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 $H_0 : \mu = 0$, with significance level $\alpha = 0.05$ [false positive rate]
  ▶ test is called significant when we reject $H_0$
  ▶ $\alpha$ 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 $i = 1, \ldots, m$ where $\mu_i$ 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: $P(\bigcup_{i=1}^{m} \{ p_i \leq \alpha \}) = 1 - P(\bigcap_{i=1}^{m} \{ p_i > \alpha \}) = 1 - (1 - \alpha)^m$
    and then $1 - (0.95)^{20} \approx 0.64$
  ▶ For dependent tests: $P(\bigcup_{i=1}^{m} \{ p_i \leq \alpha \}) \leq \sum_i P(\{ p_i \leq \alpha \}) = m \cdot \alpha$, 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 $\alpha$ such that $\alpha_{FWER} \leq b$?

- **Bonferroni correction** (most conservative one):
  - scale significance level $\alpha = b/m$
  - 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 $\alpha = 1 - (1 - b)^{1/m}$
  - 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

- 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 $\alpha$ increases $FNR = FN /(FN + TP)$
- False Discovery Rate: $FDR = FP /(FP + TP)$
  - $FDR = 0.05$ means 5% of rejected $H_0$'s are actually true
- $q$-value is $P(H_0|T \geq t)$
  - $FDR$ 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 $H_0$ 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., $U_i \sim N(0, \sigma^2)$):
    - $F$-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

- \( \mathbf{Y} = \mathbf{X} \cdot \mathbf{\beta} + \mathbf{U} \), where \( \mathbf{Y} = (Y_1, \ldots, Y_n) \), \( \mathbf{U} = (U_1, \ldots, U_n) \), and \( \mathbf{X} = (x_1, \ldots, x_n) \)
  - \( \mathbf{\beta}^T = (\alpha, \beta_1, \ldots, \beta_k) \) and \( \mathbf{x}_i = (1, x_{i1}, \ldots, x_{ik}) \)
  - Unexplained (residual) error \( \text{SSE} = S(\mathbf{\beta}) = \sum_{i=1}^{n} (y_i - \mathbf{x}_i \cdot \mathbf{\beta})^2 \)

- Null model (or intercept-only model): \( \mathbf{Y} = 1 \cdot \alpha + \mathbf{U} \)
  - Total error \( \text{SST} = S(\alpha) = \sum_{i=1}^{n} (y_i - \bar{y}_n)^2 \)  \([\text{residuals of the null model}]\)

- Explained error \( \text{SSR} = \text{SST} - \text{SSE} = \sum_{i=1}^{n} (\bar{y}_n - \mathbf{x}_i \cdot \mathbf{\beta})^2 \)

- Coefficient of determination \( R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} \)  \([\text{See Lesson 20}]\)
  - Is the model useful? Fraction of explained error

- **Is the model statistically significant?**  \([\text{vs a specific } \beta_i \text{ significant? See Lesson 29}]\)
  - \( H_0 : \beta_1 = \ldots = \beta_k = 0 \quad H_1 : \beta_i \neq 0 \) for some \( i = 1, \ldots, k \)
  - Test statistic: \( F = \frac{\text{SSR}}{\text{SSE}} \frac{n-k-1}{k} \sim F(k, n-k-1) \)

  \[\text{See R script}\]
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 \(y_{j1}, \ldots, y_{jn_j}\) for \(j = 1, \ldots, k\)
  - Assumption: normality (Shapiro-Wilk test) + homogeneity of variances (Bartlett test)
  - responses of \(k - 1\) treatments and 1 control group
  - accuracies of \(k\) classifiers over \(n_j = n\) datasets
  [one way ANOVA]
  [repeated measures/two way ANOVA]

- Linear regression model over dummy encoded \(j\):
  \[Y = \alpha + \beta_1x_1 + \ldots + \beta_{k-1}x_{k-1}\]
  - \(\alpha = \mu_k\) 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 $x^j_1, \ldots, x^j_n$ for $j = 1, \ldots, k$
  - [paired observations/repeated measures]
  - accuracies of $k$ classifiers over $n$ datasets
- Let $r^j_i$ be the rank of $x^j_i$ in $x^1_i, \ldots, x^k_i$
  - e.g., $j^{th}$ classifier w.r.t. $i^{th}$ dataset
- Average rank of classifier: $R_j = \frac{1}{n} \sum_{i=1}^{n} r^j_i$
- Under $H_0$, we have $R_1 = \ldots = R_k$ and, for $n$ and $k$ large:

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

- 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
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 $X_1, \ldots, X_k \sim N(0, 1)$. Then $Y = \sum_{i=1}^{k} X_i^2 \sim \chi^2(k)$
Common distributions

- Probability distributions at Wikipedia
- Probability distributions in R

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