Gentle Introduction to Machine Learning: **Statistics**

A Reproducible Research Workshop

Simon Stone

Research Data Services

Dartmouth College

· Reproducible · Research @ DARTMOUTH

About the Reproducible Research Group

- Joint venture of **Research Computing @ ITC** and **Research Data Services @ Library**
- Consult with **experts** on
  - research data management,
  - data visualization,
  - biomedical research support,
  - spatial data and GIS,
  - high performance and research computing,
  - statistical analysis,
  - economics and social sciences data
- **Meet** the people on campus that support your reproducible research lifecycle
- **Engage** in community discussions to learn from other researchers on campus
- Attend a workshop to **learn** practical tools and tips

About Research Data Services

Research Data Management

Data Management Plans (DMPs) for sponsored projects

Finding and using 3rd party data

Collection and cleaning of data

Organization and documentation

Publishing and Repositories

Data Analysis/Visualization

Textual, numeric, spatial data

Reproducible research workflows

Scripting in R: tidyverse core package (i.e., ggplot, dplyr, tydr, tibble, etc.)

Scripting in Python: NumPy, SciPy, Pandas, Scikit-learn, Matplotlib, Seaborn, (OpenCV, PyTorch, TensorFlow, Tesseract, NLTK, etc.)

Computational Scholarship

Computational project planning

Collections as Data

Storytelling with data and visualizations

Text and data mining

Digital Humanities support

Computational Pedagogy

Work with us

ResearchDataHelp@groups.dartmouth.edu

Jeremy Mikecz
Research Data Science Specialist
jeremy.m.mikecz@dartmouth.edu
dartgo.org/jeremyappts

Simon Stone
Research Data Science Specialist
simon.stone@dartmouth.edu
dartgo.org/meetwithsimon

Lora Leligdon
Head of Research Data Services
lora.c.leligdon@dartmouth.edu
dartgo.org/lora

Gentle Introduction to Machine Learning

Intro

Activity:

🤔 Why are you here?

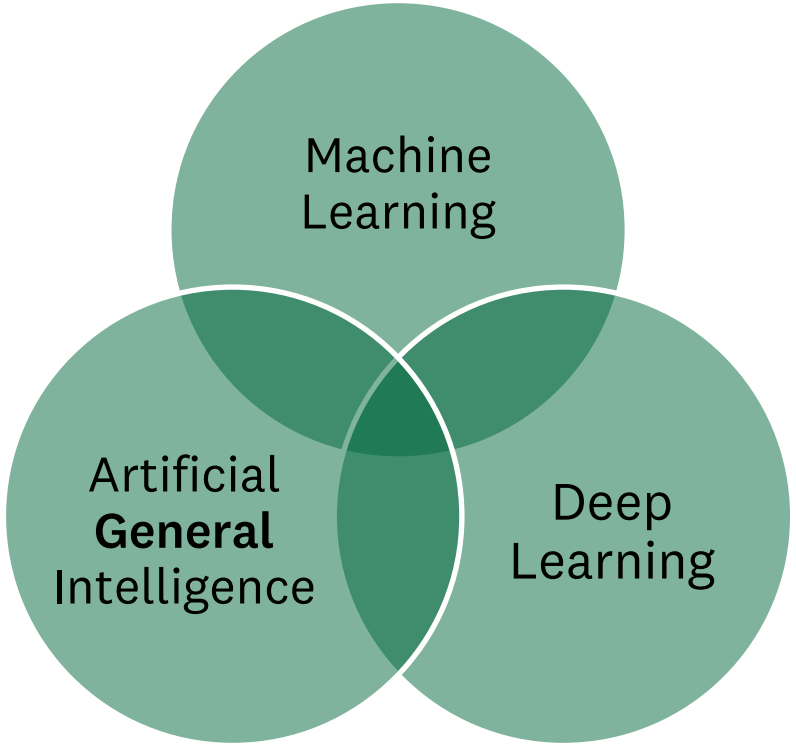
💡 What do you hope to take away from this workshop (series)?

Gentle Introduction to Machine Learning

Intro

Machine Learning is ”the field of study that gives computers the ability to learn without explicitly being programmed.”

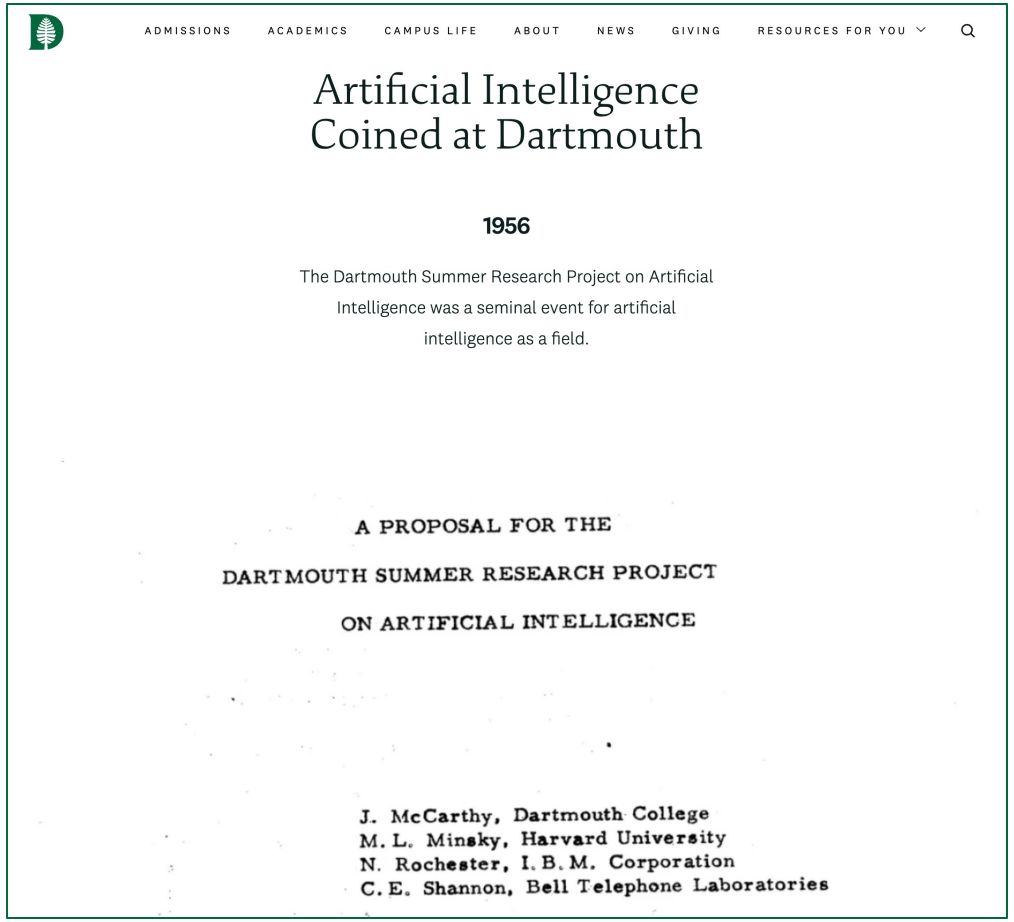
- Arthur Samuel, 1959 (paraphrased)



Machine Learning is what we should call (the current) Artificial Intelligence!

Gentle Introduction to Machine Learning

Intro



https://home.dartmouth.edu/about/artificial-intelligence-ai-coined-dartmouth

Gentle Introduction to Machine Learning

Intro

The New York Times

CURRENTS

A.I. Here, There, Everywhere

Many of us already live with artificial intelligence now, but researchers say interactions with the technology will become increasingly personalized.

By Craig S. Smith

Published Feb. 23, 2021 Updated March 9, 2021 5 MIN READ

- Machine Learning is everywhere
- We are interacting with it every day
- But how much do most of us know about our new algorithmic overlords?

Gentle Introduction to Machine Learning

Intro

Aims of this series:

🪄 **Demystify** the field a bit and give context to the buzzwords

💡 A working **mental model** how machine learning algorithms ~~think~~ calculate

🤔 To provide enough knowledge to **think critically** about “A.I.”

😎 To inspire you to **confidently use machine learning** in your work and personal life

Gentle Introduction to Machine Learning

Intro

- **Statistics**
  - 📋 A brief survey of the fundamentals for Machine Learning
- **Regression** (April 25)
  - 📈 How can an algorithm find relationships between two variables?
- **Classification** (May 9)
  - 🏷️ How can an algorithm put a label on a real-world object?

Basics

Intro

Activity:

✍️ Define “Statistics”

One definition (Upton & Cook, 2008):

“The science of collecting, displaying, and analysing data.”

In Machine Learning:

“The way that algorithms perceive the world”

Basics

Intro

Two sides of the (fair) coin:

Descriptive Statistics

🔍 **Describing** an apparent mess of data

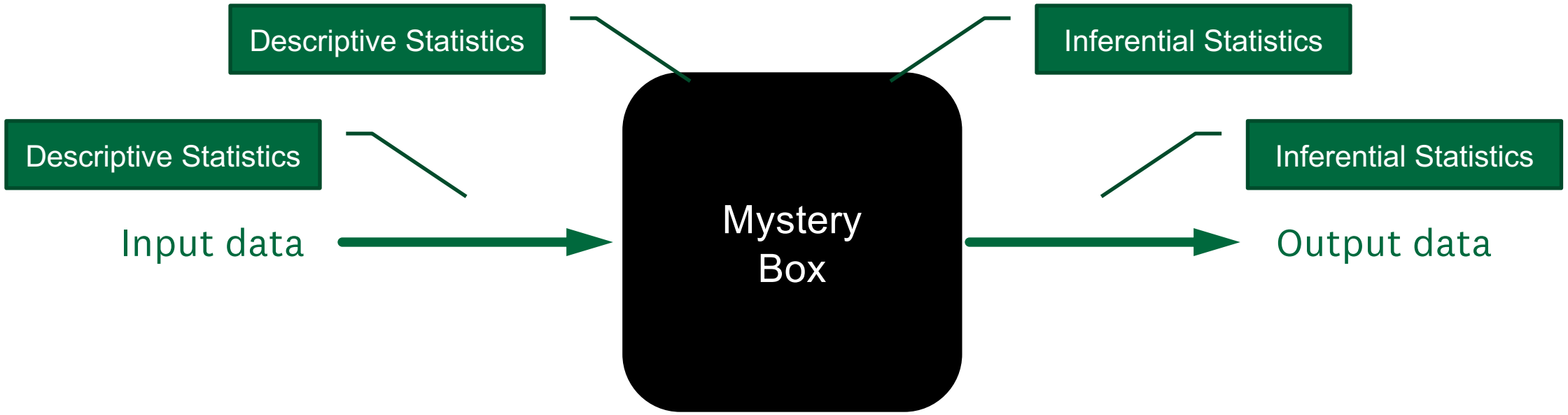
Inferential Statistics

🔮 Making **predictions** based on a mess that represents a larger mess

Basics
Intro

Descriptive Statistics
Inferential Statistics

Every machine learning model, ever:



Descriptive Statistics

Random Process, Variable, Data

Data: Recorded pieces of information

Random Variable: The underlying characteristic or quantity observed through data

Random Process:

- Assigning a value (data) to a variable randomly
- random ≠ arbitrary
- The deterministic internals of the process are not known
- The process is described by how likely it is to observe a specific value

Descriptive Statistics
Sample, Population

Population:

- The entirety of a group of interest (e.g., everyone at Dartmouth)

Sample:

- The relatively few examples from the population that were observed (e.g., survey respondents)

Descriptive Statistics

Types of data

Data: Recorded pieces of information

Variable: The underlying characteristic or quantity observed through data

Activity:

We want to statistically describe the population of people at Dartmouth to a machine learning model.

💭 Pair up and identify some variables that may help us do that

Descriptive Statistics

Types of data

- Categorical data (qualitative)
  - 🗄️ Data which can be divided into **groups**
  - 🪜 May have inherent order (**ordinal** scale) or not (**nominal** scale)
- Numerical data (quantitative)
  - Data which is in the form of numbers with **mathematical meaning**
  - Can be **discrete** (e.g., the count of books) or **continuous** (e.g., the cost of books)

Descriptive Statistics

Visualizing Data

Example:

- 🎲 Rolling a six-sided die
- 🎲 100 observations (i.e., rolls)

📊 How can we visualize the results in a diagram?

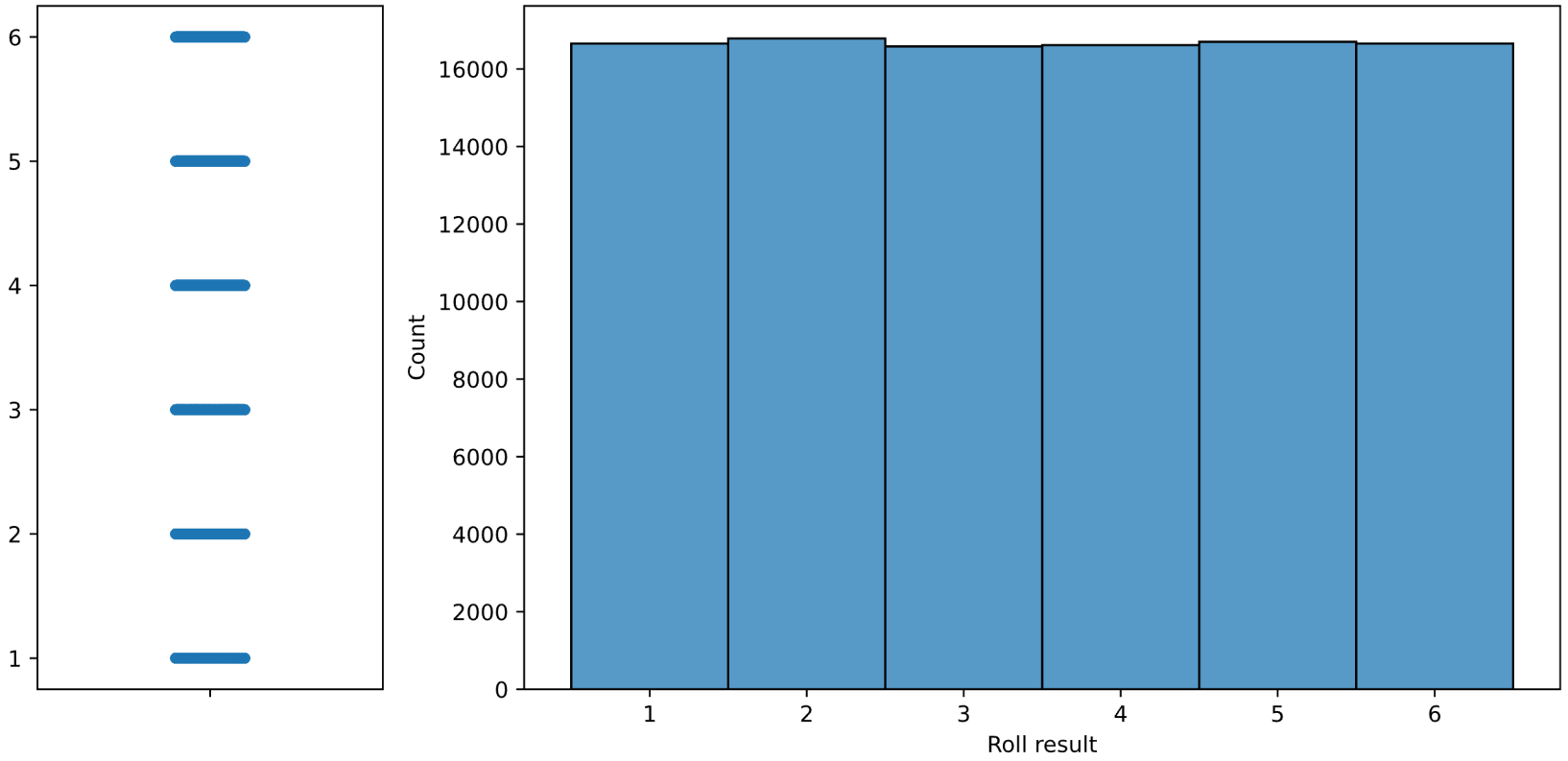
🔮 What would you expect to see?

Descriptive Statistics
Visualizing Data

Demo

Descriptive Statistics

Visualizing Data: Discrete uniform distribution



Descriptive Statistics

Visualizing Data

Example:

- 🎲 🎲 Rolling multiple six-sided dice
- 🎲 🎲 100 observations (i.e., rolls)

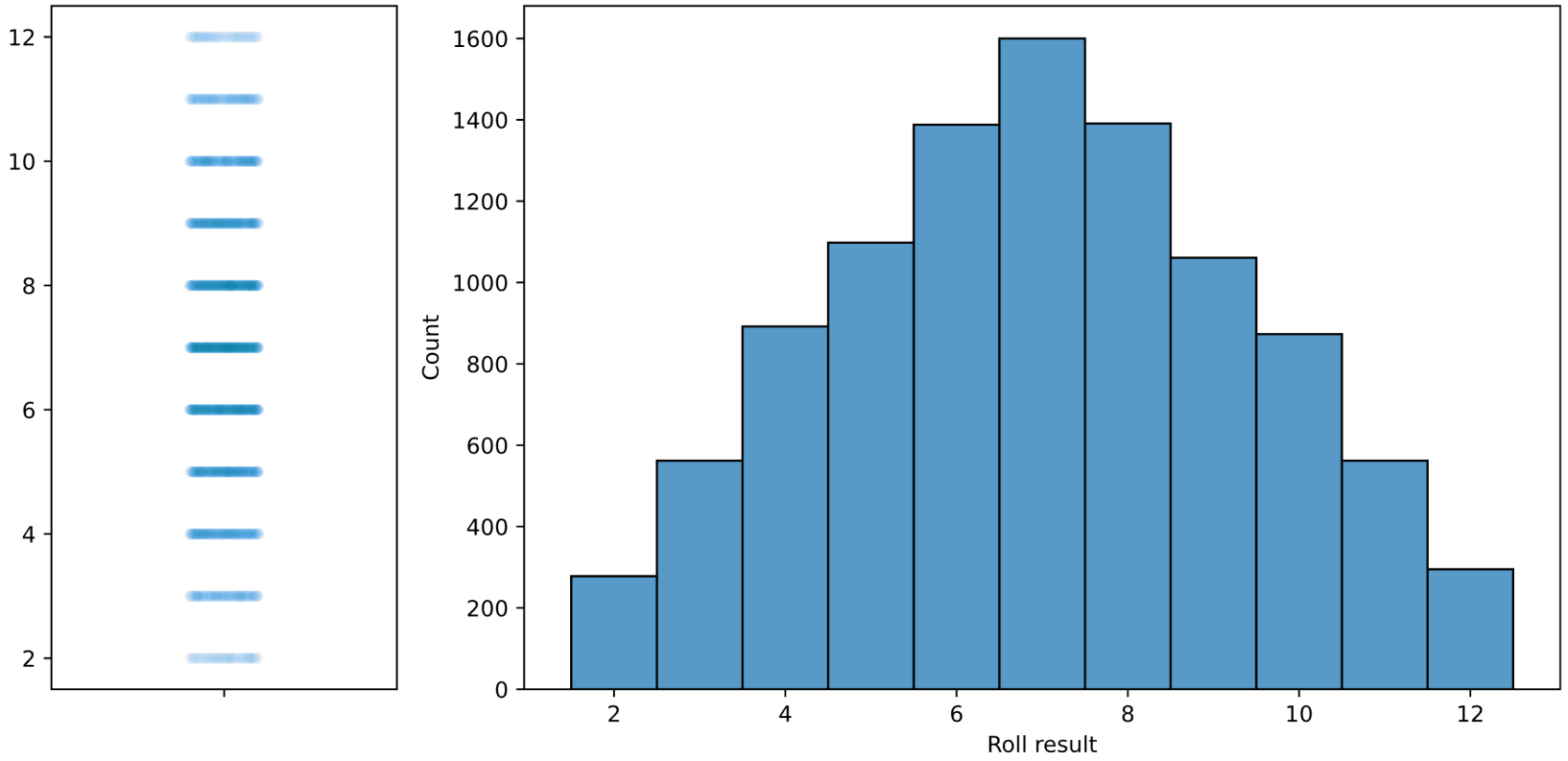
🔮 What would you expect to see now?

Descriptive Statistics
Visualizing Data

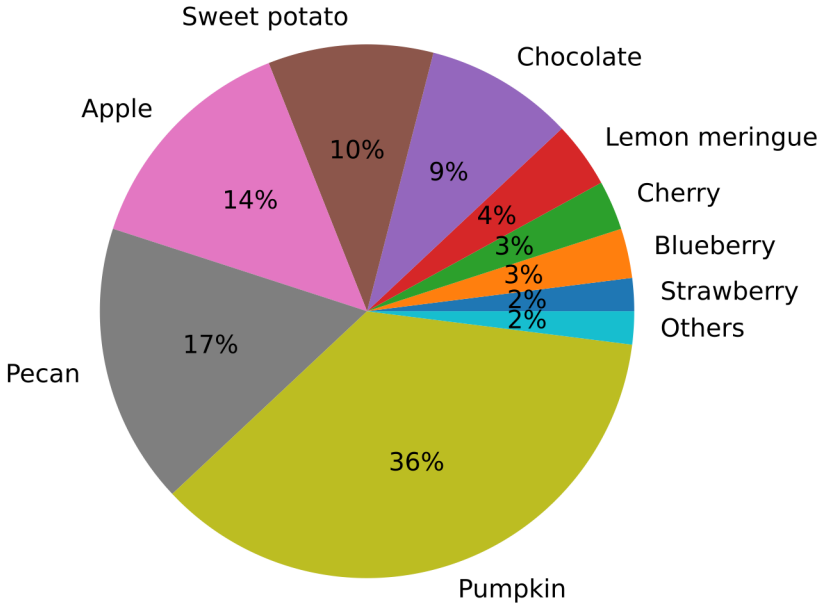
Demo

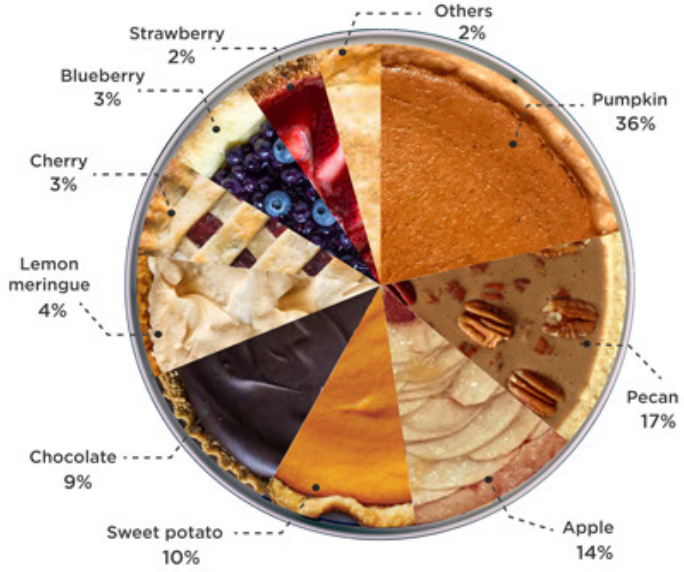
Descriptive Statistics

Visualizing Data: Discrete non-uniform distribution



Descriptive Statistics
Visualizing Data: Categorical distribution



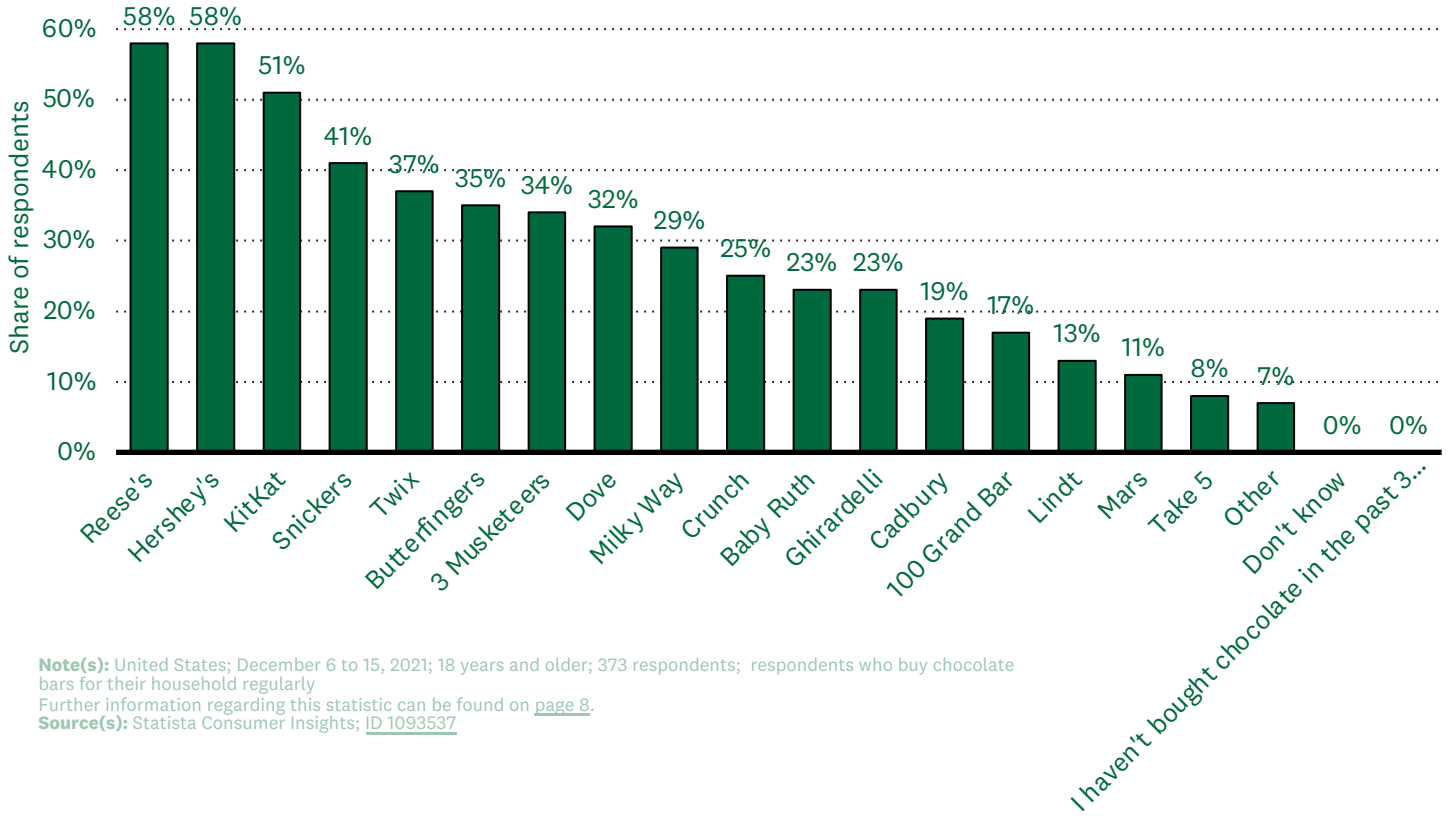


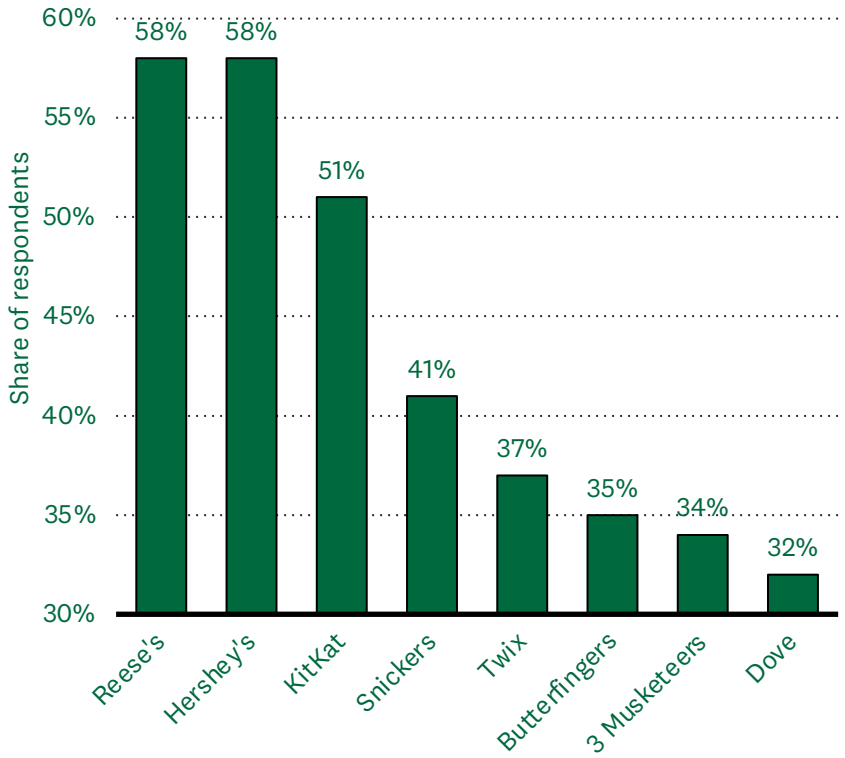
Source:
Thanksgiving Pie Survey Breaks Down America's Favorites. (2017). Delta Dental. URL: https://www.deltadental.com/us/en/about-us/press-center/2017/thanksgiving-pie-survey-breaks-down-america-s-favorites.html

Descriptive Statistics

Visualizing Data: Categorical distribution

”Which chocolate bars have you bought in the past 3 months?”





Note(s): United States; December 6 to 15, 2021; 18 years and older; 373 respondents; respondents who buy chocolate bars for their household regularly
Further information regarding this statistic can be found on page 8.
Source(s): Statista Consumer Insights; ID 1093537

Descriptive Statistics

Summarizing data

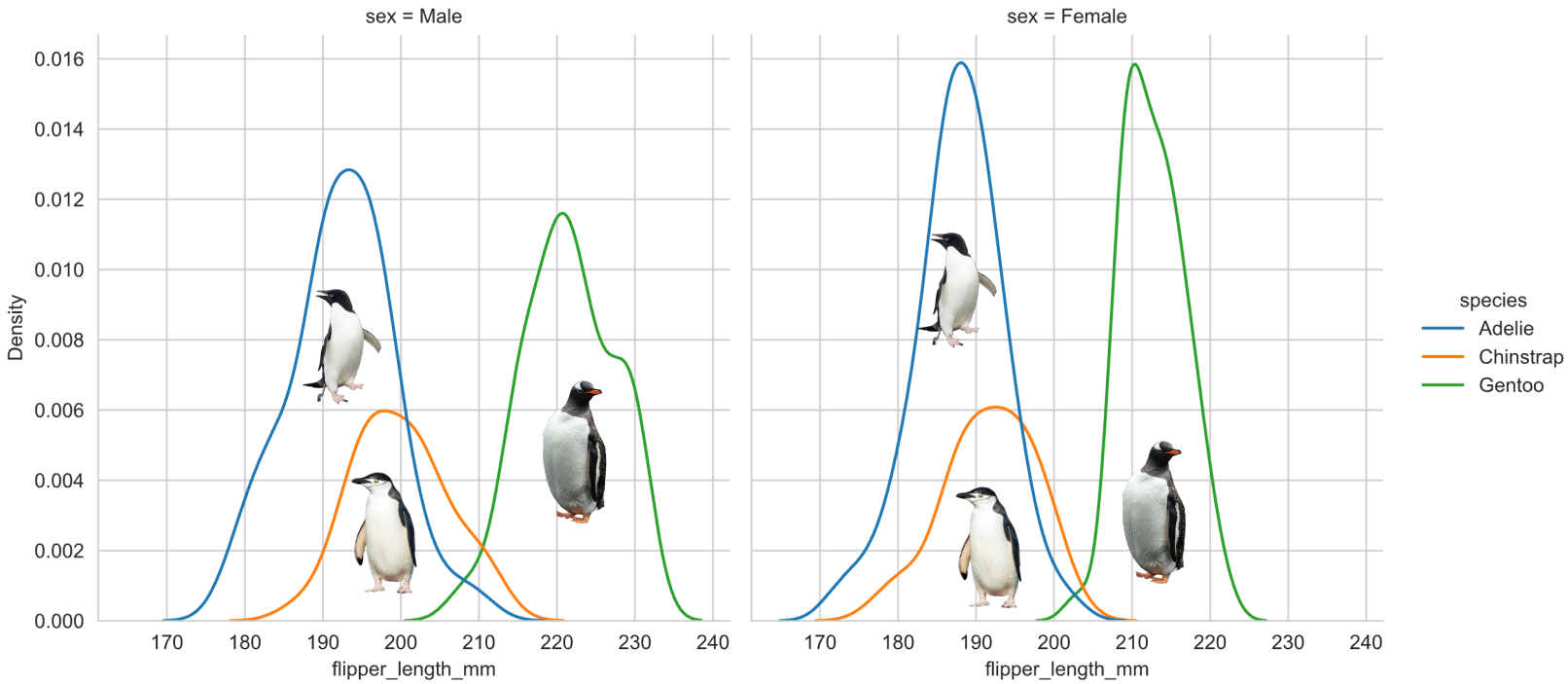
Challenge:

- 🤷‍♂️ We cannot define a deterministic function that describes the observed data
- 🗜️ How can we provide a compact representation of the sample?

Descriptive Statistics

Summarizing data

Activity: How would you describe these observed distributions?



Descriptive Statistics

Summarizing data

- **Percentiles**
  - How do the observations spread out?
- **Measures of Central Tendency**
  - What is a “typical” observation?
- **Measures of Variability**
  - How much do observations vary from the “typical” one?

Descriptive Statistics

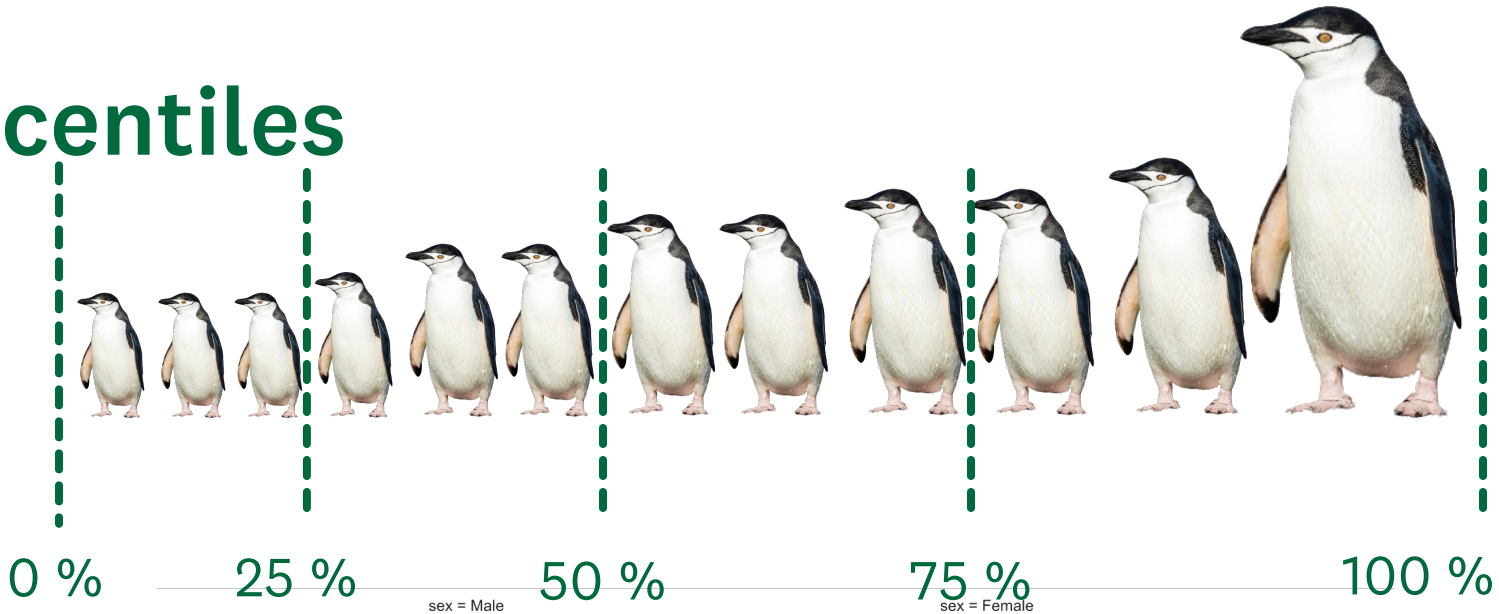
Summarizing data: Percentiles

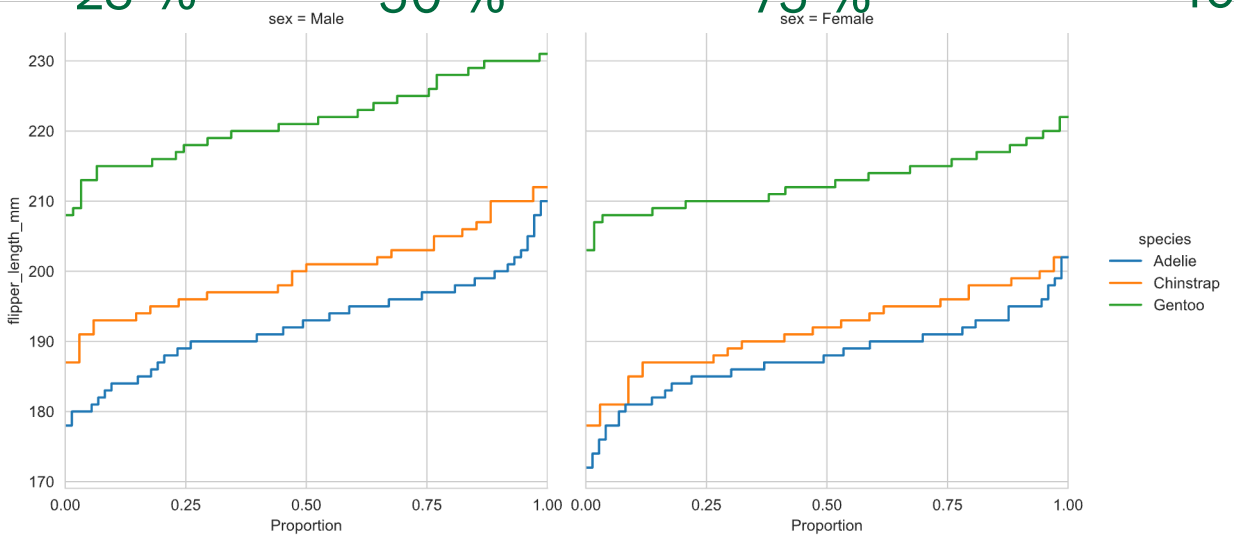
1. Sort the data ascendingly

Descriptive Statistics

Summarizing data: Percentiles

1. Sort the data ascendingly
2. Divide the data into groups
   - Four groups: quartiles
   - Five groups: quantiles
   - X groups: 1/X percentiles
3. The five magic numbers:
   - Lowest, 25 %, 50 %, 75 %, Largest

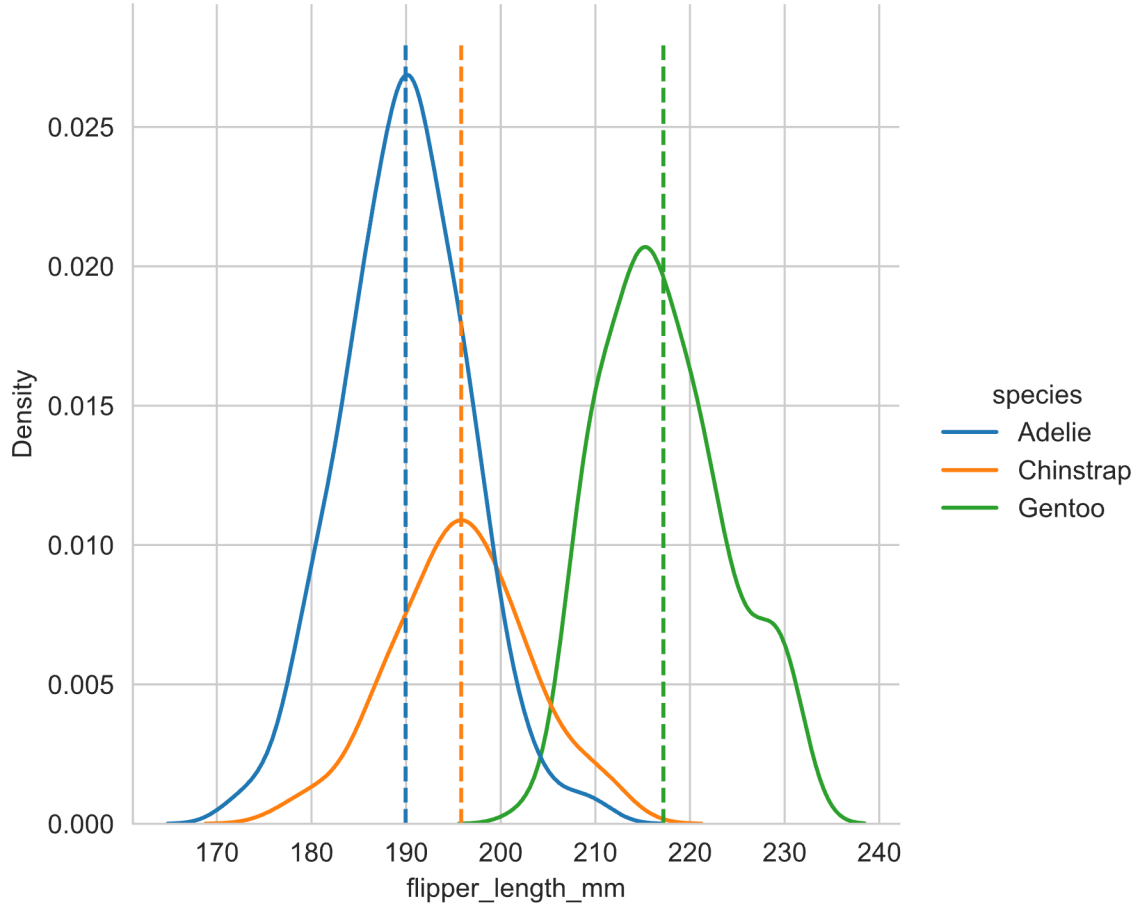




Descriptive Statistics

Summarizing data: Measures of Central Tendency

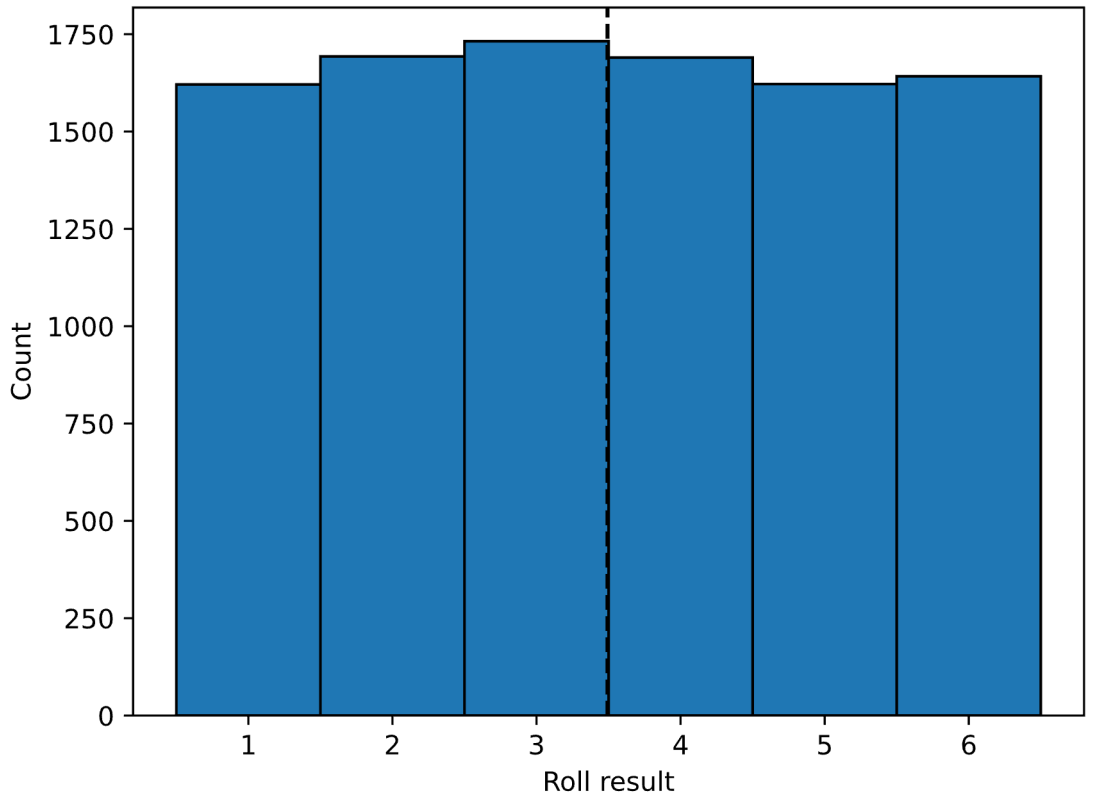
- **Mean:**
  1. Sum up all observations
  2. Divide by number of observations
- **Interpretation:**
  - Typical observation



Descriptive Statistics

Summarizing data: Measures of Central Tendency

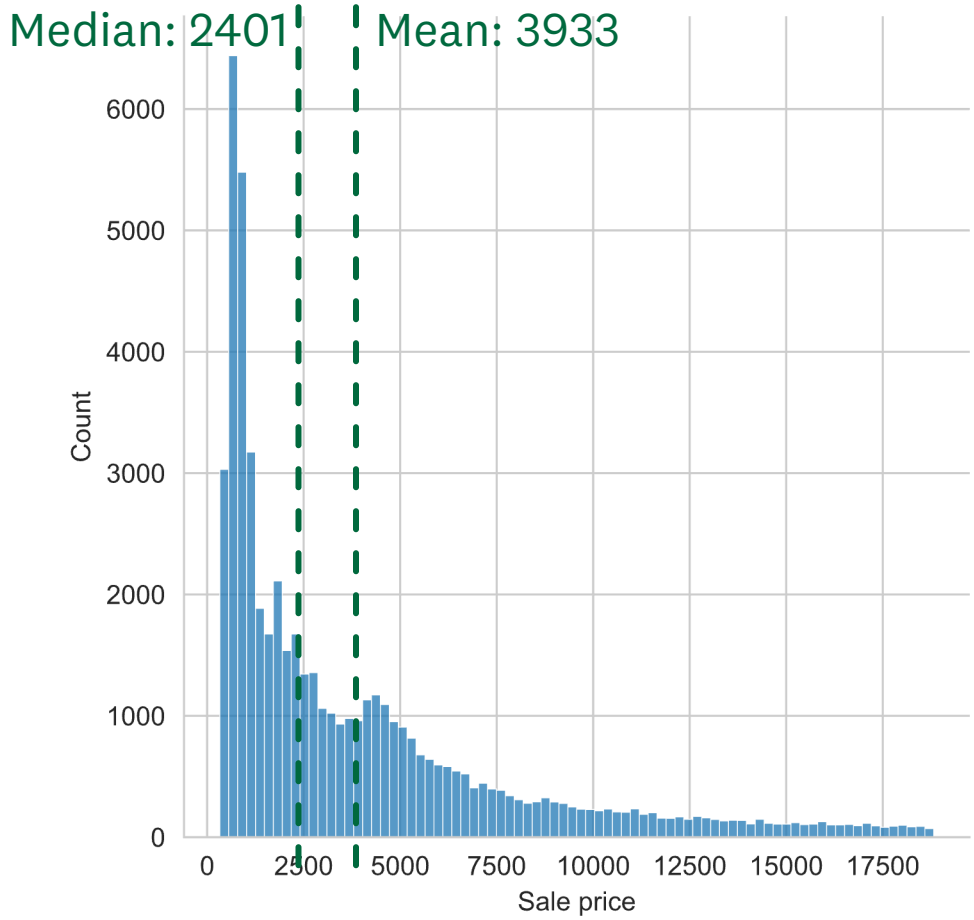
- **Mean:**
  1. Sum up all observations
  2. Divide by number of observations
- **Interpretation:**
  - Typical observation
  - Expected value: “If you expect this value, you are probably the least wrong”
  - No guarantee that this value was actually observed!



Descriptive Statistics

Summarizing data: Measures of Central Tendency

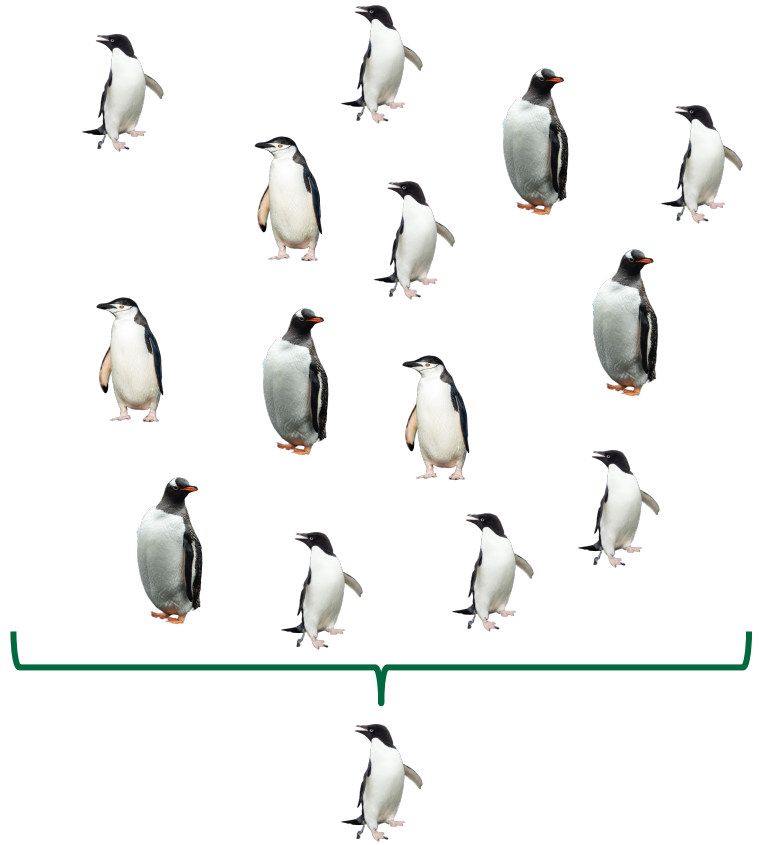
- **Median:**
  1. Sort all observations
  2. Pick the one in the middle
- **Interpretation:**
  - Typical example from the dataset
  - Half of the observations are smaller, half are bigger



Descriptive Statistics

Summarizing data: Measures of Central Tendency

- **Mode:**
  1. Count each occurrence of a value
  2. The most frequent one is the ‘mode’
- **Interpretation:**
  - The value you are most likely to observe if you pick an example from the sample at random



Descriptive Statistics

Summarizing data: Measures of Central Tendency

	Mean	Median	Mode
Interpretation:	“Least-error expectation”	Half-way point in the data	Most likely to observe this
Defined for:	Numeric data	Sortable data	Categorical

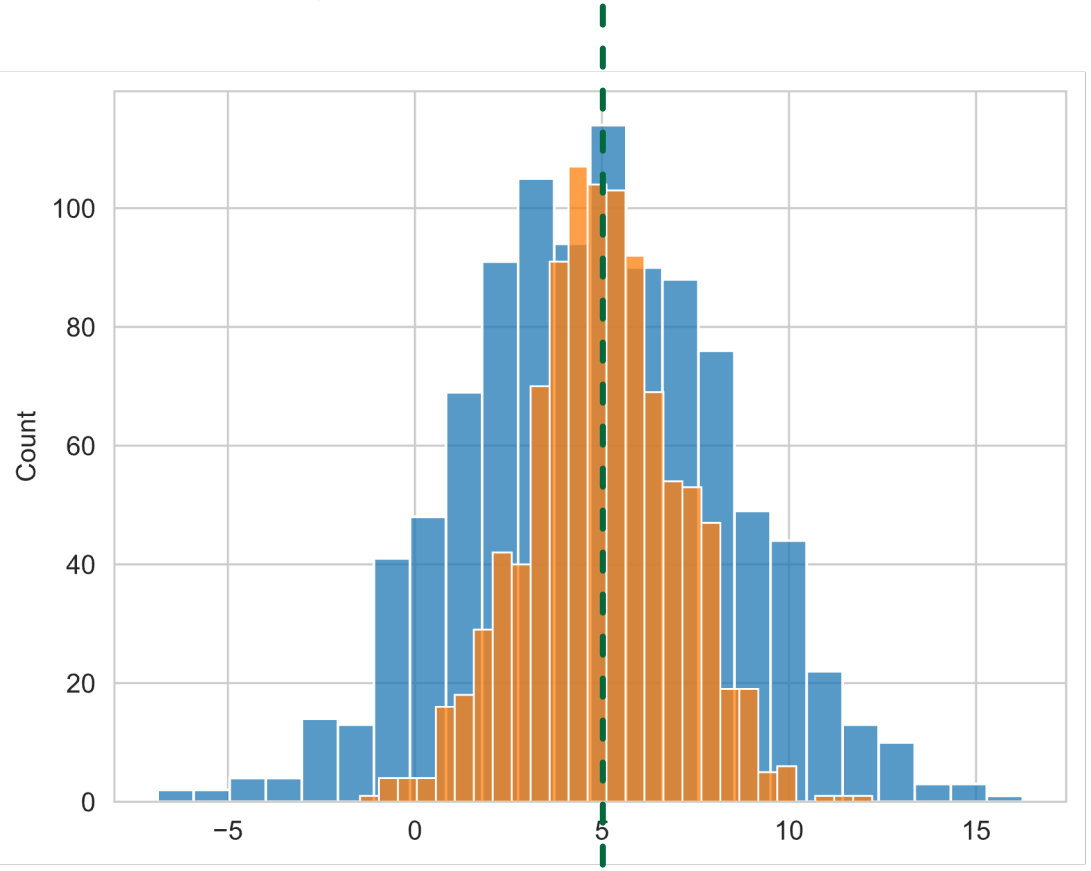
Descriptive Statistics

Summarizing data: Measures of Variability

- Two distributions with the same mean (μ = 5.0)

Question:

What makes them different?

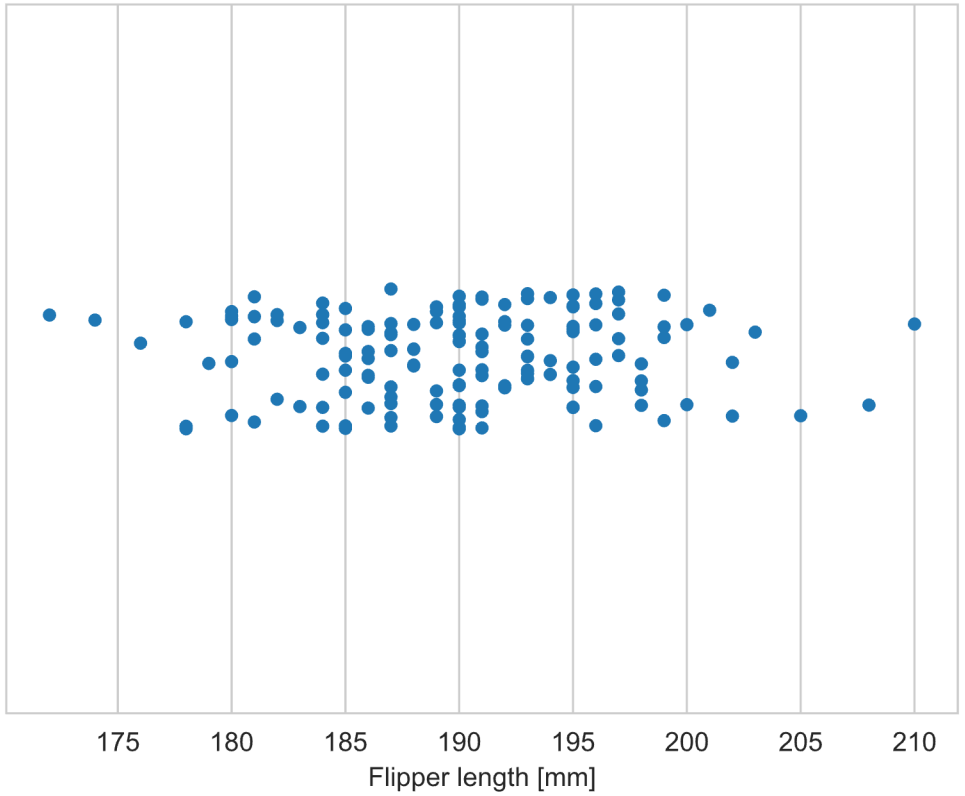


Descriptive Statistics

Summarizing data: Measures of Variability

Question:

- How would you describe the spread of the data?



Descriptive Statistics

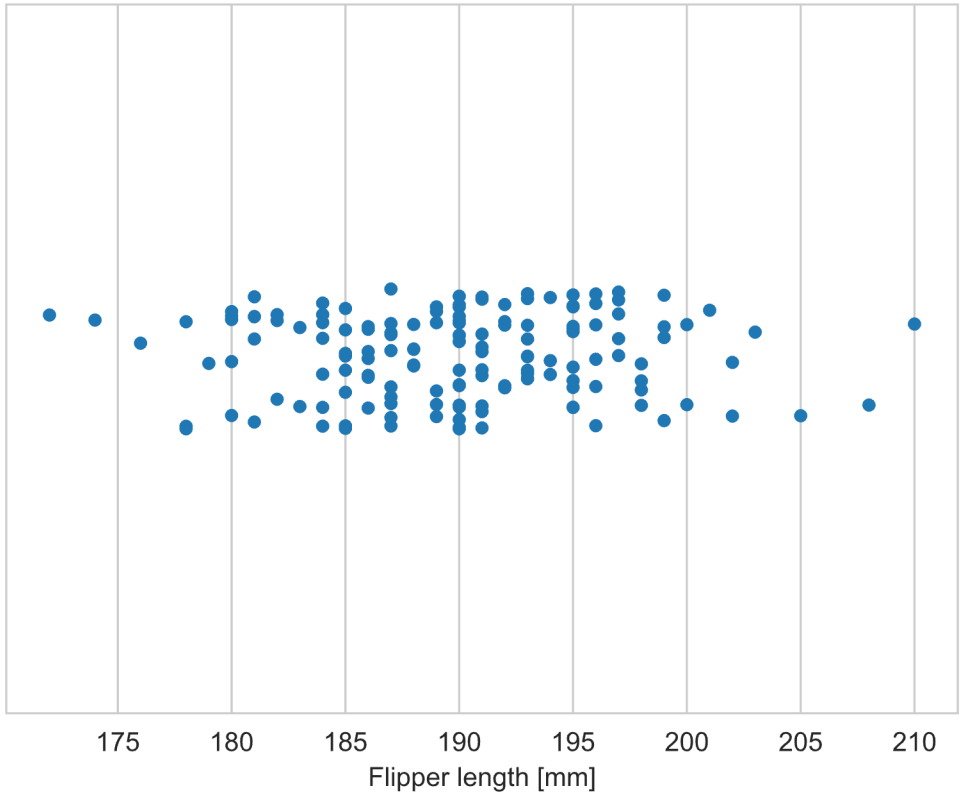
Summarizing data: Measures of Variability

Question:

- How would you describe the spread of the data?

Some options:

- Range (largest minus smallest observed value)
  - Spread of the data overall
- Interquartile range (75^{th} percentile – 25^{th} percentile)
  - Spread of the middle half of the data



Descriptive Statistics

Summarizing data: Measures of Variability

Question:

- How would you describe the spread of the data?

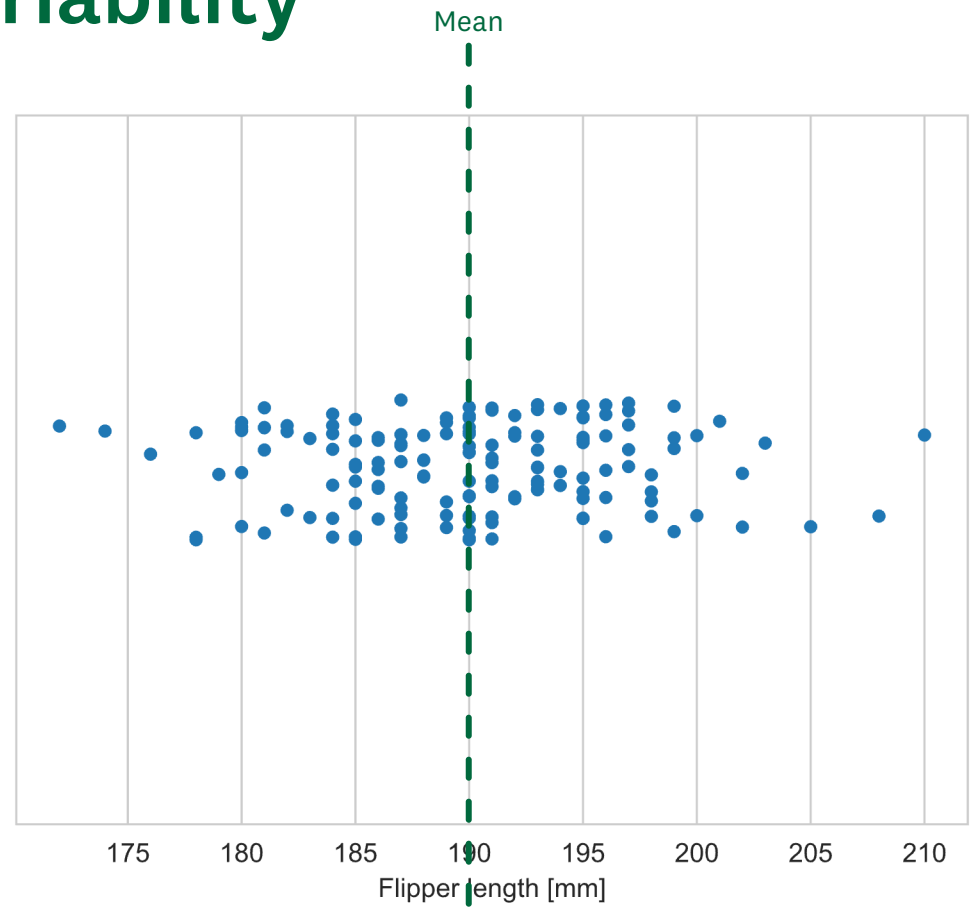
The most popular option:

Average deviation of an observation from the mean

a.k.a.

standard deviation σ

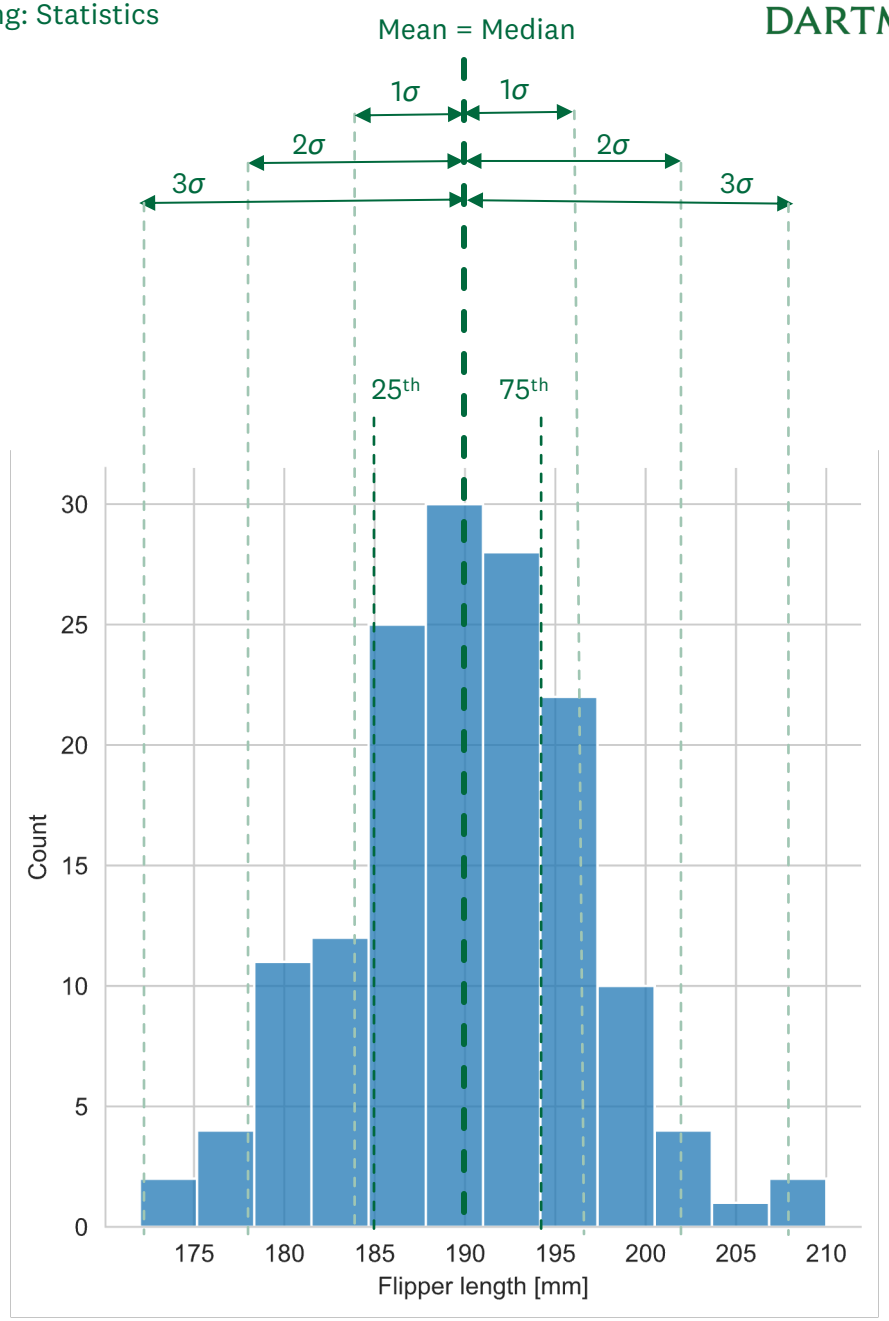
- Since we don't want positive and negative differences to cancel out, we calculate the average of the *squared* difference (the *variance* σ^2) and then take the square root again



Descriptive Statistics
Normal is special

What makes a distribution “normal”?

- The mean and the median are approximately the same
- The 25th and 75th percentile are approximately the same distance from the mean
- 68 % of data are within 1 standard deviation of the mean,
  95 % of data are within 2 standard deviations of the mean,
  99.7 % of data are within 3 standard deviations of the mean
  - The “68-95–99.7 rule”

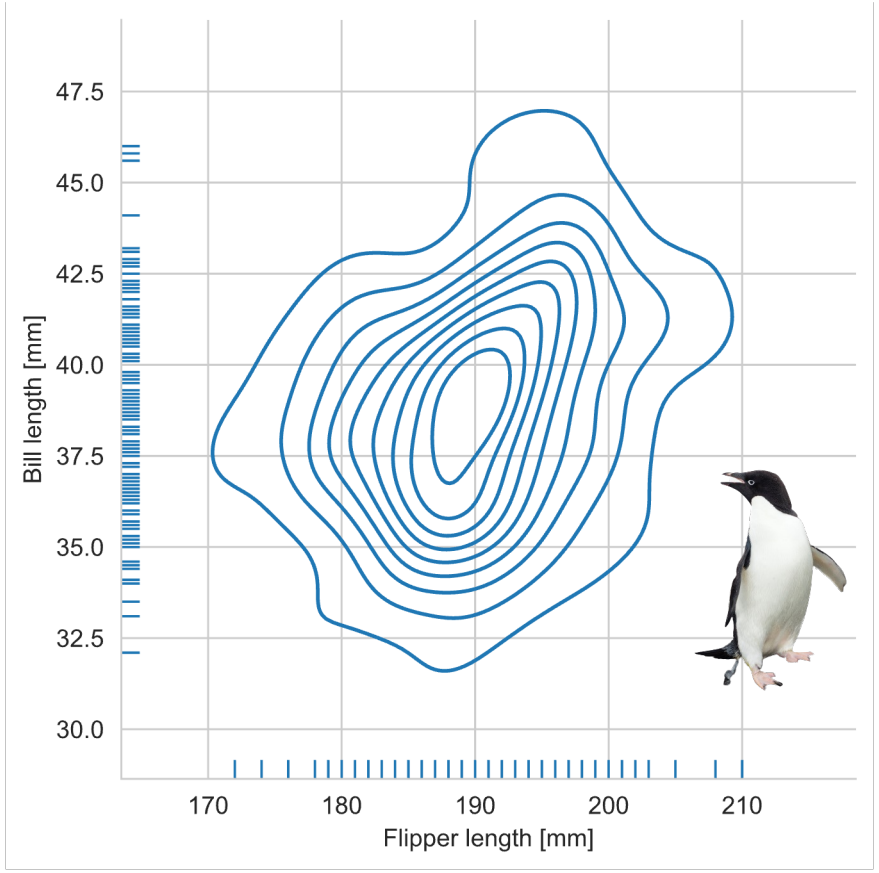


Descriptive Statistics
Multivariate distributions

- We usually describe an “object” to a machine learning model using multiple variables
- So far, we have only looked at the properties of a single variable (*univariate* statistics)
- We can also describe the distribution of multiple variables together (*multivariate* statistics)

Example:

- The distribution of flipper and bill length of Adelie penguins



Inferential Statistics

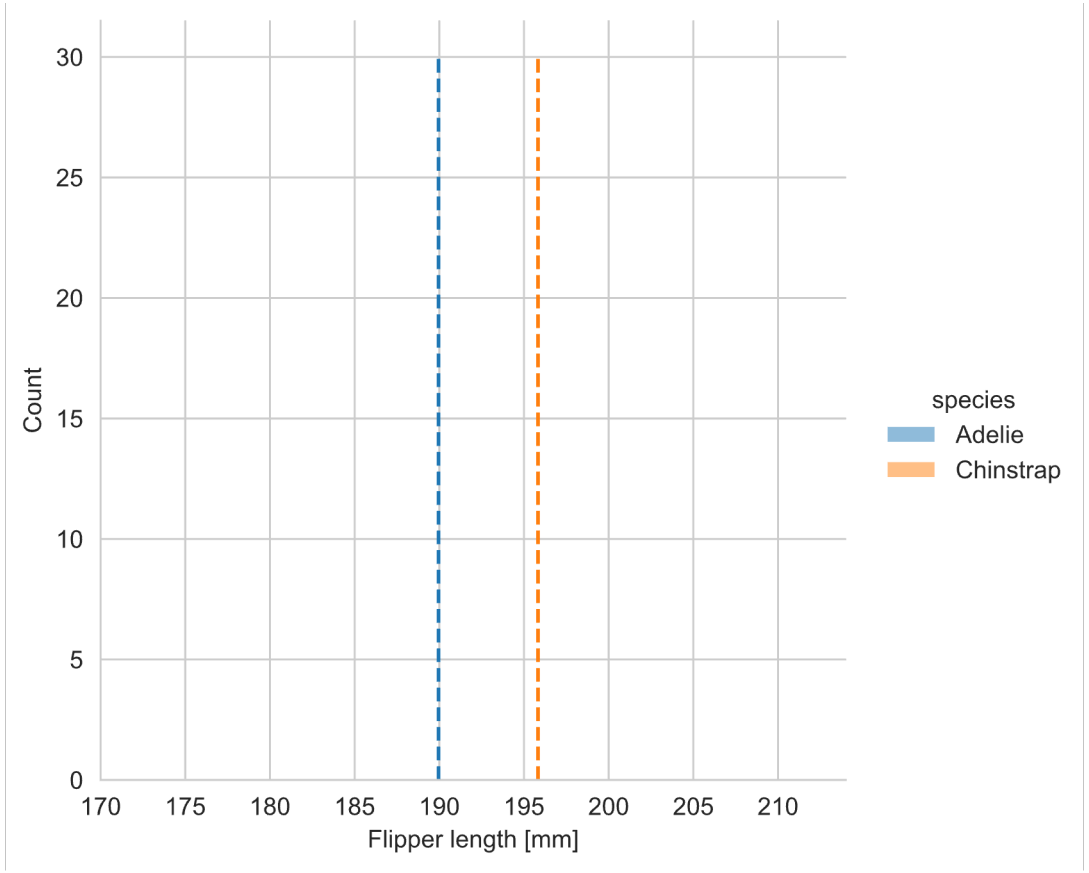
Basic Ideas

1. How can you tell if two samples are different?
   - Hypothesis testing
2. How can you make assumptions based on the data you have seen?
   - Extrapolations and predictions
   - Regression and classification

} Part of the next two sessions

Inferential Statistics
Hypothesis Testing

- You have two samples
- Are they different *on average*?



Inferential Statistics
Hypothesis Testing

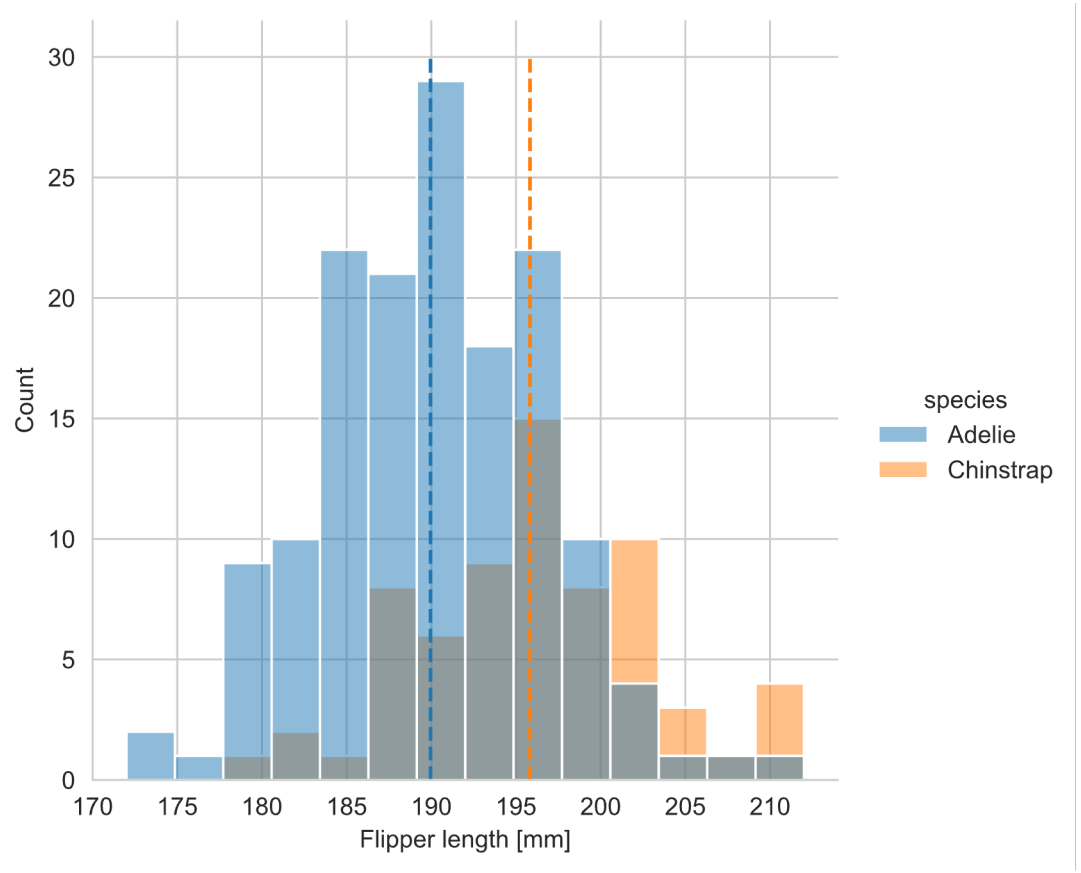
- You have two samples
- Are they different *on average*?

Hypothesis testing workflow:

1. Formulate a hypothesis

   🧐 "These samples are from the same underlying distribution and just happen to be different because of random chance"

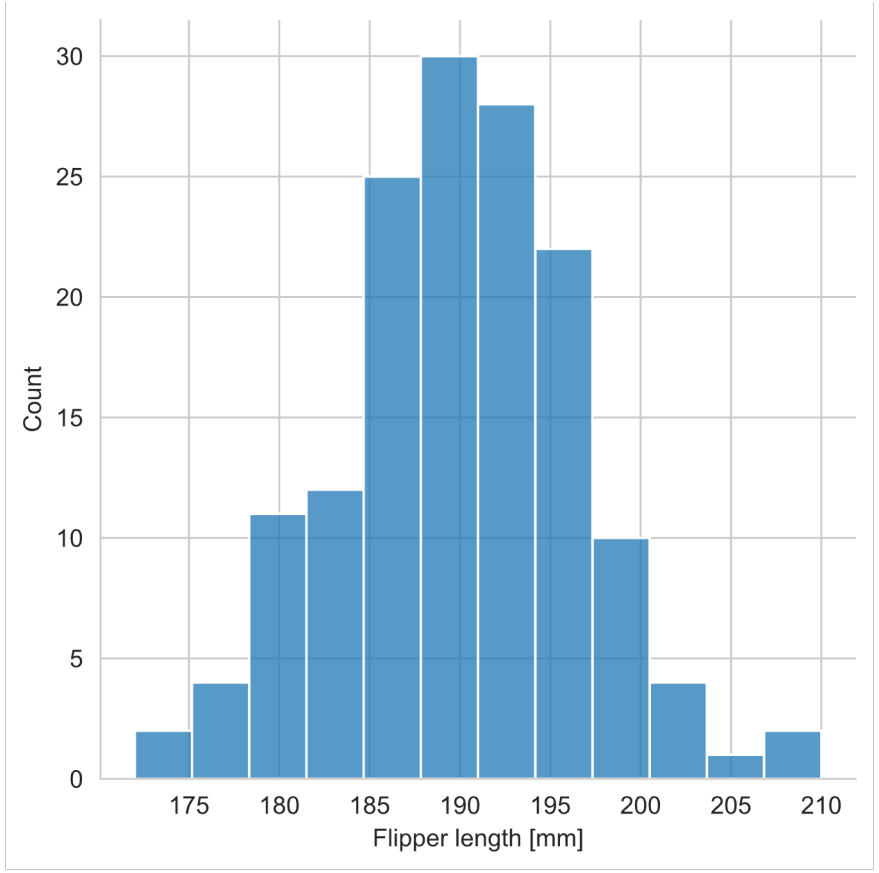
2. Calculate the difference between the means
3. Take the variance and the number of observations into account
4. Calculate how likely it is that the hypothesis is false (*p* value)
5. Compare with a threshold probability *alpha* (e.g., 5 %)
6. Call it "significant" if *p* is less than *alpha*



Inferential Statistics
Making predictions

Naïve predictor

- 🐧 Guess the flipper length of an Adelie penguin!
- Always go with the central tendency -> High bias
- Go with a random value -> High variance
- Fundamental design dilemma in machine learning: Bias-variance tradeoff



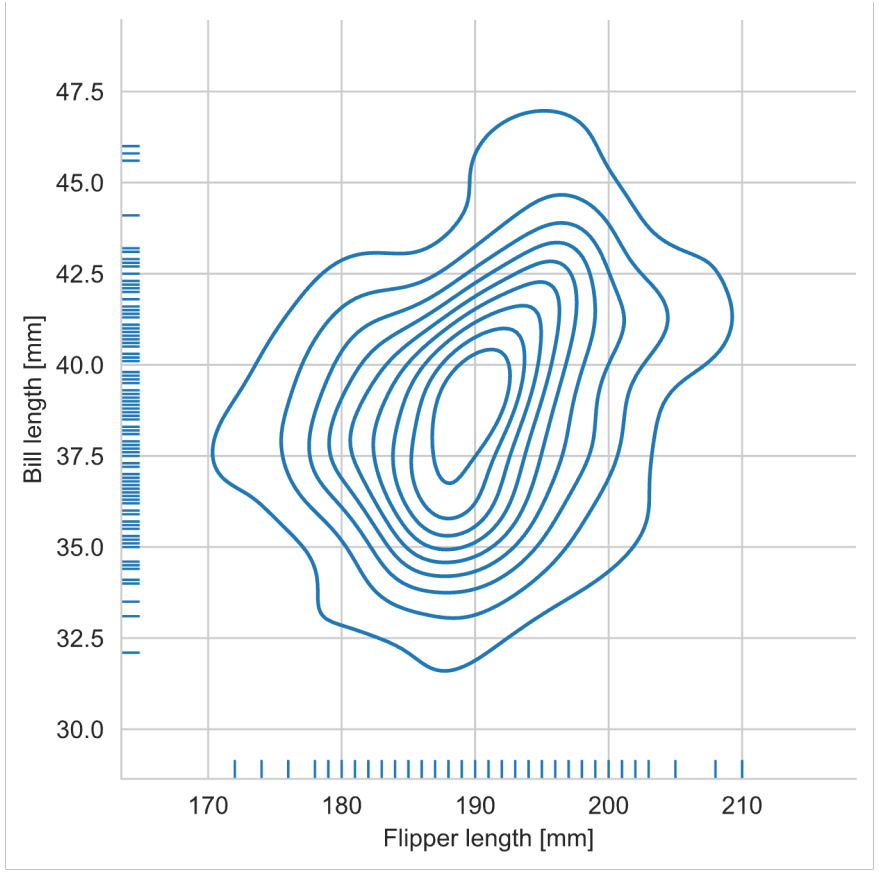
Inferential Statistics
Making predictions

Naïve predictor

- 🐧 Guess the flipper length of an Adelie penguin!
- Always go with the central tendency -> High bias
- Go with a random value -> High variance
- Fundamental design dilemma in machine learning: Bias-variance tradeoff

Conditional predictor

- 🐧 Given the bill length, guess the flipper length of an Adelie penguin!
- Take a slice through the joint distribution



Critical thinking about statistics
Visualizations

Pie charts:

- 🥧 Do the percentages add up to (about) 100 %?
- 🥧 Watch for distortions!
- 🥧 Check the sample size!

Bar charts:

- 📊 Do the units make sense when comparing bars (e.g., number of crimes versus crimes per capita)
- 📊 Is the scale appropriate or are small differences visually amplified?

Critical thinking about statistics

Biased Data

Activity:

What may cause the data or the analysis to be biased?

Some ideas:

- Is the measurement technique sufficiently precise?
- Is the observed sample representative of the population?
- Are the researchers objective? Can the analysis be “blinded”?

Critical thinking about statistics

Pitfalls in descriptive statistics

Activity:

How could a sample's statistics be misrepresented?

Some options:

- Small sample size
- Using an unsuitable measure of central tendency
- Not taking the variability into account
- Not checking for significance when comparing samples
- Normal is special!

Summary

Key take-aways

- When you cannot define a deterministic formula to explain a relationship between two variables, we need statistics
- There is no absolute certainty in statistics
- Statistics let us describe a complex object or process in simple terms
- Statistics let us extrapolate from observed and quantified data to "unseen" data

References

Upton, G., & Cook, I. (2008). *A dictionary of statistics*. Oxford University Press. DOI: 10.1093/acref/9780199541454.001.000

Rumsey, Deborah J. (2019). *Statistics Essentials*. 1st edition. Hoboken, NJ: For Dummies.