Effects of violations of model assumptions

1.1 Model assumptions for a single factor ANOVA model

1.2 Effects of various violations

1.3 Diagnostic tools

Contributors

---

1.1 Model assumptions for a single factor ANOVA model

Single factor (fixed effect) ANOVA model:

$$Y_{i_j} = \mu_i + \epsilon_{i_j}, j = 1, \ldots, n_i; i = 1, \ldots, r.$$ 

Important model assumptions

- **Normality**: $\epsilon_{i_j}$'s are normal random variables
- **Equal Variance**: $\epsilon_{i_j}$'s have the same variance ($\sigma^2$).
- **Independence**: $\epsilon_{i_j}$'s are independent random variables.

Some questions:

- What will happen if these assumptions are violated?
- How to find out whether these assumptions are violated? diagnostic tools:
  
  - residual plots: check normality, equal variance, independence, outliers, etc.
- tests for equal variance
  - What to do when these assumptions are violated? remedial measures
    - *Data transformations*
    - *Non-parametric tests*

### 1.2 Effects of various violations

- **Non-normality:**
  - It is not a big deal unless the departure from normality is extreme.
  - *(F\)--test and related procedures are pretty robust to the normality assumption, both in terms of significance level and power."

- **Unequal error variance:**
  - *(F\)--test and related analysis are pretty robust against unequal variance under an approximately balanced design.
  - One parameter inference such as pairwise comparisons of group means could be substantially affected.

- **Non-independence:**
  - It can have serious side effects (effective loss of degrees of freedom).
  - It is also hard to correct.
  - Thus it is very important to use randomization whenever necessary.

### 1.3 Diagnostic tools

Based on residuals:

- **Residuals:**
  \[ \epsilon_{i,j} = Y_{i,j} - \bar{Y}_i, j = 1, \ldots, n_i; i = 1, \ldots, r. \]

- **Studentized residuals:**
  \[ r_{i,j} = \frac{e_{i,j}}{s(e_{i,j})} \]
  where \( s(e_{i,j}) = \sqrt{MSE \times (n_i - 1)/n_i} \)

  (since \( \text{Var}(e_{i,j}) = \sigma^2(1-1/n_i) \)).

- Studentized residuals adjust for sample sizes and thus they are comparable across treatment groups when the design is unbalanced.

**Normal probability plots**

It is a graphical tool to check whether a set of quantities is approximately normally distributed.

- Each value is plotted against its "expected value under normality"
- Sort the values from smallest to largest: \(x_{(1)}, \ldots, x_{(n)}\)

- For the \(i\)-th smallest value \(x_{(i)}\), the "expected value under normality" is roughly the \(\frac{i}{n}\) percentile of the standard normal distribution (the exact definition is a bit more complex).

- A plot that is nearly linear suggests agreement with normality
- A plot that departs substantially from linearity suggests non-normality

**Check normality**

Normal probability plots of the residuals

- When sample size is small: use the combined residuals across all treatment groups.
- When sample size is large: draw separate plot for each treatment group.
- Use studentized residuals (but with MSE replaced by \(s_{(i)}^2\)'s (sample variance of the \(i\)-th treatment group) in the standard error calculation) when unequal variances are indicated and combined residuals are used. Note that \(s_{(i)}^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i}(Y_{i,j} - \bar{Y}_i)^2\)

- Normality is shown by the normal probability plots being reasonably linear (points falling roughly along the 45\(^\circ\) line when using the studentized residuals).

**Checking the equal variance assumption**

Residual vs. fitted value plots.

- When the design is approximately balanced: plot residuals \(e_{(i,j)}\)'s against the fitted values \(\bar{Y}_i\)'s.
- When the design is very unbalanced: plot the studentized residuals \(r_{(i,j)}\)'s against the fitted values \(\bar{Y}_i\)'s.
- Constancy of the error variance is shown by the plot having about the same extent of dispersion of residuals (around zero) across different treatment groups.

Other things that can be examined by residual plots:

- **Independence**: if measurements are obtained in a time/space sequence, a residual sequence plot can be used to check whether the error terms are serially correlated.
- **Outliers** are identified by residuals with big magnitude.
- Existence of other important (but un-accounted for) explanatory variables: whether residual plots shown a certain pattern.

**Example: package design**

Residuals for the package design example are given below.

**Contributors**

- Cathy Wang