MATH 205: Statistical methods

September 13th, 2021

Lecture 3: Correlation

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

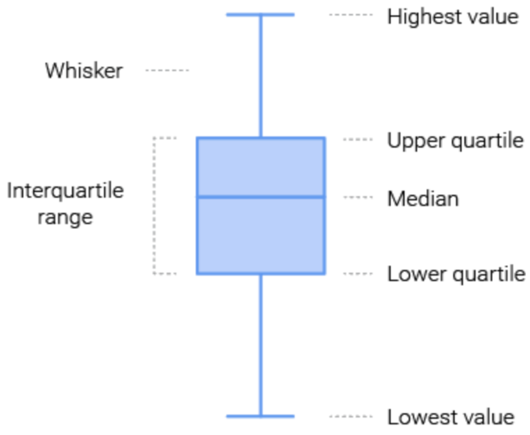
Summarizing univariate data:

- Mean
- Median
- Standard deviation
- Interquartile Range

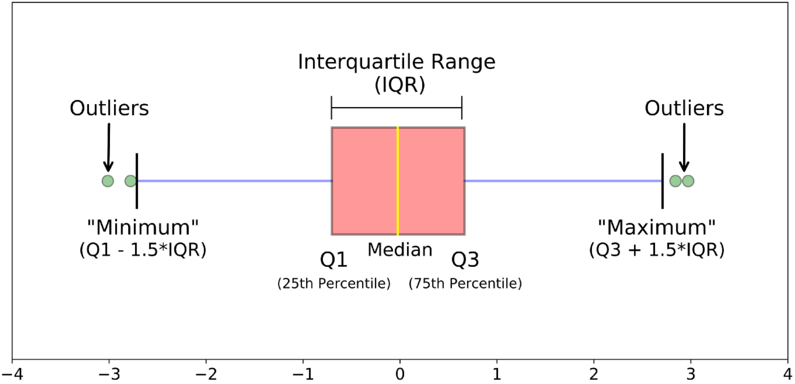
Visualizing univariate data:

- Bar chart
- Pie chart
- Histogram
- Box plot

Boxplot


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

Boxplot with outliers



Standard coordinates

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

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

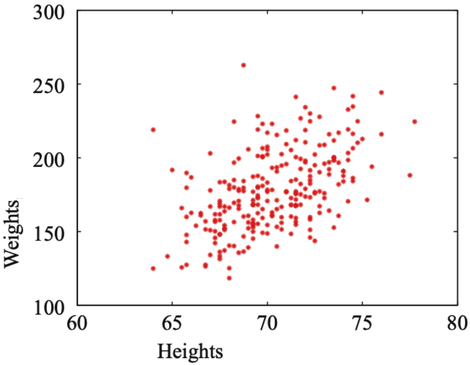
We write $\{\hat{x}\}$ for a dataset that happens to be in standard coordinates.

Chapter 2: Looking at relationship

Plotting 2D data

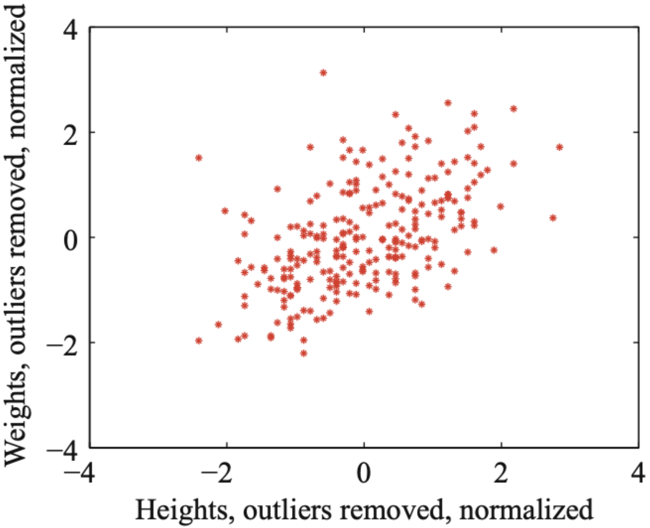
- categorical vs categorical: create a richer set of categories
- categorical vs continuous: comparative box plots
- continuous vs continuous: scatter plots

Scatter plot



Data displayed as a collection of points: the value of one variable determining the position on the x-axis and the value of the other variable determining the position on y-axis

Standard coordinates

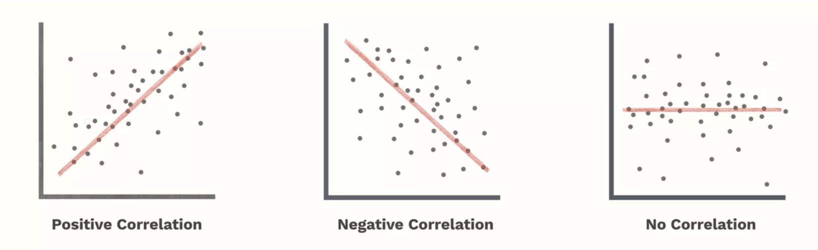


Correlations

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: properties

Useful Facts 2.1 (Properties of the Correlation Coefficient)

- The correlation coefficient is symmetric (it doesn't depend on the order of its arguments), so

$$\mathsf{corr}\left(\{(x,y)\}\right) = \mathsf{corr}\left(\{(y,x)\}\right)$$

- The value of the correlation coefficient is not changed by translating the data. Scaling the data can change the sign, but not the absolute value. For constants $a \neq 0$, b, $c \neq 0$, d we have

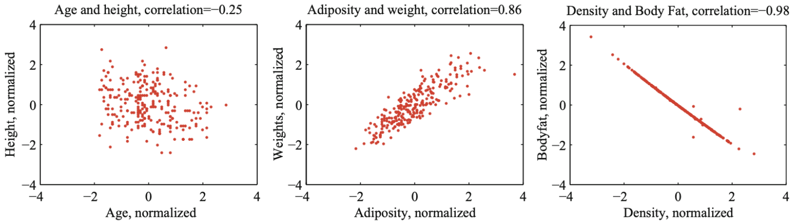
$$\mathsf{corr}\left(\{(ax+b, cx+d)\}\right) = \mathrm{sign}(ab)\mathsf{corr}\left(\{(x,y)\}\right)$$

- If $\hat{y}$ tends to be large (resp. small) for large (resp. small) values of $\hat{x}$, then the correlation coefficient will be positive.
- If $\hat{y}$ tends to be small (resp. large) for large (resp. small) values of $\hat{x}$, then the correlation coefficient will be negative.
- If $\hat{y}$ doesn't depend on $\hat{x}$, then the correlation coefficient is zero (or close to zero).
- The largest possible value is 1, which happens when $\hat{x} = \hat{y}$.
- The smallest possible value is -1, which happens when $\hat{x} = -\hat{y}$.

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

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

Using correlation to predict

Settings

- Assume that we have a two-dimensional datasets of N points:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$$

  $\rightarrow$ we can compute the correlation coefficient r
- Assume that r is close to 1, so we are confidence that we can predict y from x
- Assume that we have a new data point $(x_0, ?)$
- Question: How do we predict this unknown value '?'

Step 1: Transform data to standard coordinates

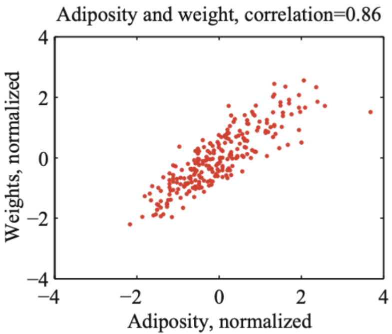
$$\hat{x}_i = \frac{1}{\text{std}(x)}(x_i - \text{mean}(\{x\}))$$

$$\hat{y}_i = \frac{1}{\text{std}(y)}(y_i - \text{mean}(\{y\}))$$

$$\hat{x}_0 = \frac{1}{\text{std}(x)}(x_0 - \text{mean}(\{x\})).$$

Idea: If we can predict the corresponding value $\hat{y}_0$, then we can transform back to the original coordinates and make prediction

Step 2: Construct $\hat{y}_0$



Idea: Maybe we could use a linear function to predict $\hat{y}$ from $\hat{x}$?

$$\hat{y}_i^p = a\hat{x}_i + b$$

and find a, b such that $\hat{y}_i - \hat{y}_i^p \approx 0$

Step 2: Construct $\hat{y}_0$

- Denote

$$u_i = \hat{y}_i - \hat{y}_i^p$$

- We want $u_i \approx 0$
- One way to do that is find a, b to ensure that

$$\text{mean}(\{u\}) = 0, \qquad \text{and} \qquad \text{std}(\{u\}) \text{ as small as possible}$$

Step 2a

$$\begin{aligned}
\text{mean}(\{u\}) &= \text{mean}(\{\hat{y} - \hat{y}^p\}) \\
&= \text{mean}(\{\hat{y}\}) - \text{mean}(\{a\hat{x}_i + b\}) \\
&= \text{mean}(\{\hat{y}\}) - a\text{mean}(\{\hat{x}\}) + b \\
&= 0 - a0 + b \\
&= 0.
\end{aligned}$$

We deduce that b should be 0.

Step 2b

$$\begin{aligned}
\text{var}(\{u\}) &= \text{var}(\{\hat{y} - \hat{y}^p\}) \\
&= \text{mean}\left(\left\{(\hat{y} - a\hat{x})^2\right\}\right) \quad \text{because } \text{mean}(\{u\}) = 0 \\
&= \text{mean}\left(\left\{(\hat{y})^2 - 2a\hat{x}\hat{y} + a^2(\hat{x})^2\right\}\right) \\
&= \text{mean}\left(\left\{(\hat{y})^2\right\}\right) - 2a\text{mean}(\{\hat{x}\hat{y}\}) + a^2\text{mean}\left(\left\{(\hat{x})^2\right\}\right) \\
&= 1 - 2ar + a^2,
\end{aligned}$$

For a fixed value of r, the optimal value for a is $a = r$ and the corresponding value for $var(\{u\})$ is $1 - r^2$.

Using correlation to predict

Procedure 2.1 (Predicting a Value Using Correlation) 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. Assume we have an x value x_0 for which we want to give the best prediction of a y value, based on this data. The following procedure will produce a prediction:

- Transform the data set into standard coordinates, to get

$$\hat{x}_i = \frac{1}{\mathsf{std}\,(x)}(x_i - \mathsf{mean}\,(\{x\}))$$

$$\hat{y}_i = \frac{1}{\mathsf{std}\,(y)}(y_i - \mathsf{mean}\,(\{y\}))$$

$$\hat{x}_0 = \frac{1}{\mathsf{std}\,(x)}(x_0 - \mathsf{mean}\,(\{x\})).$$

- Compute the correlation

$$r = \mathsf{corr}\,(\{(x, y)\}) = \mathsf{mean}\,(\{\hat{x}\hat{y}\}).$$

- Predict $\hat{y}_0 = r\hat{x}_0$.
- Transform this prediction into the original coordinate system, to get

$$y_0 = \mathsf{std}\,(y) r\hat{x}_0 + \mathsf{mean}\,(\{y\})$$

Note

By using the prediction procedure above, we have the error in prediction is

$$var(\{u\}) = 1 - r^2$$

Thus, the closer r^2 to 1, the better the prediction.

Confusion caused by correlation

Correlation does not imply causation

- When two variables are correlated, they change together. This means that one can make a reasonable prediction of one from the other.
- However, correlation does not mean that changing one variable causes the other to change (sometimes known as causation).

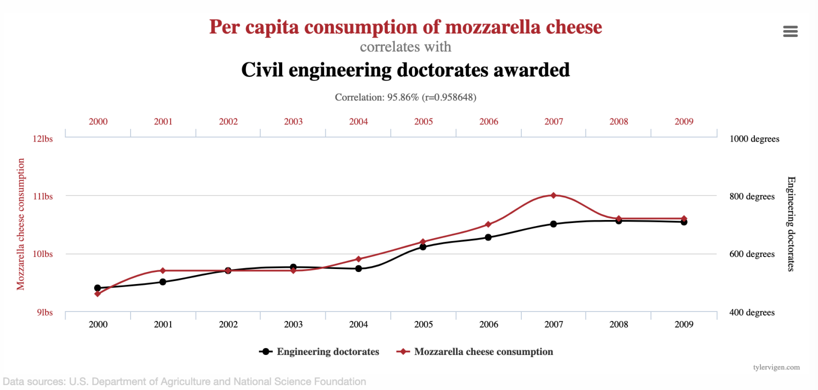
Variables could be correlated for a variety of reasons

Background (latent) variable:

- In children, shoe size is correlated with reading skills
- This doesn't mean that making your feet grow will make you read faster, or that you can make your feet shrink by forgetting how to read
- Latent variable: age. Young children tend to have small feet, and tend to have weaker reading skills

Variables could be correlated for a variety of reasons

Random chances:



Practice problem

Problem 1

Problem

In a population,the correlation coefficient between weight and adiposity is 0.9. The mean weight is 150 lb. The standard deviation in weight is 30 lb. Adiposity is measured on a scale such that the mean is 0.8, and the standard deviation is 0.1.

(a) *Using this information, predict the expected adiposity of a subject whose weight is 170 lb*

(b) *Using this information, predict the expected weight of a subject whose adiposity is 0.75*

Problem 2

Recall the definition of correlation:

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:

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

Prove that: for any numbers b, d

$$corr(\{(x + b, x + d)\}) = corr(\{(x, y)\})$$

Problem 2b

Recall the definition of correlation:

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:

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

Prove that: for any numbers a, b, c, d

$$corr(\{(ax + b, cx + d)\}) = sign(ab)corr(\{(x, y)\})$$