## Categorical Data

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

- $Y$ is a discrete random variable.
- $\epsilon$ 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 $Y_1, \dots, Y_n$ 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 Wald 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 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 $\mu$ is the overall mean and $\alpha_i$ 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:

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

**Regression Setup:** Let the control group responses be $z_1, \ldots, z_m$ and treatment group responses be $z_{m+1}, \ldots, z_{m+n}$. The model is $z_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where $x_i = 0$ for $i \leq m$ and $x_i = 1$ for $i > m$. This implies:

- Mean of control group: $E(z_i | x_i = 0) = \beta_0 = \mu_1$
- Mean of treatment group: $E(z_i | x_i = 1) = \beta_0 + \beta_1 = \mu_2$

So, testing $H_0 : \mu_1 = \mu_2$ is the same as testing $H_0 : \beta_1 = 0$.

It can be shown that:

1. The least squares estimate for $\beta_1$ is $\hat{\beta}_1 = \bar{y} - \bar{x}$.
2. The variance estimate from regression, $\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{m+n-2}$, is exactly equal to the pooled variance, $s_p^2$.
3. The t-statistic for $\beta_1$, $t = \hat{\beta}_1 / s.e.(\hat{\beta}_1)$, is identical to the two-sample t-statistic.

## ANOVA Sums of Squares

- **Total Sum of Squares (SST)**: $\sum_{i,j}(y_{ij} - \bar{\bar{y}})^2$
- **Sum of Squares Between Groups (SSB)**: Variation explained by the model.

$$SSB = \frac{mn}{m+n}(\bar{x} - \bar{y})^2$$

- **Sum of Squares Within Groups (SSW)**: Unexplained variation (residuals).

$$SSW = \sum(x_i - \bar{x})^2 + \sum(y_i - \bar{y})^2$$

And we have the decomposition $SST = SSB + SSW$.

The F-test statistic is the ratio of the mean square between groups to the mean square within groups:

$$F = \frac{SSB/(I-1)}{SSW/(N-I)} = \frac{\text{Variation between groups}}{\text{Variation within groups}}$$

If the variation between groups is much larger than the variation within groups, we conclude the group means are significantly different.

## General ANOVA Model ($I > 2$ levels)

The model is:

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

We want to test $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$. The alternative $H_a$ is that at least one $\alpha_i$ is not zero.

## The Problem of Non-Identifiability

This model is not identifiable because there are multiple sets of parameters that give the same probability distribution for $y_{ij}$. For example, if we have parameters $(\mu, \alpha_1, \ldots, \alpha_I)$, we can get the same group means with a new set of parameters $(\mu - c, \alpha_1 + c, \ldots, \alpha_I + c)$ for any constant $c$. The model has too many parameters.

In matrix form, $Y = X\beta + \epsilon$, the design matrix $X$ is not full rank.

## Fixing Non-Identifiability

To make the parameters identifiable, we must add a constraint. Common choices include:

1. **Set** $\mu = 0$. Then $\alpha_i$ represents the mean of the $i$-th level.
2. **Set** $\alpha_1 = 0$ (Treatment contrast, default in R).
   - $\mu$ represents the mean of the first level (the reference level).
   - $\alpha_i$ represents the difference between the mean of level $i$ and the mean of level 1.
3. **Set** $\sum_{i=1}^{I} n_i\alpha_i = 0$ (Sum-to-zero contrast).
   - $\mu$ represents the overall weighted mean.
   - $\alpha_i$ represents the deviation of the $i$-th group's mean from the overall mean.

The choice of constraint depends on the desired interpretation of the estimated parameters. The overall model fit and test results for $H_0$ are the same regardless of the constraint chosen.