📘 Day 10 – Distributions + Advanced Hypothesis Testing

🔷 1. Probability Distributions

A **distribution** describes how data values are spread.

Two main types:

- **Discrete** → specific values (0,1,2,...)
- **Continuous** → any value in a range

1.1 Discrete Distributions

- **Bernoulli** → Single trial, success (1) or failure (0).

  $$P(X=1)=p,\ P(X=0)=1-pP(X=1)=p,\ P(X=0)=1-p$$

  Example: Coin toss (H=1, T=0).

- **Binomial** → Repeated Bernoulli trials (n trials).

  $$P(X=k)=(nk)pk(1-p)n-kP(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$$

  Example: Probability of 4 heads in 10 tosses.

- **Poisson** → Number of events in fixed interval (rare events).

  $$P(X=k)=\lambda ke-\lambda k!P(X=k)=\frac{\lambda^k e^{-\lambda}}{k!}$$

  Example: Calls at a call center per hour.

- **Geometric** → Number of trials until first success.
- **Negative Binomial** → Trials until r successes.

- **Uniform (discrete)** → All outcomes equally likely.

1.2 Continuous Distributions

- **Normal (Gaussian)**

  Bell curve, symmetric, mean=median=mode.

  $$f(x) = 1\sigma 2\pi e - (x - \mu)22\sigma 2 f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

  Example: Heights, test scores.

- **Exponential**

  Time between events (waiting time).

  $$f(x) = \lambda e - \lambda x,\ x \geq 0 f(x) = \lambda e^{-\lambda x},\ x \geq 0$$

  Example: Time between earthquakes.

- **Uniform (continuous)**

  Equal probability in a range [a, b].

- **t-distribution**

  Used for small sample hypothesis testing. Tails fatter than normal.

- **χ^2 distribution**

  Sum of squared normal variables. Used in Chi-square test.

- **F-distribution**

  Ratio of two variances. Used in ANOVA.

1.3 Central Limit Theorem (CLT)

No matter population distribution, as sample size grows (n > 30):

- Sampling distribution of mean → **Normal**.
- Mean = μ, SD = σ/√n.

👉 This is why Z/t tests work!

🔷 2. Advanced Hypothesis Testing

- **Chi-Square Test (χ^2)**

  Tests association between categorical variables.

  $$\chi 2 = \sum (O - E) 2 E \chi^2 = \sum \frac{(O - E)^2}{E}$$

  O = observed, E = expected.

📍 Example: Is gender independent of voting preference?

- **ANOVA (Analysis of Variance)**

  Compares means of 3+ groups.

  - H_0: All group means equal.
  - H_1: At least one mean differs.

    Test statistic → F-distribution.

📍 Example: Do 3 diets give same average weight loss?

- **Non-parametric tests** (no normality assumption):
  - Mann-Whitney U test (2 groups).
  - Kruskal-Wallis test (3+ groups).
  - Wilcoxon signed-rank test (paired data).

🔷 3. Correlation (Beyond Pearson)

- **Pearson's r** → linear correlation.
- **Spearman's rank correlation (ρ)** → based on rank (good for monotonic relationships).
- **Kendall's Tau (τ)** → rank concordance measure.

🔷 4. Regression Assumptions

When using Linear Regression:

1. Linearity → relation is linear.
2. Independence → errors not correlated.
3. Homoscedasticity → equal variance of errors.
4. Normality of errors.
5. No multicollinearity (independent variables not highly correlated).

📍 If assumptions fail → use transformations, regularization, or non-parametric models.

🔷 5. Effect Size & Power Analysis

- **Effect size** → strength of a relationship.
  - Cohen's d (difference between two means in SD units).
  - η² (eta squared) in ANOVA.
- **Statistical Power** → probability of detecting a true effect.

  $Power = 1 - \beta Power = 1 - \beta$

  Higher power → lower risk of Type II error.

  Desired power = 0.8 (80%).

✅ Summary of Day 10

- Distributions → discrete (Bernoulli, Binomial, Poisson) & continuous (Normal, Exponential, t, χ², F).
- Central Limit Theorem.
- Hypothesis testing advanced → Chi-Square, ANOVA, Non-parametric.
- Correlation extensions → Spearman, Kendall.
- Regression assumptions.
- Effect size & power analysis.
