- **Data graphics (2)**
- **Experiment design**

---

*Background preparation:*

- *Kaplan (2016), Appendix, "Statistics brief reference", especially sections A.1.2 and A.3*

# 1. Data graphics

- What *kinds* of information are the following types of graphs good at showing?
  - bar graph or line graph
  - histogram
  - scatterplot

(see more examples of each type in Kaplan, sec A.1.2)

# 1. Data graphics

## Bar graph / line graph

- What kinds of information are these good at showing?

# 1. Data graphics

## Bar graph / line graph

- A good way to compare **values** (numerical) for two or more **groups** across two or more **categories**

# 1. Data graphics

## Bar graph / line graph

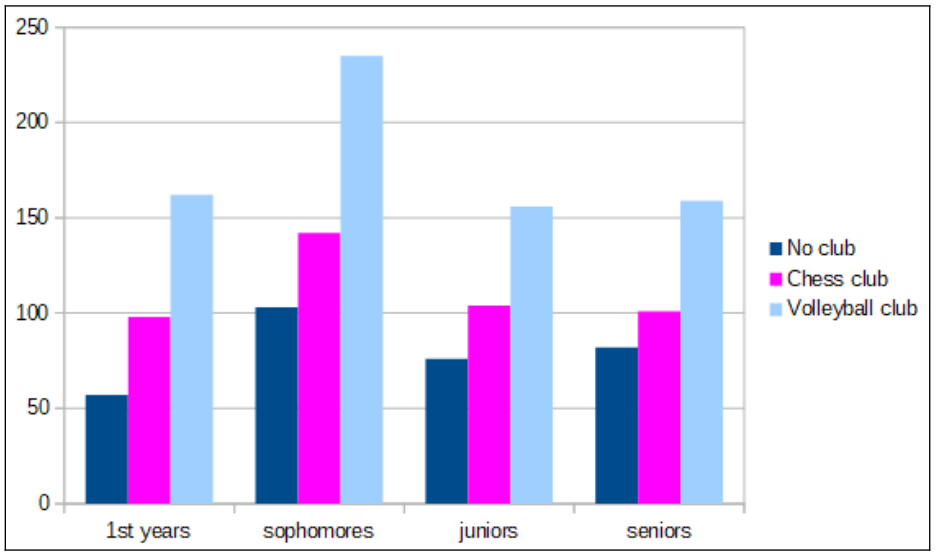
- How do we set one up?
- ***Pizza consumption*** *(invented data)*
  *Average number of slices per student per semester*

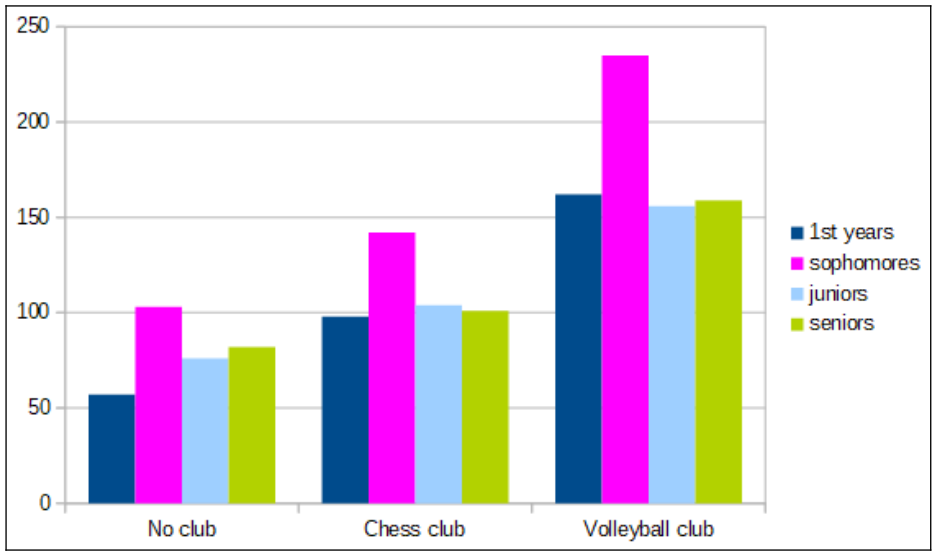
| | No club | Chess club | Volleyball club |
|---|---|---|---|
| 1st years | 57 | 98 | 162 |
| sophomores | 103 | 142 | 235 |
| juniors | 76 | 104 | 156 |
| seniors | 82 | 101 | 159 |

# 1. Data graphics

## Bar graph / line graph

- How do we set one up? — Bar graph



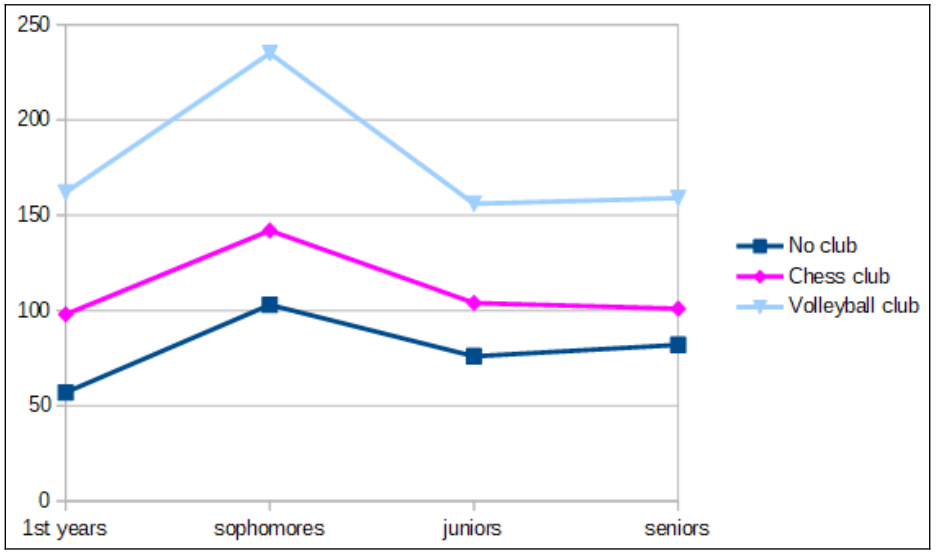



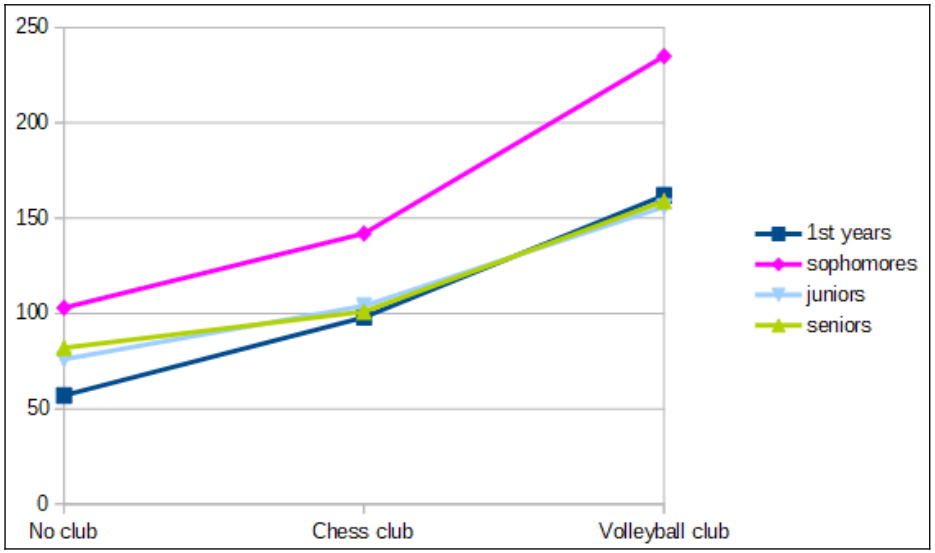

- Which comparisons are easier in each graphic? What "research question" is each better for?

# 1. Data graphics

## Bar graph / line graph

- How do we set one up? — Line graph




- Which comparisons are easier in each graphic? What "research question" is each better for?

  Note: Some researchers do not use connecting lines between points representing *categories* (because there are no intermediate values)

# 1. Data graphics

## Histogram

- What kinds of information are these good at showing?

# 1. Data graphics

## Histogram

- One way to see how the values in your data set are **distributed** over subdivisions of the range of values
  - Which values are frequent? Rare?
  - Are there two distinct groups of values in your data set (bimodal distribution)?

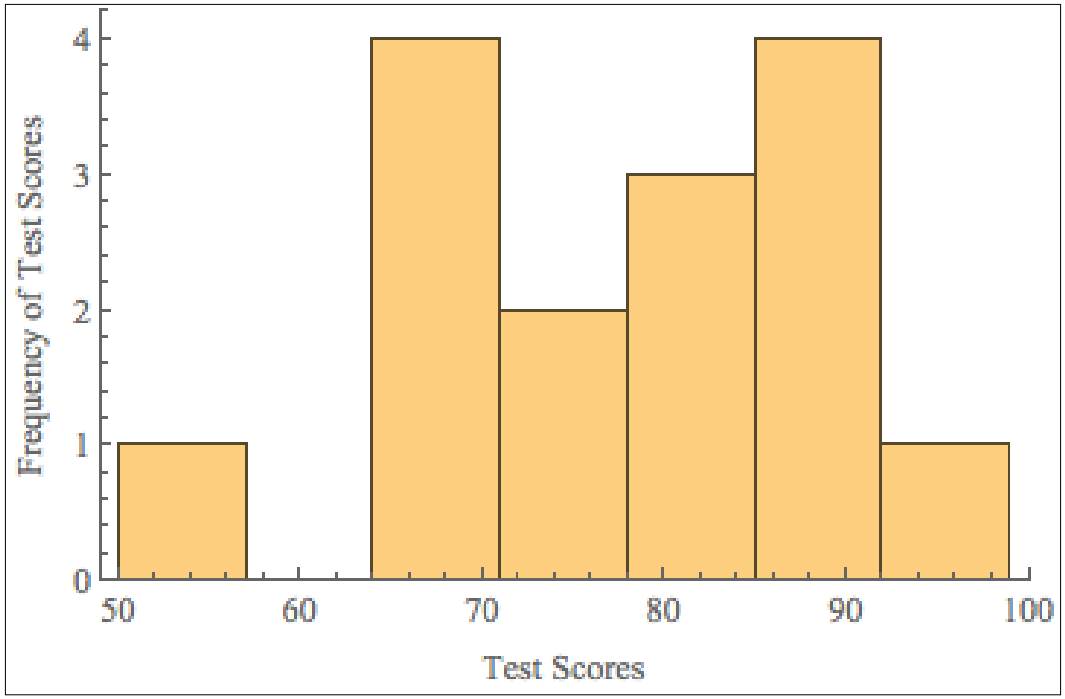
- How do we set one up?

Test scores for a geometry class:

53, 67, 69, 70, 74, 75, 79, 83, 84, 88, 88, 90, 91, 97

# 1. Data graphics

## Histogram




- Is this how you would have done a histogram for these test scores? (What is a crucial decision for histograms?)

# 1. Data graphics

## Scatterplot

- What kinds of information are these good at showing?

# 1. Data graphics

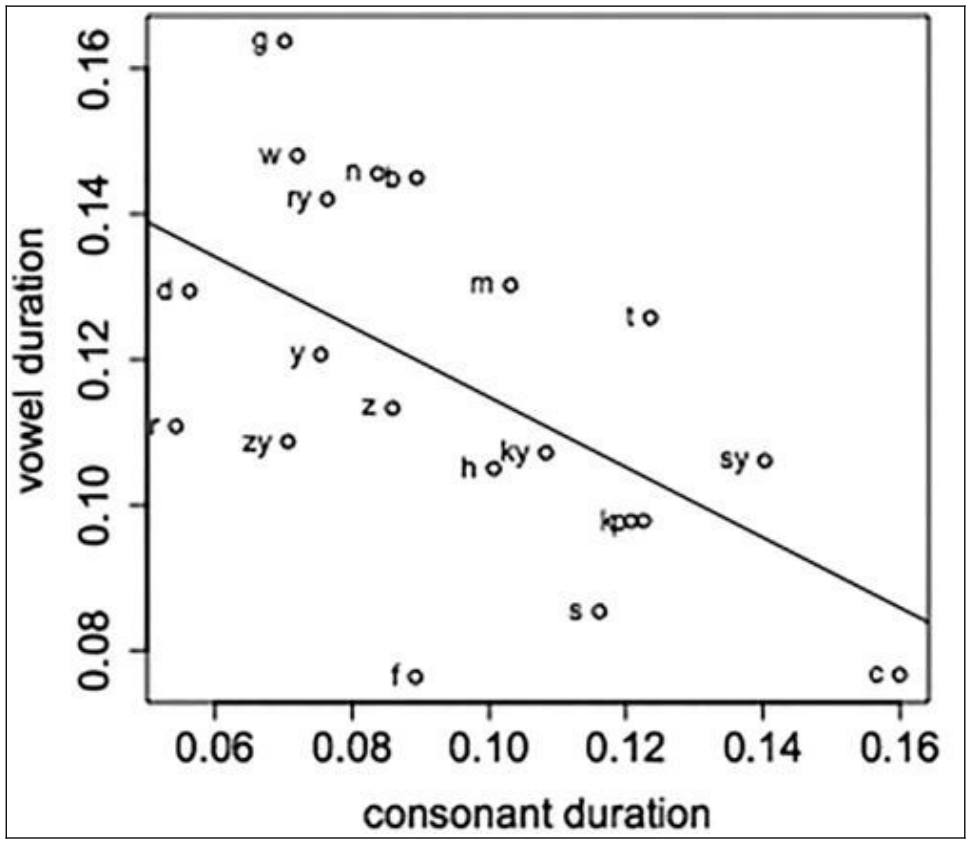
## Scatterplot

- Useful when predictor and outcome are both *continuous*
  - versus when predictor has a small number of *categories*
- Which axis to use for each variable? Usually:
  - Independent / predictor variable: *x*-axis
  - Dependent / outcome variable: *y*-axis

# 1. Data graphics

## Scatterplot

- Example from Kawahara (2017) (see today's check-in)



# 1. Data graphics

- The *Economist* (UK-based news magazine) generally has good data graphics
- Here's a blog post discussing some that could have been done better and why! https://medium.economist.com/mistakes-weve-drawn-a-few-8cdd8a42d368

# 1. Data graphics

What are some reasons to use data graphics?

# 1. Data graphics

What are some reasons to use data graphics?

- **Communication**: making your results easy for your audience to see and understand
  - Be clear on what point you most want to communicate about your data, and choose a type of data graphic that highlights that point

# 1. Data graphics

What are some reasons to use data graphics?

- **Analysis check**: Making sure you know what your data set is actually like
  - Before you start any inferential statistical analysis — look to see if there is anything going on in the data that you should take into account or be careful about

# 2. Analysis check: Anscombe's quartet

- Are mean, standard deviation (or variance), and correlation everything we need to know about a data set?
  - Famous demonstration by Anscombe (1973): four data sets

Anscombe, Francis J. 1973. Graphs in statistical analysis. *American Statistician* 27 (1): 17–21.

# 2. Analysis check: Anscombe's quartet

- Data table (https://en.wikipedia.org/wiki/Anscombe%27s_quartet)

| Data set I | | Data set II | | Data set III | | Data set 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 | 5.76 |
| 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 |

# 2. Analysis check: Anscombe's quartet

- These statistics for all four datasets are the same

(https://en.wikipedia.org/wiki/Anscombe%27s_quartet)

| Property | Value | Accuracy |
|---|---|---|
| Mean of *x* | 9 | exact |
| Sample variance of *x* | 11 | exact |
| Mean of *y* | 7.50 | to 2 decimal places |
| Sample variance of *y* | 4.125 | plus/minus 0.003 |
| Correlation between *x* and *y* | 0.816 | to 3 decimal places |

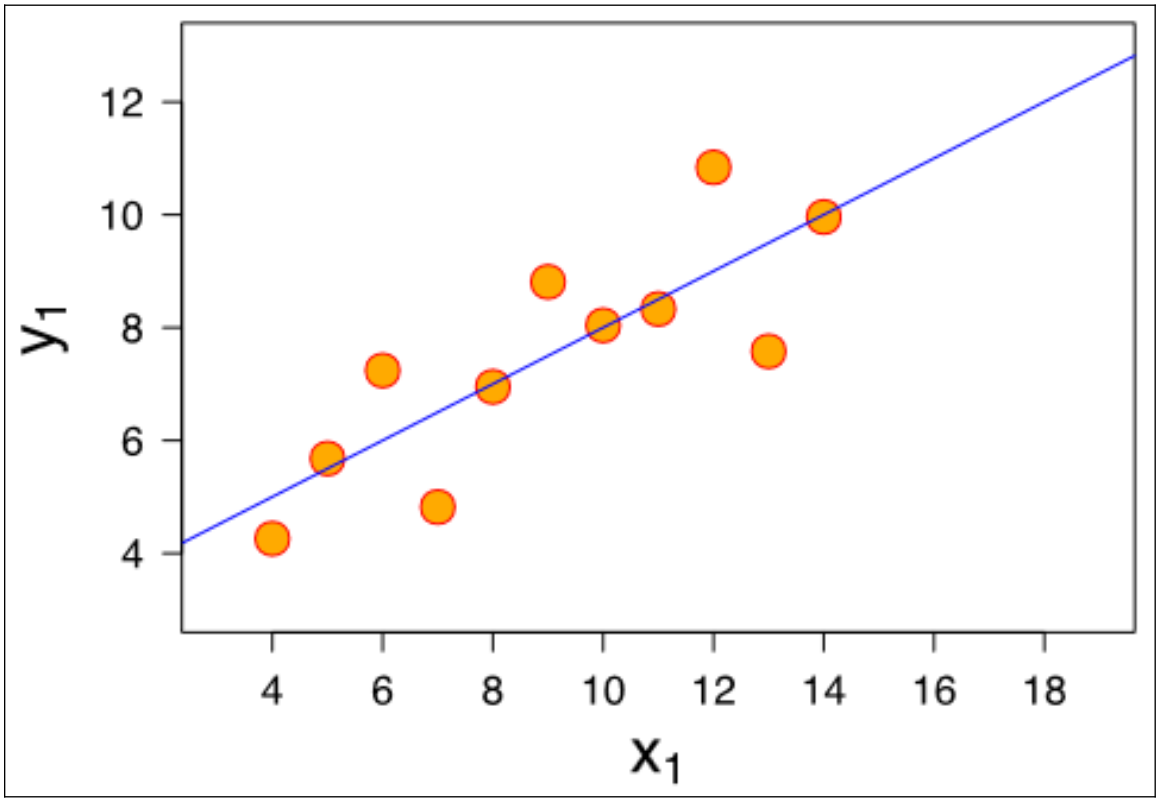
- Review: Which of these statistics are descriptive? Which are inferential?

variance = (standard deviation)$^2$

# 2. Analysis check: Anscombe's quartet

- Data graphics:
  Regression line appropriate for *x*~*y* relationship?

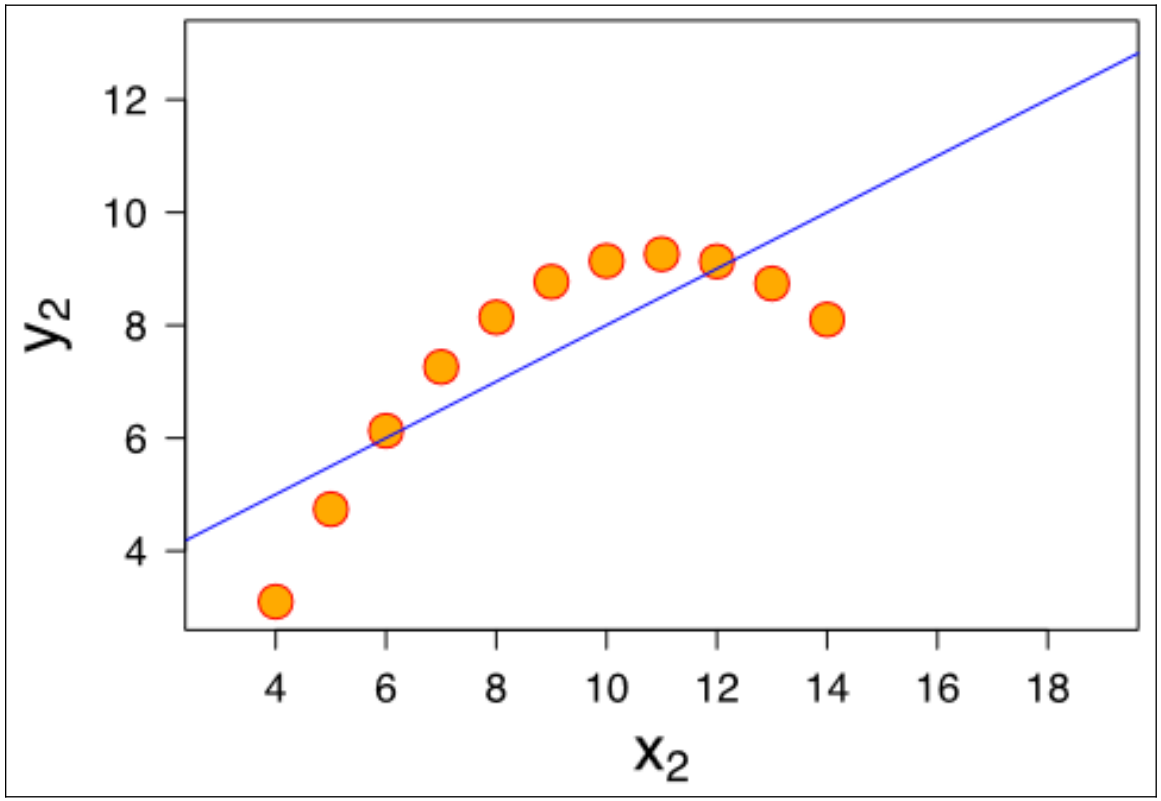
**Data set I**



# 2. Analysis check: Anscombe's quartet

- Data graphics:
  Regression line appropriate for *x*~*y* relationship?

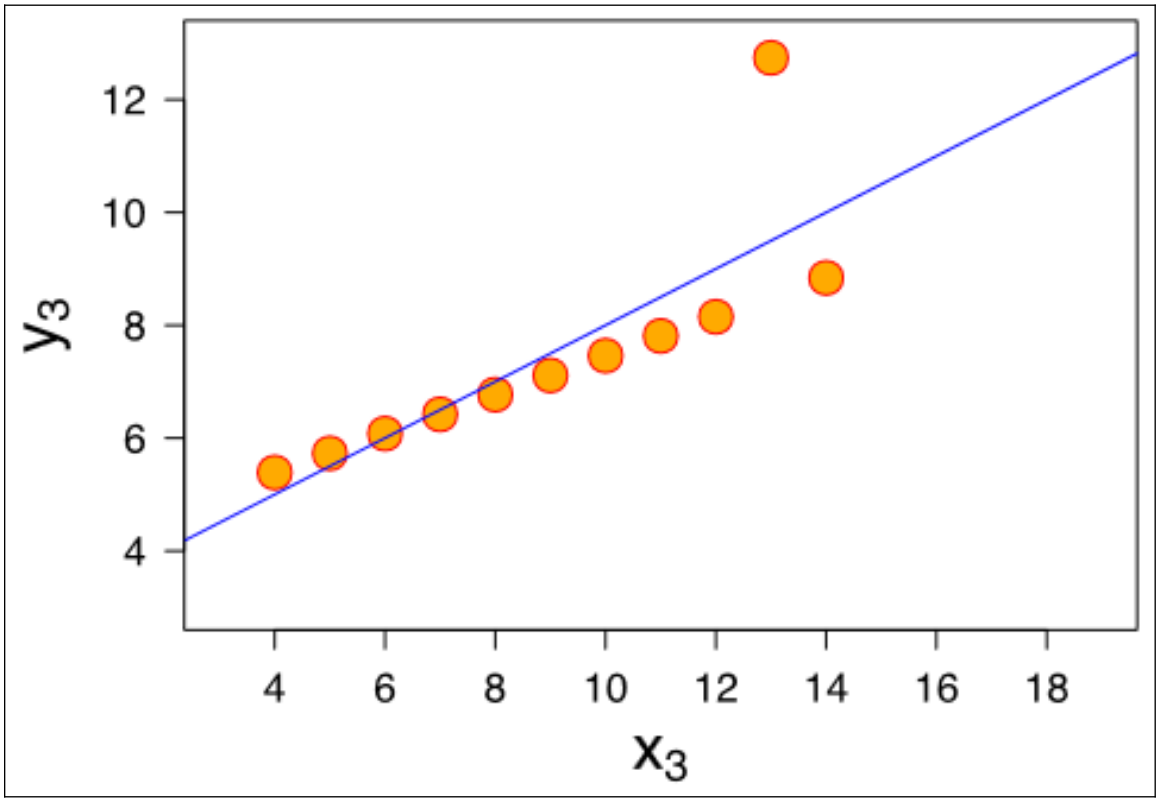
**Data set II**



# 2. Analysis check: Anscombe's quartet

- Data graphics:
  Regression line appropriate for *x*~*y* relationship?

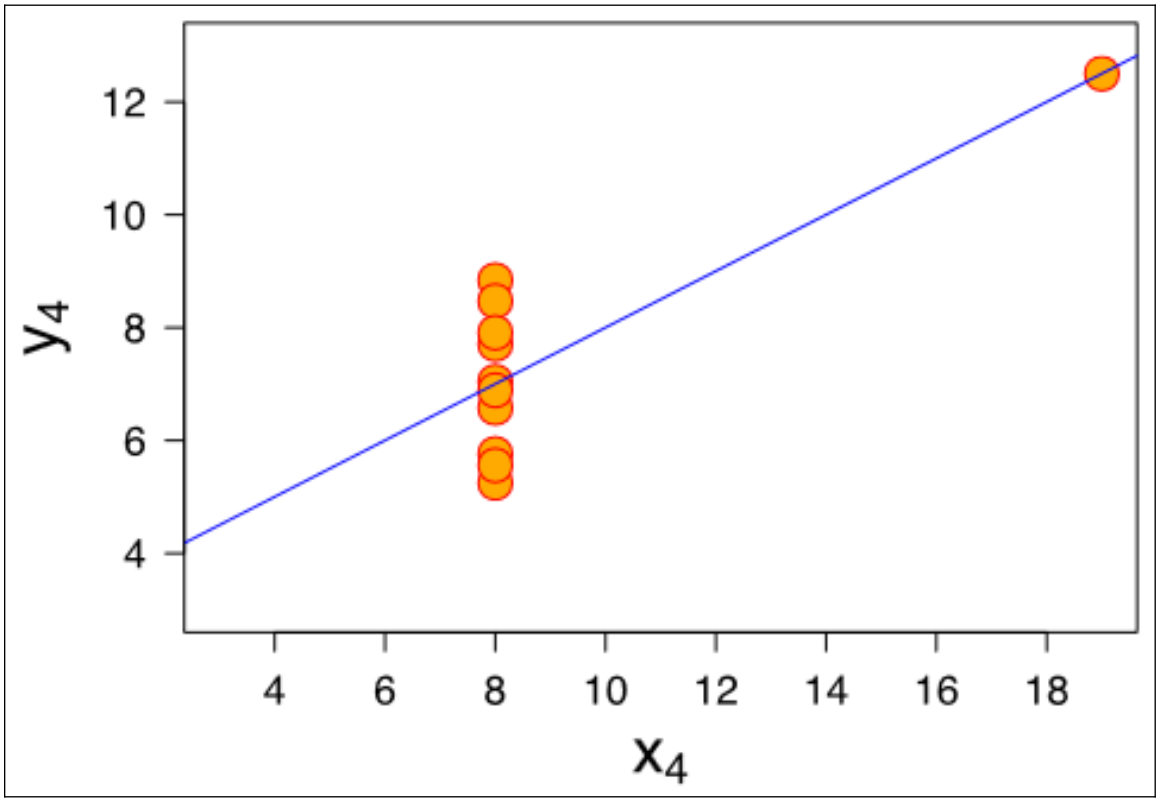
**Data set III**



# 2. Analysis check: Anscombe's quartet

- Data graphics:
  Regression line appropriate for *x*~*y* relationship?

**Data set IV**



# 2. Analysis check: Anscombe's quartet

- Some morals of this story:
  - Not all *x*~*y* relationships are linear (minor deviations from an ideal straight line)
  - Outliers can skew our interpretation of a data set

- Making data graphics is one way to discover situations like these

# 3. Experiment design

Even before data collection begins—what kinds of factors need to be considered in designing an experiment in the first place?

# 3. Experiment design

**Participants**

- Are they representative of the groups of people we want to know about?
- Many study participants in linguistics or psychology are university undergraduates…
  - WEIRD! = from ***W***estern, ***e***ducated, ***i***ndustrialized, ***r***ich, and ***d***emocratic societies (Henrich et al. 2010)
  - typically aged 18-24 or so

Henrich, J., Heine, S., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences* 33(2-3): 61-83.

# 3. Experiment design

## Confounds

- What potential confounding factors can we identify?
- Can we reduce or eliminate them, or at least include them explicitly in our analysis?

# 3. Experiment design

**Task design**

- Is the experimental task really similar to what we want to study?
- Is it likely to affect different groups of participants differently?

# 3. Experiment design

**Data analysis**

- How is the data to be coded? Is the coding protocol explicit? Is it reliable across coders? (Can we avoid coder bias?)
- Should outliers be excluded from analysis? If so, how can they be safely (objectively) identified?
- Appropriateness of statistical analysis: the right test; not too many tests ('fishing')

# 3. Experiment design

**Problems in replication and underreporting of null results**

- More exciting to publish a result than a null result
- More exciting to publish a new study than a replication
- So it might be the case that various effects are not as robust as the literature makes it seem

# 4. Summary and upshot

**Statistics**

- A little background in statistics can help us get a sense of what the results section in a research paper is saying
- Inferential statistics can help us understand whether an (apparent) numerical difference is meaningful

# 4. Summary and upshot

**Data graphics**

- Useful for communication and for checking your analysis

**Experiment design**

- It is difficult to design a good experiment
- When reading a research report, keep an eye out for some of these pitfalls

# 4. Summary and upshot

- Should we trust the science in research papers?
  - Think about some of the factors we discussed on the first day of the semester