Statistical Tools – Overview

1. Categorical Response/Data
2. Numerical Response/Data

In general any hypothesis test and its corresponding confidence interval have the same assumptions.

1. Categorical Data/Response

1.1 Contingency Tables

1.1.1 One Categorical Variable compared to theoretical distribution – χ^2 Goodness-of-fit-test

Purpose: Do the data provide sufficient evidence that the theoretical distribution is an in-appropriate description of the population?

Assumptions:

1. Simple Random Sample
2. expected frequencies for all cells are at least 5

1.1.2 Two Categorical Variables – χ^2 Test for Association

Purpose: Do the data provide sufficient evidence that the two variables are NOT independent?

Assumptions:

1. Simple Random Sample
2. expected frequencies for all cells are at least 5

1.2 Proportions or Percentages

1.2.1 One Population – **One proportion z-test**

Purpose: Do the data provide sufficient evidence that the proportion is different from a theoretical value, p_0?

Assumptions:

1. Simple Random Sample
2. both np_0 and $n(1-p_0)$ are at least 5

1.2.2 Tow Populations – **Two proportion z-test**

Purpose: Do the data provide sufficient evidence that the two proportions are different?

Assumptions:

1. Two Independent Simple Random Samples
2. in each sample the number of successes and failures are at least 5

1.3 Proportion in dependency on a number of predictors

– Logistic Regression

Purpose: Do the data provide sufficient evidence that one or more predictors affect the proportion?

Assumptions:

1. Random Sample
2. large sample size (at least 10 measurements for each parameter in the model equation)

Model equation: $\log\left(\frac{p}{1-p}\right) = \sum \beta_i x_i$

p is the probability for the event of interest and x_i, $1 \leq i \leq m$, are the predictors.

2. Numerical Response/Data

2.1 One Population (One-Sample Case)

- Parametric: **One-sample t-test**
  Purpose: To test if there is a difference between the population mean and a theoretical mean μ_0.
  Assumptions:
  1. Simple random sample.
  2. The population under study is normally distributed or sample size is large.
- Non-Parametric: **Wilcoxon signed-rank test**
  Purpose: To test if there is a difference between the population median and a theoretical median m_0.
  Assumptions:
  1. Simple random sample.
  2. distribution is symmetric.

2.2 Two populations (independent samples)

- Parametric: **Two-sample t-test**
  Purpose: To test if there is a difference between two population means
  Assumptions:
  1. Two independent Simple random samples
  2. both populations are normally distributed or both sample sizes are large
- Non-Parametric: **Wilcoxon Rank Sum test = Mann-Whitney U test**
  Purpose: To test if there is a difference between two population medians
  Assumptions:
  1. Two independent Simple random samples
  2. distributions have same shape

2.3 Two populations (paired samples)

- Parametric: **Paired t-test**
  Purpose: To test if there is a difference between two population means
  Assumptions:
  1. Paired random samples
  2. the populations of paired differences is normally distributed or the sample size is large
- Non-Parametric: **Wilcoxon Signed Rank test**
  Purpose: To test if there is a difference between two population medians
  Assumptions:
  1. Paired random samples
  2. distribution of difference is symmetric

2.4 More than two populations (independent samples)

- Parametric: **1-way ANOVA**
  Purpose: To test if there is a difference between more than two population means
  Assumptions:
  1. Independent random samples
  2. the populations are normally distributed 3. standard deviations for the populations are equal (rule-of-thumb: largest st.dev is smaller than 2 times the smallest st.dev.)
- Non-Parametric: **Kruskal-Wallis test**
  Purpose: To test if there is a difference between more than two population medians
  Assumptions:
  1. Independent random samples
  2. distributions have the same shape
  3. Sample sizes are at least 5

2.5 Effect of two factors on mean response

- Parametric: no interaction **Randomized Block Design**
  Purpose: To test if the two factors affect the mean response separately (main effects)
  Assumptions: 1. Independent random samples for each factor level combination
  2. The ANOVA model is an appropriate fit for the population (given the model equation, the error is normally distributed with equal st.dev)
- Parametric: interaction **2-way ANOVA**
  Purpose: To test if the two factors interact in their effect on the mean response and/or they affect the mean separately (main effects)
  Assumptions: 1. see RBD, 2. see RBD
- Non-Parametric: no interaction **Friedman's test**
  Purpose: To test if the two factors affect the median response separately (main effects)
  Assumptions: 1. Independent random samples
  2. distributions have the same shape
  3. number of treatments or blocks are at least 6.

2.6 Effect of more factors on mean response

Parametric: k-**way ANOVA**

Purpose: To test if a number of factors affect the mean response separately (main effects) or in interactions. Permits to test the effect of several factors on the mean response while correcting for a number of other factors.

Assumptions: 1. Independent random samples for each factor level combination

2. The ANOVA model is an appropriate fit for the population (given the model equation, the error is normally distributed with equal st.dev)

2.7 Effect of *one* numerical predictor on mean response

Parametric: **Simple Linear Regression Analysis**
Purpose: To test whether there is a linear relationship between two numerical variables and describe it **Assumptions:**
1. Random sample of independent observations
2. Linearity: The relationship between the variables must be approximately linear
3. For fixed values of the predictor the response is normal with the same st.dev.
4. no outliers (there are different kinds)

Analysis:
1. Regression ANOVA F-test
Purpose: To test whether there is a linear relationship between predictor and response
2. Regression t-test for slope
Purpose: To test whether there is a positive/negative linear relationship between predictor and response

Nonparametric: **Spearman rank correlation test**
Purpose: To test whether there is a monotone relationship between predictor and response

2.8 Effect of *several* predictors on mean response

Parametric: **Multiple Linear Regression Analysis**

Purpose: To test whether any of the predictors have a linear effect or interact in their effect on the mean response

Assumptions:

1. Random sample of independent observations
2. Linearity: The mean response is (approximately) a linear combination of the predictors and their interaction terms
3. For fixed values of the predictors the response is normal with the same st.dev.
4. no outliers (there are different kinds)

Statistical Analysis:

1. Regression ANOVA F-test

Purpose: To test whether at least one of the slopes in the model is different from 0

2. Regression t-test for each of the slopes

Purpose: To test whether there is a positive/negative linear relationship between the different predictors or the interaction terms and the mean response