PODS Lab 4: Probability Distributions, Central Tendency, and Dispersion

Hamza Alshamy

Center for Data Science, NYU
ha2486@nyu.edu

02/14/25

1. **Probability Distributions**
2. **Central Tendency and Ergodicity**
3. **Dispersion**

What Are Probability Distributions?

- **Probability distributions** describe how probabilities are assigned to different possible outcomes of a random variable.
- They are **governed by parameters**, which control their shape and properties.

Why Are They Relevant?

1. **Describing Random Variables:** Distributions characterize the behavior of random variables, assigning probabilities to their possible values.
2. **Probabilistic Modeling:** If we have data, we can use distributions to model uncertainty and make predictions.

Example of Probability Distributions

Continuous: Normal Distribution

Normal Distribution Equation:

$$P(\mathfrak{X} = x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Example of Probability Distributions

Continuous: Normal Distribution

Normal Distribution Equation:

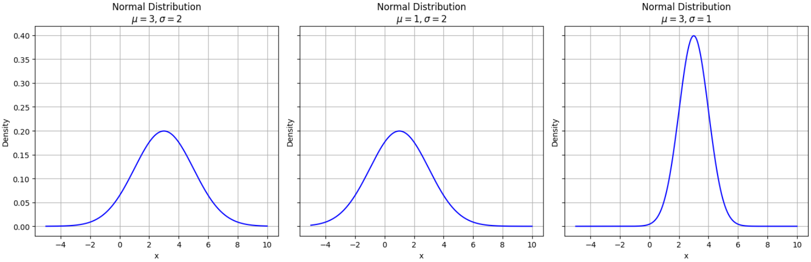
$$P(\mathfrak{X} = x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- **In English:** "The probability of the random variable $\mathfrak{X}$ taking on the value x given the parameters, the mean μ and standard deviation σ"
- **Logic:** Many natural phenomena (e.g., heights, test scores) follow a normal distribution.
- **Parameters:**
  1. μ **(mean):** Determines the center of the distribution.
  2. σ **(Standard Deviation):** Controls the spread of the distribution.

Example of Probability Distributions

Continuous: Normal Distribution

- Notice how the shape of the normal distributions differs based on the **parameters** (μ, σ)



Example of Probability Distributions

Discrete: Poisson Distribution

Poisson Distribution Equation:

$$P(\mathfrak{X} = k; \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$$

- **In English:** "The probability of the random variable $\mathfrak{X}$ taking on the value k given the rate parameter λ."
- **Logic:** Models the number of events occurring in a fixed interval of time or space when events happen independently at a constant average rate.
- **Parameter:**
  1. λ **(Rate Parameter):** Represents the expected number of occurrences in the given interval.
- **Example:** Model the number of earthquakes that will occur given a time interval.

Random Variables and Probability Distributions

Example: Measuring IQ

- Random variables can follow specific probability distributions.
- **Example:** Let $\mathfrak{X}$ represent IQ scores.

$$\mathfrak{X} \sim \mathcal{N}(\mu = 100, \sigma = 15)$$

- **In English:** "The random variable $\mathfrak{X}$ follows a normal distribution with mean $\mu = 100$ and standard deviation $\sigma = 15$."
- **Interpretation:**
  1. Most IQ scores will be close to 100, with variation controlled by $\sigma = 15$.
  2. About 68% of IQ scores fall within one standard deviation ($85 \leq \mathfrak{X} \leq 115$).
  3. Approximately 95% of IQ scores lie within two standard deviations ($70 \leq \mathfrak{X} \leq 130$).

Central Tendency, Typical Value, Average

- **Definition:** Measures of central tendency describe a typical or central value of a dataset.
- **Relevance:**
  - If you know nothing about the data, the best guess for an unknown value is often the *average*.
  - Helps summarize data with a single representative value.
- **Main Measures of Central Tendency:**
  1. **(Arithmetic) Mean:** The average of all values.
  2. **Median:** The middle value when data is ordered.
  3. **Mode:** The most frequently occurring value.

(1) The Arithmetic Mean

Best used when:

1. (Approximately) Normal or symmetrical data
2. No (or few) extreme values – Mean is not very robust
3. Not heavily skewed – Mean will be dragged towards the tail

Example:

$$\text{Set} = \{9, 3, 300, 8, 7, 10, 8, 5\}$$

- Calculation of the Mean:

(1) The Arithmetic Mean

Best used when:

1. (Approximately) Normal or symmetrical data
2. No (or few) extreme values – Mean is not very robust
3. Not heavily skewed – Mean will be dragged towards the tail

Example:

$$\text{Set} = \{9, 3, 300, 8, 7, 10, 8, 5\}$$

- Calculation of the Mean:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{9+3+300+8+7+10+8+5}{8} = \frac{350}{8} = \boxed{43.75}$$

- The mean is dragged towards the outlier (300), making it not representative of most values in the set.

(2) The Median

- **Definition:** The middle value when data is ordered.
- **Best used when:**
  1. Extreme values – Median is more robust than the mean.
  2. Data is skewed (not symmetrical).

Example:

$$\begin{aligned}\text{Set} &= \{9, 3, 300, 8, 7, 10, 8, 5\} \\ &= \{3, 5, 7, 8, 8, 9, 10, 300\} \qquad \text{(Ordered)}\end{aligned}$$

- Calculation of the Median:

$$\text{Median} = \frac{8+8}{2} = \boxed{8}$$

- Notice how the median is not affected by the extreme value (300), unlike the mean.
- Representative of the set!

(3) The Mode

- **Definition:** The most frequently occurring value in a dataset.
- **Best used when:**
  1. With categorical data.
  2. When numbers are used as labels.

Example: Diagnosing Disorders in DSM-IV

$$\text{Disorders} = \{\text{Anxiety, Depression}\}$$
$$\text{Disorder Codes} = \{300.00, 311.00\}$$

- **Incorrect Approach (Mean):**

$$\frac{300.00 + 311.00}{2} = 305.5 \quad \text{(Code for Opioid Abuse)}$$

(3) The Mode

- **Definition:** The most frequently occurring value in a dataset.
- **Best used when:**
  1. With categorical data.
  2. When numbers are used as labels.

Example: Diagnosing Disorders in DSM-IV

$$\text{Disorders} = \{\text{Anxiety, Depression}\}$$

$$\text{Disorder Codes} = \{300.00, 311.00\}$$

- **Incorrect Approach (Mean):**

$$\frac{300.00 + 311.00}{2} = 305.5 \quad \text{(Code for Opioid Abuse)}$$

  - The mean does not represent the state of the patients.
- **Correct Approach (Mode):**
  - With categorical data, the most frequent value (mode) is used to represent the typical case.

Summary of Measures of Central Tendency

Given Set:

$$\begin{aligned} \text{Set} &= \{9, 3, 300, 8, 7, 10, 8, 5\} \\ &= \{3, 5, 7, 8, 8, 9, 10, 300\} \qquad \text{(Ordered)} \end{aligned}$$

Summary Table:

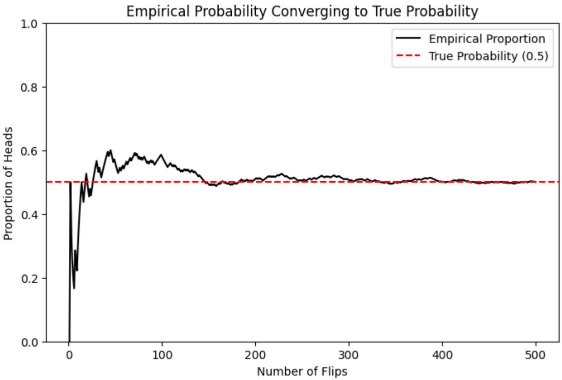
	Mean	**Median**	**Mode**
Value	43.75	8	8

- Measures of central tendency allow us to summarize data with a single representative value.
- The **median is more robust than the mean** because it is not affected by extreme values.
- The **mode is best suited for categorical data**, where taking an average does not make sense.

Note: The Mean and Ergodicity

Law of Large Numbers (LLN):

- Given enough samples, the sample mean $\bar{x}$ converges to the expected value $\mathbb{E}[X]$.
- **However, this only holds if the system is ergodic!**
  - Otherwise, the mean is not very meaningful.



Note: The Mean and Ergodicity (cont.)

What is Ergodicity?

- The average of the results obtained from a large number of independent random samples converge to the true value
- The measure of an individual (one coin flip) over time is predictive of the ensemble average (1000 coin flips)

Assumptions:

1. **Stationary:** Statistical Properties are constant over time
   - The mean, variance, and probability distribution of the system remain consistent.
2. **State Space Convergence:** The System Must Visit All Possible States
   - Over time, air molecules will spread everywhere in the room.
3. **Non-Determinism:** Future events are independent of past events.

Ergodicity: Examples

Ergodic: Roulette (Assume it's fair)

- **Total pockets:** 37 $\rightarrow$ Read: 18, Black: 18, Green: 1
- $P(red) = \frac{18}{37}, P(black) = \frac{18}{37}, P(green) = \frac{1}{37}$
- Results are about the same if you spin 1000 roulettes at the same time or soon one roulette 1000 times

Non-ergodic: Russian Roulette

- Total chambers in a revolver: 6 and 5 are empty
- **60 people playing at the same time:** Survival rate is $\frac{5}{6} \approx 83.4\%$
- **One person playing 60 times:** Survival rate is $(\frac{5}{6})^{60} \approx 1.7\%$

A Second Way to Characterize Data: Dispersion

- **Central Tendency:** If I don't know anything about the data, which single value best represents a typical value?
- **Dispersion:** How much the data deviates from the center.
- **Main Measures of Dispersion:**
  1. **Standard Deviation** (σ) – Least robust (sensitive to outliers).
  2. **Mean Absolute Deviation (MAD)** – More robust than standard deviation.
  3. **Median Absolute Deviation (MeAD)** – Most robust (resistant to extreme values).

(1) Standard Deviation (σ)

- **Arriving at the Standard Deviation (SD):**

$$\frac{1}{n}\sum_i (x_i - \bar{x}) \qquad (\text{Always } = 0)$$

$$\frac{1}{n}\sum_i (x_i - \bar{x})^2 \qquad (\text{Better, but large deviations has large effects})$$

$$\sigma = \boxed{\sqrt{\frac{1}{n}\sum_i (x_i - \bar{x})^2}} \qquad (\text{Taking the square root mitigates that it})$$

- **Interpretation:**
  - **Low SD:** Data points are clustered close to the mean
  - **High SD:** Data points are more dispersed and further away from the mean
- **Issue:** Influenced by outliers because of squaring!

Effect of an Outlier on Standard Deviation

Example: Without an Outlier

$$s = \{5, 6, 7, 8, 9\}, \qquad \text{Mean = 7}$$

- **Squared deviations:**

$$\{(5-7)^2, (6-7)^2, (7-7)^2, (8-7)^2, (9-7)^2\} = \{4, 1, 0, 1, 4\}$$

- **Variance:** $\frac{4+1+0+1+4}{5} = 2 \implies \sigma = \sqrt{2} \approx \boxed{1.41}$

Example: With an Outlier (Changing 9 to 99)

$$s = \{5, 6, 7, 8, 99\}, \quad \text{Mean =25}$$

- **Squared deviations:**

$$\{(5-25)^2, (6-25)^2, \ldots, (99-25)^2\} = \{400, 361, 324, 289, 5476\}$$

- **Variance:**

$$\frac{400+361+324+289+5476}{5} = 1470 \implies \sigma = \sqrt{1470} \approx \boxed{38.36}$$

(2) Mean Absolute Deviation (MAD)

- **Logic:** Squaring values in standard deviation **magnifies deviations**, making it sensitive to outliers.
- **Instead:** Use absolute values to measure dispersion **without over-emphasizing large deviations**.

Formula:

$$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$$

- **More robust** than standard deviation because outliers have less influence.
- Unlike standard deviation, MAD does not disproportionately weight large deviations.

(3) Median Absolute Deviation (MeAD)

- **Logic:** In Mean Absolute Deviation (MAD), we use the **mean**, which is sensitive to outliers.
- **Instead,** use the **median**, which is **more robust** to extreme values.

Formula:

$$\text{MeAD} = \text{Median}\left(\left|x_i - \text{Median}(X)\right|\right)$$

- **Most robust** measure of dispersion – outliers have **minimal influence**.
- Often used in **non-parametric statistics** where distributions may be skewed.

Summary of Measures of Dispersion and Central Tendency

Central Tendency

1. Mean
2. Median
3. Mode

Dispersion

1. Standard Deviation
2. Mean Absolute Deviation
3. Median Absolute Deviation

Example Set:

$$s = \{3, 6, 7, 8, 8, 10, 12, 25\}$$

Computed Central Tendencies:

Measure	Value
Mean	9.88
Median	8
Mode	8

Computed Dispersion Measures:

Measure	Value
Standard Deviation	6.67
Mean Absolute Deviation	4.59
Median Absolute Deviation	2