# Descriptive Statistics

Portland State University
USP 634 Data Analysis I
Spring 2018

# Outline

A general process of scientific investigation

Case study

Descriptive statistics and exploratory data analysis

# General Process of Scientific Investigation

1. Identify a question or problem.

2. Collect relevant data on the topic.

3. Analyze the data.

4. Form a conclusion.

# Identify Research Question/Problem

What?

Why?

How?

When?

# Question to keep in mind

Is there uncertain in the data generation process? Where does the uncertainty (variation) come from?

- Sample → Population
- Unobserved factors
- Chance
- Measurement error

# Data Collection

**Observational study** data are collected in a way that does not directly interfere with how the data arise. In general, observational studies can provide evidence of a naturally occurring association between variables, but they cannot by themselves show a causal connection.

**Experiments**: a sample of individuals are randomly assigned into control or treatment groups. Randomized experiments may be used to establish causality.

# Methods of sampling

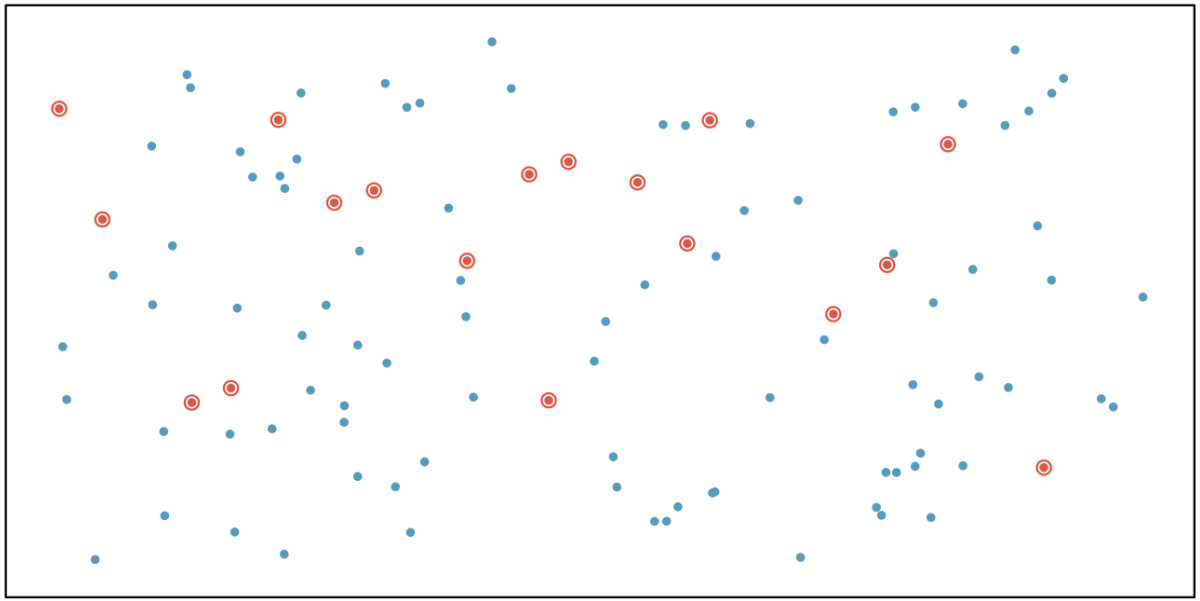
Almost all statistical methods are based on the notion of implied randomness.

If observational data are not collected in a random framework from a population, these statistical methods -- the estimates and errors associated with the estimates -- are not reliable.

Most commonly used random sampling techniques are simple, stratified, and cluster sampling.

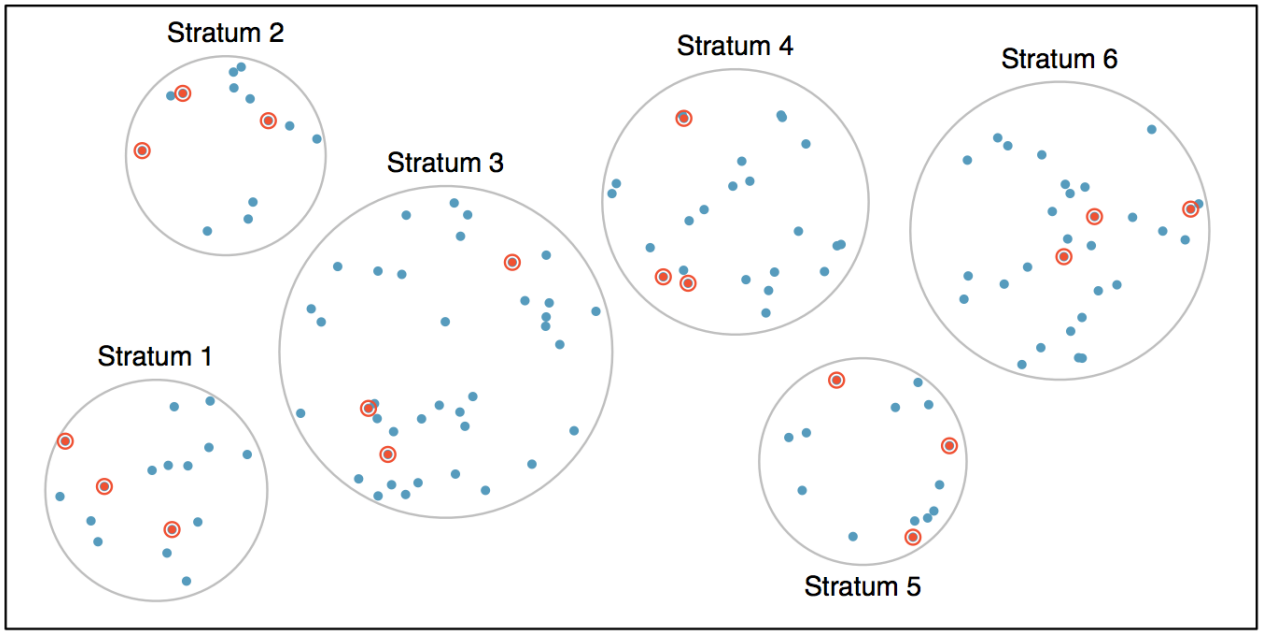
# Simple Random Sample

Randomly select cases from the population, where there is no implied connection between the points that are selected.



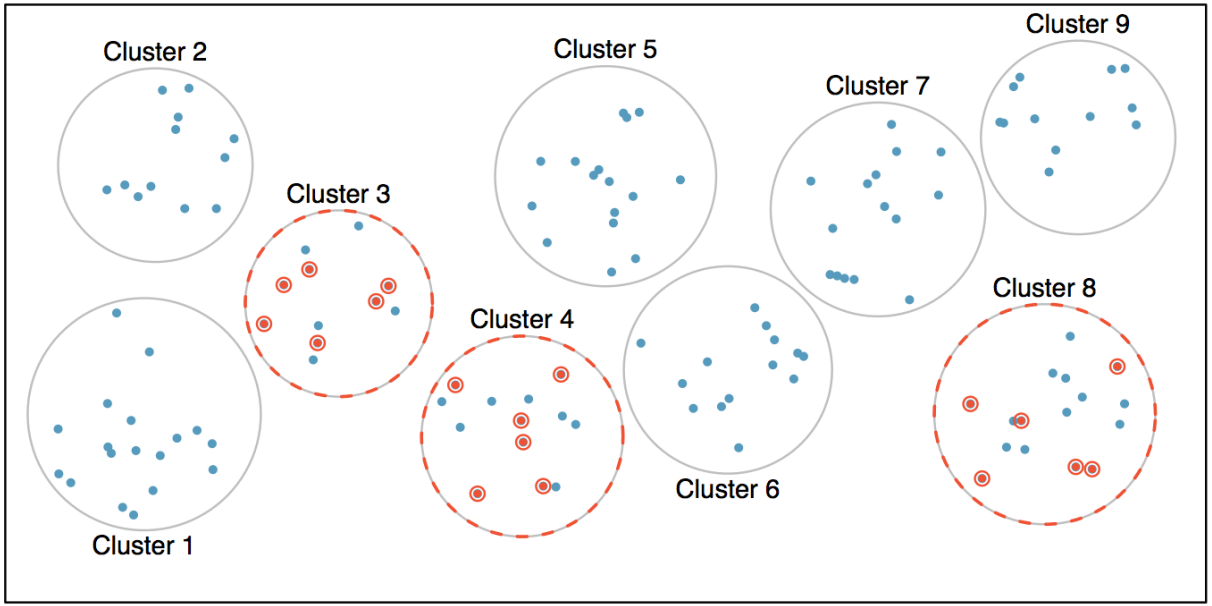
# Stratified Sample

Strata are made up of similar observations. We take a simple random sample from each stratum.



# Cluster Sample

Clusters are usually not made up of homogeneous observations, and we take a simple random sample from a random sample of clusters. Usually preferred for economical reasons.



# Sampling bias

Non-response: If only a small fraction of the randomly sampled people choose to respond to a survey, the sample may no longer be representative of the population.

Voluntary response: Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue. Such a sample will also not be representative of the population.

Convenience sample: Individuals who are easily accessible are more likely to be included in the sample.

# Case Study: Gender Discrimination

# Gender Discrimination

- In 1972, as a part of a study on gender discrimination, 48 male bank supervisors were each given the same personnel file and asked to judge whether the person should be promoted to a branch manager job that was described as “routine”.
- The files were identical except that half of the supervisors had files showing the person was male while the other half had files showing the person was female.
- It was randomly determined which supervisors got “male” applications and which got “female” applications.
- Of the 48 files reviewed, 35 were promoted.
- The study is testing whether females are unfairly discriminated against.

Is this an observational study or an experiment?

Experiment

*B.Rosen and T. Jerdee (1974), ``Influence of sex role stereotypes on personnel decisions", J.Applied Psychology, 59:9-14.*

# Data

At a first glance, does there appear to be a relationship between promotion and gender?

| | | *Promotion* | | |
|---|---|---|---|---|
| | | Promoted | Not Promoted | Total |
| *Gender* | Male | 21 | 3 | 24 |
| | Female | 14 | 10 | 24 |
| | Total | 35 | 13 | 48 |

**% of males promoted:** 21 / 24 = 87.5%

**% of females promoted:** 14 / 24 = 58.3%

A difference of 29.2%

# Practice

We saw a difference of almost 30% (29.2% to be exact) between the proportion of male and female files that are promoted. Based on this information, which of the below is true?

(a) If we were to repeat the experiment we will definitely see that more female files get promoted. This was a fluke.

(b) Promotion is dependent on gender, males are more likely to be promoted, and hence there is gender discrimination against women in promotion decisions.

(c) The difference in the proportions of promoted male and female files is due to chance, this is not evidence of gender discrimination against women in promotion decisions.

(d) Women are less qualified than men, and this is why fewer females get promoted.

# Practice

We saw a difference of almost 30% (29.2% to be exact) between the proportion of male and female files that are promoted. Based on this information, which of the below is true?

(a) If we were to repeat the experiment we will definitely see that more female files get promoted. This was a fluke.

(b) *Promotion is dependent on gender, males are more likely to be promoted, and hence there is gender discrimination against women in promotion decisions. Maybe*

(c) *The difference in the proportions of promoted male and female files is due to chance, this is not evidence of gender discrimination against women in promotion decisions. Maybe*

(d) Women are less qualified than men, and this is why fewer females get promoted.

# Two Competing Claims

“There is nothing going on.” (Null Hypothesis)
Promotion and gender are independent.
No gender discrimination.
Observed difference in proportions is simply due to chance.

“There is something going on.” (Alternative Hypothesis)
Promotion and gender are dependent.
There is gender discrimination.
Observed difference in proportions is not due to chance.

# Check for Independence and Draw Conclusion

| | | *Promotion* | | |
|---|---|---|---|---|
| | | Promoted | Not Promoted | Total |
| *Gender* | Male | 21 | 3 | 24 |
| | Female | 14 | 10 | 24 |
| | Total | 35 | 13 | 48 |

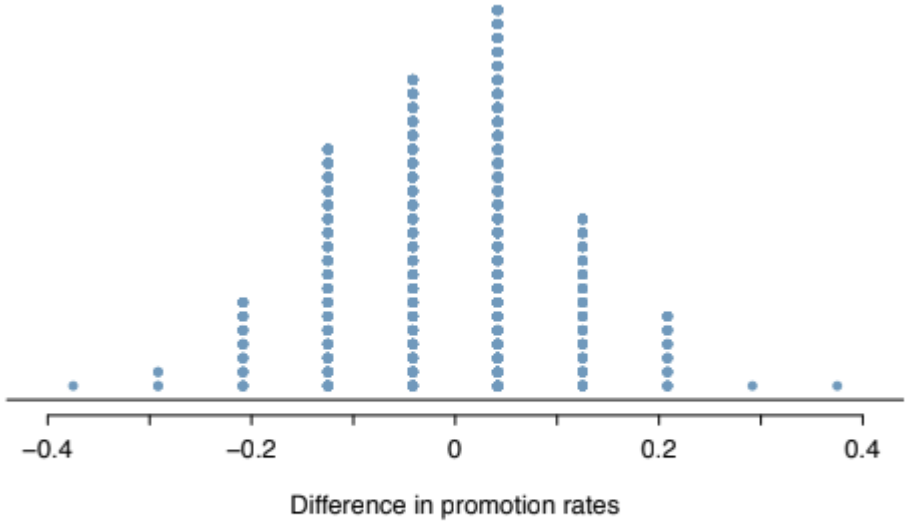



Figure 1.47: A stacked dot plot of differences from 100 simulations produced under the independence model, $H_0$, where `gender_sim` and `decision` are independent. Two of the 100 simulations had a difference of at least 29.2%, the difference observed in the study.

# Data Basics

# Data matrix

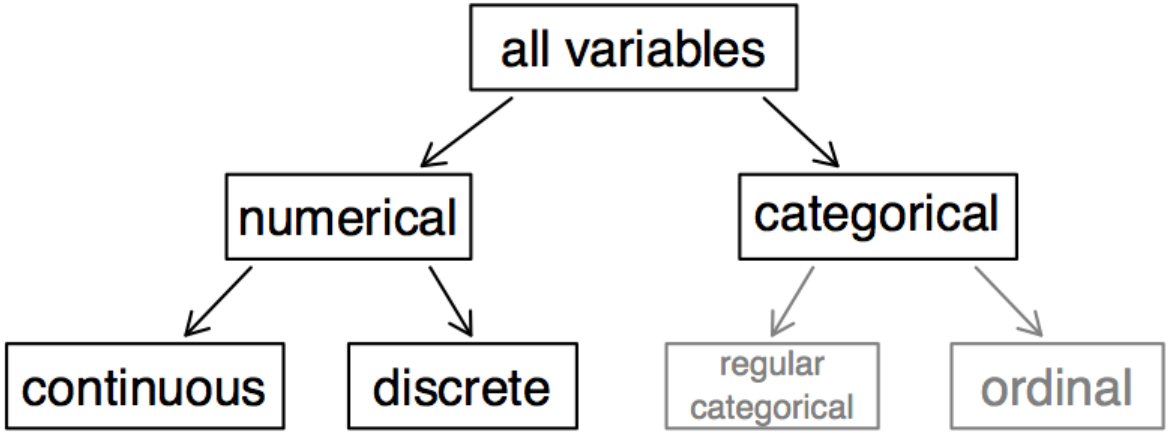
Data collected on students in a statistics class on a variety of variables:

*variable* ↓

| Stu. | gender | intro_extra | ··· | dread |
|---|---|---|---|---|
| 1 | male | extravert | ··· | 3 |
| 2 | female | extravert | ··· | 2 |
| 3 | female | introvert | ··· | 4 |
| 4 | female | extravert | ··· | 2 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 86 | male | extravert | ··· | 3 |

← *observation*

# Types of variables



# Types of variables (cont.)

| | gender | sleep | bedtime | countries | dread |
|---|---|---|---|---|---|
| 1 | male | 5 | 12-2 | 13 | 3 |
| 2 | female | 7 | 10-12 | 7 | 2 |
| 3 | female | 5.5 | 12-2 | 1 | 4 |
| 4 | female | 7 | 12-2 | | 2 |
| 5 | female | 3 | 12-2 | 1 | 3 |
| 6 | female | 3 | 12-2 | 9 | 4 |

gender - categorical

sleep - numerical, continuous

bedtime - categorical, ordinal

countries - numerical, discrete

dread - categorical, ordinal (could also be used as numerical)

# Practice

What type of variable is a telephone area code?

(a) numerical, continuous
(b) numerical, discrete
(c) categorical
(d) categorical, ordinal

# Practice

What type of variable is a telephone area code?

(a) numerical, continuous
(b) numerical, discrete
(c) *categorical*
(d) categorical, ordinal

# Descriptive Statistics

# Analyze Data: Descriptive Statistics

**Descriptive Statistics** quantitatively describes the features of a data set. Descriptive statistics are distinguished from **inferential statistics**, in that descriptive statistics aim to summarize a sample, rather than use the data to generalize information about the population.

# Analyze Data: Descriptive Statistics

**One continuous variable**

Quantitative: mean, sd, median, range, IQR; skewness, Kurtosis

Visualization: Histogram, box plot, density plot

**One Categorical variable**

Quantitative: Frequency table, mode, *mean* (ordinal)

Visualization: Bar chart, pie chart

**Two continuous variables** (X → Y):

Quantitative: measure of correlation (e.g. Pearson’s R)

Visualization: Scatter Plot

# Scatter plots

Scatter plots have two dimensions:

The independent variable (X) is plotted along the horizontal axis (which is called “the X axis”).

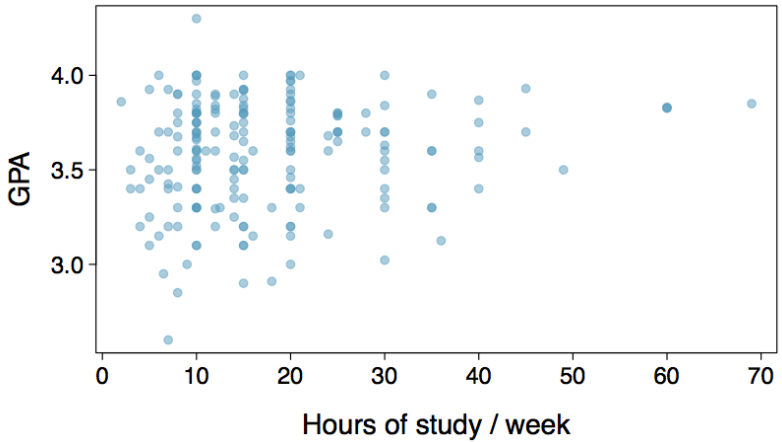
The dependent variable (Y) is plotted along the vertical axis (which is called “the Y axis”).

Each dot on a scatter plot is a case/an observation.

The dot is placed at the intersection of the case’s scores on X and Y.

# Relationships among variables

Does there appear to be a relationship between the hours of study per week and the GPA of a student?




Can you spot anything unusual about any of the data points?

There is one student with GPA > 4.0, this is likely a data error.

# Practice

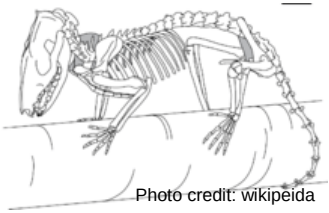
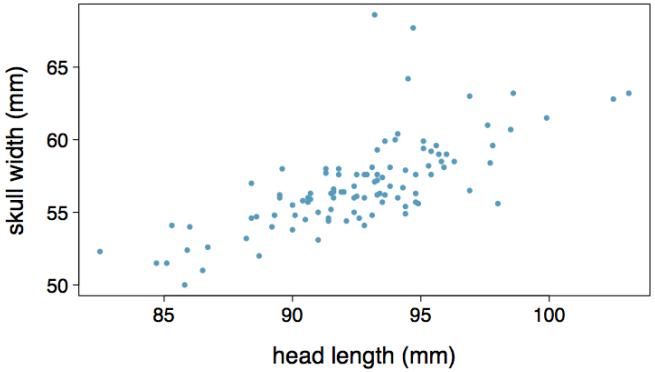


Photo credit: wikipeida

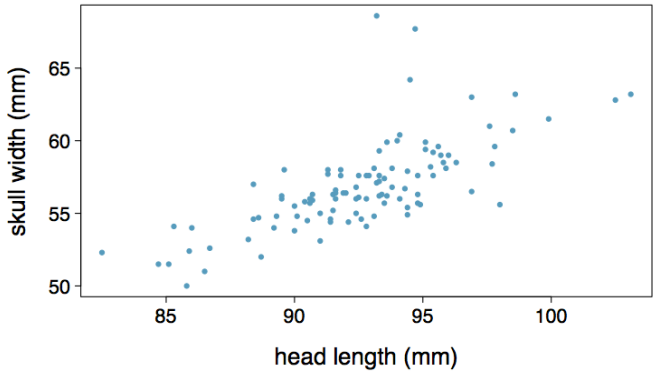
Based on the scatterplot on the right, which of the following statements is correct about the head and skull lengths of possums?




(a) There is no relationship between head length and skull width, i.e. the variables are independent.
(b) Head length and skull width are positively associated.
(c) Skull width and head length are negatively associated.
(d) A longer head causes the skull to be wider.
(e) A wider skull causes the head to be longer.

# Practice

Based on the scatterplot on the right, which of the following statements is correct about the head and skull lengths of possums?




(a) There is no relationship between head length and skull width, i.e. the variables are independent.

(b) *Head length and skull width are positively associated.*

(c) Skull width and head length are negatively associated.

(d) A longer head causes the skull to be wider.

(e) A wider skull causes the head to be longer.

# Associated vs. independent

- When two variables show some connection with one another, they are called associated variables.
  - Associated variables can also be called dependent variables and vice-versa.
- If two variables are not associated, i.e. there is no evident connection between the two, then they are said to be independent.

# How do we measure the association of X and Y?

Use a calculated regression line, if linear relationship is appropriate

Another way to measure the extent of clustering around the regression line is to using Pearson's r or $R^2$. These measures can be tested for statistical significance.

# Regression line: Strength and direction

## Strength of association

The greater the extent to which dots are clustered around the regression line, the stronger the relationship

## Direction of association

Positive: regression line rises left to right.

Negative: regression line falls left to right.

## Slope of regression line

Steeper slope implies larger “effect”—but caution: this partly an artifact of variable *units* and outliers

# Correlation: Pearson's r

AKA Pearson Product-Moment **Correlation**

Pearson's r is a measure of association for numeric variables:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}}$$

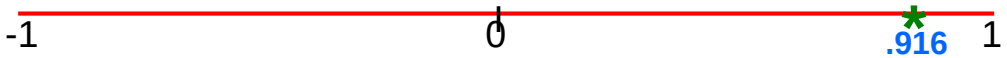
Ranges from -1 to 1:

- 0 indicates no relationship,
- -1 a perfect negative relationship
- 1 a perfect positive relationship

Limitation: No direct interpretation of intermediate values

# Correlation: Pearson's r

R code: `cor(X,Y)`




$N = 10$
$\Sigma X_i = 46$
$\Sigma X_i^2 = 256$
$\Sigma Y_i = 68.5$
$\Sigma Y_i^2 = 489.4$
$\Sigma X_i Y_i = 342.5$
$\bar{X} = 4.60$
$\bar{Y} = 6.85$

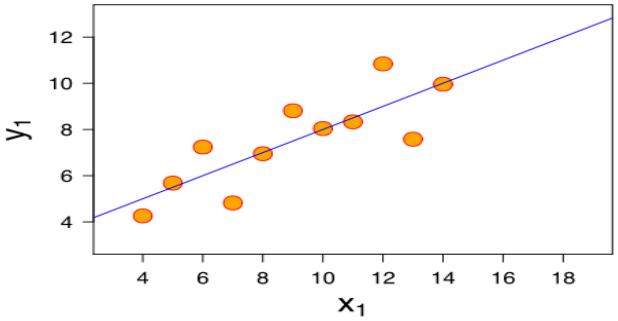
| X | Y |
|---|---|
| 1 | 4.4 |
| 2 | 5.5 |
| 3 | 5.7 |
| 4 | 5.8 |
| 4 | 7 |
| 5 | 7.2 |
| 6 | 7 |
| 6 | 9 |
| 7 | 8.4 |
| 8 | 8.8 |

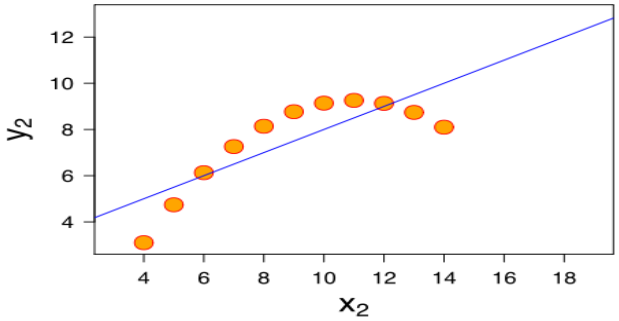
# Correlation

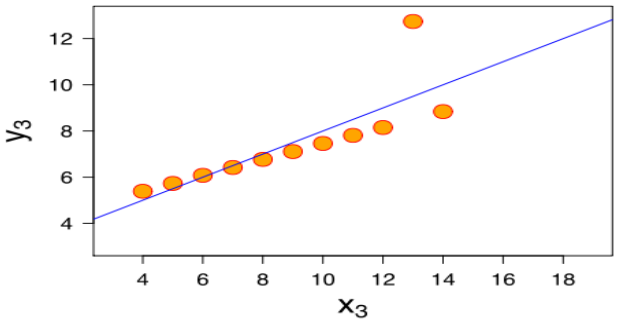
Assume a linear relationship

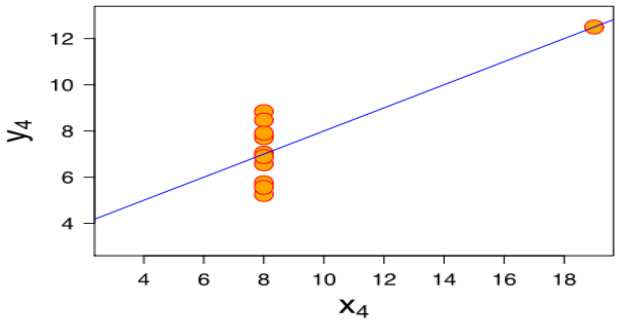
Sensitive to outliers

Always look at scatter plot, not just *r* statistic.




Four sets of data with the same correlation of 0.816

**One categorical (explanatory) and one continuous (response) variables**:

Quantitative: measure of correlation (e.g. Pearson’s R)

Visualization: grouped box plot, scatter plot, line chart, bar chart

**One Continuous (explanatory) and One categorical (response) variables**:

Quantitative: measure of correlation (e.g. Pearson’s R)

Visualization: (jittered) scatter Plot

**Two categorical variables**:

Quantitative: contingency table (cross tabulation), measure of association

Visualization: stacked bar chart, mosaic plot

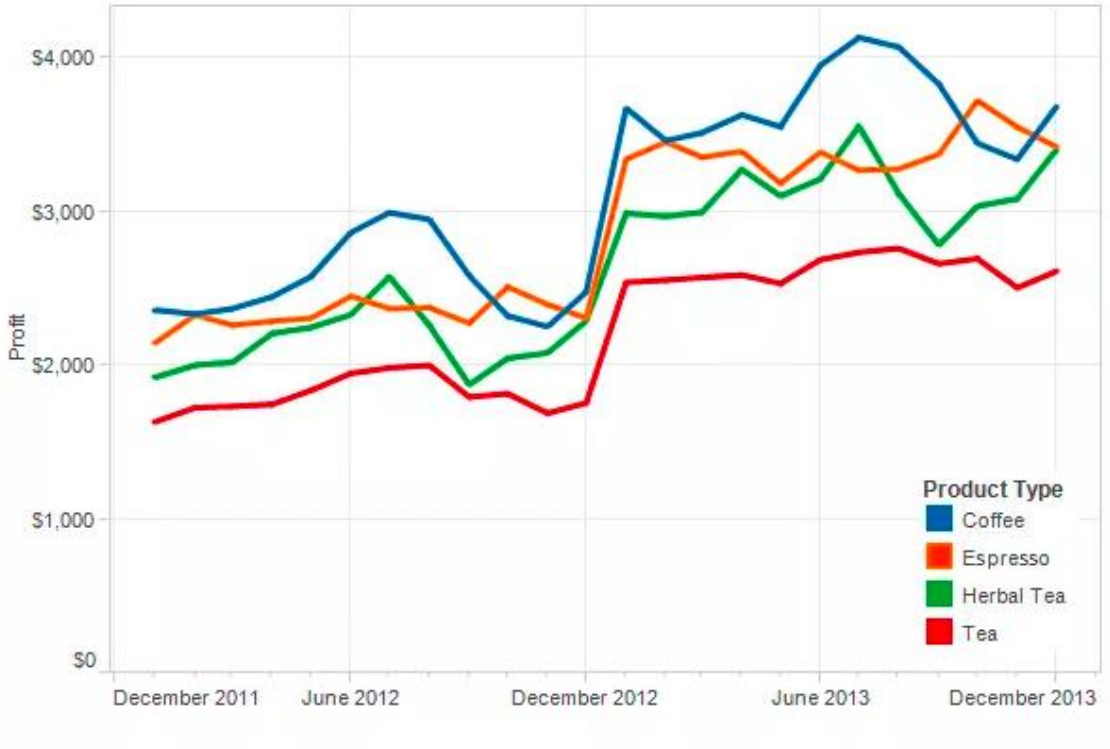
# Time Series: Continuous Variable

Continuous variable


A hypothetical coffee chain and look at their profits
Source: https://eagereyes.org/basics/data-continuous-vs-categorical

# Time series: Continuous Variable

Use colors or line styles
to differentiate categories
(categorical variable)



# Descriptive Stats Example

Barton, Bruce A. et al., 2005, The Relationship of Breakfast and Cereal Consumption to Nutrient Intake and Body Mass Index: The National Heart, Lung, and Blood Institute Growth and Health Study, Journal of the American Dietetic Association , Volume 105 , Issue 9 , 1383 - 1389