UNILAB NETWORK

DISPENSA DI

STATISTICA (CLEACC)

SECONDO PARZIALE

EDIZIONE A.A. 2020 - 2021

A cura di Giulia Barzaghi, Chiara Burastero, Martina Garavaglia

VERIFICA DI IPOTESI

È un problema di inferenza statistica (che quindi cerca di trarre conclusioni su una popolazione partendo dalle informazioni disponibili su un suo campione) in cui si vuole rispondere ad una DOMANDA DI RICERCA. La conclusione è rifiutare o non rifiutare un'ipotesi.
L'ipotesi da verificare ha a che fare con un **PARAMETRO INCOGNITO**, che può essere la **media** o la **proporzione** di una popolazione.

Caratteristiche generali del problema di verifica di ipotesi:

- IPOTESI ALTERNATIVA (H_1): è la domanda di ricerca, l'ipotesi che si vuole verificare su un parametro incognito, sulla base dei dati campionari.
  Normalmente, l'ipotesi alternativa ha a che fare con un CAMBIAMENTO – aumento, diminuzione, effetto – in seguito a qualche intervento o che conduce a un intervento successivo.
- IPOTESI NULLA (H_0): è la negazione della domanda di ricerca e rappresenta lo status quo, la situazione prima dell'intervento → l'ipotesi nulla nega l'effetto dell'intervento studiato dall'ipotesi alternativa. Si verifica sullo stesso parametro incognito dell'ipotesi alternativa.

Trattandosi di affermazioni di inferenza statistica, le conclusioni non sono mai certe al 100%, ma si ha sempre a che fare con livelli di fiducia. Un'ipotesi può essere **rifiutata** oppure può essere ritenuta **valida**, ma non si può mai affermare che essa sia falsa oppure che sia vera, perché il problema di verifica è di tipo inferenziale. Rifiutare l'ipotesi nulla significa ritenere ragionevolmente valida l'ipotesi alternativa, sulla base dell'evidenza empirica.
Sulla base della realizzazione del campione deve decidere quale delle due ipotesi sia la più ragionevole, attraverso un test statistico.

TEST STATISTICO

È una procedura che utilizza la realizzazione campionaria per decidere se rifiutare o meno l'ipotesi alternativa. Il test statistico si realizza attraverso una REGIONE DI RIFIUTO.
La regione di rifiuto è un sottoinsieme dello spazio di tutte le realizzazioni campionarie possibili in corrispondenza delle quali si rifiuta l'ipotesi nulla. Un test statistico è completamente caratterizzato da una regione di rifiuto.
→ Se la realizzazione campionaria cade nella regione di rifiuto => **si rifiuta** l'ipotesi nulla.
→ Se la realizzazione campionaria cade nel complementare della regione di rifiuto => **non si rifiuta** l'ipotesi nulla.

Dal momento che le procedure utilizzare per la verifica di ipotesi sono inferenziali, è possibile commettere errori:

1. **ERRORE DI I TIPO**: rifiutare l'ipotesi nulla, mentre in realtà essa è vera → si valuta erroneamente come FALSA l'ipotesi nulla anche se è vera.
2. **ERRORE DI II TIPO**: non rifiutare l'ipotesi nulla, mentre in realtà essa è falsa → si valuta erroneamente come VERA l'ipotesi nulla anche se è falsa.

Visto che le possibili conclusioni del test sono 2, rifiutare o non rifiutare l'ipotesi nulla, anche i tipi di errore sono 2.

La VALUTAZIONE di un test si basa sulle probabilità di commettere errori.
Un test ideale dovrebbe avere probabilità di commettere errori molto basse → Tuttavia, non è possibile ridurre contemporaneamente la probabilità di commettere errori di I e di II tipo, poiché una riduzione dell'errore di I tipo è dovuta ad una riduzione della regione di rifiuto, ma, al diminuire della regione di rifiuto aumenta la regione di accettazione, facendo così aumentare la probabilità di errore di II tipo.
In altre parole, se si riduce la regione di rifiuto diminuisce anche la probabilità di errori di I tipo (è più difficile che l'ipotesi nulla cada all'interno della regione di rifiuto ed è quindi più facile che venga ritenuta valida), ma alla riduzione della regione di rifiuto corrisponde un aumento della regione di accettazione, perché

ritenere valida l'ipotesi nulla diventa più semplice. Di conseguenza, la probabilità di commettere errori di II tipo, ritenendo erroneamente valida l'ipotesi nulla, aumenta.
Per ovviare a questo problema, si presta attenzione soltanto alla probabilità di commettere errori di I tipo, scegliendo test che abbiano tale probabilità molto bassa. La probabilità di errori di II tipo, invece, non viene controllata dal test e potrebbe anche essere molto elevata.
LIVELLO DI SIGNIFICATIVITÀ DEL TEST = α
Alfa è la probabilità di commettere errori di I tipo. → Un test di livello alfa ha probabilità di commettere errori di I tipo pari ad alfa.
Dal momento che alfa è la probabilità di commettere errori di I tipo, scegliendo alfa con valori molto bassi (pari ad esempio a 0.01, 0.05, 0.001),

- Se si rifiuta l'ipotesi nulla si ha elevata fiducia che essa sia falsa → in questo caso, il test fornisce FORTE EVIDENZA EMPIRICA CONTRO L'IPOTESI NULLA, a favore dell'ipotesi alternativa: si risponde in modo affermativo alla domanda di ricerca e si ritiene ragionevolmente valida l'ipotesi alternativa.

- Al contrario, decidendo di non rifiutare l'ipotesi nulla (non ritenendo valida l'ipotesi alternativa), non è possibile affermare con fiducia che essa sia vera, poiché la probabilità di errore del II tipo non è bassa come alfa. → in questo caso, il test non fornisce evidenza empirica significativa contro l'ipotesi nulla.

Un test non può mai fornire evidenza empirica *a favore* dell'ipotesi nulla – può soltanto fornire evidenza contro di essa oppure non fornire evidenza empirica sufficiente contro di essa.

Date le considerazioni sulle probabilità di errore, le conclusioni del test di verifica di ipotesi sono asimmetriche:
- RIFIUTO DELL'IPOTESI NULLA: è una conclusione forte e significativa, sulla cui correttezza si ha elevata fiducia dal momento che la probabilità di errore (alfa) è molto bassa.
- NON RIFIUTO DELL'IPOTESI NULLA: è una conclusione debole e non significativa, sulla cui correttezza non si può riporre fiducia. Non è una vera e propria risposta al test di verifica, perché il test non è andato nella direzione dell'ipotesi alternativa e si è rimasti nello status quo.

1. TEST DI VERIFICA DI IPOTESI SULLA MEDIA

Il punto di partenza è la media campionaria: $\bar{X} \sim N(\mu, \frac{\delta^2}{n})$

$\bar{X}$ è lo stimatore di μ (media della popolazione) e ha distribuzione normale con media μ e varianza $\frac{\delta^2}{n}$.

I test di verifica sulla media possono essere unilaterali o bilaterali.
Un test UNILATERALE ha lo scopo di verificare se la media della popolazione è aumentata o diminuita rispetto a un dato valore μ_0, mentre un test BILATERALE ha lo scopo di verificare se la media di una popolazione è variata rispetto a un dato valore iniziale μ_0.

La STATISTICA TEST è una funzione campionaria che determina lo studio della verifica di ipotesi. Viene sempre calcolata assumendo come vera l'ipotesi nulla.
→ Se la varianza della popolazione è nota, la statistica test è indicata con Z e assume sempre distribuzione normale standard → Z è la standardizzazione della media campionaria.
→ Se la varianza della popolazione è incognita, la statistica test è indicata con T e assume distribuzione T di Student → T è la studentizzazione della media campionaria.
Inoltre, se la varianza della popolazione è nota, nella statistica test compare la varianza δ della popolazione, mentre, se è incognita, nella statistica test compare lo stimatore della varianza, S.

I modi per risolvere un problema di verifica di ipotesi sono 2: utilizzando le regioni di rifiuto (R_α), oppure utilizzando il P-VALUE.

1. Con il metodo della **REGIONE DI RIFIUTO**, si traggono le conclusioni del test calcolando il valore osservato della statistica test nel campione e confrontandolo con un determinato quantile (che può variare da caso a caso).
   Per i test unilaterali, il quantile è di ordine $1-\alpha$, mentre per i test bilaterali è di ordine $1-\frac{\alpha}{2}$.
   Inoltre, se la varianza della popolazione è nota, il quantile appartiene alla distribuzione normale standard ed è indicato con z (z_α nei test unilaterali e $z_{\frac{\alpha}{2}}$ nei test bilaterali), mentre se la varianza è incognita il quantile è della distribuzione T di Student ed è indicato con t (t_α^{n-1} nei test unilaterali e $t_{\frac{\alpha}{2}}^{n-1}$ nei test bilaterali).

2. Con il metodo del **P-VALUE**, che è definito come la probabilità con cui la statistica test assume dei valori maggiori/minori/diversi da quelli osservati nel campione (a seconda del sistema di ipotesi del test), si confrontano il valore ottenuto calcolando il p-value e il livello di significatività, α. In particolare,
   - Se $p-value < \alpha$ si rifiuta l'ipotesi nulla a livello alfa
   - Se $p-value \geq \alpha$ non si rifiuta l'ipotesi nulla a livello alfa

   Il p-value misura la plausibilità dell'ipotesi nulla alla luce della realizzazione campionaria: minore è tale valore, meno è plausibile l'ipotesi nulla e, di conseguenza, si è più portati a rifiutarla. Si calcola *sempre* assumendo che l'ipotesi nulla sia *vera*.

   Il calcolo del valore del p-value cambia a seconda del sistema di ipotesi adottato e cambia se il test è unilaterale o bilaterale.

Aumentare il livello di significatività (alfa) equivale ad aumentare il rischio di commettere errori di I tipo => se non si rifiuta l'ipotesi nulla ad un dato livello, non si rifiuterà sicuramente neanche a livelli più piccoli. Allo stesso modo, se si rifiuta l'ipotesi nulla ad un dato livello, si rifiuterà sicuramente anche a livelli più grandi.
Se il p-value è un valore molto vicino a 0, si rifiuta l'ipotesi nulla per ogni livello alfa.

I. Test su una popolazione NORMALE

- con varianza NOTA

Introduzione

$x_1, x_2, \dots, x_n \sim N(\mu, \delta^2)$

Statistica test (sotto H_0): $Z = \frac{\bar{X}-\mu_0}{\delta/\sqrt{n}} \sim N(0,1)$

Realizzazione della statistica test: si calcola osservando le realizzazioni $\bar{x}$ di $\bar{X}$: $z_{oss} = \frac{\bar{x}-\mu_0}{\delta/\sqrt{n}}$

Regione di rifiuto:

a. $R_\alpha: \bar{X} > \mu_0 + k$
b. $R_\alpha: \bar{X} < \mu_0 + k$
c. $R_\alpha: \bar{X} \neq \mu_0 + k$

Dove $k = quantile\ della\ normale\ standard * \frac{\delta}{\sqrt{n}}$

→ Per i test **unilaterali**, il quantile della normale standard è z_α di ordine $1-\alpha$: è il punto che lascia alla sua destra un'area pari ad alfa nella distribuzione normale standard. Confrontando z_{oss} con z_a si impone la condizione che la probabilità di commettere errori di I tipo sia alfa.
$z_\alpha = \frac{k}{\delta/\sqrt{n}}$

→ Per i test **bilaterali**, il quantile della normale standard è $z_{\frac{\alpha}{2}}$ di ordine $1-\frac{\alpha}{2}$ è $z_{\frac{\alpha}{2}} = \frac{k}{\delta/\sqrt{n}}$

Sistemi di ipotesi:

a) Verificare se la media è aumentata rispetto a μ_0 (test unilaterale a destra)

H_0: $\mu = \mu_0$
H_1: $\mu > \mu_0$

Struttura della regione di rifiuto: è ragionevole rifiutare l'ipotesi nulla se lo stimatore assume valori sufficientemente più grandi di μ_0

$$R_\alpha: Z = \frac{\bar{X} - \mu_0}{\delta/\sqrt{n}} > z_\alpha$$

Se $z_{oss} > z_\alpha$ => si rifiuta l'ipotesi nulla H_0: c'è sufficiente evidenza empirica per affermare che la media della popolazione è aumentata rispetto a μ_0.
Se $z_{oss} \leq z_\alpha$ => non si rifiuta l'ipotesi nulla H_0: non si può affermare che la media della popolazione è aumentata rispetto a μ_0 perché non c'è sufficiente evidenza empirica.

Oppure, $p - value = P(Z > z_{oss} | \mu = \mu_0)$

Se il valore del p-value è minore di alfa, si rifiuta l'ipotesi nulla a livello alfa, altrimenti non si rifiuta.

b) Verificare se la media è diminuita rispetto a μ_0 (test unilaterale a sinistra)

H_0: $\mu = \mu_0$
H_1: $\mu < \mu_0$

Struttura della regione di rifiuto: è ragionevole rifiutare l'ipotesi nulla se lo stimatore assume valori sufficientemente minori di μ_0

$$R_\alpha: Z = \frac{\bar{X} - \mu_0}{\delta/\sqrt{n}} < -z_\alpha$$

Se $z_{oss} < -z_\alpha$ => si rifiuta l'ipotesi nulla H_0: c'è sufficiente evidenza empirica per affermare che la media della popolazione è aumentata rispetto a μ_0.
Se $z_{oss} \geq -z_\alpha$ => non si rifiuta l'ipotesi nulla H_0: non si può affermare che la media della popolazione è aumentata rispetto a μ_0 perché non c'è sufficiente evidenza empirica.

Oppure, $p - value = P(Z < z_{oss} | \mu = \mu_0)$

Se il valore del p-value è minore di alfa, si rifiuta l'ipotesi nulla a livello alfa, altrimenti non si rifiuta.

USO DI R PER I TEST UNILATERALI:
$p - value = 1 - pnorm(z_{oss})$

c) Verificare se la media è cambiata rispetto a μ_0 (test bilaterale)

H_0: $\mu = \mu_0$
H_1: $\mu \neq \mu_0$

In questo caso il test è bilaterale, quindi il quantile è $z_{\frac{\alpha}{2}}$ (di ordine $1 - \frac{\alpha}{2}$).
Bisogna considerare la statistica test in modulo.
Struttura della regione di rifiuto: è ragionevole rifiutare l'ipotesi nulla se lo stimatore assume valori sufficientemente grandi

$$R_\alpha: |Z| = \left| \frac{\bar{X} - \mu_0}{\delta/\sqrt{n}} \right|, > z_{\frac{\alpha}{2}}$$

Se $|z_{oss}| > z_{\frac{\alpha}{2}}$ => si rifiuta l'ipotesi nulla H_0: c'è sufficiente evidenza empirica per affermare che la media della popolazione è aumentata rispetto a μ_0.
Se $|z_{oss}| \leq z_{\frac{\alpha}{2}}$ => non si rifiuta l'ipotesi nulla H_0: non si può affermare che la media della popolazione è aumentata rispetto a μ_0 perché non c'è sufficiente evidenza empirica.

Oppure, $p - value = 2 * P(Z > |z_{oss}| \, |\mu = \mu_0)$

Se il valore del p-value è minore di alfa, si rifiuta l'ipotesi nulla a livello alfa, altrimenti non si rifiuta.

USO DI R PER I TEST BILATERALI:
$p - value = 2 * (1 - pnorm(z_{oss}))$

- con varianza INCOGNITA

Introduzione

$x_1, x_2, \dots, x_n iid \sim N(\mu, \delta^2)$

La varianza è da stimare tramite la varianza campionaria S.

Statistica test (sotto H_0): $T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$

Sotto l'ipotesi nulla, la statistica test T ha distribuzione T di Student con n-1 gradi di libertà.

Realizzazione della statistica test: si calcola osservando le realizzazioni campionarie $\bar{x}$ di $\bar{X}$ e le realizzazioni campionarie s di S: $t_{oss} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$

→ Per i test **unilaterali**, il quantile della distribuzione T di Student è t_{α}^{n-1} di ordine $1 - \alpha$ e con n-1 gradi di libertà.
→ Per i test **bilaterali**, il quantile della distribuzione T di Student è $t_{\frac{\alpha}{2}}^{n-1}$ di ordine $1 - \frac{\alpha}{2}$ e con n-1 gradi di libertà.

Sistemi di ipotesi:

a) Verificare se la media è aumentata rispetto a μ_0 (test unilaterale a destra)

H_0: $\mu = \mu_0$
H_1: $\mu > \mu_0$

Struttura della regione di rifiuto: è ragionevole rifiutare l'ipotesi nulla se lo stimatore assume valori sufficientemente più grandi di μ_0

$$R_{\alpha}: T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} > t_{\alpha}^{n-1}$$

Se $t_{oss} > t_{\alpha}^{n-1}$ => si rifiuta l'ipotesi nulla H_0: c'è sufficiente evidenza empirica per affermare che la media della popolazione con varianza incognita è aumentata rispetto a μ_0.
Se $t_{oss} \leq t_{\alpha}^{n-1}$ => non si rifiuta l'ipotesi nulla H_0: non si può affermare che la media della popolazione con varianza incognita è aumentata rispetto a μ_0 perché non c'è sufficiente evidenza empirica.

Oppure, $p - value = P(T > t_{oss} \, |\mu = \mu_0)$
Se il valore del p-value è minore di alfa, si rifiuta l'ipotesi nulla a livello alfa, altrimenti non si rifiuta.

b) Verificare se la media è diminuita rispetto a μ_0 (test unilaterale a sinistra)

H_0: $\mu = \mu_0$
H_1: $\mu < \mu_0$

Struttura della regione di rifiuto: è ragionevole rifiutare l'ipotesi nulla se lo stimatore assume valori sufficientemente più grandi di μ_0

$$R_\alpha: T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} < -t_\alpha^{n-1}$$

Se $t_{oss} < -t_\alpha^{n-1}$ => si rifiuta l'ipotesi nulla H_0: c'è sufficiente evidenza empirica per affermare che la media della popolazione con varianza incognita è aumentata rispetto a μ_0.
Se $t_{oss} \geq -t_\alpha^{n-1}$ => non si rifiuta l'ipotesi nulla H_0: non si può affermare che la media della popolazione con varianza incognita è aumentata rispetto a μ_0 perché non c'è sufficiente evidenza empirica.

Oppure, $p - value = P(T < t_{oss} | \mu = \mu_0)$

Se il valore del p-value è minore di alfa, si rifiuta l'ipotesi nulla a livello alfa, altrimenti non si rifiuta.

c) Verificare se la media è cambiata rispetto a μ_0 (test bilaterale)

H_0: $\mu = \mu_0$
H_1: $\mu \neq \mu_0$

Struttura della regione di rifiuto: è ragionevole rifiutare l'ipotesi nulla se lo stimatore assume valori sufficientemente più grandi di μ_0

$$R_\alpha: |T| = \left| \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \right| > t_{\frac{\alpha}{2}}^{n-1}$$

Se $t_{oss} > t_{\frac{\alpha}{2}}^{n-1}$ => si rifiuta l'ipotesi nulla H_0: c'è sufficiente evidenza empirica per affermare che la media della popolazione con varianza incognita è aumentata rispetto a μ_0.
Se $t_{oss} \leq t_{\frac{\alpha}{2}}^{n-1}$ => non si rifiuta l'ipotesi nulla H_0: non si può affermare che la media della popolazione con varianza incognita è aumentata rispetto a μ_0 perché non c'è sufficiente evidenza empirica.

Oppure, $p - value = 2 * P(T > |t_{oss}| \, | \mu = \mu_0)$

Se il valore del p-value è minore di alfa, si rifiuta l'ipotesi nulla a livello alfa, altrimenti non si rifiuta.

USO DI R PER I TEST SU POPOLAZIONI NORMALI CON VARIANZA INCOGNITA:

Funzione t.test: $t.test\ (\ dataframe\$variabile\ , mu = mu\_o\ , alternative = "two.sided")$

L'argomento alternative specifica il sistema di ipotesi del test:

a. alternative="greater" effettua il test unilaterale con coda a dx (verificare se la media è aumentata),
b. alternative="less" effettua il test unilaterale con coda a sx (verificare se la media è diminuita),
c. alternative="two.sided" effettua il test bilaterale (verificare se la media è cambiata).

Se l'argomento alternative è omesso, il test di default è quello bilaterale.
Gli output della funzione sono: t= t_{oss}, df=gradi di libertà

II. Test su una popolazione ARBITRARIA

$x_1, x_2, \dots, x_n iid$ con distribuzione qualunque

È un test asintotico e il campione deve essere abbastanza grande (n>40), così che, applicando il teorema centrale del limite, la statistica test abbia distribuzione normale standard.

Se la varianza della popolazione è nota, nella statistica test compare la varianza δ della popolazione, mentre, se è incognita, nella statistica test compare lo stimatore della varianza, S.

a) Verificare se la media è aumentata rispetto a μ_0 (test unilaterale a destra)

H_0: $\mu = \mu_0$
H_1: $\mu > \mu_0$

1. <u>Se la varianza è nota:</u>

Regione di rifiuto:

$$R_\alpha: Z = \frac{\bar{X} - \mu_0}{\delta/\sqrt{n}} > z_\alpha$$

P-value:

$$p - value = P(Z > z_{oss} | \mu = \mu_0)$$

2. <u>Se la varianza è incognita:</u>

Regione di rifiuto:

$$R_\alpha: Z = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} > z_\alpha$$

P-value:

$$p - value = P(Z > z_{oss} | \mu = \mu_0)$$

b) Verificare se la media è diminuita rispetto a μ_0 (test unilaterale a sinistra)

H_0: $\mu = \mu_0$
H_1: $\mu < \mu_0$

1. <u>Se la varianza è nota:</u>

Regione di rifiuto:

$$R_\alpha: Z = \frac{\bar{X} - \mu_0}{\delta/\sqrt{n}} < -z_\alpha$$

P-value:

$$p - value = P(Z < z_{oss} | \mu = \mu_0)$$

2. <u>Se la varianza è incognita:</u>

Regione di rifiuto:

$$R_\alpha: Z = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} < -z_\alpha$$

P-value:

$$p - value = P(Z < z_{oss} | \mu = \mu_0)$$

c) Verificare se la media è cambiata rispetto a μ_0 (test bilaterale)

H_0: $\mu = \mu_0$
H_1: $\mu \neq \mu_0$

1. <u>Se la varianza è nota:</u>

Regione di rifiuto:

$$R_\alpha: |Z| = \left| \frac{\bar{X} - \mu_0}{\delta/\sqrt{n}} \right| > z_{\frac{\alpha}{2}}$$

P-value:

$$p - value = 2 * P(Z > |z_{oss}||\mu = \mu_0)$$

2. <u>Se la varianza è incognita:</u>

Regione di rifiuto:

$$R_\alpha: |Z| = \left| \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \right| > z_{\frac{\alpha}{2}}$$

P-value:

$$p - value = 2 * P(Z > |z_{oss}||\mu = \mu_0)$$

Nei test sulla popolazione arbitraria, tutti i livelli alfa e tutti i p-value sono valori approssimati, perché il test è asintotico.

2. TEST DI VERIFICA DI IPOTESI SULLA PROPORZIONE

$x_1, x_2, \dots, x_n iid \sim Bern(p)$ con n sufficientemente grande (n>40)

La proporzione è la media della distribuzione bernoulliana.
Lo stimatore della proporzione p è la proporzione campionaria: $\hat{P} \sim N(p, \frac{p*(p-1)}{n})$

Statistica test: $Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0*(1-p_0)}{n}}}$

Realizzazione della statistica test: $z_{oss} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0*(1-p_0)}{n}}}$

Sistemi di ipotesi:

a) Verificare se la media è aumentata rispetto a μ_0 (test unilaterale a destra)

H_0: $p = p_0$
H_1: $p > p_0$

Regione di rifiuto: $R_\alpha: Z > z_\alpha$ dove $Z = \frac{\bar{X} - \mu_0}{\delta/\sqrt{n}}$, con distribuzione normale standard.

Se $z_{oss} > z_\alpha$ => si rifiuta l'ipotesi nulla H_0 a livello alfa.
Se $z_{oss} \leq z_\alpha$ => non si rifiuta l'ipotesi nulla H_0 a livello alfa.

Oppure, $p - value = P(Z > z_{oss}|p = p_0)$

b) Verificare se la media è diminuita rispetto a μ_0 (test unilaterale a sinistra)

H_0: $p = p_0$
H_1: $p < p_0$

Regione di rifiuto: $R_\alpha: Z < -z_\alpha$, dove $Z = \frac{\bar{X} - \mu_0}{\delta/\sqrt{n}}$, con distribuzione normale standard.

Se $z_{oss} < -z_\alpha$ => si rifiuta l'ipotesi nulla H_0 a livello alfa.
Se $z_{oss} \geq -z_\alpha$ => non si rifiuta l'ipotesi nulla H_0 a livello alfa.

Oppure, $p - value = P(Z < z_{oss} | p = p_0)$

c) Verificare se la media è cambiata rispetto a μ_0 (test bilaterale)

H_0: $p = p_0$
H_1: $p > p_0$

Regione di rifiuto: $R_\alpha: |Z| > z_\alpha$ dove $Z = \frac{\bar{X} - \mu_0}{\delta/\sqrt{n}}$, con distribuzione normale standard.

Se $z_{oss} > z_\alpha$ => si rifiuta l'ipotesi nulla H_0 a livello alfa.
Se $z_{oss} \leq z_\alpha$ => non si rifiuta l'ipotesi nulla H_0 a livello alfa.

Oppure, $p - value = 2 * P(Z > |z_{oss}| | p = p_0)$

3. TEST DI VERIFICA DI IPOTESI SULLA DIFFERENZA DI MEDIE

Il test riguarda 2 sottogruppi di una stessa popolazione, che vengono confrontati sulla base di una variabile di interesse comune. Tutte le variabili ($x_1, x_2, \dots, x_{nx}$ e $y_1, y_2, \dots, y_{ny}$) sono indipendenti identicamente distribuite (iid) all'interno di ciascun campione e tra i due campioni.

Esempio:
Si consideri una popolazione composta da studenti che conseguono la laurea magistrale in un'università, divisa in due sottopopolazioni: il primo gruppo (A) ha frequentato il triennio presso la medesima università, il secondo (B) presso un'altra. Lo scopo del test è confrontare il voto dei laureati nei due gruppi (in media).
Formalizziamo il nostro esempio:
Gruppo A: $x_1, x_2, \dots, x_{nA}$ → X_i è la variabile aleatoria che rappresenta l'i-esimo laureato estratto dalla popolazione da cui provengono gli studenti che hanno frequentato la stessa università.
Gruppo: $y_1, y_2, \dots, y_{nB}$ → Y_i è la variabile aleatoria che rappresenta l'i-esimo laureato estratto dalla popolazione da cui provengono gli studenti che hanno frequentato un'università diversa.

Questo test riguarda due popolazioni indipendenti, nel quale si possono verificare tre situazioni: test unilaterale a destra, test unilaterale a sinistra e test bilaterale.

I. Popolazioni NORMALI con varianze incognite e uguali

$\boldsymbol{X_1, X_2, \dots, X_{n_X}\ i.i.d \sim N(\mu_X, \sigma^2)}$
$\boldsymbol{Y_1, Y_2, \dots, Y_{n_Y}\ i.i.d \sim N(\mu_Y, \sigma^2)}$.

Sistemi di ipotesi:

a) Test unilaterale a destra

$\boldsymbol{H_0: \mu_X = \mu_Y}$ $\boldsymbol{(cioè\ \ \mu_X - \mu_Y = 0)}$
$\boldsymbol{H_1: \mu_X > \mu_Y}$ $\boldsymbol{(cioè\ \ \mu_X - \mu_Y > 0)}$

L'obiettivo del test è verificare se la media μ (in generale, la variabile di interesse) nel primo campione è maggiore rispetto a quella nel secondo.

Dal momento che la varianza è incognita e uguale in entrambi i sottogruppi e che il parametro da stimare in questo caso è $\mu_x - \mu_y$, bisogna standardizzare lo stimatore $\bar{X} - \bar{Y}$, ottenendo

$$\bar{X} - \bar{Y} \sim N(\mu_X - \mu_Y, \frac{\sigma^2}{n_X} + \frac{\sigma^2}{n_Y})$$

Ricordando che la varianza di una differenza è pari alla somma delle varianze meno la covarianza (che in questo caso è nulla poiché le due variabili sono indipendenti), per stimare la varianza incognita si utilizza il seguente stimatore:

$$S_p^2 = \frac{\sum_{i=1}^{n_X}(X_i - \bar{X})^2 + \sum_{i=1}^{n_Y}(Y_i - \bar{Y})^2}{n_X + n_Y - 2}$$

Questo stimatore viene chiamato VARIANZA CAMPIONARIA POOLED, coinvolge entrambi i campioni ed è non distorno per δ^2. Può essere anche espresso nella seguente maniera:

$$S_p^2 = \frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2}$$

Dove:
- S_x^2 = varianza campionaria di X
- S_y^2 = varianza campionaria di Y
- n_x = ampiezza del primo sottogruppo
- n_y = ampiezza del secondo sottogruppo

Sotto H_0, la statistica test ha distribuzione T di Student con $n_x + n_y - 2$ gradi di libertà e la sua espressione è:

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_p^2}{n_X} + \frac{S_p^2}{n_Y}}}$$

Regione di rifiuto di livello alfa:

$$R_\alpha: T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_p^2}{n_X} + \frac{s_p^2}{n_Y}}} > t_\alpha^{n_X + n_Y - 2}$$

Si calcola t osservato,

$$t_{oss} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_p^2}{n_X} + \frac{s_p^2}{n_Y}}}$$

Se $t_{oss} > t_\alpha^{nx+ny-2}$, si rifiuta H_0 a livello alfa.
Se $t_{oss} \leq t_\alpha^{nx+ny-2}$, non si rifiuta H_0 a livello alfa.

Oppure, $p - value = P(T > t_{oss} | \mu_x = \mu_y)$
- Se p-value<α si rifiuta l'ipotesi nulla
- Se p-value ≥α NON si rifiuta l'ipotesi nulla

b) Test unilaterale a sinistra

$H_0: \mu_X = \mu_Y$ $(cioè\ \ \mu_X - \mu_Y = 0)$
$H_1: \mu_X < \mu_Y$ $(cioè\ \ \mu_X - \mu_Y < 0)$ La statistica test è sempre la stessa, la regione di rifiuto invece è:

$$R_\alpha: T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_p^2}{n_X} + \frac{s_p^2}{n_Y}}} < -t_\alpha^{n_X + n_Y - 2}$$

Si calcola t osservato,

$$t_{oss} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_p^2}{n_X} + \frac{s_p^2}{n_Y}}}$$

Se $t_{oss} < -t_\alpha^{nx+ny-2}$, si rifiuta H_0 a livello alfa.
Se $t_{oss} \geq -t_\alpha^{nx+ny-2}$, non si rifiuta H_0 a livello alfa.

Oppure, $p - value = P(T > t_{oss} | \mu_x = \mu_y)$
- Se p-value<α si rifiuta l'ipotesi nulla
- Se p-value ≥α NON si rifiuta l'ipotesi nulla

c) Test bilaterale

$H_0: \mu_X = \mu_Y \ \ (cioè \ \ \mu_X - \mu_Y = 0)$
$H_1: \mu_X \neq \mu_Y \ \ (cioè \ \ \mu_X - \mu_Y \neq 0)$

La statistica test è sempre la stessa, mentre la regione di rifiuto è:

$$R_\alpha: |T| = \left| \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_p^2}{n_X} + \frac{s_p^2}{n_Y}}} \right| > t_{\frac{\alpha}{2}}^{n_X + n_Y - 2}$$

Si calcola t osservato,

$$t_{oss} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_p^2}{n_X} + \frac{s_p^2}{n_Y}}}$$

Se $|t_{oss}| < -t_{\frac{\alpha}{2}}^{nx+ny-2}$, si rifiuta H_0 a livello alfa.

Se $|t_{oss}| \geq -t_{\frac{\alpha}{2}}^{nx+ny-2}$, non si rifiuta H_0 a livello alfa.

Oppure, $p - value = 2 * P(T > |t_{oss}| | \mu_x = \mu_y)$

- Se p-value<α si rifiuta l'ipotesi nulla
- Se p-value ≥α NON si rifiuta l'ipotesi nulla

USO DI R PER LA VERIFICA DI IPOTESI SULLA DIFFERENZA DI MEDIE:
Funzione t.test: $t.test(V_1 \sim V_2, data = nomedeldataframe, var.equal = T, alternative = \text{two sided})$

V_1 = variabile di interesse per il test.
V_2 = variabile che divide la popolazione nei due sottogruppi.

L'argomento alternative specifica il sistema di ipotesi del test:

a. alternative="greater" effettua il test unilaterale con coda a dx (verificare se la media è aumentata),
b. alternative="less" effettua il test unilaterale con coda a sx (verificare se la media è diminuita),
c. alternative="two.sided" effettua il test bilaterale (verificare se la media è cambiata).

Se l'argomento alternative è omesso, il test di default è quello bilaterale.

II. Popolazioni ARBITRARIE con varianze incognite e uguali nei due campioni (sufficientemente grandi, con n>40)

$X_1, X_2, \dots, X_{n_X} \ i.i.d \sim N(\mu_X, \sigma^2)$
$Y_1, Y_2, \dots, Y_{n_Y} \ i.i.d \sim N(\mu_Y, \sigma^2)$.

con n_x e n_y sufficientemente grandi.

Sotto H_0, la statistica test ha distribuzione approssimativamente normale standard per il teorema centrale del limite e ha espressione:

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_p^2}{n_x} + \frac{S_p^2}{n_y}}}$$

a) Test unilaterale a destra

$H_0: \mu = \mu_0$
$H_1: \mu > \mu_0$

Regione di rifiuto:

$$R_\alpha: Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_p^2}{n_x} + \frac{S_p^2}{n_y}}} > z_\alpha$$

P-value: $p - value = P(Z > z_{oss} | \mu_x = \mu_y)$

b) Test unilaterale a sinistra

H_0: $\mu = \mu_0$
H_1: $\mu < \mu_0$

Regione di rifiuto:

$$R_\alpha: Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_p^2}{n_x} + \frac{S_p^2}{n_y}}} < -z_\alpha$$

P-value: $p - value = P(Z < z_{oss} | \mu_x = \mu_y)$

c) Test bilaterale

H_0: $\mu = \mu_0$
H_1: $\mu \neq \mu_0$

Regione di rifiuto:

$$R_\alpha: |Z| = \left| \frac{\bar{X} - \mu_0}{\delta / \sqrt{n}} \right| > z_{\frac{\alpha}{2}}$$

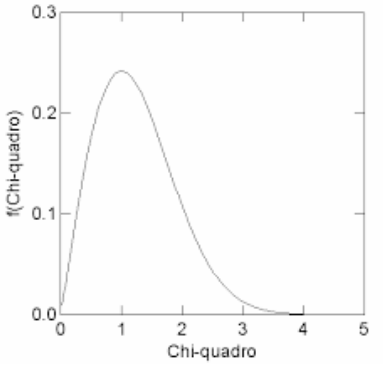
P-value: $p - value = 2 * P(Z > |z_{oss}| | \mu_x = \mu_y)$

TEST DI INDIPENDENZA "CHI-QUADRATO"

Questo tipo di test, basato su una distribuzione "chi-quadrato", serve per verificare l'associazione tra due variabili qualitative.

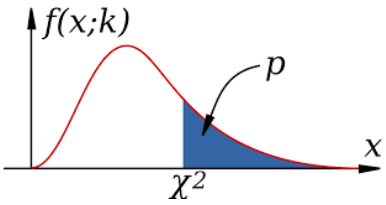
Distribuzione "chi-quadrato"

→ è una distribuzione continua, quindi con una densità, e caratterizzata da un parametro **k**, intero e positivo, denominato "**gradi di libertà**".



La sua densità è obliqua a destra e positiva solo sull'asse positivo delle ascisse (vedi grafico).

$X \sim X^2(k)$ → la variabile X ha distribuzione "chi-quadrato" con k gradi di libertà
$X^2_{\alpha,k}$ → quantile di ordine $(1-\alpha)$ della distribuzione "chi-quadrato" con k gradi di libertà



USO DI R PER LO STUDIO DELLA DISTRIBUZIONE CHI QUADRATO:
pchisq (1-α, k) → quantile di ordine (1-α) di $X \sim X^2(k)$
qchisq (z, k) → probabilità P (X< z)

Test "chi-quadrato" di indipendenza

Obbiettivo: verificare la presenza di associazione tra due caratteri **qualitativi** (è possibile considerare due caratteri quantitativi, ma con poche modalità).
Si consideri una popolazione sufficientemente grande (n>40), con due caratteri X e Y distinti. Si vuole verificare se esiste un'associazione tra il carattere X e il carattere Y.

H_0: le due variabili X e Y sono indipendenti
H_1: le due variabili X e Y sono associate

Prendendo in considerazione un campione della popolazione, bisogna verificare che i dati del campione contrastino a sufficienza l'ipotesi nulla tanto da portare al suo rifiuto. In altre parole, se le **frequenze relative congiunte** coincidono con il prodotto delle corrispondenti **frequenze marginali**, allora i due caratteri sono indipendenti.

Considerando la realizzazione di un campione e due caratteri X e Y, per prima cosa bisogna costruire la tabella a doppia entrata con le frequenze assolute → tabella delle O_{ij}.
Supponendo che le due variabili siano indipendenti, si calcolano le frequenze congiunte (relative) tramite il prodotto tra le frequenze marginali.
A questo punto, bisogna moltiplicare i valori della tabella (ossia le frequenze congiunte relative) per n, in modo da ottenere la tabella delle frequenze attese (chiamate anche frequenze assolute congiunte) → tabella delle E_{ij}.
Il test si basa sul confronto tra O_{ij} e E_{ij}.

- O_{ij}= Frequenze assolute congiunte osservate
- E_{ij}= Frequenze assolute congiunte attese

r= Numero delle righe (modalità della variabile X)
c= Numero delle colonne (modalità della variabile Y)

R_i= Frequenze assolute marginali delle righe
C_j= Frequenze assolute marginali delle colonne

X\Y	Y_1	Y_2	…	Y_j	
X_1	O_{11}	O_{12}	…	O_{1j}	R_1
X_2	O_{21}	O_{22}	…	O_{2j}	R_2
…	…	…	…	…	…
X_I	O_{i1}	O_{i2}	…	O_{ij}	R_I
	C_1	C_2	…	C_j	

X\Y	Y_1	Y_2	…	Y_j	
X_1	E_{11}	E_{12}	…	E_{1j}	R_1
X_2	E_{21}	E_{22}	…	E_{2j}	R_2
…	…	…	…	…	…
X_I	E_{i1}	E_{i2}	…	E_{ij}	R_I
	C_1	C_2	…	C_j	

$$E_{ij} = \frac{R_i \times C_j}{n}$$

Per valutare l'associazione tra le due variabili bisogna misurare la distanza tra le O_{ij} e le E_{ij}, ossia tra le frequenze osservate e le frequenze attese, attraverso la formula della statistica test X^2.

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

In particolare, tanto più è grande il valore di X^2 tanto più siamo lontani dallo stato di indipendenza, quindi c'è associazione. → X^2 ha approssimativamente distribuzione "chi-quadrato" con **(r-1)∗(c-1)** gradi di libertà, sotto H_0 (ipotesi nulla).
A questo punto, come già visto precedentemente, si calcola la regione di rifiuto della statistica test osservato e si confrontano i risultati ottenuti.

$$R_\alpha: X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} > X^2_{\alpha(r-1)(c-1)}$$

- Se $X^2_{oss} > X^2_{\alpha(r-1)*(c-1)}$ allora si rifiuta l'ipotesi nulla a livello α => è possibile affermare l'esistenza di un'associazione tra le due variabili.
- Se $X^2_{oss} \leq X^2_{\alpha(r-1)*(c-1)}$ allora non si rifiuta l'ipotesi nulla a livello α => è possibile affermare l'esistenza di uno stato di indipendenza tra le due variabili.

Sarebbe possibile ottenere lo stesso risultato anche tramite l'uso del P-value.

$$P - value = P(X_2 > X_{2\,oss} | H_0)$$

- Se P-value < α allora si rifiuta l'ipotesi nulla.
- Se P-value ≥ α allora non si rifiuta l'ipotesi nulla.

USO DI R PER IL P-VALUE NELLA DISTRIBUZIONE CHI-QUADRATO:
formula p-value: $1 - pchisq(X_{2\,oss}, k)$
tabella: $table(varX, varY)$
$summary(table(varX, varY))$

MODELLI DI REGRESSIONE

Un modello di regressione studia la DIPENDENZA, nell'ambito della popolazione, di una variabile quantitativa da altre variabili quantitative o qualitative.

MODELLO DI REGRESSIONE LINEARE SEMPLICE (2 variabili quantitative)

Data una variabile quantitativa, spiega la sua variabilità nella popolazione tramite un'altra variabile quantitativa.

→ Y = variabile quantitativa dipendente (o risposta)
→ X = variabile quantitativa indipendente (o esplicativa, o covariata)

La variabile X spiega e determina la variabile Y.
Obiettivi dell'analisi di regressione lineare:

1. **Conoscitivo**: stabilire se X ha effetto su Y e, in caso affermativo, studiare le caratteristiche di tale dipendenza.
2. **Esplicativo**: spiegare come X influenza Y (cioè, spiegare la variabilità di Y tramite X).
3. **Previsivo**: prevedere/stimare Y (o la sua media) in corrispondenza di un dato X.

Formulazione del modello di regressione lineare semplice:
con il modello di regressione, si attribuisce un valore alla variabile X (ottenendo la sua realizzazione, x) per spiegare la variabile dipendente Y.

EQUAZIONE TEORICA:

$$Y = \beta_0 + \beta_1 * x + \varepsilon$$

– Y è variabile aleatoria.
– x è la realizzazione della variabile esplicativa X.
– β_0 e β_1 sono due parametri incogniti: β_0 è l'intercetta e β_1 è il coefficiente angolare della retta che descrive la relazione tra Y e X.
Se $\beta_1 > 0$, Y aumenta all'aumentare di X, mentre se $\beta_1 < 0$, Y diminuisce all'aumentare di X.
– ε è l'errore: è una variabile aleatoria che indica tutti i fattori che, al di fuori di X, hanno effetto su Y. Se l'errore è nullo, allora Y dipende esclusivamente da X.

La variabile aleatoria Y è quindi la somma di due quantità:

- di una parte deterministica, non aleatoria, dipendente da X
- di una parte aleatoria che comprende tutti gli elementi che, oltre a X, possono avere effetto su Y.

Nel modello, i parametri incogniti β_0 e β_1 sono gli elementi più rilevanti (soprattutto β_1) perché permettono di spiegare e prevedere Y sulla base di X.

ANALISI DEL MODELLO DI REGRESSIONE LINEARE SEMPLICE

Il punto di partenza è un campione estratto dalla popolazione comporto da una n-upla di coppie, per ciascuna delle quali vale l'equazione $Y_n = \beta_0 + \beta_1 * x_n + \varepsilon_n$ e ciascuna comporta da una variabile indipendente x e una dipendente Y.

Per poter stimare i parametri incogniti, bisogna fissare delle assunzioni alla base del modello sulle due variabili aleatorie Y e ε (fare assunzioni sull'errore implica automaticamente fare assunzioni anche sulla Y) → ASSUNZIONI (dette anche IPOTESI DEBOLI o IPOTESI STANDARD):

1. Formulazione del modello: $Y_i = \beta_0 + \beta_1 * x_i + \varepsilon_i$

2. Le realizzazioni $x_1, x_2, \dots, x_n$ non sono aleatorie.

3. $E(\varepsilon_i) = 0, \forall_i = 1,2,\dots,n$
   Tutte le ε_i hanno valore atteso nullo à in media, gli errori sono pari a zero.
   Per le proprietà del valore atteso e per la relazione tra l'errore e Y, affermare che $E(\varepsilon_i) = 0$ equivale a dire che $E(Y_i) = \beta_0 + \beta_1 * x_i$, perché Y è una trasformazione lineare di ε.
   In media, il valore di Y su ogni unità campionaria è esattamente una funzione lineare del valore di x sulla stessa unità campionaria.
   Il modello di regressione dice che la variabile aleatoria Y_i è data dalla somma di una quantità deterministica ($\beta_0 + \beta_1 * x$) e di un errore aleatorio. Pertanto, date le assunzioni, il modello dice anche che in media Y_i è dato soltanto dalla relazione $\beta_0 + \beta_1 * x$.
   → Y dipende sia dalla parte deterministica dipendente da x, sia dall'errore, ma, mediamente, dipende soltanto dalla parte deterministica.

4. $Var(\varepsilon_i) = \delta^2, \forall_i = 1,2,\dots,n$
   PROPRIETÀ DI OMOSCHEDASTICITÀ (= omogeneità della varianza)
   Tutte le ε_i hanno varianza pari a sigma al quadrato à la varianza non dipende da i: tutti gli errori hanno la stessa dispersione.
   La varianza dell'errore, e, per le proprietà della trasformazione lineare, quindi anche della variabile Y, è omogenea.

5. $Cov(\varepsilon_i, \varepsilon_j) = 0, \forall_i \neq j$
   In ogni coppia le covarianze (e quindi anche il coefficiente di correlazione lineare) sono nulle. → le variabili aleatorie Y_i non sono correlate linearmente.

STIMA DEI PARAMETRI

La stima avviene sulla base dei dati campionari e viene fatta utilizzando il METODO DEI MINIMI QUADRATI, che consiste nella minimizzazione di una somma di quadrati: con questo metodo si cerca, tra tutte le rette del piano, quella che più si avvicina ai dati osservati nel campione.
→ Si confronta il valore osservato nel campione (y_i) con il valore appartenente alla retta ($\beta_0 + \beta_1 * x_i$). Poi si somma la differenza tra questi due valori, per ogni coppia (y_i, x_i) del campione, elevata al quadrato.
Dati i valori osservati nel campione, il metodo dei minimi quadrati consiste nel trovare la retta che passa più vicina al maggior numero possibile di tali punti. In altre parole, il metodo dei minimi quadrati consiste nel trovare la retta che meglio si adatta ai punti campionari.
Risolvendo questo problema di minimizzazione si trovano le stime b_0 e b_1 (STIME DEI MINIMI QUADRATI):

$$b_0 = \bar{y} - b_1 * \bar{x}$$

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) * (y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{XY}}{S_X^2} = r * \frac{S_Y}{S_X}$$

Dove S_{XY} = covarianza; r = coefficiente di correlazione lineare; S_X = deviazione standard di X_i; S_Y = deviazione standard di Y_i.
Dato che S_X e S_Y sono positivi, il segno di b1 è determinato dal segno di r:

- se $r > 0 \Rightarrow b_1 > 0$
- se $r < 0 \Rightarrow b_1 < 0$

EQUAZIONE STIMATA:

$$\hat{Y} = b_0 + b_1 X$$

Questa equazione descrive, a livello di stime, la relazione tra la variabile X e la media di Y ($E(y) = \beta_0 + \beta_1 x$)
Il parametro b_1 è la stima dell'incremento medio di Y associato ad un incremento unitario di X.

SCOMPOSIZIONE DELLA DEVIANZA (variabilità) TOTALE:

È un'uguaglianza per la quale la devianza totale è pari alla somma della devianza residua e della devianza spiegata.

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

Dove $\hat{y}_i$ = punti sulla retta di regressione con ascissa pari agli x_i delle diverse unità.

- DEVIANZA TOTALE = $\sum_{i=1}^{n}(y_i - \bar{y})^2$
  È la somma dei quadrati delle differenze tra le osservazioni e la loro media. Rappresenta la variabilità totale di Y.

- DEVIANZA RESIDUA / NON SPIEGATA = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
  È la somma dei quadrati delle differenze tra osservazioni e punti appartenenti alla retta di regressione. Non è dovuta all'effetto della variabile X su Y, ma è dovuta al fatto che il modello non è perfetto.
  La devianza residua misura la variabilità dei dati attorno alla retta di regressione → indica quanto i dati osservati si discostano dalla retta.

- DEVIANZA SPIEGATA = $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
  È la somma dei quadrati delle differenze tra i punti appartenenti alla retta di regressione e la loro media. La devianza spiegata dipende dall'effetto di X su Y (in altre parole, dipende da quanto X influenza Y).
  La devianza spiegata assume valori elevati se b_1 è grande (in modulo):

  $$devianza\ spiegata = {b_1}^2 * \sum_{i=1}^{n}(x_i - \bar{x})^2$$

Se la devianza spiegata (distanza tra la retta e $\bar{y}$) è molto alta e la devianza residua (distanza tra i punti attorno alla retta) è bassa, la variabile X spiega Y in modo consistente, quindi il modello di regressione è buono.

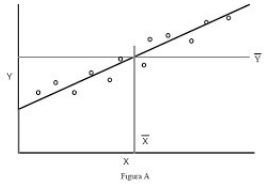


Figura A

Se la devianza spiegata è bassa e la devianza residua è elevata, la variabile X spiega poco della variabilità di Y, quindi il modello di regressione è poco utile.

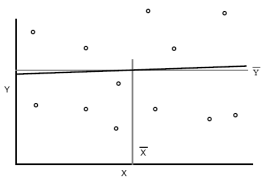


Figura B

CASI ESTREMI:

- Se tutta la devianza è spiegata, la devianza residua è nulla: $y_i = \hat{y}_i$
  Tutti i punti osservati giacciono lungo la retta di regressione e sono perfettamente allineati → tutta la variabilità di Y dipende esclusivamente da X.

- Tutta la devianza è residua, la devianza spiegata è nulla: $\hat{y}_i - \bar{y}$
  Dato che $\bar{y}$ è una costante, anche $\hat{y}_i$ è costante ed è una retta orizzontale in corrispondenza di $\bar{y}$ → non c'è alcuna correlazione tra Y e X (y_i non varia al variare di X).

COEFFICIENTE DI DETERMINAZIONE R^2

$$R^2 = \frac{\sum_{1=i}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{1=i}^{n}(y_i - \bar{y})^2} = \frac{Devianza\ spiegata}{Devianza\ totale}$$

Il coefficiente R^2 misura la capacità esplicativa del modello. In altre parole, R^2 misura la proporzione di variabilità di Y spiegata da X.

Proprietà di R^2:
- $R^2 \in [0, 1]$
- $R^2 = 1$, se la devianza residua è nulla ⇔ tutta la variabilità di Y è spiegata da X e la retta di regressione passa per i punti rilevati nel campione.
- $R^2 = 0$, se la devianza spiegata è nulla ⇔ la retta di regressione è orizzontale.
- $R^2 = r^2 = \left(\frac{S_{xy}}{S_{xy}}\right)^2$ L'indice R^2 coincide con il quadrato del coefficiente di correlazione lineare calcolato sui dati campionari. Maggiore è r, più X e Y sono correlate linearmente (quindi, il modello di regressione è buono).

ASSUNZIONI FORTI SUL MODELLO: ipotesi deboli + assunzione di normalità.
Assunzione di normalità:
$\mathcal{E}_i \sim N(0, \sigma^2)\ \ per\ ogni\ i = 1, 2, \dots n$
$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)\ per\ ogni\ i = 1, 2, \dots n$

INTERVALLI DI CONFIDENZA

Le stime b_0 e b_1, rispettivamente di β_0 e β_1, sono le realizzazioni degli stimatori corrispondenti B_0 e B_1.
Riassumendo,
- β_0 e β_1 = parametri incogniti. Rappresentano, rispettivamente, l'intercetta e il coefficiente angolare della retta di regressione.
- B_0 e B_1 = stimatori dei parametri β_0 e β_1. Sono corretti (non distorti): $E(B_1) = \beta_1$; $E(B_0) = \beta_0$
- b_0 e b_1 = stime dei parametri β_0 e β_1. Sono le realizzazioni degli stimatori.

Distribuzioni degli stimatori B_0 e B_1

Sotto le assunzioni forti,

$$B_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)\right) \qquad B_1 \sim N\left(\beta_1, \sigma^2\left(\frac{1}{n} + \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)\right)$$

INTERVALLO DI CONFIDENZA PER β_1

Equazione del modello: $\mathcal{E}_i = Y_i - (\beta_0 + \beta_1 x_i) = Y_i - E(Y_i)$

È necessario stimare σ^2, incognita, presente nella formulazione della varianza dello stimatore B_1.
Stima di σ^2:

$$s_e^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n - 2}$$

Stimatore di $s_e^2 = S_e^2$

A questo punto, si può stimare la varianza di β_1:

$$s_{b1}^2 = \frac{s_e^2}{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2}$$

Stimatore di s_{b1}^2 = S_{b1}^2

Studentizzazione di B_1: $T = \frac{B_1 - \beta_1}{S_{b1}}$
T ha distribuzione T di Student con n-2 gradi di libertà

Intervallo di confidenza di livello 1-alfa:

$$\left(b_1 - t_{\frac{\alpha}{2}}^{n-2} s_{b1}, b_1 + t_{\frac{\alpha}{2}}^{n-2} s_{b1}\right)$$

C'è fiducia pari a 1-alfa che il vero valore del parametro β_1 appartiene all'intervallo.
L'intervallo di confidenza può essere espresso anche come:

$$(stima - ME, stima + ME) = (stima - (quantile * SE), stima + (quantile * SE))$$

Standard error:

$$SE = s_{b1} = \sqrt{\frac{s_e^2}{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2}}$$

VERIFICA DI IPOTESI PER β_1

$H_0: \beta_1 = 0$ mancanza di effetto di x su y
$H_1: \beta_1 \neq 0$ presenza di effetto di x su y

Sotto l'ipotesi nulla si ottiene questa statistica test: $T = \frac{B_1}{S_{b1}}$

Regione di rifiuto: $R_\alpha: |T| = \left|\frac{B_1}{S_{b1}}\right| \geq t_{\frac{\alpha}{2}}^{n-2}$

Se $|t_{oss}| = \left|\frac{b_1}{S_{b1}}\right| \geq t_{\frac{\alpha}{2}}^{n-2}$, si rifiuta H_0 a livello alfa.
Se $|t_{oss}| = \left|\frac{b_1}{S_{b1}}\right| < t_{\frac{\alpha}{2}}^{n-2}$, non si rifiuta H_0 a livello alfa.

Oppure, utilizzando il P-value:

$$p - value = 2P(T > |t_{oss}||H_0)$$

INTERVALLI DI PREVISIONE

Uno degli obiettivi da modello di regressione è stimare Y o la sua media (ovvero E(Y)) in corrispondenza a valori fissati di x. Di conseguenza, bisogna prevedere due possibilità:

1) Stima di Y (su una singola unità con $x=x_o$): $Y = \beta_0 + \beta_1 + \varepsilon$
2) Stima di E(Y) (sulle unità della popolazione per cui $x=x_o$): $E(Y) = \beta_0 + \beta_1$

La stima del primo caso è meno precisa della seconda poiché è coinvolta anche l'incertezza legata all'errore ε. → L'errore ε costituisce la differenza tra Y ed E(Y)
Tuttavia, dal momento che che E(ε)=0 , le stime coincidono nei due casi ed in corrispondenza del valore $x=x_o$ sono date da: $\hat{y}_0 = b_0 + b_1 x_0$

$\hat{y}_0$ rappresenta la stima sia di Y sia di E(Y), ma i livelli di incertezza legati alle due stime sono diversi. Nel caso in cui si stima la singola unità, infatti, il livello di incertezza è superiore.

INTERVALLO DI PREVISIONE per E(Y)

di livello 1- α corrispondente a $x=x_o$

$$\left(\hat{y}_0 - t_{\frac{\alpha}{2}}^{n-2} s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}, \hat{y}_0 + t_{\frac{\alpha}{2}}^{n-2} s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right)$$

INTERVALLO DI PREVISIONE per Y

di livello 1- α corrispondente a $x=x_o$

$$\left(\hat{y}_0 - t_{\frac{\alpha}{2}}^{n-2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}, \hat{y}_0 + t_{\frac{\alpha}{2}}^{n-2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right)$$

La differenza tra i due intervalli sta nella presenza del +1, che riduce notevolmente il grado di certezza dell'intervallo ⇒ L'intervallo di previsione per la media avrà quindi lunghezza sempre inferiore rispetto all'intervallo di previsione per l'unità in corrispondenza di $x=x_o$.
Inoltre, la lunghezza di entrambi gli intervalli aumenta al crescere di 1- α (a parità di altri elementi).

Non è opportuno stimare Y o E(Y) in corrispondenza di valori di x che sono al di fuori del range del campione, perché ciò porterebbe a risultati inattendibili.

USO DI R PER IL MODELLO DI REGRESSIONE LINEARE SEMPLICE:

Per costruire un modello di regressione lineare la funzione da utilizzare è **lm** e si procede così:

$$nome\ oggetto <- \boldsymbol{lm}(Y \sim x, data = nome\ dataframe)$$

Le due variabili Y e X non sono precedute dal nome del dataset, perché questo viene indicato indicato nell'argomento data.
Tra gli output della fuzione c'è la sezione "Coefficients", dove

- Estimate = stima del coefficiente (b_1)
- Std. Error = standard error di b_1 (s_{b1})
- Il valore della statistica test t - value è dato da $b1/s_{b1}$
- La riga Intercept riporta le stesse quantità in relazione non a β_1 ma a β_0

Esempio:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 115.6295    19.4182   5.955 1.54e-07 ***
budget        0.8364     0.2564   3.263  0.00184 **
```

Tra gli output, la quantità indicata con Residual Standard Error è la radice quadrata di Se della stima della varianza σ^2 degli errori.

COSTRUIRE L'INTERVALLO DI CONFIDENZA:

Funzione **confint** applicata al modello: $confint(modello)$

Il livello di default è $1 - a = 0.95$
Se si vuole usare un livello differente bisogna inserire l'argomento level:
$confint(modello, level = 0.9)$

L'output della funzione confint è l'intervallo di confidenza.

RICAVARE LA SCOMPOSIZIONE DELLA DEVIANZA TOTALE:

Funzione **anova** applicata al modello: $anova(modello)$

Esempio:

```
            Df Sum Sq Mean Sq F value   Pr(>F)
budget       1  76687   76687  10.644 0.001838 **
Residuals   59 425068    7205
```

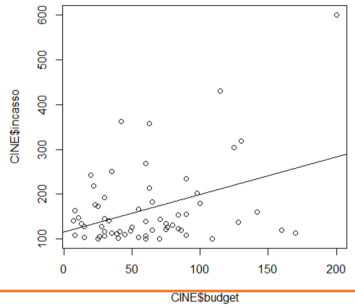
La devianza spiegata è 76687, quella residua 425068. Sommandole otteniamo la devianza totale

INSERIRE LA RETTA DI REGRESSIONE NELLO SCATTERPLOT:

Funzione **abline** applicata al modello.

Per applicare la funzione prima bisogna costruire il grafico di dispersione (scatterplot). Poi, si applica $abline(modello)$

Esempio: plot(CINE$budget,CINE$incasso)



CALCOLARE L'INTERVALLO DI PREVISIONE:

Funzione **predict**. Sintassi:

$predict(modello, newdata = data.frame(variabile = valore))$

Per calcolare l'intervallo di previsione di Y, bisogna aggiungere l'argomento interval="prediction", mentre per calcolare l'intervallo di previsione per E(Y) si ha interval="confidence"

La funzione predict calcola di default la previsione al 95%, ma per modificare il livello si può aggiungere l'argomento level.

Esempio: stimare l'incasso di un film con budget uguale a 50 e l'incasso medio dei film con budget uguale a 50.

- Per calcolare l'intervallo di previsione di Y al 95%:
  predict(modello,newdata=data.frame(budget=50),interval="prediction")
- Per calcolare l'intervallo di previsione per E(Y) al 95%:
  predict(modello,newdata=data.frame(budget=50),interval="confidence")

```
> predict(modello,newdata=data.frame(budget=50),interval="prediction")
       fit       lwr      upr
1 157.4498 -13.90566 328.8053

> predict(modello,newdata=data.frame(budget=50),interval="confidence")
       fit      lwr      upr
1 157.4498 134.7382 180.1614
```

Dall'output si può vedere come l'intervallo di confidenza calcolato per la singola unità sia pressoché inutile a causa della sua lunghezza (-13.90566,328.8053)

MODELLO DI REGRESSIONE LINEARE MULTIPLA

Questo modello studia la dipendenza di una variabile quantitativa, Y, da più variabili esplicative, X_1, X_2...X_p.
È un'estensione della regressione lineare semplice.
Gli obbiettivi di questo modello, analoghi a quelli del modello semplice, sono:

- Conoscitivo
- Esplicativo
- Previsivo

Il punto di partenza dell'analisi è l'equazione teorica del modello, nella quale risulta chiaro che la variabile dipendente, Y, è composta dalla somma di una parte deterministica (β_0, β_1,... β_p e le variabili esplicative x) e una parte aleatoria (ε).

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

Come per il modello semplice, anche in questo caso sono necessarie assunzioni deboli (le quali riprendono e hanno uguale significato di quelle del modello con una sola variabile esplicativa):

⇒ $E(\varepsilon_i)=0$ per ogni i = 1, 2,...n
⇒ $Var(\varepsilon_i)=\sigma^2$ per ogni i = 1, 2,...n
⇒ $Cov(\varepsilon_i,\varepsilon_j)=0$ per ogni $i \neq j$

STIMA DEI PARAMENTRI β_0, β_1,... β_p

Come nel caso precedente, è necessario stimare questi parametri incogniti attraverso il metodo dei minimi quadrati.

$$\sum_{i=1}^{n} (y_1 - [\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}])^2$$

Dove y_1 rappresenta la realizzazione osservata nel campione preso in considerazione.

A questo punto, una volta stimati b_0, b_1...b_p, si ottiene l'equazione stimata del modello:

$$\hat{y} = b_0 + b_1 x_1 + \cdots + b_p x_p$$

La quale risulta essere una stima di:

$$E(Y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

INTERPRETAZIONE DELLE STIME

La stima b_j di β_J ci permette di interpretare l'effetto stimato di x_j sulla media di Y, in particolare b_j rappresenta la stima della variazione media di Y, associata ad un incremento unitario di x_j (tenendo fisse le altre variabili esplicative).

$$b_j = \hat{y}_{inc} - \hat{y}$$

Dove
$\hat{y}_{inc}$ = equazione stimata di x_j+1
$\hat{y}$ = equazione stimata di x_j

OSSERVAZIONI:

⇒ L'effetto stimato dell'incremento unitario vale sempre, non importa l'intervallo che si considera
⇒ L'effetto dell'incremento, considerando un intervallo pari a c, è pari a c-volte l'incremento unitario
⇒ L'effetto dell'incremento unitario di x_j è sempre lo stesso, qualunque siano le altre variabili esplicative fissate.

USO DI R PER IL MODELLO DI REGRESSIONE LINEARE MULTIPLA:

$$Mod <- lm(var\ dipentente \sim var\ esplicativa\ 1 + var\ esplicativa2 + \cdots, data = nome\ del\ dataframe)$$

Al fine di valutare la capacità esplicativa del modello, bisogna studiare la scomposizione della varianza totale (che riprende il modello lineare). La varianza totale, infatti, è la somma della varianza non spiegata / residua e della varianza spiegata.

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2$$

COEFFICIENTI DI DETERMINAZIONE R^2 E R^2-adjusted

Il coefficiente di determinazione R^2, analogamente al modello semplice, misura la capacità esplicativa dal modello ed è rappresentato dal rapporto tra la varianza spiegata e la varianza totale.

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Le sue proprietà sono identiche a quelle del modello con una sola variabile esplicativa e sono:

- $R^2 \in [0, 1]$
- $R^2 = 1$, se la devianza residua è nulla
- $R^2 = 0$, se la devianza spiegata è nulla
- Tanto più R^2 è alto, maggiore è la capacità esplicativa del modello

L'indicatore R^2 risulta inefficace per confrontare due modelli che hanno un numero diverso di variabili esplicative. Per questo motivo, viene introdotto un altro indice chiamato **R^2- adjusted** (R^2 corretto).

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

Questo indicatore possiede 3 proprietà rilevanti:

1. $R^2_{adj} \leq R^2$
2. R^2_{adj} è un compromesso tra la **capacità esplicativa** e la **semplicità** del modello.
3. R^2_{adj} è utile per calcolare modelli con un diverso numero di variabili esplicative, se e solo se queste variabili NON sono annidiate.

INTERVALLO DI CONFIDENZA PER β_j

A questo punto, per determinare il livello di confidenza, bisogna aggiungere alle assunzioni deboli anche l'ipotesi di normalità degli errori:

$$\varepsilon_i \sim N(0, \sigma^2) \qquad i = 1,2 \dots n$$

La struttura dell'intervallo, come detto in precedenza, è (stima - ME, stima + ME) e al crescere del livello di confidenza aumenta la lunghezza dell'intervallo.

VERIFICA DI IPOTESI

H_0: $B_j = 0$
H_1: $B_j \neq 0$
Quando l'ipotesi nulla viene rifiutata, si dice che x_j è significativa. Il test considerato è basato sulla statistica test:

$$T = \frac{B_j\,(stimatore\ di\ \beta_j)}{s_{b_j}(stimatore\ della\ varianza\ di\ \beta_j)}$$

Regione di rifiuto:

$$R_\alpha: |T| = \left|\frac{B_j}{s_{b_j}}\right| \geq t_{\frac{\alpha}{2}}^{n-p-1}$$

Se $|t_{oss}| \geq t_{\alpha/2}^{n-p-1}$ si rifiuta l'ipotesi nulla, al contrario non si rifiuta.

Oppure, $p - value = 2P(T > |t_{oss}||H_0$
Se il p-value è < α, allora si rifiuta l'ipotesi nulla

MODELLO DI REGRESSIONE LINEARE MULTIPLA CON VARIABILI ESPLICATIVE QUALITATIVE

- Variabili qualitative DICOTOMICHE (dummy)

Le variabili dummy hanno due possibili modalità e vengono codificate con i valori 0 e 1, dove 1 = presenza di una caratteristica, 0 = assenza di tale caratteristica. Formalmente, una variabile dummy viene codificata nel modello come se fosse una variabile quantitativa con modalità 0 e 1.

Una variabile dummy divide la popolazione in 2 sottogruppi.
L'effetto della variabile dummy si traduce nella differenza della media di Y tra i due sottogruppi, a parità dei valori delle altre variabili) ed è valutato sulla base della differenza della E(Y).

Equazione TEORICA del modello

$$Y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \varepsilon$$

Equazione STIMATA del modello

$$\hat{y} = b_0 + b_1 * x_1 + b_2 * x_2$$

Fissata x_1, si hanno due diverse equazioni:

1. Se x_1=1: $\widehat{y_1} = b_0 + b_1 * x_1 + b_2$
2. Se x_2= 0: $\widehat{y_0} = b_0 + b_1 * x_1 + b_2 * x_2$

$\widehat{y_1}$ e $\widehat{y_0}$ sono le stime di E(Y) nelle due diverse situazioni (a seconda del valore assunto dalla variabile dummy): le due rette di regressione sono diverse, pur avendo lo stesso coefficiente angolare, perché hanno diverse intercette (b_0 e b_0+b_2).

→ Differenza tra le due equazioni: $\widehat{y_1} - \widehat{y_0} = b_2$
Una variabile dicotomica dummy produce quindi 2 equazioni distinte, che differiscono tra loro per una costante che corrisponde alla stima differenza tra la stima dell'equazione (in media) nelle due diverse situazioni.

INTERVALLO DI CONFIDENZA

$$(b_j - t_{\alpha/2}^{n-p-1} * s_{bj}\,, b_j + t_{\alpha/2}^{n-p-1} * s_{bj})$$

VERIFICA DI IPOTESI

H_0: $B_j = 0$ → la variabile dummy non ha effetto su Y.
H_1: $B_j \neq 0$ → la variabile dummy ha effetto su Y.

Rifiutare H_0 significa affermare di aver trovato sufficiente evidenza empirica di una differenza, nella popolazione, sulla media di Y nelle due situazioni identificate dalla variabile dummy.

Regione di rifiuto: $R_\alpha: |T| = \left|\frac{B_j}{S_{Bj}}\right| \geq t_{\alpha/2}^{n-p-1}$

p-value: $p - value = 2 * P(T > |t_{oss}||H_0)$

- Variabili qualitative NON DICOTOMICHE

Hanno più di 2 modalità distinte → k>2
Per inserire nel modello una variabile qualitativa con k>2 modalità, occorre introdurre k-1 variabili dummy che identificano k-1 delle k modalità totali, escludendone quindi una.

Esempio: per spiegare la quotazione (Y) di un immobile, vengono utilizzate le variabili qualitative età e zona della città. La variabile zona ha 3 modalità distinte (k=3): A, B, C
Bisogna scegliere 2 (perché k-1=3-1=2) delle modalità di zona, escludendone una: ad esempio, si scelgono B e C escludendo A. In questo modo, x_B e x_C sono due variabili dummy, che possono assumere valori 0 e 1.

L'insieme delle combinazioni di queste due variabili, però, identifica tutte e tre le modalità della variabile zona:
se x_B=1, l'immobile è in zona B, mentre se x_B=0 l'immobile non è in zona B;
se x_C=1, l'immobile è in zona C, mentre se x_C=0 l'immobile non è in zona C;
se x_B=0 e x_C=0, l'immobile è in zona A (non è né in zona B né in zona C).
(La combinazione x_B=1 e x_C=1 non è possibile).

$\widehat{y_A}$, $\widehat{y_B}$, $\widehat{y_C}$ sono le stime di E(Y) nei tre diversi casi.

$\widehat{y_B} - \widehat{y_A} = b_B$ → è la stima del coefficiente della variabile dummy che identifica la zona B.
$\widehat{y_C} - \widehat{y_A} = b_C$ → è la stima del coefficiente della variabile dummy che identifica la zona C.
b_B e b_C hanno come riferimento la zona esclusa, A.

VERIFICA DI IPOTESI:

- Caso 1: verificare se c'è differenza, nella popolazione, tra le quotazioni medie degli immobili in zona B o in zona A → H_0: $\beta_B = 0$; H_1: $\beta_B \neq 0$
- Caso 1: verificare se c'è differenza, nella popolazione, tra le quotazioni medie degli immobili in zona C o in zona A → H_0: $\beta_c = 0$; H_1: $\beta_c \neq 0$

TEST F PARZIALE

Il test F parziale per la significatività di gruppi di variabili è una procedura che serve per confrontare tra loro, a livello di popolazione, due MODELLI ANNIDATI.
Due modelli sono annidati quando uno dei due contiene tutte le variabili esplicative dell'altro più variabili aggiuntive.
Il primo modello è chiamato **RIDOTTO**, indicato formalmente con $x_1, x_2, \dots, x_q$
Il secondo è chiamato modello **COMPLETO**, indicato formalmente con $x_1, x_2, \dots, x_q, x_{q+1}, x_{q+2}, \dots, x_p$

EQUAZIONE TEORICA:

- Per il modello ridotto

$$\boldsymbol{Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q + \varepsilon}$$

- Per il modello completo:

$$\boldsymbol{Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q + \beta_{q+1} x_{q+1} + \cdots + \beta_p x_p + \varepsilon}$$

PROBLEMA DI VERIFICA DI IPOTESI:

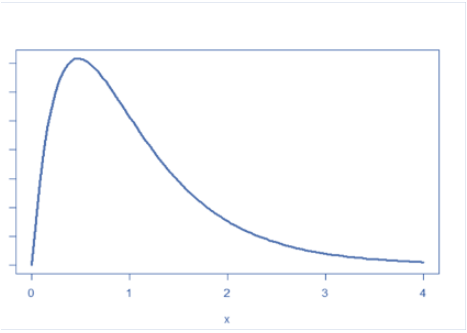
Verificare se le variabili che sono presenti nel modello completo e non in quello ridotto sono congiuntamente significative. Il problema di verifica di ipotesi è quindi:
H_0: $\beta_{q+1} = \beta_{q+2} = \cdots = \beta_q = 0$
H_1: almeno uno dei coefficienti $\beta_{q+1}, \beta_{q+2}, \dots, \beta_q$ è diverso da 0

L'ipotesi nulla consiste nell'affermare che, congiuntamente, le variabili presenti nel modello completo e non in quello ridotto hanno effetto su Y. Rifiutare l'ipotesi nulla significa affermare di aver trovato evidenza del fatto che le variabili $x_1, x_2, \dots, x_q$ hanno effetto su Y.

Il test da utilizzare per risolvere il problema è il test F parziale, che ha la seguente espressione:

$$F = \frac{(DEV.RES_{M.RID} - DEV.RES_{M.COMP})/(p-q)}{DEV.RES_{M.COMP}/(n-p-1)}$$

La statistica test F in (1) ha, sotto H_0, distribuzione di Fischer-Snedcor con $p-q$ e $n-p-1$ gradi di libertà.



Il test F ha regione di rifiuto così definita: $R_\alpha: F \geq F_{p-q,n-p-1,\alpha}$

$F_{p-q,n-p-1}$ è il quantile di ordine 1-α della distribuzione di Fischer con $p-q$ e $n-p-1$ gradi di libertà.

USO DI R PER CALCOLARE IL QUANTILE DELLA DISTRIBUZIONE DI FISCHER-SNEDCOR:
Funzione $qf(1-a, p-q, n-p-1)$

Per decidere se accettare o rifiutare l'ipotesi nulla, bisogna confrontare F_{oss} con $F_{p-q,\, n-p-1,\, a}$

se $F_{oss} \geq F_{p-q,n-p-1,\alpha}$ si rifiuta l'ipotesi nulla a livello α

se $F_{oss} < F_{p-q,n-p-1,\alpha}$ non si rifiuta l'ipotesi nulla a livello α

Oppure, $p - value = P(F > Foss \mid Ho)$
Si rifiuta l'ipotesi nulla se il p-value è minore di alfa.

ESEMPIO DELL'USO DI R PER IL TEST F PARZIALE:
Data frame Hospital.
Costruire due modelli:

- ModA con age, days, pat_cond, surgery
- ModB con age e pat_cond.

ModA è il modello completo e il ModB è quello ridotto.

MOD.A (con 5 variabili esplicative, per cui p=5):

$$Y = \beta_0 + \beta_{age} \cdot age + \beta_{days} \cdot days + \beta_{pc.average} \cdot pc.average + \beta_{pc.severe} \cdot pc.severe + \beta_{surgery} \cdot surgery + \varepsilon$$

MOD.B (con 3 variabili esplicative, per cui q=3):

$$Y = \beta_0 + \beta_{age} \cdot age + \beta_{pc.average} \cdot pc.average + \beta_{pc.severe} \cdot pc.severe + \varepsilon$$

Confrontare, attraverso un test di livello 0.01, se le variabili days e surgery sono congiuntamente significative.
Costruire i due modelli normalmente e poi applicare la funzione anova ai due modelli: anova(ModB, ModA). Bisogna mettere per primo il modello ridotto.
Output:

```
Model 1: hosp_exp ~ age + pat_cond
Model 2: hosp_exp ~ age + days + pat_cond + surgery
  Res.Df       RSS Df Sum of Sq      F    Pr(>F)
1    228 722637877
2    226 165863168  2 556774708 379.32 < 2.2e-16 ***
---
```

Nella prima colonna (Res.Df) ci sono i gradi di libertà di entrambi i modelli.
Nella seconda(RSS) c'è la devianza residua.
Nella terza (df) la differenza tra i gradi di libertà dei due modelli.
Nella quarta(SumofSq) la differenza tra le due devianze residue.
Nella quinta(F) c'è il valore del test F.
Nella sesta(Pr>F) c'è il p-value.

IN COLLABORAZIONE CON: