Fisher's Exact Test (2x2 Independent)

- **Use when:** Expected cell counts < 5 (Normal approx invalid).
- **Assumptions:** Row/Col totals fixed.
- **Hypergeometric:** $P(X = a) = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$
- **P-value:** Sum probs of tables as/more extreme.

McNemar's Test (Paired Binary)

- **Use when:** Paired data, not independent.
- **Focus:** n_A: + on A/- on B; n_B: - on A/+ on B
- **Hypothesis:** $H_0 : p_{discA} = p_{discB}$ (or $p = 0.5$).
- **Statistic:** $\chi^2 = \frac{(|n_A - n_B| - 1)^2}{n_A + n_B} \sim \chi^2_1$
- **Condition:** $\frac{n_A + n_B}{2} \geq 5$. Else use Exact Binomial.

Session 14: Non-Parametric Tests

Use when: Normality violated (skew, outliers, small n). More robust, less power.

Checks: Histogram, QQ-Plot, Shapiro-Wilk (H_0: Normal).

Sign Test (One-sample/Paired median)

- **Focus:** Median difference Δ. $H_0 : \Delta = 0$ (i.e., $p = 0.5$).
- **Method:** Count positive diffs (C). Ignore zeros (n^*).
- **Dist:** $C \sim Bin(n^*, 0.5)$.
- **Normal Approx ($n^* p(1-p) \geq 5$):** Reject H_0 if $C \geq \frac{n^*}{2} + \frac{1}{2} + z_{1-\alpha/2}\sqrt{\frac{n^*}{4}}$ or $C \leq \frac{n^*}{2} - \frac{1}{2} - z_{1-\alpha/2}\sqrt{\frac{n^*}{4}}$
- **P-value:** If $C > n^*/2$: $p = 2[1 - \Phi(\frac{C - n^*/2 - 1/2}{\sqrt{n^*/4}})]$; If $C < n^*/2$: $p = 2\Phi(\frac{C - n^*/2 + 1/2}{\sqrt{n^*/4}})$

Wilcoxon Signed-Rank (Paired continuous)

- **Assumes:** Symmetry about median. $H_0 : \Delta = 0$.
- **Method:** Rank $|d_i|$ (ignore zeros). T_+ = sum of ranks for positive diffs.
- **Test Stat (no ties):** $T = \frac{|T_+ - \frac{n^*(n^*+1)}{4}| - \frac{1}{2}}{\sqrt{n^*(n^*+1)(2n^*+1)/24}}$
- **With ties:** $T = \frac{|T_+ - \frac{n^*(n^*+1)}{4}| - \frac{1}{2}}{\sqrt{\frac{n^*(n^*+1)(2n^*+1)}{24} - \frac{\sum_{i=1}^{g}(t_i^3 - t_i)}{48}}}$ ($t_i = \#$ in tied group i)
- **Approx:** $n^* \geq 16$. Reject H_0 if $T > z_{1-\alpha/2}$.

Wilcoxon Rank-Sum / Mann-Whitney (2 Indep)

- $H_0 : F_1(t) = F_2(t)$ vs $H_1 : F_1(t) = F_2(t - \delta)$ for $\delta \neq 0$.
- **Method:** Rank pooled data. T_1 = sum of ranks for sample 1.
- **Test Stat (no ties):** $T = \frac{|T_1 - \frac{n_1(n_1+n_2+1)}{2}| - \frac{1}{2}}{\sqrt{\frac{n_1 n_2}{12}(n_1 + n_2 + 1)}}$
- **With ties:** $T = \frac{|T_1 - \frac{n_1(n_1+n_2+1)}{2}| - \frac{1}{2}}{\sqrt{\frac{n_1 n_2}{12}[n_1 + n_2 + 1 - \frac{\sum_{i=1}^{g} t_i(t_i^2 - 1)}{(n_1+n_2)(n_1+n_2-1)}]}}$
- **Approx:** $n_1, n_2 \geq 10$. Reject H_0 if $T > z_{1-\alpha/2}$.

Session 15: SLR Estimation

Assumptions (ϵ_i) - LINE:

- **L**inearity: $E(\epsilon_i) = 0$ (true linear between X and Y)
- **I**ndependence: $\epsilon_i \perp \epsilon_j$ (independent errors)
- **N**ormality: $\epsilon_i \sim N(0, \sigma^2)$ (normal errors)
- **E**qual Variance: $Var(\epsilon_i) = \sigma^2$ (equal variance)

Estimation (OLS): Minimizes $SSE = \sum(Y_i - \hat{Y}_i)^2 = \sum e_i^2$

- $\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{Cov(X,Y)}{Var(X)}$, $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
- $s^2 = MSE = \frac{SSE}{n-p-1}$ (unbiased for σ^2, for SLR $p = 1$)

Max Likelihood Estimation (MLE):

If $\epsilon_i \sim N(0, \sigma^2)$, MLE for β_0, β_1 same as OLS.

Session 16: SLR Inference

Sampling Distribution of $\hat{\beta}_1$ (Slope)

- $E(\hat{\beta}_1) = \beta_1$, $Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum(X_i - \bar{X})^2}$, $se(\hat{\beta}_1) = \sqrt{\frac{MSE}{\sum(X_i - \bar{X})^2}}$
- $t = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \sim t_{n-p-1}$, p-value two side
- CI for $\beta_1 : \hat{\beta}_1 \pm t_{n-p-1,1-\alpha/2} \cdot se(\hat{\beta}_1)$

Sampling Distribution of $\hat{\beta}_0$ (Intercept)

- $E(\hat{\beta}_0) = \beta_0$, $Var(\hat{\beta}_0) = \sigma^2(\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2})$
- $se(\hat{\beta}_0) = \sqrt{MSE(\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2})}$

Confidence Interval for Mean Response $E(Y_h|X_h)$

- $\hat{Y}_h \pm t_{n-p-1,1-\alpha/2} \cdot se(\hat{Y}_h)$
- $se(\hat{Y}_h) = \sqrt{MSE(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2})}$

Prediction Interval (PI) for ONE New Y_h

- $\hat{Y}_h \pm t_{n-p-1,1-\alpha/2} \cdot \sqrt{MSE(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2})}$
- PI is wider than CI for $E(Y_h|X_h)$.

Correlation (r)

- $r = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$. Range: $[-1, 1]$.
- $H_0 : \rho = 0$ equiv. to $H_0 : \beta_1 = 0$.

Coefficient of Determination (R^2)

- $R^2 = r^2$ (for SLR).
- Proportion of Y variation explained by X.
- $R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$

Session 17: Multiple Linear Regression (MLR)

Model: $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i$

- p predictors. Same error assumptions as SLR.

Interpretation:

- β_k: Change in Y for 1-unit increase in X_k, **holding other predictors constant.**

Categorical Predictors:

- Use $d - 1$ indicator (dummy) variables for d levels.
- One level is reference category (all dummies=0).

Interactions:

- Effect of one predictor depends on another's value.
- Included as product terms (e.g., $X_1 X_2$).
- If significant, main effects cannot be interpreted alone.
- If not significant, remove.

Confounding:

- X_2 influences $X_1 - Y$ association.
- Checking change in $\hat{\beta}_1$ when X_2 is added/removed.
- Deal with by: randomization, restriction, stratification.

Session 18: MLR ANOVA

Sum of Squares:

- $SSTO = \sum(Y_i - \bar{Y})^2$ (Total variation)
- $SSR = \sum(\hat{Y}_i - \bar{Y})^2$ (Explained by regression)
- $SSE = \sum(Y_i - \hat{Y}_i)^2$ (Residual/Error)
- $SSTO = SSR + SSE$

Global F-test: $H_0 : \beta_1 = \cdots = \beta_p = 0$.

- $F = \frac{MSR}{MSE} \sim F_{p,n-p-1}$. Reject if $F > F_{crit}$.
- For SLR ($p = 1$), $t^2 = F$.

Partial F-test (Nested Models):

- "small" (p_S predictors) vs "large" (p_L predictors).
- $F = \frac{(SSE_S - SSE_L)/(p_L - p_S)}{MSE_L} \sim F_{(p_L - p_S), n-p_L-1}$.
- Tests if additional predictors are significant.

R^2 vs Adjusted R^2: (Higher is better)

- $R^2 = 1 - \frac{SSE}{SSTO}$: Proportion of Y variance explained. Always increases with more predictors.
- $R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$: Penalizes for predictors.

Model Selection (Lowest is best):

- AIC $= n\ln(SSE/n) + 2p$: Penalizes p less.
- BIC $= n\ln(SSE/n) + p\ln(n)$: Penalizes p more.

Multiple Comparisons: Control FWER (the probability of having at least one Type 1 error among all the tests). Global test first, use adjustments, define comparisons a priori.

Session 19: MLR Diagnostics

Diagnostic Plots:

- **Residuals vs. Fitted:** Heteroscedasticity, outliers. Ideal: random cloud around 0.
- **Residuals vs. Covariate:** Linearity, heteroscedasticity.
- **Normal QQ Plot of Residuals:** Normality, outliers, heavy tails. Ideal: straight line.
- **Scale-Location:** Equal variance. Ideal: horizontal line.

Remedies for Assumptions:

- **Box-Cox:** Finds optimal power λ for Y (Y^λ, $\lambda = 0$ means $logY$).

Unusual Observations:

- **Outliers (Y):** Studentized Residuals $r_i = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$. $|r_i| > 2.5$.
- **Leverage (X):** h_{ii} from Hat Matrix (H). $h_{ii} > \frac{2p}{n}$ (high), $h_{ii} > \frac{3p}{n}$ (very high).
- **Influence:** Changes model parameters.
  - **Cook's Distance (D_i):** Combines residual and leverage. $D_i > 1$ or $D_i > \frac{4}{n}$.
  - **DFFITS:** $|DFFITS_i| > 1$ or $|DFFITS_i| > 2\sqrt{p/n}$.

Multicollinearity: Highly correlated predictors.

- **Effects:** Inflated SEs, unstable coefficients, non-significant predictors.
- **Detect:** VIF (Variance Inflation Factor). $VIF_j = \frac{1}{1-R_j^2}$. $VIF > 5$ (concern), $VIF > 10$ (serious).
- R_j is R-squared of the regression of X_j against all other predictors
- **Remedies:** Remove correlated variables, PCA, shrinkage methods.

Session 20: MLR Variable Selection

Goal: Parsimonious model (good fit, low bias, simple).

Automatic Search Procedures:

- **Backward Elimination:** Start full, remove least significant (high p-value).
- **Forward Selection:** Start empty, add most significant (low p-value).
- **Stepwise:** Combines forward/backward.

Shrinkage Methods (Regularization):

- **LASSO (L1 penalty):** Minimizes $\frac{1}{n}\sum e_i^2 + \lambda\sum|\beta_j|$.
  - Forces some β_j to zero $\implies$ **variable selection**.
  - λ controls shrinkage: large $\lambda \implies$ smaller model.
  - 'lambda.min' (best), 'lambda.1se' (simpler, similar).
- **Ridge (L2 penalty):** Minimizes $\frac{1}{n}\sum e_i^2 + \lambda\sum\beta_j^2$.
  - Shrinks β_j toward zero, but **no variable selection**.
  - Good for correlated predictors.
- **Elastic Net (L1 + L2):** Combines LASSO and Ridge (controlled by 'alpha').

Session 21: MLR Validation

Methods:

- **External Validation:** New, independent data. Evaluate with MSPE (Mean Squared Predicted Error).
- **Internal Validation (Data Splitting):**
  - Split data: **Training Set** (fit model), **Testing Set** (evaluate MSPE).
  - **k-fold Cross-Validation:** Split into k folds. Train on $k-1$, test on 1. Average MSPEs (CV_k). Prefer smaller CV_k.
  - **LOOCV:** $k = n$. Computationally expensive.

Bias-Variance Tradeoff:

- **Simple (Underfit):** High bias, low variance.
- **Complex (Overfit):** Low bias, high variance.
- **Goal:** Minimize Prediction Error ($Bias^2 + Variance + IrreducibleError$).

Session 22: WLS & Robust Regression

Weighted Least Squares (WLS):

- **Use when:** unequal error variances, σ_i^2.
- **Method:** Minimize $\sum w_i(Y_i - \hat{Y}_i)^2$, where $w_i = 1/\sigma_i^2$.
- **Estimating w_i (if σ_i^2 unknown):**
  1. Fit unweighted LS.
  2. Model e_i^2 or $|e_i|$ as function of (some subset of) predictors.
  3. Use fitted values from step 2 to get $\hat{\sigma}_i^2$ or $\hat{s}_i^2$.
  4. Calculate $w_i = 1/\hat{\sigma}_i^2$. Refit with WLS.

Robust Regression: Less affected by influential points.

- **LAD (Least Absolute Deviations):** Minimizes $\sum|Y_i - \hat{Y}_i|$. Less sensitive to outliers.
- **LMS (Least Median of Squares):** Minimizes median of squared residuals. Highly robust.
- **IRLS (Iteratively Reweighted LS):**
  1. Start with initial weights (e.g., OLS or LAD residuals). Fit WLS, get residuals.
  2. Update weights based on current residuals (large residual $\implies$ small weight).
  3. Repeat until convergence.
- **Weight Functions:** Huber, Bisquare (downweight extreme residuals).

Session 23: Lowess & Non-Linear Regression

Non-Parametric Regression: Smoothed curves without strict functional form.

Lowess (Locally Weighted Scatterplot Smoothing):

- Fits series of weighted linear regressions in local neighborhoods. Closer points get higher weights.
- Tuning parameter (span) chosen via cross-validation (smallest MSPE).

Non-Linear Regression: Not linear in parameters (e.g., $Y = \gamma_0\exp(\gamma_1 X)$).

Non-Linear Estimation:

- Numerical optimization (e.g., Gauss-Newton method).
- Iterative process, requires initial values.

Inference (Non-Linear):

- Exact methods not available. Large-sample theory gives approximate normality.
- CI: $g_k \pm t_{n-p,1-\alpha/2}\cdot s\{g_k\}$. Test: $t_{stat} = \frac{g_k - \gamma_{k0}}{s\{g_k\}}$.
- γ_k is the "true" value of the estimator g_k