

Master Program in *Data Science and Business Informatics*

Statistics for Data Science

Lesson 32 - Multiple-sample tests of the mean and applications to classifier comparison

Salvatore Ruggieri

Department of Computer Science

University of Pisa, Italy

salvatore.ruggieri@unipi.it

The multiple comparisons problem

- Single test $H_0 : \mu = 0$, with significance level $\alpha = 0.05$ [false positive rate]
 - ▶ test is called *significant* when we reject H_0
 - ▶ α is Type I error, probability of rejecting H_0 when it is true
- Multiple tests, say $m = 20$
 - ▶ E.g., $H_0^i : \mu_i = 0$ for $i = 1, \dots, m$ where μ_i is the **expectation of a subpopulation**
- What is the probability of rejecting **at least one** H_0^i when all of them are true?
 - ▶ For independent tests: $P(\cup_{i=1}^m \{p_i \leq \alpha\}) = 1 - P(\cap_{i=1}^m \{p_i > \alpha\}) = 1 - (1 - \alpha)^m$
and then $1 - (0.95)^{20} \approx 0.64$
 - ▶ For dependent tests: $P(\cup_{i=1}^m \{p_i \leq \alpha\}) \leq \sum_i P(\{p_i \leq \alpha\}) = m \cdot \alpha$, and then $\leq 20 \cdot 0.05 = 1$

Family-wise error rate (FWER)

The FWER is the probability of making at least one Type I error in a family of m tests. If the tests are independent:

$$\alpha_{FWER} = 1 - (1 - \alpha)^m$$

If the test are dependent: $\alpha_{FWER} \leq m \cdot \alpha$

Multiple comparisons: corrections

Question: what should be α such that $\alpha_{FWER} \leq b$?

- *Bonferroni correction* (most conservative one):

- ▶ scale significance level $\alpha = b/m$

[invert $b = m \cdot \alpha$]

- ▶ thus $\alpha_{FWER} \leq m \cdot \alpha = b$

Notice: $p \leq \alpha$ is equivalent to scale p-values and test $p \cdot m \leq b$

- *Šidák correction* (exact for independent tests):

- ▶ scale significance level $\alpha = 1 - (1 - b)^{1/m}$

[invert $b = 1 - (1 - \alpha)^m$]

- ▶ thus $\alpha_{FWER} = 1 - (1 - \alpha)^m = b$

Notice: $p \leq \alpha$ is equivalent to scale p-values and test $1 - (1 - p)^m \leq b$

False Discovery Rate and q -values

		True state of nature	
		H_0 is true	H_1 is true
Our decision on the basis of the data	Reject H_0	False Positive	True Positive
	Not reject H_0	True Negative	False Negative

- False Positive Rate: $FPR = FP / (FP + TN)$
 - ▶ Corrections control for FPR since $FWER = P(FP > 0 | H_0^i \ i = 1, \dots, m)$
- Drawback: acting on α increases $FNR = FN / (FN + TP)$
- False Discovery Rate: $FDR = FP / (FP + TP)$ [Korthauer et al, 2019]
 - ▶ $FDR = 0.05$ means 5% of rejected H_0 's are actually true
- **q -value** is $P(H_0 | T \geq t)$ [vs. $p = P(T \geq t | H_0)$]
 - ▶ FDR can be controlled by requiring $q \leq \text{threshold}$

See R script

Omnibus tests and post-hoc tests

- $H_0 : \theta_1 = \theta_2 = \dots = \theta_k [= 0]$
- $H_1 : \theta_i \neq \theta_j$ for some $i \neq j$
- **Omnibus tests** detect any of several possible differences
 - ▶ Advantage: no need to pre-specify which treatments are to be compared ...
... and then no need to adjust for making multiple comparisons
- If H_0 is rejected (test significant), a *post-hoc test* to find which $\theta_i \neq \theta_j$
 - ▶ Everything to everything post-hoc compare all pairs
 - ▶ One to everything post-hoc compare a new population to all the others
- We distinguish a few cases:
 - ▶ Multiple linear regression (normal errors + homogeneity of variances, i.e., $U_i \sim \mathcal{N}(0, \sigma^2)$):
 - F-test + t-test
 - ▶ Equality of means (normal distributions + homogeneity of variances):
 - ANOVA + Tukey/Dunnett
 - ▶ Equality of means (general distributions):
 - Friedman + Nemenyi

F-test for multiple linear regression

- $\mathbf{Y} = \mathbf{X} \cdot \boldsymbol{\beta} + \mathbf{U}$, where $\mathbf{Y} = (Y_1, \dots, Y_n)$, $\mathbf{U} = (U_1, \dots, U_n)$, and $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$
 - ▶ $\boldsymbol{\beta}^T = (\alpha, \beta_1, \dots, \beta_k)$ and $\mathbf{x}_i = (1, x_i^1, \dots, x_i^k)$
 - ▶ Unexplained (residual) error $SSE = S(\boldsymbol{\beta}) = \sum_{i=1}^n (y_i - \mathbf{x}_i \cdot \boldsymbol{\beta})^2$
- Null model (or intercept-only model): $\mathbf{Y} = \mathbf{1} \cdot \alpha + \mathbf{U}$
 - ▶ Total error $SST = S(\alpha) = \sum_{i=1}^n (y_i - \bar{y}_n)^2$ *[residuals of the null model]*
- Explained error $SSR = SST - SSE = \sum_{i=1}^n (\bar{y}_n - \mathbf{x}_i \cdot \boldsymbol{\beta})^2$
- Coefficient of determination $R^2 = SSR/SST = 1 - SSE/SST$ *[See Lesson 20]*
 - ▶ Is the model useful? Fraction of explained error
- **Is the model statistically significant?** *[vs a specific β_i significant? See Lesson 29]*
- $H_0 : \beta_1 = \dots = \beta_k = 0$ $H_1 : \beta_i \neq 0$ for some $i = 1, \dots, k$
- Test statistic: $F = \frac{SSR}{SSE} \frac{n-k-1}{k} \sim F(k, n-k-1)$

See R script

Equality of means: ANOVA

- $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$ *[generalization of two sample t-test]*
- $H_1 : \mu_i \neq \mu_j$ for some $i \neq j$
- datasets $y_1^j, \dots, y_{n_j}^j$ for $j = 1, \dots, k$
 - ▶ Assumption: normality (**Shapiro-Wilk test**) + homogeneity of variances (**Bartlett test**)
 - ▶ responses of $k - 1$ treatments and 1 control group *[one way ANOVA]*
 - ▶ accuracies of k classifiers over $n_j = n$ datasets *[repeated measures/two way ANOVA]*
- Linear regression model over dummy encoded j :

$$Y = \alpha + \beta_1 x_1 + \dots + \beta_{k-1} x_{k-1}$$

- ▶ $\alpha = \mu_k$ is the mean of the reference group ($j = k$)
- ▶ $\beta_j = \mu_j - \mu_k$
- ▶ in R: `lm(Y~Group)` where `Group` contains the labels of $j = 1, \dots, k$
- F -test (over linear regression): $H_0 : \beta_1 = \dots = \beta_k = 0$, i.e., $\mu_j = \mu_k$ for $j = 1, \dots, k$
- **Tukey HSD** (Honest Significant Differences) is an all-pairs post-hoc test
- **Dunnet test** is a one-to-everything test

See R script

Non-parametric test of equality of means: Friedman

- $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$
- $H_1 : \mu_1 \neq \mu_2$ for some $i \neq j$
- datasets x_1^j, \dots, x_n^j for $j = 1, \dots, k$ *[paired observations/repeated measures]*
 - ▶ accuracies of k classifiers over n datasets
- Let r_i^j be the rank of x_i^j in x_i^1, \dots, x_i^k
 - ▶ e.g., j^{th} classifier w.r.t. i^{th} dataset
- Average rank of classifier: $R_j = \frac{1}{n} \sum_{i=1}^n r_i^j$
- Under H_0 , we have $R_1 = \dots = R_k$ and, for n and k large:

$$\chi_F^2 = \frac{12n}{k(k+1)} \left(\sum_{j=1}^k R_j^2 - \frac{k(k+1)^2}{4} \right) \sim \chi^2(k)$$

- Nemenyi test is an all-pairs post-hoc test
- Bonferroni correction is a one-to-everything test
- For unpaired observations, use **Kruskal-Wallis test** instead of Friedman test

See R script

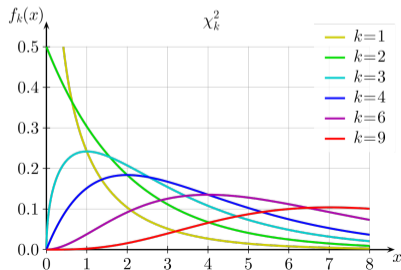
Chi-square distribution

Chi-square distribution

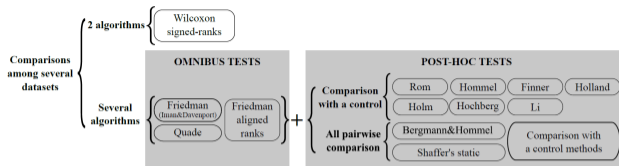
The Chi-square distribution with k degrees of freedom $\chi^2(k)$ has density:

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$$

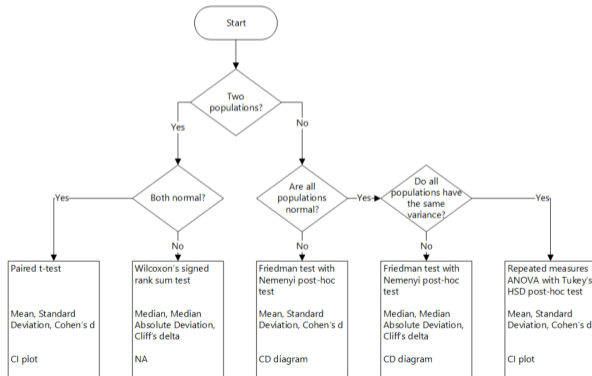
Let $X_1, \dots, X_k \sim \mathcal{N}(0, 1)$. Then $Y = \sum_{i=1}^k X_i^2 \sim \chi^2(k)$



Comparing classifiers: Summary



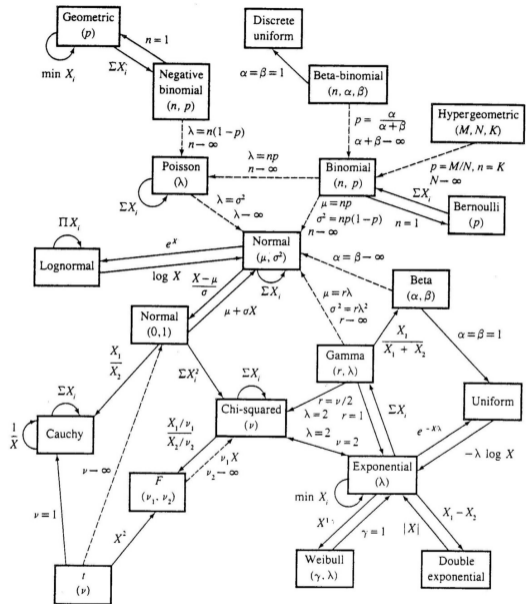
The SCMAMP package in R



The AutoRank package in Python

Common distributions


- Probability distributions at Wikipedia
- Probability distributions in R
-  C. Forbes, M. Evans, N. Hastings, B. Peacock (2010) Statistical Distributions, 4th Edition Wiley




Relationships among common distributions. Solid lines represent transformations and special cases, dashed lines represent limits. Adapted from Leemis (1986).

Optional reference

- On confidence intervals and statistical tests (with R code)

 Myles Hollander, Douglas A. Wolfe, and Eric Chicken (2014)
Nonparametric Statistical Methods.
3rd edition, *John Wiley & Sons, Inc.*

- On False Discovery Rate

 Keegan Korthauer, Patrick K. Kimes, Claire Duvallet, Alejandro Reyes, Ayshwarya Subramanian, Mingxiang Teng, Chinmay Shukla, Eric J. Alm, and Stephanie C. Hicks (2019)
[A practical guide to methods controlling false discoveries in computational biology.](#)
Genome Biology 20, article 118