Inference with Multiple Comparisons

Concerns
- Multiple comparisons refers to making several comparisons simultaneously.
- **Comparisonwise error rate (CWER)** is the Type I error rate \( \alpha \) for each comparison (i.e. the probability of false rejection for each comparison).
- Note that if each comparison has \( \alpha = 0.05 \) and suppose multiple comparisons, then the probability of having at least one significant comparison given that all \( H_0 \)'s are true is greater than 0.05.
- A crude analogy (assuming independence): toss a coin once with \( P(H) = 0.05 \), toss it twice, then \( P(\text{at least one } H) = 1 - 0.95 \times 0.95 = 0.0975 \), toss it many times, then \( P(\text{at least one } H) \) becomes much larger than 0.05.
- **Experimentwise error rate (EWER)** is the probability of at least one false rejection among multiple comparisons, given that all \( H_0 \)'s are true.

The Bonferroni method makes use of the Bonferroni idea to control EWER.

### Bonferroni idea
- The problem is that suppose \( \text{CWER} = 0.05 \), then \( \text{EWER} \) can be much larger than 0.05 if many comparisons are made. In practice, control \( \text{CWER} \), or \( \text{EWER} \), or find a compromise.
- Consider two comparisons, each with \( \text{CWER} = \alpha \).
- Let \( A \) denote the event that Type I error is made on the first comparison.
- Let \( B \) denote the event that Type I error is made on the second comparison.
- Then,
  \[
  \text{EWER} = P(\text{at least one Type I error is made}) = P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \\
  \leq P(A) + P(B) = 2\alpha
  \]
- The inequality is known as the Bonferroni inequality.
- Usually, \( P(A \text{ and } B) \) is small and thus \( \text{EWER} \approx 2\alpha \).

The ANOVA table is

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trt</td>
<td>4</td>
<td>145.94</td>
<td>36.48</td>
<td>5.09</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Error</td>
<td>30</td>
<td>214.74</td>
<td>7.16</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Total</td>
<td>34</td>
<td>360.68</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Case I: general contrasts
- Among many approaches, we consider two approaches:
  - Bonferroni method
  - Protected t-test
- The Bonferroni method makes use of the Bonferroni idea to control EWER.
- For example, consider two comparisons in the barley example:
  \[
  H_0: \mu_1 = (\mu_2 + \mu_3 + \mu_4 + \mu_5)/4 \\
  H_1: (\mu_2 + \mu_3)/2 = (\mu_4 + \mu_5)/2
  \]
- and we want to control the EWER to be 0.05.
- By the Bonferroni idea, we want \( 2\alpha \approx 0.05 \) and thus \( \alpha = 0.025 \) for each of the two comparisons. That is, for each \( H_0 \), perform a t-test and reject \( H_0 \) if the p-value is \( < 0.025 \).
- In general, suppose there are \( r \) comparisons to be made (chosen in advance) and we want an EWER to be 0.05, then we want \( r\alpha \approx 0.05 \) and thus \( \alpha = 0.05/r \) for each of the \( r \) comparisons. That is, for each \( H_0 \), perform a t-test and reject \( H_0 \) if the p-value is \( < 0.05/r \).

### Case II: all pairwise comparisons
- Among many approaches, we consider three approaches:
  - Fisher’s least significant difference (LSD).
  - Bonferroni test.
  - QD method (Q).
- We also focus on balanced data with \( n = 7 \) plants per variety is recorded. The group means are \( \bar{y}_1 = 16.3, \bar{y}_2 = 19.3, \bar{y}_3 = 14.7, \bar{y}_4 = 20.3, \bar{y}_5 = 18.5 \).
- The ANOVA table is

### Selection bias
- Consider \( k > 2 \) trt comparisons in the following way.
- Take the largest and the smallest trt means and compare them at \( \alpha \) level.
- In testing whether the corresponding population means are equal, the actual Type I error rate is larger than \( \alpha \), because we selected the test that has the highest chance of leading to rejection.
- Now we will learn how to make multiple comparisons with these concerns and ideas in mind.
Inference with Multiple Comparisons

Fisher’s LSD

1. Use protected LSD (basically the same as protected t-test).
2. Find the distance $D_k = Y_1 - Y_2$ so that this distance leads exactly to a p-value of $\alpha$:
   $$D_k = t_{\alpha/2, \text{dfErr}} \times S_A \sqrt{\frac{1}{n}}$$
   and thus the LSD is:
   $$LSD = t_{\alpha/2, \text{dfErr}} \times S_A \sqrt{\frac{1}{n}}.$$  

Remarks

- In the barley example, suppose $\alpha = 0.05$. Since $f = 5.19$ on df = (4,30) and the p-value is less than 0.01, proceed to perform all pairwise comparisons.
- Since $n = 7$, $S_A^2 = 7.16$, dfErr = 30, $t_{.025, 30} = 2.042$, we have $D_k = 2.042 \times \sqrt{7.16 \times 2/7} = 2.92$
- Group: 3 1 5 2 4
- Mean: 14.7 16.3 18.5 19.3 20.3

- That is, two group means that are within 2.92 of each other are connected with a line and are not significantly different.
- Interpretation is not transitive.
- Alternative displays are possible. See the bluebook.

Bonferroni tests

- Use the Bonferroni method for all pairwise comparisons.
- The Q-method (QD) does not involve a t-test.
- Let $D_k = Q_s \cdot \text{dfErr}^{1/2} \times S_A \sqrt{\frac{1}{n}}$ where $Q_s$ is the Q-score from Table A15 (Snedecor and Cochran’s book).

Simple Linear Regression

Objectives

In the snow fall example, the objectives are to describe the relationship between the amount of snow fall ($x$) and the time it takes to clear the streets ($y$), estimate or predict these times to clear the streets for a given amount of snowfall.

Model

- The main idea behind simple linear regression is to fit data with a straight line:
  $$y = b_0 + b_1 x$$
- Recall equation for a straight line $y = mx + b$.
- Here $b_0$ is an intercept and $b_1$ is a slope (rise/run).
- We will discuss the statistical model later.
- The goal is to find $b_0$, $b_1$ for the best fitting line.
- The approach is least squares.
Simple Linear Regression

Least squares

- Find $b_0, b_1$ that minimize the sum of squares
  \[ \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \]
  where $y_i$ is the observed value and $\hat{y}_i$ is the fitted value $\hat{y}_i = b_0 + b_1x_i$

**FACT:** The best fitting line has slope and intercept:

\[ \hat{b}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \]
\[ \hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x} \]

**Source:**

\[ \text{Source} \quad \text{df} \quad \text{SS} \quad \text{MS} \quad F \]
\[ \text{Total} \quad 6 \quad 41.13 \quad 6.855 \quad 239.13 \]
\[ \text{Error} \quad 5 \quad 0.86 \quad 0.172 \]

**Model parameters**

- $\hat{b}_0, \hat{b}_1, \sigma^2$ by estimators $\hat{b}_0, \hat{b}_1, \hat{\sigma}^2$
- $\text{MSE}_{\text{err}}$

**T-test for $H_0 : b_1 = 0$**

- Analysis of variance (ANOVA)
- $T$-test

**Simple Linear Regression**

**ANova for testing $H_0 : b_1 = 0$**

Partition sum of squares (SS):

\[ \text{SSTotal} = \text{SSReg} + \text{SSErr}, \]
\[ \text{df} = n - 1 \]
\[ \text{SSReg} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \]
\[ \text{df} = n - 1 \]
\[ \text{SSErr} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \text{SSTotal} - \text{SSReg} \]
\[ \text{SSReg} = \hat{b}_1 \left[ \sum_{i=1}^{n} x_i - \frac{1}{n} \sum_{i=1}^{n} x_i \right] \left[ \sum_{i=1}^{n} y_i - \frac{1}{n} \sum_{i=1}^{n} y_i \right] \]
\[ \text{df} = 1 \]

**Simple Linear Regression**

**SLR model**

- Model $y$ by random variable $Y$.  
  \[ E(Y|X) = b_0 + b_1X \]
- Consider the model of $Y$ conditional on $X$ ($Y|X$) such that
  \[ E(Y|X) = b_0 + b_1X holistic \]
- The formal simple linear regression (SLR) model is:
  \[ Y_i = b_0 + b_1X_i + e_i \]
  where $e_i \sim \text{N}(0, \sigma^2)$.

**Remarks**

- $\sigma^2$ is sometimes written as $\sigma^2_{y|x}$
- Equivalently, $Y$ are independent.
- $Y_i$ have homogeneous variance: $V_{Y_i} = \sigma^2$.

**Artificial**

- $\sigma^2 = 0.172$ on df = 5.
- $F_{1,5} = 23.913$.
- $\text{MSE}_{\text{err}} = \hat{\sigma}^2 = 0.172$.
- $\text{F}_{5,5} = \frac{\text{MSReg}}{\text{MSE}_{\text{err}}} = 23.913$.
- $t_{5,5} = \frac{\hat{b}_1 - \mu_{\hat{b}_1}}{\sqrt{\text{MSE}_{\text{err}}}}$
- $\mu_{\hat{b}_1} = E(\hat{b}_1) = b_1$ and
- $\text{Var}(\hat{b}_1) = \text{Var} \left( \frac{\sum_{i=1}^{n} (x_i - \bar{x})Y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) = \frac{\text{Var}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$
- $\text{F}_{5,5} = \frac{\text{SSReg}}{\text{SSErr}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$
- $\text{P}_{0.05} = 0.05$ and in the snow fall example
  $\text{t}_{5,5} = \frac{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}}{\text{MSE}_{\text{err}}}$
Simple Linear Regression

T-test for \( H_0 : b_1 = 0 \)

- Fact: Under \( H_0 : b_1 = 0 \),
  \[ T = \frac{b_1 - \mu_{b_1}}{\sigma_{b_1}} \sim T_{n-2} \]

- In the snowfall example, \( s_{b_1} = 0.0877 \) and thus the observed
  \[ t = \frac{0.258}{0.0877} = 15.46 \]
  Compare with \( T \) on df = 5, the (two-tailed) p-value is less than 0.01. 
  Reject \( H_0 \) at 5% and there is strong evidence against \( H_0 : b_1 = 0 \).

- Note that \( t^2 = (15.46)^2 = 239.34 \). Again this relation holds only for \( F \) on df = 1, something.

- In general, under \( H_0 : b_1 = b_1^*, \)
  \[ T = \frac{b_1 - b_1^*}{s_{b_1}} \sim T_{n-2} \]

- A \((1 - \alpha)\) CI for \( b_1 \) is
  \[ b_1 \pm t_{n-2/2, \alpha/2} s_{b_1} \]

Simple Linear Regression

T-test for \( H_0 : b_1 = 0 \)

- For inference of \( b_0 \), use
  \[ \hat{b}_0 = Y - b_1 x \]

- Fact: \( \hat{b}_0 \) has a normal distribution with \( E(\hat{b}_0) = b_0 \) and
  \[ \text{Var}(\hat{b}_0) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2} \right) \]

- Thus
  \[ s_{b_0} = \sqrt{\frac{\sigma^2}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}} \]

- For the snowfall example:
  \[ s_{b_0} = \sqrt{0.172 \times \frac{1}{7} + \frac{3.486}{22.37}} = 0.344 \]

Simple Linear Regression

T-test for \( H_0 : b_1 = b_1^* \)

- Fact: Under \( H_0 : b_1 = b_1^* \),
  \[ T = \frac{\hat{b}_1 - b_1^*}{s_{b_1}} \sim T_{n-2} \]

- In the snowfall example, suppose \( H_0 : b_1 = 0 \) and since \( s_{b_1} = 0.344 \), the observed
  \[ t = \frac{0.258}{0.344} = 0.75 \]
  Compare with \( T \) on df = 3, the p-value: \( P(T_3 \geq 1.00) > 0.10 \). 
  Do not reject \( H_0 \) and there is no evidence against \( H_0 : b_1 = 0 \).

- Suppose \( H_0 : b_1 = 0 \). If \( b_1 = 0 \), then the model becomes \( Y = b_0 x + e \). 
  Hence the test can be viewed as choosing between the model \( Y = b_0 x + e \) under \( H_0 \) 
  and the model \( Y = \hat{b}_0 x + \hat{e} \), under \( H_1 \).

- A \((1 - \alpha)\) CI for \( b_1 \) is
  \[ b_1 \pm t_{n-2/2, \alpha/2} s_{b_1} \]

Simple Linear Regression

Estimation vs prediction

- Consider a simpler model \( Y_i = \mu_i + e_i \), where \( e_i \sim \text{iid N}(0, \sigma^2) \).

- Then \( \hat{Y}_{\text{est}} = \hat{Y} \) estimates \( \mu \) with
  \[ \text{Var}(\hat{Y}_{\text{est}}) = \text{Var}(Y) = \frac{\sigma^2}{n} \]

- Also \( \hat{Y}_{\text{pred}} = \hat{Y} \) predicts a future observation with
  \[ \text{Var}(\hat{Y}_{\text{pred}}) = \text{Var}(Y + e) = \frac{\sigma^2}{n} + \sigma^2 \]

Simple Linear Regression

Inference of the fitted line

- Estimate (predict) \( Y \) at a given \( x^* \) of interest by
  \[ \hat{Y} = \hat{b}_0 + \hat{b}_1 x^* \]

- In the snowfall example, suppose \( x^* = 6 \), then the estimated (predicted) \( y \) is
  \[ \hat{y} = 0.345 \times 1.356 \times 6 = 8.48 \]

- But the standard error depends on the objective.

- Case 1: use \( \hat{Y} \) to estimate the true value \( \hat{b}_0 + \hat{b}_1 x^* \) for a given \( x^* \).

- Case 2: use \( \hat{Y} \) to predict a future obs for a given \( x^* \).

Simple Linear Regression

Case 1: estimation

- If \( \hat{Y} \) is an estimator of the true value \( \beta_0 + \beta_1 x^* \), then denote \( \hat{Y} = \hat{b}_0 + \hat{b}_1 x^* \) by \( \hat{Y}_{\text{est}} \).

- Then we have:
  \[ E(\hat{Y}_{\text{est}}) = E(\hat{b}_0 + \hat{b}_1 x^*) = \beta_0 + \beta_1 x^* \]
  \[ \text{Var}(\hat{Y}_{\text{est}}) = \text{Var}(\hat{b}_0 + \hat{b}_1 x^*) = \text{Var}(\hat{b}_0) + \text{Var}(\hat{b}_1 x^*) = \sigma^2 \left( \frac{1}{n} + \frac{(x^*-\bar{x})^2}{\sum(x_i - \bar{x})^2} \right) \]

- Hence
  \[ s_{\hat{Y}_{\text{est}}} = s \times \sqrt{\frac{1}{n} + \frac{(x^*-\bar{x})^2}{\sum(x_i - \bar{x})^2}} \]

- A \((1 - \alpha)\) CI for \( \beta_0 + \beta_1 x^* \) is
  \[ \hat{Y}_{\text{est}} \pm t_{n-2/2, \alpha/2} s_{\hat{Y}_{\text{est}}} \]

Simple Linear Regression

Case 2: prediction

- If \( \hat{Y} \) is a predictor of a new/future observation, then denote \( \hat{Y} = \hat{b}_0 + \hat{b}_1 x^* \) by \( \hat{Y}_{\text{est}} \).

- Then we have:
  \[ E(\hat{Y}_{\text{est}}) = \frac{E(\hat{b}_0 + \hat{b}_1 x^*)}{\sigma^2} = \frac{\beta_0 + \beta_1 x^*}{\sigma^2} \]

  \[ \text{Var}(\hat{Y}_{\text{est}}) = \text{Var}(\hat{b}_0 + \hat{b}_1 x^*) + \text{Var}(e) \]

  \[ = \sigma^2 \left( 1 + \frac{1}{n} + \frac{(x^*-\bar{x})^2}{\sum(x_i - \bar{x})^2} \right) \]

- Hence
  \[ s_{\hat{Y}_{\text{est}}} = s \times \sqrt{1 + \frac{1}{n} + \frac{(x^*-\bar{x})^2}{\sum(x_i - \bar{x})^2}} \]

- A \((1 - \alpha)\) prediction interval (PI) is
  \[ \hat{Y}_{\text{est}} \pm t_{n-2/2, \alpha/2} s_{\hat{Y}_{\text{est}}} \]
Simple Linear Regression

Remarks

- In the snowfall example, for \( x^* = 6 \),
  \[ s_{\text{pred}} = \sqrt{\frac{1}{27} \left[ 1 + 1 + (6 - 3.495)^2 \right]} \approx 0.495. \]
- A 95% PI is \( 8.48 \pm 2.571 \times 0.495 \),
  which is [7.21, 9.75] or 8.48 ± 2.571.
- How about predicting \( Y \) at \( x^* = 14 \)? Again caution against extrapolation.

Simple Linear Regression

Model fitting

- A useful quantity for assessing the overall regression fit is the coefficient of determination:
  \[ R^2 = \frac{\text{SS Regression}}{\text{SS Total}}. \]
- \( R^2 \) represents the proportion of the total SS that is explained by the regression model.
- In the snowfall example,
  \[ R^2 = \frac{41.13}{41.99} = 0.98 \]
  which is very high.

Simple Linear Regression

Model diagnostics

- Recall the four model assumptions:
  1. The model is correct:
  \[ \text{SS Regression} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2. \]
  2. Errors \( e_i \) are independent.
  3. Errors \( e_i \) have homogeneous variance: \( \text{Var}(e_i) = \sigma^2 \).
  4. Errors \( e_i \) have normal distribution: \( e_i \sim N(0, \sigma^2) \).
- Check model assumptions by examining the residuals:
  - Residual plot: \( r_i \) versus \( \hat{y}_i \).
  - The assumptions are probably OK if the residual plot is a random scatter. Otherwise various patterns may indicate problems such as wrong model, or nonhomogeneous variance, or outliers.
  - It may be hard to interpret when \( n \) is small.

Key R Commands

\[
\begin{align*}
\text{Key R Commands} & \quad \text{Simple Linear Regression} \\
\text{Remarks} & \quad \text{Simple Linear Regression} \\
\text{Model fitting} & \quad \text{Simple Linear Regression} \\
\text{Model diagnostics} & \quad \text{Simple Linear Regression}
\end{align*}
\]

Correlation

An overview

In simple linear regression, we predict \( Y \) given \( x \). Now we are interested in how to variables are related to each other and hence \( X \) and \( Y \) are treated symmetrically.

Height/weight example

Relation of height (\( X \)) and weight (\( Y \)) of adult women

- \( (\text{cm}): 166, 162, 170, 164, 157, 173, 180, 166, 151 \)
- \( (\text{kg}): 59.5, 52.5, 55.0, 50.5, 54.0, 60.5, 64.0, 55.0, 46.5 \)

Correlation

An overview

In simple linear regression, we predict \( Y \) given \( X \). Now we are interested in how to variables are related to each other and hence \( X \) and \( Y \) are treated symmetrically.

Height/weight example

Relation of height (\( X \)) and weight (\( Y \)) of adult women

- \( (\text{cm}): 166, 162, 170, 164, 157, 173, 180, 166, 151 \)
- \( (\text{kg}): 59.5, 52.5, 55.0, 50.5, 54.0, 60.5, 64.0, 55.0, 46.5 \)
Correlation

Model
- X and Y are both random and have a bivariate distribution.
- The most useful distribution is a bivariate normal distribution.
- Probability density surface can be plotted using a 3D or contour plot.
- Y|X = x is normal and so is X|Y = y.

Population correlation coefficient
- \( \rho = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}} \) is the population correlation coefficient between X and Y.
- \( \rho \) is a measure of linear relationship between X and Y.
- \( -1 \leq \rho \leq 1 \).
- \( \rho = 1 \) indicates perfect positive correlation.
- \( 0 < \rho < 1 \) indicates modest positive correlation.
- \( \rho = 0 \) indicates no linear relationship.
- \( -1 < \rho < 0 \) indicates modest negative correlation.
- \( \rho = -1 \) indicates perfect negative correlation.

Sample correlation coefficient
- Based on data \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\), the sample correlation coefficient
  \[ r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} \]
  estimates \( \rho \).
- Note the symmetry between \( x \) and \( y \) in \( r \).
- Working formula is
  \[ r = \frac{\sum_{i=1}^{n}x_iy_i - \frac{1}{n}(\sum_{i=1}^{n}x_i)(\sum_{i=1}^{n}y_i)}{\sqrt{\sum_{i=1}^{n}x_i^2 - \frac{1}{n}(\sum_{i=1}^{n}x_i)^2} \cdot \sqrt{\sum_{i=1}^{n}y_i^2 - \frac{1}{n}(\sum_{i=1}^{n}y_i)^2}} \]
- For the height-weight data, the observed \( r \) is
  \[ r = \frac{82630 - 82308.61}{\sqrt{841.22 \cdot 230.66}} = 0.876. \]
- Note that \( \hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \hat{\bar{y}})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = r \times \frac{\sqrt{\sum_{i=1}^{n}y_i^2 - \frac{1}{n}(\sum_{i=1}^{n}y_i)^2}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \]

Remark: CI uses Fisher transformation and is more involved. See the bluebook.

Correlation inference
- Assume X and Y are from a bivariate normal distribution.
- Test \( H_0: \rho = 0 \) versus \( H_A: \rho \neq 0 \).
- Use \( T = \sqrt{n-2} \cdot r \sim T_{n-2} \) under \( H_0 \).
- For the height-weight data, the observed \( t = 0.876 \times \sqrt{7} = 4.80 \) on df = 7 with a p-value < 0.01.
- Remark: \( r \) can be perfectly related \( \hat{Y} = \hat{\bar{y}} + \hat{b}x \), but not linear (e.g., \( Y \) on 1 df).
- Remark: CI uses Fisher transformation and is more involved. See the bluebook.

Correlation

Population correlation coefficient

Sample correlation coefficient

Categorical Data

An overview
- Case 1: Binomial, 1 sample
- Case 2: Multinomial, 1 sample
- Case 3: Binomial, 2 samples
- Case 4: Binomial, multiple samples

Case 1: Binomial, 1 sample
Example
- For \( Y \sim B(100, p) \), test \( H_0: p = 0.6 \) versus \( H_A: p \neq 0.6 \).
- Suppose we observe 72 heads, by normal approximation, we have \( Y_{HA} \sim N(60, 24) \) and \( \bar{p}_{HA} \sim N(0.6, 0.24) \) under \( H_0 \).
- Thus \( Z = \frac{Y_{HA} - 60}{\sqrt{24}} \sim N(0, 1) \) and the observed
  \[ z = \frac{72 - 60}{\sqrt{24}} = 2.45 \]
with p-value = 0.007 < 2 = 0.014.

Case 1: Binomial, 1 sample
New approach
- Draw the following contingency tables

<table>
<thead>
<tr>
<th>H</th>
<th>T</th>
<th>O</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>28</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>46</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

- Compute \( \chi^2 = \sum \frac{(O - E)^2}{E} \)
- FACT: If \( H_0 \) is true, then \( \chi^2 \) is approximately \( \chi^2 \) on 1 df.
- Thus the observed \( \chi^2 = \frac{(72 - 60)^2}{24} + \frac{(28 - 40)^2}{40} = 6 \)
and compared with \( \chi^2 \) on 1 df, p-value = \( P(\chi^2 \geq 6) \) [one-sided]. From Table B, the p-value is between 0.001 and 0.025.
- \( \chi^2 = 6 \) holds for \( \chi^2 \).
- Same condition as Z-test for a good approximation: np ≥ 5 and n(1 - p) ≥ 5.
Case 2: Multinomial, 1 sample

Example

A specially constructed die is such that three sides are labeled 1, the other three sides are labeled 2, 3, and 4. Roll the die 240 times for testing $H_0$: $p_1 = 1/2, p_2 = p_3 = p_4 = 1/4$ versus $H_1$: not $H_0$, where $p_i$ = probability that the die comes up with $i$.

Note that $p_1 + p_2 + p_3 + p_4 = 1$.

<table>
<thead>
<tr>
<th>Observed</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>51</td>
<td>44</td>
<td>66</td>
<td>16</td>
</tr>
</tbody>
</table>

$X^2 = \sum_{all \ possibilities} \frac{(observed - expected)^2}{expected}$

Thus the observed $X^2 = \frac{(51 - 108)^2}{108} + \frac{(22 - 40)^2}{40} + \frac{(39 - 40)^2}{40} + \frac{(66 - 40)^2}{40}$

$= 1.2 + 4.255 + 0.025 + 16.9 = 22.45$

and compared with $\chi^2$ on 3 df, p-value = $P(\chi^2 \geq 22.45) < 0.01$.

Case 2: Multinomial, 1 sample

Remarks

- In general, df = # of cells - 1.
- Formally, the model is a multinomial distribution (a generalization of binomial) with 3 assumptions:
  1. $n$ independent trials.
  2. Each trial has $k$ mutually exclusive outcomes.
  3. Constant probability for each outcome in each trial $p_i$.

Let $Y_i$ = # of $i$-th outcome in $n$ trial, $i = 1, \ldots, k$, follows a multinomial distribution.

- Conditions for $\chi^2$ test:
  1. All expected values $\geq 5$.
  2. At least 80% of the expected values $\geq 3$.

Case 3: Binomial, 2 samples

Example

- Compare two treatments A and B. For A, there are 71 successes among 105 trials. For B, there are 45 successes among 87 trials.

Let $p_1 = \frac{71}{105}$, $p_2 = \frac{45}{87}$.

We compute $p_1 = \frac{71}{105}$, $p_2 = \frac{45}{87}$.

For A and $p_1 = \frac{71}{105} = 0.676$

For B and $p_2 = \frac{45}{87} = 0.517$

Thus the z = 2.24 with a p-value of 0.025.

Case 3: Binomial, 2 samples

New approach

- FACT: If $H_0$ is true, then $X^2$ is approximately $\chi^2$ on 1 df.

Thus the observed $X^2 = \frac{(71 - 63.44)^2}{63.44} + \frac{(34 - 41.56)^2}{41.56} + \frac{(45 - 52.56)^2}{52.56} + \frac{(42 - 34.44)^2}{34.44}$

$= 5.023$

and compared with $\chi^2$ on 1 df, p-value = $P(\chi^2 \geq 5.023) = 2.42$.

df = 1 because given the marginals and the total, there is only 1 piece of independent information.

Case 3: Binomial, 2 samples

New approach

- FACT: If $H_0$ is true, then $X^2$ is approximately $\chi^2$ on 3 df.

Thus the observed $X^2 = \frac{(22 - 40)^2}{40} + \frac{(29 - 28)^2}{28} + \frac{(13 - 13)^2}{13}$

$= 2.09$

and compared with $\chi^2$ on 3 df, p-value = $P(\chi^2 \geq 2.09) = 0.36$.

Case 4: Binomial, multiple samples

Contingency tables

Compare 4 species (1-4) of pine for disease resistance in a study.

<table>
<thead>
<tr>
<th>Species</th>
<th>No disease</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22</td>
<td>14</td>
</tr>
<tr>
<td>2</td>
<td>25</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>17</td>
</tr>
<tr>
<td>4</td>
<td>31</td>
<td>18</td>
</tr>
</tbody>
</table>

Under $H_0$: $p_1 = p_2 = p_3 = p_4$.

<table>
<thead>
<tr>
<th>Species</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>21</td>
<td>24</td>
<td>26</td>
<td>22</td>
</tr>
</tbody>
</table>

$X^2 = \sum_{all \ possibilities} \frac{(observed - expected)^2}{expected}$

Case 4: Binomial, multiple samples

New approach

- FACT: If $H_0$ is true, then $X^2$ is approximately $\chi^2$ on 3 df.

Thus the observed $X^2 = \frac{(22 - 40)^2}{40} + \frac{(29 - 28)^2}{28} + \frac{(13 - 13)^2}{13}$

$= 2.09$

and compared with $\chi^2$ on 3 df, p-value = $P(\chi^2 \geq 2.09) = 0.36$.

This idea can be extended to general $r \times c$ case, where $r$ is the # of rows and $c$ is the # of columns. Then df = $(r - 1) \times (c - 1)$.

This is an overall test. It may be important to look at individual pieces.

- Conditions for the $\chi^2$ test are again:
  1. All expected values $> 1$
  2. At least 80% of the expected values $\geq 5$.
Categorical Data

Key R commands

> # case 1 binomial 1 sample
> prop.test(72, 100, p=0.6, correct=F)

1-sample proportions test without continuity correction
data: 72 out of 100, null probability 0.6
X-squared = 6, df = 1, p-value = 0.01431
95 percent confidence interval:
0.6251197 0.7986031
sample estimates:
p 0.72

> chisq.test(c(72,28), p=c(0.6,0.4), correct=F)

Chi-squared test for given probabilities
data: c(72, 28)X-squared = 6, df = 1, p-value = 0.01431

> # case 2 multinomial 1 sample
> chisq.test(c(108,27,39,66), p=c(1/2,1/6,1/6,1/6), correct=F)

Chi-squared test for given probabilities
data: c(108, 27, 39, 66)X-squared = 22.35, df = 3, p-value = 5.516e-05

> # case 3 binomial 2 samples
> prop.test(c(71,45), c(105,87), correct=F)

2-sample test for equality of proportions without continuity correction
data: c(71, 45) out of c(105, 87)X-squared = 5.0264, df = 1, p-value = 0.02496
95 percent confidence interval:
0.02097751 0.29692068
sample estimates:
prop 1 prop 2
0.6761905 0.5172414

> matrix(c(71,34,45,42),2,2)

[,1] [,2]
[1,] 71 45
[2,] 34 42

> chisq.test(matrix(c(71,34,45,42),2,2), correct=F)

Pearson's Chi-squared test
data: matrix(c(71, 34, 45, 42), 2, 2)X-squared = 5.0264, df = 1, p-value = 0.02496

> # case 4 binomial multiple samples
> prop.test(c(22,10,15,20), c(51,38,44,37), correct=F)

4-sample test for equality of proportions without continuity correction
data: c(22, 10, 15, 20) out of c(51, 38, 44, 37)X-squared = 6.8694, df = 3, p-value = 0.07618

alternative hypothesis: two.sided
sample estimates:
prop 1 prop 2 prop 3 prop 4
0.4313725 0.2631579 0.3409091 0.5405405

> matrix(c(22,29,10,28,15,29,20,17),2,4)

[1,] 22 10 15 20
[2,] 29 28 29 17

> chisq.test(matrix(c(22,29,10,28,15,29,20,17),2,4),correct=F)

Pearson's Chi-squared test
data: matrix(c(22, 29, 10, 28, 15, 29, 20, 17), 2, 4)X-squared = 6.8694, df = 3, p-value = 0.07618