Non-Parametric Tests and Interpretation

Isaac Griffith

CS 6620
Department of Informatics and Computer Science
Idaho State University

Non-parametric Tests

Non-Parametric Tests

- **Mann-Whitney**: non-parametric alternative to t-test
- **Wilcoxon**: non-parametric alternative to the paired t-test
- **Sign test**: non-parametric alternative to the paired t-test (simpler than Wilcoxon)
- **Kruskal-Wallis**: non-parametric alternative to ANOVA for one factor with more than two treatments
- **Chi-2**: family of non-parametric tests used when data are in the form of frequencies

Mann-Whitney Overview

- **Input**: samples $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_m$
- **Hypotheses**:
  - H_0: samples are from same distribution
  - H_A: they are not
- **Calculations**: rank all samples and calculate the following
  - $U = N_A N_B + \frac{N_A(N_A+1)}{2} - T$
  - $U' = N_A N_B - U$
  - $N_A = \min(n, m)$
  - $N_B = \max(n, m)$
  - T is the sum of the ranks of the smallest sample
- **Criterion**:
  - Reject H_0 if $\min(U, U')$ is less than or equal to the Mann-Whitley critical value at N_A, N_B

Mann-Whitney Example

Defect density in different programs have been compared in two projects

- Hypotheses
  - H_0: defect density distribution is the same in both projects
  - H_A: defect density distribution is not the same
- Data: Defect density results for project x and project y
  - $x = 3.42, 2.71, 2.84, 1.85, 3.22, 3.48, 3.68, 4.30, 2.49, 1.54$
  - $y = 3.44, 4.97, 4.76, 4.96, 4.10, 3.05, 4.09, 3.69, 4.21, 4.40, 3.49$
- Data Sizes
  - $N_A = \min(10, 11) = 10$
  - $N_B = \max(10, 11) = 11$

Mann-Whitney Example

- Ranks of samples:
  - Smallest sample (x): 9, 5, 6, 2, 8, 11, 4, 17, 3, 1
  - Largest sample (y): 10, 21, 19, 20, 15, 7, 14, 13, 16, 18, 12
- Calculated Values:
  - $T = 66$
  - $U = 99$
  - $U' = 11$
  - $\min(U, U') = 11$
- Results
  - Critical value for (10,11) is 26
  - Since 11 < 26, we reject H_0 with two-tailed test at 0.05 level

Wilcoxon Overview

- **Input**: Paired samples $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- **Hypotheses**:
  - H_0: If all differences, d_i, regardless or sign are ranked, then the sum of the positive differences equals the sum of the negative differences
  - H_A: they are not the same
- **Calculations**:
  - $d_i = x_i - y_i$
  - All differences, d_i are ranked regardless of sign
  - T^+ is the sum of the positive d_i's
  - T^- is the sum of the negative d_i's
- **Criterion**:
  - T_n critical value for n pairs
  - reject H_0 if $\min(T^+, T^-) \leq T_n$

Wilcoxon Example

Ten programs independently developed two different programs. They measured the effort required, as shown in the table

- Hypotheses
  - H_0: If all differences effort, d_i, regardless or sign are ranked, then the sum of the positive differences equals the sum of the negative differences
  - H_A: they are not

Programmer	1	2	3	4	5	6	7	8	9	10
Program 1	105	137	124	111	151	150	168	159	104	102
Program 2	86.1	115	175	94.9	174	120	153	178	71.3	110

Wilcoxon Example

- Calculation:
  - $T^{+} = 32$
  - $T^{-} = 23$

Pair	1	2	3	4	5	6	7	8	9	10
Difference	18.9	22	-51	16.1	23	30	15	19	32.7	9
Rank	4	6	10	3	7	8	2	5	9	1

- Statistics
  - $T_n = 8$
- Result:
  - Since $\min(T^{+}, T^{-}) = \min(32, 23) = 23 > 8$ we cannot reject H_0 with a two-tailed test at the 0.05 level

Sign Test Overview

- **Input**: Paired samples $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$
- **Hypotheses**:
  - H_0: $P(+) = P(-)$, where $+$ and $-$ are the two events that $x_i > y_i$ and $x_i < y_i$
  - Two-Sided H_A: $P(+) \neq P(-)$
  - One-Sided H_A: $P(+) < P(-)$
- **Calculations**:
  - $d_i = x_i = y_i$
    - positive differences are represented by a $+$
    - negative differences are represented by a $-$
  - $p = \frac{1}{2^N} \sum_{i=0}^{n} \binom{N}{i}$
- **Criterion**:
  - Two-Sided: reject H_0: if $p < \alpha/2$
  - One-Sided: reject H_0: if $p < \alpha$ and the $+$ event is the most rare event

Sign Test Example

Ten programs independently developed two different programs. They measured the effort required, as shown in the table

- Hypotheses
  - H_0: required effort to develop program 1 is the same as for program 2
  - H_A: it is not

Programmer	1	2	3	4	5	6	7	8	9	10
Program 1	105	137	124	111	151	150	168	159	104	102
Program 2	86.1	115	175	94.9	174	120	153	178	71.3	110

Sign Test Example

- Calculation:
  - $d = 18.9, 22, -51, 16.1, 23, 30, 15, 19, 32.7, 9$
  - $S_d = 27.358$
  - $t_0 = 0.39$
  - $df = n - 1 = 10 - 1 = 9$
- Statistics
  - $t_{0.025,9} = 2.262$
- Result:
  - Since $t_0 < t_{0.025,9}$ we cannot reject H_0 at the 0.05 level

Kruskal-Wallis Overview

- Can always be used when the assumptions ANOVA cannot be met
- **Input**: a samples: $x_{11}, x_{12}, \ldots, x_{1n_1}; x_{21}, x_{22}, \ldots, x_{2n_2}; \ldots; x_{a1}, x_{a2}, \ldots, x_{an_a}$
- **Hypotheses**:
  - H_0: the population medians of the a samples are equal
  - H_A: they are not
- **Calculations**:
  - All measures are ranked in one series (1,2,...,n_1 + n_2 + ...+ n_a)
  - The calculations are based on these ranks

Multiple Comparison

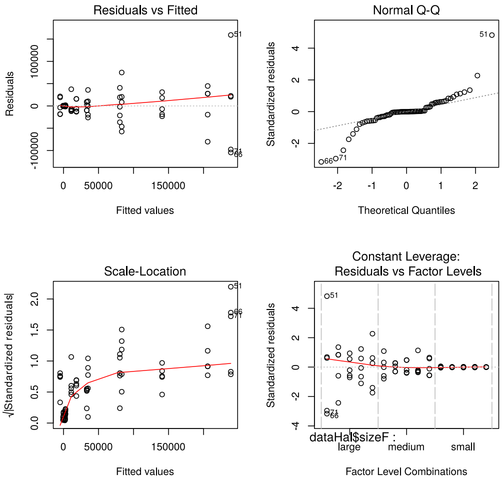
- All Pairs
  - Sigel-Tukey Test
  - Ansari-Bradley Test
  - Permutation Test on Deviants
  - Friedman MCP
  - Dwass MCP
  - Critchlow-Fligner MCP
- Control Level?
  - Steele's Comparison to Control

Model Adequacy Checking

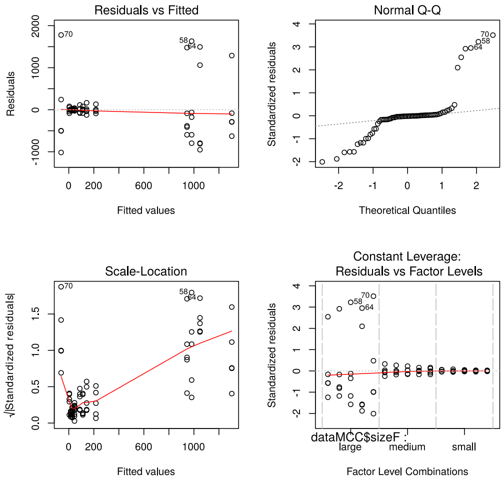
- Homogeneity of Variance
  - Levene's Test
  - Bartlett's Test
  - Brown-Forsythe's Test
  - Residuals vs. Fitted Plot
- Normality of Residuals or Data
  - Kolmorgorov-Smirnov Test
  - Anderson-Darling Test
  - Cramer-von Mises Test
  - Chi2 Goodness of Fit Test
  - Normal Q-Q plot

Visual Checks

Experiment 1



Experiment 2



Drawing Conclusions

Drawing Conclusions

- Once data analysis is complete it must be interpreted in order to draw conclusions regarding experimental outcome
- Hypothesis testing
  - If we reject H_0, we can conclude that the independent variables affect the dependent variables
    - But it still may be of little importance
  - Otherwise, no conclusion of statistical significance can be drawn
    - But lessons learned may be important
- If statistical significant differences are found
  - we may be able to make general conclusions about the relationship between independent and dependent variables
  - we can only generalize to an environment similar to the experiment
- To draw **causal inferences** we must have randomly assigned experimental units to treatments
- Care must be taken when drawing conclusions from experiments

Are there any questions?