:

$$y_i^{\text{std}} = \frac{y_i - \bar{y}}{\sigma_y}$$

# Height and Weight: Standardization

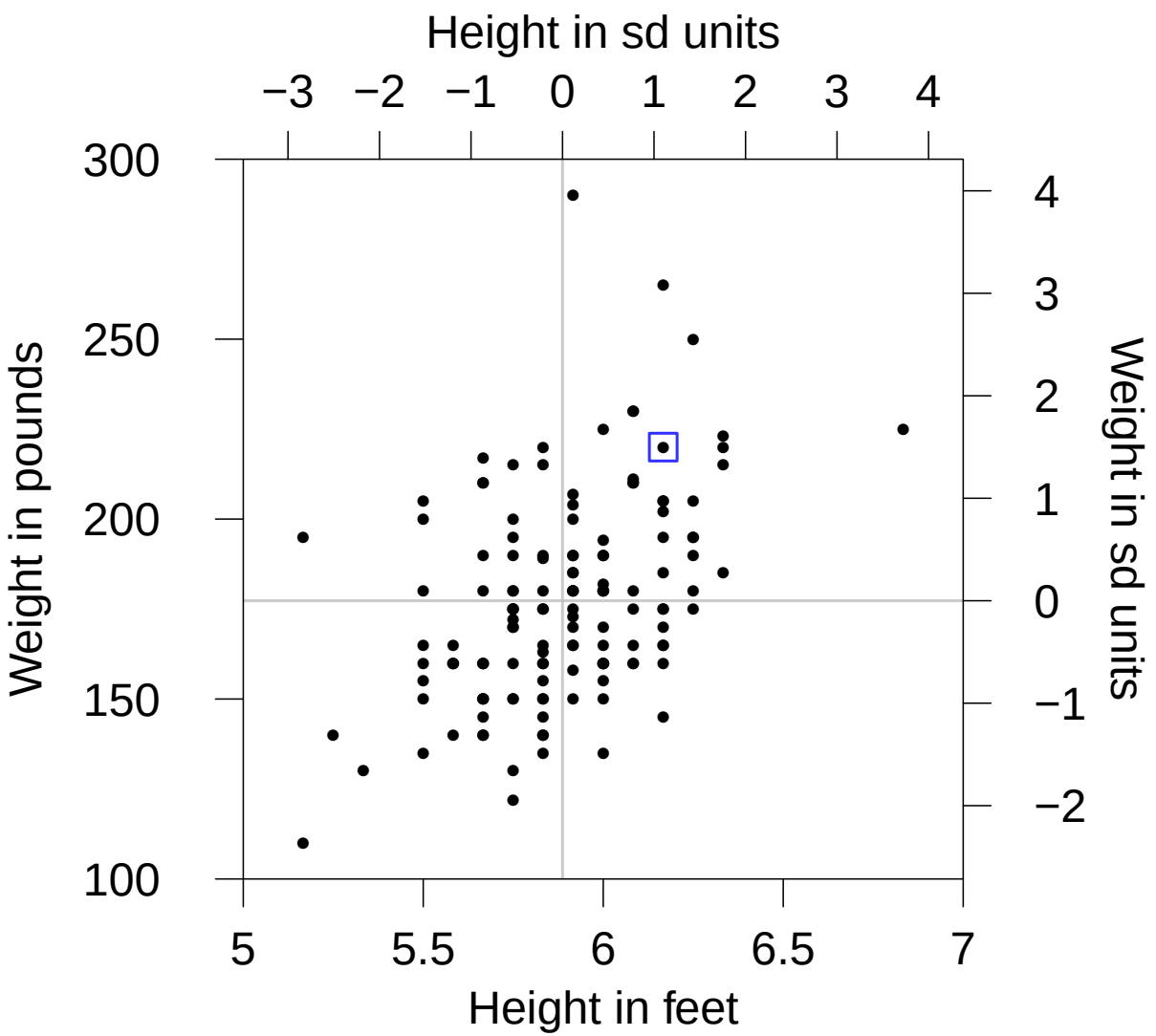



Standardized weight for blue obs:

$$y_i^{\text{std}} = \frac{y_i - \bar{y}}{\sigma_y}$$

$$+1.5 = \frac{220 - 177.3}{28.5}$$

# Height and Weight: Standardization




Standardized weight for blue obs:

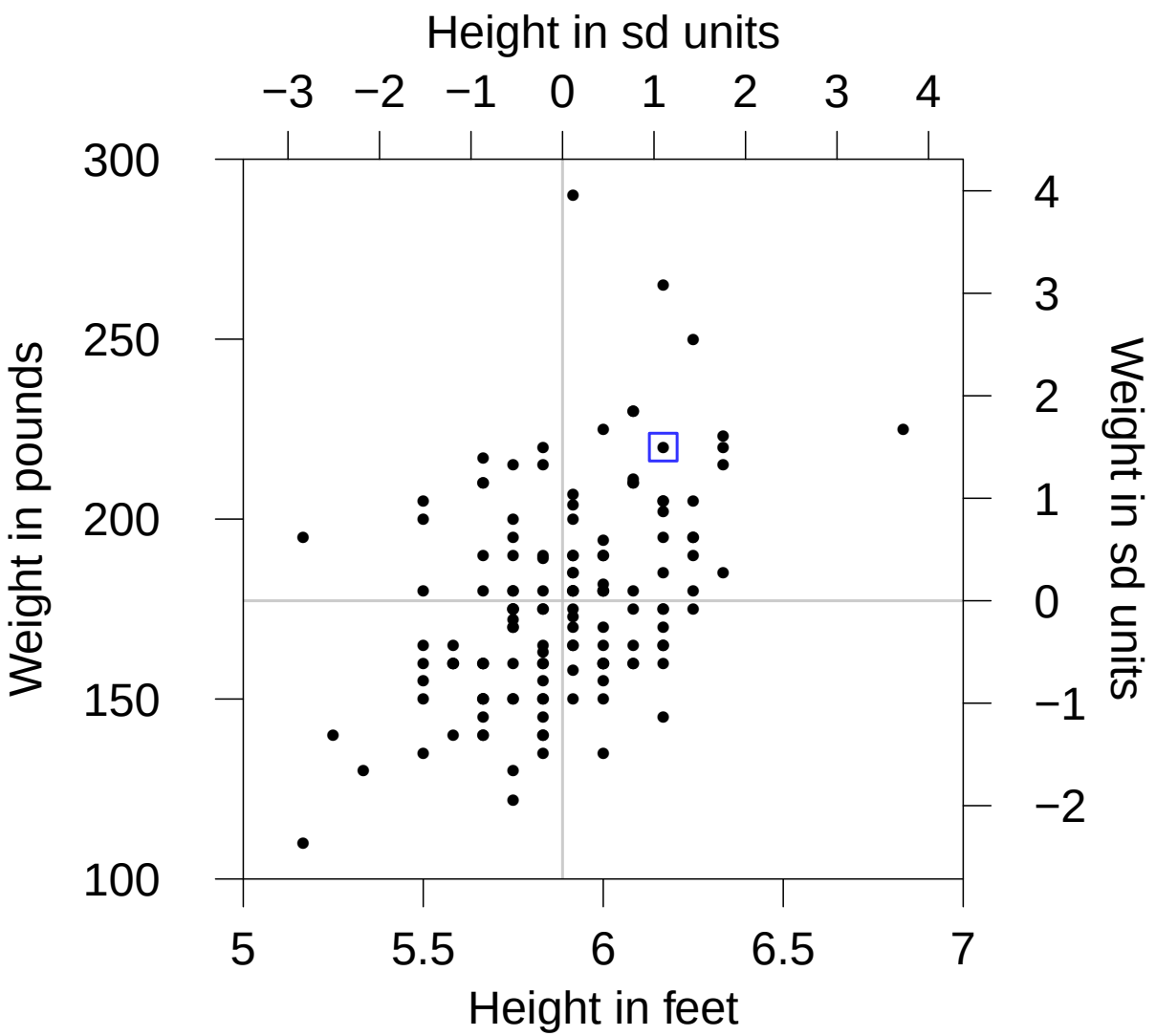
$$y_i^{\text{std}} = \frac{y_i - \bar{y}}{\sigma_y}$$

$$+1.5 = \frac{220 - 177.3}{28.5}$$

Standardized height for the blue obs:

$$x_i^{\text{std}} = \frac{x_i - \bar{x}}{\sigma_x}$$

# Height and Weight: Standardization




Standardized weight for blue obs:

$$y_i^{\text{std}} = \frac{y_i - \bar{y}}{\sigma_y}$$

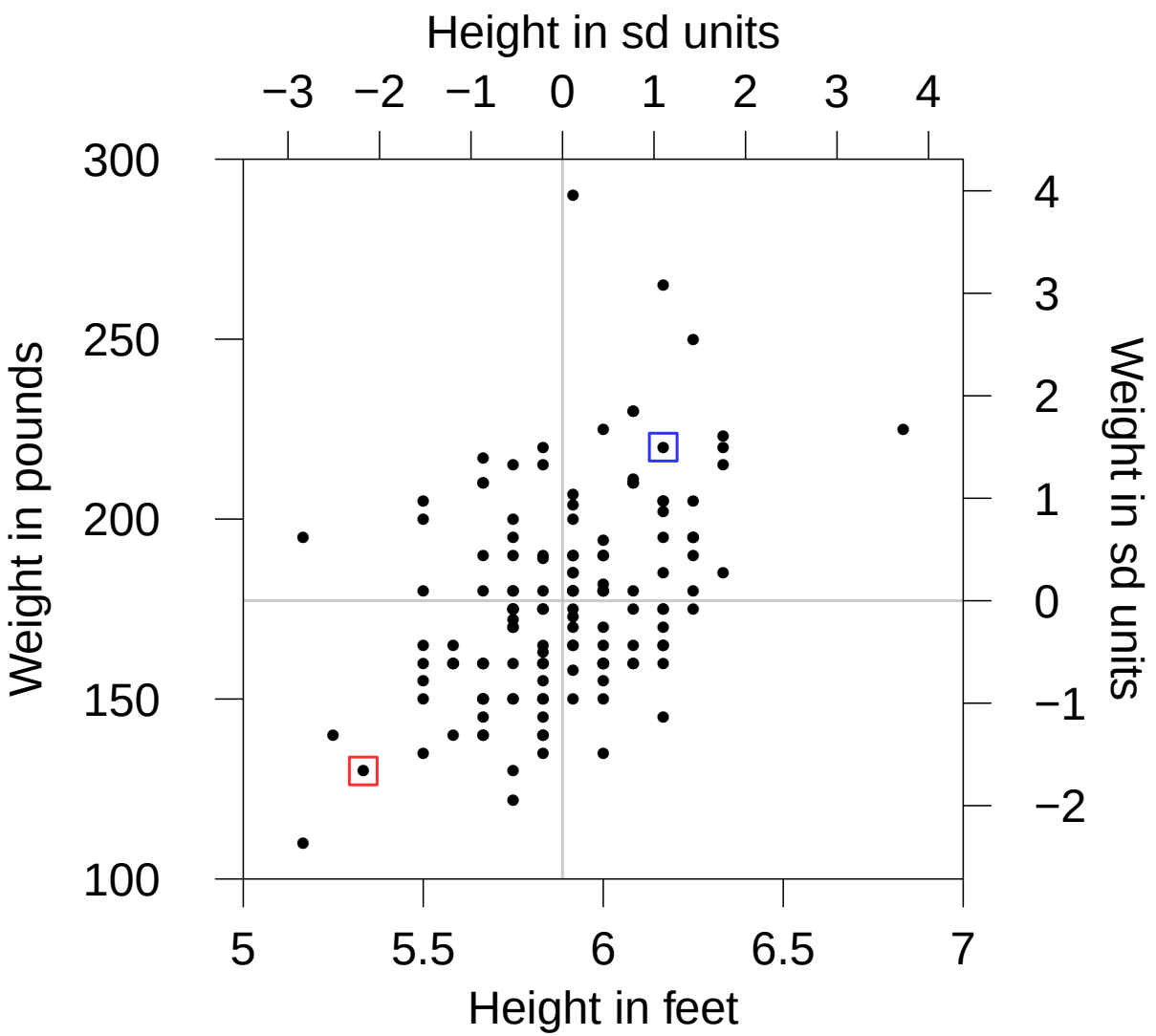
$$+1.5 = \frac{220 - 177.3}{28.5}$$

Standardized height for the blue obs:

$$x_i^{\text{std}} = \frac{x_i - \bar{x}}{\sigma_x}$$

$$+1.1 = \frac{6.17 - 5.89}{0.25}$$

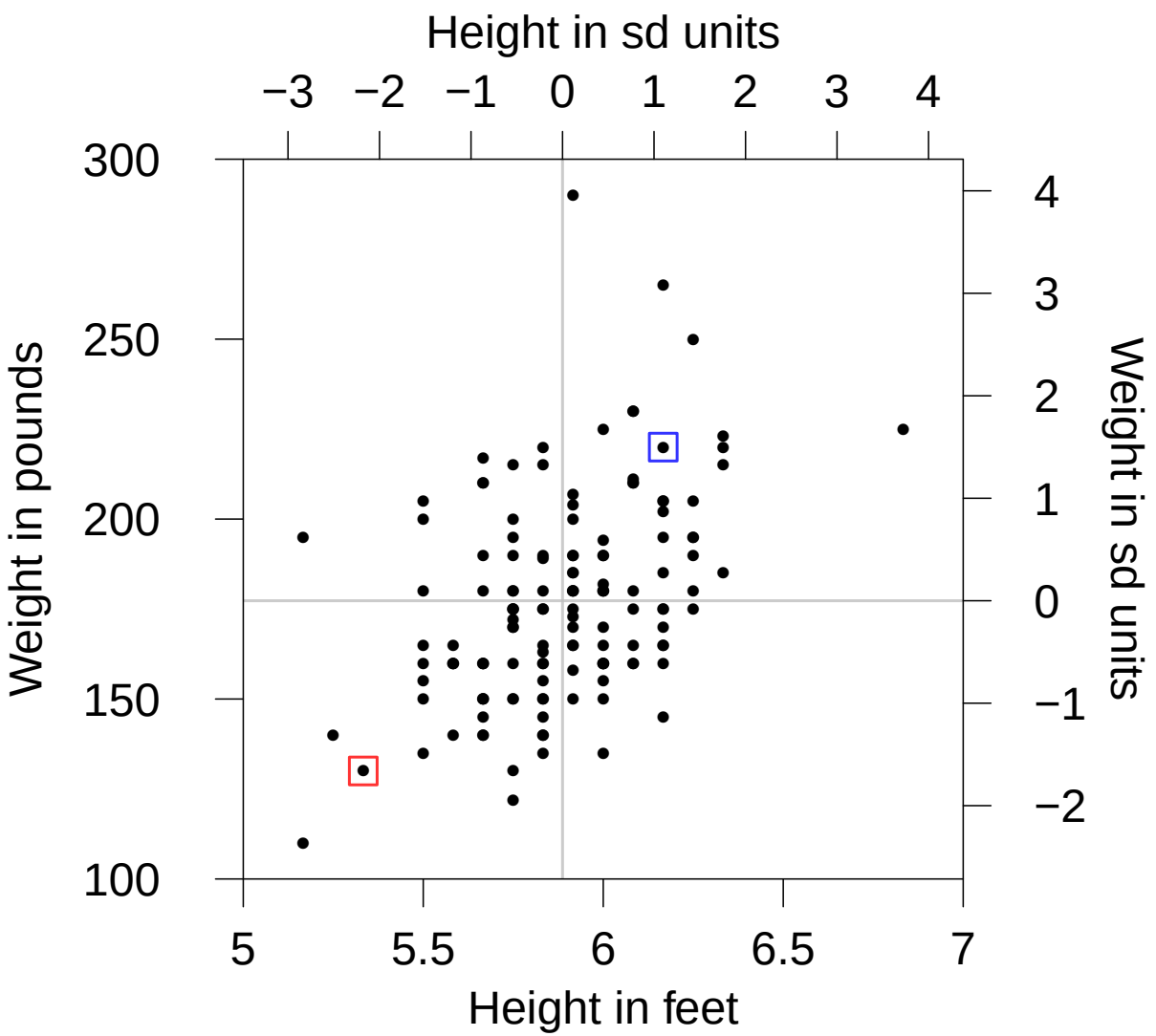
# Height and Weight: Standardization




Standardized weight for red obs:

$$y_i^{\text{std}} = \frac{y_i - \bar{y}}{\sigma_y}$$

# Height and Weight: Standardization

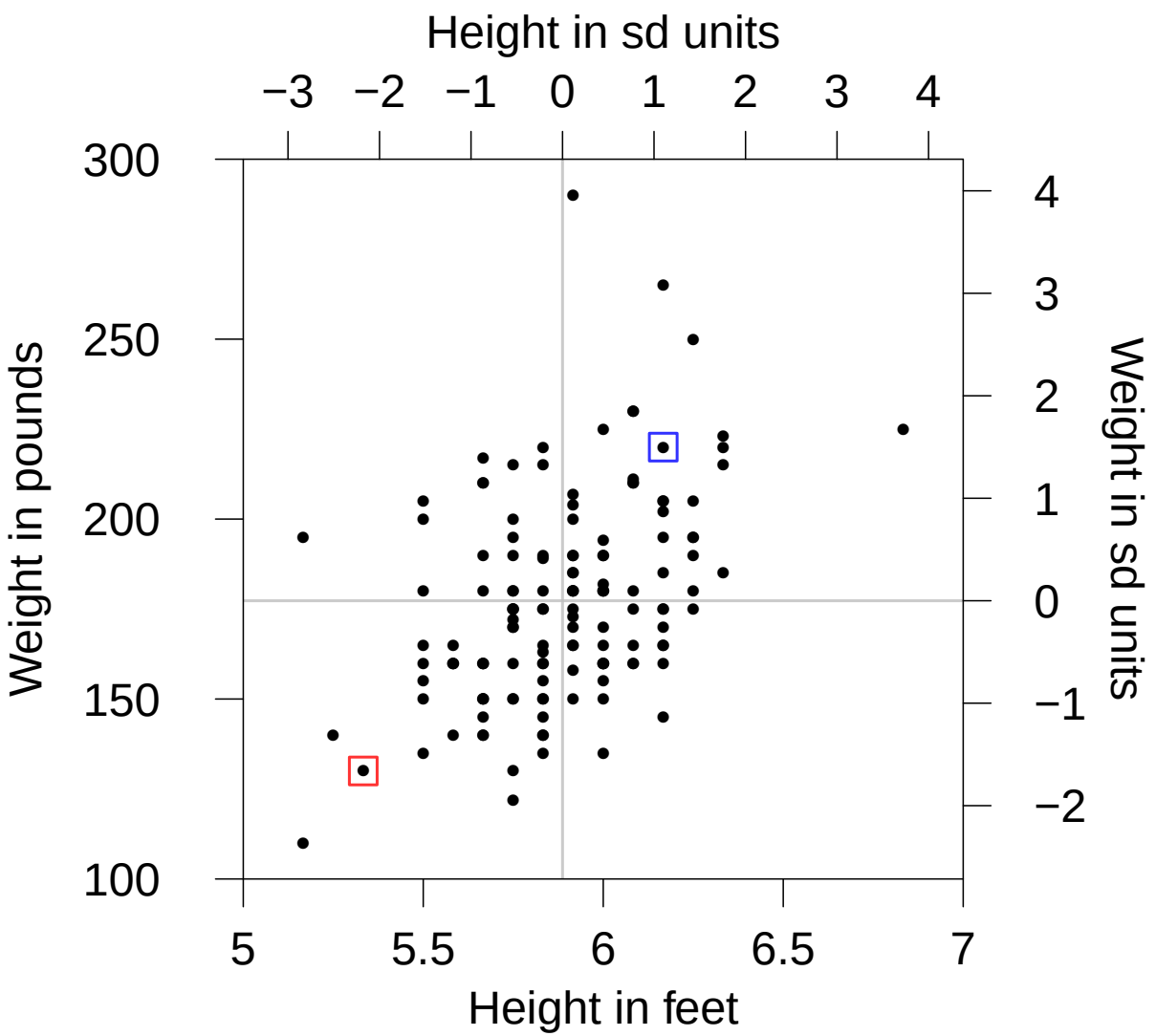



Standardized weight for red obs:

$$y_i^{\text{std}} = \frac{y_i - \bar{y}}{\sigma_y}$$

$$-1.7 = \frac{130 - 177.3}{28.5}$$

# Height and Weight: Standardization




Standardized weight for red obs:

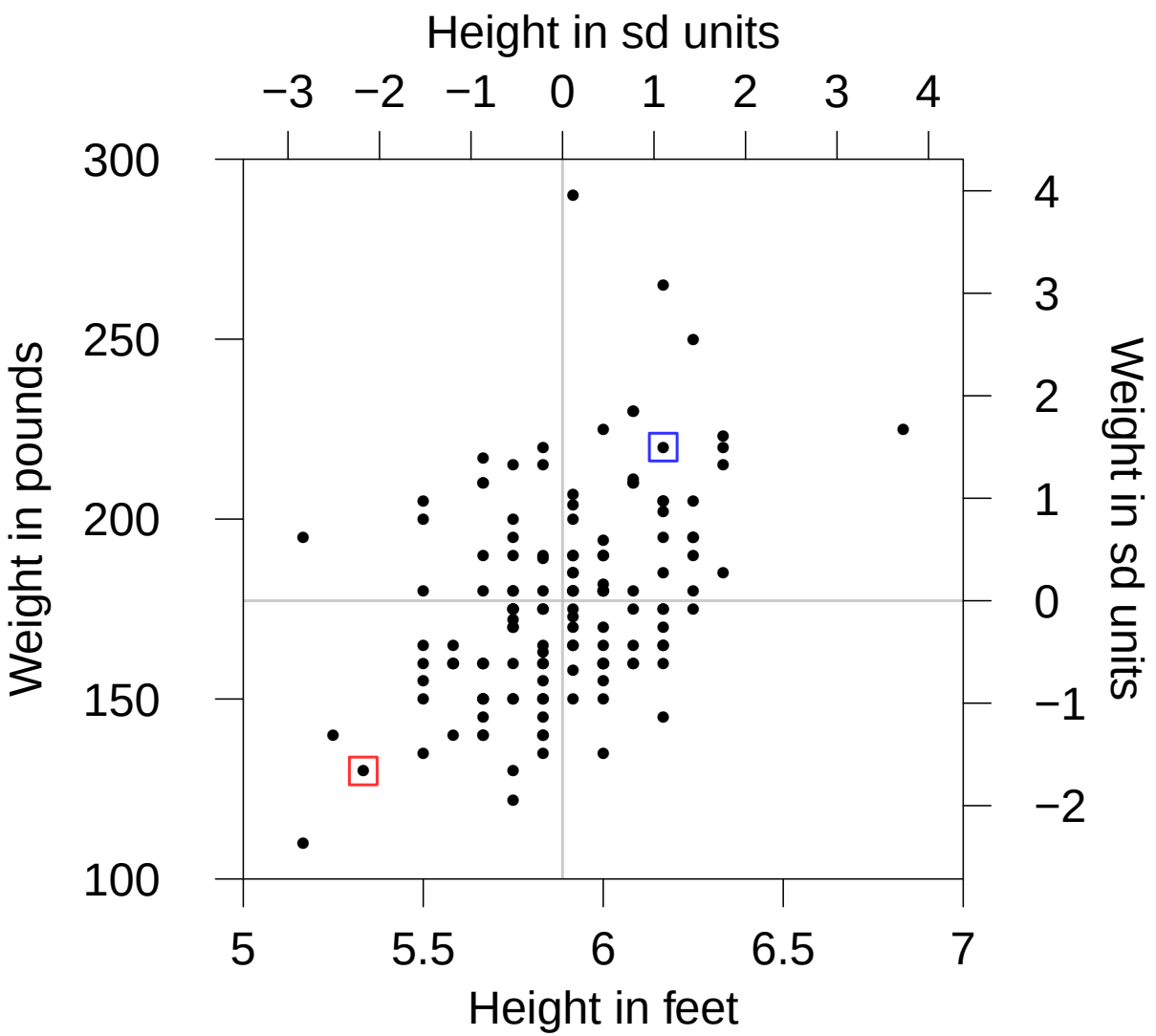
$$y_i^{\text{std}} = \frac{y_i - \bar{y}}{\sigma_y}$$

$$-1.7 = \frac{130 - 177.3}{28.5}$$

Standardized height for the red obs:

$$x_i^{\text{std}} = \frac{x_i - \bar{x}}{\sigma_x}$$

# Height and Weight: Standardization




Standardized weight for red obs:

$$y_i^{\text{std}} = \frac{y_i - \bar{y}}{\sigma_y}$$

$$-1.7 = \frac{130 - 177.3}{28.5}$$

Standardized height for the red obs:

$$x_i^{\text{std}} = \frac{x_i - \bar{x}}{\sigma_x}$$

$$-2.2 = \frac{5.33 - 5.89}{0.25}$$

## Correlation coefficient

When two variables are closely related,
their standardized values are similar

The correlation coefficient measures how similar two variables are by averaging the products of their standardized values:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} y_i^{\text{std}} x_i^{\text{std}}$$

## Correlation coefficient

When two variables are closely related,
their standardized values are similar

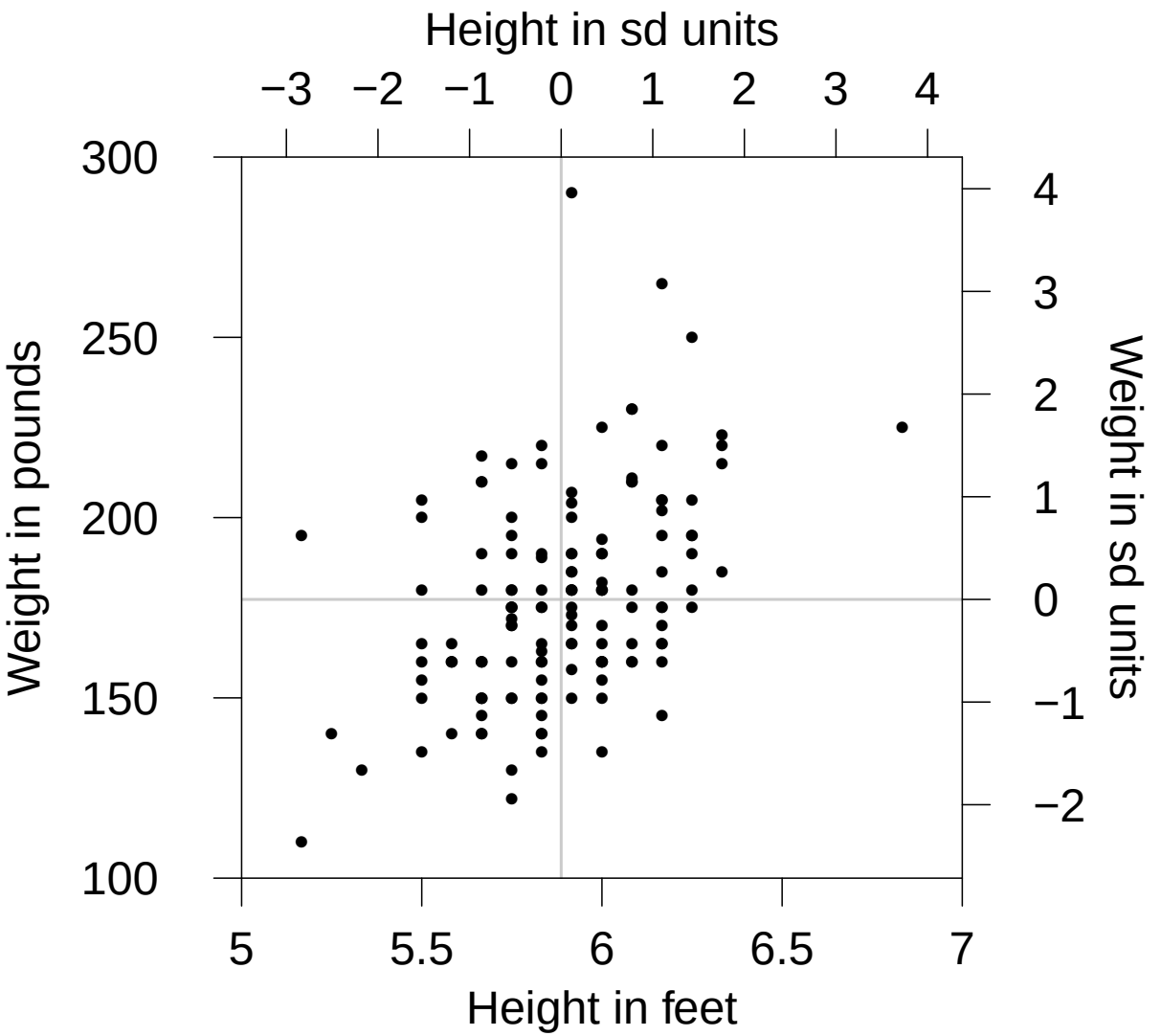
The correlation coefficient measures how similar two variables are by averaging the products of their standardized values:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} y_i^{\text{std}} x_i^{\text{std}}$$

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{y_i - \bar{y}}{\sigma_y} \right) \left( \frac{x_i - \bar{x}}{\sigma_x} \right)$$

The *correlation coefficient*, $r$, between two variables measures the strength of association between them on a $[-1, 1]$ scale

# Height and Weight: Correlation

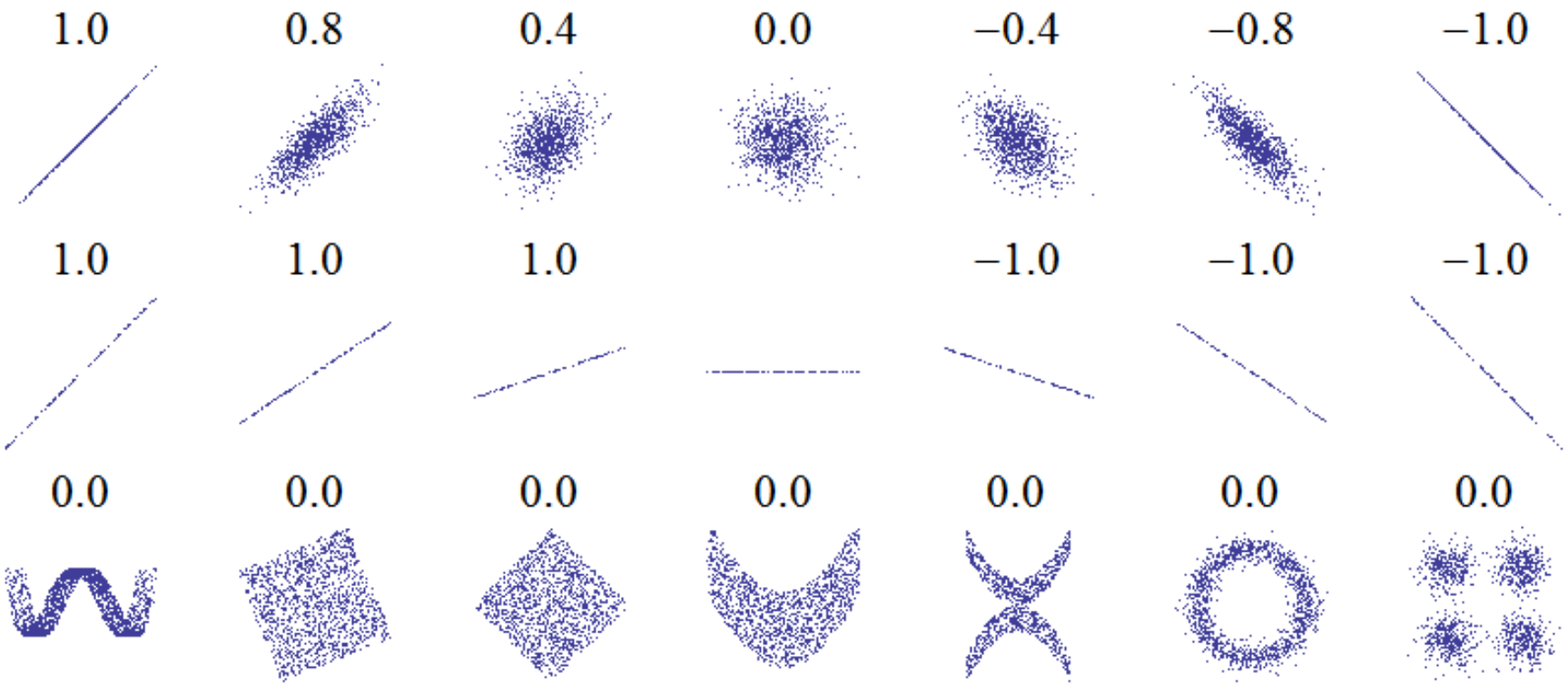



The correlation coefficient for height and weight is

$$r = 0.43$$

Is this “big”?

# Correlation examples

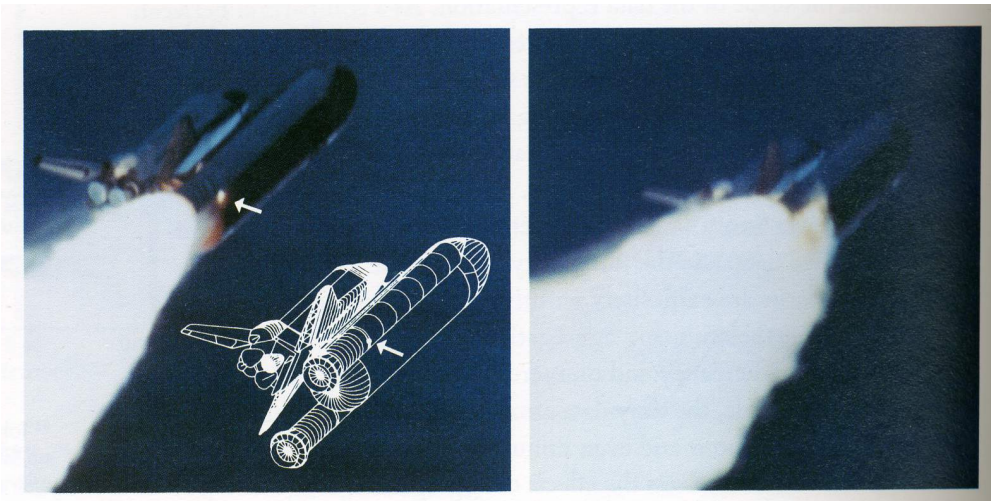


# Correlation coefficients measure strength of association

When two variables $X$ and $Y$ are highly *correlated*:

- They have $|r_{X,Y}| \approx 1$
- If we know $X$, we can narrow the expected range of $Y$ down to a small interval
- If we know $Y$, we can narrow the expected range of $X$ down to a small interval

# The *Challenger* launch decision



In 1986, the Challenger space shuttle exploded moments after liftoff

Decision to launch one of the most scrutinized in history

Failure of O-rings in the solid-fuel rocket boosters blamed for explosion

Could this failure have been forseen? Using statistics?

# The *Challenger* launch decision

Here is the data on O-ring failures at different launch temperatures

| Flights with O-ring damage | |
|---|---|
| Flt Number | Temp (F) |
| 2 | 70 |
| 41b | 57 |
| 41c | 63 |
| 41d | 70 |
| 51c | 53 |
| 61a | 79 |
| 61c | 58 |

Morton-Thiokol engineers made this table & worried about launching below 53 degrees (Why?)

O-ring would erode or have “blow-by” (2 ways to fail) in cold temp

# The *Challenger* launch decision

Here is the data on O-ring failures at different launch temperatures

| Flights with O-ring damage | |
|---|---|
| Flt Number | Temp (F) |
| 2 | 70 |
| 41b | 57 |
| 41c | 63 |
| 41d | 70 |
| 51c | 53 |
| 61a | 79 |
| 61c | 58 |

Failed to convince administrators there was a danger

(Counter-argument: “damages at low and high temps”)

# The *Challenger* launch decision

Here is the data on O-ring failures at different launch temperatures

| Flights with O-ring damage | |
|---|---|
| Flt Number | Temp (F) |
| 2 | 70 |
| 41b | 57 |
| 41c | 63 |
| 41d | 70 |
| 51c | 53 |
| 61a | 79 |
| 61c | 58 |

Are there problems with this presentation? with the use of data?

# The *Challenger* launch decision

Engineers did not consider successes, only failures;
"selection on the dependent variable" (selection bias)

| All flights, chronological order | | | |
|---|---|---|---|
| Damage? | Temp (F) | Damage? | Temp (F) |
| No | 66 | No | 78 |
| **Yes** | 70 | No | 67 |
| No | 69 | **Yes** | 53 |
| No | 68 | No | 67 |
| No | 67 | No | 75 |
| No | 72 | No | 70 |
| No | 73 | No | 81 |
| No | 70 | No | 76 |
| **Yes** | 57 | **Yes** | 79 |
| **Yes** | 63 | No | 76 |
| **Yes** | 70 | **Yes** | 58 |

Other problems?

# The *Challenger* launch decision

Engineers did not consider successes, only failures;
"selection on the dependent variable" (selection bias)

| All flights, chronological order | | | |
|---|---|---|---|
| Damage? | Temp (F) | Damage? | Temp (F) |
| No | 66 | No | 78 |
| **Yes** | 70 | No | 67 |
| No | 69 | **Yes** | 53 |
| No | 68 | No | 67 |
| No | 67 | No | 75 |
| No | 72 | No | 70 |
| No | 73 | No | 81 |
| No | 70 | No | 76 |
| **Yes** | 57 | **Yes** | 79 |
| **Yes** | 63 | No | 76 |
| **Yes** | 70 | **Yes** | 58 |

Other problems? Why sort by launch number?

# The *Challenger* launch decision

| O-ring damage pre-Challenger, by temperature at launch | | | |
|---|---|---|---|
| Damage? | Temp (F) | Damage? | Temp (F) |
| **Yes** | 53 | **Yes** | 70 |
| **Yes** | 57 | No | 70 |
| **Yes** | 58 | No | 70 |
| **Yes** | 63 | No | 72 |
| No | 66 | No | 73 |
| No | 67 | No | 75 |
| No | 67 | No | 76 |
| No | 67 | No | 76 |
| No | 68 | No | 78 |
| No | 69 | **Yes** | 79 |
| **Yes** | 70 | No | 81 |

The evidence begins to speak for itself.
What if Morton-Thiokol engineers had made this table before the launch?

# The *Challenger* launch decision

Why didn't NASA make the right decision?

Many answers in the literature:
bureaucratic politics; group think; bounded rationality, etc.

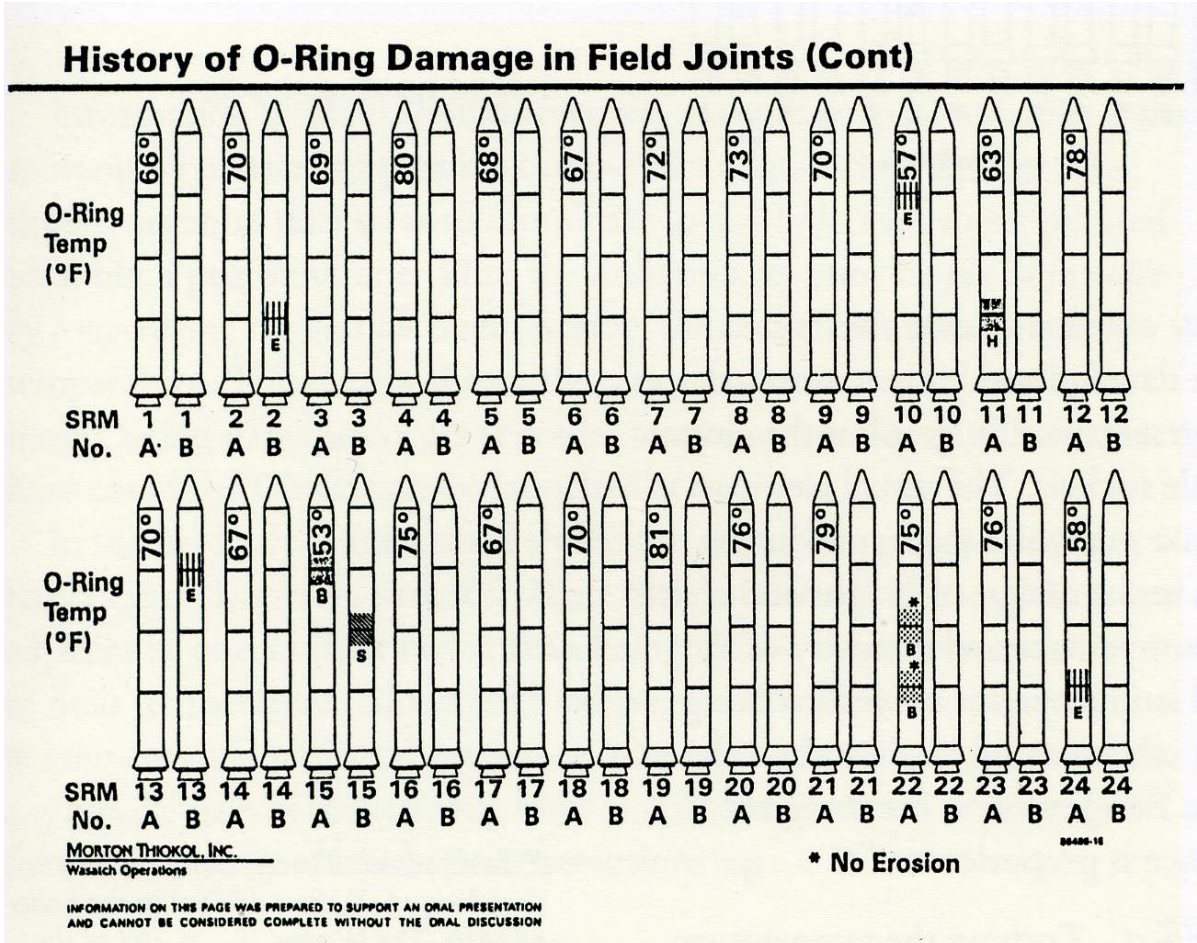
But Edward Tufte thinks it may have been a matter of presentation & modeling:

- Never made the right tables or graphics
- Selected only failure data
- Never considered a even simple statistical model

What do you think? How would you approach the data?

# The *Challenger* launch decision

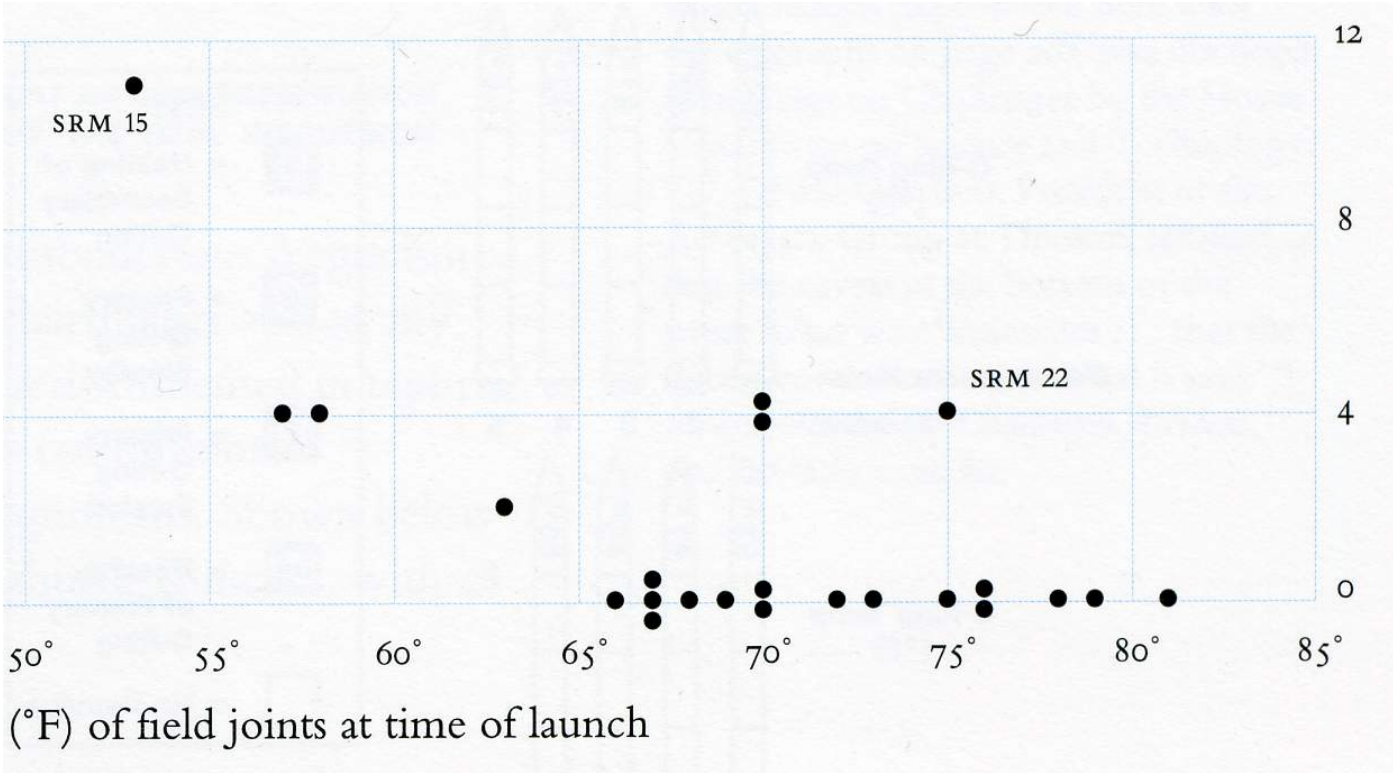
This is what Morton-Thiokol came up with to present after the disaster:



# The *Challenger* launch decision

How about a scatterplot for shuttle data? Need an additive measure of O-ring damage (Tufte's index)

Vertical axis is an O-ring damage index (due to Tufte, who made the plot)



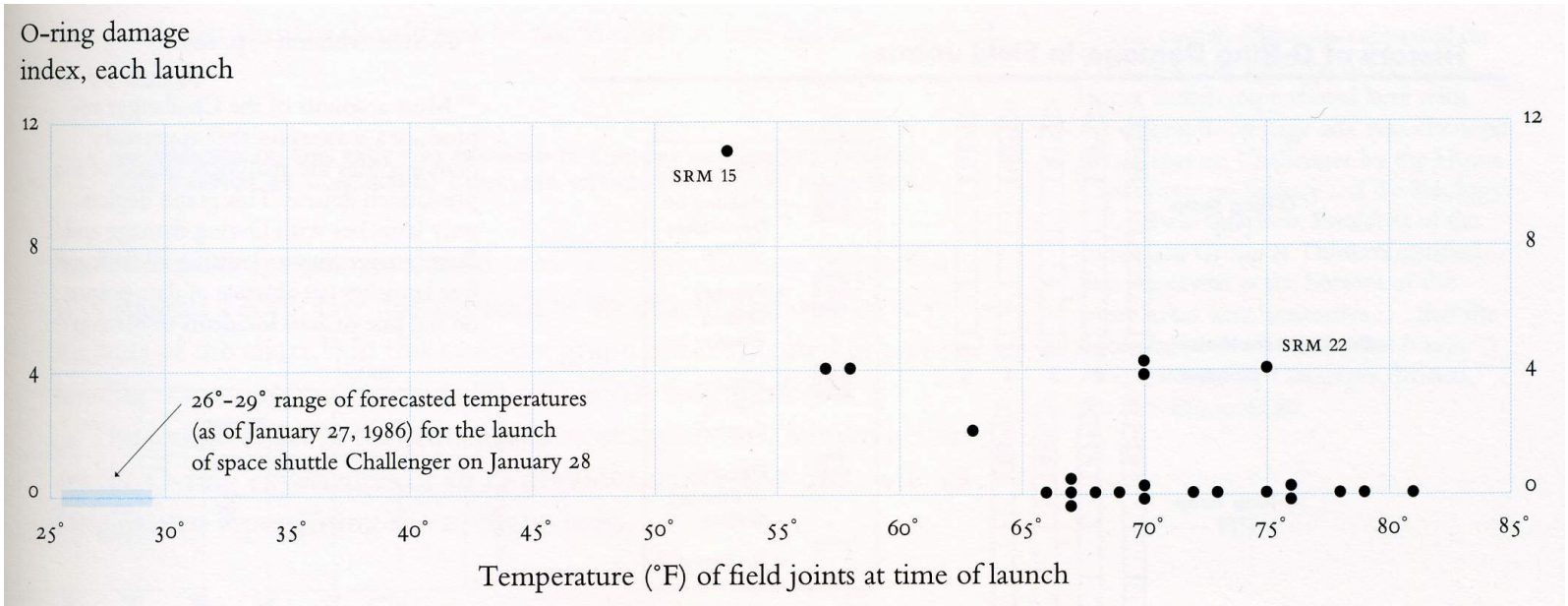

The correlation between the damage index and the temperature is $-0.64$
(What does this mean?)

# The *Challenger* launch decision

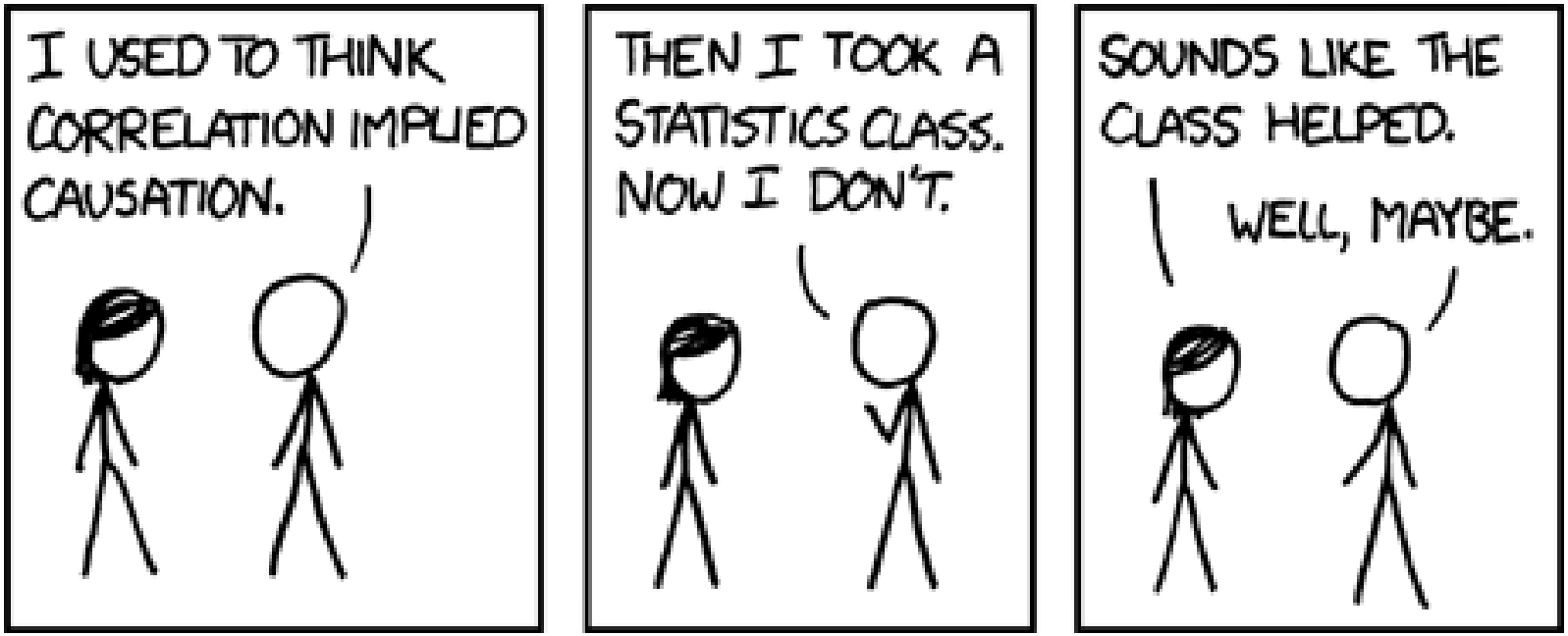
What was the forecast temperature for launch?

# The *Challenger* launch decision

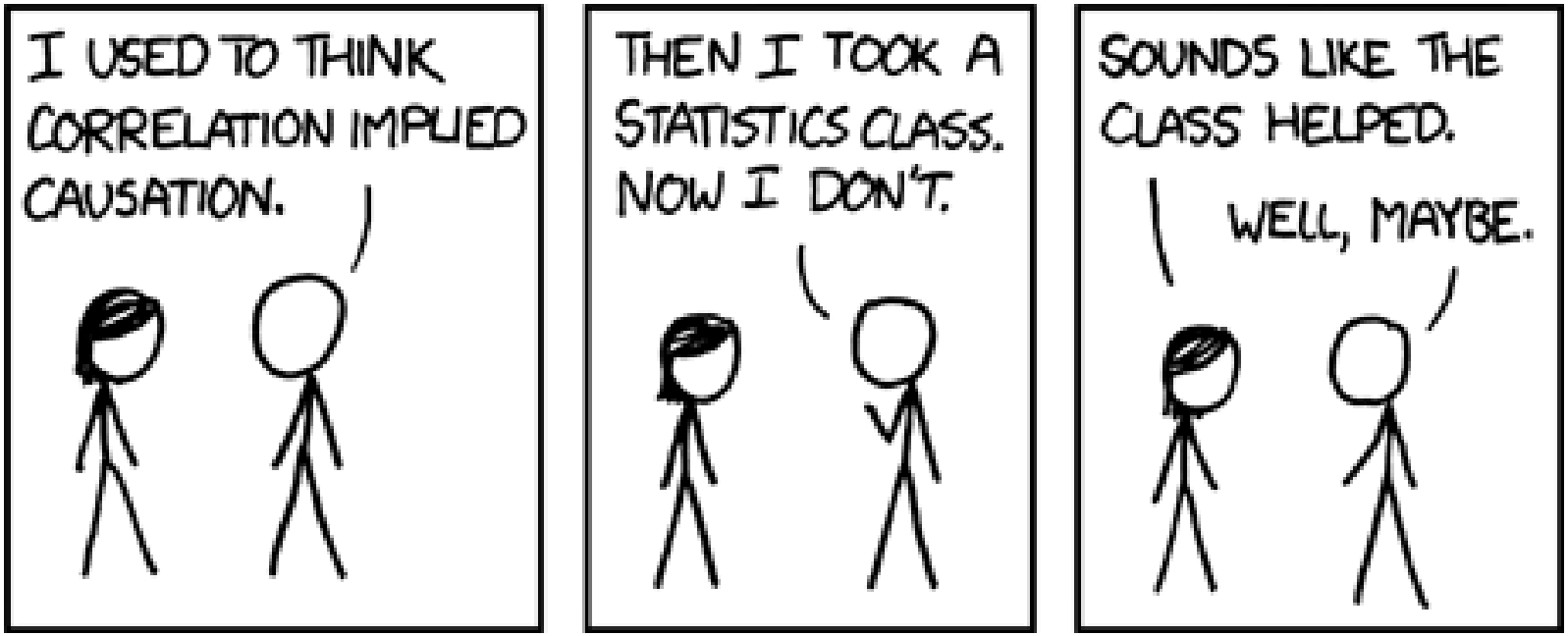
What was the forecast temperature for launch? 26 to 29 degrees Fahrenheit!




The shuttle was launched in unprecendented cold




http://xkcd.com/552/




http://xkcd.com/552/

Correlation doesn't always imply causation, but it can be a big clue. . .

# The Line of Best Fit

In the next example, we refine our scatterplots by adding a line of best fit

This line is produced by a technique called *linear regression*

Major focus of the last two weeks of 221

Key for today: understanding what a *regression coefficient* is,
and how it differs from a correlation coefficient

## Cross-national fertility Example

We have cross-national data from several sources:

**Fertility** The average number of children born per adult female, in 2000 (United Nations)

**Education Ratio** The ratio of girls to boys in primary and secondary education, in 2000 (Word Bank Development Indicators)

**GDP per capita** Economic activity in thousands of dollars, purchasing power parity in 2000 (Penn World Tables)

What are the levels of measurement of these variables?

Our question: how are these variables related to each other?

# Example: Fertility, Female Education, and Development

Specifically, we ask:

# Example: Fertility, Female Education, and Development

Specifically, we ask:

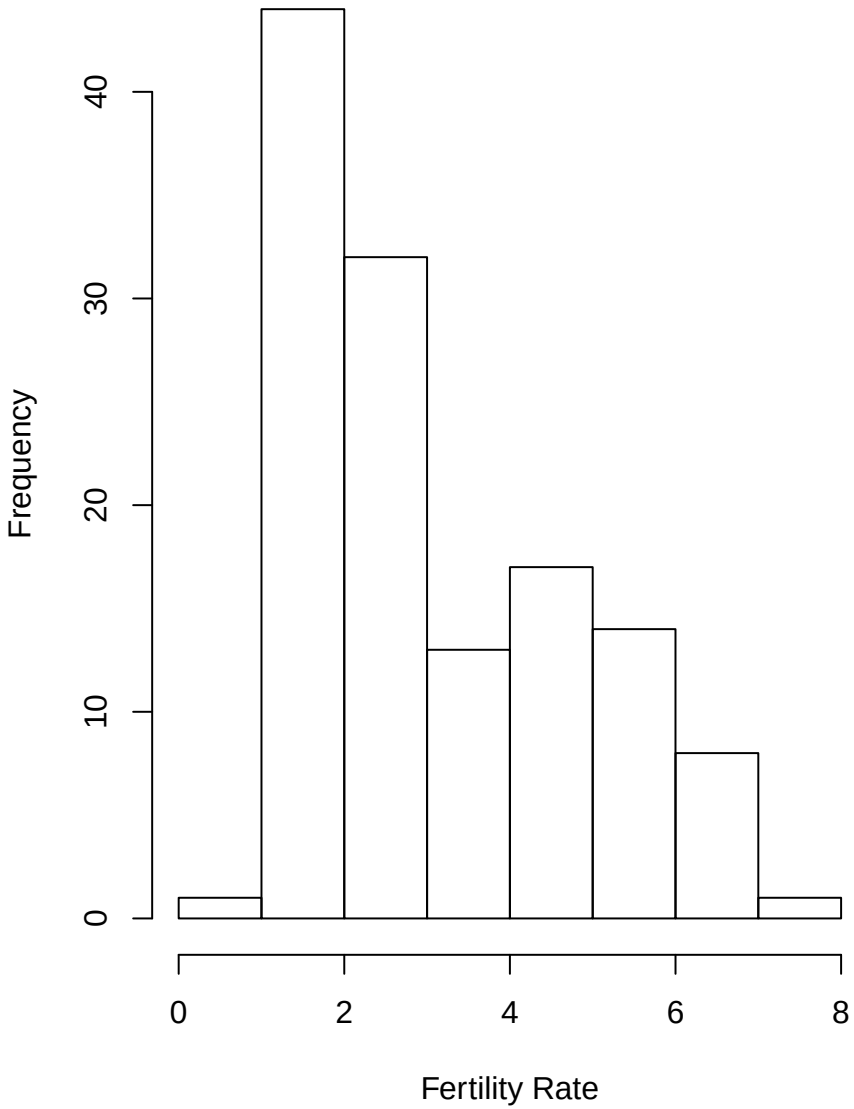
- If the level of female education changed by a certain amount, how much would we expect Fertility to change?

# Example: Fertility, Female Education, and Development

Specifically, we ask:

- If the level of female education changed by a certain amount, how much would we expect Fertility to change?
- If the level of GDP per capita changed by a certain amount, how much would we expect Fertility to change?

# Summary of Univariate Distribution: Fertility

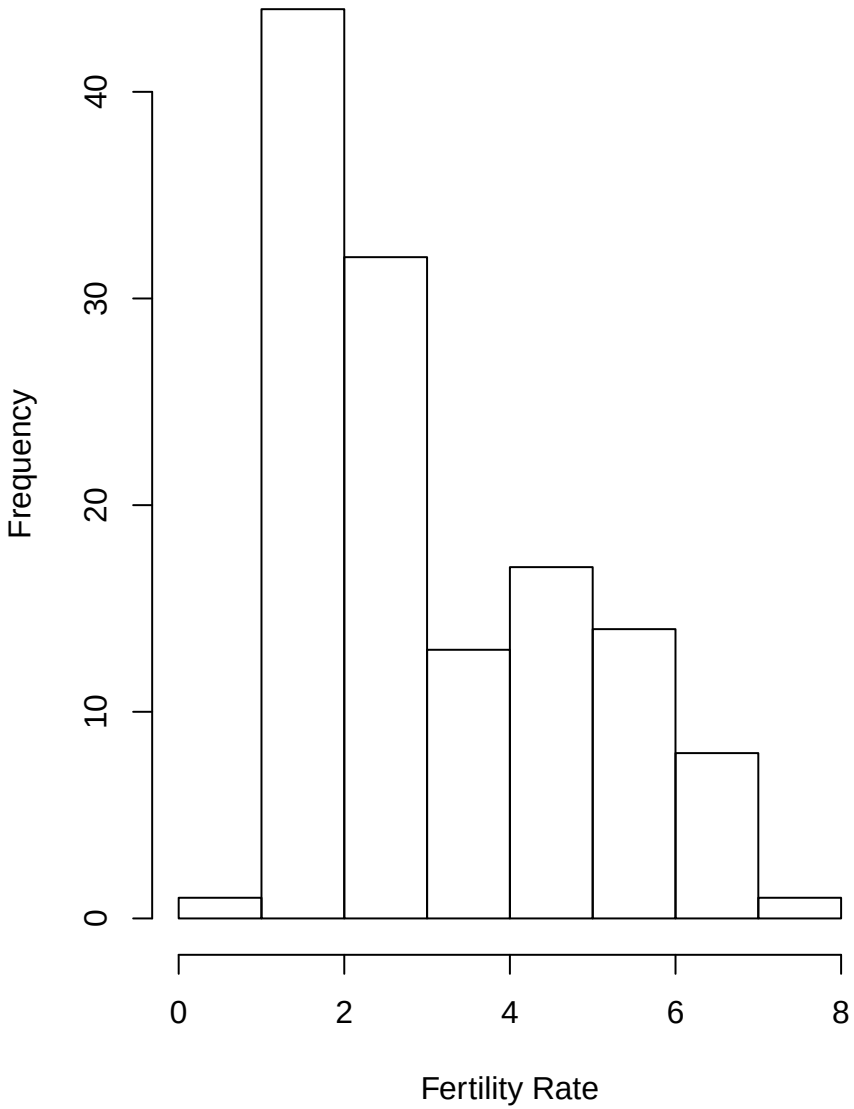



Median = 2.60

Mean = 3.12 children

std dev = 1.67 children

# Summary of Univariate Distribution: Fertility



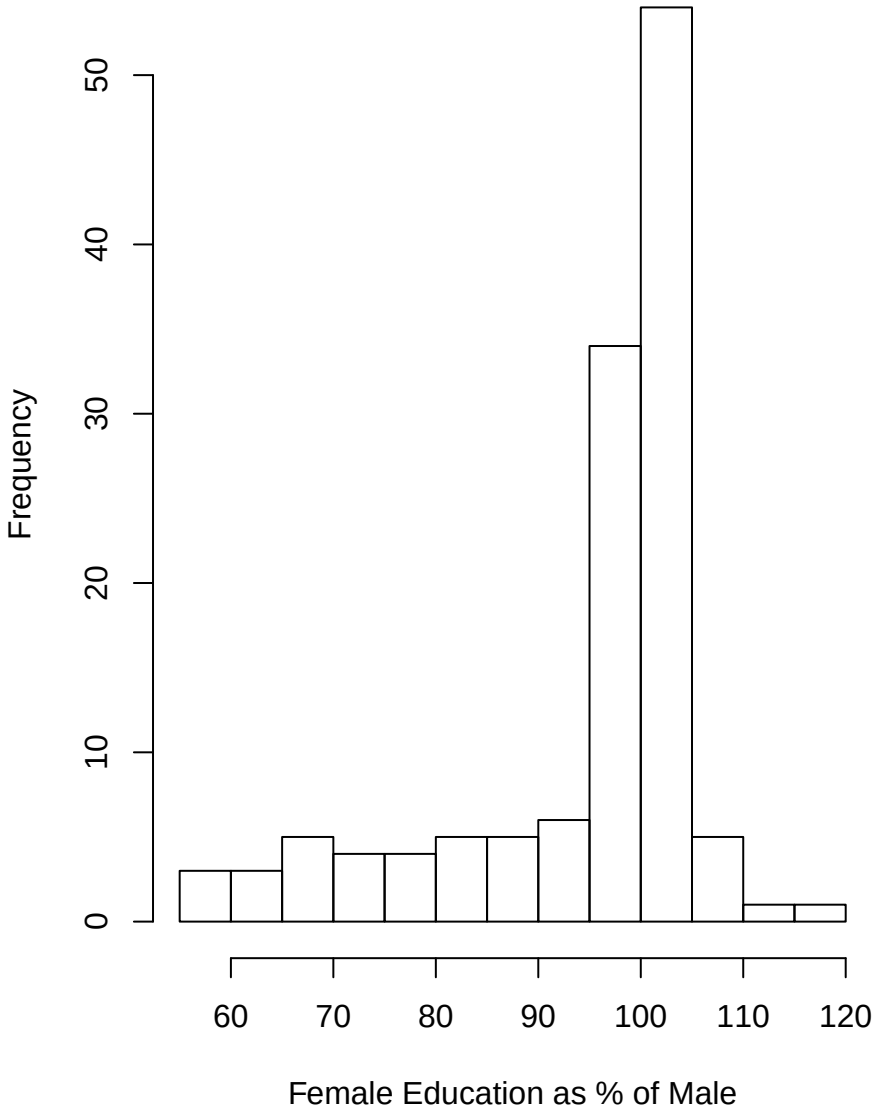

Median = 2.60

Mean = 3.12 children

std dev = 1.67 children

How would you describe this distribution?

# Summary of Univariate Distribution: Education Ratio

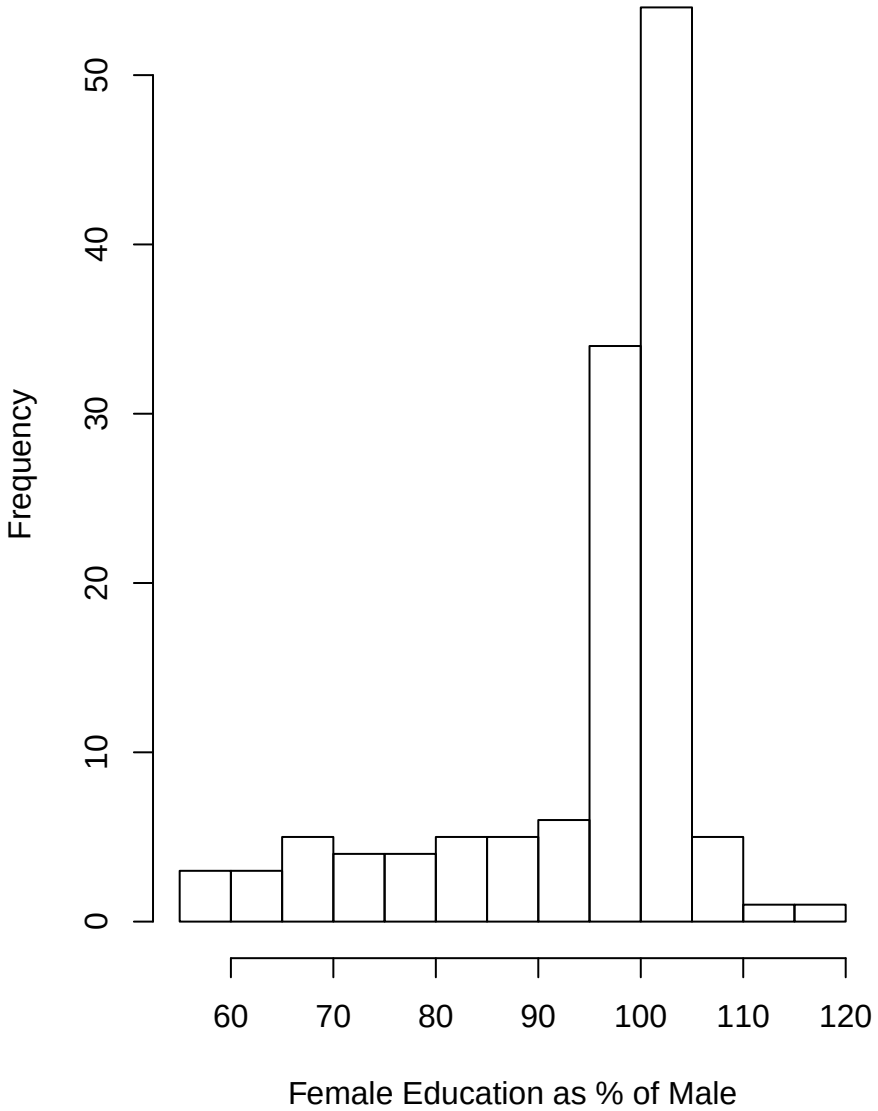



Median = 99.60%

Mean = 94.48%

std. dev. = 12.45%

# Summary of Univariate Distribution: Education Ratio



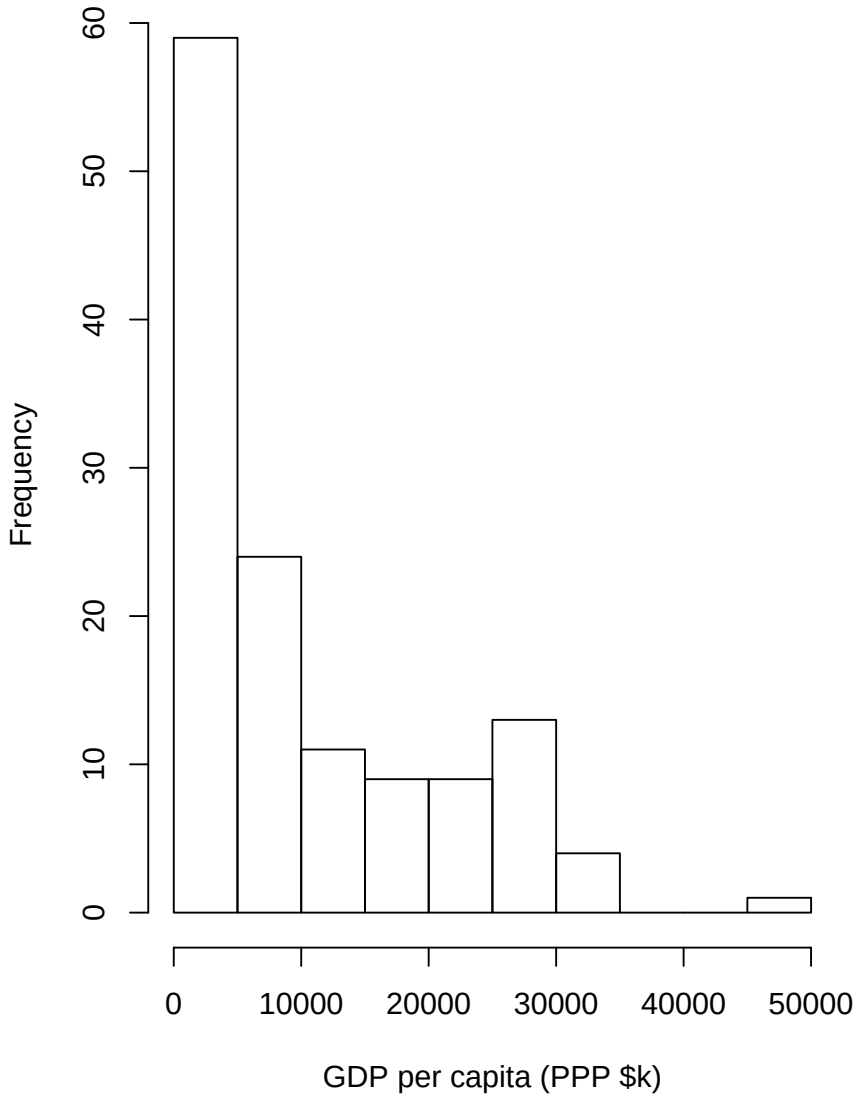

Median = 99.60%

Mean = 94.48%

std. dev. = 12.45%

How would you describe this distribution?

# Summary of Univariate Distribution: GDP per capita

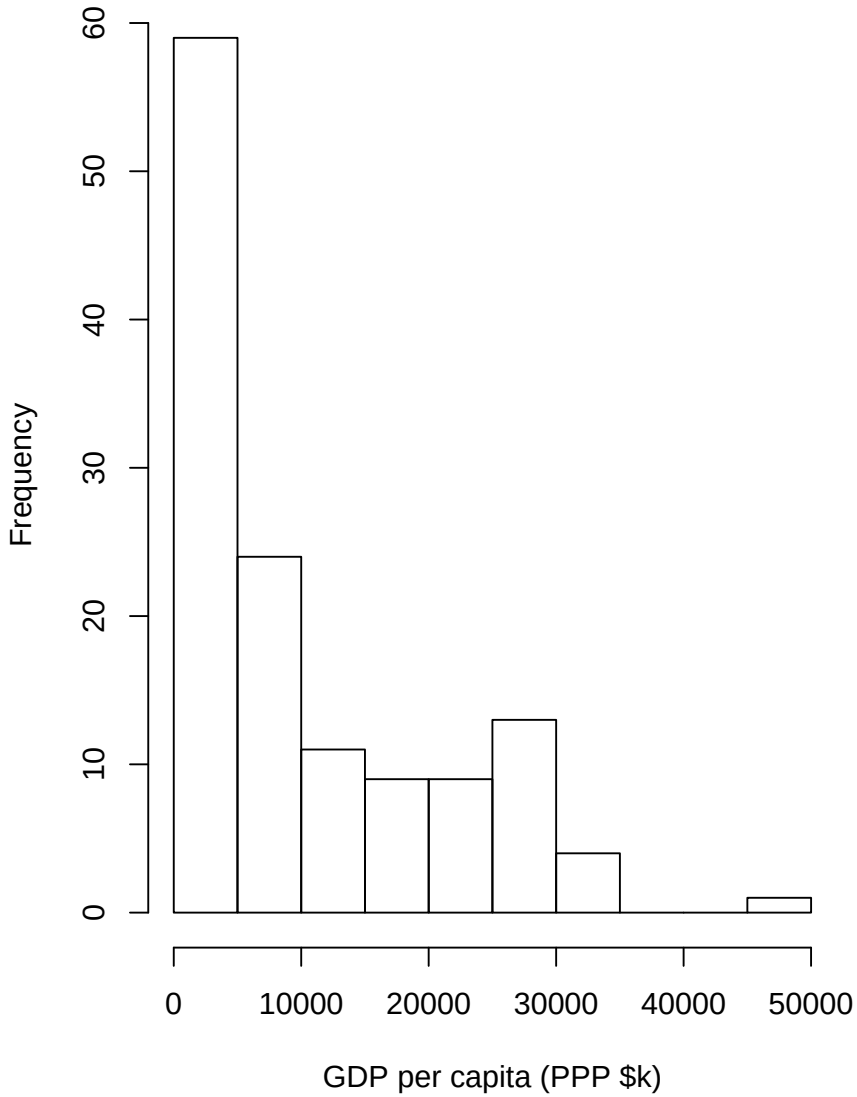



Median = $6047

Mean = $10,200

std. dev. = $10,078

# Summary of Univariate Distribution: GDP per capita

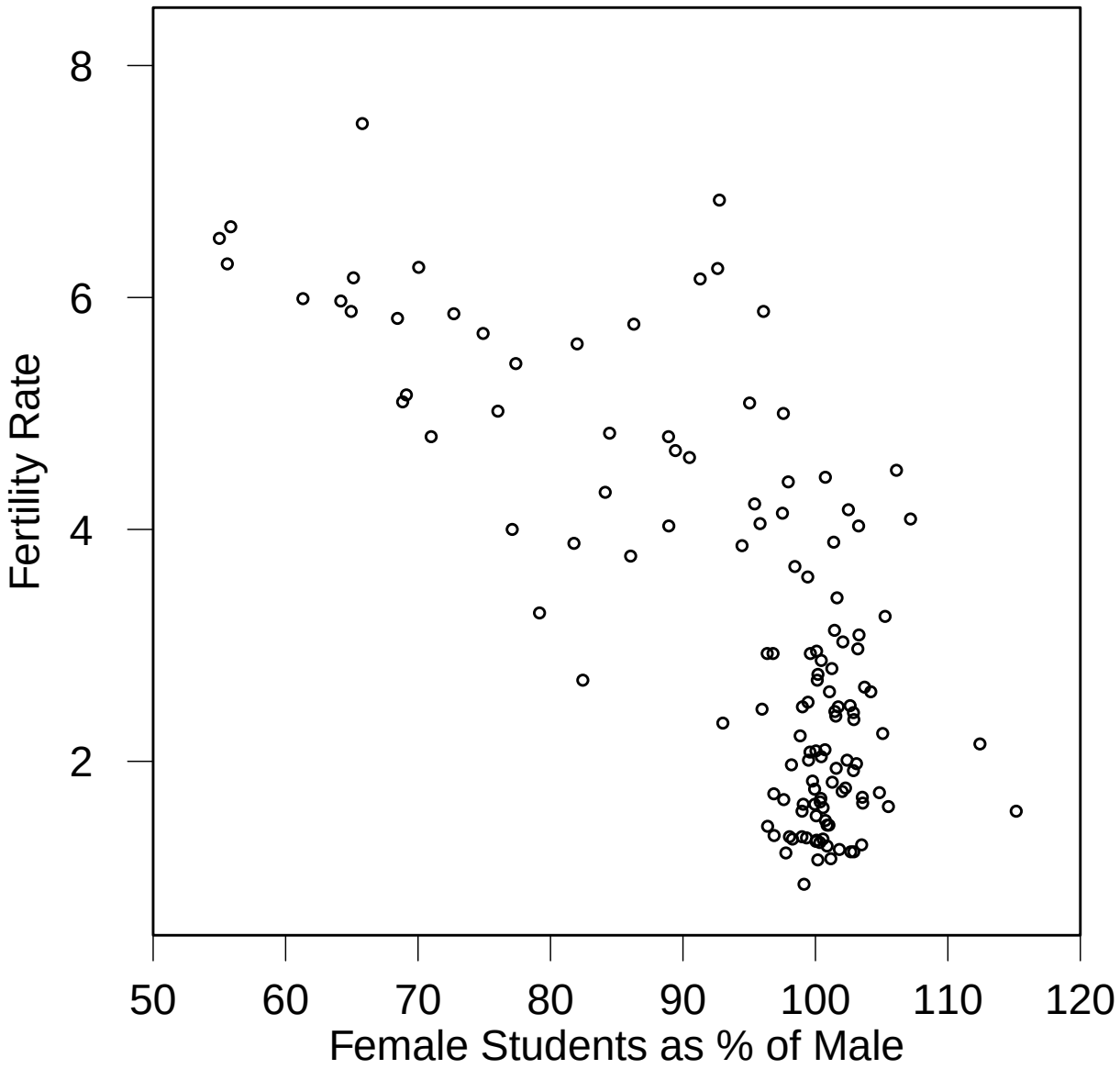



Median = $6047

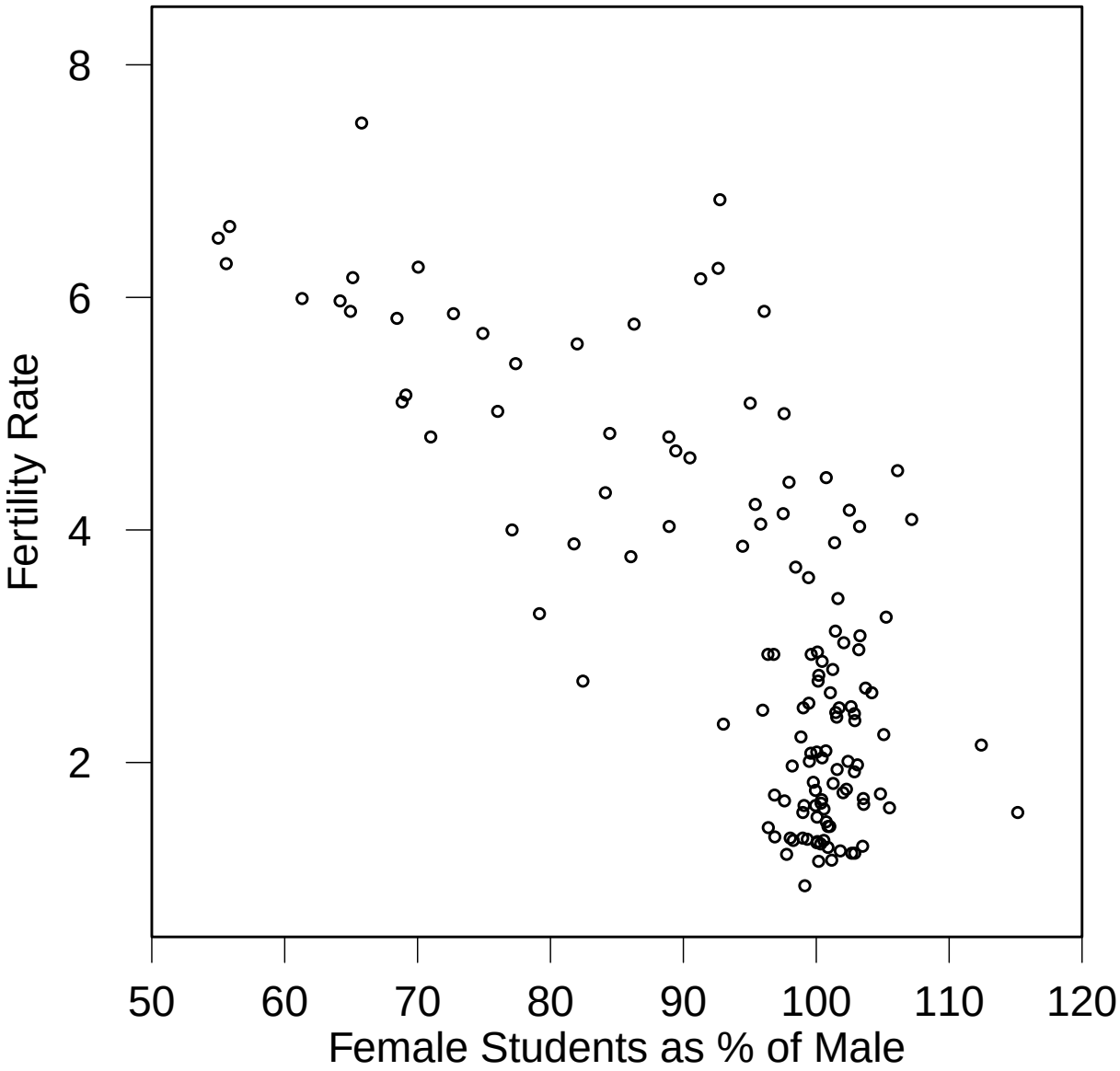
Mean = $10,200

std. dev. = $10,078

How would you describe this distribution?

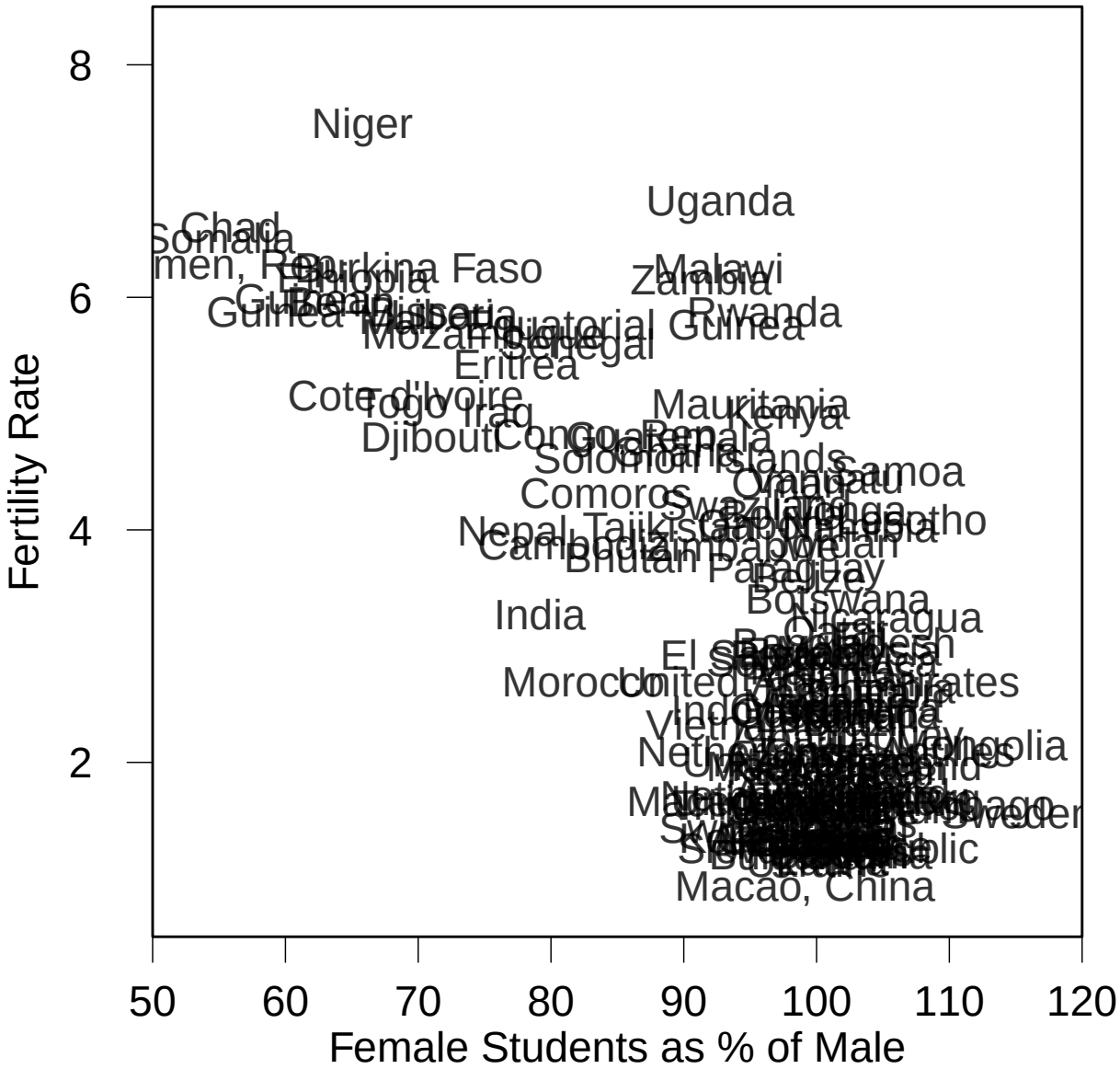



How would you describe the relationship between Fertility & Education Ratio?

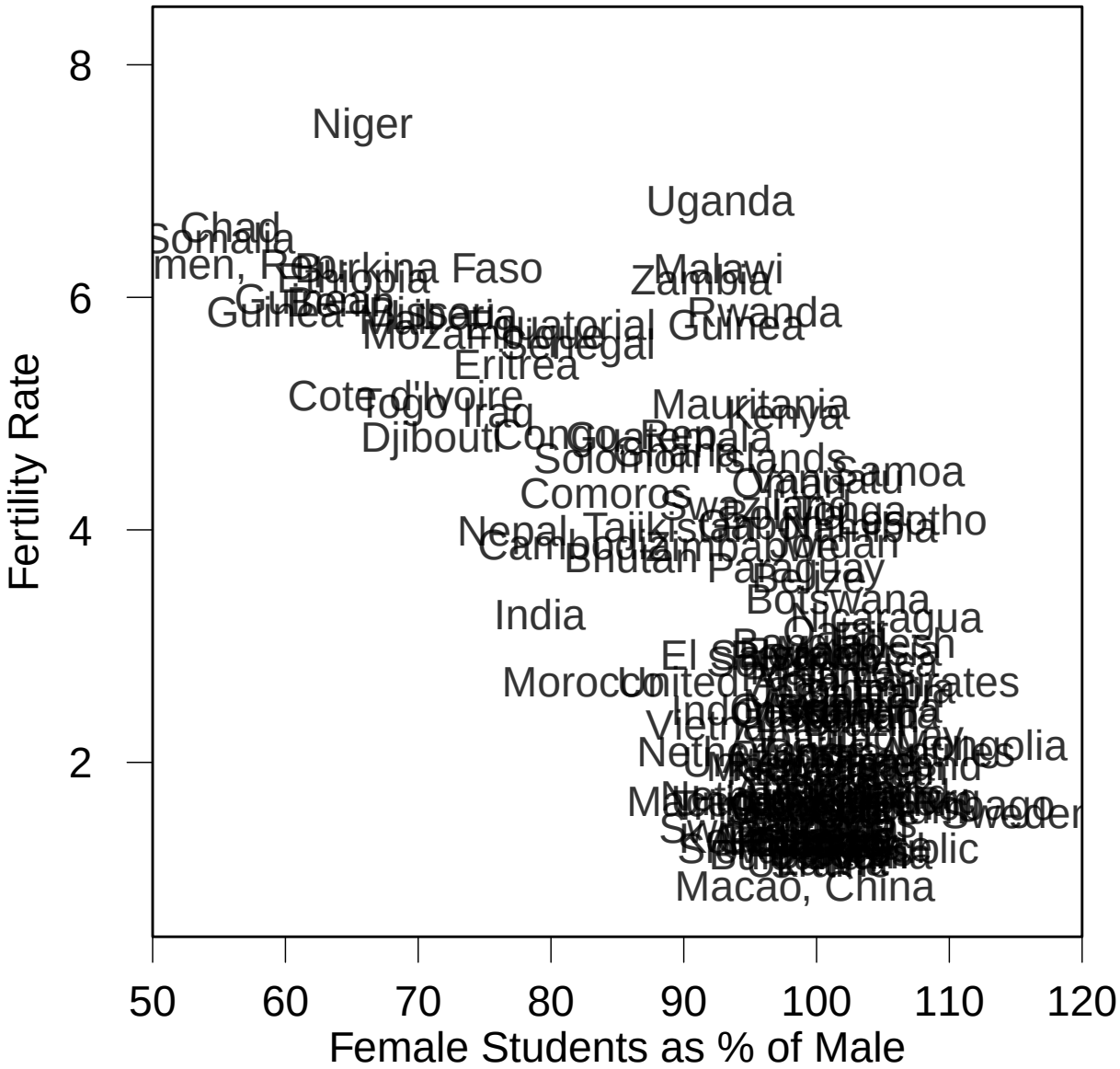



How would you describe the relationship between Fertility & Education Ratio?

If I asked you to predict Fertility for a country not sampled, how accurate do you expect your prediction to be?

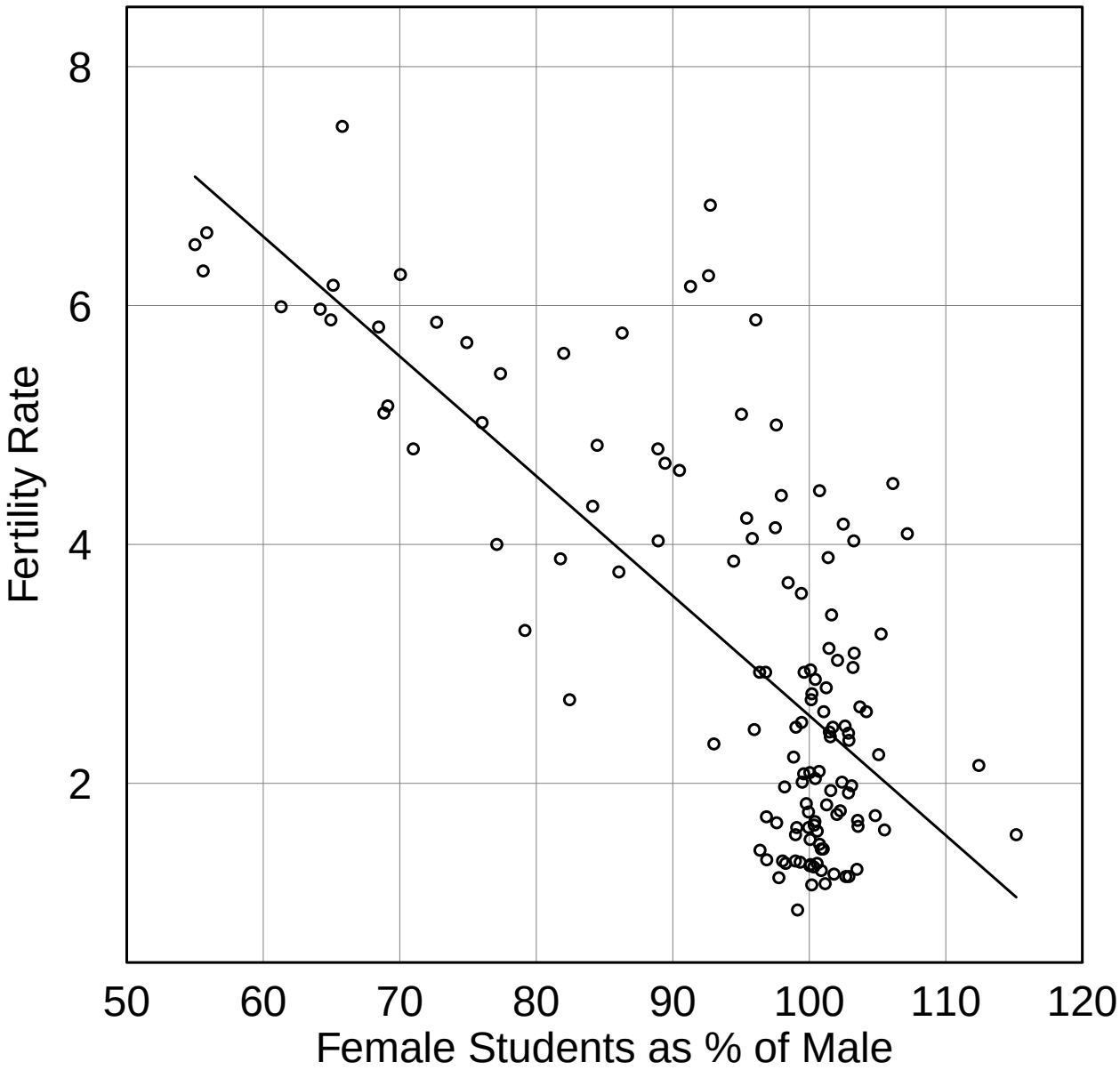



Labelling cases sometimes helps, especially for identifying *outliers*

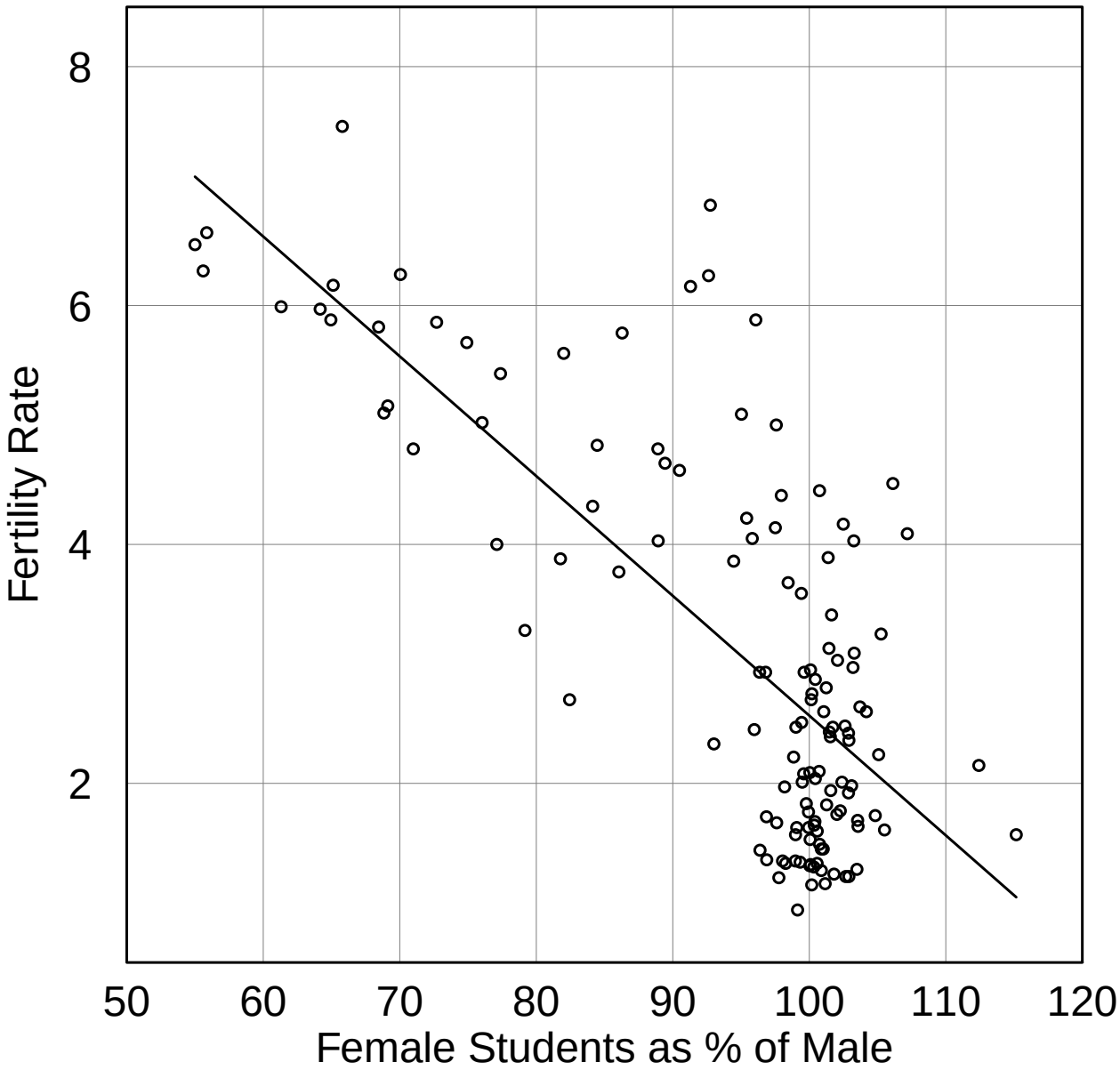



Labelling cases sometimes helps, especially for identifying *outliers*

What makes a point an outlier?




The best fit line is the line that passes closest to the majority of the points




The best fit line is the line that passes closest to the majority of the points

If we take this line to be our model of Fertility, how do we interpret it?

## Best fit lines

From high school math, a line on a plane follows this equation:

$$y = b + mx$$

where:

- $y$ is the dependent variable,
- $x$ is the independent variable,
- $m$ is the slope of the line,
  or the change in $y$ for a 1 unit change in $x$,
- and $b$ is the intercept,
  or value of $y$ when $x = 0$

## Best fit lines

Customarily, in statistics, we write the equation of a line as:

$$y = \beta_0 + \beta_1 x$$

where:

- $y$ is the dependent variable
- $x$ is the independent variable,
- $\beta_1$ is a regression coefficient. It conveys the slope of the line, or the change in $y$ for a 1 unit change in $x$,
- and $\beta_0$ is the intercept, or value of $y$ when $x = 0$

# Best fit for fertility against education ratio

$$\begin{aligned} \widehat{\text{Fertility}} &= \hat{\beta}_0 + \hat{\beta}_1 \text{EduRatio} \\ \widehat{\text{Fertility}} &= 12.59 - 0.10 \times \text{EduRatio} \end{aligned}$$

The above equation is the best fit line given by *linear regression*

The $\hat{\beta}$'s are the estimated linear regression *coefficients*

$\widehat{\text{Fertility}}$ is the *fitted value*, or model prediction, of the level of Fertility given the EduRatio

# Intrepreting regression coefficients

$$\begin{aligned} \widehat{\text{Fertility}} &= \hat{\beta}_0 + \hat{\beta}_1 \text{EduRatio} \\ \widehat{\text{Fertility}} &= 12.59 - 0.10 \times \text{EduRatio} \end{aligned}$$

Interpreting $\hat{\beta}_1 = -0.10$:

*Increasing EduRatio by 1 unit lowers Fertility by 0.10 units.*

Because EduRatio is measured in percentage points, this means a 10% increase in female education (relative to males) will lower the number of children a woman has over her lifetime by 1 on average.

# Intrepreting regression intercepts

$$\begin{aligned} \widehat{\text{Fertility}} &= \hat{\beta}_0 + \hat{\beta}_1 \text{EduRatio} \\ \widehat{\text{Fertility}} &= 12.59 - 0.10 \times \text{EduRatio} \end{aligned}$$

Interpreting $\hat{\beta}_0 = 12.59$:

*If EduRatio is 0, Fertility will be 12.59.*

If there are no girls in primary or secondary education, then women are expected to have 12.59 children on average over their lifetimes.

Can we trust this prediction?

# Intrepreting regression intercepts

$$\begin{aligned} \widehat{\text{Fertility}} &= \hat{\beta}_0 + \hat{\beta}_1 \text{EduRatio} \\ \widehat{\text{Fertility}} &= 12.59 - 0.10 \times \text{EduRatio} \end{aligned}$$

Interpreting $\hat{\beta}_0 = 12.59$:

*If EduRatio is 0, Fertility will be 12.59.*

If there are no girls in primary or secondary education, then women are expected to have 12.59 children on average over their lifetimes.

Can we trust this prediction? No.
No country has 0 female education, so this is an *extrapolation* from the model.

# Using regression coefficients to predict specific cases

$$\begin{aligned} \widehat{\text{Fertility}} &= \hat{\beta}_0 + \hat{\beta}_1 \text{EduRatio} \\ \widehat{\text{Fertility}} &= 12.59 - 0.10 \times \text{EduRatio} \end{aligned}$$

How many children do we expect women to have if girls get half the education boys do?

*If EduRatio is 50, Fertility will be* $12.59 - 0.10 \times 50 = 7.59$*.*

How many children do we expect women to have if girls get the same education boys do?

*If EduRatio is 100, Fertility will be* $12.59 - 0.10 \times 100 = 2.59$*.*

# Using regression coefficients to predict specific cases

$$\begin{aligned} \widehat{\text{Fertility}} &= \hat{\beta}_0 + \hat{\beta}_1 \text{EduRatio} \\ \widehat{\text{Fertility}} &= 12.59 - 0.10 \times \text{EduRatio} \end{aligned}$$

*If EduRatio is 100, Fertility will be* $12.59 - 0.10 \times 100 = 2.59$*.*

Does this hold exactly for any country with education parity?

# Using regression coefficients to predict specific cases

$$\begin{aligned} \widehat{\text{Fertility}} &= \hat{\beta}_0 + \hat{\beta}_1 \text{EduRatio} \\ \widehat{\text{Fertility}} &= 12.59 - 0.10 \times \text{EduRatio} \end{aligned}$$

*If EduRatio is 100, Fertility will be* $12.59 - 0.10 \times 100 = 2.59$.

Does this hold exactly for any country with education parity?

No. It holds on average. In any specific case $i$, there is some error between the expected and actual levels of Fertility

## What's the difference between correlation coefficients and regression coefficients

The correlation coefficient ($r$) measures the strength of relationship between $X$ and $Y$

Works in both directions

$[-1, 1]$ scale (standardized)

The regression coefficient ($\beta$) measures the substance of the relationship

Tells us how much $Y$ increases for a one-unit increase in $X$

One direction, and can take on any value

## Contrasting $r$ and $\beta$

Low $r$ between Fertility and Education Ratio, for example, would tell us that many other random factors besides female education intervene in causing Fertility in a particular case

High $r$ would tell us that few stochastic factors intervene in any particular case. (In this case, $r = -0.75$, which is "high" in absolute value)

Low $\beta$ would tell us that it takes a lot of female education to lower Fertility, on average

High $\beta$ would tell us that a little bit of female education lowers Fertility a lot, on average

# Tabular presentations of covariation

Scatterplots are great for showing the relationship between continuous variables

But potentially misleading if variables are discrete

What if we can only order the categories of variables, but lack additive scales?

What if we don't even know the order?

A table of one variable against another will help investigate even unordered variables

# Example: Education & Partisan Identification

We have two variables from the General Social Survey:

**Education** Highest degree attained: No degree, High School diploma, Associates Degree, Bachelors Degree, Graduate Degree

**Party Identification** Strong Democrat, Democrat, Leans Democratic, Independent, Leans Republican, Republican, Strong Republican, Other

We take these data from the 1990 and 2006 samples of the GSS

What is the level of measurement of these variables?

How can we ascertain the relationship between them?

# Monotonicity

Monotonic relationships are those which either consistently move in the same direction, or at least "stay still":

- If adding years of education always increases the *expected* probability one is Republican, or at least never lowers it, then Republican ID is *monotonically increasing* in Education
- If adding years of education always decreases the *expected* probability one is Republican, or at least never raises it, then Republican ID is *monotonically decreasing* in Education
- If adding years of education at first raises the expected probability of Republican ID, but then lowers it (or vice versa), the relationship is *non-monotonic*

# Constructing a contingency table

The simplest way to explore the relationship between two discrete variables is a *contingency table*:

1. We consider every possible combination of education and party ID
2. Total up all subjects with that combination
3. Enter the sum in a *cross-tabulation*, with one variable's categories as the columns, and the other variable's categories as the rows
4. Customarily, the "dependent variable" (to the extent we believe one variable depends on the other) is the row variable

## 2006 General Social Survey: Partisanship & Education

| | | Highest Degree Attained | | | | | |
|---|---|---|---|---|---|---|---|
| | | None | HS | Assoc | College | Grad | Sum |
| Party ID | Str. Dem. | 97 | 347 | 54 | 110 | 91 | 699 |
| | Dem. | 115 | 384 | 52 | 116 | 69 | 736 |
| | Lean Dem. | 67 | 265 | 50 | 87 | 58 | 527 |
| | Indep. | 263 | 503 | 86 | 92 | 53 | 997 |
| | Lean Rep. | 39 | 168 | 28 | 60 | 32 | 327 |
| | Rep. | 56 | 307 | 64 | 158 | 52 | 637 |
| | Str. Rep. | 40 | 256 | 37 | 118 | 44 | 495 |
| | Other | 9 | 32 | 3 | 18 | 3 | 65 |
| | Sum | 686 | 2262 | 374 | 759 | 402 | 4483 |

The above is a *contingency table* or *cross-tabulation*.

Powerful way to get the data. Can be tweaked to be more informative.

## 2006 GSS: Collapse partisans, treat leaners as independent

| | | Highest Degree Attained | | | | | |
|---|---|---|---|---|---|---|---|
| | | None | HS | Assoc | College | Grad | Sum |
| Party ID | Democrat | 212 | 731 | 106 | 226 | 160 | 1435 |
| | Independent | 369 | 936 | 164 | 239 | 143 | 1851 |
| | Republican | 96 | 563 | 101 | 276 | 96 | 1132 |
| | Other | 9 | 32 | 3 | 18 | 3 | 65 |
| | Sum | 686 | 2262 | 374 | 759 | 402 | 4483 |

The first thing we will do is collapse some similar categories

Create **Democrat** out of the old “Strong Democrat” and “Democrat”

Create **Indepedent** out of the old “Leans Democratic”, “Independent”, and “Leans Republican”

Create **Republican** out of the old “Strong Republican” and “Republican”

## 2006 GSS: Collapse partisans, treat leaners as independent

| | | Highest Degree Attained | | | | | |
|---|---|---|---|---|---|---|---|
| | | None | HS | Assoc | College | Grad | Sum |
| Party ID | Democrat | 212 | 731 | 106 | 226 | 160 | 1435 |
| | Independent | 369 | 936 | 164 | 239 | 143 | 1851 |
| | Republican | 96 | 563 | 101 | 276 | 96 | 1132 |
| | Other | 9 | 32 | 3 | 18 | 3 | 65 |
| | Sum | 686 | 2262 | 374 | 759 | 402 | 4483 |

Consolidation of categories reduces noise in each cell, but at a price: we've lost some of the fine-grained nature of our data

Introduces a trade-off between borrowing strength by pooling cells and informative measuremnt

Tabular methods pose this dilemma when applied to detailed ordered variables

## 2006 GSS: Collapse partisans, treat leaners as independent

| | | Highest Degree Attained | | | | | |
|---|---|---|---|---|---|---|---|
| | | None | HS | Assoc | College | Grad | Sum |
| Party ID | Democrat | 212 | 731 | 106 | 226 | 160 | 1435 |
| | Independent | 369 | 936 | 164 | 239 | 143 | 1851 |
| | Republican | 96 | 563 | 101 | 276 | 96 | 1132 |
| | Other | 9 | 32 | 3 | 18 | 3 | 65 |
| | Sum | 686 | 2262 | 374 | 759 | 402 | 4483 |

Collapsing Party ID has simplified our table, but it's still hard to see the relationship between the variables

What could we do?

## 2006 GSS: Collapse partisans, treat leaners as independent

| | | Highest Degree Attained | | | | | |
|---|---|---|---|---|---|---|---|
| | | None | HS | Assoc | College | Grad | Sum |
| Party ID | Democrat | 212 | 731 | 106 | 226 | 160 | 1435 |
| | Independent | 369 | 936 | 164 | 239 | 143 | 1851 |
| | Republican | 96 | 563 | 101 | 276 | 96 | 1132 |
| | Other | 9 | 32 | 3 | 18 | 3 | 65 |
| | Sum | 686 | 2262 | 374 | 759 | 402 | 4483 |

Collapsing Party ID has simplified our table, but it's still hard to see the relationship between the variables

What could we do? Perhaps percentages would be easier?

Let's divide by $N = 4483$, the total number of observations

## 2006 GSS: Percent of $N$

| | | Highest Degree Attained | | | | | |
|---|---|---|---|---|---|---|---|
| | | None | HS | Assoc | College | Grad | Sum |
| Party ID | Democrat | 4.7% | 16.3% | 2.4% | 5.0% | 3.6% | 32.0% |
| | Independent | 8.2 | 20.9 | 3.7 | 5.3 | 3.2 | 41.3 |
| | Republican | 2.1 | 12.6 | 2.3 | 6.2 | 2.1 | 25.3 |
| | Other | 0.2 | 0.7 | 0.1 | 0.4 | 0.1 | 1.4 |
| | Sum | 15.3 | 50.5 | 8.3 | 16.9 | 9.0 | 100.0 |

Does this help?

## 2006 GSS: Percent of $N$

| | | Highest Degree Attained | | | | | |
|---|---|---|---|---|---|---|---|
| | | None | HS | Assoc | College | Grad | Sum |
| Party ID | Democrat | 4.7% | 16.3% | 2.4% | 5.0% | 3.6% | 32.0% |
| | Independent | 8.2 | 20.9 | 3.7 | 5.3 | 3.2 | 41.3 |
| | Republican | 2.1 | 12.6 | 2.3 | 6.2 | 2.1 | 25.3 |
| | Other | 0.2 | 0.7 | 0.1 | 0.4 | 0.1 | 1.4 |
| | Sum | 15.3 | 50.5 | 8.3 | 16.9 | 9.0 | 100.0 |

Does this help? Not really. It's still hard to see the effects of each variable *separately*

We see that the combination of Democrat and High School is common, and Republican and College is rare, but does that mean there is an association?

That is, does being College educated make one less likely to be Republican? Or is it just that there are more High School grads than College grads?

## 2006 GSS: Percent of $N$

| | | Highest Degree Attained | | | | | |
|---|---|---|---|---|---|---|---|
| | | None | HS | Assoc | College | Grad | Sum |
| Party ID | Democrat | 4.7% | 16.3% | 2.4% | 5.0% | 3.6% | 32.0% |
| | Independent | 8.2 | 20.9 | 3.7 | 5.3 | 3.2 | 41.3 |
| | Republican | 2.1 | 12.6 | 2.3 | 6.2 | 2.1 | 25.3 |
| | Other | 0.2 | 0.7 | 0.1 | 0.4 | 0.1 | 1.4 |
| | Sum | 15.3 | 50.5 | 8.3 | 16.9 | 9.0 | 100.0 |

What can we do to zero in on the likelihood that one is Republican given that one has a College Degree?

That is, how do we estimate the conditional probability $\Pr(\text{Republican}|\text{College})$?

## 2006 GSS: Percent of $N$

| | | Highest Degree Attained | | | | | |
|---|---|---|---|---|---|---|---|
| | | None | HS | Assoc | College | Grad | Sum |
| Party ID | Democrat | 4.7% | 16.3% | 2.4% | 5.0% | 3.6% | 32.0% |
| | Independent | 8.2 | 20.9 | 3.7 | 5.3 | 3.2 | 41.3 |
| | Republican | 2.1 | 12.6 | 2.3 | 6.2 | 2.1 | 25.3 |
| | Other | 0.2 | 0.7 | 0.1 | 0.4 | 0.1 | 1.4 |
| | Sum | 15.3 | 50.5 | 8.3 | 16.9 | 9.0 | 100.0 |

What can we do to zero in on the likelihood that one is Republican given that one has a College Degree?

That is, how do we estimate the conditional probability $\Pr(\text{Republican}|\text{College})$?

How about the percentage of College grads that vote Republican in the sample?

## 2006 GSS: Column percentages

| | | Highest Degree Attained | | | | | |
|---|---|---|---|---|---|---|---|
| | | None | HS | Assoc | College | Grad | Sum |
| Party ID | Democrat | 30.9% | 32.3% | 28.3% | 29.8% | 39.8% | 32.0% |
| | Independent | 53.8 | 41.4 | 43.9 | 31.5 | 35.6 | 41.3 |
| | Republican | 14.0 | 24.9 | 27.0 | 36.4 | 23.9 | 25.3 |
| | Other | 1.3 | 1.4 | 0.8 | 2.4 | 0.7 | 1.4 |
| | Sum | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

How about the percentage of College grads that vote Republican in the sample?

That is, what if we divide each *column* by its sum, to see how people with a given level of the column variable Education get distributed on the row variable, Partisan ID?

This is called showing "column percentages". Most useful presentation of a cross-tab

## 2006 GSS: Column percentages

| | | Highest Degree Attained | | | | | |
|---|---|---|---|---|---|---|---|
| | | None | HS | Assoc | College | Grad | Sum |
| Party ID | Democrat | 30.9% | 32.3% | 28.3% | 29.8% | 39.8% | 32.0% |
| | Independent | 53.8 | 41.4 | 43.9 | 31.5 | 35.6 | 41.3 |
| | Republican | 14.0 | 24.9 | 27.0 | 36.4 | 23.9 | 25.3 |
| | Other | 1.3 | 1.4 | 0.8 | 2.4 | 0.7 | 1.4 |
| | Sum | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

Notice that with column percentages, each column sums to 100%

The interesting comparisons appear when we look *across* each row

For each row, higher values show positive relationships between that column category and the current row.

Low values within the row show negative relationships between the column category and the current row.

## 2006 GSS: Column percentages

| | | Highest Degree Attained | | | | | |
|---|---|---|---|---|---|---|---|
| | | None | HS | Assoc | College | Grad | Sum |
| Party ID | Democrat | 30.9% | 32.3% | 28.3% | 29.8% | 39.8% | 32.0% |
| | Independent | 53.8 | 41.4 | 43.9 | 31.5 | 35.6 | 41.3 |
| | Republican | 14.0 | 24.9 | 27.0 | 36.4 | 23.9 | 25.3 |
| | Other | 1.3 | 1.4 | 0.8 | 2.4 | 0.7 | 1.4 |
| | Sum | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

In this example, $\Pr(\text{Democrat})$ is higher for those without high school diplomas or with graduate degrees, but lower for those with college degrees

Republicans do best among College degree holders, and worse at the ends of the Education spectrum

That is, support for either party seems to be a *non-monotonic* function of Education

## 2006 GSS: Column percentages

| | | Highest Degree Attained | | | | | |
|---|---|---|---|---|---|---|---|
| | | None | HS | Assoc | College | Grad | Sum |
| Party ID | Democrat | 30.9% | 32.3% | 28.3% | 29.8% | 39.8% | 32.0% |
| | Independent | 53.8 | 41.4 | 43.9 | 31.5 | 35.6 | 41.3 |
| | Republican | 14.0 | 24.9 | 27.0 | 36.4 | 23.9 | 25.3 |
| | Other | 1.3 | 1.4 | 0.8 | 2.4 | 0.7 | 1.4 |
| | Sum | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

Notice that comparisons *across* rows in the column percentage cross-tab mean something different from comparisons across rows

For instance, Democrats do almost as well as Republicans in the strongest Republican category, College.

Why? College grads are more likely to be Republicans than any other education group. *But* more people on average are Dems, so even in this relatively weak category, Dems are fairly strong

## 2006 GSS: Column percentages

| | | Highest Degree Attained | | | | | |
|---|---|---|---|---|---|---|---|
| | | None | HS | Assoc | College | Grad | Sum |
| Party ID | Democrat | 30.9% | 32.3% | 28.3% | 29.8% | 39.8% | 32.0% |
| | Independent | 53.8 | 41.4 | 43.9 | 31.5 | 35.6 | 41.3 |
| | Republican | 14.0 | 24.9 | 27.0 | 36.4 | 23.9 | 25.3 |
| | Other | 1.3 | 1.4 | 0.8 | 2.4 | 0.7 | 1.4 |
| | Sum | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

What if you encounter a cross-tab "in the field"?

Check if it's in column percentages, then start looking for patterns in each row

Remember this mantra: **Sum Down, Compare Across**

## 2006 GSS: Row percentages

| | | Highest Degree Attained | | | | | |
|---|---|---|---|---|---|---|---|
| | | None | HS | Assoc | College | Grad | Sum |
| Party ID | Democrat | 14.8% | 50.9% | 7.4% | 15.7% | 11.1% | 100.0% |
| | Independent | 19.9 | 50.6 | 8.9 | 12.9 | 7.7 | 100.0 |
| | Republican | 8.5 | 49.7 | 8.9 | 24.4 | 8.5 | 100.0 |
| | Other | 13.8 | 49.2 | 4.6 | 27.7 | 4.6 | 100.0 |
| | Sum | 15.3 | 50.5 | 8.3 | 16.9 | 9.0 | 100.0 |

Why don't we use row percentages?
Because they show the conditioning of the columns on the rows, and we normally put the "dependent variable" in the rows

# Visualizing Tabular Data

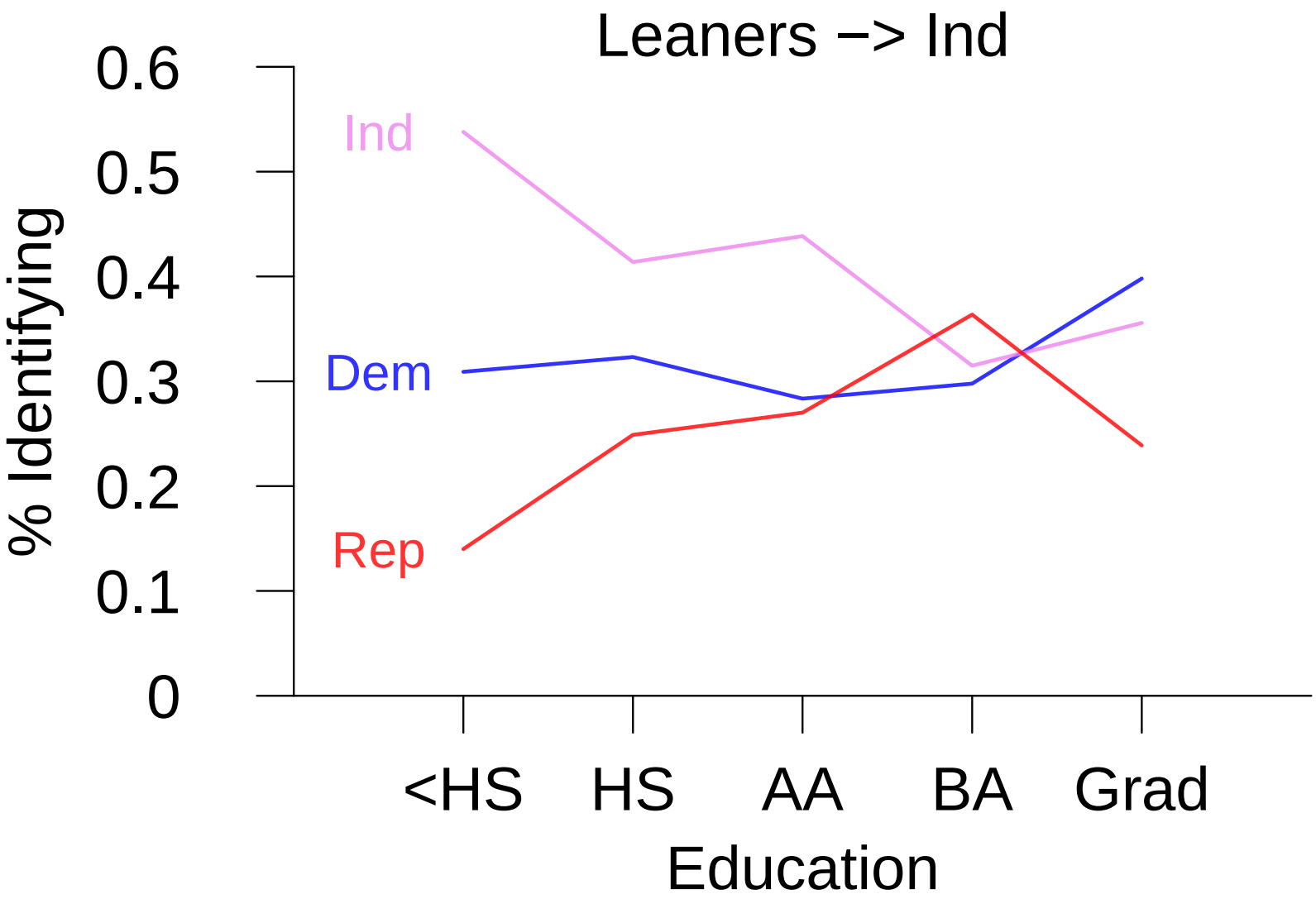
Just because our data come in a table doesn't mean we have to leave them there

A picture is often easier to sort out

But we need to plot the *right* numbers

What happens if we plot the *column percentages* from our tables?

## The table as a graph



# Exploring model sensitivity

We made several assumptions in tabulating and analyzing our data

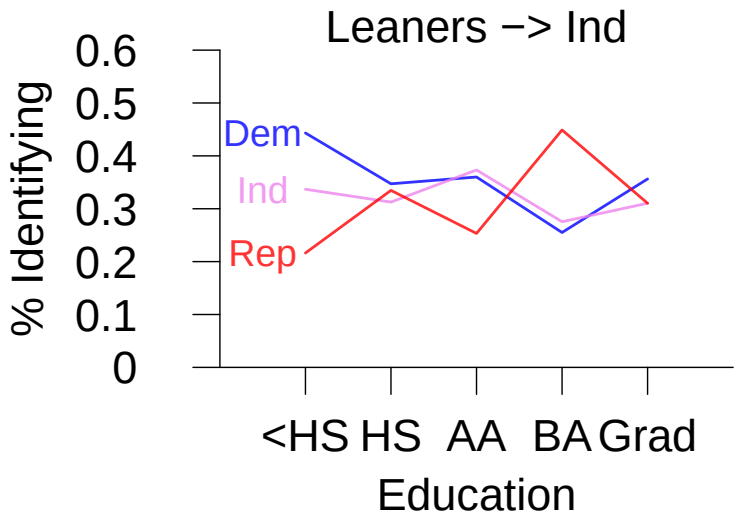
**Categorizing Leaners** We grouped leaners with other Independents. But many political scientists think they are actually intense partisans

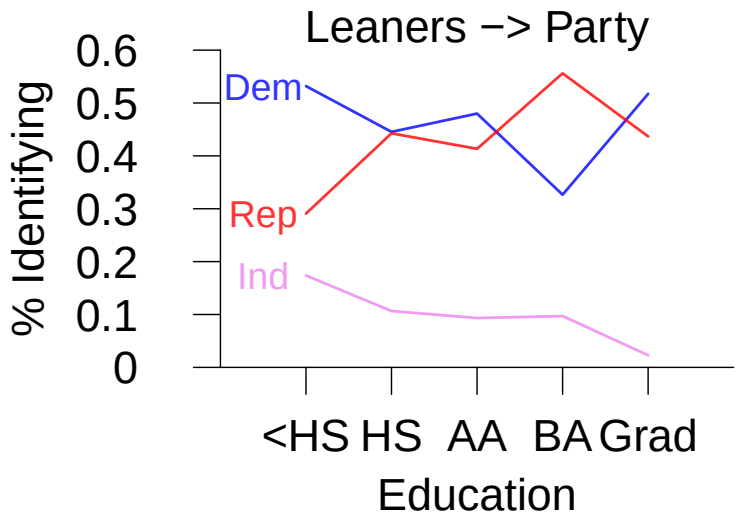
**Is 2006 special?** We looked at just one year in American politics. Do our findings hold in other years? Is there interesting variation over time?

We could make more tables categorizing the leaners as partisans, or using data from, say 1990.

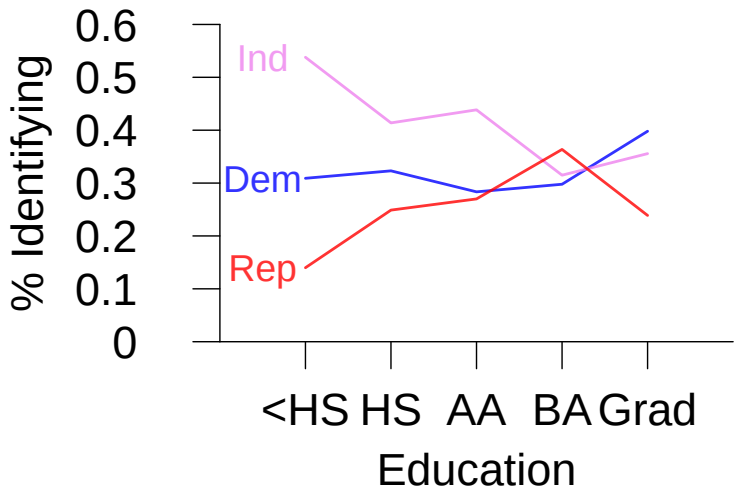
But who wants to pour over 4 cross-tabs?

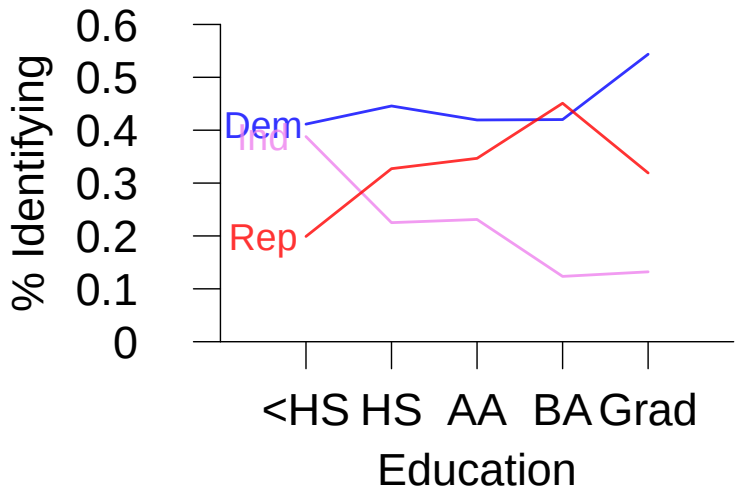
1990


Leaners −> Ind
% Identifying
0.6
0.5
0.4
0.3
0.2
0.1
0
Dem
Ind
Rep
<HS HS AA BA Grad
Education


Leaners −> Party
% Identifying
0.6
0.5
0.4
0.3
0.2
0.1
0
Dem
Rep
Ind
<HS HS AA BA Grad
Education

2006


% Identifying
0.6
0.5
0.4
0.3
0.2
0.1
0
Ind
Dem
Rep
<HS HS AA BA Grad
Education


% Identifying
0.6
0.5
0.4
0.3
0.2
0.1
0
Dem
Ind
Rep
<HS HS AA BA Grad
Education

# Multidimensional Tables

If we want to consider possible confounders, we need more than two dimensions to our table

That is, we need one dimension for every independent variable, plus one for our dependent variable

This gets tricky fast: hard to visualize, or do our column percents trick

But important to consider: if we don't include confounders, we can make very incorrect inferences about relationships

# Discrimination?

Suppose the (fictional) University of Tlon is sued for discriminatory hiring

Both sides stipulate that

- the best candidate can be determined uniquely
- should always be hired
- is equally likely to be male or female

The case turns on whether the University hired male and female candidates at the same rate

# Discrimination?

Here are the data for the university's "eclectic" departments

**Hiring data for Tlon University's "eclectic" departments**

| Departments | Men | | Women | |
|---|---|---|---|---|
| | Hired | Applied | Hired | Applied |

## Discrimination?

Here are the data for the university's "eclectic" departments

### Hiring data for Tlon University's "eclectic" departments

| | Men | | Women | |
|---|---|---|---|---|
| Departments | Hired | Applied | Hired | Applied |
| Ancient Egyptian Algebra | 2 | 8 | 1 | 5 |

# Discrimination?

Here are the data for the university's "eclectic" departments

**Hiring data for Tlon University's "eclectic" departments**

| | Men | | Women | |
|---|---|---|---|---|
| Departments | Hired | Applied | Hired | Applied |
| Ancient Egyptian Algebra | 2 | 8 | 1 | 5 |
| Navajo Cryptography | 4 | 5 | 6 | 8 |

# Discrimination?

Here are the data for the university's "eclectic" departments

**Hiring data for Tlon University's "eclectic" departments**

| | Men | | Women | |
|---|---|---|---|---|
| Departments | Hired | Applied | Hired | Applied |
| Ancient Egyptian Algebra | 2 | 8 | 1 | 5 |
| Navajo Cryptography | 4 | 5 | 6 | 8 |

The plaintiffs point out that in each dept, a greater % of men were hired:

| Departments | Men | | Women |
|---|---|---|---|
| Ancient Egyptian Algebra | 25% | $>$ | 20% |

# Discrimination?

Here are the data for the university's "eclectic" departments

**Hiring data for Tlon University's "eclectic" departments**

| | Men | | Women | |
|---|---|---|---|---|
| Departments | Hired | Applied | Hired | Applied |
| Ancient Egyptian Algebra | 2 | 8 | 1 | 5 |
| Navajo Cryptography | 4 | 5 | 6 | 8 |

The plaintiffs point out that in each dept, a greater % of men were hired:

| Departments | Men | | Women |
|---|---|---|---|
| Ancient Egyptian Algebra | 25% | $>$ | 20% |
| Navajo Cryptography | 80% | $>$ | 75% |

## Discrimination?

"But wait!" says the defense. "Look at the *totals*"

| | Men | | Women | |
|---|---|---|---|---|
| Departments | Hired | Applied | Hired | Applied |
| Ancient Egyptian Algebra | 2 | 8 | 1 | 5 |
| Navajo Cryptography | 4 | 5 | 6 | 8 |

## Discrimination?

"But wait!" says the defense. "Look at the *totals*"

| | Men | | Women | |
|---|---|---|---|---|
| Departments | Hired | Applied | Hired | Applied |
| Ancient Egyptian Algebra | 2 | 8 | 1 | 5 |
| Navajo Cryptography | 4 | 5 | 6 | 8 |
| Total | 6 | 13 | 7 | 13 |

"We actually hired *more* women at a higher rate than men!"

## Discrimination?

"But wait!" says the defense. "Look at the *totals*"

| | Men | | Women | |
|---|---|---|---|---|
| Departments | Hired | Applied | Hired | Applied |
| Ancient Egyptian Algebra | 2 | 8 | 1 | 5 |
| Navajo Cryptography | 4 | 5 | 6 | 8 |
| Total | 6 | 13 | 7 | 13 |

"We actually hired *more* women at a higher rate than men!"

The plaintiffs in a lawsuit point out that in each dept, a greater % of men were hired:

| Departments | Men | | Women |
|---|---|---|---|
| Ancient Egyptian Algebra | 25% | $>$ | 20% |
| Navajo Cryptography | 80% | $>$ | 75% |

## Discrimination?

"But wait!" says the defense. "Look at the *totals*"

| | Men | | Women | |
|---|---|---|---|---|
| Departments | Hired | Applied | Hired | Applied |
| Ancient Egyptian Algebra | 2 | 8 | 1 | 5 |
| Navajo Cryptography | 4 | 5 | 6 | 8 |
| Total | 6 | 13 | 7 | 13 |

"We actually hired *more* women at a higher rate than men!"

The plaintiffs in a lawsuit point out that in each dept, a greater % of men were hired:

| Departments | Men | | Women |
|---|---|---|---|
| Ancient Egyptian Algebra | 25% | $>$ | 20% |
| Navajo Cryptography | 80% | $>$ | 75% |
| Both departments | 46% | $<$ | 54% |

What's going on here?

# Simpson's Paradox

The Departments are different.
Perhaps AEA has much less funding that NC, and can make fewer offers.

## Simpson's Paradox

The Departments are different.
Perhaps AEA has much less funding that NC, and can make fewer offers.

Women, either by chance or by design, more often apply to Navajo Cryptography

# Simpson's Paradox

The Departments are different.
Perhaps AEA has much less funding that NC, and can make fewer offers.

Women, either by chance or by design, more often apply to Navajo Cryptography

When we look at the dept totals, we "control" for this difference in hiring difficulty

## Simpson's Paradox

The Departments are different.
Perhaps AEA has much less funding that NC, and can make fewer offers.

Women, either by chance or by design, more often apply to Navajo Cryptography

When we look at the dept totals, we "control" for this difference in hiring difficulty

When we look at the grand total, we are omitting department-level variables

# Simpson's Paradox

The Departments are different.
Perhaps AEA has much less funding that NC, and can make fewer offers.

Women, either by chance or by design, more often apply to Navajo Cryptography

When we look at the dept totals, we "control" for this difference in hiring difficulty

When we look at the grand total, we are omitting department-level variables

But these department level variables are confounders:
correlated with the outcome *and* with our explanatory variable

# Simpson's Paradox

The Departments are different.
Perhaps AEA has much less funding that NC, and can make fewer offers.

Women, either by chance or by design, more often apply to Navajo Cryptography

When we look at the dept totals, we "control" for this difference in hiring difficulty

When we look at the grand total, we are omitting department-level variables

But these department level variables are confounders:
correlated with the outcome *and* with our explanatory variable

Omitting them leads to this confusion, known as Simpson's Paradox

# Simpson's Paradox

The Departments are different.
Perhaps AEA has much less funding that NC, and can make fewer offers.

Women, either by chance or by design, more often apply to Navajo Cryptography

When we look at the dept totals, we "control" for this difference in hiring difficulty

When we look at the grand total, we are omitting department-level variables

But these department level variables are confounders:
correlated with the outcome *and* with our explanatory variable

Omitting them leads to this confusion, known as Simpson's Paradox