Statistics & Explanatory Data Analysis

Two-sample tests

dr Marcin Chlebus, dr Ewa Cukrowska - Torzewska

UNIWERSYTET WARSZAWSKI
Wydział Nauk
Ekonomicznych

Two independent samples t test

DATA TYPE:
- Dependent variable is interval/ratio & continuous
- Independent variable is binary
- Data for each population are normally distributed
- Observations between groups are independent. That is, not paired or repeated measures data
- Moderate skewness is permissible if the data distribution is unimodal without outliers
- Different statistics when variances are equal and not equal.

HYPOTHESIS:
H0: Means are equal
H1 (2 sided): Means are not equal

INTERPRETATION:
H0: Fail to reject that means are significantly different
H1 (2 sided): Means are significantly different

Equal variances:

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)\ (n_2-1)}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

Unequal variances:

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \sim t_{\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)}+\frac{s_2^4}{n_2^2(n_2-1)}}}$$

Normality assumption:
- n>30 for both samples
- fail to reject H0 about normality for both samples

Variance assumption:
- Test for equal variances (F test)

F test for equal variances

DATA TYPE:

- Dependent variable is interval/ratio & continuous
- Independent variable is binary
- Data for each population are normally distributed
- Observations between groups are independent. That is, not paired or repeated measures data

HYPOTHESIS:

H0: Variances are equal

H1 (2 sided): Variances are not equal

INTERPRETATION:

H0: Fail to reject that variances are significantly different

H1 (2 sided): Variances are significantly different

$$F = \frac{S_1^2}{S_2^2} \sim F_{(n_1-1; n_2-1)}$$

Two-sample Wilcoxon rank-sum/Mann-Whitney U Test

DATA TYPE:

- Two-sample data.
- Dependent variable is ordinal, interval, or ratio
- Independent variable is a factor with two levels.
- Observations between groups are independent.
- In order to be a test of medians, the distributions of values for each group need to be of similar shape and spread (outliers affect the spread). Otherwise the test is a test of distributions

HYPOTHESIS:

- H0: The medians of values for each group are equal (*distributions are similar in shape and spread)*/The distribution of values for each group are equal (otherwise).
- H1 (2-sided): The medians of values for each group are not equal/there is systematic difference in the distribution of values for the groups

Equal scale testing: Ansari-Bradley test

If the distributions of values of each group are similar in shape, but have outliers, then Mood's median test is an appropriate alternative

- **Procedure:**
  1. Order both samples together in ascending order
  2. Assign Ranks to each observation (if ties assign for average of ranks)

$$T = \sum_{i=1}^{n_1} R_{1i} \quad U = T - \frac{n_1(n_1+1)}{2}$$

For small samples tables

$$U \xrightarrow{aproxx} N(\mu_U, \sigma_U)$$

Two-sample Mood's Median Test

DATA TYPE:
- Two-sample data. (or more)
- Dependent variable is ordinal, interval, or ratio
- Independent variable is a factor with two levels (or more).
- Observations between groups are independent.
- Distributions of values for each group are similar in shape; however, the test is not sensitive to outliers (different variances are acceptable)

HYPOTHESIS:
- H0: The medians of values for each group are equal.
- H1 (2-sided): The medians of values for each group are not equal

Low power in comparison to Wilcoxon/M-W test, but do not require approximately equal variances (scale/spread)

Only option for data with serious outliers

- **Procedure:**
  1. Pool data for the two samples and order them in an ascending orderCalculate median for join sample
  2. Prepare contingency table (Below/Above median vs original sample ID)
  3. Perform Fischer Exact or Pearson Chi2 test

Two-sample paired t test

DATA TYPE:
- Dependent variable is interval/ratio & continuous
- Independent variable is binary
- Samples are paired. Observation in one group can be paired logically or by subject to an observation in the other group
- The distribution of the difference of paired measurements is normally distribute• Moderate skewness is permissible if the data distribution is unimodal without outliers
- Moderate skewness is permissible if the data distribution is unimodal without outliers

HYPOTHESIS:
H0: The difference between paired observations is equal to zero.
H1 (2 sided): The difference between paired observations is not equal to zero.

$$t = \frac{\overline{X}_D - \mu_D}{\sqrt{\frac{s_D^2}{n}}} \sim t_{n-1}$$

Normality assumption:
- n>30 for sample of differences
- fail to reject H0 about normality

Two-sample wilcoxon paired signed-rank test

DATA TYPE:

- Two-sample paired data. That is, one-way data with two groups only, where the observations are paired between groups.
- Dependent variable is ordinal, interval, or ratio
- Independent variable is a factor with two levels. That is, two groups
- The distribution of differences in paired samples is symmetric

If the distribution of differences between paired samples is not symmetrical, the two-sample sign test for paired data can be used.

HYPOTHESIS:

- H0: The distribution of the differences in paired values is symmetric around zero.
- H1 (2-sided): The distribution of the differences in paired values is not symmetric around zero

- **Remark: Rank Sum test is a one-sample Rank Sum test for difference of paired values from two samples**
- **Procedure:**
  - For each pair of observations (N) calculate absolute difference: $|X_i - Y_i|$
  - Drop observations with absolute difference equal to 0 (N_r observations left)
  - Order the rest in ascending order and assign ranks R_i. For tied ranks assign an average rank.
  - Calculate test statistics: $W = \sum_{i=1}^{N_r}[sgn(X_i - Y_i)R_i]$
  - Take critical values form reference table (specific distribution with $E(W) = 0$ & $VAR(W) = \frac{N_r(N_r+1)(2N_r+1)}{6}$=) or use normal approximation

Two-sample paired sign test

DATA TYPE:
- Two-sample paired data.
- Dependent variable is ordinal, interval, or ratio
- Independent variable is a factor with two levels. That is, two groups

HYPOTHESIS:
H0: The median of difference of between pairs is equal to 0
H1 (2 sided): The median of the differences between pairs is not zero

- **Data does not have to be symmetric in distribution**
- Has **smaller power** than Wilcoxon one sample test
- **Procedure:**
  - For each observation (from N) calculate difference from assumed ME: $X_i - Y_i$
  - Drop observations with absolute difference equal to 0 (N_r observations left)
  - Calculate sum of positive difference (S) and sum of negative difference (F).
  - P-Value: $P(s \leq S); S \sim binomial(n = N_r, p = 0.5)$