Master Program in *Data Science and Business Informatics*

Statistics for Data Science

Lesson 33 - Multiple-sample tests of the mean and applications to classifier comparison

Salvatore Ruggieri

Department of Computer Science
University of Pisa
salvatore.ruggieri@unipi.it

The multiple comparisons problem

- Single test $H_0 : \theta = 0$, 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., $H_0^i : \theta_i = 0$ for $i = 1, \dots, m$ where θ_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 n 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

Objective: achieve significant tests ($p \leq \alpha'$) such that $\alpha_{FWER} \leq \alpha$

- *Bonferroni correction* (most conservative one):
  - scale significance level $\alpha' = \alpha/m$ *[invert $\alpha = m \cdot \alpha'$]*
  - Notice: $p \leq \alpha'$ is equivalent to scale p-values and test $p \cdot m \leq \alpha$

  Thus $\alpha_{FWER} \leq m \cdot \alpha' = \alpha$

- *Šidák correction* (exact for independent tests):
  - scale significance level $\alpha' = 1 - (1-\alpha)^{1/m}$ *[invert $\alpha = 1 - (1-\alpha')^m$]*
  - Notice: $p \leq \alpha'$ is equivalent to scale p-values and test $1 - (1-p)^m \leq \alpha$

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

See R script

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 $FWER = P(FP > 0 | H_0^i \; i = 1, \ldots, m)$
- 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 $q \leq$ threshold

See R script

Omnibus tests and post-hoc tests

- $H_0 : \theta_1 = \theta_2 = \ldots = \theta_k$ [$= 0$]
- $H_1 : \theta_i \neq \theta_j$ 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_1 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., $U_i \sim N(0, \sigma^2)$):
    - 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

- $\boldsymbol{Y} = \boldsymbol{X} \cdot \boldsymbol{\beta} + \boldsymbol{U}$, where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, $\boldsymbol{U} = (U_1, \ldots, U_n)$, and $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$
  - $\boldsymbol{\beta}^T = (\alpha, \beta_1, \ldots, \beta_k)$ and $\boldsymbol{x}_i = (1, x_i^1, \ldots, x_i^k)$
  - Unexplained (residual) error $SSE = S(\boldsymbol{\beta}) = \sum_{i=1}^n (y_i - \boldsymbol{x}_i \cdot \boldsymbol{\beta})^2$
- Null model (or intercept-only model): $\boldsymbol{Y} = \mathbf{1} \cdot \alpha + \boldsymbol{U}$
  - Total error $SST = S(\alpha) = \sum_{i=1}^n (y_i - \bar{y}_n)^2$ *[residuals of the null model]*
- Explained error $SSR = SST - SSE = \sum_{i=1}^n (\bar{y}_n - \boldsymbol{x}_i \cdot \boldsymbol{\beta})^2$
- 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]*
- $H_0 : \beta_1 = \ldots = \beta_k = 0 \qquad H_1 : \beta_i \neq 0$ for all $i = 1, \ldots, k$
- Test statistic: $F = \frac{SSR}{SSE} \frac{n-k-1}{k} \sim F(k, n-k-1)$

See R script

Equality of means: ANOVA

- $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ *[generalization of two sample t-test]*
- $H_1 : \mu_1 \neq \mu_2$ for some $i \neq j$
- datasets $y_1^j, \ldots, y_{n_j}^j$ for $j = 1, \ldots, k$
  - 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:

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

  - $\alpha = \mu_k$ 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 $j = 1, \ldots, k$
- F-test (over linear regression): $H_0 : \beta_1 = \ldots = \beta_k = 0$, i.e., $\mu_j = \mu_k$ for $j = 1, \ldots, k$
- **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

- $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$
- $H_1 : \mu_1 \neq \mu_2$ for some $i \neq j$
- datasets $x_1^j, \ldots, x_n^j$ for $j = 1, \ldots, k$ *[paired observations/repeated measures]*
  - accuracies of k classifiers over n datasets
- Let r_i^j be the rank of x_i^j in $x_i^1, \ldots, x_i^k$
  - 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 $R_1 = \ldots = R_k$ 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

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