A very important issue in a statistical analysis is the variation within a dataset.

We can quantify the variability with the following measures:

- Range
- Variance and standard deviation
- Interquartile range
The range is the length of an interval containing all of the data, i.e.

\[
\text{Range} = (\text{Maximum value}) - (\text{Minimum value})
\]

**Example:** Consider again the wind speeds of the hurricanes and tropical storms in 2005:

40 40 45 50 65 70 70 70 105 150 155 175

Range = 175 – 40 = 135

- Easy to compute.
- Depends only on the maximum and minimum.
How far from the mean is a typical datapoint? (i.e., how spread out are typical datapoints?)

For a population, the **standard deviation** is defined as

\[
\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}
\]

and the **variance** is

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

**Note:** For a population, we use the population mean, \( \mu \), to compute the sd (standard deviation) and variance.
• Variance is mathematically convenient.

\[ \text{vs.} \]

• Standard deviation is more interpretable; it is in the same units as the data.

\[ \rightarrow \text{Standard deviation is a measure of distance from the mean to a typical datapoint, and the most common measure of spread!} \]

** Example: 2005 storms: ** Consider this to be our population of interest, thus we use \( \mu = 86.25 \).

\[
\sigma^2 = \frac{1}{12} \left[ (40 - 86.25)^2 + (40 - 86.25)^2 + (45 - 86.25)^2 + \ldots + \ldots \right] \\
= \frac{1}{12} \left[ (-46.25)^2 + (-46.25)^2 + (-41.25)^2 + \ldots + \ldots \right] \\
= \frac{1}{12} [25556.25] = 2129.69
\]

Therefore, \( \sigma = \ldots \ldots \ldots \)
In most cases, we only have a sample, not the whole population.

Then we use $\bar{x}$ to estimate $\mu$, and $s$ to estimate $\sigma$.

★ The **sample standard deviation** is

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

★ The **sample variance** is

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

★★ What is the sd (standard deviation) of the sample of August 2005 hurricanes/tropical storms? What type of sampling method is this if August was selected at random and if we are interested in the population of all hurricanes/tropical storms in 2005?

- In practice, use a fancy calculator or computer.
Empirical rule for data with a bell-shaped distribution:

- About 68% of the data will be within 1 s.d. of the mean.
- About 95% will be within 2 s.d. of the mean.
- About 99.7% will be within 3 s.d. of the mean.
Standardized Scores

• How “extreme” is an observation?
• How do we compare observations from different datasets?

* We standardize the data by subtracting its mean and dividing by its standard deviation.

\[ z = \frac{x - \mu}{\sigma} \]

→ We call \( z \) the *standardized score* or *z score*.

• What about z-score for a sample?

\[ z = \frac{x - \bar{x}}{s} \]
Examples:

→ August 2005: Hurricane Katrina \( z = \frac{175 - 86.25}{46.15} = 1.92 \)

→ August 2004: biggest storm, Karl at 145. The population mean and standard deviation for August 2004 was 90.6 and 40.92, respectively. So

\[
z = \frac{145 - 90.6}{40.92} = 1.33
\]

☆ Even after adjusting for the increased variability in 2005, Katrina stands out as more extreme.

→ SAT scores are rescaled to 200-800 to adjust for varying difficulty of the exams.

★★ For the empirical rule, if \( z < -2 \) or \( z > 2 \), the observation is unusual! Only 5% of the observations should have \( z \) values outside \((-2, 2)\).
The median separates the data into two equally-sized groups; half of the observations are above the median, half are below. So the median is the $50^{th}$ percentile.

Let’s generalize this:

- 1% of the data is below the $1^{st}$ percentile value
- 2% is below the $2^{nd}$ percentile
- 5% is below the $5^{th}$ percentile
- 90% is below the $90^{th}$ percentile

→ The 99 percentiles divide the data into 100 groups.
To find the $k^{th}$ percentile:

1. Sort the data.

2. Compute $L = \frac{k}{100} \cdot n$

3. • If $L$ is a whole number, take the average of the $L^{th}$ and $(L+1)^{th}$ values.
• Otherwise, round $L$ up to the next whole number and take that value.
**Examples:** Find the $25^{th}$ percentile.

→ August 2005 storms: 65 105 50 175 40

\[ L = \frac{25}{100} \cdot 5 = \frac{5}{4} = 1.25 \]

So, $25^{th}$ percentile wind speed is the second entry on the sorted list, i.e. 50 mph.

→ August 2004 storms: 40 45 65 70 105 120 135 145

\[ L = \frac{25}{100} \cdot 8 = \frac{8}{4} = 2 \]

So, $25^{th}$ percentile value is \( \frac{45+65}{2} = 55 \text{mph} \).
★ The 25\(^{th}\) percentile is called the 1\(^{st}\) quartile, \(Q_1\), and the 75\(^{th}\) percentile is called the 3\(^{rd}\) quartile, \(Q_3\).

★ The median is the 2\(^{nd}\) quartile.

→ The quartiles divide the data into 4 groups.

★ Interquartile range: a measure of the spread that is less affected by outliers in the data

\[ IQR = Q_3 - Q_1 \]
Boxplot: shows the distribution of the data.

- is a graphical display of centre and spread of the data.
- includes the minimum, $Q_1$, median, $Q_3$ and maximum.
A combination of a boxplot and a steam-and-leaf plot is usually the best way to display quantitative data.

→ For a very large sample, a histogram may be better.

* **Outlier**: an extreme value or highly unusual observation.
  - affect the mean, but not the median
  - affect the standard deviation and the range
  - may affect histograms

- You may want to remove outliers if they are suspected to be incorrect.
**Exploratory Data Analysis:** the use of statistical tools to investigate data sets in order to understand their important characteristics.

→ We explore

- Center
- Variation
- Distribution shape
- Outliers
- Changes over time (Linear regression)
Probability helps us to answer questions such as:

- What is the probability of getting heads when flipping a fair coin?
- What is the probability of rolling a 10 on two dice?
- What is the probability that your tuition will increase next quarter?
- What is the probability that it will rain tomorrow?
- What is the probability that the person sitting next to you will say yes if you ask them on a date?
Definitions:

★ A **simple event** is a possible outcome.
e.g. H or T

★ The **sample space** is the set of all possible simple events.
e.g. \{H, T\}, \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\},
{tuition increases, tuition doesn’t increase}, {rains, doesn’t rain},
{Yes, No}

★ An **event** is a set of simple events.
e.g. \{H\}, \{10\}, \{roll an even number\}
Classical Probability

→ if all simple events are equally likely, each has probability \( \frac{1}{n} \).

→ if event \( A \) is a set of \( s \) simple events then it has probability

\[
P(A) = \frac{s}{n}.
\]

★ ★ Examples:

• \( P(H) = \frac{1}{2} \)
• \( P(\text{roll 5 on one die}) = \frac{1}{6} \)
• \( P(\text{roll an even number on one die}) = \frac{3}{6} = \frac{1}{2} \)
• \( P(\text{roll 10 on 2 dice}) = \frac{3}{36} = \frac{1}{12} \)
• \( P(2\text{H on 3 coin flips}) = ? \)
Other definitions of probability:

- **Subjective (Bayesian) Probability**: derived from our personal knowledge and perspective.

- **Relative frequency**: probability of an event is the long-run average fraction of times it occurs.
  → This property of probability is commonly referred to as the law of large numbers.
Key Concepts!!!

- Range
- **Standard deviation/Variance**
- Quartiles
- Boxplot
- Outlier
- Probability