

Correlation

Measuring Relationships between Variables



1

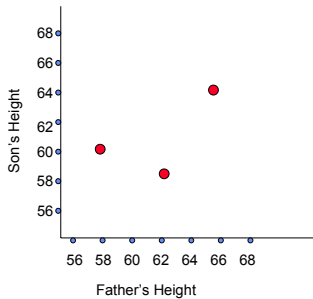
Associations Between Variables

- How strongly is a child's birthweight related to his mother's weight? Father's weight? Gestation?
- What are the relationships among nutritional contents of breakfast cereals: calories, fiber, fat, sugar, carbohydrates?
- How is a person's weight related to his height?
- How strongly is income related to education?

2

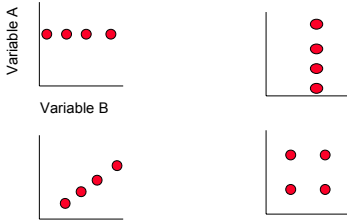
Constructing a Scatter Diagram

Father	Son
58	60
62	58
66	64



3

What Relationships Are Shown in These Scatter Diagrams?



4

Depicting Relationships with Scatter Diagrams: The Challenger Disaster

In January 1986, engineers opposed a decision to launch a space shuttle because they were worried that rubber O-rings would not seal at the cold temperature forecast for the launch day. NASA officials pressured them to reverse their recommendation. Challenger was launched the next day, it exploded, and seven astronauts died because two O-rings leaked.

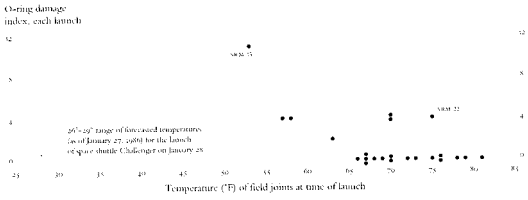
5

The Presentations Engineers Faxed to NASA

- | | |
|--|---|
| <p>CONCLUSIONS:</p> <ul style="list-style-type: none"> o TEMPERATURE OF O-RING IS NOT ONLY PARAMETER CONTRIBUTING BLOW BY o SRM IS WITH BLOW BY AND AN O-RING TEMP AT 53°F SEAL IS WITH BLOW BY AND AN O-RING TEMP AT 53°F FOUR DEVELOPMENT MOTORS WITH NO BLOW BY WERE TESTED AT O-RING TEMP OF 47° TO 52°F o DEVELOPMENT MOTORS HAD PUFFY PACKING WHICH RESULTED IN BETTER PERFORMANCE o AT ABOUT 50°F BLOW BY COULD BE EXPERIENCED IN CASE JOINTS o TEMP FOR SRM 25 ON 1-28-86 LAUNCH WILL BE 50°F 2 AM 38°F 6 PM o HAVE NO DATA THAT WOULD INDICATE SRM 25 IS DIFFERENT THAN SRM IS OTHER THAN TEMP | <p>RECOMMENDATIONS:</p> <ul style="list-style-type: none"> o O-RING TEMP MUST BE $\geq 53^\circ\text{F}$ AT LAUNCH o DEVELOPMENT MOTORS AT 47° TO 52°F WITH PUFFY PACKING HAD NO BLOW BY SRM IS (THE BEST SIMULATION) WORKED AT 53°F o PERFECT AMBIENT CONDITIONS (TEMP & WIND) TO DETERMINE LAUNCH TIME |
|--|---|

6

A More Convincing Presentation



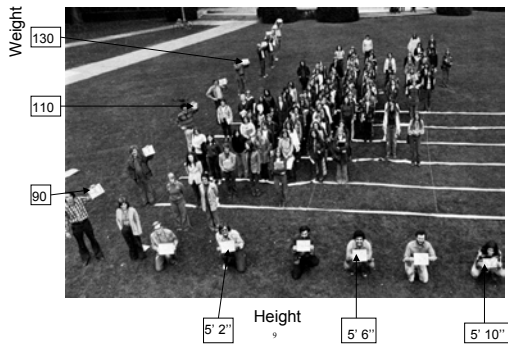
7

An Even More Convincing Presentation

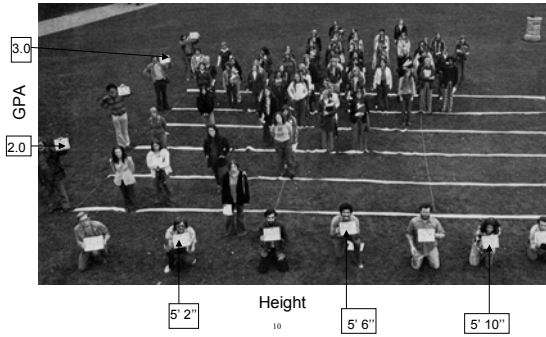


8

Those Women from Wisconsin



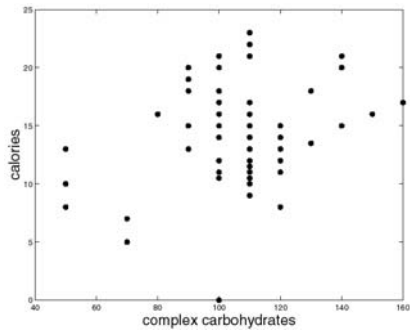
Height and GPA??

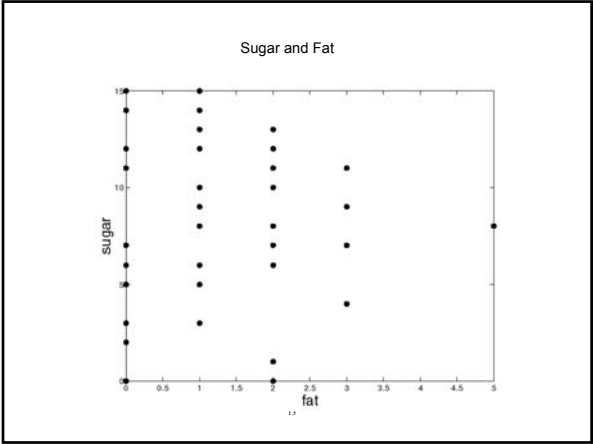


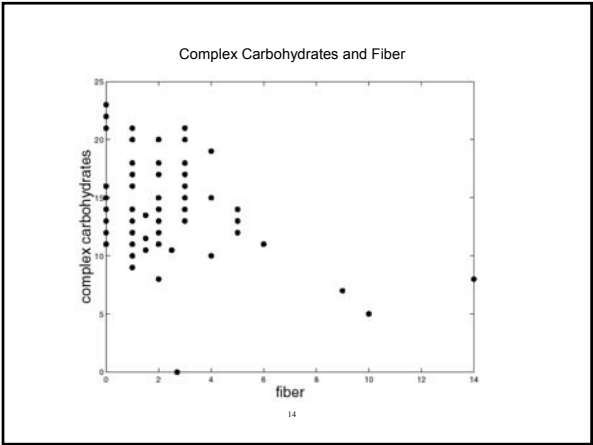
Breakfast Cereals

What are the relationships among nutritional contents of breakfast cereals: calories, fiber, fat, sugar, carbohydrates?

Calories and Carbohydrates





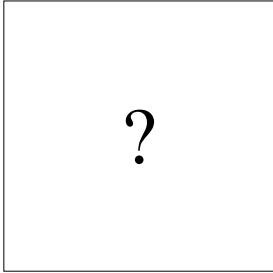


Babies' Birthweights

How strongly is a child's birthweight related to his mother's weight? Father's weight? Gestation?

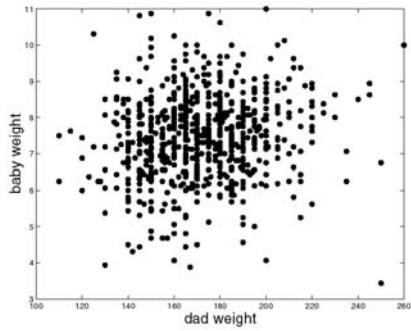
15

Birthweight and Dad's Weight



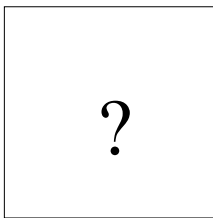
16

Birthweight and Dad's Weight

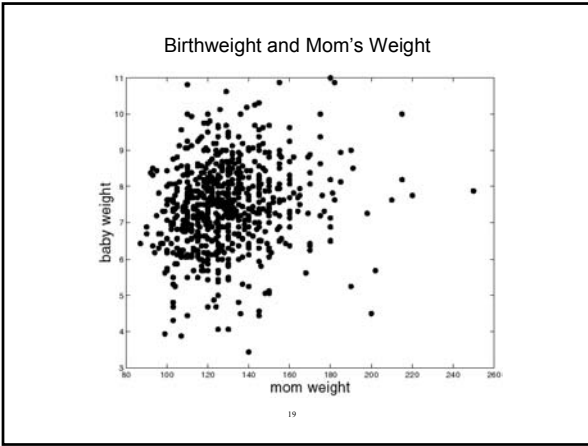


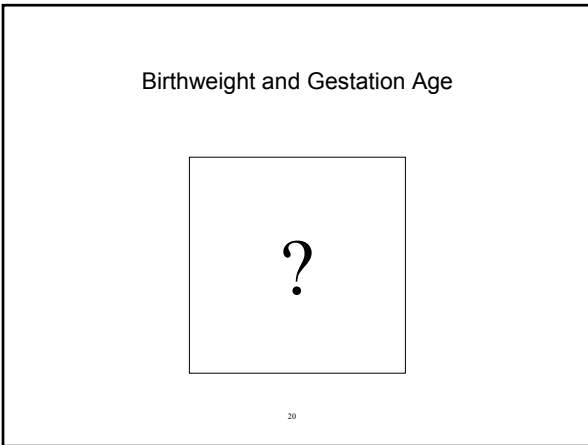
17

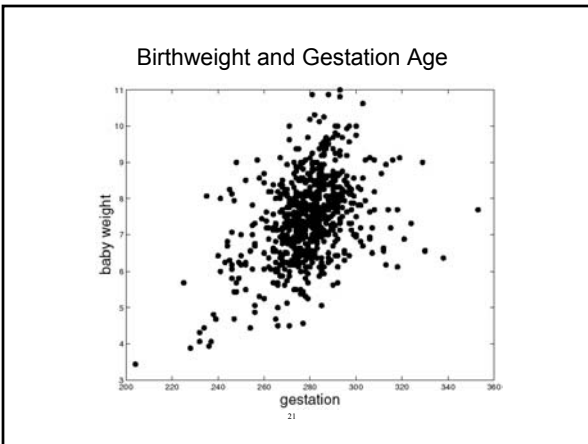
Birthweight and Mom's Weight



18





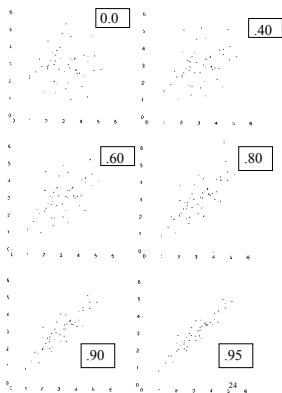


Time Sequence Plots: A Type of Scatter Diagram

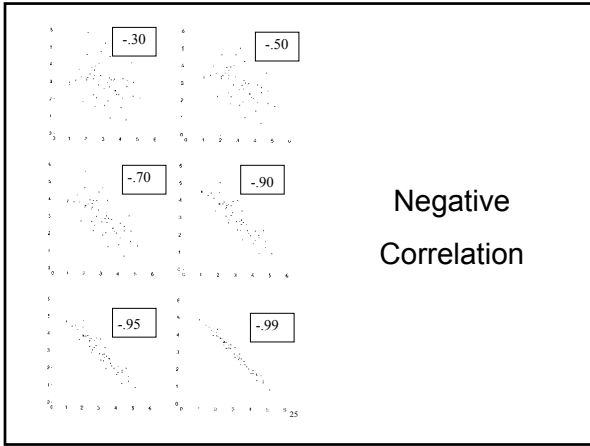
Sports Records

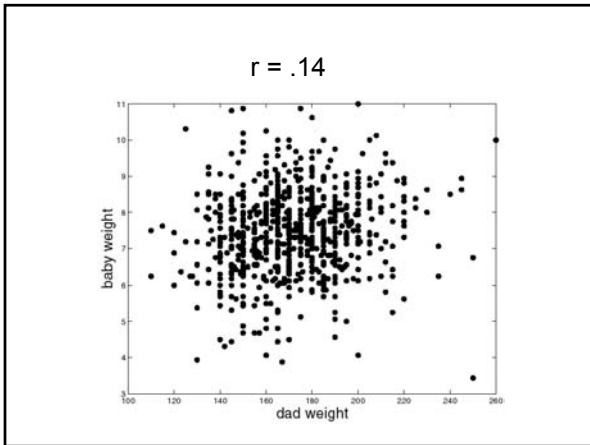
The Correlation Coefficient: r

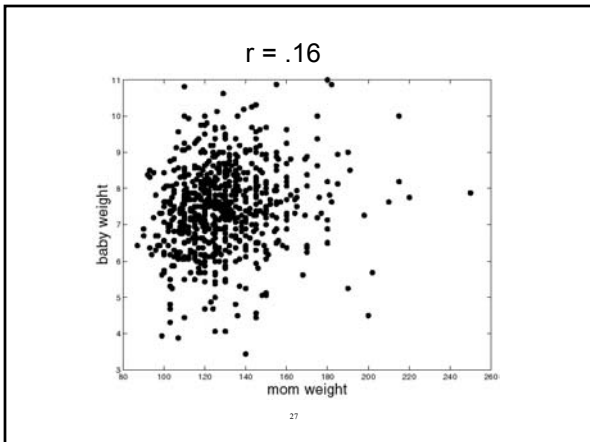
A Numerical Summary of the Strength of a Linear Relationship between Two Variables

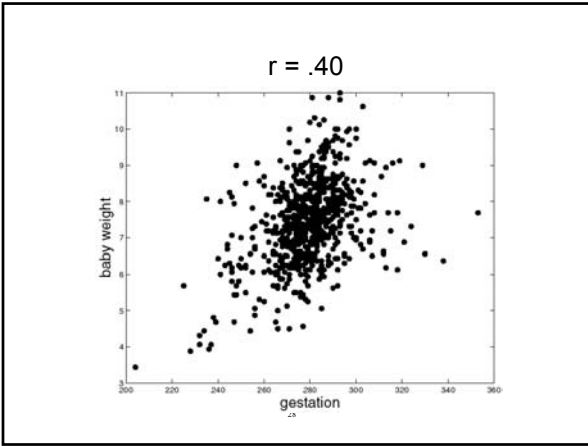


Positive Correlation








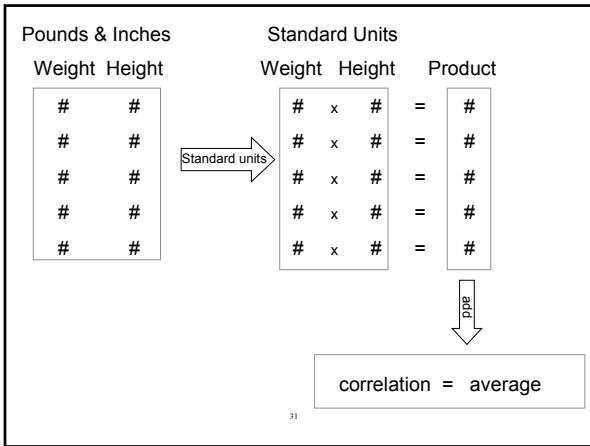






The Recipe for a Correlation Coefficient

1. Convert each variable to standard units
2. Calculate the average value of the products



Implications: The Correlation Coefficient Is Not Affected by

- *Interchanging the variables.* (r only depends on products.)
- *Adding a constant to the values of one variable.* (Adding a constant does not change their values measured in standard units.)

32

- *Multiplying the values of one variable by a constant.* (When the values are multiplied by a constant, so are the Average and SD, and the standard units are unchanged.)

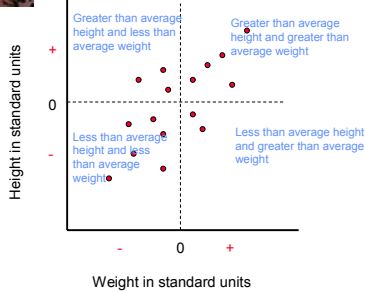
$$\text{Standard Unit} = \frac{\text{Value} - \text{Average}}{\text{SD}}$$

33

Example: The correlation between mother's height measured in *inches* and baby's weight measured in *ounces* is the same as when they are measured in *kilometers* and *tons*.



How Does the Correlation Coefficient Get Its Sign?

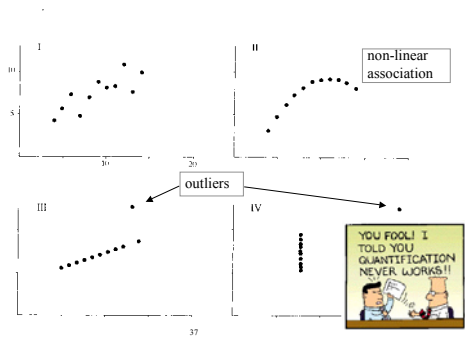




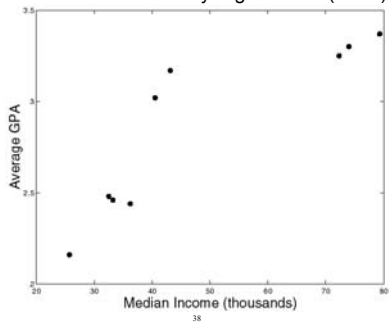
Caution: r measures *linear* relationship. These four sets of pairs all have $r = .8$

I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.26
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

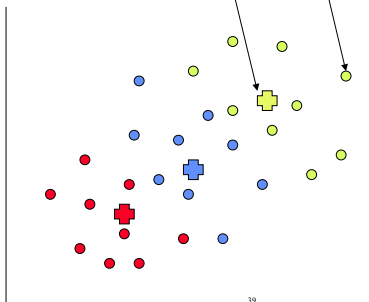
But the scatter plots differ



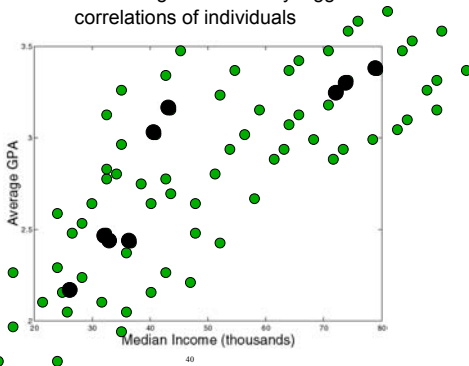
Ecological (aggregate) correlation: GPA and income by ZIP code at Berkeley High School ($r=.87$)



Correlations of *averages* are usually bigger than correlations of *individuals*



Correlations of averages are usually bigger than correlations of individuals

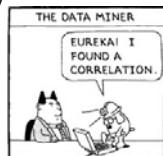


Correlation and Causation

There is strong correlation between

- The number of teachers in a school district and the number of failing students.
- The number of automobiles in California per year and the number of homicides.
- Kids' feet lengths and reading ability

Correlation does not imply causation.



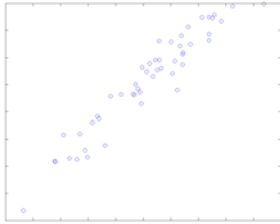
Summary

- Scatter diagrams show relationships between variables
- The correlation coefficient measures the strength of a *linear* relationship
- It measures clustering about a line
- It ranges between -1 and 1
- It is calculated as the average of products of standard units

- It measures association, not causation
- It can be misleading if there are outliers or nonlinear association
- “Ecological” correlation tend to overstate strengths of relationships for individuals

43

? 50 observations are taken on two variables and it is found that they have a correlation coefficient equal to .95. True or false and explain briefly: The scatterplot is roughly similar to the one shown below.



44

? The table below shows data for a hypothetical study where the average daily temperature (Celsius) and cumulative rainfall (mm) were measured for nine one-month periods in a particular U.S. city.

month	temp	rainfall
1	14	15
2	17	16
3	16	17
4	19	18
5	18	19
6	21	20
7	20	21
8	24	22
9	22	23

Of the following three numbers, -7, 0, and .9, which is closest to the correlation coefficient for average daily temperature and cumulative rainfall. Explain briefly —no calculations are necessary.

45

Suppose the thermometer used in the study consistently gave readings that were two degrees too high. Explain briefly, how the thermometer's bias would affect the following numbers:

The average daily temperature

The standard deviation of the average daily temperature.

The correlation coefficient of the average daily temperature and the rainfall.
