UNIVERSITÀ DEGLI STUDI DI UDINE
Francesco Davini

Formulario di Statistica

- Formulario di Statistica
  - Statistica descrittiva
    - I dati
      - unità statistiche e popolazione
      - Tipi di Statistica
      - Statistica descrittiva
      - Analisi esplorativa
      - Variabili statistiche
      - Distribuzioni di frequenza
        - Frequenza assoluta
        - Frequenza relativa
        - Frequenze cumulative
        - Altri tipi di distribuzioni di frequenza
    - Rappresentazioni grafiche
    - Diagrammi circolari
    - Diagrammi a Barre
    - Diagrammi a bastoncini
  - Calcolo delle probabilità
  - Inferenza statistica
    - Utilizzo
    - Campionamento
    - Modelli statistici paramentrici
    - Verifica del modello
    - Procedure inferenziali
    - Statistiche campionarie
    - Somma e media campionaria
    - Varianza campionaria
    - Stima puntuale - Stime
  - Formule e Esempi
    - Frequenza relativa
    - Frequenza cumulata
    - Somma e media campionaria
  - Comandi in R

Statistica descrittiva

I dati

unità statistiche e popolazione

I dati rappresentano informazioni di una **popolazione**, ovvero l'intera collezione di unità statistiche sulle quali si cerca l'informazione.

Esistono due tipi di popolazioni:

- **popolazione reale**: unità che hanno una esistenza fisica, sono popolazioni effettive e quindi finite, possono essere osservate in modo completo (**censimento**) o parziale (**campionamento**)
- **popolazione virtuale**: hanno un'esistenza concettuale e sono derivate dalla potenziale replicabilità a piacere della sperimentazione; sono potenzialmente *infinite* e quindi esaminabili solo in modo parziale

censimento: si esaminano *tutte* le unità di una popolazione reale, con riferimento a determinate caratteristiche di interesse

Per popolazioni reali i censimenti sono raramente effettuati, si opta più spesso per un campione.

campionamento: si esamina un *sottoinsieme* finito di unità statistiche, appartenenti ad una *popolazione reale* o *virtuale*, selezionate mediante l'esperimento di campionamento.

L'**esperimento di campionamento** è un particolare esperimento, assimilabile all'estrazione casuale di alcuni elementi da un'urna.

È un **esperimento casuale (aleatorio)** dal momento che risultano possibili una pluralità di esiti (campioni osservati) e prima di effettuare il campionamento non è possibile individuare con certezza quale potenziale campione verrà selezionato (**variabilità campionaria**).

Affinché il campione porti informazioni sull'intera popolazione, la sua estrazione deve essere casuale.

- il campione va scelto in modo che rifletta le **caratteristiche della popolazione**
- Esistono vari **piani di campionamento**, il più semplice è il **campionamento casuale semplice**, assimilabile all'estrazione casuale con reinserimento di elementi da un'urna.

Tipi di Statistica

I metodi statistici si possono dividere in due grandi classi.

- **Statistica descrittiva**: metodi per la descrizione, la presentazione e la sintesi dei dati disponibili, al fine di individuarne la struttura essenziale. Le finalità sono principalmente di tipo descrittivo, poichè si sintetizzano le informazioni disponibili, che riguardano la totalità della popolazione. anche quando i dati disponibili rappresentano un campione estratto da una popolazione, nella statistica descrittiva non se ne tiene conto.
- **Statistica inferenziale**: sono metodo per ricavare dai dati campionari informazioni sulla popolazione di riferimento e per quantificare la fiducia da accordare a tali informazioni. Si utilizza metodi del **calcolo delle probabilità**

Statistica descrittiva

Gli elemento più rilevanti:

- Metodi grafici e numerici per descrivere e sintetizzare i dati osservati
- Distinzione fra tecniche di **analisi univariata** (relative ad una singola caratteristica) e tecniche di **analisi multivariata**, ovvero per lo studio congiunto di due o più caratteristiche di interesse.
- alcune nozione di mase sono utili anche per la statistica inferenziale.
- Come premessa ad una analisi inferenziale, è sempre opportuno effettuare uno studio descrittivo con riferimento al particolare campione osservato.

Analisi esplorativa

Un'analisi esplorativa dei dati ha l'obbiettivo di:

- capire come i dati sono stati raccolti e se sono di natura osservazionale o sperimentale
- individuare le unità statistiche, discutere la presenza di dati mancanti ed, eventualmente *ripulire* il dataset
- codificare e riorganizzare i dati nella forma più conveniente per l'analisi
- utilizzare metodo grafici e numerici per ricavare alcune informazioni preliminari sui dati osservati

Ri suppone che i dati siano già stati acquisiti e siano disponibili nella forma di **matrice dei dati** (anche detto **data frame**), questi sono cosiddetti **dati grezzi** ad esempio:

unità	Genere	età	Livello istruzione	Dis
Andrea	M	28	3	5.0
claudio	M	17	2	7.5
Lucia	F	20	3	NA
...	...	...	...	...

con **NA** si intende un dato mancante.

La matrice dei dati fornisce informazioni sulla popolazione in esame con riferimento a:

- Genere: Maschio (M) o Femmina (F) - variabile qualitativa nominale
- Età: età in anni - variabile quantitativa discreta
- Livello di istruzione: 1 = nessuna istruzione, 2 = scuola dell'obbligo, 3 = diploma, 4 = laurea - variabile qualitativa ordinale
- Dis: distanza dal luogo di lavoro in km - variabile quantitativa continua

Ogni **riga** corripsonde ad una unità statistica e contiene i valore su essa relativi delle caratteristiche di interesse.

Ogni **colonna** corrisponde ad una caratteristica di interesse e contiene i valori di tale caratteristica rilevati sulle varie unità statistiche.

Variabili statistiche

Una **variabile** è una caratteristica delle unità statistiche che, al variare dell'unità, può assumere una pluralità di valori.

Le **modalità** di una variabile sono i valori che essa può assumere. Sono, in genere, aggettivi, valori numerici o espressioni verbali.

Le variabili si indicano con le lettere maiuscole, ad esempio Y, mentre una generica modalità si indica on *y*. l'insieme *Y* è l'insieme di tutte le possibili modalità di Y.

Le variabili si possono classificare nel seguente modo:

- **Variabili qualitative** (categoriali): se le modalità sono espresse in forma verbale, in particolare si individuano:
  - **variabili qualitative sconnesse** (normali): non è possibile individuare un ordinamento naturale delle modalità (es. genere, colore degli occhi)
  - **variabili qualitative ordinali**: è possibile invece individuare un ordinamento naturale delle modalità (es. livello di istruzione)

    - **Variabile dicotomica** (binaria): se una variabile qualitativa ha solo due modalità
- **Variabili quantitative** (numeriche): se le modalità sono espresse in forma numerica (diverse dalle codifiche numeriche). in particolare si individuano:
    - **variabili quantitative discrete**: se *Y* è un insieme finito o al più numerabile (es. numero di figli, ètà)
    - **variabili quantitative continue**: se *Y* è un insieme continuo (ad esempio distanza, altezza, reddito). si noti che la continuità ca intensa come una potenziale continuità co come opportuno **riferimento semplificativo**

Una variabile quantitativa può essere con una **scala di intervalli**, se non esiste uno zero naturare e non arbitrario(ad esempio "temperatura", perchè lo 0 è convenzionale e non ha senso dire che 30° è il doppio di 15°). Una variabile quantitativa è con una **scala di rapporti** se invece esiste uno zero con tali caratteristiche (es. peso, altezza, reddito).

Nello studio congiunto di due o più variabili si parla di analisi statistica bivariata o, in generale, multivariata.

La variabile *Y* viene rilevata su una popolazione (campione) costruita da *n* unità e si ottiene una sucessione di modalità osservate $(y_1, \ldots, y_i, \ldots, y_n)$ dove $y_i \ i = 1, \ldots, n$ è il valore assunto da *Y* con riferimento all'unità i-esima.

Si definisce **Variabile statistica** la rilevazione $(y_1, \ldots, y_i, \ldots, y_n)$ di una certa variabile *Y* su una determinata popolazione (campione). E una colonna della matrice dei dati.

Il valore che occupa la posizione i-esima, $y_{(i)}$, si dice avere **rango** *i*

Con **supporto** si intende un insieme delle modalità di Y effettivamente osservate nella popolazione, indicata con S_Y. il numero di elementi all'interno di S_Y sarà sempre minore o uguale a *n*.

Distribuzioni di frequenza

Frequenza assoluta

I dati grezzi (la variabile statistica), pur rappresentando pienamente il contenuto dell'osservazione, usualmente non permettono di cogliere in modo chiaro le caratteristiche del fenomeno in esame.

E utile passare dai dati in forma grazza ad una **tabella di frequenza** che fornisca una sintesi dei dati in un formato facile da capire.

un esempio:

Genere	Frequenza
M	5
F	7
Totale	12

Una tabella di frequenza riferita ad una **variabile statistica qualitativa** è detta **serie statistica**

Se la **variabile statistica** è **quantitativa continua**, si ossservano (a meno di effetti di arrotondamento) tante modalità distinte quante sono le unità statistiche.

è conveniente definire **classi di modalià** contigue e contare le unità che appartengono a ciascuna classe.

Le classi vanno definite in modo che: non siano ne troppe ne troppo poche. Una regola indicativa è di utilizzare circa $\sqrt{n}$ classi.

Una **tabella (distribuzione) di frequenza** con moalità raggruppate in classe:

Classi	$y_0 \dashv y_1$	$\dots$	$y_{i-1} \dashv y_i$	$\dots$	$y_{J-1} \dashv y_J$	**Totale**
Freq	f_1	$\dots$	f_i	$\dots$	f_J	$\sum_{i=1}^{J} f_i$

nota che il simbolo $\dashv$ indica che l'estremo sinistro è incluso nella classe, mentre quello destro no, in notazione matematica standard: $[y_{i-1}, y_i)$.

Frequenza relativa

La **frequenza relativa** di una modalità y_j, o di una classe di modalità $y_{j-1} \dashv y_j$, è la popolazione p_j di unità statistiche portatrici di tale modalità o classe di modalità. corrisponde a:

$$p_j = \frac{f_j}{\sum_{j=1}^{J} f_j} = \frac{f_j}{n}, \quad j = 1, \dots, J$$

Si possono definire anche le **frequenze relative percentuali** definite come $p_j 100, j = 1, \dots, J$

Qualora si abbia un solo dato, la variabile statistica è detta **degenere**

Frequenze cumulative

Quando si hanno **variabili con modalità ordinabili**, può essere utile considerare la frequenza con cui si presentano modalità di ordine inferiore o uguale ad un certo valore.

La **frequenza assoluta cumulata** F_j o, la **frequenza relativa cumulata** P_j definiscono la frequenza assoluta o relativa di modalità o classi di modalità non supeiori alla j-esima $j = 1, \ldots, J$

Si ottengon ocumulando progressivamente le frequenze, più precisaente:

$$F_j = \sum_{i=1}^{j} f_i, \quad P_j = \sum_{i=1}^{j} p_i, \quad j = 1, \ldots, J$$

Altri tipi di distribuzioni di frequenza

SERIE STORICA

Quando si misura un fenomeno nel tempo, si ottiene una distribuzione di frequenza che prende il nome di serie storica (temporale).

Esempio. Si considera il numero di occupati in Italia dal 1997 al 2001.

Anno	No. occupati (in migliaia)
1997	20207
1998	20435
1999	20692
2000	21080
2001	21514

SERIE SPAZIALE

Quando si misura un fenomeno nello spazio, si ottiene una distribuzione di frequenza che prende il nome di serie spaziale (territoriale).

Esempio. Si considera il numero di occupati in Italia nel 2002, suddivisi per ripartizione territoriale.

Ripartizione territoriale	No. occupati (in migliaia)
Nord	11461
Centro	4513
Sud e Isole	6286

Rappresentazioni grafiche

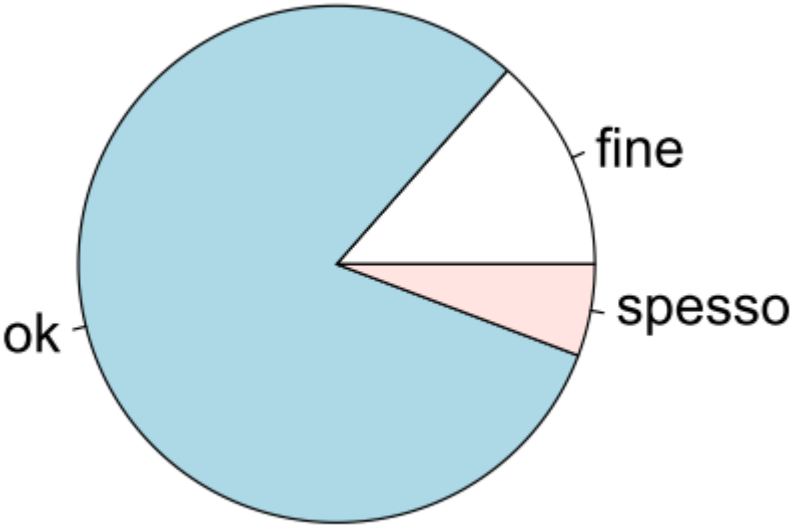
Oltre alle tabelle di frequenza, risulta utile introdurre alcune rappresentazioni grafiche, dette **diagramma statistici** (grafici)

- Per dati categoriali si possono utilizzare, ad esempio:
  - **diagrammi circilari** (o a torta - se ti vede il madda ti scanna)
  - **diagrammi a rettangoli**
  - **diagrammi a rettangoli multipli**
- Per dati numerici si possono utilizzare, ad esempio:
  - **diagrammi a bastoncini**
  - **istogrammi**
  - **poligoni di frequenza**
  - **stima della densità**
  - **funzione di ripartizione empirica**
  - **diagrammi di dispersione**
  - **box-plot**

Diagrammi circolari

I diagrammi circolari (a torta) sono utili per rappresentare serie statistiche sconnesse, riferite a dati qualitativi nominali o eventualmente ordinali (dati categoriali). L'area del settore circolare deve essere proporzionale alla frequenza della modalità corrispondente.

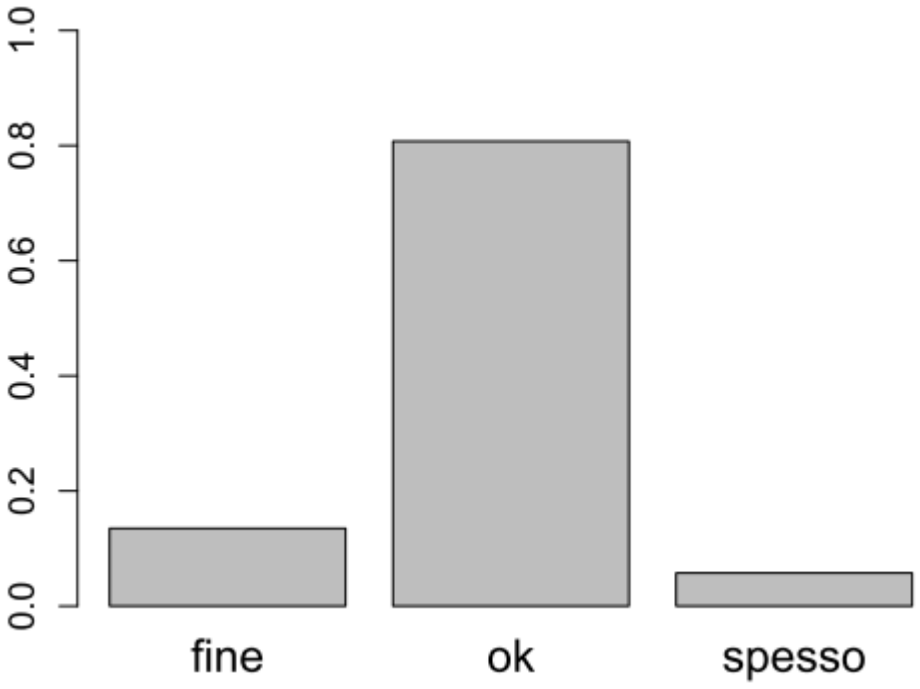
Esempio. Perni (continua). Considerando i dati riferiti alla produzione dei perni, le diverse modalità relative al diametro sono rappresentate dagli spicchi della torta, la cui dimensione Ł proporzionale alla corrispondente frequenza.



Diagrammi a Barre

I diagrammi a rettangoli (a barre) sono utili per rappresentare serie statistiche sconnesse, riferite a dati qualitativi nominali o eventualmente ordinali (dati categoriali). Le altezze dei rettangoli sono proporzionali alle frequenze delle modalità. Le basi hanno la stesa dimensione e sono separate per non implicare alcuna continuità.

Esempio. Perni (continua). Considerando i dati riferiti alla produzione dei perni, si ottiene il seguente diagramma a rettangoli.

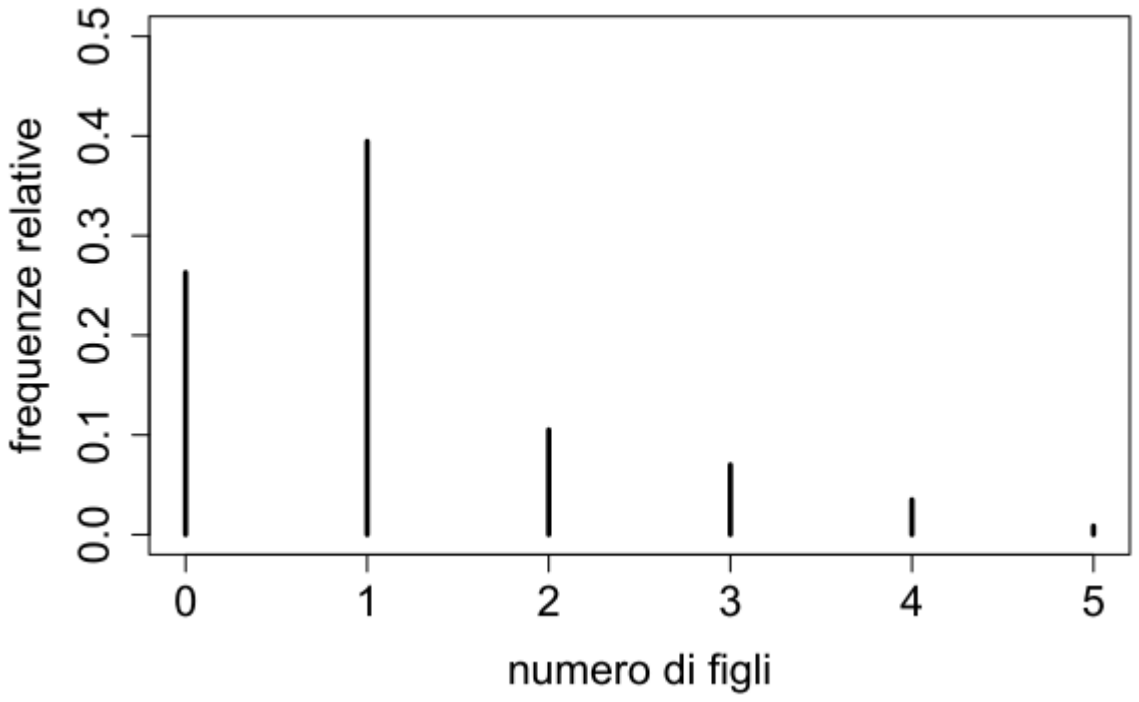


I rettangoli hanno la base uguale e sono separati per non implicare alcuna continuità. Le altezze sono proporzionali alla frequenze delle modalità; si considerano le frequenze relative anché il confronto abbia senso.

Diagrammi a bastoncini

I **diagrammi a bastoncini** servono per rappresentare **distribuzioni di frequenza assoluta o relativa**, riferite a dati qualitativi discreti. L'altezza dei bastoncini è proporzionale o pari alla frequenza, assoluta o relativa, della modalità.

Esempio. Figli. Si considera il numero di figli con riferimento alle famiglie residenti in un determinato territorio. La distribuzione di frequenza relativa Ł rappresentata con il seguente diagramma.



Calcolo delle probabilità

da fare

Inferenza statistica

Utilizzo

Viene utilizzata per studiare una popolazione in maniera parziale.

Viene usata per:

- ricavare dai dati campionari informazioni sulla popolazione di interesse
- Quantificare la fiducia da assegnare a tali informazioni

L'obiettivo è quello di arrivare a conclusioni valide per tutta la popolazione, basandosi su un campione.

Campionamento

Si effettuano indagini campionarie quando:

- presenza di vincoli di tempo e/o problemi di costo
- la popolazione di interesse può essere infinita e virtuale
- la rilevazione potrebbe distruggere le unità statistiche o essere potenzialmente dannosa
- la precisione dei risultati nelle rilevazioni censuarie potrebbe non essere adeguata

È essenziale che i dati campionari possano essere interpretati come risultato di un esperimento aleatorio, perché altrimenti verrebbe meno la rappresentatività del campione e la possibilità di ricavare informazioni utili sulla popolazione (fenomeno) di interesse (inferenza).

Campioni casuali semplici

Sono formati da n realizzazioni indipendenti (con $n \geq 1$) di un esperimento base, nelle medesime condizioni.

Le unità vengono selezionate dalla popolazione di riferimento in modo che ognuna abbia la stessa probabilità di essere scelta (con popolazioni numerose un campionamento con reinserimento o meno non cambia sostanzialmente il risultato).

I dati osservati sono riferiti ad una caratteristica di interesse, rilevata sulle n unità statistiche che costituiscono il campione.

L'**ipotesi fondamentale** su cui poggia l'inferenza statistica è che i dati campionari x sono una realizzazione di un vettore di variabili casuali $X = (X_1, X_2, \ldots, X_n)$

Nell'inferenza statistica c'Ł un rovesciamento di punto di vista. Il processo di generazione dei dati (modello probabilistico) non Ł noto in modo completo. Il processo in questione Ł, in definitiva, la popolazione (fenomeno) oggetto di indagine.

Nel **campionamento casuale semplice**, le variabili casuali $X_1, \ldots, X_n$ sono **indipendenti e identicamente distribuite**, cioè con lo stesso modello probabilistico e tali da non influenzarsi a vicenda.

Modelli statistici paramentrici

Dato un campione casuale semplice $X = (X_1, X_2, \ldots, X_n)$, **la distribuzione di probabilità** delle singole variabili casuali dipende dalla natura dei dati e del fenomeno oggetto di indagine.

La distribuzione assunta per le variabili casuali del campione dipende da costanti ignote dette **parametri**, ad esempio, $p, \mu, \sigma^2, \lambda, \ldots$

Nell'*inferenza statistica parametrica* si assume che la distribuzione delle variabili casuali del campione sia nota a meno dei valori dei parametri, che corrispondono tipicamente agli aspetti di interesse dell'analisi.

Vengono usate le seguenti assunzioni, che definiscono un **modello statistico parametrico** per i dati di un campione casuale semplice:

- le variabili casuali $X_1, X_2, \ldots, X_n$ sono indipendenti
- tutte le X_i hanno la stessa distribuzione di probabilità
- tale distribuzione è nota a meno dei valori di un numero o più parametri, indicati generalmente come $\theta = (\theta_1, \theta_2, \ldots, \theta_k), d \geq 1$

Lo scopo dell'inferenza statistica parametrica è utilizzare i dati osservati $x_1, \ldots, x_n$ per ottenere informazioni su θ, i cui possibili valori appartengono ad un certo insieme Θ detto **spazio parametrico**.

La **scelta del modello** è molto importante, poiché le conclusioni inferenziali dipendono fortemente dalle assunzioni fatte.

Nella specificazione del modello statistico parametrico, è importante considerare:

- la natura dei dati (qualitativi o quantitativi, discreti o continui, ecc.)
- gli aspetti notevoli presenti nei dati come posizione, variabilità, simmetria, curtosi, ecc.
- tutte le informazioni sul meccanismo generatore dei dati.

Esistono anche modelli per **dati dipendenti** e/o **non identicamente distribuiti** (ad esempio per serie storiche e spaziali), e modelli che prescindono dalla forma della distribuzione di probabilità delle variabili casuali del campione (modelli non parametrici)

Lo spazio parametrico è $\Theta = (0, 1)$ e lo spazio campionario è $S_X = 0, 1X \dots X0, 1 = 0, 1^n$, cioè l'insieme di tutti i possibili vettori $n-$dimensionali costituiti da 0 e 1.

Se le n osservazioni sono state effettuate in modo indipendente e nelle medesime condizioni, è ragionevole ipotizzare che $X_i, i = 1, \dots, n$ siano indipendenti con distribuzione $N(\mu, \sigma^2)$.

Si può verificare graficamente l'ipotesi di normalità considerando istogrammi e q-q plot, μ è la misura vera dell'oggetto in esame e σ^2 è riconducibile alla precisione dello strumento di misura.

Verifica del modello

Talvolta in casi complessi si necessita un **controllo empirico del modello**

alcuni strumenti possono essere:

- l'**istogramma delle frequenze relative** e la **stima della densità**
- i grafici dei quantili (**q-q plot**)

L'istogramma e la stima della densità basate sui dati campionari possono essere interpretate, in ambito inferenziale, come stime della funzione di densità

Procedure inferenziali

Si possono individuare *tre classi* generali di procedure che affrontano i problemi inferenziali, con riferimento al parametro di interesse θ:

- la **stima puntuale**: si vuole ottenere, sulla base dell'osservazione *x*, una congettura puntuale su θ
- la **stima intervallare** o **regione di confidenza**: si vuole ottenere, sulla base dell'osservazione *x*, un sottoinsieme (intervallo) di Θ in cui è plausibilmente incluso θ
- **verifica di ipotesi**: data una congettura o un'ipotesi su θ, si vuole verificare, sulla base dell'osservazione *x*, se essa è accettabile (cioè in accordo con i dati *x*).

Statistiche campionarie

La statistica inferenziale è caratterizzata da una componente di incertezza, poiché i dati campionari *x* sono interpretati come realizzazione di un vettore casuale X, ripetendo

l'esperimento si ottengono dati diversi dal primo.

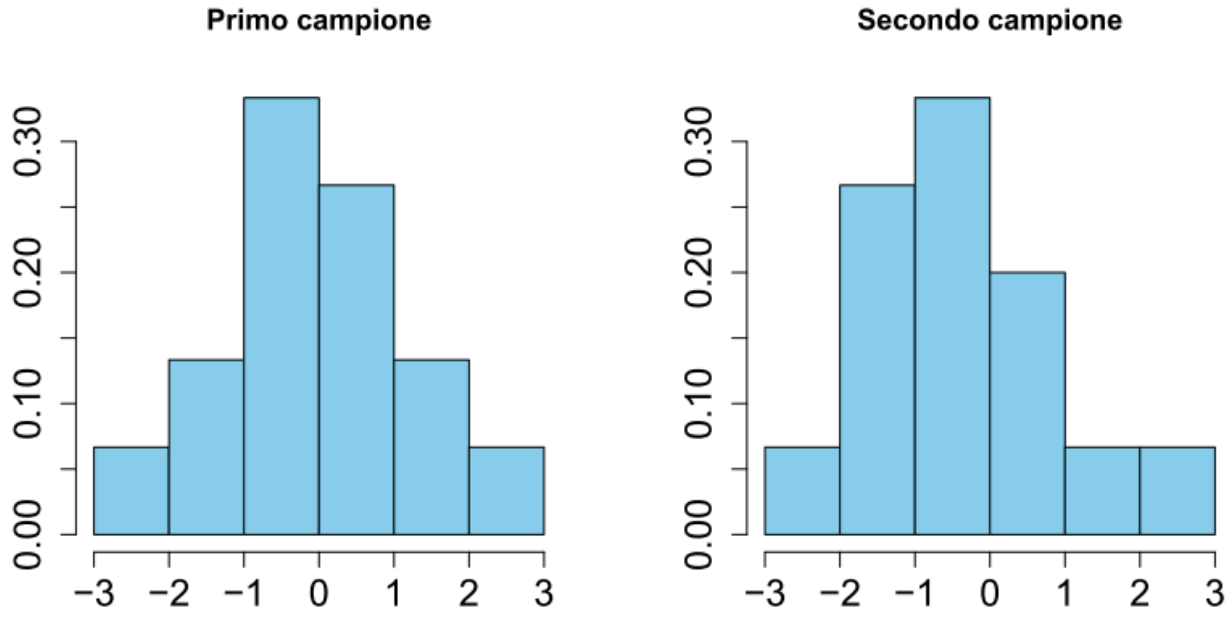
Si chiama **statistica campionaria** ogni trasformata $T = t(X_1, \ldots, X_n)$ che sintetizza opportunamente il campione $X = (X_1, \ldots, X_n)$.

La distribuzione di probabilità T, che è una funzione di $X = (X_1, \ldots, X_n)$, dipende dal parametro incognito θ. Quindi, va intesa **sotto** θ, cioè nell'ipotesi che θ sia il vero valore del parametro, qualunque esso sia.

Dato un campione casuale $X = (X_1, \ldots, X_n)$, sono esempi di statistiche campionarie:

- la **somma campionaria** $S_n = \sum_{i=1}^{n} X_i$, la media campionaria $\overline{X} = n^{-1} S_n$, la **varianza campionaria** $S^2 = n^{-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$
- le **statistiche ordinate** $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$, dove $X_{(1)}$ è la variabile casuale che occupa l'i-esima posizione del campione.
- la **variabile casuale minimo** $X_{(1)}$ = min $X_1, \ldots, X_n$ e la **variabile casuale massimo** $X_{(n)}$ = max $X_1, \ldots, X_n$
- la **mediana campionaria**, definita da $X_{((n+1)/2)}$ se *n* è dispari, e da $(X_{(n/2)} + X_{(n/2+1)})/2$ se *n* è pari
- i **momenti campionari**, centrati e non centrati: $n^{-1} \sum_{i=1}^n (X_i - \overline{X}*n)^r en^{-1}\sum*{i=1}^n X_i^r, con r \in \mathbb{N}^+$

anche se due rilevazioni campionare hanno due risultati diversi, si possono ricondurre al medesimo esperimento osservando gli istogrammi prodotti dai due campioni.



Somma e media campionaria

Sia $X_1, \ldots, X_n$ un campione casuale semplice tratto da una determinata popolazione. Si definiscono, rispettivamente, la **somma campionaria** (somma del campione) e la **media campionaria** (media del campione) le variabili casuali $$S_n = \sum_{i=1}^n X_i, \quad \overline{X}*n* = *frac{1}{n}* *sum*{i=1}^n X_i = \frac{S_n}{n}$$

Poiché $X_1, \ldots, X_n$ sono **indipendenti** e **identicamente distribuite** (quindi anche la stessa media μ e la stessa varianza σ^2), allora:

$$E(S_n) = \sum_{i=1}^{n} E(X_i) = n\mu \quad V(S_n) = \sum_{i=1}^{n} V(X_i) = n\sigma^2$$

$$E(\overline{X}_n) = \frac{E(S_n)}{n} = \mu \quad V(\overline{X}_n) = \frac{V(S_n)}{n^2} = \frac{\sigma^2}{n}$$

Se le variabili casuali $X_1, \ldots, X_n$ sono **gaussiane** $N(\mu, \sigma^2)$, allora anche somma e media campionaria sono variabili casuali gaussiane, più precisamente:

$$S_n \sim N(n\mu, n\sigma^2), \quad \overline{X}_n \sim N(\mu, \sigma^2/n)$$

Valgono, inoltre, i seguenti risultati con riferimento a variabili casuali $X_1, \ldots, X_n$ indipendenti:

- se $X_i \sim Bi(K_i, p), i = 1, \ldots, n$, allora $S_n \sim Bi(\sum_{i=1}^n K_i, p)$
- se $X_i \sim P(\lambda_i), i = 1, \ldots, n$, allora $S_N \sim P(\sum_{i=1}^n \lambda_i)$
- se $X_i \sim X^2(r_i), i = 1 \ldots, n$, allora $S_n \sim X^2(\sum_{i=1}^n r_i)$

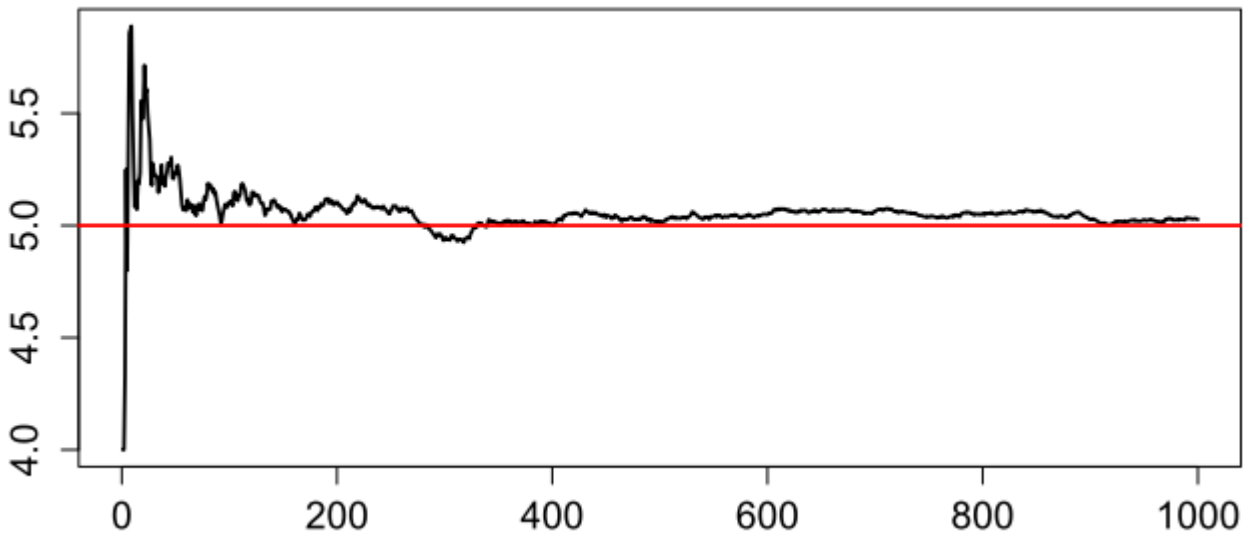
La media campionaria $\overline{X}_n$ è utile in un ambito inferenziale quando si vuole fare inferenza su μ (media della popolazione).
Quindi, la media campionaria $\overline{X}_n$ definisce uno **Stimatore** per μ e il suo valore osservato $\overline{x}_n$ che corrisponde alla media calcolata sul campione, viene utilizzato come **Stima** per μ.

Al crescere di *n*, la variabile casuale media campionaria ha una distribuzione di probabilità sempre più concentrata attorno alla media della popolazione μ.

Formalmente, si afferma che vale la **legge debole dei grandi numeri**, cioè che, nelle condizioni poste in precedenza, se $n \to \infty$:

$$\overline{X}_n \xrightarrow{p} \mu$$

Con la scrittura $\xrightarrow{p}$ si intende la **convergenza in probabilità**, una notazione di convergenza probabilistica illustrata nel seguente esempio



Al crescere di *n*, la distribuzione di probabilità della media campionaria è sempre più concentrata attorno alla media della popolazione μ.

Dice che, sotto condizioni generali, la distribuzione di probabilità della media campionaria tende ad una distribuzione normale al crescere di *n*, indipendentemente dalla distribuzione delle singole variabili casuali del campione

Per le variabili casuali **somma** e **media campionaria** vale un importante risultato: il **teorema limite centrale**.

Data una successione di variabili casuali $X_i, i \geq 1$, indipendenti e identicamente distribuite, con media μ e varianza $\sigma^2 \neq 0$ finite, allora la **Somma standardizzata** e la **media campionaria standardizzata** coincidono e sono tali che, per $n \to \infty$:

$$\frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} \overset{d}{\to} Z \sim N(0,1)$$

La scrittura $\overset{d}{\to}$ indica la **convergenza in distribuzione**: al crescere di *n* la distribuzione di probabilità è sempre più simile a quella di Z.

Per *n* fissato sufficientemente elevato (almeno n > 30), valgono le seguenti utili approssimazioni:

$$\overline{X}_n \dot{\sim} N(\mu, \sigma^2/n), \quad S_n \dot{\sim} N(n\mu, n\sigma^2)$$

dove $\dot{\sim}$ indica la distribuzione approssimata.

Per il teorema limite centrale, se *n* è sufficientemente elevato, si possono ancora utilizzare le **distribuzioni gaussiane** (approssimate), valgono le seguenti approssimazioni:

$$P(a < \overline{X}_n \leq b) \doteq \Phi\left(\frac{b-\mu}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{a-\mu}{\sigma/\sqrt{n}}\right)$$

$$P(a < S_n \leq b) \doteq \Phi\left(\frac{b-n\mu}{\sigma\sqrt{n}}\right) - \Phi\left(\frac{a-n\mu}{\sigma\sqrt{n}}\right)$$

Varianza campionaria

Sia $X_1, \ldots, X_n$ un campione casuale semplice tratto da una determinata popolazione, si definisce **varianza campionaria** la variabile casuale

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

La varianza campionaria può venire calcolata utilizzando la seguente regola di calcolo:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2$$

Poiché $X_i, i = 1, \ldots, n$ sono **indipendenti** e **identicamente distribuite** (quindi anche la stessa media μ e la stessa varianza σ^2), allora:

$$E(S^2) = \frac{n-1}{n}\sigma^2 = \sigma^2 - \frac{1}{n}\sigma^2$$

La **varianza campionaria** S^2 è utile in ambito inferenziale quando, utilizzando i dati campionari, si vuole fare inferenza su σ^2 (varianza della popolazione).

Nel caso si abbia un *n* piccolo si usa la **varianza campionaria corretta**:

$$S_c^2 = \frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{n-1}^{1}(X_i - \overline{X}_n)^2$$

Da cui:

$$E(S_c^2) = \frac{n}{n-1}E(S^2) = \frac{n}{n-1} \cdot \frac{n-1}{n}\sigma^2 = \sigma^2$$

Qualora le variabili casuali $X_1, \ldots, X_n$ siano **gaussiane** $N(\mu, \sigma^2)$, allora la varianza campionaria e la varianza campionaria corretta hanno una distribuzione di probabilità legata al modello X^2:

$$\frac{n}{\sigma^2}S^2 = \frac{n-1}{\sigma^2}S_c^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}{\sigma^2} \sim X^2(n-1)$$

Dove $X^2(n-1)$ indica un modello chi-quadrato di parametro (gradi di libertà) n-1.

La variabile casuale media si può standardizzare con:

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Se al posto di σ si considera $S_c = \sqrt{S_c^2}$ si ha la variabile casuale chiamata **media campionaria studentizzata**, e si ha che:

$$\frac{\overline{X}_n - \mu}{S_c/\sqrt{n}} \sim t(n-1)$$

dove $t(n-1)$ indica una variabile casuale *t* Student con *n* - 1 gradi di libertà.

Date due variabili casuali con distribuzione $N(\mu_X, \sigma_X^2)$ e $N(\mu_Y, \sigma_Y^2)$, le associate varianze campionarie (indipendenti), si può verificare che:

$$\frac{[nS_X^2/\sigma_X^2]/(n-1)}{[nS_Y^2/\sigma_Y^2]/(m-1)} \sim F(n-1, m-1)$$

Dove F(n-1, m-1) indica una variabile casuale F di Fisher con (n-1) e (m-1) gradi di libertà.

Stima puntuale - Stime

Formule e Esempi

Frequenza relativa

Si ottiene dividendo la frequenza assoluta per il numero totale di osservazioni.

$$p_j = \frac{f_j}{\sum_{j=1}^{J} f_j} = \frac{f_j}{n}, \quad j = 1, \ldots, J$$

ad esempio, se in un campione di 100 persone 30 sono di genere femminile, la frequenza relativa del genere femminile è:

$$p_F = \frac{30}{100} = 0.3$$

Frequenza cumulata

$$F_j = \sum_{i=1}^{j} f_i, \quad P_j = \sum_{i=1}^{j} p_i, \quad j = 1, \ldots, J$$

Esempio: Colesterolo (continua). Considerando i dati sul livello di colesterolo sierico

Liv. colesterolo (mg/100 ml)	F_j (età 25-34)	F_j (età 55-64)	P_j (età 25-34)	P_j (età 55-64)
$80 \vdash 120$	13	5	0.012	0.004
$120 \vdash 160$	163	53	0.153	0.043
$160 \vdash 200$	605	318	0.567	0.259
$200 \vdash 240$	904	776	0.847	0.632
$240 \vdash 280$	1019	1057	0.955	0.861
$280 \vdash 320$	1053	1185	0.987	0.965
$320 \vdash 360$	1062	1220	0.995	0.994
$360 \vdash 400$	1067	1227	1	1

Somma e media campionaria

Esempio: Procedura di controllo. Si è verificato un inconveniente su una linea di produzione che determina la presenza di 1/10 di pezzi difettosi. La procedura di controllo della qualità

prevede che, se si individuano almeno 5 pezzi difettosi su n ≥ 1 scelti a caso, il processo viene posto in revisione. Sia S n la somma di n ≥ 1 variabili casuali Ber(1/10) indipendenti. Si cerca il valore per n tale che ci sia una probabilità pari a 0.9 di porre il processo in revisione. Quindi, n ≥ 1 deve essere tale che

$$P(S_n \geq 5) = P\left(\frac{S_n - (n/10)}{\sqrt{n9/100}} \geq \frac{5 - (n/10)}{\sqrt{n9/100}}\right) \doteq P\left(Z \geq \frac{5 - (n/10)}{\sqrt{n9/100}}\right)$$

Sia 0.9 con $Z \sim N(0,1)$. Poiché il valore critico $z_{0.9}$= -1.282, si cerca *n* tale che $[5 - (n/10)]\sqrt{n9/100} \doteq$-1.282, con $n \geq 50$.

Si ottiene come soluzione il valore 85.58, quindi n= 86, può essere una scelta ragionevole.

Comandi in R