Chapter 2 - Part 1

Descriptive Statistics

Describing Categorical (Qualitative) Data

- A class is one of the categories into which categorical data can be classified.
- The class frequency is the number of observations in the data set that fall into a particular class.
- The class relative frequency is the class frequency divided by the total number of observations in the data set; that is,

$$\text{class relative frequency} = \frac{\text{class frequency}}{n}.$$

- The class percentage is the class relative frequency multiplied by 100; that is,

$$\text{class percentage} = \text{class relative frequency} \times 100.$$

Example: Favorite Coffee Type Survey

A campus survey asked the following question to 100 randomly selected students:
What is your favorite coffee type?

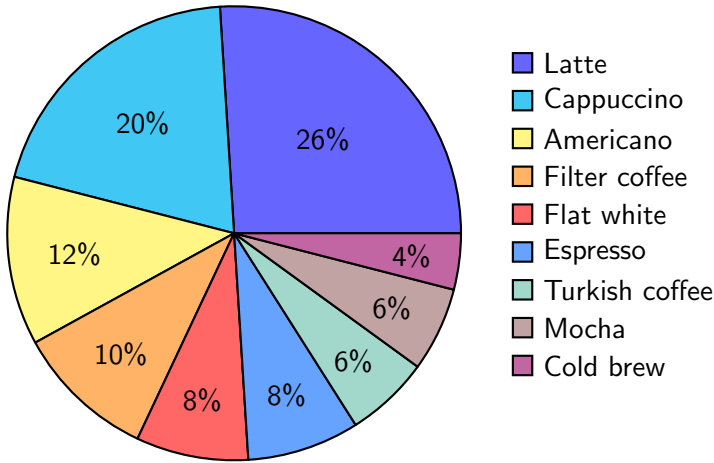
Respondent	Favorite Coffee
1	Latte
2	Cappuccino
3	Latte
4	Filter coffee
5	Americano
6	Mocha
7	Flat white
8	Cappuccino
9	Turkish coffee
10	Cappuccino
⋮	⋮
250	Latte

Frequency and Relative Frequency Table

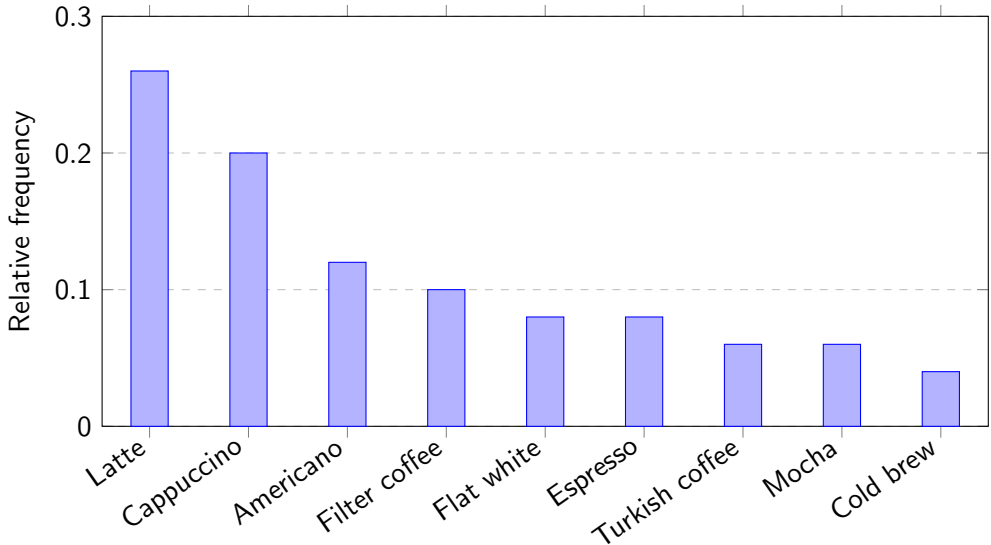
- Browsing the complete data set, we see that the classes of this qualitative variable Favorite Coffee are
  - Latte, Cappuccino, Americano, Filter coffee, Flat white, Espresso, Turkish coffee, Mocha, Cold brew
- We are ready to create the frequency and relative frequency tables.

Coffee Type	**Frequency**	**Relative Frequency**
Latte	65	0.26
Cappuccino	50	0.20
Americano	30	0.12
Filter coffee	25	0.10
Flat white	20	0.08
Espresso	20	0.08
Turkish coffee	15	0.06
Mocha	15	0.06
Cold brew	10	0.04
Total	**250**	**1.00**

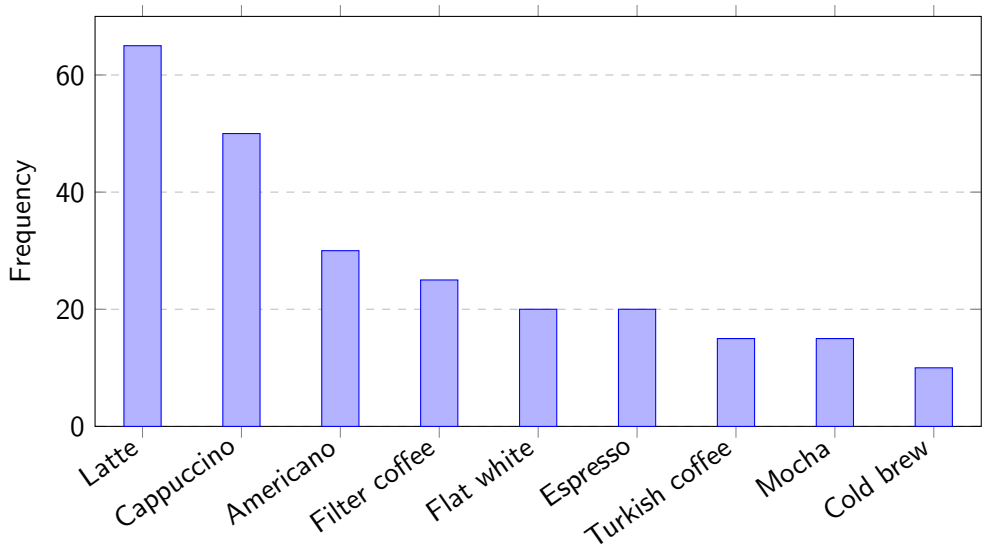
Pie Chart


26%
20%
12%
10%
8%
8%
6%
6%
4%
Latte
Cappuccino
Americano
Filter coffee
Flat white
Espresso
Turkish coffee
Mocha
Cold brew

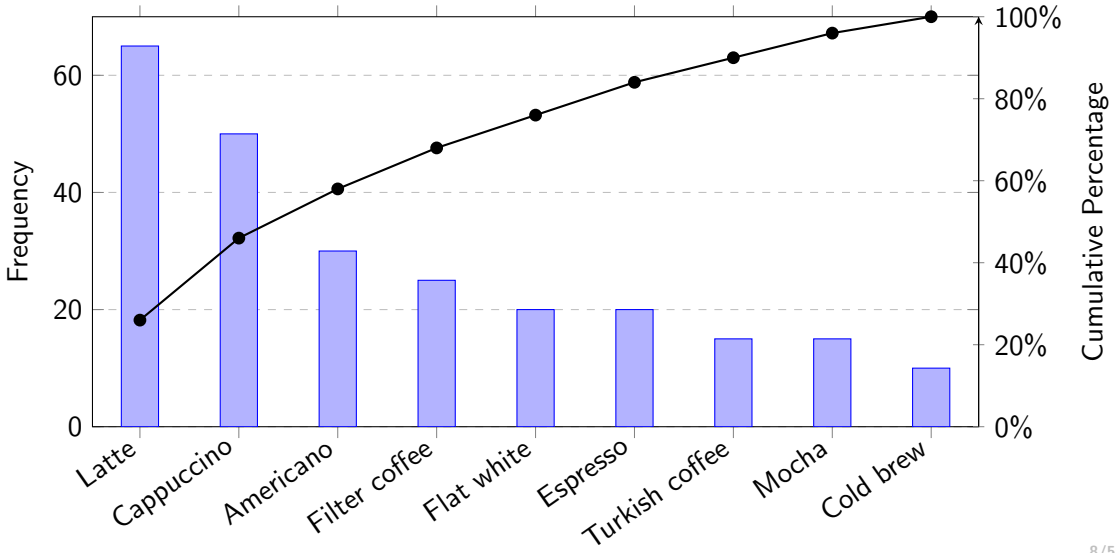
Bar Chart (Using Relative Frequencies)



Bar Chart (Using Frequencies)



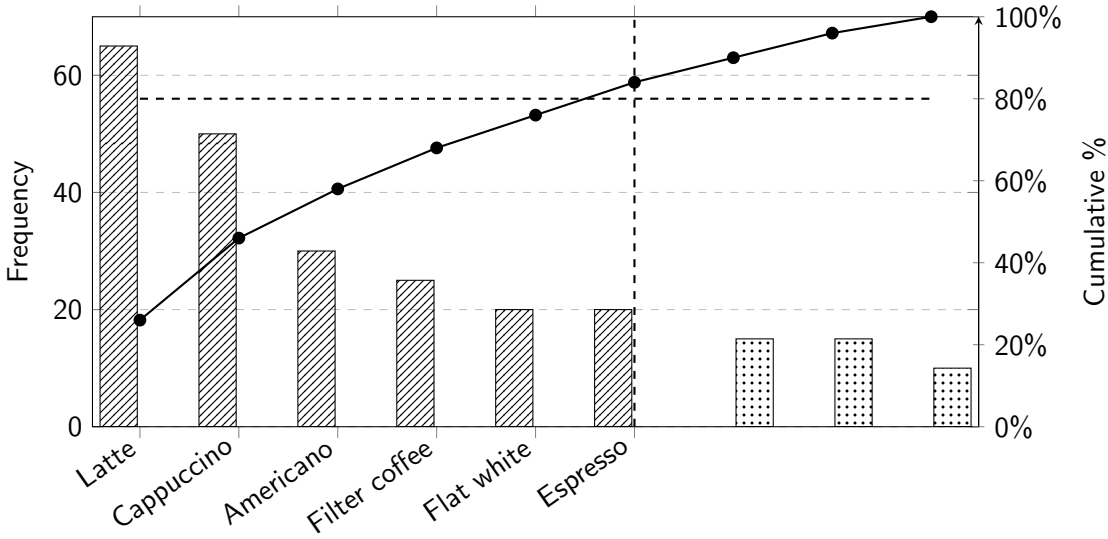
Pareto Chart



Pareto Chart

- A Pareto chart is a bar chart that displays categories in descending order of frequency, together with a cumulative percentage curve.
- It is used to identify the "vital few" categories that account for most of the total.
- Categories contributing to about 80% of the cases are referred to as the vital few.
  - In our example: Latte, Cappuccino, Americano, Filter coffee, Flat white, Espresso.
- The remaining categories are referred to as the useful many.
  - In our example: Turkish coffee, Mocha, Cold brew.

Pareto Chart



Describing Numerical (Quantitative) Data

- Two common ways of describing quantitative data are:
  - **Numerical Summaries**
    - Measures of Centrality describe the center or typical value of a data set. (mean, median, mode, etc.)
    - Measures of Variability describe how spread out or dispersed the data values are. They indicate how much the observations differ from one another and from the center. (range, interquartile range (IQR), variance, standard deviation, etc.)
  - **Graphical Summaries** (histogram, boxplot, etc.)

Mean (Average)

- The mean of a set of numerical data is the sum of the measurements, divided by the number of measurements contained in the data set.
- If data is $\{x_1, x_2, x_3, \ldots, x_n\}$, then mean is

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

- **Example:** Calculate the mean of the following five measurements: $\{7, 3, 8, 4, 7\}$

$$\bar{x} = \frac{7 + 3 + 8 + 4 + 7}{5} = \frac{29}{5} = 5.8$$

Mean (Average)

- We adopt a general policy of using Greek letters to represent numerical descriptive measures of the population and Roman letters to represent corresponding descriptive measures of the sample.
- $\bar{x}$ denotes the sample mean
- μ denotes the population mean

Mean (Average)

- **Advantages:**
  - Uses all data values
  - Easy to calculate and widely understood
- **Disadvantages:**
  - Strongly affected by extreme values (outliers)
  - May not represent a typical observation in heavily skewed data

Median

- Arrange the n measurements from the smallest to the largest.
  - If n is odd, median is the middle number.
  - If n is even, median is the mean of the middle two numbers.
- **Example:** Median of $\{1, 3, 3, 6, 7, 8, 9\}$ is 6.
- **Example:** Median of $\{1, 3, 3, 6, 7, 8, 9, 10\}$ is 6.5.

Median

- **Advantages:**
  - Not affected by extreme values
  - Works well for skewed distributions
- **Disadvantages:**
  - Ignores the actual magnitudes of most data values, using only the middle position(s)
  - More stable (robust) to extreme observations.

Mode

- The mode is the measurement that occurs most frequently in the data set.
- **Example:** Mode of $\{8, 7, 9, 6, 8, 10, 9, 9, 5, 7\}$ is 9.
- **Example:** Modes of $\{8, 7, 9, 6, 8, 10, 9, 9, 5, 7, 7\}$ are 7 and 9.

Mode

- **Advantages:**
  - Easy to understand and explain (the “most common” value)
  - Also useful for categorical data (e.g., most sold product)
- **Disadvantages:**
  - May not be unique (there can be more than one mode)
  - Not always representative of the data (especially in small or uniform data sets)

Range

- The range is the difference between the maximum and minimum values in a data set.
- **Example:** Range of $\{1, 3, 3, 6, 7, 8, 9\}$ is $9 - 1 = 8$.

Range

- **Advantages:**
  - Very simple to calculate and understand
  - Provides a quick sense of data spread (max–min)
- **Disadvantages:**
  - Based only on two values; ignores all other data
  - Extremely sensitive to outliers

Deviations

- For a data set $\{x_1, x_2, x_3, \ldots, x_n\}$ each $x_i - \bar{x}$ is called a deviation.
- **Example.** Consider the data set $\{7,\ 9,\ 10,\ 12,\ 13\}$

$$\sum_{i=1}^{5} x_i = 7 + 9 + 10 + 12 + 13 = 51, \qquad \bar{x} = \frac{51}{5} = 10.2.$$

x_i	$x_i - \bar{x}$
7	-3.2
9	-1.2
10	-0.2
12	1.8
13	2.8

- Notice that sum of the deviations is always zero!

Deviations

Sum of the deviations is zero, that is for a data set $\{x_1, x_2, x_3, \ldots, x_n\}$,

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

Proof.

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x}) &= (x_1 - \bar{x}) + (x_2 - \bar{x}) + \ldots + (x_n - \bar{x}) \\ &= (x_1 + x_2 + \ldots + x_n) - (\bar{x} + \bar{x} + \ldots + \bar{x}) \\ &= (x_1 + x_2 + \ldots + x_n) - n\bar{x} \\ &= (x_1 + x_2 + \ldots + x_n) - (x_1 + x_2 + \ldots + x_n) = 0\end{aligned}$$

□

Sample Variance

- The sample variance for a sample of n measurements is equal to the sum of the squared deviations, divided by $(n-1)$.
- The symbol s^2 is used to represent the sample variance.
- If data is $\{x_1, x_2, x_3, \ldots, x_n\}$, then sample variance is

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

  where $\bar{x}$ is the sample mean.
- It indicates how far the data values are spread out around the mean.

Example

- **Example.** Calculate the mean and (sample) variance of the following sample ($n = 5$):

$$7,\ 9,\ 10,\ 12,\ 13$$

- Recall that

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

Example

- **Example.** Calculate the mean and (sample) variance of the following sample ($n = 5$):

$$7,\ 9,\ 10,\ 12,\ 13$$

- **Solution:**

$$\bar{x} = \frac{7 + 9 + 10 + 12 + 13}{5} = \frac{51}{5} = 10.2.$$

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
7	-3.2	10.24
9	-1.2	1.44
10	-0.2	0.04
12	1.8	3.24
13	2.8	7.84
$\sum(x_i - \bar{x})^2$		22.80

Example (cont'd)

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
7	-3.2	10.24
9	-1.2	1.44
10	-0.2	0.04
12	1.8	3.24
13	2.8	7.84
$\sum(x_i - \bar{x})^2$		22.80

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{4}\sum_{i=1}^{5}(x_i - 10.2)^2 = \frac{22.8}{4} = 5.7$$

Sample Standard Deviation

- Since the unit of the sample variance is the square of the unit of the data, we take the square root of the sample variance to return to the original unit.
- This value is called the sample standard deviation.
- It measures the spread of data values around the mean in the original unit of measurement.
- The symbol s is used to represent the sample standard deviation.
- If data is $\{x_1, x_2, x_3, \ldots, x_n\}$, then sample standard deviation is

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

Example

The (sample) standard deviation of

$$7,\ 9,\ 10,\ 12,\ 13$$

is

$$s = \sqrt{s^2} = \sqrt{5.7} \approx 2.3875$$

Interpreting Standard Deviation

- The sample standard deviation measures the typical distance of data values from the mean.
  - A small standard deviation indicates that the data points are close to the mean, implying low variability.
  - A large standard deviation indicates that the data points are spread out widely around the mean, implying high variability.

Properties of Standard Deviation

- Variance and standard deviation are never negative.
- Standard deviation of a data set of the form $\{a, a, a, \ldots, a\}$ is zero.
  - For example, $\{7, 7, 7, 7, 7, 7, 7, 7\}$ has standard deviation 0.
- On the other hand, if standard deviation is zero, then data set should be of the form $\{a, a, a, \ldots, a\}$.

Example

- Consider again the sample

$$7,\ 9,\ 10,\ 12,\ 13.$$

It has mean 10.2 and deviatons are

$$-3.2,\ -1.2,\ -0.2,\ 1.8,\ 2.8$$

- Add 4 to every number in this sample:

$$11,\ 13,\ 14,\ 16,\ 17.$$

New mean is 14.2, so it is also increased by 4, and therefore deviations did not change. Hence variance and standard deviation did not change as well.

Example

- Consider again the sample

$$7,\ 9,\ 10,\ 12,\ 13.$$

It has mean 10.2 and deviatons are

$$-3.2,\ -1.2,\ -0.2,\ 1.8,\ 2.8$$

and $s^2 = 5.7$.

- Multiply every number in this sample by 3:

$$21,\ 27,\ 30,\ 36,\ 39.$$

New mean is $30.6 = 3 \times 10.2$ New deviations are:

$$-9.6,\ -3.6,\ -0.6,\ 5.4,\ 8.4$$

They are also 3 times original deviations. Hence new variance is $9 \times 5.7 = 51.3$ and new standard deviation is $\sqrt{9 \times 5.7} = 3 \times \sqrt{5.7} = 3s$.

Example

- If we multiply every number in this sample by -3. New deviations will be -3 times original deviations. Hence new variance is $9 \times 5.7 = 51.3$ and new standard deviation is $\sqrt{9 \times 5.7} = 3 \times \sqrt{5.7} = 3s$.

Standard Deviation Under a Linear Transformation

Call s for the sample standard deviation of $\{x_1, x_2, \ldots, x_n\}$. Then,

- For any constant c, the sample standard deviation of

$$\{x_1 + c, x_2 + c, \ldots, x_n + c\}$$

is s.

- For any constant b, the sample standard deviation of

$$\{bx_1, bx_2, \ldots, bx_n\}$$

is $|b|s$.

- For any constants b and c, the sample standard deviation of

$$\{bx_1 + c, bx_2 + c, \ldots, bx_n + c\}$$

is $|b|s$.

Mean Under a Linear Transformation

Call $\bar{x}$ for the mean of $\{x_1, x_2, \ldots, x_n\}$. Then,

- For any constants b and c, the mean of

$$\{bx_1 + c, bx_2 + c, \ldots, bx_n + c\}$$

is $b\bar{x} + c$.

Example

- **Example.** A temperature data has average 21.73° and standard deviation 3.84° in Celsius. Convert the data set to Fahrenheit. What are the new mean and standard deviation?
  (Degrees Fahrenheit $= 1.8\times$ Degrees Celsius $+32$. Example: 41° F $= 5^\circ$ C)
- **Solution.** Mean is

$$1.8 \times 21.73 + 32 = 71.114^\circ\text{F}$$

  standard deviation is

$$1.8 \times 3.84 = 6.912^\circ\text{F}$$

Sample Standard Deviation

- **Advantages:**
  - Standard deviation uses all data values (considers the full data set)
  - Standard deviation is expressed in the same units as the original data, unlike variance
- **Disadvantages:**
  - Standard deviation is sensitive to extreme values (outliers can inflate it)

Population Variance and Population Standard Deviation

- The population variance for a data set $\{x_1, x_2, x_3, \ldots, x_n\}$ is

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

- The population standard deviation is

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

- For large n, we have $\sigma \approx s$.

Percentiles

- Let $0 \leq r \leq 100$.
- The r^{th} percentile is a value x such that about $r\%$ of the data are at or below x (and about $(100 - r)\%$ are at or above x).
- **Example (exam scores).** "The 75^{th} percentile of midterm exam scores is 82." This means that: about 75% of the students scored ≤ 82, and about 25% scored ≥ 82.
- $Q_1 = 25^{\text{th}}$ percentile $= 1^{\text{st}}$ quartile $=$ median of the lower 50% of the data
- $Q_2 = 50^{\text{th}}$ percentile $= 2^{\text{nd}}$ quartile $=$ median
- $Q_3 = 75^{\text{th}}$ percentile $= 3^{\text{rd}}$ quartile $=$ median of the upper 50% of the data
- Thus, Q_1, Q_2, and Q_3 divide the data set into four equal parts, each containing 25% of the observations.

Interquartile Range

- We define the interquartile range (IQR) as

$$\text{IQR} = Q_3 - Q_1$$

- **Advantages:**
  - Focuses on the middle 50% of the data, so it is not influenced by extreme values
  - Provides a better picture of data spread for skewed distributions
- **Disadvantages:**
  - Ignores half of the data (the lower 25% and the upper 25%)

Describing Numerical Data Graphically

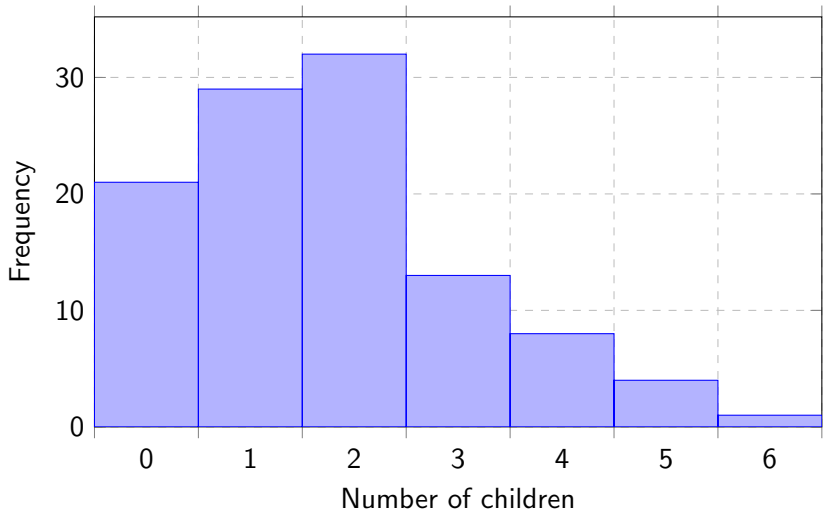
- Key Concept: Distribution
- In statistics, a distribution describes how the values of a variable are spread across possible outcomes.
- It shows which values occur and how frequently, helping us understand the shape, center, and spread of the data.

Histogram

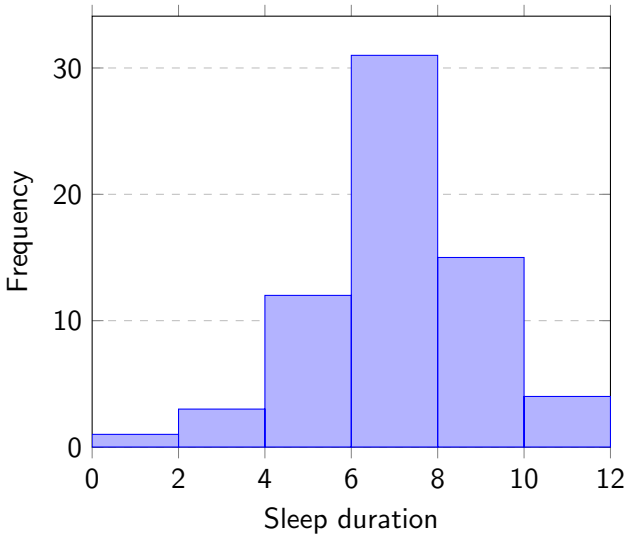
- A histogram is a graph used to display the distribution of a numerical variable.
  - Data are grouped into intervals, called bins.
  - The height of each bar represents the frequency (or relative frequency) of observations within that interval.
  - Histograms help us visualize the shape, center, and spread of the data.

Histogram (Discrete Data)

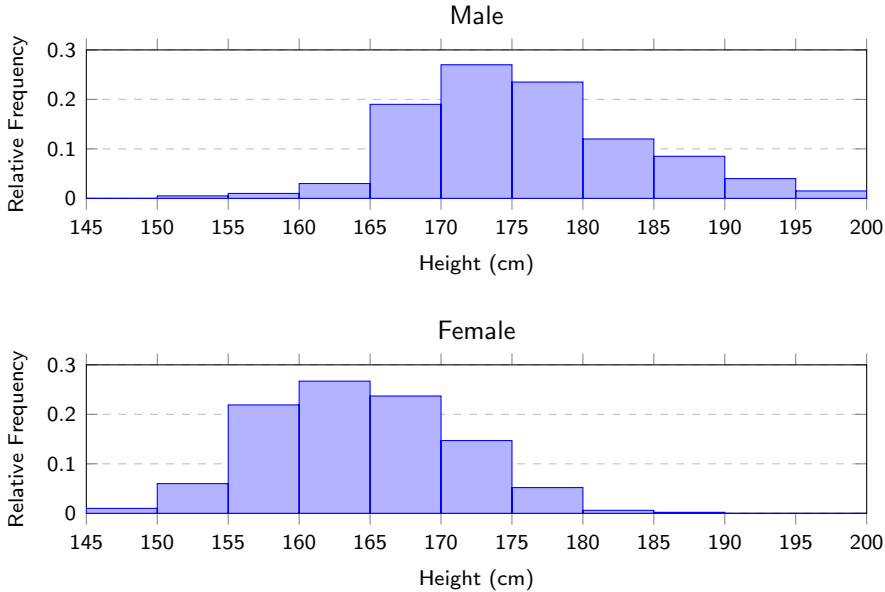
A sample of families.



Histogram (Continuous Data)

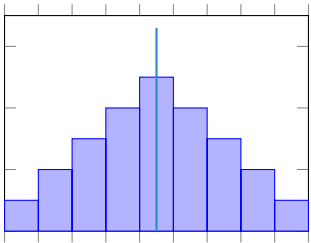


Comparing Histograms



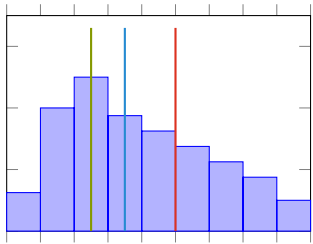
Detecting Skewness by Comparing the Mean and the Median

Symmetric



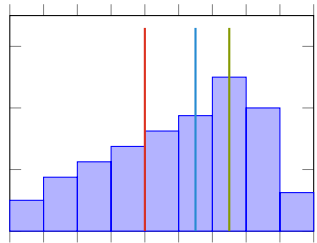
Mode=Median=Mean

Right-skewed (Positive-skewed)



Mode < Median < Mean

Left-skewed (Negative-skewed)



Mean < Median < Mode

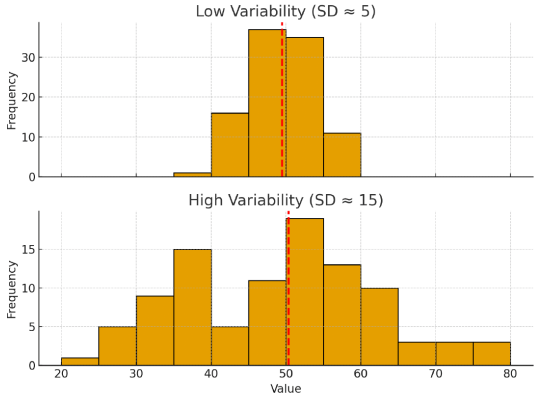
Detecting Skewness by Comparing the Mean and the Median

Sometimes basic examples are useful to understand the reasoning:

Data set	Mean	Median	
$2, 4, 6, 8, 10$	6	6	Symmetric
$0, 0, 0, 0, 10$	2	0	Right-skewed
$0, 10, 10, 10, 10$	8	10	Left-skewed

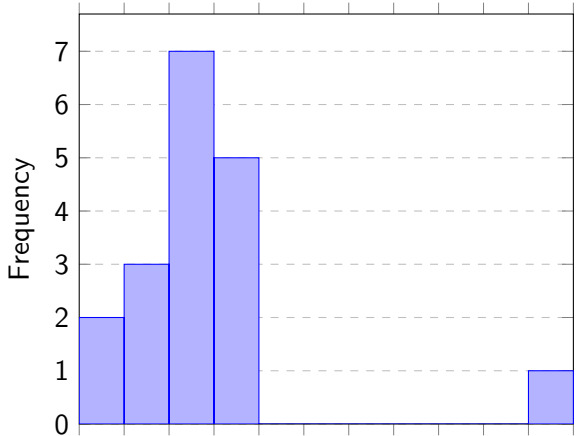
Interpreting Standard Deviation

- Small s: low variability; large s: high variability
- Two data sets with the same means but different standard deviations (SD = Standard Deviation):



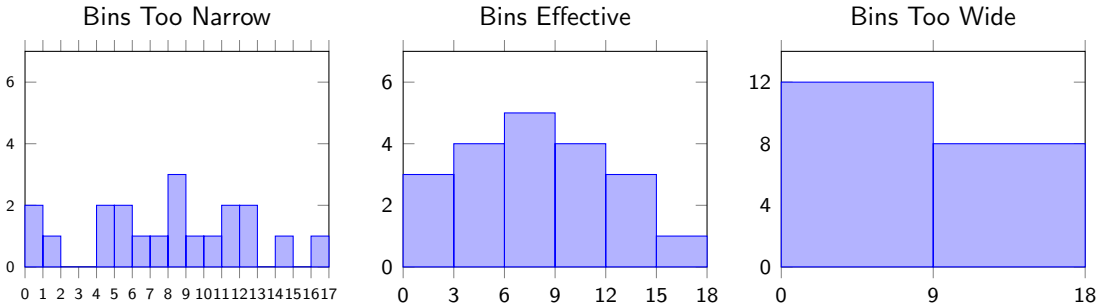
Histogram

- An outlier is a data point that lies far away from the other observations in a dataset. It may indicate an unusual case, or even an error in measurement. Outliers can strongly affect summary statistics (like the mean and standard deviation).



Histogram

- Changing the bin width of a histogram can make the graph easier or harder to interpret.



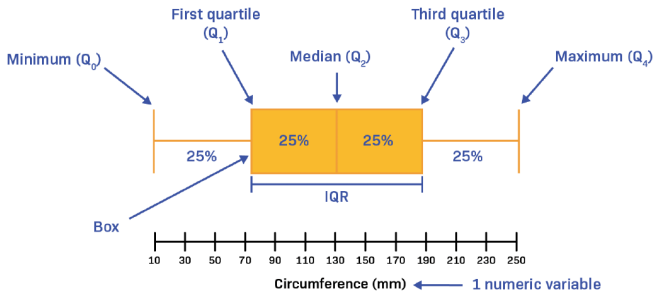
- In the left one, the bin width is too narrow that most bins have only one or two values in them.
- In the right one, the bin width is so wide that the histogram is not informative.
- In the middle one, the histogram bin width is just the right amount, so the distribution is visualized effectively.

Boxplot

- A boxplot is a graphical summary of a data set based on its quartiles.
  - The box represents the interquartile range (Q_1 to Q_3).
  - The line inside the box shows the median (Q_2).
  - The whiskers extend to the minimum and maximum values (excluding outliers).
  - Outliers, if any, are shown as separate points.
- Box plots are less detailed than histograms, but they are widely used because they provide a quick summary of the data's center, spread, skewness, and outliers in a single, compact graph.

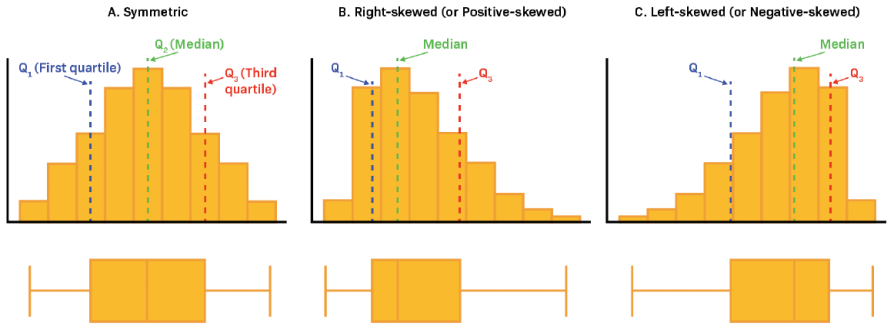
Boxplot

- The anatomy of a boxplot. Boxplots describe data sets using 5 particular numbers: the minimum, first quartile, median, third quartile, and maximum. Median and box help us understand where the data is centered. Interquartile range and range help us understand the variability in the data.



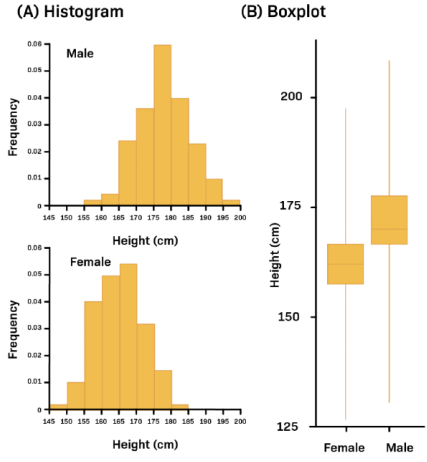
Boxplot

- Characteristic features of a distribution, such as symmetry and skewness, can be visualized with a boxplot.



Boxplot

- Boxplots are useful for making comparisons between two data sets in a compact way.



Chebyshev's Rule

Let $k > 1$ be an integer. At least

$$1 - \frac{1}{k^2}$$

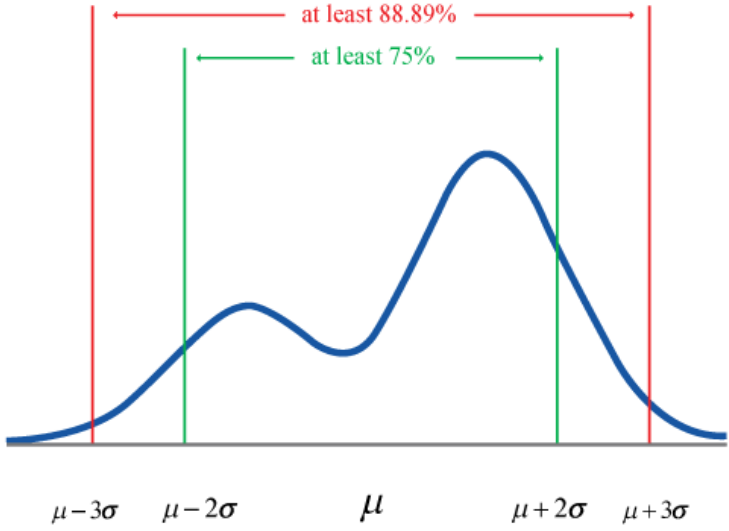
of the data lies betweem

$$\text{mean} - k \times \text{standard deviation} \quad \text{and} \quad \text{mean} + k \times \text{standard deviation}$$

To illustrate,

- $k = 2$: At least 75% of values are within 2 standard deviations of the mean.
- $k = 3$: At least 88.89% of values are within 3 standard deviations of the mean.
- $k = 4$: At least 93.75% of values are within 4 standard deviations of the mean.

Chebyshev's Rule

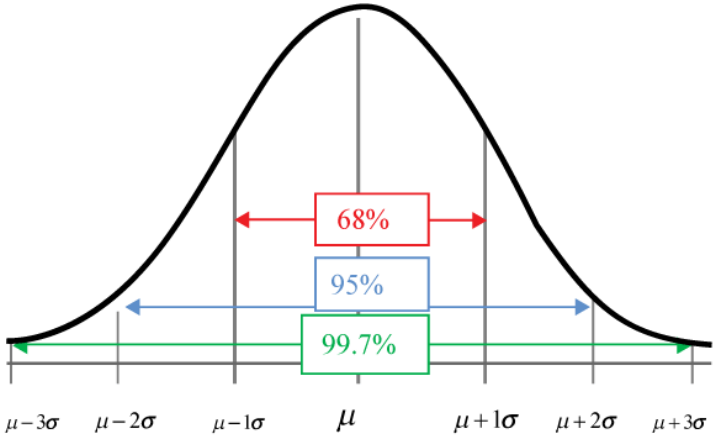


Empirical Rule

For approximately bell-shaped (normal) distributions:

- About 68% of data lies within 1 standard deviation of the mean.
- About 95% lies within 2 standard deviations of the mean.
- About 99.7% lies within 3 standard deviations of the mean.

Empirical Rule



Chebyshev's Rule vs. Empirical Rule

- Chebyshev's Rule is more conservative, since it guarantees a minimum probability, but it applies to any distribution.
- Empirical Rule is more precise, providing approximate probabilities, but it applies only to bell-shaped distributions.