- CHIAMIAMO <u>MEDIA CAMPIONARIA</u> (O SEMPLICEMENTE MEDIA) DI $m$ DATI NUMERICI $\{x_1, \dots, x_m\}$ LA QUANTITÀ

$$\bar{x} = \frac{1}{m} \sum_{j=1}^{m} x_j$$

<u>ESEMPIO</u> LE MISURE OTTENUTE SU UN CAMPIONE SONO

18 6 31 71 84 17 23 1 9 43

ABBIAMO $m=10$ E UN SEMPLICE CONTO FORNISCE $\bar{x} = 30,3$.

<u>MEDIA PER DATI RAGGRUPPATI</u>. SE I DATI SONO STATI GIÀ SUDDIVISI IN CLASSI E SI CONOSCE SOLO LA FREQUENZA, ALLORA SI PUÒ DEFINIRE LA MEDIA IN QUESTO MODO: SI ASSUME CHE I DATI NELLA SINGOLA CLASSE SIANO DISTRIBUITI IN MODO UNIFORME E CHE QUINDI IL LORO VALORE MEDIO COINCIDA CON IL VALORE MEDIO DELL'INTERVALLINO CHE REALIZZA LA CLASSE. SIA $\bar{x}_k$ TALE VALORE PER LA CLASSE $k$-ESIMA. ASSUMENDO CHE ABBIAMO $N$ CLASSI, $\bar{x}$ È OTTENUTO ATTRAVERSO UNA MEDIA PONDERATA DEGLI $\bar{x}_k$ CON PESI $f_a(k)$:

$$\bar{x} = \sum_{k=1}^{N} f_r(k)\,\bar{x}_k = \frac{1}{m} \sum_{k=1}^{N} f_a(k)\,\bar{x}_k$$

TORNANDO ALL'ESEMPIO CON IL PESO DEGLI STUDENTI, ABBIAMO

$$\bar{P} = \frac{5\cdot 61 + 18\cdot 64 + 42\cdot 67 + 27\cdot 70 + 8\cdot 73}{100} = 67,45 \text{ kg}$$

QUALCHE SEMPLICE PROPRIETÀ

- SE APPLICHIAMO UNA TRASFORMAZIONE LINEARE AI DATI

$$y_j = a x_j + b \implies \bar{y} = a\bar{x} + b$$

CIOÈ LA MEDIA COMMUTA CON LA TRASFORMAZIONE

- LA SOMMA DEGLI SCARTI DALLA MEDIA È NULLA

$$\sum_{j=1}^{n} (x_j - \bar{x}) = 0$$

- LA SOMMA DEI QUADRATI DAGLI SCARTI DELLA MEDIA È MINIMA:

$$\sum_{j=1}^{n} (x_j - x)^2 \text{ ASSUME VALORE MINIMO PER } x = \bar{x}$$

<u>OSSERVAZIONI</u> MEDIA E MEDIANA NON COINCIDONO NECESSARIAMENTE, MA SONO TANTO PIÙ VICINE QUANTO PIÙ I DATI SONO DISPOSTI UNIFORMEMENTE. LA MEDIA È PIÙ FACILE DA CALCOLARE, MA DIPENDENDO DA I VALORI DI TUTTI I DATI RISULTA SENSIBILE ALLA PRESENZA DI ANOMALIE.

LA MEDIANA È MENO SENSIBILE, E PUÒ RISULTARE UGUALE TRA DUE SET DI DATI ANCHE SE QUESTI DIFFERISCONO NOTEVOLMENTE.

IL PRINCIPALE INDICE DI DISPERSIONE È LA <u>VARIANZA</u>: DATI $m$ DATI NUMERICI $\{x_1, \dots, x_m\}$, È LA QUANTITÀ

$$\sigma^2 = \frac{1}{m} \sum_{j=1}^{m} (x_j - \bar{x})^2$$

LA RADICE QUADRATA DELLA VARIANZA SI CHIAMA <u>DEVIAZIONE STANDARD</u> O <u>SCARTO QUADRATICO MEDIO</u>

$$\sigma = \sqrt{\frac{1}{m} \sum_{j=1}^{m} (x_j - \bar{x})^2}$$

VARIANZA E DEVIAZIONE STANDARD SONO QUANTITÀ NON NEGATIVE, E SI ANNULLANO SOLO QUANTO $x_1 = \dots = x_m = \bar{x}$.

INTUITIVAMENTE, PIÙ È PICCOLA LA VARIANZA PIÙ I VALORI SI CONCENTRANO VICINO A $\bar{x}$ (PER QUESTO LA VARIANZA È UN INDICE DI DISPERSIONE).

PER IL CALCOLO DELLA VARIANZA SI PUÒ USARE LA SEGUENTE FORMULA DI FACILE VERIFICA

FORMULA, DI FACILE VERIFICA

$$\sigma^2 = \left( \frac{1}{m} \sum_{j=1}^{m} x_j^2 \right) - (\bar{x})^2$$

<u>OSSERVAZIONE</u> SE I DATI SONO DIMENSIONALI (PER ESEMPIO FORNITI IN METRI, Kg ECC) ALLORA MEDIA E DEVIAZIONE STANDARD HANNO STESSA DIMENSIONE.

<u>VARIANZA PER DATI RAGGRUPPATI.</u> SE I DATI SONO STATI GIÀ SUDDIVISI IN CLASSI E SI CONOSCE SOLO LA FREQUENZA, IL CALCOLO ESATTO NON È POSSIBILE VISTO CHE ABBIAMO PERSO PARTE DELL'INFORMAZIONE. CALCOLIAMO ALLORA LA VARIANZA APPROSSIMANDO I DATI IN QUESTO MODO: SOSTITUIAMO AI DATI NELLA CLASSE K-ESIMA (CHE NON CONOSCIAMO) $f_a(k)$ VOLTE IL VALORE MEDIO DELL'INTERVALLINO CHE INDIVIDUA LA CLASSE, CHIAMIAMOLO $\bar{x}_k$. IN FORMULE

$$\sigma^2 = \frac{1}{m} \sum_{k=1}^{N} f_a(k)(\bar{x}_k - \bar{x})^2 = \frac{1}{m} \left( \sum_{k=1}^{N} f_a(k)\, \bar{x}_k^2 \right) - (\bar{x})^2$$

BADARE CHE È COME AVERE m DATI, QUINDI DIVIDIAMO PER m E NON PER IL NUMERO DI CLASSI N.

ESEMPIO Si consideri la seguente tabella relativa alle frequenze dei pesi in kg di 100 individui adulti

| | Peso $p$ in Kg | $f_a$ |
|---|---|---|
| $\bar{x}_1 = 52{,}5$ | $50 \leq p < 55$ | 20 |
| $\bar{x}_2 = 57{,}5$ | $55 \leq p < 60$ | 15 |
| $\bar{x}_3 = 62{,}5$ | $60 \leq p < 65$ | 18 |
| $\bar{x}_4 = 67{,}5$ | $65 \leq p < 70$ | 22 |
| $\bar{x}_5 = 72{,}5$ | $70 \leq p < 75$ | 18 |
| $\bar{x}_6 = 77{,}5$ | $75 \leq p \leq 80$ | 7 |

$M = 100$

$N = 6$

Abbiamo

$$\bar{p} = \frac{1}{100}\left(20 \cdot 52{,}5 + 15 \cdot 57{,}5 + 18 \cdot 62{,}5 + 22 \cdot 67{,}5 + 18 \cdot 72{,}5 + 7 \cdot 77{,}5\right)$$

$$= 63{,}7 \text{ kg} \qquad \text{[peso medio]}$$

$$(\bar{x}_1 - \bar{p})^2 = 125{,}44 \quad (\bar{x}_2 - \bar{p})^2 = 38{,}44 \quad (\bar{x}_3 - \bar{p})^2 = 1{,}44$$

$$(\bar{x}_4 - \bar{p})^2 = 14{,}44 \quad (\bar{x}_5 - \bar{p})^2 = 77{,}44 \quad (\bar{x}_6 - \bar{p})^2 = 190{,}44$$

$$\sigma^2 = \frac{1}{100}\left(20 \cdot 125{,}44 + 15 \cdot 38{,}44 + 18 \cdot 1{,}44 + 22 \cdot 14{,}44 + 18 \cdot 77{,}44 + 7 \cdot 190{,}44\right)$$

$$\simeq 61{,}56 \text{ kg}^2 \qquad \sigma \simeq 7.85 \text{ kg} \qquad \text{[deviazione standard]}$$

TORNIAMO ALLA PROBABILITÀ. DEFINIREMO DUE NUOVE DENSITÀ PER V.A. ASSOLUTAMENTE CONTINUE. SARANNO MOLTO UTILI IN STATISTICA.

SIANO $X_1, \dots, X_m$ $m$ VARIABILI ALEATORIE INDIPENDENTI E TUTTE CON DENSITÀ GAUSSIANA STANDARD, CIOÈ $X_k \sim \mathcal{N}(0,1)$ $k=1,\dots,m$. PONIAMO $Y = \sum_{k=1}^{m} X_k^2$. LA DENSITÀ DI $Y$ VIENE CHIAMATA <u>DENSITÀ DEL $\chi$ QUADRATO AD $m$ GRADI DI LIBERTÀ</u> ED È INDICATA CON IL SIMBOLO $\chi^2(m)$.

ABBIAMO VISTO IN UNA DELLE LEZIONE SCORSE CHE SE $X \sim \mathcal{N}(0,1)$ ALLORA $Y = X^2$ HA DENSITÀ

$$g(t) = \begin{cases} \dfrac{1}{\sqrt{2\pi t}} e^{-t/2} & \text{PER } t>0 \\ 0 & \text{PER } t \leq 0 \end{cases}$$

CIOÈ $g = \Gamma(\frac{1}{2}, \frac{1}{2})$. QUESTO SIGNIFICA CHE $\chi^2(1) = \Gamma(\frac{1}{2}, \frac{1}{2})$.

ABBIAMO ANCHE VISTO CHE SOMMANDO V.A. INDIPENDENTI DI TIPO $\Gamma$ SI OTTIENE ANCORA UNA V.A. DI TIPO . PIÙ PRECISAMENTE, SE $Y_k \sim \Gamma(\alpha_k, \lambda)$, ALLORA $\sum_{k=1}^{m} Y_k \sim \Gamma\left(\sum_{k=1}^{m} \alpha_k, \lambda\right)$

NE DEDUCIAMO QUINDI CHE

$$\boxed{\chi^2(m) = \Gamma\left(\frac{m}{2}, \frac{1}{2}\right)}$$

IN PARTICOLARE, PER QUANTO VISTO SULLE V.A. DI TIPO $\Gamma$

- SE $Y \sim \chi^2(m)$ ALLORA $E[Y] = m$ E $\mathrm{Var}\, Y = 2m$ (IN QUANTO SE $Y \sim \Gamma(\alpha, \lambda)$ ALLORA $E[Y] = \alpha/\lambda$ E $\mathrm{Var}\, Y = \alpha/\lambda^2$).

IL CALCOLO DI $E[Y]$ PUÒ ESSERE FATTO ANCHE DIRETTAMENTE A MANO: TENUTO CONTO CHE $E[X_k] = 0$, ABBIAMO $E[X_k^2] = \mathrm{Var}\, X_k = 1$, DA CUI $E[Y] = \sum_{k=1}^{m} E[X_k^2] = m$.

- SE $Y_1 \sim \chi^2(m_1)$ E $Y_2 \sim \chi^2(m_2)$, E SONO INDIPENDENTI, ALLORA

$$Y_1 + Y_2 \sim \chi^2(m_1 + m_2)$$

QUESTO È DOVUTO AL RISULTATO SULLA SOMMA DI V.A. DI TIPO $\Gamma$.

PER $m$ GRANDE (DICIAMO $m > 30$) POSSIAMO APPROSSIMARE UNA V.A. DI TIPO $\chi^2(m)$ CON UNA DI TIPO GAUSSIANO. INFATTI, SE $\{X_k\}_{k \in \mathbb{N}}$ È UNA SUCCESSIONE DI V.A. INDIPENDENTI DI TIPO $\chi^2(1)$, ALLORA $Y_m = \sum_{k=1}^{m} X_k \sim \chi^2(m)$ E PER IL TEOREMA DEL LIMITE CENTRALE

$$Y_m \simeq Y \quad \text{CON} \quad Y \sim N(m, 2m)$$

(LE $X_k$ HANNO MEDIA 1 E VARIANZA 2), CIOÈ IL COMPORTAMENTO DELLA V.A. $Y_m$ CON DENSITÀ $\chi^2(m)$ È ASINTOTICAMENTE QUELLO DI UNA V.A. $Y$ CON DENSITÀ GAUSSIANA

$\mathcal{N}(m, 2m)$. IN PARTICOLARE

$$P\{Y_m \leq x\} \simeq \Phi\left(\frac{x-m}{\sqrt{2m}}\right) \qquad \forall x \in \mathbb{R}$$

COME AL SOLITO $\Phi$ INDICA LA FUNZIONE DI RIPARTIZIONE ASSOCIATA ALLA GAUSSIANA.

DATA UNA SUCCESSIONE $\{X_k\}_{k \in \mathbb{N}}$ DI VARIABILI ALEATORIE SU UN CERTO SPAZIO DI PROBABILITÀ $(\Omega, \mathcal{A}, P)$, CHIAMIAMO <u>VARIANZA CAMPIONARIA</u> LA V.A.

$$S_m^2 := \frac{1}{m-1} \sum_{k=1}^{m} (X_k - \bar{X}_m)^2 \qquad \left[\begin{array}{l}\text{BEN DEFINITA} \\ \text{SE } m \geq 2\end{array}\right]$$

AL SOLITO $\bar{X}_m = \frac{1}{m} \sum_{k=1}^{m} X_k$ INDICA LA MEDIA CAMPIONARIA.

<u>PROPOSIZIONE</u> SUPPONIAMO CHE LE $X_k$ SIANO INDIPENDENTI E CON DISTRIBUZIONE GAUSSIANA $\mathcal{N}(\mu, \sigma^2)$. ALLORA

(i) $\displaystyle\sum_{k=1}^{m} \left(\frac{X_k - \mu}{\sigma}\right)^2 \sim \chi^2(m)$

(ii) $\displaystyle\sum_{k=1}^{m} \left(\frac{X_k - \bar{X}_m}{\sigma}\right)^2 = \frac{(m-1) S_m^2}{\sigma^2} \sim \chi^2(m-1)$

(iii) LE V.A. $S_m^2$ E $\bar{X}_m$ SONO TRA LORO INDIPENDENTI

PROOF PARZIALE

IL PUNTO (i) SEGUE DAL FATTO CHE $\frac{X_n - \mu}{\sigma} \sim \mathcal{N}(0,1)$ E DALLA DEFINIZIONE DI $\chi^2(m)$.

OMETTIAMO LA DIMOSTRAZIONE DEGLI ALTRI DUE PUNTI, IN QUANTO DELICATA. OSSERVIAMO CHE (ii) DIFFERISCE DA (i) IN QUANTO AL POSTO DELLA MEDIA $\mu$ PRESENTA LA MEDIA CAMPIONARIA. IL COMPORTAMENTO È DIVERSO PERCHÉ LE $m$ V.A. $X_k - \bar{X}_m$ NON SONO INDIPENDENTI: LA LORO SOMMA È NULLA E QUINDI SONO RELAZIONATE.

<u>ESEMPIO</u> UNA DITTA CONFEZIONA KIWI. DALLO STORICO È NOTO CHE LA VARIANZA DELLE DIMENSIONI DEL FRUTTO È 1,26 $cm^2$. DOVENDO FORNIRE FRUTTI CON SIMILI DIMENSIONI, LA DITTA SCARTA UNA PARTITA DI FRUTTI SE LA VARIANZA CAMPIONARIA DI 40 PEZZI SCELTI A CASO SUPERA 2. ASSUMENDO CHE LA DIMENSIONE SEGUA UNA LEGGE NORMALE CON VARIANZA $\sigma^2 = 1,26$, CHE PROBABILITÀ C'È CHE LA PARTITA VENGA SCARTATA?

INDICHIAMO CON $X_k$, $k=1,\dots,40$, LE DIMENSIONI DEI PEZZI

SCELTI. STIAMO ASSUMENDO $X_n \sim \mathcal{N}(\mu, \sigma^2)$. SAPPIAMO CHE LA VARIANZA CAMPIONARIA $S_M^2$ SODDISFA

$$\frac{39\, S_M^2}{1{,}26} = \frac{(M-1)\, S_M^2}{\sigma^2} \sim \chi^2(M-1) = \chi^2(39)$$

QUINDI

$$P\{S_M^2 > 2\} = P\left\{\frac{39\, S_M^2}{1{,}26} > 61{,}9\right\} = 1 - P\left\{\frac{39 S_M^2}{1{,}26} \leq 61{,}9\right\} \simeq 0{,}01123$$

CON L'ULTIMO STEP OTTENUTO PER COMPUTAZIONE NUMERICA DELLA FUNZIONE DI RIPARTIZIONE

$$F(x) = \int_{-\infty}^{x} \chi^2(39)\, dt \quad \text{IN } x = 61{,}9$$

VEDIAMO UN'ULTERIORE TIPO DI V.A. ANCHE QUESTA, COME LA $\chi^2(M)$, TORNA UTILE IN STATISTICA.

SIANO $X \sim \mathcal{N}(0,1)$ E $Y \sim \chi^2(M)$ DUE VARIABILI ALEATORIE INDIPENDENTI. PONIAMO $T = \sqrt{M}\, \frac{X}{\sqrt{Y}}$. OSSERVIAMO CHE ESSENDO $\chi^2(M)$ NULLA SULLA SEMIRETTA $(-\infty, 0]$, ABBIAMO $P\{Y \leq 0\} = 0$ E QUINDI L'ESPRESSIONE PER $T$ È BEN POSTA A MENO DI INSIEMI DI PROBABILITÀ NULLA

LA DENSITÀ DI $T$ VIENE CHIAMATA DENSITÀ DI <u>STUDENT AD M GRADI DI LIBERTÀ</u> ED INDICATA CON IL SIMBOLO $t(M)$

È possibile scrivere esplicitamente la densità $t(n)$:

$$t(n)(s) = c_n \left(1 + \frac{s^2}{n}\right)^{-\frac{(n+1)}{2}} \qquad s \in \mathbb{R}$$

con l'espressione della costante $c_n$ che tralasciamo, essendo complicata e non rilevante per i nostri scopi.
Comunque $\lim_{n \to \infty} \left(1 + \frac{s^2}{n}\right)^{-\frac{(n+1)}{2}} = e^{-s^2/2}$ (per noto limite notevole), mentre $\lim_{n \to \infty} c_n = 1/\sqrt{2\pi}$. Abbiamo quindi il seguente risultato

<u>Proposizione</u> Per $n \to +\infty$, la densità $t(n)$ converge alla densità $\mathcal{N}(0,1)$

<u>Proposizione</u> Supponiamo di avere $n$ variabili aleatorie $X_1, \dots, X_n$ indipendenti e con densità gaussiana $\mathcal{N}(\mu, \sigma^2)$. Allora

$$\sqrt{n}\, \frac{\bar{X}_n - \mu}{\sqrt{S_n^2}} \sim t(n-1)$$

proof

Sappiamo che $\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$, quindi $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$.

Inoltre, per quanto visto sulla varianza campionaria $\frac{(n-1) S_n^2}{\sigma^2} \sim \chi^2(n-1)$, mentre $S_n^2$ ed $\bar{X}_n$ sono indipendenti

Quindi

$$\sqrt{n}\,\frac{\bar{X}_n-\mu}{\sqrt{S_n^2}} = \sqrt{n-1}\;\frac{\dfrac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n-1)S_n^2}{\sigma^2}}} \sim t(n-1)$$

$\blacksquare$

Nota storica

Sealy Gosset introdusse la densità di Student ai primi del 900. Il nome è lo pseudonimo usato da Gosset quando pubblicò l'articolo.