Lesson A3
Variables relationship, Research Design, Probability
GEO1001.2020

Clara García-Sánchez, Stelios Vitalis

Resources adapted from:
- David M. Lane et al. (http://onlinestatbook.com)
- Allen B. Downey et al. (https://greenteapress.com/wp/think-stats-2e/)

Lesson A3

Variables Relationship

Overview

- Bivariate data
- Correlation
- Covariance
- Pearson correlation
- Non-linear relationships
- Spearman’s rank correlation
- Correlation and causation

Overview

- **Bivariate data**
- Correlation
- Covariance
- Pearson correlation
- Non-linear relationships
- Spearman’s rank correlation
- Correlation and causation

Bivariate Data

Often when performing an experiment, more than one variable is collected. Bivariate data, consists of two quantitative variables. Two variables are related if knowing one gives you information about the other.

Let's discuss with an example:

do you think people tend to marry people of the same age?

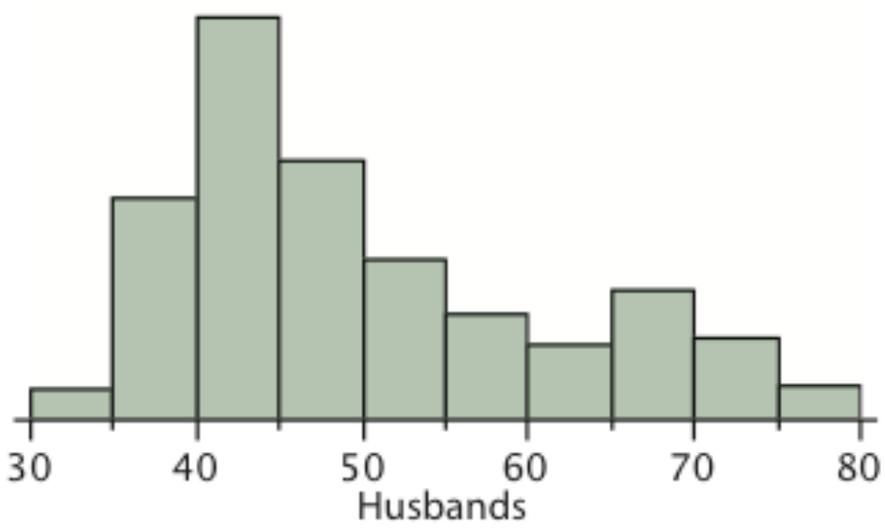
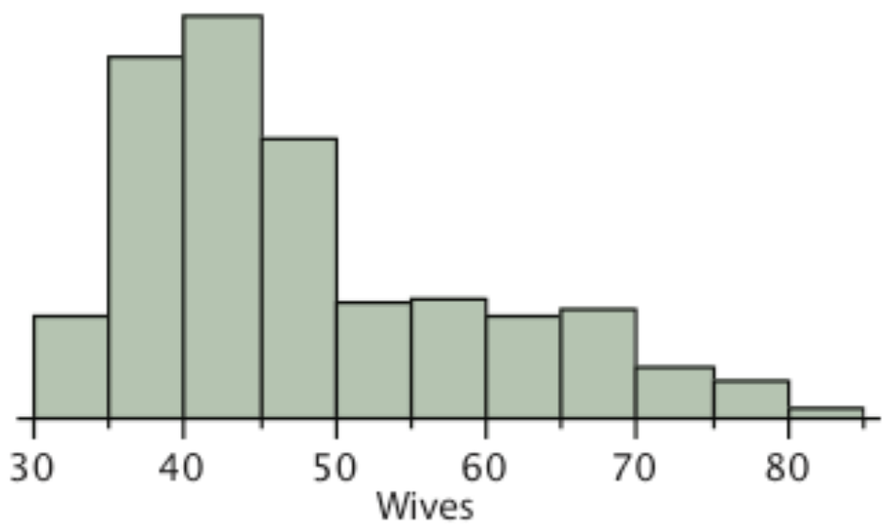


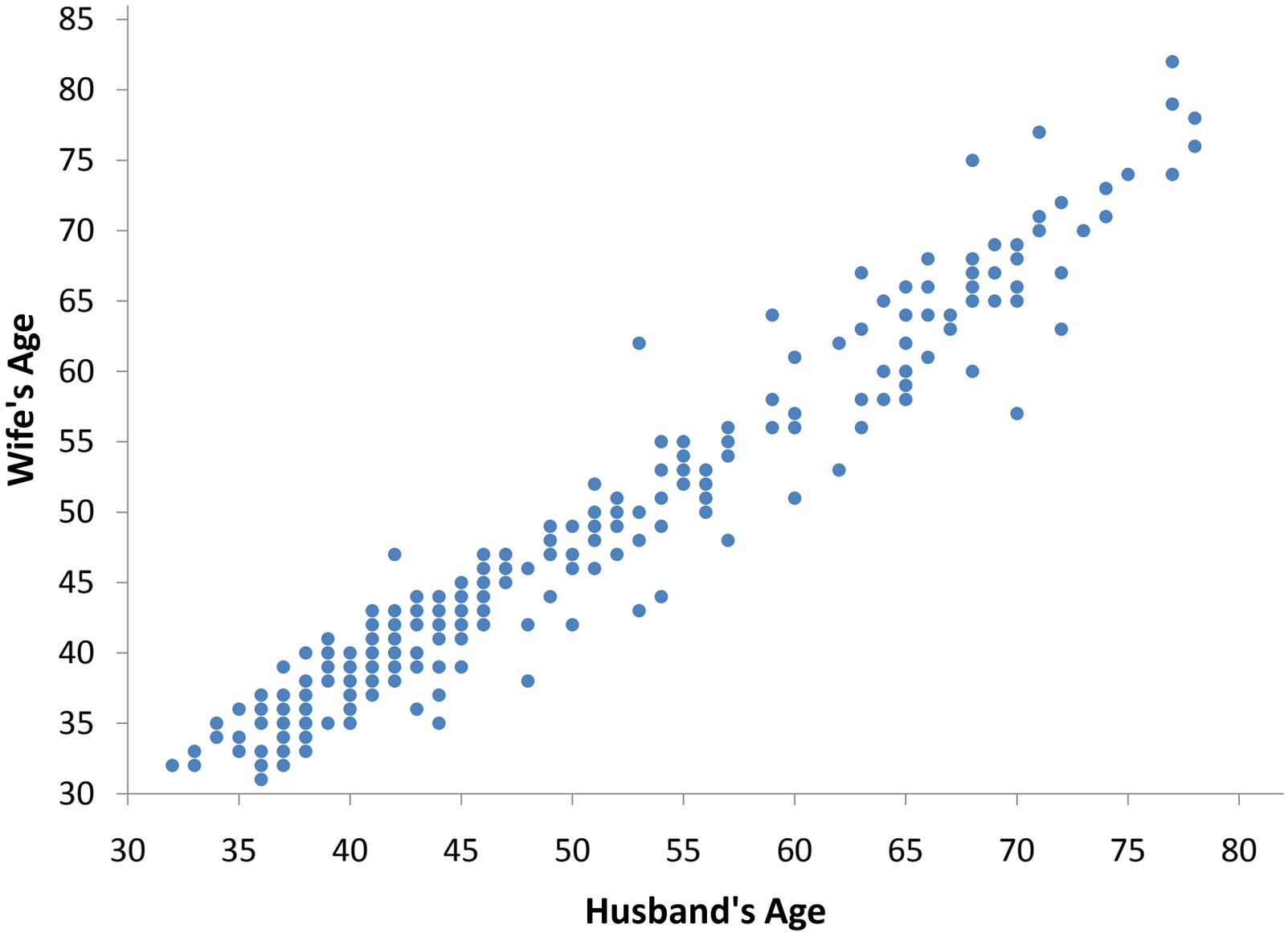
Figure 1. Histograms of spousal ages.



Histograms for 282 pairs of ages.

Bivariate Data

A better way to plot this is with a **scatter plot**:



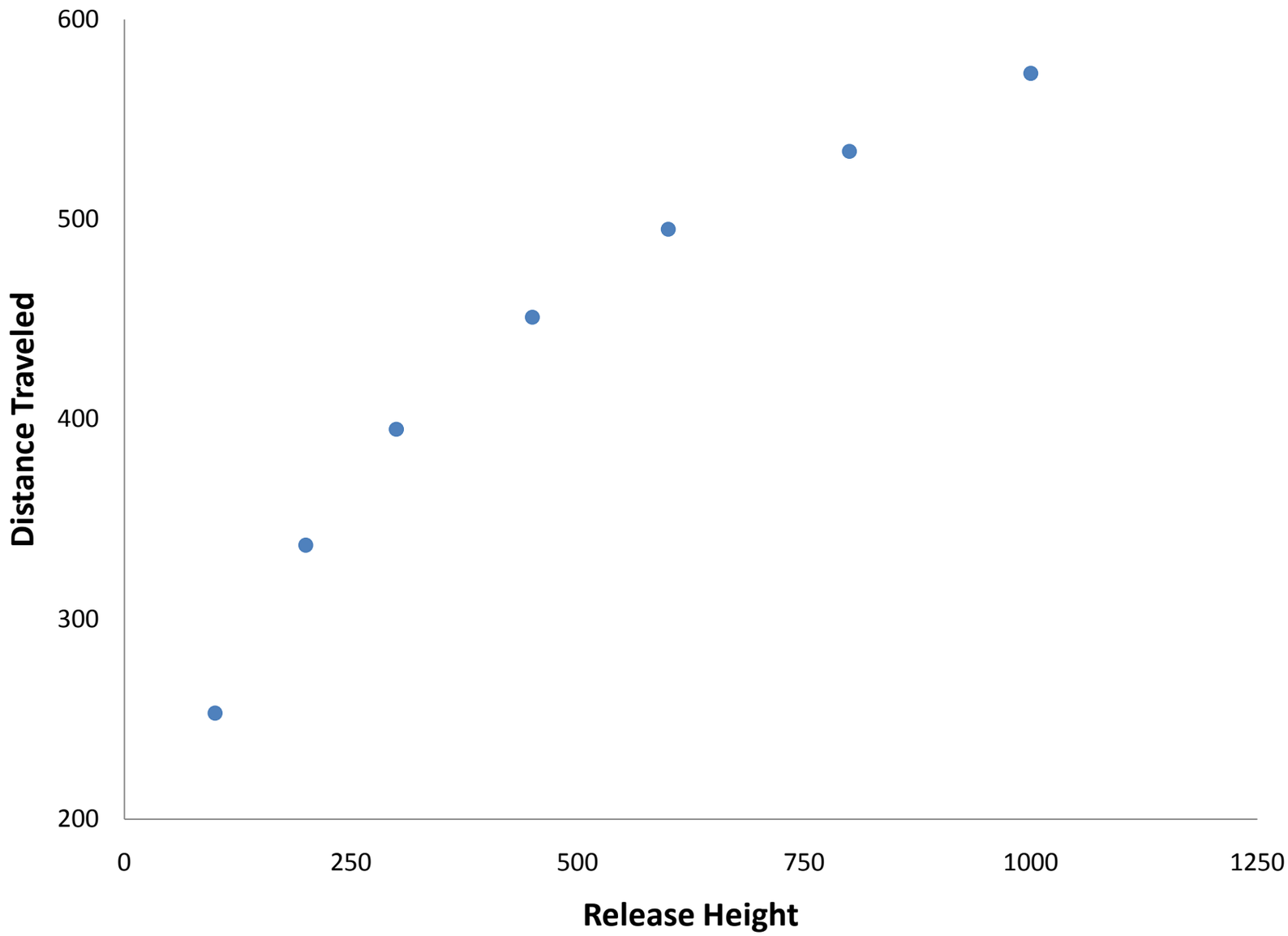
What can we deduce now?

- Both variables a **positively correlated**, when one increases the other as well.
- The points cluster along a line, which means that the **relation** between both variables is **linear**.

Bivariate Data

Be aware that not all plots show linear relations:

Galileo relating distance travelled and released height in a projectile —> fit a **parabola** (not a line)



Overview

- Bivariate data
- **Correlation**
- Covariance
- Pearson correlation
- Non-linear relationships
- Spearman’s rank correlation
- Correlation and causation

Correlation

A **correlation** is a statistic to quantify the strength of the relationship between 2 variables.

A common challenge with computing correlation is that the variables we would like to compare may be in different units, and come from different distributions. For these there are 2 solutions:

1) Transform each value to standard score, which is the number of standard deviation from the mean —> this leads to what is called "*Pearson product-moment correlation coefficient*"

$$z_i = \frac{(x_i - \mu)}{\sigma}$$

2) Transform each value to its rank, which is its index in the sorted list of values —> this leads to what is called "*Spearman rank correlation coefficient*"

Overview

- Bivariate data
- Correlation
- **Covariance**
- Pearson correlation
- Non-linear relationships
- Spearman's rank correlation
- Correlation and causation

Covariance

Covariance is a measure of the tendency of two variables to vary together. Imagine we have two series, X and Y, their deviations from the mean are:

$$dx_i = x_i - \overline{x}$$
$$dy_i = y_i - \overline{y}$$

If X and Y vary together, their deviations tend to have the same sign. If we multiply them together, the product is positive when they have the same sign, or negative otherwise, so covariance is the mean of this two products:

$$Cov(X, Y) = \frac{1}{n} \sum dx_i \cdot dy_i$$

Overview

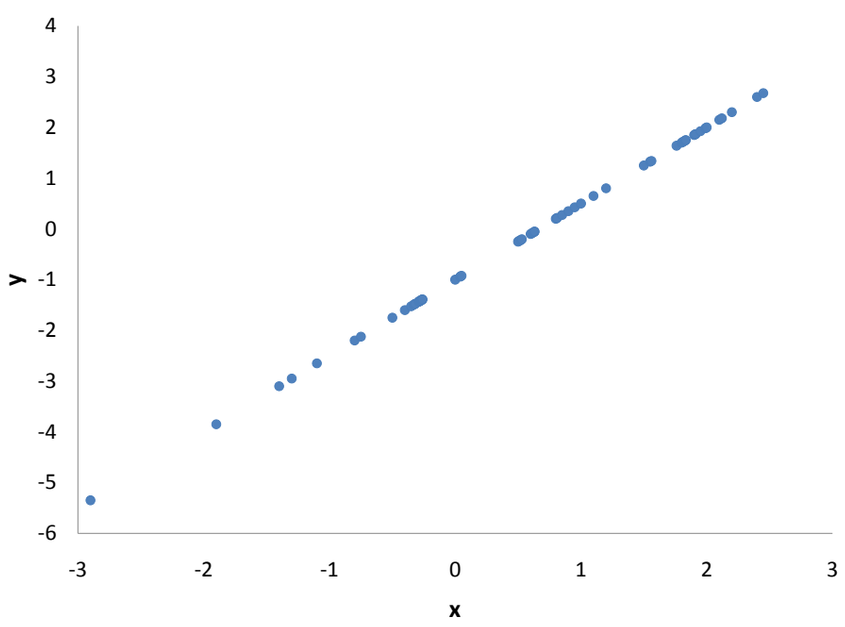
- Bivariate data
- Correlation
- Covariance
- **Pearson correlation**
- Non-linear relationships
- Spearman's rank correlation
- Correlation and causation

Pearson correlation

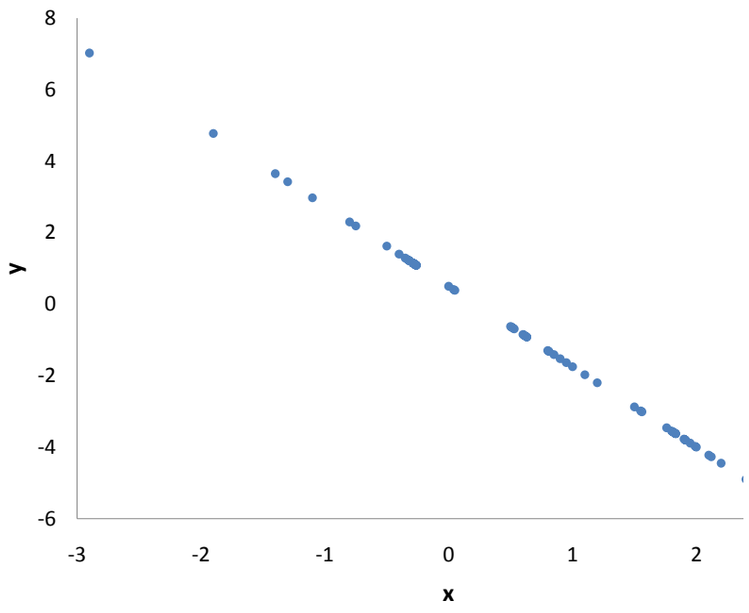
Pearson correlation: it is a coefficient that measures the strength of the linear relationship between two variables.

If the relationship between the variables is not linear, then the correlation coefficient is not representative.

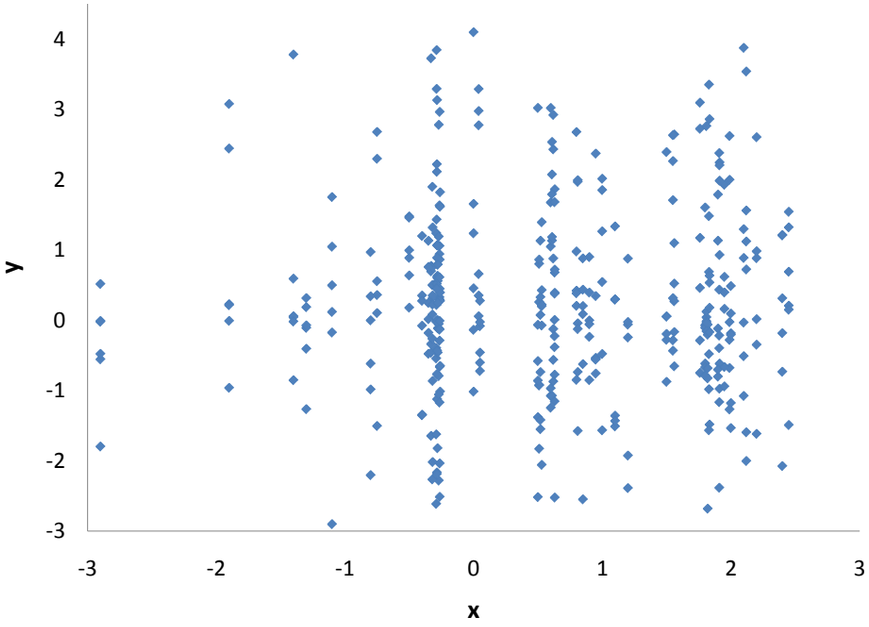
Properties: It can range from -1 to 1, it is symmetric, it is not affected by linear transformations



r=1



r=-1



r=0

Pearson correlation

$$\rho_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

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

$$\rho = \frac{1}{n}\sum p_i \qquad\qquad p_i = \frac{(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sigma_x \sigma_y}$$

Pearson correlation

http://wikipedia.org/wiki/Correlation_and_ dependence

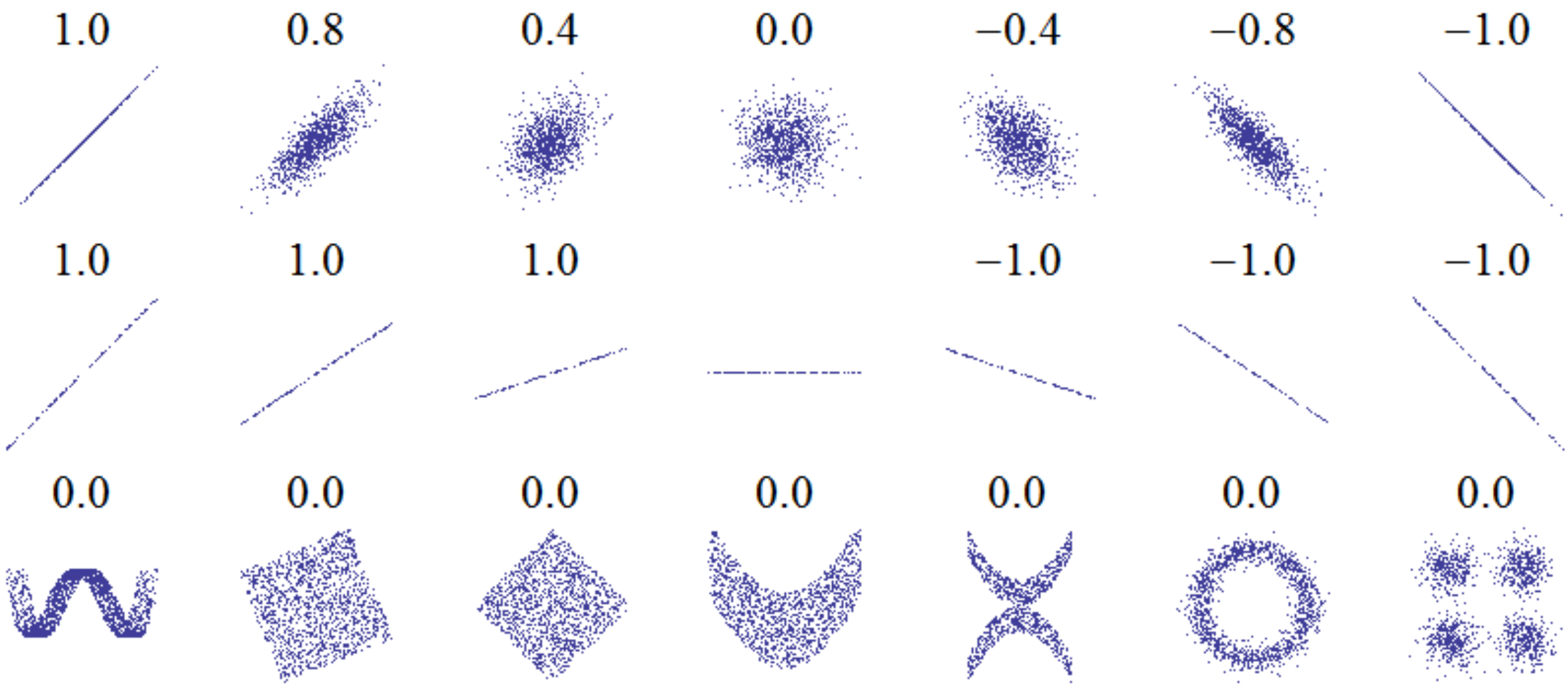


Figure 7.4: Examples of datasets with a range of correlations.

Overview

- Bivariate data
- Correlation
- Covariance
- Pearson correlation
- **Non-linear relationships**
- Spearman’s rank correlation
- Correlation and causation

Non-linear relationships

If $\rho \sim 0$ is tempting to say that there is no relationship between the variables, but that is not valid, because the Pearson's correlation only measures **linear relationships**.

What if the relationship is nonlinear?

Figure 7.4: Examples of datasets with a range of correlations.

Moral: do always a scatter plot of the data, before doing the correlation blindly. If the relationship isn't linear, Pearson's correlation will tend to underestimate the strength of the relationship.

Overview

- Bivariate data
- Correlation
- Covariance
- Pearson correlation
- Non-linear relationships
- **Spearman’s rank correlation**
- Correlation and causation

Spearman’s rank correlation

Spearman’s rank correlation is more robust if :

1. the variables are not roughly normal;

2. the relationship isn’t linear;

3. there is presence of outliers.

To compute it we need to compute the rank of each value, which is its index in the sorted sample. For example, in the sample [1, 2, 5, 7] the rank of the value 5 is 3.

Afterwards we compute the Pearson’s correlation of the ranks.

Function in python: “scipy.stats.spearmanr(a,b=None,axis=0)”

https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.spearmanr.html

Overview

- Bivariate data
- Correlation
- Covariance
- Pearson correlation
- Non-linear relationships
- Spearman's rank correlation
- **Correlation and causation**

Correlation and causation

If variables A and B are correlated there are 3 possible explanations:

1. A causes B

2. B causes A

3. Other set of factors causes both A and B

These explanations are “causal relationships” —> correlation does not imply causation, we can’t know which one of the 3 is.

How can we know causation?

1. Use time. What comes first A or B? However time does not preclude the possibility of 3.

2. Use randomness. If you divide a large sample into two groups at random and compute means of almost any variable, you expect the difference to be small. If the groups are nearly identical in all variables, but one, you can eliminate spurious relationships —> t-tests

Practice

1.Data for weight and height from ThinkStats and use standard functions in python scipy to compute the statistics for correlation and covariance between weight and height (Lecture5-VariablesRelationship.py reads data inside LectureA3 folder).

2.Make a scatter plot comparing them.

Lesson A3

Research Design

Overview

- Scientific method
- Measurement
- Basics of data collection
- Sampling bias
- Causation

Overview

- **Scientific method**
- Measurement
- Basics of data collection
- Sampling bias
- Causation

Scientific Method

There are a few descriptors that really set the scientific method:

- Data, if collected, is done systematically
- Theories in science can never be proved since one can never be 100% certain that a new empirical finding incosistent with the theory will never be found
- Scientific theories must be potentially disconfirmable
- If a hypothesis derived from theory is confirmed, then the theory has survived a test —> more useful for researchers
- The method of investigation in which a hypothesis is developed from a theory and then confirm or disconfirmed involves deductive reasoning

Overview

- Scientific method
- **Measurement**
- Basics of data collection
- Sampling bias
- Causation

Measurement

Measurements are generally required to be: reliable + valid

Reliability: revolves around whether you would get at least approximately the same result if you measure something twice with the same measurement instrument. It is the correlation between parallel forms of a test ($r_{test,test}$)

<u>True scores and error:</u>

Example: class of students taking a 100 point T/F exam. Every questions right 1 point, every question wrong -1. We assume you need to fill all questions.

A student that knew 90 questions correctly, and guessed right 7, will get a final score of: 90+(7-3)=94

$$y_{test} = y_{true} + y_{error} \quad \longleftarrow \quad \sigma^2_{test} = \sigma^2_{true} + \sigma^2_{error}$$

Measurement

Measurements are generally required to be: reliable + valid

Reliability: revolves around whether you would get at least approximately the same result if you measure something twice with the same measurement instrument. It is the correlation between parallel forms of a test ($r_{test,test}$)

<u>True scores and error:</u>

$$r_{test,test} = \frac{\sigma^2_{True}}{\sigma^2_{Test}} = \frac{\sigma^2_{True}}{\sigma^2_{True} + \sigma^2_{Error}}$$

NOTE! Reliability is not a property of a test, but a property of a test given a population, if the population properties varies, also the reliability

Measurement

Assessing Error of Measurement

$$s_{measurement} = s_{test}\sqrt{1 - r_{test,test}}$$

$s_{measurement}$ Standard error of measurement

s_{test} Standard deviation of the test scores

Increasing reliability

1) improve the quality of the items

2) to increase the number of the items

$$r_{new,new} = \frac{kr_{test,test}}{1 + (k-1)r_{test,test}}$$

k: is the factor by which the test length is increased

NOTE! Reliability won't be improved if the items added are poor quality!

Measurement - Validity

Measurements are generally required to be: reliable + valid

Validity: refers to whether the test measures what is supposed to measure. The most common types:

1) Face validity: whether the test appears to measure what it is supposed to measure.

2) Predictive validity: a test's ability to predict a relevant behaviour.

3) Construct validity: a test has construct validity if its pattern of correlations with other measures is in line with the construct it is purporting to measure.

NOTE! Reliability and predictive validity: the reliability of a test limits the size of the correlation between the test and other measures. Theoretically, it is possible for a test to correlate as high as the square root of the reliability with another measure.

Overview

- Scientific method
- Measurement
- **Basics of data collection**
- Sampling bias
- Causation

Basics of data collection

Most statistics analysis required that your data is in numerical form, therefore verbal or string data will need to be coded to be processed.

Table 1. Example Data

Student Name	Hair Color	Gender	Major	Height	Computer Experience
Norma	Brown	Female	Psychology	5’4”	Lots
Amber	Blonde	Female	Social Science	5’7”	Very little
Paul	Blonde	Male	History	6’1”	Moderate
Christopher	Black	Male	Biology	5’10”	Lots
Sonya	Brown	Female	Psychology	5’4”	Little

You should think very carefully about the scales and specify of information needed in your research before you collect it.

If you believe you might need additional information but aren’t sure, just collect it!

How to record a running track timing?

Overview

- Scientific method
- Measurement
- Basics of data collection
- **Sampling bias**
- Causation

Sampling Bias

Sampling bias refers to the method of sampling, not the sample itself.

Self-selection bias: surveys asking people to fill in, the people who "self-select" are likely to differ from he overall population.

Undercoverage bias: sample too few observations from a segment of the population.

Survivorship bias: occurs when the observations recorded at the end of the investigation are non-random set of those present at the beginning of the investigation.

Overview

- Scientific method
- Measurement
- Basics of data collection
- Sampling bias
- **Causation**

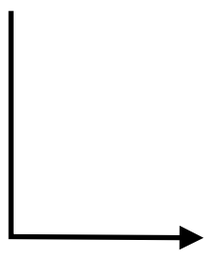
Causation

We focus in the establishment of causation in experiments.

Example: we have two groups one received an insomnia drug, the control received a placebo, and the dependent variable of interest is the hours of sleep.

An obvious obstacle to infer causality: many unmeasured variables that affect hours of sleep (stress, physiological, genetics…)

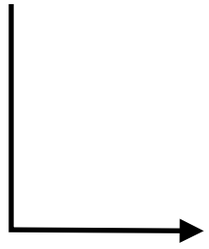
This problem seems intractable: how to measure an “unmeasured variable!!” —> but we can assess the combined effects of all unmeasured variables



Most common measure of difference: **variance**

Causation

By using the variance to assess the effects of unmeasured variables, statistical methods determine:



Probability that these unmeasured variables could produce a difference between conditions as large or larger than the experiment differences.

Probability is low —> it is inferred that the treatment had an effect!

Practice

1. If you wished to know how long since your subjects had a certain condition, it would be better to ask them_____

    a. What day the condition started and what day it is now
    b. How many days they have had the condition

2. In an experiment, if subjects are sampled randomly from a population and then assigned randomly to either the experimental group or the control group, we can be sure than the treatment caused the difference in the dependent variable.

    a. True
    b. False

Lesson A3

Probability

Overview

- Introduction
- Basics
- Permutations and combinations
- Binomial distribution
- Multinomial distribution
- Poisson distribution
- Bayes theorem

Overview

- **Introduction**
- Basics
- Permutations and combinations
- Binomial distribution
- Multinomial distribution
- Poisson distribution
- Bayes theorem

Introduction

Probability definition is not straight forward, there are many ways to define it…

- Based on **symmetrical outcomes** —> tossing a coin, what is the probability of each of the faces to lay out? What about a six-sided dice?

- Based on **relative frequencies** —> if we tossed a coin a million times, we would expect the proportion of tosses that came heads to be pretty close to 0.5.

An event with p=0 has no chance of occurring, an event with p=1 is certain to occur—>it is hard to think of any examples of interest to statistics where any of these two situations occurs.

Overview

- Introduction
- **Basics**
- Permutations and combinations
- Binomial distribution
- Multinomial distribution
- Poisson distribution
- Bayes theorem

Basics

Probability of a single event: $$p = \frac{fo}{pe - lo}$$

fo: favourable outcomes
pe: possible equally
lo: likely outcomes

Probability of two (or more) independent events:

1. Probability of A and B: $$P(A\,and\,B) = P(A) \cdot P(B)$$

2. Probability of A or B: $$P(A\,or\,B) = P(A) + P(B) - P(A) \cdot P(B)$$

Conditional probabilities: $$P(A\,and\,B) = P(A) \cdot P(B|A)$$

(dependent events)

Overview

- Introduction
- Basics
- **Permutations and combinations**
- Binomial distribution
- Multinomial distribution
- Poisson distribution
- Bayes theorem

Permutations and combinations

Possible orders: plate with 3 pieces of candy, green, yellow and red, how many different orders you can pick up the pieces?

$$Number\ of\ orders = 3!$$

Permutations: plate with 4 candies, but you only pick two pieces, how many ways are there of picking two pieces?

$$nPr = \frac{n!}{(n-r)!} = \frac{4!}{(4-2)!} = 12$$

Combinations: how many different combinations of 2 pieces could you end up with?

$$nCr = \frac{n!}{(n-r)!r!} = \frac{4!}{(4-2)!4!} = 6$$

Overview

- Introduction
- Basics
- Permutations and combinations
- **Binomial distribution**
- Multinomial distribution
- Poisson distribution
- Bayes theorem

Binomial distribution

Binomial distribution are probability distributions for which there are just two possible outcomes with fixed probabilities summing to one.

$$P(x) = \frac{N!}{x!(N-X)!}\pi^x(1-\pi)^{N-x}$$

N: trials for independent events

π: probability of success on a trial

Mean: $\mu = N \cdot \pi$

Variance: $\sigma^2 = N\pi \cdot (1-\pi)$

Overview

- Introduction
- Basics
- Permutations and combinations
- Binomial distribution
- **Multinomial distribution**
- Poisson distribution
- Bayes theorem

Multinomial distribution

It can be used to compute the probabilities in situations in which there are more than 2 possible outcomes

$$p = \frac{n!}{n_1!n_2!n_3!} p_1^{n_1} p_2^{n_2} p_3^{n_3}$$

n: total number of events
n_1: number of times outcome 1 occurs
n_2: number of times outcome 2 occurs
n_3: number of times outcome 3 occurs
p_1: probability of outcome 1
p_2: probability of outcome 2
p_3: probability of outcome 3

Overview

- Introduction
- Basics
- Permutations and combinations
- Binomial distribution
- Multinomial distribution
- **Poisson distribution**
- Bayes theorem

Poisson distribution

It can be used to calculate the probabilities of various numbers of successes based on the mean number of successes

The various events **must** be independent

$$p = \frac{e^{-\mu}\mu^{x}}{x!}$$

μ: mean number of successes
x: number of “successes” in question

Overview

- Introduction
- Basics
- Permutations and combinations
- Binomial distribution
- Multinomial distribution
- Poisson distribution
- **Bayes theorem**

Bayes Theorem

Bayes' theorem considers both the prior probability of an event and the diagnostic value of a test to determine the posterior probability of the event.

Imagine you have disease X (event X), only 2% of people in your situation have it:

- Prior probability: p(x)=0.02

The diagnostic value of the test depends on (p(t)):

- The probability that you test positive given that you have the the disease: p(t|x)

- The probability that you test positive given that you don't have it: p(t|x')

Using Bayes we can compute the probability of having the disease X having tested positive:

$$p(x|t) = \frac{p(t|x)p(x)}{p(t|x)p(x) + p(t|x')p(x')}$$

https://3d.bk.tudelft.nl/courses/geo1001/

Questions?