Describing and visualizing data

Lecture 3

Visual summary of data

Raw data vs. summary

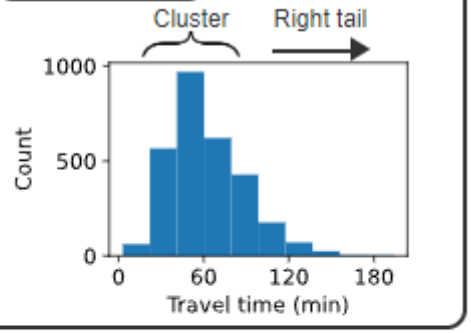
To have any meaningful insight we need to have **summaries** of data

Australian train passenger survey

ID	License	Cost	TotalTime	Satisfaction
126	0	5.0	95	1
131	1	3.6	60	3
142	1	2.5	80	4
145	1	5.5	43	3
168	1	2.3	33	3
...	...	...	...	...
22598	1	0.0	68	4
22620	1	0.0	50	4
22642	1	3.4	26	2
23003	1	1.2	35	4
23014	1	0.0	21	4

Raw data recorded as a table

Visual summary

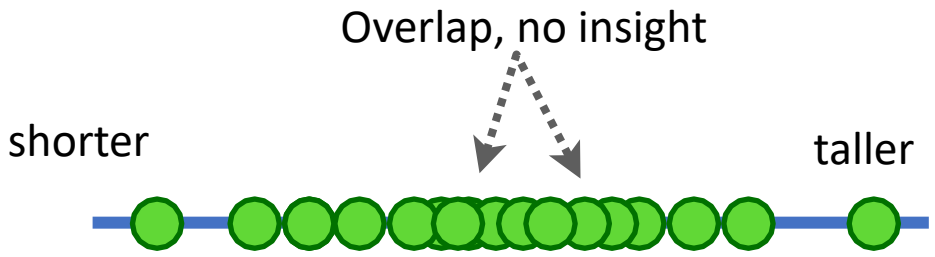


Numerical summary

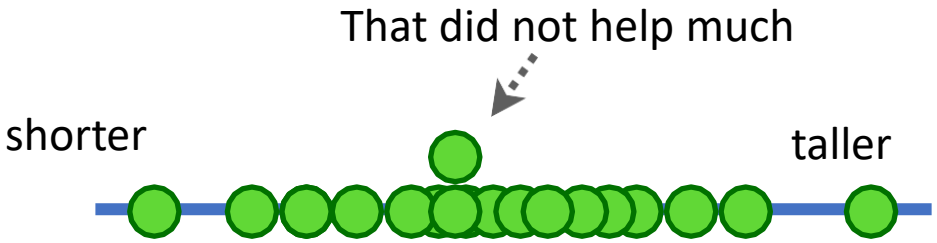
count	2927.00
mean	62.60
std	25.83
min	3.00
25%	44.50
50%	59.00
75%	78.00
max	194.00

Summarizing one (numeric) dimension

- We measured the heights of many people

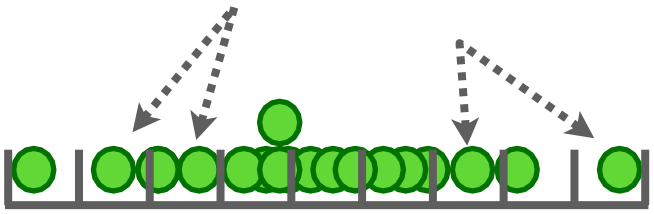


- Idea: stack the same values

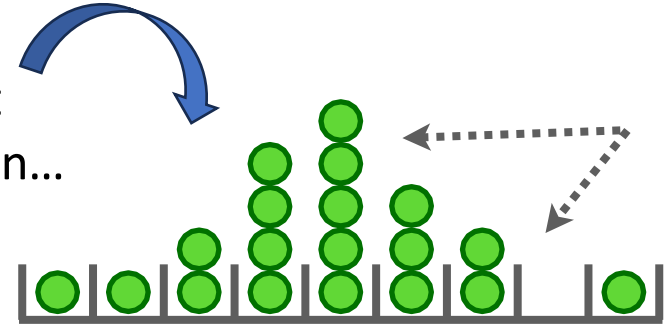


Visualizing one dimension: bins

we (artificially) divide the range of values into classes (bins)



…and stack the measurements that fall into the same bin…

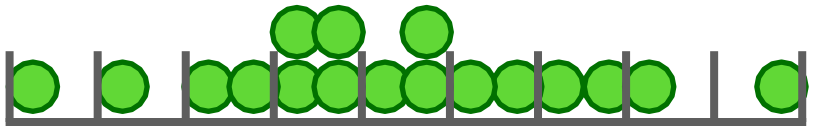


…and we end up with a **histogram**

- *Histogram* is a bar chart of the frequency distribution of a single numeric attribute

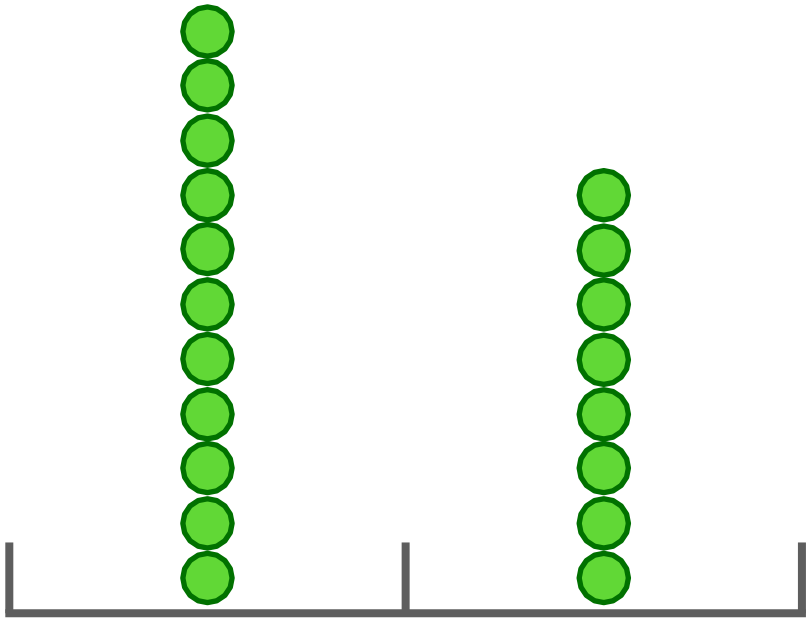
Challenge: bin size selection

…and if the bins are too narrow, then they're not much help…

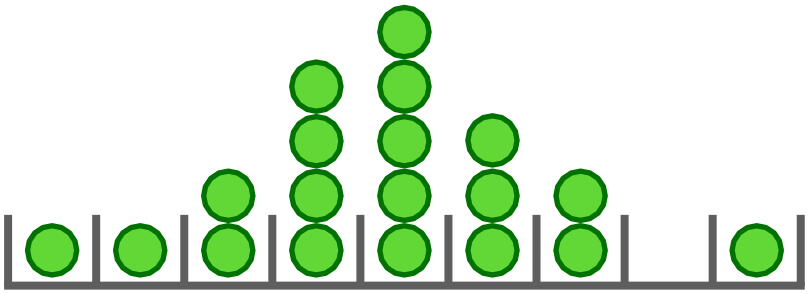


Challenge: bin size selection

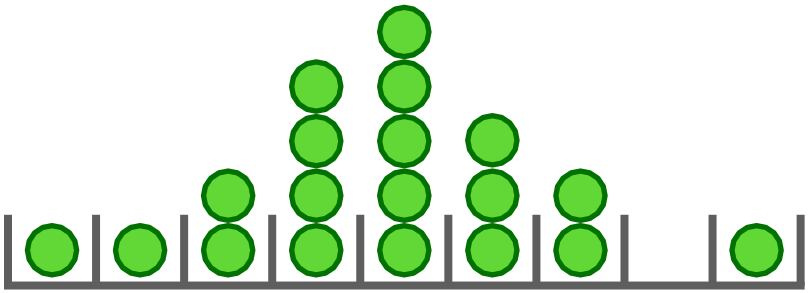
If the bins are too wide, then they're not much help either…



Challenge: bin size selection



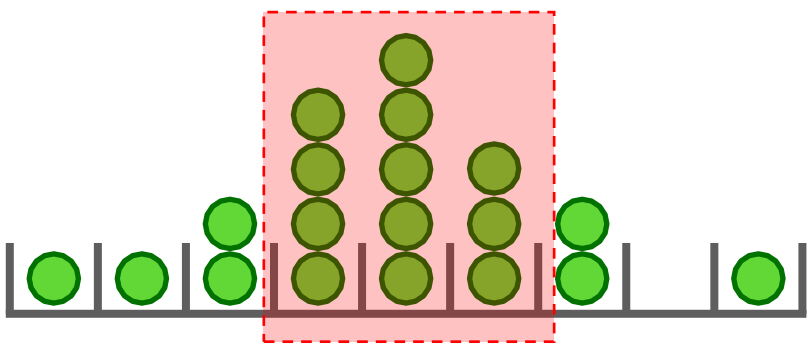
Histogram: what does it tell us



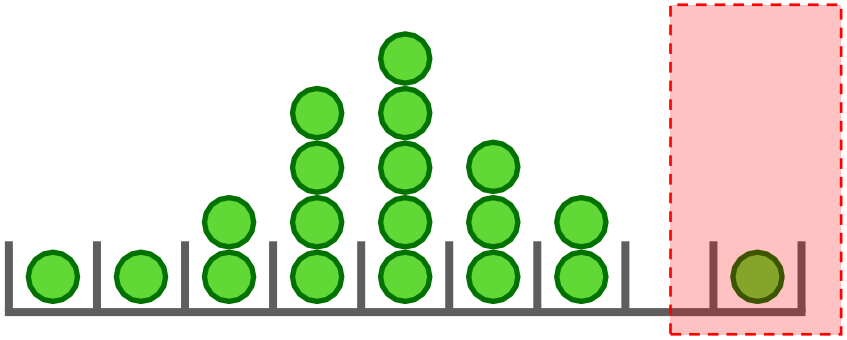
The taller the stack within a bin, the more measurements we made fell into that bin.

We could use the histogram to estimate the probability of getting future measurements.

Histogram: insights

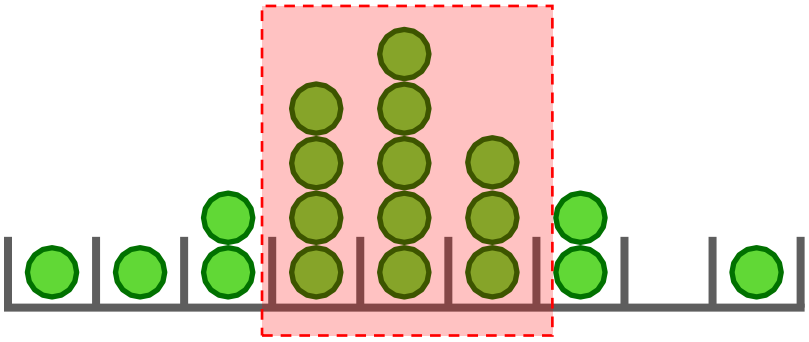


Because most of the measurements are inside this **red box**, we might be willing to bet that the next measurement we make will be somewhere in this range.



Measurements out here are rare, and less likely to happen in the future

Histograms: estimating probabilities



… This means that **63%** of the time we'll get a measurement in the **red box**.

…we count the number points *in the box* = **12**…

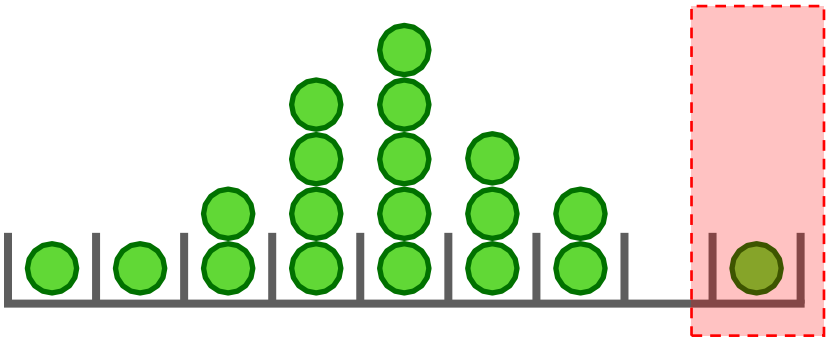
We want to estimate the probability that the next measurement will be in this **red box**…

…and divide by the *total* number of points, **19**…

$$\frac{12}{19} = 0.63$$

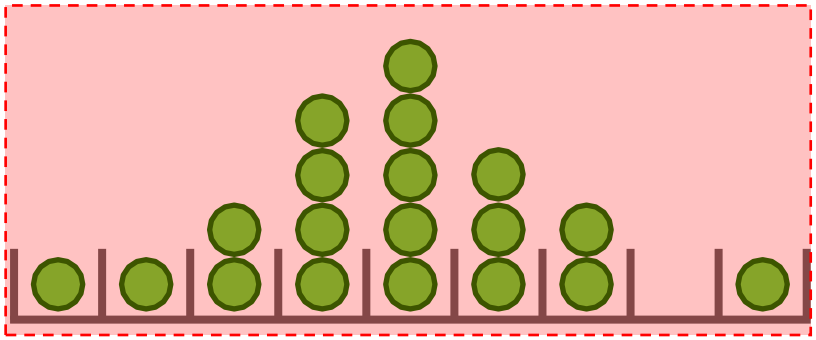
Note that the confidence of this estimate depends on the number of measurements: the more measurements you have, the more confidence you will have in the estimate.

Histograms: estimating probabilities



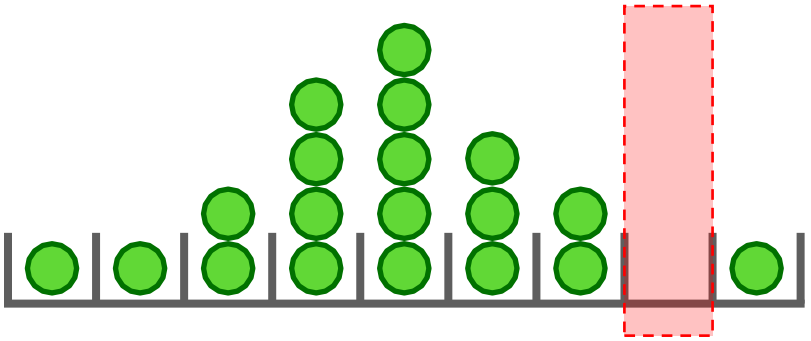
Now you try

Histograms: estimating probabilities



The probability to be in this interval (according to our data): 1.0

Histograms: estimating probabilities



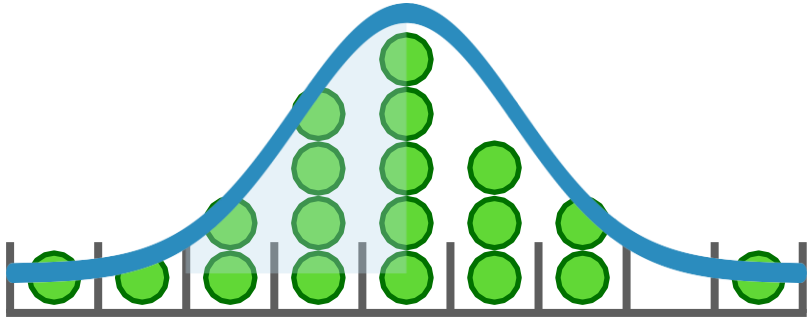
…and we get **0**. It means that we'll ***never*** get a measurement in this box. But is this true?

The probability is: ?

If we measure more people, we may find someone who fits in this bin or become more confident that it should be empty.

However, sometimes getting more measurements can be expensive, or take a lot of time. This is a problem!!!

Curve Approximation of a Histogram



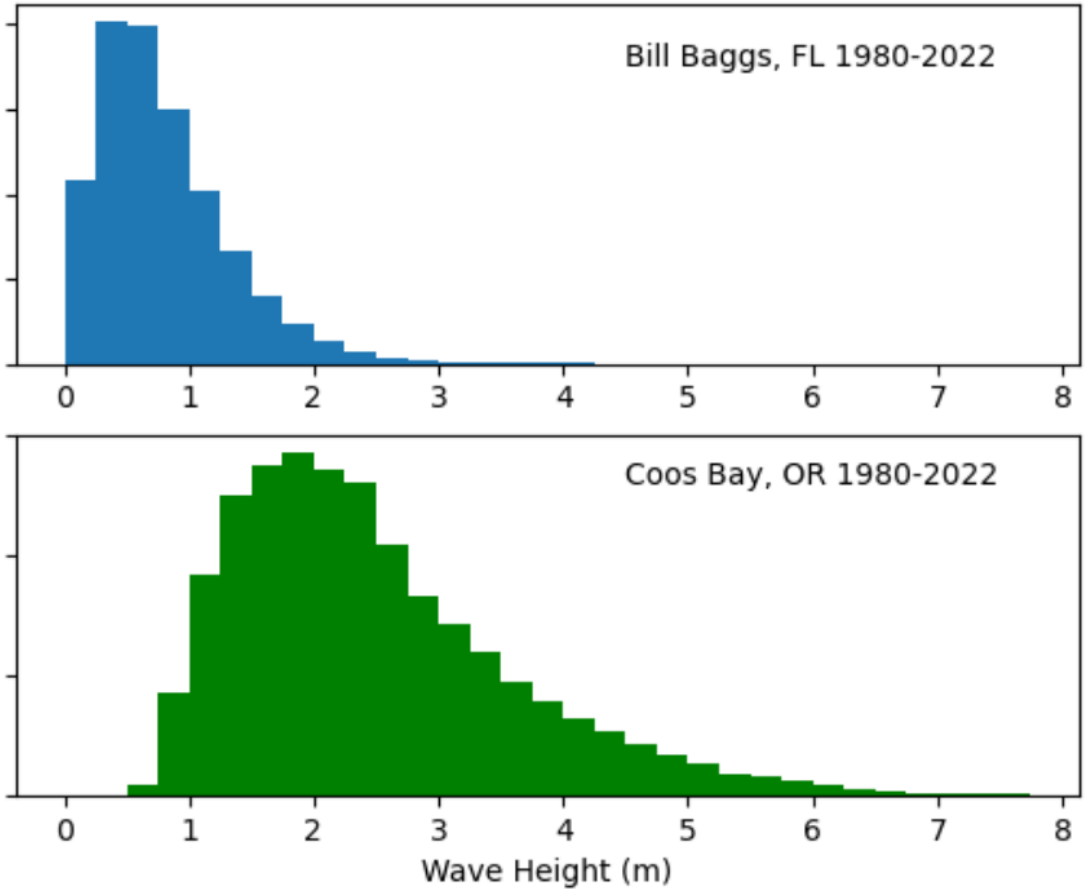
We can use a curve to approximate the shape of the histogram

We had a *discrete* frequency distribution and replaced it with a *continuous* distribution.

The shape of the histogram hinted that in case of heights, it seems to be a normal distribution.

<u>We will discuss different distributions later in the course.</u>

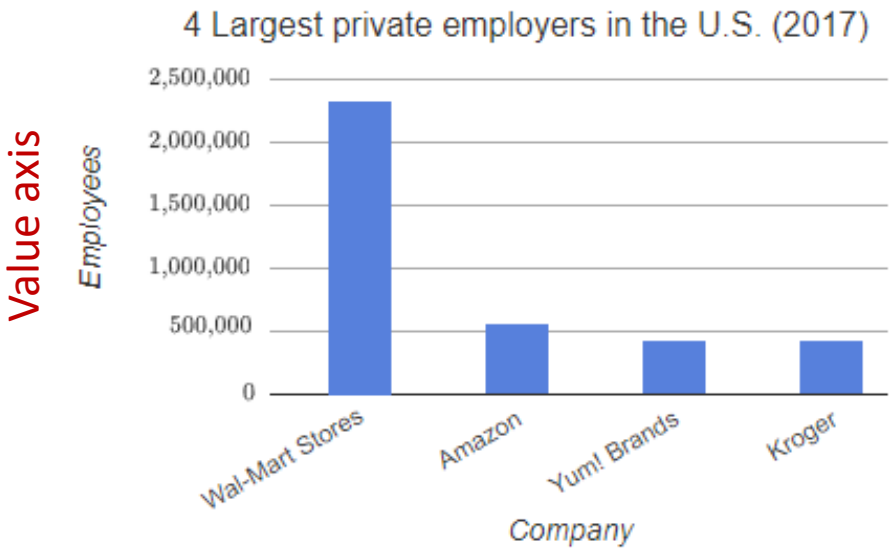
Example: comparing histograms



Wave energy converter (WEC) transforms the kinetic energy of ocean waves into electricity

Bar chart

- A bar chart depicts one numeric dimension of data for different categories, using rectangular bars having lengths proportional to the total for this category.

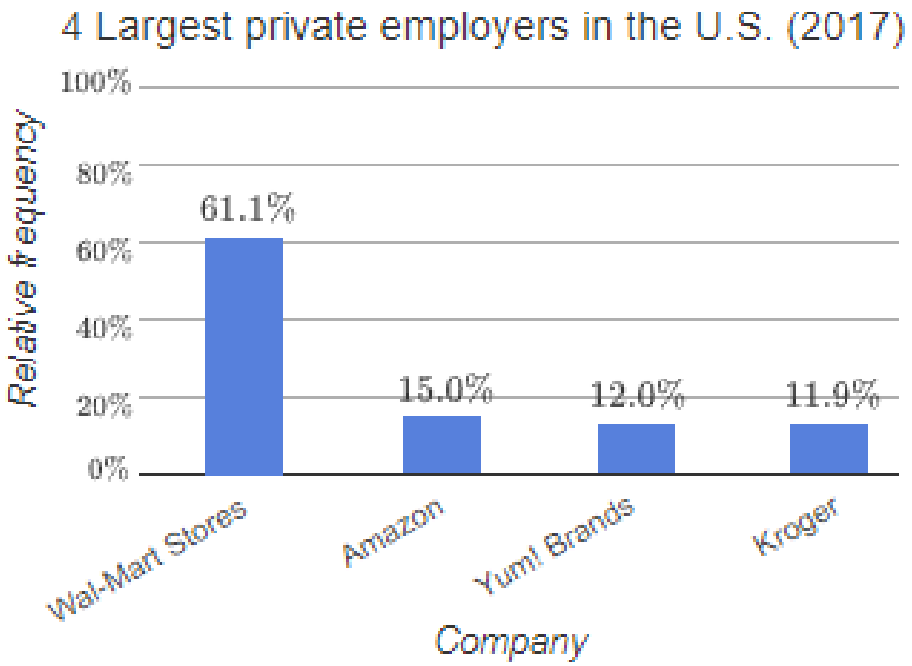


Relative-frequency bar chart

- A relative frequency bar chart shows the percent that each category is of the total

Company	Employees	
Walmart Stores	2,300,000	61.1%
Amazon	566,000	15.0%
Yum! Brands	450,000	12.0%
Kroger	449,000	11.9%
Total	3,765,000	100.0%

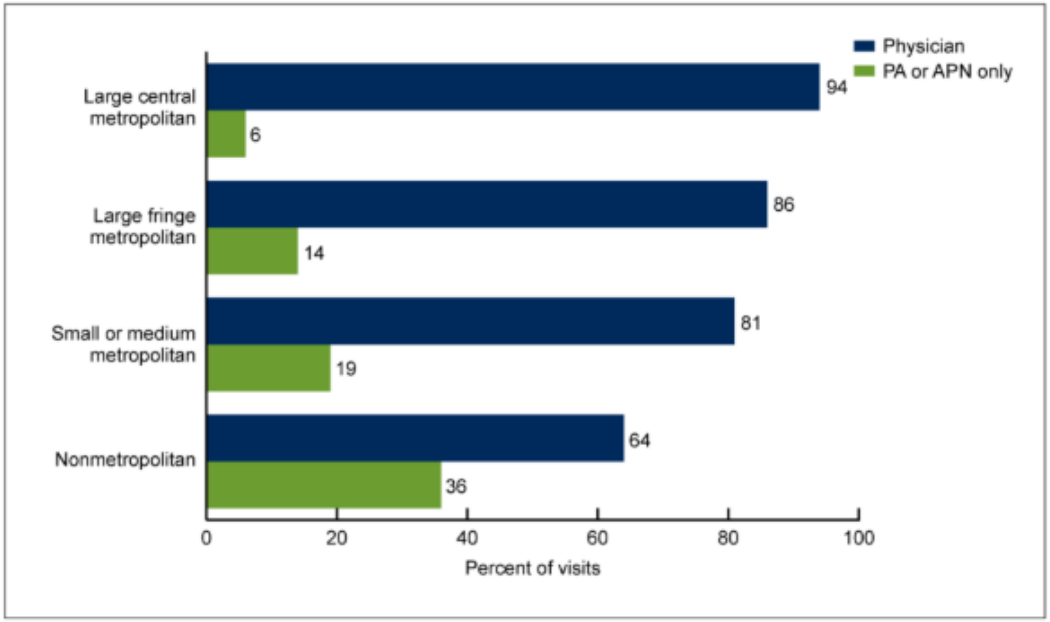
$$\frac{2{,}300{,}000}{3{,}765{,}000} \cdot 100\% = 61.1\%$$



Grouped bar chart

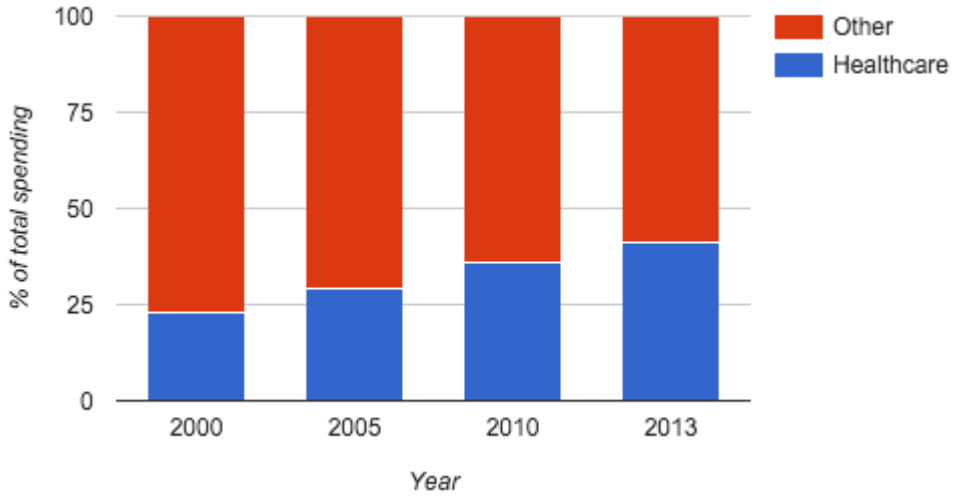
- A grouped bar chart depicts two or more additional different categories on a single bar chart, with each group using a different colored (or shaded) bar

Utilization of physician assistants (PA) and advance practice nurses (APN) by hospital location



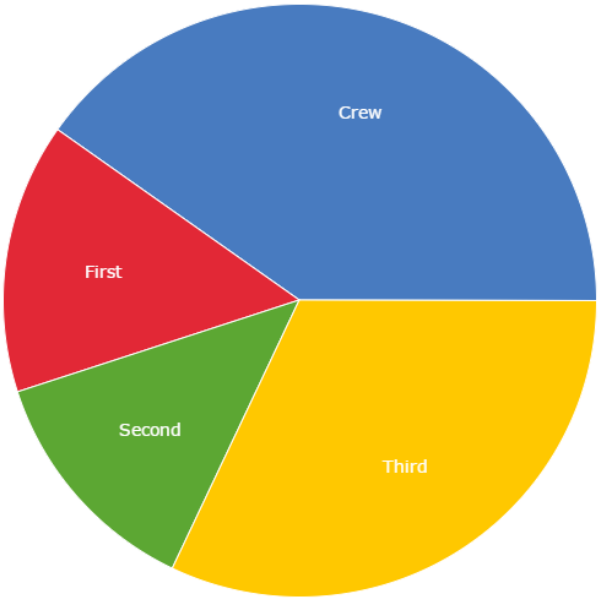
Relative frequency stacked bar chart

Massachusetts state spending on healthcare versus all other state spending

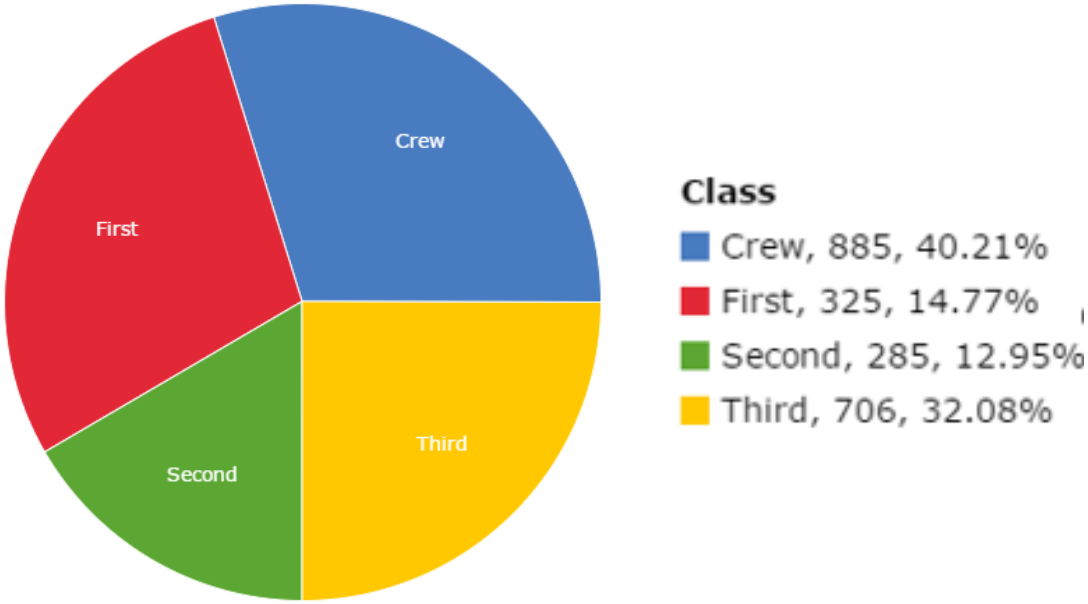


Pie charts

Titanic passengers by class



Titanic survivors by class



Numeric summary of a single numeric dimension

Measures of centrality

Mean

Median

Mode

- Provide a single, representative value of the entire dataset
- Knowing central tendency helps understand what is typical
- Help comparisons across different datasets

Mean - average value of the data

- Sample mean: $\bar{x}$

We can directly compute this. The larger the sample, the closer we get to the population mean

$$\bar{x} = \frac{\sum_{i=0}^{n-1} x_i}{n}$$

Averaged over sample size

- Population mean: μ

In most cases we cannot directly compute this.

$$\mu = \frac{\sum_{i=0}^{N-1} x_i}{N}$$

Averaged over the size of the entire population

Median - a middle value in a sorted list of data

$\{x\} = \{4, 6, 4, 1, 8, 7, 7, 2, 5, 7\}$

$\{x\}_{sorted} = \{1, 2, 4, 4, 5, 6, 7, 7, 7, 8\}$

n=10
(even)

Median is between these two central points:
median=5.5

$\{x\} = \{4, 6, 4, 1, 8, 7, 7, 2, 5\}$

$\{x\}_{sorted} = \{1, 2, 4, 4, 5, 6, 7, 7, 8\}$

n=9
(odd)

Median is in the middle position:
median=5

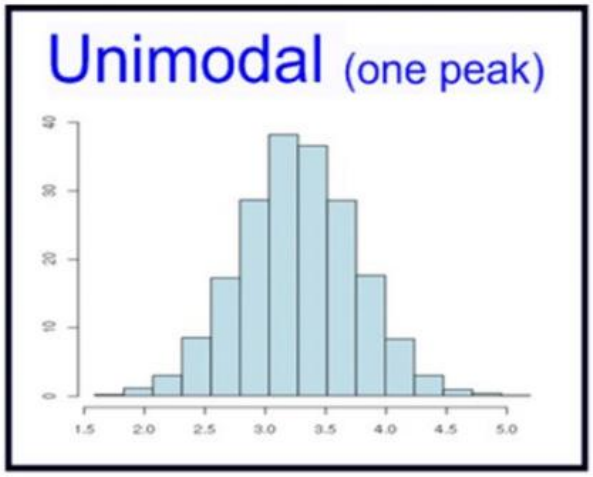
Mode - a value in a dataset that occurs most frequently

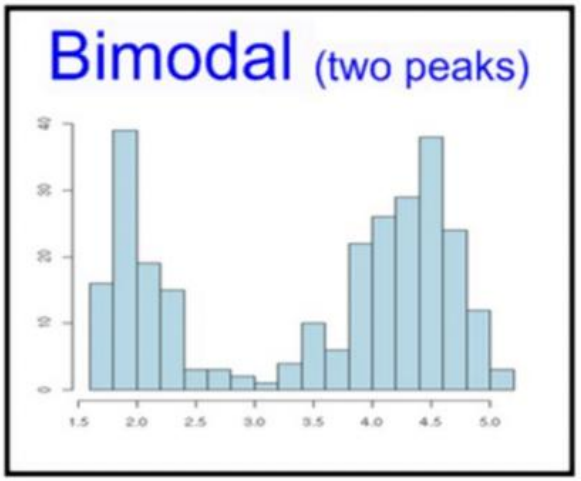
We can see it clearly if we sort the data

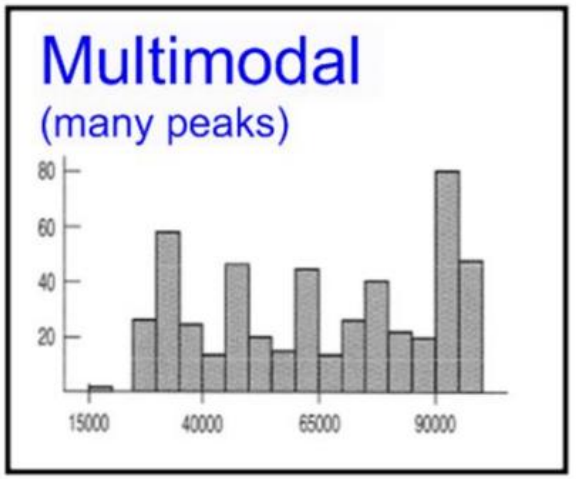
$\{x\}_{sorted}$ = {1, 2, 4, 4, 5, 6, 7, 7, 7, 8}

Mode=7

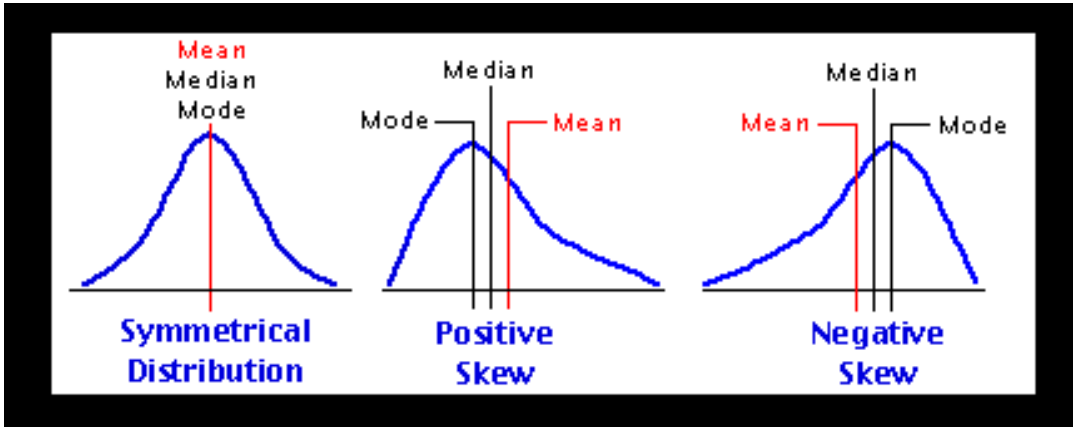
Alternative definition: mode is a peak (or bump) in a frequency distribution (histogram):





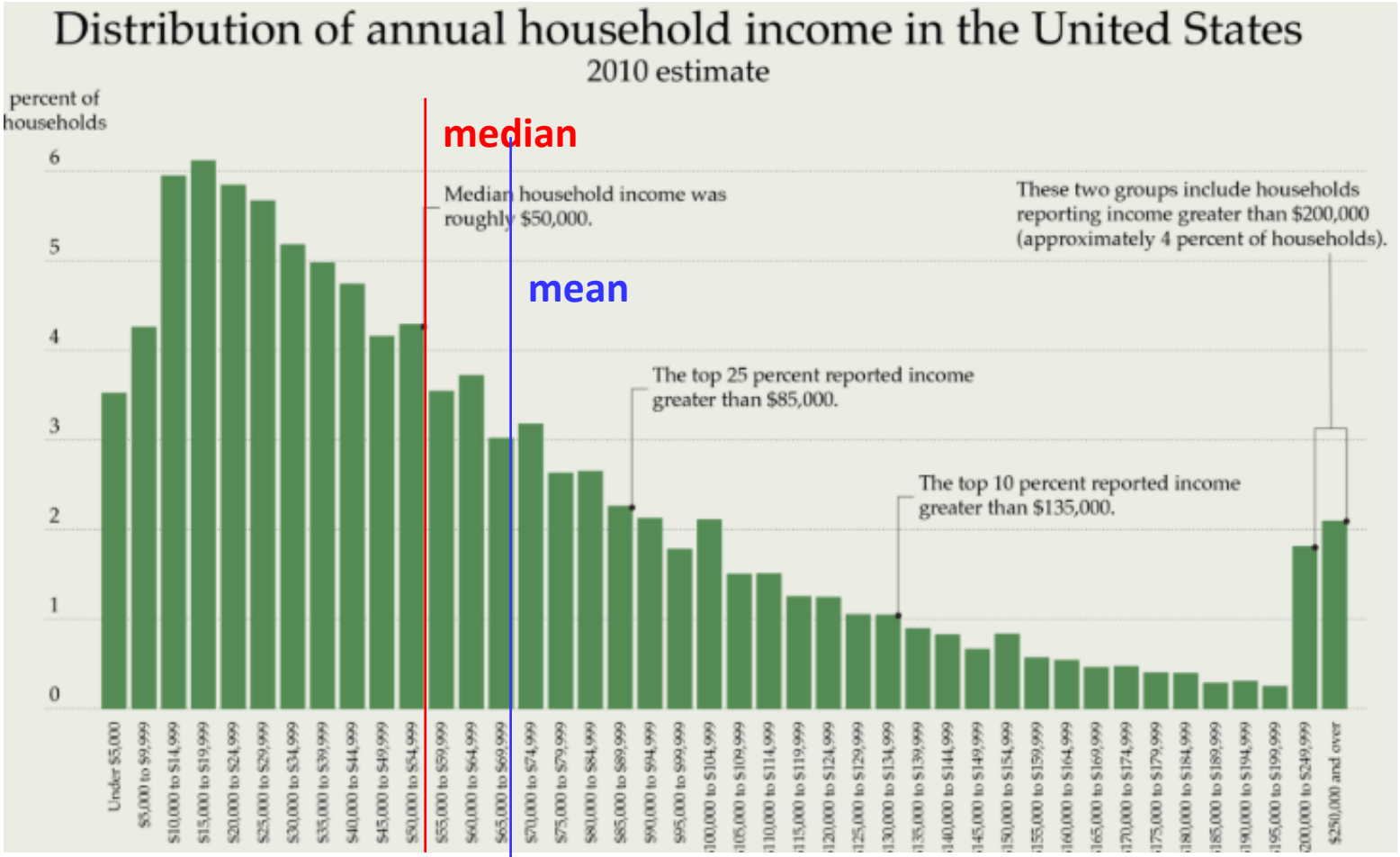


Symmetry and Skewness



- In a symmetrical distribution, the mean, median, and mode are equal.
- For a right-skewed (positively skewed) distribution, the mean is greater than the median, which is greater than the mode.
- Conversely, for a left-skewed (negatively skewed) distribution, the mode is greater than the median, which is greater than the mean.

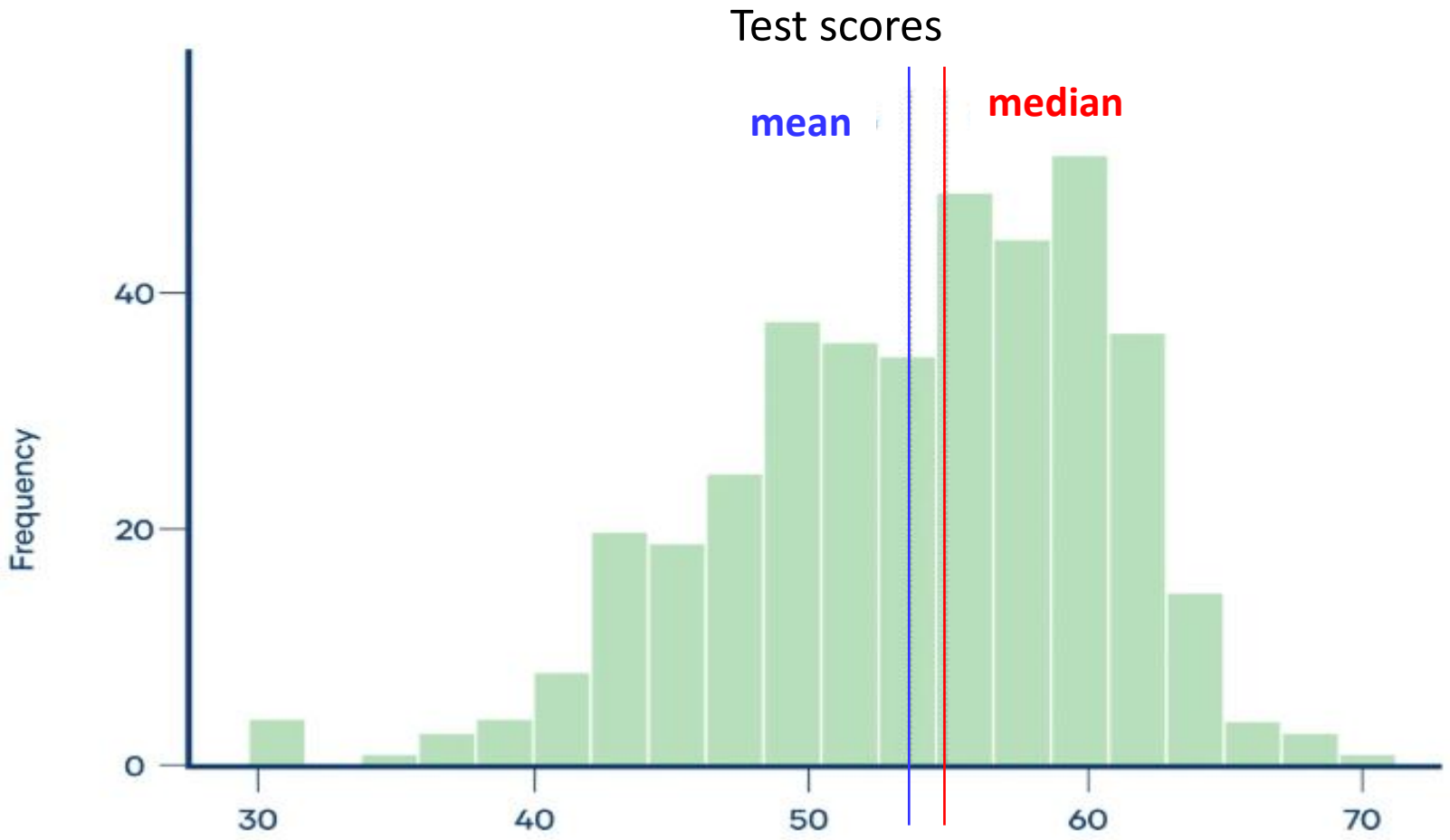
Example of positive (right) skew



High counts are are below median

Long tail to the right: represent outliers

Example of negative (left) skew



Long tail to the left: represent outliers

High counts are above median

Measures of dispersion

Variance

Standard deviation

- Quantify the variability or spread of data around a central point
- Reveal how consistent or diverse a dataset is
- Help identifying outliers

Variance - an average of a squared deviation from the mean

- Population variance: σ^2

$$\sigma^2 = \frac{\sum_{i=0}^{N-1}(x_i - \mu)^2}{N}$$

Variance - an average of a squared deviation from the mean

- Population variance: σ^2

$$\sigma^2 = \frac{\sum_{i=0}^{N-1}(x_i - \mu)^2}{N}$$

- Sample variance: s^2

$$s^2 = \frac{\sum_{i=0}^{n-1}(x_i - \overline{x})^2}{n-1}$$

Why *n* - 1?

- We want a sample variance to give an idea about population variance
- If we divide sum of deviations by n the variance in the sample will be smaller than the variance in the population: we would **underestimate** the actual variance.
- Dividing by *n*-1 gives a less biased estimator of the population variance

Standard Deviation - a square root of the variance

- Population std: $\sigma = \sqrt{\sigma^2}$
- Sample std: $s = \sqrt{s^2}$

By taking a square root we are back to the units in which the measurements were made.

Now we can summarize the sample as $\boldsymbol{\overline{x} \pm s}$

Measures of position

Number of standard deviations from the mean

Quantiles

- Indicate where a given value is located with respect to other data points
- Divide data into segments and reveal how the data is spread out.
- Summarize the key characteristics of a dataset and help identify outliers

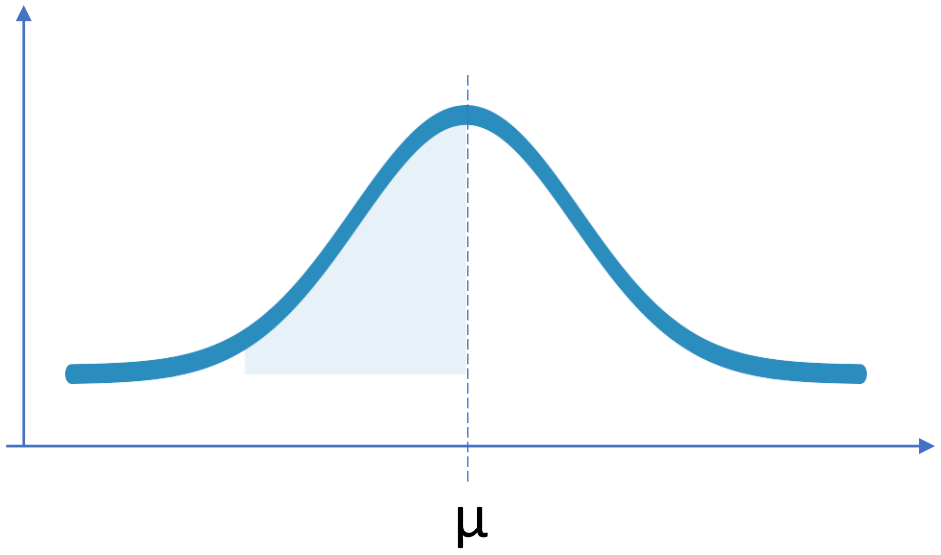
Measure of position

Number of standard deviations from the mean

Empirical rule for bell-shaped (normal) distributions

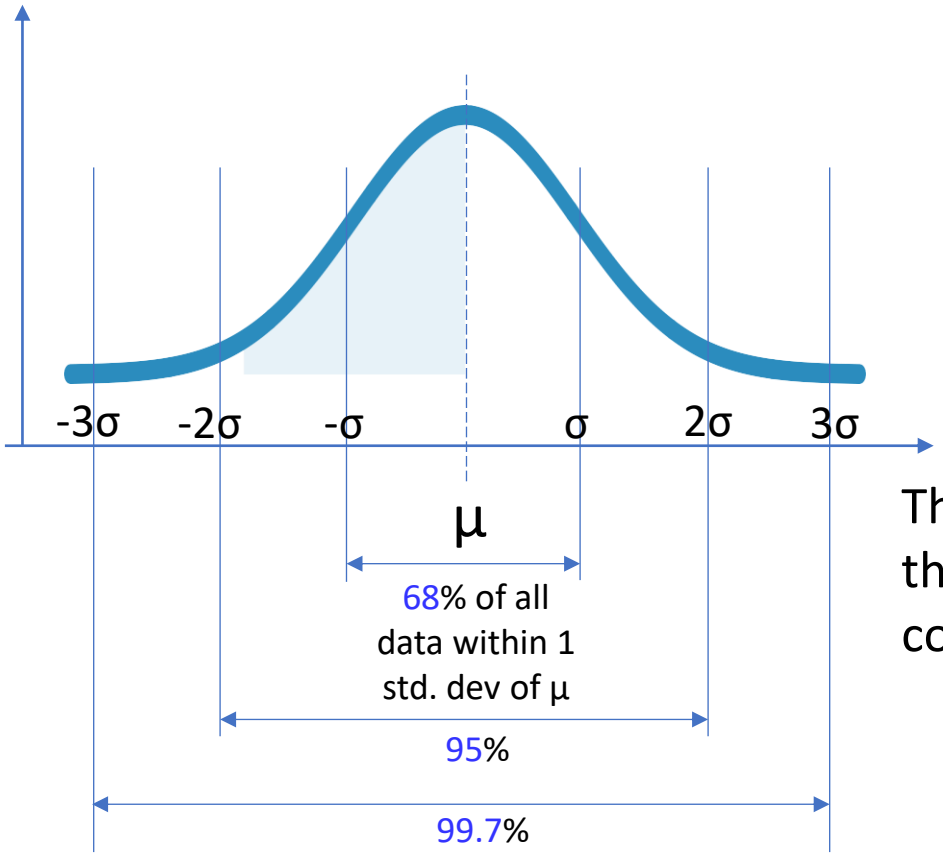
- Bell-shaped data: symmetrical around mean:

mean=median=mode



Empirical rule for bell-shaped (normal) distributions

- The rule: 68 – 95 – 99.7



The data points that fall more than 3 σ from the mean are considered outliers

Using empirical rule: example

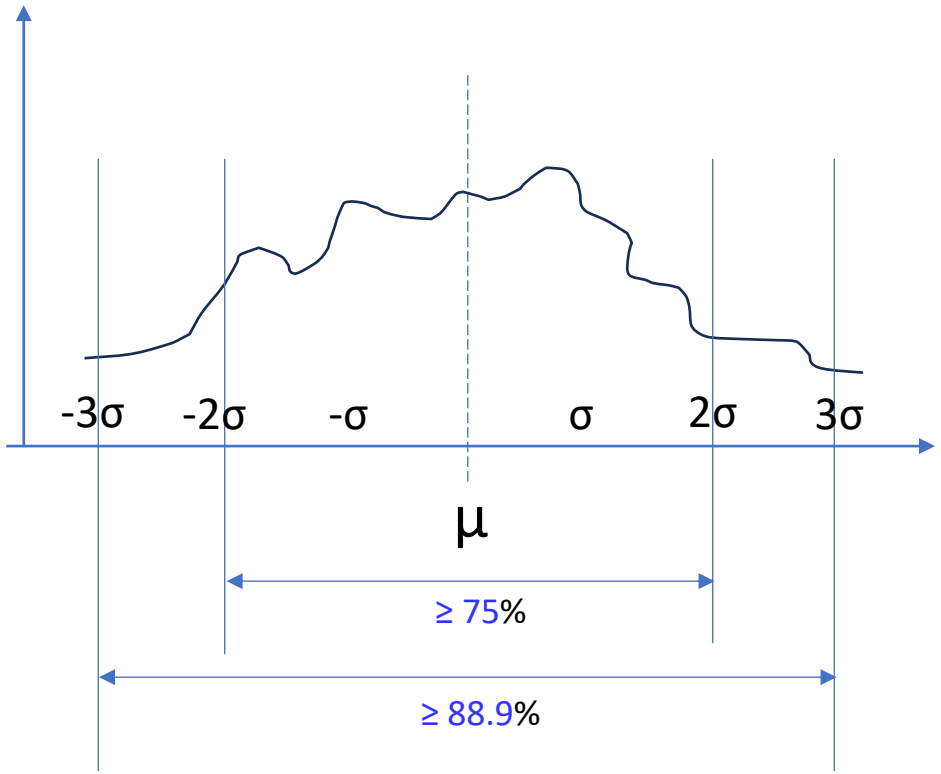
- The weight of newborn babies is normally distributed with μ=3000g and σ=500g.

What percent of babies is between 2 and 4 kg?

Answer: 95% (2 std. dev from the mean)

Chebyshev theorem: for any distribution

- For any dataset and for any distribution the % of data within *k* standard deviations from the mean is at least $1 - \frac{1}{k^2}$, for *k*>1



Number of standard deviations from the mean: normal distribution

- A value's location in the (normally distributed) dataset can be measured relative to the mean.
- A *standardized score*, or z-score of a specific data point, describes how many standard deviations from the mean —and in which direction a given value lies.
- To compute z-score of value *val*:

$$z = \frac{val - \mu}{\sigma}$$

Measure of position

Percentiles and quartiles

Measure of position: percentiles

- A *percentile* is a data value for which a specified proportion of the distribution falls at or below the value.

- The n-th percentile of a dataset is the data value such that n percent of the data falls at or below that value.

    - Example: The median is the 50-th percentile since half of the data values fall at or below the median

Quartiles

- The first quartile (Q_1) is the 25th percentile: one-quarter of the data fall at or below Q_1
- The second quartile (Q_2) is a median of the dataset
- The third quartile (Q_3) is the 75th percentile: three-quarters of the data fall at or below Q_3

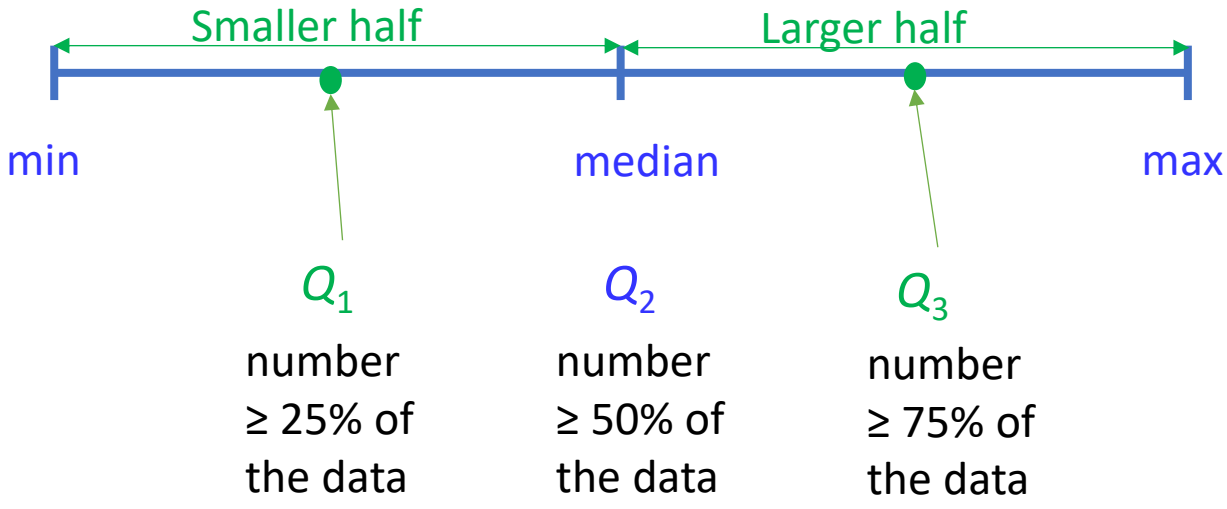
The five-number summary

- Min
- Max
- The first quartile (Q_1)
- Median (Q_2)
- The third quartile (Q_3)

min Q_1 Q_2 Q_3 max

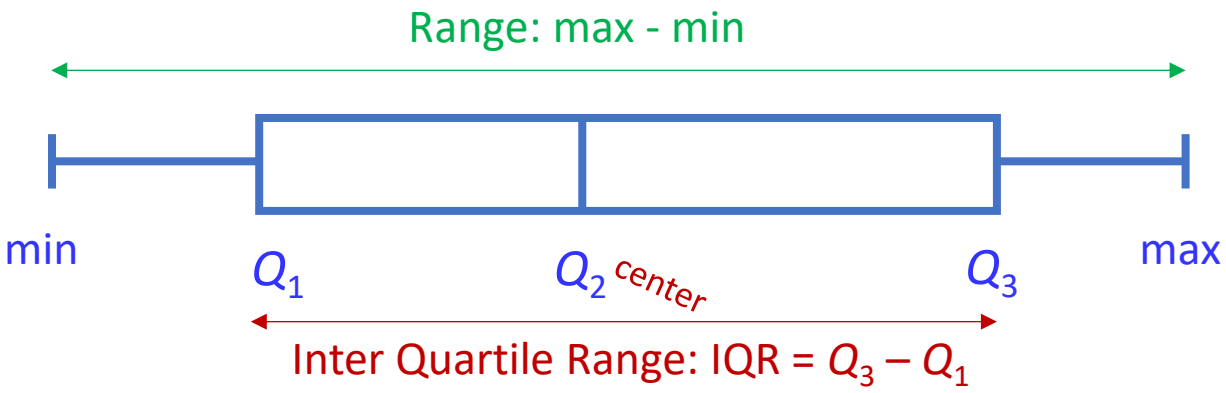
The five-number summary

- Quartiles divide dataset into 4 parts such that each part covers 25% of data:



Box plot

- Plot these five values along numeric axis → **box plot**



Example: five number summary

- Dataset: 8, 9, 2, 10, 3, 5, 7, 12, 15

sort

- 2, 3, 5, 7, 8, 9, 10, 12, 15 n=9

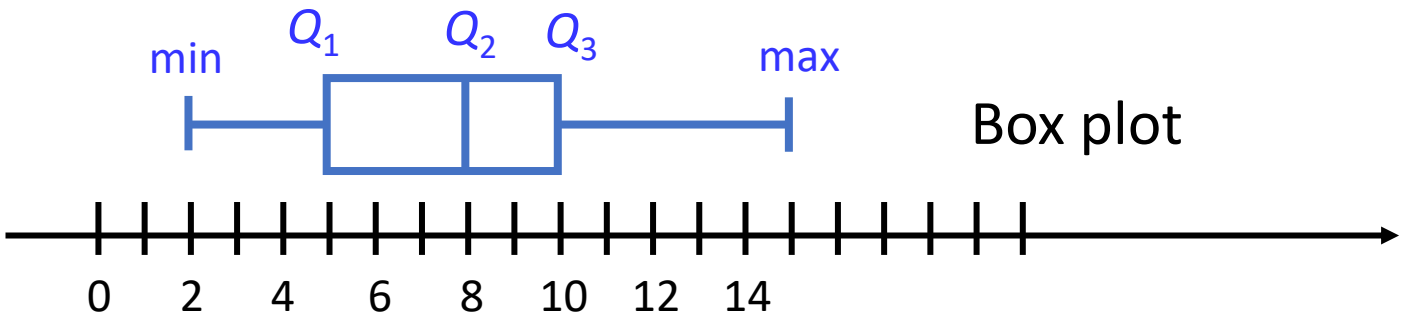
median

Q_2

Include it in both lower and upper halves

- Smaller half: 2, 3, 5, 7, 8 Larger half: 8, 9, 10, 12, 15

Q_1 Q_3



Using quartiles for outlier detection

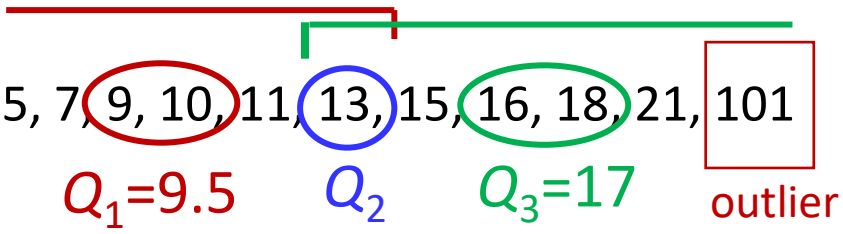
min Q_1 Q_2 Q_3 max

- Compute the Inter-Quartile Range $IQR = Q_3 - Q_1$

Outlier thresholds:

- Min threshold: $Q_1 - 1.5IQR$
- Max threshold: $Q_3 + 1.5IQR$

Example: outliers

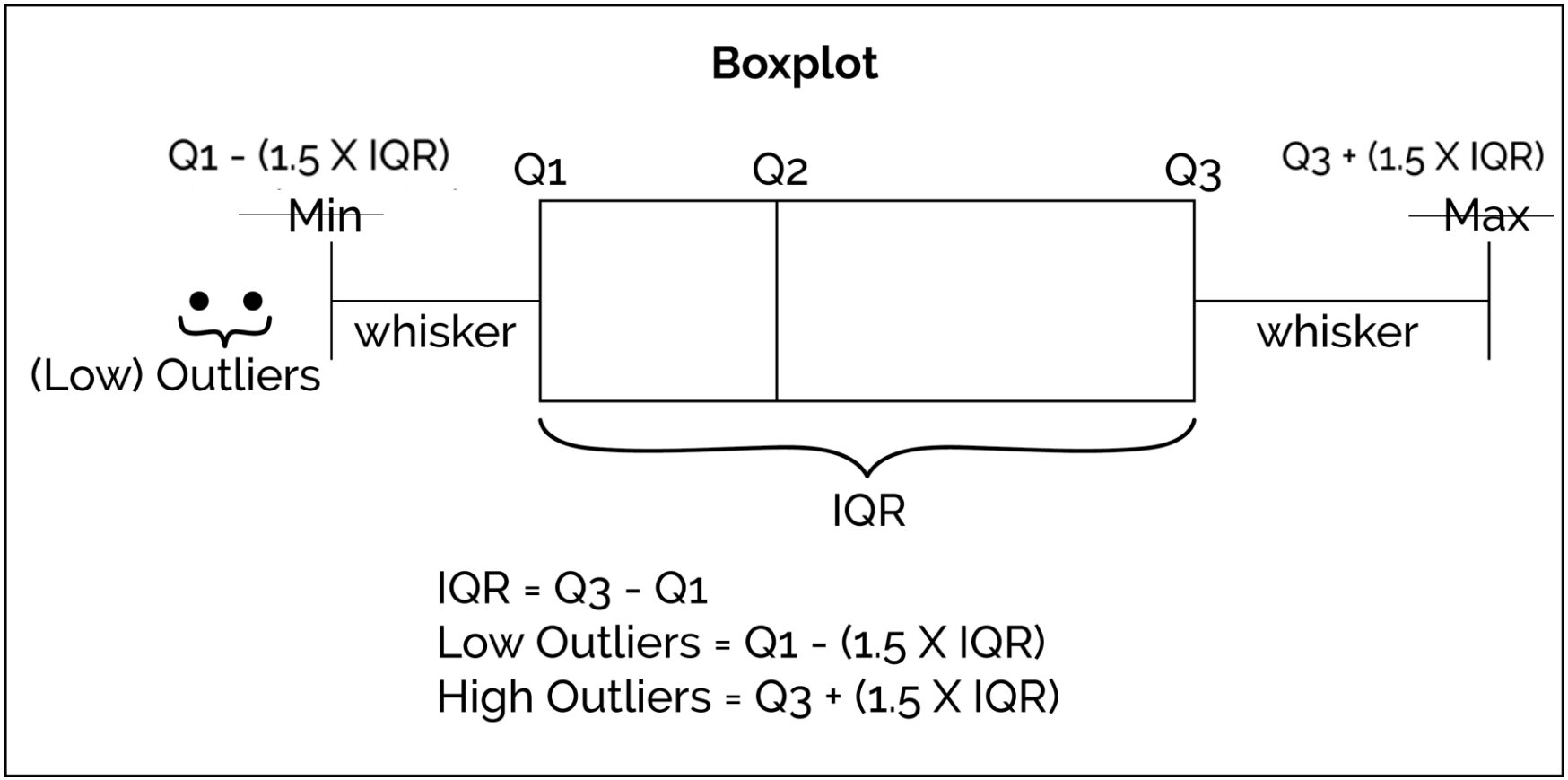


- *IQR* = 17 – 9.5 = 7.5

- Lower threshold: 9.5 – 1.5**IQR* = 9.5 – 11.25
- Upper threshold: 17 + 1.5**IQR* = 17 + 11.25 = 28.25

- 101 is an outlier

Box plot: with outliers

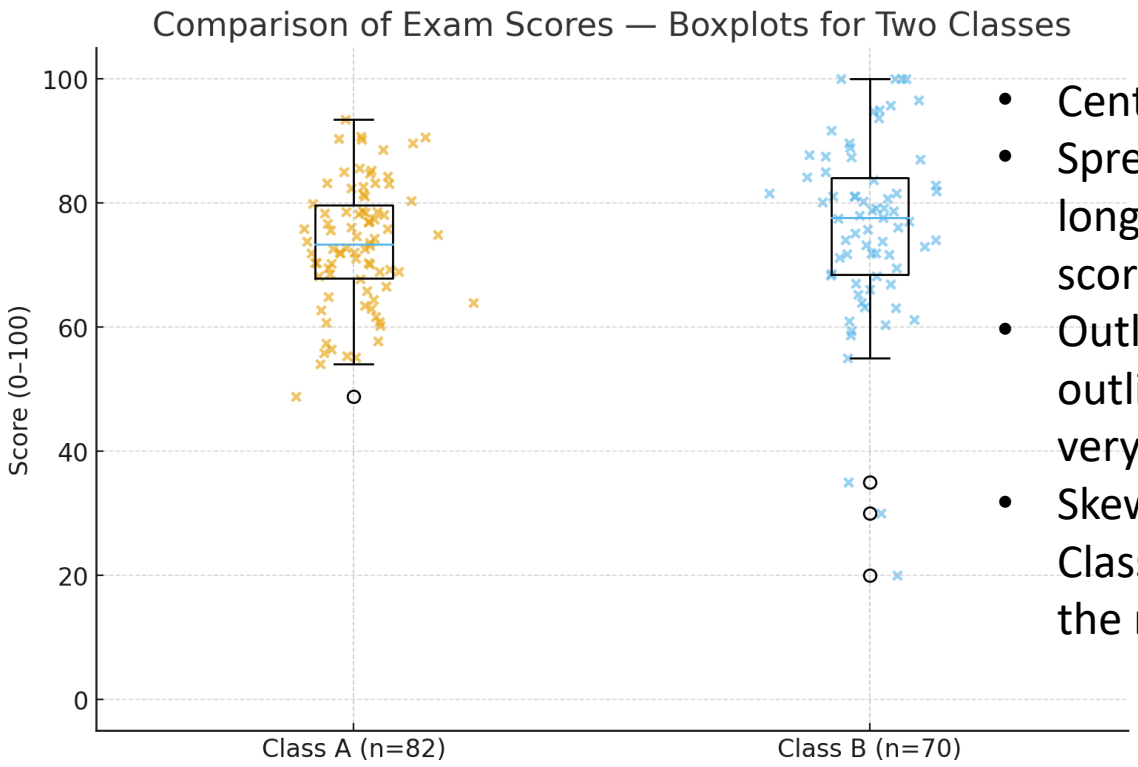


Comparing datasets: two sections exam scores

Basic statistics and the 5 number summaries

index	count	mean	std	min	25%	50%	75%	max
Class A (n=82)	82	73.3107	9.883284	48.80255	67.90886	73.38038	79.6646	93.52278
Class B (n=70)	70	75.89412	15.04578	20	68.42974	77.63265	84.04018	100

Boxplots



- Central tendency: B has higher median
- Spread: Class B has a wider IQR and longer whiskers — more variability in scores.
- Outliers: Class B has several clear low outliers (scores around 20–35) some very high scores near 100
- Skewness: Class B looks more skewed. Class A is more concentrated around the mid–high 60s/70s.

- Practical takeaway: Even if Class B's median is higher, the greater spread and presence of low outliers might motivate different interventions (e.g., targeted support for low performers in Class B).