

Master Program in *Data Science and Business Informatics*

Statistics for Data Science

Lesson 33 - Fitting distributions. Testing independence/association

Salvatore Ruggieri

Department of Computer Science

University of Pisa, Italy

salvatore.ruggieri@unipi.it

Distribution fitting and quality of fitting

- Dataset x_1, \dots, x_n realization of $X_1, \dots, X_n \sim F$
- **Distribution fitting:** What is a plausible F ?
 - ▶ Useful in Data Science for understanding the data generation process, for checking assumptions (e.g., normality of noise in LR), for checking data distribution changes, etc.

- ▶ Parametric approaches:

- Assume $F = F(\lambda)$ for some family F , and estimate λ as $\hat{\lambda}$

- Maximum Likelihood Estimation (point estimate):

[See Lesson 19]

$$\hat{\lambda} = \operatorname{argmax}_{\lambda} L(\lambda)$$

- Parametric bootstrap (p -value):

[See Lesson 28]

$$T_{ks} = \sup_{a \in \mathbb{R}} |F_n^*(a) - F_{\hat{\lambda}^*}(a)|$$

- ▶ Non-parametric approaches:

- Empirical distribution F_n

[Glivenko-Cantelli Thm]

- Kernel Density Estimation

[See Lesson 15]

- **Quality of fitting:** Among several fits F_1, \dots, F_k , which one is the best?

- ▶ Goodness of fit: measure of how good/bad is F_i in fitting the data?
- ▶ Comparison: which one between two F_1 and F_2 is better?

Quality of fitting

- Loss functions (to be minimized)
 - ▶ Akaike information criterion (AIC), balances model fit against model simplicity

$$AIC(F(\lambda)) = 2|\lambda| - 2\ell(\lambda)$$

- ▶ Bayesian information criterion (BIC), stronger balances over model simplicity

$$BIC(F(\lambda)) = |\lambda| \log n - 2\ell(\lambda)$$

- Statistics (continuous data):

- ▶ **KS test** $H_0 : X \sim F$ $H_1 : X \not\sim F$ with Kolmogorov-Smirnov (KS) statistic:

$$D = \sup_{a \in \mathbb{R}} |F_n(a) - F(a)| \sim K$$

- ▶ **LR test** $H_0 : X \sim F_1$ $H_1 : X \sim F_2$ with the likelihood-ratio test:

$$\lambda_{LR} = \log \frac{L(F_1(\lambda_1))}{L(F_2(\lambda_2))} = \ell(F_1(\lambda_1)) - \ell(F_2(\lambda_2)) \quad \text{with} \quad -2\lambda_{LR} \sim \chi^2(1)$$

See R script

Quality of fitting

- Statistics (discrete data):

- ▶ **Pearson's Chi-Square test**

$H_0 : X \sim F$ $H_1 : X \not\sim F$ with χ^2 statistic:

$$\chi^2 = \sum_{N_i > 0} \frac{(N_i - n_i)^2}{n_i} = n \cdot \sum_{N_i > 0} \frac{(N_i/n - p(i))^2}{p(i)} \sim \chi^2(df)$$

where N_i number of observations of value i , $n_i = n \cdot p(i)$ expected number of observations (rescaled), and $df = |\{i \mid N_i > 0\}| - 1$ is the number of observed values minus 1.

$\chi^2 = \infty$ if for some i : $n_i = 0$

- ▶ **Yates's correction for continuity**

It corrects for approximating the discrete probability of observed frequencies by the continuous chi-squared distribution

$$\chi^2 = \sum_{N_i > 0} \frac{(|N_i - n_i| - 0.5)^2}{n_i}$$

It increases Type II error, so do not use it!

See R script

Comparing two datasets

- Dataset x_1, \dots, x_n realization of $X_1, \dots, X_n \sim F$
- Dataset y_1, \dots, y_m realization of $Y_1, \dots, Y_m \sim G$
- $H_0 : F = G$ $H_1 : F \neq G$
 - ▶ Useful to detect **covariate drift** (data stability) from source to target datasets
- Univariate data:
 - ▶ Continuous data: KS statistics $D = \sup_{a \in \mathbb{R}} |F_n(a) - G_m(a)| \sim K$
 - KS-distance between empirical cumulative distributions
 - ▶ Discrete data: χ^2 statistics

$$\chi^2 = \sum_{R_i > 0 \vee S_i > 0} \frac{(\sqrt{\frac{m}{n}} R_i - \sqrt{\frac{n}{m}} S_i)^2}{R_i + S_i} \sim \chi^2(df)$$

where R_i (resp., S_i) is the number of observations in x_1, \dots, x_n (resp., y_1, \dots, y_m) which are equal to i , $df = |\{i \mid R_i > 0 \vee S_i > 0\}| - 1$

- ▶ Other tests in the R package **twosamples**
- Multivariate data: see classifier 2-sample test and others in the R package **Ecume**

See R script

Testing independence:

- **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
- $H_0 : X \perp\!\!\!\perp Y$ $H_1 : X \not\perp\!\!\!\perp Y$
- Test statistic:

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

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

The G-test and Mutual Information

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

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

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

- Preferable 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

Other tests of independence

- $(x_1, y_1) \dots, (x_n, y_n)$ bivariate observed dataset
- Permutation tests:
 - ▶ reduces to comparing two datasets: $(x_1, y_1) \dots, (x_n, y_n)$ and $(x_1, y_{\pi_1}) \dots, (x_n, y_{\pi_n})$, where π_1, \dots, π_n is a permutation of $1, \dots, n$ *[see slide on comparing two datasets]*
- Continuous X and Y :
 - ▶ discretize both X and Y and then apply independence tests for discrete r.v.'s, or
 - ▶ test correlation (see later), or
 - ▶ **Hoeffding's test**, see R package **independence**
- Continuous X and discrete Y :
 - ▶ discretize X and then apply independence tests for discrete r.v.'s, or
 - ▶ a direct approach **Yang and Kim**, or
 - ▶ special case Y binary: $X \perp\!\!\!\perp Y$ iff $P(X|Y) = P(X)$ iff $P(X|Y=0) = P(X|Y=1)$ *[see slide on comparing two datasets]*

Measures of association (from Lesson 16)

- *Association*: one variable provides information on the other
 - ▶ $X \perp\!\!\!\perp Y$ independent, i.e., $P(X|Y) = P(X)$: zero information
 - ▶ $Y = f(X)$ deterministic association with f invertible: maximum information
- *Correlation*: the two variables show an increasing/decreasing trend
 - ▶ $X \perp\!\!\!\perp Y$ implies $Cov(X, Y) = 0$
 - ▶ the converse is not always true

Variable Y	Variable X		
	Nominal	Ordinal	Continuous
Nominal	ϕ or λ	Rank biserial	Point biserial
Ordinal	Rank biserial	τ_b or Spearman	τ_b or Spearman
Continuous	Point biserial	τ_b or Spearman	Pearson or Spearman

ϕ = phi coefficient, λ = Goodman and Kruskal's lambda,
 τ_b = Kendall's τ_b .

Association between nominal variables: Pearson χ^2 -based

- 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**

[same 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

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:

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

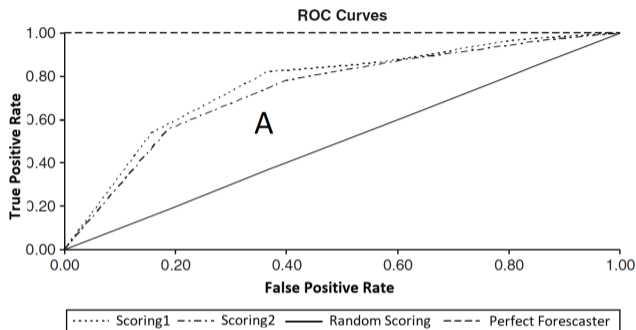
- Assumption: joint distribution of X, Y is bivariate normal (or large sample)
- $(x_1, y_1) \dots, (x_n, y_n)$ bivariate observed dataset
- $H_0 : \rho = 0 \quad H_1 : \rho \neq 0$
- Test statistics:

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

- ▶ Recall that $X \perp\!\!\!\perp Y$ implies $\rho = 0$: if H_0 can be rejected, then $X \perp\!\!\!\perp Y$ can be rejected

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

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

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

- ▶ AUC-ROC = $U/(n \cdot m)$ is an estimator of AUC
- U statistics of the **Wilcoxon rank-sum test** [See Lesson 31]
- Normal approximation, **DeLong's algorithm**, bootstrap, **Brunner-Munzel** tests and confidence intervals

See R script