Inferenza Statistica

Slides per un corso introduttivo

Leo Pasquazzi

Dipartimento di Statistica e Metodi Quantitativi
Università degli Studi di Milano - Bicocca
email: leo.pasquazzi@unimib.it
ORCID https://orcid.org/0000-0002-2467-2667

25 maggio 2023

Indice I

Indice II

L'**inferenza statistica** è la branca della statistica che si occupa di **problemi di stima** e di **verifiche d'ipotesi**.

Nei problemi di cui si occupa l'inferenza statistica si parte sempre da un **campione** e sulla base delle informazioni contenute nel campione si cerca di

- **stimare** gli **ignoti valori di parametri** che descrivono una **popolazione** di interesse
- e/o di **verificare** se una determinata **ipotesi** circa i valori dei suddetti **parametri** possa essere rifiutata o meno.

Nell'inferenza statistica si può considerare come **popolazione** un qualsiasi insieme finito o anche infinito di **unità statistiche**.

Nel caso di una **popolazione finita** è spesso possibile stilare una lista (un cosiddetto **"frame"**) di tutte le unità statistiche che appartengono alla popolazione.

Le **popolazioni infinite** sono invece sempre definite in modo concettuale specificando dei **criteri di appartenenza** (la popolazione di tutti i nascituri in una determinata regione e in un determinato periodo di tempo, la popolazione di tutte le confezioni di pasta prodotte da un determinato macchinario, la popolazione di tutti gli utenti di internet che visitano una determinata pagina web, ecc.). Per questo motivo le popolazioni infinite vengono a volte chiamate **popolazioni concettuali**.

I **parametri** che descrivono una popolazione possono essere di due tipi:

- **Indici statistici** come per esempio una **media**, **varianza** o **proporzione** (il peso medio delle mele contenute in una cassa, la varianza del peso delle mele, la proporzione di mele che sono marce, ecc.) ...
- oppure **parametri che descrivono un presunto processo che genera le unità statistiche della popolazione** (es. la probabilità di successo in un presunto processo di Bernoulli, il tasso di accadimento di un presunto processo di Poisson, le probabilità dei vari esiti possibili in un presunto esperimento multinomiale, il valore atteso comune a una successione di presunte variabili casuali i.i.d. ecc.)

Si noti che **medie, varianze e proporzioni che si riferiscono a popolazioni infinite** devono sempre essere interpretate come **limiti di successioni**.

Esempio: Per le mele che maturano in un determinato frutteto sotto determinate condizioni climatiche, di irrigazione, di concimazione ecc. (si noti che la popolazione di tutte le mele che maturano in un frutteto è una popolazione concettuale e quindi potenzialmente infinita) si può ipotizzare che all'aumentare della produzione il peso medio si stabilizzi attorno ad un determinato valore μ e questo valore di μ può dunque essere interpretato come peso medio della popolazione di tutte le infinite mele che potenzialmente potrebbero maturare nel frutteto (ovviamente il valore di μ è ignoto e deve essere stimato sulla base di un campione di mele).

Per quanto concerne il **campione**, di solito si tratta di un **sottoinsieme della popolazione** oppure di un **insieme di unità statistiche che secondo la nostra opinione sono state generate da un processo molto simile o comunque strettamente legato a quello che genera le unità statistiche della popolazione**.

Per esempio, quando la popolazione d'interesse è l'insieme delle mele contenute in una determinata cassa possiamo considerare come campione un insieme di 10 mele estratte dalla cassa oppure un insieme di 10 mele che provengono dallo stesso frutteto nel quale sono state raccolte le mele che sono presenti nella cassa.

Chiaramente, a seconda del modo in cui viene selezionato un campione, le informazioni desumibili dal campione possono essere considerate più o meno **rappresentative** per la popolazione di riferimento.

Infatti, **a seconda del modo in cui un campione viene scelto, la distribuzione congiunta delle variabili casuali campionarie cambia**, dove ...

> ... per **variabile casuale campionaria** intendiamo una qualunque variabile casuale che descrive una caratteristica di una unità statistica del campione.

Per esempio, . . .

- . . . se estraiamo in **modo casuale** 10 mele da una cassa e la cassa contiene un numero di mele molto più elevato, possiamo considerare ***realistica*** l'ipotesi che che i pesi delle 10 mele estratte siano realizzazioni di 10 variabili casuali campionarie $\widetilde{X}_1, \widetilde{X}_2, \ldots, \widetilde{X}_n$ i.i.d.[1] con distribuzione identica alla distribuzione del peso nella popolazione di tutte le mele della cassa, . . .
- mentre se scegliamo le 10 mele in **modo tale da favorire la scelta di mele pesanti** (per esempio prendiamo solo mele che ci sembrano grandi o particolarmente mature) la suddetta ipotesi sulla distribuzione delle variabili casuali campionarie $\widetilde{X}_1, \widetilde{X}_2, \ldots, \widetilde{X}_n$ non è più ***realistica***.

[1] Ricordiamo che "i.i.d." è l'acronimo di "indipendenti e identicamente distribuite"

L'esempio sulle mele ci fa capire che la distribuzione congiunta di variabili casuali campionarie dipende

- dal modo in cui viene scelto il campione . . .
- . . . e dal modo in cui è definita la popolazione di riferimento e quindi anche **dai valori degli ignoti parametri che descrivono la popolazione!!!**

Esercizio 1.1

Un macchinario produce lenti a contatto. Sia p l'ignota proporzione di lenti difettose prodotte dal macchinario e si consideri un campione casuale di $n = 10$ lenti. Per ciascuna lente del campione si definisca una variabile casuale campionaria $\widetilde{X}_i$ che assume il valore 1 se la lente di riferimento è difettosa, e che assume il valore 0 altrimenti ($i = 1, 2, \ldots, 10$).

a) Si espliciti un'ipotesi realistica per la distribuzione congiunta delle variabili casuali campionarie $\widetilde{X}_i$.

Esercizio (continua)

b) Si consideri l'ipotesi nella risposta al quesito a). Qual è la corrispondente distribuzione della variabile casuale che restituisce il numero complessivo di lenti a contatto difettose nel campione?

Esercizio (continua)

c) Si assuma ora che l'ignota proporzione p di lenti difettose sia pari a $0,01$. Qual è la probabilità che un campione di $n = 10$ lenti contenga almeno una lente difettosa?

Esercizio (continua)

d) Si assuma ancora che l'ignota proporzione p di lenti difettose sia pari a $0,01$. Qual è la probabilità che un campione di $n = 100$ lenti contenga almeno una lente difettosa?

Esercizio (continua)

e) Si assuma sempre ancora che l'ignota proporzione p di lenti difettose sia pari a $0,01$. Qual è la probabilità che un campione di $n = 1000$ lenti contenga più di 25 lenti difettose?

Esercizio (continua)

f) Si supponga ora che in un campione di $n = 1000$ lenti siano state trovate più di 25 lenti difettose. Alla luce di questo fatto è plausibile che l'ignoto valore di p sia uguale a $0,01$ oppure ad un valore più piccolo?

Esercizio 1.2

Si assuma che per un determinato tragitto il consumo di gasolio (espresso in litri) di un autotreno sia rappresentato da una variabile casuale normale con valore atteso μ ignoto e varianza $\sigma^2 = 9$.

a) Si ipotizzi che l'ignoto valore di μ sia dato da $\mu = 80$ litri. Sotto questa ipotesi, qual è la probabilità che per il tragitto in questione l'autotreno consumi più di 86 litri di gasolio?

Esercizio (continua)

b) Si ipotizzi sempre ancora che l'ignoto valore di μ sia dato da $\mu = 80$ litri e si consideri un campione di $n = 3$ autotreni (con caratteristiche simili) che percorrono il tragitto. Qual è la probabilità che il consumo complessivo di gasolio di tutti e tre gli autotreni sia maggiore di 260 litri?

Esercizio (continua)

c) Si supponga ora che $n = 3$ autotreni (con caratteristiche simili) abbiano percorso il tragitto in questione consumando complessivamente più di 260 litri di gasolio. Alla luce di questo fatto è plausibile che l'ignoto consumo atteso μ non sia superiore a 80 litri?

Esercizio (continua)

d) Come cambierebbero le risposte ai quesiti a), b) e c) se la varianza σ^2 fosse uguale $\sigma^2 = 49$ e non a $\sigma^2 = 9$?

Come si intuisce dagli esercizi che abbiamo appena visto, **le realizzazioni delle variabili casuali campionarie forniscono importanti informazioni sulla popolazione di riferimento.**

Per sfruttare queste informazioni in problemi di stima e/o in problemi di verifica d'ipotesi, di solito si calcolano delle **quantità di sintesi** che vengono chiamate **statistiche campionarie**.

Una **statistica campionaria** è una qualsiasi quantità che può essere calcolata a partire dalle realizzazioni di variabili casuali campionarie.

Esempi di statistiche campionarie sono

- la somma delle realizzazioni delle variabili casuali campionarie $\sum_{i=1}^{n} x_i$
- la **media campionaria** $\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$
- la **varianza campionaria** $v_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$ (come vedremo tra breve, di solito al posto di v_n^2 si calcola la cosiddetta **varianza campionaria corretta** $s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$)
- la **proporzione campionaria** (che indicheremo con $\widehat{p}_n$) che è un caso particolare di media campionaria che si ottiene quando le variabili casuali campionarie sono bernoulliane (assumono il valore 1 se l'unità statistica di riferimento possiede una determinata caratteristica e il valore 0 altrimenti)
- ecc. (mediana campionaria, quantili campionari, differenza interquartile campionaria, range campionario, ...)

Chiaramente, le caratteristiche di un campione possono essere considerate come esito di un esperimento casuale e **le statistiche campionarie sono quindi realizzazioni di variabili casuali**.

Spesso le **statistiche campionarie** vengono considerate come **stime puntuali** per ignoti valori di parametri e in questo caso le corrispondenti variabili casuali vengono chiamate **stimatori**.

Una **stima puntuale** è una qualsiasi statistica campionaria il cui valore viene considerato come *stima* per l'ignoto valore di un **parametro**.

Uno **stimatore** è una variabile casuale la cui realizzazione è una **stima puntuale**.

Nelle prossime *slides* elencheremo le principali proprietà di **tre stimatori molto importanti**:

- La **media campionaria** che è definita come media delle variabili casuali campionarie $\overline{X}$:
$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_i.$$
- La **varianza campionaria corretta** che è definita come
$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\widetilde{X}_i - \overline{X})^2.$$
- La **proporzione campionaria** $\widetilde{p}_n = \frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_i$ che è definita esattamente come la media campionaria ma che si riferisce soltanto al caso particolare dove le variabili casuali campionarie $\widetilde{X}_i$ sono di tipo bernoulliano.

[[Si noti che i nomi di questi **stimatori** sono identici a quelli delle **statistiche campionarie** che sono le loro realizzazioni. Quando vengono utilizzati i termini *"media campionaria"*, *"varianza campionaria (corretta)"* e *"proporzione campionaria"* bisogna quindi fare riferimento al contesto per capire se indicano uno stimatore oppure una stima puntuale.]]

Se non diversamente indicato, le proprietà che elencheremo nelle prossime *slides* si riferiscono sempre all'ipotesi di variabili casuali campionarie $\widetilde{X}_1, \widetilde{X}_2, \ldots, \widetilde{X}_n$ i.i.d.!!!

Chiaramente questa ipotesi non è sempre realistica (per esempio, non è realistica se le variabili casuali campionarie rappresentano le temperature massime di un certo numero di giorni consecutivi o più in generale delle serie storiche dove riteniamo che tra valori consecutivi debba esistere una qualche forma di dipendenza), ...

...ma fornisce le basi teoriche per trattare problemi più complicati.

M1) La media campionaria è uno stimatore **"corretto"**. Questo significa che il valore atteso della media campionaria coincide con il valore atteso comune alle variabili casuali campionarie:

$$E(\overline{X}_n) = E(\widetilde{X}_i) = \mu.$$

M2) La varianza della media campionaria è data da

$$var(\overline{X}_n) = \frac{\sigma^2}{n}$$

dove $\sigma^2 = var(\widetilde{X}_i)$ è la varianza comune alle variabili casuali campionarie.

M3) La media campionaria è uno stimatore **"consistente"**. Questo significa che all'aumentare della numerosità campionaria n la probabilità di commettere un errore di stima di qualsiasi entità $\varepsilon > 0$ tende a zero:

$$\lim_{n\to\infty} P(\{|\overline{X}_n - \mu| > \varepsilon\}) = 0 \qquad \text{per ogni } \varepsilon > 0.$$

M4) (Approssimazione normale) Per valori elevati di n la distribuzione della media campionaria è approssimativamente normale.

L'errore di approssimazione tende a zero all'aumentare della numerosità campionaria n. A parità di n, l'errore di approssimazione dipende dalla forma della distribuzione comune alle variabili casuali campionarie.

L'errore di approssimazione è nullo per ogni $n = 1, 2, \ldots$ **se la distribuzione come alle variabili casuali campionarie è normale** . . .

. . . ed è tanto maggiore quanto più la distribuzione delle variabili casuali campionarie è asimmetrica e/o con code pesanti (come nel TLC).

Esercizio 1.3

Si assuma che ritardi di un treno lungo una determinata tratta siano realizzazioni di variabili casuali i.i.d. con valore atteso μ che è ignoto. Si assuma tuttavia che lo scarto quadratico medio della distribuzione dei ritardi sia noto e pari a $\sigma = 4$ minuti. Si consideri la media campionaria di un campione di $n = 36$ ritardi.

a) Qual è la probabilità che la media campionaria si discosti dall'ignoto ritardo atteso μ per più di 1 minuto?

Esercizio (continua)

b) Come cambierebbe la risposta al quesito precedente se lo scarto quadratico medio σ fosse di 6 minuti?

Esercizio (continua)

c) Come cambierebbe la risposta al quesito precedente se la numerosità campionaria fosse di $n = 25$ ritardi?

Esercizio (continua)

d) Di quanto si deve aumentare la numerosità campionaria n se si vuole che la probabilità richiesta al punto a) sia minore di $0,05$?

Le proprietà M1 - M4 della media campionaria si riferiscono al caso di variabili casuali campionarie i.i.d..

Tuttavia, in molte applicazioni dell'inferenza statistica si considera il caso di **campionamento casuale semplice**, ovvero il caso di **campioni estratti in blocco** (oppure attraverso **estrazioni senza reimmissione**) da una **popolazione finita** e variabili casuali campionarie $\widetilde{X}_i$ che rappresentano valori di un carattere quantitativo rilevato sulle unità statistiche del campione.

[[Si consideri per esempio l'estrazione di un campione di $n = 10$ mele da una cassa che contiene $N = 1000$ mele.]]

Chiaramente, **nel caso di campionamento casuale semplice le variabili casuali campionarie non sono i.i.d.**, ma come vedremo tra breve anche in questo caso la media campionaria soddisfa spesso proprietà analoghe a quelle che abbbiamo visto nel caso di variabili casuali campionarie i.i.d..

La proprietà M1 della media campionaria rimane valida anche nel caso di **campionamento casuale semplice**, mentre le proprietà M2, M3 e M4 subiscono le seguenti modifiche:

M2') $var(\overline{X}_n) = \frac{\sigma^2}{n} \times \frac{N-n}{N-1}$, dove N è la numerosità della popolazione.

M3') La media campionaria è uno stimatore **"consistente"**. Questo significa che all'aumentare della numerosità N della popolazione e della numerosità n del campione, la probabilità di commettere un errore di stima di qualsiasi entità $\varepsilon > 0$ tende a zero:

$$\lim_{N,n\to\infty} P(\{|\overline{X}_n - \mu| > \varepsilon\}) = 0 \quad \text{per ogni } \varepsilon > 0.$$

M4') (Approssimazione normale) Se la popolazione non contiene troppi valori estremi e se N e $N - n$ sono entrambi elevati, la distribuzione della media campionaria è approssimativamente normale. L'errore di approssimazione tende a zero all'aumentare di n e $N - n$. A parità di n e $N - n$ l'errore di approssimazione dipende dalla forma della distribuzione del carattere quantitativo X nella popolazione.

Esercizio 1.4

Si consideri una popolazione di $N = 500$ studenti e un campione casuale di $n = 25$ studenti. Si assuma che con riferimento all'intera popolazione lo scarto quadratico medio σ delle stature degli studenti sia noto e pari a 6 cm.

a) Qual è la probabilità che la media campionaria si discosti dall'ignota statura media μ per più di $2,5$ cm?

Esercizio (continua)

b) Come cambierebbe la risposta al quesito precedente se lo scarto quadratico medio σ fosse pari a 10 cm?

Esercizio (continua)

c) Come cambierebbe la risposta al quesito precedente se la numerosità campionaria fosse di $n = 36$ studenti?

Esercizio (continua)

d) Di quanto si deve aumentare la numerosità campionaria n se si vuole che la probabilità richiesta al punto a) sia minore di $0,01$?

Come abbiamo già visto, la frequenza relativa campionaria $\widetilde{p}_n$ è un caso particolare di media campionaria che si ottiene quando le variabili casuali campionarie $\widetilde{X}_i$ sono di tipo bernoulliano:

$$\widetilde{p}_n = \frac{1}{n} \sum_{i=1}^{n} \widetilde{X}_i$$

Pertanto non è una sorpresa che le proprietà della frequenza relativa campionaria siano perfettamente analoghe a quelle della media campionaria.

Tuttavia, il fatto che una distribuzione bernoulliana sia univocamente determinata dal valore della probabilità di successo p comporta alcune peculiarità che evidenzieremo.

P1) La frequenza relativa campionaria è uno stimatore corretto:

$$E(\widetilde{p}_n) = E(\widetilde{X}_i) = p,$$

dove p è l'ignota probabilità di successo comune a tutte le variabili casuali campionarie $\widetilde{X}_i$ (che sono bernoulliane).

P2) La varianza della frequenza relativa campionaria è data da

$$var(\widetilde{p}_n) = \frac{p(1-p)}{n}$$

Siccome

$$\sup_{p \in (0,1)} p(1-p) = \frac{1}{4},$$

la proprietà P2 implica che

$$var(\widetilde{p}_n) = \frac{p(1-p)}{n} \leq \frac{1}{4n}.$$

P3) La frequenza relativa campionaria è uno stimatore **"consistente"**:

$$\lim_{n\to\infty} P(\{|\widetilde{p}_n - p| > \varepsilon\}) = 0 \qquad \text{per ogni } \varepsilon > 0.$$

P4) (Approssimazione normale) Per valori elevati di n la distribuzione della frequenza relativa campionaria è approssimantivamente normale.

L'errore di approssimazione è trascurabile se $np(1-p) > 5$ (come nel TLC).

Esercizio 1.5

Sia p l'ignota probabilità che un treno venga soppresso e si assuma che di giorno in giorno la decisione di sopprimere il treno venga presa in modo indipendente.

a) Si supponga che l'ignota probabilità p sia uguale a $0,10 = 10\%$ e si consideri un campione di $n = 36$ giorni. Qual è la probabilità che la frequenza relativa campionaria si discosti dall'ignota probabilità p per più di $0,02 = 2\%$?

Esercizio (continua)

b) Come cambierebbe la risposta al quesito precedente se l'ignota probabilità p fosse uguale a $0,25 = 25\%$?

Esercizio (continua)

c) Come cambierebbe la risposta al quesito precedente se al posto di un campione di $n = 36$ giorni si considerasse un campione di $n = 49$ giorni?

Esercizio (continua)

d) Di quanto si deve aumentare la numerosità campionaria n se si vuole che la probabilità richiesta al punto a) sia minore di $0,01$?

Le proprietà P1 - P4 della frequenza relativa campionaria si riferiscono al caso di variabili casuali campionarie i.i.d., ...

...ovvero al caso in cui le variabili casuali campionarie (che per ipotesi sono bernoulliane) rappresentano gli esiti delle prove di un processo di Bernoulli.

Nel caso di **campionamento casuale semplice**, ovvero nel caso in cui le variabili casuali campionarie indicano la presenza/assenza di una particolare caratterisca nelle unità statistiche ottenute mediante un'unica estrazione in blocco (oppure n estrazioni senza ripositizione) da una popolazione finita, le proprietà della frequenza relativa campionaria subiscono alcune modifiche.

La proprietà P1 della frequenza relativa campionaria rimane valida anche nel caso di **campionamento casuale semplice**, mentre le proprietà M2, M3 e P4 subiscono le seguenti modifiche:

P2') $var(\widetilde{p}_n) = \frac{p(1-p)}{n} \times \frac{N-n}{N-1}$, dove N è la numerosità della popolazione.

P3') La frequenza relativa campionaria è uno stimatore **"consistente"**. Questo significa che all'aumentare della numerosità N della popolazione e della numerosità n del campione, la probabilità di commettere un errore di stima di qualsiasi entità $\varepsilon > 0$ tende a zero:

$$\lim_{N,n\to\infty} P(\{|\widetilde{p}_n - p| > \varepsilon\}) = 0 \quad \text{per ogni } \varepsilon > 0.$$

P4') (Approssimazione normale) Se i valori di $np(1-p)$ e $(N-n)p(1-p)$ sono entrambi elevati (maggiori di 5), la distribuzione della frequenza relativa campionaria è approssimativamente normale.

Esercizio 1.6

Si consideri una popolazione composta da $N = 500$ studenti e un campione casuale di $n = 25$ studenti. Sia p l'ignota proporzione di fumatori nella popolazione.

a) Si supponga che l'ignota proporzione p sia uguale a $0,20 = 20\%$. Qual è la probabilità che la frequenza relativa campionaria si discosti dall'ignota proporzione p per più di $0,04 = 4\%$?

Esercizio (continua)

b) Come cambierebbe la risposta al quesito precedente se l'ignota proporzione p fosse uguale a $0,50 = 50\%$?

Esercizio (continua)

c) Come cambierebbe la risposta al quesito precedente se la numerosità campionaria fosse di $n = 36$ studenti?

Esercizio (continua)

d) Di quanto si deve aumentare la numerosità campionaria n se si vuole che la probabilità richiesta al punto a) sia minore di $0,01$?

Come si deduce dal seguente elenco, anche la varianza campionaria corretta

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\widetilde{X}_i - \overline{X})^2$$

soddisfa proprietà analoghe a quelle della media campionaria.

S1) La varianza campionaria corretta è uno stimatore corretto:

$$E(S_n^2) = \sigma^2,$$

dove $\sigma^2 = var(\widetilde{X}_i)$ è la varianza comune a tutte le variabili casuali campionarie $\widetilde{X}_i$.

S2) La varianza della varianza campionaria corretta è data da

$$var(S_n^2) = \begin{cases} \frac{1}{2}(\mu_4 + \sigma^4) & \text{se } n = 2 \\ \frac{\mu_4}{n} - \frac{\sigma^4(n-3)}{n(n-1)} & \text{se } n \geq 3 \end{cases}$$

dove $\mu_4 = E(\widetilde{X}_i^4)$ è il momento quarto comune a tutte le variabili casuali campionarie.

[[Questa proprietà non è richiesta all'esame.]]

S3) La varianza campionaria corretta è uno stimatore **"consistente"**:

$$\lim_{n\to\infty} P(\{|S_n^2 - \sigma^2| > \varepsilon\}) = 0 \quad \text{per ogni } \varepsilon > 0.$$

La proprietà di consistenza viene spesso utilizzata per giustificare l'approssimazione

$$S_n^2 \simeq \sigma^2 \quad \text{per } n \text{ } \textit{"sufficientemente elevato"}.$$

Si può dimostrare che questa approssimazione rimane valida **anche nel caso di campionamento casuale semplice**.

S4) (Approssimazione normale) Per valori elevati di n la distribuzione della varianza campionaria corretta è approssimativamente normale.

L'errore di approssimazione tende a zero all'aumentare di n e a parità di n dipende dalla distribuzione comune alle variabili casuali campionarie.

[[Anche questa proprietà non servirà per l'esame.]]

Nel caso particolare in cui la distribuzione comune alle variabili casuali campionarie è **normale**, si può anche esplicitare la **distribuzione *esatta* della varianza campionaria corretta**.

Infatti, nel caso di variabili casuali campionarie normali si può dimostrare che

$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1} \quad \text{per } n = 2, 3, \ldots,$$

ovvero che la variabile casuale $\frac{(n-1)S_n^2}{\sigma^2}$ ha distribuzione χ^2_{n-1} (si legge "distribuzione chi-quadrato con $n-1$ gradi di liberà" - tra breve vedremo alcune applicazioni delle distribuzioni chi-quadrato nell'inferenza statistica).

Molte tecniche dell'inferenza statistica che vedremo in questo corso sono fondate su questo risultato teorico.

[[Anche il contenuto di questa *slide* non è richiesto all'esame.]]

Stima intervallare

Introduzione

Come abbiamo visto, una **"stima puntuale" è la realizzazione di una variabile casuale che si chiama "stimatore"** e di solito questa realizzazione non coincide con il valore del parametro che si vuole stimare.

Per questo motivo una stima puntuale dovrebbe sempre essere corredata di una misura per la sua precisione.

Per fornire un'informazione sulla precisione con cui la realizzazione $\overline{x}_n$ di una media campionaria $\overline{X}_n$ stima l'ignota media μ di una popolazione si potrebbe corredare il valore di $\overline{x}_n$ con quello della sua **deviazione standard** $\sigma_{\overline{X}} = \sigma/\sqrt{n}$

[[. . . oppure di una stima puntuale della deviazione standard qualora la varianza σ^2 della popolazione fosse ignota; per ottenere una stima puntuale di $\sigma_{\overline{X}}$ basta sostituire nella formula per calcolare $\sigma_{\overline{X}}$ la stima puntuale s^2 al posto di σ^2,]] . . .

. . . ma spesso nelle applicazioni si preferisce una soluzione più elegante, ovvero quella di aggiungere e togliere al valore di una stima puntuale un cosiddetto

margine d'errore

onde ottenere un cosiddetto

intervallo di confidenza

ovvero una cosiddetta

stima intervallare.

Ma che cos'è esattamente un ***"margine d'errore"*** e come si calcola il valore di un ***"margine d'errore"***???

Nelle prossime *slides* risponderemo a queste domande . . .

Stima intervallare

Intervalli di confidenza per una media μ

Supponiamo di voler stimare l'ignota media μ di una popolazione.

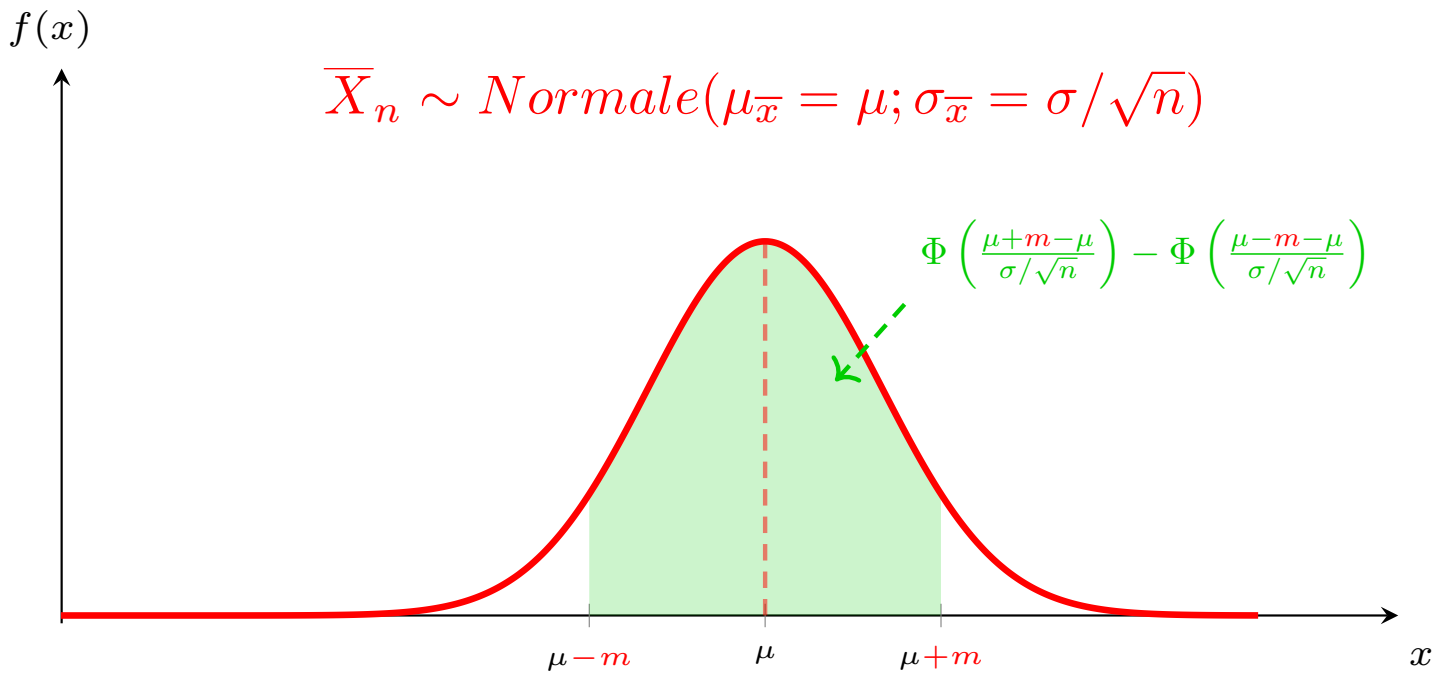
Per semplicità assumeremo inizialmente che la distribuzione comune alle variabili casuali campionarie sia normale con media μ uguale all'ignota media μ che vogliamo stimare e con varianza σ^2 che è nota.

Come consegue dalle proprietà della media campionaria $\overline{X}_n$, ...

. . . **in questo caso anche la distribuzione $\overline{X}_n$ è normale** (con media uguale all'ignota media μ della popolazione e con varianza $\sigma^2_{\overline{x}} = \sigma/\sqrt{n}$) e per qualunque $m > 0$ possiamo dunque scrivere

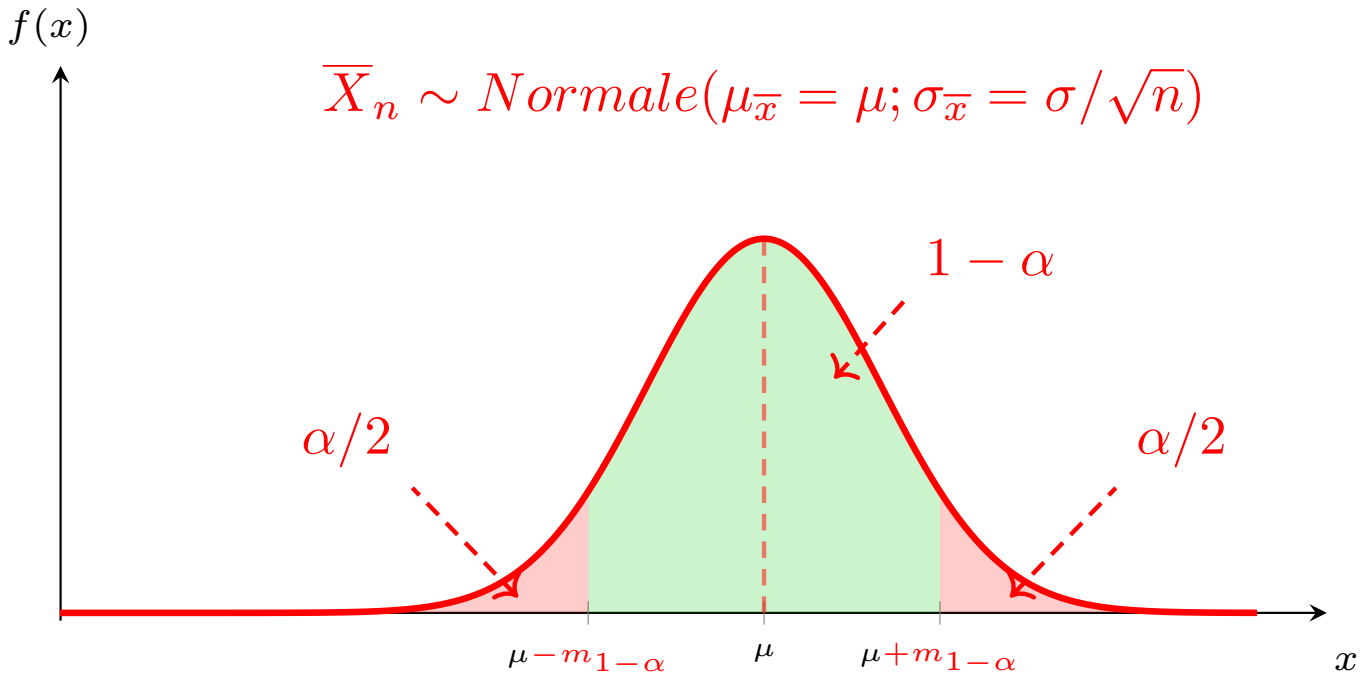
$$P\left(\left\{|\overline{X}_n - \mu| \leq m\right\}\right) = P\left(\left\{\mu - m \leq \overline{X}_n \leq \mu + m\right\}\right) =$$

$$= \Phi\left(\frac{\mu + m - \mu}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{\mu - m - \mu}{\sigma/\sqrt{n}}\right) = \cdots = 2\Phi\left(\frac{m}{\sigma/\sqrt{n}}\right) - 1.$$



L'errore di stima $m = m_{1-\alpha}$ che non viene superato con probabilità $1 - \alpha$ è quindi la soluzione dell'equazione

$$P\left(\left\{|\overline{X}_n - \mu| \leq m\right\}\right) = 2\Phi\left(\frac{m}{\sigma/\sqrt{n}}\right) - 1 = 1 - \alpha$$



. . . e questo valore di $m = m_{1-\alpha}$ viene chiamato

margine d'errore con livello di confidenza $1 - \alpha$.

Semplici ragionamenti mostrano che . . .

$$2\Phi\left(\frac{m}{\sigma/\sqrt{n}}\right) - 1 = 1 - \alpha \quad \Rightarrow \quad \Phi\left(\frac{m}{\sigma/\sqrt{n}}\right) = 1 - \alpha/2 \quad \Rightarrow$$

$$\Rightarrow \quad \frac{m}{\sigma/\sqrt{n}} = z_{1-\alpha/2} \quad \Rightarrow \quad m = m_{1-\alpha} = z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Il **margine d'errore con livello di confidenza** $1-\alpha$ è quindi l'errore di stima che non viene superato con probabilità $1-\alpha$. Aggiungendo e sottraendo il margine d'errore al valore della corrispondente stima puntuale si ottiene un cosiddetto **intervallo di confidenza (IdC) con livello di confidenza** $1-\alpha$.

Nel caso particolare di una media campionaria che si riferisce a variabili casuali campionarie normali i.i.d. il margine d'errore con livello di confidenza $1-\alpha$ è dato da

$$m = m_{1-\alpha} = z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

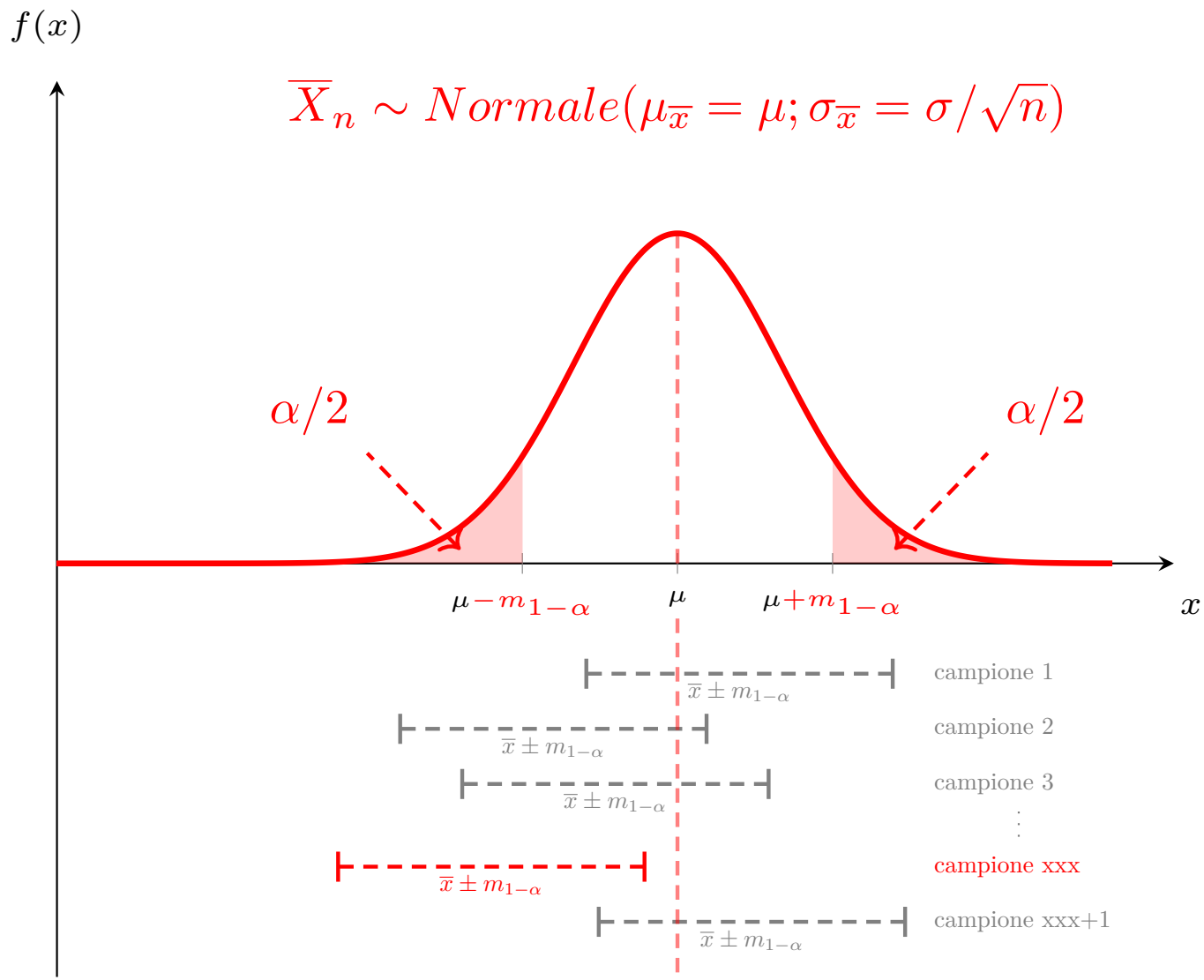
e il corrispondente IdC con livello di confidenza $1-\alpha$ è dato da

$$\overline{x}_n \pm m_{1-\alpha} = \overline{x}_n \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

E' importante tenere presente che **l'ignoto valore del parametro che si vuole stimare non è necessariamente contenuto nell'IdC che viene calcolato a partire da un dato campione!!!**. ...

Infatti, se consideriamo ancora il caso particolare di una media campionaria calcolata a partire dalle realizzazioni di variabili casuali campionarie i.i.d. con distribuzione normale vediamo che ...

. . . l'IdC con estremi dati da $\overline{x}_n \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$ contiene l'ignoto valore di μ soltanto "nell'$(1-\alpha)\times 100\%$ dei casi":



Si ricordi che per ottenere gli estremi dell'IdC

$$\overline{x}_n \pm m_{1-\alpha} = \overline{x}_n \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

abbiamo dovuto ***assumere*** che le variabili casuali campionarie siano i.i.d. con distribuzione normale e che il valore di σ sia noto.

Chiaramente, nelle applicazioni non possiamo mai essere sicuri che queste ipotesi siano soddisfatte e pertanto il **livello di confidenza effettivo** dell'intervallo di confidenza

$$\overline{x}_n \pm m_{1-\alpha} = \overline{x}_n \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

potrebbe anche differire dal **livello di confidenza nominale** $1-\alpha$!!!.

Esercizio 2.1 (IdC per durata media di batterie)

In un test effettuato su un campione di $n = 36$ batterie del tipo XXX è stata ottenuta una durata media di 54 ore. Si assuma che lo scarto quadratico medio della durata di una batteria ammonti a $\sigma = 3$ ore.

a) Si calcoli un IdC al 95% per l'ignota durata media μ di una batteria del tipo XXX.

Esercizio (continua)

b) Si consideri l'IdC calcolato nella risposta al quesito a). A quanto ammonterebbe il suo livello di confidenza effettivo se lo scarto quadratico medio della durata di una batteria fosse di 5 ore e non di 3 come ipotizzato in precedenza?

Soluzione

Siccome è ragionevole ritenere che (sotto condizioni di utilizzo simili) le durate di batterie dello stesso tipo differiscano tra di loro soltanto per **cause accidentali**, possiamo assumere che le durate siano realizzazioni di variabili casuali normali e pertanto possiamo ritenere che il livello di confidenza effettivo dell'intervallo di confidenza

$$\overline{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = 54 \pm 1,96\frac{3}{\sqrt{36}} = \begin{cases} 54,98 \\ 53,02 \end{cases}$$

sia molto prossimo a $1 - \alpha = 95\%$.

Esercizio 2.2 (IdC per statura media)

La media delle stature rilevate su un campione di $n = 100$ studenti ammonta a $\overline{x}_n = 176cm$. Si assuma che lo scarto quadratico medio della distribuzione delle stature nella popolazione di riferimento sia noto e pari a $\sigma = 5cm$.

a) Si calcoli un IdC al 99% per l'ignota media μ della distribuzione delle stature di tutti gli studenti che appartengono alla popolazione di riferimento.

Esercizio (continua)

b) Si consideri l'IdC calcolato nella risposta al quesito a). A quanto ammonterebbe il suo livello di confidenza effettivo se lo scarto quadratico medio delle stature di tutti gli studenti fosse di $6cm$ e non di $5cm$ come ipotizzato in precedenza.

Soluzione

Numerosi studi antropologici sostengono l'ipotesi che la distribuzione delle stature di esseri umani di età più o meno simile sia approssimativamente normale e pertanto possiamo ritenere che il livello di confidenza effettivo dell'intervallo di confidenza

$$\overline{x} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} = 176 \pm 2,58 \frac{5}{\sqrt{100}} = \begin{cases} 177,29 \\ 174,71 \end{cases}$$

sia molto prossimo a $1 - \alpha = 99\%$.

Finora abbiamo soltanto visto come si calcola un IdC per la media di una popolazione nel caso in le seguenti condizioni sono entrambe soddisfatte:

- le variabili casuali campionarie hanno distribuzione normale (o approssimativamente normale) con media μ uguale all'ignota media μ che si vuole stimare;
- la varianza σ^2 comune alle variabili casuali campionarie è nota.

Tuttavia, il ragionamento che ci ha condotti all'espressione del margine d'errore

$$m = m_{1-\alpha} = z_{1-\alpha/2}\sigma_{\overline{x}} = z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

rimane valido **anche se la distribuzione comune alle variabili casuali campionarie non è normale**, ma in questo caso il suddetto margine d'errore assicurerà soltanto che

$$P\left(\left\{|\overline{X}_n - \mu| \leq m\right\}\right) \simeq 1 - \alpha$$

per numerosità campionarie n sufficientemente elevate.

L'errore di approssimazione tende a zero all'aumentare di n e a parità di n l'errore di approssimazione è tanto maggiore quanto più la popolazione di riferimento è diversa da una popolazione normale.

Quindi concludiamo che ...

Anche se la distribuzione comune alle variabili casuali campionarie non è normale, la formula

$$\overline{x}_n \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

restituisce comunque gli estremi di un **IdC asintotico** per μ, ovvero di un **IdC con livello di confidenza effettivo prossimo a** $1-\alpha$ **per valori elevati di** n.

La differenza tra il **livello di confidenza effettivo** e il **livello di confidenza nominale** $1-\alpha$ è tanto più piccola quanto più elevato è il valore di n e quanto più la forma della distribuzione comune alle variabili casuali campionare è simile alla forma di una distribuzione normale. . . . (continua sulla prossima *slide*) . . .

- se la distribuzione delle variabili casuali campionarie è **approssimativamente normale**, il livello di confidenza **effettivo** è molto prossimo a quello **nominale** già a partire da $n = 10$;
- se la distribuzione delle variabili casuali campionarie è soltanto **lievemente asimmetrica e/o non presenta code troppo pesanti**, la numerosità campionaria minima per ottenere un'approssimazione accettabile aumenta a $n = 30$;
- se invece la distribuzione delle variabili casuali campionarie è **molto asimmetrica e/o una o entrambe le sue code sono molto più pesanti di quelle di una distribuzione normale**, allora la numerosità campionaria minima per ottenere un'approssimazione accettabile aumenta a $n = 50$, ma in casi estremi potrebbe anche essere molto più elevata.

Chiaramente l'IdC

$$\overline{x}_n \pm m_{1-\alpha} = \overline{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

non può essere calcolato se non si conosce il valore di σ^2.

Tuttavia, in questi casi si può calcolare un IdC molto simile che è definito come

$$\overline{x}_n \pm t_{n-1;\alpha/2}\frac{s_n}{\sqrt{n}}$$

dove

- $t_{n-1;\alpha/2}$ è il percentile di ordine $1-\alpha/2$ della **distribuzione t di Student con $n-1$ gdl** (il valore di $t_{n-1;\alpha/2}$ è reperibile in un'apposita tavola),
- e dove

$$s_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

è la deviazione standard campionaria (ovvero la radice della varianza campionaria s_n^2).

Le distribuzioni t di Student alle quali si riferiscono i percentili $t_{n-1;\alpha/2}$ che servono per calcolare gli estremi dell'IdC

$$\overline{x}_n \pm t_{n-1;\alpha/2} \frac{s_n}{\sqrt{n}}$$

hanno andamento campanulare e simmetrico molto simile a quello di una distribuzione normale standard, ma hanno code più "pesanti". Le distribuzioni t di Student formano una famiglia di distribuzioni che si distinguono per il valore di un parametro che viene chiamato

"gradi di libertà".

Il numero di **"gradi di libertà"** che identifica una distribuzione t di Student è un numero intero positivo e all'aumentare del numero di gradi di libertà (gdl) la corrispondente distribuzione t di Student si "avvicina" sempre di più alla distribuzione normale standard. La distribuzione t di Studenti con $n = 120$ gdl è praticamente indistinguibile dalla distribuzione normale standard.

La seguente tabella mostra la parte iniziale di una tavola della distribuzione t di Student:

Tavola della distribuzione t di Student						
Gradi di libertà	Area nella coda superiore					
	0,2	0,1	0,05	0,025	0,01	0,005
1	1,376	3,078	6,314	12,706	31,821	63,657
2	1,061	1,886	2,920	4,303	6,965	9,925
3	0,978	1,638	2,353	3,182	4,541	5,841
4	0,941	1,533	2,132	2,776	3,747	4,604
5	0,920	1,476	2,015	2,571	3,365	4,032
6	0,906	1,440	1,943	2,447	3,143	3,707
7	0,896	1,415	1,895	2,365	2,998	3,499
8	0,889	1,397	1,860	2,306	2,896	3,355
9	0,883	1,383	1,833	2,262	2,821	3,250
10	0,879	1,372	1,812	2,228	2,764	3,169
11	0,876	1,363	1,796	2,201	2,718	3,106
12	0,873	1,356	1,782	2,179	2,681	3,055

Ciascuna riga si riferisce alla distribuzione t di Student con un determinato numero di gradi di libertà (gdl) e fornisce alcuni percentili della corrispondente distribuzione.

L'IdC

$$\overline{x}_n \pm t_{n-1;\alpha/2} \frac{s_n}{\sqrt{n}}$$

ha caratteristiche molto simili a quello basato sul valore vero della varianza σ^2:

- entrambi gli IdC sono **esatti** (nel senso che il loro livello di confidenza effettivo è esattamente uguale a $1-\alpha$) **se le variabili casuali campionarie hanno distribuzione normale**,
- entrambi gli IdC sono soltanto **asintotici** (ovvero con livello di confidenza effettivo che si avvicina a $1-\alpha$ all'aumentare di n) **se le variabili casuali campionarie non hanno distribuzione normale**.

Esercizio 2.3

Si assuma che ritardi di un treno lungo una determinata tratta siano realizzazioni di variabili casuali i.i.d. con valore atteso μ che è ignoto. Si assuma inoltre che la somma dei ritardi delle ultime $n = 36$ corse ammonti a $\sum_{i=1}^{36} x_i = 731,21$ minuti e che la somma dei quadrati di tali ritardi sia data da $\sum_{i=1}^{36} x_i^2 = 15232,46$.

Si calcoli un IdC al 95% per l'ignoto valore atteso μ. Che cosa si può dire sul livello di confidenza effettivo dell'IdC?

Esercizio 2.4

Si consideri una popolazione finita composta da $N = 900$ studenti e si assuma che la media delle stature misurate su un campione di $n = 25$ studenti ammonti a $\bar{x} = 174cm$ e che la deviazione standard di tali stature sia di $s = 5cm$.

Si calcoli un IdC al 99% per per l'ignota media μ delle stature di tutti gli studenti che appartengono alla popolazione di riferimento. Che cosa si può dire sul livello di confidenza effettivo dell'IdC?

Chiaramente, **un IdC più ampio fornisce un'informazione meno precisa sull'ignoto valore del parametro al quale si riferisce** (nei casi che abbiamo visto finora il parametro era sempre una media μ), ...

...e come si desume dalle formule per calcolare il margine d'errore

$$m = m_{1-\alpha} = \begin{cases} z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} & \text{se il valore di } \sigma \text{ è noto,} \\ t_{\alpha/2;n-1}\frac{s}{\sqrt{n}} & \text{se si utilizza la stima } s \text{ di } \sigma, \end{cases}$$

...**l'ampiezza di un intervallo di confidenza per μ è inversamente proporzionale alla radice della numerosità campionaria n.**

$\Rightarrow$ Aumentando la numerosità campionaria n si possono quindi ottenere IdC più stretti (e quindi più precisi).

A quanto deve ammontare la numerosità campionaria $n = n^*$ affinché il margine d'errore di un IdC per μ raggiunga un determinato ***valore obiettivo*** $m^*_{1-\alpha}$ *???*

Per rispondere a questa domanda conviene partire dal caso (poco realistico) in cui **il valore di σ è noto a priori**. Come abbiamo visto, in questo caso il margine d'errore è dato da

$$m_{1-\alpha} = z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

e per raggiungere un determinato valore obbiettivo $m^*_{1-\alpha}$ bisogna quindi scegliere la numerosità campionaria n in modo tale che

$$z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = m^*_{1-\alpha} \quad \Leftrightarrow \quad n = n^* = \left(z_{1-\alpha/2}\frac{\sigma}{m^*_{1-\alpha}}\right)^2.$$

Vediamo ora invece come si può procedere nel **caso in cui il valore di σ è ignoto**.

In questo caso il margine d'errore

$$m_{1-\alpha} = t_{\alpha/2;n-1} \frac{s}{\sqrt{n}}$$

dipende dalla realizzazione s della varianza campionaria corretta S_n^2 ...

... e siccome s non assume sempre lo stesso valore, **non esiste una numerosità campionaria $n = n^*$ tale che il margine d'errore sia (con assoluta certezza) uguale ad un determinato valore obiettivo $m^*_{1-\alpha}$!!!**

Per superare questo problema, ...

. . . di solito si procede come nel caso in cui la varianza σ^2 è nota, ma al posto del valore di σ^2 (che nel caso in questione è ignoto) si sostituisce un **valore plausibile** $\widehat{\sigma}$ che può essere determinato sulla base di

- . . . una stima ottenuta da un precedente campione proveniente dalla medesima popolazione oppure da un'altra popolazione con caratteristiche simili;
- . . . una stima ottenuta da un campione preliminare di numerosità ridotta;
- . . . una qualche congettura.

  [[Per esempio, in molte situazioni si è in grado di esprimere *a priori* una congettura circa il valore massimo x_{max} e il valore minimo x_{min} all'interno della popolazione. In questi casi, come valore plausibile $\widehat{\sigma}$ viene spesso utilizzato il valore di $\frac{1}{4}(x_{max} - x_{min})$ al posto di σ.]]

Esercizio 2.5

Si assuma che ritardi di un treno lungo una determinata tratta siano realizzazioni di variabili casuali i.i.d. con valore atteso μ che è ignoto. Si assuma tuttavia che lo scarto quadratico medio della distribuzione dei ritardi sia noto e pari a $\sigma = 4$ minuti.

a) Quante rilevazioni del ritardo sono necessarie se si vuole che il margine d'errore al 95% della media campionaria sia minore di $1,5$ minuti?

Esercizio (continua)

b) Si consideri una media campionaria basata su $n = 36$ rilevazioni del ritardo. A quanto ammonta il livello di confidenza associato ad un margine d'errore di $m = 2$ minuti?

Esercizio 2.6

Si consideri la popolazione di tutti gli studenti iscritti ad un corso di studio e si assuma che le loro stature siano tutte comprese tra $x_{min} = 140cm$ e $x_{max} = 210cm$. Si supponga di essere interessati all'ignota statura media μ.

a) Di quanti studenti si deve rilevare la statura affinché con livello di confidenza pari al 99% il margine d'errore della media campionaria sia minore di $3cm$?

Esercizio (continua)

b) Si consideri una media campionaria calcolata a partire dalle stature di $n = 100$ studenti. A quanto ammonta il livello di confidenza associato ad un margine d'errore di $m = 4cm$?

Il ragionamento che conduce al margine d'errore

$$m_{1-\alpha} = z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

della media campionaria $\overline{X}_n$ può essere replicato in modo quasi identico anche con riferimento alla frequenza relativa campionaria $\widetilde{p}_n$.

[[Si ricordi infatti che **per numerosità campionarie** n **sufficientemente elevate**, la distribuzione della frequenza relativa campionaria $\widetilde{p}$ è **approssimativamente normale** con valore atteso $E(\widetilde{p}) = p$ e varianza $\sigma_p = var(\widetilde{p}) = p(1-p)/n$ (vedi le proprietà P1 e P2 della frequenza relativa campionaria).]]

Infatti, se la numerosità campionaria n è sufficientemente elevata e $m > 0$ è un qualsiasi un numero reale positivo, possiamo scrivere

$$P\left(\{|\widetilde{p}_n - p| \leq m\}\right) = P\left(\{p - m \leq \widetilde{p}_n \leq p + m\}\right) \simeq$$

$$\simeq \Phi\left(\frac{p + m - p}{\sigma_p}\right) - \Phi\left(\frac{p - m - p}{\sigma_p}\right) = \Phi\left(\frac{m}{\sigma_p}\right) - \Phi\left(\frac{-m}{\sigma_p}\right) =$$

$$= \text{ [[simmetria]] } = \Phi\left(\frac{m}{\sigma_p}\right) - \left[1 - \Phi\left(\frac{m}{\sigma_p}\right)\right] =$$

$$= 2\Phi\left(\frac{m}{\sigma_p}\right) - 1 \quad \Rightarrow \quad P\left(\{|\widetilde{p}_n - p| \leq m\}\right) = 2\Phi\left(\frac{m}{\sigma_p}\right) - 1$$

e ponendo

$$2\Phi\left(\frac{m}{\sigma_p}\right) - 1 = 1 - \alpha,$$

otteniamo il **margine d'errore**

$$m = m'_{1-\alpha} = z_{1-\alpha/2}\sigma_p = z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}.$$

Chiaramente, in un problema di stima reale il margine d'errore

$$m = m'_{1-\alpha} = z_{1-\alpha/2}\sigma_p = z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

non può essere calcolato perché il valore di p è ignoto.

Tuttavia, sostituendo al posto dell'ignoto valore di p la frequenza relativa campionaria $\widehat{p}_n$ (ovvero la realizzazione dell'omonimo stimatore $\widetilde{p}_n$) otteniamo una **stima del margine d'errore** che è data da

$$m_{1-\alpha} = z_{1-\alpha/2}\sigma_p = z_{1-\alpha/2}\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}$$

Usando questo margine d'errore "stimato" possiamo calcolare gli **estremi di un IdC con livello di confidenza** $1-\alpha$ come

$$\widehat{p}_n \pm m_{1-\alpha} = \widehat{p}_n \pm z_{1-\alpha/2}\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}.$$

Si noti che anche l'IdC

$$\widehat{p}_n \pm m_{1-\alpha} = \widehat{p}_n \pm z_{1-\alpha/2}\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}.$$

è soltanto un **IdC asintotico** nel senso che il suo **livello di confidenza effettivo è prossimo a quello nominale (ovvero a $1-\alpha$) soltanto se la numerosità campionaria n è sufficientemente elevata**.

Di solito la precisione dell'IdC viene ritenuta accettabile se $n\widehat{p}_n(1-\widehat{p}_n)$ è maggiore di 3. A volte, per un grado di precisione più elevato conviene utilizzare il valore soglia 5 al posto di 3.

$$\widehat{p}_n \pm m_{1-\alpha} = \widehat{p}_n \pm z_{1-\alpha/2}\sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}$$

sono gli estremi di un **IdC asintotico per una ignota proporzione** p.

Per accertare se il livello di confidenza effettivo è prossimo a quello nominale (ovvero a $1-\alpha$) bisogna verificare se il valore di $n\widehat{p}(1-\widehat{p})$ supera la soglia di 3. Se si desidera un grado di precisione elevato questa soglia deve essere innalzata a 5.

Esercizio 2.7

Un campione di $n = 225$ studenti contiene 60 studenti che dichiarano di essere fumatori. Si calcoli un IdC al 98% per l'ignota proporzione di studenti che sono fumatori.

Chiaramente, anche con riferimento alla frequenza relativa campionaria $\widetilde{p}_n$ si pone il problema di **determinare la numerosità campionaria** $n = n^*$ tale che il margine d'errore

$$m_{1-\alpha} = z_{1-\alpha/2}\sigma_p = z_{1-\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

sia uguale o quantomeno prossimo ad un dato **valore obiettivo** $m^*_{1-\alpha}$.

Per risolvere questo problema di solito si considera un **valore plausibile** $\widehat{\widehat{p}}$ al posto di $\widehat{p}$ (infatti, il valore di $\widehat{p}$ è noto soltanto dopo aver osservato il campione) e si determina la numerosità campionaria n^* come

$$z_{1-\alpha/2}\sqrt{\frac{\widehat{\widehat{p}}(1-\widehat{\widehat{p}})}{n}} = m^*_{1-\alpha} \quad \Rightarrow \quad n = n^* = z^2_{1-\alpha/2}\frac{\widehat{\widehat{p}}(1-\widehat{\widehat{p}})}{(m^*_{1-\alpha})^2}.$$

Per ottenere un **valore plausibile** $\widehat{\widehat{p}}$ si può procedere in modo analogo a come abbiamo visto nel caso della varianza σ^2:

- si può utilizzare una stima ottenuta da un campione precedente oppure da un campione che proviene da un'altra popolazione con caratteristiche simili;
- si può utilizzare una stima che proviene da un piccolo campione preliminare;
- si può ricorrere ad un qualche tipo di congettura:
  - se ***a priori*** si è convinti che l'ignota proporzione p non possa essere superiore a $0,10 = 10\%$, si pone $\widehat{\widehat{p}} = 0,10$;
  - se invece si è convinti che l'ignota proporzione p non possa essere minore di $0,90 = 90\%$, si pone $\widehat{\widehat{p}} = 0,90$;
  - in caso di assoluta incertezza si pone $\widehat{\widehat{p}} = 0,50$ onde ottenere il valore massimo possibile di

$$n = n^* = z^2_{1-\alpha/2} \frac{\widehat{\widehat{p}}(1-\widehat{\widehat{p}})}{(m^*_{1-\alpha})^2}.$$

[[Si noti infatti che, a parità di altre condizioni, $\widehat{\widehat{p}} = 0,5$ massimizza il valore di n^*]]

Esercizio 2.8

Un sociologo vuole stimare l'ignota proporzione p di coloro che preferiscono lavorare in modalità *smart*.

a) Quanti lavoratori si devono intervistare per ottenere un margine d'errore con livello di confidenza del 95% che non sia superiore al $2\% = 0,02$?

Esercizio (continua)

b) 450 lavoratori di un campione di $n = 1000$ lavoratori dichiarano di preferire lo *smart working*. A quanto ammonta il livello di confidenza associato ad un margine d'errore di $m = 0,03 = 3\%$?

Test statistici

Definizioni

Un **test statistico** (o test d'ipotesi) è una procedura basata su un campione che si esegue con lo scopo di decidere se una determinata **ipotesi statistica** può essere rifiutata a favore di un'altra ipotesi o meno.

Per **"ipotesi statistica"** si intende una **congettura circa i valori di uno o più parametri** che descrivono una popolazione . . .

. . . e le **ipotesi statistiche** tra le quali si deve decidere nell'ambito di un **test statistico** vengono chiamate **"ipotesi nulla"** (indicata con H_0) e **"ipotesi alternativa"** (indicata con H_a).

Chiaramente, in un **test statistico si possono commettere due tipi d'errore**:

- si decide a favore dell'ipotesi alternativa quando in realtà è vera l'ipotesi nulla;
- oppure si decide a favore dell'ipotesi nulla quando in realtà è vera l'ipotesi alternativa.

Di solito **uno di questi due errori ha conseguenze molto più gravi dell'altro** (es. si decide di somministrare una cura con pesanti effetti collaterali ad un individuo che in realtà non ne avrebbe bisogno) . . .

. . . e per questo motivo **i test statistici vengono di solito costruiti in modo tale l'errore con conseguenze più gravi possa verificarsi con probabilità non superiore ad un prefissata soglia α (prossima a zero)** che viene chiamata **"livello di significatività"** del test.

Per convenzione, nei test statistici **l'errore con conseguenze più gravi è l'errore che si commette se si decide a favore dell'ipotesi alternativa quando in realtà è vera l'ipotesi nulla** e questo errore viene chiamato **"errore di prima specie"**.

L'altro errore, ovvero quello che si commette se si decide a favore dell'ipotesi nulla quando in realtà è vera l'ipotesi alternativa, viene invece chiamato **"errore di seconda specie"**.

$\Rightarrow$ Per fare in modo che l'errore di prima specie sia effettivamente quello con conseguenze più gravi bisogna attribuire correttamente il ruolo di ipotesi nulla ad una delle due ipotesi statistiche tra le quali si deve decidere!!!

I seguenti esempi aiuteranno a capire come si deve scegliere l'ipotesi nulla.

Esempio 3.1 (Test per "dimostrare" ipotesi di ricerca)

Un ricercatore vuole dimostrare che un farmaco sia efficace, ovvero che a seguito della sua somministrazione più dell'80% degli individui malati guarisca. Per dimostrare la sua tesi, ovvero per dimostrare che l'ignota proporzione p di individui che guariscono sia maggiore di $0,80 = 80\%$, il ricercatore deve eseguire un test statistico dove

- l'**ipotesi nulla** è definita come $H_0 : p \leq 0,80$
- l'**ipotesi alternativa** è definita come $H_a : p > 0,80$.

Infatti, nel caso in questione solitamente si ritiene che l'**errore più grave** sia quello di investire ingenti somme di denaro per produrre un farmaco che in realtà è inefficace . . .

. . . e che l'**errore meno grave** sia quello di non produrre un farmaco che in realtà sarebbe efficace.

Chiaramente, possono esserci situazioni dove il secondo errore è più grave del primo. In tal caso bisogna invertire i ruoli delle due ipotesi statistiche.

Esempio 3.2 (Test sulla veridicità di affermazioni)

Secondo quanto dichiarato sulle bustine di zafferano di un determinato produttore, ciascuna bustina conterrebbe almeno $0,15$ grammi di zafferano. Per verificare se questa affermazione è veritiera, si potrebbe eseguire un test statistico che verte sul contenuto medio μ delle bustine. In particolare, si potrebbe eseguire un test statistico con

- $H_0 : \mu \geq 0,15$ grammi
- $H_a : \mu < 0,15$ grammi

Infatti, nel caso in questione solitamente si ritiene che **l'errore più grave** sia quello di accusare ingiustamente il produttore di apporre un'affermazione falsa sulle bustine, ...

...e che **l'errore meno grave** sia quello di non intraprendere azioni legali (o di altro tipo) qualora il contenuto delle bustine fosse veramente inferiore a quanto dichiarato.

Consideriamo ora un test dove le ipotesi statistiche vertono sull'ignota media μ di una popolazione.

Consideriamo inizialmente il caso in cui le ipotesi statistiche messe a confronto sono definite come

- $H_0 : \mu \geq \mu_0$ (l'ignoto valore di μ è maggiore di un dato valore μ_0)
- $H_a : \mu < \mu_0$.

Supponiamo per esempio che μ rappresenti l'ignoto contenuto medio delle bustine di zafferano di un produttore e che $\mu_0 = 0,15$ grammi sia il contenuto dichiarato sulle bustine.

Per semplicità assumeremo inizialmente che lo **scarto quadratico medio della popolazione sia noto** e pari a $\sigma^2 = 0,01$ grammi.

Supponiamo ora di aver esaminato il contenuto di un campione di $n = 36$ bustine di zafferano e che $\overline{x}_n = 0,1465$ grammi sia il contenuto medio delle bustine esaminate.

Sulla base di questo risultato campionario, possiamo ritenere di avere in mano sufficienti prove per rifiutare l'ipotesi nulla $H_0 : \mu \geq \mu_0 = 0,15$ ***grammi???***

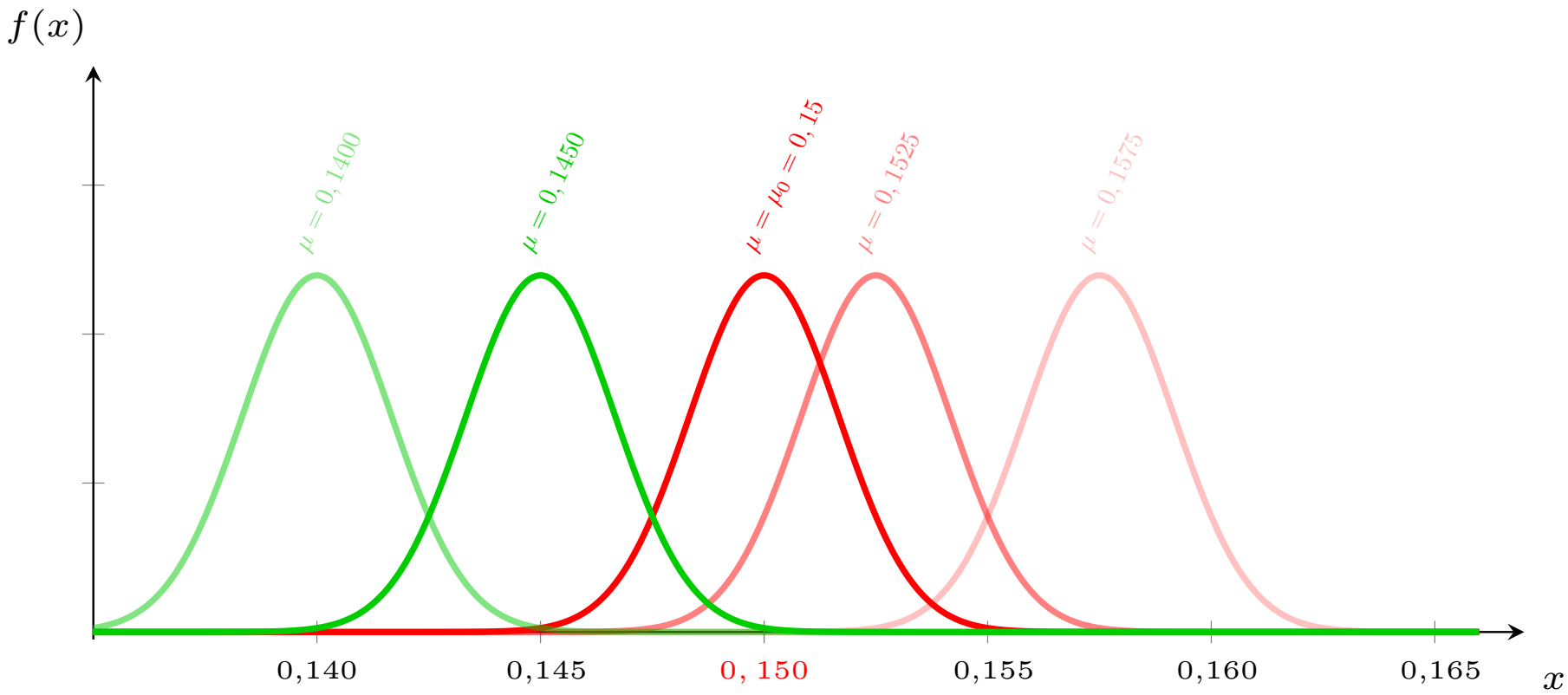
Per rispondere a questa domanda analizzeremo la distribuzione di $\overline{X}_n$. Nel caso in questione si ha

$$\overline{X}_n \sim \text{Normale}\left(\mu_{\overline{x}} = \mu = ???; \quad \sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0,01}{\sqrt{36}}\right)$$

dove

- $\mu \geq \mu_0 = 0,15$ se è vera l'ipotesi nulla $H_0 : \mu \geq \mu_0$,
- e dove $\mu < \mu_0 = 0,15$ se è vera l'ipotesi alternativa $H_a : \mu < \mu_0$.

Grafico di alcune possibili distribuzioni di $\overline{X}_n$.

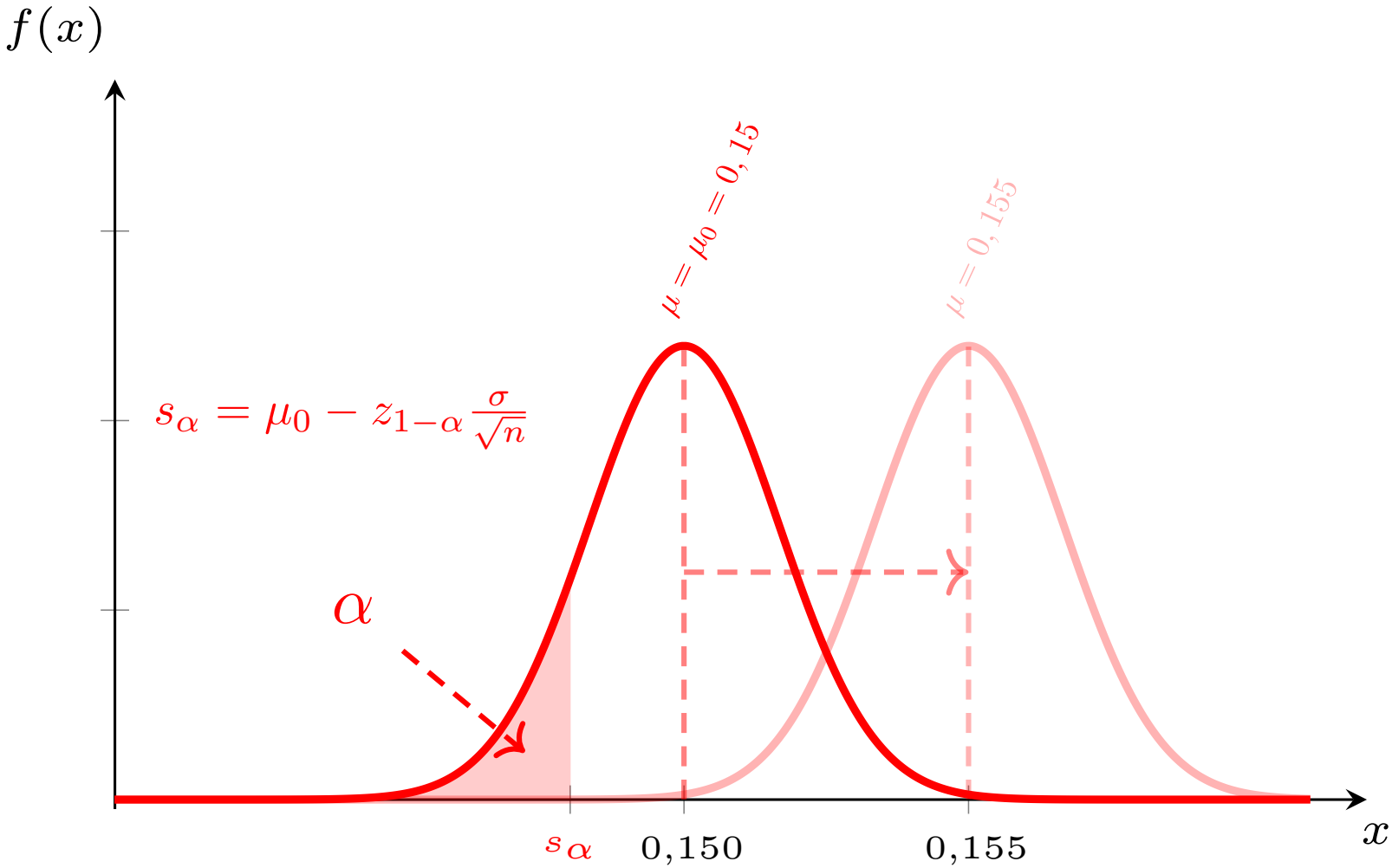


Come si desume dal grafico, nel caso in questione la media campionaria $\overline{X}_n$ è un'ottima **statistica test** su cui basare la decisione. Infatti,

- **se è vera l'ipotesi nulla** $H_0 : \mu \geq \mu_0$, allora è molto probabile osservare valori di $\overline{x}_n$ elevati ed è poco probabile osservare valori piccoli. In particolare, **è poco probabile osservare valori molto più piccoli di** $\mu_0 = 0,15$. Infatti, se scegliamo un valore di $\alpha > 0$ che è prossimo a zero (come per esempio $\alpha = 0,1$ oppure $\alpha = 0,05$) e poniamo $s_\alpha = \mu_0 - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$, otteniamo

$$P\left(\left\{\overline{X}_n < s_\alpha\right\}\right) \leq \alpha \qquad \text{per ogni } \mu \geq \mu_0$$

(vedi il grafico sulla prossima *slide*).


$f(x)$
$s_\alpha = \mu_0 - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$
α
$\mu = \mu_0 = 0,15$
$\mu = 0,155$
s_α
0,150
0,155
x

- D'altra parte, **se è vera l'ipotesi alternativa** $H_a : \mu < \mu_0$, allora è molto probabile osservare valori piccoli di $\overline{x}_n$ (ovvero minori di $s_\alpha = \mu_0 - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$) ed è poco probabile osservare valori elevati (ovvero maggiori di $s_\alpha = \mu_0 - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$).

Supponiamo ora di essere disposti a commettere l'errore di prima specie (rifiutare $H_0 : \mu \geq \mu_0$ quando in realtà è l'ipotesi vera) con probabilità non superiore ad un determinato livello di significatività α (di solito si pone $\alpha = 0,10$, $\alpha = 0,05$ oppure addirittura $\alpha = 0,01$).

Se vogliamo rispettare questo vincolo e vogliamo allo stesso tempo massimizzare la probabilità di scoprire l'ipotesi alternativa $H_a : \mu < \mu_0$ quando essa è vera, dobbiamo seguire la seguente **regola decisionale** (vedi il grafico):

- rifiutare $H_0 : \mu \geq \mu_0$ e decidere a favore di $H_a : \mu < \mu_0$ se osserviamo un campione dove

$$\overline{x}_n < s_\alpha = \mu_0 - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$$

  oppure (equivalentemente) dove

$$z = \frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} < -z_{1-\alpha} = z_{1-\alpha}.$$

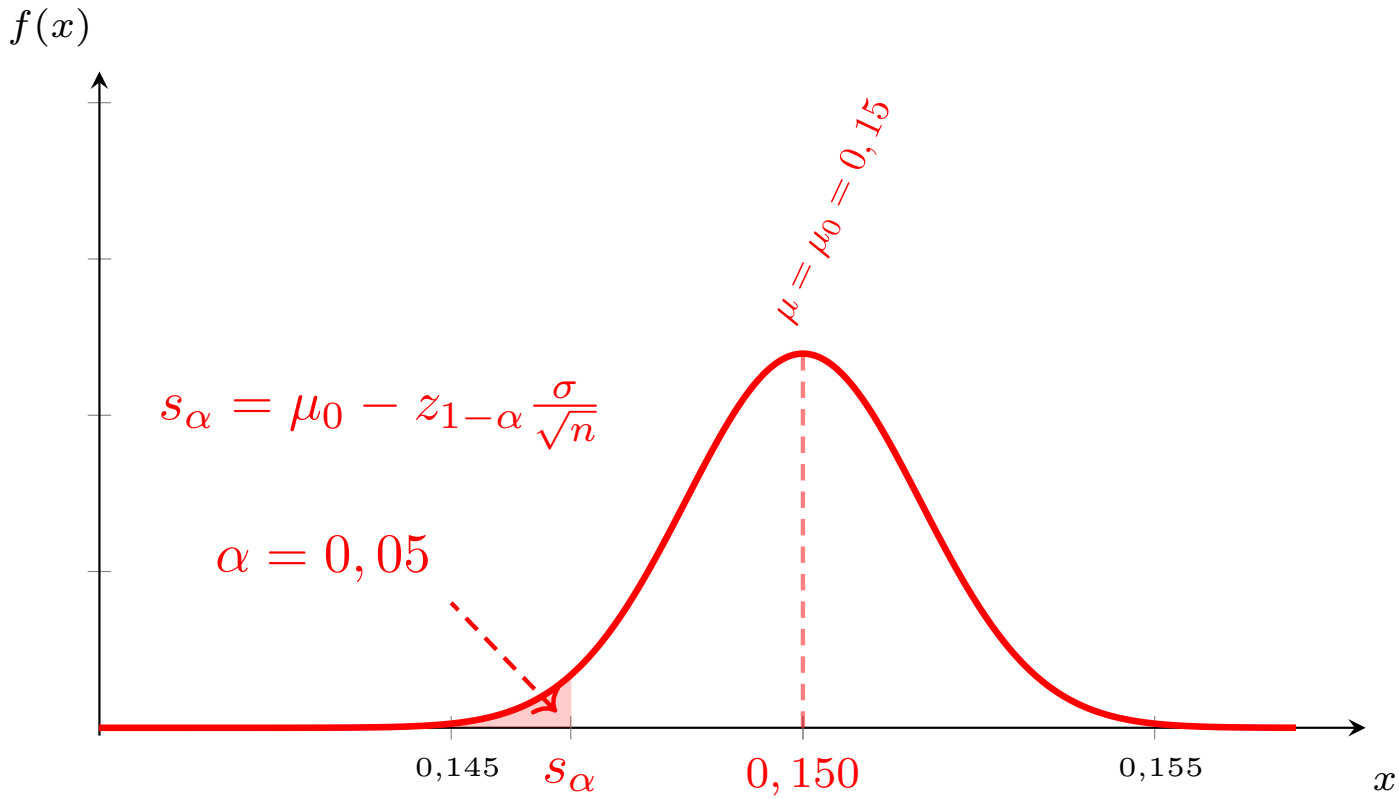
- non rifiutare $H_0 : \mu \geq \mu_0$ in caso contrario.

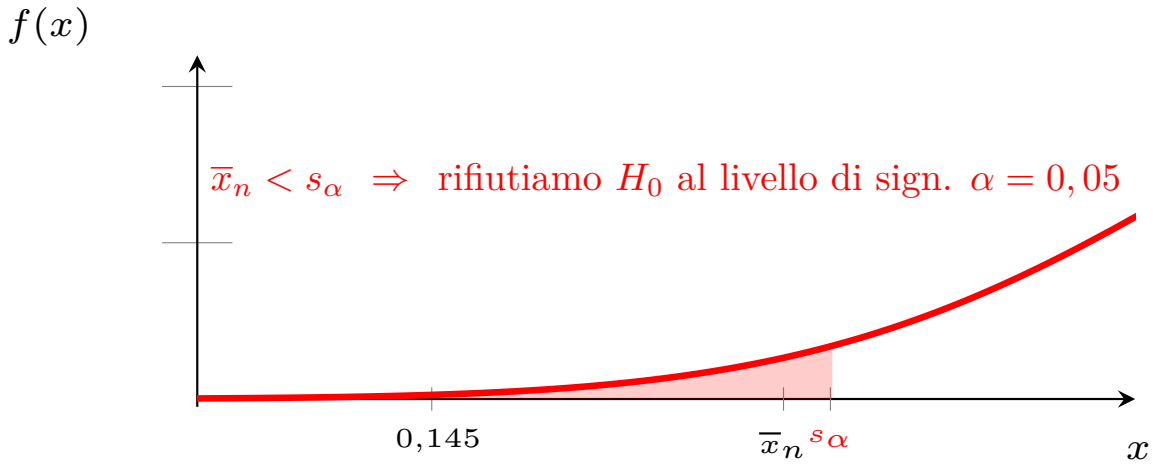
Con riferimento alle bustine zafferano, se fossimo disposti a commettere l'errore di prima specie con probabilità non superiore a $\alpha = 0,05$, ...

... allora la **soglia critica** con cui dovremmo confrontare il valore di $\overline{x}_n = 0,1465$ grammi (si ricordi che questo valore si riferisce ad un campione di $n = 36$ bustine) sarebbe data da

$$s_\alpha = \mu_0 - z_{1-\alpha}\frac{\sigma}{\sqrt{n}} = 0,15 - 1,96 \times \frac{0,01}{\sqrt{36}} = 0,1467.$$

Siccome $\overline{x}_n = 0,1465$ è minore della **soglia critica** $s_\alpha = 0,1467$, dovremmo dunque **"rifiutare l'ipotesi nulla $H_0 : \mu \geq \mu_0$ al livello di significatività $\alpha = 0,05 = 5\%$"**.


$f(x)$
$\mu = \mu_0 = 0,15$
$s_\alpha = \mu_0 - z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$
$\alpha = 0,05$
0,145
s_α
0,150
0,155
x

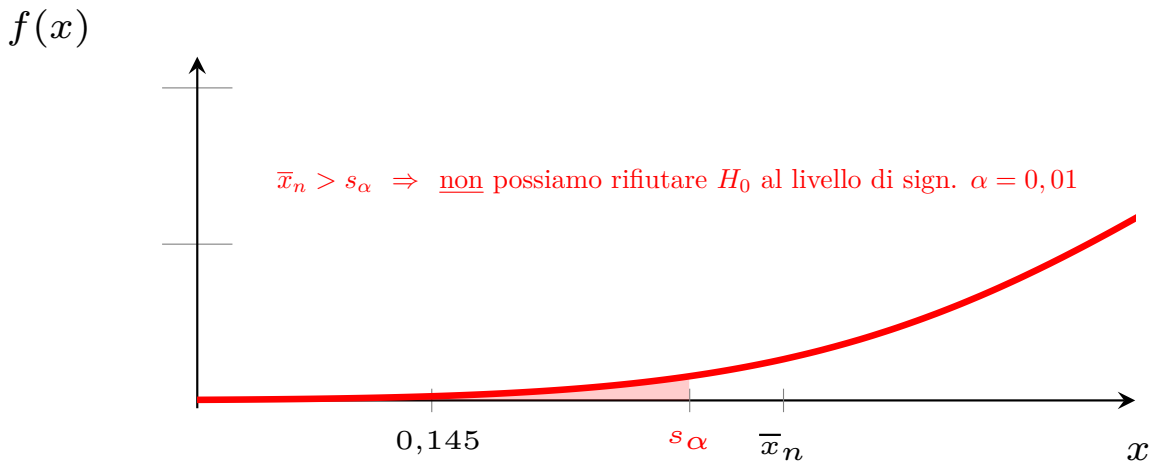

$f(x)$
$\overline{x}_n < s_\alpha \Rightarrow$ rifiutiamo H_0 al livello di sign. $\alpha = 0,05$
0,145
$\overline{x}_n$
s_α
x

Si noti tuttavia che con un livello di significatività più basso (ad esempio $\alpha = 0,01$) non avremmo rifiutato l'ipotesi nulla!!!

Infatti, con $\alpha = 0,01$ avremmo ottenuto la soglia critica

$$s_\alpha = s_{0,01} = \mu_0 - z_{1-\alpha}\frac{\sigma}{\sqrt{n}} = 0,15 - 2,33 \times \frac{0,01}{\sqrt{36}} = 0,1461$$

che è minore di $\overline{x}_n = 0,1465$ e al livello di significatività $\alpha = 0,01$ **non** avremmo dunque rifiutato $H_0 : \mu \geq \mu_0$.

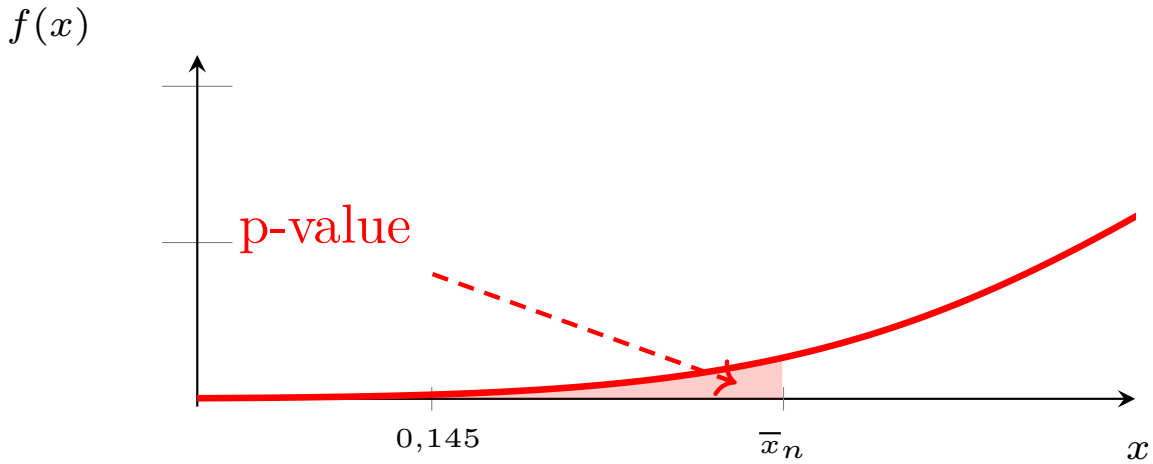


Siccome la decisione di rifiutare o meno l'ipotesi nulla dipende dal livello di significatività del test, i risultati dei test statistici vengono spesso comunicati attraverso il cosiddetto **"p-value"** che può essere definito come **il più piccolo livello di significatività α per il quale l'ipotesi nulla viene rifiutata.**

Con riferimento all'esempio sulle bustine di zafferano, il p-value è dato da (vedi il grafico)

$$\text{p-value} = P\left(\{\overline{X}_n < 0,1465\} \middle| \mu = \mu_0 = 0,15\right) = \Phi\left(\frac{0,1465 - 0,15}{0,01/\sqrt{36}}\right)$$

$$= \Phi(-2,1) = 1 - \Phi(2,1) = 1 - 0,9821 = 0,0179.$$



Si noti che il p-value può essere anche interpretato come la probabilità di ottenere, **attraverso un nuovo campione**, un nuovo valore della statistica test che è **ancora più favorevole all'ipotesi alternativa** del valore che abbiamo ottenuto sulla base del campione che abbiamo già a disposizione.

Si noti che **quanto più è piccolo il p-value, tanto più siamo inclini a pensare che sia vera l'ipotesi alternativa.** Infatti, l'ipotesi nulla viene rifiutata se e solo se

$$\text{p-value} \quad < \quad \text{livello di significatività } \alpha.$$

Finora abbiamo considerato soltanto il caso dove

$$H_0 : \mu \geq \mu_0 \quad \text{e} \quad H_a : \mu < \mu_0.$$

In questo caso si parla di un **"test sulla coda inferiore"** perché la soglia critica e il p-value vengono individuati nella coda inferiore della distribuzione della statistica test (che potrebbe essere $\overline{X}_n$ oppure anche $Z = \frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}}$; si noti che nel secondo caso la soglia critica corrispondente al livello di significatività α è semplicemente $z_{1-\alpha}$).

Chiaramente si può ragionare in modo speculare nel caso di un **"test sulla coda superiore"**, ovvero nel caso in cui le ipotesi statistiche messe a confronto fossero definite come

$$H_0 : \mu \leq \mu_0 \quad \text{e} \quad H_a : \mu > \mu_0.$$

Esercizio 3.1

Un ricercatore sostiene che la durata media di un nuovo tipo di batterie sia superiore a 80 ore. Per dimostrare la sua tesi ha rilevato le durate di un campione di $n = 64$ batterie ottenendo una durata media di $\overline{x} = 82$ ore. Sulla base dell'esperienza passata si può ritenere che lo scarto quadratico medio della durata di una batteria ammonti a 6 ore.

a) Come devono essere definite l'ipotesi nulla e l'ipotesi alternativa di un test statistico per verificare se la durata media delle nuove batterie sia effettivamente maggiore di 80 ore.

Esercizio (continua)

b) Si calcoli la soglia critica che corrisponde al livello di significatività dell'1%. Si può affermare che i dati campionari confermino la tesi del ricercatore al livello di significatività dell'1%?

Esercizio (continua)

c) Si calcoli il p-value del test.

Esercizio (continua)

d) Qual è la probabilità di commettere l'errore di seconda specie se il livello di significatività viene fissato all'1% e se la durata media delle nuove batterie fosse di 81 ore?

In molti casi si considerano anche dei **"test su due code"**, ovvero dei test sulla media dove

$$H_0 : \mu = \mu_0 \quad \text{e} \quad H_a : \mu \neq \mu_0.$$

Anche in questi casi si può fare riferimento alle statistiche test $\overline{X}_n$ oppure (in modo equivalente) $Z = \frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}}$.

L'unica differenza rispetto ai test su una singola coda consiste nel fatto che nei test su due code la soglia critica s_α viene determinata in modo tale che

$$P(\{|\overline{X}_n - \mu_0| > s_\alpha\} | \mu = \mu_0) = \alpha.$$

Il valore di s_α è quindi dato da (lasciamo al lettore il compito di verificare)

$$s_\alpha = z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

e l'ipotesi $H_0 : \mu = \mu_0$ viene dunque rifiutata se e solo se

$$|\overline{x}_n - \mu_0| > z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}},$$

ovvero se solo se $|z| = \left| \frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} \right| > z_{1-\alpha/2}$.

Esercizio 3.2

Un macchinario è stato programmato per tagliare assi di legno della lunghezza di 1 metro. Da un elevato numero di misurazioni effettuate in passato si desume che nel 99% dei casi la lunghezza di un'asse tagliata dal macchinario differisce dalla lunghezza media per meno di $1cm$. Per verificare se le operazioni di taglio stanno procedendo in modo regolare viene esaminato un **campione di $n = 16$ assi tagliate dal macchinario**.

a) Come dovrebbero essere definite le ipotesi di un test statistico per verificare se le operazioni di taglio stanno procedendo in modo regolare?

Esercizio (continua)

b) Volendo fissare il livello di significatività del test all'1%, qual è la soglia critica per la media campionaria oltre la quale bisognerebbe intervenire sul macchinario?

Esercizio (continua)

c) Si supponga che la medie delle lunghezze rilevate nel campione sia di $\overline{x} = 99,98cm$. Si calcoli il p-value associato a questo valore.

Esercizio (continua)

d) Con il p-value trovato al punto precedente, si può rifiutare l'ipotesi nulla al livello di significatività $\alpha = 0,01$?

Test d'ipotesi

Relazione tra la stima intervallare e i test d'ipotesi

Come abbiamo appena visto, nei test dove

$$H_0 : \mu = \mu_0 \quad \text{e} \quad H_a : \mu \neq \mu_0$$

l'ipotesi nulla viene RIFIUTATA se e solo se

$$|\overline{x}_n - \mu_0| > z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

oppure (equivalentemente) se e solo se

$$|z| = \left| \frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}} \right| > z_{1-\alpha/2}.$$

Com'è facile verificare queste condizioni NON sono soddisfatte se e solo se

$$\overline{x}_n - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu_0 < \overline{x}_n + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

ovvero se e solo se **l'IdC** $\overline{x} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$ **contiene il valore di** μ_0 **previsto dall'ipotesi nulla!!!**

Esercizio 3.3

Si consideri ancora l'Esercizio 3.2 sulle assi di legno. Si ricordi che nella consegna veniva ipotizzato che nel 99% dei casi la lunghezza di un'asse tagliata dal macchinario differisse dalla lunghezza media per meno di 1cm. Supponiamo ora che la media delle lunghezze rilevate in un campione di $n = 36$ assi ammonti a $\overline{x} = 99,98cm$.

a) Si calcoli un IdC al 95% per l'ignota lunghezza media μ delle assi tagliate dal macchinario.

Esercizio (continua)

b) Si utilizzi l'IdC del quesito a) onde stabilire se al livello di significatività $\alpha = 0,05$ l'ipotesi statistica $H_0 : \mu = 1$ metro deve essere rifiutata.

Test statistici

Test statistici sulla media μ di una popolazione con σ ignoto

Finora abbiamo parlato soltanto di test statistici sulla media **sotto l'ipotesi che la varianza σ^2 della popolazione sia nota**, ma chiaramente nelle applicazioni questa ipotesi non è quasi mai soddisfatta.

Tuttavia, apportando due piccole modifiche ai metodi che abbiamo visto con riferimento al caso dove la varianza della popolazione è nota, otteniamo dei test analoghi anche per il caso in cui la varianza della popolazione non è nota. Le modifiche in questione sono:

- al posto dell'ignoto valore della varianza σ^2 si sostituisce la sua stima puntuale s_n^2
- al posto del percentile della distribuzione normale standard si sostituisce il corrispondente percentile della distribuzione t di Student con $n-1$ gdl.

Le regole decisionali per i test sulla media di una popolazione con varianza σ^2 ignota sono quindi le seguenti:

- $H_0 : \mu \geq \mu_0$ contro $H_a : \mu < \mu_0$: si rifiuta H_0 (al livello di sign. α) se e solo se

$$\overline{x}_n < \mu_0 - t_{n-1;\alpha}\frac{s}{\sqrt{n}} \quad \text{ovvero se e solo se} \quad t = \frac{\overline{x}_n - \mu_0}{s/\sqrt{n}} < t_{n-1;\alpha};$$

- $H_0 : \mu \leq \mu_0$ contro $H_a : \mu > \mu_0$: si rifiuta H_0 (al livello di sign. α) se e solo se

$$\overline{x}_n > \mu_0 + t_{n-1;\alpha}\frac{s}{\sqrt{n}} \quad \text{ovvero se e solo se} \quad t = \frac{\overline{x}_n - \mu_0}{s/\sqrt{n}} > t_{n-1;\alpha};$$

- $H_0 : \mu = \mu_0$ contro $H_a : \mu \neq \mu_0$: si rifiuta H_0 (al livello di sign. α) se e solo se

$$|\overline{x}_n - \mu_0| > t_{n-1;\alpha/2}\frac{s_n}{\sqrt{n}} \text{ ovvero se e solo se } |t| = \frac{|\overline{x}_n - \mu_0|}{s_n/\sqrt{n}} > t_{n-1;\alpha/2};$$

entrambe queste condizioni sono soddisfatte se e solo se l'IdC con estremi $\overline{x}_n \pm t_{n-1;\alpha/2}\frac{s_n}{\sqrt{n}}$ contiene il valore di μ_0.

Esercizio 3.4

Per stimare l'ignoto consumo medio μ di carburante su un determinato itinerario, uno spedizioniere ha rilevato i consumi di $n = 9$ autoarticolati che lo hanno percorso. La media dei consumi rilevati ammonta a $\overline{x}_n = 120,32$ litri e la deviazione standard campionaria è di $s_n = 6$ litri.

a) Si determini il p-value di un test statistico per verificare se l'ignoto consumo medio è inferiore a $\mu_0 = 130$ litri.

Esercizio (continua)

b) Considerando un livello di significatività del 5%, si può ritenere che il consumo medio sia inferiore a $\mu_0 = 130$ litri?

In realtà, le regole decisionali che abbiamo visto finora sono tutte basate sull'ipotesi che la popolazione di riferimento sia normale.

Applicando le suddette regole decisionali con popolazioni che non sono normali, corriamo il rischio di incorrere nell'errore di prima specie con probabilità più elevata di $1 - \alpha$, ovvero **corriamo il rischio che il livello di significatività effettivo del test sia molto più elevato di quello nominale**.

Per mitigare questo rischio, soprattutto in situazioni dove *a priori* abbiamo fondati sospetti che la popolazione di riferimento sia molto asimmetrica e/o che contenga degli *outliers*, conviene, ove possibile, usare numerosità campionarie elevate e/o fissare valori molto piccoli per il livello di significatività α.

Consideriamo ora invece un test statistico dove le ipotesi messe a confronto vertono sul valore ignoto di una proporzione p. Consideriamo dunque dei test statistici dove

- $H_0 : p \geq p_0$ e $H_a : p < p_0$,
- $H_0 : p \leq p_0$ e $H_a : p > p_0$,

oppure dove

- $H_0 : p = p_0$ e $H_a : p \neq p_0$.

Attraverso ragionamenti simili a quelli che abbiamo già visto per i test statistici sul valore di una media μ, si ottengono le regole decisionali sulla prossima *slide*:

Ipotesi statistiche	Regole decisionali: si rifiuta H_0 (livello sign. α) se ...
$H_0 : p \geq p_0$ vs. $H_a : p < p_0$	$\widehat{p}_n < p_0 - z_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{n}} \quad \Leftrightarrow \quad z = \frac{\widehat{p}_n - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} < -z_{1-\alpha}$
$H_0 : p \leq p_0$ vs. $H_a : p > p_0$	$\widehat{p}_n > p_0 + z_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{n}} \quad \Leftrightarrow \quad z = \frac{\widehat{p}_n - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} > z_{1-\alpha}$
$H_0 : p = p_0$ vs. $H_a : p \neq p_0$	$\lvert\widehat{p}_n - p_0\rvert > z_{1-\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}} \quad \Leftrightarrow \quad \lvert z\rvert = \frac{\lvert\widehat{p}_n - p_0\rvert}{\sqrt{\frac{p_0(1-p_0)}{n}}} > z_{1-\alpha/2}$

Esercizio 3.5

Secondo l'esito di un sondaggio basato su un campione di $n = 625$ elettori, alle prossime elezioni il 22% degli elettori voterebbe il partito XXX.

a) Sulla base di questo sondaggio, si può affermare che più del 20% della popolazione di tutti gli elettori voterebbe il partito XXX. Si risponda calcolando il *p-value* di un opportuno test statistico.

Esercizio (continua)

b) Considerando $\alpha = 0,01$ come livello di significatività, si dovrebbe rifiutare l'ipotesi nulla del test statistico del punto precedente?

Confronto tra medie

Il caso di campioni indipendenti con varianze note a priori

Supponiamo ora di essere interessati alla **differenza** $\mu_1 - \mu_2$ **tra le medie di due popolazioni**.

Per fare un esempio concreto, consideriamo il caso dove μ_1 è la durata media delle batterie di un certo tipo e dove μ_2 è la durata media delle batterie di un altro tipo.

Per semplicità **assumeremo inizialmente che le varianze** σ_1^2 **e** σ_2^2 delle durate dei due tipi di batterie siano **note a priori**.

Chiaramente, se disponiamo di un **campione casuale per ciascuno dei due tipi di batterie** (le numerosità dei due campioni verranno indicate con n_1 e n_2), possiamo stimare l'ignoto valore della differenza $\mu_1 - \mu_2$ attraverso la **differenza tra le due medie campionarie:**

$$\overline{X}_1 - \overline{X}_2.$$

dove

$$\overline{X}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} \widetilde{X}_{1,i} \quad \text{e} \quad \overline{X}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} \widetilde{X}_{2,i}.$$

[[$\widetilde{X}_{1,i}$ sono le variabili casuali campionarie del primo campione e $\widetilde{X}_{1,i}$ sono quelle del secondo campione. Inoltre, in queste *slides* il pedice della media campionaria $\overline{X}$ identifica il campione e non più la numerosità campionaria.]]

Supponiamo ora che **i due campioni siano stati selezionati in modo indipendente**. In questo caso possiamo concludere che la deviazione standard di $\overline{X}_1 - \overline{X}_2$ sia data da

$$\sigma_{\overline{x}_1 - \overline{x}_2} = \sqrt{var(\overline{X}_1 - \overline{X}_2)} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad \ldots$$

. . . e siccome nel contesto in questione è plausibile che le distribuzioni delle durate siano normali per entrambi i tipi di batterie (**ipotesi di normalità delle popolazioni**), possiamo anche concludere che

$$\overline{X}_1 - \overline{X}_2 \sim Normale\left(\mu_{\overline{x}_1 - \overline{x}_2} = \mu_1 - \mu_2 = ???; \quad \sigma_{\overline{x}_1 - \overline{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right).$$

Sotto le suddette ipotesi possiamo dunque associare alla stima puntuale $\overline{x}_1 - \overline{x}_2$ un margine d'errore con livello di confidenza $1 - \alpha$ che è dato da

$$m_{1-\alpha} = z_{1-\alpha/2}\sigma_{\overline{x}_1-\overline{x}_2} = z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Aggiungendo e sottraendo questo margine d'errore al valore della stima puntuale $\overline{x}_1 - \overline{x}_2$, otteniamo dunque gli estremi di un IdC per $\mu_1 - \mu_2$:

$$\overline{x}_1 - \overline{x}_2 \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Supponendo che

- i due campioni siano composti da $n_1 = 10$ batterie del primo tipo e da $n_2 = 5$ batterie del secondo tipo,
- che le medie campionarie siano date da $\overline{x}_1 = 94,3$ ore e $\overline{x}_2 = 90,1$ ore
- che le varianze delle due popolazioni siano date da $\sigma_1^2 = 6^2 = 36$ e $\sigma_2^2 = 5^2 = 25$;

e ponendo il livello di confidenza uguale a $1 - \alpha = 0,98 = 98\%$, si ottiene l'IdC con estremi dati da

$$\overline{x}_1 - \overline{x}_2 \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = 94,3 - 90,1 \pm 2,33\sqrt{\frac{36}{10} + \frac{25}{5}} = \begin{cases} 11,03 \\ -2,63 \end{cases}$$

Chiaramente, a partire dalla distribuzione

$$\overline{X}_1 - \overline{X}_2 \sim \textit{Normale}\left(\mu_{\overline{x}_1-\overline{x}_2} = \mu_1 - \mu_2 = ???;\quad \sigma_{\overline{x}_1-\overline{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

si possono anche definire regole decisionali per test statistici che vertono sul valore della differenza $\mu_1 - \mu_2$ (vedi la prossima *slide*):

Ipotesi statistiche	Regole decisionali: si rifiuta H_0 (livello sign. α) se
$H_0 : \mu_1 - \mu_2 \geq D_0$ vs. $H_a : \mu_1 - \mu_2 < D_0$	$\overline{x}_1 - \overline{x}_2 < D_0 - z_{1-\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \Leftrightarrow$ $\Leftrightarrow \quad z = \frac{\overline{x}_1 - \overline{x}_2 - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < -z_{1-\alpha}$
$H_0 : \mu_1 - \mu_2 \leq D_0$ vs. $H_a : \mu_1 - \mu_2 > D_0$	$\overline{x}_1 - \overline{x}_2 > D_0 + z_{1-\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \Leftrightarrow$ $\Leftrightarrow \quad z = \frac{\overline{x}_1 - \overline{x}_2 - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > z_{1-\alpha}$
$H_0 : \mu_1 - \mu_2 = D_0$ vs. $H_a : \mu_1 - \mu_2 \neq D_0$	$\lvert\overline{x}_1 - \overline{x}_2 - D_0\rvert > z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \Leftrightarrow$ $\Leftrightarrow \quad z = \frac{\lvert\overline{x}_1 - \overline{x}_2 - D_0\rvert}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > z_{1-\alpha/2}$

Rimarchiamo che gli IdC e i test sulla differenza $\mu_1 - \mu_2$ che abbiamo appena visto sono sempre **"esatti"** (nel senso che per qualsiasi numerosità campionarie n_1 e n_2 danno luogo a livelli di confidenza/livelli di significatività effettivi che coincidono esattamente con quelli nominali) se

- i due campioni vengono selezionati in modo **indipendente**
- e le due popolazioni di provenienza sono entrambe **normali**.

Sebbene dal punto di vista teorico si possono escogitare condizioni sufficienti diverse sotto le quali i suddetti IdC/test sono comunque esatti, ...

... dal punto di vista pratico delle applicazioni queste condizioni alternative non sono mai realistiche. ...

. . . Tuttavia, conviene rimarcare che gli IdC/test sulla differenza $\mu_1 - \mu_2$ che abbiamo presentato poco sopra **sono validi in senso asintotico** se è soddisfatta soltanto l'ipotesi di indipendenza (anche se una o entrambe le popolazioni di provenienza non sono normali). . . .

. . . In questo caso il loro livello di confidenza/livello di significatività effettivo sarà prossimo a quello nominale se la numerosità di **entrambi** i campioni (oppure soltanto quella del campione che proviene dalla popolazione non normale) sono sufficientemente elevate. . . .

. . . **Nel caso in cui soltanto l'ipotesi di normalità fosse in dubbio**, si può determinare una **soglia minima per la numerosità di ciascuno dei due campioni** applicando le seguenti regole:[2]

- se a priori riteniamo che la deviazione rispetto all'ipotesi di normalità sia soltanto lieve, è sufficiente considerare un campione/due campioni con numerosità maggiore di 10 (circa);
- se a priori sospettiamo che una o entrambe le popolazioni siano asimmetriche e/o che contengano degli *outliers*, allora la numerosità del campione di riferimento dovrebbe essere almeno pari a 30 (se pensiamo che l'asimmetria sia molto pronunciata questa soglia dovrebbe essere innalzata a 50)

[2] Le regole sono del tutto analoghe a quelle che abbiamo già visto con riferimento agli IdC per la media di una popolazione.

Esercizio 4.1

Un ristoratore vuole stimare la differenza tra l'incasso medio nei giorni feriali e in quelli festivi. A tal fine considera dati ricavati da un campione di giorni feriali e da un secondo campione di giorni festivi:

numerosità campionarie	incassi medi
$n_1 = 36$ giorni feriali	$\overline{x}_1 = 3650$ euro
$n_2 = 12$ giorni festivi	$\overline{x}_2 = 4125$ euro

Basandosi su altre indagini simili, il ristoratore ritiene che la varianza degli incassi nei giorni feriali sia data da $\sigma_1^2 = 2500$ (euro^2) e che la varianza degli incassi nei giorni festivi sia invece data da $\sigma_2^2 = 6400$ (euro^2).

a) Si calcoli un IdC al 95% per l'ignota differenza tra l'incasso medio nei giorni feriali e in quelli festivi.

Esercizio (continua)

b) Si stabilisca, calcolando il p-value di un opportuno test statistico, se i dati possono essere considerati come evidenza empirica a favore dell'ipotesi che l'incasso medio nei giorni festivi sia maggiore.

Confronto tra medie

Il caso di due campioni indipendenti con varianze ignote

Fino ad ora abbiamo ipotizzato che le varianze σ_1^2 e σ_2^2 delle due popolazioni fossero note, ma nelle applicazioni questo non è quasi mai il caso.

Per adattare le suddette procedure al caso in cui le varianze delle due popolazioni non sono note, occorre effettuare due cambiamenti:

- sostituire le stime puntuali s_1^2 e s_2^2 al posto di σ_1^2 e σ_2^2;[3]
- sostituire al posto del percentile della distribuzione normale standard il corrispondente percentile della distribuzione t di Student con

$$gdl = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}.$$

[3] s_1^2 e s_2^2 sono le varianze campionarie corrette che si ottengono a partire dai due campioni.

Le procedure adattate al caso in cui le varianze delle popolazioni non sono note (vedi la *slide* precedente) **non danno luogo a livelli di confidenza/livelli di significatività effettivi che sono esattamente uguali a quelli nominali nemmeno se entrambe le popolazioni sono normali**.

Per prevenire marcate differenze tra livelli di confidenza/livelli di significatività effettivi e nominali occorre seguire le seguenti raccomandazioni:

- se a priori si è convinti che entrambe le popolazioni siano approssimativamente normali, è sufficiente assicurarsi che entrambe le numerosità campionarie siano almeno pari a 10;
- se si pensa che una o entrambe le popolazioni siano asimmetriche e/o che presentino degli *outliers* occorre assicurarsi che entrambe le numerosità campionarie siano almeno pari a 30 (questa soglia dovrebbe essere innalzata a 50 se a priori si è convinti che l'asimmetria sia molto pronunciata e/o che ci siano degli *outliers* molto lontani dalla grande massa di valori);
- se possibile, utilizzare numerosità campionarie simili.

Esercizio 4.2

Un ristoratore vuole stimare la differenza tra l'incasso medio nei giorni feriali e in quelli festivi. A tal fine considera dati ricavati da un campione di giorni feriali e da un secondo campione di giorni festivi:

numerosità campionarie	totale incassi	somma dei quadrati
$n_1 = 36$ giorni feriali	$\sum_{i=1}^{36} x_{1i} = 131256.60$ euro	$\sum_{i=1}^{36} x_{1i}^2 = 478648029$
$n_2 = 12$ giorni festivi	$\sum_{i=1}^{12} x_{2i} = 49554.40$ euro	$\sum_{i=1}^{12} x_{2i}^2 = 204725581$

a) Si calcoli un IdC al 95% per l'ignota differenza tra l'incasso medio nei giorni feriali e in quelli festivi.

Esercizio (continua)

b) Si stabilisca, calcolando il p-value di un opportuno test statistico, se i dati possono essere considerati come evidenza empirica a favore dell'ipotesi che la differenza tra l'incasso medio nei giorni festivi e in quelli feriali sia superiore a 400 euro.

Confronto tra medie

Il caso di due campioni indipendenti con varianze ignote ma uguali

A volte, quando le varianze di due popolazioni sono ignote, ricorrono **condizioni che rendono plausibile l'ipotesi che le due varianze siano uguagli:** $\sigma_1^2 = \sigma_2^2 = ???$.

In questi casi, il valore comune delle due varianze può essere stimato attraverso la **varianza campionaria congiunta**, ovvero attraverso la stima puntuale

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 1}.$$

Se le ignote varianze delle due popolazioni sono effettivamente uguali, lo stimatore S_p^2 fornisce (con probabilità elevata) una stima più precisa di quella che si ottiene da ciascuno dei due stimatori S_1^2 e S_2^2 preso singolarmente (infatti, S_1^2 e S_2^2 tengono conto soltanto delle informazioni contenute in uno dei due campioni).

Sulla base dello stimatore congiunto S_p^2 si possono costruire intervalli di confidenza e test statistici alternativi a quelli che abbiamo presentato poco

Gli estremi di un IdC alternativo che può essere utilizzato quando a priori si è convinti che le varianze delle due popolazioni siano uguali, sono dati da

$$\overline{x}_1 - \overline{x}_2 \pm t_{n_1+n_2-1;\alpha/2} \times s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

e le regole decisionali per i corrispondenti test d'ipotesi sono riportate nella prossima *slide*:

Ipotesi statistiche	Regole decisionali: si rifiuta H_0 (livello sign. α) se
$H_0 : \mu_1 - \mu_2 \geq D_0$ vs. $H_a : \mu_1 - \mu_2 < D_0$	$\overline{x}_1 - \overline{x}_2 < D_0 - t_{n_1+n_2-1;\alpha}\ s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \Leftrightarrow$ $\Leftrightarrow \quad t = \frac{\overline{x}_1-\overline{x}_2-D_0}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} < -t_{n_1+n_2-1;\alpha}$
$H_0 : \mu_1 - \mu_2 \leq D_0$ vs. $H_a : \mu_1 - \mu_2 > D_0$	$\overline{x}_1 - \overline{x}_2 > D_0 + t_{n_1+n_2-1;\alpha}\ s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \Leftrightarrow$ $\Leftrightarrow \quad t = \frac{\overline{x}_1-\overline{x}_2-D_0}{s_p\sqrt{\frac{1}{n_1}+\frac{t}{n_2}}} > t_{n_1+n_2-1;\alpha}$
$H_0 : \mu_1 - \mu_2 = D_0$ vs. $H_a : \mu_1 - \mu_2 \neq D_0$	$\lvert\overline{x}_1 - \overline{x}_2 - D_0\rvert > t_{n_1+n_2-1;\alpha/2}\ s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \Leftrightarrow$ $\Leftrightarrow \quad t = \frac{\lvert\overline{x}_1-\overline{x}_2-D_0\rvert}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} > t_{n_1+n_2-1;\alpha/2}$

Le regole decisionali in questa tabella dovrebbero essere utilizzate soltanto quando a priori si è sicuri che le varianze delle due popolazioni siano uguali (o molto simili)!!!

Esercizio 4.3

Per verificare se grazie ad un nuovo tipo di serramenti si ottiene effettivamente un risparmio energetico sono stati esaminati due campioni di famiglie che vivono in appartamenti con caratteristiche simili:

- Il primo campione è composto da 32 famiglie che vivono in appartamenti con serramenti di tipo tradizionale. Per queste famiglie è stato rilevato, nell'arco di un anno, un consumo medio di gas metano di $\overline{x}_1 = 1154 smc$ con una deviazione standard campionaria di $s_1 = 30 smc$.
- Il secondo campione è composto da 36 famiglie che vivono in appartamenti con serramenti del nuovo tipo e per queste famiglie è stato rilevato, nell'arco dello stesso anno, un consumo medio di gas metano di $\overline{x}_2 = 1045 smc$ con una deviazione standard campionaria di $s_2 = 34 smc$.

Esercizio (continua)

a) Sulla base di questi dati campionari si può ritenere che vi sia evidenza empirica a favore dell'ipotesi che con il nuovo tipo di serramenti si risparmiano più di 100*smc* all'anno? Si risponda calcolando il p-value di un opportuno test statistico.

Esercizio (continua)

b) Si calcoli anche il p-value del test statistico basato sull'ipotesi che le varianze delle due popolazioni siano uguali.

I metodi che abbiamo visto finora si riferiscono al caso in cui i due campioni sono **indipendenti** e quindi non dovrebbero essere applicati nel caso di **"campioni appaiati"**.

Per **"campione appaiato"** intendiamo un campione di n coppie di valori

$$(x_{11}, x_{21}), \quad (x_{12}, x_{22}), \quad \ldots, \quad (x_{1i}, x_{2i}), \quad \ldots, \quad (x_{1n}, x_{2n})$$

che sono state generate in modo indipendente, ...

... **ma dove sospettiamo che tra i due valori di ciascuna coppia** (x_{1i}, x_{2i}) **ci sia una certa dipendenza!!!**

Campioni appaiati si incontrano tipicamente nei contesti dove si vuole studiare l'**effetto di un trattamento** (una dieta, una cura, ecc.) ...

...e a tal fine si considerano n individui per i quali si rileva una **variabile risposta** X (il peso, lo stato di salute, ecc.) prima e dopo il trattamento.

Per fare un esempio concreto possiamo considerare un campione di 9 individui che si sono sottoposti ad una determinata dieta. Supponiamo che per ciascuno di questi individui sia stato rilevato il peso X (in kg) prima dell'inizio della dieta e dopo un mese di dieta:

individuo	x_{1i} =peso prima della dieta	x_{2i} =peso dopo un mese di dieta
1	85	79
2	91	87
3	99	91
4	88	81
5	105	99
6	93	86
7	89	87
8	94	84
9	91	85

Chiaramente, se disponiamo di un campione appaiato

$$(x_{11}, x_{21}), \quad (x_{12}, x_{22}), \quad \ldots, \quad (x_{1i}, x_{2i}), \quad \ldots, \quad (x_{1n}, x_{2n})$$

possiamo calcolare le differenze

$$d_i = x_{1i} - x_{2i}, \quad i = 1, 2, \ldots, n,$$

e la loro media campionaria

$$\overline{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1}{n}\sum_{i=1}^{n}(x_{1i} - x_{2i}) = \overline{x}_1 - \overline{x}_2$$

onde ottenere una **stima puntuale per l'ignota differenza tra la media della variabile risposta prima e dopo il trattamento**, differenza che indicheremo con $\mu_d = \mu_1 - \mu_2$ (chiaramente μ_1, μ_2 e μ_d si riferiscono alla popolazione e non al campione appaiato).

Nell'esempio sulla dieta otteniamo le differenze

individuo	x_{1i}	x_{2i}	$d_i = x_{1i} - x_{2i}$
1	85	79	6
2	91	87	4
3	99	91	8
4	88	81	7
5	105	99	6
6	93	86	7
7	89	87	2
8	94	84	10
9	91	85	6
Tot	835	779	56

e usando i totali in fondo alle colonne possiamo calcolare i valori di

$$\overline{x}_1 = \frac{835}{9} = 92,78, \quad \overline{x}_2 = \frac{779}{9} = 86,56 \quad \text{e di} \quad \overline{d} = \frac{56}{9} = 6,22$$

(ovviamente, il valore di $\overline{d}$ deve coincidere con $\overline{x}_1 - \overline{x}_2 = 92,78 - 86,56 = 6,22$).

A questo punto, onde stabilire se la dieta (ovvero il trattamento) sia efficace o meno, ...

...ovvero onde verificare se in seguito al trattamento l'ignota media della variabile risposta X diminuisce (o aumenta), ...

...**possiamo calcolare degli IdC e/o eseguire test statistici sull'ignoto valore di $\mu_d = \mu_1 - \mu_2$ applicando ai valori delle differenze $d_i = x_{1i} - x_{2i}$ esattamente le stesse procedure che abbiamo già visto con riferimento alla media μ di una singola popolazione!!!**

Per esempio, usando le differenze $d_i = x_{1i} - x_{2i}$ tra i pesi rilevati prima dell'inizio della dieta e dopo un mese di dieta, otteniamo (vedi la tabella con i dati sulla dieta)

$$\sum_{i=1}^{9} d_i^2 = 6^2 + 4^2 + 8^2 + 7^2 + 6^2 + 7^2 + 2^2 + 10^2 + 6^2 = 390 \quad \Rightarrow$$

$$\Rightarrow \quad s_d = \sqrt{\frac{1}{9-1}\left(390 - 9 \times 6,22^2\right)} = 2,29$$

(s_d è la deviazione standard campionaria calcolata sulle differenze $d_i = x_{1i} - x_{2i}$; si ricordi che $\overline{d} = 6,22$ è la media campionaria delle differenze d_i) ...

. . . e usando questo risultato possiamo (per esempio) calcolare un IdC al 99% per l'ignoto valore di $\mu_d = \mu_1 - \mu_2$:

$$\overline{d} \pm t_{n-1;0,005} \frac{s_d}{\sqrt{n}} = 6,22 \pm 3,355 \times \frac{2,29}{\sqrt{9}} = \begin{cases} 8,78 \\ 3,66 \end{cases}$$

Con livello di confidenza pari al 99% possiamo dunque affermare che l'ignota media della perdita di peso dopo un mese di dieta sia compresa tra $3,66kg$ e $8,78kg$.

Volendo, possiamo anche effettuare un **test statistico** onde verificare se il valore atteso della perdita di peso sia maggiore di una determinata soglia minima, per esempio di $4kg$:

$$H_0 : \mu_d \leq 4kg \quad \text{vs} \quad H_a : \mu_d > 4kg.$$

Per determinare il p-value di questo test statistico dobbiamo confrontare il valore della statistica test

$$t = \frac{\overline{d} - 4}{s_d/\sqrt{n}} = \frac{6,22 - 4}{2,29/\sqrt{9}} = 2,908$$

con i percentili della distribuzione t di Student con $n - 1 = 9 - 1 = 8$ gdl: siccome

$$t_{8;0,01} = 2,806 < t = 2,908 < t_{8;0,005} = 3,355$$

possiamo concludere che il p-value sia compreso tra $0,005$ e $0,01$.

Al livello di significatività $\alpha = 0,005$ non possiamo dunque rifiutare $H_0 : \mu_d \leq 4kg$, ma al livello di significatività $\alpha = 0,01$ l'ipotesi nulla $H_0 : \mu_d \leq 4kg$ viene rifiutata.

Esercizio 4.4

Per verificare se la seconda dose di un vaccino è efficace, è stato selezionato un campione di $n = 10$ individui per i quali è stato rilevato il numero di anticorpi prima e dopo la somministrazione della seconda dose. Per ciascun individuo la tabella sottostante riporta il numero di anticorpi prima (x_{1i}) e dopo (x_{2i}) la somministrazione della seconda dose del vaccino, nonché il valore della differenza $d_i = x_{1i} - x_{2i}$:

Individuo	x_{1i}	x_{2i}	d_i	Individuo	x_{1i}	x_{2i}	d_i
1	12530	16032	-3502	6	17535	22653	-5118
2	14632	18535	-3903	7	17043	25854	-8811
3	15973	16075	-102	8	12696	14378	-1682
4	19534	32404	-12870	9	8964	10754	-1790
5	9532	15694	-6162	10	7562	12690	-5128

[[con i dati della tabella si ottiene $\overline{d} = -4906.8$ e $s_d = 3749.15$]]

Si calcoli il p-value di un test statistico per verificare se la seconda dose del vaccino aumenta il numero medio di anticorpi. Si può rifiutare l'ipotesi nulla al livello di significatività $\alpha = 0,01$?

Il termine **analisi della varianza** (**ANOVA** - acronimo di "Analysis of Variance") viene utilizzato per indicare determinate analisi statistiche che hanno lo scopo di **indagare sulle cause della variabilità**.

Queste analisi sono basate su **scomposizioni della varianza** (o della devianza), da cui il loro nome.

Qui di seguito descriveremo come si può applicare l'ANOVA nell'ambito di **studi sperimentali** dove si vogliono **confrontare gli effetti di** k **trattamenti alternativi**.

Di solito, nei studi sperimentali si parte da un campione di un certo numero n di unità statistiche e si considera un **"disegno sperimentale completamente randomizzato"**, ovvero si procede all'assegnazione **casuale** di uno di k trattamenti a ciascuna unità statistica del campione.

In questo modo si ottengono k gruppi con

- n_1 unità statistiche alle quali viene assegnato il primo trattamento,
- n_2 unità statistiche alle quali viene assegnato il secondo trattamento,
- ...
- n_k unità statistiche alle quali viene assegnato il k-esimo trattamento.

Per l'assegnazione casuale dei trattamenti, solitamente si utilizza un metodo che assicura di ottenere k **gruppi di numerosità identica o quantomeno molto simile**, anche se a seconda delle circostanze potrebbe essere vantaggioso utilizzare metodi che diano luogo a gruppi di numerosità diversa.

Il motivo che suggerisce di adottare un **"disegno sperimentale completamente randomizzato"** piuttosto che procedere all'assegnazione dei trattamenti sulla base di **criteri predeterminati** è presto detto:

. . . Assegnando i trattamenti secondo **criteri predeterminati** rischiamo di concentrare unità statistiche con caratteristiche (forse anche ignote) molto omogenee nello stesso gruppo e di ottenere quindi k gruppi molto eterogenei. Le **differenze che si riscontrano nelle risposte ai** k **trattamenti** potrebbero quindi non più essere dovute ai trattamenti, ma al fatto che le unità statistiche di gruppi di diversi siano molto eterogenee.

Sulla base dei valori osservati per una **"variabile risposta"** (e/o delle medie parziali che si riferiscono ai k gruppi) non sarebbe quindi più possibile distinguere se l'efficacia dei k trattamenti è veramente diversa oppure se le differenze riscontrate sono in realtà dovute soltanto al modo in cui sono stati formati i gruppi.

Supponiamo dunque di assegnare a ciascuna delle n unità statistiche di un campione uno di k trattamenti alternativi e di osservare per ciascuna di queste n unità statistiche il valore x_{ji} che assume una variabile risposta X.

trattamento	valori osservati della **variabile risposta**
1	$x_{11}, x_{12}, \ldots, x_{1n_1}$
2	$x_{21}, x_{22}, \ldots, x_{2n_2}$
$\vdots$	$\ldots$
k	$x_{k1}, x_{k2}, \ldots, x_{kn_k}$

Si noti che le osservazioni x_{ji} della variabile risposta X sono identificate da **due pedici**:

- il pedice j che identifica uno dei k trattamenti (ovvero il gruppo)
- e il pedice i che identifica l'unità statistica all'interno del gruppo di appartenenza.

A questo punto ci chiediamo:

. . . sulla base delle osservazioni x_{ji} della variabile risposta X, possiamo ritenere che l'efficacia dei k trattamenti sia la medesima oppure dobbiamo rifiutare questa ipotesi???

Per rispondere a questa domanda possiamo effettuare un test statistico con

$$H_0: \quad \mu_1 = \mu_2 = \cdots = \mu_k$$
$$H_a: \quad \text{almeno una delle } k \text{ medie } \mu_j \text{ è diversa dalle altre.}$$

dove

- μ_1 è la media della variabile risposta dopo il primo trattamento,
- μ_2 è la media della variabile risposta dopo il secondo trattamento,
- . . . ,
- μ_k è la media della variabile risposta dopo il k-esimo trattamento.

Onde definire una **regola decisionale** per il suddetto test, di solito si utilizzano i risultati di un'opportuna **scomposizione della varianza (o devianza)**.

Nel caso in questione, la scomposizione consiste semplicemente nel calcolo della **devianza fra e nei gruppi**.

Infatti, come vedremo tra breve, attraverso il confronto di queste due devianze si ottengono regole decisionali semplici e intuitive.

Per ricordare le formule definitorie della devianza "fra" e "nei" gruppi, indicheremo con

$$\overline{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i}, \quad \overline{x}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{2i}, \quad \ldots, \quad \overline{x}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ki}$$

le **medie parziali** della variabile risposta che si riferiscono a ciascuno dei k gruppi (o trattamenti), e con

$$\overline{\overline{x}} = \frac{1}{n_1 + n_2 + \cdots + n_k}\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ji} = \frac{1}{n}\sum_{j=1}^{k} \overline{x}_j n_j$$

la **media complessiva** di tutti i valori x_{ji} del campione.

Così facendo possiamo esprimere la **devianza fra gruppi** come

$$dev_{fra}(x_{ji}) = \sum_{j=1}^{k} (\overline{x}_j - \overline{\overline{x}})^2 n_j$$

e la **devianza nei gruppi** come

$$dev_{nei}(x_{ji}) = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ji} - \overline{x}_j)^2.$$

Si ricordi che nel corso di STATISTICA DESCRITTIVA è stato dimostrato che (la **"scomposizione della devianza"**)

$$dev_{tot}(x_{ji}) = dev_{fra}(x_{ji}) + dev_{nei}(x_{ji}),$$

dove

$$dev_{tot}(x_{ji}) = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ji} - \overline{\overline{x}})^2$$

è la **devianza totale** di tutte le osservazioni x_{ji}.

Per ricordare le procedure di calcolo che conducono alla devianza "fra" e "nei" gruppi, consideriamo un esempio dove

- i k trattamenti consistono nella somministrazione di uno di k vaccini alternativi,
- e dove le osservazioni x_{ji} della variabile risposta X rappresentano il numero di anticorpi sviluppati un mese dopo la somministrazione.

vaccino	numero di anticorpi (in migliaia)				
1	30,439,	27,834,	37,654,	31,085	
2	27,628,	24,578,	20,700,	28,944,	24,713
3	26,360,	26,565,	28,355		

Con i dati sui tre vaccini otteniamo le medie parziali

$$\overline{x}_1 = \frac{1}{4}(30,439 + 27,834 + 37,654 + 31,085) = 31,753$$
$$\overline{x}_2 = \frac{1}{5}(27,628 + 24,578 + 20,700 + 28,944 + 24,713) = 25,313$$
$$\overline{x}_3 = \frac{1}{3}(26,360 + 26,565 + 28,355) = 27,093$$

e la media totale

$$\overline{\overline{x}} = \frac{1}{12}(31,753 \times 4 + 25,313 \times 5 + 27,093 \times 3) = 27,905.$$

La **devianza fra gruppi** è quindi data da è quindi data da

$$dev_{fra}(x_{ji}) = (31,753 - 27,905)^2 \times 4 + (25,313 - 27,905)^2 \times 5 + \ldots$$
$$\cdots + (27,093 - 27,905)^2 \times 3 = 94,808,$$

mentre . . .

. . . per ottenere la **devianza nei gruppi** dobbiamo prima calcolare le $k = 3$ **devianze parziali** che si riferiscono ai singoli gruppi (trattamenti) e poi sommarle. Per il primo gruppo (primo trattamento) otteniamo

$$dev_{tr1}(x_{ji}) = \sum_{i=1}^{n_1}(x_{1i} - \overline{x}_1)^2 = [[\text{ formula indiretta }]] =$$

$$= 30,439^2 + 27,834^2 + 37,654^2 + 31,085^2 - 4 \times 31,753^2 = 52,353$$

e per gli altri due gruppi (trattamenti) otteniamo

$$dev_{tr2}(x_{ji}) = 27,628^2 + 24,578^2 + 20,700^2 + 28,944^2 + 24,713^2 - 5 \times 25,313^2$$
$$= 40,622$$

e

$$dev_{tr3}(x_{ji}) = 26,360^2 + 26,565^2 + 28,355^2 - 3 \times 27,093^2 = 2,463.$$

La **devianza nei gruppi** è quindi data da

$$dev_{nei}(x_{ji}) = \sum_{j=1}^{3} \sum_{i=1}^{n_j} (x_{ji} - \overline{x}_j)^2 = \sum_{j=1}^{3} dev_{trj}(x_{ji}) =$$

$$= 52,353 + 40,622 + 2,463 = 95,438.$$

Sommando $dev_{fra}(x_{ji}) = 94,808$ e $dev_{nei}(x_{ji}) = 95,438$ dovremmo quindi ottenere

$$dev_{tot}(x_{ji}) = 94,808 + 95,438 = 190,246.$$

Per verificare questo risultato, basta calcolare la devianza di tutti di valori x_{ji} senza tener conto della suddivisione in gruppi. Usando la formula indiretta della devianza, otteniamo

$$\begin{aligned} dev_{tot}(x_{ji}) &= 30,439^2 + 27,834^2 + 37,654^2 + 31,085^2 + 27,628^2 + 24,578^2 - \\ &\quad + 20,700^2 + 28,944^2 + 24,713^2 + 26,360^2 + 26,565^2 + \\ &\quad + 28,355^2 - 12 \times 27,905^2 \\ &= 190,014. \end{aligned}$$

Come previsto, questo risultato coincide quasi esattamente con il valore di $dev_{tot}(x_{ji}) = 190,246$ che abbiamo ottenuto come somma tra $dev_{fra}(x_{ji}) = 94,808$ e $dev_{nei}(x_{ji}) = 95,438$ (la differenza è dovuta soltanto agli arrotondamenti).

Torniamo ora al test statistico con

$$\begin{aligned} H_0: &\quad \mu_1 = \mu_2 = \cdots = \mu_k \\ H_a: &\quad \text{almeno una delle } k \text{ medie } \mu_j \text{ è diversa dalle altre.} \end{aligned}$$

Come abbiamo già accennato, per definire una regola decisionale possiamo confrontare i valori $dev_{fra}(x_{ji})$ e $dev_{nei}(x_{ji})$.

Per giustificare l'impiego di regole decisionali basate su questo confronto, conviene partire dalla **scomposizione della devianza**:

$$\begin{array}{ccccc} dev_{tot}(x_{ji}) & = & dev_{fra}(x_{ji}) & + & dev_{nei}(x_{ji}) \\ \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ji}-\overline{\overline{x}})^2 & = & \sum_{j=1}^{k}(\overline{x}_j-\overline{\overline{x}})^2 n_j & + & \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ji}-\overline{x}_j)^2. \end{array}$$

Infatti, esaminando la formula della scomposizione della devianza

$$dev_{tot}(x_{ji}) = dev_{fra}(x_{ji}) + dev_{nei}(x_{ji})$$

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ji}-\overline{\overline{x}})^2 = \sum_{j=1}^{k}(\overline{x}_j-\overline{\overline{x}})^2 n_j + \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ji}-\overline{x}_j)^2.$$

possiamo trarre le seguenti conclusioni:

- Se è vera l'ipotesi nulla $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$, ci aspettiamo che le medie parziali $\overline{x}_j$ siano prossime l'una all'altra, e che siano anche prossime alla media totale $\overline{\overline{x}}$. In questo caso ci aspettiamo dunque che il valore di $dev_{fra}(x_{ji})$ spieghi soltanto una piccola quota di $dev_{tot}(x_{ji})$, ovvero che il rapporto tra $dev_{fra}(x_{ji})$ e $dev_{nei}(x_{ji})$ sia piccolo!!!
- Se, d'altra parte, l'ipotesi nulla $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ è falsa, ci aspettiamo di osservare medie parziali $\overline{x}_j$ molto più diverse l'una dall'altra e pertanto ci aspettiamo che, rispetto al caso in cui l'ipotesi nulla è vera, il valore $dev_{fra}(x_{ji})$ spieghi una quota più rilevante di $dev_{tot}(x_{ji})$, ovvero che il rapporto tra $dev_{fra}(x_{ji})$ e $dev_{nei}(x_{ji})$ sia più elevato di quello che osserveremmo nel caso in cui l'ipotesi nulla fosse vera!!!

In virtù delle considerazioni sopra esposte, nei test statistici con

$$H_0: \quad \mu_1 = \mu_2 = \cdots = \mu_k$$
$$H_a: \quad \text{almeno una delle } k \text{ medie } \mu_j \text{ è diversa dalle altre.}$$

di solito si considera la statistica test

$$F = \frac{dev_{fra}(x_{ji})/(k-1)}{dev_{nei}(x_{ji})/(n-k)}.$$

e si procede al rifiuto dell'ipotesi nulla

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

se il valore assunto dalla statistica test F supera il valore di una opportuna soglia critica s.

Per definire il valore numerico della soglia critica $s = s_\alpha$ in modo tale che il livello di significatività del test sia uguale ad un prefissato valore α, bisogna introdurre delle ipotesi sul processo che genera le osservazioni x_{ji} della variabile risposta X.

Spesso si considerano le seguenti ipotesi:

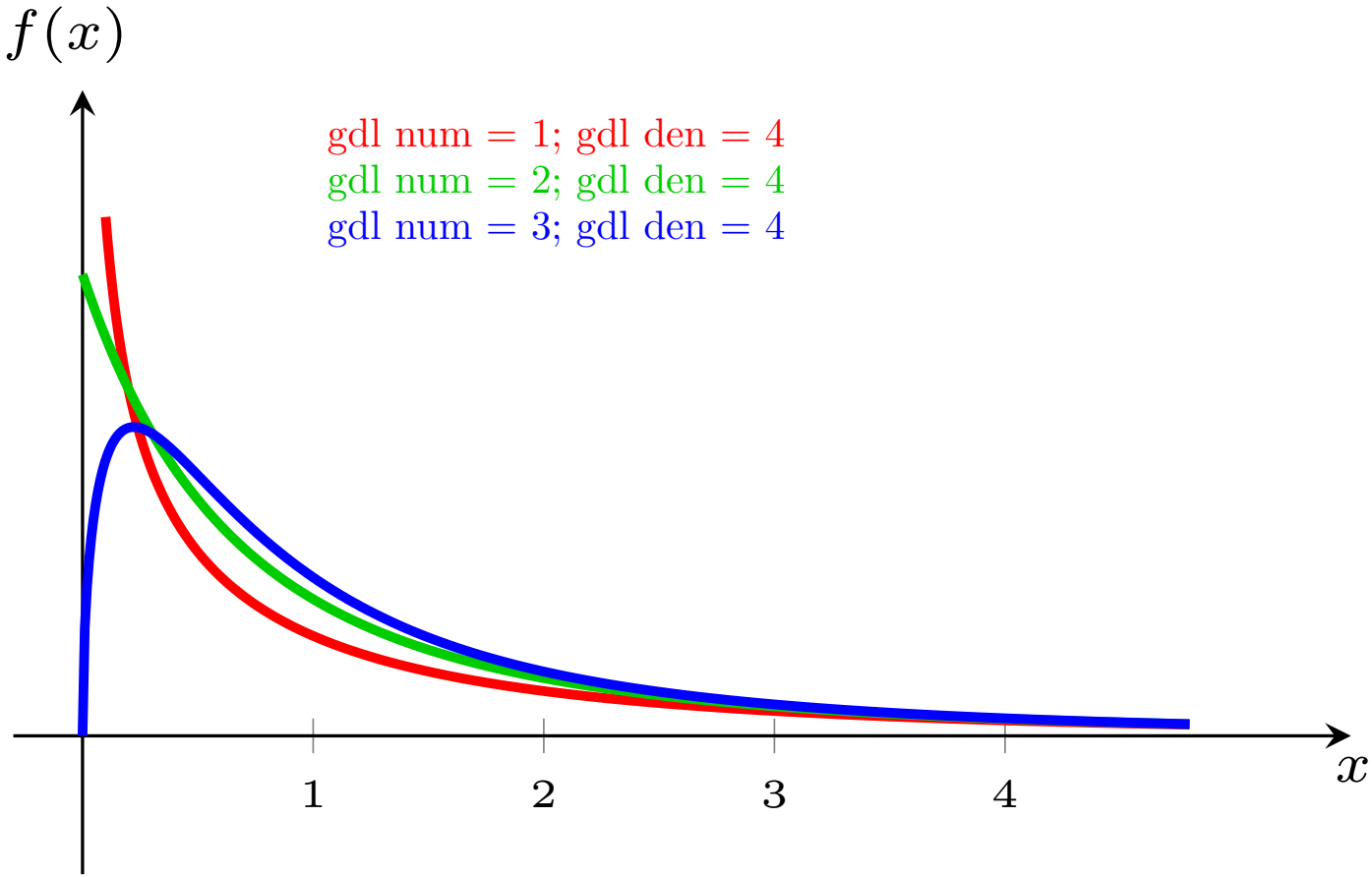
- tutti i valori x_{ji} sono realizzazioni di variabili casuali normali indipendenti con la stessa varianza σ^2;
- per ciascun gruppo (trattamento) $j = 1, 2, \ldots, k$, le variabili casuali normali che generano le osservazioni x_{ji} hanno tutte la stessa media μ_j.

Usando queste ipotesi si può dimostrare che la statistica test

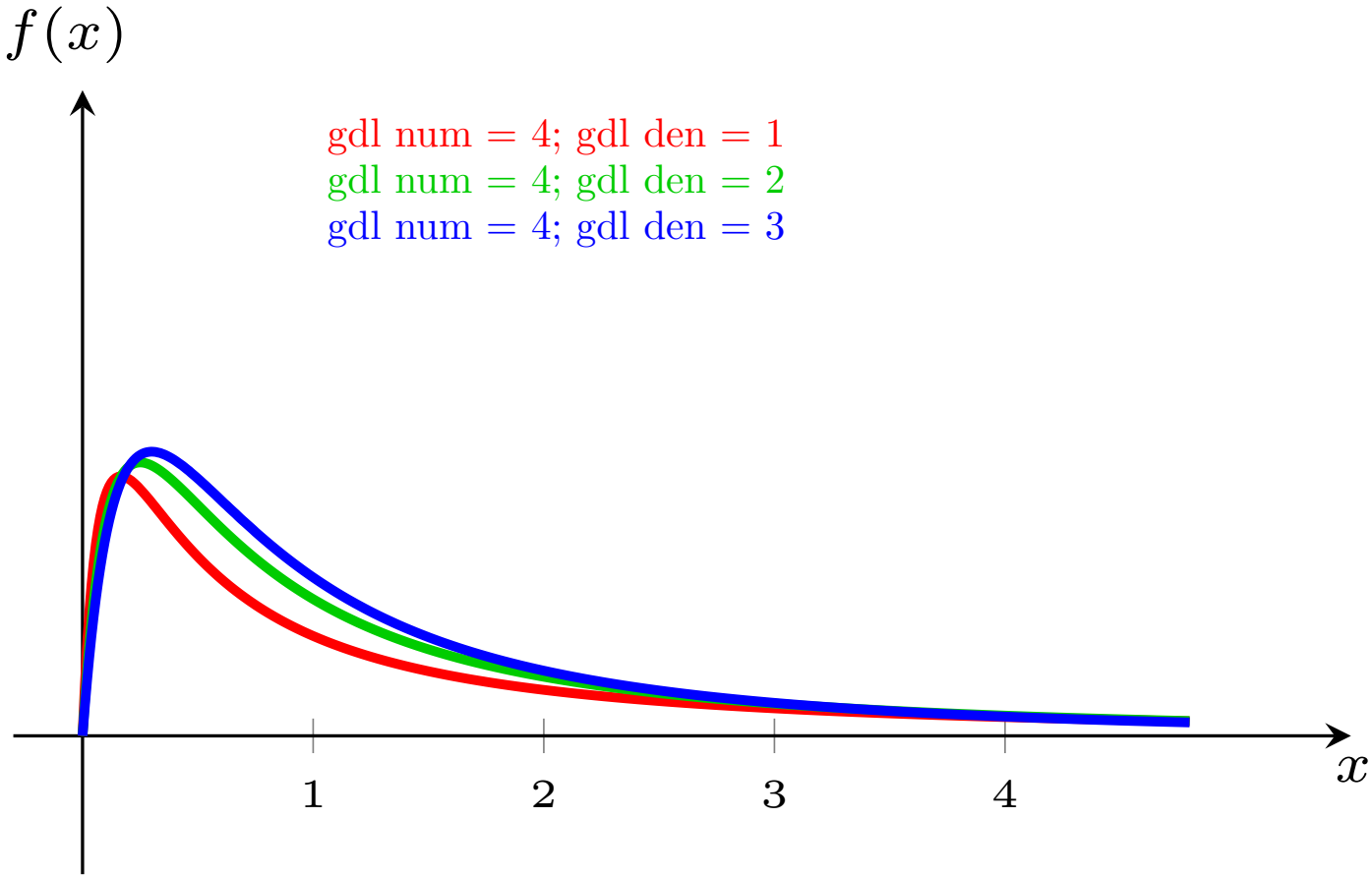
$$F = \frac{dev_{fra}(x_{ji})/(k-1)}{dev_{nei}(x_{ji})/(n-k)}$$

- ... segue la **distribuzione F di Fisher-Snedecor con $k-1$ gdl al numeratore e $n-k$ gdl al denominatore** quando l'ipotesi nulla $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ è vera, ...
- ... e che la distribuzione della statistica test F è "spostata più verso destra" quando invece l'ipotesi nulla $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ è falsa.

Grafici di alcune distribuzioni F di Fisher-Snedecor:



Grafici di alcune distribuzioni F di Fisher-Snedecor:



Onde ottenere un test con livello di significatività α, come soglia critica $s = s_\alpha$ per la statistica test

$$F = \frac{dev_{fra}(x_{ji})/(k-1)}{dev_{nei}(x_{ji})/(n-k)}$$

si considera dunque il percentile di ordine $1-\alpha$ della distribuzione F di Fisher-Snedecor con $k-1$ gdl al numeratore e $n-k$ gdl al denominatore.

In quanto segue indicheremo questo percentile con $F_{k-1;n-k;\alpha}$.

I valori dei percentili più usati delle distribuzioni F di Fisher-Snedecor sono reperibili in un'apposita tavola.

Parte iniziale della tavola delle distribuzioni F di Fisher-Snedecor:

Tavola della distribuzione F di Fisher																		
Gradi di libertà al denominatore	Area nella coda superiore	Gradi di libertà al numeratore																
		1	2	3	4	5	6	7	8	9	10	15	20	30	40	60	100	1000
1	0,1	39,863	49,500	53,593	55,833	57,240	58,204	58,906	59,439	59,858	60,195	61,220	61,740	62,265	62,529	62,794	63,007	63,296
	0,05	161,448	199,500	215,707	224,583	230,162	233,986	236,768	238,883	240,543	241,882	245,950	248,013	250,095	251,143	252,196	253,041	254,187
	0,025	647,789	799,500	864,163	899,583	921,848	937,111	948,217	956,656	963,285	968,627	984,867	993,103	1001,414	1005,598	1009,800	1013,175	1017,749
	0,01	4052,181	4999,500	5403,352	5624,583	5763,650	5858,986	5928,356	5981,070	6022,473	6055,847	6157,285	6208,730	6260,649	6286,782	6313,030	6334,110	6362,682
2	0,1	8,526	9,000	9,162	9,243	9,293	9,326	9,349	9,367	9,381	9,392	9,425	9,441	9,458	9,466	9,475	9,481	9,490
	0,05	18,513	19,000	19,164	19,247	19,296	19,330	19,353	19,371	19,385	19,396	19,429	19,446	19,462	19,471	19,479	19,486	19,495
	0,025	38,506	39,000	39,165	39,248	39,298	39,331	39,355	39,373	39,387	39,398	39,431	39,448	39,465	39,473	39,481	39,488	39,497
	0,01	98,503	99,000	99,166	99,249	99,299	99,333	99,356	99,374	99,388	99,399	99,433	99,449	99,466	99,474	99,482	99,489	99,498
3	0,1	5,538	5,462	5,391	5,343	5,309	5,285	5,266	5,252	5,240	5,230	5,200	5,184	5,168	5,160	5,151	5,144	5,135
	0,05	10,128	9,552	9,277	9,117	9,013	8,941	8,887	8,845	8,812	8,786	8,703	8,660	8,617	8,594	8,572	8,554	8,529
	0,025	17,443	16,044	15,439	15,101	14,885	14,735	14,624	14,540	14,473	14,419	14,253	14,167	14,081	14,037	13,992	13,956	13,908
	0,01	34,116	30,817	29,457	28,710	28,237	27,911	27,672	27,489	27,345	27,229	26,872	26,690	26,505	26,411	26,316	26,240	26,137
4	0,1	4,545	4,325	4,191	4,107	4,051	4,010	3,979	3,955	3,936	3,920	3,870	3,844	3,817	3,804	3,790	3,778	3,762
	0,05	7,709	6,944	6,591	6,388	6,256	6,163	6,094	6,041	5,999	5,964	5,858	5,803	5,746	5,717	5,688	5,664	5,632
	0,025	12,218	10,649	9,979	9,605	9,364	9,197	9,074	8,980	8,905	8,844	8,657	8,560	8,461	8,411	8,360	8,319	8,264
	0,01	21,198	18,000	16,694	15,977	15,522	15,207	14,976	14,799	14,659	14,546	14,198	14,020	13,838	13,745	13,652	13,577	13,475
5	0,1	4,060	3,780	3,619	3,520	3,453	3,405	3,368	3,339	3,316	3,297	3,238	3,207	3,174	3,157	3,140	3,126	3,107
	0,05	6,608	5,786	5,409	5,192	5,050	4,950	4,876	4,818	4,772	4,735	4,619	4,558	4,496	4,464	4,431	4,405	4,369
	0,025	10,007	8,434	7,764	7,388	7,146	6,978	6,853	6,757	6,681	6,619	6,428	6,329	6,227	6,175	6,123	6,080	6,022
	0,01	16,258	13,274	12,060	11,392	10,967	10,672	10,456	10,289	10,158	10,051	9,722	9,553	9,379	9,291	9,202	9,130	9,031
6	0,1	3,776	3,463	3,289	3,181	3,108	3,055	3,014	2,983	2,958	2,937	2,871	2,836	2,800	2,781	2,762	2,746	2,725
	0,05	5,987	5,143	4,757	4,534	4,387	4,284	4,207	4,147	4,099	4,060	3,938	3,874	3,808	3,774	3,740	3,712	3,673
	0,025	8,813	7,260	6,599	6,227	5,988	5,820	5,695	5,600	5,523	5,461	5,269	5,168	5,065	5,012	4,959	4,915	4,856
	0,01	13,745	10,925	9,780	9,148	8,746	8,466	8,260	8,102	7,976	7,874	7,559	7,396	7,229	7,143	7,057	6,987	6,891
7	0,1	3,589	3,257	3,074	2,961	2,883	2,827	2,785	2,752	2,725	2,703	2,632	2,595	2,555	2,535	2,514	2,497	2,473
	0,05	5,591	4,737	4,347	4,120	3,972	3,866	3,787	3,726	3,677	3,637	3,511	3,445	3,376	3,340	3,304	3,275	3,234
	0,025	8,073	6,542	5,890	5,523	5,285	5,119	4,995	4,899	4,823	4,761	4,568	4,467	4,362	4,309	4,254	4,210	4,149

Per illustrare un'applicazione del test F basato sull'ANOVA, calcoleremo ora il valore assunto dalla statistica test

$$F = \frac{dev_{fra}(x_{ji})/(k-1)}{dev_{nei}(x_{ji})/(n-k)}$$

nell'esempio sui vaccini.

Come abbiamo visto poco sopra, nell'esempio sui vaccini si ottiene

$$dev_{fra}(x_{ji}) = 94.808, \quad dev_{nei}(x_{ji}) = 95.438$$

e il valore della statistica test F è quindi dato da

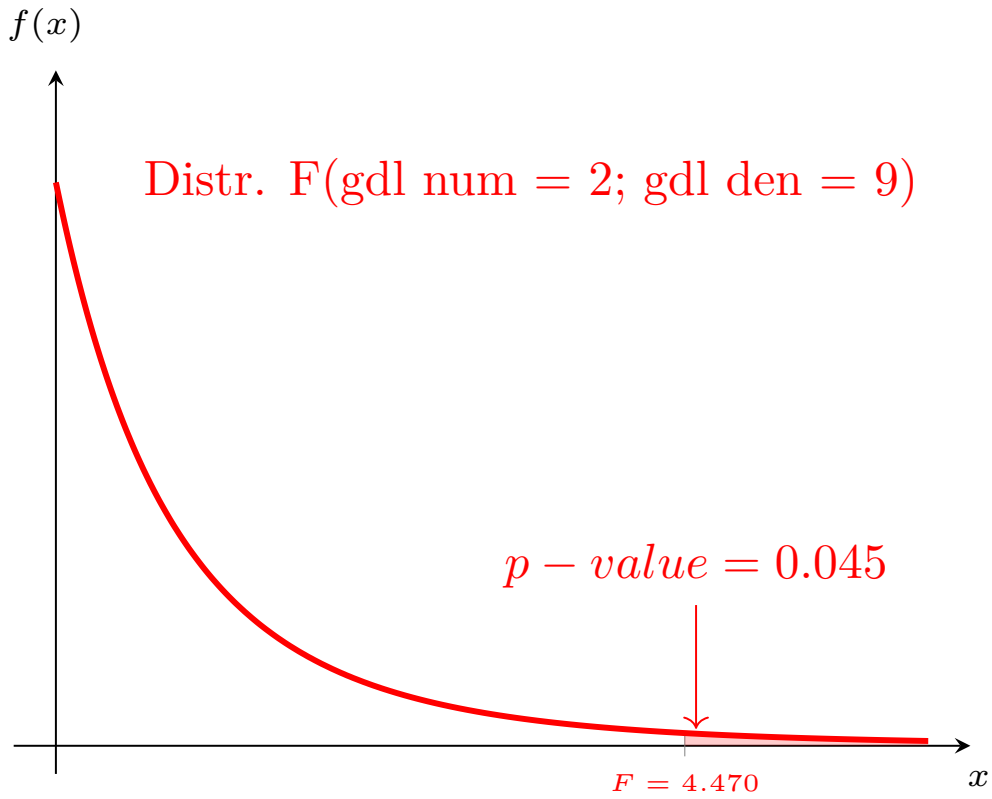
$$F = \frac{94,808/(3-1)}{95,438/(12-3)} = \frac{47,404}{10,604} = 4,470.$$

Confrontando questo valore con i percentili della distribuzione F di Fisher-Snedecor con $k-1 = 3-1 = 2$ gdl al numeratore e $n-k = 12-3 = 9$ gdl al denominatore, vediamo che

$$F_{2;9;\alpha=0,05} = 4,26 < F = 4,470 < 5,71 = F_{2;9;\alpha=0,025},$$

ovvero che il p-value è compreso tra $0,025$ e $0,05$.

Il valore esatto del p-value associato a $F = 4,468$ è riportato nel grafico sottostante:



Con questo p-value

- non possiamo rifiutare l'ipotesi nulla $H_0 : \mu_1 = \mu_2 = \mu_3$ al livello di significatività $\alpha = 0,025$,
- ma dobbiamo rifiutare $H_0 : \mu_1 = \mu_2 = \mu_3$ al livello di significatività $\alpha = 0,05$.

Analisi della Varianza come quella che abbiamo appena illustrato possono essere eseguite con molti *software*. Tipicamente, questi *software* restituiscono una tabella come quella nel grafico sottostante (che è stata ottenuta con EXCEL)

Analisi varianza: ad un fattore

RIEPILOGO

Gruppi	*Conteggio*	*Somma*	*Media*	*Varianza*
vaccino 1	4	127,012	31,753	17,4510607
vaccino 2	5	126,563	25,3126	10,1808448
vaccino 3	3	81,28	27,0933333	1,20435833

ANALISI VARIANZA

Origine della variazione	*SQ*	*gdl*	*MQ*	*F*	*e di significa*	*F crit*
Tra gruppi	94,8075111	2	47,4037555	4,46805842	0,04490693	4,25649473
In gruppi	95,4852779	9	10,6094753			
Totale	190,292789	11				

Esercizio 4.5

Nell'ambito di una ricerca di marketing è stato selezionato un campione di 23 donne. A ciascuna di esse è stata assegnata (in modo casuale) una di 4 creme di bellezza. Dopo aver provato la crema, a ciascuna donna è stato chiesto di esprimere un giudizio su una scala da 1 a 10. La tabella sottostante riassume le risposte:

crema	giudizi
AAA	6, 8, 9, 7
BBB	4, 7, 6, 6, 5
CCC	3, 5, 7, 6, 2, 4
DDD	6, 9, 6, 4, 3, 5, 6, 7

Sulla base di queste risposte, si può ritenere che una delle quattro creme sia preferita alle altre? Si risponda considerando un livello di significatività di $\alpha = 0,05$.

[[Si assuma che i giudizi siano stati espressi in modo indipendente, che le distribuzioni dei giudizi siano normali e che la varianza dei giudizi non dipenda dal tipo di crema.]]

Esercizio 4.6

Per confrontare le condizioni di vita in tre città diverse, sono stati intervistati tre campioni di famiglie. A ciascuna famiglia è stato chiesto l'ammontare della spesa settimanale per beni alimentari. Le risposte sono sintetizzate nella tabella sottostante:

città j	numerosità campionaria n_j	$\overline{x}_j$	s_j^2
1	250	120	101
2	120	118	89
3	60	122	112

- E' plausibile che i dati di questo esempio soddisfino le ipotesi per il test F dell'ANOVA? Quale di tali ipotesi potrebbe essere violata?
- Si assuma che le ipotesi per il test F dell'ANOVA siano soddisfatte. Si verifichi, attraverso il calcolo di un opportuno p-value, se le differenze tra le spese medie settimanali sono significative dal punto di vista statistico.

Confronto tra proporzioni

Inferenza sulla differenza tra due proporzioni

Supponiamo ora di essere interessati alla **differenza tra gli ignoti valori di due proporzioni** p_1 **e** p_2 che si riferiscono a **due popolazioni diverse**.

Per concretezza, assumiamo che

- p_1 sia l'ignota proporzione di fumatori nella popolazione femminile,
- e che p_2 sia l'ignota proporzione di fumatori nella popolazione maschile.

Chiaramente, la differenza tra queste due proporzioni può essere stimata selezionando

- un campione di n_1 individui all'interno della popolazione femminile,
- e un campione di n_2 individui all'interno della popolazione maschile.

Una volta rilevate le proporzioni di fumatori nei due campioni (che indicheremo con $\widehat{p}_1$ e $\widehat{p}_2$; i corrispondenti stimatori verranno invece indicati con $\widetilde{p}_1$ e $\widetilde{p}_2$), si può infatti stimare l'ignoto valore della differenza $p_1 - p_2$ attraverso la differenza tra le frequenze relative campionarie $\widehat{p}_1 - \widehat{p}_2$.

Si può dimostrare che

- **se i due campioni vengono selezionati in modo indipendente**
- e **se le numerosità di entrambi i campioni sono sufficientemente elevate** (diciamo in modo tale che $n\widehat{p}_1(1-\widehat{p}_1)$ e $n\widehat{p}_2(1-\widehat{p}_2)$ siano entrambi maggiori di 3), ...

... **allora la distribuzione di $\widetilde{p}_1 - \widetilde{p}_2$ è approssimativamente normale** con media

$$E(\widetilde{p}_1 - \widetilde{p}_2) = p_1 - p_2$$

e varianza

$$var(\widetilde{p}_1 - \widetilde{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}.$$

Ragionando come abbiamo già visto con riferimento ad una singola proporzione, possiamo dunque ottenere gli estremi di un **IdC asintotico per l'ignota differenza** $p_1 - p_2$ attraverso la formula

$$\widehat{p}_1 - \widehat{p}_2 \pm z_{1-\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}.$$

Chiaramente, il livello di confidenza nominale di questo IdC è $1 - \alpha$.

Se, per esempio, . . .

- . . . la proporzione di fumatrici in un campione di $n_1 = 81$ individui di sesso femminile fosse $\widehat{p}_1 = 27/81 = 0,33$,
- . . . e la proporzione di fumatori in un campione di $n_2 = 64$ individui di sesso maschile fosse $\widehat{p}_2 = 22/100 = 0,22$,

allora, considerando un livello di confidenza di $1 - \alpha = 0,95$ (quindi $z_{1-\alpha/2} = 1,96$), otterremmo l'IdC

$$0,33 - 0,22 \pm 1,96\sqrt{\frac{0,33(1-0,33)}{81} + \frac{0,22(1-0,22)}{100}} = \begin{cases} 0,241 \\ -0,021. \end{cases}$$

Si noti che questo IdC è molto largo. Per ottenere un IdC più stretto bisogna aumentare entrambe le numerosità campionarie.

Esercizio 5.1

Ad un esame si sono presentati 394 candidati. A ciascun candidato è stato assegnato in modo casuale uno di due temi d'esame.

- Tra i 225 candidati che hanno svolto il primo tema d'esame ci sono 25 candidati che non lo hanno superato,
- mentre tra i 169 candidati che hanno svolto il secondo tema d'esame, i candidati che non lo hanno superato sono 18.

Si calcoli un IdC al 98% per l'ignota differenza tra la probabilità di non superare l'esame con il primo e il secondo tema.

Vediamo ora invece come si eseguono dei **test statistici sulla differenza tra due proporzioni** $(p_1 - p_2)$.

Per non appesantire troppo la trattazione considereremo soltanto test statistici dove

- caso 1: $H_0 : p_1 - p_2 \geq 0$ e $H_a : p_1 - p_2 < 0$;
- caso 2: $H_0 : p_1 - p_2 \leq 0$ e $H_a : p_1 - p_2 > 0$;
- caso 3: $H_0 : p_1 - p_2 = 0$ e $H_a : p_1 - p_2 \neq 0$

. . . che sono quelli che si incontrano più frequentemente nelle applicazioni.

In questo tipo di test statistici si utilizza la statistica test

$$\widetilde{Z} = \frac{\widetilde{p}_1 - \widetilde{p}_2}{\sqrt{\widetilde{p}_m(1-\widetilde{p}_m)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

dove

$$\widetilde{p}_m = \frac{\widetilde{p}_1 n_1 + \widetilde{p}_2 n_2}{n_1 + n_2}$$

è una media ponderata delle due proporzioni campionarie $\widetilde{p}_1$ e $\widetilde{p}_2$.

Per giustificare l'impiego della statistica test

$$\widetilde{Z} = \frac{\widetilde{p}_1 - \widetilde{p}_2}{\sqrt{\widetilde{p}_m(1-\widetilde{p}_m)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

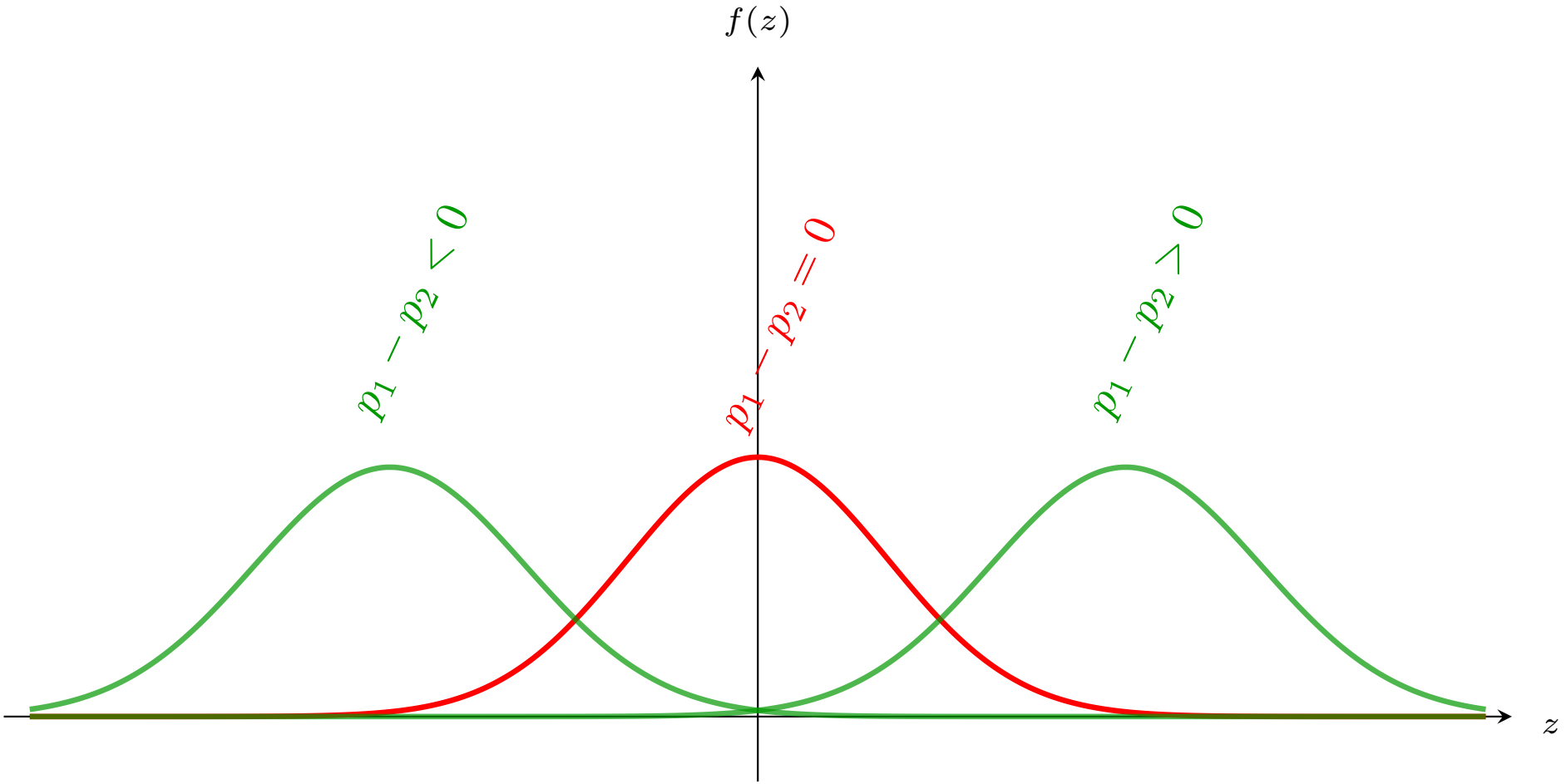
osserviamo che la sua **distribuzione** è in ogni caso (ovvero qualunque sia l'ignoto valore della differenza $p_1 - p_2$) **approssimativamente normale** (l'errore di approssimazione può essere considerato trascurabile se le quantità $n\widehat{p}_1(1-\widehat{p}_1)$ e $n\widehat{p}_2(1-\widehat{p}_2)$ sono entrambe maggiori di 3).

In particolare osserviamo che

- la distribuzione della statistica test $\widetilde{Z}$ è **approssimativamente normale standard** se $p_1 - p_2 = 0$ (si noti che in questo caso le tre ipotesi nulle $H_0 : p_1 - p_2 \geq 0$, $H_0 : p_1 - p_2 \leq 0$ e $H_0 : p_1 - p_2 = 0$ sono tutte vere),
- la distribuzione della statistica test $\widetilde{Z}$ si trova
  - **alla sinistra della distribuzione normale standard** se $p_1 - p_2 < 0$
  - **alla destra della distribuzione normale standard** se $p_1 - p_2 > 0$.

Il grafico mostra alcune possibili distribuzioni della statistica test

$$\widetilde{Z} = \frac{\widetilde{p}_1 - \widetilde{p}_2}{\sqrt{\widetilde{p}_m(1-\widetilde{p}_m)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$



Le considerazioni sulla distribuzione della statistica test

$$\widetilde{Z} = \frac{\widetilde{p}_1 - \widetilde{p}_2}{\sqrt{\widetilde{p}_m(1-\widetilde{p}_m)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

suggeriscono le seguenti regole decisionali:

Ipotesi statistiche	Regole decisionali: si rifiuta H_0 (livello sign. α) se
$H_0 : p_1 - p_2 \geq 0$ vs. $H_a : p_1 - p_2 < 0$	$z < -z_{1-\alpha}$
$H_0 : p_1 - p_2 \leq 0$ vs. $H_a : p_1 - p_2 > 0$	$z > z_{1-\alpha}$
$H_0 : p_1 - p_2 = 0$ vs. $H_a : p_1 - p_2 \neq 0$	$\lvert z \rvert > z_{1-\alpha/2}$

Volendo calcolare la realizzazione della statistica test

$$\widetilde{Z} = \frac{\widetilde{p}_1 - \widetilde{p}_2}{\sqrt{\widetilde{p}_m(1-\widetilde{p}_m)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

nell'esempio con le proporzioni di fumatrici e fumatori, otteniamo

$$\widehat{p}_1 = \frac{27}{81} = 0,33, \quad \widehat{p}_2 = \frac{22}{100} = 0,22 \quad \Rightarrow$$

$$\Rightarrow \quad \widehat{p}_m = \frac{\frac{27}{81} \times 81 + \frac{22}{100} \times 100}{81 + 100} = 0,271 \quad \Rightarrow$$

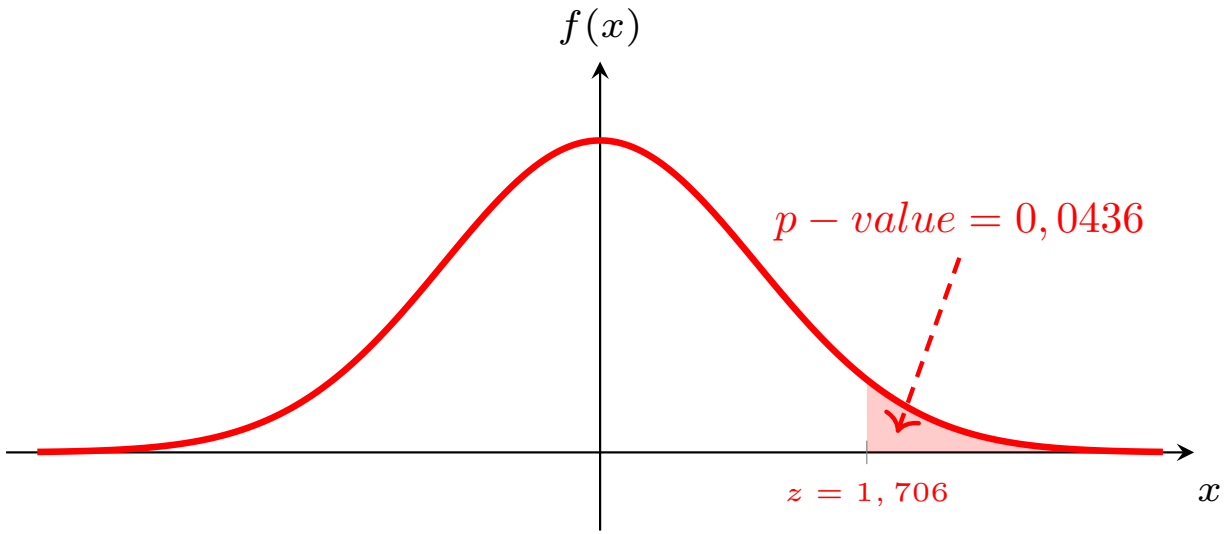
$$\Rightarrow \quad z = \frac{0,33 - 0,22}{\sqrt{0,271(1-0,271)\left(\frac{1}{81}+\frac{1}{100}\right)}} = 1,706.$$

- Se consideriamo il test con

$$H_0 : p_1 - p_2 \leq 0 \quad \text{vs.} \quad H_a : p_1 - p_2 > 0,$$

otteniamo quindi

$$\begin{aligned} p - value &= P(z > 1,706) = 1 - \Phi(1,706) \\ &\simeq 1 - \Phi(1,71) = 1 - 0,9564 = 0,0436. \end{aligned}$$



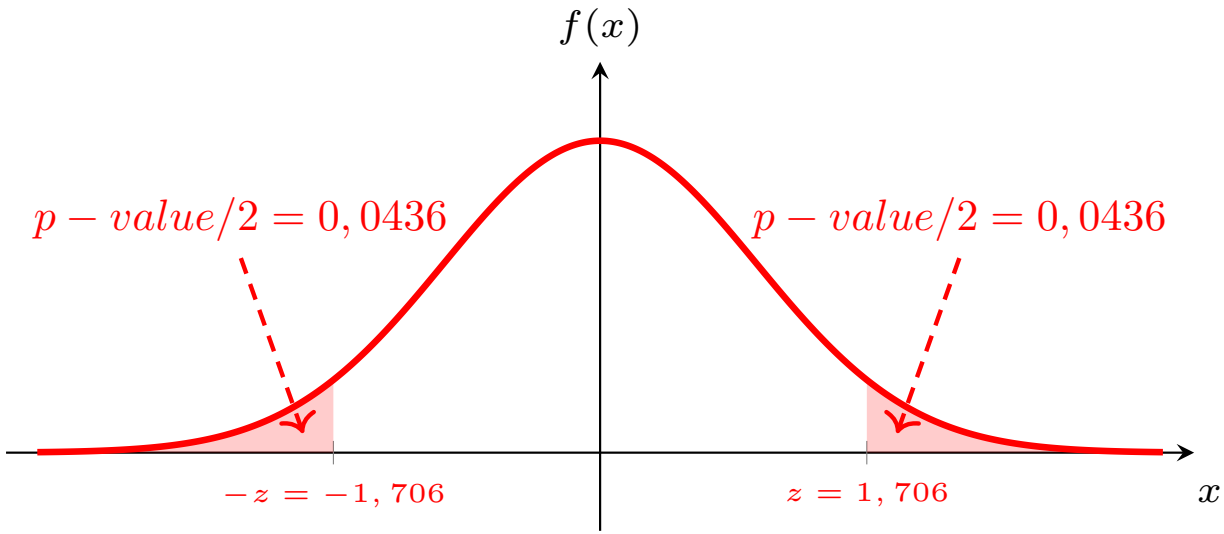
Se considerassimo $\alpha = 0,05$ come livello di significatività, dovremmo dunque rifiutare l'ipotesi nulla $H_0 : p_1 - p_2 \leq 0$.

- Se invece consideriamo il test con

$$H_0 : p_1 - p_2 = 0 \quad \text{vs.} \quad H_a : p_1 - p_2 \neq 0,$$

otteniamo

$$\begin{aligned} p - value &= P(|z| > 1,706) = 2 \times [1 - \Phi(1,706)] \\ &\simeq 2 \times [1 - \Phi(1,71)] = 2 \times [1 - 0,9564] = 0,0872. \end{aligned}$$



In questo caso l'ipotesi nulla $H_0 : p_1 - p_2 \leq 0$ non può essere rifiutata al livello di significatività $\alpha = 0,05$, ma verrebbe rifiutata al livello di significatività $\alpha = 0,10$.

Esercizio 5.2

I microchip prodotti da un'azienda provengono da due stabilimenti.

- In un campione di $n_1 = 400$ microchip prodotti nel primo stabilimento sono stati trovati 20 microchip difettosi, ...
- ... mentre in un altro campione di $n_2 = 625$ microchip prodotti nel secondo stabilimento sono stati trovati 60 microchip difettosi.

Siano p_1 e p_2 le ignote proporzioni di microchip difettosi prodotti nei due stabilimenti. Si calcoli il p-value di un test statistico con

$$H_0 : p_1 - p_2 = 0 \quad \text{vs.} \quad H_a : p_1 - p_2 \neq 0.$$

Consideriamo ora di nuovo **un'unica popolazione** le cui unità statistiche appartengono, a seconda della modalità che assume da un dato carattere X, ad una di $k \geq 2$ **categorie "incompatibili ed esaustive"**.[4]

Popolazioni di questo tipo vengono spesso chiamate **"multinomiali"**.

[4] Per definizione, k categorie sono **"incompatibili ed esaustive"** se e solo se tutte le unità statistiche della popolazione appartengono ad **una ed una sola** di tali categorie.

Per fare un **esempio di una "popolazione multinomiale"** possiamo considerare la **popolazione di tutti gli individui che parteciperanno alle prossime elezioni**.

Infatti, gli individui di questa popolazione possono essere suddivisi in $k = 3$ categorie incompatibili ed esaustive a seconda se intendono votare per il partito A, il partito B oppure un altro partito (considereremo quest'ultima come una categoria "residuale" che contiene anche gli elettori indecisi, gli elettori che hanno intenzione di votare scheda bianca, ecc.).

Si noti che le tre categorie "partito A", "partito B" e "altro partito" possono essere considerate come le tre modalità di un carattere qualitativo X che rappresenta le intenzioni di voto.

Chiaramente, con riferimento ad una popolazione multinomiale composta da k categorie incompatibili ed esaustive si possono definire k proporzioni

$$p_1, \quad p_2, \quad , \ldots , \quad p_k$$

ciascuna delle quali si riferisce ad una delle k categorie.

Siccome ciascuna unità statistica della popolazione appartiene ad una e una sola di queste k categorie, la somma delle proporzioni p_i deve essere uguale a 1.

Supponiamo ora che i **valori delle k proporzioni p_i siano ignoti**, ma di aver osservato un **campione casuale** che contiene

- n_1 unità statistiche che appartengono alla prima categoria,
- n_2 unità statistiche che appartengono alla seconda categoria,
- ...
- n_k unità statistiche che appartengono alla k-esima e ultima categoria.

Il numero complessivo di unità statistiche del campione è quindi dato da

$$n = n_1 + n_2 + \cdots + n_k.$$

Supponiamo ora di voler **utilizzare i dati di questo campione per verificare se gli ignoti valori p_i delle k proporzioni sono uguali a determinati valori p_i^***.

L'**ipotesi nulla** del suddetto test statistico è dunque definita come

$$H_0: \quad p_1 = p_1^*, \quad p_2 = p_2^*, \quad \ldots, \quad p_k = p_k^*,$$

mentre l'**ipotesi alternativa** è definita come

H_a : almeno una delle k proporzioni p_i è diversa dal valore p_i^* previsto dall'ipotesi nulla.

Con riferimento alla **popolazione multinomiale degli elettori** potremmo per esempio essere interessati a verificare se le **ignote** proporzioni di elettori che alle **prossime elezioni** hanno intenzione di votare il partito A, il partito B oppure un altro partito sono identiche alle corrispondenti proporzioni che si riferiscono alle **ultime elezioni** (ovviamente assumeremo che le proporzioni che si riferiscono alle ultime elezioni siano **note**).

In questo caso si avrebbe

$p_1 =$ ignota proporzione di elettori che voteranno per il partito A,

$p_2 =$ ignota proporzione di elettori che voteranno per il partito B,

$p_3 =$ ignota proporzione di elettori che voteranno per un altro partito.

e

$p_1^* =$ nota proporzione di elettori che alle ultime elezioni hanno votato per il partito A,

$p_2^* =$ nota proporzione di elettori che alle ultime elezioni hanno votato per il partito B,

$p_3^* =$ nota proporzione di elettori che alle ultime elezioni hanno votato per un altro partito.

Per decidere se l'ipotesi nulla

$$H_0: \quad p_1 = p_1^*, \quad p_2 = p_2^*, \quad \ldots, \quad p_k = p_k^*$$

deve essere rifiutata o meno, possiamo confrontare le **frequenze campionarie osservate** n_i con le corrispondenti **frequenze attese** che sono definite come

$$\widehat{n}_i = \text{ numerosità campionaria} \times p_i^* = n \times p_i^*, \quad i = 1, 2, \ldots, k.$$

Onde ottenere una misura per la "distanza complessiva" tra le frequenze osservate n_i e le frequenze attese $\widehat{n}_i$ possiamo calcolare la cosiddetta **statistica test "chi-quadrato"** che è definita come

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - \widehat{n}_i)^2}{\widehat{n}_i}.$$

Per illustrare il calcolo della statistica test chi-quadrato, conviene riprendere l'esempio sulla popolazione multinomiale degli elettori.

Supponiamo che alle ultime elezioni ...

- il 30% degli elettori abbia votato per il partito A, ...
- il 25% degli elettori abbia votato per il partito B ...
- e che il rimanente 45% degli elettori abbia votato per un altro partito

...e che l'ipotesi nulla sia dunque definita come

$$H_0: \quad p_1 = p_1^* = 0,30, \quad p_2 = p_2^* = 0,25, \quad p_3 = p_3^* = 0,45.$$

Supponiamo inoltre che un campione di $n = 1000$ elettori contenga

- il $n_1 = 333$ elettori che alle prossime elezioni hanno intenzione di votare per il partito A;
- il $n_2 = 260$ elettori che hanno intenzione di votare per il partito B
- e $n_3 = 410$ elettori che hanno intenzione di votare per un altro partito.

Ora, moltiplicando le proporzioni p_i^* previste dall'ipotesi nulla

$$H_0: \quad p_1 = p_1^* = 0,30, \quad p_2 = p_2^* = 0,25, \quad p_3 = p_3^* = 0,45.$$

per la numerosità campionaria $n = 1000$, otteniamo le **frequenze attese**

$$\widehat{n}_1 = 1000 \times 0,30 = 300, \quad \widehat{n}_2 = 1000 \times 0,25 = 250, \quad \widehat{n}_3 = 1000 \times 0,45 = 450$$

e sostituendo nella formula per calcolare la statistica test χ^2 otteniamo

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - \widehat{n}_i)^2}{\widehat{n}_i} = \frac{(333-300)^2}{300} + \frac{(260-250)^2}{250} + \frac{(410-450)^2}{450} = 6,956.$$

Chiaramente, **se è vera l'ipotesi nulla**

$$H_0: \quad p_1 = p_1^*, \quad p_2 = p_2^*, \quad \ldots, \quad p_k = p_k^*,$$

ci aspettiamo di ottenere una "distanza complessiva"

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - \widehat{n}_i)^2}{\widehat{n}_i}$$

piccola, ...

...mentre **se l'ipotesi nulla è falsa**, ci aspettiamo di ottenere una "distanza complessiva" χ^2 elevata.

Valori troppo elevati della statistica test χ^2 dovrebbero quindi condurre al rifiuto dell'ipotesi nulla!!!

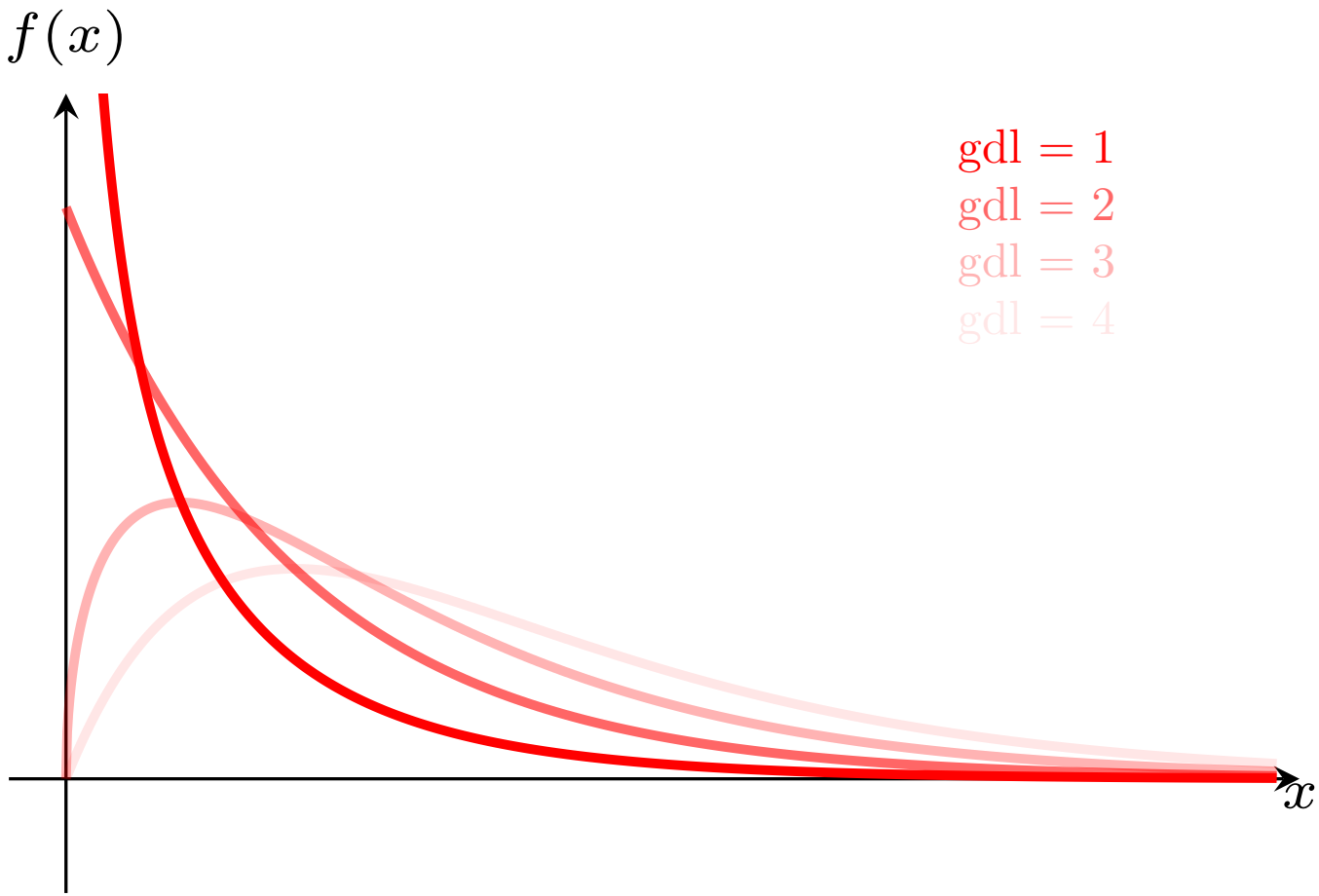
Avendo stabilito che valori troppo elevati della statistica test χ^2 dovrebbero condurre al rifiuto dell'ipotesi nulla

$$H_0: \quad p_1 = p_1^*, \quad p_2 = p_2^*, \quad \ldots, \quad p_k = p_k^*,$$

rimane soltanto più da determinare il valore di una **soglia critica** s_α in modo tale che il **livello di significatività** del test sia (almeno approssimativamente) uguale ad un prefissato valore α.

Onde determinare la soglia critica s_α, osserviamo che **se è vera l'ipotesi nulla** H_0, la distribuzione della statistica test χ^2 deve essere prossima alla cosiddetta **distribuzione chi-quadrato con $k-1$ gradi di libertà** (da cui il nome "chi-quadrato" della statistica test). L'approssimazione può essere ritenuta precisa se tutte e k le frequenze attese $\widehat{n}_i = np_i^*$ sono almeno pari a 3 (oppure a 5 se si desidera un livello di precisione maggiore).

Il grafico sottostante mostra alcune distribuzioni chi-quadrato.



Siccome **sotto l'ipotesi nulla**

$$H_0: \quad p_1 = p_1^*, \quad p_2 = p_2^*, \quad \ldots, \quad p_k = p_k^*,$$

la statistica test

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - \widehat{n}_i)^2}{\widehat{n}_i}.$$

ha una distribuzione che può essere approssimata attraverso la distribuzione chi-quadrato con $k - 1$ gdl, ...

... e siccome **sotto l'ipotesi alternativa** H_a ci aspettiamo di osservare valori elevati della statistica test χ^2 (in altre parole, se è vera l'ipotesi alternativa H_a la distribuzione della statistica test χ^2 si trova "più a destra" rispetto a dove si trova se è vera l'ipotesi nulla), ...

... possiamo ottenere un **test con livello di significatività** α se come **regola decisionale** adottiamo la regola di rifiutare H_0 quando la statistica test χ^2 supera il percentile di ordine $1 - \alpha$ della distribuzione chi-quadrato con $k - 1$ gdl (i percentili delle distribuzioni chi-quadrato sono reperibili in apposite tavole).

Parte iniziale di una tavola delle distribuzioni chi-quadrato.

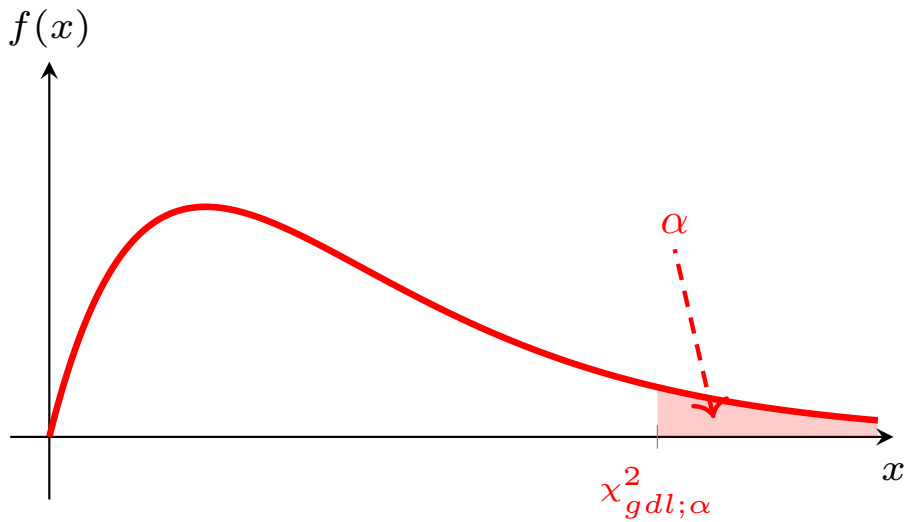


Tavola della distribuzione chi-quadrato

Area nella coda destra

gdl	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757

Vediamo dunque se nell'esempio sulla popolazione multinomiale degli elettori l'ipotesi nulla

$$H_0: \quad p_1 = p_1^* = 0,30, \quad p_2 = p_2^* = 0,25, \quad p_3 = p_3^* = 0,45$$

può essere rifiutata o meno.

Ricordiamo che per la statistica test χ^2 abbiamo ottenuto il valore $\chi^2 = 6,956$.

Siccome stiamo considerando una popolazione multinomiale composta da $k = 3$ categorie, il valore della statistica test $\chi^2 = 6,956$ deve essere confrontato con i percentili della distribuzione chi-quadrato con $k - 1 = 3 - 1 = 2$ gdl.

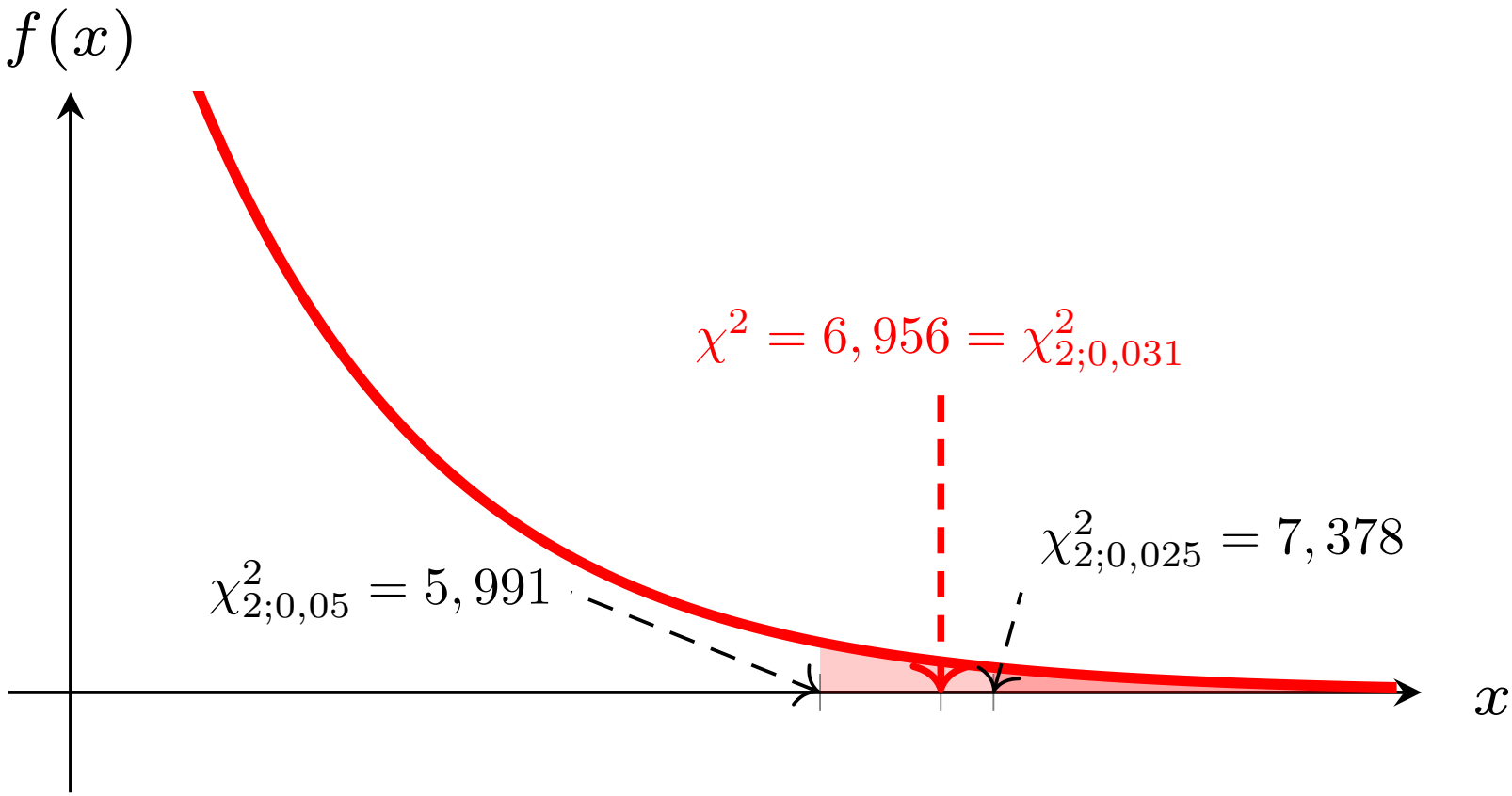
Dalla tavola della distribuzione chi-quadrato desumiamo che il valore $\chi^2 = 6.956$ della statistica test è compreso tra i percentili $\chi^2_{2;0.05} = 5.991$ e $\chi^2_{2;0.025} = 7.378$.

Tavola della distribuzione chi-quadrato

Area nella coda destra

gdl	0,995	0,990	0,975	0,950	0,900	0,100	0,050	0,025	0,010	0,005
1	0,000	0,000	0,001	0,004	0,016	2,706	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	0,211	4,605	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	0,584	6,251	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	1,064	7,779	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	1,610	9,236	11,070	12,833	15,086	16,750
6	0,676	0,872	1,237	1,635	2,204	10,645	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	2,833	12,017	14,067	16,013	18,475	20,278
8	1,344	1,646	2,180	2,733	3,490	13,362	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	4,168	14,684	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	4,865	15,987	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	5,578	17,275	19,675	21,920	24,725	26,757

Pertanto **il p-value associato alla statistica test $\chi^2 = 6,956$ è compreso tra** $0,025$ **e** $0,05$, Con l'ausilio di un software come EXCEL possiamo anche determinare il valore esatto del p-value, che ammonta a $p - value = 0,031$ (vedi il grafico sulla prossima *slide*).


$f(x)$
$\chi^2 = 6,956 = \chi^2_{2;0,031}$
$\chi^2_{2;0,025} = 7,378$
$\chi^2_{2;0,05} = 5,991$
x

Esercizio 5.3

Negli anni passati le quote di mercato per un determinato prodotto erano ripartite secondo quanto indicato nella seguente tabella:

azienda	AAA	BBB	CCC	DDD	Tot.
quota di mercato	45%	35%	15%	5%	100%

Nell'ambito di una recente indagine di mercato sono stati intervistati $n = 1500$ consumatori ai quali è stato chiesto se preferiscono il prodotto dell'azienda AAA, BBB, CCC o DDD. La tabella sottostante riporta il numero di preferenze per ciascuna delle quattro aziende:

azienda	AAA	BBB	CCC	DDD	Tot.
numero di preferenze	620	570	225	85	1500

Si può ritenere che rispetto al passato le quote di mercato si siano modificate? Si risponda determinando il p-value di un opportuno test statistico.

Assumiamo ora invece che le unità statistiche di una popolazione possano essere classificate in

- r categorie **incompatibili ed esaustive** in base alle modalità di un primo carattere X ...
- ...e in c categorie **incompatibili ed esaustive** in base alle modalità di un secondo carattere Y.

Considerando **tutti i possibili incroci tra le modalità dei due caratteri X_1 e X_2** otteniamo quindi $r \times c$ **"modalità incrociate"** alle quali corrispondono altrettante categorie incompatibili ed esaustive.

Per fare un esempio concreto possiamo considerare ancora la **popolazione degli elettori**:

- Si ricordi che avevamo suddiviso questa popolazione in $k_1 = 3$ **categorie** in base alle **intenzioni di voto** (carattere X_1 che può assumere una delle $k_1 = 3$ modalità "partito A", "partito B" o "altro partito").
- A questa prima classificazione possiamo ora **aggiungere un'altra classificazione** che si basa sull'**area geografica di residenza**. In quanto segue assumeremo che in base all'area geografica di residenza la popolazione possa essere suddivisa in $k_2 = 2$ categorie alle quali assoceremo le etichette "area est" e "area ovest".

Incrociando le modalità del carattere X_1 che rappresenta le intenzioni di voto con le modalità del carattere X_2 che rappresenta l'area geografica di residenza, otteniamo le $k_1 \times k_2 = 3 \times 2 = 6$ **modalità incrociate**

(partito A, area est), (partito A, area ovest), (partito B, area est), (partito B, area ovest), (altro partito, area est), (altro partito, area ovest).

Si noti che ad ogni elettore corrisponde **una e una sola** delle sopraelencate modalità incrociate e queste modalità incrociate suddividono quindi la popolazione in $k_1 \times k_2 = 3 \times 2 = 6$ categorie incompatibili ed esaustive.

Come abbiamo già visto nel corso di STATISTICA DESCRITTIVA, la distribuzione di una popolazione rispetto alle modalità di **due caratteri** può essere rappresentata attraverso una **tabella a doppia entrata** che per ciascun incrocio di modalità riporta il numero complessivo (ovvero la **frequenza congiunta** n_{ij}) di unità statistiche corrispondenti:

		Modalità di Y						
		y_1	y_2	$\cdots$	y_j	$\cdots$	y_c	$Tot.$
Modalità di X	x_1	n_{11}	n_{12}	$\cdots$	$\vdots$	$\cdots$	n_{1c}	$n_{1\cdot}$
	x_2	n_{21}	n_{22}	$\cdots$	$\vdots$	$\cdots$	n_{2c}	$n_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	x_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{ic}	$n_{i\cdot}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	x_r	n_{r1}	n_{r2}	$\cdots$	$\vdots$	$\cdots$	n_{rc}	$n_{r\cdot}$
	$Tot.$	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot r}$	N

Le **frequenze marginali** $n_{i\cdot}$ nella colonna marginale sono i totali di riga e rappresentano il numero complessivo di unità statistiche che presentano la modalità x_i per il carattere X (le frequenze marginali $n_{\cdot j}$ che si trovano nella riga marginale hanno significato analogo).

La distribuzione di una popolazione di elettori rispetto all'intenzione di voto (carattere X) e all'area geografica di residenza (carattere Y) potrebbe per esempio essere rappresentata attraverso la tabella

		Modalità di Y		
		area est	area ovest	*Tot.*
Modalità di X	partito A	5435	8054	13489
	partito B	6893	10354	17247
	altro partito	2403	14604	17007
	Tot.	14731	33012	47743

A partire da una tabella a doppia entrata, si possono facilmente ricavare le **frequenze relative delle distribuzioni condizionate** dei due caratteri X e Y.

Per ottenere le **frequenze relative delle distribuzioni condizionate del carattere** X (le cui modalità sono riportate per riga), basta dividere ciascuna frequenza congiunta n_{ij} per il corrispondente totale di colonna $n_{\cdot j}$:

		Modalità di Y						
		y_1	y_2	$\cdots$	y_j	$\cdots$	y_c	$\frac{n_{i\cdot}}{N}$
Modalità di X	x_1	$\frac{n_{11}}{n_{\cdot 1}}$	$\frac{n_{12}}{n_{\cdot 2}}$	$\cdots$	$\vdots$	$\cdots$	$\frac{n_{1c}}{n_{\cdot c}}$	$\frac{n_{1\cdot}}{N}$
	x_2	$\frac{n_{21}}{n_{\cdot 1}}$	$\frac{n_{22}}{n_{\cdot 2}}$	$\cdots$	$\vdots$	$\cdots$	$\frac{n_{2c}}{n_{\cdot c}}$	$\frac{n_{2\cdot}}{N}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	x_i	$\frac{n_{i1}}{n_{\cdot 1}}$	$\frac{n_{i2}}{n_{\cdot 2}}$	$\cdots$	$\frac{n_{ij}}{n_{\cdot j}}$	$\cdots$	$\frac{n_{ic}}{n_{\cdot c}}$	$\frac{n_{i\cdot}}{N}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	x_r	$\frac{n_{r1}}{n_{\cdot 1}}$	$\frac{n_{r2}}{n_{\cdot 2}}$	$\cdots$	$\vdots$	$\cdots$	$\frac{n_{rc}}{n_{\cdot c}}$	$\frac{n_{r\cdot}}{N}$
	Tot.	1	1	$\cdots$	1	$\cdots$	1	1

Si noti che la **colonna marginale** questa tabella contiene le **frequenze relative della distribuzione marginale di** X.

Allo stesso modo, dividendo ciascuna frequenza congiunta n_{ij} per il corrispondente totale di riga $n_{i\cdot}$ si ottengono le **frequenze relative delle distribuzioni condizionate del carattere** Y.

		Modalità di Y						
		y_1	y_2	$\cdots$	y_j	$\cdots$	y_c	Tot.
Modalità di X	x_1	$\frac{n_{11}}{n_{1\cdot}}$	$\frac{n_{12}}{n_{1\cdot}}$	$\cdots$	$\vdots$	$\cdots$	$\frac{n_{1c}}{n_{1\cdot}}$	1
	x_2	$\frac{n_{21}}{n_{2\cdot}}$	$\frac{n_{22}}{n_{2\cdot}}$	$\cdots$	$\vdots$	$\cdots$	$\frac{n_{2c}}{n_{2\cdot}}$	1
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	x_i	$\frac{n_{i1}}{n_{i\cdot}}$	$\frac{n_{i2}}{n_{i\cdot}}$	$\cdots$	$\frac{n_{ij}}{n_{i\cdot}}$	$\cdots$	$\frac{n_{ic}}{n_{i\cdot}}$	1
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	x_r	$\frac{n_{r1}}{n_{r\cdot}}$	$\frac{n_{r2}}{n_{r\cdot}}$	$\cdots$	$\vdots$	$\cdots$	$\frac{n_{rc}}{n_{r\cdot}}$	1
	$\frac{n_{\cdot j}}{N}$	$\frac{n_{\cdot 1}}{N}$	$\frac{n_{\cdot 2}}{N}$	$\cdots$	$\frac{n_{\cdot j}}{N}$	$\cdots$	$\frac{n_{\cdot r}}{N}$	1

Si noti che la **riga marginale** di questa tabella contiene le **frequenze relative della distribuzione marginale di** Y.

Con riferimento all'esempio sulla popolazione di elettori, otteniamo le seguenti distribuzioni condizionate per l'intenzione di voto X:

		Modalità di Y		
		area est	area ovest	$\frac{n_{i\cdot}}{N}$
Modalità di X	partito A	$\frac{5435}{14731} = 0,369$	$\frac{8054}{33012} = 0,244$	$\frac{13489}{47743} = 0,283$
	partito B	$\frac{6893}{14731} = 0,468$	$\frac{10354}{33012} = 0,314$	$\frac{17247}{47743} = 0,361$
	altro partito	$\frac{2403}{14731} = 0,163$	$\frac{14604}{33012} = 0,442$	$\frac{17007}{47743} = 0,356$
	Tot.	1	1	1

Si noti che le distribuzioni condizionate di X sono due perché il carattere Y assume soltanto due modalità diverse.

La colonna marginale contiene invece le frequenze relative della distribuzione marginale di X, ovvero della distribuzione di X che si riferisce all'intera popolazione.

Con riferimento alle **distribuzioni condizionate** di un dato carattere X, si è spesso interessati a verificare se sono tutte **"simili"** alla distribuzione marginale, ovvero a verificare se tutte le colonne nella tabella che contiene le frequenze relative delle distribuzioni **condizionate** di X sono uguali alla colonna che contiene le frequenze relative della distribuzione **marginale** di X:

$$\frac{n_{11}}{n_{.1}} = \frac{n_{12}}{n_{.2}} = \cdots = \frac{n_{1c}}{n_{.c}} = \frac{n_{1.}}{N}$$

$$\frac{n_{21}}{n_{.1}} = \frac{n_{22}}{n_{.2}} = \cdots = \frac{n_{2c}}{n_{.c}} = \frac{n_{2.}}{N}$$

$$\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots$$

$$\frac{n_{r1}}{n_{.1}} = \frac{n_{r2}}{n_{.2}} = \cdots = \frac{n_{rc}}{n_{.c}} = \frac{n_{r.}}{N}$$

Se questa condizione è soddisfatta si dice che tra i caratteri X e Y c'è **indipendenza distributiva**.

Si noti che nell'esempio sugli elettori non c'è indipendenza distributiva tra l'intenzione di voto X e l'area geografica di residenza Y:

		Modalità di Y		
		area est	area ovest	$\frac{n_{i\cdot}}{N}$
Modalità di X	partito A	$\frac{5435}{14731} = 0,369$	$\frac{8054}{33012} = 0,244$	$\frac{13489}{47743} = 0,283$
	partito B	$\frac{6893}{14731} = 0,468$	$\frac{10354}{33012} = 0,314$	$\frac{17247}{47743} = 0,361$
	altro partito	$\frac{2403}{14731} = 0,163$	$\frac{14604}{33012} = 0,442$	$\frac{17007}{47743} = 0,356$
	Tot.	1	1	1

Infatti, dalla prima riga della tabella si desume che

$$\frac{5435}{14731} = 0,369 \neq 0,244 = \frac{8054}{33012}$$

ovvero che tra gli elettori residenti nell'area est ci sono, **in termini relativi**, molti più elettori intenzionati a votare il partito A che tra gli elettori residenti nell'area ovest e che nessuna di queste due frequenze relative condizionate coincide con la corrispondente frequenza relativa marginale $\frac{13489}{47743} = 0,283$ (che si riferisce all'intera popolazione).

Non è difficile rendersi conto che **il concetto di indipendenza distributiva è "simmetrico"**, ovvero che . . .

> . . . tutte le distribuzioni condizionate di X sono "simili" alla distribuzione marginale di X se e solo se anche tutte le distribuzioni condizionate di Y sono "simili" alla distribuzione marginale di Y.

Infatti,

$$\frac{n_{ij}}{n_{\cdot j}} = \frac{n_{i\cdot}}{N} \quad \forall i, j \quad \Leftrightarrow \quad \frac{n_{ij}}{n_{i\cdot}} = \frac{n_{\cdot j}}{N} \quad \forall i, j$$

e questa doppia implicazione dice che tutte le **colonne** della tabella con le frequenze relative delle distribuzioni condizionate di X sono uguali alla **colonna marginale** che contiene le frequenze relative della distribuzione marginale di X **se e solo se** anche tutte le **righe** della tabella con le frequenze relative delle distribuzioni condizionate di Y sono uguali alla **riga marginale** che contiene le frequenze relative della distribuzione marginale di Y.

Nel caso dell'esempio sulla popolazione di elettori abbiamo già visto che nella tabella con le frequenze relative delle distribuzioni condizionate dell'intenzione di voto X, le **colonne** centrali (che contengono le frequenze relative delle distribuzioni condizionate di X) non sono tutte identiche alla **colonna marginale** che contiene le frequenze relative della distribuzione marginale di X (vedi sotto), ...

		Modalità di Y		
		area est	area ovest	$\frac{n_{i\cdot}}{N}$
Modalità di X	partito A	$\frac{5435}{14731} = 0,369$	$\frac{8054}{33012} = 0,244$	$\frac{13489}{47743} = 0,283$
	partito B	$\frac{6893}{14731} = 0,468$	$\frac{10354}{33012} = 0,314$	$\frac{17247}{47743} = 0,361$
	altro partito	$\frac{2403}{14731} = 0,163$	$\frac{14604}{33012} = 0,442$	$\frac{17007}{47743} = 0,356$
	Tot.	1	1	1

Siccome il concetto di indipendenza distributiva è "simmetrico", possiamo dunque concludere che ...

. . . nella tabella con le frequenze relative delle distribuzioni condizionate dell'area geografica di residenza Y, le **righe** centrali (che contengono le frequenze relative delle distribuzioni condizionate di Y) non possono essere identiche alla **riga marginale** che contiene le frequenze relative della distribuzione marginale di Y!!!

		Modalità di Y		
		area est	area ovest	Tot.
Modalità di X	partito A	$\frac{5435}{13489} = 0,403$	$\frac{8054}{13489} = 0,597$	1
	partito B	$\frac{6893}{17247} = 0,400$	$\frac{10354}{17247} = 0,600$	1
	altro partito	$\frac{2403}{17007} = 0,141$	$\frac{14604}{17007} = 0,859$	1
	$\frac{n_{\cdot j}}{N}$	$\frac{14731}{47743} = 0,309$	$\frac{33012}{47743} = 0,691$	1

Dalla condizione

$$\frac{n_{ij}}{n_{\cdot j}} = \frac{n_{i\cdot}}{N} \quad \forall i, j$$

che abbiamo posto come **condizione definitoria per il concetto di indipendenza distributiva**, si deduce immediatamente che si ha indipendenza distributiva se e solo se

$$n_{ij} = \frac{n_{i\cdot} \times n_{\cdot j}}{N} \quad \forall i, j.$$

Per questo motivo, le frequenze

$$\widehat{n}_{ij} = \frac{n_{i\cdot} \times n_{\cdot j}}{N}$$

vengono chiamate **frequenze teoriche di indipendenza**.

Un modo alternativo per **definire** il concetto di **indipendenza distributiva** è quindi quello di dire che due caratteri X e Y sono indipendenti in distribuzione se **l'intera tabella delle frequenze congiunte "effettive" n_{ij} è identica alla tabella con le corrispondenti frequenze teoriche di indipendenza $\widehat{n}_{ij}$.**

Com'è facile verificare, nella tabella con la distribuzione congiunta dell'intenzione di voto X e dell'area geografica di residenza Y questa condizione non è soddisfatta: infatti, la tabella con le frequenze congiunte "effettive" è data da

	n_{ij}	Modalità di Y: area est	area ovest	*Tot.*
Modalità di X	partito A	5435	8054	13489
	partito B	6893	10354	17247
	altro partito	2403	14604	17007
	Tot.	14731	33012	47743

. . . mentre la tabella con le corrispondenti **frequenze teoriche di indipendenza** è data da

	$\widehat{n}_{ij}$	Modalità di Y: area est	area ovest	*Tot.*
Modalità di X	partito A	4162,00	9327,00	13489
	partito B	5321,52	11925,48	17247
	altro partito	5247,47	11759,53	17007
	Tot.	14731	33012	47743

Si noti che nel ragionamento che ci ha condotti alla definizione del concetto di **"indipendenza distributiva"**, siamo partiti dall'ipotesi che la popolazione di riferimento fosse **finita**.

Infatti, per giungere alla definizione di "indipendenza distributiva" siamo partiti dalla tabella a doppia entrata che contiene le **frequenze congiunte assolute** n_{ij} delle modalità incrociate di due caratteri. ...

. . . Si noti, tuttavia, che la definizione del concetto di **"indipendenza distributiva"** può (in molti casi) essere estesa anche al caso in cui la popolazione di riferimento è **infinita**.

Infatti, abbiamo visto che la definizione di **"indipendenza distributiva"** è basata sul confronto tra le **frequenze relative** delle distribuzioni condizionate di un carattere e le **frequenze relative** della distribuzione marginale dello stesso carattere . . .

. . . e **anche per popolazioni infinite ha (in molti casi) senso parlare di frequenze relative**.

[[Si noti, d'altra parte, che **per una popolazione infinita ha poco senso parlare di frequenze assolute** perché per un dato carattere X potrebbero esserci due o più modalità con frequenza assoluta infinita e in questo caso non si potrebbe dire nulla sulla prevalenza di queste modalità all'interno della popolazione.]]

Per esempio, con riferimento alla popolazione infinita di tutti coloro che si infettano con un determinato virus ha senso parlare

- della frequenza relativa (marginale) di coloro che presentano il sintomo X (modalità *"presente"* del carattere X)
- della frequenza relativa (marginale) di coloro che presentano il sintomo Y (modalità *"presente"* del carattere Y)
- della frequenza relativa di coloro che presentano entrambi i sintomi (incrocio di modalità (*"presente"*, *"presente"*)),
- e anche della frequenza relativa condizionata di coloro che presentano il sintomo X tra tutti coloro che presentano il sintomo Y.

Confrontando la frequenza relativa condizionata di coloro che presentano il sintomo X tra tutti coloro che presentano il sintomo Y con frequenza relativa (marginale) di coloro che presentano il sintomo X, possiamo quindi stabilire se tra i caratteri X e Y vi sia indipendenza distributiva, ...

... ovvero possiamo stabilire se i sintomi X e Y si manifestano in modo indipendente, oppure se quando si manifesta uno dei due sintomi sia più o meno probabile che si manifesti anche l'altro.

A questo punto,

- dopo aver rivisto la definizione di indipendenza distributiva con riferimento ad una popolazione finita, . . .
- . . . e dopo aver osservato che la definizione di questo concetto ha (spesso) senso anche quando la popolazione di riferimento è infinita,

. . . affronteremo un problema legato all'indipendenza che si incontra in molte situazioni reali:

Come si può verificare l'indipendenza distributiva sulla base di un campione???

Come vedremo tra breve, per questo problema può essere affrontato con un test statistico molto simile a quello sulle proporzioni di una popolazione multinomiale.

Supponiamo dunque che la tabella

		Modalità di Y						
		y_1	y_2	$\cdots$	y_j	$\cdots$	y_c	*Tot.*
Modalità di X	x_1	n_{11}	n_{12}	$\cdots$	$\vdots$	$\cdots$	n_{1c}	$n_{1\cdot}$
	x_2	n_{21}	n_{22}	$\cdots$	$\vdots$	$\cdots$	n_{2c}	$n_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	x_i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	$\cdots$	n_{ic}	$n_{i\cdot}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
	x_r	n_{r1}	n_{r2}	$\cdots$	$\vdots$	$\cdots$	n_{rc}	$n_{r\cdot}$
	Tot.	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot r}$	n

contenga le frequenze congiunte assolute n_{ij} che si riferiscono ad un **campione** di numerosità n, ...

... e supponiamo di voler verificare se sulla base di questa tabella si può rifiutare l'ipotesi nulla secondo la quale nella **popolazione** di riferimento i caratteri X e Y sono distribuiti in modo indipendente (la popolazione può essere finita o anche infinita).

L'**ipotesi nulla** del test statistico è dunque definita come

$$H_0: \quad \text{i caratteri } X \text{ e } Y \text{ sono distribuiti in modo indipendente}$$

e l'**ipotesi alternativa** è invece definita come

$$H_a: \quad \text{i caratteri } X \text{ e } Y \underline{\text{non}} \text{ sono distribuiti in modo indipendente.}$$

Chiaramente, se è vera l'ipotesi nulla H_0, ci aspettiamo che le frequenze n_{ij} osservate nel campione siano prossime alle corrispondenti frequenze teoriche di indipendenza $\widehat{n}_{ij} = \frac{n_{i\cdot} \times n_{\cdot j}}{n}$,

. . . mentre in caso contrario ci aspettiamo che le frequenze osservate siano più lontane da quelle teoriche di indipendenza.

Per valutare la distanza tra le frequenze osservate del campione e le frequenze teoriche di indipendenza, possiamo calcolare il valore della **statistica test**

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - \widehat{n}_{ij})^2}{\widehat{n}_{ij}}$$

che viene anch'essa chiamata **statistica test chi-quadrato**.

Il nome **"chi-quadrato"** di questa statistica test non è soltanto dovuto alla sua somiglianza con la statistica test chi-quadrato che abbiamo già visto in relazione al test sulle proporzioni di una popolazione multinomiale, ma è dovuto soprattutto al fatto che . . .

. . . **sotto l'ipotesi nulla di indipendenza la distribuzione di questa statistica test è prossima alla distribuzione chi-quadrato con** $gdl = (r-1) \times (c-1)$ (si ricordi che r e c indicano, rispettivamente, il numero di righe e il numero di colonne della tabella). L'approssimazione può essere ritenuta sufficientemente precisa se tutte le frequenze teoriche di indipendenza sono maggiori di 3 (o di 5 se si vuole un grado di precisione maggiore).

D'altra parte, **sotto l'ipotesi alternativa che prevede l'assenza di indipendenza**, la distribuzione della statistica test

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - \widehat{n}_{ij})^2}{\widehat{n}_{ij}}$$

è **"spostata verso destra"**, ...

...e per questo motivo **anche i test statistici sull'indipendenza riguardano la coda superiore della distribuzione chi-quadrato**: se α è il livello di significatività di un test sull'indipendenza, l'ipotesi nulla di indipendenza viene rifiutata se il valore assunto dalla statistica test χ^2 supera la **soglia critica**

$$s_\alpha = \chi^2_{gdl;\alpha} \quad \text{con } gdl = (r-1) \times (c-1).$$

[[Ricordiamo che $\chi^2_{gdl;\alpha}$ è il percentile di ordine $1-\alpha$ della distribuzione chi-quadrato con *gdl* gradi di liberà, ovvero il valore che lascia alla sua destra il $\alpha \times 100\%$ dell'area complessiva sottesa alla suddetta distribuzione.]]

Per illustrare il test sull'indipendenza attraverso un esempio pratico, supponiamo che la seguente tabella sia stata ottenuta a partire da un **campione casuale semplice** di $n = 330$ elettori:

n_{ij}		Modalità di Y: area est	Modalità di Y: area ovest	*Tot.*
Modalità di X	partito A	45	35	80
	partito B	45	75	120
	altro partito	60	70	130
	Tot.	150	180	330

A partire dalle frequenze marginali possiamo facilmente ricavare la corrispondente tabella con le **frequenze teoriche di indipendenza**:

$\widehat{n}_{ij}$		Modalità di Y: area est	Modalità di Y: area ovest	*Tot.*
Modalità di X	partito A	36,36	43,64	80
	partito B	54,55	65,45	120
	altro partito	59,09	70,91	130
	Tot.	150	180	330

A questo punto possiamo anche calcolare il valore della statistica test

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - \widehat{n}_{ij})^2}{\widehat{n}_{ij}}.$$

A tal fine conviene costruire un'ulteriore tabella che contiene i termini della doppia sommatoria ...

	$\frac{(n_{ij}-\widehat{n}_{ij})^2}{\widehat{n}_{ij}}$	Modalità di Y area est	area ovest	*Tot.*
Modalità di X	partito A	2,051	1,709	—
	partito B	1,670	1,392	—
	altro partito	0,014	0,012	—
	Tot.	—	—	$\chi^2 = 6,849$

Come si desume da quest'ultima tabella,

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - \widehat{n}_{ij})^2}{\widehat{n}_{ij}} = 6,849.$$

Per valutare se l'ipotesi nulla di indipendenza può essere rifiutata, confronteremo ora il valore della statistica test

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - \widehat{n}_{ij})^2}{\widehat{n}_{ij}} = 6,849$$

con i percentili della distribuzione chi-quadrato con

$$gdl = (r-1) \times (c-1) = (3-1) \times (2-1) = 2.$$

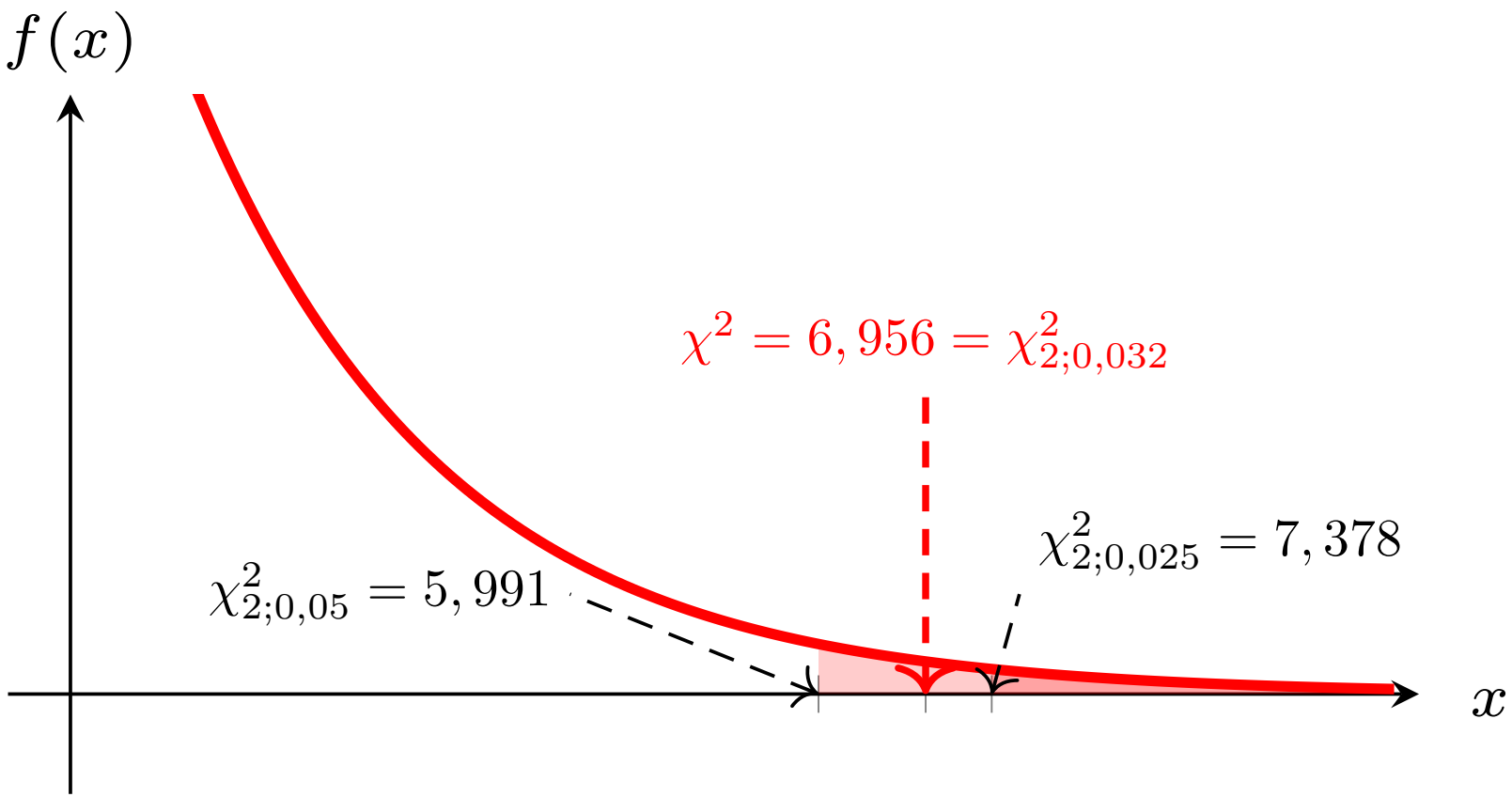
Come si desume dalla tavola della distribuzione chi-quadrato,

$$\chi^2_{2;0,05} = 5,991 < \chi^2 = 6,849 < \chi^2_{2;0,025} = 7,378$$

e **il p-value del test sull'indipendenza è quindi compreso tra** $0,025$ **e** $0,05$. Quindi concludiamo che

- al livello di significatività $\alpha = 0,025$ l'ipotesi di indipendenza non può essere rifiutata
- ma che al livello di significatività $\alpha = 0,05$ l'ipotesi di indipendenza deve essere rifiutata.

Usando un *software* come per esempio EXCEL, possiamo anche calcolare il valore esatto del p-value. Come si desume dal grafico sottostante, quest'ultimo è dato da $p - value = 0,032$.



Esercizio 5.4

Nell'ambito di un indagine di mercato è stato intervistato un campione di $n = 500$ consumatori. A ciascun individuo è stato chiesto se preferisce il prodotto dell'azienda AAA, BBB, CCC oppure DDD. La tabella sottostante riassume i risultati dell'indagine:

	AAA	BBB	CCC	DDD	Totale
età $\leq$ 30 anni	110	70	50	30	270
età $>$ 30 anni	100	50	45	45	230
Totale	210	120	95	75	500

Sulla base di questi dati campionari si può rifiutare l'ipotesi nulla secondo la quale il prodotto preferito non dipende dall'età del consumatore? Si risponda determinando il p-value di un opportuno test statistico.

Il **modello di regressione lineare semplice** è la descrizione di un processo che genera i valori di due variabili X e Y.

In realtà questo modello non dice nulla sul modo in cui vengono generati i valori x_i della variabile X, ma specifica soltanto il modo in cui vengono generati i valori y_i della variabile Y. Secondo il modello di regressione lineare semplice questi ultimi sono infatti realizzazioni di variabili casuali $\widetilde{Y}_i$ che sono definite come

$$\widetilde{Y}_i = \beta_0 + \beta_1 x_i + \widetilde{\epsilon}_i$$

dove

- β_0 e β_1 sono due costanti reali,
- e dove le $\widetilde{\epsilon}_i$ sono variabili casuali i.i.d. con distribuzione normale di media $\mu = 0$ e varianza σ^2.

Chiaramente, le ipotesi del modello di regressione lineare semplice implicano che . . .

- le variabili casuali $\widetilde{Y}_i$ sono indipendenti (visto che le variabili casuali $\widetilde{\epsilon}_i$ sono indipendenti)
- il valore atteso delle variabili casuali $\widetilde{Y}_i$ è dato da (proprietà E2 del valore atteso)

$$\begin{aligned} E(\widetilde{Y}_i | X = x_i) &= E(\beta_0 + \beta_1 x_i + \widetilde{\epsilon}_i) \\ [[\text{ proprietà E2 }]] &= \beta_0 + \beta_1 x_i + E(\widetilde{\epsilon}_i) \\ &= \beta_0 + \beta_1 x_i \end{aligned}$$

- e la varianza delle variabili casuali $\widetilde{Y}_i$ è data da (proprietà V8 della varianza)

$$\begin{aligned} var(\widetilde{Y}_i | X = x_i) &= var(\beta_0 + \beta_1 x_i + \widetilde{\epsilon}_i) \\ [[\text{ proprietà V8 }]] &= var(\widetilde{\epsilon}_i) = \sigma^2. \end{aligned}$$

La funzione

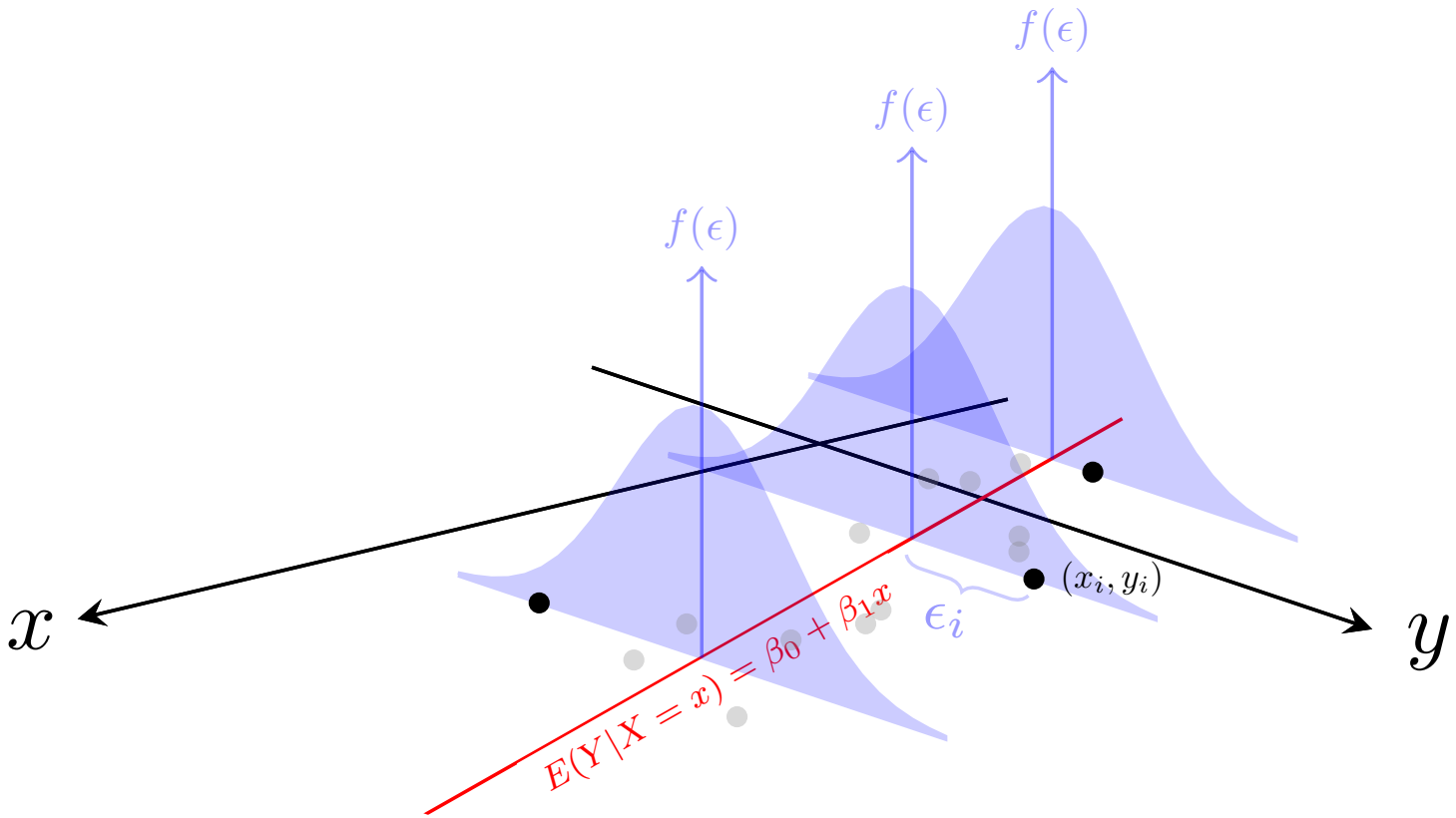
$$y = E(Y|X = x) = \beta_0 + \beta_1 x$$

che al variare di x restituisce le **medie condizionate di** Y viene chiamata **"retta di regressione"**[5] ...

...e siccome le variabili casuali $\widetilde{\epsilon}_i$ determinano delle deviazioni rispetto alla retta di regressione, spesso vengono chiamate **"termini d'errore"**.

[5] Per **media condizionata di** Y **dato un valore di** x intendiamo il valore atteso di una variabile casuale $\widetilde{Y}$ che è definita come $\widetilde{Y} = \beta_0 + \beta_1 x + \widetilde{\epsilon}_i$ dove $\widetilde{\epsilon}_i \sim Normale(\mu = 0, \sigma)$.

Il grafico sottostante mostra come un modello di regressione lineare semplice genera i valori y_i di una variabile Y:



Si noti che le distribuzioni normali dei termini d'errore $\widetilde{\epsilon}_i$ hanno tutte la stessa varianza σ^2.

Chiaramente, modificando i valori dei parametri β_0, β_1 e σ^2 si ottengono infiniti modelli di regressione lineare semplice diversi.

Il termine "modello di regressione lineare semplice" viene dunque utilizzato in due modi diversi:

1) per indicare la famiglia di tutti gli infiniti modelli di regressione lineare semplice che si possono ottenere variando i valori dei parametri β_0, β_1 e σ^2;
2) per indicare un particolare modello che si ottiene fissando i valori dei suddetti parametri.

Per fare esempi di variabili X e Y per le quali (in determinate popolazioni) le ipotesi del modello di regressione lineare semplice potrebbero essere realistiche, basti pensare alle coppie

- $X =$ statura, $Y =$ peso;
- $X =$ quantità utilizzata di un fattore produttivo, $Y =$ quantità prodotta;
- $X =$ spesa pubblicitaria; $Y =$ quantità venduta;
- ecc.

Si noti che in tutti questi esempi c'è motivo di sospettare che tra le variabili X e Y vi sia una **relazione causa-effetto**, ovvero che il valore assunto dalla variabile X abbia una qualche influenza su quello che assume la variabile Y.

Siccome tra i principali scopi del modello di regressione lineare semplice c'è proprio quello di descrivere relazioni causa-effetto, nell'ambito di un tale modello

- la variabile X viene chiamata **variabile indipendente** o **esplicativa**
- e la variabile Y viene invece chiamata **variabile dipendente**.

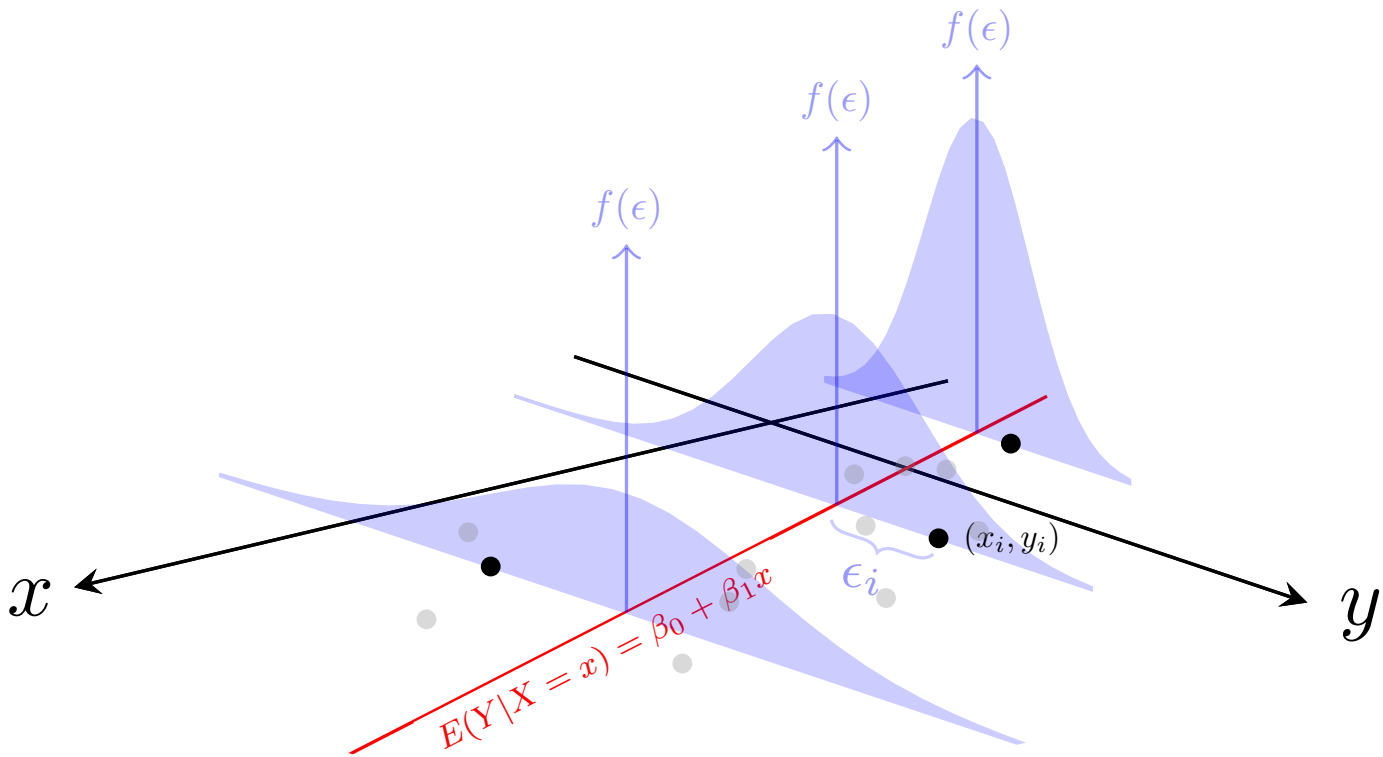
Si noti, tuttavia, che **il modello di regressione lineare semplice può essere utilizzato anche se tra X e Y non esiste nessuna relazione causa-effetto** (si pensi, per esempio, alla relazione tra l'età della sposa X e l'età dello sposo Y, alla relazione tra il consumo di gelati X e il consumo di creme solari Y, ecc).

Ovviamente, le ipotesi del modello di regressione lineare semplice forniscono soltanto una possibile descrizione del processo che genera i valori della variabile dipendente Y.

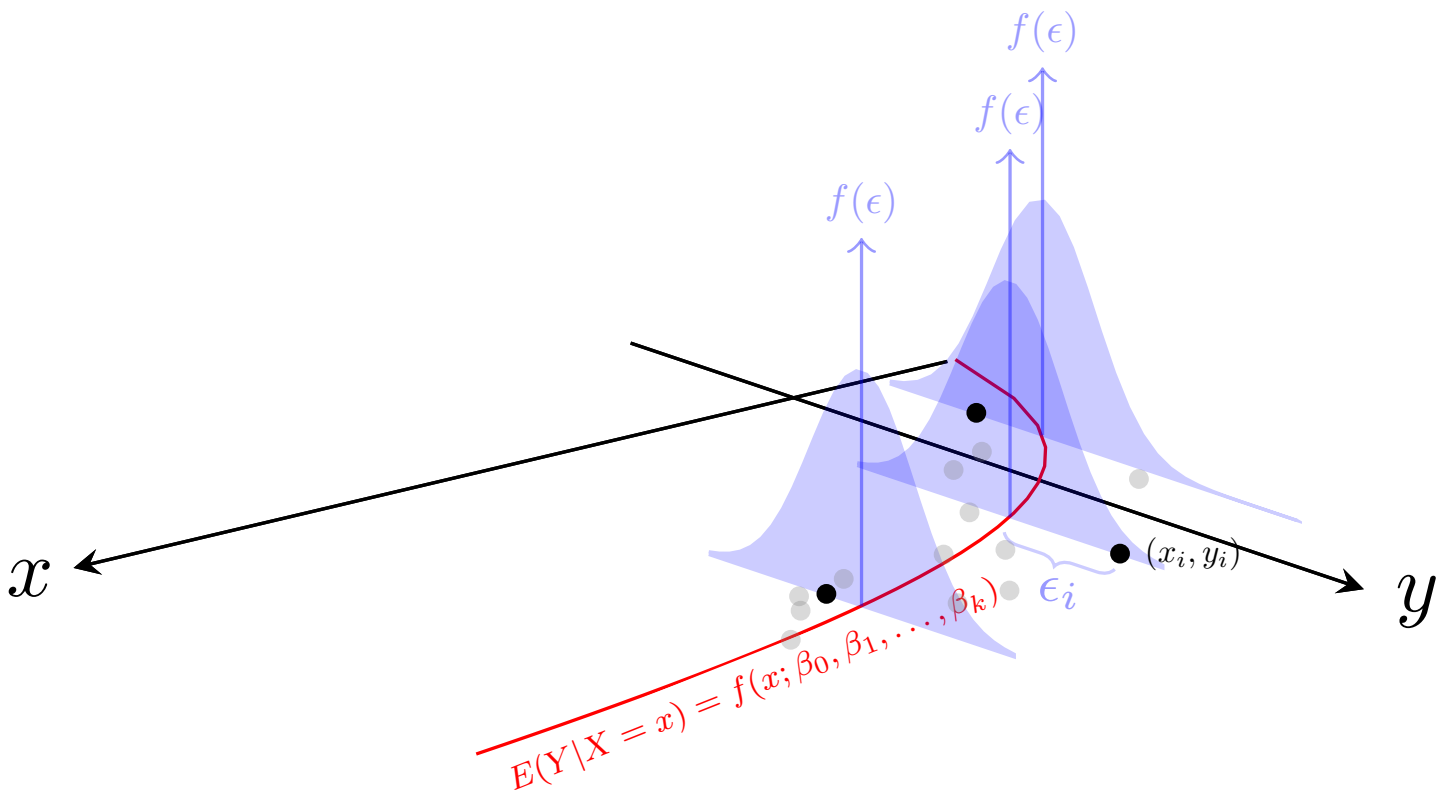
Spesso le ipotesi del modello di regressione lineare semplice non sono soddisfatte.

Per esempio, ...

. . . potrebbe darsi che le ipotesi del modello di regressione lineare semplice non siano soddisfatte perché la varianza σ^2 dei termini d'errore $\widetilde{\epsilon}_i$ aumenta all'aumentare di x (vedi il grafico sottostante)

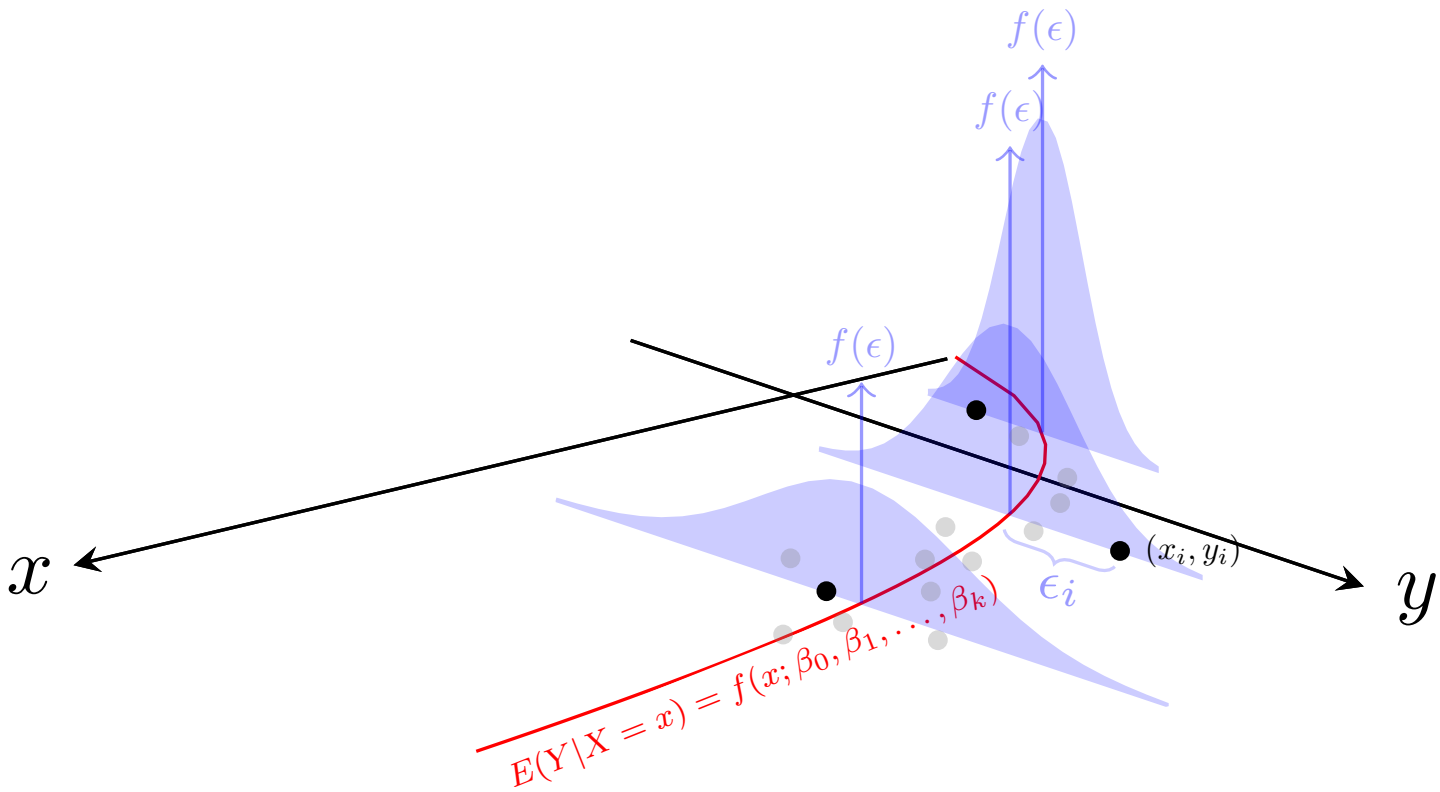


. . . oppure potrebbe darsi che le ipotesi del modello di regressione lineare semplice non siano soddisfatte perché la relazione tra la media condizionata di Y e i valori assunti dalla variabile X non è lineare (vedi il grafico sottostante), . . .



. . . oppure potrebbe anche darsi che le ipotesi del modello di regressione lineare semplice non siano soddisfatte perché

- la relazione tra le medie condizionate di Y e i valori assunti dalla variabile X non è lineare
- e la varianza σ^2 dei termini d'errore aumenta all'aumentare di x.



Si noti infine che oltre all'**ipotesi di linearità** e all'**ipotesi di costanza della varianza σ^2 (ipotesi di "omoschedasticità")**, potrebbe essere violata anche **l'ipotesi di normalità** e/o l'**ipotesi di indipendenza dei termini d'errore**.

Ovviamente, quando si osserva un campione composto da un certo numero n di coppie di osservazioni (x_i, y_i), non si è mai sicuri che siano state generate da un modello di regressione lineare semplice.

Per verificare se questa ipotesi fornisce una descrizione "sufficientemente realistica" dell'ignoto processo che genera le osservazioni di Y, si possono eseguire alcuni test grafici che ora descriveremo.

Di solito, il primo test che viene eseguito è quello basato sul **"grafico a dispersione"** (o **"scatter plot"**) che riporta i punti (x_i, y_i) del campione all'interno di un sistema di assi cartesiani e che a tali punti sovrappone la **retta ai minimi quadrati**.

La **regola decisionale (qualitativa)** in questo test grafico è ovvia e intuitiva: se la relazione di fondo tra le osservazioni delle variabili X e Y è palesemente non lineare, l'ipotesi di linearità delle medie condizionate $E(Y|X = x)$ viene rifiutata, ...

...e quindi si conclude che le ipotesi del modello di regressione lineare semplice non siano realistiche.

Promemoria 6.1 (La retta ai minimi quadrati)

Per definizione, la **retta ai minimi quadrati** che interpola n punti (x_i, y_i) è la retta

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

i cui parametri $b_0 = \widehat{\beta}_0$ e $b_1 = \widehat{\beta}_1$ minimizzano la **"distanza quadratica"**

$$D(b_0, b_1) = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2 .$$

[[$D(b_0, b_1)$ è la somma dei quadrati delle distanze verticali tra i punti (x_i, y_i) e una generica retta identificata dall'equazione $y = b_0 + b_1 x$.]]

Si può dimostrare che la distanza quadratica $D(b_0, b_1)$ ha un unico punto di minimo se e solo se $var(x_i) > 0$ (come sempre accade nelle applicazioni; altrimenti tutti i valori x_i sarebbero identici) e in tal caso il punto di minimo si trova in corrispondenza di

$$b_1 = \widehat{\beta}_1 = \frac{cov(x_i, y_i)}{var(x_i)} = \frac{codev(x_i, y_i)}{dev(x_i)} \quad \text{e} \quad b_0 = \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}.$$

Queste formule restituiscono dunque i parametri della retta ai minimi quadrati.

Promemoria (continua) (La retta ai minimi quadrati)

Per calcolare i valori dei parametri $\widehat{\beta}_1$ e $\widehat{\beta}_0$, di solito conviene calcolare

$$dev(x_i) = \sum_{i=1}^{n}(x_i - \overline{x})^2 \quad \text{e} \quad codev(x_i, y_i) = \sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})$$

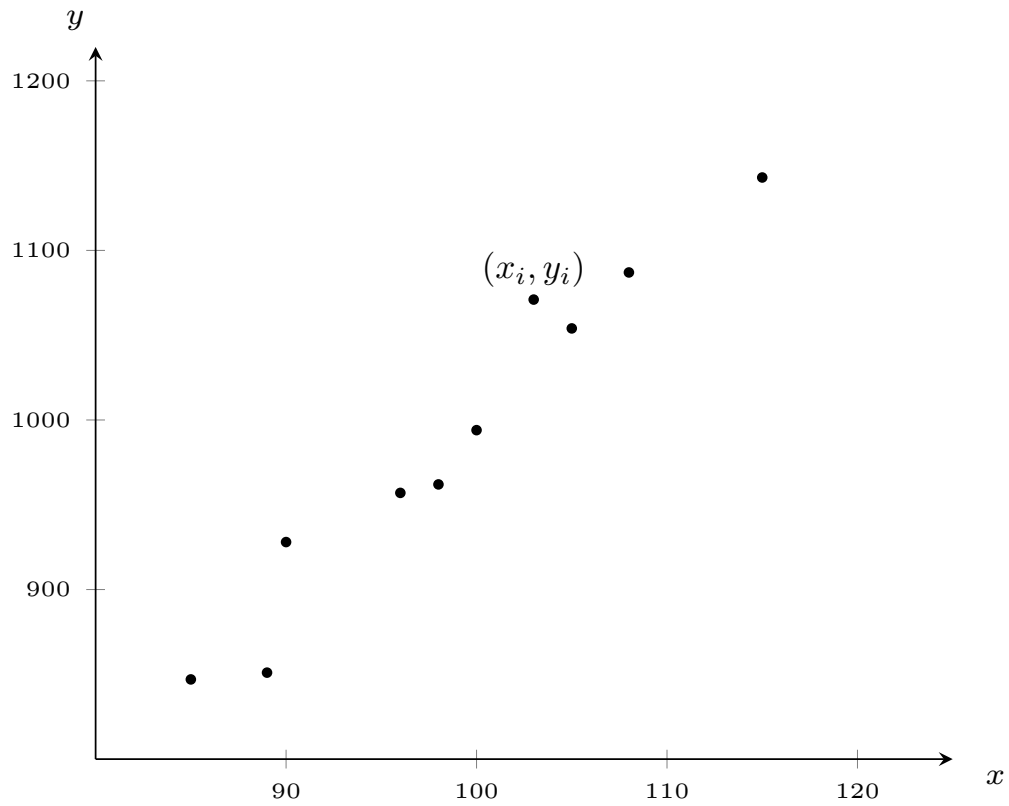
attraverso le **formule indirette**

$$dev(x_i) = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 \quad \text{e} \quad codev(x_i, y_i) = \sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}.$$

Per illustrare il **test basato sulla retta ai minimi quadrati**, consideriamo un campione di $n = 10$ appartamenti per i quali sono stati rilevati

- la **superfice** X (in mq)
- e il **consumo di gas metano** nel corso dell'ultimo anno (variabile Y con valori espressi in Smc).

x_i	y_i
100	994
90	928
105	1054
96	957
103	1071
115	1143
85	847
89	851
98	962
108	1087



Per ottenere i parametri della retta ai minimi quadrati costruiamo la tabella

Tabella 1:

x_i	y_i	x_i^2	$x_i y_i$	y_i^2
100	994	10000	99400	988036
90	928	8100	83520	861184
105	1054	11025	110670	1110916
96	957	9216	91872	915849
103	1071	10609	110313	1147041
115	1143	13225	131445	1306449
85	847	7225	71995	717409
89	851	7921	75739	724201
98	962	9604	94276	925444
108	1087	11664	117396	1181569
989	9894	98589	986626	9878098

. . . e usando i totali in fondo alle colonne otteniamo . . .

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{989}{10} = 98,9, \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{9894}{10} = 989,4,$$

$$dev(x_i) = \sum_{i=1}^{n}(x_i - \overline{x})^2 = [[\text{ formula indiretta }]] = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 =$$

$$= 98589 - 10 \times 98,9^2 = 776,9$$

$$codev(x_i, y_i) = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) =$$

$$= [[\text{ formula indiretta }]] = \sum_{i=1}^{n} x_i y_i - n\overline{xy} =$$

$$= 986626 - 10 \times 98,9 \times 989,4 = 8109,4.$$

A partire da questi indici statistici otteniamo agevolmente ...

. . . i parametri della retta ai minimi quadrati:

$$\widehat{\beta}_1 = \frac{codev(x_i, y_i)}{dev(x_i)} = \frac{8109,4}{776,9} = 10,438$$

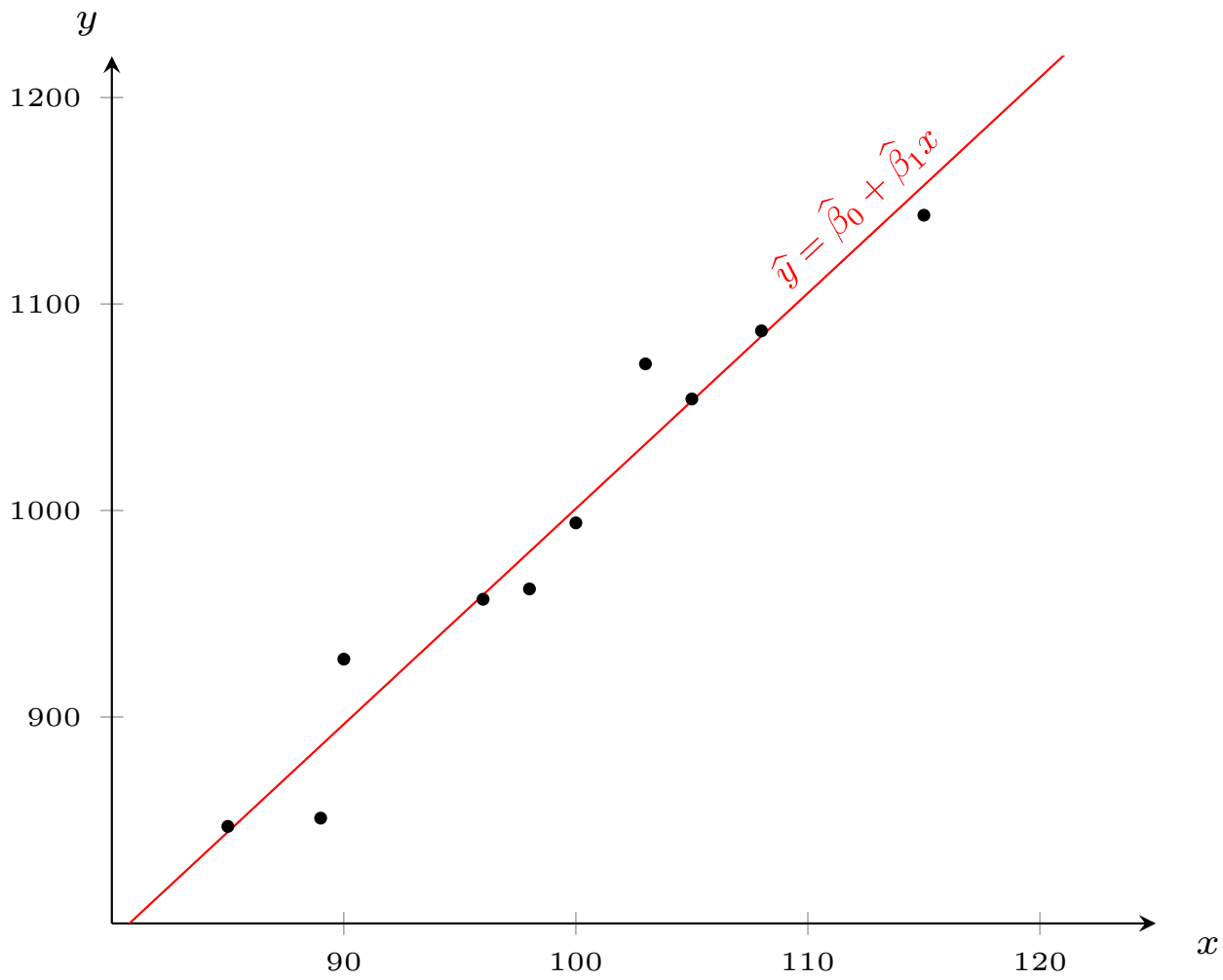
e

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x} = 989,4 - 10,438 \times 98,9 = -42,918.$$

L'equazione della retta ai minimi quadrati che interpola i punti (x_i, y_i) del campione è dunque data da

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x = -42,918 + 10,438x.$$

Il grafico sulla prossima *slide* mostra la retta ai minimi quadrati sovrapposta ai punti (x_i, y_i).


y
1200
1100
1000
900
90
100
110
120
x
$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$

Il grafico mostra che la relazione di fondo tra i valori delle variabili X e Y è lineare, . . .

. . . e sulla base di questo grafico non possiamo quindi certamente rifiutare l'ipotesi che le medie condizionate $E(Y|X = x)$ siano allineate su una retta (così come previsto dal modello di regressione lineare semplice).

Tuttavia, bisogna anche tenere presente che il grafico in questione è basato soltanto su un numero ridotto di osservazioni (x_i, y_i), . . .

. . . e che sulla base di un numero ridotto di osservazioni è molto difficile individuare violazioni dell'ipotesi di linearità delle medie condizionate $E(Y|X = x)$.

[[Più in generale, con un numero esiguo di osservazioni è difficile individuare qualsiasi violazione delle ipotesi del modello di regressione lineare semplice.]]

Di solito, al grafico a dispersione con la retta ai minimi quadrati viene anche aggiunto il valore dell'**indice di determinazione** onde valutare la bontà d'adattamento della retta ai minimi quadrati.

Ricordiamo che l'indice di determinazione può assumere soltanto valori compresi tra 0 e 1 e che ...

- l'indice di determinazione è prossimo a 0 se i punti (x_i, y_i) sono molto dispersi attorno alla retta ai minimi quadrati,
- mentre l'indice di determinazione è prossimo a 1 se i punti (x_i, y_i) sono molto prossimi alla retta ai minimi quadrati.

Valori elevati dell'indice di determinazione costituiscono dunque evidenza empirica a favore dell'ipotesi che le medie condizionate $E(Y|X = x)$ siano allineate su una retta (così come previsto dalle ipotesi del modello di regressione lineare semplice), . . .

. . . ma valori piccoli dell'indice di determinazione non sono necessariamente sintomatici per una violazione della suddetta ipotesi.

Promemoria 6.2 (Indice di determinazione)

L'indice di determinazione è definito come

$$I^2 = \frac{\text{devianza spiegata}}{\text{devianza totale}}$$

dove

- la **"devianza spiegata"** è la devianza dei **valori riprodotti**

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i,$$

- e dove la **"devianza totale"** è la devianza dei valori osservati y_i.

L'indice di determinazione

$$I^2 = \frac{\text{devianza spiegata}}{\text{devianza totale}}$$

può assumere soltanto valori compresi tra 0 e 1 perché

$$\text{devianza totale} = \text{devianza spiegata} + \text{devianza residua}$$

dove la **"devianza residua"** è la devianza dei **residui**

$$\widehat{\epsilon}_i = y_i - \widehat{y}_i.$$

Promemoria (continua) (Indice di determinazione)

Per calcolare l'indice di determinazione non è necessario calcolare i valori riprodotti

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i.$$

Infatti, si può dimostrare che l'indice di determinazione coincide con il quadrato del **coefficiente di correlazione lineare**

$$r(x_i, y_i) = \frac{cov(x_i, y_i)}{\sqrt{var(x_i)var(y_i)}} = \frac{codev(x_i, y_i)}{\sqrt{dev(x_i)dev(y_i)}},$$

ovvero che

$$I^2 = [r(x_i, y_i)]^2 = \frac{[cov(x_i, y_i)]^2}{var(x_i)var(y_i)} = \frac{[codev(x_i, y_i)]^2}{dev(x_i)dev(y_i)}.$$

Promemoria (continua) (Indice di determinazione)

Per l'interpretazione dell'indice di determinazione osserviamo che

$$I^2 = \frac{\text{devianza spiegata}}{\text{devianza totale}} = 1 - \frac{\text{devianza residua}}{\text{devianza totale}}$$

e siccome la media dei residui di una retta ai minimi quadrati è sempre nulla e la devianza residua è dunque data da

$$\text{devianza residua} = dev(\widehat{\epsilon_i}) = \sum_{i=1}^{n} (\widehat{\epsilon_i} - 0)^2 = \sum_{i=1}^{n} \widehat{\epsilon_i}^2,$$

possiamo concludere che l'indice di determinazione I^2 assume

- il valore 1 se e solo se tutti i residui $\widehat{\epsilon_i}$ sono nulli (si noti che questo è l'unico caso in cui la devianza residua è nulla), ovvero se e solo se la retta ai minimi quadrati passa per tutti i punti (x_i, y_i);
- il valore 0 se e solo se la devianza spiegata è nulla, ovvero se e solo se tutti i valori riprodotti $\widehat{y_i}$ sono uguali tra di loro e la retta ai minimi quadrati è dunque orizzontale (siccome la media dei valori riprodotti $\widehat{y_i}$ deve sempre essere uguale a $\overline{y}$, possiamo concludere che la retta ai minimi quadrati può essere orizzontale solo se il valore delle sue ordinate è $\overline{y}$);
- valori intermedi altrimenti.

Promemoria (continua) (Indice di determinazione)

Chiaramente, l'adattamento di una retta ai minimi quadrati è tanto migliore quanto più il suo indice di determinazione è prossimo a 1.

Si noti che l'indice di determinazione della retta ai minimi quadrati che spiega il consumo di gas metano in funzione della superficie degli appartamenti può essere calcolato a partire dai totali di colonna della Tabella 1:

Infatti, usando i totali di colonna della Tabella 1 abbiamo già calcolato i valori di

$$\overline{y} = 989,4, \quad dev(x_i) = 776,9 \quad \text{e di} \quad codev(x_i, y_i) = 8109,4,$$

e usando la somma dei quadrati di y_i che si trova in fondo all'ultima colonna (si noti che questa somma non ci è servita per calcolare i parametri della retta ai minimi quadrati), otteniamo anche

$$dev(y_i) = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2 = 9878098 - 10 \times 989,4^2 = 88974,4.$$

L'indice di determinazione è quindi dato da ...

$$I^2 = \frac{[codev(x_i, y_i)]^2}{dev(x_i)dev(y_i)} = \frac{8109,4^2}{776,9 \times 88974,4} = 0,951.$$

Per commentare il valore di questo indice possiamo dire che la retta ai minimi quadrati spiega il 95% della variabilità dei consumi di gas metano.

Chiaramente, questo valore dell'indice di determinazione è molto elevato e questo fatto suggerisce che l'ipotesi di linearità delle medie condizionate $E(Y|X = x)$ sia (almeno approssimativamente) soddisfatta.

Finora abbiamo visto come attraverso la retta ai minimi quadrati si possono scoprire violazioni dell'ipotesi di linearità delle medie condizionate $E(Y|X = x)$.

Per indagare sulle altre ipotesi del modello di regressione lineare semplice, si può fare riferimento al cosiddetto **"grafico dei residui"**.

Anche quest'ultimo è un grafico a dispersione, ma riporta i valori riprodotti

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i,$$

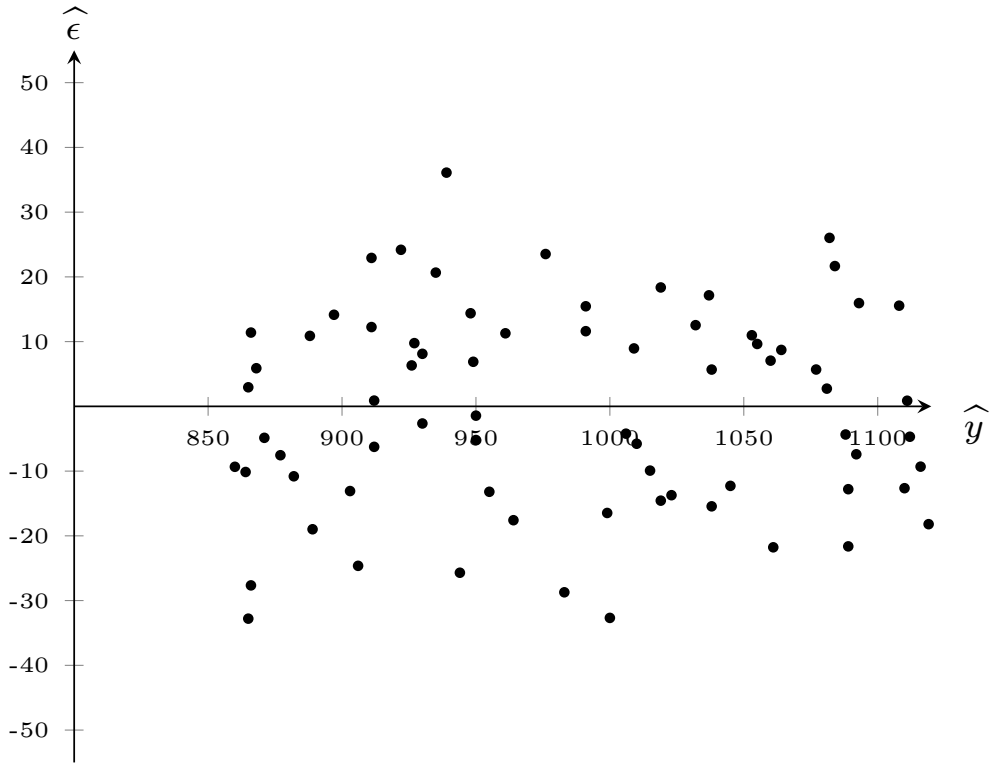
in ascissa, e i residui

$$\widehat{\epsilon}_i = y_i - \widehat{y}_i$$

in ordinata (a volte al posto dei valori riprodotti $\widehat{y}_i$ in ascissa vengono riportati i corrispondenti valori x_i).

Per farci un'idea di come appare il grafico dei residui quando le osservazioni (x_i, y_i) sono generate da un modello di regressione lineare semplice, possiamo fare riferimento al grafico sottostante:

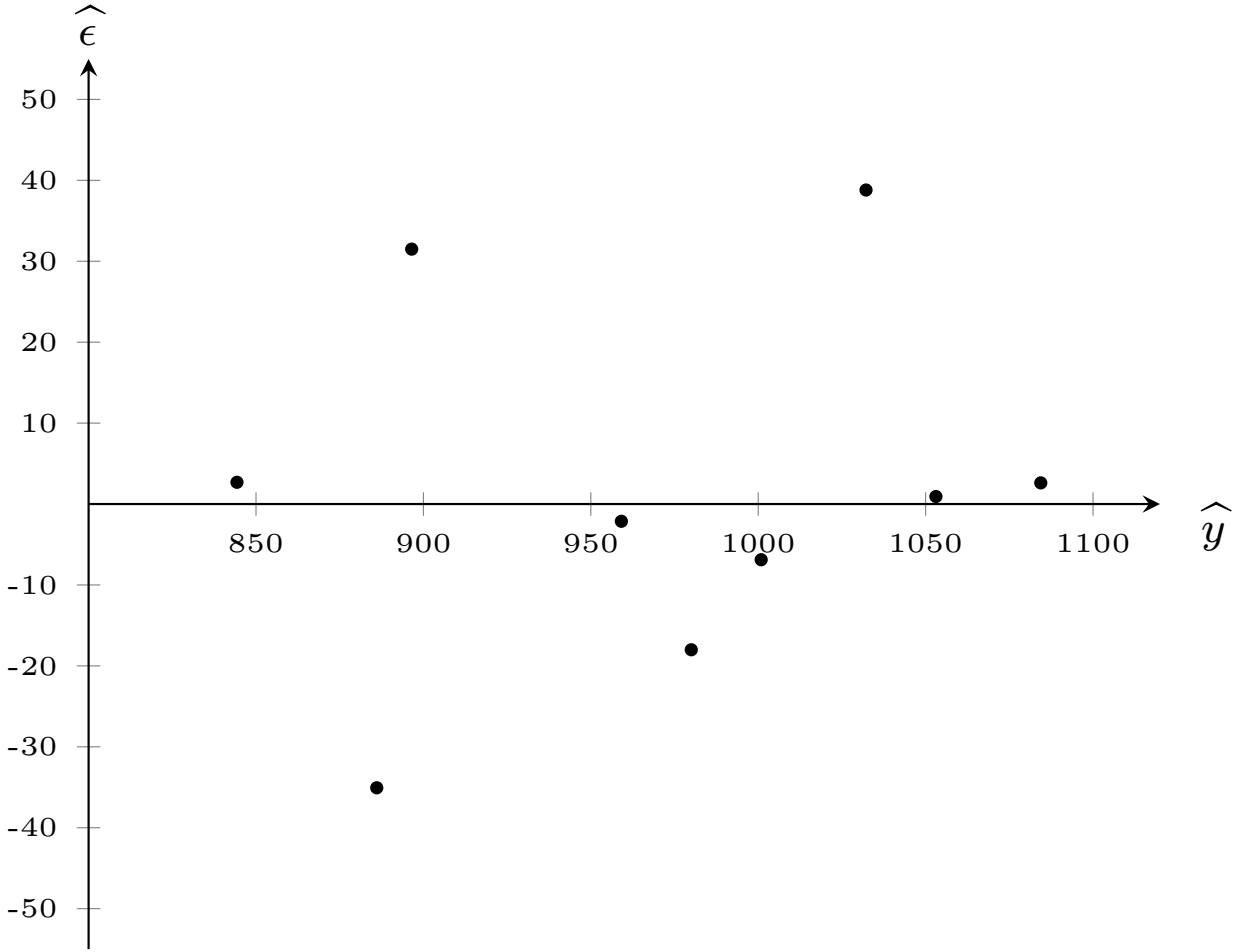
Figura 6.1:



Si noti che questo grafico non evidenzia l'esistenza di qualsiasi tipo di relazione tra i residui $\widehat{\epsilon}_i$ e i valori riprodotti $\widehat{y}_i$.

Il grafico in questa *slide* mostra invece il grafico dei residui che si ottiene a partire dalla retta ai minimi quadrati che descrive il consumo di gas metano Y in funzione della superficie X degli appartamenti:

Figura 6.2:



I principali **test che si riferiscono al grafico dei residui** sono qui di seguito descritti:

1) Il primo test è basato sulla **successione dei segni dei residui**. Se le ipotesi del modello di regressione lineare sono soddisfatte, la successione dei segni dovrebbe essere (con elevata probabilità) del tutto casuale (così come in Figura 6.1): se muovendosi da sinistra verso destra sul grafico dei residui si riscontrasse qualche tipo di **sistematicità nella successione dei segni dei residui**, questo potrebbe indicare due tipi di violazione delle ipotesi del modello di regressione lineare semplice:
   - potrebbe indicare che **le medie condizionate** $E(Y|X = x)$ **non siano allineate su una retta**,
   - e/o potrebbe indicare che **i termini d'errore non siano indipendenti**.

   Individuare la causa di un andamento sistematico dei segni dei residui non è sempre facile: a volte un andamento sistematico può essere spiegato attraverso semplici considerazioni sul processo che ha generato i dati del campione e/o attraverso una qualche teoria. Se ciò non fosse il caso, si può tentare di individuare la causa acquisendo ulteriori osservazioni delle variabili X e Y (ammesso che ciò sia possibile).

Nel grafico dei residui che abbiamo ottenuto con i dati sulle superfici degli appartamenti e sui corrispondenti consumi di gas metano (vedi Figura 6.2) la successione dei segni sembra casuale così come dovrebbe essere quando il processo che genera le osservazioni (x_i, y_i) soddisfa le ipotesi del modello di regressione lineare semplice.

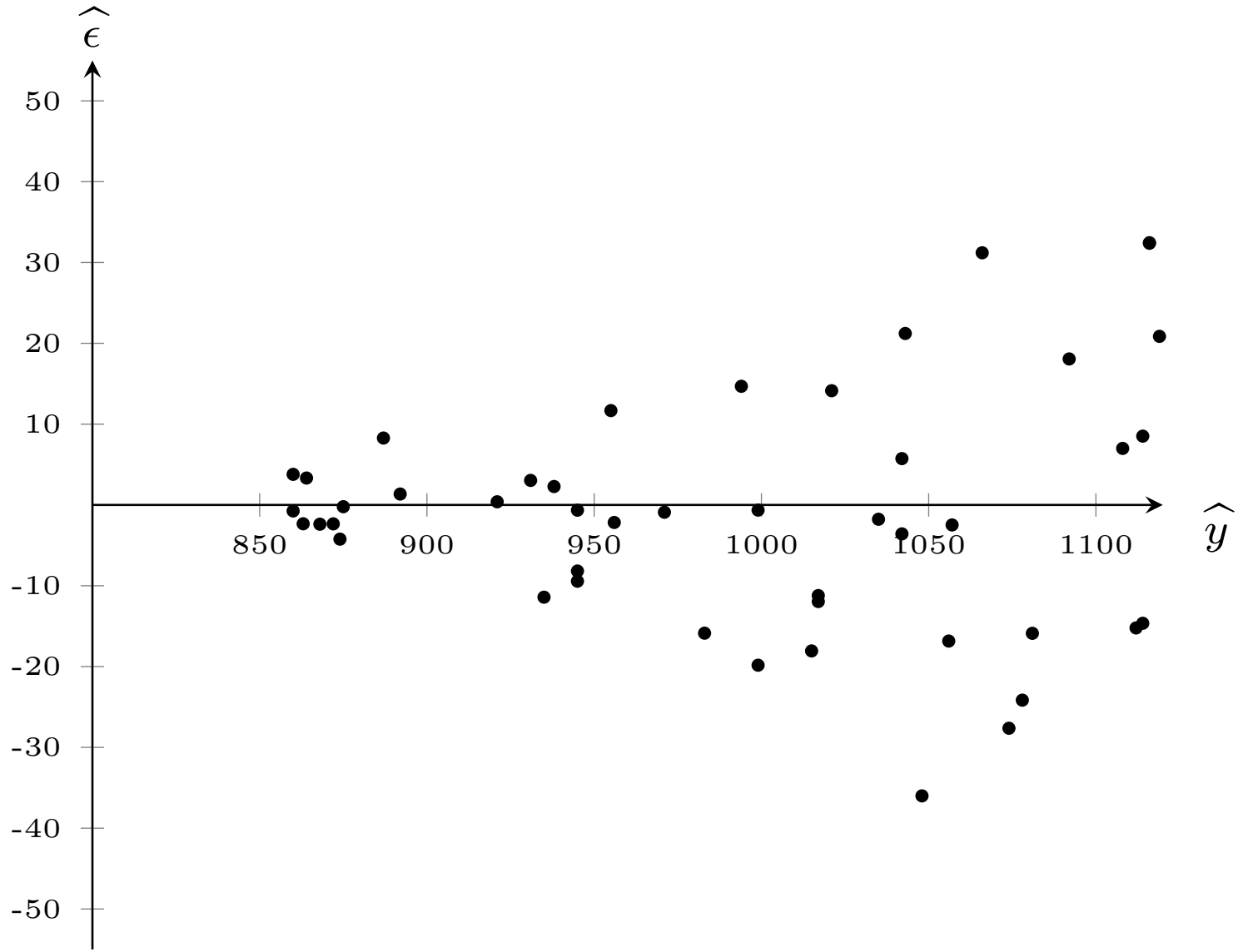
Ricordiamo, tuttavia, che il grafico dei residui in questione è basato su un numero molto esiguo di osservazioni (x_i, y_i) e che con un numero piccolo di osservazioni è molto difficile scoprire violazioni delle ipotesi del modello di regressione lineare semplice.

2) Un altro importante test basato sul grafico dei residui ha lo scopo di individuare la presenza di **"heteroschedasticità"**, ovvero di violazioni dell'ipotesi che i termini d'errore ϵ_i abbiano tutti la medesima varianza σ^2 (violazioni dell'ipotesi di "omoschedasticità").

Anche questo test è puramente visivo: se le ipotesi del modello di regressione lineare semplice sono soddisfatte, allora (con elevata probabilità) non dovrebbe esserci nessuna relazione tra i **valori assoluti dei residui** e i valori riprodotti $\widehat{y}_i$ che sono riportati sull'ascissa. Qualora il grafico dei residui mostrasse una tale relazione, si dovrebbe concludere che la varianza dei termini d'errore non è costante.

Il grafico dei residui che abbiamo ottenuto con i dati sulle superfici degli appartamenti e sui corrispondenti consumi di gas metano non è indicativo di una situazione di heteroschedasticità (vedi Figura 6.2), ma (come abbiamo già osservato) quel grafico è basato soltanto su un numero esiguo di osservazioni e con un numero esiguo di osservazioni è sempre difficile scoprire violazioni delle ipotesi del modello di regressione lineare semplice.

Il grafico sulla prossima *slide* mostra invece come potrebbe presentarsi il grafico dei residui in una situazione di heteroschedasticità.


$\widehat{\epsilon}$
50
40
30
20
10
-10
-20
-30
-40
-50
850
900
950
1000
1050
1100
$\widehat{y}$

3) L'ultimo test sul grafico dei residui che presentiamo in queste *slides* ha lo scopo di individuare **valori anomali** (cosiddetti ***"outliers"***).

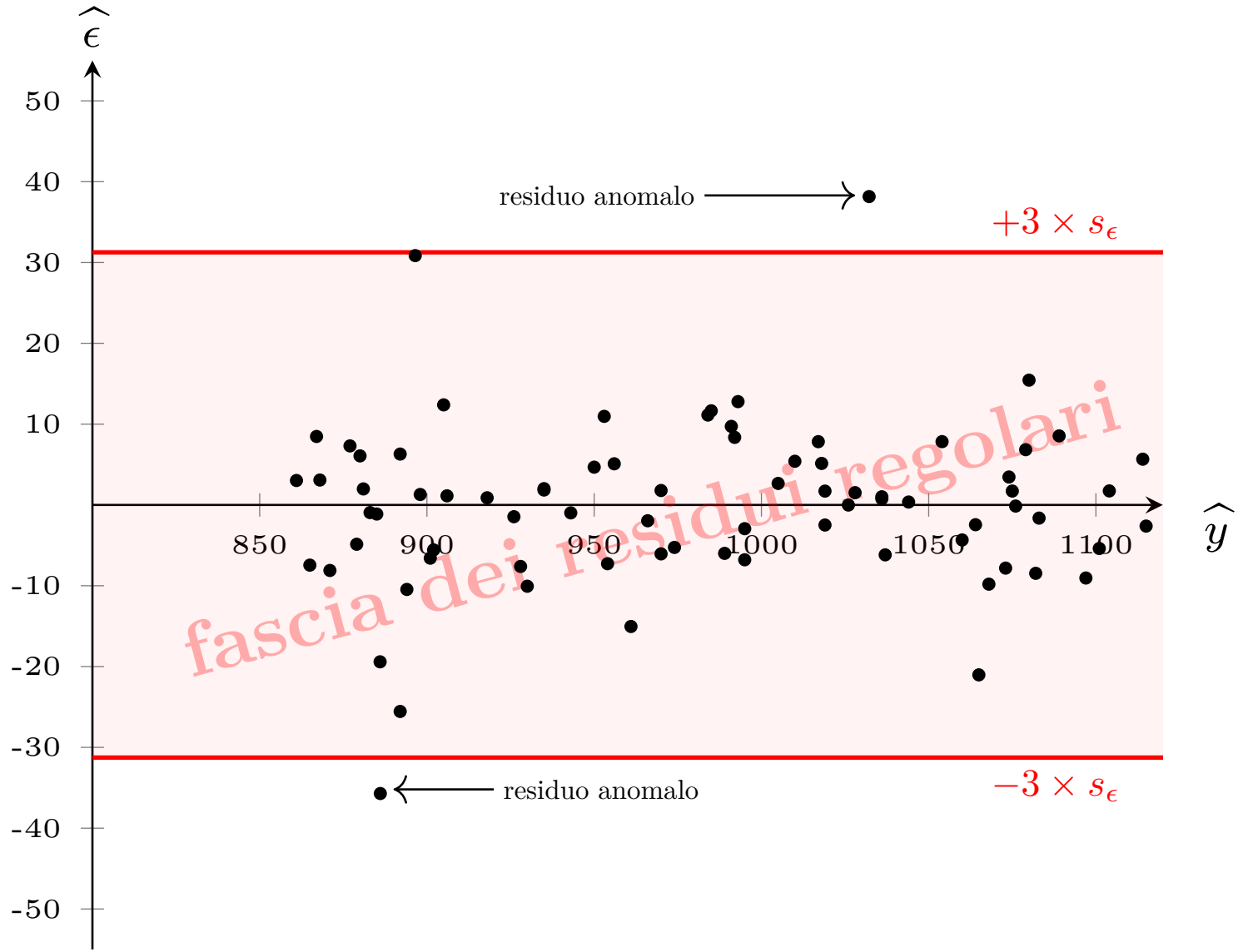
Per eseguire questo test si deve calcolare il valore di

$$s_\epsilon = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\widehat{\epsilon}_i^2} = \sqrt{\frac{dev(\widehat{\epsilon}_i)}{n-2}} = \sqrt{\frac{\text{devianza residua}}{n-2}}$$

(che viene solitamente utilizzato come stima puntuale per l'ignoto sqm σ dei termini d'errore ϵ_i) ...

... e aggiungere al grafico dei residui due rette orizzontali che intersecano l'asse delle ordinate in corrispondenza di $\pm 3 \times s_\epsilon$.

In questo modo si ottiene una fascia che contiene tutti i punti $(\widehat{y}_i, \widehat{\epsilon}_i)$ che possono essere considerati "regolari" (vedi il grafico sulla prossima slide).


$\widehat{\epsilon}$
50
40
30
20
10
-10
-20
-30
-40
-50
850
900
950
1000
1050
1100
$\widehat{y}$
residuo anomalo
$+3 \times s_\epsilon$
fascia dei residui regolari
residuo anomalo
$-3 \times s_\epsilon$

Se ci fossero dei punti $(\widehat{y}_i, \widehat{\epsilon}_i)$ che sono molto lontani da questa fascia, allora le corrispondenti coppie (x_i, y_i) dovrebbero essere considerate **"anomale"**.

In presenza di coppie (x_i, y_i) anomale, si dovrebbe indagare sulle cause che le hanno generate.

Di solito, se esistono delle coppie (x_i, y_i) anomale, si procede alla loro eliminazione e si ripetono tutti i test grafici onde eliminare l'influenza di questi punti.

Nel caso della retta ai minimi quadrati che descrive il consumo di gas metano Y in funzione della superficie dell'appartamento X otteniamo

$$dev(\widehat{\epsilon}_i) = dev(y_i) \times \left(1 - I^2\right) = 88974,4 \times (1 - 0,951) = 4359,7 \tag{1}$$

[[la formula $dev(\widehat{\epsilon}_i) = dev(y_i) \times \left(1 - I^2\right)$ può essere facilmente ottenuta a partire dalla formula $I^2 = 1 - \frac{dev(\widehat{\epsilon}_i)}{dev(y_i)}$]]

. . . e il valore di s_ϵ è quindi dato da

$$s_\epsilon = \sqrt{\frac{dev(\widehat{\epsilon}_i)}{n-2}} = \sqrt{\frac{4359,7}{10-2}} = 23,344 \tag{2}$$

Siccome le due rette orizzontali

$$\widehat{\epsilon} = \pm 3 \times s_\epsilon = \pm 3 \times 23,344 = \pm 70,032$$

si trovano ampiamente al di fuori dell'area mostrata nel grafico dei residui in Figura 6.2, possiamo concludere che nessuno dei punti (x_i, y_i) sia anomalo.

Vediamo ora invece come sulla base di un campione di n coppie di osservazioni (x_i, y_i) si possono stimare i parametri del modello di regressione lineare semplice.

D'ora in poi **assumeremo quindi che le n coppie (x_i, y_i) del campione siano effettivamente state generate da un modello di regressione lineare semplice con parametri β_0, β_1 e σ^2 che sono tutti ignoti!!!**

In realtà, abbiamo già introdotto le stime puntuali che solitamente vengono utilizzate:

- per stimare il coefficiente angolare β_1 e l'intercetta β_0 della retta di regressione (ovvero della retta che restituisce le medie condizionate $E(Y|X = x)$), di solito si utilizzano il coefficiente angolare

$$\widehat{\beta}_1 = \frac{codev(x_i, y_i)}{dev(x_i)}$$

e l'intercetta

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$$

della retta ai minimi quadrati;

- per stimare la varianza σ^2 comune a tutti i termini d'errore ϵ_i, di solito si utilizza la statistica

$$s^2_\epsilon = \frac{dev(\widehat{\epsilon}_i)}{n-2} = \frac{\text{devianza residua}}{n-2}$$

che abbiamo introdotto nel test per individuare la presenza di valori anomali (vedi formula (2)).

Praticamente tutti i ***software*** che implementano funzioni statistiche contengono una funzione che calcola le suddette stime dei parametri del modello di regressione lineare semplice.

Per presentare

- i principali test statistici sui parametri del modello di regressione lineare semplice
- e le formule per calcolare gli estremi di intervalli di confidenza,

conviene dunque partire dall'*output* di un tale *software* . . .

In virtù dell'ampia diffusione di EXCEL, considereremo l'*output* di quest'ultimo. Con i dati sulla superficie degli appartamenti e i corrispondenti consumi di gas metano EXCEL produce le seguenti tabelle:

OUTPUT RIEPILOGO						
Statistica della regressione						
R multiplo	0,97537951					
R al quadrato	0,951365188					
R al quadrato corretto	0,945285837					
Errore standard	23,25739986					
Osservazioni	10					
ANALISI VARIANZA						
	gdl	*SQ*	*MQ*	*F*	*Significatività F*	
Regressione	1	84647,14681	84647,14681	156,4912302	1,56054E-06	
Residuo	8	4327,253186	540,9066482			
Totale	9	88974,4				
	Coefficienti	*Errore standard*	*Stat t*	*Valore di significatività*	*Inferiore 95%*	*Superiore 95%*
Intercetta	-42,93319604	82,85006047	-0,518203557	0,618339365	-233,9857781	148,119386
x	10,43815163	0,834408268	12,50964549	1,56054E-06	8,514002711	12,36230055

La prima tabella (che per comodità riportiamo ancora una volta)

Statistica della regressione	
R multiplo	0,97537951
R al quadrato	0,951365188
R al quadrato corretto	0,945285837
Errore standard	23,25739986
Osservazioni	10

contiene

- il valore dell'indice di determinazione della retta ai minimi quadrati (***"R al quadrato"***)
- e il valore della stima puntuale $s_\epsilon = \sqrt{\frac{dev(\widehat{\epsilon_i})}{n-2}}$ (***"Errore standard"***).

[[Il valore di ***"R multiplo"*** è semplicemente la radice (positiva) dell'indice di determinazione, mentre il valore di ***"R al quadrato corretto"*** è una versione "corretta" dell'indice di determinazione che tiene anche conto del numero di variabili esplicative (nel nostro caso soltanto una, ma la funzione di regressione di excel può essere applicata anche per stimare i parametri di modelli di regressione **multipli**, ovvero di modelli di regressione che tengono conto di più di una variabile esplicativa).]]

Consideriamo ora invece la seconda tabella:

ANALISI VARIANZA					
	gdl	*SQ*	*MQ*	*F*	*Significatività F*
Regressione	1	84647,14681	84647,14681	156,4912302	1,56054E-06
Residuo	8	4327,253186	540,9066482		
Totale	9	88974,4			

- La colonna ***"SQ"*** questa tabella contiene i valori della

$$\text{devianza spiegata} = dev(\widehat{y}_i) = 84647,14681,$$
$$\text{devianza residua} = dev(\widehat{\epsilon}_i) = 4327,253186,$$

e della loro somma, ovvero della

$$\text{devianza totale} = dev(y_i) = 88974,4$$

Si noti che il valore della devianza residua riportato da EXCEL è leggermente diverso dal valore $dev(\widehat{\epsilon}_i) = [[\text{ formula } (1) \text{ }]] = 4359,7$ che abbiamo calcolato in precedenza. La differenza è dovuta soltanto agli arrotondamenti.

- La colonna ***"F"*** contiene invece il valore della **statistica test F** che viene utilizzata nel cosiddetto **"test sulla significatività della regressione"** (che descriveremo tra breve).
- La colonna ***"Significatività F"*** contiene invece il p-value del suddetto test.

Il **"test sulla significatività della regressione"** al quale abbiamo appena accennato, è un importante **test statistico per verificare se la variabile esplicativa X ha effettivamente un qualche potere esplicativo**.

Il test "test sulla significatività della regressione" verte quindi sul valore del coefficiente angolare β_1 della retta di regressione: chiaramente, se X non ha nessun potere esplicativo rispetto a Y, il valore di β_1 deve essere nullo; altrimenti il valore di β_1 deve essere diverso da zero.

L'ipotesi nulla e l'ipotesi alternativa del "test sulla significatività della regressione" sono dunque definite come

$$H_0 : \beta_1 = 0 \quad \text{e} \quad H_a : \beta_1 \neq 0.$$

Chiaramente, se è vera l'ipotesi nulla $H_0 : \beta_1 = 0$, allora la variabilità dei valori

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \beta_0 + \epsilon_i$$

può essere dovuta soltanto ai termini d'errore ϵ_i (e non alla variabilità dei valori x_i) e per questo motivo, con riferimento alla retta ai minimi quadrati, ci aspettiamo che i valori riprodotti

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

siano pressoché costanti e che i residui

$$\widehat{\epsilon}_i = y_i - \widehat{y}_i$$

presentino una variabilità molto più elevata.

In altre parole, se è vera l'ipotesi nulla $H_0 : \beta_1 = 0$, ci aspettiamo che la scomposizione della devianza

$$dev(y_i) = dev(\widehat{y}_i) + dev(\widehat{\epsilon}_i)$$

dia luogo ad un valore molto piccolo della devianza spiegata $dev(\widehat{y}_i)$ e ad un valore elevato della devianza residua $dev(\widehat{\epsilon}_i)$.

In virtù di questo fatto possiamo concludere che

- valori elevati della statistica test

$$F = \frac{dev(\widehat{y}_i)}{dev(\widehat{\epsilon}_i)/(n-2)} = \frac{dev(\widehat{y}_i)}{s^2_\epsilon}$$

  forniscano evidenza empirica a favore dell'ipotesi alternativa $H_a : \beta_1 \neq 0$,
- e che valori piccoli della suddetta statistica test F siano invece tipici per situazioni dove è vera $H_0 : \beta_1 = 0$.

Siccome **sotto l'ipotesi nulla $H_0 : \beta_1 = 0$ la statistica test F segue la distribuzione F di Fisher-Snedecor con 1 gdl al numeratore e $n-2$ gdl al denominatore**, possiamo definire una **regola decisionale** con associato livello di significatività α imponendo che l'ipotesi nulla $H_0 : \beta_1 = 0$ venga rifiutata se il valore della statistica test F supera il percentile di ordine $1-\alpha$ della suddetta distribuzione F di Fisher-Snedecor.

Confrontando il valore della statistica test

$$F = \frac{dev(\widehat{y}_i)}{dev(\widehat{\epsilon}_i)/(n-2)} = \frac{dev(\widehat{y}_i)}{s_\epsilon^2}$$

con i **percentili della distribuzione F di Fisher-Snedecor con 1 gdl al numeratore e $n-2$ gdl** (alcuni dei quali sono reperibili nella relativa tavola), ...

...si può quindi stabilire se per un prefissato valore del livello di significatività α l'ipotesi nulla debba essere rifiutata o meno ...

...e/o si può determinare il p-value associato al test in questione (ovviamente, attraverso la tavola della distribuzione di Fisher-Sndedcor si può di solito determinare soltanto un intervallo di valori per il p-value, ma con l'ausilio di un *software* come EXCEL si può ottenere il p-value "esatto").

Per verificare che la colonna ***"F"*** della tabella

ANALISI VARIANZA					
	gdl	*SQ*	*MQ*	*F*	*Significatività F*
Regressione	1	84647,14681	84647,14681	156,4912302	1,56054E-06
Residuo	8	4327,253186	540,9066482		
Totale	9	88974,4			

riporti effettivamente il valore della statistica test

$$F = \frac{dev(\widehat{y}_i)}{dev(\widehat{\epsilon}_i)/(n-2)}$$

basta notare che

$$F = \frac{dev(\widehat{y}_i)}{dev(\widehat{\epsilon}_i)/(n-2)} = \frac{84647,14681}{4327,253186/(10-2)} = 156,4912302$$

coincide effettivamente con il valore riportato nella colonna ***"F"***.

Confrontando il valore di $F = 156,4912302$ con i percentili della distribuzione F di Fisher-Snedecor con 1 gdl al numeratore e $n - 2 = 10 - 2 = 8$ gdl al denominatore, possiamo inoltre verificare che il valore di $F = 156,4912302$ è di gran lunga maggiore del 99-esimo percentile della suddetta distribuzione, ovvero del valore di $F_{1;8;0,01} = 11,26$ che troviamo nella tavola della distribuzione F di Fisher-Snedecor. ...

...Da questa osservazione deduciamo che il p-value del test F sulla significatività della regressione deve essere molto più piccolo di $0,01$, ma con il solo ausilio della tavola non possiamo determinare il valore esatto del p-value. ...

Tuttavia, il valore esatto del suddetto p-value è riportato nella colonna ***"Significatività F"*** della tabella

ANALISI VARIANZA					
	gdl	*SQ*	*MQ*	*F*	*Significatività F*
Regressione	1	84647,14681	84647,14681	156,4912302	1,56054E-06
Residuo	8	4327,253186	540,9066482		
Totale	9	88974,4			

dalla quale deduciamo che

$$\text{p-value} = \textbf{\textit{"Significatività F"}} = 1,56054E-06 = 0,00000156054.$$

Chiaramente, con un p-value così piccolo non si può che rifiutare l'ipotesi nulla $H_0 : \beta_1 = 0$ e quindi possiamo dire che . . .

. . . ***i dati campionari forniscono evidenza empirica molto netta a favore dell'ipotesi che le superfici degli appartamenti abbiano una qualche capacità esplicativa con riferimento ai consumi di gas metano!!!***

Per completare la descrizione della seconda tabella dell'*output* di EXCEL, tabella che per comodità riportiamo ancora una volta,

ANALISI VARIANZA					
	gdl	*SQ*	*MQ*	*F*	*Significatività F*
Regressione	1	84647,14681	84647,14681	156,4912302	1,56054E-06
Residuo	8	4327,253186	540,9066482		
Totale	9	88974,4			

dobbiamo ancora commentare le colonne ***"gdl"*** e ***"MQ"***.

- La colonna ***"gdl"*** riporta i gdl che identificano la distribuzione F di Fisher-Snedecor alla quale si deve fare riferimento per il test F sulla significatività della regressione:
  - Nella riga ***"Regressione"*** troviamo il numero di gdl al numeratore che, si noti, è uguale al numero di variabili esplicative[6], ...
  - ...mentre nella riga ***"Residuo"*** troviamo il numero di gdl al denominatore, ovvero la differenza tra il numero n di coppie (x_i, y_i) considerate nell'analisi e il numero di parametri che servono per identificare l'ignota retta di regressione (ovvero 2 perché la retta di regressione è identificata dai due parametri β_0 e β_1).

[6] Per definizione, nel caso di un modello di regressione lineare semplice il numero di variabili esplicative è sempre uguale a 1, ma la funzione di EXCEL che stima i parametri del modello di regressione lineare semplice può essere utilizzata anche con più di una variabile esplicativa e in tal caso restituisce le stime del modello di regressione "multiplo"

- La colonna ***"MQ"*** riporta invece nella riga ***"Regressione"*** il valore del rapporto

$$\frac{\text{devianza spiegata}}{\text{numero variabili esplicative}} = \frac{84647,14681}{gdl = 1} = 84647,14681,$$

e nella riga ***"Residuo"*** il valore del rapporto

$$\frac{\text{devianza residua}}{\text{n-2}} = \frac{4327,253186}{gdl = 8} = 540,9066482,$$

ovvero il valore della stima puntuale s^2_ϵ.

Si noti che il valore della statistica test

$$F = \frac{\text{devianza spiegata}}{\text{devianza residua}/(n-2)} = \frac{dev(\widehat{y}_i)}{dev(\widehat{\epsilon}_i)/(n-2)}$$

è uguale al rapporto tra i due valori riportati nella colonna ***"MQ"***.

A questo punto dobbiamo ancora commentare la terza e ultima tabella dell'*output* di EXCEL, ovvero la tabella

	Coefficienti	*Errore standard*	*Stat t*	*Valore di significatività*	*Inferiore 95%*	*Superiore 95%*
Intercetta	-42,93319604	82,85006047	-0,518203557	0,618339365	-233,9857781	148,119386
x	10,43815163	0,834408268	12,50964549	1,56054E-06	8,514002711	12,36230055

Questa tabella riporta le stime puntuali per gli ignoti valori dei parametri β_0 e β_1 che identificano la retta di regressione, nonché alcuni risultati inferenziali che si riferiscono a questi due parametri.

La prima riga si riferisce all'intercetta β_0 della retta di regressione:

- Nella colonna ***"Coefficienti"*** troviamo il valore della stima puntuale

$$\widehat{\beta}_0 = -42,93319604$$

che abbiamo già calcolato in precedenza.

- Nella colonna ***"Errore standard"*** troviamo il valore di una stima puntuale $s_{\widehat{\beta}_0}$ per lo scarto quadratico medio dello stimatore che ha generato la stima puntuale $\widehat{\beta}_0 = -42,93319604$.

  La formula per calcolare $s_{\widehat{\beta}_0}$ è data da

$$s_{\widehat{\beta}_0} = s_\epsilon \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{dev(x_i)}}$$

[[vedi risultati calcolati in precedenza]]

$$= 23,344 \times \sqrt{\frac{1}{10} + \frac{98,9^2}{776,9}} = 83,159 \simeq 82,85006047$$

(la differenza è dovuta soltanto agli arrotondamenti).

- Nella colonna ***"Stat t"*** troviamo il valore della statistica test

$$t = \frac{\widehat{\beta}_0}{s_{\widehat{\beta}_0}} = \frac{\widehat{\beta}_0}{s_\epsilon \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{dev(x_i)}}}$$

ovvero il valore di

$$t = \frac{-42,93319604}{82,85006047} = -0,518203557.$$

Questa statistica test t viene di solito utilizzata nei test d'ipotesi dove

$$H_0 : \beta_0 = 0 \quad \text{e} \quad H_a : \beta_0 \neq 0.$$

L'ipotesi nulla $H_0 : \beta_0 = 0$ viene rifiutata al livello di significatività α se il valore di $|t|$ (il modulo della suddetta statistica test t) supera il valore di $t_{n-2;\alpha/2}$. ...

. . . Usando la tavola della distribuzione t si Student vediamo che nell'esempio in questione si ha

$$|t| = |-0,518203557| < t_{n-2=10-2=8;0,20} = 0,889$$

e che il p-value è dunque maggiore di $2 \times 0,20 = 0,40$.

Il valore esatto del suddetto p-value è contenuto nella colonna ***"Valore di significatività"*** della tabella EXCEL ed è dato da

$$\text{p-value} = 0,618339365.$$

Chiaramente, con un p-value così elevato l'ipotesi nulla $H_0 : \beta_0 = 0$ non può essere rifiutata.

- Nelle colonne ***"Inferiore 95%"*** e ***"Superiore 95%"*** troviamo infine gli estremi di un intervallo di confidenza al 95% per l'ignoto valore di β_0.

  Le formule per calcolare gli estremi di un IdC per β_0 con generico livello di confidenza $1-\alpha$ sono date da

$$\widehat{\beta}_0 \pm t_{n-2;\alpha/2} \times s_{\widehat{\beta}_0} = \widehat{\beta}_0 \pm t_{n-2;\alpha/2} \times s_\epsilon \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{dev(x_i)}}.$$

  Per l'esempio sui consumi di gas metano possiamo facilmente verificare che gli estremi dell'IdC al 95% sono dati da (dalla tavola della distribuzione t di Student desumiamo che $t_{10-2;0,025} = 2,306$; i valori di $\widehat{\beta}_0$ e di $s_{\widehat{\beta}_0}$ li troviamo nelle colonne ***"Coefficienti"*** e ***"Errore standard"***)

$$-42,93319604 \pm 2,306 \times 82,85006047 = \begin{cases} 148,1190434 \\ -233,9854355 \end{cases}$$

  così come riportato nelle colonne ***"Inferiore 95%"*** e ***"Superiore 95%"***.

Vediamo ora invece il contenuto della seconda e ultima riga della tabella

	Coefficienti	*Errore standard*	*Stat t*	*Valore di significatività*	*Inferiore 95%*	*Superiore 95%*
Intercetta	-42,93319604	82,85006047	-0,518203557	0,618339365	-233,9857781	148,119386
x	10,43815163	0,834408268	12,50964549	1,56054E-06	8,514002711	12,36230055

ovvero della riga con l'intestazione ***"x"***. Questa riga contiene alcuni risultati inferenziali sul coefficiente angolare β_1 della retta di regressione.

- Nella colonna ***"Coefficienti"*** troviamo il valore della stima puntuale

$$\widehat{\beta}_1 = 10,43815163$$

che abbiamo già calcolato in precedenza.

- Nella colonna ***"Errore standard"*** troviamo il valore di una stima puntuale $s_{\widehat{\beta}_1}$ per lo scarto quadratico medio dello stimatore che ha generato la stima puntuale $\widehat{\beta}_1 = 10,43815163$.

  La formula per calcolare la stima puntuale $s_{\widehat{\beta}_1}$ è data da

$$s_{\widehat{\beta}_1} = \frac{s_\epsilon}{\sqrt{dev(x_i)}}$$

[[vedi risultati calcolati in precedenza]]

$$= \frac{23,344}{\sqrt{776,9}} = 0,838 \simeq 0,834408268$$

  (la differenza è dovuta soltanto agli arrotondamenti).

- Nella colonna ***"Stat t"*** troviamo il valore della statistica test

$$t = \frac{\widehat{\beta}_1}{s_{\widehat{\beta}_1}} = \frac{\widehat{\beta}_1}{s_\epsilon / \sqrt{dev(x_i)}}$$

ovvero il valore di

$$t = \frac{10,43815163}{0,834408268} = 12,4924639.$$

Si noti che il quadrato di questa statistica test t è dato da

$$t^2 = \frac{\widehat{\beta}_1^2}{s_\epsilon^2 / dev(x_i)} = \frac{\left[\frac{codev(x_i,y_i)}{dev(x_1)}\right]^2 \times dev(x_i)}{s_\epsilon^2} =$$
$$= \frac{\frac{[codev(x_i,y_i)]^2}{dev(x_i)dev(y_i)} \times dev(y_i)}{s_\epsilon^2} = \frac{I^2 \times dev(y_i)}{s_\epsilon^2} =$$
$$= \frac{dev(\widehat{y}_i)}{s_\epsilon^2} = \frac{dev(\widehat{y}_i)}{dev(\widehat{\epsilon}_i)/(n-2)}$$

e che t^2 coincide dunque con ...

. . . la statistica test F che (come abbiamo visto in precedenza) viene utilizzata nel test sulla significatività della regressione, ovvero nel test d'ipotesi con

$$H_0 : \beta_1 = 0 \quad \text{e} \quad H_a : \beta_1 \neq 0.$$

Nel test sulla significatività della regressione, al posto della statistica test

$$F = t^2 = \frac{\widehat{\beta}_1^2}{s_\epsilon^2 / dev(x_i)},$$

si può dunque utilizzare la statistica test

$$t = \frac{\widehat{\beta}_1}{s_\epsilon / \sqrt{dev(x_i)}}$$

il cui valore deve però essere confrontato con i percentili della distribuzione t di Studenti con $n-2$ gdl: al livello di significatività α l'ipotesi nulla $H_0 : \beta_1 = 0$ deve essere rifiutata se e solo se

$$|t| = \left| \frac{\widehat{\beta}_1}{s_\epsilon / \sqrt{dev(x_i)}} \right| > t_{n-2;\alpha/2}.$$

Siccome $t^2_{n-2;\alpha/2} = F_{1;n-2;\alpha}$, la regola decisionale basata sulla statistica test t e quella basata sulla statistica test F sono equivalenti (o si rifiuta $H_0 : \beta_1 = 0$ con entrambe le statistiche test, oppure non si rifiuta $H_0 : \beta_1 = 0$ con nessuna delle due statistiche test).

Coerentemente con quanto appena detto, notiamo che il p-value riportato nell'ultima riga della colonna ***"Valore di significatività"***, ovvero il valore

$$1,56054E-06 = 0,00000156054,$$

è identico al p-value associato alla statistica test F che è riportato nella seconda tabella dell'*output* di EXCEL.

In precedenza abbiamo già osservato che con un p-value così piccolo l'ipotesi nulla $H_0 : \beta_1 = 0$ non può che essere rifiutata.

- Nelle colonne ***"Inferiore 95%"*** e ***"Superiore 95%"*** troviamo infine gli estremi di un intervallo di confidenza al 95% per l'ignoto valore di β_1.

  Le formule per calcolare gli estremi di un IdC per β_1 con un generico livello di confidenza $1-\alpha$ sono date da

$$\widehat{\beta}_1 \pm t_{n-2;\alpha/2} \times s_{\widehat{\beta}_1} = \widehat{\beta}_1 \pm t_{n-2;\alpha/2} \times \frac{s_\epsilon}{\sqrt{dev(x_i)}}.$$

  Per l'esempio sui consumi di gas metano possiamo facilmente verificare che gli estremi dell'IdC al 95% sono dati da (dalla tavola della distribuzione t di Student desumiamo che $t_{10-2;0,025} = 2,306$; i valori di $\widehat{\beta}_1$ e di $s_{\widehat{\beta}_1}$ li troviamo nelle colonne ***"Coefficienti"*** e ***"Errore standard"***)

$$10,43815163 \pm 2,306 \times 0,834408268 = \begin{cases} 8,514006164 \\ 12,3622971 \end{cases}$$

  così come riportato nelle colonne ***"Inferiore 95%"*** e ***"Superiore 95%"***.

Oltre ai risultati inferenziali dell'*output* preimpostato di EXCEL, potrebbero interessare anche altri risultati.

A volte, per esempio, al posto del test statistico con

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_a : \beta_1 \neq 0$$

potrebbe essere più interessante un test statistico con

$$H_0 : \beta_1 \leq \beta_1^* \quad \text{vs.} \quad H_a : \beta_1 > \beta_1^*$$

oppure con

$$H_0 : \beta_1 \geq \beta_1^* \quad \text{vs.} \quad H_a : \beta_1 < \beta_1^*$$

dove β_1^* può essere un qualunque valore.

Nell'esempio sul consumo di gas metano Y, si potrebbe per esempio essere interessati a verificare se per ogni metro quadrato aggiuntivo di superficie X, il consumo medio annuo $E(Y|X = x)$ aumenti per più di $\beta_1^* = 10 Smc$.

In questo caso le ipotesi statistiche da sottoporre a verifica dovrebbero essere definite come

$$H_0 : \beta_1 \leq 10 \quad \text{vs.} \quad H_a : \beta_1 > 10$$

e la decisione dovrebbe essere basata sulla statistica test

$$t = \frac{\widehat{\beta}_1 - \beta_1^*}{s_{\widehat{\beta}_1}}.$$

Dalla terza tabella dell'*output* di EXCEL desumiamo che nel caso in questione, il valore assunto da questa statistica test t sia dato da

$$t = \frac{10,43815163 - 10}{0,834408268} = 0,525104612 \simeq 0,525.$$

Per calcolare il p-value associato a $t = 0,525$, ...

. . . bisogna confrontare $t = 0,525$ con i percentili della distribuzione t di Student con $n - 2 = 10 - 2 = 8$ gdl.

Dalla tavola della distribuzione t di Student desumiamo che

$$t = 0,525 < t_{8;0,20} = 0,889$$

e quindi possiamo concludere che il p-value è ampiamente maggiore di $0,20$.

Chiaramente, con un p-value così lontano da zero non possiamo rifiutare l'ipotesi nulla $H_0 : \beta_1 \leq 10$.

Sempre con riferimento ad un modello di regressione lineare semplice, in alcune circostanze si è interessati ad un IdC per l'ignoto valore che assume la retta di regressione

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

in corrispondenza di un dato valore $x \in \mathbb{R}$.

Si noti che la terza tabella dell'*output* di EXCEL riporta già gli estremi di un IdC al 95% per l'ignoto valore di

$$E(Y|X = 0) = \beta_0 + \beta_1 \times 0 = \beta_0.$$

Ricordiamo che gli estremi dell'IdC in questione sono dati da

$$\widehat{\beta}_0 \pm t_{n-2;\alpha/2} \times s_\epsilon \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{dev(x_i)}} \quad \text{con } \alpha = 0.05.$$

Se al posto di un IdC per $E(Y|X = 0)$ volessimo un IdC per $E(Y|X = x)$ in corrispondenza di un qualunque altro valore di $x \in \mathbb{R}$, possiamo ottenere gli estremi di un tale IdC attraverso la formula

$$\widehat{\beta}_0 + \widehat{\beta}_1 x \pm t_{n-2;\alpha/2} \times s_\epsilon \sqrt{\frac{1}{n} + \frac{(x - \overline{x})^2}{dev(x_i)}}.$$

Ovviamente, il livello di confidenza di questi IdC è $1 - \alpha$.

Per illustrare un'applicazione degli IdC in questione, possiamo utilizzare ancora i dati sulle superfici X degli $n = 10$ appartamenti e sui corrispondenti consumi annui di gas metano Y.

Come abbiamo già visto, ...

. . . con i dati del campione di appartamenti otteniamo (vedi i calcoli fatti in precedenza)

$$\overline{x} = 98,9, \quad dev(x_i) = 776,9,$$
$$\widehat{\beta}_0 = -42,918, \quad \widehat{\beta}_1 = 10,438, \quad s_\epsilon = 23,344.$$

Inoltre, dalla tavola della distribuzione t di Student desumiamo che $t_{n-2;\alpha/2} = t_{8;0,025} = 2,306$.

Gli estremi di un IdC all'$1 - \alpha = 0,95 = 95\%$ per l'ignoto consumo medio di gas metano di un appartamento di $x = 100mq$ sono dunque dati da

$$-42,918 + 10,438 \times 100 \ \pm \ 2,306 \times 23,344 \sqrt{\frac{1}{10} + \frac{(100 - 98,9)^2}{776,9}} =$$
$$= 1000,882 \ \pm \ 17,155 = \begin{cases} 1018,037 \\ 983,727. \end{cases}$$

Per concludere, affronteremo ancora un ultimo problema: quello di calcolare un cosiddetto **"intervallo di previsione"** per un nuovo valore della variabile Y in corrispondenza di un dato valore x della variabile X.

Si ricordi che sotto le ipotesi del modello di regressione lineare semplice, il nuovo valore di Y sarebbe la realizzazione di una variabile casuale $\widetilde{Y}$ che è definita come

$$\widetilde{Y} = \beta_0 + \beta_1 x + \widetilde{\epsilon}$$

dove $\widetilde{\epsilon} \sim \mathit{Normale}(\mu_\epsilon = 0; \sigma^2)$ con σ ignoto.

Dunque, sotto le ipotesi del modello di regressione lineare semplice si ha

$$\widetilde{Y} \sim \mathit{Normale}(\mu_y = \beta_0 + \beta_1 x; \sigma_y = \sigma)$$

e da questa cpnsiderazione deduciamo che

$$P\left(\left\{\beta_0 + \beta_1 x - z_{1-\alpha/2}\sigma < \widetilde{Y} < \beta_0 + \beta_1 x - z_{1-\alpha/2}\sigma\right\}\right) = 1 - \alpha.$$

Se conoscessimo i valori dei parametri β_0, β_1 e σ^2 potremmo dunque calcolare gli estremi di un **"intervallo di previsione"** con livello di confidenza $1 - \alpha$ come

$$\beta_0 + \beta_1 x \pm z_{1-\alpha/2}\sigma.$$

Tuttavia, non conoscendo i valori dei parametri β_0, β_1 e σ^2, siamo di fatto costretti a calcolare gli estremi di un **intervallo di previsione** secondo la formula

$$\widehat{\beta}_0 + \widehat{\beta}_1 x \pm t_{n-2;\alpha/2} \times s_\epsilon \sqrt{1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{dev(x_i)}}$$

che (con probabilità elevata) dà luogo ad un intervallo di previsione più ampio (infatti, gli estremi del secondo intervallo di previsione devono tener conto anche dell'incertezza sulle stime dei parametri β_0 e β_1).

Per fare un esempio, possiamo calcolare **l'intervallo di previsione al** 95% **per il consumo di gas metano di un appartamento con una superficie di** $100mq$.

Usando i dati del campione di $n = 10$ appartamenti, che per comodità riportiamo ancora una volta

$$\overline{x} = 98,9, \quad dev(x_i) = 776,9,$$
$$\widehat{\beta}_0 = -42,918, \quad \widehat{\beta}_1 = 10,438, \quad s_\epsilon = 23,344,$$

otteniamo (dalla tavola della distribuzione t di Student desumiamo che $t_{n-2;\alpha/2} = t_{8;0,025} = 2,306$)

$$-42,918 + 10,438 \times 100 \ \pm \ 2,306 \times 23,344 \sqrt{1 + \frac{1}{10} + \frac{(100 - 98,9)^2}{776,9}} =$$
$$= 1000,882 \ \pm \ 56,499 = \begin{cases} 1057,381 \\ 944,383. \end{cases}$$

Si noti che gli **intervalli di previsione** per un nuovo valore di Y sono centrati sul valore di

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x, \tag{3}$$

che può essere interpretato come una **previsione puntuale** per un nuovo valore di Y in corrispondenza di un dato valore x della variabile X.

Si ricordi che il valore di $\widehat{y}$ nella (3) è altresì una **stima puntuale** per l'ignoto valore atteso di un nuovo valore di Y, ovvero per la media condizionata

$$E(Y|X = x) = \beta_0 + \beta_1 x$$

(si ricordi che gli IdC per le medie condizionate $E(Y|X = x) = \beta_0 + \beta_1 x$ sono anch'essi centrati sul valore di $\widehat{y}$).

Si noti che l'espressione del **"margine d'errore"** di un **intervallo di previsione**, ovvero l'espressione

$$t_{n-2;\alpha/2} \times s_\epsilon \sqrt{1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{dev(x_i)}},$$

è molto simile a quella del margine d'errore di un IdC per $E(Y|X = x)$, ovvero all'espressione

$$t_{n-2;\alpha/2} \times s_\epsilon \sqrt{\frac{1}{n} + \frac{(x - \overline{x})^2}{dev(x_i)}}.$$

L'unica differenza consiste nell'aggiunta di un "1" nell'argomento della radice.

Chiaramente, questo "1" aggiuntivo rende il margine d'errore di un **intervallo di previsione per un nuovo valore di** Y maggiore del margine d'errore dell'IdC per il **valore atteso del nuovo valore di** Y, ovvero dell'IdC per $E(Y|X = x) = \beta_0 + \beta_1 x$.

Il fatto che il margine d'errore di un **intervallo di previsione per un nuovo valore di** Y sia maggiore del margine d'errore dell'IdC per il **valore atteso del nuovo valore di** Y non è una sorpresa:

Infatti, per ottenere un IdC per il valore atteso

$$E(Y|X = x) = \beta_0 + \beta_1 x,$$

è sufficiente tenere in considerazione l'incertezza sulla precisione della stima puntuale

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x,$$

mentre per prevedere un nuovo valore di Y, ovvero per prevedere la realizzazione di una variabile casuale del tipo

$$\widetilde{Y} = \beta_0 + \beta_1 x + \widetilde{\epsilon},$$

bisogna tenere in considerazione anche l'incertezza sulla realizzazione del termine d'errore $\widetilde{\epsilon}$.

Per concludere conviene aggiungere un'ultima osservazione: sia nel caso dell'IdC per il valore atteso di nuovo valore di Y, sia nel caso dell'intervallo di previsione per il nuovo valore di Y, i corrispondenti margini d'errore

$$t_{n-2;\alpha/2} \times s_\epsilon \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{dev(x_i)}}.$$

e

$$t_{n-2;\alpha/2} \times s_\epsilon \sqrt{1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{dev(x_i)}},$$

sono entrambi più stretti se si riferiscono ad un valore di x che è prossimo alla media campionaria $\overline{x}$.

Questo significa che la stima puntuale/previsione puntuale

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

è più precisa quando si riferisce ad un valore di x che è prossimo alla media campionaria $\overline{x}$.

Esercizio 6.1

La tabella sottostante si riferisce alle 10 più grandi aziende del settore GICS delle comunicazioni incluse nell'indice di borsa S&P500 (fonte: Factset - clicca qui per scaricare una cartella di lavoro EXCEL che contiene i dati).

TICKER	X = COMPENSO DIRIGENTI ANNO 2019 (in milioni di USD)	Y = UTILE NETTO ANNO 2019 ANNO 2019 (in miliardi di USD)
GOOG	281,946	34,343
FB	112,844	18,485
DIS	92,795	11,054
VZ	51,095	19,913
CMCSA	142,706	13,057
NFLX	126,379	1,867
T	80,583	14,975
TMUS	85,494	3,468
CHTR	21,511	1,668
ATVI	54,417	1,736

Le domande sono riportate sulla prossima *slide*.

Esercizio (continua)

a) Si valuti, attraverso un opportuni test grafici, se i dati presentano evidenza empirica contraria alle ipotesi del modello di regressione lineare semplice.

b) Si esegua il test sulla significatività della regressione e si commenti il risultato del test.

c) Si determini un IdC al 95% per il coefficiente angolare della retta di regressione.

d) Si calcoli un IdC al 95% per il valore atteso degli utili di un'azienda i cui dirigenti percepiscono complessivamente un compenso annuo di $x = 100$ milioni di USD.

e) Si calcoli un intervallo di previsione al 95% per il valore atteso degli utili di un'azienda i cui dirigenti percepiscono complessivamente un compenso annuo di $x = 100$ milioni di USD.

Esercizio 6.2

Un investitore vuole valutare l'esposizione al rischio di mercato del titolo AAPL. A tal fine considera gli eccessi di rendimento mensili del titolo AAPL (variabile Y) e gli eccessi di rendimento mensili dell'indice S&P500 (variabile X) del periodo gen-2017 - feb-2022 ($n = 62$ coppie di osservazioni).[a]

A partire dalle osservazioni si ottengono le seguenti quantità di sintesi:

$$\sum_{i=1}^{63} x_i = 0,1200, \quad \sum_{i=1}^{63} y_i = 1,3654,$$

$$\sum_{i=1}^{63} x_i^2 = 0,1372, \quad \sum_{i=1}^{63} y_i^2 = 0,4828, \quad \sum_{i=1}^{63} x_i y_i = 0,1660$$

[a] Gli **eccessi di rendimento** sono stati calcolati come differenze tra i rendimenti mensili del titolo AAPL/dell'indice S&P500 e i rendimenti dei titoli di Stato USA con durata residua di circa un mese.

Esercizio (continua)

a) Si valuti, attraverso un opportuni test grafici, se i dati presentano evidenza empirica contraria alle ipotesi del modello di regressione lineare semplice.

b) Si calcolino le stime puntuali per i parametri del modello di regressione lineare semplice.

c) Si determini il p-value per il test sulla significatività della regressione.

d) Si calcoli un IdC al 95% per l'ignoto valore di β_0.

e) Si determini il p-value per il test statistico che contrappone l'ipotesi nulla $H_0 : \beta_1 \leq 1$ all'ipotesi alternativa $H_a : \beta_1 > 1$.

f) Si verifichi se l'*output* della funzione "regressione" di EXCEL fronisce risultati coerenti con quelli ottenuti nelle risposte ai precedenti quesiti. (clicca qui per accedere ai dati mensili).