FIT2086 Lecture 1
Modelling for Data Analysis

Daniel F. Schmidt, with material from Geoff I. Webb

Faculty of Information Technology, Monash University

July 18, 2018

Outline

Outline

What is a "model"?

1. What is a model?
   - A mathematical description of some phenomena
2. What can we use a model for?
   - We can use it to make statements about reality
3. Where do models come from?
   - They are often learned from empirical (observational) data
4. Why is modelling important?

What is a "model"?

1. What is a model?
   - A mathematical description of some phenomena
2. What can we use a model for?
   - We can use it to make statements about reality
3. Where do models come from?
   - They are often learned from empirical (observational) data
4. Why is modelling important?

What is a "model"?

1. What is a model?
   - A mathematical description of some phenomena
2. What can we use a model for?
   - We can use it to make statements about reality
3. Where do models come from?
   - They are often learned from empirical (observational) data
4. Why is modelling important?

What is a "model"?

1. What is a model?
   - A mathematical description of some phenomena
2. What can we use a model for?
   - We can use it to make statements about reality
3. Where do models come from?
   - They are often learned from empirical (observational) data
4. Why is modelling important?

Data science is big business

Rank	Company	Capitalisation (US$ million)
1	Apple Inc	$749,124$
2	Alphabet	$628,610$
3	Microsoft	$528,778$
4	Amazon.com	$466,471$
5	Berkshire Hathaway	$418,880$
6	Johnson & Johnson	$357,310$
7	Facebook	$357,176$
8	Tencent	$344,879$
9	Exxon Mobil	$341,947$
10	JPMorgan Chase	$323,838$

Public Companies by Capitalisation (c. mid-2017)

Data science is big business

Rank	Company	Capitalisation (US$ million)
1	Apple Inc	$749,124$
2	Alphabet	$628,610$
3	Microsoft	$528,778$
4	Amazon.com	$466,471$
5	Berkshire Hathaway	$418,880$
6	Johnson & Johnson	$357,310$
7	Facebook	$357,176$
8	Tencent	$344,879$
9	Exxon Mobil	$341,947$
10	JPMorgan Chase	$323,838$

Public Companies by Capitalisation (c. mid-2017)

Data Science is Fun

- Data science lets you take data (numbers, measurements) and learn about the process that generated the data
- It lets you make predictions about the future using the past
  - Will Manchester United beat Real Madrid in the Champions League?
- It lets you quantify empirical evidence of phenomena
  - Do dogs really bite more frequently on the full moon?

Administrative Details

- Classes
  - 2 hour lecture, Monday, 15:00 - 17:00
  - 2 hour studio, as per Allocate+
- Outside class
  - Reading, assignments and self-learning
  - Note: you will be expected to learn R programming
- Text: Ross, S.M. (2014) Introduction to Probability and Statistics for Engineers and Scientists, 5th ed. Academic Press.

Subject Schedule & Assessment

Week	Topics	Assessment
1	Introduction, Modelling, Descriptive Statistics	
2	Probability and Probability Distributions	
3	Sampling, Parameter Estimation and Bias	Ass. #1 Due (10%)
4	Confidence Intervals	
5	Hypothesis Testing	
6	Linear Regression	
7	Naïve Bayes and Logistic Regression	Ass. #2 Due (20%)
8	Model Selection and Penalized Regression	
9	Trees and Nearest Neighbour Methods	
10	Introduction to Unsupervised Learning	
11	Simulation Based Statistical Methods	
12	Revision	Ass. #3 Due (20%)

- There is also an examination worth 50%.

Staff

- Lecturer
  - Dr. Daniel Schmidt (Daniel.Schmidt@monash.edu)
  - Office: 126A, Level 1, 25 Exhibition Walk
  - Consultation: Monday 13:00 – 14:00
- Tutor
  - Mr. Lachlan O'Neill

Studios

- You must prepare beforehand
  - Studio material will be released before the studio is to be run
  - Based on material covered in the current week's lecture
  - Will be using R, but we will not be teaching R programming
- The basic idea behind the studios is:
  - to get some hands-on experience analysing data
  - to use computational techniques to understand concepts

What this unit is about

- Technical overview of Data Science
  - Exposure to variety of models/methods for data science
  - Some hands-on experience with data analysis
  - Gain an understanding of data and probabilistic models
- NOT learning in depth each model, method introduced
- NOT becoming an R expert
- Realistic goals for students:
  - Familiarization with basics of a few tools
  - Learning advantages/disadvantages of main techniques/models
  - Practice data analysis
  - Exposure to fundamental ideas behind data analytic tools

What this unit is about

- Technical overview of Data Science
  - Exposure to variety of models/methods for data science
  - Some hands-on experience with data analysis
  - Gain an understanding of data and probabilistic models
- NOT learning in depth each model, method introduced
- NOT becoming an R expert
- Realistic goals for students:
  - Familiarization with basics of a few tools
  - Learning advantages/disadvantages of main techniques/models
  - Practice data analysis
  - Exposure to fundamental ideas behind data analytic tools

What this unit is about

- Technical overview of Data Science
  - Exposure to variety of models/methods for data science
  - Some hands-on experience with data analysis
  - Gain an understanding of data and probabilistic models
- NOT learning in depth each model, method introduced
- NOT becoming an R expert
- Realistic goals for students:
  - Familiarization with basics of a few tools
  - Learning advantages/disadvantages of main techniques/models
  - Practice data analysis
  - Exposure to fundamental ideas behind data analytic tools

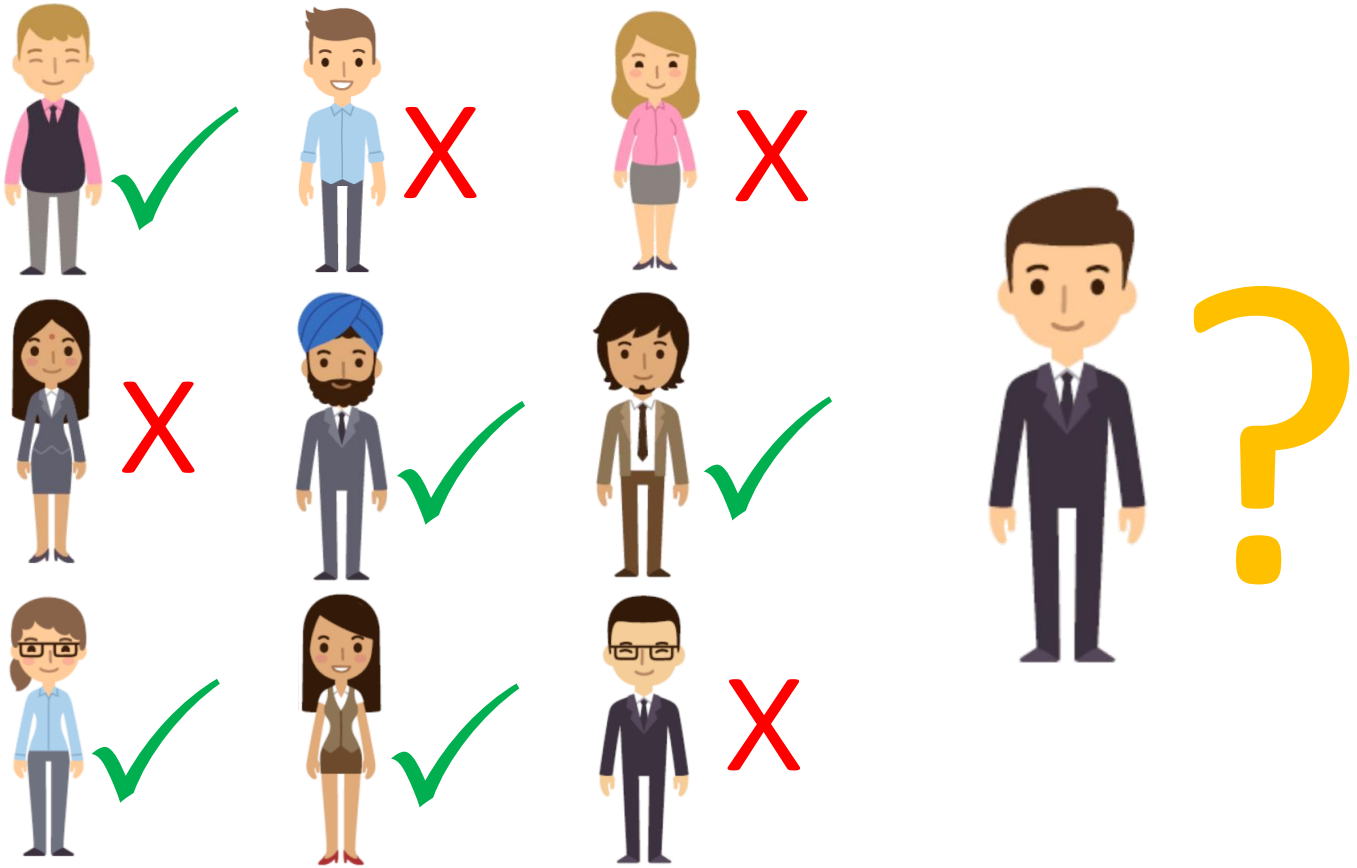
Marks and Hurdles

- Important information!
- To pass FIT2086 you must obtain:
  - 40% or more in the exam; and
  - 40% or more in the assignments; and
  - an overall unit mark of 50% or greater.
- If you get less than 40% for either exam or assignments, and the total mark is:
  - equal to or greater than 50%, a mark of 49-N will be recorded.
  - less than 50%, then the actual mark will be recorded.

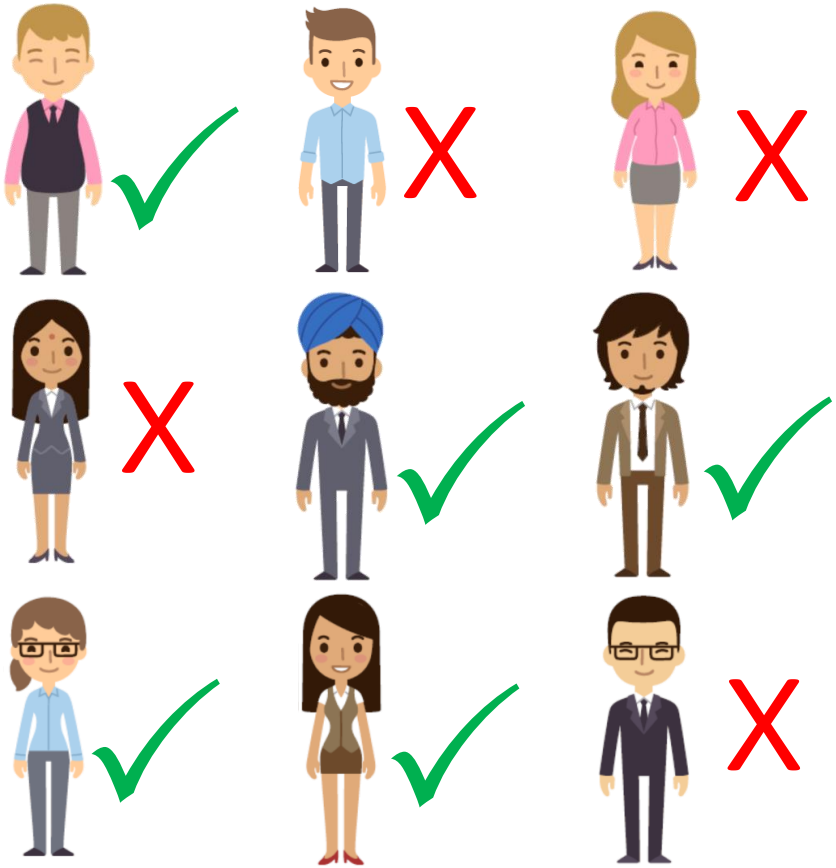
Models

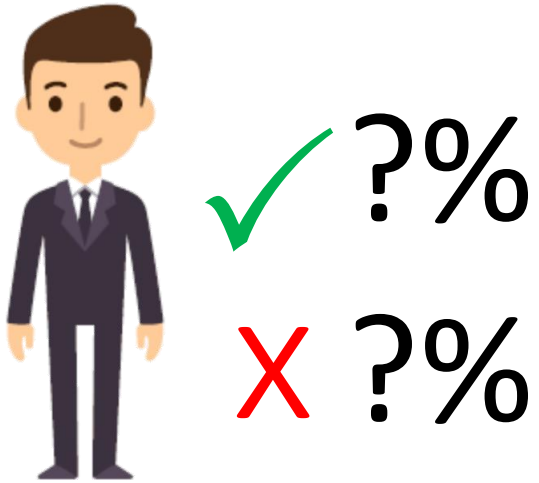
- A model is an object that represents something else
  - A model airplane, a model of a building
- Data science models are mathematical or algorithmic representations
- Models are neither correct, nor incorrect: but they can be more, or less useful for different purposes
  - One model aircraft might accurately represent the relative dimensions of the wings and body
  - An alternative model might more accurately capture the aerodynamic behaviour
- Let's take a quick tour of some models used in data science...

Classifiers



Probabilistic classifiers





Regression


100

76

84

88

92

82

45

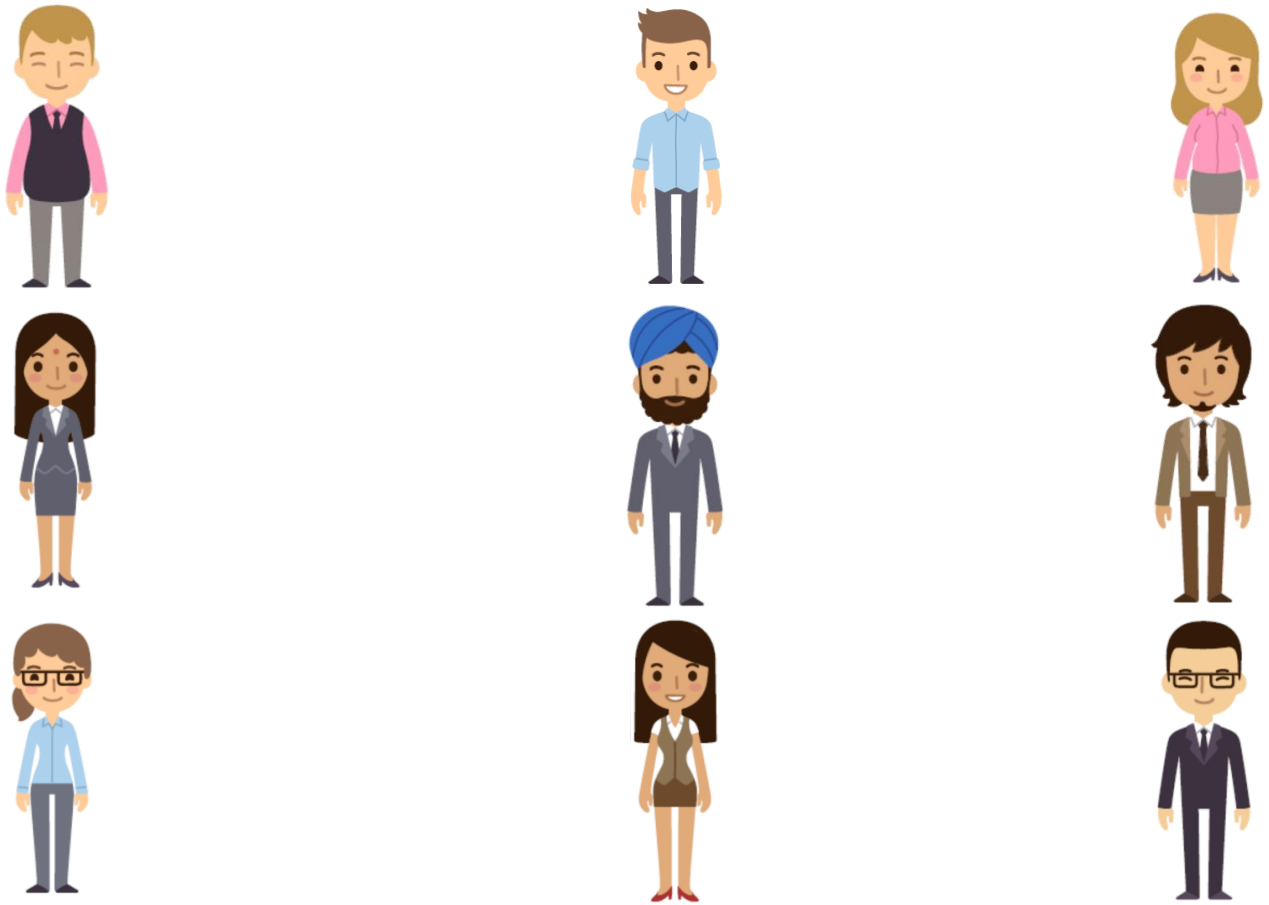
95

40

?

Forecasting

Clustering



Clustering

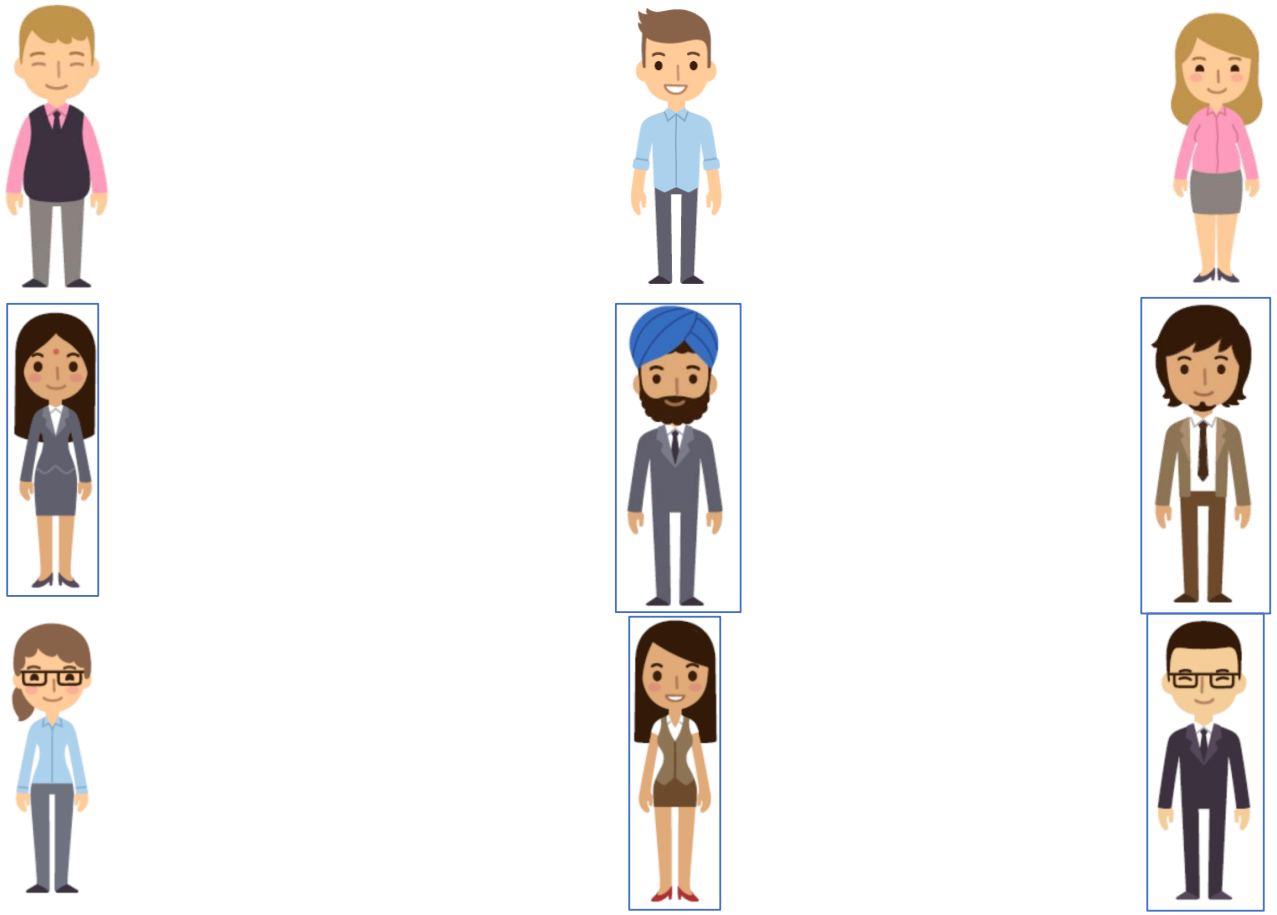








Clustering



Clustering

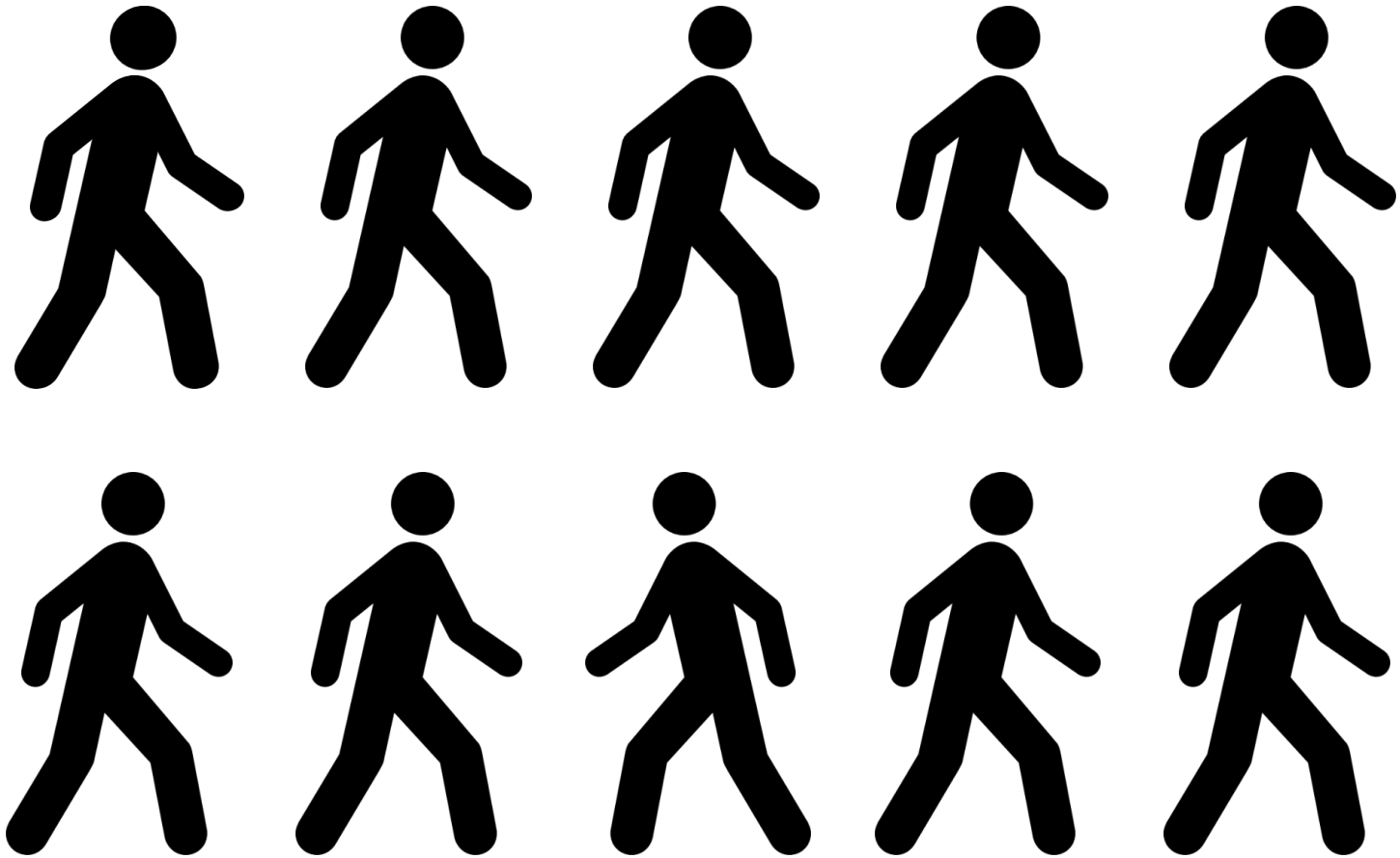




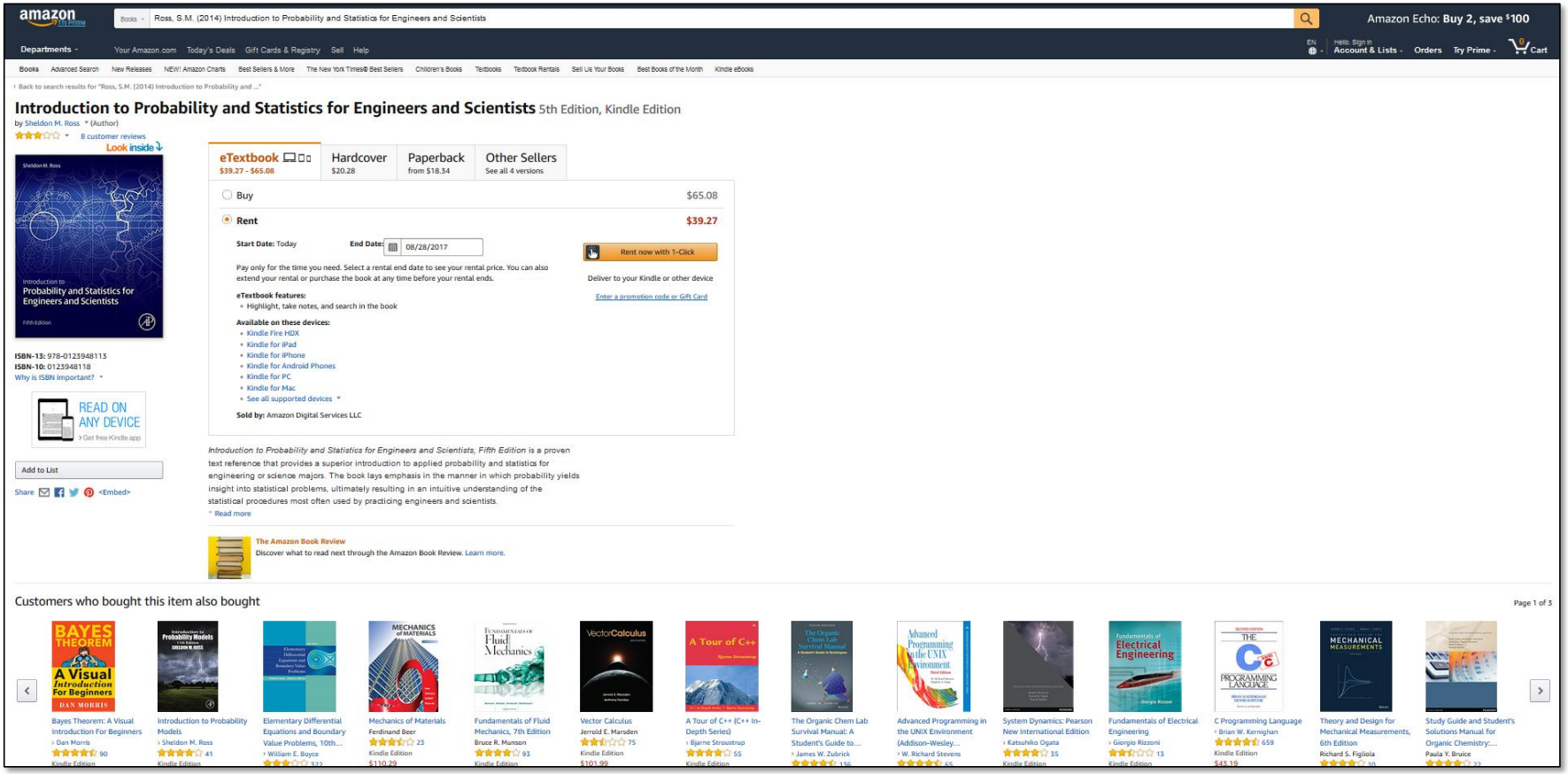




Anomaly Detection



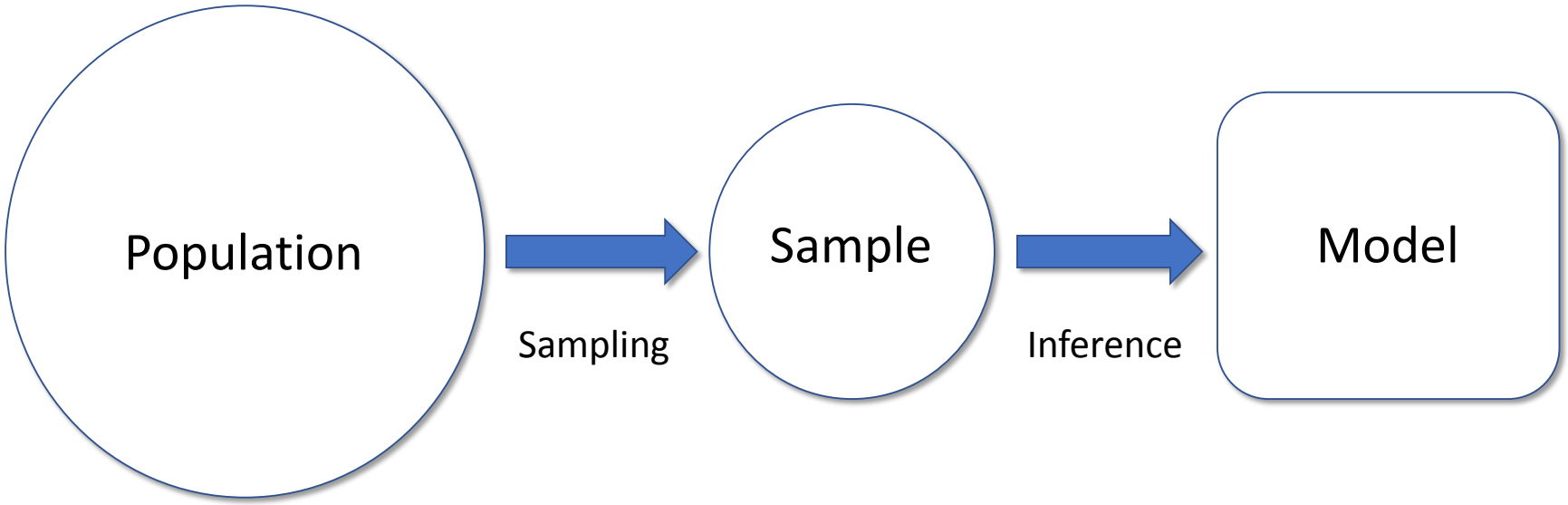
Recommendation Systems



Some Important Terms

- Population:
  - A large collection of objects/items with measurable attributes
- Sample:
  - A finite number of recordings of attributes of items from a population
- Model:
  - A mathematical or algorithmic description of the population learned/inferred from the sample

From Data to Models



Basic Types of Data

- Categorical-Nominal:
  - Discrete numbers of values, no inherent ordering
  - E.g., country of birth, sex
- Categorical-Ordinal:
  - Discrete number of states, but with an ordering
  - E.g., Education status, State of disease progression
- Numeric-Discrete:
  - Numeric, but the values are enumerable
  - e.g., Number of live births, Age (in whole years)
- Numeric-Continuous:
  - Numeric, not enumerable (i.e., real numbers)
  - E.g., Weight, Height, Distance from CBD
- Quantitative vs Qualitative:
  - Generally, categorical data is qualitative, numeric data is quantitative

Why do we Need Formal Methods for Data Science?

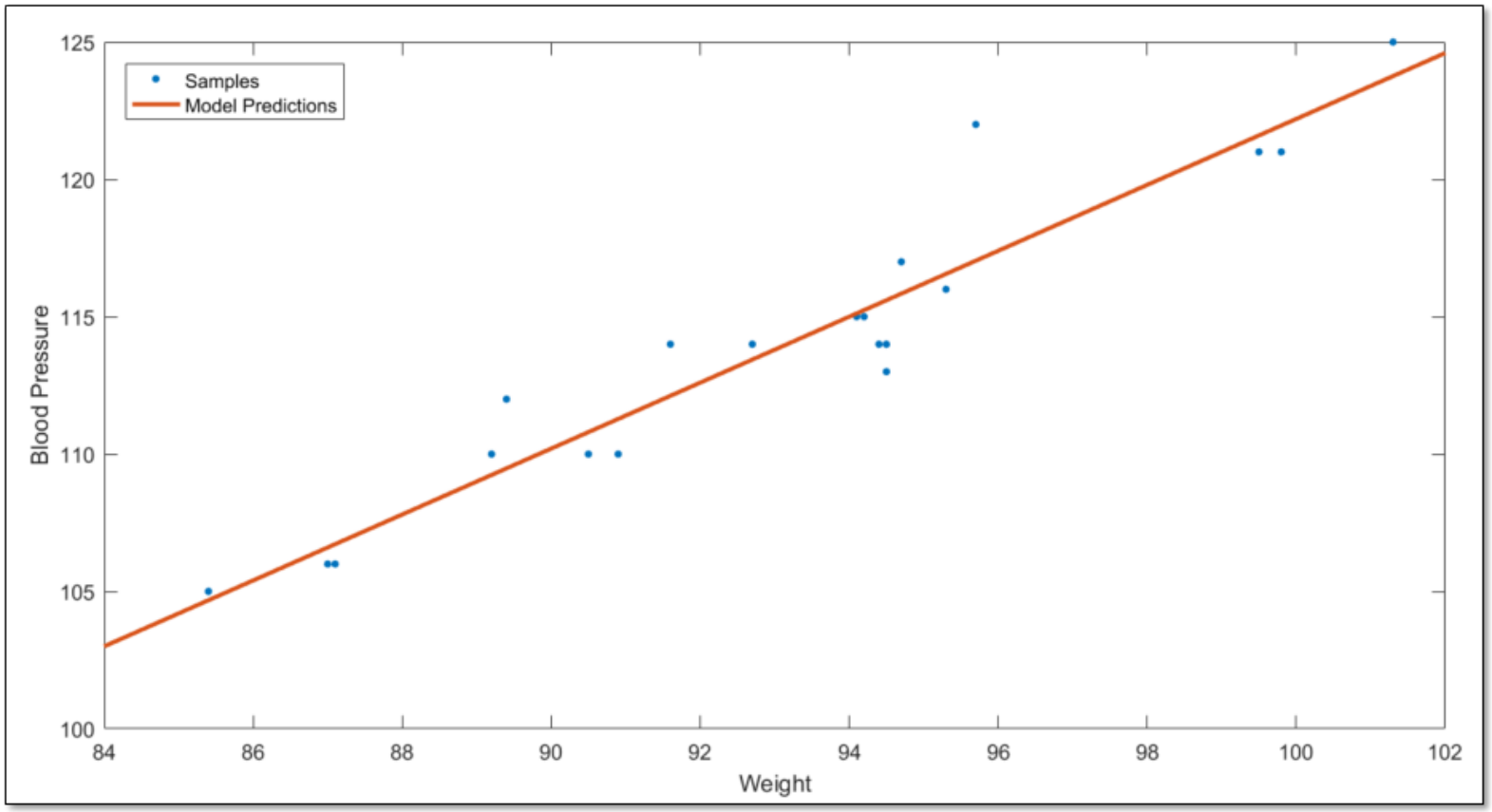
- Consider the following simple example

Pt	BP	Age	Weight	BSA	Dur	Pulse	Stress
1	105	47	85.4	1.75	5.1	63	33
2	115	49	94.2	2.10	3.8	70	14
3	116	49	95.3	1.98	8.2	72	10
4	117	50	94.7	2.01	5.8	73	99
5	112	51	89.4	1.89	7.0	72	95
6	121	48	99.5	2.25	9.3	71	10
7	121	49	99.8	2.25	2.5	69	42
8	110	47	90.9	1.90	6.2	66	8
9	110	49	89.2	1.83	7.1	69	62
10	114	48	92.7	2.07	5.6	64	35
11	114	47	94.4	2.07	5.3	74	90
12	115	49	94.1	1.98	5.6	71	21
13	114	50	91.6	2.05	10.2	68	47
14	106	45	87.1	1.92	5.6	67	80
15	125	52	101.3	2.19	10.0	76	98
16	114	46	94.5	1.98	7.4	69	95
17	106	46	87.0	1.87	3.6	62	18
18	113	46	94.5	1.90	4.3	70	12
19	110	48	90.5	1.88	9.0	71	99
20	122	56	95.7	2.09	7.0	75	99

- Knowing weight, can we build a model for blood pressure?

A Simple Model (1)

- We could "build" the following model



A Simple Model (2)

- More formally, our model is the equation:

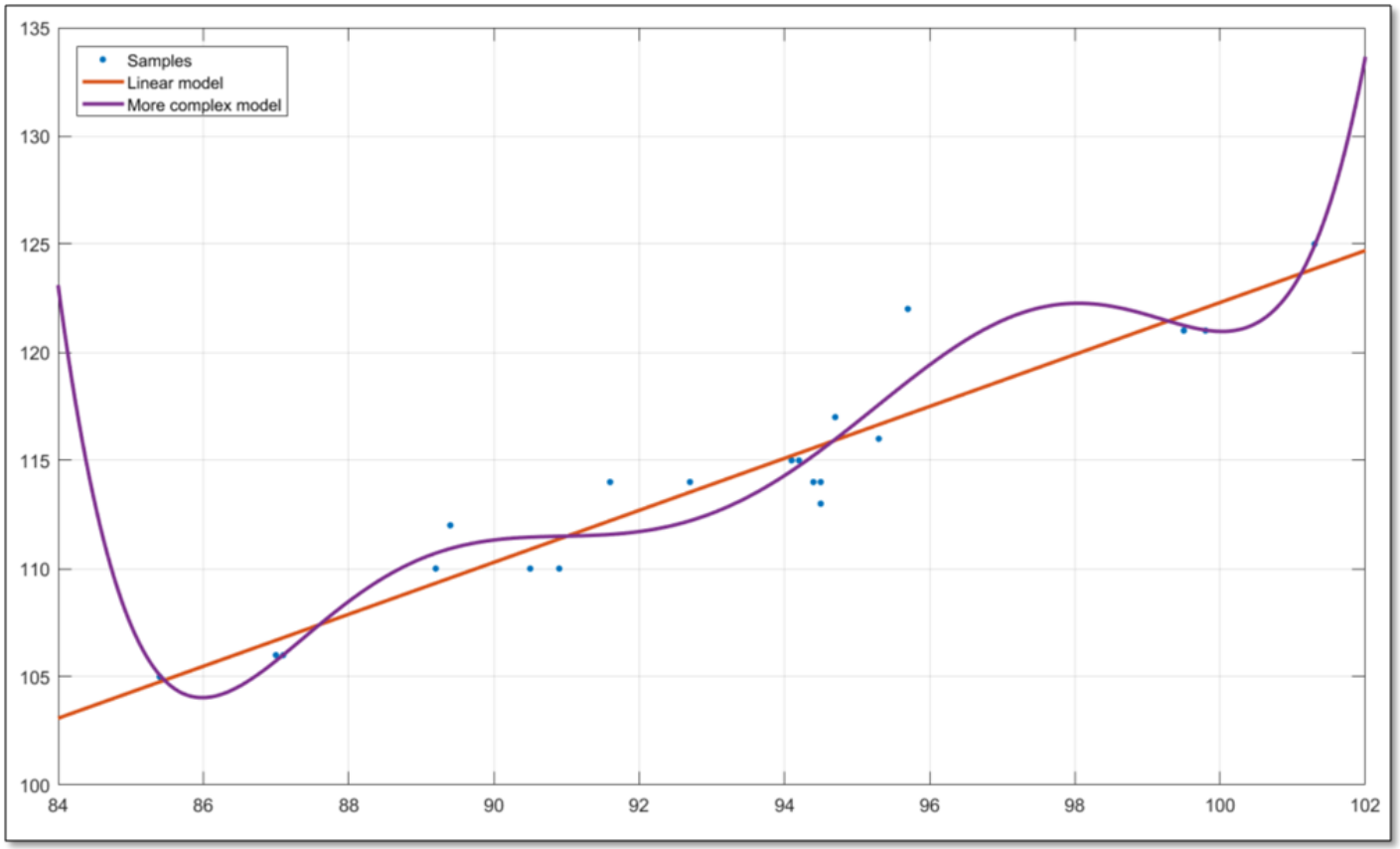
$$\text{bp} = 1.2 \times \text{weight} + 2.2 + \text{error}$$

- This model relates a person's blood pressure to their weight
  - The relationship is linear (a straight line)
  - The coefficients were learned directly from the data

- The "error" term accounts for the discrepancy between the model predictions and the measured data points
  - We handle this error by treating it as a random quantity

A Simple Model (3)

- We could build the more complex model:



- Fits the sample better – but is the improvement real?

Formal Data Science Methods

- Formal data science methods let us ...
  1. Find the coefficients of our straight line in an objective fashion
     - "Parameter estimation", learning a model
  2. Answer the question as to which of the two models we looked is the better description of the population
     - The more complex model fit the sample better, but is it warranted?
  3. Examine many variables simultaneously to find complex relationships
     - Not really possible "by hand"

Formal Data Science Methods

- Formal data science methods let us ...
  1. Find the coefficients of our straight line in an objective fashion
     - "Parameter estimation", learning a model
  2. Answer the question as to which of the two models we looked is the better description of the population
     - The more complex model fit the sample better, but is it warranted?
  3. Examine many variables simultaneously to find complex relationships
     - Not really possible "by hand"

Formal Data Science Methods

- Formal data science methods let us ...
  1. Find the coefficients of our straight line in an objective fashion
     - "Parameter estimation", learning a model
  2. Answer the question as to which of the two models we looked is the better description of the population
     - The more complex model fit the sample better, but is it warranted?
  3. Examine many variables simultaneously to find complex relationships
     - Not really possible "by hand"

Outline

Graphical representations – Refresher

- It is often useful to visualise data
  - Can sometimes quickly reveal patterns
  - However, going beyond two dimensions is problematic
- For categorical data, standard visualisations include:
  - Frequency tables
  - Bar graphs
  - Pie charts (not recommended!)
- For numeric data (continuous and discrete), we can use:
  - Histograms
  - Box-and-whisker plots

Graphical representations – Refresher

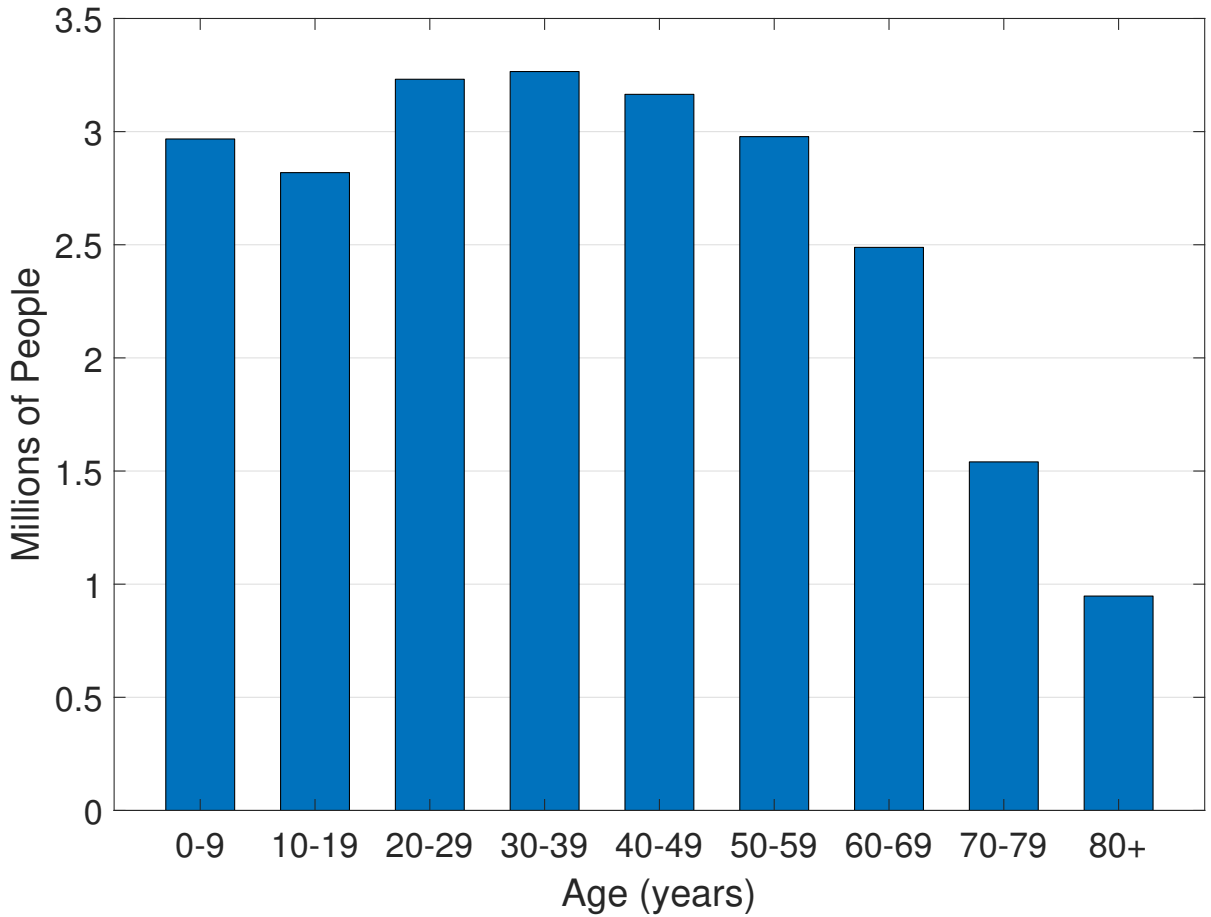
- It is often useful to visualise data
  - Can sometimes quickly reveal patterns
  - However, going beyond two dimensions is problematic
- For categorical data, standard visualisations include:
  - Frequency tables
  - Bar graphs
  - Pie charts (not recommended!)
- For numeric data (continuous and discrete), we can use:
  - Histograms
  - Box-and-whisker plots

Frequency Tables

Age (years)	Number of People
0-9	2,967,425
10-19	2,818,778
20-29	3,231,395
30-39	3,265,526
40-49	3,164,712
50-59	2,977,883
60-69	2,488,396
70-79	1,540,373
80+	947,411

Australian Population by Age (2016 Census)

Bar charts



Australian population by age (2016 Census)

Histograms

- Histograms are a special type of bar chart
  - Bar-charts only applicable to categorical data
- Group numeric data into categories by putting it bins
- If $\mathbf{y} = (y_1, \ldots, y_n)$ are our data points, we divide them between K equally spaced bins, i.e.,
  - The number of samples that fall in bin (category) k are

$$v_k = \#\{y_j \in (\min\{\mathbf{y}\} + (k-1)w, \ \min\{\mathbf{y}\} + kw)\}$$

where

$$w = \frac{\max\{\mathbf{y}\} - \min\{\mathbf{y}\}}{K}$$

is the width of the bins

$\Longrightarrow$ plot $v_1, \ldots, v_K$ using bar-chart

Histograms

- Histograms are a special type of bar chart
  - Bar-charts only applicable to categorical data
- Group numeric data into categories by putting it bins
- If $\mathbf{y} = (y_1, \ldots, y_n)$ are our data points, we divide them between K equally spaced bins, i.e.,
  - The number of samples that fall in bin (category) k are

$$v_k = \#\{y_j \in (\min\{\mathbf{y}\} + (k-1)w, \ \ \min\{\mathbf{y}\} + kw)\}$$

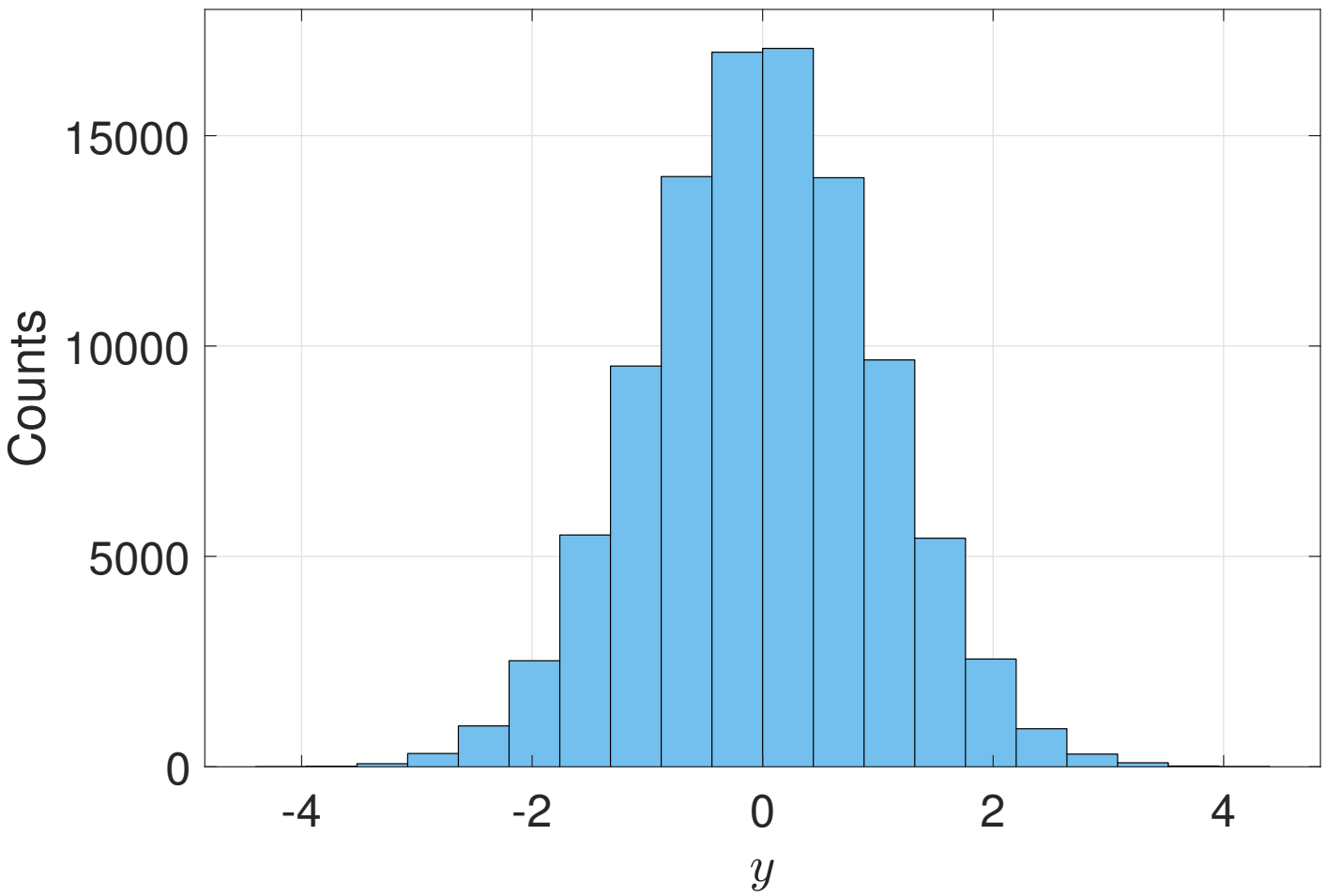
where

$$w = \frac{\max\{\mathbf{y}\} - \min\{\mathbf{y}\}}{K}$$

is the width of the bins

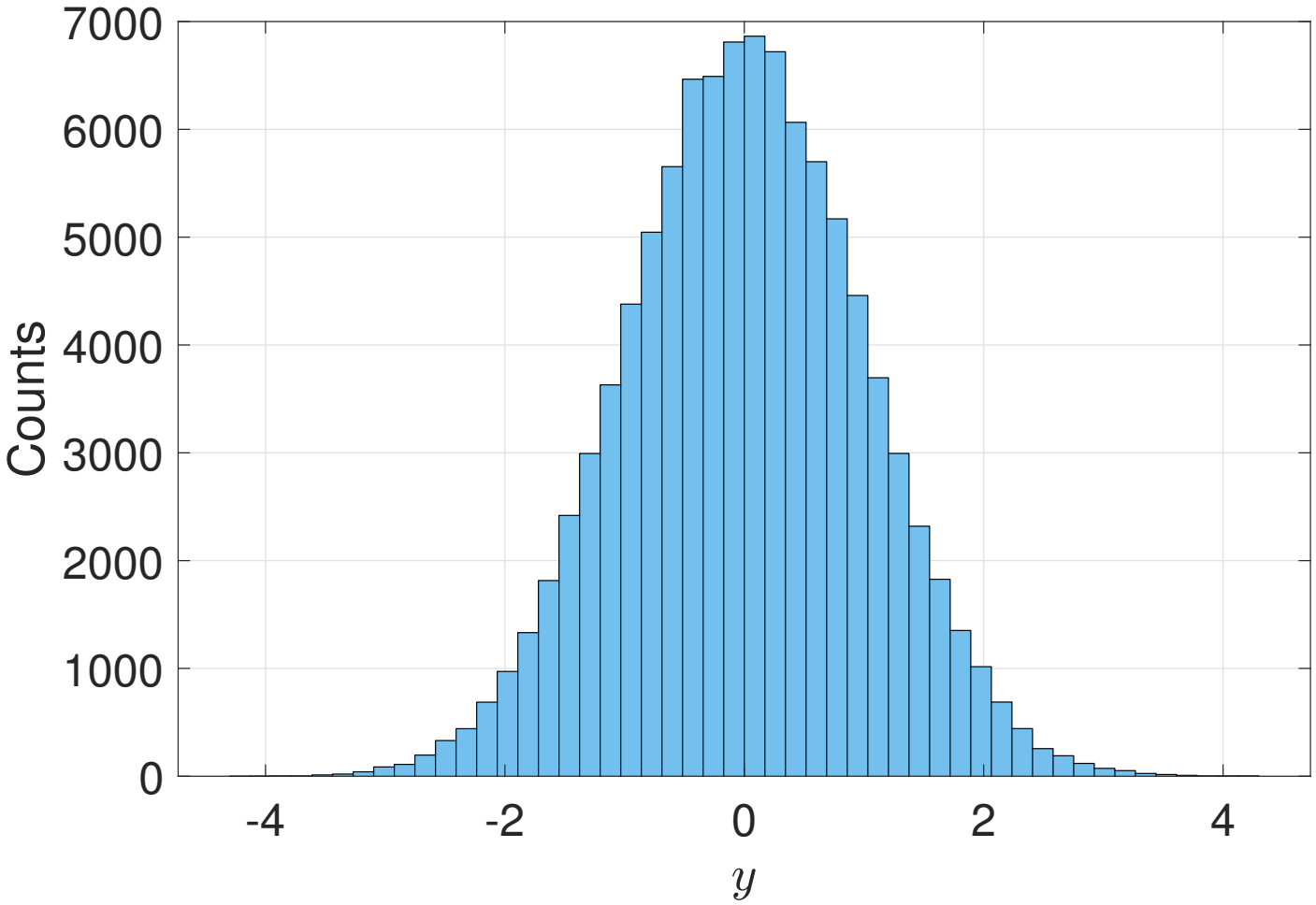
$\Longrightarrow$ plot $v_1, \ldots, v_K$ using bar-chart

Histograms: Example



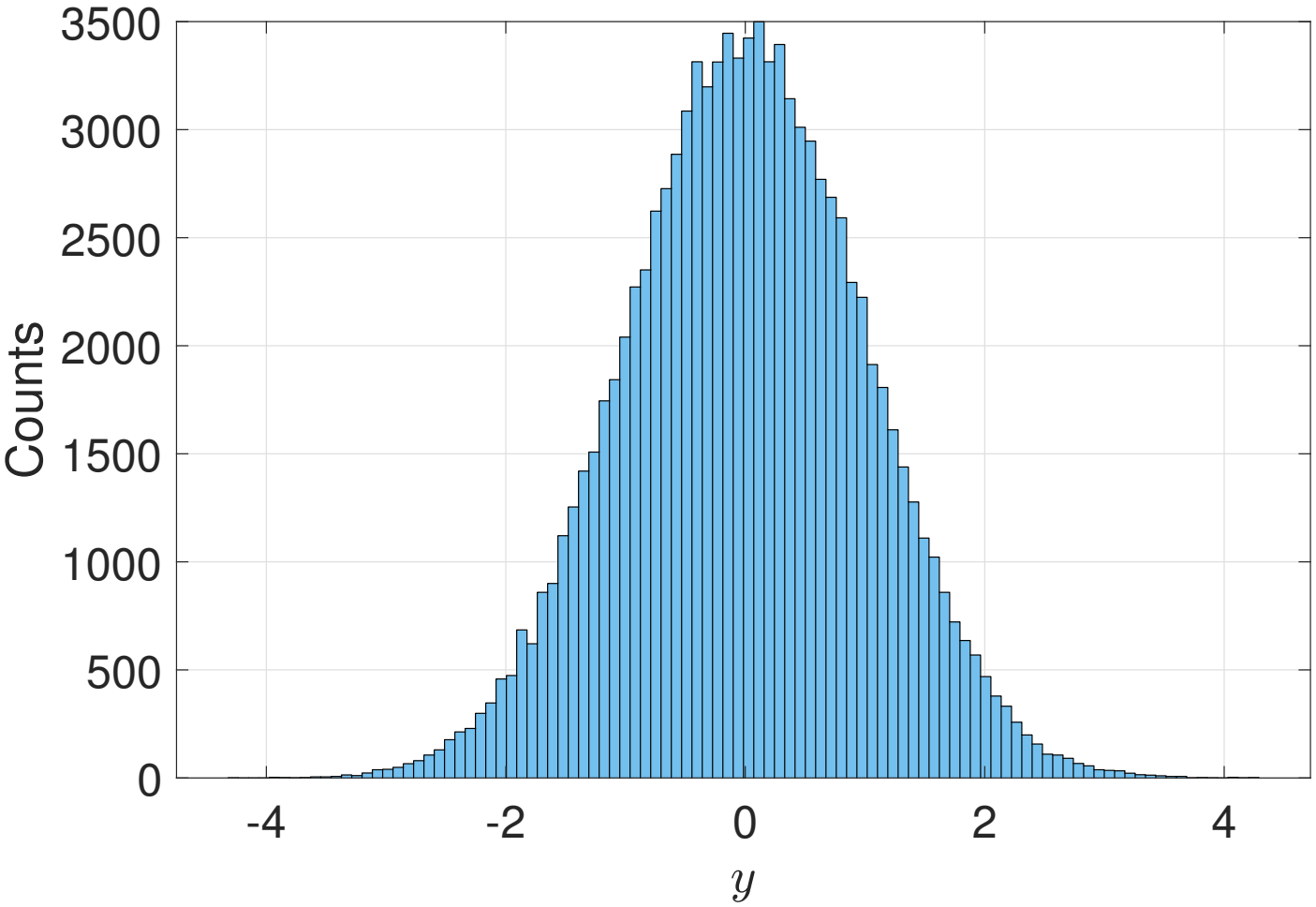
Histogram with $K = 20$ bins

Histograms: Example



Histogram with $K = 50$ bins; looks smoother

Histograms: Example



Histogram with $K = 100$; starting to look ragged

Descriptive Statistics

- Descriptive statistics summarise aspects of the data
- Usually lose information, but gain easy comprehension
- Contrast with inferential statistics

- But what is a "statistic"?
  - Let $\mathbf{y}$ denote a sample of data
  - Then a statistic is any function $s(\mathbf{y})$ of the data

- Some functions (statistics) more useful than others
  - But all describe properties of the data

Descriptive Statistics

- Descriptive statistics summarise aspects of the data
- Usually lose information, but gain easy comprehension
- Contrast with inferential statistics

- But what is a "statistic"?
  - Let $\mathbf{y}$ denote a sample of data
  - Then a statistic is any function $s(\mathbf{y})$ of the data

- Some functions (statistics) more useful than others
  - But all describe properties of the data

Measures of Centrality

- Let $\mathbf{y} = (y_1, \ldots, y_n)$ be a sample of n data points
- The the most common measure of centrality, or averageness, is the arithmetic mean

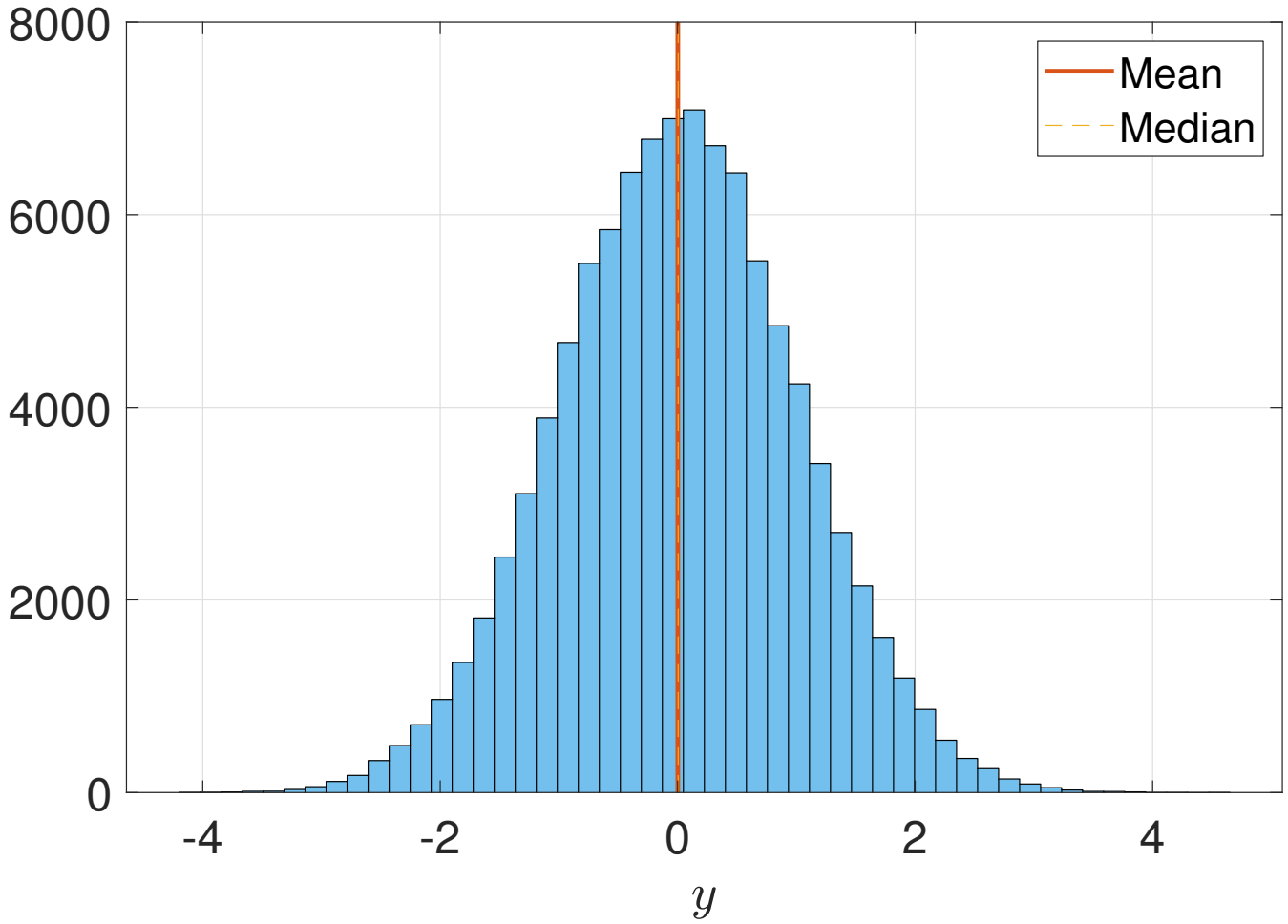
$$\bar{y} = \frac{1}{n} \sum_{j=1}^{n} y_j$$

- The mode is the most frequently occuring value in the sample
  - Of limited use for continuous numeric data
- Another common measure is the median, $\mathrm{med}(\mathbf{y})$
  - Value such that 50% of samples have values less than $\mathrm{med}(\mathbf{y})$
  - Easily found by sorting samples and finding middle sample

Mean vs Median

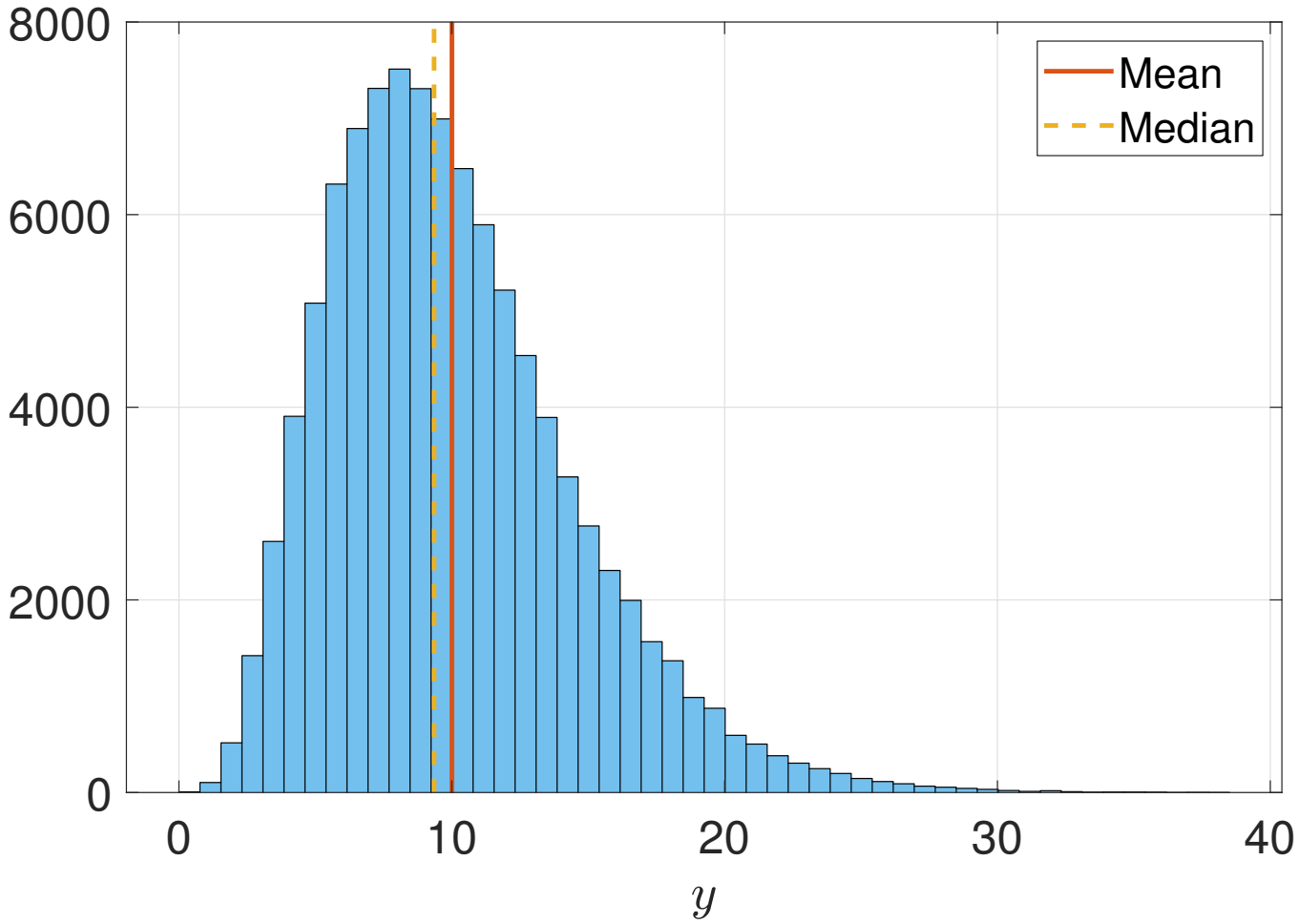
- The mean uses *all* the values of the sample
  - Any change to any sample changes the mean
  - The mean can be changed as much as desired by changing just one sample by a large enough amount
- The median uses at most two of the values of the sample
  - Is very resistant to changes to the samples not in the middle
- Example:
  - $\mathbf{y} = (1, 2, 3, 4, 5) \quad \Rightarrow \bar{y} = 3, \quad \text{med}(\mathbf{y}) = 3$
  - $\mathbf{y} = (1, 2, 3, 4, 50) \Rightarrow \bar{y} = 12, \quad \text{med}(\mathbf{y}) = 3$
- Why might we want to use mean over median then?

Mean vs Median: Symmetric Distributions



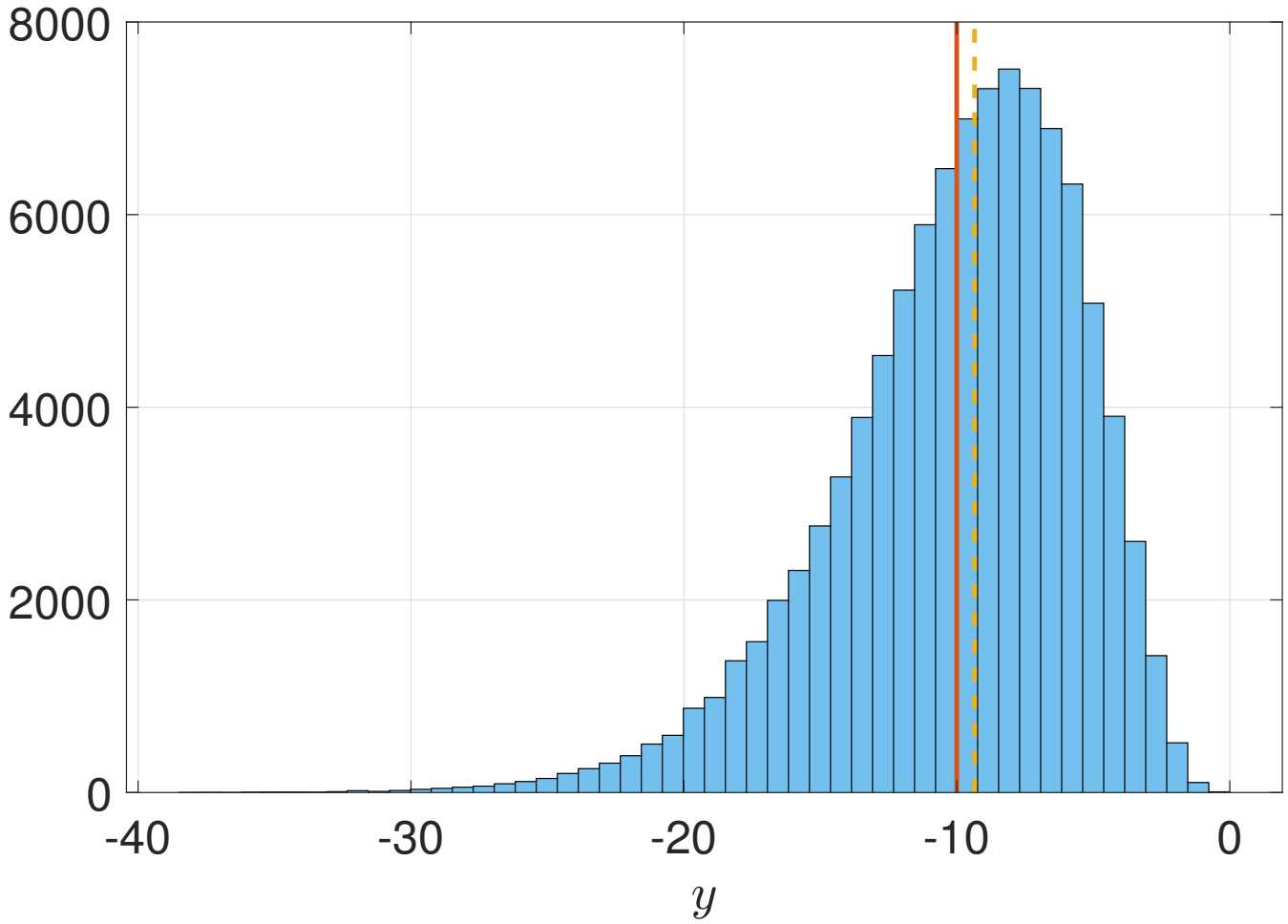
Symmetric distribution of data; mean and median (nearly) the same

Mean vs Median: Positively Skewed Data



Positively skewed data; mean greater than median

Mean vs Median: Negatively Skewed Data

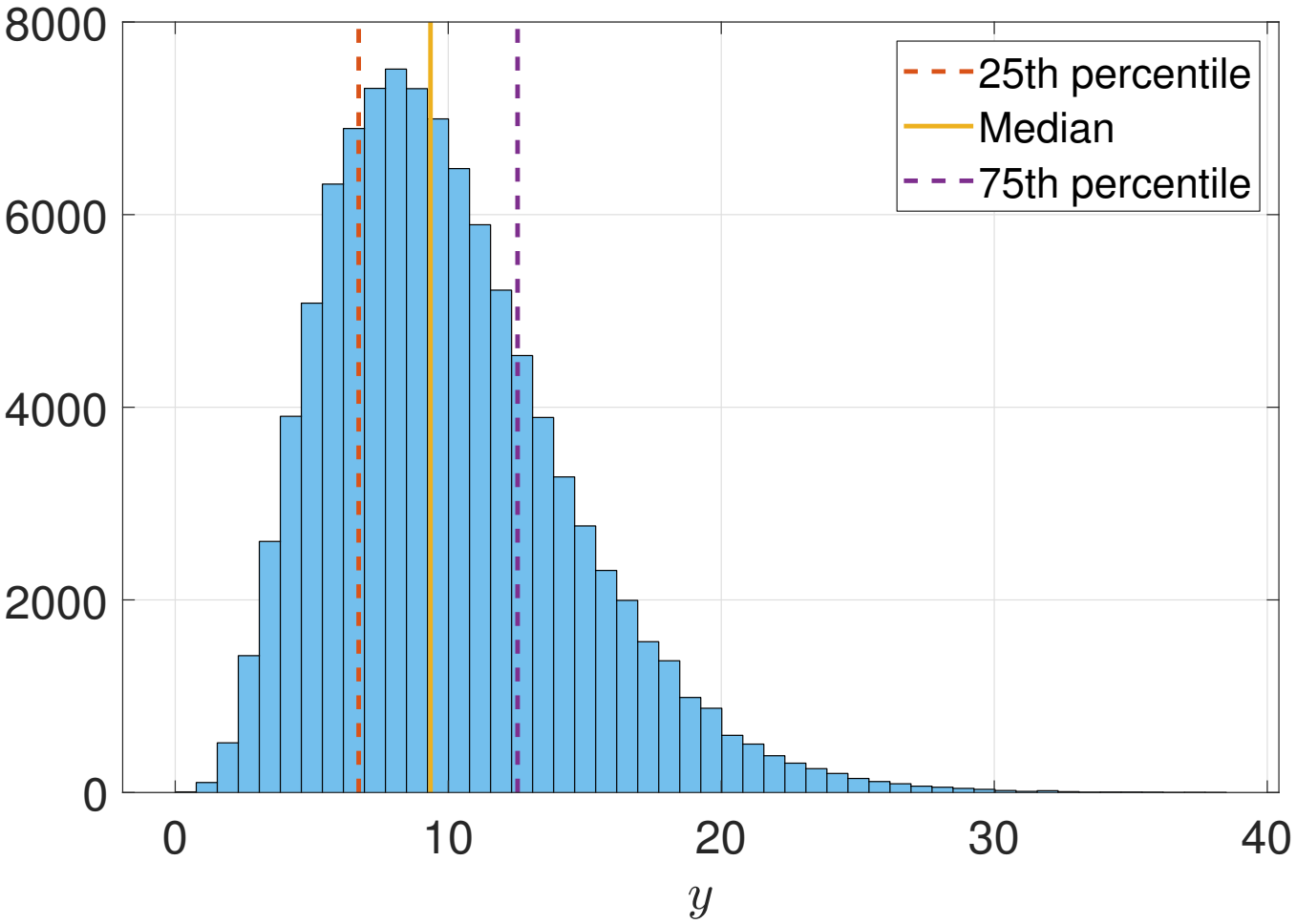


Negatively skewed data; mean less than median

Percentiles

- More generally, we can define the percentiles
  - The p-th percentile is the value, $Q(\mathbf{y}, p)$ such that $p\%$ of the values of the sample are lower than $Q(\mathbf{y}, p)$
- The median is simply the 50th percentile, $Q(\mathbf{y}, 50)$
- Other important percentiles are the 1st and 3rd quartiles
  - i.e., the 25th and 75th percentiles

Percentiles



Measures of Spread (1)

- Measures of centrality tell us about the typical value of the sample
- Measures of spread tell us how much the samples differ, on average, from the typical value
- The most straightforward is the range

$$\text{rng}(\mathbf{y}) = \max\{\mathbf{y}\} - \min\{\mathbf{y}\}$$

where

- $\min\{\mathbf{y}\}$ denotes the minimum value in the sample;
- $\max\{\mathbf{y}\}$ denotes the maximum value in the sample.

Measures of Spread (2)

- The most common measure of spread used is the sample standard deviation

$$s(\mathbf{y}) = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \bar{y})^2}$$

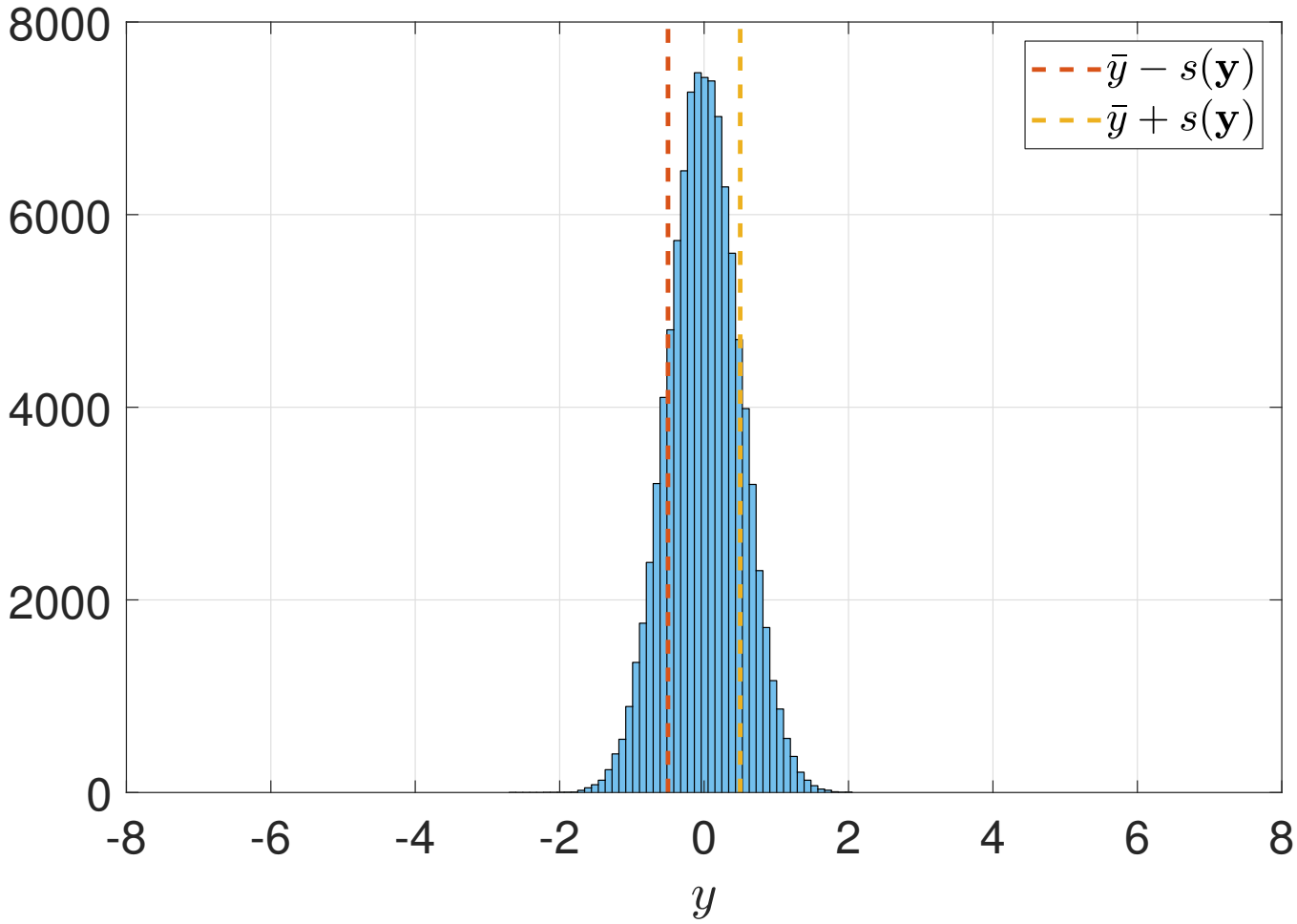
- The sample standard deviation is the arithmetic mean of the squared deviations from the sample mean
  $\Longrightarrow$ has the same unit as the data

- Like the mean, is sensitive to changes in the sample
- Often, the sample variance

$$v(\mathbf{y}) = s^2(\mathbf{y})$$

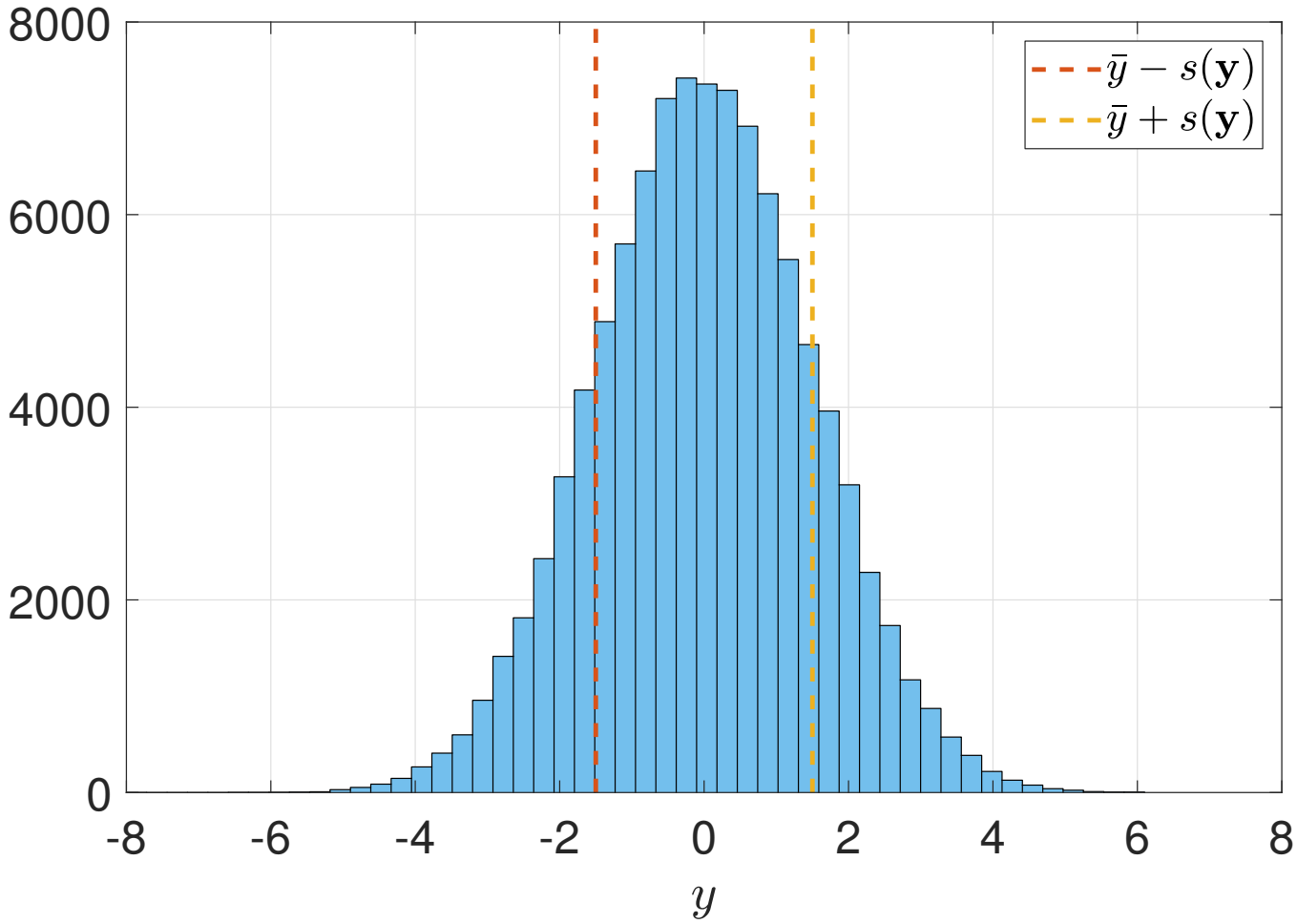
is used, as it can be easier to work with

Measures of Spread: Example



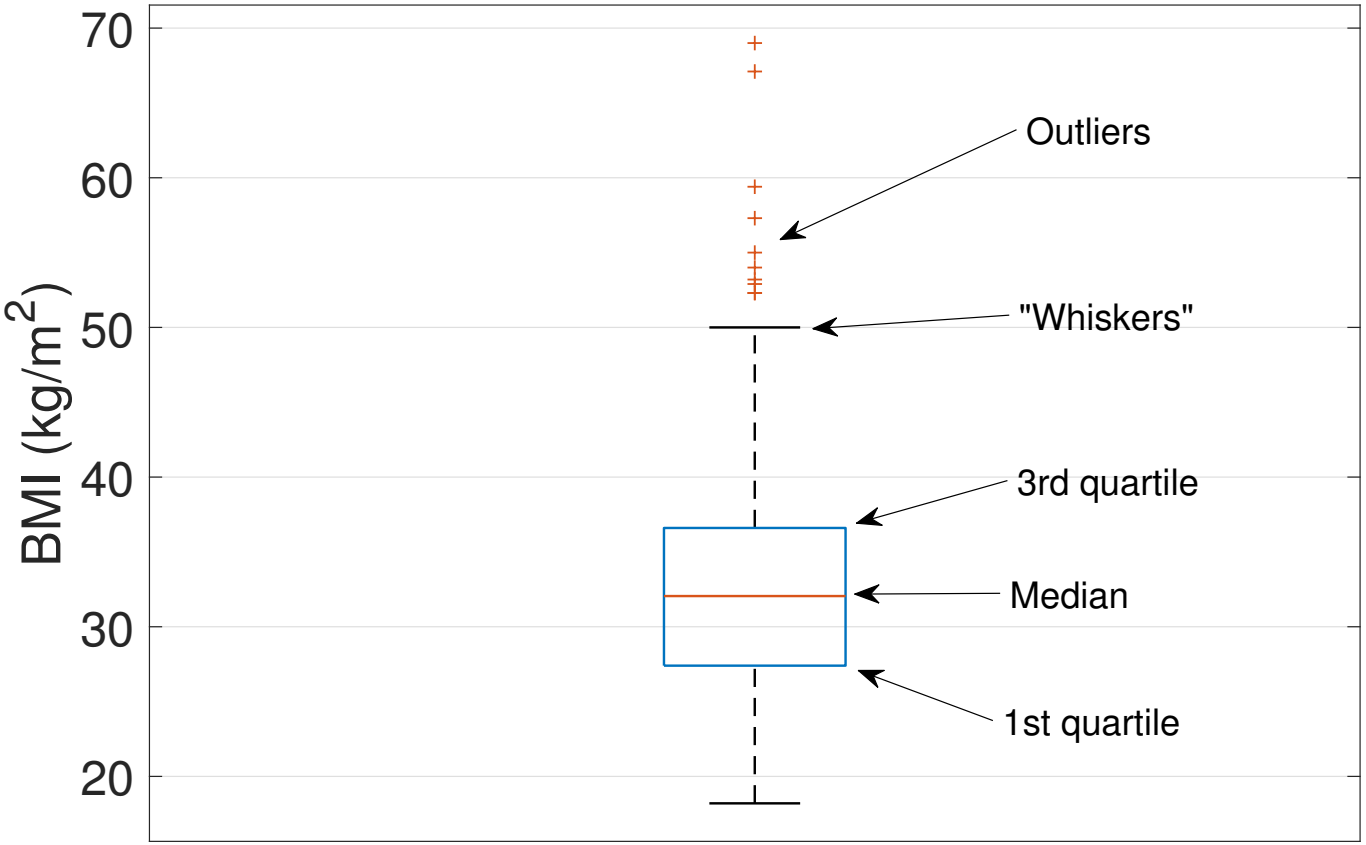
$\text{rng}(\mathbf{y}) = 4.63$ $(\min\{\mathbf{y}\} = -2.61,\ \max\{\mathbf{y}\} = 2.01)$, $s(\mathbf{y}) = 0.5$

Measures of Spread: Example



$\text{rng}(\mathbf{y}) = 13.89$ $(\min\{\mathbf{y}\} = -7.84,\ \max\{\mathbf{y}\} = 6.05)$, $s(\mathbf{y}) = 1.5$

Visualising Continuous Data: Boxplots



Boxplot graphically captures centrality, spread and skewness in one plot

Association Between Two Continuous Variables

- Let $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ be two numeric variables measured on the same objects
  - We might ask if there is an association between $\mathbf{x}$ and $\mathbf{y}$
- Pearson correlation measures linear association

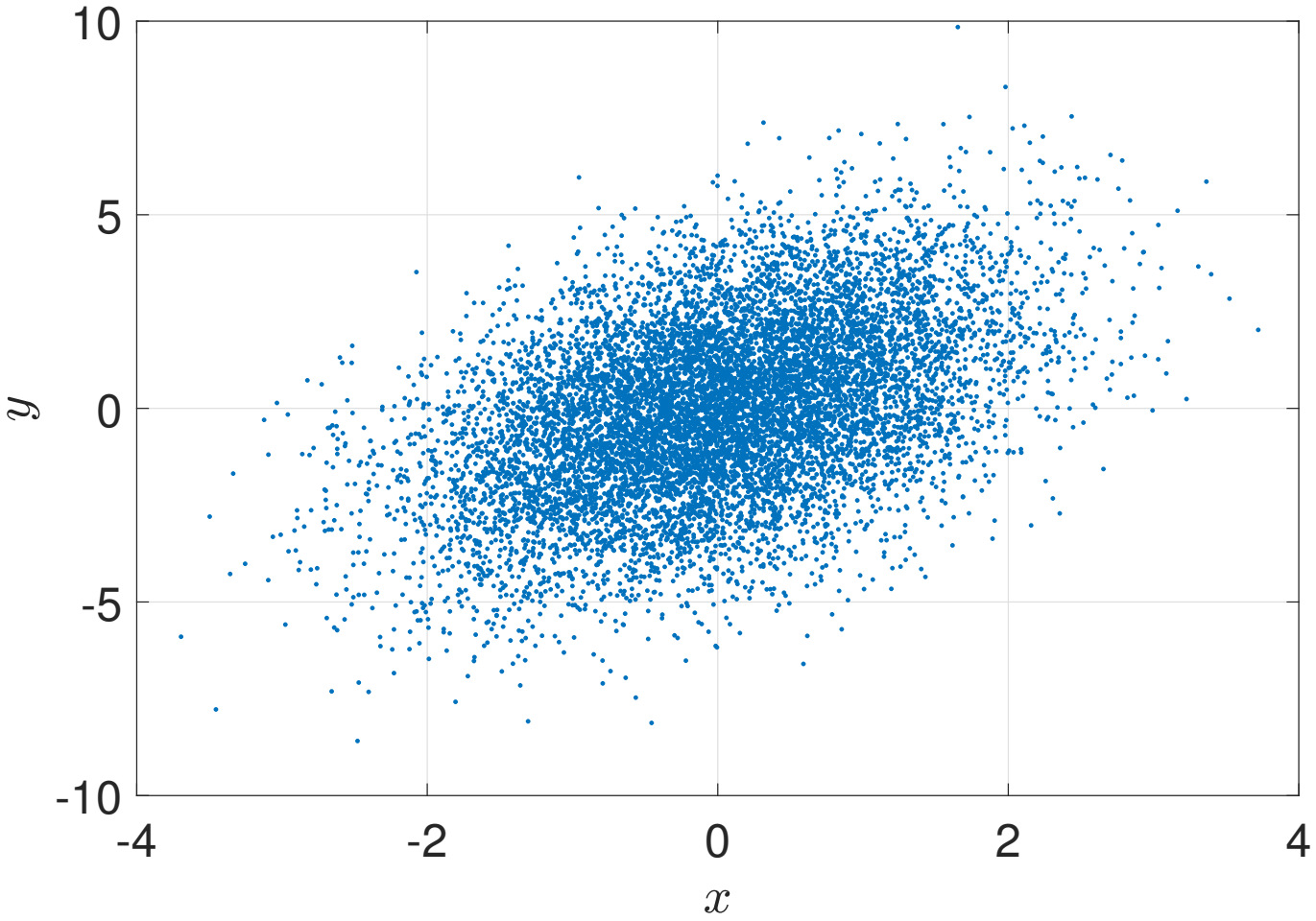
$$R(\mathbf{x}, \mathbf{y}) = \frac{\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})}{n\, s(\mathbf{x}) s(\mathbf{y})}$$

  - Correlation is always between -1 (completely negatively correlated) and 1 (completely positively correlated)
  - A correlation of zero implies there is no linear association $\Longrightarrow$ does not imply no non-linear association
- Remember: correlation $\neq$ causation!

Scatter Plots

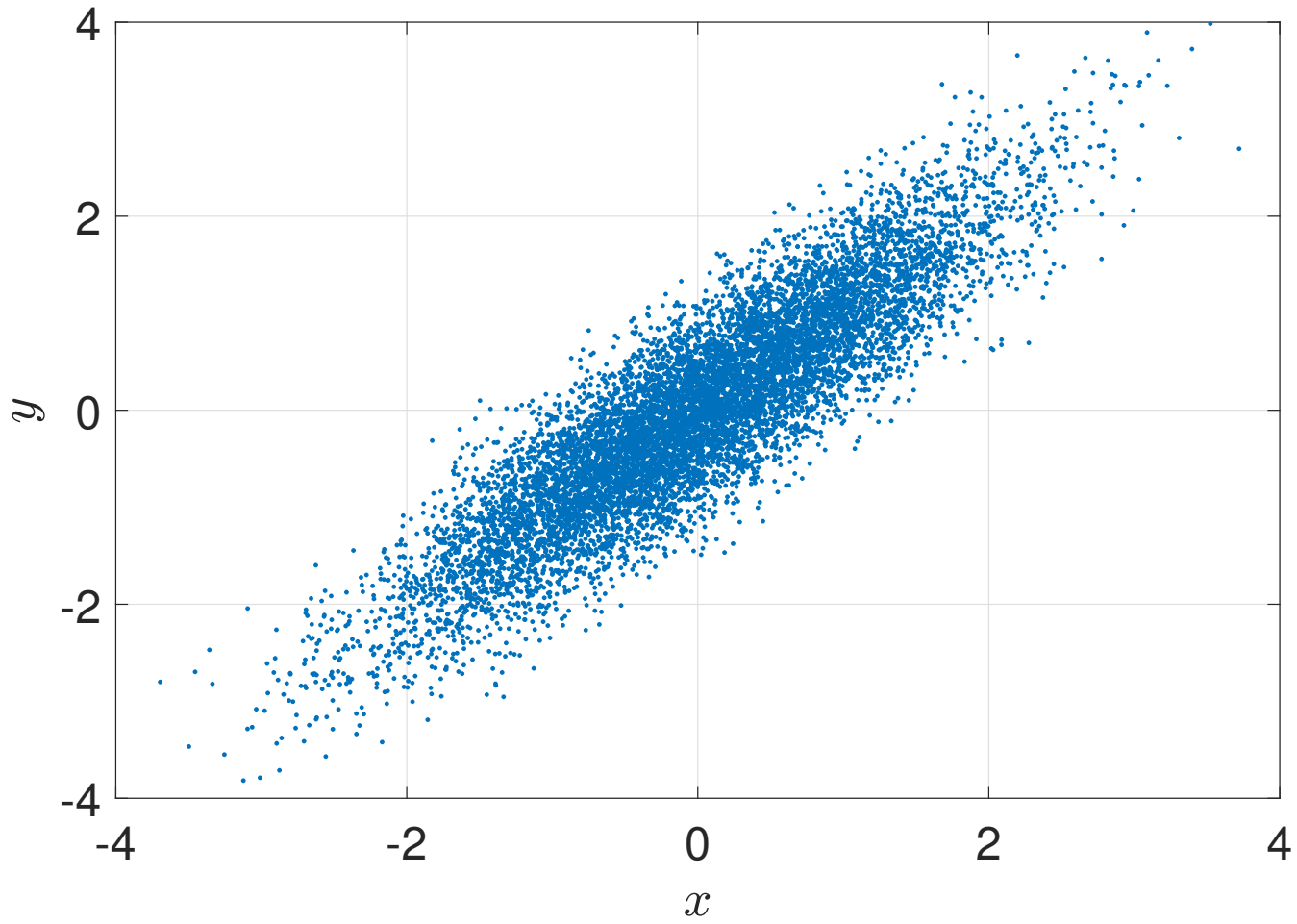
- Scatter plots help us visualise relationships between two (usually) numeric variables
  - Plot points, with one variable on x-axis and one on y-axis
- Can be used to visually look for association
- Correlation coefficients are statistics that quantatatively measure the strength of the association between two variables
  - The two can be combined for more information
- Three-variable scatter plots, like almost all three-dimensional plots, should be avoided

Correlation/Scatter Plot Example (1)



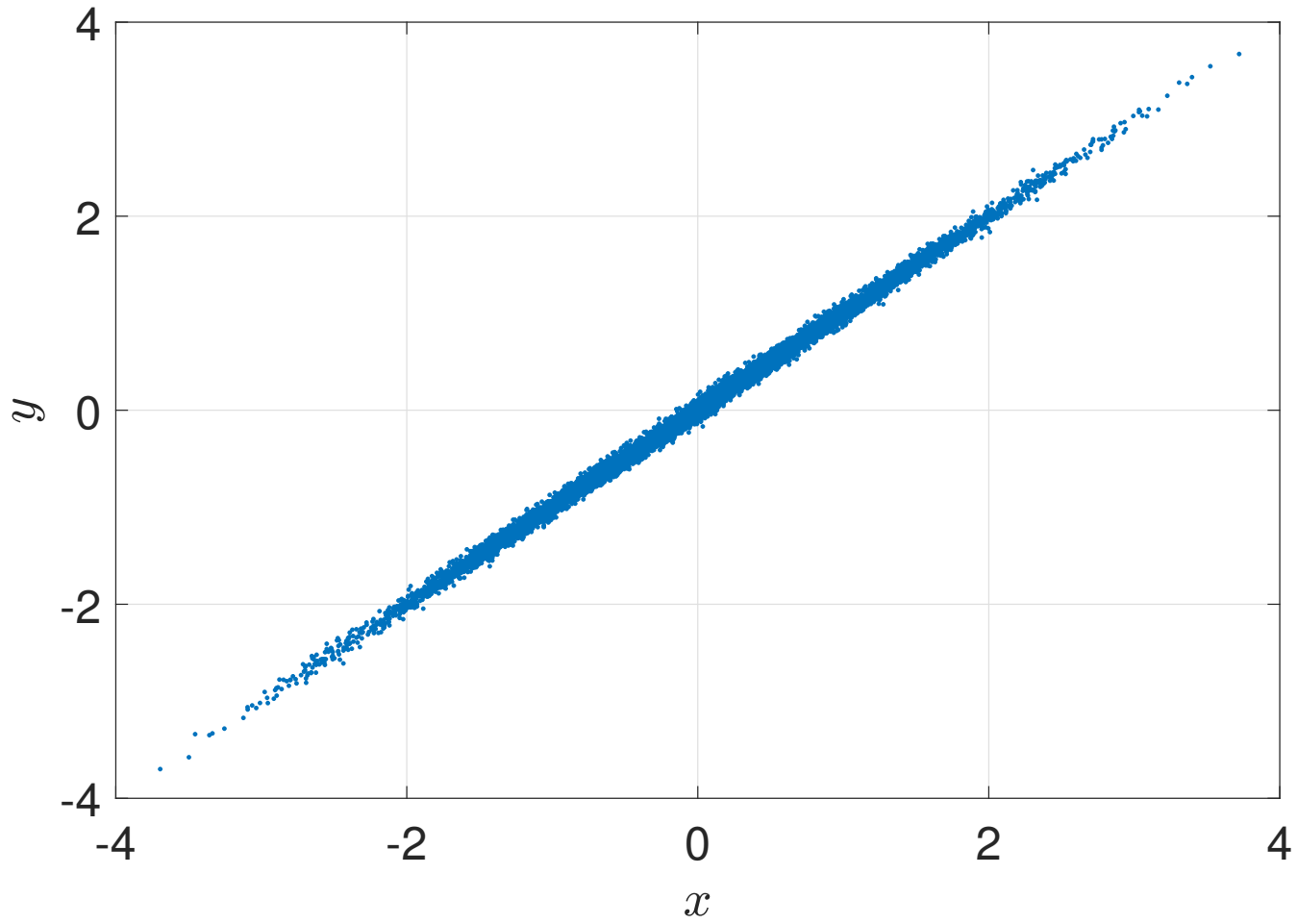
$$R \approx 0.44$$

Correlation/Scatter Plot Example (2)



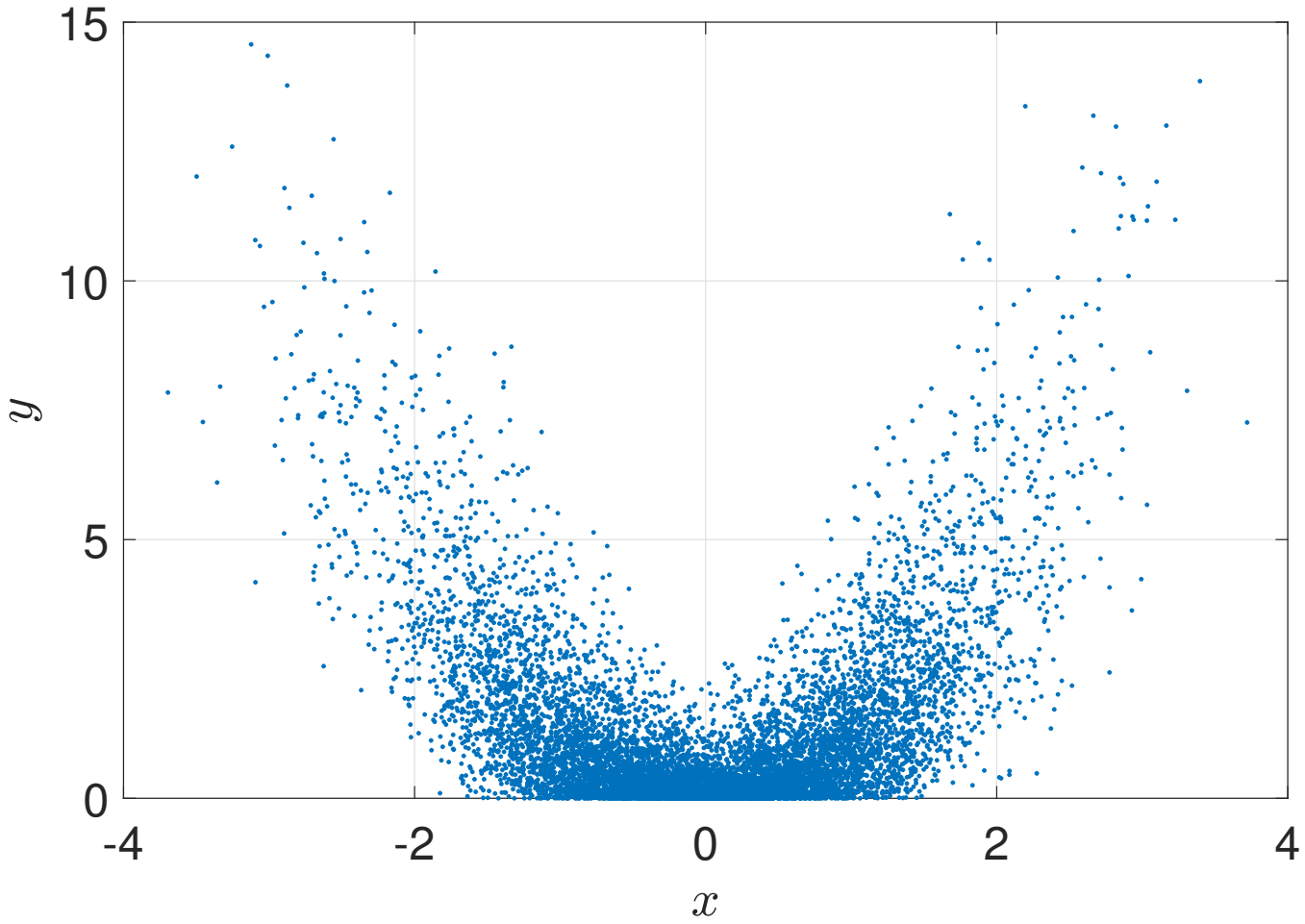
$$R = 0.9$$

Correlation/Scatter Plot Example (3)



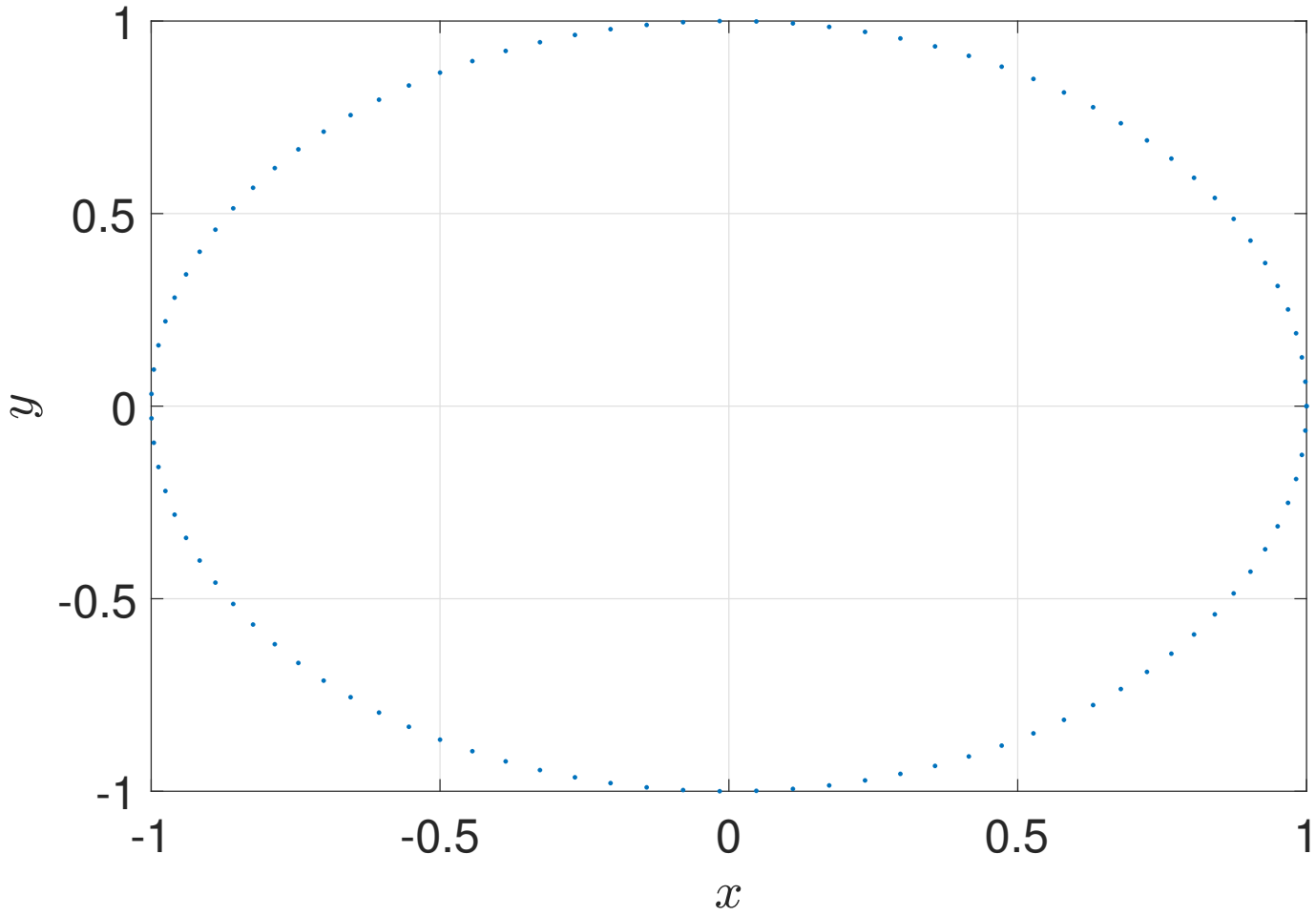
$$R \approx 0.999$$

Correlation/Scatter Plot Example (4)



$R \approx 0.01$ – though clearly associated, as $y = x^2 + \text{noise}$

Correlation/Scatter Plot Example (5)

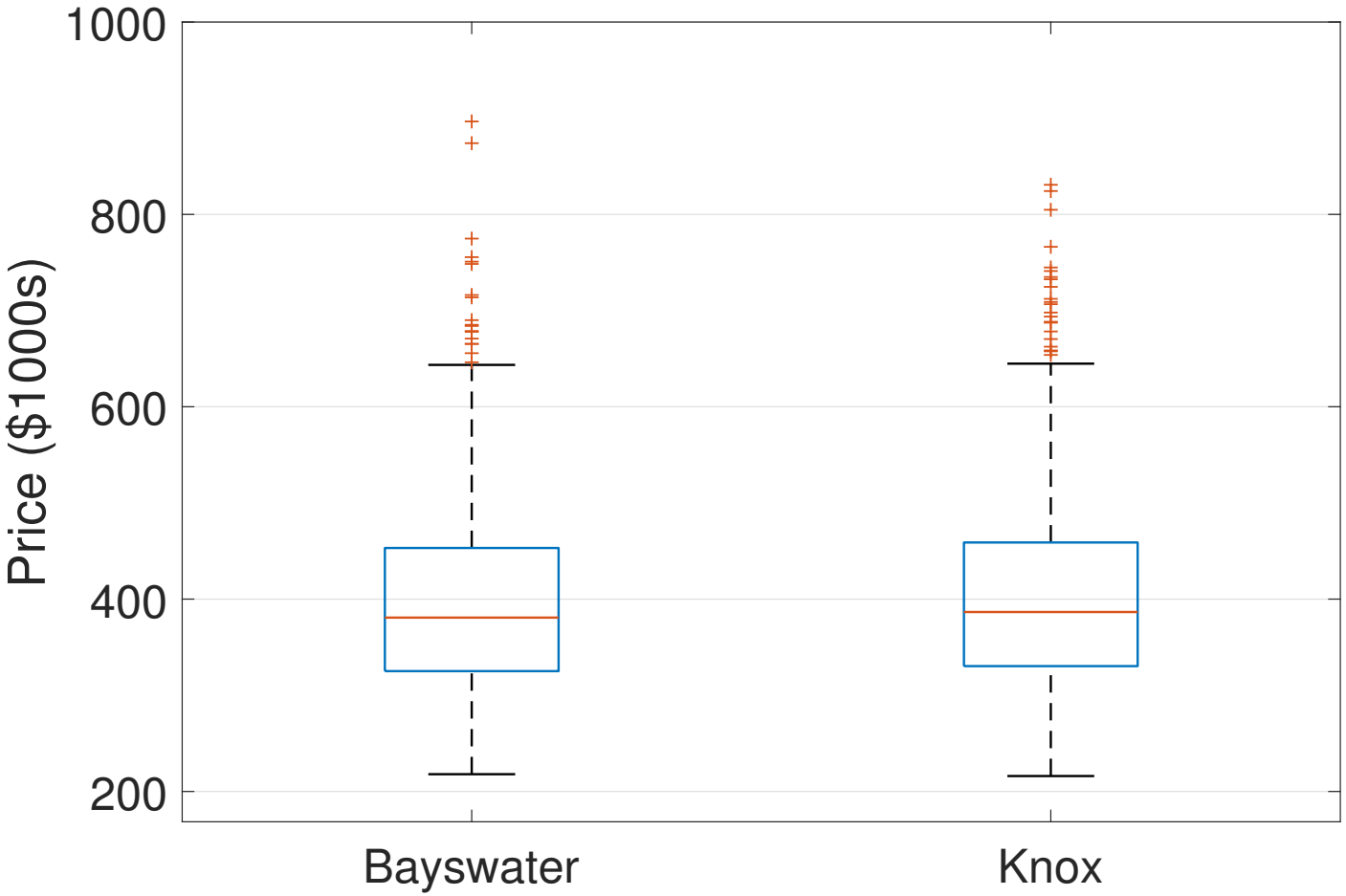


$R = 0$, though there is a deterministic association between x and y

Association Between Categorical and Numeric Variables

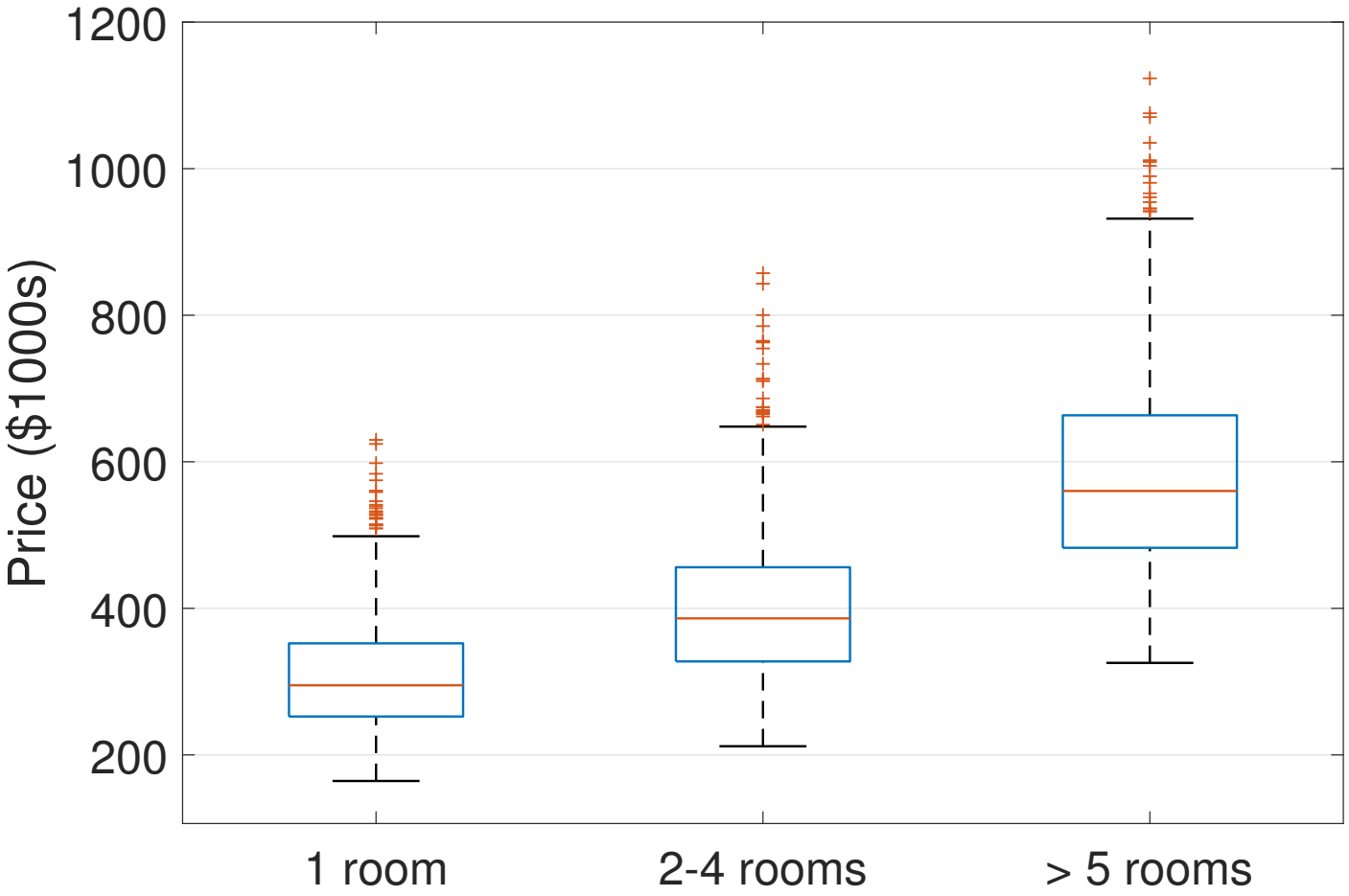
- If $\mathbf{x}$ is categorical, and $\mathbf{y}$ is numeric, how to visualise?
- A standard approach is the side-by-side boxplot
  - Divide the data between categories, then plot boxplots for each group
  - Do the boxplots look different?
- If $\mathbf{x}$ and $\mathbf{y}$ are both categorical, we can use a side-by-side bargraph instead
  - Are the distributions/bargraphs different between categories?
  - If so, there is a possible association

Example: Categorical and Numeric Variables (1)



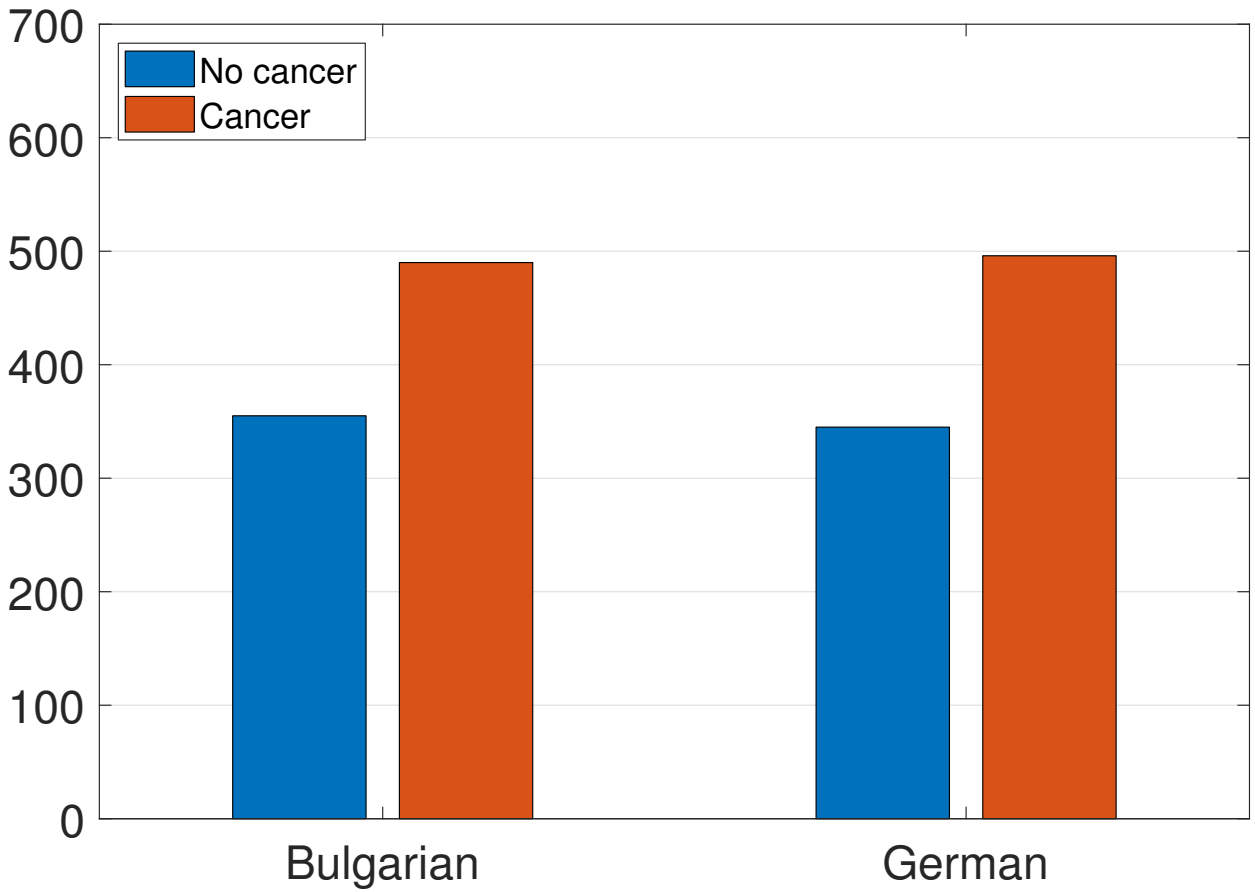
Distribution of price similar between suburbs

Example: Categorical and Numeric Variables (2)



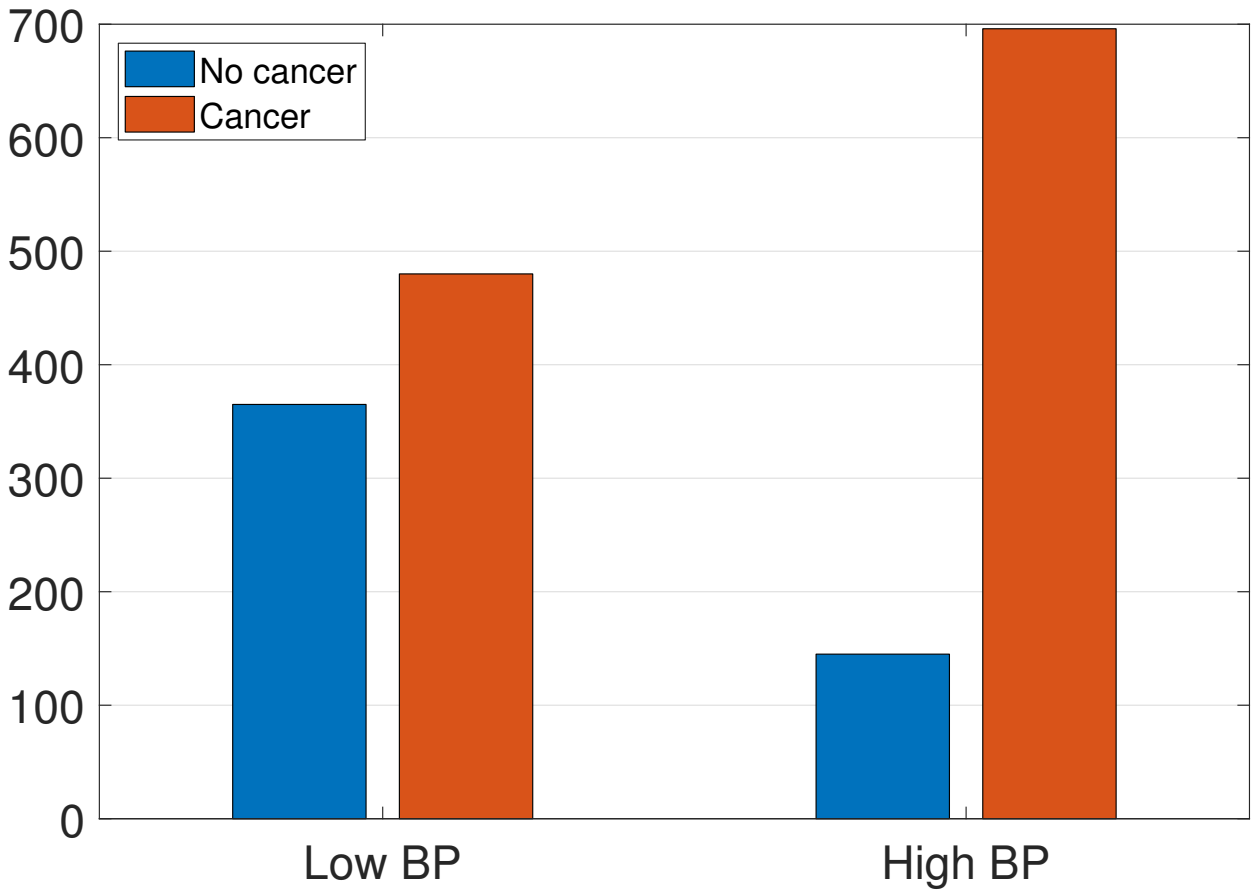
Distribution of price varies greatly with number of rooms

Example: Two Categorical Variables (1)



Frequency of cancer does not seem to change with ethnicity; unlikely to be associated

Example: Two Categorical Variables (2)



Frequency of cancer changes substantially with blood pressure; likely to be strong association

Reading/Terms to Revise

- Reading for this week: Chapter 2 of Ross.
- Terms you should know:
  - Histogram;
  - Measures of central tendency: mean, median, mode;
  - Measures of dispersion: standard deviation, variance, range;
  - Percentiles and quartiles;
  - Scatter plot;
  - Correlation coefficient