Common Statistical Methods

J. LEE

Hanyang University

Department of Applied Statistics

March 28, 2022

Table of Contents

Independent t-test

- A difference in mean with **independent** samples
- Two separate sets of independent and identically distributed samples are obtained, and one variable from each of the two populations is compared
- For example, suppose we are evaluating the effect of a medical treatment, and we enroll 100 subjects into our study, then randomly assign 50 subjects to the treatment group and 50 subjects to the control group. In this case, we have two independent samples and would use the unpaired form of the t-test

t-test

Paired t-test

- A difference in mean with **dependent** samples
- A sample of matched pairs of similar units, or one group of units that has been tested twice
- A typical example of the repeated measures t-test would be where subjects are tested prior to a treatment, say for high blood pressure, and the same subjects are tested again after treatment with a blood-pressure-lowering medication.

ANOVA

One-way ANOVA

- Test for differences among at least three groups using the F distribution
- This technique can be used only for numerical response data, the "Y", usually one variable, and numerical or (usually) categorical input data, the "X", always one variable, hence "one-way"
- An experiment to study the effect of three different levels of a factor on a response (e.g. three levels of a fertilizer on plant growth)

ANOVA

Two-way ANOVA

- Test for examining the influence of two different categorical independent variables on one continuous dependent variable
- Not only aims at assessing the main effect of each independent variable, but also if there is any interaction between them

Multi-way ANOVA

- Test for examining the influence of more than two different categorical independent variables on one continuous dependent variable

Multiple comparison or post-hoc test

- Assuming that ANOVA detects a significant effect of smoking on the pulmonary health, we can go a step further and examine whether specific population groups differ significantly from one another
- For this purpose, we need to test the differences between pairs of groups. Pairwise multiple comparisons tests, also called post hoc tests, are the right tools to address this issue

ANCOVA

- A general linear model which blends ANOVA and regression
- ANCOVA evaluates whether the means of a dependent variable are equal across levels of a categorical independent variable, while statistically controlling for the effects of other continuous variables that are not of primary interest, known as covariates or nuisance variables
- The standard linear regression assumptions hold; further we assume that the slope of the covariate is equal across all treatment groups (homogeneity of regression slopes)

Repeated measures ANOVA

- When the same subjects are used for each factor (e.g., in a longitudinal study)

Chi-squared test

- A statistical procedure used by researchers to examine the differences between **categorical variables** in the same population
- Test whether there is a statistically significant difference between the expected frequencies and the observed frequencies in one or more categories of a contingency table
- The chi-squared test applies an approximation assuming the sample is large

Fisher's exact test

- A statistical test that is used to analyze contingency tables, where contingency tables are matrices that contain the frequencies of the variables in play
- the Fisher's exact test runs an exact procedure especially for small-sized samples

Trend test

- the aim is to assess for the presence of an association between a variable with two categories and an **ordinal** variable with k categories

Correlation Analysis

Pearson correlation

- A measure of **linear** correlation between two sets of data
- Always has a value between 1 and 1

Wilcoxon rank sum

- Also known as Mann-Whitney U test
- Test whether two independent samples are likely to derive from the same population
- Useful in situations where samples are not normally distributed or small

Wilcoxon signed-rank test

- Statistical test used either to test the location of a set of samples or to compare the locations of two populations using a set of matched samples
- When applied to test the location of a set of samples, it serves the same purpose as the one-sample t-test
- On a set of matched samples, it is a paired difference test like the paired t-test
- Not assume normality

Kruskal-Wallis test

- Non-parametric method for testing whether samples originate from the same distribution
- Useful for comparing two or more independent samples of equal or different sample sizes
- The parametric equivalent is the one-way ANOVA

Spearman correlation

- Nonparametric measure of rank correlation
- Defined as the Pearson correlation coefficient between the rank variables
- While Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not)

Kendall rank correlation

- A statistic used to measure the ordinal association between two measured quantities
- It is a measure of rank correlation: the similarity of the orderings of the data when ranked by each of the quantities
- In the normal case, Kendall correlation is more robust and efficient than Spearman correlation. It means that Kendall correlation is preferred when there are small samples or some outliers.

Regression

Linear Regression

- A linear approach for modelling the relationship between a scalar response and one or more explanatory variables
- Used for prediction or explaining variation in the response variable that can be attributed to explanatory variables

Regression

Logistic Regression

- A statistical model that in its basic form uses a logistic function to model a binary dependent variable
- It can be used to make a classifier by choosing a cutoff value
- It can be extended to model several classes of events

Regression

Multinomial logistic Regression

- A classification method that generalizes logistic regression to multi-class problems
- It can be used to predict the probabilities of the different possible outcomes of a categorically distributed dependent variable, given a set of independent variables (which may be real-valued, binary-valued, categorical-valued, etc.)

Regression

Poisson Regression

- It used to model count data and contingency tables
- Assumes the response variable Y has a Poisson distribution, and assumes the logarithm of its expected value can be modeled by a linear combination of unknown parameters

Negative Binomial Regression

- Generalization of Poisson regression because it loosens the highly restrictive assumption that the variance is equal to the mean made by the Poisson model
- To remedy the problem of overdispersion in Poisson model

Zero-inflated Poisson Regression

- It concerns a random event containing excess zero-count data in unit time
- The zero-inflated Poisson model mixes two zero generating processes. The first process generates zeros. The second process is governed by a Poisson distribution that generates counts, some of which may be zero

Others

Survival Analysis

- Analyzing the expected duration of time until one event occurs, such as death in biological organisms and failure in mechanical systems
- Survival analysis involves the modelling of time to event data

Others

Time-series Analysis

- Deal with data indexed in time order
- Methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data
- Mainly focus on forecasting

Others

Spatial Analysis

- The observations typically relate to geographical locations
- Spatial analysis includes any of the formal techniques which studies entities using their topological, geometric, or geographic properties

Others

Multivariate Analysis

- It address the situations where multiple measurements are made on each experimental unit and the relations among these measurements and their structures are important
- Used for detect hidden factors, clustering

Others

Cluster Analysis

- Task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups
- A main task of exploratory data analysis, and a common technique for statistical data analysis used in many fields, including pattern recognition, image analysis, information retrieval, bioinformatics, data compression, computer graphics and machine learning

Others

Factor Analysis

- A statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors
- A rationale is that the information gained about the interdependencies between observed variables can be used later to reduce the set of variables in a dataset
- The objective is to find out the latent factors that create a commonality

Others

KNN

- A non-parametric supervised learning method used for classification and regression
- the input consists of the k closest training examples in a data set. The output depends on whether KNN is used for classification or regression

Others

Decision Tree model

- Decision Tree is a non-parametric supervised learning method used for classification and regression
- The model of computation in which an algorithm is considered to be basically a decision tree, i.e., a sequence of queries or tests that are done adaptively, so the outcome of the previous tests can influence the test is performed next
- It is called classification tree if dependent variable is discrete
- It is called regression tree if dependent variable is continuous

Others

Naive Bayes Classifier

- Naive Bayes is a simple technique for constructing classifiers
- Naive Bayes is highly scalable, requiring a number of parameters linear in the number of variables in a learning problem
- It only requires a small number of training data to estimate the parameters necessary for classification
- It is mainly used for text data

Others

Anomaly detection

- The identification of rare items, events or observations which deviate significantly from the majority of the data and do not conform to a well defined notion of normal behavior
- It is often used in preprocessing to remove anomalous data from the dataset