Concorso ISTAT 100 Collaboratori tecnici

MANUALE COMPLETO + QUIZ

PER LE PROVE SCRITTE

NLD CONCORSI

Capitolo 1 STATISTICA E PROBABILITÀ

SOMMARIO:
1. Elementi di statistica descrittiva univariata. - **1.1.** Frequenze, classi, distribuzioni di frequenza e relative rappresentazioni grafiche. - **1.2.** Indici di posizione. - **1.3**. Indici di variabilità. - **1.4.** Indici di forma. - **1.5**. Indici di concentrazione. - **2.** Elementi di statistica descrittiva multivariata. - **2.1.** Concetto di distribuzione di frequenza multivariata. - **2.2.** Indice di dipendenza in media. - **2.3.** Indice di associazione. - **2.4**. Indice di correlazione. - **3.** Teoria dei numeri indici. - **4.** Fondamenti del calcolo delle probabilità. - **5.** Variabili casuali univariate e multivariate. - **6.** Variabili casuali dipendenti e indipendenti. - **7.** Principali distribuzioni di variabili casuali discrete. - **7.1.** Uniforme discreta. - **7.2.** Bernoulliana. - **7.3.** Binomiale. - **7.4.** Poissoniana. - **8.** Principali distribuzioni di variabili casuali continue. - **8.1.** Uniforme continua. - **8.2.** Normale o Gaussiana. - **8.3.** Normale standardizzata. - **8.4.** t di Student. - **8.5.** Chi-quadrato. - **8.6.** F di Fisher. - **8.7.** Gamma. - **8.8.** Beta. - **8.9.** Weibull. - **8.10.** Esponenziale negativa. - **9.** Teoremi limite del calcolo delle probabilità. - **10.** Teoria dell'inferenza statistica. - **10.1.** Stimatori e relative proprietà. - **10.2.** Stima intervallare e relativi metodi. - **11.** Test di verifica di ipotesi- P-value- Errori di I e II tipo e potenza del test. - **11.1.** Principali test parametrici. - **11.2.** Principali test non parametrici.

1. Elementi di statistica descrittiva univariata.

La Statistica descrittiva può essere suddivisa in univariata, ovvero essa descrive un fenomeno statistico osservando un solo carattere di una popolazione mentre quella multivariata osserva e descrive due o più caratteri che assumono le relative modalità. La prima studia le distribuzioni di frequenza e le relative rappresentazioni grafiche, gli indici di posizione, che possono essere centrali e non, di variabilità, di forma e di concentrazione.

► 1.1. Frequenze, classi, distribuzioni di frequenza e relative rappresentazioni grafiche.

Si definisce distribuzione di frequenza unitaria semplice di un carattere l'insieme delle modalità o realizzazioni osservate, per ogni unità, nella popolazione di interesse. Laddove l'insieme delle modalità è riferito a più caratteri si è in presenza di distribuzione di frequenza unitaria multipla. Esempio: Si supponga di aver osservato 10 famiglie (unità statistiche) e di esse il carattere "numero dei componenti". Se questi dati vengono inseriti in una tabella composta di righe e di colonne si è in presenza di una matrice di dati. Sulle righe si dispongono le unità statistiche e sulle colonne il/i carattere/i. La *frequenza assoluta* esplicita il numero di volte che la modalità o realizzazione di un carattere si presenta nella popolazione di interesse. La *frequenza relativa* si trova facendo il rapporto tra frequenza assoluta e numero unità statistiche della popolazione N. La *frequenza relativa percentuale* è data dalla frequenza relativa moltiplicata per 100. La *frequenza cumulata assoluta* si calcola, a partire dal primo valore di frequenza assoluta, relativa ad ogni osservazione x_i , sommando progressivamente i successivi valori fino ad ottenere il loro totale. La *frequenza cumulata relativa* si ottiene dividendo la frequenza assoluta cumulata relativa ad ogni osservazione

x_i con il totale delle osservazioni N. La *frequenza cumulata relativa percentuale* è data dalla frequenza cumulata relativa moltiplicata per 100. La *frequenza retrocumulata assoluta* si calcola, a partire dall'ultimo valore di frequenza assoluta, per ogni osservazione x_i , sottraendo progressivamente i successivi valori fino ad ottenere il loro totale. La *frequenza retrocumulata relativa* si ottiene dividendo la frequenza assoluta retrocumulata relativa ad ogni osservazione x_i con il totale delle osservazioni N. La *frequenza retrocumulata relativa percentuale* è data dalla frequenza retrocumulata relativa moltiplicata per 100. Le modalità del carattere X possono essere rappresentate da valori singoli o discreti oppure suddivisi in classi. E' opportuno, quindi, esplicitare i concetti più importanti riferiti alle classi. La classe è un sub-intervallo del campo di variazione del carattere di interesse definito. L'ampiezza delle classi dipende dalle caratteristiche del fenomeno osservato e dal grado di significatività che assumono per l'analisi descrittiva e inferenziale del fenomeno stesso. La generica classe è definita come:

$$(a_{i-1}, a_i) \qquad \text{per } i = 1,2,\ldots., k \tag{1.1.1}$$

$$a_{i-1} < a \leq a_i \tag{1.1.2}$$

all'interno della quale vanno inserite tutte le modalità o realizzazioni del carattere comprese nell'intervallo reale.
Le Classi hanno un valore minimo coincidente con l'estremo sinistro, un valore massimo coincidente con l'estremo destro e possono essere Disgiunte ovvero senza sovrapposizioni, Esaustive ovvero contenente un valore minimo e massimo, Chiuse ovvero che non sono ricompresi né il valore estremo destro né quello sinistro, Chiuse a destra ovvero il valore estremo destro della classe non è ricompreso, Chiuse a sinistra ovvero il valore estremo sinistro della classe non è ricompreso, Equi-ampie ovvero aventi tutte la stessa ampiezza, Equi-frequenti ovvero aventi tutte la stessa frequenza. Nella Tabella 1.1.1 si riporta un esempio di distribuzione di frequenza per valori singoli o discreti relativa al carattere altezza di una popolazione di 100 individui che assume le modalità 155, 165, 175, 185, 195 con frequenza assoluta rispettivamente pari a 12,21,47,15,5.

Tabella 1.1.1 Distribuzione di frequenza per valori singoli o discreti

x_i	n_i	f_i	f_i*100	N_i	F_i	F_i*100	r_i	R_i	R_i*100
155	12	0,12	12%	12	0,12	12%	100	1,00	100%
165	21	0,21	21%	33	0,33	33%	88	0,88	88%
175	47	0,47	47%	80	0,80	80%	67	0,67	67%
185	15	0,15	15%	95	0,95	95%	20	0,20	20%
195	5	0,05	5%	100	1,00	100%	5	0,05	5%
Totale	100	1,00	100%				0	0	

dove: $\mathbf{x_i}$ sono i valori discreti osservati; $\mathbf{n_i}$ le frequenze assolute; $\mathbf{f_i}$ le frequenze relative; $\mathbf{f_i*100}$ le frequenze relative percentuali; $\mathbf{N_i}$ le frequenze cumulate assolute; $\mathbf{F_i}$ le frequenze cumulate relative; $\mathbf{F_i*100}$ le frequenze cumulate relative percentuali; $\mathbf{r_i}$ le frequenze retrocumulate assolute; $\mathbf{R_i}$ le frequenze retro cumulate relative; $\mathbf{R_i*100}$ frequenze retro cumulate relative percentuali.
Nella Tabella 1.1.2 si riporta un esempio di distribuzione di frequenza per valori suddivisi in classi equi ampie: 150-160;160-170;170-180;180-190 relativa al carattere altezza di una popolazione di 100 individui le con frequenza assoluta rispettivamente pari a 12,21,47,15,5.

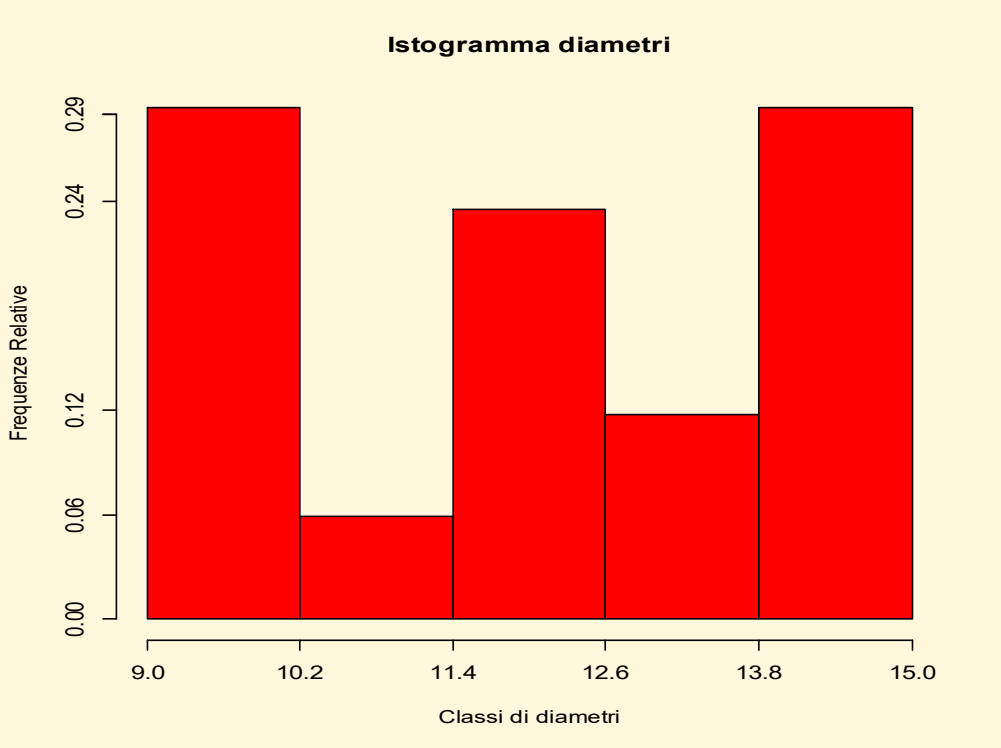
Tabella 1.1.2 Distribuzione di frequenza per valori suddivisi in classi

Classi	$\mathbf{x^{(2)}_i}$	$\mathbf{n_i}$	$\mathbf{f_i}$	$\mathbf{f_i*100}$	$\mathbf{N_i}$	$\mathbf{F_i}$	$\mathbf{F_i*100}$	$\mathbf{r_i}$	$\mathbf{R_i}$	$\mathbf{R_i*100}$
150-160	155	12	0,12	12%	12	0,12	12%	100	1,00	100%
160-170	165	21	0,21	21%	33	0,33	33%	88	0,88	88%
170-180	175	47	0,47	47%	80	0,80	80%	67	0,67	67%
180-190	185	15	0,15	15%	95	0,95	95%	20	0,20	20%
190-200	195	5	0,05	5%	100	1,00	100%	5	0,05	5%
Totale		100	1,00	100%				0	0	

La rappresentazione grafica che descrive meglio una distribuzione di frequenza, in modo particolare per valori suddivisi in classi, è quello a barre verticali contigue o non contigue o istogramma. Data i seguenti valori osservati del carattere x: 9,9.5,9.7,10,10.1,11,11.9,12,12.3,12.6,13,13.8,14,14,14.1,14.8,15 in un popolazione di 17 unità statistiche costruire la distribuzione di valori suddivisi in 5 classi di ampiezza 1,2 e rappresentarle graficamente. Le classi sono: 9,0-10,2;10,2-11,4;11,4-12,6;12,6-13.8;13,8-15.0

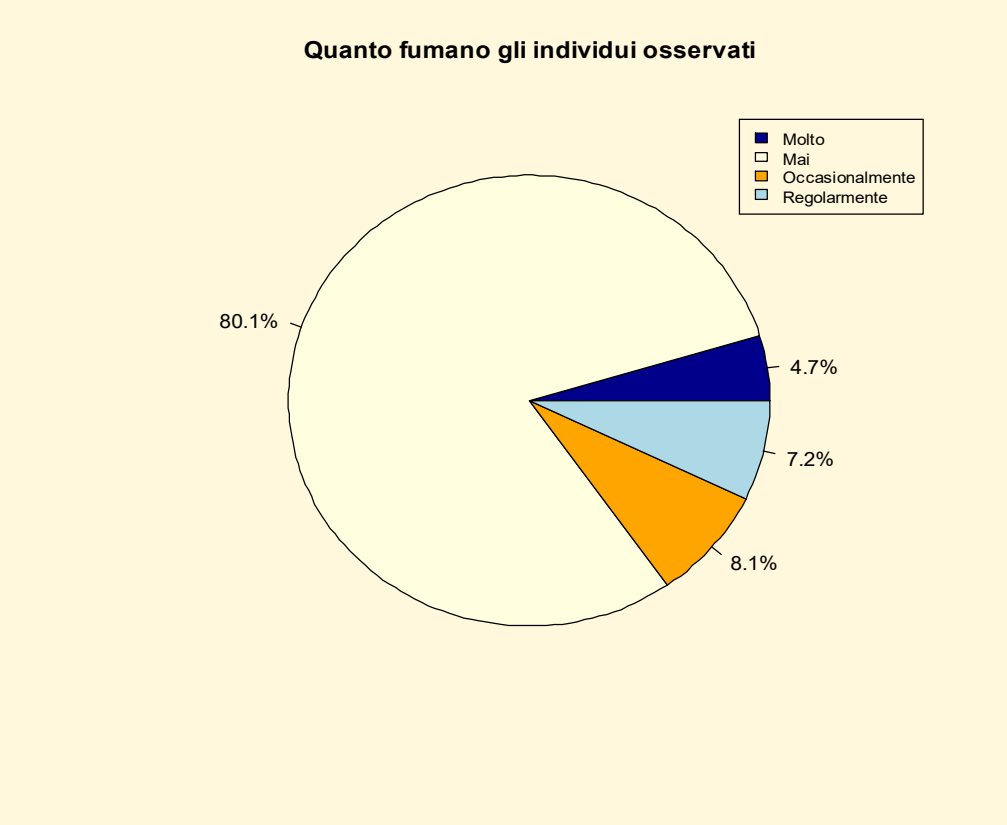
Nel Grafico 1.1.1 viene riportato il relativo istogramma:

Grafico 1.1.1 Istogramma dei diametri



La rappresentazione grafica che descrive meglio una distribuzione di frequenza territoriale o spaziale, settoriale e per prodotto è quella a torta. Nel Grafico 1.1.2 viene riportata la distribuzione del fumo:

Grafico 1.1.2 Distribuzione del fumo



Altri grafici vengono utilizzati per rappresentare distribuzioni di frequenze quali quelli *a nastri o barre orizzontali, ad anello, a bolle, radar.*

► 1.2. Indici di posizione.

Gli *indici di posizione* si suddividono in centrali quali: la *media aritmetica*, *geometrica* ed *armonica*, la *mediana* e la *moda*; la media aritmetica e geometrica, la mediana e la moda, a loro volta, possono riguardare distribuzioni di valori discreti o suddivisi in classi e non centrali quali il I e il III Quartile, il minimo ed il massimo.

La media aritmetica semplice o per valori discreti è la più comune ed importante misura di tendenza centrale di una distribuzione di un fenomeno statistico. E' un concetto che ogni individuo incontra frequentemente nella propria vita quotidiana. La relazione matematica che la definisce, secondo il Chisini, è data da

$$M_a = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (1.2.1)$$

dove x_i sono le osservazioni ed n il loro numero.

Se le osservazioni sono raggruppate in classi, come nel caso dei caratteri continui, le x_i sono rappresentate dal valore centrale di ogni classe, ovvero dalla semisomma dell'estremo inferiore e superiore della classe. Tali valori sono corretti soltanto nell'ipotesi che sia uniforme la ripartizione

delle frequenze all'interno di ciascuna classe. A questo proposito va opportunamente spiegato che il concetto di valore centrale di classe rappresenta una utile soluzione per lavorare con un valore singolo quando si è in presenza di un intervallo di valori. Ma, al contempo, questa implica un certo livello di approssimazione poiché si discosta, in una certa misura, dal valore medio delle osservazioni ricomprese nell'intervallo considerato. La notazione che esprime la media aritmetica in frequenza assoluta è la seguente:

$$M_a = \frac{\sum_{i=1}^{h} x_i n_i}{\sum_{i=1}^{h} n_i} \qquad (1.2.2)$$

mentre quella in frequenza relativa è data da:

$$M_a = \sum_{i=1}^{h} x_i f_i \qquad (1.2.3)$$

Il concetto di *media geometrica* si applica prevalentemente a valori positivi relativi a quantità in rapporto tra loro. Essa è usata spesso in campo finanziario per calcolare, ad esempio, il capitale a scadenza di un prestito o operazioni similari. Considerato un insieme *n* di valori positivi y_1, $y_2 y_n$ di un carattere *Y*, la notazione della media geometrica per valori discreti è data dalla radice ennesima dei prodotti tra i valori stessi:

$$M_{g(y)} = \sqrt[n]{y_1 . y_2 \ldots\ldots\ldots y_n} \qquad (1.2.4)$$

Se si ha una *distribuzione di frequenza per valori suddivisi in classi* il calcolo della media geometrica, così come si è visto per quella aritmetica, può essere ottenuto applicando due procedimenti: *il primo utilizza la frequenza assoluta* n_i e la notazione è la seguente:

$$M_{g(y)} = \sqrt[\sum n_k]{y_1^{n_1} \quad . y_2^{n_2} \ldots\ldots\ldots y_k^{n_k}} \qquad (1.2.5)$$

dove y_k è il numero di osservazioni o modalità del carattere e n_1, n_2,…., n_k sono le relative frequenze assolute e $\sum n_k$ la sommatoria delle stesse.

Il secondo prende in considerazione *i logaritmi* ed è espresso dalla seguente notazione:

$$Mg(y) = \frac{1}{\sum_{i=1}^{n} n_k} * \sum_{i=1}^{n} n_k \quad * \ln x_i \qquad (1.2.6)$$