Basic Statistical Concepts

Aymen Negadi

What is Statistics?

Statistics is the science of collecting, organizing, analyzing, interpreting, and presenting data. It helps us understand patterns in data and make decisions (or inferences) under uncertainty. In **Machine Learning**, statistics is essential because models learn from data distributions, variability, and relationships between variables.

10 Basic Statistical Concepts (Definitions + Examples)

1. **Mean (Average)**
   Definition: The mean is the average value of a dataset, computed by summing all values and dividing by the number of values.
   Example: Scores: $10, 12, 14, 14, 20$.

$$\mu = \frac{10 + 12 + 14 + 14 + 20}{5} = 14$$

   So, the mean score is 14.

2. **Median**
   Definition: The median is the middle value when the data is sorted.
   Example (odd count): Ages: $18, 20, 22, 25, 30$. The median is 22.
   Example (even count): Values: $18, 20, 22, 25$.

$$\text{Median} = \frac{20 + 22}{2} = 21$$

3. **Mode**
   Definition: The mode is the value that appears most frequently in a dataset.
   Example: Daily sales: $5, 7, 7, 7, 10, 12$.
   The mode is 7 (it appears most often).

4. **Standard Deviation**
   Definition: Standard deviation measures how spread out the data is around the mean.
   Example (intuition): Two classes have the same mean (10), but different spread:

$$A = \{9, 10, 11\}, \quad B = \{2, 10, 18\}$$

   Class B has a larger standard deviation because its values are more dispersed.

5. **Variance**
Definition: Variance is the average squared distance from the mean. It is the square of the standard deviation.
Example: If the standard deviation is $\sigma = 4$, then:
$$\text{Var} = \sigma^2 = 4^2 = 16$$

6. **Probability**
Definition: Probability is the likelihood that an event occurs, expressed between 0 and 1.
Example: A fair die has 6 outcomes. Probability of rolling a 3:
$$P(3) = \frac{1}{6} \approx 0.167$$

7. **Distributions**
Definition: A distribution describes the possible values of a random variable and how often they occur.
Examples:
   - **Normal distribution:** Many natural measurements (like human height) tend to cluster around a mean with fewer extreme values.
   - **Uniform distribution:** All outcomes are equally likely (e.g., selecting a random integer from 1 to 10).

8. **Hypothesis Testing**
Definition: Hypothesis testing is a procedure to evaluate a claim about a population using sample data.
Example: Claim: The average battery life is 10 hours.
$$H_0 : \mu = 10 \quad \text{(null hypothesis)}, \qquad H_1 : \mu \neq 10 \quad \text{(alternative hypothesis)}$$
Using sample data, we decide whether to reject H_0 or not.

9. **Correlation**
Definition: Correlation measures the strength and direction of a linear relationship between two variables.
Examples:
   - **Positive correlation:** Temperature $\uparrow \Rightarrow$ Ice cream sales $\uparrow$
   - **Negative correlation:** Study time $\uparrow \Rightarrow$ Free time $\downarrow$

   Correlation values are typically between -1 and $+1$:
   $$r = +1 \text{ (perfect positive)}, \quad r = 0 \text{ (no linear relation)}, \quad r = -1 \text{ (perfect negative)}$$
   Note: Correlation does not imply causation.

10. **Regression**
Definition: Regression models the relationship between a dependent variable (target) and one or more independent variables (features), often for prediction.
Example (linear regression): Predict house price from size:
$$\text{Price} = 50000 + 200 \times (\text{Size in } m^2)$$
If the size is $100\,m^2$:
$$\text{Price} = 50000 + 200 \times 100 = 70000$$

Quick Link to Machine Learning

- Mean and standard deviation are used for **feature scaling** (normalization/standardization).
- Distributions help us **understand data behavior** and detect outliers.
- Correlation can help with **feature selection**.
- Regression is a core method in **supervised learning** for prediction.