Categorical Distribution

1. What is the Categorical Distribution?

- **Generalises** the Bernoulli distribution from 2 categories to K categories.
- Models a *single* trial where the outcome is exactly one of K mutually-exclusive classes.
- Two common notations:
  - *Index form*: random variable $Y \in \{1, \dots, K\}$.
  - *One-hot form* ($\boldsymbol{X}$): vector with exactly one 1 and the rest 0. **We use this form throughout.**

$$\boldsymbol{X} = (X_1, \dots, X_K)^{\mathsf{T}}, \quad X_i \in \{0, 1\}, \quad \textstyle\sum_i X_i = 1.$$

Let $\boldsymbol{p} = (p_1, \dots, p_K)^{\mathsf{T}}$ with $p_i > 0$ and $\sum_i p_i = 1$.

2. Probability Mass Function (PMF)

$$\boxed{P(\boldsymbol{X} = \boldsymbol{x}) \;=\; \prod_{i=1}^{K} p_i^{x_i}, \qquad \text{(one-hot } \boldsymbol{x}\text{)}.}$$

Why it works – "switch" intuition. Because $\boldsymbol{x}$ is one-hot, exactly one $x_j = 1$ and all others $= 0$:

$$\prod_i p_i^{x_i} = p_j^1 \prod_{i \neq j} p_i^0 = p_j,$$

so the product "switches on" the sole probability matching the 1 entry—identical in spirit to the Bernoulli PMF when $K = 2$.

3. Worked Example

Suppose $K = 3$ categories ("red", "blue", "green"):

$$\boldsymbol{p} = \begin{pmatrix} 0.2 \\ 0.6 \\ 0.2 \end{pmatrix}, \qquad \boldsymbol{x} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \text{ (blue chosen)}.$$

Then

$$P(\boldsymbol{X} = \boldsymbol{x}) = 0.2^0 \, 0.6^1 \, 0.2^0 = 0.6.$$

4. Expectation (Mean Vector)

$$\mathbb{E}[\boldsymbol{X}] \;=\; \sum_{\boldsymbol{x}} \boldsymbol{x}\, P(\boldsymbol{X}=\boldsymbol{x}) \;=\; \boldsymbol{p}.$$

Interpretation: over many trials, the "hot 1" spends a proportion p_i of the time in slot i.

5. Covariance Matrix

$$\Sigma = \mathrm{Cov}(\boldsymbol{X}) = \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}] - \mathbb{E}[\boldsymbol{X}]\,\mathbb{E}[\boldsymbol{X}]^{\mathsf{T}} = \mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^{\mathsf{T}}.$$

Break-down.

- $\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}]$: For a one-hot $\boldsymbol{X}$ this is a matrix with a single 1 on the diagonal position that was chosen. Taking expectation over all possibilities $\Rightarrow \mathrm{diag}(\boldsymbol{p})$.
- Off-diagonal entries are *negative*. When one category turns "on", all others must be 0, so they move in opposite directions.

$$\Sigma = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \dots & -p_1p_K \\ -p_2p_1 & p_2(1-p_2) & \dots & -p_2p_K \\ \vdots & \vdots & \ddots & \vdots \\ -p_Kp_1 & -p_Kp_2 & \dots & p_K(1-p_K) \end{pmatrix}.$$

- Symmetric; every row and column sums to 0.
- Variance of each category appears on the diagonal; covariances (always ≤ 0) off the diagonal.

6. Special Case – Bernoulli Inside

If $K=2$ and $\boldsymbol{p}=(p,\,1-p)$ then

$$P(\boldsymbol{X}=\boldsymbol{x}) \;=\; p^{x_1}(1-p)^{x_2},$$

which is exactly the Bernoulli PMF written in one-hot form.

7. Where is it used?

- **Classification** targets (`one-hot` labels).
- **Generative models** (e.g. predicting next token).
- **Reinforcement learning**: policy selects an action among K possibilities.

8. Visualisation

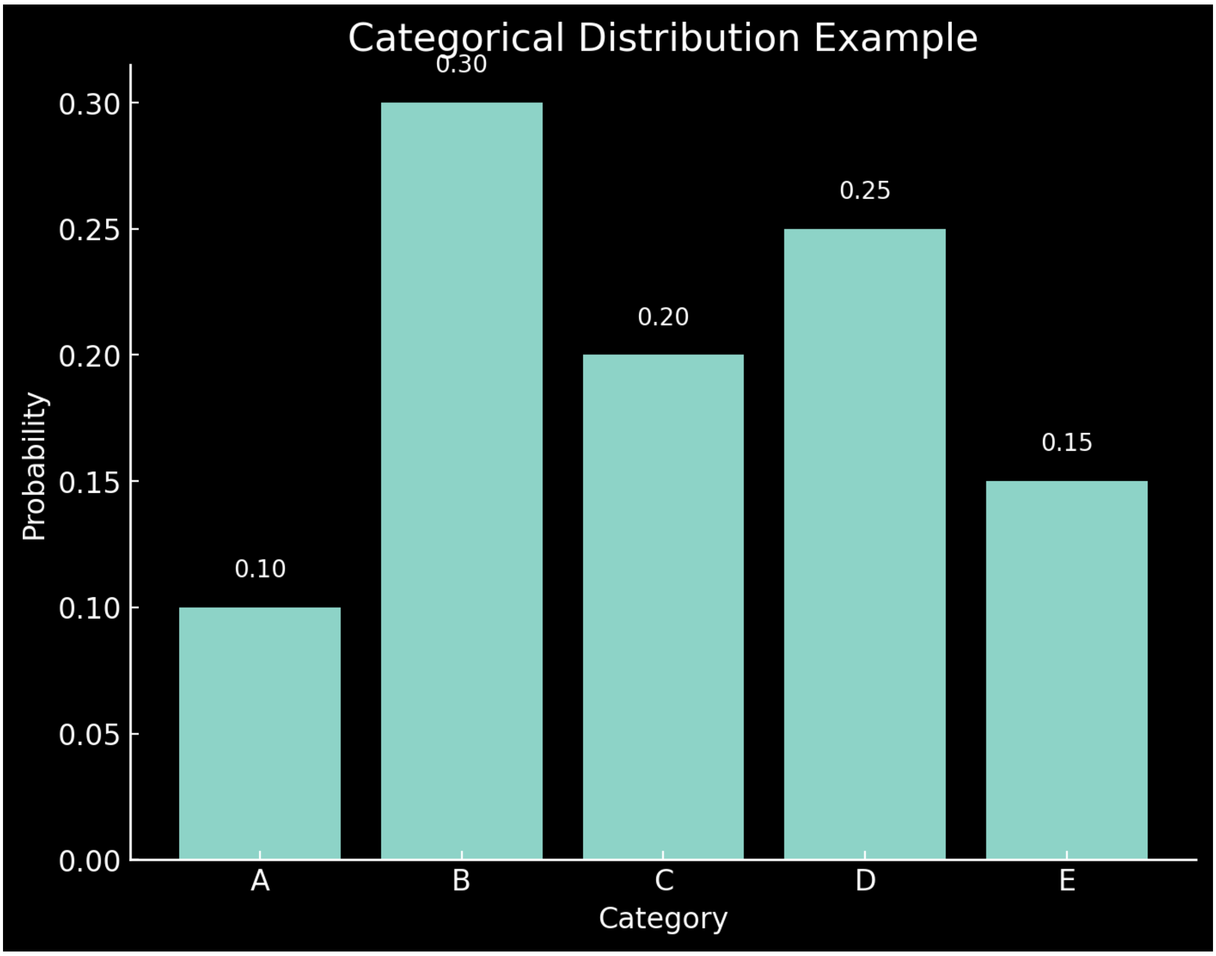

Figure 1: Categorical Distribution - taken from video

9. Summary

- Categorical($\boldsymbol{p}$) models a single draw from K classes.
- PMF: $\prod_i p_i^{x_i}$.
- Mean vector: $\boldsymbol{p}$.
- Covariance: $\operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top$.
- Reduces to Bernoulli when $K = 2$.