**1. Understanding and Writing a Linear Statistical Model for Categorical Predictor Variables**

A **linear model** relates a response variable to one or more predictors. When predictors are **categorical**, they represent groups or levels (like treatment types).

**Example:** Suppose you’re testing the effect of three treatments (A, B, and C) on tumor size in mice. Here, Treatment is a **categorical predictor** with three levels.

**Model:**

Yi=β0+β1XB,i+β2XC,i+ϵi

Where:

* Yi ​: response variable (e.g., tumor size),
* XB,i, XC,i ​: dummy variables (1 if the observation is in that group, 0 otherwise),
* β0: mean response for baseline category (Treatment A),
* β1: difference between B and A,
* β2: difference between C and A,
* ϵi ​: random error.

You don't include all 3 dummies to avoid **perfect multicollinearity** (dummy variable trap); one category is the reference.

**2. Constructing an Analysis of Variance (ANOVA) Table for Testing Hypotheses**

**ANOVA** tests whether the means of multiple groups are significantly different.

**Steps:**

1. **Null Hypothesis (H₀):** All group means are equal.
2. **Alternative Hypothesis (H₁):** At least one group mean is different.
3. **Partition total variability**:
   * **SS\_Total = SS\_Between + SS\_Within**
   * **SS\_Between**: variability due to group differences
   * **SS\_Within (Residual)**: variability within groups

**ANOVA Table Structure:**

| **Source** | **DF** | **SS** | **MS = SS/DF** | **F = MS\_between / MS\_within** |
| --- | --- | --- | --- | --- |
| Between Groups | k - 1 | SS\_Between | MS\_Between | F-statistic |
| Within Groups | N - k | SS\_Within | MS\_Within |  |
| Total | N - 1 | SS\_Total |  |  |

If **F is large**, it suggests significant differences among group means.

**3. Experimental Design – Independent Replication**

* **Experimental Unit**: The smallest unit to which a treatment is independently applied.
* **Measurement Unit**: The unit on which the measurement is taken.

**Importance**: Misidentifying these can **inflate degrees of freedom**, leading to **pseudo-replication** and invalid statistical conclusions.

**Example:** You apply a drug to a **mouse (experimental unit)** and measure **tumor volume in multiple tissue sections (measurement units)**. Treating each tissue measurement as an independent replicate would be incorrect.

**4. Experimental Design – Randomization and Blocking**

* **Randomization**: Randomly assign treatments to experimental units.
  + **Why**: Prevents bias, ensures that unknown confounding factors are equally distributed.
  + **How**: Use random number generators or software tools.
* **Blocking**: Group similar experimental units into **blocks** and randomize treatments **within blocks**.
  + **Why**: Reduces variability due to known confounding variables (e.g., batch effects).
  + **Example**: If animals come from different litters, treat each litter as a block.

**Summary Table**

| **Concept** | **Purpose** | **Example** |
| --- | --- | --- |
| Categorical Predictor in Linear Model | Model group effects | Treatment groups A, B, C in a clinical trial |
| ANOVA Table | Test mean differences | Compare treatment effects across groups |
| Independent Replication | Correctly identify experimental units | One mouse per treatment = 1 replicate |
| Randomization & Blocking | Control variability and bias | Randomize within blocks like animal litter or batch |

| **Concept** | **Example in Code** |
| --- | --- |

|  |  |
| --- | --- |
| Categorical predictor | treatment in lm() |

|  |  |
| --- | --- |
| ANOVA Table | anova(model) |

|  |  |
| --- | --- |
| Experimental unit | Each **mouse** gets one treatment |

|  |  |
| --- | --- |
| Measurement unit | Tumor volume per mouse |

|  |  |
| --- | --- |
| Blocking | + litter in the model |

|  |  |
| --- | --- |
| Randomization | (use sample()) |

| **Concept** | **Description** | **R Function / Code Example** |
| --- | --- | --- |

|  |  |  |
| --- | --- | --- |
| Post-hoc (Tukey) | Compare group means after ANOVA | TukeyHSD(aov(...)) |

|  |  |  |
| --- | --- | --- |
| Independent Replication | Identify true replicates | One mouse = one replicate |

|  |  |  |
| --- | --- | --- |
| Randomization | Avoid bias in treatment assignment | sample() |

|  |  |  |
| --- | --- | --- |
| Blocking | Control variability across known groups | lm(y ~ treatment + block, ...) |