Statistical Methods for Data Science

Lesson 24 - Testing correlation/independence, Multiple sample testing of the mean.

Salvatore Ruggieri

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

Testing independence: discrete data

- **Pearson's Chi-Square test** of independence
- X and Y discrete (finite) distributions
- $(x_1, y_1) \dots, (x_n, y_n)$ bivariate observed dataset
- $H_0 : X, Y$ independent $\quad H_1 : X, Y$ dependent
- 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

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:

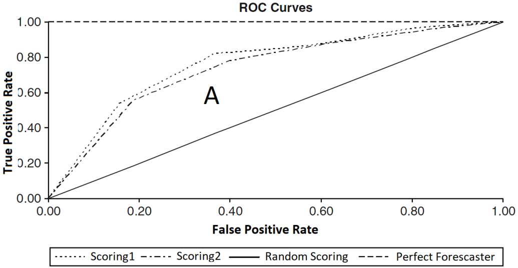
$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

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

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

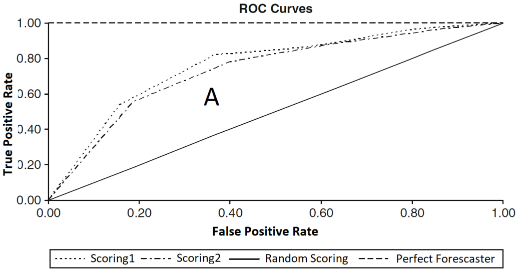
See R script

Testing AUC-ROC



- Binary classifier $f : \boldsymbol{x} \to [0,1]$ with $f(\boldsymbol{x})$ score to be positive
- Let $y_{\boldsymbol{x}}$ be the true class of $\boldsymbol{x}$, either 0 negative or 1 positive
  - calibrated classifier if $P(Y_{\boldsymbol{X}} = 1 | f(\boldsymbol{X}) = p) = p$
- ROC Curve
  - $TPR(p) = P(f(\boldsymbol{X}) \geq p | Y_{\boldsymbol{X}} = 1)$ and $FPR(p) = P(f(\boldsymbol{X}) \geq p | Y_{\boldsymbol{X}} = 0)$
  - ROC Curve is the scatter plot $TPR(p)$ over $FPR(p)$ for p ranging from 1 down to 0
  - AUC (or AUC-ROC) is the area below the curve

Testing AUC-ROC



- Linearly related to Somer's D correlation index (see Lesson 12)
- AUC (or AUC-ROC) is the probability of correct identification of the order

$$P(f(\boldsymbol{W}) < f(\boldsymbol{Z}) | Y_{\boldsymbol{W}} = 0, Y_{\boldsymbol{Z}} = 1)$$

- Related to the Wilcoxon rank-sum test W statistics (actually, equal to the U statistics of **Mann–Whitney *U* test**) $U = W - \frac{n(n+1)}{2}$, see Lesson 23
- 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): $\boldsymbol{Y} = 1 \cdot \alpha + \boldsymbol{U}$
  - 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 $R^2 = SSR/SST = 1 - SSE/SST$
  - Is the model useful? Fraction of explained error
- **Is the model statistically significant?**
- $H_0 : \beta_1 = \ldots = \beta_k = 0 \qquad H_1 : \beta_i \neq 0$ 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

- $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$
- $H_1 : \mu_1 \neq \mu_2$ for some $i \neq j$
- datasets $y^j_1, \ldots, y^j_{n_j}$ for $j = 1, \ldots, k$
  - 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 ($j = k$)
  - $\beta_j = \mu_j - \mu_k$
- F-test: $H_0 : \beta_1 = \ldots = \beta_k = 0$, i.e., $\mu_j = \mu_k$
- **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