Master Program in *Data Science and Business Informatics*

Statistics for Data Science

Lesson 33 - Testing independence/association, Multiple sample testing of the mean

Salvatore Ruggieri

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

Testing independence/association: discrete data

- **Pearson's Chi-Square test** of independence
- X and Y discrete (finite) distributions
- $(x_1, y_1) \ldots, (x_n, y_n)$ bivariate observed dataset
- $H_0 : X \perp\!\!\!\perp Y \qquad H_1 : X \not\!\perp\!\!\!\perp Y$
- Test statistic:

$$\chi^2 = \sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} = n \sum_{i,j} \frac{(O_{i,j}/n - p_{i,.}p_{.,j})^2}{p_{i,.}p_{.,j}} \sim \chi^2(df)$$

  where $O_{i,j}$ is the number of observations of value $X = i$ and $Y = j$, $E_{i,j} = np_{i,.}p_{.,j}$ where $p_{i,.} = \sum_j O_{i,j}/n$ and $p_{.,j} = \sum_i O_{i,j}/n$. $df = (n_x - 1)(n_y - 1)$ where n_x (resp., n_y) is the size of the support of X (resp., Y)
- Exact test when n is small: **Fisher's exact test**
- Paired data (e.g., before and after taking a drug): **McNemar's test**

See R script

Association between nominal variables: χ^2-based

- Association measures based on Pearson χ^2 *[See [Lesson 16]*
  - ϕ **coefficient** (or MCC, Matthews correlation coefficient)
    - □ For 2×2 contingency tables: *[Exercise. Show $\phi = |r_{xy}|$]*

$$\phi = \sqrt{\frac{\chi^2}{n}} \in [0, 1]$$

  - **Cramer's V**
    - □ For contingency tables larger than 2×2:

$$V = \sqrt{\frac{\chi^2}{n \cdot \min\{r-1, c-1\}}} \in [0, 1]$$

where r and c are the number of rows and columns

  - **Tschuprov's T** *[sames as V if $r = c$]*
    - □ For contingency tables larger than 2×2:

$$T = \sqrt{\frac{\chi^2}{n \cdot \sqrt{(r-1)(c-1)}}} \in [0, 1]$$

where r and c are the number of rows and columns

See R script

The G-test and Mutual Information

- **G-test** of independence
- X and Y discrete (finite) distributions
- $(x_1, y_1) \ldots, (x_n, y_n)$ bivariate observed dataset
- $H_0 : X \perp\!\!\!\perp Y \qquad H_1 : X \not\perp\!\!\!\perp Y$
- Test statistic:

$$G = 2 \sum_{i,j} O_{i,j} \log \frac{O_{i,j}}{E_{i,j}} = 2 \sum_{i,j} O_{i,j} \log \frac{O_{i,j}}{np_{i,.}p_{.,j}} \sim \chi^2(df)$$

  where $O_{i,j}$ is the number of observations of value $X = i$ and $Y = j$, $E_{i,j} = np_{i,.}p_{.,j}$ where $p_{i,.} = \sum_j O_{i,j}/n$ and $p_{.,j} = \sum_i O_{i,j}/n$. $df = (n_x - 1)(n_y - 1)$ where n_x (resp., n_y) is the size of the support of X (resp., Y)

- Preferrable to Chi-Squared when numbers (O_{ij} or E_{ij}) are small, asymptotically equivalent
- $G = 2 \cdot n \cdot I(O, E)$ where $I(O, E)$ is the mutual information between O and E *[See Lesson 16]*

See R script

Testing correlation: continuous data

- Population correlation:

$$\rho = \frac{E[(X - \mu_X) \cdot (Y - \mu_Y)]}{\sigma_X \cdot \sigma_Y}$$

- Pearson's correlation coefficient:

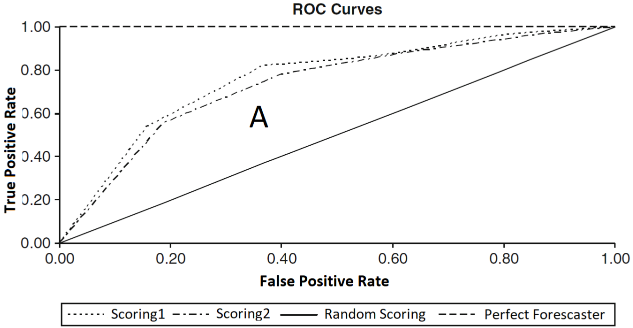
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Assumption: joint distribution of X, Y is bivariate normal (or large sample)
- $(x_1, y_1) \ldots, (x_n, y_n)$ bivariate observed dataset
- $H_0 : \rho = 0 \qquad H_1 : \rho \neq 0$
- Test statistics:

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t(n-2)$$

See R script

Testing AUC-ROC



- Binary classifier score $s_\theta(w) \in [0,1]$ where $s_\theta(w)$ estimate $\eta(w) = P_{\theta_{TRUE}}(C = 1|W = w)$
- ROC Curve
  - $TPR(p) = P(s_\theta(w) \geq p|C = 1)$ and $FPR(p) = P(s_\theta(w)|C = 0)$
  - ROC Curve is the scatter plot $TPR(p)$ over $FPR(p)$ for p ranging from 1 down to 0
  - AUC-ROC is the area below the curve **What does AUC-ROC estimate?**
  - Linearly related to Somer's D correlation index (a.k.a. Gini coefficient) *[See Lesson 16]*

Testing AUC-ROC

- AUC is the probability of correct identification of the order between two instances:

$$AUC = P_{\theta_{TRUE}}(s_\theta(W1) < s_\theta(W2) | C_{W1} = 0, C_{W2} = 1)$$

  where $(W1, C_{W1}) \sim f_{\theta_{TRUE}}$ and $(W2, C_{W2}) \sim f_{\theta_{TRUE}}$

- $s_\theta(W_1), \ldots, s_\theta(W_n) \sim F_{\theta_{TRUE}}|_{C=1}$ and $s_\theta(V_1), \ldots, s_\theta(V_m) \sim F_{\theta_{TRUE}}|_{C=0}$

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} S(s_\theta(W_i), s_\theta(V_j)) \qquad S(X, Y) = \begin{cases} 1 & \text{if } X > Y \\ 1/2 & \text{if } X = Y \\ 0 & \text{if } X < Y \end{cases}$$

  - AUC-ROC $= U/(n \cdot m)$ is an estimator of AUC

- Related to $W = U + \frac{n(n+1)}{2}$, where W is the **Wilcoxon rank-sum test statistics** *[See Lesson 34]*
- Normal approximation, DeLong's algorithm or bootstrap for confidence interval estimation

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 $y^j_1, \ldots, y^j_{n_j}$ 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.*