EDS 212: Day 4, Lecture 1

Essential summary statistics and exploration

August 6th, 2025

Data types

Data types

Quantitative

Numeric information

Further subdivided:

- Continuous: measured values, variable can take any value within a range.
- Discrete: variables can only take limited values (e.g. count).

Qualitative

Descriptions (usually words)

Two types:

- Nominal: order does not matter (classes or categories)
- Ordinal: order matters, but the difference between values isn't known or equal.
- Binary: only two possible outcomes: e.g. yes/no, true/false, 1/0.

Data types

Quantitative

Numeric information

- **Continuous:** measured values, variable can take *any value* within a range
- **Discrete:** variable only takes *limted values* (e.g. counts)

Qualitative

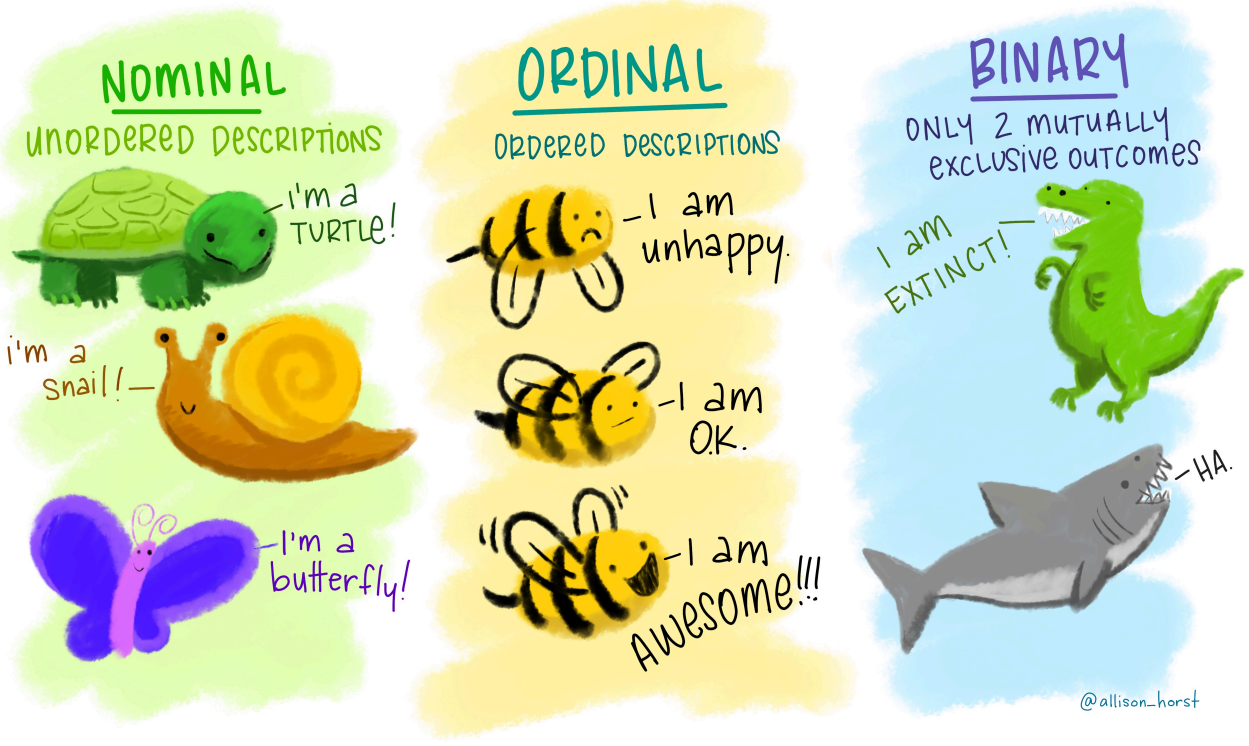
Descriptions (usually words)

- **Nominal:** order does not matter (classes or categories)
- **Ordinal:** order matters, but the difference between values isn't known or equal (e.g. Likert Scale)

- **Binary:** only two possible outcomes (yes/no, true/false, 1/0)

Quantitative data: continuous & discrete

Nominal, ordinal, binary data:



Exercise

✏️ For each of the scenarios, identify which are the dependent and independent variables, what type of variable is each one of them (quant/qual/binary/nominal/ordinal/discerete/continuous).

a. A team studying nutrient runoff wants to know whether an agricultural watershed will exceed or stay below the regulatory nitrogen threshold in a given month. They collect data on 45 watersheds. For each watershed they have recorded whether the nitrogen threshold was exceeded or not exceeded, fertilizer application rate (kg N/ha), percent agricultural land cover, and average monthly rainfall.

b. We are interested in predicting the number of air quality complaints filed in 2025 in low-income urban neighborhoods. We collect data for 60 neighborhoods over 2025. For each neighborhood we record number of complaints filed in the year, median household income, distance to the nearest industrial facility, type of nearest industrial facility (chemical plant, manufacturing, power plant), percentage of residents who are renters, and local particulate matter (PM2.5) concentration.

Exercise A: Solution

A team studying nutrient runoff wants to know whether an agricultural watershed will exceed or stay below the regulatory nitrogen threshold in a given month. They collect data on 45 watersheds. For each watershed they have recorded whether the nitrogen threshold was exceeded or not exceeded, fertilizer application rate (kg N/ha), percent agricultural land cover, and average monthly rainfall.

Dependent: nitrogen threshold exceeded:
qualitative yes / no : binary

Independent:

- fertilizer application rate : quant. , cont.
- % agr. land cover : quant. , cont.
- monthly rainfall : quant. cont.

Exercise A: Solution

A team studying nutrient runoff wants to know whether an agricultural watershed will exceed or stay below the regulatory nitrogen threshold in a given month. They collect data on 45 watersheds. For each watershed they have recorded whether the nitrogen threshold was exceeded or not exceeded, fertilizer application rate (kg N/ha), percent agricultural land cover, and average monthly rainfall.

- Dependent variable: nitrogen threshold exceeded: qualitative, binary
- Independent variables:
  - Fertilizer application rate: quantitative, continuous
  - % agricultural land cover: quantitative, continuous
  - Monthly rainfall: quantitative, continuous

Exercise B: Solution

We are interested in predicting the number of air quality complaints filed in 2025 in low-income urban neighborhoods. We collect data for 60 neighborhoods over 2025. For each neighborhood we record number of complaints filed in the year, median household income, distance to the nearest industrial facility, type of nearest industrial facility (chemical plant, manufacturing, power plant), percentage of residents who are renters, and local particulate matter (PM2.5) concentration.

- dependent : # air qual complaints
  discrete, quant.
- independent:
  - median income : quant. continuous.
  - dist. to industrial fac.: quant. cont
  - type of nearest fac.: qualitative, nominal
  - % renters : quant. cont.
  - PM 2.5 concentration: quant. cont.

Exercise B: Solution

We are interested in predicting the number of air quality complaints filed in 2025 in low-income urban neighborhoods. We collect data for 60 neighborhoods over 2025. For each neighborhood we record number of complaints filed in the year, median household income, distance to the nearest industrial facility, type of nearest industrial facility (chemical plant, manufacturing, power plant), percentage of residents who are renters, and local particulate matter (PM2.5) concentration.

- Dependent variable: number of complaints filed: quantitative, discrete
- Independent variables:
  - Median household income: quantitative, continuous
  - Distance to nearest industrial facility: quantitative, continuous
  - Type of nearest industrial facility: qualitative, nominal
  - % residents who are renters: quantitative, continuous
  - PM2.5 concentration: quantitative, continuous

Basic data visualizations

How can we describe how data are distributed?

Our starting points:

- Shape / patterns / clusters (data vizualization)
- Central tendency (mean / median)
- Spread & ~~uncertain~~nty (standard deviation / standard error / confidence interval)
- Uncertainty (SO IMPORTANT, but later)

* Update to remove std error / conf interval.

Useful data visualizations

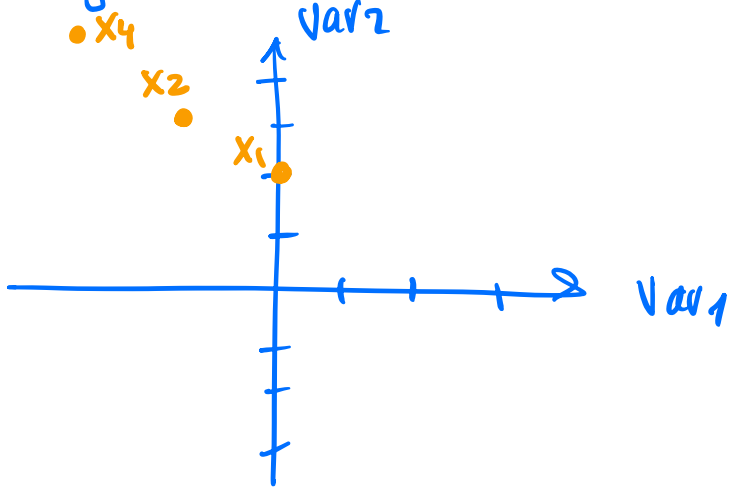
Basic ones:

- Scatterplots
- Histograms
- Boxplots

Scatterplots

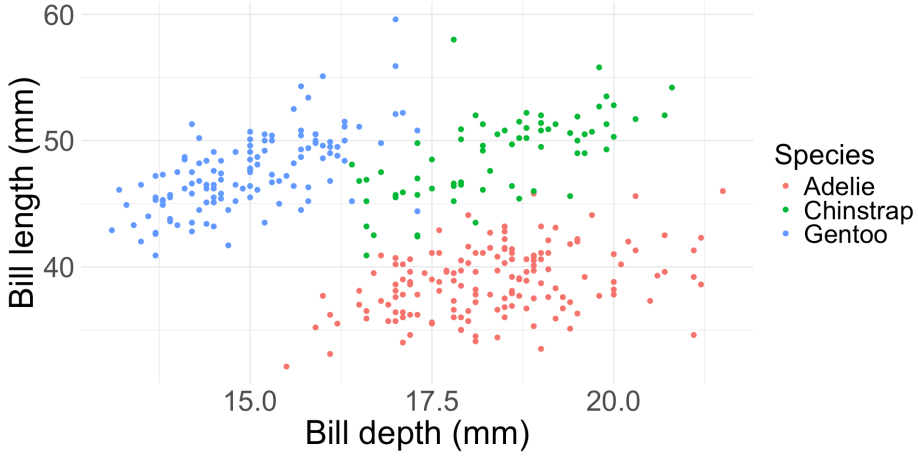
- A scatterplot shows the relationship between two numeric variables.
  - each observation becomes a point. One of the variables is the x-axis, the second is y-axis.

Observations	var_1	var_2	
x_1	0	2	$\rightarrow (0, 2)$
x_2	-1	3	$\rightarrow (-1, 3)$
x_3	-1	4	$\rightarrow (-1, 4)$



Scatterplots

A scatterplot shows the relationship between **two numeric variables**: each observation becomes a point, positioned using its value on the x-axis and its value on the y-axis. A third (categorical) variable can be mapped to color or shape.



Always, always, always look at your data. It is the only way to make a responsible decision about an appropriate type of analysis.

Histogram

- A histogram shows the distribution of a single numeric variable.
  - It groups values into bins by counting observations.

To construct one:

1. Divide the range of the data into equal-width bins.
2. Count how many observations fall into each bin
3. Draw a bar over each bin, with height equal to that count.

Histogram

A histogram shows the **distribution** of a single numeric variable, by grouping values into bins and counting observations.

To construct one:

1. Divide the range of the data into equal-width bins
2. Count how many observations fall into each bin
3. Draw a bar over each bin, with height equal to that count

Histogram example

Twelve trees are measured in a forest plot, giving these diameter at breast height (DBH) values, in cm:

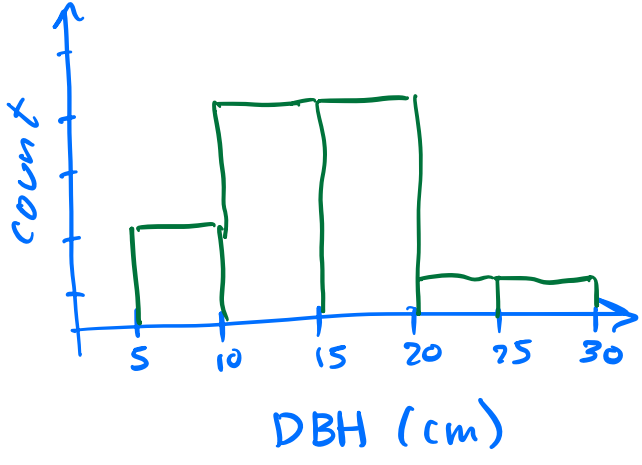
8, 9, 11, 12, 12, 14, 15, 15, 16, 19, 22, 27 ← observations.

do not include 10

include 5 in your interval

Using bins of width 5 cm:

Bin (cm)	Count
[5, 10)	2
[10, 15)	4
[15, 20)	4
[20, 25)	1
[25, 30)	1



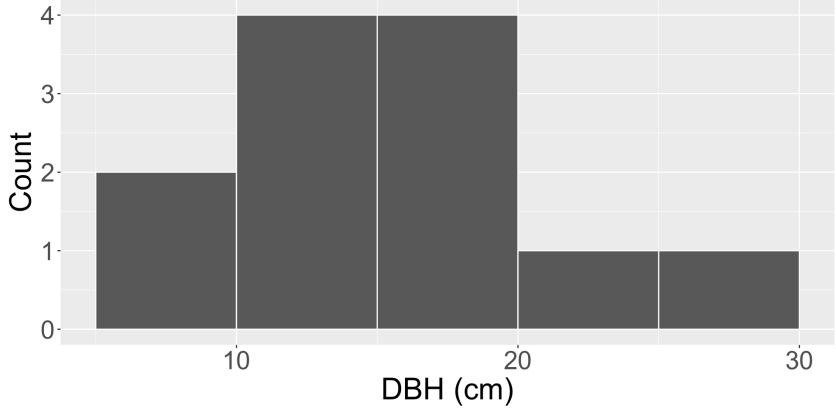
Histogram example

Twelve trees are measured in a forest plot, giving these diameter at breast height (DBH) values, in cm:

8, 9, 11, 12, 12, 14, 15, 15, 16, 19, 22, 27

Using bins of width 5 cm:

Bin (cm)	Count
[0, 5)	0
[5, 10)	2
[10, 15)	4
[15, 20)	4
[20, 25)	1
[25, 30)	1

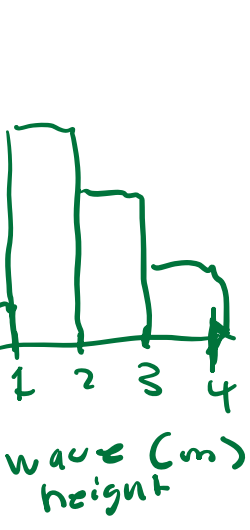


Exercise

✏️ Ten storm events produced these wave heights, in meters. Using bins of width 1 m (starting at 0), tally the counts and sketch the resulting histogram.

0.8, 1.1, 1.2, 1.5, 1.9, 2.0, 2.4, 2.6, 3.1, 3.6

Range	#
[0, 1)	1
[1, 2)	4
[2, 3)	3
[3, 4)	2

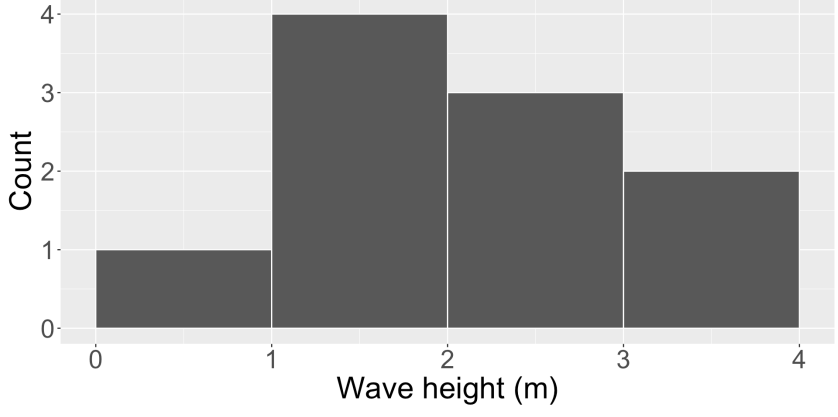


Exercise: solution

✏️ Ten storm events produced these wave heights, in meters. Using bins of width 1 m (starting at 0), tally the counts and sketch the resulting histogram.

0.8, 1.1, 1.2, 1.5, 1.9, 2.0, 2.4, 2.6, 3.1, 3.6

Bin (m)	Count
[0, 1)	1
[1, 2)	4
[2, 3)	3
[3, 4)	2



Quartiles

Quartiles divide an ordered dataset into four equal parts.

- Q_1 (1st quartile): 25% of values fall below this point
- Q_2 (2nd quartile): the **median**: 50% of values fall below this point
- Q_3 (3rd quartile): 75% of values fall below this point

even? 4 5 6 7

One simple way to find them:

1. Order the data and find the median (Q_2): this splits the data into a lower half and an upper half. If there's an even number of elements, take the average between the two "middle ones".
2. Q_1 = median of the lower half
3. Q_3 = median of the upper half

* Create blank slide

Quartiles

- Q_1 (1st quartile): 25% of values fall below this point
- Q_2 (2nd quartile): the **median**: 50% of values fall below this point
- Q_3 (3rd quartile): 75% of values fall below this point

✏️ Find Q_1, Q_2, and Q_3 for the values 4, 16, 8, 40, 2, 14, 6, 10, 12

Ordered: 2, 4, 6, 8, 10, 12, 14, 16, 40

Q_1 Q_3

- Q_2 (median): $Q_2 = 10$
- $Q_1 = 5$
- $Q_3 = 15$.

Quartiles: solution

- Q_1 (1st quartile): 25% of values fall below this point
- Q_2 (2nd quartile): the **median**: 50% of values fall below this point
- Q_3 (3rd quartile): 75% of values fall below this point

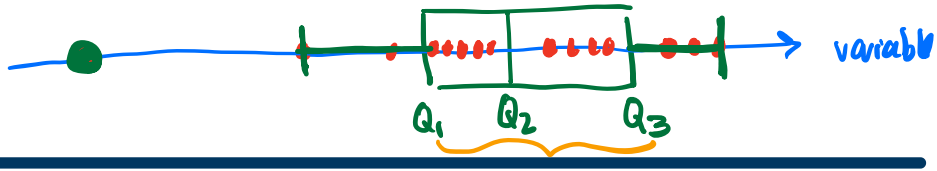
💡 Let's see a solution!

Find $Q_1, Q_2,$ and Q_3 for the values 4, 16, 8, 40, 2, 14, 6, 10, 12

Ordered: 2, 4, 6, 8, 10, 12, 14, 16, 40

- Q_2 (median): 10
- Lower half: 2, 4, 6, 8: an **even** number of values, so Q_1 is the average of the middle two: $Q_1 = \frac{4+6}{2} = 5$
- Upper half: 12, 14, 16, 40: again an even number of values: $Q_3 = \frac{14+16}{2} = 15$

Boxplot



How it's drawn:

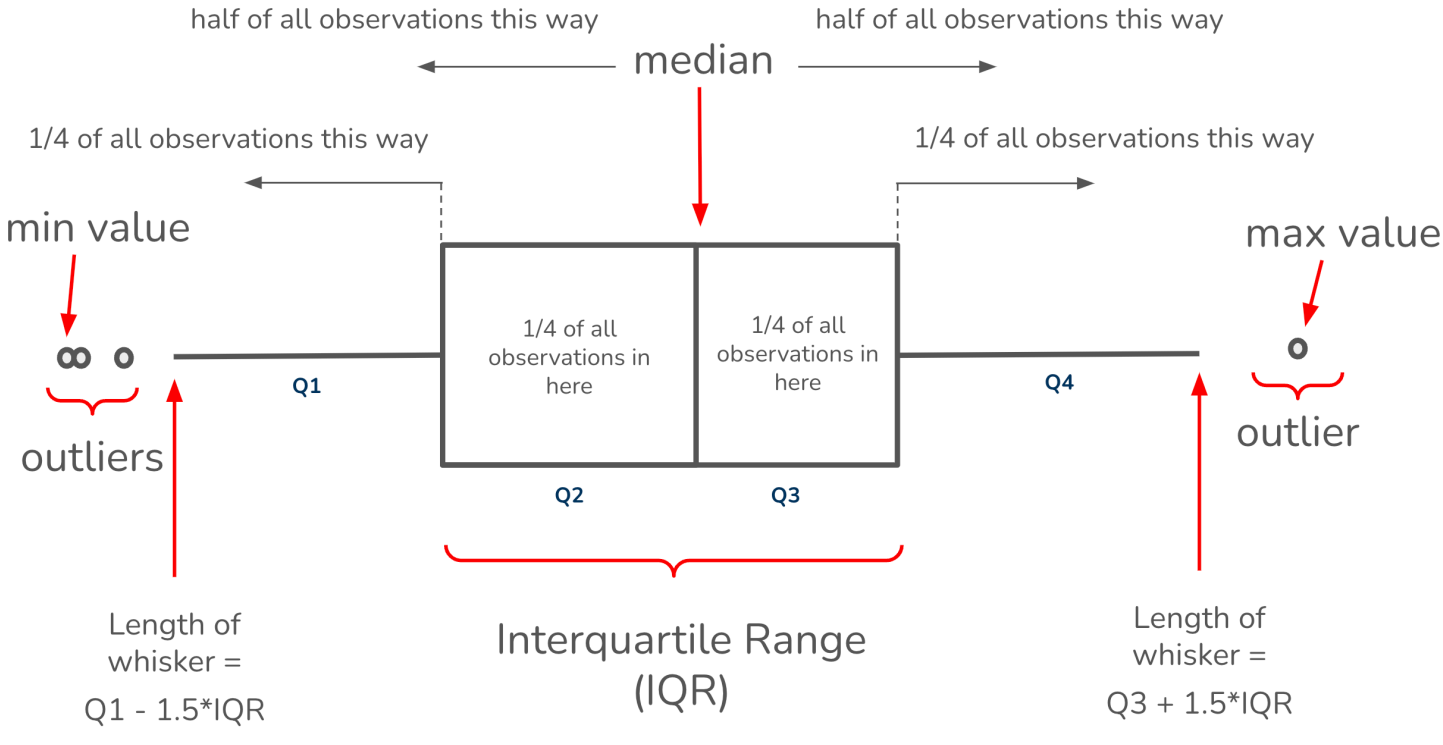
1. Draw a box from Q_1 to Q_3: this box spans the **interquartile range**:

$$IQR = Q_3 - Q_1$$

2. Draw a line inside the box at the median (Q_2)
3. Draw whiskers extending to the furthest observation within $1.5 \times IQR$ of the box
4. Plot any observation beyond the whiskers individually, as a dot (an **outlier**)

* Add space to draw boxplot.

Boxplot


half of all observations this way
median
half of all observations this way
1/4 of all observations this way
1/4 of all observations this way
min value
max value
1/4 of all observations in here
1/4 of all observations in here
Q1
Q2
Q3
Q4
outliers
outlier
Length of whisker = Q1 - 1.5*IQR
Interquartile Range (IQR)
Length of whisker = Q3 + 1.5*IQR

Exercise: draw a box plot

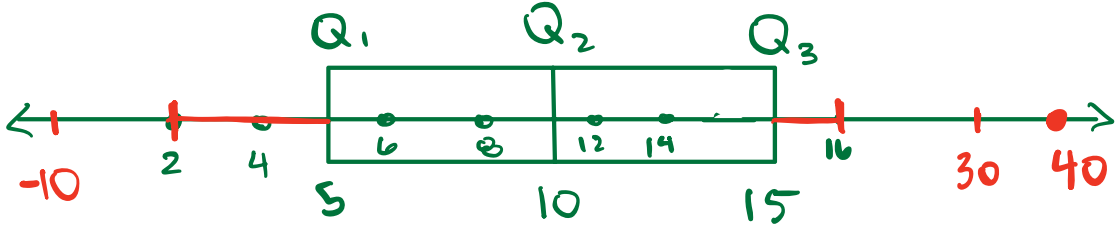
How it's drawn:

1. Draw a box from Q_1 to Q_3: this box spans the **interquartile range**:

$$IQR = Q_3 - Q_1$$

2. Draw a line inside the box at the median (Q_2)
3. Draw whiskers extending to the furthest observation within $1.5 \times IQR$ of the box
4. Plot any observation beyond the whiskers individually, as a dot (an **outlier**)

✏️ Using the quartiles you just found for 2, 4, 6, 8, 10, 12, 14, 16, 40 ($Q_1 = 5$, $Q_2 = 10$, $Q_3 = 15$), draw the box plot by hand.



$$1.5 \cdot IQR = 1.5(Q_3 - Q_1)$$

$$= 1.5(15 - 5)$$

$$= 1.5(10)$$

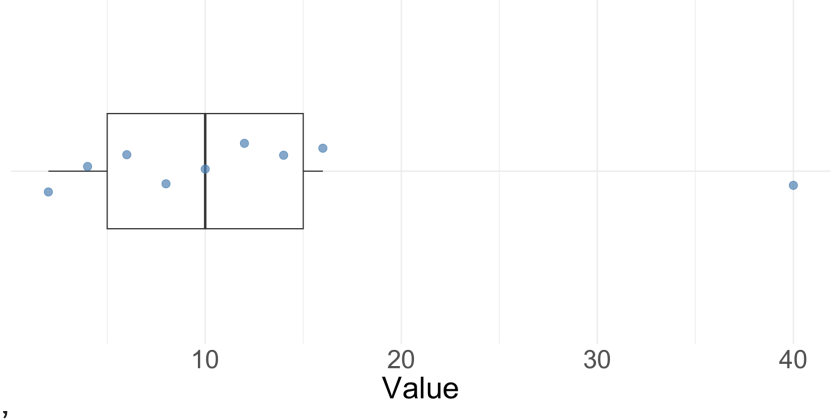
$$= 15$$

* Add blank slide for drawing.

Exercise: solution

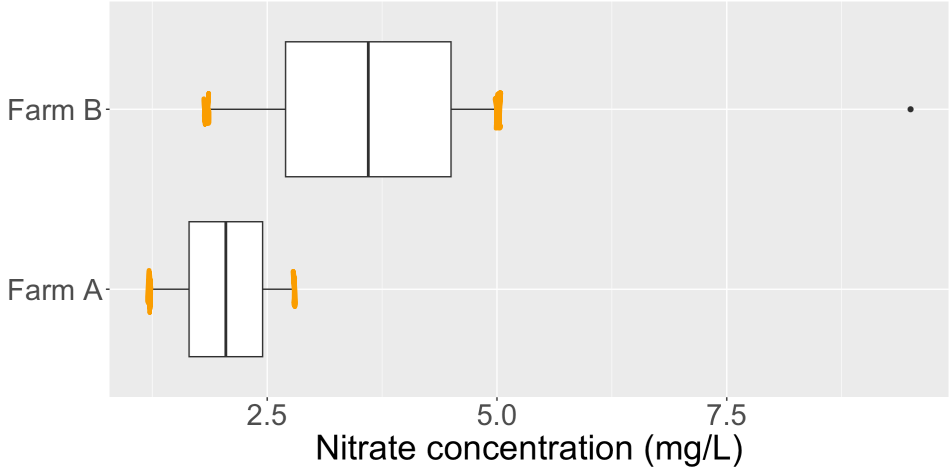
💡 Let's see a solution!

- Q_1: 5
- Median (Q_2): 10
- Q_3: 15
- $IQR = 15 - 5 = 10$
- $Q_3 + 1.5 \times IQR = 15 + 15$:
- 40 is beyond that value, so it's an **outlier**: the whisker stops at 16, the last value within $1.5 \times IQR$ of the box



Exercise: interpreting boxplots

✏️ Two farms are monitored for nitrate concentration (mg/L) in their downstream runoff. Compare the two distributions below.



- Which farm's runoff has the higher median nitrate concentration?
- Which farm's readings are more spread out?
- Is there anything unusual in either distribution?

Exercise: solution

💡 Let's discuss!

- **Median:** Farm B's median (~3.6 mg/L) is noticeably higher than Farm A's (~2.05 mg/L)
- **Spread:** Farm B's box and whiskers cover a much wider range, so its runoff is more variable
- **Outlier:** Farm B has a high outlier at 9.5 mg/L, flagged because it falls more than 1.5×IQR above the 3rd quartile. This might be worth investigating.

Let's take a 5 minute break

image: Flaticon.com

Summarizing data numerically

Summarizing data numerically

- Central tendency
- Variance and standard deviation
- Standard error
- Confidence interval

Median

Middle value when all observations are arranged in order. If you have an even number of values, the median is calculated as the average of the middle two values. E.g. $median\ of\ 3, 7, 17 = 7$

Pros:

- Less susceptible to skew and outliers

Cons:

- Doesn't take into account the magnitude of all values

Mean

Average value of sample observations, calculated by summing all observation values and dividing by the number of observations. E.g.

$$mean\ of\ 3, 7, 17 = \frac{3+7+17}{3} = 9$$

Another way: Observations: $x_1, x_2, \ldots, x_n$, then $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum x_i$.

Pros:

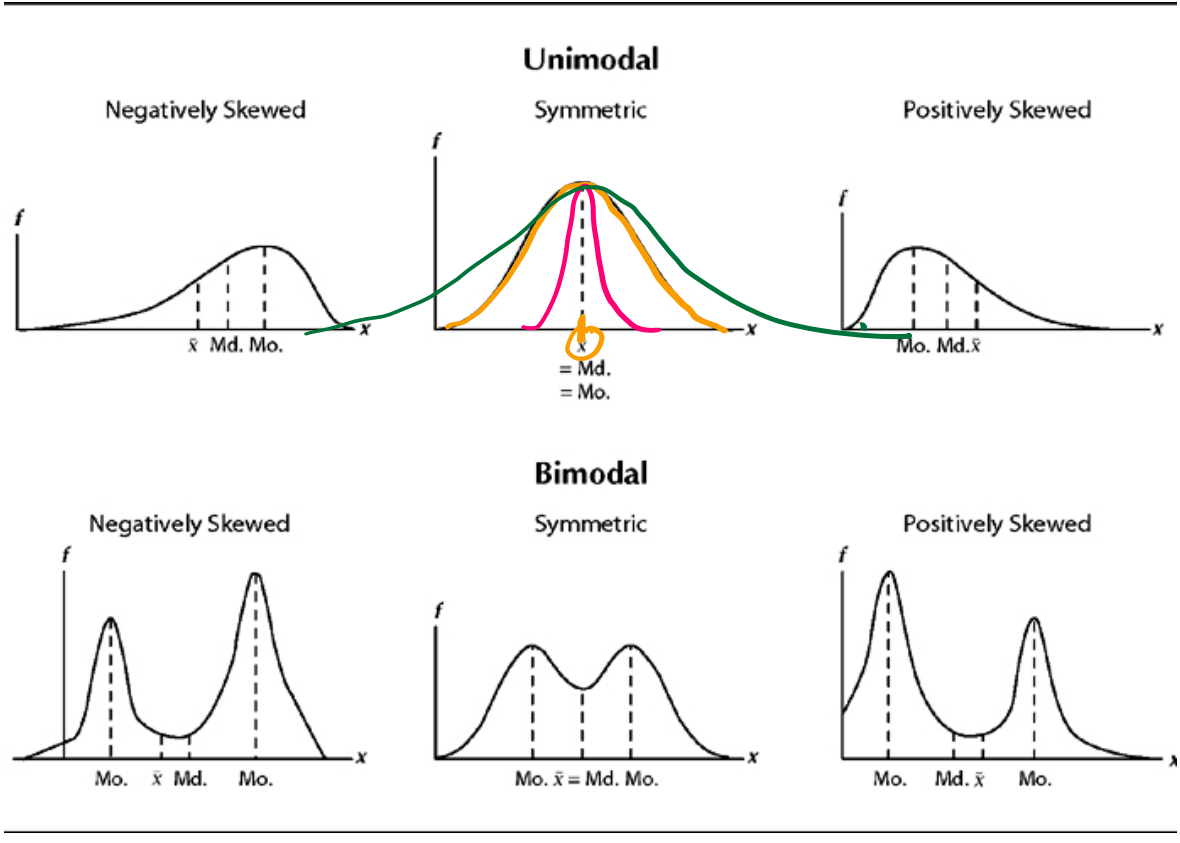
- Average value is often useful metric
- Commonly reported

Cons:

- susceptible to outliers and skew.
- subject to misinterpretation as the "most likely value"

Mode: the value that appears most frequently.

* Add Mode slide



Variance and standard deviation

Both are measures of **data spread**.

Variance

Reported in units of measurement squared

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$$

Standard deviation

Reported in units of measurement

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$$

Variance

- measures the dispersion of the data relative to the mean

observations : $x_1, x_2, \dots, x_n$

# of observations : n

mean $= \bar{x}$

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}.$$

Variance

Variance: measures the mathematical dispersion of data relative to the mean

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$$

Where S^2 is the sample variance, x_i is a sample observation value, $\bar{x}$ is the sample mean, and n is the number of observations.

* Check the n - 1 in division.

Exercise

Given these data: $2, 4, 4, 6, 9$, calculate their variance. Remember:

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$$

$\bar{x} = 5 = \frac{25}{5}$

$$s^2 = \frac{(2-5)^2 + (4-5)^2 + (4-5)^2 + (6-5)^2 + (9-5)^2}{4}$$

$$= (9 + 1 + 1 + 1 + 16) \cdot \frac{1}{4} = 28 \cdot \frac{1}{4} = 7 \text{ units}^2$$

Calculate variance by hand (1/2)

Given these data: $2, 4, 4, 6, 9$

1. Calculate the mean:

$$mean = \frac{2 + 4 + 4 + 6 + 9}{5} = \frac{25}{5} = 5$$

2. Subtract the mean from each data point and square the result:

$$(2 - 5)^2 = (-3)^2 = 9$$

$$(4 - 5)^2 = (-1)^2 = 1$$

$$(4 - 5)^2 = (-1)^2 = 1$$

$$(6 - 5)^2 = (1)^2 = 1$$

$$(9 - 5)^2 = (4)^2 = 16$$

Calculate variance by hand (2/2)

Given these data: $2, 4, 4, 6, 9$

3. Sum the squared differences:

$$9 + 1 + 1 + 1 + 16 = 28$$

4. Divide by the number of data points minus 1 (n - 1)

$$Variance = \frac{28}{5 - 1} = 7$$

If Move into single slide.

Standard deviation

Also a measure of data spread, calculated by taking the square root of the variance.

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}}$$

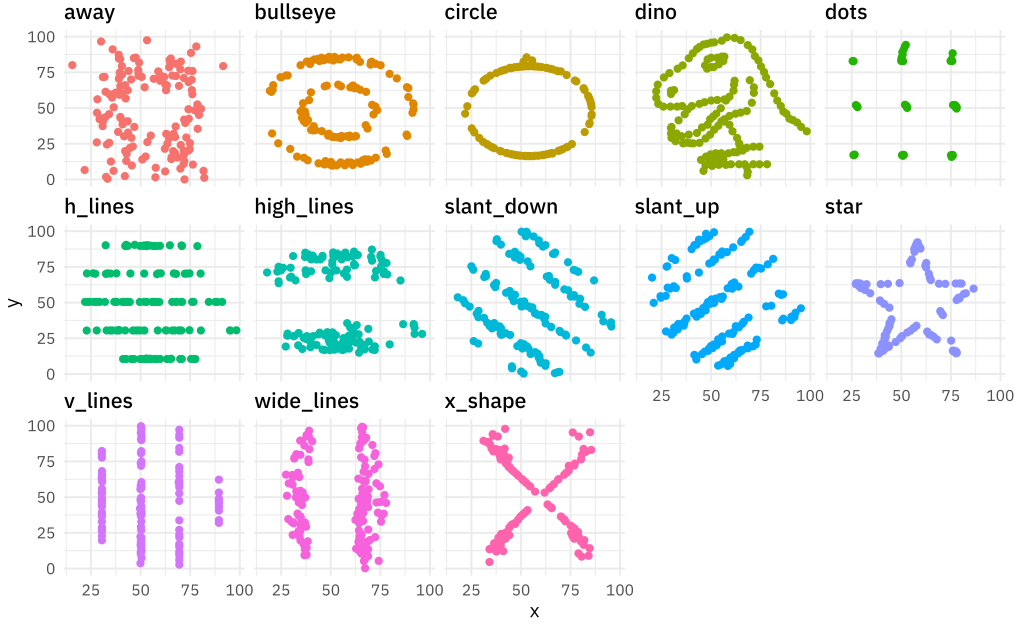
Standard deviation is expressed in the same unit of measurement as the data, and therefore can be easier to interpret – for example it's more intuitive to report that a group of people's heights has a standard deviation of 3 inches, and less intuitive to report a variance of 9 square inches.

* make this into a separate slide.

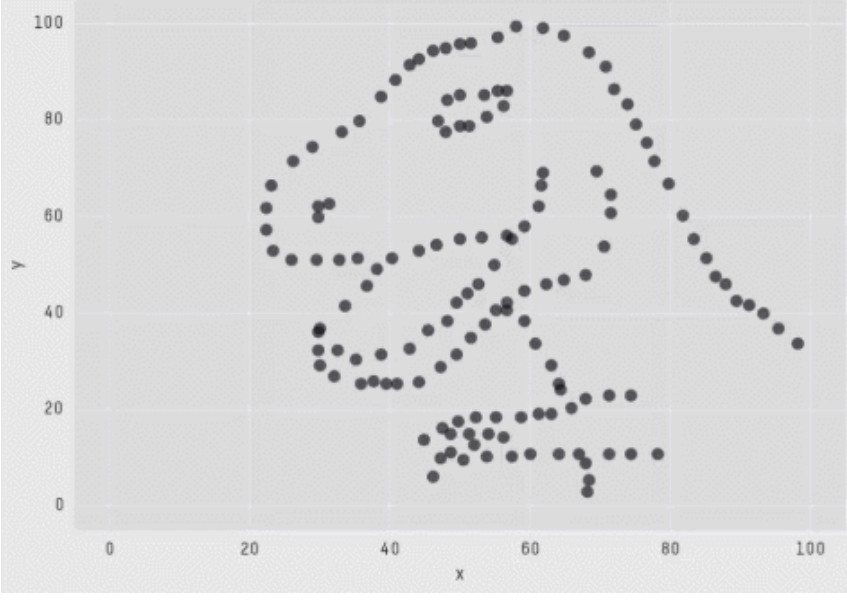
The best way to describe the distribution of the data is to present the data itself.

Beware summary statistics alone . . .

Meet the Datasaurus Dozen



Same summary statistics, different distributions



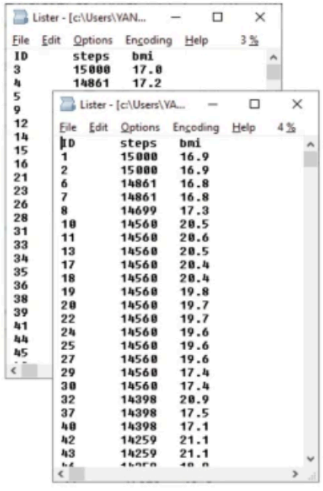
X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

selective attention test

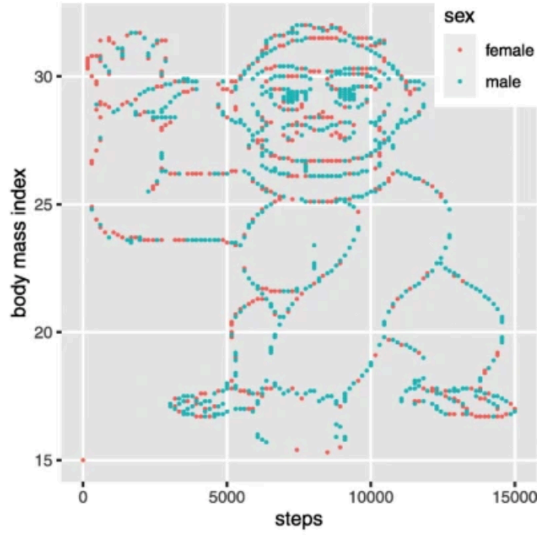
Daniel Simons

Watch on

a



b



c

	Gorilla not discovered	Gorilla discovered
Hypothesis-focused	14	5
Hypothesis-free	5	9

* get reference

Confidence intervals

*move this to next set of slides.

Confidence interval : Range of values (based on a sample) that : if we were to take multiple samples from the population and calculate the confidence interval for a statistic from each sample, then the such an interval would contain the true population parameter X percent of time (a certain percentage of time).

* Blank slide ?

Confidence interval

Confidence interval: a range of values (based on a sample) that, if we were to take multiple samples from the population and calculate the confidence interval from each, would contain the true population parameter X percent of the time.

Confidence interval

Confidence interval: a range of values (based on a sample) that, if we were to take multiple samples from the population and calculate the confidence interval from each, would contain the true population parameter X percent of the time.

What it's NOT:

"There is a 95% chance that the true population parameter is inside the confidence interval."

Confidence interval example

Mean shark length is 8.42 ± 3.55 ft (mean ± standard deviation), with a 95% confidence interval of [6.45, 10.39 ft] (n = 15).

What this **DOES NOT** mean: There is a 95% chance that the true population mean length is between 6.45 and 10.39 feet.

The true population mean is a *fixed value* and does not change – the CI either contains this true mean or it does not. There is no probability involved.

Confidence interval example

Mean shark length is 8.42 ± 3.55 ft (mean ± standard deviation), with a 95% confidence interval of [6.45, 10.39 ft] (n = 15).

What this **DOES NOT** mean: There is a 95% chance that the true population mean length is between 6.45 and 10.39 feet.

The true population mean is a *fixed value* and does not change – the CI either contains this true mean or it does not. There is no probability involved.

What this **DOES** mean: If we took a bunch of sets of samples from the population (all n = 15), then 95% of the time, the calculated mean would fall within this range.

This statement correctly describes the frequency with which we would expect CIs to capture the mean over many samples.

Communicating data summaries

- Show as much as you can for the audience you're presenting to
- Summary statistics are often useful, but are a small part of the whole data story
- Uncertainty is important! How can we responsibly communicate it?
- All summaries are strongest when accompanied by additional data communication

Artwork by Allison Horst

Wrapping up...

What we covered today

- 🧩 Data types
- 📊 Data visualization (histograms, box plots, scatterplots)
- 📐 Quartiles & the interquartile range
- 📈 Central tendency (mean & median)
- 📏 Variance & standard deviation
- 🦖 Why visualizing data matters
- 🎯 Confidence intervals
- 💬 Communicating data summaries responsibly