Categorical Data

Multiple variables: Y response, $X_1, \ldots, X_p$ predictors.

- Y is a discrete random variable.
- ϵ cannot be normal.
- $\min E(Y - X\beta)$ is not guaranteed to be discrete.

Univariate Y: discrete distributions.

2 types of categorical variables

1. **Ordinal**: values in the support of Y are ordered.
    - e.g. Letter grades in exam: $A > B > C > D$
    - e.g. Satisfaction survey (Likert scale): Excellent $>$ Good $>$ Neutral $>$ Bad $>$ Terrible
2. **Nominal**: no ordering.
    - e.g. PIN codes
    - e.g. voting preferences
    - e.g. transport taken to work
    - e.g. color

Y: Categorical.
X: some continuous & some categorical.

Binary

2 categories (nominal/ordinal same)
e.g., 0-1, S-F, H-T
$P(\text{success}) = p$.
Assume $Y_1, \ldots, Y_n$ are independent and have the same distribution.

$$Y = \sum_{i=1}^{n} Y_i \sim \text{Bin}(n, p)$$

Maximum Likelihood Estimator for p:

$$\hat{p}_{MLE} = \frac{\sum y_i}{n}$$

- $E(\hat{p}_{MLE}) = p$ (unbiased)
- $\text{Var}(\hat{p}_{MLE}) = \frac{p(1-p)}{n}$

- $s.e.(\hat{p}_{MLE}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Convergence Properties:

- $\hat{p} \xrightarrow{P} p$ (Weak Law of Large Numbers)
- $\hat{p} \xrightarrow{a.s.} p$ (Strong Law of Large Numbers)
- Central Limit Theorem (CLT):

$$\sqrt{n}\frac{\hat{p}-p}{\sqrt{p(1-p)}} \Rightarrow N(0,1)$$

Hypothesis Testing

$H_0 : p = p_0$ vs $H_a : p \neq p_0$.
Under H_0, for large n:

$$Z = \sqrt{n}\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)}} \dot{\sim} N(0,1)$$

Reject H_0 if $|Z| > z_{1-\alpha/2}$.

By Slutsky's Theorem, since $\sqrt{\hat{p}(1-\hat{p})} \xrightarrow{a.s.} \sqrt{p_0(1-p_0)}$:

$$\frac{\sqrt{n}(\hat{p}-p_0)}{\sqrt{\hat{p}(1-\hat{p})}} \Rightarrow N(0,1)$$

This is different from the exact result for normal data where $\frac{\sqrt{n}(\bar{X}-\mu)}{s} \sim t_{n-1}$.

Confidence Interval for p

To get a $(1-\alpha)100\%$ CI for p, we use the asymptotic result:

$$P\left(-z_{\alpha/2} < \frac{\sqrt{n}(\hat{p}-p)}{\sqrt{\hat{p}(1-\hat{p})}} < z_{\alpha/2}\right) \approx 1-\alpha$$

The confidence interval is:

$$\hat{p} \pm z_{\alpha/2}\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$$

Exact Hypothesis Test (any n)

Under $H_0 : p = p_0$, the test statistic $n\hat{p} = \sum Y_i$ follows an exact distribution:

$$n\hat{p} \sim \text{Bin}(n, p_0)$$

Reject H_0 if $n\hat{p}$ falls in one of the tails of the $\text{Bin}(n, p_0)$ distribution. (May need a randomized test to attain an exact significance level).

Multiple Categories

Nominal Y can take k possible values, $A_1, \ldots, A_k$.

Multinomial Categories

Assumptions: $Y_1, \ldots, Y_n$ are independent and identically distributed.

- $P(Y_i = A_j) = p_j$ for all i.
- $p_j \geq 0$ and $\sum_{j=1}^{k} p_j = 1$.

Let $X_j = \sum_{i=1}^{n} I(Y_i = A_j)$ be the number of observations in category A_j.
The random vector $X = (X_1, \ldots, X_k)^T$ follows a multinomial distribution:

$$X \sim \text{Multinomial}_k(n, p)$$

where $\sum_{j=1}^{k} X_j = n$.

The probability mass function is:

$$P(X = x) = \frac{n!}{x_1! \ldots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

The MLE of the probability vector p is $\hat{p} = \frac{1}{n}X$, which means $\hat{p}_j = \frac{x_j}{n}$. This is found by maximizing the log-likelihood:

$$l(p) = \sum_{j=1}^{k-1} x_j \ln p_j + \left(n - \sum_{j=1}^{k-1} x_j \right) \ln \left(1 - \sum_{j=1}^{k-1} p_j \right) + C$$

$$\frac{\partial l}{\partial p_j} = \frac{x_j}{p_j} - \frac{x_k}{p_k} = 0 \implies \frac{x_j}{p_j} = \frac{x_k}{p_k}$$

Solving for p_j gives:

$$\frac{x_1}{p_1} = \frac{x_2}{p_2} = \cdots = \frac{x_k}{p_k} = \frac{\sum x_j}{\sum p_j} = \frac{n}{1} \implies \hat{p}_j = \frac{x_j}{n}$$

Properties of Multinomial Distribution

- The one-dimensional marginals are Binomial:

  $$X_j \sim \text{Bin}(n, p_j)$$

  This is because $X_j = \sum_{i=1}^{n} I(Y_i = A_j)$, and each indicator is an independent Bernoulli(p_j) trial.

- The MLE $\hat{p}_j = \frac{X_j}{n}$ is unbiased for p_j.

- The variance of the estimator is $\text{Var}(\hat{p}_j) = \frac{p_j(1-p_j)}{n}$.

Covariance and Correlation

The covariance between counts is:

$$\begin{aligned}\text{Cov}(X_i, X_j) &= E(X_i X_j) - E(X_i)E(X_j) \\ &= n(n-1)p_i p_j - (np_i)(np_j) \\ &= -np_i p_j \quad \text{for } i \neq j\end{aligned}$$

The covariance between the estimators is:

$$\text{Cov}(\hat{p}_i, \hat{p}_j) = \text{Cov}\left(\frac{X_i}{n}, \frac{X_j}{n}\right) = \frac{1}{n^2}\text{Cov}(X_i, X_j) = -\frac{p_i p_j}{n}$$

The covariance matrix of $\hat{p}$ is:

$$\text{Cov}(\hat{p}) = \frac{1}{n}\begin{bmatrix} p_1(1-p_1) & -p_1p_2 & \dots & -p_1p_k \\ -p_2p_1 & p_2(1-p_2) & \dots & -p_2p_k \\ \vdots & \vdots & \ddots & \vdots \\ -p_kp_1 & -p_kp_2 & \dots & p_k(1-p_k) \end{bmatrix} = \frac{1}{n}(\text{diag}(p) - pp^T)$$

Multivariate CLT

For large n, $\sqrt{n}(\hat{p} - p) \Rightarrow N(0, \Sigma)$, where $\Sigma = \text{diag}(p) - pp^T$.

Conditional Distribution

The conditional distribution of a subset of counts, given another subset, is also multinomial. For example:

$$(X_1 | X_2 = x_2, \dots, X_k = x_k) \sim \text{Bin}\left(n - \sum_{j=2}^{k} x_j, \frac{p_1}{1 - \sum_{j=2}^{k} p_j}\right)$$

Categorical Predictors (ANOVA)

We now consider the case where the response is continuous and the predictors are categorical. These categorical predictors are often called **factors**.

One-Way ANOVA Model

This involves one categorical predictor (factor) with I categories (levels). Let y_{ij} be the j-th observation at the i-th level. The model is:

$$y_{ij} = \mu_i + \epsilon_{ij}$$

where we assume $\epsilon_{ij} \sim N(0, \sigma^2)$ are independent. This is equivalent to modeling the mean of each group.

Alternatively, the model can be written as:

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij} \quad i = 1, \dots, I, \quad j = 1, \dots, n_i$$

where μ is the overall mean and α_i is the effect of the i-th level.

Example: Drug Trial

Suppose we have 200 patients with headaches.

- **Treatment group** (50 patients): given medicine.
- **Control group** (150 patients): given a placebo (sugar pill) to control for psychological effects.

This is a single factor ("Group") with 2 levels ("Treatment", "Control"). The response Y_i is the time to recovery. We can model this with linear regression:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where $x_i = 1$ for treatment and $x_i = 0$ for control. We want to test $H_0 : \beta_1 = 0$.

Equivalence of Regression and t-test for 2 Levels

Testing $H_0 : \beta_1 = 0$ in the simple linear regression model with a single 2-level factor is **equivalent** to performing a two-sample t-test for equality of means, assuming equal variances.

Two-Sample t-test Setup:

- Sample 1: $X_1, \ldots, X_m \sim N(\mu_1, \sigma^2)$ (e.g., control group)
- Sample 2: $Y_1, \ldots, Y_n \sim N(\mu_2, \sigma^2)$ (e.g., treatment group)

Test $H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$. The test statistic is:

$$t = \frac{\bar{x} - \bar{y}}{s_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_{m+n-2}$$

where s_p^2 is the pooled variance estimator:

$$s_p^2 = \frac{\sum (x_i - \bar{x})^2 + \sum (y_i - \bar{y})^2}{m + n - 2}$$

Introduction to ANOVA

While **regression** is used when both the predictor and response variables are continuous, ANOVA is a method used when we have **categorical predictors** (also called **factors**) and a **continuous response**.

We will first discuss the case with a single factor ($k = 1$). The different categories or values the factor can take are known as **levels**. A key feature of this setup is that we can have multiple observations at each level.

The basic model looks similar to simple linear regression:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \text{where } \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

However, in the ANOVA context, x_i takes on discrete, unordered values representing the different levels of the factor.

Case Study: One Factor, Two Levels (Drug Trial)

This scenario is a common and intuitive introduction to ANOVA.

Setup

Imagine an experiment with 200 patients suffering from headaches.

- **Treatment Group**: 50 patients are given a new medicine.
- **Control Group**: The remaining 150 patients are not given the medicine. To account for psychological effects, they might be given a **placebo** (a sugar pill).

The response variable, y_i, is the time to recovery. To prevent bias, experiments can be **blinded** (patients don't know their group) or **double-blinded** (doctors also don't know).

Equivalence of Methods

For a single factor with only two levels, we can analyze the data in several equivalent ways:

1. **Linear Regression**: Fit the model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where $x_i = 1$ for the treatment group and $x_i = 0$ for the control group. We can test the one-sided hypothesis $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 < 0$ (if we expect the medicine to *reduce* recovery time) using a t-test.
2. **ANOVA F-test**: This will produce the same result as a two-sided t-test in this specific two-level case.
3. **Two-Sample t-test**: Testing for $\beta_1 = 0$ in the regression model is mathematically identical to performing a two-sample t-test for the equality of means between two normal populations, assuming they have the same variance.

The formula for the two-sample t-test is:

$$t = \frac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{m} + \frac{1}{n}}} \quad \text{where } s_p^2 = \frac{\sum(x_i - \bar{x})^2 + \sum(y_i - \bar{y})^2}{m + n - 2}$$

This tests the null hypothesis $H_0 : \mu_1 = \mu_2$ against $H_a : \mu_1 \neq \mu_2$. The equivalence can be shown by re-parameterizing the model and demonstrating that the test statistic and variance estimate ($\hat{\sigma}^2 = s_p^2$) are identical.

ANOVA for One Factor with Multiple Levels

When a factor has more than two levels (e.g., comparing three different drugs), we use the ANOVA framework.

Decomposition of Variance

ANOVA is based on partitioning the total variability in the data into two components:

- **SSB (Sum of Squares Between Groups)**: The variation of the group means around the overall mean. This reflects the effect of the factor.
- **SSW (Sum of Squares Within Groups)**: The variation of the individual observations around their respective group means. This reflects the random error or noise.

The fundamental identity of ANOVA is that the total variation is the sum of these parts: SST $=$ SSB $+$ SSW. The F-test compares the variation *between* groups to the variation *within* them. If the groups are truly different, the variation between them should be significantly larger than the variation within them.

The ANOVA Model and Identifiability

For a factor with I levels, the model is written as:

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

where y_{ij} is the j-th observation in the i-th group, μ is the overall mean, and α_i is the effect of the i-th level. The null hypothesis is that all group effects are zero: $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$.

However, this model as written is **not identifiable**, meaning there are multiple sets of parameters that produce the exact same distribution for the data. It has too many parameters (e.g., for 3 levels, we have 4 parameters: $\mu, \alpha_1, \alpha_2, \alpha_3$, but only 3 group means to estimate). In matrix form, the design matrix X is not full rank.

To solve this, we must impose a constraint on the parameters. The choice of constraint depends on the desired interpretation:

1. **Set $\boldsymbol{\mu = 0}$**: This is common in cell means models. α_i is then interpreted as the mean of the i-th level.
2. **Set $\boldsymbol{\alpha_1 = 0}$ (Reference Level Coding)**: Here, μ becomes the mean of the first level (the reference group), and each α_i represents the difference between the mean of level i and the mean of the reference level.
3. **Set $\boldsymbol{\sum \alpha_i = 0}$ (Sum-to-Zero Constraint)**: Here, μ represents the grand overall mean, and each α_i is interpreted as the deviation of the i-th group's mean from that grand mean.