MATH 10

INTRODUCTORY STATISTICS

Tommy Khoo

Your friendly neighbourhood graduate student.

It is Time for Homework! (´･ω･`)

- First homework + data will be posted on the website, under the homework tab.
- And also sent out via email.

- 30% weekly homework. Will give out first today, due the following Tues.
- Each homework might have different points assigned but carry the same weight.
- Your other homework: read and understand the relevant chapters in the textbook.

Week 2

- **Chapter 4 – Bivariate Data** ← **today's lecture**

Data with two/paired variables, Pearson correlation coefficient and its properties, general variance sum law

- **Chapter 6 – Research Design** ← **today's lecture**

Data collection, sampling bias, causation.

- **Chapter 5 – Probability**

Probability, gambler's fallacy, permutations and combinations, binomial distribution, Bayes' theorem.

Quick Recap on Important Parts of Last Week

- **Central Tendency**

- **Mean -** minimizes sum of squared deviations.
- Balances the scale because deviations of the mean sums to zero.

- **Median -** minimizes sum of absolute deviations.
- "The" Median is unique for this course but there is a general definition that allows for more than one median. More "robust" to outliers than mean.

- **Mode -** can have no modes or multiple modes.
- Common usage of the word "bi-modal" often does not strictly uses the definition of a mode.

Quick Recap on Important Parts of Last Week

- **Variability**

- **Population variance -** $\sigma^2 = \frac{\sum(X-\mu)^2}{N}$,
- Take square root to get population standard deviation.

- **Estimator of the variance -** $s^2 = \frac{\sum(X-M)^2}{N-1}$
- Dividing by N underestimates the variance. Bessel's Correction.

- **Estimator of the standard deviation** - s, is actually an overestimate!

Sum of Variance for Uncorrelated Variables → leads in to chapter 4

- Chapter 3, Section 19 : $\sigma^2_{X\pm Y} = \sigma^2_X + \sigma^2_Y$ (population version)

- Curiously, the textbook left out the sample version: $s^2_{X\pm Y} = s^2_X + s^2_Y$

- First, let's talk about the **<u>population</u>** case.
- In the context of chapter 3, X and Y are (usually) data for two different populations.
- The math works but it is actually rather contrived/artificial (will explain later).
- Context of chapter 4 makes more sense / is more natural: paired bivariate data.

Sum of Variance for Uncorrelated Variables → leads in to chapter 4

$$\sigma^2_{X\pm Y} = \sigma^2_X + \sigma^2_Y$$

- My sloppy example last lecture: let's pretend our two populations NH and CA has the same size N (*of course they don't*).
- Let $\mathrm{X} =$ income of a person in NH, and $\mathrm{Y} =$ income of a person in CA.
- When comparing two populations, we usually care about the differences in their mean.
- E.g. if we have population data, we can easily calculate $\mu_Z = \mu_X - \mu_Y$ to see if $\mu_Z > 0$.
- Later in this course you will see that most comparison of two populations are based on this approach: whether there is a difference in population means.

Sum of Variance for Uncorrelated Variables → leads in to chapter 4

$$\sigma^2_{X \pm Y} = \sigma^2_X + \sigma^2_Y$$

- New variable $Z = X - Y$. Then, $\mu_Z = \mu_X - \mu_Y$ by our calculation last lecture.
- However, $Z_i = X_i - Y_i$ depends on choosing a pairing of X_i, Y_i.
- For example, if this pairing is done with people in similar industries or job types, there would probably be some correlation (since both NH and CA are in the USA).
- You can get zero correlation by pairing them randomly but this is an artificial choice.
- Chapter 4 : paired bivariate data → the pairing is natural / given to us.

Sum of Variance for Uncorrelated Variables → leads in to chapter 4

$$s_{X \pm Y}^2 = s_X^2 + s_Y^2$$

- Now, let's talk about the **sample** case.
- **Note:** the textbook did not give the sample version of this formula in chapter 3.

- The sampling process can affect whether the variables X and Y are correlated.
- Your textbook offered a way to make sure X and Y are uncorrelated:
- Pick a X_i at random, then pick a Y_i at random. Pair them together (X_i, Y_i).
- Since the pairing is done randomly, X and Y will be uncorrelated.

Sum of Variance for Uncorrelated Variables → leads in to chapter 4

$$s_{X \pm Y}^2 = s_X^2 + s_Y^2$$

- The sampling process can affect whether the variables X and Y are correlated.
- Your textbook offered a way to make sure X and Y are uncorrelated:
- Pick a X_i at random, then pick a Y_i at random. Pair them together (X_i, Y_i).
- Since the pairing is done randomly, X and Y will be uncorrelated.

- Random pairing makes a lot more sense when it comes to samples. E.g. difference between two dice rolls.
- Not so useful for comparing two populations. As mentioned previously, in chapter 9 will see that we care more about estimating the difference (if any) between the two population means.
- We will have a powerful new tool in chapter 9 : the sampling distribution of the mean.

Chapter 4 – Bivariate Data

- Data with 2 quantitative variables for each individual / object.
- In practice, you often have *k* quantitative variables for each individual.
- E.g. using *cough* Facebook *cough*, we might have access to this set of data about each person: *(age, gender, location, number of friends)*

The pairs of ages in Table 1 are from a dataset consisting of 282 pairs of spousal ages, too many to make sense of from a table. What we need is a way to summarize the 282 pairs of ages. We know that each variable can be summarized by a *histogram* (see Figure 1) and by a mean and standard deviation (See Table 2).

Table 1. Sample of spousal ages of 10 White American Couples.

Husband	36	72	37	36	51	50	47	50	37	41
Wife	35	67	33	35	50	46	47	42	36	41

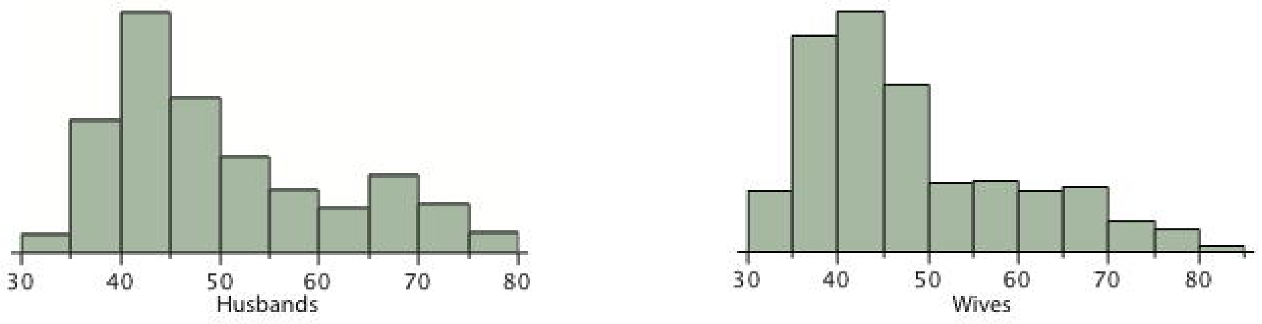


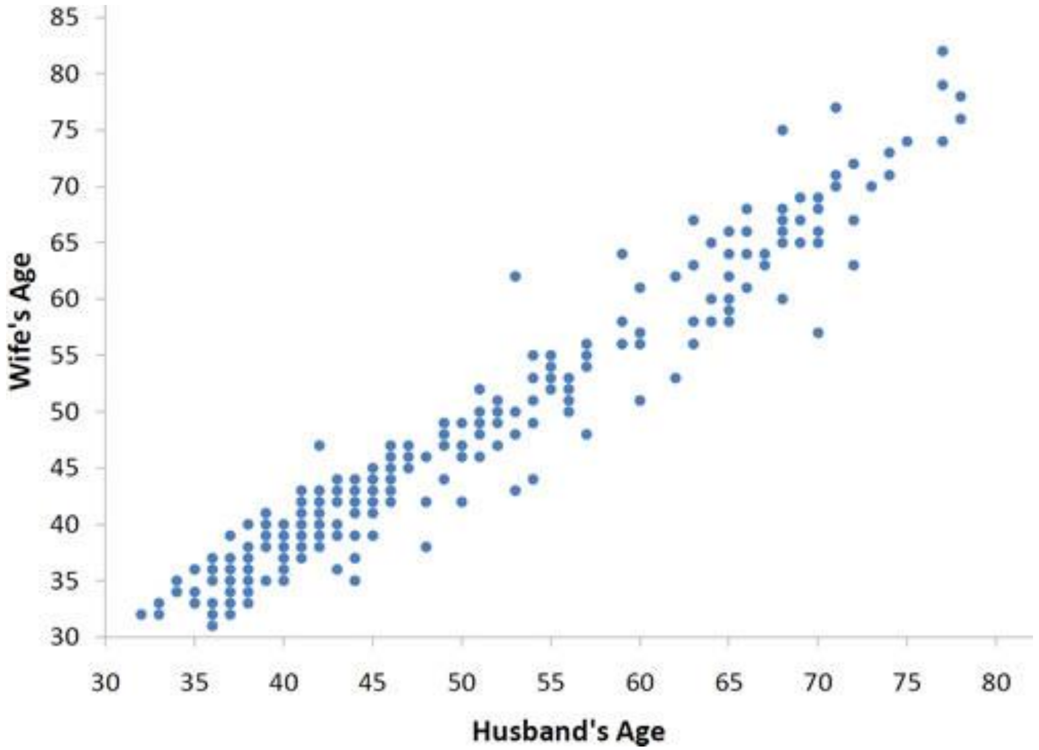
Figure 1. Histograms of spousal ages.

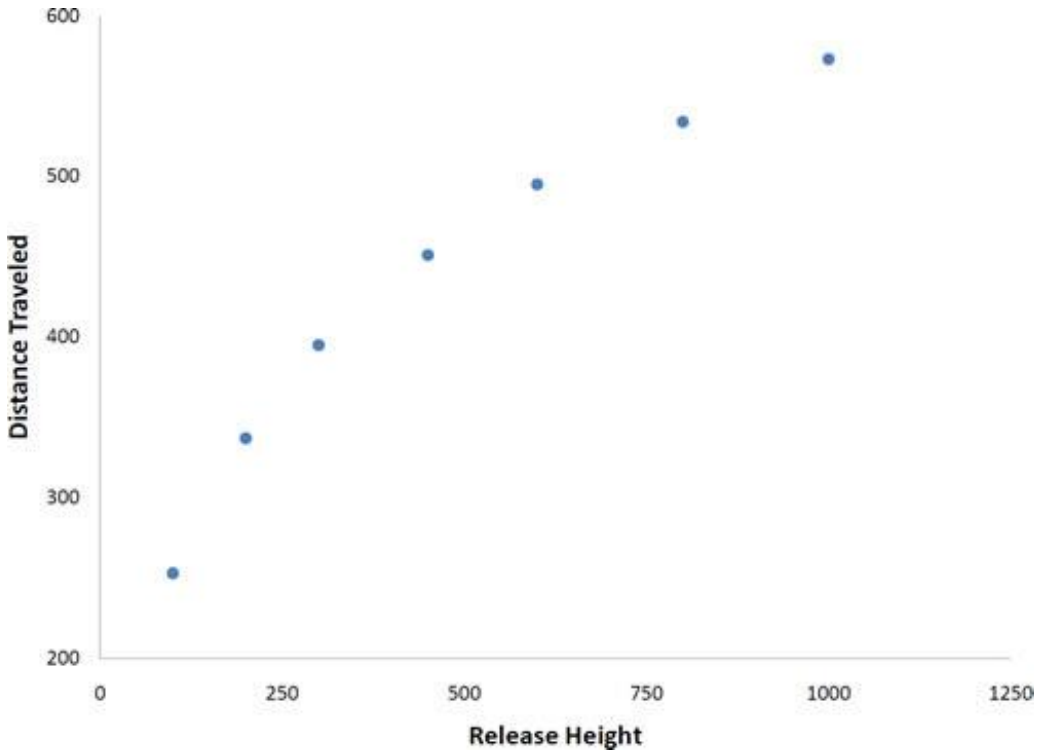
Table 2. Means and standard deviations of spousal ages.

	Mean	**Standard Deviation**
Husbands	49	11
Wives	47	11

Chapter 4 – Bivariate Data

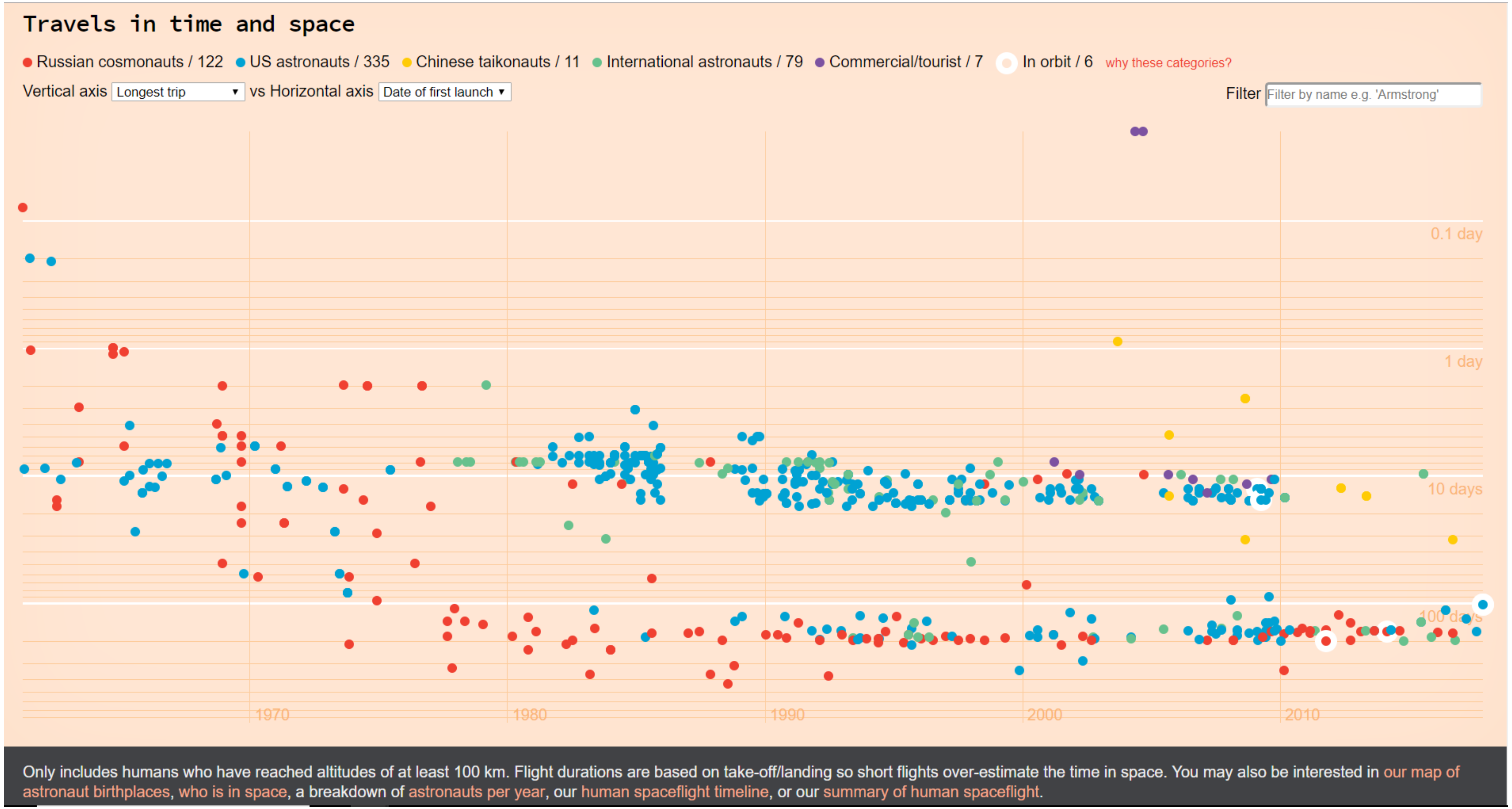
- Dot or scatter plots from chapter 2 are great for visualizing bivariate data.
- Important distinction : linear vs non-linear relationship.





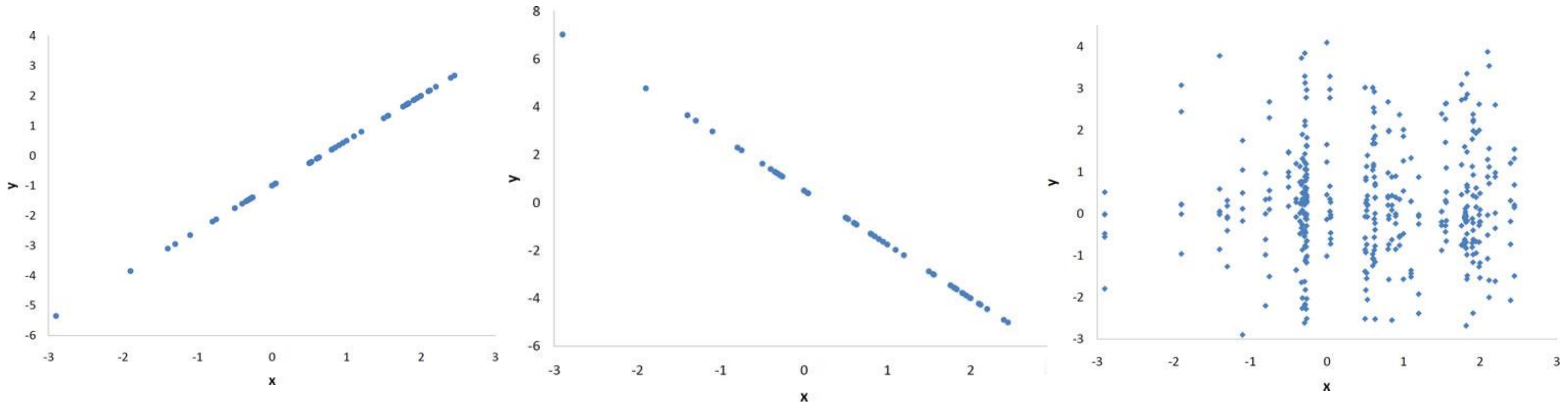
Very awesome example of bivariate data and dot/scatter plot from:

https://cosmos-book.github.io/human-spaceflight/index.html



Pearson Correlation : Chapter 4 – Section 3

- We can eyeball an intuitive difference between these three sets of data.
- How can we quantify this difference?



Pearson Correlation : Chapter 4 – Section 3

- Super official name: Pearson product-moment correlation coefficient.
- Usual name: correlation coefficient, r.

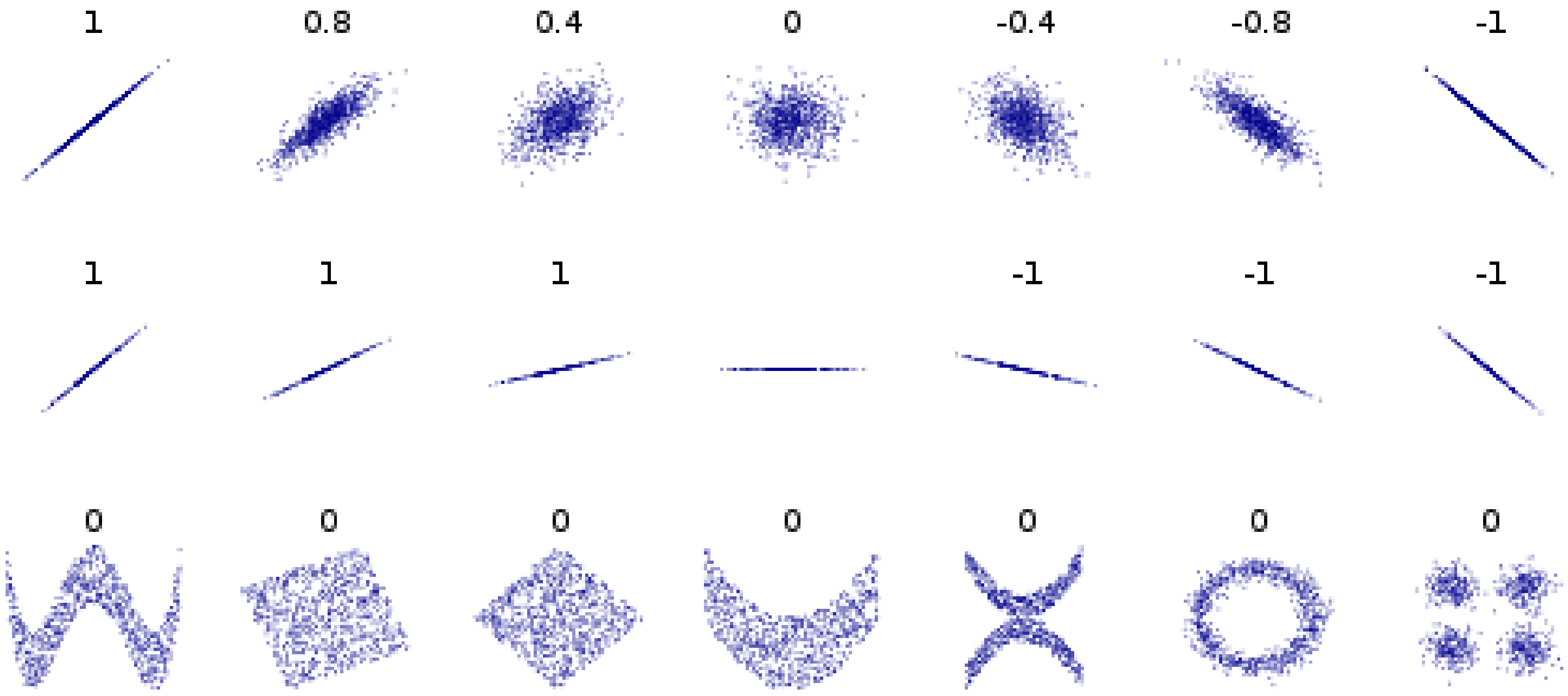
- $-1 \leq r \leq 1$

- $r = 1 \Rightarrow$ perfect positive linear relationship
- $r = -1 \Rightarrow$ perfect negative linear relationship
- $r = 0 \Rightarrow$ no linear relationship

- **Warning:** only applicable for linear relationships! → next slide

Pearson Correlation : Chapter 4 – Section 3

- **Warning:** only applicable for linear relationships!
- There are other mathematical measures of bivariate relationships.
- Guessing r : Chapter 4 – Section 4, important skill for the exam.



Properties of r : Chapter 4 – Section 5

Also, important for the exam.

- **Symmetric:** correlation coefficient of X and Y = correlation coefficient of Y and X

(you will see why this matters when we give you the formula later)

- Unaffected by linear transformations aX + b, when a > 0.

(useful when working with different units of measurement)

- **Warning:** affected by non-linear transformation!

Break Time! \o/

See you back here after 10 minutes.

Cat meme kindly provided by:
lovecuteanimals.com/lol/24268/

Computing r : Chapter 4 – Section 6

Population version

$$\rho = \frac{\sum(X-\mu_X)(Y-\mu_Y)}{\sqrt{\sum(X-\mu_X)^2}\sqrt{\sum(Y-\mu_Y)^2}}$$

Sample version

$$r = \frac{\sum(X-M_X)(Y-M_Y)}{\sqrt{\sum(X-M_X)^2}\sqrt{\sum(Y-M_Y)^2}}$$

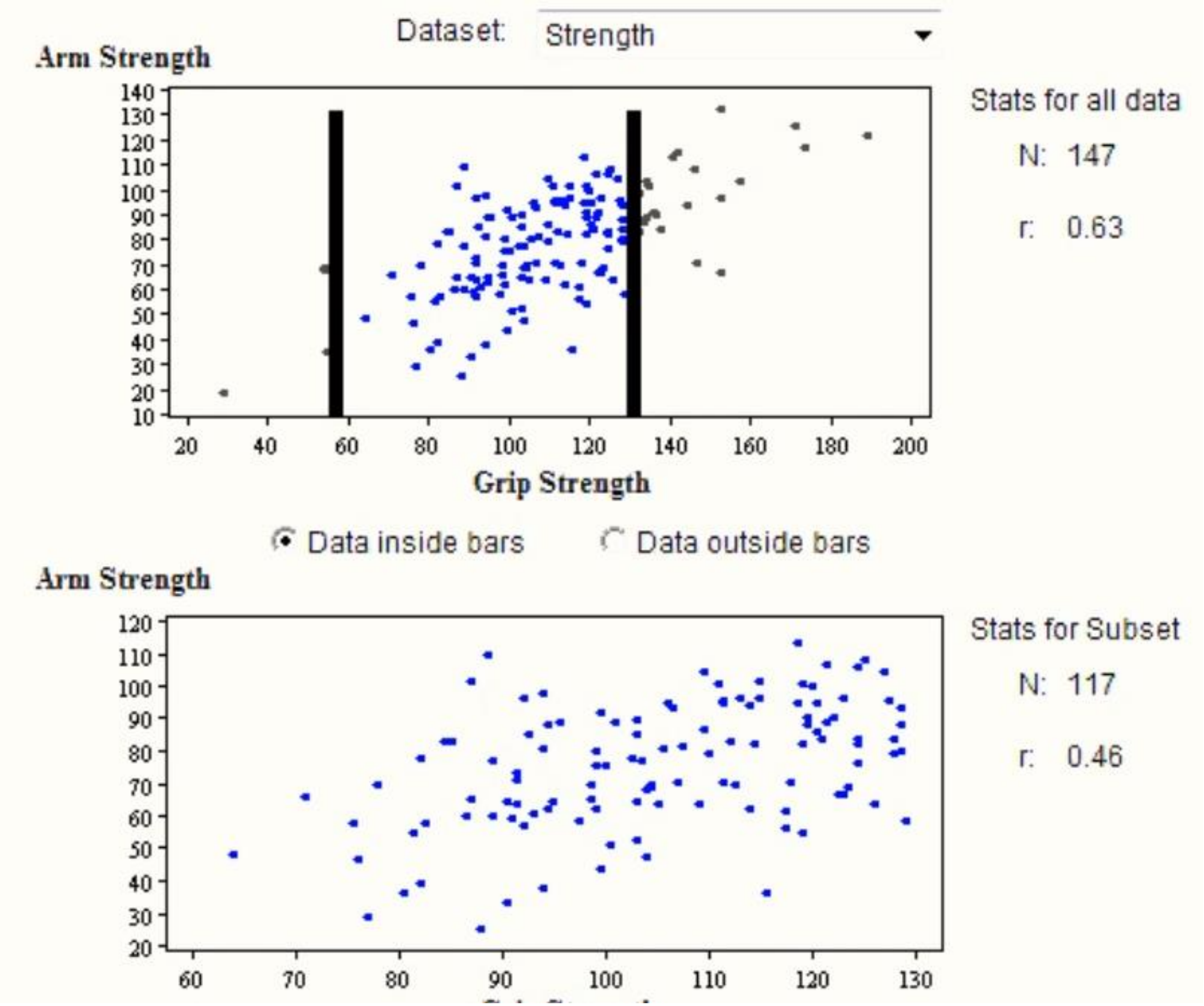
For paired data (X, Y). Curiously, the textbook uses ρ later but does not talk about computing ρ.

Chapter 4 – Section 7

A possible exam question might be: what might happen to the value of r if you throw away outliers as shown in the figure →

- Usually, more data is better than less data.
- If you are throwing away data, you need a very good reason.
- If you are throwing away data just to create stronger evidence, to support your hypothesis…don't let anyone know I taught you statistics. Please? :p

A quick word about restricting the range of your data…



Variance Sum Law : Chapter 4 – Section 8

Population version: $\sigma_{X\pm Y}^2 = \sigma_X^2 + \sigma_Y^2 \pm 2\rho\sigma_X\sigma_Y$

Sample version: $s_{X\pm Y}^2 = s_X^2 + s_Y^2 \pm 2rs_Xs_Y$

- These formulas and the exercises in the book assumed that the data comes with a pairing (bivariate data).
- We will tell you how the data is paired or it will come with an obvious pairing. I will be surprised if you had to choose a pairing in this course.

Week 2

- **Chapter 4 – Bivariate Data** ← **today's lecture**

Data with two/paired variables, Pearson correlation coefficient and its properties, general variance sum law

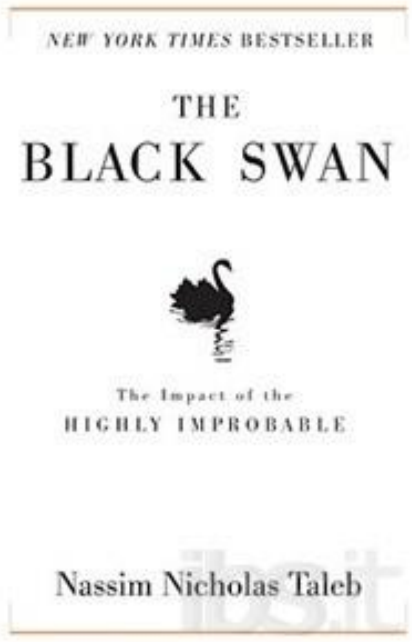
- **Chapter 6 – Research Design** ← **today's lecture**

Data collection, sampling bias, causation.

- **Chapter 5 – Probability**

Probability, gambler's fallacy, permutations and combinations, binomial distribution, Bayes' theorem.

Research Design, Chapter 6



- Theories in science can never be proved since one can never be 100% certain that a new empirical finding inconsistent with the theory will never be found.
- Scientific theories must be potentially falsifiable. If a theory can accommodate all possible results then it is not a scientific theory. Therefore, a scientific theory should lead to testable hypotheses.
- If a hypothesis derived from a theory is confirmed then the theory has survived a test and it becomes more useful and better thought of by the researchers in the field. A theory is not confirmed when correct hypotheses are derived from it.

"In so far as a scientific statement speaks about reality, it must be falsifiable; and in so far as it is not falsifiable, it does not speak about reality."

- Karl Popper

Public Service Announcement:

We are skipping Chapter 6, Section 3, titled “Measurement”.

Sampling Bias, Chapter 6, Section 5

For the exams, you do not have to memorize the names of these biases. But you should be able to recognize when a bias could occur.

- **Self – Selection Bias** : asking people to nominate themselves, not randomizing treatment and control group.

- **Under-coverage Bias** : the way you collect samples could discourage certain groups from responding. E.g. the poor not having telephones example in the textbook.

- **Survivorship Bias** : famous world war 2 aircraft armor example in the textbook, studying only people who perform well in gambling or the stock market.

Experimental Design, Chapter 6, Section 6

Public service announcement: we have decided to nuke this section from orbit (it's the only way to be sure).

Memorize: no. Understand: yes.

- **Between-Subjects :** different groups used for different possibility of the independent variable.
- **Within-Subject :** same subject tested with all possibilities of the independent variables.

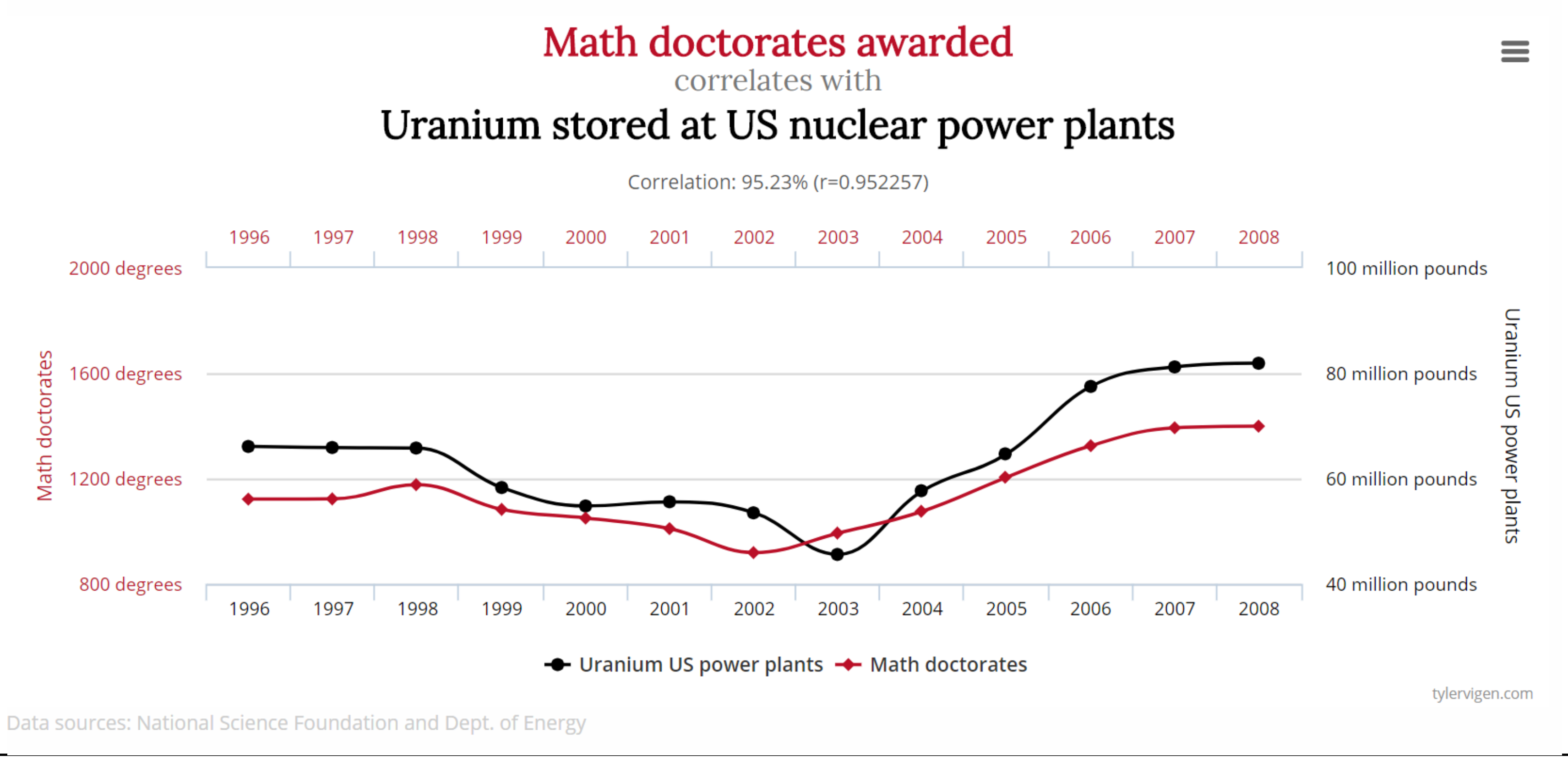
Both cases : want to study the relationship between the independent variable and the dependent variable.

- **Counter-balancing :** in a within-subject design, the order which you test the possibilities might affect the result. Need to randomize the order.
- **Multi-Factor Between Subject :** two or more independent variables.

Causation, Chapter 6, Section 7

- **Correlation does not imply causation :**

http://www.tylervigen.com/spurious-correlations



Causation, Chapter 6, Section 7

- **Correlation does not imply causation**
- **Confounding/"Third" variable**

How can we establish causation?

This is a deep philosophical question. Even defining causation is difficult.

From a statistics point of view: we use statistics to guide us towards possible theories. Then, try to explain why the theory should be true.

E.g. smoking is correlated with cancer, but this is supported by scientific understanding of the underlying mechanism.