Sia $X_1 \ldots X_n$ campione statistico e supponiamo che sia distribuito su un insieme finito $t_1 \ldots t_k$.

Conosco la distribuzione campionaria se

$\forall j = 1 \ldots k$ conosco $p_j := \mathbb{P}(X_i = t_j)$

$x_1 \ldots x_n$

Ce ne sono n_1 che assumono il valore t_1

n_2 che assumono il valore t_2

$\vdots$

n_k che assumono il valore t_k.

$$n_j \geq 0 \qquad \sum_{j=1}^{k} n_k = n$$

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i) = \prod_{j=1}^{k} p_j^{n_j}$$

$$= g(x_1 \ldots x_n \mid p_1 \ldots p_k)$$

$$p_j > 0 \qquad \sum_{j=1}^{k} p_j = 1$$

$$h(p_1 \ldots p_k) = \ln g(x_1 \ldots x_n \mid p_1 \ldots p_k) = \ln\left(\prod_{j=1}^{k} p_j^{n_j}\right) = \sum_{j=1}^{k} n_j \ln(p_j)$$

$$\longrightarrow \begin{cases} \max h(p_1 \ldots p_k) \\ \sum_{j=1}^{k} p_j = 1 \\ p_j > 0 \end{cases} \qquad p_k = 1 - \sum_{j=1}^{k-1} p_j$$

$$\sum_{j=1}^{k} n_j \ln(p_j) = \sum_{j=1}^{k-1} n_j \ln(p_j) + n_k \ln\left(1 - \sum_{j=1}^{k-1} p_j\right)$$
$$= \tilde{h}(p_1 \ldots p_{k-1})$$

$$\begin{cases} p_j > 0 \quad \forall j = 1 \ldots k-1 \\ 1 - \sum_{j=1}^{k-1} p_j > 0 \end{cases}$$

$$\tilde{h}(p_1 \ldots p_{k-1}) = \sum_{j=1}^{k-1} n_j \ln(p_j) + n_k \ln\left(1 - \sum_{j=1}^{k-1} p_j\right)$$

$$\tilde{h}(p_1 \ldots p_{k-1}) = \sum_{j=1} n_j \ln(p_j) + n_k \ln\left(1 - \sum_{j=1} p_j\right)$$

$\forall s = 1 \ldots k-1$ $\qquad \dfrac{\partial \tilde{h}}{\partial p_s} = 0$

$$\frac{\partial \tilde{h}}{\partial p_s} = n_s \cdot \frac{1}{p_s} + n_k \frac{1}{1 - \sum_{j=1}^{k-1} p_j} (-1) = 0$$

$$\frac{n_s}{p_s} - \frac{n_k}{p_k} = 0 \qquad \frac{n_s}{p_s} = \frac{n_k}{p_k} \qquad \frac{p_s}{n_s} = \frac{p_k}{n_k}$$

$$p_s = \frac{n_s}{n_k} p_k \qquad \forall s = 1 \ldots k-1$$

$$1 = \sum_{j=1}^{k} p_j = \left(\sum_{j=1}^{k-1} p_j\right) + p_k = \left(\sum_{j=1}^{k-1} \frac{n_j}{n_k} p_k\right) + p_k =$$

$$= \frac{p_k}{n_k}\left(\sum_{j=1}^{k-1} n_j\right) + p_k = \frac{p_k}{n_k}\left(\underbrace{\sum_{j=1}^{k-1} n_j + n_k}_{n}\right) = p_k \frac{n}{n_k}$$

$\Rightarrow$ $p_k = \dfrac{n_k}{n}$

$s = 1 \ldots k-1$ $\qquad p_s = \dfrac{n_s}{n_k} p_k = \dfrac{n_s}{n_k} \dfrac{n_k}{n} = \dfrac{n_s}{n}$

$\Rightarrow$ $\forall j = 1 \ldots k$ $\qquad p_j = \dfrac{n_j}{n}$ = frequenze relative delle modalità t_j nel campione $x_1 \ldots x_n$.

— o —

$Y_1 \ldots Y_n$ campione statistico distribuito sull'insieme finito $t_1 \ldots t_k$.

Siano $p_1^0 \ldots p_k^0$ numeri assegnati che soddisfano le condizioni
$$\begin{cases} p_j^0 \geq 0 & \forall j = 1 \ldots k \\ \sum_{j=1}^{k} p_j^0 = 1 \end{cases}$$

H0) La densità del campione in t_j è $p_j = p_j^0$ $\forall j = 1 \ldots k$.

$H_0)$ La densità del campione in t_j è $p_j \equiv p_j^0 \quad \forall j = 1 \ldots k$.

$H_A)$ $\exists j \in \{1 \ldots k\}$ t.c. la densità del campione in t_j è diversa da p_j^0.

— o —

Per ogni $j = 1 \ldots k$. definisco

$$X_j = \#\{i \in \{1 \ldots n\} : Y_i = t_j\} = \sum_{i=1}^{n} \mathbb{1}_{\{t_j\}}(Y_i)$$

$$\mathbb{P}\left(\mathbb{1}_{\{t_j\}}(Y_i) = 1\right) = p_j \qquad \left\{\mathbb{1}_{\{t_j\}}(Y_i) = 1\right\} = \{Y_i = t_j\}$$

$$\mathbb{P}_{\mathbb{1}_{\{t_j\}}(Y_i)} = B(p_j) \quad \forall i = 1 \ldots n \qquad \text{dove } p_j = \mathbb{P}(Y_i = t_j)$$

$$\Downarrow$$

$$\mathbb{P}_{X_j} = Bin(n, p_j) \qquad \mathbb{E}[X_j] = np_j$$

$$Var[X_j] = np_j(1 - p_j)$$

$$|X_j - np_j^0|^2$$

$$a_1 > 0 \;\_\_\_\; a_k > 0 \qquad \sum_{j=1}^{k} a_j (X_j - np_j^0)^2 < \varepsilon$$

TEOREMA DI PEARSON

Siano $X_1 \ldots X_k$ v.a. $\mathbb{P}_{X_j} = Bin(n, p_j^0) \quad \forall j = 1 \ldots k$.

Allora la v.a. $\displaystyle\sum_{j=1}^{k} \frac{(X_j - np_j^0)^2}{np_j^0}$ converge in legge ad una v.a. avente distribuzione χ^2_{k-1}.

Allora scelgo $a_j = \dfrac{1}{np_j^0} \quad \forall j = 1 \ldots k$

Accetto l'ipotesi nulla sse $\displaystyle\sum_{j=1}^{k} \frac{(x_j - np_j^0)^2}{np_j^0} < \varepsilon$

Accetto l'ipotesi nulla ssE $\displaystyle\sum_{j=1} \frac{(x_j - np_j^\circ)^2}{np_j^\circ} < \varepsilon$

dove x_j è la realizzazione di X_j sui miei dati cioè

accetto H_0 ssE $\displaystyle\sum_{j=1}^{k} \frac{(n_j - np_j^\circ)^2}{np_j^\circ} < \varepsilon$.

$$\alpha = \mathbb{P}\left(\sum_{j=1}^{k} \frac{(X_j - np_j^\circ)^2}{np_j^\circ} \geq \varepsilon \;\middle|\; p_j = p_j^\circ \quad \forall j = 1 \dots k \right) =$$

$$= 1 - \mathbb{P}\left(\sum_{j=1}^{k} \frac{(X_j - np_j^\circ)^2}{np_j^\circ} < \varepsilon \;\middle|\; p_j = p_j^\circ \quad \forall j = 1 \dots k \right)$$

$$\xrightarrow{n \to \infty} 1 - F_{\chi^2_{k-1}}(\varepsilon)$$

$$F_{\chi^2_{k-1}}(\varepsilon) = 1 - \alpha \qquad \boxed{\varepsilon = \chi^2_{k-1,\,1-\alpha}}$$

$k = 2 \quad t_1 \quad t_2$

$X_1 = \#\{ i \in \{1 \dots n\} : Y_i = t_1 \}$

$X_2 = \#\{ i \in \{1 \dots n\} : Y_i = t_2 \}$

$p_1^\circ, p_2^\circ \geq 0 \qquad p_1^\circ + p_2^\circ = 1 \qquad X_1 + X_2 = n$

$X_2 = n - X_1$

$p_2^\circ = 1 - p_1^\circ$

$$\frac{(X_1 - np_1^\circ)^2}{np_1^\circ} + \frac{(X_2 - np_2^\circ)^2}{np_2^\circ}$$

$$= \frac{(X_1 - np_1^\circ)^2}{np_1^\circ} + \frac{(\cancel{n} - X_1 - n(\cancel{1} - p_1^\circ))^2}{n(1 - p_1^\circ)}$$

$$\frac{1 - \cancel{p_1^\circ} + \cancel{p_1^\circ}}{p_1^\circ(1 - p_1^\circ)}$$

$$= \frac{(X_1 - np_1^\circ)^2}{np_1^\circ} + \frac{(X_1 - np_1^\circ)^2}{n(1 - p_1^\circ)} = \frac{(X_1 - np_1^\circ)^2}{n}\left(\frac{1}{p_1^\circ} + \frac{1}{1 - p_1^\circ} \right)$$

$$\frac{(X_1 - np_1^\circ)^2}{\dots} \left(\frac{X_1 - np_1^\circ}{\dots} \right)^2 =$$

$$= \frac{(X_1 - np_1^0)^2}{np_1^0(1-p_1^0)} = \left(\frac{X_1 - np_1^0}{\sqrt{np_1^0(1-p_1^0)}} \right)^2 =$$

$$X_1 = \sum_{i=1}^{n} \mathbb{1}_{\{t_1\}}(Y_i) \qquad = \left(\underbrace{\frac{\sum_{i=1}^{n} \mathbb{1}_{\{t_1\}}(Y_i) - n p_1^0}{\sqrt{n\, p_1^0(1-p_1^0)}}}_{} \right)^2$$

Legge

v.a Z t.c. $\mathbb{P}_Z = N(0,1)$

So che $\mathbb{P}_{Z^2} = \chi^2_1$

— o —

TEO

Sia X una v.a. con legge F. Allora la v.a $Y = F \circ X$ ha distribuzione $\mathbb{P}_Y = U([0,1])$.

DIM. Lo dimostriamo solo nel caso in cui la distribuzione sia A.C. cioè $\exists f : \mathbb{R} \to \mathbb{R}$ non negativa, funzione di Lebesgue t.c. $\mathbb{P}_X = f(x)dx$ cioè $F(t) = \int_{-\infty}^{t} f(x)\,dx \quad \forall t \in \mathbb{R}$.

$\varphi : \mathbb{R} \to \mathbb{R}$ di Borel non negativa

$$\int_{\mathbb{R}} \varphi(t)\, \underbrace{\mathbb{P}_{F\circ X}}_{}(dt) = \int_{\mathbb{R}} (\varphi \circ F)(x)\, \mathbb{P}_X(dx) = \int_{\mathbb{R}} \varphi(F(x))\, \underbrace{f(x)dx}_{} =$$

$u = F(x) \qquad du = F'(x)dx = f(x)dx$

$x \to -\infty \qquad u \to 0$

$x \to +\infty \qquad u \to 1$

densità associata alla distribuzione $U([0,1])$

$$= \int_0^1 \varphi(u)\,du = \int_{\mathbb{R}} \varphi(u)\, \overbrace{\mathbb{1}_{[0,1]}(u)}^{}\, du$$

$$= \int_0^1 f(u)\,du = \int_{\mathbb{R}} f(u)\,\mathbb{1}_{[0,1]}(u)\,du$$

$$= \int_{\mathbb{R}} f(u)\,\mathbb{P}_U(du) \qquad \Rightarrow \mathbb{P}_{F\circ X} = U([0,1])$$

dove U è una v.a. con distribuzione $U([0,1])$

— o —

$X_1 \ldots X_n$ campione statistico. Suppongo che la legge associata F sia continua.

Per ogni $t \in \mathbb{R}$ e ogni $i = 1 \ldots n$.

$$Y_i(\omega, t) = \mathbb{1}_{(-\infty, t]}(X_i(\omega)) = \begin{cases} 1 & X_i(\omega) \le t \\ 0 & X_i(\omega) > t \end{cases}$$

Le $Y_1 \ldots Y_n$ sono indipendenti

$$\mathbb{P}(Y_i(\cdot, t) = 1) = \mathbb{P}(X_i \le t) = F(t)$$

$$\mathbb{P}_{Y_i} = Ber(F(t)) \qquad \mathbb{E}[Y_i] = F(t)$$

$$Var[Y_i] = F(t)(1 - F(t)) \le \frac{1}{4}$$

$$G_n(\omega, t) = \frac{1}{n} \sum_{i=1}^n Y_i(\omega, t)$$

$$D_n(\omega) := \sup_{t \in \mathbb{R}} |G_n(\omega, t) - F(t)|$$

n.b. $F(t) = \mathbb{E}[G_n(\cdot, t)]$

$$\mathbb{P}(|G_n(\cdot, t) - F(t)| > \varepsilon) \le \frac{1}{n\varepsilon^2} Var[Y_i] \le \frac{1}{4n\varepsilon^2}$$

— o —

Legge di D_n, $d > 0$

$$\mathbb{P}(D_n \ge d) = \mathbb{P}\left(\sup_{t \in \mathbb{R}} |G_n(\cdot, t) - F(t)| \ge d\right)$$

$$= \mathbb{P}\left(\sup \left|\frac{1}{n} \#\{i \in \{1 \ldots n\} : X_i \le t\} - F(t)\right| \ge d\right) =$$

$$= \mathbb{P}\left(\sup_{t \in \mathbb{R}} \left| \frac{1}{n} \# \{ i \in \{1 \dots n\} : \underbrace{X_i \le t}_{F(X_i) \le F(t)} \} - F(t) \right| \geq d \right) =$$

$X_i \le t \iff F(X_i) \le F(t)$ se F è strettamente monotona crescente.

Monotona crescente

$X_i \le t \Rightarrow F(X_i) \le F(t)$

se $F(X_i) \le F(t)$ e $X_i > t$

è una v.a. U con distribuzione $U([0,1])$

$$\mathbb{P}\left(\sup_{t \in \mathbb{R}} \left| \frac{1}{n} \# \{ i \in \{1 \dots n\} : F(X_i) \le F(t) \} - F(t) \right| \geq d \right) =$$

$F(t)$ continua $\Rightarrow$ prende tutti i valori compresi in $(0,1)$

$$= \mathbb{P}\left(\sup_{y \in (0,1)} \left| \frac{1}{n} \# \{ i \in \{1 \dots n\} : U_i \le y \} - y \right| \geq d \right)$$

— o —

$$\lim_{n \to \infty} \mathbb{P}(D_n \sqrt{n} \le t) = \begin{cases} 0 & t \le 0 \\ 1 - 2 \sum_{j=1}^{\infty} (-1)^{j-1} \exp(-2 j^2 t^2) & t > 0 \end{cases}$$

TEST DI KOLMOGOROV - SMIRNOV

Sia $X_1 \dots X_n$ campione statistico di cui rileviamo i dati $x_1 \dots x_n$.

Sia $F_0 : \mathbb{R} \to \mathbb{R}$ funzione continua, monotona crescente,

$$\lim_{t \to -\infty} F_0(t) = 0$$

$$\lim_{t \to +\infty} F_0(t) = 1$$

$H_0)$ F_0 è la legge del campione

$H_A)$ F_0 non è legge del campione.

$$G_n(\omega, t) = \frac{1}{n}\sum_{i=1}^{n} 1\!\!1_{(-\infty,t]}(X_i(\omega))$$

$$g_n(x_1 \_ x_n, t) = \frac{1}{n}\sum_{i=1}^{n} 1\!\!1_{(-\infty,t]}(x_i) = \underbrace{\frac{1}{n}\#\{i : x_i \le t\}}$$

$$\underbrace{\sup_{t\in\mathbb{R}} \left| g_n(x_1 \_ x_n, t) - F_0(t) \right|}_{} < \varepsilon$$

è la realizzazione di D_n sui miei dati

$$\mathbb{P}(D_n \geq \varepsilon) = \mathbb{P}(D_n\sqrt{n} \geq \varepsilon\sqrt{n}) \sim 2\sum_{j=1}^{\infty}(-1)^{j-1}\exp(-2j^2\varepsilon^2 n)$$

$a_j = \exp(-2j^2\varepsilon^2 n)$ $\qquad$ $a_j > a_{j+1}$

$a_{j+1} = \exp(-2(j+1)^2\varepsilon^2 n)$

$a_1 - a_2 + a_3 - a_4 + a_5 - - - -$

$a_1 + \underbrace{(a_3 - a_2)}_{<0} + \underbrace{(a_5 - a_4)}_{<0} + \_\_\_\_ < a_1$

$$\mathbb{P}(D_n \geq \varepsilon) \leq 2\exp(-2\varepsilon^2 n) = \alpha$$

$\exp(-2\varepsilon^2 n) = \frac{\alpha}{2}$ $\qquad$ $-2\varepsilon^2 n = \ln\left(\frac{\alpha}{2}\right)$ $\qquad$ $\varepsilon^2 = \frac{1}{2n}\ln\left(\frac{2}{\alpha}\right)$

$$\varepsilon = \sqrt{\frac{1}{2n}\ln\left(\frac{2}{\alpha}\right)}$$

$$\sup_{t\in\mathbb{R}} \left| \frac{1}{n}\#\{i : x_i \le t\} - F_0(t) \right|$$

$$x_1 < x_2 < \dots < x_n$$

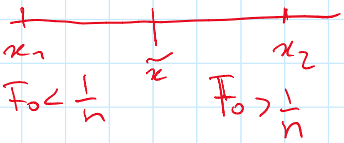
$t < x_1$ $\qquad \left|\frac{1}{n}\#\{i : x_i \le t\} - F_0(t)\right| = |F_0(t)| \le F_0(x_1)$

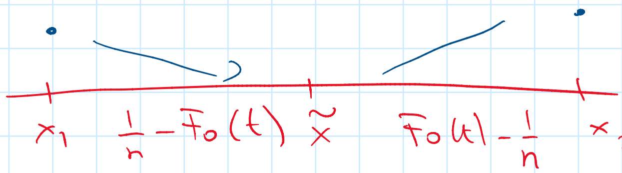
$$\sup_{t<x_1} |F_0(t)| = F_0(x_1)$$

$x_1 \le t < x_2$ $\qquad \left|\frac{1}{n}\#\{i : x_i \le t\} - F_0(t)\right| = \left|\frac{1}{n} - F_0(t)\right|$

$F_0(t) \le \frac{1}{n} \quad \forall t \in [x_1, x_2) \qquad \sup_{t\in[x_1,x_2)} \frac{1}{n} - F_0(t) = \frac{1}{n} - F_0(x_1)$

$F_0(t) \ge \frac{1}{n} \quad \forall t \in [x_1, x_2) \qquad \sup_{t\in[x_1,x_2)} F_0(t) - \frac{1}{n} = F_0(x_2) - \frac{1}{n}$





$$\sup_{t\in[x_1,x_2)} \left|\frac{1}{n} - F_0(t)\right| = \max\left\{\frac{1}{n} - F_0(x_1), F_0(x_2) - \frac{1}{n}\right\}$$

$x_2 \le t < x_3$ $\qquad \left|\frac{1}{n}\#\{i : x_i \le t\} - F_0(t)\right| = \left|\frac{2}{n} - F_0(t)\right|$

$$\max\left\{\left|\frac{2}{n} - F_0(x_2)\right|, \left|\frac{2}{n} - F_0(x_3)\right|\right\}$$

$$\sup_{t\in\mathbb{R}} \left\{\left|\frac{1}{n}\#\{i\in\{1,\dots,n\} : x_i \le t\} - F_0(t)\right|\right\} =$$

$$= \max\Big\{ F_0(x_1), \left|\frac{1}{n} - F_0(x_1)\right|, \left|\frac{1}{n} - F_0(x_2)\right|,$$

$$\left|\frac{2}{n} - F_0(x_2)\right|, \left|\frac{2}{n} - F_0(x_3)\right|, \dots$$

$$\dots \left|\frac{n-1}{n} - F_0(x_{n-1})\right|, \left|\frac{n-1}{n} - F_0(x_n)\right|,$$

$$, \left| 1 - F_0(x_n) \right| \Big\}$$