NON - PARAMETRIC TESTS

DISTRIBUTION

• KOLMOGOROV - SMIRNOV

EQUALITY OF CONTINUOUS, 1-D DISTRIBUTIONS, COMPARE SAMPLE W/REFERENCE. H_0: SAME DISTRIBUTION

• EMPIRICAL DISTRIBUTION

$F_N(x) = \frac{1}{N}\sum_{1}^{m} N_\alpha \cdot U(x - a_\alpha)$, N FREQUENZA CLASSE α ; STATISTIC $D_N = \sup_x |F_N(x) - F(x)|$

COMPARO A DISTRIBUZIONE DA TESTARE
D = DEVIAZIONE MASSIMA, USO QUANTILI
↓
DISTANZA VERTICALE

• GOODNESS - OF - FIT TEST

REJECT H_0 IF $\sqrt{N} \cdot D_N > \lambda_\alpha$; $K(\lambda_\alpha) = \sum_j (-1)^j e^{-2j^2\lambda_\alpha^2} = 1 - \alpha$ WHERE λ_α SATISFIES

FOR $N \geq 20$

• CHI - SQUARED GOF TEST

ALSO NON - CONTINUOUS DISTRIBUTIONS

N SAMPLES, m CLASSES

$$T_N = \sum_{1}^{m} \frac{N_\alpha^2}{N p_\alpha^0} - N$$

$N \geq 50$, $m \geq 5$

$H_0: p_1 = p_{10}, p_2 = p_{20}, \dots$ m-1 DOF

$T_N \geq \chi^2_{1-\alpha,\, m-1\ DOF}$

CATEGORICAL, BINNED DATA

SCOSTAMENTO FRA FREQ. EMPIRICHE E DENSITÀ TEORICHE

$$\sum_{1}^{m} \frac{(N_i - Np_i)^2}{Np_i} = \sum \frac{(\text{OBSERVED} - \text{EXPECTED})^2}{\text{EXPECTED}}$$

• CAN USE FOR CONTINUOUS DISTRIBUTIONS $p_{10} = P(a_{i-1} \leq X \leq a_i) = \int_{a_{i-1}}^{a_i} f_0(x)\,dx$

• WHEN p DEPEND ON θ

MLE ESTIMATE $\hat{\theta}_i$

$$\chi^2 = \sum^{u} \frac{[N - n\pi_i(\hat{\theta})]^2}{n\pi_i(\hat{\theta})}$$

HOMOGENEITY

VERIFICARE SE X, Y ESTRATTI DA STESSA POPOLAZIONE | POP. IDENTICHE

$H_0: F_x(t) = F_y(t)$

• ~~SIGN TEST WILCOXON SIGNED - RANK TEST~~ TEST SEGNI WILCOXON PAIRED DATA

X, Y SAME NUMEROSITY ; $n^\pm$ COUNT WHERE PAIR DIFF IS ± → SAME DISTRIBUTION $|S_N| < \sqrt{\frac{n - n_0}{2}} \cdot z_{1-\alpha}$

N_0 = COPPIE UGUALI

$S_N = n^+ - n^-$ ON PAIR

• SE UGUALE: DIFFERENZA PICCOLA

$n_0 = |x_i = y_i|$ FREQUENCY ; N° OF PAIRS

• WILCOXON ~~SIGNED-RANK~~ RANK-SUM TEST | WILCOXON MANN-WHITNEY

X, Y DIFFERENT NUMEROSITY. | SORT INCREASINGLY | RANK = ORDER NUMBER (AVG IF MORE THAN 1). S_x, S_y = SUMS OF RANKS | SYMMETRIC DISTRIBUTION

$U_x = n \cdot m + \frac{n(n+1)}{2} - S_x$; $U_y = n \cdot m + \frac{m(m+1)}{2} - S_y$, $U = \min(U_x, U_y)$. IF H TRUE $\xrightarrow[n, m \to \infty]{}$ U IS NORMAL WITH $\mu_U = \frac{n \cdot m}{2}$, $\sigma_U^2 = \frac{nm(n+m+1)}{12}$

OK WHEN $n, m \geq 4$, $n + m \geq 20$

$$Z_{n,m} = \frac{U - \overbrace{n \cdot m/2}^{\mu}}{\underbrace{\sqrt{n \cdot m \cdot (n + m + 1)/12}}_{\sigma}}$$; ACCEPT FOR $|Z_{n,m}| < Z_{1-\alpha}$

• SMIRNOV TEST

COUPLE OF DISTRIBUTIONS $F_X(t)$, $F_Y(t)$, $D = \sup_{n,m} |F_{n,x}(t) - F_{m,y}(t)|$ EMPIRICAL DISTRIBUTIONS. IF H_0 TRUE $D \to 0$ FOR INCREASING SAMPLE INDEPENDENTLY OF DISTRIBUTIONS

TWO POPULATIONS

$d_{n,m} < \sqrt{\frac{n + m}{n \cdot m}} \cdot \lambda_\alpha$, SAME OF KS TEST

SAME PRINCIPLE AS KOLMOGOROV - SMIRNOV TEST

CHI-SQUARED HOMOGENEITY TEST

STILL BINNED DATA NEEDED

SUPER-GENERAL
MULTIPLE REALIZATIONS, m CLASSI

$j = 1..N$ $\quad \sum_{1}^{M} n_{kj} = n_j = N°$ OF X_j $\quad \sum_{j=1}^{M}\sum_{k}^{m} n_{kj} = \sum_{j=1}^{M} n_j = N$, TOTAL OBSERVATIONS

RELATIVE FREQUENCIES OF CLASSES ARE SAME FOR ALL SAMPLES, $\hat{p}_k^0$, $p_1^0 + \dots + p_m^0 = 1$

$$T_N(\hat{p}^0) = N\left(\sum_{j=1}^{M}\sum_{k=1}^{m} \frac{n_{j,k}^2}{n_j \cdot n_k} - 1\right)$$

$$\hat{t}_N < \chi_{1-\alpha,(m-1)(M-1)} \text{ DOF}$$

- ALT FORMULATION

- IF $m = 2$, IN/OUT, TRUE/FALSE

$$T_N(n_A) = \frac{N}{N - n_A}\left(\frac{N}{n_A}\sum_{j=1}^{M} \frac{n_{Aj}^2}{n_j} - n_A\right)$$

- $\hat{e}_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}$, $\quad \chi^2 = \sum_i^I \sum_j^J \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$

$$\chi = \sum \frac{(OBS - EXP)^2}{EXP}$$

I POPULATIONS
J TREATMENTS/BINS/CATEGORIES

INDEPENDENCE TESTS

CHI-SQUARED INDEPENDENCE TESTS

DISCRETE RV, OR DIVIDED IN CLASSES
$(X_1, \dots X_{nx})$ $(Y_1, \dots Y_{ny})$

$n_{hk} = N°$ COPPIE $\quad \sum_{h=1}^{mx} n_{hk} = n_k$, $\quad \sum_{k=1}^{my} n_{hk} = n_h$; $\quad \sum_1^{mx}\sum_1^{my} n_{hk} = n$

- n_h, n_k MARGINAL ABSOLUTE FREQUENCIES
- $(h-1)(k-1)$ DOF

MLES $\forall h, k$ $\quad \hat{p}_h^0 = \frac{n_h}{n}$, $\hat{p}_k^0 = \frac{n_k}{n}$ → MINIMIZING $\quad \frac{n_{hk}}{n} - \hat{p}_h^0 \cdot \hat{p}_k^0 = \left(-\frac{n_{hk}}{n} + n_{hk}\right)\frac{1}{n}$

OK IF $\forall n_{hk} \geq 5$

$$T_m(\hat{p}^0) = m\left(\sum_{h=1}^{mx}\sum_{k=1}^{my} \frac{n_{hk}^2}{n_h \cdot n_k} - 1\right)$$

REJECT H_0 IF $t_m \geq \chi_{1-\alpha,(mx-1)(my-1)}$

ALT FORMULATION

$\hat{e}_{ij} = n \cdot p_i \cdot p_j = \frac{n_{i\cdot} \, n_{\cdot j}}{n}$

$\chi^2 = \sum_i^I \sum_j^J \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}$ $\quad \chi = \frac{(O-E)^2}{E}$

SPEARMAN TEST (RANK CORR COEFF)

X_k, Y_k • SORT, COMPUTE RANKS SEPARATELY $\quad d_k = Rk(X_k) - Rk(Y_k)$

$$R_S = 1 - \frac{6}{n(n^2-1)} \cdot \sum_{k=1}^{n} d_k$$

SPEARMAN CORR COEFF

- NECESSARY, NOT SUFFICIENT CONDITION FOR INDEPENDENCE

$$-1 \leq R_S \leq 1$$

H_0: X_k ~~INDEPENDENCE~~ NONCORRELATION Y_k → $AVG(R_S) = 0$

$n \geq 10$ • $T_S = R_S\sqrt{\frac{n-2}{1-R_S^2}}$ IS T-STUDENT, $n-2$ DOF

REJECT H_0 IF $|t_s| > t_{1-\alpha/2}$

RANDOMNESS TESTS

RANDOMNESS → INDEPENDENCE $F_X(x_1 \dots x_n) = F_{X_1}(x_1) \cdot F_{X_2}(x_2) \cdots F_{X_N}(x_N)$ H_0: SAMPLES ARE RANDOM

SERIAL CORRELATION TEST

IF RANDOM → ANY SUBSET (x_n, x_{n+1}) IS INDEPENDENT $R_S = \sum_{n=1}^{N} \frac{(x_n - \bar{x})(x_{n+1} - \bar{x})}{N S_X^2}$ CIRCULAR CORRELATION COEFFICIENT

- PERMUTATIONS ARE EQUIPROBABLE → $R_X = \sum_{n=1}^{N} (x_n - \bar{x})(x_{n+1} - \bar{x})$
- IF X RANDOM, $R_X \sim N$ $n \to \infty$

SE NORMALE

- $E[R_X] = -\frac{S_2}{n-1}$ $\sigma^2[R_X] = \frac{S_2^2 - S_4}{N-1}$, $S_N = (x_1 - \bar{x})^q + (x_2 - \bar{x})^q + \dots + (x_N - \bar{x})^q$ $q = 2, 4,$
- STANDARDIZE → $Z_X = \frac{R_X - E[R_X]}{\sigma[R_X]}$
- REJECT H_0 $|Z_X| > z_{1-\alpha/2}$

RUN TEST

- RUN: N° OF IDENTICAL SUBSTRINGS 'AABBAAABA...' 2 CLASSES ONLY U = TOTAL N° OF RUNS IN SEQUENCES
- IF RANDOM U WILL BE FAR FROM EXTREME VALUES DUE TO RANDOM = MAX INFORMATION
- GOOD FOR $N_A, N_B > 10$, $U \to$ NORMAL $E[U] = 1 + \frac{2 N_A N_B}{N_A + N_B}$ $\sigma_U^2 = \frac{2 N_A N_B (2 N_A N_B - N_A - N_B)}{(N_A + N_B)^2 (N_A + N_B - 1)}$
- STANDARDIZE → $Z_U = \frac{U - E[U]}{\sigma_U}$
- REJECT H_0 $|Z_U| > z_{1-\alpha/2}$
- CAN USE ON REALS; IF CONSIDER $\begin{cases} A: x < \hat{x} \text{ (MEDIAN)} \\ B: x > \hat{x} \end{cases}$