MATH 205: Statistical methods

Vu Dinh

Departments of Mathematical Sciences
University of Delaware

Lecture 2: Looking at relationship

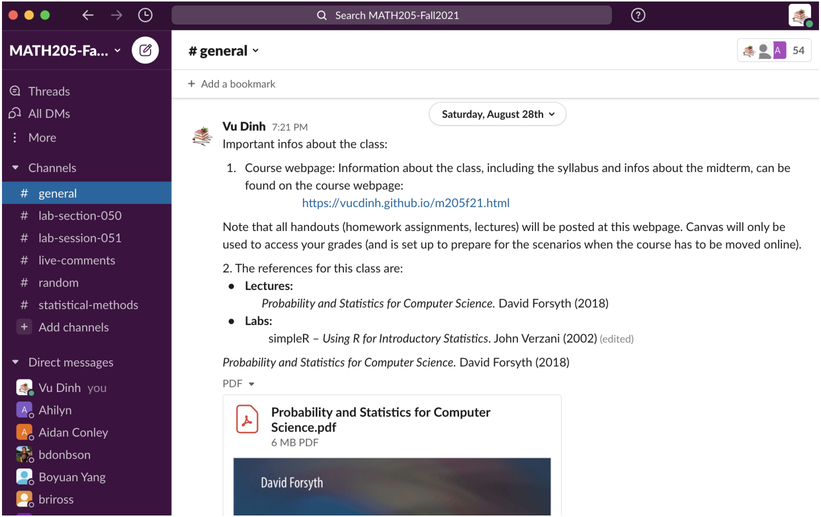
General information

- Lectures:

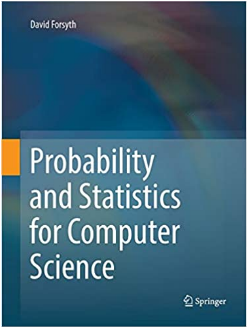
  MW 3:35pm-4:50pm, Kirkbride 205

- Labs:
  - Section 050L: M 2:30pm - 3:20pm, Ewing 101
  - Section 051L: W 2:30pm - 3:20pm, Ewing 101
- Office hours
  - TTh 2:00pm - 3:30pm, Ewing 312
  - or by appointments

Communications



Textbooks



Lectures:
Probability and Statistics for Computer Science.
David Forsyth (2018)

Labs:
simpleR – Using R for Introductory Statistics.
John Verzani (2002)

The safety of our learning environment

- Must wear a cloth mask that covers your nose and mouth
- Must not eat or drink in class
- Upon entering the classroom, wipe down your seat and desk area
- *Write down the names of the students sitting around you at the beginning of each class*

Other classroom settings

- The lectures will be recorded by UD Capture, accessible through Canvas.
  Note that there will be no camera in class, so work on the board wouldn't be seen in the records.
- The lab doesn't have UD Capture. I will provide a Zoom session for each lab.

Tentative schedule

Date	Theme/Topic	Labs	Assignments
Sep 1	Syllabus		
Sep 8	Chapter 1: Describing dataset	Section 2: Handling data	
Sep 13 - 15	Chapter 2: Looking at Relationships	Section 3: Univariate data	
Sep 20-22	Chapter 3: Basic Ideas in Probability	Section 4: Bivariate Data	Homework 1 (due 09/22)
Sep 27-29	Chapters 3-4	Section 4: Correlation	
Oct 4-6	Chapter 4: Random variables and expectations	Section 6: Random data	Homework 2 (due 10/06)
Oct 11-13	Chapter 5: Useful distributions	Section 7: The central limit theorem	
Oct 18-20	Chapter 6: Samples and populations	Section 9: Confidence interval estimation	Homework 3 (due 10/20)
Oct 25-27	Review and midterm exam		Midterm: Oct 27 (lecture), Oct 25-27 (labs)
Nov 1-3	Chapter 7: The significance of evidence	Section 10: Hypothesis testing	
Nov 8-10	Goodness of Fit	Section 12: Goodness of Fit	Homework 4 (due 11/10)
Nov 15-17	Linear Regression	Section 13: Linear regression	
Nov 22-24	Thanksgiving break		
Nov 29 - Dec 1	One-Way Analysis of Variance	Section 15: Analysis of variance	Homework 5 (due 12/01)
Dec 6-8	Selected topics + Review		
Exam week			

Chapter 1: Describing dataset

What to remember

- Categorical vs. continuous data
- Datasets as d-tuples

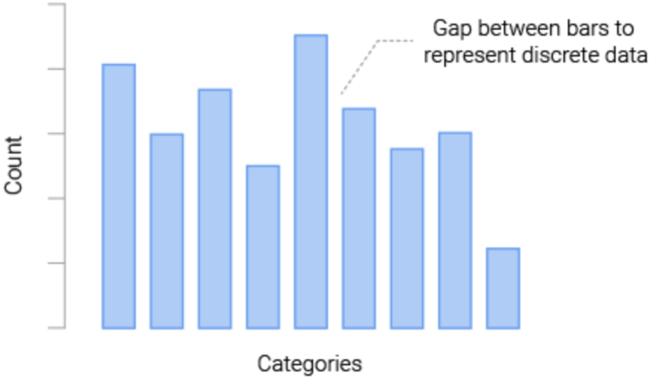
Summarizing univariate data:

- Mean
- Median
- Standard deviation
- Interquartile Range

Visualizing univariate data:

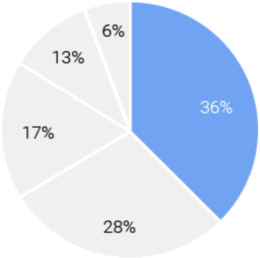
- Bar chart
- Pie chart
- Histogram
- Box plot

Bar charts


Gap between bars to
represent discrete data
Count
Categories

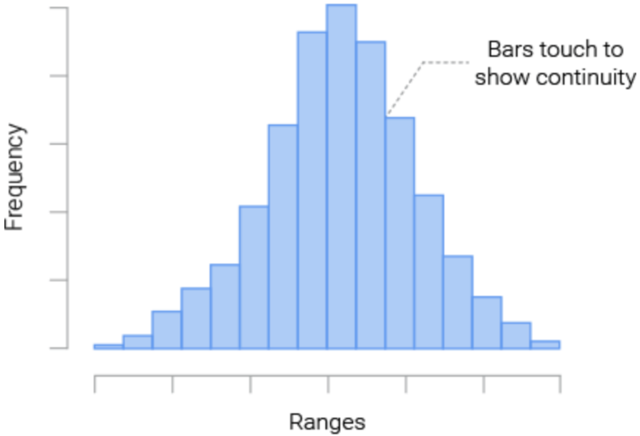
Pie charts

A Pie Chart is a special chart that shows relative sizes of data using **pie slices**.



They are good if you are trying to compare parts of a single data series to the whole.

Histograms



Summarizing univariate data

- Mean
- Median
- Standard deviation
- Variance
- Interquartile Range

Mean

Definition 1.1 (Mean) Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. Their mean is

$$\mathsf{mean}\left(\{x\}\right) = \frac{1}{N}\sum_{i=1}^{i=N} x_i.$$

Median

Definition 1.4 (Median) The median of a set of data points is obtained by sorting the data points, and finding the point halfway along the list. If the list is of even length, it's usual to average the two numbers on either side of the middle. We write

$$\text{median}\left(\{x\}\right)$$

for the operator that returns the median.

Median

The risk of developing iron deficiency is especially high during pregnancy. The problem with detecting such deficiency is that some methods for determining iron status can be affected by the state of pregnancy itself. Consider the following data on transferrin receptor concentration for a sample of women with laboratory evidence of overt iron-deficiency anemia ("Serum Transferrin Receptor for the Detection of Iron Deficiency in Pregnancy," *Amer. J. Clin. Nutrit.*, 1991: 1077–1081):

$$\begin{array}{llllll} x_1 = 15.2 & x_2 = 9.3 & x_3 = 7.6 & x_4 = 11.9 & x_5 = 10.4 & x_6 = 9.7 \\ x_7 = 20.4 & x_8 = 9.4 & x_9 = 11.5 & x_{10} = 16.2 & x_{11} = 9.4 & x_{12} = 8.3 \end{array}$$

The list of ordered values is

$$7.6 \quad 8.3 \quad 9.3 \quad 9.4 \quad 9.4 \quad 9.7 \quad 10.4 \quad 11.5 \quad 11.9 \quad 15.2 \quad 16.2 \quad 20.4$$

Since $n = 12$ is even, we average the $n/2 =$ sixth- and seventh-ordered values:

$$\text{sample median} = \frac{9.7 + 10.4}{2} = 10.05$$

Standard deviation

Definition 1.2 (Standard Deviation) Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. The standard deviation of this dataset is:

$$\begin{aligned}\mathsf{std}\,(\{x_i\}) &= \sqrt{\frac{1}{N}\sum_{i=1}^{i=N}(x_i - \mathsf{mean}\,(\{x\}))^2} \\ &= \sqrt{\mathsf{mean}\,(\{(x_i - \mathsf{mean}\,(\{x\}))^2\})}.\end{aligned}$$

Variance

Definition 1.3 (Variance) Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. where $N > 1$. Their variance is:

$$\mathsf{var}(\{x\}) = \frac{1}{N}\left(\sum_{i=1}^{i=N}(x_i - \mathsf{mean}(\{x\}))^2\right)$$

$$= \mathsf{mean}\left(\{(x_i - \mathsf{mean}(\{x\}))^2\}\right).$$

Interquartile range

Percentiles and Quartiles

Definition 1.5 (Percentile) The k'th percentile is the value such that $k\%$ of the data is less than or equal to that value. We write $\text{percentile}(\{x\}, k)$ for the k'th percentile of dataset $\{x\}$.

Definition 1.6 (Quartiles) The first quartile of the data is the value such that 25% of the data is less than or equal to that value (i.e. $\text{percentile}(\{x\}, 25)$). The second quartile of the data is the value such that 50% of the data is less than or equal to that value, which is usually the median (i.e. $\text{percentile}(\{x\}, 50)$). The third quartile of the data is the value such that 75% of the data is less than or equal to that value (i.e. $\text{percentile}(\{x\}, 75)$).

- If there are n data points, then the p quantile occurs at the position $1 + (n-1)p$ with weighted averaging if this is between integers.
- For example the .25 quantile of the numbers

$$10, 17, 18, 25, 28, 28$$

occurs at the position 1+(6-1)(.25) $= 2.25$. That is $1/4$ of the way between the second and third number which in this example is 17.25.

Interquartile range

Definition 1.7 (Interquartile Range) The interquartile range of a dataset $\{x\}$ is $\mathsf{iqr}\{x\} = \mathsf{percentile}(\{x\}, 75) - \mathsf{percentile}(\{x\}, 25)$.

Example

Consider the previous example:

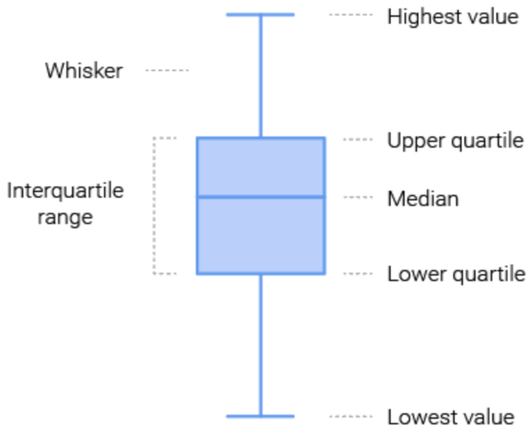
$$\begin{array}{llllll} x_1 = 15.2 & x_2 = 9.3 & x_3 = 7.6 & x_4 = 11.9 & x_5 = 10.4 & x_6 = 9.7 \\ x_7 = 20.4 & x_8 = 9.4 & x_9 = 11.5 & x_{10} = 16.2 & x_{11} = 9.4 & x_{12} = 8.3 \end{array}$$

The list of ordered values is

7.6 8.3 9.3 9.4 9.4 9.7 10.4 11.5 11.9 15.2 16.2 20.4

Compute the Interquartile range of this dataset.

Boxplot

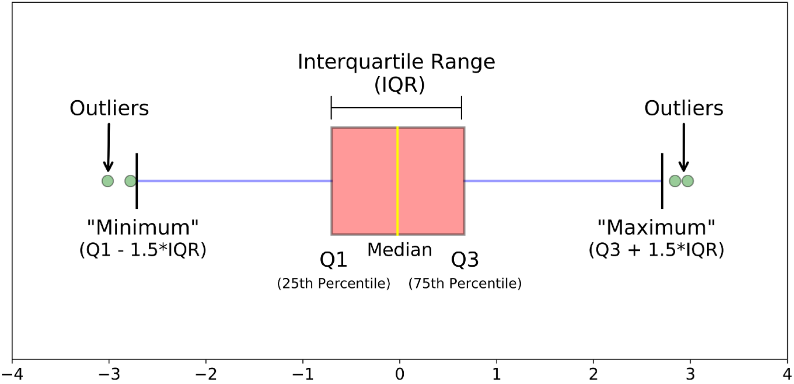


Highest value
Whisker
Upper quartile
Interquartile range
Median
Lower quartile
Lowest value

Boxplot with outliers

- Convention: any point further than 1.5*[Interquartile range] from the closest quartile is called *an outlier*
- Boxplot with outliers: The whisker is shorten to just include non-outliers. Outliers are plotted by points.

Boxplot with outliers



Final note: Standard coordinates

- It is often possible to get some useful insights about one univariate dataset from visualizations
- However, they are hard to compare because each is in a different set of units

Definition 1.8 (Standard Coordinates) Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. We represent these data items in standard coordinates by computing

$$\hat{x}_i = \frac{(x_i - \mathsf{mean}(\{x\}))}{\mathsf{std}(\{x\})}.$$

We write $\{\hat{x}\}$ for a dataset that happens to be in standard coordinates.

Problem

Definition 1.8 (Standard Coordinates) Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. We represent these data items in standard coordinates by computing

$$\hat{x}_i = \frac{(x_i - \text{mean}(\{x\}))}{\text{std}(\{x\})}.$$

We write $\{\hat{x}\}$ for a dataset that happens to be in standard coordinates.

Prove that:

- $mean(\{\hat{x}\}) = 0$
- $std(\{\hat{x}\}) = 1$

Standard coordinates

- We could then normalize the data by subtracting the location (mean) and dividing by the standard deviation (scale)
- The resulting values are unitless, and have zero mean

Chapter 2: Looking at relationship

Plotting 2D data

- We take a dataset, choose two different entries, and extract the corresponding elements from each tuple
- The result is a dataset consisting of 2-tuples, and we think of this as a two dimensional dataset
- Goal: to plot this dataset in a way that reveals relationships

Plotting 2D data

- Categorial data: The Wild West. No strict rules. Depends on the data and the user's cleverness
- Numerical data: scatter plot

Categorial vs categorical

- Common approach: create a richer set of categories
- Example: Relationship between
  - automobile class (2seater, compact, midsize, minivan, pickup, subcompact, suv)
  - drive type (front-wheel, rear-wheel, or 4-wheel drive)

Grouped bar charts

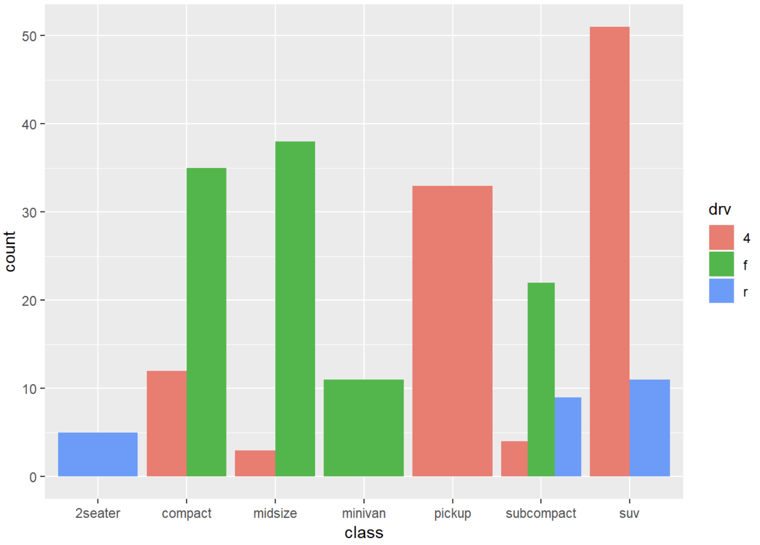
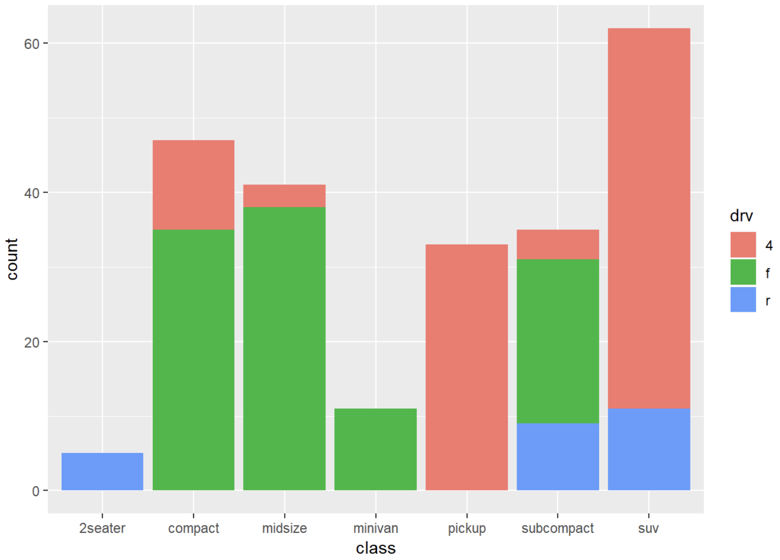
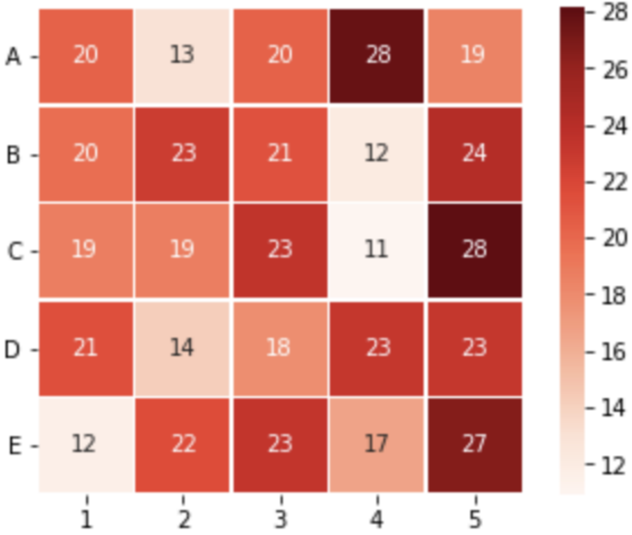


Figure 4.2: Side-by-side bar chart

Stacked bar charts

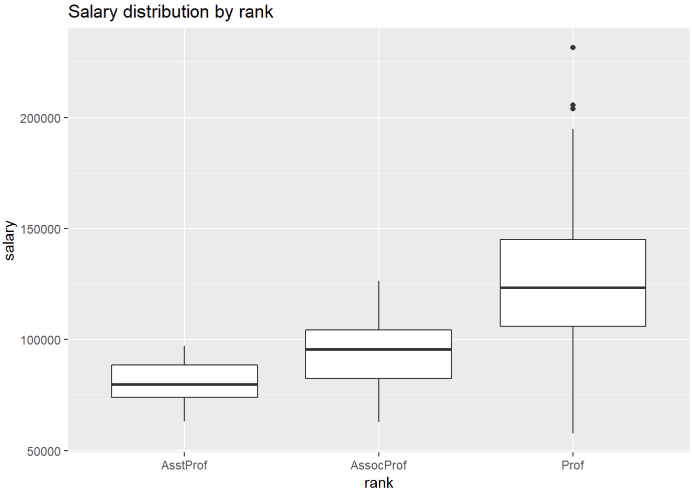


Heatmap



A
B
C
D
E
20 13 20 28 19
20 23 21 12 24
19 19 23 11 28
21 14 18 23 23
12 22 23 17 27
1 2 3 4 5
28 26 24 22 20 18 16 14 12

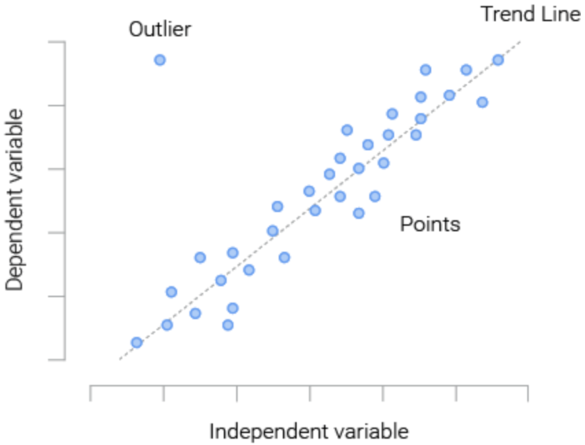
Categorial vs numerical



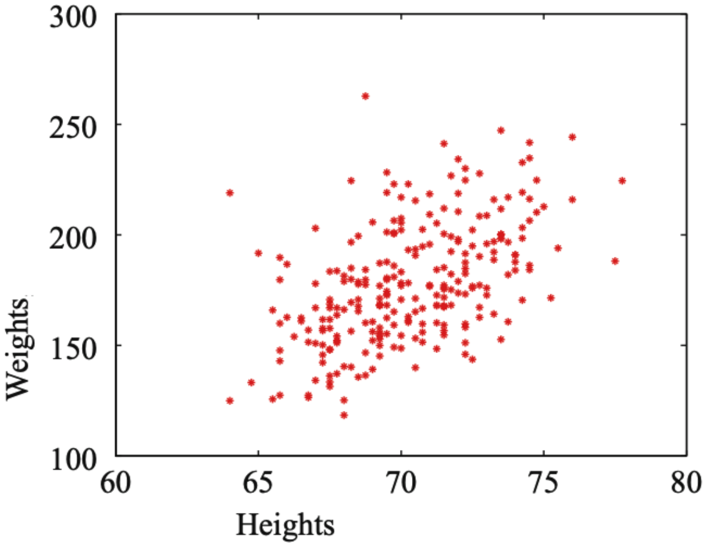
Numerical vs numerical: scatterplot

- Use Cartesian coordinates to display values for two variables for a set of data
- The data are displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis

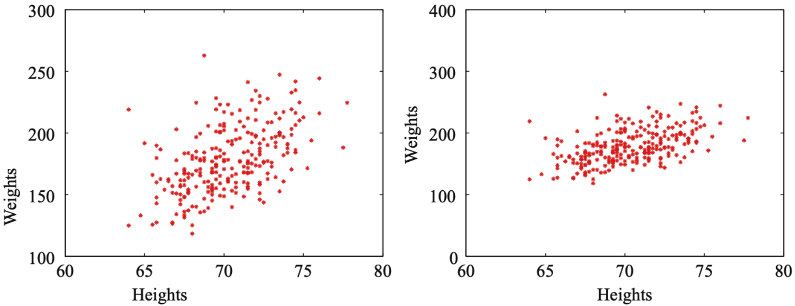
Scatterplot



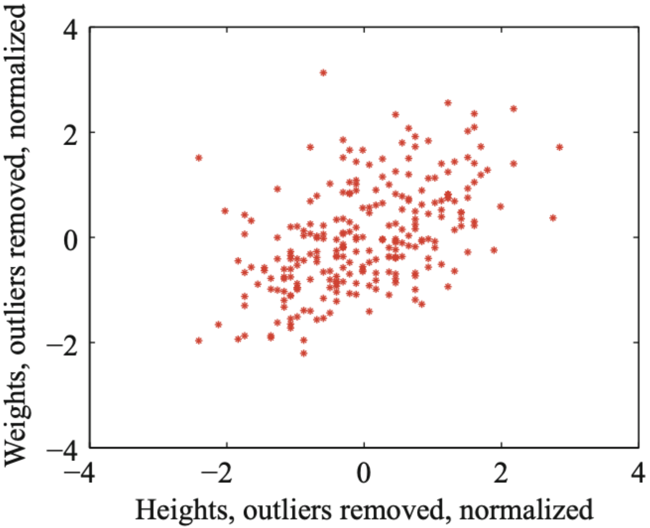
Scatterplot: example



Scale matters



Standard coordinates



Correlations

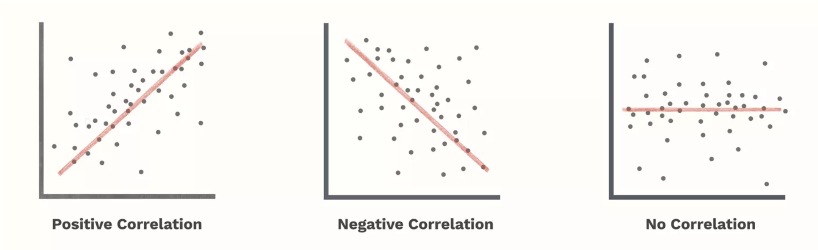
- From the figure: someone who is taller than the mean will tend to be heavier than the mean too
- This relationship is not always true for specific cases (and can not be represented by a function): some people are quite a lot taller than the mean, and quite a lot lighter

Correlations

Question: when $\hat{x}$ increases, does $\hat{y}$ tend to increase, decrease, or stay the same?

- Positive correlation: larger $\hat{x}$ values tend to appear with larger $\hat{y}$ values
- Negative correlation: larger $\hat{x}$ values tend to appear with smaller $\hat{y}$ values
- Zero correlation: no relationship

Correlations



Correlation coefficient

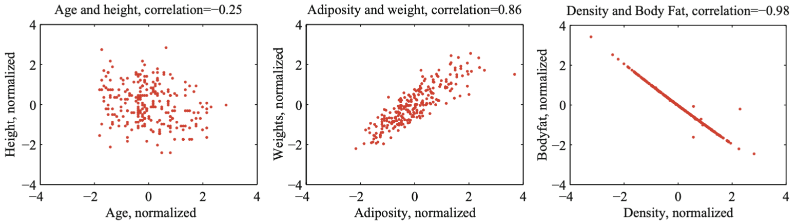
Definition 2.1 (Correlation Coefficient) Assume we have N data items which are 2-vectors $(x_1, y_1), \ldots, (x_N, y_N)$, where $N > 1$. These could be obtained, for example, by extracting components from larger vectors. We compute the correlation coefficient by first normalizing the x and y coordinates to obtain $\hat{x}_i = \frac{(x_i - \text{mean}(\{x\}))}{\text{std}(x)}$, $\hat{y}_i = \frac{(y_i - \text{mean}(\{y\}))}{\text{std}(y)}$. The correlation coefficient is the mean value of $\hat{x}\hat{y}$, and can be computed as:

$$\text{corr}(\{(x, y)\}) = \frac{\sum_i \hat{x}_i \hat{y}_i}{N}$$

Correlation coefficient

- correlation is a measure of our ability to predict one value from another
- correlation coefficient takes values between -1 and 1
- If the correlation coefficient is close to 1 or -1, then we are likely to predict very well.

Correlation coefficient



Correlation coefficient: property

Property 2.1 The largest possible value of the correlation is 1, and this occurs when $\hat{x}_i = \hat{y}_i$ for all i. The smallest possible value of the correlation is -1, and this occurs when $\hat{x}_i = -\hat{y}_i$ for all i.

Proposition

$$-1 \leq \mathsf{corr}\left(\{(x, y)\}\right) \leq 1$$